GENOME PROFILING AND ASSEMBLY OF THE NON-MODEL OILSEED CROP *Plukenetia volubilis L.*

Simón S. Villanueva*
Department of Biological Sciences
EAFIT University
Medellín, Colombia
svillanu@eafit.edu.co

Javier Correa Álvarez†
Department of Biological Sciences
EAFIT University
Medellín, Colombia
jcorre38@eafit.edu.co

August 17, 2020

ABSTRACT

Sacha inchi (*Plukenetia volubilis*) is an oilseed crop from the Peruvian's Amazonia that belongs to the spurge (Euphorbiaceae) family. Its oil, rich in antioxidants and polyunsaturated fatty acids, is considered highly valuable for food and cosmetic industries. Here, we sequenced the total DNA of Sacha inchi, using Illumina and Nanopore technologies, and present a reference-quality chloroplast genome and a draft of its nuclear genome. We detected two large inversions on the chloroplast genome not present in the previously reported sequence. Using genome profiling, we found Sacha inchi nuclear genome has a haploid size of 334 Mb, presents a high proportion of repetitive elements (47%), and outstanding heterozygosity (11.67%). Furthermore, its genomic structure suggests Sacha inchi is allotetraploid. Our draft assembly presents a high degree of completeness, both at the gene and sequence level, although the high heterozygosity prevented us to achieve a less fragmented assembly (E-size 168 kb; N50 112 kb) and a closer representation to the haploid genome. The assembled draft genome of *Plukenetia volubilis* will provide a valuable resource for the study of this promising crop, allowing the determination of the genetic bases of key phenotypic traits and enabling the use of genetic engineering in breeding programs to develop new varieties.

***K*eywords** *Plukenetia volubilis* · *Euphorbiaceae* · Chloroplast genome · Nanopore sequencing · Genome profiling · Allopolyploid

1 Introduction

Oils are the most energy-dense plant reserves, supplying humans with many of the calories and essential fatty acids required in our diet. The majority of the plant oils we consume are accumulated in the seeds of crops. World production from oilseed crops was approximately 185 million metric tons of oil in 2019 [1].

Human population growth, increased life expectancy, loss of biodiversity, climate change, and accelerated land degradation represent major challenges for agriculture regarding food safety in the following years[2]. By 2050, 90% of the food production expansion would have to come from expanding crop yields and only 10% of this will come from the use of a larger cultivated area[2]. Thus, there is a need to technify agricultural production systems and the development of genetically improved cultivars is going to be key in this process [3].

Several economically important oilseed crops have been genetically improved, such as soybean, sunflower, rapeseed, olive, palm, castor bean, among others [3]. Each of these crops have a reference genome that serves as a building block to guide genetic engineering, to establish breeding programs and to understand the biological mechanisms of the crop. All of these would be severally jeopardized in the absence of reference genomes.

*First Author
†Advisor

Sacha inchi (*Plukenetia volubilis*), a native plant from the Peruvian Amazonia, is emerging as a promising crop due to its nutritional profile. More than a half of its seed weight is oil[4], which is mostly composed by α-linolenic acid ($37.3 - 50.8\%$) linoleic acid ($33.4 - 41\%$) and oleic acid ($8.41 - 12.5\%$)[4, 5, 6, 7, 8, 9]. tocopherols[4, 6], phenolic[5] and caretonic[6] compounds and phytosferols[6]. This makes Sacha inchi's oil highly valuable for food and cosmetic industries.

Currently there is no reference genome for Sacha inchi, which limits the development of genetically engineered cultivars, wider adoption, and production growth.

Since an important part of the oil biosynthesis metabolism occurs in the plastids, endoplasmatic reticulum and oil bodies[10], it is important to have both chloroplast and nuclear genomes to have a complete genetic landscape of this process.

In 2015, seeking to incorporate oilseed crops (Sacha inchi and castor bean) as an economically viable alternative to illegal ones, the Colombian government along with a consortium of different entities, which included EAFIT university, launched a large project to increase the adoption of these alternative crops. As a continuation of that project, the objective of this work is to sequence the total DNA of *Plukenetia volubilis* and present a draft for both chloroplast and nuclear genomes.

Aiming to be the foundation for future works, this research will allow the exploration of the molecular basis of oil biosynthesis of this non-model plant and will enable more specialized crop development programs.

2 Results

2.1 Data output from sequencing platforms

After gathering of plant material and DNA extraction (see Methods), we proceeded to sequence using Illumina and Nanopore platforms at the High-Throughput Sequencing Facility of the University of North Carolina at Chapel Hill, USA.

Short reads technology Illumina Hiseq 4000 system yielded a total of 622'892508 paired-reads while Oxford Nanopore Technology the two GridION flowcells produced 5'560738 reads, with an N50 of 2 kbp. A possible explanation for the small length of the nanopore reads is the presence of an excessive amount of short fragments of DNA in the library before sequencing, which out-competed the long fragments in pore occupancy, giving a low N50 and yield (7 Gb).

2.2 Chloroplast genome assembly

The chloroplast genome contains genes that code for structural and functional components of the organelle. In most plants, it consists of fairly long circular or linear chromosome ranging from 115 to 165 kb in size[11] in land plants and 160 to 163 Kb in the *Euphorbiaceae* family[12, 13, 14, 15, 16]. Its structure is composed by two copies of a duplicated region arranged as inverted repeats (IRs), separated by a large single-copy (LSC) sequence and a small single-copy (SSC) sequence[17, 18].

Whole genome sequencing (WGS) paired-end reads were pre-assembled using Norgal [19], which takes advantage of the depth difference between nuclear and organelle reads to assemble only high depth sequences. Resulting scaffolds were filtered based on Norgal blast considering: best hit matched to a chloroplast, length (at least 1000 bp), identity (at least 95%) and alignment length (at least 200 bp). The filtered scaffold was used as seed for NOVOPlasty[20], which uses the seed to extract an initial sample of the WGS reads to assemble and start extending from those reads. When it is not able to extend more, NOVOplasty tries to circularize the sequence.

Using this strategy, we reconstruted *Plukenetia volubilis* chloroplast genome as a single circularized contig with a total length of 164111 bp, the longest among *Euphorbiaceae* chloroplasts[12, 13, 14, 15, 16]. The size of the long single copy section (LSC), the inverted repeats (IRs) and short single copy section (SSC) were 89833, 28209 and 17860 bp, respectively.

We confirmed the final quality of the chloroplast genome using Pilon. We trimmed illumina paired-end reads using Trimmomatic [21] and mapped the trimmed reads using BWA-MEM [22] to the generated genome. Even with an average depth of 40560x Pilon did not change the sequence. Furthermore, we mapped nanopore reads using Minimap2 [23] to the chloroplast genome to check its structure. A total of 956508 long reads generated a continuous smooth coverage with average 3917x depth, where the LSC, SSC and IRs could be clearly visualized with tablet[24] software (Supplemental Figure S1).

There was already an unverified *Plukenetia volubilis* chloroplast sequence reported (Genbank acc: MF062253)[25]. Our assembly differs from this sequence in genome size (164111 vs 161733), GC content (35.8% vs 36.2%) and the lengths of the LSC, IRs and SSC. However, MF062253 sequence was scaffolded using *Ricinus communis* chloroplast genome (Genbank acc: NC016736) as a reference [25], altering the natural structure of the genome.

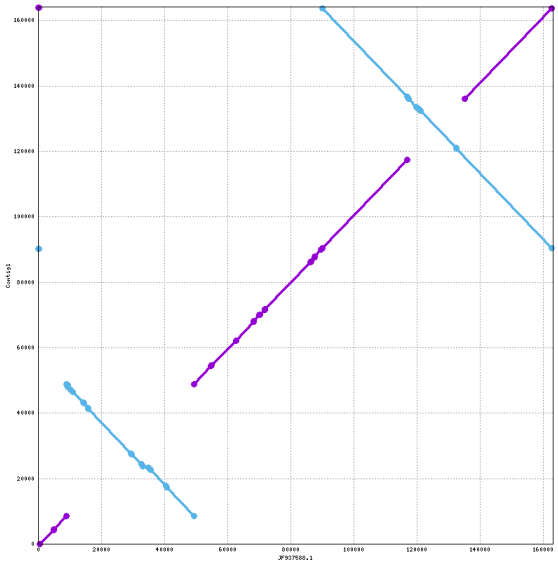


Figure 1: Dotplot of *Plukenetia volubilis* chloroplast genome vs *Ricinus communis* chloroplast genome.

Indeed, using Mummer[26], we identified two large inversions in our assembly when we compared it with *Ricinus communis* chloroplast (Figure 1). The first inversion is located in the middle of the LSC, spanning 39426 bp long. The second inversion is harder to determine as it is not clear if it comprehends only the SSC or if it includes the IRs as well. Both scenarios are indistinguishable from each other using dotplots. These structural variations were not detected previously because they are not present in MF062253 sequence.

2.3 Genome profiling

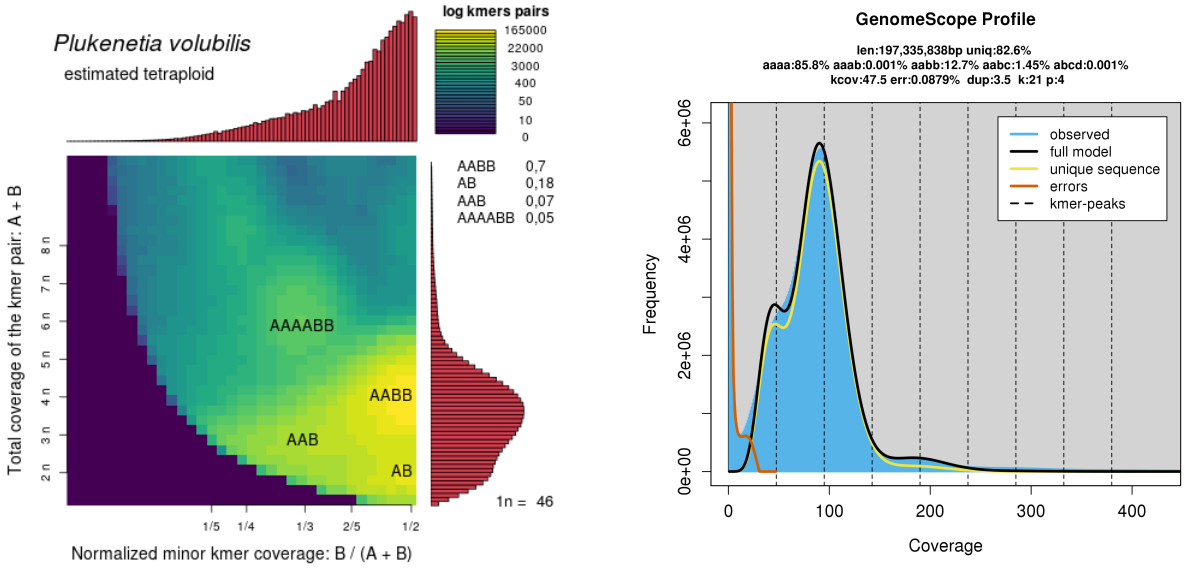
Prior to assembling the nuclear genome, we performed kmer analysis on the raw reads to estimate several of the genome features and get a better understanding of its structure.

