

# The haplotype-resolved chromosome pairs of a heterozygous diploid African cassava cultivar reveal novel pan-genome and allele-specific transcriptome features

Weihong Qi<sup>1,2,3,\*†</sup>, Yi-Wen Lim<sup>2,†</sup>, Andrea Patrignani<sup>1</sup>, Pascal Schlöpfer<sup>2</sup>, Anna Bratus-Neuenschwander<sup>1</sup>, Simon Grüter<sup>1</sup>, Christelle Chanez<sup>2</sup>, Nathalie Rodde<sup>4</sup>, Elisa Prat<sup>4</sup>, Sonia Vautrin<sup>4</sup>, Margaux-Alison Fustier<sup>4</sup>, Diogo Pratas<sup>5,6</sup>, Ralph Schlapbach<sup>1</sup> and Wilhelm Gruissem<sup>1,2,7,\*</sup>

<sup>1</sup>Functional Genomics Center Zurich, ETH Zurich and University of Zurich, Winterthurerstrasse 190, 8057, Zurich, Switzerland

<sup>2</sup>Department of Biology, Institute of Molecular Plant Biology, ETH Zurich, Universitätstrasse 2, 8092, Zurich, Switzerland

<sup>3</sup>SIB Swiss Institute of Bioinformatics, 1202, Geneva, Switzerland

<sup>4</sup>INRAE, CNRGV French Plant Genomic Resource Center, F-31320, Castanet Tolosan, France

<sup>5</sup>Department of Electronics, Telecommunications and Informatics and Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

<sup>6</sup>Department of Virology, University of Helsinki, Haartmaninkatu 3, 00014 Helsinki, Finland

<sup>7</sup>Biotechnology Center, National Chung Hsing University, 145 Xingda Road, Taichung 40227, Taiwan

\*Correspondence address. Weihong Qi. Functional Genomics Center Zurich, ETH Zurich and University of Zurich, Winterthurerstrasse 190, 8057, Zurich, Switzerland. E-mail: [weihong.qi@fgcz.ethz.ch](mailto:weihong.qi@fgcz.ethz.ch); Wilhelm Gruissem. Department of Biology, Institute of Molecular Plant Biology, ETH Zurich, Universitätstrasse 2, 8092, Zurich, Switzerland. E-mail: [wilhelm\\_gruissem@ethz.ch](mailto:wilhelm_gruissem@ethz.ch)

†Equal contributions.

## Abstract

**Background:** Cassava (*Manihot esculenta*) is an important clonally propagated food crop in tropical and subtropical regions worldwide. Genetic gain by molecular breeding has been limited, partially because cassava is a highly heterozygous crop with a repetitive and difficult-to-assemble genome.

**Findings:** Here we demonstrate that Pacific Biosciences high-fidelity (HiFi) sequencing reads, in combination with the assembler hifiasm, produced genome assemblies at near complete haplotype resolution with higher continuity and accuracy compared to conventional long sequencing reads. We present 2 chromosome-scale haploid genomes phased with Hi-C technology for the diploid African cassava variety TME204. With consensus accuracy >QV46, contig N50 >18 Mb, BUSCO completeness of 99%, and 35k phased gene loci, it is the most accurate, continuous, complete, and haplotype-resolved cassava genome assembly so far. *Ab initio* gene prediction with RNA-seq data and Iso-Seq transcripts identified abundant novel gene loci, with enriched functionality related to chromatin organization, meristem development, and cell responses. During tissue development, differentially expressed transcripts of different haplotype origins were enriched for different functionality. In each tissue, 20–30% of transcripts showed allele-specific expression (ASE) differences. ASE bias was often tissue specific and inconsistent across different tissues. Direction-shifting was observed in <2% of the ASE transcripts. Despite high gene synteny, the HiFi genome assembly revealed extensive chromosome rearrangements and abundant intra-genomic and inter-genomic divergent sequences, with large structural variations mostly related to LTR retrotransposons. We use the reference-quality assemblies to build a cassava pan-genome and demonstrate its importance in representing the genetic diversity of cassava for downstream reference-guided omics analysis and breeding.

**Conclusions:** The phased and annotated chromosome pairs allow a systematic view of the heterozygous diploid genome organization in cassava with improved accuracy, completeness, and haplotype resolution. They will be a valuable resource for cassava breeding and research. Our study may also provide insights into developing cost-effective and efficient strategies for resolving complex genomes with high resolution, accuracy, and continuity.

**Keywords:** phased chromosome pairs, haplotype heterozygosity, pan-genome, allele-specific expression

## Background

High-quality reference genomes are fundamental for genomic analyses, which have revolutionized the fields of biology and medicine. Most plant genomes are challenging to assemble with a high level of accuracy, continuity, and completeness because they vary in size, levels of ploidy, and heterozygosity [1]. Particularly, many plant species, including cassava, can be clonally propagated, which can increase the effective number of alleles and

heterozygosity [2–4]. Meanwhile, plant genomes are highly repetitive and contain abundant ancient and novel transposable elements [1, 5]. Intra-genomic heterozygosity and repeat elements are major sources of genome assembly errors [5, 6]. The cassava (*Manihot esculenta*, NCBI:txid3983) genome has a haploid genome size ~750 Mb [7–9] and is one of the most heterozygous [10] and repetitive [8] of currently sequenced plant genomes [11]. Despite continuous sequencing efforts using different technologies over

Received: October 16, 2021. Revised: January 11, 2022. Accepted: February 22, 2022

© The Author(s) 2022. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

the past decade, unresolved gaps and haplotypes persist in all chromosomes of currently available cassava genomes [7–10, 12].

Cassava is an important staple crop that is clonally propagated in tropical and subtropical regions worldwide. The starchy storage roots are an important staple food for nearly 1 billion people and used for industrial purposes. In Africa, cassava is cultivated mainly by smallholder farmers because the crop produces appreciable yields under a wide array of environmental conditions. However, production is constrained by weeds, drought, pests, and most, crucially, viral diseases. Therefore, breeding of more robust and productive cassava varieties is of high importance. Because conventional breeding of cassava is time-consuming, complete haplotype-resolved reference genomes with high accuracy will be a valuable resource for applications of genomic selection, genome editing, and improving genetic gains in cassava breeding.

Continuous long reads (CLRs) produced by Pacific Biosciences (PacBio) Single Molecule, Real-Time (SMRT) sequencing technology and other long-read sequencing technologies have been essential for generating reference-quality genome assemblies cost effectively in the past decade [13]. The African cassava cultivars TME3 and 60444 have been sequenced and assembled using 70-fold PacBio CLRs (read N50 12 kb), producing genome assemblies with contig N50 of 98 and 117 kb, respectively [8], which were lower than the continuity metric of a high-quality genome proposed by the Vertebrate Genome Project (VGP) consortium (contig N50 > 1 Mb) [14]. More than 18,000 gene loci in TME3 and 60444 were resolved with 2 haplotype alleles [8], but collapsed regions still persist throughout both assemblies, owing to the fact that assembly of error-prone long sequencing reads (hereafter referred to as long reads) homogenized sequences from different haplotype alleles, paralogous loci, and repeat elements [15]. The recently introduced PacBio high-fidelity (HiFi) sequencing technology is able to produce long (10–25 kb) and highly accurate (>99.9%) sequencing reads (hereafter referred to as HiFi reads). For several human and animal genomes, equivalent or higher continuities have been achieved with HiFi reads [15–18]. Novel genome assemblers have been developed to leverage the full potential of HiFi reads [15, 19], where the combined performance of HiFi reads and HiFi-specific genome assemblers was benchmarked in assembling human and animal genomes. Their potential in assembling plant genomes is less well studied but is gaining momentum [19, 20]. In comparison to the strawberry reference genome reconstructed from a combination of short Illumina sequencing reads and PacBio CLR [21], the HiFi assembly of *Fragaria* × *ananassa* has contig N50 values that are 10 times higher. HiFi reads also enabled the assembly of the 35.6-Gb California redwood genome [19]. The recently published haplotype-resolved potato genome [20] was generated using a combination of multiple sequencing strategies, including HiFi reads.

## Data Description

In this study, we collected PacBio CLRs (ERR5487554–ERR5487559), HiFi reads (ERR5485301), Illumina paired-end (PE) sequencing reads (hereafter referred to as Illumina PE reads) (ERR5484652), and Hi-C data (ERR5484651) for the African cassava cultivar TME204. It belongs to a group of cassava cultivars carrying the dominant monogenic CMD2 resistance locus, which provides resistance to cassava mosaic diseases (CMD) caused by African cassava mosaic viruses [22]. We benchmarked the performance of CLRs and HiFi reads in assembling this highly complex and heterozygous genome. Assembly continuity, accuracy, and haplotype resolution of different genome drafts produced by 4 CLR/HiFi as-

semblers [15, 23, 24] were evaluated using genome quality metrics proposed by the VGP consortium [25] with Illumina PE reads from the same sample. Our results demonstrate that HiFi reads are valuable in assembling a high-quality heterozygous and repetitive plant genome. The high base accuracy and long sequencing read length provide superior resolution and accuracy in resolving allele differences between haplotypes, paralogous genes, and repeat elements. By combining HiFi reads with Hi-C data we produced a highly accurate, chromosome-scale, phased assembly for a diploid African cassava cultivar. The 2 haploid assemblies (PRJNA758616 and PRJNA758615) revealed extensive haplotype heterozygosity within a cassava diploid genome and provided a systematic view of the cassava diploid genome organization with improved accuracy, completeness, and haplotype resolution. To improve genome annotation, we further generated PacBio Iso-Seq reads (ERR5489420–ERR5489422) from different tissues. In combination with public TME204 RNA-seq data from 9 tissues, *ab initio* gene prediction with experimental evidence identified 20k chromosomal novel gene loci, with enriched functionality related to chromatin organization. The close-to-complete haplotype-resolved, annotated genome also enabled pan-genome and allele-specific expression analysis, demonstrating the importance of a more complete representation of cassava genetic diversity for downstream reference-guided omics analysis and molecular breeding.

## Analyses

### Cassava TME204 genome characteristics

Illumina PE reads (Table 1) were used to estimate the overall genome characteristics of TME204, revealing a highly heterozygous diploid genome different from the reference genome of the partially in-bred South American cassava cultivar AM560 [7] and other well-studied genomes, such as the human reference genome. The peak for *k*-mers covering TME204 heterozygous sequence was as high as the peak corresponding to *k*-mers present in both haplotypes, while the *k*-mer coverage plots for the cassava and human reference genomes were dominated by their homozygous sequence peaks (Supplementary Fig. S1a). On the basis of the number of variant-induced branches in the de Bruijn assembly graph [26], the level of heterozygosity in the TME204 genome was measured at 1%, which is 1 magnitude higher than the heterozygosity level in the cassava and human reference genomes (Supplementary Fig. S1b). This value is a conservative estimate because it is based on genomic regions only with lower rates of nonstructural variations. Highly heterozygous regions introduce divergent paths with higher complexity, which cannot be resolved by conventional bubble calling algorithms used to calculate variant-induced branching rate [27]. Consequently, sequences with a high density of single-nucleotide polymorphisms (SNPs), small insertions and deletions (indels ≤50 bp), and large structural variations (SVs, e.g., indels >50 bp, duplications, inversions, and translocations) were not counted in the 1% of heterozygosity. The cassava genomes (TME204 and AM560) are more repetitive than the human reference genome (Supplementary Fig. S1c), which makes them more difficult to assemble with high quality.

### Benchmarking cassava TME204 assemblies from PacBio CLR and HiFi reads

PacBio HiFi sequencing yielded 42× HiFi reads with length N50 of 20 kb (Table 1). To assess the performance of different assemblers, the HiFi reads were assembled using 4 HiFi-specific software tools:

**Table 1:** Cassava TME204 shotgun sequencing data

Parameter	PacBio CLR	PacBio HiFi reads	Illumina PE reads
Sequencer	PacBio Sequel	PacBio Sequel II	Illumina NovaSeq
Chemistry	Sequel binding kit 3.0	Sequel II binding kit 2.0	TruSeq DNA Nano
No. of SMRT cells	6 (1M v3 cells)	1 (8M cell)	NA
No. of reads	5,037,588	1,531,543	259,505,436
No. of bases (bp)	90,586,242,030	31,312,160,541	77,851,630,800
Read length N50 (bp)	29,274	20,363	2 × 150
Estimated coverage (×) <sup>a</sup>	121	42	104
Accession Nos.	ERR5487554-ERR5487559	ERR5485301	ERR5484652

<sup>a</sup>Based on a haploid genome size of 750 Mb. NA: not applicable.

Falcon, HiCanu, hifiasm, and IPA (see Methods). For comparison of the HiFi reads with traditional long reads, we also assembled 121× PacBio CLR (Table 1) from the same DNA sample using the Falcon assembler.

With the same amount of computing resources, HiFi read assembling was ~2 orders of magnitude faster and required 10 times less data storage than CLR assembling. Each HiFi assembly was completed in only a few hours to a few days and used 20–800 GB of data storage when running on a single server with 64 CPUs and 500 GB RAM. CLR-Falcon assembly took a few weeks and used ~7 TB of disk space.

Assembled genome sizes varied on the basis of the assembly software (Fig. 1a, Supplementary Tables S1 and S2, Supplementary Results). HiFi reads improved contig continuity, doubling contig N50 and NG50 values when assembled using Falcon (Fig. 1b). According to the more comparable NG50 values (Supplementary Results), the hifiasm contig set was the most continuous (NG50 33 Mb), followed by the HiCanu contig set (23 Mb) (Fig. 1b). HiFi reads also improved accuracy and completeness of the assembled genome sequences (Supplementary Table S3). When measured using alignments of Illumina PE reads from the same sample (Table 1), both hifiasm and HiCanu achieved superior base accuracy (0.2% error rate) (Fig. 1c), structural accuracy (99.3% mapped reads were correctly paired) (Fig. 1d), and assembly completeness (99.9% mapped reads) (Fig. 1e). When measured using the Merqury *k*-mers comparisons of Illumina reads and assembled contigs [28], the hifiasm assembly was most accurate (99.997%, quality value [QV] 46.74) (Fig. 1f, Supplementary Results) and complete (98.40%) (Fig. 1e, Supplementary Results). Because a proper assembly of PacBio CLR reads requires signal level polishing to achieve satisfactory consensus accuracy [29], we further phased and polished the CLR-Falcon assembly using Falcon-Unzip [23]. The final CLR-Falcon-Unzip assembly had a QV score of 38.86, with *k*-mer completeness of 97.64%. Both measurements were still lower than those achieved by HiCanu and hifiasm with HiFi reads, while using much less computing time and resources.

