Sequencing, de novo assembly of *Ludwigia* plastomes, and comparative analysis within the Onagraceae family

Barloy-Hubler F.[3], Le Gac A.-L.[1], Boury C.[2], Guichoux E.[2], and Barloy D.[1*]

1-DECOD (Ecosystem Dynamics and Sustainability), Institut Agro, IFREMER, INRAE, Rennes, France.

2- Université de Bordeaux, INRAE, BIOGECO, Cestas, France

3- Université de Rennes 1, CNRS, UMR 6553 ECOBIO, Rennes, France.

* Corresponding author: dominique.barloy@agrocampus-ouest.fr

Abstract

The Onagraceae family, which belongs to the order Myrtales, consists of approximately 657 species and 17 genera, including the genus *Ludwigia* L., which is comprised of 82 species. There are few genomic resources for Onagraceae, which limits phylogenetic and population genetics, as well as genomic studies. In this study, new complete plastid genomes of *Ludwigia grandiflora subps. hexapetala* (*Lgh*) and *Ludwigia peploides subsp montevidensis* (*Lpm*) were generated using a combination of different sequencing technologies. These plastomes were then compared to the published *Ludwigia octovalvis* (*Lo*) plastid genome, which was re-annotated by the authors. We initially sequenced and assembled the chloroplast (cp) genomes of *Lpm* and *Lgh* using a hybrid strategy. We observed the existence of two *Lgh* haplotypes and two potential *Lpm* haplotypes. *Lgh*, *Lpm,* and *Lo* plastomes were similar in terms of genome size, gene number, structure, and inverted repeat (IR) boundaries, comparable to other species in the Myrtales order. A total of 45 to 65 SSRs (simple sequence repeats), were detected, depending on the species, with the majority consisting solely of A and T, which is common among angiosperms. Four chloroplast genes (*matK*, *accD*, *ycf2* and *ccsA)* were found under positive selection pressure, which is commonly associated with plant development, and especially in aquatic plants such as *Lgh,* and *Lpm.* Our hybrid sequencing approach revealed the presence of two *Lgh* plastome haplotypes which will help to advance phylogenetic and evolutionary studies, not only specifically for Ludwigia, but also the Onagraceae family and Myrtales order. To enhance the robustness of our findings, a larger dataset of chloroplast genomes would be beneficial.

Keywords

Water primrose, Ludwigia, Onagraceae, chloroplast genome, long and short reads, hybrid assembly, haplotype

Introduction

The Onagraceae family belongs to the order Myrtales which includes approximately 657 species of herbs, shrubs, and trees across 17 genera grouped into two subfamilies: subfam. Ludwigioideae W. L. Wagner and Hoch, which only has one genus (*Ludwigia* L.), and subfam. Onagroideae which contains six tribes and 21 genera [1]. *Ludwigia* L. is composed of 83 species [2, 3]. The current classification for Ludwigia L., which are composed of several hybrid and/or polyploid species, lists 23 sections. A recent molecular analysis is clarified and supported several major relationships in the genus but has challenged the complex sectional classification of Ludwigia L. [4].

The diploid species *Ludwigia peploides* (Kunth) Raven subsp. *montevidensis* (Spreng.) Raven (1963) (named here *Lpm*) (2n=16), and the decaploid species, *Ludwigia grandiflora* (Michx.) Greuter & Burdet (1987) subsp. *hexapetala* (Hook. & Arn.) (Nesom & Kartesz 2000) (named here *Lgh*) (2n=80), reproduce essentially by clonal propagation, which suggests that there is a low genetic diversity within the species (Dandelot et al, 2005). *Lgh* and *Lpm* are native to South America and are described as aquatic invasive plants in Europe [5]. In France, both species occupied aquatic habitats, such as static or slow-flowing waters, riversides, and have recently been observed in wet meadows [6]. The transition from an aquatic to a terrestrial habitat has led to the emergence of two *Lgh* morphotypes [7]. The appearance of metabolic and morphological adaptations could explain the ability to acclimatize to terrestrial conditions, and this phenotypic plasticity involves various genomic and epigenetic modifications [8].

Adequate genomic resources are necessary in order to be able to analyze the different (epi)genomic mechanisms of this acclimation. However, even though the number of terrestrial plant genomes has increased considerably over the last 20 years, only a small fraction (~ 0.16%) have been sequenced, with some clades being significantly more represented than others [9]. Thus, for the Onagraceae family (which includes *Ludwigia* sp.), only a handful of chloroplast

61 sequences (plastomes) are available, and the complete genome has not yet been sequenced. As
62 of April 2023, there are 10,712 reference plastomes (RefSeq) listed on GenBank, with the vast
63 majority (10,392 genomes) belonging to Viridiplantae (green plants). However, the number of
64 plastomes available for the Onagraceae family is limited, with only 36 plastomes currently
65 listed. Among these, 15 plastomes are from the tribe Epilobieae, with 11 in the Epilobium genus
66 and 4 in the Chamaenerion genus. Additionally, there are 23 plastomes from the tribe Onagreae,
67 with 17 in the Oenothera genus, 5 in the Circaea genus, and only one in the Ludwigia genus.
68 The *Ludwigia octovalvis* chloroplast genome was released in 2016 as a unique haplotype of
69 approximately 159 kb (Liu et al., 2016). *L. octovalvis* belongs to sect. *Macrocarpon* (Micheli)
70 H.Hara of the genus Ludwigia while *Lpm* and *Lgh* belong to jussieae section [10],[11].
71 Generally, the inheritance of chloroplast genomes is considered to be maternal in angiosperms.
72 However, biparentally inherited chloroplast genomes could potentially exist in approximately
73 20% of angiosperm species [12, 13] . Both maternal and biparental inheritance are described in
74 the Onagraceae family. In tribe Onagreae, *Oenothera* subsect. *Oenothera* are known to have
75 biparental plastid inheritance [1, 14]. In tribe Epilobieae, biparental plastid inheritance was also
76 reported in *Epilobium* L. with mainly maternal transmission, and very low proportions of
77 paternally transmitted chloroplasts [15].

78 The chloroplast is the symbolic organelle of plants and plays a fundamental role in
79 photosynthesis. Chloroplasts evolved from cyanobacteria through endosymbiosis and thereby
80 inherited components of photosynthesis reactions (photosystems, electron transfer and ATP
81 synthase) and gene expression systems (transcription and translation) [16]. In general,
82 chloroplast genomes (plastomes) are highly conserved in size, structure, and genetic content.
83 They are rather small (120-170 kb, [17]), with a quadripartite structure comprising two long
84 identical inverted repeats (IR, 10–30 kb) separated by large and a small single copy regions
85 (LSC and SSC, respectively). They are also rich in genes, with around 100 unique genes

86 encoding key proteins involved in photosynthesis, and a comprehensive set of ribosomal RNAs
87 (rRNAs) and transfer RNAs (tRNAs) [18]. Plastomes are generally circular but linear shapes
88 also exist [19]. Chloroplast DNA usually represents 5-20% of total DNA extracted from young
89 leaves and therefore low-coverage whole genome sequencing can generate enough data to
90 assemble an entire chloroplast genome [20].

91 If we refer to their GenBank records, more than 95% of these plastomes were sequenced by
92 so-called short read techniques (mostly Illumina). However, in most seed plants, the plastid
93 genome exhibits two large inverted repeat regions (60 to 335 kb, [20]), which are longer than
94 the short read lengths (< 300 bp). This leads to incomplete or approximate assemblies [21].
95 Recent long-read sequencing (> 1000 bp) provides strong evidence that terrestrial plant
96 plastomes have two structural haplotypes, present in equal proportions and differing in IR
97 orientation of the [22]. This shows the importance of using the so-called third generation
98 sequence (TGS, PacBio or Nanopore) to correctly assemble the IRs of chloroplasts and to
99 identify any different structural haplotypes. The current problem with PacBio or Nanopore long
100 read sequencing is the higher error rate compared to short read technology [23-25]. Thus, a
101 hybrid strategy which combines long reads (to access the genomic structure) and short reads (to
102 correct sequencing errors) could be effective [21, 26].

103 Here, we report the newly sequenced complete plastid genomes of *Ludwigia grandiflora*
104 subps. *hexapetala* (*Lgh*) and *Ludwigia peploides* subsp *montevidensis* (*Lpm*), using a
105 combination of different sequencing technologies, as well as a re-annotated comparative
106 genomic analysis of the published *Ludwigia octovalvis* (*Lo*) plastid. The main objectives of this
107 study are (1) to assemble and annotate the plastomes of two new species of Ludwigia sp., (2)
108 to reveal the divergent sequence hotspots of the plastomes in this genus and in the Onagraceae
109 (3) to identify the genes under positive selection.

110 To achieve this, we utilized long read sequencing data from Oxford Nanopore and short read
111 sequencing data from Illumina to assemble the *Lgh* plastomes and compared these assemblies
112 with those obtained solely from long reads of *Lpm*. We also compared both plastomes to the
113 published plastome of *Lo*. Our findings demonstrated the value of *de novo* assembly in reducing
114 assembly errors and enabling accurate reconstruction of full heteroplasmy. We also evaluated
115 the performance of a variety of software for sequence assembly and correction in order to define
116 a workflow that will be used in the future to assemble *Ludwigia* sp. mitochlondrial and nuclear
117 genomes. Finally, the three new Ludwigia plastomes generated by our study make it possible
118 to extend the phylogenetic study of the Onagraceae family and to compare it with previously
119 published analyses [4, 27, 28].
120

121 **Material and Methods**

122 **Plant sampling and experimental design**

123 The original plant materials were collected in June of 2018 near to Nantes (France) and
124 formal identified by D. Barloy. *L. grandiflora* subsp. *hexapetala (Lgh)* plants were taken from
125 the Mazerolles swamps (N47 23.260, W1 28.206), and *L. peploides* subsp. *montevidensis* (*Lpm*)
126 plants from La Musse (N 47.240926, W -1.788688)*)*. Plants were cultivated in a growth
127 chamber in a mixture of $^{1}/_{3}$ soil, $^{1}/_{3}$ sand, $^{1}/_{3}$ loam with flush water level, at 22°C and a 16 h/8
128 h (light/dark) cycle. A single stem of 10 cm for each species was used for vegetative
129 propagation in order to avoid potential genetic diversity. *De novo* shoots, taken three
130 centimeters from the apex, were sampled for each species. Samples for gDNA extraction were
131 pooled and immediately snap-frozen in liquid nitrogen, then lyophilized over 48 h using a
132 Cosmos 20K freeze-dryer (Cryotec, Saint-Gély-du-Fesc, France) and stored at room
133 temperature. All the plants were destroyed after being used as required by French authorities
134 for invasive plants (article 3, prefectorial decree n°2018/SEE/2423).

135 Due to high polysaccharide content and polyphenols in *Lpm* and *Lgh* tissues, genomic DNA
136 extraction was carried out using a modified version of the protocol proposed by Panova et al in
137 2016, with three purification steps [29].
138 40 mg of lyophilized buds were ground at 30 Hz for 60 s (Retsch MM200 mixer mill,
139 FISHER). The ground tissues were lysed with 1 ml CF lysis buffer (MACHEREY-NAGEL)
140 supplemented with 20 µl RNase and incubated for 1 h at 65°C under agitation. 20 µl proteinase
141 K was then added before another incubation for 1 h at 65°C under agitation. To avoid breaking
142 the DNA during pipetting, the extracted DNA was recovered using a Phase-lock gel tube as
143 described in Belser [30]. The extracts were transferred to 2 ml tubes containing phase-lock gel,
144 and an equal volume of PCIA (Phenol, Chloroform, Isoamyl Alcohol; 25:24:1) was added.
145 After shaking for 5 min, tubes were centrifuged at 11000 g for 20 min. The aqueous phase was
146 transferred into a new tube containing phase-lock gel and extraction with PCIA was repeated.
147 DNA was then precipitated after addition of an equal volume of binding buffer C4
148 (MACHEREY-NAGEL) and 99% ethanol overnight at 4°C or 1 h in ice then centrifuged at 800
149 rpm for 10 min. After removal of the supernatant, 1 ml of CQW buffer was added then the
150 pellet of DNA was re-suspended. Next, DNA purification was carried out by adding a 2 ml
151 mixture of wash buffer PW2 (MACHEREY-NAGEL), wash buffer B5 (MACHEREY-
152 NAGEL), and ethanol at 99% in equal volumes, followed by centrifugation at 800 rpm for 10
153 min. This DNA purification step was carried out twice. Finally, the DNA pellet was dried in
154 the oven at 70°C for 30 min then re-suspended in 100 µl elution buffer BE (MACHEREY-
155 NAGEL) (5 mM Tris solution, pH 8.5) after 10 min incubation at 65°C under agitation.
156 A second purification step was performed using a PCR product extraction from gel agarose
157 kit from Macherey-Nagel (MN) NucleoSpin® Gel and PCR Clean-up kit and restarting the
158 above protocol from the step with the addition of CQW buffer then PW2 buffer.

159 The third purification step consisted of DNA purification using a Macherey-Nagel (MN)
160 NucleoMag kit for clean-up and size selection. Finally, the DNA was resuspended after a 5 min
161 incubation at 65°C in 5 mM TRIS at pH 8.5.

162 The quantity and quality of the gDNA was verified using a NanoDrop spectrometer,
163 electrophoresis on agarose gel and ethidium bromide staining under UV light and Fragment
164 Analyzer (Agilent Technologies) of the University of Rennes1.

165 **Library preparation and sequencing**

166 MiSeq (Illumina) and GridION (Oxford Nanopore Technologies, referred to here as
167 ONT) sequencing were performed at the PGTB (doi:10.15454/1.5572396583599417E12). *Lgh*
168 and *Lpm* genomic DNA were re-purified using homemade SPRI beads (1.8X ratio). For
169 Illumina sequencing, 200 ng of *Lgh* DNA was used according to the QIAseq FX DNA Library
170 Kit protocol (Qiagen). The final library was checked on TapeStation D5000 screentape (Agilent
171 Technologies) and quantified using a QIAseq Library Quant Assay Kit (Qiagen). The pool was
172 sequenced on an Illumina MiSeq using V3 chemistry and 600 cycles (2x300bp). For ONT
173 sequencing, around 8 µg of *Lgh* and *Lpm* DNA were size selected using a Circulomics SRE kit
174 (according to the manufacturer's instructions) before library preparation using a SQK-LSK109
175 ligation sequencing kit following ONT recommendations. Basecalling in High Accuracy -
176 Guppy version: 4.0.11 (MinKNOW GridION release 20.06.9) was performed for the 48 h of
177 sequencing. Long reads (LR) and short reads (SR) were available for *Lgh* and only LR for *Lpm*.

178 **Chloroplast assemblies**

179 Quality controls and preprocessing of sequences were performed using Guppy v4.0.14 base
180 call software for ONT and fastp v0.20.0 [31] using the default quality filter (phred quality >=
181 Q15) for Illumina reads. After filtering and trimming, *Lgh* Illumina-reads (SR, 2*23,067,490
182 reads) were used for reference-guided assembly using GetOrganelle v1.7.0 [32] and
183 NOVOPlasty v4.2.1 [33] using the parameters recommended for green plants. As we sequenced

184 total DNA, chloroplast reads are mixed with nuclear and mitochondrial reads. Chloroplastic
185 reads were then extracted by mapping against draft *Lgh* plastomes generated by GetOrganelle
186 in order to reduce computational costs. Chloroplastic R1 and R2 reads were used with and
187 without prior error correction using ONT reads with BayesHammer [34] for *de novo* assemblies
188 using ABySS (version 2.1.5 [35, 36]), MEGAHIT (1.1.2, [37]), Velvet (version 1.2.10) using
189 VelvetOptimiser and SPAdes (version 3.15.4,[38]). The best k-mer parameters were tested
190 using kmergenie [39] and Velvetoptimizer. The best results were obtained with k=99, and was
191 therefore the parameter used for all MiSeq assemblers. For ONT reads, *Lgh* (550,516 reads)
192 and *Lpm* (68,907 reads) LR reads were self-corrected using CANU 1.8 [40] and *de novo*
193 assembly using CANU and FLYE 2.8.2 run with the option --meta and –plasmids [41]. Hybrid
194 correction and assemblies of LR reads using SR data were realized for *Lgh* using Ratatosk [42].

195 **Post plastome assembly validation**

196 As we used many assemblers and different strategies, we produced multiple contigs that
197 needed to be analyzed and filtered in order to retain only the most robust plastomes. For that,
198 all assemblies were evaluated using the QUality ASsessment Tool (QUAST) for quality
199 assessment (http://cab.cc.spbu.ru/quast/) and visualized using BANDAGE [43]. BANDAGE
200 compatible graphs (.gfa format) were created with the megahit_toolkit for MEGAHIT and with
201 gfatools for ABySS. Overlaps between fragments were manually checked and ambiguous
202 "IUPAC or N" nucleotides were also biocured with Illumina reads when available.

203 **Chloroplast genome annotation**

204 Plastomes were annotated via the GeSeq [44] using ARAGORN and tRNAscan_SE to
205 predict tRNAs and rRNAs and tRNAscan_SE to predict tRNAs and rRNAs and via Chloe
206 prediction site (https://chloe.plantenergy.edu.au). The previously reported *Lo* chloroplast
207 genome was also similarly re-annotated to facilitate genomic comparisons. Gene boundaries,
208 alternative splice isoforms, pseudogenes and gene names and functions were manually checked

209 and biocurated using Geneious (v.10). Finally, plastomes were represented using
210 OrganellarGenomeDRAW (OGDRAW) [45]. These genomes were submitted to GenBank at
211 the National Center of Biotechnology Information (NCBI) with specific accession numbers (for
212 *Lgh* haplotype 1, (LGH1) OR166254 and *Lgh* haplotype 2, (LGH2) OR166255; for *Lpm*
213 haplotype, (LPM) OR166256) using annotation tables generated through GB2sequin [46].

214 **SSRs and Repeat Sequences Analysis**

215 Simple Sequence Repeats (SSRs) were analyzed through the MISA web server [47], with
216 parameters set to 10, 5, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides,
217 respectively. Direct, reverse and palindromic repeats were identified using RepEx. Parameters
218 used were: for inverted repeats (min size 15 nt, spacer = local, class = exact); for palindromes
219 (min size 20 nt); for direct repeats (minimum size 30 nt, minimum repeat similarity 97%).
220 Tandem repeats were identified using Tandem Repeats Finder [48], with parameters set to two
221 for the alignment parameter match and seven for mismatches and indels. The IRa region was
222 removed for all these analyses to avoid over representation of the repeats.

