Long-read sequencing characterizes mitochondrial and plastid genome variants in Arabidopsis *msh1* mutants

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 provided a global view of recombination dynamics with detailed quantification of parental and crossover recombination products for both large and small repeats. We found that recombination breakpoints were distributed relatively evenly across the length of repeated sequences and detected widespread internal exchanges of sequence variants between pairs of imperfect repeats in the mitochondrial genome of *msh1* mutants. Long-read assemblies of mitochondrial genomes from seven other *Arabidopsis thaliana* wild-type accessions differed by repeat-mediated structural rearrangements similar to those observed in *msh1* mutants, but they were all in a simple low-heteroplasmy state. 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 organellar genomes and improved the sequence resolution for recombinational processes suppressed by MSH1.

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

Significance

Plant organellar genomes can undergo rapid rearrangements. Long-read sequencing provides a detailed and quantitative view of mitochondrial and plastid genome variants normally suppressed by MSH1, advancing our understanding of plant organellar genome dynamics.

Introduction

Mitochondria and plastids are major energy converters as well as central metabolic compartments of plant cells. In plants, mitochondrial (mt) and plastid (pt) genomes have followed very different evolutionary trajectories. Compared with the rather conserved structures of pt genomes (Lee et al., 2016), plant mt genomes exhibit dramatic variation among species (Wu et al., 2022). Although coding sequences in angiosperm mt genome have very low substitution rates, their gene order and

68 intergenic sequences are rarely conserved (Wolfe et al., 1987; Palmer and Herbon 1988;
69 Christensen, 2021). Angiosperm mt genomes typically contain numerous non-tandem repeats, and
70 recombination between different copies of repeat sequences causes a high level of structural
71 variation (Wynn and Christensen, 2019). These repeats show little or no sequence conservation
72 between species and are often grouped based on size ranges (Arrieta-Montiel et al., 2009). Large
73 repeats (≥1000 bp) are usually identical in sequence and recombine at high frequency, leading to
74 isomerization of the genome (Unseld et al., 1997). The recombination frequencies of intermediate-
75 sized repeats (50-1000 bp) are lower but sensitive to genetic, environmental and developmental
76 triggers that make them frequent enough to have important phenotypic effects (Gualberto and
77 Newton 2017; Mackenzie and Kundariya, 2020). Shorter repeats (<50 bp) typically do not show
78 recombination activity, but still play important roles during evolution. These short sequences are
79 often described as "microhomologies", but the very shortest of them (e.g., <10 bp) are unlikely to
80 reflect true homology, because such similarities can easily occur at random. Despite their
81 importance, repeats in sequenced mt genomes are not always annotated correctly or completely
82 even in model plant species (Wynn and Christensen, 2019). Moreover, these repeats may produce
83 complex assembly graphs, which hinder a clear characterization of mt genomes (Masutani et al.,
84 2021). Thus, mechanisms underlining the repeats and their roles in destabilizing plant mt genomes
85 require a more thorough exploration.
86 Recombination in plant mt genomes is likely involved in repairing DNA damage, such as double-
87 strand breaks (DSBs), which can disrupt genome integrity and prevent the progression of
88 replication and transcription activities. To deal with DSBs, two main branches of repair systems
89 exist: homology-directed repair (HDR) and non-homologous end joining (NHEJ) (Hussmann et al.,
90 2021). HDR requires homologous dsDNA templates for new strand synthesis. HDR involving both
91 DSB ends will lead to non-crossover repair products, which is a process known as gene conversion.
92 In contrast, HDR involving only one DSB end and non-allelic templates is known as break-induced
93 repair and leads to crossover products (Christensen, 2013).
94 Nuclear genes encode a set of mt- and mt-targeted DNA binding proteins that are involved in
95 DNA recombination, replication, and repair (RRR; Gualberto and Newton, 2017; Chevignv et al.,
96 2020; Fuchs et al., 2020). Some RRR genes, such as those in the *RECA*, *POLI*, and *SSB* families,
97 have conserved roles, while others are evolutionary innovations. Unlike bacterial MutS
98 counterparts, the dual-targeted (mt and pt) MSH1 enzyme in plants combines a mismatch-
99 recognition domain (MutS I) and a predicted DSB-generating GIY-YIG endonuclease domain in a
100 single protein (Abdelnoor et al., 2006). MSH1 has been hypothesized to participate in novel DNA

101 repair mechanisms in plant organelles (Christensen, 2013; Ayala-Garcia et al., 2018; Fukui et al,
102 2018; Wu et al., 2020; Wynn et al., 2020).

103 Arabidopsis *msh1* (=*chm*) mutants were originally isolated based on a green-white leaf
104 variegation phenotype (Rédei, 1973). Early studies focused on reproducible mt genome
105 rearrangements at a specific region near the *atp9* gene in *msh1* mutants and explored processes
106 such as asymmetric recombination and substoichiometric shifting (Martínez-Zapater et al., 1992;
107 Sakamoto et al., 1996; Abdelnoor et al., 2003; Shedge et al., 2007). They showed that new bands
108 in DNA gel blots in *msh1* mutants were similar to bands observed in other Arabidopsis accessions,
109 which existed at substoichiometric levels in wild-type individuals and could be detected by PCR.
110 Detecting pt genome rearrangements in *msh1* mutants by DNA gel blots was more difficult and
111 only successful in dissected white tissues using probes to detect rearrangements already known in
112 another mutant (Xu et al., 2011). The more limited structural effects in the Arabidopsis pt genome
113 may reflect a relative lack of dispersed repeats. An analysis of *msh1* mutants in the moss
114 *Physcomitrella patens* has revealed more extensive pt genome variation associated with repeat-
115 mediated recombination (Odahara et al., 2017).

116 The first systematic analysis across the mt genome in *msh1* mutants confirmed asymmetric
117 recombination at 23 groups of repeats using probes corresponding to 33 near-perfect intermediate-
118 sized repeats which were inferred from blast analysis of the Arabidopsis C24 mt genome (Arrieta-
119 Montiel et al., 2009). The intensity of bands detected by a specific probe varied among *msh1*
120 individuals and generations. In some cases, weak bands corresponding to secondary recombination
121 (multiple rearrangements) and loss of parental bands were observed (Shedge et al., 2007; Arrieta-
122 Montiel et al., 2009). It was also reported that lower *MSH1* expression caused mt rearrangements
123 that produced cytoplasmic male sterile (CMS) lines in tomato and allowed fertility reversion of
124 lines of *Brassica juncea* (Sandhu et al., 2007; Zhao et al., 2016, 2021). These studies demonstrated
125 that MSH1 maintains the structural integrity of the mt genome through suppressing illegitimate
126 recombination at repeats. We also recently showed increased frequencies of de novo point
127 mutations in organelle genomes of *msh1* mutants, suggesting a role of MSH1 in maintaining lower
128 mutation rates in organellar genomes compared with the nuclear genome (Wu et al., 2020).

129 Despite this extensive work on *MSH1* in Arabidopsis, we still lack a complete understanding of
130 its effect on repeat dynamics, in part due to the technical difficulties of traditional DNA gel blots,
131 quantitative PCR, and short-read sequencing. For example, a previous study classified mt genome
132 rearrangements in Arabidopsis *msh1* mutants using early-generation Illumina sequencing (Davila
133 et al. 2011). This work represented a key advance in the genome-wide characterization of mt
134 recombinational activity, including sequence-level resolution of some recombinant molecules.

135 However, the short read-lengths (36-bp) available at the time cannot span entire repeats including
136 the regions of perfect sequence identity that often separate internal variants in imperfect repeats,
137 placing significant limitations on analysis of recombination breakpoints. Short-read sequencing
138 also precludes detection of more complex structural rearrangements resulting from recombination
139 at multiple different repeats. Davila et al. (2011) did not detect recombinationally active repeats
140 smaller than 50 bp. They proposed that structural differences among Arabidopsis accessions were
141 mainly the result of repeat-mediated recombination, microhomology-mediated end joining
142 (MMEJ), and NHEJ, but they did not detect MMEJ or NHEJ activities with short reads in *msh1*
143 mutants. Davila et al. (2011) also documented discontinuous patterns of single-nucleotide variants
144 (SNVs) and indels that were exchanged between pairs of imperfect repeats. Once again, however,
145 the resolution of this important analysis was limited due to short read lengths, and it was only
146 applied to a subset of repeat pairs. Fortunately, the advent of long-read sequencing technologies
147 has created the opportunity to analyze structural variants in plant mt genome assemblies with much
148 more accuracy and scalability (Kozik et al., 2019; Yu et al., 2022).
149 In this work, we aim to comprehensively investigate mt and pt genome instabilities in
150 Arabidopsis *msh1* mutants. Using PacBio high-fidelity (HiFi) data for individual plants, we
151 identified structural variants resulting from one or more repeat-mediated rearrangements, as well
152 as lower frequency MMEJ and NHEJ events. We provide a global map of mt genome instabilities
153 with a more detailed and accurate characterization of different repeats. For imperfect repeats, we
154 systematically analyzed the distribution of breakpoints and observed the exchange of sequence
155 variants in molecules that did not exhibit any structural rearrangements, which was an unexpected
156 process suppressed by MSH1. We also show that mt genomes of seven *Arabidopsis thaliana*
157 accessions are all in a low-heteroplasmy states, despite differing by the same types of
158 rearrangements observed in *msh1* mutants. The total structural variation was very low in pt
159 genomes. But some variants show individual-specific accumulation in *msh1* mutants, suggesting a
160 strong influence of random heteroplasmic sorting. These results are important for improving mt
161 and pt genome-editing technologies and understanding the distinct patterns of evolution in these
162 organellar genomes.
163
164 **Results**
165 **Mapping of HiFi reads identified structural variants in Arabidopsis mitochondrial genomes**
166 We generated F_3 families from a cross between wild type Arabidopsis Columbia-0 (Col-0) and an
167 *msh1* mutant line (*chm1-1*, CS3372) in an earlier project (Wu et al., 2020). To analyze structural
168 rearrangements in the organellar genomes in homozygous *msh1* mutants, we collected aboveground

169 tissue from 11 individual plants (10 *msh1* individuals: M3-17-1, M3-17-2, M3-17-4, M3-17-7, M3-
170 17-8, M3-17-9, M3-43-2, M3-43-4, M3-43-5, M3-43-6; 1 wild-type individual: W3-5-2) from 3
171 different F_3 families (Table S1; Figure S1) and prepared separate libraries for each plant for PacBio
172 HiFi sequencing. The libraries generated a total of 34.36 Gb of HiFi reads (12.7-18.1 kb mean read
173 length). These reads were first mapped to the Col-0 reference genome by minimap2 (Li, 2018) to
174 separate nuclear, pt and mt proportions. After excluding one outlier with very low yield (M3-17-
175 2), the coverage of the nuclear genome ranged from 8.5× to 46.4×, the coverage of the pt genome
176 ranged from 899.0× to 3093.8×, and the coverage of the mt genome ranged from 30.9× to 247.9×
177 (Table S2). The pt (4.96%-12.85%) and mt (0.93%-2.13%) proportions were used for downstream
178 analysis (Figure S2; Table S3 and S4).
179 Published mt genomes are usually represented as a linearized sequence of the "master circle".
180 However, various sub-stoichiometric isoforms exist due to frequent repeat-mediated structural
181 rearrangements (Kozik et al., 2019). Structural rearrangements are indicated by reads whose parts
182 map to discontinuous positions in the reference genome (Figure 1a). We developed a pipeline to
183 detect these structural rearrangements (available at https://github.com/zouyinstein/hifisr), in which
184 mt genome-mapped reads were aligned against the reference sequence (NC_037304.1; Sloan et al.,
185 2018) via the NCBI blastn software. Using custom Python scripts, we grouped the reads by
186 discontinuity and mapping coordinates, and calculated alignment overlap (AO) length (i.e., repeat
187 length) and read counts for each group (Table S5). After additional manual checking and curation,
188 we were able to distinguish reads with no rearrangements (zero-rearrangement reads) from those
189 containing one, two, three, or four repeat-mediated rearrangements (≥50 bp) in a single read (Figure
190 1b; Table S6). In wild type, 7.50% of total reads accounted for rearrangements mediated by non-
191 tandem repeats, most of which were associated with the two large repeat pairs (Figure 1c). In *msh1*
192 mutants, this ratio increased to 31.63% of total reads, containing increased proportions of reads
193 with rearrangements mediated by intermediate-sized repeats and reads with multiple
194 rearrangements (Figure 1b,c; Table S7 and S8). A few low-frequency events were also detected,
195 including MMEJ (AO length between 2 and 49 bp) and NHEJ (AO length of 0 or 1 bp), replication
196 slippage associated with tandem repeats, and insertions (poly-G/C tracts in most cases) (Figure 1b;
197 Table S6). In particular, the proportion of MMEJ events also increased in *msh1* mutants compared
198 with wild-type (Figure 1b). Taken together, these data indicate that mt genome structural variants
199 in *msh1* mutants is mainly (> 96.7%) associated with non-tandem repeats. Compared to short-read
200 sequencing, our results open the opportunity for investigating multiple rearrangements in a single
201 read and low-frequency categories of structural variants.
202

203 **Global patterns of mt genome variants in *msh1* mutants**

204 By visualizing all the junctions and read counts of one-rearrangement reads, we found that
205 rearrangements in wild-type individuals were mainly associated with the Large1, Large2, Q, and V
206 repeats, as well as a few other repeats at very low frequency (Figure S3). In *msh1* mutants, the
207 activity of increased mt genome rearrangements was not randomly distributed (Figure 2a). Instead,
208 they were mostly associated with the subset of repeats that were named and annotated in previous
209 studies with the most extreme variability observed at three regions containing large repeats and two
210 additional regions with intermediate-sized repeats (Figure S4; Arrieta-Montiel et al., 2009; Davila
211 et al., 2011; Sloan et al., 2018). With more exhaustive blastn parameters, we found that many
212 smaller repeats that were not previously named were also associated with mt genome
213 rearrangements (smaller unlabeled dots in Figure 2a and S3; Table S9). However, rearrangements
214 mediated by these small repeats were supported by very few reads and were detected in fewer
215 samples (Figure 2b,c).

