



## RESEARCH ARTICLE

**REVISED** Insights into population structure of East African sweetpotato cultivars from hybrid assembly of chloroplast genomes [version 2; peer review: 2 approved, 3 approved with reservations]

Chenxi Zhou<sup>1</sup>, Tania Duarte <sup>1</sup>, Rocio Silvestre<sup>2</sup>, Genoveva Rossel<sup>2</sup>, Robert O. M. Mwanga <sup>3</sup>, Awais Khan<sup>2,4</sup>, Andrew W. George <sup>5</sup>, Zhangjun Fei<sup>6</sup>, G. Craig Yencho<sup>7</sup>, David Ellis<sup>2</sup>, Lachlan J. M. Coin <sup>1</sup>

<sup>1</sup>Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane, QLD, 4072, Australia

<sup>2</sup>International Potato Center, P.O. Box 1558, Lima 12, Peru

<sup>3</sup>International Potato Center, P.O. Box 22274, Kampala, Uganda

<sup>4</sup>Plant Pathology and Plant-Microbe Biology Section, School of Integrative Plant Science, Cornell University, Geneva, NY, 14456, USA

<sup>5</sup>Data61, CSIRO, Ecosciences Precinct, Brisbane, QLD, 4102, Australia

<sup>6</sup>Boyce Thompson Institute, Cornell University, Ithaca, NY, 14853, USA

<sup>7</sup>Department of Horticulture, North Carolina State University, Raleigh, North Carolina, 27695, USA

**v2** First published: 05 Sep 2018, 2:41  
<https://doi.org/10.12688/gatesopenres.12856.1>

Latest published: 21 Jul 2020, 2:41  
<https://doi.org/10.12688/gatesopenres.12856.2>

**Abstract**

**Background:** The chloroplast (cp) genome is an important resource for studying plant diversity and phylogeny. Assembly of the cp genomes from next-generation sequencing data is complicated by the presence of two large inverted repeats contained in the cp DNA.

**Methods:** We constructed a complete circular cp genome assembly for the hexaploid sweetpotato using extremely low coverage (<1×) Oxford Nanopore whole-genome sequencing (WGS) data coupled with Illumina sequencing data for polishing.

**Results:** The sweetpotato cp genome of 161,274 bp contains 152 genes, of which there are 96 protein coding genes, 8 rRNA genes and 48 tRNA genes. Using the cp genome assembly as a reference, we constructed complete cp genome assemblies for a further 17 sweetpotato cultivars from East Africa and an *I. triloba* line using Illumina WGS data. Analysis of the sweetpotato cp genomes demonstrated the presence of two distinct subpopulations in East Africa. Phylogenetic analysis of the cp genomes of the species from the Convolvulaceae *Ipomoea* section *Batatas* revealed that the most closely related diploid wild species of the hexaploid sweetpotato is *I. trifida*.

**Conclusions:** Nanopore long reads are helpful in construction of cp genome assemblies, especially in solving the two long inverted repeats. We are generally able to extract cp sequences from WGS data

**Open Peer Review**

Reviewer Status

	Invited Reviewers				
	1	2	3	4	5
<b>version 2</b> (revision) 21 Jul 2020					
			report	report	report
<b>version 1</b> 05 Sep 2018					
	report	report			

1. **Jun Yang** , Chinese Academy of Sciences, Shanghai, China


**Mengxiao Yan**, Chinese Academy of Sciences, Shanghai, China

2. **Aureliano Bombarely** , Virginia Polytechnic Institute and State University, Blacksburg, USA

of sufficiently high coverage for assembly of cp genomes. The cp genomes can be used to investigate the population structure and the phylogenetic relationship for the sweetpotato.

### Keywords

chloroplast, sweetpotato, genome assembly, Oxford Nanopore sequencing, Illumina sequencing, phylogenetic analysis, Convolvulaceae Ipomoea

3. **Yuki Monden** , Graduate School of Environmental and Life Science, Okayama University, Okayama, Japan

4. **Qiang Li** , Chinese Academy of Agricultural Sciences, Xuzhou, China

5. **Zongyun Li**, Institute of Integrative Plant Biology, School of Life Science, Jiangsu Normal University, Xuzhou, Jiangsu 221116, China; Jiangsu Key Laboratory of Phylogenomics and Comparative Genomics, Jiangsu Normal University, Jiangsu, China

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding author:** Lachlan J. M. Coin ([l.coin@imb.uq.edu.au](mailto:l.coin@imb.uq.edu.au))

**Author roles:** **Zhou C:** Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation; **Duarte T:** Data Curation, Writing – Review & Editing; **Silvestre R:** Data Curation, Resources, Writing – Review & Editing; **Rossel G:** Data Curation, Resources, Writing – Review & Editing; **Mwanga ROM:** Funding Acquisition, Project Administration, Writing – Review & Editing; **Khan A:** Funding Acquisition, Project Administration, Writing – Review & Editing; **George AW:** Supervision, Writing – Review & Editing; **Fei Z:** Data Curation, Funding Acquisition, Project Administration, Writing – Review & Editing; **Yencho GC:** Funding Acquisition, Project Administration, Writing – Review & Editing; **Ellis D:** Project Administration, Resources, Writing – Review & Editing; **Coin LJM:** Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** The project was financially supported by Bill & Melinda Gates Foundation (OPP1052983).  
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2020 Zhou C *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Zhou C, Duarte T, Silvestre R *et al.* **Insights into population structure of East African sweetpotato cultivars from hybrid assembly of chloroplast genomes [version 2; peer review: 2 approved, 3 approved with reservations]** Gates Open Research 2020, 2:41 <https://doi.org/10.12688/gatesopenres.12856.2>

**First published:** 05 Sep 2018, 2:41 <https://doi.org/10.12688/gatesopenres.12856.1>

**REVISED Amendments from Version 1**

A number of misspellings and incorrect claims were revised following the reviewers' comments. Some misleading and confusing sentences were corrected in the new version. [Figure 1d](#) has been slightly modified to remove red dots indicating SNPs. The figure legend has been modified accordingly.

**Any further responses from the reviewers can be found at the end of the article**

**Introduction**

The chloroplast (cp) genome has been widely used to study the phylogeography, molecular systematics and the population genetics for plants<sup>1,2</sup>. The chloroplast DNA (cpDNA) usually displays uniparental inheritance and represents a relatively high degree of conservation in genome structure and gene content<sup>2</sup>. There are over 800 complete cp sequences available for a wide variety of plants from National Center for Biotechnology Information (NCBI) repository ranging in size from 107 to 218 Kb<sup>3</sup>. The cp genomes usually contain 110–130 protein encoding genes (PEGs), about 30 transfer RNA (tRNA) genes and four ribosomal RNA (rRNA) genes, primarily participating in the process of photosynthesis<sup>3,4</sup>. The cpDNA typically forms a circular quadripartite structure with two inverted repeats (IRs), IRA and IRB, separated by one large single-copy section (LSC) and one small single-copy section (SSC)<sup>5</sup>.

The first cpDNA was sequenced from tobacco (*Nicotiana tabacum*) using the bacterial artificial chromosome (BAC) sequencing method in 1986<sup>6</sup>. The two IRs were cloned separately in order to distinguish between them. A plethora of cpDNA had since been sequenced with similar methods<sup>7–9</sup>. Besides BAC sequencing, an alternative strategy used to sequence cpDNA is whole-cp-genome amplification by rolling-circle amplification (RCA) technology<sup>10–12</sup>. However, both approaches require complicated library preparation.

The development of next-generation sequencing (NGS) technologies such as Illumina and Roche 454 facilitate faster and cheaper methods to sequence cp genomes<sup>13–15</sup>. The output of the NGS technologies is short reads of size up to a few hundred base pairs. It is difficult to assemble cp genome with short reads only, especially because of the two large IRs of tens of kilobase pairs. In order to solve this problem, a reference cp genome, normally from a related species, is usually used to anchor the contigs assembled from the short reads<sup>4,16</sup>. The long reads generated from the third-generation sequencing (TGS) technologies, such as the single-molecule real-time (SMRT) PacBio sequencing and Oxford Nanopore sequencing, can also be used to anchor the contigs and solve the repetitive regions. It is even possible to assemble cp genomes directly from long reads<sup>17</sup>. However, as the sequencing error rate of the long reads from the TGS is typically higher than 10%, it is important to introduce an error correction step to guarantee an accurate genome assembly<sup>18</sup>. The high-quality NGS short reads can be integrated for error correction to improve accuracy<sup>19,20</sup>.

The aforementioned methods to construct cp genomes from NGS or TGS data assume pure cpDNA were sequenced. More precisely, the cpDNA were isolated from the nuclear DNAs and other organelle DNAs before sequencing<sup>4,13–16</sup>. However, whole-genome sequencing (WGS) data generated from NGS or TGS technologies always contains cp sequences at various levels determined by the tissue type and library preparation. Normally we are able to gain enough coverage of cp genome for assembly even from low coverage WGS data. There have since been several studies describing assembly of cp genomes from WGS data<sup>21–27</sup>. Extraction of cp sequences from the WGS data plays a key role in these methods. The most straightforward idea is to use a reference cp genome. The cp sequences could be extracted by examining the mapping results of the WGS data to the reference cp genome<sup>21,22</sup>. An alternative strategy relies upon the fact that there are many more copies of the cpDNA than the nuclear DNA and that from other organelles. The entire WGS data is assembled to construct contigs. Contigs that represent significantly higher coverages are treated as cp contigs<sup>23–25</sup>. NOVOPlasty adopted a seed-and-extend paradigm, where the seed could be a cp read sequence, a conserved gene or a cp genome from a related species<sup>26</sup>. The start and the end of a given seed sequence are iteratively extended with reads that are overlapped with the seed until the circular genome is formed. Izan *et al.* proposed a K-mer frequency-based selection of cpDNA sequences from WGS data, which was integrated into a reference free cp genome assembler for non-model species<sup>27</sup>.