223 **Comparative chloroplast genomic analyses**

224 *Lgh* and *Lpm* plastomes were compared with the reannotated and biocurated *Lo* plastome
225 using mVISTA program [49], with the LAGAN alignment algorithm [50] and a cut-off of 70%
226 identity.

227 Nucleotide diversity (Pi) was analyzed using the software DnaSP v.6.12.01 [51, 52] with
228 step size set to 200 bp and window length to 300 bp. IRscope [53] was used for the analyses of
229 inverted repeat (IR) region contraction and expansion at the junctions of chloroplast genomes.
230 To assess the impact of environmental pressures on the evolution of these three Ludwigia
231 species, we calculated the nonsynonymous (Ka) and synonymous (Ks) substitutions and their
232 ratios (ω = Ks/Ks) using TBtools [54] to measure the selective pressure. Genes with $\omega < 1$, ω =

233 1, and 1 < ω were considered to be under purifying selection (negative selection), neutral
234 selection, and positive selection, respectively.

235 **Plastome phylogenomic analyses**

236 We reconstructed phylogenetic relationships among plastomes of Onagraceae. The FFT-
237 NS-2 method in MAFFT 7 [55] was used to align all plastomes with one of the IRs removed to
238 avoid data duplication. Phylogenetic tree analysis, based on the maximum likelihood (ML)
239 method, was conducted in RAxML 8.2.9 [56] using the GTR+G model with node support
240 assessed by fast-bootstrap (-f a) using 1,000 non-parametric bootstrap pseudo-replicates.

241 **Graphic representation**

242 Statistical analyses were performed using R software in RStudio integrated development
243 environment (R Core Team, 2015, RStudio: Integrated Development for R. RStudio, Inc.,
244 Boston, MA, http://www.rstudio.com/). Figures were realized using ggplot2, ggpubr, tidyverse,
245 dplyr, gridExtra, reshape2, and viridis packages. SNPs were represented using trackViewer [53]
246 and genes represented using gggenes packages.

247

248 **Results**

249 **Plastome short read assembly**

250 We first used GetOrganelle and Novoplasty with the short reads of *Lgh.* GetOrganelle
251 reconstructed two identical haplotypes that differ only by a "flip-flop" of the SSC region, as
252 shown in Figure 1, while NOVOplasty generated three contigs that can be reassembled to form
253 either one of the two haplotypes, as the SSC region can be oriented in both directions.

254 In order to evaluate the best strategies for assembly of *Lgh* plastomes, we used these
255 haplotypes to map and extract short (Illumina) and long (Nanopore) chloroplast reads.
256 Chloroplastic short reads, with or without prior correction using long reads, were assembled
257 using ABySS, Velvet, MEGAHIT and SPAdes. As shown in Figure 2A, the number of contigs

258 obtained is highly variable depending on the tools and strategies. Without correction, Velvet
259 and MEGAHIT generate highly fragmented assemblies with 1103 and 722 contigs respectively,
260 while ABySS and SPAdes produce a more reasonable number of contigs (65 and 12
261 respectively). In addition, all contigs produced by all assemblers except SPAdes, are
262 predominantly less than 500 nt long (Figure 2B from 68 to 95% of the contigs). By correcting
263 the reads before assembly, we saw a noticeable improvement in the number of contigs, which
264 decreased for all assemblers while increasing their sizes. However, except for SPAdes, the
265 number of contigs below 500 nt remained preponderant (from 54 to 70%, Figure 2C).

266 An assessment of the assembly quality using QUAST, and based on an alignment on *Lgh*
267 haplotype 1 generated by GetOrganelle as a reference, shows that none of the assemblers
268 produces a complete plastome. By comparing the contribution of small contigs (less than 200,
269 500 and 1000 nt), we noticed an invariant coverage rate for SPAdes (85%), while the fraction
270 of covered genome decreases drastically for Velvet (from 92% to 22%), correlated with a
271 duplication rate going from 2 to 1 (Add. Figure 1A). We noted that compared to an assembly
272 with all reads, the use of reads superior to 1 kb slightly decreases the coverage rate to reach an
273 optimal duplication ratio equal to 1. The effect of prior read correction is notable for MEGAHIT
274 and Velvet, especially concerning the increase in the size of the large alignment (Add. Figure
275 1A), loss of misassemblies, and reduction of the number of mismatches (Add. Figure 1B).

276 Investigating all possible scaffolds generated by BANDAGE visualization (Add. Figure 2)
277 shows that ABySS and SPAdes assembly graphs suggest a tripartite structure with the long
278 single-copy (LSC) region as the larger circle in the graph (blue), joined to the small single-copy
279 region (green) by one copy of the inverted repeats (IRs, red). In all graphs, the two IRs are a
280 collapsed repeat of approximately twice the coverage. For Velvet and MEGAHIT, graphs
281 confirm the significant fragmentation of the assemblies, which is improved by prior correction
282 of the reads.

283 In conclusion, SPAdes is the most efficient tool, and gives the best result using corrected short
284 reads to assemble a resolved plastome, with three contigs of 90,272 bp (corresponding to LSC),
285 19,788 bp (corresponding to SSC) and 24,762 bp (corresponding to a copy of the IR).
286
287 **Plastome long read assembly**
288 Chloroplast fractions of *Lgh* long reads were selected through mapping. Raw, self-
289 corrected reads were assembled using CANU, and short-read-corrected comparative assembly
290 was carried out using CANU or FLYE. Using raw data, CANU generates a unique contig
291 corresponding to haplotype 2, whereas FLYE makes two contigs that reconstruct haplotype 1.
292 Assemblies using long reads corrected by CANU leads to fragmentation into two (CANU
293 assembler) or three (FLYE assembler) contigs which both reconstruct haplotype 1, with an large
294 gap corresponding to one of the IR copies for the CANU assembler. Finally, correction by
295 RATATOSK allows CANU to assemble two redundant contigs, with contig 1 reproducing the
296 entire haplotype 2, while FLYE makes two contigs corresponding to haplotype 1 (Add. Figure
297 3A). In conclusion, the two haplotypes proposed by GetOrganelle using the short reads were
298 likewise reconstructed randomly using the long reads, with or without correction.
299 Consequently, both haplotypes were retained as *Lgh* plastomes and submitted to Genbank.
300 Hybrid assembly is the most complete and the most faithful, with an average of 1.6 contigs (all
301 assemblers combined), and optimal correction using RATATOSK, which permits plastomes
302 with 99.94% accuracy compared to biocurated molecules (Add. Figure 3B).
303 Unfortunately, due to the absence of short read data, we could only perform self-corrected
304 long read assembly for *Lpm* using CANU. We also compared CANU and FLYE assembler
305 efficiency, and found that assembly using CANU produces 13 contigs whereas FLYE produces
306 12 contigs. In both cases, only three contigs are required to reconstitute a complete cpDNA
307 assembly (no gap, no N), with an SSC region oriented like those of the *Lgh* haplotype 2 and the

308 *Lo* plastome. Although it is more than likely that these two SSC region orientations also exist
309 for *Lpm*, the low number of nanopore sequences generated (68907 reads) and absence of
310 Illumina short reads prevented us from demonstrating the existence of both haplotypes. As a
311 result, only the "haplotype 2" generated sequence was deposited to Genbank.

312 **Annotation and comparison of *Ludwigia* plastomes**

313 **1. General Variations**

314 Plastomes of the three species of Ludwigia sp., *Lgh*, *Lpm* and *Lo,* are circular double-
315 stranded DNA molecules (Figure 3) which are all (as shown in Table 1) approximately the same
316 size: *Lo* being slightly smallest and *Lgh* the biggest. The overall GC content is almost the same
317 for the three species and the GC contents of the IR regions are higher than those of the LSC and
318 SSC regions. Between the three species, the lengths of the total chloroplasts, LSC, SSC, and IR
319 are broadly similar (Table 2) and the three plastomes are perfectly syntenic if we orient the SSC
320 fragments the same way.

321 All three Ludwigia sp. plastomes contain the same number of functional genes (134 in total)
322 encoding 85 proteins (embracing 7 duplicated in the IR region: *ndhB*, *rpl2*, *rpl23*, *rps7*, *rps12*,
323 *ycf2*, *ycf15*), 37 tRNAs (including trnK-UUU which contains *matK*), and 8 rRNAs (16S, 23S,
324 5S, and 4.5S as duplicated sets in the IR). Among these genes, 18 contain introns, of which six
325 are tRNAs (Table 2). Only the *rps12* gene is a trans-spliced gene. A total of 46 genes are
326 involved in photosynthesis, and 71 genes related to transcription and translation, including a
327 bacterial-like RNA polymerase and 70S ribosome, as well as a full set of transfer RNAs
328 (tRNAs) and ribosomal RNAs (rRNAs). Six other protein-coding genes are involved in
329 essential functions, such as *accD,* which encodes the β-carboxyl transferase subunit of acetyl-
330 CoA carboxylase, an important enzyme for fatty acid synthesis; *matK* encodes for maturase K,
331 which is involved in the splicing of group II introns; *cemA*, a protein located in the membrane
332 envelope of the chloroplast is involved in the extrusion of protons and therby indirectly allows

333 the absorption of inorganic CO2 in the plastids; *clpP1* which is involved in proteolysis, and;
334 *ycf1*, *ycf2*, two ATPases members of the TIC translocon. Finally, a highly pseudogenized *ycf15*
335 locus was annotated in the IR even though premature stop codons indicate loss of functionality.

336 **2. Segments Contractions/Expansion**

337 The junctions between the different chloroplast segments were compared between three
338 Ludwigia sp. (*Lpm*, *Lgh* and *Lo),* and we found that the overall resemblance of Ludwigia sp.
339 plastomes was confirmed at all junctions (Figure 4A). In all three genomes, *rpl22, rps19, and*
340 *rpl2* were located around the LSC/IRb border, and *rpl2*, *trnH*, and *psbA* were located at the
341 IRa/LSC edge. The JSB (junction between IRb and SSC) is either located in the *ndhF* gene or
342 the *ycf1* gene depending on the orientation of the SSC region (Figure 4B). The *ycf1* gene was
343 initially annotated as a 1139 nt pseudogene that we biocurate as a larger gene (5302 nt) with a
344 frameshift due to a base deletion, compared to *Lg* and *Lo* which both carry a complete *ycf1*
345 gene.

346 If we compare Ludwigia sp. junctions with those of other Onagraceae plastomes (Figure
347 5), we can observe that the JLB and JLA connections are well-preserved in the whole family,
348 whereas JSB and JLA differ. Concerning JSB, in the five Onagraceae family genera with
349 available plastomes, *ndhF* is duplicated, with the exception of Circaea sp. and Ludwigia sp. For
350 *Oenothera villosa*, the first copy of *ndhF*, which is located in the IRb, overlaps the JSB border,
351 whereas for *Oenothera lindheimeri*, *Epibolium amurense* and Chamaenerion sp., *ndhF* is only
352 located in inverted repeats. Only Circaea sp. and Ludwigia sp. have a unique copy of this locus,
353 and it is found in the SSC segment (Figure 5). At the JSA border, in Circaea sp., the *ycf1* gene
354 crosses the IRa/SSC boundary and extends into the IRa region.

355 When comparing the respective sizes of chloroplast fragments (IR/SSC/LSC) in
356 Onagraceae, it can be observed that *Ludwigia* species exhibit expansions in the SSC and LSC
357 regions which are not compensated by significant contractions in the IR regions. This is likely

358 due to the relocation of the *ndhF* in the SSC region and *rps19* in the LSC region. Additionally,
359 there may be significant size variations in the intergenic region between *trnI* and *ycf2*, as well
360 as the intergenic segment containing the *ycf15* pseudogene (Add. Figure 4).

361 **3. Repeats and SSRs analysis**

362 In this study, we analyzed the nature and distribution of single sequence repeats (SSR), as
363 their polymorphism is an interesting indicator in phylogenetic analyses. A total of 65 (*Lgh*), 48
364 (*Lpm*) and 45 (*Lo*) SSRs were detected, the majority being single nucleotide repeats (38–21),
365 followed by tetranucleotides (12–10) and then di-, tri- and penta-nucleotides (Add. Figure 5A).
366 Mononucleotide SSRs are exclusively composed of A and T, indicating a bias towards the use
367 of the A/T bases, which is confirmed for all SSRs (Add. Figure 5B). In addition, the SSRs are
368 mainly distributed in the LSC region for the three species, which is probably biased by the fact
369 that LSC is the longest segment of the plastome (Add. Figure 5C). The analysis of SRR
370 locations revealed that most were distributed in non-coding regions (intergenic regions and
371 introns, Add. Figure 5D).

372 The chloroplast genomes of the three *Ludwigia* species were also screened for long repeat
373 sequences. They were counted in a non-redundant way (if smaller repetitions were included in
374 large repeats, only the large ones were considered). Four types of repeats (tandem, palindromic
375 inverted and direct) were surveyed in the three *Ludwigia* sp. plastomes. No inverted repeats
376 were detected with the criteria used.

377 For the three other types of repeats, here are their distributions:

378 ***Tandem repeats*** **(Table 3A)**: Perfect tandem repeats (TRs) with more than 15 bp were
379 examined. Twenty-two *loci* were identified in the three *Ludwigia* sp. plastomes (*Lgh*, *Lpm*, *Lo*),
380 heterogeneously distributed as shown in Table 3A: 13 loci (plus one imperfect) in *Lo*, nine loci
381 (plus one imperfect) in *Lgh* and seven loci (plus two imperfect) in *Lpm*. It can therefore be seen
382 that the TR distributions (occurrence and location) are specific to each plastome, since only

383 four pairs are common to the three species. Thus, nine TRs are unique to *Lo*, three to *Lpm* and
384 three to *Lgh.* Two pairs are common to *Lgh* and *Lpm* and one is common to *Lo* and *Lgh.* TRs
385 are mainly intergenic or intronic but are detected in two genes (*accD* and *ycf1)*. These genes
386 have accelerated substitution rates, although this does not generate a large difference in their
387 lengths. This point will be developed later in this article.

388 ***Direct repeats*** **(Table 3B)**: There are few direct (non-tandem) repeats (DRs) in the
389 chloroplast genomes of *Ludwigia* sp. A single direct repeat of 41 nt is common to the three
390 species, at 2 kb intervals, in *psaB* and *psaA* genes. This DR corresponds to an amino acid repeat
391 [WLTDIAHHHLAIA] which corresponds to a region predicted as transmembrane. We then
392 observe three direct repeats conserved in *Lpm* and *Lgh* in *ycf1*, *accD* and *clpP1* respectively,
393 two unique DRs in *Lo* (in the *accD* gene and *rps12-clpP1* intergene) and one in *Lgh* (in the
394 *clpP1* intron 1 and *clpP1* intron 2).

395 ***Palindromes*** **(Table 3C)**: Palindromic repeats make up the majority of long repetitions,
396 with the numbers of perfect repeats varying from 19, 24 and 26 in *Lo*, *Lgh* and *Lpm*,
397 respectively, and the number of quasi-palindromes (1 mutation) varying between 8, 3 and 6.
398 They are mainly found in the intronic and intergenic regions, with the exception of six genic
399 locations in *psbD*, *ndhK*, *ccsA* and *rpl22,* and two palindromic sequences in *ycf2.* These gene
400 palindromic repeats do not seem to cause genetic polymorphism in *Ludwigia* and can be
401 considered as silent.

402 Thirteen palindromes are common to the three species (including 2 with co-variations in
403 *Lo*). 13 others present in *Lpm* and *Lgh* correspond to quasi-palindromes (QPs) in *Lo* due to
404 mutated bases, and conversely, three *Lo* perfect palidromes are mutated in *Lpm* and *Lgh*.
405 Finally, only five palindromes are species specific. Two in particular are located in the
406 hypervariable intergenic spacer *ndhF-rpl32*, and are absent in *Lo* due to a large deletion of 160
407 nt.

408 **4. Repeat distribution in LSC, SSC and IR segments**

409 In the IRa/IRb regions, repeats are only identified in the first 9 kb region between *rpl2* and
410 *ycf2*: a tandem repeat in the *Lpm rpl2* intron, and a tetranucleotide repeat, [TATC]*3, located
411 in the *ycf2* gene in the 3 species. In *ycf2* we also found 1 common palindrome (16 nt), a single
412 palindrome in *Lo* (20 nt, absent following an A:G mutation in the 2 other species), as well as a
413 shared tandem repeat (24 nt), and an additional 15 nt tandem repeat in *Lo* which adds 4 amino
414 acids to protein sequence.

415 In the SSC region, the repeats are almost all located in the intergenic and/or intronic
416 regions, with a hotspot between *ndhF* and *ccsA*. There is also a shared microsatellite in *ndhF,*
417 and a palidrome (16 nt) in *ccsA* which is absent in *Lo* (due to an A:C mutation), resulting in a
418 synonymous mutation (from isoleucine to leucine). We also observed multiple and various
419 repeats in the *ycf1* gene: 3 common poly-A repeats (from 10 to 13 nt), 3 species-specific
420 microsatellites (ATAG)*3 and (ACCA)*4 in *Lgh* and (CAAC)*3 in *Lo,* as well as two direct
421 repeats of 32 nt (37 nt spacing), which were absent from *Lo* due to a G:T SNP. Two tandem
422 repeats were also observed in *Lo* and *Lgh*. Neither of these repeats are at the origin of the
423 frameshift causing the pseudogenization of *ycf1* in *Lo*, this latter being due to a single deletion
424 of an A at position 3444 of the gene.

