The Plant Cell, Vol. 26: 1398–1409, April 2014, www.plantcell.org © 2014 American Society of Plant Biologists. All rights reserved.

LARGE-SCALE BIOLOGY ARTICLE

High-Throughput Genotyping of Green Algal Mutants Reveals Random Distribution of Mutagenic Insertion Sites and Endonucleolytic Cleavage of Transforming DNA[W][OPEN]

Ru Zhang,[1] Weronika Patena,[1] Ute Armbruster, Spencer S. Gang, Sean R. Blum, and Martin C. Jonikas[2]

Carnegie Institution for Science, Department of Plant Biology, Stanford, California 94305

ORCID IDs: 0000-0002-8388-1303 (W.P.); 0000-0002-9519-6055 (M.C.J.)

A high-throughput genetic screening platform in a single-celled photosynthetic eukaryote would be a transformative addition to the plant biology toolbox. Here, we present ChlaMmeSeq (*Chlamydomonas MmeI*-based insertion site Sequencing), a tool for simultaneous mapping of tens of thousands of mutagenic insertion sites in the eukaryotic unicellular green alga *Chlamydomonas reinhardtii*. We first validated ChlaMmeSeq by in-depth characterization of individual insertion sites. We then applied ChlaMmeSeq to a mutant pool and mapped 11,478 insertions, covering 39% of annotated protein coding genes. We observe that insertions are distributed in a manner largely indistinguishable from random, indicating that mutants in nearly all genes can be obtained efficiently. The data reveal that sequence-specific endonucleolytic activities cleave the transforming DNA and allow us to propose a simple model to explain the origin of the poorly understood exogenous sequences that sometimes surround insertion sites. ChlaMmeSeq is quantitatively reproducible, enabling its use for pooled enrichment screens and for the generation of indexed mutant libraries. Additionally, ChlaMmeSeq allows genotyping of hits from *Chlamydomonas* screens on an unprecedented scale, opening the door to comprehensive identification of genes with roles in photosynthesis, algal lipid metabolism, the algal carbon-concentrating mechanism, phototaxis, the biogenesis and function of cilia, and other processes for which *C. reinhardtii* is a leading model system.

INTRODUCTION

The ability to phenotype genome-wide collections of single-celled mutants has revolutionized our understanding of cellular processes in bacteria, yeast, and animal cells (Winzeler et al., 1999; Ozawa et al., 2005; van Opijnen et al., 2009; Cipriani and Piano, 2011). An analogous platform in a photosynthetic eukaryote would open doors to rapid identification of all the genes required for photosynthesis and any other phenotype of interest and would allow grouping of genes into pathways using chemical genomics (Hillenmeyer et al., 2008). As a first step toward these goals, we present ChlaMmeSeq (*Chlamydomonas MmeI*-based insertion site Sequencing), a tool that enables high-throughput genotyping in the single-celled green alga *Chlamydomonas reinhardtii*.

C. reinhardtii has immense potential as a functional genomics platform for studying photosynthesis and other processes: (1) The similarity of its photosynthetic apparatus to that of land plants has enabled the discovery and characterization of key components of photosynthesis (Schmidt et al., 1977; Shepherd et al., 1979; Niyogi et al., 1997; Fleischmann et al., 1999). (2) Mutants deficient in photosynthesis can be maintained in the dark on acetate, unlike in most photosynthetic organisms. (3) It has a carbon-concentrating mechanism, which facilitates CO_2 fixation (Wang et al., 2011) and is of interest as a potential source of genes to enhance photosynthesis in C3 plants. (4) It accumulates lipids under stress conditions, which makes it an excellent model for studying pathways in algal lipid metabolism (Wang et al., 2009; Merchant et al., 2012), which is of interest in biofuel research. (5) Its eyespot and two flagella make it a great model to study phototaxis (Pazour et al., 1995). (6) Vegetative cells are haploid, so mutant phenotypes are visible immediately. (7) It has a short doubling time (6 to 8 h) and sexual life cycle (2 weeks). (8) Its nuclear and organellar genomes are transformable (Boynton et al., 1988; Kindle et al., 1989; Randolph-Anderson et al., 1993) and have been sequenced (GenBank accession number U03843; Maul et al., 2002; Merchant et al., 2007).

While large numbers of *C. reinhardtii* mutants can easily be generated by random insertion of a drug resistance cassette (Shimogawara et al., 1998), only a handful of insertion sites can be identified and studied at once using standard genetic techniques. Protocols based on plasmid rescue (Tam and Lefebvre, 1993), thermal asymmetric interlaced PCR (Dent et al., 2005; González-Ballester et al., 2005b), restriction enzyme site-directed amplification PCR (González-Ballester et al., 2005a), 3′-rapid amplification of cDNA ends (Meslet-Cladière and Vallon, 2012), and SiteFinding PCR (Li et al., 2012), have been successfully used to identify genomic sequences flanking the cassette, but the labor required has limited their application to dozens of insertions at a time.

Here, we present ChlaMmeSeq, a robust strategy for simultaneous genotyping of tens of thousands of *C. reinhardtii* insertional mutants. We validated the approach by in-depth analysis of 15

[1] These authors contributed equally to this work.

[2] Address correspondence to mjonikas@carnegiescience.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Martin C. Jonikas (mjonikas@carnegiescience.edu)

[W] Online version contains Web-only data.

[OPEN] Articles can be viewed online without a subscription.

www.plantcell.org/cgi/doi/10.1105/tpc.114.124099

individual mutants. We then applied ChlaMmeSeq to a pool of mutants and identified 11,478 distinct insertions, covering 39% of the 17,737 protein-coding genes annotated in the v5.3 *C. reinhardtii* nuclear genome (Merchant et al., 2007; Goodstein et al., 2012). The data reveal that insertion sites are distributed in the genome in a manner largely indistinguishable from random, which allows us to predict genome coverage for larger collections of mutants. Analysis of flanking sequences from tandem cassette insertions reveals that an endonucleolytic activity acts on the transforming DNA, leading us to propose an improved model for events during transformation. The abundance of mutants in a pool measured by ChlaMmeSeq is quantitatively reproducible, opening the door to genome-wide biological enrichment screens (Carette et al., 2011) and the generation of indexed mutant collections (Goodman et al., 2009) in green algae.

RESULTS

An *Mme*I-Based Strategy to Map Insertion Sites in High Throughput

We generated tens of thousands of insertional mutants by electroporating (Shimogawara et al., 1998) the *C. reinhardtii* CMJ030 strain with a DNA cassette encoding resistance to paromomycin (Sizova et al., 2001). Strain CMJ030 (deposited in the Chlamydomonas Resource Center as CC-4533) was isolated from the progeny of a cross between strains D66$^+$ (Schnell and Lefebvre, 1993) and 4A$^-$ (Tran et al., 2012). CMJ030 can grow photoautotrophically, mixotrophically, and heterotrophically (Supplemental Figure 1), is mating type minus (Supplemental Figure 2A), has normal swimming and lipid accumulation, has high transformation efficiency, and recovers efficiently from cryogenic storage in liquid nitrogen (Crutchfield et al., 1999) (Supplemental Figure 2B and Supplemental Table 1). These qualities make it a desirable strain for high-throughput screens for a wide range of phenotypes of interest.

To map insertion sites, we developed ChlaMmeSeq, a strategy for extracting genomic flanking sequences in parallel from tens of thousands of mutants in a pooled sample (Figure 1). ChlaMmeSeq builds upon technologies that were previously demonstrated in bacteria (Goodman et al., 2009; van Opijnen et al., 2009), with modifications to overcome challenges due to the larger genome and different insertion mechanism (see Methods for details). Genomic DNA from mutants is digested with the Type IIS restriction enzymes *Mme*I and *Bsg*I, which yield fragments containing the ends of the cassettes and 20 to 21 bp of flanking genomic DNA. An adaptor is ligated to the digested genomic DNA, and the genomic DNA flanking the cassettes is amplified by PCR and sequenced. The sequences are then used to map the genomic locations of the cassettes in mutants.

We Validated ChlaMmeSeq by Characterizing Insertion Sites in Individual Mutants

To evaluate the mapping accuracy of the tool, we analyzed 15 randomly picked and individually isolated mutants. DNA gel blot analyses with the *AphVIII* probe and two different restriction enzymes (Supplemental Figure 3) indicated that 14/15 mutants had one *AphVIII* copy per genome, and one mutant had two *AphVIII* copies.

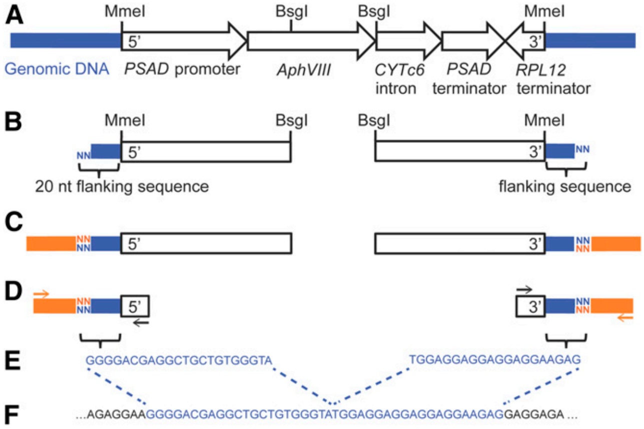


Figure 1. ChlaMmeSeq Is an *Mme*I-Based Strategy for Mapping Insertion Sites.

(A) The cassette used to transform *C. reinhardtii* cells is 2660 bp long and is composed of a *PSAD* promoter, *AphVIII* gene (conferring paromomycin resistance), *CYTc6* intron, and *PSAD* and *RPL12* terminators in opposite orientations. Restriction enzyme sites (*Mme*I and *Bsg*I) are shown. Blue lines represent genomic DNA.

(B) Double digestion of mutant genomic DNA with *Mme*I and *Bsg*I yields a 1121-bp fragment from each side of the cassette, containing 20 to 21 bp of flanking genomic DNA with a two-nucleotide overhang.

(C) An adaptor (orange) that contains both PCR and sequencing primer binding sites is ligated to the digestion products.

(D) The flanking DNA sequences are amplified with PCR primers (black and orange arrows) binding to the cassette and adaptors.

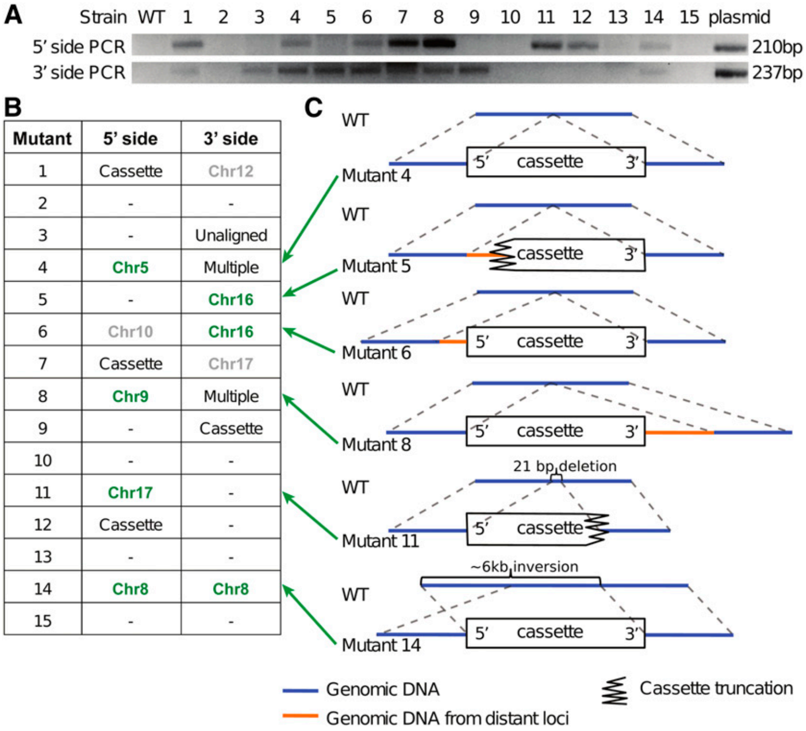
(E) The resulting PCR product is sequenced by Sanger or Illumina sequencing (for individual or pooled mutants, respectively).

(F) The 20- to 21-bp flanking DNA sequences are mapped to the *C. reinhardtii* genome to identify the site of each insertion.

We applied ChlaMmeSeq to extract flanking sequences from the 5′ and 3′ sides of the cassette in each of these 15 mutants. Eight mutants yielded 5′ side flanking sequences, and nine mutants yielded 3′ side flanking sequences (Figure 2A). These flanking sequences were aligned to the published v5.3 *C. reinhardtii* genome (Merchant et al., 2007; Goodstein et al., 2012). Of these 17 flanking sequences, 10 mapped to unique sites in the genome, two mapped to multiple sites in the genome, four mapped to the cassette sequence, and one could not be aligned (Figure 2B). Guided by the 10 flanking sequences that were mapped to unique sites in the genome, we characterized the structure of insertion sites (Figure 2C; Supplemental Figure 4).

We verified that 7/10 uniquely aligned genomic flanking sequences correctly identified the genomic insertion sites by PCR from the cassette into the flanking genome ~1 kb on either side (Figure 2; Supplemental Figure 4). The remaining three flanking sequences did not correctly identify the insertion loci: The expected genome-cassette junctions could not be amplified; instead, the corresponding wild-type loci could be amplified in the mutants (Supplemental Figures 4C, 4G, and 4H).

Nonextractable or misleading flanking sequences are likely the result of commonly observed events at insertion sites in *C. reinhardtii*. Others have reported that the ends of transformation cassettes are frequently removed during insertion into the genome



Mutant	5' side	3' side
1	Cassette	Chr12
2	-	-
3	-	Unaligned
4	Chr5	Multiple
5	-	Chr16
6	Chr10	Chr16
7	Cassette	Chr17
8	Chr9	Multiple
9	-	Cassette
10	-	-
11	Chr17	-
12	Cassette	-
13	-	-
14	Chr8	Chr8
15	-	-

Figure 2. Genomic Flanking Sequences Were Extracted from 15 Randomly Picked Individual Mutants to Validate ChlaMmeSeq.

(A) Flanking sequences were PCR amplified from 5′ and 3′ sides in 15 mutants. The wild type was included as a negative control; plasmid pMJ013b (containing the transforming cassette) served as a positive control.

(B) The mapping location of each extracted flanking sequence is given: a chromosome number if the flanking sequence was mapped uniquely to the genome, “multiple” if it was mapped to multiple genome locations, “cassette” if it was mapped to the transforming cassette, and “unaligned” if it had no alignment. A dash indicates no extractable flanking sequences. Chromosome locations confirmed by PCR are highlighted in green; those determined to inaccurately represent the insertion sites are in gray.

(C) The gene models for each mutant (bottom) and its corresponding wild type (top) are shown for the six insertion sites confirmed by PCR. Supplemental Figure 4 provides details and supporting evidence.

(Dent et al., 2005; González-Ballester et al., 2011). Consistent with this, we observed that 2/6 characterized insertion sites carried truncated cassettes (Figure 2C). Cassette truncation removes the *Mme*I binding site and makes flanking sequence extraction impossible. Others have also observed fragments of genomic DNA from distant loci inserted next to mutagenic cassettes (Meslet-Cladière and Vallon, 2012). We observed such fragments in 3/6 characterized insertion sites. Only one of these fragments caused a misleading mapping, as the other two did not yield uniquely mapping flanking sequences.

Four out of the six mutants characterized had no genomic deletions or rearrangements at the site of insertion (Figure 2C). One mutant had a 21-bp deletion (Supplemental Figure 4E). Additionally, we observed an apparent inversion of ~6 kb of genomic DNA flanking the 5′ side of the cassette in one mutant (Supplemental Figure 4F).

In summary, we expect that ~70% of the flanking sequences obtained from this collection of mutants correctly indicate the insertion sites.

We Simultaneously Mapped 11,478 Insertion Sites in a Pool of Mutants

We pooled ~40,000 mutants and extracted their 5′ flanking sequences using ChlaMmeSeq. Sequencing of the resulting sample on an Illumina Genome Analyzer IIx yielded 47 million reads total, of which 37 million contained the expected adapter and cassette sequences (the remaining 10 million reads were most likely due to nonspecific PCR amplification or sequencing errors). Similarly to our analysis of individual mutants, 9% of reads could not be mapped, 24% aligned to the cassette, 17% aligned to multiple genomic positions, and 50% aligned uniquely to a single genomic location (the ratio of multiple to unique genomic alignments is consistent with the ratio of nonunique to unique 20-bp sequences in the genome). Of these uniquely aligned reads, 96% aligned to the nuclear genome, while 3.9% aligned to the chloroplast and 0.44% to the mitochondrial genome. We have not further investigated insertions in the latter two groups because of their rarity. We identified 11,478 insertion sites distributed throughout the nuclear genome.

The Global Distribution of Insertion Positions Is Largely Indistinguishable from Random

The unprecedented scale of our insertion site data allowed us to evaluate the distribution of insertion sites in the genomes of *C. reinhardtii* mutants generated by electroporation-based transformation. We looked for regions with densities of observed insertion sites higher (hot spots) or lower (cold spots) than would be expected if the insertion sites were randomly distributed over the genome.

We compared our observed distribution of insertion sites to simulations of random insertion of cassettes into the nuclear genome. Our simulations only used insertions with uniquely mappable flanking sequences (i.e., ones that only align to a single site in the genome). This requirement made the simulation results comparable to the real insertion data, which were generated using only uniquely mappable flanking sequences. By visual inspection, it appears that most extended regions with few insertions were replicated in the simulated data, indicating that they correspond to regions of the genome with low density of uniquely mappable insertion sites, rather than being genuine cold spots (Figure 3A). To identify statistically significant insertion cold spots and hot spots, we scanned the genome with moving windows of size ranging from 1 kb to

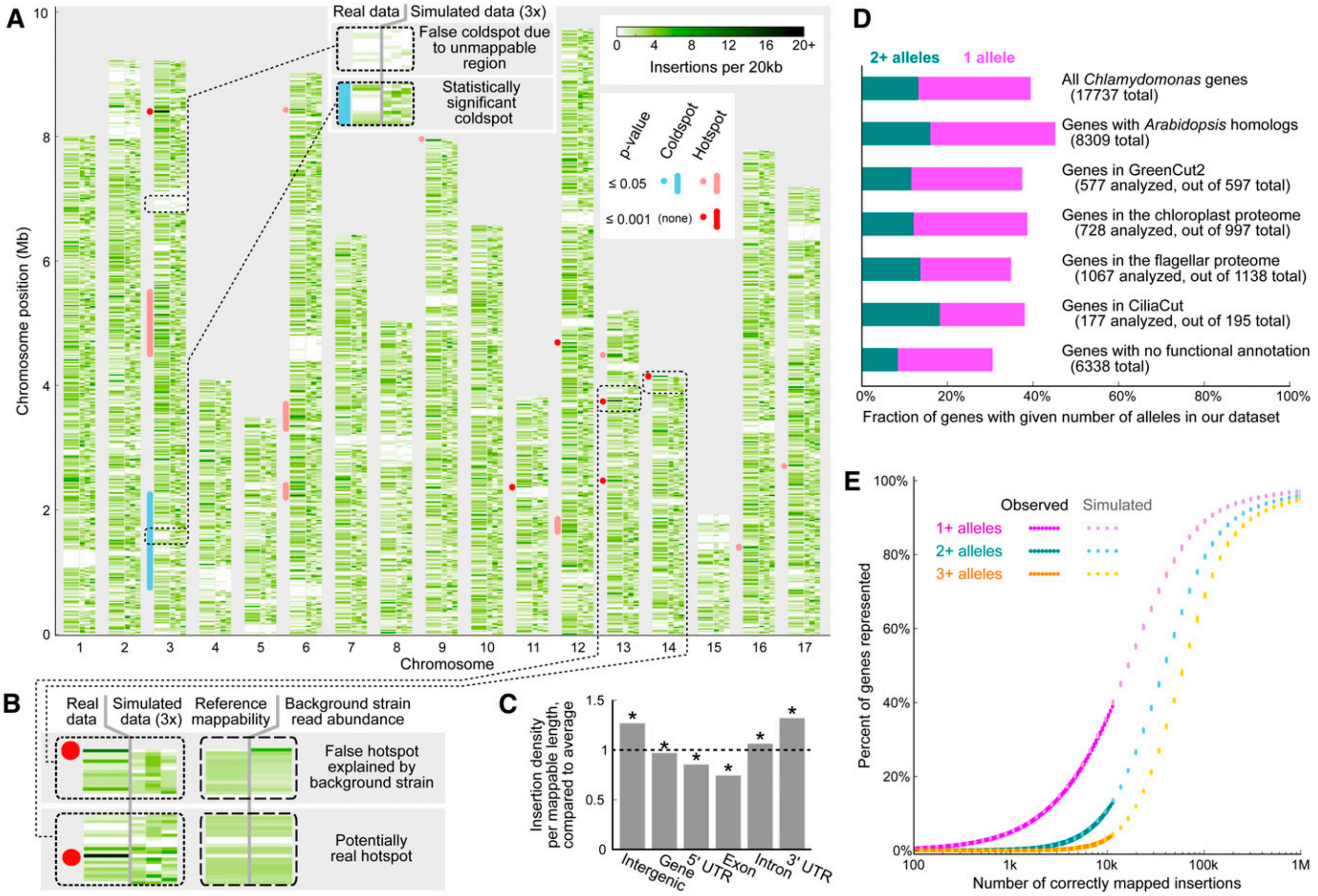


Figure 3. The Genomic Distribution of Insertions Is Largely Random, and Many Genes of Interest Are Represented.

