A detailed map of mitochondrial repeat-mediated recombination in Arabidopsis *msh1* mutants using long-read sequencing

Yi Zou[1], Weidong Zhu[1], Daniel B. Sloan[2], and Zhiqiang Wu[1,*]

[1]Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518120, China, and

[2]Department of Biology, Colorado State University, Fort Collins, CO 80523

* For Correspondence (e-mail: wuzhiqiang@caas.cn; tel.: +86 135-3040-6763).

Summary

The abundant repeats in plant mitochondrial genomes can cause rapid genome rearrangements and are also a major obstacle in short-read sequencing studies. Nuclear-encoded proteins such as MSH1 are known to suppress the generation of repeat-associated mitochondrial genome variants, but our understanding of these mechanisms has been constrained by the limitations of short-read technologies. Here, we used highly accurate long-read sequencing (PacBio HiFi) to characterize mitochondrial and plastid genome variants in *Arabidopsis thaliana msh1* mutant individuals. The HiFi reads successfully quantified parental and crossover recombination products of both large and small repeats, providing a global view of recombination dynamics. We found that recombination breakpoints were distributed relatively evenly across the length of repeated sequences and observed extensive non-crossover recombination (gene conversion) between pairs of imperfect repeats in the mitochondrial genome of *msh1* mutants. Comparison of long-read assemblies of seven other *Arabidopsis thaliana* accessions showed that mitochondrial genome structural divergence between accessions paralleled the repeat-mediated dynamics observed in *msh1* mutants. The *Arabidopsis* plastid genome generally lacks small repeats and exhibited a very different pattern of variant accumulation in *msh1* mutants compared with the mitochondrial genome. Our data illustrate the power of HiFi technology in studying repeat-mediated recombination in plant mitochondrial genomes and improved the sequence resolution for recombinational processes suppressed by MSH1.

Keywords: *Arabidopsis thaliana*, mitochondrial genome, repeat recombination, gene conversion, long-read

Significance

Plant mitochondrial genomes undergo rapid rearrangements while maintaining protein genes. Long-read sequencing provides a detailed and quantitative view of mitochondrial repeat-mediated recombination suppressed by MSH1, advancing our understanding of plant mitochondrial genome dynamics.

Introduction

Mitochondria and plastids are major energy converters as well as central metabolic compartments of plant cells. Plant mitochondrial (mt) and plastid (pt) genomes can also be the target of genetic engineering to alter phenotypes that are responsive to environmental change (Flood et al., 2020). These two organelle genomes have followed quite different evolutionary trajectories. Compared

with the rather conserved pt genomes (Lee et al., 2016), mt genomes exhibit dramatic variation among plant species (Wu et al., 2020a). Although coding sequences in angiosperm mt genome have very low substitution rates, their gene order and intergenic sequences are rarely conserved (Wolfe et al., 1987; Palmer and Herbon 1988; Christensen, 2021). Angiosperm mt genomes typically contain numerous non-tandem repeats, and recombination between different copies of repeat sequences causes a high level of structural instability and variation (Wynn and Christensen, 2019). These repeats are often grouped based on size ranges with some functional considerations. Large repeats (≥1000 bp) are usually identical in sequence and recombine at high frequency, leading to isomerization of the genome (Unseld et al., 1997). The recombination frequencies of intermediate-sized repeats (50-1000 bp) are lower but sensitive to genetic, environmental and developmental triggers that make them frequent enough to have important phenotypic effects (Gualberto and Newton 2017; Mackenzie and Kundariya, 2019). Shorter repeats (<50 bp) typically do not show recombination activity. These short sequences are often described as "microhomologies", but the very shortest of them (e.g., <10 bp) are unlikely to reflect true homology, because such similarities can easily occur at random. Despite their importance and critical roles, repeats in mt genome assemblies are not always annotated correctly or completely even in model plant species (Wynn and Christensen, 2019). Moreover, these repeats may produce complex assembly graphs, which hinder a clear characterization of mt genomes (Masutani et al., 2021). Thus, mechanisms underlining the repeats and their roles in destabilizing the mt genomes require a more thorough exploration.

Recombination in plant mt genome is likely involved in repairing DNA damage, such as double-strand breaks (DSBs), which can disrupt genome integrity and prevent the progression of replication and transcription activities. To deal with DSBs, two main branches of repair systems exist, which rely on different sets of genes: homology-directed repair (HDR) and non-homologous end joining (NHEJ) (Hussmann et al., 2021). HDR requires homologous dsDNA templates for new strand synthesis. HDR involving both DSB ends will lead to non-crossover repair products, which is a process known as gene conversion. In contrast, HDR involving only one DSB end and non-allelic templates is known as break-induced repair (BIR) and leads to crossover products (Christensen, 2013).

Nuclear genes encode a set of mitochondrial- and plastid-targeted DNA binding proteins, which are actively involved in processes such as DNA recombination, replication, and repair (RRR; Gualberto and Newton, 2017; Chevignv et al., 2020; Fuchs et al., 2020). Some RRR genes, such as *RECA*, *POLIA/B*, *SSB*, have conserved roles, while others are evolutionary innovations. Unlike bacterial MutS counterparts, the dual-targeted (mitochondrial and plastid) MSH1 enzyme in plants

102 combines a mismatch-recognition domain (MutS I) and a predicted DSB-generating GIY-YIG
103 endonuclease domain in a single protein. MSH1 has been hypothesized to participate in novel DNA
104 repair mechanisms in plant organelles (Christensen, 2013; Ayala-Garcia et al., 2018; Fukui et al,
105 2018; Wu et al., 2020b; Wynn et al., 2020).

106 Arabidopsis *msh1* (=*chm*) mutants exhibited a series of phenotypes and were originally isolated
107 based on green-white leaf variegation phenotype (Rédei, 1973). Positional cloning identified this
108 locus to encode a plant MutS homolog with dual localization (Abdelnoor et al., 2003). Based on
109 DNA gel blots of dissected white tissues from hemi-complementation plants, leaf variegation was
110 attributed to pt genome rearrangements (Xu et al., 2011). The plastid hemi-complementation plants
111 of *msh1* exhibited curly or wrinkled leaf and slower seed germination, which were thought to be
112 caused by mitochondrial genetic effects (Virdi et al., 2015, 2016). More recently, heritable
113 epigenetic effects in the nucleus were shown to cause diverse growth-related phenotypes of some
114 *msh1* individuals (Virdi et al., 2015; Yang et al., 2020; Kundariya et al., 2020). Disruption of *msh1*
115 results in ectopic recombination of intermediate-sized repeats and sub-stoichiometric shifting of
116 different genomic segments (Abdelnoor et al., 2003; Shedge et al., 2007; Arrieta-Montiel et al.,
117 2009; Davila et al; 2011). Similar results have been observed in moss organelle genomes, implying
118 that these are anciently conserved functions of *MSH1* in land plants (Odahara et al., 2017). It was
119 also reported that lower *MSH1* expression caused mitochondrial rearrangements that produced
120 cytoplasmic male sterile (CMS) lines in tomato and allowed fertility reversion of lines of *Brassica*
121 *juncea* (Sandhu et al., 2007; Zhao et al., 2016, 2021). These studies demonstrated that MSH1
122 maintains the structural integrity of the mt genome through suppressing illegitimate recombination
123 at repeats. We recently showed increases in de novo point mutations in organelle genomes of *msh1*
124 mutants by duplex sequencing and droplet digital PCR, suggesting a role of MSH1 in maintaining
125 the lower organellar synonymous substitution rates compared with the nuclear genome (Wu et al.,
126 2020b).

127 Despite this extensive work on *MSH1* in Arabidopsis, we still lack a complete understanding of
128 its effect on repeat dynamics, in part due to the technical difficulties of traditional DNA gel blots,
129 quantitative PCR, and short-read sequencing. For example, a previous study used early-generation
130 Illumina technology with 35-bp reads to detect variants in the Arabidopsis *msh1* mutants (Davila
131 et al. 2011). This work was successful in detecting repeat-mediated rearrangements, but the short
132 reads available at the time lack resolution for analyzing repeats at a detailed sequence level across
133 their entire lengths. The advent of long-read sequencing technologies has created the opportunity
134 to analyze structural variants in plant mt genome assemblies (Kozik et al., 2019; Yu et al., 2022).

135 In this work, we use PacBio high-fidelity (HiFi) sequencing to comprehensively investigate
136 mitochondrial repeat-mediated recombination in individual Arabidopsis *msh1* plants. These highly
137 accurate long reads are enough to span the entire length of repeats, including the two largest repeats
138 in the Arabidopsis mt genome (6.5 kb and 4.2 kb), enabling unambiguous resolution of structural
139 rearrangements and gene conversion events. We first developed a pipeline to characterize different
140 types of rearrangements based on blastn alignments of HiFi reads to the reference genome. We
141 were able to identify rearrangements mediated by 99 repeat groups (≥ 50 bp), of which 53 groups
142 were not shown recombination-active previously. Notably, we observed frequent gene conversion
143 between a proportion of non-identical repeats in *msh1*, an effect that can be difficult to detect and
144 quantify with earlier short-read sequencing approaches. We found that a dynamic region containing
145 *atp6-2* was consistently lost in mt genomes of 7 other *Arabidopsis thaliana* accessions and loss of
146 *atp6-1* was correlated with phenotypic variation among *msh1* individuals. We also detected novel
147 pt genome variants with HiFi reads. These results advance our understanding of plant mt genome
148 evolution and lay the foundation for engineering mitotype-associated plant phenotypes in the future.
149

150 **Results**

151 **Mapping of HiFi reads identified structural variations in Arabidopsis organelle genomes**

152 We generated F_3 families from a cross between wild type Arabidopsis Columbia-0 (Col-0) and an
153 *msh1* mutant line (*chm1-1*, CS3372) in an earlier work (Wu et al., 2020b). To analyze structural
154 rearrangements in the organelle genomes in homozygous *msh1* mutants, we collected aboveground
155 tissue from 11 individual plants from 3 different F_3 families (Table S1) and prepared separate DNA
156 libraries for each plant for PacBio HiFi sequencing. The libraries generated a total of 34.36 Gb of
157 HiFi reads (12.7-18.1 kb mean read length). These reads were first mapped to the Col-0 reference
158 genome by minimap2 (Li, 2018) to separate nuclear, pt and mt genome-mapped proportions. After
159 excluding one outlier with very low yield, the coverage of the nuclear genome ranged from 8.5× to
160 46.4×, the coverage of the pt genome ranged from 899.0× to 3093.8×, and the coverage of the mt
161 genome ranged from 30.9× to 247.9× (Table S2). The plastid (4.96%-12.85%) and mitochondrial
162 (0.93%-2.13%) proportions were used for downstream analysis (Figure 1a; Table S3 and S4).
163 The published mt genome is usually represented as a linear sequence of the "master circle".
164 However, various sub-stoichiometric isoforms exist due to frequent repeat-mediated structural
165 rearrangements (Kozik et al., 2019). Potential structural rearrangements could be indicated by reads
166 whose parts were mapped to discontinuous positions in the reference genome (Figure 1b). We
167 developed a pipeline to detect mitochondrial structural rearrangements (available at
168 https://github.com/zouyinstein/hifisr). In this pipeline, mt genome-mapped reads were aligned

169 against the reference sequence (NC_037304.1; Sloan et al., 2018) via the NCBI blastn software.
170 Raw blastn results were then processed by custom Python scripts, which can group the reads by
171 discontinuity, homologous sequence size and mapping coordinates, and calculate read counts for
172 each group (Table S5). After additional manual checking and curation, we were able to distinguish
173 zero-rearrangement events and one-, two-, three-, and four-rearrangement events mediated by
174 repeats (≥50 bp) in a single read. A few low-frequency events were also detected, including
175 microhomology-mediated end joining (MMEJ; defined as occurring at repeats between 2 and 49
176 bp in length) and non-homologous end-joining (NHEJ; defined as occurring at breakpoints sharing
177 0 or only 1 bp of sequence identity), replication slippage associated with tandem repeats, and
178 insertions (Figure 1c; Table S7). More than 90% of the identified mt genome rearrangements were
179 crossover recombination (BIR) mediated by non-tandem repeats. In wild type, 7.50% of total reads
180 accounted for rearrangements mediated by non-tandem repeats, most of which were associated with
181 the two large repeat pairs that undergo high-frequency recombination. In *msh1* mutants, this ratio
182 increased to 31.63% of total reads, with an increased frequency of rearrangement events mediated
183 by intermediate-sized repeats (Figure 1c,d; Table S9). Much lower frequencies of structural
184 variants were found in the pt genome for both wild-type and *msh1* individuals, but we still identified
185 rearrangements mediated by the larger inverted repeats (IRa and IRb), MMEJ, NHEJ, replication
186 slippage associated with tandem repeats, and insertions (Figure 1e,f; Table S6, S8 and S10). In
187 contrast to the mt genome, the Arabidopsis pt genome contains very few intermediate repeats
188 (Table S11). Only one read supported rearrangement-mediated by a pair of 123-bp imperfect
189 repeats in pt genome. And because of the large size (~26 kb), rearrangements mediated by the
190 IRa/b repeats will be underestimated because few HiFi reads span an entire repeat copy. The
191 proportions of MMEJ were low in mt and pt genomes, but both were increased in *msh1* mutant
192 compared with wild-type (Figure 1c,e). Taken together, these results demonstrate the power of
193 long-read HiFi sequencing for characterizing the increased repeat-mediated structural variants in
194 the mt genome of *msh1* mutants.
195
196 **Mitochondrial repeat annotations and recombination activity**
197 To further check whether the identified rearrangement events were indeed caused by the repeated
198 sequence elements of the mt genome, we first visualized the global patterns of the junctions and
199 relative read counts of each group of rearrangements, including both one-rearrangement events and
200 two-rearrangement events in a single read (Figure 2a-d; Table S12 and S13). These rearrangements
201 caused multiple highly dynamic regions across the mt genome, which were most severe at the
202 positions of large repeats and several repeat-dense regions (Figure 2e). However, we found that

