Recurrent evolution of cryptic triploids in cultivated enset increases yield

Yann Dussert[1,2*], James S. Borrell[3], Jonathan Stocks[3], Harriet V. Hunt[3], Oliver W. White[4], Paul Wilkin[3], Richard Buggs[1,3], Lucie Büchi[5], Sebsebe Demissew[6], Feleke Woldeyes[7], Ilia J. Leitch[3], Wendawek M. Abebe[6], Richard A. Nichols[1]

1. School of Biological and Chemical Sciences, Queen Mary University of London, London
2. Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, 91057, France
3. Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, UK
4. Natural History Museum, Cromwell Road, South Kensington, London, SW7 5BD, UK
5. Natural Resources Institute, University of Greenwich, Chatham Maritime, Kent, ME4 4TB, UK
6. Addis Ababa University, Addis Ababa, Ethiopia
7. Ethiopian Biodiversity Institute, Addis Ababa, Ethiopia

*Corresponding author: ydussert@genoscope.cns.fr

ABSTRACT

Enset is a staple crop for over 20 million people in southwestern Ethiopia. We generated a new chromosome-level genome assembly, and sequenced 65 wild and 658 cultivated individuals. We unexpectedly find that 214 cultivated individuals are triploid, from multiple origins. Landrace names consistently differentiate diploids from the triploids, which have higher pseudostem volumes and are disproportionately planted. Scientifically directed breeding of triploids could improve food security in sub-saharan Africa.

MAIN TEXT

Polyploidy is a common feature of crop genomes sometimes arising after, but more commonly before, domestication (Akagi et al., 2022; Salman-Minkov et al., 2016). Enset (*Ensete ventricosum*), belonging to the Musaceae family like bananas and plantains, is cultivated in the Ethiopian Highlands by smallholder farmers for its starch-rich corm and pseudostem, normally via clonal propagation of favoured lines (Borrell et al., 2019). These are genetically differentiated from wild enset co-occuring in the region and reproducing sexually (White et al., 2023). Prior to the current study, only diploid cytotypes (2n = 2x = 18) (Cheesman, 1947) had been reported.

We sequenced an enset plant grown from a seed of the cultivated variety 'Maze', using PacBio HiFi and Omni-C sequencing. The resulting pseudo-haploid assembly (806

contigs, total size: 534 Mb) has nine large scaffolds, ranging from 35.9 Mb to 68.0 Mb in size and containing 92% of the assembly (Supplementary Fig. 1, Supplementary Table 1). Gene prediction, informed by RNA-seq data from four different tissues, produces a total of 35,238 protein-coding gene models. The assembly and gene prediction include respectively 98.6% and 96.2% of conserved BUSCO orthologs. A k-mer analysis (Supplementary Fig. 2) showed that the sequenced individual was diploid and had high genome-wide heterozygosity (1.10%).

We conducted reduced-representation genome sequencing on 658 cultivated individuals collected in south-western Ethiopia at multiple farms along altitudinal transects, and 65 wild individuals from locations across the partly sympatric wild range (Supplementary Fig. 3), mapping the data to our genome assembly. We assessed ploidy level by examination of allelic balance (AB, i.e. the proportion of reads supporting each major allele for heterozygous sites) in each individual (Fig. 1A). Distributions for diploid individuals are expected to have a modal value around 0.5, and triploids around 0.66. Thirty individuals (4%) had aberrant distributions, generally with higher modal values (peak around 0.9 appearing to be an artifact of the cut-off parameters used for allele calling). Whole genome resequencing of six individuals followed by k-mer analysis (Fig. 1B, Supplementary Fig. 4) confirmed the diploid and triploid assignments. Resequenced aberrant individuals were either diploid or triploid, indicating that their AB data were unreliable, so all of them were excluded from subsequent analyses. All wild individuals were diploid. Among cultivated individuals, 414 were diploid and 214 were triploid. Individual inbreeding coefficients were generally highest for diploid plants and lowest for triploids (Fig. 1A).

Triploid cytotypes were significantly more common in the southerly sites than the northerly ones (Fig. 1C; multivariate regression, P-value = 10^{-14}). There was no obvious change in their proportion with altitude (Supplementary Fig. 5, P-value = 0.31). Enset is usually harvested prior to flowering, to avoid reallocation of energy from edible storage organs. We suggest that less intensive farming landscapes in the south (Borrell et al., 2020) are likely to provide more opportunities for cultivated enset to flower, leading to subsequent outcrossing and hence an increased likelihood of polyploid formation. Adaptive differences with latitude, or unidentified bioclimatic or cultural drivers may also play a role.

We identified clonal lineages among our samples using pairwise genetic distances (Fig. 1D, Supplementary Fig. 6). Within the 414 cultivated diploids there were 121 clonal lineages, of which 56 were represented by a single individual. Within the 214 triploids there were 33 clonal lineages, of which 11 were represented by a single individual. Seventy-seven (36%) of the triploids were from a single clone, suggesting preferential planting of this clone by farmers. By contrast, only 37 (9%) of diploids belonged to the most dominant diploid clone. Three diploid individuals occurring on farms appeared to be genetically wild individuals, likely being cultivated for their leaves, as wild enset is considered inedible.