Several studies suggest that *Plukenetia volubilis* might be polyploid [27, 28, 29]. Smudgeplot[30] is a novel kmer based analysis tool which uses unique kmer pairs, that differ by one nucleotide, to unravel the genome structure underlying the reads, and plot the different structures in a 2D space. Thus, we used Smudgeplot to determine *Plukenetia volubilis* ploidy (Figure 2a). Also, we used genomescope2 to analyze the kmer profile under the assumption of tetraploidy (Figure 2b). Genomescope fits a mixture model of evenly spaced negative binomial distributions to the kmer spectra to estimate genome size, repetitiveness, and heterozygosity rates[30]. Kmer counts used by both tools were generated using KMC [31].

From Figure 2a, even though the boundaries between the smudges are blurry (due to high coverage variability), it can be seen that most of the heterozygous kmer pairs map around the structure AABB, which suggests tetraploidy. There are another structures present in smaller proportions (which might have arisen from tandem repeats, genes with different copy numbers, transposable elements, among other features.)

Furthermore, normalized minor kmer coverage (x-axis) shows that most of the heterozygous kmer pairs (88%) are clumped at $1/2$, which means they are balanced in a 1 to 1 correspondence. As this proportion is not possible in odd ploidies (1n, 3n, 5n, etc.), this indicates that the ploidy has to be even, specifically either AB, AABB, AAABBB, etc.

Given the high levels of heterozygosity inferred by GenomeScope2 (Figure 2b), a diploid structure AB is unlikely as the high divergence between the alleles would raise questions about how homologous chromosome pairing is maintained, as this process is, in some extent, homology-dependent [32]. Similarly, an hexaploidy (6n) is unlikely as well, considering that a structure AAABBB would be problematic in meiosis due to improper chromosomal pairing, as non-bivalent homologous chromosome pairings are prone to aberrant chromosome segregation[33, 34]. Higher ploidies (8n, 10n,



(a) Smudgeplot with $L = 30, U = 2600$

(b) Tetraploid Genomescope model

Figure 2: Genome profiling for *Plukenetia volubilis*

12n, etc.) would have been manifested as a wider distribution of kmer pairs in the smudges of Figure 2a or more peaks in Figure 2b. Hence, we favor the initial estimation of tetraploidy.

Additionally, Genomescope2 model reports a proportion of kmer pairs AABB (12.7%) much higher than AAAB (0.001%), which is a topology expected for allotetraploids[30]. Therefore, *Plukenetia volubilis* is the result of a hybridization process, between two similar but different ancestors, that occurred at some point of the evolutionary history. The big differences between structures AABB (12.7%) and AABC (1.45 %) suggests that the event is relatively recent.

Genomescope2 model informs about other genome features, which helps to guide genome assembly. *Plukenetia volubilis* nuclear genome presents a relatively small haploid size (197 Mb), a low proportion of repetitive content (17.4%) and a high level of heterozygosity (14.2%). The last one is especially problematic for assembly, as heterozygous regions tend to complicate the assembly graph structure and make difficult to phase the haplotigs[35].

2.4 Genome assembly

Regarding assembly, we explored several strategies combining different assemblers and preprocessing tools. Most of them resulted in highly fragmented assemblies. The following is the detailed approach that led to the first draft of *Plukenetia volubilis* genome (Figure 3).

For Illumina raw data, we first used BBDuk[36] to trim usual adapters used for Truseq sequencing and others detected with FastQC, and filter synthetic molecules such as PhiX (used in Illumina sequencing). Resulting reads were processed by Brownie-corrector[37], which is a targeted error correction tool. Brownie targets reads with a specific sequence (by default, poly-A) and uses the paired-end information to cluster the reads and correct the reads within each cluster.

Unfortunately, Brownie failed to cluster the extracted reads and hence no error correction was performed. The remaining reads were error corrected using Karect[38] and then both groups of reads were merged again. We specifically avoided to correct the reads extracted by Brownie-corrector with Karect, as these regions tend to be overcorrected by global error correction tools that introduce more errors instead[37].

On the other hand, we used Guppy (Oxford Nanopore Technologies) flip-flop model to basecall raw nanopore reads. Resulting reads were processed with Yacrd[39] to reduce chimeras in nanopore reads.

A total of 622'892508 error corrected Ilumina paired reads and 4'699073 of preprocessed nanopore reads were used for assembly.

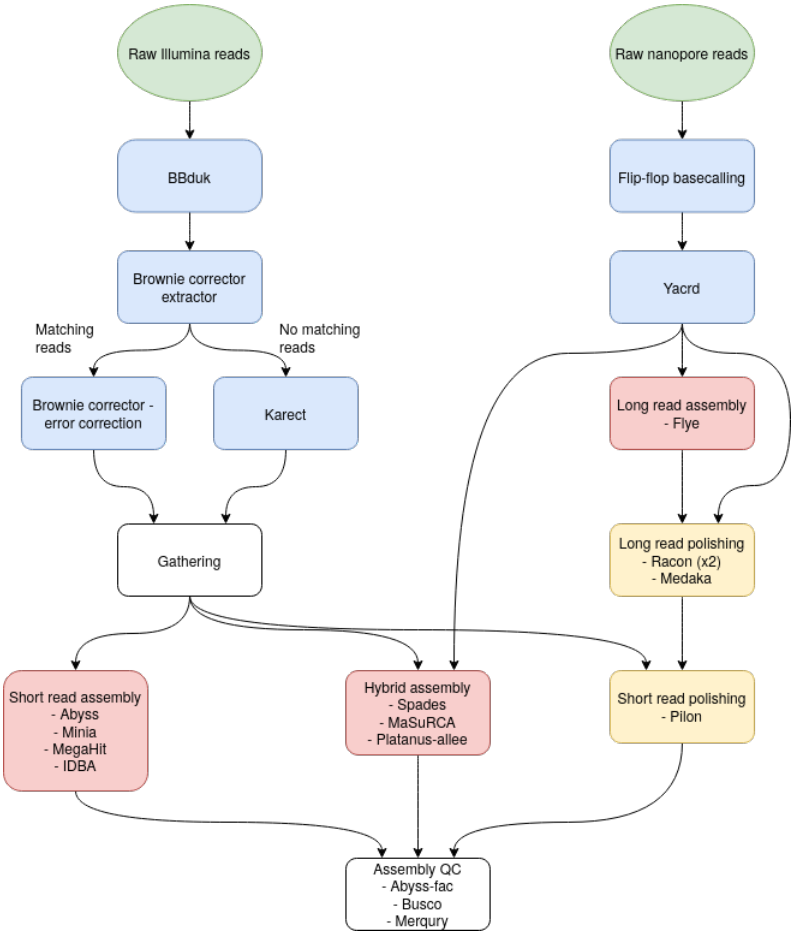


Figure 3: Workflow to assemble *Plukenetia volubilis* nuclear genome. Blue blocks represent preprocessing steps, red blocks are assembly steps and yellow blocks polishing steps

For assembly, we explored three approaches: 1) short-read assembly: we evaluated Abyss, the GATB-minia-pipeline (with no error correction, Minia[40], BESST [41]), Megahit [42] and IDBA-UD[43]; 2) long-read assembly and polishing: given our previous experiences with long-read assemblies (data not shown) we only tested Flye[44], polished with Racon[45] twice and Medaka (`https://github.com/nanoporetech/medaka`) using the nanopore reads (mapped to assembly with Minimap2 [23]), and Pilon[46] using the Illumina reads (mapped to genome with BWA-MEM [22]), in that order; 3) hybrid assembly: we evaluated Spades[47, 48], MaSuRCA [49] and Platanus-allee[50] with both the preprocessed Illumina and ONT reads.

The ideal assembly of a genome would correctly represent the sequence of each of the organism chromosomes from end to end without any gaps or ambiguities inside them. Every assessment of a *de novo* genome assembly tries to measure the deviation of this ideal in some way.

Contiguity aims to encapsulate how fragmented is the assembly using solely descriptive information about the sequences that compose it (the number of sequences on the assembly, how long they are, how many ambiguities are present in them, etc.).

Moreover, as genes are the fundamental units of genomes, an assembly might be evaluated based on its gene content. The main problems with this approach are, on the one hand, that gene prediction and annotation are a computationally intensive task; on the other hand, assessing gene annotations for a *de novo* genome might be tricky, as it is not clear beforehand what is expected to be found. BUSCO[51] circumvents both problems looking in the assembly only for highly conserved single-copy orthologs of a phylogenetically closely related dataset, which are assumed have to be present in the assembly. This fast approach provides a ceiling to the expected genes, a way to estimate gene completeness in *de novo* assemblies, and is widely used in the genomics field.

Besides, kmer-analysis is a widely used technique in bioinformatics. Its use to validate genome assemblies was introduced by Mapleson *et al.* with KAT toolkit [52]. Recently, this approach was further extended by Rhie *et al.* with Merqury [53]. In a nutshell, Merqury compares the kmer-spectra of the assembly with the kmer-spectra of the reads used in the assembly. This enables Merqury to compute concensus quality (QV), kmer-completeness and visualize copy number spectra (which is useful to identify redundancies, collapsed regions, etc).

