Classification: Biological Sciences, Evolution

Title: Stable and widespread structural heteroplasmy in chloroplast genomes revealed
by a new long-read quantification method

Weiwen Wang^a, Robert Lanfear^a

^a Research School of Biology, Australian National University, Canberra, ACT,
Australia, 2601

Corresponding Author:

Weiwen Wang, wei.wang@anu.edu.au

Robert Lanfear, rob.lanfear@anu.edu.au, +61 2 6125 2536

Keywords: Single copy inversion, flip-flop recombination, chloroplast genome
structural heteroplasmy

Abstract

The chloroplast genome usually has a quadripartite structure consisting of a large single copy region and a small single copy region separated by two long inverted repeats. It has been known for some time that a single cell may contain at least two structural haplotypes of this structure, which differ in the relative orientation of the single copy regions. However, the methods required to detect and measure the abundance of the structural haplotypes are labour-intensive, and this phenomenon remains understudied. Here we develop a new method, Cp-hap, to detect all possible structural haplotypes of chloroplast genomes of quadripartite structure using long-read sequencing data. We use this method to conduct a systematic analysis and quantification of chloroplast structural haplotypes in 61 land plant species across 19 orders of Angiosperms, Gymnosperms and Pteridophytes. Our results show that there are two chloroplast structural haplotypes which occur with equal frequency in most land plant individuals. Nevertheless, species whose chloroplast genomes lack inverted repeats or have short inverted repeats have just a single structural haplotype. We also show that the relative abundance of the two structural haplotypes remains constant across multiple samples from a single individual plant, suggesting that the process which maintains equal frequency of the two haplotypes operates rapidly, consistent with the hypothesis that flip-flop recombination mediates chloroplast structural heteroplasmy. Our results suggest that previous claims of differences in chloroplast genome structure between species may need to be revisited.

22

23 **Significance Statement**

24 Chloroplast genome consists of a large single copy region, a small single copy region,
 25 and two inverted repeats. Some decades ago, a discovery showed that there are two
 26 types chloroplast genome in some plants, which differ the way that the four regions
 27 are put together. However, this phenomenon has been shown in just a small number of
 28 species, and many open questions remain. Here, we develop a fast method to measure
 29 the chloroplast genome structures, based on long-reads. We show that almost all
 30 plants we analysed contain two possible genome structures, while a few plants contain
 31 only one structure. Our findings hint at the causes of the phenomenon, and provide a
 32 convenient new method with which to make rapid progress.

33

34

35 **Introduction**

36 Chloroplasts are organelles which are vital for photosynthesis. Most land plant
 37 chloroplast genomes are 120 – 160 kb in size (1, 2), and have a quadripartite structure
 38 consisting of a pair of identical rRNA-containing inverted repeats (hereafter referred
 39 to as IR) of ~10-30kb divided by a large single copy (LSC) region of ~80 -90 kb and a
 40 small single copy (SSC) region of ~10-20 kb.

41

42 Surprisingly, chloroplast genomes can exist two structural haplotypes differing in the
 43 orientation of single copy regions (3). The presence of this structural heteroplasmy
 44 has been confirmed in some land plants (3-6) and algae (7, 8), but its cause remains

45 unknown. One hypothesis to explain the presence of structural heteroplasmy is known
46 as flip-flop recombination (4). This hypothesis suggests that the large IRs could
47 mediate frequent intramolecular recombination, resulting in the maintenance of
48 roughly equal amounts of the two haplotypes differing only in the orientation of their
49 single copy regions.

50

51 A better understanding of chloroplast genome structural heteroplasmy is important for
52 a number of reasons. First, some recent papers have suggested that the orientation of
53 the single copy regions differ between species (9-13), but Emery *et al* (14) pointed out
54 that these studies seem to have overlooked the possibility that both orientations may
55 coexist in a single individual. Second, the relationship between the structure of the
56 chloroplast genome and the existence or otherwise of structural heteroplasmy remains
57 poorly understood. For example, if flip-flop recombination causes chloroplast
58 structural heteroplasmy, then it seems likely that the presence of two long IRs may be
59 a pre-requisite for the existence of heteroplasmy. Consistent with this, no
60 heteroplasmy was observed in a chloroplast genome with highly reduced IRs (15), but
61 the generality of this observation remains to be tested. Third, if heteroplasmy is the
62 norm rather than the exception, then this may represent a challenge for the assembly
63 of chloroplast genomes, which may be easily overcome by simply allowing for the
64 existence of two structural haplotypes during the assembly process. A large-scale
65 analysis across many different species has the potential to provide a more complete
66 picture of chloroplast heteroplasmy, further elucidating this fascinating phenomenon

67 and potentially improving genome assembly and inference. In this study, we perform
68 this large-scale analysis using a new method which we developed to quickly and
69 conveniently quantify structural heteroplasmy in chloroplast genomes from long-read
70 sequencing data.

71

72 Currently there are two methods to detect different structural heteroplasmy in
73 chloroplast genomes: Bacterial Artificial Chromosomes (BAC)-End-Sequence (BES)
74 (5), and restriction digests (4). For the BES method, one first constructs BAC libraries
75 in which large pieces of chloroplast DNA are inserted into bacterial chromosomes.
76 One then uses the BAC libraries to sequence short fragments of both ends of many
77 such pieces of chloroplast DNA to ascertain which chloroplast structures exist in a
78 given sample. BES reads can only provide information on chloroplast genome
79 structure when a single read covers an entire IR region, with one end in LSC region
80 and another end in the SSC region. Although a very useful method, BES reads may be
81 limited for detecting highly atypical chloroplast genome structures, because only a
82 short fragment from each end of a read is actually sequenced. For the restriction
83 digestion method, one uses restriction enzymes to digest the chloroplast genome, and
84 then decodes the chloroplast genome structure by studying the distribution of the
85 resulting fragment lengths using agarose gel electrophoresis. Using this approach,
86 Stein *et al* (4) found that the IR region was present in four different fragments which
87 could be separated into two groups for which the sum of lengths was equal. They
88 therefore concluded that chloroplast genome contained two equimolar isomers, which