Sweetpotato (*Ipomoea batatas*) ranks among the ten most important food crops worldwide<sup>28</sup>. The total annual production is more than 100 million metric tonnes grown on about 8.6 million hectares around the world in year 2016<sup>29</sup>. Understanding the sweetpotato genomes is of significant importance to achieve the full potential of the sweetpotato<sup>30</sup>. Sweetpotato is a hexaploid ( $2n=6x=90$ ) with genome size estimated to be between 2,200 to 3,000 Mb<sup>28</sup>. Due to the complex genome structure, the availability of sweetpotato genomic resources is lacking. Under these circumstances, the cp genome provides researchers with an easy and efficient way to study sweetpotato<sup>4,16,31,32</sup>. A number of cp genomes from the genus *Ipomoea* have been sequenced<sup>4,16,33,34</sup>. Most of them are diploid wild relatives of the sweetpotato. The genome size is around 161 Kb, and the structure represents a standard quadripartite circular with a LSC of 87 Kb, a SSC of 12 Kb and two IRs of 31 Kb<sup>4</sup>. The cp genomes were mainly used to perform phylogenetic analyses<sup>4,16,34</sup>.

In the present study, we constructed a complete cp genome assembly for the hexaploid sweetpotato cultivar Tanzania<sup>35</sup> using long reads produced by the Oxford Nanopore sequencing technology. Despite the  $<1\times$  genome coverage, we obtained approximately 270 $\times$  data coverage for the cp genome. Illumina sequencing data was integrated to improve the accuracy of the genome assembly. Using the Tanzania cp genome assembly as a reference, we constructed 19 cp genomes for a further 17 sweetpotato cultivars (including a duplicate for one cultivar) and an *I. triloba* line from paired-end whole genome Illumina sequence data. The assembled sweetpotato cp genomes were combined to perform phylogenetic analysis to investigate the population structure

of 18 East African sweetpotato cultivars. Putting together the assembled cp genomes and nine publicly available cp genomes of the sweetpotato and its wild relatives, we performed a phylogenetic analysis to investigate the phylogenetic relationship for species in Convolvulaceae *Ipomoea* section *Batatas*.

## Results

### Extraction of cp genome sequence from whole genome sequencing data

We generated high-coverage (60×) 150 bp paired-end Illumina WGS data, and low-coverage (<1×) Oxford Nanopore WGS data on a single cultivar, referred to as Tanzania<sup>35</sup> (Methods). The cultivar Tanzania was used as one of the parents to develop an F1 outcrossing mapping population (B×T) in the Genomic Tools for Sweetpotato (GT4SP) Improvement Project<sup>30</sup>. Approximately 162,000 Nanopore reads and 1.46 billion Illumina reads were generated (Supplementary Table 1). A total of 6,710 Nanopore reads were identified for cp genome by mapping to 30 publicly available cp genomes of the species from the Convolvulaceae *Ipomoea* family<sup>4,16,33,36</sup> (Methods, Supplementary Table 2). The total size is ~43.9 Mb, which represents ~270× data coverage for the cp genome. The longest read is ~30 Kb, and the average size is ~6.5 Kb (Supplementary Figure 1). We identified approximately 45 million Illumina reads for cp genome by mapping to the publicly available cp genomes summing to ~6.2 Gb, which were used for error correction for Nanopore reads and the genome assembly. The other parent for the B×T F1 outcrossing mapping population, Beauregard, was subject to whole genome sequencing at 60× coverage (Methods). A total of approximately 1.3 billion 150 bp Illumina reads were generated summing to ~164 Gb, of which approximately 52 million reads were identified as cp sequences with a total size of ~7.2 Gb (Supplementary Table 1). We performed Illumina WGS at 30× coverage for a further 16 sweetpotato cultivars—Wagabolige and New Kawogo<sup>35</sup>, Ejumula and SPK004<sup>37</sup>, NASPOT 1 and NASPOT 5<sup>38</sup>, NASPOT 7 and NASPOT 10 O<sup>39</sup>, NK259L and NASPOT 11<sup>40</sup>, Huarmeyano, Dimbuka-Bukulula and NASPOT 5/58<sup>41</sup>, Resisto<sup>42</sup>, Magabali<sup>43</sup> and Mugande<sup>44</sup>. These cultivars were used as the parental genotypes in the Mwanga Diversity Panel (MDP) which is an 8×8 diallele diversity mating panel constructed by the GT4SP project for genomic selection of the sweetpotato. While the great majority of these sweetpotato cultivars were from East African countries including Uganda and Kenya, Resisto was from USA and Huarmeyano was from Peru (Supplementary Table 3). We have duplicate samples for the cultivar NASPOT 10 O—one was from the screen-house while the other one was from the field. These two NASPOT 10 O samples were analysed separately in this research (Methods). On average, a total of approximately 75 million 251 bp reads were generated for each sample. The number and the total size of the cp reads extracted from the whole genome sequence data, on average, are ~4.4 million and ~1 Gb respectively for each sample (Supplementary Table 1).

We performed Illumina whole genome sequencing at 50× coverage for the *I. triloba* line, NCNSP-0323<sup>30</sup> (Methods). The raw whole genome sequence data consists of approximately 196 million 150bp reads summing to ~29 Gb. We extracted

approximately 13 million reads for the cp genome from the raw sequence data summing to ~2 Gb (Supplementary Table 1).

### Cp genome assembly for the sweetpotato cultivar Tanzania

We combined the Nanopore long reads with Illumina short reads to construct a cp genome assembly for the sweetpotato cultivar Tanzania (Methods). After trimming off the low-quality bases, approximately 2.2 Gb Illumina sequence data remained which was used for error correction for the Nanopore reads with Nanocorr (Supplementary Table 1). A total of 70 low quality Nanopore reads were removed after error correction and the total size reduced to approximately 43.2 Mb (Figure 1a), which was used to construct a draft genome assembly using Canu. The resulting genome assembly of approximately 218 Kb consists of three contigs of size 46 Kb, 39 Kb and 132 Kb, respectively. Compared to the published sweetpotato cp genome, the assembly is split at the boundaries of the two IRs (Figure 1b). Utilizing the overlap information between the contigs, the AMOS minimus combined the three contigs and generated a single contig of ~183 Kb (Figure 1c) (Methods). The contig contains a ~20 Kb redundancy at the ends which was removed after circularization (Figure 1d). The circularized contig is ~161 Kb, and is highly collinear with the reference cp genome assembly (Figure 1d). Application of Pilon further identified and corrected 42 single-nucleotide polymorphisms (SNPs) and small indels. To follow the paradigm of the published cp genomes, we restructured the genome assembly so that it starts from the LSC (Methods). The final genome assembly consists of a single circular contig of 161,274 bp (Figure 1e).

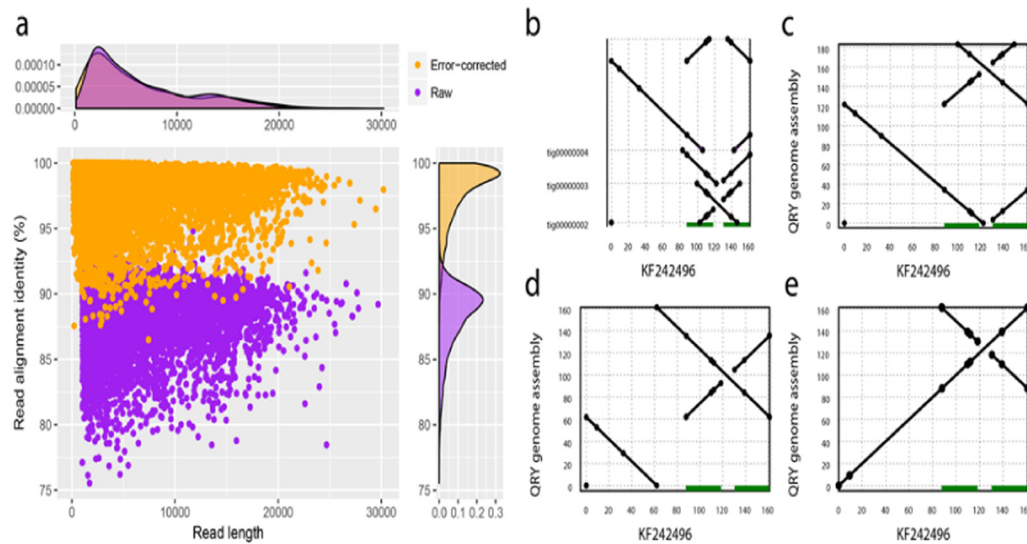
### Cp genome assembly for the other 17 sweetpotato cultivars and the *I. triloba* line NCNSP-0323

The cp sequence data was subjected to quality control before assembled with SPAdes (Methods). After trimming off the low-quality regions, the total sizes of the sequence data of the 19 samples range from approximately 267 Mb to 2.67 Gb (Supplementary Table 1). The contigs generated from SPAdes for the 19 samples vary in numbers and sizes: the minimum number of contigs is 76 for the cultivar NASPOT 7, while the maximum number is 197 for the cultivar Beauregard; and the total sizes of the genome assemblies range from ~169 Kb (cultivars Ejumula and NASPOT 7) to ~229 Kb (cultivar NK259L) (Supplementary Table 4). The SPAdes contigs were then mapped to the Tanzania cp genome assembly for anchoring (Methods). The resulting genome assemblies for the 19 samples are very similar. The largest and the smallest genome assembly is 161,509bp and 161,198bp, derived from the cultivar NASPOT 5 and Beauregard, respectively (Supplementary Table 4).

