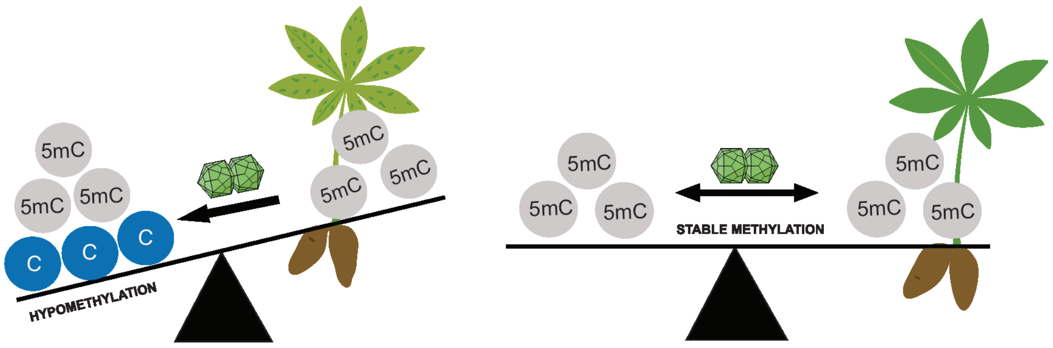
Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2025:61
Faculty of Veterinary Medicine and Animal Science

Identification of epigenetics variations influencing viral resistance in cassava farmers' field

Linking gDNA methylation profiles to phenotype

Michael Kofia Landi



Identification of epigenetics variations influencing viral resistance in cassava farmers' field

Linking gDNA methylation profiles to phenotype

Michael Kofia Landi
Faculty of Veterinary Medicine and Animal Science
Department of Animal Biosciences
Uppsala

SLU
SWEDISH UNIVERSITY
OF AGRICULTURAL
SCIENCES

DOCTORAL THESIS
Uppsala 2025

Acta Universitatis Agriculturae Sueciae
2025:61

Cover: The image illustrates DNA methylation changes in cassava following geminivirus infection. On the left, the susceptible cassava plant shows DNA hypomethylation upon infection. On the right, the tolerant cassava plant maintains stable methylation levels despite infection. The image was designed by Michael Landi.

ISSN 1652-6880
ISBN (print version) 978-91-8124-045-0
ISBN (electronic version) 978-91-8124-091-7
https://doi.org/10.54612/a.1g43ig22a9

Swedish University of Agricultural Sciences, Department of Animal Biosciences, Uppsala, Sweden

Print: SLU Grafisk service, Uppsala 2025

Identification of epigenetics variation influencing viral resistance in cassava farmers’ field

Abstract

In sub-Saharan Africa, cassava (*Manihot esculenta Crantz*) is a key food security crop, providing calories to over 800 million people. It grows well in poor soils and tolerates drought. However, its production is threatened by cassava mosaic disease (CMD), caused by African cassava mosaic virus (ACMV), which can devastate yields. While genetic loci associated with CMD resistance have been identified, the influence of epigenetic regulation on disease resilience remains largely unexplored. This thesis investigates the epigenetic basis for the variation in CMD responses between two farmer’s preferred African cassava cultivars: TMEB693 (tolerant) and TMEB117 (susceptible). In the study, we generated a reference genome of TMEB117 for the epigenetics study. This genome enabled a comparative genomic analysis that revealed a previously unknown ~9.7 Mbp repeat-rich insertion on chromosome 12. This region is enriched with *MUDR-Mutator* transposable elements and chromatin-regulating genes (HDA14, SRT2), whose presence varied across 16 cassava landraces. Although not directly linked to CMD resistance, this hypervariable region may influence epigenetic plasticity and crop adaptation. Genome-wide methylation analysis showed that TMEB117 undergoes overall hypomethylation upon infection, resembling the methylation profile of the tolerant genotype (TMEB693). These findings suggest that TMEB117 undergoes virus-induced epigenetic reprogramming, but the resulting reduced methylation level may be too low to confer effective tolerance to the virus. In contrast, the stable methylation profile of TMEB693 may reflect a defence state that is constitutively maintained. This study highlights the role of DNA methylation in CMD resistance and establishes a genomic and epigenetic framework for identifying candidate genes and regulatory factors linked to disease resilience. Such knowledge provides a foundation for integrating molecular markers into cassava breeding programs, enabling the development of cultivars with durable CMD resilience and contributing to strengthened food security across sub-Saharan Africa.

Keywords: Cassava, genome assembly, hypervariable, cassava mosaic disease, DNA methylation, hypomethylation, resistant, tolerant

Identifiering av epigenetisk variation som påverkar virusresistens i kassavaodlares fält

Abstrakt

I Afrika söder om Sahara är kassava (*Manihot esculenta* Crantz) en viktig gröda för livsmedelssäkerhet och förser över 800 miljoner människor med kalorier. Den växer bra i magra jordar och tolererar torka. Dess produktion hotas dock av kassavamosaiksjukdomen (CMD), orsakad av afrikanskt kassavamosaikvirus (ACMV), vilket kan ödelägga avkastningen. Även om genetiska loci associerade med CMD-resistens har identifierats, är inverkan av epigenetisk reglering på sjukdomsmotståndskraft fortfarande i stort sett outforskad. Denna avhandling undersöker den epigenetiska grunden för variationen i CMD-responser mellan två av jordbrukarnas föredragna afrikanska kassavakultivarer: TMEB693 (tolerant) och TMEB117 (känslig). I studien genererade vi ett referensgenom för TMEB117 för den epigenetiska studien. Detta genom möjliggjorde en jämförande genomisk analys som avslöjade en tidigare okänd ~9,7 Mbp repetitionsrik insertion på kromosom 12. Denna region är berikad med *MUDR-Mutator* transposonerbara element och kromatinreglerande gener (HDA14, SRT2), vars närvaro varierade över 16 kassavalandraser. Även om den inte är direkt kopplad till CMD-resistens, kan denna hypervariabla region påverka epigenetisk plasticitet och grödans anpassning. Genomomfattande metyleringsanalys visade att TMEB117 genomgår total hypometylering vid infektion, vilket liknar metyleringsprofilen för den toleranta genotypen (TMEB693). Dessa fynd tyder på att TMEB117 genomgår virusinducerad epigenetisk omprogrammering, men den resulterande låga metyleringsnivån kan vara för låg för att ge effektiv tolerans mot viruset. Däremot kan den stabila metyleringsprofilen för TMEB693 återspegla ett försvarstillstånd som konstitutivt upprätthålls. Denna studie belyser rollen av DNA-metylering i CMD-resistens och etablerar ett genomiskt och epigenetiskt ramverk för att identifiera kandidatgener och reglerande faktorer kopplade till sjukdomsresistens. Sådan kunskap utgör en grund för att integrera molekylära markörer i kassavaförädlingsprogram, vilket möjliggör utveckling av kultivarer med varaktig CMD-motståndskraft och bidrar till stärkt livsmedelssäkerhet i Afrika söder om Sahara.

Nyckelord: Kassava, genomsammansättning, hypervariabel, kassavamosaiksjukdom, DNA-metylering, hypometylering, resistent, tolerant

Utambuzi wa tofauti za epigenetiki zinazoathiri ukinzani wa virusi kwa muhogo

Muhtasari

Katika Afrika Kusini mwa Jangwa la Sahara, muhogo (*Manihot esculenta* Crantz) ni zao muhimu la usalama wa chakula, linatoa kalori kwa zaidi ya watu milioni 800. Inakua vizuri kwenye udongo duni na huvumilia ukame. Hata hivyo, uzalishaji wake unatishiwa na ugonjwa wa cassava mosaic (CMD), unaosababishwa na virusi vya African cassava mosaic virus (ACMV), ambavyo vinaweza kuharibu mavuno. Ingawa loci ya kijeni inayohusishwa na upinzani wa CMD imetambuliwa, ushawishi wa udhibiti wa epijenetiki juu ya ustahimilivu wa magonjwa bado haujagunduliwa. Tasnifu hii inachunguza msingi wa epijenetiki wa utofauti wa majibu ya CMD kati ya aina mbili za mihogo ya Kiafrika inayopendelewa na mkulima: TMEB693 (inayostahimili) na TMEB117 (inayoathiriwa). Katika utafiti, tulitengeneza jenomu marejeleo ya TMEB117 kwa ajili ya utafiti wa epijenetiki. Jenomu hii iliwezesha uchanganuzi linganishi wa jenomiki ambao ulifichua uwekaji wa kurudia-tajiri wa ~9.7 Mbp kwenye kromosomu 12. Eneo hili limerutubishwa na vipengele vinavyoweza kuhamishwa vya *MUDR-Mutator* na jeni zinazodhibiti kromati (HDA14, SRT2), ambazo uwepo wake ulitofautiana katika mashamba 16 ya mihogo. Ingawa haijaunganishwa moja kwa moja na upinzani wa CMD, eneo hili linaloweza kubadilika-badilika linaweza kuathiri upekee wa epijenetiki na urekebishaji wa mazao. Uchanganuzi wa methylation ya genome kote ulionyesha kuwa TMEB117 hupitia hypomethylation ya jumla juu ya maambukizo, inayofanana na wasifu wa methylation wa genotype inayostahimili (TMEB693). Matokeo haya yanapendekeza kwamba TMEB117 inapitia upangaji upya wa epigenetic unaosababishwa na virusi, lakini kiwango cha chini cha methylation kinaweza kuwa cha chini sana kutoa uvumilivu mzuri kwa virusi. Kinyume chake, wasifu thabiti wa methylation wa TMEB693 unaweza kuonyesha hali ya ulinzi ambayo inadumishwa kikamilifu. Utafiti huu unaonyesha jukumu la methylation ya DNA katika upinzani wa CMD na huanzisha mfumo wa genomic na epigenetic kwa kutambua jeni za mgombea na vipengele vya udhibiti vinavyohusishwa na ustahimilivu wa magonjwa. Ujuzi kama huo hutoa msingi wa kuunganisha alama za molekuli katika programu za ufugaji wa muhogo, kuwezesha ukuzaji wa mimea yenye ustahimilivu wa CMD na kuchangia katika kuimarisha usalama wa chakula katika Afrika Kusini mwa Jangwa la Sahara.

Dedication

To the new life quietly on the way, whose existence has given me strength and purpose in these final stages, and to my partner, Naomi, whose love and encouragement have carried me through.

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I. **Landi, M.,** Shah, T., Falquet, L., Niazi, A., Stavolone, L., Bongcam-Rudloff, E., & Gisel, A. (2023). Haplotype-resolved genome of heterozygous African cassava cultivar TMEB117 (Manihot esculenta). Sci Data, 10(1), 887. https://doi.org/10.1038/s41597-023-02800-0

II. **Landi, M.,** Carluccio, A. V., Shah, T., Niazi, A., Stavolone, L., Falquet, L., Gisel, A., & Bongcam-Rudloff, E. (2025). Genome-wide comparison reveals large structural variants in cassava landraces. *BMC genomics*, *26*(1), 362. https://doi.org/10.1186/s12864-025-11523-y

III. **Landi, M.,** Carluccio, A. V., Shah, T., Niazi, A., Stavolone, L., Falquet, L., Gisel, A., & Bongcam-Rudloff, E. (2025). Virus induced DNA hypomethylation in susceptible cassava genotype TMEB117 compared to the tolerant cassava genotype TMEB693 (manuscript)

The published papers are open access articles distributed under the terms of the Creative Commons Attribution (CC BY) licence.

The contribution of Michael Kofia Landi to the papers included in this thesis was as follows:

I. Performed the bioinformatics analysis of the data and served as the main writer of the manuscript.
II. Data compilation and took on the main responsibility for the bioinformatics analysis and writing the manuscript.
III. Performed the bioinformatics analysis of the data. Main writer of the manuscript.

In addition to paper I-III, Michael Kofia Landi contributed to the following papers during the timeframe of the doctoral project, but not included in this thesis.

I. Gisel, A., Stavolone, L., Olagunju, T., **Landi, M.**, Van Damme, R., Niazi, A., Falquet, L., Shah, T., & Bongcam-Rudloff, E. (2023). EpiCass and CassavaNet4Dev advanced bioinformatics workshop. EMBnet. journal, 29, 1045.

II. Karega, P., Mwaura, D. K., Mwangi, K. W., Wanjiku, M., **Landi, M.**, & Kibet, C. K. (2023). Building awareness and capacity of bioinformatics and open science skills in Kenya: a sensitize, train, hack, and collaborate model. Frontiers in Research Metrics and Analytics, 8, 1070390.

III. **Landi, M.**, Muzemil, S., Ondari, L. N., Adediji, A. O., Borgbara, K., Moila, A., Awoyemi, A. G., Zoclanclounon, Y. A. B., Oladimeji, T. R., Ajayi, A. D., Ebenezer, T. E., Gisel, A., & Aroworamimo, L. A. (2025). Chromosome-scale genome assembly of the white star apple Gambeya albida (Submitted)

List of tables

List of figures

Abbreviations

ACMV	African cassava mosaic virus
BLAST	Basic local alignment search tool
BUSCO	Benchmarking universal single-copy orthologs
CLR	Continuous long reads
CMD	Cassava mosaic disease
CMG	Cassava mosaic geminiviruses
CMT	Chromomethyltransferases
CpG, CHG and CHG	Cytosine methylation contexts (H = A, C, or T)
CRISPR	Clustered regularly interspaced short palindromic repeat
DMR	Differential methylated regions
DNA	Deoxyribonucleic acid
DRM	Domain rearranged methyltransferase
EM-seq	Enzymatic Methylation Sequencing
GBS	Genotyping-by-sequencing
GWAS	Genome-wide association studies
HiFi	High-fidelity
MET	Methyltransferases
PCR	Polymerase Chain Reaction
QTL	Quantitative trait locus
QV	Quality Value
RdDM	RNA-direct DNA methylation
RNA	Ribonucleic acid

ROS	Repressor of silencing
siRNA	Small interfering RNA
SMRT	Single-molecule real-time
TE	Transposable elements
TMEB	Tropical *Manihot Esculenta* breeding
TMS	Tropical manihot series
WGSB	Whole genome bisulfite sequencing

1. Introduction

Viruses are a major threat to global crop production, particularly in clonally propagated crops such as potato, sweet potato, and cassava. Cassava (*Manihot esculenta* Crantz), one of sub-Saharan Africa's most important staple crops, plays a vital role in food security, serving as the primary source of carbohydrates for over a billion people. It is valued for its adaptability to drought and nutrient-poor soils, and its flexible harvest window allows roots to be stored underground as a food reserve (Prochnik et al., 2012). However, this resilience is affected by pest and diseases. One of the threats is cassava mosaic disease (CMD), a viral infection transmitted by the whitefly (*Bemisia tabaci*) and through the reuse of infected cuttings. CMD spreads rapidly in areas with prevalent vectors and diseased planting material (Fauquet et al., 2005; Wossen et al., 2017). Yield losses in susceptible cultivars range from 20% to 95% (Fauquet & Fargette, 1990), causing estimated annual economic loss of US$1.9–2.7 billion across more than 2.6 million per km^2 pf CMD-affected area in East and Central Africa (Ndunguru et al., 2006). These yield losses are among the main factors limiting cassava productivity in sub-Saharan Africa, where average yields are 9–10 t ha^{-1}, compared to the crop's potential of 60–80 t ha^{-1} under optimal conditions (FAOSTAT, 2025). Efforts to manage CMD have included phytosanitary measures, vector control, and the deployment of CMD-resistant varieties. Given that the first two approaches are labour-intensive and difficult to sustain, deploying resistant cultivars remains the most effective control method for this devastating disease.

The source of CMD resistance is the CMD2 locus, a single dominant resistance locus first discovered in West African CMD-resistant landraces and initially mapped to chromosome 8, however, subsequent reports have investigated this locus on chromosome 12 (Rabbi et al., 2014; Rabbi et al., 2022; Ramu et al., 2017). While genetic mapping and genome-wide association studies (GWAS) have provided valuable insight into the inheritance of CMD resistance, the underlying molecular mechanisms of resistance, particularly the regulatory processes that control CMD resistance, remain poorly understood. This knowledge gap provides a strong rationale for investigating additional regulatory layers, such as epigenetic modifications, that may influence CMD resistance.

Epigenetic mechanisms such as DNA methylation alter gene expression without changing the DNA sequence and can influence plant defence responses. In cassava, whole-genome methylome analyses have shown that gene body CG methylation is positively correlated with gene expression (Wang et al., 2015), consistent with patterns observed in other plants (Li et al., 2012; Zemach et al., 2010; Zhang et al., 2006). Haplotype-resolved methylome analysis in the cultivars TME7 and TME204 revealed broadly similar global methylation levels between haplotypes, but local differences were linked to structural variants and allele-specific expression, particularly near transposable elements and intergenic regions (Zhong et al., 2023). Under abiotic stress, such as drought, methylation landscapes differ between tolerant and sensitive cassava genotypes, with differentially methylated regions (DMRs) enriched in hormone signalling and stress-response genes (Silva Filho et al., 2024). Despite these advances, the role of DNA methylation in CMD resistance has not yet been investigated. Understanding the role of DNA methylation could reveal additional regulatory layers influencing virus–host interactions and CMD resistance. This thesis addresses this gap by comparing the methylomes of two farmer-preferred cassava cultivars, TMEB693 (CMD-tolerant) and TMEB117 (CMD-susceptible), which are genetically similar but differ in their phenotypic response to CMD. By analysing uninfected and infected plants, we aim to characterize the global methylation profiles of these genotypes and identify genotype-specific DMRs in genomic features such as promoters and gene bodies that may be associated with resistance responses. This work will advance understanding of the epigenetic regulation of cassava's viral defence mechanisms and provide knowledge to support breeding strategies for CMD-resistant varieties that retain farmer-preferred varieties, thereby contributing to yield stability, income security, and food security in sub-Saharan Africa.

2. Background

2.1 Cassava

2.1.1 Origin, domestication and genetic diversity

Cassava is a perennial woody shrub belonging to the Euphorbiaceae family, within the Crotonoideae subfamily and Manihotae tribe. It falls under the Manihot genus, consisting of 300 genera and around 7,500 species, making it one of the largest family of flowering plants (A.H.M. Mahbubur Rahman, 2013). Within this genus, there are two main sections: Arborea, which includes tree species, and Fructicosae, composed of slow-growing shrubs that are adapted to grassland savannahs or desert conditions. The Fructicosae section is considered less primitive than Arborea (Jennings D.L., 2002). Among the many species within the Manihot genus, *Manihot esculenta* Crantz is the most economically significant, as it is widely cultivated for food production and industrial applications (Amelework & Bairu, 2022).

The domestication of cassava is estimated to have occurred before 4000 BC, with its origin believed to be South America (Nassar N.M.A., 2009). Recent phylogenomic analyses suggest cassava originated from its wild progenitor, *Manihot esculenta* subsp. *flabellifolia*, in tropical South America, particularly along the southern rim of the Amazon Basin (Allem, 1994, 1999; Olsen, 1999). Multiple studies have confirmed *M. esculenta* subsp. *flabellifolia*, as the closest wild relative of cassava. Genetic analyses further indicate that cultivated cassava likely derived from a single wild species, *M. esculenta* subsp. *flabellifolia*, rather than from multiple hybridizing species, as previously hypothesized (Olsen, 2004; Olsen & Schaal, 2001; Olsen, 1999; Roa A.C., 1997; Roa et al., 2000).

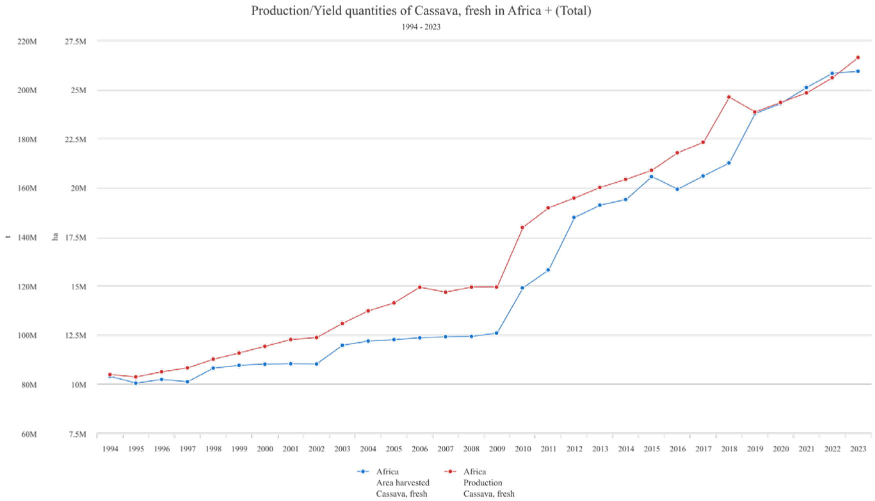
Portuguese traders introduced cassava to the western shores of Africa during the 16th century as a food source for slave ships. Several Portuguese trading posts in the Gulf of Guinea, Sierra Leone, and the coastal areas of Angola and the Democratic Republic of Congo (DRC) facilitated its introduction (Carter, 1992). Cassava's spread inland progressed gradually, but by the time European explorers ventured deeper into the continent in the

19th century, it had already become a well-established crop across West, Central, and East Africa (Jones, 1959). The introduction of cassava to East Africa is speculative. Historical accounts suggest it arrived through Portuguese trading ports along the coast, including Mozambique Island, Sofala, Kilwa, Benguela, Mombasa, Zanzibar, and Pemba, between the 17th and 18th centuries. However, its inland expansion from the East African coast was limited. Instead, cassava reached Rwanda and Burundi from the west and spread to the upper Zambezi region from Angola (Carter, 1992). Today, cassava ranks among the most broadly grown tropical crops, particularly in sub-Saharan Africa, where targeted breeding and introgression efforts have been implemented to enhance yield and improve resistance to diseases.

Cassava is clonally propagated by planting stem cuttings, a practice that has contributed to the accumulation of considerable genetic load (Kawuki et al., 2010). Although it primarily reproduces vegetatively, the plant also produces true seeds from separate male and female flowers adapted for cross-pollination (Jennings, 1963). This combination of clonal propagation and occasional sexual reproduction contributes to high levels of heterozygosity within individual clones (McKey et al., 2010) However, the overall genetic diversity of cassava populations varies by region. African germplasm, in particular, is derived from a relatively narrow genetic base and exhibits limited allelic richness compared to South American and Southeast Asian collections (Bredeson et al., 2016).This restricted diversity and reliance on vegetative propagation have contributed to the accumulation of genetic load and increased vulnerability to diseases such as cassava mosaic disease and cassava brown streak disease (Bredeson et al., 2016).

2.1.2 Cassava production and constraints

Cassava is both a subsistence crop and a source of income and is well integrated into African farming systems. Root and tuber crops play an essential role in tropical agriculture, with cassava ranking among the most widely cultivated tuber crops (Balagopalan, 2018). The starch-rich fresh roots of cassava are a vital food staple across Africa, providing calories to nearly half a billion people worldwide (Parmar et al., 2017). Between 1994 and 2023, cassava has experienced consistent production growth averaging over 3% annually (Fig. 1). On average, Africa contributes 57.2% of total global cassava production, followed by Asia with 29.9% and the Americas with 12.8%. Nigeria is the world's leading producer, where cassava is also the most widely cultivated crop, with an annual production of 45 million tonnes. Other major producers include the Democratic Republic of Congo (26 million tonnes), Brazil (22 million tonnes), Thailand (24 million tonnes), Indonesia (18 million tonnes), and Ghana (13 million tonnes) (FAOSTAT, 2025).



Source: FAOSTAT (July 16, 2025)

Figure 1 Cassava production in Africa (1994-2023, (FAOSTAT, 2025). The plot show both the production of cassava in tonnes (line in red) and the area harvested in hectares in Africa (line in blue)

To meet the foreseen rise in demand for cassava as both a food staple and an industrial raw material, there is an urgent need to increase cassava production across sub-Saharan Africa (Khandare & Choomsook, 2019; Otekunrin & Sawicka, 2019). However, several persistent constraints continue to limit productivity and undermine food security. Among the most significant challenges are pests, such as cassava green mites and whiteflies (Kalyebi et al., 2018; Koros et al., 2018), and devastating diseases caused by bacteria and viruses, including cassava bacterial blight (Fanou et al., 2017) CMD, and cassava brown streak disease (Alicai et al., 2019; Patil et al., 2015). These biotic stresses frequently lead to severe yield losses in affected regions.

Cassava contains cyanogenic glucosides, which, upon hydrolysis, release hydrogen cyanide, a compound toxic to humans if not properly processed (Akinpelu et al., 2011). From a nutritional perspective, cassava is predominantly an energy-dense crop mainly composed of starch, with relatively low levels of protein and essential micronutrients. Consequently, populations that rely heavily on cassava as their main dietary staple are at heightened risk of deficiencies in vitamin A, zinc, and iron (Gegios et al., 2010; Stephenson et al., 2010). These constraints have shaped current cassava breeding and genomics efforts to develop varieties that combine desirable traits. These efforts include achieving higher yields of dry matter per hectare to improve productivity, enhancing resistance to major diseases such as CMD, cassava brown streak disease and cassava bacterial blight, as well as key insect pests like cassava green mites and whiteflies. Efforts also focus on improving both the quantity and quality of starch, reducing cyanogenic potential to ensure food safety, and increasing the nutritional value of cassava roots through biofortification. Modern genomics approaches, including marker-assisted selection, genomic selection, GWAS, and CRISPR/Cas9 genome editing, are increasingly being integrated into these breeding programs to accelerate the development of improved cassava varieties (Andrade et al., 2019; Carmo et al., 2015; Juma et al., 2022; Rabbi et al., 2022; S. Zhang et al., 2018).

2.1.3 Cassava mosaic disease

CMD is one of the major constraints to cassava production. It is caused by 11 species of cassava mosaic geminiviruses (genus *Begomovirus*, family *Geminiviridae*) that occur across Africa and the Indian subcontinent. In Africa, the known species include ACMV, African cassava mosaic Burkina Faso virus, Cassava mosaic Madagascar virus, East African cassava mosaic Cameroon virus, East African cassava mosaic Kenya virus, East African cassava mosaic Malawi virus , East African cassava mosaic virus, East African cassava mosaic Zanzibar virus, and South African cassava mosaic virus (Fondong, 2017; Legg et al., 2015; Patil et al., 2015). CMD is transmitted by the whitefly (*Bemisia tabaci*) and through infected cuttings replanted as propagation material (Fauquet et al., 2005). A strong correlation exists between the reuse of infected cuttings and the overall incidence of CMD in fields (Wossen et al., 2017). Infection rates are highest in production areas with abundant infected planting material and whitefly populations (Wossen et al., 2017). CMD can lead to estimated yield losses in susceptible cassava cultivars ranging from 20% to 95% (Fauquet & Fargette, 1990). CMD was first reported in Tanzania (Warburg, 1894) and later identified as a viral disease (Zimmermann, 1906). Initially referred to as cassava latent virus (Bock et al., 1981), the pathogen was renamed ACMV after its genome was sequenced (Stanley & Gay, 1983). Studies indicate that the CMD pandemic affected at least nine countries across East and Central Africa, spanning an area of approximately 2.6 million square kilometres. The disease is estimated to cause annual economic losses of US$1.9–2.7 billion and has been described as the most devastating plant virus globally, linked to food insecurity (Ndunguru et al., 2006).

