Lawrence Berkeley National Laboratory

Title

Chromosome evolution and the genetic basis of agronomically important traits in greater yam

Permalink

https://escholarship.org/uc/item/6106q6wb

Journal

Nature Communications, 13(1)

ISSN

2041-1723

Authors

Bredeson, Jessen V
Lyons, Jessica B
Oniyinde, Ibukun O
et al.

Publication Date

2022

DOI

10.1038/s41467-022-29114-w

Peer reviewed

Check for updates

Chromosome evolution and the genetic basis of agronomically important traits in greater yam

Jessen V. Bredeson [1,18], Jessica B. Lyons [1,2,18], Ibukun O. Oniyinde[3], Nneka R. Okereke[4], Olufisayo Kolade[3], Ikenna Nnabue[4], Christian O. Nwadili[4], Eva Hřibová[5], Matthew Parker[6], Jeremiah Nwogha[4], Shengqiang Shu [7], Joseph Carlson[7], Robert Kariba[8,9], Samuel Muthemba [8,9], Katarzyna Knop[6], Geoffrey J. Barton [6], Anna V. Sherwood [6,16], Antonio Lopez-Montes[3,17], Robert Asiedu [3], Ramni Jamnadass[8,9], Alice Muchugi[8,9], David Goodstein [7], Chiedozie N. Egesi[3,4,10], Jonathan Featherston[11], Asrat Asfaw [3], Gordon G. Simpson[6,12], Jaroslav Doležel [5], Prasad S. Hendre [8,9], Allen Van Deynze [13], Pullikanti Lava Kumar[3], Jude E. Obidiegwu [4✉], Ranjana Bhattacharjee [3✉] & Daniel S. Rokhsar [1,2,7,14,15✉]

The nutrient-rich tubers of the greater yam, *Dioscorea alata* L., provide food and income security for millions of people around the world. Despite its global importance, however, greater yam remains an orphan crop. Here, we address this resource gap by presenting a highly contiguous chromosome-scale genome assembly of *D. alata* combined with a dense genetic map derived from African breeding populations. The genome sequence reveals an ancient allotetraploidization in the *Dioscorea* lineage, followed by extensive genome-wide reorganization. Using the genomic tools, we find quantitative trait loci for resistance to anthracnose, a damaging fungal pathogen of yam, and several tuber quality traits. Genomic analysis of breeding lines reveals both extensive inbreeding as well as regions of extensive heterozygosity that may represent interspecific introgression during domestication. These tools and insights will enable yam breeders to unlock the potential of this staple crop and take full advantage of its adaptability to varied environments.

[1] Department of Molecular & Cell Biology, University of California, Berkeley, CA 94720, USA. [2] Innovative Genomics Institute, Berkeley, CA, USA. [3] International Institute of Tropical Agriculture, PMB 5320, Oyo Road, Ibadan, Nigeria. [4] National Root Crops Research Institute (NRCRI), Umudike, Nigeria. [5] Institute of Experimental Botany of the Czech Academy of Sciences, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, CZ-77900 Olomouc, Czech Republic. [6] School of Life Sciences, University of Dundee, Dundee, UK. [7] DOE Joint Genome Institute, Berkeley, CA, USA. [8] World Agroforestry (CIFOR-ICRAF), Nairobi, Kenya. [9] African Orphan Crops Consortium, Nairobi, Kenya. [10] Cornell University, Ithaca, NY 14850, USA. [11] Agricultural Research Council, Biotechnology Platform, Pretoria, South Africa. [12] James Hutton Institute, Dundee, UK. [13] University of California, Davis, Davis, CA 95616, USA. [14] Okinawa Institute of Science and Technology, Onna, Okinawa, Japan. [15] Chan-Zuckerberg BioHub, 499 Illinois St., San Francisco, CA 94158, USA. [16] Present address: Department of Biology, University of Copenhagen, Copenhagen, Denmark. [17] Present address: International Trade Center, Accra, Ghana. [18] These authors contributed equally: Jessen V. Bredeson, Jessica B. Lyons. ✉email: ejikeobi@yahoo.com; r.bhattacharjee@cgiar.org; dsrokhsar@gmail.com

Yams (genus *Dioscorea*) are an important source of food and income in tropical and subtropical regions of Africa, Asia, the Pacific, and Latin America, contributing more than 200 dietary calories per capita daily for around 300 million people[1]. Yam tubers are rich in carbohydrates, contain protein and vitamin C, and are storable for months after harvesting, so they are available year-round[2,3]. World annual production of yam in 2018 was estimated at 72.6 million tons (FAOSTAT 2020). Over 90% of global yam production comes from the 'yam belt' (Nigeria, Benin, Ghana, Togo, and Cote d'Ivoire) in West Africa, where yam's importance is demonstrated by its vital role in traditional culture, rituals, and religion[3–5]. While yams are primarily dioecious, and hence obligate outcrossers, they are vegetatively propagated, allowing genotypes with desirable qualities (disease resistance, cooking quality, nutritional value) to be maintained over subsequent planting seasons.

Greater yam (*Dioscorea alata* L.), also called water yam, winged yam, or ube, among other names, is the species with the broadest global distribution[1]. *D. alata* is thought to have originated in Southeast Asia and/or Melanesia[2,6]. It was introduced to East Africa as many as 2000 years ago and reached West Africa by the 1500s[2,7]. Several traits of greater yam make it particularly valuable for economic production and an excellent candidate for systematic improvement. It is adapted to tropical and temperate climates, has a relatively high tolerance to limited-water environments, and no other yam comes close for yield in terms of tuber weight. Greater yam is easily propagated, its early vigor prevents weeds, and its tubers have high storability[8]. The tubers of *D. alata* possess high nutritional content relative to other *Dioscorea* spp[9,10].

Over the last two decades, global yam production has doubled, but these increases have predominantly been achieved through the expansion of cultivated areas rather than increased productivity[1] (FAOSTAT 2020). To meet the demands of an ever-growing population and tackle the threats that constrain yam production, the rapid development of improved yam varieties is urgently needed[11]. Conventional breeding for desired traits in greater yam is arduous, however, due to its long growth cycle and erratic flowering, and is further complicated by the polyploidy common in this species[12–14]. Efforts are currently underway by breeders to develop greater yam varieties with improved yield, resistance to pests and diseases, and tuber quality consistent with organoleptic preferences such as taste, color, and texture[11]. A critical challenge for greater yam is its high susceptibility to the foliar disease anthracnose, caused by the fungal pathogen *Colletotrichum gloeosporioides* Penz. Anthracnose disease is characterized by leaf necrosis and shoot dieback, and can cause losses of over 80% of production[15–18]. Anthracnose disease affects greater yam more than other domesticated yams; moderate resistance to this disease is present, however, in greater yam landraces and breeder's lines[19,20]. High-quality genomic resources and tools can facilitate rapid breeding methods for greater yam improvement with huge potential to impact food and nutritional security, particularly in Africa.

Here, we describe a chromosome-scale reference genome sequence for *D. alata* and a dense 10k marker composite genetic linkage map from five populations involving seven distinct parental genotypes. Comparison of the *D. alata* reference genome sequence with the recently sequenced genomes of the distantly related *D. rotundata*[21] and *D. zingiberensis*[22] reveals substantial conservation of chromosome structure between *D. alata* and *D. rotundata*, but considerable rearrangement relative to the more deeply divergent *D. zingiberensis* lineage. Analysis of the *D. alata* genome sequence supports the existence of ancient polyploidy events shared across Dioscoreales. Using a non-parametric statistical test for biased gene loss between subgenomes, we infer that

all *Dioscorea* share an ancient paleo-allotetraploidy, which was followed by species-specific chromosome rearrangements. We use genomic and genetic resources to identify nine QTL for anthracnose resistance and tuber quality traits. Our dense multiparental genetic map complements the maps previously used for QTL mapping for anthracnose resistance[23–25] and sex determination[26]. These tools and resources will empower breeders to use modern genetic tools and methods to breed the crop more efficiently, thereby accelerating the release of improved varieties to farmers.

Results and discussion

Genome sequence and structure. We generated a high-quality reference genome sequence for *D. alata* by assembling whole-genome shotgun sequence data from PacBio single-molecule continuous long reads (234× coverage in reads with 15.1 kb N50 read length), with short-read sequencing for polishing and additional mate-pair linkage (see Methods, Table 1, Supplementary Note 1, Supplementary Data 1). High-throughput chromatin conformation contact (HiC) data and a composite meiotic linkage map (see below) were used to organize the contigs (N50 length 4.5 Mb) into $n = 20$ chromosome-scale sequences, matching the observed karyotype, with each pair of homologous chromosomes represented by a single haplotype-mosaic sequence (Supplementary Figs. 1–3). The genome assembly spans a total of 479.5 Mb, consistent with estimates of 455 ± 39 Mb by flow cytometry[13], and 477 Mb by *k*-mer-based analyses (Table 1, Supplementary Note 1). The chromosome-scale 'version 2' assembly is available via Yam-Base (ftp://yambase.org/genomes/Dioscorea_alata) and Phytozome (https://phytozome-next.jgi.doe.gov/info/Dalata_v2_1), replacing the early 'version 1' draft released in those databases in 2019.

The genomic reference genotype, TDa95/00328, is a breeding line from the Yam Breeding Unit of the International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria. It is moderately resistant to anthracnose[23,27] and has been used as a parent

Table 1 Assembly and annotation statistics.

Assembly statistic	Value
Scaffold sequence total/count	480.0 Mb/25
Scaffold N50 length/count	24.0 Mb/9
Scaffold N90 length/count	19.5 Mb/18
Contig sequence total/count	479.5 Mb/532
Contig N50 length/count	4.5 Mb/31
Contig N90 length/count	565.0 kb/126

Annotation statistic	Value
Primary transcripts[a] (loci)	25,189
Alternate transcripts[b]	13,414
Total transcripts	3860
Primary transcripts	
Average number of exons	5.5
Median exon length (bp)	156
Median intron length (bp)	151
Number of complete genes	24,614
Number of incomplete genes with start codon	218
Number of incomplete genes with stop codon	281
Gene model support	
Number of genes with Pfam annotation	19,599
Number of genes with Panther annotation	23,183
Number of genes with KOG annotation	10,939
Number of genes with KEGG Orthology annotation	6849
Number of genes with E.C. number annotation	7654

[a]The longest transcript for each protein-coding gene.
[b]All other splice isoforms.

frequently in crossing programs. TDa95/00328 is diploid with $2n = 2x = 40$, as confirmed by chromosome counting (Supplementary Fig. 2) and genetically by segregation of AFLP[23]. The reference accession exhibits long runs of homozygosity due to recent inbreeding (Supplementary Fig. 4); outside of these segments we observe 7.9 heterozygous sites per kilobase.

To corroborate our genome assembly and provide tools for genetic analysis, we generated ten genetic linkage maps from eleven mapping populations that involved seven distinct parents segregating for relevant phenotypic traits (one of the maps combined two small, related mapping populations; Table 2, Supplementary Tables 1 and 2; see below). These mapping populations were generated from biparental crosses performed at IITA, with 32–317 progeny per cross. Genotyping was performed using sequence tags generated with DArTseq (Diversity Arrays Technology Pty), mapped to the genome assembly, and filtered (Methods, Supplementary Note 2), producing 13,584 biallelic markers that segregate in at least one of our mapping populations (Supplementary Table 3).

The 20 linkage groups derived from individual maps corroborated the sequence-based genome assembly and were particularly useful for interpreting HiC linkage between chromosome arms and determining their correct intrachromosomal orientations. These features were difficult to organize using HiC alone, due to strong 'Rabl' configurations (Fig. 1a, and Supplementary Figs. 1 and 5)—the three-dimensional chromatin structure characterized by polarized centromere or telomere clustering on the inner membranes of cell nuclei[28–30]—that led to contacts between the distal regions of chromosome arms (see below). The ten genetic maps were highly concordant (Fig. 1b; Kendall's tau correlation coefficients = 0.9091–0.9626), and we combined them into a single composite linkage map using five maps that capture the genetic diversity of the seven distinct parents (Supplementary Table 3). The composite map spans 1817.9 centimorgans, accounting for a total of 2178 meioses (1089 individuals), and includes 10,448 well-ordered (Kendall's tau = 0.9989; Supplementary Fig. 6) markers (excluding markers genotyped in individual crosses that were discordant post-imputation and/or were not phaseable) (Methods, Supplementary Note 2). This is the highest resolution genetic linkage map for *D. alata* produced to date.

The *D. alata* reference genome sequence encodes an estimated 25,189 protein-coding genes, based on an annotation that took advantage of both existing and the *D. alata* transcriptome resources generated in this study as well interspecific sequence homology (Table 1, Methods, Supplementary Note 3). With a benchmark set of embryophyte genes[31,32], we estimate that the *D. alata* gene set is 97.8% complete, with 1.5% gene fragmentation. While BUSCO methodology suggests that only 0.7% of the genes are missing, this is an overestimate, since some of these nominally-missing genes are detected by more sensitive searches (Supplementary Note 3). Our transcriptome datasets include short-read RNA-seq as well as 626,000 long, single-molecule direct-RNA sequences from twelve TDa95/00328 tissues. The transcriptome data identified 13,414 alternative transcripts. The great majority of genes have functional assignments through Pfam ($n = 19,599$) and Panther ($n = 23,183$) (Table 1).

Within chromosomes, protein-coding gene and transposable element densities are strongly anticorrelated (Pearson's $r = -0.885$), with gene loci concentrated in the highly-recombinogenic distal chromosome ends (Pearson's $r = +0.823$) and transposable elements, particularly Ty3/metaviridae and Ty1/pseudoviridae LTRs and other unclassified repeats, are enriched in the recombination-poor pericentromeres (Pearson's $r = -0.718$) (Fig. 1c, Supplementary Fig. 6, Supplementary Table 4). Homopolymers and simple-sequence repeats, however, were positively correlated with gene

Table 2 Mapping populations used in this study.