(A) For each chromosome, four columns are shown: the first, wider green column depicts the observed insertion density; the next three columns show insertion densities for three simulated data sets. The blue and red marks on the left of each chromosome indicate statistically significant hot spot and cold spot locations.

(B) The dotted rectangles contain zoomed-in portions of two significant hot spots in **(A)**. The dashed rectangles show additional data for the same two genomic regions: the density of all possible uniquely mappable positions based on the reference strain genome sequence, and the density of observed uniquely mappable 20- to 21-bp sequences from whole-genome sequencing of our background strain aligned to the reference strain sequence (Supplemental Figure 5 contains the same plot for the full genome).

(C) Insertion density differs between gene features and intergenic regions. Density is normalized to the density over the entire genome (dotted line) and is based on uniquely mappable positions only. All categories are different from the overall mutant density (P values < 0.003, exact binomial test); all category pairs except intergenic versus 3′ UTR are different from each other (P values < 0.02, χ^2 test of independence). Genes with multiple splice variants are ignored when looking at gene features.

(D) The fraction of genes with one allele or two or more alleles in our data set is shown for each of several data sets of interest. For some of the data sets, the Joint Genome Initiative protein IDs for our insertions had to be determined. The data could not be obtained for some of the genes, which are omitted from the figure (see Supplemental Methods for details.)

(E) The fraction of genes with 1+, 2+, and 3+ independent mutant alleles is shown as a function of the number of mapped insertions. The observed data (with 100 randomly chosen subsets for lower insertion numbers) and data from 10 simulations are plotted.

1 Mb and for each region compared the number of observed insertion sites to the number of uniquely mappable positions. This analysis yielded only one potential cold spot and 15 potential hot spots with P values < 0.05 after adjustment for multiple testing (Figure 3A; Supplemental Data Set 4). Overall, ~2% of all insertions are in hot spots; cold spots cover <2% of the genome. We conclude that the distribution of insertion locations in the genome is largely indistinguishable from random, on the scale observed in this study.

To investigate whether any of the potential hot spots were due to amplifications of genomic regions in our background strain in comparison to the reference genome, we sheared the genome of our background strain and sequenced the resulting fragments using Illumina sequencing. We mapped ~38 million 21-bp reads to the reference genome, using the same method as for mapping insertion flanking regions (Supplemental Figure 5). Fewer than 1% of 20-kb regions had >2× the median read density normalized to the number of uniquely mappable positions, suggesting that the technique yielded even coverage of the genome. Strikingly, we observed that 4 of the 15 potential insertion hot spots, including the most prominent one on chromosome 3, correspond precisely to regions of high read counts in the background genome sequencing data (Figure 3B; Supplemental Figure 5). This indicates that those four apparent hot spots are artifacts due either to local amplification of the genome sequence in our background strain (in comparison to the reference genome) or to possible inaccuracies in the reference genome assembly.

On a finer scale, we found that the density of insertions is higher in intergenic regions, introns, and 3′ untranslated regions (UTRs) and lower in genes, 5′ UTRs, and exons (Figure 3C). This could be due to an increased likelihood of lethality if the cassette inserts into the latter elements. Gene essentiality appears to decrease the likelihood of recovering a mutant: The 711 best BLAST hits of yeast essential genes had fewer insertions per mappable length than remaining genes ($P < 10^{-5}$, χ^2 test of independence; Supplemental Methods). We did not detect a significant difference between insertions based on position in gene (Supplemental Figure 6) or between sense and antisense insertions into genes. As expected, on average there are more insertions into longer genes (Supplemental Figure 7).

Many Genes of Interest Are Represented

Of the 11,478 insertions mapped uniquely to the nuclear genome, 10,391 (90%) are in genes. Of the insertions in genes, 3032 (29%) are in exons, 3826 (37%) in introns, 446 (4.3%) in 5′ UTRs, 1939 (19%) in 3′ UTRs, 31 (0.3%) at feature boundaries, 1104 (11%) in genes with multiple splice variants, and 13 (0.12%) in multiple features due to overlapping genes. A total of 6955 genes were represented by at least one insertion (39% of the 17,737 protein-coding genes annotated in the v5.3 *C. reinhardtii* nuclear genome; Goodstein et al., 2012). The insertion sites included 3735 (45%) genes highly conserved between *C. reinhardtii* and *Arabidopsis thaliana* (Merchant et al., 2007), 215 (37%) of the most highly conserved plant genes as defined in the GreenCut2 (Karpowicz et al., 2011), and 370 (35%) genes encoding components of cilia identified by mass spectrometry (Pazour et al., 2005) (Figure 3D).

~100,000 Correctly Mapped Mutants Will Be Needed to Cover 80% of the Genome with Two Alleles

The agreement of the observed data with our random insertion model allowed us to use the model to predict the number of *C. reinhardtii* mutants needed to achieve any desired coverage of the genome with 1+, 2+, or 3+ mutant alleles (Figure 3E). The model predicts that in a collection of ~100,000 correctly mapped mutants, ~90% of genes will be represented by 1+ allele, ~80% by 2+ alleles, and ~70% by 3+ alleles. In a screen, multiple mutant alleles in a gene are an advantage because they increase the confidence in a genotype-phenotype link.

The Flanking Sequence Abundances Are Quantitatively Reproducible

To evaluate the quantitative reproducibility of ChlaMmeSeq, we compared the flanking sequence abundances obtained from two parallel applications of the protocol on the same pool of mutants (Figure 4). Ninety-nine percent of flanking sequences with ≥200 reads in one replicate were also observed in the other replicate. The abundance of 92% of those flanking sequences was reproduced within 3× in the other replicate. This suggests that ChlaMmeSeq could be used to quantitatively track changes in abundances of mutants in a pool or for quantitative enrichment screens of mutants of interest.

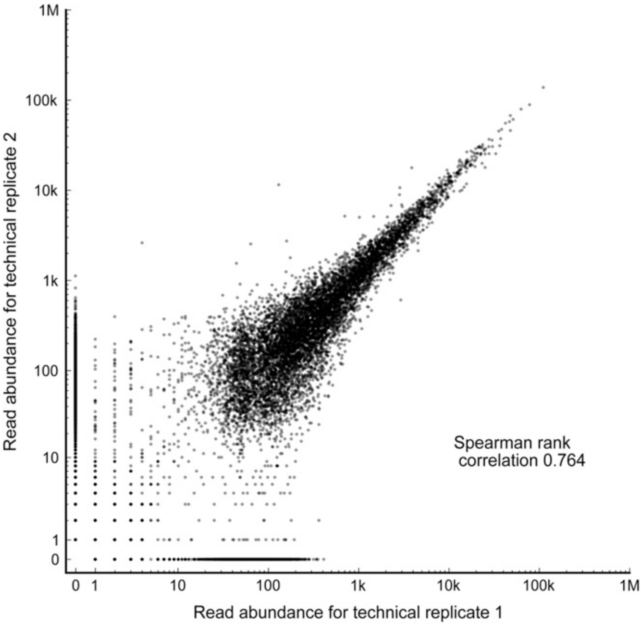


Figure 4. The Flanking Sequence Abundances Are Quantitatively Reproducible.

Reproducibility of mutant abundances between technical replicates is shown. Each dot is a mutant; the two axes are read abundance in the two replicates. The dots are 50% transparent.

Site-Specific Endonucleases Cleave the Cassette during Transformation

In previous reports (Dent et al., 2005; Aksoy et al., 2013) and in our small-scale mutant analyses (Figure 2B), fragments of the transformation cassette were found adjacent to intact cassettes in some of the mutants; however, the origin of these fragments has remained unknown. We reasoned that analysis of the large number of flanking sequences mapping to the cassette in our high-throughput data could yield insights into the mechanism of cassette fragmentation. Such insights could enable strategies to reduce fragmentation, which could increase the fraction of insertions that yield genomic flanking sequences with ChlaMmeSeq and more generally increase transformation efficiency in *C. reinhardtii*.

The transforming DNA appeared as a sharp band when evaluated by Bioanalyzer (Supplemental Figure 8), suggesting that any cassette fragments are likely generated during transformation.

We sought to determine whether some cassette positions were subject to fragmentation more often than others. We used ChlaMmeSeq on a different mutant pool with more uniform

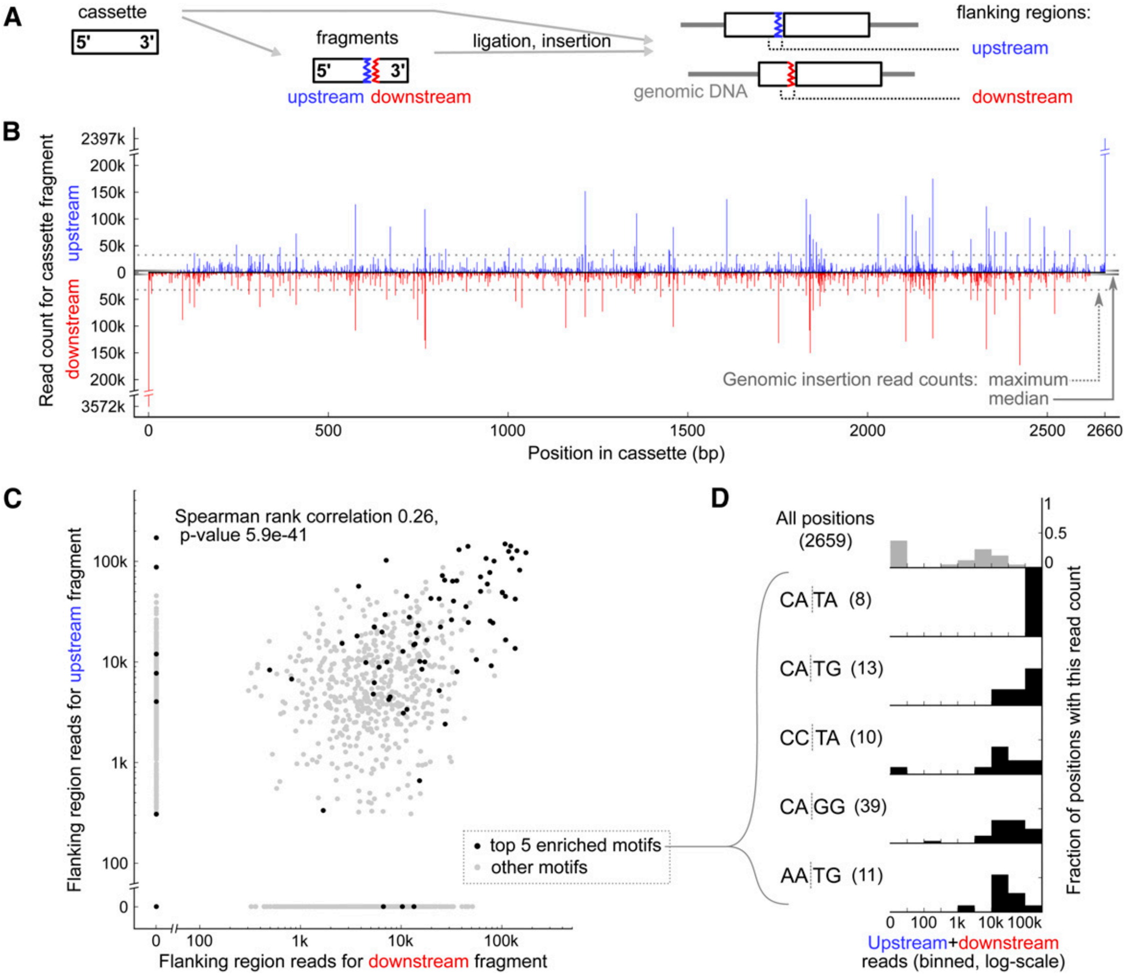


Figure 5. Site-Specific Endonucleolytic Activities Act on the Cassette during Transformation.

(A) In our model, the cassette is fragmented during transformation. A resulting fragment can be ligated to an intact cassette and inserted into the genome along with it; the end of the fragment will then be sequenced as a flanking region. We categorize cassette fragments as “upstream” (from the 5′ side of the fragmentation site) and “downstream” (from the 3′ side). Each type of fragment can be ligated to either a 5′ or a 3′ end of an intact cassette; we do not distinguish between those cases.

(B) For each position along cassette length, the number of reads mapped to that position is plotted for upstream and downstream fragments. For comparison, we show the median and maximum read count for genomic insertions from the same mutant pool (Supplemental Figure 9 contains a full comparison of cassette and genomic read count distributions).

(C) Each dot represents one cassette position; the two axes show the read counts of upstream and downstream fragments mapping to that position. Positions flanked by the five most enriched 4-bp motifs are black; all other positions are gray.

(D) Normalized binned histograms of the sum of upstream and downstream read counts are shown for all cassette positions and for positions flanked by the five most enriched fragmentation site motifs. The total number of positions matching each motif is given in parentheses.

mutant abundance, to detect cassette fragments that were present in multiple mutants. A total of 105 flanking sequences mapping to the cassette were found to be more abundant than the most abundant flanking sequence mapping to the genome (Figures 5A and 5B; Supplemental Figure 9). This strongly suggests that each of these 105 flanking sequences originated from more than one mutant. Flanking sequences mapping to the cassette 5′ and 3′ termini were by far the most abundant (16 and 11% of all cassette-mapped flanking sequences, respectively).

A single cassette fragmentation event can yield two fragments, each containing one side of the fragmentation site; we refer to them as the upstream and downstream fragments (Figure 5A). We observed a correlation between the abundance of flanking sequences originating from the upstream and downstream cassette fragments for a given fragmentation position (Figure 5C). This result suggests that both upstream and downstream fragments from frequent fragmentation sites are present at similar abundance after fragmentation. This finding indicates that the frequent fragmentation sites are not due to exonuclease activity, which one would expect to produce an uneven ratio of upstream and downstream fragments.

We observed an enrichment of specific sequence motifs flanking the fragmentation site. Fragmentation positions flanked by the motifs CA|TA and CA|TG and several one-base variants yielded significantly more flanking sequence reads than other positions (Figures 5C and 5D; Supplemental Data Set 5). The five most enriched motifs (false discovery rate–adjusted P values < 0.00001, Kruskal-Wallis test) are responsible for 15% of the reads, and all 26 statistically significantly enriched motifs (P values < 0.05) are responsible for 61% of the reads. The two independent data sets show similar cassette cleavage motifs, underscoring the reproducibility of our findings (Supplemental Figures 10A and 10B). The fact that the motifs span the fragmentation site suggests again that most fragmentation sites are not due to exonuclease activity. Instead, this result is consistent with a model in which the cassette is cleaved by site-specific endonucleolytic activities during transformation, prior to ligation into the genome.

It is worth noting that the flanking sequences mapping to the genome also show enrichment for a CA|TG insertion site motif (Supplemental Figure 10C). This could be due to action of the same endonucleolytic activities on genomic DNA from lysed cells present in the culture medium before or during electroporation. This genomic DNA would be electroporated into cells along with the transformation cassette and could yield some of the observed fragments of exogenous genomic DNA inserted next to some cassettes (Figures 2C and 6).

We Propose a Model for Events during Electroporation of *C. reinhardtii*

Our observation of sequence-specific fragmentation of the transformation cassette is consistent with the following model for events during transformation of *C. reinhardtii* by electroporation (Figure 6): (1) During transformation, the cassette and genomic DNA from lysed cells are partially digested by sequence-specific endonucleases, before or during entry of DNA into recipient cells. (2) The DNA that entered cells is ligated into double-stranded breaks in the recipient genomes. Sometimes, multiple DNA fragments are ligated together into one insertion site.

This model explains the following observations: (1) the presence of flanking regions mapping to cassette sequence and the pattern of cassette fragmentation that we observed from them; (2) insertions of DNA from other genomic loci at cassette insertion sites observed by us and others (González-Ballester et al., 2011); (3) the fact that our mutants with misleading flanking sequences had an intact copy of the genomic region corresponding to these flanking sequences; (4) improvement of transformation efficiency by adding carrier DNA (Shimogawara et al., 1998), since carrier DNA could reduce inactivation of the transforming cassette by endonucleolytic digestion (we do not recommend using carrier DNA, as it could cause additional insertions).

This model suggests possible avenues for improving transformation in *C. reinhardtii*: (1) Removal of endonuclease recognition sequences from the transforming DNA may reduce susceptibility to the endonucleases. (2) Smaller cassettes may be less susceptible to endonuclease activities. (3) Means of inactivating the endonucleases, e.g., mutation or inhibitors, would improve transformation efficiency. (4) Improved washing of cells before electroporation to remove DNA from lysed cells may reduce insertion of exogenous genomic DNA next to the cassette.

DISCUSSION

We demonstrated that ChlaMmeSeq enables mapping of insertion sites for tens of thousands of *C. reinhardtii* mutants. Insertions are randomly distributed, with remarkably few hot spots and cold

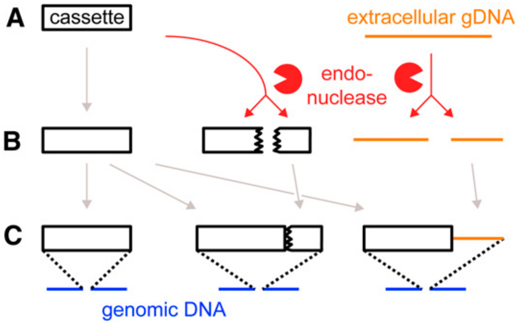


Figure 6. Our Data Are Consistent with a Model Where an Endonuclease Cleaves the Transforming DNA, Which Is Ligated into the Genome at the Site of a Double-Stranded Break.

(A) Before transformation, the extracellular medium contains the transformation cassette and genomic DNA (gDNA) from lysed cells.

(B) During transformation, some of the cassette and genomic DNA molecules are cleaved by site-specific endonucleases.

(C) After electroporation, DNA from the extracellular medium (including intact cassettes, fragmented cassettes, and fragmented genomic DNA) is ligated into a double-stranded break in the genome. In some cases, multiple fragments are ligated into a single site in the genome. When an intact cassette end is present, this cassette end can yield flanking sequences containing the neighboring DNA. In some cases, this flanking DNA will contain cassette DNA or extracellular genomic DNA that was inserted following endonuclease cleavage; in those cases, the end of the DNA neighboring the cassette would be enriched for the endonuclease's sequence preference.

spots. The high-throughput nature of our approach facilitates identification of multiple independent alleles of a gene, which alleviates commonly encountered challenges caused by incorrect flanking sequences and second-site mutations.

The reproducible quantification of flanking sequence abundances could enable the use of ChlaMmeSeq for simultaneous measurements of fitness of thousands of mutants in pooled culture under conditions of interest (van Opijnen et al., 2009) or quantitative enrichment screens (Carette et al., 2011). Our data suggest that this approach could detect a 15% growth defect after seven generations with $P < 0.08$.

We are presently using ChlaMmeSeq to generate an indexed genome-wide collection of *C. reinhardtii* mutants with known insertion sites, by means of combinatorial pooling (Goodman et al., 2009). Based on the analysis presented here, we expect that 100,000 correctly mapped mutants will cover 80% of the *C. reinhardtii* genes with at least two independent mutants per gene. Such a permanent collection could transform plant biology by increasing the phenotyping throughput of mutants in nearly all genes in *C. reinhardtii*.

More immediately, ChlaMmeSeq enables genotyping of pools of hits from *C. reinhardtii* screens on an unprecedented scale. This opens the door to comprehensive identification of genes with roles in regulation and biogenesis of the photosynthetic apparatus, algal lipid metabolism, the algal carbon concentrating mechanism, phototaxis, and other processes for which *C. reinhardtii* is a leading model system. Now that mapping flanking sequences is no longer a rate-limiting step, saturating identification of these important cellular components is within reach.

METHODS

Growth of *C. reinhardtii* Strains

Chlamydomonas reinhardtii strains were grown in Tris acetate phosphate (TAP) medium with modified trace element solution (Kropat et al., 2011) (pH 7.5). For experiments with Tris phosphate medium, the acetate was omitted and the media was buffered to pH 7.5 using HCl stock. For growth as colonies on plates, the medium was supplemented with 1.5% (w/v) agar. Unless otherwise indicated, cool white fluorescent lights (F34CW/RS/WM/ECO; Ecolux) were used for illumination.

