

The haplotype-resolved assembly of COL40 a cassava (*Manihot esculenta*) line with broad-spectrum resistance against viruses causing Cassava brown streak disease unveils a region of highly repeated elements on chromosome 12

Corinna Thoben ¹, Boas Pucker ¹, Stephan Winter,² Bethany Fallon Econopoulou,³ Samar Sheat ^{2,*}

¹Plant Biotechnology and Bioinformatics, Institute of Plant Biology and Braunschweig Integrated Centre of Systems Biology, Braunschweig 38106, Germany

²Plant Virus Department, Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Culture, Braunschweig 38104, Germany

³Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA

*Corresponding author: Plant Virus Department, Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Culture, Braunschweig 38104, Germany.
Email: samar.sheat@dsmz.de

Cassava (*Manihot esculenta* Grantz) is a vital staple crop for millions of people, particularly in Sub-Saharan Africa, where it is a primary source of food and income. However, cassava production is threatened by several viral diseases, including cassava brown streak disease (CBSD), which causes severe damage to the edible storage roots. Current cassava varieties in Africa lack effective resistance to this disease, leading to significant crop losses. We investigated the genetic diversity of cassava and identified new sources of resistance to the viruses causing CBSD. The cassava line, COL40, from a South American germplasm collection showed broad-spectrum resistance against all known strains of the viruses that cause this disease. To further understand the genetic basis of this resistance, we sequenced the genome of COL40 and produced a high-quality, haplotype-resolved genome assembly. This genomic resource provides new insights into cassava's genetic architecture, particularly in regions associated with disease resistance. The sequence reveals significant structural variation, including transposable elements, inversions, and deletions, which may contribute to the resistance phenotype. The reference genome assembly presented here will provide a valuable genomic resource for studying the cassava brown streak resistance and will help in accelerating breeding efforts to introduce virus resistance into African cassava varieties. By identifying genetic variants linked to resistance, future breeding programs can develop cassava cultivars that are more resilient to viral threats, enhancing food security and livelihoods for smallholder farmers across regions affected by the disease.

Keywords: Cassava genomes; Cassava brown streak resistance; haplotype-resolved genome assembly; transposable elements; chromosome 12

Introduction

Cassava (*Manihot esculenta*) is an important world food crop that produces starch-rich storage, which serves as a main source of food and income for people in Africa, South America, India, and Southeast Asia. The plant originated in the Amazon region of South America but is especially important in Sub-Saharan Africa, where half of the world's cassava is produced mainly by smallholder farmers. This resilient plant adapts well to diverse environments, is resistant to drought, grows in poor soils, and guarantees food security even in the most adverse environmental conditions. Nevertheless, cassava cultivation is constrained by pests and diseases, and because stem cuttings are taken to propagate the crop, diseases are maintained and disseminated by vegetative propagules. Viruses are associated with cassava cultivation in all cassava regions of the world and present serious threats to crop production. In Africa, viruses causing cassava mosaic disease are endemic, but the deployment of varieties with high virus

resistance has mitigated the impact of the diseases. In contrast, there was no means to control the spread of the viruses causing cassava brown streak disease (CBSD), which has a severe impact because of the root necrosis destroying the edible tubers of the crop. The relatively recent outbreak, first noted in 2004/2005 in Uganda (Alicai *et al.* 2007), the limited knowledge about the etiology and epidemiology of the disease, and the lack of virus-resistant cassava in Africa delayed comprehensive actions to control the disease. Cassava brown streak virus (CBSV, species *Ipomovirus brunusmanihotis*) and Ugandan cassava brown streak virus (UCBSV, species *Ipomovirus manihotis*) are closely related members of the genus *Ipomovirus* within the family *Potyviridae*. Both viruses are the causal agents of CBSD, sharing a high degree of genetic similarity and inducing similar symptoms in infected plants. However, they differ in genome sequences, virulence, and geographical distribution. Notably, CBSV is often associated with more severe symptoms compared with UCBSV, contributing

to greater yield losses in affected regions. Searching for new sources of resistance, we explored the diversity of South American cassava. We infected a subset of a germplasm collection held at International Center for Tropical Agriculture with diverse cassava brown streak viruses to identify lines that showed resistance to either isolates of the CBSV or against the UCBSV. After stringent infection and virus screening, 3 germplasm accessions COL2182, PER556, and COL40 showed broad-spectrum immunity to all species and strains of the viruses causing CBSD (Sheat et al. 2019). COL40, the most promising line, is now used as a source to provide CBSD resistance essentially in all cassava improvement programs. First prototypes from crosses with African lines are in the field, proving that high virus resistance in progenies from COL40 can be reached. Progress has been made in characterizing the resistance phenotype of COL40 (Sheat et al. 2021) and advancing field evaluation and resistance assessment (Sheat and Winter 2023). However, important steps toward molecular breeding, genomic prediction, marker-assisted selection, and a mechanistic explanation of the virus resistance in COL40 remain open challenges.

The reconstruction of complete, haplotype-resolved genome sequences of cassava genotypes is a substantial achievement in providing genome information and resources to advance our understanding of the cassava genome(s) and accelerate breeding to future-proof the crop for resilience to biotic and abiotic stresses, agronomic traits, and consumer demands. As a highly heterozygous species (Prochnik et al. 2012) (1% in TME 204; Qi et al. 2022) its genome comprises extensive amounts of repetitive elements (TMEB117 >60%; Landi et al. 2023), allelic variation, inversions, and deletions/insertions comprising large segments of the genome (Ramu et al. 2017). This genomic structural variation limits the use of a single reference genome sequence (AM560-2; Phytozome 13, v 8.1) that does not capture the full complement of sequence diversity of a crop species (Jayakodi et al. 2020) and may lack genes present in other genotypes, e.g. NBS-LRR genes relevant for disease resistance (Lozano et al. 2015).

Recently, high-quality haplotype-resolved genome sequences of African cassava lines were reported (Qi et al. 2022; Kuon et al. 2019; Mansfeld et al. 2021; Landi et al. 2023). The lines chosen were TME3 and 60444, to contrast a landrace with high resistance against viruses causing CMD, carrying the CMD2 resistance locus (TME3) and a highly susceptible line 60,444. Similarly, TME204 (Qi et al. 2022), carrying CMD2, and TMEB117 lacking (Landi et al. 2023) were chosen to provide high-resolution genome data for virus studies on resistance mechanisms and, to guide future breeding. Our study follows a similar rationale. By generating a high-quality COL40 genome sequence, we have established a reference genome, to advance our understanding about virus resistance/immunity in cassava. This resource enables the identification of genetic variants associated with resistance traits through genome-wide association studies and supports cassava resistance breeding efforts, with COL40 serving as a valuable genetic source.