89 differed in their single copy orientation. The restriction digest method has provided
 90 much of the existing information on chloroplast genome heteroplasmy, but it is
 91 limited by the availability of suitable restriction sites (which may be unknown in
 92 some species), and requires a labour-intensive hybridisation step to infer chloroplast
 93 genome heteroplasmy. Other methods have also been proposed, such as the use of
 94 PCR to attempt to amplify diagnostic regions of the chloroplast genome. However, it
 95 is not easy to generate PCR fragments which are longer than the IR region (usually
 96 10-30 kb), so while PCR-based methods could only work well with short IRs (e.g. of
 97 a few hundred bp) (15), their general application is likely to remain limited.
 98 Furthermore, PCR-mediated recombination could result in false positives when using
 99 PCR-based methods (16, 17). In short, current methods for detecting structural
 100 heteroplasmy in chloroplast genomes are labour- and time-intensive, which makes
 101 them difficult to employ for broad-scale studies. Current methods have been applied
 102 in a relatively small number of land plant species (*Beta vulgaris* (6), *Musa acuminata*
 103 (5), *Phaseolus vulgaris* (3), *Osmunda claytoniana*, *Osmunda cinnamomea* and
 104 *Osmunda regalis* (4), and some species in the Pinaceae (15)), and as a result our
 105 understanding of the patterns and causes of structural heteroplasmy in chloroplast
 106 genomes remains somewhat limited.

107

108 In this study, we develop and apply a fast, simple, and cheap method to detect and
 109 quantify structural heteroplasmy in chloroplast genomes using long-read sequencing.
 110 This method requires just a single DNA extraction and sequencing run, followed by a

111 simple bioinformatic analysis for which we provide free and open-source code
 112 (<https://github.com/asdcid/Cp-hap>). Our approach, which we call Cp-hap, relies on
 113 sequencing individual DNA molecules longer than the length of the IR region
 114 (roughly 10-30 kb in most species). Sequencing reads of this length are commonly
 115 obtained in great abundance when sequencing whole-plant DNA extractions on the
 116 Oxford Nanopore MinION device (18, 19). Our method sifts through these long
 117 sequencing reads to find those that uniquely match one of the 32 uniquely-identifiable
 118 structural haplotypes (Fig. S1), and then uses the abundance of these reads to estimate
 119 the relative abundance of each haplotype (details see Materials and Methods). Since
 120 many new plant samples are being sequenced with long-read sequencing technology,
 121 this method has the potential to provide many insights in the coming years, simply as
 122 a by-product of existing sequencing efforts.

123

124 We use Cp-hap to quantify the abundance of chloroplast genome structural haplotypes
 125 in 61 plant species. Our results suggest that most land plants contain the two
 126 commonly-found chloroplast genome haplotypes in equal abundance whenever the
 127 chloroplast genomes contain a pair of large IRs. However, when the IR is small (e.g.
 128 only a few hundred bp in size) or repeat regions are not inverted but positioned, we
 129 find only a single haplotype. Finally, we show that multiple samples taken from a
 130 single individual all contain the same two haplotypes in equal abundance, suggesting
 131 that the process which maintains equimolar concentrations of the two haplotypes
 132 operates relatively rapidly, and consistent with the suggestion that flip-flop

133 recombination is the underlying cause.

134

135 **Results**

136 **Data collection**

137 We searched the National Center for Biotechnology Information (NCBI) Sequence
138 Read Archive (SRA) database, and selected 90 land plants species which met our
139 length and data amount requirement (see Materials and Methods) to analysis (Dataset
140 S1). Two of these species (*Herrania umbratica* and *Siraitia grosvenorii*) lacked
141 suitable chloroplast reference genomes, so we therefore assembled these two
142 chloroplast genomes *de-novo* (see Materials and Methods). The chloroplast genomes
143 of *Herrania umbratica* and *Siraitia grosvenorii* were 158,475 bp and 158,757 bp in
144 size, consisting of a LSC (88,369 bp and 87,625 bp), a SSC (19,040 bp and 18,556
145 bp), and two IRs (25,533 bp and 26,288 bp), and encoding 134 and 133 genes
146 respectively (Fig. S2).

147

148 **Chloroplast structural haplotypes across the plant tree of life**

149 For these 90 species, we obtained adequate long-read sequencing data to assess
150 chloroplast genome structural heteroplasmy for 61 land plants species. Across all 61
151 species, we find just three structural haplotypes (Fig. 1, Table 1). All 58 of the
152 Angiosperms in our dataset are heteroplasmic, and contain roughly equal frequencies
153 of haplotypes A and B (binomial test for departure from equal frequency, p
154 value >0.05 in all cases). The remaining three species are not heteroplasmic (binomial
155 test for departure from equal frequency $p = 2.2 \times 10^{-16}$ in all cases). The two
156 Gymnosperms in our dataset both contain only haplotype B, and one of the

157 Pteridophytes in our dataset (*Selaginella tamariscina*) contains only haplotype C. Fig.
158 2 shows clearly that, as expected, there is a trend for datasets with larger sample sizes
159 to show relative haplotype abundances closer to exactly 50/50.
160
161 Across 8 different samples taken from a single large *Eucalyptus pauciflora* plant
162 (Table 1), we detected roughly equal frequencies of haplotypes A and B in all 8
163 samples (binomial p-values > 0.05 in all cases), no evidence for differences in relative
164 frequencies among samples (chi-square p-value > 0.05), and no evidence for
165 phylogenetic signal in the any of the relative abundances (p-value > 0.05 in all tests),
166 where the underlying topology and branch lengths are taken to be the physical
167 structure of the tree itself (see Materials and Methods).

