Received: 1 February 2021 | Accepted: 22 March 2021

DOI: 10.1002/tpg2.20100

The Plant Genome

SPECIAL ISSUE: PANGENOMES IN CROP IMPROVEMENT

The pangenome of banana highlights differences between genera and genomes

Habib Rijzaani[1,2] | **Philipp E. Bayer**[1] | **Mathieu Rouard**[3] | **Jaroslav Doležel**[4] | **Jacqueline Batley**[1] | **David Edwards**[1]

[1] School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, WA, Australia

[2] Indonesian Agency for Agricultural Research and Development, Jakarta, Indonesia

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

[4] Institute of Experimental Botany of the Czech Academy of Sciences, Centre of the Region Hana for Biotechnological and Agricultural Research, Šlechtitelů 31, Olomouc 77900, Czech Republic

Correspondence
David Edwards, School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, WA, Australia.
Email: Dave.edwards@uwa.edu.au

Assigned to Associate Editor Agnieszka Golicz.

Funding information
Australian Research Council, Grant/Award Numbers: DP1601004497, DP200100762, LP130100061, LP160100030; ERDF project, Grant/Award Number: CZ.02.1.01/0.0/0.0/16_019/0000827

Abstract

Banana (Musaceae family) has a complex genetic history and includes a genus *Musa* with a variety of cultivated clones with edible fruits, *Ensete* species that are grown for their edible corm, and monospecific *Musella* whose generic status has been questioned. The most commonly exported banana cultivars belong to Cavendish, a subgroup of *Musa* triploid cultivars, which is under threat by fungal pathogens, though there are also related species *M. balbisiana* Colla (B genome), *M. textilis* Née (T genome), and *M. schizocarpa* N. W. Simmonds (S genome), along with hybrids of these genomes, which potentially host genes of agronomic interest. Here we present the first cross-genus pangenome of banana, which contains representatives of the *Musa* and *Ensete* genera. Clusters based on gene presence–absence variation (PAV) clearly separate *Musa* and *Ensete*, while *Musa* is split further based on species. These results present the first pangenome study across genus boundaries and identifies genes that differentiate between Musaceae species, information that may support breeding programs in these crops.

1 | INTRODUCTION

The banana family (Musaceae) is a monophyletic clade comprising of three genera, *Musa*, *Ensete*, and *Musella* (Kress, 1990). Most of the identified species of this family belong to the *Musa* genus, which includes the edible fruit-bearing banana cultivars, with the latest database of *Musa* Germplasm Information System recording 6,548 accessions that are maintained in 29 collections around the world (Ruas et al., 2017). In contrast, the *Ensete* genus is comprised of cultivars with nonedible fruits (Oyen & Lemmens, 2002). The corm or the basal pseudo-stem of the *Ensete* plants is used as a starch-rich

Abbreviations: BUSCO, benchmarking universal single-copy orthologs; GO, gene ontology; NLR, nucleotide binding-site leucine rich repeat; PAV, presence–absence variation; RGA, resistance gene analog; SNP, single nucleotide polymorphism.

food source, predominantly in Ethiopia, where it plays significant agricultural and economic roles. There are six known species within the *Ensete* genus (Ploetz et al., 2007), with 387 accessions characterized by Yemataw et al. (2017). *Musella* is monotypic occurring only in higher altitudes in southwestern China (Liu et al., 2003) and may be monophyletic with *Ensete* (Liu et al., 2010).

In contrast to the limited geographical distribution of *Ensete*, banana from the genus *Musa* is widely produced across the tropics and subtropics, cultivated on more than 5.6 million ha and produces about 114 Tg of fruit annually (FAO, 2018). The main banana producing countries are from southern, eastern, and southeastern Asia; central and eastern Africa; as well as Central America and the Caribbean. Although there are hundreds of banana cultivars around the globe, few are grown commercially for large-scale production, with the main commercial banana being triploid clones from the Cavendish subgroup such as 'Grand Naine' and 'Robusta' (FAO, 2019).

Cultivation of banana is threatened by both biotic and abiotic stress, and most banana cultivars are prone to cold damage, drought, or salinity (van Asten et al., 2011). Pests and pathogens can threaten banana cultivation, including Fusarium wilt, black leaf streak disease, banana bunchy top disease, and Moko disease (Ploetz et al., 2015). These diseases can spread quickly because of the monoculture practice of banana cultivation. This is exemplified by the collapse of the 'Gros Michel' based banana trade in the 1960s caused by the spread of *Fusarium oxysporum* f. sp. *cubense* (Robinson, 1996). The Cavendish banana that replaced Gros Michel after the epidemic is now being threatened by the tropical race 4 (TR4) pathotype of the same pathogen (Ploetz et al., 2015).

Banana breeding and cultivar improvement are being hampered by banana's long generation time and low fertility. Despite this, successful hybrids were produced by conventional cross-breeding at FHIA, EMBRAPA, and some other breeding programs (Escalant et al., 2002) but the level of adoption of the hybrids has remained low (Thiele et al., 2021). Tissue-culture-based approaches such as in vitro embryo rescue, genetic engineering, or genome editing techniques could overcome some of these problems (Tripathi et al., 2015), and some reports indicate the successful production of transgenic banana with improved agronomic traits, including resistance to biotic stress (Ghag et al., 2015; Tripathi et al., 2017) and better postharvest handling (Elitzur et al., 2016). While most transgenic banana improvement remains in the laboratory (Dale et al., 2017), some field trials are underway in banana producing countries including Australia, Uganda, and the Philippines (Paul et al., 2018).

As the cost of DNA sequencing continues to decline, increasing quantities of genetic and genomic information are becoming available for banana and its relatives. There are currently seven reference genome assemblies available for the *Musa* genus. The *M. acuminata* ssp. *malaccensis* (Ridl.) N. W. Simmonds genome, derived from a doubled-haploid Pahang accession, represents the banana A genome ($n = 11$) (D'Hont et al., 2012; Martin et al., 2016), whereas *M. balbisiana* Colla, derived from a Pisang Klutuk Wulung accession, represents the B genome ($x = 11$) (Davey et al., 2013; Wang et al., 2019). There is also an assembled genome for *M. itinerans* Cheesman, a cold-tolerant banana cultivar from the Yunan region in China (Wu et al., 2016). More recently, three genomes from different subspecies of *M. acuminata*, namely ssp. *banksii* N. W. Simmonds, *zebrina* (Van Houtte ex Planch.) Nasution, and *burmannica* N. W. Simmonds were published (Rouard et al., 2018), while another publication presented a chromosome-scale assembly of the *M. schizocarpa* genome (S genome, $x = 11$) (Belser et al., 2018). This latest genome assembly benefited from the incorporation of long-read sequences from third-generation genome sequencing. These assemblies cover all the known genome types of *Musa* banana, except *Australimusa* (T genome, $x = 10$).

