DISS.ETH NO. 25224

Reconstructing cassava genomes with single-molecule technologies and chromosome conformation mapping to investigate geminivirus resistance by reverse genetics tools

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

Joel-Elias Kuon
Master of Science (M.Sc.), Agricultural Biotechnology
University of Hohenheim, Germany

Born on
08.06.1987
Citizen of
Germany

Accepted on the recommendation of
Prof. Dr. Wilhelm Gruissem, examiner
Prof. Dr. Hervé Vanderschuren, co-examiner
Prof. Dr. Clara Sánchez-Rodríguez, co-examiner
Prof. Dr. Beat Keller, co-examiner

2018

Table of Contents

Summary

Cassava is the food-security crop that feeds almost a billion people in tropical parts of the world. Cassava production mainly occurs in Sub-Saharan Africa, where the Cassava Mosaic Disease (CMD) caused by geminiviruses leads, on average, to yield loss of 24% that significantly decrease the economic income of smallholder farmers. Research over the past few years has focused on the engineering resistance against cassava mosaic geminiviruses (CMG) using the RNA interference (RNAi) technology, that triggers a sequence-specific defense mechanism against viruses. However, a recent confined field trial revealed that the current RNAi-mediated resistance is not ensured when cassava plants were challenged with a naturally occurring geminivirus population of different CMG strains. These results have prompted the research community to investigate the molecular basis involved in naturally occurring geminivirus resistance in cassava.

Cassava is mainly vegetatively propagated and different cultivars have variable flowering behavior that strongly complicate breeding for traits such as CMD resistance. To date, only three natural resistance sources have been identified that show resistance to all known CMGs. The identified sources of CMD resistance are; the recessive *CMD1*, the dominant and mono-genic *CMD2* and the *CMD3*. Because *CMD2* breeding is highly facilitated by its single-dominant nature, the *CMD2* is widely deployed in CMD resistance breeding programs. Recently, the resistance breakdown was reported for *CMD2*-type plants that went through embryogenesis, a crucial step to generate transgenic cassava plants. The limitations of cassava breeding and the instability of the RNAi-mediated resistance in the field prompt the research community to continue relying heavily on the characterization of geminivirus resistance traits present in cassava germplasm. However, the genes and mechanisms involved in resistance to CMD have remained elusive. It was the aim of this thesis to generate precise genomic resources as well as to develop tools that promote and facilitate the characterization of *CMD2*-type geminivirus resistance genes.

In the work shown in the first chapter, I present the whole genome assembly of the two high-value cassava cultivars TME 3 and 60444. TME 3 is generally known as the origin of the *CMD2* and 60444 as the cassava model cultivar. Cassava has a complex, highly repetitive, medium (750 Mb) sized and diploid genome that has shown previously to be exceptional difficult to assemble. To circumvent these limitations, a novel long-read, single-molecule whole genome sequencing platform (PacBio RSII) was used with sophisticated genome assembly algorithms that allowed the assembly of the first diploid-aware, highly contiguous cassava genomes. Sequences were further curated and improved using optical mapping, a recently published technology that can rapidly fingerprint megabase segments of a genome to generate genome-wide optical maps for sequence scaffolding and structural variation detection. For the final assembly step, the first cassava chromosome-proximity ligation data set (Hi-C) was generated that provided invaluable long-range genomic information to reconstruct chromosomal pseudo-scaffolds. Moreover, the gene space of the two cassava genomes was significantly improved using novel full-length, single-molecule transcriptome sequencing data.

Whole high-throughput transcriptome sequencing revealed a significant number of mono-allelic expressed genes. In addition, an accumulation of mutations was detected in bi-allelic genes that might be a consequence of the clonal propagation over centuries. The two cassava genomes were analyzed for protein clusters, enzymatic reactions and biosynthetic pathways that are shared or specific with other plant genomes. This revealed a highly-significant protein cluster of squalene-monooxygenase activity related proteins that potentially function in the production of antiviral compounds and might reveal an adaptation to viral pathogens induced by breeding and selection. The two high-quality cassava genomes have a near 1.3 Gb diploid genome size, reveal the repetitive DNA in detail, phase thousands of allelic variants in mega-base-pair haplotype blocks and have the highest sequence contiguity compared to other cassava genomes. It can be expected that the two genomes will revolutionize molecular breeding for geminivirus resistance and facilitate CMG resistance gene isolation during coming years.

In the second chapter, I present the genomic context of the *CMD2* resistance locus as well as the diploid-aware genome visualization tool SCEVT that was developed to visualize and compare *CMD2* associated haplotype structures. Furthermore, I present a detailed map of the 267 *de novo* annotated genes, the broad collinearity between the *CMD2* locus and the cassava genetic map, the location of *CMD2*-associated SNP markers as well as the distribution of key sequence features (i.e. repetitive elements) along the *CMD2* locus. The *CMD2* associated genes were analyzed for virus resistance related functional annotation. This revealed a Protein-Disulfide Isomerase (PDI) which are known to interact with viruses and can delay viral replication as well as a Suppressor of Gene Silencing 3 (SGS3). SGS3 genes are known to be involved in post transcriptional gene silencing (PTGS) and has been shown to directly interact with the tomato yellow leaf curl geminivirus (TYLCV) V2 protein. This chapter also revealed the major quality improvement that was achieved over each intermediate assembly step as well as their limitations in terms of assembly gaps. The final *CMD2* map can be used for future candidate gene identification, *CMD2* fine-mapping and sequence polishing (i.e. sequence gap closure) and will be instrumental for future reverse genetic candidate gene screening.

In chapter 3, I present a high-throughput reverse genetic platform and show how this platform facilitated the functional evaluation of 88 genes that were annotated to the *CMD2* locus in a previously published cassava genome. We modified the virus-induced gene silencing (VIGS) platform to make it highly efficient with targeting multi-genes as well as with using a modified agroinoculation protocol that allowed exceptional high and robust virus infection rates. This large-scale reverse genetics screening revealed genes that when silenced through VIGS had an impact on CMV replication, CMV incidence and symptom development. As an example, we identified the same PDI gene as discussed in chapter 2 that allowed the virus to replicate in *CMD2*-type resistant TME 3 plants. Stable silencing of *MePDI2-2* by constitutive expression of hairpin dsRNA in the model cultivar 60444 lines led to reduced geminivirus incidence, mild virus symptom development and decreased virus load compared to the control plants. Our pipeline demonstrates the potential of the VIGS platform to rapidly identify host genes whose modulation can alter symptom score

and geminivirus replication. This reverse genetic platform allows the in-depth characterization of the new genomic data for the *CMD2* locus that has been generated in chapter 1 and presented in chapter 2.

The work presented in this thesis, show important novel achievements and a step forward in cassava genetic research and trait discovery. The assembly pipeline shown in this thesis can be similarly applied for the assembly of other cassava genomes with the ultimate goal to generate the first cassava-pangenome that would represent the full complement of genes present in the *Manihot* clade. The full genetic diversity revealed by a pan-genome would greatly facilitate the isolation of agronomically important genes in a crop where genetic diversity is limited by breeding constraints as well as the clonal propagation. The *CMD2* locus map enable the targeted, allele-specific characterisation of *CMD2* associated candidate genes using the novel VIGS platform developed in chapter 3. Once the resistance gene(s) is known, the ultimate goal should be to generate transgenic cassava with stacked resistance genes that in turn would confer stable and durable CMV resistance in cassava cultivars.

Zusammenfassung

Cassava, auch bekannt als Maniok oder Tapioka, ist die Ernährungssicherheit-Feldfrucht, welche nahezu eine Milliarde Menschen in den tropischen Regionen der Welt ernährt. Der Cassava-Anbau findet hauptsächlich in Subsahara-Afrika statt, wobei die Cassava Mosaik Erkrankung, welche durch Cassava Mosaik Geminiviren (CMG) verursacht wird, 24 % Ernteeinbuße verursacht und dadurch zu einem signifikanten ökonomischen Verlust unter Kleinstbauern führt. Durch intensive Forschung wurde eine Resistenz gegen die Cassava Mosaik Krankheit durch den Einsatz der RNA interferenz (RNAi) Technologie erzeugt, welche eine natürliche Sequenz-spezifischer Abwehrmechanismus gegen Viren darstellt. Allerdings wurde in kürzlich veröffentlichten Feldversuchen in Kenia gezeigt, dass diese RNAi-übermittelte Resistenz ungenügend schützt, wenn eine transgene Cassava einer natürlichen Geminiviruspopulation mit verschiedenen Virusstämmen ausgesetzt wird.

Cassava wird hauptsächlich vegetativ vermehrt und zeigt starke sortenabhängige Unterschiede im Blühzeitpunkt. Diese physiologischen Eigenschaften erschweren eine konventionelle Geminivirus-Resistenzzüchtung. Zum heutigen Zeitpunkt sind lediglich drei natürliche Geminivirus-Resistenzen bekannt, welche eine stabile Resistenz im Feld vorweisen. Die identifizierten Resistenzen sind; das rezessiv vererbte *CMD1*, das mono-genetisch dominant vererbte *CMD2* und das *CMD3*. Da die Züchtung generell durch ein dominant vererbbares Gen vereinfacht wird, wurde die *CMD2* Resistenz hauptsächlich in Züchtungsprogrammen eingesetzt. Allerdings sind die molekularen Mechanismen, als auch das Gen bis zum heutigen Zeitpunkt gänzlich unbekannt. Zudem wurde kürzlich im Feld der Zusammenbruch der *CMD2* Resistenz festgestellt, sobald Cassava-Pflanzen mithilfe der Embryogenese-Technik gentechnisch transformiert wurden. Die oben genannten Probleme bei der Resistenzzüchtung, als auch die ungenügende RNAi-basierende Resistenz, hat den Forschungsfokus der letzten Jahre stark auf die genaue Charakterisierung dieser natürlichen Resistenz-Quellen gelegt.

Es war das Ziel dieser Doktorarbeit hoch-präzise genomische Ressourcen zu erschaffen und molekulare Instrumente zu entwickeln, welche eine Charakterisierung und Identifizierung der *CMD2* assoziierten Gene ermöglichen. Im Rahmen dieser Doktorarbeit wurden Genome für die Cassava-Sorten TME 3, welche als Ursprungssorte der *CMD2* Resistenz gilt, als auch für die Cassava Modell-Sorte 60444 assembliert. Im zweiten Kapitel präsentiere ich die Entwicklung einer neuen Software, mit Hilfe dessen sich der *CMD2* locus genau rekonstruieren ließ. Im letzten Teil dieser Thesis präsentiere ich die Entwicklung einer hoch-durchsatz ‚reverse genetics' Methode, welche die Charakterisierung dutzender *CMD2* Kandidatengene ermöglichte.

Im ersten Kapitel präsentiere ich wie die zwei Genome für TME 3 und 60444 entschlüsselt wurden. Cassava hat ein komplexes, hoch repetitives, mittelgroßes (~750 Mb) und diploid-heterozygotes Genom, was frühere Assemblierungen stark hinderte. Um diese Einschränkungen zu umgehen, wurde in diesem Projekt eine neuartige Genomsequenzierungs-Technologie verwendeten, welche mit Hilfe langer ‚reads' und ausgeklügelten Assemblierungsalgorithmen das erste diploide Cassava-Genom ermöglichte. Anschließend

wurden die Genome mit Hilfe genomweiter optischer Karten, eine Technologie welche durch ‚Fingerabdruck'-Technik Megabasenpaar-Segmente der DNA kartiert, kuriert. Diese optischen Karten wurden dazu verwendet die Sequenz-kontinuität zu verbessern, einzelne Haplotypen zu assemblieren und große genomische strukturelle Varianten (SVs) zu identifizieren. Im letzten Assemblierungsschritt wurde mit Hilfe der ersten 'chromosome-conformation-capture'-Sequenzierung (Hi-C) an Cassava ein vollständiger Chromosomensatz rekonstruiert. Die neuen Cassava Genome zeigen zudem eine signifikante Verbesserung der Gen-Annotation auf, da zur Gen-Annotation eine neue Sequenzierungstechnik angewandt wurde, welche die volle Länge der Messenger RNA (mRNA) sequenziert. Daraufhin wurden mit Hilfe hoch-durchsatz Transkriptom-Daten die Genome auf monoallelisch und biallelisch exprimierte Gene untersucht. Außerdem wurde eine Akkumulation von Mutationen in biallelischen Genen detektiert, welche durch eine Jahrhunderte lange klonale Vermehrung begünstigt worden sein könnte. Die beiden Genome wurden auf physische Protein-Cluster, enzymatische Reaktionen und Biosynthesewege analysiert um gemeinsame oder spezifische Eigenschaften zu detektieren. Diese Analyse identifizierte den hoch-signifikanten Protein-cluster ‚Squalen-monooxygenase Activity', welcher möglicherweise in der Herstellung antiviraler Verbindungen involviert ist und durch Züchtung und Selektion hervorgerufen worden sein könnte. Die beiden Genome spiegeln zum ersten Mal überhaupt die diploide Natur des Cassava-Genoms wieder, zeigen die als schwierig zu assemblierende repetitive DNA im Detail, decken tausende allelische Varianten auf und ermöglichen die SV Identifizierung mit Hilfe Megabasenpaare-spannende Haplotypen. Es darf angenommen werden, dass die beiden neuen Genome die zukünftige molekulare Züchtung revolutionieren und die Virusresistenzzüchtung, die Isolation von Resistenzgenen und ihre molekulare Charakterisierung beschleunigen und erleichtern.

Im zweiten Kapitel präsentiere ich den genomischen Kontext des *CMD2* locus als auch die Software SCEVT, welche im Rahmen dieser Doktorarbeit entwickelt wurde um Haplotypen und Sequenzen unkompliziert zu visualisieren und zu vergleichen. In diesem Kapitel präsentiere ich zudem eine detaillierte Karte der 267 *CMD2* assoziierten Gene, die Kollinearität zwischen dem *CMD2* locus und der generellen genetischen Karte, die genaue Lage der *CMD2* assoziierten genetischen Marker, als auch die Verteilung der Schlüssel-Sequenzmerkmale entlang des *CMD2* locus. Unter den *CMD2*-assoziierten Genen fand ich eine Protein-Disulfide Isomerase (PDI), welche mit Viren interagieren und, im Falle des HI-Virus, die Virusreplikation verzögern können. Zudem wurde ein Suppressor of Gene Silencing 3 (SGS3) Gen gefunden, welches im post-transkriptionellem Gen-Silencing involviert ist und direkt mit Proteinen des Tomaten Geminivirus interagieren kann. Dieses Kapitel hat auch die Verbesserungen gezeigt, welche durch die unterschiedlichen Assemblierungsschritte erreicht wurden. Die *CMD2* Karte bietet nun eine Plattform um Kandidatengene zu identifizieren, das *CMD2* Gen mittels klassischer Genetik genauer zu kartieren und um die Sequenzkontinuität noch weiter zu verbessern.

In Kapitel 3 präsentiere ich ein hoch-durchsatz ‚reverse genetics' Plattform, welche die funktionelle Charakterisierung von 88 *CMD2* assoziierten Gene ermöglicht hat. Dazu wurde ein Virus-Induced-Gene-Silencing (VIGS) Strategie weiterentwickelt und spezifische VIGS-Konstrukte entworfen, welche fünf

Gene je Konstrukt herunterregulieren und dadurch die Analyse von dutzenden Genen ermöglicht. Zudem wurde die Agro-inokulation der VIGS-Konstrukte weiter verbessert, um eine hohe Effektivität und Infektionsrate zu gewährleisten. Durch diese groß angelegte ‚reverse genetics' Untersuchung wurden Gene identifizierte, welche einen Einfluss nahmen auf die Virusreplikation, Virusinfektion und Symptomentwicklung. In diesem Kontext wurde ein PDI-Protein identifiziert, welches Virussymptome in der *CMD2* resistenten TME 3 verursachte. Des Weiteren führte eine stabiles ‚knock-down' des PDI durch genetische Transformation in 60444 zu einer reduzierten Symptomentwicklung als auch zu einem verringerten Virustiter in ausgewählten Linien verglichen mit den Kontroll-Linien. Dieses Kapitel zeigt das Potential dieser neuen VIGS-Plattform um schnell und effizient genetische Wechselbeziehungen zwischen Virus und Wirt zu identifizieren. Zudem könnte diese ‚reverse genetics' Strategie in Zukunft verwendet werden, um *CMD2*-assoziierte Gene, welche in Kapitel 2 thematisiert wurden, im Detail zu charakterisieren.

Die Ergebnisse dieser Doktorarbeit zeigen wichtige und neuartige Errungenschaften auf und bedeuten einen Fortschritt im Verständnis über die Genetik dieser ungemein wichtigen Kulturpflanze. Die Genom Assemblierung-Methode, welche in dieser Thesis für Cassava entwickelt wurde, kann ähnlich an weiteren Cassava-Sorten angewandt werden, um das erste Pan-Genom für Cassava zu generieren. Dieses würde die vollständige genetische Diversität in Cassava aufzeigen und die Isolation von agronomisch wichtigen genetischen Eigenschaften revolutionieren. Die erste hochauflösende und detailreiche Karte des *CMD2* locus ermöglicht eine zukünftige Charakterisierung mit Hilfe der VIGS Methode. Das ultimative Ziel sollte darin bestehen, die CMG Resistenzgene zu identifizieren und die verschiedenen Resistenzmechanismen in Sorten zu ‚stapeln', um eine dauerhafte und stabile Resistenz zu gewährleisten.

Chapter 1

General Introduction to geminivirus resistance and third-generation genome sequencing

Background

Cassava (*Manihot esculenta* Crantz, Euphorbiaceae, 2n = 36) is a woody perennial shrub that originated in the southern Amazon basin [1]. Cassava is cultivated mainly for its edible starchy tuberous roots and serves as an important food crop for a billion people in 105 countries (FAOSTAT, 2016). Thus, in the developing world cassava belongs to the top four most important crops after rice and maize, with estimated production of 277 million tons in 2016 (FAOSTAT, 2016). Cassava is grown throughout tropical and subtropical regions and has a wide range of usage. In Africa, cassava is mainly considered as a food security crop and grown primarily for food, whereas in Asia cassava root chips are commonly used at industrial levels for animal feeding and as raw material for the paper industry or biofuels [2], [3]. Due to genotype-dependent asynchronous flowering, sexual reproduction is rare and cassava is typically propagated through the use of stem cuttings [4].

In the recent years, cassava mosaic geminiviruses (CMG) have developed to become the most important agronomically threat for cassava production in Africa and the Indian subcontinent and causes over 25 million tons of yield losses that affects food security of more than 500 million people [5]–[7]. Geminiviruses represent a big family of small, circular, single-stranded (ss) DNA viruses that can infect a variety of other crops such as maize, bean, cotton and tomato [8]. For instance, the maize streak disease, cassava mosaic disease (CMD), the cotton leaf curl disease and the tomato yellow leaf curl disease (TYLCD) have a great impact on agricultural productivity and can cause yield losses, in extreme cases, from 10-100% [9]. For cassava, geminivirus incidence as well as symptom severity has strongly increased over the past decades, probably as a result of insecticide resistance, global warming and human activity [10]. Geminiviruses have small DNA genomes (2.7 kb – 3 kb) and depend heavily on host cellular machineries for viral replication, assembly, movement, transmission and symptom development [8]. They have either a genome consisting of a single component (monopartite) or with two DNA components (bipartite) which have been classified as DNA-A - and DNA-B [11]. Their small genome has limited coding capacities and encode for five to seven proteins [12]. In addition, they bear multiple silencing suppressors that alter host DNA methylation, microRNA (miRNA) and small interfering RNA (siRNA) machineries [13]–[15]. Geminiviruses require insect vectors such as whiteflies, aphids or leaf-hoppers for transmission.

The development of cultivars that are genetically resistant to geminiviruses is an efficient strategy to tackle the problems associated with the virus disease. Major achievements have been gained with genetically engineering resistance mechanisms that rely on the RNA interference (RNAi) technology [16]–[18] but naturally occurring resistance genes (*R*-genes) can provide a highly efficient barrier against viral infection as

well. In the past decades, significant gains have been achieved in the understanding of the molecular mechanisms involved in natural recessive and dominant *R*-genes but only very few *R*-genes against geminiviruses have been cloned and identified [19]–[21]. In this context, the limiting factors for *R*-gene cloning and identification are the lack in genomic resources such as high-quality genome assemblies that can be used for gene mapping and candidate gene selection, and the missing or very time-consuming reverse genetic tools which are essential for candidate gene confirmation.

In the following introduction, I discuss recent developments in natural resistance mechanisms against geminiviruses, geminivirus resistance that has been detected in cassava germplasm and the limitations for *R*-gene isolation attempts using the current cassava genomic resources. Furthermore, I introduce novel, third-generation sequencing and mapping platforms and chromosome-proximity mapping approaches that revolutionary changed the process of genome sequencing in the past few years. These novel genomic tools were used to generate the first high-quality African cassava genomes for TME 3, the source of the monogenic and dominant geminivirus resistance *CMD2,* as well as for the model cultivar 60444 in order to facilitate and speed-up the discovery of the *R*-gene(s) underlying geminivirus resistance.

Natural resistance against geminiviruses

Plants carry a unique and complex arsenal for defense against pathogens, that consists of different layers and enables plants to avoid and suppress pathogen infections. Here, often a genetically determined pathogen recognition system, controlled by a host *R*-gene, confers resistance to a pathogen that carries the corresponding avirulence gene(s) (*Avr*-gene). Such a gene-to-gene resistance mechanism can cause host defense responses, such as local cell death or hypersensitive responses (HR) that in turn limits spreading of the invading pathogen [22]. Most known plant *R*-genes contain a nucleotide binding site (NBS) and a leucine-rich repeat (LRR) domains where the latter is involved in pathogen recognition and in many cases represent single dominant resistance genes [23]. Based on this model, hundreds of *R*-gene loci were discovered that are involved in resistance to bacterial and fungal pathogens. In the special case of plant viral pathogens, only 22 *R*-genes have been identified and isolated, often by following a traditional map-based cloning strategy. *R*-genes against geminiviruses have obtained growing interest but only three have been mapped, cloned and characterized in tomato and common bean [19]–[21], [24].

The tomato yellow leaf curl disease is one of the major viral diseases of tomato worldwide infecting all cultivated tomato varieties [25]. To fight against this threat, considerable efforts have been invested in resistance breeding against the Tomato *yellow leaf curl begomovirus* (TYLCV). Several wild tomato species (e.g. *Solanum chilense*, *S. perivuanum*, *S. pimpinellifolium*) that showed TYLCV resistance were introgressed into the domesticated tomato (*S. lycopersicum*). Subsequently, five resistance loci, named *Ty-1* to *Ty-5*, have been found and mapped to the tomato chromosomes [26]. Several years later, the *Ty-1* and *Ty-3* based resistance was cloned and identified to encode a non-classical *R*-gene. Verlaan and colleagues revealed that the *Ty-1* and *Ty-3* are allelic, are members of a multigene family and encode for a tomato RNA-

dependent RNA polymerase (RDR) that leads to a complete resistant phenotype with no visible virus symptoms after virus inoculation [27]. However, low levels of virus titer were still detectable suggesting a tolerance mechanism rather than complete resistance. Furthermore, it was supposed that the RDRs might be involved in the amplification of the RNAi response and the transcriptional gene silencing against the TYCLV. To date, only a single recessive *R*-gene was cloned and identified. The recessive tomato *R*-gene *Ty-5* encodes a homolog of the messenger RNA surveillance factor Pelota (*Pelo)* and implies a completely novel mechanism acting against geminiviruses. Two SNPs were identified, one in the promoter region and one in the coding sequence of the *Pelo,* that changed the tomato plant from susceptible to resistant [20]. The protein *Pelo* is involved in the latest phase of ribosome-driven protein biosynthesis and the mutant *Pelo* must have altered host components required for a stage of the virus life cycle. However, the exact role of *Pelo* under geminivirus infection remains unclear. In contrast to geminivirus *R*-genes found in tomato, the dominant *R*-gene *CYR1* encodes a bean NBS-LRR protein and confers broad resistance against the *Mungbean yellow mosaic India virus* (MYMIV) [28], [29]. It is generally assumed that these classical R-*genes* are involved in stress signaling and pathogen recognition through the *Avr*- gene or *Avr*-gene products [24].

The few examples presented above indicate that only a very little proportion of the natural biodiversity available for geminivirus disease resistance has been exploited and genes underlying QTLs for quantitative resistance haven't been identified yet. The technical challenges associated with multi-gene mapping and cloning has set research focus on monogenic resistance genes despite their shorter durability.

Natural resistance against geminiviruses in cassava

Three types of natural resistance were identified for controlling CMD. The polygenic, recessive *CMD1*, introgressed from wild cassava relatives [30], the *CMD2,* a single-dominant gene locus conferring resistance to all known CMVs [31], and the *CMD3*, a resistance source that was recently distinguished from the *CMD2* based on a single, *CMD2*-unlinked genetic marker [32]. The *CMD2* was discovered within landraces collected from farmers' fields in Nigeria and other West African countries during the 1980s and 1990s but their breeding pedigree is unknown [33]. Because a single-dominant gene greatly facilitates breeding, the *CMD2*-type resistance became the predominant resistance source deployed in African cassava breeding programs despite its underlying molecular mechanisms remained elusive [34], [35]. Recently, the breakdown of the *CMD2* resistance was reported for plants that undergone embryogenesis, an essential step for cassava transformation [36].

Till today, only a single molecular analysis exists that attempted to investigate the molecular basis of the *CMD2* using next-generation sequencing (NGS) of the transcriptome [37]. This time course experiment revealed that overall fewer responsive transcripts were found in *CMD2*-type cultivar TME 3 as compared to virus susceptible cultivar T 200 after virus infection with *South African cassava mosaic virus* (SACMV). Moreover, the number of responsive transcripts in TME 3 declined over three time points that could be

explained with the virus-recovery phenotype that has been reported before for *CMD2*-type resistance [38]. However, the study also states the exceptional low mapping rate for the NGS reads (50.7 % for T200 and 55.06% for TME 3) when aligning them back to the cassava reference genome [39] in order to quantify the gene expression. This low mapping rate strongly suggests a reference genome bias that could have influenced the read counting and down-stream analysis drastically.

This example as well as the resistance breakdown indicate that new genomic resources are urgently needed to better characterise *CMD2*-type geminivirus resistance mechanisms. The availability of high-quality genomes for *CMD2*-type cultivars would allow the precise use of NGS platforms for the characterisation of the transcriptome and methylome changes associated with virus infection and resistance breakdown. The fact that the vast majority of geminivirus resistance breeding programs rely on the stability of the *CMD2* further indicate the great need to assemble high-quality cassava genomes with the ultimate goal to identify the corresponding *R*-genes.

Generation of novel genetic resources for cassava using single-molecule and proximity ligation mapping technologies

The extraordinary progress in high-throughput and cost-effective NGS technologies has drastically accelerated our understanding of genomic diversity and facilitated the rapid identification of genes underlying phenotypes [40], [41]. For cassava, the first draft genome was released in 2012 using a partly inbred south American genotype named AM560 [39]. Two years later, a draft genome of Asian cassava variety KU50 and of the cassava wild relative W14 (*Manihot esculenta* ssp. *flabellifolia*) was assembled [42]. Since the release of these genomes, both genetic research and crop improvement in cassava have benefitted from the partly ordered draft sequence assemblies. For instance, this resource have enabled first population genomic studies [34], [35], [43], [44], transcriptome characterization [37], [45]–[47] and whole methylome-profiling [48]. However, the current versions of the draft cassava genomes are represented as linear and haploid DNA sequence. Such a representation for a highly heterozygous genome can cause misleading results when applying read mapping sensitive applications that rely on accurate read placement. For example, whole-transcriptome sequencing reads can align falsely or even fail to map when they span challenging regions with structural variations (SV). Misplaced reads do in turn result in both missed true variants or incorrectly reported false variants and bias downstream results.

Crop plant genome sequencing is often limited because of the excessive proportion of repetitive DNA elements (RE), the high heterozygosity and the number of basepairs to sequence. In this respect, cassava has heterozygosity estimated to be among the highest found in sequenced plant genomes [42], is rich in REs and has a haploid genome size of ~750 Mb [49]. Because of these characteristics, cassava has proven difficult to assemble and previous attempts to assemble this genome yielded highly fragmented and incomplete genome assemblies [39], [42], [49]. The fact that cassava has an unfavorable genomic composition for sequencing and assembly, novel sequencing technologies have to be implemented to unravel these difficulties. Several new genomic sequencing and mapping technologies have been launched that allow to

assemble the previously inaccessible repetitive sequences, microsatellites, haplotype variants and other complex sequences. The release of novel third-generation, single-molecule and long-read sequencing platforms from Pacific Biosciences (PacBio) [50] or Oxford Nanopore [51] changed sequencing and assembling of highly complex genomes revolutionary. In contrast to the second-generation sequencing platforms (i.e. Illumina instruments) that generate sequencing reads usually of hundred nucleotides, these novel platforms are capable to generate long reads from a single molecule averaging around 10-20 kb in length. These long reads greatly facilitate whole genome assemblies (WGA) because they span most of the REs, can be used for haplotype-phasing and replace the laborious generation of various large-insert mate-pair NGS libraries that were a requirement in earlier WGA projects. However, their major limitation is the relatively high frequency of sequencing errors that can vary between 15-20% [52]. Because these sequencing errors appear randomly, they can be circumvented with the usage of sophisticated assembly algorithms to create highly accurate assemblies including haplotype phased genomes, with using only a modest sequencing depth [53]–[56]. These sequencing errors can be revised through the implementation of sequencing the same molecule various times (i.e. generating high enough genome data coverage). Having multiple sequencing reads for a single molecule will allow the algorithm to correct random base-calling errors.

In the past few years, platforms were developed that allow a large-distance scaffolding of sequences. Long-range sequence information, such as optical mapping and proximity mapping, can be used to form scaffolds by ordering and orienting contigs that in in the best-case span entire chromosomes.

Optical mapping was originally developed for ordering restriction enzyme sites through digestion and size-separation [57] and was then further developed to tag particular sequences within DNA molecules that are up to ~1 Mb long via fluorescent DNA marks. The results were stored in images that show a certain tag-pattern for each DNA molecule that run through nanochannels based arrays. Subsequently, the images were aligned to each other to assemble the location of each molecule relative to each other. This generates a cost-effective genome-wide optical map that can be used for *de novo* genome assembly, gap filling, structural variations (SV) detection and haplotype phasing of up to several Mb genomic distance [53], [58]. This technology was tuned into a high-throughput platform by the company BioNano that now allows rapid fingerprinting of megabase genomic segments within a few hours (BioNano Genomics).

Chromosome interaction mapping (Hi-C) data provide a remarkable potential for long range scaffolding and haplotype phasing of sequences [59]. Hi-C is an adaptation of the chromosome conformation capture (abbreviated to 3C) methodology [60] that uses formaldehyde cross-linked chromatin for digestion and subsequent re-ligation. This generates chimeric, circular DNA, comprised of two restriction fragments that lay initially in close spatial proximity within the nucleus. This scaffolding platform relays on the principle that the frequency of long-range chromatin interactions decay rapidly as a linear distance along a chromosome increases and reveals a genome-wide interaction matrix that can be exploited to place assembled sequences accordingly [61]. It was reported in earlier studies that this platform has been key for

the scaffolding of full chromosomes including the sequence ordering within highly repetitive and previously inaccessible genomic regions such as the centromere [62], [63].

To my best knowledge, not a single crop plant genome has been released that was assembled with the power of long-read sequencing, optical mapping and Hi-C based chromosome reconstruction. But It has been shown for the goat genome that these technologies have excellent sequence and scaffold continuity metrics [61], [64]. As a prerequisite to marker assisted breeding, allele mining and gene isolation, we decided to combine the power of these three platforms in order to generate the first long-read, optical map improved and Hi-C scaffolded genomes for two high-value cassava lines. We decided to assemble the first genomes for African cassava cultivar TME 3, the source of the *CMD2*-type resistance [31], and the cassava model cultivar 60444.

Aim of the thesis:

- Generate high-quality plant genomes for a *CMD2*-type cultivar and the cassava model cultivar 60444 using novel sequencing, mapping and gene-space annotation platforms (PacBio whole genome sequencing, PacBio RNA-Isoform sequencing, optical mapping, chromosome-proximity based sequence ordering).
- Unravel the dominant, monogenic *CMD2* geminivirus resistance locus using two *de novo* high-quality genomes and develop a visualization tool that represent the true diploid nature of an assembly by minor input data requirements.
- Establish a high-throughput forward genetics platform in cassava to assess candidate genes underlying a QTL locus (i.e. *CMD2* locus) that is cost-efficient and doesn't need expensive lab equipment.

Chapter 2

Cassava genomes assembled with single-molecule long reads, optical and Hi-C maps reveal narrow genetic diversity and mono-allelic expression

Personal contribution:
I optimised the cassava leaf-tissue DNA extraction protocol for long-read sequencing using PacBio library chemicals and instruments. Under guidance of Stefan Grob, I run the two cassava Hi-C experiments and constructed the corresponding Illumina sequencing libraries. I generated the optical genome data together with Lucy Poveda and generated and analysed the Iso-Seq data together with Weihong Qi. I assembled and quality assessed the genomes with the help of Weihong Qi and organized the raw data with the help of Matthias Hirsch-Hoffmann. I annotated the repetitive elements and visualised key genomic features. I analysed the genomes for mono-allelic expressed genes together with Matthias Hirsch-Hoffmann and performed the protein clustering assay under guidance of Pascal Schläpfer. I wrote the draft manuscript with input from Prof. Gruissem, Prof. Vanderschuren and Weihong Qi

Publication state:
This manuscript is in the final phase of editing and will be submitted to *Nature Genetics* as a research letter first.

Cassava genomes assembled with single-molecule long reads, optical and Hi-C maps reveal narrow genetic diversity and mono-allelic expression

Joel-E. Kuon[1], Weihong Qi[2], Matthias Hirsch-Hoffmann[1], Pascal Schläpfer[1], Andrea Patrignani[2], Lucy Poveda[2], Stefan Grob[3], Miyako Keller[1], Rie Shimizu-Inatsugi[4], Ueli Grossniklaus[3], Hervé Vanderschuren[5], Wilhelm Gruissem[1]

[1]Institute of Molecular Plant Biology, Department of Biology, ETH Zurich, Universitätstrasse 2, 8092 Zurich, Switzerland
[2] Functional Genomics Center Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland.
[3]Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland
[4]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[5]AgroBioChem Department, University of Liège, Passage des Déportés 2, Gembloux, Belgium

Correspondence should be addressed to kuonj@ethz.ch & wgruisse@ethz.ch

Summary
Cassava *(Manihot Esculenta* Crantz) is an important food security crop for nearly one-billion people in tropical and sub-tropical regions worldwide. But genetic improvement of cassava is constrained by the proportion of deleterious mutations in coding sequences and highly fragmented, incomplete draft genome assemblies [42], [49], [65]. Full cassava genome assemblies have not been achieved because of the excessive heterozygous genetic composition and diploid status of the genome. Here we present the first diploid-aware assemblies and annotation of genomes for two African cassava varieties (TME 3 and 60444) using single-molecule real-time sequencing, combined with high-resolution optical mapping and chromosome proximity ligation data to create chromosomal sequence scaffolds. We revised and improved the cassava *de novo* predicted gene space using full-length, single-molecule CDS sequencing and analysed the transcriptome for allele-specific expression. The two high-quality cassava genomes have a near 1.3 Gb diploid genome size, reveal the repetitive DNA proportion in detail, and phase thousands of allelic variants in mega-base-pair haplotype blocks. We expect that the high-quality genomes will facilitate targeted molecular breeding and gene isolation to improve cassava.

As a subsistence crop, cassava is valued for its starchy storage roots, especially by small-holder farmers[1]. But cassava is also becoming increasingly important as an industrial crop for the production of starch, energy (bioethanol), and as livestock feed [2], [66]. Genetic gains from breeding have been small over the last century compared to other crops [67]. The long breeding cycle, clonal propagation, and poor flowering have limited genetic improvement considerably [68]. Only recently has cassava genetics and germplasm benefited from partially ordered draft genome assemblies [39], [42], [49]. But identifying and understanding genetically- and epigenetically-controlled cassava traits based on the fragmented and incomplete draft status of the genomes remains challenging. Cassava has a complex, diploid (2n=36) genome with an estimated heterozygosity that is highest among sequenced plant genomes [42]. This and the large number of transposable elements (TEs) make it challenging to assemble the whole cassava genome [69], [70]. To date, existing cassava genomes have been assembled only from short sequence reads in haploid assemblies and miss to represent the whole information present in an heterozygous organism [39], [42], [49].

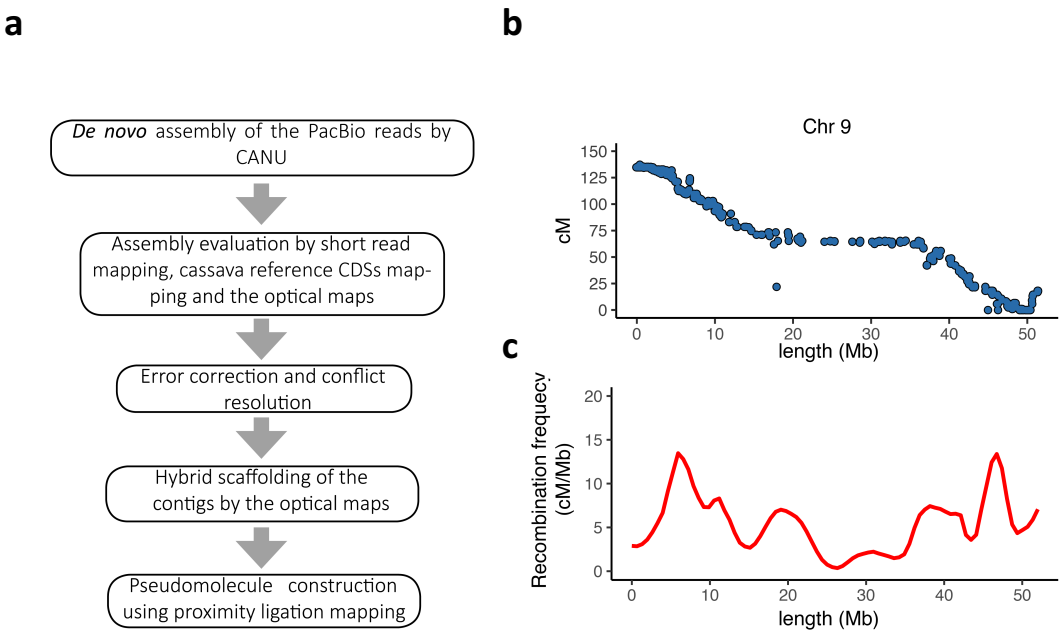


Figure 1 Assembly and validation of the 60444 and TME 3 heterozygous cassava genomes**.** (a) Overview of the processing pipeline used for the assembly of the TME 3 and 60444 genomes (see Supplementary Note for details). (b) Graphical representation of the location of SNP markers on the physical map (x-axis), as compared to their position on the composite cassava genetic map (y-axis), for the single scaffold Scaffold_176;HRSCAF=892 of the cassava TME 3 genome. Each genetic marker is depicted as a dot on the plot (937 data points). (c) Graphical representation of the mean local recombination frequencies between SNP markers along Chr 9. The x-axis represents the physical positions of the means on Chr 9, and the y-axis indicates the recombination ratio (centiMorgan (cM)/Mb) in each 1-Mb sliding window.

Table 1 Assembly statistics for the cassava TME 3 and 60444 genomes compared with previously published assemblies of cassava genomes

Cultivar	**TME 3**	**60444**	**W14**[42]	**KU50**[42]	**AM560-2**[49]
Contigs	12,971	11,459	82,335	99,509	39,574
Contig N50 (kb)	97.58	116.78	10.23	5.28	27.87
Optical map supported scaffolds	558	552	NA	NA	NA
Optical Hybrid-scaffold NG50 (Mb)	2.25	2.35	NA	NA	NA
Hi-C scaffolding N50 (Mb)	53.35	59.19	NA	NA	NA
Assembly size (Mb)	1224.5	1276.9	427.5	291.1	582.3
TE proportion (%)	64.81	64.91	36.9	25.7	50.3
Annotated protein-coding genes	33,853	34,127	34,483	38,845	33,033

Here we report the first nearly complete *de novo* assembly and annotation for two African cassava cultivars. TME 3 is an important source for the cassava mosaic virus disease resistance trait *CMD2* [31], [34], [35], and 60444 is widely used as an experimental model cultivar for gene transfer and gene editing [71]–[75] (Figure 1). With 70x whole genome shotgun, PacBio long-read, single-molecule real-time (SMRT) sequencing data**,** we assembled the TME 3 genome into 12,971 contigs with a N50 of 98 kb (i.e., 50% of the assembly consists of 98 kb or longer contigs). For 60444, we assembled reads into 11,459 contigs with a N50 of 117 kb (Table 1) (Supplementary Figure 1, Supplementary Table 1). Long-read genome assemblies generated by three different assemblers were assessed for their quality by aligning Illumina paired-end (PE) reads from the same cultivar back to the assembly. Based on this benchmarking, we found that the CANU assembler [54] generated the most robust assemblies with the highest proportion of mapped PE reads (98.4% for 60444 and 96.4% for TME 3) and the smallest proportion of discordant read pair alignments (0.11% for TME 3 and 0.09 for 60444) (Supplementary Table 2).

The high heterozygosity of the cassava genome is the consequence of interspecific admixture and past breeding [49], [76]. Optical mapping is useful to phase haplotypes especially in genomes with divergent homologous chromosomes [58]. We generated two high-coverage optical maps (150x coverage for 60444 and 130x for TME 3) using the BioNano Genomics IrysView DNA molecule imaging platform and Irys software tools. The fluorescently-labelled DNA molecules of the two cassava genomes assembled into almost exactly the same diploid genome size of 1.2 Gb (1,205 Mb for TME 3 and 1,204 Mb for 60444). Both genome maps showed a similar N50 map contiguity of 1.801 Mb and 1.875 Mb for TME 3 and 60444, respectively. Based on flow cytometry, we estimated the haploid cassava genome size to be 745 Mb for 60444 and 765 Mb for TME 3 (Supplementary Figure 2). This allowed us to calculate the number of homologous chromosome fractions that had been phased into individual haplotypes. The diploid optical map assemblies span 1.62 times the haploid cassava genome, which represents 80.08% of the diploid genome phased into true haplotype segments (Supplementary Table 3).

The Portuguese introduced cassava from South-America into Africa in the 16th and 17th century, and since then the African germplasm diversity remained exceptionally narrow [77]. Previous diversity studies relied on short-read mapping data only, but genome-wide structural variants are challenging to be detected in heterozygous and complex plant genomes [49]. We tested our optical maps for genomic diversity between the two cassava cultivars. The majority (81%) of the consensus optical maps from the TME 3 genome could be aligned with the optical maps of the 60444 genome via common label patterns, indicating an exceptional low level of genomic diversity between the two cassava genomes. We then screened the alignments for TME 3-specific insertions and deletions (INDELs) and identified clear evidence for 1,058 insertions and 1,021 deletions with average sizes of 57.4kb and 45.7kb, respectively (Supplementary Table 4). To further improve sequence contiguity and haplotype phasing, the PacBio contigs were corrected, joined, ordered, and oriented according to the optical mapping data. This generated a set of 558 optical-map-supported scaffolds spanning 634.1 Mb with a scaffold N50 of 2.25 Mb for TME 3. For 60444, we generated 552 scaffolds spanning 714.7 Mb with an even higher scaffold N50 of 2.35 Mb.

In cassava, a single bi-parental cross rarely yields enough progeny to generate a robust and dense genetic map that can be applied to chromosome anchoring. The most recent publicly available composite genetic map was generated from ten populations and anchors only 71.9% of an earlier haploid genome assembly [78]. *In vitro* proximity ligation as an application of chromosome conformation capture technologies can facilitate chromosome-scale genome assembly [59], [61], [62], [79]. To re-construct the set of cassava chromosomes independently of a composite genetic map, we constructed chromosome proximity interaction (Hi-C) libraries [80]. We combined the optical-map-improved hybrid-scaffolds with the remaining contigs and used the HiRise software pipeline (Dovetail Genomics) for scaffolding. Based on the proximity ligation data, we grouped the sequences into major chromosomal interaction bins. The HiRise pipeline could connect, orient and join 6,631 sequences in TME 3 and 5,998 in 60444, and increased sequence contiguity nearly 25-fold for a final scaffold N50 of 53.4 Mb in TME 3 and 59.2 Mb in 60444. Remarkably, Illumina sequencing of two 150 bp Hi-C libraries allowed the ordering, orienting and assembly of chromosomal arms and even whole chromosomes. To assess the quality of the Hi-C-based chromosomal scaffolds, we aligned the genetic markers from the composite genetic map [78] with the final version of the genomes. Out of 22,403 genetic markers, we were able to align 22,341 (99.7%) with the 60444 and 22,373 (99.8%) with the TME 3 genomes. To visualize and validate the chromosomal scaffolds, we plotted the genetic distance against the physical distance for each genetic marker. The data for Scaffold_176 of the TME 3 genome is shown in Figure 1b as an example. The markers in this plot were anchored to chromosome (chr) 9 of the composite genetic map with broad agreement between the physical scaffold and genetic distance. Plotting the recombination rate using a sliding window of 1 Mb across the whole Scaffold_176 revealed the expected decrease in recombination frequency in the center of the scaffold, as well as the presence of other regions with low recombination in the chromosome arms (Figure 1c). We generated similar plots for all scaffolds and confirmed that the chromosomes were assembled without large inter-chromosomal re-arrangements (Supplementary Figure 4). Based on the Hi-C data, we identified only 30 miss-assemblies in the TME 3 genome and 16 in the 60444 genome. Each miss-assembly was validated manually by testing Hi-C read-pair alignment position and alignment depth, and scaffolds were split accordingly (Supplementary Figure 5). We also found inconsistencies between the composite genetic map and our HiRise-scaffold assemblies. These inconsistencies will have to be addressed by generating a robust, dense cassava genetic map using extended mapping populations. The proximity maps presented here will be valuable for quality assessment of the composite genetic map and to improve the sequence resolution in regions that are seemingly devoid of meiotic recombination.

TEs and repeats are involved in shaping genome evolution and gene regulatory networks [81]. But short read-based assemblies often underestimate and misclassify the proportion of TE and repetitive DNA given in a genome assembly. In contrast, long-read sequencing generates reads that can span and resolve entire TEs and repeats [82]. Using *de novo* generated cassava repeat libraries, we annotated up-to 2.5 times more TEs compared to earlier reports [39], [42], [49]. In the TME 3 and 60444 assemblies, we annotated 602.90 Mb (64.81%) and 633.93 Mb (64.91%) as repetitive sequences, respectively (Figure 2a). We also investigated the spatial distribution of sequence repeats along the entire 60444 chromosomal Scaffold_1583

corresponding to the whole chromosome 9 (Figure 2b) and generated density maps for the four predominant TE categories. Long terminal repeat (LTR) retrotransposons had higher densities around the centromeric region, while non-LTR retrotransposons elements (LINE and SINE) were clustered in telomere-proximal regions. Class II DNA-transposons were more equally distributed across that scaffold. A similar distribution of TEs was reported for other complex plant chromosomes confirming the high quality of sequences ordered through Hi-C [62], [83]. The adoption of long read sequencing enabled the detailed characterization of repetitive elements and revealed a surprisingly high repetitive DNA proportion in cassava (65%) that now can be placed between other sequenced high-quality complex crop genomes such as sorghum (54%) [84], quinoa (64%) [85] or barley (81%) [62] (detailed TE annotation can be seen in Supplementary Table 6).

