# A BAC-guided haplotype assembly pipeline increases the resolution of the virus resistance locus *CMD2* in cassava

Cornet Luc*§[a], Syed Shan-e-Ali Zaidi*[a], Jia Li [b], Ngapout Yvan[c], Sara Shakir[a], Meunier Loic[d], Caroline Callot[e], William Marande[e], Hanikenne Marc[f], Stephane Rombauts [b], Yves Van de Peer [b g h], Hervé Vanderschuren§[a,c]

[a] Plant Genetics and Rhizopshere Processes Laboratory, TERRA Teaching and Research Center, Gembloux Agro-Bio Tech, University of Liège, Gembloux, Belgium
[b] Department of Plant Biotechnology and Bioinformatics, Ghent University and VIB Center for Plant Systems Biology, Ghent, Belgium
[c] Laboratory of Tropical Crop Improvement, Division of Crop Biotechnics, Biosystems Department, KU Leuven, Leuven, Belgium.
[d] InBioS, PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, Liège, Belgium
[e] CNRGV, Centre National de Ressources Génomiques Végétales, Toulouse, France
[f] InBioS, PhytoSYSTEMS, Translational Plant Biology, University of Liège, Liège, Belgium
[g] Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa.
[h] College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, China
* These authors contributed equally
§ Corresponding authors

# ABSTRACT

Cassava is an important crop for food security in the tropics where its production is jeopardized by several viral diseases, including the cassava mosaic disease (CMD) which is endemic in Sub-Saharan Africa and the Indian subcontinent. Resistance to CMD is linked to a single dominant locus, namely *CMD2*. The cassava genome contains highly repetitive regions making the accurate assembly of a reference genome challenging. In the present study, we generated BAC libraries of the CMD–susceptible cassava cultivar (cv.) 60444 and the CMD–resistant landrace TME3. We subsequently identified and sequenced BACs belonging to the *CMD2* region in both cultivars using high-accuracy long-read PacBio circular consensus sequencing (ccs) reads. We then sequenced and assembled the complete genomes of cv. 60444 and TME3 using a combination of ONT ultra-long reads and optical mapping. Anchoring the assemblies on cassava genetic maps revealed discrepancies in our, as well as in previously released, *CMD2* regions of the cv. 60444 and TME3 genomes. A BAC guided approach to assess cassava genome assemblies significantly improved the synteny between the assembled *CMD2* regions of cv. 60444 and TME3 and the *CMD2* genetic maps. We then performed repeat-unmasked gene annotation on *CMD2* assemblies and identified 81 stress resistance proteins present in the *CMD2* region, amongst which 31 were previously not reported in publicly available *CMD2* sequences.

# Introduction

Cassava (*Manihot esculenta* Crantz) is an important food security crop in the tropics (1). Despite its high yield potential (2–5) and its anticipated improved performance under climate change conditions (4, 6, 7), cassava production remains severely constrained by several biotic and abiotic stresses. In sub-Saharan Africa, two viral diseases, namely cassava mosaic disease (CMD) and cassava brown streak disease (CBSD), are widely distributed and they severely limit cassava production (8). Resistance to CMD is linked to a single dominant locus, the so-called *CMD2* locus, which was initially identified in CMD-resistant cassava landraces collected across West Africa and mapped on chromosome 8 (9–12). *CMD2*-based resistance has been extensively used to introgress CMD resistance in cassava breeding lines and released varieties (13–15). Important research programs such as NextGen Cassava are now bringing cassava breeding into a new era (16), taking advantage of High-Throughput Sequencing (HTS) technology to provide breeders and researchers with fully annotated reference cassava genomes along with single nucleotide polymorphism (SNP) information (9, 17–19). While HTS has been instrumental to generate many plant genome assemblies in a time and cost-effective manner, complex repeats and haplotype heterozygosity have remained major sources of assembly errors in released genomes (20–22). The accurate assembly of plant genomes, which can contain up to 85% of repetitive elements, remains particularly challenging (23). The recent development of long-read sequencing has opened new opportunities to improve the resolution of complex repeat-rich genomic regions (21, 24). In this context, the sequencing of ultra-long reads (ULR) using the Oxford Nanopore Technology offers new opportunities to improve the assembly of complex genomic regions as previously demonstrated for the resolution of repeat-rich regions of the human genome (25). Cassava has one of the most repetitive plant genomes as repeats are estimated to account for 61% of the total genome sequence (26). Here, we report the release of the two haplotype sequences of *CMD2* genomic region, sequenced using ULR, from two cassava genotypes contrasting for CMD resistance. A BAC-based approach was established to independently assess the quality of the assemblies in the repeat-rich *CMD2* region and to select the best assembled *CMD2* region among multiple assemblies. A comparison with the available genetic maps of the *CMD2* region shows that the *CMD2* region assembled by our ULR-based approach had a significantly better synteny and contiguity than *CMD2* regions from previously released cassava genomes. We subsequently performed gene annotation and identified the presence of additional resistance genes within our newly assembled *CMD2* regions.

# Materials and methods

## BAC sequencing

Genomic libraries and bacterial artificial chromosome (BAC) sequencing

BAC library construction and screening, and BAC clone sequencing and assembly were performed by the INRA-CNRGV. High molecular weight (HMW) genomic DNA was prepared from young frozen leaves sampled on in vitro TME3 and 60444 plantlets as described by Peterson et al. (2000) (27) and Gonthier et al. (2010) (28). Agarose embedded HMW DNA was partially digested with HindIII (New England Biolabs, Ipswich, MA, USA), subjected to two-size selection steps by pulsed-field electrophoresis using a CHEF Mapper system (Bio-Rad Laboratories, Hercules, CA, USA). DNA was eluted, ligated into the pIndigoBAC-5 HindIII-Cloning Ready vector and transformed into *Escherichia coli* electrocompetent cells. Pulsed-field migration programs, electrophoresis buffer and ligation desalting conditions were performed based on Chalhoub et al. (2004) (29). 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 José, CA, USA) using a white/blue selection. For each genotype, white colonies were arranged in 384-well microtiter plates containing LB medium with chloramphenicol (12.5 µg/mL) supplemented with 6% (v/v) glycerol (144 plates represented by 55,296 BAC clones for the TME3 genotype and 108 plates represented by 41,472 BAC clones for the 60444 genotype). The resulting libraries represent ~8.3-fold coverage of the TME3 genotype and ~7.6-fold coverage of the 60444 genotype. BAC clones were spotted on nylon membrane, screened with 27 radioactively ([α-33P]dCTP) labelled probes designed on the region of interest (Supplemental Table 1) and then analyzed with High-density filter reader program. The positive clones were verified by real-time PCR using the specific primers used for probe design. DNA were extracted from individual clones using Nucleobond Xtra midi kit (Macherey-Nagel, Düren, Nordrhein-Westfalen) and 2 µg of each individual BAC clone were used for PacBio library preparation.

Multiplexed SMRT® libraries using the standard Pacific Biosciences preparation protocol for 10 kb library with PacBio® Barcoded Adapters were prepared. Each library was then sequenced in one SMRT cell using the P6 polymerase with C4 chemistry. Sequencing was performed on a PacBio RS II sequencer. After a demultiplexing step, the sequence assembly was performed following the HGAP PacBio workflow (30), and using the SMRT® Analysis (v2.3) software suite for HGAP implementation. (https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP). BAC end sequences confirmed the position of selected clones on the cassava genome. After the first round of PacBio sequencing of selected BACs, new probes were designed at both the ends of sequenced BACs. These probes were used to screen the Mes-TM3 and Mes-60444 BAC libraries again as described above. BAC-end Sanger sequencing

was performed on the selected BACs to estimate the overlapping genomic regions between the sequenced and newly selected BACs. After several rounds of probe designing - BAC library screening - BAC-end sequencing - selection of overlapping BACs - BAC PacBio sequencing, BACs with appropriate overlapping regions were sequenced on the PacBio platform.