Thus, we evaluated the quality of the generated assemblies with three different dimensions: contiguity, gene completeness and kmer-analysis. We used abyss-fac from the Abyss assembler to generate contiguity metrics; we also used BUSCO to estimate gene completeness based on eudicotyledons-odb10 dataset; and Merqury to assess kmer completeness and QV comparing the kmer-spectra of assemblies to the kmer spectra of the raw Illumina reads. The main results are summarised in Table 1. Sequences shorter than 500 bp were filtered of assemblies before the quality assessments.

To further inspect the contiguity of the assemblies, we made the N(x) plot presented in Figure 4 with a custom python script. In a N(x) plot, the sequences of an assembly are sorted by length and their cumulative size fraction are plotted against their size in decreasing order.

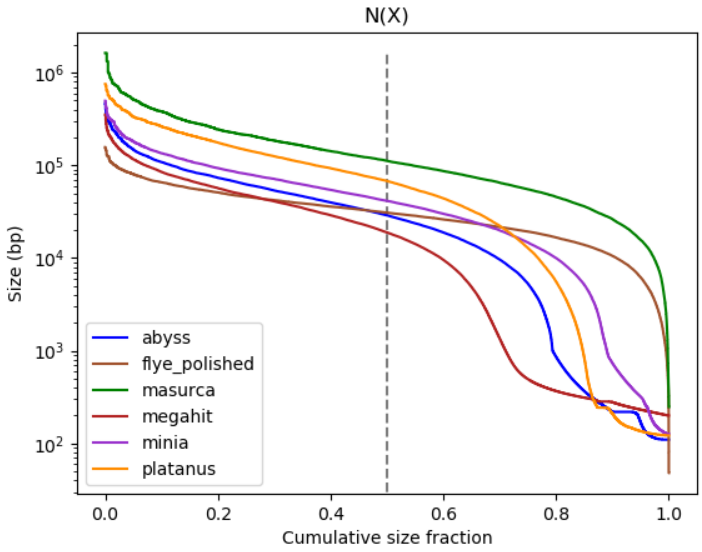


Figure 4: N(x) plots of generated assemblies.

Assembler	E-size[a]	BUSCO Comp. (%)[b]	Kmer Comp. (%)	QV[c]	Total size (Mb)
Flye	35139	47.5	36.5	27.6	301
Abyss	54000	96.1	95.6	57.3	566
Megahit	43780	95.6	96.4	50.5	587
Minia	63974	95.3	96.1	57.5	578
MaSuRCA	168206	96.4	94.7	41.5	596
Platanus-allee	119977	95.3	97.6	56	632

[a] Expected size (calculated by abyss-fac).
[b] Complete BUSCOs (single copy + duplicated).
[c] Phred score value; $QV = -10 \log P$, where P is the probability of error.

Table 1: Quality statistics for assemblies generated by the workflow to assemble *Plukenetia volubilis* nuclear genome. Sequences shorter than 500 bp were filtered before assessments.

We were not able to run Spades, as we failed to provide the resources it requested for our data. Similarly, IDBA did not run by anomalous errors that we did not manage to solve.

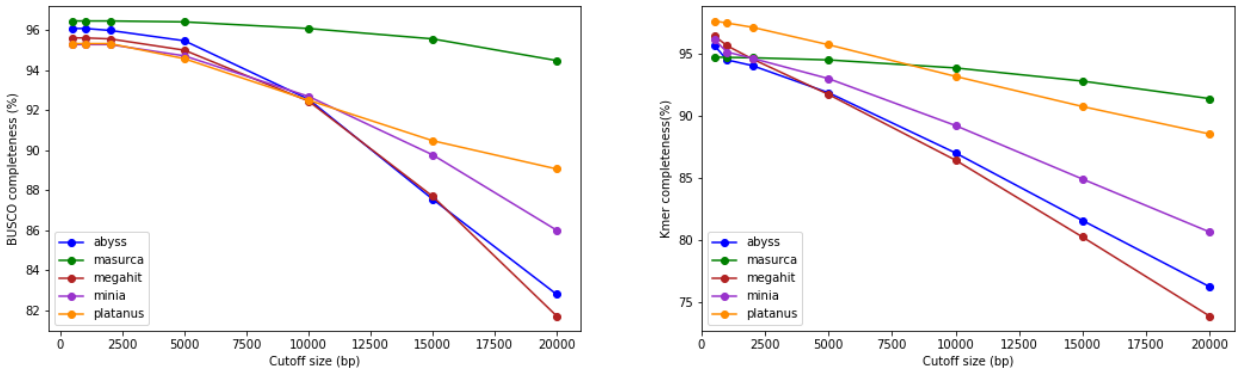
For any assembly, the higher is its curve in Figure 4 the more contiguous it is. The expected size (Table 1) is calculated as the area under this curve and is useful to summarize it in a single metric. Thus, in terms of contiguity, it is clear that the hybrid assemblers outperformed the other assembly approaches, being MaSuRCA the most contiguous, although still highly fragmented (Other contiguity metrics are available in Supplemental Table S1).

Besides Flye, every assembly has a gene completeness higher than 95%, but they also present on average a genome size 2.87 times larger than the estimated by Genomescope2 model. This suggests that these assemblies are redundant and might have several times the haploid genome. Interestingly, BUSCO gene completeness of these genomes report that more than 90% of the genes are single copy and less than 5% are duplicated (Supplemental Figure S2). This apparent contradicting evidence suggests that even though the genome is redundant, this redundancy is not evenly distributed, and there is a bias in the regions that got repeated.

Regarding the kmer-based metrics, Flye performed poorly comparing to the other assemblers. These performed very similar to each other except for MaSuRCA, which presents lower values, especially for QV. This made an interesting contrast as MaSuRCA performed best in the previous categories. However, QV computation assumes that all the kmers in the assembly should be present in the read set at least once [53]. This assumption may not apply for hybrid assemblies as these incorporate ONT reads as well, and would result in an underestimation of the QV of the assemblies. Although Platanus is a hybrid assembler, its assembly step only incorporates short reads and then it scaffolds with long reads, which explains why its QV value is comparable to short read assemblers. This result shows that the MaSuRCA assembly incorporates information (kmers) from the ONT reads that is not present in the Illumina reads. As Illumina reads present amplification bias, it is possible that some regions of the genome got under-sampled or even not sequenced [54], which reinforces the benefits of integrating data of different technologies.

Short read and Platanus assemblies appear to have a slight advantage in kmer-completeness comparing to the MaSuRCA assembly. Nevertheless, a substantial part of the genome in these assemblies is composed by short sequences, which is reflected in the earlier sharp drops in Figure 4 and a high number of sequences (n in Supplemental Table S1). Kmer diversity (or completeness) is clearly more useful if it comes from longer sequences. Also, a question that arises from fragmented assemblies is how useful the shortest sequences are, and furthermore, how to determine which to keep or filter out.

To further investigate these, we filtered sequences in assemblies by size with different cutoffs (500, 1000, 2000, 5000, 10000, 15000, 20000) and evaluated how gene completeness, kmer completeness and QV changed. These variation are presented in Figure 5a, 5b and S3, respectively. Flye assembly was excluded from this analysis to achieve better visualizations.



(a) BUSCO completeness

(b) Kmer completeness

Figure 5: Quality metrics variation of assemblies with cutoff size.

As we are removing information from the assemblies, the decreasing behaviour observed in figure 5 is expected. However, the MaSuRCA assembly stands out as it is almost not affected by the cutoff, which suggests that little information resides in the small sequences of this assembly. Indeed, it has a better kmer completeness (Figure 5b) after 10000 cutoff than the other assemblies, although at lower levels it performs worse. This means we almost certainly can get rid of these without losing information.

The same can not be said for the other assemblies, which quickly lose information with increasing cutoffs. At the 10000 cutoff the difference is notorious. Platanus assembly is the best among these, reinforcing the superiority of hybrid assemblies over short read assemblies.

In contrast to completeness, QV behaviour is almost unaffected by the cutoff, or even slightly improves (Supplemental figure S3). This makes sense as the shorter sequences might be on average of lower quality than longer ones, given that they have more breaking sites per unit of information, and errors tend to concentrate in these regions.

Given the previous results, with the highest contiguity, the highest gene completeness and the best completeness robustness, we consider that MaSuRCA assembly is the best candidate for *Plukenetia volubilis* draft genome.

2.5 Nuclear genome size mystery

The difference between genome sizes reported by genome profiling and assembly is puzzling. Although kmer-based genome size estimation has shown high accuracy, even more than established experimental methods like flow cytometry[55], there is a big gap between the estimated genome size (197 Mb) and the assembled genome size (596 Mb), being the later three times bigger than the former. Hence, either the estimated size, the assembled size or both are incorrect.

Kmer-based genome size estimates underestimated genome size before [56, 57, 58], which is suggested to be due to the exclusion of highly repetitive kmers from the analysis [55]. This is the default behaviour of KMC (count up to 10000x) and we used a cutoff of 500x with GenomeScope2 to avoid organellar kmers to affect the estimate.

For this reason, we counted again using a higher cutoff (1 billion x) to include all the repetitive kmers and run GenomeScope2 without coverage cutoff. We removed a high coverage peak in the range 1000-2000x that should correspond to organellar kmers. The re-estimate is presented in Figure 6.