Typical symptoms of CMD include leaf mosaic patterns, leaf reduction, and chlorosis. However, the expression of symptoms often varies among leaves, shoots, and entire plants within the same variety. This variability can be influenced by the specific virus strain or species involved, the sensitivity of the host genotype, plant age, and environmental conditions such as soil fertility and moisture availability (Hillocks & Thresh, 2000). Management strategies for CMD include rogueing (removal of symptomatic plants), the use of virus-free planting materials, and deploying resistant varieties (Rabbi et al., 2014). The first two methods are labour-intensive, difficult to sustain over time, and require continuous interventions. In contrast, using resistant varieties is the most effective approach for reducing CMD impact in cassava production (Akano et al., 2002; Otim-Nape et al., 1994). This strategy minimizes yield losses and lowers virus inoculum levels in the farming system, particularly when resistant varieties suppress virus accumulation (Rabbi et al., 2014).

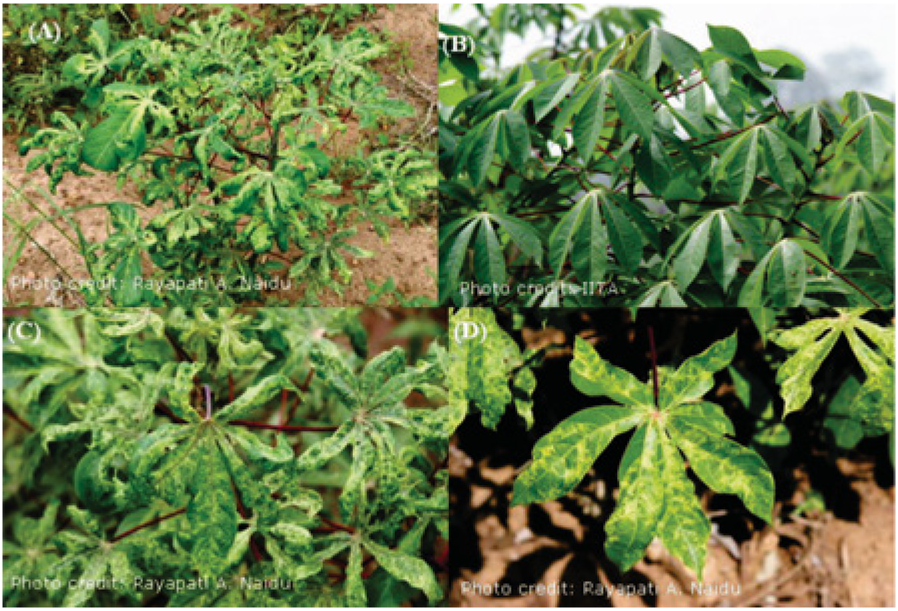

(A)
Photo credit: Rayapati A. Naidu
(B)
Photo credit: IITA
(C)
Photo credit: Rayapati A. Naidu
(D)
Photo credit: Rayapati A. Naidu

Figure 2 Casava plant showing symptoms of CMD. A - severe stunting and distortion of leaves, B – Healthy leaves, C and D – Misshapen and twisted leaflets with mosaic and mottling symptoms

2.1.4 CMD resistance in cassava

Breeding efforts have produced high-yielding cassava varieties with strong resistance or tolerance to cassava mosaic geminiviruses (CMGs). Three genetically distinct mechanisms of CMD resistance or tolerance have been identified in cassava (Okogbenin et al., 2012; Rabbi et al., 2014). CMD1 resistance was introgressed from *Manihot glaziovii* (ceara rubber) and is characterized as polygenic and recessive in inheritance (Fregene et al., 2002). This source of resistance has been used extensively in African breeding programs and contributes to the reduced infection rates and milder disease symptoms observed in many improved varieties. In contrast, CMD2 resistance originates from a single dominant genetic locus found in several West African landraces belonging to the Tropical *Manihot esculenta* (TME) series (Akano et al., 2002; Rabbi et al., 2014). Because of its simple inheritance and consistent expression of strong resistance across environments, the CMD2 locus has been widely exploited in African and Latin American breeding programs to develop genotypes that are highly resistant to a broad range of CMGs (Okogbenin et al., 2013; Rabbi et al., 2014). CMD3, has been described in the elite cultivar TMS 97/2205 (Okogbenin et al., 2012). This variety was developed through crosses between TMS 30572, which carries CMD1-type resistance, and TME 6, a CMD2-resistant landrace. Field evaluations have shown that TMS 97/2205 displays extreme resistance, with less than 1% disease incidence observed even under high CMD pressure in Nigeria. Genetic analyses of TMS 97/2205 indicate the presence of both the CMD2 locus and an additional resistance locus within the same linkage group, suggesting that combining multiple sources of resistance may further enhance durability and effectiveness (Okogbenin et al., 2012).

2.2 Cassava Genome

2.2.1 Cassava reference genome

Cassava was among the first "orphan" crops to have its genome sequenced. Over the past decade, the quality of its genome assembly has steadily improved, benefiting from continuous advances in sequencing technologies and bioinformatics tools (Lyons et al., 2022). Cassava is a diploid genome, containing 18 chromosomes (2n=36) with a haploid genome size of about 750 Mbp, a highly heterozygous and full of repetitive elements, which make up two-thirds of the genome (Prochnik et al., 2012). While these features create challenges for genome assembly and sequence analysis, they are essential for understanding the genetic basis of cassava's phenotypic traits (Elias et al., 2018).

The reference genome of cassava is the AM560-2 accession, derived from a Colombian cassava line MCol505. The AM560-2 cassava reference genome has undergone five major releases (Table 1), each reflecting significant advances in sequencing technologies and assembly algorithms. The first cassava reference genome, version 4.1 (v4.1), was released in 2009 and built using Roche 454 sequencing technology. Although fragmented, this assembly captured most of the gene models and provided the first insights into cassava's repetitive genome composition (Prochnik et al., 2012).Version 5 (v5.1) improved the genome by elevating the assembly to chromosome scale, using genetic map to order and orient 57% of the sequences into 18 chromosomes (Consortium, 2015; De Carvalho & Guerra, 2002). With the advancement of new sequencing technologies, version 6 (v6.1) was created *de novo* from deep Illumina short-read data (120× coverage), assembled into contigs, scaffolded with mate pairs and fosmid ends, and anchored to chromosomes via genetic maps (Bredeson et al., 2016). Compared to v5.1, v6.1 captured 18% more contig sequence, improved sequence continuity, and incorporated 45% more sequence into chromosomal scaffolds, becoming the reference for cassava genomics research at the time (Andrade et al., 2019; Kayondo et al., 2018; Kuon et al., 2019; Nzuki et al., 2017; Ramu et al., 2017; Wolfe et al., 2019). In 2019, version 7 (v7.1) was released on Phytozome, incorporating PacBio continuous long-read (CLR) sequences, most exceeding 10 kb. These long reads greatly improved the assembly of repetitive regions, typically fragmented or collapsed in short-read assemblies. While CLR data have higher raw error rates than Illumina reads, advances in error correction and assembly algorithms, such as Canu, have mitigated mainly this limitation (Koren et al., 2017). Using Canu, contigs were built from error-corrected reads longer than 4 kb (34× coverage) and scaffolded similarly to v6.1. This strategy increased the assembly by over 100 Mbp and improved contiguity more than 25-fold. Although the v7.1 assembly spans 669 Mbp, somewhat less than the ~750 Mbp haploid genome size estimated by flow cytometry (Kuon et al., 2019), it is consistent with estimates from shotgun sequencing, and any missing sequences are likely limited to highly repetitive regions. The recent cassava reference genome, version 8.1 (v8.1), was assembled using PacBio CLR and fosmid-end data from earlier versions, replacing genetic linkage maps with newly generated high-resolution Hi-C data and additional long-read sequences. This strategy anchored 26 Mbp more sequence than v7.1 and nearly 188 Mbp more than v6.1, substantially improving completeness and contiguity. Specifically, after gap filling, assembly contiguity increased by about 4.8X over v7.1 (Bredeson et al., 2021).

Table 1 Cassava reference genome assemblies

	V4.1	V5.1	V6.1	V7.1	V8.1
Release	2009	2014	2016	2019	2021
Sequence technology	454	454	Illumina	PacBio	PacBio
Scaffolding data	454 mate pair	Genetic map	Illumina mate pair, fosmid and genetic maps	Illumina mate pair, fosmid and genetic maps	Illumina mate-pair, fosmid and HiC
Total contig length	419 Mbp	419 Mbp	496 Mbp	667 Mpb	637 Mpb
Total scaffold length	533 Mbp	534 Mbp	582 Mbp	669 Mbp	639 Mbp
Chromosome number	-	18	18	18	18
Contig N50	11 Kbp	11 Kpb	27 Kbp	693 Kpb	34 Mpb
Annotated genes	30,666	30,666	33,033	33,849	32,447

2.2.2 African cassava genomes and resequencing

While the AM560-2 reference genome has been the reference for cassava genomics, sequencing additional cassava varieties is increasingly expanding our understanding of the crop's genetic diversity. For example, (W. Wang et al., 2014) initiated efforts of sequencing additional cassava genomes by reporting assemblies of two varieties, W14 and KU50. W14 was initially identified as *Manihot esculenta* ssp. *flabellifolia*, often called a "wild cassava" or "wild ancestor." However, subsequent analyses of the sequence data revealed that W14 is not *M. esculenta* but is more closely related to *M. glaziovii* (Bredeson et al., 2016). KU50, on the other hand, is an improved cassava cultivar widely grown in Southeast Asia (Ceballos et al., 2020). The assemblies of W14 and KU50 combined Illumina whole-genome shotgun libraries, 454 BAC clone sequencing, and linkage information from the

AM560-2 reference genome. Although these assemblies were valuable contributions to cassava genomes, they are less complete than versions of the AM560-2 reference, such as v6.1 and v7.1. The advent of long-read sequencing technologies has enabled more comprehensive assemblies of complex, outbred cassava genomes, making it possible, in principle, to reconstruct both haplotypes of the diploid genome. Early efforts to assemble African cassava lines, were the sequencing of TME3 and 60444 cassava genotypes. The genome assembly used a combination of Illumina short reads, PacBio long reads, BioNano optical mapping, and Hi-C technologies, generating assemblies with N50 values of 98 Kbp and 117 Kbp, respectively (Kuon et al., 2019). Although these assemblies were relatively contiguous, they did not fully resolve haplotypes and contained duplicated sequences within the primary assembly. In contrast, the TME7 genome was assembled using Illumina, PacBio, and Hi-C data to produce a more contiguous assembly with an N50 of approximately 320 Kbp, which was successfully deduplicated and phased (Mansfeld et al., 2021). In the following year, the TME204 genome was assembled and phased using Hi-C and PacBio high-fidelity (HiFi) sequencing, resulting in a highly contiguous assembly with an N50 exceeding 18 Mbp (Qi et al., 2022). PacBio HiFi reads, in particular, have proven highly effective for generating long, accurate sequences to tackle the challenges of assembling complex, heterozygous genomes (Hon et al., 2020). The most recent African cassava genome, TMEB117 (also known as TME117, TME 117, and ISUNIKANKIYAN), is a Nigerian landrace highly susceptible to ACMV. This genotype served as a reference for ACMV studies in this thesis and represents one of the highest-quality cassava assemblies. The chromosome-scaled haplotype-resolved genome of TMEB117 was constructed using PacBio HiFi reads and leveraged the TME204 genome as a reference to order contigs into chromosome level genome. Each haplotype achieved a contig N50 of approximately 18 Mbp and a scaffold N50 exceeding 35 Mbp, with overall base-level accuracy greater than QV 64, higher than previously sequenced African cassava genomes (Landi et al., 2023).

In addition to *de novo* assembled genomes, many cassava varieties have been resequenced using short-read sequencing technologies. Resequencing refers to generating whole-genome shotgun sequences without assembling them *de novo*. These sequences are then aligned to a reference genome to study genetic variation. Early efforts, such as the HapMap I project (Bredeson et al., 2016), resequenced 53 cultivated cassava accessions and five wild *Manihot* species. The following HapMap II project expanded this work to 241 accessions, including both cultivated varieties and wild relatives, and produced extensive single nucleotide polymorphisms (SNP) datasets that are publicly available on Cassavabase (www.cassavabase.org) (Ramu et al., 2017). Additional whole-genome resequencing of 388 cassava accessions, including 38 landraces, 33 breeding lines, and 14 wild relatives (Hu et al., 2021), was performed. 51 of these accessions were from Hap I and II projects. In the study (Hu et al., 2021), they identified 52 loci associated with 23 agronomic traits and revealed allelic variation in heterozygosity linked to cassava domestication and key traits. The study also showed that artificial selection of homozygous alleles in genes such as *MeTIR1* and *MeAHL17* contributed to increased starch content and larger storage roots. However, the latter also conferred susceptibility to cassava bacterial blight. These high-resolution datasets are valuable resources for genomic selection, breeding, and improving cassava and other highly heterozygous crops.

2.2.3 Cassava genome resources for crop improvement

The application of genomics in cassava improvement has been key in overcoming challenges posed by its biological characteristics, including a long growth cycle and a highly heterozygous genetic background, which have historically slowed progress toward breeding goals such as yield increases and disease resistance. Integrating genomics with traditional breeding approaches has shown considerable promise in enhancing adaptation to abiotic stresses. For instance, GWAS have enabled the identification of SNPs linked to key traits such as resistance to CMD, cassava green mite severity, and important yield and quality attributes, providing valuable markers and candidate genes for breeding programs (Rabbi et al., 2022).

Beyond disease resistance, improving cassava's nutritional quality has also benefited from genomic approaches. Genomic selection has proven effective for rapidly increasing provitamin A carotenoid content, offering a promising strategy for nutritional enhancement (Esuma et al., 2021). Moreover, the identification of quantitative trait loci (QTL) associated with carotenoid levels has laid the groundwork for developing cassava varieties with improved nutritional profiles (Yonis et al., 2020).

2.3 Epigenetics

2.3.1 Overview of epigenetic in plants

Aristotle first introduced the concept of epigenesis. William Harvey later expanded upon it 1650, coining "epigenetics" and defining it as a gradual developmental process that increases complexity from the initially homogeneous material found in different animals' eggs (Van Speybroeck et al., 2006). In 1942, Waddington redefined epigenetics as "the whole complex of developmental processes" that mediate between the genotype and phenotype (Ahmad et al., 2011). The redefined term "epigenetics" encompasses all mitotically and/or meiotically heritable changes in gene expression that are not encoded in the DNA sequence itself (Tsaftaris et al., 2005). In plants, such epigenetic mechanisms are essential because, although they are stationary organisms, their environment constantly changes, exposing them to a wide range of external stimuli and signals that require finely tuned responses. Consequently, plants must continuously adapt their genetic potential to cope with these challenges and have evolved sophisticated molecular functions to regulate gene expression, particularly under environmental stresses such as interactions with symbiotic and pathogenic microorganisms and changes in abiotic factors including salinity, temperature, sunlight, drought, flooding, cold, and heat (Du et al., 2024; Nawaz et al., 2023). Epigenetic studies have focused primarily on mammals, leaving plant studies underexplored. However, this has changed recently, as epigenetic research has moved to the forefront of plant biology and molecular genetics. While the field is less extensive than in animals, several plant species have become valuable models for understanding epigenetic regulation. In particular, research on *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, strawberry, among others, has revealed crucial mechanisms such as DNA methylation, histone modifications, and small interfering RNAs (siRNAs) that play key roles in development and environmental responses (Gayacharan & Joel, 2013; López et al., 2022; Schmitz & Ecker, 2012). The three main aspects of epigenetics in plants, are DNA methylation and demethylation, histone modification, and chromatin remodelling. In this thesis, we focus on DNA methylation.

2.3.2 DNA methylation in plants

DNA methylation refers to the addition of a methyl group to cytosine bases in DNA, resulting in the formation of 5-methylcytosine (Lucibelli et al., 2022). In plants, cytosine methylation occurs in three sequence contexts: the symmetric CG and CHG contexts, and the asymmetric CHH context, where H represents A, C, or T (Lucibelli et al., 2022). Methylation is most abundant in CG, followed by CHG and CHH (in percentage, calculated as the number of methylated sites divided by total coverage) (Gallego-Bartolomé, 2020). This methylation percentage distribution pattern has been consistently observed across multiple plant species, including *Arabidopsis thaliana* (CG: 24%, CHG: 6.7%, CHH: 1.7%) (Cokus et al., 2008), cassava (*Manihot esculenta*) (CG: 58.7%, CHG: 39.5%, CHH: 3.5%) (Wang et al., 2015), soybean (*Glycine max*) (CG: 63%, CHG: 44%, CHH: 5.9%) (Song et al., 2013), maize (*Zea mays*) (CG: 65%, CHG: 50%, CHH: 5%) (Regulski et al., 2013), and rice (*Oryza sativa*) (CG: 54.7%, CHG: 37.3%, CHH: 12%)(Li et al., 2012). Studies have shown that gene body methylation predominantly occurs in the CG context (Bewick & Schmitz, 2017), whereas transposable elements (TEs), which are mobile genetic elements, are characterized by high methylation levels across all three sequence contexts (Frost et al., 2005).

DNA methylation in plants involves three fundamental processes: *de novo* methylation, methylation maintenance, and DNA demethylation (Elhamamsy, 2016). *De novo* methylation is primarily catalysed by Domains Rearranged Methyltransferase 2 (Chan et al., 2005), a plant homolog of mammalian DNA methyltransferase (Kim et al., 2009). This process is guided by small RNAs through a mechanism known as RNA-directed DNA methylation (RdDM) (Henderson & Jacobsen, 2007), which depends on two plant-specific RNA polymerases, Pol IV and Pol V (Law & Jacobsen, 2010). Small RNAs, including siRNAs and miRNAs, are key in directing methylation to target sequences and regulating gene expression, especially in defence responses (Kim & Zilberman, 2014). For instance, specific miRNAs in rice have been shown to enhance resistance to *Magnaporthe oryzae* by upregulating defence-related genes (Li et al., 2014), while mutations in Argonaute 4, an essential RdDM component, increase *Arabidopsis* susceptibility to *Pseudomonas syringae* infection (Agorio & Vera, 2007; Zilberman et al., 2003). The maintenance of DNA methylation varies by sequence context. CG methylation is maintained by Methyltransferase 1 (MET1), CHG methylation by Chromomethylase 3 (CMT3), and CHH methylation by either CMT2 or continued RdDM activity (Chan et al., 2005; Law & Jacobsen, 2010). Mutations disrupting these pathways can lead to altered pathogen resistance; for example, *Arabidopsis* mutants deficient in CG methylation (MET1), or both CHG and CHH methylation, showed enhanced resistance to *P. syringae* and improved resistance to *Hyaloperonospora arabidopsidis* (Dowen et al., 2012; Luna et al., 2012; Yu et al., 2013). To balance methylation levels of the genome and maintain gene expression, plants rely on DNA demethylation, which removes 5-methylcytosine and replaces it with unmethylated cytosine. In *Arabidopsis*, this is carried out by four DNA demethylase enzymes: Demeter, Repressor of Silencing 1 (ROS1), Demeter-Like 2 (DML2), and DML3. DME is essential for maternal allele demethylation during endosperm development (Schoft et al., 2011), and ROS1 regulates both development and stress responses (Gong et al., 2002).

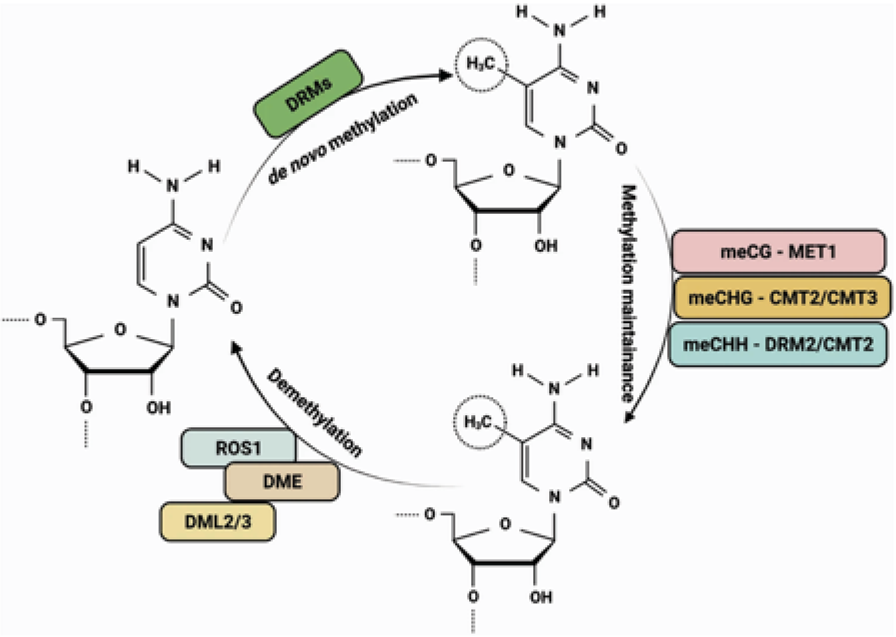


Figure 3 DNA methylation processes: de novo methylation, maintenance of methylation and demethylation processes (Arora et al., 2022).

2.3.3 Plant pathogen interaction

DNA methylation has two roles in plant–virus interactions: it influences viral adaptation and virulence, and it shapes the host's defence responses (Lang et al., 2017; Saze et al., 2003). Viral infections activate the plant's gene silencing machinery. One of the main antiviral defence strategies is RdDM. In this process, the plant produces siRNAs that help direct methylation enzymes to viral DNA, silencing it. The methylation marks are then maintained by specific enzymes, including MET1 and CMT2/3. However, viral proteins actively interfere with these processes by inhibiting enzymes responsible for de novo methylation and maintenance (Arora et al., 2022; Ashapkin et al., 2020). For example, geminiviruses encode proteins such as Transcriptional Activator Protein, which disrupt the host methylation machinery and reprogram the expression of defence-related genes (Buchmann et al., 2009; Raja et al., 2008). During infection, geminiviruses rely on specialized proteins to enter plant cells and start replicating. In bipartite viruses, DNA-A encodes AC proteins that support viral replication and AV proteins that enable movement between plant cells. Monopartite viruses have corresponding C and V proteins with similar roles (Arora et al., 2022). These viral factors help deliver the viral genome into the nucleus and suppress plant defences, including methylation-based silencing mechanisms that target viral DNA. DNA methylation also interacts with plant immune pathways, notably pattern-triggered immunity and effector-triggered immunity, by regulating the expression of defence genes (Alhoraibi et al., 2018; Nabi et al., 2024). Sometimes, hypomethylation of defence-related gene promoters leads to their activation, enhancing resistance (Lee et al., 2023; Lu et al., 2024). Conversely, hypermethylation can suppress these genes and increase susceptibility. The interplay between methylation dynamics and viral suppression of host defences highlights the complexity of plant–virus interactions and the central role of epigenetic regulation in determining infection outcomes.

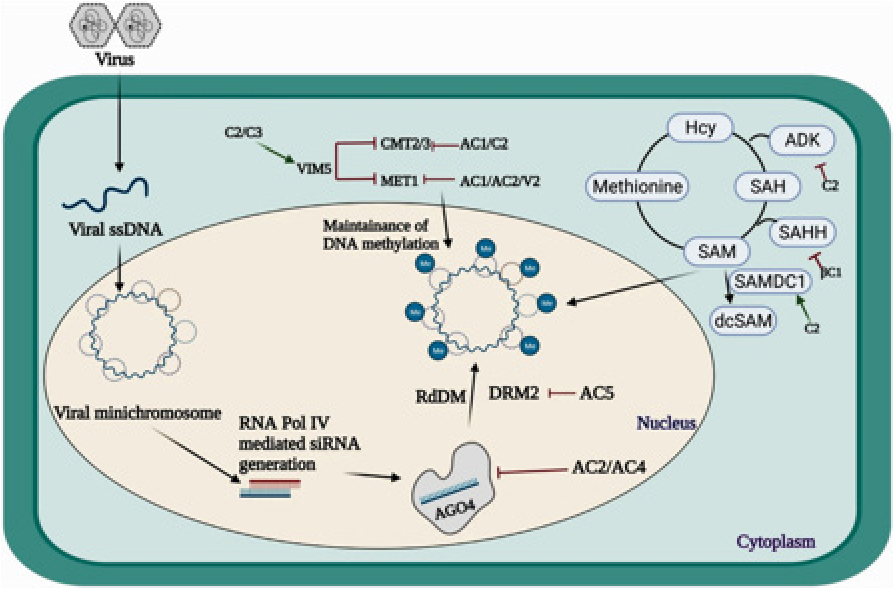


Figure 4 Schematic overview of viral suppression of DNA methylation in plants.

2.3.4 DNA methylation and gene expression

Plants adaptation to stress conditions depends on transcriptome reprogramming (Elhamamsy, 2016), and DNA methylation plays a crucial role in shaping gene expression during development and environmental challenges (Choi & Sano, 2007; Wada et al., 2004; Wang et al., 2011; Yaish, 2013). The relationship between methylation and expression is complex and influenced by genomic location, for example in promoters, gene bodies, and transposable elements (TEs) (J. Wang et al., 2014).

Promoter Methylation

Methylation within promoter regions is a primary mechanism for regulating gene activity. Generally, promoter methylation is negatively correlated with gene expression (Jones, 2012; Singer et al., 2001; Wang et al., 2015; Zhang et al., 2006). Pathogen-induced hypomethylation of promoters can activate defence-related genes and promote resistance. For example, in rice, the promoter of the *Xa21G* resistance gene is hypomethylated in a mutant line compared to the wild type, leading to higher expression and enhanced resistance to *Xanthomonas oryzae* pv. *oryzae* (Akimoto et al., 2007). Similarly, abiotic stresses like salinity and drought can induce promoter hypomethylation, upregulating stress response genes (Choi & Sano, 2007; M. Wang et al., 2014; Yaish et al., 2018). However, promoter hypomethylation is not always required for increased gene expression. Rice blast resistance gene *Pib,* for example, maintains high expression despite promoter hypermethylation. Partial demethylation of the promoter reduces *Pib* expression and compromises resistance (Li et al., 2011). These observations indicate that hypomethylation and hypermethylation can benefit plants under stress conditions. Global increases in DNA methylation can broadly suppress transcription, reducing cellular energy consumption, an essential adaptation during pathogen attack or abiotic stress. In contrast, targeted hypomethylation of resistance genes can enhance their expression, enabling a rapid adaptive response to environmental challenges (Yaish, 2013). To tackle DNA methylation effectively for disease resistance, it is essential to understand which strategy hypomethylation or hypermethylation plants deploy against specific pathogens.

