Comparative analysis of chloroplast genome structure of sweet potato and its wild relatives

Jianying Sun[1, 2], Xiaofeng Dong[1, 2], Qinghe Cao[3], Tao Xu[1, 2], Mingku Zhu[1, 2], Jian Sun[1, 2], Tingting Dong[1, 2], Daifu Ma[3], Yonghua Han[1, 2]* and Zongyun Li[1, 2]*

[1] Institute of Integrative Plant Biology, School of Life Sciences, Jiangsu Normal University, Xuzhou, China

[2] Jiangsu Key Laboratory of Phylogenomics and Comparative Genomics, Jiangsu Normal University, Xuzhou, China

[3] Jiangsu Xuhuai Regional Xuzhou Institute of Agricultural Sciences/Sweet potato Research Institute, Chinese Academy of Agricultural Sciences, Xuzhou, China

Jianying Sun and Xiaofeng Dong contributed equally to the work.

Corresponding Author:

Yonghua Han; Zongyun Li

Quanshan Campus: No.101, Shanghai Road, Tongshan District, Xuzhou, Jiangsu Province P.R.China 221116

Email address: hanyonghua@jsnu.edu.cn; zongyunli@jsnu.edu.cn

Absract

Background: Sweet potato (*Ipomoea batatas* (L.) Lam) is the seventh most important food crop in the world. In the present study, a detailed analysis of chloroplast genome in sweet potato and its ten wild relatives was conducted.

Methods: The total genomic data of the ten wild relatives was generated by next –generation sequencing and then these data were assembled. We downloaded the chloroplast genome of *I. batatas* (GenBank accession no.NC026703) and conducted a comparative analysis ~~in~~ of their chloroplast genome~~s~~ structure.

Results: The ten *Ipomoea* chloroplast genomes ranged from 161,225 bp to 161,721 bp and displayed the typical circular quadripartite structure, consisting of a pair of inverted repeat (IR) regions (30,798–30,910 bp each) separated by a large single copy (LSC) region (87,575–88,004 bp) and a small single copy (SSC) region (12,018–12,051 bp). The average guanine-cytosine (GC) content is approximately 40.5 % in the IR region, 36.1 % in the LSC, 32.2 % in the SSC regions, and 37.5 % in total length for all plastomes. The chloroplast genome~~s~~ sequences included 87 protein-coding genes, eight duplicated rRNA (rrn23, rrn16, rrn5, and rrn4.5), 37 tRNA, and *infA* was absent in all these chloroplast genomes. The boundaries of SC/IR were highly conserved in the chloroplast genomes of sweet potato and its wild relatives. We also found five relatively high variable regions (*rpl32-trnL*, *ndhH-ndhF*, *trnH-psbA*, *trnC-petN*, and *ndhA* intron), which may be used as markers for future population genetics research and species identification.

Discussion: The comparative analysis revealed that the chloroplast genomes of these eleven *Ipomoea* species were highly conserved both in sequence and structure. Generally, the detailed analysis of these chloroplast genomes provides valuable information for species identification and genetic resources in *Ipomoea* series *Batatas*.

Keywords: *Ipomoea*; chloroplast genome; divergence hotspot; genome structure

Comentado [A1]: There are more than ten sweet potato wild relatives. May be better rephrasing to “sweet potato and ten of its wild relatives”.

Comentado [A2]: Indicate “cultivar Xushu18” and the reference (Yan et al. 2015, PLoS ONE)

Comentado [A3]: It is not clear what regions are present or absent on each genome. Please rephrase for clarity

Comentado [A4]: This was already shown elsewhere (Roullier et al. 2013, Muñoz-Rodriguez et al. 2018 and the tortoise and hare papers).

Introduction

Sweet potato, *Ipomoea batatas* (L.) Lam, is one of the most important food crops in the world, grown in > 100 countries (FAO, 2017). However, the narrow genetic background and the lack of resistance genes in today's sweet potato cultivars limit its further improvement. The wild species may play an important role in providing new genes, such as those for resistance to various diseases and insects (Khoury et al., 2015). Strategies to utilize these wild *Ipomoea* germplasms in breeding programs will depend on the understanding genetic diversity and the relationships between sweet potato and its wild relatives.