425 Finally, in the LSC region, the longest segment, which consequently contains the maximum
426 number of repeats, we still observed a preferential localization in the intergenic and intronic
427 regions since only genes *atpA*, *rpoC2*, *rpoB*, *psbD*, *psbA*, *psbB*, *ndhK* and *clpP1* contain either
428 mononucleotic repeats (poly A and T), palindromes, or microsatellites (most often common to
429 the three species and without affecting the sequences of the proteins produced). As mentioned
430 earlier, the only exception is the *accD* gene, which contains several direct and tandem repeats
431 in *Lgh* and *Lpm,* corresponding to a region of 174 nt (58 amino acids) missing in *Lo* and,
432 conversely, a direct repeat of 40 nucleotides, in a region of 147 nt (49 aa), which is present in

433 *Lo* and missing in the other two species. These tandem repeats lead to the presence of four
434 copies of 9 amino acids [DESENSNEE] in *Lgh* and *Lpm*, two of which form a larger duplication
435 of 17 aa [FLSDSDIDDESENSNEE]. Similarly, the TRs present only in *Lo* generate two perfect
436 9 amino acid repeats [EELSEDGEE], included in two longer degenerate repeats of 27 nt (Add.
437 Figure 6). It should be noted that though these TRs do not disturb the open reading phases, it is
438 still possible for them to form an intron which is not translated. Different functional studies will
439 be necessary to clarify this point. The presence of polymorphisms of the *accD* gene between
440 *Lo* and the two species (*Lpm*, *Lgh*) is interesting because *accD*, that encodes a subunit of acetyl-
441 CoA carboxylase (EC 6.4.1.2). This enzyme is essential in fatty acid synthesis and also
442 catalyzes the synthesis of malonyl-CoA, which is necessary for the growth of dicots, plant
443 fitness and leaf longevity, and is involved in the adaptation to specific ecological niches [57].
444 Large *accD* expansions due to TRs have also been described in other plants such as *Medicago*
445 [58] and *Cupressophytes* [59]. Some authors have suggested that these inserted repeats are not
446 important for acetyl-CoA carboxylase activity as the reading frame is always preserved, and
447 they assume that these repeats must have a regulatory role [60].

448 **5. Sequence Divergence Analysis and Polymorphic Loci Identification**

449 Determination of divergent regions by MVista, using *Lo* as a reference, confirmed that the
450 three Ludwigia sp. plastomes are well preserved if the SSC segment is oriented in the same way
451 (Add. Figure 7). Sliding window analysis (Figure 6) indicated variations in definite coding
452 regions, notably *clpP*, *accD*, *ndh5*, *ycf1* with high Pi values, and to a lesser extent, *rps16*, *matK*,
453 *ndhK*, *petA*, *ccsA* and four tRNAs (*trnH*,*trnD*, *trnT* and *trnN*). These polymorphic *loci* could be
454 suitable for inferring genetic diversities in Ludwigia sp.

455 A comparative analysis of the sizes of protein coding genes sizes also shows that the *rps11*
456 gene initially annotated in *Lo* is shorter than those which have been newly annotated in *Lgh* and
457 *Lpm* (345 bp instead of 417 bp). Comparative analysis by BLAST shows that it is the long form

458 which is annotated in other Myrtales, and the observation of the locus in *Lo* shows a frameshift
459 mutation (deletion of a nucleotide in position 311). Functional analysis would be necessary to
460 check whether the *rps11* frameshift mutation produces shorter proteins that have lost their
461 function. And only obtaining the complete genome will verify whether copies of some of these
462 genes have been transferred to mitochondrial or nuclear genomes. Such *rps11* horizontal
463 transfers have been reported for this gene in the mitochondrial genomes of various plant
464 families [61]. This also applies to *ycf1*, found as a pseudogene in *Lo* (as specified previously),
465 although it is not known if this reflects a gene transfer or a complete loss of function [62,
466 63].Moreover, there is a deletion of nine nucleotides in the 3' region of the *rpl32* gene in *Lgh*
467 and *Lpm*, leading to a premature end of the translation and the deletion of the last 4 amino acids
468 [QRLD], which are replaced by a K. However, if we look carefully at the preserved region as
469 defined by the RPL32 domain (CHL00152, member of the superfamily CL09115), we see that
470 the later amino acids are not important for *rpl32* function since they are not found in the
471 orthologs.

472 Our results show that the Ka/Ks ratio is less than 1 for most genes (Figure 7). This indicates
473 adaptive pressures to maintain the protein sequence except for *matK* (1.17 between *Lgh* and
474 *Lpm*), *accD* (2.48 between *Lgh* and *Lo* and 2.16 between *Lpm* and *Lo*), *ycf2* (4.3 between both
475 *Lgh-Lp* and *Lo*) and *ccsA* (1.4 between both *Lgh-Lpm* and *Lo*), showing a positive selection for
476 these genes, and a possible key role in the processes of the species' ecological adaptations. As
477 we have already described the variability in the *accD* sequence, we will focus on *ycf2*, *matK,*
478 and *ccsA* variations.

479 Concerning *ccsA*, the variations observed, although significant, concern only five amino
480 acids, and modifications do not seem to affect the C-type cytochrome synthase gene function.

481 Concerning *ycf2*, our analysis shows that this gene is highly polymorphic with 256 SNPs
482 that provoke 10 deletions, 7 insertions, 21 conservative and 49 non-conservative substitutions

483 in *Lo* (Add. Figure 8), compared to *Lgh* and *Lpm* (100 % identical). This gene has been shown
484 as "variant" in other plant species such as *Helianthus tuberosus* [64].
485 The *matK* gene has been used as a universal barcoding locus to enable species
486 discrimination of terrestrial plants [65], and is often, together with the *rbcL* gene, the only
487 known genetic resource for many plants. Thus, we propose a phylogenetic tree from *Ludwigia*
488 *matK* sequences (Figure 8). It should however be noted that this tree contains only 149 amino
489 acids common to all the sequences (out of the 499 in the complete protein). As only three
490 complete *Ludwigia* plastomes are available, we cannot specify whether these barcodes are
491 faithful to the phylogenomic history of *Ludwigia* in the same way as the complete plastome. In
492 any case, for this tree, we can see that *Lo* stands apart from the other Ludwigia sp., *Lpm* and
493 *Lgh*, and that the *L. grandiflora* subsp. *hexapetala* belongs to the same branch as the species *L.*
494 *ovalis* (aquatic taxon used in aquariums [66]), *L. stolonifera* (native to the Nile, found in a
495 variety of habitats, from freshwater wetlands to brackish and marine waters) [67] and *L.*
496 *adscendens* (common weed of rice fields in Asia) [68]. *Lpm* is in a branch unique to its species
497 but close to the *L. grandiflora* subsp. *hexapetala* branch.
498
499 **Discussion**
500 In the present study, we first sequenced and *de novo* assembled the chloroplast (cp) genomes
501 of *Ludwigia peploides* (*Lpm*) and *Ludwigia grandiflora* (*Lgh*), two species belonging to the
502 Onagraceae family. We employed a hybrid strategy and demonstrated the presence of two cp
503 haplotypes in *Lgh* and one haplotype in *Lpm*, although the presence of both haplotypes in *Lpm*
504 is likely. Furthermore, we compared these genomes with those of other species in the
505 Onagraceae family to expand our knowledge of genome organization and molecular evolution
506 in these species.

507 Our findings demonstrate that the utilization of solely short reads has failed to produce complete
508 plastomes, likely due to challenges posed by long repeats and rearrangements. On the other
509 hand, relying solely on long reads resulted in a lower quality sequence due to insufficient
510 coverage and sequencing errors. After conducting our research, we discovered that hybrid
511 assembly, which incorporates both long and short read sequences, resulted in the most superior
512 complete assemblies. This innovative approach capitalizes on the advantages of both
513 sequencing technologies, harnessing the accuracy of short read sequences and the length of long
514 read sequences. Our findings corroborate with similar results obtained in studies on other
515 chloroplasts, such as those in *Eucalyptus* [21], *Falcataria* [69], *Carex* [70] or *Cypripedium* [71].
516 As a result of this strategy, we were able to successfully identify the presence of two haplotypes
517 in *Lgh*, which is a first for *Ludwigia,* as the plastome of *L. octovalvis* was only delivered in the
518 form of one haplotype [72]. Due to the unavailability of sequence data for *Ludwigia octovalvis*
519 and our exclusive use of long reads for *Ludwigia peploides*, we are unable to conclusively
520 identify the presence of these two forms in the *Ludwigia* genus. However, we believe that they
521 are likely to be present. Unfortunately, the current representation of plastomes in GenBank
522 primarily consists of short-read data, which may result in an underrepresentation of this
523 polymorphism. It is unfortunate that structural heteroplasmy, which is expected to be
524 widespread in angiosperms, has been overlooked. Notably, the existence of two plastome
525 haplotypes has been widely identified in the closely related order of Myrtales (Eucalyptus sp.),
526 and more commonly in Angiosperms, with 58 species across 16 other orders showing similar
527 patterns [22]. Recent studies have also revealed the presence of two cp haplotypes in other
528 orders, such as *Asparagales* (*Ophrys apifera* orchid, [73]), Brassicales (*Carica papaya*,
529 *Vasconcellea pubescens*, [74]), Solanales (*Solanum tuberosum*, [75]), and even in a new order,
530 Laurales (Avocado *Persea americana*, [76]). Furthermore, Wanichthanarak *et al*. (2023)
531 revisited 24 Ramanaceae published chloroplast genomes and identified a likely form of

532 heteroplasmy in *Rhamnus crenata* [77]. These emerging findings highlight the potential
533 underestimation of the occurrence of two chloroplast haplotypes in angiosperms and emphasize
534 the importance of identifying and considering such heteroplasmy during *de novo* or revisited
535 chloroplast genome assemblies.

536 The chloroplast genome sizes for the three genera of Onagraceae subfam. Onagroideae varied
537 as follows: Circaea sp. ranged from 155,817 bp to 156,024 bp, Chamaenerion sp. ranged from
538 159,496 bp to 160,416 bp, and Epilobium sp. ranged from 160,748 bp to 161,144 bp [78]. Our
539 study revealed that the size of the complete chloroplast of *Ludwigia* (Onagraceae subfamily
540 Ludwigioideae) ranged from 159,369 bp to 159,584 bp, which is remarkably similar to other
541 Onagraceae plants (average length of 162,030 bp). Furthermore, *Ludwigia* plastome sizes are
542 consistent with the range observed in Myrtales (between 152,214 to 171,315 bp, [79]). In the
543 same way, similar overall GC content was found in *Ludwigia sp*. (from 37.3 to 37.4%), *Circaea*
544 *sp.* (37.7 to 37.8%), *Chamaenerion sp.* and *Epilobium sp.* (38.1 to 38.2%, [78]) and more
545 generally for the order Myrtales (36.9–38.9%, with the average GC content being 37%, [79]).
546 Higher GC content of the IR regions (43.5%) found in *Ludwigia* sp. has already been shown in
547 the Myrtales order (39.7–43.5%) and in other families/orders such as Amaranthaceae (order
548 Caryophyllales, [80]) or Lamiaceae (order Lamiales, [81]), and is mainly due to the presence
549 of the four GC rich rRNA genes.

550 The complete chloroplast genomes of the three Ludwigia species encoded an identical set of
551 134 genes including 85 protein-coding genes, 37 tRNA genes and eight ribosomal RNAs,
552 consistent with gene content found in the Myrtales order, with a gene number varying from 123
553 to 133 genes with 77–81 protein-coding genes, 29–31 tRNA gene and four rRNA genes [79].
554 Chloroplast genes have been selected during evolution due to their functional importance [82].
555 In our current study, we made the noteworthy discovery that *matK*, *accD*, *ycf2*, and *ccsA* genes
556 were subjected to positive selection pressure. These genes have frequently been reported in

557 literature as being associated with positive selection, and are known to play crucial roles in
558 plant development conditions. *Lgh* and *Lpm* are known to thrive in aquatic environments, where
559 they grow alongside rooted emergent aquatic plants, with their leaves and stems partially
560 submerged during growth, as reported by Wagner et al. in 2007 [1]. Both species possess the
561 unique ability of vegetative reproduction, enabling them to establish themselves rapidly in
562 diverse habitats, including terrestrial habitats, as noted by Haury et al [7]. Additionally, *Lo* is a
563 wetland plant that typically grows in gullies and at the edges of ponds, as documented by
564 Wagner *et al.* in 2007 [1]. Given their ability to adapt to different habitats, these species may
565 have evolved specialized mechanisms to cope with various abiotic stresses, such as reduced
566 carbon and oxygen availability or limited access to light in submerged or emergent conditions.
567 Concerning *matK*, Barthet et al [83] demonstrated the relationship between light and
568 developmental stages, and MatK maturase activity, suggesting important functions in plant
569 physiology. This gene has recently been largely reported to be under positive selection in an
570 aquatic plant (*Anubias* sp., [84]), and more generally in terrestrial plants (*Pinus sp*, [85] or
571 *Chrysosplenium sp.* [86]). The *accD* gene has been described as an essential gene required for
572 leaf development [87] and longevity [88] in tobacco (*Nicotiana tabacum*). Under drought stress,
573 plant resistance can be increased by inhibiting *accD* [89], and conversely, enhanced in response
574 to flooding stress by upregulating *accD* accumulation [90]. Hence, we can hypothesize that the
575 positive selection observed on the *accD* gene can be explained by the submerged and emerged
576 constraints undergone by *Ludwigia* species. The *ycf2* gene seems to be subject to adaptive
577 evolution in *Ludwigia species*. Its function, although still vague, would be to contribute to a
578 protein complex generating ATP for the TIC machinery (proteins importing into the
579 chloroplasts [91, 92]), as well as plant cell survival [93, 94]. The *ccsA* gene positive selection
580 is found in some aquatic plants such as *Anubia sp.* [84], marine flowering plants as
581 *Zostera* species [95], and some species of Lythraceae [89]. The *ccsA* gene is required for

582 cytochrome c biogenesis [96] and this hemoprotein plays a key role in aerobic and anaerobic
583 respiration, as well as photosynthesis [97]. Furthermore, we showed that *Lgh* colonization is
584 supported by metabolic adjustments mobilizing glycolysis and fermentation pathways in
585 terrestrial habitats, and the aminoacyl-tRNA biosynthesis pathway, which are key components
586 of protein synthesis in aquatic habitats [8] It can be assumed that the ability of *Ludwigia* to
587 invade aquatic and wet environments, where the amount of oxygen and light can be variable,
588 leads to a high selective pressure on genes involved in respiration and photosynthesis.

589 Molecular markers are often used to establish population genetic relationships through
590 phylogenetic studies. Five chloroplasts (*rps16*, *rpl16*, trnL-trnF, trnL-CD, *trnG*) and two
591 nuclear markers (ITS, *waxy*) were used in previous phylogeny studies of *Ludwigia sp*. [4].
592 However, no SSR markers had previously been made available for the *Ludwigia* genus, or more
593 broadly, the Onagraceae. In this study, we identified 45 to 65 SSR markers depending on the
594 Ludwigia species. Most of them were AT mononucleotides, as already recorded for other
595 angiosperms [98, 99]. In addition, we identified various genes with highly mutated regions that
596 can also be used as SNP markers. Chloroplast SSRs (cpSSRs) represent potentially useful
597 markers showing high levels of intraspecific variability due to the non-recombinant and
598 uniparental inheritance of the plastomes [100, 101]. Chloroplast SSR characteristics for
599 *Ludwigia sp*. (location, type of SSR) were similar to those described in most plants. While the
600 usual molecular markers used for phylogenetic analysis are nuclear DNA markers, cpSSRs have
601 also been used to explore cytoplasmic diversity in many studies [102-104]. To conclude, the 13
602 highly variable loci and cpSSRs identified in this study are potential markers for population
603 genetics or phylogenetic studies of *Ludwigia* species, and more generally, Onagraceae.
604

605 **Conclusion**

606 In this study, we conducted the first-time sequencing and assembly of the complete plastomes
607 of *Lpm* and *Lgh*, which are the only available genomic resources for functional analysis in both
608 species. We were able to identify the existence of two haplotypes in both *Lpm* and *Lgh*, while
609 the absence of the *Lo* genome precluded further investigation for this species. Comparison of
610 all 10 Onagraceae plastomes revealed a high degree of conservation in genome size, gene
611 number, structure, and IR boundaries. However, to further elucidate the phylogenetic analysis
612 and evolution in Ludwigia and Onagraceae, additional chloroplast genomes will be necessary,
613 as highlighted in recent studies of Iris and Aristidoideae species [105].
614

615 **Declarations**

616 - Availability of data and materials

617 The datasets generated and/or analysed during the current study were available in GenBank
618 (for *Lgh* haplotype 1, (LGH1) OR166254 and *Lgh* haplotype 2, (LGH2) OR166255; for *Lpm*
619 haplotype, (LPM) OR166256).

620 - Conflict of interest disclosure

621 The authors declare that they comply with the PCI rule of having no financial conflicts of
622 interest in relation to the content of the article

623 - Funding

624 The post-doctoral research grant of Anne-Laure Le Gac was supported by the Conseil regional
625 Bretagne (SAD18001).

626 - Acknowledgements

627 We are grateful to Luis Portillo-Lemus for developing the high molecular weight genomic
628 DNA extraction protocol. All sequencing experiments were performed at the PGTB
629 (doi:10.15454/1.5572396583599417E12).

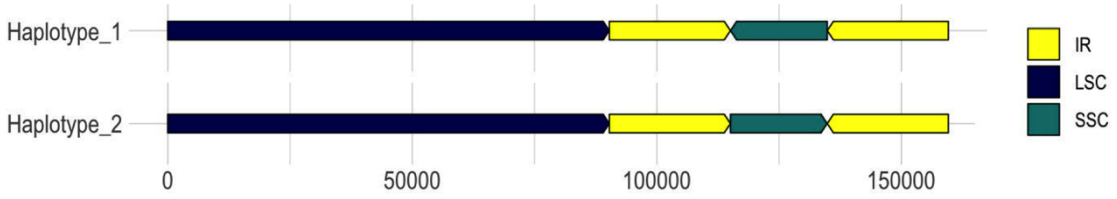


Figure 1: Two structural haplotypes of *L. grandiflora* plastomes representing the flip-flop organization of SSC segment.