In addition to the variable and larger than expected total assembled genome size of ≥1.2 Gb (Fig. 1a, Supplementary Tables S1 and S2), the BUSCO duplication rates varied but remained high (Supplementary Tables S1 and S2), underlining the difficulty of assembling the highly heterozygous and repetitive cassava genome [8–10]. For the HiCanu and hifiasm TME204 assemblies, the total sizes were both about twice the expected haploid genome size, and the BUSCO duplicate scores were >80% (Supplementary Table S2). Falcon (HiFi and CLR) and IPA produced smaller assemblies,

**Table 2:** Assembly quality assessment of Cassava TME204 haplotigs generated by hifiasm v0.15.3

Quality Category	Quality Metric	Haplotype 1	Haplotype 2
General	Contig Size (Mb)	762	706
	Contig N50 (Mb)	18	26
	Contig NG50 (Mb)	18	22
	Largest contig (Mb)	41	44
Base accuracy	QV <sup>a</sup>	45.23	48.94
	<i>k</i> -mer completeness (%) <sup>a</sup>	79.6	79.1
Structural accuracy	<i>k</i> -mer false duplications (%) <sup>a</sup>	3.0	2.4
	BUSCO duplicate (%)	4.9	4.2
	Reference asmgene duplicate (%)	1.6	0.8
	Reference gene transfer duplicate (%) <sup>b</sup>	3.8	3.2
	Reliable blocks (%)	97.2	97.5
	Congruent genetic markers (%)	99.8	99.9
	BUSCO complete (%)	99.0	98.8
Functional completeness	Reference asmgene complete (%)	95.2	95.4
	Reference gene transfer rate (%) <sup>b</sup>	96.3	96.5
	Transcript alignment rate (%)	99.3	99.4

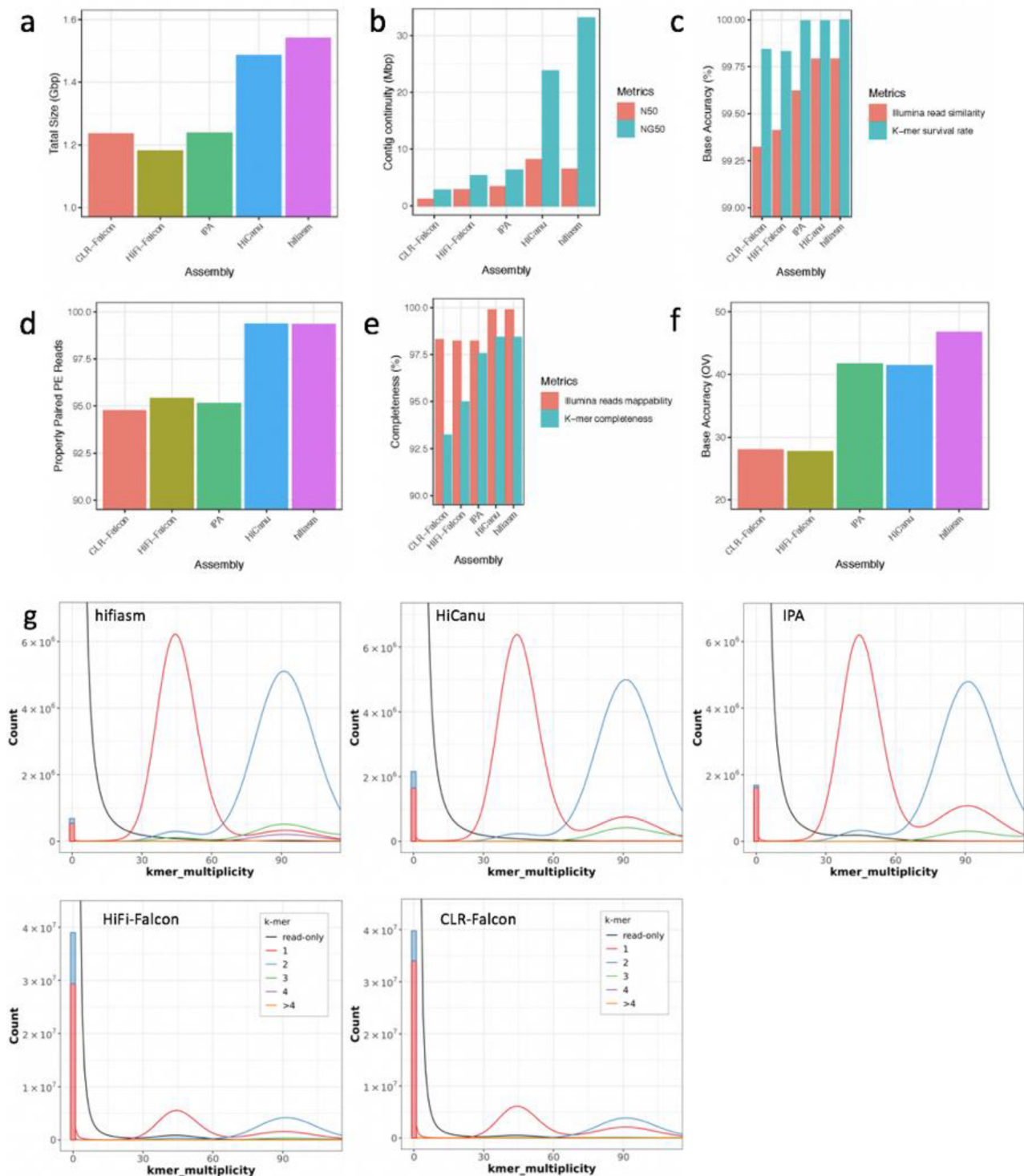
<sup>a</sup>QV and *k*-mer completeness for the combined assembly is 46.63 and 98.4%, respectively. The *k*-mer false duplication rate for the combined assembly is 3.5%.

<sup>b</sup>Calculated on the basis of the number of lifted genes regardless of completeness.

with lower BUSCO duplicate scores. Merqury *k*-mer analysis [28] confirmed that the different long reads and assemblers varied in their performance of resolving haplotypes in the TME204 genome, with HiFi reads in combination with hifiasm producing the most haplotype-resolved assembly, and CLR-Falcon the least (Fig. 1g).

### Phased, haplotype-resolved contigs of cassava TME204

Subsequently we used a newer release of the hifiasm assembler (v0.15.2) to assemble the haplotype-resolved TME204 contigs, which were also phased at the same time using Hi-C technology (Table 2). The resulting 2 sets of haplotigs (phased haplotype-resolved contigs) represent haplotype 1 (762 Mb) and haplotype 2 (706 Mb) of the diploid cassava genome, hereafter referred to as H1 and H2, respectively. The assembly has a QV score of 45.23 for H1 haplotigs, 48.94 for H2 haplotigs, and 46.63 for the combined set of sequences. For each haplotype and the combined diploid assembly, the *k*-mer completeness is 79.6, 79.1, and 98.4%, respectively, indicating that ~19% of the *k*-mers were haplotype-specific.



**Figure 1:** Benchmarking analysis of cassava TME204 assemblies from PacBio CLR and HiFi reads. (a) Assembly size of all resolved alleles. (b) Contig continuity measured as N50 and NG50. N50 is the length of the shortest contig in the set of largest contigs that make up 50% of the assembly size as shown in (a). NG50 is the length of the shortest contig in the set of largest contigs that make up 50% of the haploid genome size of 750 Mb. (c) Base accuracy of contigs, measured by sequence similarity between contigs and mapped Illumina reads, and as the fraction of  $k$ -mers found in both contigs and the Illumina reads. (d) Structure accuracy of contigs, measured by the percentage of properly paired Illumina PE reads. (e) Assembly completeness, measured by the percentage of mapped Illumina reads and the fraction of reliable Illumina  $k$ -mers retained in the contigs. (f) Phred scale quality value (QV) of contigs, calculated using the error probability  $P$  with the formula:  $QV = -10 * \log(P, 10)$ , where  $P$  is the fraction of  $k$ -mers found in the contigs but missing in the Illumina reads. (g) Completeness of resolved haplotypes measured by Merqury copy number spectrum plots. The x-axis shows  $k$ -mer multiplicity computed from the Illumina reads. The y-axis shows the abundance for  $k$ -mers with a given multiplicity, either in the Illumina reads (black) or in the contigs (colored by the number of times they are found in the underlying assembly). Red peaks at 45x represent resolved haplotype alleles, red peaks at 90x collapsed haplotype alleles. Black humps found at either 40x (heterozygotes/1-copy  $k$ -mers) or 80x (homozygotes/2-copy  $k$ -mers) represent reliable Illumina  $k$ -mers missing in contigs, corresponding to the assembly completeness in (e). Assembly-specific  $k$ -mers absent from the Illumina reads are plotted as a bar at zero  $k$ -mer multiplicity, corresponding to the error probability in (f).

Most importantly, the majority of haplotype-specific *k*-mers are present only once in the assembled sequences, while the majority of homozygous *k*-mers shared by 2 haploid genomes are present twice (Fig. 2). This would be expected for a completely haplotype-resolved genome assembly in which even homozygous segments of the genome are included in both haplotypes. Only 3% of *k*-mers were from artificial duplications (Fig. 2), which was similar to the false duplication rates measured at ~1% (reference asmgene score) to 4% (BUSCO duplication score). Functional completeness measured using plant BUSCO orthologs and TME204 Iso-Seq transcripts was  $\geq 98\%$ . The reference asmgene completeness score was slightly lower (95%), which could be due to the high level of sequence differences between AM560 and TME204 (see later comparative analysis), therefore fewer AM560 reference genes could be aligned to TME204, resulting in lower asmgene completeness and duplication scores.

To assess the structural accuracy of the assembled TME204 haplotigs, we mapped the longer PacBio CLR sequences from the same DNA sample (Table 1) to the haplotigs and analyzed the CLR read coverage along each haplotig (Supplementary Fig. S2). We defined reliably assembled sequences as those with  $\geq 10\times$  CLR read coverage. More than 97% of the assembled bases could be classified as correctly assembled with this quality metric.

To further validate the base, structural, and phasing accuracy of TME204 haplotigs, we generated complete sequences (96–128 kb) of bacterial artificial chromosomes (BACs) containing TME204 genome fragments and aligned them to both sets of haplotigs (Fig. 3, Supplementary Fig. S2). When a region is properly assembled and phased, we expect 1 continuous BAC-to-haplotig alignment for the corresponding BAC (resolved BAC). Three of the 4 sequenced BACs were resolved in H1 and 1 was resolved in H2, either perfectly or with only 1 indel difference (Supplementary Table S4), confirming the close to Q50 consensus accuracy (i.e., 1 error per 100 kb consensus sequences). The striking differences of BAC-to-haplotig alignments between the 2 haplotypes highlight the high level of haplotype differences in these regions.

The haplotigs were also compared with the cassava high-density genetic map [30]. Among the 22,403 available genetic markers, ~14,000 could be uniquely aligned to each set of the TME204 haplotigs perfectly (i.e., with full length coverage and 100% sequence identity). More than 99.8% of these unique and perfect genetic markers showed high congruence between the genetic map and assembled haplotigs (Supplementary Fig. S3). Only <0.2% of the genetic markers were found among markers with different chromosome origins. Plots of genetic versus physical distance identified 3 pairs of chromosome-scale haplotigs (chromosomes VIII, XII, and XIV) and 6 other chromosome-scale haplotigs in either H1 or H2 (Table 3). In plots of genetic versus physical distance for these chromosome-scale haplotigs (Supplementary Fig. S3), we often observed steep slopes at the haplotig ends and flat regions in their centers, which is consistent with increased recombination in chromosome arms and reduced recombination in pericentric regions of the chromosomes. Collectively, the data suggest that all of the 18 cassava chromosome pairs are already highly continuous at the haplotig level, where each chromosome is composed of only 1 to a few haplotigs.

### Pseudochromosome pairs of cassava TME204

To further scaffold haplotigs into pseudochromosomes, we first used Hi-C scaffolding, but this did not further scaffold any haplotigs in H1 (Supplementary File S1). In H2, Hi-C data produced 7 chromosomal scaffolds that were perfectly congruent with the

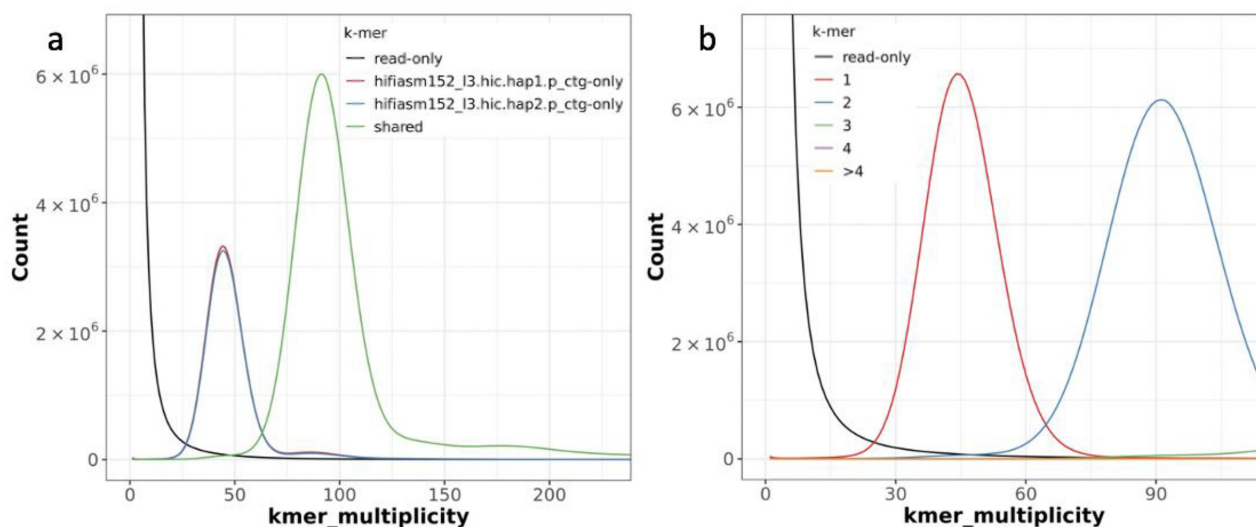
genetic map, but also misjoined haplotigs from different chromosomes (Supplementary Table S5, Methods). Together, the high congruence between haplotigs and the genetic map allowed us to reconstruct all 18 pairs of pseudochromosomes with high confidence (Fig. 4, Supplementary File S2). TME204 H1 and H2 pseudochromosomes are composed of 43 and 39 haplotigs, respectively. In total 12 pseudochromosomes are chromosome scale (Table 3). Haplotig orientations could be determined (Supplementary File S3) except for 2 small haplotigs in H2, representing the first 0.4 Mb of chromosome VII and the last 1.3 Mb of chromosome XI (Supplementary File S4). Together, 86.8 and 91.1% of the haplotig sequences could be assigned to chromosomes for H1 and H2, respectively (Table 2, Supplementary Files S3 and S4).

In both TME204 H1 and H2 assemblies, we found haplotigs that could not be scaffolded using either the genetic map (Table 3) or Hi-C technology (Supplementary Files S1 and S2). A majority of these unanchored haplotigs can be partially aligned to the pseudochromosomes with an average sequence similarity of 98% (Table 3). A few hundred AM560 genes can be transferred onto these haplotigs as well, although most (70%) were duplicated copies of genes that already transferred onto pseudochromosomes. It is clear that these haplotigs are of cassava origin and not from foreign contamination. When the assembled sequences were screened against the NCBI mitochondrial database, unanchored haplotigs representing the highly fragmented mitochondrial genome were identified in both haplotype assemblies (Table 3, Supplementary Fig. S4a). When compared to the other non-mitochondrial unanchored haplotigs, mitochondrial haplotigs have a smaller size variation (25–76 kb) and lower depth of coverage on average (Supplementary Fig. S4b). Regions similar to nuclear mitochondrial pseudogene regions (numt's) were also ubiquitous and found in both pseudochromosomes (Supplementary Fig. S4c) and unanchored haplotigs (Supplementary Table S6). Some of the non-mitochondrial unanchored haplotigs can be regions still missing from the present set of pseudochromosome pairs where the gene content completeness ranges from 91 to 98% (Table 3). They can also be results of assembly artifacts (i.e., collapsed repeats) or represent novel haplotypes from *de novo* mutations.