216 Unlike previous DNA gel blots or PCR-based analysis, we could compare the activities of
217 different repeats. According to relative recombination frequencies, 18 groups (Large1, Large2,
218 MMJS, A, B, C, D, E, F, G, N, I, L, Q, R, H, V, BB) showed high activity (read count ≥ 50 in *msh1*,
219 normalized to 10000 total mt genome-mapped reads), 17 groups (P, O, QQ, M, K, X, U, T, W, CC,
220 Z, DD, Y, LL, HNR37, KK, OO) showed intermediate activity (read count of 10 to 49 in *msh1*)
221 and the 64 remaining groups showed low activity (Figure S5a; Table S10-S13). The factors that
222 contributed to different recombination frequencies were not limited to repeat length, identity or
223 copy number, because some repeats still fell far from the regression line when we considered all
224 two-copy repeats with high sequence identity (mismatches + gaps ≤ 2; Figure S5b and Table S14)
225 and different pairs within the same three-copy repeat group showed different recombination
226 frequencies (Table S15).

227 Asymmetrical recombination in *msh1* mutants was first reported based on DNA gel blot analysis
228 and was also estimated based on read depth across one of the repeat boundaries (Arrieta-Montiel
229 et al., 2009; Davila et al., 2011). Using long reads across both repeat boundaries for a given repeat
230 group, we could estimate asymmetrical recombination events in more detail (Figure S6). The
231 polarity of asymmetrical recombination between the two *msh1* mutant families was largely
232 conserved, with small differences in intensity and accompanied loss of parental forms for some
233 repeat groups which could be caused by heteroplasmic sorting (Figure 3; Table S16). Notably, we
234 observed more pronounced asymmetrical recombination and severe loss of one parental form (2-2)
235 for Large1 in *msh1* mutants compared with wild-type (Figure 3). Such large repeats are typically
236 found to recombine symmetrically in the literature (Figure 3; Arrieta-Montiel et al., 2009).

237 Long reads enabled the identification of rearrangements that involve two or more groups of
238 repeats. For example, the increased two-rearrangement reads involving Large1 and Large2
239 indicated more fragmented and heteroplasmic mt genomes in *msh1* mutants compared with wild-
240 type (Figure S7). Together, our data provide a much more detailed and accurate view of mt genome
241 instabilities of *msh1* mutants with a quantitative comparison of different repeats.
242
243 **Base-level understanding of crossover and non-crossover recombination events in *msh1***
244 We investigated the internal single-nucleotide variants (SNVs) and indels for the 27 groups of high-
245 and intermediate-activity repeats that were not 100% identical in sequence (i.e., after excluding the
246 perfect repeats Large1, Large2, C, F, K, L, Q, and Z). Conducting such a comprehensive analysis
247 at a single-molecule level is not possible with short-read sequencing. We first examined the
248 distribution of crossover sites in one-rearrangement chimeric reads based on the SNVs/indels that
249 distinguish repeat homologs. We found that recombination breakpoints were most likely to occur
250 in the longest identical fragment (LIF) within two-copy repeats and that the proportion of
251 supporting reads was positively correlated with LIF length as a fraction of total repeat length
252 (Figure 4a; Figure S8; Table S17). This indicates that recombination breakpoints are distributed
253 relatively evenly across the length of repeats rather than being localized to a single hotspot.
254 An earlier study also observed the continuous distribution of SNVs/indels in imperfect repeats I,
255 R and X (Miller-Messmer et al 2012). In our analysis, we also identified discontinuous patterns of
256 SNVs/indels across the length of imperfect repeats in some reads (Table S17). For example, a
257 proportion of the one-rearrangement reads mediated by repeats H-1 and H-3 also contained variants
258 that derived from repeat H-2, suggesting that discontinuous SNV patterns could be related to other
259 repeat homologs not directly involved in a crossover recombination event (Figure 4c). A previous
260 study used the term "partial gene conversion" and suggested mismatch repair mechanism
261 independent of *MSH1* as a possible explanation for discontinuous SNV/indel patterns (Davila et al.,
262 2011).
263 The patterns of mt genome recombination indicate that dsDNA templates for DSB repair can be
264 allelic or non-allelic, and recombination products can be both crossover and non-crossover (Figure
265 4b). HiFi technology enabled unambiguous mapping of reads that did not contain any
266 rearrangements but still exhibited variants within repeats (Figure S9). We investigated the exchange
267 of non-allelic SNVs/indels in zero-rearrangement reads spanning the entire length of all 27 high-
268 and intermediate-activity groups of imperfect repeats. We observed numerous mosaic repeat
269 sequences apparently generated through exchange of non-allelic SNVs/indels in *msh1* mutants,
270 which seldomly occurred in wild-type (Table S18-S20). This suggested that without MSH1, non-

271 allelic dsDNA templates were used in non-crossover recombination. In *msh1* mutants, repeat
272 groups A, D, G, H, I, M, MMJS, N, R, T, and W showed a high frequency of non-allelic
273 SNVs/indels (Figure 4d and S10) and had higher frequencies of crossover recombination (Table
274 S16), while the remaining 15 analyzed repeat groups showed low frequencies of non-allelic
275 SNVs/indels (Table S18; note that E-1 and E-2 are not included in this analysis because the same
276 SNV/indel positions are included within the H repeat group). It appears that exchanges can involve
277 very short stretches because the ratios of variants at adjacent positions often differ.
278 The discontinuous patterns of SNVs/indels in one-rearrangement reads such as the examples
279 described for H-1 and H-3 above could also result from the recombination products of zero-
280 rearrangement templates that have previously undergone these internal exchanges. To investigate
281 this possibility, we compared samples with high (M3-17-1, 16.2%) and low (M3-43-2, 0%) ratios
282 of exchanged zero-rearrangement reads at two positions (the asterisk-labeled A and C in Figure 4d).
283 As predicted, we found the proportion of crossover products between repeat H-1 and H-3 that
284 contained the H-2-specific SNVs (G and T) was also higher in sample M3-17-1 (20.1%) than M3-
285 43-2 (2.0%). The ratios of the same variants in repeat group H were different among *msh1*
286 individuals, which might reflect the process of heteroplasmic sorting (Figure 4d and S11). But in
287 repeat homolog MMJS-2, the unusually high ratios of non-allelic indels in some individuals (M3-
288 17-4 and M3-43-6) were associated with the near complete loss of parental forms, i.e., very low
289 sequence coverage in this region (Figure S10 and S12). Together, these sequence-level
290 investigations suggest that *MSH1* can suppresses the exchange of non-allelic SNVs/indels between
291 imperfect repeats within the mt genome in addition to suppressing structural variation.
292
293 **Mt genome evolution in Arabidopsis involves reproducible sets of repeats**
294 We used metaFlye to reassemble the mt genomes of seven other *Arabidopsis thaliana* accessions
295 using published PacBio continuous long-read (CLR) datasets (Jiao and Schneeberger, 2020). We
296 found that all the metaFlye assembly graphs were (near) completely resolved, which suggests low
297 heteroplasmy in all accessions (Figure S13; Table S21 and S22). In contrast, the graphs were much
298 more complex with fragmented contigs and many unresolved repeats in *msh1* mutants, reflecting
299 highly rearranged and heteroplasmic mt genomes (Figure S14a). The graphs of wild-type Col-0
300 based on published HiFi datasets (Wang et al., 2021; Naish et al., 2021) were still relatively simple
301 (Figure S14b), which indicated that the complexity of mt genomes of *msh1* mutants was not caused
302 by higher sequence depth.
303 Even though the in vivo physical structures of mtDNA molecules are mostly branched linear
304 forms and circular molecules much smaller than the expected size of a complete master circle

305 (Bendich et al., 1993; Kozik et al., 2019), the "master circle" models facilitate easier comparison
306 and clearer presentation of structural variants. When comparing the pseudo-master circles of
307 multiple accessions with Col-0 mt genome, we identified structural rearrangements associated with
308 specific non-tandem repeats, MMEJ/NHEJ and tandem repeats, as well as some large indels flanked
309 by repeats or microhomologies (Figure S15; Table S23 and S24). We noticed that the repeats that
310 mediated rearrangements, and the repeats or microhomologies flanking the large insertions were
311 often shared by different accessions (Figure S15; Table S24). For example, the structures of mt
312 genomes were strikingly similar in An-1 and L*er*. The mt genomes of Col-0 and An-1 differ by five
313 rearrangements mediated by repeat group B, I, H-1/2, L and H-1/3, while that of L*er* has an
314 additional rearrangement mediated by repeat group LL (Figure 5a). The small circles generated by
315 closely located direct repeats (H-1/2, H-1/3 and LL) were lost during evolution, but in certain cases
316 they can behave like autonomously replicating episomes (Wallet et al 2015; Chevigny et al, 2022).
317 Despite this apparent reproducibility, there were also examples of accession-specific large indels,
318 including a 1105-bp sequence absent Col-0, a 1790-bp sequence absent in C24, a 981-bp sequence
319 absent in Kyoto, and a 1407-bp sequence uniquely present in Sha. The 1407-bp sequence in Sha is
320 flanked by an 18-bp microhomology and QQ-2 sequence and consists of a 1025-bp large repeat
321 plus a 400-bp sequence of unknown origin, which is thought to be associated with the CMS
322 phenotype of Sha (Gobron et al., 2013). Interestingly, we observed that the 1025-bp large repeat in
323 Sha was flanked by a pair of the same 18-bp microhomologies, which occurs 13 times in Col-0 mt
324 genome. Thus, large repeats in the mt genome may simply originate by a two-step "excision-
325 reinsertion" mechanism, in which the excision step is an MMEJ event mediated by a pair of
326 microhomologies and the reinsertion step involves a third microhomology in an unexcised copy of
327 the genome (Figure 5b).
328 The overall HiFi read coverage of different *msh1* individuals did not vary greatly compared with
329 wild-type, except for fluctuations at sporadic stretches (Figure S16). The region around repeat
330 Large1-2 showed the lowest coverage in all *msh1* individuals, which would indicate a partial loss
331 of *trnS*, *atp6-2* and *atp9* (Figure S17a-c). By analyzing the depth of CLR reads, we verified the loss
332 of two-copy genes *trnS* and *atp6-2*, but not the single-copy gene *atp9* (Figure S17d-f). Together,
333 our analyses suggest that both repeat-mediated rearrangements and the relatively random
334 MMEJ/NHEJ observed in *msh1* mutants are involved in mt genome evolution, leading to
335 rearrangements, large deletions, and generation of new repeats with selection preventing the loss
336 of essential functional genes.
337
338 **Different accumulation patterns of pt genome variants in *msh1* mutants**

339 The frequencies of structural variants were very low in the pt genome for both wild-type and *msh1*
340 mutants (Figure S18; Table S25-S27). In contrast to the mt genome, the Arabidopsis pt genome
341 contains a pair of very large repeats (IRa/b) but very few intermediate repeats (Table S28). We
342 detected only one read in a single sample that supported the previously identified rearrangement
343 mediated by a pair of 123-bp imperfect repeats (Xu et al., 2011). Therefore, this variant appears to
344 make a very small contribution towards pt genome instability. In *msh1* mutants, the proportion of
345 pt MMEJ events was increased as also observed in mt genomes, but the associated
346 microhomologies included an overrepresentation of short, partial palindromic sequences (Figure
347 S18; Table S29). A partial palindromic sequence can be viewed as a pair of repeats or
348 microhomologies with both copies at the same location but in inverted orientation. For example, in
349 MMEJ events mediated by the most active partial palindromic sequence (a 22-bp stem and a 9-nt
350 loop) in *msh1* mutants, we observed recombination breakpoints in both the stem and the loop region,
351 resembling observations in mt genomic imperfect repeats (Figure 6a; Table S30).
352 Another important feature was the sample-specific accumulation of certain variants in pt genomes
353 of *msh1* mutants. Although most NHEJ products were at low frequencies with a seemingly random
354 distribution, we observed one pt NHEJ product that occurred only in one sample (M3-43-5) but at
355 relatively high frequency (121 one-rearrangement reads versus 1385 zero-rearrangement reads),
356 resulting in a 65-bp deletion in the *petA* gene (Figure 6b; Table S31). This mutation was predicted
357 to delete a highly conserved KQFE motif in the C-terminal stromal extension of cytochrome f,
358 which would compromise photosynthesis (Choquet et al., 2003). This sample-specific
359 accumulation pattern was also observed for SNVs in the pt genome. In total, we identified 17 pt
360 SNVs that were enriched to high frequencies in one or a few *msh1* mutant samples (Figure 6c;
361 Table S32). They were predominantly AT-to-GC transitions, consistent with the mutation spectrum
362 in *msh1* mutants that was reported previously and suggested to be caused by polymerase errors (Wu
363 et al., 2020). GC variants were also recently shown to be favored in the process of heteroplasmic
364 sorting in *Arabidopsis thaliana* (Broz et al., 2022). Moreover, we found that simple sequence
365 repeats (predominantly in homopolymers) exhibited larger proportions of length variants in the pt
366 genomes of *msh1* mutants compared with wild-type (Figure S19; Tables S33). For example, for the
367 13-bp homopolymer at position 67164 (which had reads supporting length variants ranging from -
368 6 bp to +18 bp), the proportion of reference length (13 bp) was less in all *msh1* individuals than
369 wild-type (Figure 6d). Together, our sequence data suggest that MSH1 participates in removal of
370 pt genome variants. The specific pt variants in *msh1* mutants are less reproducible among samples
371 than mt variants, and heteroplasmic sorting may contribute to this variation among individuals.
372

373 **Discussion**

374 In this work, we used HiFi reads (~15 kb long, > 99% accuracy) to study the patterns of mt and
375 pt genome variants using samples from Arabidopsis *msh1* mutant individuals. HiFi reads were long
376 enough to span the entire lengths of repeats, including the two largest repeats in the Arabidopsis
377 mt genome (6.5 kb and 4.2 kb), which enabled analyses of recombination breakpoints and higher-
378 order structural rearrangements that were not previously possible with short reads. We developed
379 a mapping-based pipeline, which showed that HiFi reads performed well in resolving one to
380 multiple repeats-mediated rearrangements in a single read and studying polymorphisms between
381 long non-identical repeats. Due to the ancient conservation of the *MSH1* gene across angiosperms
382 and green plants in general (Wu et al., 2020), our analysis in Arabidopsis likely extends to mt
383 genomes in other plants. Compared to previous studies, we also identified additional repeats active
384 in low-frequency recombination. Similarly, in the bacterial *Borrelia burgdorferi*, long-read
385 sequencing successfully identified a 97% reduction in the activity of recombinational switching at
386 the *vlsE* locus in MutL-deficient line, which was much higher than the ratio (36%) estimated by
387 short-read sequencing (Castellanos et al., 2022). However, due to the low read counts, HiFi reads
388 were less cost-effective for studying low-frequency MMEJ and NHEJ variants. This cost could be
389 reduced by isolating organelles from the cells before DNA extraction for sequencing.