Genomic resources for *Ensete* ($x = 9$) are also becoming available. One draft reference genome assembly has been published for an unknown variety of *E. ventricosum* (Welw.) Cheesman, together with three assemblies for known Ethiopian cultivars: Derea, Bedadeti, and Onjamo (Harrison et al., 2014). Whole-genome sequencing data is also available for a further 17 *E. ventricosum* accessions representing 15 varieties (Yemataw et al., 2018), though no whole-genome sequence data is available for other species of *Ensete*.

To complement this public information, we have generated Illumina next-generation sequencing data for nine banana accessions with an average coverage of 55×. These accessions include *M. acuminata* species, some hybrids with B-genome banana with various ploidy levels, and a variety of *Fe'i* banana (*Australimusa/Callimusa*, T genome), which is thought to be domesticated independently from both *M. acuminata* and its hybrids with *M. balbisiana* and is characterized by its upright fruit bunch and fruit with high-carotenoid content (Sharrock et al., 2001). Using this data, we produced a draft cross-genus pangenome that captures the diversity of banana genome types and highlights the diversity of gene content across the Musaceae, identifying genes that may play a role in the evolutionary differentiation between *Ensete* and *Musa* genera as

Core Ideas

- We assembled the first banana pangenome across two genera.
- The two genera exhibit high levels of divergence.
- The banana pangenome contains very few novel disease resistance genes.

well as variable genes that could be used to improve the agronomic performance and disease resistance in this important family.

2 | MATERIALS AND METHODS

Genomic short reads of diverse banana accessions from two genera of Musaceae, *Musa* and *Ensete*, were collected from both publicly available data and newly generated from banana samples from Indonesian Agricultural Agency for Research and Development at the Indonesian Tropical Fruit Research Institute germplasm collection. The metadata of the reads used in the banana pangenome construction and analysis is listed in Supplemental Tables S1 and S2.

We applied the previously published iterative pangenome assembly pipeline (Golicz et al., 2016; Montenegro et al., 2017; Zhao et al., 2018). Bowtie2 v2.3.3.1 (-I 0 -X 1000 –end-to-end –sensitive) (Langmead & Salzberg, 2012) and SAMtools v1.2 (Li et al., 2009) were used for read mapping and subsequent extraction of unmapped reads. We assembled the pangenome in 14 iterations, first starting with the five *M. acuminata* individuals with a coverage above 10×, then merged the remaining <10× coverage *M. acuminata* individuals into one assembly in step six, then added the remaining *Musa* and *Ensete* individuals in separate steps. The order of iterations is shown in Supplemental Table S2. The assembly of additional pangenome contigs was conducted using MaSuRCA v3.1.3 (Zimin et al., 2013). Contaminant (nonplant) contigs were discarded by comparing all assembled contigs with NCBI-NR using BLAST and discarding contigs with the highest-scoring hits with nonplants.

Repeat elements were identified and masked using RepeatMasker v4.0.6 (Smit et al., 2015) and gene models were generated using MAKER v.2.31.8 (Holt & Yandell, 2011) and AUGUSTUS v3.2.2 (Stanke et al., 2006). Protein, expressed sequence tag, and transcriptome data were collected from the Sequence Read Archive and used as evidence for gene models (Supplemental Table S3). Gene ontology (GO) terms were assigned to the identified gene models using Blast2GO (Conesa et al., 2005). The GO-term enrichment used Blast2GO and topGO (Alexa & Rahnenführer, 2009). Benchmarking universal single-copy orthologs (BUSCO) (Simão et al., 2015) from OrthoDB (www.orthodb.org) analysis used 1,440 orthologs from the plant database (embryophyta_odb9) (Zdobnov et al., 2017).

Single nucleotide polymorphisms (SNPs) were called following Bowtie2 read mapping using Bcftools v1.2-63 (Li, 2011) and were quality processed (SAMtools mpileup -q 30 -Q 20). Functional annotation of the final SNPs was carried out using the SnpEff tool (Cingolani et al., 2012).

To call presence–absence variations (PAVs), reads from all samples were mapped to the pangenome using Bowtie2, and SGSGeneLoss was used to call PAVs using standard settings (Golicz et al., 2015). Samples with a coverage below 15× were excluded. PVClust (Suzuki & Shimodaira, 2006) was used for hierarchical clustering based on the gene presence–absence matrix. Core and variable gene counts were plotted using PanGP (Y. Zhao et al., 2014). This was also used for the modeling and prediction of the pangenome size. vcftools (Danecek et al., 2011) was used to calculate Weir-Cockerham's F_{ST} values at 100K nucleotide windows across the pangenome, which were visualized using R. Genes were classified into resistance gene analogs (RGAs) by using RGaugury (Li et al., 2016).

The data generated for this project is available in the Sequence Read Archive under BioProject ID PRJNA612747. All assemblies and annotations generated are available at the Banana Genome Hub (Droc et al., 2013) at https://banana-genome-hub.southgreen.fr and https://doi.org/10.26182/y00m-nb33.

3 | RESULTS AND DISCUSSION

We have assembled the first pangenome of the banana family (Musaceae) using the iterative mapping and assembly approach (Golicz et al., 2016), representing three *Musa* species, a diverse group of cultivated banana clones and one *Ensete* species. This approach requires the selection of an initial reference assembly. There are seven draft or reference genomes published for *Musa*, including reference assemblies of *M. acuminata* spp. *malaccensis* (A genome), *M. schizocarpa* (S genome), and *M. balbisiana* (B genome), a draft assemblies of *M. itinerans*, as well as three draft assemblies from different subspecies of *M. acuminata*, namely ssp. *banksii, zebrina and burmannica* (A genome) (D'Hont et al., 2012; Davey et al., 2013; Wu et al., 2016; Belser et al., 2018; Rouard et al., 2018). In addition, there are four genome assemblies for *Ensete* at the scaffold level (Yemataw et al., 2018) (Supplemental Table S4). Analysis of the assembly statistics suggests that the *M. acuminata* spp. *malaccensis* and *M. schizocarpa* genome assemblies are the most contiguous. The *M. acuminata* assembly showed the most complete BUSCO orthologs for both single-copy and duplicated genes (Supplemental Table S5; Supplemental Figure S1) and so was selected as the base reference for iterative pangenome assembly. Additional public sequence data is listed in Supplemental Table S1.

