TÜBİTAK

Turkish Journal of Agriculture and Forestry

http://journals.tubitak.gov.tr/agriculture/

Research Article

Turk J Agric For
(2015) 39:
© TÜBİTAK
doi:10.3906/tar-1406-171

Transcriptome analysis of banana (*Musa balbisiana*) based on next-generation sequencing technology

Suthanthiram BACKIYARANI*, Subbaraya UMA, Marimuthu Somasundram SARASWATHI, Asoor Santhanam SARAVANAKUMAR, Arumugam CHANDRASEKAR
ICAR-National Research Centre for Banana, Triuchirapalli, Trichy, Tamil Nadu, India

Received: 30.06.2014 • **Accepted/Published Online:** 06.01.2015 • **Printed:** 00.00.2015

Abstract: Banana (*Musa* spp.) is an important tropical fruit with high commercial potential. *Musa balbisiana* (B genome) is a progenitor of one of the most cultivated banana species and exhibits unique traits, including resistance or tolerance to many biotic and abiotic stresses. RNA sequencing of the *Musa* B genome would provide a vast array of transcriptomic information that could lead to the development of trait-specific markers and the discovery of new genes and regulatory sequences involved in resistance mechanisms. Thus, transcriptome sequencing was performed in *Musa* B genome accession Attikol using the Ion Torrent platform. This led to the generation of about 4.5 million paired-end reads, which were assembled using the MIRA assembler. The assembly produced 82,413 unique transcripts with a mean length of approximately 113 bp. The sequence similarity search against the Swiss-Prot database resulted in the identification of 35,783 unique transcripts (62.18%). Out of these, 193,826 gene ontology terms were assigned to unique transcripts. Functional annotation against PlantCYC pathway database identified 20,696 unique transcripts, which were mapped to 455 pathways. About 4780 simple sequence repeats (SSRs) were obtained from 82,413 unique transcripts. Primers could be designed for only 2628 SSRs, out of which 30 primers were randomly selected from defense-related genes to confirm their efficiency. This information will make the improvement of banana cultivars easier by facilitating the selection of resistance genes as well as the development of trait-specific markers.

Key words: Banana, Ion Torrent, *Musa balbisiana*, next-generation sequencing, transcriptome

1. Introduction

Bananas and plantains are important cash and subsistence crops in most tropical and subtropical regions of the world (Ortiz and Swennen, 2014). The varieties of both commerce and subsistence production are predominantly sterile triploids (2n = 33) originating from two important seeded wild species, namely *Musa acuminata* Colla (contributing the A genome) and *M. balbisiana* Colla (contributing the B genome) (Simmonds, 1973). Interspecific hybridization has resulted in a diverse array of cultivars with varied genomic constitutions: AA, AAA, AB, AAB, ABB, ABBB, BB, etc. Of these, cultivars rich in *Musa* B (AB, ABB, ABBB) form the basis of subsistence production (Uma et al., 2006a). Wong et al. (2002) reported that *M. balbisiana* is the most distinct species within the section *Eumusa*. Nwakanma et al. (2003) found that *M. balbisiana* has the more primitive organellar genomes.

M. acuminata might be of the most recent origin. It is also commonly accepted that hardiness is contributed by the *Musa* B genome, since *M. balbisiana* clones strive in areas experiencing pronounced dry seasons alternating with monsoons. Moreover, *M. balbisiana* has specific fruit characteristics, such as starchiness and acidity superior to those of *M. acuminata* (Simmonds, 1962a, 1962b). Moreover, the importance of B-rich genotypes (both wild and cultivated) has been well documented. The B genome imparts valuable traits such as starchiness and acidic taste (Simmonds, 1962a, 1962b), and tolerance to moisture deficit (Ravi and Uma, 2011; Ravi et al., 2013) and *Fusarium* wilt (Swarupa et al., 2013)

The current global production of more than 100×10^6 t is based on large-scale vegetative propagation of a small number of genotypes, which have been derived from only a few ancient sexual recombination events. These genetically restricted and inflexible clones are particularly susceptible to diseases, pests, and current ecological changes. The challenge for banana improvement is to produce resistant and sterile polyploid hybrids through genetic recombination of fertile diploids that meet consumer expectations for each cultivar type. The required

* Correspondence: backiyarani@gmail.com

breeding strategy will need to reproduce the sequence of crossings and selections that occurred minimally during the past 7000 years, while periodically substituting some genitors from closely related genomes selected for their level of resistance to biotic and abiotic stresses. Hence, a prerequisite for banana improvement is to reconstruct the domestication pathways of the major cultivar groups as precisely as possible (Perrier et al., 2011). Knowledge regarding the genetic basis of yield, resistance to diseases and insect pests, and abiotic stress tolerance is a basic requirement for any crop improvement program; this knowledge is lacking in banana owing to its complex genetic structure and ploidy nature. The recent application of next-generation deep sequencing provides a platform for genetic improvement of recalcitrant crops such as banana or potato (Hobert, 2010).

The *Musa* A genome was already decoded by the D'Hont group in DH-Pahang, a double haploid *M. acuminata* genotype of the subspecies *malaccensis* that contributed to the one of the three *M. acuminata* genomes of Cavendish (AAA) (D'Hont et al., 2012). Sequencing of the A genome alone will not be sufficient to improve banana and plantains, since most of the Indian commercial cultivars are allopolyploid in nature and belong to the AB, AAB, and ABB types. Thus, decoding the *Musa* B genome is also mandatory in a banana improvement program. Though B genome sequencing of Pisang Klutuk Wulung (PKW) has been completed (Davey et al., 2013), use of a dihaploid PKW would have been more informative and precise.

Because of the deep coverage and single base-pair resolution provided by next-generation sequencing instruments, RNA sequencing represents an attractive alternative to whole-genome sequencing because it only analyzes transcribed portions of the genome, while avoiding noncoding and repetitive sequences that can make up much of the genome (Margulies et al., 2005; Liu et al., 2012).

