Challenges of genomic data generation for non-model complex species

Guillaume Dore, Frédérique Barloy-Hubler, Dominique D. Barloy

Challenges of genomic data generation for non-model complex species

Guillaume DORÉ[1], Frédérique BARLOY-HUBLER[2] and Dominique BARLOY[1]

[1] UMR DECOD (Ecosystem Dynamics and Sustainability) Institut Agro, IFREMER, INRAE, 65 Rue de Saint-Brieuc, 35042, Rennes, France
[2] UMR 6553 ECOBIO, CNRS, Université de Rennes 1, 263 Avenue du Général Leclerc, 35042, Rennes, France

Introduction

Biological invasion is one of the main factor of biodiversity loss [1]. *Ludwigia grandiflora* subsp. *hexapetala* (*Lgh*) (Figure 1) is a common invasive aquatic plant, also able to colonize wet meadows [2]. To understand the (epi)genetic mechanisms involved in this acclimatation, genomic data is necessary. *Lgh* is a non-model species (Myrtales order, Onagraceae family) with no complete genome available. *Lgh* is decaploid (2n=10x=80) [3] with a very large genome (1.4 Gb). In order to generate necessary genomic resources, we chose to first assemble organelles genomes. The plastome was recently assembled as two haplotypes [4]. Obtaining the mitochondrial genome was more challenging as plant mitogenomes are highly variable in sequence and size [5], only conserved at species level, with numerous repeated sequences and possible insertions of chloroplast genes [6]. For these reasons, plant mitogenomes are poorly represented in databases compared to plastomes. In this study, we present the different strategies used to assemble and obtain *Lgh* mitogenome in two circular molecules (M1 and M2) of respectively 544,782 bp and 166,796 bp.

Figure 1: *Ludwigia grandiflora* subsp. *hexapetala* (Credit D. Barloy).

Materials and methods

- DNA long fragments were extracted from *Lgh* buds then sequenced using two types of technology: Oxford Nanopore (long reads = LR) and Illumina Mi-seq (short reads = SR) and corrected using SPAdes [7] for SR (self-correction) or Ratatosk [8] for LR using SR (hybrid) (Figure 2A).
- To *de novo* assemble *Lgh* mitogenome, we compared SR (Megahit [9]), LR (Flye [10]) and hybrid (SPAdes [7]) assembly results using Quality Assessment Tool (QUAST) [11] (Figure 2B).
- Using 39 Myrtales mitochondrial CDS (Coding DNA Sequence) as references (Figure 2C), we selected mitochondrial SR and LR (Figure 2D) and combined all assemblies to elongate and circularize (Figure 2E) using the "*de novo* assemble" tool from Geneious 10.0.9 (http://geneious.com).
- Annotations were performed with Geseq [12], BLASTx [13] and Rfam [14]. OGDraw [15] was used for graphic representation of the mitogenome. REPuter [16] identified direct and reverse repeats. Chloroplast sequences insertion were found by BLASTn [13]. Circular repeats maps were made with shinyCircos [17].

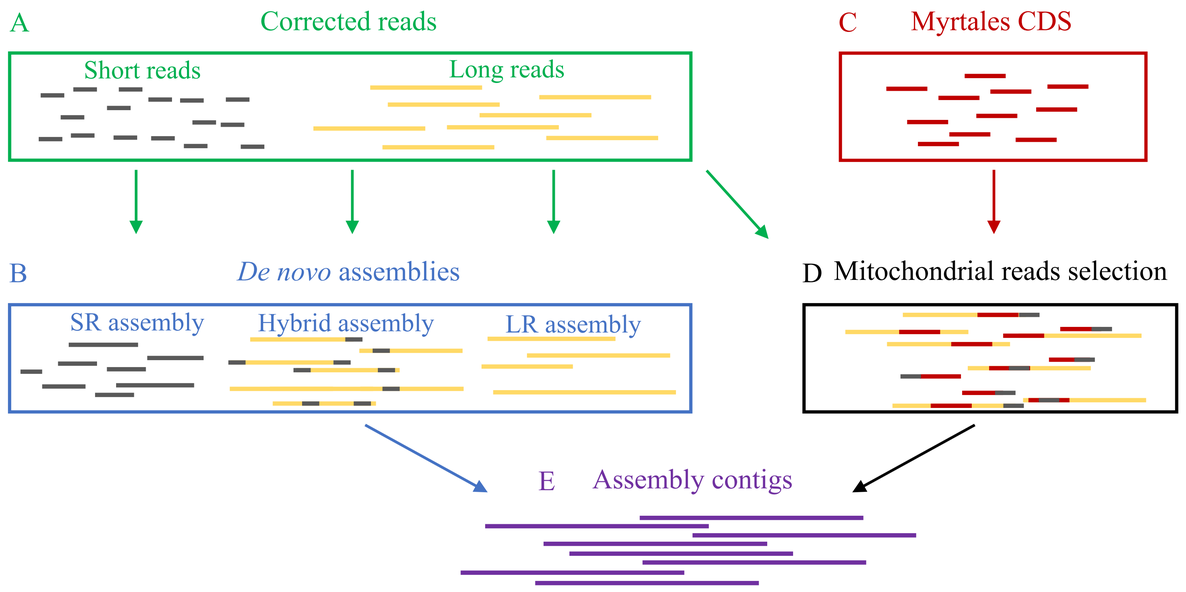

Figure 2: Assembly strategy for *Lgh* mitogenome.