168 Discussion

169 In this study, we develop a new method to detect and quantify structural
 170 heteroplasmies in the chloroplast genome. By applying this method to a broad range
 171 of plant species including representatives from Angiosperms, Gymnosperms and
 172 Pteridophytes, we show that of the large number of possible structural haplotypes,
 173 only a very small number are observed in nature. For example, species with IRs
 174 almost universally contain the same two structural haplotypes in ratios that do not
 175 depart significantly from 50/50. The only exceptions to this are species in which the
 176 repeats are not inverted (e.g. one of the Pteridophytes in our dataset), or in which the
 177 repeats are inverted but much reduced in length (e.g. both of the Gymnosperms in our
 178 dataset). In these species, we detect just one haplotype. Our results demonstrate not
 179 only that the relative abundance of the two most common haplotypes stable across the
 180 Angiosperm tree of life (i.e. all 58 Angiosperms in our dataset had a frequency that
 181 was not different from 50/50), but that it also stable within an individual plant (i.e.
 182 within a single large plant, all sampled leaves contained both haplotypes at a ratio that
 183 was not different from 50/50). This suggests that the process which maintains the
 184 equal frequencies of the two commonly-observed haplotypes operates relatively
 185 rapidly.

186

187 The most parsimonious interpretation of our results suggests that structural
 188 heteroplasmies should be extremely common across all angiosperms. In light of this, it
 189 may be necessary to re-examine many previous suggestions of structural differences

190 in the chloroplast genomes of angiosperms *between* species (9-13), as Emery *et al*(14)
 191 suggested. For example, Ibrahim *et al* (9) implied that the SSC of *Gossypium*
 192 *barbadense* was inverted when compared to *Gossypium hirsutum*. However, our study
 193 shows that both *Gossypium barbadense* and *Gossypium hirsutum* contain two
 194 structural haplotypes (haplotypes A and B (Table 1)), thus showing that there are no
 195 structural differences between these two species. Future studies may benefit from
 196 assuming at the outset that it is highly likely that the chloroplast genome of any
 197 angiosperm species, particularly if it contains two long inverted repeats, is most likely
 198 to exist in two roughly equimolar haplotypes, haplotypes A and B shown in Fig. 1.
 199 This phenomenon could also exist in Gymnosperms and Pteridophytes, but in this
 200 study we don't have enough available data to provide solid evidence to support this.

201

202 Of the 61 species for which we had sufficient data, we find just three species in which
 203 there is no evidence for structural heteroplasmy in the chloroplast genome:
 204 *Selaginella tamariscina*, *Pinus taeda* and *Picea sitchensis*. *Selaginella tamariscina* is
 205 a pteridophyte in which the repeats in the chloroplast genome are positioned in-line
 206 rather than inverted (20). The existence of a single structural haplotype in this species
 207 is therefore consistent with the hypothesis that the process which generates
 208 chloroplast structural heteroplasmy is mediated by long IRs. In the gymnosperms
 209 *Pinus taeda* and *Picea sitchensis*, the repeats are inverted but their lengths are highly
 210 reduced from the typical 10-30 kb in most species to just 399 bp and 440 bp,
 211 respectively. The existence of a single structural haplotype in these species is also

212 consistent with the flip-flop recombination theory, because it is feasible that the
213 highly reduced IRs preclude the formation of the dumbbell-like structure (4) which is
214 necessary to activate flip-flop recombination. Whether or not flip-flop recombination
215 is the underlying cause, our results confirm that chloroplast structural heteroplasmy
216 appears to require the existence of two long IRs in the chloroplast genome, and
217 therefore that groups which lack this trait would be expected to lack chloroplast
218 structural heteroplasmy. For example, chloroplast genomes in the Pinaceae and
219 Cupressophytes (two major groups of gymnosperms) usually lack one IR or have IRs
220 that are highly reduced in size (15, 21-26), and we therefore expect that these species
221 would lack chloroplast structural heteroplasmy.

222

223 In this study, we focussed on examining heteroplasmy in the quadripartite structure
224 of the chloroplast genome. However, other smaller structural heteroplasmy have
225 been observed in some chloroplast genomes, and their existence challenges the
226 conclusion that long inverted repeats are a pre-requisite for the existence of structural
227 heteroplasmy in the more general sense. Although flip-flop recombination may be
228 absent when the IR regions are small (see above), some studies have suggested that
229 homologous recombination could induce structural heteroplasmy from other much
230 smaller inverted repeats in chloroplast genomes (25, 26). For instance, the *trnQ-UUG*
231 gene (~150 bp) is duplicated in the LSC region in some *Juniperus* species, and this
232 duplication is associated with a 36 kb inversion resulting in the coexistence of two
233 isomeric chloroplast genome structures, although their abundance is unequal (0.8% -

234 5.0%) (25). Qu *et al* (26) reported a similar finding in other Cupressoideae species,
 235 and suggest that even a pair of very small inverted repeats (11 bp) could induce a 34
 236 kb inversion and resulting structural heteroplasmy in *Calocedrus macrolepis*. In light
 237 of these results, it remains unclear why the highly reduced IRs of *Pinus taeda* and
 238 *Picea sitchensis* studied here fail to induce structural heteroplasmy, while much
 239 shorter inverted repeats in regions of the chloroplast genome outside the IR region
 240 apparently do induce structural heteroplasmy. One possibility is that different
 241 processes mediate the formation of structural heteroplasms in these cases (25, 26).
 242 Another possibility is that the frequency of the alternative haplotype is too low for us
 243 to detect in *Pinus taeda* and *Picea sitchensis*. However, if this were the case the
 244 frequency of haplotype A in these species would have to be extremely low: for
 245 example in *Picea sitchensis* we detected 2845 long reads supporting haplotype B and
 246 no reads supporting haplotype A.

247
 248 The Cp-hap pipeline we present here provides a convenient method for quantifying
 249 structural heteroplasmy in chloroplast genomes. It adds to suite of existing methods
 250 such as restriction digests (3, 4) and BES (5), but differs from these methods in that it
 251 uses long sequencing reads to provide direct evidence for uniquely-identifiable
 252 chloroplast structures. It is of note that the two available long-read sequencing
 253 methods, PacBio and Oxford Nanopore sequencing, differ in their utility for detecting
 254 chloroplast structural heteroplasmy. For example, only ~63% (46 out of 73 species) of
 255 the PacBio datasets we used in this study contained enough long chloroplast reads to

adequately measure chloroplast structural heteroplasmy. In contrast, ~88% (15 out of 17 species) of Oxford Nanopore datasets provided enough long-read to measure the chloroplast genome heteroplasmy. Given the increasingly popularity of Oxford Nanopore sequencing, and ongoing improvements in the read-lengths available from PacBio sequencers, we hope that the simplicity of the Cp-hap pipeline will accelerate further work on chloroplast structural heteroplasmy.