Whole-genome Illumina sequence data was collated from 15 banana accessions, 12 of which represent the genus *Musa* with three from *Ensete*. The *Musa* accessions include six representatives of *M. acuminata* (A genome), one from *M. balbisiana* (B genome), *M. itinerans*, and *M. Fe'I* (T genome) as well as three A-B genome hybrids (one AB and two AAB). The reference-guided iterative assembly approach (Golicz

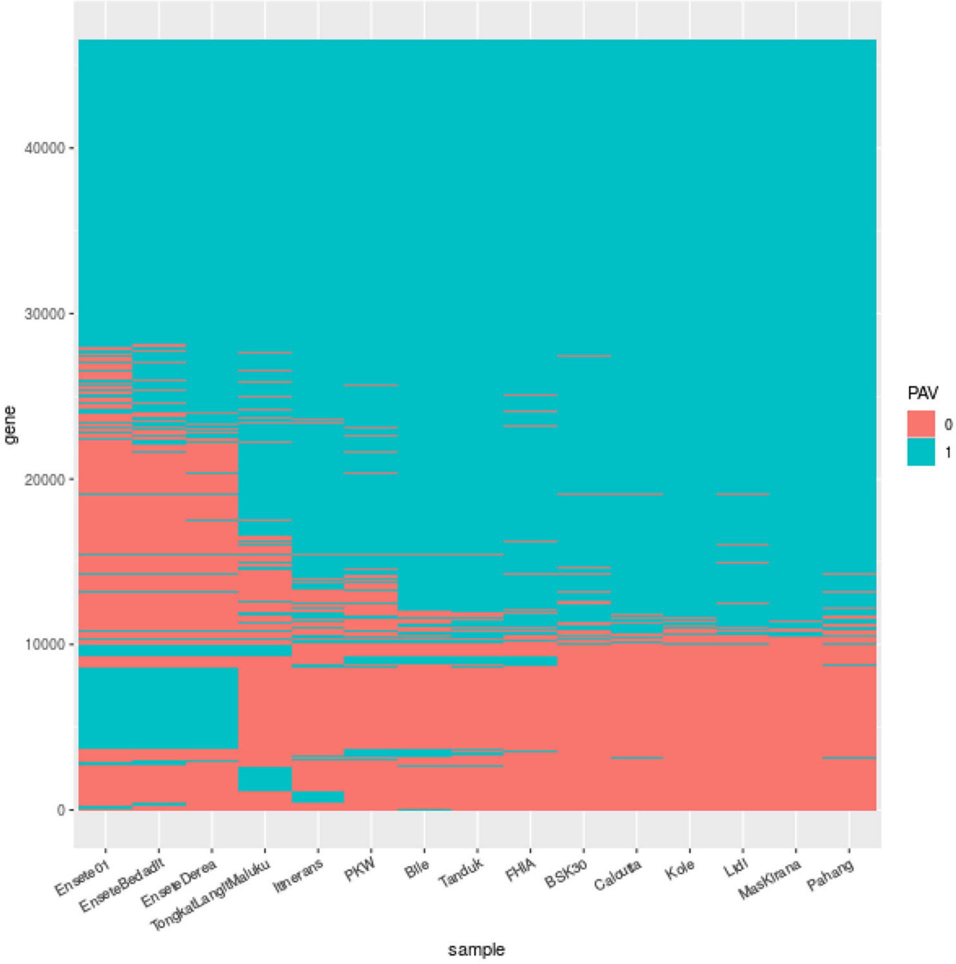

FIGURE 1 The presence or absence of genes among the banana accessions. Genes are ordered in decreasing order by the number of shared genes among all accessions

et al., 2016) produced 510,493 scaffolds with a total length of 672 Mb in addition to the *M. acuminata* reference genome of 451 Mb (A genome). The assembled pangenome produced a slight increase in BUSCO score compared with the single reference genome (Supplemental Table S6). This is in line with other pangenomes; for example, in the rape (*Brassica napus* L.) pangenome the assembled pangenome contained a single additional complete BUSCO not present in the reference genome (Hurgobin et al., 2018).

Gene prediction identified 12,310 candidate protein-coding gene models in the newly assembled contigs in addition to 35,276 gene models already identified in the *M. acuminata* genome. Other plant pangenome studies have identified a much smaller proportion of additional sequence. For example, the rice (*Oryza sativa* L.) pangenome, which was constructed from thousands of cultivars, only identified 80 Mb of novel sequence, equal to 30% of the reference genome (Zhao et al., 2018). The hexaploid wheat (*Triticum aestivum* L.) pangenome identified 350 Mb of additional sequence (Montenegro et al., 2017), an increase in size of <10%. Similarly, the kale (*B. oleracea* L.) pangenome size is ~25% larger than the reference *B. oleracea* genome (Golicz et al., 2016). These previous pangenomes were assembled within a specific genus, or with species of the same genome, and the large pangenome size we observe in banana is most likely a result of the inclusion of multiple *Musa* species as well as the *Ensete* genus. The inclusion of several cultivated *Musa* may also contribute to the expansion of the pangenome size, as their genomes diverged, and include admixture with other genomes during domestication (Martin et al., 2020). This broader approach permits the examination of gene presence or absence across diverse Musaceae and highlights the applicability of cross-genera pangenomics.

Following construction of the pangenome, we remapped the original next-generation sequencing reads to the pangenome and identified gene PAV using SGSGeneloss (Golicz et al., 2015). On average, there are 34,014 genes present in each of the Musaceae individuals, with the smallest number observed for sample Ensete01 (27,452) and highest observed for Bile (Bire) with (37,091). All *Ensete* samples

