

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 Saroussi², Moritz T. Meyer¹, Nina Ivanova², Jacob M. Robertson^{1,2}, Rebecca Yue², Ru Zhang^{2,8}, Josep Vilarrasa-Blasi², Tyler M. Wittkopp^{2,4,9}, Silvia Ramundo⁵, Sean R. Blum², Audrey Goh¹, Matthew Laudon⁶, Tharan Srikumar¹, Paul A. Lefebvre⁶, Arthur R. Grossman² 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 function³. 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 study⁹. 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 chloroplast¹⁰, 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).

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

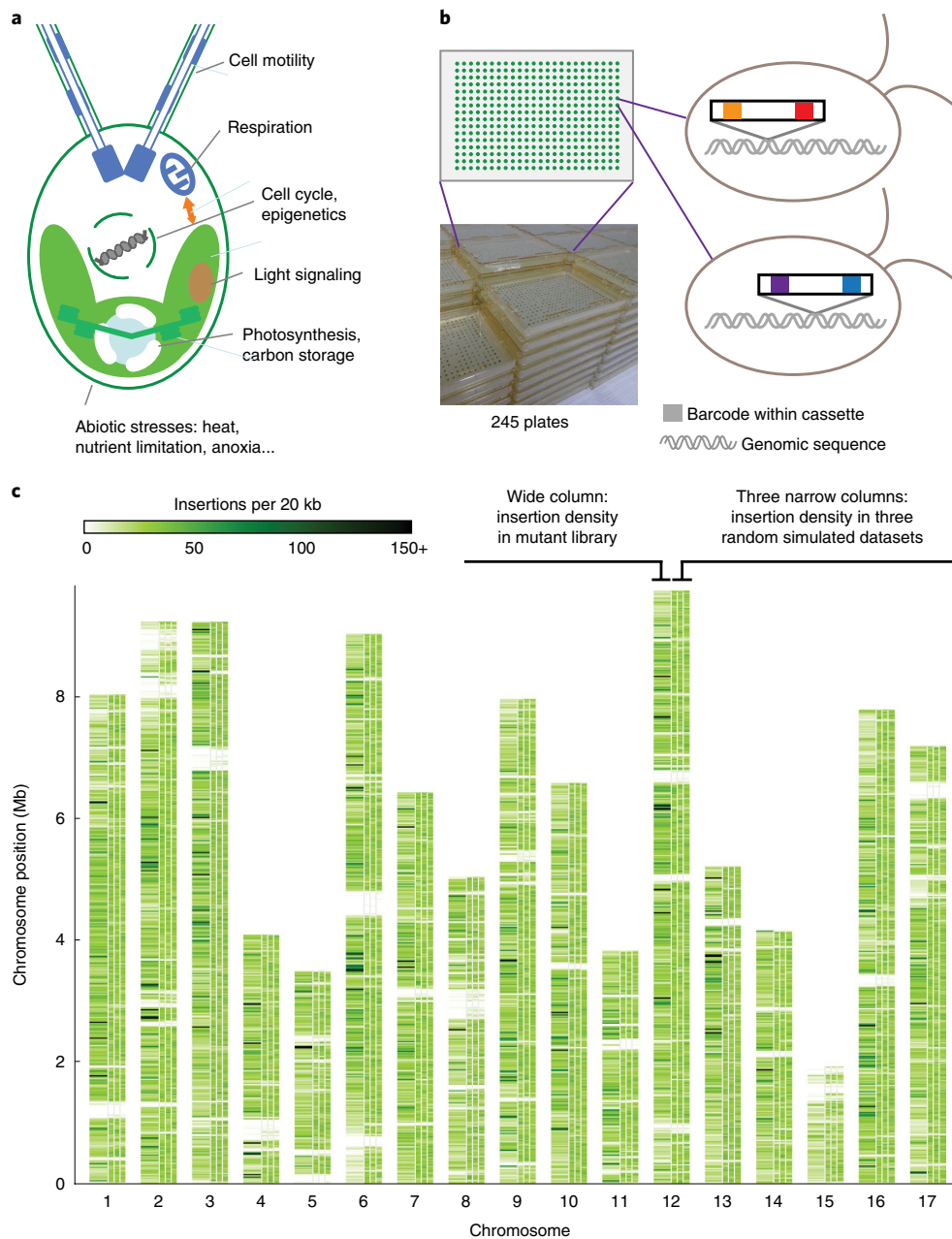


Fig. 1 | A genome-wide library of *Chlamydomonas* mutants was generated by random insertion of barcoded cassettes and mapping of insertion sites. **a, *Chlamydomonas* is used for studies of various cellular processes and organism–environment interactions. **b**, Our library contains 62,389 insertional 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

(CO₂) as the sole source of carbon and heterotrophically in the dark in Tris-acetate-phosphate (TAP) medium, where acetate provides fixed carbon and energy³ (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

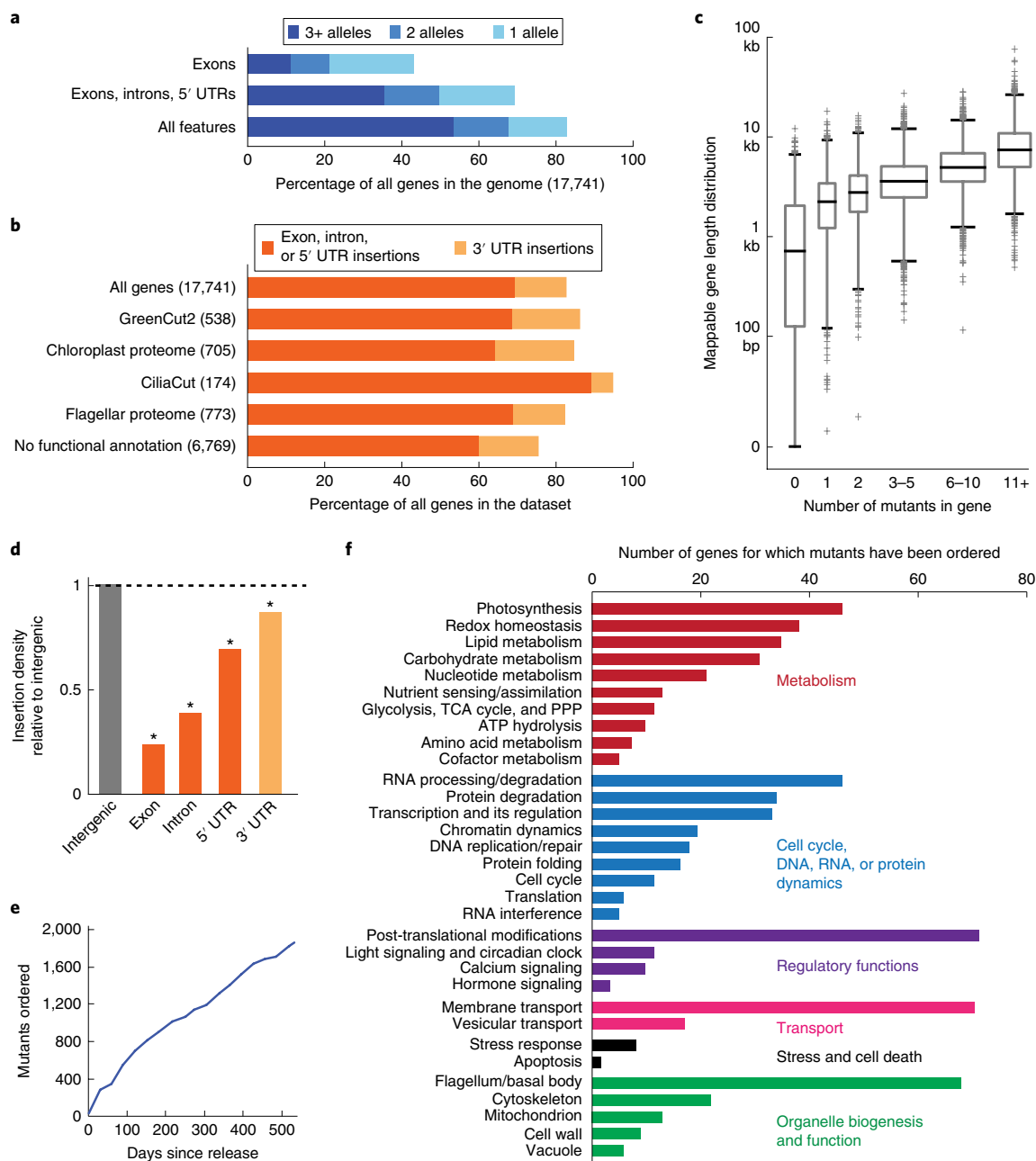


Fig. 2 | The library covers 83% of *Chlamydomonas* genes. **a**, 83% of all *Chlamydomonas* genes have one or more insertions in the library. **b**, In various functional groups, more than 75% of genes are represented by insertions in the library. **c**, The number of insertions per gene is roughly correlated 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 cluster¹³ ($P < 1 \times 10^{-11}$), are enriched for upregulation upon dark-to-light transitions¹⁴ ($P < 0.003$), and encode proteins enriched for predicted chloroplast localization ($P < 1 \times 10^{-8}$). As expected¹⁵, the candidate

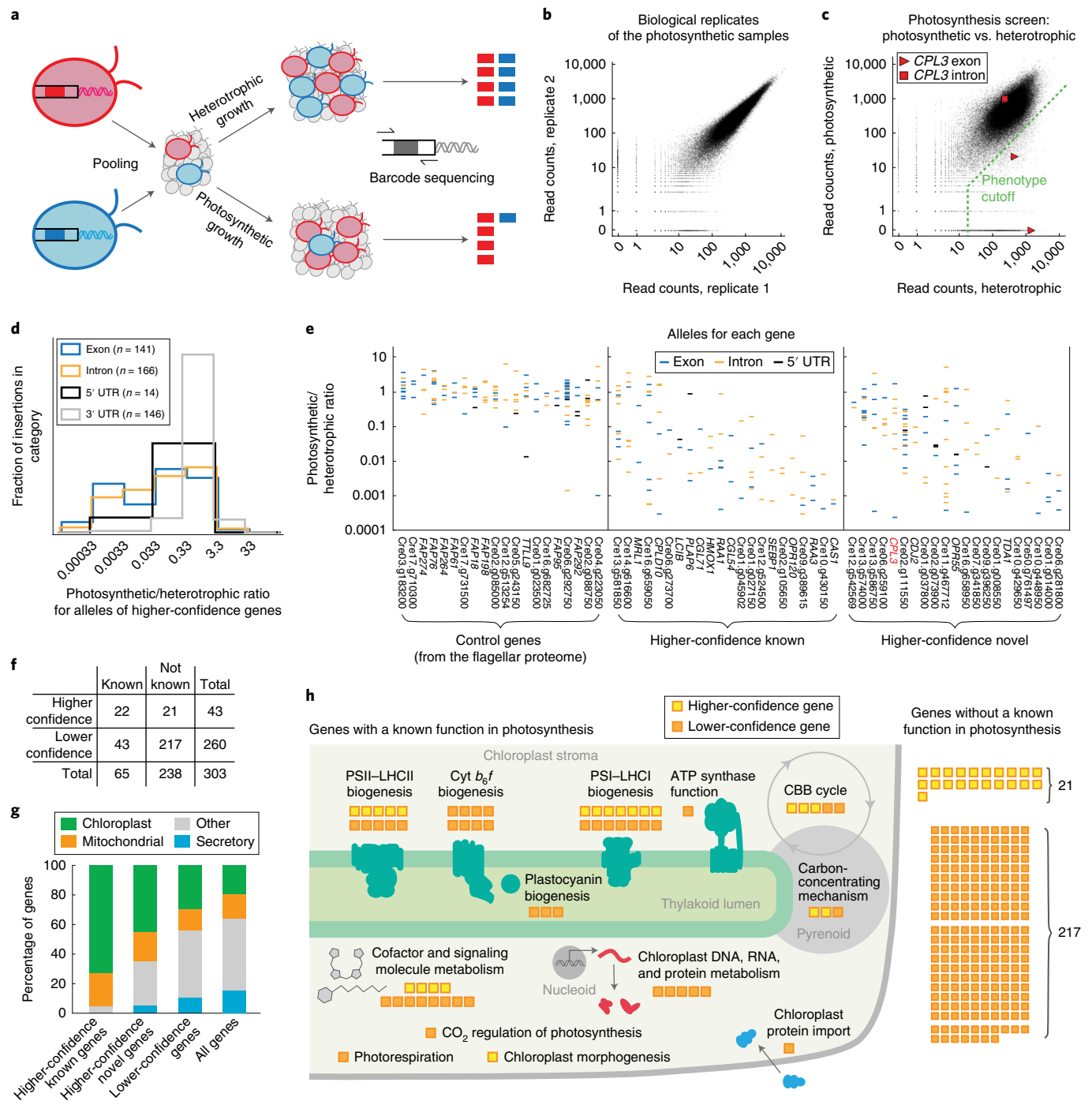


Fig. 3 | A high-throughput screen using the library identifies many genes with known roles in photosynthesis and many novel components. **a**, Unique barcodes allow screening of mutants in a pool. Mutants deficient in photosynthesis can be identified because their barcodes will be less abundant after 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: 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 *b₆f*, photosystem I (PSI), ATP synthase complexes, and a plastocyanin or cytochrome *c₆* metalloprotein, as well as small-molecule cofactors¹⁶. 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 CO₂-fixing enzyme Rubisco. In addition, most eukaryotic algae have a mechanism to concentrate CO₂ around Rubisco to enhance its activity¹⁷.