390 **MSH1 suppresses exchange of SNVs and indels between non-identical repeats**

391 There might be a common need to suppress recombination mediated by non-allelic genomic
392 repeats across systems. During meiosis in the nucleus, MutS homologs detect mismatches and
393 recruit the DSB-generating SPO11 complex to suppress crossover repair between non-identical
394 centromeric repeats (Blackwell et al., 2020; Naish et al., 2021). In the mitochondria, MSH1 is
395 similarly required to suppress crossover repair between intermediate-sized repeats (Arrieta-Montiel
396 et al., 2009; Davila et al., 2011). Gene conversion was previously proposed to explain the
397 discontinuous pattern of SNVs/indels derived from non-identical repeats in chimeric reads (Davila
398 et al., 2011). However, the short-read sequencing performed in that study makes it difficult to assess
399 whether these discontinuous patterns were present even in molecules that did not exhibit structural
400 rearrangements from recombination. In our analysis, unambiguous mapping of HiFi reads across
401 the full length of mt repeats identified the exchange of SNVs and indels between pairs of imperfect
402 repeats in zero-rearrangement reads in *msh1* mutants. MSH1 activity appears to suppress these
403 exchanges based on their relative absence in wild-type.

404 The observed variants could be introduced by non-crossover gene conversion events, but they
405 could also result from multiple rounds of crossover recombination involving the same repeats but
406 different breakpoints. It is important to note that any analysis of a standing population of genome

407 copies in a tissue sample will not only reflect the recombination events that originally generated
408 structural rearrangements but also the subsequent history replication, degradation, and
409 heteroplasmic sorting. Furthermore, such processes may be sensitive to rearing conditions (e.g.,
410 day length) and the stage in development when tissue was harvested, which should be considered
411 in any comparison between our results and previous studies.

412 There could very well be a connection between MSH1's functions in suppressing ectopic
413 recombination and mismatch repair mechanisms. Specifically, MSH1 may provide a surveillance
414 mechanism to check for mismatches in double-stranded DNA regardless of whether they are
415 introduced by polymerase errors or by strand invasion between short or non-identical repeats,
416 followed by introduction of DSBs to (re)start repair by homologous recombination. To suppress
417 recombination at short and intermediate repeats that are 100% identical in sequence, MSH1 may
418 need to detect differences in boundary sequence. During homologous recombination, the bacterial
419 RecA-ssDNA presynaptic filament checks homology in ~9 bp steps before strand exchange with
420 the dsDNA template (Huang et al., 2020). In such screening mechanisms, the first mismatches may
421 be detected at the repeat boundaries when the repeat sequences themselves are identical. Thus,
422 MSH1 may remove both crossover and non-crossover forms of ectopic recombination to ensure
423 accurate DSB repair of the genome by detecting mismatches within repeats or at repeat boundaries
424 (Figure 7a).

425

426 **Conventional recombination activity only accounts for a fraction of DSB repair events**

427 Most previous mt repeat annotations depended on the penalty parameters of mismatches and gaps
428 used in sequence alignment. We generated a global and quantitative map of rearrangements
429 mediated by repeated sequence elements, including the identification of additional low-frequency
430 repeats, which raises new questions about what contributes to the observed differences in
431 recombination activity among repeat pairs. Increased crossover recombination events involving
432 intermediate-sized repeats have been observed in wild-type plants when DSB frequencies were
433 increased by DNA damaging chemicals (Wallet et al., 2015). Therefore, repeats near DSB hotspots
434 may show higher ectopic recombination activity compared to other repeats with similar length and
435 sequence identity, which might explain cases that fall far from the regression line that defines the
436 relationship between recombination activity and repeat length (Figure S5b). It may also help to
437 study repair events systematically by inducing DSBs through mt-targeting CRISPR-Cas9 or
438 mitoTALENs (Arimura et al., 2020) and in more RRR gene mutants, such as *why2*, *reca3*, *osb1*
439 (Gualberto and Newton 2017).

440 DSBs within a repeat or in its flanking sequences would be repaired using homologous DNA
441 stretches of another intra- or intermolecular copy. After resection of DSB ends, RecA recombinases
442 bind 3' ssDNA and rapidly search for homology (Wiktor et al., 2021). It was reported that at least
443 8-nt microhomology was needed for RecA-ssDNA filaments to bind dsDNA templates stably, and
444 15-nt was sufficient to be resistant to facilitated exchange during homology search (Qi et al., 2015).
445 These studies suggest that only a small stretch of RecA-ssDNA is needed during each HDR repair
446 event. Typically, only crossover products have been used to estimate frequencies of recombinant
447 conformations (Wallet et al., 2015; Odahara et al., 2017). Such calculations do not account for the
448 fact that non-crossover recombination or gene conversion events between repeats can also occur at
449 substantial frequencies. Because *msh1* mutants exhibited exchange of sequence variants between
450 non-identical repeats even in reads with no structural rearrangements, the increases in ectopic
451 recombination activity are likely underestimated when only crossover products are quantified.
452 When calculating recombination for some three-copy or four-copy repeats locating within other
453 longer two-copy repeats, one-rearrangement reads shared by both shorter and longer annotations
454 were assigned only to the larger two-copy repeats by blastn (Figure 3). This illustrates that it may
455 be more natural to view recombination events as associated with specific genomic positions, but
456 not with repeat annotations.
457
458 **Different variant accumulation patterns in mt and pt genomes**
459 In plant organellar genomes, large repeats recombine constantly. The numerous intermediate-sized
460 repeats are the major source of altered recombination activities in *msh1* mutants. The global
461 patterns of mt recombination indicated by one-rearrangement reads (Figure 2) and fluctuation in
462 sequence depth (Figure S16) were similar among individuals, which could be a combined effect of
463 the locations of repeat pairs and the selection to maintain regions containing functional genes.
464 Despite the common loss of the region around repeat Large1-2 (containing *atp6-2*, *atp9* and *trnS*),
465 the relative dosage of parental and crossover recombination forms of each repeat pair may differ
466 among *msh1* individuals and generations, which was proposed to be caused by heteroplasmic
467 sorting in previous studies (Shedge et al., 2007; Arrieta-Montiel et al., 2009). In other Arabidopsis
468 accessions, we observed large indels flanked by repeats or microhomologies. These indels
469 sometimes coincided with the low-coverage regions observed in *msh1* mutants, including the
470 previously investigated region around Large1-2. Some of these structural rearrangements may
471 cause a CMS phenotype, such as the one reported in Sha (Figure S15; Gobron et al., 2013). In
472 natural populations, the mt genome may be transiently (within a single generation) found in a highly
473 rearranged and heteroplasmic states similar to *msh1* mutants, providing source materials for

474 functional innovation (Woloszynska, 2010; Tang et al., 2017; Johnston, 2019). However, the
475 metaFlye assembly graphs support previous observations that the mt genomes of wild-type
476 Arabidopsis accessions typically exist in low-heteroplasmy states with structural variants
477 predominantly associated with repeats or microhomologies mediated DSB repair activities (Figure
478 S13; Table S24).

479 The contrasting patterns of variation between mt and pt genomes may result from differences in
480 repeat spectrum, genome copy number per organelle, and DNA repair mechanisms. Structural
481 rearrangements in pt genome of Arabidopsis *msh1* mutants were hardly detected in earlier studies,
482 possibly due to the lack of intermediate-sized repeats (Xu et al., 2011). HiFi-identified
483 rearrangements in the pt genome were mostly MMEJ and NHEJ events. The MMEJ events
484 mediated by partial palindromic sequences were found at a higher frequency than others and
485 showed repeatability among samples. Moreover, other examples of pt genome variants occurred at
486 high frequency in individual samples regardless of whether they were MMEJ, NHEJ (Figure 6b)
487 or SNVs (Figure 6c). Such sample-specific variants were most likely due to a combination of
488 differential occurrence of the original mutations and subsequent variation introduced by
489 heteroplasmic sorting. Simple sequence repeats including homopolymers are intrinsically unstable
490 regions in pt genomes, which is also described in an earlier characterization of the mutation
491 spectrum of 51 *Oenothera* mutants (Massouh et al., 2016). Although length variants at these
492 homopolymers are found even in wild-type plants, our data suggest that the frequencies of these
493 variants are reduced by *MSH1* activity (Figure 6d and S19).

494 In conclusion, *MSH1* acts to maintain both structural and non-structural homoplasmy in plant
495 organellar genomes possibly by promoting gene conversion (with allelic DNA templates). These
496 genomes can be transiently disrupted when *MSH1* expression is downregulated, providing more
497 raw materials for further functional selection (Yang and Mackenzie, 2019). When *MSH1*
498 expression is restored, evolved versions of the genomes can rapidly rise in frequency (Shedge et
499 al., 2007). However, the two genomes can respond quite differently to the loss of the same gene
500 (Figure 7b). Organellar genome editing technologies are becoming more and more accessible
501 (Kazama et al., 2019; Arimura et al., 2020; Forner et al., 2022). However, most protocols in non-
502 model species and the recent TALEN-GDM in Arabidopsis require a callus-induction step (Forner
503 et al., 2022), during which the expression of MSH1 may be downregulated (Sun et al., 2012). Thus,
504 we need to pay attention to the rearrangements and gene conversion events in mt genomes, which
505 are not easily detected by short-read sequencing, as well as the unwanted random accumulation of
506 AT-to-GC mutations and length variants of simple sequence repeats in pt genomes. In the future,

507 we may better understand the role of cytoplasmic genetic material in crop improvement through
508 studying RRR genes including *MSH1*.
509
510
511 **Methods**
512 **DNA extraction and PacBio HiFi sequencing**
513 Generation of *msh1* mutant and wild-type F_3 families of *Arabidopsis thaliana* was described
514 previously (Wu et al., 2020). In brief, an *msh1* mutant (*chm1-1*, CS3372) was used as the pollen
515 donor in a backcross to *A. thaliana* Col-0 so that the F_1 plant would have wild-type organellar
516 genomes. Seeds of each homozygous mutant and wild-type F_2 plant was collected separately and
517 propagated as F_3 families. To extend rosette stage growth before bolting, plants were grown in a
518 short-day growth chamber (8 h day and 16 h night) for six weeks. Rosettes of 11 individuals were
519 collected to extract DNA with the CTAB method. Total DNA was purified with a QIAGEN kit
520 (Q13343) and quantified using NanoDrop and Qubit. Purified DNA (3.3~18.6 µg for each sample)
521 was used for the construction of SMRTbell libraries according to PacBio's standard protocol.
522 Sequencing was performed using the PacBio Sequel II platform in CCS mode, which generated a
523 total of 34.36 Gb HiFi reads (12.7~18.1 kb mean read length). Both the mutant and wild-type
524 individuals used in our studied may have segments originated from the L*er* accession in the nuclear
525 genome as indicated by an earlier study (Virdi et al., 2015).
526
527 **Identification of rearrangements in mt and pt genomes**
528 Raw HiFi reads from each sample were mapped to the Arabidopsis Col-0 reference genome
529 (TAIR10 for nucleus; NC_037304.1 for mitochondria; NC_000932.1 plus a 1-bp expansion at
530 position 28673 in the pt genome as described in Wu et al., 2020) with minimap2 (version 2.20,
531 r1061; parameters: -ax map-hifi; Li, 2018). Reads mapped to mt and pt genomes were separated
532 using bamtools (version 2.5.1; Barnett et al., 2011) and realigned against each genome with blastn
533 (version 2.12.0; Camacho et al., 2009). Raw blastn results were further processed by a custom
534 pipeline (available at https://github.com/zouyinstein/hifisr) to group the reads by discontinuity,
535 alignment overlap (AO) size (i.e., repeat size) and mapping coordinates, and calculate read counts
536 for input samples. Different types of genome rearrangements were manually checked, including
537 zero-, one-, two-, three-, and four-rearrangement reads, MMEJ (AO length of 2-49 bp), NHEJ (AO
538 length of 0-1 bp), and replication slippage mediated by tandem repeats. We also identified large
539 insertions, which were poly-G/C tracts in most cases. However, it was not clear whether these
540 inserted homopolymers were biological or a sequencing artefact, and they were not analyzed further.

