Mol Breeding (2017) 37:123
DOI 10.1007/s11032-017-0718-5

Candidate gene sequencing and validation of SNP markers linked to carotenoid content in cassava (*Manihot esculenta* Crantz)

Lovina I. Udoh · Melaku Gedil · Elizabeth Y. Parkes · Peter Kulakow · Adenubi Adesoye · Chiamaka Nwuba · Ismail Y. Rabbi

Received: 6 March 2017 / Accepted: 22 August 2017

Abstract Cassava is a widely grown staple in Sub-Saharan Africa and consumed as a cheap source of calories, but the crop is deficient in micronutrients including pro-vitamin A carotenoids. This challenge is currently being addressed through biofortification breeding that relies on phenotypic selection. Gene-based markers linked to pro-vitamin A content variation are expected to increase the rate of genetic gain for this critical trait. We sequenced four candidate carotenoid genes from 167 cassava accessions representing the diversity of elite breeder lines from IITA. Total carotenoid content was determined using spectrophotometer and total β-carotene was quantified by high-performance liquid chromatography. Storage root yellowness due to carotenoid pigmentation was assessed. We carried out candidate gene association analysis that accounts for population structure and kinship using genome-wide single nucleotide polymorphisms (SNPs) generated through genotyping-by-sequencing. Significant SNPs were used to design competitive allele-specific PCR assays and validated on the larger population for potential use in marker-assisted selection breeding. Candidate gene sequencing of the genes β-carotene hydroxylase (*crtRB*), phytoene synthase (*PSY2*), lycopene epsilon cyclase (*lcyE*), and lycopene beta cyclase (*lcyB*) yielded a total of 37 SNPs. Total carotenoid content, total β-carotene, and color parameters were significantly associated with markers in the *PSY2* gene. The SNPs from *lcyE* were significantly associated with color while those of *lcyB* and *crtRB* were not significantly associated with carotenoids or color parameters. These validated and breeder-friendly markers have potential to enhance the efficiency of selection for high β-carotene cassava, thus accelerating genetic gain.

Keywords Cassava · Single nucleotide polymorphism · Marker-assisted selection · Candidate gene association · Vitamin a · Biofortification

Background

Cassava (*Manihot esculenta* Crantz) is one of the most important staple crops in Africa and consumed as a cheap source of carbohydrates (FAO 2013). The crop is easy to grow and plays a major role in food security (Tonukari 2004). However, most cultivated varieties have white starchy roots which contain negligible amount of carotenoids and other micronutrients (Rodriguez-Amaya and Kimura 2004). Therefore, over-dependence on cassava-based diets may result in

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11032-017-0718-5) contains supplementary material, which is available to authorized users.

L. I. Udoh · M. Gedil · E. Y. Parkes · P. Kulakow · C. Nwuba · I. Y. Rabbi (✉)
International Institute for Tropical Agriculture (IITA), PMB 5320, Ibadan, Oyo, Nigeria
e-mail: I.Rabbi@cgiar.org

L. I. Udoh (✉) · A. Adesoye
Department of Botany, University of Ibadan, Ibadan, Nigeria
e-mail: lnkantol@yahoo.com

Published online: 11 September 2017

poor health, stunted growth, reduced capacity for physical activity, and in extreme cases, a high incidence of anemia, corneal blindness, and compromised immunity (Saltzman et al. 2013). This is especially true in regions with prolonged dry seasons that limit production and access to alternative sources of micronutrients such as fresh vegetables (von Grebmer et al. 2014).

The yellow color of the cassava storage root is strongly associated with carotenoids content. Reported Pearson's coefficient, *r*, between yellow color and total carotenoid content ranges from 0.81 to 0.89 (Iglesias et al. 1997; Akinwale et al. 2010; Sánchez et al. 2014). Strong correlation between yellow root color and total β-carotene has also been reported (Sánchez et al. 2014). Pro-vitamin A carotenoids include α-carotene, β-carotene, and β-cryptoxanthin which are precursors of vitamin A, a micronutrient essential for normal development and functioning of the human body (West and Mehra 2010). β-carotene is the major carotenoid in cassava occurring in *trans* and *cis* forms (Carvalho et al. 2016).

Biofortifying cassava with pro-vitamin A will have a significant positive impact on nutrition and general health, especially for low-income communities. Among the major approaches for biofortification are conventional breeding and genetic modification. Various initiatives, including the project that Harvest Plus is implementing, use conventional breeding for the enhancement of pro-vitamin A carotenoid in cassava roots (Pfeiffer and McClafferty 2007). Pro-vitamin A varieties that are presently available provide up to 40% of the daily recommended vitamin A intake for children less than 5 years old (De Moura et al. 2015). Nevertheless, new crosses to select varieties with an even higher content of β-carotene varieties are being generated through recurrent selection breeding scheme (Sánchez et al. 2014).

Cassava breeding through phenotype-based recurrent selection is a lengthy process owing to its 12-month cropping cycle and the low clonal multiplication rate of about 5–10 clones per generation for every parent clone propagated. This, coupled with the costly and laborious biochemical assay for quantifying pro-vitamin A carotenoid content necessitates development of alternative selection tools, particularly molecular markers, which can be used to accelerate the development of new varieties. In marker-assisted selection (MAS), individuals are selected based on the presence of favorable alleles at trait-linked markers or the target candidate genes rather than the actual trait measurement.

Moreover, MAS is also economical when several traits are simultaneously selected for, for example, pro-vitamin A combined with must have traits such as resistance to cassava mosaic disease (CMD) and cassava brown streak disease (CBSD), enabling huge reductions in the size of phenotyping trials with associated cost reductions. By reducing the number of seedlings using MAS, field breeders can allocate their limited phenotyping resources to a smaller number of selection candidates for further phenotypic evaluation of complex traits such as yield and tolerance to biotic and abiotic stresses.