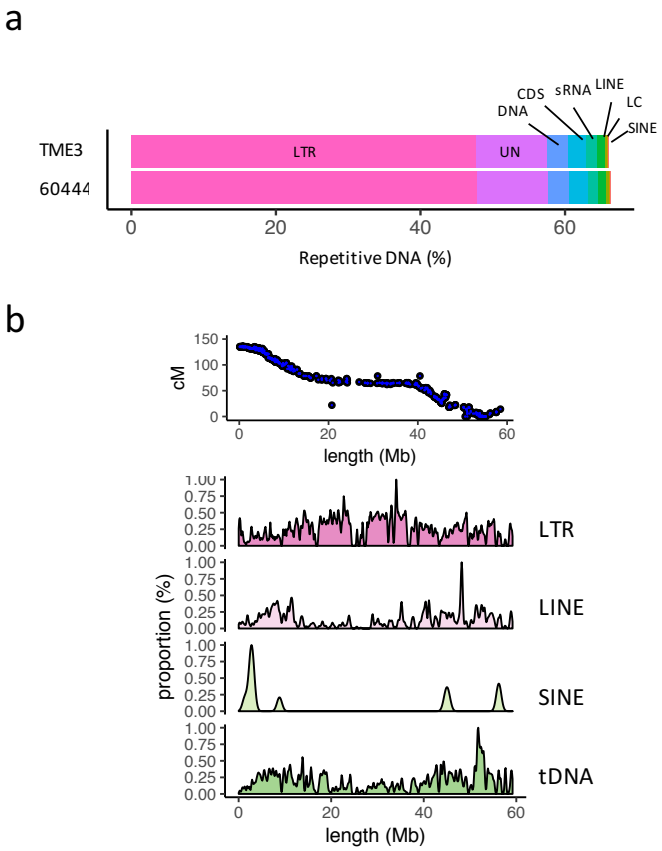


Figure 2 Distribution of key repetitive elements present in two cassava genomes. (a) Percentage of base pairs of the assembled TME3 and 60444 genomes that represent Long Terminal Repeat (LTR), Unclassified Repeat (UN), DNA transposon (tDNA), protein coding genes (CDS), short RNA (sRNA), Long Interspersed Elements (LINE), low-complexity element (LC), and Short Interspersed Nuclear Elements (SINE) sequences. (b) Graphical representation of SNP markers (top) and chromosomal density plots for the four predominant TE categories (bottom) on the physical 60444 chr9 map.

We predicted protein-coding and non-coding microRNA sequences using a combination of *ab initio* prediction and transcript evidence from publicly available cassava gene models [49]. Using Iso-Seq (high-quality, full-length cDNAs from single-molecule sequencing) data that covered 15,478 (45.7%) gene loci in

TME 3 and 16,057 (47.0%) in 60444, we determined the high accuracy of gene models (Supplementary Figure 6). The quality of the gene model annotation was assessed for 1,440 conserved plant genes using the BUSCO method [86]. We found 95% of the single-copy conserved orthologs in both genomes, with only 20 and 19 partially assembled in TME 3 and 60444, respectively (Supplementary Table 8).

To further assess the completeness of the two cassava genomes, we aligned the publicly available cassava coding DNA sequences (CDS) [49] to each of the assembled optical map-curated PacBio assemblies. Of the 41,381 CDS, 99.93 % are present in the 60444 and TME 3 genomes with only a few missing (84 and 86, respectively). We used the same CDS alignment to evaluate the haplotype phasing and allele distribution and counted when at least 50% of a CDS were aligned. Local gene duplications were excluded from this analysis. In total, we detected 18,831 and 19,501 multi-copy gene loci in TME 3 and 60444, respectively, with the vast majority of copies aligning two times (n=12,759 for TME 3 and n=13,425 for 60444) (Figure 3a). We found an increase of genes having four copies (4n=2,068 in TME 3 and 4n=2,194 in 60444), suggesting that these alleles remained present in both cassava genomes since the last whole genome duplication (WGD) event that for cassava was estimated ~35 million years ago [49]. A WGD became more evident with the analysis of synteny on a genome-wide scale (Supplementary Figure 7). Loss of one copy of a gene is common following a WGD. Remarkably, we found a high proportion of singleton CDSs (n=14,144 for TME 3 and n=13,479 for 60444), suggesting that the other gene copy was evolutionary purged because of functional redundancy or as a result of successful removal of deleterious mutant genes by recent breeding activities.

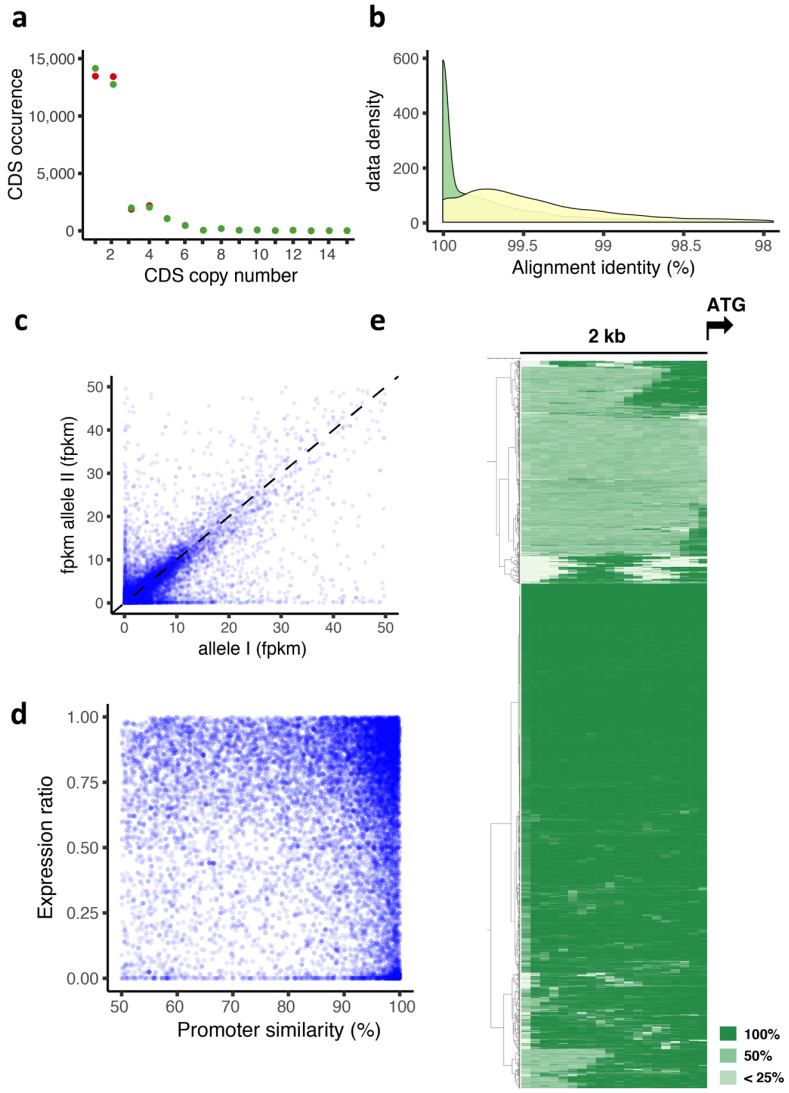


Figure 3 Allele phasing, allele nucleotide diversity and allele-specific expression analysis for diploid-aware cassava genome assemblies. (a) Cassava CDS collection (n=41,381) obtained from the AM560-2 genome and their alignment copy number distribution in the two cassava genomes 60444 (red points) and TME 3 (green points). (b) Sequence alignment properties for the bi-allelic reference CDSs (n=13,425) found in the 60444 genome. Alleles are presented as green curve and the homologous allelic counterpart as a yellow curve. Percentage of alignment identity is shown on x-axis and data density distribution on the y-axis (c) Scatterplot of allele-specific RNA read counts for 60444 measured as Fragments Per Kilobase of sequence per Million mapped reads (FPKM). A bi-allelic gene is depicted as a single blue dot and expression of one allelic copy is shown on y-axis and the expression of the homologous counterpart on the x-axis. (d) Promoter structure analysis for the same gene set. Expression ratio of 1.00 indicates an equal expression of both alleles, whereas expression ratio of <0.25 indicates mono-allelic expression. Promoter sequence similarity between the homologous promoter regions are shown on the x-axis measured in 100 bp bins for a two kb region upstream the start codon. (e) Promoter sequence comparison of all genes with mono-allelic expression (n=3,451). Promoter sequence comparisons are shown for a 2 Kb region upstream of the ATG start codon. Sequence divergence was compared in 100 bp bins. Sequence similarity ratio is shown on the right side of the plot.

The clonal propagation of cassava has resulted in a large proportion of genetically fixed deleterious mutations that affect crop vigor and limit breeding [65], [67], [87]. Purging these deleterious mutations from the cassava genome is key to maintaining and improving crop productivity. Duplicated regions are often subject to dynamic changes, including accumulation of point mutations [88]. To test this hypothesis for the multi-copy genes in the diploid 60444 and TME 3 genomes, we measured the nucleotide diversity for each allelic pair. This revealed an increase in single-base pair mutations occurring in one of the alleles (Figure 3b). To determine if the accumulation of allelic mutations has an impact on gene expression we measured the allele-specific expression using high-throughput RNA-seq analysis from eight sequencing libraries that originated from different tissues (for details see Supplementary Note). In total, we covered the expression of 18,723 alleles with two copies and identified 3,451 (14.43%) genes with strict mono-allelic expression (Figure 3c). Gene Ontology (GO) analysis of the mono-allelic expressed genes revealed an enrichment of genes involved in 'carbon-oxygen lyase activity' (GO:0016837) and 'cytochrome-C oxidase activity' (GO:0004129). We further asked if mutations within the promoter region could cause the mono-allelic expression (Figure 3d and e). However, a high proportion of the genes (44.76%) had intact promoter sequences between the alleles, indicating that monoallelic expression of these genes might be epigenetically regulated through methylation or chromatin packaging. Cassava has a more robust maintenance methylation mechanism than other plant species [48]. The high number of silenced alleles could be another property of cassava genomes that was maintained through clonal propagation of the crop over many generations. Further research is needed to determine if mono-allelic expression of genes is promoting or depressing cassava vigour and productivity.

The two high-quality genomes enabled us to investigate the gene family expansion specific for the two cassava cultivars 60444 and TME 3 using MCL clustering of all gene models present in our two assemblies, the assembly of AM 560, the assembly of *Ricinus communis* as a close relative of cassava and *Arabidopsis* as an outgroup [89], [90]. This confirmed that the two African cassava varieties are closely related (Figure 4a). For example, there were fewer gene family groups specific to 60444 or TME 3 (0.8-1.1%), whereas the number of specific gene family groups was considerably larger for *Ricinus* and *Arabidopsis*. Interestingly, there were more protein groups associated exclusively with AM560 and *Ricinus* than with *Ricinus* and either 60444 or TME 3. These trends were also seen for predicted enzymatic reactions (Figure 4b) and predicted metabolic pathways (Figure 4c) but overall species were more similar when looking at reactions and even more when considering pathways. There were 1,823 protein groups containing 4,081 gene models (2,067 for 60444 and 2,014 for TME 3) specific for the two African cassava genomes. Considering the short evolutionary time since cassava was introduced to Africa about 400 years ago, it is likely that the differences in gene divergence and expansions between AM560, 60444 and TME 3 evolved before the ancestors of the two African cassava varieties were brought to the African continent.

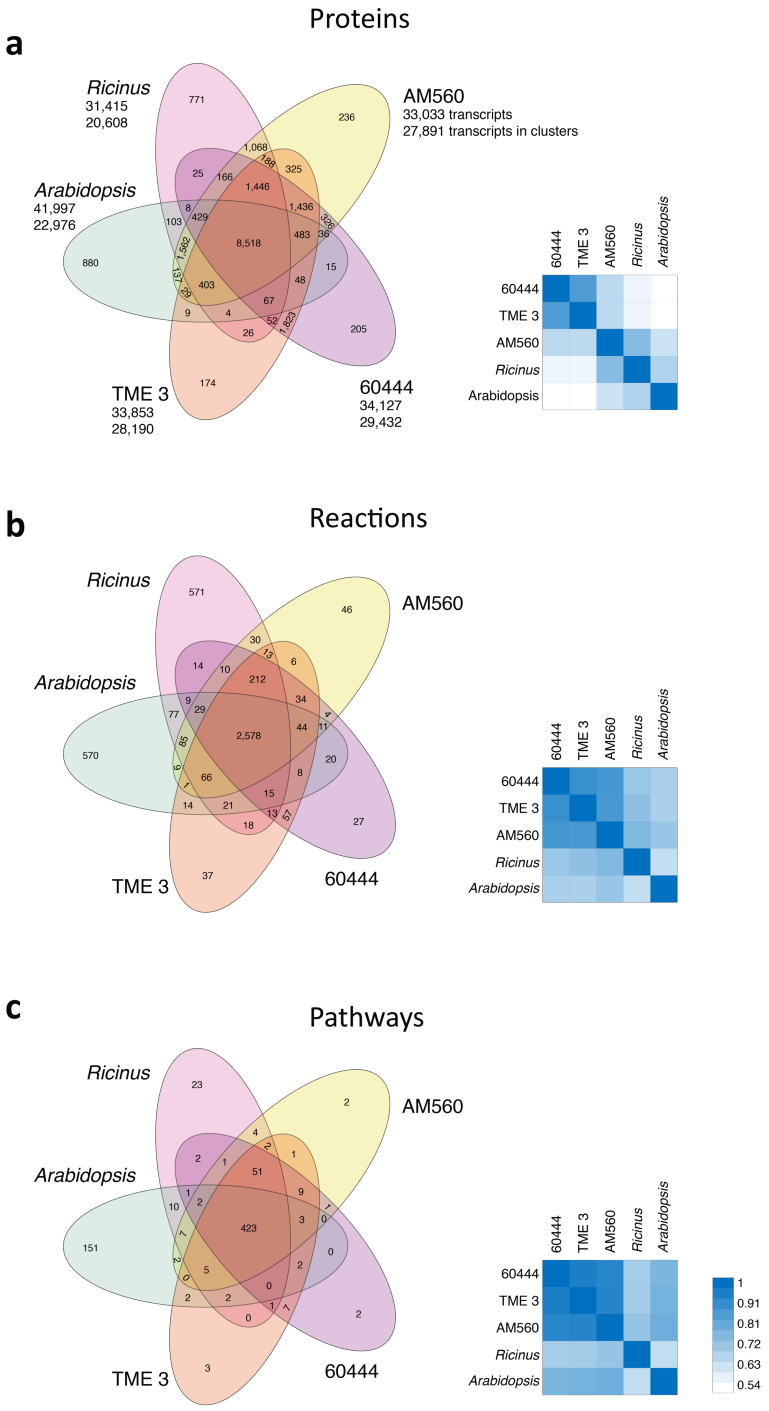


Figure 4 Expansion of gene clusters, enzymatic reactions and metabolic pathways. (a) Associations of protein groups using OrthoMCL clustering, predicted metabolic reactions (b) and metabolic pathways (c) of the three cassava genomes (AM560, TME 3 and 60444) as well as their close relative *Ricinus communis* and Arabidopsis as outgroup. Numbers in the sections of the Venn diagram correspond to the number of cluster groups. The first number below the cultivar name denotes the total number of proteins that were included into the OrthoMCL analysis. The second number indicates the number of genes in protein clusters. Left representation as a Venn diagram depicting shared and not shared elements, right representation as a heatmap where the intersection of elements between two species was divided by the union of their elements.

We then investigated the genes associated with gene families occurring in the different set of Figure 4a for over-representation of GO terms [91]. For AM560 we found cultivar specific proteins with GO terms enriched for 'polygalacturonase activity'. Among the most significantly enriched GO terms for genes that were associated exclusively with the African varieties were GO terms with categories 'structural integrity of ribosomes' (GO:0003735) and 'structural molecule activity' (GO:0005198). Another, but more specific function was squalene monooxygenase activity (GO:0004506). Squalene monooxygenase convers squalene to (3S)-2,3-epoxy-2.3-dihydrosqualene (epoxysqualene) which itself is a precursor for many specialized metabolites. Both in 60444 and TME 3, there are four metabolic pathways predicted that are involved in the metabolism of epoxysqualene leading to several specialized metabolites. Some have known antimicrobial, anti-inflammatory and/or anti-tumor activities, including the pathway producing beta-amyrin as an intermediate which can be converted to oleanolate that has antiviral activity [92]. The pathway from squalene to oleanolic acid contains three consecutive reactions, all of which have gene annotations in all three cassava varieties. The two African varieties 60444 and TME 3 that are targeted by African Cassava Mosaic Viruses, however, have an expanded gene pool for two of the three ractions in the pathway (Supplementary Figure 8). Enzymes in specialized metabolic pathways can be encoded genes that are physically co-located on chromosomes (metabolic gene clusters) [90]. Although we could not confirm such metabolic gene clusters, genes associated with 'squalene-monooxygenase activity' were not randomly distributed in the genome. Both in 60444 scaffold_1262 and TME3 scaffold_3 contain 10 or more consecutive genes predicted to encode enzymes for the same reaction, with similar duplicated genes on other scaffolds. AM560 does not contain such clusters of genes encoding the same or similar enzymes potentially producing precursors for antiviral compounds, suggesting that locus-specific expansion of these genes may have been selected by cassava farmers and breeders as adaptions to viral pathogens.

We expect that the diploid-aware assemblies of the 60444 and TME 3 cassava genomes based on optical- and proximity-maps will facilitate unlocking the limited genomic diversity of African cassava cultivars for crop improvement. The genome assembly strategy reported here can be similarly adapted to other medium-sized, non-inbred genomes with high heterozygosity and DNA repeat richness. Using the information for haplotype-phased alleles and allele-specific expression, it will be possible to characterize and to purge deleterious mutations using targeted genome editing [93] , conventional breeding, or genomic selection. Moreover, the TME 3 and 60444 genomes will greatly facilitate trait mapping and map-based cloning of agriculturally important genes in this important food security crop.

Author contributions

W.G., J.K. and H.V. conceived and designed the research, J.K prepared DNA samples for PacBio SMRT sequencing, A.P. performed PacBio library preparation and sequencing, J.K. and L.P. generated the BioNano optical genome maps, J.K. and S.G. generated Hi-C libraries with advice from U.G., W.Q performed the PacBio genome assemblies, J.K. generated the Iso-Seq libraries and A.P. performed Isoform Sequel sequencing. M.H.H. and J.K. analyzed allele-specific expression data. W.Q performed gene space annotation. J.K. performed transposable element analysis. J.K. and P.S. analyzed GO-annotation. P.S. performed enzyme and pathway prediction and their analysis. M.H.H. contributed to the data release. J.K and W.G. analyzed the data and wrote the paper.

Acknowledgement

The work was supported by the Bill & Melinda Gates Foundation. We acknowledge Anna Bratus for the technical support at the Functional Genomic Center Zurich (FGCZ). We thank John Baeten (BioNano Genomics), Bo Xue and Peifen Zhang (creation of PGDBs) for technical assistance and the Dovetail Genomics team for analyzing the *in vitro* proximity ligation data. We thank Irene Zurkichen for technical support. We acknowledge Rebecca Bart, Dario Copetti and Alexis Sarazin for helpful scientific discussions.

Methods

Further details of all methods are presented in Supplementary Note. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Long-read sequencing and sequence assembly

To sequence the two cassava genomes with long reads, we extracted high-molecular weight (HMW) genomic DNA from 3-weeks old leaf tissue of *in vitro* grown cassava lines following a modified protocol [94]. Libraries for PacBio SMRT sequencing were generated as described previously [53]. Libraries were sequenced using a PacBio RSII instrument with P6C4 sequencing reagents. We used 47 SMRT cells for TME 3 and a total of 45 SMRT cells for 60444. For 60444 we generated a total of 52.4 GB with subread bases with a mean read length of 12.8 kb. For TME3, 53.9 GB of subread bases were generated with a similar mean read length of 12.4 kb. The PacBio sequences had a > 70-fold genome coverage.

De novo assembly of the subreads was performed applying three assemblers: The PBcR-MHAP pipeline [95], the CANU-MHAP assembler [54] and the FALCON (v0.5) assemblers [52]. For FALCON, we adopted parameter sweeping and the assembly with the largest N50 was retained. For the other assemblers, default parameters were used, except the expected haploid genome size was set to values estimated by flow-cytometry as well as k-mer analysis (Supplemental Note). Quiver from SMRT Analysis v2.3.0 was run two times to polish base calling of assembled contigs [50].

Optical map constructions

Long-range scaffolding of the assembly contigs with optical mapping was achieved using the Irys optical mapping platform (BioNano Genomics). HMW DNA was isolated from 3-weeks old leaf tissue of *in vitro* grown cassava line TME 3 and 60444, embedded in thin agarose plugs according to the IrysPrep Kit and the plant tissue DNA isolation protocol (BioNano Genomics). DNA molecules were labeled using the *NT.BspQI* DNA-nicking enzyme by incorporation of fluorescent-dUTP nucleotides according to the IrysPrep nick-and-repair protocol (BioNano Genomics). DNA samples were aliquoted and quantitated using the Qubit Fluorimeter run in broad-range mode. The final samples were then loaded onto the IrysChips, linearized and visualized by the BioNano Irys molecule imaging instrument. Molecules >150kb were assembled *de novo* using the pairwise assembler provided by the IrysView software package (BioNano Genomics) with p-value threshold of 10^{-9}.

Three-dimensional genome-wide chromatin capture sequencing

Freshly harvested leaves of in vitro grown TME 3 and 60444 cassava plants were vacuum infiltrated in nuclei isolation buffer (NIB) supplemented with 2% formaldehyde. Protein-crosslinking was stopped by adding glycine and applying an additional vacuum infiltration step. Leaf tissue was snap-frozen using liquid nitrogen and ground into a fine powder, re-suspend in NIB and purified by spin-downs as described earlier [80]. Nuclei were digested with 400 units of *HindIII* as described in [80]. Digested chromatin was labeled using a fill-in reaction with 60 units of Klenow polymerase and biotin-14-dCTP. The exonuclease activity

of T4 DNA polymerase was used to remove biotin-14-dCTP from non-ligated DNA ends. Proteinase K was added to reverse the formaldehyde cross-linking and DNA was purified following phenol-chloroform extraction [80]. The Hi-C samples were quality assessed by PCR amplification of a 3C template and evaluated according to [80] (Supplementary Figure 3). Quality control passed Hi-C samples were purified following a phenol-chloroform extraction protocol introduced elsewhere [80] and mechanically sheared to fragment sizes of 300bp using a Covaris S2 sonicator. Hi-C library fragments were blunt-ended using the End Repair Mix from Illumina and finally purified using the AMPure beads according to the standard AMPure protocol. The biotinylated Hi-C samples were enriched through biotin-streptavidin-mediated pull-down and adenylated using Illumina's A-tailing mix. Illumina paired-end sequencing adaptors were ligated to the Hi-C fragments and a PCR amplification of the Hi-C library was carried on as suggested earlier [80]. Finally, PCR products were purified with AMPure beads following the standard AMPure protocol and quantified using a Q-bit device. Samples were sequenced using the Illumina HiSeq 2500 instrument. This produced 385 million pairs of 150bp reads for 60444 and 391 million reads for TME 3. Genome scaffolding was performed with Dovetail Genomics' HiRise scaffolding software.

Assembly accuracy estimation, repeat identification and gene annotation

Public available WGS Illumina paired-end reads (SRX1393211, SRX526747) were trimmed and quality filtered using Trimmomatic [96] and mapped to the draft assembly using BWA ALN (v0.7.12) [97] with default parameters. WGS read-mapping files were sorted using SAMtools SORT [98] statistics called using QUALIMAP BAMQC [99]. To assess the assembly completeness, the set of reference CDSs (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Mesculenta) was aligned to each of the assembled draft genome using GMAP [100] with option '-no fails' and 'min-identity 0.5'. Results were further filtered for alignments covering >99% of query sequence using a custom script.

Repeat families found in the two draft genome assemblies of 60444 and TME 3 were first independently discovered *de novo* and structure classified using the software package REPEATMODELER ver. 1.0.9 and REPEATMASKER ver. 4.0.7 (www.repeatmasker.org). To screen for large tandem repeats, we used the software package RefAligner from Bionano with the option '-simpleRepeat –simpeRepeatTolerance 0.1 –simpleRepeatMinEle 3'.

To annotate the gene space, we did iterative MAKER analysis. In the initiate analysis, the gene prediction tool AUGUSTUS [101] was trained with reference gene models. The predicted gene models were combined with alignment base evidence, including all ESTs from cassava found on NCBI (https://www.ncbi.nlm.nih.gov/nucest/?term=cassava%20ESTs), Iso-Seq data, and UniProt protein sequences. The initiate set of MAKER gene models were used to trained gene predictor SNAP, which was added in the second round of MAKER analysis, together with gene predictor GeneMark trained using Iso-Seq data. Putative gene functions of the final set of gene models were characterized by performing a BLAST search of the protein sequences against the Uniprot database (ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/uniprot/). PFAM domains, InterProScanID and Gene Ontology

annotation were obtain by running interproscan[102]. To annotate non-protein coding genes, the tools tRNAscan-SE [103] and Infernal [104] were used together with the Rfam version 13.0 database.

Allele-specific expression analysis and promoter region comparison

Newly generated RNA-seq datasets were derived from three key developmental stages of cassava 60444: Early stage plant with fibrous root (FR) and leaf, middle stage plant with leaf, FR and intermediate root (IR) and late stage plant with leaf, FR, IR. RNA-seq libraries were sequenced on Illumina HiSeq2000 in paired-end 2 x 100 nucleotides mode. We aligned the RNAseq reads using STAR [105] and retained the unique alignments. Read were counted using SAMtools and custom made scripts [98].

The promoter region was characterized for genes with two alleles and fpkm expression ratio >0. Sequences 2 kb upstream of the start codon were defined as promoter. A pairwise alignment was generated for each allele pair using the MUSCLE pairwise alignment tool [106]. Alignments were analysed using 100 bp bins and a similarity ratio was calculated using a custom script. and visualized using the INCHLIB cluster and heat map tools [107].

Genome wide comparison and structural variation detection

To compare the two assemblies on a genome wide scale, we used the optical maps of the two cassava lines to detect structural variations (SVs) using the RunBNG software[108]. We used the maps from 60444 as the reference and TME 3 as query. RunBNG acts as a wrapper and essentially uses the BioNanos' RefAligner for generating the alignments. Alignments were then screened by the script 'SVdetect' to detect the intergenomic SVs and to calculate the insertion size and deletion size [53]. Synteny was analyzed using the CoGe platform (https:// genomevolution.org/). Syntenic regions between 60444 and TME 3 were identified using CoGe SynMap tools.

Gene family analysis

To investigate the gene family expansion specific for the two cassava cultivars 60444 and TME 3 using OrthoMCL clustering of all gene models present in our two assemblies, the assembly of AM 560, the assembly of *Ricinus communis* as a close relative of cassava and *Arabidopsis* as an outgroup was used [89], [90]. Moreover, Only the longest protein sequence was used and datasets were filtered for internal stop codons. Pairwise sequence similarities between all input protein sequences were calculated using BLASTP[109] with an e-value cut-off of 10^{-5}. Clustering of the resulting matrix was used to define the orthology cluster with an inflation value set to 1.5. Over-and under-representation of Gene Ontology (GO) terms between the three cassava genomic compartments were calculated with a hypergeometric test using the functions GO stats and GSEABase from the Bioconductor R package [110]. The REVIGO [111] package was used to remove redundant and similar terms from long Gene Ontology lists by semantic clustering and to visualize the enrichment results.

Enzyme prediction and pathway prediction was performed as published earlier [90]. Databases can be downloaded and will subsequently further developed and refined by PMN (plantcyc.org).

Data availability

The PacBio Raw reads, the Hi-C sequences, the Iso-Seq reads, optical maps and genome annotations and gene models will be deposited at the NCBI under a specific BioProject number. All other data are available from the corresponding author upon a reasonable request.

Supplementary Notes

Plant material

We sequenced *Manihot Esculenta* (cassava) accession TME 3 (also known as Tropical Manihot Esculenta) and cassava accession 60444. TME 3 was originally collected in farmers' fields of Nigeria and other West African countries during the 1980s and 1990s [33]. TME 3 is considered as the origin of the monogenic dominant resistance gene *CMD2* conferring wide resistance against all known *cassava mosaic begomoviruses*. Because of its simplicity, *CMD2* became the predominant resistance source deployed in African cassava breeding programs despite its underlying molecular mechanisms remained unknown [34], [35]. Cassava accession 60444 is often considered as the cassava model cultivar with showing the highest transformation rate [112], [113]. Shoot cultures of 60444 and TME 3 were obtained from the ETH Zurich *in vitro* cassava germplasm collections.

Additional details for PacBio library preparation, PacBio RSII sequencing and PacBio assembly

High-molecular weight (HMW) genomic DNA was extracted from three-weeks old plantlets grown under sterile, *in-vitro* jars with CBM media [112] according a modified CTAB method [94]. DNA integrity was assessed by a standard agarose gel electrophoresis and Thermo Fisher Scientific Qubit Fluorometry (Invitrogen). PacBio 20kb SMRTbell libraries were generated as recommended previously [53]. SMRT-Libraries were sequenced using a PacBio RSII long read sequencing device with P6C4 sequencing reagents. In total, we used 47 SMRT cells for TME 3 and 45 SMRT cells for 60444. We generated 5,777,131 subreads for 60444 with a read length N50 of 12,813 kbp and 52,4 Gbp total length. For TME 3, we generated 7,650,003 subreads with 12,424 kbp read N50 and total length of 53,9 Mpb.

De novo assembly of the subreads was performed with three assemblers: The PBcR-MHAP (PBcR) pipeline [95], the CANU-MHAP (CANU) assembler [54] and the FALCON (v0.5) assemblers [52]. PBcR assembly was performed with the estimated genome size set to 500 Mb for both genomes, estimated by the assembled size of the reference genome. With the FALCON assembler, we did parameter sweep and choose parameters to maximize contig N50. For CANU assembly, the estimated genome size was set to 527 Mb and 633 Mb for 60444 and TME 3, respectively. Both values were estimated using kmer analysis of Illumina paired-end reads. Assembled drafts were benchmarked using Illumina paired-end data and reference gene models. Selected drafts were then polished using PacBio raw reads (in h5 format) with two rounds of quiver correction

Genome assembly validation

To assess the quality of the genome assemblies publicly available paired-end (PE) short reads (60444 WGS:SRX1393211, TME 3 WGS:SRX526747) were aligned to the representative drafts. In brief, sequencing adapter were trimmed using trimmomatic (v.033)[96] and PE-reads were mapped using bwa

aln [97]. Mapping statistics were collected using Samtools (v1.3) [98] and Qualimap2 (v2.2.1)[99]. Of the 409,126,944 Illumina short reads from 60444, 98.3% of the reads were successfully mapped back to the CANU assembly, with 96.6% properly paired. For TME 3, we were able to map 96.4% of the 568,006,046 reads back to the assembly with 93,4% properly paired. For the FALCON assemblies, we received overall lower mapping values. For 60444, we mapped 96,1% with 90% properly paired reads. For TME 3, we aligned 93.8% and 86% properly paired. The drafts produced by PBcR were very fragmented and showed lowest Illumina read mapping rate. Thus, it was excluded from further analysis. We then generated an *in silico* map for sequence contigs from both assemblers. The maps were aligned against the corresponding optical maps using RefAligner software (BioNano Genomics) to identify and curate potential conflicts in the contigs or in the optical maps. The result showed that the CANU assemblies had the lowest number of conflicts. When compared against FALCON assemblies, CANU introduced less assembly errors but produced a slightly more fragmented draft (lower N50 values). The parameter sweep aiming at largest N50 during FALCON assembly might have been too aggressive and thus increased error rate.

Optical map construction

Cassava plants were *in vitro* grown for three weeks and then placed in the dark for two days. HMW DNA was isolated according to the standard BioNano protocol 'IrysPrep Plant Tissue DNA Isolation User Guide'. Briefly, DNA was digested by the single-stranded nicking endonuclease Nt.BspQI and labelled with a fluorescent-dUTP nucleotide using the *Taq* polymerase. The nicks were ligated using the *Taq* DNA ligase and the DNA backbone strained with using the YOYO-1 dye for backbone staining. DNA imaging was done automatically using the BioNano Irys instrument. Molecules > 150 kb (and more than eight labels) were assembled into consensus physical maps using the BioNano IrysView analysis software. We used the IrysView pre-adjusted option 'optArguments_human' for assembly.

Genome diversity analysis

To compare the two assemblies on a genome wide scale, we used the optical maps of the two cassava lines to detect structural variations (SVs) using the RunBNG software [108] and the reference map from 60444 and TME 3 as query. RunBNG acts as a wrapper and essentially uses the BioNano program RefAligner for generating the. alignments Alignments were generated using the option '-z 1200Mb -t 1 -m 4' screened by the script 'Detect' to detect the intergenomic SVs and to calculate the insertion size and deletion size [53]. Genome synteny between the two cassava genomes was analyzed using the Snap tools (CoGe, www.genomevolution.org). To identify collinearity blocks using homologous CDS pairs the following parameters were applied: Maximum distance between two matches (-D) was set to '20'. Minimum number of aligned pairs (-A) was set to '10'. The algorithm 'Quota Align Merge' was set with Maximum distance between two blocks (-Dam) '500'.

Three-dimensional genome-wide chromatin capture sequencing

We used five grams of freshly harvested leaves of *in vitro* grown TME 3 and 60444 plantlets that had been placed in the dark 48h before tissue harvest. Leaf material was vacuum infiltrated in nuclei isolation buffer (NIB) supplemented with 2% formaldehyde. Protein-crosslinking was stopped by adding glycine and applying an additional vacuum infiltration step. Leaf tissue was snap-frozen using liquid nitrogen and ground into a fine powder, re-suspend in NIB and purified by spin-downs as described earlier [80]. Nuclei were digested with 400 units of *HindIII*[80]. Digested chromatin was labeled using a fill-in reaction with 60 units of Klenow polymerase and biotin-14-dCTP. The exonuclease activity of T4 DNA polymerase was used to remove biotin-14-dCTP from non-ligated DNA ends. Proteinase K was added to reverse the formaldehyde cross-linking and DNA was purified following phenol-chloroform extraction as described earlier [80]. The Hi-C samples were quality assessed by PCR amplification of a 3C template and evaluated following as published earlier [80] (Supplementary Figure 3). Quality control passed Hi-C samples were purified following a phenol-chloroform extraction protocol introduced elsewhere[80] and mechanically sheared to fragment sizes of 300 bp using the Covaris S2 sonicator. Hi-C library fragments were blunt-ended using the End Repair Mix from Illumina and purified using the AMPure beads according to the standard AMPure protocol. The biotinylated Hi-C samples were enriched through biotin-streptavidin-mediated pull-down and adenylated using Illumina's A-tailing mix. Illumina paired-end sequencing adaptors were ligated to the Hi-C fragments and a PCR amplification of the Hi-C library was carried according to the Illumina protocol. Finally, PCR products were purified with AMPure beads following the standard AMpure protocol and quantified using a Q-bit device. Samples were sequenced using the Illumina HiSeq 4000 instrument. This produced 385 million 151 bp paired-end reads for 60444 and 391 million reads for TME 3 providing 51.3x and 52.1x physical coverage, respectively. To assess the quality of the Hi-C sequencing, sequence reads were quality filtered using the HiCUP pipeline, a software specifically designed to filter proper Hi-C read pairs from paired-end read contaminations [114]. This revealed 17.9 million unique and valid Hi-C pairs for 60444 and 20 million valid pairs for TME 3.

Scaffolding the PacBio and BioNano assemblies with HiRise

Hi-C sequence data was used to scaffold the two cassava assemblies using HiRise, a software pipeline designed for using proximity ligation data to assemble sequences into chromosomal pseudo-molecules [115]. The mapping location of Hi-C read pairs were analyzed by HiRise to cluster sequences into large proximity bins. The read-pair position was also used to identify putative assembly errors.

Genome size and heterozygosity estimation

We measured the nuclear DNA content of the two cassava genotypes by flow cytometry. Two weeks old, in-vitro grown plants were processed together with the internal reference standard tomato (Lycopersium esculentum, cv. Stupicke with genome size of 958 Mb)(Dolezel, Sgorbati, and Lucretti 1992). The cassava

haploid genome size was estimated from a relative peak position using the CyStain PI absolute P kit and CyFlow Space provided by Partec. For 60444, we obtained a haploid genome size of 745 Mb and for TME 3, we estimated the genome size to be 768 Mb (Supplementary Figure 2).

To assess the heterozygosity of the two cassava lines, we used the public available Illumina paired-end 100bp sequencing reads from 60444, TME 3 and the AM560-2, the partly-inbred cassava reference genome, which were downloaded from NCBI Short Read Archive (SRX1393211, SRX526747, SRX1393218). Illumina reads were trimmed using the trimmomatic tools (Bolger, Lohse, and Usadel 2014). Genome properties were analysed using SGA Preqc (Simpson 2014) with default parameters**.**

Iso-Seq preparation

For the full-length transcript sequencing, RNA was extracted from the following greenhouse-grown 60444 and TME 3 samples: Top five leaves with petioles, the apical meristem, lateral meristems, stems and roots. Tissue was snap-frozen in liquid nitrogen and ground using a mortar and pestle. RNA was isolated using a modified protocol [116] and RNA integrity was tested on a Agilent 2100 BioAnalyzer and Qubit Fluorometry (Invitrogen). A subset of the RNA sample was pooled and processed according to the PacBio Protocol: Procedure & Checklist – Iso-Seq Template Preparation for Sequel Systems (11/2017). The optimal number of cycles for large-scale PCR was determined to be 14. Amplification was followed by molecule size selection using 1x AMPure beads and 0.4x AMPure beads. The two purified fractions were pooled for library construction. We used one SMRT cell for each cassava line and sequenced using the PacBio Sequel instrument. A total of 181,823 reads covering 2,779,884,989 bp and 296,109 reads covering 3,768,451,277 bp was produced for 60444 and TME 3 RNA libraries, respectively. The raw sequencing reads were processed using the Iso-seq protocol within SMRTlink (v.5.0.1.9585) to obtain full length transcripts, which were error corrected using the Arrow algorithms provided by PacBio. Isoform were aligned to the corresponding cassava genome using GMAP with option '-f samse' and '-z sense_force' and '-n 0'[100]. The isoform alignments were used as input for the gene model annotation as described in the chapter 'Gene Space Annotation'.

Repeat sequence annotation and characterization

Repeat families found in the two cassava genome assemblies were first independently identified *de novo* and classified using the software tool RepeatModeler [117]. RepeatModeler uses the programs RECON and the package RepeatScout for the *de novo* identification of repeats. After the classification process, the output data file from each of the genome assembly was used as a custom repeat library by RepeatMasker [118] for the discovery and annotation of repetitive DNA elements. Detailed results are shown in Supplementary Table 6.

Gene space annotation

Protein coding genes were annotated using iterative MAKER analysis. In the initial analysis, Augustus [101] , trained with the cassava reference gene models, was used for the *ab initio* prediction of gene models, which were combined with three different alignment base evidence, including the public available cassava ESTs from NCBI, the full-length generated from Iso-Seq and the uniprot protein sequences [119], to produce the initial set of gene models. These models were used to train the *ab initio* gene predictor SNAP, which was added in the second round of MAKER analysis. At this step, the *ab initio* gene predictor GeneMark trained with Iso-seq data was also included. The final gene models were annotated using six different evidence sources: the gene models from Augustus, SNAP and GeneMark, the cassava ESTs, the full-length transcriptome sequences and uniprot protein sequences. To assess the quality of the gene prediction, the AES scores were generated for each of the predicted genes throughout the annotation pipeline. Genes were further characterized for their putative function by performing BLASTp[109] search against the UniProt database. Gene Ontology (GO) annotation was performed using InterProScan. To annotate non-protein coding genes, the tools tRNASCAN-SE [103] (Version 2.0) and INFERNAL [104] (Version 1.1.2) were used with the Rfam database (verison 13.0) (Supplementary Table 7). Genome assembly and annotation completeness was assessed using the embryophyta_odb9 database of 1,440 single copy orthologs using BUSCO [86] run with option '-m genome -long' (Supplementary Table 8).

OrthoMCL clustering and GO over-/under-representation

Gene clusters were established from the annotated gene set of the three cassava genomes 60444, TME 3 and AM560, *Ricinus communis* and *Arabidopsis* using the OrghoMCL software tools (v2.0) [89] (www.phytozome.com). Splice variants were removed from the protein data set and proteins were filtered for internal stop codons. The input dataset comprised 33,853 TME 3 proteins, 34,127 60444 proteins and 33,033 AM560 proteins. First, pairwise sequence similarities between all input coding sequences were defined using BLASTP and a e-value cut-off of 1e-05. The markov clustering was used to define the ortholog cluster structure using the default inflation value of 1.5. A total of 101,013 proteins from the three different genomes were clustered into 17,648 gene families. A set of 11,910 clusters contained coding sequences from all three cassava genomes.

Cultivar specific genes and genes shared between 60444 and TME 3 were extracted from clusters and tested for gene ontology (GO) enrichments or under-representations using a hypergeometric testing available in the GOstats and GSEABase function from the Bioconductor R package [110]. The REVIGO tool [111] was used to remove redundant terms from long GO lists and to visualize enrichment results.

Allele specific expression analysis

For the deep transcriptome sequencing, green-house grown plant material of three key developmental stages from cassava 60444 were sampled as following: Early stage plant with fibrous root (FR) and top 3 leaves. Middle stage plant with top 3 leaves, FR and intermediate root (IR) and late stage plant with top 3 leaves, FR, IR and storage root (SR). We used three independent replicates per organ. RNA samples were prepared according to a modified protocol[116] and tested for integrity using Qubit Fluorometry (Invitrogen) and the Bioanalyzer 2100 (Agilent). High-throughput sequencing was performed on an Illumina HiSeq 2000 instrument run in paired-end 2x100 nucleotides mode. Reads were processed with Trimmomatic(v.35) [96] to remove adapter and low quality sequences (< 20 bp quality). Reads were mapped to the 60444 genome assembly using STAR (v2.5.3a) [105] and read duplicates marked using the '—bamRemoveDuplicates' with type 'UniqueIdentical'. Unspecific reads were removed from the mapping file using samtools[98] with option 'view -F0x400'. Allelic gene space was annotated using *de novo* CDSs aligned to the genome assembly using GMAP [100] run with the option '-nofails -min-identity=0.5 -f1'. Alignment positions were extracted using custom scripts and RNAseq reads counted using the samtools wrapper pysam and the module 'fetch' (https://github.com/pysam-developers/pysam) and custom python scripts. An expression ratio was calculated with dividing FPKM_alleleA by FPKM_alleleB. Mono-allelic expressed genes were defined when this expression ratio was < 0.25.

Supplementary Table 1 Assembly statistics of representative genome drafts from the three different assemblers.

		60444			TME 3	
	contigs	length (Mb)	N50(kb)	Contigs	length (Mb)	N50(kb)
CANU	11,459	975	117	12,971	947	98
FALCON	10,428	1.058	134	12,280	992	119
PBcR-MHAP	22,547	812	45	33,277	854	32

Supplementary Table 2 Assembly accuracy evaluation using public available Illumina paired-end reads

		Mapped (%)	Both mapped (%)	Properly paired (%)	Discordant reads (%)	Total reads (in Mio.)
60444	Falcon	0.96	0.95	0.90	0.050	409
	Canu	0.98	0.98	0.97	0.012	409
	PBcR-MHAP	0.92	0.90	0.85	0.056	409
TME 3	Falcon	0.94	0.92	0.87	0.047	568
	Canu	0.96	0.95	0.93	0.016	568
	PBcR-MHAP	0.90	0.87	0.82	0.052	568

Supplementary Table 3 Optical map assembly using the IrysView software provided by BioNano and using option 'optArguments_human'.

