Lawrence Berkeley National Laboratory

Title
Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity

Permalink
https://escholarship.org/uc/item/1mq3v1pt

Authors
Bredeson, Jessen V
Lyons, Jessica B
Prochnik, Simon E
et al.

Peer reviewed

RESOURCE

Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity

Jessen V Bredeson[1], Jessica B Lyons[1], Simon E Prochnik[2], G Albert Wu[2], Cindy M Ha[1,16], Eric Edsinger-Gonzales[1,16], Jane Grimwood[2,3], Jeremy Schmutz[2,3], Ismail Y Rabbi[4], Chiedozie Egesi[5], Poasa Nauluvula[6,16], Vincent Lebot[7], Joseph Ndunguru[8], Geoffrey Mkamilo[9], Rebecca S Bart[10], Tim L Setter[11], Roslyn M Gleadow[12], Peter Kulakow[4], Morag E Ferguson[13], Steve Rounsley[14] & Daniel S Rokhsar[1,2,15]

Cassava (*Manihot esculenta*) provides calories and nutrition for more than half a billion people. It was domesticated by native Amazonian peoples through cultivation of the wild progenitor *M. esculenta* ssp. *flabellifolia* and is now grown in tropical regions worldwide. Here we provide a high-quality genome assembly for cassava with improved contiguity, linkage, and completeness; almost 97% of genes are anchored to chromosomes. We find that paleotetraploidy in cassava is shared with the related rubber tree *Hevea*, providing a resource for comparative studies. We also sequence a global collection of 58 *Manihot* accessions, including cultivated and wild cassava accessions and related species such as Ceará or India rubber (*M. glaziovii*), and genotype 268 African cassava varieties. We find widespread interspecific admixture, and detect the genetic signature of past cassava breeding programs. As a clonally propagated crop, cassava is especially vulnerable to pathogens and abiotic stresses. This genomic resource will inform future genome-enabled breeding efforts to improve this staple crop.

Cassava, also known as manioc, tapioca, and yuca, is a widely grown drought-tolerant crop that can be cultivated on marginal soils and can produce high yields in favorable growing conditions. Its starch-filled storage roots provide a major source of calories in tropical regions[1]. The likely wild progenitor of cultivated cassava is *M. esculenta* ssp. *flabellifolia* (Pohl), a woody perennial shrub that is found throughout the Amazon basin[2–5]. Although domesticated over 6,000 years ago[6–10], cassava cultivation spread beyond South America only in the past 500 years, exported by European colonialists and slave traders[11]. Nowadays, cassava is one of the most widely cultivated tropical crops, especially in sub-Saharan Africa where it has undergone additional improvement through introgression and focused breeding, with the primary aims of conferring disease tolerance and increasing yield[12,13].

Cassava can outcross but is commonly clonally propagated, and harbors considerable genetic load[14]. The reliance on clonal propagation and the limited diversity of African cassava germplasm make it particularly susceptible to the spread of viral and bacterial diseases such as cassava mosaic disease (CMD), cassava brown streak disease (CBSD), and cassava bacterial blight[15,16]. In contrast to African varieties, Thai elite varieties retain considerable diversity[17]. Genetic improvement through conventional breeding in cassava is a challenging and lengthy process, owing to the 12-month cropping cycle, limited seed set of elite varieties, asynchronous flowering and most importantly, the long breeding cycle, which mainly results from the slow clonal multiplication rate (around 1:5 to 1:10 per generation), coupled with the need to obtain phenotypic data in replicated trials. Development of genomic resources, such as a chromosome-scale reference sequence, increased understanding of the cassava gene pool (including wild relatives), and insights into population structure, is expected to accelerate progress in basic biological research and genetic improvement.

We report the chromosome-scale structure of the cassava genome and its formation by an ancient whole-genome duplication that is shared with the rubber tree genus *Hevea*. To better understand the global genetic diversity of cultivated cassava and its wild relatives, we sequenced 53 cultivated and wild accessions of *M. esculenta* from South America, Africa, Asia, and Oceania using whole genome shotgun methods (median 63-fold, range 19- to 168-fold) (**Table 1**).

In this report we use "cassava" to refer to cultivated and/or domesticated varieties of *M. esculenta*, and the shorthand *M. esc. flabellifolia* for wild accessions[3]. We also shotgun-sequenced five *Manihot* accessions related to cassava, including three from the wild species *M. glaziovii* Muell. Arg., one named *M. pseudoglaziovii* Pax & K. Hoffman, and "tree" cassava, a suspected hybrid sometimes called *M. catingea* Ule[12,18]. The Ceará or India rubber tree species *M. glaziovii*, also domesticated in South America, was imported to East Africa in the early twentieth century. It is interfertile with cassava and has been used in African breeding programs to exploit the natural resistance of *M. glaziovii* to cassava pathogens[18]. To analyze genetic variation present in African varieties, we also characterized 268 cultivars of cassava using reduced representation genotyping-by-sequencing (GBS)[19] (**Table 2**).

RESULTS
Chromosome structure
To produce a high-quality chromosome-scale reference genome for cassava, we augmented our earlier draft sequence[20] of the reference genotype AM560-2 with additional whole genome shotgun sequencing and mate pair data, fosmid-end sequences, and a paired-end library developed using proximity ligation of *in vitro* reconstituted chromatin[21] (Methods and **Supplementary Note 1**). AM560-2 is an S3 line bred at Centro Internacional de Agricultura Tropical (CIAT) from MCOL1505 (also known as Manihoica P-12 (ref. 22). Compared with the previous draft[23], the contiguity of our new shotgun assembly has more than doubled (N50 length 27.7 kb vs. 11.5 kb), and an additional 135 Mb is anchored to chromosomes[23] (**Supplementary Note 1**). To organize the sequence into chromosomes we integrated the shotgun assembly with a 22,403-marker consensus genetic map[23] and two other recently published maps[24,25] to produce 18 'pseudomolecules' that represent the 18 linkage groups of cassava (**Supplementary Note 1**). This draft genome encodes 33,033 predicted protein-coding genes, based on homology and transcriptome data for a variety of tissues and conditions (**Supplementary Note 2**); of these predicted genes, 96.6% are anchored to a chromosomal position. Gypsy transposable elements containing long terminal repeats comprise more than half of the 299.3 Mb of repetitive sequence present in our assembly (**Supplementary Note 2**). An estimated 200 Mb of unassembled sequence includes highly repetitive centromeres and high copy repeats, but less than 1% of cassava genes (**Supplementary Note 1**).