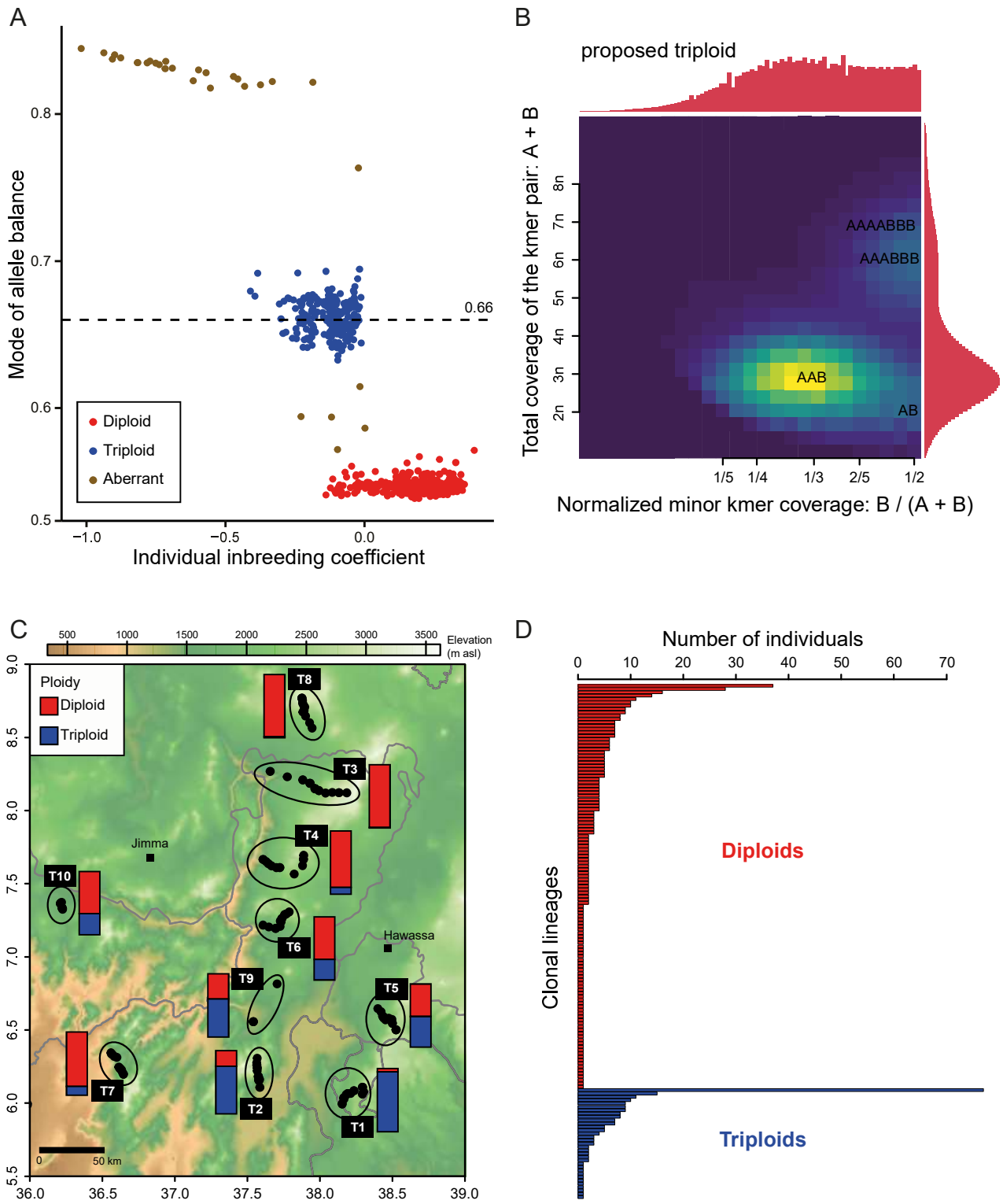

Figure 1 Ploidy variation in *Ensete ventricosum*. A: Individual inbreeding coefficient and mode of allele balance distributions for 658 cultivated and 65 wild individuals. Allele balance is defined as the number of reads supporting the most covered allele divided by the total number of reads for heterozygous genotypes. **B: K-mer profile of a triploid individual.** Heatmap colors indicate the number of heterozygous (i.e., differing by one nucleotide) 21-mer pairs in each bin, from 0 in dark purple to the maximum value in bright yellow. Histograms represent the total coverage of k-mer pairs for each axis. The brighter "smudge" indicates the ploidy of the sample, i.e. AAB for triploidy. **C: Geographic distribution of triploid individuals.** The proportion of cultivated individuals with a diploid or triploid cytotype is represented by a stacked barplot for each transect. Points on the map indicate sampling sites, and sites belonging to the same transect are encircled by an ellipse. Light

gray lines delimit regional states. **D: Distribution of cultivated individuals in clonal lineages.** Clonal lineages, represented as bars, were determined on the basis of pairwise Euclidean genetic distances, separately for diploids (in red) and triploids (in blue).

Farmers reported 177 vernacular landrace names for the sequenced cultivated plants. To some extent the names identified specific clonal lineages (Supplementary Fig. 7): 69.8% of pairwise comparisons of the same name were for the same clone. Farmers' names consistently differentiated diploids from triploids: only 2.6% of pairwise comparisons of the same name were between individuals with different ploidy (compared to a random expectation of 14%, permutation test, P-value $<10^{-4}$). Thus farmers unconsciously classify plants according to their ploidy based on phenotype alone. Farmers' names differed with geography and ethnic group. For example, the most widespread triploid clone was commonly called Ganticha in T1 and T5 (48 samples), Maze in T2 (13 samples), Wulanche or Wolanche in T6 (3 samples), Toracho in T5 (3 samples), as well as 10 other names applied to single individuals. This is unsurprising, given that at least 65 languages are spoken across the enset cultivating region.