Materials and methods

Plant cultivation

Cassava COL40 plantlets grown in tissue culture were transferred to pots and maintained at the greenhouse facility of the DSMZ Plant Virus Department in Braunschweig Germany at temperatures between 26 and 32°C and >75% relative humidity with additional light provided during German winter conditions.

DNA extraction, library preparation and sequencing

Prior to sampling, the plants were transferred to the dark for 24 h, after which fully-expanded leaves (young leaves; between 3rd and 5th leaf) were taken and flash frozen in liquid N₂. Leaves (10–15 g) were sent to Arizona Genomics Institute for high-molecular-weight (HMW) genomic DNA (gDNA) extraction, library preparation, and sequencing. HMW gDNA extraction was done essentially following the CTAB method of Doyle and Doyle, 1987 (Doyle and Doyle 1987), with minor modifications (Fu et al. 2017). The HiFi express template prep kit 2.0 was used for PacBio HiFi SMRTbell library construction. Fragments between 10 and 25-kb were selected for HiFi sequencing on the Sequel II instrument. Sequel v2.0 chemistry was used with 2 SMRT cells 8M v1 in circular consensus sequencing mode for a 30-h run time per cell.

For RNA-seq analysis, leaf, stem, and root samples were taken from COL40 and flash frozen in liquid N₂ prior to extraction of total RNA using a RNA extraction kit following the manufacturer's protocol (Epoch, United States). RNA was quantified in a QubitR fluorometer (Thermo Fisher Scientific, United States) using the Qubit RNA BR Assay Kit (Thermo Fisher Scientific, United States), and checked for size distribution in a bioanalyzer (Agilent) prior to sequencing at Novogene (Science Park, Milton, Cambridge, UK). At Novogene, mRNA was purified from total RNA preparations using poly-T magnetic beads, fragmented and subjected to first strand cDNA synthesis using random hexamer primers followed by second strand cDNA synthesis and end repair. Library construction, quality controls, and paired-end sequencing (PE 150) on an Illumina Novaseq 6000 platform followed Novogene's mRNA sequencing workflow.

Assembly generation and quality control

The HiFi reads of the COL40 cultivar were assembled by Hifiasm 0.19.8-r603 (Cheng et al. 2021, 2022), NextDenovo 2.5.0 (Hu et al. 2024), Flye 2.9.3-b1797 (Kolmogorov et al. 2019), and Shasta 0.11.1 (Shafin et al. 2020). To avoid compatibility issues with tools in downstream analysis, contig identifiers of the assemblies were cleaned using the python script `clean_genomic_fasta.py` v0.15 (Meckoni, et al. 2023). Assembly statistics were calculated with the python script `contig_stats.py` v1.31 (Pucker et al. 2016). Completeness of universal single-copy orthologs was checked with BUSCO 5.2.2 and the eudicots_odb10 dataset (Manni et al. 2021). A k-mer based analysis was performed with Merqury (Rhie et al. 2020) and k-mers generated from the cultivars HiFi reads ($k=21$). The assembly's coverage was analyzed by mapping the HiFi reads to the assembly with minimap2 2.24-r1122 and “-ax map-hifi -secondary=no” option (Li, 2018). Samtools 1.17-29-gcc18465 (Bonfield et al. 2021) was applied for conversion and sorting of the mapping file. The coverage was calculated with the “genome-cov” command of bedtools v2.30.0 (Quinlan and Hall 2010).

Scaffolding of the assembly

Genetic markers of the composite genetic map of *M. esculenta* Crantz by the International Cassava Genetic Map Consortium (ICGMC) (Supplementary File 2, International Cassava Genetic Map Consortium (ICGMC), 2014) were used for scaffolding. The genetic markers were mapped to the assemblies using the python script `genetic_map_to_fasta.py` v0.2 (<https://github.com/c-thoben/CassavaGenomicsProject>), which creates a csv map based on the best BLAST hits for each marker with a minimum similarity of 99.0% and score of 175. The scaffolding was performed by converting the CSV map with ALLMAPS (JCVI utility libraries 1.4.2)

(Tang et al. 2015) “merge” option and path construction using the “path” option. Scaffold statistics were calculated with the python script `contig_stats.py` v1.31 (Pucker et al. 2016). For the Hifiasm assembly, the scaffolding was performed for each haplotype separately. The scaffolds of both haplotypes were merged together, and for each scaffold, a suffix was added to the scaffold IDs indicating its haplotype. The coverage plot of the chromosomes including the density of the transposable elements was created with the python script `chromosome_coverage_te_plot.py` v0.4 (Pucker et al. 2016; <https://github.com/c-thoben/CassavaGenomicsProject>).

Prediction and functional annotation of polypeptide sequences

In total, 21 RNA-Seq libraries of the COL40 cultivar (Supplementary Additional File 5) and 24 further RNA-Seq libraries (Supplementary Additional File 5) from other cultivars were mapped to the COL40 assembly with HISAT2 2.2.1 (Kim et al. 2019) and “-dta” option. Samtools 1.17-29-gcc18465 (Danecek et al. 2021) was applied for conversion and sorting of the mapping files (Bonfield et al. 2021). Together with external protein hints from *M. esculenta* Crantz v8.1 (International Cassava Genetic Map Consortium (ICGMC), 2014) and the OrthoDB 11 Viridiplantae database (Kuznetsov et al. 2023), the mapped RNA-Seq reads were given to BRAKER3v3.0.6 (Stanke et al. 2006, 2008; Gabriel et al. 2021, 2024) as external hints to perform the structural annotation. Based on the mapped RNA-Seq reads, the coverage of the structural annotation was analyzed with the python script `RNAseq_cov_analysis.py` v0.1 (Meckoni et al. 2023; <https://github.com/bpucker/GenomeAssembly/>) and filtered for transcripts with a coverage >90%. Completeness of the predicted polypeptide sequences was checked with BUSCO v5.2.2 and the eudicots_odb10 (Manni et al. 2021) dataset for each haplotype separately. The polypeptide sequences were functionally annotated using InterProScan5 v5.67-99.0 (Manni et al. 2021; Blum et al. 2021).