Ipomoea batatas belongs to the genus *Ipomoea*, which is the largest genus in the Convolvulaceae family, which includes 600-700 species (Austin & Huáman, 1996). Thirteen species are considered to be closely related to *I. batatas*, and they form the *Ipomoea* series *Batatas* (Austin, 1987). The phylogenetic relationships in series *Batatas* and the evolutionary origin of *I. batatas* have always been one of the research foci. Studies conducted on morphological characteristics suggest that *Ipomoea trifida* (H.B.K.) G. Don (2X) and *Ipomoea triloba* L. are the most closely related species to *I. batatas* (Austin, 1987) and maybe they are the wild ancestors of *I. batatas*. Molecular markers, such as restriction fragment length polymorphisms (RFLPs) of genomic DNA (Jarret & Gawel, 1992), random amplified polymorphic DNAs (RAPDs) (Jarret & Austin, 1994), and inter-simple sequence repeats (ISSRs) (Huang & Sun, 2000) have been used to investigate phylogenetic relationships of *I. batatas* and its wild relatives. In addition, *β-amylase* (a nuclear-encoded gene) based method has also been employed for the phylogenetic analysis in this series (Rajapakse et al., 2004). A widely used gene in molecular phylogenetic analyses, *Waxy*, has been applied to series *Batatas*, this analysis indicated that *I. batatas* may be a hybrid between *Ipomoea littoralis* and *Ipomoea tenuissima* (Gao et al., 2011). The latest studies using both nuclear sequence and chloroplast genome sequence conducted by ~~Mun~~ Munoz-Rodr~~r~~iguez et al., (2018) suggested that the sweet potato had a single origin and the most likely ancestral species of sweet potato is *I. trifida*. Overall, these molecular approaches effectively improve the accuracy of phylogenetic research in contrast with morphological analysis and try to reveal the origin and evolution of sweet potato from different aspects. Although the chloroplast genomes of the species in the series *Batatas* have been obtained, the detailed chloroplast genome structure analysis between sweet potato and its wild relatives has not been conducted.

Comentado [A5]: Is this based on evidence and, if so, has it been reported? According to sweet potato breeders, the crop holds large genetic diversity, so it is not clear to me how "narrow" its genetic background is.

Comentado [A6]: Muñoz-Rodriguez et al. showed that this group consists of at least 16 species, including *Ipomoea batatas*.

Comentado [A7]: Too vague. Maybe better "a research foci for a long time"

The chloroplast, an organ of photosynthesis in plants, has its own genome. The chloroplast genome of *Nicotiana tabacum* was the first chloroplast genome to be sequenced in 1986 (Shinozaki et al., 1986). The length of a chloroplast genome DNA is generally 120 - 165kb (Raubeson & Jansen, 2005) , carrying a large amount of valuable evolutionary information (Carbonell-Caballero et al., 2015; Jansen et al., 2007). Chloroplast gnomes are haploid, maternal inherited, maintain a high conservation in gene content and genome structure, and have been widely used to study the evolutionary relationships at almost any taxonomic level in plants (Zhang et al., 2016; Tong et al., 2016; Jansen et al., 2007). Some difficult phylogenies can be resolved using the chloroplast genome, such as the bamboo tribe, Arundinarieae (Ma et al., 2014). In rice and cotton, chloroplast genomes were used to understand the evolutionary relationships between cultivated species and their wild relatives (Brozynska et al., 2014; Waters et al., 2012; Sotowa et al., 2013; Xu et al., 2012; Li et al., 2014). Eserman et al. (2014) recently conducted a phylogenetic analysis of morning glories based on chloroplast genomes in family Convolvulaceae, providing strong resolution in the Ipomoeeae (Eserman et al., 2014). Additionally, the chloroplast genome of a sweet potato named "Xushu 18" was sequenced in 2015 by Yan et al.(Yan et al., 2015). However, additional genomic information will be essential to better understand the origin and evolution of sweet potato and develop DNA markers for accurate species identification.

In this study, we sequenced and assembled ten wild *Ipomoea* chloroplast genomes. Our two aims were as follows: first, to understand the conservation and diversity of the *Ipomoea* chloroplast genome through comparative genomics approaches. Second, to identify appropriate DNA markers to accurately identify species and for further use in population genetic studies.

Materials & Methods

Sampling and DNA extraction

Total genomic DNA was extracted from ten species for sequencing (Table 1). ~~Samples~~ Plants were grown in the greenhouse of the Xuzhou Sweet potato Research Centre, China. Young leaves were collected from one plant of each species, subsequently frozen in liquid nitrogen and stored at -80 °C until further use. Total genomic DNA was extracted using the Takara miniBEST plant genomic DNA extraction kit (Dalian, China). The integrity of genomic was assessed by performing gel electrophoresis using a 1 % agarose gel.

Comentado [A8]: Valuable information for what?

Comentado [A9]: Often maternally inherited, but there are multiple examples of bi-parental inheritance, also investigated in Ipomoea: see for example Harris & Ingam 1991 (Taxon 40(3): 393) 10.2307/1223218.

Comentado [A10]: "Their structure is highly conserved/preserved"

Comentado [A11]: What is a "difficult phylogeny"?

Comentado [A12]: This paragraph is unclear. The authors mix information from different questions, from the first chloroplast genome sequenced to studies of the chloroplast genomes of *Ipomoea*. Also, those DNA markers have been developed already in the case of sweet potato in some papers aforementioned, Roullier´s contributions, etc.

Con formato: Fuente: Cursiva

Comentado [A13]: Genomes? Genomic sequences? Unclear.

Chloroplast genome assembling and annotation

We constructed ~~the~~ genomic ~~library~~ libraries using the TruSeq DNA Nano kit with a DNA insert size of 350 bp. Sequencing was conducted on the Illumina X Ten platform, which generated at least 21G raw data from each species. Sequence data was discarded if > 10 % of the base had a quality of < Q20 after removing the adaptor sequences. ~~On the basis of this rule, we obtained clean data.~~ Sequences were assembled according to the protocol described by Hahn et al. (Hahn et al., 2013) using the MIRA sequence assembler software with the reference genome *I. trifida* (accession number: KF242476.1). The services of library construction, sequencing, and assembly were provided by Macrogen (http://www.macrogencn.com/sy,Shenzhen, China).