The genes encoding enzymes of carotenoid biosynthesis in plants are well known and their pathway has been extensively studied (Cazzonelli 2011; Hannoufa and Hossain 2012; Giuliano 2014; Nisar et al. 2015) (Supplementary Fig. 1). The gene phytoene synthase (*PSY2*) has been reported to be associated with carotenoid accumulation in cassava (Welsch et al. 2010; Esuma et al. 2016). This gene belongs to a family with three copies in the cassava reference genome. Of these, the most important is *PSY2* which contains 6 exons and occurs in chromosome 1 … from 24,153,420 to 24,156,720 bp while the other two are found on chromosome 2 (position 6091016 to 6093529 + strand) and on chromosome 3 (position 13755566 to 13757622 + strand). The last one is not supported by expressed sequence tag data in Phytozome. Other genes such as lycopene epsilon cyclase (*lcyE*), β-carotene hydroxylase (*crtRB*), and lycopene beta cyclase (*lcyB*) have been also reported to play key roles in carotenoid variation in maize (Harjes et al. 2008; Bai et al. 2009; Yan et al. 2010; Babu et al. 2013; Fu et al. 2013). A candidate gene association mapping approach is one way to identify single nucleotide polymorphisms (SNPs) in specific genes controlling a trait of interest (Zhu and Zhao 2007; Patnala et al. 2013). This approach has been used to identify variants in the nucleotide sequence of carotenoid genes that contribute significantly to the accumulation of pro-vitamin A and total carotenoids in maize endosperm and in cassava roots (Harjes et al. 2008; Yan et al. 2010; Welsch et al. 2010; Fu et al. 2013). In maize endosperm, Yan et al. (2010) identified three polymorphic sites in the gene encoding *crtRB* accounting for 40% of the observed variation in β-carotene concentration. Also, Harjes et al. (2008) have shown that four polymorphic sites in the gene encoding *lcyE* were associated with allelic variations that lead to a threefold increase in pro-vitamin A. Two polymorphisms in the

gene encoding phytoene synthase (*PSY1)* were identified to explain 7 to 8% of the variation in total carotenoids (Fu et al. 2013), and significant effects have been detected for all the functional polymorphisms for individuals and haplotypes of selected polymorphisms of *lcyE*, *crtRB1*, and *PSY1*, using inbred lines of maize with tropical, subtropical, and temperate backgrounds (Yan et al. 2010; Azmach et al. 2013; Fu et al. 2013). In the maize breeding program, variations identified in candidate genes have already been exploited as functional markers for selection for grains that accumulate high levels of carotene (Harjes et al. 2008; Yan et al. 2010; Azmach et al. 2013; Babu et al. 2013).

Therefore, the objective of this study was to carry out candidate gene association mapping for β-carotene content in cassava. We also exploited the potential candidate SNPs as functional markers through allele-specific polymerase chain reaction assay development and validation for ultimate utilization in improving selection efficiency and accelerating genetic gain in breeding pipelines.

Methods

Plant material used and study locations

The study material is a large collection of diverse elite breeder lines present at the International Institute of Tropical Agriculture (IITA) and also referred to as the genetic gain collection (Okechukwu and Dixon 2008). It consist of more than 650 accessions that show variability in storage root color ranging from white to deep yellow. The pedigree of the collection is made up of crosses among germplasm from East and West Africa and Latin America. The population was planted using an augmented design with two checks per block and a single row of 10 plants spaced at 1 m^2. A subset of 167 accessions was selected to constitute the candidate gene association panel. Accessions were selected to ensure they were adequate representatives with respect to carotenoid content variation and population structure. Another subset of 100 accessions was selected from the candidate gene association panel to constitute the carotenoid extraction panel. These are yellow-fleshed cassava with root color chart scores of 2 and above with white accessions included as control. Because of the high cost of high-performance liquid chromatography (HPLC) and the limited number of samples that can be handled per day, a total of 100 accessions was selected for analysis. Data were collected from two field seasons (2013–2014 and 2014–2015) in two cassava growing agro-ecologies in Nigeria—Ibadan (7.40° N, 3.90° E) and Ubiaja (6.66° N, 6.38° E).

Determination of β-carotene content

Different measurements related to carotenoid content were recorded: (i) indirect color assessment through the use of a color chart on a scale of 1 (white) to 6 (deep yellow), (ii) pulp color score of 1 for white and 2 for yellow lines, and (iii) the use of a digital chromameter® (Konica Minolta CR-410 ®, Japan). The chromameter provides a precise and objective assessment of surface color. Data output in the form of the L* a* b* hunter color coordinate system was used. L* corresponds to levels of darkness/lightness between black and white. Coordinate a* signifies the balance between red and green, and b* between yellow and blue. The color intensity of each accession according to L*a*b* hunter color was determined by sampling five roots per accession. Roots were peeled, washed, cut into two, and grated from the middle portion. Grated samples were mixed thoroughly and 100 g was collected into Whirl-Pak™ sample collection bags. To minimize error from an uneven distribution of carotenoids across storage roots, four measurements were taken on different parts of the sample bag. Data represents the mean of measurements per sample. The positive b* coordinate of the chromameter was considered in this study because it measures yellow color.

Direct quantification of carotenoids in the carotenoid quantification panel was also carried out using HPLC spectrophotometer and iCheck Fluoro™ device developed by BioAnalyt (www.bioanalyt.com). Extraction of carotenoids with acetone was performed using a pestle and mortar as described by Rodriguez-Amaya and Kimura (2004). Total carotenoids content (TC) was quantified by spectrophotometer (Beckman Coulter DU® 530). Afterwards, HPLC analysis was performed twice for each sample using the modified method of Howe and Tanumihardjo (2006) to measure β-cryptoxanthin, 9-cis-β-carotene, 13-cis-β-carotene, and Trans-β-carotene. Total β-carotene (TBC) was calculated as a sum of 9-cis-β-carotene, 13-cis-β-carotene, and trans-β-carotene as described by Azmach et al. (2013).

The iCheck Fluoro™ BioAnalyt was used to quantify TC by collecting approximately 5 g of the sample into a medium-sized mortar and grinding it with 20 ml of distilled water as the solvent. The resultant solution was poured into a 50-ml falcon tube and shaken to get a homogenous slurry: 0.4 ml of the solution was injected into the iCheck reagent vial. The vial was shaken and allowed to stand for 5 min to obtain a clear upper phase and a turbid lower phase. Measurement was taken after the device was calibrated with the solid control. To get the concentration of TC in cassava roots, the result was multiplied by the dilution factor (total sample volume in water/sample weight) as described by Esuma et al. (2016).