### Molecular structure and gene content of the sweetpotato cp genome

The gene annotation of the cp genome assembly of the sweetpotato cultivar Tanzania was generated with the web tool DOGMA and further refined with MUSCLE (Methods). The circular plot of the gene annotation is depicted in Figure 2. The sweetpotato cp genome represents a common circular structure with two IRs (IRA and IRB) separating one LSC and





**Figure 1. Assembly of the Tanzania chloroplast (cp) genome.** (a) Dot plot of the Nanopore read length versus the alignment identity to reference assembly. The read alignment identity is defined as  $I = M/L$ , where  $M$  is the total number of base pairs of the exact match and  $L$  is the size of the alignment span on the reference genome. The reference genome is the 30 cp genomes downloaded from the NCBI (Supplementary Table 2). The alignment was performed with BWA MEM<sup>45</sup>. The alignment identities were calculated from the Cigar string. The purple and yellow represents before and after error correction with Illumina reads using Nanocorr<sup>20</sup>, respectively. (b) Dot plot of the reference cp genome versus the contigs produced by Canu<sup>17</sup>. (c) Dot plot of the reference cp genome versus the contigs produced by AMOS minimus<sup>46</sup> after merging Canu contigs. (d) Dot plot of the reference cp genome versus the contigs produced by AMOS minimus after circularization. (e) Dot plot of the reference cp genome versus the final cp genome assembly which was polished with Illumina reads using Pilon<sup>19</sup> and fixed the start at the LSC. For (b–e), the cp genome assembly of the *I. trifida* was used as the reference (accession number REM 753, Genbank accession number KF242496)<sup>16</sup>. The green bars on the x-axis indicate positions of the two IRs.

one SSC<sup>5</sup>. The size of the IRA, IRB, LSC and SSC is 30,874, 30,835, 87,489 and 12,076 bp, respectively. The overall GC content of the sweetpotato cp genome is 37.54%. The GC contents in different regions are highly variable. The two IRs represent significantly higher GC content than the single-copy regions: for the LSC and SSC, the GC content is 36.14% and 32.20%, respectively, whereas for the two IRs, the GC content is 40.57%. This is mainly caused by the high GC content ribosomal RNA genes in IR regions, including *rrn16*, *rrn23*, *rrn4.5* and *rrn5* (Figure 2). We identified 152 genes in the cp genome of which there are 96 protein encoding genes (PEGs), eight rRNA genes and 48 tRNA genes. Table 1 shows a full list of the functional genes. As we can see, the genes can be divided into 16 functional systems. The number of single-copy and double-copy genes is 71 and 11, respectively, and there is one triple-copy gene (*rps12*). The results are highly similar to what has been reported for the cultivar Xushu 18 cp genome<sup>4</sup>; the only difference is that the *psbZ* gene is not found in the cultivar Xushu 18 cpDNA while the *ihbA* gene is not found in the cultivar Tanzania cpDNA. It should be noted that the double-copy gene *ycf1* was not reported for the cultivar Xushu 18 cp genome<sup>4</sup>, but this was actually a miss-annotation.

#### Phylogenetic analysis of the sweetpotato cp genome

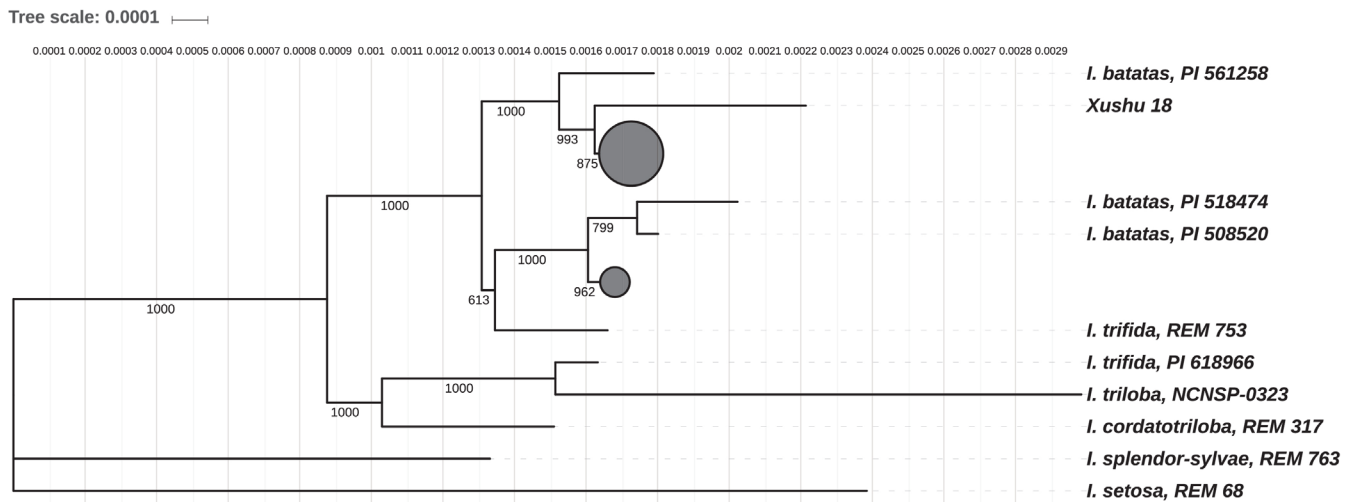
We performed a phylogenetic analysis for the Convolvulaceae *Ipomoea* section *Batatas* on the basis of the 19 cp genomes of the sweetpotato (*I. batatas*) and the cp genome of the *I. triloba* line NCNSP-0323 assembled in this research, coupled with nine publicly available cp genomes, of which, four of them are

for sweetpotato and two of them are for *I. trifida* and the other three are for *I. cordatotriloba*, *I. splendor-sylvae* and *I. setosa*, respectively<sup>4,16</sup> (Supplementary Table 2). The resulted phylogenetic tree is depicted in Figure 3. The 18 sweetpotato cultivars used as the parental genotypes for mapping populations in the GT4SP project represent two distinct clades, consisting of 12 and six cultivars, respectively. Here, the length of any branch in a clade is no greater than  $2 \times 10^{-4}$  substitutions per bp. The detailed phylogenetic relationship of the 18 sweetpotato cultivars is shown in Figure 4. As we can see, the distance between the two clades is approximately  $5 \times 10^{-4}$  substitutions per bp. In the larger clades, the cultivar Tanzania represents a relatively larger distance ( $2 \times 10^{-4}$  substitutions per bp) compared to the other cultivars. The population structure discovered here is similar to the one revealed by using simple sequence repeat primers by David *et al.* with the exception of the classification of the sweetpotato cultivars NK259L, Resisto and Mugande<sup>47</sup> (Supplementary Table 3). For the publicly available sweetpotato cp genomes, PI 561258 and Xushu 18 are closely related to the larger clade, while PI 518474 and PI 508520 have a closer relationship with the smaller clade (Figure 3). The diploid wild relative of the hexaploid sweetpotato, *I. trifida* (REM 753), displays a significantly closer relationship to the *I. batatas* compared to the other species in the Convolvulaceae *Ipomoea* section *Batatas*. The other *I. trifida* accession PI 618966, however, represents a much larger diversity to the *I. batatas* and shows a close relationship to the *I. triloba* line NCNSP-0323 assembled in this research. Interestingly, the accession PI 618966 was originally identified as *I. triloba* and was recently reidentified

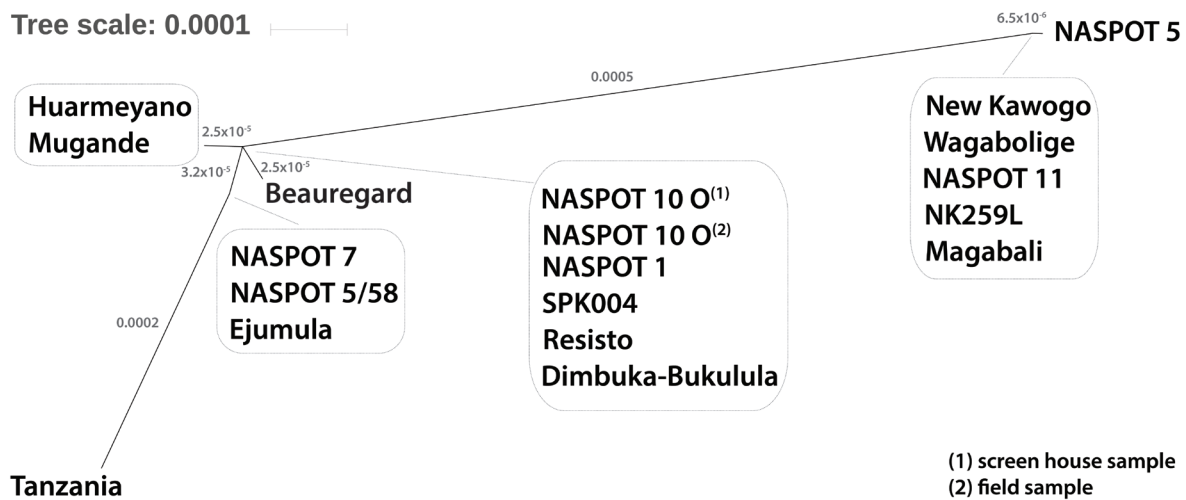


**Table 1. List of annotated genes.** The functional systems were adopted from the OGDRAW<sup>50</sup>. Bracketed superscripts represent number of copies.

Functional system	Number	Gene list
Photosystem I	7	<i>psaA, psaB, psaC, psal, psaJ, ycf3, ycf4</i>
Photosystem II	15	<i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ</i>
Cytochrome b/f complex	6	<i>petA, petB, petD, petG, petL, petN</i>
ATP synthase	6	<i>atpA, atpB, atpE, atpF, atpH, atpI</i>
NADH dehydrogenase	13	<i>ndhA, ndhB<sup>[2]</sup>, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH<sup>[2]</sup>, ndhI, ndhJ, ndhK</i>
RubisCO large subunit	1	<i>rbcL</i>
C-type cytochrome synthesis	1	<i>ccsA</i>
RNA polymerase	4	<i>rpoA, rpoB, rpoC1, rpoC2</i>
Ribosomal proteins (LSU)	9	<i>rpl2, rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36</i>
Ribosomal proteins (SSU)	16	<i>rps2, rps3, rps4, rps7<sup>[2]</sup>, rps8, rps11, rps12<sup>[2]</sup>, rps14, rps15<sup>[2]</sup>, rps16, rps18, rps19</i>
Maturase K	1	<i>matK</i>
Acetyl-CoA carboxylase carboxyltransferase	1	<i>accD</i>
Clp protease proteolytic subunit	1	<i>clpP</i>
Chloroplast envelope membrane protein	1	<i>cemA</i>
ORFs	6	<i>orf188<sup>[2]</sup>, orf42<sup>[2]</sup>, orf56<sup>[2]</sup></i>
Hypothetical chloroplast RF	8	<i>ycf1<sup>[2]</sup>, ycf15<sup>[2]</sup>, ycf2<sup>[2]</sup>, ycf68<sup>[2]</sup></i>



**Figure 3. A phylogenetic tree of the Convolvulaceae *Ipomoea* section *Batatas* on the basis of chloroplast genomes.** The numbers on the branches are bootstrap support values. The branches shorter than  $2 \times 10^{-4}$  substitutions per bp were collapsed resulting two clades consisting of 12 and 6 sweetpotato cultivars represented by a big and small solid circle respectively in the plot. The plot was generated with iTOL<sup>52</sup>.



**Figure 4. A phylogenetic tree of the East African sweetpotato cultivars used in the GT4SP project on the basis of chloroplast genomes.** This is a fine-scale representation of the two clades in Figure 3. The numbers on the branches are branch lengths given in terms of substitutions per bp.

the long reads generated from Oxford Nanopore sequencing. Nanopore reads proved to be extremely powerful in assembling the cp genome, especially in solving long repetitive regions. The sweetpotato cp genome contains two ~31 Kb IRs, which is very difficult for short-read *de novo* assemblers. With the overlapping information from long reads; however, the problem can be easily resolved. Canu<sup>17</sup> provides a useful tool set for assembling Nanopore reads, which was used in this research. It is worth noting that although the average depth of coverage of the whole sweetpotato genome is less than 1x, we obtained enough coverage of the cp genome for assembly.