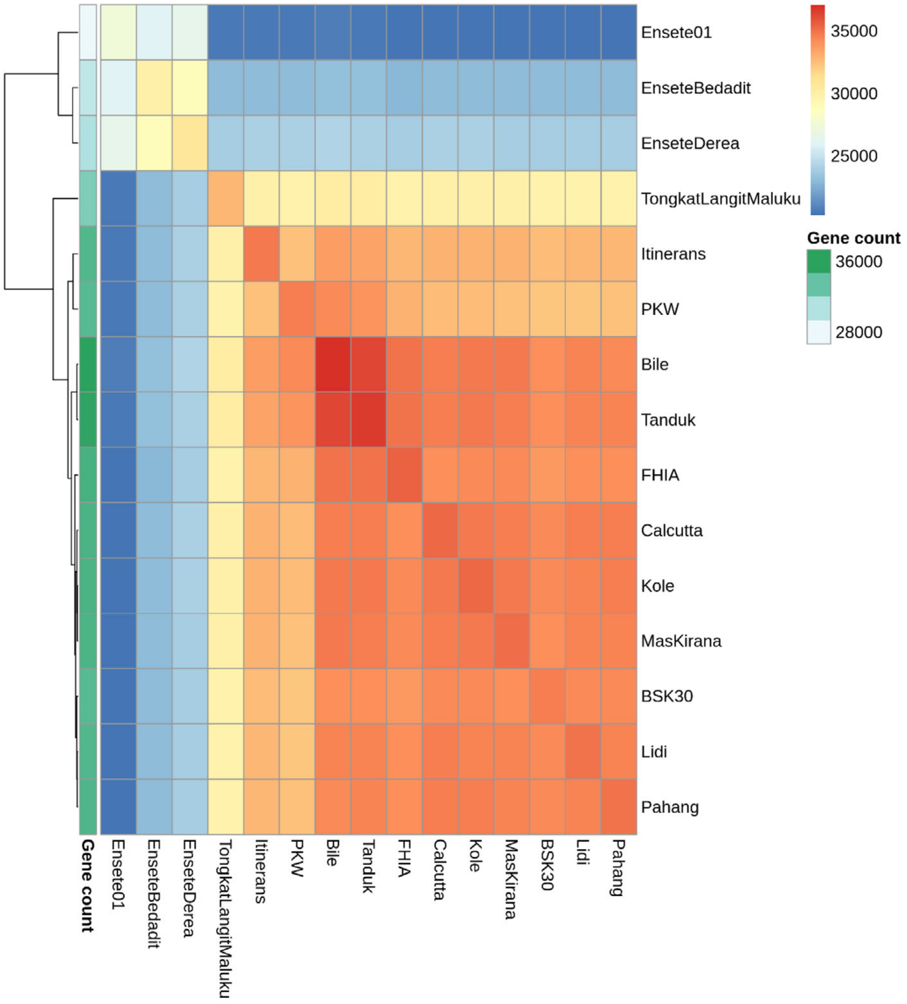

FIGURE 2 Matrix of numbers of shared genes among the banana accessions used for the pangenome study

appear to have fewer genes than *Musa* samples (Supplemental Table S7). While the number of genes identified for *Musa* species is similar to the number of genes in the *M. acuminata* reference genome, the number of predicted genes in *Ensete* species is less than the 42,749 genes reported for the draft genome of *E. ventricosum*; however, this number could be inflated because of the fragmented nature of the *E. ventricosum* draft genome assembly (Harrison et al., 2014).

Genes showed distinct presence–absence patterns that reflect the divisions of the banana genomes. The set of genes shared by *Musa* species are separate from the set shared by *Ensete* species (Figure 1), and the matrix of shared genes among banana accessions also shows two distinct clusters based on genus (Figure 2). The majority of genes (30,119) are found in both *Ensete* and *Musa*, while 5,629 are present only in the three *Ensete* individuals and 11,924 are present only in *Musa* individuals.

Gene PAV of the Musaceae pangenome showed a similar trend to previous studies among Musaceae. A study of gene content for the published *E. ventricosum* draft genome by Harrison et al. (2014) found 662 gene models of *M. acuminata* that are absent in *E. ventricosum*, and 9,967 gene models in the *E. ventricosum* genome that were not found in the banana proteome database, though, as the authors acknowledge, the published *E. ventricosum* gene model content may be inflated. Prediction of the banana pangenome size using these gene models and PAV data suggest a core gene number of 18,288 (±29). The observed number of core genes was 18,359 (Figure 3A), suggesting that we have defined the majority of the core gene content. The number of observed variable genes was 29,331 (Figure 3A). Restricting the analysis to within the *Musa* genus (12 samples) we predict a larger number of core genes (27,858 ± 69), with the difference reflecting genes that appear to be conserved in *Musa* but absent from *Ensete* (Figure 3B).

Principal component analysis of the PAV data revealed sample clustering consistent with the genus divisions, with banana samples in the *Musa* genus separate from the

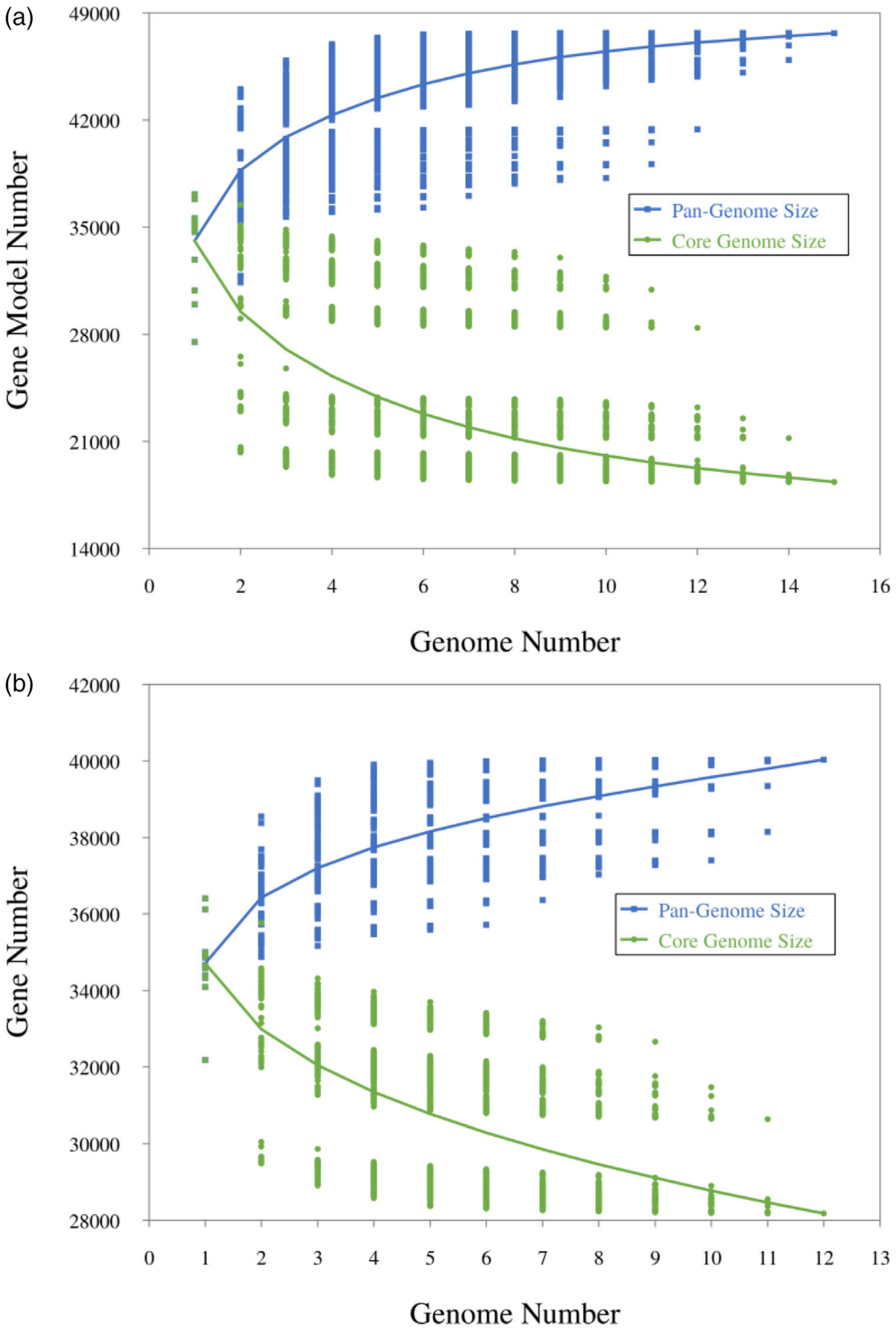

FIGURE 3 Gene content modeling showing core and pan-genome size growths. (a) All samples, both from *Musa* and *Ensete* genera, were incorporated into the model. (b) Gene content modeling based on *Musa* accessions only