The ten chloroplast genome sequences were initially annotated using the online CpGAVAS (Liu et al., 2012) software with default setting, and then manually corrected using Genious 11.0.5. The circular chloroplast genome maps were constructed using the OrganellarGenome DRAW tool (Lohse et al., 2007). We also downloaded *Ipomoea batatas* chloroplast genome sequences from GenBank (GenBank accession number: NC026703) in order to compare and analyze the divergence between sweet potato and its wild relatives in chloroplast genome.

Comentado [A14]: The authors mention sweet potato accession NC026703 in the abstract but then use an *Ipomoea trifida* accession for genome assembly. Why did they not use the *I. batatas* genome as reference? Also, the authors must be aware that this specimen, judging from Eserman et al. (2014) study, is likely **not** *I. trifida*.

Repeat structure analysis

Microsatellites (mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide repeats) were detected using the MISA-web (a web server for microsatellite prediction) (Beier et al., 2017) with default settings. REPuter (Kurtz et al., 2001) was used to visualize forward, palindrome, reverse, and complimentary sequences, with a minimum repeat size of 30 bp and a sequence identity > 90 %.

Divergence Hotspot Identification

In order to ~~observe~~ study the differences among these genomes, the 11 *Ipomoea* chloroplast genome sequences were aligned using MAFFT v7.307 (Katoh & Toh, 2010) including whole chloroplast genomes, and then sliding window analysis was conducted. The step size was set with to 200 bp, with a 600 bp as a window (Fu et al., 2017).

Phylogenetic Analysis

We downloaded 33 chloroplast genomes from GenBank almost cover all of the species in

Convolvulaceae family which have been sequenced and each coding gene was extracted from every chloroplast genome. Before ML analyses, a total of 65 coding genes shared by 43 chloroplast genomes were aligned using MAFFT plugin in Geneious 11.0.5, and then manually trimmed if necessary and compiled into a single file of "super gene". ML analyses were performed using RAxML-HPC2 on XSEDE with 1000 bootstrap replicates and the GTRGAMMA model on CPIRES Science Gateway (Miller et al., 2010) (https://www.phylo.org/).

Comentado [A15]: Muñoz-Rodriguez et al. 2018 made available the genomes of several other species of *Ipomoea* that have not been considered in this study.

Comentado [A16]: Unclear. Did the authors decide not to include non-coding regions in their study? If so, why? Please explain.

Results

Genome sequencing and assembly

At least 21 Gb raw data from each species were generated by Illumina sequencing technology then through filtering, we obtained 12.58 Gb-17.77Gb clean data. With the *I. trifida* (accession number: KF242476.1) genome as a reference, we removed the nuclear genome components from each species. Finally, there were 0.59 Gb -1.28 Gb ~~base~~ left and we assembled the chloroplast genomes of ten known *Ipomoea* species with these data. The coverage of chloroplast genome in each species comes from 3664 (in *I. splendor-sylvae*) to 7941 (in *I. cordatotriloba*) (Table S1). All ten newly sequenced chloroplast genomes were submitted to GenBank (accession numbers: MH173252-MH173254； MH173257-MH173263) (Table 1).

Genome features

Genome size and GC content

The plastomes of species in series *Batatas* were highly conserved in terms of genomic structure and size. Nucleotide sequence sizes of the newly sequenced ten *Ipomoea* chloroplast genomes ranged from 161,225 bp (*I. tabascana*; Table2, Fig. 1, Supplementary Figures) to161,721 bp (*I. splendor-sylvae*). The structure of the chloroplast genomes in these *Ipomoea* relatives displayed the typical circular quadripartite structure, ~~consisted~~ consisting of a pair of IR regions (30,798–30,910 bp) separated by an LSC region (87,575–88,004 bp), and an SSC region (12,018–12,051 bp). When comparing the proportions of three parts (LSC, SSC, IRs), we found that cultivar *I. batatas* had a relatively high proportion of LSC and SSC. On the contrary, IRs in *I. batatas* occupied a comparatively small part. The average GC content is approximately 40.5 % in the IR region, 36.1 % in the LSC, 32.2 % in the SSC regions, and

Comentado [A17]: Is this difference statistically significant?

37.5 % in the entire sequence of all plastomes (Table2).

Genes

The chloroplast genomes of ten wild species included 87 protein-coding genes, 8 duplicated rRNAs (*rrn*23, *rrn*16, *rrn*5, and *rrn*4.5), and 37 tRNAs. Based on their predicted functions, these genes can be divided into four categories, (1) genes related to photosynthesis; (2) genes related to self-replication; (3) genes related to the biosynthesis of cytochrom, protein, etc., and (4) functionally unknown *ycf* genes (Table S2). These 87 protein-coding genes are composed of 73 single-copy genes located in LSC/SSC regions and 7 two-copy genes in IRs. In the chloroplast genomes analyzed, there are 16 genes harboring introns. In these genes, 14 genes have only one intron; *ycf3* and *clpP* have two introns each (Table S3).