203 more than half of these rearrangement junctions were not associated with the subset of repeats that
204 were named and annotated in previous studies (note that annotations of repeats FF, WW, XX, YY,
205 ZZ among the mentioned repeats A-Z, AA-ZZ are missing in GenBank file NC_037304.1; Arrieta-
206 Montiel et al., 2009; Davila et al., 2011; Sloan et al., 2018). It was suggested in a recent study that
207 identifying non-tandem repeats by blastn is incomplete at default parameters and can be
208 inconsistent at repeat boundaries, especially when mismatches and gaps are abundant (Wynn and
209 Christensen, 2019). Thus, we decided to perform a blastn search with the lowest word size using
210 the mitochondrial reference sequence as both query and subject, which generated more candidate
211 coordinates of repeat boundaries for manual comparison with the junction coordinates of one-
212 rearrangement HiFi reads. In total, we found 71 groups of two-copy repeats (R2), 23 groups of
213 three-copy repeats (R3), 5 groups of four-copy repeats (R4), which accounted for most of the read
214 rearrangement junctions in *msh1* mt genome (Figure S1a; Table S14, S15, S16 and S17). These 99
215 repeat groups contained all previously reported active repeats except for repeat TT and 53
216 additional groups of recombinationally active repeats (≥ 50 bp; named HNR1 to HNR53;
217 annotations of repeats FF, WW, XX, YY, ZZ that are missing in NC_037304.1 were not
218 considered). The boundaries of some low-identity repeats in our annotation differed from previous
219 annotations. For example, the direct repeat pair MMJS-1 (767 bp) and MMJS-2 (692 bp) were
220 previously annotated as separate smaller repeats MM, J, and S. We treated short repeats as part of
221 the microhomology category (2 to 49 bp), because the blastn software can often generate
222 overlapping, inaccurate, or incomplete predictions of homology below 50 bp in size. Moreover, the
223 orientation of a repeat pair determines how they can reorganize the genome. In a hypothetical circle
224 containing only one repeat pair, crossover causes the interconversion between a large circle and
225 two smaller circles if they are in direct orientation, while causing the interconversion between two
226 isomeric circles of the same length if the repeat pair is inverted (Figure S2a). But in real mt genome
227 with multiple repeat pairs and more complex *in vivo* structures, the orientation of a given repeat
228 pair can change depending on recombination events between other repeats. Thus, the current
229 annotations of direct vs. inverted repeat orientations are arbitrary depending on the linear
230 representation of the reference "master circle" (Table S15, S16 and S17).
231 The relative recombination frequency of different repeats can be estimated by normalized read
232 counts of one-rearrangement events. Thus, we further grouped these repeats into high-,
233 intermediate-, and low-activity categories, with 18 groups showing high activity (read count ≥ 50
234 in *msh1*, normalized to 10000 total mt genome-mapped reads), 17 groups showing intermediate
235 activity (read count of 10 to 49 in *msh1*) and 64 remaining groups showing low activity (Figure
236 S1b; Table S18, S19 and S20). Previous studies have suggested a positive correlation between

237 homology length and recombination frequency of different repeat pairs (Davila et al., 2011). To
238 investigate this idea, we considered all two-copy repeats with high sequence identity (mismatch +
239 gap ≤ 2) to exclude the effects of reduced sequence identity or copy number. As indicated by HiFi
240 read counts, there was indeed a positive relationship between homology length and recombination
241 frequency (Figure S1c). Some repeats fell far from the regression line, suggesting that additional
242 factors contribute to recombination frequency, which might explain the different recombination
243 frequencies for different pairs within the same three-copy repeat group (Table S19). Asymmetrical
244 recombination or simply preferential accumulation of only one of the crossover products of
245 intermediate-sized repeats in *msh1* mutant (Arrieta-Montiel et al., 2009, Davila et al., 2011).
246 Similarly, HiFi reads could also show asymmetrical recombination with conserved polarity but
247 different intensities among samples and accompanying loss of parental forms for some repeat
248 groups (Figure S2b; Table S21). Notably, owing to the long read length, HiFi reads also indicated
249 asymmetrical recombination meditated by the two large repeat pairs, which was more pronounced
250 in *msh1* mutant than in wild-type (Figure S2b). HiFi reads also identified recombination mediated
251 by intermediate-sized repeats Q, V, B, F, and a few other repeats at very low frequency in wild-
252 type (Table S21), which were not detected by less sensitive gel blot experiments previously.
253 Together, with HiFi reads we confirmed mt genome rearrangements in *msh1* were mostly meditated
254 by repeated sequence elements of various sizes and activities, and that these rearrangements could
255 provide additional clues for manual curations of repeat annotations.
256
257 **Base-level understanding of crossover and non-crossover recombination events in *msh1***
258 The power of long reads with high accuracy enabled unambiguous mapping across mitochondrial
259 repeats, which is crucial to investigate the detailed behavior of recombination events. For crossover
260 recombination events within perfectly matched regions, the actual recombination breakpoints could
261 not be determined from the product sequence. Among the 18 high-activity and 17 intermediate-
262 activity repeat groups, 8 groups were 100% identical. For the remaining 27 groups of non-identical
263 repeats, we could examine the single-nucleotide variants (SNVs) and indels that distinguish the
264 original repeat copies within one-rearrangement chimeric reads and gain a base-level understanding
265 of crossover site distribution. For example, repeat group CC showed multiple different
266 recombination breakpoints distributed along the length of the repeats (Figure 3a). The proportion
267 of reads with a recombination breakpoint between two SNVs/indels was positively correlated with
268 the distance between those two variants as a proportion of the total length of the repeat, based on
269 the analysis of the longest identical fragment in each group of two-copy repeats (Figure 3b; Table
270 S22).

271 We also found reads with clusters of SNVs/indels (Table S22) that could be explained by short-
272 range or partial gene conversion events (Davila et al., 2011). Particularly, in the crossover products
273 of repeat H-1 and H-3, we observed a highly discontinuous pattern of SNVs indicating multiple
274 short-range gene conversion events (Figure 3c). A proportion of the one-rearrangement reads
275 mediated by repeat H-1 and repeat H-3 also contained variants derived from repeat H-2, which
276 suggested that gene conversion could happen using other homologous dsDNA templates not
277 directly involved in a crossover recombination event (Figure 3c). To get a detailed view of non-
278 allelic gene conversion events, we calculated the ratios of gene conversion sites in zero-
279 rearrangements reads spanning the entire length of 12 high-activity repeat groups and 15
280 intermediate-activity repeat groups, and indeed observed reads containing SNVs/indels-encoded
281 by non-allelic repeat homologs in *msh1* mutants, which seldomly occurred in wild-type (Figure 4a;
282 Table S23, S24, and S25). In *msh1* mutants, repeat groups A, D, G, H (SNVs/indels in E-1 and E-
283 2 were the same ones in H-2 and H-3), I, M, MMJS, N, R, T, and W showed a high frequency of
284 non-allelic gene conversion and had higher frequencies of crossover recombination (Figure 4b,c
285 and S3; Table S21), while the remaining 15 analyzed repeats groups showed a low ratio or no non-
286 allelic gene conversion (Table S23). Within repeat H-2, SNVs/indels with a higher ratio of gene
287 conversion were not the ones shared by both H-1 and H-3. Instead, they were the unique variants
288 located near the H-1 left boundary and the H-3 right boundary (Figure 4b and S4). In repeat
289 homolog MMJS-2 the unusually high ratios of gene conversion in some individuals were associated
290 with the near complete loss of parental forms, i.e., very low sequence coverage in this region
291 (Figure 4c, S2b, and S5). Together, these base-level investigations of HiFi reads indicated that in
292 addition to suppressing ectopic crossover recombination, *MSH1* also suppresses non-crossover
293 gene conversion between imperfect repeats of the mt genome. Although less likely, we cannot
294 exclude the possibility that multiple rounds of crossover recombination might lead to the same
295 effect.
296
297 **Phenotypic correlations of mt genome variants in *msh1* mutant**
298 It was established by hemi-complementation experiments that the slower germination phenotype
299 of *msh1* mutants was caused by mitochondrial genetic effects (Virdi et al., 2015). For easy
300 quantification, we used primary root length of seedlings at 5 days after sowing (DAS) to indirectly
301 reflect seed germination speed. We observed shorter primary root lengths of *msh1* seedlings at 5
302 DAS compared with wild-type, which might indicate delayed seed germination (Figure 5a,b; Table
303 S26). The three *msh1* mutant F_3 families also exhibit quite different patterns of root length
304 distribution. For example, M3-17 contained both fast- and slow-germinating individuals and M3-

305 43 contained mostly slow-germinating individuals (Figure 5b). To further check if this slow-
306 germinating phenotype was heritable, we analyzed root length distribution patterns of F_4 families.
307 As expected, the progeny of fast-germinating M3-17 F_3 individuals germinated faster than the
308 progeny of slow-germinating M3-17 F_3 individuals and M3-43 F_3 individuals (Figure S6). The
309 overall HiFi read coverage of different *msh1* individuals did not vary greatly compared with wild-
310 type, except for some fluctuations at sporadic stretches (Figure S7; Table S27). Most of these low-
311 coverage stretches contained no coding sequences, but some overlapped with functional genes,
312 including tRNA genes and multiple open reading frames (ORFs). For example, the region around
313 repeat Large1-2 showed the lowest coverage in all *msh1* individuals, which would cause a partial
314 loss of *atp6-2*, *atp9*, *trnS* and *ORF262* (Figure 5d,e). We reasoned that if any mitochondrial
315 genomic feature caused the slow-germinating phenotype, it should occur at a substantial ratio and
316 correlate with these root length distributions of different F_3 families. We found the region between
317 the boundaries of repeat V-1 and Large1-1 was most likely to meet these criteria because there was
318 a larger reduction in *atp6-1* copy number in M3-43 individuals, which also had the shortest root
319 lengths (Figure 5c). Interestingly, it has been reported that *atp6-2* knockouts could be more easily
320 obtained than *atp6-1* knockouts with mitoTALENs (Arimura et al., 2020). Therefore, we speculate
321 that the different degrees of sub-stoichiometric shifting at the *atp6-1* gene may contribute to
322 germination phenotypes in *msh1* mutants, but this hypothesis will need to be tested with more direct
323 genetic manipulations.
324
325 **Mt genomes of other accessions showed low-coverage regions similar to *msh1* mutants**
326 We also re-assembled mt genomes for our samples using HiFi reads and metaFlye, which is a fast
327 long-read assembler for bacterial metagenomes (Kolmogorov et al., 2020). The metaFlye
328 assemblies of mt genomes were much more complex with fragmented contigs and many unresolved
329 repeats in *msh1* mutants compared with wild-type, reflecting the highly rearranged and
330 heteroplasmic mt genome (Figure S8). When assembling wild-type mt genome from published
331 HiFi datasets (Wang et al., 2021; Naish et al., 2021), the information provided by assembly graphs
332 was less complete, which suggested that our mapping-based analysis had advantages in detecting
333 alternative structures and mt genome heteroplasmy (Figure S9; Table S28 and S29).
334 One recent study used a method combining Flye (not the metagenomic version) and custom
335 clustering algorithms to study mt genome heteroplasmy using the PacBio continuous long-read
336 (CLR) datasets of Col-0 and seven other *Arabidopsis thaliana* accessions (Jiao and Schneeberger,
337 2020; Masutani et al., 2021). We analyzed these same datasets with our mapping-based pipeline.
338 The resulting metaFlye assembly graphs were completely resolved for Kyoto and Sha. Only large

339 repeats remained unresolved for other accessions, which suggests low mt genome heteroplasmy in
340 all accessions, similar to observations in wild-type Col-0 (Figure S10; Table S30 and S31).
341 Although rearrangements of large repeats could be identified, the blastn alignments of corrected
342 reads to each pseudo-master circle were less continuous than the more accurate HiFi datasets (Table
343 S32-S39).
344 The mt genomes of other accessions showed clear differences in size compared to Col-0 (Table
345 S40-S42). For example, the size difference between the largest Eri-1 mt genome and Col-0 was
346 mainly caused by a 1.1 kb insertion (Figure S11). This sequence was present in all other accessions,
347 suggesting Col-0 specific deletion (Figure S12-S17). The mt genome arrangement of C24 was
348 closely related to Eri-1, except for a 1.8 kb deletion (Figure S12). These insertions and deletions
349 were often located within large syntenic regions containing multi-copy genes or no functional genes.
350 We noticed that the region containing the repeat copy Large1-2 (258638 bp to 265227 bp in Col-0
351 mt genome) was lost in other accessions except Col-0, which were coincident with the low-
352 coverage or repeat-mediated highly dynamic regions observed in *msh1* mutants (Figure 6a-c). For
353 Eri-1 and C24, the region from repeat V-1 to Large1-1 replaced the deleted region near Large1-2,
354 which formed a pair of 7.1 kb large repeats. But for accessions Cvi-0, An-1, L*er*, Kyoto and Sha,
355 this deletion caused the loss of one copy of the *trnS* and *atp6-2* genes (Figure 6b, S13-S17). In
356 particular, the smallest mt genome of Sha accession also had an additional large deletion, which
357 caused the loss of one copy of *trnK* and *rps19* in repeat A-2 (Figure 6d,e). In an earlier report based
358 on a 5.8-kb fragment (EMBL: HF543671), loss of *trnK* was experimentally verified and a 400-bp
359 insertion containing *orf117Sha* was thought to be correlated with the CMS phenotype of Sha
360 (Gobron et al., 2013). That study also reported the presence of the 1025-bp large repeats near repeat
361 R-1, which exhibited double coverage and were recombinationally active based on our analysis
362 (Figure S17; Table S42). Thus, repeat sequences caused fast intraspecies evolution of mt genomes.
363 Our analysis confirmed previous findings that evolutionary changes of Arabidopsis mt genomes
364 were constrained by gene functions with changes that are similar rearrangements observed in *msh1*
365 mutants but a low-heteroplasmy final state (Arrieta-Montiel et al., 2009; Davila et al., 2011).
366
367 **Different patterns of structural variants and mutations in *msh1* pt genome**
368 Loss of MSH1 function in mitochondria and plastids caused similar types of variants, but with quite
369 different proportions (Figure 1c,e). MMEJ and NHEJ events were detected at very low frequencies
370 in HiFi reads. They were minor proportions of mt genome rearrangements (Table S7). But with the
371 higher depth of pt genome-mapped reads, we could observe distinctive patterns of rearrangements
372 in *msh1* mutants. First, crossover recombination events mediated by short partial palindromic

