Cassava genome from a wild ancestor to cultivated varieties

Wenquan Wang, Binxiao Feng, Jingfa Xiao, Zhiqiang Xia, Xincheng Zhou, Pinghua Li, Weixiong Zhang, Ying Wang, Birger Lindberg Møller, Peng Zhang, Ming-Cheng Luo, Gong Xiao, Jingxing Liu, Jun Yang, Songbi Chen, Pablo D. Rabinowicz, Xin Chen, Hong-Bin Zhang, Hanen Ceballos, Qunfeng Lou, Meiling Zou, Luiz J.C.B. Carvalho, Changying Zeng, Jing Xia, Shixiang Sun, Yuhua Fu, Haiyan Wang, Cheng Lu, Mengbin Ruan, Shuigeng Zhou, Hui Liu, Rubini Maya Kannangara, Kirsten Jørgensen, Rebecca Louise Neale, Maya Bonde, Nanna Heinz, Wenli Zhu, Shujuan Wang, Yang Zhang, Kun Pan, Ping-An Ma, Zhengxu Li, Meizhen Hu, Wenbin Liao, Wenbin Hu, Shengkui Zhang, Jinli Pei, Anping Guo, Jianchun Guo, Jiaming Zhang, Zhengwen Zhang, Jianqi Ye, Wenjun Ou, Yaqin Ma, Xinyue Liu, Luke J. Tallon, Kevin Galens, Sandra Ott, Jie Huang, Jingjing Xue, Feifei An, Qingyun Yao, Xiaojing Liu, Martin Fregene, L Augusto Becerra López-Lavalle, Jiajie Wu, Frank M. You, Meili Chen, Songnian Hu, Guojian Wu, Silin Zhong, Peng Ling, Yeyuan Chen, Qinghuang Wang, Guodao Liu, Bin Liu, Kaimian Li & Ming Peng

Cassava is a major tropical food crop in the Euphorbiaceae family that has high carbohydrate production potential and adaptability to diverse environments. Here we present the draft genome sequences of a wild ancestor and a domesticated variety of cassava and comparative analyses with a partial inbred line. We identify 1,584 and 1,678 gene models specific to the wild and domesticated varieties, respectively, and discover high heterozygosity and millions of single-nucleotide variations. Our analyses reveal that genes involved in photosynthesis, starch accumulation and abiotic stresses have been positively selected, whereas those involved in cell wall biosynthesis and secondary metabolism, including cyanogenic glucoside formation, have been negatively selected in the cultivated varieties, reflecting the result of natural selection and domestication. Differences in microRNA genes and retrotransposon regulation could partly explain an increased carbon flux towards starch accumulation and reduced cyanogenic glucoside accumulation in domesticated cassava. These results may contribute to genetic improvement of cassava through better understanding of its biology.
Cultivated cassava, *Manihot esculenta* Crantz, originated from its wild progenitor, *Manihot esculenta* ssp. *Flabellifolia*, in tropical lowlands along the southern rim of the Amazon basin, where sunlight, heat and rainfall are plentiful, and intervals of drought are common1–3. Domestication of cassava occurred during the period of approximately 12,000 to 7,000 years ago by indigenous South Americans, as supported by DNA sequence analysis of a single locus4, and by archaeological and fossil records5–7, resulting in the modern cassava cultivars with extraordinary characteristics including high biomass and high starch yield in near optimum environments, and tolerance to drought and barren soil. Cassava is the most important root crop worldwide and provides staple food for over 700 million people in Africa (51%), Asia (29%) and South America (20%; http://faostat.fao.org). As it is highly tolerant to drought and its storage roots can be preserved in soil for a few years, cassava is considered to be an important reserve of carbohydrates to relieve global famine8. It is also an ideal feedstock crop for bioenergy, biomaterials and animal feeds because of its favourable agricultural characteristics and high-quality and -quantity starch9,10.

The cassava genome (2n = 36) is highly heterozygous because of its outcrossing nature and broad tropical distribution10,11. Conventional breeding and marker-assisted selection12–14 have so far proved ineffective in achieving its potential regarding desirable traits, such as high-quality starch, storage root yield, avoidance to postharvest biological deterioration and resistance to diseases. For instance, cassava storage root yield is approximately 13.6 t ha⁻¹ globally, which is two- to fourfold below its potential productivity. The lack of a reference genome sequence and other genomic and transcriptomic resources has limited progress in basic biological research and breeding in cassava. Therefore, the draft genome sequence of a partial inbred cassava line, AM560, has been generated and publicly released relatively recently15 (http://www.phytozome.net/cassava.php/). The sequence integrated 26- and 0.9-fold coverage of Roche 454 and Sanger reads, resulting in 530-Mb assembled scaffolds (including 410-Mb of contigs with no gaps), that cover approximately 70% of the cassava genome.