Codon usage

Chloroplast genomes of the ten *Ipomoea* species we studied contain totally 23,766-23,804 codons, possessing similar codon usage distribution, and AUU (Ile) was the most abundant codon in all samples (Table S4). The relative synonymous codon usage of the third position showed that the average frequency of GC codons was three times greater than AT.

Boundary between LSC/SSC and IRs

The junctions of LSC/IRa, SSC/IRa, and LSC/IRb are located in the IGS region between *rpl23* and *trnI*; *ndhH* and *ndhF*; and *trnI* and *trnH*, respectively and the location of SSC/IRb junction within the coding region of *ndhA* gene which made a ~~pseudogenes~~pseudogene of *ndhA* gene sizing 454 bp (Fig. 2). The distance from *ndhH* to IRA/SSC boundary and the length of *ndhA* gene located in the IRB in the eleven species is the same, whereas, the gaps of boundaries of LSC/IRA, IRB/LSC with the nearby gene were less than 25 bp.

Comentado [A18]: Is this relevant, for example for understanding species evolution? Please explain.

Repetitive Sequences Analysis

Repeat motifs are thought to have a significant impact on genome phylogeny and rearrangement (Yue et al., 2008). REPuter identified a total of 89 pairs of repeats (30 bp or longer) in ten newly sequenced *Ipomoea* species chloroplast genomes, including 41-46 palindromic repeats and 43-48 forward repeats (Table S5). The lengths of these repeats ranged from 41 to 193 bp, and copy lengths of 40–59 bp are most common while those with > 140 bp were least abundant (Fig. 3A). However, when compared to *I. batatas*, the chloroplast

Comentado [A19]: This sentence can be simplified throughout the text, it is not necessary to explain that the sequences are new every time.

genomes of these ten *Ipomoea* species chloroplast genomes contain less 60–79 bp repeats (Fig. 3A). Except for six repeats of the *acc*D gene located in the LSC region of *I. splendor-sylvae*, all other repeats could be found in the IR regions, mostly distributed in the *ycf1* gene region and the intergenic regions between *trnN*-GUU and *ycf1* or *trnI*-CAU and *ycf2*. (Table S5).

Simple sequence repeats (SSRs) are tandemly repeated nucleotides in DNA sequences. We found that ten *Ipomoea* chloroplast genome contained 47-54 SSRs that were apparently more than those identified in the *I. batatas* chloroplast genome (Fig. 3B). These ten wild relatives shared 28 SSRs, however, there were only 16 common SSRs between them and sweet potato. Among these SSRs, the majority consisted of mono-nucleotide repeats that contribute to 65 %-78 % of the SSRs in the newly sequenced *Ipomoea* chloroplast genomes, and the percentage even increased up to 92 % in the *I. batatas* chloroplast genome (Table S6). Most mono-nucleotide repeat sequences comprised A/T repeats. This is consistent with the previous findings suggest that chloroplast SSRs generally comprise short polyA or polyT repeats and rarely contain tandem G or C repeats (Kuang et al., 2011). Interestingly, almost no di- and tri-nucleotide repeat sequences exist across the compared *Ipomoea* chloroplast genomes. SSRs are different from the repetitive sequences identified by REPuter; they are almost all located in LSC regions (Table S6).

Divergence hotspot regions

The percentage of identical sites among these eleven species is 97.9 % showed highly consistent sequences. Sliding window analysis showed the mean value of the variation is 0.554 % and five relatively high variability regions with the variation rate > 2.5 % were detected, including *rpl32-trnL*, *ndhH-ndhF*, *trnH-psbA*, *trnC-petN*, and *ndhA* intron (Fig. 5). One is located on the IRa/SSC boundary (*ndhH-ndhF*), two of them are in the SSC region (*rpl32-trnL*, *ndhA* intron) and the remaining two are in the LSC region (*trnH-psbA*, *trnC-petN*). They are all from non-coding regions and these could be useful DNA markers for population genetic studies and species identification.

Phylogenetic Analysis

Using maximum likelihood (ML) analyses, we generated the topology that included the ten *Ipomoea* chloroplast genomes and 33 published Convolvulaceae chloroplast genomes, 29 of

which were of *Ipomoea* (Eserman et al., 2014; Yan et al., 2015). *Cuscuta exaltata* was included in the analysis as the outgroup taxa to perform these phylogenetic analyses. To conduct the ML analysis, 65 one-to-one protein-coding genes were extracted and aligned from among all the 43 chloroplast genomes. After elimination of poorly aligned positions and divergent regions of alignment, a super-alignment (62,522 bp long) was constructed and used for ML analysis (Figure 5). These 42 ingroups were divided into seven small groups, including *Batatas*, *Murucoides*, *Pes-caprae*, *Quamoclit*, *Cairica*, *Obscura*, and *Pes-tigridis*(Eserman et al., 2014). Our ten *Ipomoea* chloroplast genomes all clustered with *I. batatas* belonging to *Batatas* (Fig. 5).

Comentado [A20]: *Cuscuta* is known to be a group that poses extraordinary challenges in phylogenetic analysis because of their parasitic behavior. Is this the best choice as outgroup?

Comentado [A21]: Why did the authors not use the non-coding regions, despite explaining previously that those regions could be useful for phylogenetic analysis?

