



# Genome-wide association analysis reveals new insights into the genetic architecture of defensive, agro-morphological and quality-related traits in cassava

Ismail Yusuf Rabbi<sup>1</sup> · Siraj Ismail Kayondo<sup>1</sup> · Guillaume Bauchet<sup>2</sup> · Muyideen Yusuf<sup>1</sup> · Cynthia Idhigu Aghogho<sup>1</sup> · Kayode Ogunpaimo<sup>1</sup> · Ruth Uwugiaren<sup>1</sup> · Ikpan Andrew Smith<sup>1</sup> · Prasad Peteti<sup>1</sup> · Afolabi Agbona<sup>1</sup> · Elizabeth Parkes<sup>1</sup> · Ezenwaka Lydia<sup>3</sup> · Marnin Wolfe<sup>5</sup> · Jean-Luc Jannink<sup>5,6</sup> · Chiedozie Egesi<sup>1,3,4</sup> · Peter Kulakow<sup>1</sup>

Received: 10 January 2020 / Accepted: 20 July 2020 / Published online: 30 July 2020  
© The Author(s) 2020

## Abstract

**Key message** More than 40 QTLs associated with 14 stress-related, quality and agro-morphological traits were identified. A catalogue of favourable SNP markers for MAS and a list of candidate genes are provided.

**Abstract** Cassava (*Manihot esculenta*) is one of the most important starchy root crops in the tropics due to its adaptation to marginal environments. Genetic progress in this clonally propagated crop can be accelerated through the discovery of markers and candidate genes that could be used in cassava breeding programs. We carried out a genome-wide association study (GWAS) using a panel of 5130 clones developed at the International Institute of Tropical Agriculture—Nigeria. The population was genotyped at more than 100,000 SNP markers via genotyping-by-sequencing (GBS). Genomic regions underlying genetic variation for 14 traits classified broadly into four categories: biotic stress (cassava mosaic disease and cassava green mite severity); quality (dry matter content and carotenoid content) and plant agronomy (harvest index and plant type) were investigated. We also included several agro-morphological traits related to leaves, stems and roots with high heritability. In total, 41 significant associations were uncovered. While some of the identified loci matched with those previously reported, we present additional association signals for the traits. We provide a catalogue of favourable alleles at the most significant SNP for each trait-locus combination and candidate genes occurring within the GWAS hits. These resources provide a foundation for the development of markers that could be used in cassava breeding programs and candidate genes for functional validation.

**Keywords** Cassava · Breeding · Genome-wide association · Pest and disease resistance · Agro-morphology and root quality traits · Genotyping-by-sequencing

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11103-020-01038-3>) contains supplementary material, which is available to authorized users.

✉ Ismail Yusuf Rabbi  
I.Rabbi@cgiar.org

<sup>1</sup> International Institute of Tropical Agriculture (IITA), Ibadan 200001, Oyo State, Nigeria

<sup>2</sup> Boyce Thompson Institute, Ithaca, NY 14853, USA

<sup>3</sup> National Root Crops Research Institute (NRCRI), PMB 7006, Umudike 440221, Nigeria

## Introduction

Cassava (*Manihot esculenta* Crantz) is not only the most widely consumed starchy-root staple but also an emerging multi-purpose and industrial crop in Africa, Asia, and Latin America (Parmar et al. 2017). This clonally propagated

<sup>4</sup> Global Development Department, College of Agriculture and Life Sciences, Cornell University, Ithaca, NY 14850, USA

<sup>5</sup> Section on Plant Breeding and Genetics, School of Integrative Plant Sciences, Cornell University, Ithaca, NY 14850, USA

<sup>6</sup> United States Department of Agriculture - Agriculture Research Service, Ithaca, NY 14850, USA

species shows remarkable adaptation to diverse agro-ecologies and can produce reasonable yield under marginal conditions of climate and soil (Jarvis et al. 2012). In addition, its flexible harvest window allows the crop to be left in the soil as a food reserve. These properties make cassava an ideal food security crop with an increasing trend in global production (Prudencio and Al-Hassan 1994; Burns et al. 2010).

In the last four decades, cassava breeding programs across Africa, Asia, and Latin America, have developed varieties that can withstand production constraints including biotic and abiotic stresses with improved yield and starch content (Kawano 2003; Okechukwu and Dixon 2008). While phenotype-based recurrent selection has made significant progress, the rate of genetic gain has been low due to several breeding complexities associated with the biology of the crop, including asynchronous flowering, low seed set per cross, a long cropping cycle of 12 to 24 months and low multiplication rate of planting materials (Ceballos et al. 2012). These challenges hinder the breeding program's abilities to rapidly respond to changing human needs under volatile climatic and environmental conditions.

Modern breeding methods including marker-assisted selection (MAS) and genomic selection (GS) can be used to accelerate genetic improvement particularly by reducing generational interval and increasing selection intensity (Ferguson et al. 2012; Ceballos et al. 2015; García-Ruiz et al. 2016; Bredeson et al. 2016). However, integration of molecular markers as part of MAS in breeding pipelines requires an initial investment in discovery research to identify major-effect loci that serve as the targets of selection. With the rapid advances in next-generation sequencing (NGS) technologies, it is now feasible to generate genome-wide marker data in large populations. This, coupled with phenotype data, makes it possible to identify and map locations of agriculturally important genes and quantitative trait loci (QTL) at the whole genome level (Varshney et al. 2014).

Significant investment has been made in the development of genomic resources for cassava including dense multiparental linkage maps (International Cassava Genetic Map Consortium 2015), annotated reference genomes (Bredeson et al. 2016; Zhang et al. 2018) and a haplotype map of common genetic variants from deep sequencing of hundreds of diverse clones (Ramu et al. 2017). Several genome-wide association studies (GWAS) have been conducted to describe the genetic architecture of resistance against cassava mosaic disease (Wolfe et al. 2016), reduced green mite infestation (Ezenwaka et al. 2018), cassava brown streak disease (Kayondo et al. 2018) and provitamin A and dry matter content (Esuma et al. 2016; Rabbi et al. 2017; Ikeogu et al. 2019).

Using a collection of 5130 elite cassava clones derived from three cycles of recurrent selection in the International Institute of Tropical Agriculture (IITA) Cassava Breeding Program, we examined the genetic architecture of 14

continuous and categorical traits, including defense against biotic stresses, agro-morphological and quality-related traits (Table 1). Among the biotic stress-related traits, we considered cassava mosaic disease (CMD) and cassava green mite (CGM). Caused by different species of cassava mosaic geminiviruses, CMD is one of the most important biotic constraints to cassava production in Africa, India and Sri Lanka (Thottappilly et al. 2003; Alabi et al. 2015; CABI 2019) and has recently spread to the South Asian countries of Thailand and Vietnam (Uke et al. 2018). Infected plants can incur yield losses of up to 82%, which translates to more than 30 million tonnes of fresh cassava roots loss annually (Owor et al. 2004; Legg et al. 2006). Infestation by CGM (*Mononychellus tanajoa*) during the dry season causes chlorosis and restricted growth resulting in a significant negative impact on root yield. The main type of resistance to CGM is attributed to apical leaf pubescence but may also include other mechanisms (Shukla 1976; Byrne et al. 1982; Ezenwaka et al. 2018).

For quality traits, we considered total carotenoid content and dry matter content variation. Biofortification breeding to increase provitamin A carotenoids in storage roots is an important goal in cassava improvement programs around the world (Saltzman et al. 2013). Although the crop's gene pool contains accessions that are naturally rich in provitamin A, they make up a small proportion of cultivated varieties, especially in Africa (Welsch et al. 2010). We used the colour-chart based assessment of total carotenoid content variation. In cassava, the intensity of root yellowness is strongly correlated with total carotenoid content (Sánchez et al. 2014). Additionally, provitamin A is the predominant carotenoid component in cassava (Ceballos et al. 2017).

Dry matter content is a key yield component that determines varietal acceptance by farmers and processors (Bechoff et al. 2018). Cassava germplasm contains considerable variation in percentage dry matter content of the fresh roots ranging from as low as 10–45% (Kawano et al. 1987).

Among agronomic traits, harvest index and plant type were included in the present study. Harvest index reflects the partitioning of resources between the storage roots and above-ground biomass (Sinclair 1998). The desirable harvest index values for the crop range from 0.5 to 0.7 (Kawano et al. 1978, 1998). Plant type in cassava can be characterised by four general descriptive shapes, namely cylindrical, umbrella, compact, and open. Plants with cylindrical shapes do not form branches and are most preferred for mechanised farming. Umbrella plant types generally branch at a high level (above 1 m) and have fewer levels of branching whereas compact and open types are characterised by low first branching height and multiple branching levels but differ in the angle of branches and erectness of the stems (Fukuda et al. 2010).

**Table 1** Name, brief description, and classification of the 14 traits assessed in the present study

| Name of trait                              | Description                                                                                                                   | Trait class   | Time recorded                       |
|--------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|---------------|-------------------------------------|
| Mean cassava mosaic disease (CMD) severity | The visual rating of symptoms caused by cassava mosaic virus. Averages of scores done at 1 and 3 months after planting (MAP). | Biotic stress | 1 and 3 months after planting (MAP) |
| Cassava green mite (CGM) severity          | The visual rating of damage caused by cassava green mite                                                                      | Biotic stress | Peak dry season (February – March)  |
| Apical pubescence                          | Pubescence of young apical leaves                                                                                             | Morphological | 3 MAP                               |
| Leaf shape                                 | The shape of central leaf taken from a mid-height position.                                                                   | Morphological | 3 MAP                               |
| Apical leaf colour                         | Colour of unexpanded apical leaves                                                                                            | Morphological | 3 MAP                               |
| Mature leaf greenness                      | Colour of first fully expanded leaf, an indicator of leaf chlorophyll content                                                 | Morphological | 3 MAP                               |
| Petiole colour                             | Pigmentation of leaf petioles                                                                                                 | Morphological | 3 MAP                               |
| Harvest index                              | Harvest index                                                                                                                 | Agronomic     | During harvest                      |
| Plant type                                 | Plant architecture on a 1–5 scale                                                                                             | Morphological | 9 MAP                               |
| Outer stem colour                          | Stem colour nine months after planting                                                                                        | Morphological | 9 MAP                               |
| Total carotenoid (colour chart)            | Level of yellowness in cassava storage root pulp (parenchyma) due to variation in carotenoid content                          | Quality       | During harvest                      |
| Dry matter content                         | Percentage of dry matter content of storage roots                                                                             | Quality       | During harvest                      |
| Storage root periderm colour               | Colour of the outer surface of storage root periderm (outer skin)                                                             | Morphological | During harvest                      |
| Storage root cortex colour                 | Colour of the outer surface of storage root cortex (inner skin)                                                               | Morphological | During harvest                      |

We also assessed the genetic architecture of several morphological traits related to leaves (petiole colour, apical leaf colour, mature leaf greenness, leaf shape) and storage roots (periderm and cortex colour). Colour of cassava leaf petioles as well as apical leaves ranges from light green to purple due to anthocyanin pigmentation. Anthocyanins play various roles in plants, including protection against ultraviolet light, overcoming different abiotic and biotic stresses and other physiological processes such as leaf senescence (Gould et al. 2008). Mature leaf greenness is related to chlorophyll content, an indicator of a plant's photosynthetic capacity (Palta 1990). Additionally, cassava germplasm exhibits diverse leaf shapes ranging from ovoid lobes to linear forms (Fukuda et al. 2010). This variation is useful as a morphological descriptor and could also play other functional roles related to light capture (Takenaka 1994).

Cassava root periderm colour varies from cream through light brown to dark brown while that of the cortex includes cream, pink and purple. The few but predominant farmer-preferred varieties in Africa often have a pink or purple cortex and dark brown periderm. However, there is no proven genetic correlation between these traits and the culinary properties in cassava. On the other hand, industrial processing into starch and flour production generally utilise whole roots after mechanical periderm removal. These industries

prefer varieties with white/cream periderm to ensure bright-coloured products. Finally, we also assessed the genetic architecture of stem colour, another morphological descriptor used for variety identification and varies from orange to dark-brown (Fukuda et al. 2010).

In the present study, we extend previous GWAS efforts in cassava by assembling a large population of clones that were genotyped at high-density and phenotyped for multiple traits across locations and years. We provide a catalogue of genomic loci associated with the 14 traits, a list of favourable alleles for each trait-locus combination and candidate genes located within the identified loci. Understanding the genetic architecture of variation in the studied traits is an important step towards the development of molecular tools to accelerate the transfer of favourable alleles into farmer-preferred varieties.

