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

Jessen V Bredeson1, Jessica B Lyons1, Simon E Prochnik2, G Albert Wu2, Cindy M Ha1,16, Eric Edsinger-Gonzalez1,16, Jane Grimwood2,3, Jeremy Schmutz2,3, Ismail Y Rabbi4, Chiedozie Egesi5, Poasa Nauluvula6,16, Vincent Lebot7, Joseph Ndunguru8, Geoffrey Mkamilo9, Rebecca S Bart10, Tim I Setter11, Roslyn M Gleadow12, Peter Kulakow4, Morag E Ferguson13, Steve Rounsley14 & Daniel S Rokhsar1,2,15

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 Ceára 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 regions1. Although domesticated over 6,000 years ago2–5, cassava cultivation spread beyond South America only in the past 500 years, exported by European colonialists and slave traders11. 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 yield12,13.

Cassava can outcross but is commonly clonally propagated, and harbors considerable genetic load14. 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 blight15,16. In contrast to African varieties, Thai elite varieties retain considerable diversity17. 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).

1Department of Molecular and Cell Biology, University of California, Berkeley, California, USA. 2United States Department of Energy Joint Genome Institute (DOE JGI), Walnut Creek, California, USA. 3TudorAlpha Institute for Biotechnology, Huntsville, Alabama, USA. 4International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria. 5National Root Crops Research Institute (NRCRI), Umudike, Nigeria. 6Department of Agriculture, Ministry of Primary Industries, Koronivia Research Station, Fiji. 7Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Port-Vila, Vanuatu. 8Mikocheni Agricultural Research Institute (MARI), Dar es Salaam, Tanzania. 9Naliende Agricultural Research Institute (NARI), Mtwara, Tanzania. 10Donald Danforth Plant Science Center, St. Louis, Missouri, USA. 11Section of Soil and Crop Sciences, School of Integrative Plant Science, Cornell University, Ithaca, New York, USA. 12School of Biological Sciences, Monash University, Melbourne, Victoria, Australia. 13International Institute of Tropical Agriculture (IITA), Nairobi, Kenya. 14Dow AgroSciences, Indianapolis, Indiana, USA. 15Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa. 16Present addresses: Anschutz Medical Campus, University of Colorado, Denver, Colorado, USA (C.M.H.), Marine Biological Laboratory, Woods Hole, Massachusetts, USA (E.E.-G.); South Pacific Community, Kolonia, Federated States of Micronesia (P.N.). Correspondence should be addressed to D.S.R. (dsrokhsar@gmail.com) or J.V.B. (jessenbredeson@berkeley.edu).

Received 4 September 2015; accepted 10 March 2016; published online 18 April 2016; doi:10.1038/nbt.3535

NATURE BIOTECHNOLOGY ADVANCE ONLINE PUBLICATION

RESOURCES
In this report we use “cassava” to refer to cultivated and/or domesticated varieties of *Manihot esculenta*, and the shorthand *M. esculenta* for wild accessions. We also shotgun-sequenced five *Manihot* accessions related to cassava, including three from the wild species *M. glaziovii* (Arg., one named *M. pseudoglaziovii* Pax & K. Hoffman, and “tree” cassava, a suspected hybrid sometimes called *M. catingea*) from Brazil, one 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. To analyze genetic variation present in African varieties, we also characterized 268 cultivars of cassava using reduced representation genotyping-by-sequencing (GBS) (Table 2).

**RESULTS**

**Chromosome structure**

To produce a high-quality chromosome-scale reference genome for cassava, we augmented our earlier draft sequence 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 (Methods and Supplementary Note 1). AM560-2 is a S3 line bred at Centro Internacional de Agricultura Tropical (CIAT) from MCOL1505 (also known as Manihocia P-12 (ref. 22). Compared with the previous draft, 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 (Supplementary Note 1). To organize the sequence into chromosomes we integrated the shotgun assembly with a 22,403-marker consensus genetic map and two other recently published maps 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**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Location</th>
<th>Accession</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>“M. flabellifolia”</td>
<td>Brazil</td>
<td>FLA 496-1</td>
<td>Brazil43</td>
</tr>
<tr>
<td>FLA 449-1</td>
<td>Brazil43</td>
<td>FLA 503-2</td>
<td>Brazil44</td>
</tr>
<tr>
<td>FLA 433-2</td>
<td>Brazil43</td>
<td>FLA 490-1</td>
<td>Brazil45</td>
</tr>
<tr>
<td>FLA 444-1</td>
<td>Brazil43</td>
<td>FLA 502-1</td>
<td>Brazil46</td>
</tr>
<tr>
<td>FLA 488-1</td>
<td>Brazil43</td>
<td>FLA XXX-15</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Location</th>
<th>Accession</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM560-2</td>
<td>Colombia</td>
<td>CM 507-37</td>
<td>Colombia</td>
</tr>
<tr>
<td>MCOL22</td>
<td>Colombia</td>
<td>BRA 856</td>
<td>Brazil44</td>
</tr>
<tr>
<td>MBR 685</td>
<td>Brazil44</td>
<td>CM 3306-4</td>
<td>Brazil45</td>
</tr>
<tr>
<td>MCOL 1468</td>
<td>Brazil44</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>“M. flabellifolia”</td>
<td>Brazil</td>
</tr>
<tr>
<td>FLA 496-1</td>
<td>Brazil43</td>
</tr>
<tr>
<td>FLA 503-2</td>
<td>Brazil44</td>
</tr>
<tr>
<td>FLA 490-1</td>
<td>Brazil45</td>
</tr>
<tr>
<td>FLA 502-1</td>
<td>Brazil46</td>
</tr>
<tr>
<td>FLA XXX-15</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

