

What makes a banana false? How the genome of Ethiopian orphan staple *Ensete ventricosum* differs from the banana A and B sub-genomes

Sadik Muzemil

Sadik.Muzemil.1@warwick.ac.uk

University of Warwick

Parameswari Paul

University of Warwick

Laura Baxter

University of Warwick

Ana Dominguez-Ferreras

University of Warwick

Sunil Kumar Sahu

BGI Research

Allen Van Deynze

University of California, Davis

Guoqiang Mai

BGI Research

Zerihun Yemataw

Sidama Agricultural Research Institute

Kassahun Tesfaye

Bio and Emerging Technology Institute

Vardis Ntoukakis

University of Warwick

David J. Studholme

University of Exeter

Murray Grant

University of Warwick

Research Article

Keywords: *Ensete ventricosum*, *Musa*, gene families, nucleotide binding site leucine rich repeat (NLR), orphan crop

Posted Date: July 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10247519/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Ensete ventricosum, also known as the “tree against hunger” plays a key role in Ethiopian food security and farming systems, feeding more than 20 million people. Since domestication via clonal selection in the south-west Ethiopian highlands, today’s diverse enset landraces contribute multiple benefits including food, fibre by-product, animal bedding and cattle fodder to farmers and local communities. Improved genomic resources for this highly drought-tolerant plant are essential to supplement the conventional clonal selection-based breeding programme and pave the way towards targeted breeding.

Results

We sequenced the genome of enset landrace Mazia, which is partially resistant/tolerant to *Xanthomonas* wilt and predicted 38,940 protein-coding genes. The Mazia assembly (540.14 Mb) is more complete than the previously published genome assembly of landrace Bedadeti (451.28 Mb) and displayed 1.41% heterozygosity and 64.64% repetitive DNA content. Comparative analyses with the Bedadeti assembly and chromosome-level genome sequences of the two main banana progenitors (*Musa acuminata*, AA genome; *Musa balbisiana*, BB genome) unexpectedly revealed ~ 25% of the Mazia genome is unique to enset. Gene Ontology (GO) and sequence similarity search analysis of enset-specific protein-coding genes identified distinct functional signatures that underpin the lifestyle, adaptation, and corm productive quality of enset, including functions related to DNA integration, carbohydrate metabolism, disease resistance and transcriptional regulation. In contrast, *Musa*-specific genes showed enrichment for defence response, protein phosphorylation and fruit development pathways. Focusing on the classical nucleotide binding site leucine rich repeat (NLR) disease resistance genes, we identified and characterised NLRs in enset and *Musa* species genomes, revealing a considerable expansion in the *Musa acuminata* genome. We also identified unique genes in enset and banana genomes whose functional and evolutionary roles are yet to be determined.

Conclusions

Here, we report a *de novo* genome assembly for the enset (*Ensete ventricosum*) landrace Mazia and provide a high-quality annotation of both Mazia and the previously published assembly of the landrace Bedadeti. Collectively, these genomic resources provide a valuable foundation for comparative genomics within the *Musaceae* family and open new opportunities for the development of marker-assisted breeding strategies to accelerate the improvement of agronomically important traits in enset.

1. Background

The farmers' adage, "Enset is our food, our clothes, our beds, our houses, our cattle-feed, our plates" [1], captures the multifaceted importance enset (*Ensete ventricosum*), an orphan crop that serves as a staple food source for over 20 million people [2] in Africa's second most populous country. With sub-Saharan Africa's population projected to surpass 1.9 billion by 2050 [3], understanding and improving underutilised crops like enset is critical for regional food security.

Enset, also known as false banana or Abyssinian banana, is the world's largest herbaceous perennial monocarpic plant, reaching heights up to 11 m. A member of the Musaceae (banana family), enset is estimated to have diverged from *Musa* species between 51.9 [4] to 59.9 [5] million years ago, and represents one of the most socioeconomically and culturally important species. Unlike banana, which is grown for its fruit, domesticated enset is cultivated for its large starch-rich underground corm and overlapping leaf sheaths (pseudostem) (Fig. 1). These carbohydrate-rich components provide a staple food with high nutritional value: daily consumption of 0.5 kg of enset-derived products supplies approximately 20% of protein and 70% of energy requirements [6].

Enset is believed to have been domesticated approximately 10,000 years ago in Ethiopia, alongside coffee, various pulses and tef [1]. Often called "the tree against hunger", highly drought tolerant enset holds historical importance as a year-round source of food for South and South-Western Ethiopia during recurrent famine and drought periods [1]. Enset-based farming systems outperform any other Ethiopian cropping systems in energy yield per area per year [7], and its soil-enhancing and shade-providing qualities make it well-suited for intercropping with root or cash crops like coffee. Cultivation spans diverse agro-ecologies, between 1200 and 3100 m altitudes (up to 3450 m, Muzemil, S. pers. observation) [1], with smallholders growing up to 28 landraces to exploit specific traits including drought tolerance, disease resistance and preferred culinary properties [8].

The extended life cycle of enset presents both advantages and challenges for improvement. Key corm and pseudostem traits are not evident until 2–7 years from seedling establishment, with flowering taking 8–12 years and often resulting in hard, sterile seeds. This is about 2–5 times longer than banana to reach harvest. In mature enset, the pseudostem can reach 0.5–2 m circumference, with up to 17 leaves, each up to 5 m long which nourish an underground corm that can weigh up to 88 kg [9]. These traits likely drove enset's utilization by prehistoric hunter-gatherer societies [10] and its subsequent domestication. Today enset is a cornerstone of Ethiopia's economic, social, and cultural fabric [11]. Enset is harvested prior to, or immediately after, flower emergence and before flower maturation while starch reserves are maximal. Because vegetative propagation has been universally adopted, elite landraces are currently identified through time-consuming phenotypic characterization of mature, clonally propagated domesticated landraces over multiple years [9]. Thus, the capacity for beneficial trait selection would markedly benefit from genome-assisted breeding [12].

Enset productivity is threatened by multiple pests and diseases [13–16], with enset *Xanthomonas* wilt (EXW) posing a significant threat to food security [17, 18]. EXW is caused by the bacterium *Xanthomonas vasicola* pv. *musacearum* (Xvm) [19], which also causes banana *Xanthomonas* wilt (BXW) across east

and central Africa [20, 21]. Known resistance to Xvm is limited to wild banana (*M. balbisiana*) [22], with only a few genotypes showing some level of tolerance in edible banana germplasm. Importantly, certain enset landraces show reduced susceptibility to Xvm [23, 24], making them popular among enset farmers and thus comprise a considerable proportion of farmyard landraces [25]. These landraces represent potential genetic resource for enset improvement and BXW resistance breeding in banana. Plant disease resistance is often mediated by nucleotide-binding leucine-rich repeat (NLR) genes [26, 27], yet the NLR repertoire of enset and its relationship to banana NLR families has not yet been assessed.

Figure 1.

Previous studies have revealed that strong genotype × environment interactions for major agronomic traits [28, 29] have driven regional differentiation of enset landraces [9]. This has resulted in complex vernacular naming systems across ethno-linguistic communities that confound identification of landraces for breeding. Indeed, genome sequencing has revealed significant genetic diversity amongst landraces [30], reflecting local selection and environmental adaptation.

We previously published the first draft genome sequence of *E. ventricosum* [31] and subsequently sequenced 17 Ethiopian genotypes [30], providing initial genomic resources. However, these short-read assemblies had limited contiguity, constraining comparative analyses. The availability of high-quality chromosome-level genome assemblies of closely related species, notably the major diploid progenitors of *Musa* species i.e., *M. acuminata* [32] and *M. balbisiana* [33] or their triploid cultivars [34], provided the opportunity to undertake comparative genomic analyses to identify fundamental differences between these phenotypically similar but agriculturally distinct species. Despite their shared ancestry, the extent of genomic divergence between enset and banana at the gene level, and the biological processes shaped by species-specific gene evolution, remain largely unknown.

Here, we (1) sequenced and annotated the genome and transcriptome of the EXW-resistant/tolerant landrace Mazia (Fig. 1C), (2) performed comparative genomic analysis with the main progenitors of diploid banana: *M. acuminata* (MA) [32] and *M. balbisiana* (MB) [33], and (3) characterised the nucleotide-binding leucine-rich repeat (NLR) disease resistance gene repertoire of enset and banana to seek insight into potential tools that may mitigate the major impact of EXW/BXW on these clonally propagated crops. We utilised sequencing data from 20 enset genotypes [30] and 48 banana accessions [32–35] to assess presence-absence variations and identify genes or gene families unique to enset, providing new insight into their evolutionary roles.

2. Materials and Methods

2.1 Sample collection

For genomic DNA extraction ~5g of fresh unexpanded leaf material of 2 year old *E. ventricosum* (EV) landrace Mazia (Fig. 1C) was obtained from the national enset *ex-situ* germplasm at Areka Agricultural Research Center (Areka, Ethiopia) and immediately frozen in liquid nitrogen.

For RNA sequencing (RNA-seq), harvested Mazia leaf samples were immediately subjected to stress treatments (cold, heat or wounding) or left untreated before collecting 1-2g of tissue (cut into small pieces) per treatment in RNA/ater™ (Invitrogen™) according to the manufacturer's instructions. Various other samples, e.g., unexpanded and fully expanded leaves or roots were also cut in small pieces and preserved in RNA/ater™.

2.2 Nucleic acid extraction and sequencing

Genomic DNA was extracted using a modified CTAB protocol [36] and used to construct single-tube long fragment read (stLFR) libraries and generate paired-end 100 or 142-base sequence reads on a BGISEQ-500 platform.

RNA extraction was performed according to Logemann et al. [37]. RNA concentration and integrity were assessed using a 2100 bioanalyzer (Agilent) prior to library construction and Illumina sequencing. Briefly, samples from 2.1 were removed from RNA/ater™, quickly blotted dry and immediately frozen in liquid nitrogen. RNA was isolated by guanidine hydrochloride (GuHCl) extraction [37]. Samples ground to a fine powder in a liquid nitrogen were rapidly mixed with GuHCl solution (GuHCl 8 M, MES 20 mM, EDTA 20 mM) and this mixture was extracted with phenol:chloroform (1:1). The resultant aqueous phase was mixed with 0.7 vol. ethanol and 0.05 vol. acetic acid to precipitate the RNA. Following centrifugation, the RNA pellet was washed with 70% ethanol, air-dried and dissolved in sterile water. RNA samples were assessed using a plant RNA Nano assay (Agilent, Bioanalyzer 2100) to determine RNA concentration and integrity and those with a RIN > 6.5 were sent to BGI (BGI Health (HK) Company Limited) for library construction and RNA sequencing.

2.3 Enset genome assembly and annotation

2.3.1 Assessment of genome size, heterozygosity and ploidy

A k -mer ($k = 21$) based analysis was used to estimate the size and heterozygosity of the enset genome. Illumina 2x100 paired-end reads were first trimmed and filtered in TrimGalore (v.0.6.7) (<https://github.com/FelixKrueger/TrimGalore>) parameter –quality set to 30 and stLFR reads were quality filtered using stLFRdenovo (v1.0.5) (<https://github.com/BGI-biotoools/stLFRdenovo>). The frequencies of 21-mer sequences were calculated with Jellyfish (v2.3.0) [38] from the cleaned reads and used for estimating the genome size and heterozygosity in FindGSE [39] and GenomeScope v2.0 [40], respectively. Smudgeplot (v.4) [40] was used to estimate ploidy of EV genomes using quality filtered stLFR or Illumina paired-end reads.

2.3.2 *De novo* assembly of the enset genome

Adapter sequence filtered stLFR reads were assembled using Supernova (v2.1.1) [41] for stLFR reads (https://github.com/BGI-Qingdao/stlfr2supernova_pipeline#ref). Completeness of the assembled

genome was assessed by Benchmarking Universal Single-Copy Orthologs (BUSCO) v.5.2.1 [42] using land plants (embryophyta_odb10) lineage with -long parameter option.

2.3.3 Repeat identification in enset genome assembly

A repeat sequence library was first generated for the enset genome using a combination of homology, *ab initio* and transposable element (TE) structure-based prediction methods. This library was then used to identify repetitive sequences in the enset genome.

For homology-based repeat element prediction, RepeatProteinMask (v.4.1.1) was used with a *p*-value threshold of $1e^{-6}$ and command-line options -noLowSimple -engine ncbi to search against the RepBase database (RepBaseRepeatMaskerEdition-20181026) at the protein level. Then, RepeatMasker (v3.2) [43] was executed using options -nolow -no_is -norna -engine ncbi -species Viridiplantae on the RepeatProteinMask-masked genome against Dfam [44] v3.2 and the RepBase [45] database for DNA-level transposable element searching.

Subsequently, the genome assembly was further scanned for *ab initio* repetitive element prediction using Miniature Inverted-repeat Transposable Elements (MITE)-Hunter [46], Tandem Repeat Finder (TRF98 v4.09) [47], Piler (v1.0) [48] and RepeatModeler2 (v2.0.1) [49]. TRF was run with 2 5 7 80 10 50 2000 -m -h command-line options, MITE-hunter with -n 5 -S 12345678 -P 1, and PILER as described in <https://www.drive5.com/piler/>. All predicted *ab initio* repeat elements were merged and processed to exclude sequences containing protein coding genes as determined by BLASTP (e-value threshold of $1e^{-10}$) against a monocot protein database downloaded from Ensembl Plants (<https://plants.ensembl.org/index.html>) [50] and Banana Genome Hub (<https://banana-genome-hub.southgreen.fr/>) [51]. The resulting *ab initio* predicted repetitive elements not encoding proteins were combined with the repeats found by homology. Redundant sequences that did not satisfy the 80-80-80 rule [52] were removed using CD-HIT (v.4.8.1) [53] (parameters -d 0 -c 0.80 -n 5 -G 1 -g 0 -M 0).

For structure-based identification of repetitive sequences, Extensive *de-novo* Transposable Element Annotator (EDTA v1.9.9) [54], a combination of multiple TE predicting tools [55–59] was used to identify TEs. For EDTA, coding sequences of *Musa* genomes i.e. *M. acuminata* (v4.0) and *M. balbisiana* (v1.1) downloaded from Banana Genome Hub (https://banana-genome-hub.southgreen.fr) and TEs identified by RepeatProteinMask or RepeatMasker were used as input for -cds and -rmout parameters of EDTA, respectively.

TEs predicted from EDTA were finally merged with the non-redundant repeat library obtained from homology and *ab initio*-based predictions. The combined repeat library was supplied to RepeatMasker and used to identify and mask repetitive regions of the enset genome prior to gene prediction. TE families identified by RepeatMasker were further annotated using TEsorter (v1.4) [59].

2.3.4 Gene prediction in enset genome assembly

A combination of *ab initio* gene prediction, homology-based gene prediction and evidence-based gene prediction approaches from BRAKER2 (v2.1.6) [60] and MAKER (v3.01.03) [61] pipelines were applied to annotate gene structure from repeat-masked enset genomes.

BRAKER2 was used in combination with GeneMark-ETP [62–65] and AUGUSTUS (v3.4.0) [66] to predict *ab initio* gene models supported by protein homology and/or RNA-seq evidence. The raw RNA-seq data from root and young enset leaves of landrace Mazia were aligned against repeat masked genome with hisat2 (v2.1.0) [67] –max-intronlen 160000 –no-mixed –very-sensitive –no-discordant parameter settings, and the alignment processed with SAMtools to generate indexed RNA-seq alignment for input to BRAKER2. Amino acid sequences of *Arabidopsis thaliana* and 11 monocot species were obtained from Phytozome (v13) database and Banana Genome Hub and used as homology-based evidence: *Oryza sativa*, *Oryza indica*, *Sorghum bicolor*, *Dioscorea rotundata*, *Brachypodium distachyon*, *Zea mays*, *Ananas comosus*, *Musa acuminata*, *Musa balbisiana*, *Musa schizocarpa*, and *Ensete glaucum*.

Three rounds of MAKER predictions were performed to generate the final enset gene structures. The first two rounds were used to train SNAP (version 2006-07-28) [68] *ab initio* prediction and the third round for combining all the *ab initio*-, homology- and evidence-based gene models. RNA-seq data from roots and leaves of enset landrace Mazia were assembled using Trinity (v2.1.1) [69] with default parameter settings and further analysed to build splice-aware and unique transcript assemblies using PASA pipeline (v2.4.1) (<https://github.com/PASApipeline/PASApipeline>). The assembled Mazia transcripts, together with proteins from *A. thaliana* and the 11 monocots, were imported into MAKER as evidence and homology-based hits used to generate the initial gene models using MAKER settings `est2genome = 1` and `protein2genome = 1` with no *ab initio* gene predictors. SNAP was then trained with gene models from this first-round prediction and the resulting gene models were added as *ab initio* input to MAKER for second round prediction. These were then used to retrain and generate the final SNAP *ab initio* gene models for MAKER. Evidence for protein coding was determined by integrating SNAP, GENEMARK and AUGUSTUS gene models, in combination with protein homology and evidence from assembled RNA-seq transcripts in the MAKER pipeline. High-confidence gene models having an Annotation Edit Distance (AED) of less than or equal to 0.25 were retained as a final gene model set of MAKER annotation. AED [70], a gene annotation quality measure in MAKER assesses the strength of agreement between a gene annotation and the supporting evidence supplied for gene prediction.

To select the final set of gene models, the predicted protein coding gene models (AED < 0.25) were assessed for inserted TEs that might not have been detected by similarity or TE structural feature detecting tools in the repeat identification and masking steps. Differential TE insertions are prevalent in crop genomes [71] highlighting the need for manual, semi and/or automated curation of structurally annotated genes. Hence, to identify TEs containing predicted protein coding genes (AED < 0.25) of enset, we used TESorter (v1.4) [59] with parameters `-hmm rexdb-plant --pass2-rule 80-80-80` and excluded any gene models harbouring TEs and ran BUSCO to assess annotation completeness on the remaining predicted genes.

2.3.5 Functional annotation of predicted protein-coding genes

Functional annotation was performed using sequence-similarity searches against the NCBI non-redundant (NR) and the Clusters of Orthologous Groups (COGs) databases [72], SwissProt, and TrEMBL [73] using DIAMOND's [74] BLASTP (e-value < $1e^{-5}$, --ultra-sensitive) and Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups (eggNOG)-mapper (v2.1.7) [75] programs with default parameter settings. In addition, local InterProScan (v5.52-86.0) [76] was performed to scan for motifs and domains and ascribe Gene Ontology (GO) terms for annotated genes, and BUSCO was used to assess completeness of the predicted proteome. The GO terms were filtered and synchronised with the updated GO terms and in the core GO to retain their description and exclude obsolete terms (<http://geneontology.org/docs/download-ontology/>).

2.3.6 Annotation of non-coding RNA (ncRNA) genes

Enset genomes were scanned for tRNAs using tRNAscan-SE (v 2.0.9) [77] with eukaryotic parameters. Ribosomal RNA (rRNA) fragments were predicted by BLASTN (e-value < $1e^{-5}$) homology search of *Arabidopsis* (5S, 5.8S and 18S rRNA) and rice (28S rRNA) template rRNA sequences against the assembled genome. MicroRNA (miRNA) and small nuclear RNA (snRNA) sequences were identified using INFERNAL (v1.1.4) [78] cmscan with default parameter settings against the Rfam [79] database as a reference. The Rfam library of covariance model from <ftp://ftp.ebi.ac.uk/pub/databases/Rfam/14.8/Rfam.cm.gz> and clanin file from <ftp://ftp.ebi.ac.uk/pub/databases/Rfam/14.8/Rfam.clanin> were supplied to cmscan for the detection of miRNA and snRNA sequences.

2.4 Comparative analysis of enset and banana genome sequences

2.4.1 Genome assemblies data

In addition to sequence data generated during this study, we obtained data from our previous genome assembly (GenBank: GCA_000818735.3) from the GenBank [80] database via the National Center for Biotechnology Information (NCBI) [81] web portal. We also downloaded *M. acuminata* (v4.0) and *M. balbisiana* (v1.1) reference genome assemblies from the Banana Genome Hub (<https://banana-genome-hub.southgreen.fr>) [51].

2.4.2 Sequence reads data

Genomic sequence reads for 19 EV and 48 banana (*Musa* spp.) genotypes were obtained from NCBI Sequence Read Archive (SRA) [82] (Supplementary Tables 2 and 3) database using esearch (v16.2) and converted into FASTQ format in fastq-dump (v.3.03 options: --split-files --gzip --skip-technical). Long

reads of *Musa balbisiana* were obtained from CNGBdb with ID number CNR0255329 (<https://db.cngb.org/search/run/CNR0255329/>).

For genomic reads mapping, a total of 370 Gb of NGS reads from 20 enset genotypes were utilised, representing a total of 650 X coverage of the enset haploid genome. We used 942.9 Gb (1800 X) of sequence reads from 36 Musa AA-genome (AA or AAA type) genotypes, and 418.8 Gb (974 X) from 12 Musa BB-genome (BB-types). These sequence reads, including both short and long reads, are listed in Supplementary Tables 2 and 3.

In total, including the data generated in this study, enset NGS reads cover 16 cultivated landraces (including Arkiya, Bedadeti, Buffero, Derea, Lochinge, Mazia, Nechiwe, Nobo, Siute and Yako) and 2 wild genotypes (Erpha 13 and Erpha 20) originally obtained from enset growing areas in South and Southwestern Ethiopia [30]. However, two genotypes (ITC1387 and JungleSeeds [31]) had unknown origins (Supplementary Table 2).

The banana genotypes were double haploid (DH)-Pahang (AA), Pisang Kra (AA), Pisang Mas (AA), Green-Red (AA), Calypso (AA), subsp. Zebrina (AA), Cavendish (AAA), Gros Michel (AAA), DH-Pisang Klutuk Wulung (BB), Cultivar DYN018 (BB), Cultivar DYN049 (BB), Cultivar DYN019 (BB) and Cultivar DYN302 (BB) (Supplementary Table 3).

2.4.3 Quality control for sequence reads data

Downloaded Illumina paired reads were assessed for quality using FastQC (v0.11.9) [83]. Reads were trimmed on quality score and filtered for adapter sequences using TrimGalore (v0.6.7) (<https://github.com/FelixKrueger/TrimGalore>, parameter: --quality 30). The stLFR paired-end reads of landrace Mazia generated in this study were quality filtered using stLFRdenovo (v.1.0.5) (<https://github.com/BGI-biotools/stLFRdenovo>).

Pacific Bioscience high-fidelity (PacBio HiFi) long sequencing reads [84] were also cleaned from adapter sequences using cutadapt v.4.3 [85] (parameters: -b "ATCTCTCTCAACAACAACGGAGGAGGAGGAAAAGAGAGAGAT; min_overlap = 35" -b "ATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGAT; min_overlap = 35" --discard-trimmed --revcomp -e 0.1 --report= minimal --discard). PacBio contiguous long reads (CLRs) were error-corrected using Canu v2.3 and Porechop v0.2.4 (<https://github.com/rrwick/Porechop>, parameter: --adapter_threshold 95) was used to remove adapter sequences from Oxford Nanopore Technology (ONT) reads of banana genotypes (*Musa* spp.).

2.4.4 Aligning sequence reads versus reference genome assembly

Filtered and trimmed *E. ventricosum* (EV) NGS reads were mapped to *M. acuminata* (MA, AA-genome) and *M. balbisiana* (MB, BB-genome) genome assemblies using bwa-mem2 [86]. Non-aligned reads were discarded using SAMtools (v1.22) (parameter: -F 4) and PCR duplicates were filtered out using the

MarkDuplicates package of PICARD tools (v2.27.1) [87]. Similarly, quality filtered NGS reads of *Musa* AA-genome (AA or AAA type) genotypes and *Musa* BB-genome (BB-type) genotypes were mapped against the two EV genome assemblies, landraces Mazia and Bedadeti. In addition, NGS reads of EV genotypes were also mapped back to EV genome assemblies, AA or AAA type banana genotype reads were mapped to the MA genome, and BB-type reads were mapped to the MB genome.

For long read mapping of banana genotypes against EV assemblies, minimap2 (v 2.24-r1122) [88] with options -ax map-hifi, -ax map-pacbio, and -ax map-ont was used to map error corrected or adapter sequence trimmed PacBio HiFi reads, PacBio CLR reads and ONT reads, respectively. These long reads of banana genotypes were also mapped back to genome assemblies of their respective genome groups as described for NGS reads.

2.4.5 Coding sequence (CDS) alignment

Prior to CDS alignment between genomes of EV, and *Musa* spp. (MA and MB), repetitive sequences were first hard-masked with BEDtools (v2.30.0) [89] maskfasta function using genomic coordinates of non-genic regions. In addition, CDS that showed greater than 25% interspecies genomic reads mapping coverage were discarded (See section 2.4.4 for details).

Next, we aligned CDS of enset genomes against CDS of MA and MB genomes using NUCmer [90] and LASTZ [91]. NUCmer (MUMmer4 v4.0.0rc1) [90] alignment was generated first and this was parsed into coordinate representation of the aligned genomic position using show-coords (parameters: -c -d -l -r -T) in MUMmer4 [90]. LASTZ [91] was run on CDS that showed < 25% NUCmer alignment coverage setting parameters --strand=both --gapped --format = PAF options. CDS with less than 25% LASTZ alignment coverage were retained for further analysis. Scripts used for processing the alignments are available at https://github.com/sadikmz/EV_MAB_compGeno.

2.4.6 Overview of approach to finding species-specific genes in enset and banana

To assess the similarity between enset (*E. ventricosum*) versus the major diploid progenitors of cultivated banana (*Musa* spp.), we performed coding sequence (CDS) alignments both at the level of assembly versus assembly and also at the level of sequence reads versus assembly. We utilised genome sequencing reads and genome assemblies of *E. ventricosum* (EV), *M. acuminata* (MA), *M. balbisiana* (MB) and triploid banana (AAA).

Two approaches were used to determine conservation of genes between enset and banana, and thus presence / absence variations (PAVs). The first involved aligning CDS from annotated genome assemblies of one species against CDS from another. The second used alignment of genomic sequence reads from one species against another. In both approaches we examined the fraction of the gene covered by the alignment (breadth of sequence alignment) and where CDS showed less than 25% coverage, we considered that CDS (and its corresponding gene) to be absent. This enabled identification

of species-specific genes that might have evolved *de novo*, been lost in one of the species or undergone divergence from the CDS of their last common ancestor.

2.4.7 Inferring presence/absence of genes from alignment coverages

Genome-wide gene coverage of mapped genomic reads was extracted using BEDtools (v2.30.0) [89] coverage parameter. The obtained coverage values, along with merged LASTZ and NUCmer alignment coverage across coding sequences, were used to identify candidate presence-absence sets of genes among the enset and banana genomes. Breadth of CDS coverage from alignment and reads mapping, and read depth extracted from primary alignments was visualized in karyoploteR [92] R package and Jbrowse [93], respectively. Read depth of primary alignment was extracted using deepTools (v2.30.0) [94] bamCoverage. Qualimap (v.2.2.2-dev) [95] was used to summarize mapping statistics.

2.4.8 Functional annotation of differentially present/absent genes

Identified candidate EV, MA and MB specific genes were assessed for Gene Ontology (GO) [96] terms associated with “Biological” terms. Integrated Genome Viewer (IGV) [97] v2.11.9, tidyverse [98] and ggcoverage [99] R packages were used for data visualisation. Genes of presence-absence variation in enset and *Musa* species that lacked GO-terms annotation were further searched against NCBI NR and SwissProt curated proteins databases and the proteomes of *Arabidopsis thaliana* (TAIR10) and *Oryza sativa* (v7.0) obtained from Phytozome (v13) using BLASTP (e-value < $1e^{-5}$). In addition, InterProScan (v5.52-86.0) [76] and Proteinfer [100] scan was performed to detect gene motifs and conserved domains in the amino acid sequences.

2.4.9 Clustering of proteins into orthogroups

EV, MA and MB specific protein-coding genes that lack detectable homology inferred from sequence similarity with proteins of closely or distantly related species were used to cluster them into orthogroups. Clustering proteins into orthogroups was achieved using OrthoFinder (v.2.5.5) with parameters -M msa -A muscle -T raxml. The resulting output was used for protein structure prediction.

2.4.10 Predicting protein structure

The 3D structures of individual proteins were predicted using AlphaFold2 [101] in monomer model. High-confidence predicted structures showing predicted local-distance difference test (pLDDT) > 70% were further assessed with Foldseek [102–104] (<https://search.foldseek.com/search>) to infer structural similarity with other proteins. The most significant hits (with a *p* value < 0.01) against computationally or experimentally validated structures from protein database were retained. These were visualized with ChimeraX v1.8.

2.5 NLR encoding genes and phylogenetic analysis

2.5.1 Identification of NLR encoding genes in genome assemblies

Genomic locations harbouring nucleotide binding domain leucine rich repeat (NLR) genes in *E. ventricosum* (EV) landraces Mazia and Bedadeti, *Musa acuminata* (MA) and *Musa balbisiana* (MB) genomes were assessed using NLR-Annotator [105] using default parameter values. Predicted NLR loci domains were annotated using the annotateClasses.py Python script [106]. Subsequently, nucleotide sequences of predicted NLR loci in EV were subjected to a reciprocal BLASTN search (e-value < $1e^{-5}$) with enset against *Musa* genomes, and vice versa, to identify their homology.

In addition to searching for NLRs in the genomic DNA sequences, we also searched for NLRs among the amino acid sequences of predicted protein coding genes in the EV, MA and MB genomes. Briefly, a local database of NLRs from a diverse range of plants was created using predicted and experimentally validated NLRs in ANNA [107] (<https://biobigdata.nju.edu.cn/ANNA/>) and RefPlantNLR [108], respectively. The protein sequences of enset and *Musa* species genomes were then searched against the local NLR database using DIAMOND BLASTP (e-value < $1e^{-5}$ --ultra-sensitive) and hits were then aligned in Clustal Omega (1.2.1) [109] with default parameter values to build Hidden Markov Models (HMM) using hmmbuild of HMMER [110] (v.3.3.2). The HMM from the local NLR database was subsequently added to the list of NLR-associated Pfam domains [76]: NB-ARC (PF00931), TIR (PF01582, PF13676), RPW8-type (PF05659), Late blight resistance protein R1 (PF12061), Rx-type (PF18052) and LRRs (PF00560, PF13516, PF18805, PF13855, PF07725, PF12799, PF01463, PF01462, PF18831, PF18837, PF08263, PF07723, and PF13306). The combined HMMs for NLRs were then used to scan against proteins of EV, MA, and MB genomes with hmmsearch (-E 0.0001 and --domE 0.001) in HMMER (v.3.3.2). The identified NLR genes were rescanned for the NLRs highly conserved NB-ARC (PF00931) Pfam domain with hmmsearch using the same parameters to confirm their presence in each sequence. Domains detected in NLR genes were annotated with NCBI Conserved Domains Database (CDD) and InterProScan run with default parameters to identify coiled coil (CC), NB-ARC (N) and RPW8 type domains. The LRRs 1, 9 and 11 MEME (Multiple EM for Motif Elicitation) motifs [111] were used to detect LRR domains of NLR genes using Motif Alignment and Search Tool (MAST) with default parameters.

The identified NLR genes in each genome were also assessed for their overlap with NLR loci predicted by NLR-Annotator from the same genome. NLR genes were mapped back to their individual genomes using TBLASTN (e-value < $1e^{-5}$) and overlapping NLR loci from the two sources were compared using the intersect function in BEDTools (v2.30.0) [89] and a custom R script. NLR loci without a corresponding predicted protein were subjected to BLASTX (e-value < $1e^{-5}$) against the NCBI NR protein database to detect homology to NLRs in other species, thereby increasing the confidence they encoded genuine NLRs.

2.5.2 Phylogenetic analysis of NLRs

The core NB-ARC domains NLRs encoded in genomes of EV landraces Mazia and Bedadeti, MA and MB were aligned using Clustal Omega [109] and all positions with gaps > 50% in aligned sequences were removed using trimAl [112]. The evolution of NB-ARC domains of NLRs was inferred with RAxML (v8.2.12) [113] using the Maximum Likelihood method based on the JTT model and a bootstrap value of 1000 replicates. Non-plant NB-ARC containing domains from RefPlantNLR [108] were included as outgroups to root the phylogenetic tree. The resulting tree was visualized using the Interactive Tree of Life (iTOL) [114].

3. Results

3.1 The enset Mazia genome is triploid

We sequenced the genome of EXW-tolerant landrace Mazia using stLFR at 330X coverage. We estimated the Mazia genome size to be 572.23 Mb and heterozygosity at 1.41%, using the frequency distribution of 21-mers from these unassembled quality-filtered stLFR reads (Fig. 2A and Supplementary Fig. 1). K-mer-pairs of $k = 21$ differing by one SNP showed a strong signal of a triploid Smudgeplot pattern (Fig. 2B) near to 0.3 normalized coverage indicating the Mazia genome is triploid.

Figure 2

3.2 Genome assembly and annotation

The haploid genome assembly length for landrace Mazia was 540.14 Mb, which is approximately 94% of the k-mer based estimated genome size. The assembly comprises 43,459 scaffolds with an N_{50} of 290 kb. The lengths of the scaffolds ranged between 0.5 kb and 2.67 Mb, with an average size of 12.43 kb (Table 1). This represents a substantial improvement over our previous haploid assembly of landrace Bedadeti (GenBank: GCA_000818735.3), which had a total length of 451.28 Mb and an N_{50} of 21 kb.

The Mazia genome assembly had a repetitive-element content of 64.64%, compared with 51.70% for the Bedadeti genome assembly (Supplementary Table 1). We identified 38,940 and 38,623 high-confidence gene models for Mazia and Bedadeti respectively, after excluding genes containing TE insertions (Supplementary Table 4). This represents gene densities of 13.87 kb and 11.68 kb, respectively, per gene. These values are comparable to the gene densities of banana genomes, *M. acuminata* (14.12 kb) and *M. balbisiana* (14.80 kb). We assigned putative function to 83% of predicted protein sequences (Supplementary Figs. 2 and 3) in both Mazia and Bedadeti using NCBI's non-redundant protein sequence database. For Mazia and Bedadeti we annotated 141 and 137 ribosomal RNAs (rRNAs), 1634 and 1461 transfer RNAs (tRNAs), 212 and 251 micro-RNAs (miRNAs), and 209 and 197 small nuclear RNAs (snRNAs), respectively (Supplementary Tables 11 and 12).

Table 1
Assembly and annotation metrics for enset (*E. ventricosum*) genome assemblies of landraces Mazia and Bedadeti.

Description	Mazia	Bedadeti*
Assembly		
Number of scaffolds	43,459	45,745
Total length of the assembly (Mb)	540.14	451.28
Scaffolds N ₅₀ (Kbp)	290	21
Scaffolds L ₅₀	436	6010
Length of largest scaffold (Mb)	2.67	0.21
Average length of scaffolds(kb)	12.43	9.87
BUSCO (genome mode S:D in %)	82.6 : 5.9	80.7 : 3.5
Repetitive sequences content (%)	64.64	51.70
Structural and functional annotation		
Number of predicted protein coding genes	38,940	38,623
Predicted gene length (mean:median bp)	3591 : 1981	3372 : 1890
CDS length (mean:median bp)	230 : 144	229 : 142
Number of exons per spliced gene (mean:median)	5.8 : 4	5.6 : 4
BUSCO (proteins mode S:D** in %)	68.0 : 5.5	68.5 : 3.8

* GenBank: GCA_000818735.3 and protein coding genes were annotated using RNA-seq and the pipeline employed for landrace Mazia in this study.

** S=Complete and single-copy Benchmarking Universal Single-Copy Orthologs (BUSCO) [42], D=Complete and duplicated BUSCO.

3.3 Most, but not all, banana genes are conserved in enset

We next investigated the conservation of banana genes in enset by aligning enset genome sequencing reads against the reference genome sequences of the major diploid progenitors of cultivated banana: *Musa acuminata* (Fig. 3A) and *M. balbisiana* (Fig. 3B). To complement this approach, we also aligned predicted enset protein-coding DNA sequences (CDS) against the banana reference genomes (Supplementary Figs. 4 and 5). Overall, there is a broad conservation of enset and banana genes across both banana reference genomes with less conservation around the centromeres (Fig. 3). Of the enset sequence reads, 96.7% and 97.7% were aligned against the MA and MB reference genomes (Supplementary Table 5), covering respectively, 64.7% and 69.6% of the MA and MB genomes' lengths.

Most banana genes are conserved in enset. However, 875 (1%) MA and 615 (1%) MB genes were absent from enset based on the alignment of genome sequencing reads and assembled CDS. Of these, only 314 (MA 35.90%) and 413 (MB 67.15%) were annotated with GO-terms. GO-terms associated with Biological Process for these banana-specific genes included disease response (24 in MA and 8 in MB), regulation of DNA-templated transcription (19 in MA and 12 in MB) and protein phosphorylation (18 in MA and 10 in MB) (Fig. 5, Fig. 6B and Supplementary Table 9). Additionally, genes involved in abscisic acid-activated signalling pathways (4 in MA and 4 in MB) and the tricarboxylic acid cycle (3 in MA and 1 in MB) were also identified.

3.4 Enset genes that are not conserved in banana

To ascertain which predicted enset genes are conserved in banana, we aligned banana genomic reads, and banana predicted CDS, against the Mazia and Bedadeti enset genome assemblies. The banana sequence data (Supplementary Table 3) included both short reads, and PacBio/ONT long reads for MA [84] and MB [33] (Fig. 4A and 4B). About 94.50% of the AA-genome reads and 95.30% of the BB-genome banana genomic reads aligned to Mazia. The banana AA-genome reads covered 65.01% (Mazia) and 68.99% (Bedadeti) of the enset genome assemblies, whereas the BB-genome reads were slightly lower (59.93% in Mazia and 65.71% in Bedadeti) (Supplementary Tables 6 and 7). On average, 93% of the CDS in enset genomes were covered (> 25% of their breadth) by aligned banana reads and/or banana CDS. However, 2747 Mazia and 1104 Bedadeti gene predictions did not align against the banana sequences, identifying these as enset-specific genes.

Our analysis identified enset-specific genes whose GO term annotations were predicted to be involved in a range of biological processes, including DNA integration (2 genes), carbohydrate metabolic process (4 genes) and DNA-templated transcriptional regulation (7 genes) (Fig. 6A, and Supplementary Table 8).

However, approximately 1.29% (502 out of 2747 in Mazia) and 0.97% (375 out of 1104 in Bedadeti) of enset genes lacked sequence homology with genomic reads or CDS in banana AA-and BB-genomes and had no functional annotation. These unique genes most likely evolved for roles in enset-specific processes. Notably, the remaining identified enset-specific genes had predicted functions. These include four proteins associated with TEs, a homolog of disease resistance RPP8 (*Resistance to Peronospora Parasitica* 8) gene, now classified as a “helper” resistance gene, and zinc finger protein transcription factors (Supplementary Tables 13 and 14).

Figure 3

Figure 4

Figure 5

Figure 6

3.5 Unique genes

About 1.13% of the EV genes (502 in Mazia and 375 in Bedadeti) not found in MA and MB lacked detectable homology (inferred from sequence similarity with proteins of closely related or distant species) with proteins of closely related or distantly related species. In *Musa* 0.61% (225 out of 36979) of MA and 0.40% (131 out of 33021) of MB genes absent in EV lack detectable homology with proteins of other species. We refer to these sets of genes in enset and banana genomes as “orphan genes” [115–118] or “unique genes”.

To examine evolutionary relationships of these unique protein encoding genes in enset and banana with proteins in distantly related species we used *ab-initio* prediction of protein structure, allowing homology detection between distantly related proteins lacking sequence similarity [119]. Orphan genes in enset and banana genomes were clustered into orthogroups (OrthoFinder) and the structure of species-specific orthogroup proteins were predicted using AlphaFold2 [101] in monomer model. High-confidence structures showing the predicted local-distance difference test (pLDDT) > 70% were further assessed in Foldseek [102–104] to infer structural similarity with other proteins.

AlphaFold2 predicted high-confidence (pLDDT > 70%) protein structures for only 14 (out of 877) EV, 28 MA (out of 225) and 2 MB (out of 131) unique proteins (Supplementary Table 15). Two of the 14 unique EV proteins showed a very high-confidence structure (pLDDT > 90%) and most showed significant ($e\text{-value} < 1e^{-3}$) structural similarity with experimentally determined structures from proteins of unknown functions (Fig. 7A and 7B) or predicted structures from enset proteins.

Figure 7.

Interestingly, in the banana genome, a higher proportion (28 out of 225) of high-confidence structures (pLDDT > 70%) were predicted for MA unique proteins but only a single protein with pLDDT of 77.78% in MB (Supplementary Table 15). Most MB unique proteins reside on scaffolds not assigned to chromosomes and predictions at best fall into the low-confidence ($50\% < \text{pLDDT} < 77.78\%$) category, possibly representing truncated genes resulting from incomplete assembly. All unique proteins were self-aligned with the predicted structure of MA and MB proteins available in the database. A single MA protein was predicted to be an O-methyltransferase activity (pLDDT = 80.35%, Fig. 8A and 8B).

Figure 8.

3.6 Identification of NLR immune receptors

Intracellular nucleotide binding domain leucine rich repeat (NLR) immune receptors are central to plant defence against pathogens, directly or indirectly recognising pathogen effectors to trigger effector-triggered immunity (ETI). NLR gene families exhibit rapid birth-death evolution, with significant copy number variation even among closely related species and cultivars [107, 111, 120, 121]. Given the importance of NLRs in disease resistance and the emerging threat of *Xanthomonas* wilt (XW) to enset and banana cultivation, we analysed the putative NLR repertoire in enset and compared it with the well-annotated banana genomes.

Using assembled genomic sequences, NLR-Annotator [105] predicted 58 and 64 NLR loci in EV landraces Bedadeti and Mazia, respectively, approximately half the number identified in banana genomes (92 in MB and 127 in MA) (Fig. 9A). The relative patterns of NLR class distribution provide informative insights into the enset immune receptor repertoire and are useful for comparing homologous NLRs in draft EV genomes. Our analysis found 127 NLR loci in *M. acuminata* (MA) [32] plus one additional locus that was both annotated as a partial (pseudogene) and a complete (pseudogene), with most loci organised into clusters (2 on Chr3, 1 on Chr7 and 1 on Chr10) (Supplementary Fig. 6).

Cross-species comparisons identified enset-specific NLR loci lacking homologues in both *Musa* species: 6 in EV landrace Bedadeti and 5 in EV landrace Mazia (Supplementary Table 10). Of these, NLR-Annotator [105] classified two as “complete” with intact reading frames, while the remaining were classified as “complete (pseudogene)” or “partial (pseudogene)”. This could represent either a trace of pseudogenised sequence, an assembly artefact or a potential novel NLR locus. BLAST analysis of these enset-specific NLR loci against nucleotides or proteins from other species showed (9 out of 11) similarity with other disease resistance genes reinforcing that these loci are likely potential NLR encoding genes in enset, providing confidence we may have identified novel NLR loci.

Figure 9.

Complementary analysis using annotated protein coding genes identified 47 and 59 NLR-encoding genes in EV landrace Bedadeti and Mazia genomes, respectively (Fig. 9B). In the *Musa* genomes, 51 NLRs were predicted in MB protein coding genes compared with 131 in MA. These predictions were generally more conservative than those of NLR-Annotator results (which predicted 58/64 in EV landrace Bedadeti/Mazia and 92/127 in MB/MA) and represent a combination of both variation in the assembly quality and biological differences between the species. It is highly unlikely that the former accounts for the nearly two-fold higher prediction of NLRs in MB by NLR-Annotator. The notably lower number of predicted NLR genes in MB might simply reflect a stringency in the NLR gene prediction pipeline or represent potential NLR encoding genes missed in the MB annotation pipeline.

To assess our predictions, NLR genes were mapped back to their respective genomes. These NLR genes overlapped with the majority of NLR loci predicted by NLR-Annotator: 55 of 58 loci in EV Bedadeti and 61 of 64 loci NLR loci in EV Mazia, along with 83 of 92 loci in MB and 113 of 127 NLR loci in MA, showed correspondence with predicted NLR encoding genes (Supplementary Figs. 6 and 8). The remaining 66 loci without gene annotation support were queried against the NCBI non-redundant protein or nucleotide collection database using BLASTX or BLASTN search. Of these, 65 (98.5%) had significant BLAST hits, with 62 (93.9%) showing similarity to characterised disease resistance proteins, including RGA family members (25 loci), RPP13-like proteins (11 loci), generic resistance proteins (10 loci), Pik-2-like proteins (3 loci), CC-NBS-LRR (3 loci), and TIR-NBS-LRR (2 loci). The proportion of loci with disease resistance homology was high across all genomes (MA: 100%; MB: 92.3%; EV Bedadeti: 90.0%; EV Mazia: 88.9%), reflecting the majority of the predicted NLR loci encode potential disease resistance-related proteins.

3.7 Phylogenetic analysis of NLRs

To understand the evolutionary relationships among NLRs in enset and banana (MA and MB), we performed phylogenetic analysis using the highly conserved NB-ARC domains. The phylogenetic analysis revealed differences in NLR family expansion between enset and banana (Fig. 10). Enset harbours fewer monophyletic NLR clades compared with *Musa* species, with the majority of expanded clades derived from MA. This pattern is consistent with a previous report of NLR gene family expansion in MA and their reduced representation in MB [33]. Interestingly, resistance to EXW exists in MB [22]. Though fewer NLR clades were expanded in EV compared to *Musa* species, the phylogenetic analysis identified multiple clusters of enset specific NLR gene families (highlighted green or light green in Fig. 10) that may provide a possible unique reservoir that could be exploited for resistance breeding in enset.

Figure 10.

Consistent with the evolutionary loss of TIR-NLRs (TNLs) in monocots, no genomic loci harbouring either Toll/interleukin-1 receptor (TIR) domain-containing NLR (TNL) or TIR encoding NLR [122] genes were detected in enset. Although NLR-Annotator predicted four loci harbouring TNL-like loci in *M. acuminata*, detailed domain analysis revealed these loci showed no detectable TIR signatures, suggesting these as false positives. These are unlikely to be genuine TN/TNLs as TIR NLRs are typically absent or found in strictly limited numbers in monocots [123].

Among non-canonical NLR classes, we identified 4 and 5 Resistance to Powdery Mildew 8 (RPW8)-domain containing NLRs (RNLs) in enset and *Musa* gene models, respectively. RNLs, also known as "helper" NLRs, are involved in transducing signals from "sensor" NLRs [124], predominantly TNLs, to activate the effector triggered hypersensitive response in plants [125, 126]. Phylogenetic analysis indicated that these "helper" RNLs appear to have recently evolved or expanded paralogue pairs in enset and *Musa* genomes. The conservation of RNLs in enset, despite its reduced overall NLR repertoire, indicates that the core immune signalling machinery remains intact and may provide a foundation for functional disease resistance even with fewer sensor NLRs.

4. Discussion

Ensete ventricosum (EV), a unique and multipurpose orphan crop, is cultivated in south and southwestern Ethiopia, supporting over 20 million people for both food and non-food purposes. Despite its contribution to food security and resilience, genomic analysis of enset lags behind regionally important crops like banana [32, 33] and other cash crops. Surprisingly, given its immense value to Ethiopia, genomic resources for enset breeding and improvement are virtually non-existent [127–129]. To address this, we generated a genome sequence for enset landrace Mazia, which is resistant to EXW. This represents a substantial improvement on our previously published genome sequence for landrace Bedadeti and enabled comparisons between enset and two related *Musa* banana genomes.

Using our enset assemblies we undertook comparative analyses with high-quality, annotated genome assemblies of the two major progenitors of edible bananas, *M. acuminata* (MA) and *M. balbisiana* (MB).

Despite being phenotypically and phylogenetically related, the genomes of the three species vary considerably. EV shows a marked 10–12% higher abundance of repetitive elements. As enset is clonally cultivated, it is highly likely the stress associated with this could activate distinct TE families through *cis* regulatory elements, epigenetic marks [130], or other mechanisms [131–133]. Indeed, it is well documented that transposition is enhanced by pathogen infection [134, 135], wounding [133, 136] and environmental stresses [137–139]. As domesticated enset is propagated vegetatively, in the absence of the sexual recombination seen in pollinated crops, we speculate that TE activation may underlie the phenotypic plasticity and local adaptation evident in today's myriad enset landraces [140]. Further investigation of the mechanistic role of TEs in shaping the enset genome and its domestication is warranted.

Surprisingly, we found from the EV, *Musa* comparative analyses, that 25% of the enset genome was not conserved in *M. acuminata* or *M. balbisiana*. Strikingly, these unique enset sequences occupy about 7% of enset's gene space and likely reflect the biological divergence that occurred during enset and banana evolution and subsequent domestication following their split from their last common ancestor ~ 51.9 Mya [4, 5].

At the protein level, we detected a large complement of enset-specific genes we predict contribute to the lifestyle, adaptation and corm-productive qualities underpinning enset's agronomic traits. These families include proteins with putative roles in DNA integration (consistent with high TE gene content), carbohydrate metabolism (possibly underpinning corm development), and responses to abiotic and biotic stresses such as drought and disease resistance. Rather than producing edible fruit, enset generates a large corm and pseudostems that are rich in starch. These phenotypic characteristics are consistent with the identification of genes in EV that are associated with carbohydrate metabolism. Stress-induced birth of distinct transposable element families has been documented in other crop species [130–133]. We hypothesise that the enset-specific transposition genes may have driven widespread re-modelling of the enset genome, activated by the extensive physical damage to the apical meristem during clonal cultivation. Future studies warrant comparing the TE landscapes of wild enset as well as domesticated enset.

Identified genes specific to *Musa* species included genes associated with fruit development and flowering, consistent with enset's lifestyle; enset flowers once after about 8–12 years and rarely produces an edible fruit. Similarly, species-specific gene families associated with resistance and response to biotic and abiotic stresses were identified as unique to *Musa*, consistent with recent studies on *Musa* genomes [32, 33, 141], possibly reflecting different local selection pressures.

We used AlphaFold2 to identify structural homologues of enset's orphan genes in distantly related species. Surprisingly, among the genes unique to enset, AlphaFold2 predicted a structure with high confidence for only 2% of EV specific genes whereas in *M. acuminata* this was 13%. Among the proteins unique to enset, a relatively high proportion are predicted to be disordered, compared with proteins unique to *Musa* spp. Disordered proteins often function as scaffold proteins [142–144], are frequently

involved in adaptation to stress [145, 146] but are difficult to model. One may speculate that orphan genes in enset and banana that lack detectable homology may contribute to facilitating lineage-specific novel adaptations to changing environmental conditions. Indeed, it has been proposed that orphan genes contribute to lineage specific adaptation [116], tolerance to disease [147, 148] and drought [149, 150], and could thus help facilitate enset's domestication dynamics.

Enset and *Musa* species also differ in their modes of perenniality and growth morphologies. Cultivated enset has an extended juvenile-to-vegetative growth phase, taking up to 12 or more years. In contrast, *Musa* species often flower within 1.5 years [151]. Enset differs fundamentally from banana in formation of an underground, energy-rich starchy corm, which serves as an important food reserve under drought and stress conditions. Consistent with this, we identified EV specific genes implicated in biological processes associated with allocation of resources/nutrient storage, energy and survival strategies, possibly contributing to prolonging lifespan as in other perennial plant species [152–155]. In contrast, we identified *Musa* species-specific gene families implicated in metabolism associated with fruit development [33]. In summary, these species-specific sets of genes in EV, MA and MB appear to reflect fundamental phenotypic and lifestyle differences between enset and *Musa* species.

Finally, we extended our analysis to comparative profiling of nucleotide binding domain leucine rich repeat (NLR) family disease resistance (R) gene repertoires. The enset landrace Mazia genome harbours only half (46% – 50%) of the *R* gene families encoded in MA genome with a similar NLRs reduction pattern in landrace Bedadeti, and MB. This finding is consistent with previous work reporting that NLR gene family expansion is associated with host-pathogen interactions in MA, and their reduced representation in MB [33]. The reason for low numbers of NLR genes encoded by enset is unclear. Moreover, we also identified multiple enset and banana specific NLR encoding genes. With disease pressure becoming an increasing problem in enset and banana cultivation, particularly EXW and enset landrace Mazia showing high tolerance to EXW, this NLR gene repertoire provides a potentially unique resource that could be exploited for resistance breeding.

Furthermore, we identified differences between genomic and protein-based NLR predictions in counts and NLR class detection, which is expected and biologically meaningful rather than a methodological artefact in our analysis. Genomic approaches detect all sequences containing NLR-associated motifs [105], including pseudogenes with frameshift mutations or premature stop codons, unannotated genes, and partial sequences from assembly fragmentation. In contrast, protein-based approaches require complete open reading frames and depend on gene annotation quality. Interestingly, recent study reported variable coiled-coil (CC)-domain detection rates (36–76%) of CC-NLR [156]. This biological variation in CC domain detectability appears to have functional significance [157] but has important implications for NLR classification in a way that many functional CNLs may be misclassified as "NL" (lacking N-terminal domain) simply because their CC domains fall below detection thresholds, rather than representing genuinely truncated proteins. Interestingly, in the wheat genome, only 46% of genomic NLR loci corresponded to validated protein-coding genes with intact ORFs [105], a ratio similar to our observations in MB. The discrepancy is particularly pronounced in MB, where NLR-Annotator predicted

nearly twice as many loci as the protein-based approach which likely reflect both incomplete gene annotation in NLR clusters and the presence of recently pseudogenised NLR loci that may retain detectable sequence motifs. Given the draft nature of our enset assemblies the reported NLR counts likely represent minimum estimates; chromosome-scale assemblies may reveal additional NLR loci currently fragmented across contigs or located in assembly gaps. Notably, even telomere-to-telomere sequencing of Cavendish banana revealed 123–140 NLR loci across its three haploid genomes [120], compared to 128 in the MA reference, illustrating that substantial NLR copy number variation exists even within species and among haplotypes of cultivars.

This study was constrained by differences in the genome sequencing platforms on which the data were generated, subsequent differences in assembly metrics, and the reliance on putatively inferred gene functions. Our NLR analysis was entirely computational, based on sequence similarity and domain prediction. The comparative patterns observed between species nonetheless provide initial insights into enset immune receptor architecture. Consequently, more comprehensive comparisons of the enset and *Musa* species genomes, along with their domestication signatures, and the validation of putatively identified gene functions were not possible. However, with the advancement of telomere-to-telomere scale genome sequencing, future research can advance our comprehension of enset domestication, particularly the differences between wild and cultivated enset. This knowledge can then be applied to enhance our understanding of common biotic and abiotic threats and support the development of molecular-based enset trait selection strategies.

Abbreviations

EV: *Ensete ventricosum*; MA: *Musa acuminata*; MB: *Musa balbisiana*; NLR: Nucleotide-binding Leucine-rich Repeat; NB-ARC: Nucleotide-Binding adaptor shared by APAF-1, R proteins, and CED-4; LRR: Leucine-Rich Repeat; CC: Coiled-Coil; TIR: Toll/Interleukin-1 Receptor; RNL: RPW8-NB-LRR (helper NLR); CNL: CC-NB-LRR; TNL: TIR-NB-LRR; EXW: Enset Xanthomonas Wilt; BXW: Banana Xanthomonas Wilt; GO: Gene Ontology; TE: Transposable Element; LTR: Long Terminal Repeat; CDS: Coding Sequence; ORF: Open Reading Frame; stLFR: single-tube Long Fragment Read; BUSCO: Benchmarking Universal Single-Copy Orthologs; pLDDT: predicted Local Distance Difference Test.

Declarations

Data Availability

Raw sequenced reads have been uploaded to the NCBI Sequence Read Archive under accession numbers SRR27722736 for short stLFR reads and SRR27722998, SRR27722999, and SRR27723000 for RNA-seq reads. The assembled genome is available under accession number GCA_052324855.1. Annotations and other files have been submitted to Zenodo [158].

Ethical Approval (optional)

Not applicable.

Consent to Participate Declaration

Not applicable.

Competing Interests

The authors declare that they have no competing interests.

Funding

This study was supported by the University of Warwick Chancellor's International Fellowship (to SM), BGI Health (HK) Company Limited, BBSRC GCRF Accelerator Fund (BB/M017982/1) (to MG), Royal Society Challenge Grants (CHG\R1\170082) (to VN), and UKRI/BBSRC: BB/M017982/1 Warwick Integrative Synthetic Biology Centre funding (to AD-F). SKS acknowledges 10KP project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author's Contributions

SM, DJS and MG analysed and interpreted the data. SM wrote the original manuscript draft. PP, LB and AD-F contributed to methodology and investigation. SKS, AVD, GM, ZY, KT and VN provided resources. DJS and MG supervised the project.

Acknowledgements

The authors thank Southern Agricultural Research Institute for providing samples as well as necessary services and facilities during sample collection, Bio and Emerging Technology Institute (BETin) for facilitating DNA extraction and Ethiopian Biodiversity Institute (EBI) for facilitating sample export for sequencing. We used computing resources provided by the University of Warwick Scientific Computing Research Technology Platform (SCRTP).

References

1. Brandt SA, Spring A, Hiebsch C, McCabe JT, Tabogie E, Wolde-Michael G, et al. The Tree against hunger: Enset based agricultural systems in Ethiopia. American Association for the Advancement of Science; 1997.
2. Borrell JS, Biswas MK, Goodwin M, Blomme G, Schwarzacher T, Heslop-Harrison JSP et al. Enset in Ethiopia: a poorly characterized but resilient starch staple. *Annals of Botany*. Narnia, 2019, pp. 747–66. 10.1093/aob/mcy214
3. WorldBank. DataBank: Population estimates and projections filter by sub-saharan Africa (excluding high income) and using time 2025. <https://databank.worldbank.org/source/population-estimates-and-projections#advancedDownloadOptions>

4. Janssens SB, Vandelook F, De Langhe E, Verstraete B, Smets E, Vandenhouwe I et al. Evolutionary dynamics and biogeography of Musaceae reveal a correlation between the diversification of the banana family and the geological and climatic history of Southeast Asia. *New Phytologist*. 2016, pp. 1453–65. 10.1111/nph.13856
5. Fu N, Ji M, Rouard M, Yan HF, Ge XJ. Comparative plastome analysis of Musaceae and new insights into phylogenetic relationships. *BMC Genomics*. 2022;23(1):223. 10.1186/s12864-022-08454-3.
6. Pijls LTJ, Timmer AAM, Wolde-Gebriel Z, West CE, Pijls T, Ainoid J, timmer AM. Zewdie Wolde - Gwbriel, and Clive E. *J Sci Food Agric*. 1995, pp. 1–11.
7. Tsegaye A, Struik PC. Enset (*Ensete ventricosum* (Welw.) Cheesman) kocho yield under different crop establishment methods as compared to yields of other carbohydrate-rich food crops. *NJAS: Wageningen J Life Sci*. 2001;49 1:81–94. 10.1016/s1573-5214(01)80017-8.
8. Yemataw Z, Mohamed H, Diro M, Addis T, Blomme G. Ethnic-based diversity and distribution of enset (*Ensete ventricosum*) clones in southern Ethiopia. *Journal of Ecology and The Natural Environment*. Academic Journals, 2014, pp. 244–51.
9. Yemataw Z, Tesfaye K, Grant M, Studholme DJ, Chala A. Multivariate analysis of morphological variation in enset (*Ensete ventricosum* (Welw.) Cheesman) reveals regional and clinal variation in germplasm from south and south western Ethiopia. *Aust J Crop Sci*. 2018.
10. Hildebrand EA. A Tale of Two Tuber Crops: How Attributes of Enset and Yams may have Shaped Prehistoric Human-Plant Interactions in Southwest Ethiopia. *Rethinking Agriculture: Archaeological and Ethnoarchaeological Perspectives*. New York: Tylor & Francis group; 2007. pp. 273–98.
11. Daba T, Shigeta M. Enset (*Ensete ventricosum*) Production in Ethiopia: Its Nutritional and Socio-Cultural Values. *Agric Food Sci Res*. 2016.
12. Hamdan MF, Karlson CKS, Teoh EY, Lau SE, Tan BC. Genome Editing for Sustainable Crop Improvement and Mitigation of Biotic and Abiotic Stresses. *Plants* (Basel). 2022;11:19. 10.3390/plants11192625.
13. Abraham A, Winter S, Richert-Pöggeler KR, Menzel W. Molecular characterization of a new badnavirus associated with streak symptoms on enset (*Ensete ventricosum*, Musaceae). *J Phytopathol*. 2018;166(7–8):565–71. 10.1111/jph.12719.
14. Addis T, Azerefegne F, Blomme G, Kanauja K, Addis T, Azerefegne F, et al. Biology of the Enset Root Mealybug *Cataenococcus ensete* and its Geographical Distribution in Southern Ethiopia. *J Appl Biosci*. 2008;8 1:251–60.
15. Gebre SL, Woldeyohannes A, Getahun K, Regassa A. Determinants of the spatial distribution of enset (*Ensete ventricosum* Welw. Cheesman) wilt disease: Evidence from Yem special district, Southern Ethiopia. *Cogent Food Agric*. 2021;7(1):1889789. 10.1080/23311932.2021.1889789.
16. Kidane SA, Meressa BH, Haukeland S, Hvoslef-Eide T, Magnusson C, Couvreur M, et al. Occurrence of plant-parasitic nematodes on enset (*Ensete ventricosum*) in Ethiopia with focus on *Pratylenchus goodeyi* as a key species of the crop. *Nematology*. 2020;23 5:529–41. 10.1163/15685411-bja10058.

17. Yirgou D, Bradbury JF. Bacterial wilt of enset (*Ensete ventricosum*) incited by *Xanthomonas musacearum* sp.n. *Phytopathology*. 1968, pp. 111–2.
18. Yirgou D, Bradbury JF. A note on wilt of banana caused by the enset wilt organism *Xanthomonas musacearum*. *East African Agricultural and Forestry Journal*. Taylor & Francis; 1974. pp. 111–4.
19. Studholme DJ, Wicker E, Abrare SM, Aspin A, Bogdanove A, Broders K, et al. Transfer of *Xanthomonas campestris* pv. *arecae* and *X. campestris* pv. *musacearum* to *X. vasicola* (Vauterin) as *X. vasicola* pv. *arecae* comb. nov. and *X. vasicola* pv. *musacearum* comb. nov. and Description of *X. vasicola* pv. *vasculorum* pv nov *Phytopathology*®. 2020;110 6:1153–60. 10.1094/phyto-03-19-0098-le.
20. Blomme G, Dita M, Jacobsen KS, Perez Vicente L, Molina A, Ocimati W, et al. Bacterial Diseases of Bananas and Enset: Current State of Knowledge and Integrated Approaches Toward Sustainable Management. *Front Plant Sci*. 2017;8:1290. 10.3389/fpls.2017.01290.
21. Nakato V, Mahuku G, Coutinho T. *Xanthomonas campestris* pv. *musacearum*: a major constraint to banana, plantain and enset production in central and east Africa over the past decade. *Mol Plant Pathol*. 2018;19 3:525–36. 10.1111/mpp.12578.
22. Nakato GV, Christelová P, Were E, Nyine M, Coutinho TA, Doležal J, et al. Sources of resistance in *Musa* to *Xanthomonas campestris* pv. *musacearum*, the causal agent of banana xanthomonas wilt. *Plant Pathol*. 2019;49–59. 10.1111/ppa.12945.
23. Gemedá B, Tesfaye G, Simachew A, Wang A, Mekonnen A, Guadie A, et al. Responses of common enset (*Ensete ventricosum*) varieties in Ethiopia to *Xanthomonas* and virulence of *Xanthomonas* strains. *Biocatal Agric Biotechnol*. 2023;53:102872. <https://doi.org/10.1016/j.bcab.2023.102872>.
24. Muzemil S, Chala A, Tesfaye B, Studholme DJ, Grant M, Yemataw Z, et al. Evaluation of 20 enset (*Ensete ventricosum*) landraces for response to *Xanthomonas vasicola* pv. *Musacearum* infection. *Eur J Plant Pathol*. 2021;161 4:821–36. 10.1007/s10658-021-02365-x.
25. Yemataw Z, Tesfaye K, Zeberga A, Blomme G. Exploiting indigenous knowledge of subsistence farmers' for the management and conservation of Enset (*Ensete ventricosum* (Welw.) Cheesman) (*musaceae* family) diversity on-farm. *J Ethnobiol Ethnomed*. 2016;12(1):34. 10.1186/s13002-016-0109-8.
26. Jones JDG, Staskawicz BJ, Dangl JL. The plant immune system: From discovery to deployment. *Cell*. 2024;187 9:2095–116. 10.1016/j.cell.2024.03.045.
27. Maruta N, Burdett H, Lim BYJ, Hu X, Desa S, Manik MK, et al. Structural basis of NLR activation and innate immune signalling in plants. *Immunogenetics*. 2022;74 1:5–26. 10.1007/s00251-021-01242-5.
28. Yemataw Z, Chala A, Ambachew D, Studholme D, Grant M, Tesfaye K. Morphological Variation and Inter-Relationships of Quantitative Traits in Enset (*Ensete ventricosum* (welw.) Cheesman) Germplasm from South and South-Western Ethiopia. *Plants*. 2017;6 4:56. 10.3390/plants6040056.
29. Yemataw Z, Blomme G, Muzemil S, Tesfaye K. Assessing qualitative and phenotypic trait diversity in Ethiopian enset (*Ensete ventricosum* (Welw.) Cheesman) landraces. *Fruits*. Société d'éditions

- techniques continentales, 2018, pp. 310–27. 10.17660/th2018/73.6.2
30. Yemataw Z, Muzemil S, Ambachew D, Tripathi L, Tesfaye K, Chala A, et al. Genome sequence data from 17 accessions of *Ensete ventricosum*, a staple food crop for millions in Ethiopia. Data in Brief. Elsevier; 2018. pp. 285–93. 10.1016/j.dib.2018.03.026.
 31. Harrison J, Moore K, Paszkiewicz K, Jones T, Grant M, Ambacheew D, et al. A Draft Genome Sequence for *Ensete ventricosum*, the Drought-Tolerant Tree Against Hunger. Agronomy. 2014;13–33. 10.3390/agronomy4010013.
 32. Belser C, Baurens F-C, Noel B, Martin G, Cruaud C, Istace B, et al. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. Commun Biology. 2021;4(1). 10.1038/s42003-021-02559-3.
 33. Wang Z, Miao H, Liu J, Xu B, Yao X, Xu C, et al. *Musa balbisiana* genome reveals subgenome evolution and functional divergence. Nature Plants Nat Publishing Group. 2019;810–21. 10.1038/s41477-019-0452-6.
 34. Busche M, Pucker B, Viehöver P, Weisshaar B, Stracke R. Genome Sequencing of *Musa acuminata* Dwarf Cavendish Reveals a Duplication of a Large Segment of Chromosome 2. G3 (Bethesda). 2020;10 1:37–42. 10.1534/g3.119.400847
 35. Sambles C, Venkatesan L, Shittu OM, Harrison J, Moore K, Tripathi L, et al. Genome sequencing data for wild and cultivated bananas, plantains and abacá. Data Brief. 2020;33:106341. 10.1016/j.dib.2020.106341.
 36. Hulse-Kemp AM, Maheshwari S, Stoffel K, Hill TA, Jaffe D, Williams SR, et al. Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library. Hortic Res. 2018. 10.1038/s41438-017-0011-0.
 37. Logemann J, Schell J, Willmitzer L. Improved method for the isolation of RNA from plant tissues. Anal Biochem. 1987;163 1:16–20. 10.1016/0003-2697(87)90086-8.
 38. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011.
 39. Sun H, Ding J, Piednoël M, Schneeberger K. findGSE: estimating genome size variation within human and *Arabidopsis* using k-mer frequencies. Bioinformatics. 2018;34 4:550–7. 10.1093/bioinformatics/btx637.
 40. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11(1). 10.1038/s41467-020-14998-3.
 41. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27 5:757–67. 10.1101/gr.214874.116.
 42. Manni M, Berkeley MR, Seppely M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38 10:4647–54. 10.1093/molbev/msab199.
 43. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protocols Bioinf. 2009. 10.1002/0471250953.bi0410s25.

44. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* 2016;44(D1):D81–9. 10.1093/nar/gkv1272.
45. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* 2005;110(1–4):462–7. 10.1159/000084979.
46. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research.* Oxford Academic, 2010, p. e199–e. 10.1093/nar/gkq862
47. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 1999;27 2:573–80. 10.1093/nar/27.2.573.
48. Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. *Bioinformatics.* Oxford Academic, 2005, pp. i152–i8. doi: 0.1093/bioinformatics/bti1003.
49. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C et al. RepeatModeler2 for automated genomic discovery of transposable element families. *P Natl Acad Sci USA.* National Academy of Sciences, 2020, pp. 9451–7. <https://doi.org/10.1073/pnas.1921046117>
50. Bolser D, Staines DM, Pritchard E, Kersey P. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. *Plant Bioinformatics.* Springer New York; 2016. pp. 115–40. 10.1007/978-1-4939-3167-5_6.
51. Droc G, Martin G, Guignon V, Summo M, Sempéré G, Durant E, et al. The Banana Genome Hub: a community database for genomics in the Musaceae. *Hortic Res.* 2022. 10.1093/hr/uhac221.
52. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet.* 2007. 10.1038/nrg2165.
53. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics Oxf Acad.* 2006;1658–9. 10.1093/bioinformatics/btl158.
54. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology BioMed Cent.* 2019;275. 10.1186/s13059-019-1905-y.
55. Ou S, Jiang N, LTR_FINDER_parallel. Parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA.* 2019. 10.1186/s13100-019-0193-0.
56. Shi J, Liang C. Generic repeat finder: A high-sensitivity tool for genome-wide de novo repeat detection. *Plant Physiol.* 2019. 10.1104/pp.19.00386.
57. Su W, Gu X, Peterson T. TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. *Mol Plant.* 2019. 10.1016/j.molp.2019.02.008.
58. Xiong W, He L, Lai J, Dooner HK, Du C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *P Natl Acad Sci USA.* 2014;111:28. 10.1073/pnas.1410068111.

59. Zhang RG, Li GY, Wang XL, Dainat J, Wang ZX, Ou S, et al. TESorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic Res.* 2022. 10.1093/hr/uhac017.
60. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP + and AUGUSTUS supported by a protein database. *NAR Genomics Bioinf.* 2021. 10.1093/nargab/lqaa108.
61. Campbell MS, Holt C, Moore B, Yandell M. Genome Annotation and Curation Using MAKER and MAKER-P. *Curr Protocols Bioinf.* 2014. 10.1002/0471250953.bi0411s48.
62. Lomsadze A, Ter-hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. 2005, pp. 6494–506. 10.1093/nar/gki937
63. Iwata H, Gotoh O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res.* 2012.
64. Gotoh O, Morita M, Nelson DR. Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment. *BMC Bioinformatics.* 2014. 10.1186/1471-2105-15-189.
65. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods.* 2015;12 1:59–60. 10.1038/nmeth.3176.
66. Hoff KJ, Stanke M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr Protoc Bioinf.* 2019;65(1):e57. 10.1002/cpbi.57.
67. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 2019;37 8:907–15. 10.1038/s41587-019-0201-4.
68. Korf I. Gene finding in novel genomes. *BMC Bioinformatics.* 2004. 10.1186/1471-2105-5-59.
69. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 2011;29 7:644–52. 10.1038/nbt.1883.
70. Eilbeck K, Moore B, Holt C, Yandell M. Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics.* 2009;10(1):67. 10.1186/1471-2105-10-67.
71. Vitte C, Fustier MA, Alix K, Tenaillon MI. The bright side of transposons in crop evolution. *Brief Funct Genomics.* 2014;13 4:276–95. 10.1093/bfgp/elu002.
72. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science.* 1997;278 5338:631–7. 10.1126/science.278.5338.631.
73. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 2003;31 1:365–70. 10.1093/nar/gkg095.
74. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods.* 2021;18 4:366–8. 10.1038/s41592-021-01101-x.
75. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale.

- Mol Biol Evol. 2021;38 12:5825–9. 10.1093/molbev/msab293.
76. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 2021;49(D1):D344–54. 10.1093/nar/gkaa977.
77. Chan PP, Lin BY, Mak AJ, Lowe TM. TRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* 2021. 10.1093/bioinformatics/btt509.
78. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* 2013. 10.1093/nar/gkaa1047.
79. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* 2021;49(D1):D192–200. 10.1093/nar/gkaa1047.
80. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I, GenBank. *Nucleic Acids Res.* 2019;47 D1:D94–9. 10.1093/nar/gky989.
81. Sayers EW, Agarwala R, Bolton EE, Brister R, Canese J, Clark K. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2019;47(D1):D23–8. 10.1093/nar/gky1069.
82. Kodama Y, Shumway M, Leinonen R. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* 2012;40 D1:D54–6. 10.1093/nar/gkr854.
83. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010; doi:<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
84. Liu X, Arshad R, Wang X, Li W-M, Zhou Y, Ge X-J, et al. The phased telomere-to-telomere reference genome of *Musa acuminata*, a main contributor to banana cultivars. *Sci Data.* 2023;10:1. 10.1038/s41597-023-02546-9.
85. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal.* 2011;17 1:10. 10.14806/ej.17.1.200.
86. Vasimuddin M, Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 2019-05-01 2019, IEEE. 10.1109/IPDPS.2019.00041
87. Picard Toolkit. Broad Institute. GitHub Repository <https://broadinstitute.github.io/picard/>; Broad Institute. 2019.
88. Li H. Minimap2: pairwise alignment for nucleotide sequences. In: Birol I, editor. *Bioinformatics.* Oxford Academic; 2018. pp. 3094–100. 10.1093/bioinformatics/bty191.
89. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26 6:841–2. 10.1093/bioinformatics/btq033.
90. Marcais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol.* 2018;14(1):e1005944. 10.1371/journal.pcbi.1005944.

91. RS H. Improved pairwise alignment of genomic DNA (Ph.D). USA: The Pennsylvania State University; 2007.
92. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics*. 2017;33 19:3088–90. 10.1093/bioinformatics/btx346.
93. Diesh C, Stevens GJ, Xie P, De Jesus Martinez T, Hershberg EA, Leung A, et al. JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol*. 2023;24(1):74. 10.1186/s13059-023-02914-z.
94. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*. 2014;187–91. 10.1093/nar/gku365. 42 Web Server issue:W.
95. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*. 2012;28 20:2678–9. 10.1093/bioinformatics/bts503.
96. Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, et al. The Gene Ontology knowledgebase in 2023. *Genetics*. 2023;224:1. 10.1093/genetics/iyad031.
97. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14 2:178–92. 10.1093/bib/bbs017.
98. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. *J Open Source Softw*. 2019;4 43:1686. 10.21105/joss.01686.
99. Song Y, Wang J. ggcoverage: an R package to visualize and annotate genome coverage for various NGS data. *BMC Bioinformatics*. 2023;24(1):309. 10.1186/s12859-023-05438-2.
100. Sanderson T, Bileschi ML, Belanger D, Colwell LJ. ProteInfer, deep neural networks for protein functional inference. *Elife*. 2023;12. 10.7554/eLife.80942.
101. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596 7873:583–9. 10.1038/s41586-021-03819-2.
102. Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, et al. Clustering predicted structures at the scale of the known protein universe. *Nature*. 2023;622 7983:637–45. 10.1038/s41586-023-06510-w.
103. Kim W, Mirdita M, Karin EL, Gilchrist CLM, Schweke H, Söding J et al. Rapid and Sensitive Protein Complex Alignment with Foldseek-Multimer. Cold Spring Harbor Laboratory, 2024-04-14 2024.
104. Van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol*. 2024;42 2:243–6. 10.1038/s41587-023-01773-0.
105. Steuernagel B, Witek K, Krattinger SG, Ramirez-Gonzalez RH, Schoonbeek HJ, Yu G, et al. The NLR-Annotator Tool Enables Annotation of the Intracellular Immune Receptor Repertoire. *Plant Physiol*. 2020;183 2:468–82. 10.1104/pp.19.01273.

106. annotateClasses.py. <https://gist.github.com/philippbayer/0052f5ad56121cd2252a1c5b90154ed1>. 2022.
107. Liu Y, Zeng Z, Zhang YM, Li Q, Jiang XM, Jiang Z, et al. An angiosperm NLR Atlas reveals that NLR gene reduction is associated with ecological specialization and signal transduction component deletion. *Mol Plant*. 2021;14 12:2015–31. 10.1016/j.molp.2021.08.001.
108. Kourelis J, Sakai T, Adachi H, Kamoun S. RefPlantNLR is a comprehensive collection of experimentally validated plant disease resistance proteins from the NLR family. *PLoS Biol*. 2021;19 10:e3001124. 10.1371/journal.pbio.3001124.
109. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci*. 2018;27 1:135–45. 10.1002/pro.3290.
110. HMMER. Biosequence analysis using profile hidden Markov models. <http://hmmer.org/>
111. Jupe F, Pritchard L, Etherington GJ, Mackenzie K, Cock PJ, Wright F, et al. Identification and localisation of the NB-LRR gene family within the potato genome. *BMC Genomics*. 2012;13(1):75. 10.1186/1471-2164-13-75.
112. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25 15:1972–3. 10.1093/bioinformatics/btp348.
113. Stamatakis A. *Bioinformatics*. 2014;30 9:1312–3. 10.1093/bioinformatics/btu033. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.
114. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021;49 W1:W293–6. 10.1093/nar/gkab301.
115. Pereira AB, Marano M, Bathala R, Zaragoza RA, Neira A, Samano A, et al. Orphan genes are not a distinct biological entity. *BioEssays*. 2025;47(1):e2400146. 10.1002/bies.202400146.
116. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. *Nat Rev Genet*. 2011;12 10:692–702. 10.1038/nrg3053.
117. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet*. 2009;25 9:404–13. 10.1016/j.tig.2009.07.006.
118. Dujon B. The yeast genome project: what did we learn? *Trends Genet*. 1996;12 7:263–70. 10.1016/0168-9525(96)10027-5.
119. Kessel A, Ben-Tal N. *Introduction to proteins: Structure, function, and motion*. 2nd ed. Chapman and Hall/CRC; 2018.
120. Huang H-R, Liu X, Arshad R, Wang X, Li W-M, Zhou Y, et al. Telomere-to-telomere haplotype-resolved reference genome reveals subgenome divergence and disease resistance in triploid Cavendish banana. *Hortic Res*. 2023;10:9. 10.1093/hr/uhad153.
121. Jacob F, Vernaldi S, Maekawa T. Evolution and conservation of plant NLR functions. *Front Immunol*. 2013.

122. Bernoux M, Ve T, Williams S, Warren C, Hatters D, Valkov E, et al. Structural and Functional Analysis of a Plant Resistance Protein TIR Domain Reveals Interfaces for Self-Association, Signaling, and Autoregulation. *Cell Host Microbe*. 2011;9 3:200–11. 10.1016/j.chom.2011.02.009.
123. Tarr DE, Alexander HM. TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders. *BMC Res Notes*. 2009;2:197. 10.1186/1756-0500-2-197.
124. Jubic LM, Saile S, Furzer OJ, El Kasmi F, Dangl JL. Help wanted: helper NLRs and plant immune responses. *Curr Opin Plant Biol*. 2019;50:82–94. 10.1016/j.pbi.2019.03.013.
125. Kourelis J, Adachi H. Activation and Regulation of NLR Immune Receptor Networks. *Plant Cell Physiol*. 2022. 10.1093/pcp/pcac116.
126. Castel B, Ngou PM, Cevik V, Redkar A, Kim DS, Yang Y, et al. Diverse NLR immune receptors activate defence via the RPW8-NLR NRG1. *New Phytol*. 2019;222 2:966–80. 10.1111/nph.15659.
127. Bhat JA, Yu D, Bohra A, Ganie SA, Varshney RK. Features and applications of haplotypes in crop breeding. *Commun Biol*. 2021;4(1):1266. 10.1038/s42003-021-02782-y.
128. Goddard M. Genomic selection: prediction of accuracy and maximisation of long term response. 2009, pp. 245–57. 10.1007/s10709-008-9308-0
129. Rana S, Aggarwal PR, Shukla V, Giri U, Verma S, Muthamilarasan M. Genome Editing and Designer Crops for the Future. *Methods in Molecular Biology*. Springer US; 2022. pp. 37–69.
130. M.A. G. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim Biophys Acta*. 2015;1849 4:403–16. 10.1016/j.bbagr.2014.07.017
131. Dubin MJ, Zhang P, Meng D, Remigereau M-S, Osborne EJ, Paolo Casale F, et al. DNA methylation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. *eLife*. 2015;4. 10.7554/elife.05255.
132. Lu L, Chen J, Robb SMC, Okumoto Y, Stajich JE, Wessler SR. Tracking the genome-wide outcomes of a transposable element burst over decades of amplification. *Proc Natl Acad Sci U S A*. 2017;114 49:E10550–9. 10.1073/pnas.1716459114.
133. Rocheta M, Carvalho L, Viegas W, Morais-Cecílio L. Corky, a gypsy-like retrotransposon is differentially transcribed in *Quercus suber* tissues. *BMC Res Notes*. 2012;5(1):432. 10.1186/1756-0500-5-432.
134. Yu A, Lepère G, Jay F, Wang J, Bapaume L, Wang Y, et al. Dynamics and biological relevance of DNA demethylation in *Arabidopsis* antibacterial defense. *Proc Natl Acad Sci U S A*. 2013;110 6:2389–94. 10.1073/pnas.1211757110.
135. Ansari KI, Walter S, Brennan JM, Lemmens M, Kessans S, McGahern A, et al. Retrotransposon and gene activation in wheat in response to mycotoxigenic and non-mycotoxigenic-associated *Fusarium* stress. *Theor Appl Genet*. 2007;114 5:927–37. 10.1007/s00122-006-0490-0.
136. Mhiri C, Morel JB, Vernhettes S, Casacuberta JM, Lucas H, Grandbastien MA. The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. *Plant Mol Biol*. 1997;33 2:257–66. 10.1023/a:1005727132202.

137. Fujino K, Hashida SN, Ogawa T, Natsume T, Uchiyama T, Mikami T, et al. Temperature controls nuclear import of Tam3 transposase in *Antirrhinum*. *Plant J.* 2011;65 1:146–55. 10.1111/j.1365-313X.2010.04405.x.
138. Woodrow P, Pontecorvo G, Ciarmiello LF, Fuggi A, Carillo P. Ttd1a promoter is involved in DNA-protein binding by salt and light stresses. *Mol Biol Rep.* 2011;38 6:3787–94. 10.1007/s11033-010-0494-3.
139. Ramallo E, Kalendar R, Schulman AH, Martínez-Izquierdo JA. Reme1, a Copia retrotransposon in melon, is transcriptionally induced by UV light. *Plant Mol Biol.* 2008;66(1–2):137–50. 10.1007/s11103-007-9258-4.
140. Muzemil S, Olango TM, Yemataw Z, Enset. (*Ensete ventricosum*) landrace database (Version 0). 2023; <https://doi.org/10.5281/zenodo.8287638>
141. D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature.* 2012;213–7. 10.1038/nature11241.
142. Meszaros B, Tompa P, Simon I, Dosztanyi Z. Molecular principles of the interactions of disordered proteins. *J Mol Biol.* 2007;372 2:549–61. 10.1016/j.jmb.2007.07.004.
143. Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. Protein disorder -a breakthrough invention of evolution? *Curr Opin Struct Biol.* 2011;21 3:412–8. 10.1016/j.sbi.2011.03.014.
144. Tompa P, Kovacs D. Intrinsically disordered chaperones in plants and animals. *Biochem Cell Biol.* 2010;88 2:167–74. 10.1139/o09-163.
145. Murvai N, Kalmar L, Szabo B, Schad E, Micsonai A, Kardos J, et al. Cellular Chaperone Function of Intrinsically Disordered Dehydrin ERD14. *Int J Mol Sci.* 2021;22(12). 10.3390/ijms22126190.
146. Trivedi R, Nagarajaram HA. Intrinsically Disordered Proteins: An Overview. *Int J Mol Sci.* 2022;23:22. 10.3390/ijms232214050.
147. Moon H, Jeong AR, Kwon OK, Park CJ. Oryza-Specific Orphan Protein Triggers Enhanced Resistance to *Xanthomonas oryzae* pv. *oryzae* in Rice. *Front Plant Sci.* 2022;13:859375. 10.3389/fpls.2022.859375.
148. Xiao W, Liu H, Li Y, Li X, Xu C, Long M, et al. A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS ONE.* 2009;4(2):e4603. 10.1371/journal.pone.0004603.
149. Fakhar AZ, Liu J, Pajerowska-Mukhtar KM, Mukhtar MS. The Lost and Found: Unraveling the Functions of Orphan Genes. *J Dev Biol.* 2023;11:2. 10.3390/jdb11020027.
150. Mofatto LS, Carneiro Fde A, Vieira NG, Duarte KE, Vidal RO, Alekcevetch JC, et al. Identification of candidate genes for drought tolerance in coffee by high-throughput sequencing in the shoot apex of different *Coffea arabica* cultivars. *BMC Plant Biol.* 2016;16:94. 10.1186/s12870-016-0777-5.
151. Chaurasia AK, Patil HB, Krishna B, Subramaniam VR, Sane PV, Sane AP. Flowering time in banana (*Musa* spp.), a day neutral plant, is controlled by at least three FLOWERING LOCUS T homologues. *Sci Rep.* 2017;7:1. 10.1038/s41598-017-06118-x.

152. Grimberg Å, Lager I, Street NR, Robinson KM, Marttila S, Mähler N, et al. Storage lipid accumulation is controlled by photoperiodic signal acting via regulators of growth cessation and dormancy in hybrid aspen. *New Phytol.* 2018;219 2:619–30. 10.1111/nph.15197.
153. Schoonmaker AL, Hillabrand RM, Lieffers VJ, Chow PS, Landhäusser SM. Seasonal dynamics of non-structural carbon pools and their relationship to growth in two boreal conifer tree species. *Tree Physiol.* 2021;41 9:1563–82. 10.1093/treephys/tpab013.
154. Herrera-Ramírez D, Sierra CA, Römermann C, Muhr J, Trumbore S, Silvério D, et al. Starch and lipid storage strategies in tropical trees relate to growth and mortality. *New Phytol.* 2021;230 1:139–54. 10.1111/nph.17239.
155. Aoyagi Blue Y, Kusumi J, Satake A. Copy number analyses of DNA repair genes reveal the role of poly(ADP-ribose) polymerase (PARP) in tree longevity. *iScience.* 2021;24 7:102779. 10.1016/j.isci.2021.102779.
156. Guo BC, Zhang YR, Liu ZG, Li XC, Yu Z, Ping BY, et al. Deciphering Plant NLR Genomic Evolution: Synteny-Informed Classification Unveils Insights into TNL Gene Loss. *Mol Biol Evol.* 2025;42(2). 10.1093/molbev/msaf015.
157. Guo G, Zhao H, Bai K, Wu Q, Dong L, Lu L, et al. An activated wheat CC_{G10}-NLR immune receptor forms an octameric resistosome. *bioRxiv preprint.* 2025. 10.1101/2025.08.26.672026.
158. Muzemil S. Manuscript genome annotation and analysis files. *Zenodo.* 2026; doi:<https://doi.org/10.5281/zenodo.18292608>.

Figures



Figure 1

Overview of enset-based farming system in southern Ethiopia. (A) A homestead enset plantation in Dawro district containing a variety of clonally propagated landraces, each specifically selected for their preferred uses. (B) A range of clonally propagated enset plants approaching harvest growing amongst selected landraces at different developmental stages. Each landrace can be identified by its phenotypic traits such as growth habit, colour of leaf, midrib and/or leaf margins, as well as their preferred uses (e.g.

harvestable plant material or tolerance to disease and drought). (C) The 2-year-old landrace “Mazia” at Areka Agricultural Research Center (Areka, Ethiopia) sequenced in this study. Mazia is one of the least susceptible landraces to Xvm infection [24]. Photos provided by Sadik Muzemil.

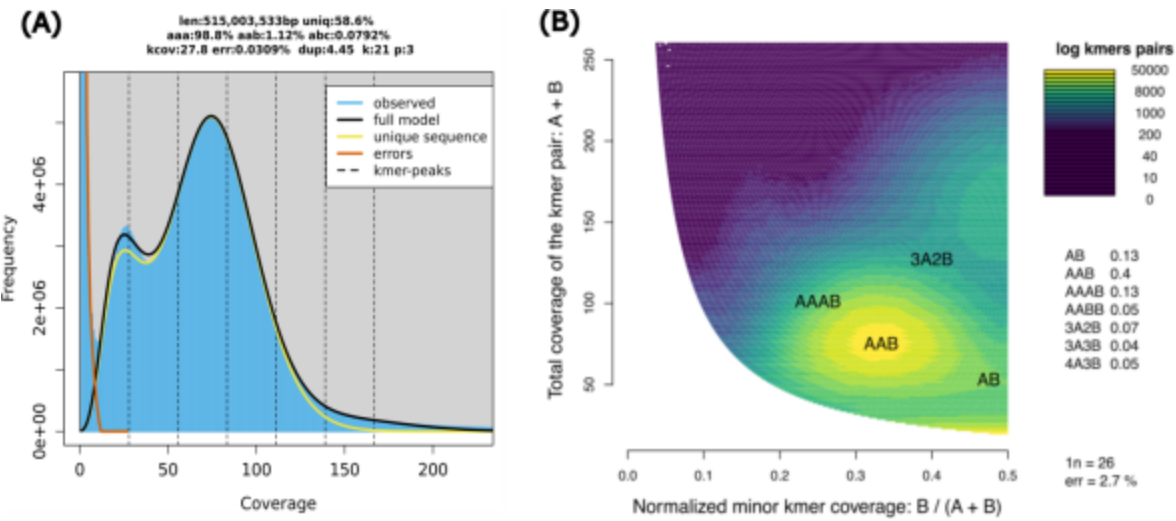


Figure 2

Kmer-based inferred genome profile of *E. ventricosum* (EV) landrace Mazia. (A) GenomeScope v2.0 [40] inferred characteristics of genome structure, including genome size, abundance of repetitive elements, and rate of heterozygosity of EV Mazia genome. (B) Smudgeplot (v.4) [40] (k=21) inferred ploidy level of EV landrace Mazia suggests it is triploid.

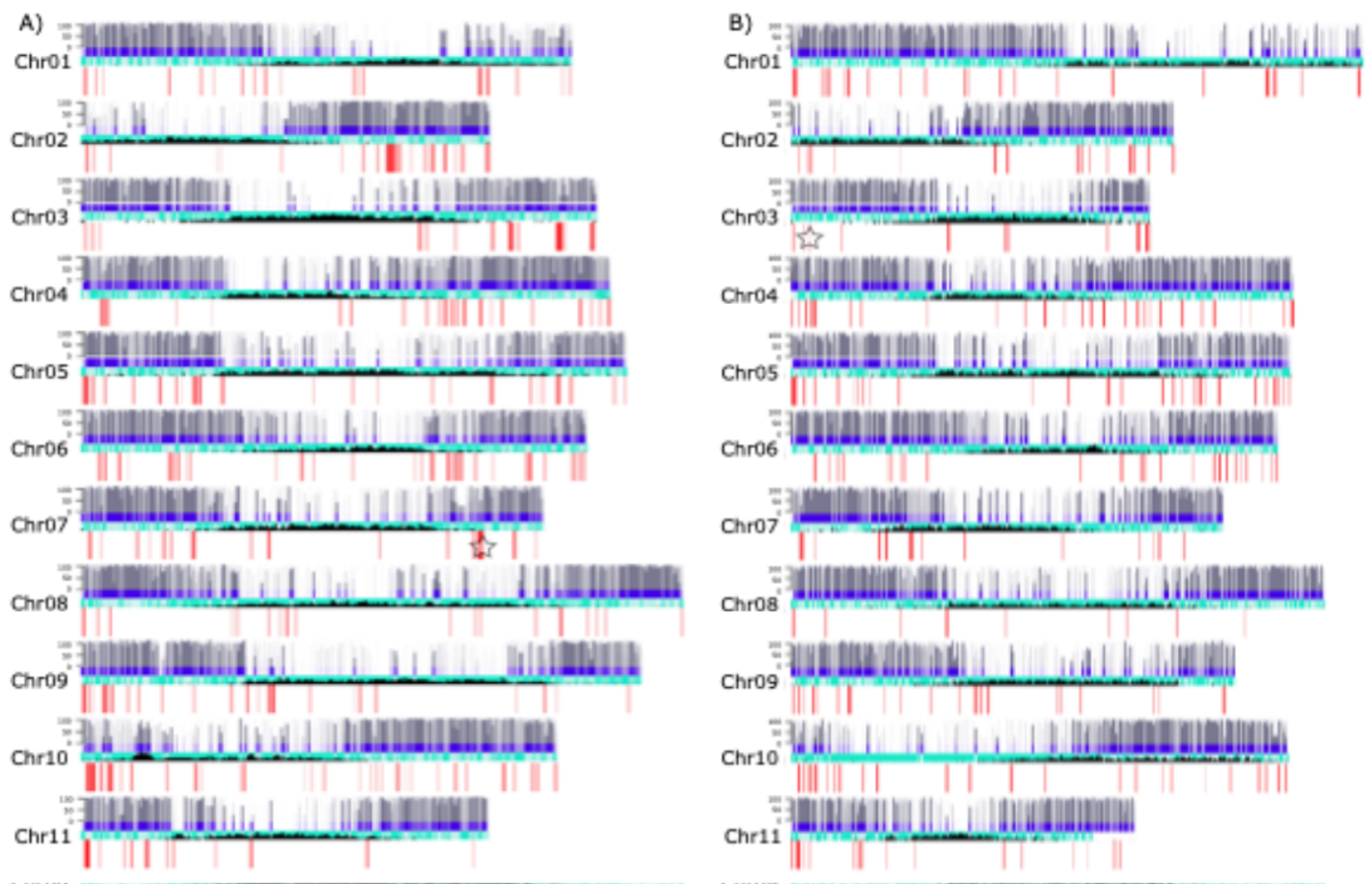


Figure 3

Sequence conservation between banana and enset. Genomic sequence reads of 20 *E. ventricosum* (EV) genotypes were mapped against *M. acuminata* (MA) and *M. balbisiana* (MB) genomes. EV coding sequences (CDS) were aligned against the MA and MB genomes. Breadth of coverage of CDS alignments and reads mapping across MA and MB genes were plotted for MA (A) and for MB (B). Reads mapped with bwa-mem2 (v.2.2.1) default parameter, LASTZ [91] and NUCmer (MUMmer4) [90] were used to produce CDS alignment. Gray coloured bar lines in the upper lane indicate breadth of coverage CDS alignment or NGS reads mapping across each gene (represented by a purple short bars). The grey bar height is proportional to the alignment or mapping fraction coverage per gene locus. Red bar lines underneath indicate gene loci that are absent from enset (<25% breadth of coverage by reads and genomic alignment). Light blue coloured lines below the blue coloured bars represent non-coding regions and black bars indicate the density of centromeric repeats in MA and MB genomes [32]. The stars on MA Chr07 and MB Chr03 indicate the genomic locations specific disease resistance encoding genes (See Figure 5 for details).



Figure 4

Enset (*Ensete ventricosum*) genomic regions with enset-specific genes. (A) A 655 kb region of EV landrace Mazia genome encodes predicted protein-coding genes that lack coverage by read mapping from *M. acuminata* (MA) and *M. balbisiana* (MB) genomes. (B) A 21.5 kb region of the genome of EV landrace Bedadeti contains enset-specific protein-coding genes. In (A) and (B), from top to bottom, the dark green coloured bars indicate mapping depth of MA ONT [32], PacBio HiFi [84], and NGS reads [SRA: ERR3606950]. The light green coloured bars were for MB PacBio [33] or NGS reads [SRA: ERR10695603], as annotated. The blue lanes represent mapping depth of EV NGS reads from domesticated landraces Bedadeti and Mazia (generated in this study), while red lanes from top to bottom indicate NGS reads from wild EV landraces reported by [30] and [31]. Predicted protein coding genes are indicated in the bottom lane as yellow and red boxes, with red boxes representing an EV specific set of genes that lack either reads mapping or CDS alignment from AA-genome and BB-genome banana sequences.

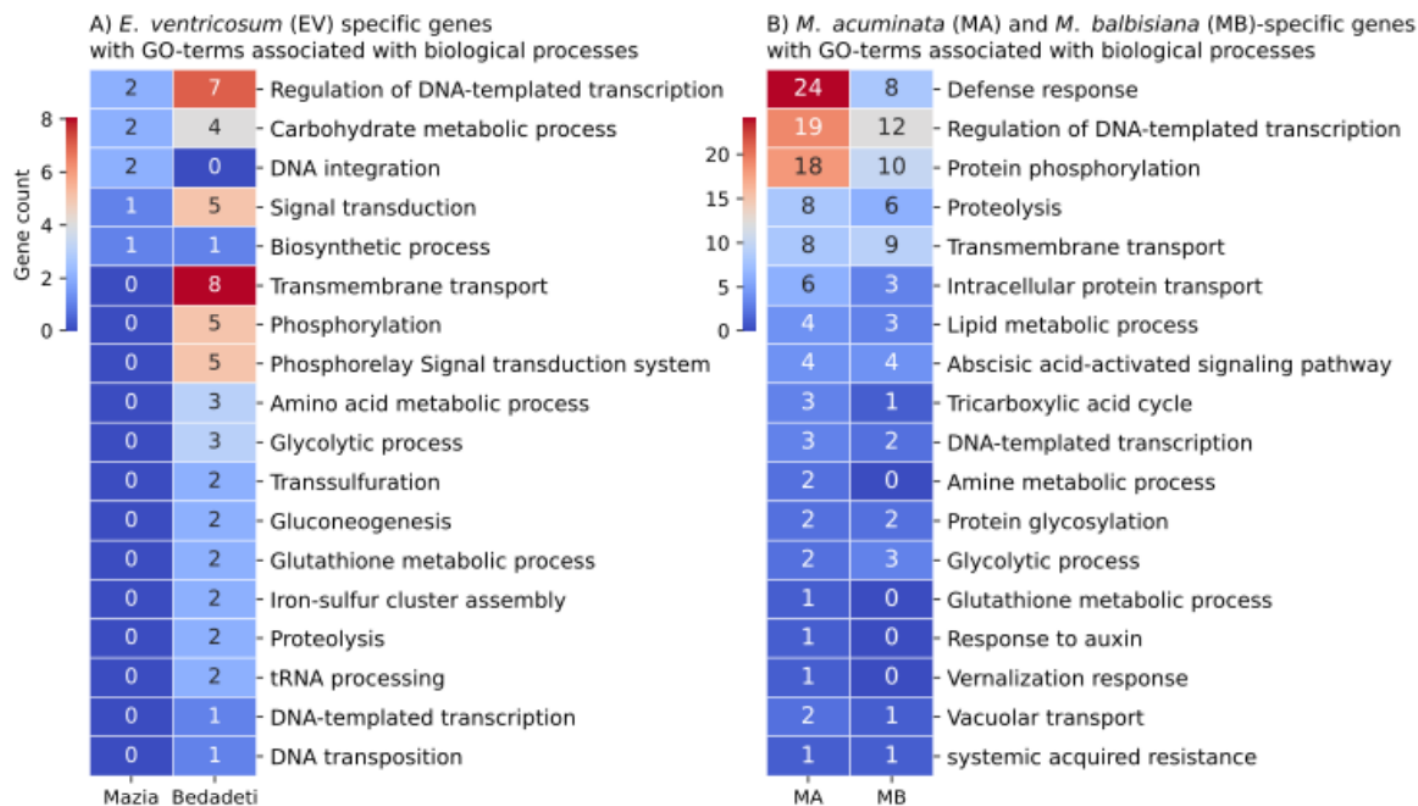


Figure 6

Annotated Biological Processes Gene Ontology (GO) terms for unique genes in Enset or MA and MB genomes. GO terms associated with biological processes in species-specific protein coding genes of EV landraces Mazia and Bedadeti (A), or MA and MB (B) specific protein coding genes. The abundance of identified genes with associated GO terms is indicated by the colour key, with dark blue representing zero and dark red the largest value depicted in A and B. Presence or absence of genes in the three species were identified based on interspecies genome alignment or read mapping across coding sequences. Genes with breadth of coverage <25% of the CDS's length were considered species-specific (see Supplementary Tables 7 and 8 for detailed lists).

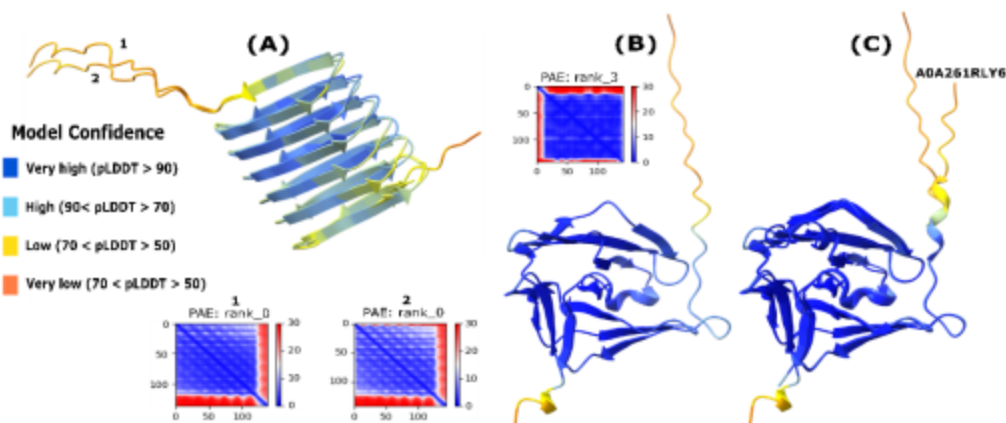


Figure 7

Predicted 3D structures of selected enset unique proteins (*Ensete ventricosum*) and their similarity with distantly related proteins. (A) Aligned protein structures of a pair of unique enset genes from Mazia (#1) and Bedadeti (#2) landraces. (B) A very high confidence (pLDDT = 91.1%) predicted structure of a Bedadeti protein showing homology with NTP_transf_9 domain-containing protein (UniProt accession ID A0A261RLY6) as indicated in (C). Protein structures were predicted by AlphaFold2 and assessed for structural similarity in Foldseek. Each residue in the sequence is colour-coded based on the model confidence score (pLDDT). The predicted aligned error (PAE) rank box indicates the selected AlphaFold2 model.

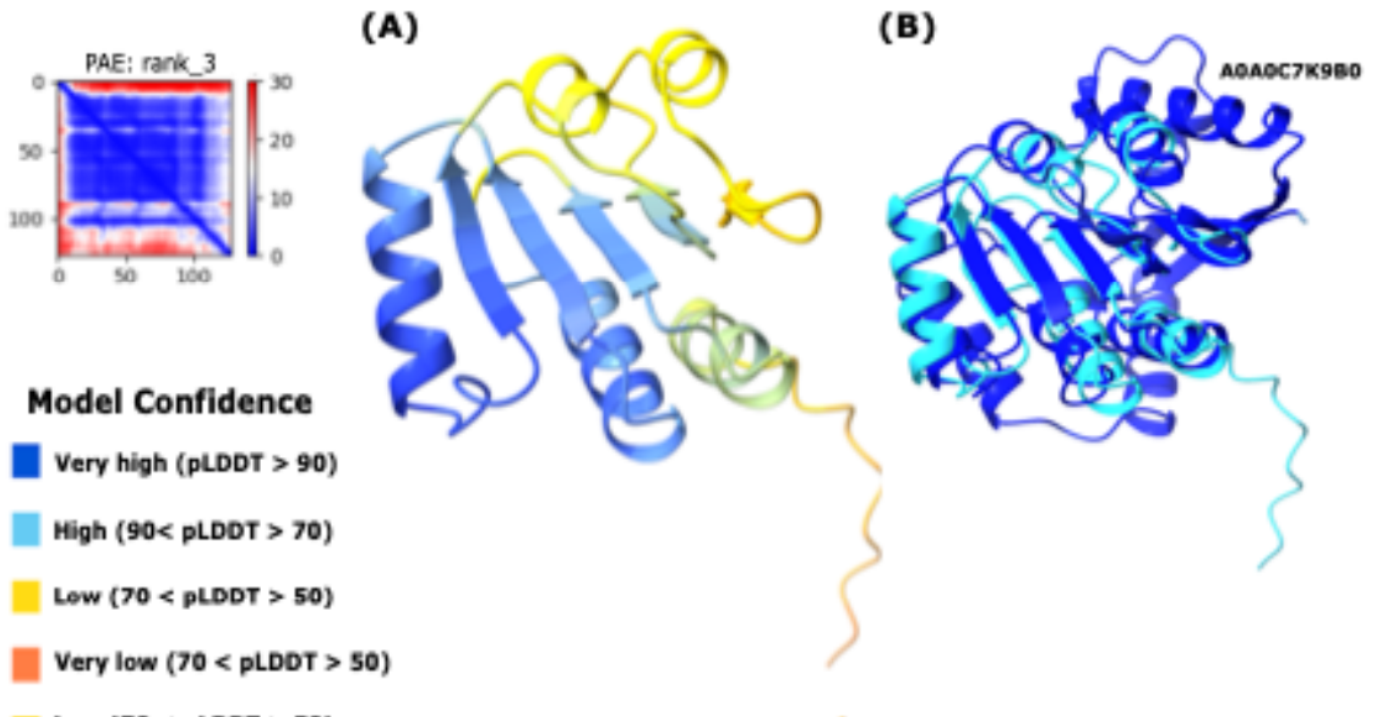


Figure 8

Predicted 3D structure of selected unique proteins in *Musa acuminata*(MA) and their similarity with distantly related proteins. A) A 127 amino acid residue protein structure (pLDDT = 80.35 %) of an MA unique gene showing disordered N-terminal structure and a relatively low confidence C-terminal structure. B) A MA unique protein prediction in Figure A (coloured vivid cyan in Figure B) showing structural similarity with O-methyltransferase (UniProt accession ID A0A0C7K9B0). Protein structures were predicted by AlphaFold2, with sequences colour coded with respect to the confidence score (pLDDT) and were assessed for structural similarity in Foldseek. The predicted aligned error (PAE) rank box in the top right of the figure indicates the selected AlphaFold2 model.

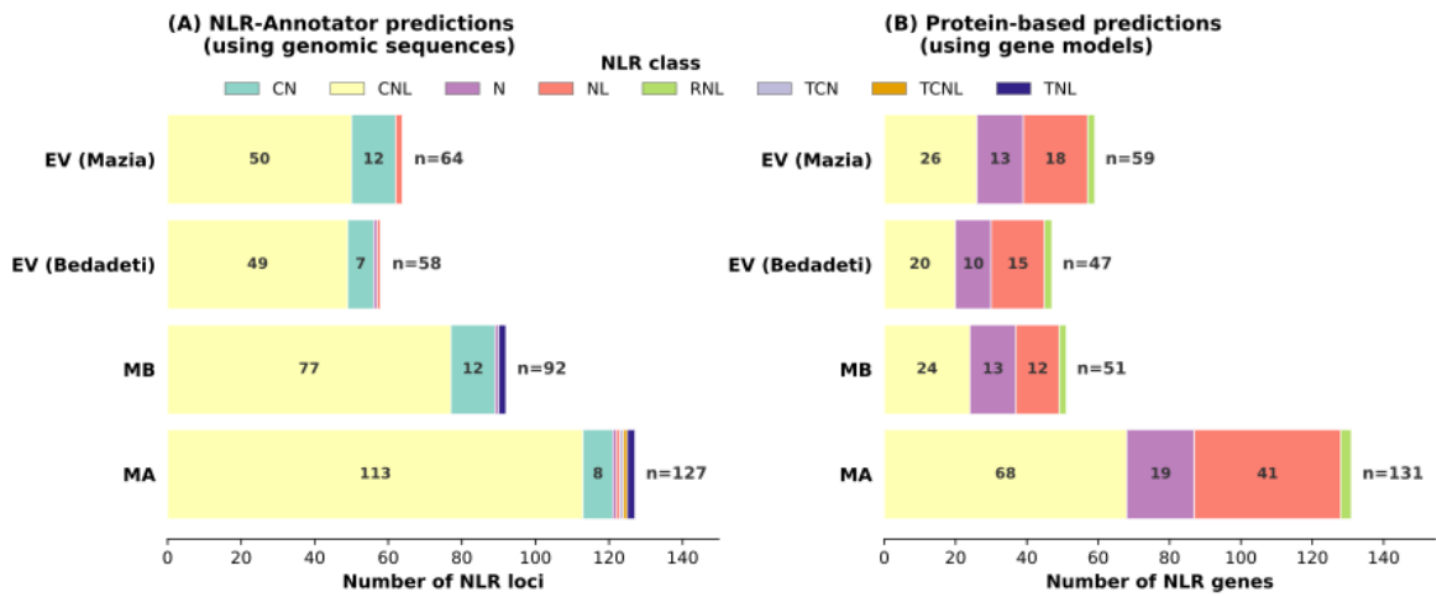


Figure 9

Nucleotide binding domain leucine rich repeat (NLR) immune receptor gene classes in enset landraces and *Musa* species. (A) NLR loci predicted by NLR-Annotator [105] from genomic sequences in *E. ventricosum* (EV) landraces, and *M. acuminata* (MA) and *M. balbisiana* (MB). (B) NLR-encoding genes predicted from annotated protein coding genes in genomes of EV landraces, MA and MB. Numbers within bars indicate counts (>5) per NLR class; total counts shown at bar ends. NLR class abbreviations: CN, CC-NB-ARC (truncated, lacking LRR); CNL, CC-NB-ARC-LRR (complete CC-type NLRs); N, NB-ARC only; NL, NB-ARC-LRR (lacking N-terminal domain—note that CC domain detection varies by CNL subclass (36-76%) [156], so some NL-classified genes may represent CNLs with undetectable CC domains); RNL, RPW8-NB-ARC-LRR; TNL, TIR-NB-ARC-LRR; TCNL, TIR-CC-NB-ARC-LRR; TCN, TIR-CC-NB-ARC. Domain abbreviations: CC, coiled-coil; NB-ARC, Nucleotide Binding site shared with APAF-1, R proteins, and CED-4; LRR, leucine-rich repeat; RPW8, Resistance to Powdery Mildew 8-type domain; TIR, Toll/interleukin-1 receptor. See Supplementary Figure 7 for NLR domain architecture schematic representation.

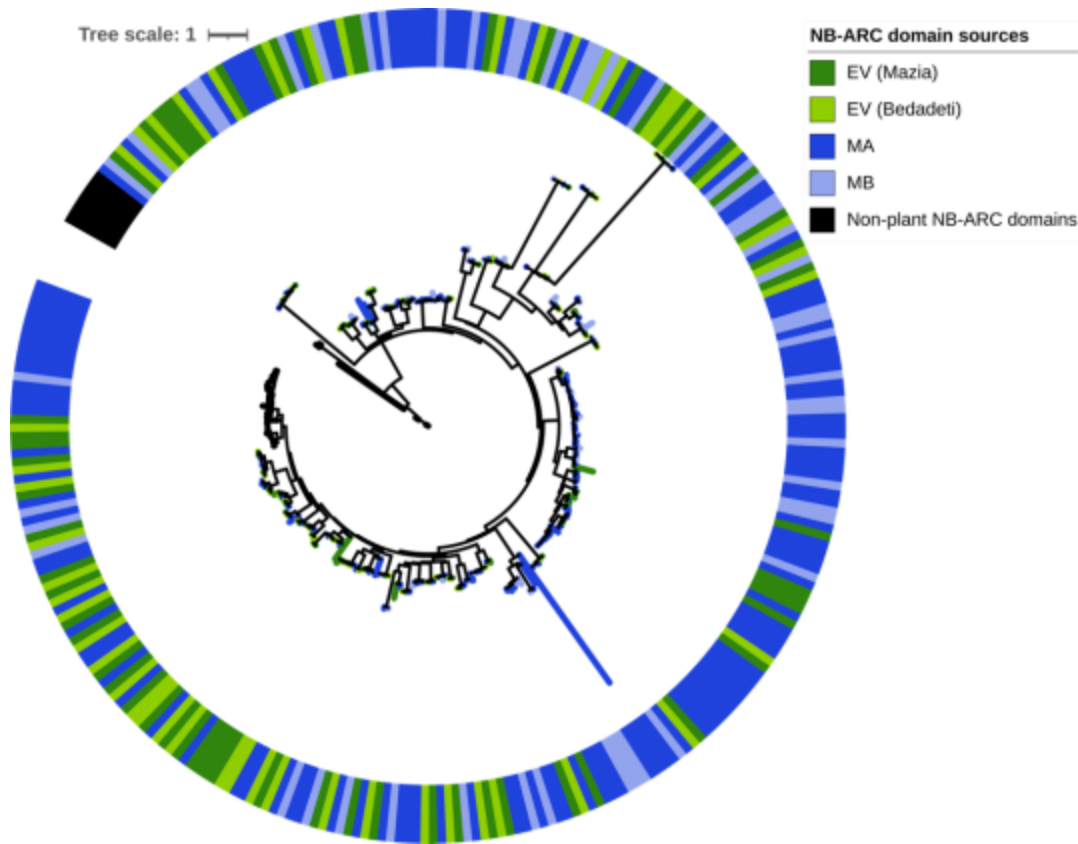


Figure 10

Phylogenetic tree of NLRs from *E. ventricosum* (EV) landraces Mazia and Bedadeti, *M. acuminata* (MA) and *M. balbisiana* (MB). The outer encircled coloured ring represents source of input NB-ARC domain from EV landrace Mazia and Bedadeti, and *Musa* species (MA and MB) as indicated in the legend label. The tree was generated using RAxML (v8.2.12) [113] applying the Maximum Likelihood method based on the JTT model and bootstrap value of 1000 replicates. For outgroups rooting, non-plant NB-ARC containing domains from RefPlantNLR [108] were included and the tree was annotated using iTOL [114].

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarytables.xlsx](#)
- [Supplementaryfigures.docx](#)