SCIENTIFIC REPORTS

OPEN

Epigenetic map and genetic map basis of complex traits in cassava population

Received July
Accepted December
Published January

Meiling Zou[1,2], Cheng Lu[2], Shengkui Zhang[2], Qing Chen[2], Xianglai Sun[2], Pingan Ma[2], Meizhen Hu[2], Ming Peng[2], Zilong Ma[2], Xin Chen[2], Xincheng Zhou[2], Haiyan Wang[2], Subin Feng[2], Kaixin Fang[2], Hairong Xie[2], Zaiyun Li[1], Kede Liu[1], Qiongyao Qin[2], Jinli Pei[2], Shujuan Wang[2], Kun Pan[2], Wenbin Hu[2], Binxiao Feng[2], Dayong Fan[3], Bin Zhou[3], Chunling Wu[3], Ming Su[3], Zhiqiang Xia[1,2], Kaimian Li[2] & Wenquan Wang[2]

Cassava (*Manihot esculenta* Crantz) is an important tropical starchy root crop that is adapted to drought but extremely cold sensitive. A cold-tolerant, high-quality, and robust supply of cassava is urgently needed. Here, we clarify genome-wide distribution and classification of CCGG hemi-methylation and full-methylation, and detected 77 much candidate QTLs^{epi} for cold stress and 103 much candidate QTLs^{epi} for storage root quality and yield in 186 cassava population, generated by crossing two non-inbred lines with female parent KU50 and male parent SC124 (KS population). We developed amplified-fragment single nucleotide polymorphism and methylation (AFSM) genetic map in this population. We also constructed the AFSM QTL map, identified 260 much candidate QTL genes for cold stress and 301 much candidate QTL genes for storage root quality and yield, based on the years greenhouse and field trials. This may accounted for a significant amount of the variation in the key traits controlling cold tolerance and the high quality and yield of cassava.

The majority of studies on many species are based on the complex traits[1–3]. Even though an increasing number of studies are investigating the heritable phenotypic variation in the model plants, such as *Arabidopsis thaliana*[2,4,5] and soybean[6], the heritable variations in cytosine methylation in non-model crops have not been investigated[7,8]. Cytosine methylation is a DNA base modification involved in the development, disease, and silencing of transposable elements and genes[9]. CG methylation is commonly found within gene bodies in plants[10]. Intraspecific surveys have revealed widespread variations in DNA methylation patterns within populations[11,12]. A key challenge in the field of population genetics is showing the changes in the genome-wide CCGG hemi-methylation and full-methylation heritable variation patterns and SNVs (single-nucleotide polymorphisms and indels) associated with heritable phenotypic variation in populations with high heterozygosity and large genomes.

To address these difficulties, we established in a cassava population with a highly heterozygous and large genome[13], and measured eight complex traits through field and greenhouse experiments. Cassava (*Manihot esculenta* Crantz), a starchy root crop, is a staple food and animal feed and serves as an important source of bioethanol[14]. As a tropical root crop, cassava is sensitive to cold. At temperatures less than 10 °C, cassava undergo chilling-injury or death, including delayed sprouting of the stem cutting, decreased yield, reduced leaf expansion and even leaf necrosis[15]. Thus, chilling injury is the most important factor limiting cassava's geographic distribution and productivity. The cold tolerance of cassava is very important for protection of the storage roots and propagation stems[15]. Cold tolerance, high quality, and a robust supply of cassava are urgently in demand. SC124 exhibits a high yield and cold tolerance but low proportion of starch. In contrast, KU50 has many elite economic traits, such as a high proportion of starch and the ability to produce high yields, but is intolerant of cold temperatures. In this study, we established a cold-tolerance cassava mapping population (KS) that descended from a cross between the two non-inbred lines, with KU50 as the female parent and SC124 as the male parent. We then used an amplified-fragment single nucleotide polymorphism and methylation approach (AFSM)[16] to concurrently identify whole genome hemi-methylation and full-methylation heritable variation pattern and SNVs in 186

Huazhong Agricultural University, Wuhan, China. The Institute of Tropical Biosciences and Biotechnology, Chinese Academy of Tropical Agriculture Sciences, Haikou, China. Guangxi Academy of Agricultural Sciences, Guilin, China. Correspondence and requests for materials should be addressed to Z.X. (email: xiazhiqiang@itbb.org.cn) or K.L. (email: likaimian@itbb.org.cn) or W.W. (email: wangwenquan@itbb.org.cn)