541

542 **Mitochondrial repeat annotations and recombination activity**

543 Candidate coordinates of repeats were generated by blastn (version 2.12.0+; parameters: -
544 word_size 4) using the mt reference sequence as both query and subject (Table S10). The resulting
545 coordinates were manually checked against the junction coordinates of one-rearrangement HiFi
546 reads. We identified 71 groups of two-copy repeats, 23 groups of three-copy repeats, 5 groups of
547 four-copy repeats, which accounted for most junctions of read alignments in *msh1* mutant mt
548 genomes (Table S11-S13). These 99 repeat groups contained all previously reported active repeats
549 except for repeat TT and 53 additional groups of recombinationally active repeats ≥ 50 bp, which
550 we named HNR1 to HNR53 (repeats FF, WW, XX, YY, and ZZ were not considered because they
551 are absent in the NC_037304.1 annotation). For consistency with blastn alignments, the boundaries
552 of some low-identity repeats in our annotation differed from previous annotations. For example,
553 the direct repeats MMJS-1 (767 bp) and MMJS-2 (692 bp) were previously annotated as separate
554 smaller repeats MM, J, and S. We treated short repeats as part of the microhomology category (2
555 to 49 bp) because blastn can often generate overlapping, inaccurate, or incomplete predictions of
556 homology below 50 bp in size. Per base ratios of rearranged reads versus total reads and read counts
557 for parental and recombinant forms were calculated with custom Python scripts based on blastn
558 outputs of the HiFi-SR pipeline and used to identify dynamic regions of the mt genome.

559

560 **Identification of non-allelic SNVs/indels between imperfect repeats**

561 To analyze patterns of SNVs/indels, zero-rearrangement reads of each sample were re-aligned to
562 the mt reference genome by PacBio pbmm2 (version 1.3.0; parameters: --sort --preset CCS). The
563 resulting BAM files were used for variant calling by GATK HaplotypeCaller (version 4.2.3.0).
564 Variant frequencies and sequencing depths for each GATK-called SNV/indel were calculated from
565 raw BAM files with the Python module pysam (version 0.17.0). Non-allelic SNVs/indels inferred
566 from putative gene conversion events between some high-activity mt imperfect repeats were not
567 always detected by GATK but were also used for calculations.

568

569 **Calculation of absolute and normalized per base coverage**

570 Total mt genome-mapped reads were re-aligned to mt reference genomes by PacBio pbmm2
571 (version 1.3.0; parameters: --sort --preset CCS). Absolute per base coverage across the entire mt
572 genome of each individual was calculated with the samtools depth subcommand (version 1.7) and
573 normalized to total bases of each sample. Annotations are based on the GenBank file NC_037304.1
574 (Sloan et al., 2018).

575

576 **metaFlye assembly and synteny analysis of mt genomes**

577 Raw PacBio CLR reads from each accession were mapped to the reference genomes (the respective
578 nuclear assembly in Jiao and Schneeberger, 2020; Col-0 NC_037304.1 for mitochondria and
579 NC_000932.1 plus a 1-bp expansion at position 28673 for plastid) with minimap2 (version 2.20,
580 r1061; parameters: -ax map-pb; Li, 2018). Reads mapped to mt genomes were separated using
581 bamtools (version 2.5.1; Barnett et al., 2011) and were re-assembled with metaFlye (Flye version
582 2.9-b1774; parameters: --meta --pacbio-hifi/--pacbio-corr --extra-params output_gfa_before_rr=1
583 --genome-size 370K/2M for mitochondria; Kolmogorov et al., 2020). Before assembling, PacBio
584 CLR reads longer than 1000 bp were extracted by Filtlong (version 0.2.1; parameters: --min_length
585 1000) and were corrected using CONSENT (version 2.2.2; parameters: --type PB; Morisse et al.,
586 2021). Assembly graphs before and after repeat resolution were visualized by Bandage (version
587 0.8.1; Wick et al., 2015). Reads from different *msh1* families were merged. To get linear
588 representations of pseudo-master circles for mapping-based analysis, we started from the
589 assemblies (--genome-size 2M) after repeat resolution, discarded low-coverage contigs, manually
590 resolved the large repeats for accessions Eri-1, C24, Cvi-0, An-1, L*er*, Sha, and re-resolved Large2
591 repeats for Kyoto with the Bandage software, followed by rotation or reverse complementation.
592 Corrected CLR reads were realigned against each pseudo-master circle using the HiFi-SR pipeline.
593 To analyze the shared synteny between mt genomes, we performed blastn alignments using the
594 Col-0 mt reference sequence as the query and pseudo-master circle of other accessions as the
595 subject. Raw blastn results were processed by a custom Python script available in HiFi-SR.
596 Coordinates of large repeats were determined based on blastn alignments using the pseudo-master
597 circle as both query and subject. Repeat-mediated rearrangements, MMEJ/NHEJ events, and large
598 insertions composed of short fragments of multiple origins were manually checked. Plots were
599 generated by the Python package pycircos (version 1.0.2).

600

601 **Identification of SNVs/indels in the pt genome**

602 To analyze patterns of SNVs/indels in the pt genome, total reads of each sample were re-aligned to
603 the pt reference genome by PacBio pbmm2 (version 1.3.0; parameters: --sort --preset CCS). The
604 resulting BAM files were used for variant calling by GATK HaplotypeCaller (version 4.2.3.0).
605 Variant frequencies and sequencing depths for each GATK-called SNV/indel were calculated from
606 raw BAM files with the Python module pysam (version 0.17.0).

607

608 **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), the U.S. National Institutes of Health (R01 GM118046), and the Chinese Postdoctoral Science Foundation (2021M693451).

Author contributions

ZW designed the research; YZ performed the experiments; YZ wrote the computer scripts and performed majority of the bioinformatic analysis; WZ analyzed the boundaries of structure rearrangements in Arabidopsis accessions; YZ, WZ, DBS, and ZW discussed the results; YZ, DBS and ZW wrote the manuscript; DBS 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 the 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), which is publicly accessible at https://ngdc.cncb.ac.cn/gsa. The Python-based pipeline for the detection of plant mt structural rearrangements used in this study is available at https://github.com/zouyinstein/hifisr.

Supporting Information

Figure S1 Seedlings of *msh1* mutants and wild-type *Arabidopsis thaliana*

Figure S2 Mapping ratio of HiFi reads in different samples

Figure S3 Patterns of one-rearrangement events detected in mt genome of wild-type

Figure S4 Variability of structural rearrangements across the mt genome

643 **Figure S5** Comparison of the recombination activities of different repeat groups
644 **Figure S6** Parental forms and recombination products can be detected according to paths
645 **Figure S7** Two-rearrangement reads involving large repeats are increased in *msh1* mutants
646 **Figure S8** Continuous distribution of SNVs in non-identical repeats CC-1/2
647 **Figure S9** Mapping of zero-rearrangement long reads identifies variants within repeats
648 **Figure S10** Imperfect repeats with relatively higher ratios of exchange of SNVs and indels in *msh1*
649 mutants
650 **Figure S11** Coverage of regions containing repeat group H in wild-type and *msh1* individuals
651 **Figure S12** Coverage of regions containing repeat group MMJS in wild-type and *msh1* individuals
652 **Figure S13** metaFlye assemblies of mt genomes in other accessions
653 **Figure S14** The metaFlye assemblies of mt genomes are more complex in *msh1* mutants than wild-
654 type
655 **Figure S15** Synteny analysis of mt genomes of other *Arabidopsis thaliana* accessions
656 **Figure S16** Normalized per base read coverage across mt genome
657 **Figure S17** Coverages around *atp6-1*, *atp6-2*, and *atp9* in *msh1* mutants and Arabidopsis
658 accessions
659 **Figure S18** The proportions of structural variants in Arabidopsis pt genomes
660 **Figure S19** Length variants at simple sequence repeats in the pt genomes
661 **Table S1** DNA content and fresh weight of individual plants used for HiFi sequencing
662 **Table S2** Mapping statistics of HiFi reads of individual plants
663 **Table S3** Statistics of mt genome-mapped reads of individual plants
664 **Table S4** Statistics of pt genome-mapped reads of individual plants
665 **Table S5** Read counts for different types of mt structural rearrangements in individual plants
666 **Table S6** Read counts for different categories of mt structural rearrangements in wild-type and
667 *msh1* plants
668 **Table S7** Alignment overlap length in mt one-rearrangement reads.
669 **Table S8** HiFi-SR outputs of two-rearrangement mt genome-mapped reads of individual plants
670 **Table S9** HiFi-SR outputs of one-rearrangement mt genome-mapped reads of individual plants
671 **Table S10** Raw blastn outputs using the mt reference sequence as both query and subject
672 **Table S11** Annotations of two-copy repeats based on blastn and one-rearrangement HiFi reads
673 **Table S12** Annotations of three-copy repeats based on blastn and one-rearrangement HiFi reads
674 **Table S13** Annotations of four-copy repeats based on blastn and one-rearrangement HiFi reads
675 **Table S14** Recombination frequencies of two-copy repeats
676 **Table S15** Recombination frequencies of three-copy repeats

677 **Table S16** HiFi read counts indicating parental forms and crossover recombination products of
678 representative repeats in wild-type and *msh1* mutants.
679 **Table S17** Counts of continuous and discontinuous one-rearrangement reads in high- and
680 intermediate-activity two-copy imperfect repeats
681 **Table S18** Ratios of predicted mt gene conversion sites in individual plants and published HiFi
682 datasets
683 **Table S19** Frequencies of GATK-identified mt SNVs in individual plants and published HiFi
684 datasets
685 **Table S20** Frequencies of GATK-identified mt indels in individual plants and published HiFi
686 datasets
687 **Table S21** Mapping statistics for PacBio CLR reads in individual plants
688 **Table S22** Statistics of corrected PacBio CLR reads used for metaFlye assembly of mt genomes
689 **Table S23** Blastn results for mt large repeats
690 **Table S24** Blastn results for syntenic relationships in mt genomes
691 **Table S25** Read counts for different types of pt structural rearrangements in individual plants
692 **Table S26** Read counts for different categories of pt structural rearrangements in wild-type and
693 *msh1* plants
694 **Table S27** Alignment overlap length in pt one-rearrangement reads
695 **Table S28** Raw blastn outputs using the pt reference sequence as both query and subject
696 **Table S29** HiFi-SR outputs of pt one-rearrangement reads mediated by short partial palindromic
697 sequences in individual plants
698 **Table S30** HiFi-SR outputs of pt one-rearrangement reads mediated by the most active short partial
699 palindromic sequence in individual plants
700 **Table S31** HiFi-SR outputs of a pt NHEJ event in individual plants
701 **Table S32** Frequencies of GATK-identified pt SNVs in individual plants and published HiFi
702 datasets
703 **Table S33** Frequencies of GATK-identified pt SNVs in individual plants and published HiFi
704 datasets
705
706 **References**
707 1. Abdelnoor RV, Yule R, Elo A, Christensen AC, Meyer-Gauen G, Mackenzie SA.
708 Substoichiometric shifting in the plant mitochondrial genome is influenced by a gene
709 homologous to MutS. **Proc Natl Acad Sci U S A 2003**, 100(10)**:** 5968-5973.

710 2. Abdelnoor RV, Christensen AC, Mohammed S, Munoz-Castillo B, Moriyama H,
711 Mackenzie SA. Mitochondrial genome dynamics in plants and animals: convergent gene
712 fusions of a MutS homologue. **J Mol Evol 2006**, 63(2)**:** 165-173.
713 3. Arimura SI, Ayabe H, Sugaya H, Okuno M, Tamura Y, Tsuruta Y, Watari Y, Yanase S,
714 Yamauchi T, Itoh T, Toyoda A, Takanashi H, Tsutsumi N. Targeted gene disruption of
715 *ATP synthases 6-1* and *6-2* in the mitochondrial genome of *Arabidopsis thaliana* by
716 mitoTALENs. **Plant J 2020**, 104(6)**:** 1459-1471.
717 4. Arrieta-Montiel MP, Shedge V, Davila J, Christensen AC, Mackenzie SA. Diversity of the
718 Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity.
719 **Genetics 2009**, 183(4)**:** 1261-1268.
720 5. Ayala-Garcia VM, Baruch-Torres N, Garcia-Medel PL, Brieba LG. Plant organellar DNA
721 polymerases paralogs exhibit dissimilar nucleotide incorporation fidelity. **FEBS J 2018**,
722 285(21)**:** 4005-4018.
723 6. Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: a C++ API
724 and toolkit for analyzing and managing BAM files. **Bioinformatics 2011**, 27(12)**:** 1691-
725 1692.
726 7. Bendich AJ. Reaching for the ring: the study of mitochondrial genome structure. **Curr**
727 **Genet 1993**, 24(4)**:** 279-290.
728 8. Blackwell AR, Dluzewska J, Szymanska-Lejman M, Desjardins S, Tock AJ, Kbiri N,
729 Lambing C, Lawrence EJ, Bieluszewski T, Rowan B, Higgins JD, Ziolkowski PA,
730 Henderson IR. MSH2 shapes the meiotic crossover landscape in relation to interhomolog
731 polymorphism in Arabidopsis. **EMBO J 2020**, 39(21)**:** e104858.
732 9. Broz AK, Keene A, Fernandes Gyorfy M, Hodous M, Johnston IG, Sloan DB. Sorting of
733 mitochondrial and plastid heteroplasmy in *Arabidopsis* is extremely rapid and depends on
734 MSH1 activity. **Proc Natl Acad Sci U S A 2022**, 119(34): e2206973119.
735 10. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.
736 BLAST+: architecture and applications. **BMC Bioinformatics 2009**, 10**:** 421.
737 11. Castellanos M, Verhey TB, Goldstein M, Chaconas G. The Putative Endonuclease Activity
738 of MutL Is Required for the Segmental Gene Conversion Events That Drive Antigenic
739 Variation of the Lyme Disease Spirochete. **Front Microbiol 2022**, 13**:** 888494.
740 12. Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, Dong L, Zhang Z, Yu C, Sun Y, Chi
741 L, Chen H, Zhai S, Sun Y, Lan L, Zhang X, Xiao J, Bao Y, Wang Y, Zhang Z, Zhao W.
742 The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data
743 Types. **Genomics Proteomics Bioinformatics 2021**.