Table 1 Whole genome shotgun sequenced *Manihot* accessions

Accession	Location	Accession	Location
M. esculenta ssp. *flabellifolia.* Brazilian accessions of the wild progenitor species of cassava. 10 sequenced, 10 distinct.			
"*M. flabellifolia*"	Brazil	FLA 496-1	Brazil[a]
FLA 449-1	Brazil[a]	FLA 503-2	Brazil[a]
FLA 433-2	Brazil[a]	FLA 490-1	Brazil[a]
FLA 444-1	Brazil[a]	FLA 502-1	Brazil[a]
FLA 488-1	Brazil[a]	FLA XXX-15	Unknown[c]
South American cassava cultivars intended to represent the diversity of the original domestication of cassava (including the AM560-2 reference accession). 7 sequenced, 7 distinct.			
AM560-2	Colombia	CM 507-37	Colombia
MCOL22	Colombia[42]	BRA 856	Brazil[a]
MBRA 685	Brazil[42]	CM 3306-4	Colombia[43]
MCOL 1468	Brazil[43]		
African landraces and improved accessions nominally derived from the original cassava germplasm transferred to Africa, and recent African breeding improvement. 20 sequenced, 16 distinct.			
Albert	Tanzania and Kenya[44]	Akena	Uganda[45]
Mkombozi	Tanzania and Kenya[44]	TME204 (TME419)	Uganda; Togo; Nigeria[23]
Nachinyaya	Tanzania[23]	Kibaha	Tanzania
Muzege	Tanzania[b]	EBW-2 (EBW-A)	Uganda[46]; Kenya and Uganda[44]
NDL06/132	Tanzania[23]	TME3 (TME7, TME14K)	Nigeria[47]; Kenya, Uganda, Tanzania[44]
Kiroba	Tanzania[23]	TME117	Nigeria[47]
Kibandameno	Tanzania and Kenya[44]	60444	West Africa[48]
Aulizaye Mjinga	Tanzania	KBH 2006/18	Kenya
Asian and Asian Pacific cassava varieties from China, Australia, Fiji, and Vanuatu. In addition, our analysis includes KU50 from Thailand[27]. 12 sequenced, 10 distinct.			
SC8	China	Merelesita	Fiji
UnkAus (TMS-I50395/Unk)	Australia	Me001Vu	Vanuatu
SMI150	Australia	Me002Vu	Vanuatu
Avoca	Australia	Me003Vu	Vanuatu
Nadelei(B) (Nadelei(U))	Fiji	Me004Vu	Vanuatu
Mixed crosses. Improved varieties known to be derived from intentional crosses between wild, South American, African, and/or Asian varieties. 4 sequenced, 4 distinct.			
TMS-I30572	Tanzania/Brazil	AR 40-6	Nigeria/Thailand[23]
TMS-I972205	Tanzania/Brazil/Nigeria	AR 37-80	Nigeria/Thailand[23]
Other *Manihot*, including *M. glaziovii* (Ceará rubber tree); *M. pseudoglaziovii*, another nominally distinct tree species from Brazil; and "tree" cassava, presumed to be an *M. glaziovii–M. esculenta* hybrid[18]. In addition, our analysis includes the "wild cassava" W14 from Wang *et al.*[27], which we identified as *M. glaziovii*. 5 sequenced, 5 distinct.			
M. glaziovii(R)	Tanzania	*M. pseudoglaziovii* PSE XXX-1	Unknown[c]
M. glaziovii(S)	Tanzania	Tree Cassava	Tanzania
M. glaziovii GLA XXX-8	Unknown[c]		

Fifty-eight accessions were sequenced in total, and 52 distinct accessions remained after clones were excluded. Accessions listed in parentheses are those we determined to be clones. Two additional accessions were obtained from Wang *et al.*[27] See **Supplementary Dataset 1** for detailed sourcing, provenance, and phenotype information; SRA BioSample accession numbers; and sequencing statistics.
[a]CIAT passport information (http://isa.ciat.cgiar.org/urg/cassavacollection.do). [b]CassavaBase (http://www.cassavabase.org/search/stocks). [c]L.A.B. López-Lavalle, personal communication.

Table 2 Cassava accessions genotyped by sequencing

Collection	Sequenced (268 total)	Distinct (215 total)
TMe (Tropical *Manihot esculenta*) cassava accessions from the IITA genebank core collection	97	62
TMEB germplasm accessions used in the IITA breeding program	10	3
TMS (Tropical Manioc Selection[49]) cassava varieties developed and selected by IITA	88	84
SEC: southern, eastern, and central African cassava diversity	73	66

IITA, International Institute of Tropical Agriculture.

Comparative analyses revealed the impact of paleotetraploidy[20,26,27] on the cassava genome (**Fig. 1a**). Analysis of the genomic distribution of paralogs reveals that the $n = 18$ linkage groups of cassava comprise five pairs of homologous chromosomes and two groups of four chromosomes that have undergone a series of breaks and fusions involving homologs. The genus *Manihot* belongs to the Euphorbiaceae, an angiosperm family that includes several other species with commercial importance including castor bean (*Ricinus communis*, $2n = 20$), physic nut (*Jatropha curcas*, $2n = 22$), and rubber tree (*Hevea brasiliensis*, $2n = 36$), which we estimate diverged from cassava ~35 million years ago (mya) (**Supplementary Note 3**). The shared chromosome number of cassava and rubber tree, roughly double the chromosome count of physic nut and castor bean, suggests that the paleotetraploidy present in cassava might be shared with *Hevea*[28,29]. Our analysis confirms this hypothesis, as both species have thousands of homologous gene pairs that diverged approximately 10 million years before the cassava-*Hevea* speciation (**Fig. 1b** and **Supplementary Note 3**). Analysis of single- or two-copy cassava genes with single-copy orthologs in *Jatropha* shows that 36.9% of genes duplicated by paleotetraploidy are retained in two copies in cassava (4,116/11,155 genes analyzed), with similar rates of retention on each of the pairs of homeologs (**Supplementary Note 3**). This phylogenetic analysis of euphorb genomes supports the early branching of the *Ricinus* lineage, agreeing with some genome-wide studies[27] but not others[30].

Global genetic diversity

We used whole genome shotgun sequencing and GBS to sample the global diversity of cassava and its wild relatives as summarized in **Table 1** and further described in **Supplementary Dataset 1**, and **Supplementary Notes 4** and **5** . We also integrated into our analyses a pair of recently published *Manihot* sequences[27]. Our first-principles approach does not depend on pre-assigned species and is alert to possible introgression.

Chloroplast sequences from the sequenced accessions separate into two deeply divergent clades representing distinct *Manihot* species (**Fig. 2a**). The *M. esculenta* clade includes only cassava and *M. esc. flabellifolia* accessions, whereas the *M. glaziovii* clade includes *M. glaziovii* and,

surprisingly, *M. pseudoglaziovii* as well as the putative "wild cassava" W14 (ref. 27; but see below). Analysis of nuclear genome variation by principal component analysis (**Fig. 2b**)[31] and model-based clustering (FRAPPE)[32] (**Fig. 2c**) reveals three distinct clusters: (i) most cultivated cassava, grouped with two *M. esc. flabellifolia* (designated "C/F"); (ii) the remaining sampled accessions of *M. esc. flabellifolia* ("F"); and (iii) *M. glaziovii* ("G"), a cluster that also includes the putative "wild cassava" W14. Several accessions (e.g., Tree Cassava) occupy intermediate positions in principal component analysis and show mixed ancestry in model-based clustering; these are discussed further below.

Accessions in the C/F cluster show a level of heterozygosity (0.84%, based on single-nucleotide variants (SNV) at callable loci, excluding runs of homozygosity) that is approximately twice the rate of homozygous differences as compared with the AM560-2 reference (**Fig. 2d** and **Supplementary Notes 6** and **7**). This is consistent with population-genetic expectation for a randomly mating population that includes the reference haplotype. Many of our nominally outbred cassava accessions show multiple short runs of homozygosity (mean 18 cM, median 8 cM), but this typically accounts for a small fraction of the genome in cassava (**Supplementary Note 6**, Supplementary Fig. 11).

