Agronomy **2014**, *4*, 13-33; doi:10.3390/agronomy4010013

OPEN ACCESS

agronomy

ISSN 2073-4395

www.mdpi.com/journal/agronomy

Article

A Draft Genome Sequence for *Ensete ventricosum*, the Drought-Tolerant "Tree Against Hunger"

James Harrison [1], Karen A. Moore [1], Konrad Paszkiewicz [1], Thomas Jones [1], Murray R. Grant [1], Daniel Ambacheew [2], Sadik Muzemil [2] and David J. Studholme [1],*

[1] College of Life and Environmental Sciences, University of Exeter, Geoffrey Pope Building, Stocker Road, Exeter EX4 4QD, UK; E-Mails: jh288@exeter.ac.uk (J.H.); K.A.Moore@exeter.ac.uk (K.A.M.); k.h.paszkiewicz@exeter.ac.uk (K.P.); tj234@exeter.ac.uk (T.J.); m.r.grant@exeter.ac.uk (M.R.G.)

[2] Southern Agricultural Research Institution (SARI). P.O. Box. 06, Hawassa, Ethiopia; E-Mails: ethiodan@gmail.com (D.A.); croprch@sari.gov.et (S.M.)

* Author to whom correspondence should be addressed; E-Mail: d.j.studholme@exeter.ac.uk; Tel.: +44-01392-72-4678; Fax: +44-1392-371-859.

Received: 15 November 2013; in revised form: 9 January 2014 / Accepted: 14 January 2014 / Published: 17 January 2014

Abstract: We present a draft genome sequence for enset (*Ensete ventricosum*) available via the Sequence Read Archive (accession number SRX202265) and GenBank (accession number AMZH01. Enset feeds 15 million people in Ethiopia, but is arguably the least studied African crop. Our sequence data suggest a genome size of approximately 547 megabases, similar to the 523-megabase genome of the closely related banana (*Musa acuminata)*. At least 1.8% of the annotated *M. acuminata* genes are not conserved in *E. ventricosum*. Furthermore, enset contains genes not present in banana, including reverse transcriptases and virus-like sequences as well as a homolog of the RPP8-like resistance gene. We hope that availability of genome-wide sequence data will stimulate and accelerate research on this important but neglected crop.

Keywords: enset; Ethiopia; drought-tolerance; Musaceae

1. Introduction

Enset (*Ensete ventricosum*) is one of the most important crop plants grown in Ethiopia, where it makes a major contribution of to the food security of the country, feeding at least 15 million people. It buffers food deficit during dry spells and recurrent drought and has been dubbed as the "tree against hunger" [1]. Enset is a multi-purpose crop, with all parts of the plant being utilized for human food, animal forage, medicine, or ornamental uses [2]. Furthermore, it has the capacity for high yield, can be stored for long periods, can be harvested at any time of the year and at any stage over a period of several years [3], thereby offering advantages over seasonal crops.

The genus *Ensete* falls within the botanical family Musaceae, which also includes bananas and plantains (genus *Musa*). Enset is susceptible to some of the same diseases that threaten banana, including bacterial wilt caused by *Xanthomonas campestris* pathovar *musacearum* [4]. Unlike banana, the main edible parts of the enset plant are the starchy corm and pseudostem. The genome of enset is diploid with $n = 9$ [5], while the recently published doubled-haploid banana genome sequence has $n = 11$ [6].

There are many clones and landraces of enset in Ethiopia [1,3]. A collection of more than 600 clones and landraces from major enset growing areas of Ethiopia has been assembled and conserved *ex situ* by the Southern Agricultural Research Institute at Areka and some of these differ in important agronomic characteristics and tolerance to disease [7]. Some attempts at molecular characterization of enset clones or landraces have been made using amplified fragment length polymorphism AFLP [8,9] and random amplified polymorphic DNA RAPD techniques [10,11], revealing the existence of genetic diversity and, therefore, the potential for improvement by breeding, if suitable markers were available. However, despite its importance and value, enset has been relatively neglected by scientific research and is arguably the least-studied African crop. There is an urgent need for efficient improvement of this crop. Our aim was to help accelerate enset research and crop improvement by providing draft genome sequence data and identifying single-nucleotide polymorphisms (SNPs) that might serve as molecular markers for marker-assisted breeding. We also aimed to investigate genetic similarity between enset and banana thus to assess the usefulness of banana genomic resources for application to enset.

2. Results and Discussion

2.1. Whole-Genome Sequencing

We generated 40.4 gigabases of whole-genome shotgun sequence data from the enset genome consisting of 202 million pairs of 100-nucleotide Illumina sequence reads. The sequence reads are freely available from the Sequence Read Archive under accession number SRX202265. Our approach was similar to that of Davey and colleagues [12] who recently re-sequenced the banana B genome (*M. balbisiana*) using 281 million pairs of 100-nucleotide Illumina sequence reads. Their attempt at *de novo* assembly yielded a highly fragmented genome assembly consisting of a large number of short contigs. However, they were able to gain insights into the B genome by aligning their sequence reads against the previously sequenced A genome (*M. acuminata*) and calling a consensus alignment [12]. Likewise, we used both *de novo* sequence assembly (that is, without using a reference genome sequence) and an approach based upon alignment of reads against the banana A-genome reference sequence as described in the sections below. Our aligned enset genomic sequence reads covered 47% of the *M. acuminata* reference genome sequence (247 out of 523 Mb). This is less than the coverage by Davey and colleagues' alignment of *M. balbisiana* reads against the same reference genome, which covered 341 out of 523 Mb (65%), perhaps not surprisingly given the larger evolutionary distance between enset and the *Musa* species.

To check for contamination, we aligned our enset genomic sequence reads against all of the 2735 available complete prokaryotic genomes [13] using the Burrows-Wheeler Aligner BWA [14]. We found that 8.27% of our sequence reads were alignable against prokaryotic bacterial sequences. The genome sequences showing the greatest coverage were *Pseudomonas fluorescens* SBW25 [15] and *Methylobacterium radiotolerans* JCM 2831 ([16], GenBank: CP001001) with sequence reads covering 30.6% and 33.5% of the lengths of their genomes, respectively. These prokaryotic sequences possibly originate from endophytes and/or epiphytes associated with the plant even though we attempted to clean and sterilize the surface of the plant material by wiping with ethanol. We note that in the study by Davey and colleagues [12] there was also some bacterial sequence present in the *M. balbisiana* genomic re-sequencing data: 3.03% of Davey's data aligned to the prokaryotic genome sequences, with coverage of 94.3% of the *Propionibacterium acnes* 266 [17] chromosome, and 60.8% of the *Serratia marcescens* WW4 [18] chromosome. Therefore, it seems that bacterial contamination of plant genome sequence data is not unique to our study. We also note that the depth of coverage of any single bacterial genome by "plant" genomic reads is very low: no more than 2.03× for the *P. fluorescens* and *M. radiotolerans* genomes and no more than 9.1× for the *P. acnes* and *S. marcescens* genomes mentioned above, and, therefore, not enough to be effectively assembled *de novo*.

2.2. Estimation of the Enset Genome Length

Based on alignment against enset nuclear DNA sequences available in the GenBank database (Table 1), we estimate the depth of coverage as 67.67×. Given that we generated a total of 37.05 gigabases of sequence data (after removing prokaryote-matching reads) this would indicate a genome size of approximately 547 megabases. This is close to the haploid genome size of 523 megabases for the closely related *M. acuminata* [6].

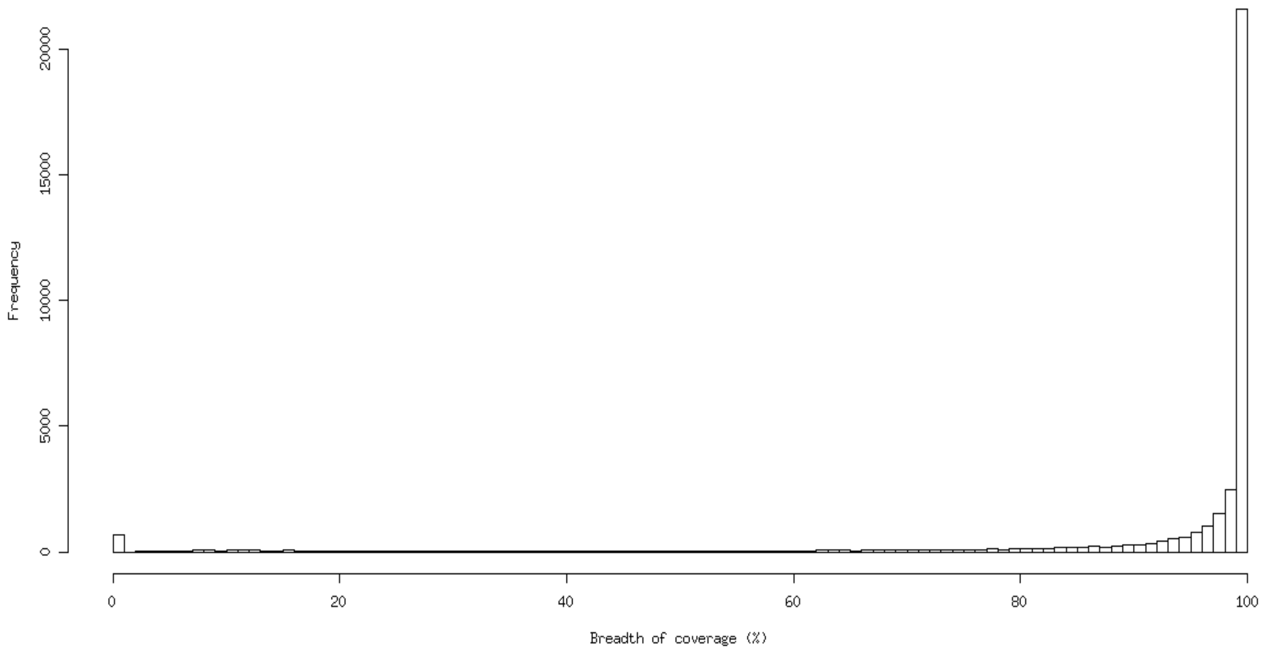
2.3. Conservation of Protein-Coding Sequences between Enset and Banana

To identify which banana protein-coding genes are conserved in enset, we aligned our enset shotgun sequence reads against the 36,542 *M. acuminata* coding sequences identified by D'Hont and colleagues [6] using BWA [19]. The advantage of this approach is that it is not confounded by incomplete assembly of or gene prediction in the enset data. The frequency distribution for breadth of coverage across these 36,542 sequences is shown in Figure 1. The breadths of coverage follow a bi-modal distribution with peaks close to zero and close to 100% coverage. The peak close to zero corresponds to banana genes that are either absent from the enset genome or else they are so divergent that the corresponding enset sequences fail to align. There are 662 (1.8%) banana protein-coding sequences that have zero coverage by the aligned enset data and are, therefore, absent, or very divergent, in enset. The Supplementary Data includes a spreadsheet indicating the breadths of coverage of each *M. acuminata* gene.