	TME 3	60444
Mapped Molecule Quantity (Mb)	64,060.011	70,148.008
Mapped Avg Size (Kb)	265	268
Avg Label Density (per 100 Kb)	9.7	9.5
Number of Consensus Genome Maps	952	926
Consensus Genome Maps Size (Mb)	1204.598	1204.106
Haploid-genome size estimation (Mb) based on flow-cytometry	765	745
Consensus Genome Maps N50 (Mb)	1.801	1.875

Supplementary Table 4 Structural variations from optical maps of two cassava lines

	60444 optical map (reference) vs TME 3 optical map (query)
Total size of genome map (Mb) TME 3	1204.6
Map aligned to 60444 genome (Mb)	974.3
Map uniquely aligned to 60444 genome (Mb)	612.03
Region in TME 3 with insertion and deletion (Mb)	107.24
Ratio of region with insertion or deletion (%)	8.9
Number of insertions	1,058
Average insertion size (bb)	57336.84
Number of deletions	1,021
Average deletion size (bb)	45615.34

Supplementary Table 5 PacBio Iso-seq full length-transcriptome sequence classification

	60444	TME 3
Number of reads of insert	181,785	296,047
Number of five prime reads	128,972	182,131
Number of three prime reads	133,388	187,096
Number of poly-A reads	123,772	173,033
Number of filtered short reads	6,028	16,526
Number of non-full-length reads	72,153	138,907
Number of full-length reads	103,604	140,614
Number of full-length non-chimeric reads	82,197	113,182
Average full-length non-chimeric read length (bp)	2,151	2,003

Supplementary Table 6 Structural annotation of transposable elements in 60444 and TME 3

		60444			TME3		
	Superfamily	Copies	Total size(Mb)	Assembly percentage (%)	Copies	Total size(Mb)	Assembly percentage (%)
LTR		**309,845**	**474.22**	48.56	**302,096**	**458.93**	48.36
	Gypsy	237,470	418.57	42.86	231,524	404.55	42.63
	Copia	60,600	44.22	4.53	59,546	43.61	4.60
	Caulimoviru	5,730	6.83	0.70	5,385	6.39	0.67
	ERVL	1,103	0.92	0.09	1,123	0.96	0.10
	unknown	4,942	3.68	0.38	4,518	3.42	0.36
SINE		**1,939**	**0.30**	**0.03**	**1,950**	**0.29**	**0.03**
	RTE	1,559	0.26	0.03	1,548	0.25	0.03
	unknown	380	0.04	0.00	402	0.04	0.00
LINE		**21,145**	**10.87**	**1.11**	**20,685**	**10.61**	**1.12**
	L1	16,153	8.21	0.84	15,806	8.00	0.84
	L1-Tx1	1,251	0.41	0.04	1,246	0.44	0.05
	Penelope	113	0.02	0.00	117	0.02	0.00
	RTE-BovB	1,290	0.22	0.02	1,271	0.21	0.02
	Tad1	2,338	2.01	0.21	2,245	1.94	0.20
Helitron		**1,990**	**1.55**	**0.16**	**1,994**	**1.20**	**0.13**
	DHH	1,990	1.55	0.16	1,994	1.20	0.13
DNA		**118,664**	**42.93**	**4.40**	**115,781**	**41.54**	**4.38**
	hAT	37,564	12.99	1.33	36,644	12.58	1.33
	CMC-EnSpm	14,405	8.08	0.83	14,001	7.75	0.82
	hAT-hATm	316	0.05	0.01	328	0.05	0.01
	hAT-Tip100	225	0.07	0.01	213	0.04	0.00
	MuLE-MuDR	2,377	1.80	0.18	2,129	1.58	0.17
	PIF-Harbinge	1,164	0.58	0.06	1,199	0.64	0.07
	TcMar-Tc1	1,927	1.31	0.13	1,944	1.31	0.14
	TcMar-Stowa	1,939	0.38	0.04	1,958	0.39	0.04
	MULE-MuDR	20,325	4.47	0.46	19,828	4.40	0.46
	Maverick	162	0.04	0.00	160	0.04	0.00
	hAT-Tag1	1,538	0.71	0.07	1,558	0.72	0.08
	hAT-Ac	35,485	12.17	1.25	34,545	11.77	1.24
	DNA	1,237	0.28	0.03	1,274	0.29	0.03
Unknown		**299,004**	**104.06**	**10.65**	**292,377**	**102.06**	**10.75**
Total		**752587.00**	**633.93**	**64.91**	**734,883**	**614.63**	**64.77**

Supplementary Table 7 non-coding RNA detected in the two cassava genomes

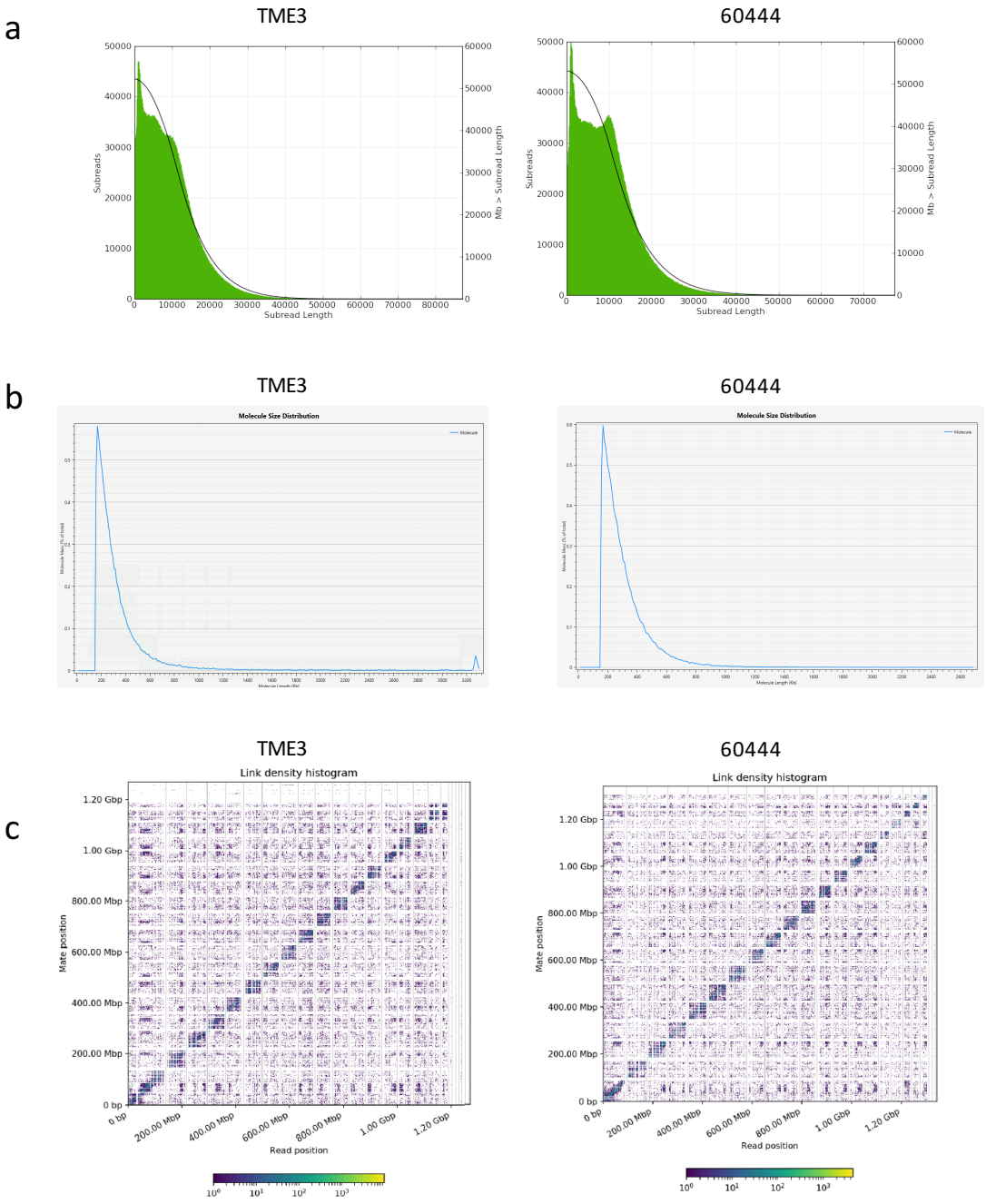
Non-coding RNA	Type	60444 Copies	TME 3 Copies
	rRNAs	706	555
	tRNAs	1,658	1,533
	miRNAs	325	333
	snRNAs	36	33

Supplementary Table 8 BUSCO analysis of genome assembly from 60444 and TME 3

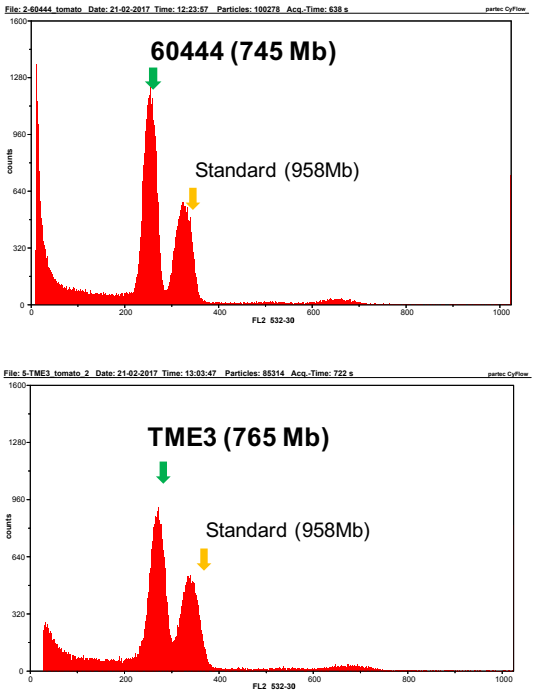
	60444	TME 3
Complete BUSCOs	1,369 (95%)	1,364 (94.8%)
Complete and single-copy BUSCOs	1,043 (72.4%)	1,081 (75.4%)
Complete and duplicated BUSCOs	326 (22.6%)	283 (19.7%)
Fragmented BUSCOs	20 (1.4%)	19 (1.3%)
Missing BUSCOs	51 (3.6%)	57 (3.9%)
Total BUSCO groups searched	1,440	1,440

Supplementary Table 8 The 18 cassava chromosomes in the *de novo* genomes

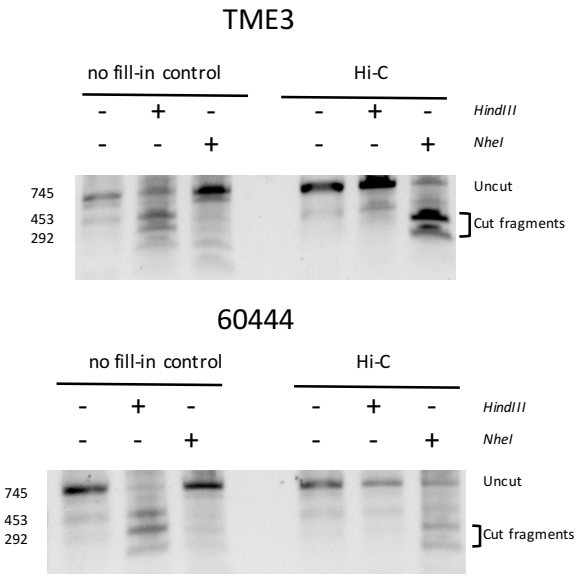
60444	**Chr.**	**TME3**	**Chr.**
Scaffold_16;HRSCAF=537	1	Scaffold_4710;HRSCAF=10556	1
Scaffold_3531;HRSCAF=8455	2	Scaffold_4710;HRSCAF=10556	2
Scaffold_3813;HRSCAF=9066	3	Scaffold_3766;HRSCAF=8561	3
Scaffold_3;HRSCAF=106	4	Scaffold_3024;HRSCAF=6918	4
Scaffold_2649;HRSCAF=6505	4	Scaffold_494;HRSCAF=1558	5
Scaffold_2579;HRSCAF=6346	5	Scaffold_4945;HRSCAF=11020	6
Scaffold_16;HRSCAF=537	5	Scaffold_6;HRSCAF=93	7
Scaffold_8;HRSCAF=202	6	Scaffold_1;HRSCAF=51	8
Scaffold_3074;HRSCAF=7427	7	Scaffold_176;HRSCAF=892	9
Scaffold_2;HRSCAF=52	8	Scaffold_3;HRSCAF=56	10
Scaffold_1583;HRSCAF=4059	9	Scaffold_14;HRSCAF=233	11
Scaffold_1262;HRSCAF=3358	10	Scaffold_16;HRSCAF=451	11
Scaffold_1;HRSCAF=40	10	Scaffold_7;HRSCAF=130	12
Scaffold_2922;HRSCAF=7074	11	Scaffold_2;HRSCAF=53	13
Scaffold_1478;HRSCAF=3800	12	Scaffold_15;HRSCAF=437	14
Scaffold_3793;HRSCAF=9016	13	Scaffold_5401;HRSCAF=12026	15
Scaffold_3881;HRSCAF=9216	14	Scaffold_11;HRSCAF=172	16
Scaffold_4;HRSCAF=126	15	Scaffold_12;HRSCAF=187	17
Scaffold_7;HRSCAF=175	16	Scaffold_3392;HRSCAF=7704	18
Scaffold_3938;HRSCAF=9338	17		
Scaffold_3237;HRSCAF=7788	18		



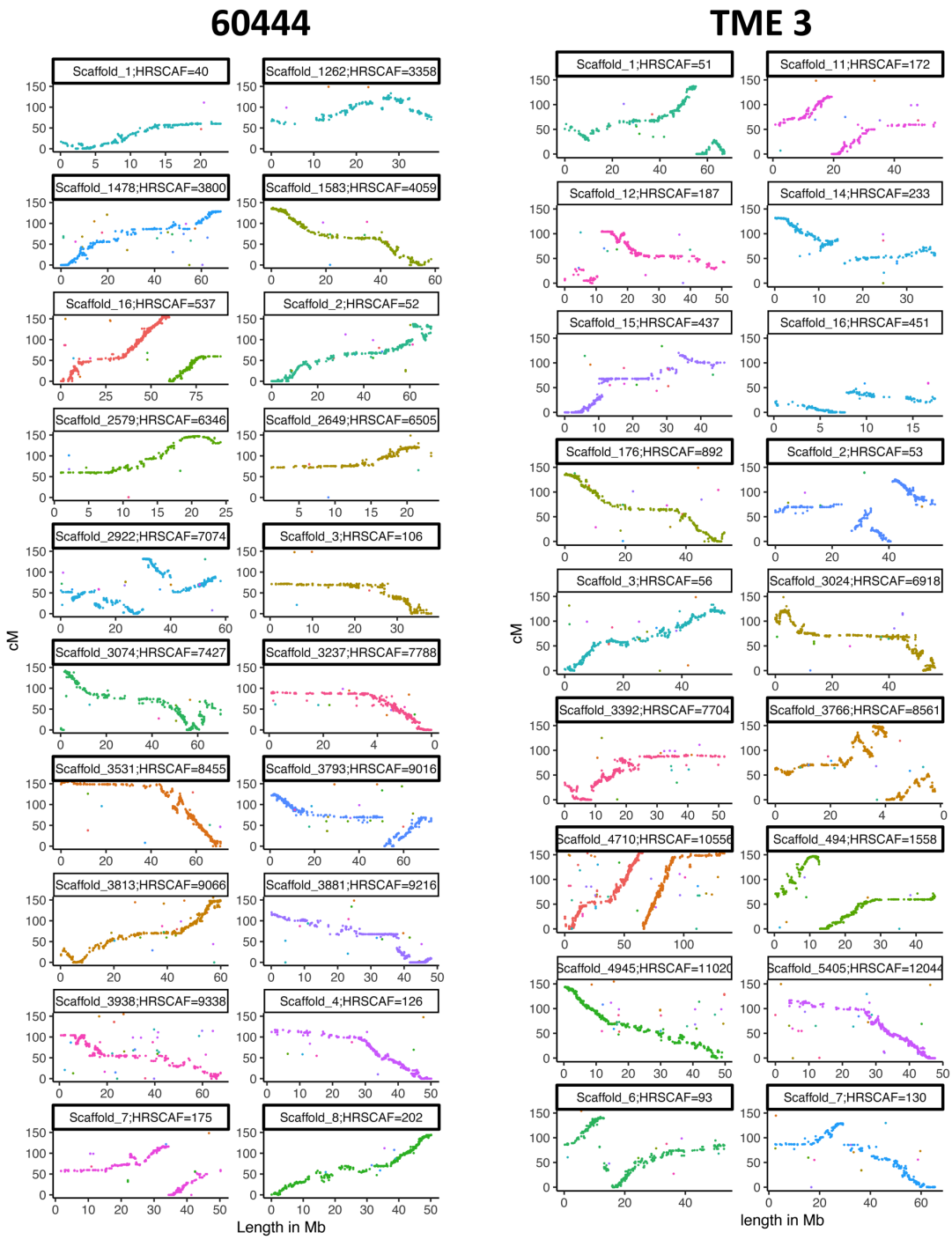
Supplementary Figure 1 Summary of data generated for genome construction. a) Size distribution of PacBio SMRT RS II subreads from single-molecule sequencing DNA from TME3 and 60444. b) Distribution of molecule lengths from BioNano Irys runs for TME3 and 60444. c) Shows the sequence binning using the proximity data. The x- and y-axes give the mapping positions of the first and second read in the read pair. The colour of each square gives the number of read pairs within that bin. White vertical and black horizontal lines have been added to show the borders between scaffolds. Scaffolds less than 1 Mb are excluded.



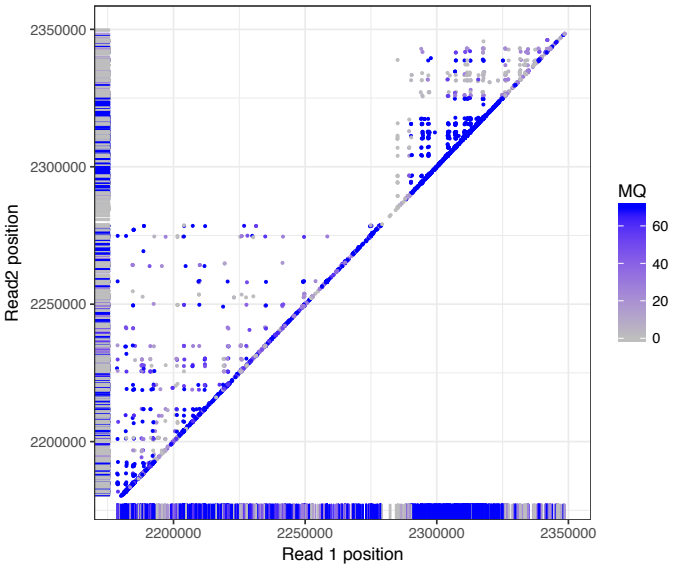
Supplementary Figure 2 Genome size estimation for the two cassava genotypes using flow cell cytometry and the tomato haploid genome reference 'Stupice'



Supplementary Figure 3 Quality controls for the Hi-C libraries for 60444 and TME 3 constructions. Control for labelling and ligation of ends in Hi-C libraries. The ligation junction of two close genomic cassava *HindIII* fragments was PCR-amplified and digested. In the no fil-in controls, no *NheI* restriction site can be generated and the *HindIII* recognition site stayed intact. In contrast, the Hi-C junctions were derived from blunt-end ligation of filled-in *HindIII* sites, and were therefore cleaved by *NheI*. DNA was separated using a standard 1.5% agarose gel. Size of the PCR-products were indicated on the left.

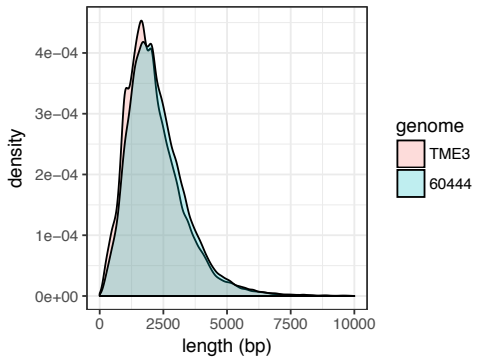


Supplementary Figure 4 Pseudo-molecule validation using the 22,403 genetic marker from the cassava composite genetic map and the 18 pseudo-chromosomes of the cassava composite genetic map. Marker were aligned to each genome using BLAT. Each dot indicates a full-length sequence match. The x-axis represents the physical map of a HiRise scaffold and the y-axis the genetic distance extracted from the cassava composite genetic map[78]. Chromosomes were visualized with different colours and chromosome number was manually written nearby the sequence scaffold. For chromosome identifiers please see Supplementary Table 8.

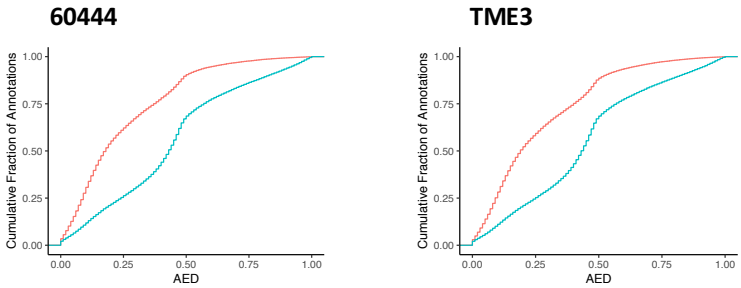


Supplementary Figure 5 Example of a misassembly identification using chromosome conformation capture read pairs. The paired-end mapping positions in the region 2,200,000-2,340,00 Mb of Super-Scaffold_123 show a sudden absence of read pairs spanning across the region at around 2,280,000 Mb. MQ: read mapping quality

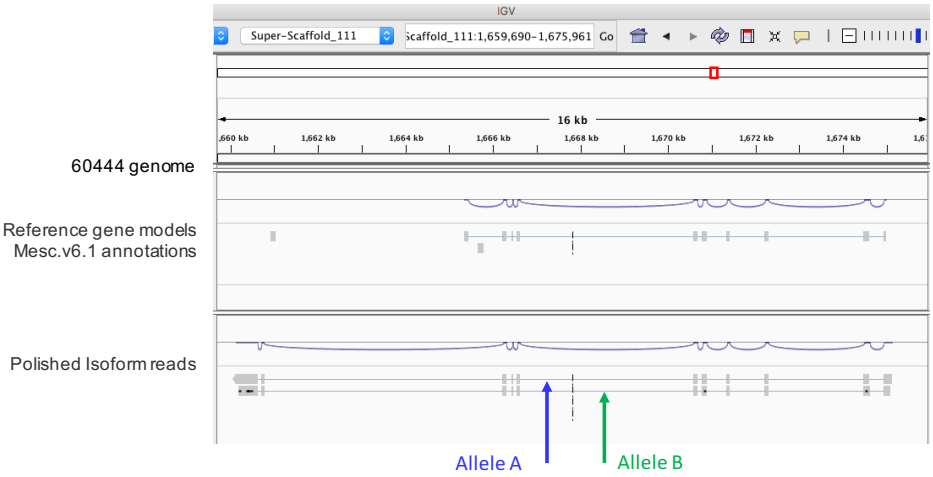
a



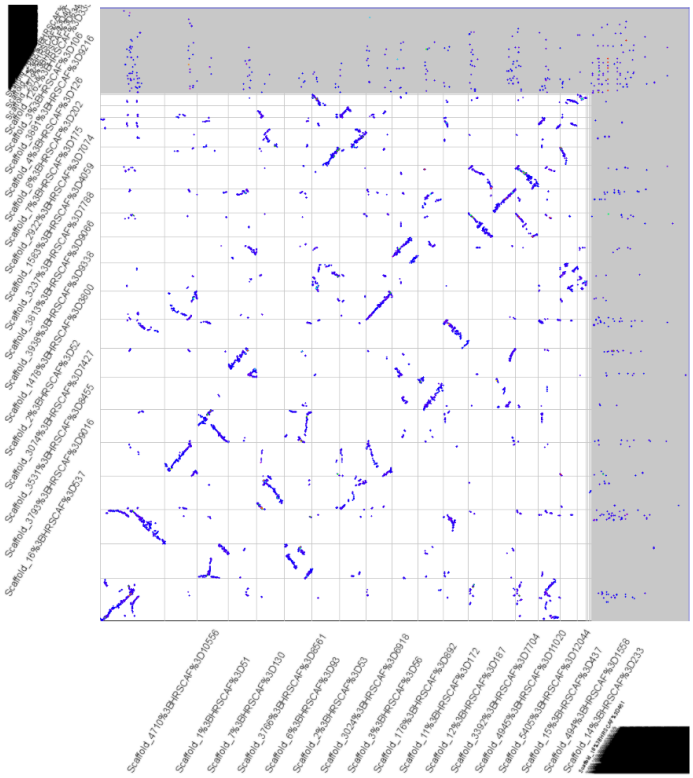
b



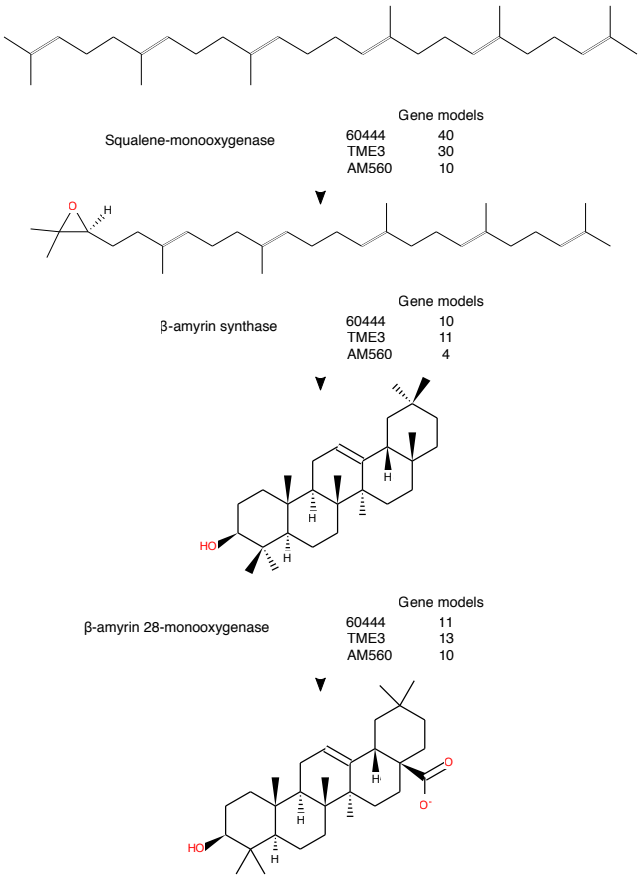
c



Supplementary Figure 6 Summary of full-length transcriptome sequencing for high-quality gene-space annotation. a) Length distribution and data density of full-length sequenced transcripts from 60444 and TME 3 RNA. b) AED analysis of the gene model prediction. Plot shows the cumulative fraction of the annotations on the y-axis and the AED scores calculated by the annotation pipeline on the x-axis. Red line represents the updated annotation that used the Iso-Seq data and green line shows the AED scored for the genes annotated without Iso-Seq. c) Improved full-length transcript supported gene space annotation for the 60444 genome assembly. The top track shows the previous gene space annotation29 (Reference gene models Mesc.v6.1 annotations). The two tracks below (Polished Isoform reads) represent sequence alignments of full-length transcript reads of 60444 RNA. Blue and green arrow indicate the two sequenced alleles aligning to that locus. Black dots in Allele B represent indels and mutations, whereas Allele A aligns with no mismatch.

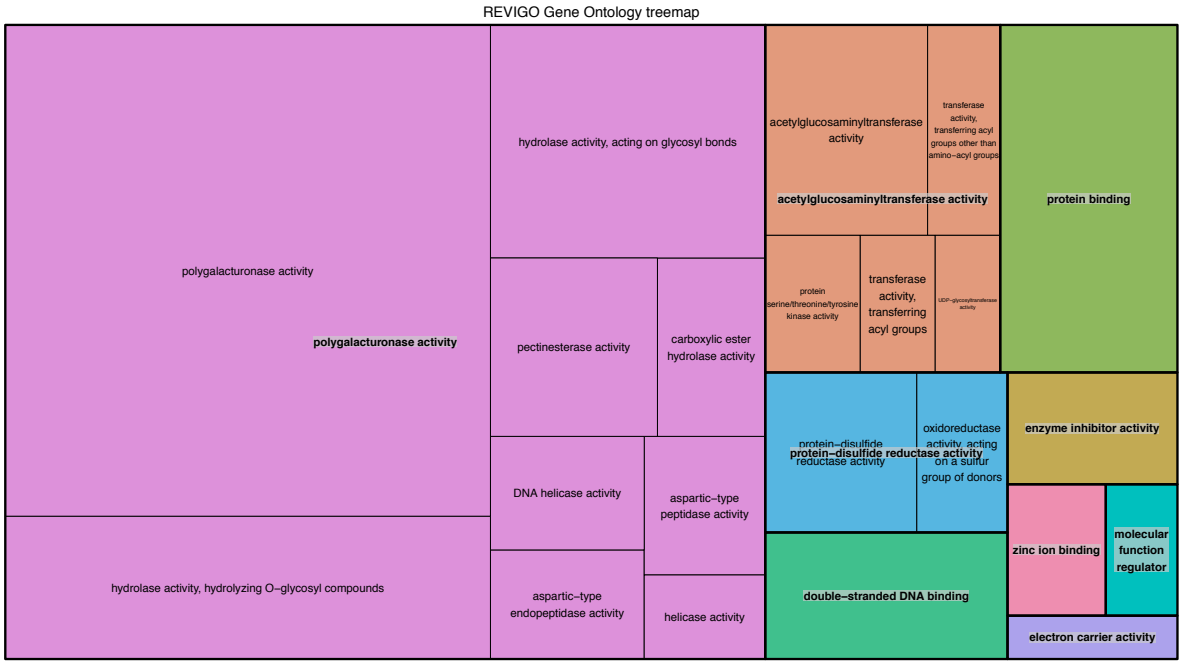


Supplementary Figure 7 Syntenic dotplot generated by SynMap (https://genomevolution.org/coge/) between cassava 60444 and TME 3. Each syntenic gene is depicted as dot. 60444 is shown on the y-axis and TME 3 on the x-axis.

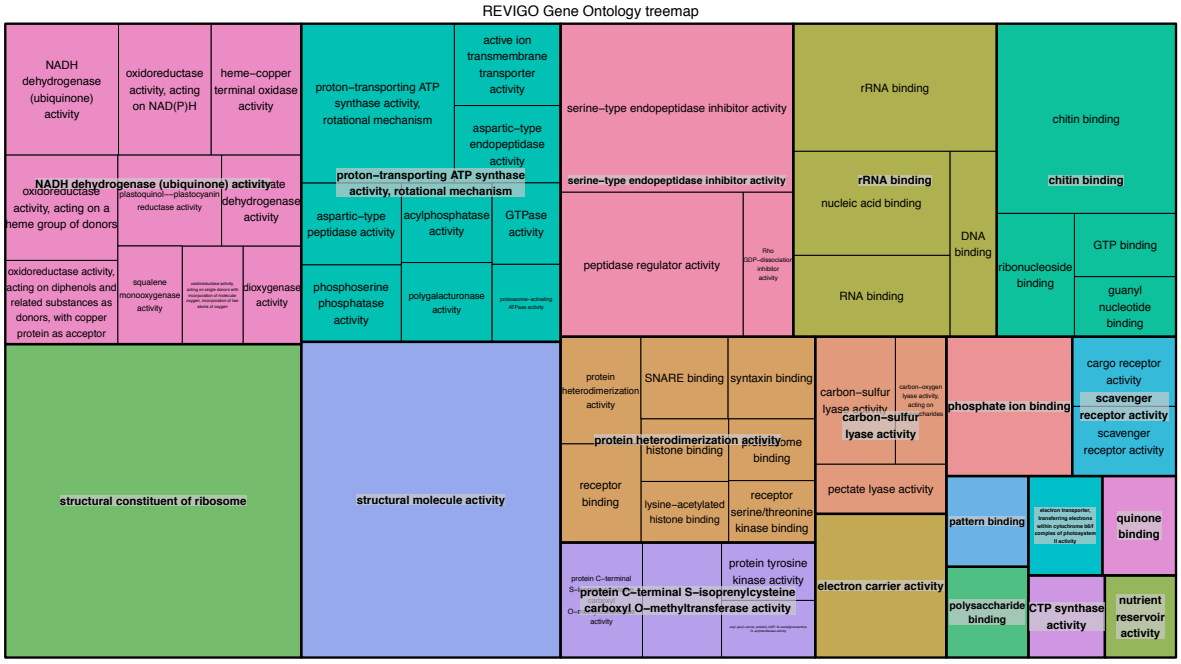


Supplementary Figure 8 Squalene monooxygenase activity (GO:0005198) pathway and the corresponding gene models found in 60444, TME 3 and AM560

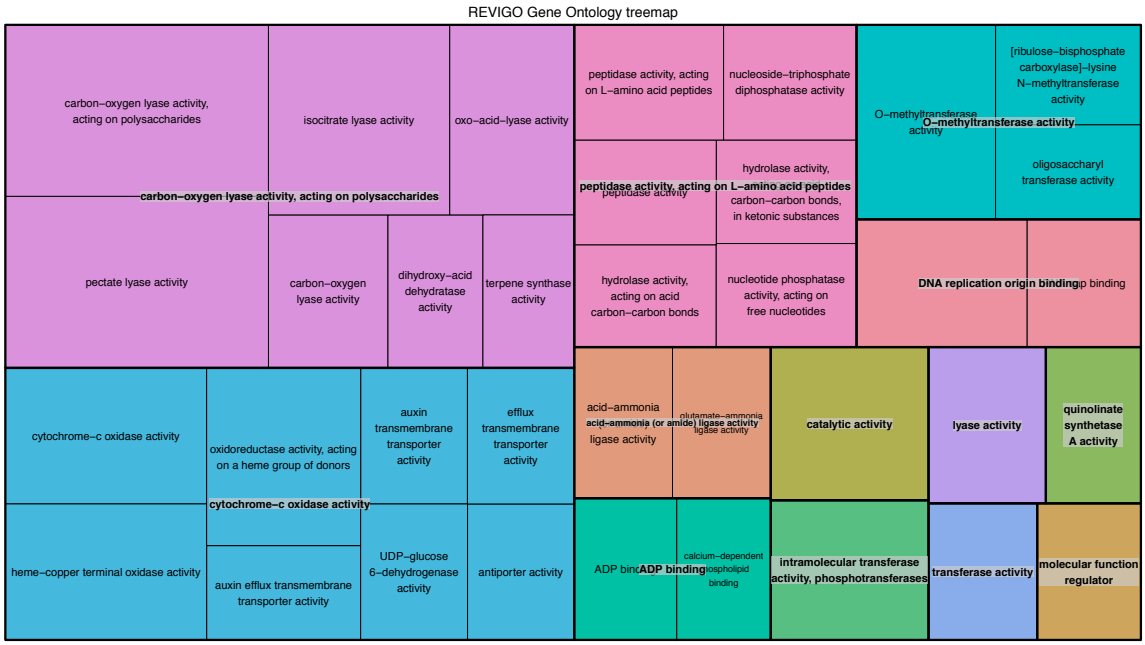
.



Supplementary Figure 9 GO enrichment analysis for the genes specific to the AM560 genome



Supplementary Figure 10 GO enrichment analysis for the genes specific to the 60444 and TME 3 genome



Supplementary Figure 11 GO enrichment analysis for mono-allelic expressed genes in TME 3

Chapter 3

Reconstruction of the cassava dominant geminivirus resistance locus *CMD2* in *CMD2*-type as well as virus susceptible cultivars using a novel diploid-genome visualization tool

Personal contribution:
I developed the QTL reconstruction pipeline using reference CDS and QTL-associated genetic markers. I conceived the SCEVT tool box and optimised together with Philipp Rogalla von Bieberstein the Python scripts and general workflow. I analysed the QTL region for CDS and reconstructed the genetic composition of the *CMD2*. I wrote the draft manuscript with input from Philipp Rogalla von Bieberstein and Prof. Vanderschuren.

Publication state:
Scripts are available on https://github.com/ and will be uploaded to https://zenodo.org/ soon. This manuscript will be submitted to *BMC genomics*.

Reconstruction of the cassava dominant geminivirus resistance locus *CMD2* in *CMD2*-type as well as virus susceptible cultivars using a novel diploid-genome visualization tool

Joel-Elias Kuon[*1], Philipp Rogalla von Bieberstein[*1], Wilhelm Gruissem[1] & Hervé Vanderschuren[2,1]

[1] Institute of Molecular Plant Biology, Department of Biology, ETH Zurich, Universitätstrasse 2, 8092 Zurich, Switzerland
[2] AgroBioChem Department, University of Liège, Passage des Déportés 2, Gembloux, Belgium

Correspondence: kuonj@ethz.ch & herve.vanderschuren@ulg.ac.be
[*]equal contributors

Abstract
Visualization of DNA sequence comparisons is instrumental to determine genotypic differences between related or unrelated species. The affordability and increasing throughput of third-generation sequencing and single-molecule mapping technologies have generated the first, diploid-aware whole genome assemblies. Recently, the first two high-quality, diploid aware genomes were released for cassava (*Manihot Esculenta Crantz*) opening the way for a new era of comparative genomics in this important food-security crop.
To estimate assembly quality, and to determine the novel haplotype structures, flexible and fast abstraction methods are required to validate and exchange genomic resources. We developed a new simple-to-use visualization tool that uses only standardized annotation files to compare the location of key genetic features (i.e., gene location, genetic markers) between diploid assembled sequences. The Scaffold and Contig Exploratory Visualization Tool (SCEVT) generates images that show shared or unique genetic features for each individual haplotype and compares their position to a reference. We applied SCEVT to reconstruct the heterozygous major geminivirus resistance dominant locus *CMD2* using two new cassava genomes having contrasting resistance to the *CMD2*. We present a detailed map of the *CMD2* locus for the cassava cultivar TME 3, which carries the *CMD2* resistance, and for the geminivirus susceptible cultivar 60444. Using SCEVT we show the major quality improvement that was achieved from long read assemblies to fully scaffolded sequences using optical mapping and proximity-ligation scaffolding. The precise *CMD2* map can be used for candidate gene identification, *CMD2* fine-mapping and further sequence polishing. The software tool SCEVT is freely available for all operating systems.

The advent of third-generation, single-molecule sequencing technologies such as the PacBio Sequel and Oxford Nanopore platforms has revolutionized whole genome assemblies from prokaryotes and eukaryotes [120]. Latest development of long-read mapping technologies (i.e. optical mapping, proximity ligation mapping) combined with sophisticated diploid-aware genome assembly algorithms have generated the first haplotype 'phased' assemblies from complex crop genomes [52], [54]. We recently generated two high-quality, diploid-aware cassava genomes (Kuon et al. in preparation) in order to elucidate the dominant geminivirus resistance locus *CMD2* that confers resistance to the cassava mosaic disease (CMD) [31], [34], [35][5]. The diploid-aware genomes were generated following a hierarchical pipeline that started with assembling PacBio long-read sequences and finished with the construction of large pseudo-molecules using long-range scaffolding with optical mapping [58] and proximity-ligation mapping (Hi-C) [61]. The gene space of the two genomes was annotated with *ab initio* predictions tools as well as evidence based data that used public available coding sequences (CDS) as well as newly generated full-length transcriptome sequencing (Isoform sequencing).

Genome visualization tools are essential to transfer and share novel genomic knowledge between researcher and research groups. But current visualization tools have several limitations restricting their use for direct and easy comparison of diploid-aware genomes and are often difficult to use for a rapid and precise evaluation of intermediate or provisional genome assemblies. For example, the Artemis Comparison Tool (ACT) [121], VISTA [122] or MAUVE [123] require either computational demanding computers with high graphical power or they cannot be applied to novel draft genome sequences because administrator rights are needed to approve the sequence. Furthermore, these tools don't support a command-line based application that would allow to run the software on a high-performance computer cluster. Other tools, such as the sequence dotter DNAplotter [124], SynMAP (https://genomevolution.org/coge/) or the MUMMER toolkits [125] demand high computational run-time, require a pre-release of the genome to an external computer platform, and handle poorly highly repetitive, GC-rich and large genomes resulting in extremely long computational run-times.

Ideally a visualization tool should use a platform-independent programming language, use only highly standardized data formats (such as FASTA and BLAST), and generate a rapid and intuitive representation of any shared or contrasting genomic feature that can be detected between two sequences. We developed the **S**caffold and **C**ontig **E**xploratory **V**isualization **T**oolkit (**SCEVT**) that is capable to draw comparative images between two genomes, can be used to reconstruct a precise map of a genomic QTL region, to compare QTL regions between different accessions, to identify haplotypic variations and to show limitations of an assembly (i.e. sequencing gaps).

We applied SCEVT to reconstruct the major cassava mosaic geminivirus (CMG) resistance locus *CMD2* and present the visualization of the *CMD2* locus across several, diploid-aware and haplotype phased incremental genome assemblies of the two cassava cultivars, TME 3 and 60444, contrasting for the resistance. This detailed map revealed *de novo* annotated genes, the broad collinearity between the *CMD2* locus and the cassava genetic map, the location of *CMD2*-associated SNP markers as well as the distribution of key sequence features (i.e. repetitive elements) along the *CMD2* locus.

Results and Discussion

SCEVT was written in the programming language python (version 2.7) and the scripts are freely accessible on github (https://github.com/pbieberstein/SCEVT). Figure 1 shows a general overview of the pipeline that can be used to visualize and reconstruct a QTL region.

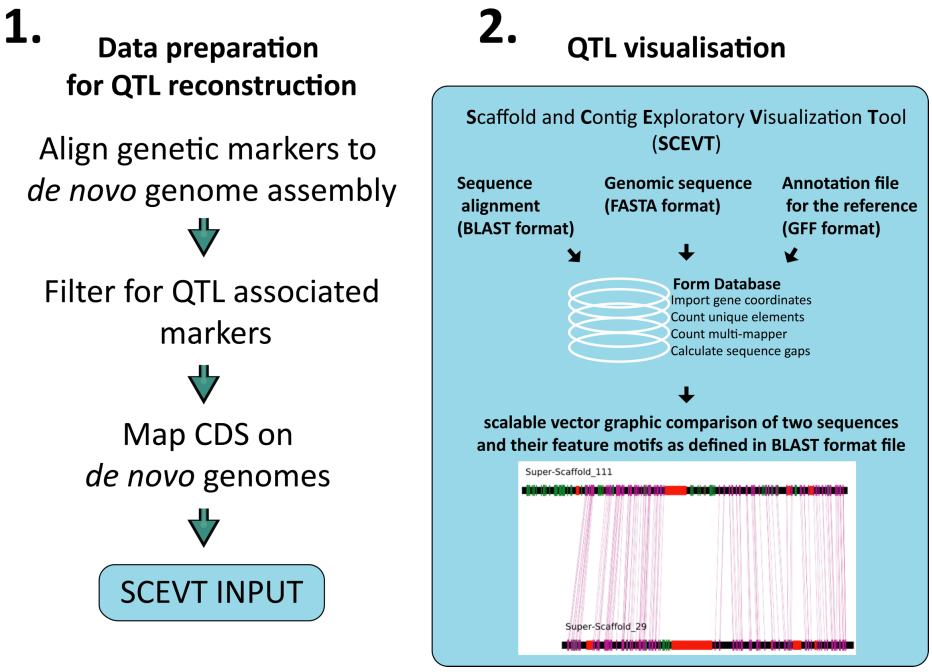


Figure 1 Overview of the SCEVT pipeline. On the left panel steps are shown to prepare the input data. In the second step (right panel), SCEVT processes the input data (i.e. BLAST, FASTA, GFF) to compare the feature space (i.e. gene space) of two sequences. An example plot for SCEVT is shown at the bottom of the right panel. The two sequences come from the 60444 genome assembly and were identified to contain *CMD2* linked genes. Green lines indicate genes that are absent on the other sequence, whereas pink lines indicate that an allele of the same gene was found on the other sequence. The red bars indicate assembly gaps. Super-Scaffold_111 and Super-Scaffold_29 spanning 2.56 Mb and 2.23 Mb, respectively.

SCEVT takes as input files the complete genome in 'FASTA'-format and the corresponding annotation file in 'gff' format. Both are common files and are usually freely available for released genome assemblies. In addition, the GMAP output of the gene mapping is needed to compare the alignment position between the *de novo* genome assembly and a reference genome 'gff' file. GMAP is a fast and resource-efficient splice-variant aware coding sequence (CDS) alignment software [100]. After generating the GMAP-database for the genome assembly using 'gmap_build' the set of reference CDSs can be aligned using the command 'gmap –f 1'. It is important to mention that the GMAP output option is correctly set since SCEVT only uses the common BLAST '.psl' format as input.

After executing the script with 'python scaphy.py', SCEVT starts with screening the sequences to be visualized and extracts the coordinates of the sequencing gaps that span sequence lengths superior to 100 bp. The user has to specify the sequences to be analyzed in a configuration file. An example .config file is given in the software distribution and can be easily modified using a common text editor. Then SCEVT scans through the GMAP output file (.psl format) and extracts the locations for all genes that were mapped to any of the scaffolds. Next, it scans through the reference genome annotation file (.gff format) and pulls out the gene positions in the reference genome. Lastly, it matches up the coordinates of all genes on the scaffolds with the corresponding locations on the reference genome and uses this information to draw the graphical vectors.

SCEVT consists of two different scripts. **Scaphy.py** (**S**caffold to **P**hysical **R**eference **M**apping) is a tool to visualize scaffolds in relation to a reference genome assembly and draws mappings to a reference sequence whenever the genes were found in the *de novo* sequence and the reference sequence. This tool is helpful to analyze genome assemblies for syntenic relation and to find structural variations (SVs). It also highlights when a gene is detected in a scaffold but absent in the specified chromosome of the reference genome, indicating new candidate genes anchored in the genomic region. **Scaco.py** (**S**caffold **C**omparison) was developed to directly compare *de novo* sequence scaffolds based on their gene annotation. This tool is particularly useful for diploid-aware genome assemblies where haplotype blocks can be directly compared. It highlights and maps the genes that are similar on two scaffolds. The tool also highlights which genes are present on one but not the other. Additionally, it also plots the gaps within the scaffolds. An example plot is shown in Figure 1, right panel.

Reconstruction of the *CMD2* using SCEVT

For reconstructing the *CMD2* in the two cassava accessions, we followed the points listed in the SCEVT description (summarized in Figure 1). Moreover, for visualization of the *CMD2,* we run the same pipeline on different intermediate assemblies. The method was applied to the long-read assemblies (CANU), the long-read assemblies plus optical map improvements (CANU-BNG), and finally, on the long-read assemblies that were improved by optical mapping and Hi-C scaffolding (Dovetail).

For input data preparation, the cassava composite genetic map[78], the cassava reference gene models (v6.1) as well as the reference annotation file ('gff' format) were downloaded from the phytozome data bases (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Mesculenta)[49]. The SNP markers as well as the reference CDS were aligned to the *de novo* cassava genomes using BLAST and GMAP. Then, the *CMD2* locus was defined by the SNP markers that had the identifiers 's5214' and 's6906' and were located between the genetic distance of 15 to 60 centi-Morgans (cM) on chromosome 12 [34], [35], [126]. The *CMD2* locus in the reference genome (v6.1) spans 2.14 Mb and carries 127 gene loci. We compared the lists of the initial set of *CMD2* associated genes (n=127, isoforms=152) with the genes that aligned on our *de novo CMD2* sequences. Figure 2 details the number of *CMD2* genes and isoforms found in the *CMD2* sequence selection for the two different cassava accessions using the incremental genome assemblies.

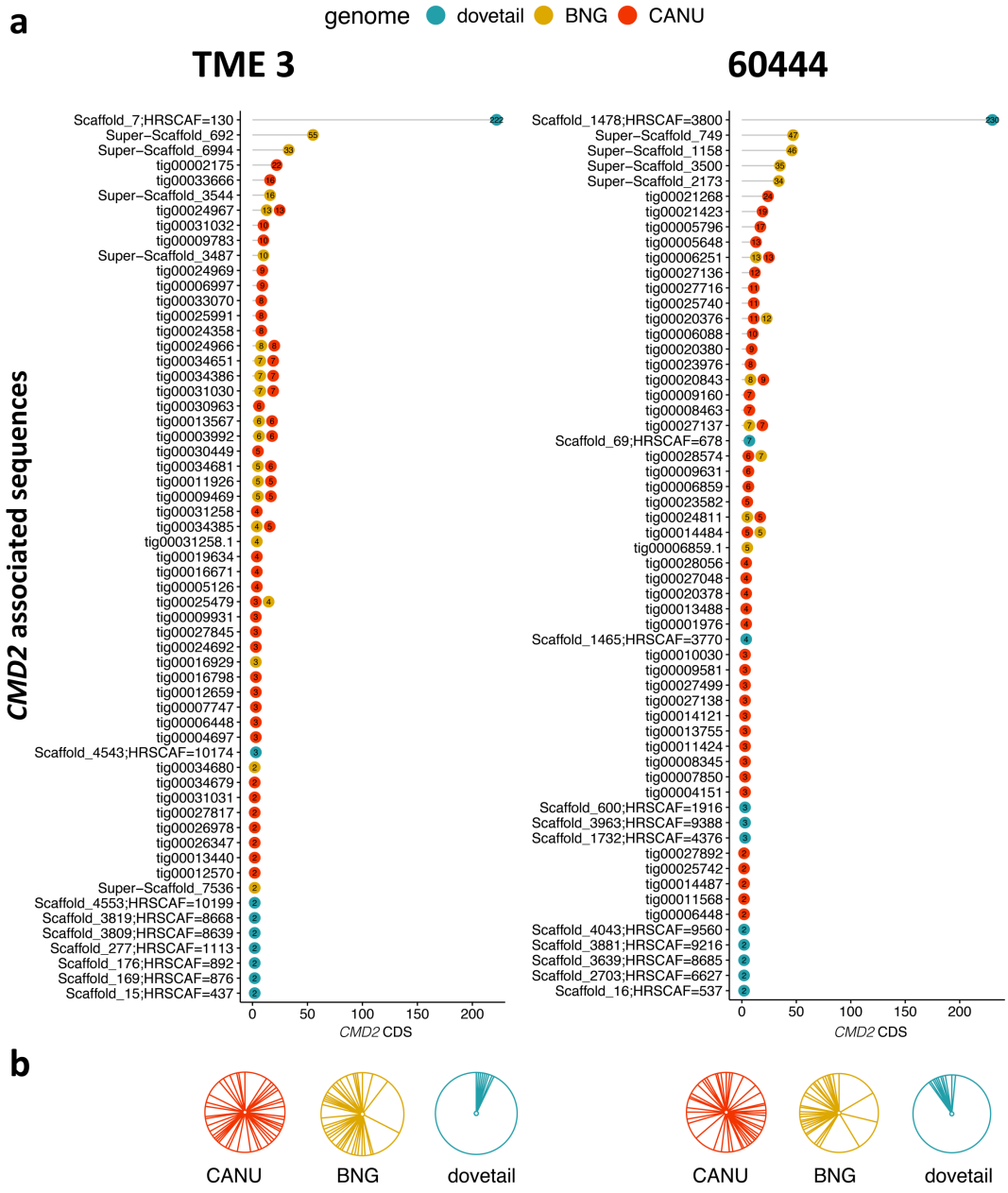


Figure 2 *CMD2* associated genes and their corresponding contigs and scaffolds in the *de novo* genomes a) On the y-axis, the sequence identifiers (IDs) are shown and on the x-axis the number of *CMD2* genes. Red circles indicate the *CMD2* contigs generated by the CANU assembler. Yellow circles indicate the optical map improved *CMD2* locus and green circles show the *CMD2* locus after Hi-C dovetail scaffolding. Numbers in the red, green or orange dots indicate the number of *CMD2* linked genes found on that contig or scaffold. b) Number of *CMD2* associated genes visualized as pie chart. Each pie segment indicates a contig or scaffold and the size of the segment shows the proportion of *CMD2* associated genes found on that sequence.

