PNAS
www.pnas.org

Supplementary Information for

MSH1 is required for maintenance of the low mutation rates in plant mitochondrial and plastid genomes

Zhiqiang Wu, Gus Waneka, Amanda K. Broz, Connor R. King, Daniel B. Sloan

Daniel B. Sloan
Email: dbsloan@rams.colostate.edu

This PDF file includes:

Supplementary text
Figures S1 to S7
Tables S1 to S11
Legends for Dataset S1
SI References

Other supplementary materials for this manuscript include the following:

Dataset S1

www.pnas.org/cgi/doi/10.1073/pnas.2001998117

Supplementary Information Text

Estimating noise threshold of duplex sequencing with *E. coli* single-colony analysis. Duplex sequencing has been used to detect rare variants by taking advantage of a reported error rate at or below 10^{-7} per bp (1). To assess the fidelity of this sequencing method in our hands, we used 2-ml liquid cultures each derived from a single colony of *Escherichia coli*. We chose these samples as our best approximation of a negative control that should be (nearly) free of true double-stranded mutations because of the low mutation rate in *E. coli* and the relatively small number of rounds of cell division required to reach saturation in a 2-ml volume (2). As such, we used these samples to estimate the duplex-sequencing error rate (while recognizing that this estimate may be conservatively high if some variants are true *de novo* mutations rather than sequencing errors). Because the standard method to fragment DNA samples with ultrasonication is known to introduce substantial oxidative damage (3, 4), we tested this approach alongside an alternative enzymatic fragmentation strategy (New England Biolabs dsDNA Fragmentase). We also performed each fragmentation strategy either with or without subsequent treatment with multiple DNA repair enzymes to eliminate common forms of single-stranded DNA damage. The unrepaired ultrasonication libraries showed the expected signature of oxidative damage dominated by single-stranded G→T errors (5), but much of these strand-specific effects could be effectively removed by enzymatic treatment (Fig. S7). For reasons that are unclear, Fragmentase treatment produced extremely high rates of single-stranded misincorporation of As (i.e., C→A, G→A, and T→A), as well as single-stranded indels, which were insensitive to subsequent repair treatment (Fig. S7). However, because these errors were generally not matched by a complementary change on the other strand, they were successfully filtered out during generation of DCS data. The average frequencies of SNVs in DCS data were statistically indistinguishable for repaired ultrasonication and Fragmentase libraries (Table S2). We used the ultrasonication methods with enzymatic repair for all subsequent experiments in this study. For this library type, the average frequency of SNVs across three *E. coli* biological replicates was 2.1×10^{-8} per bp, and there was only a single identified indel in a total of 240 Mb of mapped DCS data, confirming the extreme accuracy of duplex sequencing (Table S2).

Supplementary Materials and Methods

***Arabidopsis* lines and growth conditions.** *Arabidopsis thaliana* Col-0 was used as the wild type line used for all analyses. Existing mutant lines (Table S3) were used as pollen donors in crosses with Col-0 to generate heterozygous F1 individuals (Fig. 2). All mutant lines were originally generated in a Col-0 background with exception of the *fpg* and *ogg1* mutants, which were in a Landsberg *erecta* (L*er*) background (6, 7). F2 plants were screened with allele-specific PCR markers (Table S7) to identify individuals that were homozygous for the mutant allele and others that were homozygous for the wild type allele. Seeds were collected from each identified F2 homozygote to produce F3 full-sib families. After cold-stratification for three days, seeds were germinated and grown on ProMix BX soil mix in a growth room under a 10-hr short-day lighting conditions to extend rosette growth prior to bolting. For each candidate gene, sets of three F3 homozygous mutant families and three F3 wild type control families were grown in parallel. After seven to nine weeks of growth, approximately 35 g of rosette tissue was harvested from each family (representing approximately 60 individuals per family) and used for mitochondrial and plastid DNA purification. Additional F3 individuals in each family were left unharvested and allowed to set seed for subsequent analysis of F4 individuals.

Mitochondrial and plastid DNA isolation. Mitochondrial DNA purification was performed as described previously (8) except that initial mitochondrial pelleting spins were done at 20,000 rcf and subsequent washing spins were performed at 25,800 rcf. Plastid DNA was isolated simultaneously from the same tissue sample, using interleaved centrifugation steps during the mitochondrial DNA extraction. Pellets containing plastids (chloroplasts) from the 1500 rcf spin in the mitochondrial extraction protocol were gently resuspended with a paintbrush in a total of 6 ml of wash buffer (0.35 M sorbitol, 50 mM Tris-HCl pH 8, 25 mM EDTA) and then loaded onto two discontinuous sucrose gradients with 9 ml of 30% sucrose solution on top of 19.5 ml of 52% sucrose solution, each containing 50 mM Tris-HCl pH 8 and 25 mM EDTA. The sucrose gradients were then centrifuged at 95,400 rcf for 1.5 hr in a JS24.38 swinging bucket rotor on a Beckman-Coulter Avanti JXN-30 centrifuge. Plastids were harvested from the interface between the 30% and 52% sucrose solutions, diluted in wash buffer, and centrifuged in a JA14.50 fixed-angle rotor at 25,800 rcf for 16 min. Plastids were washed two more times by gently resuspending pellets in wash buffer using a paintbrush and centrifuging at 25,800 rcf. All organelle isolation steps were performed at 4º C in a cold room or refrigerated centrifuge. Plastid lysis and DNA purification followed the

same protocol as described previously for mitochondrial samples (8). Tissues samples were processed in pairs, with one mutant and one wild type sample in each batch.

***E. coli* growth and DNA extraction.** A glycerol stock of the *E. coli* K12 MG1655 strain was streaked onto an LB (Luria-Bertani) agar plate and grown overnight at 37º C. Single colonies were then used to inoculate each of three 2-ml liquid LB cultures, which were grown overnight at 37 ºC on a shaker at 200 rpm. Half of each culture (approximately 4×10^9 cells) was used for DNA extraction with an Invitrogen PureLink Genomic DNA Kit, following the manufacturer's protocol for gram-negative bacteria.

Duplex sequencing library construction and Illumina sequencing. Master stocks of duplex sequencing adapters were generated following the specific quantities and protocol described by Kennedy et al. (1) with the oligos in Table S8. For each mitochondrial and plastid DNA sample, a total of 100 ng was diluted in 50 µl of $T_{10}E_{0.1}$ buffer (10 mM Tris-HCl pH 8, 0.1 mM EDTA). DNA was fragmented with the Covaris M220 Focused-Ultrasonicator in microTUBE AFA Fiber Screw-Cap tubes to a target size of approximately 250 bp, using a duty cycle of 20%, peak incident power of 50 W, 200 cycles/burst, and six bouts of shearing for 20 sec each (separated by 15 sec pauses) at a temperature of 6° C. Ultrasonication settings were adapted from the protocol of Schmitt et al. (9).

After fragmentation, 80 ng of DNA was end repaired with the NEBNext End Repair Module (New England Biolabs E6050S) for 30 min at 20° C. Samples were then cleaned with 1.6 volumes of solid phase reversible immobilization (SPRI) beads. A-tailing of eluted samples was performed in a 50 µl reaction volume, containing 5 U Klenow Fragment enzyme (New England Biolabs M0212), 1 mM dATP, and 1x NEB Buffer 2, at 37°C for 1 hr, followed by clean-up with 1.6 volumes of SPRI beads. After quantification with a Qubit dsDNA HS Kit (Invitrogen), adapter ligation was performed on 20 ng of the resulting sample in a 50 µl reaction volume, using the NEBNext Quick Ligation Module (New England Biolabs E6056S) and 1 µl of a 64-fold dilution of the duplex sequencing adapter master stock, for 15 min at 20° C. Following adapter ligation, samples were cleaned with 0.8 volumes of SPRI beads. Half of the cleaned sample was then treated with a cocktail of repair enzymes to remove single-stranded damage in a 50 µl reaction volume containing 1x NEB CutSmart Buffer, 8 U Fpg (New England Biolabs M0240), 5 U Uracil-DNA Glycosylase (New England Biolabs M0280), and 10 U Endonuclease III (New England Biolabs M0268) for 30 min at 37° C. These enzymes were chosen to remove common sources of sequencing errors in DNA, such as uracil and oxidatively damaged bases. Unlike some commercial repair cocktails, we did not include a DNA polymerase. Therefore, any damaged strands should be cleaved,

effectively removing them from the sequenceable population rather than truly repairing them. Samples were then cleaned with 1.6 volumes of SPRI beads.

Following treatment with repair enzymes, the product was quantified with a Qubit dsDNA HS Kit, and 50 pg was amplified and dual-indexed using the primers shown in Table S8 and the NEBNext Ultra II Q5 Master Mix (New England Biolabs M0544) according to the manufacturer's instructions. All libraries were amplified for 19 cycles, which yielded the necessary redundancy to generate DCS families. After amplification, libraries were processed with 1 volume of SPRI beads and eluted in 20 µl $T_{10}E_{0.1}$ buffer. Libraries were assessed with an Agilent TapeStation 2200 and High Sensitivity D1000 reagents. If adapter dimers were detected (which was only the case for the batch of 12 libraries for *polIb* mutants and matched wild type controls), they were size selected with a 2% gel on a BluePippin (Sage Science), using a specified target range of 300-700 bp.