Table 1. Depths of coverage of previously published enset nuclear DNA sequences. The median depth of coverage is 67.67 times.

GenBank accession number and description	Depth
HM118700.1 TCP-1-eta subunit gene	80.71
HM118740.1 mRNA capping enzyme large subunit family protein gene	79.26
HM118605.1 electron transport protein gene	79.06
HM118577.1 ATP:citrate lyase gene	75.76
HM118779.1 succinoaminoimidazole-carboximide ribonucleotide synthetase family	74.08
HM118753.1 methylcrotonyl-CoA carboxylase beta chain-like gene	72.01
HM118766.1 annexin-like protein gene	71.61
HM118805.1 initiation factor 2B family protein gene	68.05
HM118660.1 zeaxanthin epoxidase gene	67.67
HM118646.1 CASP protein-like gene, partial sequence	65.98
HM118632.1 endoribonuclease dicer protein-like gene, partial sequence	65.39
HM118673.1 Na/H antiporter gene	65.16
HM118591.1 stomatal cytokinesis defective protein gene	64.52
HM118819.1 DNA polymerase delta catalytic subunit gene	63.05
HM118713.1 NAD+ synthase domain protein gene	61.95
HM118619.1 non-phototropic hypocotyl 3-like gene, partial sequence	61.72
HM118686.1 DUF89 family protein gene	57.14

Figure 1. Frequency distribution for breadth of coverage on 36,542 banana gene sequences by enset whole-genome shotgun sequence reads aligned against the banana genome using BWA.



2.4. Heterozygosity and Single-Nucleotide Polymorphisms (SNPs)

Single-nucleotide polymorphisms (SNPs) can be valuable markers for crop improvement [20] but have not previously been reported for enset. Given the very fragmented nature of our *de novo* assembly of the enset genome, we followed the example of Davey and colleagues [12] by performing SNP calling against the high-quality reference genome sequence of *M. acuminata* [6]. To do the alignment, we used BWA [14] and only considered sequence reads that uniquely align to a single genomic location. By aligning the enset shotgun sequence reads against this banana genome sequence, we were able to identify 30,287 sites at which there was an approximately 50:50 ratio between the two most frequent aligned nucleotides (where the most abundant base accounts for between 49% and 51% of the aligned bases and where coverage is at least 10×). These sites are distributed over the whole genome (see Figure 2) and occur on average every 17.3 kb. If we are less stringent and include all sites where the frequency of the most abundant base is between 48% and 52%, then the number of heterozygous sites increases to 76,416, a density of one site per 6.8 kb of banana genome. See Figure 3 for an example of such a locus, containing three heterozygous sites. See the Supplementary Data for a list of these heterozygous sites. The rationale for using the banana genome as a reference sequence for identifying heterozygous SNPs is that the banana reference genome sequence is much more contiguous and better annotated than the enset *de novo* genome sequence. However, one limitation of this approach is that it will fail to identify heterozygous sites that fall within enset-specific sequences. We found that alignment between enset genomic sequences reads and the banana reference genome sequence covered only 47% of the banana genome and occurred much more frequently in genes rather than intergenic regions, as also observed by Davey and colleagues [12] for alignment of *M. balbisiana* genomic reads against the same reference genome. To circumvent this limitation, we also generated lists of heterozygous sites called on the enset *de novo* assembly; these can be found in the Supplementary Data.

2.5. De Novo *Assembly of the Enset Genome Sequence*

Although alignment of raw sequence reads against the banana reference genome sequence is useful for identifying SNPs and sequences conserved between both plant species, we required a *de novo* assembly of the enset data in order to examine gene order and to identify enset sequences that are not present in the banana genome. Our assembly had a total length of 459.5 megabases. This represents 84% of the estimated enset genome-size of 547 megabases and is 97.3% of the length of the recently published banana genome assembly of 472.2 megabases [6]. Given that our estimate of the enset genome size based on sequence coverage is very approximate and assuming that the enset genome is of similar size to the banana genome, then this suggests that our *de novo* assembly represents nearly complete coverage of the enset genome.

Figure 2. Positions on the banana genome that display heterozygosity in enset. The horizontal axis indicates position on the chromosome and the vertical axis indicates the frequency of the most common base (A, C, G, or T). Only those sites are shown at which there is at least 10× coverage and at which the frequency of the most abundant base is between 49% and 51% inclusive.

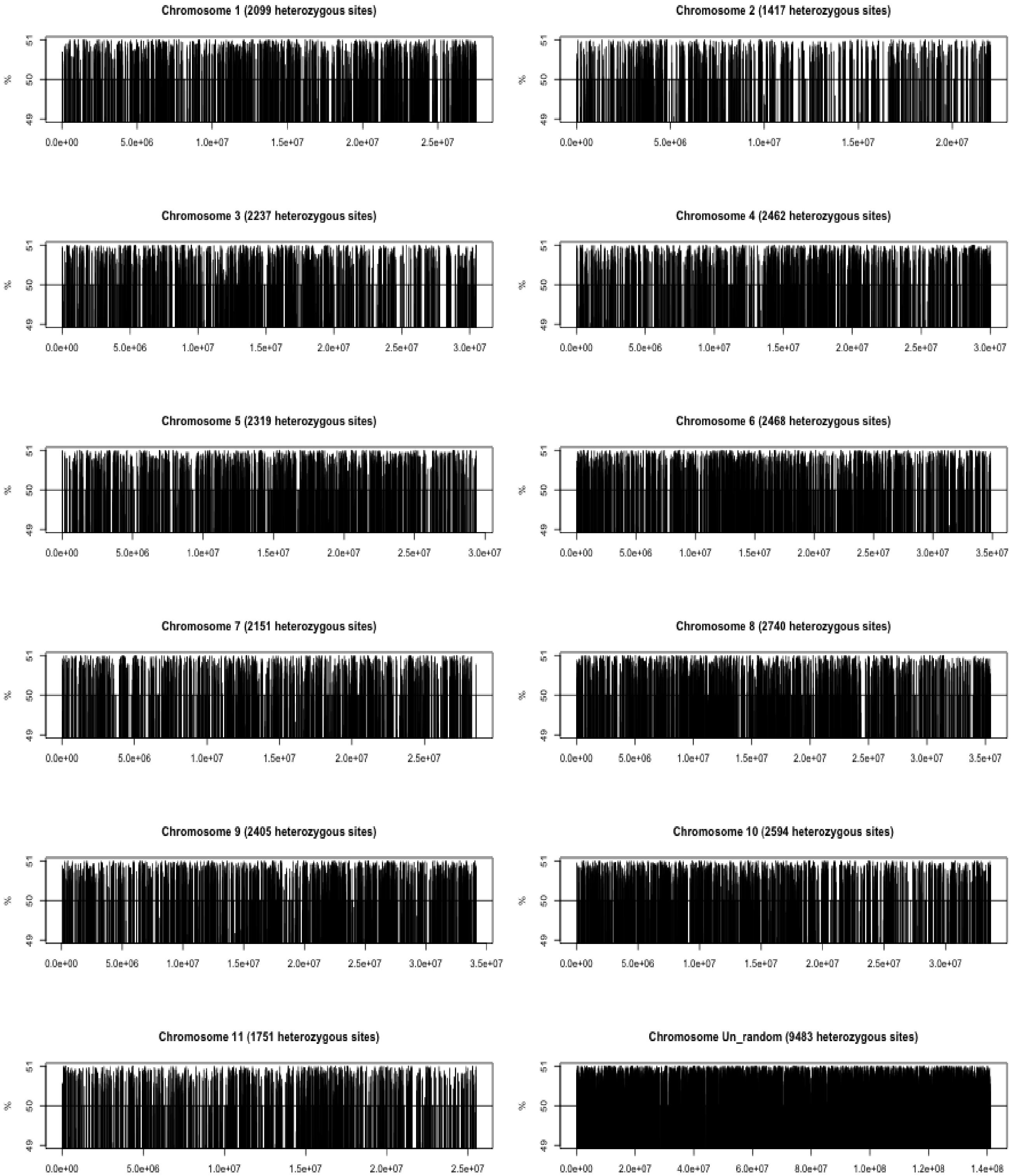
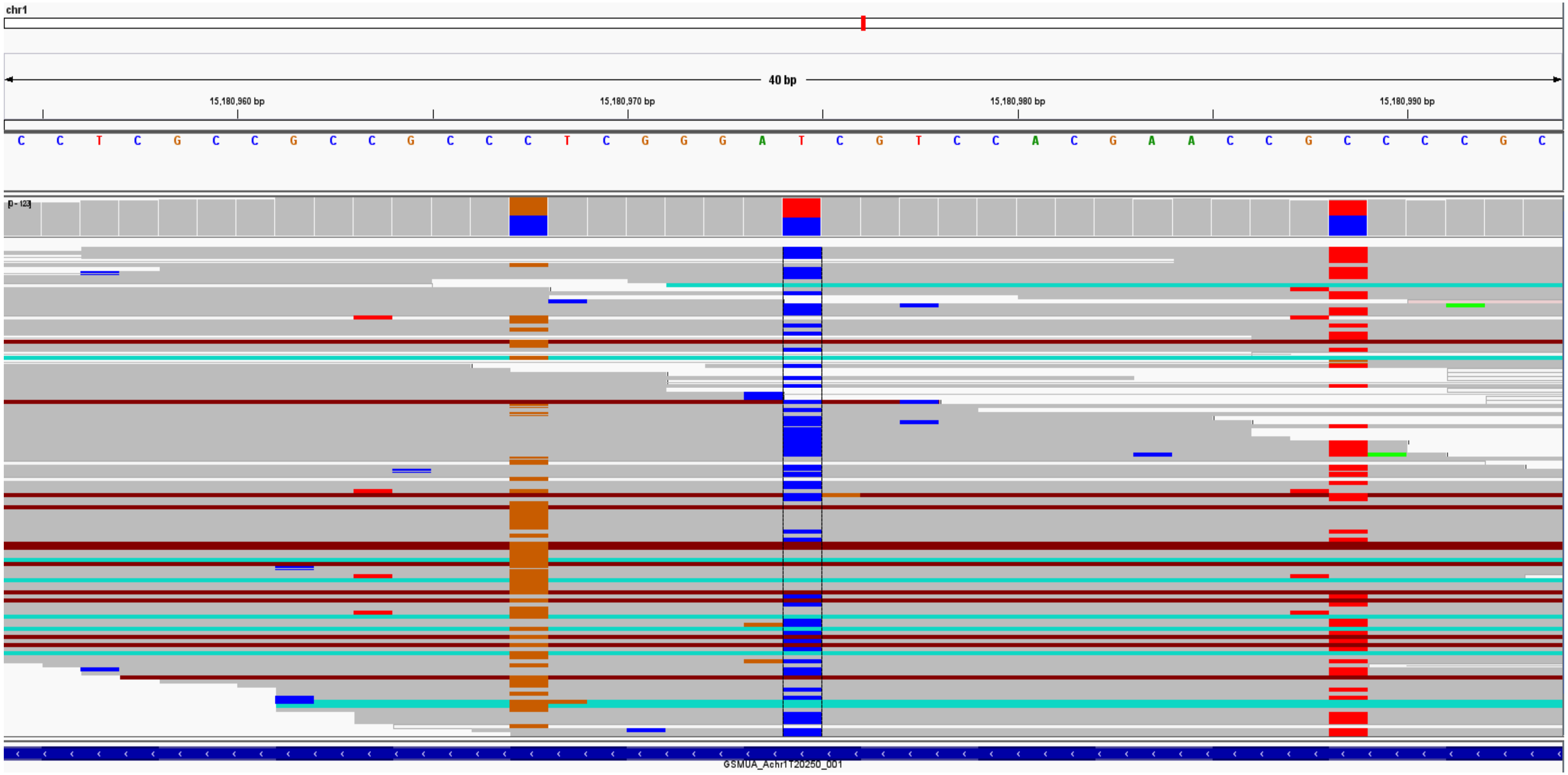