Materials and Methods

Cp-hap, a new method for quantifying chloroplast structural haplotypes using long-read sequencing data

New long-read sequencing methods such as those developed by Oxford Nanopore can routinely sequence single DNA molecules of 10s of kb in length. These molecules can be used to provide direct evidence for the existence and relative abundance of a large range of chloroplast genome haplotypes. Furthermore, long-reads from the chloroplast genome are often highly abundant in plant genome sequencing projects, because the chloroplast genome is typically present in copy numbers at least two orders of magnitude higher than the nuclear genome (27, 28). Thus, most whole genome sequencing projects of plant provide abundant information for quantifying chloroplast structural haplotypes. Here, we describe a simple approach that uses this long-read sequencing data to estimate and quantify chloroplast structural haplotypes. We call this approach the Cp-hap pipeline, and provide open source code and detailed

278 instructions for running this pipeline on GitHub at <https://github.com/asdcid/Cp-hap>.

279

280 The Cp-hap pipeline consists of three steps: (i) create a fasta file that contains all
281 haplotypes of interest; (ii) map the long-reads to the potential structures in step (i); (iii)
282 use the results of step (ii) to quantify the relative abundance of all possible haplotypes.

283 Below, we describe each of these steps in more detail.

284

285 Step (i) requires us to create a fasta file containing all of the structural haplotypes of
286 interest. To create a list of all possible haplotypes from the LSC, SSC, and IR regions,
287 we first consider that each of these regions can have four different orientations:
288 original, reversed, complement, and reverse-complement (Fig. S3A). This results in
289 256 possible structural haplotypes ($256 = 4 \times 4 \times 4 \times 4$) in which the regions retain the
290 ordering conserved across plant chloroplast genomes of LSC-IRA-SSC-IRB. These
291 256 haplotypes can be grouped into 128 identifiable structures, since it is not possible
292 to distinguish between one structure and its direct complement given the two-stranded
293 nature of the DNA molecule.

294

295 To uniquely identify one of these 128 structures, a single sequencing read would need
296 to cover at least some parts of all four regions (LSC, SSC and the two IR regions), for
297 which the read would need to be at least 30-50 kb. This is because to cover all four
298 regions, at a minimum a read must entirely cover the SSC (~20 kb) region and one IR
299 region (10-30 kb) and at least partially cover the LSC region and the other IR region.

300 But the abundance of reads of this length tends to be relatively low in most datasets,
301 meaning that in most cases it is almost impossible to uniquely identify one of the 128
302 possible structures by simply mapping a read to each structure.

303

304 To reduce the length of long-reads required to uniquely identify chloroplast structural
305 haplotypes, the Cp-hap pipeline assumes by default that the two large repeat regions
306 are always inverted. When assuming that the IR regions are always inverted, there are
307 only 32 uniquely identifiable chloroplast genome structural haplotypes (Fig. S1). In
308 this situation, a read only needs to entirely cover one IR region and partially cover the
309 two adjacent LSC and SSC regions to provide evidence to uniquely identify one of the
310 32 structures (Fig. S3B). Because of this, this approach can provide direct evidence
311 for one of the 32 possible structures using reads that are just 10-30 kb in size (i.e. at
312 least as long as the length of a single IR region). Reads of this length are often highly
313 abundant in long-read datasets, meaning that the default Cp-hap method is readily
314 applicable to many long-read datasets being produced today.

315

316 In Step (i), the default Cp-hap pipeline creates a fasta file of the 32 possible structural
317 haplotypes when assuming that the IRs are inverted by simply combining the LSC,
318 SSC and IR region sequences with different orientations. After creating each
319 structural haplotype, we then duplicate and concatenate it, such that single reads
320 which span the point at which the genome was linearised will still align successfully
321 to the relevant haplotype in the fasta file. This creates a fasta file with 32 sequences,

each of which represents one of the 32 uniquely identifiable chloroplast structural haplotypes, and each of which contains a sequence that is twice as long as the chloroplast genome of interest.

In step (ii) of the pipeline, we align our long-reads to our reference set of 32 possible structural haplotypes with Minimap2 (29), using “-secondary=no” setting to ensure that only the best alignment for each read is retained. After aligning all long-reads to the set of 32 possible structural haplotypes, we examine each mapped read, and retain only those that completely cover a whole IR region and at least 1 kb of the adjacent LSC and SSC regions. We then consider that these reads, which we term valid reads, uniquely identify one of the 32 possible structural haplotypes.

In step (iii) of the Cp-hap pipeline, we calculate the relative proportion of each of the 32 possible haplotypes. To do this, we simply divide the number of valid reads that aligned to each of the haplotypes by total number of reads that aligned to all haplotypes. These proportions provide direct estimates of the relative abundance of all 32 uniquely identifiable chloroplast genome structural haplotypes.

However, we acknowledge that the IR regions are not always the case. For example, the large repeat regions are positioned in-line instead of inverted in *Selaginella tamariscina* chloroplast genome (20). In cases such as this, Cp-hap pipeline is also able to figure out the structure. Here, we provide an example that how Cp-hap

344 pipeline confirms a chloroplast genome with atypical structure, such as a paired of
 345 in-line repeats (Fig. S3C). Since Cp-hap duplicates and concatenates all default 32
 346 structures (with IRs), reads from chloroplast genome with in-line repeats are able to
 347 map to Block A and/or C region of LSC_IR_SSC_IRrc structure (one of the 32 default
 348 structure in Cp-hap pipeline), or Block B region of LSC_IRrc_SSC_IR structure (the
 349 other default structure in Cp-hap pipeline). Those reads are impossible to map to
 350 Block B region of LSC_IR_SSC_IRrc, or Block A or C region of LSC_IRrc_SSC_IR.
 351 Therefore, if some reads only map to Block A and/or C region of LSC_IR_SSC_IRrc
 352 and Block B region of LSC_IRrc_SSC_IR, it suggests that the chloroplast genome
 353 acquires in-line repeats.