Both diploid and triploid cultivated enset were genetically differentiated from wild enset (diploids: F_{ST} = 0.198; 95% Confidence Interval: 0.190-0.208; triploids: F_{ST} = 0.209; 95% CI: 0.200-0.216). Differentiation was low between triploid and diploid cultivated clones (F_{ST} = 0.032, 95% CI: 0.030-0.034). In a PCA of SNP data, triploids overlapped with a subset of the cultivated diploids (Fig. 2A), and clustering analysis did not group them separately (Supplementary Figure 8). Analysis of kinship among samples showed that relatedness was low between most diploid-triploid pairs of lineages, and between most triploid-triploid pairs (Fig. 2B). This suggests that triploids have arisen from among the cultivated diploids through multiple independent events. None of the sampled diploids had genotypes consistent with direct parentage of the triploids, suggesting that triploidization may have occurred elsewhere, a substantial time ago or that the diploid progenitors are rare or extinct.

Using a Bayesian linear mixed-effects model, we compared the pseudostem volume (the main edible tissue) of diploids and triploids. We used ploidy as a fixed effect with age and bioclimatic variables (summarized by the first two principal components of a PCA) as covariates to the ploidy. Membership to clonal lineages and a genetic covariance matrix between lineages were applied as random effects (see Methods). Around 13% of the variation in pseudostem volume was explained by differences between clonal lineages. We found an interaction between age, ploidy and environmental PC2 (Fig. 2C, Supplementary Table 2). While the pseudostem volume as a function of age varied for diploids along PC2, it was not the case for triploids. Overall, the estimated pseudostem volume of triploids was higher: for example, a triploid harvested at 6 years old would be 42 to 75% larger than a diploid, after controlling for other factors.

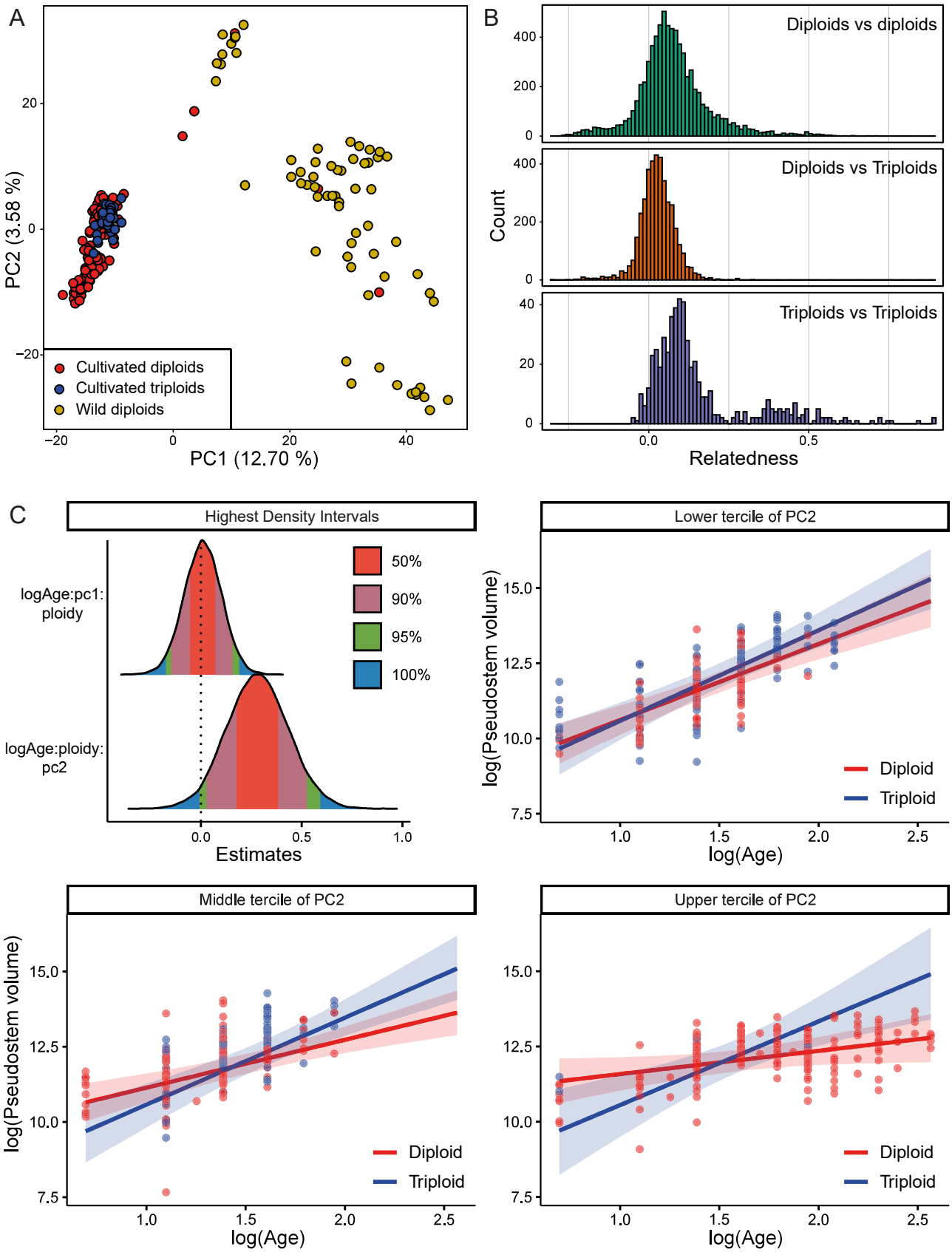

Figure 2. Origins and phenotypic variation of triploid *Ensete ventricosum*. A: Principal component analysis on the clone-corrected dataset. Diploid and triploid cultivated individuals are colored in red and blue respectively, and wild samples are in yellow. **B: Relatedness between pairs of cultivated individuals**, as measured by Huang method of moment estimator. **C: Bayesian linear mixed-effects analysis for pseudostem volume in cultivated enset.** Left panel: estimation of fixed effects of the model. Effect size for ploidy is represented relative to the diploid category. Colored

surfaces correspond to different values for the highest density interval of the full posterior distributions. Right panel: Interaction effect between plant age, ploidy, and PC2 (from a principal component analysis on environmental variables) on the plant pseudostem volume. Estimated linear relationships between the log-transformed age and log-transformed pseudostem volume are represented as lines, and their 95% credible intervals as colored areas, for diploid (in red) and triploid (in blue) individuals. Linear relationships for the three terciles of PC2 values are displayed.

