SUPPLEMENTARY INFORMATION

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

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

*e-mail: mjonikas@princeton.edu

Supplementary Notes

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 DNA[1]. 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 DNA[3]. 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 insertion[4]. 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 organisms[5-8]. 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 ESP1[9], Cre12.g521200 encodes a homolog of yeast DNA replication factor C complex subunit 1 RFC1[10], and Cre09.g400553 encodes a homolog of the yeast nutrient status sensing kinase Target of Rapamycin 2 TOR2[11]. In addition, we observed genes encoding proteins involved in acetate utilization or

respiration, such as acetyl-CoA synthetase/ligase[12] (Cre07.g353450) and components of the mitochondrial F1F0 ATP synthase[13] (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 essential[14]. This can be explained by the presence of two other isoforms of MGDG synthases in Arabidopsis but not in Chlamydomonas[15]. 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 studies[16]. 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-µL reaction mixture contained 32 µL H_2O, 10 µL Phusion GC buffer, 1.5 µL DMSO, 1 µL 10 mM dNTP, 2.5 µL 10 µM long oligo, 2.5 µL 10 µM short oligo, and 0.5 µL 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 described[1] (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-µL reaction mixture contained 38 ng 5' cassette end, 30 ng 3' cassette end, 305 ng cassette backbone, 3 µL ligase buffer, and 0.5 µL 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 µmol photons m^{-2} s^{-1} light (measured at the periphery) to a density of 1-1.5x 10^6 cells/mL. Cells were collected by centrifugation at 300-1,000*g* 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 10^8 cells/mL. 250 µL 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 µL DNA cassette CIB1 at 5 ng/µL

was added to the cell suspension and mixed by pipetting. Electroporation was performed immediately as previously described[1]. 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 cassettes[1] 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 reported[17]. 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 described[2]. Binary error-correcting codes were used to design combinatorial pooling schemes, as previously described[2]. The existence of suitable binary error-correcting codes and their mathematical construction methods were checked using an online database[18]. For colony super-pooling, the same 384-codeword subset of the [20,10,6] code as previously employed[2] was used. For plate super-pooling, the [21,11,6] code was generated by triple shortening of the [24,14,6] code[19]. 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 previously[2]. 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 described[2]. 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 described[1] 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 μL PCR mixture contained 125 ng genomic DNA, 10 μL GC buffer, 5 μL DMSO, 1 μL dNTPs at 10 mM, 1 μL (for 5') or 2 μL (for 3') $MgCl_2$ at 50 mM, 2.5 μL of each primer at 10 μM, and 1 μL 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 program[20], 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 described[2], 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 before[2] 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 alignment[21] 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 Phytozome[22]) and the cassette, and the alignments were filtered to yield a single result using deepseq_alignment_wrapper.py, as previously described[1].

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 described[2], 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 analysis[1] 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 1.7.1[20]. 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 quantitativeness 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 method[23], 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-light[3]. 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 organisms[24], 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 chloroplast[3].

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 study[25] and a dark-to-light transition study[26]. For the first one, we chose the diurnal cluster 4 in the study that had photosynthesis-related genes enriched in it[25]. For the second one, we chose the genes upregulated upon transition to light[26]. 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 method[23].

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 Fiji[27].

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).