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 *b₆f* complex subunits (PETC and PETM) and 6 cytochrome *b₆* 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, DVRI, 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), CO₂ 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 Rubisco¹⁸. 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 accumulation¹⁹, the protein encoded by Cre11.g467712 contains a starch-binding domain, Cre02.g073900 encodes a putative carotenoid dioxygenase, *VTE5* encodes a putative phosphatidate cytidyltransferase, 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 understood²⁰. 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 phenotype¹⁵. 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 reactions²¹. 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 (<i>SEBP1, SBP1</i>)	Sedoheptulose-1,7-bisphosphatase	C	3	0	0.021	AT3G55800.1 (<i>SBPASE</i>)	<i>Arabidopsis</i> ²⁹
	Cre12.g524500 (<i>RMT2</i>)	Rubisco small subunit N-methyltransferase	O	3	0	0.021	AT3G07670.1	<i>Pisum</i> ³⁰
	Cre06.g298300 (<i>MRL1, PPR2</i>)	Pentatricopeptide-repeat protein, stabilizes <i>rbcl</i> mRNA	C	1	1	1.000	AT4G34830.1 (<i>MRL1</i>)	<i>Chlamydomonas</i> and <i>Arabidopsis</i> ³¹
Carbon-concentrating mechanism	Cre12.g497300 (<i>CAS1, TEF2</i>)	Rhodanese-like calcium-sensing receptor	C	2	0	0.260	AT5G23060.1 (<i>CaS</i>)	<i>Chlamydomonas</i> ³²
	Cre10.g452800 (<i>LCIB</i>)	Low-CO ₂ -inducible protein	C	2	0	0.260	-	<i>Chlamydomonas</i> ³³
Chloroplast and thylakoid morphogenesis	Cre14.g616600	-	M	4	3	0.021	AT1G03160.1 (<i>FZL</i>)	<i>Arabidopsis</i> ³⁴
				4	3	0.018		
Cofactor and signaling molecule metabolism	Cre13.g581850	-	M	5	5	0.010	AT4G31390.1	<i>Arabidopsis</i> ³⁵
	Cre10.g423500 (<i>HMOX1, HMO1</i>)	Heme oxygenase	C	3	0	0.021	AT1G69720.1 (<i>HO3</i>)	<i>Chlamydomonas</i> ¹⁴
	Cre03.g188700 (<i>PLAP6, PLP6</i>)	Plastid lipid-associated protein, fibrillin	C	3	1	0.070	AT5G09820.2	<i>Arabidopsis</i> ³⁶
	Cre16.g659050	-	C	4	6	0.098	AT1G68890.1	<i>Chlamydomonas</i> ³⁷
PSI protein synthesis and assembly	Cre12.g524300 (<i>CGL71</i>)	Predicted protein	C	2	0	0.260	AT1G22700.1	<i>Synechocystis</i> ³⁸ , <i>Arabidopsis</i> ³⁹ , and <i>Chlamydomonas</i> ⁴⁰
	Cre01.g045902	-	C	1	1	1.000	AT3G24430.1 (<i>HCF101</i>)	<i>Arabidopsis</i> ^{41,42}
PSI RNA splicing and stabilization	Cre09.g389615	-	M	5	0	0.0002	AT3G17040.1 (<i>HCF107</i>)	<i>Chlamydomonas</i> ⁴³ and <i>Arabidopsis</i> ^{42,44 f}
	Cre01.g027150 (<i>CPLD46, HEL5</i>)	DEAD/DEAH-box helicase	M	5	1	0.0004	AT1G70070.1 (<i>EMB25, ISE2</i> ,	<i>Arabidopsis</i> ⁴⁵
	Cre09.g394150 (<i>RAA1</i>)	-	M	5	1	0.0004	-	<i>Chlamydomonas</i> ⁴⁶
	Cre12.g531050 (<i>RAA3</i>)	<i>psaA</i> mRNA maturation factor 3	C	3	0	0.021	-	<i>Chlamydomonas</i> ⁴⁷
	Cre10.g440000 (<i>OPR120</i>)	-	C	2	0	0.260	-	<i>Chlamydomonas</i> ^{48,49}
PSII protein synthesis and assembly	Cre13.g578650 (<i>CPLD10, NUOAF5</i>)	Similar to complex I intermediate-associated	C	3	3	0.260	AT1G16720.1 (<i>HCF173</i>)	<i>Arabidopsis</i> ^{42,50,51}
	Cre02.g073850 (<i>CGL54</i>)	Predicted protein	C	2	0	0.260	AT1G05385.1 (<i>LPA19, Psb27-H1</i>)	<i>Arabidopsis</i> ⁵²
	Cre02.g105650	-	C	2	0	0.260	AT5G51545.1 (<i>LPA2</i>)	<i>Arabidopsis</i> ⁵³
	Cre06.g273700 (<i>HCF136</i>)	-	C	2	0	0.260	AT5G23120.1 (<i>HCF136</i>)	<i>Arabidopsis</i> ⁴² and <i>Synechocystis</i> ⁵⁴
	Cre10.g430150 (<i>LPA1, REP27</i>)	-	C	2	0	0.260	AT1G02910.1 (<i>LPA1</i>)	<i>Arabidopsis</i> ⁵⁵

^aPrediction of protein localization by PredAlgo⁵⁶: C, chloroplast; M, mitochondrion; SP, secretory pathway; O, other. ^bThe 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. ^cThe 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. ^dThe 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 Phytozome¹². ^fAT3G17040.1 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 (<i>TRX21</i>)	ATP-binding protein; thioredoxin domain	O	3	3	0.260	AT2G18990.1 (<i>TXND9</i>)
			1	5	1.000	
Cre02.g073900	All-trans-10'-apo-β-carotenal 13,14-cleaving dioxygenase	C	3	1	0.070	AT4G32810.1 (<i>ATCCD8</i> , <i>CCD8</i> , <i>MAX4</i>)
			3	1	0.056	
Cre02.g111550	Serine/threonine kinase related	SP	10	8	<10 ⁻⁶	AT4G24480.1
			6	12	0.015	
Cre03.g185200 (<i>CPL3</i> , <i>MPA6</i>)	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 (<i>CDJ2</i>)	Chloroplast DnaJ-like protein	M	2	0	0.260	AT5G59610.1
			1	1	1.000	
Cre07.g341850 (<i>EIF2</i> , <i>INFB</i>)	Translation initiation factor IF-2, chloroplastic	C	2	0	0.260	AT1G17220.1 (<i>FUG1</i>)
			2	0	0.239	
Cre08.g358350 (<i>TDA1</i> , <i>OPR34</i>)	Fast leucine-rich domain containing ^a	C	3	2	0.152	-
			3	2	0.117	
Cre09.g396250 (<i>VTE5</i>)	Phosphatidate cytidyltransferase	SP	2	0	0.260	AT5G04490.1 (<i>VTE5</i>)
			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 (<i>ATGLR3.4</i> , <i>GLR3.4</i> , <i>GLUR3</i>)
			2	0	0.239	
Cre13.g566400 (<i>OPR55</i>)	Fast leucine-rich domain containing ^a	M	4	2	0.018	-
			4	2	0.015	
Cre13.g574000 (<i>GEF1</i> , <i>CLV1</i>)	Voltage-gated chloride channel	O	1	11	1.000	AT5G26240.1 (<i>ATCLC-D</i> , <i>CLC-D</i>)
			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 <i>mrs2</i> homolog, mitochondrial	M	2	0	0.260	AT5G22830.1 (<i>ATMGT10</i> , <i>GMN10</i> , <i>MGT10</i> , <i>MRS2-11</i>)
			2	0	0.239	

^aThe annotation of 'fast leucine-rich domain containing' cannot be confirmed by BLASTP analysis at NCBI¹⁷.

the genes required for photosynthesis remain to be characterized. This finding is notable considering that genetic studies on photosynthesis extend back to the 1950s²². 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

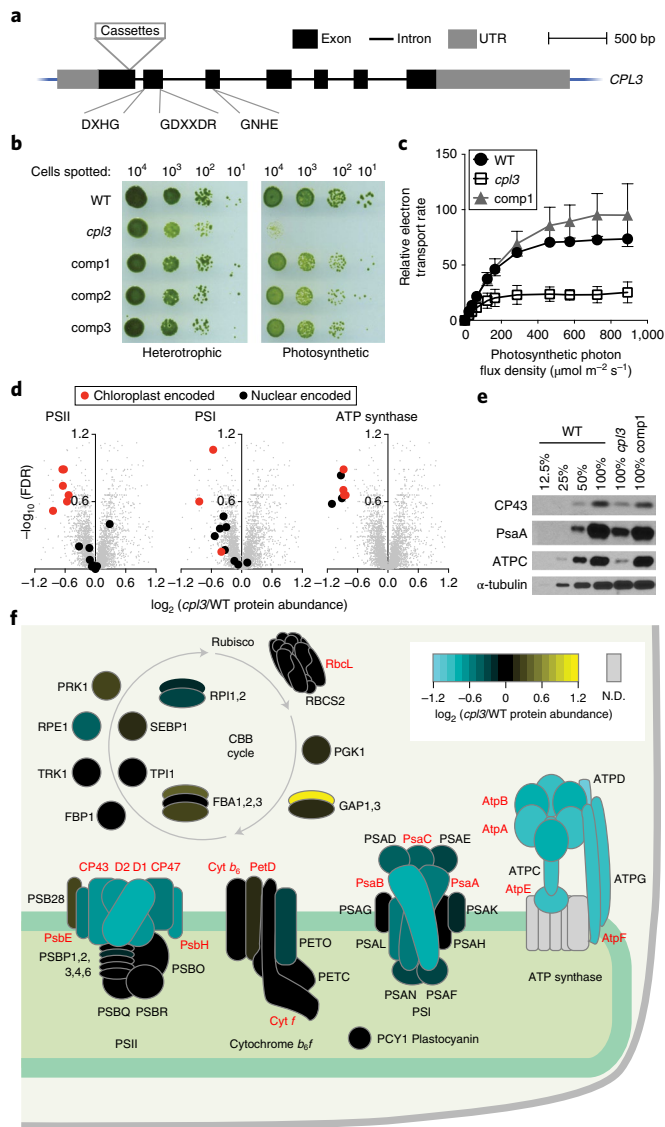


Fig. 4 | CPL3 is required for photosynthetic growth and accumulation of photosynthetic protein complexes in the thylakoid membranes.

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

- 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).
- Krishnakumar, V. et al. Araport: the *Arabidopsis* information portal. *Nucleic Acids Res.* **43**, D1003–D1009 (2015).
- Levine, R. P. Genetic control of photosynthesis in *Chlamydomonas reinhardtii*. *Proc. Natl Acad. Sci. USA* **46**, 972–978 (1960).
- Gutman, B. L. & Niyogi, K. K. *Chlamydomonas* and *Arabidopsis*. A dynamic duo. *Plant Physiol.* **135**, 607–610 (2004).
- Harris, E. H., Stern, D. B. & Witman, G. B. *The Chlamydomonas Sourcebook* (Academic Press, 2009).
- Rochaix, J. D. *Chlamydomonas reinhardtii* as the photosynthetic yeast. *Annu. Rev. Genet.* **29**, 209–230 (1995).
- 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).
- Silflow, C. D. & Lefebvre, P. A. Assembly and motility of eukaryotic cilia and flagella: lessons from *Chlamydomonas reinhardtii*. *Plant Physiol.* **127**, 1500–1507 (2001).
- 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).
- 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).
- Pazour, G. J., Agrin, N., Leszyk, J. & Witman, G. B. Proteomic analysis of a eukaryotic cilium. *J. Cell Biol.* **170**, 103–113 (2005).
- Merchant, S. S. et al. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–251 (2007).
- 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).
- Duanmu, D. et al. Retrograde bilin signaling enables *Chlamydomonas* greening and phototrophic survival. *Proc. Natl Acad. Sci. USA* **110**, 3621–3626 (2013).
- 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).
- 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).
- Giordano, M., Beardall, J. & Raven, J. A. CO₂ concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. *Annu. Rev. Plant Biol.* **56**, 99–131 (2005).
- Goldschmidt-Clermont, M. & Rahire, M. Sequence, evolution and differential expression of the two genes encoding variant small subunits of ribulose biphosphate carboxylase/oxygenase in *Chlamydomonas reinhardtii*. *J. Mol. Biol.* **191**, 421–432 (1986).
- 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).
- 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).
- 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 CO₂-concentrating mechanism by Ca²⁺-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 CO₂ 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 solanessyl diphosphate synthases in *Arabidopsis*. *Plant Cell* **27**, 2956–2971 (2015).
37. Lefebvre-Legendre, L. et al. Loss of phyloquinone in *Chlamydomonas* affects plastoquinone pool size and photosystem II synthesis. *J. Biol. Chem.* **282**, 13250–13263 (2007).
38. Wilde, A., Lunsner, 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. & Oelmüller, R. The evolutionarily conserved tetratricopeptide 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. Heinzel, 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 intergenic 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 in *Arabidopsis*. *Plant Physiol.* **160**, 2202–2218 (2012).
51. Schulz, 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.L., and X.L. generated the library. J.M.R., N.L., A.G., and R.Y. maintained, consolidated, and cryopreserved the library. X.L. developed the barcode sequencing method. N.L., 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 *cp13*. S.S. and T.M.W. performed physiological characterizations of *cp13*. M.T.M. and S.S. performed western blots on the photosynthetic protein complexes. M.T.M. performed microscopy on *cp13*. 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.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019

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-4533⁵⁸ 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 study⁹, but it consumed considerably less time because we used a simple PCR amplifying only the barcodes instead of a multistep flanking sequence extraction protocol (ChlaMmeSeq⁵⁸) 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-Seq⁹ (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 site⁹ (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 previously⁵⁸. PCR products of the expected size were verified by Sanger sequencing.

Southern blotting. Southern blotting was performed as previously described in detail⁹. 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 Stratilinker 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 region⁵⁸. Calculation of mappability, hotspot and coldspot analysis, and simulations of random insertions were performed as described previously⁵⁸, 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 genome¹². 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 genes¹¹. Comparison was performed by using Fisher's exact test with correction for multiple comparisons⁶¹ 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 genes¹¹, 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- μ m 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 (1,000g, 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- μ m 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 μ mol m⁻² s⁻¹ photons and was then increased to 500 μ mol m⁻² s⁻¹ photons after the culture reached $\sim 2 \times 10^5$ 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 (1,000g, 5 min, room temperature) for DNA extraction and barcode sequencing.

Molecular characterization of the *cpl3* mutant. Mutant genotyping PCRs were performed as previously described⁹. To complement the *cpl3* mutant, the wild-type *CPL3* gene was PCR amplified and cloned into the pRAM118 vector containing the *aph7* gene⁶², 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 ($\sim 10^6$ 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 $30 \mu\text{mol m}^{-2} \text{s}^{-1}$ photons for 1 d, under $100 \mu\text{mol m}^{-2} \text{s}^{-1}$ photons for 1 d, and then under $500 \mu\text{mol m}^{-2} \text{s}^{-1}$ photons for 4 d.

Chlorophyll *a* and chlorophyll *b* concentrations were measured as previously described⁶³ 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 *cp13* 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 (Φ_{PSII}) was quantified as previously described⁶⁴. Relative electron transport rate (rETR) was calculated according to the following equation: $\text{rETR} = \Phi_{\text{PSII}} \times I$, 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 guanidium 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 tips⁶⁵. The dried peptide mixture was separated with strong cation exchange (SCX) stage tips⁶⁶ 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 MaxQuant⁶⁷ 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 Byonic⁷⁰ 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 testing⁶¹, 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 described⁹.