## Material and methods

### Field experiments and phenotyping

The population composed of 5130 elite IITA cassava breeding genotypes were phenotyped in a total of 53 field trials (Supplementary Table 1) for the 14 traits related to biotic

stress (cassava mosaic disease and cassava green mite severity); storage root quality (dry matter content and carotenoid content); plant agronomy (harvest index and plant type); and several agro-morphological traits related to leaves, stems and roots (Table 1). Detailed trait description and ontologies are available (<https://cassavabase.org/search/traits>). The trials were planted across four different locations in Nigeria; Ubiaja (6°40'N, 6°20'E), Ibadan (7°24'N, 3°54'E), Mokwa (9°21'N 5°00'E) and Ikenne (6°52'N, 3°42'E) from 2013 to 2016. This population consisted of 715 elite lines from the genetic gain (GG) population, 2,323 full-sib progenies derived from 88 elite GG progenitors (Cycle1—C1), and 2,092 full-sib progenies derived from top 89 cycle one progenitors (Cycle2—C2). The mean family size for C1 clones was 15, ranging from 1 to 77 clones while that of C2 was 7.6 ranging from 1 to 20 clones. The GG population is a collection of the “Tropical *Manihot* Selection (TMS)” cultivars developed in the past five decades by IITA (Dixon and Ssemakula 2008; Okechukwu and Dixon 2008; Ly et al. 2013; Wolfe et al. 2016, 2017; Rabbi et al. 2017). The C1 and C2 populations are part of the IITA genomic selection breeding initiative and were previously described in (Wolfe et al. 2017).

The entire GG collection was evaluated in six clonal evaluation trials (CET) distributed across two locations (Ibadan and Ubiaja) and three seasons (2012–2013, 2013–2014, 2014–2015). In 2012–2013 season, the trial was planted as a preliminary yield trial (PYT) in Ubiaja. A subset of the GG population consisting of 88 clones, mainly parents that were used to generate C1, were evaluated in replicated multi-environment PYTs across 11 environments (three to four locations for three seasons). The C1 population was split into three sets and each evaluated as a CET in Ibadan, Ikenne and Mokwa in 2013–2014 season. In the second season (2014–2015), selections from C1 were planted in the same locations still as CET but using larger plots of 20 plants. In 2015–2016 season, selections from C1 were advanced to several multi-environment advanced yield trials (AYT) planted in one to three locations each. Similar to C1, the C2 population was also split into several subsets and planted as CETs across four locations in 2014–2015 season. Selections from these trials were established as PYT in 2015–2016 season. We also run several multi-generational field trials involving clones from C0 and C1 in CET and AYT, and C0, C1, C2 in a PYT planted across two locations. The CET plots were composed of a single row with five plants per plot, in an augmented block design, with two checks as a control. PYT plots consisted of two rows with ten plants per plot, in a randomised complete block design with two replications and five checks as a control. AYT plots contained four rows with 20 plants per plot, in a randomised complete block design with three reps and five checks as a control. Planting was performed from June to July (during

the rainy season) and harvested around the same time the following year. Spacings between rows and plants were 1 and 0.8 m in all trials, respectively, except in CETs where we used 0.5 m within rows. All field trial management was performed, whenever necessary, following the technical recommendations and standard agricultural practices for cassava (Fukuda et al. 2010; Abass et al. 2014; Atser et al. 2017).

## Genotyping

Genomic DNA was extracted from freeze-dried leaf samples following a modified Dellaporta CTAB method (Dellaporta et al. 1983). DNA quality and quantity were assessed using a Nanodrop 1,000 spectrophotometer at 260 nm absorbance. Genome-wide single nucleotide polymorphism (SNP) data was generated using the genotyping-by-sequencing (GBS) approach described by Elshire et al. (2011). Reduction in genome complexity for GBS was achieved through restriction digestion using ApeKI enzyme (Hamblin and Rabbi 2014). Sequencing reads were aligned to the cassava V6 reference genome (Bredeson et al. 2016) followed by SNP calling using TASSEL GBS pipeline V4 (Glaubitz et al. 2014). SNP calls with less than five reads were masked before imputation using Beagle V4.1 (Browning and Browning 2016). A total of 202,789 biallelic SNP markers with an estimated allelic  $r$ -squared value ( $AR^2$ ) of more than 0.4 were retained for subsequent analyses after imputation.

## Statistical analyses

### Phenotype data analyses

Our interest in this study was to identify the genetic architecture of selected traits rather than location- or year-specific QTLs. We collapsed plot observations for each genotype to a single best linear unbiased prediction (BLUP) using the following mixed linear model (MLM) with the lme4 (Bates et al. 2015) package in R (R Development Core Team 2016):  $y_{ij} = \mu + g_i + \beta_j + r_{j(l)} + \epsilon_{ijl}$ , where  $y_{ij}$  represents a vector of phenotype data,  $\mu$  is the grand mean,  $g_i$  is the random effect of genotype  $i$  with  $g_i \sim N(0, \sigma_g^2)$ ,  $\beta_j$  are the fixed effects of year–location combination  $j$ ,  $r_{j(l)}$  is a random effect of replication nested within location–year combination assumed to be distributed  $N(0, \sigma_r^2)$ ; and  $\epsilon_{ijl}$  is the residual with  $\epsilon_{ijl} \sim N(0, \sigma_e^2)$ .

Pairwise correlation between traits was determined from BLUPs using the R function “*cor*” in the “*stats*” package (R Development Core Team 2016), and visualisation of the correlation matrices was done using the ‘*corrplot*’ R package (Wei and Simko 2017). Due to the unbalanced nature of trials, we calculated the broad-sense heritability estimates

on a plot-mean basis using the formula  $H^2 \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$ . Additive genomic heritability for each trait was also estimated using a linear mixed model as implemented in Genome-Wide Complex Trait Analysis (GCTA) (Yang et al. 2011; Speed et al. 2012).

### Population structure and genetic relatedness assessment

Population structure and cryptic genetic relatedness are known to confound GWAS analysis and lead to spurious associations (Chen et al. 2016). To assess the extent of population structure, we performed a principal component analysis (PCA) using PLINK-v1.9 (Purcell et al. 2007; Rentería et al. 2013). Pairwise Kinship matrix was also calculated using PLINK and visualised using the heatmap function in R.

### Genetic architecture of studied traits

For each trait, single-locus GWAS analysis was carried out using the mixed linear model (MLM) approach implemented in GCTA software (Yang et al. 2011). Because the inclusion of the candidate marker in kinship calculation when controlling for cryptic relatedness can lead to a loss in power (Yang et al. 2011; Listgarten et al. 2012). We also used the mixed linear model which excluded candidate markers via a leave-one-chromosome-out analysis as implemented in GCTA (Yang et al. 2011; Yang et al. 2014). The first approach is referred to as MLMi (“i” for candidate marker included) and the second approach is referred to as MLMe (“e” for candidate marker excluded). Visualisation of MLMi and MLMe results in the form of Manhattan, and quantile-quantile plots were implemented in the ‘CMplot’ R package (LiLin-Yin 2019).

### SNP marker and favourable allelic prediction

To assess the genetic architecture of the studied traits, we fit a linear regression model using peak SNPs at the identified loci as independent variables against the traits’ BLUPS as the response variables. The relative allele substitution effects at each marker were assessed using pairwise *t* test and visualised using boxplots. Here, a locus was defined as a uniquely identifiable genomic region whose SNPs passed genome-wide Bonferroni significance threshold ( $\alpha = 0.05/101,521 = 4.93 \times 10^{-7}$ ).

### Candidate gene identification

Candidate loci were explored using a combination of GWAS *p* values, local linkage disequilibrium (measured as

$r^2$ ), and gene annotation using the gff3 file of the cassava genome available on phytozome v.12.1 (<https://phytozome.jgi.doe.gov/pz/portal.html>) (Goodstein et al. 2011; Batra et al. 2014). This information set was summarised for each candidate loci using Locus zoom (Pruim et al. 2010). To obtain a regional zoom plot for each candidate locus, we built a local SQLite database including the 101,521 biallelic SNP marker matrix, associated GWAS *p*-values for each trait analysed and the cassava gene-annotation following instructions available at [https://genome.sph.umich.edu/wiki/LocusZoom\\_Standalone](https://genome.sph.umich.edu/wiki/LocusZoom_Standalone). Gene codes were shortened to ease visualisation, and whenever possible, *Arabidopsis thaliana* homologues were noted. Recombination information was provided using the same approach as described in Wolfe et al. (2016). A standard interval of 100 kb (50 kb upstream, 50 kb downstream) was explored for each candidate locus and adjusted according to the extent of local linkage disequilibrium with the candidate SNP ( $r^2 < 0.8$ ).

## Results

### Variation and relationships among traits

An analysis of the phenotypic classes of the panel showed that almost all measured traits followed a normal distribution. A couple of traits were slightly skewed towards the tails, particularly the distribution of CMD severity, root periderm and cortex colour (Supplementary Fig. S1). The raw BLUPs values are presented in Supplementary Table S2. The heritability and variance component estimates associated with the studied traits are shown in Table 2. All measured traits exhibited larger than two-fold differences between the maximum and minimum values with a mean coefficient of variation (CV) of 35%. Apical pubescence was the only trait with a significantly larger CV of 113%. These large differences between the maximum and minimum values are an indication of broad genetic variability within the mapping panel. To estimate the influence of additive and non-additive genetic effects on the observed phenotypic values, we estimated broad-sense and genomic-heritabilities. Both the broad-sense and SNP-based heritability estimates were comparable, ranging from low to high: 15 to 78% and 17 to 72%, respectively. Some traits like total carotenoids, dry matter, petiole colour, apical leaf colour, and stem colour had higher SNP-heritability estimates ( $> 0.5$ ) relative to other traits like harvest index, plant type, and resistance to cassava green mite with the lower heritabilities ( $< 0.4$ ). SNP-based heritability is useful in approximating the proportion of phenotypic variance attributable to the additive genetic variation (Yang et al. 2017). The moderate to high levels of SNP-based heritabilities found for traits under active

**Table 2** Broad-sense heritability calculated on a plot-mean basis and SNP heritability estimates, variance components and coefficients of variation of 14 traits in cassava GWAS panel

| Trait                                    | SNP-h <sup>2</sup> | H <sup>2</sup> | σ <sub>g</sub> | σ <sub>g×e</sub> | σ <sub>e</sub> | CV  |
|------------------------------------------|--------------------|----------------|----------------|------------------|----------------|-----|
| CMD severity                             | 0.434              | 0.776          | 0.783          | 0.086            | 0.140          | 63  |
| CGM severity                             | 0.165              | 0.149          | 0.074          | 0.177            | 0.244          | 19  |
| Apical pubescence                        | 0.502              | 0.531          | 0.175          | 0.083            | 0.071          | 113 |
| Leaf shape                               | 0.499              | 0.510          | 0.679          | 0.475            | 0.178          | 22  |
| Apical leaf colour                       | 0.496              | 0.567          | 1.610          | 0.629            | 0.601          | 23  |
| Mature leaf greenness                    | 0.568              | 0.531          | 0.427          | 0.186            | 0.191          | 18  |
| Petiole colour                           | 0.716              | 0.754          | 3.322          | 0.602            | 0.485          | 33  |
| Harvest index                            | 0.308              | 0.538          | 0.010          | 0.002            | 0.007          | 28  |
| Plant type                               | 0.384              | 0.369          | 0.376          | 0.180            | 0.465          | 38  |
| Outer stem colour                        | 0.516              | 0.388          | 0.907          | 0.191            | 1.241          | 22  |
| Total carotenoids content (colour chart) | 0.675              | 0.726          | 0.401          | 0.066            | 0.085          | 49  |
| Dry matter content                       | 0.565              | 0.521          | 14.776         | 3.385            | 10.184         | 15  |
| Root periderm colour                     | 0.548              | 0.610          | 0.190          | 0.035            | 0.086          | 19  |
| Root cortex colour                       | 0.518              | 0.415          | 0.070          | 0.036            | 0.062          | 30  |

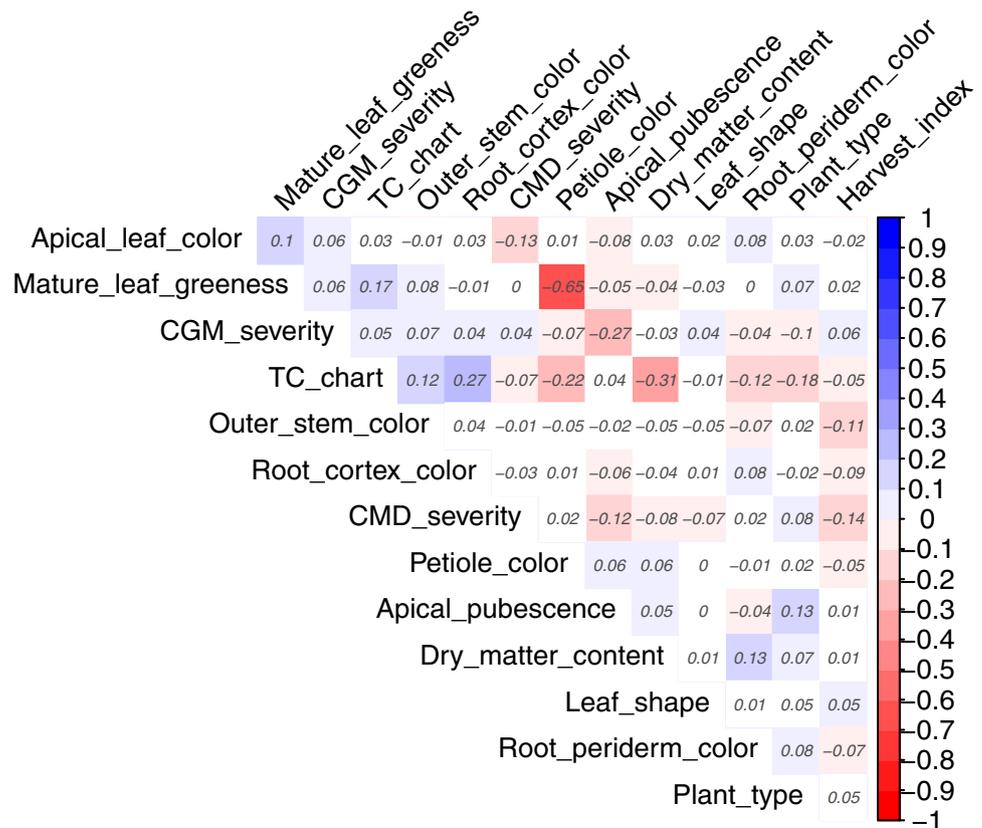
H<sup>2</sup> is the broad-sense heritability, σ<sub>g</sub> is the clonal genotypic variance, σ<sub>g×e</sub> is the variance due to genotype by environment (G × E), and σ<sub>e</sub> is the being the residual variance

selection coupled with sufficient variability in the population indicates good potential for genetic improvement of these traits through recurrent selection.

