Genome sequencing of *Musa acuminata* Dwarf Cavendish reveals a chromosome arm duplication

Mareike Busche[+], Boas Pucker[+], Prisca Viehöver, Bernd Weisshaar, Ralf Stracke*

Genetics and Genomics of Plants, Center for Biotechnology (CeBiTec), Bielefeld University, Sequenz 1, 33615 Bielefeld, Germany

MB: mbusche@cebitec.uni-bielefeld.de, 0000-0002-2114-7613

BP: bpucker@cebitec.uni-bielefeld.de, 0000-0002-3321-7471

PV: viehoeve@cebitec.uni-bielefeld.de, 0000-0003-3286-4121

BW: bernd.weisshaar@uni-bielefeld.de, 0000-0002-7635-3473

RS: ralf.stracke@uni-bielefeld.de, 0000-0002-9261-2279

+ these authors contributed equally

* corresponding author

Abstract

Different *Musa* species, subspecies, and cultivars are currently investigated to reveal their genomic diversity. Here, we compare the *Musa acuminata* cultivar Dwarf Cavendish against the previously released Pahang assembly. Numerous small sequence variants were detected and the ploidy of the cultivar presented here was determined as triploid based on sequence variant frequencies. Illumina sequencing also revealed a duplication of a large segment of chromosome 2 in the genome of the cultivar studied. Comparison against previously sequenced cultivars provided evidence that this duplication is unique to Dwarf Cavendish. Although no functional relevance of this duplication was identified, this example shows the potential of plants to tolerate such aneuploidies.

Introduction

Bananas (*Musa*) are monocotyledonous perennial plants. The edible fruit (botanically a berry) belongs to the most popular fruits in the world. In 2016, about 5.5 million hectares of land were used for the production of more than 112 million tons of bananas [1]. The majority of bananas were grown in Africa, Latin America, and Asia where they offer employment opportunities and are important export commodities [1]. Furthermore, with an annual *per capita* consumption of more than 200 kg in Rwanda and more than 100 kg in Angola, bananas provide food security in developing countries [1,2]. While plantains or cooking bananas are commonly eaten as a staple food in Africa and Latin America, the softer and sweeter dessert bananas are popular in Europe and Northern America. Between 1998 and 2000, around 47 % of the world banana production and the majority of the dessert banana production relied on the Cavendish subgroup of cultivars [2].

Even though only Cavendish bananas are traded internationally, numerous cultivars are used for local consumption in Africa and South-east Asia. Bananas went through a long domestication process which started at least 7,000 years ago [3]. The first step towards edible bananas was interspecific hybridisation between subspecies from different regions, which caused incorrect meiosis and diploid gametes [4]. The diversity of edible triploids resulted from human selection and triploidization of *Musa acuminata* and *Musa balbisiana* cultivars [4].

These exciting insights into the evolution of bananas were revealed by the analysis of genome sequences. Technological advances boosted sequencing capacities and allowed the (re-)sequencing of genomes from multiple subspecies and cultivars. *M. acuminata* can be divided into several subspecies and cultivars. The first *M. acuminata* (DH Pahang) genome sequence has been published in 2012 [5], many more genomes have been sequenced recently including: banksii, burmannica, zebrina [6], malaccensis (SRR8989632, SRR6996493), Baxijiao (SRR6996491, SRR6996491), Sucrier : Pisang_Mas (SRR6996492). Additionally, the genome sequences of other *Musa* species, *M. balbisiana* [7], *M. itinerans* [8], and *M. schizocarpa* [9], have already been published.

Here we report about our investigation of the genome of *M. acuminata* Dwarf Cavendish which revealed multiple large scale deletions compared to the DH Pahang v2 reference assembly. Moreover, we identified an increased copy number of the southern arm of chromosome 2, indicating that this region was duplicated in one haplophase.

Materials and methods

Plant material and DNA extraction. *Musa acuminata* Dwarf Cavendish tissue culture seedlings were obtained from FUTURE EXOTICS/SolarTek (Düsseldorf, Germany) (Figure 1). Plants were grown under natural daylight at 21°C. Genomic DNA was isolated from leaves following the protocol of Dellaporta *et al.* [10].