In initial tests of duplex sequencing with *E. coli* DNA sample, the above protocol was applied either with the described repair enzyme treatment or a control treatment that did not include these enzymes. It was also performed either with ultrasonication-based fragmentation or with an alternative fragmentation protocol based on dsDNA Fragmentase (New England Biolabs M0348). These protocol variations were performed in a 2×2 factorial design with three biological replicates. For the Fragmentase approach, 400 ng of DNA was incubated for 20 min at 37° C. Fragmentation was terminated by adding 5 µl of 0.5 M EDTA to each reaction. Samples were then cleaned with 1.6 volumes of SPRI beads.

Libraries were sequenced on 2×150 bp runs on an Illumina NovaSeq 6000 platform, with a target yield of 40M read pairs per library. Libraries were generally processed and sequenced in batches of 12 corresponding to each candidate gene and its matched wild type controls (2 genomes × 2 genotypes × 3 biological replicates). The only exception was the *msh1*-CS3246 set, for which the mitochondrial and plastid libraries were generated and sequenced in separate batches of six libraries. Library construction and sequencing of the original *A. thaliana* Col-0 families and the *E. coli* samples were also done in their own batches. The raw Illumina sequencing reads have been deposited to the NCBI Sequence Read Archive under BioProjects PRJNA604834 (*E. coli*) and PRJNA604956 (*Arabidopsis*). Individual accessions for each library are provided in Tables S1 and S4.

Duplex sequencing data analysis. Raw Illumina reads from duplex sequencing libraries were processed with a custom Perl-based pipeline available at https://github.com/dbsloan/duplexseq. In the first step in the pipeline, 3′ read trimming for low quality bases (q20) and adapter sequence was performed with cutadapt v1.16 (10). The

minimum length for retaining reads after trimming was set to 75, and the error tolerance for adapter trimming was set to 0.15. BBMerge (11) was then used to join overlapping paired-end reads into a single sequence where possible, with a minimum overlap of 30 bp and a maximum of five mismatches. The random duplex sequencing tags were then extracted from the resulting trimmed and merged reads, applying a stringent filter that rejected any reads with a barcode that contained a base with a quality score below 20. Reads were also filtered if they lacked the expected TGACT linker sequence built into the duplex sequencing adapters. Reads were then collapsed into single-stranded consensus sequences (SSCS), requiring a minimum of three reads to form an SSCS family. To call a consensus base, a minimum of 80% agreement was required within an SSCS family. When complementary SSCS families were available (reflecting the two different strands of an original double-stranded DNA molecule), they were used to form a DCS family. Any disagreements between the two complementary SSCS families were left as ambiguities in the DCS read, and any DCS read with ambiguities was later filtered out from downstream analyses.

The filtered DCS data were mapped using bowtie2 v2.2.3 under default parameter settings. The *E. coli* data were mapped against the corresponding K12 MG1655 reference genome (GenBank U00096.3). We later had to exclude a called SNV at position 4,296,060 and indel at position 4,296,381 because they were shared across all three replicates and appeared to reflect fixed differences between our *E. coli* line and the reference. Likewise, in *Arabidopsis*, we found that the plastid genome in our Col-0 line had a 1-bp expansion in a homopolymer at position 28,673 relative to the published reference genome (GenBank NC_000932.1). In this case, we updated the reference genome for mapping purposes such that all reported coordinates for variants reflect a 1-bp shift at that position. For the mitochondrial genome reference, we used our recently revised version (12) of the Col-0 sequence (GenBank NC_037304.1). All *Arabidopsis* samples were mapped to a database that contained both the mitochondrial and plastid genomes to avoid cross-mapping due to related sequences shared between them because of historical intergenomic transfers (MTPTs). The resulting mapping (SAM) files were parsed to extract all SNVs and simple indels, as well as coverage data. Multi-nucleotide variants (MNVs) and more complicated structural variants were not analyzed in this pipeline.

The identified variants and associated coverage data were filtered to address known sources of errors and artefacts. First, because of the DNA fragmentation and end-repair steps involved in library construction, the positions near the ends of inserts can be prone to sequencing errors that falsely appear to be true double-stranded changes (1). Therefore, we excluded any variants and sequencing coverage associated with 10 bp at each end of a DCS read. Second, contaminating sequences can easily be mistaken for *de novo* mutations. In

the case of mitochondrial and plastid genomes, one of the most likely sources of contamination is NUMT and NUPT sequences in the nucleus (13). Therefore, we used NCBI BLASTN v2.2.29+ to map each DCS reads containing an identified variant against the TAIR10 release of the *A. thaliana* Col-0 nuclear genome. Any variants that returned a perfect match to the nucleus were excluded as presumed NUMTs or NUPTs. However, one additional challenge for the mitochondrial genome is that *A. thaliana* Col-0 harbors a recent genome-scale insertion of mitochondrial DNA into Chromosome 2, only a fraction of which is accurately captured in the published nuclear genome assembly (14). Therefore, some NUMT artefacts cannot be detected using the currently available reference nuclear genome. To address this problem, we used total-cellular shotgun DNA sequencing data from *A. thaliana* Col-0 that was generated in a different lab (NCBI SRA SRR5216995) and therefore unlikely to share inherited heteroplasmic variants with our Col-0 line. We used these raw reads to generate a database of *k*-mer counts (k = 39 bp) with KMC v3.0.0 (15), and we checked all identified variants for presence in this database. We filtered all SNVs with a count of 30 or greater in the SRR5216995 dataset, as we found that this was a reliable threshold for distinguishing known NUMTs from background sequencing errors in the mitochondrial genome. Finally, an additional complication with identifying *de novo* SNVs and indels in the mitochondrial genome is that it contains an abundance of small to medium-sized repeats that can become recombinationally active in some of the mutants analyzed in this study (16, 17) and can exhibit rare recombination even in wild type genotypes (18). When chimeric sequences resulting from recombination are mapped against a reference genome, they can give the false indication that *de novo* point mutations or indels have occurred. To eliminate these false positives, we used NCBI BLASTN to map DCS reads containing identified variants against the reference mitochondrial genome with a maximum e-value of 1e-10 to check for secondary hits (i.e., related repeat sequences) that contained the exact variant and thus could have arisen by recombination between repeat copies in the genome. SNVs that met these criteria were removed from the variant call set as likely recombinants. In the case of indels, subsequent manual curation was required for all flagged candidates to confirm variants that were consistent with recombination because of the inconsistent handling of gaps in repetitive regions by BLAST.

Identified SNVs were further characterized based on the reference genome sequence and annotation to classify their location as protein-coding, rRNA, tRNA, intronic, or intergenic. For protein-coding variants, the effect (if any) on amino acid sequence was reported. To classify indels associated with expansion or contraction of short tandem repeats, we used a standalone version of SSRIT for identification of repeat positions in the reference genomes (19).

The above steps were automated with the scripts available at https://github.com/dbsloan/duplexseq. The duplexseq_batchscripts.pl script in that repository can generate a shell script for each input library for submission to standard Slurm-based queuing systems with the same parameter settings that we applied for initial read processing and raw variant calling. Scripts are also provided to aggregate and filter the raw variant files from multiple samples run in parallel. Variant frequencies were calculated from output data by dividing the total number reads with an identified variant type by the relevant DCS coverage (expressed in total bp). Reported statistical analyses were performed in R v3.4.3, and plots were generated with the ggplot2 package.

Analysis of mitochondrial and plastid genome coverage variation in *Arabidopsis* mutants. To investigate region-specific changes in copy number in mitochondrial and plastid genomes for *Arabidopsis* mutants relative to wild type, we used the DCS reads generated above to calculate sequence coverage in terms of counts per million mapped reads as described previously (8). Counts were averaged over 500-bp windows, and means were taken across three biological replicates for both mutants and matched wild type controls. Thus, when the reported ratio of these values (Figs. 3, S1 and S3) exceeds a value of 1, it indicates a region with increased relative coverage within the genome in the mutant compared to wild type. Likewise, values below 1 indicate decreased relative coverage in mutants.

Expression and intron splicing analysis for *msh1*-SALK_046763 mutants. To test the hypothesis that *msh1* SALK_046763 mutants exhibited weaker effects on leaf variegation and mutation rates because their intronic T-DNA insert only reduces but does eliminate *MSH1* expression, we sampled F4 individuals derived from F3 homozygous SALK_046763 mutant families and from their matched F3 wild type controls (Fig. 2). Four F4 individuals were sampled from each genotype (including at least one from all three F3 families for both mutants and wild types). Approximately 60-90 mg of rosette leaf tissue was collected from each plant after approximately 8 weeks of growth under 10-hr short-day lighting conditions, flash frozen with liquid nitrogen, and immediately processed using the Qiagen RNeasy Plant Mini Kit with on-column DNase digestion. For each sample, 1 µg of RNA was reverse transcribed into cDNA using Bio-Rad iScript Reverse Transcription Supermix in a 20 µl reaction volume. Quantitative PCR (qPCR) was performed with two different *MSH1* markers – one spanning the exons that flank the T-DNA insertion in intron 8 and another in exon 16 – as well as two reference gene markers (Table S9). All primer pairs were tested with conventional endpoint PCR and gel electrophoresis to ensure amplification of a single

product of expected size. Primer pair efficiency was assessed using a dilution series (Table S9). qPCR reactions (20 µl total volume) contained 10 µl of Bio-Rad 2x iTaq SYBR Green Supermix, 10 pmole of each primer, and 1 µl of cDNA. Reactions were run on the Bio-Rad CFX96 Touch Real-Time PCR System. Thermal cycling conditions included 95º C for 3 min followed by 40 cycles of 95º C for 10 sec and 60º C for 30 sec, with a final melt curve analysis ramping from 65-95º C. Three technical replicates were run for each of the four biological replicates. In addition, a single replicate of a no-reverse-transcriptase control was run for each plant sample, and one no-template control was run for each primer set. For each cDNA sample, an average threshold cycle (C_T) value was calculated from the three technical reps. The geometric mean of the two reference genes was calculated to create a single reference C_T value for each of the eight plants for normalization (calculation of ΔC_T values). Differences in *MSH1* expression between mutant and wild type genotypes were estimated using the $\Delta\Delta C_T$ method.