Although long reads are powerful in solving complex genome structures, the error-prone nature of the raw reads necessitates an extra error-correction step. Illumina reads have been widely used to assist long read error-correction<sup>19,20</sup>. The Illumina read-based correction could be performed either on the raw long reads before assembling<sup>20</sup> or on the draft genome assembly constructed from the raw reads<sup>19</sup>. In the current study, we did both. Before assembling with Canu, the Nanopore reads were corrected with Illumina reads using Nanocorr<sup>20</sup>. After assembling, the draft genome assembly was polished with Illumina reads using Pilon<sup>19</sup> (Methods). With several pipelines examined, we found that to perform error correction both before and after assembling is the best practice to construct the sweetpotato cp genome.

Assembling the cp genome from the short Illumina reads is challenging owing to the two large IRs. Since the structure of the cp genome is generally stable, reference genomes from the closely related species are usually used to perform reference-based assembling<sup>4,16</sup>. In this study, we used the genome assembly constructed from the Nanopore reads as reference to assemble cp genomes for a further 19 cp genomes including 17 sweetpotato cultivars (including a duplicate for one sample) and the *I. triloba* line NCNSP-0323. SPAdes<sup>53</sup> was used as the

*de novo* short-read assembler. The contigs generated by SPAdes were fragmented as expected. Among the 19 genome assemblies, the minimum number of contigs was 76. As the two IRs are highly homologous, there was generally only one copy of repetitive regions being assembled. In order to solve this problem, for reference-based scaffolding, we reused some single-copy contigs from the two IR regions to construct complete cp genome assemblies.

The molecular structure and gene content of the cpDNA are relatively conserved in land plants<sup>2</sup>. Many cpDNAs form a circular quadripartite structure with two IRs separated by one large and one small single-copy section<sup>2,5</sup>. All 20 cp genome assemblies constructed in this research represent this common structure. The size of the two IRs of the sweetpotato cpDNA is approximately 31 Kb each, and is much larger than the other plants such as potato<sup>10</sup>, rice<sup>54</sup>, wheat<sup>55</sup>, and maize<sup>56</sup>, of which the IRs are usually smaller than 26 Kb. This is highly likely due to gene losses in these species. By comparing the gene annotation of the sweetpotato cpDNA in this study (Figure 2) to the potato cpDNA<sup>10</sup>, we can see that, in the potato cpDNA, the boundary region of the IRA and SSC harbors a deletion of approximately 6Kb involved in the genes, *ycf1*, *rsp15* and *ndhH*. Meanwhile, these three genes are presented in the symmetric boundary region of the IRB and SSC, which explains why the size of the IRs of the potato cpDNA is approximately 6 Kb smaller than the sweetpotato cpDNA.

The cpDNA usually has uniparental inheritance and undergoes low rates of substitution and recombination, which makes it well suited for phylogenetic analysis. The cp genome has been widely used to perform phylogenetic or comparative analysis in previous studies<sup>2,10,16</sup>. In this research, we used the complete cp genome assemblies to study the phylogenetic relationship of the 18 sweetpotato potato cultivars used as the parental



genotypes for mapping populations in the GT4SP project, as well as the species from the Convolvulaceae *Ipomoea* section *Batatas*. The sweetpotato genotypes from the GT4SP project were classified into two distinct clusters, which guarantees the diversities of mapping populations derived from them. The phylogenetic analysis clearly revealed that the *I. trifida* is the most closely related diploid wild relatives to the hexaploid sweetpotato, *I. batatas*, which is consistent with conclusions from the previous studies<sup>32,57</sup>.

Almost all whole genome sequencing data contains cp sequences, from which we are usually able to obtain cp genome sequences of enough data coverage for *de novo* assembly. As we can see, all the cp genome assemblies described in this research were constructed using whole genome sequencing data. Given that the cp genome is an important resource for studying plant genomes and whole genome data has gradually become indispensable in modern genome projects, it will be a good practice to construct the cp genome assembly to gain a first insight into the plant genome we are trying to understand before moving to the complex nuclear genome.

## Methods

### Genome sequencing of the MDP parental genotypes

The 16 sweetpotato cultivars used as the parental genotypes for the MDP diversity panel were subjected to whole genome sequencing. These sweetpotato cultivars were collected from Uganda, Kenya, USA and Peru, and included Wagabolige, New Kawogo, Ejumula, SPK004, NASPOT 1, NASPOT 5, NASPOT 7, NASPOT 10 O, NK259L, NASPOT 11, Huarmeyano, Dimbuka-Bukulula, NASPOT 5/58, Resisto, Magabali and Mugande (Supplementary Table 3). Leaf tissue was ground to a fine powder using the FastPrep-24™ 5G tissue homogenizer (MP Biomedicals, Santa Ana, California) and DNA extracted from the leaf tissues following published protocols with modifications<sup>58,59</sup>. Briefly, tissue was suspended in pre-warmed (65°C) CTAB buffer (200mM Tris-CL, 50mM EDTA, 2M NaCl, 2% CTAB and 3%  $\beta$ -mercapto-ethanol), mixed and heated at 65°C for 45 min prior to extraction with chloroform:isoamyl alcohol (24:1) and precipitated with sodium acetate and ethanol. Paired-end genomic libraries were prepared using the Illumina's Genomic DNA Sample Preparation kit and sequenced on the Illumina HiSeq 2500 system with paired-end mode and read length of 251 bp (Illumina, San Diego, CA).

### 10x Genomics' Chromium sequencing of the sweetpotato cultivar Tanzania and Beauregard

The genomic DNA of Tanzania and Beauregard were extracted using the method cetyltrimethyl ammonium bromide and purified with 1x Agencourt AMPure XP beads (Beckman Coulter), according to manufacturer's instructions. Before the library preparation, 1.5  $\mu$ g purified gDNA was size selected using the BluePippin instrument (Sage Science) with the 0.75% Agarose Dye free, Marker U1 High-pass 30–40 kb vs3 protocol followed by a purification step with 0.4x AMPure XP beads. The library preparations for these two samples were done following the Chromium™ Genome Reagent Kits user guide (CG00022, Rev C). In summary, 10 ng of sample DNA was used to generate Gel Bead-In-Emulsions (GEM) in the Chromium™ Controller (10x Genomics) followed by isothermal incubation, post GEM

incubation cleanup and quality control (QC). Libraries were constructed with end-repair and A-tailing, adaptor ligation, post ligation cleanup using SPRIselect Reagent (Beckman Coulter, USA), sample index PCR, post PCR cleanup, and QC. We modified the protocol by increasing the number of PCR cycles to nice and adding 105  $\mu$ l SPRIselect reagent for the Post Sample Index PCR Cleanup, which resulted in the recovery of shorter fragments than it was expected. The libraries were sequenced using the HiSeq X Ten platform (Illumina, San Diego, CA).

### Oxford Nanopore sequencing of the sweetpotato cultivar Tanzania

Before the MinION library preparation, 5.7  $\mu$ g Tanzania pure DNA was size selected (start selection size: 8Kb) with the same protocol used in 10x Genomics' Chromium sequencing. The size selected gDNA was purified with 1x AMPure XP beads. The resulting 950 ng of Tanzania gDNA was used in MinION sequencing library preparation with the SQK-LSK108 1D ligation Sequencing kit (May 2017 version). We modified the protocol as follows: 30 min incubation each end-repair step and adapter ligation; 10 min incubation at RT in the end-repair purification step; 0.7x AMPure XP beads used after adapters ligation and ELB buffer (Oxford Nanopore Technologies) warmed up at 50°C previously to use and incubation of the eluted solution at 50°C. A library of 348 ng was loaded into a FLO-MIN106 (R.9.4 version) flowcell used in a MK1B MinION. We run the 1D protocol in the MinKnow software (version 1.5.18) and we basecalled the raw data using Albacore (version 1.1.0).

### Cp genome sequence extraction

WGS data were aligned to 30 publicly available cp genome assemblies of the species from the *Ipomoea* family<sup>4,16,33,36</sup> (Supplementary Table 2) to extract cp genome reads, using BWA MEM<sup>45</sup> (version 0.7.15). We used the option '-x ont2d' for Nanopore reads, and default options for Illumina reads. For each Nanopore read, the alignment records with at least 500 bp sequence aligned were selected to calculate the total length of the alignment. A Nanopore read was considered as a cp sequence if at least 1 Kb and 80% of the read aligned. A similar strategy was employed for Illumina reads extraction. Both of the two reads of a read pair were required to be aligned. The minimum size of the alignment block was set to 100 bp.

### Cp genome assembly from Nanopore data

We used Nanocorr<sup>20</sup> (version 0.01) to perform error correction for Nanopore reads using the Illumina reads. In order to guarantee the quality of Illumina reads, Trimmomatic<sup>60</sup> (version 0.36) was used to remove the low quality regions. We imposed the quality score of each base pair to be no less than 20 and the length of the reads no less than 100. The corrected Nanopore reads were then used to construct a draft genome assembly with Canu<sup>17</sup> (version 1.5). As the resulting draft genome assembly contained more than one contig, AMOS minimus<sup>46</sup> (version 3.1.0) was used to remove the redundancy and concatenate contigs using the overlap information. The AMOS minimus was also used to circularize the contig. We aligned the Illumina reads to the circularized contig and corrected the SNPs and small indels with Pilon<sup>19</sup> (version 1.22). In order to follow the paradigms of the published cp genomes, we aligned the genome assembly to the

published cp genomes with [MUMMER](#)<sup>61</sup> (version 3.23) to find homology regions, and let the genome assembly start from the LSC.

### Cp genome assembly from Illumina Hiseq data

The low quality regions of the extracted cp sequences were removed with [Trimmomatic](#)<sup>60</sup> (version 0.36). The minimum quality score of each base pair was set to 20 and the minimum length of the reads was set to 100. [SPAdes](#)<sup>53</sup> (version 3.10.1) was used to construct contigs from Illumina reads. We excluded the repeat resolve module from SPAdes and used the contigs before repeat resolution as it consistently missed one of the two IRs. The resulting genome assembly contains tens to hundreds of contigs. The size of the contigs ranged from several hundred base pairs to tens of kilobase pairs. Since we know the structure of cp genome is generally stable, the syntenic relationship was used for scaffolding. We mapped the SPAdes contigs to the genome assembly resulting from the Nanopore reads using [BWA MEM](#)<sup>45</sup>. The alignments were used to order the contigs. The overlap information between the neighbouring contigs was used to concatenate them.