Surprisingly, all but one (the Brazilian BRA 856) of the 39 distinct cultivated cassava accessions in our collection fall into two *M. esculenta* chloroplast (cpDNA) haplogroups that are present on all continents. Although some sharing of cpDNA haplotypes is due to the inclusion of close relatives in our sample (as detected by nuclear genome analysis; **Supplementary Note 8**), the extraordinarily limited cpDNA diversity in cultivated cassava suggests a substantial maternal bottleneck during domestication. Attempts to identify further nuclear genome substructure within the "cassava" group are described below. *M. esc. flabellifolia* accessions in the C/F cluster include FLA 433-2 from the Brazilian state of Rondônia, which has a variation profile indistinguishable from cultivated cassava (http://isa.ciat.cgiar. org/urg/cassavacollection.do; **Fig. 2e**), and cassava-like storage roots (**Supplementary Note 4**, Supplementary Fig. 5) although its cpDNA

Figure 1 *Manihot* paleotetraploidy. (**a**) Conserved synteny between five pairs of chromosomes and two sets of four chromosomes is shown. The ten chromosomes arranged in the large upper circle illustrate 1:1 synteny between five duplicated pairs of chromosomes. Chromosomes are numbered with large black text and physical positions (in Mb) are noted in small black text. The chromosomes depicted in the two smaller circles each share syntenic regions with two other chromosomes, owing to chromosomal rearrangements that occurred after the whole-genome duplication. Pericentromeric regions are shaded on each chromosome, and syntenic segments between chromosomes are connected by gray bands. (**b**) Phylogeny of euphorbs and timing of genome duplication, inferred by comparing homologous divergences within *Manihot* and *Hevea* with orthologous divergences between species. Diamonds indicate the divergence between paralogous sequences within *Manihot* (red) and *Hevea* (purple).
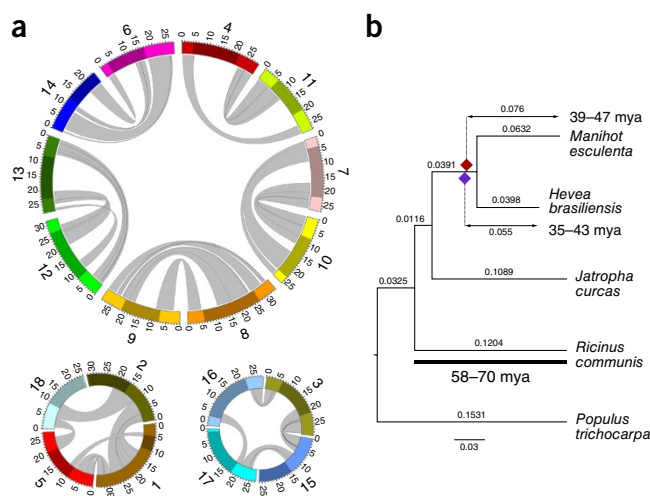
does not match either of the two common cassava haplotypes. Its grouping with cassava is consistent with the haplotype analyses of Olsen and Schaal[3], who found that cassava was domesticated in the western part of the southern Amazon region. FLA XXX-15 shares its cpDNA haplotype with cultivated cassava and also has a cassava-type

nuclear genotype and cassava-like storage roots (**Supplementary Note 4**, Supplementary Fig. 5), but its sampling site is not recorded.

Accessions in the F grouping include *M. esc. flabellifolia* samples from the more eastern portion of the southern Amazon basin. They show comparable levels of heterozygosity (0.61%) to those in C/F
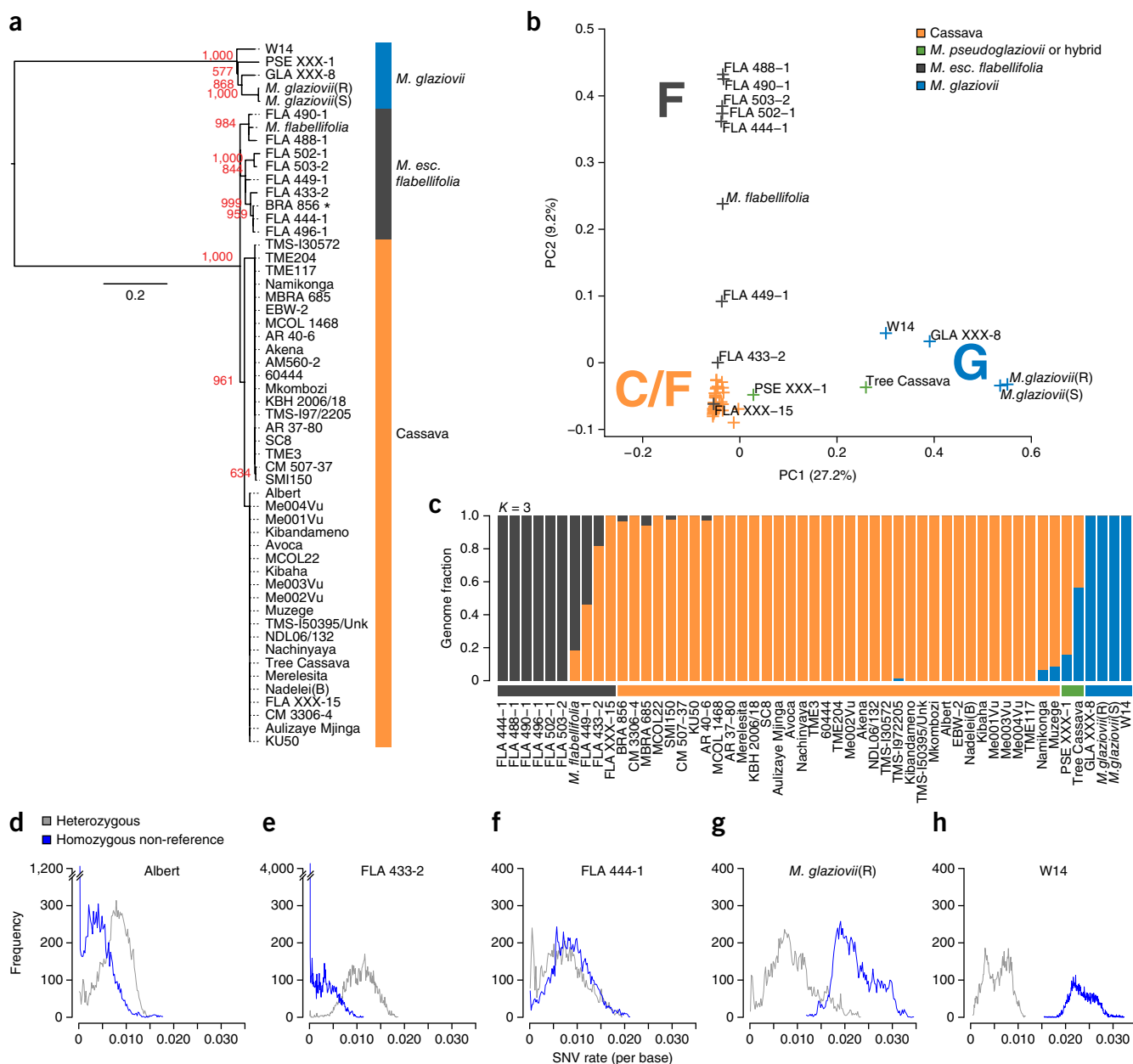
Figure 2 *Manihot* genetic diversity. (**a**) Midpoint-rooted chloroplast genome phylogeny of sequenced *Manihot* accessions. Bootstrap values for nodes with support of 500 or more (out of 1,000) shown in red. For groups of accessions with identical nuclear and chloroplast genomes, only one accession is shown. Note that *M. pseudoglaziovii* and the "wild cassava" W14 group with *M. glaziovii*, and almost all cultivated cassava in our collection have one of two cpDNA haplotypes. The *M. esc. flabellifolia* form a sister clade to cassava with much greater apparent haplotype diversity. One outlier cassava, BRA 856 (asterisked), groups among the *M. esc. flabellifolia*, suggesting possible maternal ancestry/admixing with *M. esc. flabellifolia*. (**b**) Principal component analysis based on SNVs revealing distinct clusters of nuclear genome types associated with *M. glaziovii* (blue), cultivated cassava and some *M. esc. flabellifolia* (orange), and the remaining *M. esc. flabellifolia* (gray). The fraction of population variance explained by each principal component is in parentheses. (**c**) Model-based clustering of nuclear genomes identifies the same groupings as principal component analysis, and identifies some accessions as admixed. Each vertical bar represents the fraction of an individual's genome attributable to one or more hypothetical ancestral populations. Note, for example, that Tree Cassava lies between clusters in **b** and is identified as admixed in **c**. Color key as in **b**. (**d–h**) Histograms of SNV heterozygosity (gray) and homozygous non-reference SNVs (blue) in 500 kb windows for cultivated cassava accession Albert (**d**), *M. esc. flabellifolia* FLA 433-2 (**e**), *M. esc. flabellifolia* FLA 444-1 (**f**), *M. glaziovii*(R) (**g**), and the "wild cassava" W14 (**h**). Note the similarity between *M. glaziovii* and W14, and between FLA 433-2 and Albert.