Discussion

Variations among the eleven *Ipomoea* species

In the present study, ten chloroplast genomes of *Ipomoea* species were assembled. They display~~ed~~ the typical quadripartite structure and the length of the chloroplast genome sequence in the ten wild species and *I. batatas* ranged from 161,225 to 161721 bp. The border regions of SC/IR was thought to be the main reason for the divergences in chloroplast genome size (Ravi et al., 2008). In wild species and *I. batatas*, the LSC/IRa/SSC/IRb boundary genes are highly conserved with slight structural variations and there is no significant extension/contraction of IRs between sweet potato and its wild relatives. The *Ipomoea* chloroplast genomes contained 132 genes. According to the statistic got by Gitzendanner et al. (Gitzendanner et al., 2018) some genes are often lost in plants including *accD*, *infA*, *petG*, *petL*, *psaI*, *psaJ*, *psb*, *rpl32*, *rpl36*, *ycf1*, and *ycf2*. All of these genes were presented in our chloroplast genomes except *infA*, which has been lost from all the eleven analyzed chloroplast genomes. *InfA* gene codes for translation initiation factor, almost lost in all rosid species and it also the most mobile chloroplast gene known in plants (Millen et al., 2001).

Larger and more complex repeat sequences may play an important role in the rearrangement of chloroplast genomes and sequence divergence (Timme et al., 2007; Weng et al., 2014). We, therefore, investigated and compared the numbers and distributions of dispersed, palindromic, and SSRs across the ten *Ipomoea* species and *I. batatas*. We found that dispersed and palindromic repeats in different species were usually located in the same *ycf1* and *ycf2* genes, or between *trnN*-GUU/ *ycf1* genes and *trnI*-CAU/*ycf2* genes. However, the SSRs were distributed more widely throughout these chloroplast genomes and usually located between or

within various *trn*, *atp*, *clp*, and *rpo* genes. Most of the SSRs were located in non-coding regions, however, for these lied in the genes which have no introns, they were located in the coding regions (e.x. *rpoC2*, *rpoB*, *atpB*, *ycf1*, *ycf2*, and *ndhF*). Because of their high polymorphism in the chloroplast genome, SSRs are possibly important molecular markers in the analysis of plant population genetics, evolutionary, and ecological studies (Xue et al., 2012).

In addition, nucleotide substitution (SNVs, indels, and proportions of variability) may play a critical role in the plant evolutionary processes. Our results demonstrate that nucleotide substitution rates in these chloroplast genomes are very low, suggesting that nucleotide substitution does not dominate the evolution of chloroplast genomes of the *Batatas* group. We found that IR regions were more conserved than the SC regions and that the variation was principally observed in the intergenic regions. Consistent with other angiosperms, the IR region of these species is more conserved than the LSC and SSC regions(Liu et al., 2017), possibly because of copy correction between IR sequences by gene conversion(Khakhlova & Bock, 2006).

Phylogenetic relationships

Determination of taxonomy and species in *Batatas* is particularly difficult because individuals often exhibit intermediate morphologies between descriptions of named species (Mcdonald JA, 1990; Austin, 1978), and several species may be of hybrid origin (Diaz et al., 1996). Phylogenetic analysis in series *Batatas* has been performed using DNA markers, such as RFLP, RAPD, ISSR, chloroplast restriction site variation, gene sequences, and morphological analyses (Rajapakse et al., 2004; Huang & Sun, 2000; Jarret & Austin, 1994; Jarret & Gawel, 1992). These studies have indicated the phylogenetic relationships between sweet potato and its wild relatives, however, the support values were low. Constructing the phylogenetic tree from the chloroplast genome sequences has been applied in *Ipomoea* (Eserman et al., 2014;(Muñoz-Rodríguez et al., 2018). Our tree presented different positions of some species in contrast with Mun˜ oz-Rodrı´guez et al. The newly sequenced *I. trifida* (PI 460377) and *I. trifida* (PI 618966) are the closest sister species to *I. batatas* and *I. tabascana*, whereas another *I. trifida* (REM 753) was clustered with *I. triloba*. This is different from Mun˜ oz-Rodrı´guez et al. whose studies showed that *I. trifida* was only close to *I. batatas*. In the study conducted by Eserman et al., (2014) showed that *I. trifida* (REM 753) was close to *I.*

Comentado [A22]: SSRs have been used for population genetics in *Ipomoea batatas* among other species. Please discuss and cite relevant papers.

Comentado [A23]: Unclear what the authors mean here. What is the dominant factor then, if any?

Comentado [A24]: Unnecessary sentence, repeated in next sentence.

Comentado [A25]: Phylogenetic relationships in this group have been resolved using whole chloroplast genome and 600 nuclear genes by Muñoz-Rodriguez et al. (2018), overcoming the difficulties in the studies mentioned here.

285 *cordatotriloba*, but their study didn't include *I. triloba*. *I. cordatotriloba* was also located in
286 two different positions on the phylogenetic tree. The different positions of *I. trifida* and *I.*
287 *cordatotriloba* can be explanied by their non-monphyletic origins (Eserman et al., 2014; Mun˜
288 oz-Rodrı´guez et al., 2018). Besides, *I. ramosissima* was clustered with *I. cynanchifolia* in our
289 study, however, it was grouped with *I. splendor-sylvae* by Mun˜ oz-Rodrı´guez et al. at a
290 relative basal position. The difference may be due to the different accessions we used just as
291 the *I. trifida* (REM 753) showed very different positions from the other *I. trifida* accessions.

