



OPEN ACCESS

EDITED BY

Martin Mascher,
Leibniz Institute of Plant Genetics and
Crop Plant Research (IPK), Germany

REVIEWED BY

Pankaj Bhardwaj,
Central University of Punjab, India
Xinmin An,
Beijing Forestry University, China

*CORRESPONDENCE

Shan Gao
gao_shan@mail.nankai.edu.cn
Shuangyong Yan
bioysy@139.com

†These authors have contributed
equally to this work

SPECIALTY SECTION

This article was submitted to
Plant Bioinformatics,
a section of the journal
Frontiers in Plant Science

RECEIVED 26 April 2022

ACCEPTED 12 September 2022

PUBLISHED 29 September 2022

CITATION

Liang J, Chen R, Zhang F, Wang Q,
Yang Y, Lv M, Yan S and Gao S (2022)
Full-length chloroplast genome of
Dongxiang wild rice reveals small
single-copy region switching.
Front. Plant Sci. 13:929352.
doi: 10.3389/fpls.2022.929352

COPYRIGHT

© 2022 Liang, Chen, Zhang, Wang,
Yang, Lv, Yan and Gao. This is an open-
access article distributed under the
terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other
forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the
original publication in this journal is
cited, in accordance with accepted
academic practice. No use,
distribution or reproduction is
permitted which does not comply with
these terms.

Full-length chloroplast genome of Dongxiang wild rice reveals small single-copy region switching

Jianguang Liang^{1†}, Rui Chen^{2,3†}, Fantao Zhang⁴, Qian Wang²,
Yingxia Yang², Mingjie Lv², Shuangyong Yan^{3*} and Shan Gao^{5*}

¹School of Pharmacy, Changzhou University, Changzhou, China, ²Institute of Crop Germplasm and Biotechnology, Tianjin Academy of Agricultural Sciences, Tianjin, China, ³Tianjin Institute of Crop Research, Tianjin Academy of Agricultural Sciences, Tianjin, China, ⁴College of Life Sciences, Jiangxi Normal University, Nanchang, China, ⁵College of Life Sciences, Nankai University, Tianjin, China

Background: Plant chloroplast DNA (cpDNA) typically has a circular structure, including a large single-copy region (LSC), a small single-copy region (SSC) and two inverted repeats (IR1 and IR2). The organization of these four elementary regions LSC-IR1-SSC-IR2 is highly conserved across all plant cpDNAs. Very few structural variations (SVs) occurring at the elementary-region level have been reported.

Results: In the present study, we assembled the full-length cpDNA of Dongxiang wild rice line 159 (DXWR159). Using the long PacBio subreads, we discovered a large inversion of SSC and a large duplication of IR in DXWR159 cpDNAs. Significantly, we reported for the first time forward and reverse SSCs of cpDNAs in similar proportions and named the frequent inversion of a whole SSC as SSC switching.

Conclusions: Our study helps researchers to correctly assemble the chloroplast genomes. Our recombination model explained the formation of large SVs in cpDNAs and provided insights into a novel scientific question that if there are common mechanisms in the formation or translocation of all kinds of transposon-like elements (TLEs). We propose that: (1) large inversion is the most accepted mutation type of SVs in cpDNAs; (2) SSC switching ubiquitous occurs in plant cpDNAs; and (3) further investigation of molecular mechanism underlying SSC switching may reveal new driving forces for large SVs.

KEYWORDS

structural variation, inverted repeat, asymmetric recombination, SDSA, transposon-like element

Abbreviations: DXWR, Dongxiang wild rice; cpDNA, chloroplast DNA; LSC, large single-copy region; SSC, small single-copy region; IR1, inverted repeat 1; IR2, inverted repeat 2; SSC-F, forward SSC; SSC-R, reverse SSC; InDel, small insertion and deletion; SNP, nucleotide polymorphism; STR, short tandem repeat; SSR, simple sequence repeat; CNV, copy number variation; SV, structural variation; ORF, open reading frame; NGS, next-generation sequencing.

Background

Chloroplast genomes (plastomes), also called chloroplast DNAs (cpDNAs) contain valuable markers for studying evolutionary relationships and population genetics of plants (1). In contrast to mitochondrial and nuclear genomes, cpDNAs across spermatophytes (*i.e.* seed plants) exhibit a higher degree of conservation with respect to their gene content, structure and organization (Asaf et al., 2017). Most cpDNAs encode about 80 protein-coding genes that are primarily involved in photosynthesis and other biochemical processes, along with 4 rRNA and 30 tRNA genes (Moghaddam et al., 2022). A plant cpDNA typically has a circular structure and includes a large single-copy region (LSC), a small single-copy region (SSC) and two inverted repeats (denoted as IR1 and IR2) separating the LSC and SSC. As the organization of these four elementary regions is highly conserved across all plant cpDNAs, LSC-IR1-SSC-IR2 is used a common structure for the assembly of chloroplast genomes. Although many structural variations (SVs) including duplications, deletions, insertions, and inversions have been reported at the levels of genome or gene between some of the angiosperm lineages, including *Asteraceae* (Ki-Joong et al., 2005), *Campanulaceae* (Haberle et al., 2008), *Onagraceae* (Stephan et al., 2008), *Fabaceae* (Cai et al., 2008) and *Geraniaceae* (Jansen, 2011), very few of the reported SVs occur at the elementary-region level, *e.g.*, a SV resulting in loss of a whole IR in a clade of *Papilionoideae* (Moghaddam et al., 2022), which is regarded as a rarely occurring event.

Cultivated rice (*Oryza sativa* L.), belonging to the grass family *Poaceae* (*Gramineae*), was domesticated from common wild rice (*Oryza rufipogon* Griff.). Dongxiang wild rice (DXWR) is a Chinese common wild rice (*O. rufipogon*) that was firstly discovered in Dongxiang county, Jiangxi province of China in 1978, which is northernmost (28°14'N) of the regions where many common wild rice population have been discovered globally. In our previous study (Zhang et al., 2016), we have compared the DXWR genome with the reference genome of cultivated rice (*O. sativa* ssp. *japonica*) and determined the loss/acquisition of genes during rice domestication by detection of SVs. However, we were not able to assembly the complete DXWR genome, as only short next-generation sequencing (NGS) data of DXWR genome was available. In order to accurately determine more genomics features of wild rice, we have initiated a project to obtain the full-length nuclear, mitochondrial and chloroplast genomes of DXWR using the PacBio DNA-seq (Xu et al., 2019). During the genome assembly, we unexpectedly discovered multiple large SVs in only a few seedlings. Accordingly, we report these very important findings for three main purposes: (1) to provide a new understanding of the conservation and variation in cpDNAs; (2) to help researchers to correctly assemble the chloroplast genomes; and

(3) to initiate investigation of the molecular mechanism underlying SSC switching, leading to discover new driving forces for large SVs.