In the present study, we sequence the genomes of two cassava genotypes: W14 (*Manihot esculenta* ssp. *flabellifolia*), a wild subspecies that shows low storage root yield and low root starch content; and KU50, a variety commonly cultivated in Southeast Asia that has six to eight times higher storage root yield potential and five to six times higher starch content than W14 as described in Supplementary Information. Using an integrated assembly strategy combining shotgun Illumina and Roche 454 reads, and a bacterial artificial chromosome (BAC)-based physical map with BAC-end Sanger sequences, we generate a high-quality draft genome sequence of cassava using established protocol16–20. In addition, the genome of a self-pollinated clone (s1.600) derived from the sugary cassava landrace CAS36 (ref. 21) is sequenced to 20-fold coverage, and the transcriptomes of W14, KU50 and another cultivated variety Arg7 are profiled during plant ontogeny. A comparative analysis of the three genome sequences and annotated transcriptomes enables us to better understanding genomic features underlying the evolution and domestication of cassava22–24, particularly in relation to carbon flux, starch synthesis and biosynthesis of cyanoogenic compounds. These genomic resources and findings provide a platform for advancing basic biology research, gene discovery and genomic selection-assisted breeding in cassava25,26.

Results

**Genome assembly and annotation.** The genomes of cassava lines, W14 and KU50 (Supplementary Fig. 1, Supplementary Table 1, Supplementary Note 1) were sequenced and de novo assembled using next-generation sequencing technologies and hybrid assembly approaches23 (Supplementary Note 5). For W14, a 432-Mb assembly with an N50 of 43 kb was obtained. The assembly spanned 58.2% of the 742-Mb cassava genome, estimated by the kmer-spectrum (Supplementary Note 3, Supplementary Figure 2), and 92% of the sequence were non-gapped contigs. For KU50, the assembly spanned 495 Mb representing 66.7% of the cassava genome. The N50 was 19 kb and 81% sequences were non-gapped contigs (Supplementary Note 5–8, Supplementary Figs 3–5, Supplementary Tables 3 and 6). The quality of the draft genome sequence of W14 was evaluated by aligning the Roche 454 sequences of five BAC clones (Supplementary Fig. 6). In this way, we determined that the average error rate between the BAC and the draft genome sequence was less than 0.61% (Supplementary Table 5). The detected errors were single-nucleotide mismatches and insertions/deletions. These results show that the W14 draft genome sequence assembly is of high quality in spite of its high heterozygosity (Supplementary Note 11; Supplementary Table 9) and complexity. The genome coverage and quality of these assemblies were further confirmed by 201,392 available transcript sequences of W14 and KU50. Specifically, 94.9% and 92.8% of the transcripts could be aligned to the genome assemblies of W14 and KU50, respectively (Supplementary Figs 7, 8 and 9a,b). Using transcriptomics data and the *ab initio* gene prediction, 34,483 and 38,845 genes were predicted in the W14 and KU50 genomes, respectively. Comparison to protein databases, predicted 33,310 (96.6%) protein-coding genes in W14 and 37,592 (96.8%) in KU50 (using E-value cutoff of 10⁻⁵) and tentative functions could be assigned (Supplementary Note 9 and 10, Supplementary Figs 10, 12, Supplementary Table 7).

The genome sequence assembly was searched for repetitive DNA using de novo approaches that identified 36.9% and 25.7% of the W14 and KU50 genomes as repetitive sequences, respectively. The majority of the repetitive elements were long interspersed nuclear elements and long-terminal repeat elements (LTRs, Supplementary Table 8). These results, in addition to the fact that around 35% of the genome could not be assembled, suggest that the cassava genome is highly heterochromatic. This was confirmed by chromosome *in situ* hybridization using an LTR probe (Fig. 1b).

The level of heterozygosity in cassava is among the highest found in sequenced plant genomes, as determined by the frequency of single-nucleotide variations (SNVs) and insertions and deletions (InDels) in its genome. We identified 3.8 and 3.4 SNVs per kilo-base (kb) in the W14 and the KU50 genomes, respectively (Supplementary Table 9), which are much higher levels than those found in bamboo (1.0 per kb)28, peach (1.5 per kb)29 and poplar (2.6 per kb)30, while they are comparable to those of grape (4.2 per kb)31 and potato (4.3 per kb)18. Cassava’s high heterozygosity may have important implications for the severe inbreeding depression observed in this crop.