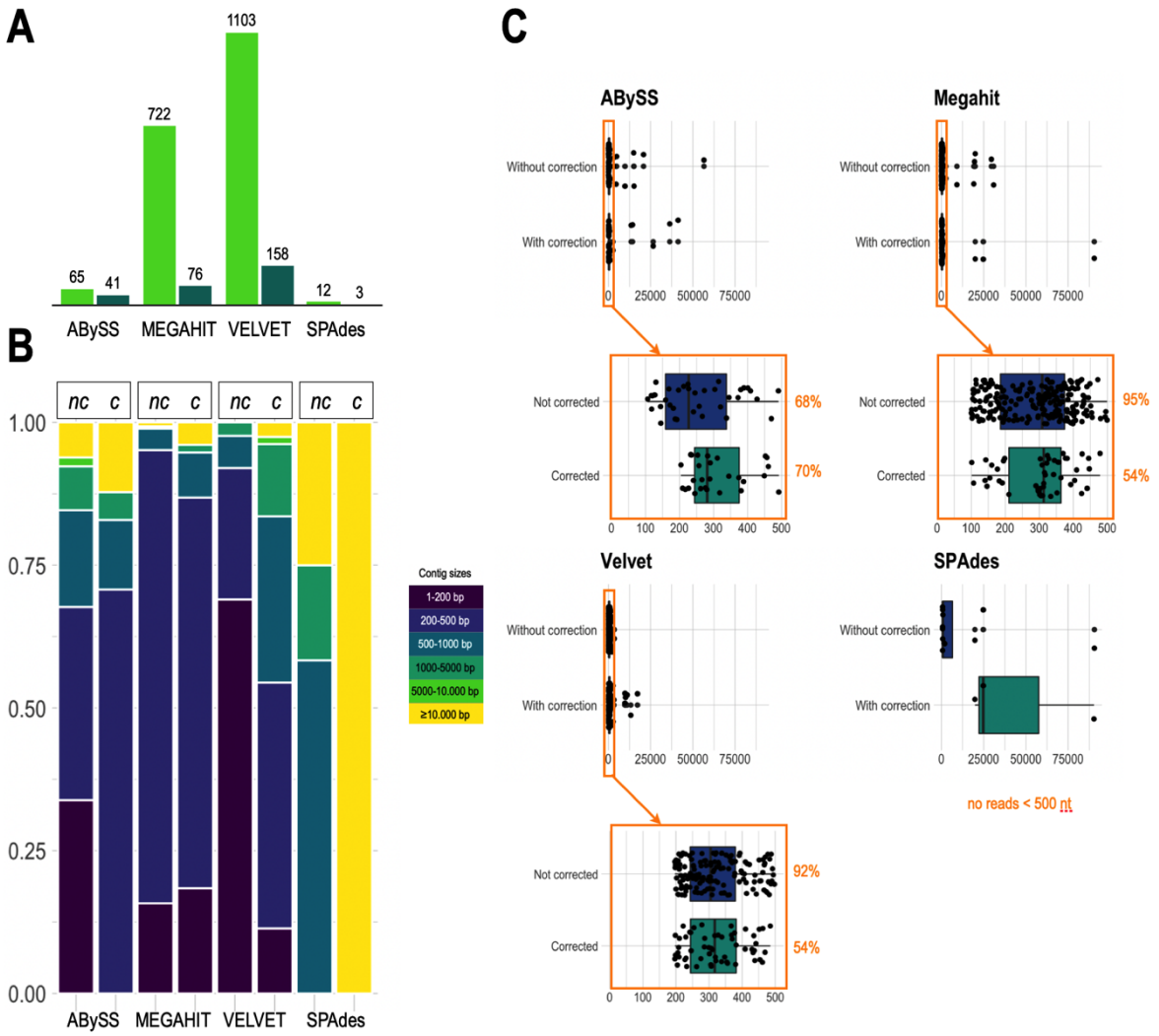
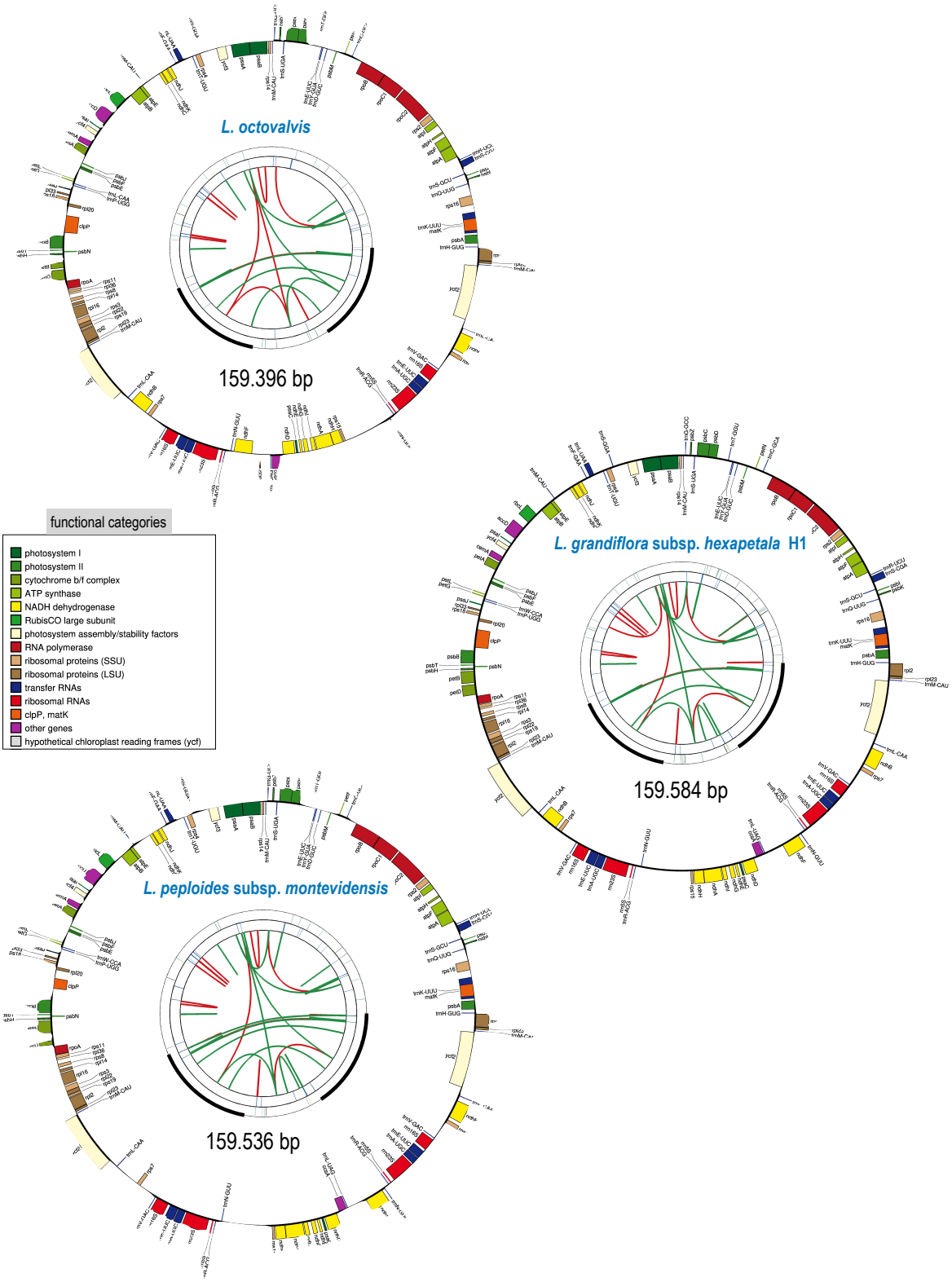


Figure 2: Comparative results of *L. grandiflora* short read (SR) assemblies. **A:** Total number of contigs obtained with the uncorrected (dark green) and corrected (light green) chloroplast SRs for the 4 assemblers (ABySS, MEGAHIT, Velvet and SPAdes). **B:** Comparison of the size of contigs assembled by the 4 tools using corrected or uncorrected SRs. **C:** Boxplot showing the distribution of these contigs by size and the improvement brought by the prior correction of the SRs with the long reads for each tool.



649

650 **Figure 3:** Circular representation of annotations plastomes in *Ludwigia octovalis*, *Ludwigia*
651 *grandiflora* and *Ludwigia peploides* using ogdraw. Each card contains four circles. From the
652 center outwards, the first circle shows forward and reverse repeats (red and green arcs,

respectively). The next circle shows tandem repeats as bars. The third circle shows the microsatellite sequences. Finally, the fourth and fifth circles show the genes colored according to their functional categories (see colored legend). Only the haplotype 1 of *L. grandiflora* is represented as haplotype 2 only diverge by the orientation of the SSC segment.

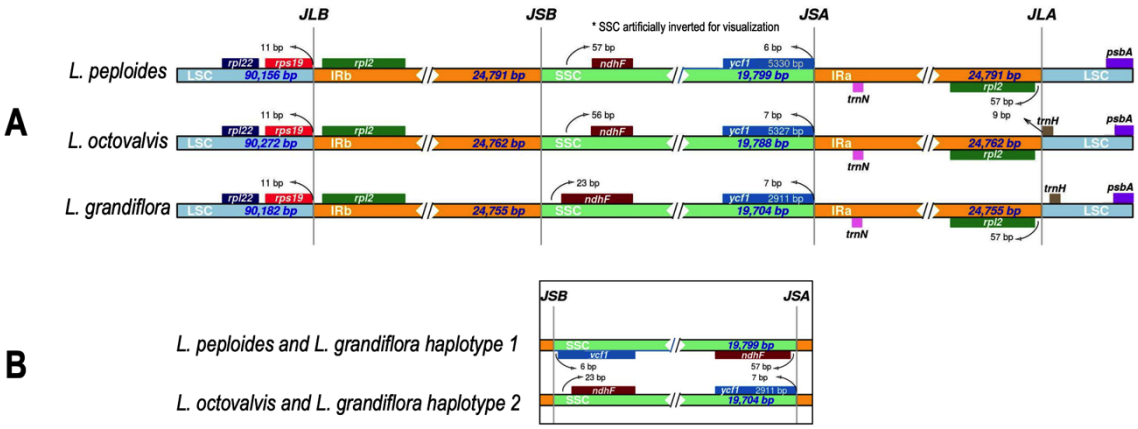


Figure 4: Comparison of the borders of LSC, SSC, and IR regions in Onograceae plastomes. **A:** Comparison of the junction between large single-copy (LSC, light blue), inverted repeat (IR, orange) and short single-copy (SSC, light green) regions among the chloroplast genomes of *L. octovalvis*, *L. peploides* and *L. grandiflora* (both haplotypes). Genes are denoted by colored boxes and the gaps between genes and boundaries are indicated by base lengths (bp). JLB: junction line between LSC and IRb; JSB: junction line between IRb and SSC; JSA: junction line between SSC and IRa; JLA: junction line between IRa and LSC. **B:** Comparison of SSC boundaries in haplotype 1 (*L. peploides* and *L. grandiflora* haplotype 1) and haplotype 2 (*L. octovalvis* and *L. grandiflora* haplotype 2) plastomes.

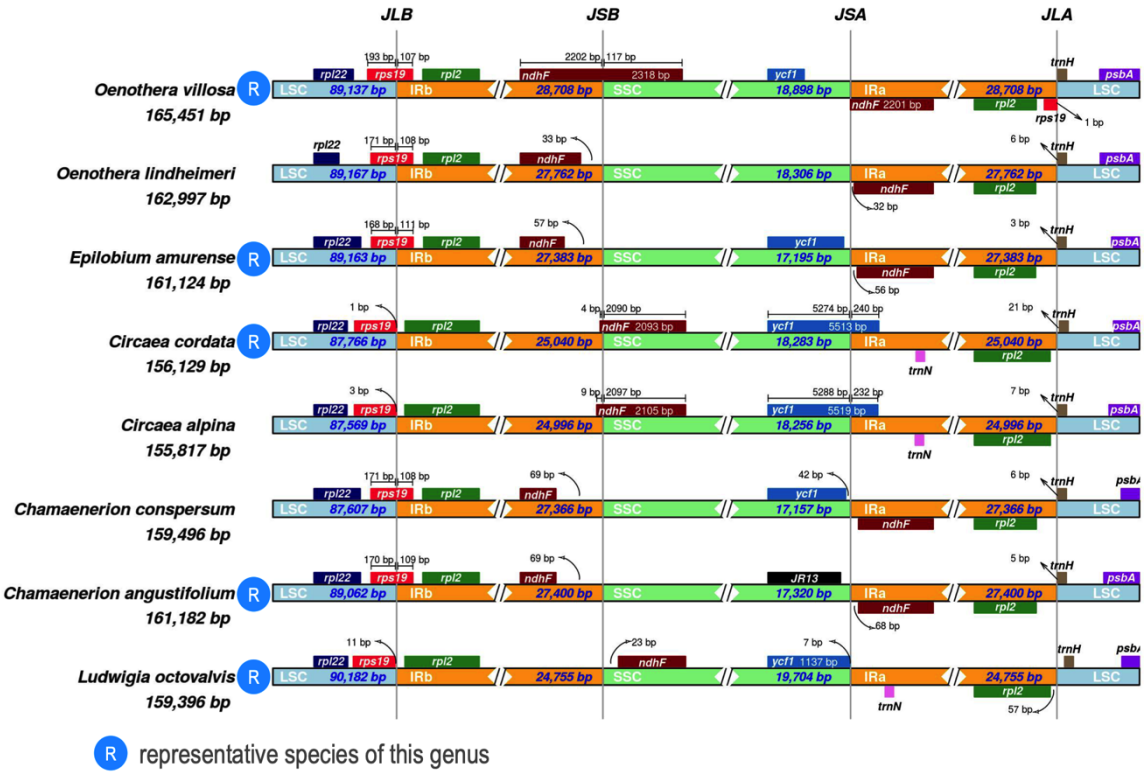
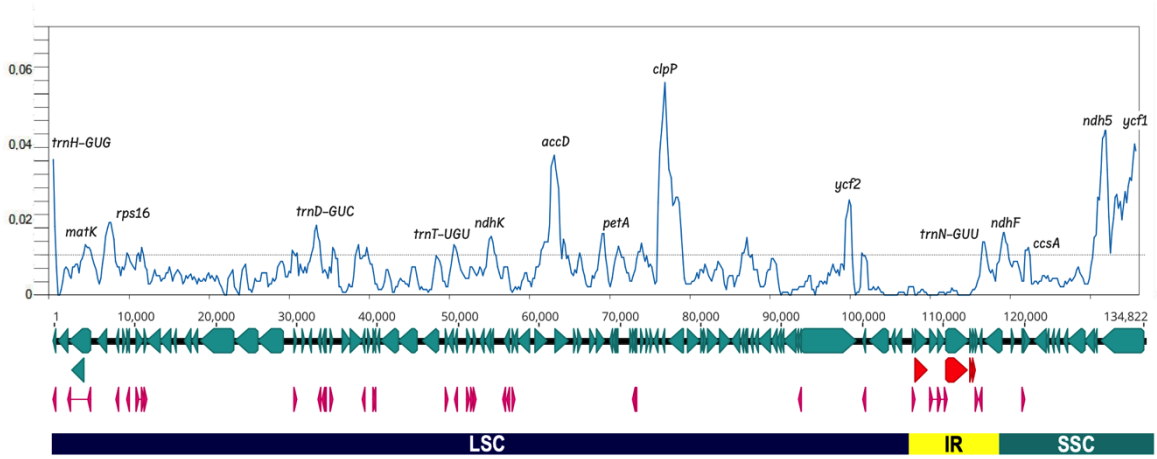


Figure 5: Comparison of LSC, SSC and IR regions boundaries in Onograceae chloroplast genomes. Representative sequences from each genus have been chosen (noted R on the diagram) except for *Oenothera lindheimeri* (only 89.35 % identity with others *Oenothera*), *Circaea alpina* (99.5 % identity but all others *Circaea* are 99.9% identical) and *Chamaenerion conspersum* (99% but all others *Chamaenerion* are ca. 99.7 identical). As shown in Figure 7, the 3 *Ludwigia* plastomas had the same structure, *L. octovalvis* was chosen as a representative of this genus.



679

680 **Figure 6:** Illustration of nucleotide diversity of the three *Ludwigia* chloroplast genome
681 sequences. The graph was generated using DnaSP software version 6.0 (windows length: 800
682 bp, step size: 200 bp). The x-axis corresponds to the base sequence of the alignment, and the y-
683 axis represents the nucleotide diversity (π value). LSC, SSC and IR segments were indicated
684 under the line representing the genes coding the proteins (in light blue) the tRNAs (in pink) and
685 the rRNAs (in red). The genes marking diversity hotspots are noted at the top of the peaks.

686

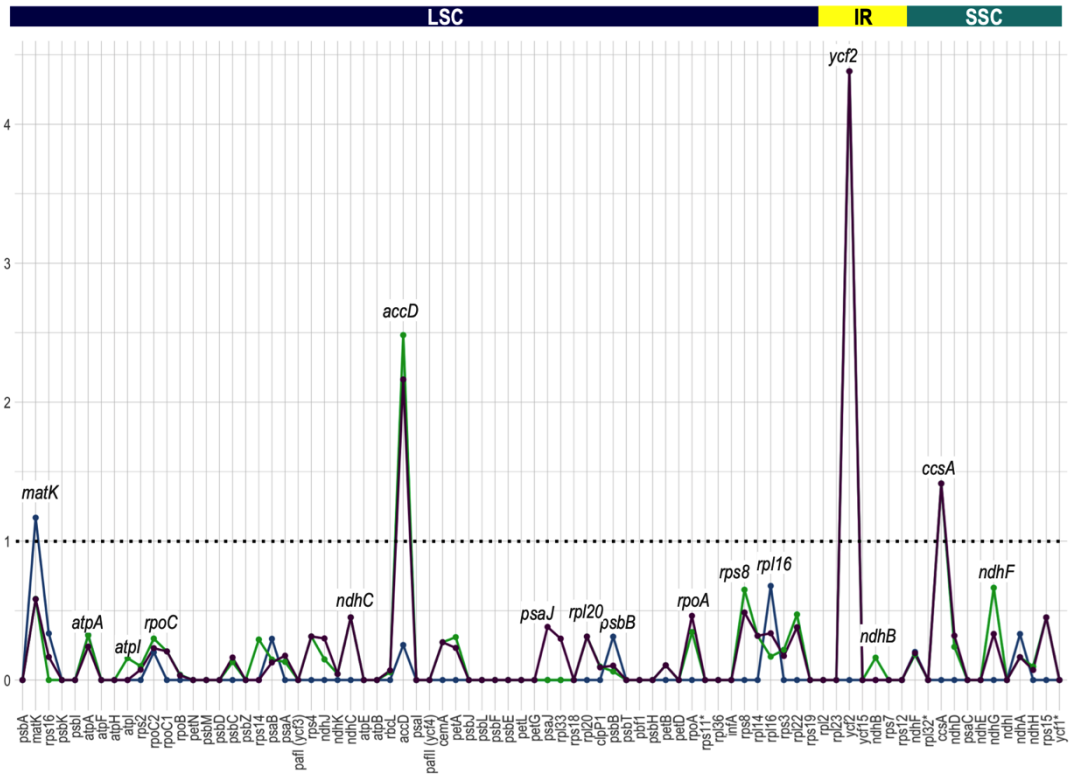


Figure 7: The Ka/Ks ratios of the 80 protein-coding genes of *Ludwigia* plastomes. The blue curve represents *L. grandiflora* versus *L. peploides*, purple curve denotes *L. grandiflora* versus *L. octovalvis* and green curve *L. peploides* versus *L. octovalvis.* Four genes (*matK*, *accD*, *ycf2* and *ccsA*) have Ka/Ks ratios greater than 1.0, whereas the Ka/Ks ratios of the other genes were less than 1.0.