For many polyploid crops, knowledge about the timing of polyploidization, and the number of parallel events involved, relies upon speculative reconstruction of events in the distant past. We show that for enset, diploids and triploids are co-occurring in the home gardens of farmers. Triploids have multiple origins and have higher pseudostem volume. Patterns of naming and clone distribution suggest that farmers can distinguish triploids from diploids. Similar processes may have occurred in the domestication of other polyploid crops (Chaïr et al., 2016; Hughes et al., 2007; Sharif et al., 2020). However, unlike polyploid crops with even numbers of chromosome sets, triploids typically have little to no fertility, so their long term success relies on clonal propagation and recurrent origins. The advanced domestication of bananas, in the sister genus to enset, relies largely on triploid lineages, of which a single clone is dominant and now threatened by disease (Drenth & Kema, 2021). Clonal lineages of enset are also vulnerable to accumulation of somatic mutations (White et al., 2023). We suggest that farm productivity and hence food security could be enhanced by the increased use of triploid enset, with the potential of breeding novel triploids, and exploration of its adaptive range. However, care should be taken to preserve the genetic diversity of domesticated diploid populations as a safeguard for the future.

METHODS

Enset reference genome development

DNA & RNA extraction and library preparation

The individual used for the genome assembly was grown from a seed collected in Wolaita (6.9°N, 37.8°E, 2,120m a.s.l.). This seed originated from a sexual cross between a plant with the vernacular name *Maze* and an unknown parent. The latter was probably a cultivated individual, since, to our knowledge, wild plants do not occur in this region. DNA was extracted using a CTAB protocol, then quantified with a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). The PacBio SMRTbell library (~20kb) for PacBio Sequel was constructed using SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park, CA, USA) using the manufacturer recommended protocol. The library was bound to polymerase using the Sequel II Binding Kit 2.0 (PacBio). Sequencing was performed on PacBio Sequel II 8M SMRT cells, generating 21.6 Gb of data. For the Dovetail Omni-C library, chromatin was fixed in place with formaldehyde in the nucleus and extracted. Fixed chromatin was digested with DNAse I, chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity ligation, crosslinks were reversed and the DNA purified. Purified DNA was treated to remove biotin that was not internal to ligated fragments. Sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments

were isolated using streptavidin beads before PCR enrichment of each library. The library was sequenced on an Illumina HiSeqX platform to produce around 30x sequence coverage.

RNA was extracted from four different tissues: leaf, midrib, pseudostem and roots. Total RNA extraction was done using the QIAGEN RNeasy Plus Kit following manufacturer protocols. Total RNA was quantified using Qubit RNA Assay and TapeStation 4200. Prior to library preparation, DNase treatment followed by AMPure bead clean up and QIAGEN FastSelect HMR rRNA depletion were performed. Libraries were prepared with the NEBNext Ultra II RNA Library Prep Kit following manufacturer protocols, and libraries were run on the Illumina NovaSeq6000 platform in 2 x 150 bp configuration.

Genome assembly

PacBio CCS reads were assembled with Hifiasm 0.15.4-r347 (Cheng et al., 2021) with default parameters. BLAST results of the Hifiasm assembly against the nt database were used as input for blobtools 1.1.1 and scaffolds identified as possible contamination were removed from the assembly. Finally, purge_dups 1.2.5 was used to remove haplotigs and contig overlaps. Scaffolding of the HiFi assembly was carried out with HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (Putnam et al., 2016). Omni-C reads were aligned to the draft input assembly using bwa (Li & Durbin, 2009). The separations of Omni-C read pairs mapped within draft scaffolds (mapping quality > 50) were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, score prospective joins, and make joins above a threshold.

A chloroplast sequence was assembled with NOVOPlasty 4.2 (Dierckxsens et al., 2017) using short reads from another sequenced landrace published previously (GenBank BioProject PRJNA252658) (Yemataw et al., 2018). This sequence was used as a query to search for chloroplastic scaffolds in our assembly with BLAST 2.10.1+ (Camacho et al., 2009), keeping only hits with a length > 500 bp. We then computed the proportion of each scaffold covered by chloroplastic sequences using the coverage command in BEDTools 2.29.2 (Quinlan & Hall, 2010). Scaffolds with > 80% BLAST hit coverage were discarded from the assembly (500 scaffolds for a total of 19.6 Mb). We also tried to assemble a mitochondrial genome using the same method, but without success. Consequently, in order to detect and discard mitochondrial sequences, we used mitochondrial gene sequences from *E. glaucum* (Wu et al., 2021) as queries for BLAST searches, keeping only genes with a minimum query coverage of 80%. Scaffolds with a valid hit (with the exception of the nine largest scaffolds) were considered as potentially mitochondrial and discarded (80 scaffolds for a total of 5.4 Mb).