Pop. ID[a]	Inst.	Seed parent	Pollen parent	Putative parental relation[b]	Trait(s) studied
TDa1401	IITA	TDa05/00015	TDa99/00048	Half avuncular	Anthracnose susceptibility (field, DLA)
TDa1402	IITA	TDa05/00015	TDa02/00012	Fourth-degree relative	Anthracnose susceptibility (field[c], DLA), tuber fresh weight, tuber dry weight, tuber flesh color, tuber oxidation, dry matter content
TDa1403	IITA	TDa00/00005	TDa02/00012	Third-degree relative	Anthracnose susceptibility (field, DLA), tuber fresh weight, tuber dry weight, tuber flesh color, tuber oxidation, dry matter content
TDa1419	IITA	TDa99/00240	TDa02/00012	Unrelated	Anthracnose susceptibility (field, DLA[c]), tuber fresh weight, tuber dry weight, tuber flesh color, tuber oxidation[c], tuber flesh color
TDa1427	IITA	TDa95/00328	TDa02/00012	Unrelated	Anthracnose susceptibility (field, DLA), tuber fresh weight, tuber flesh color, tuber oxidation, dry matter content
TDa1401B	NRCRI	TDa05/00015	TDa99/00048	Half avuncular	Anthracnose susceptibility (DLA), presence of corm, ability of corm to separate, corm type, tuber shape, tuber size[c], tuber surface texture, roots on tuber, placement of roots on tuber
TDa1506 TDa1621	NRCRI NRCRI	TDa05/00015	TDa02/00012	Fourth-degree relative	(In TDa1506) Anthracnose susceptibility (DLA), presence of corm, ability of corm to separate, corm type, tuber shape, tuber size, tuber surface texture, roots on tuber, placement of roots on tuber
TDa1512 TDa1603	NRCRI NRCRI	TDa00/00005	TDa01/00039	Parent– offspring	(In TDa1512) Anthracnose susceptibility (DLA), presence of corm, ability of corm to separate, corm type, tuber shape, tuber size, tuber surface texture, roots on tuber, placement of roots on tuber
TDa1610	NRCRI	TDa99/00240	TDa02/00012	Unrelated	–

Pop. ID mapping population identifier, *Inst.* institution that grew the plants and performed the phenotyping, *Parental Relation* parental relatedness as assessed in this study, *DLA* detached leaf assay.
[a]The first two digits in a population ID denote the year of crossing. All crosses were performed at IITA and, where applicable, progeny were sent to NRCRI as botanical seeds. For mapping populations that share parents across institutes, subsets of the progeny were sent to NRCRI. For NRCRI crosses with the same parents but different population IDs (TDa1506/1621 and TDa1512/1603), the second population ID was assigned to those individuals from a cross performed with the same parents in a subsequent year. We treated these pairs as single populations for the purposes of linkage mapping, but individually for QTL analyses.
[b]Putative parental relations derived from Fig. 4.
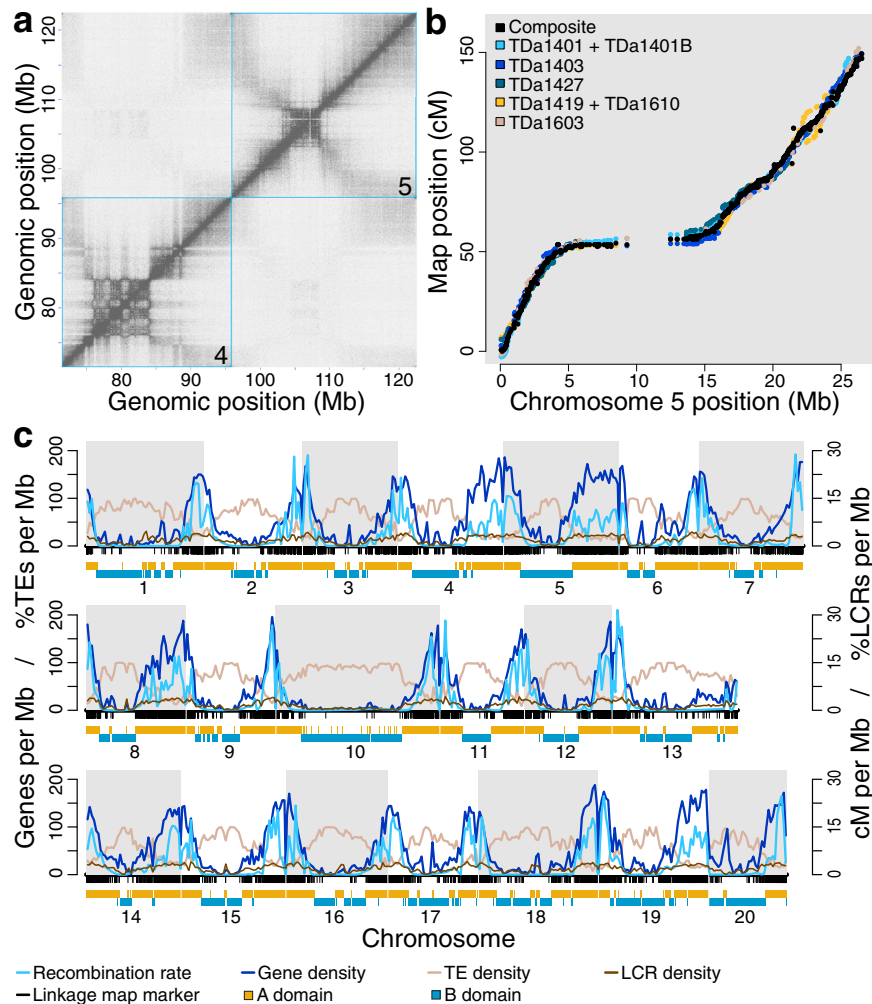[c]Traits for which significant QTL were identified (see Table 3).

Fig. 1 *D. alata* genome structure and recombination. a HiC contact matrix of TDa95/00328 chromosomes 4 and 5. Within chromosomes, the band of high contact density along the diagonal reflects the well-ordered underlying assembly. The checkerboard pattern observed between 75 and 85 Mb indicates chromatin domain A/B compartmentalization[156] within chromosome 4. The winged pattern observed within chromosomes, particularly chromosome 5, showing elevated contact densities between chromosome ends is typical of Rabl-structured chromosomes in the nucleus[29]. Chromosomes are outlined with cyan boxes. Each pixel represents the intersection between a pair of 50 kb loci along the chromosomes. The density of contacts between two loci is proportional to pixel color, with darker pixels representing more contacts and lighter representing fewer. **b** A composite genetic linkage map (black points), integrating five mapping populations (colored points, legend), is shown for chromosome 5. The maps exhibit highly concordant marker orders (Kendall's tau correlations between 0.9091 and 0.9626) and validate the large-scale correctness of the chromosome-scale assembly. The sigmoidal shape of the maps along the physical chromosome reflects suppressed recombination within the pericentromere. Individual component maps were scaled and shifted vertically to display their marker-order concordance. **c** The *D. alata* chromosome landscape is shown. Transposable elements (TEs; tan lines, left *Y*-axis) are enriched within the pericentromeres; while low-complexity repeat (LCR; brown, right *Y*-axis), protein-coding gene (dark blue line, left *Y*-axis), and meiotic recombination (cyan lines, right *Y*-axis) densities are elevated nearer the chromosome ends. Densities were computed using 500 kb bins. Composite map marker positions are shown as black ticks under the *X*-axis, with A/B chromatin compartment structure drawn below (A compartment domains in gold; B domains in dark cyan). cM centiMorgan, Mb megabase. Source data are provided as a Source Data file.

(Pearson's $r = +0.838$) and recombination (Pearson's $r = +0.728$) densities.

Analysis of chromatin conformation capture (HiC) data reveals the structure of interphase chromosomes in *D. alata* (Methods, Supplementary Note 4). We find that all chromosomes adopt a Rabl-like configuration (Supplementary Fig. 5) in which each chromosome appears 'folded' in the vicinity of the centromere, as (1) chromatin contacts are enriched among chromosome ends and (2) these chromosome ends are depleted of contacts with the pericentromeres (see also refs. [28–30]). *D. alata* chromosomes also show alternating A/B chromatin compartmentalization, as is demonstrated in several other plant species[33]. In *D. alata*, the gene-rich distal regions of each chromosome are generally spanned by open A domains (between gene density and A/B

domain status, Pearson's $r = +0.686$), while the relatively gene-poor and transposon-rich pericentromeres are characterized by closed B domains that are often punctuated by smaller A domains (Supplementary Fig. 7).

Comparative analysis and paleopolyploidy. Comparison of the *D. alata* genome sequence and protein-coding annotation with those of white yam (*D. rotundata*[21], also known as Guinea yam), bitter yam (*D. dumetorum*[34]), and peltate yam (*D. zingiberensis*[22]) highlights the completeness of our sequence and annotation and the extensive sequence divergence across the genus. Among the *Dioscorea* species sequenced to date, the annotation of *D. alata* appears to be the most complete (Supplementary Table 5,
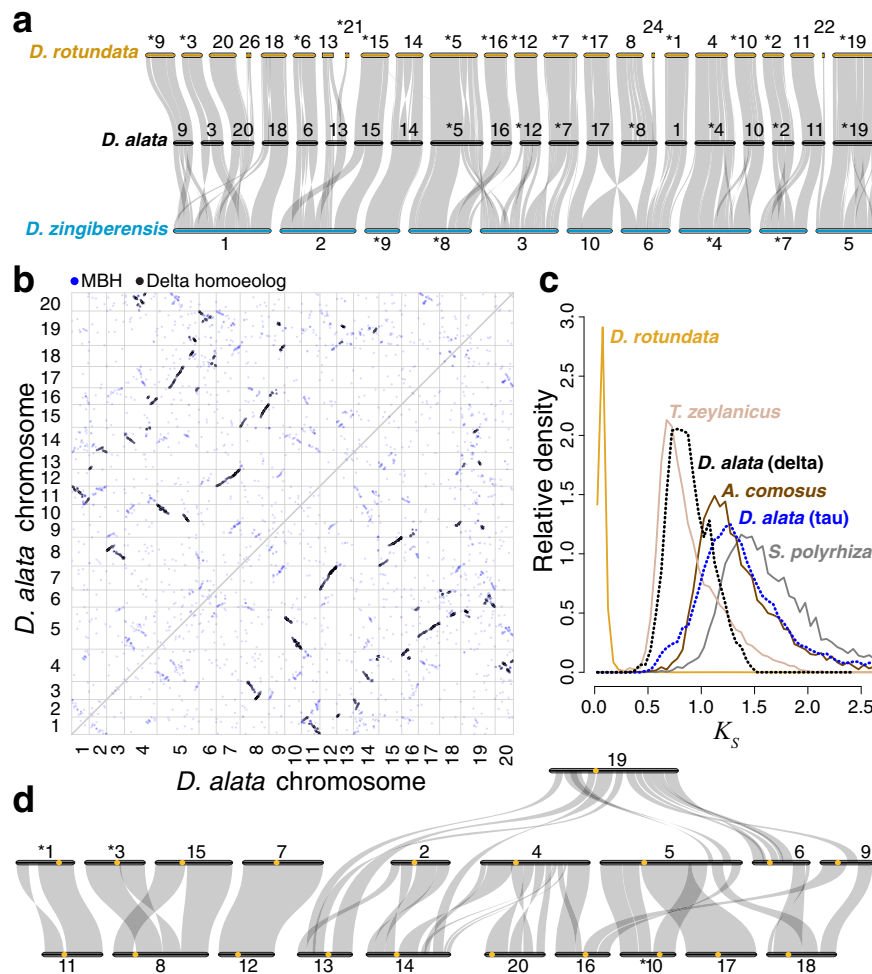
Fig. 2 Dioscoreaceae chromosome evolution. a Ribbon diagram demonstrating conserved chromosomal synteny and large-scale segmental collinearity (semi-transparent gray ribbons) between *Dioscorea alata* (black horizontal bars), *D. rotundata* (gold), and *D. zingiberensis* (cyan) one-to-one orthologous gene pairs. Only *D. rotundata* sequences with five or more collinear genes are shown. To improve visual clarity, some chromosomes, marked with asterisks, were reverse complemented with respect to their assembled sequences. Chromosome sizes are proportional to the number of annotated genes. **b** Dot plot showing evidence of two whole-genome duplications exposed by TDa95/00328 intragenomic comparison. Each point represents a mutual best-hit (MBH) gene pair and each white box (outlined in grey) represents the intersection of two chromosomes. Homoeology from the recent Dioscoreaceae delta duplication is shown in black and the ancient, core monocot tau duplication can be seen as clusters in blue (see also, Supplementary Fig. 9). **c** The synonymous substitution rate (K_S) histograms for orthologous (solid lines) or homoeologous (dotted lines) gene pairs between *D. alata* and select species comparators are shown: *D. rotundata* ($n = 14,889$), *T. zeylanicus* ($n = 9013$), *D. alata* delta ($n = 1578$), *A. comosus* ($n = 6405$), *D. alata* tau ($n = 404$), and *S. polyrhiza* ($n = 4973$). The *D. alata*–*D. rotundata* ortholog density was rescaled by 0.25 to emphasize other comparisons. **d** Shared segmental homoeology (semi-transparent gray) between *D. alata* chromosomes (black horizontal bars) resulting from the delta duplication is depicted with a ribbon diagram, as in panel **a**, but with putative centromere positions now included as gold circles (Supplementary Data 2). Source data are provided as a Source Data file.