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 correction⁶¹ 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 *cp13* mutant have been deposited to the ProteomeXchange Consortium via the PRIDE⁷¹ 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

- 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).
- Rubin, B. E. et al. The essential gene set of a photosynthetic organism. *Proc. Natl Acad. Sci. USA* **112**, E6634–E6643 (2015).
- Wetmore, K. M. et al. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio* **6**, e00306–e00315 (2015).
- 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).
- Berthold, P., Schmitt, R. & Mages, W. An engineered *Streptomyces hygroscopicus* *aph 7ⁿ* gene mediates dominant resistance against hygromycin B in *Chlamydomonas reinhardtii*. *Protist* **153**, 401–412 (2002).
- 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).
- 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).
- 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).

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](https://www.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	<input type="text" value="We empirically determined the sample sizes based on published research."/>
Data exclusions	<input type="text" value="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	<input type="text" value="We performed replicates and used orthogonal approaches where appropriate."/>
Randomization	<input type="text" value="We randomly picked mutants for validation of insertion site mapping."/>
Blinding	<input type="text" value="Blinding and randomization were not used for this study."/>

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involvement in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input type="checkbox"/> Human research participants

Methods

n/a	Involvement in the study
<input type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials	<input type="text" value="The mutants described in this manuscript are available from the Chlamydomonas Resource Center: https://www.chlamylibrary.org/"/>
----------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------

Antibodies

Antibodies used	<input type="text" value="Antibodies used in Fig. 4 are commercially available from Agrisera. The catalog numbers are provided in Methods."/>
Validation	<input type="text" value="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)	<input type="text" value="We have been using the CC-4533 Chlamydomonas strain generated in our lab."/>
Authentication	<input type="text" value="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	<input type="text" value="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

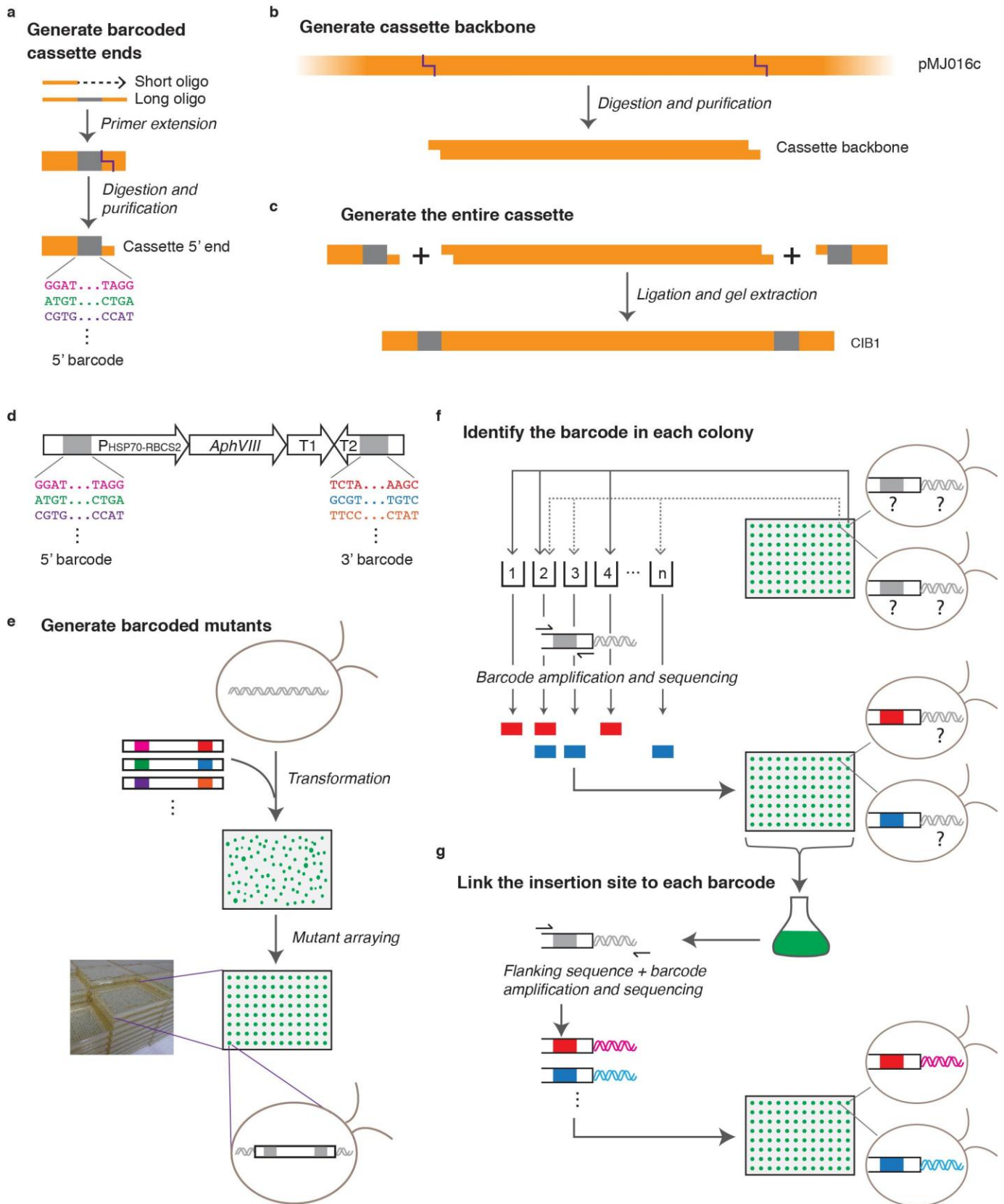
Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.

In the format provided by the authors and unedited.

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 Saroussi², Moritz T. Meyer¹, Nina Ivanova², Jacob M. Robertson^{1,2}, Rebecca Yue², Ru Zhang^{2,8}, Josep Vilarrasa-Blasi², Tyler M. Wittkopp ^{2,4,9}, Silvia Ramundo⁵, Sean R. Blum², Audrey Goh¹, Matthew Laudon⁶, Tharan Srikumar¹, Paul A. Lefebvre⁶, Arthur R. Grossman² and Martin C. Jonikas ^{1,2*}

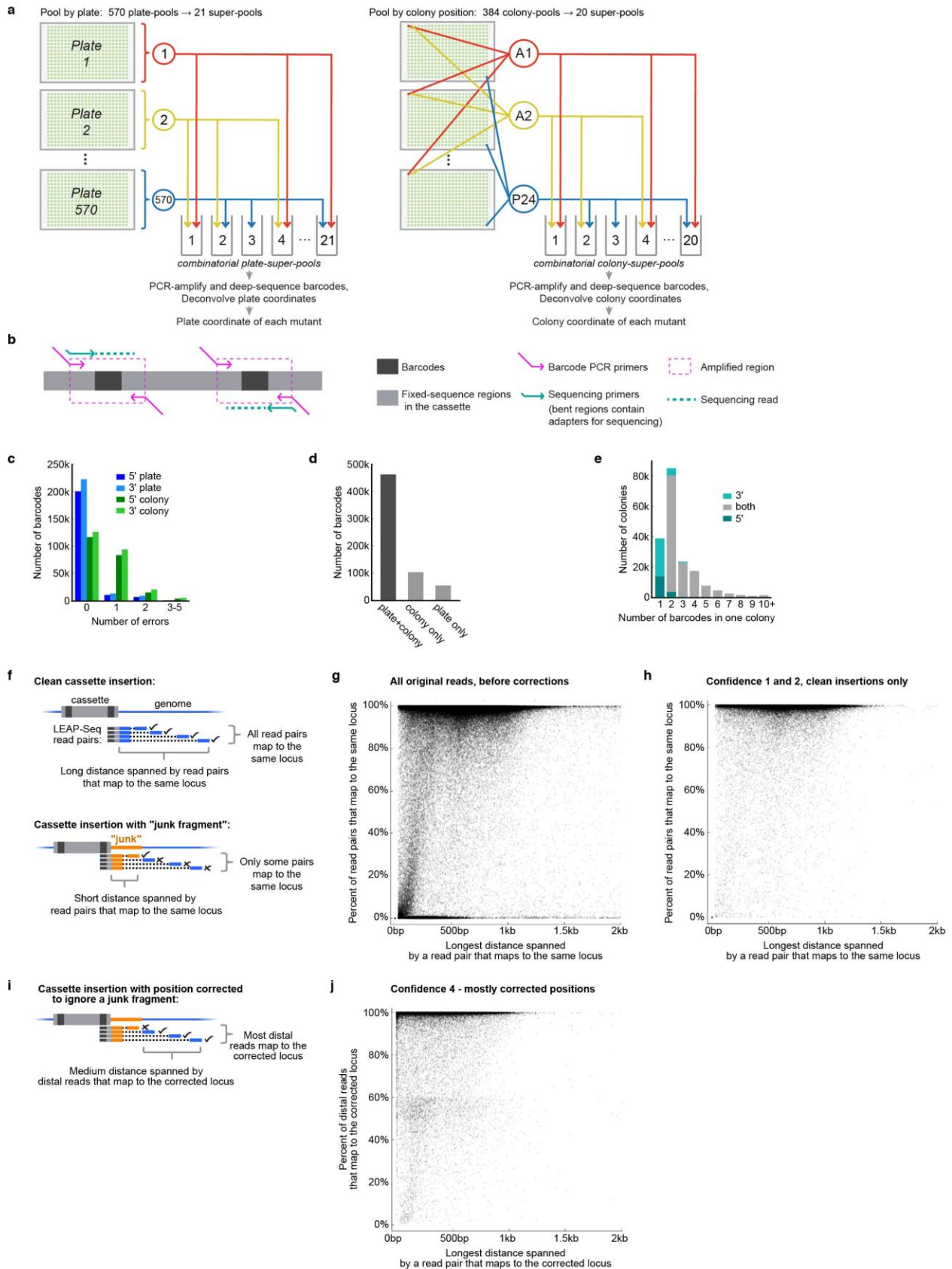
¹Department of Molecular Biology, Princeton University, Princeton, NJ, USA. ²Department of Plant Biology, Carnegie Institution for Science, Stanford, CA, USA. ³School of Life Sciences, Westlake Institute for Advanced Study, Westlake University, Hangzhou, China. ⁴Department of Biology, Stanford University,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



Supplementary Figure 1

A pipeline was developed for generating barcoded cassettes and for generating an indexed and barcoded library of insertion mutants in *Chlamydomonas*.

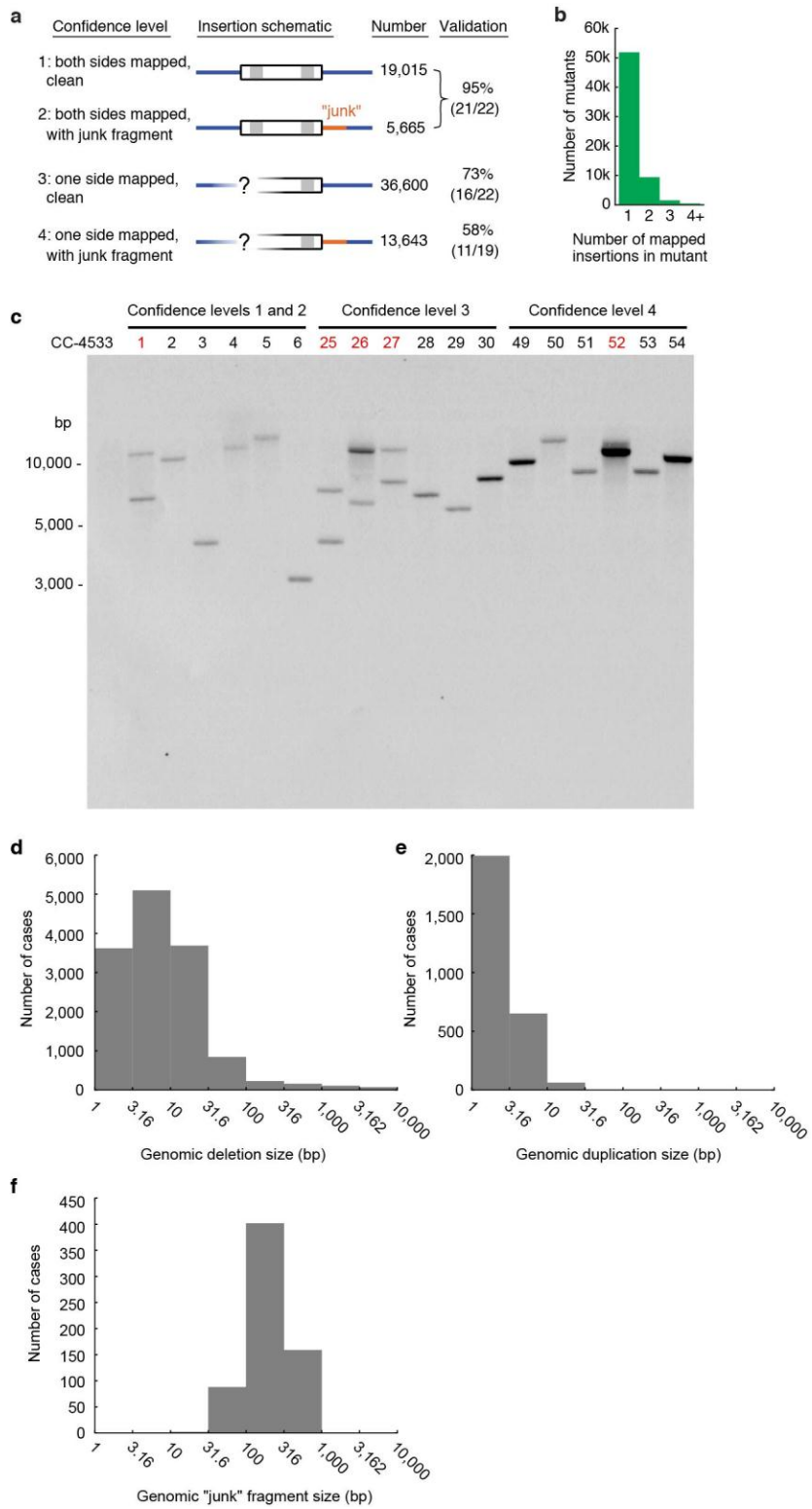
a, A long oligonucleotide primer containing a random sequence region (indicated in gray) was used as a template for the extension of a shorter oligonucleotide primer (Supplementary Table 1). The resulting double-stranded product contains a random sequence region (22 bp in length; termed "barcode"). This product was restriction digested to generate a sticky end for subsequent ligation. The above steps were performed to produce both the 5' and the 3' ends of the cassette. The 5' end of the cassette is shown as an example. **b**, The pMJ016c plasmid was digested to yield the backbone of the cassette. **c**, The 5' and 3' ends of the cassette generated above were ligated together with the cassette backbone to yield the cassette CIB1. **d**, The components of the cassette CIB1 are shown. CIB1 contains the *HSP70-RBCS2* promoter (with an intron from *RBCS2*), the *AphVIII* gene that confers resistance to paromomycin, two transcriptional terminators (T1: *PSAD* terminator; T2: *RPL12* terminator), and two barcodes (each 22 bp in length). **e**, Following transformation and arraying of individual mutants, the sequence of the barcodes contained in each insertion cassette was unique to each transformant but initially unknown for each colony. **f**, Barcodes were amplified from combinatorial pools of mutants, sequenced, and traced back to single colonies (Supplementary Fig. 2a-e; Supplementary Note). After this step, the barcode sequence for each colony was known. For simplicity, only one side of the cassette is shown. **g**, Barcodes and genomic sequences flanking the insertion cassettes were amplified from a pool of the library. By pooled next-generation sequencing, the sequence flanking each insertion cassette was paired with the corresponding barcode (Supplementary Fig. 2f). The flanking sequences were used to determine the insertion site in the genome. Because the colony location for each barcode was determined in the previous step, insertion sites could then be assigned to single colonies.