Figure 1: ***M. acuminata* Dwarf Cavendish plant.**

Library preparation and sequencing. Genomic DNA was subjected to sequencing library preparation via the TrueSeq v2 protocol as previously described [11]. Paired-end sequencing was performed on an Illumina HiSeq1500 and NextSeq500, respectively, resulting in 2x250 nt and 2x154 nt read data sets.

Read mapping, variant calling, and variant annotation. All reads were mapped to the DH Pahang v2 reference genome sequence via BWA-MEM v0.7 [12]. GATK v3.8 [13,14] was deployed to identify small sequence variants based on this read mapping. The resulting variant set was subjected to SnpEff [15] to assign predictions about the functional impact to all variants. Variants with disruptive effects were selected using a customized Python script as described earlier [11].

The genome-wide distribution of small sequence variants was assessed for small nucleotide variants (SNVs) and InDels based on previously developed scripts [16]. The length distribution

of InDels inside coding sequences was compared to the length distribution of InDels outside coding sequences using a customized Python script [11].

***De novo* genome assembly**. Trimmomatic v0.38 [17] was applied as previously described [11] to remove low quality sequences and remaining adapter sequences. Different sets of trimmed reads were subjected to SOAPdenovo2 for assembly using optimized parameters [18]. Resulting assemblies were evaluated using previously described criteria [18] including BUSCO v3 [19] and polished by removing potential contaminations and adapters as described before [18]. The DH Pahang v2 assembly [5,20] was used in the contamination detection process to distinguish between *bona fide* banana contigs and sequences of unknown origin. Contigs with high sequence similarity to non-plant sequences were removed as previously described [18]. Surviving contigs were sorted based on the DH Pahang v2 reference genome sequence and joint into pseudochromosomes to facilitate downstream analyses.

Data availability. Sequencing datasets were submitted to EBI (ERR3412983, ERR3412984, ERR3413471, ERR3413472, ERR3413473, ERR3413474). Python scripts are freely available on github (https://github.com/bpucker/banana).

Results and discussion

Structural variants. The probably most remarkable difference between the Dwarf Cavendish and Pahang genome sequence is the amplification of the southern region of chromosome 2 (Figure 2). Such a duplication was not observed in any of the other available sequencing data sets (File S1, File S2). There are multiple deletions and insertions which are not randomly distributed over the genome. An increased average coverage in the south of chromosome 2 indicates that a substantial fraction of this chromosome has been duplicated. Apparently, read mapping indicates at least four large scale deletions in Dwarf Cavendish compared to Pahang v2 on chromosomes 2, 4, 5 and 7 (Figure 2). However, analysis of the underlying sequence revealed long stretches of ambiguous bases as the cause for these low coverage regions.