Prediction and analysis of transposable elements

Transposable elements were predicted with EDTA v2.1.0 (Ou et al. 2019) and the “-overwrite 1 -sensitive 1 -anno 1 -evaluate 1” options. Annotated coding sequences predicted by BRAKER3 were provided with the “-cds” option. Annotated transposable elements were further analyzed with the script `COL40_TE_repeat_analysis.ipynb` (<https://github.com/c-thoben/CassavaGenomicsProject>) based on the R script `TMEB117TEandGeneAnnotation.R` provided by (Landi et al. 2023; <https://github.com/LandiMi2/GenomeAssemblyTMEB117>). The R package `circize` (Gu et al. 2014) was used to calculate the density of transposable elements and predicted coding sequences with the `circos` function `genomicDensity` and visualize them in a `circos` plot. The coverage plot of the chromosomes including the mean coverage over 1 kbp and the density of the transposable elements over 1 Mbp was

created with the python script `coverage_te_plot.py` v0.5 (Pucker et al. 2016; <https://github.com/c-thoben/CassavaGenomicsProject>).

Results

Assembler comparison and quality control

The HiFi dataset consisted of 3,440,114 reads with an N50 length of 18,106 and 38% GC content. All assemblies were >80% complete representations of the genome according to the k-mer analysis and, with exception of the Shasta assembly, show a high completeness regarding the single-copy universal orthologs. The Hifiasm (N50: 35 Mbp, 30 Mbp) and NextDenovo (N50: 6.7 Mbp) assemblers performed best in producing assemblies with a high continuity (Table 1).

The Hifiasm assembly had the highest k-mer completeness (98%, ~80% for each haplotype) and a good assembly consensus quality value (QV) of 58 according to the k-mer analysis. The analysis of the k-mer copy numbers (Fig. 1) demonstrates that the haplotypes are well-resolved in this assembly.

The Shasta assembly had the highest assembly quality (QV 61) in the k-mer analysis, standing out in this specific criterion, but was outperformed by the Hifiasm assembly in the remaining criteria. Therefore, the Hifiasm assembly was chosen as the representative genome sequence of COL40.

Scaffolding and structural annotation

In the scaffolding process, 16,233 out of 22,403 markers were mapped to the contigs of haplotype A and used to scaffold 23 contigs representing the 18 *M. esculenta* chromosomes (Supplementary Additional Files 1 and 2). The 18 chromosomes of haplotype B were scaffolded from 37 contigs with reference to 16,169 mapped markers. The remaining 2,582 contigs for haplotype A and 409 contigs for haplotype B remained unplaced. The structural annotation of the Hifiasm assembly predicted 82,151 polypeptide sequences for haplotype A and 72,383 polypeptide sequences for haplotype B. After filtering the transcripts based on their coverage in the mapped RNA-Seq reads, the number of sequences was reduced to 36,064 for haplotype A and 34,029 for haplotype B. The BUSCO analysis revealed a high completeness of 96.6% (D: 22.7%) for haplotype A, 96.7% (D: 21.7%) for haplotype B, and 98.6% (D: 94.6%) for both haplotypes.

Assembly coverage analysis

The coverage analysis revealed an average coverage of 40.4-fold for haplotype A and 39.7-fold for haplotype B. The coverage plot demonstrates an equal distribution of the coverage, which can be attributed to the high-resolution of the haplotypes in the assembly (Supplementary Additional Files 3 and 4). For both haplotypes, a coverage drop at the end of chromosome 12 (downstream of ~39 Mbp for haplotype A, ~37 Mbp for haplotype B) is displayed in the coverage plot (Fig. 2). This is also

Table 1. Assembly statistics and quality control of assemblies.

	Hifiasm	NextDenovo	Flye	Shasta
Assembly size	823 Mbp, 694 Mbp	804 Mbp	693 Mbp	1,015 Mbp
Number of contigs	2605, 446	394	9217	7692
N50 of contigs	35 MB, 30 Mbp	6.7 Mbp	119 kbp	1.6 Mbp
Number of scaffolds	2600, 427	224	6976	6334
N50 of scaffolds	37 MB, 36 Mbp	41 Mbp	545 kbp	41 Mbp
BUSCO	98.8% (D: 8.6%), 98.8% (D: 8.4%)	99.0% (D: 13.4%)	98.8% (D: 17.2%)	75.5% (D: 58.9%)
K-mer completeness	82, 81 (asm: 98)	86	86	82
QV	57, 63 (asm: 58)	50	36	61

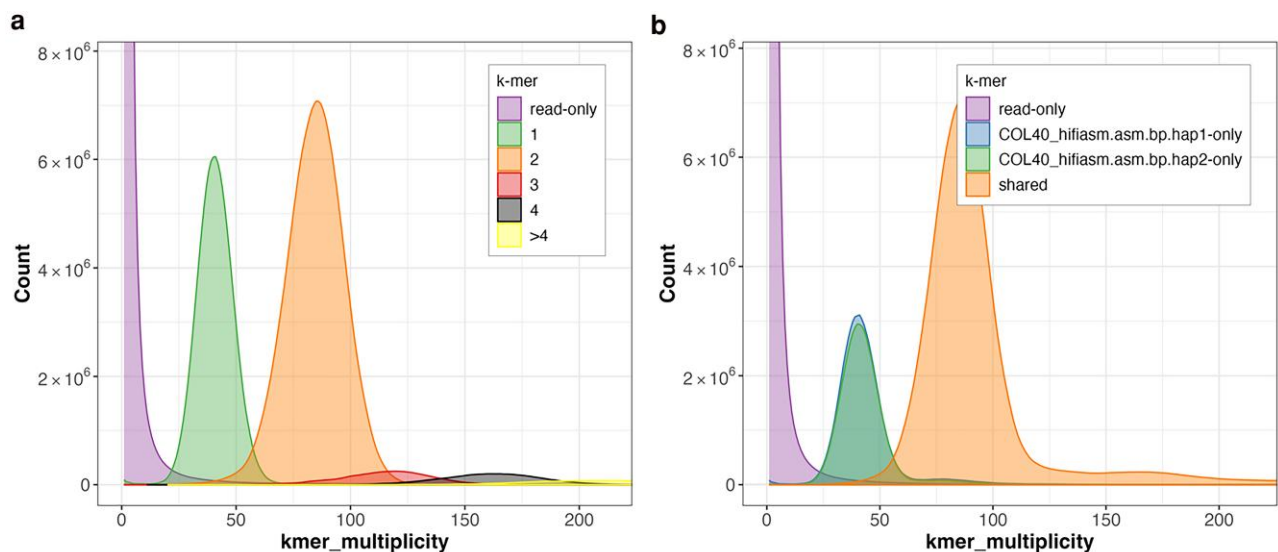


Fig. 1. Assembly completeness according to k-mer analysis of the assembly. a) Merqury copy number spectrum plot of the HiFi reads. The green peak represents 1-copy reads. The orange peak represents 2-copy reads. HiFi reads not included in the assembly are represented by the purple peak. b) Copy number spectrum plot of the k-mers in A. Heterozygous reads are colored green or blue according to the haplotype, overlapping peaks. Homozygous reads or haplotype-specific duplications are colored orange.