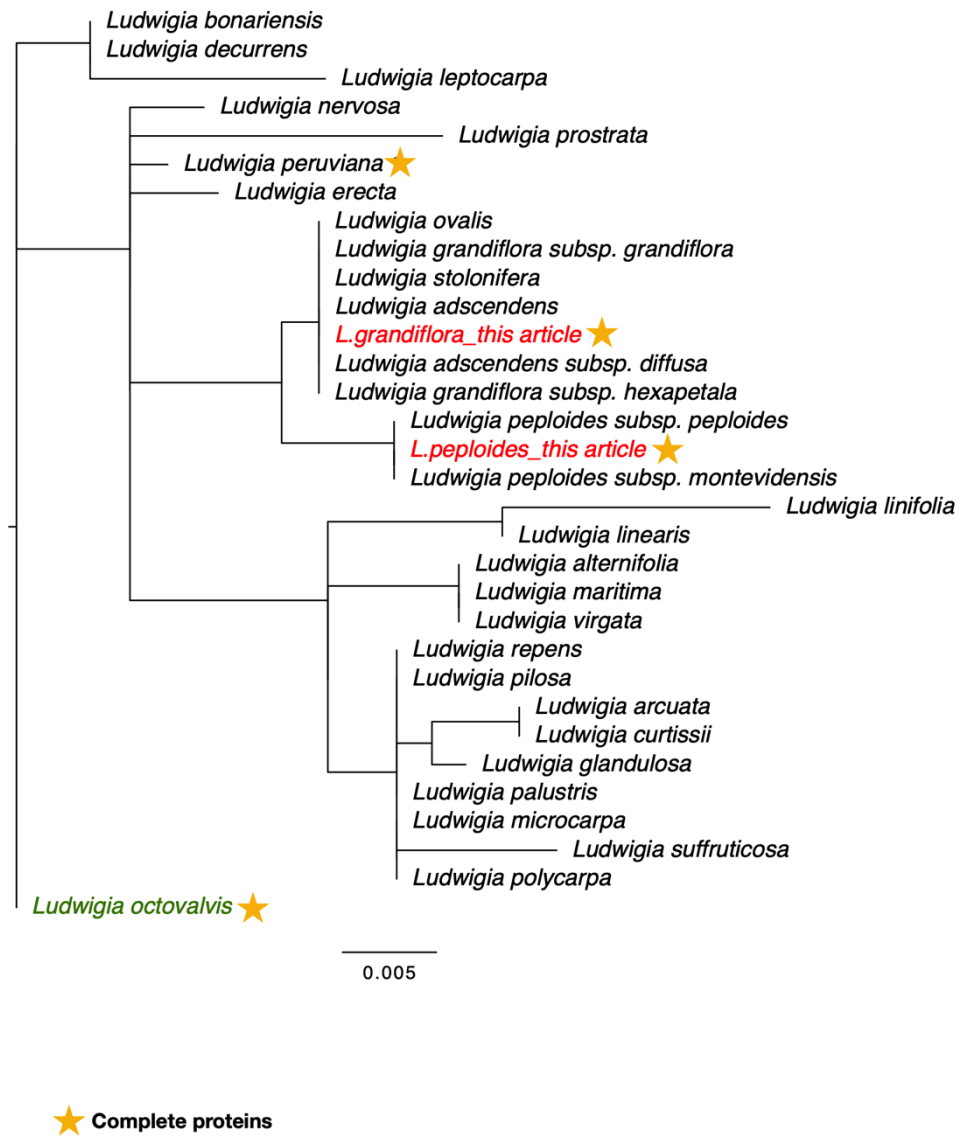


Figure 8: Phylogenetic tree based on the *Ludwigia matK* marker. Only four sequences are complete (499 aa, yellow star), the others correspond to barcode amplification ranging from 128 to 386 aa, with an average of 244 aa.

Table 1. The general characteristics of the 3 Ludwigia plastomes

	*L.octovalvis**	*L.grandiflora*	*L.peploides*
Size (bp)			
	159;396	159;584	159;537
LSC	90;183	90;272	90;156
SSC	19;703	19;788	19;799
IR	24;755	24;762	24;791
GC%			
	37;4	37;3	37;3
LSC	35;2	35;1	35;1
SSC	32	31;7	31;7
IR	43;5	43;5	43;4

* KX827312 (ref)

698

699
700

701

Table 2 : Genes present in the plastomes of *Ludwigia*

FUNCTION	NAME
Photosynthesis	
Rubisco	rbcL
Photosystem I (PSI)	psaA; psaB; psaC; psaI; psaJ
PSI assembly factors	ycf3# (pafI); ycf4 (pafII)
Photosystem II	psbA; psbB; psbC; psbD; psbE; psbF; psbH; psbI; psbJ; psbK; psbL; psbM; pbf1 (psbN) psbT; psbZ
ATP synthase	atpA; atpB; atpE; atpF#; atpH; atpI
Cytochrome *b6f*	petA; petB#; petD#; petG; petL; petN
Cytochrome biogenesis	ccsA
NADPH dehydrogenase	ndhA#; ndhB**#; ndhC; ndhD; ndhE; ndhF; ndhG; ndhH; ndhI; ndhJ
Transcription and translation	
Transcription	rpoA; rpoB; rpoC1#; rpoC2
Small ribosomal proteins	rps2; rps3; rps4; rps7**; rps8; rps11; rps12**#; rps14; rps15; rps16#; rps18; rps19
Large ribosomal proteins	rpl2**#; rpl14; rpl16#; rpl20; rpl22; rpl23**; rpl32; rpl33; rpl36
Translation initiation	infA
Ribosomal RNA	rrn5**; rrn4;5**; rrn16**; rrn23**
Transfer RNA	trnA-UGC**#;trnC-GCA;trnD-GUC;trnE-UUC;trnF-GAA;trnfM-CAU;trnG-GCC;trnG-UCC#;trnH-GUG;;trnI-CAU**;trnI-GAU**#;trnK-UUU#;trnL-CAA**;trnL-UAA#;trnL-UAG;trnM-CAU;trnN-GUU**;trnP-UGG;trnQ-UUG;trnR-ACG**;trnR-UCU;trnS-GCU;trnS-GGA;trnS-UGA;trnT-GGU;trnT-UGU;trnV-GAC**;trnV-UAC#;trnW-CCA;trnY-GUA
Other functions	
Group II intron splicing	matK
Inorganic carbon uptake	cemA
Protease	clpP1#
Fatty acid synthesis/Heat tolerance	accD
TIC machinery (protein import)	ycf1 (Tic214); ycf2**
Unknown function	ycf15**
	** duplicated in IR region; # spliced genes

702
703

Table 3:

Table 3A : Tandem repeats

Sequence	*L. octovalis (L.o)*	*L. grandiflora (L.g)*	*L. peploides (L.p)*	Length	Region	Locus	Comments
TTGTAGTCAGGGGTGTAGTACTAT				24	IRs	*ycf2*	
TAGAAGAGAGTGCAG		X	X	15	IRs	*ycf2*	15 nt deletion in L.g and L.p
ATGAAATATCGTATAATGAAGTACCACACGAGTGGATAT	X	X		39	IRs	*rpl2 intron*	39 nt deletion in L.g and L.o
AAAAATAGGATAGGAT		X	X	16	LSC	*ycf1-trnH-GUG*	56 nt deletion in L.g and L.p
TAAATTAATATCTATATA		X	X	18	LSC	psbZ-trnG-GCC	18 nt deletion in L.g and L.p
TTTTCTATCTATCTTATATCAA		X	X	22	LSC	*trnK-UUU-rps16*	22 nt deletion in L.g and L.p
AGATCCATAACATCATCAAA		X	X	20	LSC	*rps16 intron*	22 nt deletion in L.g and L.p
TATTAGTTATTAATATTATTAGA		X	X	23	LSC	trnP-UGG-psaJ	23 nt deletion in L.g and L.p
AATAATATATAATAACTTAAATA		X	X	23	LSC	*rpl33-rps18*	33 et 44 nt nt deletion in in L.g et L.p, respectively
TTTTTATTTAACATGCTATCAAATCAACAATGCCATACCGTAGGGCATCTGTT		X	X	53	LSC	*rpl20-clpP1*	107 nt deletion in L.g and L.p
ATATATTTCGATTCAATTC	X		X	19	LSC	*trnH-GUG-psbA*	3 copies in a 57 nt deletion in L.o and L.p
ATAGAAATATCAGTATTTGAGTG	X		X	23	LSC	*atpH-atpI*	23 nt deletion in L.o and L.p
TTAATTTTAATTGAAGAA	X		X	18	LSC	*psbJ-psbL*	17 and 24 nt deletion in L.o and L.p, respectively
TTAAAGAATATTAATATTC	imperfect TR			19	LSC	*trnR-UCU-atpA*	A -> C mutation in second copy in L.o
TATTATTATTATTAAT	X	X		16	LSC	*atpH-atpI*	16 nt deletion in L.g and L.o
TCTAAGGCTGAAATAAGG	X	X		18	LSC	pafI intron	18 nt deletion in L.g and L.o
TGTGAATCTATCTAT			X	15	LSC	*trnS-UGA-psbZ*	8 nt deletion in L.p
TTTTTTCTAGTA				12	LSC	*pafI intron*	
CTAGTTATTGACATGG		imperfect TR	imperfect TR	16	LSC	*psaJ-rpl33*	G -> A mutation in second in L.p et L.g
ATTTTTATTAACTCT	X		imperfect TR	15	SSC	*ycf1*	T->A mutation in first copy in L.p, other sequence in first copy in L.o
AATCAAATAGTTGAT		X	X	15	SSC	*ycf1*	other sequence in first copy of L.p and L.g
ATAATAATATATTTATTATTAATTAATA	X			28	SSC	*ndhF-rpl32*	160 nt deletion in L.o

Table 3B : Direct repeats

Sequence	*L. octovalis (L.o)*	*L. grandiflora (L.g)*	*L. peploides (L.p)*	Size (nt)	Spacers (nt)	Region	Locus	Comments
TTCAATTGGAACGGACGATTCGTCAATCATCT				32	37	SSC	*ycf1*	2 copies. In L.o, one mutation (G->A) in the second copie
CATCGATGATGAAAGTGAAAACAGTAATGAAGAGG	X			35	28 - 22 - 11	LSC	*accD*	3 perfects copies and 1 mutated (G->A) copie in *L.g* and *L.p*. Region of 174 nt deleted in *L.o*
AGATGGTGAAGAACCTTATGAAGATGGTGAAGAACCTTATG		X	X	41	22	LSC	*accD*	Region of 147 nt deleted in *L.g* and *L.p*
TATCAAATCAACAATGCCATACCGTAGGGCAT		X	X	32	22 - 21	LSC	*rps12-clpP1*	3 copies
TTAAGAGCCGTACAGGCACCTTTTGATGCATACGG	X				408 in *L.p*, 406 in *L.g*	LSC	*clpP1*	2 copies. In L.g, one mutation (C->T) in the second copie
TTAAGAGCCGTACAGGCACTTTTTGATGCATACGG	X		X	35	811	LSC	*clpP1 intron 1-intron 2*	
TGCAATAGCCAAATGATGATGAGCAATATCAGTCAGCCATA				41	2178		*psaB* & *psaA*	

Table 3C : Palindromic repeats

Common perfect palidromic repeats		
AGACTCTCATGAGAGTCT	*trnC-GCA - petN*	
ATTAAATAGAATATTCTATTTAAT	*trnE-UUC-trnT-GGU*	
TTGGTAAATTTACCAA	*psbD*	
TTCATTTCAATTTCAATTGAAATTGAAATGAA	*trnI-CAU-ycf2*	2 copies in IR
GAAAAAGGCCTTTTTC	*ycf2*	2 copies in IR
TCTCAAATGATTAATCATTTGAGA	*trnL-UAA* intron	
GGATTACTAGTAATCC	*trnD-GUC-trnY-GUA*	
TTTGAATGCATTCAAA	*trnG-UCC* intron	
ATATATTCGAATATAT	*trnG-UCC -trnR-UCU*	
TAGTAATTAATTACTA	*trnG-GCC-trnfM-CAU*	
CCAGTATGCATACTGG	*ndhK*	

Common palidromic repeats with covariation		
in L. octovalvis	*in L. grandiflora et L. peploides*	
ATAGAATCTATATTCTATTAGAATATAGATTCTAT	ATCGAATCTATATTCTATTAGAATATAGATTCGAT	*ndhC-trnV-UAC*
ATGTATATATATCGAT	ATCTATATATATAGAT	*trnE-UUC-trnT-GGU*

Common palindromic and quasi-palidromic repeats		
in L. octovalvis	*in L. grandiflora and L. peploides*	
TTTAACGAATATTAATATT t GTTAAA	TTTAACGAATATTAATATTCGTTAAA	*trnR-UCU-atpA*
TTAA c GAATATTAATATTCTTTAA	TTAAAGAATATTAATATTCTTTAA	*trnR-UCU-atpA*
AATTGTA c TTACAATT	AATTGTAATTACAATT	*ccsA*
AGGAAGATTGATCAATCTT t CT	AGGAAGATTGATCAATCTTCCT	*trnL-UAG-rpl32*
TTA c TAATATTACTAA	TTAGTAATATTACTAA	*trnK-UUU* intron
ATATAGAATAT c CTATAT	ATATAGAATATTCTATAT	*psbZ-trnG-GCC*
ACATATCATGATA g GT	ACATATCATGATATGT	*rpl22*
AATTACTAATTTCTATTACTATGTTCAATTGAACATAGTAATAGAAATTAGTAATT	AATTACTAATTTCTATTACT t TGTTCAATTGAACATAGTAATAGAAATTAGTAATT	*atpH-atpI*
TAGTTAGAATTCTAACTA	TAGTT c GAATTCTAACTA	*trnT-UGU-trnL-UAA*
TATTTTTTCTAGAAAAAATA	TATTTTTTCTAGAA g AAATA	*ycf2* 2 copies in IR
in L. octovalvis and L. peploides	*in L. grandiflora*	
CCCATCAATCATGATTG t TGGG	CCCATCAATCATGATTGATGGG	*psbN-trnD-GUC*
in L. octovalvis and L. grandiflora	*in L. peploides*	
ATGAAAAAAATCGATTTTTTTCAT	ATGATAAAAATAGATTTTT a TCAT	*trnK-UUU-rps16*
ATGAAAAAAATCGATTTTTTTCAT- ATGATAAAAATCGATTTTTATCAT	ATGATAAAAATA g ATTTTTATCAT	*trnK-UUU-rps16*

Unique palidromic repeats		
L. peploides		
TTATATATATATATATATAA	*rpl32-ndhF*	Full deletion in *L. octovalvis*, 6 bases deletion in *L. grandiflora*
L. octovalvis		
ATTGAAATTCGAATTTCAAT	*psbZ-trnG-GCC*	Full deletion in *L. grandiflora and L. peploides*
L. peploides and L. grandiflora		
AAAAAATGGATCCATTTTTT	*trnL-UAG-rpl32*	3 bases deleted and 3 bases mutated in *L. octovalvis*
AATATATTATTATAATAATATATT	*rpl32-ndhF*	Full deletion in *L. octovalvis*
TATATTTATTATTAATTAATAATAAATATA	*rpl32-ndhF*	Full deletion in *L. octovalvis*

Lo = Ludwigia octovalvis; Lgh = L. grandiflora subsp. *hexapetala; Lpm = L. peploides* subsp. *montevidensis.*

712

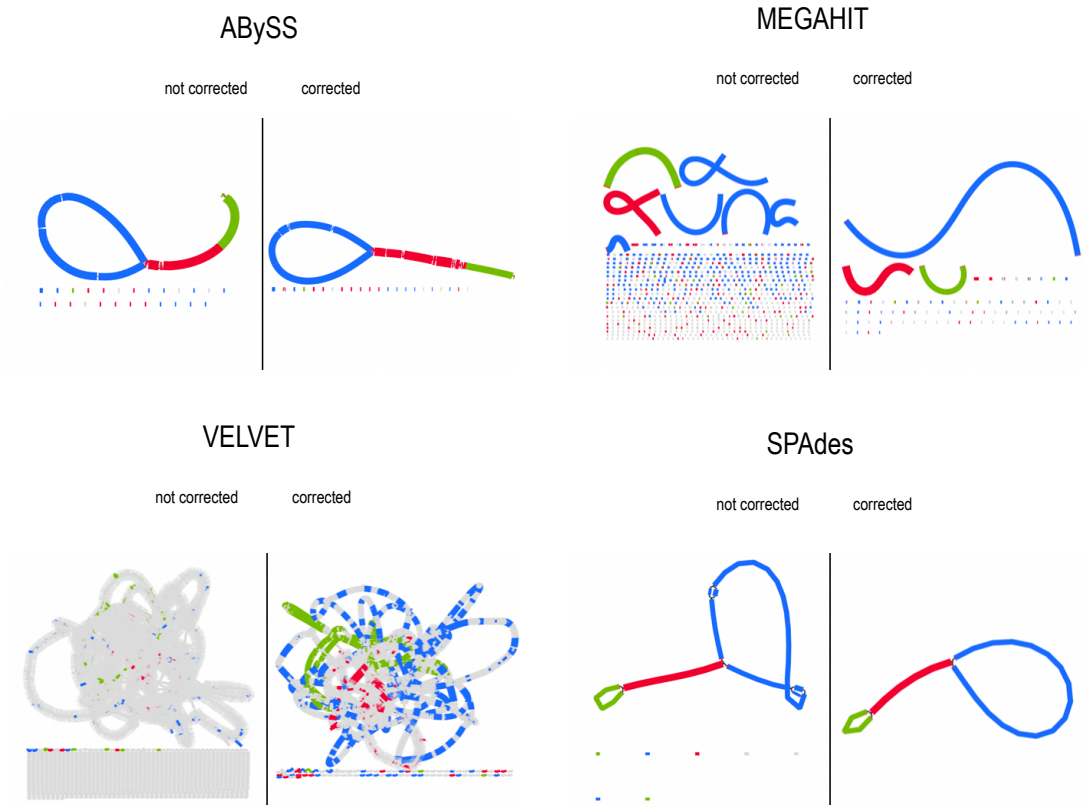
A

	ABySS		MEGAHIT		VELVET		SPAdes	
	not corrected	corrected	not corrected	corrected	not corrected	corrected	not corrected	corrected
Using all contigs								
Genome fraction (%)	86.868	85.279	86.428	85.158	91.927	86.796	84.682	84.483
Duplication ratio	1.047	1.042	1.796	1.041	2.002	1.128	1.042	1
Largest alignment	56 588	41 262	30 904	90 352	3531	17 235	90 399	90 272
Using contigs > 200 nt								
Genome fraction (%)	86.419	85.279	86.377	85.057	76.589	86.181	84.682	84.483
Duplication ratio	1.028	1.042	1.681	1.029	1.177	1.11	1.042	1
Largest alignment	56 588	41 262	30 904	90 352	3531	17 235	90 399	90 272
Using contigs > 500 nt								
Genome fraction (%)	85.564	84.517	83.503	84.774	45.468	79.279	84.682	84.483
Duplication ratio	1.009	1.012	1.041	1.004	1.015	1.054	1.042	1
Largest alignment	56 588	41 262	30 904	90 352	3531	17 235	90 399	90 272
Using contigs > 1000 nt								
Genome fraction (%)	83.701	84.199	81.256	84.545	22.194	66.438	84.563	84.483
Duplication ratio	1	1.002	1.007	1.001	1	1.011	1.026	1
Largest alignment	56 588	41 262	30 904	90 352	3531	17 235	90 399	90 272