354
 355 In sum, Cp-hap pipeline can provide direct evidence to identify one of the 32
 356 chloroplast structural haplotypes with a pair of IRs, and provide indirect evidence to
 357 identify other 96 chloroplast genome structural haplotypes with a pair of non-inverted
 358 repeats.

359 360 **Assessing chloroplast structural heteroplasmy across the plant tree of** 361 **life**

362 We used the Cp-hap pipeline to assess chloroplast genome structural heteroplasmy
 363 across the plant tree of life. To do this we searched for publicly available long-read
 364 data from land plants in the NCBI SRA database (<https://www.ncbi.nlm.nih.gov>). We
 365 only selected datasets in which the average read length was >9 kb and for which the

total amount of sequence data was >5Gb. 90 of the 164 potential species with datasets met this requirement in April, 2019.

For the 90 species, 19 of them do not have available chloroplast reference genomes.

Due to the relative conservation of land plant chloroplast genomes, we used the

chloroplast genome in the same genus as the reference genome for 17 species (Dataset

S1). For the final two species, *Herrania umbratica* and *Siraitia grosvenorii*, we

assembled and annotated chloroplast genomes *de novo* using long and short-reads

(Dataset S1) following the pipeline described in (30). The chloroplast genome of

Herrania umbratica and *Siraitia grosvenorii* have been deposited in NCBI

(MN163033 and MK279915).

Full details of the reads, accession numbers, and chloroplast genomes used for each of

the 90 species in our dataset are given in Dataset S1.

Chloroplast structural haplotype quantification

For the 90 species for which we have long-read data, we ran the Cp-hap pipeline

using default settings to generate the 32 possible haplotypes from the LSC, SSC, and

IR regions. We then ran the rest of the Cp-hap pipeline with default settings for all

species, resulting in measurements for each species of the number of valid reads that

mapped to each haplotype, and the relative abundance of each haplotype. We

calculated the relative abundance of each haplotype only if the species had more than

388 five valid reads in total. This filter led to us excluding 29 datasets that did not have
389 more than five valid chloroplast reads, resulting in a final dataset of 61 species for
390 which we have sufficient long-read data to estimate relative haplotype abundance.

391

392 For the *Eucalyptus pauciflora* dataset, we quantified the chloroplast structural
393 haplotypes not only for a single individual, but also for eight separate samples taken
394 from different branches of the same individual. This was possible because this data
395 was sequenced from leaves collected from eight different branch tips from a single
396 individual plant, and each sample contained sufficient long chloroplast reads to
397 provide reliable haplotype quantification. This provides a unique opportunity to
398 understand whether the relative abundance of chloroplast structural haplotypes varies
399 or remains constant within an individual, potentially providing insights into the
400 mechanisms which lead to the existence of multiple haplotypes.

401

402 **Statistical analyses**

403 Although our method can detect a large number of structural haplotypes, across all of
404 the species we examined with inverted repeats, we observed three haplotypes,
405 haplotypes A, B, and C (Fig. 1).

406

407 Because no species contained more than two haplotypes, we used a binomial test to
408 ask whether the abundances of haplotype A and B differ from a 1:1 ratio in each
409 species, or in each individual sample in the case of *Eucalyptus pauciflora*. A

410 significant result from a binomial test suggests that the observed haplotype
411 abundances are unlikely to be explained by an underlying 1:1 ratio of chloroplast
412 structural haplotypes.

413

414 In addition, since we have data for eight different branch tips of a single *Eucalyptus*
415 *pauciflora* individual, we used two additional approaches to test for differences
416 between samples. First, we used a chi-square test to ask whether the abundances of
417 the two haplotypes differ among samples. Second, we tested for phylogenetic signal
418 among the relative abundance of haplotypes observed in the eight branch tips. To do
419 this, we used the structure of the physical tree (with branch lengths in units of
420 centimetres) and the phylosignal package (31) in R (32). This package implements
421 five commonly-used tests of phylogenetic signal: (i) Abouheif's Cmean (33); (ii & iii)
422 Blomberg's K and K* (34); (iv) Moran's I (35); (v) Pagel's Lambda (36). These tests
423 use different approaches to ask whether the physical structure of the individual tree
424 helps to explain the observed variation in the relative haplotype abundance of the
425 eight samples. For example, if the process that generates different haplotypes operates
426 relatively slowly with respect to the age of this individual, which is roughly 40 years,
427 then we might expect haplotype frequencies to vary due to genetic drift, and thus for
428 the haplotype frequencies in neighbouring parts of the tree to be more similar than
429 would be expected by chance. This similarity would be revealed by the detection of
430 significant phylogenetic signal in the relative abundance data. However, if the process
431 that generates the different haplotypes operates very quickly with respect to the age of

432 this individual, then we would expect all haplotype frequencies to be roughly equal,
433 and thus we would expect to observe no phylogenetic signal in the haplotype
434 abundance data.

435

436 **Acknowledgments**

437 We thank the help of Chen Han in the statistical analysis. We also thank the people
438 below for providing data: Xuzhi Chao (for *Selaginella tamariscina* data), Hongyan
439 Du (for *Eucommia ulmoides* data). This work was supported by grants from the
440 Australian Research Council Future Fellowship, FT140100843.