Ensete genus (Supplemental Figure S2). Subclustering within *Musa* clearly follows subgenome delineation. For example, doubled-haploid cultivar Pahang (A genome) is separated from cultivar Pisang Klutuk Wulung (diploid, B genome) and cultivar Tongkat Langit Maluku (T genome). The SNP-based clustering shows highly similar patterns (Supplemental Figure S3). *Musa acuminata* groups are separated into two subclusters based on PAV (Supplemental Figure S4), with one group being A genome individuals and the other group being composed of cultivars Tanduk and Bile (AAB and AB respectively) as well as the improved cultivar FHIA 25 (AAB).

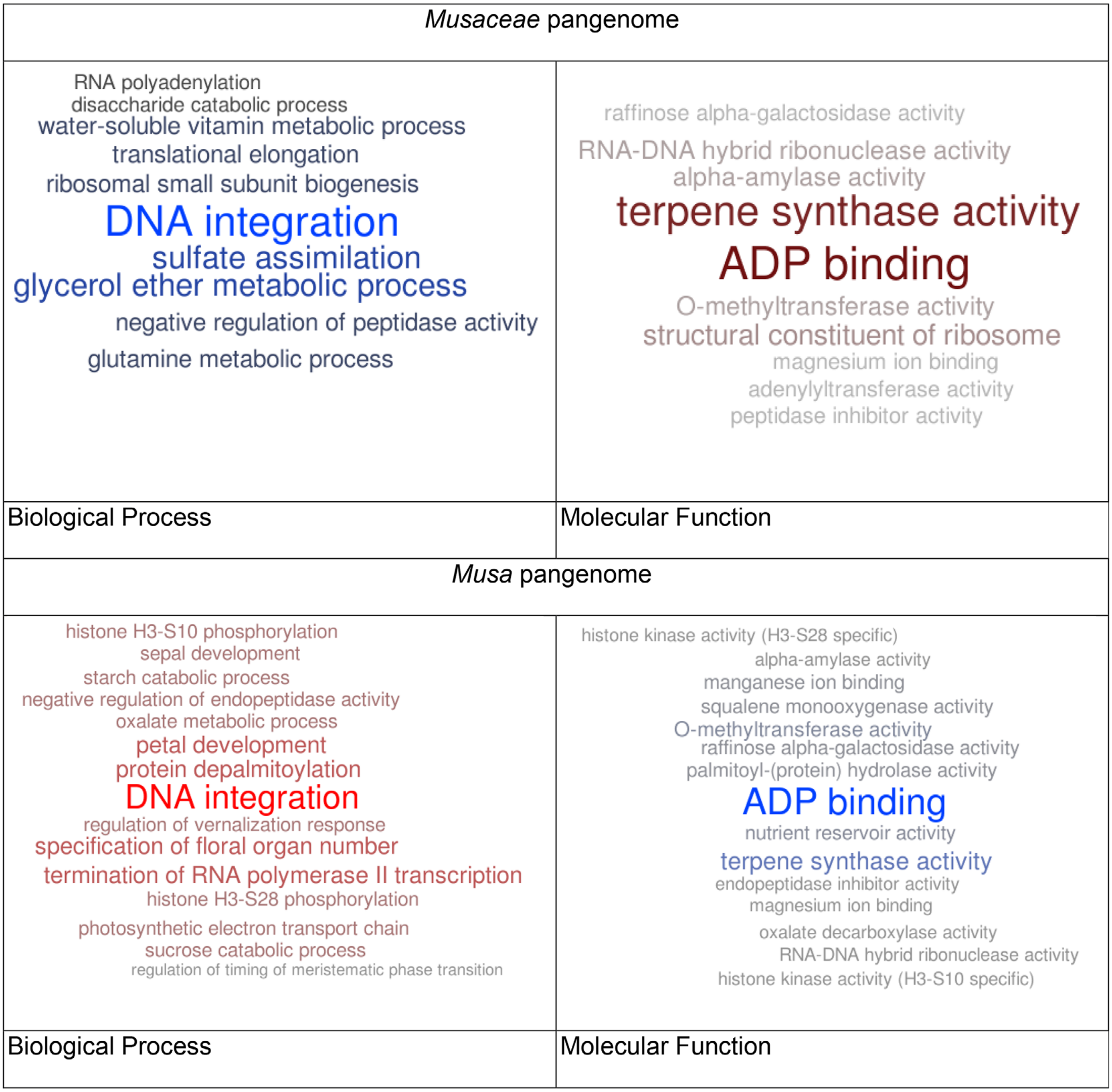
Musaceae pangenome
RNA polyadenylation
disaccharide catabolic process
water-soluble vitamin metabolic process
translational elongation
ribosomal small subunit biogenesis
DNA integration
sulfate assimilation
glycerol ether metabolic process
negative regulation of peptidase activity
glutamine metabolic process
raffinose alpha-galactosidase activity
RNA-DNA hybrid ribonuclease activity
alpha-amylase activity
terpene synthase activity
ADP binding
O-methyltransferase activity
structural constituent of ribosome
magnesium ion binding
adenylyltransferase activity
peptidase inhibitor activity
Biological Process
Molecular Function
Musa pangenome
histone H3-S10 phosphorylation
sepal development
starch catabolic process
negative regulation of endopeptidase activity
oxalate metabolic process
petal development
protein depalmitoylation
DNA integration
regulation of vernalization response
specification of floral organ number
termination of RNA polymerase II transcription
histone H3-S28 phosphorylation
photosynthetic electron transport chain
sucrose catabolic process
regulation of timing of meristematic phase transition
histone kinase activity (H3-S28 specific)
alpha-amylase activity
manganese ion binding
squalene monooxygenase activity
O-methyltransferase activity
raffinose alpha-galactosidase activity
palmitoyl-(protein) hydrolase activity
ADP binding
nutrient reservoir activity
terpene synthase activity
endopeptidase inhibitor activity
magnesium ion binding
oxalate decarboxylase activity
RNA-DNA hybrid ribonuclease activity
histone kinase activity (H3-S10 specific)
Biological Process
Molecular Function

FIGURE 4 Gene ontology terms overrepresented in the variable regions of the banana pangenome

Gene ontology enrichment analysis was performed to characterize the variable genes within the Musaceae and the *Musa* genus samples (Figure 4; Supplemental Tables S8 & S9). This highlighted diverse terms associated with metabolism across the Musaceae. Interestingly, terms related to flowering, meristem regulation, and nutrient metabolism are identified as enriched in the variable genes among *Musa* genus samples, and these functions may reflect the morphological diversity among *Musa* species as exemplified by various size, shape, color, texture, and taste of the fruit. The differences in enrichment of genes related to flowering could also reflect the difference in flowering between genus *Musa* and *Ensete*, with flowering very delayed in *Ensete* plants (Tsegaye & Struik, 2002).