Figure 3 Segmental ancestry of selected *Manihot* accessions. (**a**) Inferred ancestry of 18 admixed individuals determined from whole genome shotgun sequencing data. Orange indicates *M. esculenta* genotype (E/E); light blue indicates *M. glaziovii* (G/G); light green represents hybrid *M. glaziovii*/*M. esculenta* (G/E). Dark green or black indicates presence of a shared *M. glaziovii* haplotype proposed to be inherited from the Amani program (G$_A$). Teal segments in MBRA 685 and MCOL 1468 on chromosome 2 behave anomalously and do not fit a model of *M. glaziovii*/*M. esculenta* admixture, but are likely hybrids of *M. esculenta* and another unknown *Manihot* species (E/U) (see **b**, or **Supplementary Note 7**). Light gray segments indicate no ancestry call could confidently be made. (**b**,**c**) Clustering of *M. glaziovii* and *M. esculenta* haplotypes in chromosome 1 from 30.1 to 32.6 Mb (**b**) and chromosome 1 from 22 to 23 Mb (**c**), showing haplotype sharing among six of seven African cassava varieties and among three South American cassava varieties, respectively. (**d**) Introgression plot, as in **a**, for accessions sequenced by GBS with 1% detected introgression or greater. Accessions are divided by population. The shared Amani haplotype appears enriched in the TMe and TMS populations.
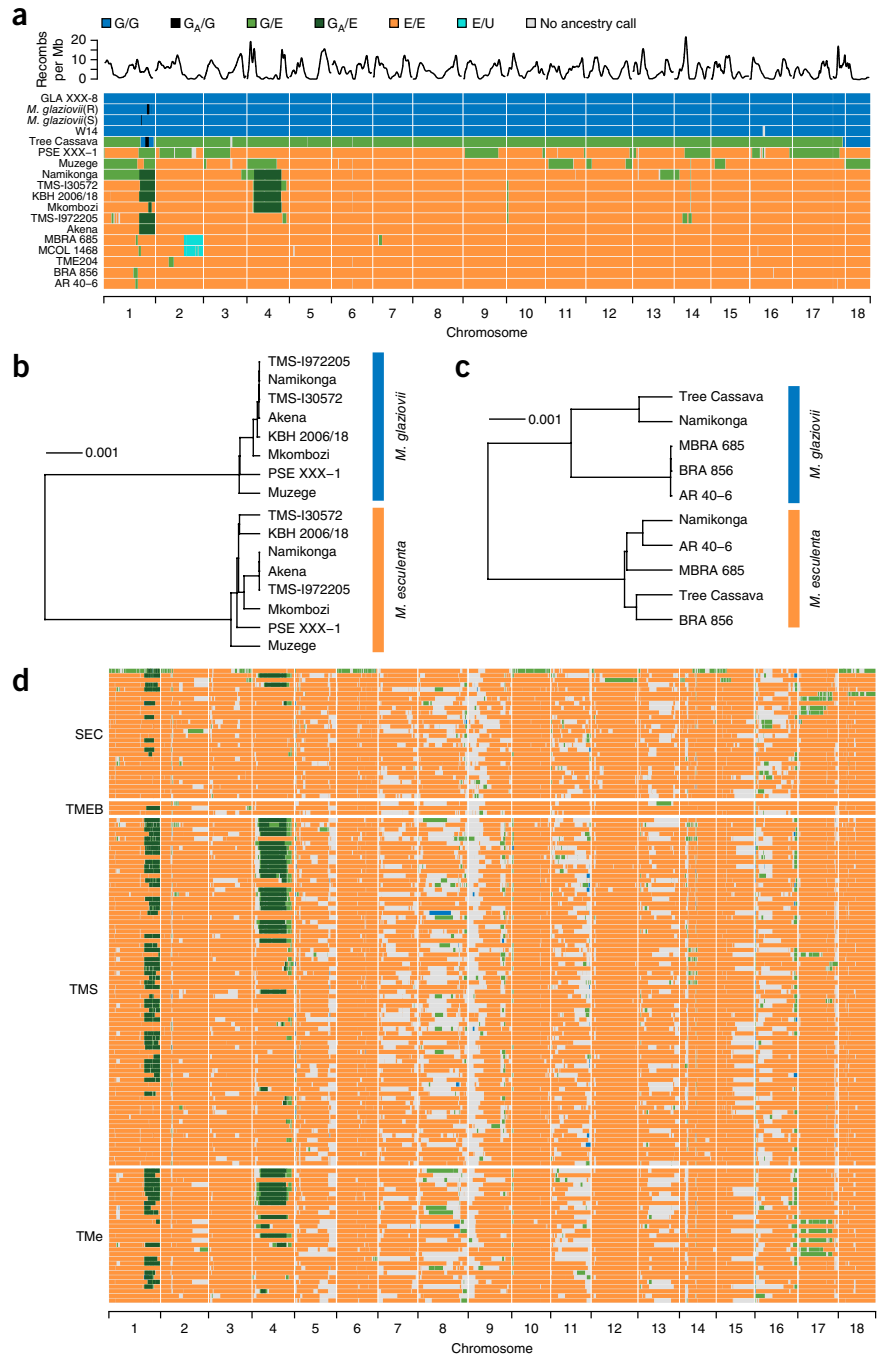
but, in contrast to the C/F group, exhibit a substantially higher level of homozygous differences relative to the cassava reference AM560-2 (0.89% for F versus 0.44% for C/F; **Fig. 2f** and **Supplementary Notes 6** and **7**). This supports the identification of F as representing a subpopulation of *M. esculenta* differentiated from cultivated cassava, although in principal component analyses they form a broad distribution and show considerable heterogeneity. The *M. esc. flabellifolia* accessions in our F group are from the central Brazilian states of Goiás and Tocantins in the southern Amazon region, which were differentiated from cassava in the studies of Olsen and Schaal[3–5]. FLA 449-1, from Mato Grosso, lies between the F and C/F groups and is a mixed type according to FRAPPE (**Fig. 2c**). The second principal component characterizes interspecific variation within *M. esculenta*, and is correlated with the distance from the center of domestication (**Supplementary Note 6**, Supplementary Fig. 12). The discrete separation between C/F and F may be an artifact[31] of our limited geographic sampling of *M. esc. flabellifolia,* and we suspect, based on the findings of Olsen and Schaal[3–5], that additional sampling would lead to a continuum representing the full intraspecific diversity of *M. esculenta*. In contrast to cultivated cassava accessions, wild *M. esc. flabellifolia* shows considerable cpDNA diversity, and no two samples in our collection share the same chloroplast haplotype, suggesting that we have not yet saturated coverage of wild *M. esculenta* cpDNA diversity.