To test whether properly spliced *MSH1* mRNA transcripts were produced in SALK_046763 homozygous mutants, we performed Sanger sequencing of an RT-PCR product spanning the junction of exons 8 and 9. cDNA was generated as described above for qPCR experiments. Endpoint PCR was performed using NEBNext 2x Master Mix, 0.25 µM of forward and reverse primers (MSH1 Exon 5F: 5′–CTGGTCTCAATCCTTTTGGTG–3′ and MSH1 Exon 10R: 5′–CAAACTCTCCCCAGCGGC–3′) and 1 µl of cDNA template in a 20 µl reaction volume. cDNAs were amplified from all four sampled SALK_046763 homozygous mutant plants. Thermal cycling conditions were as follows: 98º C for 30 sec; 35 cycles of 98º C for 10 sec, 60º C for 15 sec and 72º C for 20 sec; 72º C for 2 min. PCR products were visualized by gel electrophoresis to ensure the amplification of a single product of the expect size (~460 bp). For Sanger sequencing reactions, 2.5 µl of PCR product was treated with 1 µl of ExoSAP-IT (Thermo-Fisher) and incubated at 37º C for 15 min after which the enzymes were deactivated at 80º C for 15 min. Each treated PCR sample was sent to GeneWiz for Sanger sequencing after addition of 5 µl of MSH1 Exon 5F primer and 6.5 µl of dH_20. A single representative electropherogram for the junction between exons 8 and 9 is shown in Fig. S6, but all samples confirmed the presence of properly spliced products.

ddPCR heteroplasmy assays. To assess the possibility that observed heteroplasmies in duplex sequencing data could be transmitted across generations, we grew F4 seed collected from eight individuals from each of the three *msh1*-CS3246 mutant F3 families used in duplex sequencing. These F3 parents were siblings of the actual F3 individuals that were harvested for the duplex sequencing analysis. Approximately 80 mg of rosette leaf tissue was collected from F4 plants after approximately 8 weeks of growth under 10-hr short-day

lighting conditions. Collected tissue was either immediately processed or stored at -80° C until processing. Tissue samples were disrupted using the Qiagen TissueLyser, and total-cellular DNA was extracted using the Qiagen Plant DNeasy Mini Kit. DNA was quantified using a Qubit dsDNA HS Kit.

Locus-specific primers (Table S10) and allele-specific fluorescently labeled probes (Table S11) were designed for five different SNV targets (one mitochondrial and four plastid), which represented five of the most abundant variants in the *msh1*-CS3246 mutant lines based on duplex sequencing read counts (Dataset S1). Probes were designed with the target SNV in the center. Primers were tested by conventional endpoint PCR and gel electrophoresis to ensure that a single band was amplified for each primer set.

Each ddPCR reaction was set up in an initial 20 µl volume composed of 1x Bio-Rad ddPCR Supermix for Probes (no dUTP), 250 nM final concentration of each probe, 900 nM final concentration of each primer, 1 µl of the restriction enzyme BglII (Thermo Scientific FD0083), and 5 µl of diluted template DNA (5-500 pg depending on the SNV target and type of DNA, i.e., total cellular or organellar). The restriction enzyme was included for fragmentation of genomic DNA to improve ddPCR efficiency and was selected because it does not cut within any of the target amplicons. PCR emulsions were created with a Bio-Rad QX200 Droplet Generator according to manufacturer's instructions, using Bio-Rad DG8 Cartridges and QX200 Droplet Generation Oil for Probes. Amplification was performed in a Bio-Rad C1000 Touch Thermal Cycler with a deep-well block with the following program: enzyme activation at 95° C for 10 min, 40 cycles of 94° C for 30 sec and a variable annealing/extension temperature (see Table S10) for 1 min, and enzyme deactivation at 98° C for 10 min – with a ramp speed of 2° C per sec for all steps. Droplets were read on the Bio-Rad QX200 Droplet Reader and analyzed using QuantaSoft Analysis Software to calculate copy numbers of reference and alternative alleles in each sample. Channel thresholds were set based on initial experiments utilizing positive and negative controls.

MSH1 phylogenetic analysis. To assess the distribution of *MSH1* outside of green plants, we performed BLASTP searches with the *Arabidopsis* MSH1 protein sequence against the NCBI nr database using taxonomic filters to exclude Viridiplantae. We also used individual searches restricted to specific clades, including Bacteria, Archaea, Glaucophyta, Rhodophyta, and Opisthokonta. Candidates for plant-like MSH1 proteins were identified based on high amino-acid identity and near full-length hits that extended through the characteristic GIY-YIG domain. To further expand our search to include some of the vast amount of biological diversity that is unculturable and only detected in environmental samples, we queried a sample of 2000 metagenome assemblies from the JGI IMG/MER

repository (20). We also searched against the IMG/VR database, which houses the largest available collection of viral sequences from both sequenced isolates and environmental samples (21). In cases where MSH1-like sequences were identified on metagenomic scaffolds, we searched other proteins encoded in the flanking sequence against the NCBI nr database to infer possible origins for the scaffold.

Identified protein sequences were aligned with other select members of the MutS family (Table S6) using the E-INS-i algorithm in MAFFT v6.903b (22). The resulting alignments were trimmed with Gblocks v0.91b (http://molevol.cmima.csic.es/castresana/Gblocks.html) to remove low-quality alignment regions, using the following parameters: t=p; b1=18; b2=18; b3=10; b4=5; b5=h. Models of sequence evolution were assessed with ProtTest v3.4.2 (23), which identified LG+I+G+F as the preferred model based on the Akaike Information Criterion. A maximum-likelihood phylogenetic search was then performed in PhyML v3.3.20190321 (24) using this substitution model, an SPR search of tree space, 1000 random starts, and 1000 bootstrap replicates.

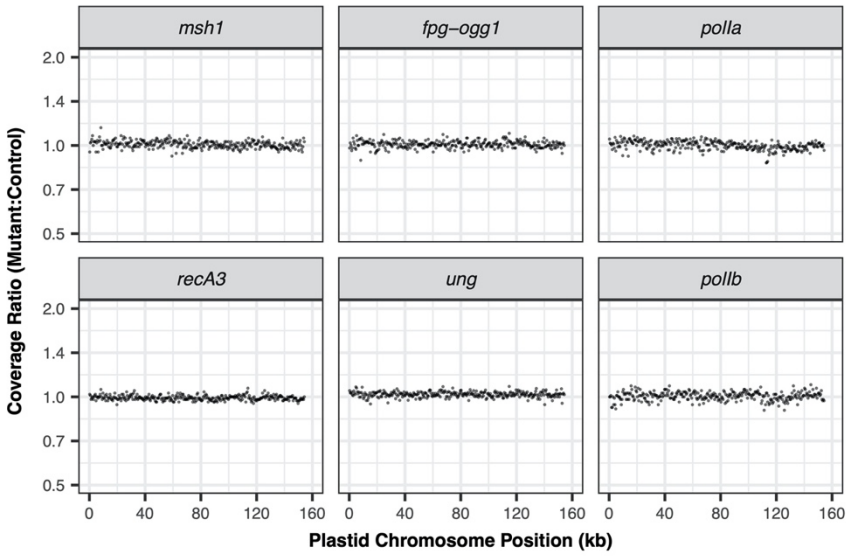

Fig. S1. Sequencing coverage variation across the plastid genome in mutants relative to their matched wild type controls. Each panel represents an average of three biological replicates, with the exception of two cases where a single outlier replicate (*ung* mutant 3 and *POLIA* wild type 3) was excluded due to what appeared to be unusually high amplification bias. The reported ratios are based on counts per million mapped reads in 500-bp windows.

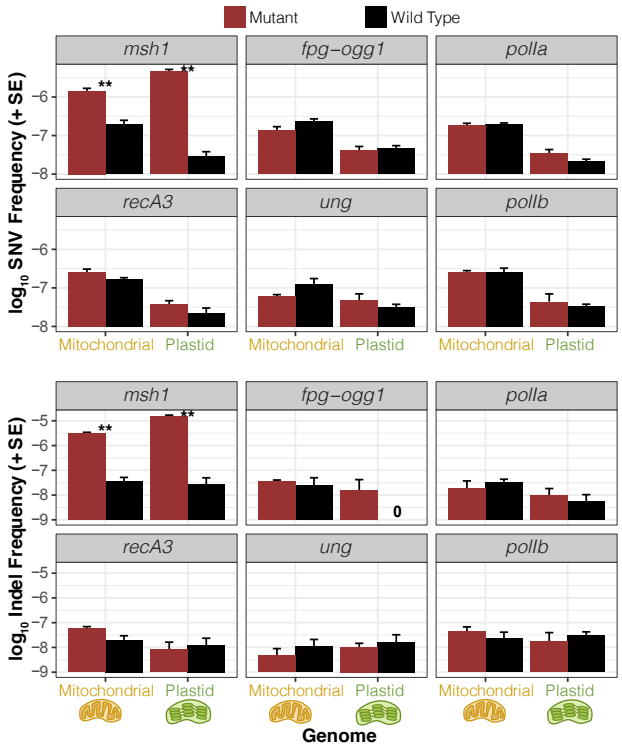

Fig. S2. The same data on SNV and indel frequencies presented in Fig. 4 but plotted on a log scale. See Fig. 4 legend for additional information.