This simple comparison shown in Figure 2 allowed to estimate the sequence continuity improvements that have been achieved with using the different sequence scaffolding technologies. For example, we identified 42 CANU contigs that represent the *CMD2* in TME 3. This high number of sequences was strongly reduced after applying optical map supported scaffolding that based on positioning of contigs according to large, often multi-Mb spanning optical maps (BioNano Genomics). Following the same example, optical maps helped to scaffold the initial set of 42 TME 3 contigs and led to a reduction from 42 to 6 optical map supported scaffolds (named as 'Super-Scaffold') that contain the majority of *CMD2* CDSs. This number of not-scaffolded sequences was drastically reduced after applying the proximity mapping data set. After the implementation of the Hi-C scaffolding, the whole *CMD2* locus was assembled into a single sequence scaffold that bears all of the initial 127 *CMD2* linked gene loci in TME 3.

CMD2 haplotype visualization using SCEVT

The SCEVT result revealed a high syntenic relation between the reference and the *de novo* sequences in both the genomes of 60444 and TME 3 (Figure 3). For example, in TME 3 we found only four CANU contigs that carried genes that did not match the *CMD2* locus of the reference (Figure 3, CANU panel). Among the four contigs, 14 genes were found by GMAP that had a different location in the reference genome. We further investigated their location in the reference genome and found that most of them (n=12) had no chromosomal location assigned leading us to the conclusion that we *de novo* anchored genes on the *CMD2* locus in cultivar TME 3. In 60444, we found 13 *de novo* anchored genes (Supplementary Figure 1). A similar pattern was observed when using SCEVT to reconstruct the optical map improved CANU genomes (BNG)(Figure 3, top BNG panel). Here we found 30 genes that originated from other locations of the reference genome. Most of those came from Super-Scaffold_692 (n=26).

Optical mapping and long-read sequencing have the ability to phase haplotypes over several Mb in distance[127]. To access this information in the *de novo* assemblies, SCEVT was used to visualize the haplotype structure of the *CMD2* in the optical map improved CANU assemblies. In TME 3, for example, the Super-Scaffold_692 and Super-Scaffold_3544 are most likely haplotypes and even show haplotypic variation (Figure 3, BNG panel). A very similar pattern was observed in 60444 where Super-Scaffold_1158 and Super-Scaffold_749 appear to span a fully haplotype phased > 2Mb genomic region of the *CMD2* (Supplementary Figure 1, BNG panel). Optical mapping strongly depends on high sequence contiguity in the initial set of contigs. In case of the *CMD2* locus, many CANU contigs were below 100 kb in length. As a consequence, the contigs had too few optical tags to anchor them precisely on the optical map. This resulted in a few major assembly gaps in the optical map supported 'Super-Scaffolds'. However, the optical map defined gap size as well as the haplotype information provides very useful information for future attempts to finish sequencing the *CMD2* locus.

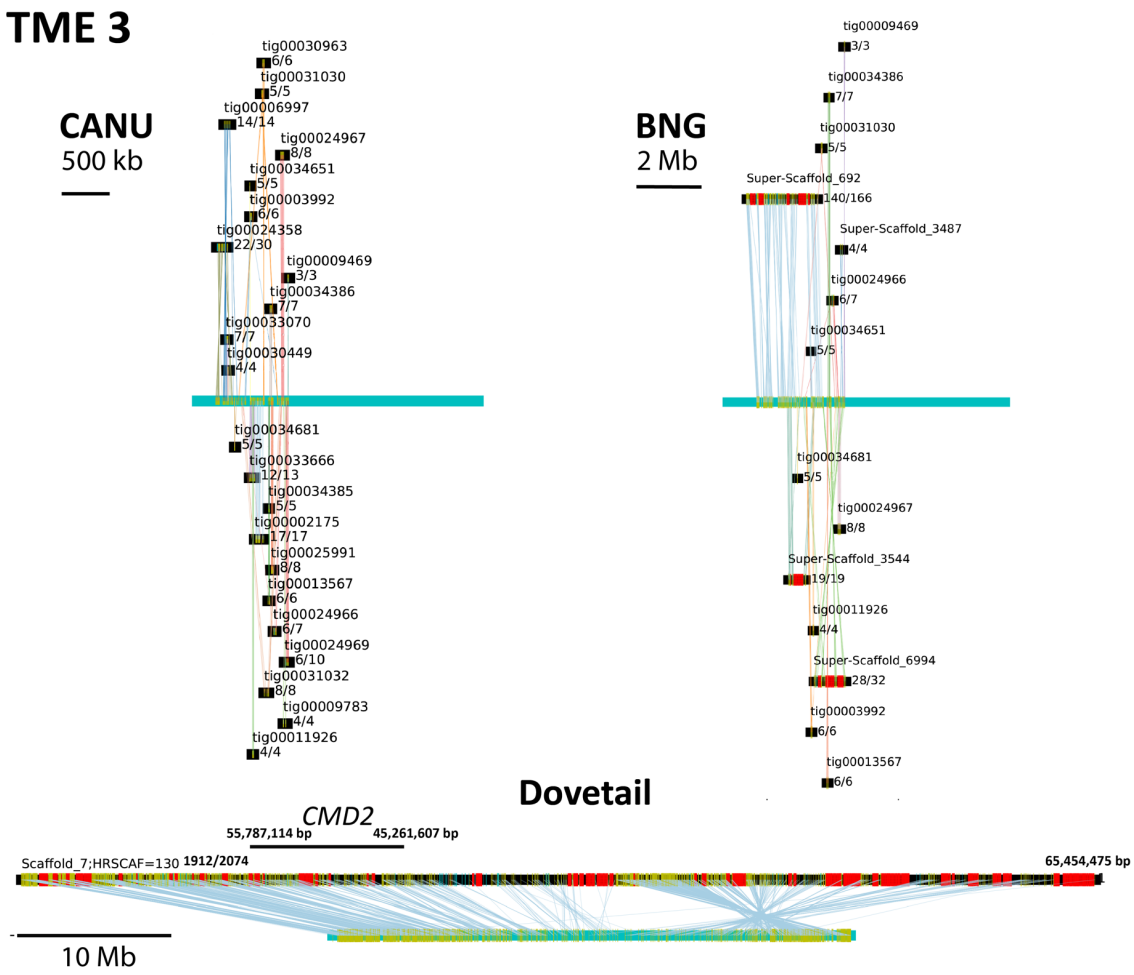


Figure 3 SCEVT output for the *CMD2* in TME 3 over the three different assemblies CANU, CANU-BNG and CANU-BNG-Hi-C (Dovetail). The cyan line indicates the physical map of chromosome 12 from the reference cassava genome AM560 [49]. Black bars indicate *CMD2* associated contigs or scaffolds from the *de novo* assembly. Above the black bars are the sequence identifiers including the number of genes found on each contig matching *CMD2* reference genes. Matching *CMD2* genes are indicated with a yellow line. Genes not matching the *CMD2* but present in a *CMD2* associated *de novo* sequence are indicated with a bright-green line. Red bars indicate the assembly gaps present in a scaffold.

In vitro proximity ligation, as an application of chromosome conformation capture technologies, can provide genomic information for scaffolding sequences and to re-construct even whole chromosomes [59], [61], [62], [79]. By analyzing the SCEVT result for the Hi-C scaffolded cassava genomes (Dovetail), a single major scaffold was identified that spanned the entire *CMD2* chromosome 12. The Hi-C based Dovetail scaffolds revealed an exceptional high syntenic collinearity between the reference and the *de novo* sequences for most of the *CMD2* region (Figure 3, Dovetail panel and Supplementary Figure 1, Dovetail panel). Interestingly, a large paracentric inversion of a ~15 Mb genomic region was identified at the opposite chromosome arm for 60444 as well as TME 3. Although this inversion doesn't affect the genomic context of the *CMD2*, it should be stressed that further comparisons and validation is needed for the chromosome 12 of the sequenced genomes to check whether the inversion is due to genomic variation or incorrect

assembly. The chromosome 12 in our Hi-C scaffolded genomes (Dovetail) carry far more sequences than the chromosome 12 of the reference genome (65.5 Mb for TME 3 vs 31.6 Mb in cassava v6.1). We previously found that the two genomes 60444 and TME 3 had ~ 15% higher amount of assembled repetitive sequences that could potentially contribute to the size differences. Moreover, the optical mapping introduced physically accurate sequencing gaps that provide a more detailed information about the chromosome size compared to the sequencing gaps in the reference genome that were introduced based on genetic rather than physical distance. However, since the SCEVT approach only identified a single major scaffold containing all the *CMD2* CDSs it is possible that Hi-C led to a more haplotype 'collapsed' genome assembly. The authors of this study want to emphasize that the optical map improved genomes provide a fully haplotype-phased representation of the *CMD2* that will be instrumental for future sequence polishing (i.e. Gap filling) or candidate gene selection. In this context, the Dovetail *CMD2* map will be highly important for future fine mapping attempts since the overall genomic context appears to be assembled correctly. This fact was further confirmed by using the scaco.py script from the SCEVT tools that allows a comparison of two sequences for syntenic features. The direct sequence comparison between the Dovetail-*CMD2* scaffold of 60444 and TME 3 revealed a high syntenic relation over a 10 Mb distance that spanned the whole *CMD2* (Supplementary Figure 2).

The highly complex nature of the CMD2

The *CMD2* reconstruction pipeline revealed for both Hi-C improved genomes a single scaffold that span the entire *CMD2* locus. We expanded the *CMD2* locus as it was defined in three mapping studies [31], [34], [35] for additional 2 Mb on each site resulting in a physical region of 10 Mb in total and analyzed this region for its collinearity to the cassava composite genetic map [128]. This revealed that the TME 3 Scaffold_7;HRSCAF=130 showed broad agreement with the genetic cassava map (Figure 4a). This analysis also revealed that the two recent *CMD2* mapping studies placed the *CMD2* locus to similar but not identical regions. This is indicated with the coloring of *CMD2* associated markers in Figure 4a where red SNP-markers indicate the *CMD2* locus published by Rabbi and colleagues and green SNP-markers show the *CMD2* locus as published by Wolfe and colleagues. We sought to find potential reasons for the six large assembly gaps in the *CMD2* locus. Assembly breaks are often direct consequences of the high abundance of repetitive elements. The Figure 4b shows a detailed map of the key genetic features of the *CMD2* locus in TME 3. This analysis revealed a highly repetitive genomic locus that contained all major sequence repeat class. We found the Long Terminal Repeat (LTR) retrotransposons to be the most abundant at the *CMD2* followed by the non-LTR retrotransposons elements (LINE). We also detected DNA satellites and Helitron hotspots at the *CMD2*.

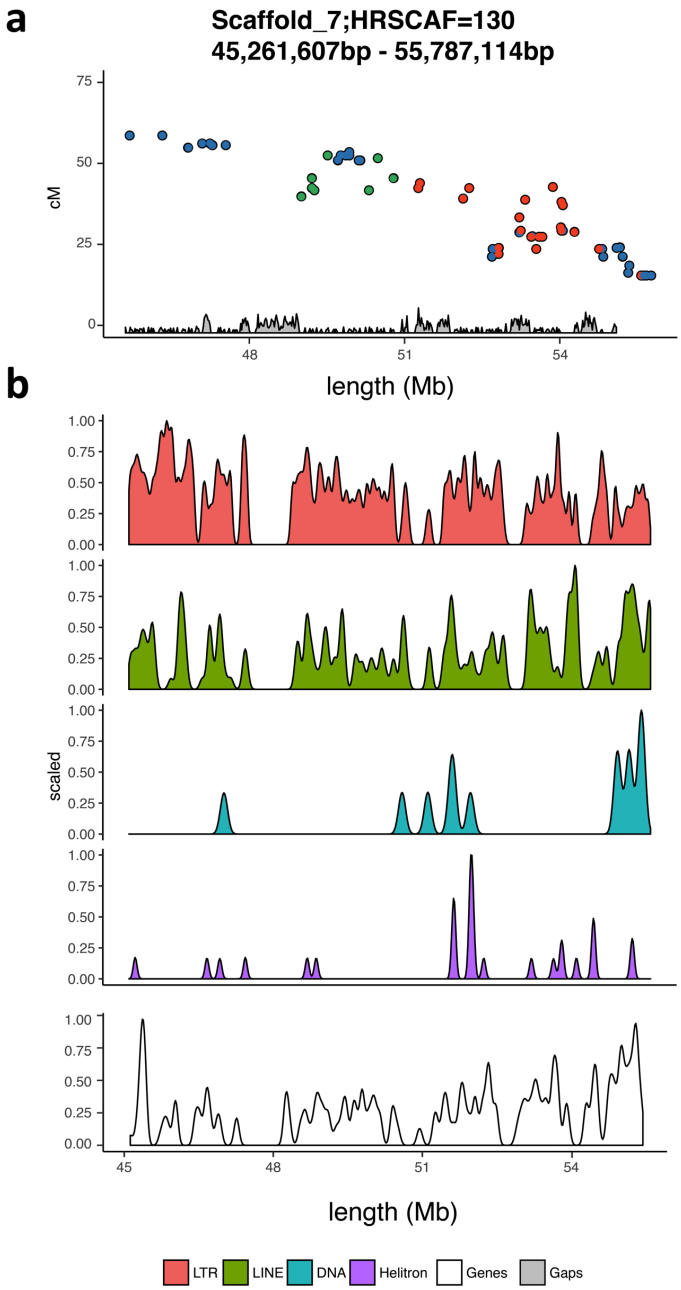


Figure 4 Key genomic features at the *CMD2* in TME 3. a) SNP-marker location at the *CMD2*. Red dots indicate *CMD2* SNP-markers released by Rabbi et al.2014 [126] and green dots indicate the SNP-markers released by Wolfe et al. 2016 [35]. The grey density curve indicates the assembly gaps. b) Key genetic features for the *CMD2*

To find *CMD2* associated genes, we predicted protein-coding sequences with a combination of *ab initio* prediction and transcript evidence from the reference cassava CDSs as well as newly generated full-length transcriptome sequencing (Isoform-sequencing) as reported earlier (Kuon et al. in preparation). For TME3, 267 *de novo* annotated genes were revealed in the 10 Mb region of the *CMD2 (*45.2 Mb-*55*.8 MB)(Supplementary Table 1). There are no classical resistance genes (e.g., nucleotide binding site-leucine rich repeat) at the *CMD2* for both genomes but two genes were identified that functional annotation can be directly linked to virus resistance in plants.

The MeTME3_00015870-RA gene at position 54,593,285 Mb encodes a protein disulfide isomerase like (PDI) 2-3. PDIs catalyze the correct folding of proteins and prevent the aggregation of unfolded or partially folded precursors. Previous genetic studies have identified *HvPDI5-1* in barley (*Hordeum vulgare L.*), the ortholog of *MePDI-2.2* as a virus susceptible factor [129] that causes resistance to the single-stranded (ss) RNA bymoviruses. The loss of function of the *HvPDI5-1* in a bymovirus resistant barley accession occurs via single-nucleotide polymorphisms (SNPs). Suppression of members of the PDI gene family can delay replication of several mammalian viruses (e.g. HIV) but their role in virus pathogenicity remains largely unknown [130], [131].

The second *CMD2* candidate gene, MeTME3_00015743-RA at positon 47,282,415 Mb, encodes for a Suppressor of Gene Silencing 3 (SGS3). SGS3 genes are involved in posttranscriptional gene silencing (PTGS) and support the RNA-directed RNA polymerase 6 (RDR6) for the dsRNA synthesis [132]. SGS3 has also been reported to be involved in the transport of the RNA-silencing signal [133]. Viruses are a direct target of the host RNA silencing machinery [134][135] and SGS3 mutants consistently displayed enhanced susceptibility to viruses [136][132]. *SlSGS3*, the tomato homolog of the *Arabidopsis* SGS3, has also been shown to directly interact with the tomato yellow leaf curl geminivirus (TYLCV) V2 protein that functions as a suppressor of silencing and counteracts the innate immune response of the host plant [13]. Both candidate genes have interesting functional properties that have to be addressed in future studies through a targeted reverse genetic screening in *CMD2*-type as well as geminivirus susceptible cassava plants.

The two *CMD2* candidate genes are separated by a large distance of 7.3 Mb and the *MePDI2.3* candidate gene is located directly within the core *CMD2* locus as it was defined earlier [126](Figure 4a). This study used a bi-parental mapping population with 180 segregating F1 plants. Due to the low segregation frequency, this mapping population seemed to be too small and allowed only a rough gene mapping. Later, the *CMD2* locus was partly confirmed with a genome wide association study (GWAS) that used 6,128 African cassava breeding lines for genotyping-by-sequencing (GBS) based marker development. This study revealed not a single geminivirus resistance locus but a large significant association on chromosome 12 that coincided with the region reported earlier [35]. However, their closest *CMD2* associated marker maps ~3 Mb away from the marker identified using the bi-parental mapping population and their complete *CMD2* region spanned ~8 Mb and appeared as two, equally significant peaks. The second peak was thought to be linked to an additional resistance locus (*CMD3*) that has been mapped on the same chromosome as *CMD2*

[33]. Overall, both genetic mapping studies indicate the great uncertainty about the exact genetic location of the single-dominant *CMD2* resistance gene. One reason for the unprecise mapping could be due to the reference genome that was assembled from the South American cassava cultivar AM560 which may not contain the *CMD2*. It is of urgent need to find closer genetic markers for the *CMD2* and we anticipate that the new genetic resources presented here will facilitate the fine mapping and identification of *CMD2* candidate gene(s).

Conclusion

In this study, we presented SCEVT, a QTL visualization pipeline that is capable to deal with highly complex and repetitive genomes and draws a precise synteny map of any desired locus in a diploid whole genome assembly. To show the potential of this new software, we applied SCEVT to reconstruct the *CMD2* locus in the *CMD2*-type TME 3 as well as in the virus susceptible 60444 accessions.

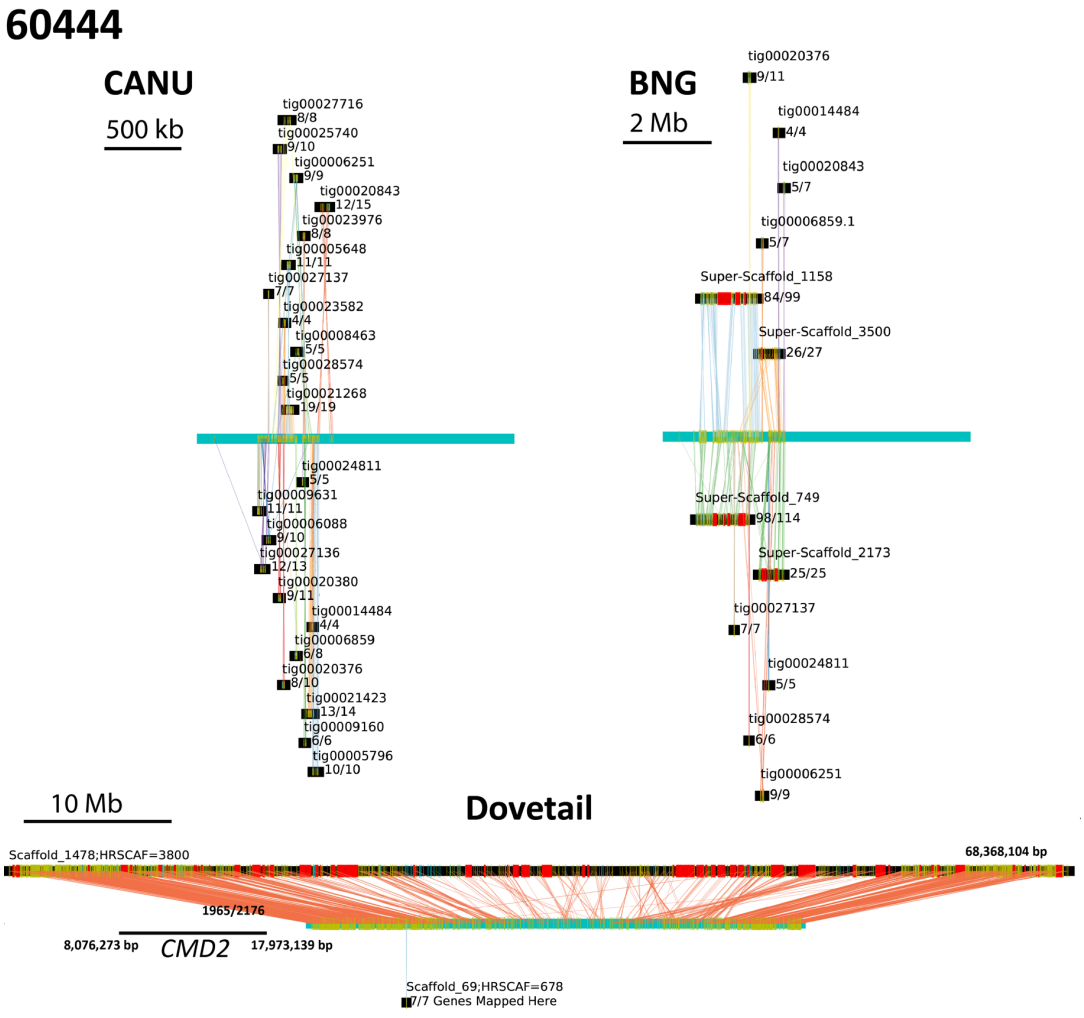
In the case of *CMD2*, SCEVT greatly facilitated the identification of new reference cassava genes linked to the *CMD2* locus. This suggests either that the *CMD2* locus in the reference genome is incompletely assembled or the *de novo* anchored genes are specific for the 60444 and TME 3 genomes. However, a precise gene space annotation can be particularly important for large-scale reverse genetic studies and new trait mapping attempts that require an accurate sequence and gene space annotation. The *de novo* gene space annotation could be further improved using long-read RNA Isoform-Sequencing (Iso-Seq) as it was achieved earlier (Kuon et al. in preparation). The Iso-Seq data were generated from leaf and stem samples and, assumable, do not cover the full gene-space. However, tissue specific Iso-Seq data could be generated to validate current gene space annotation as well as to find new *CMD2* associated genes. Using SCEVT, these new gene space annotations of the *CMD2* region could be directly compared between TME 3 and 60444 to facilitate the identification of CDS contrasting between susceptible and *CMD2*-type cultivars.

SCEVT also revealed the haplotype structure of the *CMD2* as well as the limitations of the current version of the locus (i.e. assembly gaps). With the detailed structure of the haplotypes, a fully assembled and annotated *CMD2* locus becomes more feasible. For example, the haplotype- and optical maps can be used to design haplotype-specific probes for screening bacterial artificial chromosomes (BAC) in order to fill the remaining sequencing gaps. Further attempts should be made to improve the sequence contiguity because a high-quality, near complete assembly of the *CMD2* will be key to identify sequences contributing to the major geminivirus resistance source *CMD2*. By reconstructing the *CMD2* locus, we present progress towards discovering the genetic basis for the major resistance against CMGs. However, the absence of closely linked genetic markers hamper to date the genetic map based isolation of the *CMD2* resistance. Once a dense mapping has been achieved, we believe that our highly contiguous genomes and the haplotype structures will enable the isolation of this important resistance gene.

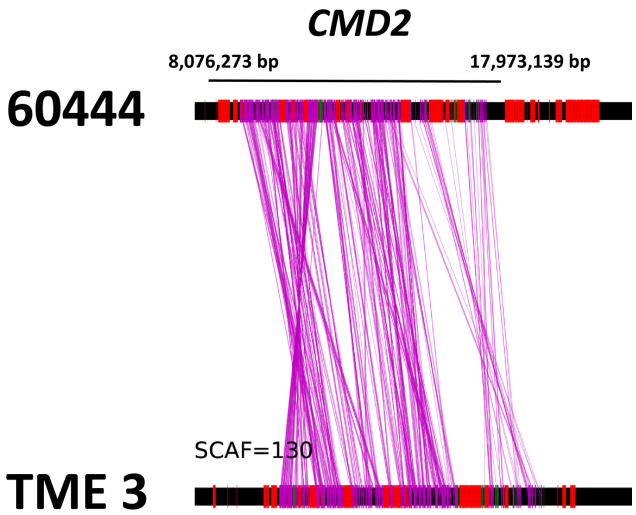
Availability and requirements

Project name: Scaffold and Contig Exploratory Visualization Tool (SCEVT)
Project home page: https://github.com/pbieberstein/SCEVT
Operating system(s): Platform independent
Programming language: Python2.7

Supplementary Notes and Figures



Supplementary Figure 1 SCEVT output for the *CMD2* locus in cassava 60444



Supplementary Figure 2 CMD2 locus comparison between 60444 and TME 3 in Dovetail Hi-C scaffolded genomes

Supplementary Table 1 CMD2 associated de novo TME 3 genes (1/6)

GeneID	start	stop	function
MeTME3_00015713-RA	45261607	45263296	Similar to XTH9: Xyloglucan endotransglucosylase/hydrolase protein 9 (Arabidopsis thaliana)
MeTME3_00015702-RA	44618534	44618997	Protein of unknown function
MeTME3_00015719-RA	45720268	45720972	Protein of unknown function
MeTME3_00015716-RA	45637992	45640048	Similar to AKR1: Probable aldo-keto reductase 1 (Glycine max)
MeTME3_00015717-RA	45683538	45694164	Similar to CDC48C: Cell division control protein 48 homolog C (Arabidopsis thaliana)
MeTME3_00015709-RA	44993636	44999664	Similar to vps18: Vacuolar protein sorting-associated protein 18 homolog (Danio rerio)
MeTME3_00015704-RA	44723148	44723426	Protein of unknown function
MeTME3_00015703-RA	44658088	44660877	Similar to PVA42: Vesicle-associated protein 4-2 (Arabidopsis thaliana)
MeTME3_00015706-RA	44776029	44776550	Similar to VQ31: VQ motif-containing protein 31 (Arabidopsis thaliana)
MeTME3_00015722-RA	45777742	45778086	Similar to PBP1: Calcium-binding protein PBP1 (Arabidopsis thaliana)
MeTME3_00015722-RB	45777835	45778086	Similar to PBP1: Calcium-binding protein PBP1 (Arabidopsis thaliana)
MeTME3_00015714-RA	45492283	45495013	Similar to At3g22104: BTB/POZ domain-containing protein At3g22104 (Arabidopsis thaliana)
MeTME3_00015712-RA	45146292	45146757	Protein of unknown function
MeTME3_00015708-RA	44944318	44945472	Similar to MYB308: Myb-related protein 308 (Antirrhinum majus)
MeTME3_00015721-RA	45749312	45753435	Similar to At4g03230: G-type lectin S-receptor-like serine/threonine-protein kinase At4g0323(
MeTME3_00015707-RA	44861677	44862733	Similar to PMRT15: Protein arginine N-methyltransferase 1.5 (Arabidopsis thaliana)
MeTME3_00015711-RA	45101864	45102142	Protein of unknown function
MeTME3_00015718-RE	45717880	45720033	Similar to PCMP-H43: Pentatricopeptide repeat-containing protein At3g12770 (Arabidopsis tha
MeTME3_00015718-RD	45717880	45720033	Similar to PCMP-H43: Pentatricopeptide repeat-containing protein At3g12770 (Arabidopsis tha
MeTME3_00015718-RC	45717880	45719625	Similar to PCMP-H43: Pentatricopeptide repeat-containing protein At3g12770 (Arabidopsis tha
MeTME3_00015718-RF	45717880	45720033	Similar to PCMP-H43: Pentatricopeptide repeat-containing protein At3g12770 (Arabidopsis tha
MeTME3_00015718-RB	45717880	45719169	Similar to PCMP-H43: Pentatricopeptide repeat-containing protein At3g12770 (Arabidopsis tha
MeTME3_00015718-RA	45717880	45719085	Similar to PCMP-H43: Pentatricopeptide repeat-containing protein At3g12770 (Arabidopsis tha
MeTME3_00015720-RA	45722907	45723968	Similar to FLA21: Fasciclin-like arabinogalactan protein 21 (Arabidopsis thaliana)
MeTME3_00015723-RD	45785967	45787562	Similar to TDC: Aromatic-L-amino-acid decarboxylase (Catharanthus roseus)
MeTME3_00015723-RE	45785967	45787562	Similar to TDC: Aromatic-L-amino-acid decarboxylase (Catharanthus roseus)
MeTME3_00015723-RC	45785967	45787562	Similar to TDC: Aromatic-L-amino-acid decarboxylase (Catharanthus roseus)
MeTME3_00015723-RB	45785967	45786923	Similar to TDC: Aromatic-L-amino-acid decarboxylase (Catharanthus roseus)
MeTME3_00015723-RA	45785967	45786899	Similar to TDC: Aromatic-L-amino-acid decarboxylase (Catharanthus roseus)
MeTME3_00015715-RA	45617453	45618048	Similar to MIP1B: B-box domain protein 31 (Arabidopsis thaliana)
MeTME3_00015705-RA	44725903	44726369	Similar to At5g08350: GEM-like protein 4 (Arabidopsis thaliana)
MeTME3_00015710-RA	45098771	45099228	Similar to At5g08350: GEM-like protein 4 (Arabidopsis thaliana)
MeTME3_00015746-RB	47637510	47645626	Similar to At5g35735: Cytochrome b561 and DOMON domain-containing protein At5g35735
MeTME3_00015746-RA	47637510	47645626	Similar to At5g35735: Cytochrome b561 and DOMON domain-containing protein At5g35735
MeTME3_00015741-RA	47228882	47233165	Similar to Calcium-dependent protein kinase SK5 (Glycine max)
MeTME3_00015743-RA	47282415	47287528	Similar to SGS3: Protein SUPPRESSOR OF GENE SILENCING 3 (Solanum lycopersicum)
MeTME3_00015725-RA	46161499	46162685	Similar to MYB4: Transcription factor MYB4 (Oryza sativa subsp. japonica)
MeTME3_00015732-RA	46787015	46795293	Protein of unknown function
MeTME3_00015729-RA	46366224	46370413	Similar to At4g03230: G-type lectin S-receptor-like serine/threonine-protein kinase At4g0323(
MeTME3_00015745-RA	47606659	47635511	Similar to CBSCBSPB3: CBS domain-containing protein CBSCBSPB3 (Arabidopsis thaliana)
MeTME3_00015734-RA	46876817	46877872	Similar to PMRT15: Protein arginine N-methyltransferase 1.5 (Arabidopsis thaliana)
MeTME3_00015730-RA	46399086	46399433	Similar to PBP1: Calcium-binding protein PBP1 (Arabidopsis thaliana)
MeTME3_00015730-RB	46399179	46399433	Similar to PBP1: Calcium-binding protein PBP1 (Arabidopsis thaliana)
MeTME3_00015742-RA	47240497	47241018	Similar to VQ31: VQ motif-containing protein 31 (Arabidopsis thaliana)
MeTME3_00015740-RA	47116964	47119521	Similar to LCB1: Long chain base biosynthesis protein 1 (Arabidopsis thaliana)
MeTME3_00015728-RA	46264042	46265102	Similar to MYB15: Transcription factor MYB15 (Arabidopsis thaliana)
MeTME3_00015735-RA	46882281	46886015	Similar to ULT1: Protein ULTRAPETALA 1 (Arabidopsis thaliana)
MeTME3_00015737-RA	46994632	46995622	Similar to MYB308: Myb-related protein 308 (Antirrhinum majus)

Supplementary Table 2 *CMD2* associated *de novo* TME 3 genes (2/6)

GeneID	start	stop	function
MeTME3_00015724-RA	46098598	46102669	Similar to PHOS34: Universal stress protein PHOS34 (Arabidopsis thaliana)
MeTME3_00015733-RA	46810565	46832571	Similar to PMRT15: Protein arginine N-methyltransferase 1.5 (Arabidopsis thaliana)
MeTME3_00015733-RB	46810565	46832571	Similar to PMRT15: Protein arginine N-methyltransferase 1.5 (Arabidopsis thaliana)
MeTME3_00015744-RA	47301642	47322068	Similar to clptm1: Cleft lip and palate transmembrane protein 1 homolog (Danio rerio)
MeTME3_00015738-RB	47024246	47078815	Similar to At5g35735: Cytochrome b561 and DOMON domain-containing protein At5g35735
MeTME3_00015738-RC	47024246	47078815	Similar to At5g35735: Cytochrome b561 and DOMON domain-containing protein At5g35735
MeTME3_00015738-RA	47024246	47078815	Similar to At5g47530: Cytochrome b561 and DOMON domain-containing protein At5g47530
MeTME3_00015738-RE	47078754	47080350	Similar to At5g35735: Cytochrome b561 and DOMON domain-containing protein At5g35735
MeTME3_00015738-RD	47024246	47080350	Similar to At5g35735: Cytochrome b561 and DOMON domain-containing protein At5g35735
MeTME3_00015739-RA	47081968	47112070	Similar to CBSCBSPB3: CBS domain-containing protein CBSCBSPB3 (Arabidopsis thaliana)
MeTME3_00015731-RB	46413668	46415191	Similar to TDC: Aromatic-L-amino-acid decarboxylase (Catharanthus roseus)
MeTME3_00015731-RA	46413668	46415191	Similar to TDC: Aromatic-L-amino-acid decarboxylase (Catharanthus roseus)
MeTME3_00015727-RA	46226994	46227215	Similar to Auxin-responsive protein SAUR50 (Helianthus annuus)
MeTME3_00015726-RA	46188132	46188311	Similar to Auxin-responsive protein SAUR50 (Helianthus annuus)
MeTME3_00015736-RA	46903413	46904734	Protein of unknown function
MeTME3_00015758-RA	49233778	49235186	Similar to COMT1: Caffeic acid 3-O-methyltransferase (Prunus dulcis)
MeTME3_00015760-RA	49256427	49259875	Similar to GLR3.4: Glutamate receptor 3.4 (Arabidopsis thaliana)
MeTME3_00015747-RB	48604499	48616428	Similar to Flad1: FAD synthase (Mus musculus)
MeTME3_00015747-RC	48604499	48616428	Similar to Flad1: FAD synthase (Mus musculus)
MeTME3_00015747-RA	48604499	48616428	Similar to SPCC1235.04c: Probable FAD synthase (Schizosaccharomyces pombe (strain 972 / A
MeTME3_00015748-RA	48634049	48636019	Similar to EPHX2: Bifunctional epoxide hydrolase 2 (Homo sapiens)
MeTME3_00015757-RA	49196174	49209232	Similar to ACO3: Aconitate hydratase 3%2C mitochondrial (Arabidopsis thaliana)
MeTME3_00015752-RD	48991595	49182854	Similar to AXR1: NEDD8-activating enzyme E1 regulatory subunit AXR1 (Arabidopsis thaliana)
MeTME3_00015752-RC	48991595	49182854	Similar to AXR1: NEDD8-activating enzyme E1 regulatory subunit AXR1 (Arabidopsis thaliana)
MeTME3_00015752-RB	48908393	48993947	Similar to TPS11: Probable terpene synthase 11 (Ricinus communis)
MeTME3_00015752-RA	48907734	48908390	Similar to TPS11: Probable terpene synthase 11 (Ricinus communis)
MeTME3_00015749-RA	48643074	48644385	Similar to EPHX2: Bifunctional epoxide hydrolase 2 (Sus scrofa)
MeTME3_00015750-RA	48669633	48673250	Similar to TPS9: Probable terpene synthase 9 (Ricinus communis)
MeTME3_00015756-RA	49178670	49179744	Similar to SKIP5: F-box protein SKIP5 (Arabidopsis thaliana)
MeTME3_00015755-RA	49119360	49122214	Similar to Bp10: L-ascorbate oxidase homolog (Brassica napus)
MeTME3_00015759-RA	49240747	49243032	Similar to RPL6: 50S ribosomal protein L6%2C chloroplastic (Arabidopsis thaliana)
MeTME3_00015753-RA	49007831	49020452	Similar to RRP6L3: Protein RRP6-like 3 (Arabidopsis thaliana)
MeTME3_00015751-RA	48694497	48697242	Similar to TPS12: Probable terpene synthase 12 (Ricinus communis)
MeTME3_00015754-RA	49057167	49057623	Similar to ARASP2: Probable membrane metalloprotease ARASP2%2C chloroplastic (Arabidopsi
MeTME3_00015761-RA	49267560	49267763	Similar to spg1: Septum-promoting GTP-binding protein 1 (Schizosaccharomyces pombe (strai
MeTME3_00015800-RA	50517127	50518734	Similar to Isocitrate lyase (Ricinus communis)
MeTME3_00015792-RA	50289005	50310891	Similar to GLR3.4: Glutamate receptor 3.4 (Arabidopsis thaliana)
MeTME3_00015763-RA	49341911	49342552	Similar to tmem97: Transmembrane protein 97 (Xenopus tropicalis)
MeTME3_00015778-RA	49942227	49949466	Similar to PEX22: Peroxisome biogenesis protein 22 (Arabidopsis thaliana)
MeTME3_00015794-RA	50358983	50359630	Similar to tmem97: Transmembrane protein 97 (Xenopus tropicalis)
MeTME3_00015769-RA	49504514	49577612	Similar to PER3: Peroxidase 3 (Arabidopsis thaliana)
MeTME3_00015799-RA	50485696	50492761	Similar to BRIX1-1: Ribosome biogenesis protein BRX1 homolog 1 (Arabidopsis thaliana)
MeTME3_00015807-RA	50881828	50888916	Similar to AXR1: NEDD8-activating enzyme E1 regulatory subunit AXR1 (Arabidopsis thaliana)
MeTME3_00015803-RA	50753759	50755996	Similar to RPL6: 50S ribosomal protein L6%2C chloroplastic (Arabidopsis thaliana)
MeTME3_00015783-RA	50060759	50064274	Similar to FH5: Formin-like protein 5 (Arabidopsis thaliana)
MeTME3_00015776-RB	49818893	49826145	Similar to At3g04600: Tryptophan--tRNA ligase%2C cytoplasmic (Arabidopsis thaliana)
MeTME3_00015776-RA	49818893	49825725	Similar to At3g04600: Tryptophan--tRNA ligase%2C cytoplasmic (Arabidopsis thaliana)
MeTME3_00015788-RA	50154890	50227348	Similar to TPS12: Probable terpene synthase 12 (Ricinus communis)
MeTME3_00015766-RA	49430207	49434512	Similar to SYP131: Putative syntaxin-131 (Arabidopsis thaliana)
MeTME3_00015770-RA	49611351	49651993	Similar to POLD1: DNA polymerase delta catalytic subunit (Oryza sativa subsp. japonica)
MeTME3_00015785-RA	50144138	50151675	Similar to At1g61730: Probable transcription factor At1g61730 (Arabidopsis thaliana)

Supplementary Table 3 *CMD2* associated *de novo* TME 3 genes (3/6)

GeneID	start	stop	function
MeTME3_00015784-RA	50113681	50115866	Similar to At1g61730: Probable transcription factor At1g61730 (Arabidopsis thaliana)
MeTME3_00015786-RA	50151943	50154356	Similar to TPS9: Probable terpene synthase 9 (Ricinus communis)
MeTME3_00015771-RA	49652432	49779551	Similar to BMY1: Beta-amylase (Glycine max)
MeTME3_00015775-RA	49800604	49817479	Similar to At3g21810: Zinc finger CCCH domain-containing protein 40 (Arabidopsis thaliana)
MeTME3_00015774-RA	49779740	49780172	Protein of unknown function
MeTME3_00015782-RA	50012021	50012813	Protein of unknown function
MeTME3_00015772-RA	49713733	49721169	Similar to PEX22: Peroxisome biogenesis protein 22 (Arabidopsis thaliana)
MeTME3_00015768-RA	49473861	49480948	Similar to BRIX1-1: Ribosome biogenesis protein BRX1 homolog 1 (Arabidopsis thaliana)
MeTME3_00015802-RA	50744355	50748865	Similar to GLR3.7: Glutamate receptor 3.7 (Arabidopsis thaliana)
MeTME3_00015804-RA	50761763	50763211	Similar to COMT1: Caffeic acid 3-O-methyltransferase (Prunus dulcis)
MeTME3_00015789-RA	50175555	50178313	Similar to TPS12: Probable terpene synthase 12 (Ricinus communis)
MeTME3_00015779-RA	49965542	49969997	Similar to BMY1: Beta-amylase (Glycine max)
MeTME3_00015797-RA	50413762	50421999	Similar to POLD1: DNA polymerase delta catalytic subunit (Glycine max)
MeTME3_00015805-RA	50785141	50798065	Similar to ACO3: Aconitate hydratase 3%2C mitochondrial (Arabidopsis thaliana)
MeTME3_00015762-RA	49299083	49304084	Similar to ATJ10: Chaperone protein dnaJ 10 (Arabidopsis thaliana)
MeTME3_00015767-RA	49444871	49446324	Similar to Isocitrate lyase (Gossypium hirsutum)
MeTME3_00015764-RA	49357140	49362907	Similar to HDG2: Homeobox-leucine zipper protein HDG2 (Arabidopsis thaliana)
MeTME3_00015795-RA	50377775	50383576	Similar to HDG2: Homeobox-leucine zipper protein HDG2 (Arabidopsis thaliana)
MeTME3_00015781-RA	50001072	50008386	Similar to At3g04600: Tryptophan--tRNA ligase%2C cytoplasmic (Arabidopsis thaliana)
MeTME3_00015801-RA	50521680	50524290	Similar to SYP132: Syntaxin-132 (Arabidopsis thaliana)
MeTME3_00015791-RA	50273884	50275440	Similar to RPL6: 50S ribosomal protein L6%2C chloroplastic (Arabidopsis thaliana)
MeTME3_00015798-RA	50439575	50479961	Similar to PER39: Peroxidase 39 (Arabidopsis thaliana)
MeTME3_00015790-RA	50257721	50265654	Similar to AXR1: NEDD8-activating enzyme E1 regulatory subunit AXR1 (Arabidopsis thaliana)
MeTME3_00015793-RA	50345965	50348353	Similar to spg1: Septum-promoting GTP-binding protein 1 (Schizosaccharomyces pombe (strai
MeTME3_00015773-RA	49751435	49752062	Similar to MIP1A: B-box domain protein 30 (Arabidopsis thaliana)
MeTME3_00015780-RA	49987480	49987797	Protein of unknown function
MeTME3_00015777-RA	49846674	49846991	Protein of unknown function
MeTME3_00015787-RA	50154468	50154739	Similar to TPS9: Probable terpene synthase 9 (Ricinus communis)
MeTME3_00015765-RA	49363053	49363262	Protein of unknown function
MeTME3_00015796-RA	50383722	50383931	Protein of unknown function
MeTME3_00015839-RA	52433570	52438111	Similar to PHYC: Phytochrome C (Oryza sativa subsp. indica)
MeTME3_00015838-RA	52425239	52427989	Similar to GNT2: Alpha-1%2C6-mannosyl-glycoprotein 2-beta-N-acetylglucosaminyltransferase
MeTME3_00015808-RA	51280539	51283796	Protein of unknown function
MeTME3_00015836-RA	52410512	52413597	Similar to KINB2: SNF1-related protein kinase regulatory subunit beta-2 (Arabidopsis thaliana)
MeTME3_00015815-RA	51758659	51759039	Similar to CERK1: Chitin elicitor receptor kinase 1 (Arabidopsis thaliana)
MeTME3_00015817-RA	51817805	51821916	Similar to Cnot9: CCR4-NOT transcription complex subunit 9 (Rattus norvegicus)
MeTME3_00015823-RA	52089201	52095329	Similar to CYP74B2: Linolenate hydroperoxide lyase%2C chloroplastic (Arabidopsis thaliana)
MeTME3_00015835-RA	52391951	52393694	Similar to CYP82C4: Cytochrome P450 82C4 (Arabidopsis thaliana)
MeTME3_00015819-RA	51848836	51851412	Protein of unknown function
MeTME3_00015833-RA	52296271	52300574	Similar to B3GALT2: Probable beta-1%2C3-galactosyltransferase 2 (Arabidopsis thaliana)
MeTME3_00015828-RA	52171699	52176300	Similar to cnot9: CCR4-NOT transcription complex subunit 9 (Xenopus tropicalis)
MeTME3_00015825-RA	52129317	52133489	Protein of unknown function
MeTME3_00015831-RA	52209464	52212034	Similar to Bp10: L-ascorbate oxidase homolog (Brassica napus)
MeTME3_00015806-RA	50835416	50908150	Similar to TPS12: Probable terpene synthase 12 (Ricinus communis)
MeTME3_00015809-RA	51327528	51329610	Similar to UP3: Stress-response A/B barrel domain-containing protein UP3 (Arabidopsis thaliar
MeTME3_00015832-RA	52266210	52270870	Similar to RRP6L3: Protein RRP6-like 3 (Arabidopsis thaliana)

Supplementary Table 4 *CMD2* associated *de novo* TME 3 genes (4/6)