Figure 3. Example of a protein-coding gene that is heterozygous in enset. We aligned enset genomic sequence reads against the banana genome using BWA. The figure shows a 40-nucleotide region of the alignment falling within a protein-coding gene (GSMUA_Achr1T20250_001), encoding a predicted acyl-transferase. This region includes three single-nucleotide polymorphisms, at which the enset genome sequence is heterozygous with approximately 50:50 frequencies for two haplotypes (C…C…T and G…T…C).



The enset genome sequence assembly is available via the GenBank database under accession number AMZH01. Due to restrictions on the numbers of contigs and supercontigs that GenBank can accept within a whole-genome shotgun project, GenBank only includes the enset contigs and super-contigs that are at least five kilobases in length. The full assembly, including contigs and super-contigs of between 200 and 5000 nucleotides, is available via Figshare [21]. Approximately 70% of the enset genome assembly is alignable against the banana genome sequence and average nucleotide sequence identity is 89.90% over the alignable sequence, as judged by the *dnadiff* tool in the MUMmer [22] software package.

Given that about 8% of our genomic sequence reads actually originated from prokaryotes rather than from the plant, we checked our *de novo* assembly for prokaryotic sequences by performing Basic Local Alignment Search Tool nucleotide (BLASTN) searches against the 2735 available complete prokaryotic genomes [13]. A total of 81,795 bp (0.018%) of the enset *de novo* assembly matched prokaryotic genome sequences. These sequences were removed from the data submitted to GenBank (accession AMZH01).

We performed a preliminary annotation of the enset genome assemblies using FGENESH [23] to predict protein-coding genes; summary statistics are given in Table 2 and the protein sequences, their genomic coordinates, results of BLASTP searches against the *M. acuminata* proteome, and the results of functional prediction using PfamScan [24] are available via Figshare [21] (the file was too large to be included in the Supplementary Data). Of 42,749 predicted proteins, 9967 did not have any significant sequence similarity to the banana proteome detectable by BLASTP. It should be noted that due to the fragmented nature of the draft *de novo* assembly, the number of predicted genes is likely to be significantly over-estimated as some gene models are split between multiple contigs. We used RfamScan [25] to identify non-coding RNA genes, including microRNAs, which are listed in Table 3, and we used RepeatMasker [26] to search for matches to repeat sequences (Table 4), as described in the Experimental Section. Overall, the enset assembly was predicted to have a greater repeat-content (32.65%) than the banana A genome (20.31%).

Gene order was highly conserved between banana and enset, at least over the scale of tens of kilobases, as exemplified in Figure 4, which shows an alignment of the longest enset super-contig against banana chromosome 5. However, we did identify some differences in gene-content between the two genomes as described in the following sections.

Table 2. Assembly statistics.

	Complete assembly	**Subset of assembly submitted to GenBank (AMZH00000000.1)**
Number of scaffolds	123,779	14,787
N_{50} scaffold length	11,149	13,657
NG_{50} scaffold length (bp)	9,954	n.a. *
Shortest scaffold (bp)	200	5,000
Longest scaffold (bp)	105,416	103,995
Sum of scaffold lengths (bp)	458,655,998	172,241,963
Mean scaffold length (bp)	3,705	15,952
Median scaffold length (bp)	1,056	13,404
Number of contigs	259,028	19,109
N_{50} contig length (bp)		8,724
NG_{50} contig length (bp)	2,428	n.a. *
Shortest contig (bp)	201	5,000
Longest contig (bp)	56,178	56,178
Sum of contig lengths (bp)	390,884,093	163,735,150
Mean contig length (bp)	1,509	8,568
Median contig length (bp)	555	7,448
Number of gene models	42,749	23,423
Mean length of predicted protein (aa)	311.64	353.84
G + C (%)	38.95	39.14

* NG_{50} lengths [27] were calculated on the basis of an estimated genome length of 50 Mb. The total length of the scaffolds submitted to GenBank (under accession AMZH00000000.1) was less than 50% of this estimated length (7.54 Mb *versus* 25 Mb); therefore, it is not possible to calculate NG_{50} length for this dataset.

Table 3. Predicted non-coding RNAs in the enset genome assembly predicted by Rfam version 11.

GenBank accession number	**Scaffold name**	**Start and end positions**	**Strand**	**Rfam ID (and accession number)**	**Rfam scan E value**
KB218331.1	scf_22030_17941	4842–4920	+	Intron_gpII (RF00029)	$2.89e^{-04}$
KB218832.1	scf_22030_39767	2365–2435	−	Intron_gpII (RF00029)	$3.47e^{-08}$
KB218412.1	scf_22030_21016	944–1028	+	mir-156 (RF00073)	$7.66e^{-17}$
KB220497.1	scf_22030_77035	4888–4971	−	mir-156 (RF00073)	$1.34e^{-17}$
KB220497.1	scf_22030_77035	4888–4971	+	mir-156 (RF00073)	$4.11e^{-09}$
KB220618.1	scf_22030_78211	2918–3003	−	mir-156 (RF00073)	$1.57e^{-14}$
KB220618.1	scf_22030_78211	2918–3003	+	mir-156 (RF00073)	$8.68e^{-09}$
KB220859.1	scf_22030_80462	10702–10791	+	mir-156 (RF00073)	$1.65e^{-17}$
KB220859.1	scf_22030_80462	10702–10791	−	mir-156 (RF00073)	$7.33e^{-09}$
KB220860.1	scf_22030_80478	14044–14147	+	mir-156 (RF00073)	$3.70e^{-17}$
KB220947.1	scf_22030_81257	2331–2413	+	mir-156 (RF00073)	$2.41e^{-16}$
KB220073.1	scf_22030_72447	11922–12159	−	MIR159 (RF00638)	$1.44e^{-35}$
KB220073.1	scf_22030_72447	11924–12161	+	MIR159 (RF00638)	$9.81e^{-22}$