Finally, the G cluster of *Manihot* genomes, which includes the three *M. glaziovii* accessions, is strongly differentiated from the cassava

reference (2.2% homozygous differences at genotyped positions; heterozygosity 0.71%; **Fig. 2g** and **Supplementary Notes 6** and **7**), and have related cpDNAs that are quite distinct (estimated divergence ~2–3 mya; **Supplementary Note 6**), from *M. esculenta*, as expected for accessions from a different species.

Notably, the "wild cassava" W14 accession, which was put forward as a genomic reference for "*M. esculenta* ssp. *flabellifolia*" by Wang *et al.*[27] groups with our G cluster of *M. glaziovii* accessions based on both nuclear and cpDNA genome analyses (**Fig. 2a–c,h**). Wang *et al.*[27] note that W14 is unusual in that it "produces a large number of fruits and is propagated only by seeds" and has a "lower rate of photosynthesis [than cassava] and very low storage root yield and starch content of the storage root." Our analysis suggests that
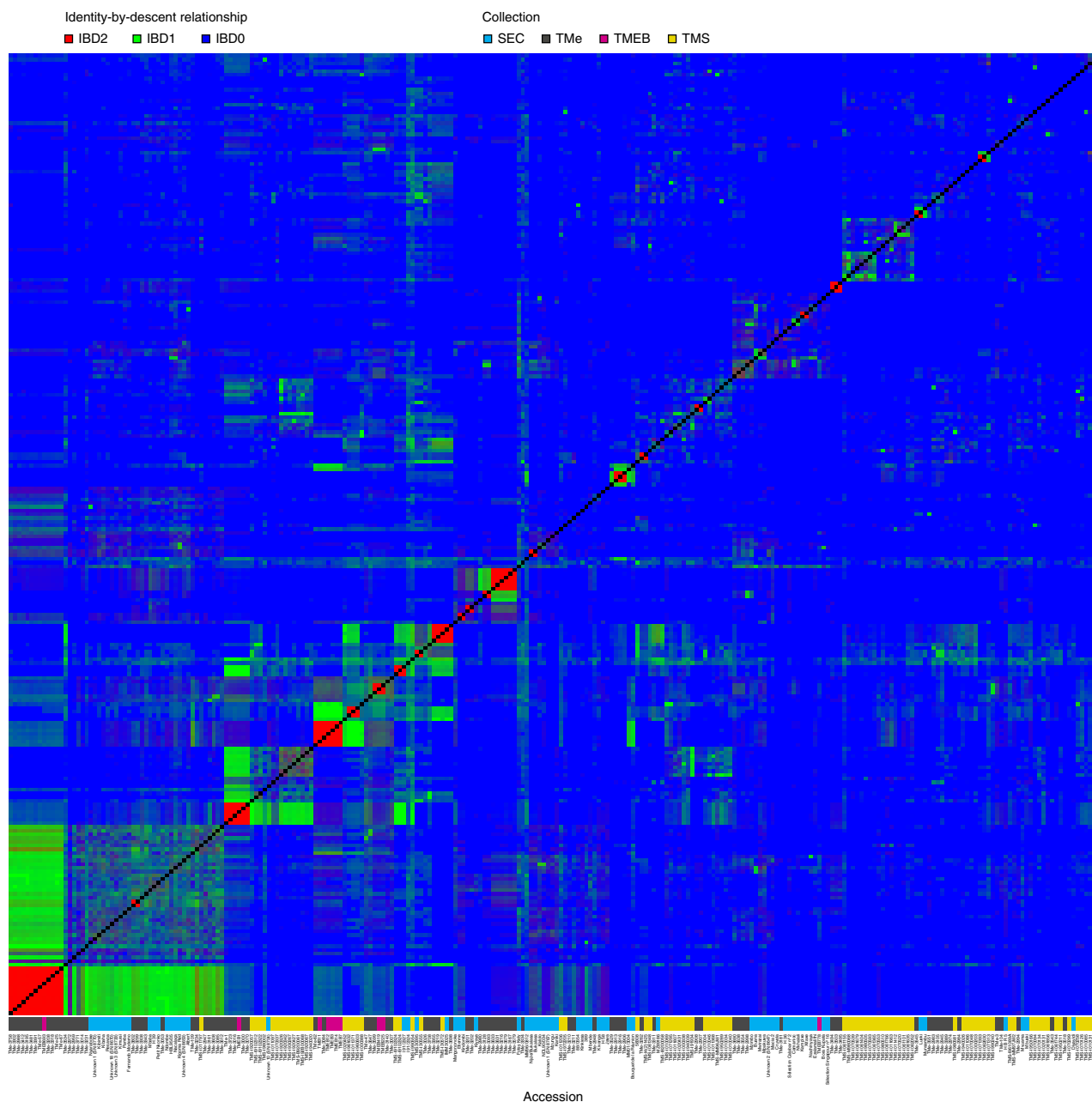
Figure 4 Identity-by-descent (IBD) relatedness between GBS samples. A heatmap is shown for IBD between 258 samples over 11,906 SNPs. More saturated colors indicate higher levels of IBD. The accessions are highlighted by collection and clustered so that those with similar relationships are closer together in the plot. Groups of samples that have identical genotypes at our markers appear as bright red boxes near the diagonal (**Supplementary Note 9**, Supplementary Table 14); bright green signals indicate likely first-degree relationships. See **Table 2** for collection descriptions.

the W14 sequence presented in Wang *et al.*[27] is in fact from an *M. glaziovii* accession, and that the diversity analysis presented in their study is dominated by interspecific variation rather than cassava domestication.

Introgression and cassava diversity

We find widespread evidence for interspecific hybridization[22] and introgression, with mixed ancestry in cassava and its relatives, based on FRAPPE (**Fig. 2c**), intermediate position in principal component analysis (**Fig. 2b**) and genomic segments of high heterozygosity (as would be expected in interspecific hybrids; **Fig. 3a**). To resolve admixture events along chromosomes, we identified 1,055,571 biallelic

ancestry-informative single-nucleotide markers that represent fixed, or nearly fixed, differences between *M. esculenta* (C/F plus F, together denoted as E) and *M. glaziovii*, and assigned segmental ancestry as either diploid *M. esculenta* (E/E), diploid *M. glaziovii* (G/G), or hybrid (G/E) using a maximum likelihood method (**Fig. 3**, **Supplementary Note 7** and **Supplementary Datasets 2** and **3**). We were unable to assemble a sufficiently comprehensive set of variants to allow assignment of C/F or F ancestry across the genome, consistent with analysis of population structure in **Supplementary Note 6**.