The relative magnitude of phenotypic correlation pairs ranged from 0.27 between total carotenoid variation by

colour chart (TC-chart) and root cortex colour to – 0.65 for mature leaf greenness and petiole colour (Fig. 1). We detected highly significant negative phenotypic correlations between dry matter and total carotenoid contents (– 0.31, p < 0.001); mean CMD severity and harvest index

**Fig. 1** Heatmap of pairwise trait correlations using BLUPS from the 14 traits



( $r = -0.14$ ,  $p < 0.001$ ); and apical pubescence and CGM severity ( $r = -0.27$ ,  $p < 0.001$ ). There were highly significant positive correlations between total carotenoids and periderm colour ( $r = 0.27$ ,  $p < 0.001$ ), leaf greenness ( $r = 0.17$ ,  $p < 0.001$ ), and stem colour ( $r = 0.12$ ,  $p < 0.001$ ).

### Distribution of SNP markers and population stratification

Genotyping of the panel enabled the identification of 202,789 imputed SNP markers. Upon filtering markers with minor allele frequency less than 1%, we retained 101,521 SNPs. All the 101 K SNPs were mapped onto the 18 chromosomes covering a total of 532.5 Mb of the cassava genome and a SNP density of 5.6 variants/Kbp (Fig. 2). Individually, the SNP coverage per chromosome ranged from 3,821 on chromosome 7 to 11,189 on chromosome 1. The average minor allele frequency was 0.196, and 80% of the SNPs had frequency greater than 5%, indicating enrichment of common SNP alleles in the population (Fig. 2). To estimate the mapping resolution for our panel, we assessed pairwise linkage disequilibrium (LD) between 101,521 SNPs across the 5130 cassava accessions. We used the mean  $r^2$  value as an estimate of LD decay using a window of 1 Mb, followed by fitting a non-linear regression curve of LD versus distance. The whole-genome LD decay peaked at  $r^2$  of 0.349 and dropped to an  $r^2$  of 0.212 at a distance of 10 kb (Fig. 2).

Principal component analysis (PCA) was conducted to visualise the extent and degree of population structure present within the panel. While the first two principal components (PC) explained 7.4% of the total phenotypic variation, PC3 to PC10 together explained about 14% of the total variation (Fig. 3). We observed a considerable overlap

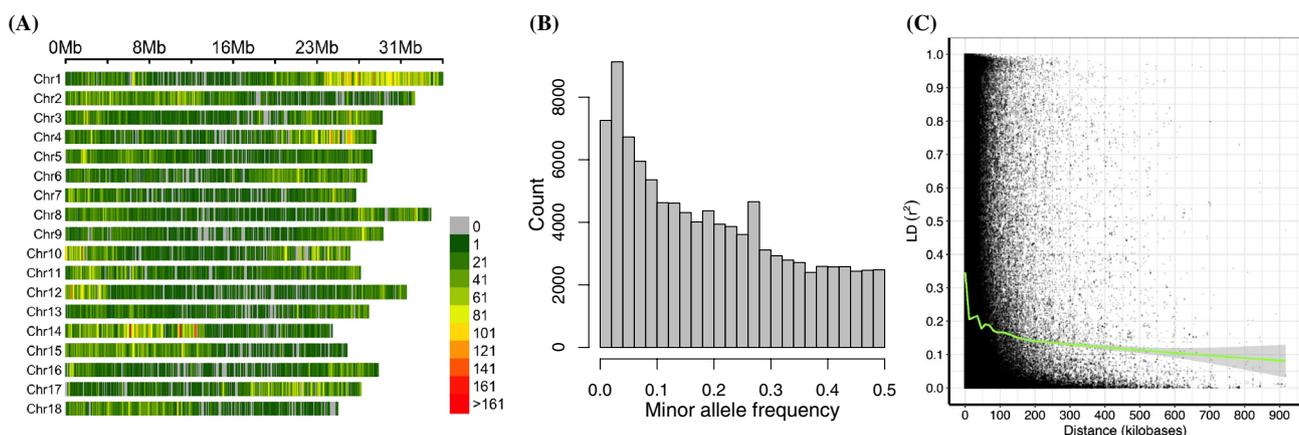
between cycles (C0, C1 and C2); hence, no distinct clusters were detected in the panel. The results of the identity-by-state distance showed a considerable range of relationship in the association panel with an average value of 0.23 and ranged from 0.02 to 0.32. We observed familial relationships along the diagonal with a few large blocks of closely related individuals. The off-diagonal part of the relationship matrix indicated low kinship (Fig. 3).

### Genome-wide association results

We identified a total of 27 unique genomic regions significantly associated with variation in the 14 studied traits following the MLMi analysis (Fig. 4). Additional loci were uncovered for a majority of the traits when we considered the MLMe approach, bringing the total number of loci to 41 (Supplementary Figure S2). The most significant trait-marker associations from MLMi and MLMe for each trait and genomic region combinations are also provided in Table 3. In the following sections, we present the results and provide discussion for each trait, starting with economically important traits to morphological traits.

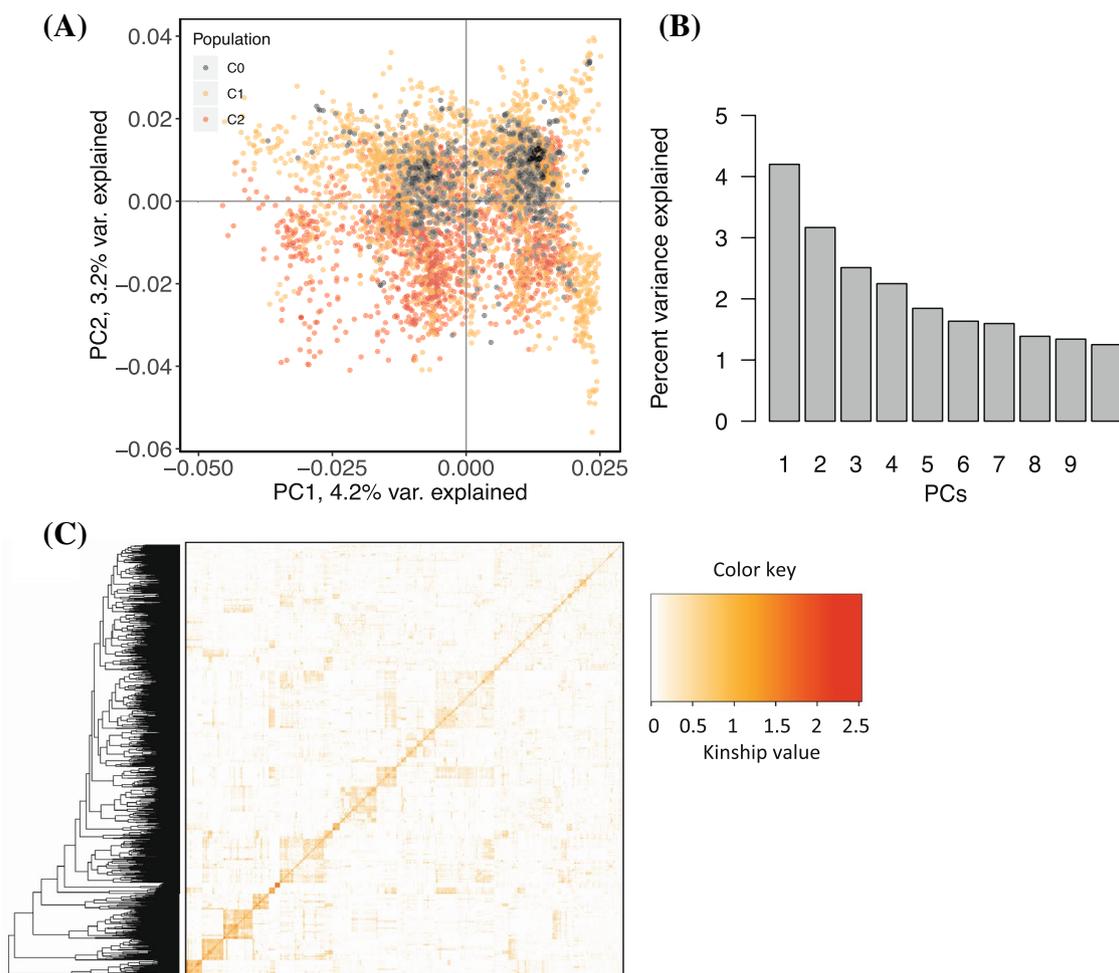
### Cassava mosaic disease severity

The genetic basis of CMD resistance has been studied extensively using bi-parental linkage mapping (Rabbi et al. 2014) and GWAS (Wolfe et al. 2016). These studies repeatedly showed that the main resistance to the disease is conferred by a single major gene on chromosome 12, which is widely known as CMD2 locus (Akano et al. 2002). In the present study, we uncovered the same locus and two other loci on chromosome 14. The CMD2 region on chromosome 12 is



**Fig. 2** Overview of SNP genotyping data. **a** The density of SNPs on the 18 chromosomes of the cassava association mapping panel within 100 kb window. **b** Histogram of minor allele frequency distribution. **c** Genome-wide linkage disequilibrium (LD) decay for the cassava

accessions in the panel showing the squared correlations ( $r^2$ ) between markers by marker physical distance (kb). The blue smoothing curve (LOESS) and the average LD were fitted to the LD decay



**Fig. 3** An assessment of population structure based on Principal Components Analysis of 101,521 SNP marker data (MAF > 1%) for 5130 individual cassava clones. **a** Scatter plot of the first two PCs; **b**

the proportion of genetic variation explained by the first 10 PCs; and **c** Heatmap showing pairwise Kinship matrix

tagged by marker S12\_7926132 ( $p$  value =  $1.7 \times 10^{-112}$ ). The favourable allele at this marker (T) was found to occur at high frequency in the population ( $\text{Freq} > 0.66$ ). The two additional loci on chromosome 14 are tagged by markers S14\_4626854 ( $p$  value =  $1.7 \times 10^{-14}$ ) and S14\_9004550 ( $p$  value =  $4.2 \times 10^{-17}$ ). The SNP at the CMD2 locus had a larger effect ( $\beta$  value 0.82) compared to the two new loci on chromosome 14 ( $\beta$  value of 0.23 and 0.28).

### Cassava green mite severity

We found four genomic regions associated with CGM resistance in our panel (Fig. 4) of which only marker S8\_6409580, occurring around 6.41 Mb region of chromosome 8, was previously reported (Ezenwaka et al. 2018). The remaining loci were found on chromosomes 1, 12 and 17 (Table 3). We note that except for the markers on chromosome 8, which were significant in both MLMi and MLMe, the remaining

loci were only significant in the MLMe analysis. However, their SNP effects ( $\beta$  statistic) were very similar in the two analyses. Association analysis of the related trait, apical pubescence identified significant loci on five chromosomes, two of which co-locate with same regions underlying resistance to CGM on chromosomes 8 (S8\_6409580) and 12 (S12\_5524524). The genetic correlation between the two traits is  $-0.51$  in the population (Supplementary Table S3). The other loci were on chromosomes 9 (S9\_1588034), 11 (S11\_5727254), and 16 (S16\_1501762).

### Carotenoid content

GWAS analysis of colour-chart based variation in root carotenoid content revealed a major locus on chromosome 1 tagged by three markers around 24.1, 24.6 and 30.5 Mb regions with the top markers being S1\_24159583, S1\_24636113 and S1\_30543962, respectively. Previous

GWAS analyses reported significant associations in this region (Esuma et al. 2016; Rabbi et al. 2017; Ikeogu et al. 2019). Besides, we uncovered five new genomic regions associated with this trait on chromosome 5 (around 3.38 Mb), 8 (two peaks at 4.31 and 25.59 Mb regions), and 15 (7.65 Mb), and 16 (0.48 Mb). We note that the regions in the last three chromosomes, including two on chromosome 8, were only detected via the MLMe analysis.