Repeat and gene annotation

A *de novo* library of repeated sequences was generated using RepeatModeler 2.0.3 (Flynn et al., 2020) with LTR structural analysis enabled. The genome sequence was then annotated and softmasked using this library with RepeatMasker 4.1.2 (Smit et al., 2015). Both RepeatModeler and RepeatMasker were used through the Dfam TE Tools container v1.5 (https://github.com/Dfam-consortium/TETools). We searched for canonical telomeric repeats (CCCTAAA / TTTAGGG) in 1 Mb windows along the genome sequence using the

nuc command in BEDTools. To detect potential centromeric regions, we looked for large tandem repeat arrays in the nine largest scaffolds using repaver (available at https://gitlab.com/gringer/bioinfscripts). Centromeres in *E. glaucum* contain large arrays of the Egcen repeat and are enriched for an interspersed nuclear element (LINE) called Nanica (Wang et al., 2022), discovered initially in the centromeres of *Musa acuminata* (D'Hont et al., 2012). To confirm our initial annotation of centromeres, the sequences of Egcen and Nanica were used as queries for a BLAST search in our assembly, and their number per 1 Mb computed with BEDTools coverage. The results of these analyses were plotted using RIdeogram (Hao et al., 2020) in R.

RNA-seq reads were trimmed using Trimmomatic 0.36 (Bolger et al., 2014) (parameters: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:5:True LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36) then mapped onto the new reference sequence with STAR 2.7.9a (Dobin et al., 2013), keeping only properly paired reads with SAMtools 1.9 (Li et al., 2009). For protein evidence, we downloaded all Embryophyte protein sequences from the OrthoDB v10 database (Kriventseva et al., 2019), and protein sequences from *M. balbisiana*, *M. itinerans*, *M. schizocarpa* and *E. glaucum* from the Banana Genome Hub website (Droc et al., 2022) (https://banana-genome-hub.southgreen.fr). RNA-seq and protein evidence were both used to annotate protein-coding genes with BRAKER 2.1.6 (Brůna et al., 2021; Hoff et al., 2016) on the softmasked genome sequence.

Quality control

Quality control was carried out through comparison of the k-mer content between the PacBio CCS reads and the final genome assembly, with KAT 2.4.2 (Mapleson et al., 2017) using a k-mer length of 21. The k-mer spectrum for the PacBio reads was also analyzed using GenomeScope 2.0 (Ranallo-Benavidez et al., 2020). The level of completeness of the genome assembly and the gene annotation was assessed using BUSCO 5.3.0 (Manni et al., 2021) with the embryophyta_odb10 database.

Sampling of enset in southwestern Ethiopia

Enset individuals were systematically sampled across eight transects (T1 to T8), 20-50 km long, across the enset growing region (Supplementary Fig. 3). These transects were oriented along environmental gradients, generally encompassing >1,000 m elevational range and therefore capturing a broad range of local climates including temperature and precipitation variation. This approach sought to capture a substantial portion of genotypic diversity and locally adapted landraces, while minimizing geographic distance within transects and associated isolation by distance effects. For each transect we sampled 10 locations, collecting tissue from all landraces present across three farms. We generally sampled a 10 x 10 cm piece of cigar leaf, and immediately dried it on silica gel. We interviewed farmers to obtain the variety name and the age of collected individuals. Additionally, for each plant we measured the pseudostem (height, basal and apical circumferences), one of the principal harvested tissues for consumption. We also used a few accessions from two other different areas (T9 et T10), collected opportunistically. Cultivated accessions were supplemented with wild individuals. These predominantly occur to the west and north of the main cultivated enset distribution with a patchy occurrence. We sought to

sample as comprehensively as possible across this range. Environmental covariates for sampling locations were sourced at a 1-km resolution from Chelsa v2.1 (Karger et al., 2017a, 2017b) and sampled using the Raster package in R software.

Reduced-representation sequencing for population genomics

After freezing leaf samples in liquid nitrogen, DNA was extracted using the DNeasy Plant Mini Kit or the DNeasy 96 Plant Kit (Qiagen) and eluted in 70 µl of AE buffer. Quantification of genomic DNA was performed using a Quantus fluorometer (Promega). Samples were sent to Eurofin Genomics, where they were sequenced with Genotyping by Random Amplicon Sequencing-Direct (GRAS-Di), a reduced representation sequencing method using random primers (Miki et al., 2020).

Adapters were trimmed from reads using Trimmomatic, and read pairs with at least one read < 36 bp were discarded. Reads were then mapped onto our new genome assembly with bwa mem 0.7.17, keeping only properly paired reads with SAMtools. Small variants were called using freebayes 1.3.1 (Garrison & Marth, 2012) (parameters: –min-mapping-quality 10 –min-base-quality 15 –no-population-priors), keeping only biallelic SNPs with a quality score > 20, less than 25% missing data and a minor allele count of 3. Low quality samples with > 45% missing data were discarded. We then discarded sites with a mean read coverage > 500 and a F_{IS} value > 0.8. Finally, we filtered out spurious SNPs potentially caused by copy number variation by detecting sites with a median allele balance (number of reads supporting the alternate allele divided by the total number of reads, for heterozygous genotypes) < 0.35 or > 0.65 and with no homozygous genotypes and sites with a proportion of heterozygous genotypes > 60%. These SNPs and any site closer than 1,000 bp from them were discarded. The final dataset included 20,173 SNPs.

Identification of triploids

Individual inbreeding coefficients were computed with VCFtools 0.1.17 (Danecek et al., 2011). The distribution of allele balance (defined here as the number of reads supporting the most covered allele divided by the total number of reads, considering only heterozygous genotypes) for each individual was determined with vcfR (Knaus & Grünwald, 2017) for sites with a read coverage > 20, and the mode of these distributions was computed with modeest (Poncet, 2019) in R. We expected diploid accessions to have a mode around 0.5, since heterozygous sites should show a 1:1 allele coverage ratio. Values of 0.66 (1:2 ratio) would be associated with triploidy, and so on. Plotting of individual inbreeding coefficients against the allele balance modes for all samples revealed the existence of three groups (see Fig. 1A in the Results section). After visual inspection of the raw allele balance distributions, we classified individuals as diploid, triploid or aberrant.