## Nanopore sequencing

Nanopore sequencing of cassava 60444 and TME3 genomic DNA was performed using MinION ultra-long reads as described (25). HMW DNA was extracted from young leaf tissues of 60444 and TME3 cassava, grown at a 28°C/25°C day/night rhythm with a 16 h/8 h photoperiod, following the protocol provided by Oxford Nanopore Technologies, “High molecular weight gDNA extraction from plant leaves" downloaded from the ONT Community in October, 2018; using QIAGEN Genomic tip 500/G (QIAGEN Cat. No. 10262). DNA quality was measured on 1% agarose gel and with a NanoDrop spectrophotometer. DNA was quantified using a Quantus Fluorometer (Promega). All the following steps involving handling DNA were performed with wide-bore pipet tips. To obtain ultra-long reads, the standard Rapid Adapters (RAD004) protocol (SQK-RAD004 Rapid Sequencing Kit, Oxford Nanopore Technologies (ONT), Oxford, United Kingdom) for genomic DNA was modified as described by Jain et al., 2018 (25). MinION sequencing was performed as per manufacturer's guidelines using R9 flow cells (FLO-MIN106D, ONT) using a MinION sequencer Mk1B. Twelve flow cells were used for 60444 ULR sequencing and seven for TME3; these flow cells generated 39,525 MB (6,221,230 reads) and 41,340 MB (6,549,320 reads) sequencing data for 60444 and TME3, respectively, with 22,782 reads longer than 8 kb.

## Bionano sequencing

### Ultra High Molecular Weight DNA extraction

To generate the optical map, Ultra High Molecular Weight (UHMW) DNA were purified from 1g of fresh dark treated very young leaves according to the Bionano Prep Plant Tissue DNA Isolation Base Protocol (30068 - Bionano Genomics) with the following specifications and modifications. Briefly, the leaves were fixed in fixing buffer containing formaldehyde. After 3 washes, leaves were cut in 2mm pieces and disrupt with rotor stator in homogenization buffer containing spermine, spermidine and beta-mercaptoethanol. Nuclei were washed, purified using a density gradient and then embedded in agarose plugs. After overnight proteinase K digestion in Lysis Buffer (BNG) and 1-hour treatment with RNAse A (Qiagen, MD,USA), plugs were washed 4 times in 1×Wash Buffer (BNG) and 5 times in 1× TE Buffer (ThermoFisher Scientific, Waltham, MA). Then, plugs were melted for 2 minutes at 70◦C and solubilized with 2 µL of 0.5 U/µL AGARase enzyme (ThermoFisher Scientific, Waltham, MA) for 45

minutes at 43◦C. A dialysis step was performed in 1× TE Buffer (ThermoFisher Scientific, Waltham, MA) for 45 minutes to purify DNA from any residues. The DNA samples were quantified by using the Qubit dsDNA BR Assay (Invitrogen, Carlsbad, CA, USA). Quality of megabase-size DNA was validated by pulsed-field gel electrophoresis.

### Data collection and optical map construction

Labelling and staining of the UHMW DNA were performed according to the Direct Label and Stain (DLS) protocol (BNG). Briefly, labelling was performed by incubating 750 ng genomic DNA with 1× DLE-1 Enzyme (BNG) for 2 hours in the presence of 1× DL-Green (BNG) and 1× DLE-1 Buffer (BNG). Following proteinase K digestion and DL-Green clean-up, the DNA backbone was stained by mixing the labelled DNA with DNA Stain solution (BNG) in the presence of 1×Flow Buffer (BNG) and 1× DTT (BNG), and incubating overnight at room temperature. The DLS DNA concentration was measured with the Qubit dsDNA HS Assay (Invitrogen, Carlsbad, CA, USA).

Labelled and stained DNA was loaded on Saphyr chips. Loading of the chips and running of the BNG Saphyr System were all performed according to the Saphyr System User Guide. Digitalized labelled DNA molecules were assembled to optical maps using the BNG Access software.

## Genome assembly and BAC mapping

The assembly process was optimized with the reads from cv. 60444, using three different assemblers. *NanoFilt (31)* was used to generate a dataset with long reads (> 8kb). *CANU* V2.3 (32), with the options stopOnLowCoverage=5 and cnsErrorRate=0.25, was used with all reads and on the >8kb read dataset. *Flye* V2.19.b1774 (33, 34), with default settings, was used with the same two datasets. *wtdbg2* V2.1 (35) was used on all reads with default settings and read length option set to 300, 2000, 3000, 5000, 8000, 10000. All final assemblies were polished using illumina short reads, downloaded from the NCBI (SRX1393211 and SRX526747), with *pilon* V1.24 (36), using default settings, after mapping of the reads with *bwa mem* V0.7.17 (37) and *samtools* V1.13 (38).

The BAC sequences were mapped on the genomes using *Blasr* V5.1 (39). *Blasr,* whose initial purpose is the mapping of corrected PacBio reads (39), was used on the full BAC sequences in fasta format, with default settings. Only the longest alignment per BAC was used to compute the percentage of BAC mapping on the genome. The same approach was used to compare public cassava genomes with our best assemblies. The results of BAC mapping are available in Supplemental Table 2. The quality of the assemblies was assessed with *BUSCO* V5.3.0 (40), using auto-lineage settings. The BAC mapping showed that the *CMD2* region was better resolved using Flye with default settings and reads longer than 8kb. Based on BUSCO scores, the

best assembly of cv. 60444 was generated with CANU (32) using all reads (Supplemental Table 2). However, this assembly, better in completeness, was outperformed by the one generated with Flye using reads above 8kb when assessed by BAC mapping.

The TME3 reads were filtered with *NanoFilt* (31) to generate the >8kb read dataset. The assembly of TME3 reads was done using the best approaches as previously assessed by BAC mapping on cv.60444 assemblies (i.e. *Flye* with default settings using all reads and reads longer than 8kb). The selected TME3 assembly was polished using Pilon V1.24 (36). The BAC mapping was performed by *Blasr* (39).

Bionano optical mapping was used to scaffold the selected cv.60444 and TME3 genomes (i.e. *Flye* with default setting using > 8kb reads). Genomic statistics on these final genomes, and on public genomes, were computed using QUAST (41), with default settings. *BUSCO* V5.3.0 (40), with auto-lineage settings, was used to estimate completeness and duplication (based on duplication value reported by BUSCO of its single gene marker set).

## Markers mapping

The 64 genetic markers from Rabbi et al (10) were mapped on the *CMD2* haplotypes produced in this study and previously available in public databases (ref) using blastn V2.10.0 (42) with an e-value cut off of 10e-3. The top hit was taken to determine the position of each marker on the chromosome.

The *CMD2* locus were mapped using blastn (BLAST 2.9.0 +) searches of 101 markers on several genomes of cassava (60444 H1 & H2 GCA_963409065.1 Cornet, TME3 H1 & H2 GCA_963409055.1 Cornet, TME3 GCA_003957995.1 Kuon, and 60444 GCA_003957885.1 Kuon). These 101 markers (Supplemental Table 6) included 6 classical markers (RFLP and SSR markers) published by Akano et al. 2002 (43), Lokko et al. 2005 (44), Okogbenin et al. 2007 (45) and Okogbenin et al. 2012 (46), 31 SNP markers on chromosome 12 from Rabbi et al. 2022 (12) (The 31 markers were selected by taking the SNP with the highest -log p value plus 15 markers upstream stream and 15 markers downstream of this peak SNP marker) and 64 other SNP markers published by Wolfe et al. 2016 (11).