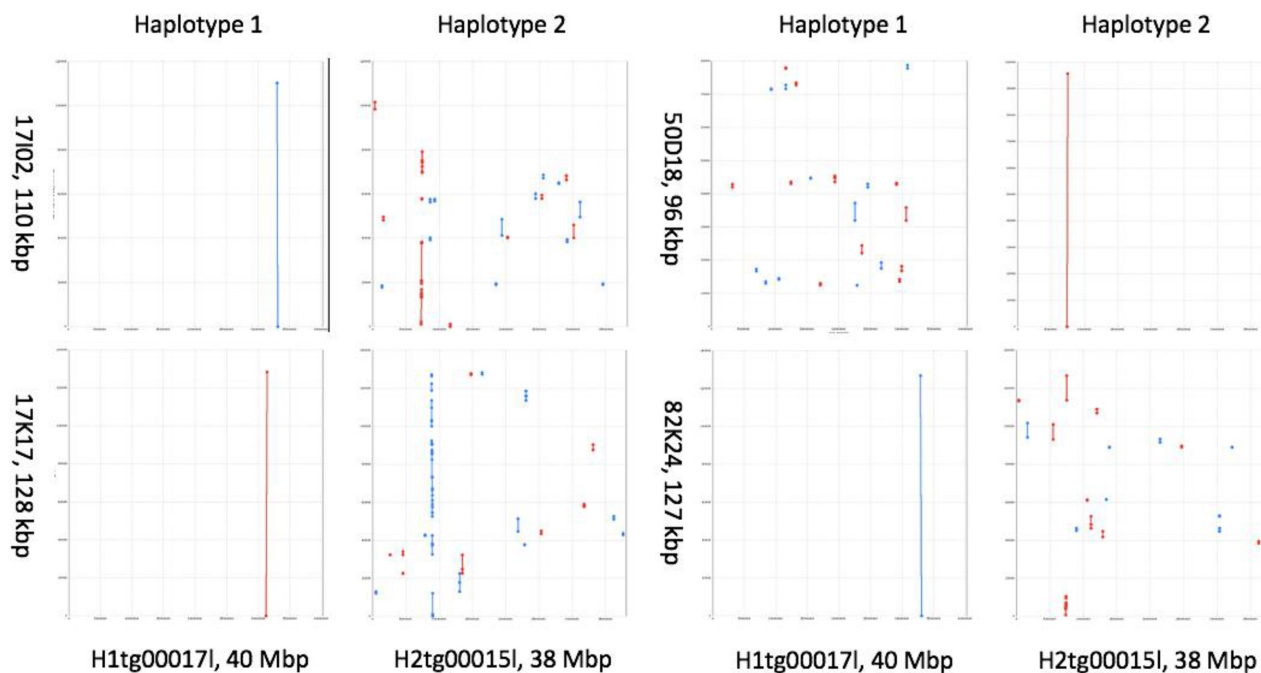
### Repeat and gene landscape of cassava TME204 genome

*De novo* repeat modeling using all resolved allelic sequences identified 1,431 repeat families, with 1,016 families representing novel unclassified repeats, which make up 20% of the TME204 genome (Supplementary Fig. S5). The distributions of family sizes and sequence lengths among the novel repeat families are similar to those in LTR families (Supplementary Fig. S5b), which make up 38% of TME204 genome. In total, >60% of each TME204 haploid genome can be masked as repeats, without counting small RNA and low-complexity sequences (Supplementary Fig. S5).

During the past 10 years, continuous efforts have been made to improve the assembly and annotation of the cassava reference genome AM560 [7, 12, 30]. The set of AM560 reference gene models [31] is widely used in the research field. We therefore first annotated the TME204 genome (annotation release v1.0) by transferring well-established cassava reference gene models to TME204 H1 and H2 assemblies (Table 3). A total of 96–97% of the 32,805 AM560 gene loci could be lifted completely to TME204 H1 and H2 assemblies with a duplication rate of 3–4%, which is similar to BUSCO complete and duplicate scores (Table 2). Comparison of orthologous gene pairs revealed high gene synteny (99%) between



**Figure 2:** Merqury assembly and copy number spectrum plots of cassava TME204 haplotigs. (a) For the Merqury assembly plot, k-mers are colored by their uniqueness in the Illumina PE reads (black), haplotype 1 (red) and haplotype 2 (blue) assemblies. Shared k-mers are shown in green. At the heterozygotes peak (45 $\times$ ), the second haplotype has only slightly fewer k-mers (blue) compared to the first haplotype (red), indicating that the reconstruction of heterozygous variants was almost complete. Red hump and blue shoulder  $\sim 90\times$  are haplotype-specific k-mers that are actually from homozygotes sequences, green shoulder  $\sim 45\times$  is due to shared k-mers belonging to heterozygotes. These shoulders are all very small, suggesting a very low level of collapsed homozygous regions and artificial duplications. (b) In the copy number spectrum plot, the majority of heterozygous k-mers appear once (red peak at 45 $\times$ ) and the majority of homozygous k-mers twice (blue peak at 90 $\times$ ), confirming that the assembly is close to complete haplotype-resolved and even the homozygous part of the genome is included in both haplotypes. High k-mer completeness is supported by the lack of black humps at 45 $\times$  or 90 $\times$ . Low artificial duplication is revealed by the barely detectable humps (green, purple, orange) of duplicated k-mers. The bars at zero k-mer multiplicity are low in both plots, suggesting that most k-mers in the assemblies are also present in Illumina reads and therefore the assembled sequences are of high consensus accuracy.



**Figure 3:** Phasing accuracy of cassava TME204 haplotigs validated by BAC-to-haplotig alignments. Each dot plot shows the alignment of 1 BAC (y-axis) with 1 haplotig (x-axis). Forward alignments are plotted as red lines/dots, reverse alignments in blue. A line represents an undisturbed segment of alignment. When a region is correctly assembled and phased, the corresponding BAC sequence will align continuously (a resolved BAC). Three of the 4 BACs were resolved in the TME204 H1 assembly, and 1 (50D18) was resolved in the TME204 H2 assembly. For each BAC, the striking differences of BAC-to-haplotig alignments between haplotypes reveal the high level of haplotype differences in these regions.

**Table 3:** Phased chromosome pairs in TME204 diploid genome assembly

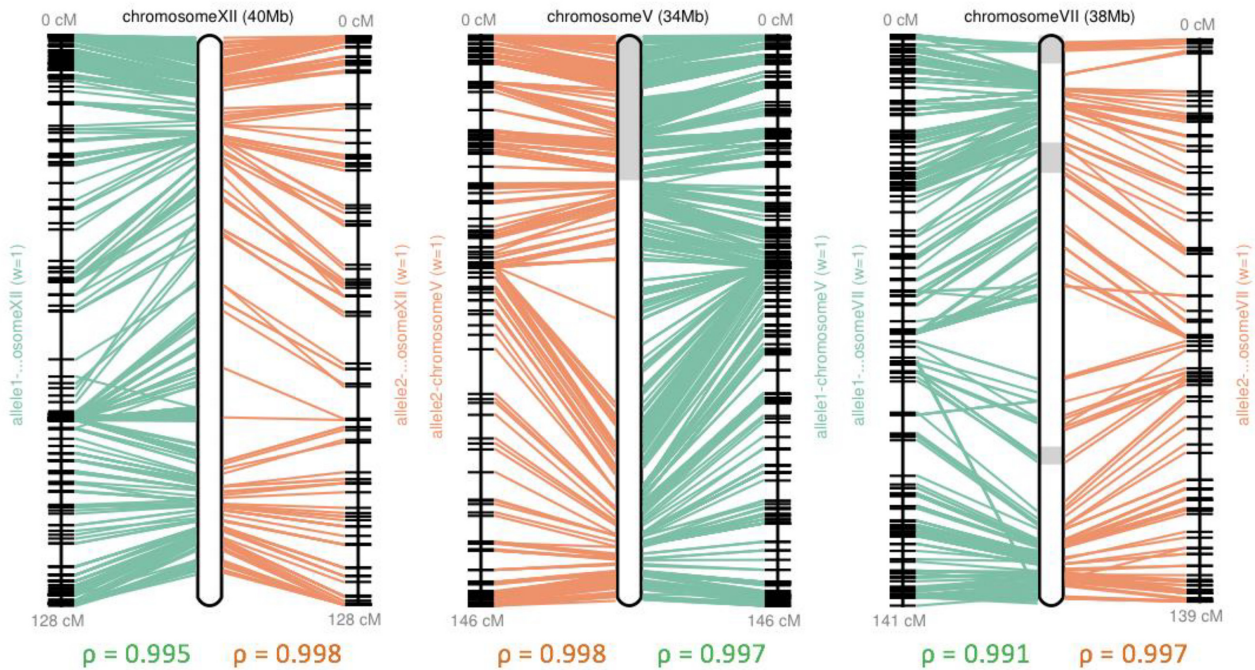
Chromosome	Haplotype 1			Haplotype 2		
	Length (Mb)	Haplotigs (n)	Reference gene transfer rate (%) <sup>a</sup>	Length (Mb)	Haplotigs (n)	Reference gene transfer rate (%) <sup>a</sup>
I	44.60	5	97.8	43.73	1	98.1
II	39.76	2	96.2	40.67	2	97.9
III	34.10	2	95.5	33.73	2	96.1
IV	35.07	1	97.2	35.26	3	93.6
V	33.99	1	97.3	33.40	2	97.7
VI	32.06	2	95.0	32.36	3	96.0
VII	37.89	6	92.5	36.62	5	91.5
VIII	40.94	1	97.0	42.44	1	97.1
IX	39.49	3	95.5	36.52	5	94.4
X	33.53	3	93.3	31.79	1	94.1
XI	34.31	2	94.9	33.74	2	95.4
XII	40.28	1	96.6	38.12	1	96.4
XIII	39.96	2	94.9	38.46	1	95.6
XIV	31.29	1	97.1	29.54	1	96.4
XV	35.50	2	97.8	34.23	2	97.9
XVI	34.02	1	95.6	34.23	2	95.9
XVII	37.53	3	93.1	33.98	2	92.1
XVIII	37.65	3	91.4	34.55	3	93.5
Total No. of haplotigs		1,439		770		
Total length of haplotigs (bp)		762,392,783		706,328,643		
Total No. of anchored haplotigs		43		39		
Total length of pseudochromosomes (bp)		661,977,943		643,362,786		
No. of unanchored haplotigs		1,396		731		
No. (%) of unanchored haplotigs aligned to pseudochromosomes		1,154 (82.7)		688 (94.1)		
Length of unanchored haplotigs (bp)		100,414,840		62,965,857		
Length (%) of unanchored haplotigs aligned to pseudochromosomes		63,212,546 (63.0)		45,142,471 (71.7)		
Annotated genes in unanchored haplotigs (% duplicate)		374 (80.6)		443 (69.1)		
No. of unanchored, mitochondrial haplotigs		281		53		
Length of unanchored, mitochondrial haplotigs		12,170,354		2,557,482		

<sup>a</sup>Percentage of AM560 genes that were lifted to the corresponding chromosome in each TME204 haplotype assembly, regardless of copy number and completeness

pseudochromosomes of AM560 and the TME204 H1/H2 assembly. Nine inverted regions involving 109 genes distributing among pseudochromosomes III, VI, VII, VIII, X, and XVIII were found between AM560 and the TME204 H1 assembly. There are 7 inversions between the AM560 and TME204 H2 assemblies involving 203 genes on chromosomes VI, VII, X, XI, and XVIII (Fig. 5a).

To complement the reference gene models, we also predicted genes (annotation release v1.1) in TME204 H1 and H2 assemblies using the AUGUSTUS software tool, with experimental evidence from TME204 Iso-Seq transcripts (Supplementary Table S7) and RNA-seq data [32] (Supplementary Table S8). In contrast to the 53k protein-coding transcripts transferred from reference annotation, *ab initio* gene prediction with extrinsic evidence identified 93k and 84k protein-coding transcripts in the TME204 H1 and H2 assemblies, respectively. BUSCO scores for the predicted proteomes and assembled genomes were similar, suggesting that the functional completeness of the annotated proteomes well represents the underlying genomes (Supplementary Fig. S6). Figure 5b shows that 93% of the previously transferred gene loci overlapped with *ab initio* predicted gene models, indicating a good recall rate of the

*ab initio* gene prediction process. Among the >30k predicted novel gene loci per assembly, which do not overlap with any transferred reference genes, 27k are distributed uniformly across the 18 pseudo-chromosomes (Fig. 5c). The other 21k and 13k arise from unanchored haplotigs in the TME204 H1 and H2 assemblies, respectively (Fig. 5d). Close to 80% of the chromosomal predicted transcripts (both novel and known) could be functionally annotated using the InterPro protein database, while only 10% of the novel predicted transcripts from unanchored haplotigs showed significant matches to InterPro protein families (Fig. 5e). Novel predicted genes on chromosomes and unanchored haplotigs are functionally distinct. Chromosomal novel genes are enriched for GO biological process (BP) term “cellular components of DNA packaging complex,” including nucleosome, chromatin, chromosome, and DNA-protein complex (Fig. 5f). Novel genes from unanchored haplotigs are enriched for GO BP term “cellular components of chloroplast thylakoids” (Fig. 5g). Similarly, other GO BP terms were enriched as well (Supplementary Fig. S7a and b). Chromosomal novel genes are enriched for “chromatin organization,” “meristem development and maintenance,” and “cell response to stress



**Figure 4:** Reconstruction of pseudochromosomes in the cassava TME204 H1 assembly using the genetic map. For each pseudochromosome, the panel shows the physical positions on the reconstructed pseudochromosome and the map positions connecting by lines. Adjacent contigs within the reconstructed pseudochromosome are shown as boxes with alternating shades. The  $\rho$ -value under each map measures the Pearson correlation coefficient, with values in the range of  $-1$  to  $1$ , and values closer to  $-1$  and  $1$  indicate near-perfect collinearity. Chromosome XII is composed of a single chromosomal haplotig, the same as for chromosomes IV, VIII, XIV, and XVI. Chromosome V is composed of 2 contigs, the same as for chromosomes II, III, VI, IX, X, XI, XIII, XV, XVII, and XVIII. Chromosome VII is composed of 6 haplotigs, which is the most fragmented chromosome in the TME204 H1 assembly, followed by chromosome I, which has 4 haplotigs. Plots for all chromosomes in both H1 and H2 assemblies are provided in Supplementary File S5.

and stimuli.” Novel genes on unanchored haplotigs are enriched for “cytochrome complex assembly” and “related metabolism processes.”

### Tissue-specific differentially expressed transcripts

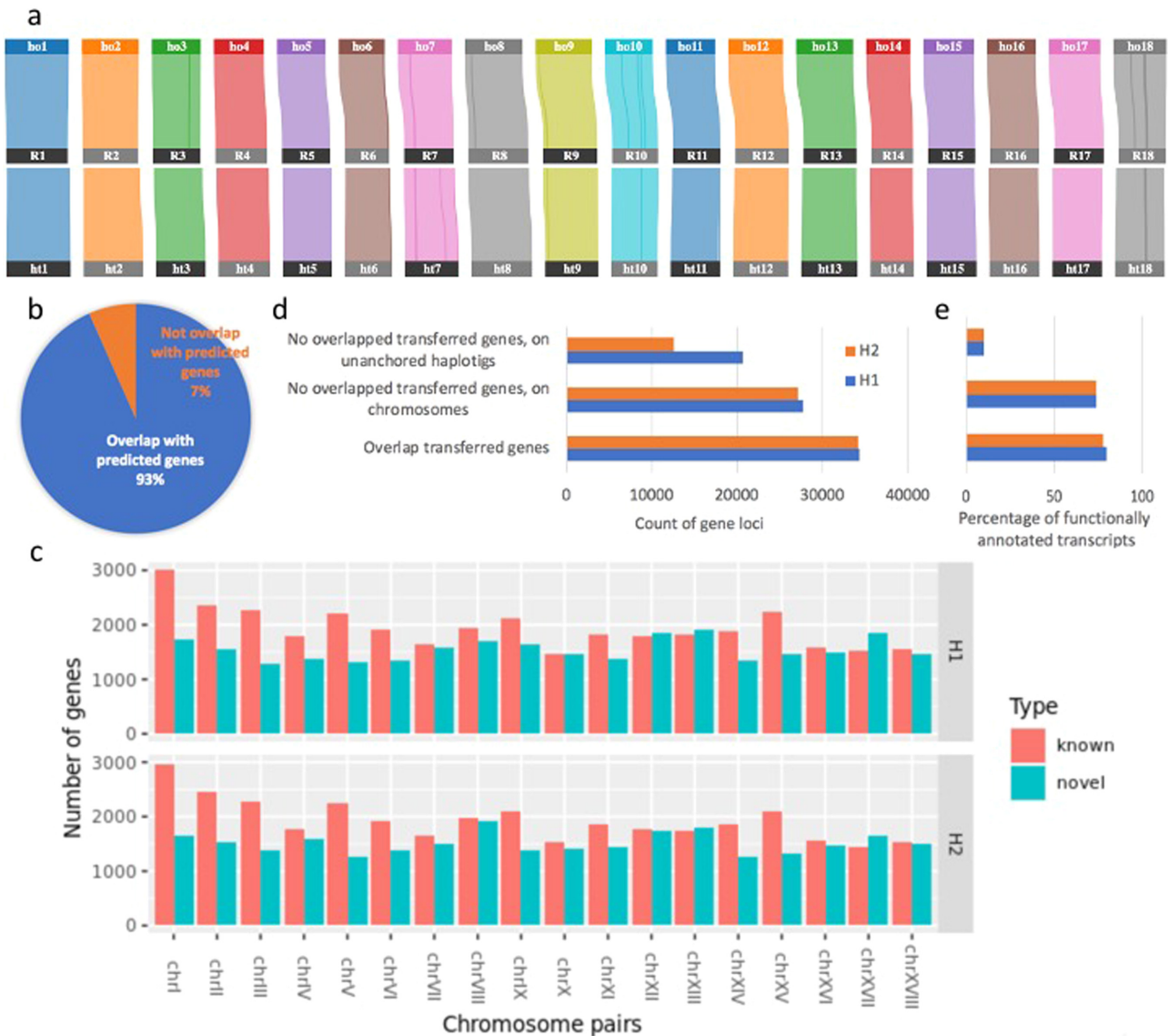
*Ab initio* genome annotation identified 94k and 84k transcripts in TME204 H1 and H2 assemblies, respectively. To construct the haplotype-resolved reference transcriptome, identical transcripts between haplotypes and within 1 haplotype were collapsed and only 1 copy of the sequences was kept, yielding 147,503 unique transcript sequences. More than 81% of predicted transcripts have different sequences between haplotypes (Fig. 6a). In such cases, analyzing RNA-seq data using 1 haploid set of genes/transcripts as the reference could potentially miss haplotype-specific, novel expression patterns. Therefore we reanalyzed previously published [32] RNA-seq data (Supplementary Table S8) generated from 9 different tissues of TME204. Based on the newly calculated transcript expression values, biological replicates cluster closely together in the principal component analysis analysis of Fig. 6b, which is similar to the previous analysis result based on the AM560 reference genome [32]. Across the 9 tissues, 60,839 transcripts (41%) were expressed (i.e.,  $\geq 2$  replicates per pairwise comparison have TPM value  $\geq 1$ ). Among all pairwise comparisons against stem, in total 9,437 (6%) transcripts showed significant difference in expression (a fold change [FC] of  $\log_2|FC| > 2$ , adjusted  $P < 0.00001$ ) (Fig. 6c). Clustering analysis of expression values of the differentially expressed transcripts (DETs) grouped the samples into 2 major distinct clusters: storage root, fibrous root, and root apical meristem (RAM) in 1 cluster, shoot apical meristem (SAM), lateral bud, leaf,