Supplementary Figure 2

Combinatorial pooling, barcode deconvolution to colony and determination of insertion sites.

a, To determine which plate each barcode was on, each plate of mutants was pooled into one of 570 plate-pools. The plate-pools were then further combinatorially pooled into 21 plate-super-pools, in such a way that each plate-pool was in a unique combination of plate-super-pools. The barcodes present in each plate-super-pool were determined by deep sequencing, and the barcodes were assigned to plates based on the combination of plate-super-pools they were found in. A similar process was applied to the colony positions of each barcode. Combining the plate and colony data yielded a specific position for each barcode. **b**, The barcodes on the 5' and 3' sides of the cassette were sequenced separately, each with a single-end Illumina read. With the sequencing primers we used (indicated on the cassette), the reads start with the barcode sequence and extend into the cassette. **c**, Most barcode colony positions were identified with no errors, i.e. were found in one of the expected combinations of super-pools. Some were found in a combination of super-pools that had one or more differences from any expected combination, but the positions could still be identified due to the redundancy built into our method. The much higher number of one-error cases in the colony data compared to plate data is due to a loss of one of the colony-super-pools for a significant fraction of the samples (Supplementary Note). **d**, Both a plate and a colony position were identified for most barcodes. **e**, The number of barcodes mapped to an individual colony varied, with 2 being the most common. For colonies with two mapped barcodes, the large majority had one 5' and one 3' barcode, likely derived from two sides of one cassette. **f**, LEAP-Seq reads are paired-end reads with the proximal read containing the cassette barcode and immediate flanking genomic sequence, and the distal read containing flanking genomic sequence a variable distance away from the insertion site. During transformation, short fragments of genomic DNA, likely originating from lysed cells, are often inserted between the cassette and the true flanking genomic DNA. We refer to these short DNA fragments as "junk fragments" (Zhang, R. *et al.*, *Plant Cell*. **26**, 1398-1409, 2014 and Li, X. *et al.*, *Plant Cell*. **28**, 367-387, 2016). Such junk fragments can lead to incorrect insertion mapping if only the immediate flanking genomic sequence is obtained. LEAP-Seq data can be used to detect presence of junk fragments at an insertion junction based on two key characteristics: 1) the number of read pairs where both sides aligned to the same locus and 2) the longest distance spanned by such read pairs. **g**, The two key characteristics are plotted for the original full library, before any mapping corrections were applied. **h**, The same two characteristics are plotted for confidence level 1 and 2 insertions. For confidence level 2 insertions, only the side with no junk fragment is shown; for confidence level 1 insertions, one randomly chosen side is shown. **i**, LEAP-Seq data can be used to correct cases of probable junk fragment insertions and determine the most likely correct insertion position. The corrected data can be visualized using two modified key characteristics: the number of distal reads aligned to the corrected location, and the distance spanned by such reads. **j**, The modified characteristics are plotted for confidence level 4 insertions.

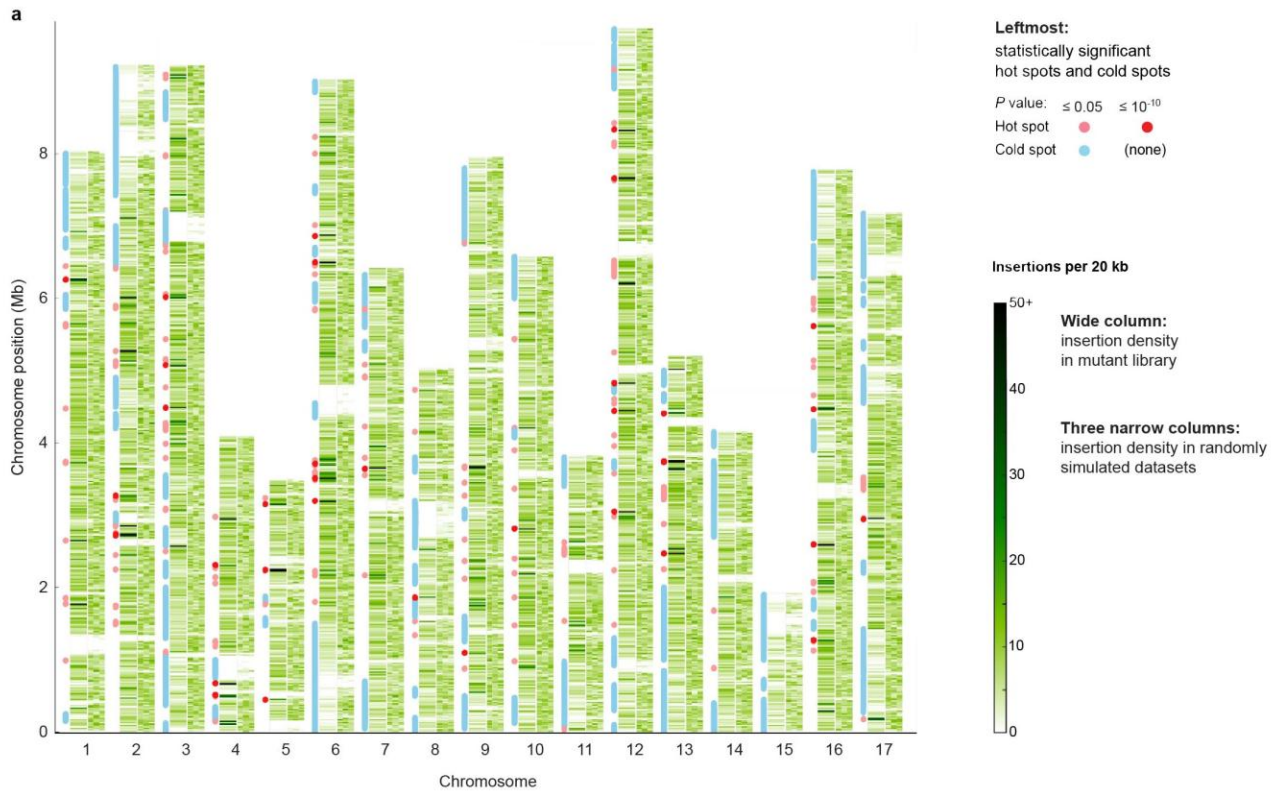


Supplementary Figure 3

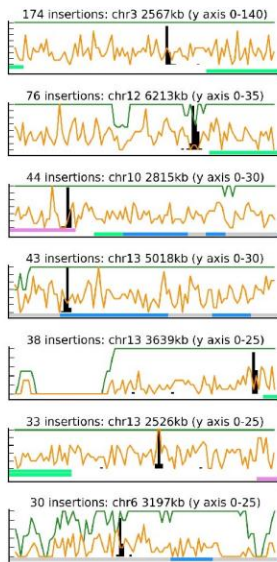
Characterization of genomic disruptions in mutants in the library.

a, Mutants in the library were divided into four confidence levels, corresponding to different mapping scenarios. The insertion sites of a

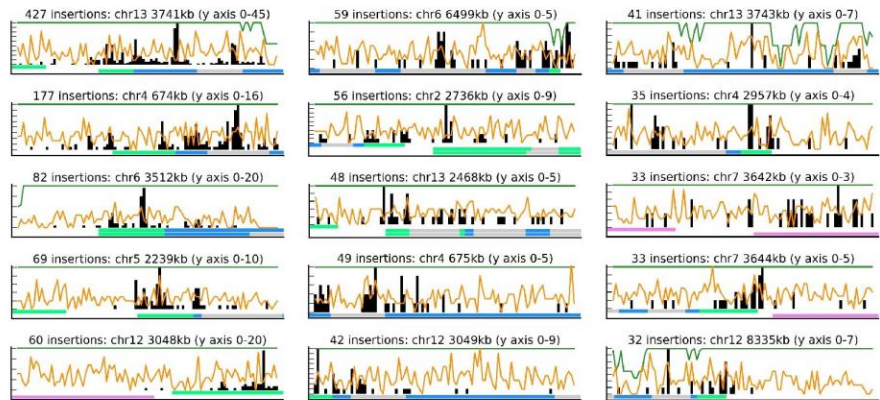
number of randomly chosen mutants in each category were verified by PCR (mutants from confidence levels 1 and 2 were assayed as one group; Supplementary Table 6). The numbers and percentages of confirmed insertions are shown in the last column. **b**, Most mutants have a single mapped insertion, and < 20% contain two or more mapped insertions. **c**, Eighteen randomly selected mutants from the four confidence levels were analyzed by Southern blotting using the coding sequence of *AphVIII* as the probe. Mutants are numbered and the details of their insertion sites are presented in Supplementary Table 6. The mutant number is highlighted in red when the Southern blot was interpreted to indicate at least two insertions in that mutant. The wild-type strain CC-4533 (WT) was included as a negative control. **d**, Most genomic deletions accompanying cassette insertions are smaller than 100 bp, but deletions up to 10 kb are present in some mutants. Deletions larger than 10 kb may also be present, but there were not enough of them to be clearly detected based on the aggregate numbers. **e**, Most genomic duplications accompanying cassette insertion are smaller than 10 bp, but they can be up to 30 bp. Larger duplications may be present, but these are not common enough to be detected based on the aggregate numbers. **f**, The distribution of junk fragment lengths was determined using a dataset of 651 insertions of two cassettes surrounding a junk fragment, allowing us to precisely map both ends of the junk fragment using LEAP-Seq. Most junk DNA fragments are smaller than 320 bp, but we have detected some up to 1 kb in size. Larger junk fragments may be present, but are not common enough to be detected based on the aggregate numbers. Note that the x-axes for **d-f** are set to the logarithmic scale. Data presented in this figure are described in a Supplementary Note.



b 9 narrow hot spots, ~20bp in size:



c 13 hot spots with insertions spread over a wider area:



Each plot represents one hot spot.

X axis: genome position:
1kb surrounding each hot spot, in 10 bp bins

Y axis - different for each type of data plotted:

- insertion density, with different Y scale for each graph, noted at the top along with the total number of insertions
- ⬆ background strain whole-genome sequencing density, with Y scale 0-10 reads per 10 bp
- ⬆ genome mappability, with Y scale 0-100%

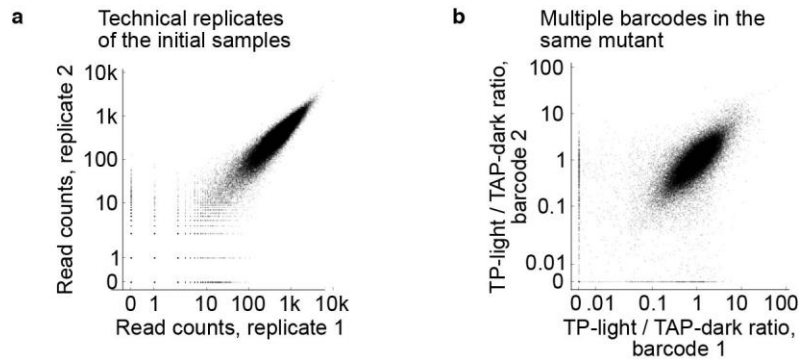
Bottom of graph: gene structure
(multiple genes/transcripts shown if present):

- exon — intron — 5' UTR — 3' UTR

Supplementary Figure 4

The distribution of insertions in the genome is largely random, and the hotspots fall into two classes.

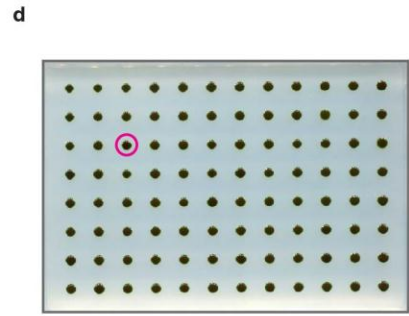
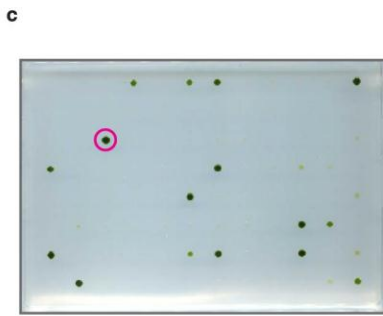
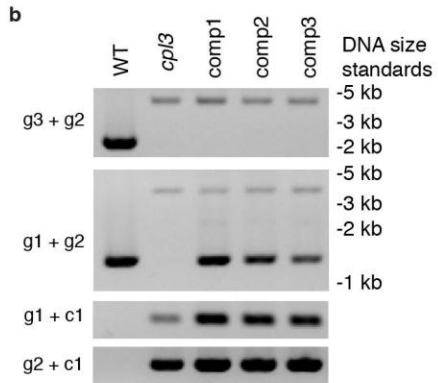
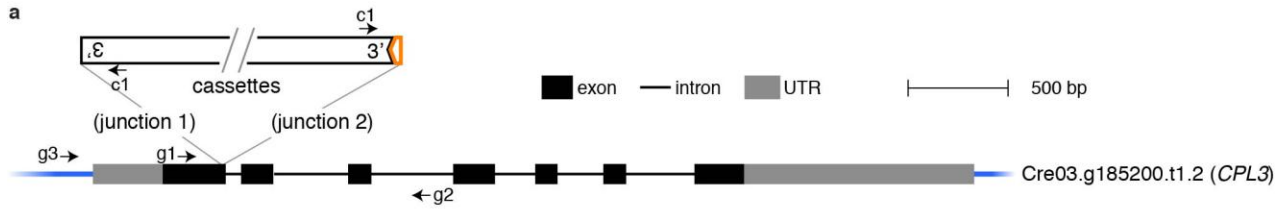
a, For each chromosome, the observed insertion density is shown as a heatmap in a wide column, followed by three narrow columns depicting three simulated datasets in which insertions were placed in randomly chosen mappable genomic locations. The simulated data provide a visual guide to the amount of variation expected from a random distribution. The large white areas present in both the observed and simulated data correspond to repetitive genomic regions in which insertions cannot be mapped uniquely. The red and blue circles/lines to the left of each chromosome show statistically significant insertion hotspots and cold spots, respectively. To ensure that we are showing true insertion density rather than artifacts caused by junk fragments or other mapping inaccuracies, the plot of insertion site distribution and identification of hot/cold spots are based on confidence level 1 insertions only. In contrast, Fig. 1c shows the distribution of insertions of all confidence levels over the genome. **b** and **c**, Each plot represents a 1-kb genomic region surrounding one hot spot, showing multiple features of that region, as listed in the legend. The plots shown are the 22 1-kb regions with the highest total insertion number. The total number of insertions for each region is listed above each plot, along with the genomic position and the y-axis range. **b**, 7 of the top 22 hot spots are narrow, with 20 or more insertions in a 10-bp area, and a total width of 20-30 bp with few or no additional insertions in the surrounding 1 kb. **c**, 15 of the top 22 hot spots are wider, with multiple peaks of high insertion density spanning at least hundreds of base pairs. In either class, the insertion density peaks do not appear to reliably correlate with any of the other genomic features shown. Data presented in this figure are described in a Supplementary Note.