GeneID	start	stop	function
MeTME3_00015827-RA	52150197	52152867	Similar to WRKY42: WRKY transcription factor 42 (Arabidopsis thaliana)
MeTME3_00015829-RA	52179405	52181502	Protein of unknown function
MeTME3_00015821-RA	51900127	51984533	Similar to At3g21620: CSC1-like protein At3g21620 (Arabidopsis thaliana)
MeTME3_00015820-RA	51852052	51858564	Similar to HPL: Fatty acid hydroperoxide lyase%2C chloroplastic (Solanum lycopersicum)
MeTME3_00015824-RA	52096003	52098518	Protein of unknown function
MeTME3_00015813-RA	51653515	51655568	Similar to B'ZETA: Serine/threonine protein phosphatase 2A 59 kDa regulatory subunit B' zeta
MeTME3_00015822-RA	51984653	51988215	Similar to CSC1: Calcium permeable stress-gated cation channel 1 (Arabidopsis thaliana)
MeTME3_00015822-RB	51988215	51989673	Similar to At4g15430: CSC1-like protein At4g15430 (Arabidopsis thaliana)
MeTME3_00015816-RA	51793575	51795521	Similar to WRKY31: Probable WRKY transcription factor 31 (Arabidopsis thaliana)
MeTME3_00015811-RA	51598318	51599087	Similar to CERK1: Chitin elicitor receptor kinase 1 (Arabidopsis thaliana)
MeTME3_00015810-RA	51594265	51596914	Similar to CERK1: Chitin elicitor receptor kinase 1 (Arabidopsis thaliana)
MeTME3_00015826-RA	52135044	52139153	Similar to CNOT9: CCR4-NOT transcription complex subunit 9 (Pongo abelii)
MeTME3_00015814-RA	51752967	51777562	Similar to MAPKKK17: Mitogen-activated protein kinase kinase kinase 17 (Arabidopsis thaliana
MeTME3_00015812-RA	51636906	51645078	Protein of unknown function
MeTME3_00015818-RA	51823635	51830564	Protein of unknown function
MeTME3_00015834-RA	52321516	52321972	Similar to ARASP2: Probable membrane metalloprotease ARASP2%2C chloroplastic (Arabidopsi
MeTME3_00015837-RA	52418654	52421899	Similar to PCMP-H35: Putative pentatricopeptide repeat-containing protein At5g09950 (Arabid
MeTME3_00015830-RA	52203124	52205237	Protein of unknown function
MeTME3_00015851-RA	52713297	52716734	Similar to RIN4: RPM1-interacting protein 4 (Arabidopsis thaliana)
MeTME3_00015849-RA	52689477	52691427	Similar to sll1770: Uncharacterized protein sll1770 (Synechocystis sp. (strain PCC 6803 / Kazu
MeTME3_00015895-RA	54092364	54093940	Similar to At2g01630: Glucan endo-1%2C3-beta-glucosidase 3 (Arabidopsis thaliana)
MeTME3_00015888-RA	54001131	54008675	Similar to EMB3004: Bifunctional 3-dehydroquinate dehydratase/shikimate dehydrogenase%2C
MeTME3_00015840-RA	52506701	52510386	Similar to SAP: Transcriptional regulator STERILE APETALA (Arabidopsis thaliana)
MeTME3_00015879-RA	53768204	53775614	Similar to COP1: E3 ubiquitin-protein ligase COP1 (Arabidopsis thaliana)
MeTME3_00015892-RA	54061058	54066161	Similar to AUG2: AUGMIN subunit 2 (Arabidopsis thaliana)
MeTME3_00015892-RB	54061058	54066161	Similar to AUG2: AUGMIN subunit 2 (Arabidopsis thaliana)
MeTME3_00015859-RA	53301708	53306739	Similar to At1g05000: Probable tyrosine-protein phosphatase At1g05000 (Arabidopsis thalian
MeTME3_00015870-RA	53593285	53598411	Similar to PDIL2-3: Protein disulfide isomerase-like 2-3 (Oryza sativa subsp. japonica)
MeTME3_00015864-RA	53423624	53428545	Similar to AGPS1: Glucose-1-phosphate adenylyltransferase small subunit%2C chloroplastic (Br
MeTME3_00015848-RA	52685002	52687165	Similar to sll1770: Uncharacterized protein sll1770 (Synechocystis sp. (strain PCC 6803 / Kazu
MeTME3_00015865-RA	53448668	53452620	Similar to CAX5: Vacuolar cation/proton exchanger 5 (Arabidopsis thaliana)
MeTME3_00015858-RA	53245376	53252726	Similar to At1g04990: Zinc finger CCCH domain-containing protein 3 (Arabidopsis thaliana)
MeTME3_00015894-RA	54081166	54081833	Protein of unknown function
MeTME3_00015880-RA	53813232	53816912	Similar to KINB2: SNF1-related protein kinase regulatory subunit beta-2 (Arabidopsis thaliana)
MeTME3_00015891-RA	54054082	54059553	Similar to NAT1: Nucleobase-ascorbate transporter 1 (Arabidopsis thaliana)
MeTME3_00015881-RA	53820222	53830208	Similar to PCMP-H35: Putative pentatricopeptide repeat-containing protein At5g09950 (Arabid
MeTME3_00015850-RA	52704593	52709827	Similar to AAT1: Acetyl-CoA acetyltransferase%2C cytosolic 1 (Arabidopsis thaliana)
MeTME3_00015861-RA	53366876	53368906	Similar to LIP2: Triacylglycerol lipase 2 (Arabidopsis thaliana)
MeTME3_00015893-RA	54079280	54080108	Protein of unknown function
MeTME3_00015890-RA	54018270	54028531	Protein of unknown function
MeTME3_00015868-RA	53527733	53530343	Similar to At3g27390: Uncharacterized membrane protein At3g27390 (Arabidopsis thaliana)
MeTME3_00015854-RA	52813052	52822730	Protein of unknown function
MeTME3_00015842-RA	52572180	52575959	Similar to At4g00590: Putative threonine aspartase (Arabidopsis thaliana)
MeTME3_00015855-RA	52839347	52841619	Similar to PUB3: U-box domain-containing protein 3 (Arabidopsis thaliana)
MeTME3_00015843-RA	52580208	52583085	Similar to RLP12: Receptor-like protein 12 (Arabidopsis thaliana)
MeTME3_00015841-RA	52561074	52563968	Similar to RLP12: Receptor-like protein 12 (Arabidopsis thaliana)
MeTME3_00015876-RA	53696451	53697228	Similar to HIPP16: Heavy metal-associated isoprenylated plant protein 16 (Arabidopsis thaliana
MeTME3_00015845-RB	52659612	52660171	Similar to NFYB3: Nuclear transcription factor Y subunit B-3 (Arabidopsis thaliana)
MeTME3_00015845-RA	52659603	52660141	Similar to NFYB3: Nuclear transcription factor Y subunit B-3 (Arabidopsis thaliana)
MeTME3_00015847-RA	52683217	52684409	Protein of unknown function
MeTME3_00015885-RA	53977521	53984132	Protein of unknown function

Supplementary Table 5 *CMD2* associated *de novo* TME 3 genes (5/6)

GeneID	start	stop	function
MeTME3_00015889-RA	54010844	54013323	Protein of unknown function
MeTME3_00015863-RA	53402191	53406952	Similar to COL9: Zinc finger protein CONSTANS-LIKE 9 (Arabidopsis thaliana)
MeTME3_00015853-RA	52766830	52773146	Similar to IGPS: Indole-3-glycerol phosphate synthase%2C chloroplastic (Arabidopsis thaliana)
MeTME3_00015873-RA	53646619	53651541	Similar to AGPS1: Glucose-1-phosphate adenylyltransferase small subunit%2C chloroplastic (B
MeTME3_00015887-RA	53987177	53990670	Similar to At2g32990: Endoglucanase 11 (Arabidopsis thaliana)
MeTME3_00015860-RA	53313075	53314738	Similar to ACO1: 1-aminocyclopropane-1-carboxylate oxidase (Prunus mume)
MeTME3_00015878-RA	53746184	53752510	Similar to At1g04990: Zinc finger CCCH domain-containing protein 3 (Arabidopsis thaliana)
MeTME3_00015886-RA	53983665	53986029	Similar to NPF6.4: Protein NRT1/ PTR FAMILY 6.4 (Arabidopsis thaliana)
MeTME3_00015862-RA	53376580	53378614	Similar to LIP2: Triacylglycerol lipase 2 (Arabidopsis thaliana)
MeTME3_00015846-RA	52666420	52668137	Similar to NUDT2: Nudix hydrolase 2 (Arabidopsis thaliana)
MeTME3_00015877-RA	53719584	53724626	Similar to At1g05000: Probable tyrosine-protein phosphatase At1g05000 (Arabidopsis thalian
MeTME3_00015857-RA	52849837	52857678	Similar to RNE: Ribonuclease E/G-like protein%2C chloroplastic (Arabidopsis thaliana)
MeTME3_00015874-RA	53654571	53656371	Similar to AIR3: Subtilisin-like protease SBT5.3 (Arabidopsis thaliana)
MeTME3_00015884-RA	53961388	53968492	Similar to EBS: Chromatin remodeling protein EBS (Arabidopsis thaliana)
MeTME3_00015884-RB	53965241	53968492	Similar to EBS: Chromatin remodeling protein EBS (Arabidopsis thaliana)
MeTME3_00015875-RA	53658628	53659724	Similar to AIR3: Subtilisin-like protease SBT5.3 (Arabidopsis thaliana)
MeTME3_00015852-RA	52717332	52722212	Protein of unknown function
MeTME3_00015882-RA	53855914	53868977	Similar to EBS: Chromatin remodeling protein EBS (Arabidopsis thaliana)
MeTME3_00015882-RB	53858145	53868977	Similar to EBS: Chromatin remodeling protein EBS (Arabidopsis thaliana)
MeTME3_00015872-RA	53629579	53633911	Similar to CAX5: Vacuolar cation/proton exchanger 5 (Arabidopsis thaliana)
MeTME3_00015869-RA	53534367	53537038	Protein of unknown function
MeTME3_00015867-RA	53520701	53524086	Similar to At5g03795: Probable glycosyltransferase At5g03795 (Arabidopsis thaliana)
MeTME3_00015883-RA	53878430	53882976	Similar to EBS: Chromatin remodeling protein EBS (Arabidopsis thaliana)
MeTME3_00015871-RB	53600743	53608375	Similar to FPA: Flowering time control protein FPA (Arabidopsis thaliana)
MeTME3_00015871-RA	53600743	53608317	Similar to FPA: Flowering time control protein FPA (Arabidopsis thaliana)
MeTME3_00015866-RA	53470081	53470878	Similar to HIPP16: Heavy metal-associated isoprenylated plant protein 16 (Arabidopsis thalian
MeTME3_00015856-RA	52846788	52847666	Protein of unknown function
MeTME3_00015844-RA	52590236	52590463	Protein of unknown function
MeTME3_00015922-RA	55186315	55190992	Similar to VPS11: Vacuolar protein-sorting-associated protein 11 homolog (Arabidopsis thalian
MeTME3_00015915-RA	55062753	55066445	Similar to SAP: Transcriptional regulator STERILE APETALA (Arabidopsis thaliana)
MeTME3_00015924-RA	55204701	55215592	Similar to NUP133: Nuclear pore complex protein NUP133 (Arabidopsis thaliana)
MeTME3_00015923-RA	55197833	55202304	Protein of unknown function
MeTME3_00015923-RB	55198130	55202304	Protein of unknown function
MeTME3_00015906-RA	54759975	54766311	Similar to IGPS: Indole-3-glycerol phosphate synthase%2C chloroplastic (Arabidopsis thaliana)
MeTME3_00015952-RA	55654366	55657839	Similar to At1g04970: Putative BPI/LBP family protein At1g04970 (Arabidopsis thaliana)
MeTME3_00015896-RA	54218159	54224939	Similar to COP1: E3 ubiquitin-protein ligase COP1 (Arabidopsis thaliana)
MeTME3_00015947-RA	55569372	55575386	Similar to FIM5: Fimbrin-5 (Arabidopsis thaliana)
MeTME3_00015945-RA	55538107	55540440	Similar to GDI1: Rho GDP-dissociation inhibitor 1 (Arabidopsis thaliana)
MeTME3_00015955-RA	55679362	55682276	Similar to IBR5: Protein-tyrosine-phosphatase IBR5 (Arabidopsis thaliana)
MeTME3_00015926-RA	55261482	55264070	Similar to OBE2: Protein OBERON 2 (Arabidopsis thaliana)
MeTME3_00015902-RA	54637939	54647043	Similar to RNE: Ribonuclease E/G-like protein%2C chloroplastic (Arabidopsis thaliana)
MeTME3_00015913-RA	54860746	54862372	Similar to NUDT2: Nudix hydrolase 2 (Arabidopsis thaliana)
MeTME3_00015949-RA	55621745	55622152	Protein of unknown function
MeTME3_00015925-RA	55245422	55252217	Similar to HSP90-6: Heat shock protein 90-6%2C mitochondrial (Arabidopsis thaliana)

Supplementary Table 6 *CMD2* associated *de novo* TME 3 genes (6/6)

GeneID	start	stop	function
MeTME3_00015907-RA	54770518	54806304	Protein of unknown function
MeTME3_00015907-RB	54801046	54806304	Protein of unknown function
MeTME3_00015940-RA	55463715	55465397	Similar to At5g47530: Cytochrome b561 and DOMON domain-containing protein At5g47530
MeTME3_00015938-RA	55450726	55452240	Similar to At5g35735: Cytochrome b561 and DOMON domain-containing protein At5g35735
MeTME3_00015927-RA	55294916	55303574	Similar to AHK2: Histidine kinase 2 (Arabidopsis thaliana)
MeTME3_00015929-RA	55322670	55324271	Similar to PAT1: Scarecrow-like transcription factor PAT1 (Arabidopsis thaliana)
MeTME3_00015918-RA	55109423	55113171	Similar to At4g00590: Putative threonine aspartase (Arabidopsis thaliana)
MeTME3_00015939-RA	55454340	55458512	Similar to At5g47530: Cytochrome b561 and DOMON domain-containing protein At5g47530
MeTME3_00015933-RA	55357097	55358080	Similar to WNK11: Probable serine/threonine-protein kinase WNK11 (Arabidopsis thaliana)
MeTME3_00015899-RA	54277617	54279312	Similar to LIP2: Triacylglycerol lipase 2 (Arabidopsis thaliana)
MeTME3_00015897-RA	54238110	54240144	Similar to LIP2: Triacylglycerol lipase 2 (Arabidopsis thaliana)
MeTME3_00015898-RA	54256281	54258046	Similar to LIP2: Triacylglycerol lipase 2 (Arabidopsis thaliana)
MeTME3_00015917-RA	55102295	55103497	Similar to RLP12: Receptor-like protein 12 (Arabidopsis thaliana)
MeTME3_00015916-RA	55100604	55102181	Similar to RLP12: Receptor-like protein 12 (Arabidopsis thaliana)
MeTME3_00015932-RA	55350248	55352916	Similar to Es2: Protein DGCR14 homolog (Drosophila melanogaster)
MeTME3_00015901-RA	54632213	54632719	Similar to XERICO: Probable E3 ubiquitin-protein ligase XERICO (Arabidopsis thaliana)
MeTME3_00015946-RB	55546013	55546486	Similar to FLA7: Fasciclin-like arabinogalactan protein 7 (Arabidopsis thaliana)
MeTME3_00015946-RA	55546013	55546474	Similar to FLA7: Fasciclin-like arabinogalactan protein 7 (Arabidopsis thaliana)
MeTME3_00015908-RA	54806760	54810159	Similar to RIN4: RPM1-interacting protein 4 (Arabidopsis thaliana)
MeTME3_00015957-RA	55690023	55693594	Protein of unknown function
MeTME3_00015948-RA	55576377	55580286	Similar to At2g04740: BTB/POZ domain-containing protein At2g04740 (Arabidopsis thaliana)
MeTME3_00015910-RA	54833911	54835862	Similar to sll1770: Uncharacterized protein sll1770 (Synechocystis sp. (strain PCC 6803 / Kaz
MeTME3_00015928-RA	55306863	55309176	Similar to Transmembrane protein 256 homolog (Bufo gargarizans)
MeTME3_00015951-RA	55633386	55648711	Protein of unknown function
MeTME3_00015944-RA	55525581	55526461	Protein of unknown function
MeTME3_00015911-RA	54838316	54840480	Similar to sll1770: Uncharacterized protein sll1770 (Synechocystis sp. (strain PCC 6803 / Kaz
MeTME3_00015921-RA	55157677	55160444	Similar to At2g05160: Zinc finger CCCH domain-containing protein 18 (Arabidopsis thaliana)
MeTME3_00015909-RA	54813409	54818665	Similar to AAT1: Acetyl-CoA acetyltransferase%2C cytosolic 1 (Arabidopsis thaliana)
MeTME3_00015935-RA	55379787	55381540	Similar to TK: Thymidine kinase (Oryza sativa subsp. japonica)
MeTME3_00015941-RA	55465592	55469907	Similar to EMB2761: Threonine--tRNA ligase%2C chloroplastic/mitochondrial 2 (Arabidopsis th
MeTME3_00015942-RA	55479849	55480172	Protein of unknown function
MeTME3_00015954-RA	55667542	55672073	Similar to At3g58140: Phenylalanine--tRNA ligase%2C chloroplastic/mitochondrial (Arabidopsi
MeTME3_00015900-RA	54285803	54287831	Similar to LIP2: Triacylglycerol lipase 2 (Arabidopsis thaliana)
MeTME3_00015934-RA	55366617	55370280	Similar to At2g04865: Protein MAIN-LIKE 2 (Arabidopsis thaliana)
MeTME3_00015931-RA	55337984	55340689	Similar to WRKY1: WRKY transcription factor 1 (Arabidopsis thaliana)
MeTME3_00015953-RA	55660327	55665114	Similar to RDM4: RNA-directed DNA methylation 4 (Arabidopsis thaliana)
MeTME3_00015937-RA	55392048	55442180	Similar to At5g47530: Cytochrome b561 and DOMON domain-containing protein At5g47530
MeTME3_00015937-RB	55392048	55442180	Similar to At5g47530: Cytochrome b561 and DOMON domain-containing protein At5g47530
MeTME3_00015905-RA	54713753	54723486	Protein of unknown function
MeTME3_00015936-RA	55382662	55391983	Similar to At2g04850: Cytochrome b561 and DOMON domain-containing protein At2g04850
MeTME3_00015956-RA	55686843	55688844	Protein of unknown function
MeTME3_00015904-RA	54654826	54657102	Similar to PUB3: U-box domain-containing protein 3 (Arabidopsis thaliana)
MeTME3_00015930-RA	55336444	55337721	Similar to WRKY1: WRKY transcription factor 1 (Arabidopsis thaliana)
MeTME3_00015912-RA	54841072	54842263	Protein of unknown function
MeTME3_00015914-RA	54874300	54874857	Similar to NFYB3: Nuclear transcription factor Y subunit B-3 (Arabidopsis thaliana)
MeTME3_00015914-RB	54874330	54874866	Similar to NFYB3: Nuclear transcription factor Y subunit B-3 (Arabidopsis thaliana)
MeTME3_00015950-RA	55629467	55629832	Similar to Eukaryotic translation initiation factor 1A (Onobrychis viciifolia)
MeTME3_00015950-RB	55629482	55629832	Similar to Eukaryotic translation initiation factor 1A (Onobrychis viciifolia)
MeTME3_00015950-RD	55629506	55629832	Similar to Eukaryotic translation initiation factor 1A (Onobrychis viciifolia)
MeTME3_00015950-RC	55629506	55629712	Similar to Eukaryotic translation initiation factor 1A (Onobrychis viciifolia)
MeTME3_00015950-RE	55629524	55629712	Similar to Eukaryotic translation initiation factor 1A (Onobrychis viciifolia)
MeTME3_00015919-RA	55116540	55119418	Similar to RLP12: Receptor-like protein 12 (Arabidopsis thaliana)
MeTME3_00015943-RA	55507046	55520170	Similar to At3g07870: F-box protein At3g07870 (Arabidopsis thaliana)
MeTME3_00015903-RA	54647971	54650115	Protein of unknown function
MeTME3_00015920-RA	55142482	55142709	Protein of unknown function
MeTME3_00015959-RA	55741025	55742994	Similar to At2g04570: GDSL esterase/lipase At2g04570 (Arabidopsis thaliana)
MeTME3_00015960-RA	55764290	55773322	Similar to ZW10: Centromere/kinetochore protein zw10 homolog (Arabidopsis thaliana)
MeTME3_00015964-RA	55815583	55819764	Protein of unknown function
MeTME3_00015966-RA	55850863	55854246	Similar to PEX13: Peroxisomal membrane protein 13 (Arabidopsis thaliana)
MeTME3_00015962-RA	55791889	55799712	Protein of unknown function
MeTME3_00015968-RA	55864438	55865898	Protein of unknown function
MeTME3_00015965-RA	55820380	55827621	Similar to At4g12770: Auxilin-related protein 2 (Arabidopsis thaliana)
MeTME3_00015963-RA	55806148	55814676	Similar to RLT3: Homeobox-DDT domain protein RLT3 (Arabidopsis thaliana)
MeTME3_00015958-RA	55729844	55736777	Similar to LPXB: Probable lipid-A-disaccharide synthase%2C mitochondrial (Arabidopsis thalian
MeTME3_00015967-RA	55856073	55860578	Similar to gpn3: GPN-loop GTPase 3 (Danio rerio)
MeTME3_00015961-RA	55786832	55787683	Similar to TMN12: Transmembrane 9 superfamily member 12 (Arabidopsis thaliana)
MeTME3_00015961-RB	55787114	55787956	Similar to TMN12: Transmembrane 9 superfamily member 12 (Arabidopsis thaliana)

Chapter 4

A high-throughput reverse genetic platform to study genes from virus resistance locus in cassava

Personal contribution:
I identified the QTL-associated genes and designed the specific target sequences. I cloned the various VIGS construct and infected plants using the modified agro-inoculation protocol. I confirmed the gene silencing and tracked the infection incidence and infection phenotypes. Under my supervision Marius Rohner run the experiments using single-gene VIGS constructs. I wrote the draft manuscript with input from Prof. H. Vanderschuren.

Publication state:
We aim to submit this chapter to *Plant Methods* or *BMC Plant Biology*

A high-throughput reverse genetics platform to study genes from virus resistance locus in cassava

Joel-Elias Kuon[1], Marius Rohner[1], Simon Bull[1], Wilhelm Gruissem[1] & Hervé Vanderschuren[12]

[1] Institute of Molecular Plant Biology, Department of Biology, ETH Zurich, Universitätstrasse 2, 8092 Zurich, Switzerland
[2] AgroBioChem Department, University of Liège, Passage des Déportés 2, Gembloux, Belgium

Correspondence should be addressed to herve.vanderschuren@ulg.ac.be

Abstract

Cassava geminiviruses (CGMs) are DNA viruses that severely affect the production of the food security crop cassava (*Manihot esculenta* Crantz) in Africa and on the Indian subcontinent. The mitigation of CGMs-associated disease, the so-called Cassava Mosaic Disease (CMD), requires the identification and characterization of genetic sources of CMD resistance for their rapid deployment in farmer-, industry- and consumer-preferred cassava varieties. For cassava, only few natural resistance sources have so far been identified but their molecular mechanisms have remained elusive. The identified sources of CMD resistance are; the recessive *CMD1*, the dominant mono-genic *CMD2* and the *CMD3*. Recently, it was found that transgenic *CMD2*-type cassava regenerated via somatic embryogenesis becomes highly susceptible to CMD. Recent advances in cassava genomics and genetics allowed the *CMD2* mapping and the identification of 88 genes annotated within the *CMD2* locus. We implemented a reverse genetic approach in order to identify the *CMD2*-located genes whose alteration of expression impacts symptom score and virus replication. We identified four genes whose reduction of transcript levels by Virus-Induced Gene Silencing (VIGS) alters virus symptom spreading, symptom development and virus incidence in the CMD susceptible model cultivar 60444 and the *CMD2*-type cassava cultivar TME 3. Among the four *CMD2* candidate genes, the Protein Disulfide Isomerase (PDI) appeared as the only VIGS targeted gene enabling virus replication in the cassava cultivar TME 3, whereas silencing of PDI in 60444 led to no visible changes in virus incidence. Stable silencing of *MePDI2-2* by constitutive expression of hairpin dsRNA in the model cultivar 60444 lines led to reduced geminivirus incidence, mild virus symptom development and decreased virus load compared to the control plants. Our results suggest that *MePDI2-2* has a contrasting role in virus replication for CGM resistant and susceptible cassava cultivars. Our pipeline demonstrates the potential of the VIGS platform to rapidly identify host genes whose modulation can alter symptom score and geminivirus replication.

Keywords: VIGS, geminivirus resistance, ACMV, geminivirus, cassava, reverse genetics

Cassava was introduced into Africa by the Portuguese slave traders around 500 years ago [137] and was initially spread into less accessible interior along the major rivers of West and Central Africa. Cassava has been introduced to East Africa from Madagascar late 18th century and was virtually grown throughout much of sub-Saharan Africa by the 20th century [5]. Today, cassava is the staple crop for an estimated 800 million people worldwide (FAO, 2016). In Africa, cassava is mainly grown by smallholders' farmers under poor growing conditions, including unfertilized marginal soils and unpredictable rainfall. Although cassava has a remarkable ability to tolerate unfavorable conditions, on-farm productivity remains very low on the African continent (FAOSTAT, 2016) due to various production constraints.

The Cassava mosaic disease (CMD), caused by eleven species of cassava mosaic geminiviruses (CMGs), belongs to the most economically important cassava diseases in Africa as it severely impacts cassava yield [6], [7], [138]. As a consequence CMD causes more than 25 million tons of yield loss annually and affects the food security of more than half billion people [5]. Geminiviruses belong to the genus begomoviruses and have a single-stranded DNA (ssDNA) bipartite genome [139] and require an insect vector for transmission [140]. Geminiviruses alter the cell cycle of infected hosts to promote the replication of viral as well as plant DNA, modulate host gene expression, inhibit cell death pathways, molecule trafficking and infer with cell signaling to block host defense mechanisms [141][8]. Genetic studies and conventional breeding for natural virus resistance has identified 22 resistance genes (*R*-genes) that have been successfully isolated and identified via map-based cloning strategies. Of these, only two are responsible for dominant resistance to geminiviruses and were identified in tomato (*Solanum lycopersicum*) [24]. The geminivirus resistance genes *Ty-1*/*Ty-3* are allelic and encode for a tomato RNA-dependent RNA polymerase (RDR) [21] and the recessive tomato *Ty-5* was identified to encode for the homolog of the messenger RNA surveillance factor Pelota (*Pelo)*[20]. Moreover, natural resistance sources against geminiviruses were also identified in pepper [142], cotton [143] and bean [144] but their underlying genes and molecular basis are still unknown and need to be characterized.

For cassava, three types of natural resistance have been identified for controlling CMGs, the recessive *CMD1*, transmitted from wild cassava relatives [138] and the single-dominant *CMD2* locus that confers resistance to all known CMGs [31]. The *CMD2* was discovered within landraces collected from farmers' fields in Nigeria and other West African countries during the 1980s and 1990s [33]. Recently, an additional resistance source was described , named *CMD3*, that was genetically mapped to the *CMD2* region but hypothesized to be unlinked to *CMD2* based on a single genetic marker [33]. *CMD2* breeding is facilitated by its dominant nature and therefore has become the major resistance source deployed in African cassava breeding programs. However, *CMD2* has not yet been molecularly identified and cloned and the exact pedigree of the *CMD2* is also unknown. The *CMD2* was mapped using a bi-parental mapping population of 180 F1 individuals to a > 1 Mb region of chromosome 12 in the cassava reference genome [34], [49]. In another study, the *CMD2* locus was mapped to a similar location by a genome-wide association study

(GWAS) that genotyped 6,128 cassava breeding lines [35].

Recently, it was found that the *CMD2* resistance breaks down during embryogenesis, a crucial step for cassava genetic transformation [36], [112]. Plant tissue culture has been particularly prone to somaclonal variation which can result in gene mutation or changes in epigenetic marks [145][146]. Somaclonal variation had been observed previously in cassava [112], [113]. Importantly, the loss of the *CMD2* mediated resistance occurred uniformly in *CMD2*-type transgenic TME 204 [36]. The identification of the *CMD2* genes would enable further investigation of the molecular changes associated with the loss of *CMD2* resistance during embryogenesis. Moreover, identification of CMD resistance genes would provide additional options to genetically engineer geminivirus resistance in cassava. The current generation of engineered geminivirus resistance uses double stranded (ds) RNA-based approaches [16][147]. However, such approaches have limitations when exposed to high virus diversity in the field [148]. In contrast, the *CMD2* based geminivirus resistance shows stable resistance to all known CMGs [31], [149].

Reverse genetics is a powerful tool to identify genes underlying phenotypes. The silencing of candidate genes using RNA interference (RNAi) has largely been used in reverse genetics and involves sequence-specific alteration of gene expression. The gene function(s) can be subsequently analyzed based on the phenotypes that result from the change in gene expression. RNAi functions as a sequence-specific RNA degradation mechanism that is triggered by ds RNA and has been shown to be the primarily mechanism underlying host responses to viral infections [150][151][152]. A virus carrying a host gene fragment triggers silencing of the host gene in a sequence-specific manner by the host machinery. During VIGS infection, dsRNA corresponding to the host gene are produced and subsequently cleaved into small interfering RNA (siRNA) of 21-24 nucleotides (nt). Those siRNAs are incorporated into the RNA-induced-silencing complex (RISC) to degrade the target mRNA [153]. In cassava, the production of stable genetic transformation is challenging to establish, time-consuming and genotype-dependent[112]. In contrast, VIGS has appeared as a method of choice to bypass these limitations and to investigate gene functions using a genotype-independent system, provided the use of CMV-susceptible genotypes.

We recently established an *African cassava mosaic virus* isolate ACMV-[NOg] DNA-A based VIGS vector for agroinoculation of cassava plantlets [16] (Lentz et al. under review). In the present work, we established an optimized VIGS system for a rapid inoculation and multiple gene silencing to screen 88 candidate genes located in the *CMD2* locus.

Identification of the scaffolds associated with CMD2

In cassava, conventional breeding and quantitative trait (QTL) mapping is challenging due to the complex breeding cycle with the non-synchronous flowering and the common incompatibility between genotypes [67]. Several genetics studies have achieved narrowing down the *CMD2*-bearing chromosome region [31], [34], [35], however, a *CMD2* co-segregating marker has not yet been generated. To reconstruct the *CMD2* region, a BLASTn [109] (Word length= 11, Expect threshold= -1) search with the set of genetic markers highly associated with *CMD2* were conducted using the public available draft whole genome assembly (v4.1) of a South-American cassava cultivar AM560 available on phytozome [39]. Table 1 shows the list of genetic markers used and their alignment position on the draft genome. Two sequencing scaffolds (scaffold5214 and scaffold6906), genetically anchored between 23.57-52.61 cM of chromosome 12 [78], were received. The two scaffolds span a physical genomic region of 1.846 megabase pairs (Mb) (scaffold6906 spans 382,376 bp and scaffold5214 spans 1,463,792 bp) and harbor 88 gene loci in total (Supplementary Table 1). Remarkably, it was reported that thirty-five out of the 88 genes were responding to the CMV infection in a transcriptome study of the virus susceptible cassava cultivar T 200 inoculated with the *South African Cassava Mosaic Virus* [37]. The 88 genes and their functional annotations were compared to databases of known dominant virus resistance genes [29] but no candidate gene could be identified. Provided that all cassava genes located in the *CMD2* genomic region have been reported in the reference genome, it suggests that a novel and totally unknown resistance mechanism underlays the *CMD2*.

Development of VIGS clones

The design and inoculation of VIGS clones followed the pipeline depicted in Figure 1. First, the QTL region was identified and the annotated coding DNA sequence (CDS) of the 88 candidate genes were received from the public phytozome data base (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Mesculenta). Then target sequences were gradually optimized using BLASTn (Word length = 11, Expect threshold=-1) against the complete set of cassava CDSs to identify gene sequence between 250-500 bp long that are specific to limit potential off-targeting. To minimize possible interference of secondary DNA structures (i.e. palindromes or inverted repeats) on the viral replication life-cycle, the selected sequences were tested for secondary DNA motifs using public available online tools (http://emboss.bioinformatics.nl/cgi-bin/emboss/). In order to increase the specificity, target sequences and potential off-targets were validated using short-read based genome assemblies from genotypes that were used for the virus inoculation. To assess the *CMD2* located genes, the cassava genotype TME 3, generally known as one of the original sources of the *CMD2* [35], and the cassava genotype 60444, the model and virus susceptible cultivar, were used. The provisional 250-500 bp target sequences were aligned to the assemblies of TME 3 and 60444 using BLASTn. The genome assemblies as well as the BLASTn searches were conducted using the CLC genomic workbench under default assembly

and BLASTn parameters. Alignments of target gene sequences were manually screened for the best 100 nt sequences harboring >21 nt stretches perfectly matching (100% homology) the sequences between 60444 and TME 3. It has been shown previously that 100 bp of the target sequence can generate high level of gene silencing [154], [155]. The lower limit was set at 21 nt as they represent the shortest siRNA produced by the antiviral RNAi plant immune system [153].

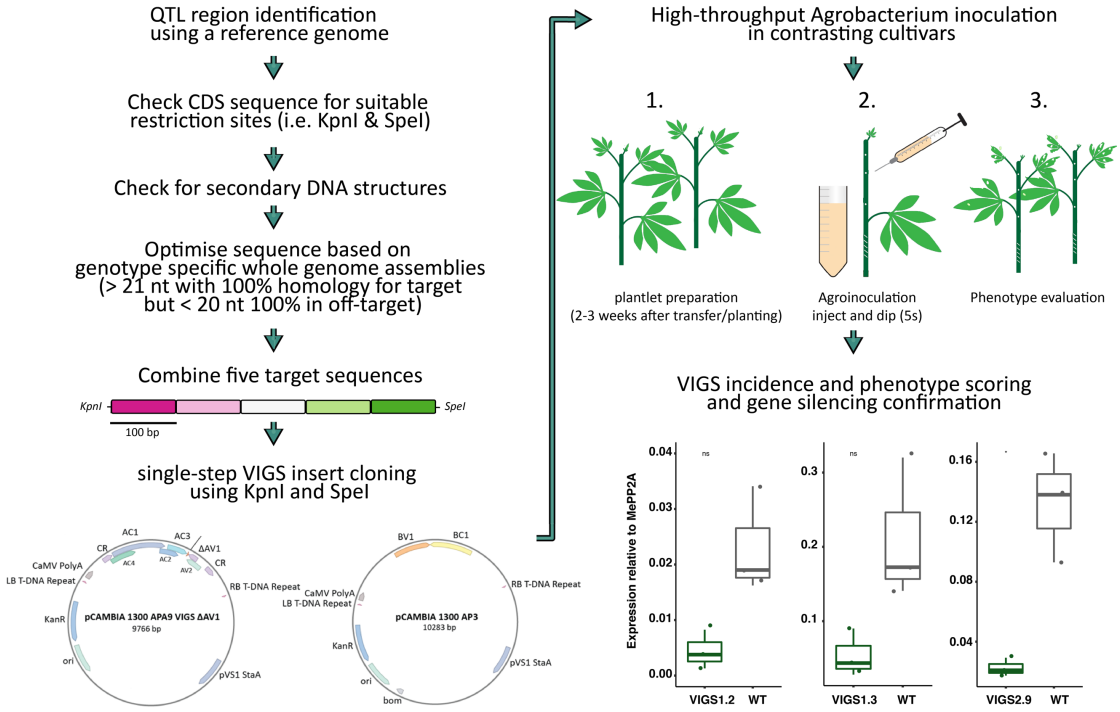


Figure 1 Overview scheme of the optimized VIGS experiment for high-throughput gene analysis

Table 1 Summary of the known markers associated with *CMD2* resistance in cassava and their chromosome number

Marker	Primer sequence	Chr (genome v6.1)	scaffold (genome v4.1)	Study
S8_7762525	GBS-SNP*	12	scaffold06906	Wolfe et al. 2017
S5214_780931	GBS-SNP*	12	scaffold05214	Rabbi et al. 2014
SSR_NS158	Fw:GTGCGAAATGGAAATCAATG	12	scaffold06906	Okogbenin et al. 2007
	Rev:TGAAATAGTGATACATGCAAAAGGA	12	scaffold06906	
SSR_NS169	Fw:GTGCGAAATGGAAATCAATG	12	scaffold06906	Okogbenin et al. 2007
	Rev:GCCTTCTCAGCATATGGAGC	12	scaffold06906	
SSRY28	Fw:TTGACATGAGTGATATTTTCTTGAG	12	scaffold05214	Akano et al. 2002
	Rev:GCTGCGTGCAAAACTAAAAT	12	scaffold05214	

*sequence location can be extracted from the corresponing publication and genetic map

The chimeric VIGS inserts have been chemically synthesized in blocks of five, on the genome physically co-localized genes (5x100bp). In total, 19 VIGS inserts were designed to target the 88 genes. In addition, VIGS constructs were designed that targeted a single gene (VIGS 2.9) as well as four genes (VIGS 2.4). To allow the comparison of infection rates between different VIGS constructs, a control VIGS insert harboring 500 bp sequence fragments from the green-fluorescent protein (GFP) gene (250bp) and the β-glucuronidase (GUS) gene (250bp) was generated. The target sequences were inserted into the multiple-cloning sites (MCS) of the VIGS vector and the sequence confirmed using PCR and Sanger-sequencing. Positive clones were subsequently introduced into the hyper-virulent *Agrobacterium tumefaciens* strain AGL 1, that allows high virus infection rates (Lentz et al. under review). Target sequences as well as target genes are provided in Supplementary Tables 2 and 3.

Agrobacterium-based VIGS inoculation

The use of the high-throughput VIGS vector inoculation platform allowed a rapid and cost-effective assessment of the VIGS clones in the two cassava genotypes 60444 and TME 3. The robustness of the inoculation method was manifested by the high infection rates over the two assays. For example, in 60444 we were able to visually detect VIGS mediated symptoms for 91% of the plants at six weeks post inoculation (wpi) (Figure 2a). In contrast, the resistant TME 3 plants rarely displayed virus symptoms over the two assays (6 % infection rate at 6 wpi). A VIGS vector for visual validation of gene silencing was also developed. It included a partial sequence from the Mg^{2+}-chelatase gene (Manes.17G053100), that encodes the first enzyme of the chlorophyll biosynthesis pathway (*Chl1*, Mg2+-chelatase subunit I). Silencing of the Mg^{2+}-chelatase transcript leads to a chlorotic phenotype in tissues with reduced levels of Mg^{2+}-chelatase enzyme [156]. As expected, *Chl1* silencing was more prominent in CMV-susceptible 60444 plants and only constricted chlorotic symptoms could be observed in VIGS-*Chl1*-inoculated TME 3 plants (Figure 2b, top panel) (Supplementary Figure 1b).

Several VIGS constructs displayed contrasting infection rate and CMD symptom development. For example, the vector VIGS 2.3 did not produce virus-like symptoms in both 60444 and TME 3 plantlets. In contrast, VIGS 2.8 generated the highest infection rate for both genotypes and even infected a total of 10 out of 15 TME 3 plants (67% infection rate). We found also VIGS constructs that had low infection rates in both cassava genotypes. For example, VIGS 1.9 infected only very few 60444 plants and no TME 3 plantlets. A common feature of *CMD2*-type cultivars is a virus symptom recovery phenotype, by which the severity of CMD symptoms displayed on new growth reduces over time, until newly formed leaf tissues are free of visible viral symptoms [38]. This recovery phenotype was observed for the few symptomatic TME 3 plants but not for 60444 plantlets that remained infected.

Silencing of an UNCHARACTERIZED RING ZINC FINGER-CONTAINING PROTEIN causes Hypersensitive – Response (HR) -like black necrotic tissues

We found that the vector VIGS 2.3, which targets the genes cassava4.1_026906m, cassava4.1_012052m, cassava4.1_016758m, cassava4.1_007335m and cassava4.1_002986m, triggered the appearance of black areas on young, emerging leaves of infected 60444 plants (Figure 2b, middle panel). Noticeably, VIGS 2.3 inoculated plants remained free from visible virus symptoms in 60444 as well as in TME 3 plants. In order to determine whether the development of black areas on young leaves was due to the silencing of single or multiple gene sequences present in the VIGS 2.3 vector, we generated new VIGS clones that targeted each of the five genes separately. Using single target VIGS vector the gene cassava4.1_026906m was identified as the unique sequence whose silencing by VIGS causes the development of necrotic tissue. By approximately two-month post infection, the black area/necrotic symptoms in newly emerging leaves were attenuated and inoculated plants eventually became free from necrotic symptoms (Figure 2b, middle panel, Supplementary Figure 1). The cassava4.1_026906m gene encodes for 199 amino acids (aa) protein with no functional information provided by the reference genome annotation. However, a computational characterization for conserved domains (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) revealed an intact RING finger motif with a structural zinc finger (pfam13920, C3HC4- type). Proteins with a similar motif pattern can bind DNA, RNA, proteins or lipids [157] and are involved in numerous biological processes including photomorphogenesis [158], light signaling [159], secretory pathways [160], peroxisome biogenesis [161], stress tolerance and even disease resistance [162]. Moreover, it is assumed that such proteins play an important role in the ubiquitination, protein location and protein degradation pathways [163].

Silencing of two peroxidases leads to severe virus symptoms and growth reduction

The VIGS 2.9 vector targeted the cassava genes cassava4.1_011768m, cassava4.1_029175m, cassava4.1_012316m, cassava4.1_012330m, cassava4.1_022227m and caused severe virus symptoms in the inoculated plants that led to reduced and stunted growth (Figure 2b, bottom panel). Usually such a phenotype with strong leaf curling and exceptionally slow growth indicates high virus infection pressure or a hyper-susceptible host that provides the perfect environment for viral replication and pathogen spreading. We subsequently generated VIGS vectors with unique target sequence in order to identify the gene(s) whose deregulation provokes severe CMD symptoms. The downregulation of two peroxidases (cassava4.1_029175m and cassava4.1_011768m) appeared to cause the development of severe symptoms. The two peroxidases cluster together in a narrow 17 kilobases (kb) genomic region suggesting that a functional gene cluster may exist at that locus. A pairwise protein sequence comparison using MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/) revealed large sequence differences between the two peroxidases, although there were only very little differences detected in protein length (328 aa for cassava4.1_029175m and 326 aa for cassava4.1_011768m). Both proteins were annotated with the same functional KEGG orthology annotation (KEGGORTH K00430). This KEGG entry is linked to the

phenylpropanoid pathway, to metabolic pathways in general as well as the biosynthesis of secondary metabolites (http://www.genome.jp/dbget-bin/www_bget?ko:K00430). Plant peroxidases can be found in plants as well as in fungi, bacteria an yeasts and are involved in numerous cellular processes such as development as well as stress responses [164]. While their role in host defense against viral infection remains unknown, it has been reported for tomato that peroxidases activity is induced in juvenile leaves under whitefly mediated geminivirus infection [165].

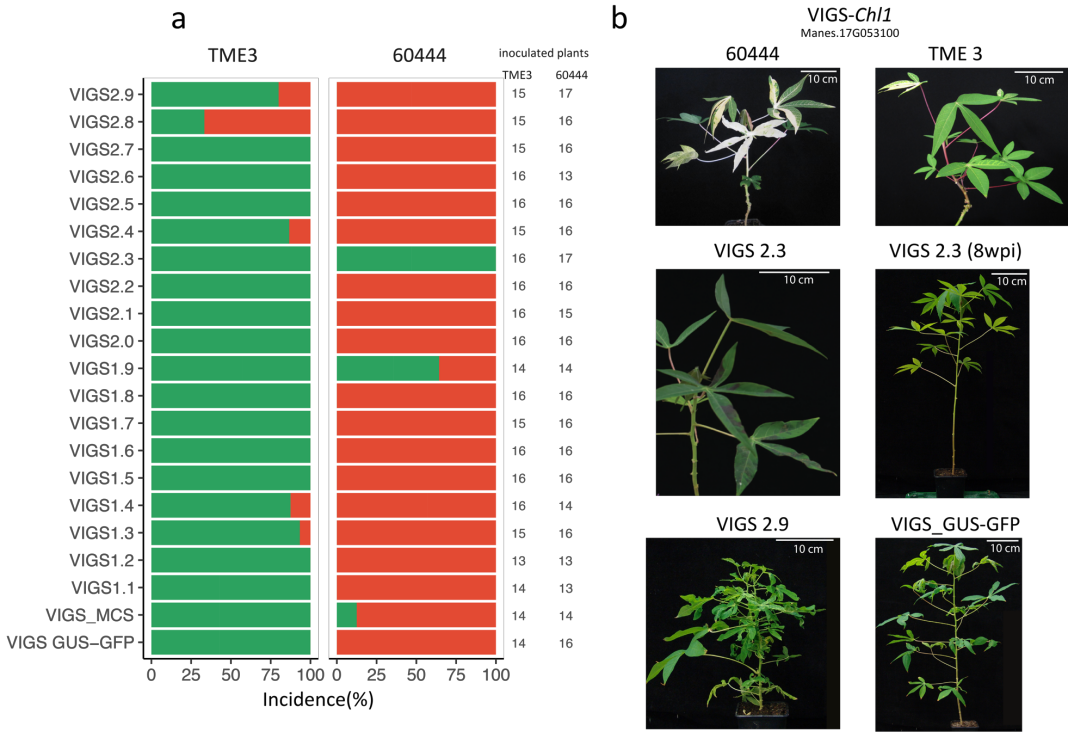


Figure 2 VIGS characterization of the *CMD2* associated genes. a) Virus incidence scores, defined by Vanderschuren et at. 2007, six weeks post infection (wpi). An average incidence score from two independent assays is shown as percentage of symptomatic plantlets. The red bar indicates infected plants and green bars represent the proportion of plants with no visible CMD symptoms. Numbers behind the bars indicate number of plants infected for TME 3 and 60444, respectively b) Representative images of the VIGS vector inoculated cassava plants showing silencing of the *Chl1* gene encoding the Mg^{2+}-chelatase enzyme. Below, representative images for constructs with phenotypically abnormalities were shown for cultivar 60444.

Silencing of a *Protein Disulfide Isomerase like 2.2* gene allows virus replication in virus resistant TME 3

VIGS 2.8 targeted the *MePDI-2.2* (cassava4.1_07986m) gene that encodes a Protein Disulfide Isomerase (PDI). VIGS 2.8 vector was the only construct that displayed a high infection rate in TME 3 (Figure 2a). It was previously shown for a barley ortholog (*HvPDI5-1*) that a loss of function via a single-nucleotide polymorphism (SNP) led to complete resistance to the RNA virus bymoviruses [129]. Protein disulfide isomerases (PDIs) catalyze the correct folding of proteins and prevent the aggregation of unfolded or partially folded precursors [131]. Other studies have shown that the suppression of members of the PDI gene family can delay virus replication of several human and animal viruses (e.g., HIV) [130], [131], [166]. This

data as well as the previous reports indicate that PDIs are involved in the host-virus interaction, however, their functional interactions with viruses remain largely unknown.

Polymorphism detection and expression analysis of the MePDI2.2

In order to characterize allelic variation, Sanger-Amplicon cDNA sequencing as well as genome wide sequencing data (Kuon et al. in preparation) were analyzed to test the four candidate genes for *CMD2*-specific mutations. Homologous genes were identified using BLASTn search of the genomic region extracted from the cassava reference genome against a whole genome assembly data set of 60444 and TME 3. Sequences were compared using MUSCLE [106] and the pairwise sequence alignment tool provided by the CLC-Genomics workbench. Polymorphisms could be detected in the *CMD2* associated genes from TME 3, 60444 and the reference genome AM560-2. However, most mutations were shared between TME 3 and 60444. In case of the two peroxidases, no amino-acid changing (non-synonymous) polymorphisms were detected in the coding sequence between 60444 and TME 3. The occurrence of no *CMD2* specific mutations for cassava4.1_011768m was also reported earlier where the authors analyzed the coding sequences of the peroxidases using short-read alignments [35]. Furthermore, no non-synonymous mutation was detected for cassava4.1_026906m that caused the HR-like phenotype in virus susceptible between 60444 and TME3 as well. In contrast, the sequence analysis for the *MePDI2-2* revealed *CMD2*-specific, non-synonymous as well as synonymous mutations in TME 3 that were not found in 60444. In TME3, four heterozygous point mutations were found of which three mutations caused a non-synonymous amino acid change (Figure 3a). The *MePDI2.2* protein sequence is organized in three thioredoxin-like conserved domains (TRX) with two of them being affected by the location of the non-synonymous single-nucleotide polymorphisms (SNPs).

Because *CMD2*-type resistance breakdown occurs in cassava plantlets regenerated through somatic embryogenesis [36], we investigated whether *MePDI2.2* expression is altered by this tissue culture process. For this purpose, we used green-house grown, independent transgenic 60444 and virus susceptible *CMD2*-type TME 7 lines together with the wildtype (wt) controls to measure the relative expression of the *MePDI2.2* using quantitative real-time (qRT) PCR. Due to highly variable gene expression, no statistically significant differences were measured between virus-susceptible and wildtype plantlets (Figure 3b) suggesting that the expression of *MePDI2.2* in leaves is not affected through somatic embryogenesis.