Table 3. *Cont.*

GenBank accession number	Scaffold name	Start and end positions	Strand	Rfam ID (and accession number)	Rfam scan E value
KB220655.1	scf_22030_78562	4140–4330	−	MIR159 (RF00638)	$1.15e^{-37}$
KB220655.1	scf_22030_78562	4142–4332	+	MIR159 (RF00638)	$2.01e^{-21}$
KB218508.1	scf_22030_25031	13232–13319	+	mir-160 (RF00247)	$3.76e^{-23}$
KB218508.1	scf_22030_25031	13231–13319	−	mir-160 (RF00247)	$1.52e^{-09}$
KB219059.1	scf_22030_50116	8622–8711	+	mir-160 (RF00247)	$3.16e^{-23}$
KB219059.1	scf_22030_50116	8622–8711	−	mir-160 (RF00247)	$1.35e^{-11}$
KB218046.1	scf_22030_5366	30669–30758	−	mir-160 (RF00247)	$7.21e^{-21}$
KB218046.1	scf_22030_5366	30669–30756	+	mir-160 (RF00247)	$3.20e^{-08}$
KB219346.1	scf_22030_59171	24014–24101	+	mir-160 (RF00247)	$1.18e^{-20}$
KB219346.1	scf_22030_59171	24014–24101	−	mir-160 (RF00247)	$6.30e^{-09}$
KB218895.1	scf_22030_42834	6184–6270	−	MIR164 (RF00647)	$5.38e^{-19}$
KB218895.1	scf_22030_42834	6184–6270	+	MIR164 (RF00647)	$3.11e^{-12}$
KB219508.1	scf_22030_63187	11271–11378	+	MIR164 (RF00647)	$1.12e^{-18}$
KB219508.1	scf_22030_63187	11271–11378	−	MIR164 (RF00647)	$1.02e^{-12}$
KB218104.1	scf_22030_8363	10326–10443	−	MIR164 (RF00647)	$6.46e^{-23}$
KB218104.1	scf_22030_8363	10326–10443	+	MIR164 (RF00647)	$6.71e^{-16}$
KB217991.1	scf_22030_2485	3315–3401	−	mir-166 (RF00075)	$5.93e^{-21}$
KB217991.1	scf_22030_2485	3315–3401	+	mir-166 (RF00075)	$2.53e^{-10}$
KB218022.1	scf_22030_4161	21528–21639	+	mir-166 (RF00075)	$3.99e^{-20}$
KB218022.1	scf_22030_4161	21528–21639	−	mir-166 (RF00075)	$1.31e^{-10}$
KB219071.1	scf_22030_50479	2432–2530	−	mir-166 (RF00075)	$2.04e^{-22}$
KB219071.1	scf_22030_50479	2432–2530	+	mir-166 (RF00075)	$1.27e^{-12}$
KB219643.1	scf_22030_65797	40153–40244	−	mir-166 (RF00075)	$2.40e^{-22}$
KB219643.1	scf_22030_65797	40153–40244	+	mir-166 (RF00075)	$9.30e^{-12}$
KB220445.1	scf_22030_76496	6198–6315	−	mir-166 (RF00075)	$2.47e^{-23}$
KB220445.1	scf_22030_76496	6198–6315	+	mir-166 (RF00075)	$5.31e^{-12}$
KB220707.1	scf_22030_79012	6213–6322	−	mir-166 (RF00075)	$2.17e^{-24}$
KB220707.1	scf_22030_79012	6213–6322	+	mir-166 (RF00075)	$8.47e^{-13}$
KB221155.1	scf_22030_81490	17577–17697	+	mir-166 (RF00075)	$6.47e^{-17}$
KB221155.1	scf_22030_81490	17577–17697	−	mir-166 (RF00075)	$4.00e^{-08}$
KB218667.1	scf_22030_31606	22038–22152	+	MIR167_1 (RF00640)	$6.27e^{-22}$
KB218667.1	scf_22030_31606	22039–22153	−	MIR167_1 (RF00640)	$4.21e^{-16}$
KB218973.1	scf_22030_46697	19560–19671	+	MIR167_1 (RF00640)	$2.76e^{-17}$
KB218973.1	scf_22030_46697	19561–19672	−	MIR167_1 (RF00640)	$9.11e^{-14}$
KB220367.1	scf_22030_75599	1–83	+	MIR167_1 (RF00640)	$1.83e^{-11}$
KB220367.1	scf_22030_75599	1–81	−	MIR167_1 (RF00640)	$5.81e^{-09}$
KB220896.1	scf_22030_80878	14228–14335	+	MIR168 (RF00677)	$1.12e^{-22}$
KB220896.1	scf_22030_80878	14227–14333	−	MIR168 (RF00677)	$2.28e^{-14}$
KB218337.1	scf_22030_18159	17587–17690	−	MIR169_2 (RF00645)	$1.07e^{-26}$
KB218337.1	scf_22030_18159	13143–13246	−	MIR169_2 (RF00645)	$2.24e^{-21}$
KB218337.1	scf_22030_18159	12902–12993	−	MIR169_2 (RF00645)	$3.40e^{-21}$
KB218337.1	scf_22030_18159	17589–17692	+	MIR169_2 (RF00645)	$2.10e^{-15}$
KB218337.1	scf_22030_18159	12904–12995	+	MIR169_2 (RF00645)	$2.36e^{-15}$
KB220127.1	scf_22030_72989	786–899	−	MIR169_2 (RF00645)	$9.28e^{-18}$

Table 3. *Cont.*

GenBank accession number	Scaffold name	Start and end positions	Strand	Rfam ID (and accession number)	Rfam scan E value
KB220321.1	scf_22030_74988	935–1052	+	MIR169_2 (RF00645)	7.84e^{-18}
KB220321.1	scf_22030_74988	933–1050	−	MIR169_2 (RF00645)	9.12e^{-11}
KB218337.1	scf_22030_18159	17584–17696	−	MIR169_5 (RF00865)	3.86e^{-08}
KB218337.1	scf_22030_18159	17583–17695	+	MIR169_5 (RF00865)	5.88e^{-08}
KB220127.1	scf_22030_72989	780–906	+	MIR169_5 (RF00865)	1.94e^{-19}
KB220127.1	scf_22030_72989	781–907	−	MIR169_5 (RF00865)	1.46e^{-06}
KB220321.1	scf_22030_74988	928–1058	−	MIR169_5 (RF00865)	7.73e^{-20}
KB220321.1	scf_22030_74988	927–1057	+	MIR169_5 (RF00865)	9.15e^{-06}
KB220807.1	scf_22030_80059	3863–3990	+	MIR169_5 (RF00865)	4.61e^{-11}
KB218810.1	scf_22030_38865	27461–27559	+	MIR171_1 (RF00643)	1.79e^{-16}
KB218810.1	scf_22030_38865	27459–27557	−	MIR171_1 (RF00643)	8.90e^{-14}
KB220711.1	scf_22030_79061	2105–2214	+	MIR171_1 (RF00643)	2.74e^{-19}
KB220711.1	scf_22030_79061	2103–2212	−	MIR171_1 (RF00643)	4.15e^{-13}
KB219420.1	scf_22030_61010	2619–2748	−	mir-172 (RF00452)	2.11e^{-19}
KB219420.1	scf_22030_61010	2619–2748	+	mir-172 (RF00452)	1.03e^{-15}
KB218089.1	scf_22030_7511	28886–28982	−	mir-287 (RF00788)	3.04e^{-04}
KB218983.1	scf_22030_47118	10649–10756	−	MIR390 (RF00689)	1.99e^{-21}
KB218983.1	scf_22030_47118	10649–10756	+	MIR390 (RF00689)	1.75e^{-14}
KB219488.1	scf_22030_62701	16710–16837	+	MIR390 (RF00689)	3.68e^{-23}
KB219488.1	scf_22030_62701	16710–16837	−	MIR390 (RF00689)	8.85e^{-12}
KB218810.1	scf_22030_38865	36369–36475	+	MIR394 (RF00688)	9.23e^{-14}
KB219360.1	scf_22030_59359	18185–18287	−	mir-395 (RF00451)	5.48e^{-14}
KB219360.1	scf_22030_59359	18185–18287	+	mir-395 (RF00451)	6.44e^{-11}
KB219922.1	scf_22030_70572	3837–3927	+	MIR396 (RF00648)	1.03e^{-20}
KB219922.1	scf_22030_70572	1415–1528	+	MIR396 (RF00648)	1.35e^{-17}
KB219922.1	scf_22030_70572	3836–3926	−	MIR396 (RF00648)	2.37e^{-15}
KB219922.1	scf_22030_70572	1414–1527	−	MIR396 (RF00648)	2.41e^{-13}
KB219961.1	scf_22030_71131	9924–10008	−	MIR396 (RF00648)	1.30e^{-15}
KB219961.1	scf_22030_71131	9925–10009	+	MIR396 (RF00648)	3.38e^{-12}
KB220512.1	scf_22030_77233	7423–7504	+	MIR396 (RF00648)	1.50e^{-20}
KB220512.1	scf_22030_77233	7422–7503	−	MIR396 (RF00648)	6.96e^{-17}
KB221106.1	scf_22030_81441	12748–12911	+	MIR408 (RF00690)	2.85e^{-09}
KB219476.1	scf_22030_62392	5876–5979	+	MIR535 (RF00714)	4.25e^{-19}
KB219838.1	scf_22030_69379	8499–8600	+	MIR535 (RF00714)	1.44e^{-23}
KB219838.1	scf_22030_69379	8497–8598	−	MIR535 (RF00714)	1.83e^{-17}
KB220694.1	scf_22030_78899	5550–5652	−	MIR535 (RF00714)	3.74e^{-18}
KB220154.1	scf_22030_73255	538–819	+	Plant_SRP (RF01855)	1.43e^{-24}
KB220490.1	scf_22030_76954	17439–17650	+	Plant_U3 (RF01847)	2.04e^{-36}
KB219898.1	scf_22030_70290	25811–25954	+	snoF1_F2 (RF00482)	1.49e^{-19}
KB218033.1	scf_22030_4706	9374–9436	−	snoJ33 (RF00315)	4.02e^{-07}
KB219471.1	scf_22030_62284	16444–16526	−	snoJ33 (RF00315)	5.63e^{-09}
KB219426.1	scf_22030_61169	69226–69316	−	snoR11 (RF00349)	1.31e^{-17}
KB219685.1	scf_22030_66563	26216–26343	−	snoR111 (RF01228)	1.27e^{-14}
KB220857.1	scf_22030_80459	12071–12174	−	snoR113 (RF01420)	4.15e^{-20}