Supplementary Figure 5

The barcode sequencing method is robust.

a, The barcode sequencing read counts (normalized to 100 million total reads) for each insertion were highly reproducible between technical replicates, with a Spearman's correlation of 0.978. 94% of barcodes showed a normalized read count of no more than a 2-fold difference between the two replicates. **b**, The TP-light/TAP-dark ratios of multiple barcodes in the same mutant are consistent, with a Spearman's correlation of 0.744. Only 4% of insertion pairs had a greater than 5x difference between ratios. See also Fig. 3, b and c.



e

```

CPL3      MALGMQRQLRGHQRTAPAPVLPVVRPR-----ATRATGPSASRGSRRHLLQOIAGATL--
PPI       -----MNKIY-----CLAVLSLTLTLLSPLALANTATEFDGPYV-----ITPISGQSTAY
          *::                . * :: *          ** .  ** .          : *:* :

CPL3      -LVHARSVADPSSVASASATLAAPTEEASTSTTVL--GNSALDPPTYVTATGRIIAIGDL
PPI       WICDNRL--KTTSIE--KLQVNRPEHCGLDPETKLSSEIKQIMPDTYL--GIKKVVALSDV
          : . * . :*: . : * . . . * * . : * **: . :*:*:*:

CPL3      HGDLDKAVEALKLGRVISVSDEGEVSWVGGDTVVVQLGDVLDRGDVEIGIINLLRYLDTE
PPI       HGQYDVLTLTLLKQKIID----SDGNWAFGEGHMVMTGDI FDRGHQVNEVLWFMYQLDQQ
          **: * : * * ::* . : . * . * : * **:***. : : : * * :

CPL3      ARKQGGAVYMLNGNHESLNVCDFRYVTPGAFSAESALYAGLSESDLKDWQLVAKVRYSLY
PPI       ARDAGGMVHLLMGNHEQMVLGDDLRYVHQRYDIATTL-----INRPYNKLY
          ** . ** *::* *::* . : : *::*** : : * : : : : * *

CPL3      KPGGDLAREFSRNPTVLVVDNDFVFAHGGLLPTHVEYGIERLNSEVAAWMRGDDIP-DGN-
PPI       GADTEIGQWLRSKNTI IKINDVLYMHGGISSEWISRELTLDKA-NALYRANVDASKKSLK
          . : : : : : *:: : **:: *:: *:: : . : : : * : . * ..

CPL3      -KAQPPFLAMGDANSVMWNRTLSKERFATPYERYHACNALKQALAKVRGKRLVVGHTPQL
PPI       ADDLLNFLFFG--NGPTWYRGYFSETFTEA-----ELDTILQHFNVNHIVVGHTSQE
          .   ** : * . * * . * * :          * . * : . : : * * * * *

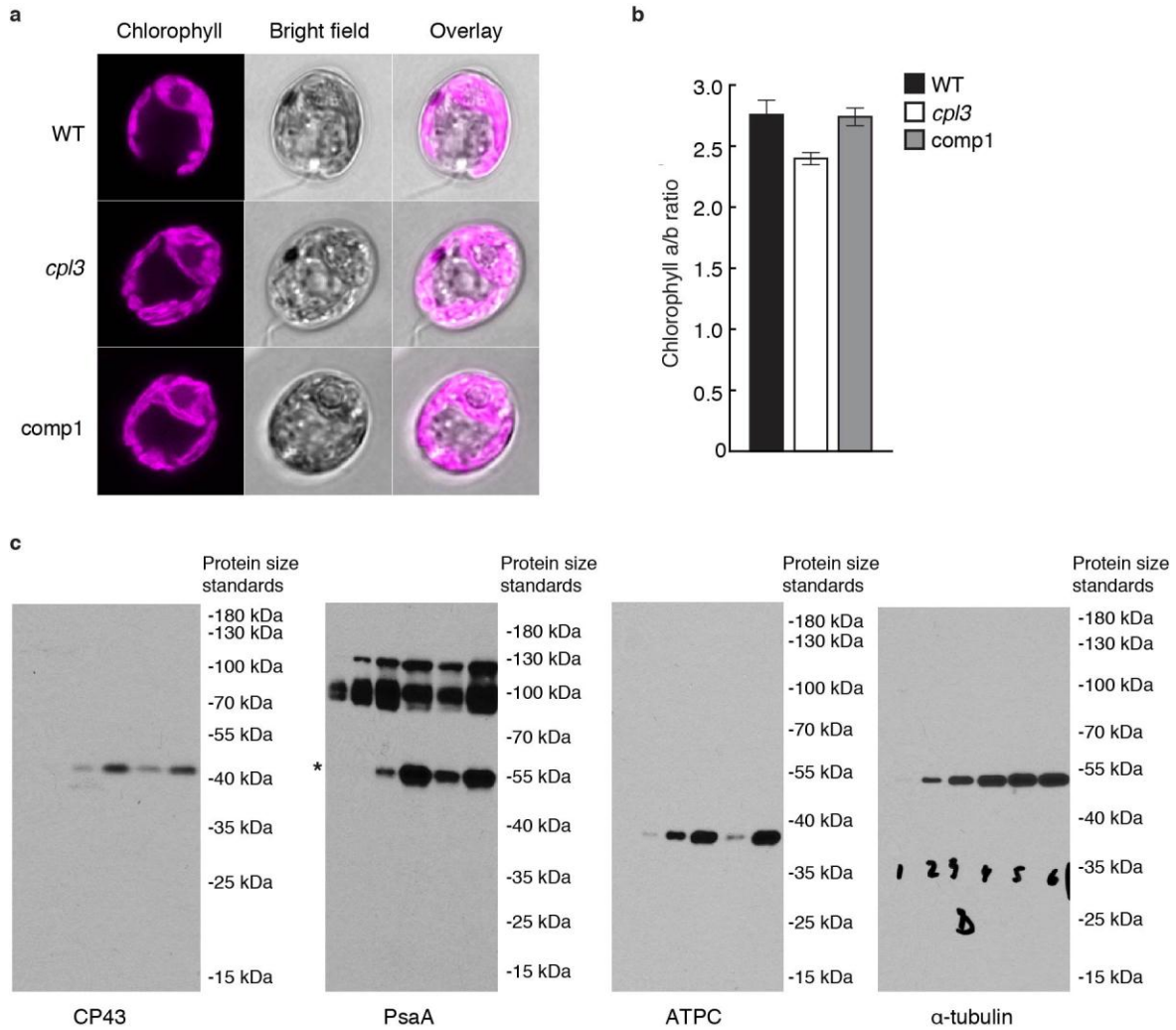
CPL3      GGVNCECENQVWRIDVGMYSYGLNRPVQVIEIVPPEEGGDDAKVRVIRNTPNSMSSADDD
PPI       RVLGL-FHNKVIADVSSIKVKGSGELL--LLENRLIRGLYDGTRETLO--ENSLNQ----
          : . . *:* : * . . . * . . : : * . * * . . : : * * : *

CPL3      ITIASNL
PPI       -----
  
```

Supplementary Figure 6

Molecular characterization of the *cp13* mutant.

a, The cassette insertion site is indicated on a model of the *CPL3* gene from the *Chlamydomonas* v5.5 genome. Two cassettes are inserted in opposite orientations, with one of them truncated on the 3' side (indicated by a notch); the 5' ends may be intact or truncated. The orange box arrow indicates insertion of a small fragment of unknown origin. Binding sites for primers g1, g2, g3, and c1 are indicated. **b**, PCR genotyping results of *cp13* and complemented lines. PCR with the primer pair "g1 + g2" indicated presence of an insertion within the *CPL3* gene in the *cp13* mutant and presence of wild-type *CPL3* sequence in the wild-type (from the native *CPL3* locus) and in the complemented lines (from the complementation construct inserted at a random site in the genome of each line). PCR with the primer pair "g3 + g2" demonstrated the disruption of the native *CPL3* locus in the *cp13* and *comp1-3* lines, as the binding site for primer g3 is present only in the native *CPL3* locus and not in the complementation construct. PCR with primer pairs "g1 + c1" and "g2 + c1" showed the presence of a cassette inserted into the *CPL3* gene in *cp13* as well as the complemented lines. **c**, *cp13* mutants transformed with the *CPL3* gene were arrayed and grown photosynthetically in the absence of acetate for one day under 100 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ light and four additional days under 500 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ light before imaging. The colony circled was a positive control strain that grows photosynthetically. **d**, The same transformants were grown for five days in the presence of acetate in the medium under 50 $\mu\text{mol photons m}^{-2} \text{s}^{-1}$ light. All colonies grew similarly. **e**, *CPL3* contains conserved tyrosine phosphatase motifs. Sequences of *CPL3* in *Chlamydomonas* and its homolog psychrophilic phosphatase I (PPI) in *Shewanella* sp. were aligned using Clustal Omega (Sievers, F. *et al.*, *Mol Syst Biol.* **7**, 539, 2011). Asterisks (*), colon (:), and period (.) indicate conserved, strongly similar, and weakly similar amino acid residues, respectively. The motifs that are conserved among multiple protein phosphatases (Tsuruta, H. *et al.*, *J Biochem.* **137**, 69-77, 2005) are boxed. Data in panels **a-d** are described in a Supplementary Note. See also Fig. 4.



Supplementary Figure 7

Phenotypic characterization of the *cp13* mutant.

a, *cp13*, the wild-type strain (WT), as well as the complemented line (comp1), contain a normal cup-shaped chloroplast. Representative images of confocal chlorophyll fluorescence, bright field, and an overlay are shown for each strain. **b**, *cp13* has a lower chlorophyll *a/b* ratio than WT and comp1 ($P < 0.03$, Student's *t*-test). Error bars indicate standard deviations ($n = 3$). **c**, western blots show that *cp13* accumulates lower levels of the PSII subunit CP43, the PSI subunit PsaA, and the chloroplast ATP synthase subunit ATPC. For PsaA, bands with a higher molecular weight have been observed when its antibody was used on *Chlamydomonas* (see the product sheet for Agrisera antibody AS06-172-100). An asterisk is used to indicate the band at the expected PsaA molecular weight. α -tubulin served as a loading control. Major bands cropped from this panel are also presented in Fig. 4e.

Supplementary Notes

Accuracy of insertion mapping and number of insertions per mutant	3
Deletions, duplications, and junk fragments associated with insertions are small	3
Insertion sites are randomly distributed with mild cold spots and a small number of hot spots.....	4
Absence of insertions identifies over 200 genes potentially essential for growth under the propagation conditions used.....	4
Deleterious mutations rather than differential chromatin configuration are the major cause of insertion density variation.....	5
Disruption of <i>CPL3</i> is the cause of the photosynthetic deficiency in the <i>cpl3</i> mutant	5
Supplementary Methods.....	7
References	18

Accuracy of insertion mapping and number of insertions per mutant.

In *Chlamydomonas* insertional mutants, short “junk fragments” of genomic DNA (likely from lysed cells) are often inserted between the cassette and flanking genomic DNA¹. The difficulty in distinguishing these junk fragments from true flanking genomic DNA can lead to inaccurate mapping of the insertion to a genomic location^{1,2}. Additionally, some cassettes are truncated during insertion, preventing mapping of the flanking sequence on one side. We sought to help users prioritize mutants for characterization by classifying insertions into categories that reflect our confidence in the mapping accuracy, based on two criteria: (1) whether flanking sequences from both sides of the cassette mapped to the same genomic region; and (2) whether the LEAP-Seq reads contained sequences from multiple genomic regions, suggesting the presence of junk DNA fragments inserted next to the cassette (Supplementary Fig. 3a and Supplementary Fig. 2f-j).

A confidence level of 1 was assigned to 19,015 insertions in which both cassette-genome junctions mapped to the same genomic region and were free of junk fragments. A confidence level of 2 was assigned to 5,665 insertions in which both cassette-genome junctions mapped to the same genomic region, after correcting for the presence of a junk fragment at one junction. A mapping confidence level of 3 was assigned to 36,600 insertions in which only one cassette-genome junction could be identified, with the likelihood of junk DNA insertion determined to be low based on fewer than 40% of LEAP-Seq reads containing sequence from multiple genomic regions. A mapping confidence level of 4 was assigned to 13,643 insertions in which only one junction could be identified, and that junction was likely to contain a junk fragment, or the flanking sequence could not be mapped to a unique genomic location. The mapping for these insertions was adjusted to reflect the most likely correct insertion site.

Approximately 95% of confidence level 1 and 2 insertions are mapped correctly based on PCR validation of randomly chosen mutants, compared to ~73% of confidence level 3 and ~58% of confidence level 4 (Supplementary Table 6; Supplementary Note).

Our bioinformatic analyses suggest that over 80% of the mutants harbor only one mapped insertion (Supplementary Fig. 3b), consistent with Southern blot data from randomly chosen mutants (Supplementary Fig. 3c).