Results

Genome sequencing, assembly and annotation

For the *de novo* assembly of full-length nuclear, mitochondrial and chloroplast genomes of Dongxiang wild rice line 159 (DXWR159), one 500 bp and one 10 Kb DNA libraries were prepared using fresh leaves from a few (<4) seedlings of DXWR159 and sequenced on the Illumina and PacBio platforms, respectively. In the subsequent data analyses, 217,263 subreads extracted from the PacBio DNA-seq data were used to assemble the DXWR159 cpDNA with a total length of 134,509 bp at an extremely high depth of approximately 13,441 X. Then, the two IRs, long poly (GC), low complexity, and other repeat regions were exactly determined by manual curation (Methods and Materials). Long (> 20 Kb) PacBio subreads were used to validate the structure of DXWR159 cpDNA and all detected SVs. As the draft genome using high-depth PacBio data may still contain two types of errors in the low complexity and STR regions, respectively (Chang et al., 2022), 1,223,905 pairs of Illumina DNA-seq reads were properly aligned to the DXWR159 cpDNA and only one error in a STR region was corrected. The DXWR159 cpDNA is a full-length genome (Supplementary File 1), as defined to has no gaps and ambiguous nucleotides (Chang et al., 2022).

According to the NCBI RefSeq database, 4 rRNA, 30 tRNA, 75 protein-coding genes and 15 open reading frames (ORFs) have been annotated in the rice reference cpDNA (RefSeq: NC_001320) of *O. sativa* ssp. *japonica* (Nipponbare). Both DXWR159 and Nipponbare have almost identical chloroplast rRNA, tRNA and protein-coding genes (Figure 1). In a recent study (Moghaddam et al., 2022), 4 rRNA, 30 tRNA and 76 protein-coding genes (Table 1) have been well annotated in the *Onobrychis gaubae* cpDNA (GenBank: LC647182) with a length of 122,688 bp and also in the *O. viciifolia* cpDNA (GenBank: MW007721) with a length of 121,932 bp. By sequence alignment, we identified 73 of the 76 *Onobrychis* protein-coding genes as common genes which are present in both rice (DXWR159 and Nipponbare) and *Onobrychis* (*O. gaubae* and *O. viciifolia*) cpDNAs, while the three other protein-coding genes (*accD*, *ycf1* and *ycf2*) are absent in rice cpDNAs. Furthermore, we updated the annotations of chloroplast genes in rice and *Onobrychis* with correction, particularly: (1) the annotation of *psbF* [missed in the previous study (Moghaddam et al., 2022)] was added into the *Onobrychis* cpDNAs; (2) ORF44 in rice was identified as *psaI*; and (3) *ORF23*, *ORF28*, *ORF56*, *ORF72*, *ORF82*, *ORF85*, *ORF100*, and *ORF137* in rice were removed.