Phenotype data analysis

The following mixed linear model was fitted using the *lme4* package in R to generate best linear unbiased predictor (BLUP)

$$y_{l,i.j} = \mu + c_l + \beta_i + \varepsilon_{l,i,j}$$

Here, $y_{l,i,j}$ represents raw phenotypic observations, μ is the grand mean, c_l is a random effects term for accession, β_i is a fixed effect term for the combination of location and year harvested, and $\varepsilon_{l,i,j}$ is the residual variance, assumed to be random and normally distributed. Broad-sense heritability for chromameter b* reading, color chart score, TC quantified by spectrophotometer, and TC quantified by iCheck Fluoro were calculated according to Ly et al. (2013). Pearson's correlation coefficient was calculated to test the relationship between color components and quantified carotenoids.

Candidate genes sequencing

Candidate genes sequenced from 167 cassava lines are β-carotene hydroxylase (*crtRB*), phytoene synthase (*PSY2*), lycopene epsilon cyclase (*lcyE*), and lycopene beta cyclase (*lcyB*). The respective genes were retrieved from the Phytozome database (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Mesculenta). BatchPrimer3 v1.0, a high throughput web application for PCR and sequencing primer design (You et al. 2008), was used to design primers around the four genes. Two sets of primers were designed for each gene to cover both 3′ and 5′ UTRs (untranslated regions) of the genes to sequence the full length by Sanger sequencing. Also, in previous findings, useful polymorphisms have been identified in the UTR regions (Harjes et al. 2008). Selected primers were submitted to the National Center for Biotechnology Information (NCBI) BLASTn algorithm (Altschul et al. 1997) for similarity search and ensure that primers align to target genes. Those that gave spurious hits were excluded.

Total RNA extraction, Sanger sequencing, and alignment

Approximately 0.1 g of fresh leaf was collected from the 167 candidate gene association panel into sterile 2-ml Eppendorf tubes and immediately put into liquid nitrogen to avoid sample degradation. Total RNA was extracted using the cetyl trimethyl ammonium bromide (CTAB) method, modified from Lodhi et al. (1994). First strand cDNA was synthesized from 200 ng of total RNA using the Thermo Scientific Maxima H Minus First Strand cDNA Synthesis Kit according to manufacturer's instruction. The cDNA synthesized served as the template for PCR amplification with gene-specific primers.

The PCR reaction was performed in volumes of 25 µl reaction with 200 ng/µl cDNA template, 10× Kcl Taq buffer, 25 mM $MgCl_2$, and 2.5 µM dNTPs, 0.01unit of Taq polymerase, 4–5% DMSO, and 10 pM of forward and reverse primers. Amplification was carried out with a touchdown program performed on Biorad Peltier thermal cycler. Initial DNA denaturation was 94°C for 2 min, followed by 9 cycles of 93°C for 20s, at −1°C per cycle, and 72°C for 45 s followed by 35 cycles of 93°C for 20s, 55–65°C for 30s, and 72°C for 45 s. A final extension step was performed at 72°C for 5 min. PCR reaction products with the correct length were confirmed using 1.5% agarose gel containing 0.5 mg/ml ethidium bromide in 1× TBE buffer, and visualized using the gel image system (ENDURO™ GDS Labnet international, Inc).

PCR products with single clear band obtained after a purification stage were sequenced in both forward and reverse directions at Iowa State University DNA sequencing facility, USA, with the same PCR primers on ABI Prism 3130X1 Genetic Analyzer (Applied Biosystems) and BigDye terminator V3.1 kit (Applied Biosystems Inc.). Raw sequences were edited in CodonCode Aligner (v3.7.1) by trimming both ends, removing bad chromatograms, and calling SNPs. Nucleotide sequences were aligned by ClustalW

(Thompson et al. 1994). Sequences were examined for SNPs and insertion deletions.

Population structure and candidate gene association mapping

Candidate gene SNP-trait association was performed using four models: (i) a naïve model without correction for population structure, (ii) a general linear model (GLM) model with principal component analysis (PCA) as a covariate (Price et al. 2010), and (ii) a Mixed linear Model (MLM) implementing PCA as fixed effects and a kinship matrix as random genotype effects (VanRaden 2008). The PCA-based population structure and kinship matrix were calculated using previously reported genome-wide SNP data from the IITA elite breeder cassava lines (Wolfe et al. 2016). Principal component analysis was calculated using PLINK and Kinship matrix using TASSEL (Bradbury et al. 2007), from 37,205 SNPs with minor allele frequency of 0.05 or more in the candidate gene association panel.

Association mapping models were evaluated based on visual observation of the quantile-quantile plots of observed—log10 p values versus expected—log10 p values. In this study, the first approach to identify the association signal was based on smallest p values obtained from mixed models and the threshold was set at $p < 0.01$, given the small number of candidate gene SNPs. Secondly, test p values were considered significant when more extreme than the Bonferroni threshold (with experiment-wise type I error rate of 0.05).

Validation of candidate SNPs using KASP SNP assay

Following the association analysis of candidate gene SNPs, a total of six SNPs were selected for competitive allele-specific PCR assay (KASPar™) development (LGC Genomics, UK) for validation in the entire elite breeder lines in the genetic gain population. Two *PSY2* SNPs were selected based on their significance according to the Bonferroni threshold, and four *lcyE* SNPs were selected based on the smallest p value set at a threshold of $p < 0.1$. Fifty nucleotide bases flanking the SNP of interest on each side were used for the KASP assay design (Supplementary Table 1). DNA from the elite breeder lines was extracted as described in Rabbi et al. (2014) and genotyped using KASP. The allele detection is based on fluorescence resonance energy transfer quencher cassette which allows bi-allelic discrimination of known SNPs. A no template control was included in the SNP genotyping.

Results

Phenotypic variation for carotenoids and color components

Total carotenoids content was quantified by spectrophotometry (TC SPEC) and by iCheck Fluoro™ (TC iCheck). Carotenoid content was also indirectly assessed by estimating the yellow pigmentation of storage roots, by color chart and digital color measurement using a chromameter (Minolta CR-410 ®). The average color chart score was 2.02 and ranged from 1 to 6. Chromameter b* value reading showed a range of variation from 11.9 to 40.8 and an average of 22.9.