Deletions, duplications, and junk fragments associated with insertions are small.

Random insertions in *Chlamydomonas* are sometimes also associated with deletions and duplications of neighboring genomic DNA³. To further help users understand the quality of mutants in this library, we characterized these deletions and duplications by examining the sequences across both junctions of confidence level 1 insertions (Supplementary Note). Of these insertions, 11% had no deletions or duplications, 74% harbored genomic deletions and 15% had genomic duplications. The great majority (98%) of genomic deletions were less than 100 bp, but some were as large as 10 kb. While 98% of the genomic duplications were shorter than 10 bp, some extended to 30bp (Supplementary Fig. 3, d and e). Both the deletions and duplications likely resulted from non-homologous end joining repair that occurs during cassette insertion⁴. Additionally, examining the 651 insertions in which a junk fragment separated two cassettes inserted in the same location allowed us to estimate the typical junk fragment length. Most (73%) junk fragments were

shorter than 300 bp, but some were as large as 1,000 bp (Supplementary Fig. 3f). If larger deletions, duplications or junk fragments were present, they were not sufficiently frequent to allow us to identify them reliably.

Insertion sites are randomly distributed with mild cold spots and a small number of hot spots.

While a random insertion model produced a distribution of insertion sites broadly similar to the observed distribution (Fig. 1c and Supplementary Fig. 4a), we did detect some cold spots and hot spots where insertion density differed significantly from the random insertion model (Supplementary Fig. 4a; Supplementary Table 7; Methods). Cold spots cover 26% of the genome and on average show a 48% depletion of insertions. Hot spots cover 1.5% of the genome and contain 16% of insertions (Methods).

Hot spots fell into two distinct classes that differed in the local distribution of insertions (Supplementary Fig. 4, b and c). In one class, dozens of insertions were found within a region of 20-40 bp. In the other class, the insertions were distributed over a much larger region of 200-1,000 bp. Our observations suggest that hot spots could be caused by two distinct mechanisms; however, we did not observe a correlation between specific features of the genome (e.g. sequence, exon, intron, UTR, mappability) and the occurrence of either class of hot spots.

Absence of insertions identifies over 200 genes potentially essential for growth under the propagation conditions used.

Identification of essential genes in bacteria, fungi, and mammals has revealed important molecular processes in these organisms⁵⁻⁸. We sought to take advantage of the very large set of mapped mutations in the library to identify candidate essential *Chlamydomonas* genes based on the absence of insertions in those genes (Methods). We note that our approach does not allow testing of gene essentiality under all possible conditions. Therefore, it is likely that some of the candidate essential genes we identify in this approach are required specifically for growth under our propagation conditions, but not under all conditions. For example, mutants in respiratory genes would be identified as essential if these mutants were not recovered under our propagation conditions (in the dark on acetate media), although the same mutants could have grown if recovery were under photosynthetic conditions.

Given our average density of insertions, we were able to detect a statistically significant (FDR < 0.05) lack of insertions for genes with a mappable length greater than 5 kb. We identified 203 candidate essential genes (Supplementary Table 9). We caution that this is a conservative list for two reasons: (1) if a gene has a mappable length smaller than 5 kb and has no insertion, its underrepresentation is not statistically significant; (2) some essential genes were not detected because there are insertions incorrectly mapped to them.

Many of these predicted essential genes have homologs that have been shown to be essential in other organisms. For example, Cre01.g029200 encodes a homolog of the yeast cell cycle protease separase ESP1⁹, Cre12.g521200 encodes a homolog of yeast DNA replication factor C complex subunit 1 RFC1¹⁰, and Cre09.g400553 encodes a homolog of the yeast nutrient status sensing kinase Target of Rapamycin 2 TOR2¹¹. In addition, we observed genes encoding proteins involved in acetate utilization or

respiration, such as acetyl-CoA synthetase/ligase¹² (Cre07.g353450) and components of the mitochondrial F1F0 ATP synthase¹³ (Cre15.g635850 and Cre07.g340350). As discussed above, these genes may be essential under the conditions of library propagation, in which acetate serves as the energy source.

We also observed genes on the list with nonessential homologs in other organisms. One example is Cre13.g585301, which encodes monogalactosyldiacylglycerol (MGDG) synthase and whose Arabidopsis homolog MGD1 is not essential¹⁴. This can be explained by the presence of two other isoforms of MGDG synthases in Arabidopsis but not in Chlamydomonas¹⁵. Comparison of our candidate Chlamydomonas essential genes with those of other organisms can provide insights into evolutionary differences across the tree of life.

Deleterious mutations rather than differential chromatin configuration are the major cause of insertion density variation.

One caveat for our above prediction of essential genes is that the lack of insertions could be caused by low chromatin accessibility at those loci to insertional mutagenesis. We reasoned that if chromatin accessibility influenced insertion density, the 3' UTRs of these genes would also be less represented; while if low insertion density primarily reflected essentiality, we would still see many insertions in the 3' UTRs of these genes, because 3' UTR insertions typically do not disrupt gene function (Fig. 3, d and e). For all genes in the genome, we observed an insertion density of 1.1 insertions per mappable kb in exons and introns and 4.7 insertions per mappable kb in 3' UTRs. For the candidate essential genes, despite a lack of insertions in exons and introns, the insertion density in 3' UTRs is 4.1 insertions per mappable kb, similar to that of all genes. We thus conclude that low insertion density in our candidate essential genes is largely caused by mutations that impair mutant fitness instead of low chromatin accessibility to insertional mutagenesis.

Disruption of *CPL3* is the cause of the photosynthetic deficiency in the *cpl3* mutant.

We sought to confirm and characterize the *cpl3* insertion in detail. Our high-throughput LEAP-Seq data suggested that *cpl3* contained an insertion of two back-to-back cassettes. Specifically, the *cpl3* mutant contains two insertion junctions from 3' ends of two cassettes in opposite orientations, within the *CPL3* gene. Junction 1 is confidence level 3 (no junk fragment), and junction 2 is confidence level 4 (with a junk fragment, corrected) (Supplementary Fig. 6a). We successfully confirmed both junctions by PCR (Supplementary Fig. 6b). Sequencing of the product from junction 2 revealed that the end of the cassette has a 10-bp truncation and a 10-bp fragment of unknown origin inserted between the cassette and the *CPL3* gene. The genomic flanking sequence of junction 2 overlaps with the flanking sequence in junction 1 by 2 bp. When we amplified across the insertion site, *cpl3* yielded a product ~3 kb larger than the product from wild type (Supplementary Fig. 6b). Based on these results, the most likely model for this insertion is that two copies of the cassette (at least one truncated) inserted together into the *CPL3* gene in opposite orientations, with a 2-bp genomic duplication at the site of insertion.

To confirm the involvement of *CPL3* in photosynthesis, we cloned *CPL3* genomic DNA and transformed it into the *cpl3* mutant. Based on colony size, photoautotrophic growth was rescued in approximately 14% of the transformants (Supplementary Fig. 6, c

and d), a percentage consistent with previous *Chlamydomonas* genetic studies¹⁶. Three rescued transformants, named comp1-3, were chosen at random for phenotypic confirmation (Fig. 4b) and genotyping. PCR with primers “g3 + g2” demonstrated the disruption of the endogenous *CPL3* locus in the *cpl3* and comp1-3 lines (Supplementary Fig. 6a, b). In comp1-3, PCR across the insertion site of the *cpl3* mutation with primers “g1 + g2” yielded products (expected size: 1,311 bp) that indicate presence of wild-type *CPL3* sequence from the wild-type *CPL3* in the complementation construct, and weak ~4 kb bands from the endogenous *CPL3* locus disrupted by the cassette insertion (Supplementary Fig. 6b). To further confirm that comp1-3 still contained the original insertion in *CPL3*, we amplified the two insertion junctions in the complemented lines with primers “g1 + c1” and “g2 + c1”. These genetic complementation results demonstrate that the disruption of *CPL3* is the cause of the growth defect of the mutant.

Supplementary Methods

This section contains method details that are omitted from the Online Methods section.

Generation of insertion cassettes. The insertion cassette designated Cassette containing Internal Barcodes 1 (CIB1) was generated in four steps: (1) generating double-stranded DNAs containing random sequences (Supplementary Fig. 1a); (2) digesting the double-stranded DNAs to yield cassette ends (Supplementary Fig. 1a); (3) obtaining the backbone from digestion of plasmid pMJ016c that contains the sequences between the two barcodes (Supplementary Fig. 1b); (4) ligating the two cassette ends with the cassette backbone (Supplementary Fig. 1c).

Step 1: To generate each end of the cassette that contains barcodes, a long oligonucleotide primer (Supplementary Fig. 1a and Supplementary Table 1) containing a random sequence region of 22 nucleotides was used as a template for the extension of a shorter oligonucleotide primer. Each 50- μ L reaction mixture contained 32 μ L H₂O, 10 μ L Phusion GC buffer, 1.5 μ L DMSO, 1 μ L 10 mM dNTP, 2.5 μ L 10 μ M long oligo, 2.5 μ L 10 μ M short oligo, and 0.5 μ L Phusion HS II DNA polymerase (F549L, Thermo Fisher). The reaction mixtures were subjected to a single thermal cycle: 98°C for 40 sec, 97°C to 63°C ramp (-1°C every 10 sec), 63°C for 30 sec, 72°C for 5 min.

Step 2: The double-stranded product yielded from Step 1 was digested using *Bsa*I (R0535L, New England Biolabs). For the 5' side primer extension product, the digestion yielded two bands of 87 bp (plus 4 nt of overhang) and 31 bp (plus 4 nt of overhang). For the 3' side, they were 68 bp and 31 bp. The larger band from each digestion was purified from a 2.5% agarose gel using D-tubes (71508-3, EMD Millipore) as previously described¹ (Supplementary Fig. 1a).

Step 3: The synthesized plasmid pMJ016c, which contains the *HSP70-RBCS2* promoter, the paromomycin resistance gene *AphVIII*, and the *PSAD* and *RPL12* terminators, was digested using *Bsa*I. Two bands of 2064 bp and 3363 bp were obtained. The 2064 bp band (cassette backbone) was purified from a 0.8% agarose gel using the QIAquick Kit (28106, Qiagen) according to the manufacturer's instructions (Supplementary Fig. 1b).

Step 4: The two fragments and the cassette backbone were ligated using T4 DNA ligase (M0202L, New England Biolabs) (Supplementary Fig. 1c). Each 30- μ L reaction mixture contained 38 ng 5' cassette end, 30 ng 3' cassette end, 305 ng cassette backbone, 3 μ L ligase buffer, and 0.5 μ L ligase. The double-stranded product of 2,223 bp was gel purified using D-tubes and used for mutant generation. The sequence of the CIB1 cassette generated (Supplementary Fig. 1d) has been uploaded to the mutant ordering website: <https://www.chlamylibrary.org/showCassette?cassette=CIB1>.

Mutant generation, mutant maintenance, and medium recipes. *Chlamydomonas* CC-4533 strain was grown in Tris-Acetate-Phosphate (TAP) medium in a 20-L container under 100 μ mol photons m⁻² s⁻¹ light (measured at the periphery) to a density of 1-1.5x 10⁶ cells/mL. Cells were collected by centrifugation at 300-1,000g for 4 min. Pellets were washed once with 25 mL TAP medium supplemented with 40 mM sucrose, and then resuspended in TAP supplemented with 40 mM sucrose at 2x 10⁸ cells/mL. 250 μ L of cell suspension was then aliquoted into each electroporation cuvette (Bio-Rad) and incubated at 16°C for 5-30 min. For each cuvette, 5 μ L DNA cassette CIB1 at 5 ng/ μ L

was added to the cell suspension and mixed by pipetting. Electroporation was performed immediately as previously described¹. After electroporation, cells from each cuvette were diluted into 8 mL TAP supplemented with 40 mM sucrose and shaken gently in dark for 6 h. After incubation, cells were plated on TAP containing 20 µg/mL paromomycin (800 µL per plate) and incubated in darkness for approximately two weeks before colony picking.

Approximately 210,000 total mutants were picked using a Norgren CP7200 colony picking robot and maintained on 570 agar plates, each containing a 384-colony array. We propagated this original, full library by robotically passaging the mutant arrays to fresh 1.5% agar solidified TAP medium containing 20 µg/mL paromomycin using a Singer RoToR robot (Singer Instruments)². The full collection was grown in complete darkness at room temperature and passaged every four weeks. In this collection, 127,847 of the mutants were mapped. Colonies that failed to yield barcodes or flanking sequences may contain truncated insertion cassettes¹ that have lost the primer binding sites used for barcode amplification or LEAP-Seq analysis. By removing the mutants that were not mapped, mutants that did not survive propagation, and some of the mutants in genes with 20 or more insertions, we consolidated 62,389 mutants into 245 plates of 384-colony arrays for long-term robotic propagation.

The TAP medium was prepared as previously reported¹⁷. The TP medium used in this research was similar to TAP except that HCl instead of acetic acid was used to adjust the pH to 7.5.

Combinatorial pooling. For combinatorial pooling and barcode determination for each mutant colony, 570 plate-pools (each containing all mutants on one plate) and 384 colony-pools (each containing all mutants in the same colony position across all plates) were generated from two separate sets of the library as previously described². Binary error-correcting codes were used to design combinatorial pooling schemes, as previously described². The existence of suitable binary error-correcting codes and their mathematical construction methods were checked using an online database¹⁸. For colony super-pooling, the same 384-codeword subset of the [20,10,6] code as previously employed² was used. For plate super-pooling, the [21,11,6] code was generated by triple shortening of the [24,14,6] code¹⁹. In order to ensure detection of cases of two colonies derived from a single mutant, which could otherwise cause incorrect colony locations to be identified for such mutants, the subset of codewords with a bit sum of 10 (708 codewords) was taken from the [21,11,6] code, using the `choose_codewords_by_bit_sum` function. Both subsets of codewords were checked for the possibility of such sister colony conflicts using the `clonality_conflict_check` function: no conflicts were detected up to 2 errors, meaning any incorrect result due to a sister colony case would have at least 2 differences compared to any expected correct result. The final subset of 570 codewords for plate super-pooling was chosen as previously². The final codeword lists are provided as Supplementary Tables 2 and 3.

Generation of plate-super-pools and colony-super-pools from the plate-pools and colony-pools was performed using the Biomek FX liquid handling robot (Beckman Coulter) as previously described². The instruction files for the Biomek robot were generated using the `robotic_plate_transfer.py` program.