Supplementary Note 3). For example, *D. alata* has the fewest missing conserved gene families in cross-species comparisons within Dioscoreaceae (53 in *D. alata* compared with 385 for *D. zingiberensis* and 595 for *D. rotundata*) and in cross-monocot comparisons (7 in *D. alata* compared with 99 in *D. zingiberensis* and 110 in *D. rotundata*) (Supplementary Fig. 8). These metrics combine genome assembly completeness and accuracy with exon-intron structure predictions based, in part, on transcriptome resources.

At the nucleotide level, *D. alata* coding sequences exhibit 97.4%, 93.6%, and 86.5% identity with *D. rotundata*, *D. dumetorum*, and *D. zingiberensis*, corresponding to median synonymous substitution (K_S) rates of 0.064, 0.163, and 0.389, respectively. These measures are consistent with *D. zingiberensis* being a deeply branching outgroup to the clade formed by *D. alata*, *D. rotundata*, and *D. dumetorum* (see also Supplementary Table 6), and highlights the ~60 My old divergences within the genus *Dioscorea*.

The medicinal plant *Trichopus zeylanicus* (common name 'Arogyappacha' in India, meaning 'the green that gives strength')[35] is a more distantly related member of the Dioscoreaceae family, with 77.9% identity and median K_S of 0.804.

The ($n = 20$) chromosome sequences of *D. alata* and *D. rotundata*[21,36,37] are in 1:1 correspondence, and are highly collinear (Fig. 2a, Supplementary Fig. 9a). The few intra-chromosome differences observed could represent bona fide rearrangements between species or, possibly, imperfections in the *D. rotundata* v2 assembly[21] that could have arisen from the reliance on linkage mapping to order and orient *D. rotundata* scaffolds, especially in recombination-poor pericentromeric regions of the genome. Under the assumption that *D. rotundata* chromosomes are in 1:1 correspondence with *D. alata* chromosomes, we can provisionally assign four large but unmapped *D. rotundata* scaffolds to chromosomes (Fig. 2a). We found one inter-chromosome difference (not present in the *D. rotundata* v1

assembly[37]), which requires further study (Supplementary Fig. 9a). While the draft *D. dumetorum* genome assembly is not organized into chromosomes, comparison with the *D. alata* reference sequence shows that the two genomes are locally collinear on the scale of the *D. dumetorum* contigs, with only one discordance (Supplementary Fig. 9b). This observation suggests a provisional organization of the *D. dumetorum* contigs into probable chromosomes. Notably, the distantly related *D. zingiberensis* has a haploid complement of $n = 10$ (ref. [38]), compared with $n = 20$ found in *D. alata*, *D. rotundata*[21,39], and *D. dumetorum*[2,40]. We find that the *D. zingiberensis* chromosomes[22] were formed from ancestral, *D. alata*-like chromosomes and/or chromosome arms by combinations of end-to-end and centric fusions and translocations (Fig. 2a, Supplementary Fig. 9c).

We found evidence for two ancient paleotetraploidies in the *D. alata* lineage. These duplications evidently preceded the origin of the genus, since all *Dioscorea* genome sequences show one-to-one orthology (Supplementary Fig. 9a–c, Supplementary Note 5). The most recent paleotetraploidy is apparent from extensive collinear paralogy in *D. alata* (Fig. 2b) and coincides with the genome duplication recently described in *D. zingiberensis*[22] and previously identified based on transcriptome analysis of *D. villosa* in the context of one thousand plant transcriptomes as DIV1-alpha[41], but not found in an earlier analysis that included the *D. opposita* transcriptome[42]. Following the common use of Greek letters to denote plant polyploidies, we designate this *Dioscorea* lineage duplication as 'delta.' The median sequence divergence between 1,578 delta paralogs in *D. alata* is $K_S = 0.869$ substitutions/site (Fig. 2c). While comparisons with the draft genome assembly of *T. zeylanicus* ($K_S = 0.804$ to *D. alata*) further suggest that the delta paleotetraploidy may have preceded the origin of the family Dioscoreaceae, the fragmentation of the *T. zeylanicus* assembly precludes a definitive assessment. The timing of the delta duplication (estimated to be 64 Mya[22]) is contemporaneous with the K/T boundary and a cluster of other successful paleopolyploidies[43].

Analysis of the *D. alata* genome sequence reveals large-scale genomic reorganization after the delta duplication. *D. alata* chromosomes preserve long collinear paralogous segments arising from the delta paleotetraploidy event, and the genomic organization of these segments reveals large-scale rearrangements after whole-genome duplication (Fig. 2d, Supplementary Data 2). These include cases of one-to-one whole-chromosome paralogs, (chromosomes 1 and 11; 7 and 12) as well as examples of centric insertion (e.g., the paralog of chromosome 3 was inserted within the paralog of chromosome 15 to form chromosome 8; the paralog of chromosome 17 was inserted into the paralog of chromosome 10 to form most of the chromosome 5). Other large-scale rearrangements are evident, including apparent end-to-end 'fusions' (or more properly translocations[44]). Taken together, these paralogies provide further evidence for the delta duplication.

Genome duplication can occur by two distinct evolutionary mechanisms[45]: allotetraploidy (genome duplication after hybridization of two distinct diploid progenitors) or autotetraploidy (genome duplication within a single species). Since hybridization brings together genomes with distinct epigenetic properties[46], a hallmark of ancient allotetraploidy is differential evolution of the homoeologous chromosome sets ('subgenomes') inherited from distinct progenitor species. In particular, paleo-allotetraploids may exhibit asymmetric gene loss (or conversely, gene retention) between subgenomes, often referred to as 'biased fractionation'[45,47,48]. While the observation of asymmetric gene retention is considered positive evidence for paleo-allotetraploidy[45], a lack of detectable asymmetry in gene loss can be consistent either with autotetraploidy or with allotetraploidy that is recent and/or involved hybridization of closely related progenitors species.

To test for patterns of differential gene retention that are diagnostic of paleo-allotetraploidy, we analyzed 15 robust pairs of paralogous *D. alata* segments (each with more than 40 paralogous genes) from the delta duplication, drawn from 11 distinct chromosome pairs. We observe a bimodal distribution of retention rates across these 30 chromosomal segments relative to the inferred unduplicated gene complement (Methods, Supplementary Note 5), with peaks at 0.63 and 0.48 (Supplementary Fig. 10). Importantly, for each of the 11 homoeologous chromosome pairs, one paralog has a high retention rate and the other low (Supplementary Table 7). Such a paired distribution of high and low-retention chromosomes is unexpected under the null (autotetraploid) model of uncorrelated gene loss ($p = 2.9 \times 10^{-3}$; $k = 11$, $n = 11$) (Supplementary Table 8, Supplementary Note 5). Analysis of the other *Dioscorea* genomes yields consistent results (Supplementary Tables 7 and 8).

Our finding of consistent patterns of differential gene retention between homoeologous chromosomes (1) allows us to reject the autotetraploid hypothesis, and (2) provides positive support for a paleo-allotetraploid scenario for the ancient delta genome duplication in *Dioscorea*. Under this paleo-allotetraploid scenario, the high- and low-retention chromosomes of *Dioscorea* spp. represent the descendants of the ancestral chromosomes of the two progenitors (now subgenomes). Since our method does not require an extant relative of the unduplicated progenitors[49] it can be applied to other ancient genome duplications, with the caveat that not all allotetraploidizations may trigger asymmetric gene loss[48,50].

In addition to delta, the *D. alata* genome sequence also displays relicts of a more ancient genome-wide duplication in the form of nearly-collinear ancient paralogous segments with median $K_S = 1.21$ substitutions per site (Fig. 2b, c). We identify this duplication with the famed 'tau' duplication shared by other core monocots, including grasses[50], pineapple (*Ananas comosus*[51]), oil palm (*Elaeis guineensis*[52]), and asparagus (*Asparagus officinalis*[53]) but not duckweed (*Spirodela polyrhiza*[54]). The tau duplication has also been noted in transcriptome analyses[41,42]. The clear 2:2 pattern of orthology between yam, pineapple, and oil palm (Supplementary Fig. 9d, e) confirms that these three lineages have each experienced one lineage-specific whole-genome duplication (delta, sigma, and p, respectively) since they diverged from each other. This pattern implies that relicts of any earlier duplications observed in these species must represent shared events. Since Dioscoreales is one of several early-branching core monocot lineages (only Petrosaviales branches earlier), the discovery of tau in yam implies that this duplication likely preceded the divergence of the core monocot clade (Supplementary Figs. 9f and 11). (Since tau occurred close in time to the divergence of the non-Petrosaviales core monocots, the combination of tau and the respective lineage-specific duplications produce 4:4 patterns of paralogy in dot plots. See Supplementary Fig. 9d, e)

QTL mapping. To demonstrate the utility of our dense linkage maps and high-quality *D. alata* reference genome sequence for advancing greater yam breeding, we searched for quantitative trait loci (QTL) for resistance to anthracnose disease and several tuber quality traits (dry matter, oxidation, tuber color, corm type, and other traits). Our mapping populations were generated in controlled crosses by yam breeders at IITA, Nigeria, using parents from the yam breeding program (Table 2, Supplementary Table 1). Phenotyping was performed in Nigeria at IITA Ibadan and NRCRI in Umudike (Methods, Supplementary Note 6). Leveraging the ability to clonally propagate individuals, we
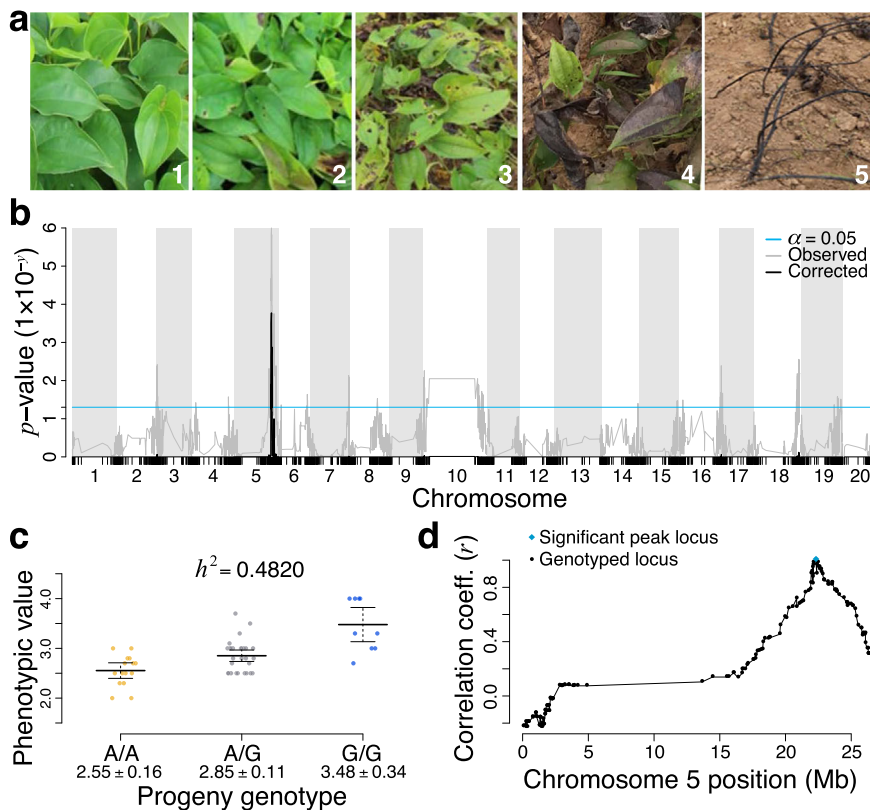
Fig. 3 Quantitative trait locus for anthracnose resistance. a Exemplars of the yam anthracnose disease (YAD) field assessment severity rating scale (scored 1–5) used at IITA in Ibadan, Nigeria. **b** Genome-wide QTL association scan for YAD resistance in the TDa1402 genetic population ($n = 53$ biologically independent samples) for the year 2017. A statistically significant association (corrected $p = 1.69 \times 10^{-4}$) was found on chromosome 5, at 23.3 Mb. Per-locus Wald statistic-based logistic regression significance values (gray line) were corrected for multiple testing (black line) via max(T) adjustment with 1×10^6 permutations. The minimum significance threshold ($\alpha = 0.05$) is represented with a cyan horizontal line. **c** Effect plot for the peak locus on chromosome 5 at 23.3 Mb, the genotypes (X-axis) of which explain 48.2% of the observed phenotypic variance (i.e., narrow-sense heritability, h^2), suggests that an increased dose of the 'A' allele is associated with lower severity of YAD. Centerline and whisker plots, and their corresponding statistics (X-axis), represent the mean ± 95% confidence intervals. **d** Plot showing the strength of linkage disequilibrium (LD) between the peak marker (cyan diamond) and other loci (black points) in chromosome 5. LD was calculated as Pearson's correlation (r) between alleles. Source data are provided as a Source Data file.

measured multiple traits over the years 2016–2019. Our QTL analyses exploited the imputed genotypes derived from our dense linkage maps. In total, we found eight distinct QTL: three for anthracnose resistance and five for tuber traits (Fig. 3, Table 3, Supplementary Figs. 12–13).

QTL for anthracnose resistance. Yam Anthracnose Disease (YAD), or yam dieback, is a major disease afflicting yams caused by the fungus *Colletotrichum gloeosporioides*[15,18]. Greater yam is particularly susceptible to YAD, although resistance has been shown to vary among *D. alata* genotypes[55]. We sought QTL for YAD resistance using field trials in five mapping populations and detached leaf assays in eight mapping populations (Table 2, Methods, Supplementary Note 6). While most of these populations did not show significant QTL, we found three significant anthracnose resistance QTL in two of them.