373 sequences were over-represented in *msh1* pt genome, compared with the mt genome (Table S43).
374 A partial palindromic sequence is also a specific pair of inverse repeats or microhomologies at the
375 same location. For example, we could observe crossover site distribution in chimeric reads caused
376 by rearrangement-mediated by partial palindromic sequence U-8 in *msh1* (Figure 7a; Table S44).
377 Second, although most NHEJ events were at low frequencies with a seemingly random distribution,
378 we could observe one NHEJ event which occurred only in one sample (M3-43-5) at relatively high
379 frequency (121 one-rearrangement reads versus 1385 zero-rearrangement reads), resulting in a 65-
380 bp deletion in the *petA* gene (Figure 7b; Table S45). This mutation was predicted to delete a strictly
381 conserved KQFE motif in the C-terminal stromal extension of cytochrome f, which would
382 compromise photosynthesis (Choquet et al., 2003). This indicated a different accumulation pattern
383 compared with the repeatable mitochondrial high-frequency rearrangements among samples, which
384 might suggest that different yellow sectors of leaves in *msh1* individuals could be caused by
385 different random MMEJ/NHEJ events of the pt genome. This kind of sample-specific accumulation
386 pattern of plastid variations was also reproduced when analyzing the SNVs in *msh1*. In total, we
387 identified 17 SNVs that were enriched in one or a few samples at high ratios (Figure 7c; Table S46).
388 They were predominantly AT-to-GC transitions, consistent with the mutation spectrum in *msh1*
389 mutants that was reported previously and suggested to be caused by polymerase errors (Wu et al.,
390 2020b). Third, in contrast to the indel variants caused by gene conversion between imperfect
391 repeats observed in the mt genome of *msh1* mutants, the 85 identified indel mutations in the pt
392 genome were mostly mediated by homopolymers (Tables S47). For example, the 13-bp
393 homopolymer at position 67164 contained the most variant types detected by GATK, with higher
394 variant frequencies in *msh1* mutants than the wild-type individual (Figure 7d). Indel variant
395 frequencies were also much higher in *msh1* individuals compared with wild-type at most other
396 positions (Figure 7e). Together, these results show that loss of *MSH1* causes different patterns of
397 structural variants and mutation accumulation in pt and mt genomes.
398
399
400 **Discussion**
401 In this work, we used HiFi reads (~15 kb long, > 99% accuracy) to study the behaviors of
402 mitochondrial repeats using samples from Arabidopsis *msh1* mutant individuals. HiFi reads were
403 long enough to span the entire lengths of repeats, enabling analyses that were not previously
404 possible with short reads. We developed a mapping-based analysis, which showed that HiFi reads
405 performed well in identifying more than one structural variation in a single read and studying
406 polymorphisms between long non-identical repeats. They were also sufficiently accurate to

407 distinguish mt genome sequences from those associated with the large nuclear mitochondrial
408 insertion in Chromosome 2, although that structure has not been fully captured in the reference
409 nuclear genome assembly, so it may not be possible to fully separate reads from these two different
410 sources (Stupar et al., 2001). Compared to previous studies, we also identified additional repeats
411 active in recombination at lower frequencies. However, due to the low number of read counts, HiFi
412 reads were not cost-effective for studying low-frequency MMEJ and NHEJ variants. This cost
413 could be reduced by isolating organelles from the cells before DNA extraction for sequencing.
414

415 **MSH1 suppresses gene conversion between non-identical repeats**

416 In meiosis, MutS homologs detect mismatches and recruit the DSB-generating SPO11 complex
417 to suppress crossover repair between non-identical centromeric repeats (Blackwell et al., 2020;
418 Naish et al., 2021). SPO11 also repairs double DSB (dDSB) gaps independent of MSH2 at DSB
419 hotspots (Prieler et al., 2021). It was proposed that DNA mismatches detected by organelle-targeted
420 MSH1 are converted to DSBs by the GIY-YIG endonuclease domain at its C-terminus and directed
421 to DSB repair pathways (Christensen, 2014). In this analysis, HiFi reads were unambiguously
422 mapped across each mitochondrial repeat homolog, leading to the identification of high-frequency
423 non-crossover recombination or gene conversion in zero-rearrangement reads in *msh1* mutants,
424 including both SNVs (as in repeats A, H, etc.) and indels (as in repeat MMJS). Gene conversion
425 was previously observed to explain the discontinuous pattern of SNVs/indels between non-identical
426 repeats in chimeric reads (Davila et al., 2011), the limitations of short reads have generally
427 prevented comprehensive analyses across repeats of different lengths and quantitative comparisons
428 between wild-type and *msh1* mutants.

429 Numerous earlier studies have focused on the massive increase in repeat-mediated crossover
430 recombination and mt genome rearrangements in *msh1* mutants (Abdelnoor et al., 2003; Shedge et
431 al., 2007; Arrieta-Montiel et al., 2009; Davila et al; 2011). More recently, we found a role of MSH1
432 in suppressing organellar point mutations and indels (Wu et al., 2020b). At first glance, it might
433 seem contradictory that MSH1 could be involved in suppressing ectopic recombination between
434 short repeats, while also inducing homologous recombination as part of a mismatch repair pathway
435 (Christensen, 2014). However, there could very well be a connection between these two processes.
436 Specifically, MSH1 may provide a surveillance mechanism to check for mismatches in double-
437 stranded DNA regardless of whether they are introduced by polymerase errors or by strand invasion
438 between short or non-identical repeats, followed by introduction of DSBs to (re)start repair by
439 homologous recombination. To suppress recombination at short and intermediate repeats that are
440 100% identical in sequence, MSH1 may need to detect differences in boundary sequence. During

441 homologous recombination, the bacterial RecA-ssDNA presynaptic filament checks homology in
442 ~9 bp steps before strand exchange with the dsDNA template (Huang et al., 2020). In such
443 screening mechanisms, the first mismatches may be detected at the repeat boundaries when the
444 repeat sequences themselves are identical. Thus, MSH1 may suppress both crossover and non-
445 crossover forms of ectopic recombination to ensure accurate DSB repair of the genome by detecting
446 mismatches within repeats or at repeat boundaries.
447
448 **Conventional recombination activity only accounts for a proportion of DSB repair events**
449 Most previous mitochondrial repeat annotations depended on the penalty parameters of mismatches
450 and gaps of raw sequence alignment results. We showed that through comparing chimeric HiFi
451 reads with blastn results, it was possible to identify additional low-frequency repeats, raising
452 questions about what contributes to the observed differences in recombination activity of each
453 repeat pair. Increased crossover recombination events of intermediate-sized repeats have been
454 observed in wild-type plants when DSB frequencies were increased by DNA damaging chemicals
455 (Wallet et al., 2015). Therefore, repeats near DSB hotspots may show higher ectopic recombination
456 activity compared to other repeats with similar length and sequence identity, which might explain
457 cases that fall far from the regression line that defines the relationship between recombination
458 activity and repeat length (Figure S1c). It may also help to study repair events systematically by
459 inducing DSBs through mitochondrial-targeting CRISPR-Cas9 or mitoTALENs (Arimura et al.,
460 2020) and in more RRR gene mutants, such as *why2*, *reca3*, *osb1* (Gualberto and Newton 2017).
461 DSBs within a repeat or in the flanking sequence would be repaired using homologous DNA
462 stretches of another intra- or intermolecular copy. After resection of DSB ends, RecA recombinases
463 bind 3' ssDNA and rapidly search for homology (Wiktor et al., 2021). It was reported that at least
464 8-nt microhomology was needed for RecA-ssDNA filaments to bind dsDNA templates stably, and
465 15-nt was sufficient to be resistant to facilitated exchange during homology search (Qi et al., 2015).
466 These studies suggested that only a small stretch of RecA-ssDNA was needed during each HDR
467 repair event. Typically, only crossover or BIR products have been used to estimate frequencies of
468 recombinant conformations (Wallet et al., 2015; Odahara et al., 2017). Such calculations do not
469 account for the fact that gene conversion events between repeats can also occur at substantial
470 frequencies. These gene conversion events likely explain why larger repeats tend to be 100%
471 identical. In wild-type plants, only the two large repeat pairs showed high recombination activities.
472 The observed difference in *msh1* mutants was the increased proportion of uncontrolled gene
473 conversion between non-identical repeats and BIR proportion of HDR repair. The lower detected
474 recombination activity of large repeats in *msh1* could be due to repair of DSBs by other repeats

475 (Figure 1d). When calculating recombination of some three-copy or four-copy repeats locating in
476 other longer two-copy repeats, BIR-indicating reads shared by both shorter and longer annotations
477 were assigned to the larger two-copy repeats (Figure S2b). This indicated that it was more natural
478 to view recombination events as associated with specific genomic regions per se, but not with repeat
479 annotations.
480
481 **Phenotypic contributions of cytoplasmic genetic materials**
482 During seed germination, normal mitochondrial biogenesis and mtDNA copy number are tightly
483 controlled (Paszkiewicz et al., 2017). This connection between seed germination and mitochondrial
484 defects was also reported in many other studies (Han et al., 2010; Virdi et al., 2015; Van Aken et
485 al., 2016; Heidorn-Czarna et al., 2018; Golin et al., 2020). We showed that variation in root lengths
486 at 5 DAS was heritable among different *msh1* mutant families. Based on sequencing coverage
487 distributions, the degree of slow seed germination was potentially associated with reduction in
488 *atp6-1* copy number. Generally, coverage analysis indicated similar patterns of fluctuation in
489 sequence depth across *msh1* individuals, which could be a combined effect of the locations of repeat
490 pairs and the selection to maintain regions containing functional genes (Figure S7). Interestingly,
491 the most dynamic region in *msh1* containing Large1-2 was lost in other Arabidopsis accessions.
492 Despite the very rapid intraspecific structural rearrangements of mt genomes, the various
493 Arabidopsis accessions were all in a low-heteroplasmy state. During the process of
494 substoichiometric shifting (Woloszynska, 2009; Tang et al., 2016; Johnston, 2019), the mt genome
495 may be transiently in a highly rearranged and heteroplasmic state similar to *msh1* mutants,
496 providing source materials for functional innovation.
497 In comparison to plant mt genomes, the pt genome is smaller and contains a pair of very long
498 inverted repeats. The contrasting patterns of variation between mt and pt genomes may result from
499 differences in repeat spectrum, genome copy number per organelle, and DNA repair mechanisms.
500 In the pt genome of *msh1* mutants, HiFi-identified rearrangements were mostly MMEJ and NHEJ
501 events. MMEJ events mediated by partial palindromic sequences were found at a higher frequency
502 than others and showed repeatability among samples. However, a few other pt genome variants
503 occurred at high frequency in individual samples (Figure 7b). These individual high-frequency
504 variants might reflect mutations in different yellow patches, which can often be often in a
505 homoplasmic state because of random sorting early in leaf primordia (Malinova et al., 2021). Leaf
506 variegation and albino phenotypes are common in other RRR gene mutants (for example, moc1
507 single mutant, *why1 why3 pol1b* and *why1 why3 reca1* triple mutants) and transgenic plants with

508 plastid-targeted restriction endonucleases (Zampini et al., 2015; Kobayashi et al., 2017; Sugimoto
509 et al., 2020).
510 In conclusion, *MSH1* acts to control both structural and non-structural variants in plant organellar
511 genomes. The genomes could be transiently uncontrolled when *MSH1* expression was
512 downregulated, providing more raw materials for further functional selection (Yang and Mackenzie,
513 2019). However, the two genomes responded quite differently to the loss of the same gene. In the
514 future, through understanding the dynamics of these repeats and the detailed functions of RRR
515 genes in DSB repair processes, we may better harness the cytoplasmic genetic material for crop
516 improvement with recently developed mt genome editing technologies (Kazama et al., 2019;
517 Arimura et al., 2020).
518
519
520 **Methods**
521 **Plant Materials and Growth Conditions**
522 Generation of *msh1* mutant and wild-type F_3 families of *Arabidopsis thaliana* used for HiFi
523 sequencing was described previously (Wu et al., 2020b). Individuals showing clear long and short
524 root lengths at 5 days after sowing (DAS) were selected to get corresponding F_4 families. For
525 estimating the speed of seed germination, seeds were sowed on solid ½ Murashige and Skoog media
526 with vitamins and 1% sucrose, stratified in darkness (4°C, 2 d), and grown in a light incubator
527 (22°C, 16 h day and 8 h night) for another three days, e.g., 5 DAS. Photos of seedlings at 5 DAS
528 were analyzed with ImageJ to measure primary root lengths.
529
530 **DNA extraction and PacBio HiFi sequencing**
531 Above-ground tissues of 11 individual plants grown in a short-day grown chamber (8 h day and 16
532 h night) for six weeks were collected to extract DNA with the CTAB method. Total DNA was
533 purified with a QIAGEN kit (Q13343) and quantified using NanoDrop and Qubit. Purified DNA
534 (3.3~18.6 μg for each sample) was used for the construction of SMRTbell libraries according to
535 PacBio's standard protocol. Sequencing was performed using the PacBio Sequel II platform in CCS
536 mode, which generated a total of 34.36 Gb HiFi reads (12.7~18.1 kb mean read length).
537
538 **Identification of rearrangements in mt and pt genomes**
539 Raw HiFi reads of each sample were mapped to the Arabidopsis Col-0 reference genome (TAIR10
540 for nucleus; NC_037304.1 for mitochondria; NC_000932.1 plus a 1-bp expansion at 28673 for
541 plastid as described in Wu, et al., 2020) with minimap2 (version 2.20, r1061; parameters, -ax map-