midvein, petiole, and stem in the other. The sample clustering is consistent with the previous result as well [32]. When comparing single-copy with multi-copy transcripts (i.e., same transcripts from duplicated gene loci within 1 haplotype), multi-copy transcripts have a relatively lower fraction of expressed transcripts and DETs (Fig. 6a). This is consistent with previous findings [33] that single-copy genes are generally more highly expressed than multi-copy genes. Transcripts from novel predicted gene loci on chromosomes are also less expressed and differentially regulated in different tissues (Fig. 6d). Classified by the haplotype origin of the transcripts, only 141 DETs are collapsed homozygous isoforms between haplotypes, 4,651 are from the H1 assembly, and 4,644 are H2-specific. Functional enrichment analysis of the 3 sets of DETs identifies a common GO BP term “developmental process,” which is expected because different tissues are compared here. DETs from H1 contribute to more diverse biological processes than H2 DETs (Fig. 6e), with the most significantly enriched BP terms being “transcription regulation, DNA templated,” and “photosynthesis, light harvesting,” which are also enriched in H2 DETs. The DETs between haplotypes do have different functions during tissue development. Our analysis also identifies 663 DETs from novel predicted gene loci on chromosomes. Functional enrichment analysis reveals that these novel DETs are mainly involved in biological processes of “response to stress, stimuli, and defense,” “DNA modification,” “methylation/demethylation,” and “protein phosphorylation” (Fig. 6f).

### Isoform allele-specific expression

For isoforms that are common to both haplotypes, we further investigated allele-specific expression (ASE) differences between

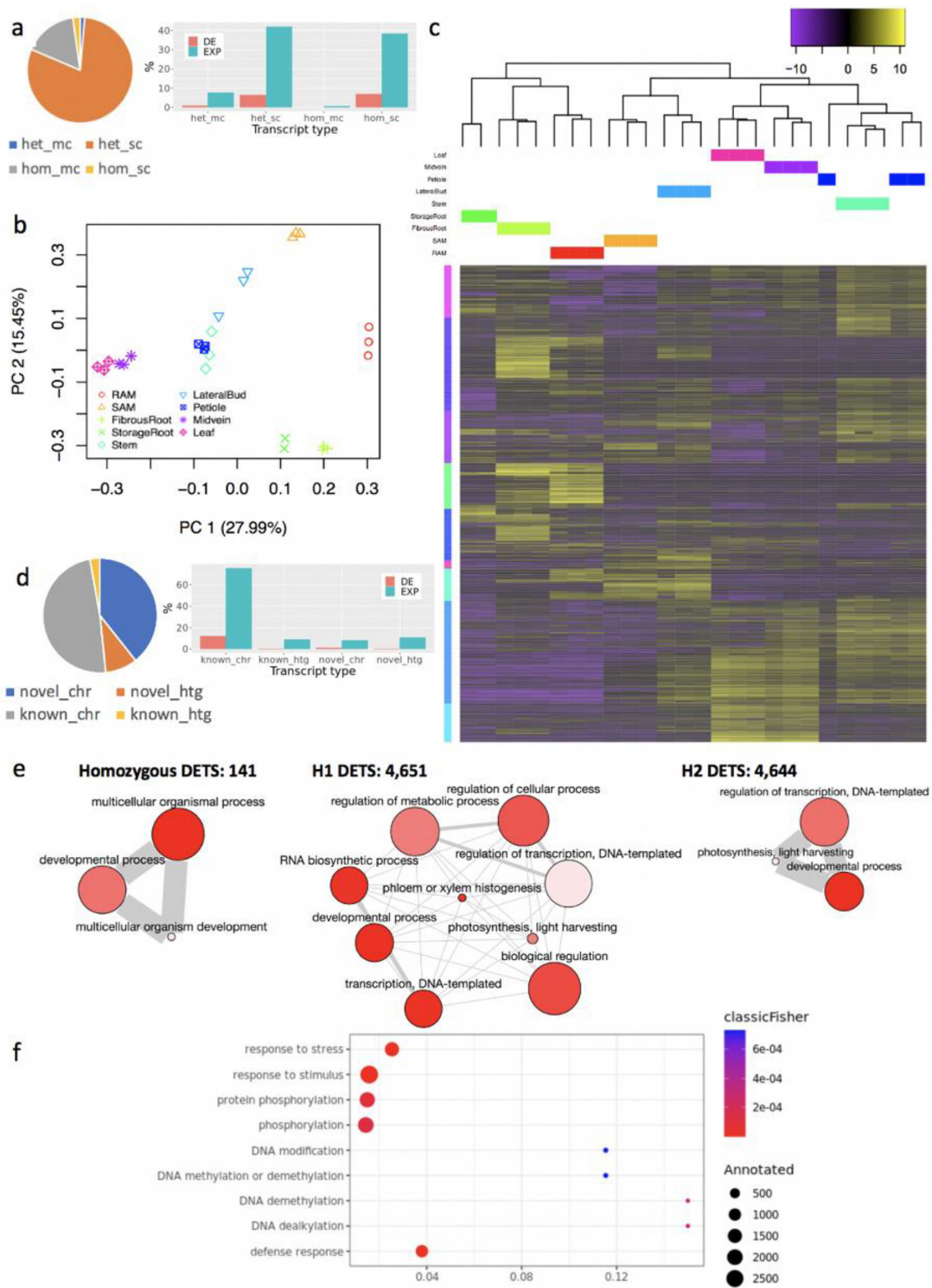




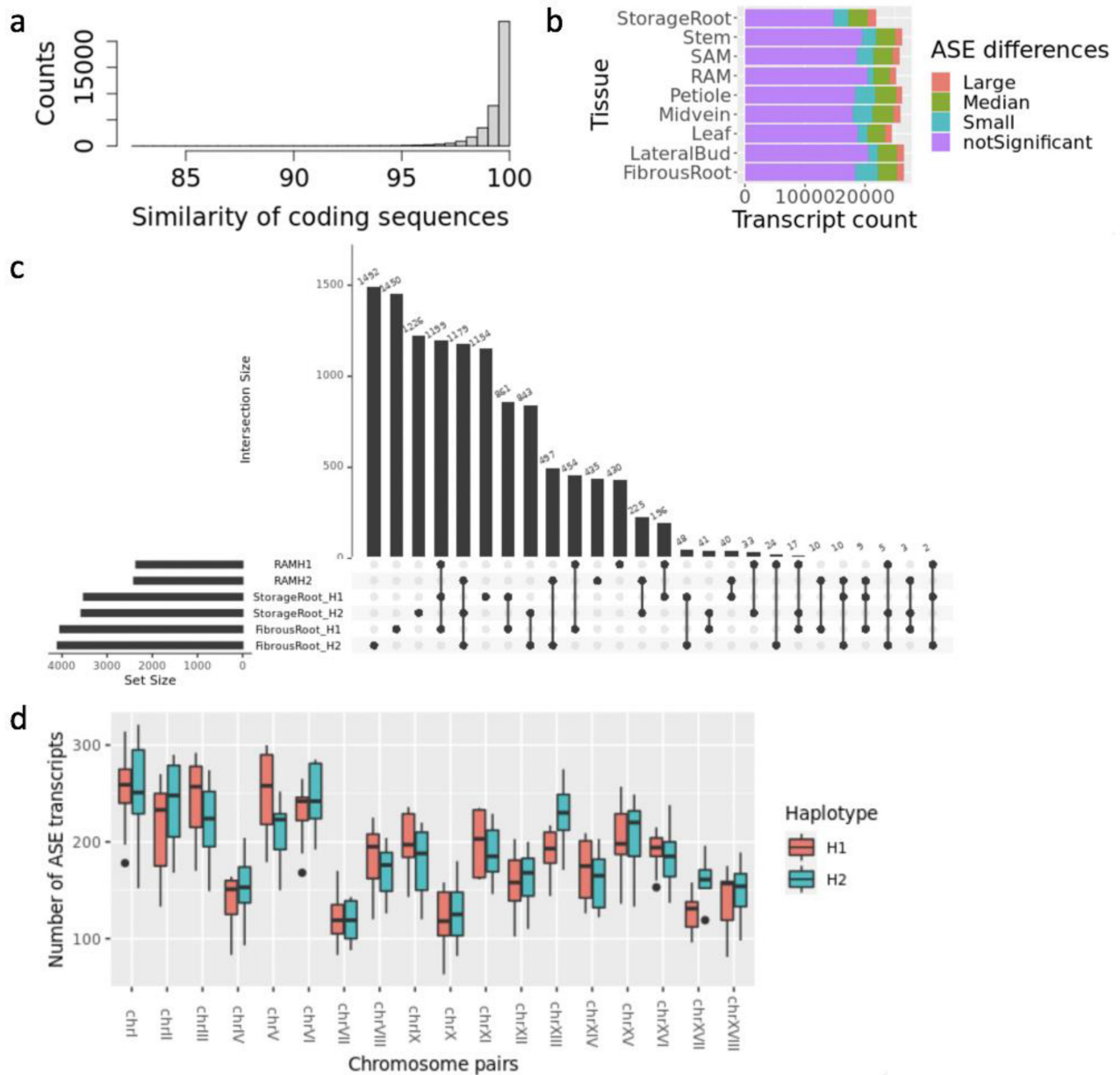
**Figure 5:** Cassava TME204 genome annotation. (a) Gene synteny (99%) between AM560 and the TME204 pseudochromosome pairs revealed by orthologous pairs of transferred reference genes. “ho” and “ht” encode “haplotype one” and “haplotype two” of TME204, respectively. “R” encodes “reference.” Color lines highlight the inverted regions. (b) Recall rate (93%) of the transferred reference genes by *ab initio* gene prediction in the TME204 H1 assembly. For the TME204 H2 assembly the recall rate is 94%. (c) Distribution of *ab initio* predicted gene loci among TME204 chromosome pairs. “known” represents predicted gene loci overlapped with transferred reference gene models. “novel” represents predicted gene loci without any overlapping transferred reference gene models. (d) Amount of novel predicted gene loci on unanchored haplotigs in comparison to that of chromosomal gene loci in TME204 assemblies. In comparison to the H1 assembly, the H2 assembly has less unanchored haplotigs (60 Mb instead of 100 Mb, Table 2). As a result, the number of novel predicted gene loci from these sequences is also lower. (e) Fractions of functionally annotated transcripts, which are grouped similarly as predicted gene loci, as described in c and colored in d. Bubble plots of enriched cellular component (CC) terms in novel chromosomal genes (f) and novel genes from unanchored haplotigs (g). The colors of the bubbles are illustrated from blue to red in descending order of  $-\log_{10}(P\text{-value})$ . The sizes of the bubbles are from small to large in ascending order of total gene counts annotated with the CC terms shown on the y-axis. The x-axis represents the ratio of novel/total gene counts. Results shown are from the TME204 H1 assembly. The enriched CC terms in the TME204 H2 assembly are almost identical (Supplementary Fig. S7c and d).

the 39,028 bi-allelic transcripts from 35,264 orthologous gene loci (see Methods). This is the largest set of bi-allelic transcripts analyzed for cassava so far [8, 9]. Most of these bi-allelic pairs maintained high levels of coding sequence similarity (Fig. 7a). In each tissue (Supplementary Table S8), ~20–30% of the expressed alleles showed significant (adjusted  $P < 0.05$ ) differences in expression between allelic pairs, within the range of previously reported values of 14% [8] and 34% [9]. The ASE differences are mostly small to median fold changes ( $<|8|$  fold). RAM has the least number of alleles with ASE differences, while fibrous root, midvein,

and petiole have the most abundant alleles with ASE differences (Fig. 7b, Supplementary Fig. S8a). When alleles with ASE biases were compared across all 9 tissues, ~8% were consistent such that the ASE was biased towards 1 allele in all tissues, 33% were inconsistent and tissue-specific, and the rest were inconsistent and the ASE bias persists only in some but not all tissues (Supplementary Fig. S8a). Among ASE transcripts in the 3 more closely related subterranean tissues (storage root, fibrous root, and RAM), 20% were consistent, 52% were inconsistent and tissue specific, and 1.6% were inconsistent and showed direction-shifting (Fig. 7c).



**Figure 6:** Haplotype-resolved analysis of differentially expressed transcripts during cassava TME204 tissue development. (a) Clustering analysis of transcripts from both TME204 H1 and H2 assemblies. “hom” encodes transcripts with identical sequences between haplotypes. “het” encodes transcripts with different sequences between haplotypes. “mc” represents multi-copy transcripts within 1 haplotype assembly. “sc” represents single-copy transcripts in 1 haplotype assembly. Identical transcripts (“hom” and “mc”) are collapsed and present only once in the haplotype-resolved reference transcriptome. Multi-copy (“het\_mc” and “hom\_mc”) transcripts are less highly expressed (“EXP”) and differentially regulated (“DE,” adjusted  $P < 0.00001$  and a fold change of  $\log_2|FC| > 2$ ) during TME204 tissue development. (b) Principal component analysis (PCA) of samples based on transcript expression levels. (c) Sample and transcript clustering analysis of the 9,436 differentially expressed transcripts (DETs) across the 9 tissues (adjusted  $P < 0.00001$  and  $\log_2|FC| > 2$ ). The color scale represents row-centered expression levels. (d) Transcripts from chromosomal, novel predicted gene loci are also less highly expressed (“EXP”) and differentially regulated (“DE,” adjusted  $P < 0.00001$  and a  $\log_2|FC| > 2$ ) in different TME204 tissues. (e) ReviGo graphs of enriched biological process (BP) terms in DETs with different haplotype origins: homozygous DETs common to both H1 and H2, DETs from the H1 assembly, and DETs from H2 assembly. Each GO term is a node. Related GO terms are connected by edges between the nodes. Node color indicates the Fisher exact test P-value. The lighter the color, the lower the P-value. Node size corresponds to the frequency of the GO term in the whole UniProt database. (f) Bubble plot of enriched BP terms in the 663 chromosomal, novel predicted DETs.