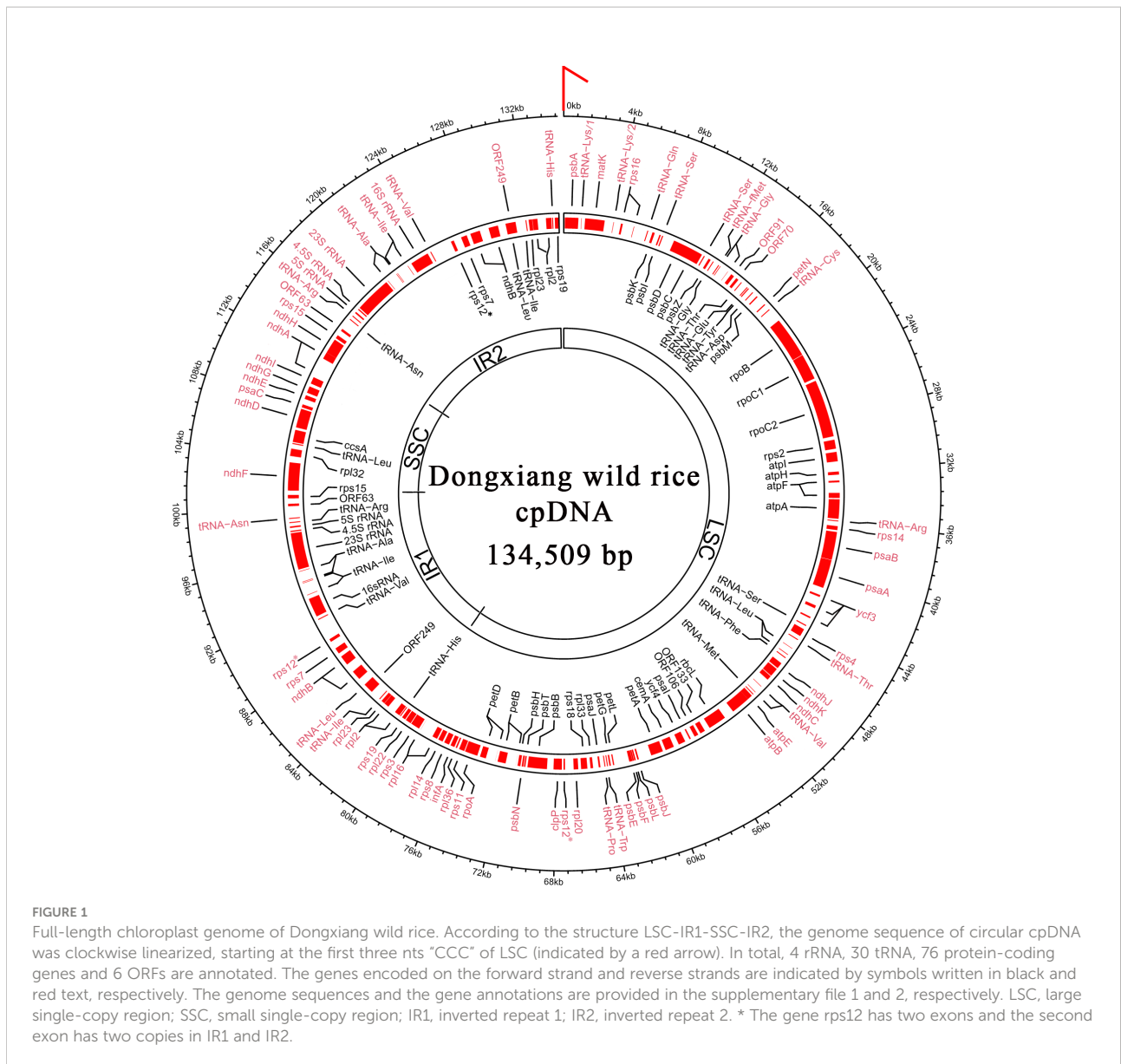


FIGURE 1

Full-length chloroplast genome of Dongxiang wild rice. According to the structure LSC-IR1-SSC-IR2, the genome sequence of circular cpDNA was clockwise linearized, starting at the first three nts “CCC” of LSC (indicated by a red arrow). In total, 4 rRNA, 30 tRNA, 76 protein-coding genes and 6 ORFs are annotated. The genes encoded on the forward strand and reverse strands are indicated by symbols written in black and red text, respectively. The genome sequences and the gene annotations are provided in the supplementary file 1 and 2, respectively. LSC, large single-copy region; SSC, small single-copy region; IR1, inverted repeat 1; IR2, inverted repeat 2. * The gene rps12 has two exons and the second exon has two copies in IR1 and IR2.

After correction, 4 rRNA, 30 tRNA, 76 protein-coding genes and 6 ORFs (*ORF63*, 70, 91, 106, 133, and 249) were annotated in the DXWR159 and Nipponbare cpDNAs (Supplementary File 2). The 76 rice protein-coding genes include 73 common genes which are present in both rice and *Onobrychis* cpDNAs, while the three other protein-coding genes (*infA*, *rpl22* and *rps16*) are absent in *Onobrychis* cpDNAs. Among the 76 protein-coding genes in rice, 10 multi-exon genes are *rpl2*, *rpl16*, *rps12*, *ndhA*, *ndhB*, *petB*, *petD*, *atpF*, and *rps16* with two exons, and *ycf3* with three exons (Table 1). The same 4 rRNA and 30 tRNA genes are present in both rice and *Onobrychis* cpDNAs. Among the 30 tRNAs, six are multi-exon genes which contain two exons, and they are tRNA^{Lys}(AAA), tRNA^{Gly}(GGA), tRNA^{Leu}(UUA), tRNA^{Val}(GUA), tRNA^{Ile}(AUC), tRNA^{Ala}(GCA). In addition,

tRNA^{Leu} (UUA), tRNA^{Leu} (CUA), tRNA^{Lys} (AAA), tRNA^{Ser} (UCC), tRNA^{Ser} (UCA), tRNA^{Ser} (AGC), tRNA^{Tyr} (UAC) have irregular secondary structures.

Short tandem repeats between individuals

Blasting the DXWR159 cpDNA sequence to the NCBI NT database, we found that the best hit is the cpDNA (GenBank: CP056064) of Zhenshan97, a cultivar of *O. sativa* ssp. *indica*. The length of DXWR159 cpDNA was determined to be 134,509 bp, which is very close to the Zhenshan97 and Nipponbare cpDNA lengths of 134,501 bp and 134,525 bp, respectively. The

TABLE 1 Annotated genes in cpDNAs.

Category	Group	Genes
Self-replication§	Large subunit of ribosomal proteins (8)	<i>rpl2*</i> , <i>rpl14</i> , <i>rpl16*</i> , <i>rpl20</i> , <i>rpl23</i> , <i>rpl32</i> , <i>rpl33</i> , <i>rpl36</i>
	Small subunit of ribosomal proteins (11)	<i>rps2</i> , <i>rps3</i> , <i>rps4</i> , <i>rps7</i> , <i>rps8</i> , <i>rps11</i> , <i>rps12*</i> , <i>rps14</i> , <i>rps15</i> , <i>rps18</i> , <i>rps19</i>
	DNA-dependent RNA polymerase (4)	<i>rpoA</i> , <i>rpoB</i> , <i>rpoC1</i> , <i>rpoC2</i>
	Ribosomal RNA genes (4)	rRNA 16S, 23S, 4.5S, 5S
	Transfer RNA genes (30)	30 tRNA genes (6 contain an intron)
Genes for photosynthesis§	Subunits of NADH-dehydrogenase (11)	<i>ndhA*</i> , <i>ndhB*</i> , <i>ndhC</i> , <i>ndhD</i> , <i>ndhE</i> , <i>ndhF</i> , <i>ndhG</i> , <i>ndhH</i> , <i>ndhI</i> , <i>ndhJ</i> , <i>ndhK</i>
	Subunits of photosystem I (5)	<i>psaA</i> , <i>psaB</i> , <i>psaC</i> , <i>psaI</i> , <i>psaJ</i>
	Subunits of photosystem II (15)	<i>psbA</i> , <i>psbB</i> , <i>psbC</i> , <i>psbD</i> , <i>psbE</i> , <i>psbF</i> , <i>psbH</i> , <i>psbI</i> , <i>psbJ</i> , <i>psbK</i> , <i>psbL</i> , <i>psbM</i> , <i>psbN</i> , <i>psbT</i> , <i>psbZ</i>
	Subunits of cytochrome b/f complex (6)	<i>petA</i> , <i>petB*</i> , <i>petD*</i> , <i>petG</i> , <i>petL</i> , <i>petN</i>
	Subunits of ATP synthase (6)	<i>atpA</i> , <i>atpB</i> , <i>atpE</i> , <i>atpF*</i> , <i>atpH</i> , <i>atpI</i>
Others§	Subunit of rubisco (1)	<i>rbcL</i>
	Maturase K	<i>matK</i>
	Envelope membrane protein	<i>cemA</i>
	Subunit of Acetyl-CoA-carboxylase	<i>accD</i>
	C-type cytochrome synthesis gene	<i>ccsA</i>
Unknown function§	Protease	<i>clpP</i>
	Conserved hypothetical ORFs (4)	<i>ycf1</i> , <i>ycf2</i> , <i>ycf4</i> , <i>ycf3**</i>
Only present in Rice	Genes (3) and ORFs (6)	<i>infA</i> , <i>rpl22</i> , <i>rps16*</i> , <i>ORF63</i> , 70, 91, 106, 133, and 249