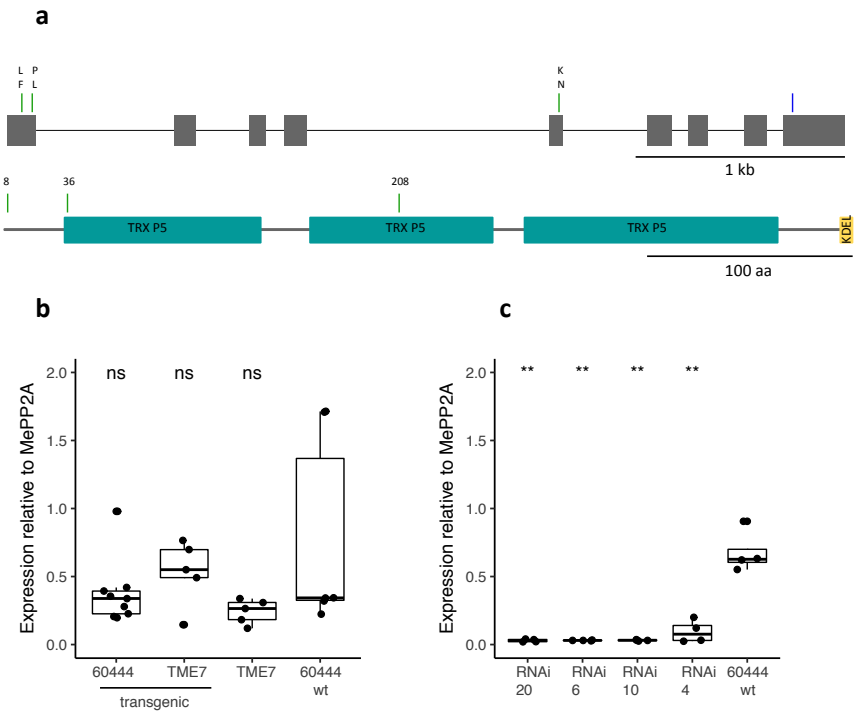


Figure 3 Genomic structure of *MePDI2.2* gene and expression analysis in wild-type and transgenic cassava. a). Genomic structure of *MePDI2.2* and the nine exons are shown as grey boxes. TME 3-specific non-synonymous SNPs are indicated with green lines and their amino acid alteration is shown in letters above, whereas synonymous SNPs are shown as blue line. Below, *MePDI2.2* coding sequence (1,304kb) and the annotated conserved domains. The green box indicates the Thioredoxin-domain. b) *MePDI2.2* expression data of transgenic and wt plants using qRT-PCR c) qRT-PCR assay for testing the expression of 60444 *MePDI2.2* in RNAi lines compared to wt 60444. The qRT-PCR data were normalized to the endogenous expression of *MePP2A*. S.D., biological replicates (n=3), technical replicates (n=2) and statistical variation was assessed using Tukey's multiple comparisons test ($p < 0.01$ = **).

Downregulation of MePDI-2.2 reduces CMD incidence and symptom development

Because VIGS 2.8 led to high virus incidence in 60444 as well as TME 3 cassava plantlets, virus replication and virus incidence was tested in stably silenced *MePDI2-2* transgenic 60444, the cassava cultivar amenable for genetic transformation. Because *CMD2*-type resistant cultivars lose their CMD resistance during somatic embryogenesis, it was not possible to validate the higher susceptibility of TME 3 plants whose *MePDI2.2* expression is altered. Silencing of the *MePDI2.2* was achieved in cassava 60444 through the expression of a 35S promoter-driven hairpin RNA cassette (*MePDI-2.2*-RNAi). One hundred cotyledons were successfully regenerated into plantlets and more than 75% of those rooted on selection media. Transgenic 60444 plantlets were subsequently selected for molecular characterization using PCR amplification of the hygromycin selectable marker. A set of 20 PCR-positive plants were characterized with Southern blot and four independent transgenic lines were selected for *MePDI2.2* silencing confirmation (Supplementary Figure 2b). Significant transcript reduction was detected in plants expressing *MePDI-2.2*-RNAi using qRT-PCR measurement (Figure 3c). After confirming the silencing phenotype, four independent transgenic lines along with the wildtype (wt) and transgenic (pCAMBIA) control lines were tested for virus resistance against infectious virus clone of ACMV-NOg under greenhouse conditions.

A minimum of fifteen agroinoculated plants per line were screened for disease incidence and disease severity over a period of four weeks. The infection rates in control lines ranged from 80% to 100% indicating that the inoculation procedure leads to high infection rates in wild-type plants and pCAMBIA transgenic plants. The transgenic lines displayed on average a lower symptom score as compared to the control lines, however the large variation in symptom scores between independent plants made those differences not significant (Figure 4a). Investigation of the viral load in transgenic lines revealed a significant lower viral load in the two transgenic lines, RNAi-20 and RNAi-4, as compared to the control lines (Figure 4c).

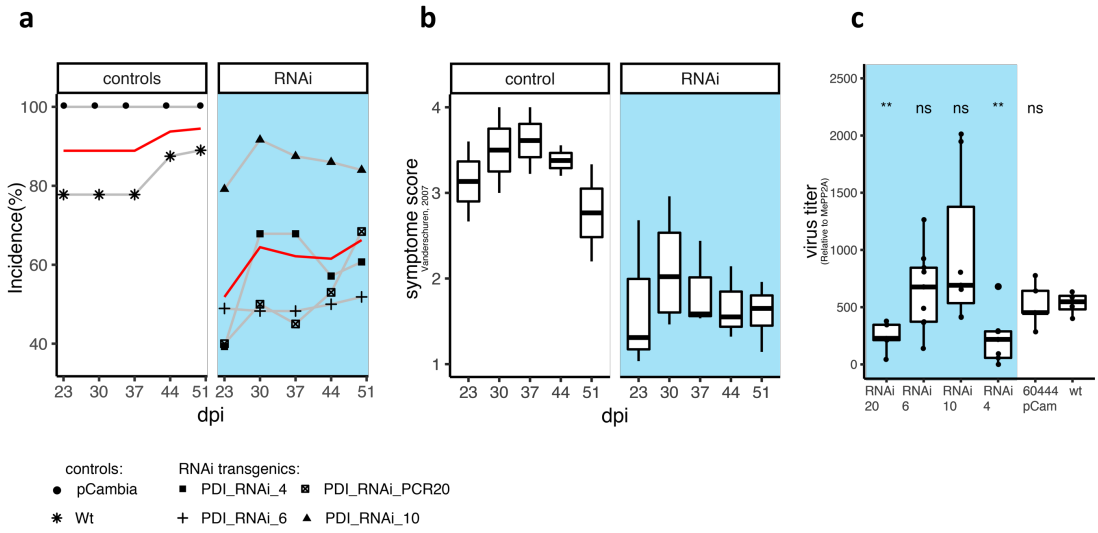


Figure 4 Results of glasshouse ACMV agro-inoculation of *MePDI2.2*-RNAi transgenics and controls. a) Disease incidence as percentage of symptomatic plants over the period of days post infection (dpi). Average Incidence scores are shown as red line b) Average symptom score of all transgenic RNAi-plants over period of infection monitoring. c) Virus titer quantification by qPCR on samples harvested at 51 dpi. The data was normalized to the endogenous *MePP2A* DNA. Statistical variation (biol.replicates n=10, technical replicates=3) was assessed using Tukey's multiple comparisons test ($p < 0.01$ = **).

Conclusion

This study used a modified, high-throughput VIGS system that enabled characterization of QTL associated genes in cassava. The rapid and easy-to-handle agro-inoculation method allowed screening dozens of *CMD2* associated genes in a single screening experiment. Moreover, the modularity of the system allowed screening all *CMD2* associated genes in a single infection assay and the direct comparison of silencing phenotypes.

Because high-quality genome sequences and gene space annotation of a *CMD2*-type cultivar for the *CMD2* genome region was not available, a reference genome bias can not be excluded. The candidate gene selection and target sequence development strongly relied on the gene space annotation of the *CMD2* locus in the genome of AM560-2 cultivar. However, the presence of the gene(s) and/or sequence features associated with *CMD2*-type resistance remains hypothetical in the CMD susceptible AM560-2 cultivar. Moreover, the two *CMD2* linked scaffolds consist of numerous unassembled regions and sequencing gaps [128]. A high-

quality *de novo* genome assembly and *de novo* gene space annotation from a *CMD2*-type cassava genotype would help generating a complete set of candidate genes and features present in the *CMD2* locus. Genetic studies proved that the *CMD2* is transmitted as dominant, heterozygous trait[35]. Therefore, a diploid aware genome assembly would also uncover the allelic variation in the *CMD2* located genes and features and help designing allele-aware VIGS experiments.

Furthermore, stable transgenic cassava lines overexpressing and downregulating candidate genes in susceptible as well as *CMD2*-type genotypes are required for validation. Stable downregulation of the *MePDI2-2* in transgenic cassava lines appeared to be associated with reduced symptom score and viral load in selected transgenic lines, suggesting that *MePDI2.2* plays a role in geminivirus infection. The *MePDI2.2* protein contains an endoplasmic reticulum (ER) KDEL retention motif. Interestingly, the ER seems to be the crucial organelle that supports viral entry, translation, replication and assembly [167]. PDIs is a multifunctional redox chaperone of the ER and studies in mammalian systems have revealed that these proteins are prone to redox-dependent post translational modifications under specific disease states [168][169]. Interestingly, an accumulation of reactive oxygen species (ROS) and nitric oxide (NO) was found in geminivirus infected pepper and Java jute [142][170] that potentially could trigger such a confirmation change of the PDIs. Further investigations are clearly needed in this area, to unravel the role of *MePDI2-2* under geminivirus pathogenesis.

Identifying *CMD2* and understanding why it breaks down is of major importance since several large-scale breeding projects and cassava transformation projects rely on stable and robust geminivirus tolerance mediated through the *CMD2*. The ultimate goal should be to transfer the gene(s) to virus susceptible, high value cassava cultivars. Moreover, the *CMD2* could be pyramided with other resistance sources such as the *CMD1* and *CMD3* based resistance, or even combined with genetic engineered strategies (i.e. RNAi constructs that target the virus genome) to generate a robust and durable virus resistance in the field.

Acknowledgement:

We thank Irene Zurkirchen for taking care of the plants in the greenhouse. We thank Dr. Ezequiel Matias Lentz, Dr. Adrian Alder for helpful support during the VIGS experiments. We want to express a special thanks to Dr. Ravi Bodampalli for providing support during the cassava transformation and molecular characterization experiments. This work was supported by grants from the Swiss national science foundation and the Bill & Melinda Gates Foundation.

Author Contributions:

J-E. K., M.R and S.E.B performed the experiments. J-E.K, W.G. and H.V. analyzed the data and J-E. K. and H.V. wrote the manuscript.

VIGS plasmid construction and target design

The cassava reference genome assembly (v.4.1) was deployed for designing target sequences for specific gene silencing. Each *CMD2* linked gene was inspected for a specific 100 bp target site using BLASTN. Target sequences were tested for their specificity in 60444 and TME 3. Whole genome shot-gun sequencing reads were trimmed and *de novo* assembled in CLC Genomics Workbench Version 6.5 (CLC Bio, www.clcbio.com) using default parameters. VIGS targets were aligned against the assemblies using BLASTN and sequence selected when > 21nt aligned between TME 3 and 60444.

Five genes were combined to a single target block when they were neighboring or in close proximity based on the genome assembly gene order. To avoid any secondary structures that may take an influence on the VIGS performance, we screened each target sequence using EINVERTED, PALINDROME and EQUICKTANDEM from the emboss bioinformatics tool box. (http://emboss.bioinformatics.nl/cgi-bin/emboss). The 500 bp DNA fragments were chemically synthesized (Thermo Fisher) and ligated to the pJet1.2 (Lifetechnology) intermediate vector and fully sequenced. Fragments were inserted into the VIGS vector using *KpnI* and *SpeI* restriction enzymes. The final constructs were used to electroporate *Agrobacterium tumefaciens* strain AGL1. Electroporation was confirmed by performing PCR using VIGS vector specific primers. The primers and target sequences are listed in Supplementary Table. 6.

Virus inoculation and symptom scoring

For the VIGS and ACMV inoculation experiments, four weeks old cassava plants were used for agroinoculation. *Agrobacterium tumefaciens* stain AGL1 containing the different VIGS constructs were cultured for 48 h at 28 °C in 5 ml YEB (5 g/L tryptone, 1 g/L yeast extract, 5 g/L nutrient broth, 5 g/L sucrose, 2mM MgSO4) containing 100 mg/L carbenicillin, 20 mg/L rifampicin and 50 mg/L kanamycin. Two ml of the starter culture were then added to 200 ml YEB with the same antibiotic composition and grown at 28 °C till an OD_{600nm} of 1.5-2 was reached. Cells were pelleted by centrifuging 10 min with 5,000x and washed twice with using sterile deionized water. Then the washed bacterial pellet was re-suspended in 10 ml inoculation medium (10 mM MES pH 5.6, 10 mM $MgCl_2$, 0.25 mM acetosyringone) and inoculated for two hours under constant shaking. Then the suspension was adjusted to OD_{600nm} 2 using the inoculation medium. Equal volumes of the suspension of Agrobacterium carrying the VIGS – vector / DNA-A and the suspension of Agrobacterium carrying DNA-B vector was prepared prior to inoculation.

For inoculation, all leaves were removed and stem and auxiliary bud pricked using a syringe (0.33 mm / 29 G / 12.7 mm). Then the plant was dipped into the Agrobacterium solution for 5 seconds and covered in a Plexiglas box for one week. Hereafter, plants were grown under greenhouse conditions (28 °C, 16-h day length, 22 klx, 50% humidity). Virus symptoms or VIGS phenotypes in the top five leaves were scored from four to eight weeks post infection (PI) with a scale of 0 – 4 as described in Vanderschuren et al. [147]. For virus titer quantification, the top five fully grown leaves were sampled from each plant at eight weeks PI and DNA was extracted using a modified protocol [94].

Virus quantification

Virus quantitation was performed by relative quantification qPCR on 20 ng of total DNA extracts derived from the top five leaves using ACMV DNA A specific primers and MePP2A genomic DNA reference primers as listed in Supplementary Table 6. Symptomatic leaves of ten plants per line were harvested and pooled into five DNA samples. Two technical replicates were used per pooled sample.

Plasmid construction and cassava transformation

The expression binary vector pRNAi-MePDI2-2 was constructed based on an RNAi plasmid described earlier [16][171]. Primer were designed based on the coding sequence of the *MePDI-2.2* from phytozome (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Mesculenta). The African Mosaic Virus-NOg AC1 sequence was replaced with the *MePDI-2.2* sequence from position 899 to 1112 in the reverse and the forward orientations. The *MePDI-2.2-RNAi* expression was controlled by the Cauliflower Mosaic Virus (*CaMV*) 35S promoter and terminator sequence. The vector was transformed into chemically component *Escherichia Coli* (TOP10 competent cells, Invitrogen) and grown on LB agar plates containing 50 mg/L kanamycin antibiotics. Final plasmids were evaluated using PCR and amplicon Sanger sequencing. The resulting construct was mobilized into *Agrobacterium tumefaciens* strain LBA4404 for transformation of cassava 60444 following Bull et al. [172].

Molecular characterization of transgenic cassava lines

Cassava genomic DNA was extracted from liquid-nitrogen frozen leaf tissue according to a modified protocol [94]. DNA integrity and quantity was determined by using Nanodrop (ThermoFisher Scientific,Waltham, MA, United States). 10 µg DNA was digested using *HindIII* (New England Biolabs, Ipswich, MA, United States) for 16 h and subsequently ethanol precipitated and re-suspended in 20 µl sterile, nuclease-free water. Sample was loaded on a 1 % TAE agarose gel including a DIG-labelled marker (Roche, Basel, Switzerland). DNA was transferred to nylon membrane using Southern blotting and hybridized with a DIG-labelled probe targeting the *hptII* gene. T-DNA integration events were assessed using exposure to autoradiograph film.

Supplementary Table 1 Cassava genes harboring the two *CMD2* scaffolds scaffold05214 and scaffold06906

Scaffold	cassava transcript	Tair orthologs	funtional annotation
6906	cassava4.1_029206m	AT1G43760	DNAse I-like superfamily protein
6906	cassava4.1_011768m	AT1G05260	Peroxidase superfamily protein
6906	cassava4.1_029175m	AT3G01190	Peroxidase superfamily protein
6906	cassava4.1_012316m	AT1G52930	Ribosomal RNA processing Brix domain protein
6906	cassava4.1_012330m	AT1G52930	Ribosomal RNA processing Brix domain protein
6906	cassava4.1_022227m	AT3G21720	isocitrate lyase
6906	cassava4.1_012418m	AT3G03800	syntaxin of plants 131
6906	cassava4.1_025392m	AT2G46150	Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family
6906	cassava4.1_025765m	AT3G21710	
6906	cassava4.1_002166m	AT1G05230	homeodomain GLABROUS 2
6906	cassava4.1_017777m	AT2G32380	Transmembrane protein 97, predicted
6906	cassava4.1_026431m	AT3G21700	Ras-related small GTP-binding family protein
6906	cassava4.1_031195m	AT1G05200	glutamate receptor 3.4
6906	cassava4.1_001288m	AT2G32400	glutamate receptor 5
6906	cassava4.1_015589m	AT1G05190	Ribosomal protein L6 family
6906	cassava4.1_030515m	AT3G21680	
6906	cassava4.1_028951m	AT3G51880	high mobility group B1
6906	cassava4.1_031311m	AT1G33030	O-methyltransferase family protein
6906	cassava4.1_000903m	AT2G05710	aconitase 3
5214	cassava4.1_026844m	AT1G05180.1	NAD(P)-binding Rossmann-fold superfamily protein
5214	cassava4.1_015989m	AT2G32415.1	Polynucleotidyl transferase, ribonuclease H fold protein with HRDC domain
5214	cassava4.1_028772m	AT2G32415.1	Polynucleotidyl transferase, ribonuclease H fold protein with HRDC domain
5214	cassava4.1_025566m	AT2G32415.1	Polynucleotidyl transferase, ribonuclease H fold protein with HRDC domain
5214	cassava4.1_008793m	AT1G05170.2	Galactosyltransferase family protein
5214	cassava4.1_033288m	AT1G26320.1	Zinc-binding dehydrogenase family protein
5214	cassava4.1_031223m	AT2G32480.1	ARABIDOPSIS SERIN PROTEASE
5214	cassava4.1_030300m	AT2G27820.1	prephenate dehydratase 1
5214	cassava4.1_021025m	AT1G11790.1	arogenate dehydratase 1
5214	cassava4.1_019925m	AT4G04610.1	APS reductase 1
5214	cassava4.1_005161m	AT1G76160.1	SKU5 similar 5
5214	cassava4.1_002113m / cassava4.1_002645m /	AT1G05120.1	Helicase protein with RING/U-box domain
5214	cassava4.1_015726m	AT5G42570.1	B-cell receptor-associated 31-like
5214	cassava4.1_002340m	AT4G24970.1	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase family protein
5214	cassava4.1_021361m	AT2G32500.1	Stress responsive alpha-beta barrel domain protein
5214	cassava4.1_008192m	AT5G55090.1	mitogen-activated protein kinase kinase kinase 15
5214	cassava4.1_033355m	AT3G20800.1	Cell differentiation, Rcd1-like protein
5214	cassava4.1_004950m	AT4G20720.1	dentin sialophosphoprotein-related
5214	cassava4.1_017288m (primary), cassava4.1_017290m	AT2G32580.1	Protein of unknown function (DUF1068)
5214	cassava4.1_006101m	AT4G15440.1	hydroperoxide lyase 1
5214	cassava4.1_002188m	AT3G21620.1	ERD (early-responsive to dehydration stress) family protein
5214	cassava4.1_023563m	AT3G21630.1	chitin elicitor receptor kinase 1
5214	cassava4.1_021633m	AT3G21630.1	chitin elicitor receptor kinase 1
5214	cassava4.1_027340m		
5214	cassava4.1_004296m	AT2G32640.1	Lycopene beta/epsilon cyclase protein
5214	cassava4.1_025142m	AT4G15415.1	Protein phosphatase 2A regulatory B subunit family protein
5214	cassava4.1_008294m	AT3G21650.1	Protein phosphatase 2A regulatory B subunit family protein
5214	cassava4.1_017968m cassava4.1_018804m	AT4G22140.1	PHD finger family protein / bromo-adjacent homology (BAH) domain-containing protein

5214	cassava4.1_015952m cassava4.1_015965m cassava4.14.1_015965m	AT4G22140.2	PHD finger family protein / bromo-adjacent homology (BAH) domain-containing protein
5214	cassava4.1_020323m cassava4.1_020350m	AT5G55140.1	ribosomal protein L30 family protein
5214	cassava4.1_008304m	AT2G32990.1	glycosyl hydrolase 9B8
5214	cassava4.1_030256m	AT3G06350.1	dehydroquinate dehydratase, putative / shikimate dehydrogenase, putative
5214	cassava4.1_020803m	AT5G55125.2	Ribosomal protein L31
5214	cassava4.1_025873m	AT2G05755.1	Nodulin MtN21 /EamA-like transporter family protein
5214	cassava4.1_005568m	AT2G05760.1	Xanthine/uracil permease family protein
5214	cassava4.1_012736m	AT2G32980.1	
5214	cassava4.1_002922m	AT2G32970.2	
5214	cassava4.1_006972m	AT2G05790.1	O-Glycosyl hydrolases family 17 protein
5214	cassava4.1_004363m	AT2G05810.2	ARM repeat superfamily protein
5214	cassava4.1_029493m	AT4G12140.1	RING/U-box superfamily protein
5214	cassava4.1_026906m	AT4G03965.1	RING/U-box superfamily protein
5214	cassava4.1_012052m	AT1G05010.1	ethylene-forming enzyme
5214	cassava4.1_016758m	AT1G05000.1	Phosphotyrosine protein phosphatases superfamily protein
5214	cassava4.1_007335m	AT1G04990.1	Zinc finger C-x8-C-x5-C-x3-H type family protein
5214	cassava4.1_002986m	AT2G32950.1	Transducin/WD40 repeat-like superfamily protein
5214	cassava4.1_024759m	AT5G14180.1	Myzus persicae-induced lipase 1
5214	cassava4.1_023985m	AT5G14180.1	Myzus persicae-induced lipase 1
5214	cassava4.1_028362m		
5214	cassava4.1_029968m	AT5G14180.1	Myzus persicae-induced lipase 1
5214	cassava4.1_007986m	AT1G04980.1	PDI-like 2-2
5214	cassava4.1_001190m	AT4G12640.1	RNA recognition motif (RRM)-containing protein
5214	cassava4.1_024415m	AT1G55730.1	cation exchanger 5
5214	cassava4.1_005518m	AT5G48300.1	ADP glucose pyrophosphorylase 1
5214	cassava4.1_002555m	AT5G59810.1	Subtilase family protein
5214	cassava4.1_018748m	AT3G07600.1	Heavy metal transport/detoxification superfamily protein
5214	cassava4.1_033257m	AT3G07600.1	Heavy metal transport/detoxification superfamily protein
5214	cassava4.1_025682m	AT5G03795.1	Exostosin family protein
5214	cassava4.1_004304m	AT4G12680.1	
5214	cassava4.1_014939m	AT3G07640.1	
5214	cassava4.1_029560m	AT2G04235.1	
5214	cassava4.1_027980m	AT3G07650.1	CONSTANS-like 9
5214	cassava4.1_032247m		
5214	cassava4.1_029590m	AT2G04240.1	RING/U-box superfamily protein
5214	cassava4.1_004625m	AT4G12700.1	
5214	cassava4.1_022789m	AT1G06980.1	
5214	cassava4.1_002515m (primary) cassava4.1_002643m	AT3G07660.1	Kinase-related protein of unknown function (DUF1296)
5214	cassava4.1_032232m	AT5G48220.1	Aldolase-type TIM barrel family protein
5214	cassava4.1_020020m	AT2G04400.1	Aldolase-type TIM barrel family protein
5214	cassava4.1_029284m		

Supplementary Table 2 VIGS constructs and the target genes

VIGS_constr.	CDS1	CDS2	CDS3	CDS4	CDS5
1.1	cassava4.1_008793m	cassava4.1_031223m	cassava4.1_014939m	cassava4.1_027980m	cassava4.1_027340m
1.2	cassava4.1_005161m	cassava4.1_008304m	cassava4.1_029968m	cassava4.1_033257m	cassava4.1_008192m
1.3	cassava4.1_033288m	cassava4.1_030300m	cassava4.1_021025m	cassava4.1_019925m	cassava4.1_015726m
1.4	cassava4.1_028951m	cassava4.1_031311m	cassava4.1_028772m	cassava4.1_002113m	cassava4.1_026431m
1.5	cassava4.1_026844m	cassava4.1_015989m	cassava4.1_002340m	cassava4.1_004296m	cassava4.1_005518m
1.6	cassava4.1_001288m	cassava4.1_015589m	cassava4.1_030515m	cassava4.1_000903m	cassava4.1_025566m
1.7	cassava4.1_031195m	cassava4.1_021759m	cassava4.1_034346m	cassava4.1_006675m	cassava4.1_021339m
1.8	cassava4.1_021361m	cassava4.1_033355m	cassava4.1_004950m	cassava4.1_017288m	cassava4.1_006101m
1.9	cassava4.1_012418m	cassava4.1_025392m	cassava4.1_025765m	cassava4.1_002166m	cassava4.1_017777m
2	cassava4.1_002188m	cassava4.1_025142m	cassava4.1_008294m	cassava4.1_017968m	cassava4.1_015952m
2.1	cassava4.1_020323m	cassava4.1_030256m	cassava4.1_020803m	cassava4.1_025873m	cassava4.1_005568m
2.2	cassava4.1_012736m	cassava4.1_002922m	cassava4.1_006972m	cassava4.1_004363m	cassava4.1_029493m
2.3	cassava4.1_026906m	cassava4.1_012052m	cassava4.1_016758m	cassava4.1_007335m	cassava4.1_002986m
2.4	cassava4.1_024759m	cassava4.1_023985m	cassava4.1_028362m	cassava4.1_001190m	
2.5	cassava4.1_024415m	cassava4.1_002555m	cassava4.1_018748m	cassava4.1_025682m	cassava4.1_004304m
2.6	cassava4.1_029560m	cassava4.1_027980m	cassava4.1_032247m	cassava4.1_029590m	cassava4.1_004625m
2.7	cassava4.1_029206m	cassava4.1_002515m	cassava4.1_032232m	cassava4.1_020020m	cassava4.1_029284m
2.8	cassava4.1_07986m				
2.9	cassava4.1_011768m	cassava4.1_029175m	cassava4.1_012316m	cassava4.1_012330m	cassava4.1_022227m

Supplementary Table 3 VIGS inserts that have been synthesized

>41 (VIGS1.1)
GCGGGTACCCTGTTGTTGTTGCTGCTGCTCCCGGGAACTGATTGTTTGCCTGTACAAGAGGCCTTTCCTGGTGTGCTTGTGCCTGAAGTTCGAGTTTATTCGGCAGCTTC
CCGAGATGGGTTGCTTCCTGGTGATGCGCTGCCAGAAGCCAACGGCCTGTTAGTCCCCGTCGGCCACCATCCCCGATGAATGTTTCCTTCCATGGCCGAGCTAAAGAGTT
TGAATCCCAGAAGCGAAAACTTCACATTATTAAAATGCAGTTATGCTGTCCTGGAACTAAAGGTCCTGCAGTTTCTGGAGATGATGATCTATATGAGGACTTCAATATGG
ATGAAGCATGTCGAAGGATACCTTGAATTGTCAGCGAAAACAAAGATATATTTTGCTACTGCTGCAGCTGTATGGGATGCAGATTTCTATGTCAAAGTTGATGATGGGCT
TGTGCTTTAAGTGAAGTCTTGACCCAAATAATAGGGGAAGCATCAGCAGGCCTACCAACACTTGTAGTTCCTTTTATTGGTAGTACTAGTGCG

>42 (VIGS1.2)
GCGGGTACCCTGTTGTTGTTGCTGCTGCTCCCGGGTCAGGTAACATCAGAAGATGGTTATATTCTTAGCCTTCAGAGAATGCCTGCCGAGCGGTCCGGCAAGTTAGCAGA
CAATCCACCAGTCTTGCTGCAACATGTTGTATCACCGTCCTCGCTGCAATTGCCTCTTACTGGGATCCAATTCTCATCCCAATCCCAACTGGGTCCCTCTGAAAGTAAGG
ACAGCCGTCTGATCCTGGCCTGGCCTTGTGTGGGTGGCGCTCCTTATCAGTGTATCTATGAGCTTAAGCAGCCAGTGGGGCTTTACCAGGAGCCCTGCTACTGTTCCATC
ATGTAACATGGAGTGTCATCGAATTCCGGGATCAGATCGCTCGCGCCGGAGAATTAGAACACGCATTAGAGGCTATTAAATGGGGGACTGATTATTTCATCAAAGCATGA
TACAGCTCCTTTACACTGTAATCTTTGCCGAAATGGCCTTGATTTTGACGTTTCTGTTCAAAACCCCTTTGAGGAAGCACTAGTGCG

>45 (VIGS1.3)
GCGGGTACCTTTGGATTTGGTACTTTGGACTTAGAGTAGATGGCTAGGGAGTTGAAATGGTGTTCCTTGTGTATGTATTAGTATGAAAATGATGGAATGGTTAGGGTTAG
AAACGCATCGTATTTACTTTGGATGAAGGCCCAGGAGTGCTGTTTAAGGCCTTGGCAGTGTTTGCATTGAGGGACATAAATTTGACAAAGATAGAAAGTCGGCCAACATA
CAAGTTCTTTGATTCTGTGGAGAAACACTATAACAAAAGGCTATTCTCTGTCTATGCTGGGATGACAAGCCAGAGAAGAGATGAGCCTCCGGAATGAGGAAGCATTTAAG
GCTGTTGAATTATGGCTAGCTGATAAAGCAGTTCTTCCAATTGAATGTTCTATAGCTGGAAGCATTCATCGCAACTATGATTTAGAGTCATTGGAAGTGCTGGATCTGAT
GAAAAGATAAAGCTGTTAAAGGAGGAATTTGGATACGACGATGCTTTCAACTACAAGAAAGAAAAAGATTTTGACATCACTAGTGCG

>46 (VIGS1.4)
GCGGGTACCTTTGGAAGCATATGAGAAGCAGAAGCTTAGTCACAGTGGTGACAATGAGGAATCTGAAAAATCCACTTCTGAGATCCACGATGATGCTGAGCAGGAAGCCA
GCTCTCAGACAGCCCCCCATGTACTAGACCGTATCCTCCTCGTCCTAGCTAGCAATGGTATTCTAACTTGCTCTGCAACTGCAACCATCGGCACCGAAGATGGCCAAAAC
TTGTTGATGACAATCCCCCTGCCATAATGCTTCTTTTTGAACCAAAGGGTCGACCTGAAGATGAAGACAATGATTTTTACATCCAAAGTAAGAACAAGTGCAACCTACAA
TATTAATGTAAATAAGATCTTCAAATTTGTAACAGCAAAGCTCTTTGACCTGCCATGGACACCTGAGCGAAATCTCACTATTGGTGTGGGAAATTTGGGAAGAAGAGTAT
GATAAATGGATTGATGAACACTTGACGGATTATGTTGATTTGGACCATCAACATGGTATTGTGAATGAAGCGGTCATCACTAGTGCG

>47 (VIGS1.5)
GCGGGTACCTTTGGATTCTAATTCTTCAGACTTCTGGGTGATGGTGGCTGCTCTAAAGGAGTTCATTGTCAATAAAGGTGTGGGGGAGGCACCTCTTGAGGGTTCAATAC
CAGATTGTGAATCAGTTGCAGATACCACTGAAGCATGTCTATTTAAGCGAACTAGCTTAAATCCAACTGAAATTTGTGGTAATAGCAGCACCATTTCACCTTCTCAAATC
TGAACTTGCTAATGGCGAGTATGGCGGCCATCGGAGTTCCGAGAGTACCGTCTTCTTCGACTTCATCTTCTTCACAGTCCAATTCGTCGAATCTCTTATTAGAAGACTAT
TGGGATTTGATGCCTAAATATCAGGGAGTTTCCCTTGAAAATCTGGAGATTCTGAGAGTTGTATATGGTATTTTCCCTACGTATCATAAATTATCCTAGACATGAAGGCA
GTGCTCTGCCTGCAAATAATTCAAAGCTATCGCTATTAGGACATGGACCACATGGGAAGCAAGTCGTGGAACATCATCACTAGTGCG

>49 (VIGS1.7)
GCGGGTACCTCCTGTTCCTCGCACGTATATTCTATATGGAGATGGAAAGAGGAACCCAAGCTACAATGGACTCGTCGATGCAGTTGCTCAAAATACTTATGATGCAGCTT
ATATACAGATTATTGATCCTCGGAAAGCACAAAATTTGTCAATTCTTCTAAGAGCTTTGGATGTATCTACAGAAGAAGTCGTCAATGCCTTGCAGAATGCTAAACCCCAT
GATTTGAAATCTTTTGAGCTCTCGAAAAAGTTCTGGGATGGTGATAAAAATTATGAAAAAAATGGTACTGTAAATCGGAGTGGCAAGGTCTACCAAACCAATCAAGCAAA
AAATTTGGCTTTGTTAGAGCTAGCAAAGTTGGATTATAATCTAGTGCAGTCGGTGTATCAAACGGAGCTAAAAGAGCTTGGTTGGGATTGGCATATCGCTTCGAGAAGGA
CATTAAGAGTGTTCTTGATAGATTTGTAGATTCAACAGGATGTAATTTGTTGAACCACAAATCTCTTCAACTAGTGCG

>50 (VIGS1.8)
GCGGGTACCACTAAAGCCCTTGTCGGAGCCGACCCATCAAACTCGCCTGAAATTGAGAAAAATGGGTTCGCCATGCTCGACCGATGGCTGGCCCTCCAGCTCCTCCCCAC
ATCCTCCGCTTGCTCTCCTTGCGTTTGTGATTGCTCTTCTCAGCCGCTTCTCACCATCCCTCAAGGATTGAGCAATGCTTCTTTTGCGGATTGTGCAAACTGGTTTGAAG
GTGATCTATGGAGGAAGTCTGATACTGGGGTAGCTGTCCAGGAAGATCTGTCTGAAGTTCCTGTAAATGCAAAAGATCACAGAACAGATTCACTGTCTTCTGCACATCTC
ACAGAAAGAGTATCAAACCGAGTGTGTAATGCACTTGCTTTACTTCAGTATGTGAACATCTTGTTCACTTGGACATTGATAGCATTTGATGAGAGTGTATCTGGTGCACA
TGTAGAAGATGCCTTGTTATCTCTTGAAAAGTTGACAAAAGAATTTCCATCCTTAGTTGTCCAGTCTACACTAGTGCG

>51 (VIGS1.9)
GCGGGTACCAGATACAAGATGTTGAGAAACAGGTTGACAAGGTCTCTGGACTTCTCAAGAACTTGAAGGAAGCTAATGAGGAGTCAAAGTCTGTAACAAAGGCATCTTCG
TCAATTATCATGAAGATCTTGTCGCCGAAATTCCCATTGATGGATGCTTTGTTCCTTCTCATTCCACGGTTAATATCCCAACTTCCGGTGCGCTAATGGTTTTCTTACCA
CTGATATGAATCCCAGGAAATCAAATGATCATTTCAATAGTTCCCTGAGTAAAGAAGGTGGGTTAGAGCATTATGGGAGAGTAGAGGAGGAATTGGACCAAAACCTTCTG
GTTTTAAATGCGAGGCTTCAAGAGAAACTGCTGTGGTTATCATGAATCACATCAACCTGGTTGAGTATCTCATGGATGTGATCGCGGTGGCGGCTCCGTTACTCGACGCG
CAGACGTGTCTCCCATCCAGCTACTTCCCAGAAGTCTTGATCGATTTGAAGAGTTGGTACAGCGAAGAAACTAGTGCG

>52 (VIGS2.0)
GCGGGTACCTCCACCAGTCTGCAAATGAGTATGTTCTTTTAAACTTGTCTTTTGGATTGAATGCTCCTATTAAAACAATCTATTATCTGGCTGGGCATGCATTGGAGATA
GATGAAGTGAAGGAGAGAGAAATACAGGAGAAGAGAGAATCAATATGGAAAAGACTGGAGGATGTAGCGGCCTCTAAGGCCATAAGCAATGAGGCTGTGGGAGGATGAAC
CCACCATGGAGCCGGCATGGCCTCATCTTCAAATGGTGTATGAGTTCCTTCTTAGATTTGTGGCTTCAACAGAGACTGATGCCAAGCTTGGAGCTGAGGACTACTTCTGT
AGATTTGAATACAAGGCTGCCACTGGCGGGTTCACCCCTGATCGGGTAGCTGTGTACTGTAAATGTGAGATGCCATACAAGCCAAAAAATTGGATCACTTTCTGTGTTTT
GACTGTTCCTCTGATGATAATGCCAAAAGAACTTTGAACGCGTTCCCAGTATCACCATCTGTTGAGGGCACTAGTGCG

>53 (VIGS2.1)
GCGGGTACCTGTACAAGGCCCGCAAACAAAATGTGGCAAACCATCAAGCTTTGCGTCCCCCATTGGTCGTGAACCACTCACCTGCGCACTCAAGCAATTCTCCCTAGAGA
AAGTGTATGGCATTATTGGCAAGCCTGTTGGCCATAGCAAATCTCCTACTTTATACAATGAAGCATTCAAGTCTGTTGGATTCAATGGTGTGTATATGCAGTTTCTAACA
AAATGAAGAAAGGAATTCACCCACAGAGGCAGTGGATATCTTACGTGACCCAGAGTGGAAGGTTGATGCACATTATGATGACAAAAATTGATTGGTTGGTTTCGGATTGT
TTCAACGAGGATTTCCTGTATTGGGAAATTGCATGGAAATCGGATTGCCAATGCTGTTTTTGGTTATTGGGTTGTCACAACAGTTCATTATCTGGATATGGGGTGGATCC
AGATACTCAGGCATATTGTGCATGGCCTTATCATCAATGATCTACTTTGCAATGGAAGTTCTTTCTGATACTAGTGCG

>54 (VIGS2.2)
GCGGGTACCTGGGGTGCCATGTAGTGGCAGACACGGATATTACCGATGAGCACTTGATTGTCGGAGATCCCATTGATCTGACAATGATTATAGACGAAGTGCCCCAAGAG
TTATGCGAGCTATCAGCTCTCTTTCAGTCTCAGATTCCACGGCTCGGATTCTGTCATCATCGACCTTATTCGTTATACGCCTTAGCGAATTCACCAAACTCCCAACGAGC
AAAAAGTTTATGACATCCCTTTTACAGTGGAGGGCCTGAAGAATTACACGGACCGCCGATCACCGGTATCCGGTGGTCAAAAGGTGAATGGGAATTTGATAGCATAGAGC
AAGCTGCAAGTTGCATATCTTTGTGGTTCTCAGGAACAGTTTTTGAGCAGCTACATTTGAAAGAGTATCTAGAATCTGCCTAGCAGGAGAATCAGACTGTATGACTCCTC
CTTGGAGAAATGAATCTAGTTTTGATGACTTGGCAATCAAAAGCCTAAGAAAGCAAGAGATTGAGCAGCACTAGTGCG

>55 (VIGS2.3)
GCGGGTACCGCCTTCAACATTAGATAGGAATAGTGGTGGTGTGATCAGCAGTTCTCTAAATGCGAAAGGAGGGTTGACTGCAGGGAATCTTCCAACTAAGACAATGGATC
AGTCTCATTTAACATTCTTGGTCTTCCTATGCGCCAGGATGAAAAATCATGTGCTTATTACATGCGGACTGGATCATGTAAATTTGGAGTGGCATGTAACTATCATATGT
ATGTGCCCTGAGCCTTATCCGGAGCAAAACAATGATTTTCTTAATGCCAATGGGATTAGACTTTTTCAGTTTGGGATTGAAGGTTATAAAGATGACAGGGTCAGTGGCCT
TCAGCTCCTGAAAGATGGGCAATGGATTGATGTGCCTCCTATGCGCCACTCCATTGTTGTTAACCTTGGAGACCAGCTTGAGCGATCGACCTTTGGCCACTATCATCGCC
AATGGACGTGAAATTCCGGTGGAGAACAGTAACAGCGTAGCGCTGGCGCCTTCAGGGATGAATCTCCGGACTAGTGCG

>56 (VIGS2.4)

GCGGGTACCTGGCTTTCATATTAGCAGACAATGGATACGATGTGTGGATTGCTAATACCCGTGGAAGTAGATTCAGCCGCGGACACACCTCTCTTACTCCCTATGATCCA
CCAATCTTATCATCAACTTCTTCTCGTGCTGTTCGCTGTGAGGAGCTCTTCTTGTCATCCCAACAACGGATTAAGAGTCGTACGAGGAATGCTTATGATCGGGAATTTGT
TCCACAAGGGCAGGCTGCTGCCAAGCTTATTGAAGATATATGCAGTAAGCATGGTGTGAACTGCTTAAACTTAGTGCAAGCTTTAACTGAAGCGTGTAGGTGTTGGTGGA
TACGGGTATCCAGCTCTGATAGCATTGAATGTGAAGAAAGGAGCATATGCCCCACTCAAAAGCGCATTCGAGCTTGAACAAGGGCATGTTGCATCCTGGTGGAAGCAACA
CATTTGATAATTGGAGGTTTGGGGAAGAACTAGGACCACCACCAGATGTGTATGAACGGCGTGGCAGTCACTAGTGCG

>56 (VIGS2.4)
GCGGGTACCTGGCTTTCATATTAGCAGACAATGGATACGATGTGTGGATTGCTAATACCCGTGGAAGTAGATTCAGCCGCGGACACACCTCTCTTACTCCCTATGATCCA
CCAATCTTATCATCAACTTCTTCTCGTGCTGTTCGCTGTGAGGAGCTCTTCTTGTCATCCCAACAACGGATTAAGAGTCGTACGAGGAATGCTTATGATCGGGAATTTGT
TCCACAAGGGCAGGCTGCTGCCAAGCTTATTGAAGATATATGCAGTAAGCATGGTGTGAACTGCTTAAACTTAGTGCAAGCTTTAACTGAAGCGTGTAGGTGTTGGTGGA
TACGGGTATCCAGCTCTGATAGCATTGAATGTGAAGAAAGGAGCATATGCCCCACTCAAAAGCGCATTCGAGCTTGAACAAGGGCATGTTGCATCCTGGTGGAAGCAACA
CATTTGATAATTGGAGGTTTGGGGAAGAACTAGGACCACCACCAGATGTGTATGAACGGCGTGGCAGTCACTAGTGCG

>57 (VIGS2.5)
GCGGGTACCGCTATTCCCTGCTGTTCTCCACTCCACAAGAACAGAACTGCAGTTTGGGAAGTCTGAGTTGGCTCTTTCAAGGTTTAGCAGCTGTGTCATGCTGGCTGCAG
CACCATGGACAGGGAGTTTCCGAGTTATGTGACTCTTGGCAATGACATGACCTTAAAGGGAGAAAGTTTATCAAGAAAGGCCTTGCCAAAGGACAAGTATGAGAAAGAAA
TTTGCGACTAAAACTTGCTGCCTGAAAAAGAAAAAAGGGCATGCAACGCTGGTAACTGTAGAGGAGATCAAGAAGCAGCAACCAACAAATCTGATGAAGTTGGTGTTAAT
GCAACAGCAAAAGTTATAAAGAGATACAGTAGCTTAGCGAAAGTGGAAGCAAGCTTGGCTAAAGCGAGGTCTGCCATAATACAGTTGTTGCACTCTGGAAAAGCCCATAC
ATGCTGTTCCAAGGATGGAAAAGGCTATTAGAGGACTTGATGGGCAGAGAAGGGCCATTCCTGGAGACGACTAGTGCG

>58 (VIGS2.6)
GCGGGTACCTCAGAAGTAGTCCTAATTTCAAAGTTTTTGAGCCTTCTCCCCTTGCATATTCATTGAAAAATGGAATTGAAAAATCAAGACTTAGGTTATCAAAGCTCCGC
AAATTTCTGTGTCAGAGTTCACTTGTTGCGTAAGGCTCCATTCAATTGCAGTTCCATTTGGAAATTTGTTGTTCTTTTCTTGATTTCATTTTTTTGGGTATTTGAAACAA
TGTGTAGCTATGAACAGTCTCAGCAACATGACTGCTCAGTTTGCTTGACACAATTCGAGCCAGACTCGGAGATAAACTGCTTATCCTGTGTGCTTTTACCTGTTGTAGGG
GAAACTGTGAATGATTCACTTCCAGTGGTTGAATCAGAGAAGTCGTTTAGTCGTGGCAAGTACATTATTTACAGTGGTGTTGTTCTAGCCAAGCCTGAGGTGTTCAGGAG
ACCTTGGGATTCGGTGGTCCGGCCGGAGAAGATTCTCACCCCGGGACACAGATTCTTTCTTGTCCCACTACTAGTGCG

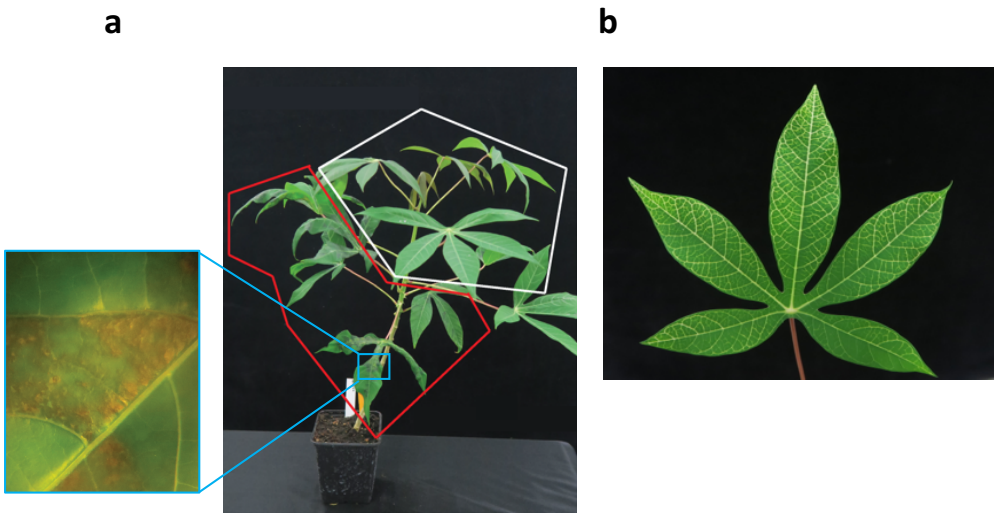
>59 (VIGS2.7)
GCGGGTACCTTTCCGCAGGCTCATTTTGTGCATGAATCTTTGGTGGAATTAGATTATAGGCCTTTGCTACTATTGTTAAATCCTTTTCCAACCAAGCATACCTTTTCATG
TGGCTGCTGAGGAGGGAGCTTATCCTGATCATCCAGAGTCACCTTCTCATGCGCCTGAAAATTTAACTGGTGAGAGCAAAGTCTCATCCAGTGCAATAAGTTGGACCTTT
TGAATTTCGCATACAGAACGAGGGTAATACCCCCAGGAACATCCTCGAGGAAATCATATGGCACAAGGACACTGAAGTCTCCCAATTAAGTTTTGGTTGGGGAGTCGATT
GTGAAACAAAATGACCCTGCCAAGGGAATAACTGGACTTTTTGGTAAAGAAATTTCATCATGAGTTGAAACTAGGAAGTTTGATTGACAATGTGGATAAGGACATGTCTC
CTTCAACAATTACAGAATTTATACACAGACAAACTTCAATCTCAGTTCGAGTATTTGTTTTGCCAAGTTACTAGTGCG

>VIGS_60 (VIGS2.9)
GCGGGTACCAGGGAGTCACCAACCCAAACCTTCGAGGGAAAAGCTTGGCTTCCCTCTTGGCAGATGCAATGGCTGCTGGAAAAACAGGAAATGAACTTCAAGCCCTGGAG
GATAATTGGCTCGCAGCAGCTCAAGTGTACAACGCCTAATGATAATACCACCATTGTTGAGATGGATCCTGGAAGTAGAAAAACATTTGATCTTAGCTATTACTCTAATT
TGCTCAAGAGAAGAGGACTTTTCCAATCAGATTCTGATACCTTAGCCTTAGTAGCTCGAGATGCAGTTTCAATGATTGGGGGACCATTTTGGGATGTTAAAACTGGACGG
AGAGATGGAAGAGTGTCAATTGCCTCGGAGGCTTTAACACAGCTGCCATCAAAGTAGTAAAGGCATGGCTCTTAATGAGCTCGTTGAGCTGAAGAGCTGCTCTTCTTGCT
TATTCTTCGAGTGTAGGAAACATAAGGATCTGTATTTATGGATGGCAAAGTGCCCTGGTGGTCCAAACTAGTGCG