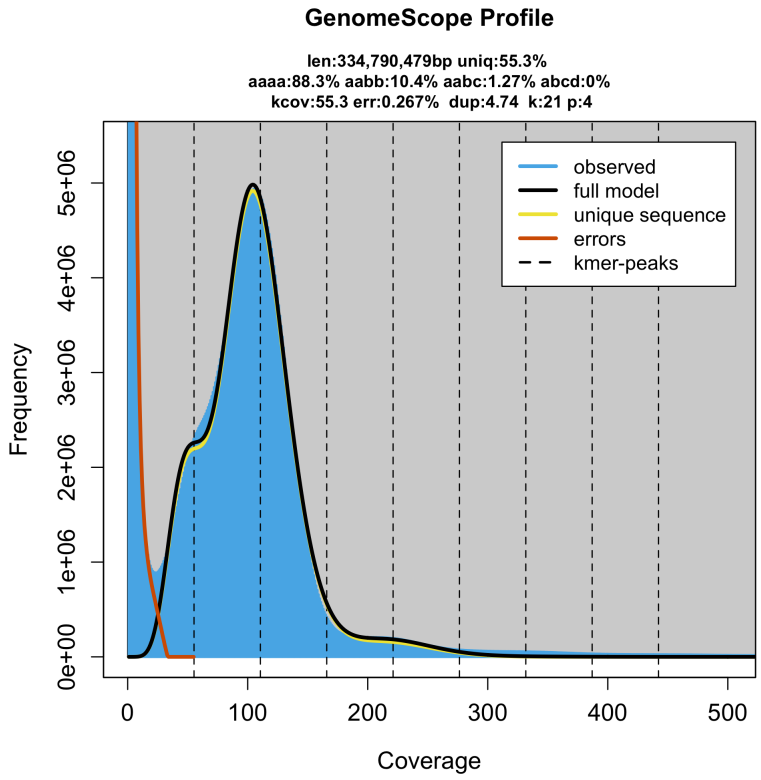


Figure 6: Genomescope model with all kmer, except chloroplast kmers.

The inclusion of the long tail of the histogram (highly repetitive kmers) in the GenomeScope2 model changed mainly the genome size and the repetitive content estimates. Genome size jumped from 197 Mb to 334 Mb, which means that 137 Mb of repetitive content was missing from the original estimate. The exclusion of the high coverage peak, only affected marginally the genome size estimate (12 Mb; Supplemental Figure S4). Taking this into account, the repetitive content changed from 17.4% to 44.7%.

We confirmed this repetitive content using the Extensive De novo TE Annotator (EDTA)[59] to annotate the MaSuRCA assembly, obtaining a total of 282 Mb of repeats, which represents around 47.25% of the genome.

Nevertheless, this does not fully explain the nuclear size mystery, as the genome assembly size is still 1.78 times bigger than estimated. Accordingly, the other part of the explanation has to involve the assembly. As discussed earlier, the big genome size is common to all the generated assemblies except Flye (Table 1). As this is the case for all the these assemblies, the root of the discrepancy has to be systemic.

One of the features inferred by Genomescope2, is the extremely high level of heterozygosity of the genome (reestimated to 11.67%). Much lower levels of heterozygosity have been reported to complicate the assembly process and inflate genomes to sizes higher than expected [60, 61, 62]. Thus, considering the heterozigosity of the genome, the MaSuRCA assembly is likely to present most of both haplotypes. Interestingly, as BUSCO reported 91.4% of orthologs as single copy and just 5% as duplicated, most of these genes are located at homozygous regions, which raises questions of how prevalent is the heterozygosity among other genes. Furthermore, it also suggests that the heterozygosity is not evenly distributed in the genome.

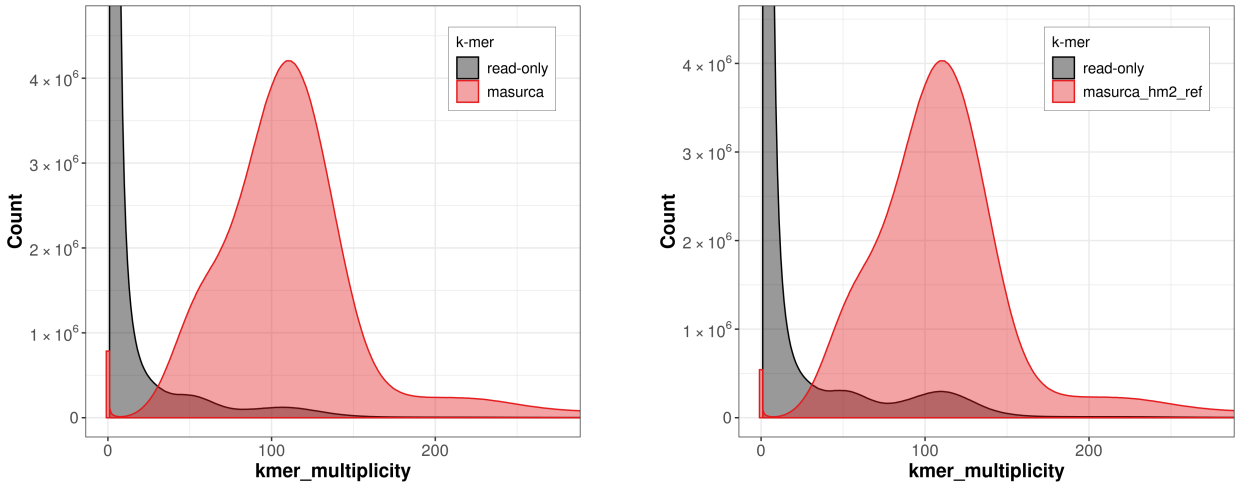
To address this, we used Haplomerger2[63], a post-assembly pipeline that aims to recover haplotigs from highly polymorphic assemblies. We also evaluated Purge Haplotigs[64], a tool to curate higly heterozygous genome assemblies identifying and classifying allelic contigs to separate the haplotigs. We assessed both outputs as we did with the previous assemblies. Results are summarized in Table 2. MaSuRCA assembly is included for comparison.

Additionally, assembly kmer spectra of MaSuRCA and Haplomerger2, generated by Merqury, are presented in Figure 7 (assembly spectra of Purge Haplotigs is presented in supplemental Figure S5). In this figure, the kmer spectra of the assembly (red) is plotted along with the kmer spectra of kmers that are present in the reads but not in the assembly (grey). Ideally, for a non-polymorphic genome, the grey spectra would follow a negative exponential distribution, as the probability of finding erroneous kmers decreases with higher multiplicities, which would also mean that assembly kmer spectra has all the non-erroneous kmers from the reads. For polymorphic genomes, if only one haplotig is assembled, part of the heterozygous kmers is expected to be absent.

Method	# Sequences	E-size[a]	BUSCO Comp. (%)[b]	Kmer Comp. (%)	QV[c]	Total size (Mb)
MaSuRCA	13713	168206	96.4	94.7	41.5	596
Purge Haplotigs	10632	171397	96.4	94.03	42.6	581
Haplomerger2	9145	199626	96.1	91.8	42.8	550

[a] Expected size (calculated by abyss-fac).
[b] Complete BUSCOs (single copy + duplicated).
[c] Phred score value; $QV = -10 \log P$, where P is the probability of error.

Table 2: Quality statistics for Haplomerger2 input and output.



(a) MaSuRCA assembly

(b) Haplomerger2 main haplotype

Figure 7: Kmer spectra plots of assembly and reads

From Table 2, we observed different results from both pipelines. Purge Haplotigs was able to reduce the number of sequences, filtering those that were classified either as artifacts or haplotypes, while preserving completeness. However, this had little effect in genome size (only 15 Mb), which suggests that most of the haplotigs are still present. On the other hand, Haplomerger2 reduced the number of sequences even more and improved the assembly contiguity, although it did not reduce the genome size substantially either (45 Mb). Alas, it reduced slightly gene completeness, which suggests that it discarded some genes and more notoriously kmer completeness.

Some of the kmer completeness reduction is expected as these pipelines aim to separate the haplotypes and we are assessing only one of them. Indeed, when we inspect the kmer spectra of the input and output of Haplomerger2 (Figure 7a and 7b, respectively.), we found an increase on the missing kmers (grey distributions). However, there is just a small increase in the first hillock (around 50x), which corresponds to heterozygous kmers, while there is a much substantial increase in the second hillock (around 120x), which corresponds to homozygous kmers. This suggests that Haplomerger2 is confusing homozygous regions with heterozygous regions and removing the former from the assembly instead of the latter. In contrast, kmer spectra of Purge Haplotigs resembles almost identically the kmer spectra of the MaSuRCA assembly (supplemental Figure S5).

Overall, post-processing pipelines showed mixed results, both deteriorating and improving the MaSuRCA assembly. Nevertheless, none of the assessed tools were able to resolve the heterozygosity issue and the draft assembly remains with mixed haplotigs.

3 Discussion

In this work, we sequenced *Plukenetia volubilis* total DNA from leaf tissue. With the sequencing data generated we were able to assemble completely *de novo* its chloroplast genome with reference-quality, infer several characteristics of its nuclear genome through genome profiling, and propose the first draft nuclear genome of *Plukenetia volubilis*.

After confirming both sequence and structure of *Plukenetia volubilis* chloroplast genome, we found two inversions that were not previously detected. The existence of rearrangements in segments of chloroplast genomes may be useful as phylogenetic markers within genera or even within families, becoming a potential tool for understanding the evolution of plant species[65]. Chloroplast genomes tend to be conserved in structure, especially in the same plant family, as happens in other *Malpighiales* families (*e.g.* Salicaceae [66] , Chrysobalanaceae [67]). This trend seemed to apply in the *Euphorbiaceae* family as well[65], until a 30 kb inversion was found in *Hevea brasiliensis*[15]. Our findings further raise the question of how conserved is the chloroplast genomic structure in the *Euphorbiaceae* family. Future sequencing work and analysis of other chloroplasts in the family will help to elucidate if this pattern is specific to some genera or has a richer evolutionary history.

Using genome profiling we were able to infer several biological features of *Plukenetia volubilis* nuclear genome from unassembled short reads. In fact, this approach guided the assembly process and allowed us to better evaluate and interpret the assembly results.

Among its features we found a relatively small haploid genome size (334 Mb), a considerably high repeat content (47%), and a high level of heterozigosity (11.67%). This positions *Plukenetia volubilis* genome as an outlier of a trend in plants of increasing repetitive content with genome size [68, 69], having a repetitive content much higher than expected. Moreover, our initial genome profiling estimates were flawed due to the exclusion of highly repetitive kmers. These kmers accounted for 41% of the haploid genome size and 27.3% of the repetitive content. These kmers come from repeats that exhibit a high degree of conservation. Under this observation, we speculate that *Plukenetia volubilis* genome underwent a recent repeat expansion, possibly through class-1 retrotransposon activity [70, 71]. Further research is needed to establish the existence of such event and the implications it may have for the species.