Isolation of Background Strain CMJ030

C. reinhardtii strains D66$^+$ and 4A$^-$ were crossed (Jiang and Stern, 2009). Progeny were screened for photoautotrophic and heterotrophic growth, greening in the dark, normal swimming ability, high transformation efficiency (Shimogawara et al., 1998), and efficient recovery from cryogenic storage in liquid nitrogen. One progeny strain that showed all of these properties was isolated and named CMJ030. The cell wall phenotype of CMJ030 was evaluated by treatment with 1% (w/v) Nonidet P-40 and 1% (w/v) SDS (separately), and in both cases the cells swelled and exploded, leaving behind no visible sacks, suggesting that the cells are at least *cw15* wall-less. The mating type of CMJ030 was determined by PCR (Werner and Mergenhagen, 1998). CMJ030 appeared to mate efficiently.

Large-Scale Transformation

The transformation cassette was prepared by digestion of plasmid pMJ013b (GenBank accession number KJ572788) with *Mly*I (New England Biolabs) at 37°C for 1 h. Digestion products were separated on a 1% (w/v) agarose gel, and the 2660-bp digested DNA fragment was extracted using a QIAquick gel purification kit (Qiagen), according to the manufacturer's instructions. The purified fragment was analyzed with an Agilent Bioanalyzer (Supplemental Figure 8) and used for transformation.

CMJ030 was grown in TAP medium in a 20-liter container under 100 μmol photons m^{-2} s^{-1} cool white fluorescent light, with continuous stirring and bubbled air, until it reached a cell density of 1.5×10^6 cells/mL. Cells were collected as follows. Bubbling was stopped and the 20-liter container was transferred into a 32-gallon garbage bin and illuminated from the top by four cool white fluorescent bulbs for 2 h. This caused the cells to settle to the bottom of the 20-liter container. The top 15 liters were removed by aspiration, and the lower 5 liters were centrifuged in RC5C centrifuges (Sorvall Instruments) with GS3 rotors for 4 min at 1000*g*. Pellets were resuspended in TAP supplemented with 40 mM sucrose at 2×10^8 cells/mL. Transformation was performed by electroporation according to Shimogawara et al. (1998) with some modifications. Transforming DNA (144 μL) at 52 ng/μL was added to a sterile 50-mL Falcon tube with 50 mL of concentrated cells (37.5 ng DNA per 250 μL concentrated cells) in 40 mM sucrose. The concentrated cells were incubated with transforming DNA at 16°C for at least 20 min before electroporation. The cell/DNA mix was then aliquoted into sterile electroporation cuvettes (4-mm gap, 1.5-mL Micro Cuvette, two Clear Sides, E&K Scientific) at 250 μL/cuvette. Cells were electroporated (Bio-Rad; Gene Pulser2 electroporation system) with pulse settings of 800 V and 25 μF, followed by immediate decanting into a 15-mL Falcon tube containing 13 mL of TAP supplemented with 40 mM sucrose. The 15-mL Falcon tubes were shaken gently under low light (5 μmol photons m^{-2} s^{-1}) for 6 h. Cells were then collected by centrifugation at 1000*g* for 4 min, most of the supernatant was decanted, and the cells were resuspended in the remaining 500 μL of supernatant. Resuspended cells were gently plated onto 2% (w/v) TAP agar plates containing 20 μg/mL paromomycin. These plates were stored at 5 μmol photons m^{-2} s^{-1} light for 2 weeks, until transformant colonies appeared.

Flanking Sequence Extraction from Pooled Mutants

Our protocol for flanking sequence extraction from pooled *C. reinhardtii* mutants was built upon technologies that were previously demonstrated in bacteria (Goodman et al., 2009; van Opijnen et al., 2009) with modifications to overcome the following challenges: (1) The bacterial genomes (6.3 and 2 Mb, respectively) are smaller than the 121 Mb *C. reinhardtii* genome; and (2) both previous methods used in vitro transposon mutagenesis of genomic DNA, followed by homologous recombination of the mutagenized DNA into the recipient genomes, whereas our *C. reinhardtii* mutants were generated by random insertion of linear transforming DNA (likely by nonhomologous end joining).

Our protocol is most similar to that of Goodman et al. (2009) with the following major changes: (1) We used phenol/chloroform to extract DNA, whereas they used DNeasy columns. (2) We performed digestions with both *Mme*I and *Bsg*I to generate the same size fragments from both full and truncated cassettes, whereas they only did *Mme*I digestion. (3) Our PCR protocol was optimized for GC-rich DNA templates of *C. reinhardtii*. (4) Placement of the *Mme*I sequence at the very ends of the cassette allowed us to extract 20/21 bp of flanking sequences to map insertion sites, whereas their sites were recessed and yielded only 16/17 bp (which is adequate for small genomes but insufficient for the *C. reinhardtii* genome). (5) Each step of our protocol was optimized several times to increase the quantitative character of the tool. Below is the detailed protocol.

Transformants were scraped from 80 transformation plates when colonies were ~1 mm in diameter, pooled together, and grown in TAP in the dark for 1 week in a 1-liter photobioreactor from Photon Systems Instruments at a constant cell density of 1×10^6 cells/mL, with constant bubbling with air. Samples of 50 mL were harvested by centrifugation at 3000*g* for 10 min. The pellet was used for extraction of genomic DNA by phenol:chloroform:isoamyl alcohol (Phenol:CIA, 25:24:1; Sigma-Aldrich).

Genomic DNA was digested as follows. The 500-μL reactions were assembled with 3.5 μg DNA, 50 μL 10× NEB4 buffer, 1.25 μL 32 mM *S*-adenosyl methionine, 5 μL 10 mg/mL BSA, 40 μL 2 units/μL *Mme*I, and 1.25 μL 5 units/μL *Bsg*I (NEB). Reactions were incubated at 37°C for 1 h. Digestion products were phenol/chloroform-extracted, ethanol-precipitated, and dissolved in water. Double digestion of genomic DNA by *Mme*I and *Bsg*I yielded 1100-bp DNA fragments containing either the 5′ or 3′ end of the cassette, and 20 to 21 bp of flanking genomic DNA. The digestion products were run on a 1% (w/v) agarose gel in an Owl D3 tray (BioExpress) at 100 V for 4 h, and DNA fragments in the range of 1 to 1.2 kb were cut out and gel extracted by D-Tube Dialyzer Maxi (MWCO 3.5 kD; EMD Biosciences). DNA was precipitated at –20°C for at least 30 min with 9 μL 20 mg/mL glycogen, 210 μL 3 M NaOAC at pH 5.2 (0.1 volume), and 2100 μL isopropanol, and then centrifuged at 4°C at 16,000*g* for 20 min, followed by a wash with 1 mL 70% (v/v) ethanol, and then by a wash with 1 mL of 100% ethanol. The DNA pellet was dissolved in water and quantified by Qubit.

Adaptors (50 μM) were prepared by mixing equal volumes of oMJ082 and oMJ083 at 100 μM each in water, placing the mixture in a heat block (E&K Scientific D-1200 AccuBlock Digital Dry Bath) at 96°C for 2 min, then placing the metal insert of the heat block containing the samples on the bench at room temperature and letting it cool for 1 h.

Ligations were performed as follows. Thirty-microliter ligation reactions contained 16 μL DNA template (7 ng/μL) from the previous step, 3 μL T4 DNA ligase buffer, 10 μL 50 μM adaptors, and 1 μL 2000 units/μL T4 DNA ligase. The reactions were incubated at 16°C overnight. The high concentration of adaptors increased ligation efficiency, but also interfered with PCR. As a result, a second gel extraction using a D-Tube Dialyzer Maxi was performed to remove extra adaptors.

PCR was performed as follows. Ninety percent of the DNA from the previous step was used as template for each sample. The DNA template was diluted to 0.1 to 0.3 ng/μL and heated at 65°C for 20 min before PCR. We found that this step helped reduce nonspecific amplification. For each sample, 48 PCR reactions of 50 μL each were assembled. Each reaction contained 10 μL 5× Phusion GC buffer, 1 μL 10 mM deoxynucleotide triphosphates, 0.5 μL 100% DMSO, 0.2 μL 50 mM $MgCl_2$, 0.25 μL of each primer (oMJ079, oMJ081, and oMJ191) at 100 μM, 10 μL DNA template at 0.1 to 0.3 ng/μL, 25.55 μL water, and 2 μL 2 units/μL Phusion Hot Start II High-Fidelity DNA Polymerase (New England Biolabs). Cycling parameters were as follows: 3 min at 98°C, 10 cycles of 10 s at 98°C, 25 s at 61°C, 15 s at 72°C; then 14 cycles of 10 s at 98°C, 40 s at 72°C; then a final extension of 2 min at 72°C. Primer oMJ079 binds to the adaptor, oMJ081 binds the 5′ end of the cassette to amplify the 5′ side flanking sequence, and oMJ191 binds to the 3′ end of the cassette to amplify the 3′ side flanking sequence. The products were run on a 1.8% (w/v) agarose gel at 100 V for 2 h. PCR products of the expected size (190 bp for flanking sequences from both sides) were gel extracted with QIAquick and submitted for deep sequencing by Illumina Genome Analyzer IIx at the Stanford Sequencing Service Center.

Data in Figure 5 are from a second collection of mutants generated by a similar protocol, with the following differences, which were applied to reduce the variation in mutant abundance in the pool. Transformants were picked individually from transformation plates using a CP7200 colony picking robot (Norgren Systems), arrayed at 384 colonies/plate on fresh TAP plates containing 20 μg/mL paromomycin, and grown up under 5 μmol photons m^{-2} s^{-1} light for 3 weeks. The plates were then propagated by Singer Rotor HDA Robot to fresh TAP plates containing 20 μg/mL paromomycin. After growth under 5 μmol photons m^{-2} s^{-1} light for 1 week, most mutants on these 384 plates had approximately similar colony sizes. Mutants were scraped from plates and pooled together without further growth, and flanking sequences were extracted using the protocol above for pooled mutants.

PCR-Based Characterization of the Insertion Sites Indicated by the Flanking Genomic DNA

PCRs were attempted only for the insertions that yielded flanking sequences that were mapped uniquely (10 out 17 total flanking sequences from 15 individual mutants). Primers were designed to amplify the cassette-genome junction on both sides of each insertion site indicated by the 20- to 21-bp flanking sequences, based on the genome sequence v5.3 from Phytozome (Goodstein et al., 2012). Primer binding sites were 1 to 1.3 kb away from the flanking sequences. PCR products were amplified using the Taq PCR core kit (Qiagen). The 25-μL PCR reactions included: 5 μL 5× Q-solution, 2.5 μL 10× PCR buffer, 1.25 μL 100% DMSO, 0.5 μL 10 μM deoxynucleotide triphosphates, 12.65 μL water, 0.1 μL Taq DNA polymerase, 1.25 μL of each primer at 10 μM, and 0.5 μL 50 ng/μL *C. reinhardtii* genomic DNA. PCR cycling parameters were: 5 min at 95°C, 40 cycles of 30 s at 95°C, 45 s at 58°C, 2 min at 72°C, followed by a final extension of 10 min at 72°C. For oligos and template used for each check PCR, see Supplemental Data Set 2. PCR products of the expected size were gel extracted and submitted for Sanger sequencing by ELIM Biopharmaceuticals. The resulting sequences were aligned against the v5.3 genome sequence from Phytozome.

Code Availability

All custom software written for this project is available at github.com/Jonikas-Lab/Zhang-Patena-2014.git and as Supplemental Data Set 6.

Inferring Insertion Details from Pooled Mutant Deep-Sequencing Data

The workflow is summarized in Supplemental Figure 11. Adaptor and cassette sequences were removed from reads to obtain the flanking sequences; reads without the expected adaptor or cassette sequences were discarded. Flanking sequences were aligned to the *C. reinhardtii* nuclear, chloroplast, and mitochondrial genomes and to the insertion cassette sequence using bowtie (Langmead et al., 2009), allowing at most one mismatch. Flanking sequences with no alignments or with multiple genomic alignments were discarded. Flanking sequences aligned to the same position in the same orientation were combined into single insertion positions and annotated with gene and feature information based on *C. reinhardtii* v5.3 genome data from Phytozome.

Insertion positions with very low abundance were removed. We believe that these low abundance reads are likely the result of sequencing or PCR errors. The genomic read count histograms in Supplemental Figures 12A and 12B show a tall peak at around one read per insertion and then a second peak at around 50 reads per insertion. We suspect that the second peak is the number of reads resulting from a single DNA molecule as a PCR template, so any “insertions” below that peak are likely to be due to PCR and/or sequencing errors. Therefore, we removed any insertions below a custom cutoff positioned between the first two peaks (15 reads for replicate 1 and 20 reads for replicate 2, marked on Supplemental Figures 12A and 12B). The data used in Figure 5 was processed in a similar way (the cutoff is marked on Supplemental Figures 12C and 12D).

The two technical replicate data sets were merged together at this point. The single replicates were used to show replicate correlation (Figure 4) and reproducibility; the merged data set was used for all other analyses.

Adjacent insertion positions that probably came from a single real insertion were merged. In some cases, a single insertion could generate two flanking sequences that map to different (adjacent) positions. The main scenarios where this could happen are (Supplemental Figure 13): (1) 1-bp insertions/deletions during PCR or sequencing; (2) a single insertion of two cassette copies ligated together, in opposite orientations. We merged such pairs until their level was reduced to that expected randomly. This merging process impacted <100 insertions, but omitting it can cause bias in the distribution of insertion positions and inflate the number of insertions per gene. The merging was not done for the cassette-aligned flanking sequences.

The insertion position data for all the data sets are available as Supplemental Data Sets 7 to 11.

Determining Genome Mappability

To meaningfully compare the observed and expected insertion positions and densities, we determined which of all possible genomic insertion positions would yield sequencing reads uniquely mapped to that position. Each 20- and 21-bp slice of the genome sequence was categorized as unique or nonunique (77% of 20 bp slices and 78% of 21 bp slices in the *C. reinhardtii* genome are unique). The results were summed to yield "mappable lengths" for the whole genome, each gene, and other regions of interest; these are proportional to the expected density of insertions in a region if insertion positions were purely random.

Generating Simulated Random Insertion Data Sets

We generated 10 simulated data sets with the same number of insertions as the real data set, with the location of each insertion randomly chosen out of all the mappable positions in the genome. Furthermore, in order to estimate how much of the genome would be covered by larger numbers of insertions (Figure 3E), we used the same method to generate 10 simulated data sets of one million mappable insertions each.

Locating Statistically Significant Insertion Density Hot Spots and Cold Spots

We used the binomial test with correction for multiple testing to detect regions of the genome with more (hot spots) or fewer (cold spots) insertions than would be expected if the insertion positions were random. We looked for hot spots/cold spots in a large range of sizes: 1 kb, 5 kb, 20 kb, 100 kb, 200 kb, 400 kb, and 1 Mb. We sliced the genome into windows of each size, using evenly spaced offsets to get two to four overlapping sets of windows. For each region, we used the exact binomial test to determine the probability of obtaining the observed number of insertions given that region's mappable length and the total number of observed insertions in the genome, assuming that the insertions were uniformly randomly distributed over the mappable genome positions. The resulting P values were corrected for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) (as implemented in R with the p.adjust function) separately for each region size and offset combination.

A single real hot spot or cold spot is likely to generate multiple statistically significant results in overlapping windows with different sizes or offsets. Therefore, the results were clustered into 15 distinct groups of overlapping hot spots and 1 group of cold spots. For plotting (Figures 3A and 3B; Supplemental Figure 5), each overlapping cluster was reduced to a nonredundant subset of one to three regions with the lowest P values. Supplemental Data Set 4 contains the full data.

We performed the same analysis on the simulated data sets to ensure that the method worked as expected. The 10 simulated data sets had only 0 to 2 distinct hot spot and cold spot groups, with the lowest adjusted P value much higher than those of most of the hot spots detected in the real data set. This shows that a few of the potential hot spots and cold spots in the real data set may be due to noise, but the majority of the hot spots are almost definitely nonrandom.

Background Strain Genome Sequencing and Analysis

Genomic DNA from our background strain was extracted in technical duplicate and then submitted for Illumina library preparation and sequencing. To mimic the process of mapping insertion flanking regions, we extracted the first 20 or 21 bp from one end of each read and aligned it to the reference *C. reinhardtii* genome using bowtie. The differences between the two replicates and the 20/21-bp versions were minimal; they were added together for the final display of sequenced fragment density (Supplemental Figure 5).

Cassette Fragmentation Motif Enrichment

For each possible 4-bp motif, we determined the number of times it occurs in the cassette, and the observed read counts for upstream and downstream fragments for each of the positions in which it occurred. We calculated the average enrichment by dividing the average read count for those fragments by the average read count for all observed cassette fragments. We used the nonparametric Kruskal-Wallis test to determine whether the read counts for each motif originate from a different distribution than all the cassette fragment read counts and corrected the resulting P values for multiple testing using the Benjamini-Hochberg procedure. The full list of motifs, average read counts, P values, and other information are available in Supplemental Data Set 5.

Accession Numbers

Sequence data from this article can be found in the *Arabidopsis* Genome Initiative or GenBank/EMBL databases under accession number KJ572788 for the insertion cassette pMJ013b along with its cloning vector, annotated with gene features, primer binding sites, and relevant restriction sites.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. The Background Strain CMJ030 Can Grow Photoautotrophically, Mixotrophically, and Heterotrophically.

Supplemental Figure 2. CMJ030 Is Mating Type Minus and Recovers Efficiently from Cryogenic Storage in Liquid Nitrogen.

Supplemental Figure 3. Cassette Copy Number Was Determined in 15 Individual Mutants by DNA Gel Blot.

Supplemental Figure 4. PCR Analysis Revealed the Structure of Insertion Sites in Individual Mutants.

Supplemental Figure 5. Some of the Potential Hot Spots Can Be Explained by Local Amplifications of the Genome of the Background Strain.

Supplemental Figure 6. Insertion Density Does Not Vary along the Gene Length.

Supplemental Figure 7. Longer Genes Have More Insertions.

Supplemental Figure 8. The Transforming DNA Is Intact before Transformation.

Supplemental Figure 9. The Cassette Read Count Distribution Has a Longer Tail Than the Genomic Read Count Distribution.

Supplemental Figure 10. Cassette and Genomic Insertion Positions Are Enriched for the CA|TG Motif.

Supplemental Figure 11. Our Computational Pipeline Determines Insertion Sites Based on Deep-Sequencing Data.

Supplemental Figure 12. Read Count Cutoffs Are Set Based on Each Data Set's Read Count Distribution.

Supplemental Figure 13. We Merge Flanking Sequences Mapped to Adjacent Positions If We Suspect That They Originated from a Single Insertion.

Supplemental Table 1. Summary of Phenotypes of the Background Strain CMJ030.

Supplemental Methods.

The following materials have been deposited in the DRYAD repository under accession number http://doi.org/10.5061/dryad.50p47.

Supplemental Data Set 1. Mapped Insertion Sites of the Individual Mutants Presented in Figure 2.

Supplemental Data Set 2. Primer Pairs Used for the Mutant Insertion Site Characterization by PCR Presented in Supplemental Figure 4.

Supplemental Data Set 3. Oligonucleotide Sequences Used in This Manuscript.

Supplemental Data Set 4. Full Fist of Hot Spots and Cold Spots with FDR-Adjusted P Values < 0.05.

Supplemental Data Set 5. List of Cassette Fragmentation Motifs and Enrichment Statistics.

Supplemental Data Set 6. All the Programs Written in Our Lab That Were Used for This Project.

Supplemental Data Set 7. Insertion List for the Data Set Used for Figures 3 and 4, Replicate a, Raw.

Supplemental Data Set 8. Insertion List for the Data Set Used for Figures 3 and 4, Replicate b, Raw.

Supplemental Data Set 9. Insertion List for the Data Set Used for Figures 3 and 4, Both Replicates, Filtered.

Supplemental Data Set 10. Insertion List for the Data Set Used for Figure 5, 5′ Side, Raw.

Supplemental Data Set 11. Insertion List for the Data Set Used for Figure 5, 3′ Side, Raw.

ACKNOWLEDGMENTS

We thank K. Wemmer for help with assessing the swimming phenotypes of strains; P. Lefebvre for assessing the cell wall phenotype of CMJ030; G. Euskirchen and Z. Weng for help with Illumina sequencing; members of the Jonikas lab for helpful discussion; M. Terashima for help with mutant pooling; G. Watrous for help with mapping insertion sites in individual mutants; A. Grossman, W. Yang, M. Heinnickel, and C. Catalanotti (Carnegie Institution for Science) for strains D66$^+$ and 4A$^-$ as well as protocols and advice on DNA extraction and crossing; and W. Frommer, V. Walbot, M. Terashima, L. Mackinder, L. Pallesen, E. Freeman, and X. Li for critical reading of the article. We thank the Barton, Frommer, and Ehrhardt labs for allowing us to use their lab space. This project was funded by the Carnegie Institution for Science, by U.S. Air Force Office of Scientific Research Grant FA9550-11-1-0060, and National Science Foundation Grant EF-1105617 to M.C.J. as well as by Deutsche Forschungsgemeinschaft research fellowship AR 808/1-1 to U.A.