744 13. Chevigny N, Weber-Lotfi F, Le Blevenec A, Nadiras C, Fertet A, Bichara M*, et al.* RADA-
745 dependent branch migration has a predominant role in plant mitochondria and its defect
746 leads to mtDNA instability and cell cycle arrest. **PLoS Genet 2022**, 18(5)**:** e1010202.
747 14. Choquet Y, Zito F, Wostrikoff K, Wollman FA. Cytochrome f translation in
748 Chlamydomonas chloroplast is autoregulated by its carboxyl-terminal domain. **Plant Cell**
749 **2003**, 15(6)**:** 1443-1454.
750 15. Christensen AC. Plant mitochondrial genome evolution can be explained by DNA repair
751 mechanisms. **Genome Biol Evol 2013**, 5(6)**:** 1079-1086.
752 16. Christensen AC. Plant Mitochondria are a Riddle Wrapped in a Mystery Inside an Enigma.
753 **J Mol Evol 2021**, 89(3)**:** 151-156.
754 17. Davila JI, Arrieta-Montiel MP, Wamboldt Y, Cao J, Hagmann J, Shedge V, Xu YZ, Weigel
755 D, Mackenzie SA. Double-strand break repair processes drive evolution of the
756 mitochondrial genome in Arabidopsis. **BMC Biol 2011**, 9**:** 64.
757 18. Forner J, Kleinschmidt D, Meyer EH, Fischer A, Morbitzer R, Lahaye T*, et al.* Targeted
758 introduction of heritable point mutations into the plant mitochondrial genome. **Nat Plants**
759 **2022**, 8(3)**:** 245-256.
760 19. Fuchs P, Rugen N, Carrie C, Elsasser M, Finkemeier I, Giese J, Hildebrandt TM, Kuhn K,
761 Maurino VG, Ruberti C, Schallenberg-Rudinger M, Steinbeck J, Braun HP, Eubel H,
762 Meyer EH, Muller-Schussele SJ, Schwarzlander M. Single organelle function and
763 organization as estimated from Arabidopsis mitochondrial proteomics. **Plant J 2020**,
764 101(2)**:** 420-441.
765 20. Fukui K, Harada A, Wakamatsu T, Minobe A, Ohshita K, Ashiuchi M, Yano T. The GIY-
766 YIG endonuclease domain of *Arabidopsis* MutS homolog 1 specifically binds to branched
767 DNA structures. **FEBS Lett 2018**, 592(24)**:** 4066-4077.
768 21. Gobron N, Waszczak C, Simon M, Hiard S, Boivin S, Charif D, Ducamp A, Wenes E,
769 Budar F. A cryptic cytoplasmic male sterility unveils a possible gynodioecious past for
770 *Arabidopsis thaliana*. **PLoS One 2013**, 8(4)**:** e62450.
771 22. Gualberto JM, Newton KJ. Plant Mitochondrial Genomes: Dynamics and Mechanisms of
772 Mutation. **Annu Rev Plant Biol 2017**, 68**:** 225-252.
773 23. Huang X, Lu Y, Wang S, Sui M, Li J, Ma J, Ma D, Jia Q, Hu S, Xu C, Li M. Mismatch
774 sensing by nucleofilament deciphers mechanism of RecA-mediated homologous
775 recombination. **Proc Natl Acad Sci U S A 2020**, 117(34)**:** 20549-20554.
776 24. Hussmann D, Starnawska A, Kristensen L, Daugaard I, Thomsen A, Kjeldsen TE, Hansen
777 CS, Bybjerg-Grauholm J, Johansen KD, Ludvigsen M, Kristensen T, Larsen TS, Moller

778 MB, Nyvold CG, Hansen LL, Wojdacz TK. *IGHV*-associated methylation signatures more
779 accurately predict clinical outcomes of chronic lymphocytic leukemia patients than *IGHV*
780 mutation load. **Haematologica 2021**.
781 25. Jiao WB, Schneeberger K. Chromosome-level assemblies of multiple Arabidopsis
782 genomes reveal hotspots of rearrangements with altered evolutionary dynamics. **Nat**
783 **Commun 2020**, 11(1)**:** 989.
784 26. Johnston IG. Tension and Resolution: Dynamic, Evolving Populations of Organelle
785 Genomes within Plant Cells. **Mol Plant 2019**, 12(6)**:** 764-783.
786 27. Kazama T, Okuno M, Watari Y, Yanase S, Koizuka C, Tsuruta Y, Sugaya H, Toyoda A,
787 Itoh T, Tsutsumi N, Toriyama K, Koizuka N, Arimura SI. Curing cytoplasmic male sterility
788 via TALEN-mediated mitochondrial genome editing. **Nat Plants 2019**, 5(7)**:** 722-730.
789 28. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan
790 J, Polevikov E, Smith TPL, Pevzner PA. metaFlye: scalable long-read metagenome
791 assembly using repeat graphs. **Nat Methods 2020**, 17(11)**:** 1103-1110.
792 29. Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, Christensen AC.
793 The alternative reality of plant mitochondrial DNA: One ring does not rule them all. **PLoS**
794 **Genet 2019**, 15(8)**:** e1008373.
795 30. Kundariya H, Yang X, Morton K, Sanchez R, Axtell MJ, Hutton SF, Fromm M, Mackenzie
796 SA. MSH1-induced heritable enhanced growth vigor through grafting is associated with
797 the RdDM pathway in plants. **Nat Commun 2020**, 11(1)**:** 5343.
798 31. Lee J, Cho CH, Park SI, Choi JW, Song HS, West JA, Bhattacharya D, Yoon HS. Parallel
799 evolution of highly conserved plastid genome architecture in red seaweeds and seed plants.
800 **BMC Biol 2016**, 14**:** 75.
801 32. Li H. Minimap2: pairwise alignment for nucleotide sequences. **Bioinformatics 2018**,
802 34(18)**:** 3094-3100.
803 33. Mackenzie SA, Kundariya H. Organellar protein multi-functionality and phenotypic
804 plasticity in plants. **Philos Trans R Soc Lond B Biol Sci 2020**, 375(1790)**:** 20190182.
805 34. Martínez-Zapater JM, Gil P, Capel J, Somerville CR. Mutations at the Arabidopsis CHM
806 locus promote rearrangements of the mitochondrial genome. **Plant Cell 1992**, 4(8)**:** 889-
807 899.
808 35. Massouh A, Schubert J, Yaneva-Roder L, Ulbricht-Jones ES, Zupok A, Johnson MT,
809 Wright SI, Pellizzer T, Sobanski J, Bock R, Greiner S. Spontaneous Chloroplast Mutants
810 Mostly Occur by Replication Slippage and Show a Biased Pattern in the Plastome of
811 Oenothera. **Plant Cell 2016**, 28(4)**:** 911-929.

812 36. Masutani B, Arimura SI, Morishita S. Investigating the mitochondrial genomic landscape
813 of *Arabidopsis thaliana* by long-read sequencing. **PLoS Comput Biol 2021**, 17(1)**:**
814 e1008597.
815 37. Members C-N, Partners. Database Resources of the National Genomics Data Center, China
816 National Center for Bioinformation in 2022. **Nucleic Acids Res 2022**, 50(D1)**:** D27-D38.
817 38. Miller-Messmer M, Kuhn K, Bichara M, Le Ret M, Imbault P, Gualberto JM. RecA-
818 dependent DNA repair results in increased heteroplasmy of the Arabidopsis mitochondrial
819 genome. **Plant Physiol 2012**, 159(1): 211-226.
820 39. Morisse P, Marchet C, Limasset A, Lecroq T, Lefebvre A. Scalable long read self-
821 correction and assembly polishing with multiple sequence alignment. **Sci Rep 2021**, 11(1)**:**
822 761.
823 40. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmucker A, Mandakova
824 T, Jamge B, Lambing C, Kuo P, Yelina N, Hartwick N, Colt K, Smith LM, Ton J, Kakutani
825 T, Martienssen RA, Schneeberger K, Lysak MA, Berger F, Bousios A, Michael TP, Schatz
826 MC, Henderson IR. The genetic and epigenetic landscape of the *Arabidopsis* centromeres.
827 **Science 2021**, 374(6569)**:** eabi7489.
828 41. Odahara M, Kishita Y, Sekine Y. MSH1 maintains organelle genome stability and
829 genetically interacts with *RECA* and *RECG* in the moss *Physcomitrella patens*. **Plant J**
830 **2017**, 91(3)**:** 455-465.
831 42. Palmer JD, Herbon LA. Plant mitochondrial DNA evolves rapidly in structure, but slowly
832 in sequence. **J Mol Evol 1988**, 28(1-2)**:** 87-97.
833 43. Qi Z, Redding S, Lee JY, Gibb B, Kwon Y, Niu H, Gaines WA, Sung P, Greene EC. DNA
834 sequence alignment by microhomology sampling during homologous recombination. **Cell**
835 **2015**, 160(5)**:** 856-869.
836 44. Rédei GP. Extra-chromosomal mutability determined by a nuclear gene locus in
837 Arabidopsis. **Mutat Res 1973**, 18: 149–162.
838 45. Sakamoto W, Kondo H, Murata M, Motoyoshi F. Altered mitochondrial gene expression
839 in a maternal distorted leaf mutant of Arabidopsis induced by chloroplast mutator. **Plant**
840 **Cell 1996**, 8(8)**:** 1377-1390.
841 46. Sandhu AP, Abdelnoor RV, Mackenzie SA. Transgenic induction of mitochondrial
842 rearrangements for cytoplasmic male sterility in crop plants. **Proc Natl Acad Sci U S A**
843 **2007**, 104(6)**:** 1766-1770.

844 47. Shedge V, Arrieta-Montiel M, Christensen AC, Mackenzie SA. Plant mitochondrial
845 recombination surveillance requires unusual *RecA* and *MutS* homologs. **Plant Cell 2007**,
846 19(4)**:** 1251-1264.
847 48. Sloan DB, Wu Z, Sharbrough J. Correction of Persistent Errors in Arabidopsis Reference
848 Mitochondrial Genomes. **Plant Cell 2018**, 30(3)**:** 525-527.
849 49. Sun P, Arrieta-Montiel MP, Mackenzie SA. Utility of in vitro culture to the study of plant
850 mitochondrial genome configuration and its dynamic features. **Theor Appl Genet 2012**,
851 125(3)**:** 449-454.
852 50. Tang H, Zheng X, Li C, Xie X, Chen Y, Chen L, Zhao X, Zheng H, Zhou J, Ye S, Guo J,
853 Liu YG. Multi-step formation, evolution, and functionalization of new cytoplasmic male
854 sterility genes in the plant mitochondrial genomes. **Cell Res 2017**, 27(1)**:** 130-146.
855 51. Unseld M, Marienfeld JR, Brandt P, Brennicke A. The mitochondrial genome of
856 *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. **Nat Genet 1997**, 15(1)**:**
857 57-61.
858 52. Virdi KS, Laurie JD, Xu YZ, Yu J, Shao MR, Sanchez R*, et al. Arabidopsis* MSH1
859 mutation alters the epigenome and produces heritable changes in plant growth. **Nat**
860 **Commun 2015**, 6**:** 6386.
861 53. Wallet C, Le Ret M, Bergdoll M, Bichara M, Dietrich A, Gualberto JM. The RECG1 DNA
862 Translocase Is a Key Factor in Recombination Surveillance, Repair, and Segregation of the
863 Mitochondrial DNA in Arabidopsis. **Plant Cell 2015**, 27(10)**:** 2907-2925.
864 54. Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, Dong Q, Ye
865 K. High-quality *Arabidopsis thaliana* Genome Assembly with Nanopore and HiFi Long
866 Reads. **Genomics Proteomics Bioinformatics 2021**.
867 55. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo*
868 genome assemblies. **Bioinformatics 2015**, 31(20)**:** 3350-3352.
869 56. Wiktor J, Gynna AH, Leroy P, Larsson J, Coceano G, Testa I, Elf J. RecA finds
870 homologous DNA by reduced dimensionality search. **Nature 2021**, 597(7876)**:** 426-429.
871 57. Wolfe KH, Li WH, Sharp PM. Rates of nucleotide substitution vary greatly among plant
872 mitochondrial, chloroplast, and nuclear DNAs. **Proc Natl Acad Sci U S A 1987**, 84(24)**:**
873 9054-9058.
874 58. Woloszynska M. Heteroplasmy and stoichiometric complexity of plant mitochondrial
875 genomes--though this be madness, yet there's method in't. **J Exp Bot 2010**, 61(3)**:** 657-671.

