The chloroplast genome of *Rosa rugosa × Rosa sertata* (Rosaceae): genome structure and comparative analysis

Yuan Niu¹ ², Yanyan Luo¹, Chunlei Wang¹, Qiong Xu² and Weibiao Liao¹

¹Gansu Agricultural University, College of Horticulture, Lanzhou, China.
²Lanzhou Agro-technical research and Popularization Center, Lanzhou, China

Abstract

*Rosa rugosa × Rosa sertata*, which belongs to the family Rosaceae, is one of the native oil-bearing roses in China. Most research has focused on its essential oil components and medicinal values. However, there have been few studies about its chloroplast genome. In this study, the whole chloroplast genome of *R. rugosa × R. sertata* was sequenced, analyzed, and compared to other genus *Rosa* species. The chloroplast genome of *R. rugosa × R. sertata* is a circular structure and 157,120 bp in length. The large single copy and small single copy is 86,173 bp and 18,743 bp in size, respectively, and the inverted repeats are 26,102 bp in size. The GC content of the whole genome is 37.96%, while those of regions of LSC, SSC, and IR are 35.20%, 31.18%, and 42.73%, respectively. There are 130 different genes annotated in this chloroplast genome, including 84 protein coding genes, 37 tRNA genes, 8 tRNA genes, and 1 pseudogene. Phylogenetic analysis of 19 species revealed that *R. rugosa × R. sertata* belong to the Sect. *Cinnamomeae*. Overall, this study, providing genomic resources of *R. rugosa × R. sertata*, will be beneficial for species identification and biological research.

Keywords: Rosaceae; *Rosa rugosa × Rosa sertata*; chloroplast genome; phylogenetic relationship.

Received: October 7, 2021; Accepted: July 15, 2022.
150 model. Generated 17,902,347 paired-end raw reads and the sequencing data was first filtered by Trimmomatic (version 0.36), low-quality reads were discarded and the reads contaminated with adaptor sequences were trimmed. The clean reads and reference sequence as *R. acicularis* (Chen et al., 2019) (GenBank accession no. MK714016.1) were used to extract chloroplast-like reads, which aligned to the database built by Genepioneer Biotechnologies (Nanjing, China) using Bowtie2 v2.2.4 (Langmead and Salzberg, 2012) and SPAdes v3.10.1 (Bankevich et al., 2012). Then, the sequences with the cp-like reads were assembled with NOVOPlasty (Dierckxsens et al., 2017). Annotation of the assembled chloroplast sequence was conducted with two methods. Firstly, the CDS, tRNA and rRNA were predicted with Prodigal v2.6.3 (Hyatt et al., 2010), hmmer v3.1b2 (Prakash et al., 2017) and Aragorn v1.2.38 (Laslett and Canback, 2004), respectively. Secondly, blast v2.6 (Johnson et al., 2008) was used to compare the gene sequences of the assembled one and the reference species. To determine the final annotation, the above two results were manually checked to remove the redundant and determine the multiple exon boundaries. A circular map of *R. rugosa × R. sertata* plastid genome was generated using the Chloroplot program (Zheng et al., 2020).

The whole chloroplast genome sequence of *R. rugosa × R. sertata* was determined and deposited to GenBank under accession number: MT845214. The size of the complete chloroplast genome is 157,120 bp, near other *Rosa* chloroplast genome level. It displayed a typical quadripartite structure, possessing a LSC region (86,176 bp), an SSC region (18,743 bp) and a pair of IR region (52,204 bp) (Figure 1). The overall GC content is 37.22%, and the order of GC content in different regions is 42.73% in IR regions, 35.20% in LSC region and 31.18% in SSC region (Table S1). It is a normal phenomenon that the highest GC content exists in the IR regions in different plants. There has been studies that show that such GC skewness can be indicators of replication origins, replication terminals, DNA lead chains or lag chains (Tillier and Collins, 2000; Necșulea and Lobry, 2007).