In field trials of the TDa1402 population, we found a major QTL on chromosome 5 ($p = 1.69 \times 10^{-4}$) that explains 48.2% of phenotypic variance in the 2017 data, with an additive effect (Fig. 3a–c), and a minor QTL on chromosome 19 (Supplementary Fig. 12a–c) that explains 29.9% of the variance in the 2018 data ($p = 1.25 \times 10^{-2}$). Although anthracnose response and resistance are poorly understood in yams, studies in other species suggest potential candidate genes overlapping these QTL intervals, including a gene (Dioal.05G183500) on chromosome 5 that

encodes a receptor-like EIX1/2 protein, which is a member of the LRR (leucine-rich-repeat) superfamily of plant disease resistance proteins[56], and genes on chromosome 19 that encode members of the EMSY-LIKE family of immune regulators of fungal disease resistance[57,58] (Dioal.19G063700), three NB-ARC domain-containing *R*-gene analog (RGA) disease resistance protein-encoding genes[59] (Dioal.19G073100, Dioal.19G074700, and Dioal.19G084600), and two genes (Dioal.19G066100 and Dioal.19G066200) encoding proteins of unknown function that contain C-terminal domains of the ENHANCED DISEASE RESISTANCE 2 (EDR2) family that are negative regulators of plant-pathogen response[60,61]. These QTL are candidates for use in marker-assisted breeding and provide leads for further molecular characterization of anthracnose disease response in yam. However, since variation in levels of infestation, overall plant vigor, and timing and amount of rainfall influence disease severity in field trials, validation of these QTL is required.

In detached leaf assays of the TDa1419 population, performed under varying conditions over three years (Methods), we found a QTL of smaller effect (7.3% of phenotypic variance) on chromosome 6 (Supplementary Fig. 12d–g). While this QTL was marginally significant ($p = 1.28 \times 10^{-2}$), it was found only using three-year averages, and the locus was not significantly associated with YAD in the data from individual years. Furthermore, anthracnose disease levels, as measured by detached leaf assay, were not significantly correlated across genotypes over

Table 3 Significant QTL identified in this study.

Pop. ID	Trait	QTL peak position	n	p-value	Variant	h^2	Significance Window[a]
TDa1402	Anthracnose susceptibility (Field 2017)	Chr5: 22,308,637	53	1.69×10^{-4}	A/A,A/G,G/G	0.4820	21,931,073 22,825,712
TDa1402	Anthracnose susceptibility (Field 2018)	Chr19: 8,369,514	49	1.25×10^{-2}	T/T,T/C	0.2986	3,732,307 17,565,140
TDa1419	Anthracnose DLA 3-yr mean	Chr6: 61,001	243	1.28×10^{-2}	C/C,C/T	0.0734	38,157 1,418,849
TDa1419	Dry matter	Chr18: 25,069,928	150	2.27×10^{-2}	C/C,C/T	0.1020	24,779,355 25,415,124
TDa1419	Oxidation after 30 min[b]	Chr18: 26,496,992	151	5.86×10^{-3}	T/T,T/A,A/A	0.1367	26,199,630 26,749,589
TDa1419	Oxidation after 180 min[b]	Chr18: 26,496,992	151	1.38×10^{-2}	T/T,T/A,A/A	0.1188	26,199,630 26,749,589
TDa1427	Oxidation after 30 min	Chr18: 24,495,033	97	4.52×10^{-6}	A/A,A/G	0.3127	24,034,264 24,938,398
TDa1401B	Tuber size	Chr12: 310,852	53	4.19×10^{-2}	T/T,T/C,C/C	0.2894	76,400 489,583
TDa1512	Tuber shape	Chr7: 3,115,608	43	3.17×10^{-2}	A/A,A/G	0.3406	1,798,899 5,707,988

Pop. ID mapping population identifier, n the number of genotyped and phenotyped progeny used in QTL analysis, p-value empirical significance ($\alpha = 0.05$) of the genotype-phenotype association at the peak locus, calculated by Wald statistic-based logistic regression and corrected for family-wise multiple testing by the max(T) method, Variant alleles segregating at QTL peak position, h^2 narrow-sense heritability.
[a]Calculated as haplotypic linkage disequilibrium ≥0.9 relative to the peak QTL marker.
[b]Same QTL for both oxidation time points in TDa1419.

years. These observations suggest that variation in YAD may be dominated by non-genetic factors.

While previous studies identified two significant anthracnose QTL using EST-SSRs[25] and three QTL using GBS-SNPs[62], none of these colocalize with the QTL in our study. This discrepancy (and the variability seen among different years in our work) may be due to differences in the parental yam genotypes, differences in anthracnose strain and/or inoculation rate in these field studies, and possible genotype-by-environment interactions. Although our parental lines show evidence suggesting past introgression (see below), we did not find any overlaps between these putatively introgressed blocks and our QTL, as might be expected if disease resistance was brought into cultivated yam from a related wild species.

QTL for tuber quality traits. Post-harvest oxidation causes browning of yam tuber flesh and flavor changes that reduce crop value[63]. We found an additive-effect QTL for tuber oxidation after peeling at both 30 min ($p = 5.86 \times 10^{-3}$) and 180 min ($p = 1.38 \times 10^{-2}$) on chromosome 18 in the TDa1419 population (Supplementary Fig. 13a–f). The QTL explained 13.67% and 11.88% of the phenotypic variance at 30 and 180 min after peeling, respectively. In the TDa1427 population, a closely linked QTL ($p = 4.52 \times 10^{-6}$), located 2 Mb upstream on the same chromosome, explained 31.3% of the phenotypic variance in oxidation after 30 min (Supplementary Fig. 13g–i). Although enzymatic browning in yam remains poorly understood, polyphenol oxidases and peroxidases are active during browning of *D. alata* and *D. rotundata*[64], and inhibition of this activity has been shown to reduce browning in Chinese yam (*D. polystachya*)[65]. We find a cluster of three peroxidase-encoding genes (Dioal.18G098800, Dioal.18G099400, and Dioal.18G100900) on chromosome 18 at 26.23–26.36 Mb, within ~200 kb of the oxidation QTL at 26.50 Mb in TDa1419 and within 2 Mb of the oxidation QTL in TDa1427, raising the possibility that oxidation is affected by genetic variation in peroxidase activity.

Dry matter (principally starch) content is an important measure of yam yield[66]. We found a single, minor QTL (explaining 10.2% of the phenotypic variance for the dry matter)

on chromosome 18 (Supplementary Fig. 13j–l) in population TDa1419, at position Chr18:25,069,928 ($p = 2.27 \times 10^{-2}$), with genotypes segregating in the population in a pseudo-testcross configuration. Lastly, we identified two QTL for tuber size ($p = 4.19 \times 10^{-2}$) and shape ($p = 3.17 \times 10^{-2}$) in populations TDa1401B and TDa1512, respectively, accounting for 28.9% and 34.1% of their phenotypic variances (Supplementary Fig. 13m–r). While three loci associated with dry matter content and two associated with oxidative browning were previously identified via a genome-wide association study (GWAS)[67], these QTL do not colocalize with those found here, which may be due to differences in the parental yam genotypes or possible genotype-by-environment interactions.

Genetic variation within D. alata. To enable future genetic analyses, we developed a catalog of nearly 3.05 million biallelic single-nucleotide variants (SNVs) in *D. alata*, based on whole-genome shotgun resequencing (Supplementary Note 7, Supplementary Data 1, Supplementary Fig. 14) of breeding lines representing the seven parents of our biparental mapping populations and an additional breeding line (TDa95-310). Of the 3.05 million biallelic SNVs, in our collection, 1.89 million could be confidently genotyped across all individuals. Included within the larger set are 305.5k coding SNVs (251.5k in the reduced set) with predicted effect, 127.1k of which introduce nonsynonymous amino acid changes.

We used these dense SNVs to determine the relationships among the eight breeding lines (Fig. 4a, Supplementary Table 1, Supplementary Data 3) by estimating the fractions of their genomes they shared as identical by descent (IBD). We identified six parent-child relationships (i.e., IBD1, one haplotype shared across the entire genome; relatedness coefficients ~0.50) and five second-degree relationships (i.e., coefficients of ~0.25). All second-degree relations showed unusually high values of IBD1, and both first- and second-degree relations shared substantial IBD2, suggesting a history of recent inbreeding. The relationships inferred are consistent with available pedigree records (Supplementary Table 1), with the addition of several previously unrecorded grandparent-grandchild relationships. Although the
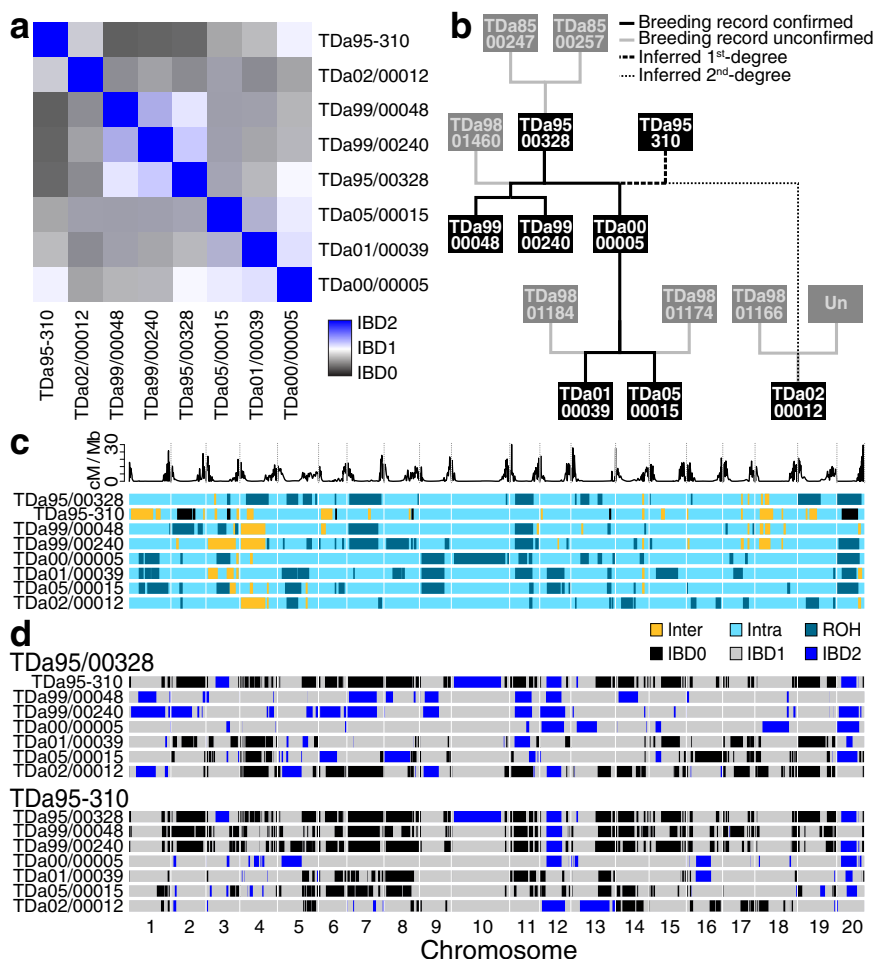
Fig. 4 Relationships between eight deeply sequenced *D. alata* breeding lines. a Matrix of identity-by-descent (IBD) relatedness between all pairs of individuals (Supplementary Data 3). Blue represents the degree of diploid genome identity (IBD2); white, one haplotype (IBD1); and black, no shared haplotypes (IBD0). **b** Pedigree of relationships. Sequenced individuals are represented in black boxes with white text, while individuals not sequenced are in gray. Relationships known via IITA records (Supplementary Table 1) are drawn with solid lines. Relationships that could be confirmed using direct sequence comparison are highlighted with solid black lines, and those that could not be are colored grey. Inferred cryptic relationships are indicated with broken lines (first- and second-degree relations are represented as thick dashed and thin dotted lines, respectively). Unexpectedly, TDa95-310 is a parent of TDa00/00005 and a likely second-degree relative of TDa02/00012. **c** Regions of heterozygosity, autozygosity, and possible introgression. Within a background of intraspecific genetic variation (light cyan), large homozygous blocks (runs of homozygosity [ROH], dark cyan) appear common in the resequenced individuals, suggesting autozygosity from historical inbreeding. In addition, large blocks of exceptionally high heterozygosity (yellow) can also be observed, indicating possible introgressions (interspecific variation introduced via hybridization) in one or more of the unsampled pedigree founders. The recombination rate along each chromosome is shown in the track above. **d** Haplotype sharing between TDa95/00328 and all other resequenced individuals, and TDa95-310 and all others. Regions of the genome where an individual shares two haplotypes (i.e., they are IBD2) with TDa95/00328 (or TDa95-310) are highlighted in blue, one shares haplotype (IBD1) in gray, or shares no haplotypes (IBD0) in black. Source data are provided as a Source Data file.

use of highly related parents in breeding programs limits the diversity of alleles available for selection, we note that, as a practical matter, yam crosses are limited to genotypes that flower appropriately, consistently, and profusely.

Unexpectedly, our identity-by-descent analysis shows that TDa95-310 shares a parent-child relationship to TDa00/00005 and a grandparent-grandchild relationship to TDa01/00039 and TDa05/00015. This finding implies that TDa95-310 and the individual TDa98/00150, which appears in the corresponding position in pedigrees, are clones, or that TDa98/00150 is not a parent of TDa00/00005. TDa95-310 is a landrace from Cote d'Ivoire that is likely derived from an accession known as 'Brazo-Fuerte' ('strong arm') introduced from Latin America. It is susceptible to anthracnose and has been used as parent material for crossing[68,69]. We find that TDa95-310 is a second-degree

relative of TDa02/00012. Based on the reported pedigree (Fig. 4b), TDa95-310 must be (1) a parent of either (a) TDa98/01166 or (b) the unknown pollen parent of TDa02/00012, or (2) TDa95-310 also shares one of them as parents. Additional genotyping will resolve this mystery and prevent accidental inbreeding using TDa95-310.