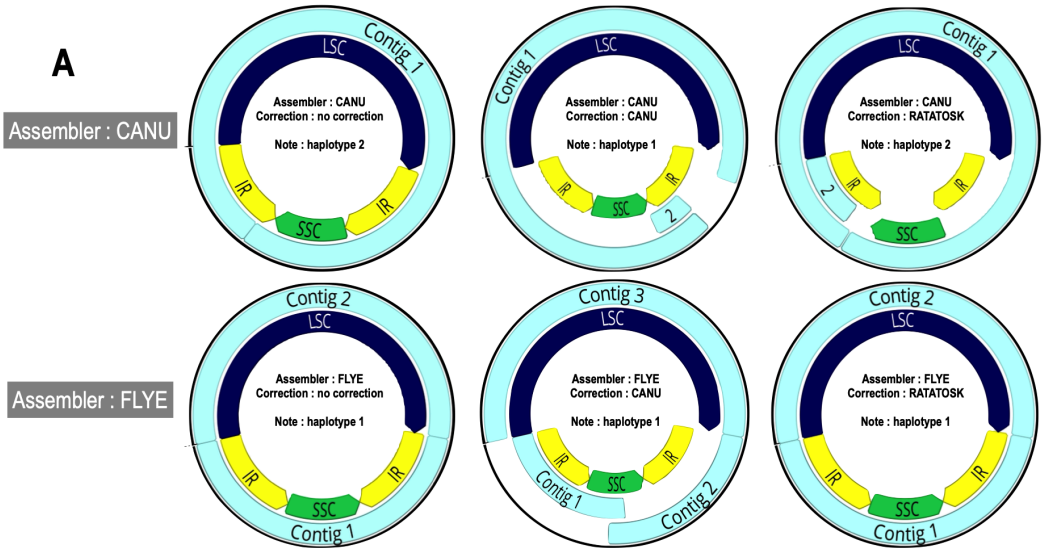
B

	ABySS		MEGAHIT		VELVET		SPAdes	
	not corrected	corrected	not corrected	corrected	not corrected	corrected	not corrected	corrected
Using all contigs								
NGA50	15 215	26 577	19 986	90 352	469	2796	90 399	90 272
LGA50	3	3	3	1	93	9	1	1
Misassemblies								
# misassemblies	0	0	4	0	0	0	0	0
Misassembled contigs length	0	0	1595	0	0	0	0	0
Mismatches								
# mismatches per 100 kbp	109.53	107.19	1036.93	45.24	499.16	229.11	96.57	0
# indels per 100 kbp	12.4	10.58	62.99	16.26	27.92	74.88	19.17	0
# N's per 100 kbp	0	0	0	0	0	6.1	0	0
Using contigs > 500 nt								
NGA50	15 215	26 577	19 986	90 352	–	2796	90 399	90 272
LGA50	3	3	3	1	–	9	1	1
Misassemblies								
# misassemblies	0	0	1	0	0	0	0	0
Misassembled contigs length	0	0	665	0	0	0	0	0
Mismatches								
# mismatches per 100 kbp	62.39	46.17	123.32	8.1	221.33	148.48	96.57	0
# indels per 100 kbp	2.9	2.2	4.33	1.47	28.51	63.74	19.17	0
# N's per 100 kbp	0	0	0	0	0	7.22	0	0
Using contigs > 1000 nt								
NGA50	15 215	26 577	19 986	90 352	–	2796	90 399	90 272
LGA50	3	3	3	1	–	9	1	1
Misassemblies								
# misassemblies	0	0	0	0	0	0	0	0
Misassembled contigs length	0	0	0	0	0	0	0	0
Mismatches								
# mismatches per 100 kbp	0	0	37.51	0.74	64.94	61.6	67.87	0
# indels per 100 kbp	1.5	0.74	0	0.74	25.41	56.93	19.49	0
# N's per 100 kbp	0	0	0	0	0	9.25	0	0

713
714 **Supp. Figure 1:** QUAST evaluation of performance of the four assembly tools (using corrected or
715 uncorrected SRs). **A:** Comparison of plastome fraction, duplication rate and size of the largest alignment
716 obtained. **B:** Comparison of classic metrics (NGA50 and LGA50), number of errors (misassemblies and
717 mismatches) produced.
718



Supp. Figure 2: BANDAGE visualization of the *L. grandiflora* plastome assembly graphs on corrected or uncorrected SRs. Contigs are colored according to their BLAST match to the LSC (blue), SSC (green), and IR (red) segments



B

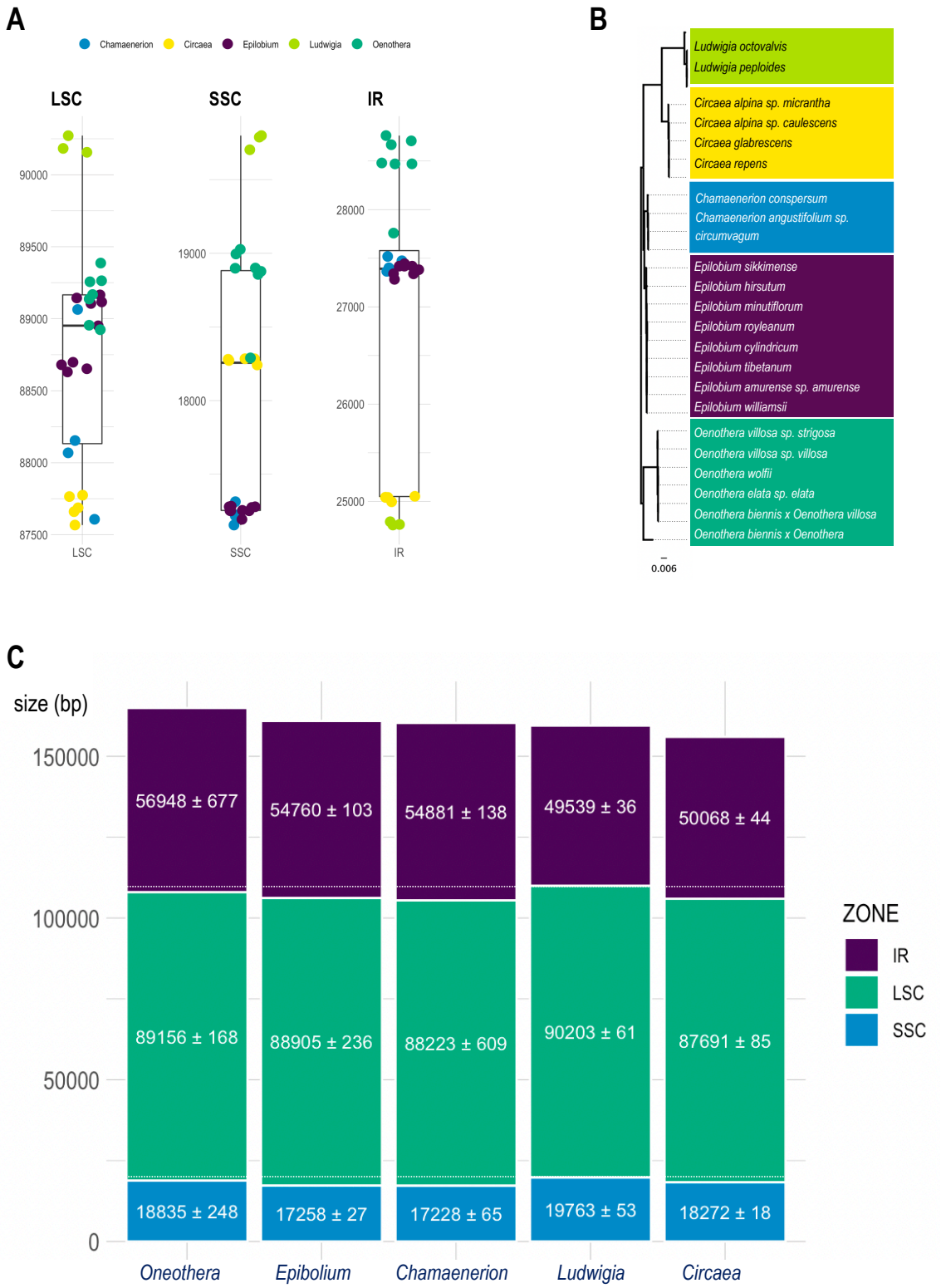
Assembler : CANU

	LSC	IR	SSC
No correction	99.793 (187 nt)	99.960 (30 nt)	99.778 (27 nt)
Canu-correction	99.811 (171 nt)	99.899 (25 nt)	99.828 (34 nt)
Ratatosk-correction	99.952 (43 nt)	99.879 (10 nt)	99.864 (27 nt)

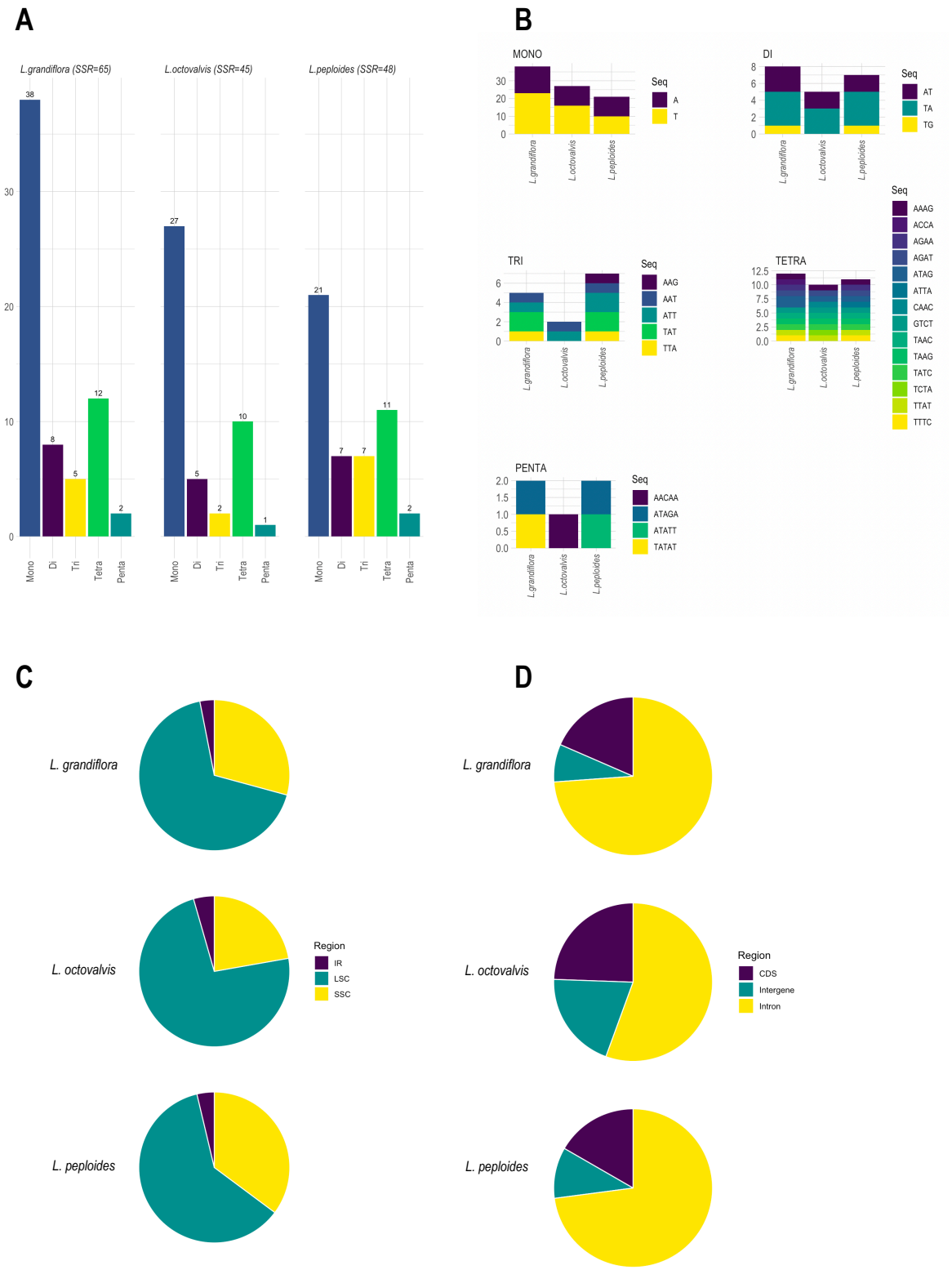
Assembler : FLYE

	LSC	IR	SSC
No correction	99.798 (182 nt)	99.903 (24 nt)	99.783 (50 nt)
Canu-correction	99.831 (153 nt)	99.907 (23 nt)	99.843 (31 nt)
Ratatosk-correction	99.975(23 nt)	99.988 (3 nt)	99.995 (3 nt)

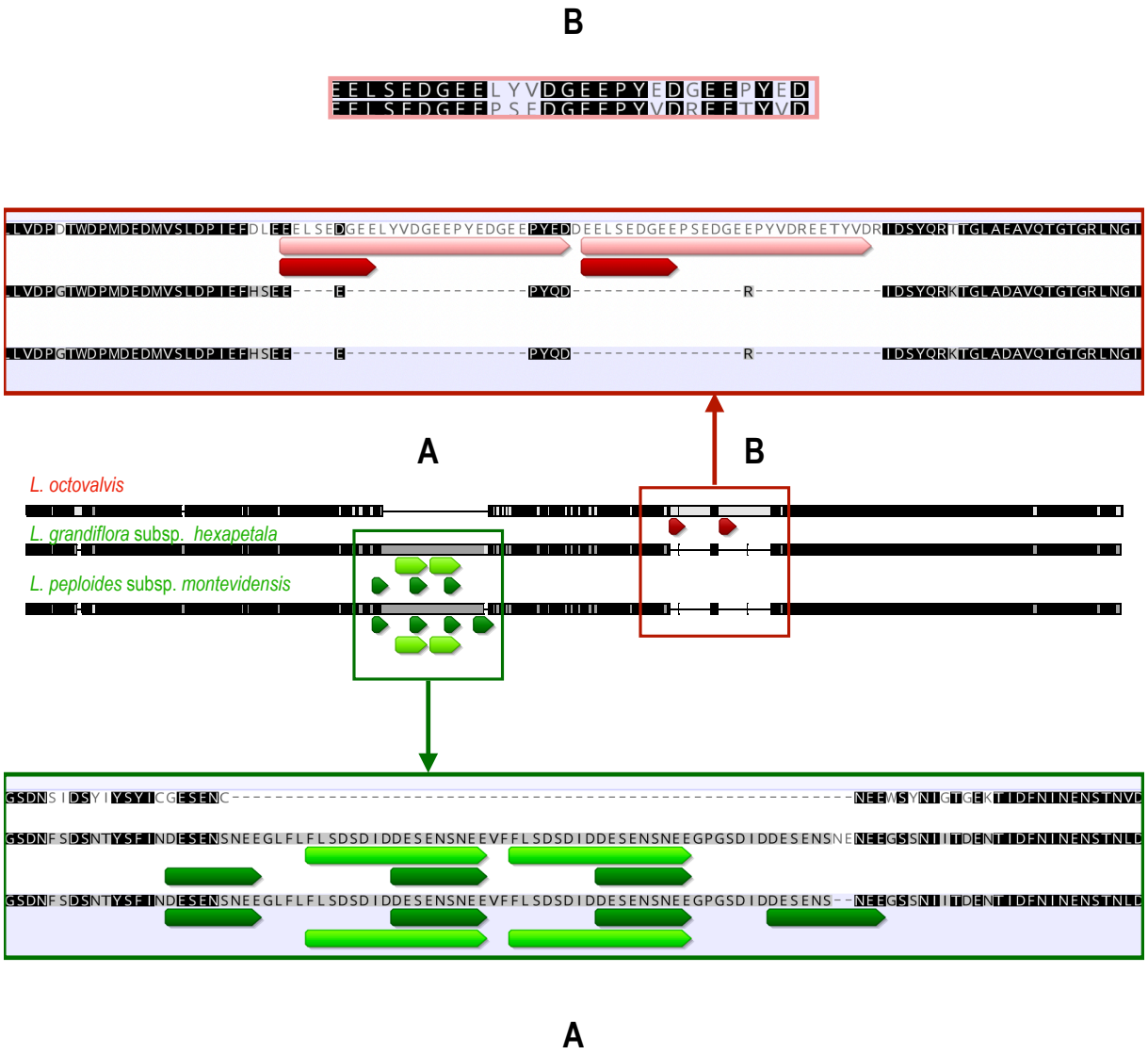
Supp. Figure 3: Graphs representing the assemblies of *L. grandiflora* long reads. **A:** Contigs are represented in light blue and the three segments (LSC, SSC and IR) in dark blue, green and yellow, respectively. B: Comparative effectiveness of CANU and RATATOSK correctors.