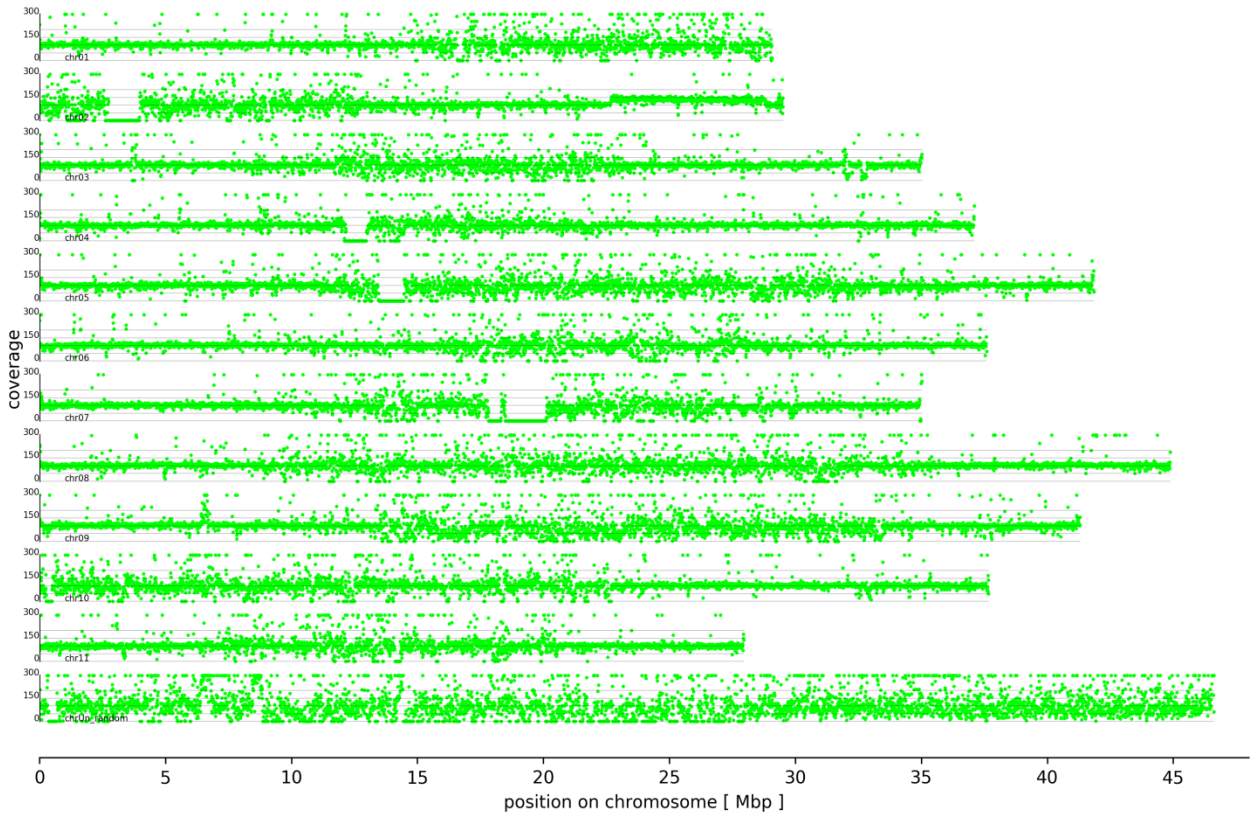

Figure 2: Coverage distribution. Chromosomes are ordered by increasing number with the north end on the left hand side. Mapping of *M. acuminata* Dwarf Cavendish reads against the DH Pahang v2 reference sequence revealed a tetraploid region in the southern part of chromosome 2 in Dwarf Cavendish. Apparent large scale deletions are technical artifacts caused by large stretches of ambiguous bases in the Pahang assembly that cannot be covered by reads.

Ploidy of *M. acuminata* Dwarf Cavendish. Based on the coverage of small sequence variants, the ploidy of Dwarf Cavendish was identified as triploid (Figure 3). Many heterozygous variants display a frequency of the reference allele close to 0.33 or close to 0.66. This can be explained by two copies of the reference allele and one copy of a different allele or *vice versa*. Deviation from the precise values can be explained by random fluctuation. Since the peak around 0.66 is substantially higher than the peak around 0.33, it is possible that two haplophases are very similar to the reference, while one haplophase is different. This hypothesis could be tested in a high continuity phased assembly.