Average direct estimate of total carotenoids using spectrophotometry (TC SPEC) was 3.75 μg/g and ranged from 0.07 to 13.34 μg/g. The average level of total carotenoids as measured using iCheck (TC iCheck) was found to be higher than in TC SPEC (6.09 μg/g). Total β-carotene (TBC) estimated using HPLC ranged from 0.07 to 10.14 μg/g with an average of 2.73 μg/g (Table 1). Among the carotenoid components quantified through HPLC, all-trans-β-carotene had the highest mean value of 1.51 μg/g ranging from 0.03 to 6.70. Root color assessment by chromameter b* reading showed a binomial distribution where b* values from 10 to 20 are associated with the white lines while values from 20 to 40 are associated with variations in the yellow lines. Total carotenoids as quantified by iCheck Fluoro showed a normal distribution; although most of the elite breeder cassava lines were white-fleshed, others had a variation of yellow color as seen in the distribution of the color chart score (Supplementary Fig. 2). Total β-carotene had high correlation coefficient with TC SPEC ($r = 0.90$) and TC iCheck ($r = 0.75$). Also, TBC had a moderate correlation coefficient with color assessment by color chart score $r = 0.64$ and chromameter (b*) reading $r = 0.62$.

Sequence analysis and identification of nucleotide variants

The four candidate genes (phytoene synthase, lycopene beta cyclase, lycopene epsilon cyclace, and β-carotene hydroxylase) were identified by BLAST analysis using

Table 1 Carotenoid content variation in the elite breeder lines

Data set	Trait[a]	Number	Min	Max	SD	Mean
Validation set	Pulpcol	2858	1.00	2.00	0.50	1.51
	Color chart	2860	1.00	6.00	1.25	2.02
	Chromameter b*	1360	11.88	40.78	7.76	22.90
	TC iCheck (μg/g)	642	0.00	16.74	3.07	6.09
Candidate gene panel	TC SPEC (μg/g)	252	0.07	13.34	2.45	3.75
	βcryp (μg/g)	252	0.01	1.93	0.15	0.10
	13cisβc (μg/g)	252	0.02	2.06	0.31	0.50
	Transβc (μg/g)	252	0.03	6.70	1.17	1.51
	9cisβc (μg/g)	252	0.02	2.77	0.49	0.72
	TBC (μg/g)	252	0.07	10.14	1.79	2.73

Pulpcol pulp-color score, *TC SPEC* total carotenoid by spectrophotometer, *TC iCheck* total carotenoid by iCheck Fluoro, *βcryp* β-cryptoxanthin, *13cisβc* 13-cis-β-carotene, *Transβc* all-trans-β-carotene, *9cisβc* 9-cis-β-carotene, *TBC* total β-carotene, *SD* standard deviation, *N* number of samples, *Min* minimum, *Max* maximum

[a] Carotenoid content parameters in the candidate gene panel was measured using HPLC and spectrophotometry while in the larger validation population, the indirect estimation methods were used, except for the TC iCheck portable spectrophotometry

previously published cassava genes and those from *Arabidopsis thaliana*, rice (*Oryza sativa*), and castorbean (*Ricinus communis*) against the cassava genome assembly (version 6.1, available at https://phytozome.jgi.doe.gov/pz/portal.html). We confirmed previous findings by Welsch et al. (2010) and uncovered three *Phytoene synthase* genes (PSY1-Manes.02G081700, *PSY2*-Manes.01G124200, and PSY3-Manes.03G084700). Of these, only *PSY2* was found to be expressed and supported by full-length PASA (program to assemble spliced alignments) aligned cDNA sequence data. Three lycopene beta cyclase (Manes.16G099600, Manes.09G008200, and Manes.11G163700) were uncovered in the cassava reference genome. Of these, only Manes.16G099600 occurring on chromosome16 from the 25,581,180 to 25,586,577 bp region was found to expressed and supported by full-length PASA-aligned cDNA. One full-length copy of the *lycopene epsilon cyclase* (Manes.09G008200) was uncovered in the reference genome. Two copies of *β-carotene hydroxylase* (Manes.06G152200 and Manes.02G018300) were uncovered using BLAST search of which only Manes.06G152200 was supported by gene expression data.

The genomic location of the candidate genes selected for re-sequencing, their heterozygosity, and number identified SNPs are presented in Supplementary Table 2. A total of 37 SNP sites were discovered across candidate genes as follows: 5 in *crtRB*, 6 in *PSY2*, 13 in *lcyE*, and 13 in *lcyB*. Frequency of SNP occurrence on average was every 150 bp for *lcyE*, 123 bp for *lcyB*, 143 bp for *crtRB*, and 165 bp for *PSY2*. On average, the *lcyB* gene was more heterozygous (56.2%) while the *crtRB* gene was least (7.6%). Nucleotide sequences were translated to amino acid to determine the number of synonymous and non-synonymous changes. In the *PSY2* gene, five polymorphic sites were synonymous except SNP A/C at position 572 which had a non-synonymous amino acid exchange from alanine-A (a non-polar, neutral amino acid) to aspartic acid-D (an acidic and polar amino acid). Translation of nucleotide sequences to amino acid in *lcyE* showed nine non-synonymous amino acid changes and five in the *lcyB* gene. All observed polymorphism in the *crtRB* gene were synonymous (Supplementary Table 2).

Due to its importance in the carotenoid biosynthesis in cassava, further bioinformatics analysis was carried out for *PSY2*. Multiple sequence alignment of the three cassava PSY homologs as well as those from other species was performed using ClustalW in JalView (Clamp et al. 2004). Also included in the alignment were the sequences from two white- and two yellow-root cassava clones as examples. We found that the non-synonymous change from alanine to aspartic acid at position 191 of *PSY2* from the reference genome (Manes.01G124200) occurs in the conserved region of the gene (Supplementary Fig. 3). Furthermore, the change from A to D is only found in cassava and not