We find extended runs of homozygosity among our eight sequenced lines, as expected based on their high degree of relatedness (Fig. 4c). Long blocks of homozygosity generally stretch across pericentromeric regions, consistent with the low-recombination rates in these regions (Figs. 1 and 4). Although our sampling is not random, the extensive homozygosity (and identity across genotypes) suggests that there may have been selection for the haplotype on chromosome 20 that appears in a homozygous state in six of our eight breeding lines, as well as some other common haplotypes seen in Fig. 4d. The reduced

genetic variation present in these breeding lines suggests a strong need for the introduction of additional diversity in yam breeding programs at IITA and other national institutes.

Conversely, we find that multiple genomes contain several long runs of unusually high heterozygosity (Fig. 4c, Supplementary Fig. 4). While the typical rate of single-nucleotide heterozygosity across 100 kb blocks is ~7–10 SNVs per kb (excluding runs of homozygosity), these highly heterozygous runs have more than 17.5 SNVs/kb (Supplementary Fig. 4c, d, f–g). In cassava and citrus, blocks of high heterozygosity exceeding 10 SNVs/kb variation have been demonstrated to be due to interspecific introgression[70,71]. The co-cultivation of related yam species (Supplementary Fig. 15, Supplementary Note 8, Supplementary Data 4) by growers and breeders suggests that these blocks (some of which are found overlapping low-recombination-rate pericentromeric regions, e.g., on chromosome 4) are the result of past interspecific introgression. Since the Pacific yam *D. nummularia* is the only other yam species shown to be interfertile with *D. alata*[20], we speculate that it is the source of introgression into greater yam breeding lines, possibly before introduction to Africa. The retention of these hybrid sequences in this germplasm suggests that they may confer some possible adaptive advantage, as has been hypothesized in cassava (*Manihot esculenta* Crantz)[70]. Wolfe et al.[72] showed that *Manihot glaziovii* Muell. Arg. segments introgressed into and maintained as heterozygous in the cassava genome are associated with preferred traits. In the future, a comparison of these highly heterozygous regions with sequences from related *Dioscorea* spp. should reveal the source of these interspecific contributions to the greater yam germplasm.

Conclusion. The near-complete and contiguous chromosome-scale assembly of *D. alata* reported here, along with the associated genetic and genomic resources, opens new avenues for improving this important staple crop. We demonstrated the utility of these resources by finding eight QTL for anthracnose disease resistance and tuber quality traits. The genome sequence and associated resources will facilitate future marker-assisted breeding efforts in this crop. A major hurdle for breeders is the difficulty of making successful crosses in *D. alata* due to lack of flowering, limited seed set, and differences in flowering time. Genome-enabled methods such as marker-assisted selection, GWAS, and genomic selection will allow breeders to make the most out of each cross and use fewer resources to maintain genotypes that are less likely to be useful. By analyzing the diversity of popular breeding lines, we found that they are highly related and, in some cases, have long runs of homozygosity that reduce the genetic diversity available for selection but may represent genomic regions fixed for desirable traits. Analysis of a broader sampling of African greater yam germplasm will prove valuable to avoiding inbreeding depression associated with inbreeding elite lines[73]. Conversely, we found regions of presumptive interspecific hybridization, pointing to the potential value of broader crosses that may enable the transfer of valuable traits from other yam species while minimizing linkage drag with genome-assisted selection. Similarly, the genome sequence also enables the application of gene editing to directly alter genotypes in a targeted manner, preserving genetic backgrounds that confer cohorts of desirable traits. The small genome of *D. alata* and the advent of rapid long-read technologies open the door to rapidly assemble additional accessions to discover and leverage structural variants for breeding. Such variants have been shown to control important traits, such as plant development[74], and contribute to reproductive isolation[75].

Greater yam has a high potential for increased yield and broader cultivation, with advantages compared with other root-tuber-banana crops due to its superior nutritious content and low glycemic index[76,77]. Greater yam's ability to grow in tropical and sub-temperate regions around the world suggests that it is highly adaptable to its environment and that there may be adaptive traits (and associated alleles) that could be exploited in different global contexts. It establishes itself vigorously, is higher yielding than other domesticated yam species, and is highly tolerant to marginal, poor soil and drought conditions, and thus likely nutrient use efficient[8]. These traits will be valuable assets in a changing climate. Greater yam is also highly tolerant of the most significant yam virus, yam mosaic virus[19]. By leveraging QTL and genome-wide association for disease resistance and tuber quality, as well as marker-aided breeding strategies and genome editing, yam breeders are poised to rapidly generate disease-resistant, high-performing, farmer-/consumer-preferred, climate-resilient varieties of greater yam.

Methods

Reference accession. The breeding line TDa95/00328, from the International Institute of Tropical Agriculture (IITA) yam breeding collection, was chosen as the *D. alata* reference genome accession because it is moderately resistant to anthracnose (a fungal disease caused by *Colletotrichum gloeosporioides*) and was confirmed to be diploid by marker segregation analysis[23,27]. Chromosome number ($2n = 40$) was further confirmed through chromosome counting (Supplementary Note 1, Supplementary Fig. 2).

Genome sequencing. High molecular weight DNA for Pacific Biosciences (PacBio, Menlo Park, USA) Single-Molecule Real-Time (SMRT) continuous long-read (CLR) sequencing was isolated as described in Supplementary Note 1. PacBio library preparation and sequencing were performed at the University of California Davis Genome and Biomedical Sciences Facility. Three libraries were constructed as per manufacturer protocol, with fragments smaller than 7, 15, and 20 kb, respectively, excluded using Blue Pippin. In total, one RSII and 20 Sequel SMRT cells of CLR data were generated for a combined 235× sequence depth. Half of the 112.4 Gb of generated bases were sequenced in reads 14.5 kb or longer.

For HiC chromatin conformation capture, suspensions of intact nuclei from *D. alata* (TDa95/00328) were prepared from young leaves and apical parts of the stem according to ref. [78]. at the Institute of Experimental Botany, Olomouc, Czech Republic, with modifications as described in Supplementary Note 1. These nuclei were sent to Dovetail Genomics for HiC library preparation[79], which were sequenced on an Illumina HiSeq 4000 to produce 358.5 million 151 bp paired-end reads.

For genome sequence polishing, a 625 bp insert-size Illumina TruSeq library was made and sequenced on a HiSeq 2500 at UC Berkeley's Vincent J. Coates Genomics Sequencing Lab (VCGSL), yielding 131 million 251 bp paired reads (137× depth). For contig linking, three Nextera mate-pair libraries (insert sizes ~2.5 kb, 6 kb, and 9 kb) were prepared and sequenced as 151 bp paired-end reads on a HiSeq 4000 at the UC Davis Genome and Biomedical Sciences Facility. More details are described in Supplementary Note 1. A listing of all TDa95/00328 sequencing data, and corresponding NCBI Sequence Read Archive (SRA) accession numbers, may be found in Supplementary Data 1.

Genome assembly. We assembled the *D. alata* genome sequence with Canu[80] v1.7-221-gb5bffcf from the longest 110× of PacBio CLR reads (50.228 Gb in reads 19.8 kb or longer). Contigs were filtered down to a single mosaic haplotype in JuiceBox[81,82] v1.9.0, considering median contig depth (Supplementary Fig. 3), sequence similarity, and HiC contacts. Non-redundant contigs were scaffolded into chromosomes using SSPACE[83] v3 and 3D-DNA[84] commit 2796c3b. Misassemblies were corrected manually with the aid of genetic maps and JuiceBox HiC visualization. The assembly was polished twice with Arrow[85] v2.2.2 (SMRT Link v6.0.0.47841) followed by two rounds of Illumina-based polishing with FreeBayes[86] v1.1.0-54-g49413aa and custom scripts (Supplementary Note 1).

DArTseq genotyping. DNA was isolated at IITA and NRCRI from their respective mapping populations and parents using modified CTAB methods (Supplementary Note 2). DNA samples were genotyped by Integrated Genotyping Service and Support (IGSS, BecA-ILRI hub, Nairobi, Kenya) or DArT (Canberra, Australia) using the 'high-density' DArTseq reduced-representation method. DArTseq genotype datasets were deposited in Dryad [https://doi.org/10.6078/D1DQ54][87]. Lists of sequence data used for DArTseq genotyping, and corresponding NCBI Sequencing Read Archive (SRA) accession numbers, are provided in Supplementary Data 1.

Genetic linkage mapping. DArTseq genotyping datasets were mapped to the v2 genome sequence, then filtered for a minimum 90% genotyping completeness and F_1 Mendelian segregation via χ^2 goodness-of-fit tests ($\alpha = 1 \times 10^{-2}$) on allele and

genotype frequencies using MapTK[88] v1.4.1-11-g19a5f3a (https://bitbucket.org/rokhsar-lab/gbs-analysis) and VCFtools[89]. Half-sibs, off-types, and sample errors were detected via clustering as in ref. [88]. and removed. Parental genotypes from one dataset were substituted when a sample by the same name was found to be inconsistent in another. Genotypes were phased and imputed using AlphaFamImpute[90] v0.1 and parent-averaged linkage maps constructed in JoinMap[91,92] v4.1 with the maximum-likelihood mapping function for cross-pollinated populations, which were then integrated into a composite map using LPmerge[93] v1.7. Further detail regarding genetic linkage mapping can be found in Supplementary Table 2 and Supplementary Note 2. All linkage maps were deposited in Dryad [https://doi.org/10.6078/D1DQ54][87].

RNA sequencing. RNA was extracted at ICRAF from 12 tissues from a single TDa95/00328 plant grown onsite in Nairobi, Kenya. Tissues included leaf petiole, roots, various stages of leaves (initial sprouting leaf, leaf bud, young leaf, semi-matured leaf, matured leaf, fifth leaf), bark, stem, first internode, and middle vine as described in Supplementary Note 3. RNA samples were pooled for sequencing by two technologies.

Illumina RNA-seq libraries were prepared using the TruSeq stranded mRNA preparation kit (Illumina cat# 20020594) and sequenced at the Agricultural Research Council Biotechnology Platform (ARC-BTP) in Pretoria, South Africa on an Illumina HiSeq 2500 as 125 bp paired ends (SRA: SRR13683865 [https://www.ncbi.nlm.nih.gov/sra/SRR13683865]).

Oxford Nanopore Technologies (ONT) Direct-RNA Sequencing (Nanopore DRS) and data processing were performed at the University of Dundee, Dundee, UK. The Nanopore DRS library was prepared using the SQK-RNA001 kit (ONT)[94], using 5 µg of total RNA as input for library preparation, and sequenced on R9.4 SpotON Flow Cells (ONT) using a 48 h runtime. Nanopore DRS reads (SRA: SRR13683864) were base-called using Guppy v2.3.1 (ONT), then corrected using proovread[95] v2.14.1 without sampling. Transcript assemblies were generated with Pinfish (ONT) v0.1.0 from corrected reads aligned to the v2 genome sequence with Minimap2 v2.8 (ref. [96]). More details on Nanopore transcriptome sequencing are in Supplementary Note 3.

Protein-coding gene annotation. Transcript assemblies (TAs) were constructed with PERTRAN[97] v2.4 from 107 M pairs of Illumina RNA-seq reads, combining our data with those from Wu et al.[98] (SRA: SRR1518381 and SRR1518382) and Sarah et al.[99] (SRA: SRR3938623) along with 44k 454 ESTs from Narina et al.[68] (SRA: SAMN00169815, SAMN00169801, SAMN00169798). A merged set of 86,399 TAs were constructed by PASA[100] v2.0.2 from the above RNA-seq TAs along with 53k assemblies from corrected Nanopore DRS reads, and 18 full-length cDNAs collected from NCBI.

Protein-coding genes were predicted with the DOE-JGI Integrated Gene Call[101] (IGC) v5.0 annotation pipeline, which integrates TA evidence and ab initio gene predictions. Briefly, gene loci were determined by TA alignments and/or EXONERATE[102] v2.4.0 peptide alignments from *Arabidopsis thaliana*[39] TAIR10, *Glycine max*[103] Wm82.a4.v1, *Sorghum bicolor*[104] v3.1.1, *Oryza sativa*[105] v7.0, *Setaria viridis*[106] v2.1, *Amborella trichopoda*[107] v1.0, *Zostera marina*[108] v2.2, *Musa acuminata*[109] v1, *Ananas comosus*[51] v3, and *Vitis vinifera*[110] v2.1 proteomes obtained from Phytozome[111] v13 (https://phytozome-next.jgi.doe.gov) and Swiss-Prot[112] proteins (2018, release 11). Gene models were predicted using FGENESH + [113] v3.1.1, FGENESH_EST v2.6, EXONERATE v2.4.0, PASA (v2.0.2) assembly-derived ORFs, and AUGUSTUS v3.3.3 via BRAKER1 v1.9 (ref. [114].). After selecting the best-scoring predictions at each locus (Supplementary Note 3), UTRs and alternative transcripts were added with PASA. Functional annotations were predicted with InterProScan[115] v5.17-56.0. The annotation completeness of this and other Dioscoreaceae species (Supplementary Table 5) were measured using BUSCO[31] v3.0.2-11-g1554283 with the Embryophyta OrthoDB[32] v10 database.

Genomic repeat annotation. Repeat annotation was performed twice (see Supplementary Note 3) with RepeatMasker[116] v4.1.1. The initial round annotated de novo repeats inferred from the preliminary v1 assembly by RepeatModeler[117] v1.0.11, combined with *Dioscorea* repeats deposited in RepBase[118]. The second round used a repeat library inferred by RepeatModeler v2.0.1 (-LTRstruct) from the more complete v2 genome sequence.