## Genome annotation

### Repetitive elements and noncoding RNAs annotation

We used two strategies to predict the repeat sequences in the two genotypes. The first strategy was *ab initio* prediction. RepeatModeler v2.0 (http://www.repeatmasker.org/RepeatModeler/) and LTR_FINDER v1.0.7 (47) were employed to generate custom libraries. Then, the assembled sequences of these two

genotypes were mapped against the libraries, respectively, to generate the *ab initio* annotation results. The second strategy was an homology-based prediction. The assembled sequences were aligned to the RepBase v23.05 ((http://www.girinst.org/repbase/) by using RepeatMasker v4.1.0 (48). Then the information from the above two methods was integrated into non-redundant results.

Transfer RNAs (tRNAs) were predicted by tRNAscan-SE v1.43 (49) with default parameters. To predict ribosomal RNAs (rRNAs), the genome assemblies were aligned against the RNA families (Rfam) v14.1 database (50) by the Blastn program (42).

## Gene annotation

Three different strategies were used to predict the gene set of the two cassava genotypes:

*Homology annotation*. We aligned the protein sequences from five published genomes, including *Manihot esculenta* (AM560) (17), *Ricinus communis* (https://bioinformatics.psb.ugent.be/plaza/versions/plaza_v4_5_dicots/), and *Jatropha curcas* (NCBI accession: GCA_000696525.1), respectively, against our assemblies to predict genes based on homology. The potential homology-based genes were searched by GeMoMa v1.7 (51).

*RNA-seq annotation*. We removed the redundancies in a public Iso-seq cassava dataset (52) using Cupcake v12.1.0 (https://github.com/Magdoll/cDNA_Cupcake/) and got the unique isoforms that were used as input for the Assemble Spliced Alignments (PASA) pipeline v2.4.1 (53). For Illumina RNA-seq data, SRR25338832, Trinity (54) was used to assemble the data. Then we used PASA to identify the potential gene structures.

*de novo annotation*. In order to discard pseudo gene predictions, we used “N” to substitute the repetitive elements of the assembled sequences. Then we extracted the complete genes with multiple exons and start/stop codons from the predicted genes of homologous annotation and RNA-seq annotation to build the training set for the do novo gene predictors. We identified 7,553 complete genes in cv. 60444 and 7,394 complete genes in TME3 to construct the training set, respectively. Subsequently, AUGUSTUS v3.3.3 (55), SNAP v2006-07-28 (56), and GlimmerHMM v3.0.4 (57) were trained based on the training set and then operated to identify the potential gene models. GeneMark-ES v4.57_lic (58) were executed to predict genes with default parameter.

For the integration of the annotation results, we employed the Evidencemodeler v1.1.1 (59) to generate non-redundant and comprehensive gene sets. After that, we used the PASA pipeline again to correct the potential errors to generate the final gene sets for the two genotypes. Then we used BUSCO v5.0.0 (40) and the embryophyta_odb10 dataset to assess the quality of the predicted gene sets.

Functional annotation of the predicted genes was operated by running BlastP (60) setting an e-value cut-off of 1e05 against the public protein function databases Uniprot/SwissProt (61) and NCBI NR (62) (RefSeq non- redundant protein record). Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms were classified using eggNOG v2.1.6 (63).

## Resistance gene

The *CMD2* SNP markers of Rabbi et al.,2014 (10) and Rabbi et al 2022 (12), were used to define the *CMD2* region on our haplotypes. All proteins of this region were then analyzed with prPred (64) to predict the probability that these proteins contributed to biotic stress resistance. 81 Proteins with a probability above 0.1 were retained for both haplotypes of TME3.

The 81 proteins were searched in the cv. 60444 genome (GCA_963409065.1 Cornet), the NCBI TME3 genome (GCA_003957885.1 Kuon), and IsoSeq data (personal communication from Kuon et al., 2018 (52) of the two genotypes by orthologous enrichment with the program Forty-two V0.1416 (65, 66). The generated orthologous group (OGs) files were aligned with mafft (67), with default settings. The alignments were then cleaned of mis-predicted stretches with HMMCleaner (68), using leave-on-out profiles, with c1=0.40 and default weights for c2 to c4. Individual phylogenetic trees were produced with Phyml v3.1, through seaview (69), and subtrees with orthologous sequences were manually created. The presence in genomes and Isoseq data were determined based on presence in these subtrees. The 81 proteins were also analyzed by Interproscan V5.48-83 (70).

# Results and Discussion

In order to optimize the assembly of the cassava genomes, we implemented an assessment by Bacterial Artificial Chromosome (BAC) mapping. We first generated BAC libraries of cassava cv. 60444 and TME3, named Mes-60444 and Mes-TME3, that contained 41,472 clones (108 384-wells microplates) and 55,296 clones (144 384-wells microplates), respectively. We then screened BAC libraries using high-density filter hybridization with radioactively labeled probes. *CMD2*-specific probes were designed based on simple-sequence repeat (SSR) (43–45, 71) and genome-wide SNP markers (10, 11) (Supplementary table 1). After several rounds of probe design - BAC library screening - BAC-end Sanger sequencing - selection of overlapping BACs - BAC PacBio sequencing, we identified 13 and 16 BACs of cv. 60444 and TME3 covering the *CMD2* region, respectively. These BAC sequences from repeat-rich regions were subsequently used to assess the accuracy of ULR-based genome assemblies and provided a powerful parameter to select the best performing assembler (**Figure 1A**).

Ultra-Long Read (ULR) sequencing (25) of HMW cassava DNA on MinION flow cells generated 39,525 Mb (6,221,230 reads) and 41,340 Mb (6,549,320 reads) sequencing data for 60444 and TME3, respectively, with 22,782 reads above 8 kb of length. Three multiple long read assemblers were used to perform the assemblies and assess the impact of increasing read lengths on the quality of the *CMD2* region assembly: *wtdbg2* (35) with 6 different minimal read lengths from 0,3 to 10 kb, *CANU (32)* with all reads and only reads longer than 8kb, and *Flye* (33, 34) with reads longer than 8kb. We implemented a method computing the contiguity of BAC sequences in the assembled *CMD2* regions to determine the quality of the assembled *CMD2*. Each assembly was thus assessed by BAC mapping. The contiguity values ranged from 44,9 % to 81,45 % of BAC mapping on the assemblies (Supplemental Table 2), showing an important variation depending on the read length and the assembler used. Our results indicated that mapping of high-accuracy BACs to the assembled genome could be used as a parameter to assess the quality of the assembled CMD2 regions. Assemblies generated using Flye (33, 34) and a cut-off threshold for reads smaller than 8kb displayed the best contiguity with BAC sequences despite a decrease in genome coverage (Supplemental Table 2). These assemblies mapped 81,45% and 80,2 % of contiguous BACs sequence, for cv. 60444 (GCA_963409065.1 Cornet) and TME3 (GCA_963409055.1 Cornet), respectively (Figure 2), against 73,15% and 61,08% for the previously released cassava genomes of cv. 60444 (GCA_003957885.1 Kuon) and TME3 (GCA_003957995.1 Kuon) respectively (52). To assess the quality of the *CMD2* region with an independent approach, we took advantage of the publicly available genetic markers for the *CMD2* region (10, 11). Markers indicated the presence of two haplotypes of the CMD2 region in our two genome assemblies (Figure 1B). The length of the *CMD2* region (Figure 1B) in these haplotypes varied moderately when compared to the length of *CMD2* (1,88 Mb) estimated on marker mapping reported by Rabbi and colleagues (10): 2,18 Mb (1,16x) and 2,31Mb (1,23x) for cv. 60444 (GCA_963409065.1 Cornet) and 2,22 Mb (1,18x) and 2,25 Mb (1,19x) for TME3 (GCA_963409055.1 Cornet). Importantly, our assemblies had the same marker order as in the genetic map (Figure 1B). When analyzing the markers in the *CMD2* region from the public released cassava genomes (52), we noticed they contain a single genomic region for *CMD2* in which the two haplotypes are merged, with a consistent increase of the theoretical length calculated by genetic mapping (10) [i.e. 5,95 Mb (3,16x) for 60444 (GCA_003957885.1 Kuon) and 5,08 Mb (2,70x) for TME3 (GCA_003957995.1 Kuon)]. Moreover, the *CMD2* region of the publicly released genomes contained 48 inversion events as compared to the *CMD2* genetic map (Figure 1B).