441

442 **References**

- 443 1. S. Wicke, G. M. Schneeweiss, C. W. dePamphilis, K. F. Muller, D. Quandt, The
444 evolution of the plastid chromosome in land plants: gene content, gene order, gene
445 function. *Plant Mol Biol* **76**, 273-297 (2011).
- 446 2. X. M. Zheng *et al.*, Inferring the evolutionary mechanism of the chloroplast genome
447 size by comparing whole-chloroplast genome sequences in seed plants. *Scientific*
448 *Reports* **7**, 1555 (2017).
- 449 3. J. D. Palmer, Chloroplast DNA exists in two orientations. *Nature* **301**, 92 (1983).
- 450 4. D. B. Stein, J. D. Palmer, W. F. Thompson, Structural evolution and flip-flop
451 recombination of chloroplast DNA in the fern genus *Osmunda*. *Current Genetics* **10**,
452 835-841 (1986).
- 453 5. G. Martin, F. C. Baurens, C. Cardi, J. M. Aury, A. D'Hont, The complete chloroplast
454 genome of banana (*Musa acuminata*, Zingiberales): insight into plastid
455 monocotyledon evolution. *PLoS One* **8**, e67350 (2013).
- 456 6. T. Brears, C. L. Schardl, D. M. Lonsdale, Chloroplast genome organisation in sugar

- 457 beet and maize. *Plant Mol Biol* **6**, 171-177 (1986).
- 458 7. H. J. Bohnert, W. Löffelhardt, Cyanelle DNA from *Cyanophora paradoxa* exists in two
459 forms due to intramolecular recombination. *FEBS Letters* **150**, 403-406 (1982).
- 460 8. J. Aldrich, B. Cherney, E. Merlin, C. Williams, L. Mets, Recombination within the
461 inverted repeat sequences of the *Chlamydomonas reinhardtii* chloroplast genome
462 produces two orientation isomers. *Current Genetics* **9**, 233-238 (1985).
- 463 9. R. I. Ibrahim, J. Azuma, M. Sakamoto, Complete nucleotide sequence of the cotton
464 (*Gossypium barbadense* L.) chloroplast genome with a comparative analysis of
465 sequences among 9 dicot plants. *Genes Genet Syst* **81**, 311-321 (2006).
- 466 10. M. Yang *et al.*, The complete chloroplast genome sequence of date palm (*Phoenix*
467 *dactylifera* L.). *PLoS One* **5**, e12762 (2010).
- 468 11. Y. Liu *et al.*, Complete chloroplast genome sequences of Mongolia medicine *Artemisia*
469 *frigida* and phylogenetic relationships with other plants. *PLoS One* **8**, e57533 (2013).
- 470 12. Y. Zhang, L. Li, T. L. Yan, Q. Liu, Complete chloroplast genome sequences of *Praxelis*
471 (*Eupatorium catarium* Veldkamp), an important invasive species. *Gene* **549**, 58-69
472 (2014).
- 473 13. M. Wang *et al.*, Comparative Analysis of Asteraceae Chloroplast Genomes: Structural
474 Organization, RNA Editing and Evolution. *Plant Molecular Biology Reporter* **33**,
475 1526-1538 (2015).
- 476 14. J. F. Walker, R. K. Jansen, M. J. Zanis, N. C. Emery, Sources of inversion variation in
477 the small single copy (SSC) region of chloroplast genomes. *Am J Bot* **102**, 1751-1752
478 (2015).

- 479 15. C. S. Wu, C. P. Lin, C. Y. Hsu, R. J. Wang, S. M. Chaw, Comparative chloroplast
480 genomes of pinaceae: insights into the mechanism of diversified genomic
481 organizations. *Genome Biol Evol* **3**, 309-319 (2011).
- 482 16. R. Cronn, M. Cedroni, T. Haselkorn, C. Grover, J. F. Wendel, PCR-mediated
483 recombination in amplification products derived from polyploid cotton. *Theor Appl*
484 *Genet* **104**, 482-489 (2002).
- 485 17. D. J. Lahr, L. A. Katz, Reducing the impact of PCR-mediated recombination in
486 molecular evolution and environmental studies using a new-generation high-fidelity
487 DNA polymerase. *Biotechniques* **47**, 857-866 (2009).
- 488 18. C. Belser *et al.*, Chromosome-scale assemblies of plant genomes using nanopore
489 long reads and optical maps. *Nature Plants* **4**, 879-887 (2018).
- 490 19. S. Deschamps *et al.*, A chromosome-scale assembly of the sorghum genome using
491 nanopore sequencing and optical mapping. *Nature Communications* **9**, 4844 (2018).
- 492 20. Z. Xu *et al.*, Genome Analysis of the Ancient Tracheophyte *Selaginella tamariscina*
493 Reveals Evolutionary Features Relevant to the Acquisition of Desiccation Tolerance.
494 *Mol Plant* **11**, 983-994 (2018).
- 495 21. C. S. Wu, Y. N. Wang, C. Y. Hsu, C. P. Lin, S. M. Chaw, Loss of different inverted
496 repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and
497 influence of heterotachy on the evaluation of gymnosperm phylogeny. *Genome Biol*
498 *Evol* **3**, 1284-1295 (2011).
- 499 22. C. S. Wu, S. M. Chaw, Large-Scale Comparative Analysis Reveals the Mechanisms
500 Driving Plastomic Compaction, Reduction, and Inversions in Conifers II

501 (Cupressophytes). *Genome Biol Evol* **8**, 3740-3750 (2016).

502 23. L. A. Raubeson, R. K. Jansen, A rare chloroplast-DNA structural mutation is shared by
503 all conifers. *Biochemical Systematics and Ecology* **20**, 17-24 (1992).

504 24. X. Yi, L. Gao, B. Wang, Y. J. Su, T. Wang, The complete chloroplast genome
505 sequence of *Cephalotaxus oliveri* (Cephalotaxaceae): evolutionary comparison of
506 cephalotaxus chloroplast DNAs and insights into the loss of inverted repeat copies in
507 gymnosperms. *Genome Biol Evol* **5**, 688-698 (2013).

508 25. W. Guo *et al.*, Predominant and substoichiometric isomers of the plastid genome
509 coexist within *Juniperus* plants and have shifted multiple times during cupressophyte
510 evolution. *Genome Biol Evol* **6**, 580-590 (2014).