Comparisons with other monocot genomes. Orthologous genes were clustered with OrthoFinder[119] v2.4.1 across the available assembled Dioscoreaceae species: *D. alata*, *D. rotundata*[21] (GCA_009730915.1), *D. dumetorum*[34] (GCA_902712375.1), *D. zingiberensis*[22] (GCA_014060945.1), and *Trichopus zeylanicus*[35] (GCA_005019695.1). This procedure produced 5,454 clusters of genes in strict 1:1:1:1 correspondence among the *Dioscorea* species of which 99.9% (n = 5451), 90.5% (n = 4937), and 99.1% (n = 5404) were localized to chromosome-scale scaffolds in *D. alata*, *D. rotundata*, and *D. zingiberensis*, respectively. We also used OrthoFinder to compare a broader set of monocots (*D. alata*, *D. rotundata*, *D. dumetorum*, *D. zingiberensis*, *T. zeylanicus*, *Xerophyta viscosa*[120] (GCA_002076135.1), *Apostasia shenzhenica*[121] (GCA_002786265.1), *Dendrobium catenatum*[122] (GCF_001605985.2), *Asparagus officinalis*[53] (GCF_001876935.1), *Elaeis guineensis*[52] (GCF_000442705.1), *Phoenix*

dactylifera[123] (GCF_000413155.1), *Musa acuminata*[109] (GCF_000313855.2), *Oriza sativa*[124] (GCF_001433935.1), *Zea mays*[125] (GCF_000005005.2), *Ananas comosus*[51] (GCF_001540865.1), *Spirodela polyrhiza*[54,126] (GCA_000504445.1, GCA_001981405.1), *Zostera marina*[108] (GCA_001185155.1)) with *Arabidopsis thaliana*[39,127] (GCF_000001735.4) and *Amborella trichopoda*[107] (GCF_000471905.2) as outgroups. These results are presented graphically in Supplementary Fig. 8 using the ClusterVenn[128] online tool (https://orthovenn2.bioinfotoolkits.net/cluster-venn). See Supplementary Note 3 and Supplementary Data 4 for more detail.

Chromosome landscape, Rabl chromatin structure, and centromere estimates. The A/B compartment structure (Supplementary Fig. 7) for each chromosome was inferred at 100 kb resolution with Knight-Ruiz (KR)-balanced MapQ30 intra-chromosomal HiC count matrices using a custom script (call-compartments v0.1.2-67-g18fff4a; https://bitbucket.org/bredeson/artisanal). Centromeric positions were estimated in JuiceBox (v1.9.0) following the principles described by Varoquaux et al.[129]. Rabl chromatin structure (Supplementary Note 4) was extracted in R[130] v3.5.3 using the prcomp function (chr-structure.R v1.0; https://github.com/bredeson/Dioscorea-alata-genomics) on KR-balanced MapQ30 inter-chromosomal HiC count matrices, with chromosome 2 as the reference comparator. Pearson's correlations (r) between gene count, low-complexity and transposable element repeat densities, recombination rate, and A/B compartment domain status were computed using 500 kb non-overlapping windows with BEDtools[131] v2.28.0 and R[130] v3.5.3 (Supplementary Note 4). Putative centromere sequences and loci (Supplementary Data 2) were determined using a combination of HiC and tandem-repeat finding approaches (Supplementary Note 4).

Synteny and comparative genomics. We used BLASTP[132,133] (BLAST + v2.10.0) to search for homologous proteins between *Dioscorea alata* and each comparator species: *Ananas comosus*[51] (GCF_001540865.1), *D. rotundata*[21] (GCA_009730915.1), *D. dumetorum*[34], *D. zingiberensis*[22] (GCA_014060945.1), *Elaeis guineensis*[52] (GCF_000442705.1), *Spirodela polyrhiza*[54,126] (GCA_000504445.1, GCA_001981405.1), and *Trichopus zeylanicus*[35] (GCA_005019695.1). DIALIGN-TX[134] v1.0.2 and the kaks function from the SeqinR[135] v3.6-1 R[130] (v3.5.3) package were used to calculate synonymous substitution (K_S) rates. Runs of collinear loci (Supplementary Data 2) were inferred using custom filtering and clustering scripts (run-collinearity.sh v1.0, https://github.com/bredeson/Dioscorea-alata-genomics; cluster-collinear-bedpe v0.1.2-67-g18fff4a, https://bitbucket.org/bredeson/artisanal). See Supplementary Note 5 for more details. All ribbon diagrams were generated with the jcvi.graphics.karyotype module in MCscan[136] v1.0.14-0-g58b7710b.

Mapping populations at IITA. Phenotyping of five mapping populations was performed at IITA from 2016–2019. In 2016, mapping populations were planted in single pots and grown in the screenhouse for seed tuber multiplication and screening of anthracnose disease in a controlled environment. In 2017, individual mini-tubers of each mapping population were pre-planted in pots to ensure germination, and one-month-old seedlings were transplanted in the field using a ridge-and-furrow system. Land preparation, weeding, staking and harvesting were carried out following standard field operating protocol for yam[137]. In 2018 and 2019, harvested tubers were cut into mini-sets of 100 g each, treated with pesticide to prevent rotting, and planted in the field as above. More detail on the planting scheme used at IITA may be found in Supplementary Note 6.

Phenotyping for anthracnose disease. Populations were assessed for yam anthracnose disease (YAD) at the International Institute for Tropical Agriculture (IITA, Ibadan, Nigeria) and the National Root Crops Research Institute (NRCRI, Umudike, Nigeria). More detailed descriptions of phenotyping for YAD may be found in Supplementary Note 6; all YAD phenotyping datasets were deposited in Dryad [https://doi.org/10.6078/D1DQ54][87].

For the five IITA populations (TDa1401, TDa1402, TDa1403, TDa1419 and TDa1427), each plant was visually scored in the field in 2017 and 2018 for YAD severity at 3 months after planting (MAP) and 6 MAP using a 1–5 scale as follows: Score 1 = No symptoms, Score 2 = 1–25%, Score 3 = 25–50%, Score 4 = 50–75%, Score 5 ≥ 75%. Detached leaf assays (DLA) were performed at IITA on plants grown in the screenhouse in 2016, and on plants grown in the field in 2017 and 2018, following a modified protocol of Green et al.[138] and Nwadili et al.[139].

At NRCRI, site-specific *C. gloeosporioides* isolates were collected and evaluated, as described in Supplementary Note 6. The most virulent isolate was used for anthracnose severity evaluation of NRCRI *D. alata* mapping populations using DLA[139].

Phenotyping for post-harvest tuber traits. Tuber dry matter content was phenotyped at IITA. After harvest, healthy yam tubers were sampled in each replication for dry matter determination. The tubers of each genotype were cleaned with water to remove soil particles. Thereafter, the tubers were peeled and grated for easy oven drying; 100 g of freshly grated tuber flesh sample was weighed, put into a Kraft paper bag, and dried at 105 °C for 16 h. After drying, the weight of each

sample was recorded and the dry matter content was determined using Eq. 1:

$$\% \text{ Dry matter content} = 100 \cdot \frac{\text{weight of dry sample } (g)}{\text{weight of fresh sample } (g)} \quad (1)$$

Tuber flesh color and oxidation/oxidative browning were phenotyped at IITA. After harvest, one well-developed and mature representative tuber was sampled in each replication. The sampled tuber was peeled, cut, and chipped with a hand chipper to get small thickness size pieces. A chromameter (CR-410, Konica Minolta, Japan) was used to read the total color of sampled pieces placed on a petri dish immediately and exposure to air at 0, 30, and 180 min. The lightness (L^*), red/green coordinate (a^*), and yellow/blue coordinate (b^*) parameters were recorded for each chromameter reading for the determination of the total color difference. A reference white porcelain tile was used to calibrate the chromameter before each determination[140]. Tuber whiteness was calculated with Eq. 2:

$$f_L = \frac{L^{*2}}{L^{*2} + a^{*2} + b^{*2}} \quad (2)$$

where ΔL^* = difference in lightness and darkness ([+] = lighter, [−] = darker), Δa^* = difference in red and green ([+] = redder, [−] = greener), and Δb^* = difference in yellow and blue ([+] = yellower, [−] = bluer) (http://docs-hoffmann.de/cielab03022003.pdf).

Tuber flesh oxidation was estimated from the total variation from the difference in the final and initial color reading, as in Eq. 3:

$$\text{Tuber flesh oxidation} = E_{\text{final}} - E_{\text{initial}} \quad (3)$$

where ΔE_{final} = color reader value at the final time (30 min) and $\Delta E_{\text{initial}}$ = Initial color reader value (at 0 min).

Tubers were evaluated post-harvest at NRCRI. Of the three populations evaluated at NRCRI, 172 progeny survived. As soon as the yam tubers were harvested, eight traits were assessed using the descriptors from Asfaw[137]: presence or absence of corm (CORM: 0 = absent; 1 = present), the ability of corm to separate (CORSEP: 0 = no; 1 = yes), type of corm (CORTYP: 1 = regular; 2 = transversally elongated; 3 = branched), tuber shape (TBRS: 1 = spherical/round; 2 = oval; 3 = cylindrical; 5 = irregular), tuber size (TBRSZ: 1 = small, length less than 15 cm; 2 = medium, length between 15 and 25 cm; 3 = big, length longer than 25 cm), tuber surface texture (TBRST: 1 = smooth; 2 = rough), roots on tuber (RTBS: 0 = no roots; 2 = few; 3 = many) and position of roots on tuber (PRTBS: 1 = lower; 2 = middle; 3 = upper; 4 = entire tuber). Tuber trait phenotyping datasets for all mapping populations were deposited in Dryad [https://doi.org/10.6078/D1DQ54][87].

QTL analysis. QTL association analyses integrated linkage maps, imputed genotype data, and phenotype data into Binary PED files using PLINK[141,142] v1.90b6.16. Only progeny samples with both genotype and phenotype data were retained per trait. Some traits were initially scored using a discrete 0–2 system, which PLINK assumes are missing/case/control phenotypes; these trait values were shifted out of the 0–2 range before analysis by adding an offset of 1 or 2 to all values (depending on initial data range). An independent QTL association analysis was performed for each trait using logistic regression. Per-locus Wald statistic p-values were adjusted for multiple testing by max(T) correction[141,143] with 1×10^6 phenotype label-swap permutations. A locus was considered significant if the empirical max(T)-corrected p-value was less than $\alpha = 0.05$. Two dry matter phenotype measurements were excluded from the TDa1419 population: TDa1419_485 (a likely typographical error in data collection) and TDa1419_142 (an extreme outlier value).

For each identified QTL, an effect plot was generated to determine the dominance pattern and estimate narrow-sense heritability (h^2) at the peak marker. Effect plots and h^2 were calculated as described by Broman and Sen[144] (pg. 122) using a custom R[130] script (plot-qtl-gxp.R v1.0, https://github.com/bredeson/Dioscorea-alata-genomics). The effect status (i.e., dominance) for chromosomes 6 and 19 anthracnose QTL could not be determined because the alleles at these loci are segregated in pseudo-testcross configurations. The interval around each QTL peak (Table 3) was determined by expanding the interval boundaries upstream and downstream of the peak marker until another marker with linkage disequilibrium (LD) below 0.9 was encountered (plot-qtl-ld.R v1.0, https://github.com/bredeson/Dioscorea-alata-genomics). The gene loci contained within these intervals, and their functional annotations, are provided in Dryad [https://doi.org/10.6078/D1DQ54][87]. In addition to the predicted functional annotations (Supplementary Note 3) for each *D. alata* gene, protein descriptions were included from the best BLASTP[133] (-seg yes -lcase_masking -soft_masking true -evalue 1e-6) hits to the NCBI RefSeq proteomes (release 207, 2021-07-15) of *Arabidopsis thaliana*, *Gossypium hirsutum*, *Ipomoea batatas*, *Malus domestica*, *Medicago truncatula*, *Musa acuminata*, *Nicotiana tabacum*, *Oryza sativa Japonica*, *Solanum lycopersicum*, *Solanum tuberosum*, *Vitis vinifera*, and *Zea mays* when searching for causal gene candidates within QTL intervals.

WGS Illumina sequencing. DNA samples from the breeding lines listed in Supplementary Table 1 were isolated at IITA (Supplementary Note 7). TruSeq Illumina libraries were constructed and sequenced at the VCGSL. Inferred insert sizes ranged from 247–876 bp. These libraries were sequenced on HiSeq 2500 or HiSeq 4000 with reading lengths ranging from 150–251 bp, yielding combined sample

depths of 19–230×. Supplementary Data 1 lists all Illumina sequence data from our breeding lines, including external data, and accompanying summary statistics.

WGS variant calling. Single-nucleotide variants (SNVs) were called from the whole-genome resequencing datasets listed in Supplementary Data 1. Briefly, Illumina reads were screened for TruSeq adapters with fastq-mcf (ea-utils[145] tool suite) v1.04.807-18-gbd148d4, then aligned with BWA-MEM[146] v0.7.17-11-g20d0a13 to a TDa95/00328 v2 genome index containing *D. alata* plastid (GenBank: MZ848367.1 [https://www.ncbi.nlm.nih.gov/nuccore/MZ848367.1]) and mitochondrial (GenBank: OK106275.1 [https://www.ncbi.nlm.nih.gov/nuccore/OK106275.1]) sequences and a *Pseudomonas fluorescens* chromosome (GenBank: CP081968.1 [https://www.ncbi.nlm.nih.gov/nuccore/CP081968.1]) as bait for contaminant reads. BAM files were processed with SAMtools[147] v1.9-93-g0ca96a4 to fix mate information, mark duplicates, sort, merge, and filter for properly-paired reads. Initial SNVs and indels were called with the Genome Analysis ToolKit[148] (GATK; v3.8-1-0-gf15c1c3ef) HaplotypeCaller and GenotypeGVCFs tools. False-positive variant and genotype calls were filtered using individual-specific minimum- and maximum-depth cutoffs, allele-balance binomial test thresholds ($\alpha = 0.001$; Supplementary Fig. 14), a read depth mask, and annotated repeat masks. See Supplementary Note 7 for a more complete description of the filtering protocol used. Only biallelic SNVs were used in downstream analyses and effect predictions were annotated with SnpEff[149] v5.0.c2020-11-25.