---

**Figure 1** – Gene map of the *R. rugosa × R. sertata* chloroplast genome. The genes with different functional classification are color coded and the pseudogene is marked asterisks. The genes are shown inside and outside the outermost layer represented with transcription directions clockwise and
A total of 130 functional genes were detected in *R. rugosa × R. sertata* chloroplast genome, including 84 protein-coding genes, 37 tRNA genes, and 8 rRNA genes. In the IR regions, there were 12 protein-coding genes, 14 tRNA genes and 8 rRNA genes. In the LSC and SSC region, there were 60 and 12 protein-coding genes, and 22 and 1 tRNA genes, respectively (Table 1, Table S2). In addition, the *ycf1* was interpreted as pseudogene in our study as it contains several internal stop codons. Studies have shown that splicing introns are the hallmark of eukaryotic genes (Niu and Yang, 2011). In this study, 22 genes with intron structure were detected in this chloroplast genome, including 8 tRNA genes and 13 protein-coding genes (Table S3). Among them, 11 genes are in the LSC regions: 1 gene in the SSC region and 10 genes in the IR region. Genes *ycf3* and *clpP* contain two introns, which is consistent with other chloroplast genomes (Li *et al.*, 2018; Xue *et al.*, 2019; Liu *et al.*, 2020).

### Table 1 – List of genes annotated in the *R. rugosa × R. sertata* chloroplast genome.

<table>
<thead>
<tr>
<th>Category</th>
<th>Gene group</th>
<th>Gene name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photosynthesis</td>
<td>photosystem I</td>
<td>psaA,psaB,psaC,psal,psaJ</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>cytochrome b/f complex</td>
<td>petA,petB*,petD*,petG,petL,petN</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>ATP synthase</td>
<td>atpA,atpB,atpE,atpF,atpH,atpI</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Large subunit of rubisco</td>
<td>rbcL</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>photochlorophyllide reductase</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Self-replication</td>
<td>large ribosomal subunit</td>
<td>rpl14,rpl16*,rpl2*2,rpl20,rpl22,rpl23(2),rpl32,rpl33,rpl36</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>small ribosomal subunit</td>
<td>rps11,rps12**2,rps14,rps15,rps16*,rps18,rps19,rps2,rps3,rps4,rps7(2),rps8</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>RNA polymerase</td>
<td>rpoA,rpoB,rpoC1*,rpoC2</td>
<td>4</td>
</tr>
<tr>
<td>Ribosomal RNAs</td>
<td>rmL16(2),rmL23(2),rmL4.5(2),rmL5(2)</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>transfer RNAs</td>
<td>trnA-UGC*(2),trnC-GCA,trnD-GUC,trnE-UUC,trnF-GAA,</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>trnG-GCC*,trnG-UCC,trnG-HUG,trnM-CAU(2),trnM-GAU*(2),</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>trnK-UUU*,trnL-CAA(2),trnL- UUU*,trnL- UAG,trnM-CAU,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>trnN-GUU(2),trnP-UGC,trnQ-UUG,trnR-ACG(2),trnR-UCU,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>trnS-GCU,trnS-GGA,trnS-UGA,trnT-GTT,trnT-GUG,trnV-GAC(2),trnV-UAC*,trnW-CCA,trnY-GUA,trnM-CAU</td>
<td></td>
</tr>
<tr>
<td>Other Genes</td>
<td>Maturase</td>
<td>matK</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Protease</td>
<td>clpP**</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Envelope membrane protein</td>
<td>cemA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Acetyl-CoA carboxylase</td>
<td>accD</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>c-type cytochrome synthesis gene</td>
<td>ccsA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Translation initiation factor</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>Genes of unknown function</td>
<td>Conserved hypothetical chloroplast ORF</td>
<td>#ycf1,ycf1,ycf2(2),ycf3**,ycf4</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>130</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Gene*: Gene with one introns; Gene**: Gene with two introns; #Gene: Pseudo gene; Gene(2): Number of copies of multi-copy genes; - : Nonexistent gene