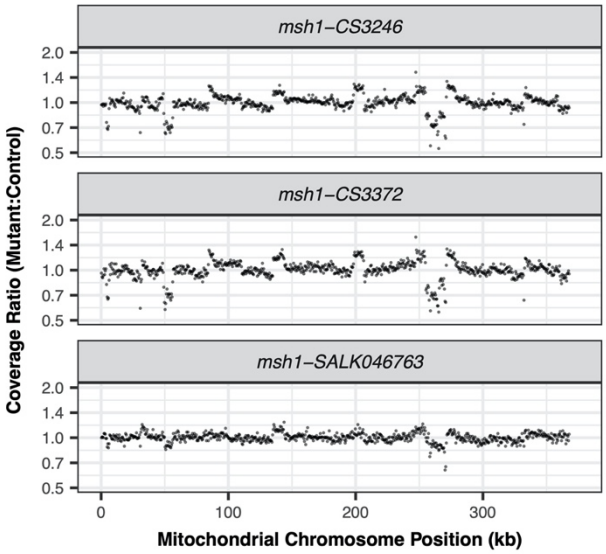

Fig. S3. Sequencing coverage variation across the mitochondrial genome in three different *msh1* mutants relative to their matched wild type controls. Each panel represents an average of three biological replicates. The reported ratios are based on counts per million mapped reads in 500-bp windows. The weaker effects of SALK_046763 likely reflect the fact that this allele has a reduced expression level but is not a full functional knockout.

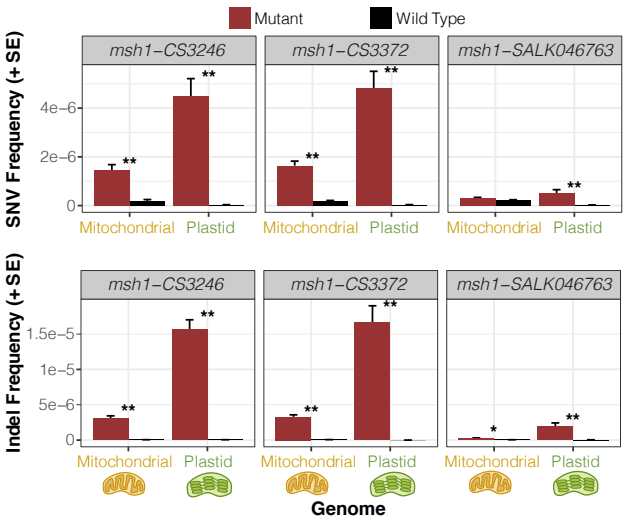

Fig. S4. Observed frequency of mitochondrial and plastid SNVs (top) and indels (bottom) based on duplex sequencing in three different *Arabidopsis msh1* mutant backgrounds compared to matched wild type controls. Variant frequencies are calculated as the total number of observed mismatches or indels in mapped duplex consensus sequences divided by the total bp of sequence coverage. Means and standard errors are based on three replicate F3 families for each genotype (see Fig. 2). Significant differences between mutant and wild type genotypes at a level of $P < 0.05$ or $P < 0.01$ (t-tests on log-transformed values) are indicated by * and **, respectively. The weaker effects of SALK_046763 likely reflect the fact that this allele has a reduced expression level but is not a full functional knockout.

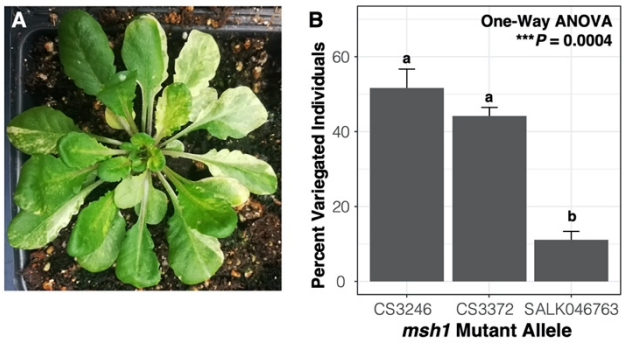

Fig. S5. Extent of leaf variegation observed for different *msh1* mutant alleles. **A.** An example of an *msh1* mutant (CS3372) individual with a leaf-variegation phenotype. **B.** Values represent the percentage of individuals in an F3 family from a homozygous mutant F2 parent that showed visible leaf variegation at time of harvest for mitochondrial and plastid DNA extraction. Means and standard errors are from three replicate F3 families from each mutant line (see Fig. 2). Between 45 and 66 individuals were scored for each family. Lowercase letters indicate significant differences between alleles based on a Tukey's HSD test. Consistent with its lower rate of observed sequence and structural variation in cytoplasmic genomes, the SALK_046763 *msh1* mutant line exhibited less severe phenotypic effects.

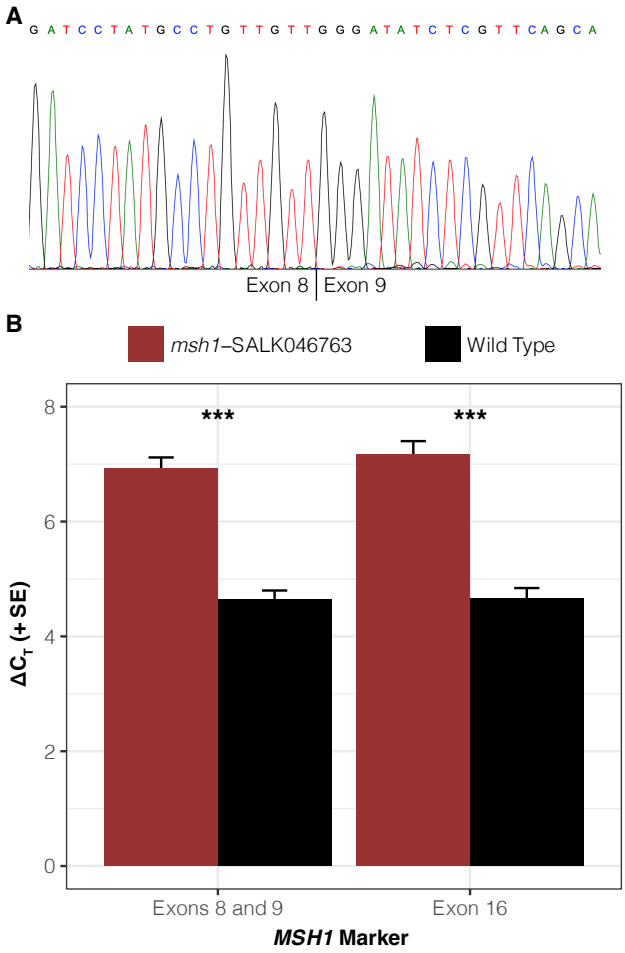

Fig. S6. Intact *MSH1* transcripts but reduced expression level in homozygous SALK_046763 *msh1* mutants. **A.** Sanger trace from cDNA sequencing confirms that properly spliced transcripts are present in SALK_046763 *msh1* mutants despite the large T-DNA insertion in intron 8. The vertical line below the trace indicates the location of the expected splice junction between exons 8 and 9. **B.** ΔC_T values are calculated based on the difference in quantitative reverse-transcriptase PCR (qRT-PCR) threshold cycle value for each indicated *MSH1* marker and the geometric mean of the threshold cycle values from two reference genes (*UBC* and *UBC9*). Means and standard errors are from four biological replicates (F4 plants derived from crossing design described in Fig. 2), each of which is based on the mean of three technical replicates. The SALK_046763 mutants exhibit higher ΔC_T (indicating lower *MSH1* expression). Both *MSH1* markers indicate a similar shift in ΔC_T values (2.3 cycles for exons 8/9 and 2.5 cycles for exon 16), corresponding to an approximately 5-fold difference in transcript abundance. Significant differences between mutant and wild type genotypes at a level of $P < 0.001$ (*t*-tests) are indicated by ***.