Gene body methylation

Compared to promoters, the role of gene body methylation in gene regulation is less well defined. In many plant species, gene body methylation in the CG context is generally positively correlated with gene expression. Genome-wide studies in *Arabidopsis* revealed that moderately expressed genes are most likely to be methylated within their coding regions, whereas genes with very low or very high expression tend to be less methylated (Henderson & Jacobsen, 2007; Zilberman et al., 2007). Similar patterns have been reported in cassava (Wang et al., 2015), rice (Li et al., 2012; X. Wang et al., 2017), and soybean (Kim et al., 2015). By contrast, methylation in CHG and CHH contexts tends to be negatively correlated with expression, as seen in tomato (González et al., 2011) and *Arabidopsis* (You et al., 2012). The location of gene body methylation can also influence gene function. In maize and *Arabidopsis*, CHG and CHH methylation often occurs at intron–exon boundaries and may regulate alternative splicing by inhibiting RNA splicing machinery (Regulski et al., 2013). This positional effect creates additional layers of transcriptional and post-transcriptional control. Resistance genes such as NLRs can display atypical methylation patterns. In common bean, some NLRs are methylated across CG, CHG, and CHH contexts, resembling transposon-like methylation patterns (Richard et al., 2018). Although the regulatory significance of this is still unclear, it suggests that gene body methylation could have unique roles in defence gene expression.

TE methylation

TEs regulate gene expression through multiple mechanisms that depend on their activation, which is mainly controlled by their DNA methylation status (Yoder et al., 1997). Genome-wide hypomethylation activated by biotic or abiotic stress can activate transposons, increasing their mobility within disease-related genes and thereby altering gene expression levels (Biémont & Vieira, 2006; Forestan et al., 2016; Grandbastien et al., 2005; C. Wang et al., 2017). These mechanisms used by TEs to regulate gene expression include: (i) insertion into coding regions, which can change expression levels or silence genes, often without severe detrimental effects (ii) insertion or deletion in promoters, resulting in the formation of novel promoters (iii) alteration of the methylation status of pre-existing TEs in promoters and (iv) insertion upstream or downstream of genes, modifying their transcription (Hirsch & Springer, 2017). Although the underlying processes by which these mechanisms operate during pathogen interactions are still not fully understood, several examples demonstrate that pathogen-induced changes in TE methylation play essential roles in regulating defence responses (Dowen et al., 2012; Yu et al., 2013). For example, in *Arabidopsis*, the *rdd* (ros1 dml2 dml3) triple demethylase mutant shows increased susceptibility to *Fusarium oxysporum*, associated with hypermethylated TEs in the promoters of downregulated stress-response genes (Le et al., 2014). In rice, the methylation status of TEs in the promoters of resistance genes can determine disease resistance and yield losses. The *Pigm* locus, which confers broad-spectrum resistance to rice blast disease, includes NLR genes such as *PigmR* and *PigmS*. In leaves, highly methylated MITEs in the *PigmS* promoter silence its expression through the RdDM pathway, while in pollen, *PigmS* remains active. This tissue-specific methylation pattern balances strong resistance with reduced yield loss (Deng et al., 2017). Therefore, the methylation state of TEs can be a critical factor in the coordinated regulation of multiple genes. Managing TE methylation patterns may represent a promising strategy in breeding programs to enhance desirable agronomic traits while minimizing undesirable effects.

3. Aim of the thesis

This thesis aimed to identify epigenetic variation influencing cassava responses to African cassava mosaic virus (ACMV), to generate genomic and epigenetic insights to support future breeding strategies for more resilient, farmer-preferred cultivars. We hypothesised that DNA methylation profiles differ between susceptible and tolerant cassava genotypes, and that ACMV infection induces genotype-specific changes in DNA methylation patterns.

Specific objectives were to:

Study I: To generate a high-quality genome assembly of the cassava genotype TMEB117, susceptible to African cassava mosaic virus, and use it as a reference for CMD study.

Study II: To explore structural variation in the TMEB117 genome. We identified a previously unreported large insertion on chromosome 12. This region was further analysed through genome-wide comparisons across 16 cassava genotypes to assess its presence and variability in other cultivars.

Study III: To compare genome-wide DNA methylation patterns between TMEB117 and TMEB693 under uninfected and infected conditions, and to identify differentially methylated regions (DMRs) in genomic features such as promoters, gene bodies, and transposable elements, potentially indicating genes regulated by methylation that contribute to the plant's defence mechanisms.

4. Summary of the papers

This chapter summarises the methods and results of paper I-III. More elaborate explanation can be found on the corresponding papers.

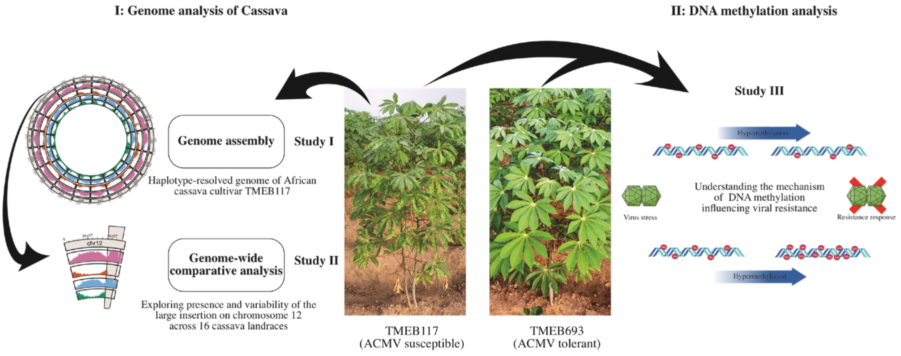

I: Genome analysis of Cassava
II: DNA methylation analysis
Genome assembly
Study I
Haplotype-resolved genome of African cassava cultivar TMEB117
chr12
Genome-wide comparative analysis
Study II
Exploring presence and variability of the large insertion on chromosome 12 across 16 cassava landraces
TMEB117 (ACMV susceptible)
TMEB693 (ACMV tolerant)
Study III
Hypomethylation
Me
Virus stress
Understanding the mechanism of DNA methylation influencing viral resistance
Resistance response
Hypermethylation

Figure 5 Overview of the study in this thesis from study I-III

4.1 Paper I

The paper presents a high-quality, chromosome-scaled, haplotype-resolved genome assembly of the TMEB117 genotype. Cassava plants of the TMEB117 genotype were obtained from the IITA GenBank (Paliwal et al., 2021) and grown in a controlled screen house. Genomic DNA was extracted from young, fully expanded leaves using a modified CTAB protocol. Sequencing was performed using the PacBio Sequel II platform, generating long, HiFi reads from two single-molecule real-time cells. Raw reads were filtered and cleaned using Fastp and HiFiAdapterFilt (Sim et al., 2022) (Chen et al., 2018), with additional quality control steps triggered by abnormal peaks in the GC content of the raw reads, prompting further analysis. Using read mapping and BLAST analysis, we identified fungal contamination from *Alternaria alternata* in some reads. These contaminant reads were subsequently removed. De novo genome assembly was carried out using the hifiasm assembler (Cheng et al., 2021), and contigs were ordered and scaffolded using RagTag (Alonge et al., 2022) with the TME204 genome as a reference. The assembly resulted in two haplotypes: Hap1, with a total length of 694 Mbp and a contig N50 of 18.6 Mbp, and Hap2, with 665 Mbp and a contig N50 of 17.3 Mbp. After scaffolding, the chromosome-level N50 were 37.6 Mbp and 35.7 Mbp for Hap1 and Hap2, respectively. The quality and completeness of the TMEB117 genome assembly was confirmed by both Merqury and BUSCO assessments (Rhie et al., 2020; Simao et al., 2015). Each haplotype achieved 98.9% completeness, with quality value (QV) scores exceeding 64, an indication of the assembly's high accuracy. Analysis of repetitive elements using the EDTA pipeline (Ou et al., 2019) showed that transposable elements (TEs) occupy a substantial portion of the genome, 65.3% in Hap1 and 60.3% in Hap2. This repeat landscape is consistent with previous findings in other sequenced African cassava genotypes. Gene prediction and annotation were carried out using the Funannotate pipeline (Palmer & Stajich, 2020), integrating RNA-seq data, *ab initio* prediction tools, and protein homology. The annotation resulted in 47,138 gene models for Hap1 and 49,163 for Hap2. BUSCO analysis of the predicted protein sequences revealed 90% completeness, supporting the overall gene annotation quality. Further ortholog analysis using OrthoVenn2 online tool (https://orthovenn2.bioinfotoolkits.net/) compared TMEB117 to other cassava genomes (TME204, TME7, and AM560-2), revealing a large number of shared gene clusters, while also identifying unique gene families specific to TMEB117, emphasizing its genomic distinctiveness. The final genome assembly and annotation provide a fully phased, chromosome-scale reference for cassava. This genomic resource will help investigate, for example, the genetic and epigenetic basis of virus susceptibility, which is the main objective of this thesis, and support future molecular breeding efforts.

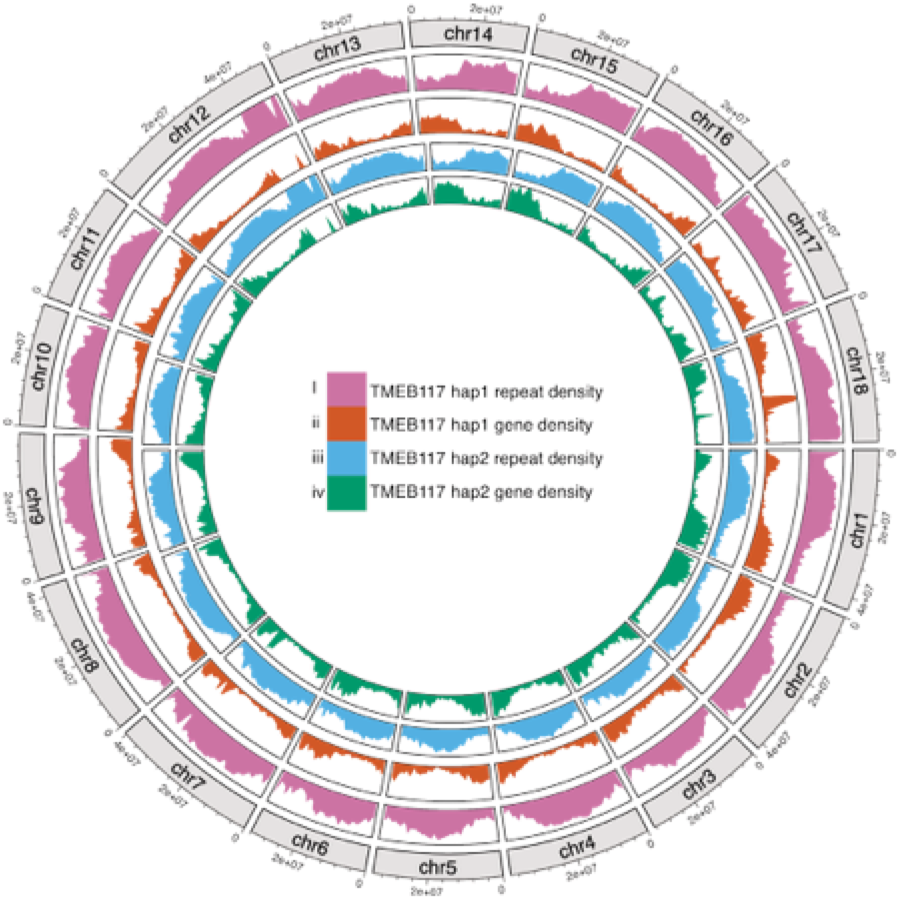


Figure 6 This circos plot illustrates gene and repeat density across the 18 chromosomes of TMEB117 for both haplotypes.

4.2 Paper II

This study identifies a previously unreported 9.7 Mbp highly repetitive insertion on chromosome 12 in the cassava genotype TMEB117. The TMEB117 genome was assembled using three tools, HiCanu, Flye, and hifiasm, to confirm that the presence of the insertion region was not an assembly artifact. We sequenced Illumina short reads for TMEB117, TMEB419, and TMEB693 as part of this study, and downloaded short-read data for an additional 13 cassava genotypes from the Sequence Read Archive (SRA). These short-read sequences from 16 cassava cultivars were mapped to chromosome 12 of TMEB117 using BWA and SAMtools (Li & Durbin, 2009; Li et al., 2009) to analyse read coverage within the insertion region. Although the insertion was absent in some final genome assemblies, raw read data suggested that the region was present in all cultivars, albeit with variable read coverage. Some cassava landraces showed full coverage across the insertion, while other cultivars displayed reduced coverage. Unique read mapping further confirmed variation in the within gene features within this region. PCR amplification analysis verified presence of fragments of the insertion region. The region had low gene content with high repeat level (Fig. 7). Gene ontology showed enrichment of two genes HDA14 and SRT2, with their function associated with histone deacetylase activity. TE annotation showed that the region comprises over 90% TEs, with a significant overrepresentation of the *MUDR-Mutator* superfamily. Comparative genome-wide analysis using SyRI (Goel et al., 2019) revealed additional structural variations, including large inversions, translocations, and duplications across TMEB117 and other cassava genomes. These findings suggest that the chromosome 12 insertion is a hypervariable locus contributing to cassava's genomic diversity. The high repeat content of the insertion likely promotes recombination and structural changes, and we highlight the need for future studies using long-read sequencing to resolve this region's complexity fully. This work provides new insights into large-scale structural variation in cassava and its potential implications for genome function and diversity.

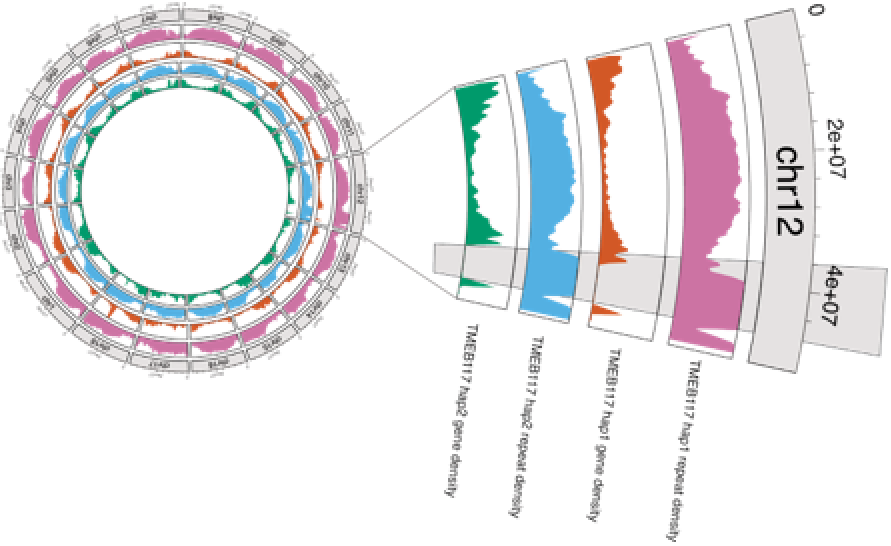


Figure 7 Zoomed-in view of chromosome 12 of the TMEB117 genome, showing high repeat density and low gene density within the insertion region.

4.3 Paper III

The study investigated DNA methylation patterns in two cassava genotypes, TMEB693 (CMD-tolerant) and TMEB117 (CMD-susceptible), to identify DNA methylation variations influencing viral resistance in these genotypes. The study was conducted at the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria. Certified healthy cassava in vitro plants of TMEB117 and TMEB693 genotypes were used as planting material for the study. After six months, virus infection was confirmed, and leaf samples from infected and uninfected plants were collected in three biological replicates per genotype per condition (n = 12 total). Total DNA was extracted using a CTAB-based method with spin-column purification. Genome-wide methylation profiling was performed with enzymatic methyl-seq (EM-seq) using Illumina sequencing. To improve mapping accuracy, a TMEB693 pseudo-reference genome was constructed by aligning TMEB693 short reads to the fully haplotype-resolved TMEB117 genome and replacing reference alleles with high-confidence SNPs and indels identified by both FreeBayes and BCFtools. This yielded ~800,000 SNPs and ~3,000 indels across both haplotypes. Read quality control was conducted using FastQC and fastp, followed by bisulfite-aware mapping (ERNE-BS5) via the EpiDiverse WGBS pipeline. Non-conversion rates were assessed using unmethylated lambda DNA. The TMEB693 infected replicate with poor mapping efficiency (39%) and high non-conversion rate (31.8%) was excluded. Methylation calls were generated with MethylDackel, retaining cytosines with ≥5× coverage and excluding extreme coverage values (>99.9th percentile). Differentially methylated regions (DMRs) were identified using the EpiDiverse DMR pipeline (metilene v2-8) with an FDR ≤ 0.05 and ≥25% methylation difference. Six pairwise comparisons were performed: within-genotype (uninfected vs. infected for TMEB117 and TMEB693) and across-genotype (uninfected vs. infected states in all combinations). Global methylation analysis showed that TMEB693 maintained stable methylation across CpG, CHG, and CHH contexts regardless of infection, while TMEB117 showed a reduction in methylation upon ACMV infection. This loss of methylation in TMEB117 is consistent with the stable methylation levels observed in TMEB693. The largest number of DMRs occurred in the TMEB117 uninfected vs. infected comparison, 100% of the DMRs were hypermethylated in the uninfected state and in CHH context. Comparisons between infected TMEB117 and TMEB693 yielded few DMRs, consistent with infection-induced convergence of methylation profiles. Chromosomal mapping revealed genotype and context-specific DMR clustering, notably dense CpG/CHG DMR blocks on chromosomes 1 and 3 in TMEB117 upon infection, and widespread CHH hypermethylation in uninfected TMEB117. The previously reported large insertion on chromosome 12 lacked DMRs, due to low cytosine coverage in this TE-rich region. Intersection analysis of DMR-associated genes showed that TMEB117 uninfected vs. infected had the highest number of uniquely hypermethylated genes in both promoters and gene bodies, suggesting overall demethylation of these genes during infection. These results support our hypothesis that TMEB117 initiates epigenetic reprogramming in response to viral stress, shifting toward a TMEB693-like methylation profile. However, the reactive effectiveness of this shift may limit its defensive efficiency, whereas TMEB693's stable methylome may reflect a primed defence state. This work provides the first genome-wide comparison of DNA methylation responses to ACMV in contrasting cassava genotypes, offering a foundation for integrative epigenomics and transcriptomic studies to identify early-acting defence regulators and inform breeding strategies for CMD resistance.

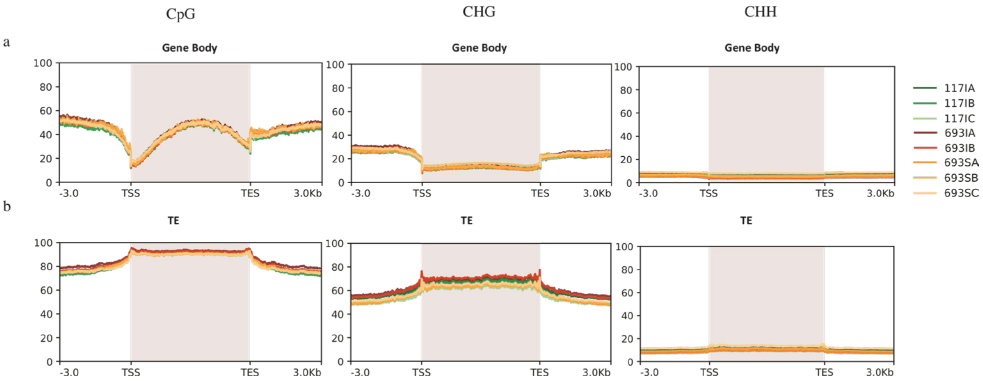


Figure 8 DNA methylation patterns on gene bodies and transposons of infected TMEB117 samples and TMEB693 samples (infected and healthy). TMEB117 infected samples are shown in green, TMEB693 infected samples in red and TMEB693 healthy sample in orange.

5. General Discussion

Plants' ability to withstand biotic and abiotic stress is influenced by epigenetic regulation, with DNA methylation playing a central role in shaping these responses (Springer & Schmitz, 2017; H. Zhang et al., 2018). DNA methylation changes are critical, as rapid and targeted methylation adjustments can trigger defence mechanism, whereas delayed changes may compromise defence effectiveness (Dowen et al., 2012). Although this regulatory layer has been widely studied in model species such as *Arabidopsis* (López Sánchez et al., 2016; Yu et al., 2013) its contribution to disease resistance in some tropical crops remains poorly understood. Cassava, in particular, offers an essential yet underexplored to investigate how DNA methylation dynamics influence pathogen outcomes.

Across Africa, CMD remains one of the most serious constraints to production, with yield losses reported from 20% to as high as 95% in severely affected fields (Legg & Thresh, 2000; Thresh et al., 1997). While some cultivars suffer severe losses, others can tolerate infection with relatively little yield reduction (Legg et al., 2015). GBS and GWAS have revealed essential loci associated with CMD resistance (Rabbi et al., 2014; Rabbi et al., 2022; Wolfe et al., 2019; S. Zhang et al., 2018), but breeding efforts have largely focused on the genetic basis of resistance. In contrast, the role of epigenetic regulation, including DNA methylation, to CMD outcomes in cassava remains poorly understood, with only limited research addressing this aspect.

To address this, we focused on the contrasting CMD responses of two African cassava cultivars: TMEB693 (tolerant) and TMEB117 (susceptible). Our hypothesis was that virus infection triggers DNA methylation reprogramming in susceptible plants, shifting their epigenetic landscape toward that of tolerant genotypes. We first generated a haplotype-resolved, chromosome-scale genome for TMEB117 to serve as the reference for DNA methylation analysis. The assembly achieved exceptional quality (QV > 64, N50 > 35 Mbp, 98.9% BUSCO). The genome was highly repetitive, with >60% of the genome repetitive and containing over 45,000 predicted protein-coding genes per haplotype. Using a genotype specific reference for our

DNA methylation analysis ensured we minimized mapping biases that is often observed between genotypes and calling methylation differences reflected true biological variation rather than artefacts of sequence divergence.

During the genome assembly, we identified an unexpectedly large structural feature: a ~9.7 Mbp insertion on chromosome 12 absent from earlier cassava genome assemblies. This TE-rich block (>90% repeats, enriched for *MUDR-Mutator* elements) harbours two histone deacetylase genes (HDA14, SRT2) associated with chromatin regulation. Comparative analysis across 16 cassava genotypes revealed the insertion to be widespread but variable in coverage, marking it as a hypervariable locus. The study confirmed its presence in some CMD-susceptible and CMD-tolerant cultivars, indicating that it is not directly linked to CMD resistance. However, its density of transposable elements, particularly those of the *MUDR-Mutator* superfamily, and the presence of genes that regulate chromatin structure suggest that this region could influence some unknown epigenetic responsiveness. The large insertion can be considered a valuable resource for pan-genome analysis of the cassava genome, offering insights into its unique features, including its high degree of heterozygosity, and providing a starting point for exploring the potential interplay between transposable elements, chromatin state, and trait variation.

Using the TMEB117 genome, we characterized DNA methylation changes following ACMV infection. TMEB117 depicted extensive CHH hypermethylation across gene bodies and promoter regions in the uninfected state. Following infection, the cultivar underwent substantial demethylation in CpG, CHG, and CHH contexts, with its methylation profile converging toward that of the tolerant TMEB693. While this shift suggests virus-induced epigenetic reprogramming, its role and effectiveness relative to defence activation remain to be defined. In contrast, TMEB693 displayed stable methylation profiles under both infected and uninfected conditions, a pattern that may reflect a more stable epigenetic state, though this too requires experimental validation.

6. Concluding Remarks

We provide a baseline understanding of DNA methylation variation in two contrasting cassava genotypes (tolerant and susceptible) to better explain how epigenetic regulation influences ACMV infection outcomes. By generating a high-quality, haplotype-resolved genome assembly for the susceptible cultivar TMEB117, we deliver a reference for ACMV methylome studies and expand the genomic resources available for cassava research. Comparative genome analysis uncovered a ~9.7 Mbp repeat-rich insertion on chromosome 12, enriched in transposable elements and chromatin-regulating genes, variably present across genotypes and potentially influencing epigenetic plasticity. This structural feature represents a valuable addition to cassava pan-genome resources and a foundation for future studies aimed at understanding its role in crop adaptation. Genome-wide methylation profiling revealed that TMEB117 starts from a highly methylated state and undergoes extensive demethylation only after infection, reaching the stable methylation profile of the tolerant cultivar TMEB693. This shift and reduced methylation level of TMEB117 may be too low to trigger effective resistant to the virus. These findings demonstrate that CMD resistance is shaped not only by genetic architecture but also by epigenetic regulation. The genomic and methylation resources developed here provide a platform for functional studies linking structural variants, epigenetic markers, and gene expression to disease outcomes, offering targets for molecular breeding and accelerating the development of cassava cultivars with more durable CMD resistance and broader stress resilience.

7. Future Perspectives

TMEB117 and TMEB693 are farmer-preferred cassava cultivars in Africa, with TMEB693 also known for its high yield. In contrast, TMEB117's yield potential has been significantly reduced by its high susceptibility to cassava mosaic disease (CMD). Understanding the molecular basis of these contrasting responses is critical for preserving the agronomic value of both cultivars while enhancing disease resilience. This study provides genomic and epigenetic resources that begin to unravel how DNA methylation influences CMD outcomes. The next step is a transcriptome analysis of TMEB117 and TMEB693 across different infection time points. This will confirm the hypothesis that DNA demethylation in TMEB117 occurs upon ACMV infection and identify the specific genes whose expression is altered in this response. Linking these gene expression changes to methylation patterns will enable the identification of methylation-regulated genes that may influence CMD tolerance. TMEB693, in which such genes are expressed under both infected and uninfected conditions, will serve as a valuable control for identifying defence-related pathways. Building on this, investigating the small RNA expression profiles for cassava plants selected for CMD resistance and correlating these patterns with DMRs will provide insights into the regulatory mechanisms driving epigenetic changes. Further correlations between DMRs and the molecular functionalities they may affect could highlight key biological processes linked to CMD defence. Finally, these combined analyses could lead to the identification and proposal of molecular and functional markers for testing in breeding environments. In addition to breeding efforts, exploring CRISPR/Cas9 technology to demethylate/methylate regions within these targets to trigger CMD resistance in farmer-preferred cassava varieties that are currently susceptible. This approach could provide a direct path toward protecting high-value cultivars while preserving traits important to farmers.

To maximise the utility of this work, a public repository of the generated data could be developed, and existing public visualisation platforms should be populated with these new results to support open science and facilitate global cassava research efforts. Incorporating these molecular insights into breeding programs could accelerate the development of CMD-resistant cassava varieties that retain farmer-preferred traits, reduce yield losses, safeguard rural incomes, and strengthen food security for the millions who rely on cassava as a dietary staple in Africa.

References

A.H.M. Mahbubur Rahman, M. A. (2013). Taxonomy and Medicinal Uses of Euphorbiaceae (Spurge) Family of Rajshahi, Bangladesh. *Research in Plant Sciences,, Vol. 1, No. 3, 74-80*. https://doi.org/10.12691/plant-1-3-5

Agorio, A., & Vera, P. (2007). ARGONAUTE4 is required for resistance to Pseudomonas syringae in Arabidopsis. *The Plant Cell*, *19*(11), 3778-3790.