Table 3. *Cont.*

GenBank accession number	Scaffold name	Start and end positions	Strand	Rfam ID (and accession number)	Rfam scan E value
KB218307.1	scf_22030_16452	15390–15476	−	snoR118 (RF01424)	$1.15e^{-15}$
KB218657.1	scf_22030_31300	24736–24824	+	snoR14 (RF01280)	$8.40e^{-14}$
KB218015.1	scf_22030_3847	11974–12060	−	snoR16 (RF00296)	$1.39e^{-18}$
KB218015.1	scf_22030_3847	12491–12577	−	snoR16 (RF00296)	$1.11e^{-17}$
KB220504.1	scf_22030_77091	17217–17303	−	snoR16 (RF00296)	$4.81e^{-19}$
KB220504.1	scf_22030_77091	16789–16875	−	snoR16 (RF00296)	$9.43e^{-19}$
KB220539.1	scf_22030_77514	2858–2933	+	snoR160 (RF00203)	$1.40e^{-15}$
KB219378.1	scf_22030_59710	15789–15866	+	snoR28 (RF00355)	$4.91e^{-22}$
KB218307.1	scf_22030_16452	15543–15617	−	snoR66 (RF00202)	$2.49e^{-16}$
KB219947.1	scf_22030_70993	16528–16659	+	snoR80 (RF01224)	$2.92e^{-20}$
KB220353.1	scf_22030_75402	20181–20308	−	snoR86 (RF00303)	$1.06e^{-24}$
KB219338.1	scf_22030_58993	16769–16872	−	snoR97 (RF01215)	$1.30e^{-18}$
KB219443.1	scf_22030_61493	32748–32838	−	SNORD15 (RF00067)	$2.00e^{-09}$
KB219661.1	scf_22030_66054	15711–15796	−	SNORD25 (RF00054)	$5.96e^{-22}$
KB219661.1	scf_22030_66054	15482–15566	−	SNORD25 (RF00054)	$5.50e^{-21}$
KB219661.1	scf_22030_66054	14874–14958	−	SNORD25 (RF00054)	$2.14e^{-20}$
KB219661.1	scf_22030_66054	15075–15159	−	SNORD25 (RF00054)	$9.04e^{-17}$
KB219898.1	scf_22030_70290	25498–25585	+	SNORD33 (RF00133)	$5.82e^{-16}$
KB218015.1	scf_22030_3847	12999–13097	−	SNORD43 (RF00221)	$7.53e^{-11}$
KB220504.1	scf_22030_77091	17701–17798	−	SNORD43 (RF00221)	$6.80e^{-12}$
KB220504.1	scf_22030_77091	17915–18012	−	SNORD43 (RF00221)	$9.20e^{-11}$
KB219898.1	scf_22030_70290	25347–25436	+	snoU31b (RF01285)	$4.66e^{-17}$
KB220870.1	scf_22030_80641	5915–5999	+	snoU36a (RF01302)	$5.82e^{-21}$
KB219426.1	scf_22030_61169	68869–68977	−	snoZ152 (RF00350)	$2.58e^{-16}$
KB219947.1	scf_22030_70993	16107–16211	+	snoZ157 (RF00333)	$1.58e^{-18}$
KB219898.1	scf_22030_70290	25690–25775	+	snoZ196 (RF00134)	$2.75e^{-14}$
KB220870.1	scf_22030_80641	6066–6159	+	snoZ223 (RF00135)	$1.98e^{-19}$
KB218327.1	scf_22030_17743	7560–7631	+	snoZ266 (RF00332)	$8.06e^{-09}$
KB219338.1	scf_22030_58993	17401–17516	−	snoZ278 (RF00201)	$1.76e^{-16}$
KB219338.1	scf_22030_58993	17113–17226	−	snoZ278 (RF00201)	$9.06e^{-13}$
KB219250.1	scf_22030_57131	12714–12875	−	U1 (RF00003)	$9.36e^{-39}$
KB219770.1	scf_22030_68191	6294–6455	+	U1 (RF00003)	$3.43e^{-41}$
KB220529.1	scf_22030_77416	6949–7110	+	U1 (RF00003)	$5.34e^{-36}$
KB220746.1	scf_22030_79451	5096–5256	+	U1 (RF00003)	$2.21e^{-27}$
KB218084.1	scf_22030_7289	6288–6438	−	U12 (RF00007)	$1.92e^{-27}$
KB219620.1	scf_22030_65416	19689–19820	−	U2 (RF00004)	$2.10e^{-17}$
KB220509.1	scf_22030_77120	23424–23564	−	U4 (RF00015)	$1.19e^{-08}$
KB218936.1	scf_22030_44766	5102–5143	+	U5 (RF00020)	$2.13e^{-09}$
KB218979.1	scf_22030_47021	19677–19800	+	U5 (RF00020)	$4.89e^{-10}$
KB218084.1	scf_22030_7289	12644–12761	−	U5 (RF00020)	$4.29e^{-18}$
KB220567.1	scf_22030_77768	17710–17830	+	U5 (RF00020)	$3.52e^{-11}$
KB217934.1	scf_22030_16	16123–16225	−	U6 (RF00026)	$1.54e^{-10}$
KB218759.1	scf_22030_36539	4240–4337	+	U6 (RF00026)	$2.72e^{-11}$

Figure 4. BLASTN alignment of an enset supercontig (GenBank: KB219804) against banana chromosome 5, displayed using the Artemis Comparison Tool (ACT).

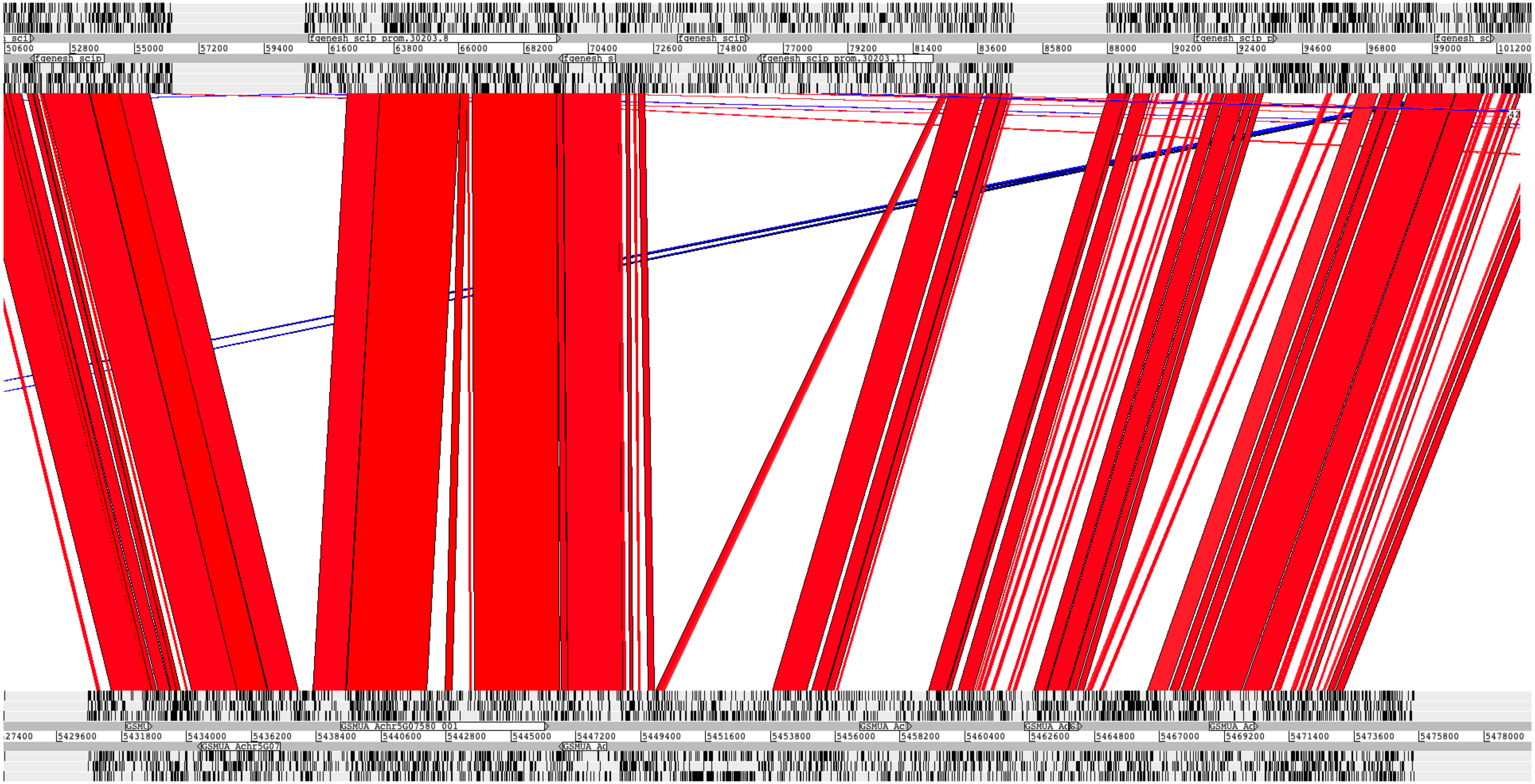


Table 4. Overview and classification of the repeats present in the enset genome and comparison with those in the *M. acuminata* genome.