A total of 37 long repeats were detected by REPuter software, including 16 forward repeats, 19 palindromic repeats, and 2 reverse repeats (Figure S1). More than half of the long repeats (51%) were distributed in intergenic spaces (IGSs), 10.81% in both genes and IGSs, and 37.83% in genes. In addition, the distribution of these repeats in different regions is distinct. The number of repeats in regions of LSC, SSC, IRa, and IRb is 19, 3, 13, and 13, respectively. Some repeats, such as genes of *ycf1*, *ycf2*, *ndhB*, *trnG-GCC*, and *trnS-UGA*, existed in two regions simultaneously. In total, 260 SSRs were detected in *R. rugosa × R. sertata* chloroplast genome by software MISA. Among them, 65.55%, 19.33% and 15.13% were in the regions of LSC, IR, and SSC, respectively. Besides, from the perspective of the relationship with location of genes, 52.52%, 34.45%, and 13.03% were found in the IGSs, coding regions and introns, respectively. Types with a number of 20 or more were A (8), T (8), and T (9), while types with numbers between 5 and 20 were A (9), A (10), T (10), T (11), TA (5), TAA (3), TTA (3), and TTC (3). The number of other types was less than 5. The frequencies of mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide were 62.69%, 5.00%,
26.92%, 4.23%, 0.38%, and 0.77%, respectively. Among the identified mononucleotide SSRs, A/T types (92.64%) was dominant compared with G/C types (7.36%).

Gene flow between species or genetic diversity within a species is often measured by comparison of the chloroplast sequences. To determine differences in the chloroplast genome sequences of *R. rugosa*, *R. odorata var. gigantea*, *R. multiflora*, *R. luciae*, *R. canina* and *R. rugosa × R. sertata*, sequence identity was calculated for these species’ chloroplast sequence using the online program mVISTA with *R. chinensis cultivar Old Blush* as a reference (Figure S2, Table S4). Consistent with other studies, the region of greatest divergence is LSC, in which the noncoding regions possess higher divergence than coding regions. The chloroplast genome of *R. rugosa × R. sertata* is closer to *R. rugosa*, and the significant variation between them could be found in the intergenic regions of *psbM-trnD*, *trnD-trnY*, *rbcL-accD*, *petB-petD*, *petD-rpoA*, *rps3-rpl22*, *trnL-ndhB* and *ndhF-rpl32* (Supplementary Data). It would be considered valuable to utilize the identification of these higher-resolution loci for species identification.

In the long term of evolution, the change of the IR region at the borders plays a critical role. In our study, the genetic architecture of seven *Rosa* genomes was mapped at the junction of the IR region, LSC region, and SSC region by IRscope (Figure S3). Gene location and gene order were relatively conservative in *Rosa*. In *R. canina*, *R. odorata*, *R. rugosa*, *R. chinensis*, and *R. rugosa × R. sertata*, the coding region of *ycf1* was at the border of SSC/IRa, and spanned the SSC and IRa region, while in *R. luciae* and *R. multiflora*, it was at the border of SSC/IRb and spanned the SSC and IRb region. It is noteworthy that in *R. rugosa* and *R. rugosa × R. sertata*, the pseudogene *ycf1* was located in IRb, while in *R. luciae* and *R. multiflora*, it was located in IRa. The mutation region of pseudogene *ycf1* in IRa/SSC or IRb/SSC region was 1106-1111 bp.

The phylogenetic analysis was performed based on complete chloroplast genome sequences from 19 taxa, including 18 *Rosa* species and one outgroup (*Vitis vinifera*, MN561034.1), all of which were downloaded from the NCBI database except the *R. rugosa × R. sertata*. All the sequences from these 19 species were aligned by MAFFT v7.455 (Katoh and Standley, 2013) and trimmed by trimAl (Capella-Gutiérrez et al., 2009). A maximum likelihood (ML) analysis was performed by IQtree (Nguyen et al., 2015), and a bootstrap test was set with 1000 repetitions. The result of phylogenetic analysis was visualized by MEGA v7.0 (Kumar et al., 2016) (Figure 2). The chloroplast genomes play a significant role in understanding the evolutionary relationship and history of plant species (Jansen et al., 2007). Here, as expected, 14 species from the *Rosa* genus formed a monophyletic clade composed of seven branches, which were consistent with the seven subgroups obtained by morphological classification. *R. rugosa × R. sertata* was mostly related to *R. rugosa*, with bootstrap support value of 100%. They all belong to the Sect. *Cinnamomeae*. The availability of a completed *R. rugosa × R. sertata* chloroplast genome sequence will provide useful information for the phylogenetic study among *Rosa*.