We could also infer that *Plukenetia volubilis* genome is allotetraploid with a high degree of heterozygosity coming from its homoeologous chromosomes. We discard other ploidies with this level of heterozygosity for several reasons:

Besides the already discussed (see Genome profiling), it is hard to imagine a sexual reproducing species developing such high levels of polymorphism only via evolution mechanisms. Although some degree of heterozygosity is present on the population as genetic diversity, the main driver of diversification in sexual species is recombination [72, 73, 74]. In fact, high levels of heterozygosity can reduce recombination of homologous chromosomes [75], possibly by hindering homologous pairing [32]. This imposes a limit to the degree of heterozygosity a sexual species can accumulate with evolution mechanisms. However, depending on the divergence between parentals, this limit can be bypassed through hybridization, where the genetic diversification occurs in a single event.

We argue that the level of heterozygosity of *Plukenetia volubilis* further supports that its genome is product of hydridization. This also complements our initial analysis regarding its ploidy, as allopolyploids (tetraploid in this case)

tend to have higher leves of heterozygosity than homoploids (diploid) [76]. The high level of heterozygosity suggests that the hybridization event was relatively recent, but more work is needed to locate it in the evolutionary history.

It is worth mentioning though, that there is an alternative model that could explain our genome profiling results: A highly homozygous diploid structure of the genome. Under this model, the genome is so homozygous that the signal observed in the smudgeplot (Figure 2a) is dominated not by heterozygous loci but rather by other sources, such as paralogous genes, copy number variations, tandem repeats, etc. Nevertheless, we do not think this is the case as the assembly stage would had yielded, in general, better results. One way to definitively rule out this alternative, a homology search of annotated genes against the whole genome can be performed. If the genome is allotetraploid, the similarity distribution of the secondary hits should exhibit a peak around the expected divergence of the homoeologous (10%), while a homozygous diploid genome would yield a random distribution.

Thanks to our available data, we were able to evaluate different preprocessing and assembly approaches. Most of them produced highly fragmented assemblies, and were not described in this work for brevity. In short, we evaluated short read assembly, hybrid assembly and long read assembly with polishing.

Even though long read assemblers are expected to outperform short read assemblers[77], it was not surprising that hybrid assemblers were superior, considering the low quality of our nanopore data (read N50 $2kb$) and low depth (total data less than 7 Gb). Thus, long read assemblers did not have enough information to work with. Yet, there was still useful information in the data, giving hybrid assemblers better chances to find a good representation of the genome than short read assemblers.

In our assessment of the different assemblies, we found that the MaSuRCA assembly had the best quality, exhibiting high and robust completeness at both gene and kmer level. The high levels of heterozygosity prevented us to achieve higher contiguity and a closer representation of the haploid genome. Although we evaluated post-processing tools that handle highly heterozygous assemblies, we were not able to improve it significantly. Hence, the assembly remains with mixed haplotigs, which might be problematic for some downstream analysis if not taken into account properly[78].

Still, our draft genome assembly represent a useful resource for the study of this promising crop. It will facilitate and accelerate the genetic improvement of *Plukenetia volubilis* through molecular breeding and exploitation of genetic resources. It also will help to get a better understanding of the molecular bases of the phenotypic traits.

In addition to the aforementioned, several avenues opened for future work. From an evolutionary point of view, the reported genomes will help to get a better view of the *Euphorbiaceae* family, given the plethora of new markers. Further refinement of the genome will be helpful for downstream analysis, opening of new possibilities. Heterozygosity handling, improving contiguity, and disambiguate collapsed repeat content are of special interest in this regard. Gene annotation and metabolic analysis will be key to comprehend the biological basis of *Plukenetia volubilis* oil composition, including the accumulation of polyunsaturated fatty acids. This understanding will represent a major step towards the development of genetic improvement programs.

4 Methods

4.1 Plant material

This research was conducted at Plant Biotechnology Laboratory and Molecular Biology Laboratory of the Department of Biological Sciences at Universidad EAFIT, Medellin, Colombia. Both leaves and seeds were collected from one *Plukenetia volubilis* cultivar in Santa Rosa, Antioquia (Colombia) farm, having a permit for this gathering issued by the National Authority for Environmental Licenses (ANLA acronym in Spanish), covered in resolution 1516 of 2014 (modified through resolution 1312 of 2015). It was not possible to further trace back the origin of the cultivar, although it is likely it was introduced from Peru at some point. Then, sample was washed completely with distilled water, wiped and packed in a bag to avoid light degradation, and stored at -20 C.

4.2 DNA extraction and sequencing

For Nanopore and Illumina sequencing, high molecular weight genomic DNA from leaves of cultivar 1 (C1) was isolated following *Ramírez-Ríos et al.* protocol[79] with some adaptations for plant DNA, described in supplemental methods.

Extract quality was evaluated using gel electrophoresis for size estimation, spectrophotometry (A260/A280 and A260/A230 ratios) for purity estimation, and Qbit for total DNA extracted. DNA samples with a A260/A280 ratio close to 2 and a A260/A230 ratio above 1.5 were conserved.

4.3 Computing resources

All bioinformatic analysis presented in this work were executed in the center of scientific computing Apolo at EAFIT university.

Each analysis was perfromed in either a standard node, with 32 cpu threads and 64 Gb of memory; or a big memory node, with 24 cpu threads and 378 Gb of memory.

4.4 Chloroplast genome assembly

Raw Illumina reads were used as input to Norgal (v1.0.0) with default settings. Resulting scaffolds were filtered with a custom python script based on Norgal blast considering: best hit matched a chloroplast, length (at least 1000 bp), identity (at least 95%) and alignment length (at least 200 bp).

Novoplasty(v2.7.2), which was used to assemble the same WGS paired-end reads, using the survivor scaffold after filtering as seed, 100-200 kb as *genome range*, a *kmer* length of 39, an *insert size* of 370, and the remaining parameters as default.

Raw paired-end reads were trimmed using Trimommatic-PE (v0.39) with a *sliding window* of width 4 and quality 25 and a *min length* of 50. All surviving (paired and unpaired) reads were mapped using BWA-MEM(v0.7.17). Samtools(v1.9) was used to filter the unmapped reads and sort the resulting alignments. These were provided to Pilon, along with the generated sequence, with options *fix*set to all and *min depth* set to 60.

Nanopore raw reads were basecalled with Guppy (v2.3.5) using the flip-flop model (currently renamed as high accuracy model) with default settings. Passing reads (mean $QV > 7$) were mapped using Minimap2 (v2.17-r941), in mode *map-ont* and remaining parameters set to default. Samtools(v1.9) was used to filter the unmapped reads and sort the resulting alignments.

Nucmer, from the Mummer (v3.23) program, was used to align *Ricinus communis* chloroplast genome and the assembled chloroplast genome, with default parameters. Resulting delta file was used to generate the dotplot using mummerplot with default parameters.

4.5 Genome profiling

Kmer counting of 21-mers was performed on the raw Illumina data using KMC (v3.1) with default parameters. Resulting kmer spectra was used as input for Genomescope2 web-portal, with *max coverage* set to 500 and *ploidy* to 4; and Smudgeplot (v0.2), with lower and upper thresholds set to 30 and 2600, respectively.

4.6 Genome Assembly

4.6.1 Preprocessing:

Cleaning of Illumina raw data was carried out using BBduk (v37.62), from the BBtools package. Adapters commonly used in preparation of Illumina libraries, PhiX adapters (both included in BBTools package) and adapters found by FastQC (v0.11.8) were removed at both sides of the reads, setting *k* to 23 and *mink* to 21. Another round of BBduk was used to filter reads matching PhiX174 virus genome sequence (NC001422.1), with *k* set to 31.

Error correction of clean Illumina data was performed using Brownie corrector (v0.1), with default extraction motif (15 bp long poly-A) and default parameters. Reads not processed in the previous step were corrected using Karect (v1.0), with *celltype* set to diploid and *matchtype* to hamming.

Passing reads after basecalling (section 4.3) were used to compute all vs all overlaps using Minimap2 (v2.17-r941) in *ava-ont* mode. These overlaps were used to further preprocess the reads using Yacrd (v0.5.1) and remove chimeric reads, using default settings.

4.6.2 Assembly and polishing:

For short read assembly the Illumina error corrected paired-end reads were used as input to the following assemblers, with default parameters unless stated otherwise:

- Abyss (v2.1.5) was used with several *kmer* values $(39, 50, 70, 90, 110)$. The best assembly was kept for downstream analysis.

- GATB-minia-pipeline, which uses Minia (v3.2.1) for assembly and BESST (v2.2.8) for scaffolding, was used without error correction and evaluating different *kmer* values $(21, 39, 57, 75, 93, 111, 127)$.
- Megahit (v1.2.8) was used with *min-kmer* value set to 27, *min-count* set to 4 and option *no mercy*.
- IDBA-UD (v1.1.3) was used with *min-count* set to 3 and *no-correct* option.

For hybrid assembly, the same reads used for short read assembly and the preprocessed nanopore reads were used as input to the following assemblers, with default parameters unless stated otherwise:

- Spades (v3.13.1) was used in *hybrid* mode, kmer values of $21, 39, 57, 75$ and option *cov-cutoff* set to auto.
- Platanus-allee (v2.0.2) was used in three steps: first, platanus assemble mode using Illumina data as input and a *minimum-kmer coverage* of 3; second, platanus phase using both Illumina and ONT reads as input to phase the assembly generated in the previous step in 3 *iterations*; finally, platanus concensus using both Illumina and ONT reads as input to make a concensus over the phased blocks found in the previous step.
- MaSuRCA (v3.3.3) was used setting its Illlumina reads mean insert size to 270 and standard deviation to 90 (determined with BBmerge), *limit jump mates* set to 300, *kmer count threshold* set to 2, and options *use linking mates* and *close gaps*.

For long read assembly and polishing, Flye (v2.4.2) was used with the nanopore reads aforementioned as nano-raw input and *genome size* set to 200 Mb. The generated assembly was polished twice using Racon (v1.4.3), with parameters *match* set to 8, *mismatch* to -6, *gap* to -8 and *window size* to 500. Mapping of ONT reads to assembly for this step was performed using Minimap2 in *map-ont* mode. The assembly was further polished with the ONT reads using medaka (v0.8.1) with model r941_flip235. For polishing with Illumina reads, reads were mapped to assembly using BWA-MEM (v0.7.17) and resulting alignments sorted with Samtools (v) which were then used as input to Pilon (v1.23) with parameter *fix* set to all and option *diploid*.