Ahmad, A., Dong, Y., & Cao, X. (2011). Characterization of the PRMT gene family in rice reveals conservation of arginine methylation. *PLoS ONE*, *6*(8), e22664. https://doi.org/10.1371/journal.pone.0022664

Akano, A., Dixon, A., Mba, C., Barrera, E., & Fregene, M. (2002). Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. *Theoretical and Applied Genetics*, *105*(4), 521-525.

Akimoto, K., Katakami, H., Kim, H.-J., Ogawa, E., Sano, C. M., Wada, Y., & Sano, H. (2007). Epigenetic inheritance in rice plants. *Annals of botany*, *100*(2), 205-217.

Akinpelu, A., Amamgbo, L., Olojede, A., & Oyekale, A. (2011). Health implications of cassava production and consumption. *Journal of Agriculture and Social Research (JASR)*, *11*(1).

Alhoraibi, H., Bigeard, J., Rayapuram, N., Colcombet, J., & Hirt, H. (2018). Plant immunity: the MTI-ETI model and beyond.

Alicai, T., Szyniszewska, A. M., Omongo, C. A., Abidrabo, P., Okao-Okuja, G., Baguma, Y., Ogwok, E., Kawuki, R., Esuma, W., & Tairo, F. (2019). Expansion of the cassava brown streak pandemic in Uganda revealed by annual field survey data for 2004 to 2017. *Scientific Data*, *6*(1), 327.

Allem, A. C. (1994). The origin of Manihot esculenta Crantz (Euphorbiaceae). *Genetic Resources and Crop Evolution 41:133-150, 1994.*

Allem, A. C. (1999). The closest wild relatives of cassava (Manihot esculenta Crantz). *Euphytica*, *107: 123–133, 1999.*

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome biology*, *23*(1), 258.

Amelework, A. B., & Bairu, M. W. (2022). Advances in Genetic Analysis and Breeding of Cassava (Manihot esculenta Crantz): A Review. *Plants (Basel)*, *11*(12). https://doi.org/10.3390/plants11121617

Andrade, L. R. B. d., Sousa, M. B. e., Oliveira, E. J., Resende, M. D. V. d., & Azevedo, C. F. (2019). Cassava yield traits predicted by genomic selection methods. *PLoS ONE*, *14*(11), e0224920.

Arora, H., Singh, R. K., Sharma, S., Sharma, N., Panchal, A., Das, T., Prasad, A., & Prasad, M. (2022). DNA methylation dynamics in response to abiotic and pathogen stress in plants. *Plant cell reports*, *41*(10), 1931-1944.

Ashapkin, V. V., Kutueva, L. I., Aleksandrushkina, N. I., & Vanyushin, B. F. (2020). Epigenetic mechanisms of plant adaptation to biotic and abiotic stresses. *International Journal of Molecular Sciences*, *21*(20), 7457.

Balagopalan, C. (2018). *Cassava in food, feed and industry*. CRC press.

Bewick, A. J., & Schmitz, R. J. (2017). Gene body DNA methylation in plants. *Current opinion in plant biology*, *36*, 103-110.

Biémont, C., & Vieira, C. (2006). Junk DNA as an evolutionary force. *Nature*, *443*(7111), 521-524.

Bock, K., Guthrie, E., & FIGUEIREDO, G. (1981). A strain of cassava latent virus occurring in coastal districts of Kenya. *Annals of Applied Biology*, *99*(2), 151-159.

Bredeson, J. V., Lyons, J. B., Prochnik, S. E., Wu, G. A., Ha, C. M., Edsinger-Gonzales, E., Grimwood, J., Schmutz, J., Rabbi, I. Y., Egesi, C., Nauluvula, P., Lebot, V., Ndunguru, J., Mkamilo, G., Bart, R. S., Setter, T. L., Gleadow, R. M., Kulakow, P., Ferguson, M. E., . . . Rokhsar, D. S. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat Biotechnol*, *34*(5), 562-570. https://doi.org/10.1038/nbt.3535

Bredeson, J. V., Shu, S., Berkoff, K., L., J.B., Caccamo, M., Santos, B., Ovalle, T., B., R.S., Lopez-Lavalle, A. B., L., C. Y., M., A., E., Wenzl, P., J., J.-L., D., S., & Rokhsar, D. S. (2021). An improved reference assembly for cassava (Manihot esculenta Crantz)".In preparation.

.

Buchmann, R. C., Asad, S., Wolf, J. N., Mohannath, G., & Bisaro, D. M. (2009). Geminivirus AL2 and L2 proteins suppress transcriptional gene silencing and cause genome-wide reductions in cytosine methylation. *Journal of virology*, *83*(10), 5005-5013.

Carmo, C. D. d., Silva, M. S. d., Oliveira, G. A. F., & Oliveira, E. J. d. (2015). Molecular-assisted selection for resistance to cassava mosaic disease in Manihot esculenta Crantz. *Scientia Agricola*, *72*(6), 520-527.

Carter, S. E. F., Louise O; Jones, Peter G; Fairbairn, James N. (1992). *An atlas of cassava in Africa : Historical, agroecological and demographic aspects of crop distribution* . . Centro Internacional de Agricultura Tropical (CIAT), Cali.

Ceballos, H., Rojanaridpiched, C., Phumichai, C., Becerra, L. A., Kittipadakul, P., Iglesias, C., & Gracen, V. E. (2020). Excellence in cassava breeding: perspectives for the future. *Crop Breeding, Genetics and Genomics*, *2*(2).

Chan, S. W.-L., Henderson, I. R., & Jacobsen, S. E. (2005). Gardening the genome: DNA methylation in Arabidopsis thaliana. *Nature Reviews Genetics*, *6*(5), 351-360.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*, *18*(2), 170-175. https://doi.org/10.1038/s41592-020-01056-5

Choi, C.-S., & Sano, H. (2007). Abiotic-stress induces demethylation and transcriptional activation of a gene encoding a glycerophosphodiesterase-like protein in tobacco plants. *Molecular Genetics and Genomics*, *277*, 589-600.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., & Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, *452*(7184), 215-219.

Consortium, I. C. G. M. (2015). High-resolution linkage map and chromosome-scale genome assembly for cassava (Manihot esculenta Crantz) from 10 populations. *G3: Genes, Genomes, Genetics*, *5*(1), 133-144.

De Carvalho, R., & Guerra, M. (2002). Cytogenetics of Manihot esculenta Crantz (cassava) and eight related species. *Hereditas*, *136*(2), 159-168.

Deng, Y., Zhai, K., Xie, Z., Yang, D., Zhu, X., Liu, J., Wang, X., Qin, P., Yang, Y., & Zhang, G. (2017). Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. *Science*, *355*(6328), 962-965.

Dowen, R. H., Pelizzola, M., Schmitz, R. J., Lister, R., Dowen, J. M., Nery, J. R., Dixon, J. E., & Ecker, J. R. (2012). Widespread dynamic DNA methylation in response to biotic stress. *Proceedings of the National Academy of Sciences*, *109*(32), E2183-E2191.

Du, B., Haensch, R., Alfarraj, S., & Rennenberg, H. (2024). Strategies of plants to overcome abiotic and biotic stresses. *Biological Reviews*, *99*(4), 1524-1536.

Elhamamsy, A. R. (2016). DNA methylation dynamics in plants and mammals: overview of regulation and dysregulation. *Cell Biochemistry and Function*, *34*(5), 289-298.

Elias, A. A., Rabbi, I., Kulakow, P., & Jannink, J.-L. (2018). Improving genomic prediction in cassava field experiments using spatial analysis. *G3: Genes, Genomes, Genetics*, *8*(1), 53-62.

Esuma, W., Ozimati, A., Kulakow, P., Gore, M. A., Wolfe, M. D., Nuwamanya, E., Egesi, C., & Kawuki, R. S. (2021). Effectiveness of genomic selection for improving provitamin A carotenoid content and associated traits in cassava. *G3*, *11*(9), jkab160.

Fanou, A. A., Zinsou, V. A., & Wydra, K. (2017). Cassava bacterial blight: A devastating disease of cassava. In *Cassava*. IntechOpen.

FAOSTAT. (2025). *Statistical database of the Food and Agriculture Organisation of the United Nations.* http://www.fao.org/faostat/en/#home

Fauquet, C., & Fargette, D. (1990). African cassava mosaic virus: etiology, epidemiology and control. *Plant Dis*, *74*(6), 404-411.

Fauquet, C. M., Mayo, M. A., Maniloff, J., Desselberger, U., & Ball, L. A. (2005). *Virus taxonomy: VIIIth report of the International Committee on Taxonomy of Viruses*. Academic Press.

Fondong, V. N. (2017). The search for resistance to cassava mosaic geminiviruses: how much we have accomplished, and what lies ahead. *Frontiers in Plant Science*, *8*, 408.

Forestan, C., Aiese Cigliano, R., Farinati, S., Lunardon, A., Sanseverino, W., & Varotto, S. (2016). Stress-induced and epigenetic-mediated maize transcriptome regulation study by means of transcriptome reannotation and differential expression analysis. *Scientific Reports*, *6*(1), 30446.

Fregene, M., Puonti-Kaerlas, J., Hillocks, R., & Thresh, J. (2002). Cassava: Biology, production and utilization. *Cassava Biotechnology*, 179-207.

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, *3*(9), 722-732.

Gallego-Bartolomé, J. (2020). DNA methylation in plants: mechanisms and tools for targeted manipulation. *New Phytologist*, *227*(1), 38-44.

Gayacharan, & Joel, A. J. (2013). Epigenetic responses to drought stress in rice (Oryza sativa L.). *Physiology and Molecular Biology of Plants*, *19*, 379-387.

Gegios, A., Amthor, R., Maziya-Dixon, B., Egesi, C., Mallowa, S., Nungo, R., Gichuki, S., Mbanaso, A., & Manary, M. J. (2010). Children consuming cassava as a staple food are at risk for inadequate zinc, iron, and vitamin A intake. *Plant Foods for Human Nutrition*, *65*, 64-70.

Goel, M., Sun, H., Jiao, W.-B., & Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome biology*, *20*(1), 277.

Gong, Z., Morales-Ruiz, T., Ariza, R. R., Roldán-Arjona, T., David, L., & Zhu, J.-K. (2002). ROS1, a repressor of transcriptional gene silencing in Arabidopsis, encodes a DNA glycosylase/lyase. *Cell*, *111*(6), 803-814.

González, R. M., Ricardi, M. M., & Iusem, N. D. (2011). Atypical epigenetic mark in an atypical location: cytosine methylation at asymmetric (CNN) sites within the body of a non-repetitive tomato gene. *BMC Plant Biology*, *11*, 1-11.

Grandbastien, M.-A., Audeon, C., Bonnivard, E., Casacuberta, J., Chalhoub, B., Costa, A.-P., Le, Q. H., Melayah, D., Petit, M., & Poncet, C. (2005). Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenetic and genome research*, *110*(1-4), 229-241.

Henderson, I. R., & Jacobsen, S. E. (2007). Epigenetic inheritance in plants. *Nature*, *447*(7143), 418-424.

Hillocks, R., & Thresh, J. (2000). Cassava mosaic and cassava brown streak virus diseases in Africa: a comparative guide to symptoms and aetiologies. *Roots*, *7*(1), 1-8.

Hirsch, C. D., & Springer, N. M. (2017). Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1860*(1), 157-165.

Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., & Steiner, C. C. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, *7*(1), 399.

Hu, W., Ji, C., Liang, Z., Ye, J., Ou, W., Ding, Z., Zhou, G., Tie, W., Yan, Y., & Yang, J. (2021). Resequencing of 388 cassava accessions identifies valuable loci and selection for variation in heterozygosity. *Genome biology*, *22*, 1-23.

Jennings, D. (1963). Variation in pollen and ovule fertility in varieties of cassava, and the effect of interspecific crossing on fertility. *Euphytica*, *12*(1), 69-76.

Jennings D.L., I. C. (2002). Breeding for crop improvement. In T. A. M. Hillocks R.J., Bellotti A.C. (Ed.), *Cassava: Biology, Production and Utilization.* (pp. 149–166). CABI; Wallingford, UK.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, *13*(7), 484-492.

Jones, W. O. (1959). *Manioc in Africa.* Stanford University Press, Stanford, California.

Juma, B. S., Mukami, A., Mweu, C., Ngugi, M. P., & Mbinda, W. (2022). Targeted mutagenesis of the CYP79D1 gene via CRISPR/Cas9-mediated genome editing results in lower levels of cyanide in cassava. *Frontiers in Plant Science*, *13*, 1009860.

Kalyebi, A., Macfadyen, S., Parry, H., Tay, W. T., De Barro, P., & Colvin, J. (2018). African cassava whitefly, Bemisia tabaci, cassava colonization preferences and control implications. *PLoS ONE*, *13*(10), e0204862.

Kawuki, R., Nuwamanya, E., Labuschagne, M., Herselman, L., & Ferguson, M. (2010). Genetic effects of inbreeding on harvest index and root dry matter content in cassava.

Kayondo, S. I., Pino Del Carpio, D., Lozano, R., Ozimati, A., Wolfe, M., Baguma, Y., Gracen, V., Offei, S., Ferguson, M., & Kawuki, R. (2018). Genome-wide association mapping and genomic prediction for CBSD resistance in Manihot esculenta. *Scientific Reports*, *8*(1), 1549.

Khandare, V., & Choomsook, P. (2019). Cassava export of Thailand: Growth performance and composition. *Int. J. Res. Anal. Rev*, *6*, 847-857.

Kim, J., Samaranayake, M., & Pradhan, S. (2009). Epigenetic mechanisms in mammals. *Cellular and molecular life sciences*, *66*, 596-612.

Kim, K. D., El Baidouri, M., Abernathy, B., Iwata-Otsubo, A., Chavarro, C., Gonzales, M., Libault, M., Grimwood, J., & Jackson, S. A. (2015). A comparative epigenomic analysis of polyploidy-derived genes in soybean and common bean. *Plant physiology*, *168*(4), 1433-1447.

Kim, M. Y., & Zilberman, D. (2014). DNA methylation as a system of plant genomic immunity. *Trends in plant science*, *19*(5), 320-326.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, *27*(5), 722-736.

Koros, J., Runo, S., Yusuf, M., & Orek, C. O. (2018). Screening selected cassava cultivars for resistance against cassava viruses and cassava green mites under advanced yield trials in Kenya. *IOSR J. Biotechnol. Biochem*, *4*, 37-52.

Kuon, J.-E., Qi, W., Schläpfer, P., Hirsch-Hoffmann, M., von Bieberstein, P. R., Patrignani, A., Poveda, L., Grob, S., Keller, M., & Shimizu-Inatsugi, R. (2019). Haplotype-resolved genomes of geminivirus-resistant and geminivirus-susceptible African cassava cultivars. *BMC biology*, *17*, 1-15.

Landi, M., Shah, T., Falquet, L., Niazi, A., Stavolone, L., Bongcam-Rudloff, E., & Gisel, A. (2023). Haplotype-resolved genome of heterozygous African cassava cultivar TMEB117 (Manihot esculenta). *Sci Data*, *10*(1), 887. https://doi.org/10.1038/s41597-023-02800-0

Lang, Z., Wang, Y., Tang, K., Tang, D., Datsenka, T., Cheng, J., Zhang, Y., Handa, A. K., & Zhu, J.-K. (2017). Critical roles of DNA demethylation in the activation of ripening-induced genes and inhibition of ripening-repressed genes in tomato fruit. *Proceedings of the National Academy of Sciences*, *114*(22), E4511-E4519.

Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, *11*(3), 204-220.

Le, T.-N., Schumann, U., Smith, N. A., Tiwari, S., Au, P. C. K., Zhu, Q.-H., Taylor, J. M., Kazan, K., Llewellyn, D. J., & Zhang, R. (2014). DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in Arabidopsis. *Genome biology*, *15*(9), 458.

Lee, S., Choi, J., Park, J., Hong, C. P., Choi, D., Han, S., Choi, K., Roh, T.-Y., Hwang, D., & Hwang, I. (2023). DDM1-mediated gene body DNA methylation is associated with inducible activation of defense-related genes in Arabidopsis. *Genome biology*, *24*(1), 106.

Legg, J. P., Kumar, P. L., Makeshkumar, T., Tripathi, L., Ferguson, M., Kanju, E., Ntawuruhunga, P., & Cuellar, W. (2015). Cassava virus diseases: biology, epidemiology, and management. In *Advances in virus research* (Vol. 91, pp. 85-142). Elsevier.

Legg, J. P., & Thresh, J. (2000). Cassava mosaic virus disease in East Africa: a dynamic disease in a changing environment. *Virus research*, *71*(1-2), 135-149.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, X., Zhu, J., Hu, F., Ge, S., Ye, M., Xiang, H., Zhang, G., Zheng, X., Zhang, H., & Zhang, S. (2012). Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC genomics*, *13*, 1-15.

Li, Y., Lu, Y.-G., Shi, Y., Wu, L., Xu, Y.-J., Huang, F., Guo, X.-Y., Zhang, Y., Fan, J., & Zhao, J.-Q. (2014). Multiple rice microRNAs are involved in immunity against the blast fungus Magnaporthe oryzae. *Plant physiology*, *164*(2), 1077-1092.

Li, Y., Xia, Q., Kou, H., Wang, D., Lin, X., Wu, Y., Xu, C., Xing, S., & Liu, B. (2011). Induced Pib Expression and Resistance to Magnaporthe grisea are Compromised by Cytosine Demethylation at Critical Promoter Regions in Rice. *J Integr Plant Biol*, *53*(10), 814-823. https://doi.org/10.1111/j.1744-7909.2011.01070.x

López, M.-E., Roquis, D., Becker, C., Denoyes, B., & Bucher, E. (2022). DNA methylation dynamics during stress response in woodland strawberry (Fragaria vesca). *Horticulture research*, *9*, uhac174.

López Sánchez, A., Stassen, J. H., Furci, L., Smith, L. M., & Ton, J. (2016). The role of DNA (de) methylation in immune responsiveness of Arabidopsis. *The Plant Journal*, *88*(3), 361-374.

Lu, X., Liu, Y., Xu, J., Liu, X., Chi, Y., Li, R., Mo, L., Shi, L., Liang, S., & Yu, W. (2024). Recent progress of molecular mechanisms of DNA methylation in plant response to abiotic stress. *Environmental and Experimental Botany*, *218*, 105599.

Lucibelli, F., Valoroso, M. C., & Aceto, S. (2022). Plant DNA methylation: an epigenetic mark in development, environmental interactions, and evolution. *International Journal of Molecular Sciences*, *23*(15), 8299.

Luna, E., Bruce, T. J., Roberts, M. R., Flors, V., & Ton, J. (2012). Next-generation systemic acquired resistance. *Plant physiology*, *158*(2), 844-853.

Lyons, J. B., Bredeson, J. V., Mansfeld, B. N., Bauchet, G. J., Berry, J., Boyher, A., Mueller, L. A., Rokhsar, D. S., & Bart, R. S. (2022). Current status and impending progress for cassava structural genomics. *Plant Mol Biol*, *109*(3), 177-191. https://doi.org/10.1007/s11103-020-01104-w

Mansfeld, B. N., Boyher, A., Berry, J. C., Wilson, M., Ou, S., Polydore, S., Michael, T. P., Fahlgren, N., & Bart, R. S. (2021). Large structural variations in the haplotype-resolved African cassava genome. *The Plant Journal*, *108*(6), 1830-1848.

McKey, D., Elias, M., Pujol, B., & Duputié, A. (2010). The evolutionary ecology of clonally propagated domesticated plants. *New Phytologist*, *186*(2), 318-332.

Nabi, Z., Manzoor, S., Nabi, S. U., Wani, T. A., Gulzar, H., Farooq, M., Arya, V. M., Baloch, F. S., Vlădulescu, C., & Popescu, S. M. (2024). Pattern-Triggered Immunity and Effector-Triggered Immunity: crosstalk and cooperation of PRR and NLR-mediated plant defense pathways during host–pathogen interactions. *Physiology and Molecular Biology of Plants*, *30*(4), 587-604.

Nassar N.M.A., O. R. (2009). Cassava genetic resources: Manipulation for crop improvement. *Plant Breeding Reviews*. https://doi.org/https://doi.org/10.1002/9780470593783.ch5

Nawaz, M., Sun, J., Shabbir, S., Khattak, W. A., Ren, G., Nie, X., Bo, Y., Javed, Q., Du, D., & Sonne, C. (2023). A review of plants strategies to resist biotic and abiotic environmental stressors. *Science of The Total Environment*, *900*, 165832.

Ndunguru, J., Legg, J. P., Fofana, I., Aveling, T., Thompson, G., & Fauquet, C. M. (2006). Identification of a defective molecule derived from DNA-A of the bipartite begomovirus of East African cassava mosaic virus. *Plant Pathology*, *55*(1), 2-10.

Nzuki, I., Katari, M. S., Bredeson, J. V., Masumba, E., Kapinga, F., Salum, K., Mkamilo, G. S., Shah, T., Lyons, J. B., & Rokhsar, D. S. (2017). QTL mapping for pest and disease resistance in cassava and coincidence of some QTL with introgression regions derived from Manihot glaziovii. *Frontiers in Plant Science*, *8*, 1168.

Okogbenin, E., Egesi, C., Olasanmi, B., Ogundapo, O., Kahya, S., Hurtado, P., Marin, J., Akinbo, O., Mba, C., & Gomez, H. (2012). Molecular marker analysis and validation of resistance to cassava mosaic disease in elite cassava genotypes in Nigeria. *Crop Science*, *52*(6), 2576-2586.

Okogbenin, E., Moreno, I., Tomkins, J., Fauquet, C., Mkamilo, G., & Fregene, M. (2013). Marker-assisted breeding for cassava mosaic disease resistance. *Translational genomics for crop breeding: biotic stress*, *1*, 291-325.

Olsen, K. M. (2004). SNPs, SSRs and inferences on cassava's origin. *Plant Molecular Biology, 56: 517–526.*

Olsen, K. M., & Schaal, B. A. (2001). Microsatellite variation in cassava (Manihot esculenta, Euphorbiaceae) and its wild relatives: further evidence for a southern Amazonian origin of domestication. *American Journal of Botany*, *88*(1), 131-142. https://doi.org/10.2307/2657133

Olsen, K. M. S., B. A. (1999). Evidence on the origin of cassava: Phylogeography of Manihot esculenta. *Proc. Natl. Acad. Sci. USA, 96, pp. 5586–5591.*

Otekunrin, O. A., & Sawicka, B. (2019). Cassava, a 21st century staple crop: How can Nigeria harness its enormous trade potentials. *Acta Scientific Agriculture*, *3*(8), 194-202.

Otim-Nape, G., Shaw, M., & Thresh, J. (1994). The effects of African cassava mosaic geminivirus on the growth and yield of cassava in Uganda.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., & Peterson, T. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology*, *20*, 1-18.

Paliwal, R., Adegboyega, T. T., Abberton, M., Faloye, B., & Oyatomi, O. (2021). Potential of genomics for the improvement of underutilized legumes in sub-Saharan Africa. *Legume Science*, *3*(3), e69.

Palmer, J. M., & Stajich, J. (2020). *Funannotate v1.8.1: Eukaryotic genome annotation (v1.8)* https://doi.org/https://doi.org/10.5281/zenodo.1134477

Parmar, A., Sturm, B., & Hensel, O. (2017). Crops that feed the world: Production and improvement of cassava for food, feed, and industrial uses. *Food Security*, *9*(5), 907-927. https://doi.org/10.1007/s12571-017-0717-8

Patil, B. L., Legg, J. P., Kanju, E., & Fauquet, C. M. (2015). Cassava brown streak disease: a threat to food security in Africa. *Journal of General Virology*, *96*(5), 956-968.

Prochnik, S., Marri, P. R., Desany, B., Rabinowicz, P. D., Kodira, C., Mohiuddin, M., Rodriguez, F., Fauquet, C., Tohme, J., & Harkins, T. (2012). The cassava genome: current progress, future directions. *Tropical plant biology*, *5*, 88-94.

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

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

Rabbi, I. Y., Kayondo, S. I., Bauchet, G., Yusuf, M., Aghogho, C. I., Ogunpaimo, K., Uwugiaren, R., Smith, I. A., Peteti, P., & Agbona, A. (2022). Genome-wide association analysis reveals new insights into the genetic architecture of defensive, agro-morphological and quality-related traits in cassava. *Plant Molecular Biology*, *109*(3), 195-213.

Raja, P., Sanville, B. C., Buchmann, R. C., & Bisaro, D. M. (2008). Viral genome methylation as an epigenetic defense against geminiviruses. *Journal of virology*, *82*(18), 8997-9007.

Ramu, P., Esuma, W., Kawuki, R., Rabbi, I. Y., Egesi, C., Bredeson, J. V., Bart, R. S., Verma, J., Buckler, E. S., & Lu, F. (2017). Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nature genetics*, *49*(6), 959-963.

Regulski, M., Lu, Z., Kendall, J., Donoghue, M. T., Reinders, J., Llaca, V., Deschamps, S., Smith, A., Levy, D., & McCombie, W. R. (2013). The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome research*, *23*(10), 1651-1662.

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*, *21*(1), 245. https://doi.org/10.1186/s13059-020-02134-9

Richard, M. M., Gratias, A., Thareau, V., Kim, K. D., Balzergue, S., Joets, J., Jackson, S. A., & Geffroy, V. (2018). Genomic and epigenomic immunity in common bean: the unusual features of NB-LRR gene family. *DNA Research*, *25*(2), 161-172.

Roa A.C., M. M. M., Duque M.C., Tohme J., Allem A., Bonierbale M.W. (1997). AFLP analysis of relationships among cassava and other Manihot species. *Theor Appl Genet*, *95 : 741—750*.

Roa, A. C., Chavarriaga-Aguirre, P., Duque, M. C., Maya, M. M., Bonierbale, M. W., Iglesias, C., & Tohme, J. (2000). Cross-species amplification of cassava (Manihot esculenta) (Euphorbiaceae) microsatellites: allelic polymorphism and degree of relationship. *American Journal of Botany*, *87*(11), 1647-1655. https://doi.org/10.2307/2656741

Saze, H., Scheid, O. M., & Paszkowski, J. (2003). Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. *Nature genetics*, *34*(1), 65-69.

Schmitz, R. J., & Ecker, J. R. (2012). Epigenetic and epigenomic variation in Arabidopsis thaliana. *Trends in plant science*, *17*(3), 149-154.

Schoft, V. K., Chumak, N., Choi, Y., Hannon, M., Garcia-Aguilar, M., Machlicova, A., Slusarz, L., Mosiolek, M., Park, J.-S., & Park, G. T. (2011). Function of the DEMETER DNA glycosylase in the Arabidopsis thaliana male gametophyte. *Proceedings of the National Academy of Sciences*, *108*(19), 8042-8047.