Overall, the complete chloroplast genome of *R. rugosa × R. sertata*, an endemic oil-bearing rose species in China, was firstly reported and analyzed. The characteristics of quadripartite structure, genome size, GC content, and gene order of the plastid genome of *R. rugosa × R. sertata* were shown to be similar with that of other genus *Rosa* species. There were 37 long repeats sequences and 260 SSRs detected in this plastid genome. Besides, reconstructed phylogenetic relationships among 19 species found *R. rugosa × R. sertata* to be closely related to *R. rugosa*. These results combined with the comparison with the whole chloroplast genome of other genus *Rosa* species have provided the worthy information and will bring insight into developing DNA markers suitable for identification of species within this genus.

![Figure 2](image-url) – Maximum likelihood (ML) phylogenetic tree of 18 species of Rosaceae constructed using their chloroplast genomes. *Vitis vinifera* was used as the outgroup.
The chloroplast genome for a rose

Acknowledgments
This research was funded by the National Natural Science Foundation of China (Nos. 32072559, 31860568, 31560563, and 31160398); the National Key Research and Development Program (2018YFD0100800); the Research Fund of Higher Education of Gansu, China (No. 2018C-14 and 2019B-082); the Natural Science Foundation of Gansu Province, China (Nos. 1606RJZA073 and 1606RJZA077); the Science and Technology planning project of Gansu Province, China (No. 20YF8NA138).

Conflict of interest
The authors declare that there is no conflict of interest that could be perceived as prejudicial to the impartiality of the reported research.

Author Contributions
YN performed the experiments, analyzed the data and wrote the manuscript, YYL and YN designed the figures, YYL and CLW performed data curation, QX conceived the study, WBL supervise the project and reviewed the manuscript. All authors read and approved the final version.

References

Conflict of interest
The authors declare that there is no conflict of interest that could be perceived as prejudicial to the impartiality of the reported research.

Author Contributions
YN performed the experiments, analyzed the data and wrote the manuscript, YYL and YN designed the figures, YYL and CLW performed data curation, QX conceived the study, WBL supervise the project and reviewed the manuscript. All authors read and approved the final version.

References


Rehder A (1949) Bibliography of cultivated trees and shrubs hardy in the cooler temperate regions of the Northern Hemisphere. Arnold Arboretum of Harvard University, Boston.


Supplementary Material

The following online material is available for this article:

Table S1 - Base composition in the R. rugosa × R. sertata chloroplast genome.

Table S2 - The number of genes in the R. rugosa × R. sertata chloroplast genome.

Table S3 - The length of exons and introns in genes in the R. rugosa × R. sertata chloroplast genome.

Table S4 - Statistics of the chloroplast genomes of R. rugosa × R. sertata and five other Rosaceae species.

Figure S1 - Analysis of long repeat sequences and simple sequence repeats (SSRs) in R. rugosa × R. sertata chloroplast genome.

Figure S2 - Sequence identity plot of 6 Rosa chloroplast genomes by mVISTA.

Figure S3 - Comparison of LSC, SSC and IR regions in chloroplast genomes.

Supplementary Data - Sequence information of eight Intergenic Regions in the R. rugosa × R. sertata and R. rugosa chloroplast genome.

Associate Editor: Guilherme Correa de Oliveira

License information: This is an open-access article distributed under the terms of the Creative Commons Attribution License (type CC-BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original article is properly cited.