### Dry matter content

GWAS for variation in dry matter content following the MLMi analysis revealed two major loci, of which one was previously reported (Rabbi et al. 2017). The most significant locus occurs on chromosome 1 around 24.64 Mb region and is tagged by marker S1\_24636113 ( $p$  value =  $5.0 \times 10^{-8}$ ). The second locus was tagged by marker S6\_20589894 ( $p$  value =  $3.4 \times 10^{-7}$ ). Additional loci on chromosomes 12 (S12\_5524524,  $p$  value =  $8.0 \times 10^{-12}$ ), 15 (S15\_1012346,  $p$  value =  $4.0 \times 10^{-17}$ ), and 16 (S16\_25663808,  $p$  value =  $4.2 \times 10^{-12}$ ) were detected from the MLMe analysis. Along with the new locus on chromosome 6, these regions have not been previously reported to be associated with dry matter content variation. They are suitable for further genetic studies including identification of underlying candidate genes.

### Harvest index

The MLMi-based GWAS for harvest index, the ratio of fresh root weight to total plant weight, uncovered two genomic regions that are significantly associated with the trait. The first peak is in chromosome 2, tagged by SNP S2\_2809137 ( $p$  value =  $3 \times 10^{-8}$ ). The second locus occurred on chromosome 12 with SNP S12\_6055806 showing the strongest association with the trait ( $p$  value =  $5.4 \times 10^{-24}$ ). Analysis of the same trait using the MLMe approach uncovered several other regions scattered across chromosomes 3, 4, 6, 8, 9, 14, 15, 16, and 18.

### Morphological traits

GWAS for outer cortex colour uncovered two association signals located on chromosomes 1 (3.05 Mbp region) and 2 (6.56 Mbp region) and tagged by SNPs S1\_3047840 ( $p$  value  $1.6 \times 10^{-92}$ ) and S2\_6566608 ( $p$  value  $5.0 \times 10^{-83}$ ), respectively. A single genomic region on chromosome 3 (4.54 Mbp) was found to be linked to periderm colour. The most significant SNP at this locus (S3\_4545411) had a  $p$  value of  $1.4 \times 10^{-123}$ .

Two association signals were detected for plant type from MLMe while no marker passed the Bonferroni threshold in the MLMi analysis. The detected loci jointly

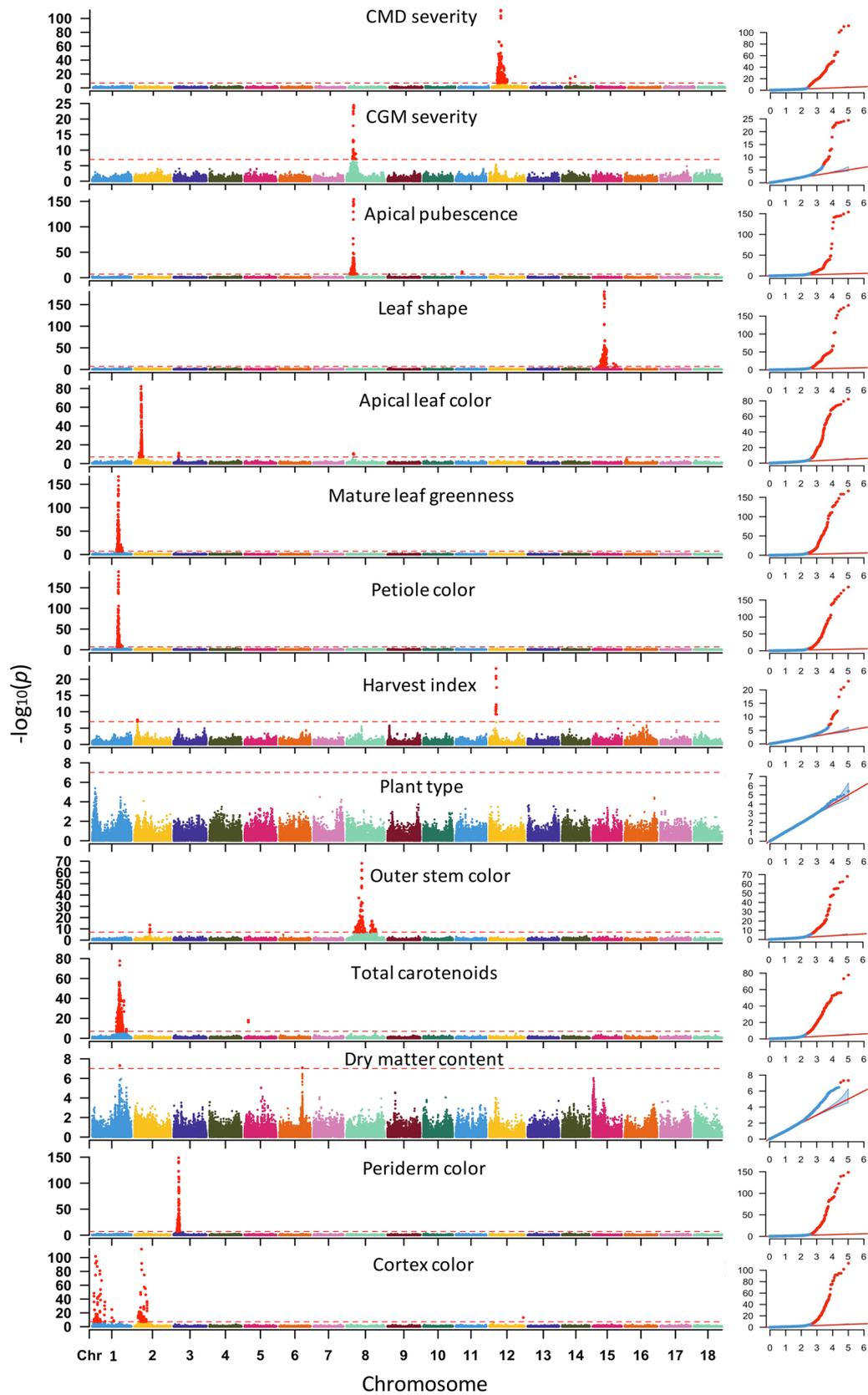
occurred on chromosome 1 at 2.19 Mbp region (tagged by SNP S1\_3192405,  $p$  value  $3.82 \times 10^{-8}$ ) and 25.30 Mbp region (S1\_25303195,  $5.25 \times 10^{-14}$ ). Three loci were found to be associated with stem colour variation, one in chromosome 2 and two on chromosome 8. The most significant locus was around 13.6 Mbp region of chromosome 8 and is tagged by marker S8\_13604799 ( $p$  value  $8.3 \times 10^{-69}$ ). The other two loci were tagged by SNPs S2\_13928566 ( $p$  value  $3.6 \times 10^{-14}$ ) and S8\_22630799 ( $p$  value  $3.3 \times 10^{-17}$ ).

Genetic architecture of leaf morphology traits showed that one to three major loci control them, indicating simple genetic architecture. We found a single genomic region around 23.45 Mbp of chromosome 1 to be associated with leaf petiole colour and is tagged by SNP S1\_23452638 ( $p$  value  $9.8 \times 10^{-180}$ ). Notably, the same SNP was found to underlie the variation in mature leaf greenness. It is therefore not surprising that these traits are negatively correlated in our population. SNP effect analysis showed that while allele “T” at S1\_23452638 had a positive effect on petiole colour, the same allele showed a negative effect on leaf greenness. Regression of the marker on the traits for leaf colour and petiole colour returned an  $R^2$  0.57 and 0.62, respectively.

Variation in the colour apical leaves was found to be associated with three loci occurring on chromosomes 2, 3, and 8. The most significant marker was S2\_6086714 ( $p$  value  $6.1 \times 10^{-83}$ ) followed by the markers on chromosome 8 (S8\_6061421,  $p$  value  $1.9 \times 10^{-11}$ ) and 3 (S3\_4745233,  $p$  value  $4.4 \times 10^{-9}$ ). Multiple regression returned an  $R^2$  of 0.31 for this trait suggesting either a more complex architecture or imprecise scoring of the variation in the trait. The GWAS result for leaf shape uncovered two major loci on chromosome 15 occurring around 10.27 Mbp and 20.57 Mbp regions. The first peak tagged by SNP S15\_10273255 was highly significant ( $3.7 \times 10^{-174}$ ) while the second peak was tagged by S15\_20573383 ( $p$  value  $1.8 \times 10^{-11}$ ). Fitting linear model with the two top SNPs for this trait returned  $R^2$  of 0.40.

### Genetic architecture of studied traits

To assess the genetic architecture of the studied traits, we fit a linear model using peak SNPs at the identified loci (Table 3) as independent variables against the traits BLUPS as the response variables. Peak SNPs at the identified loci explained approximately 34% of the trait variation on average, and  $R^2$  ranged from 5 to 62% (Fig. 5). Markers for cassava green mite severity, harvest index, plant type and dry matter content had the lowest predictive ability ( $R^2 < 0.1$ ). Most of the morphology and colour related traits for leaves and roots had between 1 and 3 peaks of association except apical leaf pubescence with six loci. Peaks associated with



**Fig. 4** Manhattan plots for GWAS for 14 traits of 5130 cassava accessions using MLM analysis approach. A total of 101,521 SNP markers were used for the GWAS analyses with the red horizontal line representing Bonferroni adjusted genome-wide significance threshold ( $\alpha = 0.05/101,521 = 4.93 \times 10^{-7}$ ). The QQ-plots inset—right with observed  $p$  values on the  $y$ -axis and expected  $p$  values on the  $x$ -axis

variation in total carotenoid content and resistance to CMD had a large effect ( $R^2 = 0.60$  and  $0.45$ , respectively). The major loci controlling these traits had known candidate genes reported previously. Still, the new loci identified in these as well as the other traits are attractive candidates for follow-up studies.

### Identification of favourable alleles

To identify favourable alleles for traits under selection in the breeding pipeline, the most significantly associated SNP (lowest  $p$ -value) at each major-effect locus were chosen. Allelic substitution effects for these markers are shown in Supplementary Figure S3. Selected traits include resistance to CMD and CGM, increased dry matter and carotenoid content. For CMD resistance, the haplotypes at the top SNPs S12\_7926132 (allele G/T), S14\_4626854 (A/G) and S14\_9004550 (T/C) are *T-A-C*, respectively. Two hundred fifty-one genotypes were fixed for the favourable alleles in the population. Their average CMD severity BLUP was the lowest among all haplotype combinations (mean =  $-0.56$ ,  $SD = 0.29$ ). We also note dominance allele effect at the first SNP which is linked to the CMD2 locus, which agrees with previous results from biparental QTL studies (Akano et al. 2002; Rabbi et al. 2014).

The CGM resistance-linked haplotype for SNPs S1\_28672656 (A/T), S8\_6409580 (C/G), S12\_5524524 (C/T) and S17\_23749968 (A/G) are *T-C-C-G*, respectively. Alleles associated with increased pubescence particularly for loci that co-locates with CGM resistance of chromosomes 8 (S8\_6409580 (C/G)) and 12 (S12\_5524524 (C/T)) are *C-T*.

Several SNP loci from chromosome 1 (S1\_24159583, S1\_24636113, S1\_30543962), 5 (S5\_3387558), 8 (S8\_4319215, S8\_25598183), 15 (S15\_7659426), and 16 (S16\_484011) were associated with increased carotenoid content. The favourable haplotype for chromosome 1 SNPs are *T-G-G*, respectively. In contrast, that for chromosome 5 is *T*. The favourable haplotype for the two loci on chromosome 8 is *A-T* while for chromosomes 15 and 16 SNPs it is *T-T*, respectively. No individuals in the population were fixed for the favourable alleles across all the SNPs.

The favourable allele at the most significant dry matter locus on chromosome 1 (S1\_24636113 allele G/A) is *A*. We note that this marker also co-located with the major locus for carotenoid content and allele *A* has a non-favourable effect in carotenoid content. The favourable haplotypes

at the other loci on chromosomes 6 (S6\_20589894 G/A), 12 (S12\_5524524, C/T), 15 (S15\_1012346 C/T) and 16 (S16\_25663808 T/C) are *G-C-T-C*. A total of 61 genotypes fixed for the favourable alleles at the SNPs on chromosomes 1, 6 and 12. Their average dry matter BLUP was the highest among all haplotype combinations (mean =  $2.62$ ,  $SD = 3.66$ ). Finally, pairwise comparisons of genotypic classes at each locus were found to be significant ( $t$  test;  $p$  value  $\leq 0.05$ ) across all loci except for SNP S16\_484011 for carotenoid content and S16\_25663808 for dry matter content (Supplementary Figure S3).