Asian and Asian Pacific cassava varieties from China, Australia, Fiji, and Vanuatu. In addition, our analysis includes KU50 from Thailand27, 12 sequenced, 10 distinct.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC8</td>
<td>China</td>
</tr>
<tr>
<td>SM1150</td>
<td>Australia</td>
</tr>
<tr>
<td>Avoca</td>
<td>Australia</td>
</tr>
<tr>
<td>Nadele(B) (Nadele(U))</td>
<td>Fiji</td>
</tr>
</tbody>
</table>

Mixed crosses, Improved varieties known to be derived from intentional crosses between wild, South American, African, and/or Asian varieties. 4 sequenced, 4 distinct.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR 40-6</td>
<td>Tanzania/Nigeria</td>
</tr>
<tr>
<td>AR 37-80</td>
<td>Nigeria/Thailand</td>
</tr>
</tbody>
</table>

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. In addition, our analysis includes the “wild cassava” W14 from Wang et al., which we identified as *M. glaziovii*. 5 sequenced, 5 distinct.

**Table 2**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. glaziovii(R)</td>
<td>Tanzania</td>
</tr>
<tr>
<td>M. glaziovii(S)</td>
<td>Tanzania</td>
</tr>
<tr>
<td>M. glaziovii GLA XXX-8</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

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. See Supplementary Dataset 1 for detailed sourcing, provenance, and phenotype information; SRA BioSample accession numbers; and sequencing statistics.

© 2016 Nature America, Inc. All rights reserved.
Comparative analyses revealed the impact of paleotetraploidy 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 \( \sim 35 \) 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 homologs (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 homoyzgosity) that is approximately twice the rate of homoyzgous 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

Table 2  Cassava accessions genotyped by sequencing

<table>
<thead>
<tr>
<th>Collection</th>
<th>Sequenced (268 total)</th>
<th>Distinct (215 total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMe (Tropical Manihot esculenta) cassava accessions from the IITA genebank core collection</td>
<td>97</td>
<td>62</td>
</tr>
<tr>
<td>TMEB germplasm accessions used in the IITA breeding program</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>TMS (Tropical Manioc Selection) cassava varieties developed and selected by IITA</td>
<td>88</td>
<td>84</td>
</tr>
<tr>
<td>SEC: southern, eastern, and central African cassava diversity</td>
<td>73</td>
<td>66</td>
</tr>
</tbody>
</table>

IITA, International Institute of Tropical Agriculture.
The grouping with cassava is consistent with the haplotype analyses of Olsen and Schaal, 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 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.
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 Schaal3–5. 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 of our limited geographic sampling of *M. esc. flabellifolia*, and we suspect, based on the findings of Olsen and Schaal3–5, 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
the W14 sequence presented in Wang et al. 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.

**Introggression and cassava diversity**

We find widespread evidence for interspecific hybridization and introgression, with mixed ancestry in cassava and its relatives, based on FRAPPE (Fig. 2c) 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*. 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, 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 genus34,35. 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 cassava35-36 (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 Muzenge; Fig. 3a). Six of the seven (all but Muzenge) 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 trait12. 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 programs38,39, 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 crop38,39. Large-scale breeding efforts, such as the NextGen Cassava program40,41, 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 assembled in this paper is also available at Phytozome (https://phytozome.jgi.doe.gov/Mesocentella).

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

ACKNOWLEDGMENTS

We thank K. Swaminathan for advice on and protocols for DNA isolation; J. Burke for collecting the AM560-2 material, and J. Vrebalov for preparing the DNA used for fosmid, mate pair, and Dovetail “Chicago” libraries; M. Hall for early project planning; M. Chung, J. Choi, K. Lundy, and other members of the VCGSL at UC Berkeley for advice and technical assistance with Illumina library preparation and sequencing; J. Galina-Mehlman and J. Still at the University of Arizona Genetics Core for library preparation and sequencing; R. McEwan and C. Evans for sequencing performed at Dow AgroSciences; I.B. Boston for mate pair library construction at HudsonAlpha; N. Putnam and J. Stites at Dovetail Genomics for performing HiSeq assembly; R. Keough and Lucien for fosmid library construction and sequencing; R. Hyde for providing cassava tissue from the Setter laboratory; M. Cohn and the Staskawicz laboratory for cassava tissue; C. Hershey and L.A.B. López-Lavalle for permission to sequence cassava plants for the Setter Lab; M. Cohn and the Staskawicz laboratory for cassava tissue; C. Hershey and L.A.B. López-Lavalle for permission to sequence cassava plants for the Setter Lab; 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 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.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/.


29. de Carvalho, R. & Guerra, M. Cytogenetics of Manihot esculenta Crantz (cassava) and eight related species. Hereditas 136, 159–168 (2002).


ONLINE METHODS
Sequencing and assembly of AM560-2. Four Illumina whole genome shotgun fragment libraries were paired-end sequenced from cassava accession AM560-2 DNA left over from Prochnik et al., 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)23, and anchoring to a composite genetic map23 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 PASA23 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 accesses 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 PhyML52 maximum-likelihood phylogenetic tree was constructed from Malvidae chloroplast sequences aligned with DIALIGN23, allowing timing of the divergence of M. esculenta and M. glaziovii and M. esculenta (Supplementary Note 6). A minimal “pants” model54 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-MEM55 and genotyping with the HaplotypeCaller tool from GATK56,57. smartpca51 and FRAPPE52 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 F were calculated with PLINK58 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 GBS53 with BWA59 and the HaplotypeCaller tool from the GATK software package. IBD was calculated with PLINK (Supplementary Note 9).