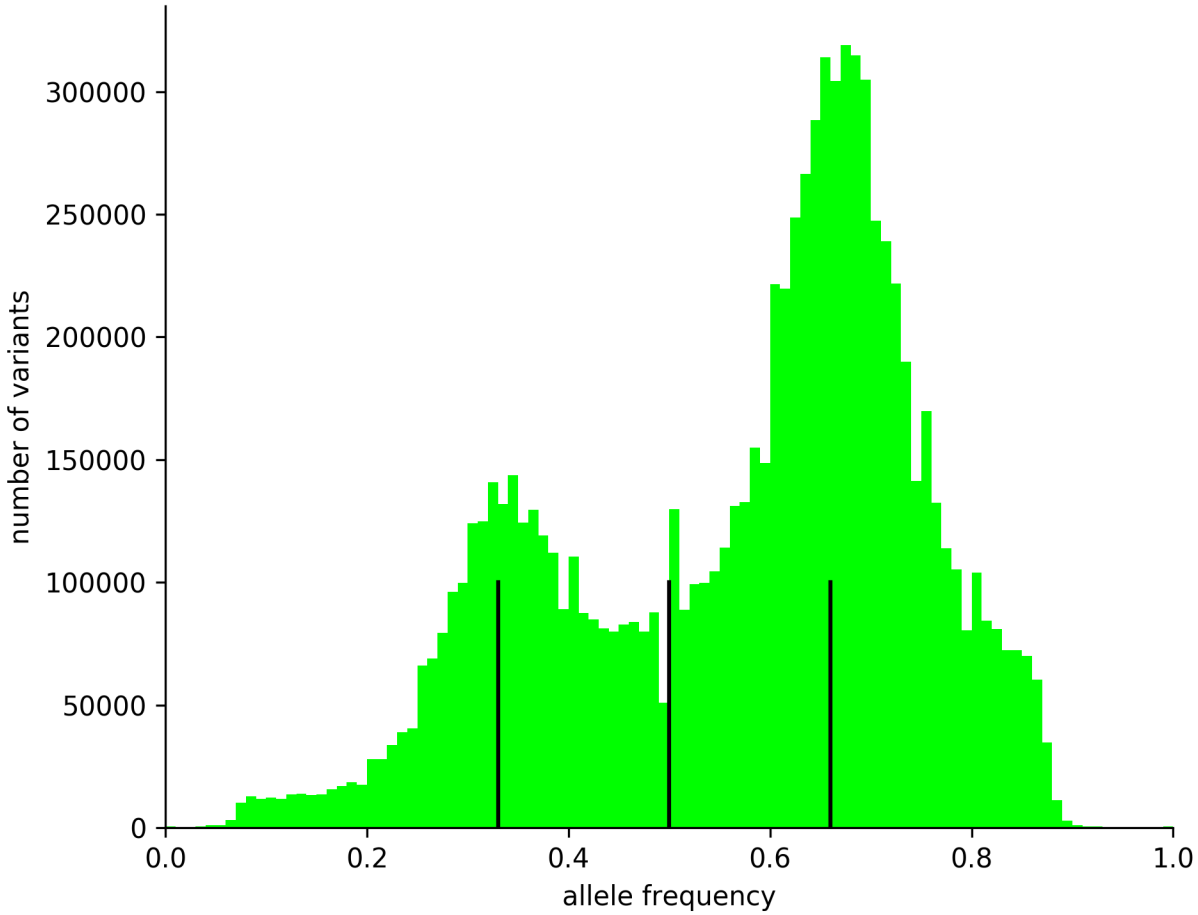

Figure 3: Allele frequency histogram. A mapping of Illumina reads against the Pahang v2 reference sequence was used for the identification of small sequence variants. The frequencies of the reference allele at small variant positions are displayed here, excluding those positions at which the Pahang v2 reference sequence deviates from an invariant sequence position of Dwarf Cavendish. Black vertical lines indicate 0.33, 0.5, and 0.66, respectively.

Genome-wide distribution of small sequence variants. In total, 10,627,677 single nucleotide variants and 1,496,000 InDels of up to 894 bp were identified between the Dwarf Cavendish reads and the Pahang v2 assembly (http://doi.org/10.4119/unibi/2936278). As previously observed in other re-sequencing studies [11], the number of SNVs exceeds the number of InDels substantially. Moreover, InDels are more frequent outside of annotated coding regions. Inside coding regions, InDels show an increased proportion of lengths which are divisible by 3, because no deleterious frameshifts are caused.

SnpEff predicted 4,184 premature stop codons, 3,252 lost stop codons, and 8,159 frameshifts based on this variant set (File S3). Even given the larger genome size, these numbers are substantially higher than high impact variant numbers observed in re-sequencing studies of homozygous species before [11,21]. One explanation could be the presence of three alleles for

each locus leading to compensation of disrupted alleles. Since banana plants are propagated vegetatively, breeders do not suffer inbreeding depressions.