To confirm the ploidy results found with the GRAS-Di data, we randomly chose one diploid, three triploid and two aberrant samples and resequenced their whole genome on an Illumina NovaSeq sequencer (2 x 150 bp paired-end reads). After adapter trimming with Trimmomatic, the k-mer spectrum for each individual was determined with Jellyfish 2.2.6 (Marçais & Kingsford, 2011), and their ploidy was further explored with Smudgeplot 0.2.5 (Ranallo-Benavidez et al., 2020), using a k-mer size of 21. Additionally, reads were mapped onto the reference sequence as described above, and SNPs discovered with the GRAS-di data

were genotyped again with freebayes to check the allele balance distributions for the resequencing data. Due to the lack of availability of a flow cytometer in Ethiopia, and restrictions on export of living material, we were unable to estimate ploidy level from genome size measurements by flow cytometry using fresh plant material. While we tested this approach by rehydrating silica dried samples, as used successfully for some plant materials (Zarrei et al., 2012), we were unsuccessful, most likely because samples were too old (2-3 years).

To ascertain whether ploidy level was correlated with vernacular landrace names, we calculated the proportion of pairs of individuals sharing the same name that also shared the sample ploidy: P_p .The observed value of P_p was compared with the distribution generated by randomly relocating the landrace names to individuals (10,000 iterations).

Population genetic analysis of mixed ploidy dataset

For analyses of datasets with mixed ploidy, we followed the recommendations of (Meirmans et al., 2018). First, SNP calling was carried out again for triploid individuals with freebayes, only using sites detected during the previous SNP calling and a ploidy of three. Clonal lineages were determined separately for diploid and triploid as follows: after filtering out sites with a minor allele frequency (MAF) inferior to 5%, more than 50% missing data, and selecting a random SNP for each genomic window of 100 kb, individuals belonging to the same clonal lineage were grouped together with the mlg.filter function in the R package poppr 2.9.3 (Kamvar et al., 2014), on the basis of pairwise Euclidean distances with the farthest neighbor method. The cut-off values were chosen visually as suggested by (Arnaud-Haond et al., 2007). The number of clonal lineages was compared between diploid and triploid individuals after correcting sample size by rarefaction, using the R package vegan 2.6-4 (Oksanen et al., 2022). To check whether the proportion of triploid individuals in each site was correlated with latitude, longitude and elevation, we fitted a linear regression model using these three variables in R.

We investigated the population genetic structure of cultivated and wild enset and the origin of triploid individuals using multiple approaches. We obtained a clone-corrected dataset by keeping the sample with the least missing data in each clonal lineage, with the exception of three cultivated individuals which were grouped with wild samples and retained for further analyses. Sites with a MAF < 5% were discarded, and, for each genomic window of 100 kb we kept the site with the higher MAF. Relatedness between clonal lineages was assessed using a method of moment estimator designed for mixed ploidy data (Huang et al., 2014, 2015) and implemented in PolyRelatedness v1.11b. Population genetic structure in the clone-corrected dataset was assessed with principal component analysis (PCA) based on individual allele frequencies. F_{ST} values between our three groups (cultivated diploids, cultivated triploids, and wild individuals) were computed with the R package StAMPP 1.6.3 (Pembleton et al., 2013), and their 95% confidence intervals (CI) were determined with 1,000 bootstraps across loci.

Genotype likelihoods for each SNP were computed with updog 2.1.2 (Gerard et al., 2018; Gerard & Ferrão, 2020) in R, from the read counts in freebayes outputs and with default parameters. Genotype likelihoods were then used to investigate population structure for the entire clone-corrected dataset with entropy (Shastry et al., 2021), a Structure-like

algorithm developed for mixed ploidy data. The software was run with a number of populations K ranging from 2 to 15, with five MCMC chains (5,000 burn-in steps, 10,000 MCMC steps, thinning of 10) for each K value to assess convergence. The fit of the data to each value of K was calculated using the Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010).

Analysis of phenotypic data for diploid and triploid cultivated individuals

To assess the potential effect of ploidy on enset phenotype, here the pseudostem volume, we used a Bayesian linear mixed-effects approach implemented in the R package brms 2.19.0 (Bürkner, 2017; Carpenter et al., 2017). First, the pseudostem volume for each individual was calculated according to the shape of a truncated cone, using the measured pseudostem height and basal and apical circumferences. To reduce dimensionality and avoid correlated variables, the environmental data was transformed using PCA. Plant age, ploidy and the two first principal components of the PCA were used as fixed effects, as well as interactions between ploidy, age, and each of the two principal components. The response variable (pseudostem volume) and age values were log-transformed. Membership to clonal lineages was set as a random effect, with the covariance between lineages computed from pairwise relatedness coefficients, to take into account correlation between clonal lineages. Individuals with missing data for any variable and four one-year-old individuals were discarded, leaving us with 555 samples. We fitted the model with four chains of 10,000 iterations, including 2,500 warm-up iterations, using default priors. Convergence was assessed visually by plotting the traces of the chains, and by checking the R-hat values and the effective sample sizes (Bulk-ESS and Tail-ESS). Model fit was evaluated using posterior predictive checks.

ACKNOWLEDGMENTS

We thank the numerous farmers, agricultural extension agents and drivers who generously facilitated our fieldwork. We also thank Romain Pigeault for his help with the mixed-effects model analysis. This research was supported by the Global Challenges Research Fund Agrisystems awards entitled ‘Landscape scale genomic-environment data to enhance the food security of Ethiopian agri-systems’ (Grant No. BB/S014896/1) and ‘Modelling and genomics resources to enhance exploitation of the sustainable and diverse Ethiopian starch crop enset and support livelihoods’ (Grant No. BB/P02307X/1). This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT (http://doi.org/10.5281/zenodo.438045).