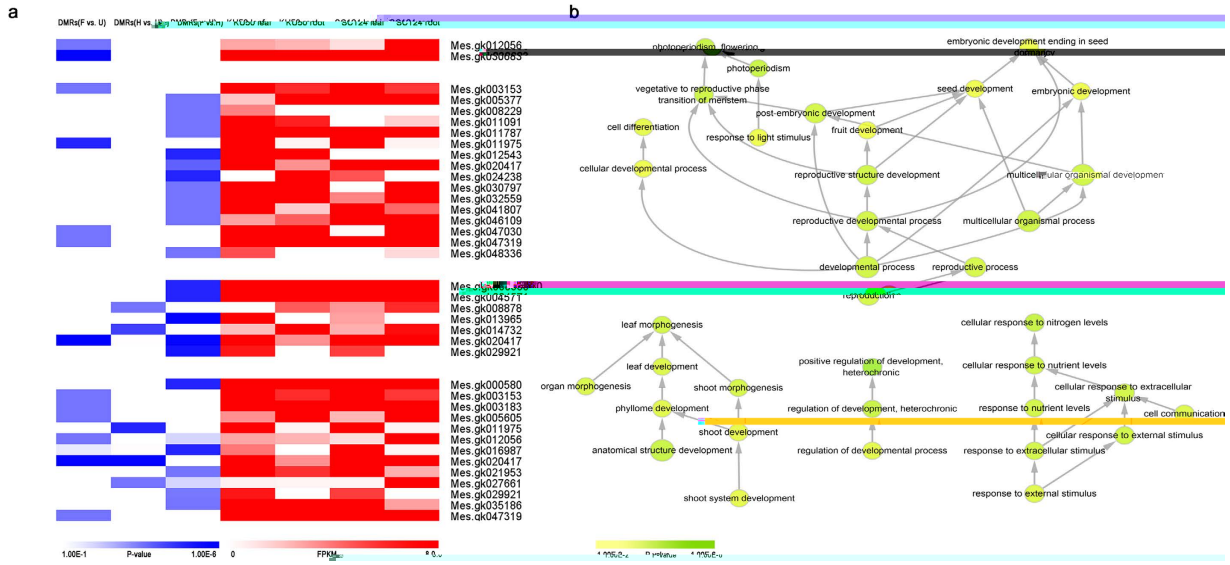

Figure 2. Differentially methylated and differentially expressed genes (DMR-DEGs). (**a**) A heatmap representation of genes with differential methylation and differentially expressed (DMR-DEGs) between the leaf and root in SC124 and KU50, based on a cuffdiff pairwise analysis (false discovery rate [FDR] < 0.001). The related genes were categorized; (**b**) Part of functional category enrichment was calculated using BiNGO[19] ($P < 0.001$, χ^2 text) for these DMR-DEGs relative to the KU50 genes[14].

we observed that polymorphic methylation, especially polymorphic fully methylation, tended to exhibit enrichment in gene bodies in population (Fig. 1b).

Effects of DNA methylation on gene expression. To assess the function of the differentially methylated regions (DMRs) in cassava, and to clarify the relationship between them and differentially expressed genes, we constructed RNA-seq libraries and performed deep sequencing on the transcriptome of the parents' roots and leaves belonging to the KS population as well as AFSM sequencing. A total of 130,707,837 high-quality (filtered) reads were generated (including 29,905,212 reads for the KU50 leaf, 34,029,215 reads for the KU50 root, 33,174,195 reads for the SC124 leaf, and 33,599,215 reads for the SC124 root). A total of 1,482 DMRs (SC124 leaf vs. KU50 leaf, SC124 root vs. KU50 root, leaf vs. root in KU50, and leaf vs. root in SC124) were detected in gene bodies and promoters (Fig. S4). Among these, 38 DMRs were considered significantly differentially expressed (DMR-DEGs) based on a cuffdiff pairwise analysis (false discovery rate (FDR) < 0.001) (Fig. 2). The functional category enrichment was then assessed using BiNGO[19] ($P < 0.001$, χ^2 text) for these DMR-DEGs relative to the KU50 genes[14]. These DMR-DEGs tended to exhibit enrichment of cellular components, including small nucleolar ribonucleoprotein complex and intracellular membrane-bounded organelle. Enriched molecular function for DMR-DEGs included translation factor activity, nucleic acid binding, ATPase activity. For biological processes, organ morphogenesis, cell differentiation, response to light stimulus and cellular response to extracellular stimulus (Table S4, Fig. 2b).

QTLs^{epi} mapping and relative genes involved in cold tolerance and yielding. We found 318 methylated QTL genes that were significantly associated with cold tolerance (CT-QTLGs^{epi}, including 11 repeatable CT-QTLGs^{epi}) and 524 that were significantly associated with quality and yield (QY-QTLGs^{epi}, including 105 repeatable QY-QTLGs^{epi}), based on the correlation analysis (x^2 test, $p < 0.01$, Fig. 3 and Table S5). Among these CT-QTLGs^{epi}, we observed *glycoside hydrolase family 2 β-mannosidase* (*GH family 2β-mannosidase*, Mes.gk016414) was significantly associated with CTIG, CTIF-4, LFIF-2 (Fig. S7a). Significantly correlated polymorphic methylation sites (SCPMs) in the promoter and coding region of *GH family 2β-mannosidase* were found, including these SCPMs in the CpG island of 3′ coding region. *Photosystem II protein D1* (Mes.gk020417) was observed significantly associated with LFIF-2 (Fig. S7b). Near the 5′ initiation codon area, several SCPMs were detected. Interesting, we also found SCPMs in photosystem II protein D1 CT-QTL^{epi} transcription factor binding sites in DNA sequence (reCT-QTL^{epi}TFBSs). This gene has previously been shown to a general adaptive response to environmental extremes[20]. Photosystem II D1 protein degradation speed will be greater than the synthesis under various stress conditions, resulting the damage of photosystem II reaction center. Under cold stress, decreased membrane fluidity affected the diffusion rate of the damaged Photosystem II D1 protein, hindering the insert of newly synthesized Photosystem II D1 protein. Plant photosynthetic activity decreased obviously, as photosystem II D1 protein damage-repair turnover was interferenced by low temperature stress[21]. The SC124 *Photosystem II protein D1* genes expressions were higher than KU50 in the leaves and roots, which in accordance with their cold tolerance. Two *peroxidase superfamily protein* (*POD*, Mes.gk031011 and Mes.gk045763) were significantly associated with RIF-2 and RIG, respectively (Fig S7c and d). We found SCPMs in the coding region in both genes. Previous study showed these *POD* gene family members in response to environmental stimulus, including low-temperature stimulus[22,23]. In accordance with previous studies showed that *mitogen-activated*