### Cp genome annotation

We used the web tool [Dual Organellar GenoMe Annotator](#) (DOGMA)<sup>48</sup> to generate the preliminarily gene annotations. For each particular gene, we used [MUSCLE](#)<sup>49</sup> (version 3.8.31) to align the genuine protein sequences of the gene gained from the [NCBI GenBank](#) to the genome assembly to decide the exact boundary positions. The web tool [Organellar Genome DRAW](#) (OGDRAW)<sup>50</sup> was used to generate the circular annotation plot of the genome assembly. The hypothetical cp open-reading frame *ycf1* was not identified by DOGMA initially. It was added to the annotation on the basis of the MUSCLE alignment results.

### Phylogenetic analysis

Phylogenetic analysis was performed on the 18 sweetpotato cultivars used as the parental genotypes for constructions of mapping populations in GT4SP project as well as the Convolvulaceae *Ipomoea* section *Batatas* including the cp genome assemblies constructed in this research and nine publicly

available cp genome assemblies. [MAFFT](#)<sup>62</sup> (version 7.310) was employed to perform the multiple sequence alignment (MSA) for cp genomes. The phylogenetic structure was constructed with [PhyML](#)<sup>63</sup> (version 3.1). Branch certainty was evaluated with 1000 replications of bootstrap resampling. The phylogenetic tree depicted in this research was constructed with the web tool [iTOL](#) (version 4)<sup>52</sup>.

### Data availability

#### Underlying data

Nanopore and Illumina reads and the cp genome assemblies are deposited at NCBI BioProject repository, accession number PRJNA438020: <http://identifiers.org/bioproject/PRJNA438020>.

#### Extended data

**Supplementary Figure 1. Size distribution of the Nanopore sequencing data of the total DNA.** <https://doi.org/10.26188/12652034.v2><sup>64</sup>

**Supplementary Table 1. Statistics of the chloroplast (cp) sequencing data.** <https://doi.org/10.26188/12652067.v1><sup>65</sup>

**Supplementary Table 2. List of the 30 publicly available Ipomoea chloroplast (cp) genomes in the NCBI repository.** <https://doi.org/10.26188/12652079.v1><sup>66</sup>

**Supplementary Table 3. Description of the parental genotypes of the Mwanga Diversity Panel (MDP).** <https://doi.org/10.26188/12652085.v1><sup>67</sup>

**Supplementary Table 4. Statistics of the chloroplast (cp) genome assemblies of the 18 sweetpotato cultivars and the *I. triloba* line NCNSP-0323.** <https://doi.org/10.26188/12652094.v1><sup>68</sup>

### Acknowledgements

We would like to thank Rick Miller for his valuable suggestions on the phylogenetic analysis.

## References

- Martin W, Rujan T, Richly E, *et al.*: **Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus.** *Proc Natl Acad Sci U S A*. 2002; **99**(19): 12246–12251. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Shaw J, Lickey EB, Schilling EE, *et al.*: **Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III.** *Am J Bot*. 2007; **94**(3): 275–288. [PubMed Abstract](#) | [Publisher Full Text](#)
- Daniell H, Lin CS, Yu M, *et al.*: **Chloroplast genomes: diversity, evolution, and applications in genetic engineering.** *Genome Biol*. 2016; **17**(1): 134. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yan L, Lai X, Li X, *et al.*: **Analyses of the complete genome and gene expression of chloroplast of sweet potato [*Ipomoea batata*].** *PLoS One*. 2015; **10**(4): e0124083. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sandelius AS, Aronsson H: **The chloroplast: interactions with the environment.** (Springer Science & Business Media). 2008; **13**. [Publisher Full Text](#)
- Shinozaki K, Ohme M, Tanaka M, *et al.*: **The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression.** *EMBO J*. 1986; **5**(9): 2043–2049. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bausher MG, Singh ND, Lee SB, *et al.*: **The complete chloroplast genome sequence**

- of *Citrus sinensis* (L.) Osbeck var 'Ridge Pineapple': organization and phylogenetic relationships to other angiosperms. *BMC Plant Biol.* 2006; 6: 21. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Samson N, Bausher MG, Lee SB, *et al.*: The complete nucleotide sequence of the coffee (*Coffea arabica* L.) chloroplast genome: organization and implications for biotechnology and phylogenetic relationships amongst angiosperms. *Plant Biotechnol J.* 2007; 5(2): 339–353. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  9. Saski C, Lee SB, Fjellheim S, *et al.*: Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. *Theor Appl Genet.* 2007; 115(4): 591. [PubMed Abstract](#) | [Publisher Full Text](#)
  10. Daniell H, Lee SB, Grevich J, *et al.*: Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. *Theor Appl Genet.* 2006; 112(8): 1503–18. [PubMed Abstract](#) | [Publisher Full Text](#)
  11. Daniell H, Wurdack KJ, Kanagaraj A, *et al.*: The complete nucleotide sequence of the cassava (*Manihot esculenta*) chloroplast genome and the evolution of *atp* in Malpighiales: RNA editing and multiple losses of a group II intron. *Theor Appl Genet.* 2008; 116(5): 723–37. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  12. Saski C, Lee SB, Daniell H, *et al.*: Complete chloroplast genome sequence of *Gycine max* and comparative analyses with other legume genomes. *Plant Mol Biol.* 2005; 59(2): 309–322. [PubMed Abstract](#) | [Publisher Full Text](#)
  13. Moore MJ, Dhingra A, Soltis PS, *et al.*: Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol.* 2006; 6: 17. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  14. Wu J, Liu B, Cheng F, *et al.*: Sequencing of chloroplast genome using whole cellular DNA and solexa sequencing technology. *Front Plant Sci.* 2012; 3: 243. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  15. Cronn R, Liston A, Parks M, *et al.*: Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* 2008; 36(19): e122. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  16. Eserman LA, Tiley GP, Jarret RL, *et al.*: Phylogenetics and diversification of morning glories (tribe Ipomoeae, Convolvulaceae) based on whole plastome sequences. *Am J Bot.* 2014; 101(1): 92–103. [PubMed Abstract](#) | [Publisher Full Text](#)
  17. Koren S, Walenz BP, Berlin K, *et al.*: Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 2017; 27(5): 722–736. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  18. Chin CS, Alexander DH, Marks P, *et al.*: Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods.* 2013; 10(6): 563–569. [PubMed Abstract](#) | [Publisher Full Text](#)
  19. Walker BJ, Abeel T, Shea T, *et al.*: Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014; 9(11): e112963. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  20. Goodwin S, Gurtowski J, Ethel-Sayers S, *et al.*: Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res.* 2015; 25(11): 1750–1756. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  21. Soomri A, Haak D, Zaitlin D, *et al.*: Organelle\_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. *BMC Genomics.* 2017; 18(1): 49. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  22. Wang W, Messing J: High-throughput sequencing of three *Lemnoideae* (duckweeds) chloroplast genomes from total DNA. *PLoS One.* 2011; 6(9): e24670. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  23. Kim K, Lee SC, Lee J, *et al.*: Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. *Sci Rep.* 2015; 5: 15655. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  24. Antipov D, Hartwick N, Shen M, *et al.*: plasmidspades: assembling plasmids from whole genome sequencing data. *bioRxiv.* 2016; 048942. [Publisher Full Text](#)
  25. Zhang T, Zhang X, Hu S, *et al.*: An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. *Plant Methods.* 2011; 7: 38. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  26. Dierckxsens N, Mardulyn P, Smits G: NOVOPlasty: *de novo* assembly of organellar genomes from whole genome data. *Nucleic Acids Res.* 2017; 45(4): e18. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  27. Izan S, Esselink D, Visser RGF, *et al.*: *De Novo* Assembly of Complete Chloroplast Genomes from Non-model Species Based on a K-mer Frequency-Based Selection of Chloroplast Reads from Total DNA Sequences. *Front Plant Sci.* 2017; 8: 1271. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  28. Loebenstein G, Thottappilly G: *The sweetpotato*. (Springer Science & Business Media). 2009. [Publisher Full Text](#)
  29. Faostat F: *Agriculture organization of the united nations statistics division (2014)*. [Review date: April 2015]. 2016. [Reference Source](#)
  30. Mwangi ROM, Andrade MI, Carey EE, *et al.*: Sweetpotato (*Ipomoea batatas* L.). (Springer International Publishing, Cham). 2017; 181–218. [Publisher Full Text](#)
  31. Roullier C, Rossel G, Tay D, *et al.*: Combining chloroplast and nuclear microsatellites to investigate origin and dispersal of New World sweet potato landraces. *Mol Ecol.* 2011; 20(19): 3963–3977. [PubMed Abstract](#) | [Publisher Full Text](#)
  32. Huang J, Sun M: 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. *Theor Appl Genet.* 2000; 100(7): 1050–1060. [Publisher Full Text](#)
  33. Hoshino A, Jayakumar V, Nitasaka E, *et al.*: Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*. *Nat Commun.* 2016; 7: 13295. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  34. Muñoz-Rodríguez P, Carruthers T, Wood JR, *et al.*: Reconciling conflicting phylogenies in the origin of sweet potato and dispersal to Polynesia. *Curr Biol.* 2018; 28(8): 1246–1256.e12. [PubMed Abstract](#) | [Publisher Full Text](#)
  35. Mwangi ROM, Odongo B, p'Obwoya CO, *et al.*: Release of five sweetpotato cultivars in uganda. *HortScience.* 2001; 36(2): 385–386. [Publisher Full Text](#)
  36. McNeal JR, Kuehl JV, Boore JL, *et al.*: Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. *BMC Plant Biol.* 2007; 7: 57. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  37. Mwangi ROM, Odongo B, Niringiye C, *et al.*: Release of two orange-fleshed sweetpotato cultivars, 'spk004' ('kakamega') and 'ejumula', in uganda. *HortScience.* 2007; 42(7): 1728–1730. [Publisher Full Text](#)
  38. Mwangi ROM, Odongo B, Turyamureeba G, *et al.*: Release of six sweetpotato cultivars ('naspot 1' to 'naspot 6') in uganda. *HortScience.* 2003; 38(3): 475–476. [Publisher Full Text](#)
  39. Mwangi ROM, Odongo B, Niringiye C, *et al.*: 'naspot 7', 'naspot 8', 'naspot 9 o', 'naspot 10 o', and 'dimbuka-bukulula' sweetpotato. *HortScience.* 2009; 44(3): 828–832. [Reference Source](#)
  40. Mwangi ROM, Niringiye C, Alajo A, *et al.*: 'naspot 11', a sweetpotato cultivar bred by a participatory plant-breeding approach in uganda. *HortScience.* 2011; 46(2): 317–321. [Publisher Full Text](#)
  41. Mwangi ROM, Kyalo G, Ssemakula GN, *et al.*: 'naspot 12 o' and 'naspot 13 o' sweetpotato. *HortScience.* 2016; 51(3): 291–295. [Publisher Full Text](#)
  42. Jones A, Dukes PD, Schalk JM, *et al.*: 'resisto' sweet potato. *HortScience.* 1983; 18(2): 251–252. [Reference Source](#)
  43. Gruneberg WJ, Ma D, Mwangi ROM, *et al.*: Advances in sweetpotato breeding from 1992 to 2012. (CABI International). 2015. [Publisher Full Text](#)
  44. Abidin PE, van Eeuwijk FA, Stam P, *et al.*: Adaptation and stability analysis of sweet potato varieties for low-input systems in uganda. *Plant Breed.* 2005; 124(5): 491–497. [Publisher Full Text](#)
  45. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25(14): 1754–1760. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  46. Sommer DD, Delcher AL, Salzberg SL, *et al.*: Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics.* 2007; 8: 64. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  47. David MC, Diaz FC, Mwangi ROM, *et al.*: Gene pool subdivision of east african sweetpotato parental material. *Crop Science.* 2018; 58(6): 2302–2314. [Publisher Full Text](#)
  48. Wyman SK, Jansen RK, Boore JL: Automatic annotation of organellar genomes with DOGMA. *Bioinformatics.* 2004; 20(17): 3252–3255. [PubMed Abstract](#) | [Publisher Full Text](#)
  49. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004; 32(5): 1792–1797. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
  50. Lohse M, Drechsel O, Bock R: OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr Genet.* 2007; 52(5–6): 267–274. [PubMed Abstract](#) | [Publisher Full Text](#)
  51. Li Q, Li Y, Song J, *et al.*: High-accuracy *de novo* assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy. *New Phytol.* 2014; 204(4): 1041–1049. [PubMed Abstract](#) | [Publisher Full Text](#)