**Figure 7:** Allele-specific expression in cassava TME204. (a) Coding sequence similarity of the 39,028 bi-allelic transcript pairs in TME204. (b) Number of the bi-allelic transcripts expressed in 9 cassava tissues. The expressed transcripts were further classified into 4 categories: “notSignificant”: no significant allelic expression differences ( $P \geq 0.05$ ); “Small”: significant allelic expression differences with  $\log_2|FC| < 1$  ( $P < 0.05$ ); “Median”: significant allelic expression differences with  $\log_2|FC| \geq 1$  and  $\log_2|FC| < 3$  ( $P < 0.05$ ); and “Large”: significant allelic expression differences with  $\log_2|FC| \geq 3$  ( $P < 0.05$ ). Storage root has fewer expressed transcripts than all other tissues, which is partially due to fewer biological replicates (2 instead of 3). (c) Upset plot showing allele-specific expression (ASE) among the TME204 subterranean tissues. Transcripts with ASE biased towards H1 or H2 alleles in fibrous root, storage root, and RAM are shown as sets at the bottom of the graph, sorted by the number of ASE transcripts in each tissue and haplotype (i.e., set size). Transcripts overlapping between tissues and haplotypes are connected as dots with vertical lines (i.e., transcript intersection categories). The number of transcripts within each intersection category is plotted above as a vertical black bar, labeled with the transcript count. (d) Distribution of transcripts with ASE differences between pseudochromosome pairs in the TME204 diploid genome. Transcripts with ASE biased towards H1 alleles were counted for H1 pseudochromosomes, transcripts with ASE biased towards H2 alleles for H2 pseudochromosomes. Box plots summarize the average across all 9 tissues. The lower and upper hinges correspond to the 25th and 75th percentile. The upper whisker extends from the hinge to the largest value no further than  $1.5 \times IQR$  from the hinge, where IQR is the inter-quartile range (i.e. the distance between the 25th and 75th percentile). The lower whisker extends from the hinge to the smallest value at most  $1.5 \times IQR$  of the hinge. Outlying points beyond the end of the whiskers are plotted individually.

Across the more distantly related tissues, fibrous root, midvein, and petiole, which harbor the most abundant ASE transcripts, similar distributions were observed and the fraction of inconsistent ASE transcripts with direction-shifting was  $\sim 2\%$  (Supplemen-

tary Fig. S8b). The number of upregulated alleles between allelic chromosome pairs was similar across the genome, suggesting that the allelic expression tends to be balanced between haplotypes (Fig. 7d).

**Table 4:** Inter- and intra-genomic diversity of cassava revealed by comparative analysis of assembled contig sequences

Parameter	Inter-genomic	Inter-genomic	Intra-genomic
Reference	AM560	AM560	TME204 H1
Query	TME204 H1	TME204 H2	TME204 H2
Too divergent to be aligned (Mb)	220 (29%)	186 (26%)	181 (24%)
Uniquely aligned (Mb)	387	393	420
Sequence similarity (%) in uniquely aligned regions	98.75	98.79	98.85
No. of SNPs	2,720,699	2,679,237	2,720,467
No. of single-nucleotide indels	1,874,181	1,855,394	1,867,232
No. of Assemblytics small indels (20–50 bp) (% expansion/contraction of repeats)	13,605 (3%)	13,467 (3%)	13,332 (3%)
No. of Assemblytics large indels (50 bp to 10 kb) (% expansion/contraction of repeats)	13,387 (67%)	13,073 (66%)	13,213 (67%)

### Intra- and inter-genomic diversity of cassava genomes

Based on *k*-mer analysis, each TME204 haplotype harbors close to 20% of haplotype-specific *k*-mers. However, analysis of orthologous pairs of coding sequences revealed high gene synteny and coding sequence similarity on average. To systematically investigate sequence differences between the TME204 haplotypes and between cassava cultivars, different methods were applied. In the first approach, we produced reliable alignments between assembled sequences longer than 500 bp, with exact matches >100 bp [34, 35]. With this method, 24–29% of sequences per haploid genome were too divergent to be aligned and thus not accounted for the comparative analysis (Table 4). Between the 2 TME204 haploid genomes, the average level of sequence differences was 1.12%, including 2,526,852 SNPs and 1,733,059 single-nucleotide indels, 13,332 small indels (20–50 bp), and 13,213 large indels (50–10,000 bp). A total of 67% of the large indels were expansion/contraction of repetitive elements, while only 3% of the small indels were of the same types (Table 4, Supplementary Fig. S9). The levels and characteristics of inter-genomic differences between the 2 cassava cultivars (TME204 vs AM560) were similar to those within TME204 diploid genome (Table 4, Supplementary Fig. S9).

As a complementary approach, we also compared HiFi reads directly to AM560 contigs and TME204 haplotigs, which identified not only indels, but also inversions and breakpoints of other complex SVs such as translocations, and so forth. If the TME204 genome was assembled error-free, all sequence variants between 1 TME204 haplotype assembly and HiFi reads would have been heterozygous and representing intra-genomic diversity. Indeed, only <1% of structural variants (SVs) reported by HiFi read alignments were homozygous. They could have resulted from misassemblies and/or misalignments. Most of the SVs (>99%, 115,000) were heterozygous between TME204 haplotypes, confirming that the TME204 haplotigs are structurally accurate and do harbor a high level of intra-genomic sequence differences between haplotypes. The very high number of reported SVs was due to the high sensitivity of the analysis method because SVs supported by 3 or more HiFi reads could be identified with high confidence.

Similarly, between the TME204 diploid genome and the AM560 genome, 198,000 SVs were identified by the HiFi read alignments, of which 70.5% were heterozygous and thus specific to only 1 of the TME204 haplotypes (Fig. 8a). On average, the number of SVs between 1 TME204 haplotype and AM560 haploid genome reached 128,000, which is again very similar to the number of intra-genomic SVs (115,000) between TME204 haplotypes. In addition to the much higher sensitivity, analysis of HiFi read alignments was also able to identify very small inversions such as those from 100 bp to a few kb (Fig. 8b), which were not captured by gene synteny analysis. Consequently, the number of inversions reported with this method was much higher and not directly comparable with the counts from gene synteny analysis.

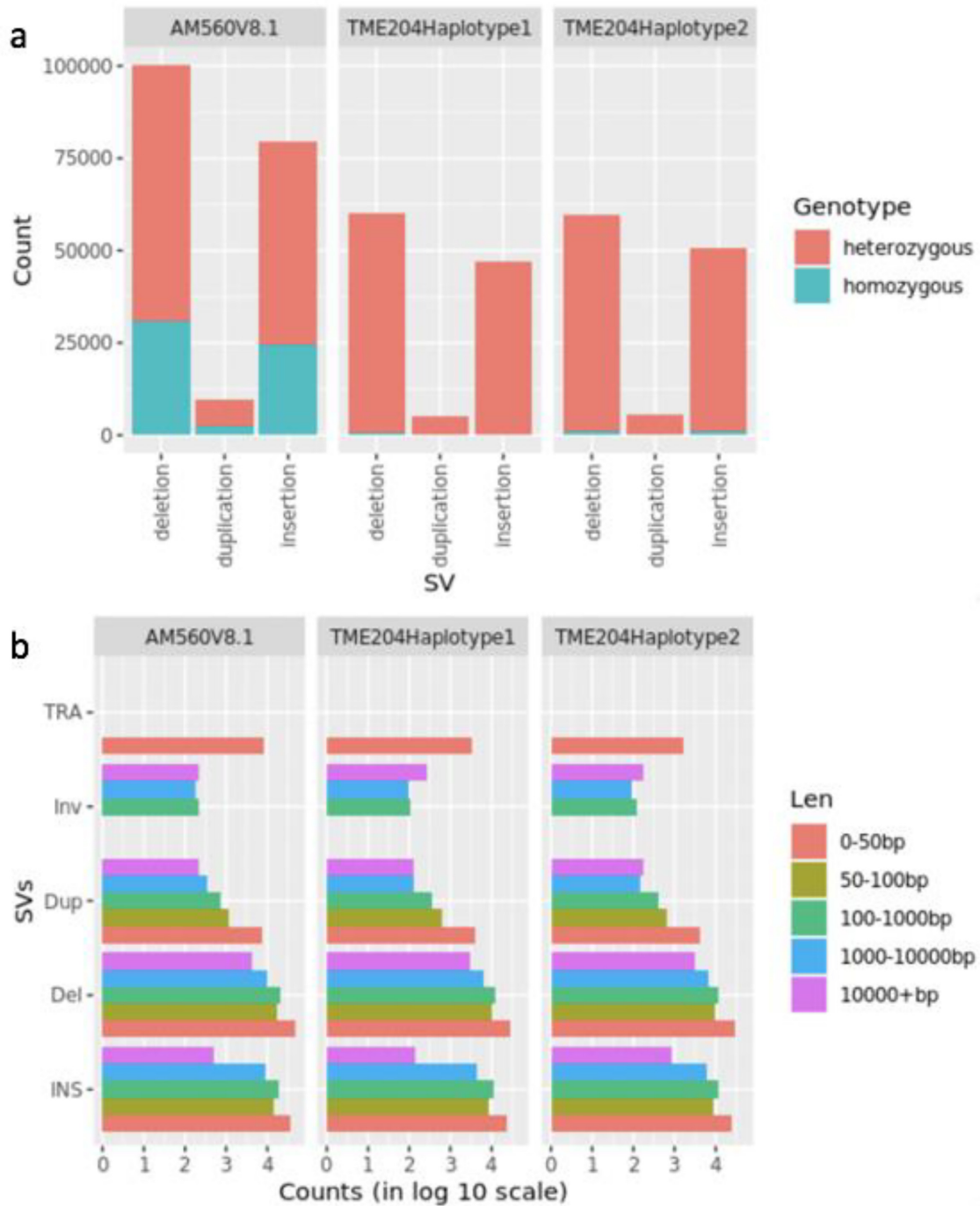
We also compared the TME204/AM560 pseudochromosome pairs by identifying and examining regions that shared information content [36], which is more robust in comparing sequences with low sequence identity and where the linear order of homologs is not preserved [37]. The analysis revealed that each cassava pseudochromosome consists of islands of conserved regions flanked by regions with more degenerated sequences. Although the order of these conserved regions was mostly kept between each pseudochromosome pair, extensive genomic rearrangements still exist (Fig. 9). In total, >2,500 inversions were detected between each pair of cassava haploid genomes with this method (Supplementary File S6).

### Cassava pan-genome

The presence of haplotype-specific *k*-mers and abundant SVs between the cassava haploid assemblies suggests that any of the linear reference genomes of 1 haplotype, either the AM560 pseudohaplotype or TME204 H1 or H2, cannot represent the sequence diversity of cassava populations and may miss haplotype-specific sequences. To overcome this limitation, we built a pan-genome graph from TME204 H1 and H2, and also one including the reference AM560 pseudo-haplotype. Starting with each initial reference haplotype (TME204 H1 or AM560), haplotype-specific large SVs (100 bp and 100 kb) were identified in the query haplotype and subsequently amended to the reference haplotype for pan-genome graph reconstruction. We found 114,773,684 bases representing 40,776 such large SVs in TME204 H2 that were divergent from TME204 H1 (Fig. 10a). In comparison to the linear TME204 H1/H2 as the only reference genome, using the TME204 pan-genome as reference allowed us to map more Illumina reads from the same TME204 sample with higher accuracy (i.e., mapping quality  $\geq 20$ ) (Fig. 10b). In the pan-genome that includes the AM560 genome and the 2 TME204 haplotypes, we found 198,028,264 bases representing 53,098 large SVs in the 2 TME204 haplotypes that were divergent from AM560. As reported above by the Assemblytics analysis, where a majority of large indels (50 bp to 10 kb) are expansion/contraction of repeats, the SV harboring divergent sequences in both pan-genomes are enriched for repeats, especially LTR elements (Fig. 10c), suggesting that most SVs captured by pan-genome graphs are related to LTR retrotransposons.

### Discussion

By comparing PacBio CLR and HiFi sequencing technologies and benchmarking 4 HiFi assemblers [15, 23, 24], we demonstrate that HiFi reads are extremely effective in producing a nearly complete and accurate haplotype-resolved assembly of the complex diploid cassava genome. The combination of high base accuracy and long read length greatly simplified the data analysis work-

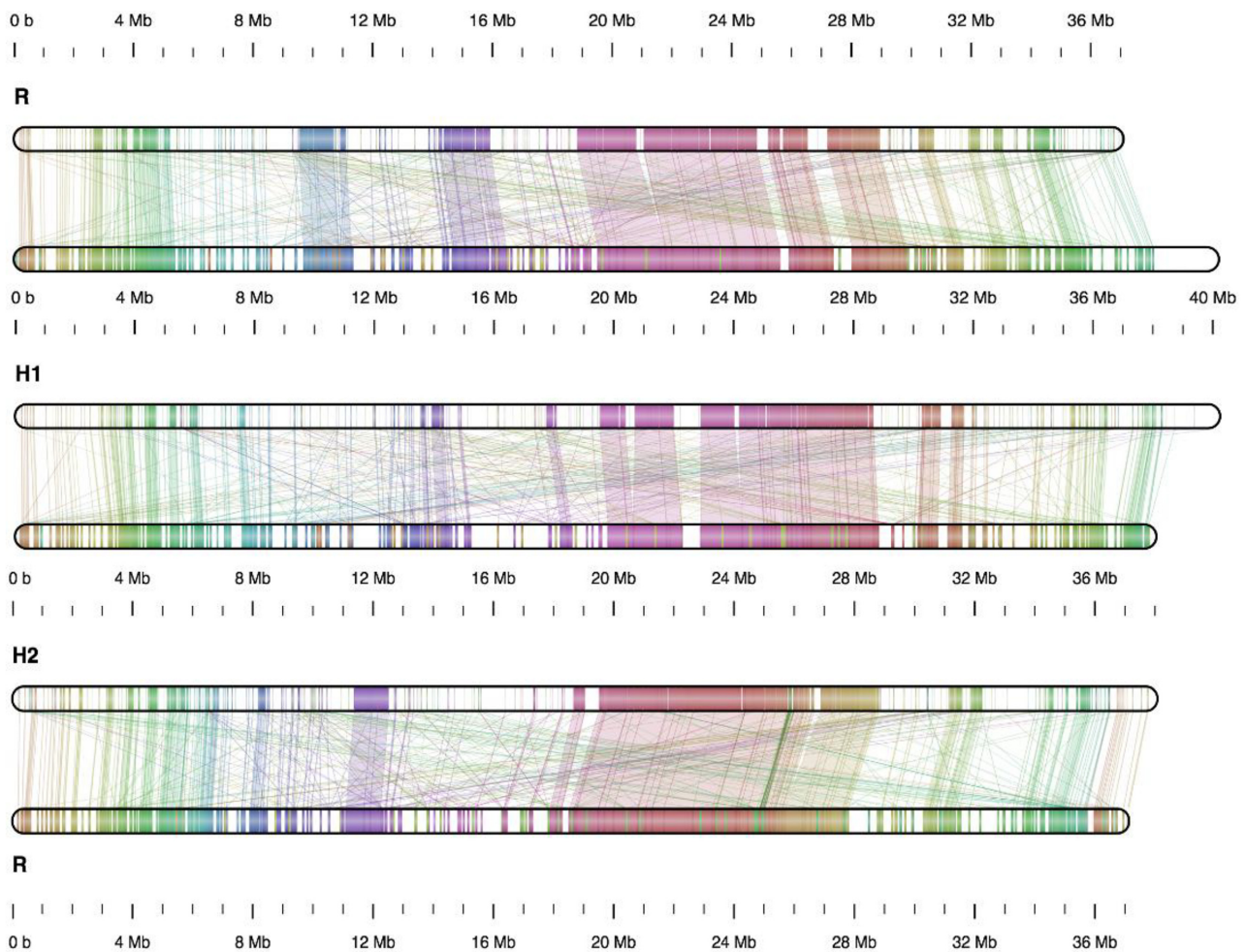


**Figure 8:** Identification of structural variants in cassava TME204 and AM560 genomes by HiFi reads. (a) Classification and counts of SVs by genotypes. (b) Classification and counts of SVs by variant types and length. INS: insertions; Del: deletions; Dup: duplications; INV: inversions; TRA: breakpoints of complex variants with unknown sizes, such as translocations.

flow, decreased data footprints, shortened data analysis time, and improved the assembly quality. CLR-Falcon assembly starts with read self-correction, which is not only computationally expensive but can also mix reads from different haplotype alleles, paralogous gene members, or repetitive elements. In contrast, HiFi reads have higher resolution and accuracy in resolving these sequence variants. All HiFi TME204 assemblies reached consensus accuracy between Q40 (99.99%) and Q50 (99.999%). The CLR-Falcon contig sequences were less accurate even after extensive polishing using signal-level data, which also has the risk of introducing novel

errors because current polishing pipelines cannot accurately differentiate reads from different haplotype alleles and repeat copies [25].