511 26. X.-J. Qu, C.-S. Wu, S.-M. Chaw, T.-S. Yi, Insights into the Existence of Isomeric
512 Plastomes in Cupressoideae (Cupressaceae). *Genome Biology and Evolution* **9**,
513 1110-1119 (2017).

514 27. S. A. Morley, B. L. Nielsen, Chloroplast DNA Copy Number Changes during Plant
515 Development in Organelle DNA Polymerase Mutants. *Front Plant Sci* **7**, 57 (2016).

516 28. H. Daniell, C. S. Lin, M. Yu, W. J. Chang, Chloroplast genomes: diversity, evolution,
517 and applications in genetic engineering. *Genome Biol* **17**, 134 (2016).

518 29. H. Li, Minimap and miniasm: fast mapping and de novo assembly for noisy long
519 sequences. *Bioinformatics* **32**, 2103-2110 (2016).

520 30. W. Wang *et al.*, Assembly of chloroplast genomes with long- and short-read data: a
521 comparison of approaches using *Eucalyptus pauciflora* as a test case. *BMC*
522 *Genomics* **19**, 977 (2018).

- 523 31. F. Keck, F. Rimet, A. Bouchez, A. Franc, phylsignal: an R package to measure, test,
524 and explore the phylogenetic signal. *Ecol Evol* **6**, 2774-2780 (2016).
- 525 32. R Core Team, R: A language and environment for statistical computing. *R Foundation*
526 *for Statistical Computing* (2013).
- 527 33. E. Abouheif, A method for testing the assumption of phylogenetic independence in
528 comparative data. *Evolutionary Ecology Research* **1**, 895-909 (1999).
- 529 34. S. P. Blomberg, T. Garland, Jr., A. R. Ives, Testing for phylogenetic signal in
530 comparative data: behavioral traits are more labile. *Evolution* **57**, 717-745 (2003).
- 531 35. P. A. Moran, Notes on continuous stochastic phenomena. *Biometrika* **37**, 17-23
532 (1950).
- 533 36. M. Pagel, Inferring the historical patterns of biological evolution. *Nature* **401**, 877-884
534 (1999).

535

536

537 **Figure Legends**

538

539

540 Fig. 1. The three different structural haplotypes of chloroplast genomes detected in
541 this study. The green region is the Long Single Copy (LSC) region and blue region is
542 the Short Single Copy (SSC) region. The two black regions are the Inverted Repeat
543 (IR) regions. The arrow denotes 5'-3' orientation. *psbA* is in the minus strand of LSC
544 region, whereas *rrn23* is in the plus strand of IR regions. *ndhF* is in the minus strand
545 of SSC region, while *ccsA* is in the plus strand of SSC region. For ease of
546 communication, we use the relative order of three genes (*psbA* in LSC, and *ndhF* and
547 *ccsA* in the SSC) to label these two haplotypes 'A' and 'B'. In haplotype A, these
548 genes are ordered *psbA-ndhF-ccsA*. In haplotype B these genes are ordered
549 *psbA-ccsA-ndhF*. For haplotype C, these three genes are ordered the same as
550 haplotype A, but the repeat regions are in-line rather than inverted.

Fig. 2. The relationship between haplotype counts and haplotype frequencies. Each point is one species. A blue dot means binomial p-value > 0.05 , whereas the orange dot means binomial p-value ≤ 0.05 . Two species, *Pinus taeda*, *Picea sitchensis* only showed evidence of haplotype B, therefore the frequencies of them were 0. For the remaining 58 species, they show higher level of equal frequency between haplotype A and B with the increase of sample size. One species (*Selaginella tamariscina*) is omitted from this figure because it contained only a third haplotype (haplotype C, Figure 1).

Table

Table 1. The frequency of existing chloroplast genome haplotypes from 61 species.

Division	Order	Species	Size (bp)	LSC (bp)	SSC (bp)	IR (bp)	HA	HB	Frequency	p-value
Angiospermae	Poales	<i>Andropogon gerardii</i>	139,712	81,548	12,592	22,786	31	31	0.50	1.00
		<i>Brachypodium</i> † <i>mexicanum</i>	~135,199	~79,447	~12,668	~21,542	403	373	0.52	0.30
		<i>Eleusine coracana</i> †	~135,151	~80,667	~12,646	~20,919	14	11	0.56	0.69
		<i>Hordeum vulgare</i>	136,462	81,671	12,701	21,045	347	329	0.51	0.51
		<i>Oropetium thomaeum</i> †	~133,880	~79,181	~12,591	~21,054	27	28	0.49	1.00
		<i>Oryza coarctata</i>	134,750	80,816	12,334	20,800	40	38	0.51	0.91
		<i>Oryza longistaminata</i>	134,567	80,596	12,357	20,807	8	4	0.67	0.39
		<i>Oryza meyeriana</i>	136,133	81,832	12,499	20,901	5	5	0.50	1.00
		<i>Panicum miliaceum</i>	140,048	81,841	9,603	24,302	37	22	0.63	0.07
		<i>Saccharum officinarum</i>	141,176	83,046	12,540	22,795	51	56	0.48	0.70
		<i>Saccharum spontaneum</i>	141,168	83,046	12,544	22,789	25	28	0.47	0.78
		<i>Sorghum bicolor</i>	140,754	83,733	12,503	22,259	94	71	0.57	0.09
		<i>Zea mays</i>	140,384	82,352	12,536	22,748	564	555	0.50	0.30
		<i>Zoysia japonica</i>	135,854	81,348	12,582	20,962	35	54	0.39	0.06
	Zingiberales	<i>Musa schizocarpa</i> †	~169,503	~87,828	~11,487	~35,094	192	210	0.48	0.40
	Asparagales	<i>Yucca aloifolia</i> †	~157,785	~86,135	~18,292	~26,679	87	105	0.45	0.22
		<i>Yucca filamentosa</i>	157,785	86,135	18,292	26,679	200	202	0.50	0.96
	Ranunculales	<i>Papaver somniferum</i>	152,931	83,029	17,920	25,991	12	13	0.48	1.00
	Ericales	<i>Actinidia eriantha</i>	156,964	88,759	20,541	23,832	122	94	0.56	0.11
	Lamiales	<i>Osmanthus fragrans</i>	155,326	86,564	17,370	25,696	65	63	0.51	0.93