52. Letunic I, Bork P: **Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees.** *Nucleic Acids Res.* 2016; **44**(W1): W242–W245.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
53. Bankevich A, Nurk S, Antipov D, *et al.*: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.** *J Comput Biol.* 2012; **19**(5): 455–477.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
54. Tang J, Xia H, Cao M, *et al.*: **A comparison of rice chloroplast genomes.** *Plant Physiol.* 2004; **135**(1): 412–420.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
55. Gogniashvili M, Naskidashvili P, Bedoshvili D, *et al.*: **Complete chloroplast DNA sequences of zanduri wheat (*Triticum* spp.).** *Genet Resour Crop Evol.* 2015; **62**(8): 1269–1277.  
[Publisher Full Text](#)
56. Strittmaier G, Kössel H: **Cotranscription and processing of 23S, 4.5S and 5S rRNA in chloroplasts from *Zea mays*.** *Nucleic Acids Res.* 1984; **12**(20): 7633–7647.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
57. Srisuwan S, Sihachakr D, Siljak-Yakovlev S: **The origin and evolution of sweet potato (*Ipomoea batatas* Lam.) and its wild relatives through the cytogenetic approaches.** *Plant Sci.* 2006; **171**(3): 424–433.  
[PubMed Abstract](#) | [Publisher Full Text](#)
58. Dellaporta SL, Wood J, Hicks JB: **A plant DNA miniprep: version ii.** *Plant Mol Biol Report.* 1983; **1**(4): 19–21.  
[Publisher Full Text](#)
59. Mace ES, Buhariwalla KK, Buhariwalla HK, *et al.*: **A high-throughput DNA extraction protocol for tropical molecular breeding programs.** *Plant Mol Biol Report.* 2003; **21**(4): 459–460.  
[Publisher Full Text](#)
60. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics.* 2014; **30**(15): 2114–2120.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
61. Kurtz S, Phillippy A, Delcher AL, *et al.*: **Versatile and open software for comparing large genomes.** *Genome Biol.* 2004; **5**(2): R12.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
62. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol.* 2013; **30**(4): 772–780.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
63. Guindon S, Dufayard JF, Lefort V, *et al.*: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol.* 2010; **59**(3): 307–321.  
[PubMed Abstract](#) | [Publisher Full Text](#)
64. Zhou C: **Nanopore read length distribution.** University of Melbourne. Figure. 2020.  
<http://www.doi.org/10.26188/12652034.v2>
65. Zhou C: **Statistics of the chloroplast genome sequencing data.** University of Melbourne. Dataset. 2020.  
<http://www.doi.org/10.26188/12652067.v1>
66. Zhou C: **List of the 30 publicly available *Ipomoea* chloroplast genomes in the NCBI repository.** University of Melbourne. Dataset. 2020.  
<http://www.doi.org/10.26188/12652079.v1>
67. Zhou C: **Description of the parental genotypes of the Mwanga Diversity Panel (MDP).** University of Melbourne. University of Melbourne. Dataset. 2020.  
<http://www.doi.org/10.26188/12652085.v1>
68. Zhou C: **Statistics of the chloroplast genome assemblies of the 18 sweetpotato cultivars and the *I. triloba* line NCNSP-0323.** University of Melbourne. Dataset. 2020.  
<http://www.doi.org/10.26188/12652094.v1>



## Open Peer Review

Current Peer Review Status: ? ? ✓ ? ✓

### Version 2

Reviewer Report 07 October 2020

<https://doi.org/10.21956/gatesopenres.14357.r29627>

© 2020 Li Z. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



#### Zongyun Li

Institute of Integrative Plant Biology, School of Life Science, Jiangsu Normal University, Xuzhou, Jiangsu 221116, China; Jiangsu Key Laboratory of Phylogenomics and Comparative Genomics, Jiangsu Normal University, Jiangsu, China

Summary of the key results:

The authors sequenced and assembled 19 cp genomes of sweetpotato cultivars and one wild species using Oxford Nanopore and Illumina sequencing. With the published data, the authors constructed a phylogeny tree and proposed that *I. trifida* is the most closely related diploid species of the sweetpotato. In addition, 18 sweetpotato cp genomes demonstrated the presence of the two distinct subpopulations in East Africa.

Overall evaluation:

This manuscript provided more information for sweetpotato genome resources which is very important for us to learn more about the genetic diversity, origin, and evolution of sweetpotato. According to the contents of results and discussion, some comments were listed as follows:

1. Park *et al.* (2018<sup>1</sup>) and Sun *et al.* (2019<sup>2</sup>) also published some *Ipomoea* cp genomes. It would be better to cite their works.
2. There are more genes annotated in this article than in other studies (Eserman *et al.*, 2014<sup>3</sup>; Yan *et al.*, 2015<sup>4</sup>; Park *et al.*, 2018<sup>1</sup>; Sun *et al.*, 2019<sup>2</sup>). Can the authors discuss the reasons? Which genes were not annotated in previous studies?
3. I thought one of the highlights of this paper was the assembly of 18 sweetpotato cp genomes. The authors demonstrated the presence of the two distinct subpopulations in East Africa using these cp genomes. However, no other more detailed analysis and discussion about these 18 cp genomes. Would it be better to add more detailed sequences analysis? For example, the factors which impact the genome size, some specific loci



between the two subpopulation, etc.

Overall, this manuscript is valuable for indexing.

### References

1. Park I, Yang S, Kim WJ, Noh P, et al.: The Complete Chloroplast Genomes of Six Ipomoea Species and Indel Marker Development for the Discrimination of Authentic Pharbitidis Semen (Seeds of *I. nil* or *I. purpurea*). *Front Plant Sci.* 2018; **9**: 965 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Sun J, Dong X, Cao Q, Xu T, et al.: A systematic comparison of eight new plastome sequences from *Ipomoea* L. *PeerJ.* 2019; **7**: e6563 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Eserman LA, Tiley GP, Jarret RL, Leebens-Mack JH, et al.: Phylogenetics and diversification of morning glories (tribe Ipomoeae, Convolvulaceae) based on whole plastome sequences. *Am J Bot.* 2014; **101** (1): 92-103 [PubMed Abstract](#) | [Publisher Full Text](#)
4. Yan L, Lai X, Li X, Wei C, et al.: Analyses of the complete genome and gene expression of chloroplast of sweet potato [*Ipomoea batata*]. *PLoS One.* 2015; **10** (4): e0124083 [PubMed Abstract](#) | [Publisher Full Text](#)

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 07 October 2020

<https://doi.org/10.21956/gatesopenres.14357.r29625>

© 2020 Li Q. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Qiang Li**

Xuzhou Institute of Agricultural Sciences in Jiangsu Xuhuai District, Key Laboratory for Biology and Genetic Breeding of Sweetpotato (Xuzhou), Ministry of Agriculture, Sweet Potato Research Institute (SPRI), Chinese Academy of Agricultural Sciences, Xuzhou, China

Overall evaluation:

I think that the manuscript is acceptable. It gives us a new method to insight into the population structure of plants, especially sweetpotato with complicated genome. However, I have a number of suggestions to the article.

#### 1. Materials:

1. Just the author's response to Prof. Aureliano Bombarely, "the primary focus of this study was to investigate the population structure of the East African sweetpotato cultivars used in the GT4SP project", 16 cultivars for MDP and 2 cultivars (Tanzania and Beauregard) for constructing F1 mapping population are suitable for this research. The *I. triloba* was selected. You can give us reasons why you chose this wild species, not *I. trifida*, or other wild relatives.
2. The authors assembled cp genomes 2 times of cultivar NASPOT 10 O - one was from the screen-house while the other one was from the field. What was the aim? You can give us more detail about the difference or consistency between them.
3. There are total 20 samples, including 19 samples from 18 sweetpotato cultivars (2 samples of the cultivar NASPOT 10 O), and 1 sample of wild relative. I always confuse the number in the article, cultivars or samples.