292 These phylogenetic relationships derived in the current study support that *I. trifida* and *I.*
293 *tabascana* are the two sister species closest to *I. batatas* in the *Batatas* group (Rajapakse et
294 al., 2004; Huang & Sun, 2000; Jarret & Austin, 1994). RFLP (Jarret & Gawel, 1992) and
295 RAPD analyses (Jarret & Austin, 1994) indicate that *I. tabascana* is more similar to *I. batatas*
296 than *I. trifida*. Conversely, a more recent study, based on analysis of a nuclear gene (*β-*
297 *amylase*) supported that *I. trifida* as the species most closely to *I. batatas* based on exon
298 analysis, whereas the intron analysis indicated that *I. tabascana* is more closely related to *I.*
299 *batatas* (Rajapakse et al., 2004). *Ipomoea tabascana* was clustered with *I. batatas* in our
300 study which indicated that *I. tabascana* is closer to *I. batatas* than *I. trifida* and this result is
301 consistent with Mun˜ oz-Rodrı´guez et al., (2018).

302 **Conclusions**

303 In this study, we sequenced and assembled ten chloroplast genomes from wild relatives of *I.*
304 *batatas* with high coverage. Based on whole -chloroplast genome sequencing, phylogenetic
305 analysis of the series *Batatas* were conducted. Our results showed the chloroplast genome of
306 *I. batatas* and its wild relatives is highly consistent in sequence and structure and we also
307 identified five valuable genetic markers for investigating the population genetics and
308 biogeography of closely related *Ipomoea* species.
309
310 **References:**
311 **Austin, D.F. 1978**. *Ipomoea-Batatas* complex .1. Taxonmy. *Bulletin of the torrey botanical*
312 *club* **105**:114-129 DOI 10.2307/2484429.
313 **Austin, D.F. 1987.** The taxonomy evolution and genetic diversity of sweetpotatoes and related
314 wild species.: International Potato Center, Lima, Peru. p 27-60.

Comentado [A26]: Ipomoea trifida is monophyletic in Muñoz-Rodríguez et al. (2018) and in all previous studies using multiple specimens except Eserman et al. (2014). This most likely means that specimen KF242476 in Eserman's paper is misidentified.

Con formato: Español (España)

315 **Austin, D.F., and Huáman, Z. 1996**. A Synopsis of *Ipomoea* (Convolvulaceae) in the
316 Americas. *Taxon* **45**:3-38.
317 **Beier, S., Thiel, T., Munch, T., Scholz, U., and Mascher, M. 2017**. MISA-web: a web server
318 for microsatellite prediction. *Bioinformatics* **33**:2583-2585 DOI
319 10.1093/bioinformatics/btx198.
320 **Brozynska, M., Omar, E.S., Furtado, A., Crayn, D., Simon, B., Ishikawa, R., and Henry,**
321 **R.J. 2014**. Chloroplast Genome of Novel Rice Germplasm Identified in Northern Australia.
322 *Tropical Plant Biology* **7**:111-120 DOI 10.1007/s12042-014-9142-8.
323 **Carbonell-Caballero, J., Alonso, R., Ibanez, V., Terol, J., Talon, M., and Dopazo, J. 2015**.
324 A Phylogenetic Analysis of 34 Chloroplast Genomes Elucidates the Relationships between
325 Wild and Domestic Species within the Genus *Citrus*. *Molecular Biology and Evolution*
326 **32**:2015-2035 DOI 10.1093/molbev/msv082.
327 Crops (FAO, accessed 1 August 2017); http://www.fao.org/faostat/en/#data/QC.
328 **Diaz, J., Schmiediche, P., and Austin, D.F. 1996**. Polygon of crossability between eleven
329 species of Ipomoea: Section *Batatas* (convolvulaceae). *Euphytica* **88**:189-200 DOI
330 10.1007/BF00023890.
331 **Eserman, L.A., Tiley, G.P., Jarret, R.L., Leebens-Mack, J.H., and Miller, R.E. 2014**.
332 Phylogenetics and diversification of morning glories (tribe Ipomoeeae, Convolvulaceae) based
333 on whole plastome sequences. *American Journal of Botany* **101**:92-103 DOI
334 10.3732/ajb.1300207.
335 **Fu, C., Li, H., Milne, R., Zhang, T., Ma, P., Yang, J., Li, D., and Gao, L. 2017**. Comparative
336 analyses of plastid genomes from fourteen *Cornales* species: inferences for phylogenetic
337 relationships and genome evolution. *BMC Genomics* **18** DOI 10.1186/s12864-017-4319-9.
338 **Gao, M., Ashu, G.M., Stewart, L., Akwe, W.A., Njiti, V., and Barnes, S. 2011**. *Wx* intron
339 variations support an allohexaploid origin of the sweetpotato [*Ipomoea batatas* (L.) Lam].
340 *Euphytica* **177**:111-133 DOI 10.1007/s10681-010-0275-z.
341 **Gitzendanner, M.A., Soltis, P.S., Wong, G.K., Ruhfel, B.R., and Soltis, D.E. 2018**. Plastid
342 phylogenomic analysis of green plants: A billion years of evolutionary history. *American*
343 *Journal of Botany* DOI 10.1002/ajb2.1048.
344 **Hahn, C., Bachmann, L., and Chevreux, B. 2013**. Reconstructing mitochondrial genomes
345 directly from genomic next-generation sequencing reads--a baiting and iterative mapping
346 approach. *Nucleic Acids Research* **41**:e129 DOI 10.1093/nar/gkt371.