	Ensete Ventricosum			**Musa Acuminata**		
Class	**Count**	**Bp**	**%**	**Count**	**Bp**	**%**
Ty1/Copia	17,446	6,064,590	1.36	5,053	2,476,355	0.75
Copia/Angela	102,430	39,177,431	8.78	15,025	10,764,293	3.24
Copia/SIRE1Maximus	102,464	27,386,896	6.14	37,446	26,594,658	8.01
Copia/Tnt1	10,144	4,915,981	1.10	2,869	3,300,009	0.99
Ty3/Gypsy	24,694	11,556,851	2.59	5,047	4,552,048	1.37
Gypsy/CRM	3,740	2,246,235	0.50	542	534,904	0.16
Gypsy/Galadriel	12,452	6,626,137	1.49	1,874	2,210,611	0.67
Gypsy/Galadriel-lineage	16	734	0.00	5	237	0.00
Gypsy/Reina	65,858	23,579,479	5.29	6,170	4,243,784	1.28
Gypsy/Tekay	14,043	5,490,598	1.23	4,351	3,031,464	0.91
LINE	5,833	1,346,085	0.30	1,745	552,483	0.17
RE	31,224	4,967,551	1.11	9,005	2,824,122	0.85
Satellite/Type1	178	69,579	0.02	20	30,828	0.01
Satellite/Type2	9,516	3,563,409	0.80	18	29,902	0.01
clDNA	6,590	1,126,726	0.25	2,652	430,368	0.13
DNA/hAT	2,910	783,511	0.18	1,916	637,668	0.19
Total	409,538	138,901,793	31.14	93,738	62,213,734	19.74

2.6. Enset—Specific Genes Include Reverse Transcriptases, Viral Sequences, and a Putative Disease-Resistance Gene

Among the enset genes not conserved in the *M. acuminata* genome [6], are several predicted to encode reverse transcriptases (Pfam accession PF00078). Reverse transcriptases are characteristic of several classes of mobile elements, including retroviruses, such as the banana streak virus. The phylogenetic relationships of these reverse transcriptases are shown in Figure 5, which indicates that they fall into two distinct clades. One of these clades (in the lower part of Figure 5) includes two genes from banana along with two from enset. However, the other clade (the upper part of Figure 5) includes no known sequences from *Musa* species, but includes sequences from several other monocot and dicot plants.

Similarly, the enset genome encodes at least 14 predicted proteins containing the integrase core domain (Pfam: PF00665) while the banana genome [6] encodes only one (see Figure 6). The integrase core domain is involved in integration of a copy of a viral genome into the host chromosome. The enset genome also encodes at least 19 predicted retrotransposon gag proteins (Pfam: PF03732) with no closely related sequence in banana (Figure 7).

Figure 5. Maximum-Likelihood phylogenetic tree for enset reverse transcriptase-domain proteins. Protein sequences from *E. ventricosum* are indicated by circles. The sequences from *M. acuminata* are indicated by diamonds. Bootstrap values of greater than 50% are indicated as numbers on the branches.

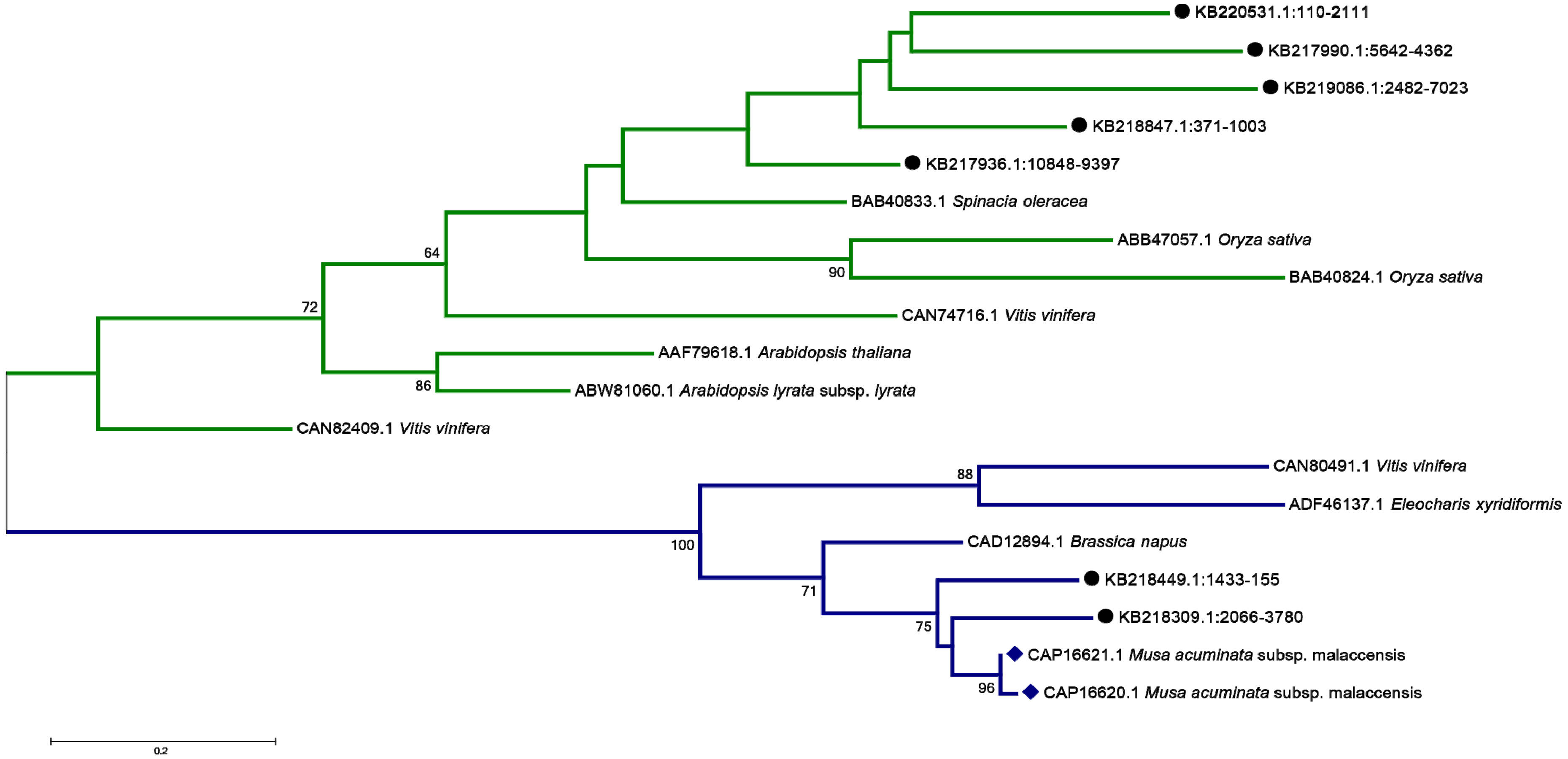

KB220531.1:110-2111
KB217990.1:5642-4362
KB219086.1:2482-7023
KB218847.1:371-1003
KB217936.1:10848-9397
BAB40833.1 *Spinacia oleracea*
ABB47057.1 *Oryza sativa*
64
90
BAB40824.1 *Oryza sativa*
72
CAN74716.1 *Vitis vinifera*
AAF79618.1 *Arabidopsis thaliana*
86
ABW81060.1 *Arabidopsis lyrata* subsp. *lyrata*
CAN82409.1 *Vitis vinifera*
88
CAN80491.1 *Vitis vinifera*
ADF46137.1 *Eleocharis xyridiformis*
100
CAD12894.1 *Brassica napus*
71
KB218449.1:1433-155
75
KB218309.1:2066-3780
CAP16621.1 *Musa acuminata* subsp. malaccensis
96
CAP16620.1 *Musa acuminata* subsp. malaccensis
0.2

Figure 6. Maximum-Likelihood phylogenetic tree for enset integrase core-domain proteins. Protein sequences from *E. ventricosum* are indicated by circles. Bootstrap values of greater than 50% are indicated as numbers on the branches.