### Candidate gene identification

The most significant GWAS peaks were further investigated for the presence of potential candidate genes using local linkage disequilibrium, coupled to gene annotation. We matched the highly significant SNP markers identified in the MLME analysis with gene annotation in the regions up and downstream derived from phytozome online database. Seventeen candidate genes that colocalised with the identified putative genomic loci on height chromosomes were retrieved from the cassava genome available on phytozome v.12.1 and are presented in Table 4.

Several of these identified genes were further selected and highlighted based on their significance within a given biological pathway. Additionally, we provide regional Manhattan plots for each locus-trait combination in Supplementary Figure S4. These plots include candidate genes within 100 kb of the top SNP marker (50 kb upstream, 50 kb downstream) with some adjustments based on the extent of local linkage disequilibrium with the candidate SNP.

Three genes associated with total carotenoid content were identified on chromosomal regions 1, 5 and 15. Of these, Manes.01G124200 (Phytoene synthase) occurred within the previously reported genomic region (Esuma et al. 2016; Rabbi et al. 2017). Manes.01G124200 gene has a transferase enzymatic activity critical in the carotenoid biosynthesis pathway. The other two genes are novel: Manes.05G051700 and Manes.15G102000 (both of which are beta-carotene dioxygenases) located at 3.87 Mb on chromosome 5 and 7.58 Mb on chromosome 15, respectively, and are also known to play major roles in carotenoid biosynthesis.

For dry matter content, we identified two genes involved in starch and sucrose metabolism occurring 600 Kbp away from top SNP S1\_24636113 in chromosome 1: Manes.01G123000 (UDP-glucose pyrophosphorylase) and Manes.01G123800 (sucrose synthase) previously reported (Rabbi et al. 2017). Although these genes are a few hundred Kb away from the top SNP, this particular genomic region of chromosome 1 is known to harbour extensive LD in Africa cassava germplasm (Rabbi et al. 2017). Other candidate genes found close to the top SNPs on chromosomes 6, 15

**Table 3** Summary statistics of most significant SNP markers at each major trait linked locus for the 14 traits

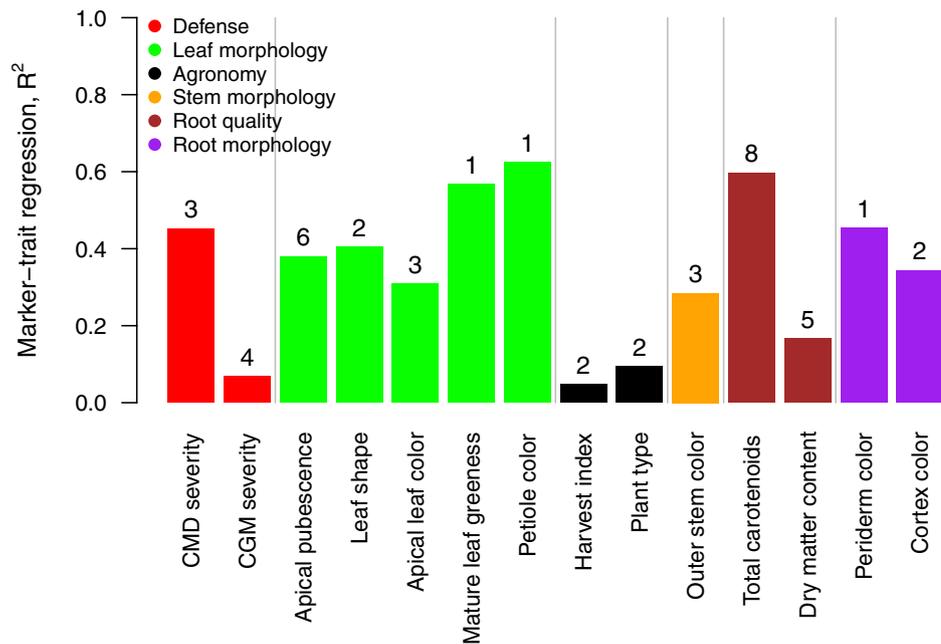
| Trait                           | SNP          | Minor allele | Major allele | MAF  | Beta  | SE   | p value                                  | Beta* | SE*  | p value*                                 |
|---------------------------------|--------------|--------------|--------------|------|-------|------|------------------------------------------|-------|------|------------------------------------------|
| CMD severity                    | S12_7926132  | G            | T            | 0.44 | 0.82  | 0.04 | <b><math>1.7 \times 10^{-112}</math></b> | 0.89  | 0.02 | <b><math>P \approx 0.0</math></b>        |
| CMD severity                    | S14_4626854  | A            | G            | 0.40 | -0.23 | 0.03 | <b><math>1.7 \times 10^{-14}</math></b>  | -0.23 | 0.03 | <b><math>1.0 \times 10^{-14}</math></b>  |
| CMD severity                    | S14_9004550  | T            | C            | 0.17 | 0.28  | 0.03 | <b><math>4.2 \times 10^{-17}</math></b>  | 0.27  | 0.03 | <b><math>1.7 \times 10^{-16}</math></b>  |
| CGM severity                    | S1_28672656  | A            | T            | 0.28 | 0.09  | 0.04 | $1.4 \times 10^{-2}$                     | 0.10  | 0.02 | <b><math>1.1 \times 10^{-10}</math></b>  |
| CGM severity                    | S8_6409580   | C            | G            | 0.41 | -0.18 | 0.02 | <b><math>3.6 \times 10^{-25}</math></b>  | -0.18 | 0.01 | <b><math>4.7 \times 10^{-37}</math></b>  |
| CGM severity                    | S12_5524524  | C            | T            | 0.38 | -0.07 | 0.02 | $6.3 \times 10^{-5}$                     | -0.07 | 0.01 | <b><math>1.3 \times 10^{-7}</math></b>   |
| CGM severity                    | S17_23749968 | A            | G            | 0.31 | 0.09  | 0.02 | $1.5 \times 10^{-5}$                     | 0.08  | 0.01 | <b><math>1.1 \times 10^{-7}</math></b>   |
| Apical pubescence               | S8_6409580   | C            | G            | 0.41 | 0.41  | 0.02 | <b><math>1.1 \times 10^{-154}</math></b> | 0.42  | 0.01 | <b><math>4.3 \times 10^{-293}</math></b> |
| Apical pubescence               | S9_1588034   | A            | G            | 0.47 | 0.08  | 0.02 | <b><math>4.2 \times 10^{-7}</math></b>   | 0.08  | 0.01 | <b><math>2.0 \times 10^{-14}</math></b>  |
| Apical pubescence               | S11_5727254  | A            | G            | 0.09 | -0.19 | 0.03 | <b><math>1.9 \times 10^{-12}</math></b>  | -0.19 | 0.03 | <b><math>1.8 \times 10^{-12}</math></b>  |
| Apical pubescence               | S12_5524524  | C            | T            | 0.38 | 0.07  | 0.02 | $2.3 \times 10^{-4}$                     | 0.09  | 0.01 | <b><math>6.8 \times 10^{-18}</math></b>  |
| Apical pubescence               | S16_1501762  | G            | A            | 0.30 | -0.06 | 0.02 | $2.8 \times 10^{-3}$                     | -0.06 | 0.01 | <b><math>8.1 \times 10^{-8}</math></b>   |
| Leaf shape                      | S15_10273255 | A            | G            | 0.03 | 2.70  | 0.10 | <b><math>3.7 \times 10^{-174}</math></b> | 2.90  | 0.08 | <b><math>P \approx 0.0</math></b>        |
| Leaf shape                      | S15_20573383 | G            | C            | 0.02 | 0.73  | 0.11 | <b><math>1.8 \times 10^{-11}</math></b>  | 1.47  | 0.10 | <b><math>1.6 \times 10^{-52}</math></b>  |
| Apical leaf colour              | S2_6086714   | A            | T            | 0.43 | -1.22 | 0.06 | <b><math>6.1 \times 10^{-83}</math></b>  | -1.26 | 0.04 | <b><math>5.7 \times 10^{-267}</math></b> |
| Apical leaf colour              | S3_4745233   | G            | A            | 0.23 | -0.35 | 0.06 | <b><math>4.4 \times 10^{-9}</math></b>   | -0.34 | 0.04 | <b><math>1.7 \times 10^{-18}</math></b>  |
| Apical leaf colour              | S8_6061421   | G            | C            | 0.41 | -0.32 | 0.05 | <b><math>1.9 \times 10^{-11}</math></b>  | -0.27 | 0.03 | <b><math>2.4 \times 10^{-16}</math></b>  |
| Mature leaf greenness           | S1_23452638  | T            | A            | 0.16 | -1.11 | 0.04 | <b><math>2.7 \times 10^{-167}</math></b> | -1.24 | 0.03 | <b><math>P \approx 0.0</math></b>        |
| Petiole colour                  | S1_23452638  | T            | A            | 0.16 | 2.73  | 0.10 | <b><math>9.8 \times 10^{-180}</math></b> | 2.73  | 0.10 | <b><math>9.8 \times 10^{-180}</math></b> |
| Harvest index                   | S2_2809137   | G            | T            | 0.09 | -0.04 | 0.01 | <b><math>3.0 \times 10^{-8}</math></b>   | -0.04 | 0.01 | <b><math>5.6 \times 10^{-15}</math></b>  |
| Harvest index                   | S12_6055806  | A            | G            | 0.24 | -0.03 | 0.00 | <b><math>5.4 \times 10^{-24}</math></b>  | -0.02 | 0.00 | <b><math>2.0 \times 10^{-17}</math></b>  |
| Plant type                      | S1_3192405   | T            | C            | 0.28 | 0.20  | 0.05 | $1.7 \times 10^{-5}$                     | 0.22  | 0.04 | <b><math>3.8 \times 10^{-8}</math></b>   |
| Plant type                      | S1_25303195  | C            | A            | 0.48 | 0.23  | 0.05 | $3.4 \times 10^{-5}$                     | 0.27  | 0.04 | <b><math>5.3 \times 10^{-14}</math></b>  |
| Outer stem colour               | S2_13928566  | C            | G            | 0.39 | -0.44 | 0.10 | <b><math>3.6 \times 10^{-14}</math></b>  | -0.45 | 0.06 | <b><math>6.5 \times 10^{-15}</math></b>  |
| Outer stem colour               | S8_13604799  | G            | A            | 0.34 | -1.06 | 0.06 | <b><math>8.3 \times 10^{-69}</math></b>  | -1.06 | 0.05 | <b><math>8.7 \times 10^{-89}</math></b>  |
| Outer stem colour               | S8_22630799  | G            | A            | 0.25 | -0.58 | 0.07 | $3.3 \times 10^{-17}$                    | -0.76 | 0.06 | <b><math>1.3 \times 10^{-37}</math></b>  |
| Total carotenoids (color chart) | S1_24159583  | T            | C            | 0.30 | 0.27  | 0.02 | <b><math>5.3 \times 10^{-57}</math></b>  | 0.37  | 0.01 | <b><math>1.3 \times 10^{-219}</math></b> |
| Total carotenoids (color chart) | S1_24636113  | G            | A            | 0.23 | 0.49  | 0.03 | <b><math>1.8 \times 10^{-78}</math></b>  | 0.57  | 0.02 | <b><math>1.3 \times 10^{-270}</math></b> |
| Total carotenoids (color chart) | S1_30543962  | G            | A            | 0.10 | 0.18  | 0.03 | <b><math>6.3 \times 10^{-8}</math></b>   | 0.40  | 0.02 | <b><math>2.4 \times 10^{-76}</math></b>  |
| Total carotenoids (color chart) | S5_3387558   | T            | C            | 0.10 | 0.21  | 0.02 | <b><math>5.5 \times 10^{-17}</math></b>  | 0.20  | 0.02 | <b><math>2.0 \times 10^{-16}</math></b>  |
| Total carotenoids (color chart) | S8_4319215   | A            | C            | 0.01 | 0.22  | 0.05 | $5.2 \times 10^{-5}$                     | 0.23  | 0.04 | <b><math>3.4 \times 10^{-8}</math></b>   |
| Total carotenoids (color chart) | S8_25598183  | T            | G            | 0.03 | 0.18  | 0.04 | $6.3 \times 10^{-6}$                     | 0.18  | 0.03 | <b><math>1.1 \times 10^{-8}</math></b>   |
| Total carotenoids (color chart) | S15_7659426  | G            | T            | 0.26 | -0.07 | 0.02 | $2.0 \times 10^{-3}$                     | -0.06 | 0.01 | <b><math>3.2 \times 10^{-7}</math></b>   |
| Total carotenoids (color chart) | S16_484011   | G            | T            | 0.35 | 0.09  | 0.03 | $7.6 \times 10^{-4}$                     | 0.05  | 0.01 | <b><math>8.9 \times 10^{-8}</math></b>   |
| Dry matter content              | S1_24636113  | G            | A            | 0.23 | -1.32 | 0.24 | <b><math>5.0 \times 10^{-8}</math></b>   | -1.68 | 0.14 | <b><math>1.7 \times 10^{-33}</math></b>  |
| Dry matter content              | S6_20589894  | G            | A            | 0.48 | 0.85  | 0.17 | <b><math>3.4 \times 10^{-7}</math></b>   | 0.78  | 0.09 | <b><math>1.7 \times 10^{-16}</math></b>  |
| Dry matter content              | S12_5524524  | C            | T            | 0.37 | 0.68  | 0.17 | $9.8 \times 10^{-5}$                     | 0.69  | 0.10 | <b><math>8.0 \times 10^{-12}</math></b>  |
| Dry matter content              | S15_1012346  | C            | T            | 0.44 | -0.95 | 0.20 | $2.2 \times 10^{-6}$                     | -0.84 | 0.10 | <b><math>4.0 \times 10^{-17}</math></b>  |
| Dry matter content              | S16_25663808 | T            | C            | 0.35 | -0.48 | 0.18 | $6.8 \times 10^{-3}$                     | -0.69 | 0.10 | <b><math>4.2 \times 10^{-12}</math></b>  |
| Periderm colour                 | S3_4545411   | G            | C            | 0.43 | -0.38 | 0.02 | <b><math>1.4 \times 10^{-123}</math></b> | -0.42 | 0.01 | <b><math>7.3 \times 10^{-288}</math></b> |
| Cortex colour                   | S1_3047840   | T            | G            | 0.01 | 1.08  | 0.05 | <b><math>1.6 \times 10^{-92}</math></b>  | 1.18  | 0.05 | <b><math>1.8 \times 10^{-135}</math></b> |
| Cortex colour                   | S2_6566608   | C            | T            | 0.01 | 0.81  | 0.04 | <b><math>5.0 \times 10^{-83}</math></b>  | 0.89  | 0.04 | <b><math>2.0 \times 10^{-113}</math></b> |