542 hifi; Li, 2018). Reads mapped to mt and pt genomes were separated using bamtools (version 2.5.1)
543 and realigned against each genome with blastn (version 2.12.0). Raw blastn results were further
544 processed by a custom pipeline (available at https://github.com/zouyinstein/hifisr) to group the
545 reads by discontinuity, homologous sequence size and mapping coordinates, and calculate read
546 counts for input samples. Different types of genome rearrangements were manually checked,
547 including zero-, one-, two-, three-, and four-rearrangement events in a single read, microhomology-
548 mediated end joining (MMEJ; 2-49 bp homology sequence size) and non-homologous end-joining
549 (NHEJ; 0-1 bp homology sequence size), replication slippage mediated by tandem repeats, and
550 insertions.
551
552 **Mitochondrial repeat annotations and recombination activity**
553 Candidate coordinates of repeats were generated by blastn (version 2.12.0+; parameters, -
554 word_size 4) using the mitochondrial reference sequence as both query and subject. The resulting
555 coordinates were manually checked against the junction coordinates of one-rearrangement HiFi
556 reads. In some cases, the boundaries of low-identity repeats were different from previous
557 annotations. Per base reads counts ratios of zero-rearrangement reads versus total reads indicating
558 dynamic regions of mt genome, zero-rearrangement and one-rearrangement read counts indicating
559 parental and recombinant forms of repeat groups were calculated based on blastn outputs of the
560 custom HiFi-SR pipeline.
561
562 **Identification of gene conversion between imperfect repeats**
563 To analyze patterns of SNVs/indels, zero-rearrangement reads of each sample were re-aligned to
564 the mitochondrial reference genome by PacBio pbmm2 (version 1.3.0; parameters, --sort --preset
565 CCS). The resulting BAM files were used for variant calling by GATK HaplotypeCaller (version
566 4.2.3.0). Depths of A, T, G, C and read counts of different indel sizes, and the percentage of
567 reference cases of each GATK-called SNVs/indels were calculated from raw BAM files with the
568 Python module pysam (version 0.17.0). SNVs/indels inferred from gene conversion events between
569 some high-activity mitochondrial imperfect repeats were also used for calculations, which were not
570 always detected by GATK.
571
572 **Calculation of absolute and normalized per base coverage**
573 Total mt genome-mapped reads were re-aligned to mitochondrial reference genomes by PacBio
574 pbmm2 (version 1.3.0; parameters, --sort --preset CCS). Absolute per base coverage across the
575 entire mt genome of each individual and the published datasets was calculated by samtools depth

576 subcommand (version 1.7) and normalized to total bases of each sample. Annotations are based on
577 the GenBank file NC_037304.1 (Sloan et al., 2018).
578
579 **metaFlye assembly and synteny analysis of mt genomes**
580 Raw PacBio CLR reads of each accession were mapped to reference genome (the respective nuclear
581 assembly in Jiao and Schneeberger, 2020; Col-0 NC_037304.1 for mitochondria and NC_000932.1
582 plus a 1-bp expansion at 28673 for plastid) with minimap2 (version 2.20, r1061; parameters, -ax
583 map-pb; Li, 2018). Reads mapped to mt genomes were separated using bamtools (version 2.5.1)
584 and were re-assembled with metaFlye (Flye version 2.9-b1774; parameters, --meta --pacbio-hifi/--
585 pacbio-corr --extra-params output_gfa_before_rr=1 --genome-size 370K/2M for mitochondria).
586 Before assembling, PacBio CLR reads longer than 1000 bp were extracted by Filtlong (version
587 0.2.1; parameters, --min_length 1000) and were corrected using CONSENT (version 2.2.2;
588 parameters, --type PB). Assembly graphs before and after repeat resolution were visualized by
589 Bandage (version 0.8.1). Reads from different *msh1* families are merged. To get linear
590 representations of pseudo-master circles for mapping-based analysis, we started from the
591 assemblies (--genome-size 2M) after repeat resolution, discarded low-coverage contigs and
592 manually resolved the large repeats for accessions Eri-1, C24, Cvi-0, An-1, L*er*, Sha, and re-
593 resolved Large2 repeats for Kyoto with the Bandage software, followed by rotation or reverse
594 complementation. Corrected CLR reads were realigned against each pseudo-master circle using the
595 HiFi-SR pipeline. To analyze the shared synteny between mt genomes, we performed blastn
596 alignments using the Col-0 mitochondrial reference sequence as query and pseudo-master circle of
597 the other accession as subject. Raw blastn results were processed by a custom Python script
598 available in HiFi-SR. Coordinates or large repeats were determined based on blastn alignments
599 using the pseudo-master circle as both query and subject. Results were generated by the Python
600 package pycircos (version 1.0.2).
601
602 **Identification of SNVs/indels pt genome**
603 To analyze patterns of SNVs/indels in pt genome, total reads of each sample were re-aligned to
604 plastid reference genome by PacBio pbmm2 (version 1.3.0; parameters, --sort --preset CCS). The
605 resulting BAM files were used for variant calling by GATK HaplotypeCaller (version 4.2.3.0).
606 Depths of A, T, G, C and read counts of different indel sizes, and the percentage of reference cases
607 of each GATK-called SNVs/indels were calculated from raw BAM files with the Python module
608 pysam (version 0.17.0).
609

Acknowledgments

We thank Xiaojing Zhang from Prof. Li-Zhen Tao's laboratory at South China Agricultural University for assistance with plant growth. This work was supported by grants from National Natural Science Foundation of China (32170238), the Science, Technology and Innovation Commission of Shenzhen Municipality (RCYX20200714114538196), and the U.S. National Institutes of Health (R01 GM118046).

Author contributions

ZW designed the research; YZ performed the experiments; YZ wrote the computer scripts and performed majority of the bioinformatic analysis; YZ, WZ, DS, and ZW discussed the results; YZ, DS and ZW wrote the manuscript; DS and ZW revised the manuscript.

ORCID

Yi Zou: 0000-0003-3117-3898

Weidong Zhu: 0000-0002-9613-354X

Daniel B. Sloan: 0000-0002-3618-0897

Zhiqiang Wu: 0000-0002-4238-7317

Conflict of Interest

The authors have no conflict of interest to declare.

Data Availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Chen et al, 2021) in National Genomics Data Center (C.-N. Members and Partners, 2022), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA006060) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa. The Python-based pipeline for the detection of plant mitochondrial structural rearrangements used in this study is available at https://github.com/zouyinstein/hifisr.

Supporting Information

Figure S1 99 repeat groups supported by HiFi reads in the mt genome of *msh1*

Figure S2 Asymmetrical recombination phenomena in mt genome

Figure S3 Imperfect repeats with relatively higher ratios of gene conversion events

Figure S4 Coverages of regions containing repeat group H in wild-type and *msh1* individuals

644 **Figure S5** Coverages of regions containing repeat group MMJS in wild-type and *msh1* individuals
645 **Figure S6** Primary root length of wild-type and *msh1* seedlings at 5 days after sowing
646 **Figure S7** Normalized per base read coverage across mt genome
647 **Figure S8** The metaFlye assemblies of mt genomes are more complex in *msh1* than wild-type
648 **Figure S9** Assembly graphs and mapping-based analysis of published HiFi datasets
649 **Figure S10** The metaFlye assemblies of mt genomes are simple as wild-type in other accessions
650 **Figure S11** Synteny and coverage analysis of Eri-1 mt genome
651 **Figure S12** Synteny and coverage analysis of C24 mt genome
652 **Figure S13** Synteny and coverage analysis of Cvi-0 mt genome
653 **Figure S14** Synteny and coverage analysis of An-1 mt genome
654 **Figure S15** Synteny and coverage analysis of L*er* mt genome
655 **Figure S16** Synteny and coverage analysis of Kyoto mt genome
656 **Figure S17** Synteny and coverage analysis of Sha mt genome
657 **Table S1** DNA content and fresh weight of individual plants used for HiFi sequencing
658 **Table S2** Mapping statistics of HiFi reads of individual plants
659 **Table S3** Statistics of mt genome-mapped reads of individual plants
660 **Table S4** Statistics of pt genome-mapped reads of individual plants
661 **Table S5** Read counts of different types of mitochondrial structural rearrangements of individual
662 plants
663 **Table S6** Read counts of different types of plastid structural rearrangements of individual plants
664 **Table S7** Read counts of different categories of structural rearrangements in mt genome of wild-
665 type and *msh1* plants
666 **Table S8** Read counts of different categories of structural rearrangements in pt genome of wild-
667 type and *msh1* plants
668 **Table S9** Overlapping region length of the two alignments in mitochondrial one-rearrangement
669 events detected by HiFi reads.
670 **Table S10** Overlapping region length of the two alignments in plastid one-rearrangement events
671 detected by HiFi reads.
672 **Table S11** Raw blastn outputs using the plastid reference sequence as both query and subject
673 **Table S12** HiFi-SR outputs of one-rearrangement mt genome-mapped reads of individual plants
674 **Table S13** HiFi-SR outputs of two-rearrangement mt genome-mapped reads of individual plants
675 **Table S14** Raw blastn outputs using the mitochondrial reference sequence as both query and
676 subject
677 **Table S15** Annotations of two-copy repeats based on blastn and one-rearrangement HiFi reads

678 **Table S16** Annotations of three-copy repeats based on blastn and one-rearrangement HiFi reads
679 **Table S17** Annotations of four-copy repeats based on blastn and one-rearrangement HiFi reads
680 **Table S18** Recombination frequencies of two-copy repeats
681 **Table S19** Recombination frequencies of three-copy repeats
682 **Table S20** Recombination frequencies of three-copy repeats
683 **Table S21** HiFi read counts indicating parental forms and crossover recombination products of
684 representative repeats in wild-type and *msh1*.
685 **Table S22** Counts of continuous and discontinuous one-rearrangement reads in high- and
686 intermediate-activity two-copy imperfect repeats
687 **Table S23** Ratios of predicted gene conversion sites of individual plants and the published HiFi
688 datasets in mt genome
689 **Table S24** Ratios of GATK-identified SNVs of individual plants and the published HiFi datasets
690 in mt genome
691 **Table S25** Ratios of GATK-identified indels of individual plants and the published HiFi datasets
692 in mt genome
693 **Table S26** Primary root length frequency of F3 families and selected F4 progenies at 5 days after
694 sowing
695 **Table S27** Absolute and normalized per base read coverage across mt genome of each individual
696 and the published HiFi datasets
697 **Table S28** HiFi-SR outputs of one-rearrangement mt genome-mapped reads of wild-type XJTU
698 **Table S29** HiFi-SR outputs of one-rearrangement mt genome-mapped reads of wild-type CEN
699 **Table S30** Mapping statistics of HiFi reads of individual plants
700 **Table S31** Statistics of corrected PacBio CLR reads used for metaFlye assembly of mt genomes
701 **Table S32** HiFi-SR outputs of one-rearrangement reads of Col-0 CLR dataset
702 **Table S33** HiFi-SR outputs of one-rearrangement reads of An-1 CLR dataset
703 **Table S34** HiFi-SR outputs of one-rearrangement reads of C24 CLR dataset
704 **Table S35** HiFi-SR outputs of one-rearrangement reads of Cvi-0 CLR dataset
705 **Table S36** HiFi-SR outputs of one-rearrangement reads of Eri-1 CLR dataset
706 **Table S37** HiFi-SR outputs of one-rearrangement reads of Kyoto CLR dataset
707 **Table S38** HiFi-SR outputs of one-rearrangement reads of L*er* CLR dataset
708 **Table S39** HiFi-SR outputs of one-rearrangement reads of Sha CLR dataset
709 **Table S40** Blastn results of large repeats in mt genomes
710 **Table S41** Blastn results of syntenic relationships of mt genomes

711 **Table S42** Absolute and normalized per base read coverage across mt genome of each *Arabidopsis*
712 *thaliana* accessions from the published PacBio CLR datasets
713 **Table S43** HiFi-SR outputs of one-rearrangement mediated by short partial palindromic sequences
714 in pt genomes of individual plants
715 **Table S44** HiFi-SR outputs of one-rearrangement mediated by short partial palindromic sequence
716 U-8 in pt genomes of individual plants
717 **Table S45** HiFi-SR outputs of a NHEJ event in pt genomes of individual plants
718 **Table S46** Ratios of GATK-identified SNVs of individual plants and the published HiFi datasets
719 in pt genome
720 **Table S47** Ratios of GATK-identified indels of individual plants and the published HiFi datasets
721 in pt genome
722
723 **References**

724 1. Abdelnoor RV, Yule R, Elo A, Christensen AC, Meyer-Gauen G, Mackenzie SA.
725 Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene
726 homologous to MutS. **Proc Natl Acad Sci U S A 2003**, 100(10)**:** 5968-5973.
727 2. Arimura SI, Ayabe H, Sugaya H, Okuno M, Tamura Y, Tsuruta Y, Watari Y, Yanase S,
728 Yamauchi T, Itoh T, Toyoda A, Takanashi H, Tsutsumi N. Targeted gene disruption of
729 *ATP synthases 6-1* and *6-2* in the mitochondrial genome of *Arabidopsis thaliana* by
730 mitoTALENs. **Plant J 2020**, 104(6)**:** 1459-1471.
731 3. Arrieta-Montiel MP, Shedge V, Davila J, Christensen AC, Mackenzie SA. Diversity of the
732 Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity.
733 **Genetics 2009**, 183(4)**:** 1261-1268.
734 4. Ayala-Garcia VM, Baruch-Torres N, Garcia-Medel PL, Brieba LG. Plant organellar DNA
735 polymerases paralogs exhibit dissimilar nucleotide incorporation fidelity. **FEBS J 2018**,
736 285(21)**:** 4005-4018.
737 5. Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: a C++ API
738 and toolkit for analyzing and managing BAM files. **Bioinformatics 2011**, 27(12)**:** 1691-
739 1692.
740 6. Blackwell AR, Dluzewska J, Szymanska-Lejman M, Desjardins S, Tock AJ, Kbiri N,
741 Lambing C, Lawrence EJ, Bieluszewski T, Rowan B, Higgins JD, Ziolkowski PA,
742 Henderson IR. MSH2 shapes the meiotic crossover landscape in relation to interhomolog
743 polymorphism in Arabidopsis. **EMBO J 2020**, 39(21)**:** e104858.