AUTHOR CONTRIBUTIONS

R.Z. and M.C.J. designed the experiments and developed the initial version of the ChlaMmeSeq protocol. R.Z. optimized ChlaMmeSeq. R.Z. and S.S.G. isolated and characterized the CMJ030 strain. U.A., S.S.G., and S.R.B. optimized the transformation protocol. S.R.B. and S.S.G. generated the mutant pools. R.Z. applied ChlaMmeSeq to mutant pools to generate deep sequencing data and characterized individual mutants by PCR to validate ChlaMmeSeq. U.A. performed DNA gel blots, spot tests, and mating-type PCRs. W.P. designed and implemented computational data analyses. R.Z., W.P., and M.C.J. wrote the article.

Received February 7, 2014; revised February 7, 2014; accepted March 18, 2014; published April 4, 2014.

REFERENCES

Aksoy, M., Pootakham, W., Pollock, S.V., Moseley, J.L., González-Ballester, D., and Grossman, A.R. (2013). Tiered regulation of sulfur deprivation responses in *Chlamydomonas reinhardtii* and identification of an associated regulatory factor. Plant Physiol. **162:** 195–211.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B **57:** 289–300.

Boynton, J.E., et al. (1988). Chloroplast transformation in *Chlamydomonas* with high velocity microprojectiles. Science **240:** 1534–1538.

Carette, J.E., Guimaraes, C.P., Wuethrich, I., Blomen, V.A., Varadarajan, M., Sun, C., Bell, G., Yuan, B., Muellner, M.K., Nijman, S.M., Ploegh, H.L., and Brummelkamp, T.R. (2011). Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. Nat. Biotechnol. **29:** 542–546.

Cipriani, P.G., and Piano, F. (2011). RNAi methods and screening: RNAi based high-throughput genetic interaction screening. Methods Cell Biol. **106:** 89–111.

Crutchfield, A., Diller, K., and Brand, J. (1999). Cryopreservation of *Chlamydomonas reinhardtii* (Chlorophyta). Eur. J. Phycol. **34:** 43–52.

Dent, R.M., Haglund, C.M., Chin, B.L., Kobayashi, M.C., and Niyogi, K.K. (2005). Functional genomics of eukaryotic photosynthesis using insertional mutagenesis of *Chlamydomonas reinhardtii.* Plant Physiol. **137:** 545–556.

Fleischmann, M.M., Ravanel, S., Delosme, R., Olive, J., Zito, F., Wollman, F.A., and Rochaix, J.D. (1999). Isolation and characterization of photoautotrophic mutants of *Chlamydomonas reinhardtii* deficient in state transition. J. Biol. Chem. **274:** 30987–30994.

González-Ballester, D., de Montaigu, A., Galván, A., and Fernández, E. (2005a). Restriction enzyme site-directed amplification PCR: a tool to identify regions flanking a marker DNA. Anal. Biochem. **340:** 330–335.

González-Ballester, D., de Montaigu, A., Higuera, J.J., Galván, A., and Fernández, E. (2005b). Functional genomics of the regulation of the nitrate assimilation pathway in *Chlamydomonas.* Plant Physiol. **137:** 522–533.

González-Ballester, D., Pootakham, W., Mus, F., Yang, W., Catalanotti, C., Magneschi, L., de Montaigu, A., Higuera, J.J., Prior, M., Galván, A., Fernández, E., and Grossman, A.R. (2011). Reverse genetics in *Chlamydomonas*: a platform for isolating insertional mutants. Plant Methods **7:** 24–36.

Goodman, A.L., McNulty, N.P., Zhao, Y., Leip, D., Mitra, R.D., Lozupone, C.A., Knight, R., and Gordon, J.I. (2009). Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell Host Microbe **6:** 279–289.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. **40:** D1178–D1186.

Hillenmeyer, M.E., et al. (2008). The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science **320:** 362–365.

Jiang, X., and Stern, D. (2009). Mating and tetrad separation of *Chlamydomonas reinhardtii* for genetic analysis. J. Vis. Exp. **30:** e1274.

Karpowicz, S.J., Prochnik, S.E., Grossman, A.R., and Merchant, S.S. (2011). The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage. J. Biol. Chem. **286:** 21427–21439.

Kindle, K.L., Schnell, R.A., Fernández, E., and Lefebvre, P.A. (1989). Stable nuclear transformation of *Chlamydomonas* using the

Chlamydomonas gene for nitrate reductase. J. Cell Biol. **109:** 2589–2601.

Kropat, J., Hong-Hermesdorf, A., Casero, D., Ent, P., Castruita, M., Pellegrini, M., Merchant, S.S., and Malasarn, D. (2011). A revised mineral nutrient supplement increases biomass and growth rate in *Chlamydomonas reinhardtii.* Plant J. **66:** 770–780.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10:** R25.

Li, X., Moellering, E.R., Liu, B., Johnny, C., Fedewa, M., Sears, B.B., Kuo, M.H., and Benning, C. (2012). A galactoglycerolipid lipase is required for triacylglycerol accumulation and survival following nitrogen deprivation in *Chlamydomonas reinhardtii.* Plant Cell **24:** 4670–4686.

Maul, J.E., Lilly, J.W., Cui, L., dePamphilis, C.W., Miller, W., Harris, E.H., and Stern, D.B. (2002). The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. Plant Cell **14:** 2659–2679.

Merchant, S.S., Kropat, J., Liu, B., Shaw, J., and Warakanont, J. (2012). TAG, you're it! *Chlamydomonas* as a reference organism for understanding algal triacylglycerol accumulation. Curr. Opin. Biotechnol. **23:** 352–363.

Merchant, S.S., et al. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science **318:** 245–250.

Meslet-Cladière, L., and Vallon, O. (2012). A new method to identify flanking sequence tags in *Chlamydomonas* using 3′-RACE. Plant Methods **8:** 21.

Niyogi, K.K., Björkman, O., and Grossman, A.R. (1997). *Chlamydomonas* xanthophyll cycle mutants identified by video imaging of chlorophyll fluorescence quenching. Plant Cell **9:** 1369–1380.

Ozawa, T., Nishitani, K., Sako, Y., and Umezawa, Y. (2005). A high-throughput screening of genes that encode proteins transported into the endoplasmic reticulum in mammalian cells. Nucleic Acids Res. **33:** e34.

Pazour, G.J., Agrin, N., Leszyk, J., and Witman, G.B. (2005). Proteomic analysis of a eukaryotic cilium. J. Cell Biol. **170:** 103–113.

Pazour, G.J., Sineshchekov, O.A., and Witman, G.B. (1995). Mutational analysis of the phototransduction pathway of *Chlamydomonas reinhardtii.* J. Cell Biol. **131:** 427–440.

Randolph-Anderson, B.L., Boynton, J.E., Gillham, N.W., Harris, E.H., Johnson, A.M., Dorthu, M.P., and Matagne, R.F. (1993). Further characterization of the respiratory deficient dum-1 mutation of *Chlamydomonas reinhardtii* and its use as a recipient for mitochondrial transformation. Mol. Gen. Genet. **236:** 235–244.

Schmidt, G.W., Matlin, K.S., and Chua, N.H. (1977). A rapid procedure for selective enrichment of photosynthetic electron transport mutants. Proc. Natl. Acad. Sci. USA **74:** 610–614.

Schnell, R.A., and Lefebvre, P.A. (1993). Isolation of the *Chlamydomonas* regulatory gene NIT2 by transposon tagging. Genetics **134:** 737–747.

Shepherd, H.S., Boynton, J.E., and Gillham, N.W. (1979). Mutations in nine chloroplast loci of *Chlamydomonas* affecting different photosynthetic functions. Proc. Natl. Acad. Sci. USA **76:** 1353–1357.

Shimogawara, K., Fujiwara, S., Grossman, A., and Usuda, H. (1998). High-efficiency transformation of *Chlamydomonas reinhardtii* by electroporation. Genetics **148:** 1821–1828.

Sizova, I., Fuhrmann, M., and Hegemann, P. (2001). A *Streptomyces rimosus* aphVIII gene coding for a new type phosphotransferase provides stable antibiotic resistance to *Chlamydomonas reinhardtii.* Gene **277:** 221–229.

Tam, L.W., and Lefebvre, P.A. (1993). Cloning of flagellar genes in *Chlamydomonas reinhardtii* by DNA insertional mutagenesis. Genetics **135:** 375–384.

Tran, P.T., Sharifi, M.N., Poddar, S., Dent, R.M., and Niyogi, K.K. (2012). Intragenic enhancers and suppressors of phytoene desaturase mutations in *Chlamydomonas reinhardtii.* PLoS ONE **7:** e42196.

van Opijnen, T., Bodi, K.L., and Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat. Methods **6:** 767–772.

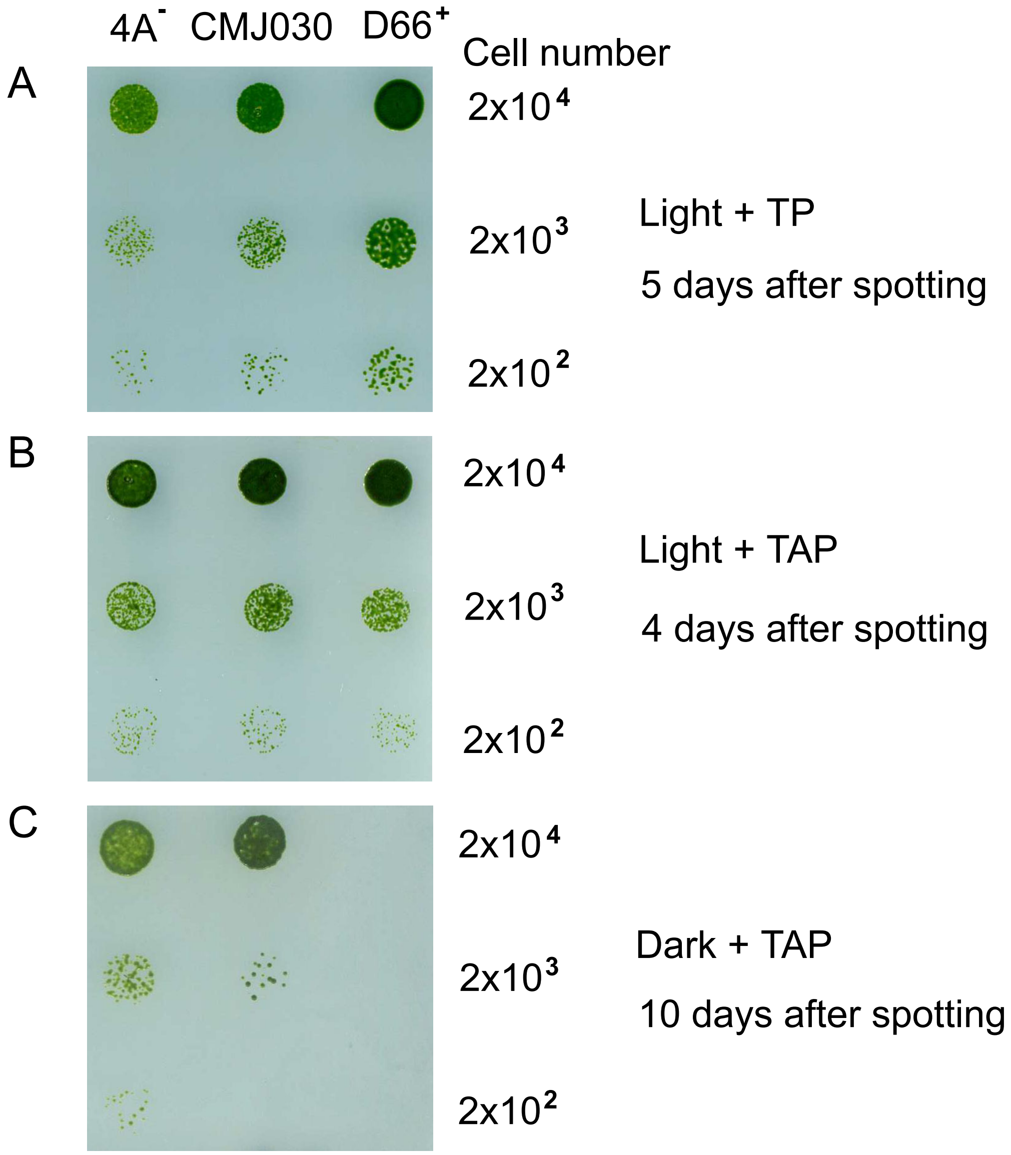
Wang, Y., Duanmu, D., and Spalding, M.H. (2011). Carbon dioxide concentrating mechanism in *Chlamydomonas reinhardtii*: inorganic carbon transport and CO_2 recapture. Photosynth. Res. **109:** 115–122.

Wang, Z.T., Ullrich, N., Joo, S., Waffenschmidt, S., and Goodenough, U. (2009). Algal lipid bodies: stress induction, purification, and biochemical characterization in wild-type and starchless *Chlamydomonas reinhardtii.* Eukaryot. Cell **8:** 1856–1868.

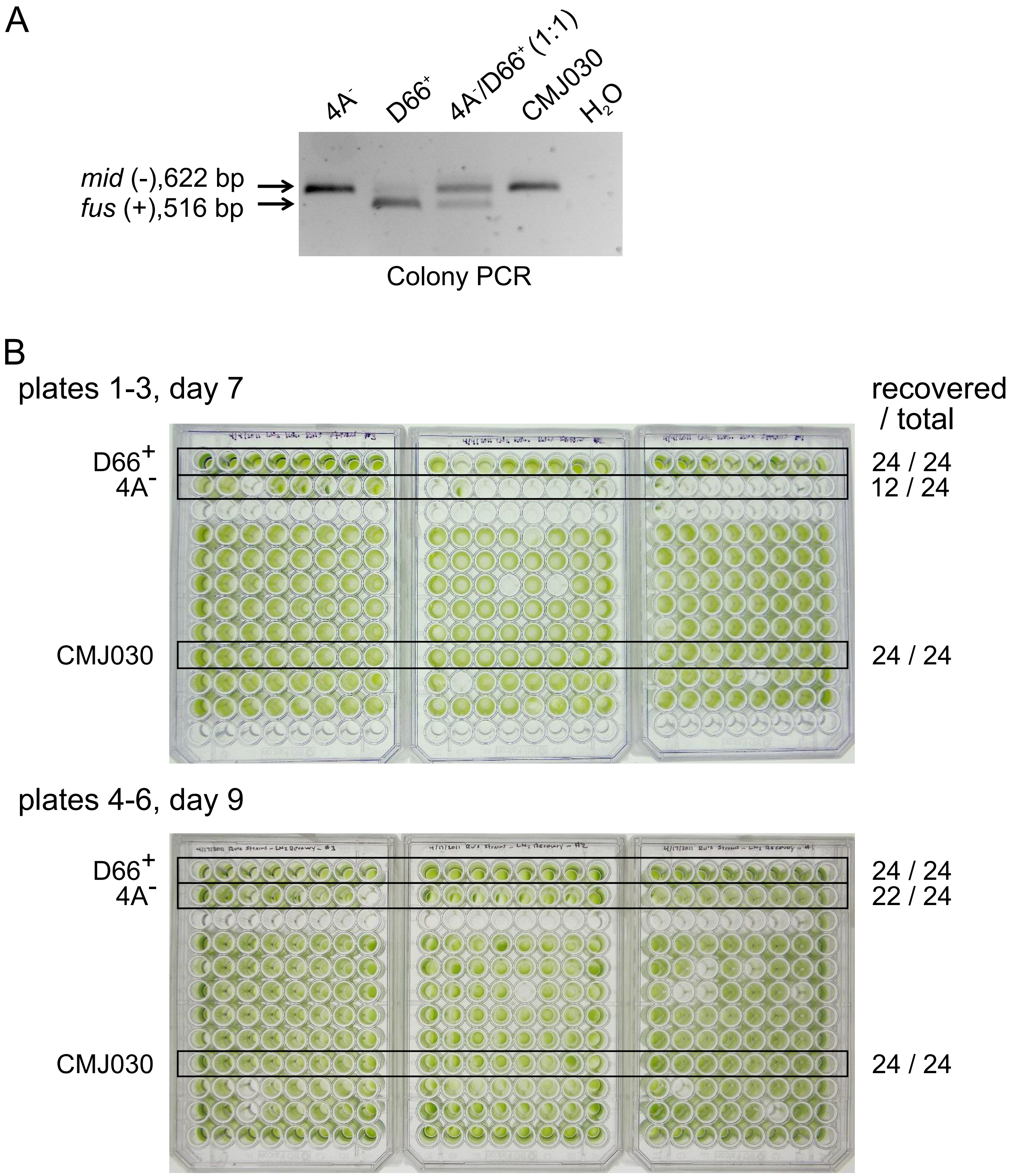
Werner, R., and Mergenhagen, D. (1998). Mating Type Determination of *Chlamydomonas reinhardtii* by PCR. Plant Mol. Biol. Rep. **16:** 295–299.

Winzeler, E.A., et al. (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science **285:** 901–906.

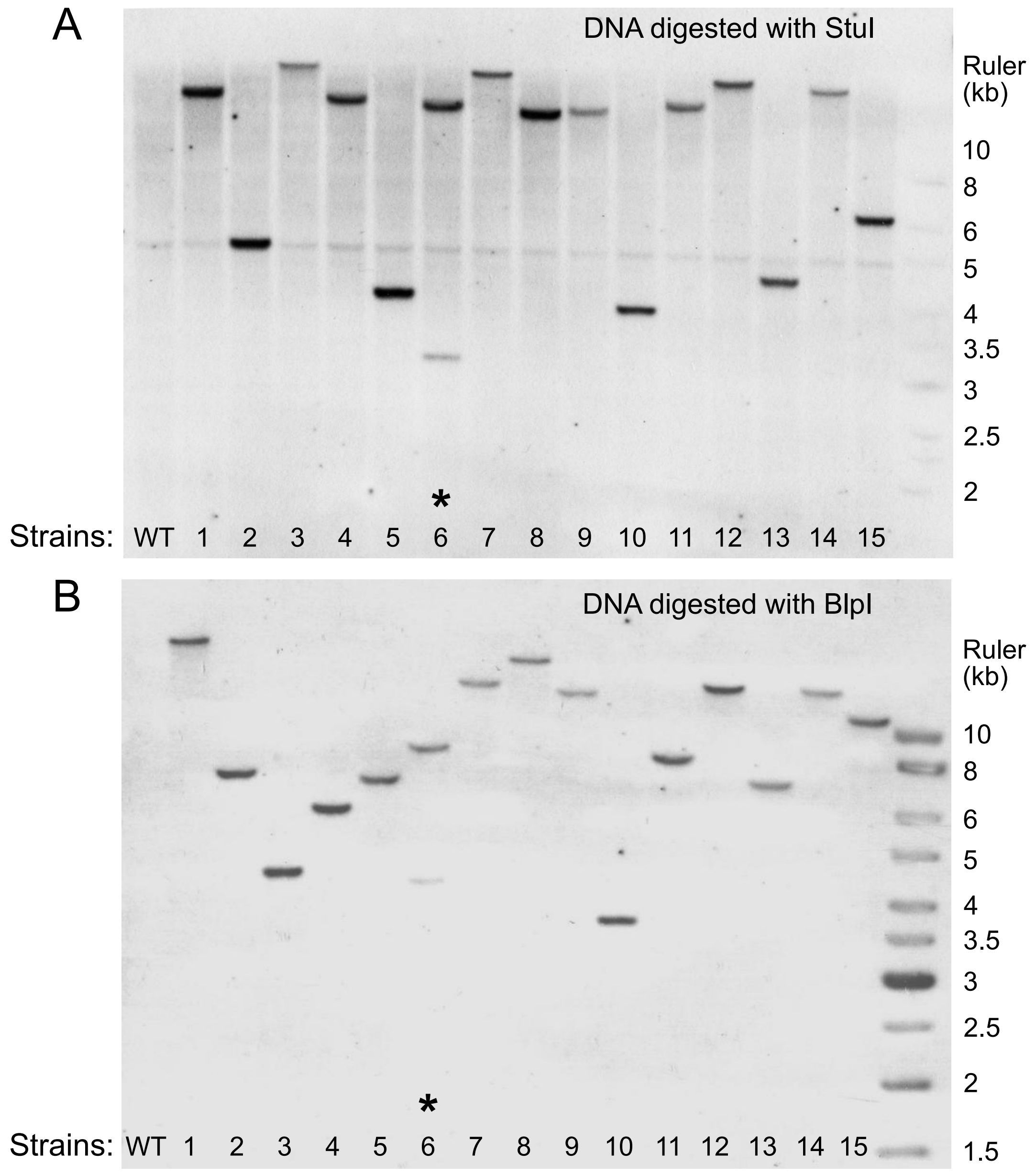
Supplemental Figures:



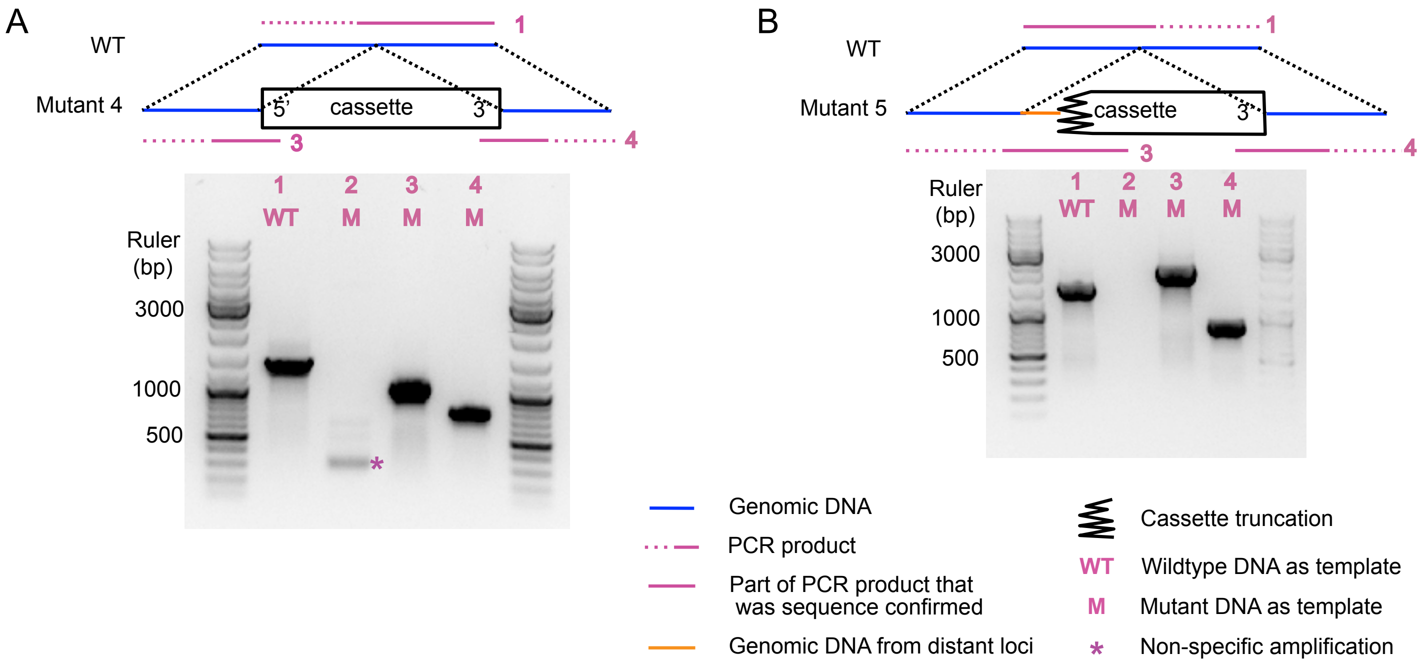
Supplemental Figure 1. The background strain CMJ030 can grow photoautotrophically, mixotrophically, and heterotrophically. CMJ030 is the progeny of 4A$^-$ and D66$^+$. TP is Tris Phosphate medium; TAP is Tris Acetate Phosphate medium. The indicated number of cells were spotted onto solid media and grown under conditions as follows: (A) TP under 50 μmol photons m^{-2} s^{-1} light, (B) TAP under 50 μmol photons m^{-2} s^{-1} light, (C) TAP grown in the dark. Plates were imaged after the indicated number of days.