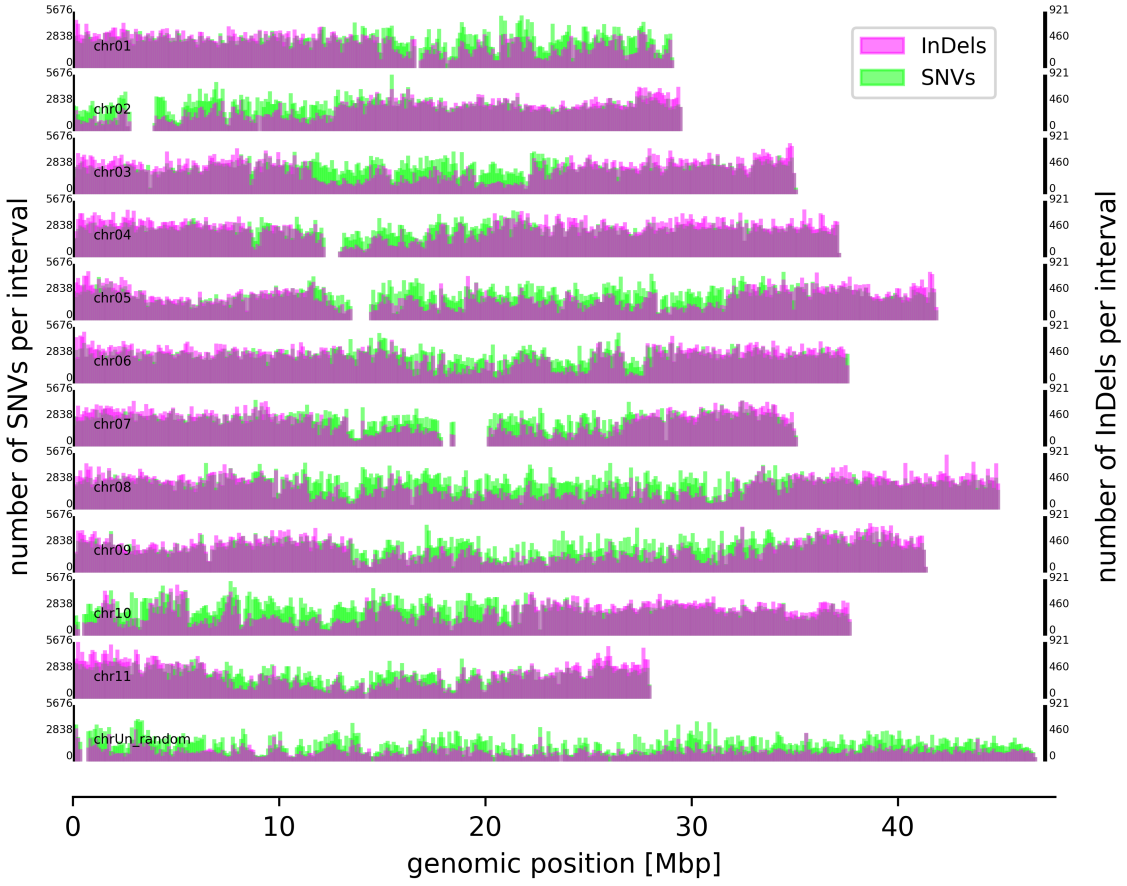

Figure 4: Genome-wide distribution of small sequence variants. Single nucleotide variants (SNVs, green) and small InDels (magenta) distinguish Dwarf Cavendish from Pahang. Variants were counted in 100 kb windows and are displayed on two different y-axes to allow maximal resolution [11].

***De novo* genome assembly**. To facilitate wet lab applications like oligonucleotide design and validation of amplicons, the genome sequence of Dwarf Cavendish was assembled *de novo*. The assembly comprises 404,889 contigs with an N50 of 4.7 kb (Table 1). Differences between the three haplophases are one possible explanation for the low assembly contiguity. Since the short reads are insufficient to resolve entire haplophases, the assembly size is only slightly exceeding the size of one haplotype. Due to the low contiguity of this assembly and only minimal above 50% complete BUSCOs (Benchmarking Universal Single-Copy Orthologs) [19], annotation was omitted.

Table 1: *M. acuminata* Dwarf Cavendish *de novo* genome assembly statistics.

Parameter	Value
Number of contigs	404,889
Maximal contig length	240,314 bp
Assembly size	1,068,694,790 bp
GC content	38.87 %
N50	4,709
N90	1,006

Acknowledgments

We thank Joachim Weber for great technical assistance.

Author contributions

BP, MB and RS planned the experiment. PV did the library preparation and sequencing. BP performed bioinformatic analyses. MB and BP wrote the initial draft. MB, BP, BW and RS revised the manuscript. All authors read and approved the final manuscript version.