According to Wang et al. (2009, 2011), transcriptome analysis is important when interpreting the functional elements of the genome. Thus, decoding the *M. balbisiana* transcriptome is vital to understanding the genetic basis of the B genome's unique traits; this information can be further exploited in a banana improvement program.

2. Materials and methods

2.1. Plant material and RNA extraction

Different tissues, such as leaf, root, sheath, flower bud, flower bract, pulp, and rhizome, of *M. balbisiana* 'Attikol', which is being maintained at the National Research Centre for Banana, were used to extract RNA for transcriptome sequencing. Two grams of tissues were weighed and frozen in liquid nitrogen, and total RNA was extracted using the Agilent Plant RNA Isolation Mini Kit (Agilent Technologies, Inc., USA). RNA integrity of the total RNA was checked using a Bioanalyzer 2100 (Agilent Technologies, Inc.). Equal amounts of RNA derived from different tissues were pooled together for library construction.

2.2. Library construction for Ion Torrent sequencing

The transcriptome library for sequencing was constructed according to the Ion Total RNA-Seq Kit User Guide (Part # 4467098; Rev. A; May 2011). Poly(A) RNA was purified from 10 µg of intact total RNA using a micro poly(A) purist kit. Samples of 1.3 µg were taken from fruit, root, leaf, flower, sheath, rhizome, and bract to make a total RNA pool of 10 µg. The purified mRNA was fragmented with RNase III and cleanup was done using the RNeasy Micro Kit (QIAGEN, USA). The fragmented RNA was hybridized and ligated with adapter and reverse-transcribed using ArrayScript reverse transcriptase. The cDNA was cleaned up using Agencourt Ampure XP SPRI beads (Beckman Coulter, USA). The purified products were amplified to create the final cDNA library. The prepared library was quantified using a NanoDrop spectrophotometer (Thermo Scientific, USA) and validated for quality by running an aliquot on a High Sensitivity Bioanalyzer Chip (Agilent Technologies, Inc.). Library concentration was quantified using a bioanalyzer (Agilent Technologies, Inc.) and 5×10^9 molecules were used for sequencing preparation. The library was amplified and the resulting templates were sequenced with the Ion OneTouch. Sequence reads were processed in the Torrent server and genomic data analysis was performed using the Partek Genomics Suite.

2.3. Data analysis and assembly

The raw reads were cleaned by removing adapter sequences, low-quality sequences (reads with ambiguous bases 'N'), and reads with a base quality smaller than 20 (Q20 bases). All sequences smaller than 100 bases were eliminated based on the assumption that small reads might represent sequencing artifacts (Meyer et al., 2009). The reads were submitted to the NCBI-SRA (Accession No. SRP050239). The quality reads were assembled into unigenes using MIRA-3.4.0, which recovers more full-length transcripts across a broad range of expression levels, with a sensitivity similar to that of methods that rely on genome alignments (Grabherr et al., 2011). The assembled unigenes were mapped with a banana reference genome (*Musa acuminata*) sequence (D'Hont et al., 2012). The overlap settings used for this assembly were 31 bp and 80% similarity, with all other parameters set to their default values.

2.4. Sequence annotation

The optimal assembly results were chosen according to the assembly evaluation. Clustering analysis was performed

to create a unigene database of the potential alternative splicing transcripts. Simple sequence repeat (SSR) analysis of unigenes longer than 100 bp was performed using MISA software. The assembled sequences were compared against the Swiss-Prot database using BLASTx (version 2.2.14) with an E-value of 10^{-5} (Altschul et al., 1997). Gene names were assigned to each assembled sequence based on the best BLAST hit (highest score). To increase computational speed, the search was limited to the first 10 significant hits for each query. Unannotated sequences were searched against the TrEMBL database using BLASTx with an E-value of 10^{-5} (Apweiler et al., 2004). To annotate the assembled sequences with GO terms describing biological processes, molecular functions, and cellular components, the Swiss-Prot BLAST results were imported into a GO database (http://www.geneontology.org/) that retrieves GO terms, allowing gene functions to be determined and compared. These GO terms are assigned to query sequences, producing a broad overview of groups of genes catalogued in the transcriptome for each of three ontology vocabularies, biological processes, molecular functions, and cellular components. The unigene sequences were also aligned to the KOG database (http://www.ncbi.nlm.nih.gov/COG) to predict and classify functions. Plant metabolic pathways were assigned to assembled sequences using the PlantCYC local database (http://www.plantcyc.org).

2.5. Transcriptome-SSR detection

All the unigenes of *Musa* obtained in this study were subjected to SSR detection using the MISA Perl Program (http://pgrc.ipk-gatersleben.de/misa/). The parameters were adjusted for the identification of perfect mono-, di-, tri-, tetra-, penta-, and hexanucleotide motifs with a minimum of 10, 6, 5, 5, 5, and 5 repeats, respectively, as described by Temnykh et al. (2001). The report of this search included the total number of sequences containing SSRs among the submitted unigenes, sequence ID, SSR motifs, number of repeats, repeat length, SSR starts, and SSR ends. Primers were designed for these SSRs using BatchPrimer3 v1.0 (http://probes.pw.usda.gov/batchprimer3/). Thirty genic-SSRs potentially related to defense responses were selected based upon annotation. Marker amplification and allele length polymorphisms were evaluated using 20 diploid (B genome) *M. balbisiana* accessions. A standard 100-bp molecular size marker was added to each gel to enable allele size estimation for PCR products run on 3% agarose gel along with ethidium bromide. Locus polymorphism was calculated using the polymorphism information content (PIC) calculator (http://w3.georgikon.hu/pic/english/default.aspx). All 37 genotypes were clustered with UPGMA analysis and the SAHN procedure of NTSYS-PC v2.10t.