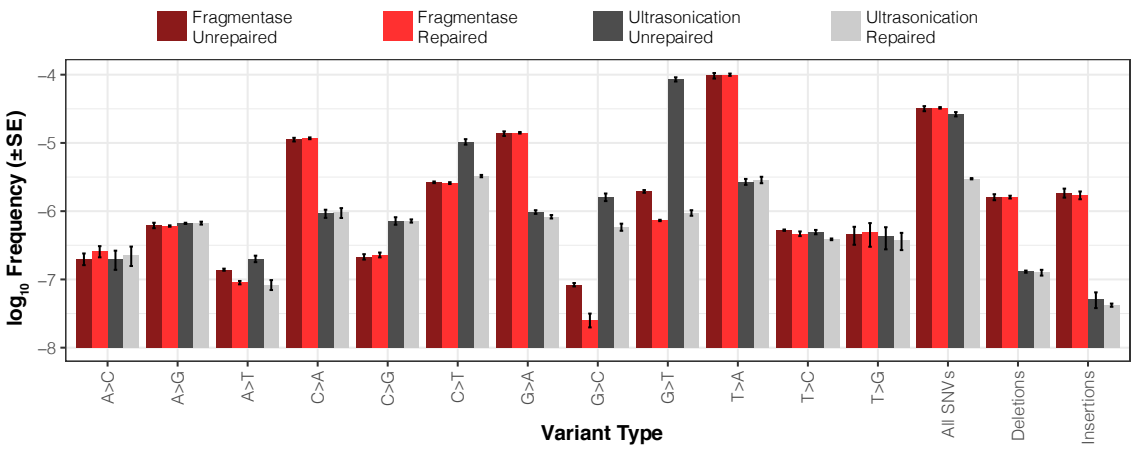

Fig. S7. Strand-specific variant frequencies from *E. coli* SSCS (not DCS) data for different library construction preparation methods (shearing by dsDNA Fragmentase or Covaris ultrasonication, either with or without subsequent enzymatic repair treatment). Values are based on means and standard errors from three biological replicates.

Table S1. *E. coli* duplex sequencing library summary

Library (SRA)	Frag. Method	Repair Treatment	Sample	Read Pairs (M)	Raw Sequence (Gb)	Mapped DCS (Mb)
SRR11018568	Fragmentase	Yes	1	30.7	9.2	74.8
SRR11018564	Fragmentase	Yes	2	35.9	10.8	79.1
SRR11018562	Fragmentase	Yes	3	27.0	8.1	75.6
SRR11018567	Fragmentase	No	1	36.4	10.9	95.9
SRR11018563	Fragmentase	No	2	39.6	11.9	97.8
SRR11018561	Fragmentase	No	3	28.7	8.6	86.2
SRR11018560	Ultrasonication	Yes	1	37.0	11.1	80.2
SRR11018558	Ultrasonication	Yes	2	27.2	8.1	68.9
SRR11018566	Ultrasonication	Yes	3	28.7	8.6	90.9
SRR11018559	Ultrasonication	No	1	32.6	9.8	98.6
SRR11018557	Ultrasonication	No	2	21.4	6.4	69.9
SRR11018565	Ultrasonication	No	3	28.9	8.7	65.5
Total				**374.1**	**112.2**	**983.2**

Table S2. Variants detected in *E. coli* duplex consensus sequence data for different library construction preparation methods (shearing by dsDNA Fragmentase or Covaris ultrasonication, either with or without enzymatic repair treatment). Three different DNA samples were collected and then subjected to each of the methods.

Method	Sample	Coverage (bp)	AT>CG	AT>GC	AT>TA	GC>AT	GC>CG	GC>TA	All SNVs	SNV Freq	Indels
Fragmentase Unrepaired	A	9.6E+07	0	0	0	0	0	0	0	0	0
	B	9.8E+07	0	0	0	0	0	1	1	1.0E-08	1
	C	8.6E+07	0	0	0	1	0	1	2	2.3E-08	0
	Total	2.8E+08	0	0	0	1	0	2	3	1.1E-08	1
Fragmentase Repaired	A	7.5E+07	0	0	0	1	0	0	1	1.3E-08	0
	B	7.9E+07	1	0	0	0	0	0	1	1.3E-08	0
	C	7.6E+07	0	0	0	0	0	0	0	0	0
	Total	2.3E+08	1	0	0	1	0	0	2	8.7E-09	0
Ultrasonication Unrepaired	A	9.9E+07	1	1	0	0	2	2	6	6.1E-08	0
	B	7.0E+07	0	0	0	0	2	2	4	5.7E-08	0
	C	6.5E+07	0	0	0	0	4	5	9	1.4E-07	0
	Total	2.3E+08	1	1	0	0	8	9	19	8.1E-08	0
Ultrasonication Repaired	A	8.0E+07	0	0	0	0	3	0	3	3.7E-08	0
	B	6.9E+07	0	1	0	0	0	0	1	1.5E-08	0
	C	9.1E+07	0	0	0	1	0	0	1	1.1E-08	1
	Total	2.4E+08	0	1	0	1	3	0	5	2.1E-08	1

Table S3. Mutant lines used for analysis of candidate genes involved in DNA replication, recombination, and repair.

Gene	AGI	Mutant Allele	Source	Ref
MSH1	At3g24320	CS3246 (chm1-2)	ABRC	(25)
MSH1	At3g24320	CS3372 (chm1-1)	ABRC	(25)
MSH1	At3g24320	SALK_046763	ABRC	(25)
RECA3	At3g10140	CS872520 (recA3-1)	ABRC	(16)
POLIA	At1g50840	SALK_022638 (polIa-1)	ABRC	(26)
POLIB	At3g20540	SALK_134274 (polIb-1)	ABRC	(26)
FPG	At1g52500	fpg	Dolores Córdoba-Cañero	(6, 7)
OGG1	At1g21710	ogg1	Dolores Córdoba-Cañero	(6, 7)
UNG	At3g18630	CS308297 (GK-440E07)	ABRC	(27)

Table S4. *Arabidopsis* duplex sequencing library summary

							Mapped DCS (Mb)	
Library (SRA)	**Gene/Line**	**Genotype**	**Purification**	**Rep**	**Read Pairs (M)**	**Raw Seq (Gb)**	**Mito**	**Plastid**
SRR11025108	Col-0	.	cpDNA	1	26.1	7.8	5.4	54.6
SRR11025107	Col-0	.	cpDNA	2	22.9	6.9	3.4	41.3
SRR11025106	Col-0	.	cpDNA	3	21.8	6.6	10.3	61.4
SRR11025105	Col-0	.	mtDNA	1	23.6	7.1	39.3	3.6
SRR11025175	Col-0	.	mtDNA	2	25.8	7.7	33.7	4.8
SRR11025174	Col-0	.	mtDNA	3	22.1	6.6	30.8	2.1
SRR11025121	fpg-ogg1	Mutant	cpDNA	1	29.8	8.9	6.4	53.4
SRR11025120	fpg-ogg1	Mutant	cpDNA	2	23.9	7.2	3.9	35.3
SRR11025119	fpg-ogg1	Mutant	cpDNA	3	30.5	9.2	5.7	61.9
SRR11025118	fpg-ogg1	Mutant	mtDNA	1	22.7	6.8	42.8	1.0
SRR11025117	fpg-ogg1	Mutant	mtDNA	2	33.7	10.1	58.9	2.8
SRR11025116	fpg-ogg1	Mutant	mtDNA	3	30.9	9.3	50.4	3.0
SRR11025114	fpg-ogg1	Wild Type	cpDNA	1	39.4	11.8	11.8	74.3
SRR11025113	fpg-ogg1	Wild Type	cpDNA	2	33.2	10.0	7.0	56.6
SRR11025112	fpg-ogg1	Wild Type	cpDNA	3	27.6	8.3	6.0	44.9
SRR11025111	fpg-ogg1	Wild Type	mtDNA	1	33.7	10.1	57.8	1.6
SRR11025110	fpg-ogg1	Wild Type	mtDNA	2	32.0	9.6	32.0	1.6
SRR11025109	fpg-ogg1	Wild Type	mtDNA	3	32.7	9.8	39.8	2.4
SRR11025100	msh1-CS3246	Mutant	cpDNA	1	42.4	12.7	28.7	60.8
SRR11025099	msh1-CS3246	Mutant	cpDNA	2	32.4	9.7	25.4	71.5
SRR11025098	msh1-CS3246	Mutant	cpDNA	3	43.9	13.2	25.9	106.7
SRR11025093	msh1-CS3246	Mutant	mtDNA	1	32.6	9.8	103.1	2.7
SRR11025092	msh1-CS3246	Mutant	mtDNA	2	44.5	13.3	125.4	4.6
SRR11025091	msh1-CS3246	Mutant	mtDNA	3	41.5	12.5	98.8	2.4
SRR11025097	msh1-CS3246	Wild Type	cpDNA	1	41.3	12.4	30.0	53.6
SRR11025095	msh1-CS3246	Wild Type	cpDNA	2	38.1	11.4	30.3	73.3
SRR11025094	msh1-CS3246	Wild Type	cpDNA	3	35.5	10.6	18.7	98.0
SRR11025090	msh1-CS3246	Wild Type	mtDNA	1	38.9	11.7	126.6	1.5
SRR11025089	msh1-CS3246	Wild Type	mtDNA	2	40.2	12.1	117.9	2.2
SRR11025088	msh1-CS3246	Wild Type	mtDNA	3	36.3	10.9	143.7	1.4
SRR11025159	msh1-CS3372	Mutant	cpDNA	1	40.2	12.1	12.4	98.4
SRR11025158	msh1-CS3372	Mutant	cpDNA	2	39.8	11.9	12.9	95.7
SRR11025157	msh1-CS3372	Mutant	cpDNA	3	36.1	10.8	17.3	77.4
SRR11025156	msh1-CS3372	Mutant	mtDNA	1	33.5	10.1	79.8	8.3
SRR11025155	msh1-CS3372	Mutant	mtDNA	2	32.8	9.8	64.4	3.2
SRR11025154	msh1-CS3372	Mutant	mtDNA	3	36.6	11.0	62.0	16.2
SRR11025152	msh1-CS3372	Wild Type	cpDNA	1	38.6	11.6	6.2	79.2
SRR11025151	msh1-CS3372	Wild Type	cpDNA	2	33.6	10.1	7.7	76.6
SRR11025104	msh1-CS3372	Wild Type	cpDNA	3	38.4	11.5	11.8	90.5
SRR11025103	msh1-CS3372	Wild Type	mtDNA	1	35.6	10.7	56.2	2.4
SRR11025102	msh1-CS3372	Wild Type	mtDNA	2	44.7	13.4	78.1	1.8
SRR11025101	msh1-CS3372	Wild Type	mtDNA	3	32.0	9.6	48.7	15.9
SRR11025087	msh1-SALK046763	Mutant	cpDNA	1	29.1	8.7	5.2	24.8
SRR11025086	msh1-SALK046763	Mutant	cpDNA	2	26.2	7.9	14.5	39.8
SRR11025078	msh1-SALK046763	Mutant	cpDNA	3	36.2	10.9	17.5	64.0
SRR11025083	msh1-SALK046763	Mutant	mtDNA	1	30.1	9.0	61.2	1.0
SRR11025082	msh1-SALK046763	Mutant	mtDNA	2	34.9	10.5	70.6	1.2
SRR11025149	msh1-SALK046763	Mutant	mtDNA	3	29.9	9.0	82.4	1.7
SRR11025080	msh1-SALK046763	Wild Type	cpDNA	1	35.0	10.5	14.1	55.8
SRR11025079	msh1-SALK046763	Wild Type	cpDNA	2	32.8	9.8	13.9	63.1
SRR11025084	msh1-SALK046763	Wild Type	cpDNA	3	27.6	8.3	9.4	28.3
SRR11025077	msh1-SALK046763	Wild Type	mtDNA	1	33.8	10.1	84.7	2.2
SRR11025150	msh1-SALK046763	Wild Type	mtDNA	2	30.7	9.2	62.8	1.1
SRR11025081	msh1-SALK046763	Wild Type	mtDNA	3	27.9	8.4	59.3	1.0