For example, "tree" cassava, grown around homesteads in Africa and whose leaves are eaten as a vegetable, is widely believed to be a natural hybrid of cassava and *M. glaziovii*[12,18,22]. Our analysis

confirms this ancestry, with (at least for our Tree Cassava from Tanzania) cassava as the maternal parent, consistent with FRAPPE and principal component analysis. Whereas most of the genome is a hybrid of *M. esculenta*/*M. glaziovii*, the right arms of chromosome 1 and 18 are derived only from *M. glaziovii* (**Fig. 3a**). This is consistent with a widespread introgression of *M. glaziovii* into African cassava, as detailed below.

Surprisingly, we find that the genome of a Brazilian accession designated "*M. pseudoglaziovii* Pax. & Hoffm.," which was thought to be a separate species[33], is an interspecific admixture of *M. esculenta* and *M. glaziovii*. The evidence from our investigation is consistent with a second-generation backcross into *M. esculenta* from an *M. glaziovii* maternal great-grandmother (**Supplementary Note 7**). *Manihot* taxonomists have described up to 98 separate species in the genus[34,35]. Our results raise the possibility that some of these species may be interspecific hybrids or admixtures.

Two outliers in our analyses are the South American cassavas MBRA 685 and MCOL 1468, which both have long segments (overlapping over 13.2 Mb of chromosome 2) whose ancestry could not be confidently assigned based on our collection of *M. esculenta* and *M. glaziovii* alleles. These segments are (i) highly heterozygous (mean 2.2%) and (ii) enriched in variant alleles that are not found elsewhere within our collection (0.93% of genotyped sites in segments), but are shared between the two accessions (56.3% of rare alleles are shared in the overlapping region) (**Supplementary Note 7**, Supplementary Fig. 17). These segments may be introgressions of an as-yet unidentified third *Manihot* species into cassava[3,36] (teal segments, **Fig. 3a**). The unique variants shared by these two cassavas can be used to query future collections of *Manihot* sequences.

Introgression of *M. glaziovii* into cassava

We find that seven cultivated African cassava accessions arose by introgression of *M. glaziovii* into *M. esculenta* (Namikonga, Akena, Mkombozi, TMS-I972205, KBH 2006/18, TMS-I30572, and Muzege; **Fig. 3a**). Six of the seven (all but Muzege) share a common *M. glaziovii* haplotype on chromosome 1 (**Fig. 3b**); four of these (all of these except TMS-I972205 and Akena) also share a common *M. glaziovii* haplotype on chromosome 4 (**Supplementary Note 7**). In the 1930s and 1940s, the Amani breeding program in Tanzania intentionally introgressed *M. glaziovii* into cassava germplasm with the aim of transferring CMD resistance; CBSD resistance was a secondary trait[12]. Of our sequenced accessions, the CBSD-resistant but CMD-susceptible Namikonga, the CBSD-susceptible but CMD-tolerant TMS-I30572 (ref. 37), and the TMS-I30572 descendent TMS-I972205 are known to be derived from the Amani program. Our analysis suggests that the other introgressed African cassava accessions also derive from Amani germplasm. The number and size of the *M. glaziovii*/*M. esculenta* hybrid segments of many of these accessions are consistent with having one or two *M. glaziovii* great-great-grandparents. Our Tree Cassava, isolated from Tanzania, appears to be a cross between *M. glaziovii* and an introgressed cassava, because in this region of the genome both haplotypes are of *M. glaziovii* type. Tree Cassava and two escaped East African *M. glaziovii* also possess short segments of the Amani haplotype (**Fig. 3a**), consistent with shared ancestry.

Unexpectedly, three South American cassava cultivars (BRA 856, MBRA 685, and MCOL 1468), and one known derivative of crosses between South American and Nigerian germplasm (AR 40-6), also show *M. glaziovii* introgression (**Fig. 3a**), but with a smaller fraction of admixture than the African Amani-derived cultivars. Three of the four

(AR 40-6, BRA 865, MBRA 685), however, share a common *M. glaziovii* haplotype in the 22–23 Mb region on chromosome 1 (**Fig. 3c**). Thus, it is possible that *M. glaziovii* introgression has also occurred as part of South American breeding programs[36], or that these programs have incorporated undocumented introgressed African germplasm.

Comparing these *M. glaziovii* markers to our collection of 268 genotyped African cassava accessions, we find that the same introgressed Amani segments are widespread among TMS elite lines, TMEB breeder lines, and TMe landraces, but are rare in farmer varieties from southern, eastern, and central Africa (SEC collection), presumably because those accessions arose from farmer selection rather than breeding programs (**Fig. 3d**). In most cases, these introgressed accessions share a common haplotype. We hypothesize that these shared segments, which include 285 and 206 genes on chromosomes 1 and 4, respectively (**Supplementary Datasets 4** and **5**), may contain desirable *M. glaziovii* CMD/CBSD resistance gene(s) transferred in the Amani program, although the differential disease resistance among these cultivars may also implicate other introgressed segments, and other traits may be involved. *M. glaziovii* alleles in these regions can be used as markers to track these segments in further breeding efforts.

DISCUSSION

Our analyses reveal relationships among cultivated cassava that will aid in developing diverse germplasm for breeding. Many differently named accessions are near-clones based on genome-wide identity, although they may harbor accumulated somatic mutations (**Supplementary Note 8**). Other accessions are common first- or second-degree relatives and are hubs in the relatedness network (**Supplementary Note 8**, Supplementary Table 13, Supplementary Fig. 20). GBS-based analysis of a broader sampling of African accessions confirms the prevalence of first- or second-degree identity by descent (**Fig. 4** and **Supplementary Note 9**). The recurrent use of a small number of genotypes as parents in breeding efforts, in part due to poor flowering in many landraces or cultivars, has reduced the genetic diversity of cassava, especially in Africa. Knowledge of these relationships will guide breeding decisions to restore lost variation.

Early in its domestication cassava experienced a strong maternal bottleneck, as revealed by limited global chloroplast diversity relative to the wild progenitor species. Interspecific introgression, however, has injected new variation into the nuclear genome, both through organized breeding programs and through what appears to be natural introgression. In Africa, specific *M. glaziovii* haplotypes introduced by organized breeding programs are widespread among preferred varieties (**Fig. 3d** and **Supplementary Note 9**, Supplementary Fig. 22), and they likely encode desired traits. These haplotypes are also found in farmer varieties from throughout Africa, presumably spread by undocumented crosses. These introgressed segments span substantial fractions of chromosomes, and additional effort will be needed to break these linkages and pinpoint causal variants. At least one unknown species of *Manihot* has contributed to the genetic diversity of cultivated South American cassava, suggesting the profitability of exploring additional interspecific breeding.