Our assembly process, which was optimized for the contiguity of the *CMD2* region, also generated whole-genome assemblies. Although our genome assemblies produced from ULR displayed a more accurate assembly and contiguity for the *CMD2* region, the completeness scores (72,9 % for 60444, 95,8% for TME3) as indicated by BUSCO (40) (Supplemental Table 3), remained lower than previously released public assemblies. A further characterization of the assemblies revealed that assemblies with higher completeness had been generated by our BAC guided approach, but they were

not selected due to their lower BACs mapping percentage. For instance, the assembly of cv. 60444 with CANU (32) using all reads had a completeness score of 98% but a BAC mapping score of 67,46 % (Supplemental Table 2). While our results clearly indicate that ONT reads used with BAC-guided assessment of the assemblies provide a significant improvement of genomic regions with high levels of repetitive elements, the use of long reads with high base-level resolution (72) could help achieving both high completeness of the genome and high accuracy in repeat-rich regions.

*CMD2*-based resistance has recently been shown to co-segregate with a single nucleotide polymorphism in the *MePOLD1* gene (73). Noticeably an analysis focusing on 101 cassava markers associated with the *CMD2*-based resistance indicated that the *MePOLD1* gene was not located in the *CMD2* region where the majority of the markers mapped (Figure 1C, Supplemental Figures 1 & 2). Previous breeding work demonstrated that *CMD2*-based resistance is linked to a single dominant locus (10, 11, 43). Annotation of our two newly produced *CMD2* haplotypes from the resistant landrace (TME3) indicated the presence of 79 genes for *CMD2* region of contig 100004 and 78 genes for *CMD2* region of contig 100036, respectively (same id used in Figure 1B. In the cv. 60444, the *CMD2* haplotypes contained 90 genes for contig 4632 and 93 genes for contig 100051, respectively). Among the 157 genes from the *CMD2* region of TME3, 81 genes had a non-null probability of contributing to CMD2 resistance (see methods). The 81 genes were then searched for in the public 60444 (GCA_003957885.1 Kuon) and TME3 (GCA_003957995.1 Kuon) genomes from the NCBI (Supplemental Table 4). Thanks to the better resolution of our *CMD2* region assemblies, 31 additional genes (out of 81) were present in our *CMD2* haplotypes but were absent in the publicly available TME3 (GCA_003957995.1 Kuon) genome. Twenty-one of these 31 genes, present on TME3 (GCA_963409055.1 Cornet), did not have reported expression in the cv. 60444 Isoseq dataset (52) (Supplemental Table 4) and might represent additional candidate genes for virus resistance potentially complementary to the previously reported *MePOLD1* gene associated with CMD resistance. We analyzed the expression profile of the 81 genes in a public IsoSeq dataset for cv. 60444 and TME3 (52). Nine genes, 3 of them absent from the TME3 (GCA_003957995.1 Kuon) genome, were expressed in TME3 and non-expressed or absent in 60444 (Figure 1C). Among these genes, 5 have been reported to be associated with biotic stress resistance, including mosaic virus, in other plants such as Arabidopsis or tobacco (Figure 1C; Supplemental table 5). Additional wet lab experiments would be required to assess the possibility of multiple biotic stress resistance genes contributing to CMD2 resistance. Because the IsoSeq transcriptome profiling was not performed in virus-infected plants, we cannot exclude that some of the 81 resistance genes reported here would have a better significance under viral conditions than the one highlighted above. Indeed, currently, nine genes have an expression in TME3 IsoSeq data with a non-expression in 60444 IsoSeq data, experiment under viral conditions might might provide more insight into the link with CMD2 resistance of these 81 genes. Our BAC-assessed approach to assemble

complex genomes and genomic regions provides a framework to improve the plant genome assemblies that have been released so far. The BAC-guided assembly of cassava genomes has significantly improved the contiguity and accuracy of the repeat-rich *CMD2* leading to the identification of additional sequences potentially relevant for virus resistance. A better resolution of the *CMD2* haplotypes will pave the way to a better understanding of resistance to cassava mosaic disease.

# Data availability

L Cornet and H Vanderschuren have summitted raw data to the National Center for Biotechnology Information (NCBI) under the Bioproject accession no. PRJNA981703, Raw nanopore reads no. SRR25338822 for 60444 and no. SRR25338821 for TME3, bionano omptical maps accession no. SUPPF_0000005490 fo 60444 and no. SUPPF_0000005489 for TME3, RNAseq reads accession no. SRR25338832. L Cornet and H Vanderschuren have summitted genome assemblies to European Nucleotide Archive (ENA) under the Bioproject PRJEB65447, genome accession GCA_963409065.1 for 60444 and GCA_963409055.1 for TME3. CMD2 regions and corresponding proteins are available in the figshare repository https://figshare.com/projects/CMD2_in_cassava/177291.

# Funding

The authors acknowledge financial support from the Belgian FNRS grant M.i.S. F.4515.17 to H.V. and grant 1.B456.20 to S.S.Z. and H.V. YVdP acknowledges funding from the European Research Council (ERC) under the European Union's 501 Horizon 2020 research and innovation program (No. 833522) and from Ghent University (Methusalem 502 funding, BOF.MET.2021.0005.01).

# Acknowledgment

We thank David Colignon and the CÉCI for help with computing cluster usage.

# Authors contribution

CL, SZ, JL, NY, YVDP, HV conceived the study and wrote the initial manuscript. SS, SZ, LC produced ONT data. SS, SZ produced BACs. LC & LM ran bioinformatics analysis. LC, NY made figures. CC and WM produced optical maps and corrected

genomes. JL, SR, YVDP made protein predictions. LC ran resistance protein analysis. All authors read and accepted the final manuscript.

# References

1. Sayre,R., Beeching,J.R., Cahoon,E.B., Egesi,C., Fauquet,C., Fellman,J., Fregene,M., Gruissem,W., Mallowa,S., Manary,M., *et al.* (2011) The BioCassava Plus Program: Biofortification of Cassava for Sub-Saharan Africa. *Annual Review of Plant Biology*, **62**, 251–272.

2. Fermont,A.M., van Asten,P.J.A., Tittonell,P., van Wijk,M.T. and Giller,K.E. (2009) Closing the cassava yield gap: An analysis from smallholder farms in East Africa. *Field Crops Research*, **112**, 24–36.

3. Kreye,C., Hauser,S., Pypers,P. and Vanlauwe,B. (2020) Intensification options of small holders' cassava production in South-west Nigeria. *Agronomy Journal*, **112**, 5312–5324.

4. Moreno-Cadena,P., Hoogenboom,G., Cock,J.H., Ramirez-Villegas,J., Pypers,P., Kreye,C., Tariku,M., Ezui,K.S., Becerra Lopez-Lavalle,L.A. and Asseng,S. (2021) Modeling growth, development and yield of cassava: A review. *Field Crops Research*, **267**, 108140.

5. El-Sharkawy,M.A. (2004) Cassava biology and physiology. *Plant Mol Biol*, **56**, 481–501.

6. Lobell,D.B., Burke,M.B., Tebaldi,C., Mastrandrea,M.D., Falcon,W.P. and Naylor,R.L. (2008) Prioritizing Climate Change Adaptation Needs for Food Security in 2030. *Science*, **319**, 607–610.

7. Rosenthal,D.M., Slattery,R.A., Miller,R.E., Grennan,A.K., Cavagnaro,T.R., Fauquet,C.M., Gleadow,R.M. and Ort,D.R. (2012) Cassava about-FACE: Greater than expected yield stimulation of cassava (Manihot esculenta) by future CO2 levels. *Global Change Biology*, **18**, 2661–2675.