A
4A-
D66+
4A-/D66+ (1:1)
CMJ030
H2O
mid (-),622 bp
fus (+),516 bp
Colony PCR
B
plates 1-3, day 7
recovered / total
D66+
4A-
CMJ030
24 / 24
12 / 24
24 / 24
plates 4-6, day 9
D66+
4A-
CMJ030
24 / 24
22 / 24
24 / 24

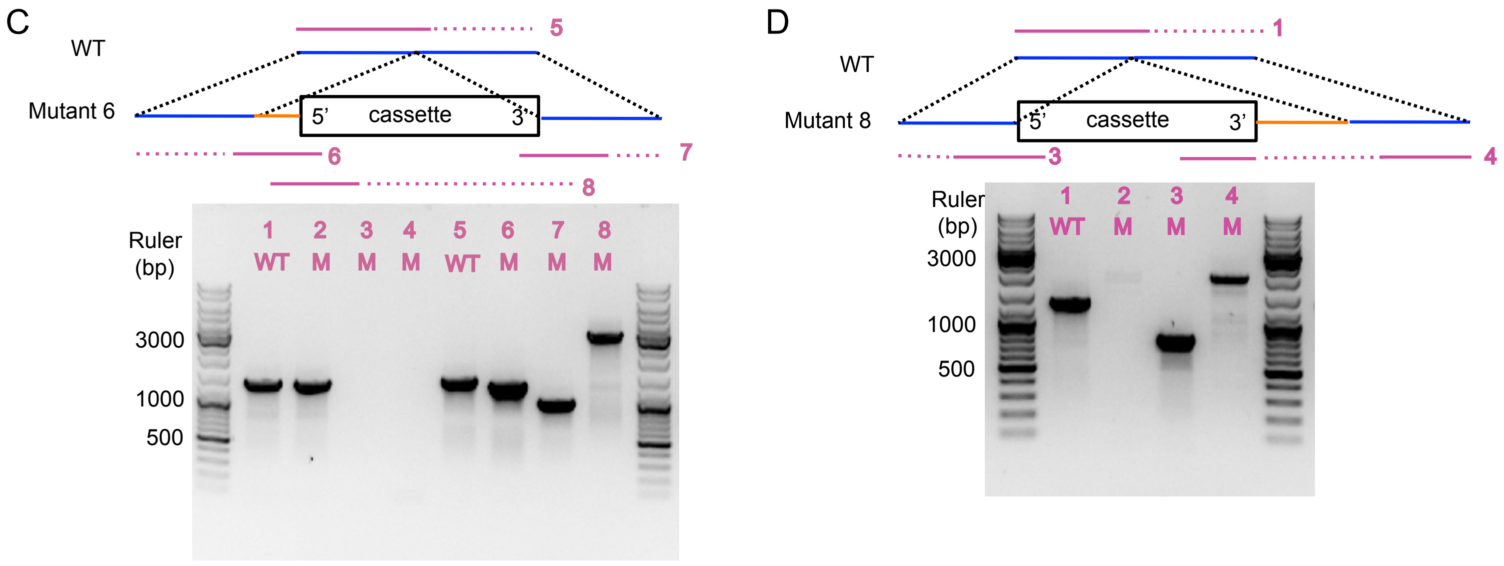
Supplemental Figure 2. CMJ030 is mating type *minus* and recovers efficiently from cryogenic storage in liquid nitrogen. (A) To determine the mating type of CMJ030, colony PCR was performed on the indicated strains using two primer pairs in each reaction: mating type *plus* specific primers (*mid*-specific forward and reverse primers), and mating type *minus* specific primers (*fus1*-specific forward and reverse primers). 4A^{-} and D66^{+} serve as positive controls for mating type *minus* and *plus*, respectively. An equal mixture of D66^{+} and 4A^{-} DNA was used as template for PCR in the third sample from the left. (B) CMJ030 and its background strains 4A^{-} and D66^{+} were thawed from cryogenic storage in liquid nitrogen and recovered in TAP under 50 μmol photons m^{-2} s^{-1} light. Two independent experiments were performed (plate 1-3, plate 4-6). After 7 or 9 days, both CMJ030 and D66^{+} recovered 100% but 4A^{-} appeared to recover slowly and incompletely. Unmarked rows were other progeny from the D66^{+}/4A^{-} cross.



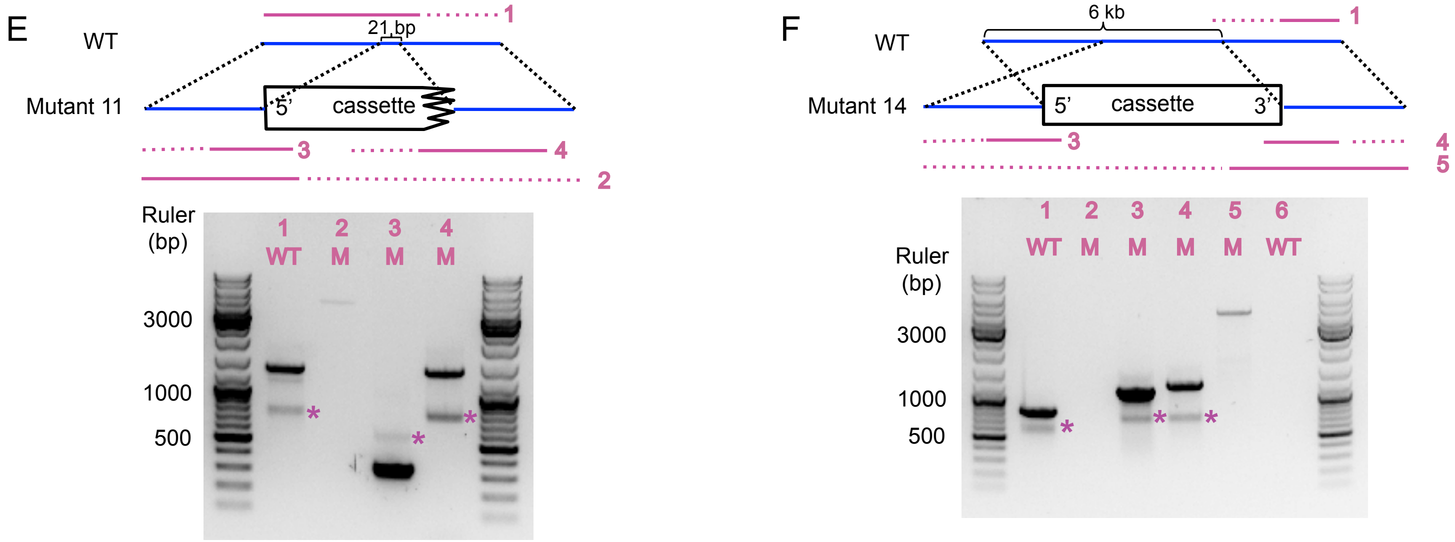
Supplemental Figure 3. Cassette copy number was determined in 15 individual mutants by DNA gel blot. (A) Genomic DNA from 15 individual mutants was digested with StuI, and a DNA gel blot was performed with a probe against the *AphVIII* gene. The lane showing two insertions is marked with *. (B) The DNA gel blot was repeated with a different enzyme, BlpI.



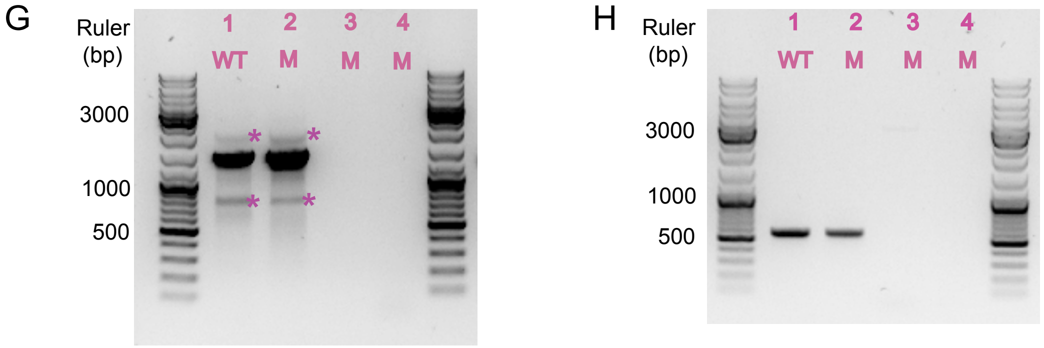
Supplemental Figure 4. PCR analysis revealed the structure of insertion sites in individual mutants. For each mutant, a model of the insertion locus is presented. Numbers by the purple lines correspond to the lane numbers on the gel images. The mapping sites of the mutants are listed in Supplemental Dataset 1; primers used for each PCR reaction are listed in Supplemental Dataset 2. (A) Mutant 4 yielded flanking sequences from both sides. The 5' side flanking sequence was mapped to chromosome 5, the 3' side flanking sequence mapped to multiple locations. PCR amplification of the genomic region around the 5' side flanking sequence yielded the expected product in WT (wildtype, lane 1) but not in mutant 4 (lane 2). Mutant 4 yielded PCR products amplifying the junctions between the cassette and the flanking genomic DNA (lanes 3 & 4). Sequencing of these PCR products confirmed an intact cassette and perfect match of the 5' and 3' side flanking sequences. (B) Mutant 5 yielded a flanking sequence from only the 3' side of the cassette, which was mapped to chromosome 16. PCR amplification of the genomic region around the 3' side flanking sequence yielded the expected product in WT (lane 1) but not mutant 5 (lane 2). Mutant 5 yielded PCR products containing the junctions between the cassette and the flanking genomic DNA (lanes 3 & 4). Sequencing of PCR products indicated a 418 bp truncation of the 5' side of the cassette, and a 365 bp fragment from chromosome 5 inserted between the 5' side of the cassette and the surrounding genomic DNA. *(Supplemental Figure 4 is continued on the next page.)*



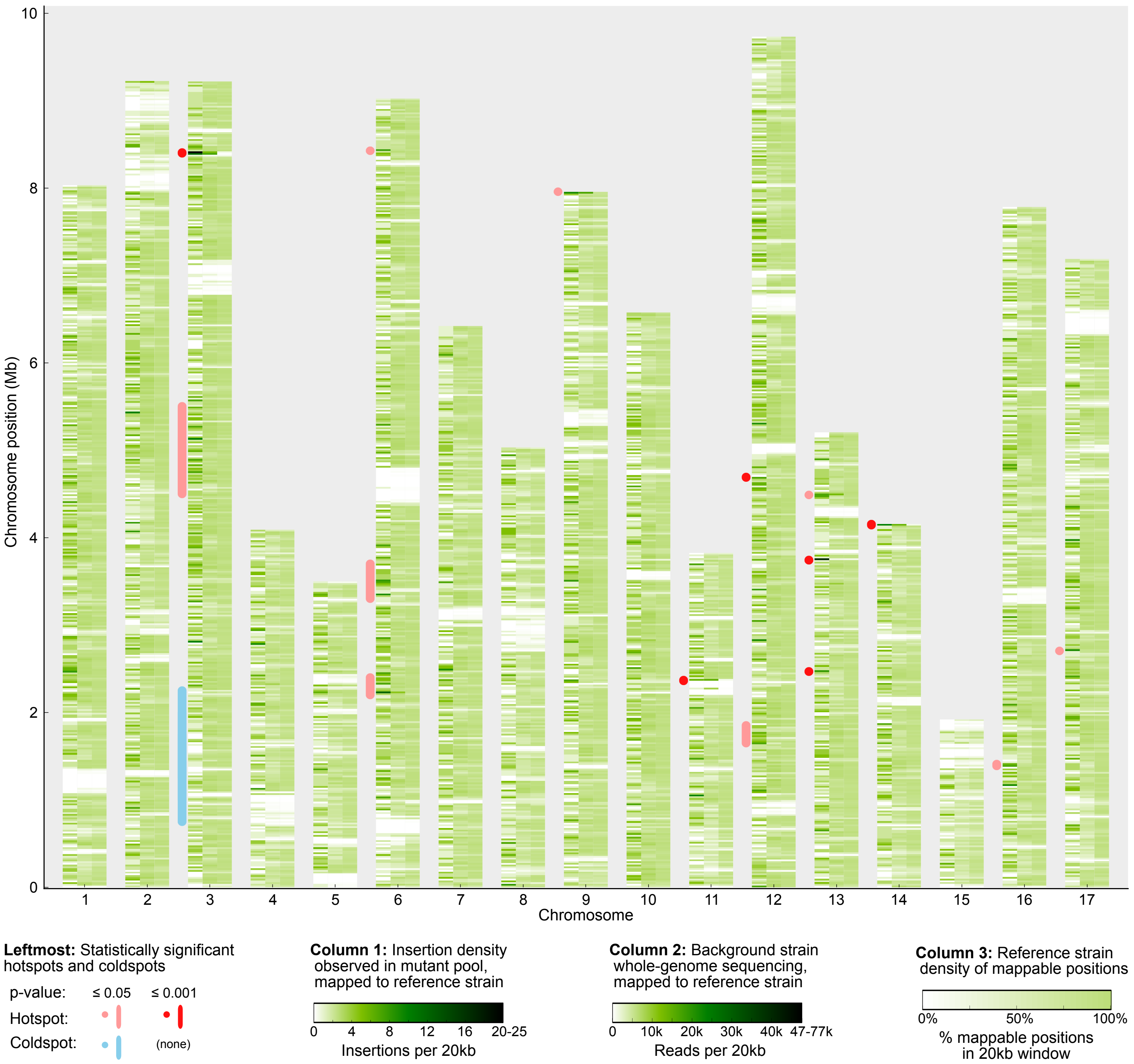
Supplemental Figure 4 *continued.* (C) Mutant 6 yielded flanking sequences from both sides; the 5' side flanking sequence mapped to chromosome 10, the 3' side flanking sequence mapped to chromosome 16. Cassette-genome junctions could not be amplified in mutant 6 using primers based on the 5' side flanking sequence (lanes 3 & 4); instead, a PCR product containing the genomic region around the 5' side flanking sequence was amplified in both WT (lane 1) and mutant 6 (lane 2). These results indicate a misleading 5' side flanking sequence. PCR amplification of the genomic region around the 3' side flanking sequence yielded the expected product in WT (1.4 kb, lane 5) and a longer product (3 kb) in mutant 6 (lane 8), as would be expected from a cassette insertion. Mutant 6 yielded PCR products containing the junctions between the cassette and the flanking genomic DNA using primers based on the 3' side flanking sequence (lanes 6 & 7). Sequencing of PCR products indicated a 134 bp fragment from chromosome 10 followed by a 414 bp fragment from chromosome 17, inserted between the 5' end of the cassette and the surrounding genomic DNA. (D) Mutant 8 yielded flanking sequences from both sides; the 5' side was mapped to chromosome 9, the 3' side mapped to multiple locations. PCR amplification of the genomic region around the 5' side flanking sequence yielded the expected product in WT (lane 1) but not mutant 8 (lane 2). Mutant 8 yielded PCR products containing the junctions between the cassette and the flanking genomic DNA based on the 5' side flanking sequence (lanes 3 & 4). Sequencing of these PCR products indicated a ~1,000 bp DNA fragment inserted between the 3' side of cassette and the surrounding genomic DNA, which could not be sequenced. *(Supplemental Figure 4 is continued on the next page.)*



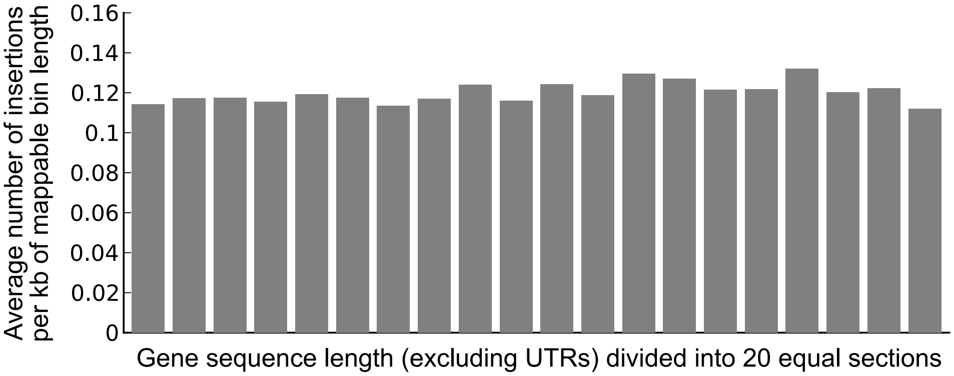
Supplemental Figure 4 *continued.* (E) Mutant 11 yielded a flanking sequence from the 5' side only. PCR amplification of the genomic region around the 5' side flanking sequence yielded the expected product in WT (lane 1, 1.5 kb) and a much longer product in mutant 11 (lane 2, faint band on the top, >3 kb). Mutant 11 yielded PCR products containing the junctions between the cassette and the flanking genomic DNA using primers based on the 5' side flanking sequence (lanes 3 & 4). Sequencing of PCR products indicated a 508 bp truncation of the 3' side of the cassette and a 21 bp deletion in the genomic DNA at the site of the insertion. (F) Mutant 14 yielded flanking sequences from both sides. Both flanking sequences were mapped to chromosome 8 at positions 6 kb apart, in orientations consistent with a local inversion of the genomic DNA on one side of the cassette. PCR amplification of the genomic region around the 3' side flanking sequence yielded the expected product in WT (lane 1) but not in mutant 14 (lane 2). Amplification of the genomic region around the 5' side flanking sequence was not successful in either WT or mutant 14 after many trials, possibly because of poor quality of the reference genome sequence in this region. Mutant 14 yielded PCR products containing the junctions between the cassette and the flanking genomic sequences on both sides (lane 3 & 4). Using primers based on both flanking sequences, a 4 kb product spanning the cassette was amplified in mutant 14 (lane 5) but not in WT (lane 6). PCR and sequencing results indicate that the cassette was intact and flanking sequences from both sides were accurate, but there was an inversion of the genomic DNA on the 5' side of cassette which was not present in WT. *(Supplemental Figure 4 is continued on the next page.)*