WGS population analyses. Using 1.89 million SNVs with 75% or more of individuals genotyped, pairwise genome-wide relatedness estimates were obtained with VCFtools[89] v0.1.16-16-g954e607. The resulting relatedness network and origination year encoded in each sample's identifier were used to verify IITA pedigrees. The intrinsic heterozygosity and autozygosity of each individual, as well as the pairwise segmental (5000 SNV windows, 1000 SNV step) identity-by-descent (IBD) of each, were estimated with custom tools (snvrate and IBD tools v1.0-26-g4cf73ab, https://bitbucket.org/rokhsar-lab/wgs-analysis). A 100 kb sliding window (10 kb step) was called autozygous if the rate of intrinsic heterozygosity was less than 2×10^{-4}. This threshold was determined empirically (Supplementary Fig. 4, Supplementary Note 7).

Mitochondrial and plastid sequence assemblies and phylogenetics. Mitochondrial and plastid DNA sequences were assembled using de novo and comparative methods (Supplementary Note 8). The IboSweet3 *D. dumetorum* plastid was extracted from the Siadjeu et al.[34] assembly. Our Dioscoreaceae DNA phylogeny was built from plastid long single-copy regions using MAFFT[150,151] FFT-NS-i v7.427 (--6merpair --maxiterate 1000), Gblocks v0.91b, and PhyML[152] v3.3.20190909 (--leave_duplicates --freerates -a e -d nt -b 1000 -f m -o tlr -t e -v e). The monocot plastid phylogeny was constructed using OrthoFinder[119,153,154] v2.4.1 (MAFFT v7.427 alignment and IQ-TREE[155] v2.0.3 phylogenetic reconstruction) single-copy orthologs. All trees were visualized with FigTree v1.4.4 (https://github.com/rambaut/figtree).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
A reporting summary for this article is available as a Supplementary Information file. The genome sequence, annotation, and SNP data are browsable at Phytozome [https://phytozome-next.jgi.doe.gov/info/Dalata_v2_1] or YamBase [https://yambase.org/organism/Dioscorea_alata/genome]. The *D. alata* TDa95/00328 nuclear genome (GCA_020875875.1), transcriptome (GJIX00000000.1), plastid (MZ848367.1), and mitochondrion (OK106275.1) assemblies, and *Pseudomonas fluorescens* chromosome (CP081968.1) were deposited in the NCBI GenBank database. *D. rotundata* TDr96_F1 and *D. dumetorum* IboSweet3 plastid sequences were also deposited in the NCBI GenBank database under accessions MZ848368.1 and MZ848369.1, respectively. All sequencing read data generated for this work were deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA666450; see Supplementary Data 1 for individual sample SRA metadata. The genetic linkage maps, phenotype datasets, and DArTseq genotype datasets for all populations, as well as functional annotations for all genes within QTL intervals, were deposited in Dryad [https://doi.org/10.6078/D1DQ54][87]. Source Data files are provided with this work. Source data are provided with this paper.

Code availability
Analysis scripts used throughout this work are available at Github [https://github.com/bredeson/Dioscorea-alata-genomics] (tag 'v1.0') and Bitbucket: [https://bitbucket.org/rokhsar-lab/wgs-analysis] (v1.0-26-g4cf73ab), [https://bitbucket.org/rokhsar-lab/gbs-analysis] (v1.4.1-11-g19a5f3a), and [https://bitbucket.org/bredeson/artisanal] (v0.1.2-67-g18fff4a).

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-022-29114-w

ARTICLE

References

1. Mignouna, H. D., Abang, M. M. & Asiedu, R. In *Genomics of Tropical Crop Plants* (eds. Moore, P. H. & Ming, R.) 549–570 (Springer New York, 2008).
2. Lebot, V. *Tropical Root and Tuber Crops*, 2nd edn. (CABI, 2019).
3. Coursey, D. G. *Yams. An account of the nature, origins, cultivation and utilisation of the useful members of the Dioscoreaceae* (Longmans, Green and Co. Ltd, London, 1968).
4. Zannou, A. et al. Yam and cowpea diversity management by farmers in the Guinea-Sudan transition zone of Benin. *NJAS Wagening. J. Life Sci.* **52**, 393–420 (2004).
5. Obidiegwu, J. E. & Akpabio, E. M. The geography of yam cultivation in southern Nigeria: exploring its social meanings and cultural functions. *J. Ethn. Foods* **4**, 28–35 (2017).
6. Power, R. C., Güldemann, T., Crowther, A. & Boivin, N. Asian crop dispersal in Africa and late Holocene human adaptation to tropical environments. *J. World Prehistory* **32**, 353–392 (2019).
7. Hahn, S. K. Yams. In *Evolution of crop plants* (eds. Smartt, J. & Simmonds, N. W.) 112–120 (Wiley-Blackwell, 1995).
8. Sartie, A. & Asiedu, R. Segregation of vegetative and reproductive traits associated with tuber yield and quality in water yam (*Dioscorea alata* L.). *Afr. J. Biotechnol.* **13**, 2807–2818 (2014).
9. Muzac-Tucker, I., Asemota, H. N. & Ahmad, M. H. Biochemical composition and storage of Jamaican yams (*Dioscorea* sp). *J. Sci. Food Agric.* **62**, 219–224 (1993).
10. Obidiegwu, J. E., Lyons, J. B. & Chilaka, C. A. The *Dioscorea* genus (yam)-an appraisal of nutritional and therapeutic potentials. *Foods* **9**, 1304 (2020).
11. Darkwa, K., Olasanmi, B., Asiedu, R. & Asfaw, A. Review of empirical and emerging breeding methods and tools for yam (*Dioscorea* spp.) improvement: status and prospects. *Plant Breed.* **139**, 474–497 (2020).
12. Malapa, R., Arnau, G., Noyer, J. L. & Lebot, V. Genetic diversity of the greater yam (*Dioscorea alata* L.) and relatedness to *D. nummularia* Lam. and *D. transversa* Br. as revealed with AFLP markers. *Genet. Resour. Crop Evol.* **52**, 919–929 (2005).
13. Arnau, G., Nemorin, A., Maledon, E. & Abraham, K. Revision of ploidy status of *Dioscorea alata* L. (Dioscoreaceae) by cytogenetic and microsatellite segregation analysis. *Theor. Appl. Genet.* **118**, 1239–1249 (2009).
14. Arnau, G. et al. Yams. In *Root and Tuber Crops* (ed. Bradshaw, J. E.) 127–148 (Springer New York, 2010).
15. Winch, J. E., Newhook, F. J., Jackson, G. V. H. & Cole, J. S. Studies of *Colletotrichum gloeosporioides* disease on yam, *Dioscorea alata*, in Solomon Islands. *Plant Pathol.* **33**, 467–477 (1984).
16. Nwankiti, A. O., Okpala, E. U. & Odurukwe, S. O. Effect of planting dates on the incidence and severity of anthracnose/blotch disease complex of *Dioscorea alata* L., caused by *Colletotrichum gloeosporioides* Penz., and subsequent effects on the yield. *Beitr. Trop. Landwirtsch. Veterinarmed.* **22**, 288–292 (1984).
17. Mignucci, J. S., Hepperly, P. R., Green, J., Torres-López, R. & Figueroa, L. A. Yam protection II. Anthracnose, yield, and profit of monocultures and interplantings. *J. Agric. Univ. Puerto Rico* **72**, 179–189 (1988).
18. Abang, M. M., Winter, S., Mignouna, H. D., Green, K. R. & Asiedu, R. Molecular taxonomic, epidemiological and population genetic approaches to understanding yam anthracnose disease. *Afr. J. Biotechnol.* **2**, 486–496 (2003).
19. Egesi, C. N., Odu, B. O., Ogunyemi, S., Asiedu, R. & Hughes, J. Evaluation of water yam (*Dioscorea alata* L.) germplasm for reaction to yam anthracnose and virus diseases and their effect on yield. *J. Phytopathol.* **155**, 536–543 (2007).
20. Lebot, V., Abraham, K., Kaoh, J., Rogers, C. & Molisalé, T. Development of anthracnose resistant hybrids of the Greater Yam (*Dioscorea alata* L.) and interspecific hybrids with *D. nummularia* Lam. *Genet. Resour. Crop Evol.* **66**, 871–883 (2019).
21. Sugihara, Y. et al. Genome analyses reveal the hybrid origin of the staple crop white Guinea yam (*Dioscorea rotundata*). *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.2015830117 (2020).
22. Cheng, J. et al. The origin and evolution of the diosgenin biosynthetic pathway in yam. *Plant Commun.* **2**, 100079 (2021).
23. Mignouna, H. et al. A genetic linkage map of water yam (*Dioscorea alata* L.) based on AFLP markers and QTL analysis for anthracnose resistance. *Theor. Appl. Genet.* **105**, 726–735 (2002).
24. Petro, D., Onyeka, T. J., Etienne, S. & Rubens, S. An intraspecific genetic map of water yam (*Dioscorea alata* L.) based on AFLP markers and QTL analysis for anthracnose resistance. *Euphytica* **179**, 405–416 (2011).
25. Bhattacharjee, R. et al. An EST-SSR based genetic linkage map and identification of QTLs for anthracnose disease resistance in water yam (*Dioscorea alata* L.). *PLoS ONE* **13**, e0197717 (2018).
26. Cormier, F. et al. A reference high-density genetic map of greater yam (*Dioscorea alata* L.). *Theor. Appl. Genet.* **132**, 1733–1744 (2019).
27. Mignouna, H. D., Abang, M. M., Green, K. R. & Asiedu, R. Inheritance of resistance in water yam (*Dioscorea alata*) to anthracnose (*Colletotrichum gloeosporioides*). *Theor. Appl. Genet.* **103**, 52–55 (2001).
28. Cowan, C. R., Carlton, P. M. & Cande, W. Z. The polar arrangement of telomeres in interphase and meiosis. *Rabl Organ. Bouquet Plant Physiol.* **125**, 532–538 (2001).
29. Mascher, M. et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
30. Muller, H., Gil, J. Jr & Drinnenberg, I. A. The impact of centromeres on spatial genome architecture. *Trends Genet.* **35**, 565–578 (2019).
31. Waterhouse, R. M. et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
32. Kriventseva, E. V. et al. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
33. Dong, P. et al. 3D chromatin architecture of large plant genomes determined by local A/B compartments. *Mol. Plant* **10**, 1497–1509 (2017).
34. Siadjeu, C., Pucker, B., Viehöver, P., Albach, D. C. & Weisshaar, B. High contiguity de novo genome sequence assembly of trifoliate yam (*Dioscorea dumetorum*) using long read sequencing. *Genes* **11**, 274 (2020).
35. Chellappan, B. V. et al. High quality draft genome of Arogyapacha (*Trichopus zeylanicus*), an important medicinal plant endemic to western Ghats of India. *G3 Genes Genomes Genet.* **9**, 2395–2404 (2019).
36. Scarcelli, N., Daïnou, O., Agbangla, C., Tostain, S. & Pham, J.-L. Segregation patterns of isozyme loci and microsatellite markers show the diploidy of African yam *Dioscorea rotundata* (2n = 40). *Theor. Appl. Genet.* **111**, 226–232 (2005).
37. Tamiru, M. et al. Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. *BMC Biol.* **15**, 86 (2017).
38. Huang, X. & Guo, H. Karyotype of different ploidy *Dioscorea zingiberensis* CH Wright. *J. Trop. Subtrop. Bot.* **20**, 256–262 (2012).
39. Lamesch, P. et al. The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).
40. Baquar, S. R. Chromosome behaviour in Nigerian yams (*Dioscorea*). *Genetica* **54**, 1–9 (1980).
41. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
42. Ren, R. et al. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Mol. Plant* **11**, 414–428 (2018).
43. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).
44. Schubert, I. & Lysak, M. A. Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.* **27**, 207–216 (2011).
45. Garsmeur, O. et al. Two evolutionarily distinct classes of paleopolyploidy. *Mol. Biol. Evol.* **31**, 448–454 (2014).
46. Shi, T. et al. Distinct expression and methylation patterns for genes with different fates following a single whole-genome duplication in flowering plants. *Mol. Biol. Evol.* **37**, 2394–2413 (2020).
47. Langham, R. J. et al. Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* **166**, 935–945 (2004).
48. Cheng, F. et al. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat. Plants* **4**, 258–268 (2018).
49. Edger, P. P., McKain, M. R., Bird, K. A. & VanBuren, R. Subgenome assignment in allopolyploids: challenges and future directions. *Curr. Opin. Plant Biol.* **42**, 76–80 (2018).
50. Jiao, Y., Li, J., Tang, H. & Paterson, A. H. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* **26**, 2792–2802 (2014).
51. Ming, R. et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* **47**, 1435–1442 (2015).
52. Singh, R. et al. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature* **500**, 335–339 (2013).
53. Harkess, A. et al. The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nat. Commun.* **8**, 1279 (2017).
54. Wang, W. et al. The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nat. Commun.* **5**, 3311 (2014).
55. Egesi, C. N., Onyeka, T. J. & Asiedu, R. Severity of anthracnose and virus diseases of water yam (*Dioscorea alata* L.) In Nigeria I: effects of yam genotype and date of planting. *Crop Prot.* **26**, 1259–1265 (2007).
56. Ron, M. & Avni, A. The receptor for the fungal elicitor ethylene-inducing xylanase is a member of a resistance-like gene family in tomato. *Plant Cell* **16**, 1604–1615 (2004).
57. Eulgem, T. et al. EDM2 is required for RPP7-dependent disease resistance in *Arabidopsis* and affects RPP7 transcript levels. *Plant J.* **49**, 829–839 (2007).
58. Tsuchiya, T. & Eulgem, T. EMSY-like genes are required for full RPP7-mediated race-specific immunity and basal defense in *Arabidopsis*. *Mol. Plant. Microbe Interact.* **24**, 1573–1581 (2011).