4.6.3 Assembly quality evaluation

The following methods were used to evaluate quality of the generated assemblies:

Contiguity metrics were computed with abyss-fac from the Abyss assembler. Gene completeness was estimated using BUSCO (v3.0.2) comparing to the eudicotyledons-odb10 dataset. For Merqury (v2020-01-29) analysis, a Meryl (v1.0) database of the raw Illumina reads was created with a kmer size of 19 and each genome was compared to this database.

Assembly filtering by size and Figures 4, 5 and S3 were made with custom python scripts using Biopython [80], Pandas [81, 82], Matplotlib [83] and Seaborn [84].

4.7 Nuclear size mistery

Kmer counting of 21-mers was performed on the raw Illumina data using KMC with counts up to 1000000000x. Resulting kmer spectra was used as input for Genomescope2 web-portal, without *max coverage* (set to -1) and *ploidy* to 4.

EDTA (v1.6.4) pipeline was used, and thus all its dependencies [85, 86, 87, 88, 89, 90, 91, 92], to annotate the repetitive content of the MaSuRCA assembly, with options *sensitive*, *anno* and *evaluate*.

Illumina error corrected reads were used as input for both HaploMerger2 (v20180603) and Purge Haplotigs (v1.1.0). MaSuRCA assembly was masked with windowmasker (v1.0.0) prior to HaploMerger2 processing, which was run with default parameters. Reads were aligned to assembly using BWA-MEM and sorted with Samtools, before Purge Haplotigs processing: *depth* was set to 300 for *hist* step and *low*, *mid* and *high* were set to 16, 80 and 240 for *cov* step; remaining parameters were set as default.

5 Acknowledgments

We would like to thank Apolo supercomputing center team for their support through the development of this project. We would also like to thank Kamil S. Jaron for thorough discussions about genomes and genome profiling. This work was supported by EAFIT university internal research grant 828-000156.

References

[1] Foreign Agricultural Service. Production supply and distribution. `https://apps.fas.usda.gov/psdonline/app/index.html#/app/home`. Accessed: 2020-08-17.

[2] FAO. *How to feed the world in 2050*. Expert meeting on how to feed the world in 2050, 2009.

[3] Diego Villanueva-Mejia and Javier Correa Alvarez. Genetic improvement of oilseed crops using modern biotechnology. *Advances in Seed Biology*, pages 295–317, 2017.

[4] Luis A. Follegatti-Romero, Carla R. Piantino, Renato Grimaldi, and Fernando A. Cabral. Supercritical CO2 extraction of omega-3 rich oil from Sacha inchi (Plukenetia volubilis L.) seeds. *Journal of Supercritical Fluids*, 49(3):323–329, 2009.

[5] Chiara Fanali, Laura Dugo, Francesco Cacciola, Marco Beccaria, Simone Grasso, Marina Dachà, Paola Dugo, and Luigi Mondello. Chemical characterization of Sacha inchi (Plukenetia volubilis L.) oil. *Journal of Agricultural and Food Chemistry*, 59(24):13043–13049, 2011.

[6] Rosana Chirinos, Gledy Zuloeta, Romina Pedreschi, Eric Mignolet, Yvan Larondelle, and David Campos. Sacha inchi (Plukenetia volubilis): A seed source of polyunsaturated fatty acids, tocopherols, phytosterols, phenolic compounds and antioxidant capacity. *Food Chemistry*, 141(3):1732–1739, 2013.

[7] Luis Felipe Gutiérrez, Lina María Rosada, and Álvaro Jiménez. Chemical composition of Sacha Inchi (Plukenetia volubilis L.) seeds and characteristics of their lipid fraction. *Grasas y Aceites*, 62(1):76–83, 2011.

[8] Diego Leandro Castaño T, María del Pilar Valencia G, Elizabeth Murillo P, Jonh Jairo Mendez A, and Jordi Eras Joli. Fatty Acid Composition of Inca Peanut (Plukenetia volúbilis Linneo) and its Relationship with Vegetal Bioactivity. *Revista Chilena de Nutricion*, 39(11):45–52, 2012.

[9] W. Carrillo, M. F. Quinteros, C. Carpio, D. Morales, G. Vásquez, M. Álvarez, and M. Silva. Identification of fatty acids in sacha inchi oil (Plukenetia volubilis) from Ecuador. *Asian Journal of Pharmaceutical and Clinical Research*, 11(2):18–20, 2018.

[10] Philip D Bates, Sten Stymne, and John Ohlrogge. Biochemical pathways in seed oil synthesis. *Current opinion in plant biology*, 16(3):358–364, 2013.

[11] Jeffrey D Palmer. Plastid chromosomes: structure and evolution. *The molecular biology of plastids*, 7:5–53, 1991.

[12] Maximo Rivarola, Jeffrey T Foster, Agnes P Chan, Amber L Williams, Danny W Rice, Xinyue Liu, Admasu Melake-Berhan, Heather Huot Creasy, Daniela Puiu, MJ Rosovitz, et al. Castor bean organelle genome sequencing and worldwide genetic diversity analysis. *PloS one*, 6(7):e21743, 2011.

[13] Mehar H Asif, Shrikant S Mantri, Ayush Sharma, Anukool Srivastava, Ila Trivedi, Priya Gupta, Chandra S Mohanty, Samir V Sawant, and Rakesh Tuli. Complete sequence and organisation of the jatropha curcas (euphorbiaceae) chloroplast genome. *Tree Genetics & Genomes*, 6(6):941–952, 2010.

[14] Henry Daniell, Kenneth J Wurdack, Anderson Kanagaraj, Seung-Bum Lee, Christopher Saski, and Robert K Jansen. The complete nucleotide sequence of the cassava (manihot esculenta) chloroplast genome and the evolution of atpf in malpighiales: Rna editing and multiple losses of a group ii intron. *Theoretical and Applied Genetics*, 116(5):723, 2008.

[15] Sithichoke Tangphatsornruang, Pichahpuk Uthaipaisanwong, Duangjai Sangsrakru, Juntima Chanprasert, Thippawan Yoocha, Nukoon Jomchai, and Somvong Tragoonrung. Characterization of the complete chloroplast genome of hevea brasiliensis reveals genome rearrangement, rna editing sites and phylogenetic relationships. *Gene*, 475(2):104–112, 2011.

[16] Ying Zhang, Yancai Shi, Na Duan, Bing-Bing Liu, and Jia Mi. Complete chloroplast genome of euphorbia tirucalli (euphorbiaceae), a potential biofuel plant. *Mitochondrial DNA Part B*, 4(1):1973–1974, 2019.

[17] Masahiro Sugiura, Tetsuro Hirose, and Mamoru Sugita. Evolution and mechanism of translation in chloroplasts. *Annual review of genetics*, 32(1):437–459, 1998.

[18] Masahiro Sugiura. The chloroplast genome. *Plant molecular biology*, 19(1):149–168, 1992.

[19] Kosai Al-Nakeeb, Thomas Nordahl Petersen, and Thomas Sicheritz-Pontén. Norgal: extraction and de novo assembly of mitochondrial dna from whole-genome sequencing data. *BMC bioinformatics*, 18(1):510, 2017.

[20] Nicolas Dierckxsens, Patrick Mardulyn, and Guillaume Smits. Novoplasty: de novo assembly of organelle genomes from whole genome data. *Nucleic acids research*, 45(4):e18–e18, 2016.

[21] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[22] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

[23] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[24] Iain Milne, Micha Bayer, Linda Cardle, Paul Shaw, Gordon Stephen, Frank Wright, and David Marshall. Tablet—next generation sequence assembly visualization. *Bioinformatics*, 26(3):401–402, 2010.

[25] Xiao-Di Hu, Bang-Zhen Pan, Qiantang Fu, Mao-Sheng Chen, and Zeng-Fu Xu. The complete chloroplast genome sequence of the biofuel plant sacha inchi, plukenetia volubilis. *Mitochondrial DNA Part B*, 3(1):328–329, 2018.

[26] Arthur L Delcher, Steven L Salzberg, and Adam M Phillippy. Using mummer to identify similar regions in large sequence sets. *Current protocols in bioinformatics*, (1):10–3, 2003.

[27] Lauren Ancel Meyers and Donald A Levin. On the abundance of polyploids in flowering plants. *Evolution*, 60(6):1198–1206, 2006.

[28] ALL Vanzela, PM Ruas, and MA Marin-Morales. Karyotype studies of some species of dalechampia plum.(euphorbiaceae. *Botanical Journal of the Linnean Society*, 125(1):25–33, 1997.

[29] ZQ Cai, T Zhang, and HY Jian. Chromosome number variation in a promising oilseed woody crop, plukenetia volubilis l.(euphorbiaceae). *Caryologia*, 66(1):54–58, 2013.

[30] Timothy Rhyker Ranallo-Benavidez, Kamil S Jaron, and Michael C Schatz. Genomescope 2.0 and smudgeplots: Reference-free profiling of polyploid genomes. *BioRxiv*, page 747568, 2019.

[31] Marek Kokot, Maciej Długosz, and Sebastian Deorowicz. Kmc 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761, 2017.

[32] Denise Zickler and Nancy Kleckner. Recombination, pairing, and synapsis of homologs during meiosis. *Cold Spring Harbor perspectives in biology*, 7(6):a016626, 2015.

[33] SC Le Comber, ML Ainouche, A Kovarik, and AR Leitch. Making a functional diploid: from polysomic to disomic inheritance. *New Phytologist*, 186(1):113–122, 2010.

[34] Jacob Sybenga. *Meiotic configurations: a source of information for estimating genetic parameters*, volume 1. Springer Science & Business Media, 2012.

[35] Fay-Wei Li and Alex Harkess. A guide to sequence your favorite plant genomes. *Applications in plant sciences*, 6(3), 2018.

[36] Brian Bushnell. Bbmap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.