Silva Filho, J., Pestana, R. K. N., Silva Junior, W. J. D., Coelho Filho, M. A., Ferreira, C. F., de Oliveira, E. J., & Kido, E. A. (2024). Exploiting DNA methylation in cassava under water deficit for crop improvement. *PLoS ONE*, *19*(2), e0296254. https://doi.org/10.1371/journal.pone.0296254

Sim, S. B., Corpuz, R. L., Simmonds, T. J., & Geib, S. M. (2022). HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC genomics*, *23*(1), 157.

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212. https://doi.org/10.1093/bioinformatics/btv351

Singer, T., Yordan, C., & Martienssen, R. A. (2001). Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). *Genes & development*, *15*(5), 591-602.

Song, Q.-X., Lu, X., Li, Q.-T., Chen, H., Hu, X.-Y., Ma, B., Zhang, W.-K., Chen, S.-Y., & Zhang, J.-S. (2013). Genome-wide analysis of DNA methylation in soybean. *Molecular plant*, *6*(6), 1961-1974.

Springer, N. M., & Schmitz, R. J. (2017). Exploiting induced and natural epigenetic variation for crop improvement. *Nature Reviews Genetics*, *18*(9), 563-575.

Stanley, J., & Gay, M. R. (1983). Nucleotide sequence of cassava latent virus DNA. *Nature*, *301*(5897), 260-262.

Stephenson, K., Amthor, R., Mallowa, S., Nungo, R., Maziya-Dixon, B., Gichuki, S., Mbanaso, A., & Manary, M. (2010). Consuming cassava as a staple food places children 2-5 years old at risk for inadequate protein intake, an observational study in Kenya and Nigeria. *Nutrition journal*, *9*, 1-6.

Thresh, J., Otim-Nape, G., Legg, J., & Fargette, D. (1997). African cassava mosaic virus disease: the magnitude of the problem.

Tsaftaris, A. S., Polidoros, A. N., Koumproglou, R., Tani, E., Kovacevic, N., & Abatzidou, E. (2005). Epigenetic mechanisms in plants and their implications in plant breeding. *In the wake of the double helix: from the green revolution to the gene revolution. Avenue Media, Bologna*, 157-171.

Van Speybroeck, L., de Waele, D., & Van de Vijver, G. (2006). Theories in Early Embryology. *Annals of the New York Academy of Sciences*, *981*(1), 7-49. https://doi.org/10.1111/j.1749-6632.2002.tb04910.x

Wada, Y., Miyamoto, K., Kusano, T., & Sano, H. (2004). Association between up-regulation of stress-responsive genes and hypomethylation of genomic DNA in tobacco plants. *Molecular Genetics and Genomics*, *271*, 658-666.

Wang, C., Yang, Q., Wang, W., Li, Y., Guo, Y., Zhang, D., Ma, X., Song, W., Zhao, J., & Xu, M. (2017). A transposon-directed epigenetic change in ZmCCT underlies quantitative resistance to Gibberella stalk rot in maize. *New Phytologist*, *215*(4), 1503-1515.

Wang, H., Beyene, G., Zhai, J., Feng, S., Fahlgren, N., Taylor, N. J., Bart, R., Carrington, J. C., Jacobsen, S. E., & Ausin, I. (2015). CG gene body DNA methylation changes and evolution of duplicated genes in cassava. *Proceedings of the National Academy of Sciences*, *112*(44), 13729-13734.

Wang, J., Marowsky, N. C., & Fan, C. (2014). Divergence of gene body DNA methylation and evolution of plant duplicate genes. *PLoS ONE*, *9*(10), e110357.

Wang, M., Qin, L., Xie, C., Li, W., Yuan, J., Kong, L., Yu, W., Xia, G., & Liu, S. (2014). Induced and constitutive DNA methylation in a salinity-tolerant wheat introgression line. *Plant and Cell Physiology*, *55*(7), 1354-1365.

Wang, W., Feng, B., Xiao, J., Xia, Z., Zhou, X., Li, P., Zhang, W., Wang, Y., Møller, B. L., & Zhang, P. (2014). Cassava genome from a wild ancestor to cultivated varieties. *Nature communications*, *5*(1), 5110.

Wang, W.-S., Pan, Y.-J., Zhao, X.-Q., Dwivedi, D., Zhu, L.-H., Ali, J., Fu, B.-Y., & Li, Z.-K. (2011). Drought-induced site-specific DNA methylation and its association with drought tolerance in rice (Oryza sativa L.). *Journal of experimental botany*, *62*(6), 1951-1960.

Wang, X., Zhang, Z., Fu, T., Hu, L., Xu, C., Gong, L., Wendel, J. F., & Liu, B. (2017). Gene-body CG methylation and divergent expression of duplicate genes in rice. *Scientific Reports*, *7*(1), 2675.

Warburg, O. (1894). *Die kulturpflanzen usambaras*. na.

Wolfe, M. D., Bauchet, G. J., Chan, A. W., Lozano, R., Ramu, P., Egesi, C., Kawuki, R., Kulakow, P., Rabbi, I., & Jannink, J.-L. (2019). Historical introgressions from a wild relative of modern cassava improved important traits and may be under balancing selection. *Genetics*, *213*(4), 1237-1253.

Wossen, T., Girma, G., Abdoulaye, T., Rabbi, I., Olanrewaju, A., Alene, A., Feleke, S., Kulakow, P., Asumugha, G., & Abass, A. (2017). The cassava monitoring survey in Nigeria. *Report International Institute of Tropical Agriculture, Ibadan, Nigeria*.

Yaish, M. W. (2013). DNA methylation-associated epigenetic changes in stress tolerance of plants. In *Molecular stress physiology of plants* (pp. 427-440). Springer.

Yaish, M. W., Al-Lawati, A., Al-Harrasi, I., & Patankar, H. V. (2018). Genome-wide DNA Methylation analysis in response to salinity in the model plant caliph medic (Medicago truncatula). *BMC genomics*, *19*, 1-17.

Yoder, J. A., Walsh, C. P., & Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in genetics*, *13*(8), 335-340.

Yonis, B. O., Pino del Carpio, D., Wolfe, M., Jannink, J.-L., Kulakow, P., & Rabbi, I. (2020). Improving root characterisation for genomic prediction in cassava. *Scientific Reports*, *10*(1), 8003.

You, W., Tyczewska, A., Spencer, M., Daxinger, L., Schmid, M. W., Grossniklaus, U., Simon, S. A., Meyers, B. C., Matzke, A. J., & Matzke, M. (2012). Atypical DNA methylation of genes encoding cysteine-rich peptides in Arabidopsis thaliana. *BMC Plant Biology*, *12*, 1-15.

Yu, A., Lepère, G., Jay, F., Wang, J., Bapaume, L., Wang, Y., Abraham, A.-L., Penterman, J., Fischer, R. L., & Voinnet, O. (2013). Dynamics and biological relevance of DNA demethylation in Arabidopsis antibacterial defense. *Proceedings of the National Academy of Sciences*, *110*(6), 2389-2394.

Zemach, A., Kim, M. Y., Silva, P., Rodrigues, J. A., Dotson, B., Brooks, M. D., & Zilberman, D. (2010). Local DNA hypomethylation activates genes in rice endosperm. *Proceedings of the National Academy of Sciences*, *107*(43), 18729-18734.

Zhang, H., Lang, Z., & Zhu, J.-K. (2018). Dynamics and function of DNA methylation in plants. *Nature reviews Molecular cell biology*, *19*(8), 489-506.

Zhang, S., Chen, X., Lu, C., Ye, J., Zou, M., Lu, K., Feng, S., Pei, J., Liu, C., & Zhou, X. (2018). Genome-wide association studies of 11 agronomic traits in cassava (Manihot esculenta Crantz). *Frontiers in Plant Science*, *9*, 503.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W.-L., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., & Jacobsen, S. E. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, *126*(6), 1189-1201.

Zhong, Z., Feng, S., Mansfeld, B. N., Ke, Y., Qi, W., Lim, Y. W., Gruissem, W., Bart, R. S., & Jacobsen, S. E. (2023). Haplotype-resolved DNA methylome of African cassava genome. *Plant Biotechnol J*, *21*(2), 247-249. https://doi.org/10.1111/pbi.13955

Zilberman, D., Cao, X., & Jacobsen, S. E. (2003). ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science*, *299*(5607), 716-719.

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., & Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet*, *39*(1), 61-69. https://doi.org/10.1038/ng1929

Zimmermann, A. (1906). Die krauselkrankheit des maniok. *Pflanzer*, *2*, 145.

Popular science summary

Cassava is a staple food for hundreds of millions of people in sub-Saharan Africa, thriving in poor soils and under drought, where many other crops fail. However, its production is threatened by cassava mosaic disease, a viral infection that can destroy up to 95% of yields in susceptible varieties. Some cassava varieties fight the disease better than others, but the reasons for this difference are not fully understood. This thesis explores one possible explanation, epigenetics, primarily focusing on DNA methylation. These natural chemical “switches” control when genes are turned on or off without changing their DNA sequence, and can adjust rapidly in response to stress, potentially helping plants fight infection. We first generated a high-quality reference genome for the CMD-susceptible variety TMEB117 to investigate this, providing a basis for DNA methylation studies. During this work, we discovered a large insertion on chromosome 12, a highly repetitive region enriched with transposable elements and genes linked to epigenetic regulation. While not directly associated with CMD resistance, this region could influence cassava’s adaptability to stress and warrants further study. We then compared TMEB117 with the CMD-tolerant variety TMEB693. DNA methylation analysis revealed that TMEB693 maintained stable methylation patterns regardless of infection, whereas TMEB117 began with higher methylation and shifted toward TMEB693’s methylation profile after infection, possibly too late to provide full protection. These results suggest that CMD resilience depends not only on the plant’s genes but also on the epigenetic changes. The findings provide a basis for further investigation of the genes involved, enabling the identification of specific gene markers to understand CMD resistance better. This knowledge can open new opportunities for plant breeders to develop cassava varieties with stronger, more durable CMD resistance, helping safeguard harvests and strengthen food security across Africa.

Populärvetenskaplig sammanfattning

Kassava är en basföda för hundratals miljoner människor i Afrika söder om Sahara och frodas i magra jordar och under torka, där många andra grödor misslyckas. Produktionen hotas dock av kassavamosaiksjukdomen, en virusinfektion som kan förstöra upp till 95 % av skördarna hos mottagliga sorter. Vissa kassavasorter bekämpar sjukdomen bättre än andra, men orsakerna till denna skillnad är inte helt klarlagda. Denna avhandling utforskar en möjlig förklaring, epigenetik, med fokus främst på DNA-metylering. Dessa naturliga kemiska "strömbrytare" styr när gener slås på eller av utan att ändra deras DNA-sekvens, och kan anpassa sig snabbt som svar på stress, vilket potentiellt hjälper växter att bekämpa infektion. Projektet genererade först ett högkvalitativt referensgenom för den CMD-mottagliga sorten TMEB117 för att undersöka detta, vilket gav grunden för DNA-metyleringsstudien. Under detta arbete gjordes upptäckten att det fanns en stor insättning på kromosom 12, en mycket repetitiv region berikad med transposon-element och gener kopplade till epigenetisk reglering. Även om den inte är direkt associerad med CMD-resistens, kan denna region påverka kassavas anpassningsförmåga till stress och motiverar ytterligare studier. Inom projektet jämfördes det sedan TMEB117 med den CMD-toleranta sorten TMEB693. DNA-metyleringsanalysen visade att TMEB693 bibehöll stabila metyleringsmönster oavsett infektion, medan TMEB117 började med högre metylering och skiftade mer mot TMEB693s metyleringsprofilen efter infektion, möjligen sker detta för sent för att ge fullt skydd. Dessa resultat tyder på att CMD-motståndskraft inte bara beror på växtens gener utan också på de epigenetiska förändringarna. Resultaten ger en grund för vidare undersökning av de involverade generna, vilket möjliggör identifiering av specifika genmarkörer för att bättre förstå CMD-resistens. Denna kunskap kan öppna nya möjligheter för växtförädlare att utveckla kassavasorter med starkare och mer hållbar CMD-resistens, vilket hjälper till att skydda skördar och stärka livsmedelssäkerheten i hela Afrika..

Acknowledgements

This work was carried out at the Department of Animal Biosciences (HBIO) at the Swedish University of Agricultural Sciences (SLU) and IITA, Nairobi, Kenya. It was supported by funding from the Swedish Research Council (VR) through the EpiCass and CassavaNet4Dev projects.

I am deeply grateful to my supervisors: **Erik Bongcam-Rudloff, Andreas Gisel, Trushar Shah, Adnan Niazi, Livia Stavolone, and Laurent Falquet** for their invaluable guidance, mentorship, and the many insightful discussions that shaped this thesis. My sincere thanks also go to the lab team at the Institute of Sustainable Plant Protection, CNR, in Bari, Italy, special thanks to Anna Vittoria Carluccio, whose hands-on support was key to all the lab work reported here. I extend my appreciation to the Bioinformatics team at SLU and the SLU Bioinformatics Infrastructure (SLUBI) for their technical support and the stimulating discussions that helped me progress. Special thanks to Renaud, Ellena, and Cano at SLU, whose kind assistance made each of my visit at SLU smooth and welcoming with administrative and IT matters at campus.

I thank PhD students, Sallam, Fotios, Net, Tomas, and Farok at SLU for their friendly interactions along the way. At IITA Nairobi, I am especially grateful to Dan Waweru and Eve Nyawira for their valuable friendship and the shared experiences that made the journey more meaningful. Thanks to my colleague at IITA bioinformatics section, Laurah Ondari, for the exciting bioinformatics discussions that enriched this work.

Above all, I thank God whose grace, strength, and guidance carried me through every step of this journey. Finally, I thank my friends and family for your unwavering support, love, and encouragement throughout. I sincerely thank you all.

www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

Haplotype-resolved genome of heterozygous African cassava cultivar TMEB117 *(Manihot esculenta)*

Michael Landi[1,2] ✉, Trushar Shah[2], Laurent Falquet[3,4], Adnan Niazi[1], Livia Stavolone[5,6], Erik Bongcam-Rudloff[1] & Andreas Gisel[5,7] ✉

Cassava (*Manihot esculenta Crantz*) is a vital tropical root crop providing essential dietary energy to over 800 million people in tropical and subtropical regions. As a climate-resilient crop, its significance grows as the human population expands. However, yield improvement faces challenges from biotic and abiotic stress and limited breeding. Advanced sequencing and assembly techniques enabled the generation of a highly accurate, nearly complete, haplotype-resolved genome of the African cassava cultivar TMEB117. It is the most accurate cassava genome sequence to date with a base-level accuracy of QV > 64, N50 > 35 Mbp, and 98.9% BUSCO completeness. Over 60% of the genome comprises repetitive elements. We predicted over 45,000 gene models for both haplotypes. This achievement offers valuable insights into the heterozygosity genome organization of the cassava genome, with improved accuracy, completeness, and phased genomes. Due to its high susceptibility to African Cassava Mosaic Virus (ACMV) infections compared to other cassava varieties, TMEB117 provides an ideal reference for studying virus resistance mechanisms, including epigenetic variations and smallRNA expressions.

Background & Summary

Plants exhibit remarkable genetic diversity, often as a mosaic of different variants within a single individual. Crops like cassava, mango, and rubber tree are often highly heterozygous because of either outcrossing or clonal propagation[1–3]. Plants propagated clonally through methods such as stem cutting retain genetic variation, making it challenging to create high-quality reference genomes. Despite this challenge, the recent advancements in sequencing technologies have made it possible for researchers to explore the complex genomic architecture of these crops[4]. By uncovering and analyzing the genomic diversity of plants, we can fully harness their potential and facilitate innovations in the fields like breeding, agronomy, and food security.

Cassava (*Manihot esculenta Crantz*) is a vital crop for subsistence farmers in tropical and subtropical regions across the globe, providing a source of food and industrial purposes. Cassava is utilized to produce various products such as starch, bioethanol, and other bio-based products such as feed, medicine, cosmetics, and biopolymers[5]. Subsistence farmers in Africa prefer cultivating cassava as it yields substantial harvests under diverse environmental conditions[6]. Additionally, cassava roots have an ideal harvesting age and can be harvested flexibly, offering the benefits of a longer in-ground storage[7]. However, the crop faces pests, diseases, drought, weeds, and environmental factors that limit its productivity. Developing more resilient and productive cassava varieties through conventional breeding is time-consuming. Therefore, having a complete haplotype-resolved genome with high accuracy can be a valuable resource in cassava breeding and genomics.

[1]Department of Animal Breeding and Genetics, Bioinformatics, Swedish University of Agricultural Sciences, Uppsala, Sweden. [2]International Institute of Tropical Agriculture, Nairobi, Kenya. [3]Department of Biology, University of Fribourg, Fribourg, Switzerland. [4]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [5]International Institute of Tropical Agriculture, Ibadan, Nigeria. [6]Institute for Sustainable Plant Protection, Consiglio Nazionale delle Ricerche, Bari, Italy. [7]Institute for Biomedical Technologies, Consiglio Nazionale delle Ricerche, Bari, Italy. ✉e-mail: michael.landi@slu.se; m.landi@cgiar.org; a.gisel@cgiar.org; andreas.gisel@cnr.it

Description	Contig-level assembly statistics hap1	Chromosome-level assembly statistics hap1	Contig-level assembly statistics hap2	Chromosome-level assembly statistics hap2
Number of contigs	362	299	159	96
Number of contigs > = 25000 bp	317	262	156	96
Number of contigs > = 50000 bp	100	52	103	45
Largest contig (bp)	43,434,407	51,416,241	35,731,055	50,183,282
Total length (bp)	693,971,781	693,957,521	664,959,903	664,966,403
N50	18,674,865	37,612,488	17,299,599	35,761,448
N90	9,202,869	32,042,468	6,549,198	30,936,428
L50	13	9	13	9
L90	33	17	33	17
GC (%)	37.81	37.81	37.62	37.62
N's per 100 Kbp	0	0.49	0	0.53
N's	0	3412	0	3492

Table 1. Assembly quality metrics generated by QUAST for the assembly produced by hifiasm before and after ordering and scaffolding.

Cassava has a haploid genome of about 750 Mbp, a highly heterozygous and repetitive plant genome[8]. Despite the use of various sequencing technologies over time, there are unresolved gaps in the genome. The reference genome AM560-2, derived from a Colombian cassava line MCol505, has undergone steady improvement over a decade and has had five major releases, with the current version being AM560-2 version 8[4]. While this reference genome benefits the cassava community, it does not capture the genetic diversity in African cassava cultivars grown by smallholder farmers due to its homozygous nature. Recently, attempts have been made to assemble genomes of African cassava lines such as TME3 and 60444, using a combination of Illumina short reads, PacBio long reads, bio-nano optical mapping, and chromatin conformation capture (Hi-C) sequence technologies producing assemblies of N50 of 98 and 117 Kbp. The assembled genomes had large contiguous assemblies but lacked haplotypic separation, containing multiple copies of duplicated sequences in the primary assembly[9]. The TME7 genome was assembled using a combination of Illumina, PacBio, and Hi-C sequencing technology to generate a contiguous genome assembly of N50 of approximately 320 Kbp. This genotype was successfully deduplicated and phased using Hi-C sequence data[10]. The most recent African cassava genotype to be assembled is TME204, which was phased using Hi-C technologies and PacBio high-fidelity (HiFi) sequencing reads, resulting in a highly contiguous assembly of N50 > 18 Mbp[11]. PacBio HiFi sequencing technology has proven effective in creating long and highly accurate reads for assembling complex genomes[12]. Recent studies demonstrate its potential in assembling high-quality plant genomes, including *Populus tomentosa* Carr, the 35.6 Gb California redwood genome and *Bletilla striata*[13–15].

In this study, we have generated a haplotype-resolved diploid assembly of TMEB117, a farmer-preferred cassava cultivar, using PacBio HiFi reads. TMEB117 (also called TME117, TME 117, and ISUNIKANKIYAN) is a Nigerian cassava landrace highly susceptible to African cassava mosaic virus (ACMV)[16]. This genotype served as a reference for ACMV studies[16] and a high-resolution genome will pave the way for future investigations in epigenetics and small RNA expression analysis to learn more about the mechanisms of ACMV resistance in cassava to support future breeding programs. The TMEB117 hap1 assembled genome had a total size of 694 Mbp and hap2 665 Mbp with a contig N50 length of 18 Mbp (hap1) and 17 Mbp (hap2) (Table 1). These assembled haplotigs were further ordered and scaffolded using the TME204 reference genome to produce a chromosome-scaled genome for TMEB117 and enhanced contiguity with an improved N50 length exceeding 35 Mbp in both haplotypes (Table 1). The haplotype-specific annotations for TMEB117 hap1 and hap2 genomes resulted in 47,138 and 49,163 gene models, respectively. Within the TMEB117 hap 1 genome, a total of 442 Mbp (65.34%) was occupied by repetitive elements, whereas hap2, 408 Mbp (60.32%), encompassed by transposable elements. Evaluation of the final genome exhibited a high completeness of 98.9%, according to BUSCO[17]. The two haplotype genomes attained a high base-level accuracy of QV > 64. Furthermore, in the raw data, we detected reads that closely matched the entire genome of the fungus *Alternaria alternata*, despite the plants being healthy and showing no symptoms. We eliminated these contaminant reads and excluded them from the final assembly. The phased and annotated homologous chromosomes provide a comprehensive perspective of cassava's heterozygous genome organization with improved accuracy and completeness at a haplotype-resolved level (Fig. 1a). These chromosome pairs are anticipated to be a valuable resource for cassava breeders and essential for functional analysis to characterize molecular mechanisms important agronomical.

Methods

Sampling, sequencing, sequence quality and contamination check. Cassava plants of the TMEB117 genotype, obtained from the International Institute of Tropical Agriculture (IITA) Genebank collection[18], were grown in pots in a screen house (Fig. 1b). Third and fourth fully-expanded leaves of a potted plantlet hardened from *in vitro* culture were used for genomic DNA extraction using an optimized version of the CTAB (2% CTAB, 2% PVP-40, 20 mM Tris-HCl, pH 8.0, 1.4 M NaCl, 20 mM ethylenediaminetetraacetic acid) total nucleic acid extraction protocol as described by Carluccio, A. V. *et al.*[19]. After RNase treatment, the resulting DNA was cleaned using the Genomic DNA clean and concentrator kit (Zymo Research) according to the manufacturer's instructions. We improved the gene annotation step by utilizing RNA data from another project, previously extracted from leaves

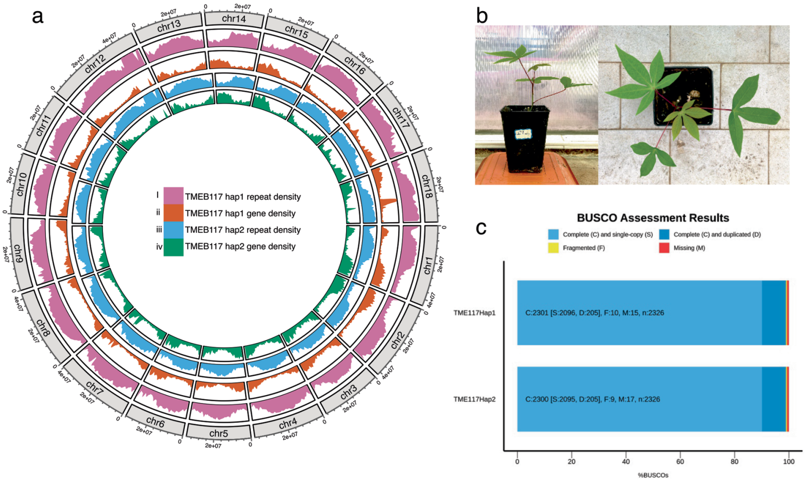


Fig. 1 Overview of the cassava cultivar TMEB117 genome. (**a**) Circos plot displays repeat and gene densities for the two haplotypes visualized in 1 Mbp sliding windows. The tracks from the outer to inner show, (i) Repeat density for hap1 genome (ii) Gene density for hap1 genome (iii) Repeat density for hap2 genome (iv) Gene density for hap2 genome. (**b**) Cassava plant in a pot from the screen house. (**c**) BUSCO score of the TMEB117 genome.

of approximately two-month-old greenhouse plants at IITA Ibadan. The extraction process employed a combination of CTAB and spin column-based purification methods. The total RNA was then sequenced using the Illumina HiSeq. 2500 with a paired-end 2 × 100 cycle approach. The total DNA sample was sequenced with PacBio technology (PacBio Sequel II platform) using two SMRT cells. In the first cell, 875,686 reads and 1,163,062 reads in the second cell were generated. The raw sequence reads obtained from the two SMRT cells were combined. Fastp version 0.23.1[20] (parameters: --length_required 10000 --length_limit 30000) was used to filter the raw sequencing reads, with acceptable read lengths between 10 Kbp to 30 Kbp. Adapter-contaminated reads were removed using HiFiAdapterFilt[21] with the default setting. Filtering and adapter removal resulted in 2,029,912 retained out of 2,038,748 raw reads (99.57%). The quality of the remaining reads was assessed with the default setting of FastQC version 0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and the GC content graphs from the FastQC outputs were further investigated. GC content showed three peaks (Fig. S1a) (see figure deposited at Figshare)[22]. It is generally anticipated that the distribution of GC content follows a normal distribution close to the theoretical distribution. However, the peaks deviate from the theoretical normal distribution in this case. The first peak at ~35% and the second peak at 44.5% mean GC contents represent the GC content of nuclear and mitochondrial genomes, respectively[23,24]. Since the GC % of the chloroplast genome is similar to the nuclear genome, the third peak at ~51% mean GC content could be explained by contamination. Pbbm2 version 1.10.0 (https://github.com/PacificBiosciences/pbmm2), an SMRT minimap2 version 2.15[25] (parameters:–unmapped–log-level INFO–log-file) wrapper was used to then map the raw reads to the cassava reference genome AM560-2 version 8[26]. BLAST search was conducted on unmapped reads to detect the presence of contaminants. BLAST results indicated that the third peak of the GC content originates from reads of *A. alternata*[27]. We further mapped the filtered reads to the *A. alternata* genome. A total 58,703 reads (2.89%) mapped with a consensus of 23.7 Mbp, roughly 69% of the *A. alternata* genome of 34.38 Mbp. The GC content of the *A. alternata* genome is 51%, confirming the third peak of the GC content plot. We extracted all the reads mapping to the *A. alternata* genome, resulting in 1,971,209 sequence reads with GC content having two peaks in the GC content plot, which we process as clean reads for the assembly (Fig. S1b) (see figure deposited at Figshare)[22]. Extracting mapped and unmapped reads was done using samtools v1.15.1 (parameters: samtools fastq -F 4 & -f 4)[28].