SRR11025172	polIa	Mutant	cpDNA	1	32.0	9.6	4.3	60.8
SRR11025171	polIa	Mutant	cpDNA	2	42.1	12.6	13.7	121.1
SRR11025170	polIa	Mutant	cpDNA	3	47.5	14.2	7.5	130.6
SRR11025169	polIa	Mutant	mtDNA	1	35.6	10.7	45.9	7.4
SRR11025168	polIa	Mutant	mtDNA	2	55.6	16.7	97.9	14.3
SRR11025167	polIa	Mutant	mtDNA	3	40.6	12.2	84.2	4.4
SRR11025166	polIa	Wild Type	cpDNA	1	39.1	11.7	6.0	96.8
SRR11025165	polIa	Wild Type	cpDNA	2	47.6	14.3	9.6	135.0
SRR11025163	polIa	Wild Type	cpDNA	3	41.9	12.6	14.4	105.2
SRR11025162	polIa	Wild Type	mtDNA	1	54.1	16.2	60.4	7.1
SRR11025161	polIa	Wild Type	mtDNA	2	35.4	10.6	61.8	13.3
SRR11025160	polIa	Wild Type	mtDNA	3	50.2	15.1	115.3	3.1
SRR11025178	polIb	Mutant	cpDNA	1	23.1	6.9	11.1	31.3
SRR11025177	polIb	Mutant	cpDNA	2	40.1	12.0	27.7	103.7
SRR11025164	polIb	Mutant	cpDNA	3	23.7	7.1	14.5	22.5
SRR11025153	polIb	Mutant	mtDNA	1	38.5	11.6	117.6	2.2
SRR11025096	polIb	Mutant	mtDNA	2	39.3	11.8	87.3	1.7
SRR11025085	polIb	Mutant	mtDNA	3	44.6	13.4	102.8	1.0
SRR11025148	polIb	Wild Type	cpDNA	1	21.9	6.6	14.6	28.1
SRR11025137	polIb	Wild Type	cpDNA	2	50.6	15.2	37.2	114.9
SRR11025126	polIb	Wild Type	cpDNA	3	24.2	7.3	15.4	22.8
SRR11025115	polIb	Wild Type	mtDNA	1	38.9	11.7	143.4	2.4
SRR11025176	polIb	Wild Type	mtDNA	2	43.5	13.1	105.5	1.6
SRR11025173	polIb	Wild Type	mtDNA	3	52.1	15.6	123.9	1.4
SRR11025147	recA3	Mutant	cpDNA	1	62.7	18.8	26.3	197.4
SRR11025146	recA3	Mutant	cpDNA	2	55.0	16.5	21.2	171.4
SRR11025145	recA3	Mutant	cpDNA	3	57.3	17.2	21.7	179.7
SRR11025144	recA3	Mutant	mtDNA	1	56.7	17.0	127.1	5.9
SRR11025143	recA3	Mutant	mtDNA	2	53.6	16.1	144.7	5.3
SRR11025142	recA3	Mutant	mtDNA	3	53.7	16.1	131.5	13.6
SRR11025141	recA3	Wild Type	cpDNA	1	55.2	16.6	26.5	175.2
SRR11025140	recA3	Wild Type	cpDNA	2	51.4	15.4	19.1	162.3
SRR11025139	recA3	Wild Type	cpDNA	3	64.1	19.2	34.0	184.7
SRR11025138	recA3	Wild Type	mtDNA	1	56.8	17.0	167.2	4.8
SRR11025136	recA3	Wild Type	mtDNA	2	60.7	18.2	175.2	4.7
SRR11025135	recA3	Wild Type	mtDNA	3	47.3	14.2	170.8	7.2
SRR11025134	ung	Mutant	cpDNA	1	69.3	20.8	28.5	129.0
SRR11025133	ung	Mutant	cpDNA	2	62.5	18.8	21.8	121.3
SRR11025132	ung	Mutant	cpDNA	3	55.9	16.8	16.0	109.1
SRR11025131	ung	Mutant	mtDNA	1	56.4	16.9	124.7	2.6
SRR11025130	ung	Mutant	mtDNA	2	57.3	17.2	106.8	14.2
SRR11025129	ung	Mutant	mtDNA	3	56.3	16.9	103.1	2.1
SRR11025128	ung	Wild Type	cpDNA	1	63.2	19.0	26.3	122.5
SRR11025127	ung	Wild Type	cpDNA	2	61.4	18.4	19.3	113.1
SRR11025125	ung	Wild Type	cpDNA	3	56.3	16.9	11.2	91.2
SRR11025124	ung	Wild Type	mtDNA	1	47.2	14.1	125.5	2.3
SRR11025123	ung	Wild Type	mtDNA	2	45.8	13.8	73.2	2.3
SRR11025122	ung	Wild Type	mtDNA	3	73.2	22.0	142.7	10.7
Total					**4137.3**	**1241.2**	**5459.2**	**4700.7**

Table S5. ddPCR data corresponding to results reported in Fig. 6. Samples shown in italics and labeled F4-sib are not included in Fig. 6 and represent follow-up assays performed on siblings of F4 individuals that were identified in the initial analysis as having the mutant allele at detectable frequencies. Follow-up assays were also performed for the plastid 36873 SNV on 16 additional F4 individuals not listed in this table that were from families in the M2 line (*msh1*-CS3246 mutant, biological replicate 2) other than the M2-1-21-4 family that exhibited apparent homoplasy for the mutant allele. All of these 16 additional samples exhibited mutant allele frequencies well below the noise threshold of ~0.2%

SNV	Sample Type	Sample ID	Mutant Allele Freq (%)
Mito 91017	F3 Pool	M3 mtDNA	14.3
Mito 91017	F4	M3-1-34-1-16	0.0
Mito 91017	F4	M3-1-34-2-18	66.7
Mito 91017	*F4-sib*	*M3-1-34-2-19*	*50.2*
Mito 91017	F4	M3-1-34-3-8	0.1
Mito 91017	F4	M3-1-34-4-10	0.1
Mito 91017	F4	M3-1-34-5-14	24.4
Mito 91017	*F4-sib*	*M3-1-34-5-11*	*0.1*
Mito 91017	*F4-sib*	*M3-1-34-5-12*	*5.6*
Mito 91017	F4	M3-1-34-6-1	58.7
Mito 91017	F4	M3-1-34-7-2	0.2
Mito 91017	F4	M3-1-34-8-3	0.0
Mito 91017	Control	M1-1-15-1-2	0.0
Mito 91017	Control	M1-1-15-2-1	0.0
Mito 91017	Control	M1-1-15-3-2	0.1
Mito 91017	Control	M1-1-15-4-2	0.2
Mito 91017	Control	M2-1-21-2-27	0.0
Mito 91017	Control	M2-1-21-4-15	0.2
Mito 91017	Control	M2-1-21-5-20	0.0
Mito 91017	Control	M2-1-21-6-4	0.1
Plastid 29562	F3 Pool	M1 cpDNA	2.5
Plastid 29562	F4	M1-1-15-1-2	0.0
Plastid 29562	F4	M1-1-15-2-1	0.0
Plastid 29562	F4	M1-1-15-3-2	0.0
Plastid 29562	F4	M1-1-15-4-2	0.0
Plastid 29562	F4	M1-1-15-5-2	0.0
Plastid 29562	F4	M1-1-15-6-1	0.0
Plastid 29562	F4	M1-1-15-7-1	0.0
Plastid 29562	F4	M1-1-15-8-2	0.0
Plastid 29562	Control	M2-1-21-2-27	0.0
Plastid 29562	Control	M2-1-21-4-15	0.0
Plastid 29562	Control	M2-1-21-5-20	0.0
Plastid 29562	Control	M2-1-21-6-4	0.0
Plastid 29562	Control	M3-1-34-1-16	0.0
Plastid 29562	Control	M3-1-34-3-8	0.0
Plastid 36873	F3 Pool	M2 cpDNA	3.5
Plastid 36873	F4	M2-1-21-1-22	0.0
Plastid 36873	F4	M2-1-21-2-17	0.0
Plastid 36873	F4	M2-1-21-3-12	0.0
Plastid 36873	F4	M2-1-21-4-15	100.0
Plastid 36873	*F4-sib*	*M2-1-21-4-16*	*99.9*
Plastid 36873	*F4-sib*	*M2-1-21-4-17*	*100.0*
Plastid 36873	*F4-sib*	*M2-1-21-4-18*	*100.0*
Plastid 36873	*F4-sib*	*M2-1-21-4-19*	*100.0*
Plastid 36873	F4	M2-1-21-5-20	0.0