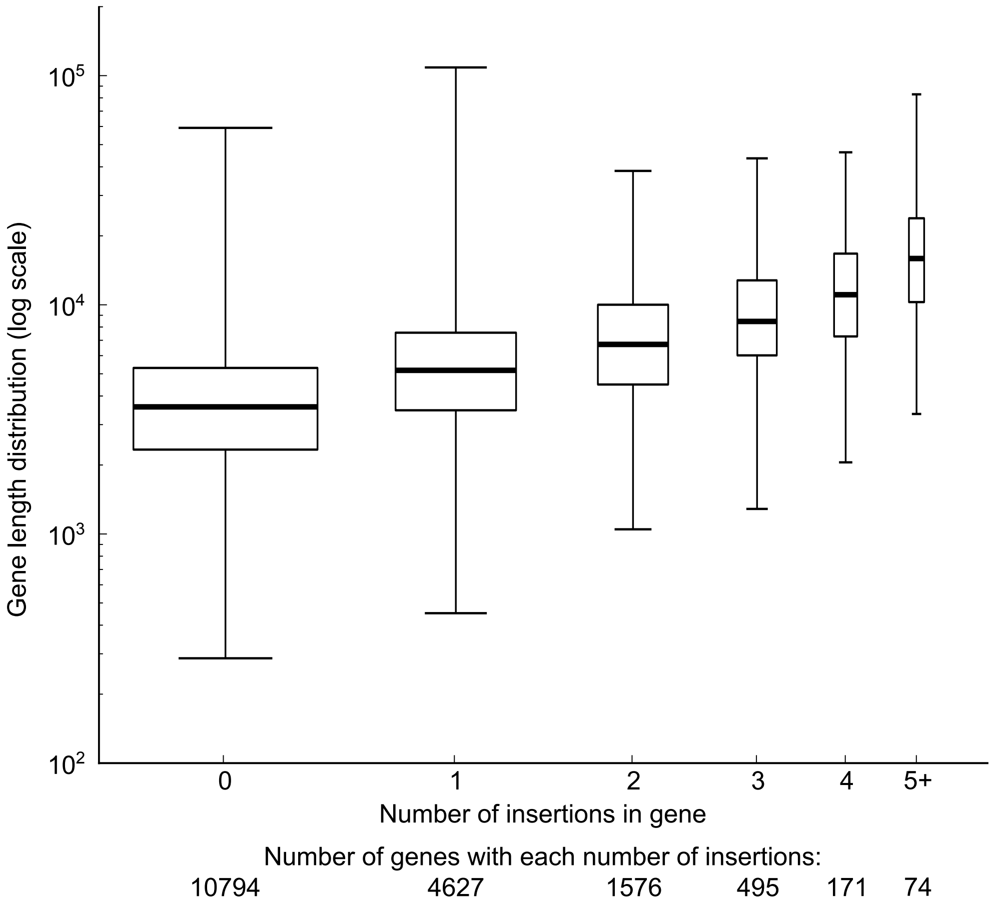
Supplemental Figure 4 ***continued.*** (G) Mutant 1 yielded flanking sequences from both sides, but the 5’ side was mapped to the cassette. Amplification of the genomic region around the 3’ side flanking sequence yielded the same PCR product in both WT (lane 1) and mutant 1 (lane 2). No junctions between the cassette and the flanking genomic sequences could be amplified from either side using primers based on the 3’ side flanking sequence (lanes 3 & 4), suggesting a misleading flanking sequence. (H) Mutant 7 yielded flanking sequences from both sides, but the 5’ side flanking sequence was mapped to the cassette. Amplification of the genomic region around the 3’ side flanking sequence yielded the same product in both WT (lane 1) and mutant 7 (lane 2). No junctions between the cassette and the flanking genomic sequences could be amplified from either side using primers based on 3’ side flanking sequence (lanes 3 & 4), suggesting a misleading flanking sequence.


Chromosome position (Mb)
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Chromosome
Leftmost: Statistically significant hotspots and coldspots
p-value:
≤ 0.05
≤ 0.001
Hotspot:
Coldspot:
(none)
Column 1: Insertion density observed in mutant pool, mapped to reference strain
0
4
8
12
16
20-25
Insertions per 20kb
Column 2: Background strain whole-genome sequencing, mapped to reference strain
0
10k
20k
30k
40k
47-77k
Reads per 20kb
Column 3: Reference strain density of mappable positions
0%
50%
100%
% mappable positions in 20kb window

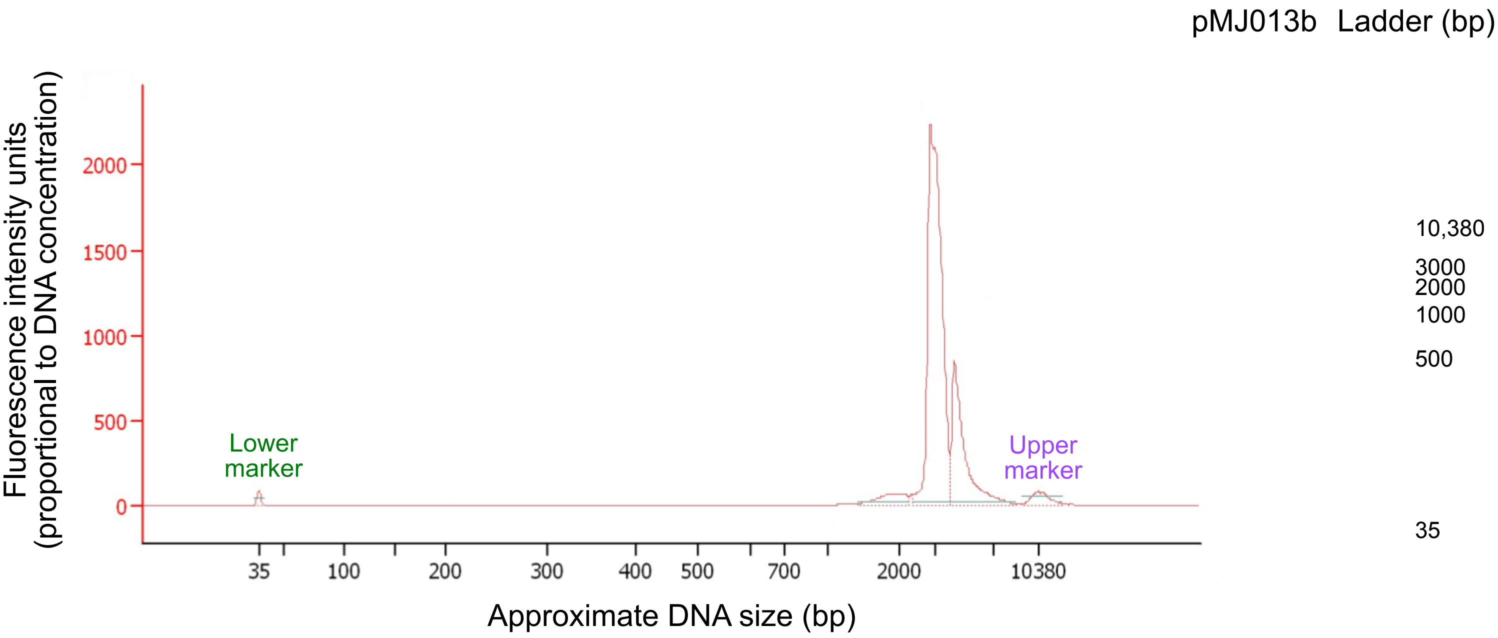
Supplemental Figure 5. Some of the potential hotspots can be explained by local amplifications of the genome of the background strain. The first green column for each chromosome shows the observed insertion density (same as in Figure 3A); the second column shows the density of observed uniquely mappable 20–21 bp sequences from whole-genome sequencing of our background strain aligned to the reference strain genome sequence; the third column shows the density of all possible uniquely mappable positions based on the genome sequence of the reference strain. The colorbars for the three datasets were scaled for easier visual comparison so that the median color is the same for all three. The blue and red marks on the left of each chromosome show a summary of statistically significant hotspot and coldspot locations (same as in Figure 3A).



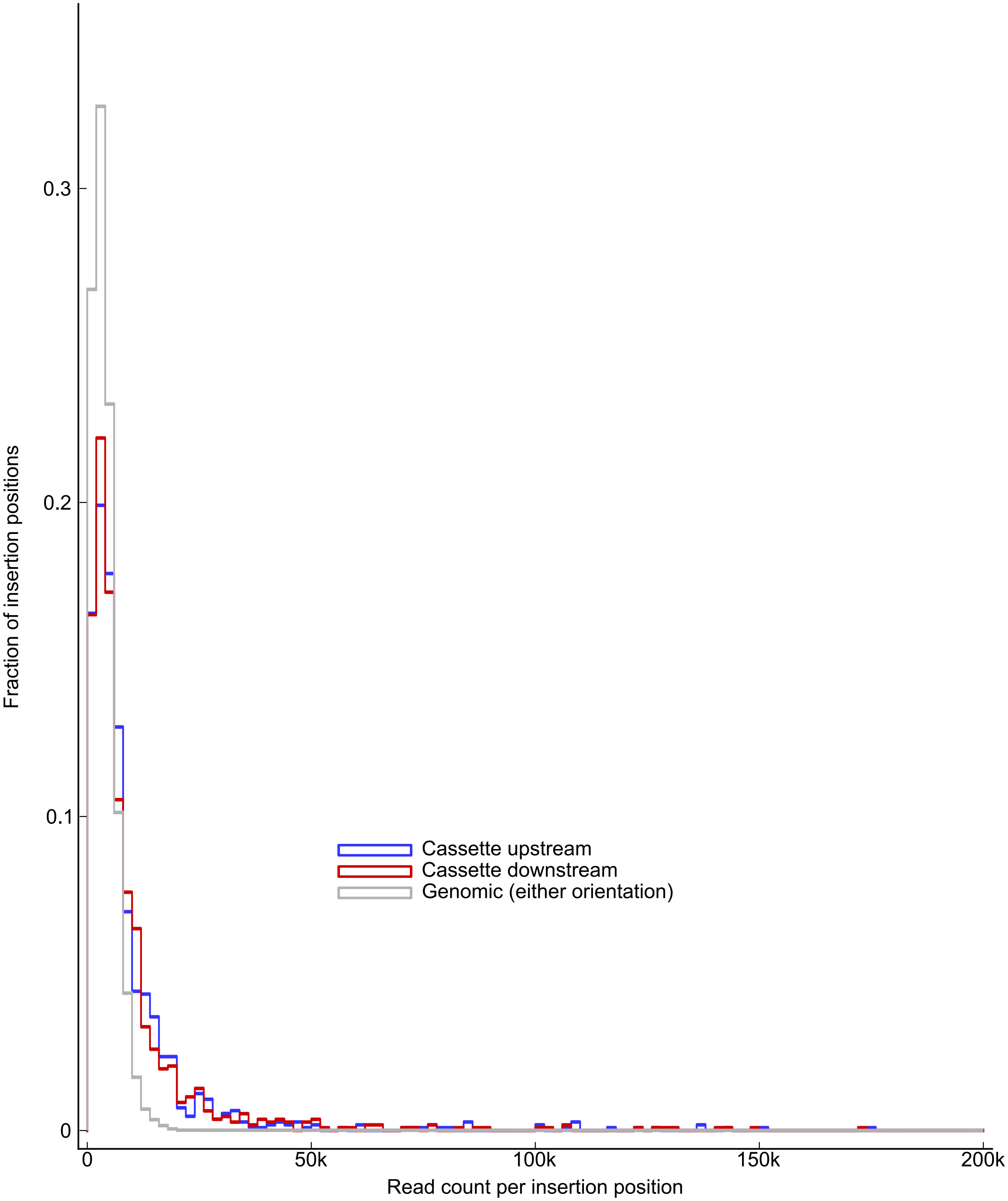
Supplemental Figure 6. Insertion density does not vary along the gene length. Each gene (excluding 5' and 3' UTRs, but including introns) was divided into 20 equal length sections. The insertion density normalized to uniquely mappable length was calculated for each section. For each section, the insertion density was averaged across all genes and plotted. None of the results are statistically significantly different from the overall insertion density (p-values 0.09 or higher, exact binomial test with Benjamini-Hochberg correction for multiple testing).



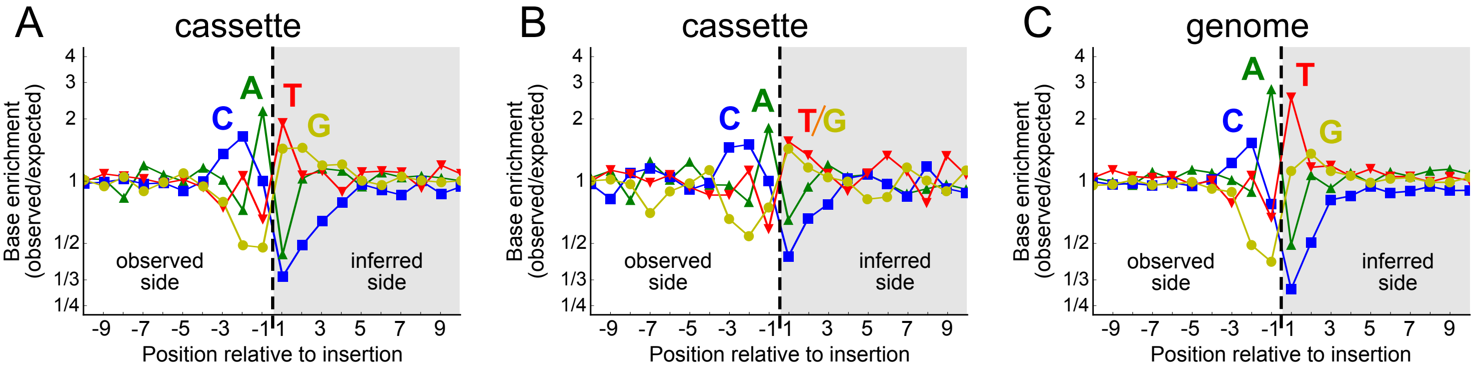
Supplemental Figure 7. Longer genes have more insertions. The boxplot shows the gene length distribution for genes with 0, 1, 2, 3, 4 and 5+ insertions. The whiskers indicate min/max values, the box indicates the 25/75 percentiles, and the middle line indicates the median. The boxplot width is proportional to the square root of the number of genes in each group.



Supplemental Figure 8. The transforming DNA is intact before transformation. 10 μl of transforming DNA (2.66 kb digestion product from pMJ013b) at 5 ng/μl was submitted for the Agilent Bioanalyzer High Sensitivity DNA Assay, and the electropherogram of the bioanalyzer result is shown. The gel images on the right were created based on the electropherogram trace by the Agilent Bioanalyzer software. The result shows the transforming DNA has a single clean peak around 2.6 kb with a lower shoulder around 4 kb. The 4 kb shoulder is probably due to incomplete digestion of the plasmid, but the backbone of the plasmid has no alignments to the cassette, so incomplete digestion products cannot explain the cassette fragments observed in the flanking sequence data.


Fraction of insertion positions
0.3
0.2
0.1
0
0
50k
100k
150k
200k
Read count per insertion position
Cassette upstream
Cassette downstream
Genomic (either orientation)

Supplemental Figure 9. The cassette read count distribution has a longer tail than the genomic read count distribution. This histogram shows the read count distributions for insertion positions mapped to the cassette and the genome, in the dataset used in Figure 5. The non-parametric Kruskal-Wallis test was used to compare the read count distributions: the p-values for the comparison between cassette-mapped fragment read counts and genome-mapped position read counts are 1.32×10^{-79} for upstream cassette fragments and 7.92×10^{-72} for downstream fragments.

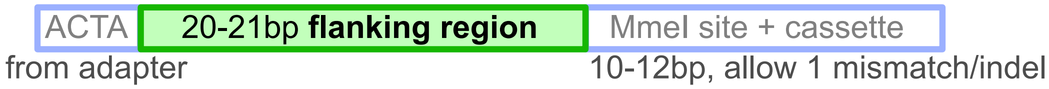


dataset used for Figure 5

dataset used for Figures 3 and 4

Supplemental Figure 10. Cassette and genomic insertion positions are enriched for the CAITG motif. For each insertion position, we took 10 bp of the sequenced flanking region preceding the insertion position, as well as the following 10 bp sequence from the other side of the mapped insertion position in the genome or cassette. We calculated the average A,C,T,G base content for each position, averaging together all the insertions from each dataset. We also calculated the average base content for the +/– 10 bp window around each theoretically mappable insertion position in the genome or cassette. We plotted the ratio between the observed and expected base fractions, for each base at each position. The insertion occurs between positions –1 and 1: the left side (white background) is the sequenced flanking region, and the right side (grey background) is inferred based on the genome sequence. (A) The base content enrichment is plotted for cassette insertions from the dataset used in Figure 5. Each insertion position is weighed by its abundance, since the data indicate that high-abundance cassette positions correspond to multiple mutants (see Figure 5B). (B) The same plot as A is shown for the dataset used in Figures 3 and 4. (C) The base content enrichment is plotted for genomic insertions in the dataset used in Figures 3 and 4. Each insertion position is counted once, since the vast majority of genomic insertions correspond to single mutants.

1) Remove adapter and cassette sequence → get flanking region;
filter out reads that don't match the expected structure



(Optionally collapse identical flanking region sequences together
to speed up downstream processing - keep the original counts)

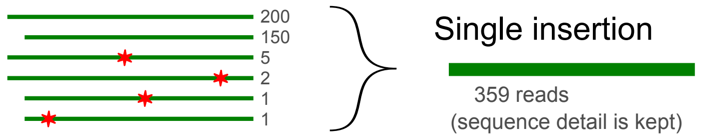
2) Align reads to genome and cassette (allow one mismatch),
separate into categories

Alignment result for each read
(how many genome/cassette
locations it aligned to)
→ category:

genome:	cassette:	
-	-	→ unaligned
0+	1	→ **cassette**
1	-	→ **genomic-unique**
2+	-	→ genomic-multiple

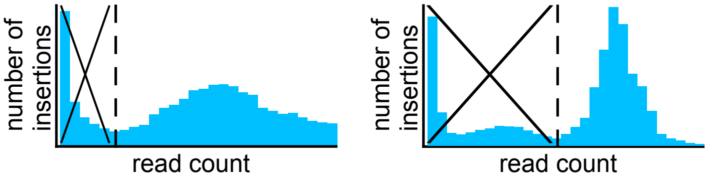
3) Join reads into insertions based on alignment position;
annotate with gene information.