MAF frequency of the reference allele, *B*SNP effect, *SE* standard error of SNP effect, *P* marker-trait association p value

Bold fonts represent markers that are significant at the Bonferroni threshold of  $0.05/101,521 = 4.93 \times 10^{-7}$

\*SNP effect, standard error and p values obtained from GCTA MLME model

**Fig. 5** Multiple marker-trait regression barplot across traits and the proportion of phenotypic variance explained. Numbers above barplot denote the number of loci in the regression model



and 16 are Manes.06G103600 (Bidirectional sugar transporter Sweet4-Related); Manes.15G011300 (Sweet17 homologue, which mediates fructose transport across the tonoplast of roots) and Manes.16G109200 (Hexokinase), respectively.

Our search for candidate genes related to CMD severity uncovered two peroxidase genes: Manes.12G076200 and Manes.12G076300 occurring less than 45 Kbp away from marker S12\_7926132 in chromosome 12. These two candidate genes were previously reported by Wolfe et al. (2016) and Rabbi et al. (2014). Peroxidases have been reported to play a role in activating plant defence systems upon pathogen infections (Ye et al. 1990; Wu et al. 1997; Gonçalves et al. 2013).

Our analyses further detected three loci that colocalised with three candidate genes associated with CGM severity. Specifically, SNP marker S8\_6409580 fell in the coding region of Manes.08G058000 gene, a homologue of AtMYB16, encoding a MIXTA-like MYB gene which regulates cuticle development and trichome branching (Oshima et al. 2013). Structural traits including trichomes and waxy cuticles are known to act as a physical barrier to arthropod pest attachment, feeding and oviposition (Mitchell et al. 2016).

Harvest index GWAS analysis highlighted two regions on chromosome 2 and 12, respectively. On chromosome 2, the candidate SNP (S2\_2809137) is located 24 Kbp away Manes.02G035900 a homologue of BFRUCT4 (vacuolar invertase) a key enzyme in sucrose hydrolysis and involved in the export of reduced carbon sink organs such as roots (Haouazine-Takvorian et al. 1997; Nägele et al. 2010). For leaf shape, a candidate gene Manes.15G136200, homologous to KNOX1 that is implicated in the expression of

diverse leaf shapes in plants was found around 186 Kb away from the major locus on chromosome 15 tagged by SNP S15\_10273255 (Furumizu et al. 2015). Candidate gene search around the major locus for petiole colour and leaf greenness tagged by SNP S1\_23452638 revealed the presence of a Myb transcription factor homologue Manes.01G115400, which occurred 30.7 Kb away from the top marker. Myb genes are known to play a crucial role in regulating pigment biosynthesis pathway in plants (Nesi et al. 2001; Kobayashi et al. 2002; Himi and Noda 2005; Allan et al. 2008; Furumizu et al. 2015).

## Discussion

Understanding the genetic architecture of key breeding-goal traits is a critical step towards more efficient and accelerated genetic improvement. This study builds on and extends previous cassava GWAS efforts by analysing a large breeding population phenotyped extensively over successive years and stages of selection in multi-environment field trials. The population showed large phenotypic variation among clones with respect to all traits (Supplementary Figure S1). Furthermore, the population was derived from two successive cycles of recurrent selection using elite parents with good breeding values for yield, dry matter content and resistance against CMD (Wolfe et al. 2016). The collection is expected to be enriched for favourable alleles from major- and minor-effect loci underlying these traits and thus well suited to conduct a marker-trait association study efficiently.

The observed and consistent trend in the magnitude of both broad-sense and SNP-based heritability estimates

**Table 4** Summary information of the potential candidate genes identified in the vicinity of the GWAS hits for analysed traits

| Trait                             | SNP          | Chromosome | Position (bp) | Candidate gene  | Protein name                                   | Upstream (kb) | Downstream (kb) | Homologues                                                  |
|-----------------------------------|--------------|------------|---------------|-----------------|------------------------------------------------|---------------|-----------------|-------------------------------------------------------------|
| CMD severity                      | S12_7926132  | S12        | 7,926,132     | Manes.12G076200 | Peroxidase                                     | 26.90         | 28.4            | PER27 P27, At3g01190, T4P13.12                              |
| CMD severity                      | S12_7926132  | S12        | 7,926,132     | Manes.12G076300 | Peroxidase                                     | 44.04         | 45.4            | PRX74 Os05g0499300                                          |
| CMD severity                      | S14_4626854  | S14        | 4,626,854     | Manes.14G058400 | TCP family transcription factor                | 6.37          | 5.3             | D0Y65_016060                                                |
| CGM severity                      | S8_6409580   | S8         | 6,409,580     | Manes.08G058000 | myb proto-oncogene protein                     | 1.33          | 0.2             | MYB16                                                       |
| CGM severity                      | S17_23749968 | S17        | 23,749,968    | Manes.17G094500 | Myb-like DNA-binding domain                    | 39.59         | 36.57           | AtMYB011, AtMYB017 and AtMYB023 (Arabidopsis)               |
| Total carotenoid                  | S1_24159583  | S1         | 24,159,583    | Manes.01G124200 | Phytoene synthase                              | 6.16          | 2.9             | PSY, PSY3 (Maize), PSY1 (Tomato), PSY7B (PSY7B)             |
| Total carotenoid                  | S5_3387558   | S5         | 3,387,558     | Manes.05G051700 | Beta-carotene Dioxygenase                      | 481.47        | 484.7           | CCD7 (Arabidopsis), MAX3 (Arabidopsis), NCED7 (Arabidopsis) |
| Total carotenoid                  | S15_7659426  | S15        | 7,659,426     | Manes.15G102000 | Beta-carotene Dioxygenase                      | 82.73         | 80.9            | NCED2 (Arabidopsis), NCED4 (maize), NCED1 (tomato)          |
| Dry matter content                | S1_24636113  | S1         | 24,636,113    | Manes.01G123000 | UTP-glucose pyrophosphorylase                  | 577.55        | 571.4           | UGP1 At5g17310                                              |
| Dry matter content                | S1_24636113  | S1         | 24,636,113    | Manes.01G123800 | Sucrose synthase                               | 496.09        | 491.5           | SUS5 At5g37180 (Arabidopsis)                                |
| Dry matter content                | S6_20589894  | S6         | 20,589,894    | Manes.06G103600 | Bidirectional sugar transporter sweet4-related | 794.31        | 795.5           | 25,495,183 MTR_6g007623, LOC107801067, LOC103483429         |
| Dry matter content                | S15_1012346  | S15        | 1,012,346     | Manes.15G011300 | RAG1 activating protein                        | 23.39         | 16.8            | ES319_A13G194900v1                                          |
| Dry matter content                | S16_25663808 | S16        | 25,663,808    | Manes.16G109200 | Helicase                                       | 797.11        | 803.7           | At3g20040, MAL21.4 (Arabidopsis)                            |
| Leaf shape                        | S15_10273255 | S15        | 10,273,255    | Manes.15G136200 | KNOX1 and KNOX2 domain                         | 186.22        | 187.3           | AT1G14760.3—KNOX, AT1G14760.2—KNOX                          |
| Petiole colour and Leaf greenness | S1_23452638  | S1         | 23,452,638    | Manes.01G115400 | Myb proto-oncogene protein                     | 30.76         | 32.1            | AtMYB075, AtMYB090, AtMYB113 and AtMYB114                   |

**Table 4** (continued)

| Trait         | SNP        | Chromosome | Position (bp) | Candidate gene  | Protein name            | Upstream (kb) | Downstream (kb) | Homologues                                                    |
|---------------|------------|------------|---------------|-----------------|-------------------------|---------------|-----------------|---------------------------------------------------------------|
| Harvest index | S2_2809137 | S2         | 2,809,137     | Manes.02G035900 | Beta-fructofuranosidase | 26.69         | 24.2            | BFRUCT3<br>BETAFRUCT3,<br>At1g62660,<br>F23N19.3,<br>T3P18.21 |
| Harvest index | S2_2809137 | S2         | 2,809,137     | Manes.02G037700 | Glycogen phosphorylase  | 115.16        | 116.9           | Alpha-1,4 glucan phosphorylase (citrus hybrid cultivar)       |

further emphasises the significant contribution of additive genetic factors in the expression of some of these traits. The heritability estimates recorded in our study also indicate a good repeatability and reproducibility of the experimental procedures. The heritability estimates we found are comparable to those previously reported in other studies for these traits (Oliveira et al. 2014, 2015; Njoku et al. 2015; Silva et al. 2016; Favour et al. 2017; Rao et al. 2018).

Accounting for population structure and genetic relatedness in GWAS is necessary to reduce false positives. PCA did not reveal the presence of substantial population stratification in our GWAS panel. This is not surprising since extensive inter-crosses are routinely carried out as part of the generation of new genetic variation in the IITA Cassava Breeding program. Moreover, individuals from C1 and C2 largely overlapped with each other and also with the founder population (GG). GWAS analysis using PCA and the kinship matrix gave very similar results to the analysis that considered kinship alone in controlling for spurious associations. For this reason, as well as for computational efficiency, we used the MLM model for the full analysis.

Our data replicated the previously identified associations for CMD and CGM resistance traits, dry matter and carotenoid content. Also, they showed more robust evidence of a major gene effect in this larger sample. In addition to the confirmed loci, we uncovered additional genomic regions with significant associations that were previously not reported. For example, two additional regions on chromosome 14 were found to contribute to increased resistance to CMD virus and present useful targets for further genetic analysis. Genotypes carrying favourable resistance alleles at these as well as the CMD2 major effect locus on chromosome 12 exhibit high levels of resistance against CMD and rarely show symptoms. However, despite the concerted global efforts to identify the causal gene underlying the CMD2 locus in cassava, there has not been any breakthrough in cloning and functionally identifying the actual gene. We hope that these resources will guide the cassava community

in narrowing down on the candidate genes to carry out the functional analysis.

Although our primary interest was in traits that are under active breeding selection, we also measured several morphological traits that we knew to be heritable. Among the morphological traits presented here, to our best knowledge, only leaf shape and apical leaf pubescence were previously reported (Ezenwaka et al. 2018; Zhang et al. 2018). While our study confirmed major locus for apical pubescence locus on chromosome 8, we did not replicate the results of Zhang et al. (2018) for leaf aspect ratio suggesting a possibility of different genetic factors.

Many traits of interest to crop improvement are positively or negatively correlated. Such correlations can cause unfavourable changes in traits that are important but that are not under direct selection. Alternatively, one trait can be used to indirectly select for another positively correlated trait which is more difficult to phenotype. Genetic correlations among traits can arise due to linkage disequilibrium or pleiotropy (a single gene having multiple otherwise unrelated biological effects, or shared regulation of multiple genes) (Chen and Lübberstedt 2010). Correlations due to linkage disequilibrium tend to be temporary and are generally considered to be less critical than pleiotropy. Of particular importance is the negative correlation observed between total carotenoid content variation and dry matter content which confirms previous findings (Ceballos et al. 2013; Njoku et al. 2015; Esuma et al. 2016; Ceballos et al. 2017; Rabbi et al. 2017; Okeke et al. 2018). Both genetic linkage and pleiotropy are plausible reasons for the inverse relationship. The linkage hypothesis is supported by the co-location of major QTLs for these traits on chromosome 1 and the presence of major genes in the biochemical pathways in close proximity.