8. Rey,C. and Vanderschuren,H. (2017) Cassava Mosaic and Brown Streak Diseases: Current Perspectives and Beyond. *Annual Review of Virology*, **4**, 429–452.

9. Ramu,P., Esuma,W., Kawuki,R., Rabbi,I.Y., Egesi,C., Bredeson,J.V., Bart,R.S., Verma,J., Buckler,E.S. and Lu,F. (2017) Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nature Genetics*, **49**, 959–963.

10. Rabbi,I.Y., Hamblin,M.T., Kumar,P.L., Gedil,M.A., Ikpan,A.S., Jannink,J.-L. and Kulakow,P.A. (2014) High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding. *Virus Research*, **186**, 87–96.

11. Wolfe,M.D., Rabbi,I.Y., Egesi,C., Hamblin,M., Kawuki,R., Kulakow,P., Lozano,R., Carpio,D.P.D., Ramu,P. and Jannink,J.-L. (2016) Genome-Wide Association and Prediction Reveals Genetic Architecture of Cassava Mosaic Disease Resistance and Prospects for Rapid Genetic Improvement. *The Plant Genome*, **9**, plantgenome2015.11.0118.

12. Rabbi,I.Y., Kayondo,S.I., Bauchet,G., Yusuf,M., Aghogho,C.I., Ogunpaimo,K., Uwugiaren,R., Smith,I.A., Peteti,P., Agbona,A., *et al.* (2022) Genome-wide association analysis reveals new insights into the genetic architecture of defensive, agro-morphological and quality-related traits in cassava. *Plant Mol Biol*, **109**, 195–213.

13. Malik,A.I., Kongsil,P., Nguyễn,V.A., Ou,W., Sholihin, Srean,P., Sheela,M.N., López-Lavalle,L.A.B., Utsumi,Y., Lu,C., *et al.* (2020) Cassava breeding and agronomy in Asia: 50 years of history and future directions. *Breeding Science*, **70**, 145–166.

14. Legg,J.P., Lava Kumar,P., Makeshkumar,T., Tripathi,L., Ferguson,M., Kanju,E., Ntawuruhunga,P. and Cuellar,W. (2015) Chapter Four - Cassava Virus Diseases: Biology, Epidemiology, and Management. In Loebenstein,G., Katis,N.I. (eds), *Advances in Virus Research*, Control of Plant Virus Diseases. Academic Press, Vol. 91, pp. 85–142.

15. Ige,A.D., Olasanmi,B., Mbanjo,E.G.N., Kayondo,I.S., Parkes,E.Y., Kulakow,P., Egesi,C., Bauchet,G.J., Ng,E., Lopez-Lavalle,L.A.B., *et al.* (2021) Conversion and Validation of Uniplex SNP Markers for Selection of Resistance to Cassava Mosaic Disease in Cassava Breeding Programs. *Agronomy*, **11**, 420.

16. Maxmen,A. (2019) How African scientists are improving cassava to help feed the world. *Nature*, **565**, 144–146.

17. Bredeson,J.V., Lyons,J.B., Prochnik,S.E., Wu,G.A., Ha,C.M., Edsinger-Gonzales,E., Grimwood,J., Schmutz,J., Rabbi,I.Y., Egesi,C., *et al.* (2016) Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature Biotechnology*, **34**, 562–570.

18. Prochnik,S., Marri,P.R., Desany,B., Rabinowicz,P.D., Kodira,C., Mohiuddin,M., Rodriguez,F., Fauquet,C., Tohme,J., Harkins,T., *et al.* (2012) The Cassava Genome: Current Progress, Future Directions. *Tropical Plant Biol.*, **5**, 88–94.

19. Qi,W., Lim,Y.-W., Patrignani,A., Schläpfer,P., Bratus-Neuenschwander,A., Grüter,S., Chanez,C., Rodde,N., Prat,E., Vautrin,S., *et al.* (2022) The haplotype-resolved chromosome pairs of a heterozygous diploid African cassava cultivar reveal novel pan-genome and allele-specific transcriptome features. *GigaScience*, **11**, giac028.

20. Rhie,A., McCarthy,S.A., Fedrigo,O., Damas,J., Formenti,G., Koren,S., Uliano-Silva,M., Chow,W., Fungtammasan,A., Kim,J., *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, **592**, 737–746.

21. Tørresen,O.K., Star,B., Mier,P., Andrade-Navarro,M.A., Bateman,A., Jarnot,P., Gruca,A., Grynberg,M., Kajava,A.V., Promponas,V.J., *et al.* (2019) Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, **47**, 10994–11006.

22. Koren,S. and Phillippy,A.M. (2015) One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, **23**, 110–120.

23. Kress,W.J., Soltis,D.E., Kersey,P.J., Wegrzyn,J.L., Leebens-Mack,J.H., Gostel,M.R., Liu,X. and Soltis,P.S. (2022) Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proceedings of the National Academy of Sciences*, **119**, e2115640118.

24. Amarasinghe,S.L., Su,S., Dong,X., Zappia,L., Ritchie,M.E. and Gouil,Q. (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, **21**, 30.

25. Jain,M., Koren,S., Miga,K.H., Quick,J., Rand,A.C., Sasani,T.A., Tyson,J.R., Beggs,A.D., Dilthey,A.T., Fiddes,I.T., *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, **36**, 338–345.

26. Mansfeld,B.N., Boyher,A., Berry,J.C., Wilson,M., Ou,S., Polydore,S., Michael,T.P., Fahlgren,N. and Bart,R.S. (2021) Large structural variations in the haplotype-resolved African cassava genome. *The Plant Journal*, **108**, 1830–1848.

27. Peterson,D.G., Tomkins,J.P., Frisch,D.A. and Paterson,A.H. CONSTRUCTION OF PLANT BACTERIAL ARTIFICIAL CHROMOSOME (BAC) LIBRARIES: AN ILLUSTRATED GUIDE.

28. Gonthier,L., Bellec,A., Blassiau,C., Prat,E., Helmstetter,N., Rambaud,C., Huss,B., Hendriks,T., Bergès,H. and Quillet,M.-C. (2010) Construction and characterization of two BAC libraries representing a deep-coverage of the genome of chicory (Cichorium intybus L., Asteraceae). *BMC Research Notes*, **3**, 225.

29. Chalhoub,B., Belcram,H. and Caboche,M. (2004) Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size. *Plant Biotechnology Journal*, **2**, 181–188.

30. Chin,C.-S., Alexander,D.H., Marks,P., Klammer,A.A., Drake,J., Heiner,C., Clum,A.,

Copeland,A., Huddleston,J., Eichler,E.E., *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, **10**, 563–569.

31. De Coster,W., D'Hert,S., Schultz,D.T., Cruts,M. and Van Broeckhoven,C. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, **34**, 2666–2669.

32. Koren,S., Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.

33. Kolmogorov,M., Bickhart,D.M., Behsaz,B., Gurevich,A., Rayko,M., Shin,S.B., Kuhn,K., Yuan,J., Polevikov,E., Smith,T.P.L., *et al.* (2020) metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, **17**, 1103–1110.

34. Kolmogorov,M., Yuan,J., Lin,Y. and Pevzner,P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, **37**, 540–546.

35. Ruan,J. and Li,H. (2020) Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, **17**, 155–158.

36. Walker,B.J., Abeel,T., Shea,T., Priest,M., Abouelliel,A., Sakthikumar,S., Cuomo,C.A., Zeng,Q., Wortman,J., Young,S.K., *et al.* (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, **9**, e112963.

37. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589–595.

38. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

39. Chaisson,M.J. and Tesler,G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**, 238.

40. Manni,M., Berkeley,M.R., Seppey,M., Simao,F.A. and Zdobnov,E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *arXiv:2106.11799 [q-bio]*.