744 7. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.
745 BLAST+: architecture and applications. **BMC Bioinformatics 2009**, 10**:** 421.
746 8. Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, Dong L, Zhang Z, Yu C, Sun Y, Chi
747 L, Chen H, Zhai S, Sun Y, Lan L, Zhang X, Xiao J, Bao Y, Wang Y, Zhang Z, Zhao W.
748 The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data
749 Types. **Genomics Proteomics Bioinformatics 2021**.
750 9. Chevigny N, Schatz-Daas D, Lotfi F, Gualberto JM. DNA Repair and the Stability of the
751 Plant Mitochondrial Genome. **Int J Mol Sci 2020**, 21(1).
752 10. Choquet Y, Zito F, Wostrikoff K, Wollman FA. Cytochrome f translation in
753 Chlamydomonas chloroplast is autoregulated by its carboxyl-terminal domain. **Plant Cell**
754 **2003**, 15(6)**:** 1443-1454.
755 11. Christensen AC. Plant mitochondrial genome evolution can be explained by DNA repair
756 mechanisms. **Genome Biol Evol 2013**, 5(6)**:** 1079-1086.
757 12. Christensen AC. Genes and junk in plant mitochondria-repair mechanisms and selection.
758 **Genome Biol Evol 2014**, 6(6)**:** 1448-1453.
759 13. Christensen AC. Plant Mitochondria are a Riddle Wrapped in a Mystery Inside an Enigma.
760 **J Mol Evol 2021**, 89(3)**:** 151-156.
761 14. Davila JI, Arrieta-Montiel MP, Wamboldt Y, Cao J, Hagmann J, Shedge V, Xu YZ, Weigel
762 D, Mackenzie SA. Double-strand break repair processes drive evolution of the
763 mitochondrial genome in Arabidopsis. **BMC Biol 2011**, 9**:** 64.
764 15. Flood PJ, Theeuwen T, Schneeberger K, Keizer P, Kruijer W, Severing E, Kouklas E,
765 Hageman JA, Wijfjes R, Calvo-Baltanas V, Becker FFM, Schnabel SK, Willems LAJ,
766 Ligterink W, van Arkel J, Mumm R, Gualberto JM, Savage L, Kramer DM, Keurentjes
767 JJB, van Eeuwijk F, Koornneef M, Harbinson J, Aarts MGM, Wijnker E. Reciprocal
768 cybrids reveal how organellar genomes affect plant phenotypes. **Nat Plants 2020**, 6(1)**:**
769 13-21.
770 16. Fuchs P, Rugen N, Carrie C, Elsasser M, Finkemeier I, Giese J, Hildebrandt TM, Kuhn K,
771 Maurino VG, Ruberti C, Schallenberg-Rudinger M, Steinbeck J, Braun HP, Eubel H,
772 Meyer EH, Muller-Schussele SJ, Schwarzlander M. Single organelle function and
773 organization as estimated from Arabidopsis mitochondrial proteomics. **Plant J 2020**,
774 101(2)**:** 420-441.
775 17. Fukui K, Harada A, Wakamatsu T, Minobe A, Ohshita K, Ashiuchi M, Yano T. The GIY-
776 YIG endonuclease domain of *Arabidopsis* MutS homolog 1 specifically binds to branched
777 DNA structures. **FEBS Lett 2018**, 592(24)**:** 4066-4077.

778 18. Gobron N, Waszczak C, Simon M, Hiard S, Boivin S, Charif D, Ducamp A, Wenes E,
779 Budar F. A cryptic cytoplasmic male sterility unveils a possible gynodioecious past for
780 *Arabidopsis thaliana*. **PLoS One 2013**, 8(4)**:** e62450.
781 19. Golin S, Negroni YL, Bennewitz B, Klosgen RB, Mulisch M, La Rocca N, Cantele F,
782 Vigani G, Lo Schiavo F, Krupinska K, Zottini M. WHIRLY2 plays a key role in
783 mitochondria morphology, dynamics, and functionality in *Arabidopsis thaliana*. **Plant**
784 **Direct 2020**, 4(5)**:** e00229.
785 20. Gualberto JM, Newton KJ. Plant Mitochondrial Genomes: Dynamics and Mechanisms of
786 Mutation. **Annu Rev Plant Biol 2017**, 68**:** 225-252.
787 21. Han L, Qin G, Kang D, Chen Z, Gu H, Qu LJ. A nuclear-encoded mitochondrial gene
788 *AtCIB22* is essential for plant development in *Arabidopsis*. **J Genet Genomics 2010**,
789 37(10)**:** 667-683.
790 22. Heidorn-Czarna M, Domanski D, Kwasniak-Owczarek M, Janska H. Targeted Proteomics
791 Approach Toward Understanding the Role of the Mitochondrial Protease FTSH4 in the
792 Biogenesis of OXPHOS During Arabidopsis Seed Germination. **Front Plant Sci 2018**, 9**:**
793 821.
794 23. Huang X, Lu Y, Wang S, Sui M, Li J, Ma J, Ma D, Jia Q, Hu S, Xu C, Li M. Mismatch
795 sensing by nucleofilament deciphers mechanism of RecA-mediated homologous
796 recombination. **Proc Natl Acad Sci U S A 2020**, 117(34)**:** 20549-20554.
797 24. Hussmann D, Starnawska A, Kristensen L, Daugaard I, Thomsen A, Kjeldsen TE, Hansen
798 CS, Bybjerg-Grauholm J, Johansen KD, Ludvigsen M, Kristensen T, Larsen TS, Moller
799 MB, Nyvold CG, Hansen LL, Wojdacz TK. *IGHV*-associated methylation signatures more
800 accurately predict clinical outcomes of chronic lymphocytic leukemia patients than *IGHV*
801 mutation load. **Haematologica 2021**.
802 25. Jiao WB, Schneeberger K. Chromosome-level assemblies of multiple Arabidopsis
803 genomes reveal hotspots of rearrangements with altered evolutionary dynamics. **Nat**
804 **Commun 2020**, 11(1)**:** 989.
805 26. Johnston IG. Tension and Resolution: Dynamic, Evolving Populations of Organelle
806 Genomes within Plant Cells. **Mol Plant 2019**, 12(6)**:** 764-783.
807 27. Kazama T, Okuno M, Watari Y, Yanase S, Koizuka C, Tsuruta Y, Sugaya H, Toyoda A,
808 Itoh T, Tsutsumi N, Toriyama K, Koizuka N, Arimura SI. Curing cytoplasmic male sterility
809 via TALEN-mediated mitochondrial genome editing. **Nat Plants 2019**, 5(7)**:** 722-730.

810 28. Kobayashi Y, Misumi O, Odahara M, Ishibashi K, Hirono M, Hidaka K, Endo M,
811 Sugiyama H, Iwasaki H, Kuroiwa T, Shikanai T, Nishimura Y. Holliday junction
812 resolvases mediate chloroplast nucleoid segregation. **Science 2017**, 356(6338)**:** 631-634.
813 29. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan
814 J, Polevikov E, Smith TPL, Pevzner PA. metaFlye: scalable long-read metagenome
815 assembly using repeat graphs. **Nat Methods 2020**, 17(11)**:** 1103-1110.
816 30. Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, Christensen AC.
817 The alternative reality of plant mitochondrial DNA: One ring does not rule them all. **PLoS**
818 **Genet 2019**, 15(8)**:** e1008373.
819 31. Kundariya H, Yang X, Morton K, Sanchez R, Axtell MJ, Hutton SF, Fromm M, Mackenzie
820 SA. MSH1-induced heritable enhanced growth vigor through grafting is associated with
821 the RdDM pathway in plants. **Nat Commun 2020**, 11(1)**:** 5343.
822 32. Lee J, Cho CH, Park SI, Choi JW, Song HS, West JA, Bhattacharya D, Yoon HS. Parallel
823 evolution of highly conserved plastid genome architecture in red seaweeds and seed plants.
824 **BMC Biol 2016**, 14**:** 75.
825 33. Li H. Minimap2: pairwise alignment for nucleotide sequences. **Bioinformatics 2018**,
826 34(18)**:** 3094-3100.
827 34. Mackenzie SA, Kundariya H. Organellar protein multi-functionality and phenotypic
828 plasticity in plants. **Philos Trans R Soc Lond B Biol Sci 2020**, 375(1790)**:** 20190182.
829 35. Malinova I, Zupok A, Massouh A, Schottler MA, Meyer EH, Yaneva-Roder L, Szymanski
830 W, Rossner M, Ruf S, Bock R, Greiner S. Correction of frameshift mutations in the *atpB*
831 gene by translational recoding in chloroplasts of *Oenothera* and tobacco. **Plant Cell 2021**,
832 33(5)**:** 1682-1705.
833 36. Massouh A, Schubert J, Yaneva-Roder L, Ulbricht-Jones ES, Zupok A, Johnson MT,
834 Wright SI, Pellizzer T, Sobanski J, Bock R, Greiner S. Spontaneous Chloroplast Mutants
835 Mostly Occur by Replication Slippage and Show a Biased Pattern in the Plastome of
836 *Oenothera*. **Plant Cell 2016**, 28(4)**:** 911-929.
837 37. Masutani B, Arimura SI, Morishita S. Investigating the mitochondrial genomic landscape
838 of *Arabidopsis thaliana* by long-read sequencing. **PLoS Comput Biol 2021**, 17(1)**:**
839 e1008597.
840 38. Members C-N, Partners. Database Resources of the National Genomics Data Center, China
841 National Center for Bioinformation in 2022. **Nucleic Acids Res 2022**, 50(D1)**:** D27-D38.

842 39. Morisse P, Marchet C, Limasset A, Lecroq T, Lefebvre A. Scalable long read self-
843 correction and assembly polishing with multiple sequence alignment. **Sci Rep 2021**, 11(1)**:**
844 761.
845 40. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmucker A, Mandakova
846 T, Jamge B, Lambing C, Kuo P, Yelina N, Hartwick N, Colt K, Smith LM, Ton J, Kakutani
847 T, Martienssen RA, Schneeberger K, Lysak MA, Berger F, Bousios A, Michael TP, Schatz
848 MC, Henderson IR. The genetic and epigenetic landscape of the *Arabidopsis* centromeres.
849 **Science 2021**, 374(6569)**:** eabi7489.
850 41. Odahara M, Kishita Y, Sekine Y. MSH1 maintains organelle genome stability and
851 genetically interacts with *RECA* and *RECG* in the moss *Physcomitrella patens*. **Plant J**
852 **2017**, 91(3)**:** 455-465.
853 42. Palmer JD, Herbon LA. Plant mitochondrial DNA evolves rapidly in structure, but slowly
854 in sequence. **J Mol Evol 1988**, 28(1-2)**:** 87-97.
855 43. Paszkiewicz G, Gualberto JM, Benamar A, Macherel D, Logan DC. Arabidopsis Seed
856 Mitochondria Are Bioenergetically Active Immediately upon Imbibition and Specialize via
857 Biogenesis in Preparation for Autotrophic Growth. **Plant Cell 2017**, 29(1)**:** 109-128.
858 44. Prieler S, Chen D, Huang L, Mayrhofer E, Zsoter S, Vesely M, Mbogning J, Klein F. Spo11
859 generates gaps through concerted cuts at sites of topological stress. **Nature 2021**,
860 594(7864)**:** 577-582.
861 45. Qi Z, Redding S, Lee JY, Gibb B, Kwon Y, Niu H, Gaines WA, Sung P, Greene EC. DNA
862 sequence alignment by microhomology sampling during homologous recombination. **Cell**
863 **2015**, 160(5)**:** 856-869.
864 46. Redei GP. Extra-chromosomal mutability determined by a nuclear gene locus in
865 Arabidopsis. **Mutat Res 1973**, 18: 149–162.
866 47. Sandhu AP, Abdelnoor RV, Mackenzie SA. Transgenic induction of mitochondrial
867 rearrangements for cytoplasmic male sterility in crop plants. **Proc Natl Acad Sci U S A**
868 **2007**, 104(6)**:** 1766-1770.
869 48. Shedge V, Arrieta-Montiel M, Christensen AC, Mackenzie SA. Plant mitochondrial
870 recombination surveillance requires unusual *RecA* and *MutS* homologs. **Plant Cell 2007**,
871 19(4)**:** 1251-1264.
872 49. Sloan DB, Wu Z, Sharbrough J. Correction of Persistent Errors in Arabidopsis Reference
873 Mitochondrial Genomes. **Plant Cell 2018**, 30(3)**:** 525-527.
874 50. Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, Jiang J. Complex mtDNA
875 constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2:

876 implication of potential sequencing errors caused by large-unit repeats. **Proc Natl Acad**
877 **Sci U S A 2001**, 98(9)**:** 5099-5103.
878 51. Sugimoto H, Hirano M, Tanaka H, Tanaka T, Kitagawa-Yogo R, Muramoto N, Mitsukawa
879 N. Plastid-targeted forms of restriction endonucleases enhance the plastid genome
880 rearrangement rate and trigger the reorganization of its genomic architecture. **Plant J 2020**,
881 102(5)**:** 1042-1057.
882 52. Tang H, Zheng X, Li C, Xie X, Chen Y, Chen L, Zhao X, Zheng H, Zhou J, Ye S, Guo J,
883 Liu YG. Multi-step formation, evolution, and functionalization of new cytoplasmic male
884 sterility genes in the plant mitochondrial genomes. **Cell Res 2017**, 27(1)**:** 130-146.
885 53. Unseld M, Marienfeld JR, Brandt P, Brennicke A. The mitochondrial genome of
886 *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. **Nat Genet 1997**, 15(1)**:**
887 57-61.
888 54. Van Aken O, Ford E, Lister R, Huang S, Millar AH. Retrograde signalling caused by
889 heritable mitochondrial dysfunction is partially mediated by ANAC017 and improves plant
890 performance. **Plant J 2016**, 88(4)**:** 542-558.
891 55. Virdi KS, Laurie JD, Xu YZ, Yu J, Shao MR, Sanchez R, Kundariya H, Wang D, Riethoven
892 JJ, Wamboldt Y, Arrieta-Montiel MP, Shedge V, Mackenzie SA. *Arabidopsis* MSH1
893 mutation alters the epigenome and produces heritable changes in plant growth. **Nat**
894 **Commun 2015**, 6**:** 6386.
895 56. Virdi KS, Wamboldt Y, Kundariya H, Laurie JD, Keren I, Kumar KRS, Block A, Basset
896 G, Luebker S, Elowsky C, Day PM, Roose JL, Bricker TM, Elthon T, Mackenzie SA.
897 MSH1 Is a Plant Organellar DNA Binding and Thylakoid Protein under Precise Spatial
898 Regulation to Alter Development. **Mol Plant 2016**, 9(2)**:** 245-260.
899 57. Wallet C, Le Ret M, Bergdoll M, Bichara M, Dietrich A, Gualberto JM. The RECG1 DNA
900 Translocase Is a Key Factor in Recombination Surveillance, Repair, and Segregation of the
901 Mitochondrial DNA in Arabidopsis. **Plant Cell 2015**, 27(10)**:** 2907-2925.
902 58. Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, Dong Q, Ye
903 K. High-quality *Arabidopsis thaliana* Genome Assembly with Nanopore and HiFi Long
904 Reads. **Genomics Proteomics Bioinformatics 2021**.
905 59. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo*
906 genome assemblies. **Bioinformatics 2015**, 31(20)**:** 3350-3352.
907 60. Wiktor J, Gynna AH, Leroy P, Larsson J, Coceano G, Testa I, Elf J. RecA finds
908 homologous DNA by reduced dimensionality search. **Nature 2021**, 597(7876)**:** 426-429.

909 61. Wolfe KH, Li WH, Sharp PM. Rates of nucleotide substitution vary greatly among plant
910 mitochondrial, chloroplast, and nuclear DNAs. **Proc Natl Acad Sci U S A 1987**, 84(24)**:**
911 9054-9058.
912 62. Woloszynska M. Heteroplasmy and stoichiometric complexity of plant mitochondrial
913 genomes--though this be madness, yet there's method in't. **J Exp Bot 2010**, 61(3)**:** 657-671.
914 63. Wu Z, Liao X, Zhang X, Tembrock LR, Broz A. Genomic architectural variation of plant
915 mitochondria—A review of multichromosomal structuring. **J Syst Evol 2020a**, 60(1): 160-
916 168.
917 64. Wu Z, Waneka G, Broz AK, King CR, Sloan DB. *MSH1* is required for maintenance of
918 the low mutation rates in plant mitochondrial and plastid genomes. **Proc Natl Acad Sci U**
919 **S A 2020b**, 117(28)**:** 16448-16455.
920 65. Wynn E, Purfeerst E, Christensen A. Mitochondrial DNA Repair in an *Arabidopsis*
921 *thaliana* Uracil N-Glycosylase Mutant. **Plants (Basel) 2020**, 9(2).
922 66. Wynn EL, Christensen AC. Repeats of Unusual Size in Plant Mitochondrial Genomes:
923 Identification, Incidence and Evolution. **G3 (Bethesda) 2019**, 9(2)**:** 549-559.
924 67. Xu YZ, Arrieta-Montiel MP, Virdi KS, de Paula WB, Widhalm JR, Basset GJ, Davila JI,
925 Elthon TE, Elowsky CG, Sato SJ, Clemente TE, Mackenzie SA. MutS HOMOLOG1 is a
926 nucleoid protein that alters mitochondrial and plastid properties and plant response to high
927 light. **Plant Cell 2011**, 23(9)**:** 3428-3441.
928 68. Yang X, Mackenzie SA. Many Facets of Dynamic Plasticity in Plants. **Cold Spring Harb**
929 **Perspect Biol 2019**, 11(10).
930 69. Yang X, Sanchez R, Kundariya H, Maher T, Dopp I, Schwegel R, Virdi K, Axtell MJ,
931 Mackenzie SA. Segregation of an *MSH1* RNAi transgene produces heritable non-genetic
932 memory in association with methylome reprogramming. **Nat Commun 2020**, 11(1)**:** 2214.
933 70. Yu R, Sun C, Zhong Y, Liu Y, Sanchez-Puerta MV, Mower JP, Zhou R. The minicircular
934 and extremely heteroplasmic mitogenome of the holoparasitic plant *Rhopalocnemis*
935 *phalloides*. **Curr Biol 2022**, 32(2)**:** 470-479 e475.
936 71. Zampini E, Lepage E, Tremblay-Belzile S, Truche S, Brisson N. Organelle DNA
937 rearrangement mapping reveals U-turn-like inversions as a major source of genomic
938 instability in *Arabidopsis* and humans. **Genome Res 2015**, 25(5)**:** 645-654.
939 72. Zhao N, Li Z, Zhang L, Yang X, Mackenzie SA, Hu Z, Zhang M, Yang J. MutS
940 HOMOLOG1 mediates fertility reversion from cytoplasmic male sterile *Brassica juncea*
941 in response to environment. **Plant Cell Environ 2021**, 44(1)**:** 234-246.

942 73. Zhao N, Xu X, Wamboldt Y, Mackenzie SA, Yang X, Hu Z, Yang J, Zhang M. MutS
943 HOMOLOG1 silencing mediates *ORF220* substoichiometric shifting and causes male
944 sterility in *Brassica juncea*. **J Exp Bot 2016**, 67(1)**:** 435-444.
945
946
947 **Figure Legends**
948 **Figure 1** Mapping of HiFi reads identified structural variants in Arabidopsis organelle genomes
949 (a) Ratios of HiFi reads mapped to nuclear (nc), plastid (pt), mitochondrial (mt) genomes by
950 minimap2; un denotes the ratio of unmapped reads. (b) Schematics of reads alignment to the
951 reference genome. Reads mapped to two discontinuous positions indicates one-rearrangement
952 events. The overlapping region of two alignments indicated sequence homologies such as non-
953 tandem repeats. (c) Proportions of different types of mt genome rearrangements detected by HiFi
954 reads in wild-type and *msh1* mutant; ref, repeat, repeat_double, repeat_triple, and repeat_quadra
955 denote ratios of zero-, one-, two-, three-, four- rearrangement events; MMEJ, microhomology-
956 mediated end joining (2-49 bp homology sequence size); NHEJ, non-homologous end-joining (0-
957 1 bp homology sequence size); tandem, large-scale tandem repeats-mediated variants; insertion,
958 large-scale nucleotide insertions (mostly are poly-G and poly-C tracts). (d) Distribution of
959 overlapping region length of the two alignments in mitochondrial one-rearrangement events
960 detected by HiFi reads. Reads from different *msh1* individuals are merged. Read counts are
961 normalized to 10000 total reads. (e) Proportions of different types of pt genome rearrangements
962 detected by HiFi reads in wild-type and *msh1* mutant. (f) Distribution of overlapping length of the
963 two alignments in mitochondrial one-rearrangement events detected by HiFi reads. Reads from
964 different *msh1* individuals are merged. Read counts are normalized to 100000 total reads per sample.
965
966 **Figure 2** Repeat-mediated mt genome rearrangements identified by HiFi reads
967 (a, b) Global patterns of the junctions and relative read counts of mitochondrial one-rearrangement
968 events in (a) wild-type and (b) *msh1* mutant. In wild-type, most recombination events are mediated
969 by large repeats, and intermediate repeats Q, V near Large1-2. In *msh1*, recombination events
970 mediated by intermediate-sized repeats were common. Reads from different *msh1* individuals are
971 merged. There are 5544 one-rearrangement reads out of 24162 total reads in all *msh1* samples.
972 Colors denote repeat sizes. The sizes of dots denote read counts normalized to 10000 total reads
973 per sample. (c, d) Global patterns of the junctions and relative read counts of mitochondrial two-
974 rearrangement events in (c) wild-type and (d) *msh1* mutant. Reads from different *msh1* individuals
975 are merged. There are 1668 two-rearrangement reads out of 24162 total reads in all *msh1* samples.

976 Colors denote repeat sizes. The thickness of lines linking two dots denotes read counts normalized
977 to 10000 total reads per sample (line color is the same as the larger dot). (e) Ratios of zero-
978 rearrangement reads versus total reads indicate multiple highly dynamic regions across the mt
979 genome in wild-type and *msh1*. Reads from different *msh1* families are merged. Bottom is
980 annotations of repeats as in NC_037304.1.
981
982 **Figure 3** Continuous and discontinuous distribution of SNVs/indels in non-identical repeats
983 (a) Continuous change of SNVs in one-rearrangement HiFi reads indicate possible crossover site
984 distribution of recombination events between repeat CC-1 and CC-2 in *msh1*. Numbers in brackets
985 denote the sum of read counts in all samples without normalization. The shaded region denotes the
986 longest identical fragment in repeat CCs (b) The ratio of read counts between the longest identical
987 fragment was correlated with the ratio of length versus total length in each group of two-copy
988 repeats. Blue circles denote data points from non-identical two-copy repeats. Blue line denotes the
989 linear regression line. The diagonal line denotes equal ratios. (c) Discontinuous patterns of SNVs
990 in one-rearrangement chimeric reads of repeat H-1 and H-3 indicate partial gene conversion events
991 in *msh1*. Asterisks denote SNVs introduced by gene conversion from repeat homolog (black) H-1
992 or H-3, and (red) H-2. Numbers in brackets denote the sum of read counts in all samples without
993 normalization. Vertical bars in (a) and (c) denote A (green), T (red), G (brown), C (blue) for bases
994 and I (gray) for indel sites.
995
996 **Figure 4** Gene conversion events between imperfect repeats are suppressed by *MSH1*
997 (a) Schematic graph showing mapping of zero-rearrangement reads to two non-identical repeats
998 with one SNV. Both long and short reference reads are correctly mapped based on the SNV in the
999 repeated sequence. Long gene-conversion (GC) reads could be correctly mapped based on flanking
1000 sequences, enabling the identification of gene conversion events. Short GC reads failed to identify
1001 gene conversion events because of incorrect mapping to the other repeat homolog. (b,c) Proportions
1002 of gene conversion events at predicted sites in repeat group (b) H and (c) MMJS, which illustrated
1003 unsuppressed gene conversion events in samples of *msh1* compared to wild-type. The stacked bar
1004 height in each box indicated the ratio of a given base for A (green), T (red), G (brown), C (blue),
1005 and I (gray) for indel sites. Reference bases are indicated below. Results were calculated from zero-
1006 rearrangement reads. SNVs and indels are arranged from small to large coordinates in the reference.
1007 Coordinates of the first and last SNVs/INDs (black triangles) are shown. H-2 and H-3 are in the
1008 reverse orientation compared to H-1, and MMJS-2 is in the reverse orientation compared to MMJS-

1009 1. Asterisks in (b) denote predicted gene conversion sites corresponding to two other repeat
1010 homologs.
1011
1012 **Figure 5** Phenotypic correlations of mt genome variants in *msh1*
1013 (a) Photos showing wild-type and *msh1* seedlings of different families at 5 days after sowing (DAS).
1014 (b) Distributions of primary root length frequency of F_3 families at 5 DAS indicate delayed seed
1015 germination ($120 > n \geq 100$). (c-e) Normalized per-base read coverage of mt genome regions: (c)
1016 49201-51200, (d) 263901-265900, (e) 270501-271500. Blue arrows denote coding regions of *atp6-*
1017 *1*, *atp6-2* and *atp9*. Purple arrows denote tRNA and ORF genes. Orange arrows denote different
1018 mitochondrial repeats. Annotations are based on the GenBank accession NC_037304.1. Purple and
1019 orange lines denote normalized per-base read coverage of M3-17 and M3-43 mutant individuals.
1020 Green lines denote normalized per-base read coverage of W3-5-2, and the published datasets WT-
1021 XJTU and WT-CEN. Red horizontal lines in c to e denote 100% coverage. The white bar in (a)
1022 represents 10 mm.
1023
1024 **Figure 6** Loss of mt genome regions in other *Arabidopsis thaliana* accessions
1025 (A-E) Normalized per-base read coverage of mt genome regions: (A) 49201-51200, (B) 263901-
1026 265900, (C) 270501-271500, (D) 19201-21200, (E) 345801-347800. Blue arrows denote coding
1027 regions of *atp6-1*, *atp6-2*, *atp9*, *rps3*, and *rps19*. Purple arrows denote tRNA and ORF genes.
1028 Orange arrows denote different mitochondrial repeats. Annotations are based on the GenBank file
1029 NC_037304.1. Green lines denote normalized per-base read coverage of published Col-0 CLR
1030 dataset. Purple lines denote normalized per-base read coverage of published Eri-1 and C24 datasets.
1031 Orange lines denote normalized per-base read coverage of published Cvi-0, An-1, L*er*, Kyoto and
1032 Sha datasets. Red horizontal lines in C-E denote 100% coverage.
1033
1034 **Figure 7** Mutations and structural variants of in pt genome detected by HiFi reads
1035 (A) A MMEJ event mediated by a partial palindromic sequence (U-8) in *msh1*. A partial
1036 palindromic sequence is also a specific pair of inverse repeats or microhomologies at the same
1037 location. Numbers in brackets denote the sum of read counts showing crossover site distribution in
1038 chimeric reads in all samples without normalization. The ratio of one-rearrangement reads versus
1039 zero-rearrangement reads across U-8 is 160/14786 (1.1% of total reads) in *msh1* and 0/1966 in
1040 wild-type. (B) A NHEJ event occurring only in the M3-43-5 individual was predicted to disrupt the
1041 *petA* gene. The ratio of one-rearrangement reads versus zero-rearrangement reads across this
1042 deletion is 121/1385 (8.0% of total reads). (C) Proportions of AT-to-GC transitions in 17 GATK-