The variants and population structure described here are essential inputs for marker-assisted and genomic selection-based approaches to improving disease resistance and yield for this staple crop[38,39]. Large-scale breeding efforts, such as the NextGen Cassava program[40,41], will need to incorporate the impact of common introgressions in predictive genotype–phenotype models to realize the full power of genome-enabled approaches.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession Codes. All *Manihot* whole genome shotgun sequence, plus mate pair and fosmid sequence used for AM560-2 genome assembly, as well as the v6.1 AM560-2 genome assembly itself, may be found under BioProject PRJNA234389. Diversity GBS sequence is deposited in BioProject PRJNA234391. The v6.1 AM560-2 genome assembly described in this paper is also available at Phytozome (https://phytozome.jgi.doe.gov/Mesculenta).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS

D.S.R., S.R., and M.E.F. designed the study and provided scientific leadership of project. D.S.R., J.V.B., J.B.L., and S.E.P. coordinated sequencing and analysis efforts. I.Y.R., C.E., P.N., V.L., J.N., G.M., R.S.B., T.L.S., R.M.G., and M.E.F. provided *Manihot* samples for sequencing. J.B.L. led molecular biology associated with the project with early contributions from E.E.-G. J.G. and J.S. were responsible for AM560-2 mate-pair sequencing. J.V.B. assembled the reference genome, integrated genetic maps with assembly, defined variant genotypes, performed population genetic analyses, analyzed cpDNAs, and analyzed admixture. S.E.P. annotated genome and performed paleotetraploidy analyses. G.A.W. performed whole genome shotgun sequencing admixture analysis, developed interspecific phasing and haplotype sharing method, and contributed to population structure analysis and population genetic modeling. J.V.B., S.E.P., and C.M.H. analyzed GBS data. D.S.R., J.V.B., J.B.L., S.E.P., and G.A.W. wrote the manuscript with input from S.R., M.E.F., P.K., R.M.G., R.S.B., C.E., and I.Y.R. All authors reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Howeler, R., Lutaladio, N. & Thomas, G. *Save and Grow: Cassava: a Guide to Sustainable Production Intensification* (Food and Agriculture Organization of the United Nations, 2013).
2. Allem, A.C. in *Cassava: Biology, Production and Utilization* (eds. Hillocks, R.J., Thresh, J.M. & Bellotti, A.C.) 1–16 (CABI, 2002).
3. Olsen, K.M. & Schaal, B.A. Evidence on the origin of cassava: phylogeography of *Manihot esculenta*. *Proc. Natl. Acad. Sci. USA* **96**, 5586–5591 (1999).
4. Olsen, K. & Schaal, B. Microsatellite variation in cassava (*Manihot esculenta*, Euphorbiaceae) and its wild relatives: further evidence for a southern Amazonian origin of domestication. *Am. J. Bot.* **88**, 131–142 (2001).
5. Olsen, K.M. SNPs, SSRs and inferences on cassava's origin. *Plant Mol. Biol.* **56**, 517–526 (2004).
6. Nassar, N.M. Conservation of the genetic resources of cassava (*Manihot esculenta*) determination of wild species localities with emphasis on probable origin. *Econ. Bot.* **32**, 311–320 (1978).
7. Nassar, N.M. Cassava, *Manihot esculenta* Crantz, genetic resources: origin of the crop, its evolution and relationships with wild relatives. *Genet. Mol. Res.* **1**, 298–305 (2002).
8. Ugent, D., Pozorski, S. & Pozorski, T. Archaeological manioc (*Manihot*) from coastal Peru. *Econ. Bot.* **40**, 78–102 (1986).
9. Rival, L. & McKey, D. Domestication and diversity in manioc (*Manihot esculenta* Crantz ssp. *esculenta*, Euphorbiaceae). *Curr. Anthropol.* **49**, 1119–1128 (2008).
10. Clement, C.R., de Cristo-Araújo, M., Coppens D'Eeckenbrugge, G., Alves Pereira, A. & Picanço-Rodrigues, D. Origin and domestication of native Amazonian crops. *Diversity (Basel)* **2**, 72–106 (2010).
11. Jones, W.O. *Manioc in Africa* (Stanford University Press, Stanford, USA, 1959).
12. Jennings, D.L. in *African Cassava Mosaic* (ed. Nestel B.L.) 39–44 (International Development Research Centre, Bogota, 1976).
13. Nweke, F.I. *New Challenges in the Cassava Transformation in Nigeria and Ghana*, Vol. 118 (Intl Food Policy Res Inst, 2004).
14. Kawuki, R., Nuwamanya, E., Labuschagne, M., Herselman, L. & Ferguson, M. Genetic effects of inbreeding on harvest index and root dry matter content in cassava. Second RUFORUM Biennial Regional Conference on "Building capacity for food security in Africa," Entebbe, Uganda, 20–24 September 2010 (eds. Adipala, E., Tusiime, G. & Majaliwa, J.G.M.) 377–381 (RUFORUM, 2010).
15. Alabi, O.J., Kumar, P.L. & Naidu, R.A. Cassava mosaic disease: A curse to food security in Sub-Saharan Africa. *Online. APSnet Features.* doi:10.1094/APSnetFeature-2011-0701 (2011).
16. Lozano, J. & Sequeira, L. Bacterial blight of cassava in Colombia: epidemiology and control. *Phytopathology* **64**, 8 (1974).
17. Fu, Y.-B., Wangsomnuk, P.P. & Ruttawat, B. Thai elite cassava genetic diversity was fortuitously conserved through farming with different sets of varieties. *Conserv. Genet.* **15**, 1463–1478 (2014).
18. Nichols, R. Breeding cassava for virus resistance. *East Afr. Agric. J.* **12**, 184–194 (1947).
19. Elshire, R.J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**, e19379 (2011).
20. Prochnik, S. *et al.* The cassava genome: current progress, future directions. *Trop. Plant Biol.* **5**, 88–94 (2012).
21. Putnam, N.H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
22. Second, G. & Iglesias, C. in *Broadening the Genetic Base of Crop Production* (eds. Cooper, H.D., Spillane, C. & Hodgkin, T.) 201221 (CABI, 2001).
23. International Cassava Genetic Map Consortium (ICGMC). High-resolution linkage map and chromosome-scale genome assembly for cassava (*Manihot esculenta* Crantz) from 10 populations. *G3 (Bethesda)* **5**, 133–144 (2014).
24. Rabbi, I. *et al.* Genetic mapping using genotyping-by-sequencing in the clonally propagated cassava. *Crop Sci.* **54**, 1384–1396 (2014).
25. Rabbi, I.Y. *et al.* High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding. *Virus Res.* **186**, 87–96 (2014).
26. Umanah, E.E. & Hartmann, R.W. Chromosome numbers and karyotypes of some Manihot species. *J. Am. Soc. Hortic. Sci.* **98**, 272–274 (1973).
27. Wang, W. *et al.* Cassava genome from a wild ancestor to cultivated varieties. *Nat. Commun.* **5**, 5110 (2014).
28. Jennings, D. Variation in pollen and ovule fertility in varieties of cassava, and the effect of interspecific crossing on fertility. *Euphytica* **12**, 69–76 (1963).
29. De Carvalho, R. & Guerra, M. Cytogenetics of *Manihot esculenta* Crantz (cassava) and eight related species. *Hereditas* **136**, 159–168 (2002).
30. Rahman, A.Y. *et al.* Draft genome sequence of the rubber tree *Hevea brasiliensis*. *BMC Genomics* **14**, 75 (2013).
31. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
32. Tang, H., Peng, J., Wang, P. & Risch, N.J. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* **28**, 289–301 (2005).
33. Pax, F. & Hoffman, K. in *Das Pflazenreich IV*, Vol. 147 XVI. (ed. Engler A.) 196 (Wilhelm Engelmann, Leipzig, 1924).
34. Rogers, D. & Appan, S. Flora neotropica monograph no. 13. Manihot, Manihotoides (Euphorbiaceae). *New York: Hafner 275p. Illustrations, portraits, dot maps* (1973).