§76 protein-coding genes are present in *Onobrychis* cpDNAs. In the second column, the number in parentheses after the group name is the gene number of this group. In the third column, the number of asterisks after the gene names indicates the number of introns contained in the genes. The duplicated genes in two IRs are counted once. Three genes *accD*, *ycf1* and *ycf2* are absent in rice cpDNAs, while three genes *infA*, *rpl22* and *rps16* are absent in *Onobrychis* cpDNAs. In addition, *ORF63*, 70, 91, 106, 133, and 249 were predicted in rice cpDNAs. The sequences of all these genes are provided in the [Supplementary File 1](#).

GC contents of the DXWR159, Zhenshan97, and Nipponbare cpDNAs are also very close (approximately 39%). In addition, the LSC, IR1, SSC, and IR2 of the DXWR159 cpDNA had lengths of 80,553, 20,805, 12,346, and 20,805 bp, and shared nt sequence identities of 99.93% (80505/80565), 99.99% (20804/20805), 99.99% (12346/12347), and 99.99% with those of the Zhenshan97 cpDNA, respectively. Likewise, the LSC, IR1, SSC, and IR2 of the DXWR159 cpDNA shared the nucleotide (nt) sequence identities of 99.54% (80338/80709), 99.81% (20779/20817), 99.76% (12320/12350), and 99.81% with those of the Nipponbare cpDNA, respectively. These results indicated that cpDNAs across rice species exhibit an extreme high degree of conservation with respect to their gene content, structure and organization. Multiple sequence alignment of the three cpDNAs demonstrated that about 76% (1519/2000) of the variation sites were single nucleotide polymorphisms (SNPs), while the others were small insertions and deletions (InDels). Almost all the InDels were associated with copy number variation (CNV) of STRs. STRs (Chen et al., 2020), which are widely used by forensic geneticists and genealogy experts, are often referred to as simple sequence repeats (SSRs) by plant geneticists or microsatellites by oncologists. The minimum repeat unit length of STR is evidently 1 bp; this type of STR is predominantly referred to as a polynucleotide (e.g. polyAs and polyTs).

Subsequently, we detected the variation sites across two lines of DXWR, DXWR159 and DXWR line 44 (DXWR44). Aligning the NGS data of DXWR44 (Materials and Methods) to the DXWR159

cpDNA, we found that DXWR159 and DXWR44 share an identical chloroplast genome without copy number variation (CNV) of STRs or SNPs between them. We then detected CNV of STRs and SNPs in cpDNAs of DXWR159 and DXWR44, respectively. Most of the detected STRs with CNV were polyAs or polyTs, but SNPs were not detected. Among these polyAs and polyTs, at least nine were shared by DXWR159 and DXWR44. They were 11291[T]₇₋₈, 31462[T]₇₋₈, 36488[T]₉₋₁₀, 49216[T]₁₀₋₁₅, 63462[A]₆₋₉, 80673[A]₂₋₃, 102303[A]₆₋₈, 107081[A]₆₋₇ and 111165[A]₆₋₈, where a STR (e.g. 11291[T]₇₋₈) is described by its genomic position in numbers (e.g. 11291 in 11291 [T]₇₋₈), the repeat unit [in brackets] (e.g. [T] in 11291[T]₇₋₈), and its copy numbers as subscripts (e.g. 7-8 in 11291[T]₇₋₈), as described in our previous study (Chen et al., 2020). We inferred that CNV of STRs, particularly polyAs or polyTs occurs more frequently than we expected, and accounts for cpDNA diversity within an individual of plants, just as it does to mtDNA diversity within an individual of tick, insect and human, which has been reported in our previous study (Chen et al., 2020).

Large structural variations and SSC switching

The comparative genomics analysis revealed many large (> 10 Kb) SVs between rice and *Onobrychis* cpDNAs. Among these SVs, three large inversions (named as SV1, SV2 and SV3) were notable in size: SV1 resulted in a reverse orientation of 26 genes

in the LSCs and the loss/acquisition of two genes (*rps16* and *accD*) (Figure 2A); SV2 resulted in a reverse orientation of 10 genes (Figure 2A); and SV3 resulted in a reverse orientation of *Onobrychis* SSC, compared to the rice SSC (Supplementary File 2). The other SVs were much smaller than the three SVs in size. Among these smaller SVs, two deletions resulted in the loss of *ycf1* and *ycf2*. The two genes (*ycf1* and *ycf2*) are absent in rice cpDNAs, while *ycf3* and *ycf4* are present in both rice and *Onobrychis* cpDNAs. The genes *ycf1-4* are the most common members of the hypothetical chloroplast reading frames. Both Nipponbare and *Onobrychis gaubae* have *ycf3* with a comparatively high nt sequence identity of 89% (447/505), and *ycf4* with a very low identity (far less than 70%). As the only protein-coding gene containing three exons in these cpDNAs (Table 1), *ycf3* is highly conserved in its nt sequences across kingdoms, indicating that *ycf3* may have important biological functions.

Subsequently, two important large SVs (named as SV4 and SV5) were discovered in DXWR159 cpDNAs. SV4 was a large

inversion of SSC, resulting in a reverse orientation of SSC. According to the orientation of SSC, cpDNAs can be classified into the forward SSC (SSC-F) and reverse SSC (SSC-R) type cpDNAs, which were also designated as wild-type cpDNAs and mutants (Figure 2B), respectively. The organization of Nipponbare cpDNA was defined as the SSC-F type structure (denoted as LSC-IR1-SSC-IR2). Accordingly, the organization of a cpDNA with a SSC reverse to that of the Nipponbare cpDNA was defined as the SSC-R type structure (denoted as LSC-IR1-ssc-IR2). According to the designations, the two *Onobrychis* cpDNAs (Described above) have the SSC-R structures. As the most significant finding, both SSC-F and SSC-R type cpDNAs were detected in only several seedlings of DXWR159. Then, we used long (> 20 Kb) PacBio subreads as junction reads spanning the LSC-IR1-SSC and LSC-IR1-ssc regions (Figure 2B) for validation (Materials and methods). As a result, these junction reads (Supplementary File 3) demonstrated the presence of both SSC-F (n = 82) and SSC-R type (n = 99) cpDNAs in similar proportions of approximately 45% to 55%. The SSC/ssc ratio of