Results and discussion

(1) Assembly strategies were complementary for *Lgh* mitogenome *de novo* reconstruction.

Table 1: Comparison of different assemblies performance with *Lgh* mitogenome as reference using QUAST. For each statistics, blue color display the best score, red the worst.

Genome statistics	Flye	Megahit	Spades
Genome fraction (%)	92.42	89.086	94.942
Duplication ratio	1.005	1.012	1.039
Largest alignment	105 008	38 614	191 601
Total aligned length	659 815	637 933	692 850
NGA50	53 304	15 088	115 267
LGA50	5	15	3
Misassemblies			
# misassemblies	1	1	11
Misassembled contigs length	30 320	4916	575 770
Mismatches			
# mismatches per 100 kbp	69.03	41.33	130.85
# indels per 100 kbp	10.95	4.26	30.49
# N's per 100 kbp	0	0	0
Statistics without reference			
# contigs	37	70	20
Largest contig	105 008	43 938	191 601
Total length	660 613	648 165	806 438
Total length (>= 1000 bp)	650 062	646 177	801 733
Total length (>= 10000 bp)	619 455	414 136	780 610
Total length (>= 50000 bp)	420 228	0	752 738

Interests of different assemblies :

- SPAdes : Best length statistics
- Megahit : Less mistakes generated
- Flye : Less duplication rate

The presence of many repeats as in other plant mitogenomes [18] makes the combination of SR, LR and both sets (hybrid) efficient and needful.

(2) *Lgh* mitogenome is composed of two circular molecules.

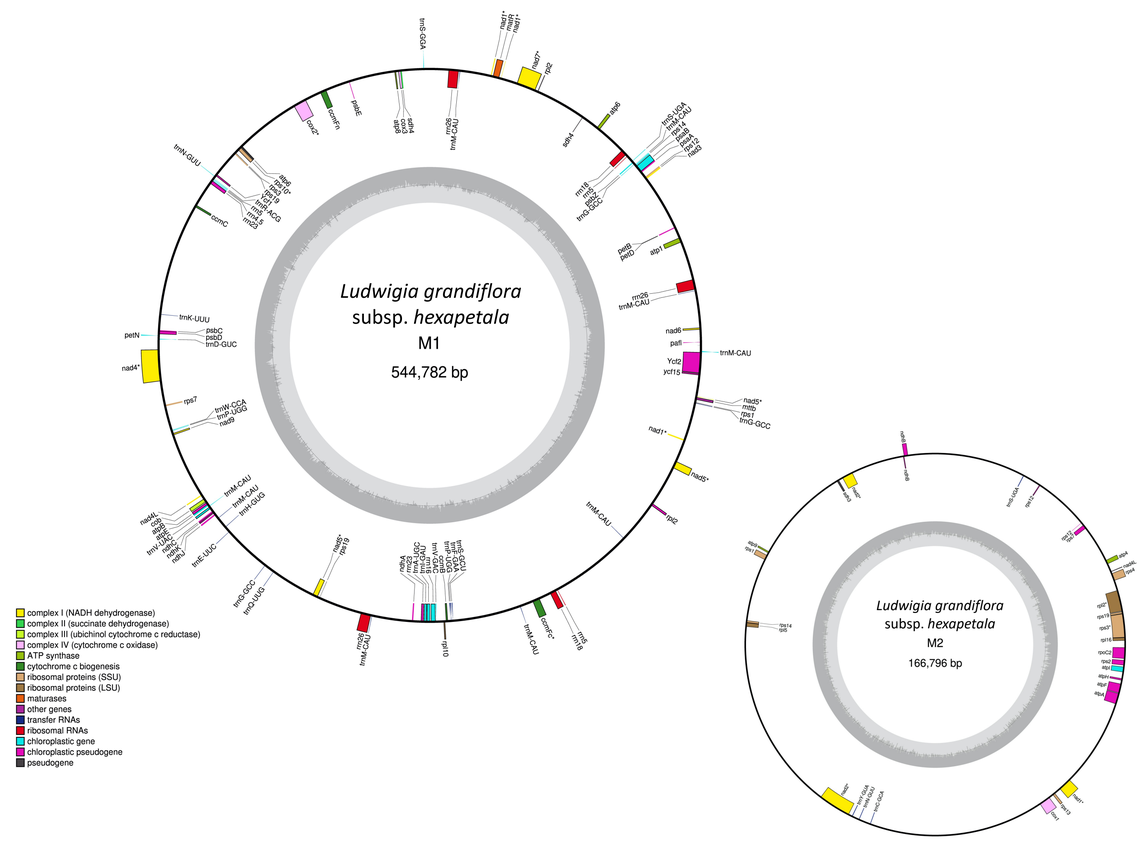

Figure 3: Circular representation of *Ludwigia grandiflora* subsp. *hexapetala* mitochondrial genome using OGDraw. Genes outside the circle are transcribed in counter clockwise direction, genes on the inside are transcribed in clockwise direction. Spliced genes are marked with "*". The grey inner circle shows the GC content. Color chart shows gene family, origin or pseudogenes.