Genome assembly, scaffolding, and assembly quality check. Clean HiFi sequence reads of 45x estimated coverage with read length N50 of 17,513 bp (Table S1) (see table deposited in Figshare)[29] were assembled *de novo* using hifiasm v0.16.1-r375 default settings, HiCanu v2.3 (parameters:-p out -pacbio-hifi genomeSize = 750 m -useGrid = false -merylThreads = 4 -merylMemory = 8 corOverlapper = ovl), and Flye v2.9.1-b1780 (parameters: --pacbio-hifi -o out --genome-size 750 m) assembly tools[13,30,31]. Utilizing outcomes from benchmarking analysis of these tools[11] and outputs of these three assemblies, we opted to use the hifiasm assembly based on contiguity and achievement of haplotype resolved assemblies. Assembly statistics were compiled using QUAST[32] default setting. The assembly comprised two haplotigs, hap1 and hap2. Before scaffolding, the two haplotigs assembly metrics showed hap1 and hap2 to consist of 362 contigs with a total length of 694 Mbp with a contig N50 length of 18 Mbp and 159 contigs with a total size of 665 Mbp with contig N50 length of 17 Mbp, respectively (Table 1). Scaffolding and ordering of the contigs were improved by RagTag[33] (parameter: scaffold -o out -t 12) using TME204 reference, a chromosome-scaled genome[11]. The contiguity of the TMEB117 chromosome-scaled genome was improved, with hap1 N50 of 37 Mbp and hap2 N50 of 35 Mbp (Table 1). Eighteen pseudo-molecules representing the chromosomes were compiled for further annotation analysis. The unplaced contigs were separated from the sequences of chromosomes. These unplaced contigs of both haplotypes were then mapped to the chloroplast and mitochondrial genomes[23,24]. The alignment revealed a complete 100% coverage to the chloroplast genome and 90.39% and 62.63% coverage for hap1 and hap2, respectively, to the mitochondrial genome. These haplotigs provided representative sequences for the mitochondrial and chloroplast genome. The final scaffolded assembly of TMEB117 eighteen chromosomes was utilized for downstream analysis.

Repeat landscape and gene annotation. The annotation of transposable elements (TE) was accomplished by using the Extensive *de novo* TE Annotator (EDTA)[34] (parameters:–genome–overwrite 1–sensitive 1–anno 1–evaluate 1), combining structure and homology-based detection to identify predominant TEs in the assembled genome. The pipeline applies various tools, such as HelotronScanner, LTR_FINDER, LTRHarvest, LTR_retriever, TIR-Learner, RepeatModeler2, and RepeatMasker[35–41], to classify novel TE sequences. We screened the outputs of EDTA using R and tidyverse package, resulting in non-redundant TE annotations and visualizations for both haplotypes. The generated repeat-masked genome was subsequently used for gene prediction. Consistent with other African cassava genomes[10,11], over 50% of the genome constitutes repetitive elements. Specifically, in this study, 65.34% and 60.32% of the genome in hap1 and hap2 are transposable elements. The long terminal repeats – retrotransposons (LTR-RTs) are the most abundant, covering 57.37% (hap1) and 54.42% (hap2) of the genome size (Fig. 2a,b). *Gypsy* was the most abundant retrotransposons superfamily, occupying 41.14% (hap1) and 38.35% (hap2) of the genome (Table S2a,b) (see tables deposited at Figshare)[29]. The annotations are classified as families and superfamilies. Between the two haplotype genomes, there is a minimal difference in transposable element annotation percentage (Fig. 2a,b). However, the distribution of the TEs across the chromosomes differs between the two haplotypes (Fig. 2c,d). We used the Funannotate v1.8.9 singularity pipeline (see the script in the code availability section) to annotate the TMEB117 genome. The annotation pipeline involves three primary steps: genome masking, gene prediction, and functional annotation. Prior to annotation, the genome assembly as an output of EDTA annotation underwent soft-masked using scripts provided by the EDTA tool *make_masked.pl*. PASA alignment tool[42] was used to generate an initial set of gene models by integrating RNA-seq data and protein homology to improve the accuracy of gene models. We used a set of 568,002 reviewed and curated protein sequences from a diverse array of species found in the UniProtKB/Swiss-Prot database release 2022_03 for the gene prediction step. Gene prediction was conducted using *ab initio* gene prediction tools, Augustus v3.3, SNAP v2013-02-16, and GlimmerHMM v3.0.4[43,44] were employed for gene prediction. EVidenceModeler v1.1.1[45] integrated gene models from various gene predictors and generated a consensus of the gene models. These gene models were used to generate protein sequences, which underwent filtering to remove proteins with less than 50 amino acids and to check for homology to transposable elements. The predicted genes were functionally annotated using the EggNOG database, UniProtKB, MEROPS, and CAZYmes[46–49], giving insights into the biological functions and pathways. The resulting annotations were manually curated to correct errors and adjust gene models as necessary using the Funnannotate interface. Non-overlapping tRNA genes were predicted using tRNAscan-SE v2.0.9[50]. Transcript evidence was generated by Trinity v2.11.0[51] through *de novo* transcript assembly, which was used in correcting, enhancing, and updating the predicted gene models. The haplotype-specific annotations for hap1 and hap2 genomes resulted in 47,138 and 49,163 genes, 53,264 and 55,222 transcripts, 836 and 814 tRNA, and 52,428 and 54,408 proteins, respectively. BUSCO analysis reveals 90% protein sequence completeness for both haplotypes.

Orthologue analysis. The predicted protein sequences from the study were further analyzed by comparing them to other phased African cassava genomes, namely TMEB7 and TME204, with the AM560-2 v8.1 genome. Each haplotype was analyzed separately with the AM560-2 v8.1 genome. The OrthoVenn2 online tool (https://orthovenn2.bioinfotoolkits.net/) was used to identify orthologous protein groups across the genomes. The analysis was conducted with default parameters, including an inflation value of 1.5 for the Markov clustering algorithm and a BLASTP e-value of 1e-2. The resulting orthologous groups were visualized using the OrthoVenn2 web interface, which displayed Venn diagrams that indicate the number of unique and shared groups across the genomes. The OrthoVenn2 analysis identified 37,384 and 37,518 orthologous clusters, with 18,770 and 19,588 core genome orthologs for hap1 and hap2, respectively, indicating the presence of conserved groups across the genomes. In the TMEB117 hap1 and hap2 genomes, we observe fewer unique protein sequences (931 and 1042) than other cassava genomes, as shown in Fig. 3a,b.

Data Records

Raw PacBio Hifi reads utilized for the assembly can be accessed from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under BioProject PRJNA1002255, with accession numbers SRR25517176[52] and SRR25517175[53]. The two chromosome-scaled haploid genomes have been submitted under distinct BioProject identifiers within NCBI, PRJNA1002865 for hap1 and PRJNA1002864 for hap2 with accession numbers JAWPHJ000000000[54] and JAWPHK000000000[55], respectively. The transcriptome data employed to annotate the genomes are available at NCBI with accession numbers SRR25537339[56], SRR25537340[57], and SRR25537338[58]. The genome annotation files are uploaded to Zenodo[59].

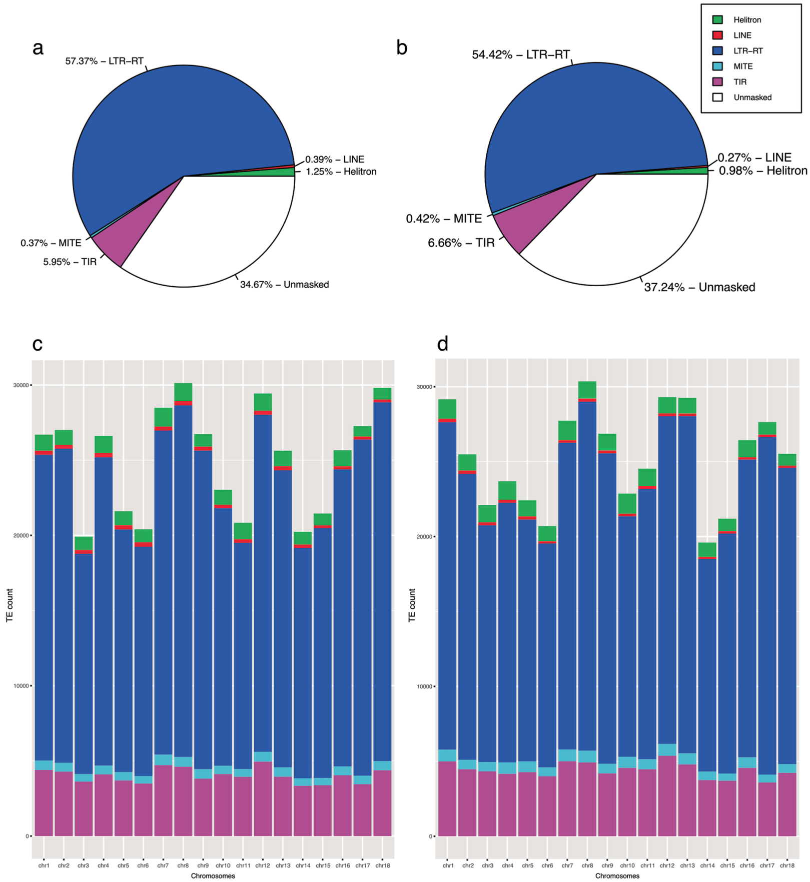


Fig. 2 Illustrate the proportion and distribution of TEs across the chromosomes, as annotated by EDTA. (**a**) Shows the proportion of TEs identified in the hap1 genome, with the most abundant type LTR-RT (represented in the blue segment in the pie chart), covering 57.37% of the genome. (**b**) In hap2, LTR-RT remains the predominant TE family covering 55.07% of the genome. (**c**) Provides an overview of the distribution of TE families across all cassava chromosomes. (**d**) Slight difference in the distribution of TEs families annotated in the chromosomes of the hap2 genome compared to the hap1 genome.

Technical Validation

The genome was validated by ensuring that the constructed assembly conforms with the data used to generate it. Over 99% of the raw sequence reads mapped to both the haplotigs. To evaluate the completeness, we used BUSCO v5.3.2 (parameters: -m genome eudicots_odb10) with orthologs from the eudicots lineage datasets, which included 2326 reference sets of genes specific to plants. BUSCO completeness score for the hap1 and hap2 assemblies was 98.9% (Fig. 1c). We employed Blobtools[60] using the default setting to ascertain the absence of contamination by blasting the nucleotide database against the assembly and mapping the coverage of the assembly using HiFi reads to generate blobs. Blobs in the blob plot (Fig. S2) (see figure deposited at Figshare)[22] were plotted at expected GC content percentages consistent with the GC plot after contamination removal.

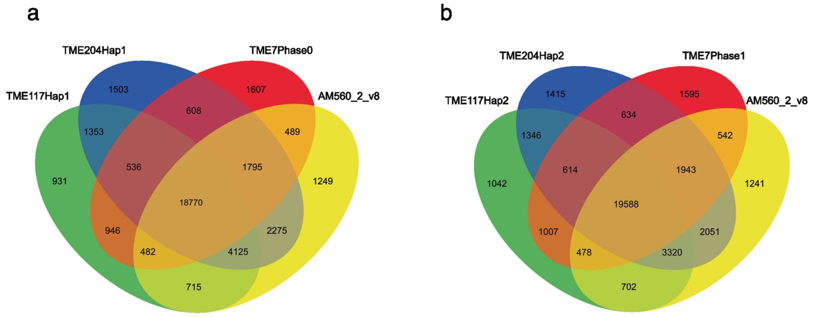


Fig. 3 Venn diagram of the number of gene families shared among and unique to the haplotype genomes of three African cassava cultivars: TMEB117 (hap1 and hap2), TME204 (hap1 and hap2), TME7 (hap1 and hap2), in comparison to the reference genome AM560-2 v8. (**a**) 18,770 core gene families shared among the first haplotigs comparison with the reference genome AM560-2 v8. The second comparison (**b**)19,588 core genes on the second haplotig comparison with the reference genome AM560-2 v8. 931 gene families were unique in TMEB117 hap1 genome and 1042 in the hap2 genome.

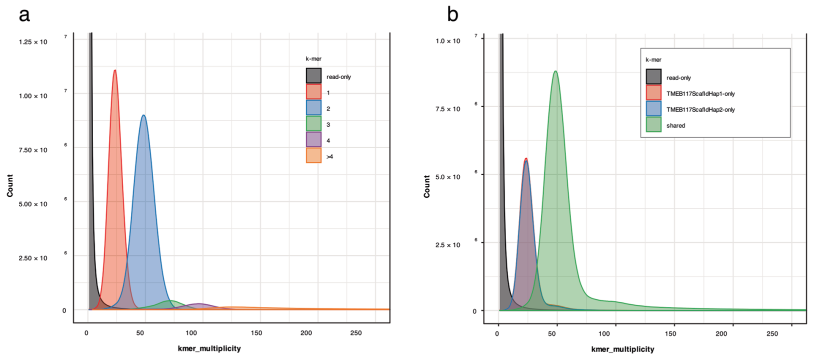


Fig. 4 The completeness of resolved haplotypes assessed by Merqury copy number spectrum plots (**a**) and assembly plots (**b**). The x-axis represents the k-mer multiplicity, while the y-axis shows the abundance of k-mers. The grey region. represents the abundance of k-mers in the HiFi reads missing in the scaffold of the genome. (**a**) Copy number spectrum plot - the red peak observed at ~ 25x indicates heterozygous k-mers (1-copy k-mers), while the blue peak at ~ 50x represents the homozygous k-mers (2-copy k-mers). The other peaks show low levels of duplicated k-mers. (**b**) Assembly plot – k-mers coloured by their uniqueness, red peak (hap1), blue peak (hap2). At the heterozygous peak (25x), there is a slight difference in the k-mers indicating reconstruction of heterozygous variants was almost complete. Shared k-mers are shown in green which is at the 50x k-mer multiplicity.

BLAST hits from the blob plot showed Streptophyta, which is a clade of plants. Therefore, we conclude that the assembly was free from contaminants. Haplotypic separation and assembly quality were achieved by performing a *k-mer*-based analysis using Merqury[61] (*k-mer* = 21). The assembly has a quality value (QV) score of 64.38 for hap1 and 67.99 for hap2. The *k-mer* completeness for each haplotype assembly and the combined set was 78.63%, 77.95%, and 98.79%, respectively. This was approximately 20% of *k-mers* being haplotype-specific. So far, this genome is of better quality assembly, based on the QV score, compared to already assembled African cassava cultivars[10,11] (Table 2). Figure 4 illustrates that the assembled sequence resulted in a nearly completely haplotype-resolved genome, as indicated by the copy number and assembly *k-mer* plots. Most heterozygous haplotype-specific *k-mers* were observed once in the assembled sequence, and the majority of homozygous

	TMEB117	TME204	TME7
Primary assembly (hap1)	64.38	45.23	34.1
Alternative assembly (hap2)	67.99	48.94	34.4

Table 2. Quality value scores comparison table for TMEB117 and previously reported haplotype-resolved cassava genome assemblies.

k-mers were shared by the two genome haplotypes (Fig. 4a). In the heterozygous peak, slightly fewer *k-mers* differed between the two haplotypes (Fig. 4b), confirming that the reconstruction of heterozygous variants was almost thorough. We assessed the completeness of the predicted protein sequences within the eudicots lineage using BUSCO v5.3.2 (parameters: -m proteins eudicots_odb10) to validate the gene annotations. BUSCO analysis reveals 90% protein sequence completeness for both haplotypes (Fig. 1c). Subsequently, we performed a conditional reciprocal BLAST[62], extracting the predicted gene model sequences from both haplotypes and compared them to the gene sequences of AM560-2, the cassava reference genome. Out of the predicted 47,138 genes in hap1 and 49,163 genes in hap2, we identified 30,456 in hap1 and 30,370 in hap2 reference gene sequences. Particularly, the reference AM560-2 genome, with a total of 32,805 genes, exhibited similarity to most predicted genes of both haplotypes of the TMEB117 genome.

Code availability

No custom programming or coding was used. Instead, the analysis utilized bash commands and the corresponding scripts stored within the GitHub repository accessible at: https://github.com/LandiMi2/GenomeAssemblyTMEB117.

Received: 16 August 2023; Accepted: 29 November 2023;
Published online: 09 December 2023

References

1. Wang, P. *et al*. The genome evolution and domestication of tropical fruit mango. *Genome Biol* **21** (2020).
2. Tang, C. *et al*. The rubber tree genome reveals new insights into rubber production and species adaptation. *Nat Plants* **2** (2016).
3. Bredeson, J. V. *et al*. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat Biotechnol* **34**, 562–570 (2016).
4. Lyons, J. B. *et al*. Current status and impending progress for cassava structural genomics. *Plant Molecular Biology* vol. 109, 177–191, https://doi.org/10.1007/s11103-020-01104-w (2022).
5. Li, S. *et al*. The industrial applications of cassava: current status, opportunities and prospects. *Journal of the Science of Food and Agriculture* **97**, 2282–2290, https://doi.org/10.1002/jsfa.8287 (2017).
6. Ceballos, H. H., Iglesias, C. A., Pe´rezpe´rez, J. C. & Dixon, A. G. O. *Cassava breeding: opportunities and challenges*.
7. Uchechukwu-Agua, A. D., Caleb, O. J. & Opara, U. L. Postharvest Handling and Storage of Fresh Cassava Root and Products: a Review. *Food and Bioprocess Technology* **8**, 729–748, https://doi.org/10.1007/s11947-015-1478-z (2015).
8. Prochnik, S. *et al*. The Cassava Genome: Current Progress, Future Directions. *Tropical Plant Biology* **5**, 88–94, https://doi.org/10.1007/s12042-011-9088-z (2012).
9. Kuon, J. E. *et al*. Haplotype-resolved genomes of geminivirus-resistant and geminivirus-susceptible African cassava cultivars. *BMC Biol* **17**, 1–15 (2019).
10. Mansfeld, B. N. *et al*. Large structural variations in the haplotype-resolved African cassava genome. *Plant Journal* **108**, 1830–1848 (2021).
11. Qi, W. *et al*. The haplotype-resolved chromosome pairs of a heterozygous diploid African cassava cultivar reveal novel pan-genome and allele-specific transcriptome features. *Gigascience* **11** (2022).
12. Hon, T. *et al*. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* **7** (2020).
13. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
14. An, X. *et al*. High quality haplotype-resolved genome assemblies of Populus tomentosa Carr., a stabilized interspecific hybrid species widespread in Asia. *Mol Ecol Resour* **22**, 786–802 (2022).
15. Jiang, L. *et al*. Haplotype-resolved genome assembly of Bletilla striata (Thunb.) Reichb.f. to elucidate medicinal value. *Plant Journal* **111**, 1340–1353 (2022).
16. Quantification of African cassava mosaic virus (ACMV) and East African cassava mosaic virus (EACMV-UG) in single and mixed infected Cassava (Manihot esculenta Crantz) using quantitative PCR - 1-s2.0-S0166093415003262-main.
17. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
18. Paliwal, R., Adegboyega, T. T., Abberton, M., Faloye, B. & Oyatomi, O. Potential of genomics for the improvement of underutilized legumes in sub-Saharan Africa. *Legume Science* **3,** https://doi.org/10.1002/leg3.69 (2021).
19. Carluccio, A. V. *et al*. Set up from the beginning: The origin and early development of cassava storage roots. *Plant Cell Environ* **45**, 1779–1795 (2022).
20. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. in *Bioinformatics* vol. 34 i884–i890 (Oxford University Press, 2018).
21. Sim, S. B., Corpuz, R. L., Simmonds, T. J. & Geib, S. M. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* **23** (2022).
22. Landi, M. Supplementary figures cassava TMEB117 genome. *Figshare*. https://doi.org/10.6084/m9.figshare.23792292.v2 (2023).
23. Tao, Q., Cao, J., Zhu, L. & Lin, H. The complete mitochondrial genome of an important root crop cassava (Manihot esculenta). *Mitochondrial DNA B Resour* **4**, 1081–1082 (2019).
24. Daniell, H. *et al*. The complete nucleotide sequence of the cassava (Manihot esculenta) chloroplast genome and the evolution of atpF in Malpighiales: RNA editing and multiple losses of a group II intron. *Theoretical and Applied Genetics* **116**, 723–737 (2008).
25. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
26. Bredeson, J. V. *et al*. 'An improved reference assembly for cassava (Manihot esculenta Crantz)'. In preparation.
27. Gai, Y. *et al*. Chromosome-scale genome sequence of Alternaria alternata causing alternaria brown spot of citrus. *Molecular Plant-Microbe Interactions* **34** (2021).

28. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
29. Landi, M. Supplementary tables. *Figshare.* https://doi.org/10.6084/m9.figshare.23792298.v1 (2023).
30. Koren, S. *et al*. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res* **27**, 722–736 (2017).
31. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546 (2019).
32. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
33. Alonge, M. *et al*. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol* **23**, (2022).
34. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20**, (2019).
35. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc Natl Acad Sci USA* **111**, 10263–10268 (2014).
36. Ou, S. & Jiang, N. LTR_FINDER_parallel: Parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA* **10** (2019).
37. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, (2008).
38. Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* **176**, 1410–1422 (2018).
39. Su, W., Gu, X. & Peterson, T. TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. *Mol Plant* **12**, 447–460 (2019).
40. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457 (2020).
41. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics,* https://doi.org/10.1002/0471250953.bi0410s25 (2009).
42. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654–5666 (2003).
43. Korf, I. *Gene finding in novel genomes*. http://www.biomedcentral.com/1471-2105/5/59 (2004).
44. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
45. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9** (2008).
46. Drula, E. *et al*. The carbohydrate-active enzyme database: Functions and literature. *Nucleic Acids Res* **50**, D571–D577 (2022).
47. Huerta-Cepas, J. *et al*. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314 (2019).
48. Rawlings, N. D. *et al*. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res* **46**, D624–D632 (2018).
49. Bateman, A. *et al*. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480–D489 (2021).
50. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. in *Methods in Molecular Biology* **1962**, 1–14 (Humana Press Inc., 2019).
51. Haas, B. J. *et al*. *De novo transcript sequence recostruction from RNA-Seq: reference generation and analysis with Trinity*. *Nature protocols* **8** (2013).
52. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25517176 (2023).
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25517175 (2023).
54. Landi, M. *et al*. The genome information of African cassava cultivar TMEB117 genome (Hap1). *GenBank.* https://identifiers.org/ncbi/insdc:JAWPHJ000000000 (2023).
55. Landi, M. *et al*. The genome information of African cassava cultivar TMEB117 genome (Hap2). *GenBank.* https://identifiers.org/ncbi/insdc:JAWPHK000000000 (2023).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25537339 (2023).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25537340 (2023).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR25537338 (2023).
59. Landi, M. *et al*. Genome annotation of African cassava cultivar TMEB117 genome. *Zenodo.* https://zenodo.org/doi/10.5281/zenodo.10013084 (2023).
60. Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit - interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics* **10**, 1361–1374 (2020).
61. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21** (2020).
62. Aubry, S., Kelly, S., Kümpers, B. M. C., Smith-Unna, R. D. & Hibberd, J. M. Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis. *PLoS Genet* **10** (2014).

Acknowledgements

This work was funded by the Swedish Research Council (VR) funded projects, EpiCass (Identification of epigenetic variations influencing viral resistance and yield in cassava (Manihot esculenta) farmers' fields) grant number 2020-04457 and Cassavanet4Dev (In a race to prevent hunger: C assava Bioinformatics for African cassava biotechnology) grant number 2021-05105. This study was supported by the SLU Bioinformatics Infrastructure (SLUBI), Scilifelab, NGI, and NBIS Uppsala, Sweden. We thank the Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden, and the International Institute of Tropical Agriculture (IITA) for providing research facilities and other support for this study. ML acknowledges using the High-Performance Computing services of the International Institute of Tropical Agriculture (IITA) and the International Livestock Research Institute (ILRI).

Author contributions

A.G. and E.B.R. conceived the idea and led the grant proposal writing. L.S. handled the preparation of the plant materials and sequencing. ML analyzed the data and wrote the original draft of the manuscript. T.S., L.F. and A.N. provided oversight throughout the process. The final manuscript underwent review and approval by all authors. Correspondence and requests for materials should be addressed to M.L. and A.G.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.L. or A.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Landi *et al. BMC Genomics* (2025) 26:362
https://doi.org/10.1186/s12864-025-11523-y

BMC Genomics

RESEARCH

Open Access

Genome-wide comparison reveals large structural variants in cassava landraces

Michael Landi[1,2*], Anna Vittoria Carluccio[4], Trushar Shah[2], Adnan Niazi[1], Livia Stavolone[3,4], Laurent Falquet[5,6], Andreas Gisel[3,7*] and Erik Bongcam-Rudloff[1*]

Abstract

Background Structural variants (SVs) are critical for plant genomic diversity and phenotypic variation. This study investigates a large, 9.7 Mbp highly repetitive segment on chromosome 12 of TMEB117, a region not previously characterized in cassava (*Manihot esculenta* Crantz). We aim to explore its presence and variability across multiple cassava landraces, providing insights into its genomic significance and potential implications.

Results We validated the presence of the 9.7 Mbp segment in the TMEB117 genome, distinguishing it from other published cassava genome assemblies. By mapping short-read sequencing data from 16 cassava landraces to TMEB117 chromosome 12, we observed variability in read mapping, suggesting that while all genotypes contain the insertion region, some exhibit missing segments or sequence differences. Further analysis revealed two unique genes associated with deacetylase activity, HDA14 and SRT2, within the insertion. Additionally, the *MUDR-Mutator* transposable element was significantly overrepresented in this region.

Conclusions This study uncovers a large structural variant in the TMEB117 cassava genome, highlighting its variability among different genotypes. The enrichment of HDA14 and SRT2 genes and the *MUDR-Mutator* elements within the insertion suggests potential functional significance, though further research is needed to explore this. These findings provide important insights into the role of structural variations in shaping cassava genomic diversity.

Keywords Cassava, Structural variants, Chromosome 12, Large highly repetitive insert

Background

Exploring comparative genomics across plant species holds great significance, offering valuable insights into genetic diversity, gene functionalities, and evolutionary studies. This exploration proves relevant in detecting various genomic variations with profound implications for the plants' phenotypic traits. The understandings gained in these investigations are essential for improving crops and creating varieties with enhanced nutritional content, resilience against diseases, and adaptability to diverse environmental conditions.

DNA structural variants (SVs) exceeding 50 bps in plant genomes are still being explored despite their crucial influence on genomic diversity. These diverse SV

*Correspondence:
Michael Landi
michael.landi@slu.se
Andreas Gisel
andreas.gisel@cnr.it
Erik Bongcam-Rudloff
erik.bongcam@slu.se
[1] Department of Animal Biosciences, Bioinformatics Section, Swedish University of Agricultural Sciences, Uppsala, Sweden
[2] International Institute of Tropical Agriculture, Nairobi, Kenya
[3] International Institute of Tropical Agriculture, Ibadan, Nigeria
[4] Institute of Sustainable Plant Protection, CNR, Bari, Italy
[5] Department of Biology, University of Fribourg, Fribourg, Switzerland
[6] Swiss Institute of Bioinformatics, Lausanne, Switzerland
[7] Institute of Biomedical Technologies, CNR, Bari, Italy

types include insertions, deletions, translocations, and inversions [1]. Such structural variations can significantly impact the structure and function of genomes. Various mechanisms can contribute to forming SVs [2]. One mechanism involves TEs generated as insertion/deletion polymorphisms spanning several kilobase pairs characterized by their repetitive sequences, which can mediate ectopic recombination events, leading to even larger SVs [3]. Single nucleotide polymorphisms (SNPs) capture specific genomic variations. In contrast, SVs account for a larger proportion of heritable genetic variation and significantly impact the genome's structure and function [4, 5]. SVs have become valuable in plant breeding [6, 7]. Within plants, SVs have been linked to variations in critical phenotypic traits like fruit color [1], fruit shape [8], and leaf size [9]. In studies involving tomatoes [6] and grapes [1], substantial impacts on important agricultural traits may arise from large genomic SVs. One illustrative instance is found in the Chardonnay grape, where white berries are suggested to result from a significant inversion and deletion, leading to hemizygosity at the MybA locus [1]. These examples motivate the need for a thorough understanding of SVs, which is vital for unraveling the diverse phenotypes observed in plants [10, 11]. However, identifying such variations relies on the availability of a high-quality reference genome.