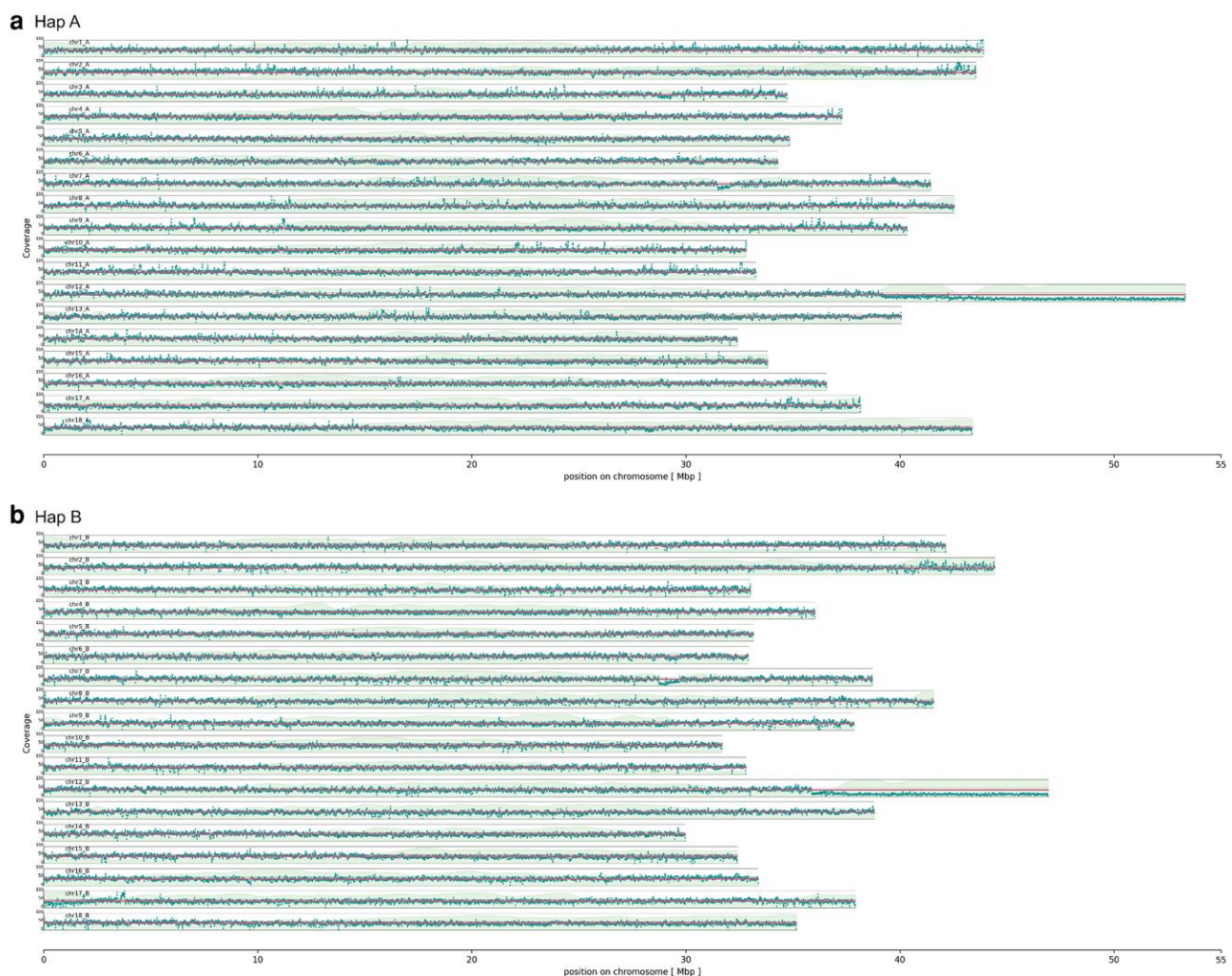


Fig. 2. Coverage plot of the chromosome-resolved Hifiasm assembly. The HiFi reads were mapped to the haplotype A a) and haplotype B b) of the scaffolded assembly and the mean coverage was calculated for each haplotype (red line) and block-wise over 1 kbp intervals for each chromosome (dark green dots). The density of transposable elements was calculated block-wise per 1 Mbp interval (light green area).