3. Results and discussion

3.1. Sequence analysis and assembly

To obtain a global overview of the *Musa balbisiana* transcriptome and gene activity at nucleotide resolution, a pooled cDNA sample representing various tissues of *M. balbisiana* 'Attikol' was prepared and sequenced using the Ion Torrent genome analyzer. Pooling of RNA of different tissues and then sequencing with next-generation sequencing covered the entire transcript of the genome of the *M. balbisiana*, which was more informative for further analysis. After stringent quality assessment and data filtering, 4.5 million reads (519.87 million base pairs) with 85.66% Q20 bases (those with a base quality of >20) were selected as high-quality reads for further analyses. In pineapple, Ong et al. (2012) obtained 4.7 million reads, which is 4% higher than *Musa balbisiana* whole-transcriptome reads. The number of quality reads obtained in this study is greater than the *Musa* leaf transcriptomic reads (0.85 million) obtained using the 454 GS FLX platform (Passos et al., 2013) and less than the *Musa* root transcriptomic reads (26.6 million, two rounds of paired ends) obtained using Illumina's HiSeq 2000 system (Li et al., 2012). The tremendous variation in the number of quality reads may mainly be due to the difference in the platforms used for sequencing (Deschamps et al., 2012). The greatest read length was obtained with the Roche 454FLX (800 bp), followed by Ion Torrent (100 bp) and Illumina's HiSeq 2000 system (96 bp).

Though the number of reads obtained in this study is nearly 5 times less than the banana root transcriptomes, the number of contigs obtained in this study (82,413) is two times greater than the root transcriptome (47,411) developed by Li et al. (2012). The number of contigs gained in this study might be due to the sequencing of pooled RNA of different tissues like leaf, root, sheath, flower bud, flower bract, pulp, and rhizome (whole transcriptome) of the plant. Quail et al. (2012) reported that sequences generated by Ion Torrent and Illumina platforms generally display near-perfect coverage behavior on GC-rich and neutral genomes. The result obtained for nonredundant unigenes was on par with the Illumina platform, which could be attributed to the neutral position of *Musa* for GC content (Lescot et al., 2008; Backiyarani et al., 2013). The present study also confirmed banana's neutral GC content (49.7%). An overview of the sequencing is presented in Table 1.

The whole transcriptome consisted of 82,413 contigs, whose maximum length was 3402 bp. The assembled data revealed that most contigs (70,708) had an average length of 201–500 bp, while only 5 contigs had a length of >3 kbp. A summary of the Ion Torrent transcriptome assembly for *M. balbisiana* and the distribution of contigs are shown in Table 2 and Figure 1.

Table 1. Summary of Ion Torrent transcriptome sequencing of *Musa balbisiana.*

Sample	Total no. of bases (Mbp)	No. of reads	GC (%)	Q20 (%)
Banana	519.87	4,598,181	49.7	85.66

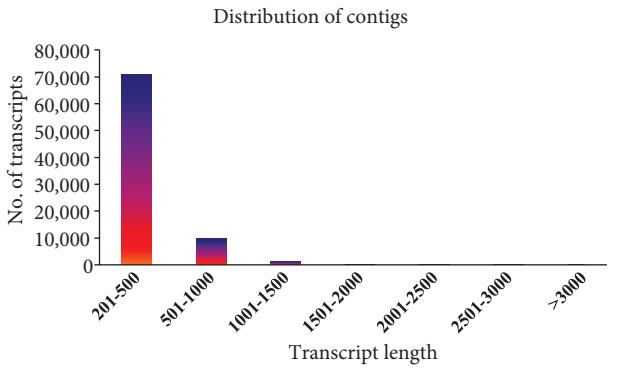

Figure 1. Overview of the *Musa balbisiana* transcriptome sequencing and assembly.

Table 2. Summary of Ion Torrent transcriptome assembly for *Musa balbisiana.*

Assembly quality parameters	
Contigs generated	82,413
Maximum contig length	3402
Minimum contig length	201
Average contig length ± SD	354.418 ± 216.44
Median contig length	444
Total contigs length	29,208,621
Total number of non-ATGC characters***	52,225
Percentage of non-ATGC characters***	0.1788
Contigs 201–500 bp	70,708
Contigs 501 bp–1 kbp	9862
Contigs 1–2 kbp	1737
Contigs 2–3 kbp	101
Contigs >3 kbp	5
N50 value	351

***: IUPAC nucleotide codes.

3.2. Sequence annotation

Several bioinformatic resources have been used to annotate the assembled sequences. The unigenes were annotated by aligning them with those in the public domain of diverse protein databases including UniProt/Swiss-Prot, UniProt/TrEMBL, the Cluster of Orthologous Groups of Proteins (COG), and the Plant Metabolic Pathway Database (PMN/PlantCyc). An E-value of less than 10^{-5} was used as the selection criterion for annotation (Altschul et al., 1997) and the details of functional annotation are given in Table 3.

Out of 82,413 unigenes, 57,540 (69.81%) were successfully annotated in the Swiss-Prot, TrEMBL, COG, and PlantCYC databases, while a maximum of 35,783 (62.18%) had similarity to proteins in the Swiss-Prot databases. The present transcriptome study on *Musa* revealed that more than 30% of unhit unigenes could be successfully exploited for the discovery of new genes specific to the *Musa* genome as reported by Meyer et al. (2009).

Organism distribution based on the BLASTx analysis of the unigenes showed that the transcripts hit a range of plant species. *Arabidopsis* had the highest number of hits (65%), followed by *Oryza sativa* (23.48%), *Musa acuminata* (2.88%), *Zea mays* (2.52%), *Solanum lycopersicum* (1.74%), *Solanum tuberosum* (1.62%), *Nicotiana tabacum* (1.50%), and *Glycine max* (1%). The extent of hits with *Vitis vinifera* was only 0.95%, while hits to nonplant organisms made up a total of 0.25% (Figure 2). Only a small number of unigenes (2373) matched with those of *Musa acuminata*, indicating that the *Musa balbisiana* transcriptome has identified a large group of genes that were unidentified in *Musa acuminata.*

Table 3. Functional annotation of the *M. balbisiana* transcriptome.