References

1. FAO. FAOSTAT [Internet]. [cited 23 Jun 2019]. Available: http://www.fao.org/faostat/en/#data/QC

2. Arias P, Dankers C, Liu P, Pilkauskas P. The World Banana Economy, 1985-2002. Food and Agriculture Organisation (FAO) of the United Nations 2003. Available: http://www.fao.org/3/y5102e/y5102e00.htm

3. Denham TP, Haberle SG, Lentfer C, Fullagar R, Field J, Therin M, et al. Origins of Agriculture at Kuk Swamp in the Highlands of New Guinea. Science. 2003;301: 189–193. doi:10.1126/science.1085255

4. Perrier X, Langhe ED, Donohue M, Lentfer C, Vrydaghs L, Bakry F, et al. Multidisciplinary perspectives on banana (Musa spp.) domestication. Proc Natl Acad Sci. 2011;108: 11311–11318. doi:10.1073/pnas.1102001108

5. D'Hont A, Denoeud F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature. 2012;488: 213–217. doi:10.1038/nature11241

6. Rouard M, Droc G, Martin G, Sardos J, Hueber Y, Guignon V, et al. Three New Genome Assemblies Support a Rapid Radiation in Musa acuminata (Wild Banana). Genome Biol Evol. 2018;10: 3129–3140. doi:10.1093/gbe/evy227

7. Davey MW, Gudimella R, Harikrishna JA, Sin LW, Khalid N, Keulemans J. "A draft Musa balbisiana genome sequence for molecular genetics in polyploid, inter- and intra-specific Musa hybrids." BMC Genomics. 2013;14: 683. doi:10.1186/1471-2164-14-683

8. Wu W, Yang Y-L, He W-M, Rouard M, Li W-M, Xu 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. doi:10.1038/srep31586

9. Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants. 2018;4: 879. doi:10.1038/s41477-018-0289-4

10. Dellaporta SL, Wood J, Hicks JB. A plant DNA minipreparation: Version II. Plant Mol Biol Report. 1983;1: 19–21. doi:10.1007/BF02712670

11. Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B. A *De Novo* Genome Sequence Assembly of the *Arabidopsis thaliana* Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny. PLOS ONE. 2016;11: e0164321. doi:10.1371/journal.pone.0164321

12. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. 2013; Available: http://arxiv.org/abs/1303.3997

13. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

14. Auwera GAV der, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy□Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc Bioinforma. 2013;43: 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

15. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6: 80–92. doi:10.4161/fly.19695

16. Baasner J-S, Howard D, Pucker B. Influence of neighboring small sequence variants on functional impact prediction. bioRxiv. 2019; 596718. doi:10.1101/596718

17. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

18. Pucker B, Feng T, Brockington S. Next generation sequencing to investigate genomic diversity in Caryophyllales. 2019; doi:https://doi.org/10.1101/646133

19. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma Oxf Engl. 2015;31: 3210–3212. doi:10.1093/bioinformatics/btv351

20. Martin G, Baurens F-C, Droc G, Rouard M, Cenci A, Kilian A, et al. Improvement of the banana “Musa acuminata” reference sequence using NGS data and semi-automated bioinformatics methods. BMC Genomics. 2016;17: 243. doi:10.1186/s12864-016-2579-4

21. Xu Y-C, Niu X-M, Li X-X, He W, Chen J-F, Zou Y-P, et al. Adaptation and Phenotypic Diversification in Arabidopsis through Loss-of-Function Mutations in Protein-Coding Genes. Plant Cell. 2019;31: 1012–1025. doi:10.1105/tpc.18.00791

Additional files

File S1. List of public sequence read samples used for comparison against Dwarf Cavendish based on the DH Pahang reference.

File S2. Coverage plots of public sequence read samples for comparison against Dwarf Cavendish based on the DH Pahang reference. Order of samples is: *Musa acuminata*, AYP_BOSN_r1, *Musa banksia*, *Musa burmannica*, Cavendish BaXiJiao, Gros Michel, *Musa malaccensis*, Sucrier : Pisang Mas, sucrier pisang mas 1998-2307, *Musa zebrina*, Pisang_Klutuk_Wulung, *Musa itinerans*, *Musa schizocarpa*

File S3. Selected high impact variants between Dwarf Cavendish and DH Pahang with effects predicted by SnpEff.