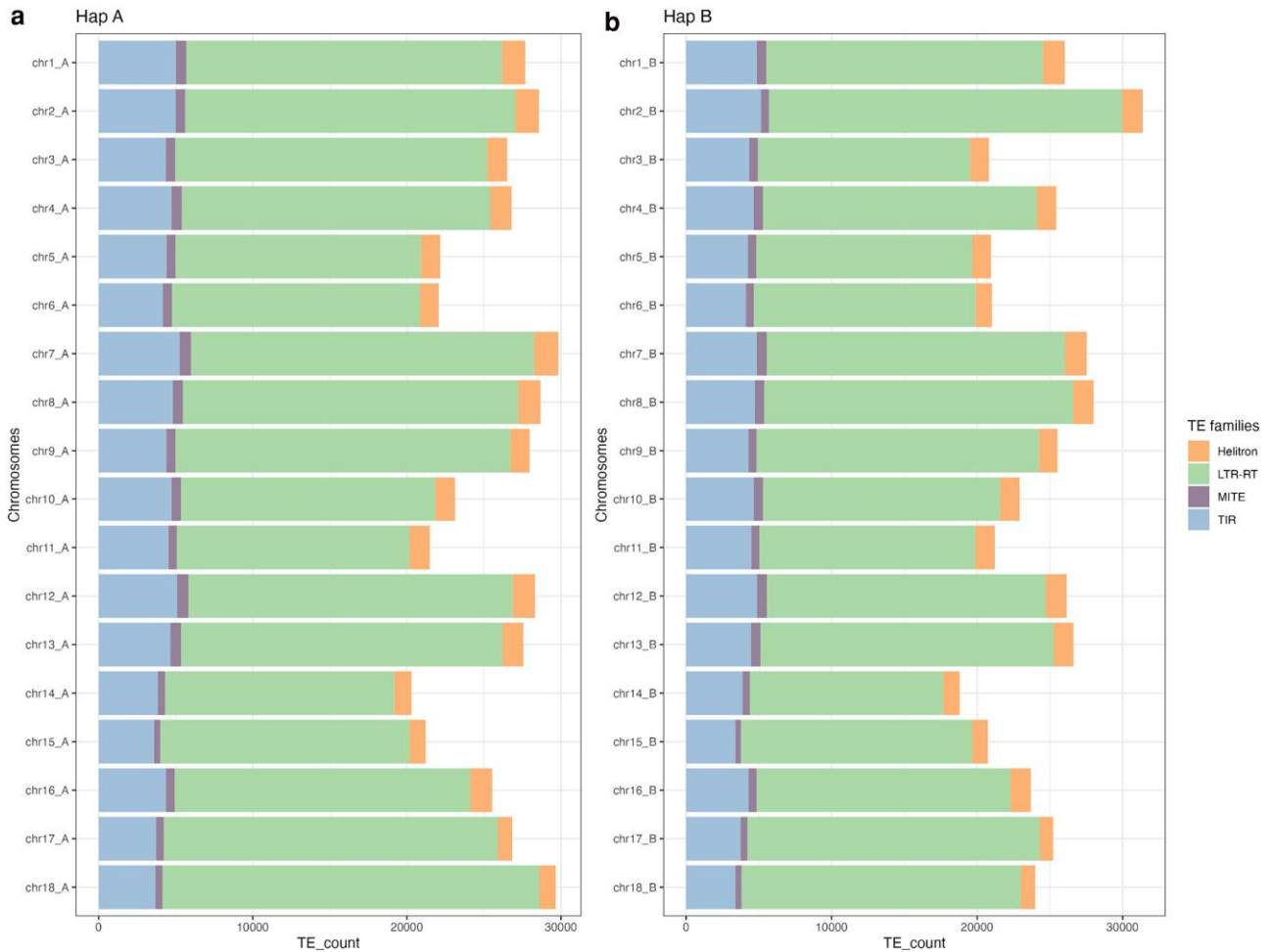


Fig. 3. Distribution of transposable element families for pseudochromosomes of haplophase A a) and haplophase B b) in the Hifiasm assembly. Unclassified transposable elements were removed from the annotation. For all pseudochromosomes, LTR-RTs are the most abundant repeats for both haplophases.

reflected in the coverage histogram of chromosome 12, which shows a second peak around a coverage of 15 for both haplophases (Supplementary Additional Files 3 and 4). For both haplophases, the density of transposable elements is enhanced in these regions, which is further discussed below (Figs. 2 and 3).

Annotation of transposable elements

In total, 3,027 repeat regions were annotated, which cover 58.04% of the COL40 assembly. The most abundant transposable elements (TE) are long terminal repeats—retrotransposons (LTR-RTs), mostly Gypsy LTR-RT, which cover 51.88% of the complete assembly and are the most abundant family on all chromosomes (Fig. 3).

As described above, the density of TE increases at the end of chromosome 12 in both haplophases (Fig. 4). This increase can be attributed to fragmented transposable elements, as the density of structurally intact transposable elements decreases. Furthermore, the density of predicted coding sequences is distinctly reduced in this region. The scaffolding results (Supplementary Additional Files 1 and 2) show that no markers of the *M. esculenta* Crantz genetic map by the International Cassava Genetic Map Consortium (ICGMC) [Supplementary File 2, International Cassava Genetic Map Consortium (ICGMC), 2014] were mapped to the region, suggesting that the transposable elements in this area are either not represented in the map or that

markers could not be developed due to the high level of repetitive sequences present in the region.

Discussion

In this study, we presented a haplotype-resolved genome assembly of the cassava line COL40, a South American germplasm line with high and broad-spectrum resistance to the viruses causing CBSD. The assembly of COL40 joins a growing collection of high-quality genome assemblies from other cassava cultivars, such as TMEB117, TME204, and TME3 (Kuon et al. 2019; Qi et al. 2022; Landi et al. 2023), and highlighting the increasing role of advanced sequencing technologies in improving crop genomics.

Using PacBio HiFi reads and after a comparative evaluation of assembly tools, we selected Hifiasm as the optimal assembler due to its superior continuity and completeness. Hifiasm also generated best assembly in other cassava genome studies (Kuon et al. 2019; Landi et al. 2023) with haplophase sizes between 665 and 823 Mbp (Table 2).

The high accuracy of HiFi data was crucial for achieving a robust genome assembly with phased haplotypes. Similarly, PacBio HiFi sequencing reported for the assembly of other plant genomes, including *Haloxylon ammodendron* (Wang et al. 2022), and *Rhododendron vialii* (Chang et al. 2023) allowed generation of highly contiguous genome assemblies for diploid and polyploid

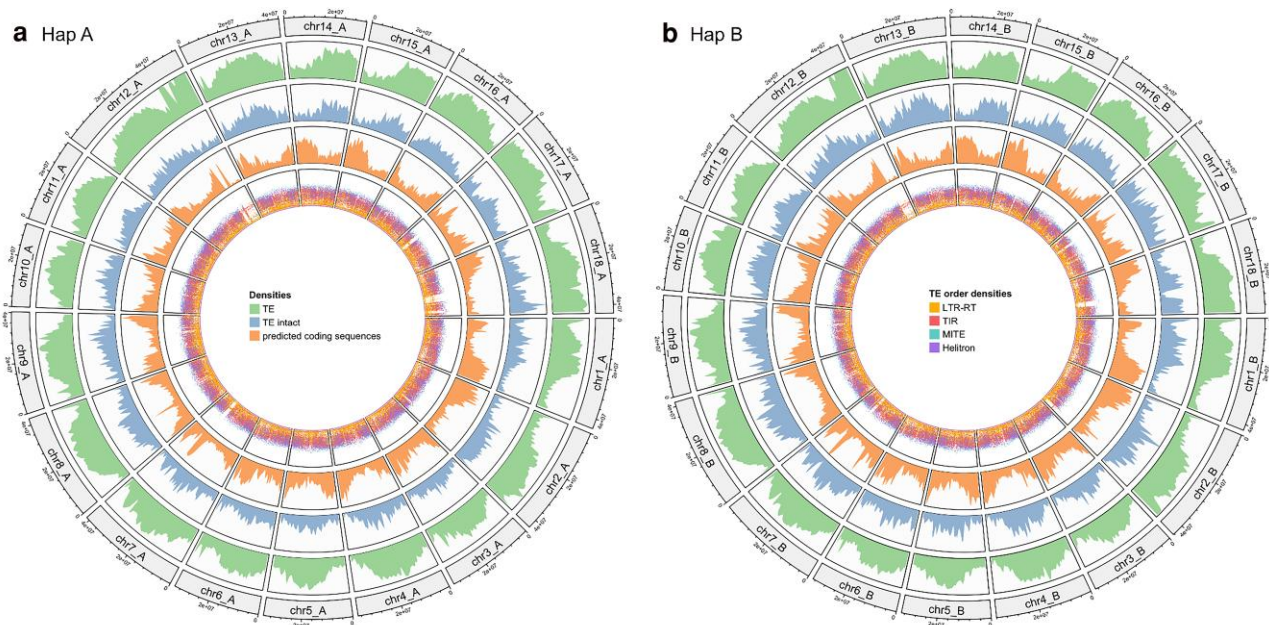


Fig. 4. Density of transposable elements and predicted coding sequences across the pseudochromosomes of haplophase A a) and haplophase B b) in the Hifiasm assembly. For all tracks, the density was calculated over a window size of 1 Mbp. The density of transposable elements is shown in the outer track and reflects the percent of the window covered by the input regions (green). The second track shows the density of structurally intact transposable elements (blue). The density of predicted coding sequences is displayed in the third track (orange). In the inner track, a rainbow plot represents the minimal distance ($\log_{10}$ -transformed) to the neighboring regions. The color of the region displays the transposable elements order.

species and made it ideal for resolving the complex structure of cassava's genome.