Annotated databases	All sequences	≥300 bp	≥1000 bp
Swiss-Prot annotation	35,783	19,781	1504
TrEMBL annotation	56,967	30,053	1832
Go annotation	34,781	19,246	1474
KOG annotation	32,179	18,076	1251
Hypothetical protein	21,890	11,933	743
Total	57,540	33,621	3628

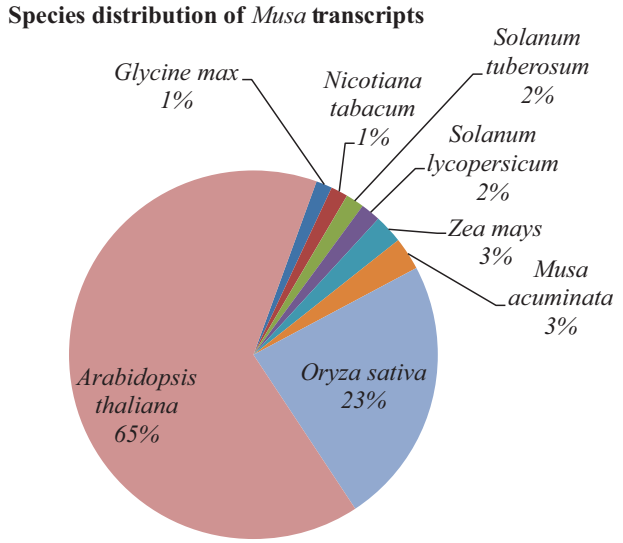

Figure 2. Species distribution of *Musa balbisiana* unigenes.

3.3. Functional characterization by GO annotation

The unigenes were annotated using the BLASTx program against the Swiss-Prot database and then mapped against the GO database to retrieve the GO terms. A total of 193,826 GO terms were assigned to all 35,783 mapped unigenes, with an average of one unigene assigned to three main categories. This is the least occurrence of GO terms compared to cultivars rich in the A genome, namely Calcutta 4 (AA) (341,244) (Passos et al., 2013) and Grand Naine (AAA) (351,220) (Li et al., 2012). Of these, the majority of GO terms were assigned to biological processes (72,012 (37.2%)), followed by molecular function (62,685 (32.4%)); the fewest were categorized as cellular components (59,039 (30.5%)). In our study, biological process ontology distribution showed that the unigenes were mainly assigned to "transcription DNA-dependent" (3235), "defense" (2434), and "salt stress" responses (1798). This information will be helpful for studies investigating the regulatory factors that control different promoter sites, as well as to detect the defense-related genes and salt stress tolerance genes present in the B genome. Under the molecular function ontology, many unigenes were assigned to "ATP binding" (6433), while under the cellular component ontology, unigenes were assigned to "nucleus" (5692) and "integral to membrane" (62,685) (Figure 3). Molecular functions generally correspond to activities that can be performed by individual gene products, or sometimes by assembled complexes of gene products, like catalytic activity, transporter activity, or binding. This specific molecular function GO term can be very useful for functional evaluation. They allow for more detailed host–pathogen interactions and enable a more focused

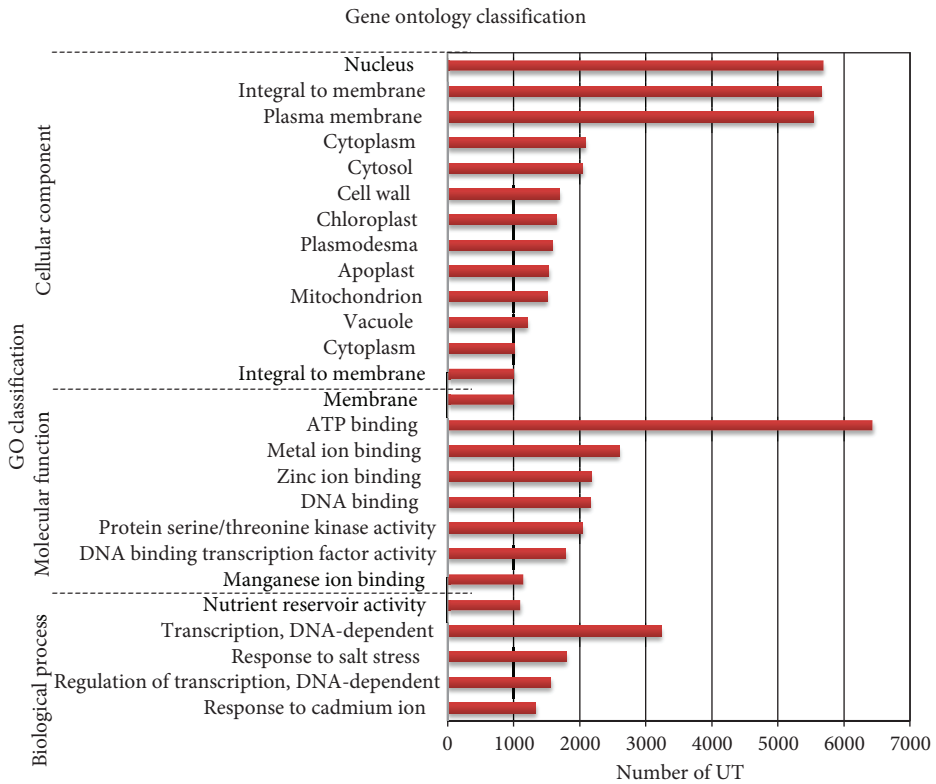

Figure 3. Histogram representation of gene ontology classification of *Musa balbisiana*.

molecular strategy for the experimental validation. All these categories indicate that the *Musa* B genome of banana undergoes multiple processes of defense mechanisms, developmental processes, and stress responses.