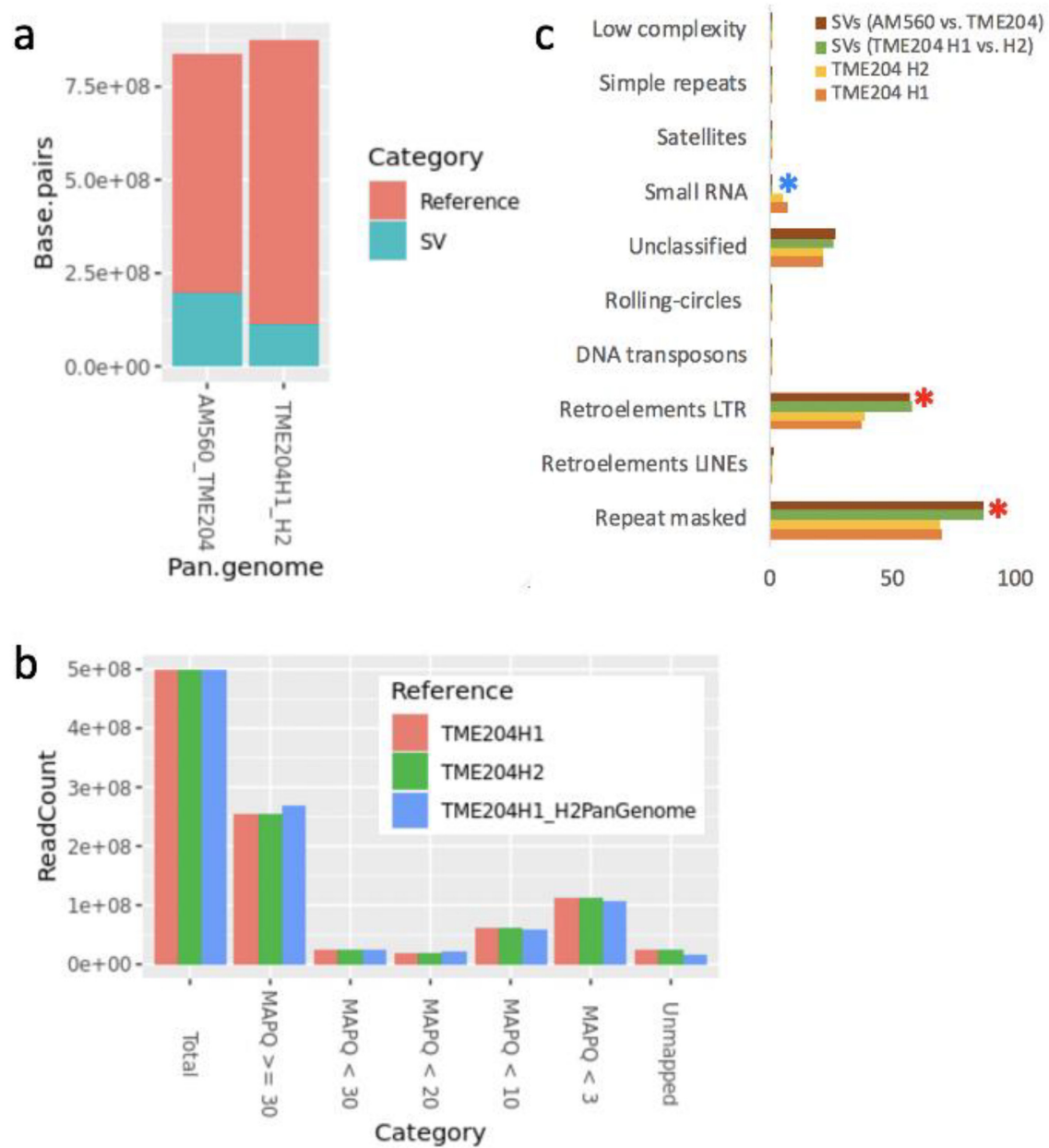
Among the compared HiFi assemblers, hifiasm generated the most completely haplotype-resolved TME204 genome assembly. The haplotigs reached NG50 of 18 Mb, with consensus accuracy of QV45. Three pairs of pseudochromosomes are haplotigs without sequencing gaps. These values satisfy the 6.7.Q40 and 7.C.Q50 genome assembly quality metrics, which are measures for close-to-finished genome qualities as proposed by the VGP consortium



**Figure 9:** Chromosome XII maps show extensive genomic rearrangements between cassava chromosome pairs. “R” indicates the pseudo-chromosome from the reference AM560 v8.0 assembly, “H1” pseudo-chromosome from the TME204 haplotype 1 assembly, “H2” pseudo-chromosome from the TME204 haplotype 2 assembly. Shared regions between chromosome pairs are shown as color segments and connected by color lines between chromosomes. Shared regions with similar sequence information content were detected by Smash++ with parameters adjusted for highly repetitive genomes (Methods). White segments represent regions that are degenerated between a chromosome pair. Such accumulation of degenerated genomic sequences was observed between all pseudo-chromosome pairs (Supplementary File S6), both within the TME204 diploid genome, and between each TME204 haplotype and the AM560 haploid genome.

[25]. In comparison to a recently published correspondence on a chromosome-level phased assembly of cassava cultivar SC205 [9], the TME204 HiFi assembly is more continuous (contig N50 18 vs 1.1 Mb in SC205) and more complete (BUSCO complete score of 99% instead of 88% in SC205). The high accuracy and completeness of assembled sequences improved gene annotation, with 20k novel chromosomal gene loci being predicted with experimental evidence of Iso-Seq transcripts and RNA-seq data. A total of 35,264 bi-allelic gene loci were phased, making the TME204 assembly the most haplotype-resolved cassava genome so far, in comparison to 24,128 in SC205 and 18,723 in 60444. ASE analysis of this largest set of bi-allelic transcripts across 9 TME204 tissues confirmed that most alleles are coordinately expressed during tissue development, as previously reported for cassava [8, 9] and recently reported for ginger [38]. On average, 20–30% of the expressed transcripts showed ASE differences in  $\geq 1$  tissue. The expression bias is often inconsistent and tissue specific. ASE bias with direction-shifting was observed in 2% of the expressed transcripts. This is also similar to the patterns observed in ginger and tea plant [38, 39]. RNA-seq data analysis against the reference transcriptome where haplotype origin of transcripts is known also enabled us to

determine that DETs from different haplotypes may play different roles during cassava tissue development. However, the current TME204 H1 and H2 assembly is still a random mixture of different parental chromosomes because with Hi-C technology alone it is not possible to phase across chromosomes [19]. Trio-binning [40] using 2 parental genomes will be needed to completely separate parental chromosomes in the offspring genome and to assist in the analysis of monoallelic expression of parentally imprinted genes in offspring. Between allelic chromosome pairs, the number of ASE transcripts biased towards each haplotype was similar across the genome (Fig. 7d). Finding ASE transcripts will thus still hold true after reshuffling of pseudo-chromosomes between haplotypes. The number of DETs between allelic chromosome pairs was also similar (Supplementary Fig. S10a). Functional enrichment analysis of DETs per chromosome identified chromosomes with uniquely enriched BP terms (Supplementary Fig. S10a and b). The existence of such chromosomes suggests that haplotype-specific enriched functions in DETs will hold true no matter how the chromosomes are shuffled between haplotypes, although the specific set of DETs and actual enriched GO terms per haploid genome will change. Given the high level of sequence accuracy,



**Figure 10:** Properties of cassava pan-genomes. (a) Cassava pan-genomes across different assemblies and large SVs (100 bp to 100 kb) detected by pan-genome graphs. The pan-genome size decreased when AM560 was included because there were fewer 1-to-1 orthogonal regions between AM560, TME204 H1, and H2. (b) The pan-genome of TME204 H1 and H2 improved mapping rate and mapping quality of Illumina PE reads collected from the same DNA sample. (c) SVs in cassava pan-genomes are enriched with repeats ( $\chi^2$  test  $P < 0.05$ ), especially LTR elements ( $P < 0.05$ ), and are deprived of small RNAs ( $P < 0.005$ ).

functional completeness, and haplotype resolution, the TME204 genome and transcriptome will be powerful resources and tools for establishing new technologies, such as novel marker identifications and genome editing for cassava trait improvement and breeding.

The HiFi sequencing strategy in combination with Hi-C not only enabled the assembly of haplotype-resolved chromosome pairs but also allowed reconstruction of >300 mitochondrial haplotigs with lengths varying between 25 and 76 kb. Plant mitochondrial genomes are known to be highly fragmented, with total lengths varying from 200 to 2,000 kb [41]. The 53 mitochondrial haplotigs in the TME204 H2 assembly added up to a total size of 2 Mb (Table 3), which can represent a complete mitochondrial genome.

Interestingly, there were still 281 mitochondrial haplotigs (with a total length of 12 Mb) in the TME204 H1 assembly, suggesting the presence of different sequence variants of the mitochondrial genome. This result strongly supports the recent discovery of plant mitochondrial genomes as a complex and dynamic mixture of sequence variants [42]. It signifies that the highly accurate base information over very long stretches of DNA molecules provided by the combination of HiFi sequencing with Hi-C technology is powerful in resolving the complexity of multiple haplotypes and isoforms, which will revolutionize and fundamentally improve future assemblies of plant genomes.

Extensive SVs and divergent sequences per haploid genome are dispersed throughout both TME204 haplotypes, and the levels of

intra-genomic (TME204) and inter-genomic (TME204 vs AM560) diversity are similar in cassava. Genome regions with SVs are enriched with repeats, especially LTR elements. Accumulation of SVs and hemizygous sequences have been recently reported for other crops such as grapes, potatoes, and rice and are considered a major force contributing to the cost of domestication [20, 43, 44]. Analysis of SVs in cassava TME204 population samples will help to reveal to what extent SV is driving cassava genome evolution. Our study demonstrates that reference-guided analysis of HiFi read alignment is more sensitive in identifying SVs than comparative analysis of assembled consensus sequences, and thus HiFi sequencing will be a cost-effective method for population-scale analysis of SVs.

The high degree of genomic variations in cassava cultivars also highlights the importance of building a pan-genome [45–48] for research and breeding. Underrepresentation of genetic diversity by any linear haploid cassava genome will limit our understanding of genetic variations in reference-guided analysis, especially when samples are sequenced using Illumina short reads, e.g., in genotyping-by-sequencing and RNA-seq experiments. Haplotype-specific short reads may remain unmapped; thus important genome information may be left undiscovered. Technically, large SVs are a frequent source of errors in aligning short Illumina reads, which may lead to misinterpretation of data [49]. We demonstrate that using a pan-genome reference did increase mapping rate and mapping quality of Illumina reads in comparison to using a conventional linear haploid reference. Detailed investigation of a cassava pan-genome, including more cultivars, and its influence on interpretations of omics data is ongoing and will be reported in the near future.

## Potential implications

Using the HiFi sequencing strategy in combination with Hi-C, we reconstructed 2 chromosome-scale haploid genomes for the diploid cassava TME204 with the highest accuracy and completeness achieved so far, which allowed us to study the sequence, gene content, gene expression, and genome structure with unprecedented resolution. The haplotype-resolved genome and transcriptome will be a valuable resource for cassava breeding and research. The ability to resolve the high complexity of multiple haplotypes and isoforms demonstrated in our study will provide insights for future work on plant genomics.

## Methods

### DNA extraction and Illumina shotgun sequencing

Leaves were collected from 6- to 8-week-old *in vitro*-grown TME204 plants. Genomic DNA was extracted using DNeasy Plant Mini Kit (QIAGEN, Hilden, Germany). The TruSeq DNA Nano Sample Prep Kit v2 (Illumina, San Diego, CA, USA) was used for library preparations according to the manufacturer's instructions (Supplementary Methods). The NovaSeq 6000 (Illumina NovaSeq 6000 Sequencing System, [RRID:SCR\\_016387](#)) was used for cluster generation and sequencing according to the standard protocol for PE sequencing at 2 × 150 bp.

### High molecular weight DNA extraction

Fresh leaves were harvested from *in vitro*-grown TME204 plants kept in the dark for 12–24 hours before harvest, and the petiole and basal midrib were removed with a sterile pair of scissors. A 1-g sample of leaf tissue was then snap-frozen in liquid nitro-

gen and homogenized to a powder with a mortar and pestle. Lysis buffer (9.5 mL of G2 buffer from the Blood & Cell Culture DNA Midi Kit [QIAGEN, Hilden, Germany] and 19 μL of RNase A [100 mg/mL, Sigma Aldrich, Burlington, MA, USA]) was added to the homogenized tissue in a 50-mL conical centrifuge tube (Falcon, Sigma Aldrich, Burlington, MA). Then 500 μL of Proteinase K (20 mg/mL, Roche, Basel, Switzerland) was added to the sample and the mixture was vortexed for 10 seconds. The sample was incubated at 50°C (Memmert Incubator, Büchenbach, Germany) on a laboratory roller for 3 hours. Afterwards the sample was centrifuged for 10 minutes at 20°C at 1,800g. The supernatant was then used for high molecular weight (HMW) genomic DNA extraction according to the Genomic-tips protocol (100/G, Blood & Cell Culture DNA Midi Kit, QIAGEN, Hilden, Germany).

### PacBio CLR and HiFi library preparation and sequencing

The concentration of HMW genomic DNA was measured using a Qubit Fluorometer dsDNA Broad Range assay (Thermo Fisher Scientific, Waltham, MA, USA). The CLR and HiFi library preparations started with 8 and 15 μg HMW DNA, respectively, using the SM-RTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA) according to the manufacturer's instructions (Supplementary Methods). The CLR SMRT bell template-polymerase complex was sequenced on a PacBio Sequel instrument using the Sequel Sequencing Kit 3.0 (PacBio, Menlo Park, CA, USA) with 6 Sequel™ SMRT® Cells 1M v3 (PacBio, Menlo Park, CA, USA), taking a 10-hour movie per cell. The HiFi SMRT bell template-polymerase complex was sequenced on a PacBio Sequel II instrument (PacBio Sequel II System, [RRID:SCR\\_017990](#)) using the Sequel II Sequencing Kit 2.0 (PacBio, Menlo Park, CA, USA) and 1 Sequel™ II SMRT Cell 8M (PacBio, Menlo Park, CA), taking a 30-hour movie.

### Hi-C library preparation and sequencing

A 2-g sample of fresh leaf tissue was harvested from *in vitro*-grown TME204 plants and flash-frozen in liquid nitrogen. The leaf tissue was then shipped on dry ice to Arima Genomics (San Diego, CA, USA) for Hi-C library preparation (Supplementary Methods). The DNA library was sequenced by Arima on the Illumina HiSeq X (Illumina HiSeq X Ten, [RRID:SCR\\_016385](#)) following the manufacturer's protocols, yielding 727,211,240 read pairs (2 × 150 bp) (accession No. ERR5484651).

### RNA isolation, PacBio Iso-Seq library preparation and sequencing

Three different tissues were collected from greenhouse-grown TME204 plants: the top 5 leaves with petioles, apical and lateral meristems including the stem, and fibrous roots. The various tissues were flash-frozen in liquid nitrogen and homogenized with a mortar and pestle. RNA was isolated with the Spectrum Plant Total RNA kit (Sigma-Aldrich, Burlington, MA, USA) according to Protocol A. The quantity and quality of total RNA samples were measured using Qubit RNA BR Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA) and Agilent TapeStation 4200 with RNA-specific tapes (Agilent Technologies, Santa Clara, CA, USA), respectively. Samples with RNA integrity numbers ≥7 were used for Iso-Seq library preparation and sequencing.