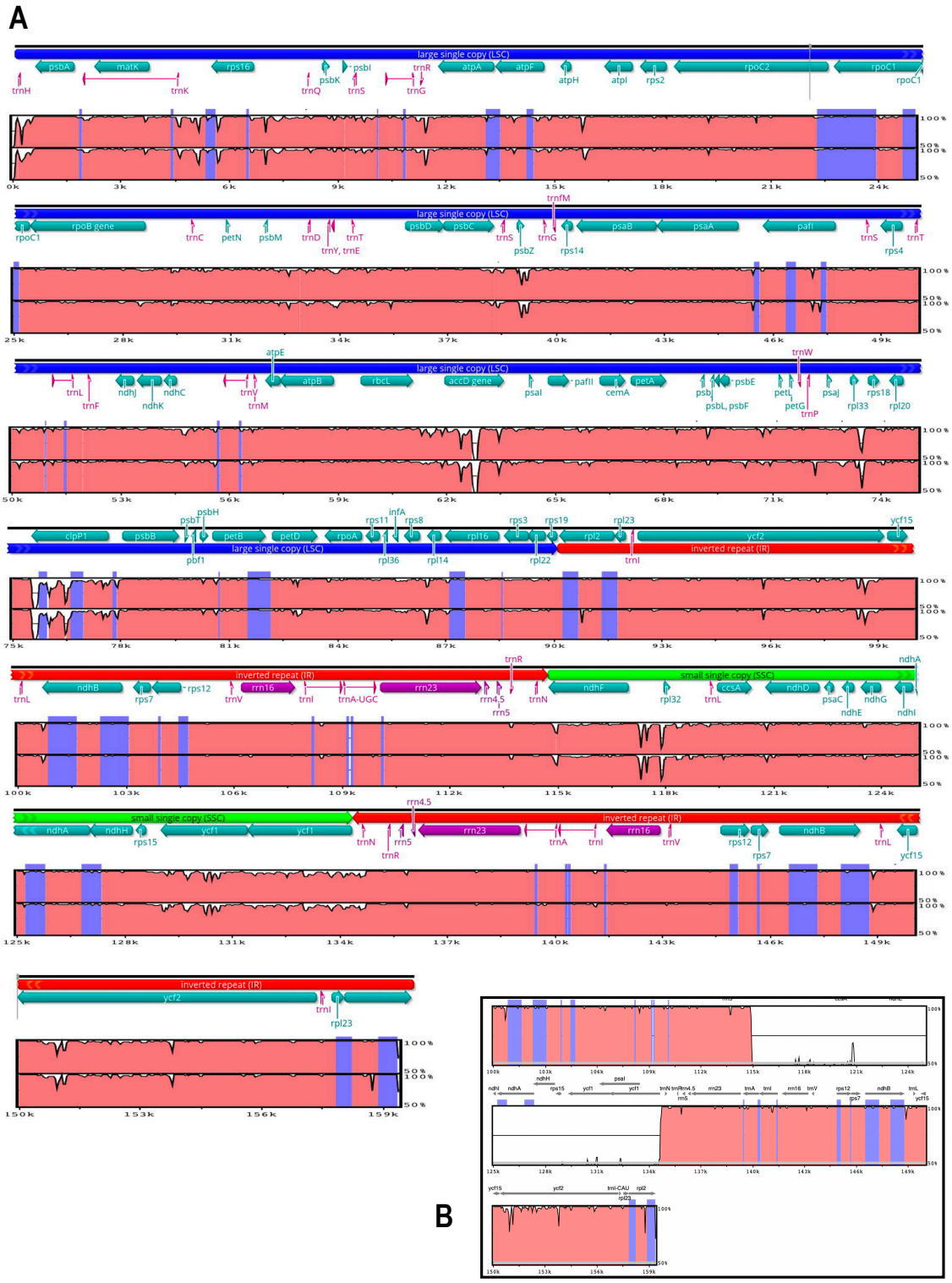
Supp. Figure 4: Comparison of LSC, SSC and IR sizes in the Onagraceae. **A:** Comparison of the sizes of LSC, SSC and IR segments in the Onograceae family (*Chamaenerion* in blue, *Circaea* in yellow, *Epibolium* in dark purple, *Ludwigia* in light green and *Oenothera* in dark green). **B:** Maximum likelihood tree made using RAxML (model GTR-GAMMA, algorithm Rapid Hill-climbing) on multiple sequences alignment of Onograceae plastomes made using MAFFT. **C:** Average size of the different chloroplast segments (LSC, SSC and IR) for the 5 genres of Onograceae. IR size corresponds to the sum of the two copies.



Supp. Figure 5: Comparative analysis of Simple-Sequence Repeats (SSRs) in *Ludwigia* chloroplast genomes. A: SSR numbers detected in the three species, by repeat class types (mono, di-, tri-, tetra and pentanucleotides). **B:** Frequency of SSR motifs by repeat class types. **C:** Frequency of SSRs in LSC, SSC and IR regions. **D:** Repartition of SSRs in intergenic, protein-coding and intronic regions.

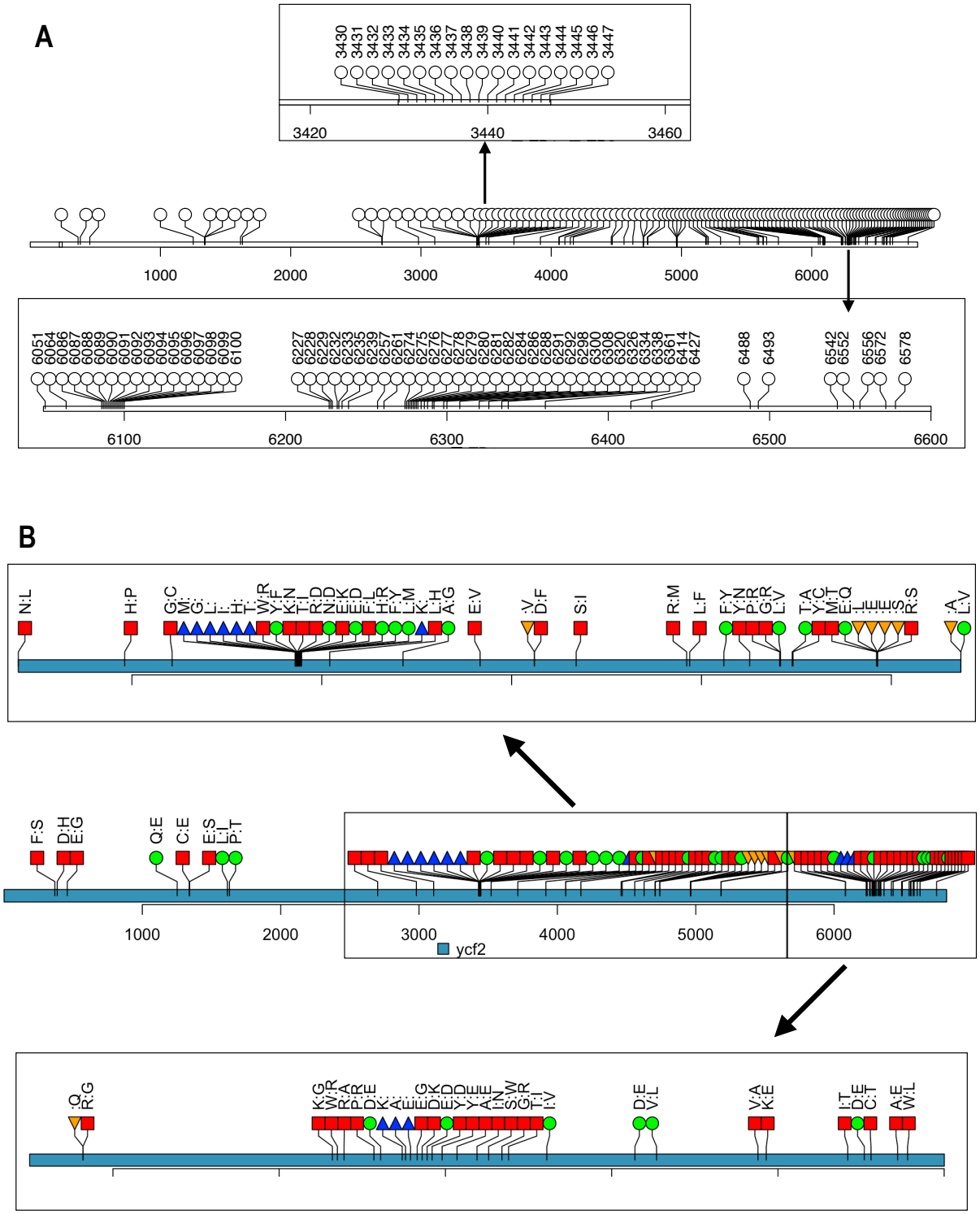


Supp. Figure 6: Diagram showing the position of tandem repeats in the *accD* gene. *L. octovalis* (in red) and *L. peploides* and *L. grandiflora* (in green). We also observe the consequences of these repetitions on the insertion of amino acids, also repeated.



Supp. Figure 7: Comparison of the three *Ludwigia* plastomes using mVISTA, with the *L. octovalvis* as a reference. **A:** The y-axis represents the identity percentage (between 50 and 100%). The arrows show the genes (in green: proteins genes, in purple: rRNAs and in fuchsia: tRNAs). Blue blocks indicate exonic regions. LCS, IR and SSC regions are also distinguished (in dark blue, red and green, respectively). The second line corresponds to *L. grandiflora* haplotype 2 (For this haplotype, SSC segment is oriented like *L. octovalvis*) and the third line corresponds to *L. peploides* for which the SSC region has been artificially oriented in the same way as the two other plastomes to allow comparison. **B:** Small box showing a part of

762 the alignment and presenting the consequences if we do not artificially orient the SSC segments in the same
763 direction for the analysis.
764



Supp. Figure 8: Lollipop diagram allowing the visualization of SNPs and their translational effects on the *ycf2*. **A:** localization of the 256 single nucleotide polymorphisms (SNP) observed by comparing *L. grandiflora-L. peploides* with *L. octovalvis.* Two regions particularly dense in SNPs (between 3420 and 3460 and between 6100 and 6600) have been zoomed into to allow better reading. **B:** Effect of these SNPs on the translated sequence of *L. octovalvis*, compared to Ycf2 of the other two species: non conservative mutation: red square; conservative mutation: circle green; deletion: triangle_point_up blue and insertion: triangle_point_down, orange. As for A, two regions were zoomed into in order to distinguish each mutation.

773

774 **References**

775 1. Wagner WL, Hoch PC, Raven PH: **Revised classification of the Onagraceae**.
776 *Systematic Botany Monographs* 2007.

777 2. Levin RA, Wagner WL, Hoch PC, Nepokroeff M, Pires JC, Zimmer EA, Sytsma KJ:
778 **Family-level relationships of Onagraceae based on chloroplast rbcL and ndhF**
779 **data**. *American Journal of Botany* 2003, **90**(1):107-115.

780 3. Levin RA, Wagner WL, Hoch PC, Hahn WJ, Rodriguez A, Baum DA, Katinas L,
781 Zimmer EA, Sytsma KJ: **Paraphyly in tribe Onagreae: insights into phylogenetic**
782 **relationships of Onagraceae based on nuclear and chloroplast sequence data**.
783 *Systematic Botany* 2004, **29**(1):147-164.

784 4. Liu S-H, Hoch PC, Diazgranados M, Raven PH, Barber JC: **Multi-locus phylogeny of**
785 **Ludwigia (Onagraceae): insights on infra-generic relationships and the current**
786 **classification of the genus**. *Taxon* 2017, **66**(5):1112-1127.

787 5. Hussner A, Windhaus M, Starfinger U: **From weed biology to successful control: an**
788 **example of successful management of Ludwigia grandiflora in Germany**. *Weed*
789 *Research* 2016, **56**(6):434-441.

790 6. Lambert E, Dutartre A, Coudreuse J, Haury J: **Relationships between the biomass**
791 **production of invasive Ludwigia species and physical properties of habitats in**
792 **France**. *Hydrobiologia* 2010, **656**:173-186.

793 7. Haury J, Druel A, Cabral T, Paulet Y, Bozec M, Coudreuse J: **Which adaptations of**
794 **some invasive Ludwigia spp.(Rosidae, Onagraceae) populations occur in**
795 **contrasting hydrological conditions in Western France?** *Hydrobiologia* 2014,
796 **737**:45-56.

797 8. Billet K, Genitoni J, Bozec M, Renault D, Barloy D: **Aquatic and terrestrial**
798 **morphotypes of the aquatic invasive plant, Ludwigia grandiflora, show distinct**
799 **morphological and metabolomic responses**. *Ecol Evol* 2018, **8**(5):2568-2579.
800 9. Marks RA, Hotaling S, Frandsen PB, VanBuren R: **Representation and participation**
801 **across 20 years of plant genome sequencing**. *Nat Plants* 2021, **7**(12):1571-1578.
802 10. Zardini E, Raven PH: **A new section of Ludwigia (Onagraceae) with a key to the**
803 **sections of the genus**. *Systematic botany* 1992:481-485.
804 11. Hoch PC, Wagner WL, Raven PH: **The correct name for a section of Ludwigia**
805 **L.(Onagraceae)**. *PhytoKeys* 2015(50):31.
806 12. Hu Y, Zhang Q, Rao G: **Occurrence of plastids in the sperm cells of Caprifoliaceae:**
807 **biparental plastid inheritance in angiosperms is unilaterally derived from**
808 **maternal inheritance**. *Plant and cell physiology* 2008, **49**(6):958-968.
809 13. Zhang Q: **Why does biparental plastid inheritance revive in angiosperms?** *Journal*
810 *of plant research* 2010, **123**(2):201-206.
811 14. Cleland RE: **Oenothera. Cytogenetics and evolution**. 1972.
812 15. Schmitz UK, Kowallik K-V: **Plastid inheritance in Epilobium**. *Current genetics* 1986,
813 **11**(1):1-5.
814 16. Sato N: **Are Cyanobacteria an Ancestor of Chloroplasts or Just One of the Gene**
815 **Donors for Plants and Algae?** *Genes (Basel)* 2021, **12**(6).
816 17. Gualberto JM, Mileshina D, Wallet C, Niazi AK, Weber-Lotfi F, Dietrich A: **The plant**
817 **mitochondrial genome: dynamics and maintenance**. *Biochimie* 2014, **100**:107-120.
818 18. Tonti-Filippini J, Nevill PG, Dixon K, Small I: **What can we do with 1000 plastid**
819 **genomes?** In., vol. 90: Wiley Online Library; 2017: 808-818.

820 19. Oldenburg DJ, Bendich AJ: **The linear plastid chromosomes of maize: terminal**
821 **sequences, structures, and implications for DNA replication**. *Current genetics* 2016,
822 **62**:431-442.
823 20. Twyford AD, Ness RW: **Strategies for complete plastid genome sequencing**.
824 *Molecular ecology resources* 2017, **17**(5):858-868.
825 21. Wang W, Schalamun M, Morales-Suarez A, Kainer D, Schwessinger B, Lanfear R:
826 **Assembly of chloroplast genomes with long-and short-read data: a comparison of**
827 **approaches using Eucalyptus pauciflora as a test case**. *BMC genomics* 2018, **19**:1-
828 15.
829 22. Wang W, Lanfear R: **Long-reads reveal that the chloroplast genome exists in two**
830 **distinct versions in most plants**. *Genome Biology and Evolution* 2019, **11**(12):3372-
831 3381.
832 23. Ferrarini M, Moretto M, Ward JA, Šurbanovski N, Stevanović V, Giongo L, Viola R,
833 Cavalieri D, Velasco R, Cestaro A: **An evaluation of the PacBio RS platform for**
834 **sequencing and de novo assembly of a chloroplast genome**. *BMC genomics* 2013,
835 **14**(1):1-12.
836 24. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD,
837 Dilthey AT, Fiddes IT: **Nanopore sequencing and assembly of a human genome with**
838 **ultra-long reads**. *Nature biotechnology* 2018, **36**(4):338-345.
839 25. Rang FJ, Kloosterman WP, de Ridder J: **From squiggle to basepair: computational**
840 **approaches for improving nanopore sequencing read accuracy**. *Genome biology*
841 2018, **19**(1):90.
842 26. Scheunert A, Dorfner M, Lingl T, Oberprieler C: **Can we use it? On the utility of de**
843 **novo and reference-based assembly of Nanopore data for plant plastome**
844 **sequencing**. *PLoS One* 2020, **15**(3):e0226234.

845 27. Bedoya AM, Madriñán S: **Evolution of the aquatic habit in Ludwigia (Onagraceae):**
846 **Morpho-anatomical adaptive strategies in the Neotropics**. *Aquatic Botany* 2015,
847 **120**:352-362.
848 28. Liu S-H, Yang H-A, Kono Y, Hoch PC, Barber JC, Peng C-I, Chung K-F:
849 **Disentangling Reticulate Evolution of North Temperate Haplostemonous**
850 **Ludwigia (Onagraceae) 1, 2**. *Annals of the Missouri Botanical Garden* 2020,
851 **105**(2):163-182.
852 29. Panova M, Aronsson H, Cameron RA, Dahl P, Godhe A, Lind U, Ortega-Martinez O,
853 Pereyra R, Tesson SV, Wrange A-L: **DNA extraction protocols for whole-genome**
854 **sequencing in marine organisms**. *Marine genomics: Methods and protocols* 2016:13-
855 44.
856 30. Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, Genete M, Berrabah
857 W, Chèvre A-M, Delourme R: **Chromosome-scale assemblies of plant genomes using**
858 **nanopore long reads and optical maps**. *Nature plants* 2018, **4**(11):879-887.
859 31. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor**.
860 *Bioinformatics* 2018, **34**(17):i884-i890.
861 32. Jin J-J, Yu W-B, Yang J-B, Song Y, DePamphilis CW, Yi T-S, Li D-Z: **GetOrganelle:**
862 **a fast and versatile toolkit for accurate de novo assembly of organelle genomes**.
863 *Genome biology* 2020, **21**:1-31.
864 33. Dierckxsens N, Mardulyn P, Smits G: **NOVOPlasty: de novo assembly of organelle**
865 **genomes from whole genome data**. *Nucleic acids research* 2017, **45**(4):e18-e18.
866 34. Nikolenko SI, Korobeynikov AI, Alekseyev MA: **BayesHammer: Bayesian clustering**
867 **for error correction in single-cell sequencing**. In: *BMC genomics: 2013*. Springer: 1-
868 11.