3.4. COG classification

COGs were delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. The current COG database contains both prokaryotic clusters (COGs) as well as eukaryotic clusters (KOGs), which are classified into 25 functional classifications. In order to find out the predicted function and classification of *M. balbisiana* unigenes, all were subjected to a search against the KOG database. Annotation of 82,413 sequences with the Swiss-Prot database revealed that 36,370 sequences had high homology with the known proteins of other species. The Swiss-Prot database could be used to assign KOG classifications (Figure 4). Putative KOG-annotated proteins were functionally classified into at least 25 protein families involved in cellular structure, metabolism, molecular processing, signal transduction, etc. (Figure 4). The cluster for general function prediction (6154 (16.92%)) represented the largest group in banana and a similar trend was also reported by Liu et al. (2012) in bamboo.

General function prediction (6154) was followed by posttranslational modifications, protein turnover, chaperones (4340 (11.93%)), signal transduction mechanisms (3446 (9.47%)), intracellular trafficking, secretion, vesicular transport (2298 (6.32%)), translation, ribosomal structure and biogenesis (2145 (5.90%)), carbohydrate transport and metabolism (2093 (5.75%)), transcription (1685 (4.63%)), function unknown (1676 (4.61%)), and amino acid transport and metabolism (1549 (4.26%)). Only a few unigenes were assigned to cell motility and extracellular structures (13 and 78 unigenes, respectively). In addition, 567 unigenes were assigned to cell wall/membrane/envelope biogenesis and 217 unigenes were assigned to defense mechanisms in KOG classification (Figure 4). Comparison of GO terms and KOG analysis resulted in maximum difference which confirmed and emphasized the uniqueness of banana unigenes with special reference to defense-related genes.

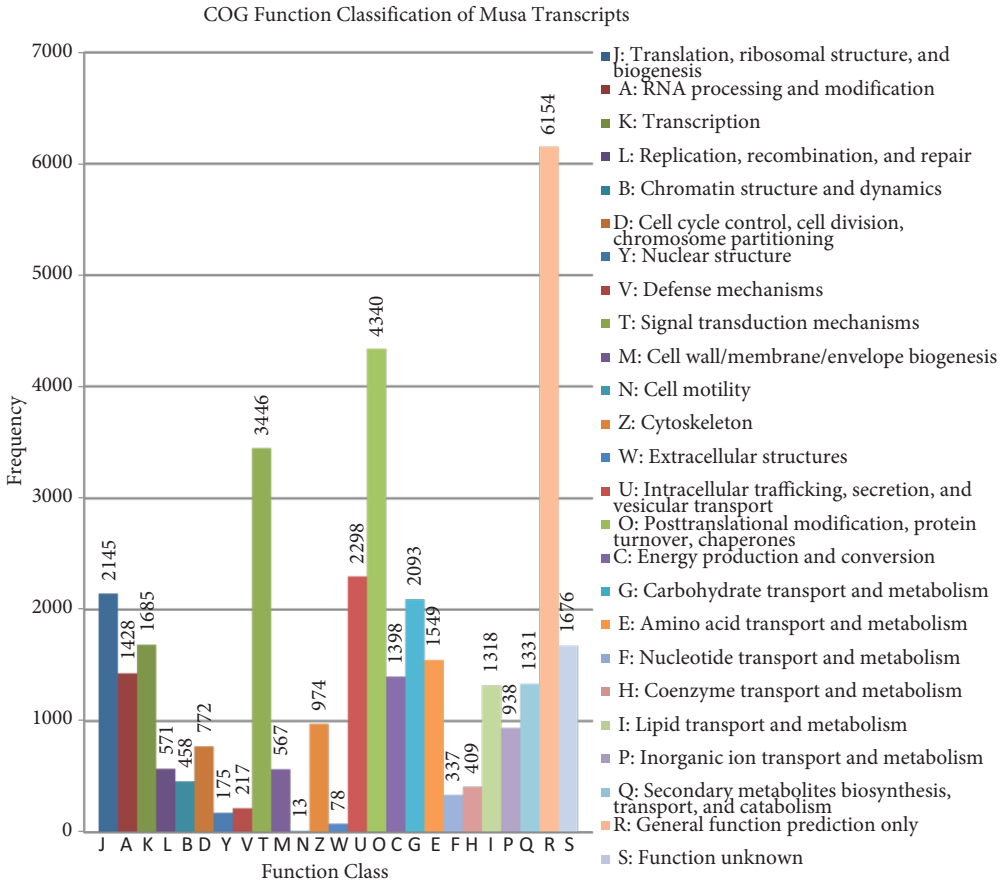

Figure 4. Clusters of Orthologous Groups (COG) classification. In total, 36,370 of the 82,413 sequences with Nr hits were grouped into 25 COG classifications.

3.5. Functional classification by PlantCYC

Annotations of *Musa* unigenes were fed into the PlantCYC database to categorize gene functions with an emphasis on biochemical/metabolic pathways (Zhang et al., 2010). Later, the assembled unigenes were assigned according to the Enzyme Commission, (EC), which resulted in 21,399 unigenes being assigned to 600 enzymes and grouped into 445 plant biochemical pathways. These predicted pathways represented the majority of plant biochemical pathways for compound biosynthesis, degradation, utilization, assimilation, and pathways involved in the processes of detoxification and generation of precursor metabolites and energy.

Higher portions of enzymes were found to participate in biosynthesis (241 pathways), biodegradation (56 pathways), photosynthetic cycles (15 pathways), general pathways (12 pathways), and others (121 pathways). Enzymes catalyzing almost all steps in several major plant metabolic pathways, including glycolysis (963), superoxide radical degradation (664), ethylene biosynthesis (594), jasmonic acid biosynthesis (369), Calvin cycle (344), gluconeogenesis (279), the pentose phosphate pathway (119), starch biosynthesis (108), anthocyanin biosynthesis (53), and several important secondary metabolite biosynthesis pathways including the thioredoxin pathway (11), could be represented by unigenes derived from the *Musa* dataset. Moreover, genes involved in several signaling pathways, including the mitogen-activated protein kinase signaling pathway, were also found in the unigene collection. This result indicates that there is a great deal of cell wall activity, as well as starch and sucrose synthesis, involved in the defense mechanisms of banana. A similar trend was also observed in other fruit crops, like apple (Newcomb et al., 2006) and kiwi (Crowhurst et al., 2008).