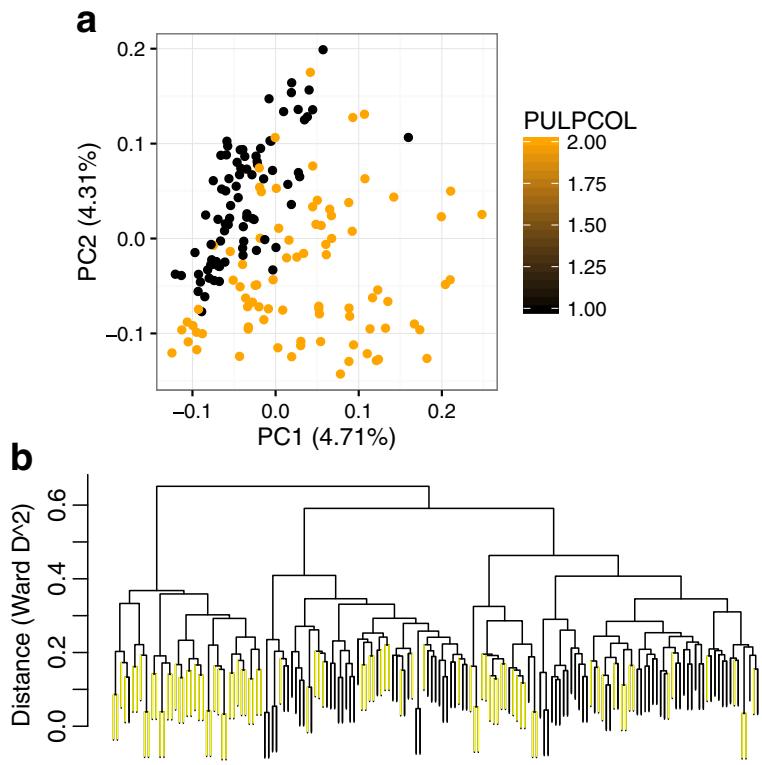

Fig. 1 The population structure of 167 elite breeder cassava lines used for association analysis. **a** Principal component analysis based on genome-wide SNPs and *numbers in parenthesis* indicate proportion of variance explained by the respective axis. **b** Neighbor-joining dendrogram of the breeder cassava lines and *yellow color* highlights yellow accessions

in any of the other 170 homologs from the sequences Viridiplantae available in Phytozome.

Population structure and linkage disequilibrium

The population structure of the selected 167 accessions used for candidate gene association analysis was analyzed using 37,205 genome-wide SNPs after filtering out those with minor allele frequency (MAF) of less than 0.05. The population structure is shown by PCA plot and indicates no clustering but a genetic differentiation between the white and yellow accessions. Neighbor-joining dendrogram from pairwise identity-by-state distance coefficients among all pairs of accessions shows a cluster of yellow root accessions and mixed ancestry between white and yellow lines (Fig. 1).

Linkage disequilibrium (LD) was not observed between the candidate genes in this study. All pairwise squared correlations (R^2) between the SNPs located in different genes were less than 0.2. High LD was observed between SNPs within the *lcyB* gene while the SNPs within the other candidate genes *lcyE*, *crtRB*, and *PSY2* exhibited low LD (Supplementary Fig. 4).

Choice of model and candidate gene association mapping

To control for false-positives resulting from effects of population stratification and relatedness, four models were compared. The GLM (naïve model) had the highest inflation rate of p values as determined by quantile-quantile (Q-Q) plot (Supplementary Fig. 5). The model correcting for population structure using PCA as covariate (P) also showed a high level of inflation. The MLM method had the lowest inflation factor as determined by QQ plots and therefore the lowest false-discovery rate. Thus, the results shown are only from the MLM model correcting for kinship and three PCs (P + K).

A summary of the most significant association from the MLM analysis is presented in Table 2. A total of 37 SNP sites across four candidate genes were tested for association with directly quantified carotenoids and color parameters. The *PSY2*-572 SNP was associated with root pulp color (p value = 6.80×10^{-23}, R^2 = 0.79), chromameter b* reading (p value = 1.17×10^{-16}, R^2 = 0.51), and TBC (p value = 1.29×10^{-04}, R^2 = 0.15), while the *PSY2*-549 SNP was associated

Table 2 Summary of significant SNPs from candidate gene association

Marker	Trait	SNP	Amino acid substitution	*p* value	Marker R^2
PSY2_572	Pulpcol	A/C	Ala/Asp	6.80×10^{-23}	0.79
	b*			1.17×10^{-16}	0.51
	Color chart			7.73×10^{-14}	0.37
	TBC			1.29×10^{-04}	0.15
	TC SPEC			2.83×10^{-04}	0.14
	TC iCheck			3.43×10^{-04}	0.13
PSY2_549	Pulpcol	T/C		1.83×10^{-12}	0.35
	b*			3.34×10^{-09}	0.24
	Color chart			6.53×10^{-08}	0.19
	TC iCheck			0.001512	0.10
	TC SPEC			0.002175	0.10
	TBC			0.005574	0.08
PSY2_870	b*	A/C		1.44×10^{-05}	0.13
	Pulpcol			1.58×10^{-05}	0.13
	Color chart			7.90×10^{-05}	0.10

Pulpcol pulp color score, *b** chromameter b*coordinate, *TBC* total β-carotene, *PVAC* pro-vitamin A carotenoid, *TC SPEC* spectrophotometer, *TC iCheck* total carotenoid estimated by iCheck Fluoro™

with root pulp color (p value = 1.83×10^{-12}, $R^2 = 0.35$) (Fig. 2). The two SNPs were significantly associated with most of the tested parameters. Considering the significant threshold of $p < 0.01$ in the GLM model, the *lcyE*-1015 SNP was associated with TC quantified using iCheck Fluoro ($R^2 = 0.08$) and *lcyE*-1294 SNP was associated with chromameter b* reading ($R^2 = 0.06$). Only the *PSY2*-572 SNP caused a non-synonymous amino acid substitution between yellow and white cassava accessions and also explained most of the variation for directly and indirectly estimated carotenoids.