41. Gurevich,A., Saveliev,V., Vyahhi,N. and Tesler,G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

42. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

43. Akano,A., Dixon,A., Mba,C., Barrera,E. and Fregene,M. (2002) Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. *Theor Appl Genet*, **105**, 521–525.

44. Lokko,Y., Danquah,E.Y., Offei,S.K., Dixon,A.G.O. and Gedil,M.A. (2005) Molecular markers associated with a new source of resistance to the cassava mosaic disease. *African Journal of Biotechnology*, **4**.

45. Okogbenin,E., Porto,M. c. m., Egesi,C., Mba,C., Espinosa,E., Santos,L. g., Ospina,C., Marín,J., Barrera,E., Gutiérrez,J., *et al.* (2007) Marker-Assisted Introgression of Resistance to Cassava Mosaic Disease into Latin American Germplasm for the Genetic Improvement of Cassava in Africa. *Crop Science*, **47**, 1895–1904.

46. Okogbenin,E., Egesi,C.N., Olasanmi,B., Ogundapo,O., Kahya,S., Hurtado,P., Marin,J., Akinbo,O., Mba,C., Gomez,H., *et al.* (2012) Molecular Marker Analysis and Validation of Resistance to Cassava Mosaic Disease in Elite Cassava Genotypes in Nigeria. *Crop Science*, **52**, 2576–2586.

47. Xu,Z. and Wang,H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, **35**, W265–W268.

48. Tarailo-Graovac,M. and Chen,N. (2009) Using RepeatMasker to Identify Repetitive

Elements in Genomic Sequences. *Current Protocols in Bioinformatics*, **25**, 4.10.1-4.10.14.

49. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research*, **25**, 955–964.

50. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Research*, **36**, D154–D158.

51. Keilwagen,J., Hartung,F. and Grau,J. (2019) GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. In Kollmar,M. (ed), *Gene Prediction: Methods and Protocols*, Methods in Molecular Biology. Springer, New York, NY, pp. 161–177.

52. Kuon,J.-E., Qi,W., Schläpfer,P., Hirsch-Hoffmann,M., von Bieberstein,P.R., Patrignani,A., Poveda,L., Grob,S., Keller,M., Shimizu-Inatsugi,R., *et al.* (2019) Haplotype-resolved genomes of geminivirus-resistant and geminivirus-susceptible African cassava cultivars. *BMC Biology*, **17**, 75.

53. Haas,B.J., Delcher,A.L., Mount,S.M., Wortman,J.R., Smith,R.K., Hannick,L.I., Maiti,R., Ronning,C.M., Rusch,D.B., Town,C.D., *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*, **31**, 5654–5666.

54. Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L., Raychowdhury,R., Zeng,Q., *et al.* (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol*, **29**, 644–652.

55. Stanke,M., Keller,O., Gunduz,I., Hayes,A., Waack,S. and Morgenstern,B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*, **34**, W435-439.

56. Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

57. Majoros,W.H., Pertea,M. and Salzberg,S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.

58. Borodovsky,M. and Lomsadze,A. (2011) Gene Identification in Prokaryotic Genomes, Phages, Metagenomes, and EST Sequences with GeneMarkS Suite. *Current Protocols in Bioinformatics*, **35**, 4.5.1-4.5.17.

59. Haas,B.J., Salzberg,S.L., Zhu,W., Pertea,M., Allen,J.E., Orvis,J., White,O., Buell,C.R. and Wortman,J.R. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*, **9**, R7.

60. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

61. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.

62. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D., *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, **44**, D733–D745.

63. Huerta-Cepas,J., Forslund,K., Coelho,L.P., Szklarczyk,D., Jensen,L.J., von Mering,C. and Bork,P. (2017) Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, **34**, 2115–2122.

64. Wang,Y., Wang,P., Guo,Y., Huang,S., Chen,Y. and Xu,L. (2021) prPred: A Predictor to Identify Plant Resistance Proteins by Incorporating k-Spaced Amino Acid (Group) Pairs. *Frontiers in Bioengineering and Biotechnology*, **8**.

65. Irisarri,I., Baurain,D., Brinkmann,H., Delsuc,F., Sire,J.-Y., Kupfer,A., Petersen,J., Jarek,M., Meyer,A., Vences,M., *et al.* (2017) Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat Ecol Evol*, **1**, 1370–1378.

66. Simion,P., Philippe,H., Baurain,D., Jager,M., Richter,D.J., Di Franco,A., Roure,B., Satoh,N., Quéinnec,É., Ereskovsky,A., *et al.* (2017) A Large and Consistent Phylogenomic

Dataset Supports Sponges as the Sister Group to All Other Animals. *Current Biology*, **27**, 958–967.

67. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

68. Di Franco,A., Poujol,R., Baurain,D. and Philippe,H. (2019) Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol Biol*, **19**, 21.

69. Gouy,M., Guindon,S. and Gascuel,O. (2010) SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution*, **27**, 221–224.

70. Mulder,N. and Apweiler,R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol*, **396**, 59–70.

71. Mohan,C., Shanmugasundaram,P., Maheswaran,M., Senthil,N., Raghu,D. and Unnikrishnan,M. (2013) Mapping New Genetic Markers Associated with CMD Resistance in Cassava (Manihot esculenta Crantz) Using Simple Sequence Repeat Markers. *JAS*, **5**, p57.

72. Hon,T., Mars,K., Young,G., Tsai,Y.-C., Karalius,J.W., Landolin,J.M., Maurer,N., Kudrna,D., Hardigan,M.A., Steiner,C.C., *et al.* (2020) Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data*, **7**, 399.

73. Lim,Y.-W., Mansfeld,B.N., Schläpfer,P., Gilbert,K.B., Narayanan,N.N., Qi,W., Wang,Q., Zhong,Z., Boyher,A., Gehan,J., *et al.* (2022) Mutations in DNA polymerase δ subunit 1 co-segregate with CMD2-type resistance to Cassava Mosaic Geminiviruses. *Nat Commun*, **13**, 3933.

74. Zhu,F., Zhang,Q.-P., Che,Y.-P., Zhu,P.-X., Zhang,Q.-Q. and Ji,Z.-L. (2021) Glutathione contributes to resistance responses to TMV through a differential modulation of salicylic acid and reactive oxygen species. *Molecular Plant Pathology*, **22**, 1668–1687.

75. Koch,A., Kang,H.-G., Steinbrenner,J., Dempsey,D.A., Klessig,D.F. and Kogel,K.-H. (2017) MORC Proteins: Novel Players in Plant and Animal Health. *Frontiers in Plant Science*, **8**.

76. Pandian,B.A., Sathishraj,R., Djanaguiraman,M., Prasad,P.V.V. and Jugulam,M. (2020) Role of Cytochrome P450 Enzymes in Plant Stress Response. *Antioxidants*, **9**, 454.

77. Yuan,W., Jiang,T., Du,K., Chen,H., Cao,Y., Xie,J., Li,M., Carr,J.P., Wu,B., Fan,Z., *et al.* (2019) Maize phenylalanine ammonia-lyases contribute to resistance to Sugarcane mosaic virus infection, most likely through positive regulation of salicylic acid accumulation. *Molecular Plant Pathology*, **20**, 1365–1378.