3.6. Annotation of defense genes and pathways

To understand the defense system of the banana *Musa* B genome, defense-related genes alone were short-listed from the blast analysis of *Musa* B transcriptome data. Based on an E-value of $<10^{-5}$, a total of 3301 unigenes were shortlisted as defense-related genes and subjected to further metabolic pathway analysis using the PlantCYC database. This analysis revealed that these unigenes were significantly enriched in various resistance-relevant metabolic or signaling pathways (Figure 5). This suggested that such defense-related genes and pathways were highly conserved among various genera. Similar results were also reported in cotton (Zhang et al., 2010) and banana (Li et al., 2012).

The selected pathways included perception of pathogen-associated molecular patterns (PAMPs) by pattern recognition receptors (PRRs), effector-triggered immunity (ETI), ion fluxes, transcription factors, oxidative burst, pathogenesis-related proteins, programmed cell death, plant hormones, and cell wall modification, among others. Among all defense-related transcripts, the highest number of transcripts (953) were involved in PAMP. PAMP-triggered immunity is a branch of plant immunity that involves interactions between host PRRs and PAMPs (Nürnberger and Kemmerling, 2009). Similarly, it was observed that 816 and 346 transcripts were hit with plant hormones biosynthesis and in cell-wall modification genes, respectively. The presence of 54 transcripts of ETI suggested that the high number of transcripts related to various

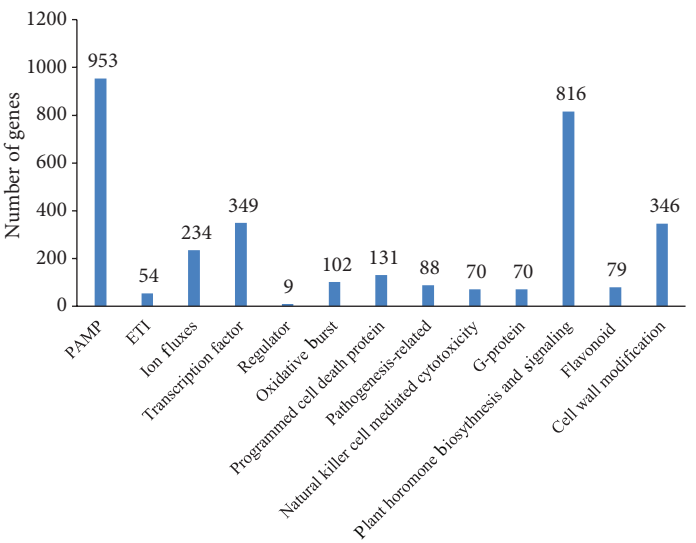

Figure 5. Pathway assignments of defense-related genes in whole transcriptome of *Musa balbisiana*.

defense-related pathways might be due to coevolution of plant resistance R-protein receptors and specific pathogen effector molecules responsible for disease resistance (Jones et al., 2006). Similarly, Passos et al. (2013) reported that the majority of unigenes are potentially involved in ETI, PAMP-triggered immunity, and PAMP pathways in the transcriptome analysis of AA genome. In the same way, many transcription factors, such as WRKY, bHLH, ethylene-responsive transcription factor 1, UNE12, RAP1, HBP-1a, BIM, ATB2, ILR3, ATR2, TCP20, HY5, BIM1, BIM2, and GLABRA 3, are responsible for resistance to the majority of potential pathogens. The present study gave a clear picture about the defense genes of *M. balbisiana*, which could lead to further in-depth study on developing banana cultivar lines resistant to various biotic and abiotic stresses.

3.7. SSR discovery

SSRs are highly informative and widely used in genetics, evolution, and breeding studies. To explore SSR profiles in the unigenes of *M. balbisiana*, the 82,413 unigene sequences were submitted to the MISA Perl program to identify SSRs. A total of 4780 SSRs were obtained, with 348 sequences containing more than one SSR. Similarly, Passos et al. (2013) also reported the presence of SSRs in Calcutta-4 (4068) and Cavendish (4095) clones belonging to the A genome.

About 7.3% of the *Musa* unigene contigs possessed more than one SSR, similar to the ESTs of wheat (7.41%); higher than grapes (2.5%), barley (2.8%), and flax (3.5%); and lower than coffee (18.5%) (Scott et al., 2000; Varshney et al., 2006). Exploration of SSR profiles in both genomes (AA and BB) is necessary to understand the evolution of *Musa* species, since most of the commercial cultivars are interspecific hybrids of the aforementioned genomes. This enormous number of *Musa* genic-SSRs obtained from AA and BB genomes will be helpful for understanding the functional genetic variation, which will be further utilized for DNA fingerprinting and evolutionary studies. The genic-SSR frequency also depends on the parameters used in exploring SSR markers, i.e. the repeat length and number of repeat unit thresholds. Margulies et al. (2005) implied that mononucleotide repeats should not be considered for the sequence data obtained through the Ion Torrent platform because of the quality problem associated with homopolymers. Thus, in this study, only di-, tri-, tetra-, penta-, and hexanucleotide repeats were taken into consideration. The abundance of genic-SSR (kbp/SSR) in *Musa* was 8.4, compared to rice (3.4), wheat (5.4), soybean (7.4), tomato (11.1), *Arabidopsis* (14.0), and cotton (20.0) (Cardle et al., 2000; Peng et al., 2005). The trinucleotide pattern was the most abundant, accounting for 57.64% of repeats, followed by dinucleotide (33%), tetranucleotide (1.23%), pentanucleotide (0.25%), and hexanucleotide (0.66%). The total number of compound SSRs was 281 (Table 4) (Passos et al., 2012).