However, we recognize that Oxford Nanopore Technology has made significant advancements, particularly in read length and base-calling accuracy, allowing for highly contiguous and scaffolded genome assemblies. As sequencing technologies continue to evolve, leveraging the strengths of both approaches may offer the most comprehensive and accurate genome reconstructions.

Assemblies comparison among published data and our data present a high completeness of >98% in both BUSCO and k-mer analysis and around 80% for each haplophase in k-mer analysis (Table 2).

The observed duplication of BUSCOs among the predicted proteins is a result of the assembly being derived from the merging of haplophase A and haplophase B. Since most regions of the assembly are represented in both haplophases, the assembly predominantly exhibits a diploid nature, leading to duplication. Additionally, regions that were already duplicated within one or both haplophases before merging naturally appear as triploid or tetraploid in the final assembly. This explains the higher number of duplicate BUSCOs detected in the predicted proteins compared with those directly identified on the scaffolds.

Our results are consistent with other assemblies of complex plant genomes like that of the European common bean (Carrère et al. 2023). This level of completeness indicates that our assembly has captured the vast majority of the gene space, an essential feature for downstream functional studies, particularly for understanding CBS resistance mechanisms.

Comparing the QV of different cassava assemblies (Table 2), we observed that the COL40 genome assembly had a QV of 57 for one haplotype and 63 for the other, which is close to the TMEB117 assembly (Landi et al. 2023). In contrast, earlier assemblies, such as TME204 (Qi et al. 2022), had significantly lower QVs. These improvements in genome quality metrics reflect the advancements

made by HiFi sequencing, as highlighted by Sun et al., (2022), where HiFi reads produced near-perfect assemblies for various complex plant genomes.

Another outstanding feature of our assembly is that we resolved high-quality polypeptide sequences, with a completeness of around 97% for both haplotypes. These results align with other high-quality plant genome assemblies produced using HiFi data, such as the assembly of *Salix wilsonii* (Han et al. 2022) and the bean genome (Carrère et al. 2023). These findings underscore the value of using HiFi reads, as they produce long, highly accurate reads that are crucial for resolving complex regions and providing reliable gene annotations.

The COL 40 assembly provides critical insights into the repetitive regions and structural variations in cassava genomes. The coverage analysis along the pseudochromosomes after scaffolding of the COL40 assembly revealed a region with reduced coverage at the end of chromosome 12 on both haplophases (Fig. 2). The annotation of this assembly region showed a high density of fragmented transposable elements TE (Fig. 4), which was also found in chromosome 12 of TMEB117 (Landi et al. 2023). This region is not covered by genetic markers in the composite genetic map of *M. esculenta* Crantz by the International Cassava Genetic Map Consortium (ICGMC) (Supplementary Additional File 2), suggesting that the TEs in this region are either not represented or, that the high frequency of repetitive sequences hindered the development of respective markers. Nevertheless, this region may not be present in all cassava genomes but it is also likely that earlier assemblies based on short sequencing reads may have not resolved repetitive regions and consequently created gaps and fragmentation of the transposable elements. Highly TE-rich regions present challenges to the completeness of genome assemblies, for cassava and other species (Benham et al. 2024).

The COL40 assembly focuses on haplotype resolution and genome completeness. In contrast, the telomere-to-telomere

Table 2. Comparison of recently published cassava assemblies.

	COL40	TMEB117	TME204	TME3
Chromosomes resolved	yes	yes	yes	No
Haplotype-resolved	yes	yes	yes	No
Assembly size	823 MB, 694 Mbp	694 MB, 665 Mbp	762 MB, 706 Mbp	1,225 Mbp
N50 of scaffolds	37 MB, 36 Mbp	38 MB, 36 Mbp	18 MB, 26 Mbp	53 Mbp
BUSCO	98.8% (D: 8.6%), 98.8% (D: 8.4%)	98.9% (D: 8.8%), 98.9% (D: 8.8%)	99% (D: 4.9%), 98.8% (D: 4.2%)	94.8% (D: 19.7%)
K-mer completeness	82%, 81% (asm: 98%)	78.63%, 77.95% (asm: 98.79%)	79.6%, 79.1% (asm: 98.4%)	–
QV	57, 63	64, 68	45, 49	–
BUSCO (Predicted polypeptide sequences)	96.6% (D: 22.7%), 96.7% (D: 21.7%)	~90% both	96.7% (D: 17.4%), 96.7% (D: 16%)	–
TE proportion	58.04%	57.37%, 54.42%	>60%	64.81%
Reference	This study	Landi et al. 2023 https://doi.org/10.1038/s41597-023-02800-0	Kuon et al. 2019 https://doi.org/10.1186/s12915-019-0697-6	Qi et al. 2022 https://doi.org/10.1093/gigascience/giac028

assembly of the cassava Xinxuan 048 delved deeper into epigenetic analysis (Xu et al. 2023), highlighting that the structural variation between the haploid genomes of Xinxuan 048 was mainly due to TE insertions. The variation led to differences in CG methylation of alleles, thus resulting in differential allelic expression. These findings suggest that differential gene expression in Xinxuan 048 could be influenced by methylation changes caused by varying TE insertions.

The proportion of TE elements in the genome of COL 40 is around 60%, and this is consistently found in all cassava genome assemblies (Kuon et al. 2019; Landi et al. 2023). These TE-rich regions may hold essential genes for interesting traits, such as those associated with CBSD resistance.

In conclusion, the high-quality, haplotype-resolved genome assembly of cassava COL40 contributes significantly to cassava genomic resources. It lays the groundwork for further comparative analyses between cassava cultivars and provides an invaluable resource to advance our understanding of virus resistance and immunity.