Huang, J.C., and Sun, M. 2000. Genetic diversity and relationships of sweetpotato and its wild relatives in *Ipomoea* series *Batatas* (Convolvulaceae) as revealed by inter-simple sequence repeat (ISSR) and restriction analysis of chloroplast DNA. *Theoretical and Applied Genetics* **100(7)**:1050-1060 DOI 10.1007/s001220051386.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., Depamphilis, C.W., Leebens-Mack, J., Muller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee, S.B., Peery, R., McNeal, J.R., Kuehl, J.V., and Boore, J.L. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc Natl Acad Sci U S A* **104**:19369-19374 DOI 10.1073/pnas.0709121104.

Jarret, R.L., and Austin, D.F. 1994. Genetic diversity and systematic relationships in sweetpotato (*Ipomoea batatas* (L.) Lam.) and related species as revealed by RAPD analysis. *Genetic Resources and Crop Evolution* **41**:165-173 DOI 10.1007/BF00051633.

Jarret, R.L., and Gawel, N.W.A. 1992. Phylogenetic Relationships of the Sweetpotato [*Ipomoea batatas* (L.) Lam. *J. Amer. Soc. Hort. Sci.* **117**:633-637.

Katoh, K., and Toh, H. 2010. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**:1899-1900 DOI 10.1093/bioinformatics/btq224.

Khakhlova, O., and Bock, R. 2006. Elimination of deleterious mutations in plastid genomes by gene conversion. *Plant Journal* **46**:85-94 DOI 10.1111/j.1365-313X.2006.02673.x.

Khoury, C.K., Heider, B., CastaÃ Eda-Ã Lvarez, N.P., Achicanoy, H.A., Sosa, C.C., Miller, R.E., Scotland, R.W., Wood, J.R.I., Rossel, G., Eserman, L.A., Jarret, R.L., Yencho, G.C., Bernau, V., Juarez, H., Sotelo, S., Haan, S.D., and Struik, P.C. 2015. Distributions, ex situ conservation priorities, and genetic resource potential of crop wild relatives of sweetpotato [*Ipomoea batatas* (L.) Lam., *I.* series *Batatas*]. *Frontiers in Plant Science* **6** DOI 10.3389/fpls.2015.00251.

Kuang, D., Wu, H., Wang, Y., Gao, L., Zhang, S., and Lu, L. 2011. Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. *Genome* **54**:663-673 DOI 10.1139/G11-026.

Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* **29**:4633-4642 DOI.

Li, P., Li, Z., Liu, H., and Hua, J. 2014. Cytoplasmic diversity of the cotton genus as revealed by chloroplast microsatellite markers. *Genetic Resources and Crop Evolution* **61**:107-119 DOI 10.1007/s10722-013-0018-9.

Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., and Guan, X. 2012. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics* **13**:715 DOI 10.1186/1471-2164-13-715.

Liu, L., Li, R., Worth, J.R.P., Li, X., Li, P., Cameron, K.M., and Fu, C. 2017. The Complete Chloroplast Genome of Chinese Bayberry (*Morella rubra*, Myricaceae): Implications for Understanding the Evolution of *Fagales*. *Frontiers in Plant Science* **8** DOI 10.3389/fpls.2017.00968.

Lohse, M., Drechsel, O., and Bock, R. 2007. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* **52**:267-274 DOI 10.1007/s00294-007-0161-y.

Ma, P.F., Zhang, Y.X., Zeng, C.X., Guo, Z.H., and Li, D.Z. 2014. Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe Arundinarieae (poaceae). *Systematic Biology* **63**:933-950 DOI 10.1093/sysbio/syu054.

Mcdonald JA, A.D. 1990. Changes and additions in *ipomoea* section *batatas* (Convolvulaceae). *Brittonia* **42**:116-120.

Millen, R.S., Olmstead, R.G., Adams, K.L., Palmer, J.D., Lao, N.T., Heggie, L., Kavanagh, T.A., Hibberd, J.M., Gray, J.C., Morden, C.W., Calie, P.J., Jermiin, L.S., and Wolfe, K.H. 2001. Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell* **13**:645-658.

Miller, M.A., Pfeiffer, W., and Schwartz, T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. 2010 Gateway Computing Environments Workshop (GCE)

. p 1-8.

Muñoz-Rodríguez, P., Carruthers, T., Wood, J.R.I., Williams, B.R.M., Weitemier, K., Kronmiller, B., Ellis, D., Anglin, N.L., Longway, L., Harris, S.A., Rausher, M.D., Kelly, S., Liston, A., and Scotland, R.W. 2018. Reconciling Conflicting Phylogenies in the Origin of Sweet Potato and Dispersal to Polynesia. *Current Biology* DOI 10.1016/j.cub.2018.03.020.