The advent of high-throughput sequencing technologies has revolutionized the field of genomics, enabling the sequencing and assembly of high-quality plant genomes. These advancements have enabled the sequencing and assembly of high-quality plant genomes, including cassava. Cassava is a vital crop, especially for subsistence farmers in developing regions [12]. It serves as a source of nutrition for close to a billion people residing in the tropical regions of Africa, South America, and Asia [13]. Understanding cassava genetics is essential for addressing food security challenges and harnessing the crop's potential for improvement. The release of several high-quality, chromosome-scale, haplotype-resolved genome assemblies of African cassava provides a powerful tool for exploring its genetic diversity, offering more profound insights into cassava genetics. Recently released genomes include TMEB117, TME204, and TME7 [14–16]. The TME204 and TMEB117 genomes were sequenced using Pacific Biosciences high-fidelity (HiFi) sequencing reads, while the TME7 genome was sequenced using the PacBio RSII system. HiFi sequence reads have proven to be of high quality and have produced highly accurate genomes [17]. These genomes have already established a high level of heterozygosity in cassava, and there is apparent substantial genetic variability between these cassava cultivars. In the study of the TME7 genome, a comparison was made with the cassava reference AM560-2 v6.1 genome [18].The TME7 genome revealed many SVs, with over 10,000 large SVs (50 – 10,000 bp) covering > 15.99 Mb of sequence. Additionally, > 5,000 large haplotype-specific SVs were discovered within the genome. These SVs in TME7 include insertions, deletions, tandem duplications, expansions, and contractions of repetitive elements. These findings show a substantial contribution of SVs to shaping the genomic landscape and potentially influencing phenotypic diversity in TME7 [16].

Our study focuses on whole-genome alignment using previously published cassava genomes of TMEB117, TME204, and AM560-2 v8.1 [19] to identify both haplotype-specific SVs and SVs across different cassava cultivars through comparative genomics. Haplotype-resolved assemblies are invaluable for delving into haplotype-specific structural variations within genomes and identifying genetic differences across other cassava genomes. Our results reveal structural diversity, including a large 9.7 Mbp insertion on chromosome 12 of TMEB117, absent in previously published genomes. This insertion is highly repetitive and contains unique genetic features, such as overrepresented transposable elements and genes associated with deacetylase activity. Understanding SVs, like this insertion, is crucial for comprehending the genetic landscape of cassava and its impact on phenotypic traits. This research provides new insights into cassava genetics, paving the way for developing improved cassava varieties with enhanced nutritional and agronomic characteristics.

Results

Large 9.7 mbp fragment on chromosome 12 of the TMEB117 genome

Our recent study produced a high-quality haplotype-resolved genome for the TMEB117 cassava cultivar [14]. We examined this genome's chromosome sizes and genomic features, as illustrated in (Fig. 1). Chromosome 12 in TMEB117 exhibits a larger size of approximately 51 Mbp, contrasting with the sizes observed in other cassava cultivars such as TME204 (40 Mbp), TME7 (36 Mbp), and AM560-2 (37 Mbp) (Fig. 1a). Conducting a pairwise alignment of chromosome 12 of TMEB117 with an African cultivar, TME204, and with the reference AM560-2 v8.1, we identified a large inserted fragment of 9.7 Mbp in the TMEB117 genome (Fig. 1b-d). 90% of the large insertion, located in a region characterized by low gene density and high repeat content in both haplotypes (Fig. 1f), is composed of transposable elements (TEs). Of these TEs, the terminal inverted repeats of the *MUDR-Mutator* superfamily accounted for 76% of the total annotated TEs in this region (Table 1). To identify TE superfamilies enriched in this insertion relative to the genome, we performed a Fisher's exact test on the counts of TEs. This analysis uncovered significant enrichment for the *MUDR-Mutator*, Helitron, and CACTA superfamilies (Additional file 2: Table S2).

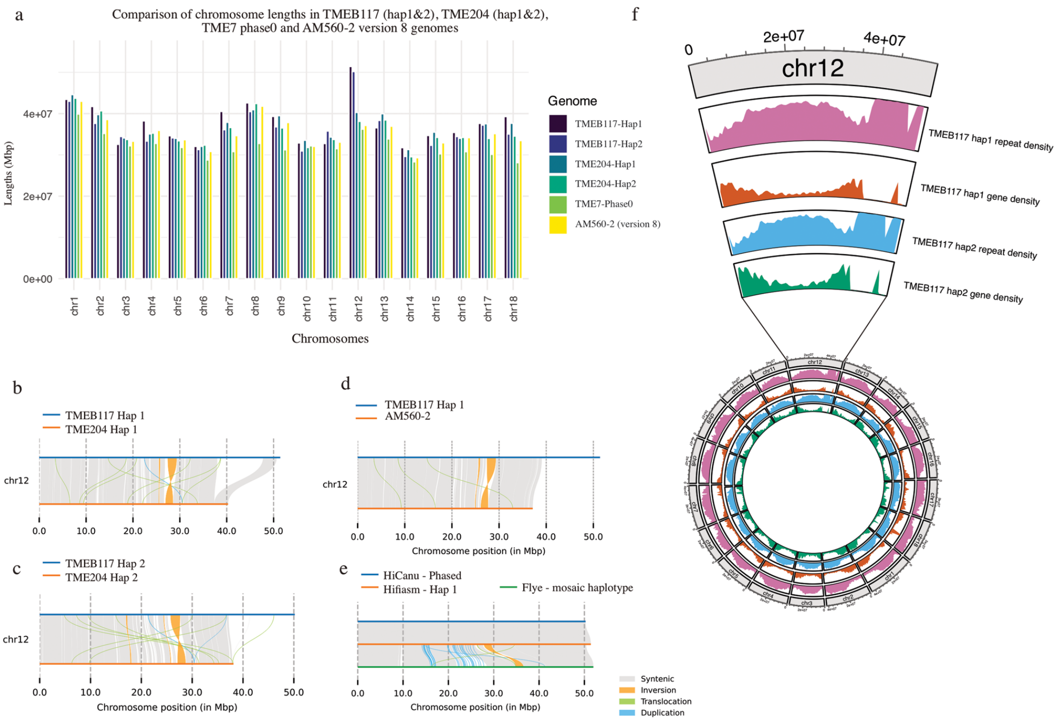


Fig. 1 Chromosome lengths and pairwise comparison of chromosome 12 across cassava cultivars (**a**) Stacked bar plots display the chromosome lengths of TMEB117 (both haplotypes), TME204 (both haplotypes), TME7 (primary haplotype), and AM560-2 version 8 genome. The plot uncovers that TMEB117's chromosome 12 exceeds 50 Mbp, exhibiting a considerable size difference compared to other cassava genome assemblies. (**b**, **c**) Pairwise synteny plots comparing chromosome 12 haplotypes (hap1 and hap2) of TMEB117 and TME204. (**d**) Comparison of TMEB117 chromosome 12 hap1 with the cassava reference genome AM560-2 (version 8). (**e**) Assembly of TMEB117 chromosome 12 using HiCanu (blue, primary haplotype), hifiasm (orange, hap1), and Flye (green, mosaic haplotype), (**f**) Zoomed in view of chromosome 12 from Landi et al., 2023, showing high repeat density and low gene density near the chromosome's end.

Table 1 The proportion of different classes of annotated transposons occupied in the hap1 and hap2 genomes at the insertion region (Coordinates: hap1 chr12:37986388–47748555; hap2 chr12:36042535–45804701)

Family	Superfamily	Total % - Hap 1	Total % - Hap2
TIR	MUDR-Mutator	76.30	76.5
LTR-RT	Gypsy	10.30	10.40
LTR-RT	Unknown	5.38	5.16
LTR-RT	Copia	0.82	0.85
Helitron	Helitron	0.08	0.16
TIR	CACTA	0.08	0.08
TIR	hAT	0.08	0.06
TIR	PIF-Harbinger	0.04	0.01
TIR	Tc1-Mariner	0.01	0.03
LINE	Unknown	0.02	0.03
MITE	CACTA	0.02	0
MITE	MUDR-Mutator	0.02	0.02
MITE	hAT	0.02	0.03
MITE	PIF-Harbinger	0.01	-

Assembling TMEB117 chromosome 12 using three different assemblers (HiCanu, Flye, and hifiasm) to confirm the large insertion and resolve assembly errors

The unique genomic characteristic of chromosome 12 within TMEB117 prompted additional analysis to confirm the presence of this large fragment. We assembled the TMEB117 genome employing three different assembly tools: Hifiasm, HiCanu, and Flye [20–22]. Hifiasm was utilized to generate phased genomes of TMEB117. HiCanu produced a diploid genome that was then phased to a primary haplotype, and Flye produced a collapsed assembly, representing the diploid genome as a single mosaic haplotype. Both hifiasm and HiCanu assemblies showed better assembly statistics and completeness than Flye, with better contiguity, fewer contigs, and minimal duplication. While Flye achieved a higher N50, its genome size was overestimated (Additional file 1: Fig. S8; additional file 2: Table S6). The size of chromosome 12 from the phased HiCanu assembly was 50.26 Mbp, 51.42

Mbp for Hifiasm, and 51.98 Mbp for Flye. Synteny blocks spanning the entire chromosome 12 were observed while comparing the phased genomes obtained with HiCanu and Hifiasm assemblers, successfully capturing the whole insertion region in chromosome 12. However, the Flye assembly, which generated a mosaic haplotype genome, displayed duplications (Fig. 1e). Despite this, synteny was also observed towards the insertion region in the Flye assembly. Furthermore, pairwise alignment showed a lack of similarity between any portion of chromosome 12 sequences and other chromosomes in the genome, unplaced contigs, and the *Alternaria alternata* genome [23] (Additional file 1: Fig. S1), previously seen as a major contaminant.

Coverage analysis of chromosome 12 with other cassava cultivars

The read mapping coverage analysis for chromosome 12 haplotypes across 16 cassava cultivars revealed several differences in read mapping patterns between cultivars (Additional file 1: Fig. S2). The mapped reads showed a high mean mapping quality (>21) on both haplotypes (Additional file 2: Tables S1a and S1b). Uniform coverage within the insertion region across both haplotypes was observed in cultivars such as TME60444, COL2182, CUB40, TMEB117, TME3, TME7, TME14K, and TMS961089A indicating a high degree of similarity and the full-length insertion region in these cultivars with TMS961089A having the lowest and most uneven coverage profile (Additional file 1: Fig. S2 and Additional file 2: Table S1a & b). The read mapping patterns for the cassava reference genome AM560-2, Tree cassava (natural hybrid between *Manihot esculenta* and *Manihot glaziovii*), *Manihot esculenta subsp. flabellifolia* (FLA 496-1) and PER226 showed a slope towards the end of chromosome 12 of approximately 4 Mbp on hap1, with a similar pattern on hap2. However, on hap2, reads were mapped to the last portion of chromosome 12 after the slope on these cultivars (Additional file 1: Fig. S2– left panel). Even though these cultivars (AM560-2, FLA 469-1, PER226, and Tree cassava) had a drop of reads mapping towards the end of chromosome 12, reads covered the insertion regions within the specified coordinates of the insertion (Additional file 1: Fig. S2– right panel). Cultivars like TMEB419, TME204, and ECU41 displayed drops and peaks in read coverage within the insertion region, with a depression pattern. TMEB693 mapping coverage within the insert showed a clear depression of reads mapping in the insertion region. Additionally, hap2 of these cultivars (TMEB419, TME204, TMEB419, and ECU41) exhibited a slight slope of reads coverage at the end of chromosome 12. For ECU41, the read coverage showed a depression and a slope at the last portion of chromosome 12 on hap1. Meanwhile, hap2 showed a pattern similar to PER226, Tree cassava, FLA 496-1, and AM560-2, mapping reads to the last portion of the chromosome. While several South American cultivars (AM560-2, FLA 469-1, PER226, and Tree cassava) exhibited a noticeable drop in the read coverage toward the end of chromosome 12, this pattern was inconsistent across all studied cultivars. In contrast, African cultivars, whether they showed depressed read coverage at the insertion region or uniform coverage across chromosome 12, generally maintained consistent read mapping toward the end of chromosome 12.

Unique read count analysis on gene features within the insertion

Read coverage analysis provided insight into multiple mapped reads within a given genomic region, indicating the extent of sequencing depth even when only a single read is mapped at a specific position. We performed additional mapping count analysis, selecting only uniquely aligned reads within the 8 Mb insertion region (chr12:39,000,000–47,000,000 on hap1 and chr12:37,000,000–45,000,000 on hap2) and compared it to a control region 8 Mb away from the insert (chr12:1,000,000–18,000,000– same coordinates for both haplotypes). We analyzed 22 gene features on hap1 and 17 on hap2 within the insert region and 324 and 349 gene features outside the insert region on hap1 and hap2, respectively. Normalizing the read counts against the whole genome allowed for accurate comparison, accounting for differences in sequencing depth. Our analysis (Fig. 2) revealed that the AM560-2 cultivar showed the highest read counts on both haplotypes within the insertion region. Additionally, other genotypes also had higher mapping counts compared to TMEB117. Consistent with our coverage plot (Additional file 1: Fig. S2– right panel), cultivars such as TME204, TMEB419, ECU41, and TMEB693, which showed depression characteristics in read mapping, had the lowest average read counts. Of these, TMEB693 had slightly higher read counts in this region on hap1. These four cultivars had the lowest counts also on hap2, where the read counts were somewhat higher than those on hap1 (Fig. 2a). As expected, the average read counts across all cultivars outside the insertion region were similar on both haplotypes (Fig. 2b).

Further investigation into gene-by-gene read counts indicated that the AM560-2 cultivar had the highest read counts on both haplotypes (Fig. 3). Some gene features appeared repetitive due to lacking unique reads mapped across all cultivars. As already demonstrated above, TMEB419, TME693, TME204, and ECU41 had the lowest read counts, although certain gene features in these cultivars exhibited higher read counts (Fig. 3). Gene-by-gene plots on gene features outside the insertion region

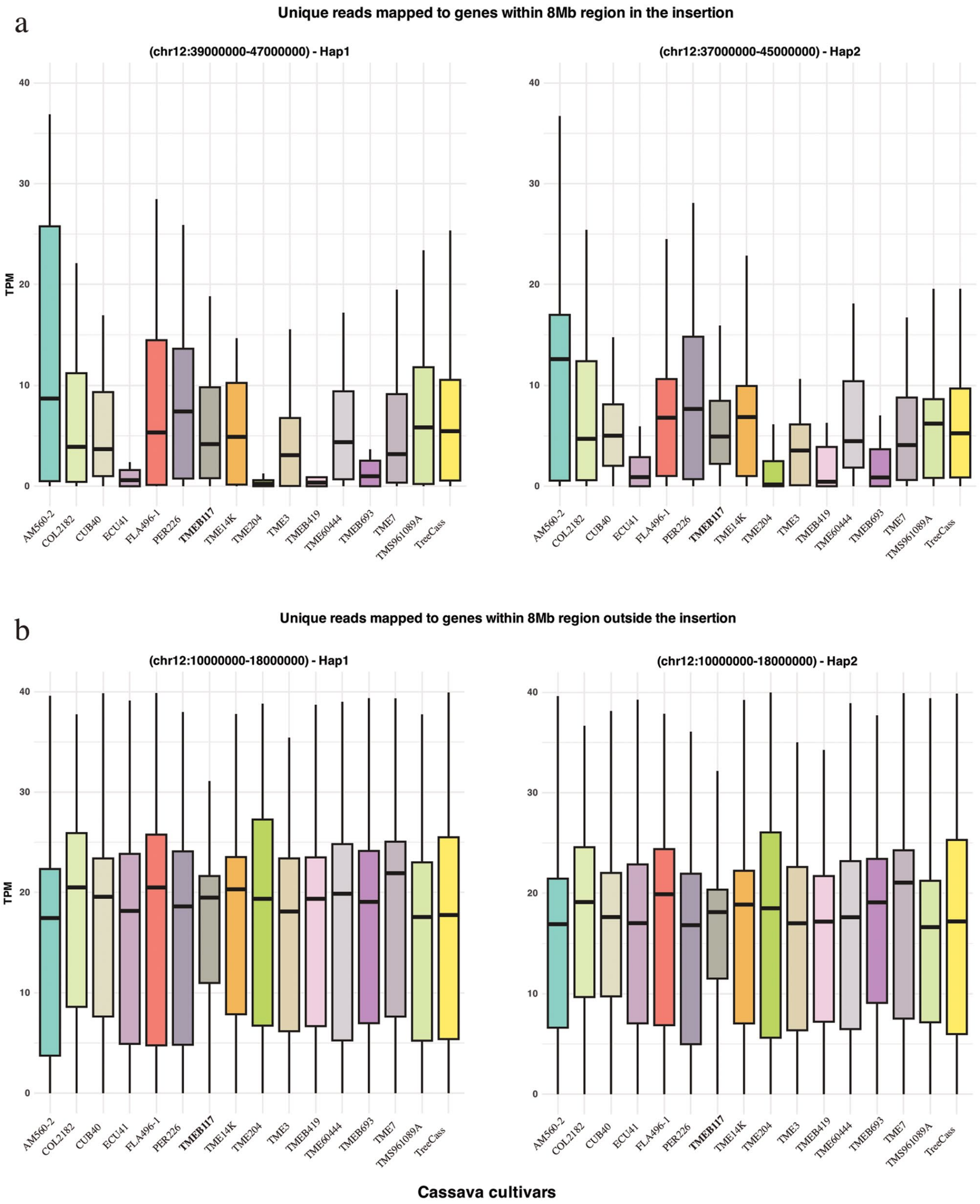


Fig. 2 Boxplots of unique reads mapped to gene features within the 8 Mb insertion region and outside the inserted region. TMEB117 is highlighted in hold. (**a**) Normalized read counts (TPM) for gene features within the insertion region: chr12:39,000,000-47,000,000 on hap1 (left panel) and chr12:37,000,000-45,000,000 on hap2 (right panel). (**b**) Unique reads mapped to gene features outside the insertion region: chr12:1,000,000-18,000,000 on both haplotypes serving as a control.

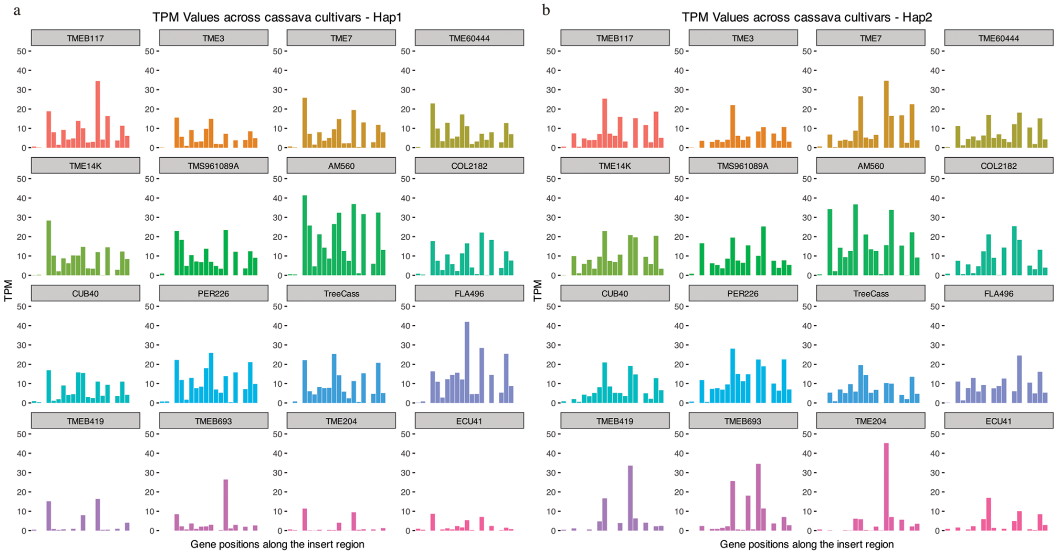


Fig. 3 Gene-by-gene TPM values of read counts across all cultivars within the insertion region. (**a**) Distribution of TPM read count values for 22 gene features across all cultivars on hap1. (**b**) TPM read count values distributed for 17 gene features across all cultivars on hap2.

exhibited similar read counts across all cultivars (Additional file 1: Fig. S3).

PCR confirms the presence of insertion region fragments in TMEB117 genomes

Following the mapping of unique reads, we observed high variability in read coverage within the insertion region across the cassava cultivars in this study. Based on the availability of plant material, we selected three cultivars - TMEB117, TMEB419, and TME693 - to confirm the presence of the insert region fragments. The repetitive nature of the insertion region made it challenging to obtain suitable primers. Nonetheless, using the TMEB117 hap1 insertion coordinates (chr12:37,986, 388 – 47,748,555), we designed three unique primer sets to selectively amplify different fragments of the insertion region (Additional file 2: Table S3). We conducted multiplex PCR to verify the presence of these specific fragments in the cassava genomes of the TMEB117, TME693, and TMEB419 genotypes, incorporating the *Protein phosphatase 2 A* (PP2A) and *GTP-binding protein* (GTPb) positive control genes in the same reaction. We successfully amplified the insertion fragments A, B, and C (Additional file 2: Table S3) and obtained products of the correct sizes. (Additional file 1: Fig. S6) The results confirmed that all three fragments are present in the TMEB117 and are also detected in the TMEB693 and TMEB419 genotypes (Additional file 1: Fig. S6). The fragments amplified from the TMEB117 genome were sequenced to confirm the correct amplification.

Gene enrichment analysis for the known genes in the insertion region

We observed the presence of 147 gene features in hap1 and 159 in hap2 by investigating insertion regions defined by specific coordinates (hap1 chr12:37986388–47748555; hap2 chr12:36042535–45804701). The hap1 insertion region had 37 known genes and 110 genes annotated as potential. Similarly, hap2 had 36 known genes and 123 potential genes. We ran a gene ontology (GO) enrichment analysis with the known genes from both haplotypes and identified significant gene sets. The GO analysis identified three significant enriched molecular function terms among the genes. The terms included “histone deacetylase activity” (GO:0004407), “protein lysine deacetylase activity” (GO:0033558), and “deacetylase activities” (GO:0019213) in both haplotypes. For each GO term, both haplotypes showed enrichment of two genes, HDA14 and SRT2, as indicated by the gene count of 2 (Additional file 2: Table S7a, b). To explain the unknown genes, BLAST searches were performed on the gene sequences extracted from the potential gene set for both haplotypes. The results revealed homologous sequences in *Manihot esculenta* chromosome 12, annotated with Manes IDs, and *Hevea brasiliensis* (Additional file 2: Table S4 and S5).

Genome-wide comparative analysis across cassava cultivars

Large haplotypic structural variation within the TMEB117 genome

We compared the two haplotypes to analyze differences within the TMEB117 genome. Our analysis showed unique structural variations specific to each haplotype. Structural annotations with predominant syntenic regions and observed large inversions are comprehensively illustrated (Fig. 4). A total of 779 syntenic regions were identified, spanning 551,743,349 bps (79.5%) in TMEB117 Hap1 and 547,212,985 bps (82.3%) in TMEB117 Hap2 (Fig. 4a). We identified 110 large inversions distributed across chromosomes. These inversions collectively span 18,588,121 bps (2.7%) in hap1 and 17,017,386 bps (2.6%) in hap2 genomes. Specifically, large inversions of more than 1Mbp were observed in chromosomes 4, 7, 15, and 18, with chromosome 4 having the largest inversion of 6.5Mbp in size (Fig. 4b). In addition to these variations, translocations, duplications, and non-aligned regions demonstrate the structural complexity of the TMEB117 haplotypes. The genome exhibited diverse variations, including 1,608,991 single nucleotide variants and 104 copy number variations, encompassing both copy gains and losses (Additional file 1: Fig. S4). Additionally, chromosome 12 had the highest average length of synteny blocks, measuring 2.1 Mbp (Additional file 1: Fig. S5). It also comprised the most extended synteny blocks in both haplotypes, with lengths of 16.3 Mbp and 16.2 Mbp (hap1 and hap2) towards the end of the chromosome (chr12:35034285–51416241 on hap1 and chr12:33191782–49458244 on hap2).

Comparative analysis of AM560-2 v8.1, TME204 and TMEB117 cassava genomes

As shown in (Fig. 1b), there is a notable observation of a huge 9.7 Mbp insertion region in chromosome 12. To inspect this finding and other structural variations, we conducted a comparative analysis with the whole genome of TME204 haplotypes and cassava reference genome AM560-2 v8.1 (Fig.5). The comparative synteny analysis uncovered distinctive genomic patterns,

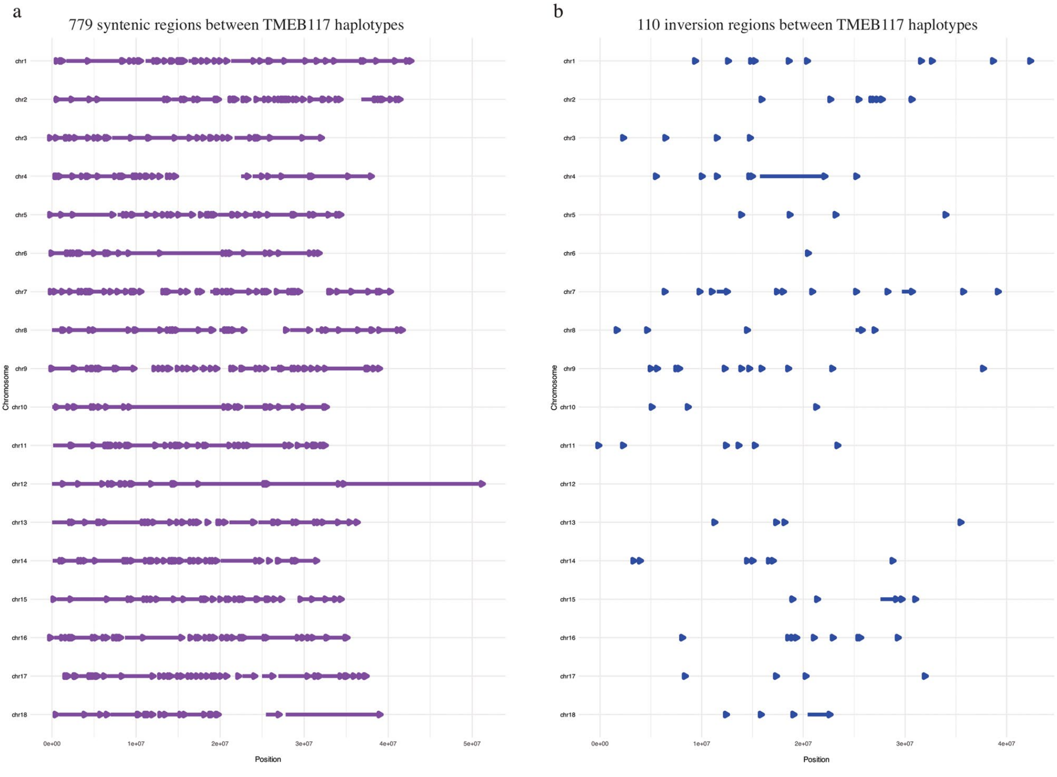


Fig. 4 Syntenic and inversion regions within the TMEB117 genome across its chromosome. (a) Purple arrows between coordinates represent 779 syntenic regions per chromosome of TMEB117 haplotypes. (b) 110 inversion regions highlighted in blue arrows in TMEB117 haplotypes per chromosome.