#### 2. Methods:

1. "In the present study, we constructed a complete cp genome assembly for the hexaploid sweetpotato cultivar Tanzania using long reads produced by the Oxford Nanopore sequencing technology." Is this a new method? Is it the first report? It is an efficient tool to assemble complicated genomes in my mind. It is an important part of this article. I suggest the method should be reflected in the title.
2. Give more detail in the discussion part about this new tool compared to other methods.

#### 3. Results:

1. "The sweetpotato cp genome of 161,274 bp contains 152 genes, of which there are 96 protein coding genes, 8 rRNA genes and 48 tRNA genes..." this is cp genome of Tanzania. There are a little difference among other cultivars.
2. Suggest to compare the phylogenetic tree by using cp genome data and nuclear

genome data, and to validate the method.

4. Others:

1. "The only difference is that the *psbZ* gene is not found in the cultivar Xushu 18 cpDNA while the *ihbA* gene is not found in the cultivar Tanzania cpDNA. It should be noted that the double-copy gene *ycf1* was not reported for the cultivar Xushu 18 cp genome<sup>4</sup>, but this was actually a miss-annotation." The difference between Xushu 18 and Tanzania are *psbZ* and *ihbA*, why the *ycf1* was actually a miss-annotation. You should give more information about it.
2. Just a suggestion from Dr. Yang, the important references should be added.

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Reviewer Report 05 October 2020

<https://doi.org/10.21956/gatesopenres.14357.r29622>

© 2020 Monden Y. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Yuki Monden** 

Graduate School of Environmental and Life Science, Okayama University, Okayama, Japan

I think this paper is well-written, and the authors have properly revised the manuscript by referring to the comments of the reviewer. This paper performed a complete circular cp genome assembly using NGS and TGS technologies in the hexaploid sweetpotato. Phylogenetic analysis using the cp genomes revealed that there are two distinct clusters of sweetpotato in East Africa and *I. trifida* is the most closely related diploid wild species to the *I. batatas* hexaploid sweetpotato. The results of this paper provide insights into the genetic relationships and the population structure of the species from the Convolvulaceae *Ipomoea* section *Batatas*. Besides, despite the complexity of the cp genomes by the presence of two large inverted repeats, this research demonstrates the possibility of building the cp genomes using extremely low coverage (<1x) Oxford Nanopore WGS data combined with Illumina short reads. Other comments are shown below.

1. Table 1:

The copy number of rps12 gene should be three, but the bracketed superscript of this gene is two. Please make sure.

2. Phylogenetic analysis using nuclear genomes

Is it possible to compare the results of phylogenetic analyses based on the cp genomes and the nuclear genomes using the same materials? I think such comparative analysis should provide new insight into evolutionary dynamics on cp and nuclear genomes of *Ipomoea* species. Do you have any plans for such work?

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Plant genetics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

---

**Version 1**

Reviewer Report 20 November 2018

<https://doi.org/10.21956/gatesopenres.13938.r26743>

© 2018 Bombarely A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Aureliano Bombarely**

Department of Horticulture, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

**Summary of the key results**

The manuscript titled “Insights into population structure of East African sweetpotato cultivars from hybrid assembly of chloroplast genomes” presents the chloroplast genome assembly and annotation of the sweet potato cultivar Tanzania and its comparison with seventeen other cultivars and the different species *Ipomoea triloba*. The final genome assembly of the Tanzania cultivar was 161,274 bp. No major findings were reported except for the deletion of the gene *psbZ* in the cultivar Xushu 18 and *ihbA* gene in the Tanzania. The phylogenetic analysis of these cultivars and nine publicly available *Ipomoea* chloroplast pointed that the diploid *I. trifida* species is close related to the hexaploid sweet potato (*I. batatas*) than other *Ipomoea* species.

**Overall evaluation**

This manuscript presents the assembly and analysis of the chloroplast genome of the sweet potato cultivar Tanzania. The *I. batatas* chloroplast genome was already published in 2015 by Yan et. al. (PLoS One 10:4) so the novelty of the results presented in this manuscript are limited from the point of new of a “new” chloroplast genome. The use of ON for the sequencing of the Tanzania cultivar and the addition of the resequencing data of seventeen other cultivars potentially could add some interesting findings. Nevertheless, the analysis that the authors performed failed in the development of attractive results. Personally, I would propose several analysis that it may help to increase the impact of the manuscript:

- Improved phylogenetic analysis partitioning the alignments per gene (or in bins) and using a Bayesian framework (e.g. BEAST). The phylogenetic analysis could include the dating of the divergency time of the different taxa.
- Nuclear gene mining. One of the most interesting questions about the polyploids is the origin of those. The use of the resequencing data could potentially derived in the mining of nuclear copy single gene that could help to elucidate a different evolutionary trajectory for these accessions. Additionally, it is interesting the result in which some genes are missing. Maybe they have been transferred to the nuclear genome. It could be interesting that hypothesis.



- Positive selection. Each of the chloroplast nuclear genes could be tested for positive selection using PAML and the Ks/Kn ratio.

In terms of the manuscript organization and writing, I found confusing some parts. For example, the material and methods are not aligned with the results presented in the manuscript. For example, the section "Extraction of cp genome sequence from whole genome sequencing data" describe the chloroplast data mining from ON and Illumina for the Tanzania accession and then for the Beaugregard accession, but the material and methods also describe the use of 10X Genomics Chromium that I am not sure where it comes from. Do the authors used 10X Genomics also? Probably for the genome assembly, the comparison of the Canu assembly with Organelle\_PBA (Soorni et al. 2017) could be interesting, to see if the authors obtain only one contig representing the whole chloroplast.

Overall, I think that the manuscript is okay, but there are some space for improvement in the structure of the manuscript and in the results that are presented.

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 12 Jul 2020

**Lachlan Coin**, Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane, Australia

We thank the reviewer for their thoughtful review of our manuscript.

**Reviewers comment** 1. This manuscript presents the assembly and analysis of the chloroplast genome of the sweet potato cultivar *Tanzania*. The *I. batatas* chloroplast genome was already published in 2015 by Yan *et. al.* (*PLoS One* 10:4) so the novelty of the results presented in this manuscript are limited from the point of new of a “new” chloroplast genome. The use of ON for the sequencing of the *Tanzania* cultivar and the addition of the resequencing data of seventeen other cultivars potentially could add some interesting findings. Nevertheless, the analysis that the authors performed failed in the development of attractive results. Personally, I would propose several analysis that it may help to increase the impact of the manuscript:

- Improved phylogenetic analysis partitioning the alignments per gene (or in bins) and using a Bayesian framework (e.g. BEAST). The phylogenetic analysis could include the dating of the divergency time of the different taxa.
- Nuclear gene mining. One of the most interesting questions about the polyploids is the origin of those. The use of the resequencing data could potentially derived in the mining of nuclear copy single gene that could help to elucidate a different evolutionary trajectory for these accessions. Additionally, it is interesting the result in which some genes are missing. Maybe they have been transferred to the nuclear genome. It could be interesting that hypothesis.
- Positive selection. Each of the chloroplast nuclear genes could be tested for positive selection using PAML and the Ks/Kn ratio.

**Response:** The primary focus of this study was to investigate the population structure of the East African sweetpotato cultivars used in the GT4SP project. We strongly agree that these suggested analyses will largely increase the impact of the manuscript. However, it is difficult to integrate these suggested analyses in the current study within the constraints of the data. We will include these suggestions in the future directions of the project.

**Reviewers Comment:** In terms of the manuscript organization and writing, I found confusing some parts. For example, the material and methods are not aligned with the results presented in the manuscript. For example, the section “Extraction of cp genome sequence from whole genome sequencing data” describe the chloroplast data mining from ON and Illumina for the *Tanzania* accession and then for the *Beauregard* accession, but the material and methods also describe the use of 10X Genomics Chromium that I am not sure where it comes from. Do the authors used 10X Genomics also? Probably for the genome assembly, the comparison of the Canu assembly with Organelle\_PBA (Soorni *et al.* 2017) could be interesting, to see if the authors obtain only one contig representing the whole chloroplast.

**Response:** We have amended the manuscript to address these concerns. 10X Genomics was indeed used to perform the whole genome sequencing for the sweetpotato accessions *Tanzania* and *Beauregard*. However, the linked reads information was not utilized in construction of the cp genome assemblies for them. Instead, the sequence data was simply used as paired-end reads to create contigs. The contigs were then used to construct whole cp genome assembly with a cp reference genome.

**Competing Interests:** Not applicable

Reviewer Report 13 September 2018

<https://doi.org/10.21956/gatesopenres.13938.r26651>

© 2018 Yang J et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Jun Yang** 

State Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Shanghai Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China

**Mengxiao Yan**

Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Plant Science Research Center (SCPSRC), Shanghai Chenshan Botanical Garden, Chinese Academy of Sciences, Shanghai, China

Zhou *et al.* sequenced 16 sweet potato cultivars in GT4SP project supported by Bill & Melinda Gates Foundation. Here authors presented only partial data about chloroplast (cp) genome assemblies.

I found this study will be a good supplementary work of previous publication in *Current Biology*<sup>1</sup>. Due to the lack of awareness of this publication, the claims in the manuscript are incorrect and need to be revised.

In this case, the finding of this study about two distinct cp subpopulations in East Africa cultivars is reasonable.

#### Other comments

- "In recent years, the development of next-generation sequencing (NGS) technologies such as Illumina and Roche 454 facilitate faster and cheaper methods to sequence cp genomes 13-15."  
To my knowledge, the Roche 454 already left the market.
- "By examining the mapping results of the WGS data, we are able to extract cp sequences 21,22."  
We? Who are we?
- "Sweetpotato is a hexaploid ( $2n=6x=90$ ) with genome size estimated to be between 2,200 to 3,000 Mb<sup>28</sup>."  
How about the C-values?
- "Due to the complex genome structure, the availability of sweetpotato genomic resources is lacking."  
We do have a published genome, right?