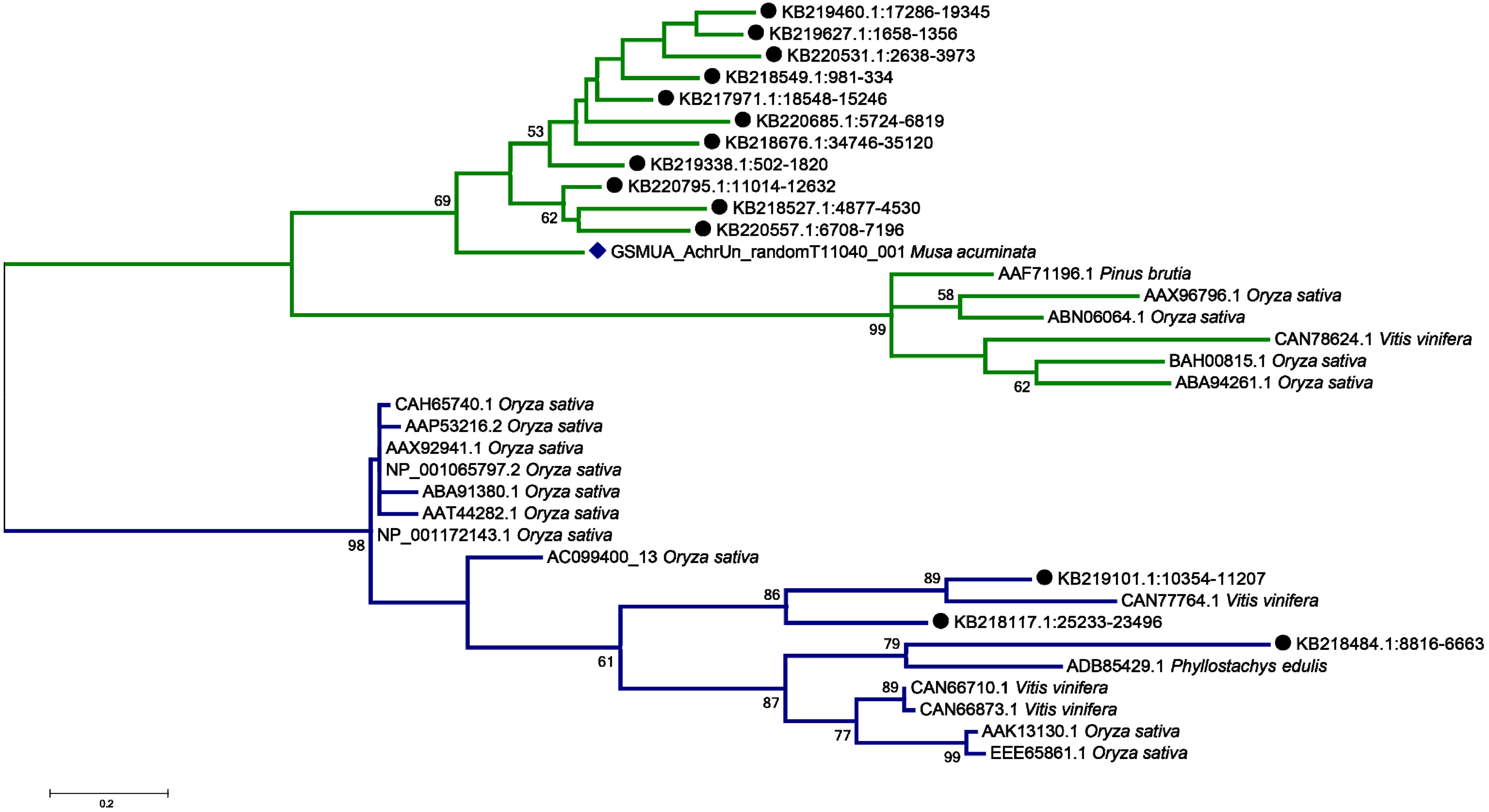
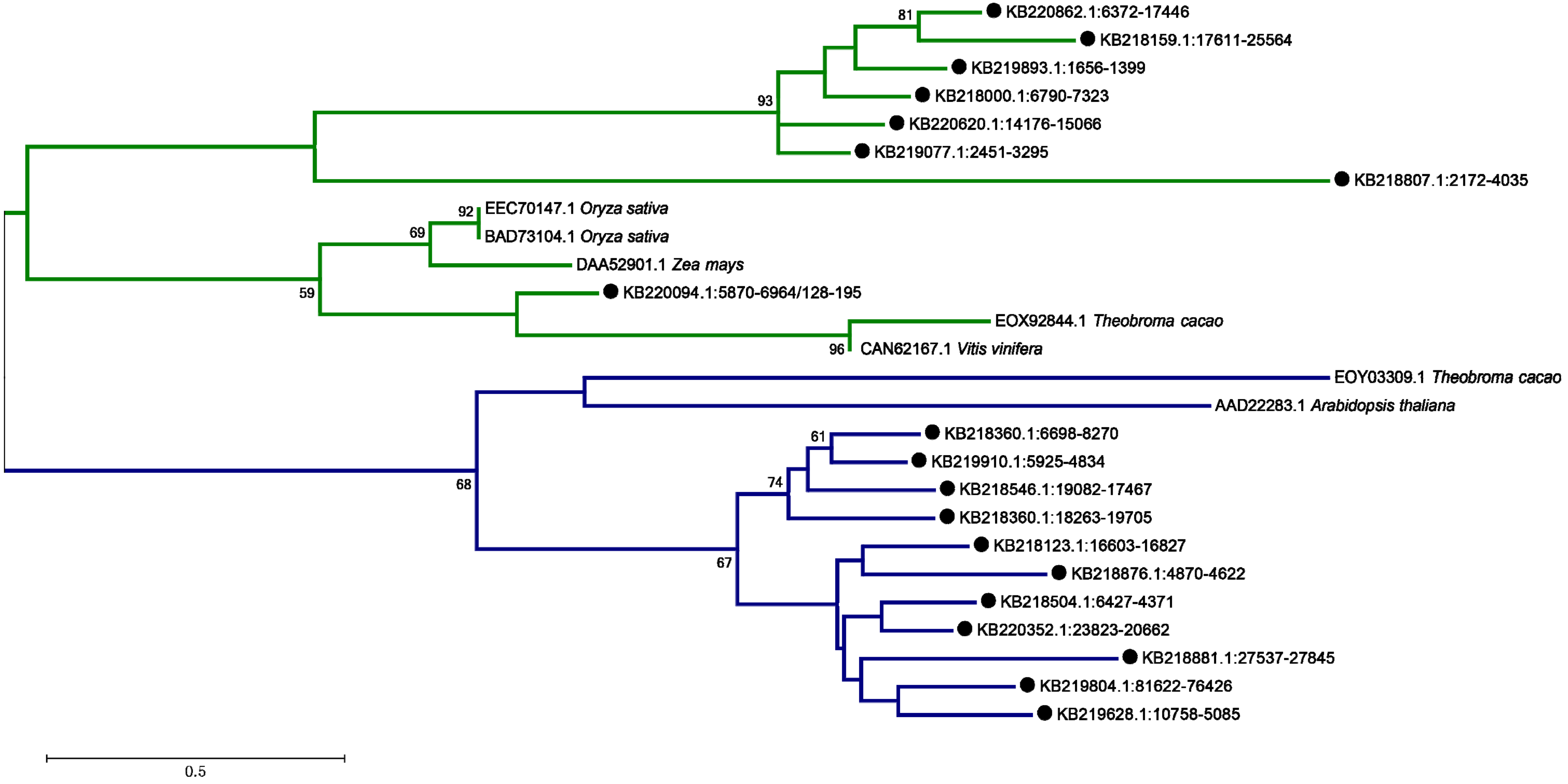

● KB219460.1:17286-19345
● KB219627.1:1658-1356
● KB220531.1:2638-3973
● KB218549.1:981-334
● KB217971.1:18548-15246
● KB220685.1:5724-6819
53
● KB218676.1:34746-35120
● KB219338.1:502-1820
● KB220795.1:11014-12632
69
62
● KB218527.1:4877-4530
● KB220557.1:6708-7196
◆ GSMUA_AchrUn_randomT11040_001 *Musa acuminata*
AAF71196.1 *Pinus brutia*
58
AAX96796.1 *Oryza sativa*
ABN06064.1 *Oryza sativa*
99
CAN78624.1 *Vitis vinifera*
BAH00815.1 *Oryza sativa*
62
ABA94261.1 *Oryza sativa*
CAH65740.1 *Oryza sativa*
AAP53216.2 *Oryza sativa*
AAX92941.1 *Oryza sativa*
NP_001065797.2 *Oryza sativa*
ABA91380.1 *Oryza sativa*
AAT44282.1 *Oryza sativa*
NP_001172143.1 *Oryza sativa*
98
AC099400_13 *Oryza sativa*
89
● KB219101.1:10354-11207
86
CAN77764.1 *Vitis vinifera*
● KB218117.1:25233-23496
79
● KB218484.1:8816-6663
61
ADB85429.1 *Phyllostachys edulis*
89
CAN66710.1 *Vitis vinifera*
87
CAN66873.1 *Vitis vinifera*
77
AAK13130.1 *Oryza sativa*
99
EEE65861.1 *Oryza sativa*
0.2

Figure 7. Maximum-Likelihood phylogenetic tree for enset integrase core-domain proteins. Proteins sequences from *E. ventricosum* are indicated by circles. The sequence from *M. acuminata* is indicated by a diamond. Bootstrap values of greater than 50% are indicated as numbers on the branches.



It has been shown that the genomes of some *Musa* species contain endogeneous retroviruses that are integrated into the host chromosome [28]. The genome of *E. ventricosum* contains several sequences that resemble retrovirus sequences and therefore may represent endogeneous integrated viruses. Specifically, a *M. balbisiana* sequence containing eBSOLV (endogeneous *Obino l'Ewai virus*) sequence (GenBank: HE983609 [28]) is highly conserved in *E. ventricosum*, though this sequence is absent from the *M. acuminata* genome [6]. Similarly, *E. ventricosum* contains sequences with 86% nucleotide identity to a 2.25-kb fragment of banana streak UA virus (GenBank: AEC49874) and 79% identity to a 1.1-kb fragment of the sugarcane bacilliform virus (SCBV) BT20231 (GenBank: FJ439799 [29]). It is not clear whether any of these virus sequences represent viruses that can become infectious as they can in *Musa* species [28].

Other enset proteins not found in the banana genome include a protein (GenBank: KB218027) that shares 42% amino-acid identity with *Arabidopsis thaliana* protein At1g53350, annotated as an RPP8-like resistance protein. Examples such as this are candidates for future studies on disease resistance in enset and perhaps even for introgression into banana.

3. Experimental Section

The *E. ventricosum* plant was grown from seed purchased from Jungle Seeds (Wallington, UK). We extracted genomic DNA using the DNAEasy Plant Minikit supplied by Qiagen (Manchester, UK). We sequenced genomic DNA using an Illumina HiSeq 2500, according to the manufacturer's instructions. We used a single lane of an eight-lane flowcell and generated 202 million pairs of 100-nucleotide reads with a mean insert-length of approximately 350 nucleotides.

For alignment of sequence reads against reference sequences, we used BWA version 0.7.5a-r405 [14] and visualized BWA alignments using the Integrative Genomics Viewer IGV [30]. For *de novo* assembly we used SOAPdenovo version 1.05 [31]. Prior to assembly, we removed all sequence reads that contained "N"s. Calculations of N_{50} and NG_{50} were based on the definitions of these two statistics stated by Assemblathon [27].