Data availability

The data that support the findings of this study have been deposited into HiFi reads of the COL40 cultivar are available from the European Nucleotide Archive (ENA): ERS20926319. The 21 COL40 RNA-Seq data sets are also available via ENA: ERS20926298, ERS20926299, ERS20926300, ERS20926301, ERS20926302, ERS20926303, ERS20926304, ERS20926305, ERS20926306, ERS20926307, ERS20926308, ERS20926309, ERS20926310, ERS20926311, ERS20926312, ERS20926313, ERS20926314, ERS20926315, ERS20926316, ERS20926317, ERS20926318, and ERS20926319. Genome assembly/annotation on LeoPARD at <https://doi.org/10.24355/dbbs.084-202409230712-0>. The assembly and the corresponding annotation are published in NCBI accession nr PRJNA1172196 https://www.ncbi.nlm.nih.gov/bioproject?LinkName=bioproject_bioproject_d2u&from_uid=1172196. Scripts used for the generation and analysis of the genome sequence assembly are available via GitHub repository: <https://github.com/c-thoben/CassavaGenomicsProject>.

Supplemental material available at G3 online.

Acknowledgments

We are grateful for all assistance in the experiments provided by Marianne Koerbler, Michelle “Skadi” Falk and Agnes Pietruszka

for the excellent and skillful handling of all cassava plant materials, tissue cultures and nucleic acid preparations. This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

Funding

This study was funded by the Bill & Melinda Gates Foundation through the Next Generation Cassava Breeding Project, agreement no. 84941-11220 (under prime agreement no. OPP1175661).

Conflicts of interest

The author(s) declare no conflicts of interest.

Author contributions

SS, SW, and BP conceived the idea; CT and BP devised the bioinformatic approach; SS and SW handled the plant materials and RNA-seq; BFE handled PacBio HiFi sequencing; CT analyzed the data and compiled the workflows. All authors contributed to the original draft of the manuscript, and reviewed and approved the final manuscript.

Literature cited

- Alicai T, Omongo C, Maruthi M, Hillocks R, Baguma Y, Kawuki R, Bua A, Otim-Nape G, Colvin J. 2007. Re-emergence of cassava brown streak disease in Uganda. *Plant Dis.* 91(1):24–29. doi:[10.1094/PD-91-0024](https://doi.org/10.1094/PD-91-0024).
- Benham PM, Cicero C, Escalona M, Beraut E, Fairbairn C, Marimuthu MPA, Nguyen O, Sahasrabudhe R, King BL, Thomas WK, et al. 2024. Remarkably high repeat content in the genomes of sparrows: the importance of genome assembly completeness for transposable element discovery. *Genome Biol Evol.* 16(4):evae067. doi:[10.1093/gbe/evae067](https://doi.org/10.1093/gbe/evae067).
- Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49(D1):D344–D354. doi:[10.1093/nar/gkaa977](https://doi.org/10.1093/nar/gkaa977).
- Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, Keane T, Davies RM. 2021. HTSlib: C library for reading/writing