Development and validation of competitive allele-specific PCR assays

To facilitate MAS for β-carotene in cassava, we converted a total of six SNPs to KASP markers, which include two highly significant *PSY2* SNPs (*PSY2*_572 and *PSY2*_549) and four *lcyE* SNPs (*lcyE*_829,

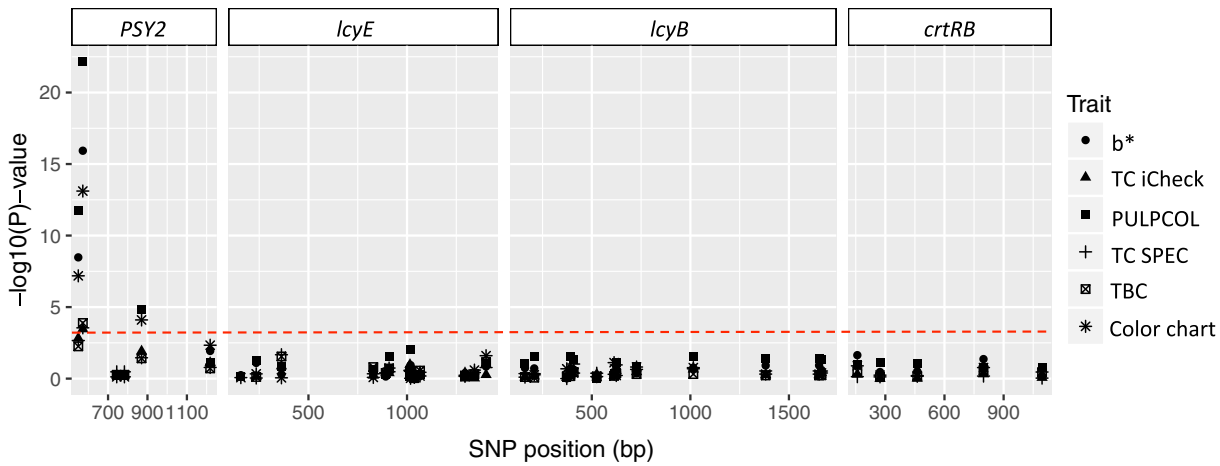

Fig. 2 Manhattan plot of MLM (P + K) model for carotenoid content and color scores. Candidate genes are displayed at the top according to pathway order and *red horizontal lines* indicate Bonferroni threshold

Table 3 Summary results of validated SNP markers on the larger breeding collection

Marker	MAF	Het	PIC	Trait	*p* value	Marker R^2
PSY2_572	0.81	0.16	0.26	Color chart	3.26×10^{-198}	0.75
				b*	7.38×10^{-199}	0.78
				TC SPEC	1.95×10^{-20}	0.62
				TBC	3.96×10^{-18}	0.57
PSY2_549	0.76	0.25	0.30	Color chart	3.64×10^{-146}	0.63
				b*	7.77×10^{-120}	0.59
				TC SPEC	1.08×10^{-19}	0.59
				TBC	5.58×10^{-17}	0.54
lcyE_1066	0.73	0.32	0.32	TBC	9.63×10^{-04}	0.13
				TC SPEC	0.00322	0.11
				Color chart	0.00262	0.02
				b*	0.01152	0.01
lcyE_1294	0.98	0.04	0.04	Color chart	3.93×10^{-06}	0.03
				b*	1.94×10^{-07}	0.04
lcyE_1015	0.96	0.05	0.07	Color chart	1.36×10^{-04}	0.03
				b*	8.86×10^{-06}	0.04
lcyE_829	0.82	0.19	0.21	Color chart	0.01377	0.01

MAF major allele frequency, *Het* heterozygosity, *PIC* polymorphic information content

*lcyE*_1066, *lcyE*_1015, and *lcyE*_1294) selected based on the significant threshold of $p < 0.01$ across all carotenoid parameters. Designed assays were validated on 650 genetic gain cassava accessions. A total of 4122 allele calls were made with a call rate of 98.0%. Missing data across the markers upon genotyping ranged from 0.7% in *lcyE*_1294 to 7.5% in *PSY2*_572. Summary diversity statistics of the validated SNPs including frequencies of their major alleles and diversity are presented in Table 3. Polymorphic information content, a measure of polymorphism for each marker, ranged from 0.04 in *lcyE*_1294 to 0.32 in *lcyE*_1066, with an average of 0.21. Observed heterozygosity values for each SNP ranged from 0.04 in *lcyE*_1294 to 0.39 in *lcyE*_1066 with an average of 0.25 (Table 3). The alleles generated by KASP genotyping were used to test association between KASP markers and carotenoids as estimated by spectrophotometer, color assessment by chromameter b* reading, and color chart scores. A GLM was used for the association analysis because of the small number of SNP markers involved. Marker *PSY2*_572 explained most of the proportion of phenotypic variation for the chromameter b* reading (R^2 of 0.78) and TC SPEC (R^2 of 0.62) (Table 3). Marker *lcyE*_829 explained the lowest of proportion of phenotypic variation.

Discussion

Cassava is an important staple crop for food and livelihood security in Africa. To alleviate pro-vitamin A deficiency where cassava is the main staple, the development of improved cultivars with high levels of pro-vitamin A is a major breeding goal in the continent. Deployment of molecular markers in the breeding pipeline can increase efficiency and accelerate the rate of genetic improvement with respect to pro-vitamin A biofortification, especially when done in parallel with the marker-assisted selection for other traits such as resistance to virus diseases and increase in dry matter content. Use of functional markers will allow breeders to preselect the best candidates that combine desired characteristics for subsequent field evaluation. In the present study, we conducted a candidate gene association study to uncover SNP markers linked to increased carotenoids in cassava storage roots. The significant SNP variants were converted to KASP PCR assays. The markers were validated by using them to genotype a larger collection of elite breeder lines of IITA.

Good phenotyping for nutritional quality is invaluable in carrying out downstream analysis. In cassava, HPLC is the standard method for accurate quantification of carotenoids but the procedure is not only expensive but also very low-throughput for analyzing breeding populations

that can be large (from hundreds to thousands plots per trial). Given the rapid deterioration of cassava roots and stems, rapid evaluation and selection methods are required. In this study, we found a strong correlation between color parameters and direct quantification of carotenoids through spectrophotometer, iCheck Fluoro, and HPLC-derived TBC. However, total carotenoid quantified by the iCheck Fluoro device showed a moderate correlation with color assessment by color chart and pulp color score. The chromameter b* reading was consistent in its strong relationship between total carotenoids and TBC. Thus, in searching for a precise, more accurate, and faster means of quantifying TC especially in large populations, the chromameter b* has proved to be a reliable instrument. Other studies have also reported a strong correlation between intensity of yellow color in storage roots and TC content as well as TBC (Chávez et al. 2005; Marín Colorado et al. 2009; Akinwale et al. 2010; Sánchez et al. 2014; Esuma et al. 2016).