Genomic variation in the form of SNPs were predicted, and we identified 10,926,656 candidate genic SNPs, with 1,082,854 found in variable genes and 9,843,802 in core genes. A greater proportion of SNPs in variable genes were predicted to have an impact on protein structure and function, suggesting reduced selective pressure for these genes (Supplemental Figure S5). There was relatively wide genetic differentiation between *Ensete* and *Musa* genera as apparent from average Weir-Cockerham's F_{ST} value of 0.29. Selective sweep analysis further identified regions in the pangenome that may explain the genetic differentiation among the two genera. Some genes in the highly differentiated regions have GO terms related to flavonoid biosynthesis, response to stress including defence response, meristem initiation, cell division, protein transport and refolding, as well as metabolite transport (Figure 4). The genes with the highest F_{ST} values are listed in Supplemental Table S10. The selective sweep analysis identified regions that may play roles in evolution and selection in the two genera and includes genes related to drought tolerance and auxin biosynthesis. *Ensete* plants are known to be tolerant to drought stress (Borrell et al., 2019), and understanding the genomic basis for this important trait may support the breeding of more drought-tolerant *Musa* banana species.

Disease is a constant threat to banana cultivation, and the identification of resistance is a major target for breeding new varieties. A total of 965 RGAs were identified, including 702 receptor-like kinase genes, 136 nucleotide binding-site leucine rich repeat (NLR) genes, and 127 receptor-like proteins (Supplemental Table S11). Of the 965 RGAs, the majority are found in the reference assembly and only 70 (5.7%) are located in the newly assembled contigs. These numbers are much lower than other pangenomes (Hurgobin et al., 2018; Zhao et al., 2018; Bayer et al., 2019) suggesting that there is very low diversity in disease resistance genes across *Musa* and *Ensete*. In total, 717 Musaceae RGAs were core genes (74%) and the remaining 248 were variable (26%). A dendrogram based on RGA PAVs results in distinct *Musa* and *Ensete* groups (Supplemental Figure S6). The presence of these genus-specific RGAs can be explained by their evolutionary distance and differential pathogen pressure during their evolution. Wu et al. (2016) identified 117, 93, and 62 NLRs in *M. acuminata*, *M. balbisiana*, and *M. itinerans*, respectively, and the authors suggest the decreasing number of NLRs among these three species happened as a result of their geographical distribution, where during the transition from humid tropical to cool subtropical habitats, some NLRs may have become less abundant as a result of reduced selection pressure from tropical pathogens.

Recent studies on banana resistance to bacterial pathogens indicate variable responses and uncover some potential new sources of resistance among the studied plants. For example, three landraces of *Ensete* among 20 studied were found to be resistant to *Xanthomonas vasicola* pv. *musacearum* (Muzemil et al., 2019). A similar study from 72 diverse *Musa* cultivars identified several resistant genotypes, mostly with BB or hybrid AB genome types (Nakato et al., 2019). The B genome is known to be a source of disease resistance in banana breeding programs (Ssekiwoko et al., 2006; Tripathi et al., 2008), though some A genome *Musa* genotypes also showed tolerant reactions to *Xanthomonas vasicola* (Nakato et al., 2019).

In summary, the banana pangenome provides a cross-genus census of the conserved and disposable gene content for banana. We found a core gene content of 18,288 genes across *Musa* and *Ensete* species, while the core gene content for *Musa* alone is 27,858 indicating the extent of the two genomes' divergence. Variable genes were enriched for annotations related to flowering and meristem regulation and we identified candidate regions for drought resistance, meristem initiation, and stress resistance. This information provides the foundation for broader diversity and evolutionary studies and is a resource for the application of genomics for the improvement of these important crops.

ACKNOWLEDGMENTS

This work is funded by the Australia Research Council (Projects LP130100061, DP1601004497, LP160100030, and DP200100762). J.D. has been supported by the ERDF project "Plants as a tool for sustainable global development" No. CZ.02.1.01/0.0/0.0/16_019/0000827. Dr. Philipp Bayer acknowledges the support of the Forrest Research Foundation. This work was undertaken with the assistance of resources provided at the Pawsey Supercomputing Centre.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Habib Rijzaani https://orcid.org/0000-0001-5981-1869
Philipp E. Bayer https://orcid.org/0000-0001-8530-3067
Mathieu Rouard https://orcid.org/0000-0003-0284-1885
Jaroslav Doležel https://orcid.org/0000-0002-6263-0492
Jacqueline Batley https://orcid.org/0000-0002-5391-5824
David Edwards https://orcid.org/0000-0001-7599-6760

REFERENCES

Alexa, A., & Rahnenführer, J. (2009). *Gene set enrichment analysis with topGO*. Bioconductor Improv.

Bayer, P. E., Golicz, A. A., Tirnaz, S., Chan, C. K. K., Edwards, D., & Batley, J. (2019). Variation in abundance of predicted resistance genes in the *Brassica oleracea* pangenome. *Plant Biotechnology Journal*, *17*, 789–800. https://doi.org/10.1111/pbi.13015

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F. C., Falentin, C., Genete, M., Berrabah, W., Chèvre, A. M., Delourme, R., Deniot, G., Denoeud, F., Duffé, P., Engelen, S., Lemainque, A., Manzanares-Dauleux, M., Martin, G., Morice, J., Noel, B., … Aury, J. M. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants*, *4*, 879–887. https://doi.org/10.1038/s41477-018-0289-4

Borrell, J. S., Biswas, M. K., Goodwin, M., Blomme, G., Schwarzacher, T., Heslop-Harrison, J. S., Wendawek, A. M., Berhanu, A., Kallow, S., Janssens, S., & Molla, E. L. (2019). Enset in Ethiopia: A poorly characterized but resilient starch staple. *Annals of Botany*, *123*(5), 747–766.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*, 80–92. https://doi.org/10.4161/fly.19695

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*, 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Dale, J., James, A., Paul, J. Y., Khanna, H., Smith, M., Peraza-Echeverria, S., Garcia-Bastidas, F., Kema, G., Waterhouse, P., Mengersen, K., & Harding, R. (2017). Transgenic Cavendish bananas with resistance to Fusarium wilt tropical race 4. *Nature Communications*, *8*, 1496. https://doi.org/10.1038/s41467-017-01670-6