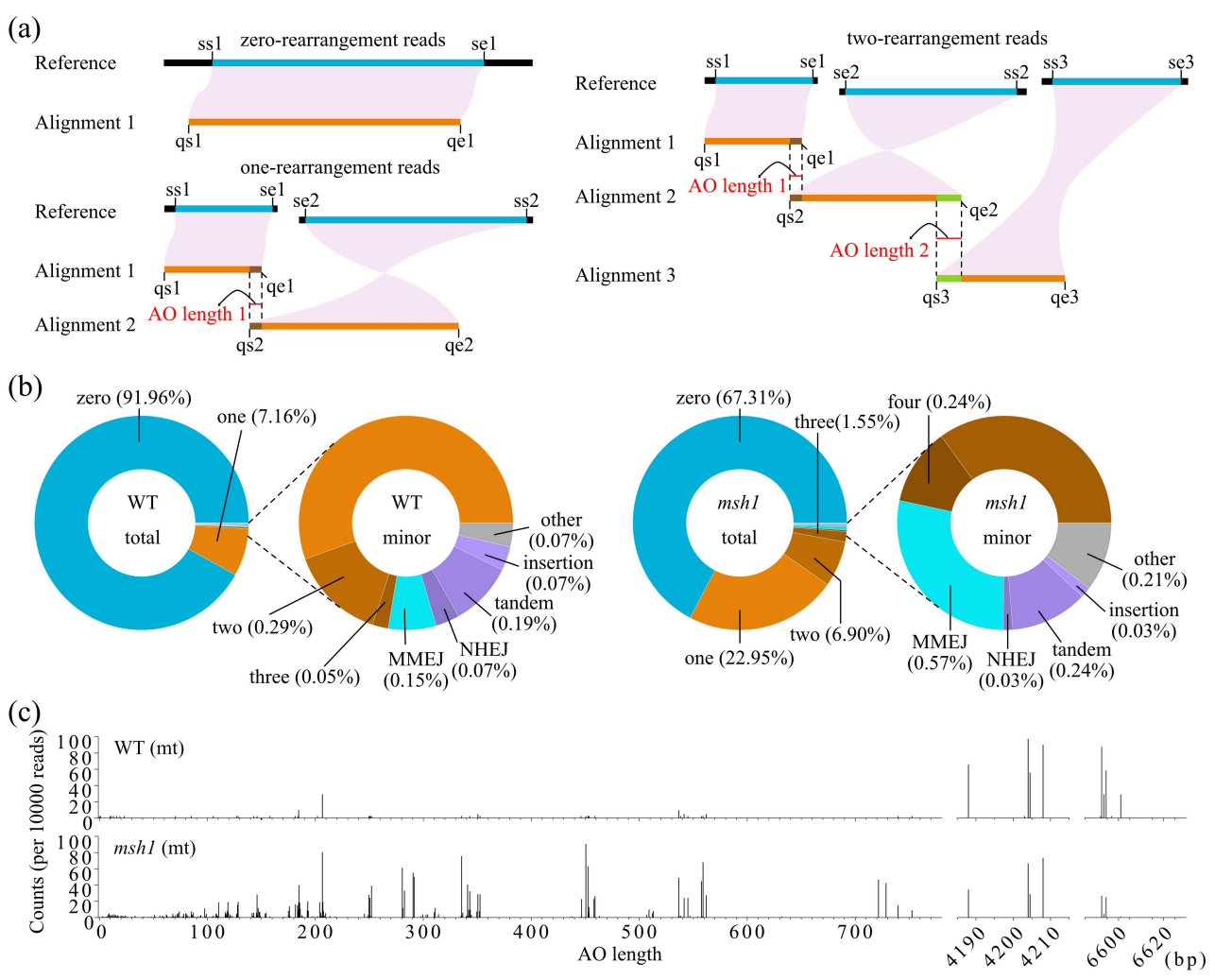
876 59. Wu Z, Liao X, Zhang X, Tembrock LR, Broz A. Genomic architectural variation of plant
877 mitochondria—A review of multichromosomal structuring. **J Syst Evol 2022**, 60(1): 160-
878 168.
879 60. Wu Z, Waneka G, Broz AK, King CR, Sloan DB. *MSH1* is required for maintenance of
880 the low mutation rates in plant mitochondrial and plastid genomes. **Proc Natl Acad Sci U**
881 **S A 2020**, 117(28)**:** 16448-16455.
882 61. Wynn E, Purfeerst E, Christensen A. Mitochondrial DNA Repair in an *Arabidopsis*
883 *thaliana* Uracil N-Glycosylase Mutant. **Plants (Basel) 2020**, 9(2).
884 62. Wynn EL, Christensen AC. Repeats of Unusual Size in Plant Mitochondrial Genomes:
885 Identification, Incidence and Evolution. **G3 (Bethesda) 2019**, 9(2)**:** 549-559.
886 63. Xu YZ, Arrieta-Montiel MP, Virdi KS, de Paula WB, Widhalm JR, Basset GJ, Davila JI,
887 Elthon TE, Elowsky CG, Sato SJ, Clemente TE, Mackenzie SA. MutS HOMOLOG1 is a
888 nucleoid protein that alters mitochondrial and plastid properties and plant response to high
889 light. **Plant Cell 2011**, 23(9)**:** 3428-3441.
890 64. Yang X, Mackenzie SA. Many Facets of Dynamic Plasticity in Plants. **Cold Spring Harb**
891 **Perspect Biol 2019**, 11(10).
892 65. Yu R, Sun C, Zhong Y, Liu Y, Sanchez-Puerta MV, Mower JP, Zhou R. The minicircular
893 and extremely heteroplasmic mitogenome of the holoparasitic plant *Rhopalocnemis*
894 *phalloides*. **Curr Biol 2022**, 32(2)**:** 470-479 e475.
895 66. Zhao N, Li Z, Zhang L, Yang X, Mackenzie SA, Hu Z, Zhang M, Yang J. MutS
896 HOMOLOG1 mediates fertility reversion from cytoplasmic male sterile *Brassica juncea*
897 in response to environment. **Plant Cell Environ 2021**, 44(1)**:** 234-246.
898 67. Zhao N, Xu X, Wamboldt Y, Mackenzie SA, Yang X, Hu Z, Yang J, Zhang M. MutS
899 HOMOLOG1 silencing mediates *ORF220* substoichiometric shifting and causes male
900 sterility in *Brassica juncea*. **J Exp Bot 2016**, 67(1)**:** 435-444.
901
902
903 **Figure Legends**
904 **Figure 1** Mapping of HiFi reads identified structural variants in Arabidopsis mt genomes
905 (a) Schematics show read alignment to the reference genome by blastn. Reads mapped in their
906 entirety to one position indicate zero-rearrangement events. Reads mapped to two discontinuous
907 positions indicate a single rearrangement. Start and end coordinates of blastn queries are shown as
908 qs1, qe1 for Alignment 1, and qs2, qe2 for Alignment 2. Start and end coordinates of corresponding
909 blastn subjects are shown as ss1, se1 for Alignment 1, and ss2, se2 for Alignment 2. Alignment

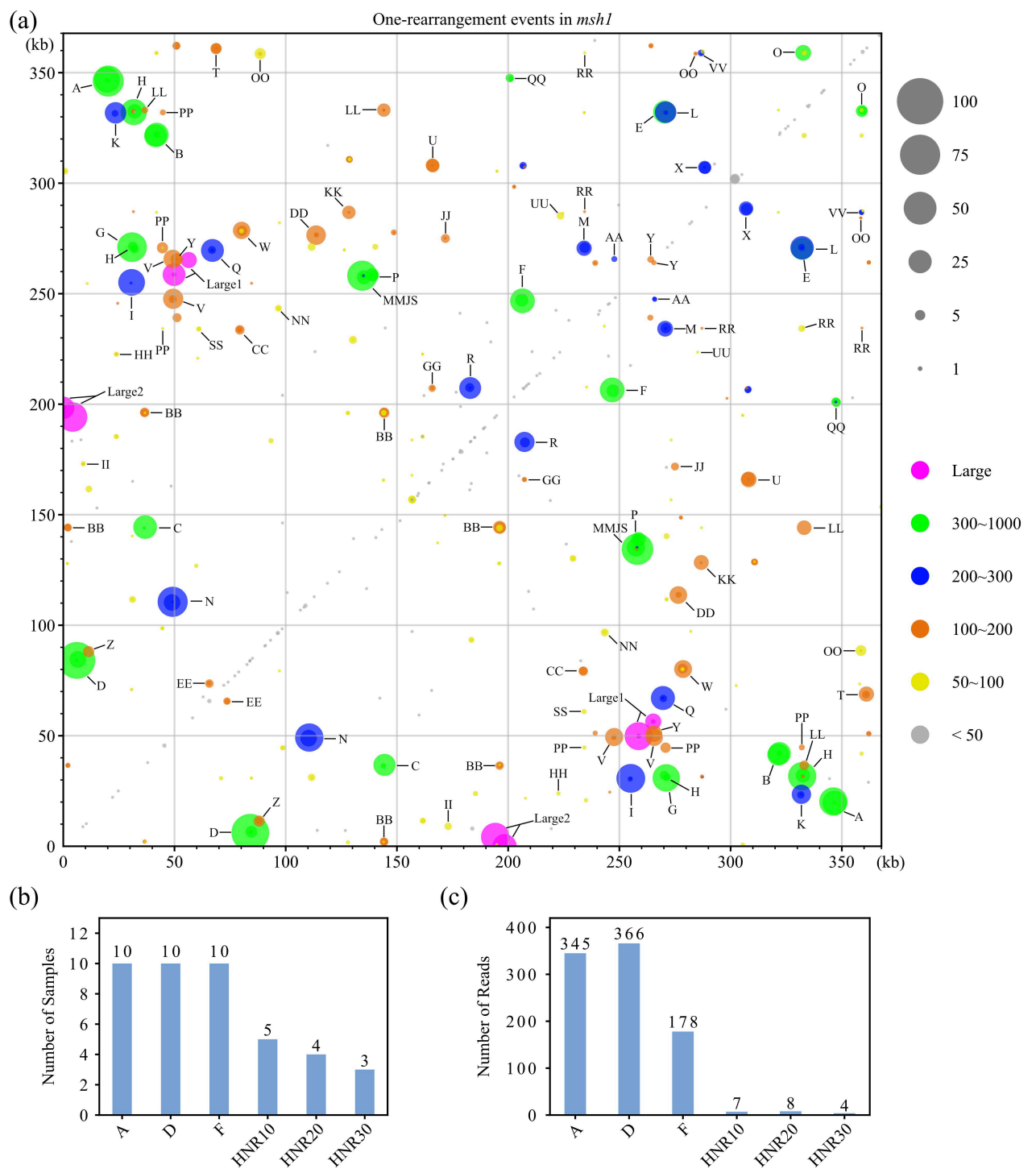
910 overlap length (AO length) 1 in red color equals "qe1 - qs2 + 1". The overlapping regions of two
911 alignments of successive parts of the same read correspond to sequence homologies such as non-
912 tandem repeats. Coordinates of se1 and ss2 in the genome correspond to repeat boundaries of the
913 rearrangement event. Reads mapped to three discontinuous positions indicate two rearrangements
914 in a single read. (b) Plots are proportions of different types of mt genome rearrangements detected
915 by HiFi reads in wild-type and *msh1* mutants; zero, one, two, three, and four denote ratios of zero-,
916 one-, two-, three-, four- rearrangement events; MMEJ, microhomology-mediated end joining (2-
917 49 bp homology sequence size); NHEJ, non-homologous end-joining (0-1 bp homology sequence
918 size); tandem, large-scale tandem repeats-mediated variants; insertion, large insertions (mostly
919 poly-G and poly-C tracts). The percentages of rearrangement categories are shown in parentheses.
920 The total circles represent all reads, and the minor circles represent the last 2% of total reads to
921 show low-frequency categories. (c) The distribution of AO length in mt one-rearrangement events
922 detected by HiFi reads. Counts from read groups with equal AO lengths are summed. Reads from
923 different *msh1* individuals are merged. Read counts are normalized to 10000 total mt genome-
924 mapped reads.
925
926 **Figure 2** Patterns of one-rearrangement mt reads detected in *msh1* mutants
927 (a) Dot plots show global patterns of the junctions and relative read counts of mt one-rearrangement
928 reads in *msh1* mutants. Signals from previously named repeats are labeled. In *msh1*, recombination
929 events mediated by intermediate-sized repeats are common. Reads from different *msh1* individuals
930 are merged. There are 5544 one-rearrangement reads out of 24162 total reads in all *msh1* samples.
931 Colors represent categories of alignment overlap length (AO length). The x position of each dot is
932 the end coordinate of the blastn subject for Alignment 1 (se1). The y position of each dot is the start
933 coordinate of the blastn subject for Alignment 2 (ss2). The sizes of dots denote read counts
934 normalized to 10000 total mt genome-mapped reads per sample. The near symmetrical appearance
935 of the plot is a result of HiFi reads aligned in both forward and reverse directions relative to the
936 reference genome. The spots in mirrored positions should not be interpreted as quantifications of
937 the two reciprocal recombination products. (b, c) Rearrangements mediated by newly named
938 repeats are supported by (c) fewer reads and are observed in (b) fewer samples.
939
940 **Figure 3** Asymmetrical recombination events and accompanied loss of parental forms
941 Stacked bar plots show HiFi read counts indicating parental forms (1-1, 2-2, 3-3) and crossover
942 recombination products (1-2, 2-1, 1-3, 3-1, 2-3, 3-2) of representative repeats (Large1, Large2,
943 MMJS, D, F, Q, E, H) in wild-type and *msh1* mutants. The unequal counts 1-2 versus 2-1 or 1-3