So far, in cassava, mostly SSR markers have been applied in MAS breeding (Ferguson et al. 2012). Njoku et al. (2014) validated two SSR markers, namely NS717 and SSR301, in a molecular marker analysis study of F_1 progenies and their parents for inheritance of carotenoids in some African cassava. These SSR markers accounted for 0.19 and 0.20, respectively, of phenotypic variance for carotenoids. However, SNP markers have not been applied in cassava breeding for increased pro-vitamin A. To carry out candidate gene sequencing, a total of 167 elite breeder cassava lines were analyzed and 37 SNP sites were generated across the studied candidate genes. The *PSY2* enzyme is known to convert geranyl-geranyl pyrophosphate in the biosynthesis pathway to phytoene and has been reported to play a major role in carotenogenesis in cassava from Latin America (Welsch et al. 2010). Our research, which focused on African germplasm, indicates that the same locus underlies a large part of the variation in carotenoid content. On the other hand, only moderate association between SNPs in *lcyE* and carotenoid content was uncovered, despite the presence of many non-synonymous variations in this and other genes in the present study. This may be due to the small population that was analyzed for total carotenoids and TBC. Carotenoid accumulation in other staples, such as maize, has been shown to be largely mediated by differential expression of genes encoding *lcyB* and *lcyE* (Harjes et al. 2008; Bai et al. 2009). Cyclization of lycopene is the first branch point of the carotenoid pathway where the action of lycopene beta cyclase (*lcyB*) at both ends produces a molecule with two β rings. On the other hand, the action of both *lcyB* and lycopene epsilon cyclase (*lcyE*) generates a β,ε-carotene that is a precursor to lutein (Ruiz-Sola and Rodríguez-Concepción 2012). Nevertheless, the simple genetic architecture of carotenoid variation in the studied cassava germplasm is in agreement with previous studies (Iglesias et al. 1997; Chávez et al. 2005).

Moving from discovery to breeding application through MAS, a total of six significant SNPs consisting of two from the *PSY2* gene and four from the *lcyE* gene were used to design singleplex KASP SNP assays. In maize breeding for increased β-carotene, SNP markers are currently being utilized (Harjes et al. 2008; Yan et al. 2010; Azmach et al. 2013; Fu et al. 2013). We therefore propose to deploy the discovered SNPs markers from the present study in cassava biofortification breeding programs in Africa.

In the present study, we preselected candidate genes based on a priori knowledge and available gene annotations in the cassava reference genome. To uncover additional trait-linked markers, we recommend a follow-up study using the genome-wide association mapping approach in a larger population genotyped at genome-wide SNP markers. Such study has the potential to uncover novel genes that are not in the present reference assembly.

Conclusions

Biofortification of cassava with pro-vitamin A through breeding is important especially in sub-Saharan Africa where the crop is widely grown and consumed as a major staple. There is a need to utilize molecular markers such as single nucleotide polymorphisms to facilitate breeding for pro-vitamin A in cassava through marker-assisted selection. In this study, we sequenced four candidate carotenoid genes (*lcyE*, *PSY2*, *lcyB*, and *crtRB*) to identify SNP variants that could be associated with trait variation. Through association analysis, we identified those SNPs that are significantly linked to carotenoid content and exploited them as gene-based markers. These breeder-friendly markers developed in this study can greatly increase the efficiency of selection in breeding for high pro-vitamin A cassava and genetic gain.

Acknowledgements We acknowledge the support of Ruth Uwugiaren and Tessema Gezahegn in the laboratory work and Andrew Ikpan and the staff of the Cassava Breeding Unit of IITA

for conducting field trials. The Next Generation Cassava Breeding project (www.nextgencassava.org) is appreciated for providing the genome-wide information on the SNPs of the population used in this study. This research was part of a PhD project supported by the HarvestPlus Project and the CGIAR Research Program on Roots, Tubers, and Bananas (CRP-RTB).

Authors' contributions LU conducted the field and laboratory experiments, carried out statistical analysis, and prepared the draft manuscript. MG supervised the laboratory work and revised the manuscript. EP and PK contributed in securing funds for the project and revised the manuscript. AA contributed to supervision of the work and revised the manuscript. CN contributed to the laboratory and field work and revised the manuscript. IR conceived and led the project, took part in statistical analysis, and contributed to drafting the manuscript.

Compliance with ethical standards

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and materials The datasets generated during and/or analyzed during the current study are available at www.cassavabase.org, ftp://ftp.cassavabase.org/manuscripts/MolecularBreeding_Udoh_et_al_2017.zip.

Competing interest The authors declare that they have no competing interest.

References

Akinwale MG, Aladesanwa RD, Akinyele BO et al (2010) Inheritance of β-carotene in cassava (Manihot esculenta crantza). Int J Genet Mol Biol 2:198–201

Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Azmach G, Gedil M, Menkir A, Spillane C (2013) Marker-trait association analysis of functional gene markers for provitamin A levels across diverse tropical yellow maize inbred lines. BMC Plant Biol 13:227. https://doi.org/10.1186/1471-2229-13-227

Babu R, Rojas NP, Gao S et al (2013) Validation of the effects of molecular marker polymorphisms in LcyE and CrtRB1 on provitamin A concentrations for 26 tropical maize populations. Theor Appl Genet 126:389–399. https://doi.org/10.1007/s00122-012-1987-3

Bai L, Kim EH, Dellapenna D, Brutnell TP (2009) Novel lycopene epsilon cyclase activities in maize revealed through perturbation of carotenoid biosynthesis. Plant J 59:588–599. https://doi.org/10.1111/j.1365-313X.2009.03899.x

Bradbury PJ, Zhang Z, Kroon DE et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Carvalho LJ, Agustini MA, Anderson JV et al (2016) Natural variation in expression of genes associated with carotenoid biosynthesis and accumulation in cassava (Manihot esculenta Crantz) storage root. BMC Plant Biol 16:133. https://doi.org/10.1186/s12870-016-0826-0

Cazzonelli CI (2011) Goldacre review: carotenoids in nature: insights from plants and beyond. Funct Plant Biol 38:833. https://doi.org/10.1071/FP11192

Chávez AL, Sánchez T, Jaramillo G et al (2005) Variation of quality traits in cassava roots evaluated in landraces and improved clones. Euphytica 143:125–133. https://doi.org/10.1007/s10681-005-3057-2

Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. Bioinformatics 20:426–427. https://doi.org/10.1093/bioinformatics/btg430

De Moura FF, Moursi M, Lubowa A et al (2015) Cassava intake and vitamin A status among women and preschool children in Akwa-Ibom, Nigeria. PLoS One 10:e0129436. https://doi.org/10.1371/journal.pone.0129436

Esuma W, Herselman L, Labuschagne MT et al (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

FAO (2013) Save and Grow: Cassava. A guide to sustainable production intensification. FAO, Rome, pp 130. http://www.fao.org/ag/save-and-grow/cassava/index_en.html

Ferguson M, Rabbi I, Kim DJ et al (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

Fu Z, Chai Y, Zhou Y et al (2013) Natural variation in the sequence of PSY1 and frequency of favorable polymorphisms among tropical and temperate maize germplasm. Theor Appl Genet 126:923–935. https://doi.org/10.1007/s00122-012-2026-0

Giuliano G (2014) Plant carotenoids: genomics meets multi-gene engineering. Curr Opin Plant Biol 19:111–117. https://doi.org/10.1016/j.pbi.2014.05.006

Hannoufa A, Hossain Z (2012) Regulation of carotenoid accumulation in plants. Biocatal Agric Biotechnol 1:198–202. https://doi.org/10.1016/j.bcab.2012.03.004

Harjes CE, Rocheford TR, Bai L et al (2008) Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319:330–333. https://doi.org/10.1126/science.1150255

Howe JA, Tanumihardjo SA (2006) Evaluation of analytical methods for carotenoid extraction from biofortified maize (*Zea mays sp.)* J Agric Food Chem 54:7992–7997. https://doi.org/10.1021/jf062256f

Iglesias C, Mayer J, Chavez L, Calle F (1997) Genetic potential and stability of carotene content in cassava roots. Euphytica 94:367–373. https://doi.org/10.1023/A:1002962108315

Lodhi MA, Ye G-N, Weeden NF, Reisch BI (1994) A simple and efficient method for DNA extraction from grapevine cultivars and Vitis species. Plant Mol Biol Report 12:6–13. https://doi.org/10.1007/BF02668658

Ly D, Hamblin M, Rabbi I et al (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

Marín Colorado JA, Ramírez H, Fregene M (2009) Genetic mapping and QTL analysis for carotenes in a S1 population of cassava. Acta Agron Univ Nac Colomb 58:15–21

Nisar N, Li L, Lu S et al (2015) Carotenoid metabolism in plants. Mol Plant 8:68–82. https://doi.org/10.1016/j.molp.2014.12.007

Njoku DN, Gracen VE, Offei SK, Asante IK, Danquah EY, Egesi CN, Okogbenin E (2014) Molecular marker analysis of F1 progenies and their parents for carotenoids inheritance in African cassava (Manihot esculenta Crantz). Afr J Biotechnol 13(40):3999-4007

Okechukwu RU, Dixon AGO (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. https://doi.org/10.1080/15427520802212506

Patnala R, Clements J, Batra J (2013) Candidate gene association studies: a comprehensive guide to useful in silico tools. BMC Genet 14:39. https://doi.org/10.1186/1471-2156-14-39

Pfeiffer WH, McClafferty B (2007) HarvestPlus: breeding crops for better nutrition. Crop Sci 47:S-88. https://doi.org/10.2135/cropsci2007.09.0020IPBS

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459–463. https://doi.org/10.1038/nrg2813

Rabbi I, Hamblin M, Gedil M et al (2014) Genetic mapping using genotyping-by-sequencing in the clonally propagated cassava. Crop Sci 54:1384–1396. https://doi.org/10.2135/cropsci2013.07.0482

Rodriguez-Amaya DB, Kimura M (2004) HarvestPlus handbook for carotenoid analysis. International Food Policy Research Institute (IFPRI), Washington, pp 58. http://www.harvestplus.org/sites/default/files/tech02.pdf

Ruiz-Sola MÁ, Rodríguez-Concepción M (2012) Carotenoid biosynthesis in Arabidopsis: a colorful pathway. Arabidopsis Book 10:e0158. https://doi.org/10.1199/tab.0158

Saltzman A, Birol E, Bouis HE et al (2013) Biofortification: progress toward a more nourishing future. Glob Food Sec 2:9–17. https://doi.org/10.1016/j.gfs.2012.12.003

Sánchez T, Ceballos H, Dufour D et al (2014) Prediction of carotenoids, cyanide and dry matter contents in fresh cassava root using NIRS and Hunter color techniques. Food Chem 151:444–451. https://doi.org/10.1016/j.foodchem.2013.11.081

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tonukari NJ (2004) Cassava and the future of starch. Electron J Biotechnol 7:12–15. https://doi.org/10.2225/vol7-issue1-fulltext-i02

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423. https://doi.org/10.3168/jds.2007-0980

von Grebmer K, Saltzman A, Birol E, Wiesmann D, Prasai N, Yin S, Yohannes Y, Menon P, Thompson J, Sonntag A (2014) 2014 Global Hunger Index: The Challenge of Hidden Hunger. Welthungerhilfe, International Food Policy Research Institute, and Concern Worldwide, Bonn, Washington, D.C., and Dublin, pp 50. http://dx.doi.org/10.2499/9780896299580

Welsch R, Arango J, Bär C et al (2010) Provitamin A accumulation in cassava (Manihot esculenta) roots driven by a single nucleotide polymorphism in a phytoene synthase gene. Plant Cell 22:3348–3356. https://doi.org/10.1105/tpc.110.077560

West KP, Mehra S (2010) Vitamin a intake and status in populations facing economic stress. J Nutr 140:201S–207S. https://doi.org/10.3945/jn.109.112730

Wolfe MD, Rabbi IY, Egesi C et al (2016) Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. Plant Genome 9:1–248. https://doi.org/10.3835/plantgenome2015.11.0118

Yan J, Kandianis CB, Harjes CE et al (2010) Rare genetic variation at Zea mays crtRB1 increases beta-carotene in maize grain. Nat Genet 42:322–327. https://doi.org/10.1038/ng.551

You FM, Huo N, Gu Y et al (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinformatics 9:253. https://doi.org/10.1186/1471-2105-9-253

Zhu M, Zhao S (2007) Candidate gene identification approach: progress and challenges. Int J Biol Sci 3:420–427