PacBio Iso-Seq templates were prepared using the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (New England BioLabs, Ipswich, MA, USA) and PacBio Iso-Seq Express Template Switching Oligos (TSO) (PacBio, Menlo Park, CA, USA), following the PacBio Iso-Seq protocol "Procedure & Checklist—



Iso-Seq Express Template Preparation for Sequel and Sequel II Systems” (PN 101-763-800) (Supplementary Methods). The Iso-seq SMRT bell template-polymerase complex was sequenced on a PacBio Sequel II instrument using Sequel II Sequencing Kit 2.0 (PacBio, Menlo Park, CA, USA) and single Sequel™ II SMRT Cell 8M (PacBio, Menlo Park, CA, USA) taking a 30-hour movie.

### Bacterial artificial chromosome clone library construction, screening, sequencing, and assembly

HMW DNA was prepared from TME204 young leaves as previously described [50, 51]. Agarose-embedded HMW DNA was partially digested with HindIII (New England Biolabs, Ipswich, MA, USA), sized through 2 size selection steps by pulsed-field gel electrophoresis (CHEF Mapper system, Bio-Rad Laboratories, Hercules, CA, USA), and ligated into the pAGIBAC-5 HindIII-Cloning vector. Pulsed-field migration programs, electrophoresis buffer, and ligation desalting conditions were done according to [52]. The insert size of the BAC clones was assessed using the FastNot I restriction enzyme (New England Biolabs, Ipswich, MA, USA) and analyzed by pulsed-field gel electrophoresis. Colony picking was carried out using a robotic workstation QPix2 XT (Molecular Devices, San Jose, CA, USA) using a white/blue selection. White colonies were arranged in 144 384-well (55,296 BAC clones) microtiter plates containing LB medium with chloramphenicol (12.5  $\mu\text{g}/\text{mL}$ ) supplemented with 6% (v/v) glycerol.

Individual BAC clones were selected using radiolabeled ( $\alpha$ -33P)dCTP probes. DNA was extracted from individual clones using Nucleobond Xtra midi kit (Macherey-Nagel) and used for PacBio library preparation by the French Plant Genomic Resources Center (CNRGV) of the French National Research Institute for Agriculture, Food and Environment (INRAE). PacBio sequencing was performed on the Sequel II system with a movie time of 30 hours with 120-min pre-extension step by Gentyane Genomic Platform (INRAE). Circular consensus sequence (CCS) reads per BAC clone were generated using SMRT Analysis Software SMRT Link v9.0.0 (SMRT-Analysis, [RRID:SCR\\_002942](#)) and assembled using hifiasm v0.12.0 (hifiasm, [RRID:SCR\\_021069](#)). More details on BAC clone screening, sequencing, and assembly can be found in Supplementary Methods.

### Sequencing data quality control

The technical quality and potential sample contamination in Illumina PE reads were evaluated using FastQC v0.11.8 (FastQC, [RRID:SCR\\_014583](#)) and FastQ Screen v0.11.1 (FastQ Screen, [RRID:SCR\\_00141](#)), respectively. The technical quality of PacBio raw data was checked using the “QC module” in the PacBio SMRT Analysis Software SMRT Link version 8.0 (SMRT-Analysis, [RRID:SCR\\_002942](#)). Iso-Seq reads were clustered into high-quality (accuracy 99.9%, HQ) transcripts using the “Iso-Seq Analysis” Application in PacBio SMRT Analysis Software (SMRT Link v10.1.0.119588). The technical quality of Hi-C data was checked using HiCUP v0.8.0 (HiCUP, [RRID:SCR\\_005569](#)) [53].

### Estimation of genome properties

Genome complexities such as repeat content and the level of heterozygosity were evaluated with  $k$ -mers in the Illumina PE reads using Preqc in SGA v0.10.15 (SGA, [RRID:SCR\\_001982](#)) [26, 54]. Analyzed datasets and their accessions are as follows: human (ERR091571-ERR091574) [26], cassava AM560 (SRR2847385), cassava TME204 (ERR5484652), cassava 60444 (ERR5484654) (8), cassava TME3 (ERR5484653) (8).

### PacBio CLR and HiFi whole-genome assembly

PacBio CLR reads were assembled using Falcon (Falcon, [RRID:SCR\\_016089](#)) [23] in pb-assembly (v0.06). PacBio HiFi reads were assembled using multiple HiFi-specific assemblers, including Falcon in pb-assembly (v0.0.8), Improved Phased Assembler IPA (v1.0.5) [55], hifiasm v0.7 (hifiasm, [RRID:SCR\\_021069](#)) [19], and HiCanu (v2.0, [RRID:SCR\\_015880](#)) [15]. Default options were used unless otherwise noted. Improved phased assembly (IPA) was run with both phasing and polishing included.

### Benchmarking analysis of assembly accuracy and completeness

Assembly statistics were collected using QUAST v4.5 (QUAST, [RRID:SCR\\_001228](#)) [56]. NG50 [57] was calculated using the haploid genome size of 750 Mb. Base-level accuracy and completeness was measured using both mapping-based method and alignment-free,  $k$ -mer-based method Merqury (v1.1) [28] (Supplementary Methods).

For evaluation of structural accuracy, Merqury  $k$ -mer analysis results were first used to compute false duplication rates, where  $k$ -mers that appeared more than twice in each haploid assembly were used to identify artificial duplications. PacBio CLR reads were then aligned to each haploid genome and the coverage was analyzed using Asset software [58]. Assembled regions supported by  $\geq 10$  PacBio CLR reads were identified as reliable regions.

Functional completeness was measured using BUSCO v5 (BUSCO, [RRID:SCR\\_015008](#)) completeness of single-copy orthologs discovered in plants (Viridiplantae Odb10) [59], and alignment rates of reference genes and TME204 Iso-Seq transcripts. The AM560 reference genome (v8.0) and gene annotation (v8.1) [31] were downloaded from JGI Phytozome 13 [60]. Reference coding sequences (CDSs) were aligned to TME204 haplotigs using minimap2 (v2.15r905, -cxssplice -C5) (Minimap2, [RRID:SCR\\_018550](#)) [61]. The “asmgene” completeness and duplication scores [19] were calculated using the “paftools” script from the minimap2 package, based on CDSs mapped at  $\geq 97\%$  identity over  $\geq 99\%$  of the CDS length. Iso-Seq data collected from TME204 transcriptomes of fibrous root, stem meristems, and leaves were spliced aligned using minimap2 (v2.15r905, x splice:hq). Alignment statistics were collected using alignqc [62].

### Haplotype-resolved, phased contig assembly using HiFi reads integrated with Hi-C technology

Two sets of haplotype-resolved, phased contig (haplotig) assemblies were generated using hifiasm (v0.15.3) with a combination of HiFi reads and PE Hi-C reads. Haplotigs were first validated against the high-density genetic map of cassava [30], which contains 22,403 SNP markers with allele numbers ranging from 2 to 6. Allelic sequences (50 nt upstream sequence + allele sequence + 50 nt downstream sequence) were aligned to haplotigs using BLAT v3.2.1 (BLAT, [RRID:SCR\\_011919](#)) [63]. For each haplotig, correlation plots of genetic vs physical distance based on uniquely and perfectly aligned alleles were generated for visual inspection. Sequences of BACs were also aligned to haplotigs using BLAT (v3.2.1). The best BAC-to-haplotig alignment was manually inspected to identify resolved BACs, where 1 continuous BAC-to-haplotig alignment was produced.

### Construction of pseudochromosomes

Hi-C reads were mapped back to each set of haplotigs independently using the Arima mapping pipeline [64] and were used to

further scaffold haplotigs with SALSA2 (v2.2, assisted by the assembly graph, resolved misassemblies, 5 iterations) [65]. No haplotigs in TME204 H1 were further scaffolded with Hi-C data after 5 rounds of iteration (Supplementary File S1). Thirty haplotigs in TME204 H2 were scaffolded into 13 scaffolds, of which 7 were chromosome-scale and consistent with the genetic map (Supplementary Table S5). One scaffold was apparently a technical artifact based on genetic markers, reaching 107 Mb long and joining haplotigs from different chromosomes (I, XVI, and XVIII) together (Supplementary Table S5). Chromosomes VII, IX, and XI were not reconstructed in H2. Because Hi-C scaffolding did not generate results for H1 and the results for H2 were not all satisfactory, ALLMAPS v0.8.12 (ALLMAPS, [RRID:SCR\\_021171](#)) [66] was used to reconstruct pseudochromosomes for both sets of haplotigs based on the genetic map [30]. Given the observed high congruence between the map and haplotigs, as well as between the map and Hi-C scaffolds, the choice of scaffolding strategy was reasonable and sound.

### Repeat modeling, genome masking and annotation

Starting with the assembly of all resolved alleles (i.e., primary plus alternate contigs), repeat elements were predicted using RepeatModeler v2.0.1 (RepeatModeler, [RRID:SCR\\_015027](#)), with dependency on TRF (v4.09) [67], RECON v1.08 (RECON, [RRID:SCR\\_021170](#)) [68], RepeatScout v1.0.6 (RepeatScout, [RRID:SCR\\_014653](#)) [69], and RepeatMasker v4.1.0 (RepeatMasker, [RRID:SCR\\_012954](#)) [70]. Analysis of long terminal repeats (LTRs) was enabled with GenomeTools v1.5.9 (GenomeTools, [RRID:SCR\\_016120](#)) [71], LTR\_Retrieve v2.9.0 (LTR\_Retrieve, [RRID:SCR\\_017623](#)) [72], Ninja (v0.95-cluster\_only) [73], MAFFT v7.471-with-extensions (MAFFT, [RRID:SCR\\_011811](#)) [74], and CD-HIT v4.8.1 (CD-HIT, [RRID:SCR\\_007105](#)) [75]. Among the 1,436 predicted repeat families, 1,021 were unknown/novel according to RepeatClassifier (V2.0.1) [76]. Five predicted repeat families with significant hits to plant genes were identified and removed from the repeat library using ProExcluder (v1.1) [77], and each TME204 haplotype assembly was then masked using RepeatMasker (-no\_is -nolow -norma -xsmall) (v4.1.0).

Genome annotation was first performed by transferring reference gene models from AM560 v8.1 to TME204 haplotype assemblies using liftoff (v1.6.1) [78]. Finding extra copies of the same genes was enabled with a minimum sequence identity of 95% in exons/CDSs. Synteny analysis of orthologous pairs of transferred genes was performed using MCScanX [79] and visually inspected using SynVisio [80].

Complementary to the transferred reference gene models, *ab initio* gene prediction was performed using AUGUSTUS (Augustus, [RRID:SCR\\_008417](#)) [81] with experimental evidence from RNA-seq data, Iso-Seq transcripts, and AM560 v8.1 protein/transcript sequences (Supplementary Methods). Genomic locations of predicted gene loci were compared to those of transferred gene loci using BEDTools v2.29.2 (BEDTools, [RRID:SCR\\_006646](#)). Predicted protein sequences were compared against AM560 v8.1 protein sequences and Uniprot/Swiss-Prot (release 2021\_03) using blastp v2.10.1+ (BLASTP, [RRID:SCR\\_001010](#)) and InterPro using InterProScan v5.52–86.0 (InterProScan, [RRID:SCR\\_005829](#)). The best protein matches from AM560 v8.1, Uniprot/Swiss-Prot, and InterPro, plus Gene Ontology (GO) terms and pathways, were used to functionally annotate predicted genes.

### Haplotype-resolved transcriptome analysis of differentially expressed transcripts

Transcripts annotated in TME204 H1 and H2 were pooled and de-duplicated using cd-hit-est (v4.8.1) [75] to generate the haplotype-resolved reference transcriptome, for expression quantification and differential expression analysis of transcripts. Transcripts that are identical between TME204 H1 and H2 were counted as homozygous alleles (hom). Transcripts with different sequences between TME204 H1 and H2 were counted as heterozygous alleles (het). Transcripts duplicated within 1 haploid genome were counted as multi-copy alleles (mc). Transcripts without an identical copy within the same haploid genome were counted as unique/single-copy alleles (sc). Duplicated transcripts (homozygous and/or multi-copy) were collapsed and represented only once in the reference transcriptome. RNA-seq reads from 9 tissues with 3 biological replicates [32] were mapped to the haplotype-resolved reference transcriptome using kallisto v0.46.1, un-stranded (kallisto, [RRID:SCR\\_016582](#)) [82]. All tissues were compared against stem tissue. In each pair-wise comparison of 6 samples, a transcript was considered to be expressed if the expression value of  $\geq 2$  samples exceeded 1 TPM (transcript per million). Differentially expressed transcripts (DETs) between tissues were identified using DESeq2 v1.32.0 (DESeq2, [RRID:SCR\\_015687](#)) [83] as those with a  $\log_2|FC| > 2$  and adjusted  $P < 0.00001$ .

### Analysis of allele-specific expression

For ASE, bi-allelic transcripts were identified by reciprocal blastn (v2.10.0) (BLASTN, [RRID:SCR\\_001598](#)) comparison of H1 and H2 transcripts. A unique, bi-directional best-matched transcript pair was considered as allele A and B. Expression values for bi-allelic transcripts were a subset from the master quantification table including all resolved alleles. ASE was determined using the same package DESeq2, with adjusted  $P < 0.05$ . ASE transcripts overlapped between tissues and haplotypes were analyzed using upSetR [84].

### Comparative genomics

For alignment-based sequence similarity analysis, the cassava reference genome AM560 v8.0 was first disassembled into contig sequences using the utility function "split\_scaffold" in IDBA (v1.1.3) [85]. Each set of TME204 haplotigs was then aligned to the AM560 reference contigs and against each other using nucmer (-maxmatch -l 100 -c 500) in MUMmer v4.0.0beta2 (MUMmer, [RRID:SCR\\_018171](#)) [35], which reported all sequence alignments longer than 500 bp with each exact match longer than 100 bp. Contigs rather than pseudochromosomes were used to prevent false-positive results when the padding Ns in the query did not match perfectly to the distance in the reference. Sequence alignments were further analyzed using dnadiff in MUMmer [35] and Assemblytics [34] for identification of SNPs, single-nucleotide indels, and large indels (20 bp to 10 kb).

For SV analysis using HiFi reads, TME204 HiFi reads were aligned to reference contigs of AM560 and TME204 haplotigs using minimap2 v2.15r905 (Minimap2, [RRID:SCR\\_018550](#)) [61]. SVs were called using PacBio structural variant calling and analysis tools PBSV [86]. Summary statistics of SVs were collected using SURVIVOR (v1.0.7) [87].

For chromosome-level comparisons, the alignment-free method smash++ (v20.04) [36] was used to identify similar/shared regions and genomics rearrangements larger than

10 kb between pseudochromosome pairs. Parameters adjusted for analyzing highly repetitive genomes were as follows: filter scale = large, filter size = 50,000, filter type = blackman, threshold = 1.0, minimum segment size = 10,000.

### Pan-genome analysis

Pan-genomes were constructed using minigraph (v0.15-r426) [49]. Large SVs (100 bp to 100 kb) were identified and extracted from each pan-genome graph using gfatools (0.4-r214) [88].

### Gene ontology enrichment analysis

For all selected gene sets, GO enrichment analysis was performed using topGO v2.44.0 (topGO, RRID:SCR\_014798) [89, 90] with Fisher exact test P-value cut-off set to 0.00001. The only exception was for the 141 homozygous DETs, where the P-value cut-off was set to 0.001. GO annotation of the *ab initio* predicted gene models were used as the background gene set.