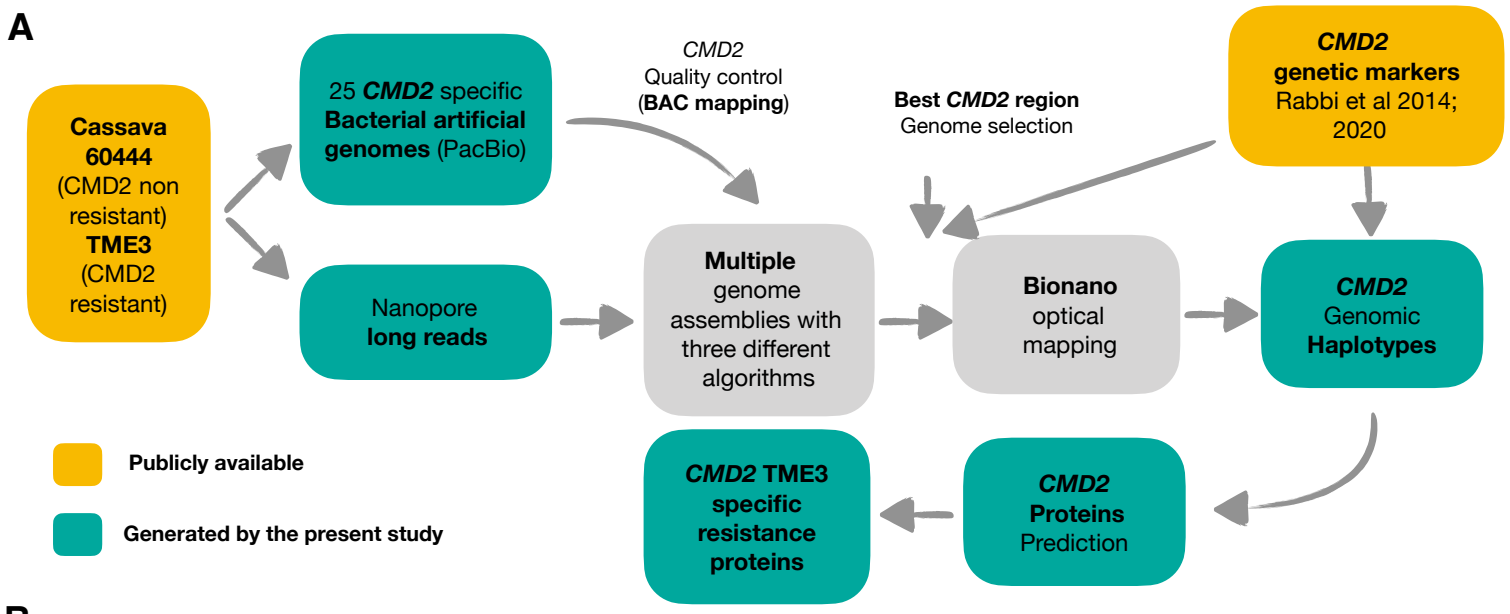
78. Chong,J., Baltz,R., Schmitt,C., Beffa,R., Fritig,B. and Saindrenan,P. (2002) Downregulation of a Pathogen-Responsive Tobacco UDP-Glc:Phenylpropanoid Glucosyltransferase Reduces Scopoletin Glucoside Accumulation, Enhances Oxidative Stress, and Weakens Virus Resistance. *The Plant Cell*, **14**, 1093–1107.

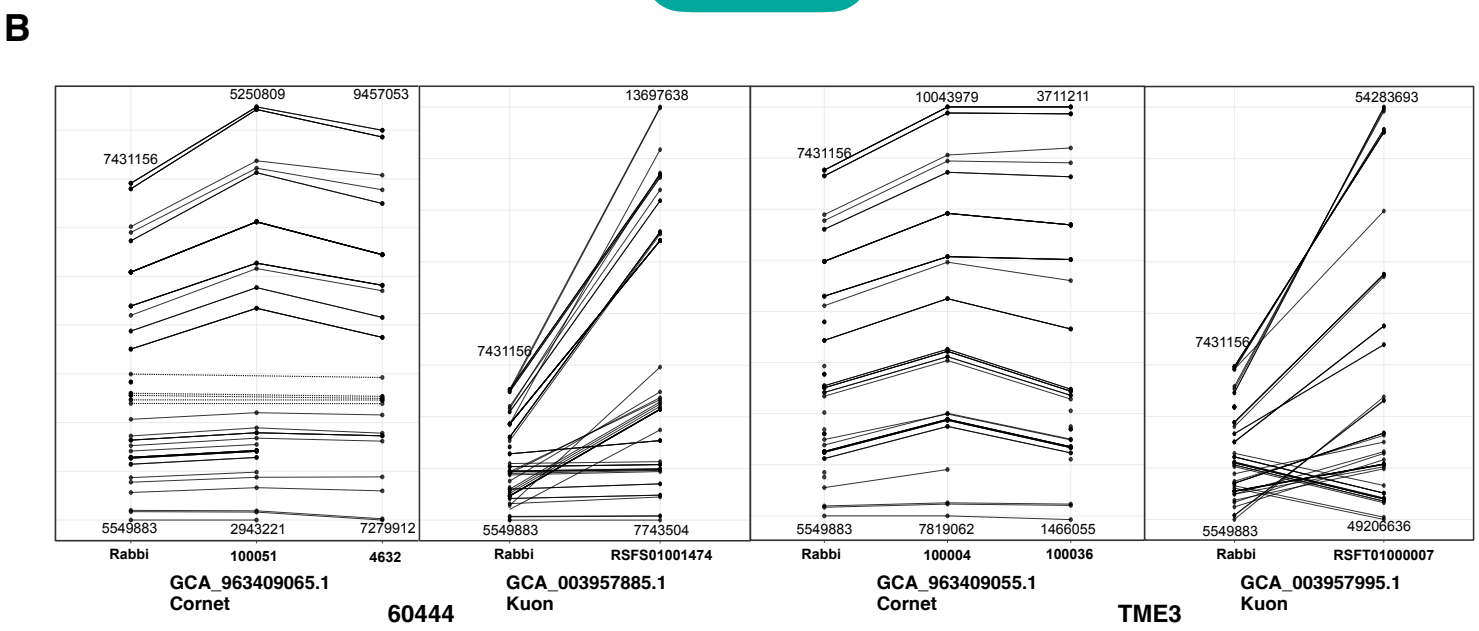
# Figures

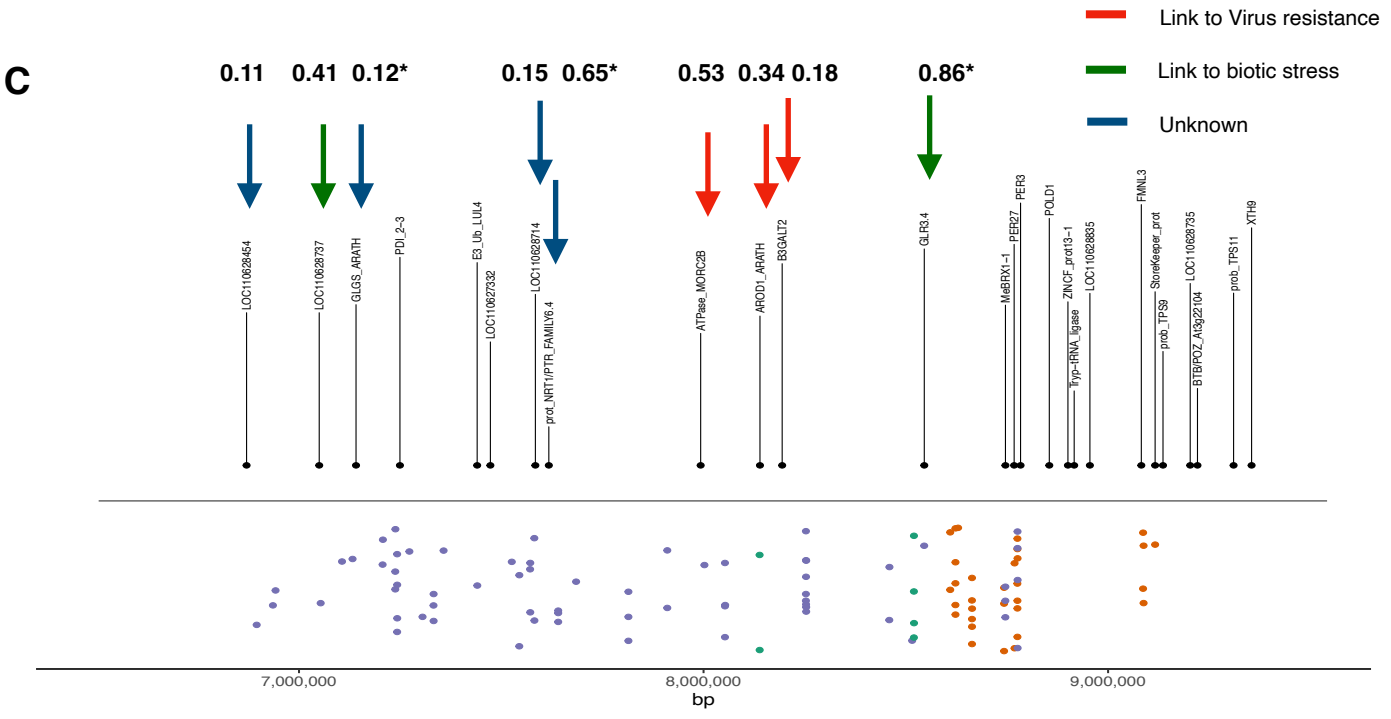
**Figure 1: A. Pipeline of the study. B. Mapping of 64 genetic markers.** The markers from Rabbi et al 2014 (10) were mapped by blastn. First hits of each marker were plotted with ggplot2 according to their coordinate **C. Markers and genes mapping.** The extended CMD2 locus in chromosome 12 of the AM560-2 V8 reference genome. The black dots (above the horizontal black line) show the alignment positions of genes that can be found in the *CMD2* from the genome AM560-2 V8 reference genome. The colored dots (below the horizontal black line) indicate various molecular markers associated with the CMD resistance. The Green dots indicate classical markers (RFLP and SSR markers) previously published by Akano et al. 2002 (43), Lokko et al. 2005 (44), Okogbenin et al. 2007 (45) and Okogbenin et al. 2012 (46).Orange dots indicate *CMD2* SNP markers published by Rabbi et al. 2014 (10), and violet dots indicate markers published by Wolfe et al. 2016 (11). The x-axis of the plot indicates the base pair (bp) position on chromosome 12 of AM560-2**.** Putative resistance proteins with prPred prediction above 0.11 percent are indicated by arrows (prPred prediction are indicated above each arrow).