Plastid 36873	F4	M2-1-21-6-4	0.0
Plastid 36873	F4	M2-1-21-7-6	0.0
Plastid 36873	F4	M2-1-21-8-8	0.0
Plastid 36873	Control	M1-1-15-1-2	0.0
Plastid 36873	Control	M1-1-15-2-1	0.0
Plastid 36873	Control	M3-1-21-2-18	0.0
Plastid 36873	Control	M3-1-21-4-10	0.0
Plastid 36873	Control	M3-1-21-5-14	0.0
Plastid 36873	Control	M3-1-21-6-1	0.0
Plastid 48483	F3 Pool	M3 cpDNA	2.4
Plastid 48483	F4	M3-1-34-1-16	0.0
Plastid 48483	F4	M3-1-34-2-18	0.0
Plastid 48483	F4	M3-1-34-3-8	0.1
Plastid 48483	F4	M3-1-34-4-10	0.0
Plastid 48483	F4	M3-1-34-5-15	0.0
Plastid 48483	F4	M3-1-34-6-1	0.1
Plastid 48483	F4	M3-1-34-7-2	0.1
Plastid 48483	F4	M3-1-34-8-3	0.0
Plastid 48483	Control	M1-1-15-5-2	0.0
Plastid 48483	Control	M1-1-15-6-1	0.0
Plastid 48483	Control	M1-1-15-7-1	0.1
Plastid 48483	Control	M1-1-15-8-2	0.0
Plastid 48483	Control	M2-1-21-2-27	0.0
Plastid 48483	Control	M2-1-21-4-15	0.0
Plastid 48483	Control	M2-1-21-5-20	0.0
Plastid 48483	Control	M2-1-21-6-4	0.1
Plastid 72934	F3 Pool	M3 cpDNA	1.4
Plastid 72934	F4	M3-1-34-1-16	0.0
Plastid 72934	F4	M3-1-34-2-18	0.0
Plastid 72934	F4	M3-1-34-3-8	0.0
Plastid 72934	F4	M3-1-34-4-10	0.0
Plastid 72934	F4	M3-1-34-5-14	0.0
Plastid 72934	F4	M3-1-34-6-1	0.1
Plastid 72934	F4	M3-1-34-7-2	0.0
Plastid 72934	F4	M3-1-34-8-3	0.1
Plastid 72934	Control	M1-1-15-5-2	0.1
Plastid 72934	Control	M1-1-15-6-1	0.0
Plastid 72934	Control	M1-1-15-7-1	0.1
Plastid 72934	Control	M1-1-15-8-2	0.0
Plastid 72934	Control	M2-1-21-2-27	0.1
Plastid 72934	Control	M2-1-21-4-15	0.1

Table S6. MSH1 and other MutS sequences used for phylogenetic analysis. Targeting predictions were generated with TargetP v2.0 (28).

Sequence	Clade	Source	Accession	TargetP Prediction
Arabidopsis thaliana	Viridiplantae	GenBank	AAO49798	Mitochondrial
Oryza sativa	Viridiplantae	GenBank	XP_015636674	Plastid
Selaginella moellendorffii	Viridiplantae	GenBank	XP_024525391	Mitochondrial
Physcomitrella patens	Viridiplantae	GenBank	XP_024362883	Mitochondrial
Marchantia polymorpha	Viridiplantae	GenBank	PTQ35957	None
Klebsormidium nitens	Viridiplantae	GenBank	GAQ89439	None
Chlamydomonas eustigma	Viridiplantae	GenBank	GAX83118	Plastid
Chloropicon primus	Viridiplantae	GenBank	QDZ22437	None
Coccomyxa subellipsoidea C-169	Viridiplantae	GenBank	XP_005644580	None
Ostreococcus lucimarinus	Viridiplantae	GenBank	XP_001416776	None
Putative Bathycoccaceae (Tara Oceans metagenome)	Viridiplantae	JGI IMG/MER	Ga0211588_10010061	None
Putative Chlorophyta (Trout Bog Lake metagenome)	Viridiplantae	JGI IMG/MER	Ga0164297_100118835	None
Guillardia theta	Cryptophyta	GenBank	XP_005824881	None
Chrysochromulina tobinii	Haptophyta	GenBank	KOO23129	Mitochondrial
Emiliania huxleyi CCMP 1516	Haptophyta	JGI IMG	2508101639	Mitochondrial
Babesia microti	Alveolata	GenBank	XP_021338715	Mitochondrial
Lingulodinium polyedra	Alveolata	GenBank	QDO16335	None
Vitrella brassicaformis	Alveolata	GenBank	CEL93038	None
Putative diatom (Ellis Fjord metagenome)	Stramenopile	IMG/MER	Ga0307978_10152622	None
Thalassiosira pseudonana	Stramenopile	GenBank	XP_002294874	None
Fragilariopsis cylindrus	Stramenopile	GenBank	OEU09203	None
Phaeodactylum tricornutum	Stramenopile	GenBank	XP_002177023	None
Nannochloropsis gaditana	Stramenopile	GenBank	XP_005855452	Mitochondrial
Agarilytica rhodophyticola	Gammaproteobacteria	GenBank	WP_086930438	None
Endobugula sertula	Gammaproteobacteria	GenBank	ODS23082/ODS23083	None
Putative bacterium (Damariscotta River mineral coupon metagenome)	Gammaproteobacteria	JGI IMG/MER	Ga0316202_100001462	.
Putative virus (North Sea metagenome)	Virus	JGI IMG/VR	Ga0115564_1000006915	.
Putative virus (San Pedro Channel metagenome)	Virus	JGI IMG/VR	Ga0181430_100039610	.
Putative virus (Ellis Fjord metagenome)	Virus	JGI IMG/MER	Ga0307978_10013766	.
Unknown Scaffold (Western Arctic Ocean metagenome)	Undetermined	JGI IMG/MER	Ga0302121_100079582	.
Sarcophyton glaucum	MutS7	GenBank	O63852	.
Klosneuvirus KNV1	MutS7	GenBank	ARF11760	.
Campylobacterales bacterium 16-40-21	MutS7	GenBank	OYZ35361	.
Saccharomyces cerevisiae	Fungal MSH1	GenBank	AJU17961	Mitochondrial
Escherichia coli	MutS1	GenBank	AIZ90229	.

Table S7. Primers for genotyping of mutant alleles for candidate genes.

Gene/Line	Fwd Primer	Rev Primer	Ref
MSH1-CS3246 WT	ATATTGAACTCAATTTCTTTGATTTTGGTGTTTGGT	GAAGAGTAGATGGTTTCTTACATGTCTGCAATCAC	(16)
MSH1-CS3246 Mut	ATATTGAACTCAATTTCTTTGATTTTGGTGTTTGGT	TGAAGAGTAGATGGTTTCTTACATGTCTGCAATTTT	(16)
MSH1-CS3372 WT	TTAAAAGATGGAAACCTCAACTGGGAGATGTTAC	TGTGAGTAAGCTTGAAATTCAAAAAGGATTGTGTAC	
MSH1-CS3372 Mut	TTAAAAGATGGAAACCTCAACTGGGAGATGTTAT	TGTGAGTAAGCTTGAAATTCAAAAAGGATTGTGTAC	
MSH1-SALK046763 WT	CGACAGACTTTGGATGACCT	CATACAATACCCCCTTGCTG	
MSH1-SALK046763 Mut	CGACAGACTTTGGATGACCT	ATTTTGCCGATTTCGGAAC	
RECA3 WT	GATTCCATTAGCCATGAACCGA	TCTTTCCAGATGCTTCTTTTCCG	(16)
RECA3 Mut	TAGCATCTGAATTTCATAACCAATCTCGATACAC	TCTTTCCAGATGCTTCTTTTCCG	(16)
POLIA WT	GGATCTGAAGGGAAAATCGT	CAAAACATCCCCACCTACAG	(29)
POLIA Mut	GGATCTGAAGGGAAAATCGT	ATTTTGCCGATTTCGGAAC	(29)
POLIB WT	TTACCAAAAGCATCATCCTGG	AGAGTTTTCGTGTTCCCCATC	(29)
POLIB Mut	TTACCAAAAGCATCATCCTGG	ATTTTGCCGATTTCGGAAC	(29)
UNG WT	ACTTGGAGAAGGTAAAGCAATTCA	CCATACAAAATATAATACACCACCACTC	(27)
UNG Mut	ACTTGGAGAAGGTAAAGCAATTCA	ATATTGACCATCATACTCATTGC	(27)
FPG WT	AACGAAGCAATAAAAGGCGC	CCACTCCTCTGAGTCCTTTACAGC	(6)
FPG Mut	ATTTTGCCGATTTCGGAAC	CCACTCCTCTGAGTCCTTTACAGC	(6)
OGG1 WT	GATGAAGAGACCTCGACCTAC	CTCTTCCTCAGAAACCCTAGATAA	(6)
OGG1 Mut	CGATTACCGTATTTATCCCGTTC	CTCTTCCTCAGAAACCCTAGATAA	(6)