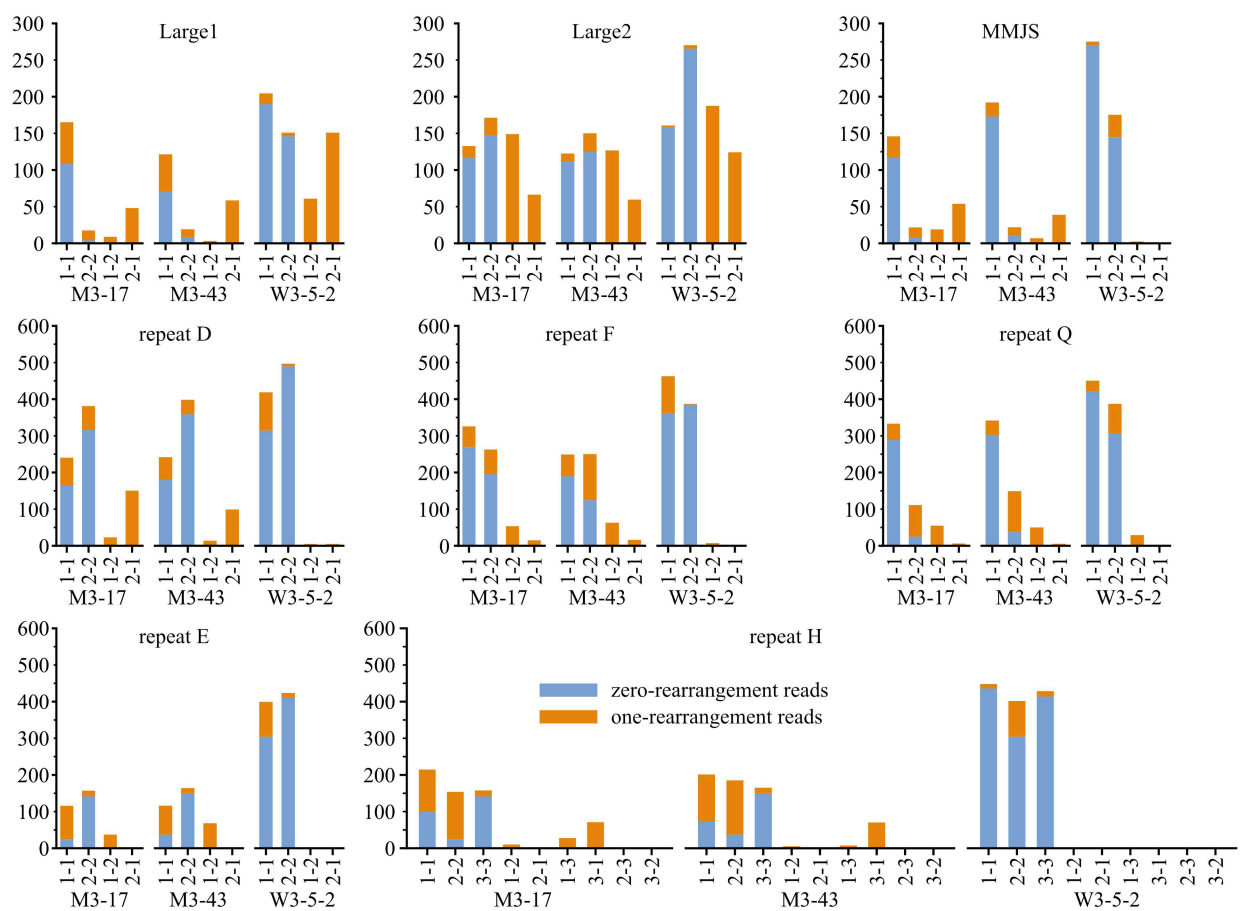
944 versus 3-1 in each repeat group indicate asymmetrical recombination or simply preferential
945 accumulation of only one of the crossover products, which is accompanied by a striking loss of 2-
946 2 versus 1-1 parental forms in Large1, MMJS, and Q. Due to the locations of H-2 and H-3 within
947 E-1 and E-2, recombination events between H-2 and H-3 were assigned to the larger two-copy
948 repeats E-1 and E-2 and, thus, appear as zero values in this plot. Both zero-rearrangement reads
949 (blue bars) and one-rearrangement reads (orange bars) were used for calculation. Read counts are
950 normalized to 10000 total mt genome-mapped reads per sample. Reads from different *msh1*
951 families are merged.
952
953 **Figure 4** Exchange of SNVs/indels between imperfect repeats is suppressed by *MSH1*
954 (a) Blue circles denote data points from each group of two-copy imperfect repeats. The x coordinate
955 represents the length of the longest identical fragment (LIF) as a percentage of the total length of
956 the repeat. The y coordinate represents the corresponding percentage of recombinant reads for that
957 repeat pair. Blue line denotes the linear regression line. The diagonal 1:1 line denotes equal ratios.
958 (b) Double-strand breaks (DSBs) in DNA can be repaired by allelic or non-allelic dsDNA templates.
959 The purple and green boxes are a pair of imperfect repeats that mediate ectopic homology-directed
960 repair (HDR). The events are not detected (unseen) if allelic templates are involved.
961 Rearrangements or gene conversion events involving non-allelic templates can be detected by
962 sequencing. (c) Discontinuous patterns of SNVs/indels in a proportion of one-rearrangement
963 chimeric reads of repeat H-1 and H-3 may indicate complex patterns of exchange in *msh1* mutants
964 (data from more reads found according to Table S9). The colors of horizontal bars indicate the most
965 parsimonious explanations for the SNVs/indels according to repeat flanking regions. Asterisks
966 denote SNVs potentially introduced by gene conversion from repeat homolog H-2 in these H-1/H-
967 3 mediated one-rearrangement reads. Numbers in parentheses denote the sum of read counts in all
968 samples without normalization. The top three lines in panel c are to scale with vertical bars
969 indicating bases of the same colors as the letters. The lines below are not drawn to scale to avoid
970 crowding. Vertical bars in panels a and c denote A (green), T (red), G (brown), C (blue) for SNVs
971 and I (gray) for indel sites. (d) Square-shaped stacked bar plots are proportions of non-allelic
972 SNVs/indels at predicted sites in repeat group H, which illustrate unsuppressed accumulation in
973 *msh1* mutant samples compared to wild-type (W3-5-2, this study; XJTU, Wang et al., 2021; CEN,
974 Naish et al., 2021). The relative heights in each square-shaped box indicate the ratios of given bases
975 in zero-rearrangement reads for A (green), T (red), G (brown), C (blue), and I for reference (gray)
976 or non-allelic (purple) sequences at indel sites. Reference bases are indicated below. SNVs and
977 indels are arranged from small to large coordinates in the reference. Coordinates of the first and

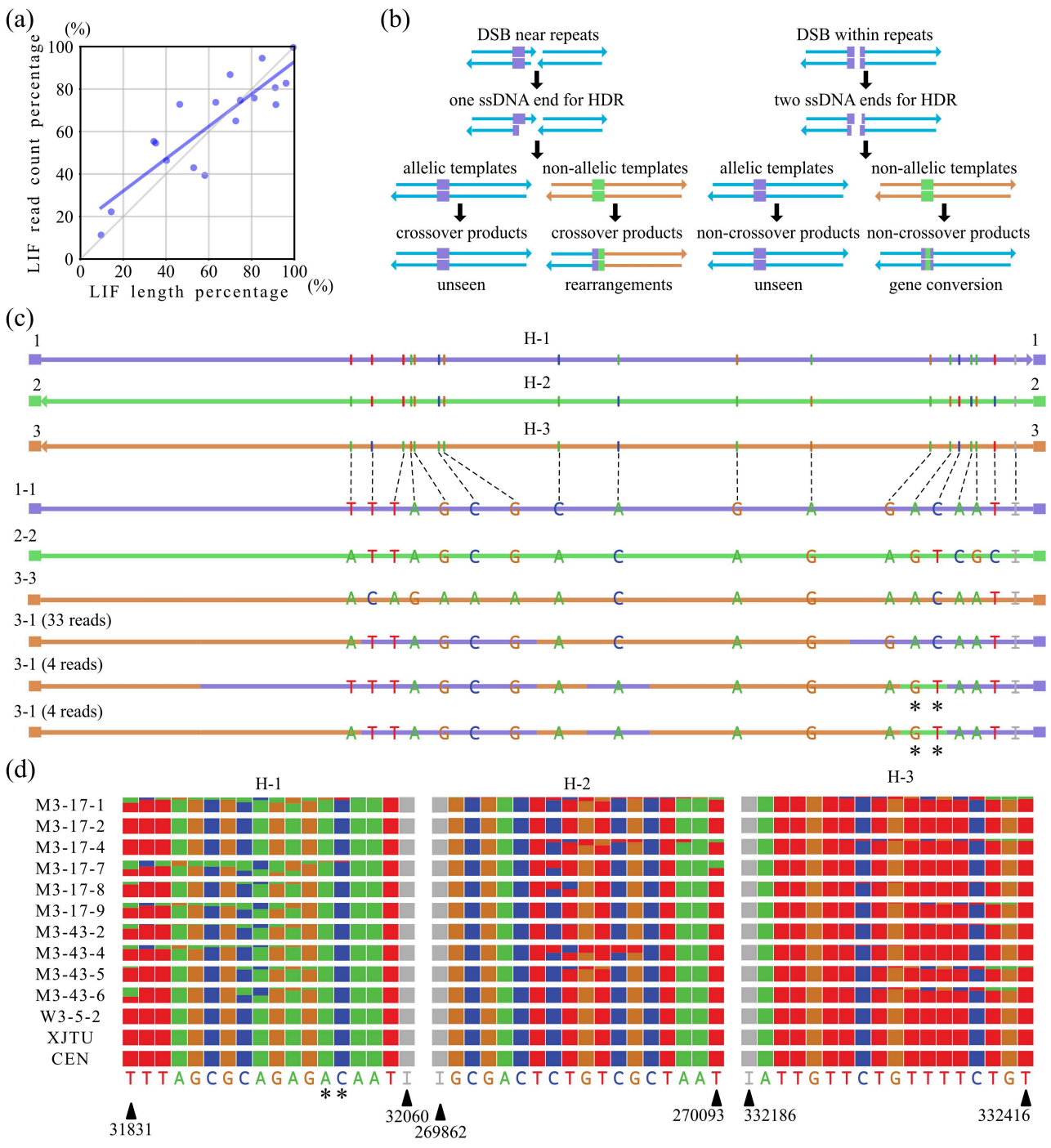
978 last SNVs/indels (black triangles) are shown. H-2 and H-3 are in the reverse orientation compared
979 to H-1. Asterisks denote the same labeled positions as in panel c.
980
981 **Figure 5** Proposed models for the generation of new mt genome configurations and large repeats
982 (a) Five rearrangements can explain the structural differences and three large deletions (1002-bp,
983 15652-bp, 55-bp) that distinguish the Col-0 and An-1 mt genomes. An additional rearrangement is
984 needed to explain the large deletions (5920-bp, 15652-bp, 944-bp) in L*er*. Note that all three copies
985 of repeat group H are involved. The 1105-bp sequence absent in Col-0 was not considered in this
986 model. (b) A two-step "excision-reinsertion" mechanism for the generation of large repeats in mt
987 genomes. Note that microhomologies with three copies (or more) are needed.
988
989 **Figure 6** Mutations and structural variants in the pt genome detected by HiFi reads
990 (a) The partial palindromic sequence contains a stem (orange arrow) and a loop (black arrow) and
991 is also a pair of repeats or microhomologies with both copies at the same location but in inverted
992 orientation. The original sequence can be viewed as repeat homolog 1 (1-1), and at the same time
993 as repeat homolog 2 (2-2) in the opposite direction. MMEJ events mediated by the most active
994 partial palindromic sequence in *msh1* mutants were shown. Numbers in brackets denote the sum of
995 read counts showing crossover site distribution in chimeric reads in all samples without
996 normalization. Note that there are four reads showing the breakpoint in the middle of the loop. The
997 ratio of one-rearrangement reads versus zero-rearrangement reads across this partial palindromic
998 sequence is 160/14786 (1.1% of total reads) in *msh1* and 0/1966 in wild-type. (b) A NHEJ event
999 occurring only in the M3-43-5 individual was predicted to disrupt the *petA* gene. The ratio of one-
1000 rearrangement reads versus zero-rearrangement reads across this deletion is 121/1385 (8.0% of
1001 total reads). (c) Square-shaped stacked bar plots are proportions of AT-to-GC transitions in 17
1002 GATK-called SNVs in wild-type and *msh1* mutant individuals, which were enriched in one or a
1003 few samples at high ratios. The relative heights in each square-shaped box indicate the ratio of a
1004 given base for A (green), T (red), G (brown), and C (blue). Reference bases are indicated below.
1005 (d) Heatmap shows the proportions of length variants at a 13-bp homopolymer started near position
1006 67164.
1007
1008 **Figure 7** Proposed models of mt/pt genome evolution controlled by fluctuating *MSH1* levels
1009 (a) MSH1 appears to suppress both crossover and non-crossover forms of ectopic recombination
1010 by detecting mismatches within repeats or at repeat boundaries if the two copies of repeats are
1011 identical. (b) In the wild-type (Step 1), MSH1 maintains the structural and sequence-level