Multiple sequences
at same position:
some with mismatches,
varied read numbers



4) Remove low-read "insertions" - likely noise or contamination

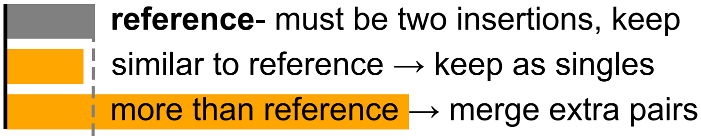
Set cutoff below
main "real" peak
(where we think that is
depends on the dataset)



5) Merge replicate datasets together if desired
(just match insertions by position and add readcounts together)

6) Merge reads that probably come from a single real insertion

Number of adjacent insertion pairs by category → decision:

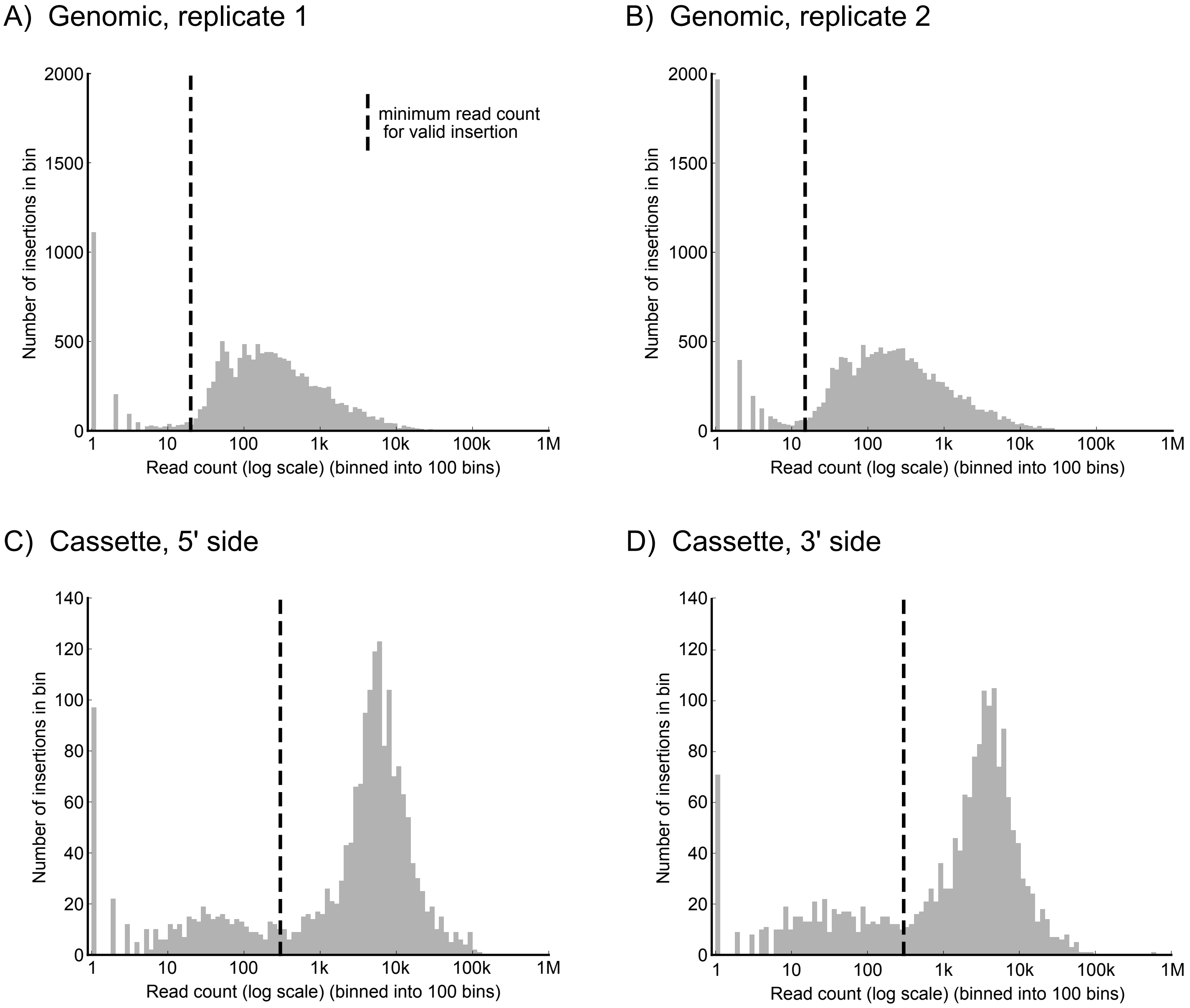


1bp distance, opposite strand — **reference**- must be two insertions, keep

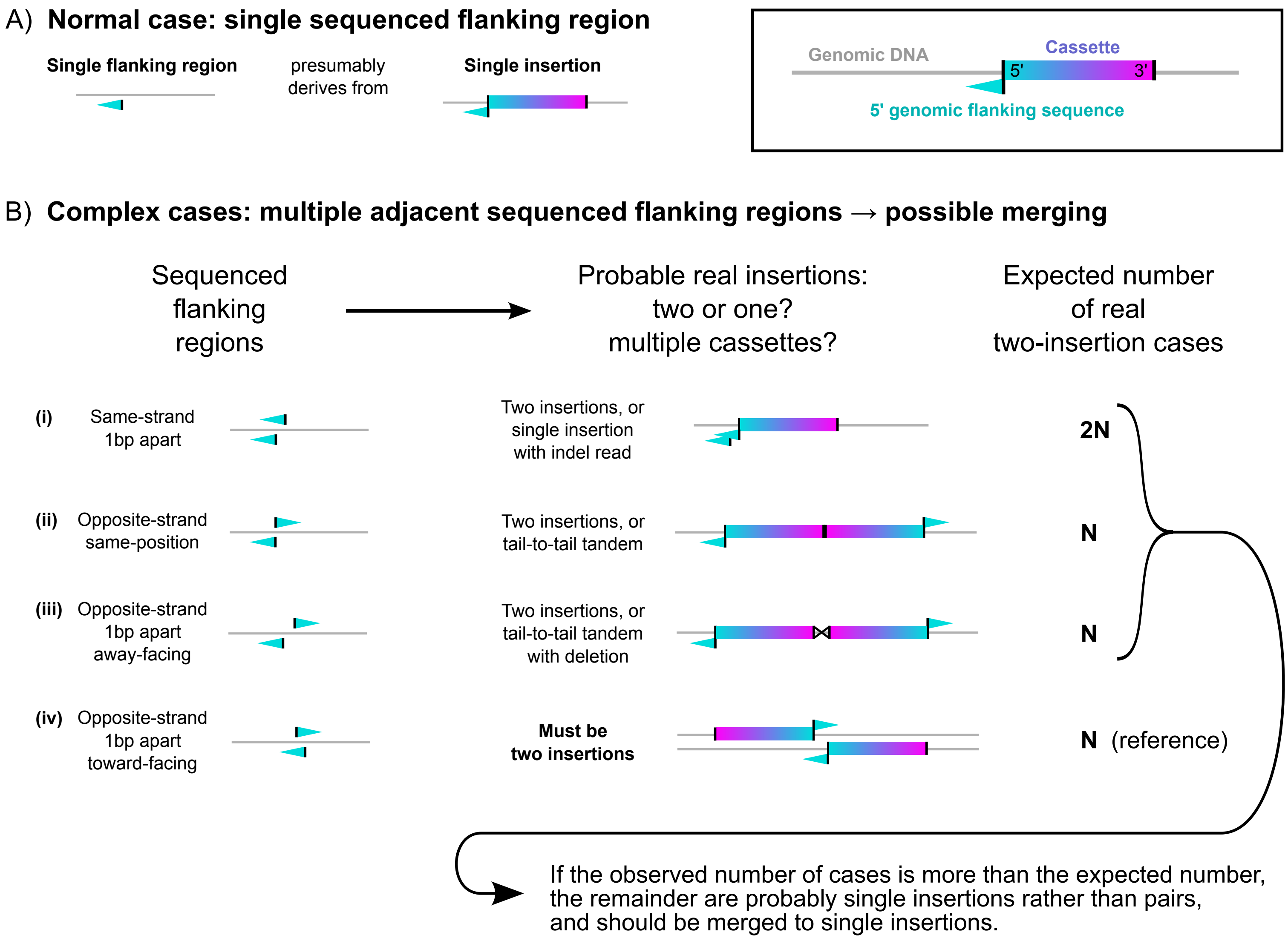
1bp distance, same strand — similar to reference → keep as singles

same position, opposite strand — more than reference → merge extra pairs

Supplemental Figure 11. Our computational pipeline determines insertion sites based on deep-sequencing data. See methods section "Inferring insertion details from pooled mutant deep-sequencing data" for detailed description.



Supplemental Figure 12. Read count cutoffs are set based on each dataset's read count distribution. See methods section "Inferring insertion details from pooled mutant deep-sequencing data" point 4 for detailed explanation. Read count distributions and cutoffs are plotted for each dataset: (A,B) genomic insertions for the two technical replicates from the dataset used in Figures 3 and 4; (C,D) cassette insertions (5' and 3' sides) from the dataset used in Figure 5.


A) Normal case: single sequenced flanking region
Single flanking region
presumably derives from
Single insertion
Genomic DNA
Cassette
5'
3'
5' genomic flanking sequence
B) Complex cases: multiple adjacent sequenced flanking regions → possible merging
Sequenced flanking regions
Probable real insertions: two or one? multiple cassettes?
Expected number of real two-insertion cases
(i) Same-strand 1bp apart
Two insertions, or single insertion with indel read
2N
(ii) Opposite-strand same-position
Two insertions, or tail-to-tail tandem
N
(iii) Opposite-strand 1bp apart away-facing
Two insertions, or tail-to-tail tandem with deletion
N
(iv) Opposite-strand 1bp apart toward-facing
Must be two insertions
N (reference)
If the observed number of cases is more than the expected number, the remainder are probably single insertions rather than pairs, and should be merged to single insertions.

Supplemental Figure 13. We merge flanking sequences mapped to adjacent positions if we suspect that they originated from a single insertion. (A) Usually, each mapped insertion position is well separated from others, and we can assume that each mapped insertion position corresponds to a distinct insertion event. (B) Multiple mapped insertion positions within 1 bp of each other could be a result of two independent insertions, or could be derived from a single insertion. See methods section "Extracting insertion details from deep-sequencing data" point 6 for detailed description of the different cases.

Supplemental Tables:

Supplemental Table 1. Summary of phenotypes of the background strain CMJ030.

Phenotypes	**D66$^+$**	**4A$^-$**	**CMJ030**
Photoautotrophic growth	✔	✔	✔
Heterotrophic growth	✗	✔	✔
Mixotrophic growth	✔	✔	✔
Green in the dark	✗	✔	✔
Normal swimming	✔	✗	✔
Efficient electroporation	✔	✔	✔
Mating type	*plus*	*minus*	*minus*
Efficient mating	✔	✔	✔
Recovery from cryogenic storage	✔	slow	✔
Cell wall	*cw15*	intact	*cw15*-like

Supplemental Methods:

Spot tests

Cultures of CMJ030 and its parental strains, D66$^+$ and 4A$^-$, were grown in the same growth medium and light intensity before and after spotting. Cells were grown in liquid until they reached a cell density of 2×10^6 cells/ml. Cell concentration was determined with a hemocytometer. Three serial dilutions were performed using fresh growth medium, and 10 μl of each dilution containing the indicated number of cells (3×10^4, 3×10^3, 3×10^2) were spotted onto plates of TAP or TP. The growth conditions were as follows: TP plates under 50 μmol photons m^{-2} s^{-1} light (Supplemental Figure 1A online); TAP plates under 50 μmol photons m^{-2} s^{-1} light (Supplemental Figure 1B online); and TAP plates in the dark (Supplemental Figure 1C online). Plates were imaged on the indicated day after spotting.

Flanking sequence extraction for individual mutants

To validate our tool, 15 colonies, representing 15 different mutants, were randomly picked from transformation plates, streaked out onto new TAP plates, and grown under low light (5 μmol photons m^{-2} s^{-1} light). A single colony of each mutant was picked and grown up in 30 ml TAP to a cell density of 3–4 x 10^6 cells/ml. Cells were harvested by centrifugation at 3,000 g for 10 min. The supernatant was decanted, and the pellet was used for extraction of genomic DNA by phenol-chloroform (Green and Sambrook, 1989). The cell pellet was lysed by addition of 800 μl buffer containing 1% (wt/vol) SDS, 200 mM NaCl, 20 mM ethylenediaminetetraacetic acid (EDTA), and 50 mM Tris-HCl pH 8.0, followed by vortexing. The lysate was split into two 1.5 ml tubes, each containing 500 μl of phenol:chloroform:isoamyl (Phenol:CIA, 25:24:1, Sigma). The mix was vortexed for 2 min and then centrifuged at 20,000 g for 5 mins in an Eppendorf centrifuge (5415 D). The aqueous phase was transferred to a new tube with 1.6 μl RNase A (5 PRIME) at 100 mg/ml, mixed by vortexing, and then incubated at 37°C for 30 min. The above Phenol:CIA extraction was repeated three times, followed by one extraction with chloroform:isoamyl (CIA, 24:1). The DNA was precipitated by addition of 2.5x 100% ethanol and incubation at –20°C for at least 30 min, then centrifuged at 16,000 g for 20 min at 4°C. The supernatant was discarded, and the pellet was washed with 1 ml 70% (vol/vol) ethanol. The

DNA was again centrifuged at 16,000 g for 20 min at 44°C, the supernatant was discarded, then the pellet was air-dried and resuspended in water. DNA concentration was quantified with the Qubit® 2.0 Fluorometer using the Qubit dsDNA HS assay kit (Invitrogen). All DNA concentration quantification was performed with this method.

DNA was digested by MmeI (New England Biolabs, NEB) as follows. 60 μl of genomic DNA at 25 ng/μl were combined with 20 μl 10x NEB buffer 4, 0.32 μl S-Adenosyl Methionine at 32 mM, 108 μl double-distilled water (ddH$_2$O), and 12 μl MmeI. The contents were mixed by vortexing and incubated at 37°C for 30 min, then treated with 1 μl Calf Intestinal Phosphatase (New England Biolabs, 10 units/μl) at 37°C for 1 hour to remove 5' phosphates to prevent self-ligation. Digested DNA was phenol-chloroform-extracted (1 round of Phenol:CIA extraction, followed by 1 round of CIA extraction, each extraction using the same procedure as above), ethanol-precipitated (same as above) and dissolved in water.

50 μM adaptors were prepared by mixing equal volumes of oMJ041 and oMJ042 at 100 μM each in water, placing the mixture in a heat block (E&K Scientific D-1200 AccuBlock Digital Dry Bath) at 96°C for 2 minutes, then placing the metal insert of the heat block containing the samples on the bench at room temperature and letting it cool for 1 hr.

Ligation reactions were prepared on ice, and contained 3 μl T4 DNA ligase buffer (NEB), 1 μl 50 μM adaptors, 25 μl digested DNA at 10 ng/μl, and 1 μl of 2,000 U/μl T4 DNA ligase (NEB). Ligations were incubated overnight (10–16 h) at 16°C in a PCR machine with a lid set at 25°C.

5' and 3' side flanking sequences were amplified in separate PCR reactions, using the adaptor-ligated samples as template. Primer oMJ044 anneals to the adaptor, and primers oMJ005 and oMJ155 anneal to the 5' and 3' sides of the transforming cassette, respectively. Primers oMJ044 and oMJ005 were used to amplify 5' side flanking sequences, and primers oMJ044 and oMJ155 were used to amplify 3' side flanking sequences. PCR products were amplified using Phusion® Hot Start II High-Fidelity DNA Polymerase (New England Biolabs). 50 μl PCR reactions contained 10 μl 5x Phusion GC Buffer, 1 μl 10 mM dNTPs, 1.5 μl 100% DMSO, 0.1 μl of each primer at 100 μM, 1.5 μl ligation product, 35.3 μl water, and 0.5 μl of 2 U/μl Phusion. PCR cycling parameters were as follows: 30 s at 98°C; 40 cycles of 10 s at 98°C, 25 s at 65°C, 15 s at 72°C; followed by a final extension of 2 min at 72°C.

The PCR products were run on a 1.8% (wt/vol) agarose gel, and bands with the expected size (210 bp for 5' flanking sequences, 237 bp for 3' flanking sequences) were gel extracted and cloned using the Zero-Blunt kit (Invitrogen K2700-20), then transformed into One Shot® TOP10 cells (following the Invitrogen protocols). *E. coli* transformants were selected on LB plates containing 50 μg/ml kanamycin. Six colonies of each cloned PCR product were randomly picked from the kanamycin LB plates and submitted for colony PCR sequencing using the M13 forward primer.

Sequences of PCR products contain adaptor sequences, 20–21 bp flanking genomic DNA, and sequence from the 5' or 3' end of the cassette.

The sequences of 20–21 bp flanking genomic DNA were mapped to the *Chlamydomonas* genome (v5.3 from Phytozome (Goodstein et al., 2012; Merchant et al., 2012)) and the cassette sequence with the program Bowtie (Langmead et al., 2009) to determine the insertion sites for these mutants (bowtie parameters "-v2 --all --best --tryhard --strata").

DNA gel blots

Transformation cassette copy number was determined by DNA blot hybridization analysis (Southern, 2006). Specifically, 5 μg of total genomic DNA from each mutant was digested with StuI or BlpI (New England Biolabs), and resulting fragments were separated on a 0.7% (wt/vol) agarose gel and blotted onto nitrocellulose membranes (Biorad). Hybridization was carried out with an alkaline phosphatase-labeled probe against the full *AphVIII* gene generated by PCR of the transformation cassette vector (AlkPhos Direct labeling system, Biorad) and specific signals were detected by chemiluminescence of the CDP-Star reagent (Biorad) on X-ray films (Thermo).

Inferring insertion details from pooled mutant deep-sequencing data

The original data were in Illumina fastq format, pass-filter reads only; the two technical replicates had 21 and 26 million reads. The replicates were processed separately, using a combination of existing tools and custom software (Supplemental Dataset 6 online, also available at github.com/Jonikas-Lab/Zhang-Patena-2014.git), following the workflow below (see Supplemental Figure 11 online for schematic).

(1) Adaptor and cassette sequences were removed from reads to obtain the flanking sequences; reads without the expected adaptor or cassette sequences were discarded. The expected read structure is: ACTA (from the adaptor), then 20–21 bp of the flanking region (due to MmeI cutting 20 or 21 bp from the binding site), then the cassette sequence starting with the MmeI binding site. First, ACTA was removed from the beginning of each sequence, and sequences not starting with ACTA were counted and discarded. Second, cutadapt (Martin, 2011) was used to trim the cassette sequence from the end of the reads, requiring at least 10 bp overlap with the cassette, allowing 10% errors, and requiring the result to be 20–21 bp long; sequences not trimmed correctly were counted and discarded. Third, any multiple identical sequences were collapsed to single ones using fastx_collapser (http://hannonlab.cshl.edu/fastx_toolkit), keeping track of the original count, to speed up the alignment; the sequence quality information was discarded. A custom wrapper script was used to accomplish all these steps; the full shell command was "deepseq_preprocessing_wrapper.py -C".

(2) Flanking sequences were aligned to the *Chlamydomonas* genome and to the cassette sequence and separated into categories based on mapping results: unaligned, aligned to cassette, multiple genomic alignments, unique genomic alignments. Bowtie (Langmead et al., 2009) version 1.0.0 was used for alignment, with “-v1 -k10 --strata --best --tryhard” options (allowing up to one mismatch; bowtie doesn't allow indels). Two separate alignment runs were done for each sample: one against our insertion cassette (GenBank accession number KJ572788, insertion cassette feature only), and one against the *Chlamydomonas* v5.3 genome (from Phytozome, no repeat masking) with the chloroplast and mitochondrial genomes added as separate chromosomes (from the National Center for Biotechnology Information (NCBI) website: NC_005353 and NC_001638). Both result files were parsed in parallel to categorize the reads: reads that did not align to either reference were counted and discarded; reads that aligned to the cassette were used for the cassette dataset, regardless of any genome alignment; reads that aligned uniquely to the genome were used for the genomic dataset; reads that aligned to multiple genomic locations were counted and discarded (the cassette and genomic datasets were separately processed further). A custom wrapper script was used to accomplish all these steps – the full shell command was “deepseq_alignment_wrapper.py -c”.

(3) Flanking sequences were combined into insertion positions based on alignment position, and were annotated with gene information. The genomic and cassette alignment files were parsed (in python, using HTSeq, http://www-huber.embl.de/users/anders/HTSeq), the bowtie flanking sequence alignment locations were converted to the cassette insertion locations (by adjusting the position and strand based on which end of the flanking region was adjacent to the cassette), and all flanking sequences with the same insertion position and orientation were joined into single insertions. Distinct flanking sequences for the same insertion are due to 20/21 bp flanking region lengths (due to MmeI variability) and to mismatches (presumably due to PCR/sequencing errors). Annotation files from the Phytozome bulk downloads page for the *Chlamydomonas* v5.3 genome were used to get the gene and feature positions (Creinhardtii_236_gene.gff3, parsed in python using BCBio.GFF, http://github.com/chapmanb/bcbb/tree/master/gff) and gene annotation information (Creinhardtii_236_annotation_info.txt); the insertion locations were matched to genes to get the gene ID, feature (exon/intron/UTR), and orientation (sense or antisense compared to gene direction) for each insertion. Most of the work was accomplished using a custom program – the full shell command was "mutant_count_alignments.py -C -u -g Creinhardtii_236_gene.gff3 -aA Creinhardtii_236_annotation_info.txt".

(4) Insertion positions with very low abundance were removed. We believe that these low abundance reads are likely the result of sequencing or PCR errors. The genomic read count histograms in Supplemental Figure 12A,B online show a tall peak at around 1 read per insertion, and then a second peak at around 50 reads per insertion. We suspect that the second peak is the number of reads resulting from a single DNA molecule as a PCR template, so any "insertions" below that peak are likely to be due to PCR and/or sequencing errors. Therefore we removed any insertions below a custom cutoff positioned between the first two peaks (15 reads for replicate 1 and 20 reads for replicate 2, marked on Supplemental Figure 12 online). Supplemental Figures 12C,D online shows the same processing for the cassette dataset used in Figure 5: in that case there was a sharp peak around 1k–10k reads, which we think corresponds to the real insertions, and a lower peak around 10–100 reads with a sharper increase at 1 read, which we think is due to noise. For the dataset used in Figure 5, we removed all insertions with fewer than 300 reads (marked on Supplemental Figure 12 online).

This low-abundance insertion removal was done in the interactive python shell, with the command: “dataset.remove_mutants_below_readcount(min_readcount)” (where dataset is a mutant_analysis_classes.Insertional_mutant_pool_dataset instance, loaded from the .pickle file generated by mutant_count_alignments.py in the previous step).

(5) The two technical replicate datasets were merged together. This was done in the python interactive shell, using the command “dataset_a.merge_other_dataset(dataset_b)” which simply matches the insertions by location and merges their read counts and other data (dataset_a and dataset_b are loaded from the .pickle files as above). The single replicates were used to show replicate correlation (Figure 4) and reproducibility; the merged dataset was used for all other analyses.