	<i>Salvia splendens</i> †	~153,995	~84,573	~17,590	~25,916	15	11	0.58	0.56
Solanales	<i>Cuscuta campestris</i>	86,749	50,978	7,063	14,354	5	8	0.38	0.58
	<i>Nicotiana attenuata</i>	155,886	86,602	18,518	25,383	3	6	0.33	0.51
Asterales	<i>Lactuca sativa</i>	152,765	84,103	18,596	25,033	39	31	0.56	0.40
Cucurbitales	<i>Siraitia grosvenorii</i>	158,757	87,625	18,556	26,288	5	4	0.56	1.00
Fabales	<i>Ammopiptanthus nanus</i>	154,084	84,070	18,014	26,000	31	43	0.42	0.24
	<i>Arachis hypogaea</i>	156,391	85,946	18,797	25,824	17	17	0.50	1.00
Malpighiales	<i>Vigna unguiculata</i>	152,415	81,822	17,425	26,584	4	3	0.43	1.00
	<i>Populus trichocarpa</i>	157,033	85,129	16,600	27,652	58	78	0.39	0.40
	<i>Salix purpurea</i>	155,590	84,454	16,220	27,458	41	47	0.47	0.59
Rosales	<i>Fragaria vesca</i>	155,691	85,606	18,173	25,956	31	25	0.55	0.51
	<i>Fragaria ananassa</i>	155,549	85,532	18,145	25,936	4	2	0.67	0.69
Fagales	<i>Prunus avium</i> †	~157,859	~85,977	~19,120	~26,381	114	128	0.47	0.40
	<i>Rosa chinensis</i>	156,546	85,727	18,754	26,030	5	4	0.56	1.00
	<i>Rubus occidentalis</i> †	~155,760	~85,430	~18,768	~25,781	9	7	0.56	0.80
	<i>Betula pendula</i>	161,147	89,428	19,607	26,056	22	16	0.58	0.42
Brassicales	<i>Arabidopsis thaliana</i>	154,478	84,170	17,780	26,264	135	150	0.47	0.41
	<i>Brassica napus</i>	152,860	83,030	17,760	26,035	51	38	0.57	0.20
Malvales	<i>Brassica oleracea</i>	153,364	83,136	17,834	26,197	1,714	1,773	0.49	0.33
	<i>Brassica rapa</i>	153,483	83,282	17,775	26,213	1,097	1,077	0.51	0.68
	<i>Bombax ceiba</i>	158,997	89,022	21,111	24,432	71	96	0.43	0.06
	<i>Gossypium arboreum</i>	160,230	88,722	20,274	25,617	38	47	0.45	0.39
	<i>Gossypium barbadense</i>	160,317	88,841	20,294	25,591	11	14	0.44	0.69
	<i>Gossypium bickii</i>	159,422	88,073	20,183	25,583	333	330	0.50	0.94
	<i>Gossypium darwinii</i>	160,378	88,906	20,266	25,603	93	95	0.49	0.94
	<i>Gossypium hirsutum</i>	160,301	88,817	20,280	25,602	623	586	0.52	0.30

		<i>Gossypium longicalyx</i>	160,241	88,667	20,278	25,648	205	213	0.49	0.73
		<i>Gossypium mustelinum</i>	160,313	88,824	20,271	25,609	50	52	0.49	0.92
		<i>Gossypium raimondii</i>	160,161	88,654	20,205	25,651	48	38	0.56	0.33
		<i>Gossypium tomentosum</i>	160,433	88,932	20,273	25,614	42	61	0.41	0.08
		<i>Theobroma cacao</i>	160,619	89,333	20,194	25,546	86	77	0.52	0.53
	Myrtales	<i>Eucalyptus albens</i> †	~160,076	~88,828	~18,476	~26,386	20	25	0.44	0.55
		<i>Eucalyptus marginata</i>	160,076	88,828	18,476	26,386	13	16	0.45	0.71
		<i>Eucalyptus melliodora</i>	160,386	89,073	18,557	26,378	371	373	0.50	0.97
		<i>Eucalyptus pauciflora</i>	159,942	88,787	18,421	26,367	1,431	1,459	0.50	0.62
		Branch tip A					101	110	0.48	0.63
		Branch tip B					186	184	0.50	0.96
		Branch tip C					203	191	0.51	0.58
		Branch tip D					132	114	0.54	0.28
		Branch tip E					186	189	0.50	0.92
		Branch tip F					234	255	0.48	0.37
		Branch tip G					141	150	0.48	0.64
		Branch tip H					248	266	0.48	0.45
	Sapindales	<i>Acer yangbiense</i> †	~156,262	~86,018	~18,072	~26,086	104	94	0.53	0.52
		<i>Citrus maxima</i>	160,133	87,739	18,395	26,999	2	6	0.25	0.29
		<i>Xanthoceras sorbifolium</i>	161,231	85,299	18,692	28,620	110	119	0.48	0.60
Gymnosperm	Pinales	<i>Pinus taeda</i>	121,530	66,444	54,288	399	0	898	0.00	2.2e-16*
		<i>Picea sitchensis</i>	124,049	67,411	55,758	440	0	2,845	0.00	2.2e-16*
Pteridophytes	Selaginellales	<i>Selaginella tamariscina</i> ‡	126,399	53,176	47,573	12,825	0	0	-	-

*: p-value ≤ 0.05

†: The long-reads were mapped to the chloroplast genome from the same genus species due to the lack of chloroplast genome of this species, so the lengths of LSC, SSC and IR are around (~).

568 ‡: *Selaginella tamariscina* only present one haplotype, haplotype C, with 174 counts, and p-value is 2.2e-16.
569 **Size:** the chloroplast genome size
570 **HA/B:** Haplotype A/B, the number of supported reads of LSC and SSC having identical/opposite orientation.
571 **Frequency:** the proportion of the count of haplotype A in the count of haplotype A+B

572