On the other hand, genetically engineered cassava to produce and accumulate carotenoids in storage roots was found in one study to have reduced dry matter content which indicates a pleiotropic effect (Beyene et al. 2017). Besides the region on chromosome 1, we identified additional loci on chromosomes 5, 8, 15, and 16 that additively explain more

than 60% of the phenotypic variation for total carotenoids. These and other minor effect loci could explain the lack of inverse relationship between carotenoids and dry matter content in other populations (Ceballos et al. 2013). Likewise, we also found several significant regions associated with dry matter content, although they only explain 16% of the trait variation. Other notable negative but the favourable correlation in the study population was between apical pubescence and CGM severity. Generally, genotypes with glabrous apical leaves are more affected by the pest than pubescent ones (Raji et al. 2008; Chalwe et al. 2015; Ezenwaka et al. 2018). For both traits, we identified a common major locus on chromosome 8 that is associated with variation in the degree of pubescence as well as CGM damage severity. We also found other loci that were unique to either trait suggesting additional factors may be contributing to the resistance against CGM besides the presence of trichomes. Significant genetic and phenotypic correlations were not observed between harvest index, dry matter content and plant type, implying that they are amenable to concurrent improvement.

### Opportunities and implications

The present study is based predominantly on germplasm developed by the IITA breeding program and therefore represents a subset of the available diversity worldwide. While some traits such as yield and yield components, are universally considered, other traits, especially biotic and abiotic stresses, are region-specific. Moreover, restricted germplasm movement due to quarantine regulations makes it nearly impossible to evaluate the same population in multiple regions simultaneously. Further studies using germplasm from other regions, including east Africa and Latin America—the centre of origin and diversity of cassava—are expected to reveal region-specific/population-specific large effect alleles. Such efforts are expected to enrich the catalogue of major effect loci available for molecular breeding.

A major objective of this study was to provide breeders with a catalogue of major loci for marker-assisted selection. Many of the previous QTL studies using bi-parental mapping populations in cassava have had limited value due to low marker densities and poor genetic resolution (Ferguson et al. 2012; Ceballos et al. 2015; Hershey 2017). Access to high-density genome-wide SNP markers through GBS coupled with GWAS mapping approach has resulted in higher mapping resolution, close to known candidate genes for several traits in the present study. If converted to allele-specific high-throughput SNP assays, SNPs tagging major loci from the present study can be used to screen and identify individuals carrying favourable alleles during early stages of selection. However, further validation of these loci is required to ensure they are useful across environments (genotype-by-environment interaction) and populations (different genetic

background) before large-scale deployment. For other traits such as harvest index, the identified loci only contributed to a small proportion of trait variation suggesting additional genes with minor effects. Thus, such traits are more likely to benefit from genomic selection (GS) to estimate breeding values from genome-wide marker data. GS relies on high-density marker coverage such that each QTL is expected to be in linkage disequilibrium (LD) with at least one marker (Goddard and Hayes 2007). A tandem approach incorporating MAS and GS is likely to increase breeding efficiency, reduce breeding cycle and cost (Zhao et al. 2014). MAS can be used to screen a large number of individuals at seedling nursery stages to cull accessions that do not carry favourable alleles. Reduced number of lines can then be genotyped at higher density for GS and allocated to field testing plots for evaluation and selection for complex traits.

### Conclusions

In this study, we demonstrate that the use of a diverse association mapping panel consisting of landraces and improved cultivars from multiple breeding initiatives can identify SNP variants associated with agronomically essential traits in cassava. The power of this study to discover additional markers associated with measured traits was derived from extensive multi-locational testing over 4 years at four IITA testing locations in Nigeria. The SNP markers we identified provide a useful reference catalogue not only for cassava breeding programs but also studies aimed at uncovering unique biological pathways necessary for advancing genetic transformation studies. To realise the full potential of these marker-trait associations in population improvement, all top markers presented in Table 3 have been converted to allele-specific PCR assays. They have been made freely available through several commercial genotyping service vendors, including LGC Genomics (UK) and Intertek (Sweden, Australia). We hope that these markers will be of benefit to the cassava research and breeding community and will be included in the breeders' toolbox.

**Acknowledgements** This work was funded by The Bill & Melinda Gates Foundation and the Department for International Development of the United Kingdom (award number OPP1048542) through the “Next Generation Cassava Breeding Project”; and the CGIAR-Research Program on Roots, Tubers and Bananas. Thanks to the technical staff at IITA Cassava Breeding Program for implementing the field experiments. We also thank Excellence-in-Breeding Platform of the CGIAR for supporting the conversion of the trait-linked markers to allele-specific PCR assays.

**Author contribution** IYR, CE, JLJ, and PK conceived and designed the study; IYR, SIK, GB, AA, and MY performed analyses and wrote the manuscript; CE, EL, EP, MW, JLJ, and PK edited the manuscript; CA,

KO, RU, ASI, and PP Implemented field trials, generated and curated data; and PK Provided overall coordination and leadership.

**Data availability** The raw phenotypic data that supports this research is openly accessible on the cassava breeding database ([www.cassavabas.org](http://www.cassavabas.org)) and can be downloaded using the trial names in Supplementary Table S1. Clone names can be linked to clone GBS ID as in Supplementary Table 2. The GBS data can be downloaded from [ftp://ftp.cassavabase.org/manuscripts/Rabbi\\_et\\_al\\_2020/](ftp://ftp.cassavabase.org/manuscripts/Rabbi_et_al_2020/).

## Compliance with ethical standards

**Conflict of interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abass AB, Towo E, Mukuka I, Okechukwu R, Ranavivon R, Tarawali G, Kanju E (2014) Growing cassava: a training manual from production to postharvest
- Akano A, Dixon A, Mba C, Barrera E, Fregene M (2002) Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. *Theor Appl Genet* 105:521–525. <https://doi.org/10.1007/s00122-002-0891-7>
- Alabi OJ, Mulenga RM, Legg JP (2015) Cassava mosaic: virus diseases of tropical and subtropical crops. In: Tennant P, Fermin G (eds) *Virus diseases of tropical and subtropical crops*. CAB International, Wallingford, pp 42–55
- Allan AC, Hellens RP, Laing WA (2008) MYB transcription factors that colour our fruit. *Trends Plant Sci* 13:99–102
- Atser G, Dixon A, Ekeleme F, Chikoye D, Dashiell KE, Ayankanmi T, Hauser S, Agada M, Okwusi M, Sokoya G (2017) The ABC of weed management in cassava production in Nigeria: a training manual
- Bates D, Maechler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *J Stat Softw* 67:1–48. <https://doi.org/10.18637/jss.v067.i01>
- Batra S, Carlson J, Hayes R, Shu S, Schmutz J, Rokhsar D (2014) Phytozome Comparative Plant Genomics Portal, pp 1–2
- Bechoff A, Tomlins K, Westby A, Fliedel G, Hershey C, Dufour D, Becerra Lopez-lavalle LA (2018) Cassava traits and end-user preference: relating traits to consumer liking, sensory perception, and genetics. *Crit Rev Food Sci Nutr* 58(4):547–567
- Beyene G, Chauhan RD, Ilyas M, Wagaba H, Fauquet CM, Miano D, Alicai T, Taylor NJ (2017) A virus-derived stacked RNAi construct confers robust resistance to cassava brown streak disease. *Front Plant Sci* 7:1000. <https://doi.org/10.3389/fpls.2016.02052>
- Bredeson JV, Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, Grimwood J, Schmutz J, Rabbi IY, Egesi C, Nauluvula P, Lebot V, Ndunguru J, Mkamilo G, Bart RS, Setter TL, Gleadow RM, Kulakow P, Ferguson ME, Rounsley S, Rokhsar DS (2016) Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat Biotechnol* 34:562–570. <https://doi.org/10.1038/nbt.3535>
- Browning BL, Browning SR (2016) Genotype imputation with millions of reference samples. *Am J Hum Genet* 98:116–126. <https://doi.org/10.1016/j.ajhg.2015.11.020>
- Burns A, Gleadow R, Cliff J, Zacarias A, Cavagnaro T (2010) Cassava: the drought, war and famine crop in a changing world. *Sustainability* 2:3572–3607. <https://doi.org/10.3390/su2113572>
- Byrne DH, Guerrero JM, Bellotti AC, Gracen VE (1982) Behavior and development of *Mononychellus tanajoa* (Acari: Tetranychidae) on resistant and susceptible cultivars of Cassava 1. *J Econ Entomol* 75:924–927. <https://doi.org/10.1093/jee/75.5.924>
- CABI (2019) Invasive Species Compendium. <https://www.cabi.org/isc>. Accessed 11 Dec 2019
- Ceballos H, Kulakow P, Hershey C (2012) Cassava breeding: current status, bottlenecks and the potential of biotechnology tools. *Trop Plant Biol* 5:73–87. <https://doi.org/10.1007/s12042-012-9094-9>
- Ceballos H, Morante N, Sánchez T, Ortiz D, Aragón I, Chávez AL, Pizarro M, Calle F, Dufour D (2013) Rapid cycling recurrent selection for increased carotenoids content in cassava roots. *Crop Sci*. <https://doi.org/10.2135/cropsci2013.02.0123>
- Ceballos H, Kawuki RS, Gracen VE, Yencho GC, Hershey CH (2015) Conventional breeding, marker-assisted selection, genomic selection and introbreeding in clonally propagated crops: a case study for cassava. *Theor Appl Genet*. <https://doi.org/10.1007/s00122-015-2555-4>
- Ceballos H, Davrieux F, Talsma EF, Belalcazar J, Chavarriaga P, Andersson MS (2017) Carotenoids in cassava roots. In: *Carotenoids*. InTech, Rijeka
- Chalwe A, Melis R, Shanahan P, Chiona M (2015) Inheritance of resistance to cassava green mite and other useful agronomic traits in cassava grown in Zambia. *Euphytica* 205:103–119. <https://doi.org/10.1007/s10681-015-1404-5>
- Chen Y, Lübberstedt T (2010) Molecular basis of trait correlations. *Trends Plant Sci* 15:454–461
- Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celedón JC, Redline S, Papanicolaou GJ, Thornton TA, Laurie CC, Rice K, Lin X (2016) Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am J Hum Genet* 98:653–666. <https://doi.org/10.1016/j.ajhg.2016.02.012>
- Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA miniprep: version II. *Plant Mol Biol Rep* 1:19–21. <https://doi.org/10.1007/BF02712670>
- Dixon AGO, Ssemakula G (2008) Prospects for cassava breeding in Sub-Saharan Africa in the next decade. *J Food Agric Environ* 6:256–262
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:1–10. <https://doi.org/10.1371/journal.pone.0019379>
- Esuma W, Herselman L, Labuschagne MT, Ramu P, Lu F, Baguma Y, Buckler ES, Kawuki RS (2016) Genome-wide association mapping of provitamin A carotenoid content in cassava. *Euphytica* 212:97–110. <https://doi.org/10.1007/s10681-016-1772-5>
- Ezenwaka L, Del Carpio Dunia P, Jannink J-L, Rabbi I, Danquah E, Asante I, Danquah A, Blay E, Egesi C (2018) Genome-wide association study of resistance to cassava green mite pest and related traits in cassava. *Crop Sci* 58:1907–1918
- Favour E, Emeka N, Chiedozi E, Bunmi O, Emmanuel O (2017) Genetic variability, heritability and variance components of some yield and yield related traits in second backcross population (BC2) of cassava. *Afr J Plant Sci* 11:185–189. <https://doi.org/10.5897/AJPS2015.1324>