(6) Adjacent insertion positions that probably came from a single real insertion were merged. We suspect that in some cases, a single insertion could generate two flanking sequences that map to different (adjacent) positions. The main scenarios where this could happen are (Supplemental Figure 13 online): (i) 1 bp insertions/deletions during PCR or sequencing could cause a very small proportion of the flanking regions for an insertion to align 1 bp away from the main alignment position, in the same orientation; (ii) a single insertion of two cassette copies ligated together, in opposite orientations, would cause the flanking region from both sides to be read as the 5' flanking region, resulting in two flanking regions on two sides of the same insertion position, thus mapping to the same insertion position with opposite orientations and similar read numbers; (iii) the same situation in combination with a 1 bp deletion could result in two flanking regions mapping to insertion positions 1 bp apart, with opposite orientations, facing "away" from each other. In contrast, we assume that (iv) two flanking regions that map 1 bp apart in opposite orientations facing "toward" each other (with the flanking region sequences partially overlapping) must originate from two distinct mutant strains, since if they originated from a single mutant with two adjacent cassettes, the flanking region from each cassette would overlap the other cassette rather than mapping to the genome.

In order to determine the prevalence of cases of single insertions yielding different flanking region positions or orientations, we compared the numbers of insertion pairs in the four categories. If none of the phenomena described above occur, we would expect the same number of opposite-orientation flanking region pairs in all three categories (ii = iii = iv), and the same number of same-orientation and opposite-orientation pairs 1 bp apart (i = iii+iv). Therefore, significantly different numbers of flanking sequence pairs in different categories would indicate that some of the flanking sequence pairs likely do not correspond to distinct insertions, but instead come from a single insertion, and should be merged. In the joint genomic dataset used for Figure 3, we observed 10 same-orientation pairs 1 bp apart, 88 opposite-orientation pairs in the same location, 7 opposite-orientation away-facing pairs 1 bp apart, and 9 opposite-orientation toward-facing pairs 1 bp apart. These numbers indicate the need to merge most of the opposite-orientation same-location pairs, but no need to merge the other categories. For the purpose of finding hits, we would have merged all of the opposite-orientation same-location pairs to avoid false cases of two insertions per gene; however, for the purpose of getting an overview of insertion locations, we chose to leave as many pairs as would be expected based on the other categories, to avoid bias. The pairs chosen for merging were the ones with the most similar read counts, since flanking sequences resulting from two sides of one cassette tandem insertion would be expected to give similar read counts. Merging was performed iteratively until the numbers satisfied the constraints above; after merging, the numbers were 10, 6, 5 and 8 respectively. It is worth noting that before removing low-read insertions in step 4 above, we saw several hundred same-orientation pairs 1 bp apart (i), but the vast majority was removed by applying the low-read-count cutoff in step 4, leaving a number that would be expected randomly.

This adjacent-insertion-merging process impacts relatively few insertions, but omitting it can cause two problems: inflating the number of genes with two independent insertion alleles, and distorting the distribution of gaps between adjacent insertion locations.

The merging was only done for genomic datasets: for cassette-aligned flanking regions, the reasoning behind merging opposite-orientation same-position insertions does not apply, and the insertion density is too high to meaningfully merge same-orientation 1 bp apart insertion pairs.

The merging was done in the python interactive shell, using commands “dataset.merge_adjacent_mutants(1, 'auto', None, 'by_ratio')" and "dataset.merge_opposite_tandem_mutants('auto', None, 'by_ratio')", with results checked using the "dataset.summary.adjacent_mutant_summary()" command).

The final insertion position data for all the datasets (both replicates and the joint and filtered final version of the dataset used for Figure 3, and the 5' and 3' cassette reads for the dataset used for Figure 5) are available as Supplemental Datasets 7–11 online.

Determining genome mappability

It's important to note that not all genomic insertion positions can be mapped – if an identical sequence occurs in multiple places in the genome, flanking regions derived from it will not be mapped uniquely and thus will be discarded in the alignment stage. Thus, to meaningfully compare the observed and expected insertion positions and densities, we need to account for non-unique "unmappable" regions. To do this, we determine which of all possible genomic insertion positions would result in mappable reads – this is called the "mappability" of the genome (based purely on the reference genome sequence), and corresponds to the expected density of insertions if insertion positions were purely random.

In order to determine the mappability of each genomic position or region, each 20 bp and 21 bp slice of the genome sequence was categorized as unique (that sequence occurs only once in the genome) or non-unique (the sequence occurs more than once in the genome, and thus when it's read as a flanking region, it cannot be mapped to a single locus with certainty). Out of the *Chlamydomonas* genome, (v5.3 from Phytozome, non-repeat-masked), 77% of 20 bp slices were unique, and 78% of 21 bp slices; the 20 bp and 21 bp data were very similar, as expected. These data allowed us to determine which of all possible insertion positions would result in uniquely mappable reads, and which would not, making it possible to compare the observed and expected insertion densities.

The genome mappability analysis was done in the interactive python shell using mutant_simulations.genome_mappable_slices and mutant_simulations.genome_mappable_insertion_sites functions.

Generating simulated random insertion datasets, and randomly chosen subsets

In order to compare the observed insertion density distribution to a completely random one, we generated ten simulated datasets with the same number of insertions as the real dataset, and with the location of each insertion randomly chosen out of all the mappable positions in the genome. The first three of those simulated datasets are shown in Figure 3A.

Further, in order to estimate how much of the genome would be covered by larger numbers of insertions, we used the same method to generate ten simulated datasets of 1M mappable insertions each, and determined which gene each insertion fell in. These ten datasets, and randomly chosen smaller subsets of them of sizes ranging from 100 to 1M insertions, are shown as the "simulated" data in Figure 3E; the "observed" data in the same figure is our real dataset, and 100 randomly chosen smaller subsets of it, ranging from 100 to 10k insertions.

The simulated datasets were generated in python with the mutant_simulations.simulate_dataset_from_mappability function; their gene locations were determined with the mutant_simulations.find_genes_for_simulated function; the Figure 3E plots were made with the mutant_plotting_utilities.genes_with_N_mutants function.

Locating statistically significant insertion density hotspots and coldspots

We used the binomial test with correction for multiple testing to detect regions of the genome with more or fewer insertions than would be expected if the insertion positions were random. We call regions with more insertions “hotspots”, and regions with fewer insertions “coldspots”. To make sure the analysis was thorough, we looked for hotspots/coldspots in a large range of sizes; for each size, we sliced the genome into adjacent windows of that size, using evenly spaced offsets to get 2–4 overlapping sets of windows. The sizes and offsets were: 1 kb (offsets 0 and 500bp), 5 kb (offsets of 0 and 2.5 kb), 20 kb (offsets of 0 and 10 kb), 100 kb (offsets of 0, 25, 50, 75 kb), 200kb (offsets of 0, 50, 100, 150 kb), 400 kb (offsets of 0, 100, 200, 300 kb), and 1 Mb (offsets of 0, 250, 500, 750 kb). For each region, we used the exact binomial test to determine the probability of getting the observed number of insertions given that region's mappable length and the total number of observed insertions in the genome, assuming that the insertions were uniformly randomly distributed over the mappable genome positions. The resulting p-values were corrected for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) (as implemented in R with the p.adjust function). Since each region size and offset combination uses the same genome sequence and thus isn't independent data, we performed the multiple testing correction for each combination separately (resulting in under-correction and lower adjusted p-values), rather than for all of them together (which would result in over-correction). We chose to under-correct, as this provides the most conservative statistical support of our conclusion significant hot/coldspots are rare. The analysis was done in python, using the mutant_simulations.find_hot_cold_spots function.

Treating each region size and offset separately, we counted a total of 64 hotspots and 4 coldspots with adjusted p-value 0.05 or lower. However, a single real hotspot or coldspot is likely to generate multiple statistically significant results in overlapping windows with different sizes or offsets – therefore, the results were clustered into 15 distinct groups of overlapping hotspots and 1 group of coldspots. For plotting (Figure 3A,B and Supplemental Figure 5), each overlapping cluster was reduced to a non-redundant subset of 1–3 regions with the lowest p-values (if region A completely contained B, they were considered redundant, and only the one with the lower p-value was chosen for the non-redundant subset; partially overlapping regions were kept.). Supplemental Dataset 4 shows the full list, along with positions and expected and observed insertion numbers, and which ones were plotted.

We performed the same analysis on the simulated datasets with the same number of insertions, to ensure that the method worked as expected and to check the degree of under-correction. The ten simulated datasets had 0–5 hotspots and 0–1 coldspots with an adjusted p-value 0.05 or lower (clustered into 0–2 distinct groups of overlapping hotspots or coldspots); the lowest adjusted p-value was 0.00272. For comparison, the real dataset had 41 hotspots and no coldspots with adjusted p-value 0.0027 or lower. This shows that a few of the hotspots and coldspots in the real dataset may be noise showing up as statistically significant because of under-correction for multiple testing, but the majority of the hotspots are almost definitely non-random.

Determining the exact boundaries of each hotspot/coldspot is difficult, since we only have data for groups of partially overlapping specific-size windows. To be conservative, we used the full range from the start of the first region to the end of the last region for each hotspot/coldspot cluster to determine how much of the genome was covered by the hotspots and coldspots.

Using this method, the results are: 4,027.5 kb covered by hotspots (~3.6% of the 111 Mb genome), and 1.5 Mb covered by coldspots (~1.4% of the genome); if we only include the smaller non-redundant region set for each of the 15 hotspots and 1 coldspot, the numbers are 1,846 kb for hotspots (~1.8% of the genome) and 1 Mb for coldspots (~0.9% of the genome). A similar analysis was done to determine how many insertions more than the expected number were seen in all the hotspots together, and how many fewer than the expected number in all the coldspots together. Counting the full range of overlapping regions for each hotspot/coldspots, the result was 238 “extra” insertions in hotspots (2.1% of the total 11,478 insertions), and 41 “missing” insertions in coldspots (~0.36%); counting only the non-redundant subsets, the result was 206 (~1.8%) in hotspots and 40 (~0.35%) in coldspots.

Background strain genome sequencing and analysis

Genomic DNA from our background strain was extracted in technical duplicate, using the same protocol as above for DNA extraction for individual mutants, then submitted for standard Illumina sequencing library preparation (DNA fragment size 300–400 bp) and sequenced on the Illumina HiSeq (multiplexed, paired-end, 101 bp on both ends). We obtained 20M and 18M pass-filter reads from the two replicates.

In order to mimic the process of mapping insertion flanking regions as closely as possible, we extracted the first 20 bp or 21 bp from one end of each read and aligned it to the reference *Chlamydomonas* genome using bowtie with “--trim3 80 -v1 -m1 --strata --best --tryhard” options: “--trim3 80” removes the last 80 bp, leaving the first 21 bp; the remaining options are the same as for insertion mapping, except that “-m1” is used instead of “-k10” to discard multiple alignments directly instead of doing it later in the post-processing stage. The read positions were extracted using the HTSeq python package. The process was repeated for both replicates, with 21 bp and 20 bp sequences; the differences between the four result sets were minimal, and they were added together for the final display of sequenced fragment density (Supplemental Figure 5 online).

Examining the influence of gene essentiality on recovered insertion positions

We obtained a list of 1110 essential Saccharomyces cerevisiae genes from the Database of Essential Genes (Giaever et al., 2002; Zhang and Lin, 2009). We used BLAST (Altschul et al., 1997) version 2.2.26 to align these genes against all *Chlamydomonas* protein-coding genes (Creinhardtii_169_peptide.fa file from Phytozome). We obtained a list of 711 potentially essential *Chlamydomonas* genes by taking the top hit for each yeast gene if its BLAST alignment e-value was below 1e-5. This is a rough way of detecting homologs, but it suffices for the purpose of obtaining a set enriched in essential genes.

We then calculated the total number of insertions and the total number of positions without insertions (i.e. the mappable length minus the number of insertions) for this set of genes and for all remaining *Chlamydomonas* genes, and compared them using the chi-square test of independence (the p-value was 9.38e-06). The insertion density for the set of potentially essential genes was 0.10 insertions/kb; for the remaining genes it was 0.13 insertions/kb. This shows that we recover fewer mutants in the set including more essential genes, presumably because mutants in essential genes are not viable.

The fact that we recover many insertions mapped to our set of potentially essential genes can be explained by two factors: 1) we don't expect all of those genes to be essential, since we used a rough method of finding homologs of essential yeast genes, and there is no guarantee that even true homologs of essential yeast genes are essential in *Chlamydomonas*; 2) approximately 30% of insertions are mapped incorrectly due to additional genomic DNA fragments inserted along with the cassette.

Finding JGI gene IDs for our insertions, and comparison to other datasets of interest

Since many of the datasets of interest to the scientific community still use the older JGI (Joint Genome Institute) gene/protein IDs, we obtained JGI v4 and v3 protein IDs for all our genomic insertion locations.

We obtained JGI v4 and v3 gene location files from the JGI Chlre4 download website (Merchant et al., 2007) (http://genome.jgi-psf.org/Chlre4/Chlre4.download.ftp.html). In order to obtain JGI v4 protein IDs for our insertion locations, we re-aligned our data to the v4 JGI genome sequence (same as Phytozome v4.3), using the same methods as in the v5.3 alignment, and mapped the resulting insertion locations to the JGI v4 gene location file

(Chlre4_best_genes.gff). The JGI v3 annotation uses a still older genome assembly – rather than re-align our raw reads to that assembly, we obtained JGI v3 protein IDs for our insertion locations by comparing them to the FrozenGeneCatalog_20080828_genes.gff gene location file, which gives JGI v3 genes re-mapped onto the v4 genome sequence.

The resulting JGI v3 protein IDs for our insertions were used to determine how many genes in GreenCut2 (Karpowicz et al., 2011), CiliaCut (Merchant et al., 2007) and the chloroplast proteome (Terashima et al., 2011) are present in our dataset (Figure 3*D*). Each of the datasets had a subset of protein IDs that were not present in the original gene location file we used, so we could not determine whether they were present in our dataset: that was the case for 20/597 GreenCut2 genes, 213/997 chloroplast proteome genes, and 18/195 CiliaCut genes; in addition, 55/997 chloroplast proteome genes had no JGI protein IDs at all. All these genes were excluded from the analysis.

For the flagellar proteome (Pazour et al., 2005), we compared the JGI v4 IDs of our insertions to the JGI v4 IDs from the "Mapping to JGI v4" file from the database linked in the paper (only from lines where the "Present in CrFPv2" field value was 1) (Figure 3D). The file contained 1067 IDs, which were not given direct matches to the IDs in the main database; the total number of proteins in the main database is 1,138, implying that approximately 71 genes were not included in the analysis, although the correspondence between v2 and v4 genes may not be exactly one-to-one.

For the sections of Figure 3D comparing our dataset to the set of all *Chlamydomonas* genes, genes with Arabidopsis homologs, and genes with no functional annotation, the *Chlamydomonas* genome v5.3 from Phytozome was used directly. Genes with Arabidopsis homologs are genes that have a best Arabidopsis hit name, symbol or defline listed in the Phytozome downloadable annotation file; Genes with no functional annotation are genes that lack any annotation in the Phytozome downloadable annotation or defline files.

Cassette fragmentation motif enrichment

For each possible cassette fragmentation site, we determined the read count for the resulting upstream and downstream fragmentation sites (adding together read counts originating from extraction of 5' and 3' flanking regions, since they're essentially replicates in this context). For each possible 4 bp motif (136 total after collapsing reverse-complements), we determined the number of times it occurs in the cassette, and the observed read counts for upstream and downstream fragments for each of the positions in which it occurred.

For each motif, we calculated the average enrichment by dividing the average read count for those fragments by the average read count for all observed cassette fragments. We then used the non-parametric Kruskal-Wallis test to check whether the read counts for each motif originate from a different distribution than all the cassette fragment read counts, and corrected the resulting p-values for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The motifs for which the resulting adjusted p-value was <0.05 were categorized as enriched (if average enrichment >1) or depleted (if average enrichment <1); the most highly enriched motifs were chosen based on the average enrichment, not p-value. Motifs that appeared in the cassette fewer than 5 times were ignored, since the Kruskal-Wallis test is not reliable for low values of N. The full list of motifs, average read counts, p-values, and other information is available in Supplemental Dataset 5.

Additional software used for the analysis

Several python packages were used at various stages in the analysis: NumPy (Oliphant, 2007) was used throughout the code; matplotlib (Hunter, 2007) was used for visualization; SciPy (http://www.scipy.org) and RPy2 (http://rpy.sourceforge.net) were used for statistical functions; Biopython (Cock et al., 2009) was used for sequence file parsing.

References for Supplemental Methods

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25,** 3389-3402.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) **57,** 289–300.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25,** 1422-1423.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, A.P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.D., Flaherty, P., Foury, F., Garfinkel, D.J., Gerstein, M., Gotte, D., Güldener, U., Hegemann, J.H., Hempel, S., Herman, Z., Jaramillo, D.F., Kelly, D.E., Kelly, S.L., Kötter, P., LaBonte, D., Lamb, D.C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S.L., Revuelta, J.L., Roberts, C.J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D.D., Sookhai-Mahadeo, S., Storms, R.K., Strathern, J.N., Valle, G., Voet, M., Volckaert, G., Wang, C.Y., Ward, T.R., Wilhelmy, J., Winzeler, E.A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J.D., Snyder, M., Philippsen, P., Davis, R.W., and Johnston, M. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature **418,** 387-391.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res **40,** D1178-1186.

Green, M.R., and Sambrook, J. (1989). Molecular Cloning: A Laboratory Manual (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory).

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Computing In Science & Engineering **9,** 90-95.

Karpowicz, S.J., Prochnik, S.E., Grossman, A.R., and Merchant, S.S. (2011). The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage. J Biol Chem **286,** 21427-21439.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10,** R25.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal **17,** 10-12.

Merchant, S.S., Kropat, J., Liu, B., Shaw, J., and Warakanont, J. (2012). TAG, you're it! *Chlamydomonas* as a reference organism for understanding algal triacylglycerol accumulation. Curr Opin Biotechnol **23,** 352-363.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Marechal-Drouard, L., Marshall, W.F., Qu, L.H., Nelson, D.R., Sanderfoot, A.A., Spalding, M.H., Kapitonov, V.V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., Lucas, S.M., Grimwood, J., Schmutz, J., Cardol, P., Cerutti, H., Chanfreau, G., Chen, C.L., Cognat, V., Croft, M.T., Dent, R., Dutcher, S., Fernández, E., Fukuzawa, H., González-Ballester, D., González-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P.A., Lemaire, S.D., Lobanov, A.V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J.V., Moseley, J., Napoli, C., Nedelcu, A.M., Niyogi, K., Novoselov, S.V., Paulsen, I.T., Pazour, G., Purton, S., Ral, J.P., Riaño-Pachón, D.M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S.L., Allmer, J., Balk, J., Bisova, K., Chen, C.J., Elias, M., Gendler, K., Hauser, C., Lamb, M.R., Ledford, H., Long, J.C., Minagawa, J., Page, M.D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A.M., Yang, P., Ball, S., Bowler, C., Dieckmann, C.L., Gladyshev, V.N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R.T., Brokstein, P., Dubchak, I., Goodstein, D., Hornick, L.,

Huang, Y.W., Jhaveri, J., Luo, Y., Martinez, D., Ngau, W.C., Otillar, B., Poliakov, A., Porter, A., Szajkowski, L., Werner, G., Zhou, K., Grigoriev, I.V., Rokhsar, D.S., and Grossman, A.R. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science **318,** 245-250.

Oliphant, T.E. (2007). Python for Scientific Computing. Computing In Science & Engineering **9,** 10-20.

Pazour, G.J., Agrin, N., Leszyk, J., and Witman, G.B. (2005). Proteomic analysis of a eukaryotic cilium. J Cell Biol **170,** 103-113.

Southern, E. (2006). Southern blotting. Nat Protoc **1,** 518-525.

Terashima, M., Specht, M., and Hippler, M. (2011). The chloroplast proteome: a survey from the *Chlamydomonas reinhardtii* perspective with a focus on distinctive features. Curr Genet **57,** 151-168.

Zhang, R., and Lin, Y. (2009). DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res **37,** D455-458.

High-Throughput Genotyping of Green Algal Mutants Reveals Random Distribution of Mutagenic Insertion Sites and Endonucleolytic Cleavage of Transforming DNA

Ru Zhang, Weronika Patena, Ute Armbruster, Spencer S. Gang, Sean R. Blum and Martin C. Jonikas

Plant Cell 2014;26;1398-1409; originally published online April 4, 2014;
DOI 10.1105/tpc.114.124099

Supplemental Data	**http://www.plantcell.org/content/suppl/2014/05/15/tpc.114.124099.DC1.html**
References	This article cites 41 articles, 23 of which can be accessed free at: http://www.plantcell.org/content/26/4/1398.full.html#ref-list-1
Permissions	https://www.copyright.com/ccc/openurl.do?sid=pd_hw1532298X&issn=1532298X&WT.mc_id=pd_hw1532298X
eTOCs	Sign up for eTOCs at: http://www.plantcell.org/cgi/alerts/ctmain
CiteTrack Alerts	Sign up for CiteTrack Alerts at: http://www.plantcell.org/cgi/alerts/ctmain
Subscription Information	Subscription Information for *The Plant Cell* and *Plant Physiology* is available at: http://www.aspb.org/publications/subscriptions.cfm