Barcode amplification from super-pools. DNA was extracted from super-pool samples as previously described¹ and the barcodes were amplified (Supplementary Fig. 1f) using the Phusion HSII PCR system. For either 5' or 3' barcode amplifications, one primer (5' R1 or 3' R1; sequences provided in Supplementary Table 1) used in the PCR was common for all super-pools; the other primer (5' R2-1, 5' R2-2,...; 3' R2-1, 3' R2-2,...;) contained an index sequence that allows multiplexed sequencing, i.e. combining of multiple samples in one sequencing lane. Each 50 μ L PCR mixture contained 125 ng genomic DNA, 10 μ L GC buffer, 5 μ L DMSO, 1 μ L dNTPs at 10 mM, 1 μ L (for 5') or 2 μ L (for 3') MgCl₂ at 50 mM, 2.5 μ L of each primer at 10 μ M, and 1 μ L Phusion HSII polymerase. The reaction mixtures were incubated at 98°C for 3 min, followed by 10 three-step cycles (10 sec at 98°C, 25 sec at 58°C or 63°C for 5' and 3' barcodes respectively, and 15 sec at 72°C), and then 8 two-step cycles (10 sec at 98°C, and 40 sec at 72°C). Similar amount of products from three to eight super-pools were combined, purified using MinElute columns (28006, Qiagen), and the product bands (235 bp for 5' and 209 bp for 3') were gel purified. The purified products were sequenced using the Illumina HiSeq platform from a single end with a custom primer (5' Seq and 3' Seq, Supplementary Table 1).

Deconvolution of super-pool sequencing data. The barcode sequences were extracted from the Illumina sequencing data from each super-pool using the cutadapt command-line program²⁰, with a 13 bp expected cassette sequence, allowing 1 alignment error, and taking the trimmed barcode reads between 21 and 23 bp in length. The command for 5' sequences was “cutadapt -a GGCAAGCTAGAGA -e 0.1 -m 21 -M 23”, and for 3' sequences “cutadapt -a TAGCGCGGGGCGT -e 0.1 -m 21 -M 23”. A barcode was found in 97-99% of the sequences in each super-pool.

The reads for each distinct barcode sequence in each super-pool were counted (Supplementary Table 4). Many of the sequenced barcodes are likely to contain PCR or sequencing errors. Such barcodes were left uncorrected, because they are very unlikely to appear in enough super-pools to be deconvolved and included in the final data. The deconvolution based on the read count table was performed as previously described², for 5' and 3' data separately. A single set of optimized (N, x) parameters was chosen for each dataset, with m = 0 in all cases: N = 8 and x = 0.14 for 5' plate-super-pool data, N = 8 and x = 0.16 for 3' plate-super-pool data, N = 6 and x = 0.12 for 5' colony-super-pool data, N = 6 and x = 0.1 for 3' colony-super-pool data. Note that data for colony-super-pool 14 are missing for plates 351-570, which caused imperfections in the deconvolution process, but the missing data were dispensable due to the error-correction capability built into the pooling scheme.

LEAP-Seq. To connect the flanking sequence with the corresponding barcode for each insertion, we performed LEAP-Seq as reported before² except that barcodes in addition to the flanking sequences were included in the amplicons (Supplementary Fig. 1g, and Supplementary Fig. 2f). Genomic DNA of mutants in the library was used as the template for the extension of a biotinylated primer that anneals to the insertion cassette. The primer extension products were purified by binding to streptavidin-coupled magnetic beads and then ligated to a single-stranded DNA adapter. The ligation products were then

used as templates for PCR amplification. The PCR products were gel-purified before being submitted for deep sequencing.

We tried different combinations of primers and attempted to perform LEAP-Seq either on six sub-pools (each containing mutants from one-sixth of the library) separately or on the entire library in a single reaction (Supplementary Table 1). Sequencing results from all the samples were used in the analyses below.

Basic LEAP-Seq data analysis. The LEAP-Seq samples were sequenced with Illumina Hi-Seq, yielding paired-end reads. Each read pair has a proximal side, containing the barcode, a part of the cassette sequence, and the immediate genomic flanking sequence; and a distal side, containing the genomic sequence a variable distance away (Supplementary Fig. 2f-j).

A newly developed method was used to separate cassette sequence from the proximal reads and thus identify the barcode and genomic flanking sequence even in cases where the cassette was truncated. This was done using the `deepseq_strip_cassette.py` script, which uses local bowtie2 alignment²¹ to detect short cassette sequence. A bowtie2 alignment was performed against the expected cassette sequence (GGAGACGTGTTTCTGACGAGGGCTCGTGTGACTAGTGAGTCCAAC for 5' reads and ACTGACGTCGAGCCTTCTGGCAGACTAGTTGCTCCTGAGTCCAAC for 3' reads), using the following bowtie2 options: “--local --all --ma 3 --mp 5,5 --np 1 --rdg 5,3 --rfg 4,3 --score-min C,20,0 -N0 -L5 -i C,1,0 -R5 -D30 --norc --reorder”. The alignments for each proximal read were filtered to only consider cases where the cassette aligns after a 21-23 bp barcode, at most 5 bp of expected initial cassette sequence are missing, and at least 10 bp of expected cassette sequence are aligned with at most 30% errors. Out of the filtered alignments, the best one was chosen in a maximally deterministic manner, in order to ensure that multiple reads of the same insertion junction yield the same result. The alignment with the highest alignment score is chosen (the bowtie scoring function was customized to distinguish between as many cases as possible); if there were multiple alignments with the same score, the one with the longer alignment was chosen.

The resulting cassette alignment was then removed from each proximal read, with the section before the cassette being considered the barcode and the section after the cassette being considered the genomic flanking region. The resulting genomic proximal reads and the raw genomic distal reads were trimmed to 30 bp using the `fastx_trimmer` command-line utility (http://hannonlab.cshl.edu/fastx_toolkit), aligned to the *Chlamydomonas* genome (version 5.5 from Phytozome²²) and the cassette, and the alignments were filtered to yield a single result using `deepseq_alignment_wrapper.py`, as previously described¹.

The barcode sequences and proximal and distal alignment results were merged into a single dataset, with data grouped into insertion junctions based on the barcode, using the `add_RISCC_alignment_files_to_data` function. Data relating to barcodes that were not present in the combinatorial deconvolution results were discarded. The gene-related information for each insertion junction was added using the `find_genes_for_mutants` and `add_gene_annotation` functions. All functions in this paragraph are methods of the `Insertional_mutant_pool_dataset` class in the `mutant_IB_RISCC_classes.py` module.

Detecting pairs of flanking sequences that correspond to two sides of the same insertion (confidence levels 1 or 2). Pairs of insertion junctions likely derived from two sides of the same insertion were identified using the `deconvolution_utilities.get_matching_sides_from_table` function, using the method previously described², with an additional distance bin of 1-10 kb. The resulting pair counts were as follows:

	0 bp	1-10 bp	11-100 bp	101 bp - 1 kb	1-10 kb	10+ kb
Inner-cassette (toward-facing)	3935	17708	7866	737	339	540
Outer-cassette (away-facing)	-	5010	188	560	58	494
Same-direction	13	17	40	158	133	1520

Additionally, there were 22,247 pairs in which the two junctions were mapped to different chromosomes.

The number of inner-cassette pairs is significantly larger than 50% of the number of same-direction pairs in all size ranges up to 10 kb, implying that most of the inner-cassette pairs in those size ranges are derived from a single insertion with a genomic deletion corresponding to the distance. This can be further confirmed by looking at the indicators of the probability of correct mapping for the insertion junctions: insertions with both sides mapped to the same region are almost certainly correctly mapped, and therefore independent indications of their correct mapping should be higher than for other insertions. As expected, the inner-cassette pairs up to 10 kb have a higher fraction of very high confidence insertion pairs (with both sides having 70% or more read pairs mapping to the same locus, and 500 bp or higher longest distance spanned by such read pairs): for size ranges up to 10 kb, 37-41% of the pairs are very high confidence, while for 10+ kb the number is only 16%.

The number of outer-cassette pairs is significantly larger than 50% of the number of same-direction pairs in size ranges between 1 bp and 1 kb, implying that most of the outer-cassette pairs in those size ranges are derived from a single insertion. There are two possible physical interpretations of a single insertion yielding an outer-cassette pair of insertion junctions: (1) an insertion with a genomic duplication causing the same genomic DNA sequence to be present on both sides of the cassette (potentially due to single-strand repair); and (2) an insertion of two cassettes flanking a “junk” fragment of genomic DNA. The 1-10 bp cases must be a genomic duplication, since a 1-10 bp “junk” fragment could not yield a 30 bp flanking sequences aligning to the genome. This is confirmed by 41% of the pairs being very high confidence. The 101 bp-1 kb cases are almost certainly insertions of two cassettes flanking a “junk” fragment, based on only 3.8% of them being very high confidence. The 188 11-100 bp cases, with a 27% very high confidence, are likely split between the two categories; based on previous analysis¹ we used 30 bp as the cutoff between cases 1 and 2 for outer-cassette pairs. The case 2 pairs, i.e. insertions of two cassettes flanking a junk fragment, were used to determine the typical range of lengths of junk fragments (Supplementary Fig. 3f).

Based on this analysis, all insertion junction pairs likely to be derived from two sides of the same insertion (inner-cassette up to 10 kb and outer-cassette up to 30 bp) were categorized as confidence level 1 (extremely likely to be correctly mapped) because their mapping position is derived from two independent flanking sequences. They were annotated in Supplementary Table 5 as confidence level 1, and the “if_both_sides” column was set to “perfect” for the 0 bp distance cases, “deletion” for the remaining inner-cassette cases, and “duplication” for the outer-cassette cases.

A similar type of analysis was performed to look for pairs of insertion junctions derived from two sides of an insertion with a junk fragment. For each pair of insertion junctions in one colony (except pairs of insertion junctions already identified as two sides of the same insertion), we looked at the distance and relative orientation between the proximal read of the first junction and each distal read from the second junction; cases where the distal read was mapped to within 10 kb of the proximal read were counted as matches. We repeated the process with the first and second junctions swapped. To simplify the analysis, two cases were ignored: colonies with matches between more than two insertions (~12% of match cases), and insertion pairs where the proximal read of one insertion was a match to multiple distal reads of the other insertion with different orientations (~3% of match cases). We then took the distance to the closest distal read, and counted the cases by orientation and distance, as before:

	0-10 bp	11-100 bp	101 bp - 1 kb	1-10 kb
Inner-cassette (toward-facing)	11	5072	5787	289
Outer-cassette (away-facing)	28	140	152	82
Same-direction	6	185	283	195

Note that the distances are expected to be higher in this case, because if we are looking at a case of two sides of one insertion with a junk fragment, the distal read will be a variable distance away from the junk-genome junction which is the actual insertion location. So even for insertions with no genomic deletion/duplication, the distance between the proximal read on one side and the nearest distal read on the other side will not be 0 bp.

The number of inner-cassette cases up to 1 kb is more than 10x larger than the number of same-direction cases, so these insertion pairs are extremely likely to be two sides of one insertion with a junk fragment (and possibly a genomic deletion). Thus, all the pairs in this category were identified as confidence level 2, which are extremely likely to be correctly mapped.

The number of inner-cassette cases with a distance of 1-10 kb and the number of outer-cassette cases with a distance of 0-10 bp is also higher than the expected 50% of the same-direction cases, suggesting that many of them are also two sides of the same insertion, but the differences are less dramatic and thus the number of false positives would be too high for us to be comfortable identifying all these pairs as confidence level 2.

The insertion position information for junk fragment sides of confidence level 2 insertions originally reflected the junk fragment rather than the actual genomic insertion position. We corrected it to show the nearest distal read matching the non-junk side: the flanking sequence and position was changed to that of that distal read; the

“LEAPseq_distance” field was changed to the longest distance between two distal reads that mapped to the presumed real insertion position (i.e. to the same region as the proximal read of the insertion junction from the other side); the remaining LEAPseq fields were likewise changed to reflect the numbers of distal reads and positions mapped to the presumed real insertion position. For confidence level 2 insertions, the “if_both_sides” column was set to “with-junk”; for the sides with a junk fragment, the “if_fixed_position” column was set to “yes_nearest_distal”, and for the sides without a junk fragment it was kept as “no”.

The confidence level 1 and 2 insertions (counting only the non-junk side of the confidence level 2 insertions) appear to be of high quality (Supplementary Fig. 2h).

Categorizing the remaining insertions and correcting junk fragments (confidence levels 3 and 4). After identifying the highest-confidence insertion junctions, i.e. those with two matching sides of the same insertion, we sought to separate the remaining insertions (with only one side mapped) into a set with a high likelihood of having correctly mapped genomic insertion positions and a set with insertion positions likely to reflect junk fragments. We considered two factors to separate these two sets: (1) the percentage of read pairs that map to the same locus, and (2) the longest distance spanned by such a read pair (Supplementary Fig. 2, i and j). We decided to solely use the first factor based on the fact that nearly all of the insertions with low distances but high percentage of read pairs mapped to the same locus were ones with relatively few LEAP-Seq reads, indicating that their short distances spanned are likely due to them having few reads (and thus a lower chance of a long read) rather than to a junk fragment. Therefore we decided to use the percentage of read pairs mapping to the same locus as the only factor in distinguishing the higher and lower confidence insertion sets, because that factor is independent of the number of reads. To determine what cutoff would be appropriate, we took advantage of the already known confidence level 1 insertions. We calculated the fraction of confidence level 1 pairs among all the colonies with exactly two insertions (two insertions are required for a confidence level 1 pair) as an approximate lower bound on the number of correctly mapped insertions. Over the entire dataset, this fraction is 65%; when calculated only on insertions with at least 50% read pairs mapping to the same locus, it's 78%; for insertions with at least 60%, 70%, 80% and 90% read pairs mapping to the same locus, it is 79%. Thus it is clear that using a cutoff anywhere in the 50-90% range significantly improves the quality of the dataset, regardless of the exact position of the cutoff. This makes sense, because the 50-90% range constitutes a very small fraction of all insertions. We opted to use 60% as the cutoff for confidence level 3, i.e. insertions with only one mapped side but with LEAP-Seq data indicating very likely correct mapping.

The remaining insertions, with below 60% read pairs mapping to the same locus and thus with the proximal LEAP-Seq read likely to be part of a junk fragment, were analyzed further to identify the most likely true insertion position. The same analysis was applied to all insertions with the proximal LEAP-Seq read with no genomic alignment (possibly due to a very short junk fragment resulting in the 30 bp proximal read being a hybrid of the junk fragment sequence and genomic sequence from the real insertion position, or simply due to PCR or sequencing errors yielding an unmappable sequence), or with multiple equally good genomic alignments (which could be derived

from the real genomic location, but in a non-unique region of the genome, requiring the use of distal reads to determine the correct insertion location), or mapped to the insertion cassette (indicating a second cassette fragment inserted between the first cassette and the genome, which can be treated the same way as a junk genomic DNA fragment).