- Ferguson M, Rabbi I, Kim DJ, Gedil M, Lopez-Lavalle LAB, Okogbenin E (2012) Molecular markers and their application to cassava breeding: past, present and future. *Trop Plant Biol* 5:95–109. <https://doi.org/10.1007/s12042-011-9087-0>
- Fukuda WMG, Guevara CL, Kawuki R, Ferguson ME (2010) Selected morphological and agronomic descriptors for the characterization of cassava. International Institute of Tropical Agriculture (IITA), Ibadan
- Furumizu C, Alvarez JP, Sakakibara K, Bowman JL (2015) Antagonistic roles for KNOX1 and KNOX2 genes in patterning the land plant body plan following an ancient gene duplication. *PLOS Genet* 11:e1004980. <https://doi.org/10.1371/journal.pgen.1004980>
- García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassel CP (2016) Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proc Natl Acad Sci USA* 113:E3995. <https://doi.org/10.1073/pnas.1519061113>
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9:e90346–e90346. <https://doi.org/10.1371/journal.pone.0090346>
- Goddard ME, Hayes BJ (2007) Genomic selection. *J Anim Breed Genet* 124:323–330. <https://doi.org/10.1111/j.1439-0388.2007.00702.x>
- Gonçalves L, Rodrigues R, Diz M, Robaina R, Júnior A, Carvalho A, Gomes V (2013) Peroxidase is involved in pepper yellow mosaic virus resistance in *Capsicum baccatum* var. *pendulum*. *Genet Mol Res* 12:1411–1420
- Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N (2011) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40:D1178–D1186
- Gould K, Davies KM, Winefield C (2008) Anthocyanins: biosynthesis, functions, and applications. Springer, Heidelberg
- Hamblin MT, Rabbi IY (2014) The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: a study in cassava. *Crop Sci* 54(6):2603–2608. <https://doi.org/10.2135/cropsci2014.02.0160>
- Haouzine-Takvorian N, Tymowska-Lalanne Z, Takvorian A, Tregear J, Lejeune B, Lecharny A, Kreis M (1997) Characterization of two members of the *Arabidopsis thaliana* gene family, Atβfruct3 and Atβfruct4, coding for vacuolar invertases. *Gene* 197:239–251
- Hershey CH (2017) Marker-assisted selection in cassava breeding: Ismail Y. Rabbi, International Institute of Tropical Agriculture (IITA), Nigeria. In: Achieving sustainable cultivation of cassava, vol 2. Burleigh Dodds Science Publishing, Cambridge, pp 121–127
- Himi E, Noda K (2005) Red grain colour gene (R) of wheat is a Myb-type transcription factor. *Euphytica* 143:239–242
- Ikeogu UN, Akdemir D, Wolfe MD, Okeke UG, Amaefula C, Jannink J-L, Egesi CN (2019) Genetic correlation, genome-wide association and genomic prediction of portable NIRS predicted carotenoids in cassava roots. *Front Plant Sci* 10:1570–1570
- International Cassava Genetic Map Consortium (2015) High-resolution linkage map and chromosome-scale genome assembly for cassava (*Manihot esculenta* and *Crantz*) from 10 populations. *G3 Genes Genomes Genet* 5:133–144
- Jarvis A, Ramirez-Villegas J, Herrera Campo BV, Navarro-Racines C (2012) Is cassava the answer to African climate change adaptation? *Trop Plant Biol* 5:9–29. <https://doi.org/10.1007/s12042-012-9096-7>
- Kawano K (2003) Thirty years of cassava breeding for productivity—biological and social factors for success. *Crop Sci* 43:1325–1335
- Kawano K, Daza P, Amaya A, Rios M, Goncalves WMF (1978) Evaluation of cassava germplasm for productivity. *Crop Sci* 18:377–380
- Kawano K, Maria W, Fuluda G, Cenpuldee U, Maria W, Fuknda G, Cenpnkdee U, Fnknda G, Cenpukdee U (1987) Genetic and environmental effects on dry matter content of cassava root. *Crop Sci* 27:69–74. <https://doi.org/10.2135/cropsci1987.0011183X002700010018x>
- Kawano K, Narintaraporn K, Narintaraporn P, Sarakarn S, Limsila A, Limsila J, Suparhan D, Sarawat V, Watananonta W (1998) Yield improvement in a multistage breeding program for cassava. *Crop Sci* 38:325–332. <https://doi.org/10.2135/cropsci1998.0011183X003800020007x>
- Kayondo SI, Pino Del Carpio D, Lozano R, Ozimati A, Wolfe M, Baguma Y, Gracen V, Offei S, Ferguson M, Kawuki R, Jannink J-L (2018) Genome-wide association mapping and genomic prediction for CBD resistance in *Manihot esculenta*. *Sci Rep* 8:1549–1549. <https://doi.org/10.1038/s41598-018-19696-1>
- Kobayashi S, Ishimaru M, Hiraoka K, Honda C (2002) Myb-related genes of the Kyoho grape (*Vitis labruscana*) regulate anthocyanin biosynthesis. *Planta* 215:924–933
- Legg J, Owor B, Sseruwagi P, Ndunguru J (2006) Cassava mosaic virus disease in East and Central Africa: epidemiology and management of a regional pandemic. *Adv Virus Res* 67:355–418
- LiLin-Yin (2019) CMplot: Circle Manhattan Plot. <https://github.com/YinLiLin/CMplot>
- Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D (2012) Improved linear mixed models for genome-wide association studies. *Nat Methods* 9:525
- Ly D, Hamblin M, Rabbi I, Melaku G, Bakare M, Gauch HG, Okechukwu R, Dixon AGO, Kulakow P, Jannink JL (2013) Relatedness and genotype × environment interaction affect prediction accuracies in genomic selection: a study in cassava. *Crop Sci* 53:1312–1325. <https://doi.org/10.2135/cropsci2012.11.0653>
- Mitchell C, Brennan RM, Graham J, Karley AJ (2016) Plant defense against herbivorous pests: exploiting resistance and tolerance traits for sustainable crop protection. *Front Plant Sci* 7:1132. <https://doi.org/10.3389/fpls.2016.01132>
- Nägele T, Henkel S, Hörmiller I, Sauter T, Sawodny O, Ederer M, Heyer AG (2010) Mathematical modeling of the central carbohydrate metabolism in *Arabidopsis* reveals a substantial regulatory influence of vacuolar invertase on whole plant carbon metabolism. *Plant Physiol* 153:260–272. <https://doi.org/10.1104/pp.110.154443>
- Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L (2001) The *Arabidopsis* TT2 gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *Plant Cell* 13:2099–2114
- Njoku DN, Gracen VE, Offei SK, Asante IK, Egesi CN, Kulakow P, Ceballos H (2015) Parent-offspring regression analysis for total carotenoids and some agronomic traits in cassava. *Euphytica* 206:657–666. <https://doi.org/10.1007/s10681-015-1482-4>
- Okechukwu R, Dixon AG (2008) Genetic gains from 30 years of cassava breeding in Nigeria for storage root yield and disease resistance in elite cassava genotypes. *J Crop Improv* 22:181–208
- Okeke UG, Akdemir D, Rabbi I, Kulakow P, Jannink JL (2018) Regional heritability mapping provides insights into dry matter content in African white and yellow cassava populations. *Plant Genome*. <https://doi.org/10.3835/plantgenome2017.06.0050>
- Oliveira EJ, Aidar SD, Morgante CV, Chaves AR, Cruz JL, Coelho Filho MA (2015) Genetic parameters for drought-tolerance in cassava. *Pesqui Agropecuária Bras* 50:233–241
- Oliveira EJ, Santana FA, Oliveira LA, Santos VS (2014) Genetic parameters and prediction of genotypic values for root quality traits in cassava using REML/BLUP. *Genet Mol Res* 13:6683–6700. <https://doi.org/10.4238/2014.August.28.13>
- Oshima Y, Shikata M, Koyama T, Ohtsubo N, Mitsuda N, Ohme-Takagi M (2013) MIXTA-like transcription factors and WAX

- INDUCER1/SHINE1 coordinately regulate cuticle development in *Arabidopsis* and *Torenia fournieri*. *Plant Cell* 25:1609–1624
- Owor B, Legg J, Okao-Okuja G, Obonyo R, Ogenga-Latigo M (2004) The effect of cassava mosaic geminiviruses on symptom severity, growth and root yield of a cassava mosaic virus disease-susceptible cultivar in Uganda. *Ann Appl Biol* 145:331–337
- Palta JP (1990) Leaf chlorophyll content. *Remote Sens Rev* 5:207–213. <https://doi.org/10.1080/02757259009532129>
- 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* 9:907–927
- Prudencio YC, Al-Hassan R (1994) The food security stabilization roles of cassava in Africa. *Food Policy* 19:57–64. [https://doi.org/10.1016/0306-9192\(94\)90008-6](https://doi.org/10.1016/0306-9192(94)90008-6)
- Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Glied TP, Boehnke M, Abecasis GR, Willer CJ (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26:2336–2337
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575. <https://doi.org/10.1086/519795>
- R Development Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, R Foundation for Statistical Computing, Vienna
- Rabbi IY, Hamblin MT, Kumar PL, Gedil MA, Ikpan AS, Jannink JL, Kulakow PA (2014) High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding. *Virus Res* 186:87–96. <https://doi.org/10.1016/j.virusres.2013.12.028>
- Rabbi IY, Udoh LI, Wolfe M, Parkes EY, Gedil MA, Dixon A, Ramu P, Jannink J-L, Kulakow P (2017) Genome-wide association mapping of correlated traits in cassava: dry matter and total carotenoid content. *Plant Genome* 0:0–0. <https://doi.org/10.3835/plantgenome2016.09.0094>
- Raji A, Ladeinde O, Dixon A (2008) Screening landraces for additional sources of field resistance to cassava mosaic disease and green mite for integration into the cassava improvement program. *J Integr Plant Biol* 50:311–318
- Ramu P, Esuma W, Kawuki R, Rabbi IY, Egesi C, Bredeson JV, Bart RS, Verma J, Buckler ES, Lu F (2017) Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat Genet* 49:959–963. <https://doi.org/10.1038/ng.3845>
- Rao BB, Swami D, Ashok P, Babu BK, Ramajayam D, Sasikala K (2018) Estimation of genetic variability and heritability for yield and its related components in cassava (*Manihot esculenta* Crantz) genotypes. *Int J Curr Microbiol Appl Sci* 7:287–297
- Rentería ME, Cortes A, Medland SE (2013) Using PLINK for genome-wide association studies (GWAS) and data analysis. In: Gondro C, van der Werf J, Hayes B (eds) *Methods Mol Biol*. Humana Press, Totowa, pp 193–213
- Saltzman A, Birol E, Bouis HE, Boy E, De Moura FF, Islam Y, Pfeiffer WH (2013) Biofortification: progress toward a more nourishing future. *Glob Food Security* 2:9–17. <https://doi.org/10.1016/j.gfs.2012.12.003>
- Sánchez T, Ceballos H, Dufour D, Ortiz D, Morante N, Calle F, Zum Felde T, Dominguez M, Davrieux F (2014) Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and Hunter color techniques. *Food Chem* 151:444–451
- Shukla P (1976) Preliminary report on the green mite (*Mononychellus tanajoa*, Bonder) resistance in Tanzanian local cassava varieties. *East Afr Agric For J* 42:55–59
- Silva RS, Moura EF, Neto JTF, Sampaio JE (2016) Genetic parameters and agronomic evaluation of cassava genotypes. *Pesqui Agropecu Bras* 51:834–841. <https://doi.org/10.1590/S0100-204X2016000700006>
- Sinclair TR (1998) Historical changes in harvest index and crop nitrogen accumulation. *Crop Sci* 38:638–643
- Speed D, Hemani G, Johnson MR, Balding DJ (2012) Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* 91:1011–1021
- Takenaka A (1994) Effects of leaf blade narrowness and petiole length on the light capture efficiency of a shoot. *Ecol Res* 9:109–114
- Thottappilly G, Thresh JM, Calvert LA, Winter S (2003) Cassava. In: Loebenstein G, Thottappilly G (eds) *Virus and virus-like diseases of major crops in developing countries*. Springer, Dordrecht, pp 107–165
- Uke A, Hoat TX, Quan MV, Liem NV, Ugaki M, Natsuaki KT (2018) First report of Sri Lankan cassava mosaic virus infecting cassava in Vietnam. *Plant Dis* 102:2669. <https://doi.org/10.1094/PDIS-05-18-0805-PDN>
- Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PLOS Biol* 12:e1001883. <https://doi.org/10.1371/journal.pbio.1001883>
- Wei T, Simko V (2017) R package “corrplot”: visualization of a correlation matrix (Version 0.84). <https://github.com/taiyun/corrplot/>
- Welsch R, Arango J, Bär C, Salazar B, Al-Babili S, Beltrán J, Chavarriaga P, Ceballos H, Tohme J, Beyer P (2010) Provitamin A accumulation in cassava (*Manihot esculenta*) roots driven by a single nucleotide polymorphism in a phytoene synthase gene. *Plant Cell* 22:3348. <https://doi.org/10.1105/tpc.110.077560>
- Wolfe MD, Rabbi IY, Egesi C, Hamblin M, Kawuki R, Kulakow P, Lozano R, Carpio DPD, Ramu P, Jannink J-L (2016) Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. *Plant Genome* 9:0–0. <https://doi.org/10.3835/plantgenome2015.11.0118>
- Wolfe MD, Del Carpio DP, Alabi O, Ezenwaka LC, Ikeogu UN, Kayondo IS, Lozano R, Okeke UG, Ozimati AA, Williams E, Egesi C, Kawuki RS, Kulakow P, Rabbi IY, Jannink J-L (2017) Prospects for genomic selection in cassava breeding. *Plant Genome*. <https://doi.org/10.3835/plantgenome2017.03.0015>
- Wu G, Shortt BJ, Lawrence EB, Leon J, Fitzsimmons KC, Levine EB, Raskin I, Shah DM (1997) Activation of host defense mechanisms by elevated production of H<sub>2</sub>O<sub>2</sub> in transgenic plants. *Plant Physiol* 115:427–435
- Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88:76–82. <https://doi.org/10.1016/j.ajhg.2010.11.011>
- Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* 46:100–106. <https://doi.org/10.1038/ng.2876>
- Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM (2017) Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet* 49:1304
- Ye X, Pan S, Kuc J (1990) Activity, isozyme pattern, and cellular localization of peroxidase as related to systemic resistance of tobacco to blue mold (*Peronospora tabacina*) and to tobacco mosaic virus. *Phytopathology* 80:1295–1299
- 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). *Front Plant Sci* 9:503–503
- Zhao Y, Mette MF, Gowda M, Longin CFH, Reif JC (2014) Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. *Heredity* 112:638–645. <https://doi.org/10.1038/hdy.2014.1>