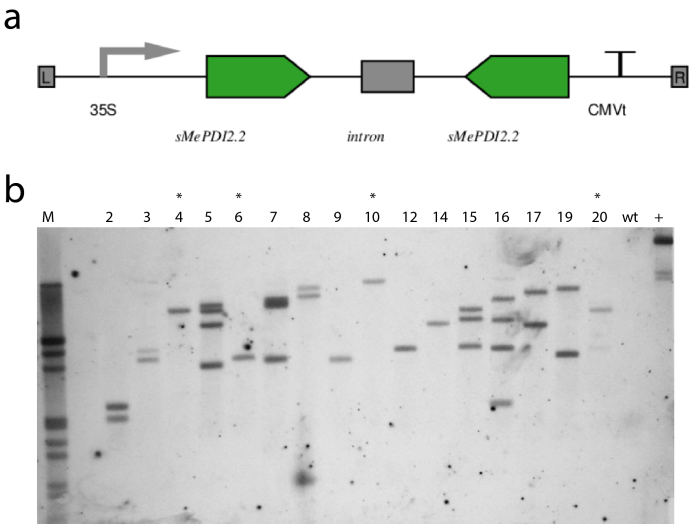
>MePDI2.2 (VIGS2.8)
GCGGGTACCAAGTGACTGAGCTAACTGGCCCAGACGTAATGGAAGAGAAGTGTGGTTCTGCTGCCATTTGTTTTGTTGCTTTCTTACCTGACATTTTGGACTCCAAGGCA
GAAGGAAGGAACAAGTACCTTGAGCAGTTGTTATCAGTTGCTGAGAAGTTCAAAAGCAATCCATACAGCTATGTTTGGACAGCTGCAGGTAAGCAGCCAGATCTTGAGAA
GCGTGTAGGTGTTGGTGGATACGGGTATCCAGCTCTGATAGCATTGAATGTGAAGAAAGGAGCATATGCCCCACTCAAAAGCGCATTCGAGCTTGAACATATTATAGAGT
TTGTTAAAGAAGCTGGGCGTGGCGGAAAGGGGAATTTGGCTTTGGGCGGTACACCAGAAATAGTGAAGACTGAGCCATGGGACGGCAAAGATGGAGAGATCATTGAAGAT
GATGAGTTCTCTCTTGAAGAACTAATGGGAGAAGATGCTGGAAGTAAGGATGAGCTATACTAGTGCG

Supplementary Table 4 VIGS assays conducted to analyze *CMD2* associated genes

genotype	SeqID	VIGS_ID	infected	not_infected	total	assay	infected	not_infected	total	infected total	plants total	incidence (%)
60444	4.1	VIGS1.1	6	0	6	#1	7	0	7 #2	13	13	100
60444	4.2	VIGS1.2	6	0	6	#1	7	0	7 #2	13	13	100
60444	4.5	VIGS1.3	8	0	8	#1	8	0	8 #2	16	16	100
60444	4.6	VIGS1.4	8	0	8	#1	6	0	6 #2	14	14	100
60444	4.7	VIGS1.5	8	0	8	#1	8	0	8 #2	16	16	100
60444	4.8	VIGS1.6	8	0	8	#1	8	0	8 #2	16	16	100
60444	4.9	VIGS1.7	8	0	8	#1	8	0	8 #2	16	16	100
60444	5.0	VIGS1.8	8	0	8	#1	8	0	8 #2	16	16	100
60444	5.1	VIGS1.9	3	5	8	#1	2	4	6 #2	5	14	36
60444	5.2	VIGS2.0	8	0	8	#1	8	0	8 #2	16	16	100
60444	5.3	VIGS2.1	8	0	8	#1	7	0	7 #2	15	15	100
60444	5.4	VIGS2.2	8	0	8	#1	8	0	8 #2	16	16	100
60444	5.5	VIGS2.3	0	8	8	#1	0	9	9 #2	0	17	0
60444	5.6	VIGS2.4	8	0	8	#1	8	0	8 #2	16	16	100
60444	5.7	VIGS2.5	8	0	8	#1	8	0	8 #2	16	16	100
60444	5.8	VIGS2.6	8	0	8	#1	5	0	5 #2	13	13	100
60444	5.9	VIGS2.7	8	0	8	#1	8	0	8 #2	16	16	100
60444	PDI5	VIGS2.8	8	0	8	#1	8	0	8 #2	16	16	100
60444	2xKin	VIGS2.9	8	0	8	#1	9	0	9 #2	17	17	100
60444	GUS-GFP	VIGS_GUS-GFP	8	0	8	#1	8	0	8 #2	16	16	100
60444	Mcs	VIGS_Mcs	4	2	6	#1	6	2	8 #2	10	14	71
TME3	4.1	VIGS1.1	0	6	6	#1	0	8	8 #2	0	14	0
TME3	4.2	VIGS1.2	0	6	6	#1	0	7	7 #2	0	13	0
TME3	4.5	VIGS1.3	1	7	8	#1	0	7	7 #2	1	15	7
TME3	4.6	VIGS1.4	1	7	8	#1	1	7	8 #2	2	16	13
TME3	4.7	VIGS1.5	0	8	8	#1	0	8	8 #2	0	16	0
TME3	4.8	VIGS1.6	0	8	8	#1	0	8	8 #2	0	16	0
TME3	4.9	VIGS1.7	0	8	8	#1	0	7	7 #2	0	15	0
TME3	5.0	VIGS1.8	0	8	8	#1	0	8	8 #2	0	16	0
TME3	5.1	VIGS1.9	0	8	8	#1	0	6	6 #2	0	14	0
TME3	5.2	VIGS2.0	0	8	8	#1	0	8	8 #2	0	16	0
TME3	5.3	VIGS2.1	0	8	8	#1	0	8	8 #2	0	16	0
TME3	5.4	VIGS2.2	0	8	8	#1	0	8	8 #2	0	16	0
TME3	5.5	VIGS2.3	0	8	8	#1	0	8	8 #2	0	16	0
TME3	5.6	VIGS2.4	1	7	8	#1	1	6	7 #2	2	15	13
TME3	5.7	VIGS2.5	0	8	8	#1	0	8	8 #2	0	16	0
TME3	5.8	VIGS2.6	0	8	8	#1	0	8	8 #2	0	16	0
TME3	5.9	VIGS2.7	0	8	8	#1	0	7	7 #2	0	15	0
TME3	PDI5	VIGS2.8	6	2	8	#1	4	3	7 #2	10	15	67
TME3	2xKin	VIGS2.9	2	6	8	#1	1	6	7 #2	3	15	20
TME3	GUS-GFP	VIGS_GUS-GFP	0	6	6	#1	0	8	8 #2	0	14	0
TME3	Mcs	VIGS_Mcs	0	6	6	#1	0	8	8 #2	0	14	0



Supplementary Figure 1 VIGS 2.3 prevents symptom development in freshly emerging leaves. Red box indicates symptomatic plant tissue with its characteristic necrotic damages. Necrotic leaf tissue is shown in the example image on the left. The white box indicates the symptom free young leaf tissue. b) Example of a leaf inoculated with VIGS -*Chl1* in TME 3. A faint chlorophyll loss was observed.



Supplementary Figure 2 Downregulation of the *MePDI2.2* using RNAi. a) Scheme of the intron hairpin MePDI-2.2-RNAi used for transformation. b) Southern blot assay for determining the number of T-DNA integration events per transgenic cassava line. *lines were used for virus inoculation experiments.

Supplementary Table 5 *MePDI2.2* mutations in TME 3 and 60444

Manes.12G068300.1 (*MePDI2.2*)					
Position*	Ref.	Alt.	Exon	non-syn	present in 60444
28	**G**	**T**	**1**	**yes**	**no**
111	**C**	**T**	**1**	**yes**	**no**
1060	G	A	2	no	yes
3293	T	C	5	no	yes
3354	G	C	5	no	yes
3360	**G**	**C**	**5**	**yes**	**no**
3927	T	G	6	yes	yes
3937	G	A	6	yes	yes
4029	G	A	6	no	yes
4281	C	G	7	no	yes
4781	G	A	9	no	no
4859	T	G	9	no	yes

*Position from start codon ATG

Supplementary Table 6 Primer sequences

Primer Name	**Sequence**	**Use**
mePP2A_genomic_F	CGC TGT GGA AAT ATG GCA TCA	Cassava qPCR reference gene
mePP2A_genomic_R	CTG GCT CAA ACT GCA GGA TCA A	
CMV_qPCR_F	GGT CCT GGA TTG CAG AGG AAG ATA GTG GG	Cassava geminiviral DNA quantitation
CMV_qPCR_R	GGT ACA ACG TCA TTG ATG ACG TCG ATC CC	
PP2A-cDNA fw	TGCAAGGCTCACACTTTCATC	Quantification of PP2A (Manes.09G039900)
PP2A-cDNA rv	CTGAGCGTAAAGCAGGGAAG	
VIGS_MCS_Fw	TGG GTC GCT GAT AAT GTT AGG	VIGS-insert confirmation
VIGS_MCS_Rev	GGA GAT ATC ATC ATT TCC ACT CC	
4.1_033257m_F	AGG CGA GGT CGA CTC TGT AGA G	Quantification of cassava4.1_033257m
4.1_033257m_R	TGC TGG GCT TTA GCT TCA TCT T	
4.1_015726mQ_F	GAA TGT GAG GGA AAA GCG AAA C	Quantification of 4.1_015726m
4.1_015726mQ_R	AGT CTG ATC TAT CGA CTC CAA TTG G	
q4.1_023563mFw	ATG CTC TTG ATG AAG CTG ATG CT	Quantification of cassava4.1_023563m
q4.1_023563mRev	TCA TAC TGG GTC GTA ATT GAG GAT T	

Chapter 5

General Discussion & Recommendations

General Discussion & Recommendations

To date, only natural resistance against geminiviruses has been confirmed to confer stable resistance under field conditions and it appeared that the diversity of geminiviruses strongly limits a RNAi-mediated resistance. As example, the first field tests that used RNAi-mediated geminivirus resistance were conducted in Brazil and Cuba and revealed geminivirus resistance in tomato and bean [173], [174]. However, in the case of tomato a single non-symptomatic virus was detected to evade the repressive sequence-specific action of the RNAi transgene. Moreover, these studies lack to expose the transgenic plants to a diverse virus population. For cassava, a confined field trial in Kenya revealed that the RNAi transgenic cassava plants appear to accumulate geminivirus species sharing the lowest similarity with the hairpin RNA expressed in the transgenic cassava. This observation suggests the current RNAi approach is only suitable when transgenic RNAi plants are exposed to viral population with limited genetic diversity [148]. Those results prompt research community not to neglect the use of natural geminivirus resistance traits present in cassava germplasm that has to date ensured stable field-proven geminivirus resistance against all known cassava-infecting geminiviruses (CGMs). To speed up molecular characterization and resistance gene (*R*-gene) discovery, high-quality cassava genomes were assembled and annotated using sequencing platforms of the third-generation. To facilitate candidate gene confirmation, a high-throughput reverse genetic platform was developed using a Virus Induced Gene Silencing (VIGS) platform. In the following sections, future directions and recommendations are discussed that should be considered for the improvement of genome as well as *CMD2* locus assemblies. Technical possibilities that could help to precisely map and evaluate *CMD2* candidate genes are also discussed.

General considerations and future perspectives to improve cassava genomes of 60444 and TME 3

In past whole genome sequencing (wgs) projects, the workload as well as the financial burden were shared between many, highly specialized labs that were often organized into big genome-sequencing consortiums. With the advent of high-throughput, cost-effective, third-generation sequencing and mapping technologies, this has changed drastically [120]. Nowadays, a single specialized lab can produce genomes of high quality within a year. Because of this remarkable evolution on sequencing methodology, it becomes realistic to expect a release of the first high-quality cassava pan-genome within the next years. A pan-genome, as it was achieved in soybean (Glycine soja) [175] or maize (*Zea mays*)[176], would entail sequencing and *de novo* assembling of several high-value cassava genomes to capture the total variability of the *Manihot esculenta* species. The experiences and methods shown in this thesis can help to optimize and plan the sequencing of many more cassava cultivars and facilitate to achieve this important milestone in cassava genetic research.

The basis of every genome sequencing attempt is the sequencing of DNA and the subsequent assembly of the raw reads into contigs. Cassava DNA was sequenced to generate a high coverage raw data (> 70X) using the long-read sequencing PacBio RS II instrument. However, in the meantime PacBio replaced the PacBio RS II instrument with the new PacBio Sequel platform that allows a ~7-times higher data throughput at the

same financial costs. It is likely that a deeper long-read sequencing would have resulted in a more continues and more haplotype phased genome assembly. Additional cost-effective alternative devices such as the MinION long-read sequencing instruments (Oxford Nanopore) are emerging on the market and the first complex genomes were assembled with using Nanopore data only [51], [177], [178]. Nanopore is especially price-effective as it currently costs $ 500 (USD) for a flow cell and $ 215 to prepare the sequencing library. In contrast, a PacBio RS II flow cell usually generates ~ 1X cassava genome coverage that would cost ~ $ 1,000 including the library preparation (www.pacb.com). Cassava has a genome of medium-size (< 1Gb), and the cost of sequencing its full genome to >70-fold genome coverage would be below $ 12,500 when using the Nanopore instruments [177]. To improve the current versions of cassava genomes, the flexibility of the Nanopore platform could be used to hunt for the ultra-long sequencing reads (> 100 kb) that are particularly useful to fill the remaining assembly gaps, especially in highly repetitive regions. We attempted to Nanopore sequence the ultra-pure hmw DNA that was generated during the optical mapping experiments using the IrysPrep Plant Tissue DNA Isolation Kit (30104Rev.A, www.bionanogenomics.com) but failed to produce data fulfilling the minimum quality requirements. Optical mapping of this DNA revealed an exceptionally high molecule N50 of 167.3 kb in TME 3 and sequencing this sample could potentially yield in very long sequencing reads. It is general known that secondary metabolites and phenols can interact with the sequencing reagents of the Nanopore instrument that could cause the failure. However, we also tried to isolate hmw DNA using another protocol (IrysPrep High Polysaccharides Plant Tissue DNA Isolation kit) from BioNano but this experiment failed as well and generated only highly fragmented DNA molecules. These examples show the current limitations of the Nanopore system and indicate the need for further optimization to isolate hmw cassava DNA that is compatible with the Nanopore instrument.

We generated the first high-quality genomes of a crop plant by combining the power of three novel sequencing and assembly methods. We followed the assembly steps as shown in previous successful examples [59], [62] and started by assembling long-reads into contigs, subsequently combining contigs into scaffolds using optical maps, and finally recreating the chromosomes with Hi-C based proximity data. The PacBio reads were assembled into contigs ranging between N50 of 116,78 kb to N50 of 97,578 kb in 60444 and TME 3, respectively. The optical maps then scaffolded the contigs to scaffolds of several Mb in size indicating a sequence contiguity improvement of ~20 fold. However, we observed that a considerable proportion of contigs were not scaffolded after applying optical mapping (537 Mb for 60444 and 564 Mb for TME 3). Due to the limited number of optical recognition sites, optical maps usually do not align to short contigs (< 80 kb) unless they are scaffolded with additional sequences such as mate-pair sequencing reads or Hi-C reads. Future analysis could be placing the Hi-C scaffolding method before the optical scaffolding that might allow a proportion of correctly placed short contigs to be confirmed or corrected by the optical maps. This in turn could potentially reduce the amount of sequences that have not been scaffolded after optical mapping, increase the sequence continuity and reduce the number of assembly gaps.

Cassava has an exceptional heterozygous genomic composition and we were interested to see how the chromosome-proximity Hi-C data can unwind the haplotypes. In theory, the Hi-C data should carry all the

information to phase haplotypes accurately along each chromosome [179]. However, we found that different Hi-C scaffolding tools produced different outputs for the haplotype phasing. In that respect, the software tool SALSA [64] generated clear haplotype structures for the *CMD2* locus that had not been observed with other tools such as LACHESIS [64], [180] or HiRise. In contrast, the commercially available software HiRise (Dovetail Genomics) provided the highest scaffold contiguity and accuracy when validated with the composite genetic map [78] but had an overall more collapsed haplotype structure as compared to other tools. To date, only four tools are currently available for sequence scaffolding and haplotype phasing (LACHESIS, SALSA, HiRISE, GRAAL)[181]. It is expected that the number of long-read assemblies for heterozygous and large crop genomes will increase in the coming years that in turn will increase the demand for standardized haplotype-phasing tools [182]. Future projects should attempt to develop novel scaffolding tools that exploits the full potential of Hi-C or optical mapping for haplotype phasing.

The two cassava genomes are of high-quality but are not complete and carry a substantial amount of sequencing- and assembly-gaps. For future research, these assembly gaps have to be addressed with sophisticated software tools, additional whole genome sequencing (i.e. using the Nanopore platform) or with a more targeted approach such as bacterial artificial chromosomes (BAC) sequencing. The two genomes were sequence polished using the tool QUIVER [50] that aligns the PacBio reads to the final assembly for error corrections. More computational approaches exist and should be tested in the future for error correction and gap closure. For example, in an recent genome assembly project of the goat genome, 681 from 1,439 gaps could be closed with the combination of computational tools [59]. The software tools PBjelly2 or GMcloser could be run with all raw long sequencing reads to correct erroneously scaffolded contigs and to close the remaining gaps [183], [184]. Also, PILON should be tested for gap closure and sequence correction [185]. However, future research has to show if efficient gap closure can also be achieved in a highly repetitive crop plant genome. Since long PacBio reads carry a substantial amount of sequencing errors (15-20%) [186], an initial error correction of the long PacBio reads could potentially improve the genome assembly and reduce the number of gaps. The software tool PROOVREAD uses the accurate short sequencing reads for correction by aligning them back to the long raw PacBio read [187]. The resulting error corrected sequences have to be carefully tested since concerns are reported that such hybrid tools can generate unreliable corrections, especially within highly repetitive sequences [182]. Since cassava has a highly-repetitive genome, in fact >65% of the genome contain repeats, such a hybrid-error correction attempt should only be considered when too low long-read genome coverage is available (<70-fold coverage). However, these *in silico* approaches have the advantage to rapidly improve the current version of the genomes with only very little effort. In the frame of the cassava genome sequencing project, whole genome shotgun BAC libraries for the two cassava genotypes 60444 and TME 3 have been generated and screening BACs using *CMD2*-associated markers is ongoing (H. Vanderschuren, personal communication). BAC libraries are useful for genome validation and genome polishing and can help filling sequencing gaps manually. Because assembly gaps mainly occur in repeat-rich regions, a targeted BAC-based gap filling can be challenging since the platform relies on specific gap flanking sites. It is also possible to perform a non-targeted characterization of BAC libraries by short-read mate-pair sequencing of the BAC-ends [188]. To use the full potential of BACs,

future direction should attempt to generate such long-distance mate-pair sequencing reads that would allow a precise validation of scaffolds and facilitate the scaffolding of the remaining contigs.

Future perspectives for genetic mapping of the *CMD2*

A map-based or positional cloning approach for a gene can be summarized into basic steps, starting with identifying a marker tightly linked to the gene using a mapping population, finding BACs to which the marker probe hybridizes, creating new co-segregating markers from the BAC clone, perform genetic complementation (transformation) to rescue the wild-type phenotype and finally sequence the gene and determine if the function is known [189]. In context of the *CMD2*, neither the bi-parental full-sib F1 mapping population (n=180) from Rabbi and colleagues, nor the large-scale GWAS study conducted by Wolfe and colleagues, generated a dense enough mapping to allow the precise location of the *CMD2* within a narrow genetic region [34], [35]. These two recent studies indicate the limitation of the current gene-discovery approaches and show the great need to improve crossing capability and to develop new gene-mapping platforms for cassava.

Recently, a novel breeding option was published by Bull and colleagues, where they triggered an early-flowering phenotype in cassava 60444 by over-expressing the *Arabidopsis FLOWERING LOCUS T* (*AtFT*)[73]. This induced flowering system could be used to enable mating of genotypes with asynchronous flowering but with valuable traits such as CMD resistance. For example, by applying this system a *CMD2* segregating population from crossings between 60444 and TME 3 could be achieved. However, it is difficult to estimate if such a platform could generate a large-enough number of offspring plants. The recently published 'speed breeding' platform could potentially further facilitate breeding and mating of these cassava plants but future investigations have to prove that the light-mediated breeding tool can be effectively applied in a tropical plant system [190].

To further facilitate trait mapping in cassava, novel genetic marker systems and gene mapping platforms should be considered and tested. For instance, mapping-by-sequencing has emerged as a powerful platform for genetic mapping in several plant and animal species and was termed by Schneeberger and Weigel as the 'fast-forward genetics' tool [41]. It uses a combination of bulked segregants analysis (BSA) with high-throughput shallow sequencing for the rapid detection of mutant alleles [191]. This gene-discovery platform was successfully applied on crop plant species and led to the discovery of agronomically important genes [40], [191]–[194]. The expected segregation of the *CMD2* is 1:1 in the F1 population because it is reported to be inherited as a monogenic, heterozygous and a dominant gene [34]. Following the mapping-by-sequencing approach, phenotypically DNA pools of segregating plants could be generated and shallow sequenced. As an example, exome capture assays in combination with mapping-by-sequencing revealed many agronomic relevant genes [195]–[197]. For instance, this platform has been successfully deployed in crop systems such as barley and wheat to identify the *many noded dwarf* gene [198] as well as the wheat stem rust resistance genes *Sr22* and *Sr45* [199]. For the *Sr45* gene, a targeted enrichment was applied for leucine-rich repeat containing proteins (NLRs). However, because not all resistance genes are NLRs and no

gene was found at the *CMD2* locus with such an annotation, this method might not be suitable for future *CMD2* cloning attempts. Another interesting approach would be to sort chromosomes prior to shallow sequencing. Such an approach was successfully applied in wheat for a rapid isolation of the wheat *Pm2* and barley *Eceriferium-q* gene required for epicuticular aliphatic wax accumulation [200]. However, only a specialized laboratory can provide sufficient capacity to satisfy the demand for chromosome sorting. For cassava, only a single cytogenetic study has been published that used conventional staining and cytological markers (DAPI-staining) [201]. The same study also revealed that mitotic karyotypes show similar chromosome sizes and have a variable, but distinguishable number of satellite DNA elements. This information could facilitate the development of a chromosome sorting platform for cassava to specifically shallow sequence the chromosome 12 from plants segregating for *CMD2*.

High-throughput gene screening for cassava

It was the aim of this thesis to generate a flexible and high-throughput candidate gene screening platform for cassava that allows to test QTL associated gene(s). The VIGS clone previously developed in the lab (Lentz et al., under review), was tuned into a highly-flexible and cost-effective gene screening platform that allows functional screening of docents of candidate gene within a single assay.

The gene characterization work, that has been done for *CMD2* candidate genes, relied strongly on the cassava genome AM560, a partially inbred line derived from a Latin American cassava genotype. Considering that CMD is not endemic to South America, the AM560 reference genome may not contain the functional sources of CMD resistance. However, the four candidate genes revealed by VIGS were also detected in the *CMD2* locus of the two new genomes underpinning the validity of the *CMD2* investigation conducted in this thesis. By using VIGS, we identified the *MePDI2-2* at the *CMD2* locus that allowed virus replication in *CMD2*-type TME 3. Based on functional prediction, the gene homolog in TME 3 (MeTME3_00015870-RA) catalyzes the correct folding of proteins and prevent the aggregation of unfolded or partially folded precursors. In contrast to our results, the barley (*Hordeum vulgare L.*) ortholog of *MePDI-2.2* (*HvPDI5-1*), that carries several non-synonymous SNPs, act as a virus susceptibility factor [129] and causes resistance to bymoviruses, a single stranded RNA virus. Moreover, the suppression of members of the PDI gene family can delay replication of several mammalian viruses (e.g. HIV) but their role in interactions with viruses remains largely unknown [130], [131]. Sanger-sequencing and haplotype analysis revealed TME 3 – specific non-synonymous SNPs that affect the thioredoxin conserved domains of the gene. The recent finding of PDIs being involved in virus susceptibility stand in contrast to the VIGS results found for *MePDI2-2* where silencing of *MePDI2-2* in virus susceptible plant 60444 did not lead to a reduced virus incidence. Furthermore, the *MePDI2-2* knock-down through RNAi led to a reduced virus incidence and symptom score in virus susceptible 60444 plants that, however, supported the function as a virus susceptible factor. To solve the discrepancy between VIGS- and RNAi- plants after virus-inoculation, a gene-knock out and *MePDI2-2* overexpressing transgenic lines should be generated for a *CMD2*-type

cassava plants, although the experimental setup could be complicated by the loss of CMD resistance in *CMD2*-type genotypes following induced *in vitro* embryogenesis [36]. The sequencing, *de novo* assembly and gene-space annotation performed in this thesis will be instrumental to further advance the VIGS work on *CMD2* associated genes presented in chapter 3.

Concluding remarks

The two high-quality cassava genomes that have been assembled in the course of this work will be instrumental to improve our understanding of cassava genomics and future characterization of cassava diversity. It is expected that sequencing and assembly price will further decrease over the coming years and the assembly pipeline used in this work can provide helpful guidance for future cassava genome sequencing project. The ultimate goal should be to generate the first high-quality, diploid-aware cassava pan-genome that includes full genome information for the other two geminivirus resistance sources *CMD1* and *CMD3* identified in cassava germplasm [33], [38].

The two high-quality genomes were used to reconstruct the major geminivirus resistance locus *CMD2*. This revealed a high syntenic relation between *CMD2*-type cultivar TME 3 and CMD susceptible cultivar 60444. In the scope of this chapter, the first list of *de novo* annotated *CMD2* associated genes were presented that now facilitate targeted candidate gene screening using either high-throughput methods (VIGS) or targeted reverse genetics approaches. The *CMD2* locus visualization revealed also the detailed location of the remaining assembly gaps that have to be targeted by BAC sequencing or additional Nanopore long-read, full genome sequencing. The complete assembly of the *CMD2* will be key to confirm gene space annotation as well as to identify the cause of the *CMD2* breakdown after *in vitro* induced embryogenesis.

Since CMGs are evolving fast over the past decades, breeding and gene-mapping platforms have to be reconsidered and improved. Future *CMD2* mapping attempts should use the new *CMD2*-type genomes together with novel marker-systems (i.e. mapping-by-sequencing approach) in order to narrow down the number of candidate genes. Because of the loss of the *CMD2* after embryogenesis, cassava transformation platforms have to use alternative resistance sources (*CMD3* or *CMD1*) to ensure the CMG resistance in transgenic cassava. Once the mono-genic and dominant *CMD2* is isolated and identified, the resistance source should be stacked with other resistance sources such as *CMD1* or *CMD3* to generate durable and stable CMG resistance in the field.

Bibliography

[1] A. C. Allem, "The origins and taxonomy of cassava.," *Cassava Biol. Prod. Util.*, pp. 1–16, 2002.

[2] M. Balat and H. Balat, "Recent trends in global production and utilization of bio-ethanol fuel," *Appl. Energy*, vol. 86, no. 11, pp. 2273–2282, 2009.

[3] P. M. Schmitz and A. Kavallari, "Crop plants versus energy plants-On the international food crisis," *Bioorganic Med. Chem.*, vol. 17, no. 12, pp. 4020–4021, 2009.

[4] M. E. Halsey, K. M. Olsen, N. J. Taylor, and P. Chavarriaga-aguirre, "Reproductive Biology of Cassava (Manihot esculenta Crantz) and Isolation of Experimental Field Trials," no. February, pp. 49–58, 2008.

[5] J. P. Legg and J. M. Thresh, "Cassava mosaic virus disease in East Africa: a dynamic disease in a changing environment," *Virus Res.*, vol. 71, no. 1–2, pp. 135–149, 2000.

[6] J. P. Legg, P. L. Kumar, T. Makeshkumar, L. Tripathi, M. Ferguson, E. Kanju, P. Ntawuruhunga, and W. Cuellar, *Cassava Virus Diseases : Biology , Epidemiology , and Management*, 1st ed., vol. 91. Elsevier Inc., 2015.

[7] J. P. Legg, B. Owor, P. Sseruwagi, and J. Ndunguru, "Cassava Mosaic Virus Disease in East and Central Africa: Epidemiology and Management of A Regional Pandemic," *Adv. Virus Res.*, vol. 67, no. 06, pp. 355–418, 2006.

[8] L. Hanley-Bowdoin, E. R. Bejarano, D. Robertson, and S. Mansoor, "Geminiviruses: masters at redirecting and reprogramming plant processes.," *Nat. Rev. Microbiol.*, vol. 11, no. 11, pp. 777–88, Nov. 2013.

[9] H. Vanderschuren, M. Stupak, J. Fütterer, W. Gruissem, and P. Zhang, "Engineering resistance to geminiviruses--review and perspectives.," *Plant Biotechnol. J.*, vol. 5, no. 2, pp. 207–20, Mar. 2007.

[10] D. N. Shepherd, D. P. Martin, E. Van Der Walt, K. Dent, A. Varsani, and E. P. Rybicki, "Maize streak virus: An old and complex 'emerging' pathogen," *Mol. Plant Pathol.*, vol. 11, no. 1, pp. 1–12, 2010.

[11] W. Zhang, N. H. Olson, T. S. Baker, L. Faulkner, M. Agbandje-McKenna, M. I. Boulton, J. W. Davies, and R. McKenna, "Structure of the maize streak virus geminate particle," *Virology*, vol. 279, no. 2, pp. 471–477, 2001.

[12] J. Navas-Castillo, E. Fiallo-Olivé, and S. Sánchez-Campos, "Emerging Virus Diseases Transmitted by Whiteflies," *Annu. Rev. Phytopathol.*, vol. 49, no. 1, pp. 219–248, Aug. 2011.

[13] E. Glick, A. Zrachya, Y. Levy, A. Mett, D. Gidoni, E. Belausov, V. Citovsky, and Y. Gafni, "Interaction with host SGS3 is required for suppression of RNA silencing by tomato yellow leaf curl virus V2 protein (vol 105, pg 157, 2007)," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 11, p. 4571, 2009.

[14] J. Zhang, J. Dong, Y. Xu, and J. Wu, "V2 protein encoded by Tomato yellow leaf curl China virus is an RNA silencing suppressor," *Virus Res.*, vol. 163, no. 1, pp. 51–58, 2012.

[15] I. Amin, K. Hussain, R. Akbergenov, J. S. Yadav, J. Qazi, S. Mansoor, T. Hohn, C. M. Fauquet, and R. W. Briddon, "Suppressors of RNA Silencing Encoded by the Components of the Cotton Leaf Curl Begomovirus-BetaSatellite Complex," *Mol. Plant-Microbe Interact.*, vol. 24, no. 8, pp. 973–983, 2011.

[16] H. Vanderschuren, A. Alder, P. Zhang, and W. Gruissem, "Dose-dependent RNAi-mediated geminivirus resistance in the tropical root crop cassava.," *Plant Mol. Biol.*, vol. 70, no. 3, pp. 265–72, Jun. 2009.

[17] K. Bonfim, J. C. Faria, E. O. P. L. Nogueira, E. a Mendes, F. J. L. Aragão, É.

a Mendes, F. J. L. Aragão, E. Recursos, P. W. Norte, and U. De Brasília, "RNAi-Mediated Resistance to Bean golden mosaic virus in Genetically Engineered Common Bean (Phaseolus vulgaris)," *Mol. Plant. Microbe. Interact.*, vol. 20, no. 6, pp. 717–726, 2007.

[18] A. Fuentes, P. L. Ramos, E. Fiallo, D. Callard, Y. Sánchez, R. Peral, R. Rodríguez, and M. Pujol, "Intron-hairpin RNA derived from replication associated protein C1 gene confers immunity to tomato yellow leaf curl virus infection in transgenic tomato plants," *Transgenic Res.*, vol. 15, no. 3, pp. 291–304, 2006.

[19] D. De Ronde, P. Butterbach, R. Kormelink, and R. K. Richardkormelinkwurnl, "Dominant resistance against plant viruses : Supplementary Material Correspondence :," no. 0, pp. 1–10.

[20] M. Lapidot, U. Karniel, D. Gelbart, D. Fogel, D. Evenor, Y. Kutsher, Z. Makhbash, S. Nahon, H. Shlomo, L. Chen, M. Reuveni, and I. Levin, "A Novel Route Controlling Begomovirus Resistance by the Messenger RNA Surveillance Factor Pelota," *PLoS Genet.*, vol. 11, no. 10, pp. 1–19, 2015.

[21] M. G. Verlaan, S. F. Hutton, R. M. Ibrahem, R. Kormelink, R. G. F. Visser, J. W. Scott, J. D. Edwards, and Y. Bai, "The Tomato Yellow Leaf Curl Virus Resistance Genes Ty-1 and Ty-3 Are Allelic and Code for DFDGD-Class RNA-Dependent RNA Polymerases," *PLoS Genet.*, vol. 9, no. 3, 2013.

[22] P. N. Dodds, G. J. Lawrence, A.-M. Catanzariti, T. Teh, C.-I. A. Wang, M. A. Ayliffe, B. Kobe, and J. G. Ellis, "Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes," *Proc. Natl. Acad. Sci.*, vol. 103, no. 23, pp. 8888–8893, 2006.

[23] V. Nicaise, "Crop immunity against viruses: outcomes and future challenges," *Front. Plant Sci.*, vol. 5, no. November, pp. 1–18, 2014.

[24] D. de Ronde, P. Butterbach, and R. Kormelink, "Dominant resistance against plant viruses.," *Front. Plant Sci.*, vol. 5, no. June, p. 307, Jan. 2014.

[25] K. B. G. Scholthof, S. Adkins, H. Czosnek, P. Palukaitis, E. Jacquot, T. Hohn, B. Hohn, K. Saunders, T. Candresse, P. Ahlquist, C. Hemenway, and G. D. Foster, "Top 10 plant viruses in molecular plant pathology," *Mol. Plant Pathol.*, vol. 12, no. 9, pp. 938–954, 2011.

[26] I. Anbinder, M. Reuveni, R. Azari, I. Paran, S. Nahon, H. Shlomo, L. Chen, M. Lapidot, and I. Levin, "Molecular dissection of Tomato leaf curl virus resistance in tomato line TY172 derived from Solanum peruvianum," *Theor. Appl. Genet.*, vol. 119, no. 3, pp. 519–530, 2009.

[27] M. G. Verlaan, S. F. Hutton, R. M. Ibrahem, R. Kormelink, R. G. F. Visser, J. W. Scott, J. D. Edwards, and Y. Bai, "The Tomato Yellow Leaf Curl Virus resistance genes Ty-1 and Ty-3 are allelic and code for DFDGD-class RNA-dependent RNA polymerases.," *PLoS Genet.*, vol. 9, no. 3, p. e1003399, Mar. 2013.

[28] S. Maiti, S. Paul, and A. Pal, "Isolation, characterization, and structure analysis of a non-TIR-NBS-LRR encoding candidate gene from MYMIV-resistant Vigna mungo," *Mol. Biotechnol.*, vol. 52, no. 3, pp. 217–233, 2012.

[29] S. Maiti, J. Basak, S. Kundagrami, A. Kundu, and A. Pal, "Molecular marker-assisted genotyping of Mungbean Yellow Mosaic India Virus resistant germplasms of mungbean and urdbean," *Mol. Biotechnol.*, vol. 47, no. 2, pp. 95–104, 2011.

[30] E. Okogbenin, M. C. M. Porto, C. Egesi, C. Mba, E. Espinosa, L. G. Santos, C. Ospina, J. Marín, E. Barrera, J. Gutiérrez, I. Ekanayake, C. Iglesias, and M. A. Fregene, "Marker-assisted introgression of resistance to cassava mosaic

disease into latin American germplasm for the genetic improvement of cassava in Africa," *Crop Sci.*, vol. 47, no. 5, pp. 1895–1904, 2007.
[31] O. Akano, O. Dixon, E. Barrera, and M. Fregene, "Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease.," *Tag Theor. Appl. Genet. Theor. Und Angew. Genet.*, vol. 105, no. 4, pp. 521–525, Sep. 2002.
[32] V. N. Fondong, "The Search for Resistance to Cassava Mosaic Geminiviruses: How Much We Have Accomplished, and What Lies Ahead," *Front. Plant Sci.*, vol. 8, no. March, pp. 1–19, 2017.
[33] E. Okogbenin, C. N. Egesi, B. Olasanmi, O. Ogundapo, S. Kahya, P. Hurtado, J. Marin, O. Akinbo, C. Mba, H. Gomez, C. De Vicente, S. Baiyeri, M. Uguru, F. Ewa, and M. Fregene, "Molecular marker analysis and validation of resistance to cassava mosaic disease in elite cassava genotypes in Nigeria," *Crop Sci.*, vol. 52, no. 6, pp. 2576–2586, 2012.
[34] I. Y. Rabbi, M. T. Hamblin, P. L. Kumar, M. a. Gedil, A. S. Ikpan, J. L. Jannink, and P. a. Kulakow, "High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding," *Virus Res.*, vol. 186, pp. 87–96, Jun. 2014.
[35] M. D. Wolfe, I. Y. Rabbi, C. Egesi, M. Hamblin, R. Kawuki, P. Kulakow, R. Lozano, D. P. Del Carpio, P. Ramu, and J.-L. Jannink, "Genome-wide association and prediction reveals the genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement," *Plant Genome*, vol. 9, no. 2, pp. 1–13, 2016.
[36] G. Beyene, R. D. Chauhan, H. Wagaba, T. Moll, T. Alicai, D. Miano, J. C. Carrington, and N. J. Taylor, "Loss of CMD2-mediated resistance to cassava mosaic disease in plants regenerated through somatic embryogenesis," *Mol. Plant Pathol.*, vol. 17, no. 7, pp. 1095–1110, 2016.
[37] F. Allie, E. J. Pierce, M. J. Okoniewski, and C. Rey, "Transcriptional analysis of South African cassava mosaic virus-infected susceptible and tolerant landraces of cassava highlights differences in resistance, basal defense and cell wall associated genes during infection.," *BMC Genomics*, vol. 15, no. 1, p. 1006, Nov. 2014.
[38] O. E., M. I., T. J., F. C. M., M. G., and F. M., "Marker-Assisted Breeding for Cassava Mosaic Disease Resistance," *Translational Genomics for Crop Breeding*. 11-Oct-2013.
[39] S. Prochnik, P. R. Marri, B. Desany, P. D. Rabinowicz, C. Kodira, M. Mohiuddin, F. Rodriguez, C. Fauquet, J. Tohme, T. Harkins, D. S. Rokhsar, and S. Rounsley, "The Cassava Genome: Current Progress, Future Directions," *Trop. Plant Biol.*, vol. 5, no. 1, pp. 88–94, Jan. 2012.
[40] K. Schneeberger, "Using next-generation sequencing to isolate mutant genes from forward genetic screens," *Nat. Rev. Genet.*, vol. 15, no. 10, pp. 662–676, 2014.
[41] K. Schneeberger and D. Weigel, "Fast-forward genetics enabled by new sequencing technologies," *Trends Plant Sci.*, vol. 16, no. 5, pp. 282–288, 2011.
[42] W. Wang, B. Feng, J. Xiao, Z. Xia, X. Zhou, P. Li, W. Zhang, Y. Wang, B. L. Møller, P. Zhang, M.-C. Luo, G. Xiao, J. Liu, J. Yang, S. Chen, P. D. Rabinowicz, X. Chen, H.-B. Zhang, H. Ceballos, Q. Lou, M. Zou, L. J. C. B. Carvalho, C. Zeng, J. Xia, S. Sun, Y. Fu, H. Wang, C. Lu, M. Ruan, S. Zhou, Z. Wu, H. Liu, R. M. Kannangara, K. Jørgensen, R. L. Neale, M. Bonde, N. Heinz, W. Zhu, S. Wang, Y. Zhang, K. Pan, M. Wen, P.-A. Ma, Z. Li, M. Hu, W. Liao, W. Hu, S. Zhang, J. Pei, A. Guo, J. Guo, J. Zhang, Z. Zhang, J. Ye,

W. Ou, Y. Ma, X. Liu, L. J. Tallon, K. Galens, S. Ott, J. Huang, J. Xue, F. An, Q. Yao, X. Lu, M. Fregene, L. A. B. López-Lavalle, J. Wu, F. M. You, M. Chen, S. Hu, G. Wu, S. Zhong, P. Ling, Y. Chen, Q. Wang, G. Liu, B. Liu, K. Li, and M. Peng, "Cassava genome from a wild ancestor to cultivated varieties," *Nat. Commun.*, vol. 5, p. 5110, 2014.

[43] S. I. Kayondo, D. P. Del Carpio, R. Lozano, A. Ozimati, M. Wolfe, Y. Baguma, V. Gracen, S. Offei, M. Ferguson, R. Kawuki, and J. L. Jannink, "Genome-wide association mapping and genomic prediction for CBSD resistance in Manihot esculenta," *Sci. Rep.*, vol. 8, no. 1, pp. 1–11, 2018.

[44] E. A. Masumba, F. Kapinga, G. Mkamilo, K. Salum, H. Kulembeka, S. Rounsley, J. V. Bredeson, J. B. Lyons, D. S. Rokhsar, E. Kanju, M. S. Katari, A. A. Myburg, N. A. van der Merwe, and M. E. Ferguson, "QTL associated with resistance to cassava brown streak and cassava mosaic diseases in a bi-parental cross of two Tanzanian farmer varieties, Namikonga and Albert," *Theor. Appl. Genet.*, vol. 130, no. 10, pp. 2069–2090, 2017.

[45] T. Amuge, D. K. Berger, M. S. Katari, A. A. Myburg, S. L. Goldman, and M. E. Ferguson, "A time series transcriptome analysis of cassava (Manihot esculenta Crantz) varieties challenged with Ugandan cassava brown streak virus," *Sci. Rep.*, vol. 7, no. 1, pp. 1–21, 2017.

[46] R. S. Bart, M. C. Wilson, A. M. Mutka, A. W. Hummel, J. Berry, R. D. Chauhan, A. Vijayaraghavan, N. J. Taylor, D. F. Voytas, D. H. Chitwood, and R. S. Bart, "Rapid report Gene expression atlas for the food security crop cassava," 2017.

[47] R. B. Anjanappa, D. Mehta, M. J. Okoniewski, A. Szabelska-Bereseqicz, W. Gruissem, and H. Vanderschuren, "Molecular insights into Cassava brown streak virus susceptibility and resistance by profiling of the early host response," *Mol. Plant Pathol.*, pp. 1–14, 2017.

[48] H. Wang, G. Beyene, J. Zhai, S. Feng, N. Fahlgren, N. J. Taylor, R. Bart, J. C. Carrington, S. E. Jacobsen, and I. Ausin, "CG gene body DNA methylation changes and evolution of duplicated genes in cassava," *Proc. Natl. Acad. Sci.*, vol. 112, no. 44, pp. 13729–13734, 2015.

[49] J. V Bredeson, J. B. Lyons, S. E. Prochnik, G. A. Wu, C. M. Ha, E. Edsinger-Gonzales, J. Grimwood, J. Schmutz, I. Y. Rabbi, C. Egesi, P. Nauluvula, V. Lebot, J. Ndunguru, G. Mkamilo, R. S. Bart, T. L. Setter, R. M. Gleadow, P. Kulakow, M. E. Ferguson, S. Rounsley, and D. S. Rokhsar, "Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity," *Nat. Biotechnol.*, no. April, 2016.

[50] C. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner, and J. Korlach, "Nonhybrid , finished microbial genome assemblies from long-read SMRT sequencing data," vol. 10, no. 6, 2013.

[51] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose, "Nanopore sequencing and assembly of a human genome with ultra-long reads," *Nat. Biotechnol.*, vol. 36, no. 4, 2018.

[52] C. Chin, P. Peluso, F. J. Sedlazeck, M. Nattestad, G. T. Concepcion, A. Clum, C. Dunn, R. O. Malley, R. Figueroa-balderas, A. Morales-cruz, G. R. Cramer, M. Delledonne, C. Luo, J. R. Ecker, D. Cantu, D. R. Rank, and M. C. Schatz, "Phased diploid genome assembly with single-molecule real-time sequencing," vol. 13, no. 12, 2016.

[53] M. Pendleton, R. Sebra, A. W. C. Pang, A. Ummat, O. Franzen, T. Rausch, A. M. Stütz, W. Stedman, T. Anantharaman, A. Hastie, H. Dai, M. H.-Y. Fritz, H. Cao, A. Cohain, G. Deikus, R. E. Durrett, S. C. Blanchard, R. Altman, C.-S. Chin, Y. Guo, E. E. Paxinos, J. O. Korbel, R. B. Darnell, W. R. McCombie, P.-Y. Kwok, C. E. Mason, E. E. Schadt, and A. Bashir, "Assembly and diploid architecture of an individual human genome via single-molecule technologies," *Nat. Methods*, vol. 12, no. 8, pp. 780–786, 2015.

[54] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, "Canu : scalable and accurate long-read assembly via adaptive k - mer weighting and repeat separation," pp. 722–736, 2017.

[55] K. Berlin, S. Koren, C.-S. Chin, J. P. Drake, J. M. Landolin, and A. M. Phillippy, "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing," *Nat. Biotechnol.*, vol. 33, no. 6, 2015.

[56] J. Chu, H. Mohamadi, R. L. Warren, C. Yang, and I. Birol, "Innovations and challenges in detecting long read overlaps: An evaluation of the state-of-the-art," *Bioinformatics*, vol. 33, no. 8, pp. 1261–1270, 2017.

[57] D. C. Schwartz, X. Li, L. I. Hernandez, S. P. Ramnarain, E. J. Huff, and Y. K. Wang, "Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping," *Science (80-.).*, vol. 262, no. 5130, p. 110 LP-114, Oct. 1993.

[58] E. T. Lam, A. Hastie, C. Lin, D. Ehrlich, S. K. Das, M. D. Austin, P. Deshpande, H. Cao, N. Nagarajan, M. Xiao, and P. Kwok, "Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly," *Nat. Biotechnol.*, vol. 30, no. 8, pp. 771–776, 2012.

[59] D. M. Bickhart, B. D. Rosen, S. Koren, B. L. Sayre, A. R. Hastie, S. Chan, J. Lee, E. T. Lam, I. Liachko, S. T. Sullivan, J. N. Burton, H. J. Huson, J. C. Nystrom, C. M. Kelley, J. L. Hutchison, Y. Zhou, J. Sun, A. Crisà, F. A. Ponce de León, J. C. Schwartz, J. A. Hammond, G. C. Waldbieser, S. G. Schroeder, G. E. Liu, M. J. Dunham, J. Shendure, T. S. Sonstegard, A. M. Phillippy, C. P. Van Tassell, and T. P. L. Smith, "Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome," *Nat. Genet.*, vol. 49, no. 4, pp. 643–650, 2017.

[60] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Chromosome Conformation," *Adv. Sci.*, vol. 295, no. 5558, pp. 1306–1311, 2012.

[61] J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, and J. Shendure, "Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions.," *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1119–25, 2013.