Generally, the shorter the nucleotide core sequence, the greater the number of repeats observed. In case of *M. balbisiana*, an average of 9.4 repeats for di-, 5.9 for tri-, 5.8 for tetra-, 5.2 for penta-, and 5.3 for hexanucleotide motifs were observed, which is higher than the results obtained in *M. acuminata* accessions, namely Calcutta 4 (AA). More repeating nucleotide base pairs were observed in Grande Naine (AAA) (Backiyarani et al., 2013; Passos et al., 2013) than in other genomic groups. Petit et al. (2005) reported that SSR loci with more repeats generally exhibited higher mutation rates, probably because DNA slippage increases in proportion to the number of repeats. Hence, the *Musa* B genome might have more diversity than AA owing to higher mutation rates in the lengthy SSR loci.

The AG/CT motifs accounted for 79% of the total number of dinucleotide SSRs, similar to that of *Huperzia serrata* Thunb. (Luo et al., 2010). The most common motif for trinucleotide repeats of SSRs was AGG/CCT (25%) and AAG/CTT (24.8%), as in case of *M. acuminata* (Passos et al., 2012).

SSRs were developed as powerful molecular markers for comparative genetic mapping and genotyping. The SSRs were selected since they are ubiquitous in transcriptomes, typically locus-specific and codominant, multiallelic, highly polymorphic, and transportable among species and within genera (Varshney et al., 2005). EST databases have been a rich source of SSRs for genotyping numerous species

Table 4. Summary of SSR types found in the *M. balbisiana* transcriptome.

No. of repeat motifs	Dinucleotide	Trinucleotide	Tetranucleotide	Pentanucleotide	Hexanucleotide
Highest repeat motifs	AG/CT (1023)	AGG/CCT (569)	AAGG/CCTT (10)	AAGAG/CTCTT (2)	AAGACG/CGTCTT (3)
Lowest repeat motifs	CG/CG (6)	ACT/AGT (23)	AAAC/GTTT (1)	ACACT/AGTGT (1)	AACAGC/CTGTTG (1)
Total	1287	2248	48	10	26
Percentage	33%	57.64%	1.23%	0.25%	0.66%

of flowering plants (Yu et al., 2004). The unigenes obtained from *Musa* have provided a good resource for SSR mining and their applications in research and molecular marker-assisted breeding.

A total of 2628 SSR primer pairs were designed for the above-mentioned SSR-containing sequences using BatchPrimer3 v1.0 software (You et al., 2008). This is in line with the findings of Passos et al. (2012), who were able to design primers for only 50% of the SSR motifs in AA and AAA genomes of *Musa* spp. This could be due to the smaller size of ESTs with SSR regions or the presence of SSR motifs towards the end of sequences.

A total of 30 SSRs were validated to assess the polymorphism in 37 banana accessions across various genomes and in *Ensete*. Only 14 primers (46.6%) amplified products resulting in discrete, repeatable amplicons and were considered for analysis. They produced 56 alleles with a mean of 3.92 alleles per primer based upon the presence (1) and absence (0) of alleles. SSRs 14 and 23 had the highest number (7) of alleles and 17 and 26 had the lowest (2). PIC values ranged from 0.63 (SSR 26) to 0.88 (SSRs 14 and 21) with an average of 0.78 (Tables 5 and 6). The dendrogram delineated the 37 accessions into 2 major clusters with similarity near 65% (Figure 6). Cluster 1 contained B genome accessions, including 2 subclusters, while cluster 2 included accessions with the *Musa* A genome and other wild exotic accessions, along with the genus *Ensete*, and consisted of 2 subclusters.

Cluster 1 consisted of *M. balbisiana* accessions sharing 35% dissimilarities, which indicated that each of the accessions tested are unique, as reported earlier in morphotaxonomic (Sotto and Rabara, 2000) and molecular characterizations (Uma et al., 2006b). Cluster 2 consisted of the accessions belonging to the sections *Eumusa*, *Rhodochlamys*, and *Australimusa* and the genus *Ensete*. *Rhodochlamys* and *Eumusa* accessions clustered together and proved their genetic proximity. This is in accordance with the previous results of Wong et al. (2002) using AFLP markers. Although further studies are needed to select markers that show polymorphism, these data will provide a powerful tool for the identification of markers linked to specific traits. Furthermore, SSR markers can contribute to the construction of genetic linkage maps, genetic identification, and lineage analysis in *Musa* species.

To our knowledge, this is the first report to investigate the whole transcriptome of *M. balbisiana* where the assembly of the reads has been done using banana AA genome sequences as a reference map. The dataset is expected to improve our understanding of the molecular mechanisms involved in plant defense, biosynthesis,

Table 5. Validation of microsatellite loci isolated from *M. balbisiana* transcriptome.

Transcript ID	SSR name	SSR repeat motif	SSR locus length (bp)	Product size (bp)	Number of alleles	H. value	PIC value
banana_c8129	SSR4	(TC)11	22	172	4	0.84	0.82
banana_c7374	SSR6	(GAC)7	21	191	3	0.79	0.76
banana_c12500	SSR9	(GAC)7	21	151	3	0.76	0.72
banana_rep_c50441	SSR10	(GGA)7	21	153	3	0.79	0.76
banana_c5116	SSR11	(GAC)7	21	154	4	0.82	0.80
banana_rep_c52057	SSR12	(CTC)7	21	131	6	0.87	0.86
banana_c69	SSR13	(CCT)7	21	150	5	0.84	0.82
banana_c9829	SSR14	(TTC)8	24	128	7	0.89	0.88
banana_rep_c61381	SSR17	(TCC)7	21	142	2	0.74	0.69
banana_c19964	SSR21	(ATA)9	27	159	7	0.89	0.88
banana_c773	SSR23	(CAG)10	30	158	4	0.83	0.80
banana_c40221	SSR26	(AAG)9	27	180	2	0.68	0.63
banana_c564	SSR27	(TC)15	30	141	3	0.82	0.79
banana_c9112	SSR29	(GT)13	26	151	3	0.79	0.76
Average					4	0.81	0.78

Table 6. Transcriptome-derived SSR primer details with annotation results.