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

Davey, M. W., Gudimella, R., Harikrishna, J. A., Sin, L. W., Khalid, N., & Keulemans, J. (2013). A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter-and intra-specific *Musa* hybrids. *BMC Genomics*, *14*, 683. https://doi.org/10.1186/1471-2164-14-683

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

Droc, G., Lariviere, D., Guignon, V., Yahiaoui, N., This, D., Garsmeur, O., Dereeper, A., Hamelin, C., Argout, X., Dufayard, J. F., & Lengelle, J. (2013). The banana genome hub. *Database*, *2013*, bat035. https://doi.org/10.1093/database/bat035

Elitzur, T., Yakir, E., Quansah, L., Zhangjun, F., Vrebalov, J., Khayat, E., Giovannoni, J. J., & Friedman, H. (2016). Banana *MaMADS* transcription factors are necessary for fruit ripening and molecular tools to promote shelf-life and food security. *Plant Physiology*, *171*, 380–391. https://doi.org/10.1104/pp.15.01866

Escalant, J. V., Sharrock, S., Frison, E., Carreel, F., Jenny, C., Swennen, R., & Tomekpe, K. (2002). *The genetic improvement of Musa using conventional breeding, and modern tools of molecular and cellular biology*. IPGRI.

FAO. (2018). *Banana market review 2017*. http://www.fao.org/fileadmin/templates/est/COMM_MARKETS_MONITORING/Bananas/Documents/web_Banana_Review_2018_Final_DV.pdf

FAO. (2019). *Banana market review preliminary results for 2019*. http://www.fao.org/3/ca7567en/ca7567en.pdf

Ghag, S. B., Shekhawat, U. K. S., & Ganapathi, T. R. (2015). Small RNA profiling of two important cultivars of banana and overexpression of miRNA156 in transgenic banana plants. *PLoS ONE*, *10*, e0127179. https://doi.org/10.1371/journal.pone.0127179

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., Chan, C. K. K., Severn-Ellis, A., McCombie, W. R., Parkin, I. A. P., Paterson, A. H., Pires, J. C., Sharpe, A. G., Tang, H., Teakle, G. R., Town, C. D., Batley, J., & Edwards, D. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, *7*, 13390. https://doi.org/10.1038/ncomms13390

Golicz, A. A., Martinez, P. A., Zander, M., Patel, D. A., Van De Wouw, A. P., Visendi, P., Fitzgerald, T. L., Edwards, D., & Batley, J. (2015). Gene loss in the fungal canola pathogen *Leptosphaeria maculans*. *Functional & Integrative Genomics*, *15*, 189–196. https://doi.org/10.1007/s10142-014-0412-1

Harrison, J., Moore, K., Paszkiewicz, K., Jones, T., Grant, M., Ambacheew, D., Muzemil, S., & Studholme, D. (2014). A draft genome sequence for *Ensete ventricosum*, the drought-tolerant "tree against hunger." *Agronomy*, *4*, 13–33. https://doi.org/10.3390/agronomy4010013

Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*, 491. https://doi.org/10.1186/1471-2105-12-491

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C. K. K., Tirnaz, S., Dolatabadian, A., Schiessl, S. V., Samans, B., Montenegro, J. D., Parkin, I. A. P., Pires, J. C., Chalhoub, B., King, G. J., Snowdon, R., Batley, J., & Edwards, D. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnology Journal*, *16*, 1265–1274. https://doi.org/10.1111/pbi.12867

Kress, W. J. (1990). The phylogeny and classification of the Zingiberales. *Annals of the Missouri Botanical Garden*, 698–721. https://doi.org/10.2307/2399669

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357. https://doi.org/10.1038/nmeth.1923

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

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

Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S., & You, F. M. (2016). RGAugury: A pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics*, *17*, 852. https://doi.org/10.1186/s12864-016-3197-x

Liu, A. Z., Kress, W. J., & Li, D. Z. (2010). Phylogenetic analyses of the banana family (Musaceae) based on nuclear ribosomal (ITS) and chloroplast (*trnL-F*) evidence. *Taxon*, *59*, 20–28. https://doi.org/10.1002/tax.591003

Liu, A. Z., Kress, W. J., & Long, C. L. (2003). The ethnobotany of *Musella lasiocarpa* (Musaceae), an endemic plant of southwest China. *Economic Botany*, *57*, 279–281. https://doi.org/10.1663/0013-0001(2003)057%5b0279:TEOMLM%5d2.0.CO;2

Martin, G., Baurens, F. C., Droc, G., Rouard, M., Cenci, A., Kilian, A., Hastie, A., Doležel, J., Aury, J. M., Alberti, A., Carreel, F., & D'hont, A. (2016). Improvement of the banana "*Musa acuminata*" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics*, *17*, 243. https://doi.org/10.1186/s12864-016-2579-4

Martin, G., Cardi, C., Sarah, G., Ricci, S., Jenny, C., Fondi, E., Perrier, X., Glaszmann, J. C., d'Hont, A., & Yahiaoui, N. (2020). Genome ancestry mosaics reveal multiple and cryptic contributors to cultivated banana. *The Plant Journal*, *102*, 1008–1025. https://doi.org/10.1111/tpj.14683

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., & Edwards, D. (2017). The pangenome of hexaploid bread wheat. *The Plant Journal*, *90*, 1007–1013. https://doi.org/10.1111/tpj.13515

Muzemil, S., Chala, A., Tesfaye, B., Studholme, D. J., Grant, M., Yemataw, Z., & Olango, T. M. (2019). Evaluation of 20 enset

(*Ensete ventricosum*) landraces for response to *Xanthomonas vasicola* pv. *musacearum* infection. *bioRxiv*, 736793. https://doi.org/10.1101/736793

Nakato, G. V., Christelova, P., Were, E., Nyine, M., Coutinho, T. A., Doležel, J., Uwimana, B., Swennen, R., & Mahuku, G. (2019). Sources of resistance in *Musa* to *Xanthomonas campestris* pv. *musacearum*, the causal agent of banana xanthomonas wilt. *Plant Pathology*, *68*, 49–59. https://doi.org/10.1111/ppa.12945

Oyen, L., & Lemmens, R. (2002). *Plant resources of tropical Africa. Precursor*. PROTA Programme.