[62] M. Mascher, H. Gundlach, A. Himmelbach, S. Beier, S. O. Twardziok, T. Wicker, V. Radchuk, C. Dockter, P. E. Hedley, J. Russell, M. Bayer, L. Ramsay, H. Liu, G. Haberer, Q. Zhang, Q. Zhang, R. A. Barrero, L. Li, S. Taudien, M. Groth, M. Felder, A. Hastie, H. Sta, J. Vr, S. Chan, R. Ounit, S. Wanamaker, D. Bolser, C. Colmsee, T. Schmutzer, L. Aliyeva-, S. Grasso, J. Tanskanen, A. Chailyan, D. Sampath, D. Heavens, L. Clissold, S. Cao, B. Chapman, F. Dai, Y. Han, H. Li, X. Li, C. Lin, J. K. Mccooke, C. Tan, P. Wang, S. Wang, S. Yin, G. Zhou, J. A. Poland, M. I. Bellgard, L. Borisjuk, A. Houben, J. Dole, S. Ayling, S. Lonardi, P. Kersey, P. Langridge, G. J. Muehlbauer, M. D. Clark, M. Caccamo, A. H. Schulman, K. F. X. Mayer, M. Platzer, T. J. Close, U. Scholz, M. Hansson, G. Zhang, I. Braumann, M. Spannagl, C. Li, R. Waugh, N. Stein, P. Genetics, G. Centre, P. Genome, S. Biology, E. Health, M. Biology, L. Sciences, A. Export, G. Innovation, S.

Perth, C. Genomics, P. Genetics, S. Diego, E. Botany, R. Han, C. Republic, P. Sciences, C. Science, E. Molecular, E. Sciences, G. Technology, V. Plant, S. Centre, W. Genetics, P. Pathology, P. Biology, and A. Botany, "A chromosome conformation capture ordered sequence of the barley genome," *Nat. Publ. Gr.*, vol. 544, no. 7651, pp. 1–43, 2017.

[63] D. M. Bickhart, B. D. Rosen, S. Koren, B. L. Sayre, A. R. Hastie, S. Chan, J. Lee, E. T. Lam, I. Liachko, S. T. Sullivan, J. N. Burton, H. J. Huson, C. M. Kelley, J. L. Hutchison, Y. Zhou, J. Sun, A. Crisa, F. A. Ponce de Leon, J. C. Schwartz, J. A. Hammond, G. C. Waldbieser, S. G. Schroeder, G. E. Liu, M. J. Dunham, J. Shendure, T. S. Sonstegard, A. M. Phillippy, C. P. Van Tassell, and T. P. L. Smith, "Single-molecule sequencing and conformational capture enable de novo mammalian reference genomes," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2016.

[64] J. Ghurye, M. Pop, S. Koren, D. Bickhart, and C. Chin, "Scaffolding of long read assemblies using long range contact information," pp. 1–11, 2017.

[65] P. Ramu, W. Esuma, R. Kawuki, I. Y. Rabbi, C. Egesi, J. V Bredeson, R. S. Bart, J. Verma, E. S. Buckler, and F. Lu, "Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation," *Nat. Genet.*, no. August 2016, pp. 1–7, 2017.

[66] A. Parmar, B. Sturm, and O. Hensel, "Crops that feed the world : Production and improvement of cassava for food , feed , and industrial uses," *Food Secur.*, pp. 1–22, 2017.

[67] H. Ceballos, C. A. Iglesias, J. C. Pérez, and A. G. O. Dixon, "Cassava breeding: Opportunities and challenges," *Plant Mol. Biol.*, vol. 56, no. 4, pp. 503–516, 2004.

[68] H. Ceballos, J. C. Pérez, O. Joaqui Barandica, J. I. Lenis, N. Morante, F. Calle, L. Pino, and C. H. Hershey, "Cassava Breeding I: The Value of Breeding Value," *Front. Plant Sci.*, vol. 7, no. August, pp. 1–12, 2016.

[69] E. E. Alkan C, Sajjadian S, "Limitation of next generation genome sequence assembly," *Nat Methods*, vol. 8, no. 1, pp. 61–65, 2011.

[70] M. C. Schatz, A. L. Delcher, and S. L. Salzberg, "Assembly of large genomes using second-generation sequencing," *Genome Res.*, vol. 20, no. 9, pp. 1165–1173, 2010.

[71] J. Odipio, T. Alicai, I. Ingelbrecht, D. A. Nusinow, R. Bart, and N. J. Taylor, "Efficient CRISPR/Cas9 Genome Editing of Phytoene desaturase in Cassava," *Front. Plant Sci.*, vol. 8, no. October, pp. 1–11, 2017.

[72] M. A. Gomez, Z. D. Lin, T. Moll, C. Luebbert, R. D. Chauhan, A. Vijayaraghavan, K. Renninger, G. Beyene, N. J. Taylor, J. Carrington, B. Staskawicz, and R. Bart, "Simultaneous CRISPR/Cas9-mediated editing of cassava eIF4E isoforms nCBP-1 and nCBP-2 confers elevated resistance to cassava brown streak disease," 2017.

[73] S. E. Bull, A. Alder, C. Barsan, M. Kohler, L. Hennig, W. Gruissem, and H. Vanderschuren, "FLOWERING LOCUS T Triggers Early and Fertile Flowering in Glasshouse Cassava (Manihot esculenta Crantz)," 2017.

[74] H. Vanderschuren, I. Moreno, R. B. Anjanappa, I. M. Zainuddin, and W. Gruissem, "Exploiting the combination of natural and genetically engineered resistance to cassava mosaic and cassava brown streak viruses impacting cassava production in Africa.," *PLoS One*, vol. 7, no. 9, p. e45277, Jan. 2012.

[75] E. Ogwok, J. Odipio, M. Halsey, E. Gaitán-solís, A. Bua, N. J. Taylor, C. M. Fauquet, and T. Alicai, "Transgenic RNA interference (RNAi) -derived field resistance to cassava brown streak disease," vol. 13, pp. 1019–1031, 2012.

[76] L. Rival and D. McKey, "Domestication and Diversity in Manioc (Manihot

esculenta Crantz ssp . esculenta , Euphorbiaceae),” *Curr. Anthropol.*, vol. 49, no. 6, pp. 1119–1128, 2008.
[77] M. D. Wolfe, P. Kulakow, I. Y. Rabbi, and J.-L. Jannink, “Marker-Based Estimates Reveal Significant Non-additive Effects in Clonally Propagated Cassava (Manihot esculenta): Implications for the Prediction of Total Genetic Value and the Selection of Varieties,” *G3 Genes|Genomes|Genetics*, p. g3.116.033332, 2016.
[78] I. C. G. M. C. (ICGMC), “High-Resolution Linkage Map and Chromosome-Scale Genome Assembly for Cassava (Manihot esculenta Crantz) from Ten Populations,” *G3*, vol. 5, no. 1, pp. 133–144, 2015.
[79] E. Lieberman-aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, and L. A. Mirny, “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome,” vol. 33292, no. October, pp. 289–294, 2009.
[80] S. Grob and U. Grossniklaus, “Chromatin Conformation Capture-Based Analysis of Nuclear Architecture,” in *Plant Epigenetics: Methods and Protocols*, I. Kovalchuk, Ed. Boston, MA: Springer US, 2017, pp. 15–32.
[81] R. K. Slotkin and R. Martienssen, “Transposable elements and the epigenetic regulation of the genome,” *Nat. Rev. Genet.*, vol. 8, no. 4, pp. 272–285, 2007.
[82] R. VanBuren, D. Bryant, P. P. Edger, H. Tang, D. Burgess, D. Challabathula, K. Spittle, R. Hall, J. Gu, E. Lyons, M. Freeling, D. Bartels, B. Ten Hallers, A. Hastie, T. P. Michael, and T. C. Mockler, “Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum.,” *Nature*, vol. 527, no. 7579, pp. 508–11, 2015.
[83] N. Daccord, J.-M. Celton, G. Linsmith, C. Becker, N. Choisne, E. Schijlen, H. van de Geest, L. Bianco, D. Micheletti, R. Velasco, E. A. Di Pierro, J. Gouzy, D. J. G. Rees, P. Guérif, H. Muranty, C.-E. Durel, F. Laurens, Y. Lespinasse, S. Gaillard, S. Aubourg, H. Quesneville, D. Weigel, E. van de Weg, M. Troggio, and E. Bucher, “High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development,” *Nat. Genet.*, no. October 2016, 2017.
[84] A. H. Paterson, J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, J. Schmutz, M. Spannagl, H. Tang, X. Wang, T. Wicker, A. K. Bharti, J. Chapman, F. A. Feltus, U. Gowik, I. V. Grigoriev, E. Lyons, C. A. Maher, M. Martis, A. Narechania, R. P. Otillar, B. W. Penning, A. A. Salamov, Y. Wang, L. Zhang, N. C. Carpita, M. Freeling, A. R. Gingle, C. T. Hash, B. Keller, P. Klein, S. Kresovich, M. C. McCann, R. Ming, D. G. Peterson, Mehboob-Ur-Rahman, D. Ware, P. Westhoff, K. F. X. Mayer, J. Messing, and D. S. Rokhsar, “The Sorghum bicolor genome and the diversification of grasses,” *Nature*, vol. 457, no. 7229, pp. 551–556, 2009.
[85] D. E. Jarvis, Y. S. Ho, D. J. Lightfoot, S. M. Schmöckel, B. Li, T. J. A. Borm, H. Ohyanagi, K. Mineta, C. T. Michell, N. Saber, N. M. Kharbatia, R. R. Rupper, A. R. Sharp, N. Dally, B. A. Boughton, Y. H. Woo, G. Gao, E. G. W. M. Schijlen, X. Guo, A. A. Momin, S. Negrão, S. Al-Babili, C. Gehring, U. Roessner, C. Jung, K. Murphy, S. T. Arold, T. Gojobori, C. G. van der Linden, E. N. van Loo, E. N. Jellen, P. J. Maughan, and M. Tester, “The genome of Chenopodium quinoa,” *Nature*, pp. 1–6, 2017.
[86] F. A. Sima, R. M. Waterhouse, P. Ioannidis, E. V Kriventseva, and E. M. Zdobnov, “Genome analysis BUSCO : assessing genome assembly and

annotation completeness with single-copy orthologs," vol. 31, no. June, pp. 3210–3212, 2015.
[87] M. C. Rojas, J. C. Pérez, H. Ceballos, D. Baena, N. Morante, and F. Calle, "Analysis of inbreeding depression in eight S1 cassava families," *Crop Sci.*, vol. 49, no. 2, pp. 543–548, 2009.
[88] M. Sémon and K. H. Wolfe, "Consequences of genome duplication," *Curr. Opin. Genet. Dev.*, vol. 17, no. 6, pp. 505–512, 2007.
[89] L. Li, C. J. J. Stoeckert, and D. S. Roos, "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes -- Li et al. 13 (9): 2178 -- Genome Research," *Genome Res.*, vol. 13, no. 9, pp. 2178–2189, 2003.
[90] P. Schläpfer, P. Zhang, C. Wang, T. Kim, M. Banf, L. Chae, K. Dreher, A. K. Chavali, R. Nilo-Poyanco, T. Bernard, D. Kahn, and S. Y. Rhee, "Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants," *Plant Physiol.*, vol. 173, no. 4, pp. 2041–2059, 2017.
[91] Gene Ontology Consortium, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Res.*, vol. 32, no. 90001, p. 258D–261, 2004.
[92] L. Kong, S. Li, Q. Liao, Y. Zhang, R. Sun, X. Zhu, Q. Zhang, J. Wang, X. Wu, X. Fang, and Y. Zhu, "Oleanolic acid and ursolic acid: Novel hepatitis C virus antivirals that inhibit NS5B activity," *Antiviral Res.*, vol. 98, no. 1, pp. 44–53, 2013.
[93] P. Horvath, R. Barrangou, P. Horvath, and R. Barrangou, "CRISPR/Cas, the Immune System of Bacteria and Archaea," *Source Sci. New Ser.*, vol. 327, no. 5962, pp. 167–170, 2010.
[94] J. J. Doyle and J. L. Doyle, "A rapid total DNA preparation procedure for fresh plant tissue," *Focus (Madison).*, vol. 12, pp. 13–15, 1990.
[95] K. Berlin, S. Koren, C.-S. Chin, J. P. Drake, J. M. Landolin, and A. M. Phillippy, "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing," *Nat. Biotechnol.*, vol. 33, no. 6, 2015.
[96] A. M. Bolger, M. Lohse, and B. Usadel, "Genome analysis Trimmomatic : a flexible trimmer for Illumina sequence data," vol. 30, no. 15, pp. 2114–2120, 2014.
[97] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows – Wheeler transform," vol. 25, no. 14, pp. 1754–1760, 2009.
[98] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, G. P. Data, and T. Sam, "The Sequence Alignment / Map format and SAMtools," vol. 25, no. 16, pp. 2078–2079, 2009.
[99] K. Okonechnikov, A. Conesa, and F. Garcı, "Genome analysis Qualimap 2 : advanced multi-sample quality control for high-throughput sequencing data," vol. 32, no. October 2015, pp. 292–294, 2016.
[100] T. D. Wu and C. K. Watanabe, "Sequence analysis GMAP : a genomic mapping and alignment program for mRNA and EST sequences," vol. 21, no. 9, pp. 1859–1875, 2005.
[101] M. Stanke, M. Diekhans, R. Baertsch, and D. Haussler, "Sequence analysis Using native and syntenically mapped cDNA alignments to improve de novo gene finding," vol. 24, no. 5, pp. 637–644, 2008.
[102] A. Conesa, S. Götz, J. M. García-gómez, J. Terol, M. Talón, D. Genómica, I. Valenciano, D. I. Agrarias, and U. P. De Valencia, "Blast2GO : a universal tool for annotation , visualization and analysis in functional genomics research," vol. 21, no. 18, pp. 3674–3676, 2005.
[103] T. M. Lowe and P. P. Chan, "tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes," *Nucleic Acids Res.*, vol. 44, no.

W1, pp. W54–W57, 2016.
[104] E. P. Nawrocki and S. R. Eddy, “Infernal 1.1: 100-fold faster RNA homology searches,” *Bioinformatics*, vol. 29, no. 22, pp. 2933–2935, 2013.
[105] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, “STAR: Ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.
[106] R. C. Edgar, “MUSCLE: Multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004.
[107] C. Škuta, P. Bartu˚něk, and D. Svozil, “InCHlib - Interactive cluster heatmap for web applications,” *J. Cheminform.*, vol. 6, no. 1, pp. 1–9, 2014.
[108] Y. Yuan, P. E. Bayer, H. Lee, and D. Edwards, “Sequence analysis runBNG : a software package for BioNano genomic analysis on the command line,” no. June, pp. 1–3, 2017.
[109] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool.,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–10, 1990.
[110] R. Gentleman, V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Yang, and J. Zhang, “Bioconductor: open software development for computational biology and bioinformatics.,” *Genome Biol.*, vol. 5, no. 10, p. R80, 2004.
[111] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, “Revigo summarizes and visualizes long lists of gene ontology terms,” *PLoS One*, vol. 6, no. 7, 2011.
[112] S. E. Bull, J. A. Owiti, M. Niklaus, J. R. Beeching, W. Gruissem, and H. Vanderschuren, “Agrobacterium -mediated transformation of friable embryogenic calli and regeneration of transgenic cassava,” *Nat. Protoc.*, vol. 4, no. 2, pp. 1845–1854, 2009.
[113] N. J. Taylor, M. V. Masona, R. Carcamo, T. Ho, C. Schöpke, and C. M. Fauquet, “Production of embryogenic tissues and regeneration of transgenic plants in cassava (Manihot esculenta Crantz),” *Euphytica*, vol. 120, no. 1, pp. 25–34, 2001.
[114] S. Wingett, P. Ewels, M. Furlan-magaril, T. Nagano, S. Schoenfelder, P. Fraser, and S. Andrews, “HiCUP : pipeline for mapping and processing Hi-C data [version 1 ; referees : 2 approved , 1 approved with reservations] Referee Status :,” vol. 1310, pp. 1–12, 2015.
[115] N. H. Putnam, B. O. Connell, J. C. Stites, B. J. Rice, P. D. Hartley, C. W. Sugnet, D. Haussler, and D. S. Rokhsar, “Chromosome-scale shotgun assembly using an in vitro method for long-range linkage arXiv : 1502 . 05331v1 [q-bio . GN] 18 Feb 2015,” pp. 1–25, 2016.
[116] S. Chang, J. Puryear, and J. Cairney, “A simple and efficient method for isolating RNA from pine trees,” *Plant Mol. Biol. Report.*, vol. 11, no. 2, pp. 113–116, 1993.
[117] A. F. . Smit and R. Hubley, “RepeatModeler Open-1.0.,” *http://www.repeatmasker.org*.
[118] A. F. A. Smit, R. Hubley, and P. Green, “RepeatMasker Open-4.0.,” *http://www.repeatmasker.org*.
[119] R. Apweiler, “UniProt: the Universal Protein knowledgebase,” *Nucleic Acids Res.*, vol. 32, no. 90001, p. 115D–119, 2004.
[120] W. Jiao and K. Schneeberger, “ScienceDirect The impact of third generation genomic technologies on plant genome assembly,” *Curr. Opin. Plant Biol.*, vol. 36, pp. 64–70, 2017.

[121] T. J. Carver, K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell, and J. Parkhill, "ACT: The Artemis comparison tool," *Bioinformatics*, vol. 21, no. 16, pp. 3422–3423, 2005.
[122] K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak, "VISTA: Computational tools for comparative genomics," *Nucleic Acids Res.*, vol. 32, no. WEB SERVER ISS., pp. 273–279, 2004.
[123] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna, "Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements," pp. 1394–1403, 2004.
[124] T. Carver, N. Thomson, A. Bleasby, M. Berriman, and J. Parkhill, "DNAPlotter: Circular and linear interactive genome visualization," *Bioinformatics*, vol. 25, no. 1, pp. 119–120, 2009.
[125] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, "Versatile and open software for comparing large genomes.," *Genome Biol.*, vol. 5, no. 2, p. R12, 2004.
[126] I. Rabbi, M. Hamblin, M. Gedil, P. Kulakow, M. Ferguson, A. S. Ikpan, D. Ly, and J.-L. Jannink, "Genetic Mapping Using Genotyping-by-Sequencing in the Clonally Propagated Cassava," *Crop Sci.*, vol. 54, no. 4, p. 1384, 2014.
[127] A. Minio, J. Lin, B. S. Gaut, and D. Cantu, "How Single Molecule Real-Time Sequencing and Haplotype Phasing Have Enabled Reference-Grade Diploid Genome Assembly of Wine Grapes," *Front. Plant Sci.*, vol. 8, no. May, pp. 1–6, 2017.
[128] I. Cassava and G. Map, "High-Resolution Linkage Map and Chromosome-Scale Genome Assembly for Cassava (Manihot esculenta Crantz) from 10 Populations.," *G3 (Bethesda).*, vol. 5, no. 1, pp. 133–44, Jan. 2014.
[129] P. Yang, T. Lüpken, A. Habekuss, G. Hensel, B. Steuernagel, B. Kilian, R. Ariyadasa, A. Himmelbach, J. Kumlehn, U. Scholz, F. Ordon, and N. Stein, "PROTEIN DISULFIDE ISOMERASE LIKE 5-1 is a susceptibility factor to plant viruses.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 6, pp. 2104–9, Feb. 2014.
[130] J. Gilbert, W. Ou, J. Silver, and T. Benjamin, "Downregulation of Protein Disulfide Isomerase Inhibits Infection by the Mouse Polyomavirus," *J. Virol.*, vol. 80, no. 21, pp. 10868–10870, 2006.
[131] S. Parakh and J. D. Atkin, "Novel roles for protein disulphide isomerase in disease states: a double edged sword?," *Front. Cell Dev. Biol.*, vol. 3, no. May, pp. 1–11, 2015.
[132] F. Li, Y. Wang, and X. Zhou, "SGS3 cooperates with RDR6 in triggering geminivirus-induced gene silencing and in suppressing geminivirus infection in Nicotiana Benthamiana," *Viruses*, vol. 9, no. 9, 2017.
[133] E. M. Maine, "A conserved mechanism for post-transcriptional gene silencing?," *Genome Biol.*, vol. 1, no. 3, p. REVIEWS1018, 2000.
[134] Y. Stram and L. Kuzntzova, "Inhibition of viruses by RNA interference," *Virus Genes*, vol. 32, no. 3, pp. 299–306, 2006.
[135] A. Eamens, M.-B. Wang, N. A. Smith, and P. M. Waterhouse, "RNA Silencing in Plants: Yesterday, Today, and Tomorrow," *Plant Physiol.*, vol. 147, no. 2, pp. 456–468, 2008.
[136] P. Mourrain, C. Béclin, T. Elmayan, F. Feuerbach, C. Godon, J.-B. Morel, D. Jouette, A.-M. Lacombe, S. Nikic, N. Picault, K. Rémoué, M. Sanial, T.-A. Vo, and H. Vaucheret, "Arabidopsis SGS2 and SGS3 Genes Are Required for Posttranscriptional Gene Silencing and Natural Virus Resistance," *Cell*, vol. 101, no. 5, pp. 533–542, 2000.

[137] S. E. Carter, L. O. Fresco, P. G. Jones, and J. N. Fairbairn, “An atlas of cassava in Africa: historical, agroecological and demographic aspects of crop distribution,” *CIAT publication ; no. 206 (CIAT)*. 1992.
[138] S. K. Hahn, E. R. Terry, and K. Leuschner, “Breeding cassava for resistance to cassava mosaic disease,” *Euphytica*, vol. 29, no. 3, pp. 673–683, 1980.
[139] C. Rey, “Cassava Mosaic and Brown Streak Diseases : Current Perspectives and Beyond,” 2017.
[140] J. P. Legg, P. L. Kumar, T. Makeshkumar, L. Tripathi, M. Ferguson, E. Kanju, P. Ntawuruhunga, and W. Cuellar, *Cassava Virus Diseases : Biology , Epidemiology , and Management*, 1st ed. Elsevier Inc., 2014.
[141] C. Gutierrez, “Geminiviruses and the plant cell cycle,” pp. 763–772, 2000.
[142] M. a García-Neria and R. F. Rivera-Bustamante, “Characterization of Geminivirus resistance in an accession of Capsicum chinense Jacq.,” *Mol. Plant. Microbe. Interact.*, vol. 24, no. 2, pp. 172–182, 2011.
[143] R. Z. Naqvi, S. S. E. A. Zaidi, K. P. Akhtar, S. Strickler, M. Woldemariam, B. Mishra, M. Shahid Mukhtar, B. E. Scheffler, J. A. Scheffler, G. Jander, L. A. Mueller, M. Asif, and S. Mansoor, “Transcriptomics reveals multiple resistance mechanisms against cotton leaf curl disease in a naturally immune cotton species, Gossypium arboreum,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–15, 2017.
[144] Y. S. Seo, P. Gepts, and R. L. Gilbertson, “Genetics of resistance to the geminivirus, Bean dwarf mosaic virus, and the role of the hypersensitive response in common bean,” *Theor. Appl. Genet.*, vol. 108, no. 5, pp. 786–793, 2004.
[145] M. Ong-Abdullah, J. M. Ordway, N. Jiang, S.-E. Ooi, S.-Y. Kok, N. Sarpan, N. Azimi, A. T. Hashim, Z. Ishak, S. K. Rosli, F. A. Malike, N. A. A. Bakar, M. Marjuni, N. Abdullah, Z. Yaakub, M. D. Amiruddin, R. Nookiah, R. Singh, E.-T. L. Low, K.-L. Chan, N. Azizi, S. W. Smith, B. Bacher, M. a. Budiman, A. Van Brunt, C. Wischmeyer, M. Beil, M. Hogan, N. Lakey, C.-C. Lim, X. Arulandoo, C.-K. Wong, C.-N. Choo, W.-C. Wong, Y.-Y. Kwan, S. S. R. S. Alwee, R. Sambanthamurthi, and R. a. Martienssen, “Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm,” *Nature*, no. 11, 2015.
[146] H. Stroud, B. Ding, S. a. Simon, S. Feng, M. Bellizzi, M. Pellegrini, G. L. Wang, B. C. Meyers, and S. E. Jacobsen, “Plants regenerated from tissue culture contain stable epigenome changes in rice,” *Elife*, vol. 2013, no. 2, pp. 1–14, 2013.
[147] H. Vanderschuren, R. Akbergenov, M. M. Pooggin, T. Hohn, W. Gruissem, and P. Zhang, “Transgenic cassava resistance to African cassava mosaic virus is enhanced by viral DNA-A bidirectional promoter-derived siRNAs,” *Plant Mol. Biol.*, vol. 64, no. 5, pp. 549–557, 2007.
[148] D. Mehta, M. Hirsch-Hoffmann, A. Patrignani, W. Gruissem, and H. Vanderschuren, “CIDER-Seq: unbiased virus enrichment and single-read, full length genome sequencing,” *bioRxiv*, p. 168724, 2017.
[149] M. Fregene, A. Bernal, M. Duque, and A. Dixon, “AFLP analysis of African cassava (Manihot esculenta Crantz) germplasm resistant to the cassava mosaic disease (CMD),” pp. 678–685, 2000.
[150] P. M. Waterhouse, M. Wang, and T. Lough, “Gene silencing as an adaptive defence against viruses,” *Nature*, vol. 411, pp. 834–842, 2001.
[151] F. G. Ratcliff, “Gene Silencing without DNA: RNA-Mediated Cross-Protection between Viruses,” *Plant Cell Online*, vol. 11, no. 7, pp. 1207–1216, 1999.
[152] A. J. Hamilton and D. Baulcombe, “A species of small antisense RNA in

posttranscriptional gene silencing in plants," *Sci. (New York, NY)*, vol. 286, no. 5441, pp. 950–952, 1999.
[153] J. Schuck, T. Gursinsky, V. Pantaleo, J. Burgyán, and S. E. Behrens, "AGO/RISC-mediated antiviral RNA silencing in a plant in vitro system," *Nucleic Acids Res.*, vol. 41, no. 9, pp. 5090–5103, 2013.
[154] E. Liu and J. E. Page, "Optimized cDNA libraries for virus-induced gene silencing (VIGS) using tobacco rattle virus," *Plant Methods*, vol. 4, no. 1, pp. 1–13, 2008.
[155] B. Zhou and L. Zeng, "Elucidating the role of highly homologous Nicotiana benthamiana ubiquitin E2 gene family members in plant immunity through an improved virus-induced gene silencing approach," *Plant Methods*, vol. 13, no. 1, pp. 1–17, 2017.
[156] J.-B. Hiriart, E.-M. Aro, and K. Lehto, "Dynamics of the VIGS-mediated chimeric silencing of the Nicotiana benthamiana ChlH gene and of the tobacco mosaic virus vector.," *Mol. Plant. Microbe. Interact.*, vol. 16, no. 2, pp. 99–106, 2003.
[157] J. M. Berg and Y. Shi, "The Galvanization of Biology: A Growing Appreciation for the Roles of Zinc," *Science (80-.).*, vol. 271, no. 5252, p. 1081 LP-1085, Feb. 1996.
[158] A. G. Von Arnim and X. W. Deng, "Ring finger motif of Arabidopsis thaliana COP1 defines a new class of zinc-binding domain," *J. Biol. Chem.*, vol. 268, no. 26, pp. 19626–19631, 1993.
[159] A. E. Pepper and J. Chory, "Extragenic Suppressors of the Arabidopsis det1 Mutant Identify Elements of Flowering-Time and Light-Response Regulatory Pathways," *Genetics*, vol. 145, no. 4, p. 1125 LP-1137, Apr. 1997.
[160] N. Matsuda, T. Suzuki, K. Tanaka, and a Nakano, "Rma1, a novel type of RING finger protein conserved from Arabidopsis to human, is a membrane-bound ubiquitin ligase.," *J. Cell Sci.*, vol. 114, no. Pt 10, pp. 1949–57, 2001.
[161] U. Schumann, G. Wanner, M. Veenhuis, M. Schmid, and C. Gietl, "AthPEX10, a nuclear gene essential for peroxisome and storage organelle formation during Arabidopsis embryogenesis," *Proc. Natl. Acad. Sci.*, vol. 100, no. 16, pp. 9626–9631, 2003.
[162] Y.-S. Wang, L.-Y. Pi, X. Chen, P. K. Chakrabarty, J. Jiang, A. L. De Leon, G.-Z. Liu, L. Li, U. Benny, J. Oard, P. C. Ronald, and W.-Y. Song, "Rice XA21 Binding Protein 3 Is a Ubiquitin Ligase Required for Full Xa21-Mediated Disease Resistance," *Plant Cell Online*, vol. 18, no. 12, pp. 3635–3646, 2006.
[163] D. Komander and M. Rape, "The Ubiquitin Code," *Annu. Rev. Biochem.*, vol. 81, no. 1, pp. 203–229, 2012.
[164] L. Almagro, L. V. Gómez Ros, S. Belchi-Navarro, R. Bru, A. Ros Barceló, and M. A. Pedreño, "Class III peroxidases in plant defence reactions," *J. Exp. Bot.*, vol. 60, no. 2, pp. 377–390, 2009.
[165] H. Dieng, T. Satho, A. A. Hassan, A. T. Aziz, R. E. Morales, S. A. Hamid, F. Miake, and S. Abubakar, "Peroxidase Activity after Viral Infection and Whitefly Infestation in Juvenile and Mature Leaves of Solanum lycopersicum," *J. Phytopathol.*, vol. 159, no. 11–12, pp. 707–712, 2011.
[166] A. Gallina, T. M. Hanley, R. Mandel, M. Trahey, C. C. Broder, G. A. Viglianti, and H. J. P. Ryser, "Inhibitors of protein-disulfide isomerase prevent cleavage of disulfide bonds in receptor-bound glycoprotein 120 and prevent HIV-1 entry," *J. Biol. Chem.*, vol. 277, no. 52, pp. 50579–50588, 2002.

[167] J. Verchot, “Plant virus infection and the ubiquitin proteasome machinery: Arms race along the endoplasmic reticulum,” *Viruses*, vol. 8, no. 11, 2016.
[168] T. Nakamura and S. A. Lipton, “S-Nitrosylation of Critical Protein Thiols Mediates Protein Misfolding and Mitochondrial Dysfunction in Neurodegenerative Diseases,” *Antioxid. Redox Signal.*, vol. 14, no. 8, pp. 1479–1492, 2011.
[169] C. Muller, J. Bandemer, C. Vindis, C. Camaré, E. Mucher, F. Guéraud, P. Larroque-Cardoso, C. Bernis, N. Auge, R. Salvayre, and A. Negre-Salvayre, “Protein Disulfide Isomerase Modification and Inhibition Contribute to ER Stress and Apoptosis Induced by Oxidized Low Density Lipoproteins,” *Antioxid. Redox Signal.*, vol. 18, no. 7, pp. 731–742, 2013.
[170] T. S. Sarkar, U. Majumdar, A. Roy, D. Maiti, A. M. Goswamy, A. Bhattacharjee, S. K. Ghosh, and S. Ghosh, “Production of nitric oxide in host-virus interaction: A case study with a compatible begomovirus-kenaf host-pathosystem,” *Plant Signal. Behav.*, vol. 5, no. 6, pp. 668–676, 2010.
[171] M. Pooggin, P. V. Shivaprasad, K. Veluthambi, and T. Hohn, “RNAi targeting of DNA virus in plants,” *Nat. Biotechnol.*, vol. 21, no. 2, pp. 131–132, 2003.
[172] S. E. Bull, J. Ndunguru, W. Gruissem, J. R. Beeching, and H. Vanderschuren, “Cassava: Constraints to production and the transfer of biotechnology to African laboratories,” *Plant Cell Rep.*, vol. 30, no. 5, pp. 779–787, 2011.
[173] F. J. L. Arago and J. C. Faria, “First transgenic geminivirus-resistant plant in the field,” *Nat. Biotechnol.*, vol. 27, no. 12, pp. 1086–1088, 2009.
[174] A. Fuentes, N. Carlos, Y. Ruiz, D. Callard, Y. Sánchez, M. E. Ochagavía, J. Seguin, N. Malpica-López, T. Hohn, M. R. Lecca, R. Pérez, V. Doreste, H. Rehrauer, L. Farinelli, M. Pujol, and M. M. Pooggin, “Field Trial and Molecular Characterization of RNAi-Transgenic Tomato Plants That Exhibit Resistance to Tomato Yellow Leaf Curl Geminivirus,” *Mol. Plant-Microbe Interact.*, vol. 29, no. 3, pp. 197–209, 2016.
[175] Y. H. Li, G. Zhou, J. Ma, W. Jiang, L. G. Jin, Z. Zhang, Y. Guo, J. Zhang, Y. Sui, L. Zheng, S. S. Zhang, Q. Zuo, X. H. Shi, Y. F. Li, W. K. Zhang, Y. Hu, G. Kong, H. L. Hong, B. Tan, J. Song, Z. X. Liu, Y. Wang, H. Ruan, C. K. L. Yeung, J. Liu, H. Wang, L. J. Zhang, R. X. Guan, K. J. Wang, W. Bin Li, S. Y. Chen, R. Z. Chang, Z. Jiang, S. A. Jackson, R. Li, and L. J. Qiu, “De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits,” *Nat. Biotechnol.*, vol. 32, no. 10, pp. 1045–1052, 2014.
[176] F. Lu, M. C. Romay, J. C. Glaubitz, P. J. Bradbury, R. J. Elshire, T. Wang, Y. Li, Y. Li, K. Semagn, X. Zhang, A. G. Hernandez, M. A. Mikel, I. Soifer, O. Barad, and E. S. Buckler, “High-resolution genetic mapping of maize pan-genome sequence anchors,” *Nat. Commun.*, vol. 6, 2015.
[177] M. H. Schmidt, A. Vogel, A. K. Denton, B. Istace, A. Wormit, H. van de Geest, M. E. Bolger, S. Alseekh, J. Maß, C. Pfaff, U. Schurr, R. T. Chetelat, F. Maumus, J.-M. Aury, S. Koren, A. R. Fernie, D. Zamir, A. Bolger, and B. Usadel, “De novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing,” *Plant Cell*, p. tpc.00521.2017, 2017.
[178] N. R. Hofmann, “Nanopore Sequencing Comes to Plant Genomes,” *Plant Cell*, vol. 29, no. November, p. tpc.00863.2017, 2017.
[179] J. F. Flot, H. Marie-Nelly, and R. Koszul, “Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures,” *FEBS Lett.*, vol. 589, no. 20, pp. 2966–2974, 2015.
[180] S. Reyes-Chin-Wo, Z. Wang, X. Yang, A. Kozik, S. Arikit, C. Song, L. Xia, L. Froenicke, D. O. Lavelle, M.-J. Truco, R. Xia, S. Zhu, C. Xu, H. Xu, X.

Xu, K. Cox, I. Korf, B. C. Meyers, and R. W. Michelmore, “Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce,” *Nat. Commun.*, vol. 8, p. 14953, 2017.
[181] H. Marie-Nelly, M. Marbouty, A. Cournac, J.-F. Flot, G. Liti, D. P. Parodi, S. Syan, N. Guillén, A. Margeot, C. Zimmer, and R. Koszul, “High-quality genome (re)assembly using chromosomal contact data,” *Nat. Commun.*, vol. 5, p. 5695, 2014.
[182] F. J. Sedlazeck, H. Lee, C. A. Darby, and M. C. Schatz, “Piercing the dark matter: bioinformatics of long-range sequencing and mapping,” *Nat. Rev. Genet.*, 2018.
[183] A. C. English, S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin, D. M. Muzny, J. G. Reid, K. C. Worley, and R. a. Gibbs, “Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology,” *PLoS One*, vol. 7, no. 11, pp. 1–12, 2012.
[184] S. Kosugi, H. Hirakawa, and S. Tabata, “GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments.,” *Bioinformatics*, no. August, p. btv465-, 2015.
[185] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl, “Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement,” *PLoS One*, vol. 9, no. 11, 2014.
[186] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. DeWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, “Real-time DNA sequencing from single polymerase molecules,” *Science (80-.).*, vol. 323, no. 5910, pp. 133–138, 2009.
[187] T. Hackl, R. Hedrich, J. Schultz, and F. Forster, “proovread: large-scale high-accuracy PacBio correction through iterative short read consensus,” *Bioinformatics*, vol. 30, no. 21, pp. 3004–3011, 2014.
[188] L. Briñas, C. Orvain, C. Belser, C. Cruaud, K. Labadie, L. Bertrand, V. Barbe, J.-M. Aury, P. Wincker, and A. Alberti, “BAC ends library generation for Illumina sequencing ,” Jun. 2015.
[189] J. L. Peters, F. Cnudde, and T. Gerats, “Forward genetics and map-based cloning approaches,” *Trends Plant Sci.*, vol. 8, no. 10, pp. 484–491, 2003.
[190] A. Watson, S. Ghosh, M. J. Williams, W. S. Cuddy, J. Simmonds, M. D. Rey, M. Asyraf Md Hatta, A. Hinchliffe, A. Steed, D. Reynolds, N. M. Adamski, A. Breakspear, A. Korolev, T. Rayner, L. E. Dixon, A. Riaz, W. Martin, M. Ryan, D. Edwards, J. Batley, H. Raman, J. Carter, C. Rogers, C. Domoney, G. Moore, W. Harwood, P. Nicholson, M. J. Dieters, I. H. Delacy, J. Zhou, C. Uauy, S. A. Boden, R. F. Park, B. B. H. Wulff, and L. T. Hickey, “Speed breeding is a powerful tool to accelerate crop research and breeding,” *Nat. Plants*, vol. 4, no. 1, pp. 23–29, 2018.
[191] K. Schneeberger, S. Ossowski, C. Lanz, T. Juul, A. H. Petersen, K. L. Nielsen, J. E. Jørgensen, D. Weigel, and S. U. Andersen, “SHOREmap: Simultaneous mapping and mutation identification by deep sequencing,” *Nat. Methods*, vol. 6, no. 8, pp. 550–551, 2009.
[192] H. Takagi, A. Abe, K. Yoshida, S. Kosugi, S. Natsume, C. Mitsuoka, A.

Uemura, H. Utsushi, M. Tamiru, S. Takuno, H. Innan, L. M. Cano, S. Kamoun, and R. Terauchi, “QTL-seq: Rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations,” *Plant J.*, vol. 74, no. 1, pp. 174–183, 2013.

[193] X. Yang, X. Xia, Z. Zhang, B. Nong, Y. Zeng, F. Xiong, Y. Wu, J. Gao, G. Deng, and D. Li, “QTL Mapping by Whole Genome Re-sequencing and Analysis of Candidate Genes for Nitrogen Use Efficiency in Rice,” *Front. Plant Sci.*, vol. 8, no. September, pp. 1–10, 2017.

[194] M. Mascher, M. Jost, J. E. Kuon, A. Himmelbach, A. Abfalg, S. Beier, U. Scholz, A. Graner, and N. Stein, “Mapping-by-sequencing accelerates forward genetics in barley,” *Genome Biol.*, vol. 15, no. 6, pp. 1–15, 2014.

[195] M. Choi, U. I. Scholl, W. Ji, T. Liu, I. R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Ozen, S. Sanjad, C. Nelson-Williams, A. Farhi, S. Mane, and R. P. Lifton, “Genetic diagnosis by whole exome capture and massively parallel DNA sequencing,” *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 19096–19101, 2009.

[196] M. Mascher, T. A. Richmond, D. J. Gerhardt, A. Himmelbach, L. Clissold, D. Sampath, S. Ayling, B. Steuernagel, M. Pfeifer, M. D’Ascenzo, E. D. Akhunov, P. E. Hedley, A. M. Gonzales, P. L. Morrell, B. Kilian, F. R. Blattner, U. Scholz, K. F. X. Mayer, A. J. Flavell, G. J. Muehlbauer, R. Waugh, J. A. Jeddeloh, and N. Stein, “Barley whole exome capture: A tool for genomic research in the genus Hordeum and beyond,” *Plant J.*, vol. 76, no. 3, pp. 494–505, 2013.

[197] I. M. Henry, U. Nagalakshmi, M. C. Lieberman, K. J. Ngo, K. V. Krasileva, H. Vasquez-Gross, A. Akhunova, E. Akhunov, J. Dubcovsky, T. H. Tai, and L. Comai, “Efficient Genome-Wide Detection and Cataloging of EMS-Induced Mutations Using Exome Capture and Next-Generation Sequencing,” *Plant Cell*, vol. 26, no. 4, pp. 1382–1397, 2014.

[198] M. Mascher, M. Jost, J.-E. Kuon, A. Himmelbach, A. Aßfalg, S. Beier, U. Scholz, A. Graner, and N. Stein, “Mapping-by-sequencing accelerates forward genetics in barley.,” *Genome Biol.*, vol. 15, no. 6, p. R78, 2014.

[199] B. Steuernagel, S. K. Periyannan, I. Hernández-Pinzón, K. Witek, M. N. Rouse, G. Yu, A. Hatta, M. Ayliffe, H. Bariana, J. D. G. Jones, E. S. Lagudah, and B. B. H. Wulff, “Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture,” *Nat. Biotechnol.*, vol. 34, no. 6, pp. 652–655, 2016.

[200] J. Sánchez-Martín, B. Steuernagel, S. Ghosh, G. Herren, S. Hurni, N. Adamski, J. Vrána, M. Kubaláková, S. G. Krattinger, T. Wicker, J. Doležel, B. Keller, and B. B. H. Wulff, “Rapid gene isolation in barley and wheat by mutant chromosome sequencing,” *Genome Biol.*, vol. 17, no. 1, pp. 1–7, 2016.

[201] L. J. C. B. Carvalho, E. A. Vieira, J. de F. Fialho, and C. R. B. de Souza, “A genomic assisted breeding program for cassava to improve nutritional quality and industrial traits of storage root,” *Crop Breed. Appl. Biotechnol.*, vol. 11, no. 4, pp. 289–296, 2011.

Acknowledgements

The work presented in this PhD thesis could only be completed because of the support of very many people. The nice atmosphere in and outside lab was key to reach goals that were set at the beginning of this thesis. In this respect, I want emphasize a special thanks to my colleagues Ravi Bodampalli, Devang Metha, Wilfred Elegba, Simon Bull, Simrat Pal Singh, Ting-Ying Wu, Ima Zainuddin, Kumar Vasudevan, Kulaporn Bunyaves, Tiago Dias Cruz, Pascal Schläpfer and Sebastian Petersen for their support. Without you this work wouldn't have been that enjoyable.

I want to give a special thanks to Prof. Hervé Vanderschuren for accepting me as a PhD student. Hervé's guidance as a mentor and supervisor in the early phase of my PhD was key for the success of this thesis.

I thank Prof. Willi Gruissem for the opportunity to spend the last four years in the Plant Biotechnology Lab at ETH Zurich. I'm grateful that Willi gave me the opportunity and belief to develop novel approaches that helped my science to grow.

I owe a special thanks to the people from the Functional Genomic Center Zurich (FGCZ). I'm special grateful to Weihong Qi for supporting the various genome projects and for answering every tiny question whenever I got lost in the jungle of sequencing data. I want to thank Lucy Poveda, Catharine Aquino, Andrea Patrignani, Anna Bratus-Neuenschwander for their tireless support that created the foundation on which I could build upon.

I thank Matthias Hirsch-Hoffmann for his patience and goodwill to show me the secrets of command-line based LINUX operation and his way in managing and organizing big data.

I thank Prof. Sánchez-Rodríguez and Prof. Beat Keller for accepting to be the co-referee of my PhD thesis.

I want to thank the Thursday-Team for interesting scientific and non-scientific discussions as well as my *bianchi infinito cv* to keep my spirit high.

The SNF 'SAVUCA' and the Bill & Melinda Gates Foundation are acknowledged for their financial support to the project.

Last but not least, I want to thank Mattea for her support and patience during these years.

CURRICULUM VITAE

Personal Details

Joel-Elias Kuon
Department of Biology, ETH Zurich
LFW E14, Universitätstrasse 2, 8002 Zurich, Switzerland
Nationality: German
Date of Birth: 08.06.1987
Email: kuonj@ethz.ch / joel.kuon@gmail.com

Education

Dr. Sc. (ETH): 2018	Plant Biotechnology Certificate in Science & Policy ETH Zurich, Switzerland
Master of Science: 2011-2013	Agricultural Biotechnology University of Hohenheim, Germany IPK Gatersleben, Germany
Bachelor of Science: 2008-2011	Agricultural Biology University of Hohenheim, Germany

Research & Work Experience

PhD Thesis: ETH Zurich Jan. 2014-2018 Prof. W. Gruissem Prof. H. Vanderschuren	*Reconstructing cassava genomes with single-molecule technologies and chromosome conformation mapping to investigate geminivirus resistance by reverse genetics tools* Whole genome sequencing and assembly of two high-value cassava genomes using SMRT sequencing and single-molecule mapping, software development for diploid-aware QTL visualization, Development of a high-throughput gene discovery platform for virus resistance gene discovery.
Master's Thesis: IPK Gatersleben (GED lab) Mar.2012-Dec.2013 Dr. N. Stein Prof. A. Graner Prof. K. Schmid	*Identification of a many-noded-dwarf gene by using mapping-by-sequencing in barley* Phenotyping of a segregating mapping population, construction of a genetic map using Exome-capture next-generation-sequencing of phenotypically pooled plants, Identification of candidate genes through re-sequencing and confirmation of a candidate gene through re-sequencing independent alleles.
Summer Internships: R&D Selecta&Sohn Ornamental Plant Breeding Aug.2012-Oct.2012 Aug.2011-Oct.2011	*Implementation and development of breeding-schemes for novel flowering species and investigation of a genotype-depended durability of flowering behaviour in Dianthus species* Hands-on experience in plant breeding and plant biotechnology, involved in the process of development and release of new varieties at one of the world's leading ornamental plant breeding company.

Volunteerism:

Jul.2010 – Sept.2010	Vegetable cultivation at 'Hofgut Rengoldshausen', Bodensee, Germany
Jan.2008 – Apr.2008	Social-ecological Internship: Fundación Centro de Capacitación. Fernandez, Argentina

Student Supervision

Master Thesis	Jenny Brown	MSc. Biology, ETH Zurich
	Marius Rohner	MSc. Biology, ETH Zurich
Semester students	Philipp Rogalla von Bieberstein	

Academic Publications (peer- reviewed)

1. Mascher, M., Jost, M., **Kuon, J. E**., Himmelbach, A., Aßfalg, A., Beier, S., ... & Stein, N. (2014). Mapping-by-sequencing accelerates forward genetics in barley. *Genome biology*, *15*(6), R78.

Conference Talks & Presentations

1. J. Kuon, W. Qi, W. Gruissem and H. Vanderschuren. Decoding complex cassava genomes using single-molecule technologies. Poster: PSC-Syngenta Symposium 2017, Basel, Switzerland 30 August 2017
2. J. Kuon, W. Qi, W. Gruissem and H. Vanderschuren. Chromosome-level assembly of farmer preferred cassava varieties using single-molecule sequencing (SMRT) technology and chromosome-conformation capture mapping. Poster: Plant and Animal Genome XXV Conference. San Diego, US. 13-18 January 2017
3. J. Kuon, W. Gruissem and H. Vanderschuren. Exploiting genetic resources: How reverse/forward genetics and genomics facilitate precision plant breeding. Selecta Internal Symposium. Stuttgart, Germany. 7 December 2016
4. J. Kuon, W. Gruissem and H. Vanderschuren. The search for the monogenic natural resistance to the cassava mosaic disease. DPG-Plant Virology Symposium, Hannover, Germany. 7-8 March, 2016
5. J. Kuon, E. Lentz, W. Gruissem and H. Vanderschuren. Assessing *CMD2* geminivirus resistance genes through agro-bacterium based virus induced gene silencing (VIGS) in cassava. International Cassava Conference, Guangxi, China. 19 January 2016
6. J. Kuon and H. Vanderschuren. Mapping-by-sequencing for trait discovery in complex plant genomes. Cassava molecular breeding workshop, Ghent, Belgium. 8 September 2014
7. J. Kuon and H. Vanderschuren. Next-generation cassava with enhanced agronomic and industrial performances for southern Africa. Poster: ETH D-BIOL Symposium, Davos, Switzerland, June 2014