1043 called SNVs in wild-type and *msh1* individuals, which were enriched in one or a few samples at
1044 high ratios. The stacked bar height in each box indicated the ratio of a given base for A (green), T
1045 (red), G (brown), and C (blue). Reference bases are indicated below. (D) Proportions of an IND
1046 mutation caused by a 13-bp homopolymer started near position 67164 with wild-type showing the
1047 larger proportion of reference genotype. Multiple indel sizes (ranging from 6-bp deletion to 18-bp
1048 insertion) were calculated from raw BAM-alignment files by the pysam package. (E) Colors
1049 indicate ratios of variant genotypes of individual samples minus the mean ratios of variant
1050 genotypes of three wild-type samples of the 85 GATK-called indel mutations. The proportions of
1051 variant genotype at most positions were much higher in *msh1* individuals compared with wild-type,
1052 which supported the previously proposed role of *MSH1* in maintaining low mutation rates to some
1053 extent.
1054
1055

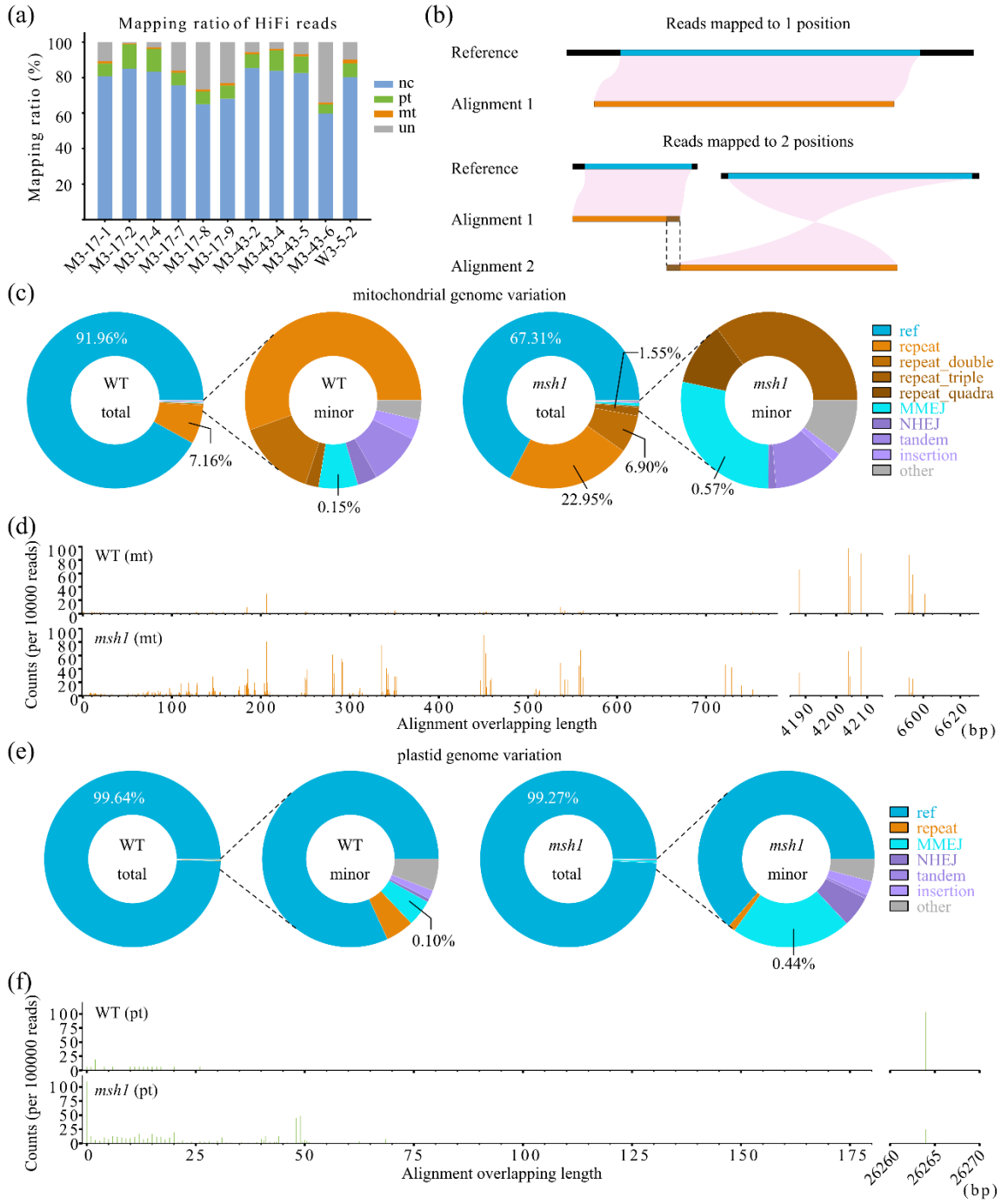

Figure 1 Mapping of HiFi reads identified structural variants in Arabidopsis organelle genomes

1056

1057

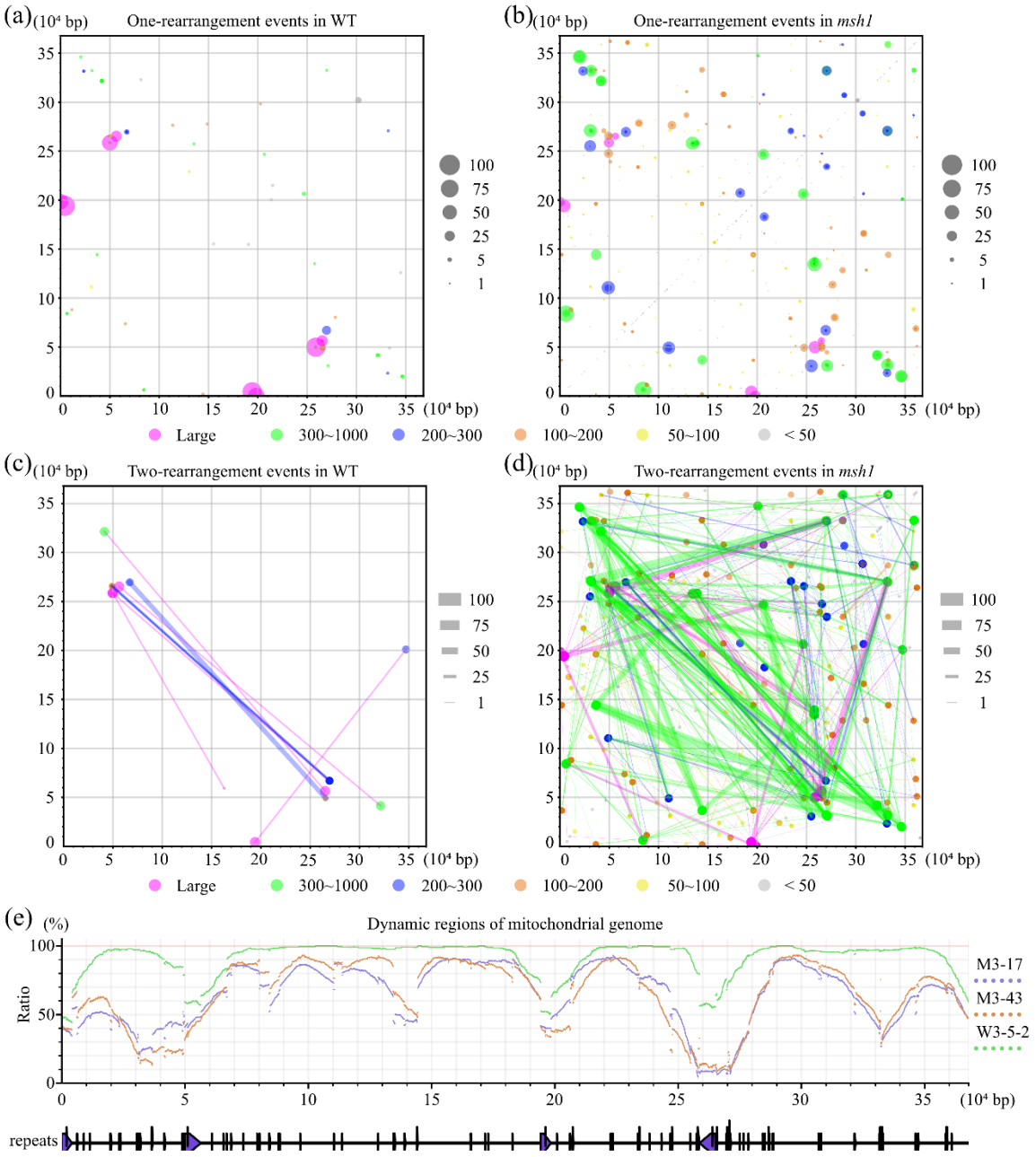

Figure 2 Repeat-mediated mt genome rearrangements identified by HiFi reads

1058

1059

1060

1061

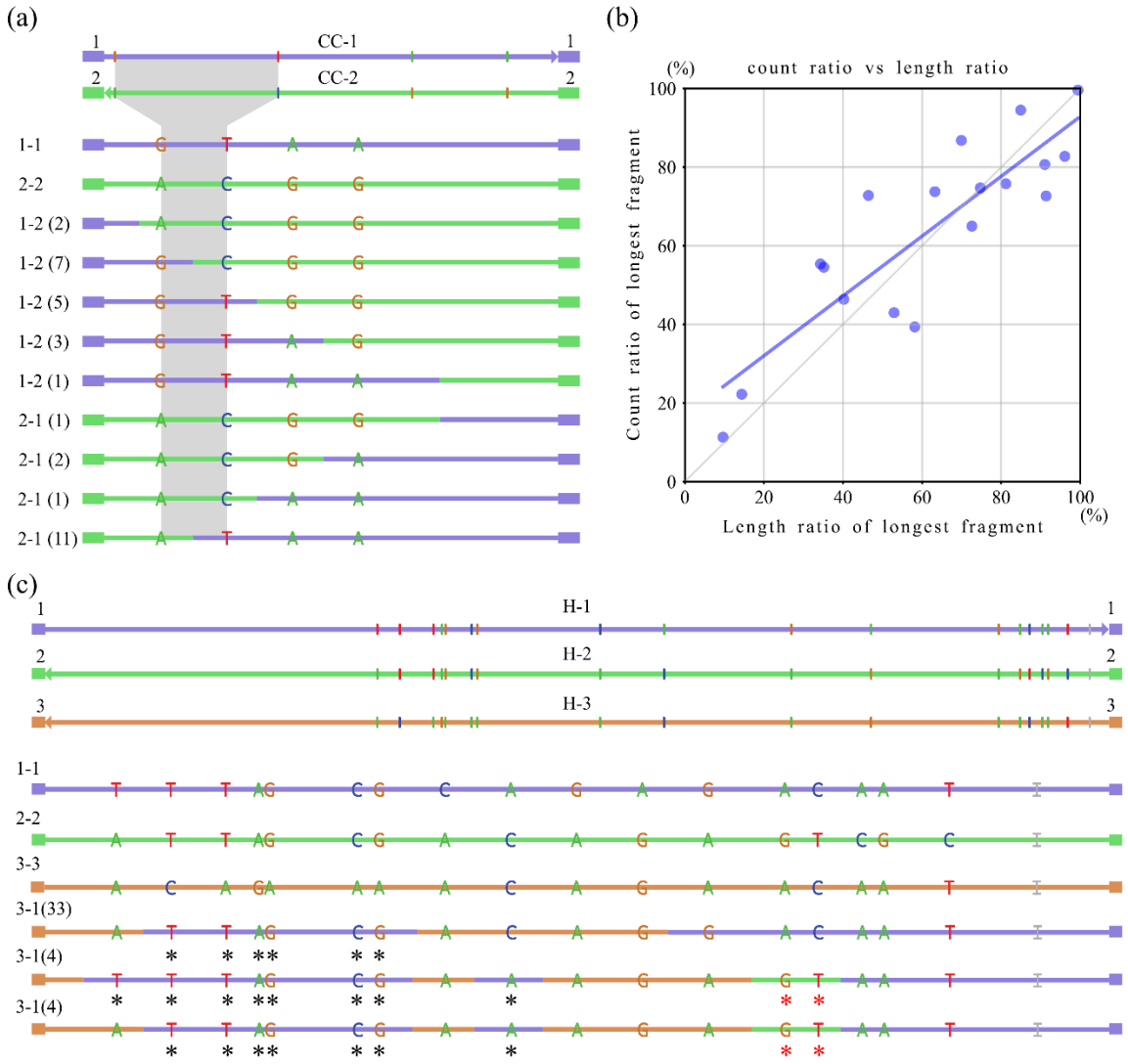

Figure 3 Continuous and discontinuous distribution of SNVs/indels in non-identical repeats