DATA AVAILABILITY

The enset genome assembly and its associated data have been deposited in GenBank BioProject PRJNA899044, and is also available on the Banana Genome Hub website (https://banana-genome-hub.southgreen.fr/). Raw reads from resequencing individuals have been deposited in GenBank BioProject PRJNAXXXXXX, and other data have been deposited in Zenodo (DOI: 10.5281/zenodo.14866251). Scripts used for analysis are available from https://github.com/YannDussert/enset_triploidy_analyses.

REFERENCES

Akagi, T., Jung, K., Masuda, K., & Shimizu, K. K. (2022). Polyploidy before and after domestication of crop species. *Current Opinion in Plant Biology*, *69*, 102255. https://doi.org/10.1016/j.pbi.2022.102255

Arnaud-Haond, S., Duarte, C. M., Alberto, F., & Serrão, E. A. (2007). Standardizing methods to address clonality in population studies. *Molecular Ecology*, *16*(24), 5115–5139. https://doi.org/10.1111/j.1365-294X.2007.03535.x

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borrell, J. S., Biswas, M. K., Goodwin, M., Blomme, G., Schwarzacher, T., Heslop-Harrison, J. S. (Pat), Wendawek, A. M., Berhanu, A., Kallow, S., Janssens, S., Molla, E. L., Davis, A. P., Woldeyes, F., Willis, K., Demissew, S., & Wilkin, P. (2019). Enset in Ethiopia: A poorly characterized but resilient starch staple. *Annals of Botany*, *123*(5), 747–766. https://doi.org/10.1093/aob/mcy214

Borrell, J. S., Goodwin, M., Blomme, G., Jacobsen, K., Wendawek, A. M., Gashu, D., Lulekal, E., Asfaw, Z., Demissew, S., & Wilkin, P. (2020). Enset-based agricultural systems in Ethiopia: A systematic review of production trends, agronomy, processing and the wider food security applications of a neglected banana relative. *Plants, People, Planet*, *2*(3), 212–228. https://doi.org/10.1002/ppp3.10084

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, *3*(1), lqaa108. https://doi.org/10.1093/nargab/lqaa108

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421–421. https://doi.org/10.1186/1471-2105-10-421

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Chaïr, H., Traore, R. E., Duval, M. F., Rivallan, R., Mukherjee, A., Aboagye, L. M., Rensburg, W. J. V., Andrianavalona, V., Carvalho, M. A. A. P. de, Saborio, F., Prana, M. S., Komolong, B., Lawac, F., & Lebot, V. (2016). Genetic diversification and dispersal of taro (*Colocasia esculenta* (L.) Schott). *PLOS ONE*, *11*(6), e0157712. https://doi.org/10.1371/journal.pone.0157712

Cheesman, E. E. (1947). Classification of the bananas: The genus *Ensete* Horan. *Kew Bulletin*, *2*(2), 97–106. https://doi.org/10.2307/4109206

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), Article 2. https://doi.org/10.1038/s41592-020-01056-5

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., Da Silva, C., Jabbari, K., Cardi, C., Poulain, J., Souquet, M., Labadie, K., Jourda, C., Lengellé, J., Rodier-Goud, M., … Wincker, P. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*, *488*(7410), 213–217. https://doi.org/10.1038/nature11241

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, *45*(4), e18.

https://doi.org/10.1093/nar/gkw955
Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635
Drenth, A., & Kema, G. (2021). The vulnerability of bananas to globally emerging disease threats. *Phytopathology*, *111*(12), 2146–2161. https://doi.org/10.1094/PHYTO-07-20-0311-RVW
Droc, G., Martin, G., Guignon, V., Summo, M., Sempéré, G., Durant, E., Soriano, A., Baurens, F.-C., Cenci, A., Breton, C., Shah, T., Aury, J.-M., Ge, X.-J., Harrison, P. H., Yahiaoui, N., D'Hont, A., & Rouard, M. (2022). The banana genome hub: A community database for genomics in the Musaceae. *Horticulture Research*, *9*, uhac221. https://doi.org/10.1093/hr/uhac221
Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117
Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*, Article arXiv:1207.3907. https://doi.org/10.48550/arXiv.1207.3907
Gerard, D., & Ferrão, L. F. V. (2020). Priors for genotyping polyploids. *Bioinformatics*, *36*(6), 1795–1800. https://doi.org/10.1093/bioinformatics/btz852
Gerard, D., Ferrão, L. F. V., Garcia, A. A. F., & Stephens, M. (2018). Genotyping polyploids from messy sequencing data. *Genetics*, *210*(3), 789–807. https://doi.org/10.1534/genetics.118.301468
Hao, Z., Lv, D., Ge, Y., Shi, J., Weijers, D., Yu, G., & Chen, J. (2020). RIdeogram: Drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Computer Science*, *6*, e251. https://doi.org/10.7717/peerj-cs.251
Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, *32*(5), 767–769. https://doi.org/10.1093/bioinformatics/btv661
Huang, K., Ritland, K., Guo, S., Dunn, D. W., Chen, D., Ren, Y., Qi, X., Zhang, P., He, G., & Li, B. (2015). Estimating pairwise relatedness between individuals with different levels of ploidy. *Molecular Ecology Resources*, *15*(4), 772–784. https://doi.org/10.1111/1755-0998.12351
Huang, K., Ritland, K., Guo, S., Shattuck, M., & Li, B. (2014). A pairwise relatedness estimator for polyploids. *Molecular Ecology Resources*, *14*(4), 734–744. https://doi.org/10.1111/1755-0998.12217
Hughes, C. E., Govindarajulu, R., Robertson, A., Filer, D. L., Harris, S. A., & Bailey, C. D. (2007). Serendipitous backyard hybridization and the origin of crops. *Proceedings of the National Academy of Sciences*, *104*(36), 14389–14394. https://doi.org/10.1073/pnas.0702193104
Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, *2014*(1), 1–14. https://doi.org/10.7717/peerj.281
Karger, D., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017a). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, *4*(1), Article 1. https://doi.org/10.1038/sdata.2017.122
Karger, D., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R., Zimmermann, N., Linder, H., & Kessler, M. (2017b). Data from: Climatologies at high resolution for the earth's land surface areas. *Dryad Digital Repository*. http://dx.doi.org/doi:10.5061/dryad.kd1d4
Knaus, B. J., & Grünwald, N. J. (2017). vcfr: A package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, *17*(1), 44–53. https://doi.org/10.1111/1755-0998.12549