Lgh mitogenome characteristics :

- A complete mitogenome size of 711,578 bp with a M1 molecule size = 544,782 bp (77% of total size) and a M2 molecule size = 166,796 bp (23% of total size)
- Intra-M1 repeats represented 10% of M1 molecule whereas intra-M2 is only 0.2%. Repeats between M1 and M2 represented 0.9% of the mitogenome.

Usually plant mitogenomes are represented as a single circular molecule [19] but multi-chromosomal structures in plants mitogenome have already been discovered in other species like sugarcane [20] or *Populus simonii* [21]. *Lgh* mitogenome is the first described with a multi-molecular structure among seven available Myrtales order species mitogenomes.

(3) Chloroplast sequence insertions represented 6.4% of the mitogenome.

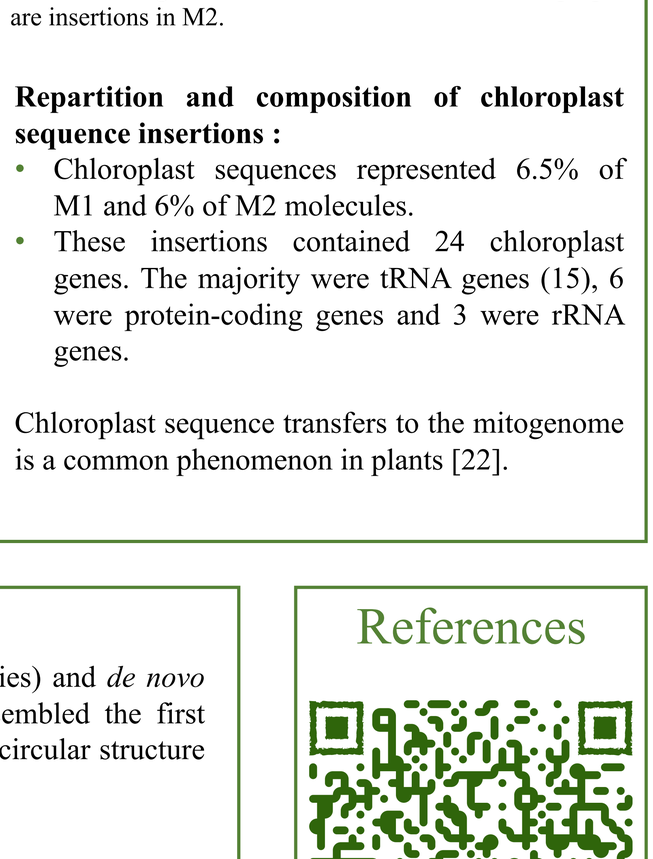

Figure 4: Map of chloroplast sequence insertions (≥ 100 bp) in the *Lgh* mitogenome. Curved lines connect chloroplast insertions in the mitogenome to their origin in the plastome and width is proportional to insertions size. Links in orange are insertions in M1 and links in purple are insertions in M2.

Repartition and composition of chloroplast sequence insertions :

- Chloroplast sequences represented 6.5% of M1 and 6% of M2 molecules.
- These insertions contained 24 chloroplast genes. The majority were tRNA genes (15), 6 were protein-coding genes and 3 were rRNA genes.

Chloroplast sequence transfers to the mitogenome is a common phenomenon in plants [22].

Conclusion

Despite the large size of the mitogenome and the high ploidy of *Lgh*, the combination of *a priori* approach (using CDS from Myrtales order species) and *de novo* assembly approach allowed the successful acquisition of the *Lgh* mitochondrial genome, even in the absence of a reference genome. We assembled the first mitogenome in the *Ludwigia* genus, the fourth in the Onagraceae family and the seventh in the Myrtales order. We highlighted a multi-chromosomal circular structure in *Lgh* mitogenome. We also observed chloroplast gene transfers into the mitogenome.

Perspectives

This highly effective genomic assembly strategy will be used to generate a reference transcriptome and a nuclear genome draft. One of the next step is to verify, with RNA-Seq analysis, if the two molecules are functional (transcribed) and to find if their expression depends of the organ, the development state or the environment. These RNA-Seq analysis will help to estimate the originated chloroplast gene expression and compare it to their expression in the chloroplast. Furthermore, we will try to understand the evolutive history of chloroplast transfer in *Lgh* and in the Myrtales order.

References