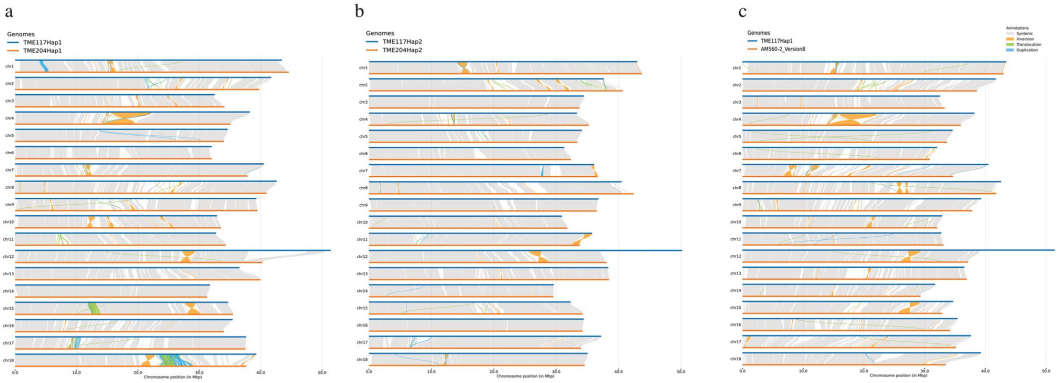


Fig. 5 Synteny plots comparing haplotypes and genomes. (a) TMEB117 hap1 vs. TME204 hap1. (b) TMEB117 hap2 vs. TME204 hap2. (c) TMEB117 hap1 vs. AM560-2 v8. Blue and orange lines represent different genomes, with grey regions showing synteny and gaps indicating differences.

shedding light on structural variations among cassava cultivars. TMEB117 hap1 vs. TME204 hap1 and TMEB117 hap1 vs. AM560-2 v8.1 comparisons (Fig.5 a and c) revealed genomic signatures. The former showed 801 syntenic regions spanning 533,328,857 bps (76.9%) in TMEB117 hap1 and 527,567,427 bps (76%) in AM560-2 v8.1, accompanied by 105 inversions and 657 translocations. The latter unveiled 758 syntenic regions, covering 549,669,393 bps (79.2%) in TMEB117 hap1 and 551,866,350 bps (79.5%) in TME204 hap1, along with 96 inversions and 682 translocations. In the TMEB117 hap2 vs. TME204 hap2 comparison (Fig. 5b), 489 syntenic regions were identified, covering 578,212,474 bps (86.9%) in TMEB117 hap2 and 586,667,066 bps (83%) in TME204 hap2, 80 inversions, and 358 translocations.

Discussion

Large structural variations observed among cassava genotypes emphasize the significant genomic diversity within the species. This study identified a major structural difference in chromosome 12 of the TMEB117 genome: 9.7 Mbp insertion region missing in earlier published cassava genomes. This insertion is characterized by a high repeat content and low gene density, reflecting the complexity of the cassava genome. We validated the presence of this insertion through multiple methods, including genome assembly using three tools (hifiasm, HiCanu, and Flye), pairwise alignment of chromosomal regions, and PCR confirmation of specific fragments. These methods consistently confirmed the insertion and ruled out the possibility of assembly artifacts due to misassembly or scaffolding. The insertion region comprises over 90% TEs, with a significant overrepresentation of the *MUDR-Mutator* TE superfamily [24, 25]. Transposable elements can influence gene regulation and genomic stability, and their abundance in this region may have important implications for genome function and evolution [26].

Comparative mapping coverage analysis of chromosome 12 haplotypes across 16 cassava cultivars using short-read sequencing revealed variability in the insertion region. Some cultivars, such as TME60444, COL2182, CUB40, TMEB117, TME3, TME7, and TME14K, displayed uniform read coverage across the insertion, indicating a conserved genomic structure. In contrast, other cultivars, including AM560-2, Tree cassava, FLA 496-1, and PER226, showed decreased read coverage toward the end of chromosome 12, suggesting potential structural differences. Compared to these, African landraces generally maintained consistent read mapping toward the end of chromosome 12, irrespective of whether they showed uniform or depressed coverage in the insertion region. This highlights the differences in genomic architecture between African and some South American cultivars at this locus. Although the insertion was absent in the final assemblies of some cultivars like AM560-2 and TME7 (Fig. 5a-c, S7), raw data indicated that reads cover the entire region, unlike TME204, which lacked coverage entirely. This suggests difficulties assembling repetitive sequences.

Further analysis of uniquely mapped reads provided more detailed insights into the insertion region's gene features and variability. AM560-2 had a higher read count within the insertion region than other cultivars. However, the sequencing data indicated fewer unique reads in TMEB693, TME204, TMEB419, and ECU41 cultivars. PCR analysis confirmed the presence of fragments of the insertion region in TMEB693 and TMEB419. Due to the non-quantitative nature of the PCR analysis, we could not confirm the lower read counts found by the unique read count analysis in TMEB693 and TMEB419. In fact, in these PCR reactions, the amplification of detectable regions reaches saturation independently of the template amount used. This discrepancy between the sequencing depth and PCR amplification is due to PCR's sensitivity to low-abundance sequences, leading to saturation. To clarify these differences, future studies should employ quantitative PCR (qPCR) [27] to measure the copy number of the insertion sequences accurately. This approach would provide precise quantification and help resolve inconsistencies between sequencing and PCR results, leading to a clearer understanding of the genomic structure of the region.

Our analysis revealed that although no genotype completely lacks the insertion, many exhibit partial coverage. The variability in read mapping patterns suggests that this insertion is a highly polymorphic region, possibly driven by its repetitive sequence content, which may promote recombination events and structural changes. This defines the insertion as a hypervariable [28] locus contributing to cassava cultivars' overall genomic diversity. Future studies should incorporate long-read sequencing technologies across diverse cassava genotypes to better understand this region's complexity. Such approaches could provide more accurate assemblies of highly repetitive regions and help clarify the structural variation observed in this insertion.

Gene ontology enrichment analysis shows two significant unique genes, HDA14 and SRT2, within the genome's insertion region. HDA14, as a histone deacetylase, contributes to removing acetyl groups from histone proteins, promoting a closed chromatin conformation [29]. On the other hand, SRT2, a member of the sirtuin family, also possesses deacetylase activity targeting acetylated proteins, including histones, to regulate chromatin structure by removing acetyl groups from lysine residues on histones [30]. Histone deacetylases induce the deacetylation of histone lysine residues, forming a tighter chromatin structure. The presence of these genes in the insertion region could imply their potential role in influencing the structural arrangement of the genome. Their roles in chromatin modification and plant epigenetics regulations may influence the inserted region's stability and organization. Further exploration of their specific interactions and downstream effects within the insertion region could disclose unknown insights into the functional consequences of large structural variations in the genome.

A genome-wide comparison of the TMEB117 cassava genome with other cultivars revealed extensive structural variation, underlining the complexity and diversity within cassava genomes. Structural variations represent a large component of genetic diversity within the genomes of eukaryotic organisms and can potentially impact the organism's fitness [31]. Our synteny analysis of the TMEB117 cassava genome revealed haplotype-specific structural variations, including large inversions, translocations, and duplications. These findings highlight the structural complexity of the TMEB117 genome. Comparative analysis with other cassava genomes, such as AM560-2 and TME204, further underlines the structural diversity among cultivars. The large 9.7 Mbp insertion on chromosome 12 in TMEB117, absent in the final assemblies of other cultivars but present in raw reads, suggests that such structural variants could be more common in cassava than previously recognized. This large insertion and extensive synteny blocks within this region, consistent across both haplotypes and homologous sequences of the TMEB117 genome (Fig.4 a), also confirm that these are genuine genomic features rather than assembly artifacts.

Conclusions

While this study does not explore the detailed implications of the identified large insertion and other SVs, it does confirm the presence of the large insertion. These findings are a step towards exploring the potential impact of SVs, highlighting the genetic complexity within plant genomes and emphasizing the importance of delving deeper into structural variations to unravel the complexities of genetic diversity and evolutionary processes. The large insertion can be considered a valuable resource for pan-genome analysis of the cassava genome, offering insights into its unique features, primarily its high degree of heterozygosity. Additionally, possible analyses with different traits, such as genome-wide association studies at the SV level, could further explain the functional impact of this region on various phenotypes. A comprehensive understanding of the landscape of SVs offers confidence in uncovering the mechanisms underlying plant adaptation, resilience, tolerance to pathogens, and agronomic traits, guiding informed breeding strategies and crop improvement efforts.

Methods

DNA extraction for sequencing of TMEB117 and TMEB693

To prepare the DNA library, we extracted genomic DNA (gDNA) from expanded leaves of the genotypes TMEB117 and TMEB693 using the NucleoBond® High Molecular Weight DNA kit, following the manufacturer's instructions. We modified the protocol by extending the incubation at 50 °C to 45 min, dramatically increasing the final yield. The extracted gDNA was then sent for library preparation and subsequent sequencing.

DNA extraction for PCR

gDNA was extracted from leaves of TMEB117, TMEB419 and TMEB693 cassava plants by combining the cetyltrimethylammonium bromide (CTAB) extraction method and spin-column-based purification methods.

Total nucleic acid was extracted by using the CTAB (2% CTAB,2% PVP-40, 20 mM Tris–HCl, pH 8.0, 1.4 M NaCl, 20 mM ethylene-diaminetetraacetic acid) extraction protocol as described by Li et al. [32] with some modifications. Approximately 500 mg samples were ground in liquid nitrogen and mixed with 1 ml pre-heated CTAB extraction buffer. Samples were incubated at 65 °C for 15 min with intermittent vortexing, then subjected to centrifugation (16 000 g at 4 °C for 5 min). The supernatant was mixed with an equal volume of cold chloroform: isoamyl alcohol (24:1) before centrifugation (16 000 g at 4 °C for 10 min). The supernatant was added to 0.6 volumes of cold isopropanol and gently mixed to precipitate the nucleic acids. The pellet was washed with 70% ethanol, air-dried, and dissolved in nuclease-free water. After RNaseA treatment, the resulting gDNA was cleaned using the kit DNA Clean & Concentrator™ (Zymo Research) according to the manufacturer's instructions. The gDNA was then sequenced.

PCR validation of the insertion region

The coordinates for the start and end points of the insertion regions in both genome haplotypes were derived using the SyRi (version– 1 0.6.3) [33] outputs obtained from the minimap2 (version– 2.26-r1175) alignments between TMEB117 and TME204 haplotypes. We extracted all the insertion regions labeled 'INS' on chromosome 12 and got the coordinates for the largest insertion within this specific chromosome for both haplotypes. Using hap1 insertion coordinates, we designed primers using the primer3 [34] online tool to get a unique specific primer set that could amplify fragments within the region. It is important to note that these coordinates are based solely on the pairwise alignment of TMEB117 and TME204, and therefore, the exact coordinates of the insertion region may differ in other cultivars.

PCR amplification was conducted in 20 µl of PCR mixture containing 20ng of cassava gDNA, 1x GoTaq Buffer with 1.5 mM $MgCl_2$, 0.04 mM of dNTP, 0.4 µM of each primer, and 0.25 units of GoTaq DNA polymerase (Promega). The thermal cycling conditions were initial denaturation at 94 °C for 1 min, 35 cycles of 30 s at 94 °C, 30 s at 58 °C, and 2 min at 72 °C, and final extension at 72 °C for 5 min. The PCR primers used in this study are shown in Additional file 2: Table S3.

Assembling TMEB117 chromosome 12 using three different assemblers (HiCanu, Flye, and hifiasm)

The TMEB117 genome was assembled using HiCanu (version– 2.3) Flye (version– 2.9.3-b1797). HiCanu is a diploid-aware assembler that produces diploid genomes. In contrast, the Flye assembler generated a single "collapsed" assembly represented a mosaic of both haplotypes. HiCanu's diploid genome was further separated using Purge Haplotigs [35], resulting in a phased-diploid assembly. Hifiasm (version– 0.16.1-r375) haplotype-resolved TMEB117 genome, previously published, was also utilized. BUSCO (version– 5.3.2) and QUAST (version– 5.1) analyses were performed to assess assembly metrics, and these assemblies were used to validate the presence of the insertion region identified through comparative analysis. We aligned chromosome 12 assemblies of the three assembly tools using minimap2 [36], visualized the alignment using plotsr (version– 1.1.2) [37]. Since we have two haplotypes from the hifiasm assembly. Hap1 was compared against the primary assemblies of HiCanu and Flye assemblers to confirm that sequences of the haplotypes are represented within the insertion region. In addition, we compared using dot plots TMEB117 hap1 chromosome 12 to all TMEB117 hap1 chromosomes, unplaced contigs, and *Alternaria alternata* genome [23] to confirm that the existence of the insertion was not a result of assembly error.

Coverage analysis of chromosome 12 with other cassava cultivars

To compare coverage across other cassava genotypes, we collected raw short-read sequences of 13 cassava genotypes from previously published data for our study. These genotypes include wild-type subspecies FLA 496-1 (*Manihot esculenta ssp. fabellifolia*), Tree cassava, a natural hybrid between *Manihot esculenta* and *Manihot glaziovii*, five South American (AM560-2, CUB40, ECU41, COL2182, and PER226), and five Tropical *Manihot esculenta* genotypes (TME204, TME3, TME60444, TME7, and TME14K), one Tropical Manihot Selection (TMS961089A) [15, 16, 19, 38]. We sequenced Illumina short-reads for the TMEB117, TMEB693, and TMEB419 and mapped all the 16 cassava cultivars to the whole genome haplotypes using Burrows-Wheeler Alignment tool (BWA) (version– 0.7.13-r1126) with default parameters [39]. We then extracted the alignment of chromosome 12. We used the mapped short-read data to investigate the coverage of the insertion region and the entirety of chromosome 12 in the TMEB117 genotype using samtools (version– 1.15.1) [40] (parameters: coverage -m, default bin size 40). The coverage is the percentage of positions within a given window with at least one aligned read. Samtools calculates this by computing the number of bases that align to each position within the region and generating a histogram (Additional file 1: Fig. S2).

Read count analysis on gene features within the insertion

In addition to the coverage analysis, we investigated unique reads mapping to gene features within two regions: an 8 Mbp insertion region (chr12:39,000,000–47,000,000 on hap1 and chr12:37,000,000–45,000,000 on hap2) and an 8 Map control region outside the insertion region (chr12:1,000,000–18,000,000). We selected the 8 Mbp length to avoid the borders of the 9.7 Mbp insertion. We maintained a 1 Mbp buffer on either side to prevent including sequences too close to the boundaries, which might be less reliable. Using BWA alignment from coverage analysis, we extracted unique reads mapping the whole genome by excluding alternative and supplementary alignments and filtering out reads with a mapping quality of 0. We systematically analyzed unique mapping reads per gene for each cultivar. First, we extracted the unique mapping reads per gene from BAM files using the featureCounts tool (version– 2.0.6) [41], resulting in a BED file with chromosome, start, end, and count for each gene. Next, we calculated each gene's Reads Per Kilobase (RPK). We then computed a genome-wide scaling factor by summing the RPK values of all genes. We used this factor to normalize the RPK per gene, yielding Transcripts Per Million (TPM), ensuring the total TPM values are summed to a million. Finally, we extracted genes from the two specific regions to compare their TPM values across all cultivars.

Gene enrichment analysis

The genomic gene features were obtained from both haplotypes of the TMEB117 genome on chromosome 12. We extracted all genes in the insertion region, distinguishing known genes from potential ones to enable further examination of function. The known genes within the insertion region underwent a gene ontology (GO) enrichment analysis using the Arabidopsis thaliana database through R's enrichGo function of the clusterProfiler package (version– 4.10.1) [42]. The gene IDs were reformatted to ensure compatibility with the database, and the resulting set of genes underwent enrichment analysis to reveal potential associations with their functions. A BLAST search was also conducted for the potential genes from our annotations. The BLAST outputs were filtered based on percentage identity, query coverage (>90%), and e-value (<1e-05). The resulting RefSeq IDs from the filtered blast outputs were used as inputs for the Entrez database [43] to determine the functions of these genes.

Enrichment analysis of transposable elements superfamilies

Fisher's exact test was performed to determine whether specific transposable element (TE) superfamilies were overrepresented in the insertion region of chromosome 12 compared to the whole genome TE annotations. The null hypothesis assumed a random distribution of TEs across the genome, with no specific enrichment in the insertion region. For each TE superfamily, a contingency table was generated to compare the counts of TEs within the insertion region against those in the rest of the genome. The contingency table for each superfamily was constructed as follows: the number of TEs of the superfamily within the insertion region, the total number of TEs in the insertion region minus the count of the superfamily's TEs, the number of TEs of the superfamily in the genome, and the total number of TEs in the genome minus the count of the superfamily's TEs. Fisher's exact test was then applied to each contingency table to assess the statistical significance of enrichment. The resulting p-values were adjusted using the Benjamini-Hochberg method to account for multiple tests.

Genome-wide comparative analysis across cassava cultivars

Plotsr [37] was employed for visualizing the genome-wide comparisons between TMEB117 haplotypes (hap1 and hap2), followed by comparisons of these haplotypes with three other genomes - AM560-2 v8.1 and TME204 cassava cultivars. Plotsr utilizes minimap2 and SyRI outputs for this analysis. Minimap2 is used for whole-genome alignment, and SyRI [33] is used to identify conserved synteny and structural rearrangements across the genomes. The outputs of SyRI provided insights into the extent of structural variations observed within the TMEB117 genome and across different cassava cultivars.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11523-y.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

We acknowledge Stephan Winter and Cornell University for providing the data for the South American cassava cultivar and TMS961089A genotypes. This study was supported by the SLU Bioinformatics Infrastructure (SLUBI). We thank the Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden, and the International Institute of Tropical Agriculture (IITA) for providing research facilities and other support for this study.

Author contributions

AG and EBR conceived and designed the research. ML analysed the data and wrote the original draft of the manuscript. AVC and LS prepared plant materials and conducted experiments. LF, TS, and AN provided oversight of the analysis and edited the manuscript. The final manuscript underwent review and approval by all authors.

Funding

Open access funding provided by Swedish University of Agricultural Sciences. This work was funded by the Swedish Research Council (VR) funded projects, EpiCass (Identification of epigenetic variations influencing viral resistance and yield in cassava (Manihot esculenta) farmers' fields) grant number 2020–04457 and Cassavanet4Dev (In a race to prevent hunger: Cassava Bioinformatics for African cassava biotechnology) grant number 2021–05105.

Data availability

Raw Illumina reads for TMEB117, TMEB693, and TMEB419 can be accessed from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under BioProject PRJNA1149535, with accession numbers SRR30294079, SRR30294078 and SRR30294077. Additionally, data for other cassava cultivars reused in this study are available under the following accessions: Tree cassava (SRR2847469), AM560-2 (SRR2847385), TME60444 (SRR2847379), FLA 496-1 (SRR2847408), TME3 (SRR1261610), TME7 (SRR16021941), TME14K (SRR2847461) and TME204 (ERR5484651). Stephan Winter's team (Leibniz Institute DSMZ) provided data for South American cultivars (CUB40, ECU41, COL2182, and PER226). Ramu Punna from Cornell University provided data for the TMS961089A genotype. These data are currently restricted to this manuscript and will be available once all teams involved complete their analysis.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable

Competing interests
The authors declare no competing interests.

Received: 28 October 2024 / Accepted: 24 March 2025
Published online: 10 April 2025

References

1. Zhou Y, Minio A, Massonnet M, Solares E, Lv Y, Beridze T, et al. The population genetics of structural variants in grapevine domestication. Nat Plants. 2019;5(9):965–79.
2. Lam HY, Mu XJ, Stutz AM, Tanzer A, Cayting PD, Snyder M, et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol. 2010;28(1):47–55.
3. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14(1):49–61.
4. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10(1):1784.
5. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14(2):125–38.
6. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell. 2020;182(1):145–61. e23.
7. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017;546(7659):524–7.
8. Guan J, Xu Y, Yu Y, Fu J, Ren F, Guo J, et al. Genome structure variation analyses of Peach reveal population dynamics and a 1.67 mb causal inversion for fruit shape. Genome Biol. 2021;22(1):13.
9. Horiguchi G, Gonzalez N, Beemster GT, Inze D, Tsukaya H. Impact of segmental chromosomal duplications on leaf size in the grandifolia-D mutants of Arabidopsis Thaliana. Plant J. 2009;60(1):122–33.
10. Saxena RK, Edwards D, Varshney RK. Structural variations in plant genomes. Brief Funct Genomics. 2014;13(4):296–307.
11. Marroni F, Pinosio S, Morgante M. Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol. 2014;18:31–6.
12. Ceballos Hn, Iglesias CÁ rez JCP, Dixon AGO. Cassava breeding: opportunities and challenges. 2004.
13. Parmar A, Sturm B, Hensel O. Crops that feed the world: production and improvement of cassava for food, feed, and industrial uses. Food Secur. 2017;9(5):907–27.
14. Landi M, Shah T, Falquet L, Niazi A, Stavolone L, Bongcam-Rudloff E, et al. Haplotype-resolved genome of heterozygous African cassava cultivar TMEB117 (Manihot esculenta). Sci Data. 2023;10(1):887.
15. Qi W, Lim YW, Patrignani A, Schlapfer P, Bratus-Neuenschwander A, Gruter S et al. The haplotype-resolved chromosome pairs of a heterozygous diploid African cassava cultivar reveal novel pan-genome and allele-specific transcriptome features. Gigascience. 2022;11.
16. Mansfeld BN, Boyher A, Berry JC, Wilson M, Ou S, Polydore S, et al. Large structural variations in the haplotype-resolved African cassava genome. Plant J. 2021;108(6):1830–48.
17. Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data. 2020;7(1):399.
18. Bredeson JV, Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, et al. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nat Biotechnol. 2016;34(5):562–70.
19. Bredeson JV, Shu S, Berkoff K, Lyons JB, Caccamo M, Santos B, Ovalle T, Bart RS, Lopez-Lavalle AB, Yepes LC, Aranzales M, Wenzl E, Jannink P, Dyer J-L, Rokhsar S. D.S. An improved reference assembly for cassava (Manihot esculenta Crantz). In preparation.
20. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de Novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–5.
21. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 2020;30(9):1291–305.
22. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6.
23. Gai Y, Ma H, Chen Y, Li L, Cao Y, Wang M, et al. Chromosome-Scale genome sequence of alternaria alternata causing alternaria brown spot of citrus. Mol Plant Microbe Interact. 2021;34(7):726–32.
24. Lisch D. Mutator and MULE Transposons. Microbiology Spectrum. 2015.
25. Dupeyron M, Singh KS, Bass C, Hayward A. Evolution of mutator transposable elements across eukaryotic diversity. Mob DNA. 2019;10:12.
26. Rahaman MM, Chen D, Gillani Z, Klukas C, Chen M. Advanced phenotyping and phenotype data analysis for the study of plant growth and development. Front Plant Sci. 2015;6:619.
27. Chen Z, Halford NG, Liu C, Real-Time Quantitative PCR. Primer design, reference gene selection, calculations and statistics. Metabolites. 2023;13(7).
28. González-Candelas F, López-Labrador FX. Hypervariable Region. Brenner's Encyclopedia of Genetics2013. pp. 603-5.
29. Park SY, Kim JS. A short guide to histone deacetylases including recent progress on class II enzymes. Exp Mol Med. 2020;52(2):204–12.
30. Maxwell MM, Tomkinson EM, Nobles J, Wizeman JW, Amore AM, Quinti L, et al. The Sirtuin 2 microtubule deacetylase is an abundant neuronal protein that accumulates in the aging CNS. Hum Mol Genet. 2011;20(20):3986–96.
31. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2020;21(3):171–89.
32. Li R, Mock R, Huang Q, Abad J, Hartung J, Kinard G. A reliable and inexpensive method of nucleic acid extraction for the PCR-based detection of diverse plant pathogens. J Virol Methods. 2008;154(1–2):48–55.
33. Goel M, Sun H, Jiao WB, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20(1):277.
34. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3–new capabilities and interfaces. Nucleic Acids Res. 2012;40(15):e115.
35. Roach MJ, Schmidt SA, Borneman AR. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19(1):460.
36. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
37. Goel M, Schneeberger K, Robinson P. Plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics. 2022;38(10):2922–6.
38. Kuon JE, Qi W, Schlapfer P, Hirsch-Hoffmann M, von Bieberstein PR, Patrignani A, et al. Haplotype-resolved genomes of geminivirus-resistant and geminivirus-susceptible African cassava cultivars. BMC Biol. 2019;17(1):75.
39. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and samtools. Bioinformatics. 2009;25(16):2078–9.
41. Liao Y, Smyth GK, Shi W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30.

42. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov (Camb). 2021;2(3):100141.

43. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. Nucleic Acids Res. 2005;33(Database issue):D54–8.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2025:61

This thesis explores epigenetic variation influencing responses to African cassava mosaic virus in two farmer-preferred African cassava genotypes: TMEB117 (susceptible) and TMEB693 (tolerant). We generated a reference genome for TMEB117, allowing comparative analysis that revealed a large repeat-rich insertion on chromosome 12. Genome-wide methylation profiling showed DNA hypomethylation in TMEB117 upon infection and stable methylation in TMEB693. These findings will support targeted breeding efforts to develop CMD-resistant cassava varieties, ultimately improving food security in Africa.

Michael Kofia Landi received his doctoral studies in the Department of Animal Biosciences. He earned his master's and bachelor's degrees in the Department of Biochemistry at Pwani University in Kenya.

Acta Universitatis Agriculturae Sueciae presents doctoral theses from the Swedish University of Agricultural Sciences (SLU).

SLU generates knowledge for the sustainable use of biological natural resources. Research, education, extension, as well as environmental monitoring and assessment are used to achieve this goal.

ISSN 1652-6880
ISBN (print version) 978-91-8124-045-0
ISBN (electronic version) 978-91-8124-091-7