Paul, J. Y., Harding, R., Tushemereirwe, W., & Dale, J. (2018). Banana21: From gene discovery to deregulated golden bananas. *Frontiers in Plant Science*, *9*, 558. https://doi.org/10.3389/fpls.2018.00558

Ploetz, R. C., Kema, G. H. J., & Ma, L. J. (2015). Impact of diseases on export and smallholder production of banana. *Annual Review of Phytopathology*, *53*, 269–288. https://doi.org/10.1146/annurev-phyto-080614-120305

Ploetz, R. C., Kepler, A. K., Daniells, J., & Nelson, S. C. (2007). Banana and plantain—An overview with emphasis on Pacific Island cultivars *Musaceae* (banana family). *Species Profiles for Pacific Island Agroforestry*. http://www.bananenzeug.ch/wp-content/uploads/2018/06/banana-plantain-overview.pdf

Robinson, J. C. (1996). *Bananas and plantains*. CAB International.

Rouard, M., Droc, G., Martin, G., Sardos, J., Hueber, Y., Guignon, V., Cenci, A., Geigle, B., Hibbins, M. S., Yahiaoui, N., Baurens, F. C., Berry, V., Hahn, M. W., D'Hont, A., & Yahiaoui, N. (2018). Three new genome assemblies support a rapid radiation in *Musa acuminata* (wild banana). *Genome Biology and Evolution*, *10*, 3129–3140. https://doi.org/10.1093/gbe/evy227

Ruas, M., Guignon, V., Sempere, G., Sardos, J., Hueber, Y., Duvergey, H., Andrieu, A., Chase, R., Jenny, C., Hazekamp, T., Irish, B., Jelali, K., Adeka, J., Ayala-Silva, T., Chao, C. P., Daniells, J., Dowiya, B., Effa effa, B., Gueco, L., … Hazekamp, T. (2017). MGIS: managing banana (*Musa* spp.) genetic resources information and high-throughput genotyping data. *Database*, *2017*, bax046. https://doi.org/10.1093/database/bax046

Sharrock, S., Horry, J., & Frison, E. (2001). The state of the use of *Musa* diversity. In H. D. Cooper, C. Spillane, & T. Hodgkin (Eds.), *Broadening the genetic base of crop production*. (pp. 223–243). IPGRI/FAO.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smit, A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0. http://www.repeatmasker.org

Ssekiwoko, F., Tushemereirwe, W. K., Batte, M., Ragama, P. E., & Kumakech, A. (2006). Reaction of banana germplasm to inoculation with *Xanthomonas campestris pv musacearum*. *African Crop Science Journal*, *14*. https://doi.org/10.4314/acsj.v14i2.46165

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research*, *34*, W435–W439. https://doi.org/10.1093/nar/gkl200

Suzuki, R., & Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, *22*, 1540–1542. https://doi.org/10.1093/bioinformatics/btl117

Thiele, G., Dufour, D., Vernier, P., Mwanga, R. O. M., Parker, M. L., Schulte Geldermann, E., Teeken, B., Wossen, T., Gotor, E., Kikulwe, E., Tufan, H., Sinelle, S., Kouakou, A. M., Friedmann, M., Polar, V., & Hershey, C. (2021). A review of varietal change in roots, tubers and bananas: Consumer preferences and other drivers of adoption and implications for breeding. *International Journal of Food Science & Technology*, *56*, 1076–1092. https://doi.org/10.1111/ijfs.14684

Tripathi, J. N., Oduor, R. O., & Tripathi, L. (2015). A high-throughput regeneration and transformation platform for production of genetically modified banana. *Frontiers in Plant Science*, *6*, 1025. https://doi.org/10.3389/fpls.2015.01025

Tripathi, L., Atkinson, H., Roderick, H., Kubiriba, J., & Tripathi, J. N. (2017). Genetically engineered bananas resistant to *Xanthomonas* wilt disease and nematodes. *Food and Energy Security*, *6*, 37–47. https://doi.org/10.1002/fes3.101

Tripathi, L., Odipio, J., Tripathi, J. N., & Tusiime, G. (2008). A rapid technique for screening banana cultivars for resistance to *Xanthomonas* wilt. *European Journal of Plant Pathology*, *121*, 9–19. https://doi.org/10.1007/s10658-007-9235-4

Tsegaye, A. and Struik, P. C. (2002). Analysis of enset (*Ensete ventricosum*) indigenous production methods and farm-based biodiversity in major enset-growing regions of southern Ethiopia. *Experimental Agriculture*, *38*, 291–315. https://doi.org/10.1017/S0014479702003046

Van Asten, P. J. A., Fermont, A. M., & Taulya, G. (2011). Drought is a major yield loss factor for rainfed East African highland banana. *Agricultural Water Management*, *98*, 541–552. https://doi.org/10.1016/j.agwat.2010.10.005

Wang, Z., Miao, H., Liu, J., Xu, B., Yao, X., Xu, C., Zhao, S., Fang, X., Jia, C., Wang, J., Zhang, J., Li, J., Xu, Y., Wang, J., Ma, W., Wu, Z., Yu, L., Yang, Y., Liu, C., … Jin, Z. (2019). *Musa balbisiana* genome reveals subgenome evolution and functional divergence. *Nature Plants*, *5*, 810–821. https://doi.org/10.1038/s41477-019-0452-6

Wu, W., Yang, Y. L., He, W. M., Rouard, M., Li, W. M., Xu, M., Roux, N., & Ge, X. J. (2016). Whole genome sequencing of a banana wild relative *Musa itinerans* provides insights into lineage-specific diversification of the *Musa* genus. *Scientific Reports*, *6*, 31586. https://doi.org/10.1038/srep31586

Yemataw, Z., Chala, A., Ambachew, D., Studholme, D., Grant, M., & Tesfaye, K. (2017). Morphological variation and inter-relationships of quantitative traits in Enset (*Ensete ventricosum* (welw.) Cheesman) germplasm from South and South-Western Ethiopia. *Plants*, *6*, 56. https://doi.org/10.3390/plants6040056

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

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simão, F. A., Ioannidis, P., Seppey, M., Loetscher, A., & Kriventseva, E. V. (2017). OrthoDB v9. 1: Cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Research*, *45*, D744-D749. https://doi.org/10.1093/nar/gkw1119

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., Wang, Y., Fan, D., Zhao, Y., Wang, Z., Zhou, C., Chen, J., Zhu, C., Li, W., Weng, Q., … Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, *50*, 278–284. https://doi.org/10.1038/s41588-018-0041-z

Zhao, Y., Jia, X., Yang, J., Ling, Y., Zhang, Z., Yu, J., Wu, J., & Xiao, J. (2014). PanGP: A tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics*, *30*, 1297–1299. https://doi.org/10.1093/bioinformatics/btu017

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, *29*, 2669–2677. https://doi.org/10.1093/bioinformatics/btt476

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Rijzaani H, Bayer P E, Rouard, M, Doležel J, Batley, J, Edwards D. The pangenome of banana highlights differences between genera and genomes. *Plant Genome*. 2021;e20100. https://doi.org/10.1002/tpg2.20100