Comparative genomics analysis revealed a considerable amount of genome diversity (SNVs and InDels) in W14, KU50 and CAS36 when compared with the reference genome of AM560. We identified 6.9 SNVs and 0.8 InDels per kb in W14, whereas 0.7 SNVs and 0.08 InDels per kb in the KU50. The number of SNVs discovered by comparison between the W14, KU50, CAS36 and AM560 genomes ranged from 2.84 to 4.81 millions (Supplementary Note 15, Supplementary Tables 4, 10, 11 and 12). Of these, 570,695 were shared by the genomes of wild and cultivated varieties, and 200,908 were found in genic regions. These SNVs constitute valuable markers for genotyping, genetic analysis and genomics-based breeding in cassava (Fig. 1a, Supplementary Note 12).
A total of 63 microRNA (miRNA) families consisting of 147 miRNAs were identified in the cassava genome, including 22 previously reported (Supplementary Note 21 and Supplementary Data 7). Other noncoding RNAs, including transfer RNAs (tRNAs, 861 in W14 and 707 in KU50), ribosomal RNAs (rRNAs, including 18S, 26S, 5.8S and 5S; 337 in W14 and 192 in KU50), small nuclear RNAs, small nucleolar RNAs, signal recognition particle RNAs and long noncoding RNAs, were also found in the wild and cultivated cassava genomes (Supplementary Note 21, Supplementary Table 21).

**Genome variation.** Alignment of the larger scaffolds revealed that there is a significant similarity among the three cassava genomes, as expected, more substantial syntenic blocks are found between *M. esculenta* and *Ricinus communis* than between *M. esculenta* and *Arabidopsis thaliana* (Fig. 1c). Of the 15,636 gene families identified in *M. esculenta*, 2,043 were present in cassava but absent in other sequenced Euphorbiaceae genomes (*R. communis* and *Jatropha curcas*) or the outgroup species (*Vitis vinifera*; Supplementary Note 13, Supplementary Fig. 13). Further gene model comparisons among cassava and 12 more distantly related genomes revealed that 8,414 gene models were unique to cassava and 3,710 were specific to Euphorbiaceae (Supplementary Note 13, Supplementary Figs 14 and 15). Using 71 chloroplast genes from eight different plant species (Fig. 1d), we estimated that cassava diverged from rubber tree (*Hevea brasiliensis*) 5.1 million years ago (MYA), 6.4 MYA from physic nut (*J. curcas*) and 14.8 MYA from castor bean (*R. communis*). Therefore, the cassava lineage from which cultivated cassava was originated diverged from a common ancestor approximately 0.5 MYA. These results are in agreement with the reported divergence time between *Manihot* and other genera in the Euphorbiaceae family (Supplementary Note 13).

We compared all the predicted genes from the genomes of W14, KU50 and AM560. A total of 28,302 independent gene models were confirmed, although copy number variations existed and were more frequently observed in the cultivated varieties than in W14 (Supplementary Data 1, Supplementary Table 16). Among the gene models, 1,584 were unique to W14 or lost in KU50 and

---

**Figure 1 | Cassava comparative genomes.** (a) Venn diagram of SNVs/InDels diversity of the cassava genomes of W14, KU50 and CAS36 sequenced in this study with comparison to the AM560 genome sequences previously released. The number of SNVs is listed and the number of InDels is shown in parentheses. (b) Chromosome in situ hybridization showing the repeated occurrence of 45S (Nucleolus organizer, NOR), LTR and chromosome numbers (*2n* = 36) of cultivar KU50. (c) A CirCOS (http://circos.ca/) figure showing synteny between three paralogous cassava genomic regions and their putative orthologues present in *R. communis* and *A. thaliana* genomes. Coloured lines connect the cassava scaffolds to the *A. thaliana* chromosomes and *R. communis* scaffolds. The line distances across different scaffolds denote the similarities of the segments, with a longer line indicating a higher similarity. (d) Gene tree showing the divergence time of the wild ancestor subspecies to cultivars, referenced to neighbour species in the Euphorbiaceae family inferred from sequence comparison to 71 chloroplast genes from eight different plant species. Mtr: *Medicago truncatula*, Csa: *Cucumis sativus*, Ptr: *Populus trichocarpa*, Pni: *Populus nigra*, Rco: *Ricinus communis*, Ees: *Euphorbia esula*, Jcu: *Jatropha curcas*, Mef-W14: *Manihot esculenta* ssp. *flabellifolia* (W14), Mes-KU50: *Manihot esculenta* ssp. *esculenta* (KU50), Mes-AMS60: *Manihot esculenta* ssp. *esculenta* (cultivar AMS60).
AM560, whereas another 1,678 genes were specific to the cultivated varieties, and 20,133 homologous genes (including 16,219 high-confidence orthologues) were shared among the three draft genome sequences (Supplementary Note 14). The majority of the present and absent variation genes could be assigned to six Gene Ontology (GO) functional categories, including ‘catalytic activity’, ‘binding’, ‘metabolic process’, ‘cellular process’, ‘cell’ and ‘cell part’, and those genes with significant copy number variation were mainly ascribed into the first three functional categories (Supplementary Figs 17–19 and 41–43). Significant differences in the average SNVs were detected between W14 and the cultivated varieties (1.7%) and between the two cultivars (0.5%; Supplementary Note 16, Supplementary Fig. 24).