### Data Availability

Raw sequencing read data from PacBio (HiFi, CLR, and Iso-Seq) and Illumina (Hi-C and shotgun) underlying this article are available in the European Nucleotide Archive (ENA) database and can be accessed with accession No. PRJEB43673 (or ERP127652 as the secondary accession number in ENA). Assembled genome sequences of TME204 H1 and H2 are available in the NCBI database and can be accessed with accession No. PRJNA758616 and PRJNA758615, respectively. Assembled BAC clone sequences are available in the NCBI GenBank database and can be accessed with accession Nos. MZ959795, MZ959796, MZ959797, and MZ959798. All supporting data and materials are available in the GigaScience GigaDB database [91].

### Abbreviations

ASE: allele-specific expression; BAC: bacterial artificial chromosome; BP: biological process; bp: base pairs; CCS: circular consensus sequence; CDS: coding sequence; CLR: continuous long reads; CMD: Cassava Mosaic Diseases; CPU: central processing unit; DE: differentially expressed/differential expression; DET: differentially expressed transcript; ENA: European Nucleotide Archive; GO: gene ontology; HiFi: high-fidelity; HMW: high molecular weight; Indel: insertion and deletion; IPA: Improved Phased Assembler; kb: kilobase pairs; Mb: megabase pairs; MF: molecular function; NCBI: National Center for Biotechnology Information; numt's: nuclear mitochondrial pseudogene regions; PacBio: Pacific Biosciences; PE: paired-end; QV: quality value; RAM: root apical meristem; SAM: shoot apical meristem; SMRT: Single Molecule Real-Time; SNP: single-nucleotide polymorphism; SV: structural variation; TPM: transcript per million; VGP: the Vertebrate Genome Project.

### Consent for Publication

The cassava TME204 (Tropical Manihot esculenta 204) cultivar used in our study was obtained by ETH Zurich from the International Institute of Tropical Agriculture (IITA) in Nigeria in 2003 prior to the implementation of the International Treaty on Plant Genetic Resources for Food and Agriculture [92]. TME204 has been part of the ETH Zurich cassava germplasm collection since 2003. As a major crop, non-genetically modified cassava, including the wild-type TME204 cultivar, is exempt from the Cartagena Protocol

on Biosafety to the Convention on Biological Diversity [93]. The study reported in our article follows all Swiss and international guidelines and legislation.

### Competing Interests

The authors declare that they have no competing interests.

### Funding

This work was supported by the Bill & Melinda Gates Foundation (INV-008213), ETH Zurich, and the Functional Genomics Center Zurich (FGCZ). D.P. is funded by national funds through FCT (Fundação para a Ciência e a Tecnologia, I.P.) under the Institutional Call to Scientific Employment Stimulus (reference CEECINST/00026/2018). W.G. is supported by a Yushan Scholarship of the Ministry of Education in Taiwan.

### Authors' Contributions

W.Q., Y.L., A.P., R.S., and W.G. designed the study. Y.L. and C.C. prepared DNA and RNA samples for sequencing. A.P., S.G., and A.B. prepared CLR, HiFi, and Iso-Seq libraries and performed PacBio sequencing. Y.L., N.R., E.P., S.V., and M.F. generated the BAC sequences. W.Q., Y.L., P.S., D.P., and W.G. analyzed data. W.Q., Y.L., A.P., A.B., P.S., and W.G. wrote the manuscript. All authors reviewed the final manuscript before submission.

### Acknowledgments

We thank the high-throughput sequencing team at FGCZ for Illumina sequencing service, Arima Genomics for Hi-C service, Dr. David Stucki and Deborah Moine from PacBio for their technical support, Dr. Haoyu Cheng from Harvard Medical School and Alaina Shumate from the Johns Hopkins University School of Medicine for insightful discussion. We thank Jay Tracy from FGCZ for reviewing the manuscript for English writing and clarity.

### References

1. Claros, MG, Bautista, R, Guerrero-Fernández, D, et al. Why assembling plant genome sequences is so challenging. *Biology* 2012;**1**(2):439–59.
2. Birky, CW Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics* 1996;**144**(1):427–37.
3. Balloux, F, Lehmann, L, de Meeùs, T. The population genetics of clonal and partially clonal diploids. *Genetics* 2003;**164**(4):1635–44.
4. Meloni, M, Reid, A, Caujapé-Castells, J, et al. Effects of clonality on the genetic variability of rare, insular species: the case of *Ruta microcarpa* from the Canary Islands. *Ecol Evol* 2013;**3**(6):1569–79.
5. Michael, TP, VanBuren, R. Building near-complete plant genomes. *Curr Opin Plant Biol* 2020;**54**:26–33.
6. Tørresen, OK, Star, B, Mier, P, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res* 2019;**47**(21):10994–1006.
7. Bredeson, JV, Lyons, JB, Prochnik, SE, et al. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat Biotechnol* 2016;**34**(5):562–70.
8. Kuon, J-E, Qi, W, Schläpfer, P, et al. Haplotype-resolved genomes of geminivirus-resistant and geminivirus-susceptible African cassava cultivars. *BMC Biol* 2019;**17**(1):75.

9. Hu, W, Ji, C, Shi, H, et al. Allele-defined genome reveals biallelic differentiation during cassava evolution. *Mol Plant* 2021;**14**(6):851–4.
10. Wang, W, Feng, B, Xiao, J, et al. Cassava genome from a wild ancestor to cultivated varieties. *Nat Commun* 2014;**5**:5110.
11. Chen, F, Song, Y, Li, X et al. Genome sequences of horticultural plants: past, present, and future. *Hortic Res* 2019;**6**:112.
12. Prochnik, S, Marri, PR, Desany, B, et al. The cassava genome: current progress, future directions. *Trop Plant Biol* 2012;**5**(1):88–94.
13. van Dijk, EL, Jaszczyszyn, Y, Naquin, D, et al. The third revolution in sequencing technology. *Trends Genet* 2018;**34**(9):666–81.
14. A reference standard for genome biology. *Nat Biotechnol* 2018;**36**:1121.
15. Nurk, S, Walenz, BP, Rhie, A, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* 2020;**30**(9):1291–305.
16. Wenger, AM, Peluso, P, Rowell, WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019;**37**(10):1155–62.
17. Vollger, MR, Logsdon, GA, Audano, PA, et al. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Ann Hum Genet* 2020;**84**(2):125–40.
18. Porubsky, D, Ebert, P, Audano, PA, et al. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat Biotechnol* 2021;**39**(3):302–8.
19. Cheng, H, Concepcion, GT, Feng, X, et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 2021;**18**(2):170–5.
20. Zhou, Q, Tang, D, Huang, W, et al. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat Genet* 2020;**52**(10):1018–23.
21. Edger, PP, Poorten, TJ, VanBuren, R, et al. Origin and evolution of the octoploid strawberry genome. *Nat Genet* 2019;**51**(3):541–7.
22. Rabbi, IY, Hamblin, MT, Kumar, PL, et al. High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding. *Virus Res* 2014;**186**:87–96.
23. Chin, C-S, Peluso, P, Sedlazeck, FJ, et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* 2016;**13**(12):1050–4.
24. Cheng, H, Concepcion, GT, Feng, X, et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 2021;**18**(2):170–5.
25. Rhie, A, McCarthy, SA, Fedrigo, O, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 2021;**592**(7856):737–46.
26. Simpson, JT. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* 2014;**30**(9):1228–35.
27. Iqbal, Z, Caccamo, M, Turner, I, et al. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* 2012;**44**(2):226–32.
28. Rhie, A, Walenz, BP, Koren, S, et al. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* 2020;**21**(1):245.
29. Chin, C-S, Alexander, DH, Marks, P, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013;**10**(6):563–9.
30. International Cassava Genetic Map Consortium (ICGMC). High-resolution linkage map and chromosome-scale genome assembly for cassava (*Manihot esculenta* Crantz) from 10 populations. *G3 (Bethesda)* 2014;**5**(1):133–44.
31. Phytozome. Phytozome info: *Manihot esculenta* v8.1. [https://phytozome-next.jgi.doe.gov/info/Mesculenta\\_v8\\_1](https://phytozome-next.jgi.doe.gov/info/Mesculenta_v8_1). Accessed 14 February 2022.
32. Wilson, MC, Mutka, AM, Hummel, AW, et al. Gene expression atlas for the food security crop cassava. *New Phytol* 2017;**213**(4):1632–41.
33. Smet, RD, Adams, KL, Vandepoele, K, et al. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A* 2013;**110**(8):2898–903.
34. Nattestad, M, Schatz, MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* 2016;**32**(19):3021–3.
35. Marçais, G, Delcher, AL, Phillippy, AM, et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol* 2018;**14**(1):e1005944.
36. Hosseini, M, Pratas, D, Morgenstern, B, et al. Smash++: an alignment-free and memory-efficient tool to find genomic rearrangements. *Gigascience* 2020;**9**(5):doi:10.1093/gigascience/giaa048.
37. Zielezinski, A, Girgis, HZ, Bernard, G, et al. Benchmarking of alignment-free sequence comparison methods. *Genome Biol* 2019;**20**(1):144.
38. Cheng, S-P, Jia, K-H, Liu, H, et al. Haplotype-resolved genome assembly and allele-specific gene expression in cultivated ginger. *Hortic Res* 2021;**8**(1):188.
39. Zhang, X, Chen, S, Shi, L, et al. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nat Genet* 2021;**53**(8):1250–9.
40. Koren, S, Rhie, A, Walenz, BP, et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* 2018;**36**(12):1174–82.
41. Morley, SA, Nielsen, BL. Plant mitochondrial DNA. *Front Biosci (Landmark Ed)* 2017;**22**(6):1023–32.
42. Kozik, A, Rowan, BA, Lavelle, D, et al. The alternative reality of plant mitochondrial DNA: one ring does not rule them all. *PLoS Genet* 2019;**15**(8):e1008373.
43. Zhou, Y, Minio, A, Massonnet, M, et al. The population genetics of structural variants in grapevine domestication. *Nat Plants* 2019;**5**(9):965–79.
44. Kou, Y, Liao, Y, Toivainen, T, et al. Evolutionary genomics of structural variation in Asian rice (*Oryza sativa*) domestication. *Mol Biol Evol* 2020;**37**(12):3507–24.
45. Bayer, PE, Golicz, AA, Scheben, A, et al. Plant pan-genomes are the new reference. *Nat Plants* 2020;**6**(8):914–20.
46. Della Coletta, R, Qiu, Y, Ou, S, et al. How the pan-genome is changing crop genomics and improvement. *Genome Biol* 2021;**22**(1):3.
47. Li, J, Yuan, D, Wang, P, et al. Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biol* 2021;**22**(1):119.
48. Sun, X, Jiao, C, Schwaninger, H, et al. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat Genet* 2020;**52**(12):1423–32.
49. Li, H, Feng, X, Chu, C. The design and construction of reference pangene graphs with minigraph. *Genome Biol* 2020;**21**(1):265.
50. Peterson, DG, Tomkins, JP, Frisch, DA, et al. Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide. *J Agric Genomics* 2000;**5**:1–3.
51. Gonthier, L, Bellec, A, Blassiau, C, et al. Construction and characterization of two BAC libraries representing a deep-coverage of the genome of chicory (*Cichorium intybus* L., Asteraceae). *BMC Res Notes* 2010;**3**:225.

52. Chalhouh, B, Belcram, H, Caboche, M. Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size. *Plant Biotechnol J* 2004;**2**(3):181–8.
53. Wingett, S, Ewels, P, Furlan-Magaril, M, et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* 2015;**4**:1310.
54. Simpson, JT, Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* 2012;**22**(3):549–56.
55. Pacific Biosciences. IPA HiFi Genome Assembler. Accessed: June 28, 2020.
56. Gurevich, A, Saveliev, V, Vyahhi, N, et al. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013;**29**(8):1072–5.
57. Earl, D, Bradnam, K, St John, J, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* 2011;**21**(12):2224–41.
58. Guan, D. Asset. <https://github.com/dfguan/asset>. Accessed: 2 June, 2021.
59. Seppey, M, Manni, M, Zdobnov, EM, et al. Gene prediction: methods and protocols. *Methods Mol. Biol* 1962:227–45.
60. Goodstein, DM, Shu, S, Howson, R, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 2012;**40**(D1):D1178–86.
61. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**(18):3094–100.
62. Weirather, JL, de Cesare, M, Wang, Y, et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res* 2017;**6**:100.
63. Kent, WJ. BLAT—The BLAST-Like Alignment Tool. *Genome Res* 2002;**12**(4):656–64.
64. Arima. Arima mapping pipeline. [https://github.com/ArimaGenomics/mapping\\_pipeline](https://github.com/ArimaGenomics/mapping_pipeline). Accessed: January 8, 2019.
65. Ghurye, J, Rhie, A, Walenz, BP, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol* 2019;**15**(8):e1007273.
66. Tang, H, Zhang, X, Miao, C, et al. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol* 2015;**16**(1):3.
67. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;**27**(2):573–80.
68. Bao, Z, Eddy, SR. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* 2002;**12**(8):1269–76.
69. Wootton, JC, Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 1993;**17**(2):149–63.
70. Smit, A, Hubley, R, Green, P. RepeatMasker Open-4.0. 2013-2015. <https://www.repeatmasker.org/RepeatMasker/>, April 1st, 2021.
71. Gremme, G, Steinbiss, S, Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinform* 2013;**10**(3):645–56.
72. Ou, S, Jiang, N. LTR\_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* 2018;**176**(2):1410–22.
73. Wheeler, T. Large-scale neighbor-joining with NINJA. In: SL Salzberg, T Warnow, eds. *Algorithms in Bioinformatics*. Berlin, Heidelberg: Springer; 2009:375–89.
74. Katoh, K, Standley, DM. MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. *Mol Biol Evol* 2013;**30**(4):772–80.
75. Fu, L, Niu, B, Zhu, Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**(23):3150–2.
76. Flynn, JM, Hubley, R, Goubert, C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* 2020;**117**(17):9451–7.
77. Campbell, MS, Law, M, Holt, C, et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol* 2014;**164**(2):513–24.
78. Shumate, A, Salzberg, SL. Liftoff: an accurate gene annotation mapping tool. *Bioinformatics* 2020;**37**(12):1639–43.
79. Wang, Y, Tang, H, Debarry, JD, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 2012;**40**(7):e49.
80. Bandi, VK. SynVisio: A Multiscale Tool to Explore Genomic Conservation. MA Thesis. University of Saskatchewan; 2020. <http://hdl.handle.net/10388/12895>. Accessed: December 1st, 2021.
81. Hoff, KJ, Stanke, M. Predicting genes in single genomes with AUGUSTUS. *Curr Protoc Bioinformatics* 2019;**65**(1):e57.
82. Bray, NL, Pimentel, H, Melsted, P, et al. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;**34**(5):525–7.
83. Love, MI, Huber, W, Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**(12):550.
84. Conway, JR, Lex, A, Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 2017;**33**(18):2938–40.
85. Peng, Y, Leung, HCM, Yiu, SM, et al. IDBA - a practical iterative de Bruijn graph de novo assembler. In: B Berger, ed. *Research in Computational Molecular Biology*. RECOMB 2010. Berlin, Heidelberg: Springer; 2010. doi:10.1007/978-3-642-12683-3\_28.
86. Pacific Biosciences. PacificBiosciences/pbsv. <https://github.com/PacificBiosciences/pbsv>. Accessed: April 20, 2021.
87. Jeffares, DC, Jolly, C, Hoti, M, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* 2017;**8**:14061.
88. Li, H. lh3/gfatools. <https://github.com/lh3/gfatools>. Accessed: April 1, 2021.
89. Alexa, A, Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology. 2021. R package version 2.46.0.
90. Alexa, A, Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology. Bioconductor version: Release (3.14). <https://bioconductor.org/packages/release/bioc/html/topGO.html>. Accessed: December 19, 2021.
91. Qi, W, Lim, Y, Patrignani, A, et al. The haplotype-resolved chromosome pairs and transcriptome data of a heterozygous diploid African cassava cultivar. *GigaScience Database* 2022. <http://doi.org/10.5524/102193>.
92. International Treaty on Plant Genetic Resources for Food and Agriculture. <https://www.fao.org/3/i0510e/i0510e.pdf>. Accessed: June 25, 2019.
93. Cartagena Protocol on Biosafety to the Convention on Biological Diversity. <https://www.cbd.int/doc/legal/cartagena-protocol-en.pdf>. Accessed: June 25, 2019.