We used BLAST [32] and MUMMER [22] for pairwise alignments of assembled sequences and reference sequences and visualized BLAST alignments using the Artemis Comparison Tool (ACT) [33]. We used MEGA5 [22] for phylogenetic analysis.

To identify repeat sequences, we used RepeatMasker version open-4.0.1 [26,34,35] in default mode run with RMBLAST version 2.2.27+ against the customized library of *M. acuminata* repeats (1903 sequences) from Hřibová and colleagues [36,37]. This is the same library of banana-specific repeats used in the *M. balbisiana* genome re-sequencing project [12].

For *ab initio* gene prediction from our de novo genome assembly, we used FGENESH v.3.1.1 [22] with parameters tuned for 'monocot plant'.

4. Conclusions

Here we present the first genome-wide sequencing study of enset (*Ensete ventricosum*). We have identified more than 1000 candidate SNPs, and by using less stringent criteria, many more candidates could be identified. These data will be useful as a reference sequence for future "omics studies" on this neglected crop. Armed with this initial draft genome sequence, we can now extend our studies to genotypic variation among different Ethiopian varieties of enset, both cultivated and wild.

Acknowledgments

This work was funded in part by the Wellcome Trust Biomedical Informatics Hub at the University of Exeter. We are grateful to Eva Hřibová for making available the library of banana repeat sequences.

Conflicts of Interest

The authors declare no conflict of interest.

References and Notes

1. Brandt, S.A.; Spring, A.; Hiebsch, C.; McCabe, J.T.; Tabogie, E.; Diro, M.; Wolde-Michael, G.; Yntiso, G.; Shigeta, M.; Tesfaye, S. *The "Tree Against Hunger" Enset-Based Agricultural Systems in Ethiopia*; American Association for the Advancement of Science: Washington, DC, USA, 1997; pp. 1–58.
2. Pijls, L.T.J.; Timmer, A.A.M.; Wolde-Gebriel, Z.; West, C.E. Cultivation, preparation and consumption of ensete *(Ensete ventricosum*) in Ethiopia. *J. Sci. Food Agric.* **1995**, *67*, 1–11.
3. Asfaw, B.T. *Studies on Landraces Diversity* in vivo *and* in vitro *Regeneration of Enset: (*Enset ventricosum *Welw.)*; Köster: Milan, Lombardy, Italy, 2002; p. 127.
4. Biruma, M.; Pillay, M.; Tripathi, L.; Blomme, G.; Abele, S.; Mwangi, M.; Bandyopadhyay, R.; Muchunguzi, P.; Kassim, S.; Nyine, M.; *et al.* Banana *Xanthomonas* wilt: A review of the disease, management strategies and future research directions. *Afr. J. Biotechnol.* **2007**, *6*, 953–962.
5. Cheesman, E. Classification of the bananas: The genus ensete horan. *Kew Bull.* **1947**, *2*, 97–106.
6. D'Hont, A.; Denoeud, F.; Aury, J.-M.J.; Baurens, F.-C.F.; D'Hont, A.; Carreel, F.; Garsmeur, O.; Noel, B.; Bocs, S.; Droc, G.; *et al.* The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **2012**, *488*, 213–217.
7. Ethiopian Institute of Agricultural Research (EIAR). Enset Research and Development Experiences in Ethiopia. In Proceedings of Enset National Workshop, Wolkite, Ethiopia, 19–20 August 2010; Yesuf, M., Hunduma, T., Eds.; Ethiopian Institute of Agricultural Research (EIAR): Addis Ababa, Ethiopia, 2012.
8. Tsegaye, A. On Indigenous Production, Genetic Diversity and Crop Ecology of Enset (*Ensete ventricosum* (Welw.) Cheesman). Ph.D. Thesis, Wageningen University, Wageningen, The Netherlands, 22 April 2002; p. 198.
9. Negash, A.; Niehof, A. The significance of enset culture and biodiversity for rural household food and livelihood security in southwestern Ethiopia. *Agric. Human Values* **2004**, *21*, 61–71.
10. Birmeta, G.; Nybom, H.; Bekele, E. RAPD analysis of genetic diversity among clones of the Ethiopian crop plant *Ensete ventricosum*. *Euphytica* **2002**, *124*, 315–325.
11. Birmeta, G.; Nybom, H.; Bekele, E. Distinction between wild and cultivated enset (*Ensete ventricosum*) gene pools in Ethiopia using RAPD markers. *Hereditas* **2004**, *140*, 139–148.

12. Davey, M.W.; Gudimella, R.; Harikrishna, J.A.; Sin, L.W.; 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*, doi:10.1186/1471-2164-14-683.
13. National Center for Biotechnology Information. Available online: ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/ (accessed on 22 December 2013).
14. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760.
15. Silby, M.W.; Cerdeño-Tárraga, A.M.; Vernikos, G.S.; Giddens, S.R.; Jackson, R.W.; Preston, G.M.; Zhang, X.-X.; Moon, C.D.; Gehrig, S.M.; Godfrey, S.A.C.; *et al*. Genomic and genetic analyses of diversity and plant interactions of *Pseudomonas fluorescens*. *Genome Biol.* **2009**, *10*, R51.
16. Copeland, A.; Lucas, S.; Lapidus, A.; Glavina del Rio, T.; Dalin, E.; Tice, H.; Bruce, D.; Goodwin, L.; Pitluck, S.; Kiss, H.; *et al*. US DOE Joint Genome Institute, Walnut Creek, CA, USA. Unpublished work, 2008.
17. Brzuszkiewicz, E.; Weiner, J.; Wollherr, A.; Thürmer, A.; Hüpeden, J.; Lomholt, H.B.; Kilian, M.; Gottschalk, G.; Daniel, R.; Mollenkopf, H.-J.; Meyer, T.F.; Brüggemann, H. Comparative genomics and transcriptomics of *Propionibacterium acnes*. *PLoS One* **2011**, *6*, e21581.
18. Chung, W.-C.; Chen, L.-L.; Lo, W.-S.; Kuo, P.-A.; Tu, J.; Kuo, C.-H. Complete genome sequence of *Serratia marcescens* WW4. *Genome Announc.* **2013**, *1*, e0012613.
19. Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **2010**, *26*, 589–595.
20. Mammadov, J.; Aggarwal, R.; Buyyarapu, R.; Kumpatla, S. SNP markers and their impact on plant breeding. *Int. J. Plant Genomics* **2012**, *2012*, 728398.
21. Studholme, D. *Ensete ventricosum* Genome Sequence. Available online: http://figshare.com/articles/Ensete_ventricosum_genome_sequence/894306 (accessed on 6 January 2014).
22. Kurtz, S.; Phillippy, A.; Delcher, A.L.; Smoot, M.; Shumway, M.; Antonescu, C.; Salzberg, S.L. Versatile and open software for comparing large genomes. *Genome Biol.* **2004**, *5*, R12.
23. Solovyev, V. Statistical Approaches in Eukaryotic Gene Prediction. In *Handbook of Statistical Genetics*; John Wiley & Sons, Ltd.: Chichester, West Sussex, UK, 2004; pp. 97–159.
24. Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; *et al*. Pfam: The protein families database. *Nucleic Acids Res.* **2013**, *42*, D222–D230.
25. Gardner, P.P.; Daub, J.; Tate, J.; Moore, B.L.; Osuch, I.H.; Griffiths-Jones, S.; Finn, R.D.; Nawrocki, E.P.; Kolbe, D.L.; Eddy, S.R.; *et al*. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.* **2011**, *39*, D141–D145.
26. Tempel, S.; Repeatmasker, U. Using and understanding RepeatMasker. *Methods Mol. Biol.* **2012**, *859*, 29–51.
27. Earl, D.; Bradnam, K.; St John, J.; Darling, A.; Lin, D.; Fass, J.; Yu, H.O.K.; Buffalo, V.; Zerbino, D.R.; Diekhans, M.; *et al*. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res.* **2011**, *21*, 2224–2241.

28. Chabannes, M.; Baurens, F.-C.; Duroy, P.-O.; Bocs, S.; Vernerey, M.-S.; Rodier-Goud, M.; Barbe, V.; Gayral, P.; Iskra-Caruana, M.-L. Three infectious viral species lying in wait in the banana genome. *J. Virol.* **2013**, *87*, 8624–8637.
29. Muller, E.; Dupuy, V.; Blondin, L.; Bauffe, F.; Daugrois, J.-H.; Nathalie, L.; Iskra-Caruana, M.-L. High molecular variability of sugarcane bacilliform viruses in Guadeloupe implying the existence of at least three new species. *Virus Res.* **2011**, *160*, 414–419.
30. Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): High-Performance genomics data visualization and exploration. *Briefings Bioinforma.* **2013**, *14* , 178–192.
31. Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y.; *et al.* SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **2012**, *1*, doi:10.1186/2047-217X-1-18.
32. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.
33. Carver, T.J.; Rutherford, K.M.; Berriman, M.; Rajandream, M.-A.; Barrell, B.G.; Parkhill, J. ACT: The Artemis Comparison Tool. *Bioinformatics* **2005**, *21*, 3422–3423.
34. Tarailo-Graovac, M.; Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **2009**, *4*, doi:10.1002/0471250953.bi0410s25.
35. RepeatMasker. Available online: http://www.repeatmasker.org (accessed on 20 December 2013).
36. Hribová, E.; Neumann, P.; Matsumoto, T.; Roux, N.; Macas, J.; Dolezel, J. Repetitive part of the banana (*Musa acuminata*) genome investigated by low-depth 454 sequencing. *BMC Plant Biol.* **2010**, *10*, 204.
37. Institute of Experimental Botany. Available online: http://wwwueb.asuch.cas.cz/Olomouc1/banana-sequencing-data/BananaREP.tar.gz (accessed on 20 December 2013).