Rajapakse, S., Nilmalgoda, S.D., Molnar, M., Ballard, R.E., Austin, D.F., and Bohac, J.R. 2004. Phylogenetic relationships of the sweetpotato in *Ipomoea* series *Batatas* (Convolvulaceae)

based on nuclear β-amylase gene sequences. *Molecular Phylogenetics and Evolution* **30**:623-632 DOI 10.1016/S1055-7903(03)00249-5.

Raubeson, L.A., and Jansen, R.K. 2005. Chloroplast genomes of plants. In: Henry, R.J., ed. *Plant diversity and evolution: Genotypic and phenotypic variation in higher plants*, 44-68.

Ravi, V., Khurana, J.P., Tyagi, A.K., and Khurana, P. 2008. An update on chloroplast genomes. *Plant Systematics and Evolution* **271**:101-122 DOI 10.1007/s00606-007-0608-0.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B.Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Kato, A., Tohdoh, N., Shimada, H., and Sugiura, M. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *Embo Journal* **5**:2043-2049

Sotowa, M., Ootsuka, K., Kobayashi, Y., Hao, Y., Tanaka, K., Ichitani, K., Flowers, J.M., Purugganan, M.D., Nakamura, I., Sato, Y., Sato, T., Crayn, D., Simon, B., Waters, D.L., Henry, R.J., and Ishikawa, R. 2013. Molecular relationships between Australian annual wild rice, *Oryza meridionalis*, and two related perennial forms. *Rice (N Y)* **6**:26 DOI 10.1186/1939-8433-6-26.

Timme, R.E., Kuehl, J.V., Boore, J.L., and Jansen, R.K. 2007. A comparative analysis of the Lactuca and Helianthus (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats. *AMERICAN JOURNAL OF BOTANY* **94**:302-312 DOI 10.3732/ajb.94.3.302.

Tong, W., Kim, T., and Park, Y. 2016. Rice Chloroplast Genome Variation Architecture and Phylogenetic Dissection in Diverse *Oryza* Species Assessed by Whole-Genome Resequencing. *Rice* **9** DOI 10.1186/s12284-016-0129-y.

Waters, D.L., Nock, C.J., Ishikawa, R., Rice, N., and Henry, R.J. 2012. Chloroplast genome sequence confirms distinctness of Australian and Asian wild rice. *Ecology and Evolution* **2**:211-217 DOI 10.1002/ece3.66.

Weng, M.L., Blazier, J.C., Govindu, M., and Jansen, R.K. 2014. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Molecular Biology and Evolution* **31**:645-659 DOI 10.1093/molbev/mst257.

Xu, Q., Xiong, G., Li, P., He, F., Huang, Y., Wang, K., Li, Z., and Hua, J. 2012. Analysis of complete nucleotide sequences of 12 *Gossypium* chloroplast genomes: origin and evolution of allotetraploids. *PLoS One* **7**:e37128 DOI 10.1371/journal.pone.0037128.

Xue, J., Wang, S., and Zhou, S. 2012. Polymorphic chloroplast microsatellite loci in nelumbo (Nelumbonaceae). *American Journal of Botany* **99**:E240-E244 DOI 10.3732/ajb.1100547.

Yan, L., Lai, X., Li, X., Wei, C., Tan, X., and Zhang, Y. 2015. Analyses of the Complete Genome and Gene Expression of Chloroplast of Sweet Potato [*Ipomoea batata*]. *PLoS One* **10**:e124083 DOI 10.1371/journal.pone.0124083.

Yue, F., Cui, L., DePamphilis, C.W., Moret, B.M., and Tang, J. 2008. Gene rearrangement analysis and ancestral order inference from chloroplast genomes with inverted repeat. *BMC Genomics* **9 Suppl 1**:S25 DOI 10.1186/1471-2164-9-S1-S25.

Zhang, Y., Du L, Liu, A., Chen, J., Wu, L., Hu, W., Zhang, W., Kim, K., Lee, S.C., Yang, T.J., and Wang, Y. 2016. The Complete Chloroplast Genome Sequences of Five *Epimedium* Species: Lights into Phylogenetic and Taxonomic Analyses. *Frontiers in Plant Science* **7**:306 DOI 10.3389/fpls.2016.00306.

Figure legends

Fig.1. Chloroplast genome map of *Ipomoea tabascana*. The inside genes of the outer circle are transcribed counterclockwise while the genes outside are transcribed clockwise. LSC: large single copy; SSC: short single copy; IR: inverted repeats.

Fig.2. Comparison of the boundary between LSC/SSC and IR regions among the eleven *Ipomoea* chloroplast genomes.

Fig.3. Repeated sequences in eleven *Ipomoea* chloroplast genomes. A. Number of the five repeat types; B. Number of different SSR types.

Fig.4. Percentages of variable sites in homologous regions among the eleven *Ipomoea* chloroplast genomes.

Fig. 5. Phylogenetic tree reconstruction of ten new sequenced cp genomes and 33 previous sequenced chloroplast genomes based on 65 protein sequences.