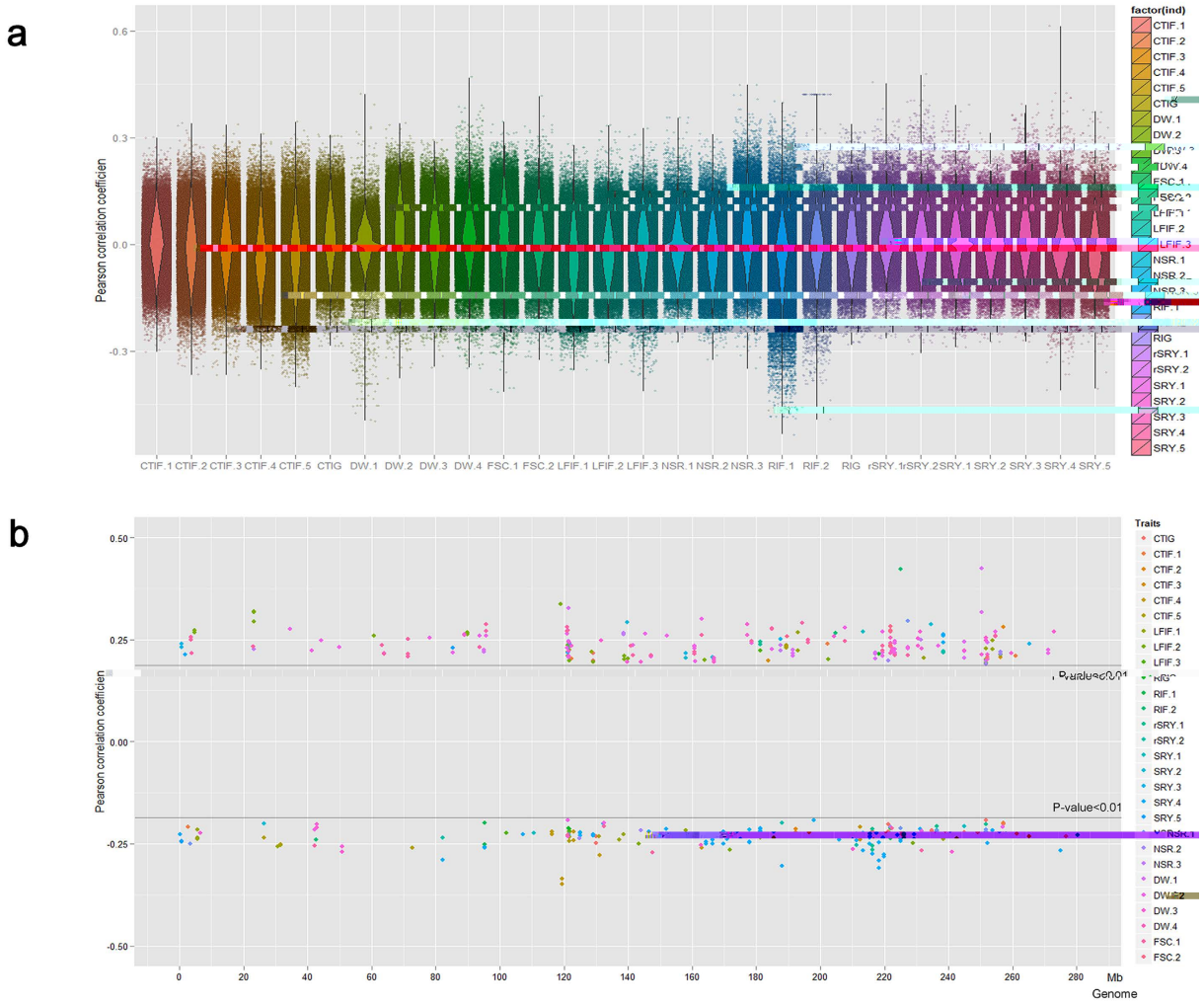

Figure 3. QTL[e] a g e KS a . (**a**) As a violin plot shows the distribution of all the Pearson correlation coefficient for epigenetic QTLs with cold tolerance (CT-QTLs[epi]) and quality and yield (QY-QTLs[epi]); (**b**) Distribution of epigenetic QTLs significantly associated with cold tolerance (CT-QTLs[epi]) and quality and yield (QY-QTLs[epi]). The y-axis indicates the correlation coefficient (p-value < 0.01, *t* test) and x-axis indicates the length of the genome contigs (Mb).

protein kinase kinase 4 (*MKK4*) were activated and rising its mRNA levels by cold stress in a mitogen-activated protein kinase pathway, suggested it could be a plant stress signaling[24], we found *MKK4* (Mes.gk046937) was significantly associated with rSRY-2 (Fig. S7e). SCPMs were detected in *MKK4* coding region.