869 35. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G,
870 Khan H, Coombe L, Warren RL: **ABySS 2.0: resource-efficient assembly of large**
871 **genomes using a Bloom filter**. *Genome research* 2017, **27**(5):768-777.
872 36. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel**
873 **assembler for short read sequence data**. *Genome research* 2009, **19**(6):1117-1123.
874 37. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W:
875 **MEGAHIT v1. 0: a fast and scalable metagenome assembler driven by advanced**
876 **methodologies and community practices**. *Methods* 2016, **102**:3-11.
877 38. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
878 Nikolenko SI, Pham S, Prjibelski AD: **SPAdes: a new genome assembly algorithm**
879 **and its applications to single-cell sequencing**. *Journal of computational biology* 2012,
880 **19**(5):455-477.
881 39. Chikhi R, Medvedev P: **Informed and automated k-mer size selection for genome**
882 **assembly**. *Bioinformatics* 2014, **30**(1):31-37.
883 40. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable**
884 **and accurate long-read assembly via adaptive k-mer weighting and repeat**
885 **separation**. *Genome research* 2017, **27**(5):722-736.
886 41. Kolmogorov M, Yuan J, Lin Y, Pevzner PA: **Assembly of long, error-prone reads**
887 **using repeat graphs**. *Nature biotechnology* 2019, **37**(5):540-546.
888 42. Holley G, Beyter D, Ingimundardottir H, Møller PL, Kristmundsdottir S, Eggertsson
889 HP, Halldorsson BV: **Ratatosk: hybrid error correction of long reads enables**
890 **accurate variant calling and assembly**. *Genome Biology* 2021, **22**(1):1-22.
891 43. Wick RR, Schultz MB, Zobel J, Holt KE: **Bandage: interactive visualization of de**
892 **novo genome assemblies**. *Bioinformatics* 2015, **31**(20):3350-3352.

893 44. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S:
894 **GeSeq–versatile and accurate annotation of organelle genomes**. *Nucleic acids*
895 *research* 2017, **45**(W1):W6-W11.
896 45. Greiner S, Lehwark P, Bock R: **OrganellarGenomeDRAW (OGDRAW) version 1.3.**
897 **1: expanded toolkit for the graphical visualization of organellar genomes**. *Nucleic*
898 *acids research* 2019, **47**(W1):W59-W64.
899 46. Lehwark P, Greiner S: **GB2sequin-A file converter preparing custom GenBank files**
900 **for database submission**. *Genomics* 2019, **111**(4):759-761.
901 47. Beier S, Thiel T, Münch T, Scholz U, Mascher M: **MISA-web: a web server for**
902 **microsatellite prediction**. *Bioinformatics* 2017, **33**(16):2583-2585.
903 48. Benson G: **Tandem repeats finder: a program to analyze DNA sequences**. *Nucleic*
904 *acids research* 1999, **27**(2):573-580.
905 49. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: **VISTA: computational**
906 **tools for comparative genomics**. *Nucleic acids research* 2004, **32**(suppl_2):W273-
907 W279.
908 50. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou
909 S, Program NCS: **LAGAN and Multi-LAGAN: efficient tools for large-scale**
910 **multiple alignment of genomic DNA**. *Genome research* 2003, **13**(4):721-731.
911 51. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-
912 Onsins SE, Sánchez-Gracia A: **DnaSP 6: DNA sequence polymorphism analysis of**
913 **large data sets**. *Molecular biology and evolution* 2017, **34**(12):3299-3302.
914 52. Rozas J, Rozas R: **DnaSP version 3: an integrated program for molecular**
915 **population genetics and molecular evolution analysis**. *Bioinformatics (Oxford,*
916 *England)* 1999, **15**(2):174-175.

917 53. Amiryousefi A, Hyvönen J, Poczai P: **IRscope: an online program to visualize the**
918 **junction sites of chloroplast genomes**. *Bioinformatics* 2018, **34**(17):3030-3031.
919 54. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R: **TBtools: an**
920 **integrative toolkit developed for interactive analyses of big biological data**.
921 *Molecular plant* 2020, **13**(8):1194-1202.
922 55. Katoh K, Misawa K, Kuma Ki, Miyata T: **MAFFT: a novel method for rapid multiple**
923 **sequence alignment based on fast Fourier transform**. *Nucleic acids research* 2002,
924 **30**(14):3059-3066.
925 56. Stamatakis A: **The RAxML v8. 2. X Manual**. *Heidleberg Institute for Theoretical*
926 *Studies*.
927 57. Konishi T, Sasaki Y: **Compartmentalization of two forms of acetyl-CoA carboxylase**
928 **in plants and the origin of their tolerance toward herbicides**. *Proc Natl Acad Sci U*
929 *S A* 1994, **91**(9):3598-3601.
930 58. Wu S, Chen J, Li Y, Liu A, Li A, Yin M, Shrestha N, Liu J, Ren G: **Extensive genomic**
931 **rearrangements mediated by repetitive sequences in plastomes of Medicago and**
932 **its relatives**. *BMC Plant Biol* 2021, **21**(1):421.
933 59. Li J, Su Y, Wang T: **The Repeat Sequences and Elevated Substitution Rates of the**
934 **Chloroplast accD Gene in Cupressophytes**. *Front Plant Sci* 2018, **9**:533.
935 60. Gurdon C, Maliga P: **Two distinct plastid genome configurations and**
936 **unprecedented intraspecies length variation in the accD coding region in Medicago**
937 **truncatula**. *DNA Res* 2014, **21**(4):417-427.
938 61. Richardson AO, Palmer JD: **Horizontal gene transfer in plants**. *J Exp Bot* 2007,
939 **58**(1):1-9.
940 62. de Vries J, Sousa FL, Bolter B, Soll J, Gould SB: **YCF1: A Green TIC?** *Plant Cell*
941 2015, **27**(7):1827-1833.

942 63. Filip E, Skuza L: **Horizontal Gene Transfer Involving Chloroplasts**. *Int J Mol Sci*
943 2021, **22**(9).
944 64. Zhong Q, Yang S, Sun X, Wang L, Li Y: **The complete chloroplast genome of the**
945 **Jerusalem artichoke (Helianthus tuberosus L.) and an adaptive evolutionary**
946 **analysis of the ycf2 gene**. *PeerJ* 2019, **7**:e7596.
947 65. Antil S, Abraham JS, Sripoorna S, Maurya S, Dagar J, Makhija S, Bhagat P, Gupta R,
948 Sood U, Lal R *et al*: **DNA barcoding, an effective tool for species identification: a**
949 **review**. *Mol Biol Rep* 2023, **50**(1):761-775.
950 66. Li J, Wang Y, Cui J, Wang W, Liu X, Chang Y, Yao D, Cui J: **Removal effects of**
951 **aquatic plants on high-concentration phosphorus in wastewater during summer**. *J*
952 *Environ Manage* 2022, **324**:116434.
953 67. Soliman AT, Hamdy RS, Hamed AB: **Ludwigia stolonifera (Guill. & Perr.) PH**
954 **Raven, insight into its phenotypic plasticity, habitat diversity and associated**
955 **species**. *Egyptian Journal of Botany* 2018, **58**(3):605-626.
956 68. Kamoshita A, Ikeda H, Yamagishi J, Lor B, Ouk M: **Residual effects of cultivation**
957 **methods on weed seed banks and weeds in Cambodia**. *Weed Biology and*
958 *Management* 2016, **16**(3):93-107.
959 69. Anita VPD, Matra DD, Siregar UJ: **Chloroplast genome draft assembly of Falcataria**
960 **moluccana using hybrid sequencing technology**. *BMC Res Notes* 2023, **16**(1):31.
961 70. Xu S, Teng K, Zhang H, Gao K, Wu J, Duan L, Yue Y, Fan X: **Chloroplast genomes**
962 **of four Carex species: Long repetitive sequences trigger dramatic changes in**
963 **chloroplast genome structure**. *Front Plant Sci* 2023, **14**:1100876.
964 71. Guo YY, Yang JX, Li HK, Zhao HS: **Chloroplast Genomes of Two Species of**
965 **Cypripedium: Expanded Genome Size and Proliferation of AT-Biased Repeat**
966 **Sequences**. *Front Plant Sci* 2021, **12**:609729.

967 72. Liu S-H, Edwards C, Hoch PC, Raven PH, Barber JC: **Complete plastome sequence**
968 **of Ludwigia octovalvis (Onagraceae), a globally distributed wetland plant**. *Genome*
969 *announcements* 2016, **4**(6):e01274-01216.
970 73. Bateman RM, Rudall PJ, Murphy ARM, Cowan RS, Devey DS, Perez-Escobar OA:
971 **Whole plastomes are not enough: phylogenomic and morphometric exploration at**
972 **multiple demographic levels of the bee orchid clade Ophrys sect. Sphegodes**. *J Exp*
973 *Bot* 2021, **72**(2):654-681.
974 74. Lin Z, Zhou P, Ma X, Deng Y, Liao Z, Li R, Ming R: **Comparative analysis of**
975 **chloroplast genomes in Vasconcellea pubescens A.DC. and Carica papaya L**. *Sci*
976 *Rep* 2020, **10**(1):15799.
977 75. Lihodeevskiy GA, Shanina EP: **The Use of Long-Read Sequencing to Study the**
978 **Phylogenetic Diversity of the Potato Varieties Plastome of the Ural Selection**.
979 *Agronomy* 2022, **12**(4):846.
980 76. Nath O, Fletcher SJ, Hayward A, Shaw LM, Masouleh AK, Furtado A, Henry RJ, Mitter
981 N: **A haplotype resolved chromosomal level avocado genome allows analysis of**
982 **novel avocado genes**. *Hortic Res* 2022, **9**:uhac157.
983 77. Wanichthanarak K, Nookaew I, Pasookhush P, Wongsurawat T, Jenjaroenpun P,
984 Leeratsuwan N, Wattanachaisaereekul S, Visessanguan W, Sirivatanauksorn Y,
985 Nuntasaen N: **Revisiting chloroplast genomic landscape and annotation towards**
986 **comparative chloroplast genomes of Rhamnaceae**. *BMC Plant Biology* 2023,
987 **23**(1):59.
988 78. Luo Y, He J, Lyu R, Xiao J, Li W, Yao M, Pei L, Cheng J, Li J, Xie L: **Comparative**
989 **Analysis of Complete Chloroplast Genomes of 13 Species in Epilobium, Circaea,**
990 **and Chamaenerion and Insights Into Phylogenetic Relationships of Onagraceae**.
991 *Front Genet* 2021, **12**:730495.

992 79. Zhang XF, Landis JB, Wang HX, Zhu ZX, Wang HF: **Comparative analysis of**
993 **chloroplast genome structure and molecular dating in Myrtales**. *BMC Plant Biol*
994 2021, **21**(1):219.
995 80. Xu J, Shen X, Liao B, Xu J, Hou D: **Comparing and phylogenetic analysis**
996 **chloroplast genome of three Achyranthes species**. *Sci Rep* 2020, **10**(1):10818.
997 81. Lian C, Yang H, Lan J, Zhang X, Zhang F, Yang J, Chen S: **Comparative analysis of**
998 **chloroplast genomes reveals phylogenetic relationships and intraspecific variation**
999 **in the medicinal plant Isodon rubescens**. *PLoS One* 2022, **17**(4):e0266546.
1000 82. Mohanta TK, Mishra AK, Khan A, Hashem A, Abd Allah EF, Al-Harrasi A: **Gene Loss**
1001 **and Evolution of the Plastome**. *Genes (Basel)* 2020, **11**(10).
1002 83. Barthet MM, Hilu KW: **Expression of matK: functional and evolutionary**
1003 **implications**. *American Journal of Botany* 2007, **94**(8):1402-1412.
1004 84. Li L, Liu C, Hou K, Liu W: **Comparative Analyses of Plastomes of Four Anubias**
1005 **(Araceae) Taxa, Tropical Aquatic Plants Endemic to Africa**. *Genes (Basel)* 2022,
1006 **13**(11).
1007 85. Zeb U, Wang X, AzizUllah A, Fiaz S, Khan H, Ullah S, Ali H, Shahzad K:
1008 **Comparative genome sequence and phylogenetic analysis of chloroplast for**
1009 **evolutionary relationship among Pinus species**. *Saudi J Biol Sci* 2022, **29**(3):1618-
1010 1627.
1011 86. Wu Z, Liao R, Yang T, Dong X, Lan D, Qin R, Liu H: **Analysis of six chloroplast**
1012 **genomes provides insight into the evolution of Chrysosplenium (Saxifragaceae)**.
1013 *BMC Genomics* 2020, **21**(1):621.
1014 87. Kode V, Mudd EA, Iamtham S, Day A: **The tobacco plastid accD gene is essential**
1015 **and is required for leaf development**. *Plant J* 2005, **44**(2):237-244.

1016 88. Madoka Y, Tomizawa K, Mizoi J, Nishida I, Nagano Y, Sasaki Y: **Chloroplast**
1017 **transformation with modified accD operon increases acetyl-CoA carboxylase and**
1018 **causes extension of leaf longevity and increase in seed yield in tobacco**. *Plant Cell*
1019 *Physiol* 2002, **43**(12):1518-1525.
1020 89. Gu H, Wang Y, Xie H, Qiu C, Zhang S, Xiao J, Li H, Chen L, Li X, Ding Z: **Drought**
1021 **stress triggers proteomic changes involving lignin, flavonoids and fatty acids in tea**
1022 **plants**. *Sci Rep* 2020, **10**(1):15504.
1023 90. Bharadwaj B, Mishegyan A, Nagalingam S, Guenther A, Joshee N, Sherman SH, Basu
1024 C: **Physiological and genetic responses of lentil (Lens culinaris) under flood stress**.
1025 *Plant Stress* 2023:100130.
1026 91. Kikuchi S, Asakura Y, Imai M, Nakahira Y, Kotani Y, Hashiguchi Y, Nakai Y, Takafuji
1027 K, Bedard J, Hirabayashi-Ishioka Y *et al*: **A Ycf2-FtsHi Heteromeric AAA-ATPase**
1028 **Complex Is Required for Chloroplast Protein Import**. *Plant Cell* 2018, **30**(11):2677-
1029 2703.
1030 92. Schreier TB, Clery A, Schlafli M, Galbier F, Stadler M, Demarsy E, Albertini D, Maier
1031 BA, Kessler F, Hortensteiner S *et al*: **Plastidial NAD-Dependent Malate**
1032 **Dehydrogenase: A Moonlighting Protein Involved in Early Chloroplast**
1033 **Development through Its Interaction with an FtsH12-FtsHi Protease Complex**.
1034 *Plant Cell* 2018, **30**(8):1745-1769.
1035 93. Drescher A, Ruf S, Calsa T, Jr., Carrer H, Bock R: **The two largest chloroplast**
1036 **genome-encoded open reading frames of higher plants are essential genes**. *Plant J*
1037 2000, **22**(2):97-104.
1038 94. Xing J, Pan J, Yi H, Lv K, Gan Q, Wang M, Ge H, Huang X, Huang F, Wang Y *et al*:
1039 **The plastid-encoded protein Orf2971 is required for protein translocation and**
1040 **chloroplast quality control**. *Plant Cell* 2022, **34**(9):3383-3399.

1041 95. Chen J, Zang Y, Shang S, Yang Z, Liang S, Xue S, Wang Y, Tang X: **Chloroplast**
1042 **genomic comparison provides insights into the evolution of seagrasses**. *BMC Plant*
1043 *Biol* 2023, **23**(1):104.
1044 96. Xie Z, Merchant S: **The plastid-encoded ccsA gene is required for heme attachment**
1045 **to chloroplast c-type cytochromes**. *J Biol Chem* 1996, **271**(9):4632-4639.
1046 97. Kranz R, Lill R, Goldman B, Bonnard G, Merchant S: **Molecular mechanisms of**
1047 **cytochrome c biogenesis: three distinct systems**. *Mol Microbiol* 1998, **29**(2):383-396.
1048 98. Maheswari P, Kunhikannan C, Yasodha R: **Chloroplast genome analysis of**
1049 **Angiosperms and phylogenetic relationships among Lamiaceae members with**
1050 **particular reference to teak (Tectona grandis L.f)**. *J Biosci* 2021, **46**.
1051 99. Zhang Y, Du L, Liu A, Chen J, Wu L, Hu W, Zhang W, Kim K, Lee SC, Yang TJ *et al*:
1052 **The Complete Chloroplast Genome Sequences of Five Epimedium Species: Lights**
1053 **into Phylogenetic and Taxonomic Analyses**. *Front Plant Sci* 2016, **7**:306.
1054 100. Huang LS, Sun YQ, Jin Y, Gao Q, Hu XG, Gao FL, Yang XL, Zhu JJ, El-Kassaby YA,
1055 Mao JF: **Development of high transferability cpSSR markers for individual**
1056 **identification and genetic investigation in Cupressaceae species**. *Ecol Evol* 2018,
1057 **8**(10):4967-4977.
1058 101. Leontaritou P, Lamari FN, Papasotiropoulos V, Iatrou G: **Exploration of genetic,**
1059 **morphological and essential oil variation reveals tools for the authentication and**
1060 **breeding of Salvia pomifera subsp. calycina (Sm.) Hayek**. *Phytochemistry* 2021,
1061 **191**:112900.
1062 102. Snoussi M, Riahi L, Ben Romdhane M, Mliki A, Zoghlami N: **Chloroplast DNA**
1063 **Diversity of Tunisian Barley Landraces as Revealed by cpSSRs Molecular**
1064 **Markers and Implication for Conservation Strategies**. *Genet Res (Camb)* 2022,
1065 **2022**:3905957.

1066 103. Song SL, Lim PE, Phang SM, Lee WW, Hong DD, Prathep A: **Development of**
1067 **chloroplast simple sequence repeats (cpSSRs) for the intraspecific study of**
1068 **Gracilaria tenuistipitata (Gracilariales, Rhodophyta) from different populations**.
1069 *BMC Res Notes* 2014, **7**:77.
1070 104. Wheeler GL, Dorman HE, Buchanan A, Challagundla L, Wallace LE: **A review of the**
1071 **prevalence, utility, and caveats of using chloroplast simple sequence repeats for**
1072 **studies of plant biology**. *Appl Plant Sci* 2014, **2**(12).
1073 105. Feng JL, Wu LW, Wang Q, Pan YJ, Li BL, Lin YL, Yao H: **Comparison Analysis**
1074 **Based on Complete Chloroplast Genomes and Insights into Plastid Phylogenomic**
1075 **of Four Iris Species**. *Biomed Res Int* 2022, **2022**:2194021.
1076