- "A number of cp genomes from the *Ipomoea* family have been sequenced<sup>16,33</sup>."  
Dose *Ipomoea* family mean genus *Ipomoea*? Or genus *Ipomoea* Series *Batatas*?
- "Most of them are diploid wild relatives of the sweetpotato. To the best of our knowledge, to date, four cp genomes have been completely sequenced for the hexaploid sweetpotato<sup>4,16</sup>; the genome size is around 161 Kb, and the structure represents a standard quadripartite circular with a LSC of 87 Kb, a SSC of 12 Kb and two IRs of 31 Kb<sup>4</sup>. The cp genomes were mainly used to perform phylogenetic analyses<sup>4,16</sup>."  
The mentioned Current Biology paper has provided hundreds of cp genome sequences of sweet potato and its wild relatives.
- "The circularized contig is ~161 Kb, and is highly collinear with the published sweetpotato cp genome assembly (Figure 1d)."  
In Figure 1d, it is an *I. trifida* cp genome, not a published sweet potato cp genome.
- "The sweetpotato cp genome represents a common circular structure with two IRs (IRA and IRB) separating one LSC and one SSC2."  
Where does the 2 in SSC2 come from? Convert into right format if it is a citation.
- "The red dots represent SNPs between the two cp genomes. The green bars on the x-axis indicate positions of the two IRs"  
No red dots there, only black dots.
- "It should be noted that the doublecopy gene *ycf1* was not reported for the cultivar Xushu 18 cp genome<sup>4</sup>"  
Convert into right format if it is a citation.
- "Interestingly, the accession PI 618966 was originally identified as *I. triloba* and was recently reidentified as *I. trifida* by the GRIN National Genetic Resources Program."  
The identification of PI 618966 needs to be checked carefully. All individuals of *I. trifida* formed a monophyletic clade closely related to *I. batatas* according to Current Biology paper. As the progenitor of sweet potato, it's quite strange that *I. trifida* is much closer to other species in Series *Batatas* than *I. batatas*.
- Figure 3 & 4  
It will be much clear to add the tip labels rather than collapsed clades on the tree. Figure 4 will be no more informative in this case.  
If the tree is not that complicated, it is not suggested to collapse the two clades. Since information about the relationship between within-clade sample and out-clade sample is not visible when one collapse clade. This information will not be illustrated in Figure 4. Clades can be labeled in different colors if one wants to highlight the clades.  
Furthermore, it is not clear to me which place each sample nested on in Figure 4.
- "In this study, we used the genome assembly constructed from the Nanopore reads as reference to assemble cp genomes for a further 19 cp genomes including..."  
Misleading sentence, authors do rely on published cp genome rather than *de novo* Nanopore assembly.

- "In order to solve this problem, for reference-based scaffolding, we reused some single-copy contigs from the two IR regions to construct complete cp genome assemblies." In which cultivar(s), did author investigate the influence on the tree structure?

I agree the population structure of East African sweet potato cultivars is important for GT4SP project. Also obviously, the data organization and visualization could be largely improved to meet the indexing standards.

### References

1. Muñoz-Rodríguez P, Carruthers T, Wood JRI, Williams BRM, et al.: Reconciling Conflicting Phylogenies in the Origin of Sweet Potato and Dispersal to Polynesia. *Curr Biol.* 2018; **28** (8): 1246-1256.e12 [PubMed Abstract](#) | [Publisher Full Text](#)

**Is the work clearly and accurately presented and does it cite the current literature?**

Partly

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Yes

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Plant genetics

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

Author Response 12 Jul 2020

**Lachlan Coin**, Institute for Molecular Bioscience, University of Queensland, St Lucia, Brisbane, Australia

We thank the reviewer for their thoughtful review.

**Comment 1.** I found this study will be a good supplementary work of previous publication



in Current Biology (Mu, Pablo, *et al.*, 2018). Due to the lack of awareness of this publication, the claims in the manuscript are incorrect and need to be revised.

**Response:** This publication was cited in the revised version. The incorrect claim was revised (see reviewer's comment 7).

**Comment 2.** "In recent years, the development of next-generation sequencing (NGS) technologies such as Illumina and Roche 454 facilitate faster and cheaper methods to sequence cp genomes<sup>13-15</sup>."

To my knowledge, the Roche 454 already left the market.

**Response:** In the revised version, "In recent years" was deleted to make it more precise.

**Comment 3.** "By examining the mapping results of the WGS data, we are able to extract cp sequences<sup>21,22</sup>."

We? Who are we?

**Response:** In the revised version, this sentence was rewritten to "The cp sequences could be extracted by examining the mapping results of the WGS data to the reference cp genome<sup>21, 22</sup>."

**Comment 4.** "Sweetpotato is a hexaploid ( $2n=6x=90$ ) with genome size estimated to be between 2,200 to 3,000 Mb<sup>28</sup>."

How about the C-values?

**Response:** The nuclear genome size is not the key point of this study. The C-value was not investigated.

**Comment 5.** "Due to the complex genome structure, the availability of sweetpotato genomic resources is lacking."

We do have a published genome, right?

**Response:** Even though there is a sweetpotato reference genome published recently (Yang, Jun, *et al.*, 2017), we think the availability of the sweetpotato genome resources is still lacking.

**Comment 6.** "A number of cp genomes from the *Ipomoea* family have been sequenced<sup>16,33</sup>."

Dose *Ipomoea* family mean genus *Ipomoea*? Or genus *Ipomoea* Series *Batatas*?

**Response:** "*Ipomoea* family" means "genus *Ipomoea*". This was made clear in the revised version.

**Comment 7.** "Most of them are diploid wild relatives of the sweetpotato. To the best of our knowledge, to date, four cp genomes have been completely sequenced for the hexaploid sweetpotato<sup>4, 16</sup>; the genome size is around 161 Kb, and the structure represents a standard quadripartite circular with a LSC of 87 Kb, a SSC of 12 Kb and two IRs of 31 Kb<sup>4</sup>. The cp genomes were mainly used to perform phylogenetic analyses<sup>4, 16</sup>."

The mentioned Current Biology paper has provided hundreds of cp genome sequences of sweet potato and its wild relatives.

**Response:** The claim "To the best of our knowledge, to date, four cp genomes have been completely sequenced for the hexaploid sweetpotato<sup>4, 16</sup>;" was removed in the revised version.

**Comment 8.** “The circularized contig is ~161 Kb, and is highly collinear with the published sweetpotato cp genome assembly (Figure 1d).”

In Figure 1d, it is an *I. trifida* cp genome, not a published sweet potato cp genome.

**Response:** This sentence was corrected as “The circularized contig is ~161 Kb, and is highly collinear with the reference cp genome assembly (Figure 1d).”

**Comment 9.** “The sweetpotato cp genome represents a common circular structure with two IRs (IRA and IRB) separating one LSC and one SSC2.”

Where does the 2 in SSC2 come from? Convert into right format if it is a citation.

**Response:** This was a citation. It was corrected in the revised version.

**Comment 10.** “The red dots represent SNPs between the two cp genomes. The green bars on the x-axis indicate positions of the two IRs”

No red dots there, only black dots.

**Response:** The red dots are hard to see due to the resolution of the image. Since the SNPs are not important in this genome assembly section and are further discussed in the phylogenetic analysis section, the SNPs (red dots) were removed from the figure 1d in the revised version.

**Comment 11.** “It should be noted that the doublecopy gene *ycf1* was not reported for the cultivar Xushu 18 cp genome4”

Convert into right format if it is a citation.

**Response:** This was a citation. It was corrected in the revised version.

**Comment 12.** “Interestingly, the accession PI 618966 was originally identified as *I. triloba* and was recently reidentified as *I. trifida* by the GRIN National Genetic Resources Program.”

The identification of PI 618966 needs to be checked carefully. All individuals of *I. trifida* formed a monophyletic clade closely related to *I. batatas* according to Current Biology paper. As the progenitor of sweet potato, it's quite strange that *I. trifida* is much closer to other species in Series *Batatas* than *I. batatas*.

**Response:** We agree with the reviewer that the identification of PI 618966 needs to be checked carefully. According to the phylogenetic structure identified in this study (Fig. 3), PI 618966 has a closer phylogenetic relationship to *I. triloba* instead of *I. trifida*. However, it was recently reidentified as *I. trifida* by the GRIN National Genetic Resources Program. A further study is required to fully clarify this.

**Comment 13.** Figure 3 & 4

It will be much clear to add the tip labels rather than collapsed clades on the tree. Figure 4 will be no more informative in this case.

If the tree is not that complicated, it is not suggested to collapse the two clades. Since information about the relationship between within-clade sample and out-clade sample is not visible when one collapse clade. This information will not be illustrated in Figure 4. Clades can be labeled in different colors if one wants to highlight the clades.

Furthermore, it is not clear to me which place each sample nested on in Figure 4.

**Response:** The two sweetpotato clades showed in Figure 3 were collapsed since the phylogenetic distances are too small. It will be impossible to see the detail phylogenetic

structure of the two East African sweetpotato subpopulations if incorporate Figure 4 into Figure 3.

**Comment 14.** "In this study, we used the genome assembly constructed from the Nanopore reads as reference to assemble cp genomes for a further 19 cp genomes including..."  
Misleading sentence, authors do rely on published cp genome rather than *de novo* Nanopore assembly.

**Response:** The cp genome assembly of the sweetpotato cultivar *Tanzania* was constructed from the Nanopore reads coupled with a published cp reference genome. The cp genome assembly of the sweetpotato cultivar *Tanzania* from the Nanopore reads was then used as reference to constructed cp genome assemblies for a further 19 cp genomes.

**Comment 15.** "In order to solve this problem, for reference-based scaffolding, we reused some single-copy contigs from the two IR regions to construct complete cp genome assemblies."

In which cultivar(s), did author investigate the influence on the tree structure?

**Response:** Single-copy contigs were reused in construction of cp genome assembly for all cultivars. The genome assembler SPAdes collapsed the contigs from the two IR regions as they are almost identical. In order to construct the whole cp genome, the contigs from two IR regions need to be reused. This has no influence on the tree structure.

**Competing Interests:** Not applicable



Minerva Access is the Institutional Repository of The University of Melbourne

**Author/s:**

Zhou, C; Duarte, T; Silvestre, R; Rossel, G; Mwanga, ROM; Khan, A; George, AW; Fei, Z;  
Yencho, GC; Ellis, D; Coin, LJM

**Title:**

Insights into population structure of East African sweetpotato cultivars from hybrid assembly of chloroplast genomes.

**Date:**

2018

**Citation:**

Zhou, C., Duarte, T., Silvestre, R., Rossel, G., Mwanga, R. O. M., Khan, A., George, A. W., Fei, Z., Yencho, G. C., Ellis, D. & Coin, L. J. M. (2018). Insights into population structure of East African sweetpotato cultivars from hybrid assembly of chloroplast genomes.. Gates Open Res, 2, pp.41-. <https://doi.org/10.12688/gatesopenres.12856.2>.

**Persistent Link:**

<http://hdl.handle.net/11343/272066>

**File Description:**

Published version

**License:**

CC BY