Among these QY-QTLGs[epi], *phosphoserine aminotransferase* (*PSAT*, Mes.gk000126) was significantly associated with SRY-4, DW-2 and FSC-1 (Fig. S7f). SCPMs were detected in *PSAT* coding region. In previous studies observed serine took part in senescence, protein degradation, and Inhibit the growth of plants[25,26]. In accordance with previous study we found the *PSAT* gene expression level in the roots and leaves of KU50 were lower than that in SC124. *Snrnp auxiliary factor, small subunit* (Mes.gk030683) was significantly associated with DW-2, DW-3, FSC-1 (Fig. S7g), and *Zinc finger* (*RING-H2-type finger*, Mes.gk024887) was significantly associated with DW-2, FSC-1 (Fig. S7h). Besides, we found *translocon at the outer envelope membrane of chloroplasts 33* (*TOC33*, Mes.gk016669), *chaperonin 20* (*CPN20*, Mes.gk033730), *NADH-dependent glutamate synthase 1* (*GLT1*, Mes.gk036403) were significantly associated with DW-2, DW-3 (Fig. S7i,j and k). SCPMs were detected in *PSAT*, *Snrnp auxiliary factor*, *RING-H2-type finger*, *CPN20*, *GLT1* gene coding regions, while were detected in *TOC33* intron region. Besides, we found SCPMs in *Ribulose-1,5 bisphosphate carboxylase/oxygenase large subunit N-methyltransferase* (Mes.gk014095) and *RNA methyltransferase* (Mes.gk015292) QY-QTL[epi]TFBSs, and their adjacent genes expressions were both higher in SC124 than KU50. These data may reflect a metastable phenomenon in the heritable hemi-methylation and full-methylation patterns in cassava.

High dense AFSM linkage group maps based on KS population. In addition to methylation, 573,557 single nucleotide variants (SNVs; SNP and indel markers) were identified using this approach. Among the 573,557 SNVs, 10,627 were distributed in the CDS region, 22,439 in the gene region, and 8,709 in the promoter region. A high density KS genetic map was produced for the KS cassava mapping population using JoinMap version 4.1 to

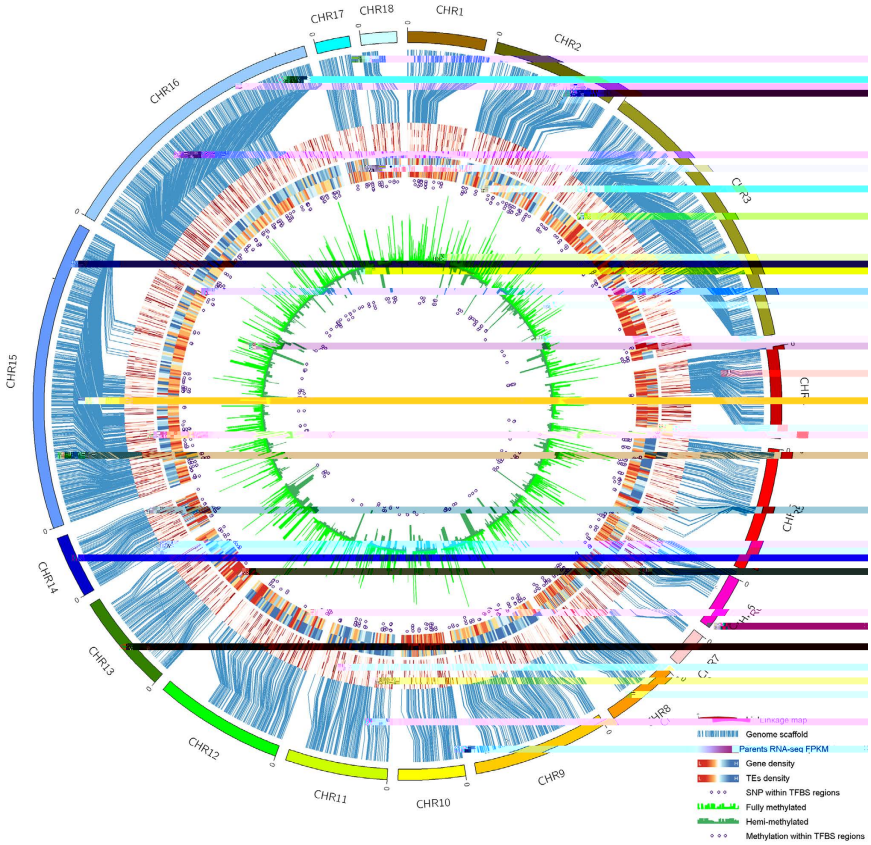

Figure 4. A high-density KS genetic linkage map. Outside cylindrical, KS genetic linkage groups; Blue lines, synteny between the scaffold and genetic linkage groups; Red lines, expressed genes from outside to inside, are root of SC124, leaf of SC124, root of KU50, and leaf of KU50; Blue and red lines, gene and TEs density (the outside for gene, the inside for TEs); Purple circles in the middle, SNPs within predict transcription factor binding sites in DNA sequence (TFBSs); Light green, Fully-methylated sites; Dark green, hemi-methylated sites; Purple circles in the center, Methylated sites within TFBSs.

estimate the linkage, map order, and distance (Fig. 4 and Table S8). This study is attempt to develop the biggest cassava genetic linkage map using only AFSM markers. Four thousand six hundred and forty-eight out of 69,141 filtered AFSM markers (consisting of 4,437 SNVs and 211 methylation polymorphic markers) were assigned to 18 linkage groups (LGs). A total of 2,605 of the markers were located within gene regions. The identification of 18 LGs is consistent with cytological studies reporting the diploid number of chromosomes in cassava[27]. The lengths of the LGs varied from 38.39 (LG 1) to 235.26 cM (LG 15), and the KS genetic map spanned a total of 2190.34 cM. A total of 4,648 AFSM linkage group markers were associated with 2,734 KU50 scaffolds. Then, 38,380 AFSM makers were mapped to these scaffolds. Using the single copies, the KU50, AM560 and W14 genomes were associated with each other. We combined 461 AM560 scaffolds and 933 W14 scaffolds into 18 linkage groups using 1,634 and 1,408 AFSM linkage group markers, respectively (Fig. S8 and Table S8).

Fine mapping of QTLs and relative genes for cold tolerance, yield and other economic traits. Our linkage mapping obtained 574 repeatable cold-tolerance QTLs and 499 repeatable quality and yield QTLs (Fig. 5 and Table S8). To explore the biological functions of these QTLs, we entered them into the MapMan software for pathway analysis. We detected 260 genes with repeatable cold-tolerance QTLs (reCT-QTLGs) and 301 genes with repeatable quality and yield QTLs (reQY-QTLGs) in these pathways (Tables S9 and S10). Phytohormones play important roles in the adjustment to adapt to the environmental stresses[22,28,29]. Consistent with these findings, ABA biosynthesis reCT-QTLG (*nine-cis-epoxycarotenoid dioxygenase 3*, *NCED3*), jasmonate biosynthesis reCT-QTLG (*OPDA reductase 3*, *OPR3*) and ethylene biosynthesis reCT-QTLGs (1-aminocyclopropane-1-carboxylate synthase 7 and ethylene signal DNA binding/transcription factor) were detected in the hormone synthesis and metabolism pathway (Fig. S9a). *Calmodulin-dependent protein kinase* (*CDPK*), *CBL4*, *CaLB* and *IQ-domain 10* (*IQD10*) reCT-QTLGs were found in the calcium/calmodulin-mediated signaling network (Fig. S9a), indicating that these QTLs might be related to the cold tolerance of cassava. These findings contrast with the previously reported result that calcium/calmodulin-mediated genes play important roles in cold stress response for plant[30,31]. In a general adaptive response to cold stress, plants always grow slow or stop growing. In this study, we found SNVs within *auxin response factor 2* (ARF 2) (Mes.gk011443) reCT-QTLTFBS, and its adjacent gene expression in SC124 is significantly lower than that in KU50. This may be

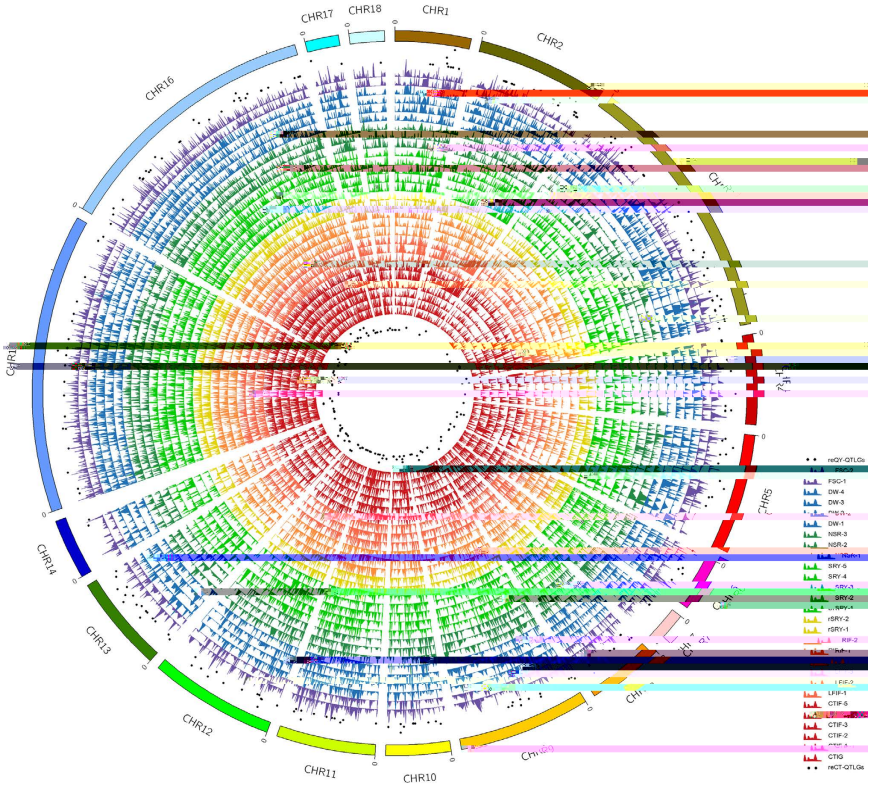

Figure 5. QTL landscape of repeatable cold-tolerance (CTRT) and repeatable yield (YQRT) KS plants. QTL mapping profiles for four independent CTRTs measurements [Red: Cold-tolerance index in greenhouse (CTIG) and Cold-tolerance index in field (CTIF) 1–5; Dark orange: Leaf fall index (LFIF) 1–3; Light orange: Recovery index in greenhouse (RIG) and Recovery index in field (RIF) 1–2; Yellow: Relative storage root yield (rSRY) 1–2], as well as four independent YQRTs measurements [Light green: Storage root yield (SRY) 1–5; Dark green: Number of storage root (NSR) 1–3; Blue: Storage root dry weigh (DW) 1–4; Purple: Fresh root starch content (FSC) 1–2].

one of strategies for SC124 to be more cold resistant than KU50, by slowing or stopping growth during chilling stress. WD-40 repeat family protein is associated with biomass accumulation. We found SNVs within *WD-40 repeat family protein* (Mes.gk025413) reQY-QTLTFBS, and its adjacent gene expression in SC124 is significantly higher than in KU50.

In addition, we identified 25 repeatable cold-tolerance QTL transcription factors (reCT-QTLTFs, Fig. S9b), out of 46 CT-QTL transcription factors (CT-QTLTFs). *WRKY* genes showed strong and rapid induction in cold stress response in previous studies[32,33]. Consistent with these findings, we identified WRKY reCT-QTLTFs (*WRKY31* and *WRKY9*). *Myb domain protein 55* (*MYB55*), *MYB70*, *MYB40*, *helix–loop–helix* (*bHLH*), *auxin response factor 2* (*ARF 2*) and *zinc finger protein 6* (*ZFP6*) reCT-QTLTFs were also detected in our study (Fig. S9b). Moreover, we identified *sucrose phosphate synthase 2F* (*SPS2F*), *sucrose synthase 6* (*Susy6*), *neutral invertases*, *fructokinase*, and *beta-amylase* (*BAM*) reQY-QTLTFs in the starch and sucrose metabolism pathway (Fig. S9c,d and e). These QTLs might play important roles in improving the cold tolerance, high quality and yield of cassava.

Conclusion

In summary, we combined novel genome-wide CCGG hemi-methylation and full-methylation analysis and the SNP discovery sequencing approach AFSM, the transcriptome deep sequencing method RNA-seq and years of independent field and greenhouse trials in the cassava KS population and the parents to enable the high-resolution genome-wide characterization of a map with CCGG DNA methylation sites, Gene and TEs density, SNPs, repeatable QTLs, QTLs^{epi} and related gene expression profiles. In addition to a large number of SNVs, we also identified thousands of hemi-methylated and fully methylated sites using the same genotypes. We described the distribution of these methylated sites in the cassava genome and grouped them into twelve classes based on the hemi-methylated or fully methylated site heritability in the KS population. Furthermore, we provide the description of the DMR-DEGs in cassava.

The complete set of whole-genome CCGG DNA methylation and gene expression data can be downloaded from the NCBI Sequence Read Archive (SRA) (SRX1674579 and SRX53531 for AFSM data, SRR2361999, SRR2404206, SRR2495947 and SRR2496326 for RNA-seq data). The related software including AFSM Perl scripts can be downloaded at (http://afsm.strikingly.com/). We build a new SNP and methylation genome browse website for cassava at (http://192.64.83.141/JBrowse-1.11.5/?data=test). We showed 77 much candidate QTLs^{epi} for cold stress and 103 much candidate QTLs^{epi} for storage root quality and yield (Tables S6 and S7). Besides, 260

much candidate QTL genes for cold stress and 301 much candidate QTL genes for storage root quality and yield were also presented at Tables S9 and S10. These new tools and genome-wide resources should serve as the molecular basis for future cassava marker-assisted breeding programs and highlight the discovered cold-tolerance and high-quality $QTLs^{epi}$ and QTLs that may reduce timelines. Finally, the whole genome approaches developed here should be useful for future studies with many other organisms with large complex genomes and complex traits.

Methods

KS mapping population development. In this study, the KS mapping population of 186 progenies was generated by crossing two non-inbred lines with differentially cold-resistant. The female parent, KU50, possess many elite economic traits with high rate of starch and high-yielding, and is sensitive to chilling. The male parent, SC124 yielded high and is tolerant of chilling. Seeds were disinfected using sodium hypochloride and stratified in sterile water. Then the seeds were germinated in sterilized garden soil and transplanted two month after sowing. At maturity, the stem cutting were planted at Haikou, Hainan Provience (HK09), Wenchang, Hainan Provience (WC10, WC11, WC12, WC13), Guiling, Guangxi Provience (GL11, GL12), Hezhou, Guangxi Provience (HZ13) from 2009–2013.

Cold-tolerance index: cold-tolerance index in greenhouse (CTIG) and cold-tolerance index in field (CTIF). For each seedling line, eight seedlings were grown on 15X 15 cm pots containing sterilized garden soil. After two months, the strong and uniform seedlings were selected for chilling injury treatment in a phytotron at photon flux density of 800 μmol m^{-2} s^{-1} PAR, 60–80% relative humidity. The 20 °C/18 °C day-night (12 h/12 h) treatment was carried out for 24 h. After 15 °C/12 °C day-night (12 h/12 h) chilling treatment was performed for 24 h, then 6 °C/4 °C day-night cycle (12 h/12 h) chilling treatment was carried out by exposure to cool air for 5 d. The temperature was returned to 28 °C after the chilling treatment and the plants was allowed to recover for 24 h. The CTIG was calculated as follows:

$$\mathrm{CTIG} = \frac{\sum(n_i X\ \mathrm{CTLG}_i)}{n} \quad (1)$$

ie9.1(e n)15(l)-4In this equation, $CTLG_i$ was the cold-tolerance level in greenhouse (0–4 levels), n_i was the number 5.8(h0 0 9 253.2751

In this equation, RLF_i was the recovery level in field (0–4 levels), n_i was the number of plants with the same recovery level.

Relative storage root yield (rSRY). Storage root samples were harvested between February and March (11–12 months old plants). Two rSRY measurements (rSRY1–2) were performed. Measurement rSRY1 was root tuber yield from Guilin field relative to the root tuber yield from Wenchang field in 2012, and measurement rSRY2 was root tuber yield from Hezhou field relative to the root tuber yield from Wenchang field in 2013.

Storage root yield (SRY). We performed five SRY measurements (SRY1–5). Measurements SRY1, SRY2 and SRY5 come from Wenchang field experiments in 2010, 2012 and 2013, SRY3 measurement from a Guilin field experiment in 2012 and, SRY4 measurement from a Hezhou field experiment in 2013.

Number of storage root: (NSR). Three measurements of NSR1–3 were performed. Measurements NSR1 and NSR3 come from Wenchang field experiments in 2010 and 2013, and NSR2 measurement from Hezhou field experiment in 2013.

Storage root dry weigh (DW). DW was determined using the method of Benesi *et al.*[34]. Six undamaged roots were randomly selected. The medial sections of selected fresh roots were shredded into thin slices, mixed thoroughly and duplicate of 200 g (w1) were oven dried at 65 °C for 72 h. Dry matter content was weighed immediately (w2). Four measurements (DW1–4) were performed for four years (2010, 2011, 2012 and 2013). The percentage of storage root dry weigh (DW%) was calculated via the following equation:

$$\mathrm{DW}\,\% = \frac{w2}{w1} \times 100\% \tag{6}$$

This was done within 12 h after harvest to avoid post harvest changes through physiological deterioration or moisture loss of the root.

Fresh root starch content (FSC). FSC measurements (FSC1–2) were done in Wenchang over 2 years (2011 and 2013). Starch content was analyzed using a Total Starch Assay kit (Megazyme International, Wicklow, Ireland), and spectrophotometric readings were conducted using a Spectronic 1201 spectrophotometer (Milton Roy Company, Ivyland, PA, USA) using glucose as sugar control and maize starch as the starch control. The starch content of the storage roots was first calculated as a percentage per dry weight basis, and later converted to a percentage per fresh weight basis for analysis.

Principal coordinate analysis. A principal coordinate analysis (PCoA) was constructed based on the binary character matrix using the princomp in R[35]. The PCoA plot revealed the structeristics of a binary data matrix of complex traits in the entire population. All complex traits were approximately scatted in two areas (Fig. S6): The cold-tolerance related traits (CTRTs, right blue circle) and the yield and quality related traits (YQRTs, left green circle).

AFSM library construction and sequencing. DNA from the fully expanded leaves of 5-month-old crops was extracted using Plant DNeasy Maxi Kit (QIAGEN, Valencia, CA), with two biological replicates were pooled for each individual. AFSM libraries was constructed using AFSM method[16], which can concurrently identify whole genome SNPs, indels, fully-methylation and hemi-methylation sites for 186 samples from KS population. AFSM libraries were sequenced using Illumina HiSeq2500 with Pair-end 150 bp lengths.

RNA-seq library construction and sequencing. Total RNA was extracted from the fully expanded leaves and storage roots of the parents of KS population at the same developmental stage (120 days), using RNA plant reagent kit (Tiangen Company). The RNA-seq libraries were performed according to Illumina manufacturer's instructions (Illumina). Then, using Illumina HiSeq2500, 51 bp sequencings were performed.

Alignments of Illumina AFSM reads and data analysis. The raw Illumina AFSM sequence reads were processed using custom Perl scripts[16] and then aligned to the MK_v1 cassava genome using Bowtie2[17], allowing one mismatch. SNPs were identified using the SAMtools and VCFtools_v0.1.9 (http://vcftools.sourceforge.net/). Besides, custom algorithms were used for methylation analyses as decribed by previously[16].

Methylation analyses. Analyses of the AFSM methylation results were based on comparisons of the EcoRI-HpaII- and EcoRI-MspI-assembled sequences with methylated cytosines at the 5′-CCGG sites using custom Perl scripts (http://afsmseq.sourceforge.net/) for individual plants, described as Xia *et al.*[16]. In this study, a CCGG methylated site was definited as that present at more than 4 reads (at least 2 HpaII-reads and 2 MspI-reads). For each individual assembled sequence, it was first determined whether those with CCGG sites were: (1) present only in the HpaII cleavage sites of the EcoRI-HpaII products and the body sequences of the EcoRI-MspI products but not in the MspI-cleaved sites; (2) present only in the MspI cleavage sites of the EcoRI-MspI products and the body sequences of the EcoRI-HpaII products but not in the HpaII cleavage sites; (3) present in the body sequences of both the EcoRI-HpaII and EcoRI-MspI products but not in the HpaII or MspI cleavage sites. Condition (1) denotes a hemi-mCCGG methylated state, and conditions (2) and (3) correspond to fully CmCGG and fully mCCGG methylated states. Only methylated sites that were similarly methylated at least in two samples were remained in this study.

Methylated density was computed as described in Weiss[36] and each region was split into 80 equal windows, with the average alignment depth calculated for each window. Methylated genes were determined only when a

methylated site similarly methylated at least in two samples in cassava gene body or promoter (2 kb upstream) regions.

We refer to the level of methylation of genomic regions in this study. To compute this level within a bin, we summed the number of methylated CCGG (the number of methylated CCGG sites multiplied by the number of reads from methylated fragments within a bin), and divided the summed number of sequenced bases covering all CCGG (the number of CCGG sites multiplied by the number of reads from all fragments within a bin).

DMRs were identified using our pipeline. Each methylation was scanned genome-wide requiring at least 5 methylated CCGG differences within a given window. DMRs were identified by comparison of the leaf and root in SC124 and KU50 methylations(SC124 leaf vs. KU50 leaf, SC124 root vs. KU50 root, leaf vs. root in KU50 and leaf vs. root in SC124). A DMR was identified if the P-value from Chi-squared test was ≤ 0.05.

Alignments of Illumina RNA-seq reads and data analysis. Adapters were removed from raw Illumina RNA-seq sequence reads using FASTX-toolkit pipeline, version 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/). Sequence quality was examined using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were mapped to the cassava genome (MK_v1)[14] using Tophat v. 2.0.10[37]. Gene level expression and differential expression analysis were performed using the program Cufflinks version 0.8.0[38] with default settings. To test DE with unambiguous mapping data DEGseq[39] was used.

Functional annotation.

Analysis of parentally-derived QTL. We used the framework maps to search for QTLs on parental and consensus maps. First, the nonparametric Kruskal–Wallis (KW) rank-sum test was carried out at each marker on raw phenotypic data with MapQTL 5.0[46] for muscat score and each raw monoterpene content (with a first type error rate of a = 0.001 for each individual test). We used MapQTL for KW tests at each marker, with a = 0.001 for individual tests, for parental and consensus maps. Then KW tests, Simple Interval Mapping (SIM) and Composite Interval Mapping (CIM) were performed on ln-transformed monoterpene contents. Composite interval mapping (CIM) was performed to identify QTL controlling LL using the MQM method with the program MapQTL 5.0 and the appropriate cofactor selection. The permutation test[47] was performed with 1000 runs to determine the P = 0.05 genomewide significance level for declaring QTL for LL significant, according to the complementarities of these softwares.

References

1. Schmitz, Robert J. The secret garden-Epigenetic alleles underlie complex Traits. *Science* **343**, 1082–1083 (2014).
2. Johannes, Frank *et al.* Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet*. **5**, e1000530, doi: 10.1371/journal.pgen.1000530 (2009).
3. Roux, Fabrice *et al.* Genome-wide epigenetic perturbation jump-starts patterns of heritable variation found in nature. *Genetics* **188**, 1015–1017 (2011).
4. Cortijo, Sandra *et al.* Mapping the epigenetic basis of complex traits. *Science* **343**, 1145–1148 (2014).
5. Becker, Claude *et al.* Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature* **480**, 245–249 (2011).
6. Schmitz, Robert J., Stacey & Ecker, J. R. *et al.* Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res*. **23**(**10**), 1663–74 (2013).
7. Jin, Huajun *et al.* Alterations in cytosine methylation and species-specific transcription induced by interspecific hybridization between Oryza sativa and O. officinalis. *Theor. Appl. Genet*. **117**, 1271–1279 (2008).
8. Zhao, Y., Yu, S., Xing, C., Fan, S. & Song, M. DNA methylation in cotton hybrids and their parents. *Mol. Biol*. **42**, 195–205 (2008).
9. Lister, Ryan *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523–536 (2008).
10. Schmitz, Robert J. *et al.* Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334**, 369–373 (2011).
11. Schmitz, Robert J. *et al.* Patterns of population epigenomic diversity. *Nature* **495**, 193–198 (2013).
12. Reinders, Jon *et al.* Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. *Genes Dev*. **23**, 939–950 (2009).
13. Chen, X., Xia, Z., Fu, Y., Lu, C. & Wang, W. Constructing a Genetic Linkage Map Using an F1 Population of Non-inbred Parents in Cassava (Manihot esculenta Crantz). *Plant Mol. Biol. Rep*. **28**, 676–683 (2010).
14. Wang, W. *et al.* Cassava genome from a wild ancestor to cultivated varieties. *Nature Commun*

39. Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 136–138 (2010).
40. Kanehisa, Minoru *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**, 354–357 (2006).
41. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, 182–185 (2007).
42. Thimm, Oliver *et al.* MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **37**, 914–939 (2004).
43. Mulder, Nicola J. *et al.* New developments in the InterPro database. *Nucleic Acids Res.* **35**, 224–228 (2007).
44. Van Ooijen, J. W. Multipoint maximum likelihood mapping in a full- sib family of an outbreeding species. *Genetics Res.* **93**, 343–349 (2011).
45. Krzywinski, Martin *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
46. Van Ooijen, J. W. *MapQTL 5, Software for the mapping of quantitative trait loci in experimental populations*. (Kyazma B. V., Wageningen, Netherlands 2004).
47. Churchill, G. A. & Doerge, R. W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971 (1994).

Acknowledgements

This work was supported by National Science Foundations of China (31301102, 31261140363, 31171230), National Nonprofit Institute Research Grant of CATAS-ITBB (ITBB2015ZY10), A new type of molecular breeding technology based on the second generation sequencing (P153020016) and Chinese Agriculture Research System (CARS-12-HNwwq). Thanks to M.C. Luo of UC Davis, USA, who provided constructive suggestions. Thanks to Dr Nicole D. of American Journal Experts for editing this paper.

Author Contributions

Z.X. and W.W. conceived and designed the studies. Z.X., K.L. and M.Z. analyzed the data as a whole. M.Z. wrote the paper. Z.X. developed the original protocol for AFSM library preparation, and M.Z. developed modifications of the protocol. Z.X. created bioinformatics scripts and conducted sequence analysis and functional annotation. B.F. created part of the bioinformatics scripts. M.Z. and Z.X. constructed the cassava genetic linkage map, comparative map, QTL map and QTL^{epi} map. M.Z., Z.X. and S.Z. performed DNA preparation and prepared libraries for Illumina sequencing. M.Z. and Z.X. performed transcriptome analyses. C.L. developed the cassava KS mapping population. C.L., X.S., M.Z., Q.C., S.Z., P.M., M.H., M.P., Z.M., X.C., X.Z., H.W., S.F., K.F., H.X., Z.L., K.L., Q.Q., J.P., S.W., K. P., W.H., D.F., B.Z., C.W. and M.S. planted and performed survey of years geenhouse and field trials from 2009–2013. W.W. supervised the whole study.

Additional Information

Acce c de : Sequence Read Archive (SRA) database SRX1674579, SRX53531, SRR2361999, SRR2404206, SRR2495947 and SRR2496326. SNP and methylation genome browse website for cassava Browse website: http://192.64.83.141/JBrowse-1.11.5/?data=test.

S e e a f a accompanies this paper at http://www.nature.com/srep

C e g a c a e e : The authors declare no competing financial interests.

H c e h a c e: Zou, M. *et al.* Epigenetic map and genetic map basis of complex traits in cassava population. *Sci. Rep.* **7**, 41232; doi: 10.1038/srep41232 (2017).

P b he e: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.