In order to determine the best method of identifying the true insertion location based on the full distal LEAP-Seq read data, we grouped the distal LEAP-Seq reads for each insertion into regions no more than 3 kb in size. For each such group, we calculated three measures that we thought might be the best method of identifying the real insertion location: (1) the number of reads in the group, (2) the number of unique genomic positions to which reads in the group were mapped, and (3) the distance spanned by the reads. LEAP-Seq reads mapped to the insertion cassette, or with no unique mapping to the genome, were excluded. In order to determine which method was the best, we used the junk fragment sides of confidence level 2 insertions, since for those the distal reads corresponding to the true genomic insertion locations had already been determined by an independent method (i.e. by matching the proximal read of the other side of the insertion). For each of the three methods listed above, the insertion location predicted by the method was compared to the known insertion location of each confidence level 2 insertion with a junk fragment. The results were as follows: 90% of the known insertion positions were correctly predicted by taking the region with the most total distal reads, 84% by taking the region with the most unique mapping positions, and 84% by taking the region with the longest distance spanned by the reads. Thus, the total number of distal reads was chosen as the most likely measure to yield the correct genomic insertion position of insertions with a junk fragment.

This method was then applied to all the insertions listed in the previous paragraph, yielding the most likely true location for each insertion; insertions with only a single LEAP-Seq distal read in each region were excluded, because one read did not provide enough data to determine the insertion position with any confidence. For some insertions, the region with the most distal LEAP-Seq reads also included the proximal LEAP-Seq read - in those cases, the original insertion position based on the proximal LEAP-Seq read was left unchanged. It is still possible that this position reflects a relatively long junk fragment rather than the true genomic insertion position, but we did not have enough data to distinguish those cases from high confidence. Likewise, it is possible that the corrected position with the most distal LEAP-Seq reads that do not match the proximal read reflects a second long junk fragment inserted after the first junk fragment which contains the proximal read (we know that insertions with multiple junk fragments can happen), but given the limited length of Illumina-sequenced LEAP-Seq reads, we cannot detect those cases with certainty, and have to limit ourselves to finding putative insertion positions that have a reasonably high probability of being correct.

Additionally, it turned out that many corrected positions for insertions originally mapped to the insertion cassette did not appear to be high-quality, with only a small fraction of distal reads mapped to the putative real insertion position. After looking at several such cases in detail, we concluded that they had not been analyzed correctly. They had single LEAP-Seq reads mapped to multiple distant locations on many chromosomes, compared to 100+ reads mapped to many cassette locations, with the putative real insertion position identified due to two or three single LEAP-Seq reads mapped close together on one chromosome. The uniformly low read numbers of genome-

mapped reads compared with the high read numbers of cassette-mapped reads led us to conclude that the genome-mapped reads were results of PCR or sequencing errors or other artifacts, rather than being derived from real LEAP-Seq products, which should usually yield more than one read. Thus, those appeared to be cases where no LEAP-Seq products sequenced past the additional cassette fragment - this could be expected, because the full cassette is >2.2 kb in length, whereas vanishingly few LEAP-Seq reads are over 1.5 kb. In contrast, junk genomic DNA fragments are mostly smaller than 500 bp and all identified ones were below 1 kb, so this problem would not be expected to be common in genomic junk fragment cases. Indeed a cluster of low-matching-read-percent insertions was not observed in the corrected insertion positions in that category. We decided to exclude this category of incorrectly mapped insertions by only including corrected originally cassette-mapped insertions if >50% of the distal LEAP-Seq reads mapped to the putative correct insertion location.

All the insertions included in the final results of this analysis were annotated as confidence level 4. The final confidence level 4 insertions are of a relatively high quality (Supplementary Fig. 2j). The positions, flanking sequences and LEAP-Seq data of the corrected confidence level 4 insertions in Supplementary Table 5 were changed to reflect the new insertion position, in the same way as for the junk fragment sides of the confidence level 2 insertions above. An additional complication of the new corrected insertion positions was presented by the fact that the position of the nearest distal LEAP-Seq read is always at some distance from the true insertion position, depending on the length of the LEAP-Seq read. We attempted to correct for this by using confidence level 1 insertions to determine the average distance between the proximal read (reflecting the true insertion position) and the nearest distal read, separately for 5' and 3' datasets, depending on the total number of LEAP-Seq reads for the insertion (binned into ranges: 1, 2, 3, 4-5, 6-10, 11-20, 21+ total reads). For each confidence level 4 insertion with a corrected position, the position was further adjusted by the average distance for the correct side and number of reads as calculated above. This distance was appended as a number to the value in the "if_fixed_position" field for each insertion in Supplementary Table 5.

Barcode sequencing and data analysis for pooled screens. Barcodes were amplified and sequenced using the Illumina HiSeq platform as performed on the combinatorial super-pools in library mapping (Supplementary Fig. 1f). Initial reads were trimmed using cutadapt version 1.7.1²⁰. Sequences were trimmed using the command "cutadapt -a <seq> -e 0.1 -m 21 -M 23 input_file.gz -o output_file.fastq", where seq is GGCAAGCTAGAGA for 5' data and TAGCGCGGGGCGT for 3' data. Barcodes were counted by collapsing identical sequences using "fastx_collapser" (http://hannonlab.cshl.edu/fastx_toolkit). The barcode read counts for each dataset were normalized to a total of 100 million (Supplementary Table 10).

For evaluation of the quantitiveness of our barcode sequencing method, barcodes obtained from two technical replicate aliquots of the same initial pool were compared in read counts (Supplementary Fig. 5a). Barcodes obtained from the two TP-light cultures at the end of growth were compared to assess consistency between biological replicates (Fig. 3b).

To detect deficiency in photosynthetic growth, we compared mutant abundances in TP-light with TAP-dark at the end of growth (Fig. 3c). As a quality control, different barcodes in the same mutant were compared in the ratio of the TP-light read count to TAP-dark read count. Highly consistent ratios were observed (Supplementary Fig. 5b).

For the identification of photosynthetically deficient mutants, each barcode with at least 50 normalized reads in the TAP-dark dataset was classified as a hit if its ratio of normalized TP-light:TAP-dark read counts was 0.1 or lower, or a non-hit otherwise. The fraction of hit barcodes was 3.3% in replicate 1 and 2.9% in replicate 2. These barcodes represent 2,638 and 2,369 mutants showing a growth defect in the TP-light-I and TP-light-II replicates, respectively. A total of 3,109 mutants covering 2,599 genes showed a growth defect in either of the TP-light sample.

Identification and annotation of the hit genes from the screen. To evaluate the likelihood that a gene is truly required for photosynthesis, we counted the number of alleles for this gene with and without a phenotype, including exon/intron/5'UTR insertions. If the insertion was on the edge of one of those features, or in one of the features in only one of the splice variants, it was still counted. We excluded alleles with insertions in the 3' UTRs, which we observed to less frequently cause a phenotype (Fig. 3, d and e). In cases of multiple barcodes in the same mutant (likely two sides of one insertion), the one with a higher TAP-dark read count was used for the calculation of normalized TP-light:TAP-dark read counts, to avoid double-counting a single allele. For each gene, a *P* value was generated using Fisher's exact test comparing the numbers of alleles in that gene with and without a phenotype to the numbers of all insertions in the screen with and without a phenotype (Supplementary Table 11). A false discovery rate (FDR) correction was performed on the *P* values using the Benjamini-Hochberg method²³, including only genes with at least 2 alleles present in the screen. Thus, genes with a single allele have a *P* value but lack a FDR.

This process was performed for both TP-light replicates. The list of higher-confidence genes was generated by taking genes with FDR of 0.27 or less in either replicate - this threshold includes all genes with 2 hit alleles and 0 non-hit alleles. The resulting list of hits included 37 genes in replicate 1, 34 in replicate 2, 44 total. The FDR values for the higher-confidence genes in both replicates are shown in Tables 1 and 2. Additionally, the list of lower-confidence genes was generated by taking genes with a *P* value of 0.058 or less - this value was chosen to include genes with only one allele with a phenotype and no alleles without a phenotype, but to exclude genes with one allele with and one without a phenotype. The resulting list included 264 genes total (210 in replicate 1, 196 in replicate 2).

One gene in the original higher-confidence list and four genes in the original lower-confidence list encode subunits of the plastidic pyruvate dehydrogenase. Mutants in these genes require acetate to grow because they cannot generate acetyl-CoA from pyruvate but can generate acetyl-CoA from acetate. This requirement for acetate, rather than a defect in photosynthesis, likely explains why mutants in this gene showed a growth defect in TP-light³. Removal of these genes led to a final list of 43 higher-confidence genes and 260 lower-confidence genes (Fig. 3f, Tables 1 and 2, and Supplementary Table 12).

We identified 65 (22 higher-confidence and 43 lower-confidence) out of the 303 hit genes as “known” genes based on genetic evidence: mutation of this gene in *Chlamydomonas* or another organism caused a defect in photosynthesis. Among the remaining 238 “candidate” genes (21 higher-confidence ones and 217 lower-confidence ones), some genes appear to be related to photosynthesis because of their predicted chloroplast localization or evolutionary conservation among photosynthetic organisms²⁴, despite lack of solid genetic evidence. For three of the candidate genes (*CGL59*, *CPL3*, and *VTE5*), mutants with insertions adjacent to them or in their 3' UTRs were previously found to be acetate-requiring or hypersensitive to oxidative stress in the chloroplast³.

Analysis of candidate gene enrichment in reported transcriptional clusters related to photosynthesis. Two transcriptome datasets in *Chlamydomonas* were used in this analysis: a diurnal regulation study²⁵ and a dark-to-light transition study²⁶. For the first one, we chose the diurnal cluster 4 in the study that had photosynthesis-related genes enriched in it²⁵. For the second one, we chose the genes upregulated upon transition to light²⁶. In each case, the number of candidate genes included and not included in the regulated gene sets was compared to the total number of *Chlamydomonas* genes included and not included in the cluster, using Fisher's exact test. The resulting *P* values were FDR-adjusted using the Benjamini-Hochberg method²³.

Microscopy. Cells were grown under the TAP-dark condition to log phase and concentrated ten-fold before microscopic analysis. Aliquots were deposited at the corner of a poly-L-lysine coated microslide well (Martinsried) and spread over the bottom of the well by overlaying with TAP-1% agarose at low temperature (<30°C), to minimize cell motion during image acquisition. Cells were imaged at room temperature though a Leica TCS SP5 laser scanning confocal microscope and an inverted 100x NA 1.46 oil objective. Chlorophyll fluorescence signal was generated using 514 nm excitation, and 650-690 nm collection. All images were captured using identical laser and magnification settings (4x zoom and single-slice through the median plane of the cell). Composite images (chlorophyll fluorescence overlay with bright field) were generated with Fiji²⁷.

References

1. 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).
2. Li, X. *et al.* An Indexed, Mapped Mutant Library Enables Reverse Genetics Studies of Biological Processes in *Chlamydomonas reinhardtii*. *Plant Cell* **28**, 367-87 (2016).
3. 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).
4. Vu, G.T. *et al.* Repair of Site-Specific DNA Double-Strand Breaks in Barley Occurs via Diverse Pathways Primarily Involving the Sister Chromatid. *Plant Cell* **26**, 2156-2167 (2014).
5. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387-91 (2002).
6. Peters, J.M. *et al.* A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. *Cell* **165**, 1493-506 (2016).
7. Rubin, B.E. *et al.* The essential gene set of a photosynthetic organism. *Proc Natl Acad Sci U S A* **112**, E6634-43 (2015).
8. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-101 (2015).
9. Baum, P., Yip, C., Goetsch, L. & Byers, B. A yeast gene essential for regulation of spindle pole duplication. *Mol Cell Biol* **8**, 5386-97 (1988).
10. Cullmann, G., Fien, K., Kobayashi, R. & Stillman, B. Characterization of the five replication factor C genes of *Saccharomyces cerevisiae*. *Mol Cell Biol* **15**, 4661-71 (1995).
11. Kunz, J. *et al.* Target of rapamycin in yeast, TOR2, is an essential phosphatidylinositol kinase homolog required for G1 progression. *Cell* **73**, 585-96 (1993).
12. Spalding, M.H. The CO₂-Concentrating Mechanism and Carbon Assimilation. in *The Chlamydomonas Sourcebook*, Vol. 2 (eds. E.H., H., E.B., S. & G.B., W.) 257-301 (Academic Press, 2009).
13. Devenish, R.J., Prescott, M. & Rodgers, A.J. The structure and function of mitochondrial F1F0-ATP synthases. *Int Rev Cell Mol Biol* **267**, 1-58 (2008).
14. Jarvis, P. *et al.* Galactolipid deficiency and abnormal chloroplast development in the Arabidopsis MGD synthase 1 mutant. *Proc Natl Acad Sci U S A* **97**, 8175-9 (2000).
15. Riekhof, W.R., Sears, B.B. & Benning, C. Annotation of genes involved in glycerolipid biosynthesis in *Chlamydomonas reinhardtii*: discovery of the betaine lipid synthase BTA1Cr. *Eukaryot Cell* **4**, 242-52 (2005).
16. Wang, L. *et al.* Chloroplast-mediated regulation of CO₂-concentrating mechanism by Ca²⁺-binding protein CAS in the green alga *Chlamydomonas reinhardtii*. *Proc Natl Acad Sci U S A* **113**, 12586-12591 (2016).
17. Kropat, J. *et al.* A revised mineral nutrient supplement increases biomass and growth rate in *Chlamydomonas reinhardtii*. *Plant J* **66**, 770-80 (2011).
18. Grassl, M. Bounds on the minimum distance of linear codes and quantum codes. Vol. 2017 (<http://www.codetables.de/>, 2015).
19. Simonis, J. The [23; 14; 5] Wagner code is unique. *Discrete Mathematics* **213**, 269-282 (2000).
20. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10-12 (2011).
21. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
22. Merchant, S.S. *et al.* The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245-251 (2007).
23. 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).
24. 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**, 21427-39 (2011).

25. 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* (2015).
26. Duanmu, D. *et al.* Retrograde bilin signaling enables *Chlamydomonas* greening and phototrophic survival. *Proc Natl Acad Sci U S A* **110**, 3621-6 (2013).
27. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-82 (2012).