[37] Mahdi Heydari, Giles Miclotte, Yves Van de Peer, and Jan Fostier. Illumina error correction near highly repetitive dna regions improves de novo genome assembly. *BMC Bioinformatics*, 20(1):298, Jun 2019.

[38] Amin Allam, Panos Kalnis, and Victor Solovyev. Karect: accurate correction of substitution, insertion and deletion errors for next-generation sequencing data. *Bioinformatics*, 31(21):3421–3428, 2015.

[39] Pierre Marijon, Rayan Chikhi, and Jean-Stéphane Varré. yacrd and fpa: upstream tools for long-read genome assembly. *bioRxiv*, page 674036, 2019.

[40] Rayan Chikhi and Guillaume Rizk. Space-efficient and exact de bruijn graph representation based on a bloom filter. *Algorithms for Molecular Biology*, 8(1):22, 2013.

[41] Kristoffer Sahlin, Francesco Vezzi, Björn Nystedt, Joakim Lundeberg, and Lars Arvestad. Besst-efficient scaffolding of large fragmented assemblies. *BMC bioinformatics*, 15(1):281, 2014.

[42] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.

[43] Yu Peng, Henry CM Leung, Siu-Ming Yiu, and Francis YL Chin. Idba-ud: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, 2012.

[44] Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin, and Pavel A Pevzner. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*, 37(5):540–546, 2019.

[45] Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research*, 27(5):737–746, 2017.

[46] Bruce J Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, 9(11):e112963, 2014.

[47] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477, 2012.

[48] Dmitry Antipov, Anton Korobeynikov, Jeffrey S McLean, and Pavel A Pevzner. hybridspades: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, 32(7):1009–1015, 2016.

[49] Aleksey V Zimin, Guillaume Marçais, Daniela Puiu, Michael Roberts, Steven L Salzberg, and James A Yorke. The masurca genome assembler. *Bioinformatics*, 29(21):2669–2677, 2013.

[50] Rei Kajitani, Dai Yoshimura, Miki Okuno, Yohei Minakuchi, Hiroshi Kagoshima, Asao Fujiyama, Kaoru Kubokawa, Yuji Kohara, Atsushi Toyoda, and Takehiko Itoh. Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nature communications*, 10(1):1–15, 2019.

[51] Felipe A Simão, Robert M Waterhouse, Panagiotis Ioannidis, Evgenia V Kriventseva, and Evgeny M Zdobnov. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, 2015.

[52] Daniel Mapleson, Gonzalo Garcia Accinelli, George Kettleborough, Jonathan Wright, and Bernardo J Clavijo. Kat: a k-mer analysis toolkit to quality control ngs datasets and genome assemblies. *Bioinformatics*, 33(4):574–576, 2017.

[53] Arang Rhie, Brian P Walenz, Sergey Koren, and Adam M Phillippy. Merqury: reference-free quality and phasing assessment for genome assemblies. *BioRxiv*, 2020.

[54] Daniel Aird, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome biology*, 12(2):1–14, 2011.

[55] Yamkela Mgwatyu, Allison Anne Stander, Stephan Ferreira, Wesley Williams, and Uljana Hesse. Rooibos (aspalathus linearis) genome size estimation using flow cytometry and k-mer analyses. *Plants*, 9(2):270, 2020.

[56] Ying Hu, Marcio FR Resende, Aureliano Bombarely, Maria Brym, Elias Bassil, and Alan H Chambers. Genomics-based diversity analysis of vanilla species using a vanilla planifolia draft genome and genotyping-by-sequencing. *Scientific reports*, 9(1):1–16, 2019.

[57] Richard J Edwards, Daniel Enosi Tuipulotu, Timothy G Amos, Denis O'Meally, Mark F Richardson, Tonia L Russell, Marcelo Vallinoto, Miguel Carneiro, Nuno Ferrand, Marc R Wilkins, et al. Draft genome assembly of the invasive cane toad, rhinella marina. *GigaScience*, 7(9):giy095, 2018.

[58] Dennis Hedgecock, Patrick M Gaffney, Philippe Goulletquer, Ximing Guo, Kimberly Reece, and Gregory W Warr. The case for sequencing the pacific oyster genome. *Journal of Shellfish research*, 24(2):429–441, 2005.

[59] Shujun Ou, Weija Su, Yi Liao, Kapeel Chougule, Jireh RA Agda, Adam J Hellinga, Carlos Santiago Blanco Lugo, Tyler A Elliott, Doreen Ware, Thomas Peterson, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology*, 20(1):1–18, 2019.

[60] Leszek P Pryszcz, Tibor Németh, Attila Gácser, and Toni Gabaldón. Genome comparison of candida orthopsilosis clinical strains reveals the existence of hybrids between two distinct subspecies. *Genome biology and evolution*, 6(5):1069–1078, 2014.

[61] Jade P Vinson, David B Jaffe, Keith O'Neill, Elinor K Karlsson, Nicole Stange-Thomann, Scott Anderson, Jill P Mesirov, Nori Satoh, Yutaka Satou, Chad Nusbaum, et al. Assembly of polymorphic genomes: algorithms and application to ciona savignyi. *Genome research*, 15(8):1127–1135, 2005.

[62] Kerrin S Small, Michael Brudno, Matthew M Hill, and Arend Sidow. A haplome alignment and reference sequence of the highly polymorphic ciona savignyi genome. *Genome biology*, 8(3):1–14, 2007.

[63] Shengfeng Huang, Mingjing Kang, and Anlong Xu. Haplomerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics*, 33(16):2577–2579, 2017.

[64] Michael J Roach, Simon A Schmidt, and Anthony R Borneman. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC bioinformatics*, 19(1):460, 2018.

[65] Luiz A Cauz-Santos, Carla F Munhoz, Nathalie Rodde, Stephane Cauet, Anselmo A Santos, Helen A Penha, Marcelo C Dornelas, Alessandro M Varani, Giancarlo CX Oliveira, Helene Berges, et al. The chloroplast

genome of passiflora edulis (passifloraceae) assembled from long sequence reads: structural organization and phylogenomic studies in malpighiales. *Frontiers in Plant Science*, 8:334, 2017.

[66] Zhiqiang Wu. The new completed genome of purple willow (salix purpurea) and conserved chloroplast genome structure of salicaceae. *J Nat Sci*, 1:e49, 2015.

[67] Pierre-Jean G Malé, Léa Bardon, Guillaume Besnard, Eric Coissac, Frédéric Delsuc, Julien Engel, Emeline Lhuillier, Caroline Scotti-Saintagne, Alexandra Tinaut, and Jérôme Chave. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Molecular Ecology Resources*, 14(5):966–975, 2014.

[68] Wen-Biao Jiao and Korbinian Schneeberger. The impact of third generation genomic technologies on plant genome assembly. *Current opinion in plant biology*, 36:64–70, 2017.

[69] Paul Julian Kersey, James E Allen, Irina Armean, Sanjay Boddu, Bruce J Bolt, Denise Carvalho-Silva, Mikkel Christensen, Paul Davis, Lee J Falin, Christoph Grabmueller, et al. Ensembl genomes 2016: more genomes, more complexity. *Nucleic acids research*, 44(D1):D574–D580, 2016.

[70] Sung-Il Lee and Nam-Soo Kim. Transposable elements and genome size variations in plants. *Genomics & informatics*, 12(3):87, 2014.

[71] Peter Civáň, Miroslav Švec, and Pavol Hauptvogel. On the coevolution of transposable elements and plant genomes. *Journal of Botany*, 2011, 2011.

[72] Sarah Perin Otto and Nick H Barton. The evolution of recombination: removing the limits to natural selection. *Genetics*, 147(2):879–906, 1997.

[73] Austin Burt. Perspective: sex, recombination, and the efficacy of selection—was weismann right? *Evolution*, 54(2):337–351, 2000.

[74] Julie G Hussin, Alan Hodgkinson, Youssef Idaghdour, Jean-Christophe Grenier, Jean-Philippe Goulet, Elias Gbeha, Elodie Hip-Ki, and Philip Awadalla. Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nature genetics*, 47(4):400–404, 2015.

[75] Rhona H Borts, SR Chambers, and MFF Abdullah. The many faces of mismatch repair in meiosis. *Mutation research/Fundamental and molecular mechanisms of mutagenesis*, 451(1-2):129–150, 2000.

[76] Ovidiu Paun, Félix Forest, Michael F Fay, and Mark W Chase. Hybrid speciation in angiosperms: parental divergence drives ploidy. *New Phytologist*, 182(2):507–518, 2009.

[77] Todd P Michael and Robert VanBuren. Building near-complete plant genomes. *Current opinion in plant biology*, 54:26–33, 2020.

[78] Nathan D Olson, Steven P Lund, Rebecca E Colman, Jeffrey T Foster, Jason W Sahl, James M Schupp, Paul Keim, Jayne B Morrow, Marc L Salit, and Justin M Zook. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in genetics*, 6:235, 2015.

[79] Viviana Ramírez-Ríos, Nicolás D Franco-Sierra, Javier Correa Alvarez, Clara I Saldamando-Benjumea, and Diego F Villanueva-Mejía. Mitochondrial genome characterization of tecia solanivora (lepidoptera: Gelechiidae) and its phylogenetic relationship with other lepidopteran insects. *Gene*, 581(2):107–116, 2016.

[80] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[81] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[82] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[83] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[84] Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017.

[85] David Ellinghaus, Stefan Kurtz, and Ute Willhoeft. Ltrharvest, an efficient and flexible software for de novo detection of ltr retrotransposons. *BMC bioinformatics*, 9(1):18, 2008.

[86] Zhao Xu and Hao Wang. Ltr_finder: an efficient tool for the prediction of full-length ltr retrotransposons. *Nucleic acids research*, 35(suppl_2):W265–W268, 2007.

[87] Shujun Ou and Ning Jiang. Ltr_finder_parallel: parallelization of ltr_finder enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA*, 10(1):1–3, 2019.

[88] Shujun Ou and Ning Jiang. Ltr_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant physiology*, 176(2):1410–1422, 2018.

[89] Weijia Su, Xun Gu, and Thomas Peterson. Tir-learner, a new ensemble method for tir transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Molecular plant*, 12(3):447–460, 2019.