59. Song, J. et al. Gene RB cloned from *Solanum bulbocastanum* confers broad spectrum resistance to potato late blight. *Proc. Natl Acad. Sci. USA* **100**, 9128–9133 (2003).

60. Tang, D., Ade, J., Frye, C. A. & Innes, R. W. Regulation of plant defense responses in *Arabidopsis* by EDR2, a PH and START domain-containing protein. *Plant J.* **44**, 245–257 (2005).

61. Vorwerk, S. et al. EDR2 negatively regulates salicylic acid-based defenses and cell death during powdery mildew infections of *Arabidopsis thaliana*. *BMC Plant Biol.* **7**, 35 (2007).

62. Agre, P. A. et al. Identification of QTLs controlling resistance to anthracnose disease in water yam (*Dioscorea alata*). *Genes.* **13**, 1–15 (2022).

63. Martin, F. W. & Ruberte, R. Polyphenol of *Dioscorea alata* (yam) tubers associated with oxidative browning. *J. Agric. Food Chem.* **24**, 67–70 (1976).

64. Akissoe, N., Mestres, C., Hounhouigan, J. & Nago, M. Biochemical origin of browning during the processing of fresh Yam (*Dioscorea* spp.) into dried product. *J. Agric. Food Chem.* **53**, 2552–2557 (2005).

65. Jia, G.-L., Shi, J.-Y., Song, Z.-H. & Li, F.-D. Prevention of enzymatic browning of Chinese yam (*Dioscorea* spp.) using electrolyzed oxidizing water. *J. Food Sci.* **80**, C718–C728 (2015).

66. Goenaga, R. J. & Irizarry, H. Accumulation and partitioning of dry matter in water yam. *Agron. J.* **86**, 1083–1087 (1994).

67. Gatarira, C. et al. Genome-wide association analysis for tuber dry matter and oxidative browning in water yam (*Dioscorea alata* L.). *Plants* **9**, 969 (2020).

68. Narina, S. S. et al. Generation and analysis of expressed sequence tags (ESTs) for marker development in yam (*Dioscorea alata* L.). *BMC Genomics* **12**, 100 (2011).

69. Saski, C. A., Bhattacharjee, R., Scheffler, B. E. & Asiedu, R. Genomic resources for water yam (*Dioscorea alata* L.): Analyses of EST-sequences, de novo sequencing and GBS libraries. *PLoS ONE* **10**, e0134031 (2015).

70. Bredeson, J. V. et al. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat. Biotechnol.* **34**, 562–570 (2016).

71. Wu, G. A. et al. Genomics of the origin and evolution of *Citrus*. *Nature* **554**, 311–316 (2018).

72. Wolfe, M. D. et al. Historical introgressions from a wild relative of modern cassava improved important traits and may be under balancing selection. *Genetics* **213**, 1237–1253 (2019).

73. Sharif, B. M. et al. Genome-wide genotyping elucidates the geographical diversification and dispersal of the polyploid and clonally propagated yam (*Dioscorea alata*). *Ann. Bot.* **126**, 1029–1038 (2020).

74. Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e23 (2020).

75. Todesco, M. et al. Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* https://doi.org/10.1038/s41586-020-2467-6 (2020).

76. Ihediohanm, N. C., Onuegbu, N. C., Peter-Ikec, A. I. & Ojimba, N. C. A comparative study and determination of Glycemic Indices of three yam cultivars (*Dioscorea rotundata*, *Dioscorea alata* and *Dioscorea domentorum*). *Pak. J. Nutr.* **11**, 547–552 (2012).

77. Oko, A. O. & Famurewa, A. C. Estimation of nutritional and starch characteristics of *Dioscorea alata* (water yam) varieties commonly cultivated in the South-Eastern Nigeria. *Br. J. Appl. Sci. Technol.* **6**, 145–152 (2014).

78. Doležel, J., Sgorbati, S. & Lucretti, S. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol. Plant.* **85**, 625–631 (1992).

79. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

80. Koren, S. et al. Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

81. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).

82. Dudchenko, O., Shamim, M. S., Batra, S. S. & Durand, N. C. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *Biorxiv.* https://www.biorxiv.org/content/10.1101/254797v1 (2018).

83. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).

84. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).

85. Chin, C.-S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).

86. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv.* https://arxiv.org/abs/1207.3907 (2012).

87. Bredeson, J. V. et al. Chromosome evolution and the genetic basis of agronomically important traits in greater yam. *Dryad. Dataset.* https://doi.org/10.6078/D1DQ54 (2021).

88. International Cassava Genetic Map Consortium (ICGMC). High-resolution linkage map and chromosome-scale genome assembly for cassava (*Manihot esculenta* Crantz) from 10 populations. *G3 Genes Genomes Genet.* **5**, 133–144 (2015).

89. Danecek, P. et al. The Variant Call Format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

90. Whalen, A., Gorjanc, G. & Hickey, J. M. AlphaFamImpute: High accuracy imputation in full-sib families from genotype-by-sequencing data. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa499 (2020).

91. Van Ooijen, J. W. *JoinMap 4: Software for the calculation of genetic linkage maps in experimental populations of diploid species* (Plant Research International BV and Kayazma BV, 2006).

92. Van Ooijen, J. W. Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genet. Res.* **93**, 343–349 (2011).

93. Endelman, J. B. & Plomion, C. LPmerge: An R package for merging genetic maps by linear programming. *Bioinformatics* **30**, 1623–1624 (2014).

94. Parker, M. T. et al. Nanopore direct RNA sequencing maps the complexity of *Arabidopsis* mRNA processing and m6A modification. *Elife* **9**, e49658 (2020).

95. Hackl, T., Hedrich, R., Schultz, J. & Förster, F. proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004–3011 (2014).

96. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

97. Lovell, J. T. et al. The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nat. Commun.* **9**, 5213 (2018).

98. Wu, Z.-G. et al. Transcriptome analysis reveals flavonoid biosynthesis regulation and simple sequence repeats in yam (*Dioscorea alata* L.) tubers. *BMC Genomics* **16**, 346 (2015).

99. Sarah, G. et al. A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. *Mol. Ecol. Resour.* **17**, 565–580 (2017).

100. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).

101. Shu, S., Rokhsar, D., Goodstein, D., Hayes, D. & Mitros, T. *JGI Plant Genomics Gene Annotation Pipeline.* https://www.osti.gov/biblio/1241222 (2014).

102. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

103. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).

104. McCormick, R. F. et al. The *Sorghum bicolor* reference genome: Improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* **93**, 338–354 (2018).

105. Ouyang, S. et al. The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* **35**, D883–D887 (2007).

106. Mamidi, S. et al. A genome resource for green millet *Setaria viridis* enables discovery of agronomically valuable loci. *Nat. Biotechnol.* **38**, 1203–1210 (2020).

107. Amborella Genome Project. The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).

108. Olsen, J. L. et al. The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature* **530**, 331–335 (2016).

109. D'Hont, A. et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213–217 (2012).

110. The French–Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).

111. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).

112. UniProt Consortium, T. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).

113. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).

114. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).

115. Jones, P. et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

116. Smit, A. F. A., Hubley, R. & Green, P. *RepeatMasker Open-4.0.* https://www.repeatmasker.org/ (2013–2015).

117. Smit, A. F. A. & Hubley, R. *RepeatModeler Open-1.0.* https://www.repeatmasker.org/ (2008–2015).

118. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).

119. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

120. Costa, M.-C. D. et al. A footprint of desiccation tolerance in the genome of *Xerophyta viscosa*. *Nat. Plants* **3**, 17038 (2017).

121. Zhang, G.-Q. et al. The *Apostasia* genome and the evolution of orchids. *Nature* **549**, 379–383 (2017).

122. Zhang, G.-Q. et al. The *Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. *Sci. Rep.* **6**, 19029 (2016).

123. Al-Mssallem, I. S. et al. Genome sequence of the date palm *Phoenix dactylifera* L. *Nat. Commun.* **4**, 2274 (2013).

124. Kawahara, Y. et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**, 4 (2013).

125. Jiao, Y. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).

126. Michael, T. P. et al. Comprehensive definition of genome features in *Spirodela polyrhiza* by high-depth physical mapping and short-read DNA sequencing strategies. *Plant J.* **89**, 617–635 (2017).

127. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000).

128. Xu, L. et al. OrthoVenn2: A web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **47**, W52–W58 (2019).

129. Varoquaux, N. et al. Accurate identification of centromere locations in yeast genomes using Hi-C. *Nucleic Acids Res.* **43**, 5331–5339 (2015).

130. R Core Team. *R: A language and environment for statistical computing.* (Foundation for Statistical Computing, 2013).

131. Quinlan, A. R. BEDTools: The Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1–34 (2014).

132. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

133. Camacho, C. et al. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

134. Subramanian, A. R., Kaufmann, M. & Morgenstern, B. DIALIGN-TX: Greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol. Biol.* **3**, 6 (2008).

135. Charif, D. & Lobry, J. R. In *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* (eds. Bastolla, U., Porto, M., Roman, H. E. & Vendruscolo, M.) 207–232 (Springer Berlin Heidelberg, 2007).

136. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).

137. Asfaw, A. Standard operating protocol for yam variety performance evaluation trial. Vol. **27** (IITA, Ibadan, Nigeria, 2016).

138. Green, K. R., Abang, M. M. & Iloba, C. A rapid bioassay for screening yam germplasm for response to anthracnose. *Tropical Sci.* **40**, 132–138 (2000).

139. Nwadili, C. O. et al. Comparative reliability of screening parameters for anthracnose resistance in water yam (*Dioscorea alata*). *Plant Dis.* **101**, 209–216 (2017).

140. Tenorio Cavalcante, P. M. et al. The influence of microstructure on the performance of white porcelain stoneware. *Ceram. Int.* **30**, 953–963 (2004).

141. 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).

142. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

143. Browning, B. L. PRESTO: Rapid calculation of order statistic distributions and multiple-testing adjusted *P*-values via permutation for one and two-stage genetic association studies. *BMC Bioinformatics* **9**, 309 (2008).

144. Broman, K. W. & Sen, S. A Guide to QTL mapping with R/qtl. *Stat. Biol. Health* https://doi.org/10.1007/978-0-387-92125-9 (2009).

145. Aronesty, E. Comparison of sequencing utility programs. *The Open Bioinformatics Journal* **7**, 1–8 (2013).

146. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* https://arxiv.org/abs/1303.3997 (2013).

147. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

148. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *Cold Spring Harb. Lab.* https://doi.org/10.1101/201178 (2017).

149. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

150. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

151. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

152. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

153. Emms, D. M. & Kelly, S. STRIDE: Species Tree Root Inference from Gene Duplication Events. *Mol. Biol. Evol.* **34**, 3267–3278 (2017).

154. Emms, D. M. & Kelly, S. STAG: Species Tree Inference from All Genes. *Cold Spring Harb. Lab.* https://doi.org/10.1101/267914 (2018).

155. Minh, B. Q. et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

156. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

Acknowledgements

Author contributions

Conceived, designed, and led study: D.S.R., R.B., J.E.O., J.V.B., J.B.L. Genome assembly and chromatin structure, chromosome landscape, comparative genomics, chromosome evolution, population genetic, and phylogenetic analyses: J.V.B. (lead), D.S.R. Genome sequencing planning and coordination: D.S.R., J.V.B., A.V.D., J.B.L. Genetic mapping: J.V.B. (lead), J.B.L. QTL analysis: J.V.B. Overall project management: J.B.L. Mapping population development: A.L.M., A.A. Mapping population management/propagation: R.B., I.O.O., A.A., J.N., I.N. Development of and info on breeding lines: R.A., A.L.M. Phenotyping of mapping populations: I.O.O., O.K., A.A., P.L.K., N.R.O., C.O.N., I.N., J.N. Preparation of cell nuclei for HiC analysis; karyotype and chromosome counting: J.D. (lead), E.H. Nanopore DRS sequencing and analysis: M.P., K.K., A.V.S., G.J.B., G.G.S. (lead). DNA isolation for reference genome, sequencing of breeding lines, and genotyping: I.O.O., N.R.O., J.N., R.K., S.M., P.S.H. RNA isolation: R.K., S.M., P.S.H. (lead). Provision of RNA-seq data: J.F. Wrote manuscript: D.S.R., J.V.B., J.B.L., J.E.O., O.K., N.R.O., C.N., R.B., E.H. with input from A.V.D., G.G.S., J.D. Annotation and database management: D.G. (lead), S.S., J.C. Other project planning/site-specific supervision: I.O.O., C.N.E., R.J., AM.

Competing interests

D.S.R. is a member of the Scientific Advisory Board of, and a minor shareholder in, Dovetail Genomics LLC, which provides as a service the high-throughput chromatin conformation capture (HiC) technology used in this study. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-29114-w.

Correspondence and requests for materials should be addressed to Jude E. Obidiegwu, Ranjana Bhattacharjee or Daniel S. Rokhsar.

Peer review information *Nature Communications* thanks Todd Michael and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints