The Banana Genome Hub: a community database for genomics in the Musaceae

Gaëtan Droc[1,2,3*], Guillaume Martin[1,2,3], Valentin Guignon[3,4], Marilyne Summo[1,2,3], Guilhem Sempéré[3,5,6], Eloi Durant[3,7,8], Alexandre Soriano [1,2,3], Franc-Christophe Baurens[1,2], Alberto Cenci[3,4], Catherine Breton[3,4], Trushar Shah[9], Jean-Marc Aury[10], Xue-Jun Ge[11,12], Pat Heslop Harrison[11,13], Nabila Yahiaoui[1,2], Angélique D'Hont[1,2], Mathieu Rouard[3,4*]

[1] CIRAD, UMR AGAP Institut, F-34398 Montpellier, France

[2] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, , F-34398 Montpellier, France

[3] French Institute of Bioinformatics (IFB) - South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, F-34398 Montpellier France

[4] Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier, France

[5] CIRAD, UMR INTERTRYP, F-34398 Montpellier, France

[6] INTERTRYP, Université de Montpellier, CIRAD, IRD, 34398 Montpellier, France

[7] Syngenta Seeds SAS, Saint-Sauveur, 31790, France

[8] DIADE, Univ Montpellier, CIRAD, IRD, Montpellier, 34830, France

[9] IITA, Nairobi P.O. Box 30709-00100, Kenya

[10] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 2 rue Gaston Crémieux, 91057 Evry, France

[11] Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China, 510520.

[12] Center of Conservation Biology, Core Botanical Gardens, Chinese Academy of Sciences, Guangzhou, China, 510520

[13] Department of Genetics and Genome Biology, University of Leicester, Leicester LE1 7RH, UK

E-mail addresses: Gaëtan Droc (gaetan.droc@cirad.fr), Guillaume Martin (guillaume.martin@cirad.fr), Valentin Guignon (v.guignon@cgiar.org), Marilyne Summo (marilyne.summo), Guilhem Sempere (guilhem.sempere@cirad.fr), Eloi Durant (eloi.durant@syngenta.com), Alexandre Soriano (alexandre.soriano@cirad.fr), Franc-Christophe Baurens (franc-christophe.baurens@cirad.fr), Alberto Cenci (a.cenci@cgiar.org), Catherine Breton (c.breton@cgiar.org), Trushar Shah (t.shah@cgiar.org), Jean-Marc Aury (jmaury@genoscope.cns.fr), Xue-Jun Ge (xjge@scib.ac.cn), Pat Heslop Harrison (phh4@leicester.ac.uk), Nabila Yahiaoui (nabila.yahiaoui@cirad.fr), Angélique D'Hont (dhont@cirad.fr), Mathieu Rouard (m.rouard@cgiar.org)

***Corresponding authors:** E-mail: m.rouard@cgiar.org Tel./Fax: 33 (0) 467 612 908 / +33 (0) 467 610 334 and E-mail: gaetan.droc@cirad.fr Tel./Fax: +33 (0)467 614 912 / +33 (0) 467 615 605

Running title: The Banana Genome Hub

Abstract

The Banana Genome Hub provides centralized access for genome assemblies, annotations, and the extensive related omics resources available for bananas and banana relatives. A series of tools and unique interfaces are implemented to harness the potential of genomics in bananas, leveraging the power of comparative analysis, while recognizing the differences between datasets. Besides effective genomic tools like BLAST and the JBrowse genome browser, additional interfaces enable advanced gene search and gene family analyses including multiple alignments and phylogenies. A synteny viewer enables the comparison of genome structures between chromosome-scale assemblies. Interfaces for differential expression analyses, metabolic pathways and GO enrichment were also added. A catalogue of variants spanning the banana diversity is made available for exploration, filtering, and export to a wide variety of software. Furthermore, we implemented new ways to graphically explore gene presence-absence in pangenomes as well as genome ancestry mosaics for cultivated bananas. Besides, to guide the community in future sequencing efforts, we provide recommendations for nomenclature of locus tags and a curated list of public genomic resources (assemblies, resequencing, high density genotyping) and upcoming resources—planned, ongoing or not yet public. The Banana Genome Hub aims at supporting the banana scientific community for basic, translational, and applied research and can be accessed at https://banana-genome-hub.southgreen.fr

Keywords: Banana, database, genome assemblies, pangenome, SNP, visualization, gene annotations, biodiversity, genomes, plantain, evolution

Introduction

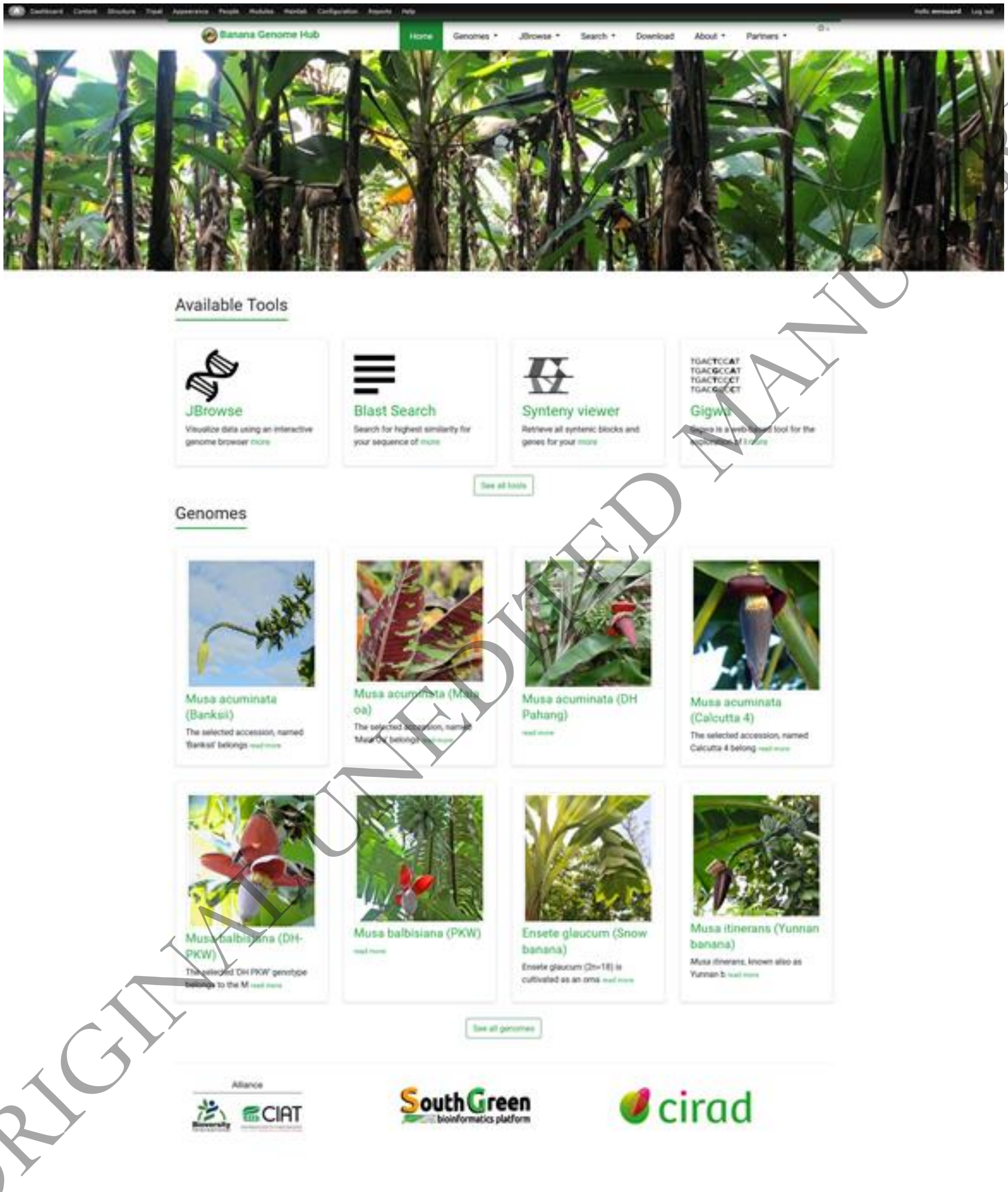
The *Musaceae,* known as the banana family, belongs to the monocotyledons, that comprise crops of great economic value as well as ornamental plants. Notably, *Musaceae* includes the genus *Musa* with bananas, a top-ten crop for food security, and arguably the favorite fruit worldwide [1]. Its sister genus, Ensete, contains *Ensete ventricosum*, an important crop for food security in Ethiopia [2] and ornamental plants like *Ensete glaucum* widely distributed in Asia. The final monospecific genus in *Musaceae* includes *Musella lasiocarpa* from southwest China and possibly extinct in the wild. Wild species within *Musaceae* are diploids, with basic chromosome numbers of x=9, 10 and 11. The *Musa* cultivars grown for fruit result from hybridization between different wild diploid *Musa* species and subspecies. They are parthenocarpic, sterile or poorly fertile and mostly cultivated as vegetatively propagated triploids (2n=3x=33) although some cultivars are diploids or tetraploids, most of cultivars bear large structural variations in their chromosomes, transmitted from different wild ancestors. All these features make banana breeding very complex. Genomic characterization has a great potential to significantly contribute to better conservation strategies, improved use of banana genetic resources and increased sustainability of crop production [4]. Increasing the availability of genomic resources and facilitating their use has been much needed [5,6].

In 2012, the first Musaceae reference genome, representative of *Musa acuminata* (A genome), was published [7] alongside the Banana Genome Hub [8] (https://banana-genome-hub.southgreen.fr). In the last decade, this reference was iteratively improved [9,10] while a number of new genome assemblies of different *Musaceae* species have also been generated. The next sequenced genome was that of *Musa balbisiana* (B genome) [11], first as a draft genome and later as a chromosome-scale assembly from a double haploid [12]. In the meantime, draft assemblies of *Musa itinerans* [13], *Ensete ventricosum* [14], *Musa textilis* [15] and other subspecies of *Musa acuminata* were produced [16]. A pangenome composed of the 15 individuals belonging to Ensete and Musa was also developed [17]. Benefiting from easier and cheaper access to long reads sequencing technologies and scaffolding methods, chromosome scale genome assemblies were released for *Musa schizocarpa* [18], *Ensete glaucum* [19] and a telomere-to-telomere assembly of *Musa acuminata* was published [10]. Thanks to available reference genomes, a broad range of studies have been conducted to explore multiple aspects including genetic diversity [20], plant genome evolution [21–23], chromosome structural variation [24], gene family analyses [25–28], trait-phenotype [29,30], association genetics [31–33] and genetic engineering [34]. All these topics need access to various types of datasets and related query or visualisation interfaces.

Here, we present an overhauled and enriched version of the Banana Genome Hub (BGH), a community database that serves as a central online platform for whole genome sequences and related omics data on *Musaceae*. We detail the implemented interfaces, and the way data were collected and curated. Finally, we list and discuss the status of sequencing projects and propose a locus name nomenclature for future projects about the genomics of *Musaceae*.

Tools and interfaces

We implemented a list of web interface and collected data to facilitate functional and comparative genomics-oriented data analyses (**Figure 1**). Some interfaces focus on exploration of individual genes or of a list of genes to check their location on the genome, presence in gene families, their expression patterns, their functional annotations (i.e. Gene Ontologoy (GO)) as well as associated SNP markers. Other tools enable a more global exploration of chromosome structures by looking at synteny, presence absence variation and genome ancestry mosaics. From a technical perspective, the BGH core has been developed with the Tripal toolkit (i.e. Drupal v7, Tripal v3), an open-source project supporting the development of biological databases [8,35,36] complemented by the development of additional modules [37]. All these elements are further described below.



Visualize data using an interactive genome browser more

Search for highest similarity for your sequence of more

Retrieve all syntenic blocks and genes for your more

Gigwa is a web-based tool for the exploration of l more

The selected accession, named 'Banksii' belongs read more

The selected accession, named 'Maia Oa' belongs read more

The selected accession, named Calcutta 4 belong read more

The selected 'DH PKW' genotype belongs to the M read more

Ensete glaucum (2n=18) is cultivated as an orna read more

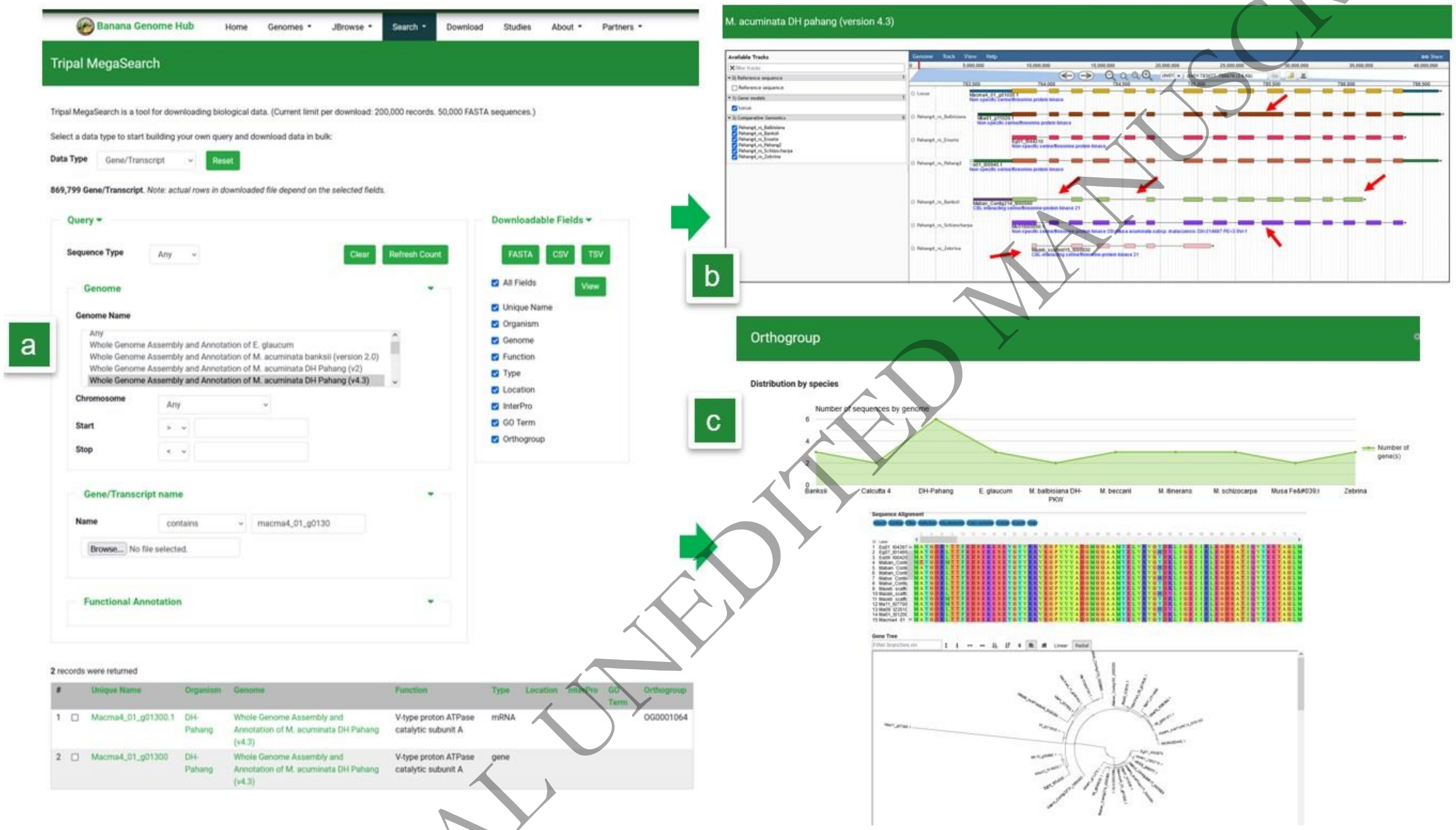
Musa itinerans, known also as Yunnan b read more

Figure 1. Screenshot of the Banana Genome Hub homepage showing a subset of available genome sequence and visualisation and analytical tools.

Gene(s) query including orthogroups and omics-related datasets

Users have multiple ways to search for genes in the system, either using a gene locus (or a list of them), keywords, genomic coordinates powered by MegaSearch [38] or using the BLAST graphical interface searches from Sequenceserver [39] (**Figure 2a**). Results are connected to genome browsers [37] specific to each genome. Comparisons between genomes are facilitated by tracks showing gene annotations projected on other genomes using the lift-over tool. It allows at a glance to see missing genes and investigate possible errors in the prediction of structural gene annotation [40] (**Figure 2b**).

Any gene search result lists several information including gene membership to orthogroups or gene families in *Musaceae*. The three versions corresponding to the *M. acuminata* reference genome ('DH Pahang' v1, v2 and v4) were conserved in the system for traceability. To enable orthogroup visualization, we developed extension modules that support visualisation of multiple genome alignment and phylogenetic tree with all functionalities provided by MSAviewer [41] and PhyloTree [42] respectively (**Figure 2c**)



Banana Genome Hub | Home | Genomes ▾ | JBrowse ▾ | Search ▾ | Download | Studies | About ▾ | Partners ▾

Tripal MegaSearch

Tripal MegaSearch is a tool for downloading biological data. (Current limit per download: 200,000 records. 50,000 FASTA sequences.)

Select a data type to start building your own query and download data in bulk:

Data Type: Gene/Transcript | Reset

869,799 Gene/Transcript. *Note: actual rows in downloaded file depend on the selected fields.*

Query ▾

Sequence Type: Any | Clear | Refresh Count

Genome

Genome Name:
Any
Whole Genome Assembly and Annotation of E. glaucum
Whole Genome Assembly and Annotation of M. acuminata banksii (version 2.0)
Whole Genome Assembly and Annotation of M. acuminata DH Pahang (v2)
Whole Genome Assembly and Annotation of M. acuminata DH Pahang (v4.3)

Chromosome: Any

Start: >

Stop: <

Gene/Transcript name

Name: contains | macma4_01_g0130

Browse... No file selected.

Functional Annotation

Downloadable Fields ▾

FASTA | CSV | TSV

☑ All Fields | View
☑ Unique Name
☑ Organism
☑ Genome
☑ Function
☑ Type
☑ Location
☑ InterPro
☑ GO Term
☑ Orthogroup

2 records were returned

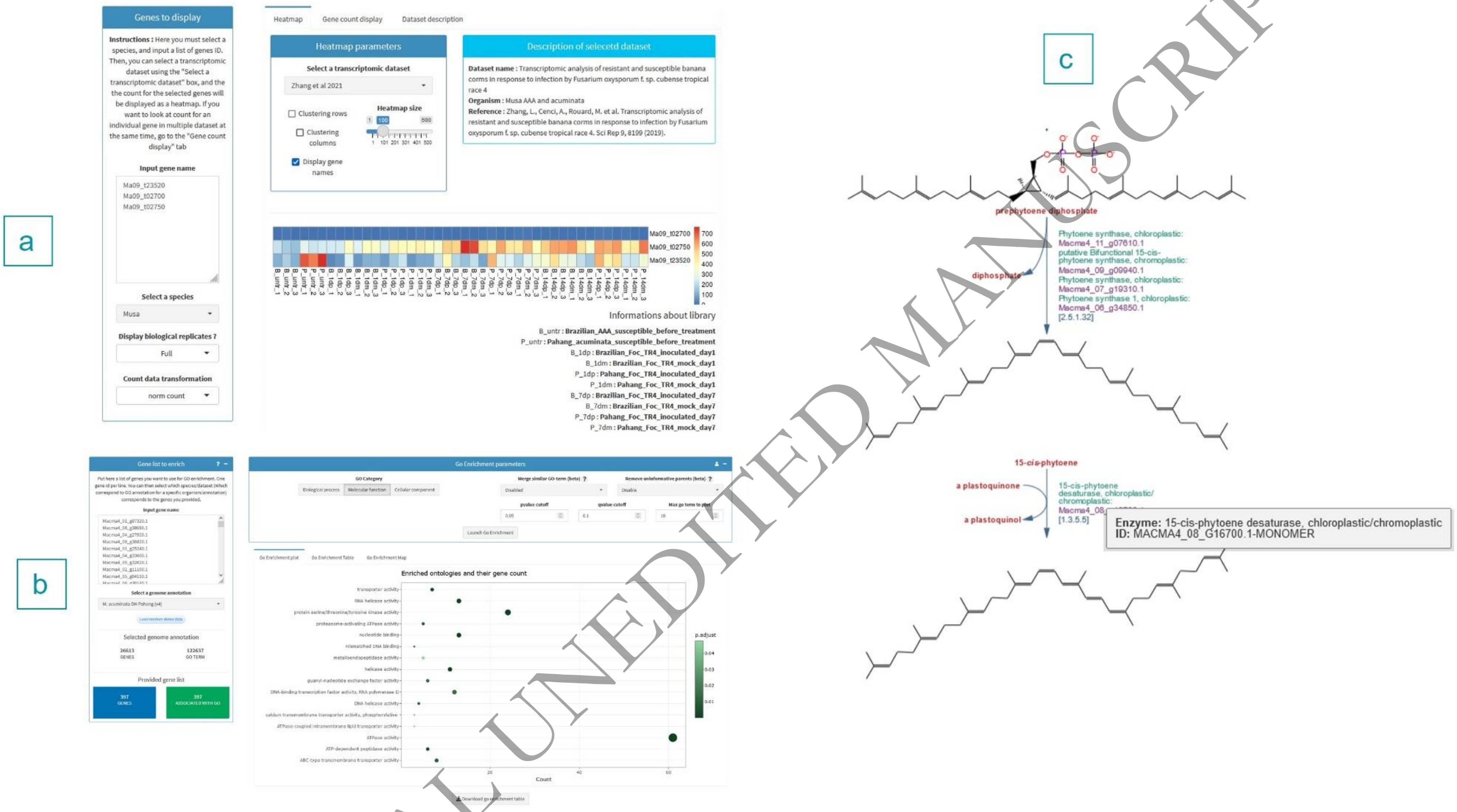
#		Unique Name	Organism	Genome	Function	Type	Location	InterPro	GO Term	Orthogroup
1	□	Macma4_01_g01300.1	DH-Pahang	Whole Genome Assembly and Annotation of M. acuminata DH Pahang (v4.3)	V-type proton ATPase catalytic subunit A	mRNA				OG0001064
2	□	Macma4_01_g01300	DH-Pahang	Whole Genome Assembly and Annotation of M. acuminata DH Pahang (v4.3)	V-type proton ATPase catalytic subunit A	gene				

Figure 2. (A) Gene search interface enabling access results hits that can be visualized in (B) genome browser (JBrowse) with Liftoff tracks. Red arrows indicate region that are inconsistent between gene prediction and that might need curation and (C) in an orthogroup context with associated multiple alignments and phylogenetic tree

For users interested in gene expression patterns for specific gene(s), we built interactive interfaces based on the shiny apps technology (R package) to enable manipulation of data results from published studies [29,43,44]. For instance, it is possible to search for genes annotated as RGA2, a putative nucleotide-binding and leucine-rich repeat (NB-LRR)-type resistance (R) gene known to be involved in the resistance to Fusarium wilt when overexpressed [45], and to check their level of expression in a study linked to Fusarium wilt [29] (**Figure 3a**).

Also, additional datasets can be uploaded in the Diane suite [46] to perform differential gene expression analyses, expression-based clustering and gene regulatory network analyses in which *Musa* references genomes were added. Besides, when a list of genes is identified, users can quickly test in a few clicks for Gene Ontology enrichment for several genomes and without the need to extract functional annotations and use external software (**Figure 3b**).

With regards to other OMICS, there have been increasing numbers of proteomics and metabolomics experiments in banana [30,47–50]. To complement these resources and enable various options like experimental data overlay on metabolic pathways, we set up the latest version of PathwayTools v25 [51], named MusaCyc, that comprises a comprehensive set of interfaces to cover user needs. For instance, the carotenoid pathway has been actively studied in banana [52–54] and the Phytoene desaturase (PDS) enzyme, that can cause albinism when disrupted, was used as a proof of concept for gene editing. Using MusaCyc, the PDS gene can be easily found (**Figure 3c**).



Genes to display

Instructions : Here you must select a species, and input a list of genes ID. Then, you can select a transcriptomic dataset using the "Select a transcriptomic dataset" box, and the the count for the selected genes will be displayed as a heatmap. If you want to look at count for an individual gene in multiple dataset at the same time, go to the "Gene count display" tab

Input gene name: Ma09_t23520, Ma09_t02700, Ma09_t02750

Select a species: Musa

Display biological replicates ?: Full

Count data transformation: norm count

Heatmap parameters

Select a transcriptomic dataset: Zhang et al 2021

☐ Clustering rows

☐ Clustering columns

☑ Display gene names

Heatmap size: 100

Description of selecetd dataset

Dataset name : Transcriptomic analysis of resistant and susceptible banana corms in response to infection by Fusarium oxysporum f. sp. cubense tropical race 4

Organism : Musa AAA and acuminata

Reference : Zhang, L., Cenci, A., Rouard, M. et al. Transcriptomic analysis of resistant and susceptible banana corms in response to infection by Fusarium oxysporum f. sp. cubense tropical race 4. Sci Rep 9, 8199 (2019).

Informations about library

B_untr : **Brazilian_AAA_susceptible_before_treatment**
P_untr : **Pahang_acuminata_susceptible_before_treatment**
B_1dp : **Brazilian_Foc_TR4_inoculated_day1**
B_1dm : **Brazilian_Foc_TR4_mock_day1**
P_1dp : **Pahang_Foc_TR4_inoculated_day1**
P_1dm : **Pahang_Foc_TR4_mock_day1**
B_7dp : **Brazilian_Foc_TR4_inoculated_day7**
B_7dm : **Brazilian_Foc_TR4_mock_day7**
P_7dp : **Pahang_Foc_TR4_inoculated_day7**
P_7dm : **Pahang_Foc_TR4_mock_day7**

Gene list to enrich

Put here a list of genes you want to use for GO enrichment. One gene id per line. You can then select which species/dataset (which correspond to GO annotation for a specific organism/annotation) corresponds to the genes you provided.

Input gene name: Macma4_01_g07320.1, Macma4_06_g38080.1, Macma4_04_g27920.1, Macma4_08_g30820.1, Macma4_01_g25240.1, Macma4_04_g33600.1, Macma4_05_g32020.1, Macma4_02_g11100.1, Macma4_05_g04110.1

Select a genome annotation: M. acuminata DH Pahang (v4)

Selected genome annotation: 26613 GENES, 122637 GO TERM

Provided gene list: 397 GENES, 397 ASSOCIATED WITH GO

Go Enrichment parameters

GO Category: Molecular function

Merge similar GO-term (beta): Disabled

Remove uninformative parents (beta): Disable

pvalue cutoff: 0.05

qvalue cutoff: 0.1

Max go term to plot: 16

Enzyme: 15-cis-phytoene desaturase, chloroplastic/chromoplastic
ID: MACMA4_08_G16700.1-MONOMER

Figure 3. (A) Transcriptomic interface with a list of RGA2 genes from *M. acuminata* ‘DH Pahang’ submitted to visualize their level of expression for a study on Fusarium wilt. (B) GO enrichment interface with a list of genes submitted. (C) First steps of the carotenoid pathways with Phytoene desaturase (PDS) identified by MusaCyc in the *Musa acuminata* genome

Genetic variant search and usage

This section, powered by the GIGWA tool [55,56], gives access to a range of studies related to genetic diversity [57], GWAS [31,33], Genomic selection or chromosome structure exploration [58,59]. Notably, available studies include SNPs of the diploid banana panel that was designed specifically for GWAS analyses [31] while corresponding plant material for this panel can be ordered for phenotyping at the International Transit Center (ITC) via the *Musa* Germplasm Information System (MGIS) website [60,61]. After filtering with advanced functionality, the datasets can be exported in multiple formats for subsequent analyses such as genetic diversity studies or directly visualized in JBrowse, IGV, Flapjack (and flapjack-bytes) (**Figure 4**). In addition, this catalogue of variants is compliant with BrAPI v1 & v2 [62] and can be accessed programmatically and used in third party client or databases.

Pangenome viewer and exploration

A single reference genome is not enough to capture genetic diversity in a species or a genus [63,64]. To capture the diversity of gene content across *Musaceae*, a draft cross genus (*Musa-Ensete*) pangenome was built. It revealed distinct presence/absence patterns between genera [17]. While global results were analysed, exploration of specific regions along pan-chromosomes is still to be done. To make this easier, we implemented an instance of the Panache software [65] which enables the exploration of gene presence/absence variations (PAV) within pan-chromosomes. With it, users can automatically search for PAV areas and visualize them in the interface, where each line corresponds to one of the re-sequenced individuals (**Figure 5a**). Multiple sorting options (taxonomy, presence or absence of a given gene, etc.) are proposed to guide users toward genomic regions rich in PAV or showing a particular pattern.

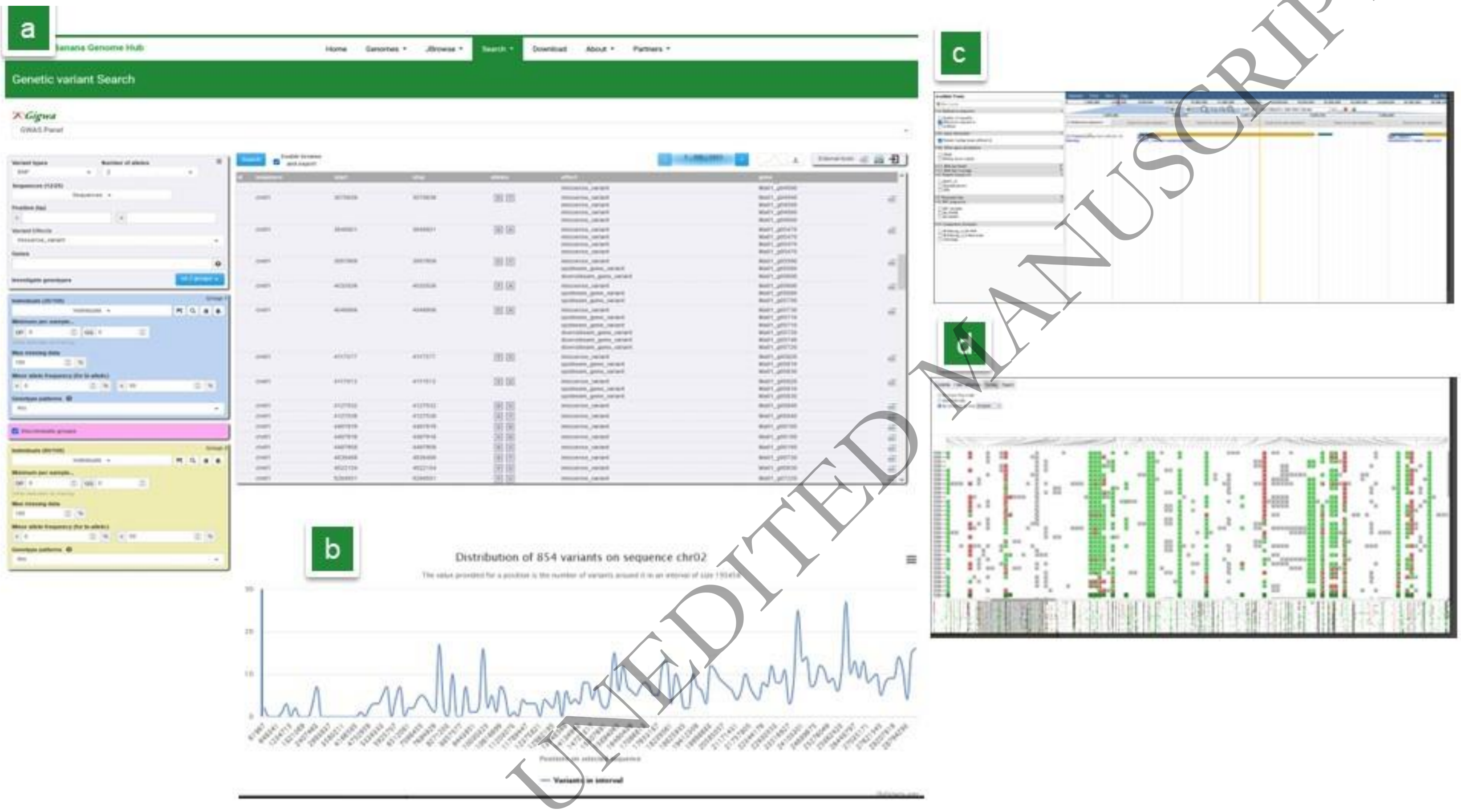


Figure 4. Overview of the genetic variant interface powered by GIGWA (A) Main interface for the GWAS panel with discriminated variants between 2 groups (seeded vs non-seeded) (B) Statistics of SNPs along Chromosome 2. (C) SNP visualization in JBrowse from the GIGWA interface (D) Data export online for graphical previews of genotype data in Flapjack-bytes.

Genome ancestry mosaics viewer

Cultivated bananas result from a relatively limited number of sexual events with inter(sub)specific hybridizations and recombination [67]. The different ancestral contributions can be represented as genomic segments of distinct origin along the chromosomes. To provide access to recent studies that reported recombination between A and B genomes [59] and genome ancestry mosaics for a panel of diploid and triploid bananas [66], we embedded a new tool, called GeMo [67]. By selecting an samples like ‘Grande Naine’ (AAA), an autotriploid cultivar belonging to the Cavendish subgroup, users can immediately spot the ancestral contributors of the *M. acuminata* subspecies, predominantly ‘banksii’, ‘zebrina’, ‘malaccensis’ (**Figure 5b**). This viewer is intended to become a registry for any future studies performing in silico chromosome painting on Musaceae individuals but also enable user to manipulate their own data in a non-persistent way.

Synteny viewer

The Zingiberales order evolution was shaped by lineage specific ancient whole genome duplications [7,22] and within the *Musaceae*, for which the crown age was estimated at 59.19 Ma [68], a large number of chromosome rearrangements occurred [24,69]. As an example, *M. acuminata* and *M. balbisiana* differ by a large translocation on chromosome1/3 and a large inversion on chromosome 5 [12]. To explore the chromosome structure between genome assemblies, SynVisio [70] was implemented for syntenic block visualization. It enables the comparison of two or more genomes (**Figure 5c**) and supports multi-resolution analysis and interactive filtering. Users can compare genomes one to one or in multi-genome mode. Conveniently, it also allows downloading high-quality images. Such a tool will be increasingly relevant as new assemblies are produced to visualize and understand fusion and fission events between chromosomes in *Musaceae* where different basic chromosome numbers exist (from 7 to 11 haploid chromosomes).

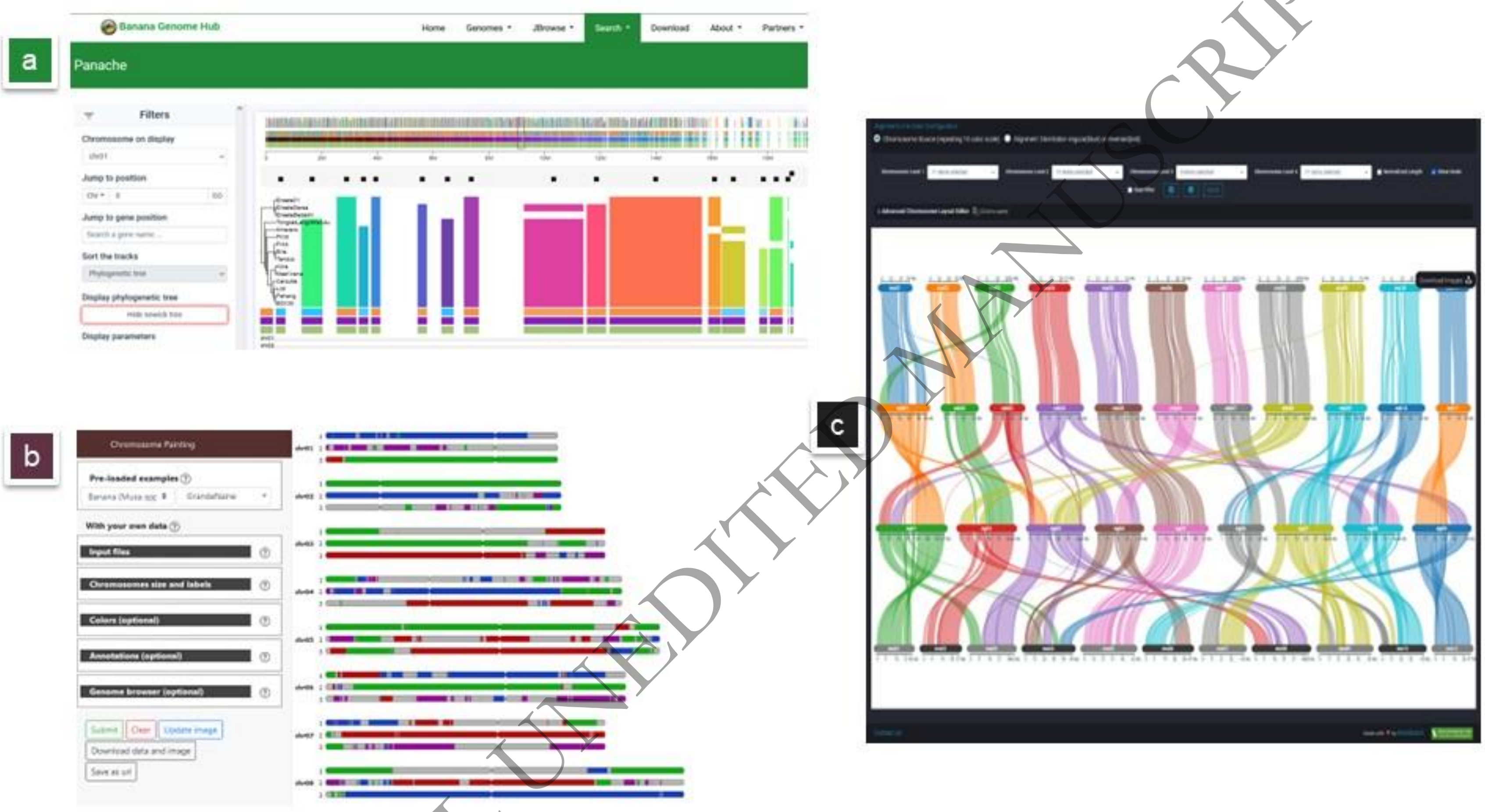


Figure 5. Overview of proposed web interfaces for comparative genomics within *Musaceae*. (A) Overview of the *Musaceae* Pangenome represented with the Panache interface. (B) Examples of genome ancestry mosaics. (C) Synteny between *Ensete glaucum*, *Musa acuminata*, *Musa balbisiana* and *M. schizocarpa* using Synvisio.

Database construction and content

Collection of genome assemblies and gene annotation

We collected 16 publicly released *Musaceae* nuclear genome sequences (8 high-quality and 8 draft sequences) that were released publicly (**Table 1**) as well as 91 chloroplast assemblies [68,71–75]. Functional annotations from InterPro were obtained using InterProScan [76]. Gene ontology (GO) were retrieved by combining results from interpro2go and BlastP on SwissProt and TrEMBL[77]. For each assembly, they were compared and mapped using Liftoff [40]. When available, TE annotations from published studies were inserted into JBrowse.

Species	**Genotype**	**Version**	**Technology**	**Status**	**Comments**	**References**
Musa acuminata	DH Pahang	1	Sanger + Illumina SR	High quality draft	1st reference (A genome)	[7]
Musa acuminata	DH Pahang	2	Illumina SR + optical map	Improved high quality draft		[9]
Musa acuminata	DH Pahang	4	Nanopore LR + Illumina	Telomere to telomere	Final version	[10]
Musa acuminata	Banksii	2	Illumina + PacBio LR	Draft	CS in progress	[16]
Musa acuminata	Maia Oa	1	Illumina SR	Draft	CS in progress	[16]
Musa acuminata	Calcutta 4	1	Illumina SR	Draft	CS in progress	[16]
Musa balbisiana	PKW	1	Illumina SR	Draft		[11]
Musa balbisiana	DH PKW	1.1	Illumina SR, PacBio LR + Hi-C	Chromosome scale	B genome reference	[12]
Musa itinerans	-	1	Illumina SR	Draft		[13]
Musa schizocarpa	-	1	Nanopore LR + Bionano	Chromosome scale	S genome	[18]
Ensete	-	1	Nanopore LR +	Chromosome		[19]

glaucum			Hi-C	scale		
Ensete ventricosum	Bedadeti	3	Illumina SR	Draft (download only)		[14,78]
Musa textilis (abaca)	abuad	-	Illumina SR PacBio LR	Draft (download only)	CS in progress	[15]
Musa acuminata	Dwarf Cavendish	1	Illumina SR	Draft (download only)		[79]
Musa troglodytarum	Karat	1	Nanopore LR + Illumina SR + PacBio LR + Hi-C	Chromosome scale		[80]
Musa beccarii		1	Nanopore LR + Hi-C	Chromosome scale		Early advance

Table 1. List of genome sequence assemblies accessible via Banana Genome Hub. (CS: chromosome scale; SR: short reads; LR: long reads)

Only minimal modifications of the assemblies or annotations from their description in publications are intended, to facilitate comparisons and traceability. In some cases, however, we improved the gene annotation: in agreement with data providers, we filtered *Musa balbisiana* PKW for TE and released a new annotation; we also released a new annotation for *Musa balbisiana* ‘DH PKW’ where we reversed some chromosomes to be consistent with the orientation in *Musa acuminata* ‘DH Pahang’ and *Musa schizocarpa*.

Transcriptomics and pathway related datasets

Transcriptomics data supplied by the community were included [12,43,44,79,81]. RNAseq data were mapped using STAR [82] and added in JBrowse as mapped tracks and in the download section. Whenever possible, derived reads count from published transcriptomics studies were collected and connected to the transcriptomics interface [29,43,44]. For pathway related information, enzymes and metabolic pathways were predicted from the protein-coding genes of *Musa acuminata* ‘DH Pahang’ v4. Enzyme Classification (EC) numbers were predicted combining both tools PRIAM [83] and BlastKOALA [84]. As a result, data were inferred for 774 pathways, 6,762 enzymatic reactions and 97 transport reactions. A total of 8,220 enzymes have been annotated and are available in the pathway tools section of the BGH.

Comparative genomic analysis

We identified syntenic genes in the five chromosome scale assemblies available for *Musaceae*. Protein-coding genes were processed to identify reciprocal best hits (RBH) with BLASTP (e-value 1e-10) followed by MCScanX (e-value 1e-5, max gaps 25) [85].

Gene family identification

Protein-coding genes from *E. glaucum* v1, *M. acuminata* ('DH Pahang' v2, Zebrina 'Maia Oa', 'Calcutta 4' and 'Banksii'), *M. balbisiana* v1.1 and *M. schizocarpa* v1 were processed using OrthoFinder v2.5.2 [86] with default parameters. We built the alignments and gene trees by applying our phylogenomic workflow, as implemented in GreenPhylDB [87].

Genetic variants

SNP markers from multiple studies were retrieved and inserted into the GIGWA v2 genotyping database [55]. Quality checks, read mapping on reference genomes, SNP calling and variant effect in genic regions were conducted as described in [1]. The outputs of the analyses were produced in the variant call format (VCF), then loaded in GIGWA with associated metadata [55].

Pangenome

Pangenome assembly, gene annotation and PAV matrix were collected from [17]. The study was based on 15 accessions across *Musa* and *Ensete* sequenced with short read technologies. To define the presence-absence of genes in the different accessions, they assembled the pangenome iteratively and annotated the genes in the new contigs, then proceeded with read mapping.

Genome and transcriptome sequencing status

The curated list of SRA genomic resources was searched on NCBI SRA [88] by filtering on Taxonomic ids for *Musa* and *Ensete* and metadata was extracted from BioSample metadata descriptions. Information on ongoing projects was obtained by personal communications and interactions within the scientific community.

Discussion and perspectives

The Banana Genome Hub is a comprehensive platform dedicated to the genomics of a specific plant family – the *Musaceae* - as it has been developed for other families such as the Rosaceae [89] or the Juglandaceae [90]. The core functionalities are similar by providing access to genome datasets via JBrowse [91], BLAST, synteny and gene families viewers. However, the BGH has some specificities taking into account the nature of the plant and the existing ecosystems of tools and databases in the community.

An innovative pangenomics-related interface, Panache [65], has been implemented to support exploration of presence-absence variation (PAV). Both provides possible valuable resources for the design and exploration of precision genetics studies being conducted in the genus *Musa* [52,92]. Besides, as a vegetatively propagated plant with low fertility, unravelling the genome ancestry mosaics of cultivated bananas has been initiated to decipher it complex domestication history [66] and we provide a unique way to store and visualize, through GeMo, future work in that direction. For functional oriented studies, users have now access to handy interface to check gene expression and functional enrichment.

Furthermore, the BGH intends to complement other databases on bananas and contribute to a better conservation and use of *Ensete* and *Musa* genetic resources. Contrary to the other portal [89,90], the BGH does not intend to develop its own breeding module but rather proposes to implement BrAPI standards [62] to increase interoperability with the Banana instance of Breedbase [93]; which has been specifically designed for this purpose and that is actively supported by some banana breeding programs. Like GDR [89], a catalogue of variants is curated to provide facilitated access to data for SNP-based published studies. This catalogue, maintained by a different system, is shared with the Musa Germplasm Information System (MGIS) [60] to connect with the existing diversity of genetic resources conserved and documented in genebanks.

While the *Musaceae* family contains 80 species classified in three genera, the Banana Genome Hub includes all publicly available whole genomes for eight species from two genera. Therefore, the BGH is designed to hold more whole genomes, and still has high potential to grow and to propose new tools to efficiently exploit new datasets considering specificities of the crop (e.g., polyploidy, structural variations). We will continue to curate and add new genome assemblies and related OMICS data as they become publicly available. Given the level of structural variation including chromosome rearrangements that are now well documented between the six species, high quality (N50 nearing average chromosome length) genome sequences (currently supported by Hi-C and/or long-molecule sequencing and genetic mapping data) are required as references.

To guide sampling for future sequencing projects and in an attempt to manage redundancy in data generation, we compile information from public sources or gleaned in conferences or from personal communications that will be regularly updated online (https://banana-genome-hub.southgreen.fr/content/sequencing-status). The first observation is that if no genome assembly of known *Musa* cultivars, mostly triploids, has been released at chromosome-scale, some are underway as well as for additional wild species. Increasing accuracy of long-molecule sequencing is important to assembling haplotypes in triploid hybrids that are so important regionally and in trade. High quality whole-genome assemblies underpin exploitation of survey sequence data for allele mining or GWAS (Genome Wide Association Studies) to identify functional variants. Re-sequencing is ongoing in several germplasm collections, which will help identifying allelic and potentially copy number variation. Also, assemblies are available for chloroplast genomes on wild species, sometimes redundantly, and future effort might focus on cultivated groups and systematically cover the diversity of the family.

Whenever possible, plant material used to generate genomic data should be deposited in genebanks or national collections (**Table 2**) where passport data, possibly associated with phenotype information, is documented and material distribution processes are streamlined. For instance, use of accessions from the International Transit Center (ITC) [60,61] or the CRB Plantes Tropicales Antilles CIRAD-INRAe can facilitate traceability, reproducibility, and data integration with previous and future experiments since accessions can be sent internationally, virus indexed and free of charge for research purposes. Furthermore, missing accessions of interest can be also proposed to ITC for conservation.

Collection name	**Country**	**# Available Accessions**	**Distribution**	**Conditions**	**Access**
International Transit Center (ITC)	Belgium	990	International	Free of charges (SMTA)	https://www.crop-diversity.org/mgis/moos/how-to-order
CRB Plantes Tropicales Antilles CIRAD-INRAe (CRB-PT)	Guadeloupe, France	381	International	Free except transport (SMTA)	http://crb-tropicaux.com/Portail
International Institute of Tropical Agriculture (IITA)	Nigeria	275	Regional (Africa)	Free of charges (SMTA)	https://www.genesys-pgr.org

Table 2. Examples of genebanks or germplasm collection where material can be requested for research purposes.

Regarding gene annotation, we recommend adopting a defined nomenclature for locus tag that would consider the wide range of wild *Musaceae* species (**Table S1**). However, we acknowledge that further work is necessary to address the case of groups and subgroups in cultivated bananas.

Finally, we encourage scientists generating genomics data in Musaceae to contact us or the Genomics Thematic group of MusaNet (https://musanet.org) early in the publication process to make sure that general standards (chromosome orientation, gene locus) are consistent with existing resources and eventually to get support to create dedicated pages and associated tools (BLAST, JBrowse, download).

Data availability statement

For data download, the BGH is structured by organism with regards to individual genome assemblies and also by studies that provide directory listing of the related datasets. A global download section, supported by Drupal Filebrowser module, provides FTP-like browsing capabilities for datasets (e.g., FASTA, GFF, BAM/CRAM, VCF). The catalogue of variants can also be accessed using Breeding API (BrAPI)[62]. The BGH is proposed as a FAIR (Findable, Accessible, Interoperable and Re-usable) compliant resource [94] (https://bio.tools/Banana_Genome_Hub), and according to FAIR checker (https://fair-checker.france-bioinformatique.fr/check), it scored a high level in terms of accessibility and findability (**Figure S1**).

Acknowledgements

This work was partially supported of the CGIAR Research Program on Roots, Tubers and Bananas (RTB), the Agropolis Foundation (ID 1504-006) ‘GenomeHarvest’ project through the French Investissements d’avenir programme (Labex Agro: ANR-10-LABX-0001-01). XJ. G. acknowledges support of the National Natural Science Foundation of China (No. 32070237, 31261140366). This work is technically supported by the South Green Bioinformatics platform and the CIRAD - UMR AGAP HPC Data Center. We warmly thank all data providers who proactively enrich the BGH with datasets and feedback including Sebastien Carpentier, Julie Sardos, Sijun Zheng, Nicolas Roux (Alliance Bioversity International - CIAT), David Studholme (University of Exeter), Boas Pucker (CeBiTec), Chunyan Xu, Xiaodong Fang (BGI), Ana Almeida (California State University East Bay), Wei Hu (CATAS), Mark Davey (KU Leuven), Dave Edwards, Philipp Bayer (University of Western Australia), Jose de Vega (Earlham Institute). We are grateful to Gabriel Sachter-Smith, Pat Heslop-Harrison, Julie Sardos, Ziwei Wang and Megan Hansen who provided the beautiful pictures for the homepage.

Contributions

M.R. and G.D. designed and managed the project. G.D. constructed the core database; V.G., M.S., E.D., G.S. developed additional modules. G.D., G.M., F-C.B., C.B. and M.R. collected and analysed datasets. P. H-H., T.S., XJ. G., N.Y., A.DH. supported the Hub with key resources.

M.R. drafted the manuscript, and all authors were involved in manuscript revision and approved the submitted version.

Conflict of interests

The authors declare that they have no conflict of interest.

References

1. Rouard M, Sardos J, Sempéré G *et al.* A digital catalog of high-density markers for banana germplasm collections. *PLANTS, PEOPLE, PLANET* 2021, DOI: https://doi.org/10.1002/ppp3.10187.

2. Borrell JS, Goodwin M, Blomme G *et al.* Enset-based agricultural systems in Ethiopia: A systematic review of production trends, agronomy, processing and the wider food security applications of a neglected banana relative. *PLANTS, PEOPLE, PLANET* 2020;**2**:212–28.

3. MusaNet, Laliberte B. *Global Strategy for the Conservation and Use of Musa Genetic Resources*. Bioversity International, 2016.

4. Ortiz R, Swennen R. From crossbreeding to biotechnology-facilitated improvement of banana and plantain. *Biotechnol Adv* 2013, DOI: 10.1016/j.biotechadv.2013.09.010.

5. Borrell JS, Biswas MK, Goodwin M *et al.* Enset in Ethiopia: a poorly characterized but resilient starch staple. *Ann Bot*, DOI: 10.1093/aob/mcy214.

6. Chen F, Song Y, Li X *et al.* Genome sequences of horticultural plants: past, present, and future. *Horticulture Research* 2019;**6**:112.

7. D'Hont A, Denoeud F, Aury J-M *et al.* The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* 2012;**488**:213.

8. Droc G, Lariviere D, Guignon V *et al.* The Banana Genome Hub. *Database* 2013;**2013**:bat035–bat035.

9. Martin G, Baurens F-C, Droc G *et al.* Improvement of the banana "Musa acuminata" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics* 2016;**17**:243.

10. Belser C, Baurens F-C, Noel B *et al.* Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. *Commun Biol* 2021;**4**:1–12.

11. Davey MW, Gudimella R, Harikrishna JA *et al.* A draft Musa balbisiana genome sequence for molecular genetics in polyploid, inter- and intra-specific Musa hybrids. *BMC Genomics* 2013;**14**:683.

12. Wang Z, Miao H, Liu J *et al.* Musa balbisiana genome reveals subgenome evolution and functional divergence. *Nature Plants* 2019;**5**:810–21.

13. Wu W, Yang Y-L, He W-M *et al.* Whole genome sequencing of a banana wild relative Musa itinerans provides insights into lineage-specific diversification of the Musa genus. *Sci Rep* 2016;**6**:31586.

14. Harrison J, Moore KA, Paszkiewicz K *et al.* A Draft Genome Sequence for Ensete ventricosum, the Drought-Tolerant "Tree Against Hunger." *Agronomy* 2014;**4**:13–33.

15. Galvez LC, Koh RBL, Barbosa CFC *et al.* Sequencing and de Novo Assembly of Abaca (Musa textilis Née) var. Abuab Genome. *Genes* 2021;**12**:1202.

16. Rouard M, Droc G, Martin G *et al.* Three New Genome Assemblies Support a Rapid Radiation in Musa acuminata (Wild Banana). *Genome Biology and Evolution* 2018;**10**:3129–40.

17. Rijzaani H, Bayer PE, Rouard M *et al.* The pangenome of banana highlights differences between genera and genomes. *The Plant Genome* 2021;**n/a**:e20100.

18. Belser C, Istace B, Denis E *et al.* Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants* 2018;**4**:879.

19. Wang Z, Rouard M, Biswas MK *et al.* A chromosome-level reference genome of Ensete glaucum gives insight into diversity and chromosomal and repetitive sequence evolution in the Musaceae. *GigaScience* 2022;**11**:giac027.

20. Christelová P, Langhe ED, Hřibová E *et al.* Molecular and cytological characterization of the global Musa germplasm collection provides insights into the treasure of banana diversity. *Biodivers Conserv* 2017;**26**:801–24.

21. Wendel JF, Jackson SA, Meyers BC *et al.* Evolution of plant genome architecture. *Genome Biology* 2016;**17**:37.

22. Garsmeur O, Schnable JC, Almeida A *et al.* Two Evolutionarily Distinct Classes of Paleopolyploidy. *Molecular Biology and Evolution* 2014;**31**:448–54.

23. Sass C, Iles WJD, Barrett CF *et al.* Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ* 2016;**4**:e1584.

24. Martin G, Carreel F, Coriton O *et al.* Evolution of the banana genome (Musa acuminata) is impacted by large chromosomal translocations. *Mol Biol Evol* 2017, DOI: 10.1093/molbev/msx164.

25. Cenci A, Guignon V, Roux N *et al.* Genomic analysis of NAC transcription factors in banana (Musa acuminata) and definition of NAC orthologous groups for monocots and dicots. *Plant Mol Biol* 2014;**85**:63–80.

26. Hu W, Zuo J, Hou X *et al.* The auxin response factor gene family in banana: genome-wide identification and expression analyses during development, ripening, and abiotic stress. *Front Plant Sci* 2015;**6**:742.

27. Backiyarani S, Anuradha C, Thangavelu R *et al.* Genome-wide identification, characterization of expansin gene family of banana and their expression pattern under various stresses. *3 Biotech* 2022;**12**:101.

28. Miao H, Sun P, Liu Q *et al.* Molecular identification of the key starch branching enzyme-encoding gene SBE2.3 and its interacting transcription factors in banana fruits. *Horticulture Research* 2020;**7**:1–15.

29. Zhang L, Cenci A, Rouard M *et al.* Transcriptomic analysis of resistant and susceptible banana corms in response to infection by Fusarium oxysporum f. sp. cubense tropical race 4. *Scientific Reports* 2019;**9**:8199.

30. Wesemael J, Hueber Y, Kissel E *et al.* Homeolog expression analysis in an allotriploid non-model crop via integration of transcriptomics and proteomics. *Scientific Reports* 2018;**8**:1353.

31. Sardos J, Rouard M, Hueber Y *et al.* A Genome-Wide Association Study on the Seedless Phenotype in Banana (Musa spp.) Reveals the Potential of a Selected Panel to Detect Candidate Genes in a Vegetatively Propagated Crop. *PLOS ONE* 2016;**11**:e0154448.

32. Nyine M, Uwimana B, Swennen R *et al.* Trait variation and genetic diversity in a banana genomic selection training population. *PLoS One* 2017;**12**, DOI: 10.1371/journal.pone.0178734.

33. Nyine M, Uwimana B, Akech V *et al.* Association genetics of bunch weight and its component traits in East African highland banana (Musa spp. AAA group). *Theor Appl Genet* 2019;**132**:3295–308.

34. Naim F, Dugdale B, Kleidon J *et al.* Gene editing the phytoene desaturase alleles of Cavendish banana using CRISPR/Cas9. *Transgenic Res* 2018, DOI: 10.1007/s11248-018-0083-0.

35. Ficklin SP, Sanderson L-A, Cheng C-H *et al.* Tripal: a construction toolkit for online genome databases. *Database (Oxford)* 2011;**2011**, DOI: 10.1093/database/bar044.

36. Sanderson L-A, Ficklin SP, Cheng C-H *et al.* Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. *Database* 2013;**2013**:bat075–bat075.

37. Staton M, Cannon E, Sanderson L-A *et al.* Tripal, a community update after 10 years of supporting open source, standards-based genetic, genomic and breeding databases. *Briefings in Bioinformatics* 2021, DOI: 10.1093/bib/bbab238.

38. Jung S, Cheng C-H, Buble K *et al.* Tripal MegaSearch: a tool for interactive and customizable query and download of big data. *Database* 2021;**2021**:baab023.

39. Priyam A, Woodcroft BJ, Rai V *et al.* Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. *Molecular Biology and Evolution* 2019;**36**:2922–4.

40. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. *Bioinformatics* 2021;**37**:1639–43.

41. Yachdav G, Wilzbach S, Rauscher B *et al.* MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics* 2016;**32**:3501–3.

42. Shank SD, Weaver S, Kosakovsky Pond SL. phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC Bioinformatics* 2018;**19**:276.

43. Zorrilla-Fontanesi Y, Rouard M, Cenci A *et al.* Differential root transcriptomics in a polyploid non-model crop: the importance of respiration during osmotic stress. *Sci Rep* 2016;**6**:22583.

44. Cenci A, Hueber Y, Zorrilla-Fontanesi Y *et al.* Effect of paleopolyploidy and allopolyploidy on gene expression in banana. *BMC Genomics* 2019;**20**:244.

45. Dale J, James A, Paul J-Y *et al.* Transgenic Cavendish bananas with resistance to Fusarium wilt tropical race 4. *Nat Commun* 2017;**8**:1496.

46. Cassan O, Lèbre S, Martin A. Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite. *BMC Genomics* 2021;**22**:387.

47. Drapal M, Carvalho EB de, Rouard M *et al.* Metabolite profiling characterises chemotypes of Musa diploids and triploids at juvenile and pre-flowering growth stages. *Scientific Reports* 2019;**9**:4657.

48. Drapal M, Amah D, Schöny H *et al.* Assessment of metabolic variability and diversity present in leaf, peel and pulp tissue of diploid and triploid Musa spp. *Phytochemistry* 2020;**176**:112388.

49. Price EJ, Drapal M, Perez-Fons L *et al.* Metabolite database for root, tuber, and banana crops to facilitate modern breeding in understudied crops. *The Plant Journal* 2020;**101**:1258–68.

50. Du L, Song J, Forney C *et al.* Proteome changes in banana fruit peel tissue in response to ethylene and high-temperature treatments. *Horticulture Research* 2016;**3**:16012.

51. Karp PD, Midford PE, Billington R *et al.* Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics* 2021;**22**:109–26.

52. Paul J-Y, Khanna H, Kleidon J *et al.* Golden bananas in the field: elevated fruit pro-vitamin A from the expression of a single banana transgene. *Plant Biotechnol J* 2017;**15**:520–32.

53. Amah D, van Biljon A, Brown A *et al.* Recent advances in banana (musa spp.) biofortification to alleviate vitamin A deficiency. *Critical Reviews in Food Science and Nutrition* 2019;**59**:3498–510.

54. Kozicka M, Elsey J, Ekesa B *et al.* Reassessing the Cost-Effectiveness of High-Provitamin A Bananas to Reduce Vitamin A Deficiency in Uganda. *Front Sustain Food Syst* 2021;**5**, DOI: 10.3389/fsufs.2021.649424.

55. Sempéré G, Pétel A, Rouard M *et al.* Gigwa v2—Extended and improved genotype investigator. *Gigascience* 2019;**8**, DOI: 10.1093/gigascience/giz051.

56. Sempéré G, Larmande P, Rouard M. Managing High-Density Genotyping Data with Gigwa. In: Edwards D (ed.). *Plant Bioinformatics: Methods and Protocols*. New York, NY: Springer US, 2022, 415–27.

57. Sardos J, Breton C, Perrier X *et al.* Wild to domesticates: genomes of edible diploid bananas hold traces of several undefined genepools. *bioRxiv* 2021:2021.01.29.428762.

58. Baurens F-C, Martin G, Hervouet C *et al.* Recombination and Large Structural Variations Shape Interspecific Edible Bananas Genomes. *Mol Biol Evol* 2019;**36**:97–111.

59. Cenci A, Sardos J, Hueber Y *et al.* Unravelling the complex story of intergenomic recombination in ABB allotriploid bananas. *Annals of Botany* 2021;**127**:7–20.

60. Ruas M, Guignon V, Sempere G *et al.* MGIS: managing banana (Musa spp.) genetic resources information and high-throughput genotyping data. *Database (Oxford)* 2017;**2017**, DOI: 10.1093/database/bax046.

61. Van den houwe I, Chase R, Sardos J *et al.* Safeguarding and using global banana diversity: a holistic approach. *CABI Agriculture and Bioscience* 2020;**1**:15.

62. Selby P, Abbeloos R, Backlund JE *et al.* BrAPI—an application programming interface for plant breeding applications. *Bioinformatics* 2019;**35**:4147–55.

63. Yang X, Lee W-P, Ye K *et al.* One reference genome is not enough. *Genome Biology* 2019;**20**:104.

64. Khan AW, Garg V, Roorkiwal M *et al.* Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends in Plant Science* 2020;**25**:148–58.

65. Durant É, Sabot F, Conte M *et al.* Panache: a web browser-based viewer for linearized pangenomes. *Bioinformatics* 2021, DOI: 10.1093/bioinformatics/btab688.

66. Martin G, Cardi C, Sarah G *et al.* Genome ancestry mosaics reveal multiple and cryptic contributors to cultivated banana. *The Plant Journal* 2020;**102**:1008–25.

67. Summo M, Comte A, Martin G *et al.* GeMo: a web-based platform for the visualization and curation of genome ancestry mosaics. *Database* (accepted).

68. Fu N, Ji M, Rouard M *et al.* Comparative plastome analysis of Musaceae and new insights into phylogenetic relationships. *BMC Genomics* 2022;**23**:223.

69. Wang Z, Rouard M, Biswas MK *et al.* A chromosome-level reference genome of Ensete glaucum gives insight into diversity, chromosomal and repetitive sequence evolution in the Musaceae. 2021:2021.11.23.469474.

70. Bandi V, Gutwin C. Interactive Exploration of Genomic Conservation. *46th Graphics Interface Conference on Proceedings of Graphics Interface 2020 (GI'20)*. Waterloo, Canada: Canadian Human-Computer Communications Society, 2020.

71. Martin G, Baurens F-C, Cardi C *et al.* The Complete Chloroplast Genome of Banana (Musa acuminata, Zingiberales): Insight into Plastid Monocotyledon Evolution. *PLoS ONE* 2013;**8**:e67350.

72. Li W, Liu Y, Gao L-Z. The complete chloroplast genome of the endangered wild Musa itinerans (Zingiberales: Musaceae). *Conservation Genet Resour* 2017:1–3.

73. Shetty SM, Shah MUM, Makale K *et al.* Complete Chloroplast Genome Sequence of Musa balbisiana Corroborates Structural Heterogeneity of Inverted Repeats in Wild Progenitors of Cultivated Bananas and Plantains. *The Plant Genome* 2016;**9**:plantgenome2015.09.0089.

74. Song W, Ji C, Chen Z *et al.* Comparative Analysis the Complete Chloroplast Genomes of Nine Musa Species: Genomic Features, Comparative Analysis, and Phylogenetic Implications. *Frontiers in Plant Science* 2022;**13**.

75. Wu C-S, Sudianto E, Chiu H-L *et al.* Reassessing Banana Phylogeny and Organelle Inheritance Modes Using Genome Skimming Data. *Front Plant Sci* 2021;**12**:713216.

76. Zdobnov EM, Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 2001;**17**:847–8.

77. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011;**2011**, DOI: 10.1093/database/bar009.

78. Yemataw Z, Muzemil S, Ambachew D *et al.* Genome sequence data from 17 accessions of Ensete ventricosum, a staple food crop for millions in Ethiopia. *Data in Brief* 2018;**18**:285–93.

79. Busche M, Pucker B, Viehöver P *et al.* Genome Sequencing of Musa acuminata Dwarf Cavendish Reveals a Duplication of a Large Segment of Chromosome 2. *G3: Genes, Genomes, Genetics* 2020;**10**:37–42.

80. Li Z, Wang J, Fu Y *et al.* The Musa troglodytarum L. genome provides insights into the mechanism of non-climacteric behaviour and enrichment of carotenoids. *BMC Biology* 2022;**20**:186.

81. Sambles C, Venkatesan L, Shittu OM *et al.* Genome sequencing data for wild and cultivated bananas, plantains and abacá. *Data in Brief* 2020:106341.

82. Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21.

83. Claudel-Renard C, Chevalet C, Faraut T *et al.* Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Research* 2003;**31**:6633.

84. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* 2016;**428**:726–31.

85. Wang Y, Tang H, DeBarry JD *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 2012;**40**:e49.

86. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;**16**:157.

87. Guignon V, Toure A, Droc G *et al.* GreenPhylDB v5: a comparative pangenomic database for plant genomes. *Nucleic Acids Research* 2021;**49**:D1464–71.

88. Leinonen R, Sugawara H, Shumway M *et al.* The Sequence Read Archive. *Nucleic Acids Research* 2011;**39**:D19–21.

89. Jung S, Ficklin SP, Lee T *et al.* The Genome Database for Rosaceae (GDR): year 10 update. *Nucleic Acids Res* 2014;**42**:D1237-1244.

90. Guo W, Chen J, Li J *et al.* Portal of Juglandaceae: A comprehensive platform for Juglandaceae study. *Hortic Res* 2020;**7**:1–8.

91. Buels R, Yao E, Diesh CM *et al.* JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* 2016;**17**:66.

92. Tripathi JN, Ntui VO, Shah T *et al.* CRISPR/Cas9-mediated editing of DMR6 orthologue in banana (Musa spp.) confers enhanced resistance to bacterial disease. *Plant Biotechnology Journal* 2021;**19**:1291–3.

93. Morales N, Ogbonna AC, Ellerbrock BJ *et al.* Breedbase: a digital ecosystem for modern plant breeding. *G3 Genes|Genomes|Genetics* 2022:jkac078.

94. Wilkinson MD, Dumontier M, Aalbersberg IjJ *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 2016;**3**:160018.