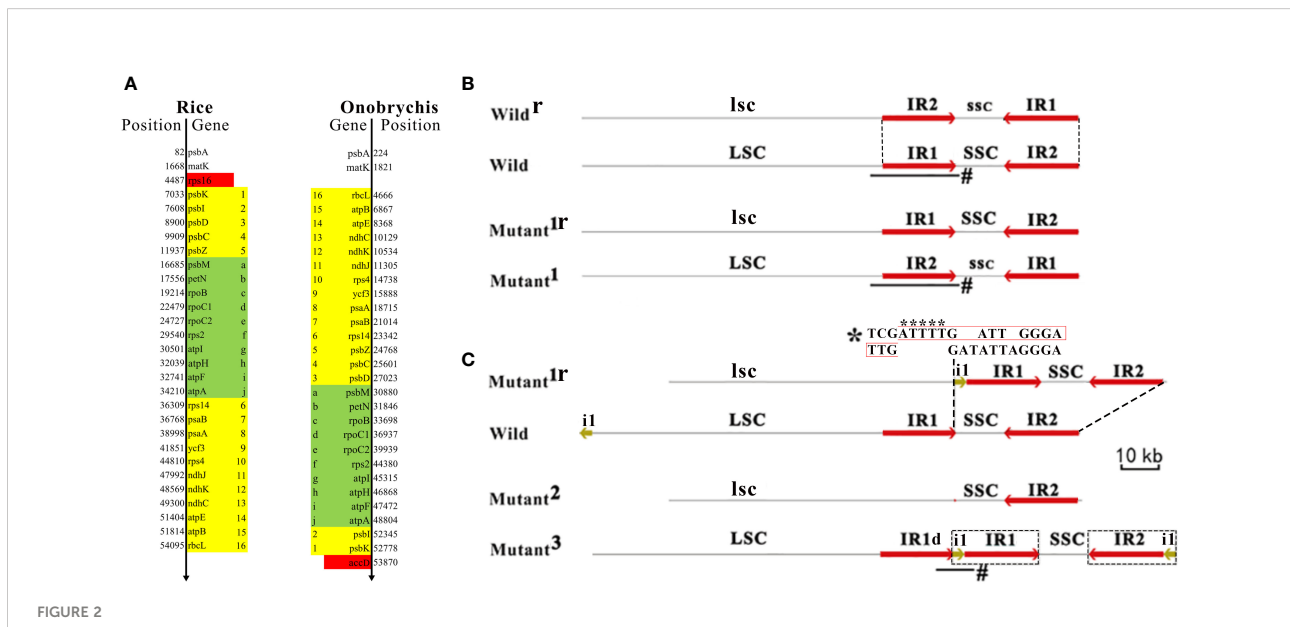


FIGURE 2

Large structural variations in cpDNAs and their formation. Plant chloroplast DNA (cpDNA) typically has a circular structure, including a large single-copy region (LSC), a small single-copy region (SSC) and two inverted repeats (IR1 and IR2). The organization of these four elementary regions LSC-IR1-SSC-IR2 is highly conserved across all plant cpDNAs. (A) A large inversion resulted in a reverse orientation of 26 genes (from *psbK* to *rbcL*) in the LSCs and the loss/acquisition of *rps16* and *accD* (in red color). Another large inversion resulted in a reverse orientation of 10 genes (in green color). The genomic positions of the genes were annotated according to the Nipponbare cpDNA (RefSeq: NC_001320) and the *Onobrychis gaubae* cpDNA (GenBank: LC647182). (B) A large inversion of SSC results in a reverse orientation of SSC (denoted as ssc). According to the orientation of SSC, cpDNAs can be classified into the forward SSC (SSC-F) and reverse SSC (SSC-R) type cpDNAs, which were also designated as wild-type cpDNAs (denoted as wild) and mutants (denoted as mutant¹). The frequent inversion of a whole SSC was named as SSC switching. Wild^r and mutant^{1r} represents cpDNAs that have reverse complementary sequences to wild and mutant¹, respectively. LSC and SSC are reverse complementary to lsc and ssc, respectively. As IR1 is identical to IR2, both LSC-IR1-ssc-IR2 and LSC-IR2-ssc-IR1 represent the same sequence (C) The recombination between mutant^{1r} and wild resulted in a cpDNA (mutant²) that lost a IR and another (mutant³) that acquired a IR. Although the IR-lacking cpDNA (mutant²) was not detected in the present study, it had been previously reported in *Onobrychis spp* (Moghaddam et al., 2022). While the 319-bp deleted segment was at the 3' end of the IR1 the 22810-bp insert includes five nucleotides "ATTTT" plus a 2000-bp segment of LSC (denoted as i1) and a whole IR1. The 319-bp lacking IR1 is denoted as IR1d. A high-score segment pair (HSP) that may be involved as recombination sites (indicated by *) was detected between the 5' end of the 319-bp deleted segment (the below sequence) and that of i1 (the above sequence). Junction reads (indicated by #) were used to validate these structural variations (Materials and methods). The sequence (in black box) synthesized during recombination contain an insert of five nucleotides "ATTTT" (indicated by asterisks below) that resulted in a repeat TTGAATTTTGATT.

0.82 (0.45/0.55) suggested that the inversion of a whole SSC is not a rarely occurring event and may occur frequently. We used SSC switching to name the frequent inversion of a whole SSC, as it is analogous to the mating-type (MAT) region in a yeast genome (Haber, 2012). MAT switching and SSC switching occur in similar structures IR-MAT-IR and IR-SSC-IR, respectively. The MAT region has a length of ~19 Kb between two IRs, while the SSC has a length of ~12.4 Kb between two IRs. During MAT or SSC switching (Figure 2B), the two IRs recombine, inverting MAT or SSC relative to the rests of the chromosomes. Neither MAT nor SSC switching changes the sequences of genes, as the structure of IR-MAT-IR or IR-SSC-IR is symmetric. By MAT switching, yeasts initiate the transcription of some mating related genes, determining the sexual identity of a cell. The molecular mechanism underlying MAT switching has already been revealed by scientists (Haber, 2012). However, the molecular mechanism underlying SSC switching is unknown.

Further analysis showed that rice cpDNAs are highly conserved on the junction site between IR1 and SSC (IR1-SSC) or that between SSC and IR2 (SSC-IR2). As rice cpDNAs are linearized for their representation, starting at the first three nts “CCC” of LSC, almost all the IR1-SSC junction sites contain the highly conserved sequences of TGGAAAAATCG|GCAAATAGGAAA and TGGAAAAAATCG|CGGAAAACCGAA for the SSC-F and SSC-R types, respectively. This feature was then used to quickly screen rice cpDNAs from the NCBI GenBank database. The results revealed that the structures of most rice cpDNAs belong to the SSC-F type, whereas very few (e.g. Zhenshan97) belong to the SSC-R type. The notable SSC-F type cpDNAs are the cpDNAs of *O. sativa* L. spp. *indica* (GenBank: JN861110), *O. nivara* (KF359901), *O. glaberrima* (KF359903), *O. barthii* (KF359904), *O. glumaepatula* (KF359905), *O. meridionalis* (KF359906), *O. punctata* (MT726932) and *O. brachyantha* (MT726939). However, The SSCs in some of the SSC-F type cpDNAs are chimeras. For instance, *O. nivara* and *O. punctata* have IR1-SSC junction sites of the SSC-R type cpDNAs, but SSC bodies of the SSC-F type cpDNAs (Supplementary file 1). This suggested that these SSCs had been assembled using hybrid data from both the SSC-F and SSC-R type cpDNAs. Based on the above results, we proposed that: (1) SSC switching ubiquitous occurs in plant cpDNAs, resulting in the presence of both SSC-F and SSC-R type cpDNAs in different proportions; and (2) an assembled chloroplast genome only represents the dominant type of cpDNAs. However, most of cpDNAs have been “*de novo*” assembled using short NGS or Sanger sequencing data and the orientations of SSCs have been determined according to the structure LSC-IR1-SSC-IR2 without validation using junction reads. According to the classical definition, *de novo* assembly refers to the process of reconstructing a genome without knowing its structure. For instance, the Zhenshan97 cpDNA (GenBank: CP056064) was *de novo* assembled using PacBio DNA-seq data. We inferred that the SSC-F type cpDNAs of

Zhenshan97 are supposed to be present as the SSC-R type cpDNAs, in a comparatively smaller proportion.

SV5 was a large duplication of IR in cpDNAs of several DXWR159 seedlings (Figure 2C). Compared to the wild-type cpDNA, the mutant acquired a 319-bp deletion (denoted as d1) and a 22810-bp insertion. While the 319-bp segment was deleted from the 3' end of the IR1 (Figure 2C), the 22810-bp insert included five nucleotides “ATTTT” plus a 2000-bp segment of LSC (denoted as i1) and a whole IR1. Six long (> 20 Kb) PacBio subreads (Supplementary File 1) were used as junction reads for validation (Materials and Methods). As a result, these junction reads spanned the IR1d-i1-IR1 region (Figure 2C), validating this new finding. Using the blast program, a high-score segment pair (HSP) that may be involved as recombination sites (Figure 2C) was detected between the 5' end of the 319-bp deleted segment and that of i1. This inspired us to propose a homologous recombination model to explain the formation of SV4 and SV5. This model includes symmetric recombination and asymmetric recombination. Symmetric recombination is reversible, while asymmetric recombination is irreversible. Symmetric recombination occurs frequently in symmetric regions (e.g. IR1-SSC-IR2 and IR1-ssc-IR2) of cpDNAs (Figure 2B), repeatedly resulting in the SSC-F and SSC-R structures (i.e., SSC switching). Therefore, both structures are ubiquitous present in plant cpDNAs. However, recombination rarely occurs in asymmetric regions (e.g. i1-IR1-SSC-IR2 and d1-SSC-IR2), which was defined as asymmetric recombination (Figure 2C). Asymmetric recombination results in loss/acquisition of large segments of cpDNAs. According to this model, an asymmetric recombination between two cpDNAs resulted in an exchange between i1+IR1 and d1. Consequently, one of the two cpDNAs acquired a large duplication of IR, whereas the other lost an IR, presumably resulting in an IR-lacking cpDNA. Although IR-lacking cpDNAs were not detected in the present study, an IR-lacking cpDNA had been previously reported in *Onobrychis* spp (Moghaddam et al., 2022), confirming the prediction of our model. Our recombination model explained the formation of large SVs in cpDNAs and revealed the association between symmetric recombination and asymmetric recombination.

One of two additional findings is that the i1-IR1 region is the invert repeat of the IR2-i1 region, indicating that the second large SV resulted in IR expansion. Asymmetric recombination maybe a cause of IR expansion. The other additional finding is that the sequence synthesized during recombination contain an insert of five nucleotides “ATTTT” that resulted in a tandem repeat TTGATTTTGATT (Figure 2C), which merits further investigation. Further analysis revealed that the tandem repeat may has a breakpoint “TGG|ATT” for enzyme cleavage at 5' end of the IR1. Both SSC switching and MAT switching occur in featured DNA regions, which were defined as DNA transposon-like elements (TLEs), as they have similar structures as DNA transposon elements (TEs). We defined that a TLE is composed

of an internal coding region bounded by two flanking IRs but does not include transposase genes in the coding region. As both IR-SSC-IR and IR-MAT-IR are TLEs, we addressed a novel scientific question that if there are common mechanisms in the formation or translocation of all kinds of TLEs. To answer this question, the first step is to validate our hypothesis that symmetric recombination of TLEs results in TLE switching, while asymmetric recombination of TLEs results in loss/acquisition of large segments and possible IR expansion. The validation requires intensive research of a variety of TLEs in nuclear, chloroplast and mitochondrial genomes. One of our ongoing studies is to detect the switching of a transposon-like element 1 (TLE1) in tick mitochondrial genomes using PacBio DNA-seq. TLE1, as the first reported TLE in mitochondrial genomes (Chen et al., 2020), includes genes ND1, tRNA^{Leu}, 16S rRNA, tRNA^{Val}, 12S rRNA, CR1, tRNA^{Ile}, tRNA^{Gln} in the coding region with two flanking IRs. The IRs contains the breakpoint-like sequence “TGCA|” at their 5' ends. Based on the above results, we proposed that SSC switching, MAT switching or even more TLEs may share common mechanisms. The most possible mechanism is synthesis-dependent strand-annealing (SDSA), which has been well studied in MAT switching in previous studies (Haber, 2012). SDSA in yeasts initiates when HO endonuclease makes a double-stranded DNA break at the MAT locus (Haber, 2012). As DNA synthesis in SDSA does not use all the factors employed in normal replication, the possible errors in synthesized DNA can not be corrected or removed. The insert of five nucleotides “ATTTT” may be a mutation caused by errors in synthesized DNA by SDSA. Another possible mechanism may depend on the transposase system, which has not been reported in the formation or translocation of any TLE. However, as recombinases, transposases and other enzymes with similar functions contain common structural domains (e.g. homeodomain), the complexity of the system goes beyond what is known based on existing knowledge. Thus, further studies of symmetric and asymmetric recombination may provide a new research direction to unravel the underlying mechanisms.

Conclusions

In the present study, we assembled the full-length chloroplast genome of Dongxiang wild rice, a Chinese common wild rice. The two main findings included the CNV of STRs between individuals of DXWR159 and two important large SVs at the elementary-region level. Large SVs and CNV of STRs, particularly polyAs or polyTs occur more frequently than we expected within an individual of plants, which merits further investigation in future studies. As CNV of STRs occurs frequently in an individual, STRs in cpDNAs may not be useful as molecular markers in phylogenetic studies,

particularly at low taxonomic levels (e.g., between individuals, lines or species). As the most significant finding, both SSC-R and SSC-F type cpDNAs were detected in similar proportions in several DXWR159 seedlings. The frequent inversion of a whole SSC was named as SSC switching. SSC switching occurs much more frequently than we expected within an individual of plants.

We proposed a homologous recombination model to explain the formation of two important large SVs. This model includes symmetric recombination and asymmetric recombination. Symmetric recombination is reversible, while asymmetric recombination is irreversible. Symmetric recombination results in SSC switching, while asymmetric recombination results in loss/acquisition of large segments of cpDNAs and possible IR expansion. The SSC/ssc ratio may indicate the activities of the enzymes that are responsible for SSC switching and asymmetric recombination. Our recombination model explained the formation of large SVs in cpDNAs and provided insights into a novel scientific question that if there are common mechanisms in the formation or translocation of all kinds of TLEs. We propose that: (1) large inversion is the most accepted mutation type of SVs in cpDNAs; (2) SSC switching ubiquitous occurs in plant cpDNAs; and (3) further investigation of molecular mechanism underlying SSC switching may reveal new driving forces for large SVs. As most species may have both SSC-F and SSC-R type cpDNAs, we recommend researchers to assemble the reference cpDNA using the SSC-F structure of the species rather than using the dominant structure of their samples. In addition, researchers can report the proportions of SSC-F and SSC-R type cpDNAs by counting the junction reads (Materials and Methods).

Materials and methods

All specimen used in the present study was identified by Fantao Zhang. DXWR159 and DXWR44 are two rice lines isolated from the Dongtangshang and Anjiashan populations of Dongxiang wild rice, respectively. DNA extraction and quality control were performed as described in our previous study (Wang et al., 2016). A 500 bp DNA library of DXWR44 was constructed as described in our previous study (Zhang et al., 2016) and sequenced on the Illumina HiSeq 2000 platform to produce 90-bp paired-end data. A 350 bp DNA library of DXWR159 was constructed and sequenced on the Illumina HiSeq X Ten platform to produce 150-bp paired-end data. A 10 Kb DNA library of DXWR159 using fresh leaves from a few (<4) seedlings was constructed and sequenced on the PacBio Sequel platform, according to the manufacturer's instructions. The cleaning and quality control of PacBio DNA-seq data were performed with the software SMRTlink v5.0, while the cleaning and quality control of Illumina data was performed with the software Fastq_clean (Zhang et al., 2014) v2.0.

The PacBio DNA-seq data of DXWR159 was used to assemble the chloroplast genome with the software CANU

(Koren et al., 2017) v2.2. The alignment of Illumina and PacBio DNA-seq data were performed with the software BWA (Li and Durbin, 2010) v0.7.10. Perl scripts for processing PacBio data in a special folder were integrated into Fastq_clean v2.0. Genome graphs (i.e., Figure 1) were plotted using the software Circos (Krzywinski et al., 2009) v0.66. Statistics and plotting were conducted using the software R v2.15.3 with Bioconductor packages (Gao et al., 2014). Using the software Tablet (Milne et al., 2013) v1.17, manual curation of the two IRs, long poly (GC), low complexity, and other repeat regions were performed. Long (> 20 Kb) PacBio subreads were used as junction reads to validate the structure of DXWR159 cpDNA and all detected SVs. To validate SVs (e.g. LSC-IR1-SSC and LSC-IR1-ssc), junction reads are required to be aligned to the SSC-F and SSC-R cpDNAs of DXWR159 with more than 90% of their lengths and cover 100% of IR1 with two 200-bp long overhangs that can be aligned to LSC and SSC/ssc at the 5' and 3' ends of IR1, respectively (Figure 2B). These aligned junctions reads were provided in SAM format with the SSC-F and SSC-R cpDNAs of DXWR159 (Supplementary file 3), which can be observed using the software Tablet.

Data availability statement

The datasets presented in this study can be found in online repositories. The NGS data are available in the NCBI SRA database with ID SRP070627. All the supporting data are included as additional files.

Ethics statement

All research on the rice lines detailed in this manuscript comply with the IUCN Policy Statement on Research Involving Species at Risk of Extinction and the Convention on the Trade in Endangered Species of Wild Fauna and Flora. All lines of Dongxiang wild rice (DXWR) are preserved *ex situ* at Jiangxi Academy of Agricultural Sciences, Nanchang, China (<http://www.jxaas.com/index.html>), and the seeds of DXWR are freely available for scientific research.

Author contributions

SG conceived this project. SG and SY supervised this study. JL, QW, and YY analyzed the data. RC performed the programming. FZ conducted experiments. SG drafted the main manuscript. SG revised the manuscript. ML prepared all the figures, tables and additional files. All authors have read and approved the manuscript.

Funding

This work was supported by the Natural Science Foundation of Jiangxi Province, China (20202ACB205002) to Fantao Zhang and the Innovation Research and Experiment Program for Youth Scholar, Tianjin Academy of Agricultural Sciences (2021023) to Qian Wang. The funding bodies played no role in the study design, data collection, analysis, interpretation or manuscript writing.

Acknowledgments

We appreciate the help from Professor Jiankun Xie from Jiangxi Normal University and Professor Wenjun Bu in College of Life Sciences, Nankai University. This manuscript was online as a preprint on March 20nd, 2022 at Research Square with the DOI 10.21203/rs.3.rs-1470820/v1.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2022.929352/full#supplementary-material>

SUPPLEMENTARY DATA SHEET S1

Contains all the sequences to support the results.

SUPPLEMENTARY DATA SHEET S2

Contains the corrected annotations of the DXWR159 and Nipponbare cpDNAs.

SUPPLEMENTARY DATA SHEET S3

Contains all junction reads (in the ssc-f.sam and ssc-r.sam files) aligned to the SSC-F and SSC-R cpDNAs (in the ssc-f.fa and ssc-r.fa files) of DXWR159 to validate LSC-IR1-SSC and LSC-IR1-ssc.

References

- Asaf, S., Khan, A. L., Khan, M. A., Imran, Q. M., and Lee, I. J. (2017). Comparative analysis of complete plastid genomes from wild soybean (*Glycine soja*) and nine other glycine species. *PLoS One* 12 (8), e0182281. doi: 10.1371/journal.pone.0182281
- Cai, Z., Guisinger, M., Kim, H. G., Ruck, E., Blazier, J. C., Mcmurtry, V., et al. (2008). Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *J. Mol. Evol.* 67 (6), 696–704. doi: 10.1007/s00239-008-9180-7
- Chang, J., Bei, J., Qi, S., Wang, H., Fan, H., Yau, T. O., et al. (2022). Full-length genome of an *ogataea polymorpha* strain CBS4732 *ura3Δ* reveals large duplicated segments in subtelomeric regions. *Front. Microbiol.* 13, 1–10. doi: 10.3389/fmicb.2022.855666
- Chen, Z., Xuan, Y., Liang, G., Yang, X., Yu, Z., Barker, S. C., et al. (2020). Precise annotation of tick mitochondrial genomes reveals multiple copy number variation of short tandem repeats and one transposon-like element. *BMC Genomics* 21 (488), 1–11. doi: 10.1186/s12864-020-06906-2
- Gao, S., Ou, J., and Xiao, K. (2014). *R language and bioconductor in bioinformatics applications (Chinese edition)* (Tianjin: Tianjin Science and Technology Translation Publishing Ltd).
- Haber, J. E. (2012). Mating-type genes and MAT switching in *Saccharomyces cerevisiae*. *Genetics* 191 (1), 33–64. doi: 10.1534/genetics.111.134577
- Haberle, R. C., Fourcade, H. M., Boore, J. L., and Jansen, R. K. (2008). Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *J. Mol. Evol.* 66 (4), 350–361. doi: 10.1007/s00239-008-9086-4
- Jansen, R. K. (2011). Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: Rearrangements, repeats, and codon usage. *Mol. Biol. Evol.* 28 (1), 583–600. doi: 10.1093/molbev/msq229
- Ki-Joong, K., Keung-Sun, C., and Jansen, R. K. (2005). Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). *Mol. Biol. Evol.* 9, 1783–1792. doi: 10.1093/molbev/msi174
- Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., and Bergman, N. H. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27 (5), 722–736. doi: 10.1101/gr.215087.116
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circo: An information aesthetic for comparative genomics. *Genome Res.* 19 (9), 1639–1645. doi: 10.1101/gr.092759.109
- Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698
- Liang, J., Chen, R., Zhang, F., Wang, Q., Yang, Y., Lv, M., et al. (2022). Full-length chloroplast genome of *Dongxiang* wild rice reveals single-copy region switching in cpDNAs. *Preprint*. doi: 10.21203/rs.3.rs-1498134/v1
- Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., et al. (2013). Using tablet for visual exploration of second-generation sequencing data. *Briefings Bioinf.* 14 (2), 193–202. doi: 10.1093/bib/bbs012
- Moghaddam, M., Ohta, A., Shimizu, M., Terauchi, R., and Kazempour-Osaloo, S. (2022). The complete chloroplast genome of *Onobrychis gaubae* (Fabaceae-papilionoideae): Comparative analysis with related IR-lacking clade species. *BMC Plant Biol.* 21 (1), 1–15. doi: 10.1186/s12870-022-03465-4
- Stephan, G., Xi, W., Uwe, R., Silber, M. V., Klaus, M., Jörg, M., et al. (2008). The complete nucleotide sequences of the five genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: I. sequence evaluation and plastome evolution. *Nucleic Acids Res.* 36 (7), 2366–2378. doi: 10.1093/nar/gkn081
- Wang, Y., Wang, Z., Chen, X., Zhang, H., Guo, F., Zhang, K., et al. (2016). The complete genome of *Brucella suis* 019 provides insights on cross-species infection. *Genes* 7 (2), 1–12. doi: 10.3390/genes7020007
- Xu, X., Ji, H., Jin, X., Cheng, Z., Yao, X., Liu, Y., et al. (2019). Using pan RNA-seq analysis to reveal the ubiquitous existence of 5' and 3' end small RNAs. *Front. Genet.* 10, 1–11. doi: 10.3389/fgene.2019.00105
- Zhang, F., Xu, T., Mao, L., Yan, S., Chen, X., Wu, Z., et al. (2016). Genome-wide analysis of *Dongxiang* wild rice (*Oryza rufipogon* Griff.) to investigate lost/acquired genes during rice domestication. *BMC Plant Biol.* 16, 1–11. doi: 10.1186/s12870-016-0788-2
- Zhang, M., Zhan, F., Sun, H., Gong, X., Fei, Z., and Gao, S. (2014). "Fastq_clean: An optimized pipeline to clean the Illumina sequencing data with quality control, Bioinformatics and Biomedicine (BIBM)," in *2014 IEEE International Conference on, IEEE*, (Netherlands: IEEE), 44–48.