Table S8. Oligos for duplex sequencing adapters and library amplification

Primer	Sequence[a]
i7 Library Amplification Primer	CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
i5 Library Amplification Primer	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXACACTCTTTCCCTACACGACGCTCTTCCGATC*T
Duplex Adapter Oligo 1[b]	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Duplex Adapter Oligo 2	TCTTCTACAGTCANNNNNNNNNNNNAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

[a]In primer sequences, Xs indicate conventional i5 and i7 library barcodes, Ns indicate random duplex sequencing barcodes, and * indicates a phosphorothioate bond.
[b]The Duplex Adapter Oligo 1 is shortened relative to the standard protocol (1) to facilitate i5 barcoding during library amplification.

Table S9. Primers for *msh1*-SALK046763 qPCR expression analysis and efficiency statistics from dilution series

Gene/Line	Fwd Primer	Rev Primer	R^2	Efficiency	Ref
MSH1 exons 8/9	GCATGCACATCCAGGAAGTC	GAGCTTGGTAACTAAGGCTTC	0.979	1.35	
MSH1 exon 16	GGGCGTCTGATACAATTGGTG	GCTAAAGATAAAGCCTCAGCTG	0.995	0.85	
UBC9 (At4g27960)	TCACAATTTCCAAGGTGCTGC	TCATCTGGGTTTGGATCCGT	0.993	0.95	(30)
UBC (At5g25760)	CTGCGACTCAGGGAATCTTCTAA	TTGTGCCATTGAATTGAACCC	0.996	0.91	(30)

Table S10. Primers for ddPCR heteroplasmy assays

SNV Target	Fwd Primer	Rev Primer	Annealing Temp (° C)
Plastid 29562	TCTTTCCTTGGTTGAATTCGA	GAGATACTGTATGGGGTTTCC	57
Plastid 36873	AATAATTGAAGGAGCCCCTC	ACAAGATCAAGCTGGTAAGG	57
Plastid 48483	AGGAAAGGTTAAATGAGTTCCG	ACTGGGAATGAATAAATAAGATCGG	57
Plastid 72934	CTTTCAGGAGTGGCTTGCTTCG	TTTGTATAGACGTTGAGGCGGACG	65
Mito 91017	CGTCATCGTCTCAACTACC	CAAAGACGACATCCTGAGG	57

Table S11. Allele-specific probes for ddPCR heteroplasmy assays (synthesized by Integrated DNA Technologies [IDT])

SNV Target	Reference Probe[a]*	Alternative Probe[b]*
Plastid 29562	/56-FAM/CCTATT+CC+A+T+TTT+C+CT/3IABkFQ/	/5HEX/CC+T+ATT+CC+A+C+TTTCC/3IABkFQ/
Plastid 36873	/56-FAM/C+CAT+T+CT+G+T+CTAAATAG/3IABkFQ/	/5HEX/C+CATTCT+G+C+CTA+A+ATA/3IABkFQ/
Plastid 48483	/56-FAM/TT+GCC+C+T+TCAA+C+TAT/3IABkFQ/	/5HEX/TTGCC+C+C+TCAAC+TAT/3IABkFQ/
Plastid 72934	/56-FAM/CAA+A+ACC+C+T+CCACG+CC/3IABkFQ/	/5HEX/CAAA+ACC+C+C+CCACGC/3IABkFQ/
Mito 91017	/56-FAM/TCAACTAG+A+A+TT+C+C+CTT/3IABkFQ/	/5HEX/CAA+CTAG+A+G+TTC+CCT/3IABkFQ/

[a]Each probe for the reference allele carries a 5′ 6-FAM fluorescent modification and a 3′ Iowa Black fluorescent quencher.
[b]Each probe for the alternative allele carries a 5′ HEX fluorescent modification and a 3′ Iowa Black fluorescent quencher.
*Bases preceded by a + symbol indicate locked nucleic acids (LNA).

Dataset S1 (separate file). Detailed summary information for each variant call. Coordinates are based on the *A. thaliana* Col-0 mitochondrial reference genome (NC_037304.1) and a modified version of the plastid reference genome (NC_000932.1) that includes a 1-bp insertion at position 28,673. (DatasetS1.xlsx)

SI References

1. S. R. Kennedy *et al.*, Detecting ultralow-frequency mutations by Duplex Sequencing. *Nature Protocols* **9**, 2586-2606 (2014).
2. J. Jee *et al.*, Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature* **534**, 693-696 (2016).
3. L. Chen, P. Liu, T. C. Evans, L. M. Ettwiller, DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **355**, 752-756 (2017).
4. M. Costello *et al.*, Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research* **41**, e67 (2013).
5. B. Arbeithuber, K. D. Makova, I. Tiemann-Boege, Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications. *DNA Research* **23**, 547-559 (2016).
6. T. M. Murphy, What is base excision repair good for?: knockout mutants for FPG and OGG glycosylase genes in Arabidopsis. *Physiologia Plantarum* **123**, 227-232 (2005).
7. D. Córdoba-Cañero, T. Roldán-Arjona, R. R. Ariza, Arabidopsis ZDP DNA 3′-phosphatase and ARP endonuclease function in 8-oxoG repair initiated by FPG and OGG 1 DNA glycosylases. *Plant Journal* **79**, 824-834 (2014).
8. Z. Wu, G. Waneka, D. B. Sloan, The tempo and mode of angiosperm mitochondrial genome divergence inferred from intraspecific variation in Arabidopsis thaliana. *G3: Genes, Genomes, Genetics* **In Press** (2020).
9. M. W. Schmitt *et al.*, Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 14508-14513 (2012).
10. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).
11. B. Bushnell, J. Rood, E. Singer, BBMerge–accurate paired shotgun read merging via overlap. *PloS One* **12**, e0185056 (2017).
12. D. B. Sloan, Z. Wu, J. Sharbrough, Correction of persistent errors in Arabidopsis reference mitochondrial genomes. *Plant Cell* **30**, 525-527 (2018).
13. E. Hazkani-Covo, R. M. Zeller, W. Martin, Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genetics* **6**, e1000834 (2010).
14. R. M. Stupar *et al.*, Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: implication of potential sequencing errors caused by large-unit repeats. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5099-5103 (2001).
15. M. Kokot, M. Długosz, S. Deorowicz, KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759-2761 (2017).
16. V. Shedge, M. Arrieta-Montiel, A. C. Christensen, S. A. Mackenzie, Plant mitochondrial recombination surveillance requires unusual RecA and MutS homologs. *Plant Cell* **19**, 1251-1264 (2007).
17. M. Miller-Messmer *et al.*, RecA-dependent DNA repair results in increased heteroplasmy of the Arabidopsis mitochondrial genome. *Plant Physiology* **159**, 211-226 (2012).
18. M. P. Arrieta-Montiel, V. Shedge, J. Davila, A. C. Christensen, S. A. Mackenzie, Diversity of the Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity. *Genetics* **183**, 1261-1268 (2009).
19. S. Temnykh *et al.*, Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Research* **11**, 1441-1452 (2001).
20. I.-M. A. Chen *et al.*, IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research* **47**, D666-D677 (2019).

21. D. Paez-Espino *et al.*, IMG/VR v. 2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Research* **47**, D678-D686 (2019).
22. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780 (2013).
23. F. Abascal, R. Zardoya, D. Posada, ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-2105 (2005).
24. S. Guindon *et al.*, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**, 307-321 (2010).
25. R. V. Abdelnoor *et al.*, Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene homologous to MutS. *Proceedings of the National Academy of Sciences* **100**, 5968-5973 (2003).
26. J. S. Parent, E. Lepage, N. Brisson, Divergent roles for the two PolI-like organelle DNA polymerases of Arabidopsis. *Plant Physiology* **156**, 254-262 (2011).
27. D. Córdoba-Cañero, E. Dubois, R. R. Ariza, M. P. Doutriaux, T. Roldan-Arjona, Arabidopsis uracil DNA glycosylase (UNG) is required for base excision repair of uracil and increases plant sensitivity to 5-fluorouracil. *Journal of Biological Chemistry* **285**, 7475-7483 (2010).
28. J. J. A. Armenteros *et al.*, Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance* **2**, e201900429 (2019).
29. J. D. Cupp, B. L. Nielsen, Arabidopsis thaliana organellar DNA polymerase IB mutants exhibit reduced mtDNA levels with a decrease in mitochondrial area density. *Physiologia Plantarum* **149**, 91-103 (2013).
30. T. Czechowski, M. Stitt, T. Altmann, M. K. Udvardi, W.-R. Scheible, Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiology* **139**, 5-17 (2005).