- high-throughput sequencing data. *Gigascience*. 10(2):giab007. doi:[10.1093/gigascience/giab007](https://doi.org/10.1093/gigascience/giab007).
- Carrère S, Mayjonade B, Lalanne D, Gaillard S, Verdier J, Chen NW. 2023. First whole genome assembly and annotation of a European common bean cultivar using PacBio HiFi and iso-seq data. *Data Brief*. 48:109182. doi:[10.1016/j.dib.2023.109182](https://doi.org/10.1016/j.dib.2023.109182).
- Chang Y, Zhang R, Ma Y, Sun W. 2023. A haplotype-resolved genome assembly of *Rhododendron vialii* based on PacBio HiFi reads and Hi-C data. *Sci Data*. 10(1):451. doi:[10.1038/s41597-023-02362-1](https://doi.org/10.1038/s41597-023-02362-1).
- Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 18(2):170–175. doi:[10.1038/s41592-020-01056-5](https://doi.org/10.1038/s41592-020-01056-5).
- Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, Li H. 2022. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol*. 40(9):1332–1335. doi:[10.1038/s41587-022-01261-x](https://doi.org/10.1038/s41587-022-01261-x).
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience*. 10(2):giab008. doi:[10.1093/gigascience/giab008](https://doi.org/10.1093/gigascience/giab008).
- Doyle JJ, Doyle JL. 1987. A Rapid DNA Isolation Procedure for Small Quantities of Fresh Leaf Tissue. *worldveg.tind.io: Phytochemical Sect., Botanical Soc. of America*.
- Fu Z-Y, Song J-C, Jameson PE. 2017. A rapid and cost effective protocol for plant genomic DNA isolation using regenerated silica columns in combination with CTAB extraction. *J Integr Agric*. 16(8):1682–1688. doi:[10.1016/S2095-3119\(16\)61534-4](https://doi.org/10.1016/S2095-3119(16)61534-4).
- Gabriel L, Bruna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M. 2024. BRAKER3: fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res*. 34(5):769–777. doi:[10.1101/gr.278090.123](https://doi.org/10.1101/gr.278090.123).
- Gabriel L, Hoff KJ, Bruna T, Borodovsky M, Stanke M. 2021. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics*. 22(1):566. doi:[10.1186/s12859-021-04482-0](https://doi.org/10.1186/s12859-021-04482-0).
- Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. (2014). “Circlize” implements and enhances circular visualization in R. *Bioinformatics*. 2014;30(19):2811–2812. doi:[10.1093/bioinformatics/btu393](https://doi.org/10.1093/bioinformatics/btu393).
- Han F, Qu Y, Chen Y, Xu LA, Bi C. 2022. Assembly and comparative analysis of the complete mitochondrial genome of *Salix wilsonii* using PacBio HiFi sequencing. *Front Plant Sci*. 13:1031769. doi:[10.3389/fpls.2022.1031769](https://doi.org/10.3389/fpls.2022.1031769).
- Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper D, Ye K. 2024. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol*. 25(1):107. doi:[10.1186/s13059-024-03252-4](https://doi.org/10.1186/s13059-024-03252-4).
- International Cassava Genetic Map Consortium (ICGMC). 2014. High-resolution linkage map and chromosome-scale genome assembly for cassava (*Manihot esculenta* crantz) from 10 populations. *G3 (Bethesda)*. 5(1):133–144. doi:[10.1534/g3.114.015008](https://doi.org/10.1534/g3.114.015008).
- Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, et al. 2020. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*. 588(7837):284–289. doi:[10.1038/s41586-020-2947-8](https://doi.org/10.1038/s41586-020-2947-8).
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 37(8):907–915. doi:[10.1038/s41587-019-0201-4](https://doi.org/10.1038/s41587-019-0201-4).
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 37(5):540–546. doi:[10.1038/s41587-019-0072-8](https://doi.org/10.1038/s41587-019-0072-8).
- Kuon J-E, Qi W, Schläpfer P, Hirsch-Hoffmann M, von Bieberstein PR, Patrignani A, Poveda L, Grob S, Keller M, Shimizu-Inatsugi R. 2019. Haplotype-resolved genomes of geminivirus-resistant and geminivirus-susceptible African cassava cultivars. *BMC Biol*. 17(1):75. doi:[10.1186/s12915-019-0697-6](https://doi.org/10.1186/s12915-019-0697-6).
- Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, Evgenia V. & Zdobnov, E. M. (2023). OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res*. 51(D1), D445–D451. doi:[10.1093/nar/gkac998](https://doi.org/10.1093/nar/gkac998).
- Landi M, Shah T, Falquet L, Niazi A, Stavalone L, Bongcam-Rudloff E, Gisel A. 2023. Haplotype-resolved genome of heterozygous African cassava cultivar TMEB117 (*Manihot esculenta*). *Sci Data*. 10(1):887. doi:[10.1038/s41597-023-02800-0](https://doi.org/10.1038/s41597-023-02800-0).
- Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34(18):3094–3100. doi:[10.1093/bioinformatics/bty191](https://doi.org/10.1093/bioinformatics/bty191).
- Lozano R, Hamblin MT, Prochnik S, Jannink JL. 2015. Identification and distribution of the NBS-LRR gene family in the Cassava genome. *BMC Genomics*. 16(1):360. doi:[10.1186/s12864-015-1554-9](https://doi.org/10.1186/s12864-015-1554-9).
- Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021. BUSCO: assessing genomic data quality and beyond. *Curr Protoc*. 1(12):e323. doi:[10.1002/cpz1.323](https://doi.org/10.1002/cpz1.323).
- Mansfeld BN, Boyher A, Berry JC, Wilson M, Ou S, Polydore S, Michael TP, Fahlgren N, Bart RS. 2021. Large structural variations in the haplotype-resolved African cassava genome. *Plant J*. 108(6):1830–1848. doi:[10.1111/tpj.15543](https://doi.org/10.1111/tpj.15543).
- Meckoni SN, Nass B, Pucker B. 2023. Phylogenetic placement of *Ceratophyllum submersum* based on a complete plastome sequence derived from nanopore long read sequencing data. *BMC Res Notes*. 16(1):187. doi:[10.1186/s13104-023-06459-z](https://doi.org/10.1186/s13104-023-06459-z).
- Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 20(1):275. doi:[10.1186/s13059-019-1905-y](https://doi.org/10.1186/s13059-019-1905-y).
- Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, Rodriguez F, Fauquet C, Tohme J, Harkins T, et al. 2012. The Cassava genome: current progress, future directions. *Trop Plant Biol*. 5(1):88–94. doi:[10.1007/s12042-011-9088-z](https://doi.org/10.1007/s12042-011-9088-z).
- Pucker B, Holtgräwe D, Rosleff Sørensen T, Stracke R, Viehöver P, Weisshaar B. 2016. A De Novo genome sequence assembly of the *Arabidopsis thaliana* accession Niederzenz-1 displays presence/absence variation and strong synteny. *PLoS One*. 11(10):e0164321. doi:[10.1371/journal.pone.0164321](https://doi.org/10.1371/journal.pone.0164321).
- Qi W, Lim YW, Patrignani A, Schlapfer P, Bratus-Neuenschwander A, Gruter S, Chanez C, Rodde N, Prat E, Vautrin S, et al. 2022. The haplotype-resolved chromosome pairs of a heterozygous diploid African cassava cultivar reveal novel pan-genome and allele-specific transcriptome features. *Gigascience*. 11:giac028. doi:[10.1093/gigascience/giac028](https://doi.org/10.1093/gigascience/giac028).
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26(6):841–842. doi:[10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033).
- Ramu P, Esuma W, Kawuki R, Rabbi IY, Egesi C, Bredeson JV, Bart RS, Verma J, Buckler ES, Lu F. 2017. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat Genet*. 49(6):959–963. doi:[10.1038/ng.3845](https://doi.org/10.1038/ng.3845).
- Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merquary: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 21(1):245. doi:[10.1186/s13059-020-02134-9](https://doi.org/10.1186/s13059-020-02134-9).

- Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, *et al.* 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol.* 38(9):1044–1053. doi:[10.1038/s41587-020-0503-6](https://doi.org/10.1038/s41587-020-0503-6).
- Sheat S, Fuerholzner B, Stein B, Winter S. 2019. Resistance against Cassava brown streak viruses from Africa in cassava germplasm from South America. *Front Plant Sci.* 10:567. doi:[10.3389/fpls.2019.00567](https://doi.org/10.3389/fpls.2019.00567).
- Sheat S, Margaria P, Winter S. 2021. Differential tropism in roots and shoots of resistant and susceptible cassava (*Manihot esculenta* Crantz) infected by Cassava Brown Streak viruses. *Cells.* 10(5): 1221. doi:[10.3390/cells10051221](https://doi.org/10.3390/cells10051221).
- Sheat S, Winter S. 2023. Developing broad-spectrum resistance in cassava against viruses causing the cassava mosaic and the cassava brown streak diseases. *Front Plant Sci.* 14:1042701. doi:[10.3389/fpls.2023.1042701](https://doi.org/10.3389/fpls.2023.1042701).
- Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics.* 24(5):637–644. doi:[10.1093/bioinformatics/btn013](https://doi.org/10.1093/bioinformatics/btn013).
- Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics.* 7(1): 62. doi:[10.1186/1471-2105-7-62](https://doi.org/10.1186/1471-2105-7-62).
- Sun Y, Shang L, Zhu Q-H, Fan L, Guo L. 2022. Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.* 27(4):391–401. doi:[10.1016/j.tplants.2021.10.006](https://doi.org/10.1016/j.tplants.2021.10.006).
- Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, Lu J. 2015. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* 16(1):3. doi:[10.1186/s13059-014-0573-1](https://doi.org/10.1186/s13059-014-0573-1).
- Wang M, Zhang L, Tong S, Jiang D, Fu Z. 2022. Chromosome-level genome assembly of a xerophytic plant, *Haloxylon ammodendron*. *DNA Res.* 29(2):dsac006. doi:[10.1093/dnares/dsac006](https://doi.org/10.1093/dnares/dsac006).
- Xu X-D, Zhao R-P, Xiao L, Lu L, Gao M, Luo Y-H, Zhou Z-W, Ye S-Y, Qian Y-Q, Fan B-L. 2023. Telomere-to-telomere assembly of cassava genome reveals the evolution of cassava and divergence of allelic expression. *Hortic Res.* 10(11):uhad200. doi:[10.1093/hr/uhad200](https://doi.org/10.1093/hr/uhad200).

Editor: J. Udall