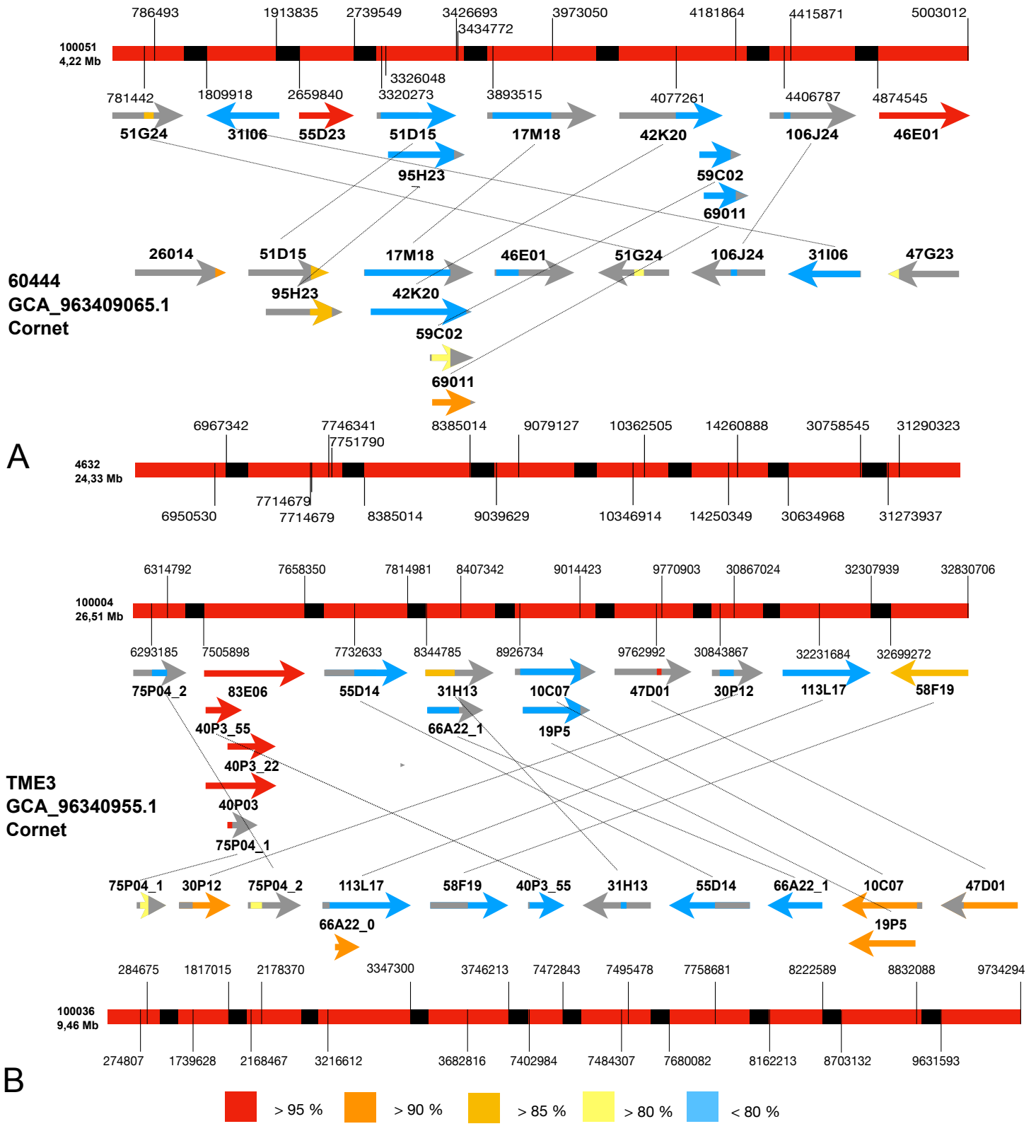
**Figure 2: BACs mapping on haplotypes**

**A. 60444 haplotypes. B. TM3 Haplotypes**. Haplotypes are represented in red with break in black. Coordinates of the BAC mapping regions are indicated on the haplotypes. BAC are represented by arrows and non-mapping regions are represented in gray. Identity of the BAC mapping is represented by colors of the BAC.




786493
1913835
2739549
3426693
3434772
3973050
4181864
4415871
5003012
100051
4,22 Mb
781442
1809918
2659840
3320273
3326048
3893515
4077261
4406787
4874545
51G24
31I06
55D23
51D15
17M18
42K20
106J24
46E01
95H23
59C02
69011
26014
51D15
17M18
46E01
51G24
106J24
31I06
47G23
60444
GCA_963409065.1
Cornet
95H23
42K20
59C02
69011
A
6967342
7746341
7751790
8385014
9079127
10362505
14260888
30758545
31290323
4632
24,33 Mb
6950530
7714679
7714679
8385014
9039629
10346914
14250349
30634968
31273937
6314792
7658350
7814981
8407342
9014423
9770903
30867024
32307939
32830706
100004
26,51 Mb
6293185
7505898
7732633
8344785
8926734
9762992
30843867
32231684
32699272
75P04_2
83E06
55D14
31H13
10C07
47D01
30P12
113L17
58F19
40P3_55
66A22_1
19P5
40P3_22
40P03
75P04_1
TME3
GCA_96340955.1
Cornet
75P04_1
30P12
75P04_2
113L17
58F19
40P3_55
31H13
55D14
66A22_1
10C07
47D01
66A22_0
19P5
284675
1817015
2178370
3347300
3746213
7472843
7495478
7758681
8222589
8832088
9734294
100036
9,46 Mb
274807
1739628
2168467
3216612
3682816
7402984
7484307
7680082
8162213
8703132
9631593
B
> 95 %
> 90 %
> 85 %
> 80 %
< 80 %