1062

1063

1064

1065

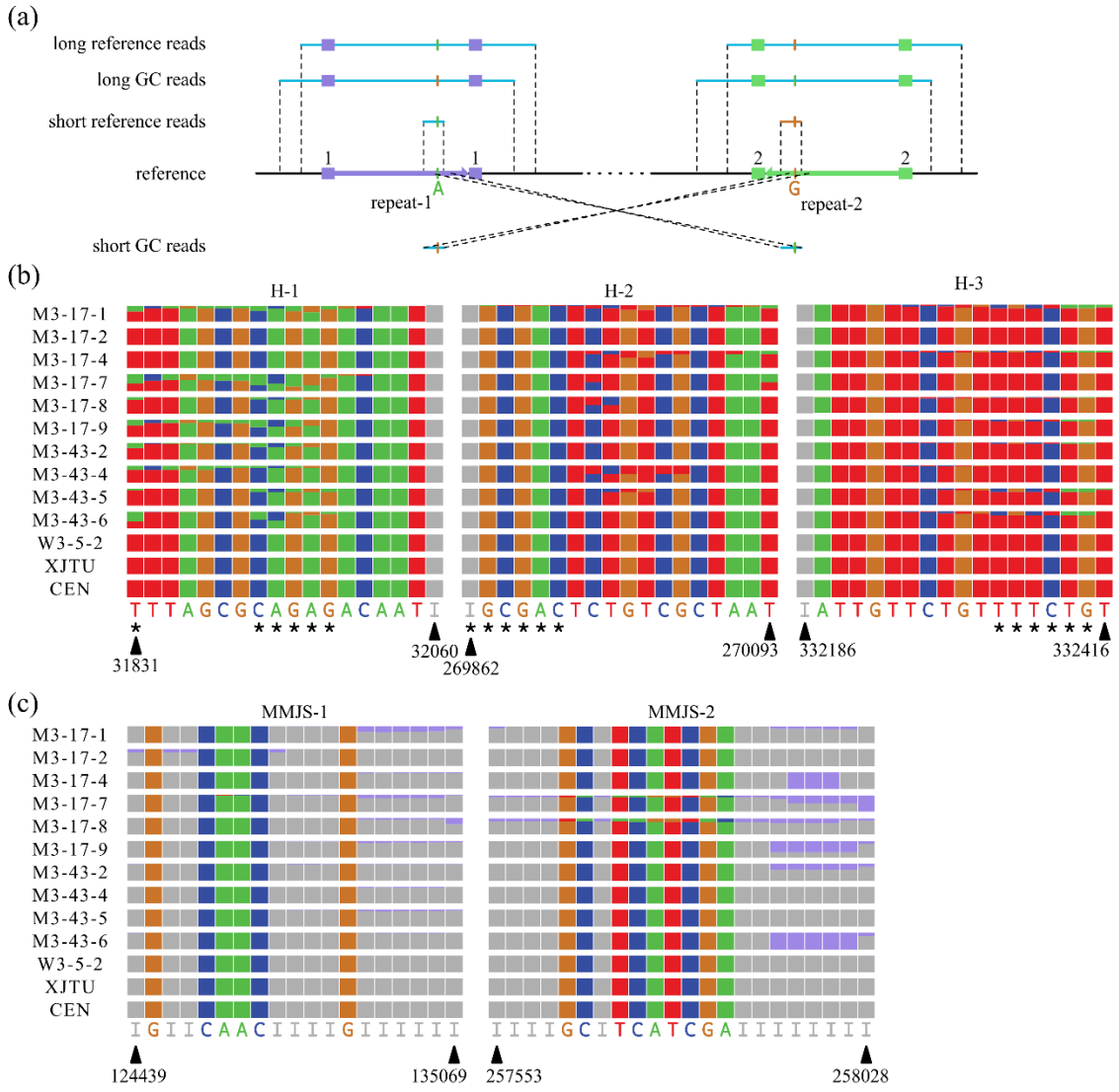

Figure 4 Gene conversion events between imperfect repeats are suppressed by *MSH1*

1066

1067

1068

1069

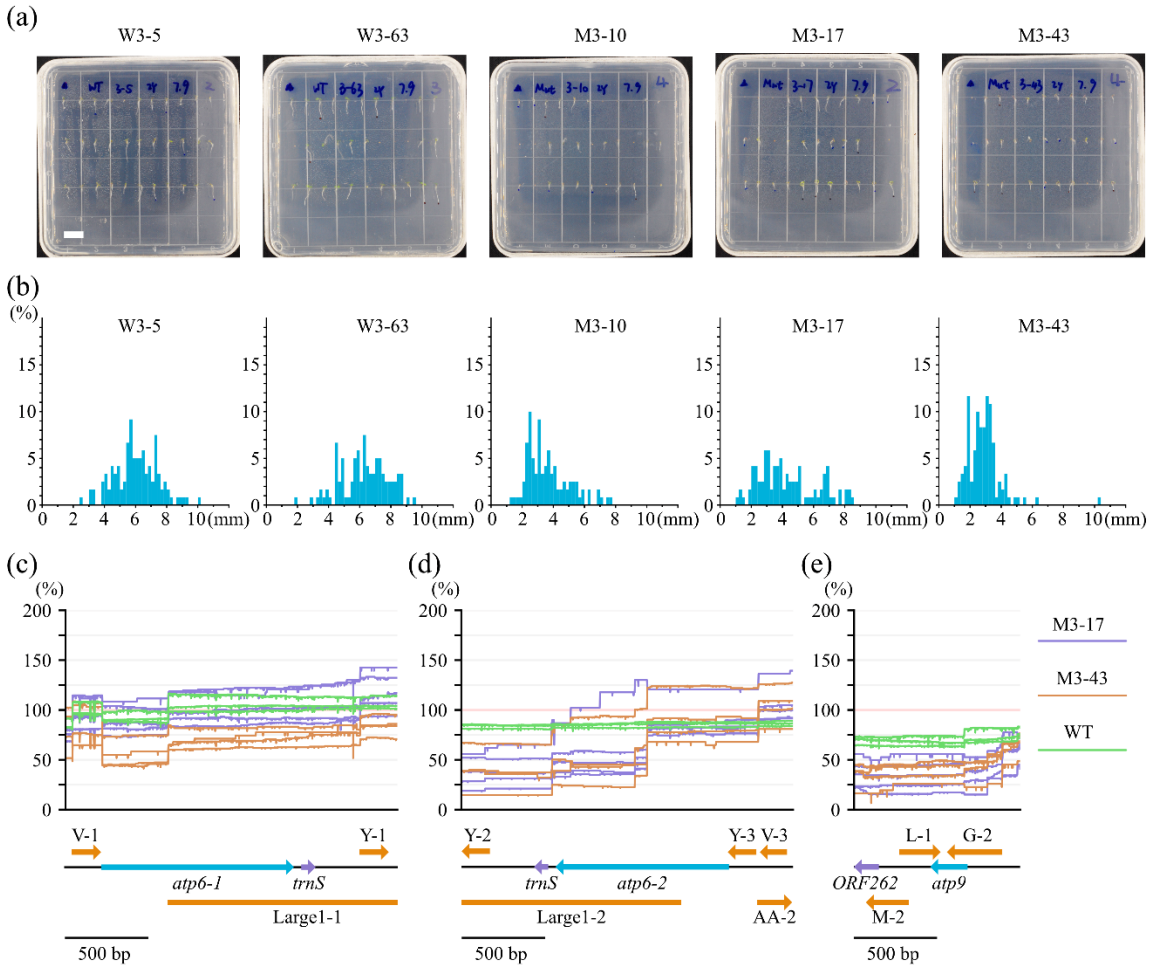

Figure 5 Phenotypic correlations of mt genome variants in *msh1*

1070
1071
1072
1073

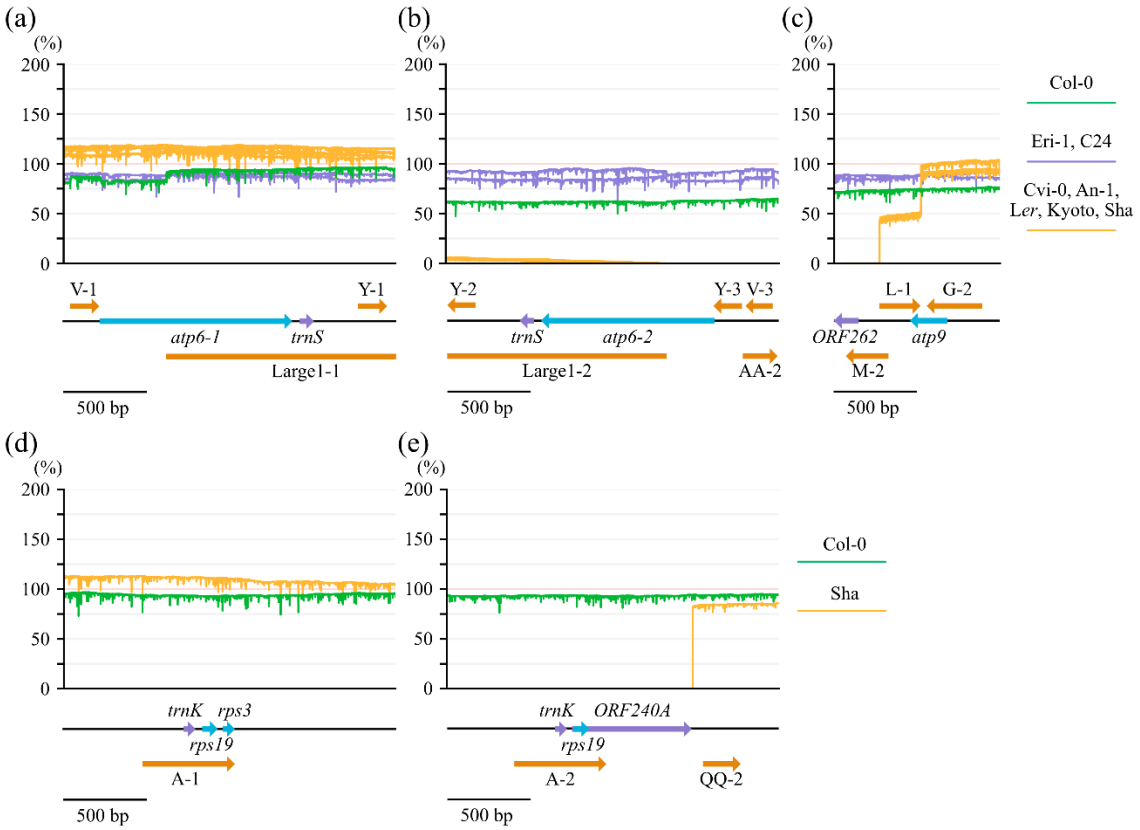

Figure 6 Loss of mt genome regions in other *Arabidopsis thaliana* accessions

1074

1075

1076

1077

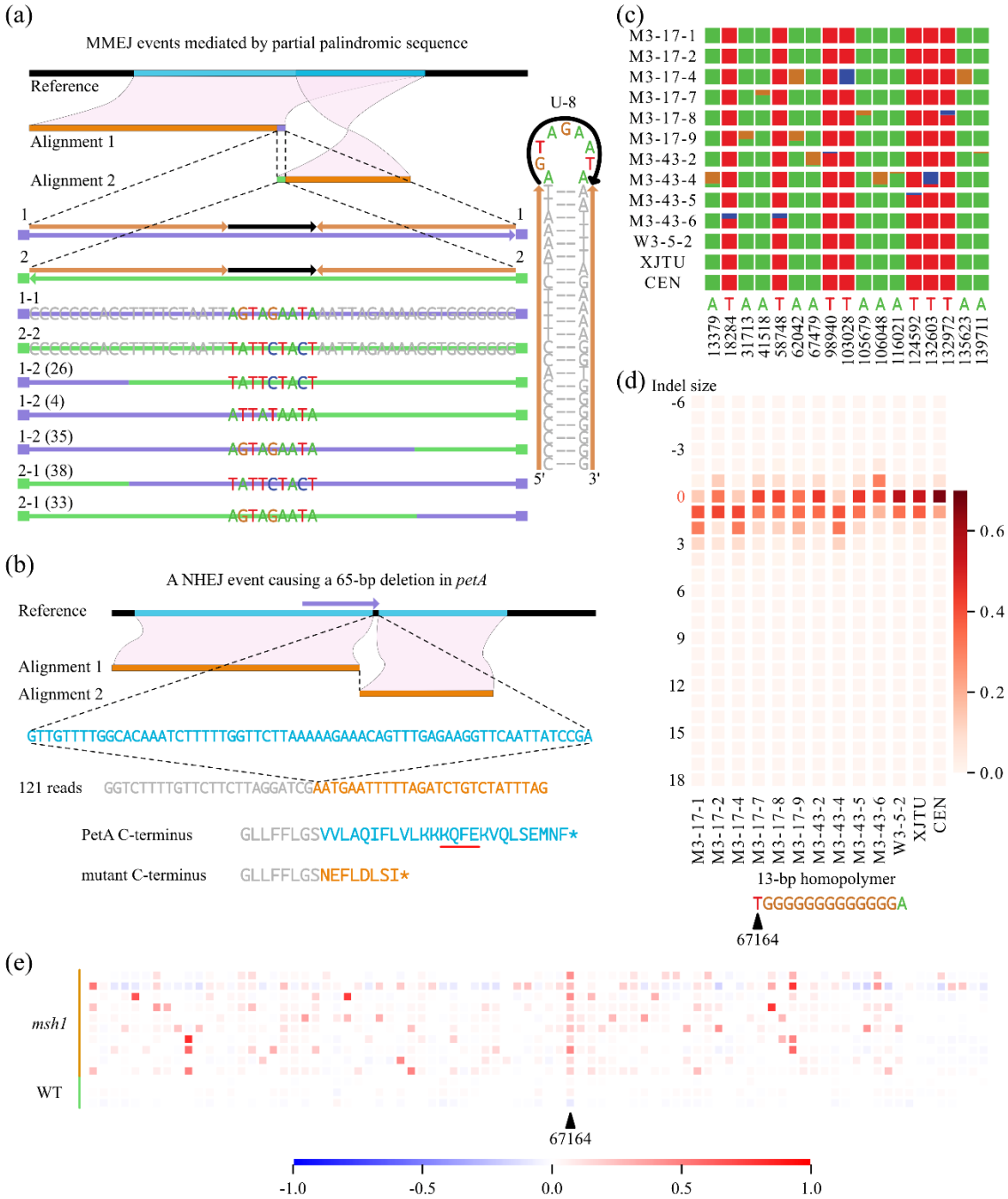

Figure 7 Mutations and structural variants of in pt genome detected by HiFi reads

1078

1079