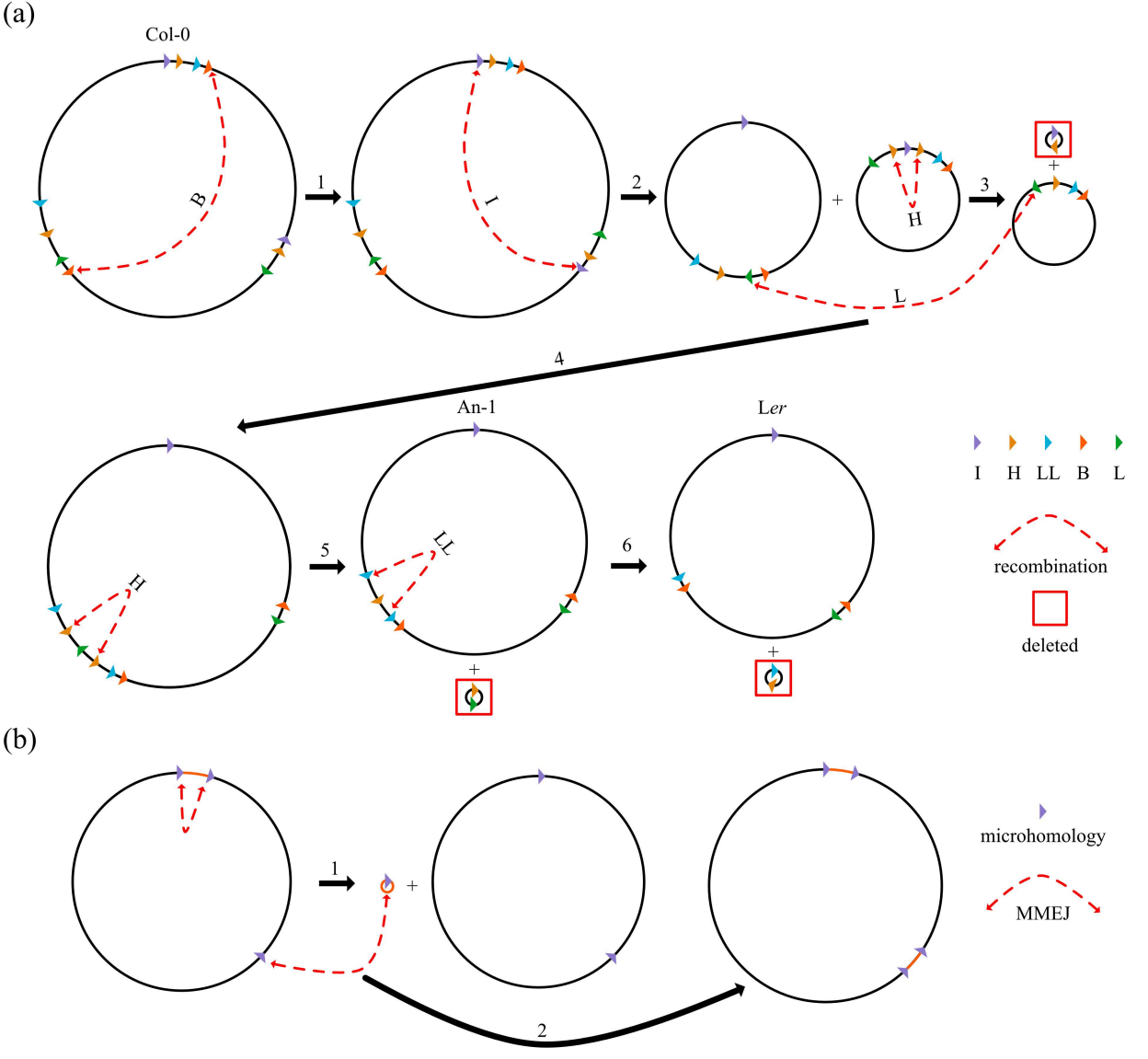
1012 homoplasmy by repressing ectopic recombination (and by removing other types of variants
1013 resulting from polymerase misincorporations and/or DNA damage). It has been previously
1014 hypothesized that MSH1 performs these functions by recognizing mismatches in DNA and
1015 introducing DSBs to (re)initiate recombinational repair. When *MSH1* is disrupted either genetically
1016 or environmentally (Step 2), organellar genomes become highly heteroplasmic. When MSH1
1017 activity is restored (Step 3), both genomes return to a low-homoplasmy state. The new mt genome
1018 undergoes rearrangements and large deletions involving reproducible sets of repeats. Given the low
1019 abundance of intermediate-sized repeats in Arabidopsis pt genome, altered pt genome often differ
1020 at random sites relating to MMEJ/NHEJ, AT-to-GC mutation, or length variants at simple sequence
1021 repeats.

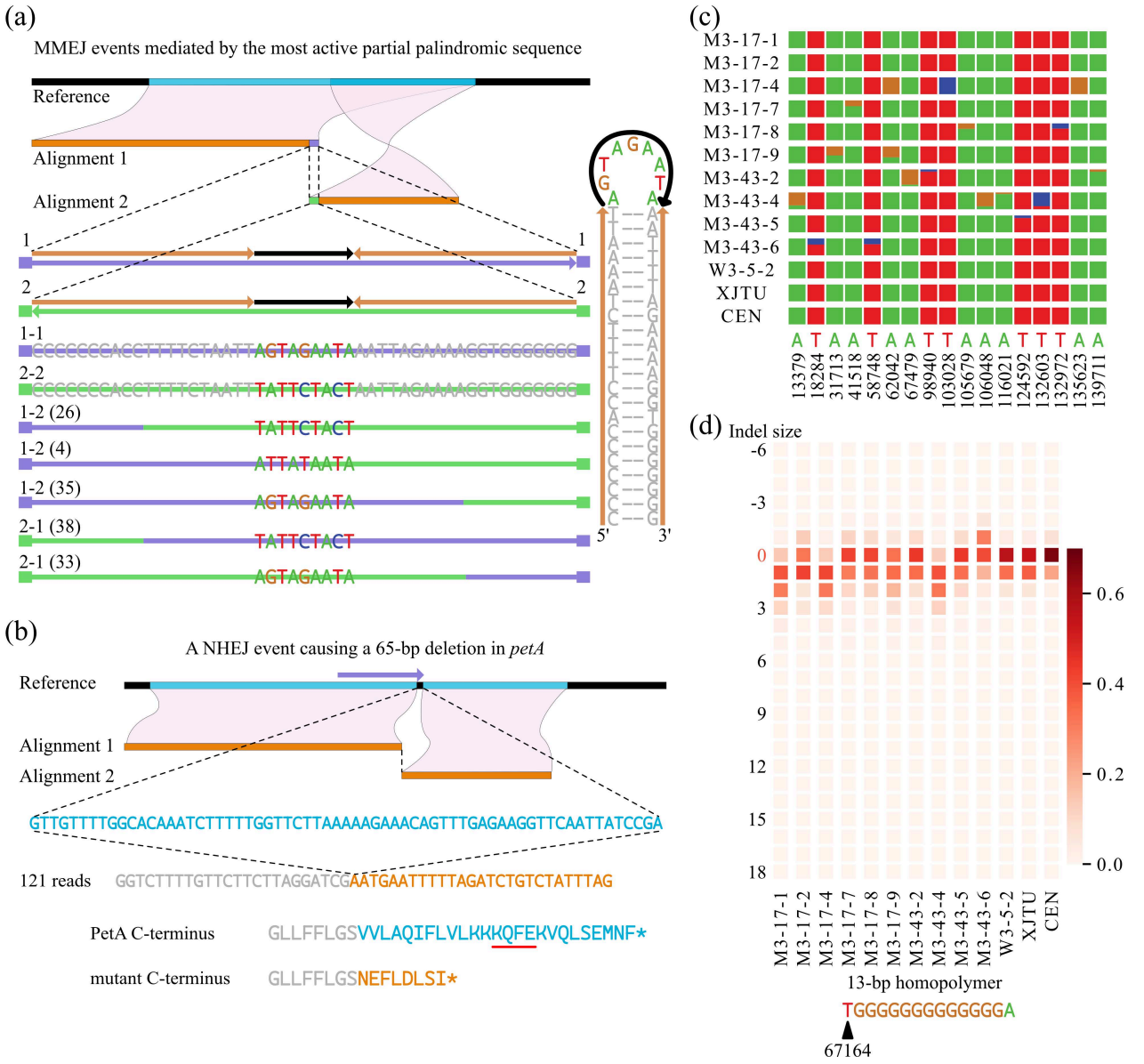

(a)
zero-rearrangement reads
Reference
ss1
se1
Alignment 1
qs1
qe1
one-rearrangement reads
Reference
ss1
se1
se2
ss2
Alignment 1
qs1
qe1
AO length 1
Alignment 2
qs2
qe2
two-rearrangement reads
Reference
ss1
se1
se2
ss2
ss3
se3
Alignment 1
qs1
qe1
AO length 1
Alignment 2
qs2
qe2
AO length 2
Alignment 3
qs3
qe3
(b)
zero (91.96%)
one (7.16%)
WT
total
two (0.29%)
three (0.05%)
WT
minor
other (0.07%)
insertion (0.07%)
tandem (0.19%)
NHEJ (0.07%)
MMEJ (0.15%)
zero (67.31%)
three(1.55%)
four (0.24%)
msh1
total
one (22.95%)
two (6.90%)
msh1
minor
other (0.21%)
insertion (0.03%)
tandem (0.24%)
NHEJ (0.03%)
MMEJ (0.57%)
(c)
Counts (per 10000 reads)
WT (mt)
msh1 (mt)
0
20
40
60
80
100
0
100
200
300
400
500
600
700
AO length
4190
4200
4210
6600
6620
(bp)


(a)
One-rearrangement events in msh1
(kb)
Large1
Large2
100
75
50
25
5
1
Large
300~1000
200~300
100~200
50~100
< 50
(b)
Number of Samples
10
10
10
5
4
3
A
D
F
HNR10
HNR20
HNR30
(c)
Number of Reads
345
366
178
7
8
4
A
D
F
HNR10
HNR20
HNR30

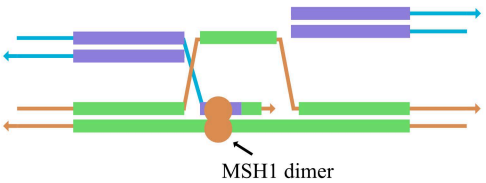

Large1
Large2
MMJS
repeat D
repeat F
repeat Q
repeat E
repeat H
zero-rearrangement reads
one-rearrangement reads
M3-17
M3-43
W3-5-2
1-1
2-2
1-2
2-1
1-3
3-1
2-3
3-2
3-3


(a)
(%)
LIF read count percentage
LIF length percentage
(%)
(b)
DSB near repeats
one ssDNA end for HDR
allelic templates
non-allelic templates
crossover products
crossover products
unseen
rearrangements
DSB within repeats
two ssDNA ends for HDR
allelic templates
non-allelic templates
non-crossover products
non-crossover products
unseen
gene conversion
(c)
H-1
H-2
H-3
1-1
2-2
3-3
3-1 (33 reads)
3-1 (4 reads)
3-1 (4 reads)
(d)
H-1
H-2
H-3
M3-17-1
M3-17-2
M3-17-4
M3-17-7
M3-17-8
M3-17-9
M3-43-2
M3-43-4
M3-43-5
M3-43-6
W3-5-2
XJTU
CEN
TTTAGCGCAGAGACAATI
IGCGACTCTGTCGCTAAT
IATTGTTCTGTTTTCTGT
31831
32060
269862
270093
332186
332416


(a)
Col-0
B
1
I
2
+
H
3
+
L
4
An-1
H
5
LL
+
6
Ler
+
I
H
LL
B
L
recombination
deleted
(b)
1
+
2
microhomology
MMEJ

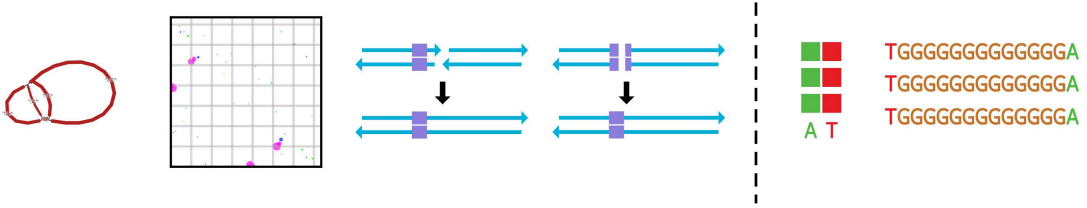

(a)
MMEJ events mediated by the most active partial palindromic sequence
Reference
Alignment 1
Alignment 2
1-1
2-2
1-2 (26)
1-2 (4)
1-2 (35)
2-1 (38)
2-1 (33)
5'
3'
(b)
A NHEJ event causing a 65-bp deletion in petA
Reference
Alignment 1
Alignment 2
GTTGTTTTGGCACAAATCTTTTTGGTTCTTAAAAAGAAACAGTTTGAGAAGGTTCAATTATCCGA
121 reads
GGTCTTTTGTTCTTCTTAGGATCGAATGAATTTTTAGATCTGTCTATTTAG
PetA C-terminus
GLLFFLGSVVLAQIFLVLKKKQFEKVQLSEMNF*
mutant C-terminus
GLLFFLGSNEFLDLSI*
(c)
M3-17-1
M3-17-2
M3-17-4
M3-17-7
M3-17-8
M3-17-9
M3-43-2
M3-43-4
M3-43-5
M3-43-6
W3-5-2
XJTU
CEN
A T A A T A A T T A A A T T T A A
13379 18284 31713 41518 58748 62042 67479 98940 103028 105679 106048 116021 124592 132603 132972 135623 139711
(d)
Indel size
-6
-3
0
3
6
9
12
15
18
0.6
0.4
0.2
0.0
M3-17-1 M3-17-2 M3-17-4 M3-17-7 M3-17-8 M3-17-9 M3-43-2 M3-43-4 M3-43-5 M3-43-6 W3-5-2 XJTU CEN
13-bp homopolymer
TGGGGGGGGGGGGGA
67164

(a)

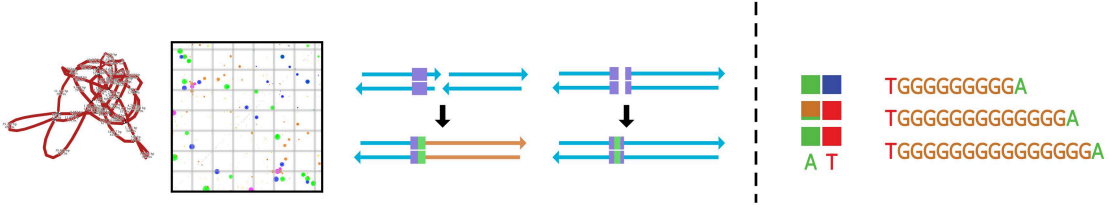


(b)

1 MSH1 maintains the structural and sequence-level homoplasmy of mt/pt genomes



2 Distinctive patterns of genome instabilities in mt/pt genomes without MSH1



3 Evolved versions of mt/pt genomes after MSH1 restoration