35. Nassar, N.M., Hashimoto, D.Y. & Fernandes, S.D. Wild *Manihot* species: botanical aspects, geographic distribution and economic value. *Genet. Mol. Res.* **7**, 16–28 (2008).
36. Nassar, N.M. Broadening the genetic base of cassava, *Manihot esculenta* Crantz, by interspecific hybridization. *Can. J. Plant Sci.* **69**, 1071–1073 (1989).
37. Lokko, Y., Dixon, A., Offei, S. & Danquah, E. Genetic relationships among improved cassava accessions and landraces for resistance to the cassava mosaic disease. *J. Food Agric. Environ.* **7**, 156–162 (2009).
38. Ferguson, M. *et al.* Molecular markers and their application to cassava breeding: past, present and future. *Trop. Plant Biol.* **5**, 95–109 (2012).
39. Jannink, J.L., Lorenz, A.J. & Iwata, H. Genomic selection in plant breeding: from theory to practice. *Brief. Funct. Genomics* **9**, 166–177 (2010).
40. Fessenden, M. A cassava revolution could feed the world's hungry http://www.scientificamerican.com/article/a-cassava-revolution-could-feed-the-worlds-hungry/ *Sci. Am.* (24 March 2014).
41. Ceballos, H., Kawuki, R.S., Gracen, V.E., Yencho, G.C. & Hershey, C.H. Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. *Theor. Appl. Genet.* **128**, 1647–1667 (2015).
42. Sanchez, G. *et al.* AFLP assessment of genetic variability in cassava accessions (*Manihot esculenta*) resistant and susceptible to the cassava bacterial blight (CBB). *Genome* **42**, 163–172 (1999).
43. Duque, L. *Cassava Drought Tolerance Mechanisms Re-Visited: Evaluation of Drought Tolerance in Contrasting Cassava Genotypes Under Water Stressed Environments.* PhD thesis, Cornell Univ. (2012).
44. Nyaboga, E. *et al.* Unlocking the potential of tropical root crop biotechnology in east Africa by establishing a genetic transformation platform for local farmer-preferred cassava cultivars. *Front. Plant Sci.* **4**, 526 (2013).
45. Turyagyenda, L. *et al.* Genetic diversity among farmer-preferred cassava landraces in Uganda. *Afr. Crop Sci. J.* **20** (suppl. s1), 15–20 (2012).
46. Ogwok, E. *et al.* Transgenic RNA interference (RNAi)-derived field resistance to cassava brown streak disease. *Mol. Plant Pathol.* **13**, 1019–1031 (2012).
47. Kabeya, M.J., Kabeya, U.C., Bekele, B.D. & Ingelbrecht, I.L. Genetic Analysis of Selected Cassava (*Manihot esculenta*) Genetic Pool in Africa Assessed with Simple Sequence Repeats. *World J. Agric. Sci.* **8**, 637–641 (2012).
48. Taylor, N., Chavarriaga, P., Raemakers, K., Siritunga, D. & Zhang, P. Development and application of transgenic technologies in cassava. *Plant Mol. Biol.* **56**, 671–688 (2004).
49. Nweke, F., Spencer, D. & Lynam, J. *The Cassava Transformation: Africa's best kept secret* (Michigan State University, East Lansing, 2002).

ONLINE METHODS

Sequencing and assembly of AM560-2. Four Illumina whole genome shotgun fragment libraries were constructed from cassava accession AM560-2 DNA left over from Prochnik *et al.*[20], and sequenced on Illumina HiSeq with 250-bp forward and 200-bp reverse reads. Leaves were collected from AM560-2 plants and high molecular weight DNA prepared for fosmid, mate pair and Dovetail "Chicago" libraries. The former two of these were sequenced on Illumina MiSeq and the latter on HiSeq. Assembly of shotgun, mate-pair and fosmid sequences with Platanus (v1.2.1)[50]; further scaffolding by Dovetail Genomics (Santa Cruz, CA)[21], and anchoring to a composite genetic map[23] generated an assembly on 18 chromosomes. The shotgun assembly captures more than 98.5% of cassava's protein-coding genes based on comparison with EST sequences. See **Supplementary Note 1** for more detail.

Annotation. *De novo* repeat finding in the assembly was performed with RepeatModeler v1.0.8 (http://www.repeatmasker.org/RepeatModeler.html), followed by masking with Repeatmasker (http://www.repeatmasker.org). RNA-seq data, together with 454 and Sanger ESTs, were used to reconstruct transcripts which were combined with homology-based gene predictions with PASA[51] to make gene models (**Supplementary Note 2**). Of the 33,033 predicted protein-coding genes, 11,872 and 29,274 have evidence for transcription or homology, respectively, over more than 50% of their length. 31,895 predicted protein-coding genes (96.6%) and 518.5 Mb (89.0% of the assembled sequence) are mapped to a chromosomal position.

Whole genome duplication. Homologous segments were identified in the cassava genome by comparing all cassava proteins to each other and looking for runs of two or more paralogous genes (with up to six intervening genes) in separate regions of the cassava genome. Cassava genes in these duplicated regions were compared to proteins in *Ricinus*, *Hevea*, *Jatropha*, and *Populus*, and average corrected fourfold degenerate transversion (4DTv) rates were calculated between the species allowing reconstruction of a neighbor-joining phylogenetic tree and timing of species divergences, calibrated by fossil evidence. Average 4DTv from *Hevea* and cassava paralog pairs was used to place the whole genome duplication before speciation (**Supplementary Note 3**).

Global *Manihot* diversity. Tissue or DNA was obtained from 58 accessions of cassava and related *Manihot* from collections including South American, African, Asian, and Oceanian diversity (**Supplementary Note 4**). Whole genome shotgun fragment libraries were paired-end sequenced using Illumina HiSeq. The majority of libraries were sequenced with reads 200 bp or longer (**Supplementary Note 5**).

***Manihot* relatedness and haplotype ancestry.** A PhyML[52] maximum-likelihood phylogenetic tree was constructed from Malvidae chloroplast sequences aligned with DIALIGN[53], allowing timing of the divergence of *M. glaziovii* and *M. esculenta* (**Supplementary Note 6**). A minimal "pants" model[54] was used to calculate population genetic parameters of this divergence (**Supplementary Note 10**). SNVs were called by aligning reads to the reference genome with BWA-MEM[55] and genotyping with the HaplotypeCaller tool from GATK[56,57]. smartpca[31] and FRAPPE[32] software were used to estimate ancestral proportions (**Supplementary Note 6**). Pure individuals were used to identify ancestry-diagnostic SNVs. These SNVs were used to determine admixture in cassava accessions (**Supplementary Note 7**). IBD and $\hat{\pi}$ were calculated with PLINK[58] software to classify relatedness (e.g., parent-offspring, full sibling; see **Supplementary Note 8**).

Genotyping-by-sequencing of diverse African cassava. SNV genotypes were called from 271 accessions from three collections using GBS[23] with BWA[59] and the HaplotypeCaller tool from the GATK software package. IBD was calculated with PLINK (**Supplementary Note 9**).

50. Kajitani, R. *et al.* Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
51. Haas, B.J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
52. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).
53. Morgenstern, B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211–218 (1999).
54. Wu, G.A. *et al.* Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.* **32**, 656–662 (2014).
55. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at http://arxiv.org/abs/1303.3997 (2013).
56. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
57. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
58. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
59. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).