[90] Jieming Shi and Chun Liang. Generic repeat finder: a high-sensitivity tool for genome-wide de novo repeat detection. *Plant physiology*, 180(4):1803–1815, 2019.

[91] Wenwei Xiong, Limei He, Jinsheng Lai, Hugo K Dooner, and Chunguang Du. Helitronscanner uncovers a large overlooked cache of helitron transposons in many plant genomes. *Proceedings of the National Academy of Sciences*, 111(28):10263–10268, 2014.

[92] Ren-Gang Zhang, Zhao-Xuan Wang, Shujun Ou, and Guang-Yuan Li. Tesorter: lineage-level classification of transposable elements using conserved protein domains. *bioRxiv*, page 800177, 2019.

[93] Miriam Schalamun and Benjamin Schwessinger. High molecular weight gdna extraction after mayjonade et al. optimised for eucalyptus for nanopore sequencing. `dx.doi.org/10.17504/protocols.io.khkct4w`, 2017. Online; accessed 29 January 2014.

6 Appendix

6.1 Supplemental methods

6.1.1 DNA extraction and sequencing

Plant DNA extraction was performed following a modification of *Ramirez-Rios et al.* [79]. Lysis buffer from *Schalamun and Schwessinger* was also used [93]. Gravities and time were modified in the centrifugation steps, to decrease the mechanical shearing stress over the DNA, ensuring an unchanged decanting time.

Around 150 mg of freshly harvested leaves were ground in liquid nitrogen with a mortar and pestle and immediately transferred to 1 ml, previously heated to 64 °C, of either CTAB (100 mM Tris-HCl pH 8.0, 1.4 M NaCl, 20 mM EDTA, 2x CTAB, 0.1% (v/v) β-mercaptoethanol) or lysis (2% PVP 40, 500 mM NaCl, 100 mM TRIS-HCl pH 8, 50 mM EDTA, 1.25% SDS, 1% (w/v) sodium metabisulfite, 5 mM Dithiothreitol (DTT)) buffer, and incubated at 37 °C for 30 min. The extract was added with 1 μl of proteinase K (10 mg/ml) and incubated another 30 min at same temperature. After a 5 min cool down, 0.3 volumes of $5M$ potassium acetate were added and the extract was centrifuged ($2000g$ for 48 min at 4 °C). Supernatant was transferred to a fresh tube, 1 volume of chloroform-isoamylalcohol (24:1) was added, and gently mixed by inversion before centrifuging ($2000g$ for 48 min at 4 °C). The previous step was performed twice. The aqueous phase was transferred to a fresh tube, 0.1 volumes of $3M$ sodium acetate were added and the tube was gently mixed by inversion before 1 equivalent volume of isopropanol was added. The extract was incubated at -20 °C overnight and then centrifuged ($10000g$ for 10 min at 4 °C). The resulting DNA pellet was washed with 200 μl of 70% ethanol and centrifuged ($10000g$ for 10 min at 4 °C) twice before left to dry at room temperature for 30 min. The pellet was resuspended in 40 μl of nuclease-free water and 1 μl of RNAse A (1 mg/ml) was added before a final incubation at 37 °C for 1 hour.

6.2 Supplemental figures and tables

Figure S1: Chloroplast genome coverage with ONT reads.

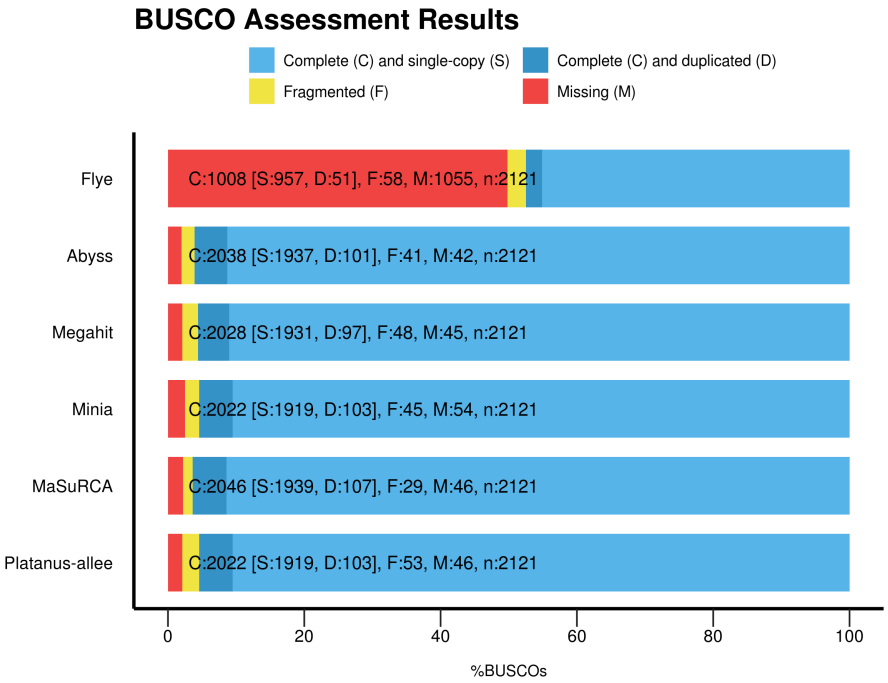


Figure S2: BUSCO gene completeness for *Plukenetia volubilis* nuclear genome assemblies.

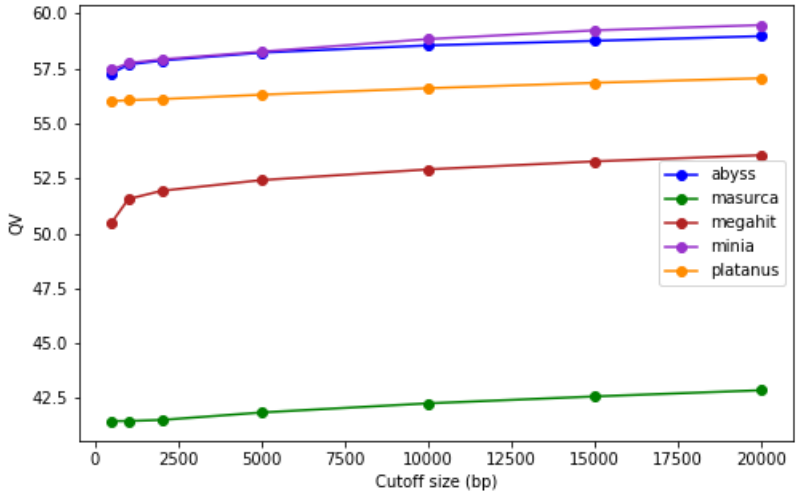


Figure S3: QV variation of adssemblies with cutoff size.

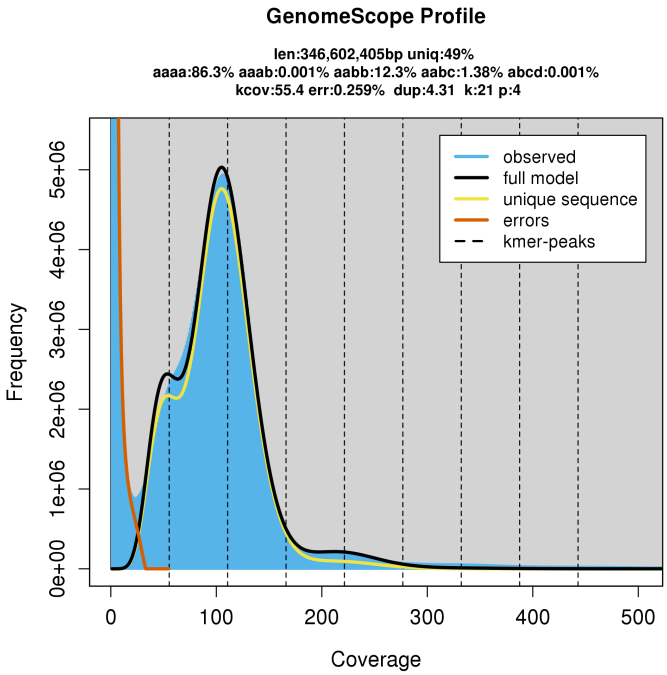


Figure S4: Genomescope model with all kmer

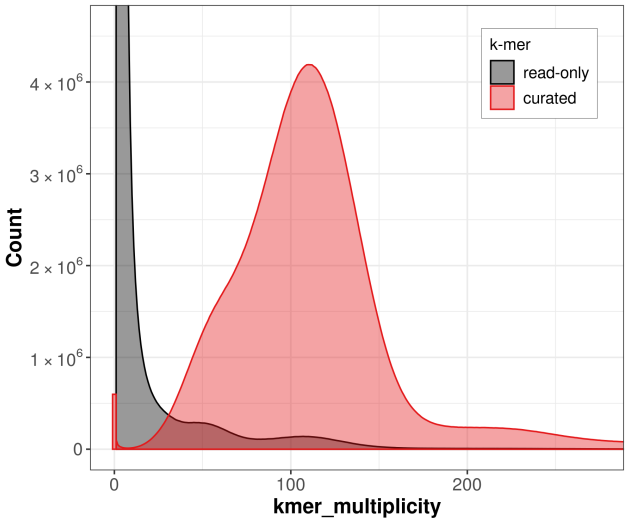


Figure S5: Kmer spectra plot of Purge Haplotigs curated assembly.

Assembler	n[a]	n:2000[b]	L50	N75	N50	N25	Max	Size (Mb)[c]
Flye	18738	14216	3184	19616	31064	45848	156402	297.7
Abyss	669643	25539	3683	19958	40670	74530	464213	537.3
Megahit	775739	29681	4340	17458	35558	62757	351667	538.6
Minia	290756	22588	3169	24016	48992	87791	494742	551.6
MaSuRCA	13713	10722	1428	55625	112275	213463	1627529	594.3
Platanus-allee	688159	20551	1898	37869	88042	169274	750764	616.9

[a] Number of sequences
[b] Number of sequences larger than 2000 bp
[c] Genome size calculated as the sum of the n:2000 scaffolds
[d] Complete BUSCOs (single copy + duplicated)

Table S1: Contiguity and gene completeness for different assemblers