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, *47*(D1), D807–D811. https://doi.org/10.1093/nar/gky1053

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., & Clavijo, B. J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, *33*(4), 574–576. https://doi.org/10.1093/bioinformatics/btw663

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

Meirmans, P. G., Liu, S., & van Tienderen, P. H. (2018). The analysis of polyploid genetic data. *Journal of Heredity*, *109*(3), 283–296. https://doi.org/10.1093/jhered/esy006

Miki, Y., Yoshida, K., Enoki, H., Komura, S., Suzuki, K., Inamori, M., Nishijima, R., & Takumi, S. (2020). GRAS-Di system facilitates high-density genetic map construction and QTL identification in recombinant inbred lines of the wheat progenitor *Aegilops tauschii*. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-78589-4

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., Caceres, M. D., Durand, S., … Weedon, J. (2022). *vegan: Community ecology package* (Version 2.6-4) [Computer software]. https://cran.r-project.org/web/packages/vegan/index.html

Pembleton, L. W., Cogan, N. O. I., & Forster, J. W. (2013). StAMPP: An R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources*, *13*(5), 946–952. https://doi.org/10.1111/1755-0998.12129

Poncet, P. (2019). *modeest: Mode estimation* (Version 2.4.0) [Computer software]. https://cran.r-project.org/web/packages/modeest/index.html

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., Haussler, D., Rokhsar, D. S., & Green, R. E. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research*, *26*(3), 342–350. https://doi.org/10.1101/gr.193474.115

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1432. https://doi.org/10.1038/s41467-020-14998-3

Salman-Minkov, A., Sabath, N., & Mayrose, I. (2016). Whole-genome duplication as a key factor in crop domestication. *Nature Plants*, *2*(8), 1–4. https://doi.org/10.1038/nplants.2016.115

Sharif, B. M., Burgarella, C., Cormier, F., Mournet, P., Causse, S., Van, K. N., Kaoh, J.,

Rajaonah, M. T., Lakshan, S. R., Waki, J., Bhattacharjee, R., Badara, G., Pachakkil, B., Arnau, G., & Chaïr, H. (2020). Genome-wide genotyping elucidates the geographical diversification and dispersal of the polyploid and clonally propagated yam (*Dioscorea alata*). *Annals of Botany*, *126*(6), 1029–1038. https://doi.org/10.1093/aob/mcaa122

Shastry, V., Adams, P. E., Lindtke, D., Mandeville, E. G., Parchman, T. L., Gompert, Z., & Buerkle, C. A. (2021). Model-based genotype and ancestry estimation for potential hybrids with mixed-ploidy. *Molecular Ecology Resources*, *21*(5), 1434–1451. https://doi.org/10.1111/1755-0998.13330

Smit, A. F. A., Hubley, R., & Green, P. (2015). *RepeatMasker Open-4.0. 2013–2015*.

Wang, Z., Rouard, M., Biswas, M. K., Droc, G., Cui, D., Roux, N., Baurens, F.-C., Ge, X.-J., Schwarzacher, T., Heslop-Harrison, P. (J S., & Liu, Q. (2022). A chromosome-level reference genome of *Ensete glaucum* gives insight into diversity and chromosomal and repetitive sequence evolution in the Musaceae. *GigaScience*, *11*, giac027. https://doi.org/10.1093/gigascience/giac027

Watanabe, S. (2010). Asymptotic aquivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(116), 3571–3594.

White, O. W., Biswas, M. K., Abebe, W. M., Dussert, Y., Kebede, F., Nichols, R. A., Buggs, R. J. A., Demissew, S., Woldeyes, F., Papadopulos, A. S. T., Schwarzacher, T., Heslop-Harrison, P. J. S., Wilkin, P., & Borrell, J. S. (2023). Maintenance and expansion of genetic and trait variation following domestication in a clonal crop. *Molecular Ecology*, *32*(15), 4165–4180. https://doi.org/10.1111/mec.17033

Wu, C.-S., Sudianto, E., Chiu, H.-L., Chao, C.-P., & Chaw, S.-M. (2021). Reassessing banana phylogeny and organelle inheritance modes using genome skimming data. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.713216

Yemataw, Z., Muzemil, S., Ambachew, D., Tripathi, L., Tesfaye, K., Chala, A., Farbos, A., O'Neill, P., Moore, K., Grant, M., & Studholme, D. J. (2018). Genome sequence data from 17 accessions of *Ensete ventricosum*, a staple food crop for millions in Ethiopia. *Data in Brief*, *18*, 285–293. https://doi.org/10.1016/j.dib.2018.03.026

Zarrei, M., Wilkin, P., Ingrouille, M. J., Leitch, I. J., Buerki, S., Fay, M. F., & Chase, M. W. (2012). Speciation and evolution in the *Gagea reticulata* species complex (Tulipeae; Liliaceae). *Molecular Phylogenetics and Evolution*, *62*(2), 624–639. https://doi.org/10.1016/j.ympev.2011.11.003