The synonymous \((K_s)\) and nonsynonymous substitution rate \((K_a)\) and selection pressure \((K_a/K_s)\) of the gene set were used to describe evolutionary signatures of the cassava genome\(^{35,36}\) (Supplementary Note 16, Supplementary Fig. 26). Approximately 2,818 genes were strictly positively selected (Fig. 2a, \(K_a/K_s > 1\)), 436 genes were negatively selected (Fig. 2a, \(K_a/K_s < 1\)), and 2,260 genes were under neutral selection (Fig. 2a, \(K_a/K_s = 1\)).

**Figure 2 | Selection pressure and carbon flux diversification in cassava.** (a) Chart for synonymous substitution \((K_s)\) and nonsynonymous substitution rate \((K_a)\) and selection pressure \((K_a/K_s)\) between wild W14 and cultivated variety (WC) and between cultivated varieties (CC). \(K_a/K_s = 1\) indicates genes with neutral selection, \(K_a/K_s > 1\) indicates positive selection and \(K_a/K_s < 1\) indicates negative selection. (b) The differential expression patterns of genes involved in photosynthesis, Calvin cycle, sugar transport and starch synthesis in storage roots and leaves between cultivated varieties (KU50 and Arg7) and wild ancestor (W14) revealed by digital transcriptome sequencing. (c) A model of high-efficient starch accumulation in the tuber roots of domesticated cassava. Red arrows present the carbon flux directions in cultivar and blue arrows indicate the carbon flux directions in wild W14. The width of the arrow indicates the strength of carbon flux. The gene symbol marked in red shows genes with copy number expansion in cultivars. cPGM, cytoplasmic phosphor-glucomutase; GPI, glucose-6-phosphate isomerase; G6PT, glucose-6-phosphate/phosphate translocator; pPGM, phospho-glucomutase; SBE, starch branching enzyme; SS, starch synthase; SUT, sucrose transporter; TPT, triosephosphate translocator; UTP, uridine triphosphate.
Carbon flux diversification. The high carbon accumulation in the form of starch in the storage root is an extraordinary feature of cultivated cassava. The transcriptome annotation showed that a considerable number of genes involved in photosynthesis and the Calvin cycle in leaves, and sucrose transport and starch synthesis in storage roots were preferentially expressed in the two domesticated varieties when compared with the wild W14 (Fig. 2b). This is consistent with the higher vigour and yield potential showed by KU50 and Arg7 relative to W14 (Supplementary Table 1, Supplementary Fig. 1). These results were confirmed by reverse transcriptase–quantitative PCR of selected genes (Supplementary Note 20, Supplementary Fig. 34). Also, we found an alternative starch synthesis pathway relying on plastid phosphorylase (Pho1), which was expressed at a higher level in cultivated varieties than in W14 in the storage roots. This pathway allows glucose 1-phosphate to be directly transferred into amyloplasts, as shown in rice grain and potato tubers. The expression level of genes involved in cell wall synthesis and secondary metabolism are significantly decreased in leaves and storage roots of cultivated varieties in comparison to the wild subspecies (Supplementary Note 18, Supplementary Figs 32–33). These results agree with the observation that SWEET genes controlling sucrose efflux into the cell wall show reduced expression in KU50 and Arg7, but not in W14. At the genome level, copy number expansion and alternative splicing were found in several key genes in the cultivars, such as aldolase, phosphoglycerate kinase and ribulose bisphosphate carboxylase, which are involved in photosynthesis. Genes involved in starch synthesis and accumulation in amyloplasts of storage roots such as sucrose transporters, sucrose synthases, ADP glucose pyrophosphorylase (Apl), starch branching enzymes and phospho-glucomutase showed similar copy number and alternative splicing differences (Supplementary Data 5). These genes have been identified as the key genes strongly associated with cassava storage root development. Furthermore, our result suggests that miRNAs may play a role in regulating storage root formation and growth as well as starch synthesis. At least nine miRNAs could target genes that were highly expressed in the storage roots of cultivars and were involved in the photosynthesis and carbon metabolism pathways (Supplementary Note 21, Supplementary Table 19 and Supplementary Fig. 36) as observed in other plants. For example, miR394 directly downregulates APL2, a key gene in starch synthesis at the late stage of storage root development in cultivated varieties; likewise miR319, miR159, miR160, miR166 and miR396 negatively regulate their targets, such as MYB33 and ARF10, which control starch synthesis through ABA signalling (Supplementary Note 9). miR167, miR169 and miR156 positively regulate transcription factors RD19, NF-YA3 and SPL13B, respectively, which are involved in storage root and leaf development. Remarkably, target genes such as MYB33 and ARF10 are known to bind to cis-elements in the upstream regions of genes related to starch metabolism, such as SuSy, APL and genes involved in photosynthesis (Supplementary Note 22, Supplementary Figs 35 and 37–40). Taken together, the observed gene expression patterns, enhanced agronomic phenotypes, copy number expansions and miRNA- and cis-element-mediated regulation of key genes suggest that carbon flux could have been shifted as a result of domestication in cassava. Based on these findings, we suggest a model of the efficient accumulation of starch in cassava (Fig. 2c).

Cyanogenesis differentiation. The latent toxicity caused by cyanogenesis in cassava is clearly a potential health hazard when it is consumed as food. The pathway for cyanogenic glucoside biosynthesis in cassava and the genes encoding the enzymes involved have been elucidated in recent years. We...
determined the linamarin and lotaustralin content in cultivated KU50, Arg7 and wild W14, and found that the linamarin content was reduced six- to tenfold in storage roots and three- to fourfold in leaves of KU50 and Arg7 relative to W14 (Fig. 3a, Supplementary Note 19, Supplementary Table 18). Remarkably, the expression of the genes CYP79D1, CYP79D2, CYP71E7, CYP71E11, UGT85K4 and UGT85K5 that encode the enzymes catalysing linamarin and lotaustralin formation, all exhibited five- to tenfold lower expression levels in the storage roots and leaves of KU50 relative to W14, further suggesting a potential outcome of domestication. Different classes of DNA retrotransposons, like miniature inverted-repeat transposable elements (MITEs) and LTR transposable elements, have been shown to influence the expression of proximal genes, especially if simultaneously situated downstream and upstream of the same gene. In general, gene expression is suppressed by the presence of these elements. To investigate potential effects of transposons on gene expression in cassava, the 1-kb upstream regions of orthologous genes present in the W14, KU50 and AM560 genomes were analysed for the presence of MITEs. A total of 553 MITEs were found, of which 310 and 243 were uniquely present within the genomes of AM560 and W14, respectively. Among the 310 AM560-specific MITE insertions, 96 (34.5%) showed significantly lower expression and 32 (11.5%) had significantly higher expressions in storage roots or leaves of cultivated varieties when compared with W14 (Supplementary Data 6). We compared the genomic regions containing CYP71E11, CYP71E7 and UGT85K4, and found that these three genes were positioned in a linear array as CYP71E11, CYP71E7 and UGT85K4, there were more transposable or retrotransposable elements in the gene 1-kb upstream regions of cultivated species KU50 and AM560 than wild subspecies W14. CDS, Coding sequence.

**Discussion**

We produced and annotated two draft genomes of cassava, a cultivated variety and a wild ancestor. Comparative analysis provided new insights into cassava genome evolution and genetic events that may have occurred during domestication. Gene
models specific to either wild or cultivated cassava were elucidated. We found a high degree of heterozygosity between the analysed cassava genomes and gene sets that have been strictly selected during the process of evolution and, potentially, domestication. Genes responding to stimulus such as light, high temperature, water stress and oxidative stress were highly expressed in domesticated cassava, most likely reflecting their adaptation to tropical and dryer growth conditions. On the other hand, some genes involved in ion transport and starch metabolism that could enhance starch yield potential, and the other leading to a dwindling of cell wall and secondary metabolism, including cyanogenic compounds. This carbon flux shift towards starch accumulation would be desirable in cultivated varieties, whereas stress tolerance may not be so critical in cultivation conditions, as biotic and abiotic stress can be milder in natural environments. Therefore, we propose that a pathway that prioritizes starch accumulation versus cyanogenesis has been selected in cultivated cassava. In addition, the development of substantial new genomic resources, including millions of SNVs, which are available in a public database (http://www.cassava-genome.cn/), will promote development of toolkits for enhanced cassava breeding.

Methods

Genomic DNA isolation. To reduce organelle contamination in genomic DNA, nuclei were isolated from fresh young leaves of W14 and KU50, as described by Zhang et al.52 Briefly, approximately 100 g of tissues were ground into a fine powder in liquid nitrogen and transferred to a beaker containing ice-cold 1% homogenization buffer plus 0.5% Triton X-100 and 0.15% β-mercaptoethanol. After filtering the homogenate through cheesecloth and Miracloth, the nuclei were washed with the same buffer and centrifuged. This step was repeated until the nuclei pellet became white. DNA was extracted from the nuclei as described by Kidwell and Osborn51. The purified DNA was dissolved in 1 ml of TE-buffer for Illumina and 454 sequencing.

Genome sequencing. Genome sequence data of W14 and KU50 were produced with the Illumina HiSeq2000 (Illumina) and Roche/454 GS FLX platforms (Roche) at the Beijing Institute of Genomics and Qingdao Bioenergy and Process Institute of the Chinese Academy of Sciences. A total of 2.6 Gb high-quality DNA sequence for W14 and 34.43 Gb for KU50 representing 103- and 46-fold coverage of the 742-Mb cassava genome, respectively. The BAC libraries of wild W14 and cultivated KU50 were fingerprinted, 58,244 clones of AM560-2 and 24,784 clones from W14 were analyzed to identify novel miRNAs and profile miRNA expression following the method previously documented66. The qualified reads, the ones that carried the adaptor and were longer than 17-nt, were mapped to the known miRNAs to detect their expression and determine their expression abundance. The qualified reads not mapped to the known miRNAs were mapped to a cassava genome to identify novel miRNAs. The conservation and specificity of all newly identified and known cassava miRNAs were analysed across the cassava genomes and eight diverse plant species. The sequences of these miRNAs were aligned to the genomes of the cassava cultivars and the other eight plants using BLAST with the P-value threshold set to 1e-10; the alignment results were further manually examined to determine homologue to a miRNA.

Genome assembly and annotation. The small RNA-seq data of W14 and KU50 were analysed to identify novel miRNAs and profile miRNA expression following the method previously documented66. The qualified reads, the ones that carried the adaptor and were longer than 17-nt, were mapped to the known miRNAs to detect their expression and determine their expression abundance. The qualified reads not mapped to the known miRNAs were mapped to a cassava genome to identify novel miRNAs. The conservation and specificity of all newly identified and known cassava miRNAs were analysed across the cassava genomes and eight diverse plant species. The sequences of these miRNAs were aligned to the genomes of the cassava cultivars and the other eight plants using BLAST with the P-value threshold set to 1e-10; the alignment results were further manually examined to determine homologue to a miRNA.

MicroRNA and noncoding RNA annotation. The small RNA-seq data of W14 and KU50 were analysed to identify novel miRNAs and profile miRNA expression following the method previously documented66. The qualified reads, the ones that carried the adaptor and were longer than 17-nt, were mapped to the known miRNAs to detect their expression and determine their expression abundance. The qualified reads not mapped to the known miRNAs were mapped to a cassava genome to identify novel miRNAs. The conservation and specificity of all newly identified and known cassava miRNAs were analysed across the cassava genomes and eight diverse plant species. The sequences of these miRNAs were aligned to the genomes of the cassava cultivars and the other eight plants using BLAST with the P-value threshold set to 1e-10; the alignment results were further manually examined to determine homologue to a miRNA.

RNA-Seq analysis. The RNA-Seq reads were de novo assembled by Trinity56 and Velvet-Oases62-63, and those RNA-Seq reads and their assembled transcripts were mapped to the W14, KU50 and AM560 draft sequences using TopHat and BLAT. And the ab initio alignment to itself assembly draft genome using TopHat v2.0.6 and cufflinks v2.0.2, statistical analysis was completed with cummeRbund v1.2.0 to assist. And the whole-genome and genome annotated to curated set of gene collection in NCBI non-coded repeats database (NT), non-code repeats database (RT), non-coding ESTs and Rfam database. The Consensus approach of the ab initio gene prediction on the repeat-masked genome, via protein similarity and transcript reconstruction to build optimal gene models using the PASA and EVM pipelines56,64. All genome assembly and annotation data of W14 and KU50 could be found at http://www.cassava-genome.cn/data.html with visionV1.0.

Cyanogen analysis. The cyanogenic glucoside contents of roots and leaves of the wild W14 and cultivated KU50 was determined by liquid chromatography–mass spectrometry. Five plants were analysed separately for each of them. A leaf disc was cut from the first unfolded leaf of each plant by snap-closing the 2 ml-culture tubes (Nunc) using the tweezers. Five leaves were homogenized in 300 μl of pre-warmed 85% (v/v) methanol for leaf and tuber, respectively. After closing the tube and securing the lid with a cap lock, the samples were boiled in a water bath at 100 °C for 3 min (leaf) or 5 min (tuber). Then, the MeOH extract was transferred into a new tube, lyophilized to dryness, re-suspended in water in a total volume of 200 μl and filtered through a 0.45-μm filter. Analytical liquid chromatography–mass spectrometry was carried out using an Agilent 1100 Series LC (Agilent Technologies).
References

1. Allem, C. A. C. The origin of Manihot esculenta Crantz (Euphorbiaceae). Genet. Resour. Crop Evol. 41, 133–150 (1994).
2. Allem, C. A. C. The closest wild relatives of cassava (Manihot esculenta Crantz). Euphytica 107, 1–12 (1999).
3. Olsen, K. M. & Schaal, B. A. Evidence on the origin of cassava: phyllogeography of Manihot esculenta. Proc. Natl Acad. Sci. USA 96, 5586–5591 (1999).
4. Bartlett, A. S., Barghoorn, E. S. & Berger, R. Moss faunia from panama. Science 165, 389–390 (1969).
5. Gibbons, A. New view of early Amazonia: recent findings suggest complex culture was indigenous to the Amazon basin-upsetting some received opinions about environment and culture. Science 248, 1488–1490 (1990).
6. Raheem, D. & Chukwuma, C. Foods from cassava and their relevance to Nigeria and other African countries. Agr. Hum. Values 18, 383–390 (2001).
7. Balat, M. & Balat, H. Recent trends in global production and utilization of bioethanol fuel. Appl. Energy 86, 2273–2282 (2009).
8. Schmitz, P. M. & Kavallari, A. Crop plants versus energy plants—On the international food crisis. BioEnergy. Med. Chem. 17, 4020–4021 (2009).
9. Carvalho, R. D. & Guerra, M. Cytogenetics of Manihot esculenta Crantz (cassava) and eight related species. Hereditas 136, 159–168 (2002).
10. Fregene, M. A. et al. Simple sequence repeat marker diversity in cassava landraces: genetic diversity and differentiation in an axenopagely propagated crop. Theor. Appl. Genet. 107, 1083–1093 (2003).
11. Siqueira, M. V. B. M. et al. Microsatellite polymorphisms in cassava landraces from the Cerrado biome. Mato Grosso do Sul, Brazil. Biochem. Genet. 45, 879–895 (2010).
12. Okogbenin, E. & Fregene, M. Genetic mapping of QTLs affecting productivity of Manihot esculenta. Proc. Natl Acad. Sci. USA 106, 20339–20346 (2009).
13. Qu, Q. et al. The yak genome and adaptation to life at high altitude. Nat. Genet. 41, 946–949 (2012).
14. Jain, M. A. et al. A draft genome sequence of the pulse crop chickpea (Cicer arietinum L.). Plant J. 74, 715–729 (2013).
15. Sato, H. et al. Mutation of the plastidial z-glucan phosphorylase gene in rice affects the synthesis and structure of starch in the endosperm. Plant Cell 20, 1833–1849 (2008).
16. Fettke, J. et al. Glucose 1-phosphate is efficiently taken up by potato (Solanum tuberosum) tuber parenchyma cells and converted to reserve starch granules. N. Phytol. 185, 663–675 (2009).
17. Yang, J., An, D. & Zhang, P. Expression profiling of cassava storage roots reveals an active process of glycolysis/glucogenesis. J. Integr. Plant Biol. 53, 193–211 (2011).
18. Hu, Y.-F. et al. Binding of ABI4 to a CACCG motif mediates the ABA-induced expression of the ZmSS1 gene in maize (Zea mays L.) endosperm. J. Exp. Bot. 63, 5979–5989 (2012).
19. Peng, T. et al. Characterization and expression patterns of microRNAs involved in rice grain filling. PLoS ONE 8, e54118 (2013).
20. Bernoux, M. et al. RD19, an Arabidopsis cysteine protease proteome for RRS1-R mediated resistance, is relocalized to the nucleus by the ralstonia solanacearum PopP2 effector. Plant Cell 20, 2252–2264 (2008).
21. Nardini, M. et al. Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination. Cell 152, 132–143 (2013).
22. Yang, Z. et al. Comparative study of SBP-box gene family in Arabidopsis and rice. Gene 407, 1–11 (2008).
23. Zhou, S.-R., Yin, L.-L. & Xue, H.-W. Functional genomics based understanding of rice endosperm development. Curr. Opin. Plant Biol. 16, 236–246 (2013).
24. Jørgensen, K. et al. Biosynthesis of the cyanogenic glucosides linamarin and lotaustralin in cassava: isolation, biochemical characterization, and expression pattern of CYP71F7, the oxime-metabolizing cytochrome P450 enzyme. Plant Physiol. 155, 282–292 (2011).
25. Kodangara, R. et al. Characterization and expression profile of two UDP-glucosyltransferases, UGT85K4 and UGT85K5, catalyzing the last step in cyanogenic glucoside biosynthesis in cassava. Plant J. 68, 287–301 (2011).
26. Gleadow, R. M. & Müller, B. L. Cyanogenic glycosides: synthesis, physiology, and phenotypic plasticity. Ann. Rev. Plant Biol. 65, 155–185 (2014).
27. Wang, X., Weigel, D. & Smith, L. M. Transposon variants and their effects on gene expression in Arabidopsis. PLoS Genet. 9, e1003255 (2013).
28. Zhang, H.-B., Zhao, X., Ding, X., Paterson, A. H. & Wing, R. A. Preparation of megabase-size DNA from plant nuclei. Plant J. 7, 175–185 (1995).
29. Kidwell, K. K. & Osborn, T. C. in Plant Genomes: Methods for Genetic and Physical Mapping (eds Beckman, J. S. & Osborn, T. C.) 1–13 (Kluwer Academic Publishers, 1992).
30. Tao, Q., Wang, A. & Zhang, H.-B. One large-insert plant transformation-competent BIBAC library and three BAC libraries of Japonica rice for genome research in rice and other grasses. Theor. Appl. Genet. 105, 1058–1066 (2002).
31. Luo, M. et al. High-throughput fingerprinting of bacterial artificial chromosomes using the SNAPSHOT labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82, 378–389 (2003).
32. Nelson, W. M. et al. Efficacy of clone fingerprinting methodologies. Genomics 89, 160–165 (2007).
33. Gu, Q. et al. Construction of physical map for Brachypodium distachyon and its comparative analysis with rice. BMC Genom. 10, 496 (2009).
34. van Oeveren, J. et al. Sequence-based physical mapping of complex genomes by whole genome profiling. Genome Res. 21, 618–625 (2011).
35. Engler, F. W., Hatfield, J., Nelson, W. & Soderlund, C. A. Locating sequence on FPC maps and selecting a minimal tiling path. Genome Res. 13, 2152–2163 (2003).
36. Nelson, W. & Soderlund, C. Integrating sequence with FPC fingerprint maps. Nucleic Acids Res. 37, 3048–3058 (2009).
37. Haas, B. J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31, 5654–5666 (2003).
38. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9, R7 (2008).
39. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnol. 29, 644–652 (2011).
40. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829 (2008).
63. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086–1092 (2012).
64. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnol. 31, 46–53 (2013).
65. Zhang, W. et al. Multiple distinct small RNAs originate from the same microRNA precursors. Genome Biol. 11, R81 (2010).
66. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964 (1997).
67. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. Bioinformatics 25, 1335–1337 (2009).

Acknowledgements

The project was funded partly by the National Basic Research Program of China (973 Program, 2010CB126600) and by the Chinese Academy of Tropical Agriculture Sciences (CATAS). The physical map of AM560 integrated in this study was obtained based on funds by the Generation Challenge Program grant G3007.03. Thanks to the DOE Joint Genome Institute (JGI) and Steve Rounsley who contributed the draft assembly of AM560 (http://www.phytozome.net). Thanks to the Beijing Institutes of Life Sciences, Chinese Academy of Sciences, for providing computational resources for this project, Jian Xu of Qingdao Institute of Bioenergy and Bioprocess Technology, CAS for part of genome sequencing. Thanks to Dr Wilhelm Gruissem of ETH in Switzerland and Dr Hans Bohnert of Illinois University, USA, who provided constructive suggestions, and thanks to all members of the 973 programme and other scientists for discussions. This project was organized and performed under the Cassava Genome Consortium, thanks to all participates for their contributions.

Author contributions

W.W., M.P. conceived the project, designed the studies and contributed to the original concept of the manuscript. J.Xiao, K. Li and B. Liu analyzed the data as a whole and wrote the manuscript. B. Feng developed the de novo assembly pipeline and performed de novo genome assembly. X. Chen, Y. Zhang, K. Pan and Q. Yao performed DNA preparation. Z. Xia, P. Li, S. Zhong, J. Liu and J. Zhang performed transcriptome (RNA-seq and cDNA) analyses. X. Zhou, G. Wu, P. Ling and J. Guo performed the repetitive sequence analysis and tRNA, tRNA, small nucleolar RNA annotation. Z. Li, M. Hu, S. Wang, W. Liao, W. Hu, S. Zhang, M. Zou, M. Wen, J. Pri, P. Ma and M. Ruan completed the Q-PCR validation of selected genes. M.C. Luo, P. Rabinowicz, J. Wu, Y. Ma, X. Liu, L. Tallon, K. Galens, S. Ott, F. You and Y. Fu performed construction of BAC libraries and physical map. H. Zhang and Y. Fu produced part of BAC clones used for BAC pooling sequencing and H. Zhang also contributed for edition of in language. Q. Lou, H. Wang, C. Lu and A. Guo performed cytogenetics studies. S. Zhou, S. Hu, Z. Wu, H. Liu, S. Sun provided IT support. G. Liu, Y. Chen and Q. Wang coordinated the project. Y. Wang, G. Xiao, L.J. Carvalho and S. Chen performed the comparative genomics analyses across the species. W. Zhang, J. Xia and C. Zeng completed the annotation of miRNAs. B. Möller, R. Kannangara, K. Jørgensen performed the analysis of cyanide metabolism and contributed to the writing of the manuscript. H. Ceballos, M. Fregene, L.A.B. López-Lavalle, R. Neale, N. Heinz, M. Bonde and P Zhang, gave the revision of manuscript.

Additional information

Accession numbers: Cassava genome sequence data have been deposited at DDBJ/EMBL/GenBank under the accession code JPQF00000000 for W14 and JPQE00000000 for KU50. RNA-seq reads have been deposited in GenBank/EMBL/DBJ sequence read archive under the accession codes SRX551093, SRX553797, SRX553798, SRX553799, SRX553800, SRX553801, SRX553802, SRX553803, SRX553804, SRX553805, SRX553806 and SRX553807.

Supplementary information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions

How to cite this article: Wang, W. et al. Cassava genome from a wild ancestor to cultivated varieties. Nat. Commun. 5:5110 doi: 10.1038/ncomms6110 (2014).