SSR name	Accession number	SSR	Primer sequence	Product size (bp)	Predicted function
SSR1	banana_rep_c53328	(CT)11	TGATGGAGGTAATGGACGAGA CAATCTAAAACGAGGAGGAGGA	143	1-aminocyclopropane-1-carboxylate oxidase
SSR2	banana_rep_c49856	(CT)11	CCACCTTCGGAGCTCTTCTAT GGGTTCTCCCTCAGGATCTC	118	Chalcone synthase 8
SSR3	banana_rep_c132357	(CT)11	ACGTCTCGAATTTCAGTGCAG GGGAGGCTTCGATTCATCTTA	157	Chitinase 3
SSR4	banana_c8129	(TC)11	CACTTCTCTCCCTGCCTTCTC CAGTGTAGCCTCTACGCAGGA	172	Cinnamoyl-CoA reductase 2
SSR5	banana_c24462	(AG)12	TACGATGGATTCCACCTCATC GTACACCTTCACCCCCATTCT	150	Glutathione S-transferase 3
SSR6	banana_c7374	(GAC)7	CTCTTACCCTCCGACGATACC AACTCCCTCTCCTCCTCCTCT	191	Indole-3-acetaldehyde oxidase
SSR7	banana_c10989	(CGA)7	CACGAACACCCCTGCTTC AAGAAAGGCAGCTCTGTGATG	193	Malate dehydrogenase, mitochondrial
SSR8	banana_rep_c70529	(TGA)8	CGAGTGGAAGAAGTGGAACTG CCTGACACATTGAGCCTAGCA	147	Pectate lyase
SSR9	banana_c12500	(GAC)7	GTCGAAGCGGAAGAGAACAG CCCCGAAGGTTTAAACAAAGA	151	Probable beta-1,3-galactosyltransferase 20
SSR10	banana_rep_c50441	(GGA)7	AGCTGCTGAGGTACCAAAACG CTCTTCGACTTCGGATCTCCT	153	Probable leucine-rich repeat receptor-like protein kinase
SSR11	banana_c5116	(GAC)7	TGACGATGAGGAAGAGGAAGA CACTTGGCGGTGATACTCCT	154	Protein kinase G11A
SSR12	banana_rep_c52057	(CTC)7	TCCTTCTGCTTCTTCCTTCAA AGGTCGTCGAAGTCTCACAGA	131	S-adenosylmethionine synthase 2
SSR13	banana_c69	(CCT)7	CTTTCTCGGATGGCTCTTCTT GTTGGTGGGATCGGTGAG	150	Somatic embryogenesis receptor kinase 1
SSR14	banana_c9829	(TTC)8	ACCTCTTTGTTTTCCGGTTTC CCACACATCCCCCATACATTA	128	Spore coat protein, putative
SSR15	banana_c9222	(AG)11	GTTCAGGTTAGGCATTTGGTG CCAGCTTCAGCTTCACCATC	149	Squalene synthase
SSR16	banana_c11645	(GA)11	TCCATTGGTACCAGAGATTGC CAGCTTACAACCTCCTCTCTCTC	173	Ubiquitin-conjugating enzyme
SSR17	banana_rep_c61381	(TCC)7	TGGTGGGAAATAGGAAACG GGGATCGACGAAGATGAAGAT	142	UDP-arabinopyranose mutase 1
SSR18	banana_rep_c120459	(GTA)10	CCCCCTCAAGAAGCTTTATCT GACCAGGTCTGGATGTTGATG	154	ATP-dependent Clp protease proteolytic subunit 3
SSR19	banana_rep_c197185	(AGA)9	AGGAGCATGGGGAGAAGAAG GGAAGGAGATGAGAGGAGGAA	156	Beta-galactosidase 1 (Lactase 1)
SSR20	banana_c21865	(GTA)10	AAGCTTTATCTCGCTTCAGCA CAAGGTTGATTGATGATCAGG	150	Peroxidase 70
SSR21	banana_c19964	(ATA)9	GCAAATTACAAGATCGGCAAG GCTCAGAACCAGTCTCTTTCG	159	Phosphoenolpyruvate carboxykinase [ATP] 1
SSR22	banana_c3790	(AG)19	AAGCCACTTAAACCACAACGA CTGATGCCTGGTACAGAGAGG	155	Probable galacturonosyltransferase 15
SSR23	banana_c773	(CAG)10	TATCTCCATTGCTCCACTTCG TTGTGGTGTTGCTTACACTCG	158	Probable xyloglucan endotransglucosylase/ hydrolase
SSR24	banana_rep_c46035	(GA)15	TGTGTGTGTGTGTTGTTGTGC CTCTCACACACACACACAGCA	140	Protein-tyrosine-phosphatase MKP1
SSR25	banana_c5296	(CAG)9	ATCGCTTCCTTCTGGAGAGTG GCTCCTCGAACTCTCCTTCC	151	UDP-glucose 4-epimerase 2
SSR26	banana_c40221	(AAG)9	GTGTGGCAACTGAGAAGCACT TGCCATGACTTCATTCACAAG	180	Uridine cytidine kinase I, putative
SSR27	banana_c564	(TC)15	TTAGGTGAGATGGCAGCATTC CAACCTTCAGCCATGCACTAT	141	Probable UDP-N-acetylglucosamine--peptide N-acetylglucosaminyltransferase
SSR28	banana_rep_c79367	(CT)11	GATTTCAGGCAACAGGGTACA ACAAGAGAGCCAGCGACATTA	150	Putative vesicle-associated membrane protein 726
SSR29	banana_c9112	(GT)13	TCACAGCAGCCAACGATAGTA GGAATTCGATCAGAGTTTCCA	151	Probable WRKY transcription factor 71
SSR30	banana_rep_c72184	(TTA)12	GACTGCGGCTACGATTACAAG GCACCCGGTGTTTGTTTTATT	190	Zinc finger A20 and AN1 domain-containing stress-associated protein 5

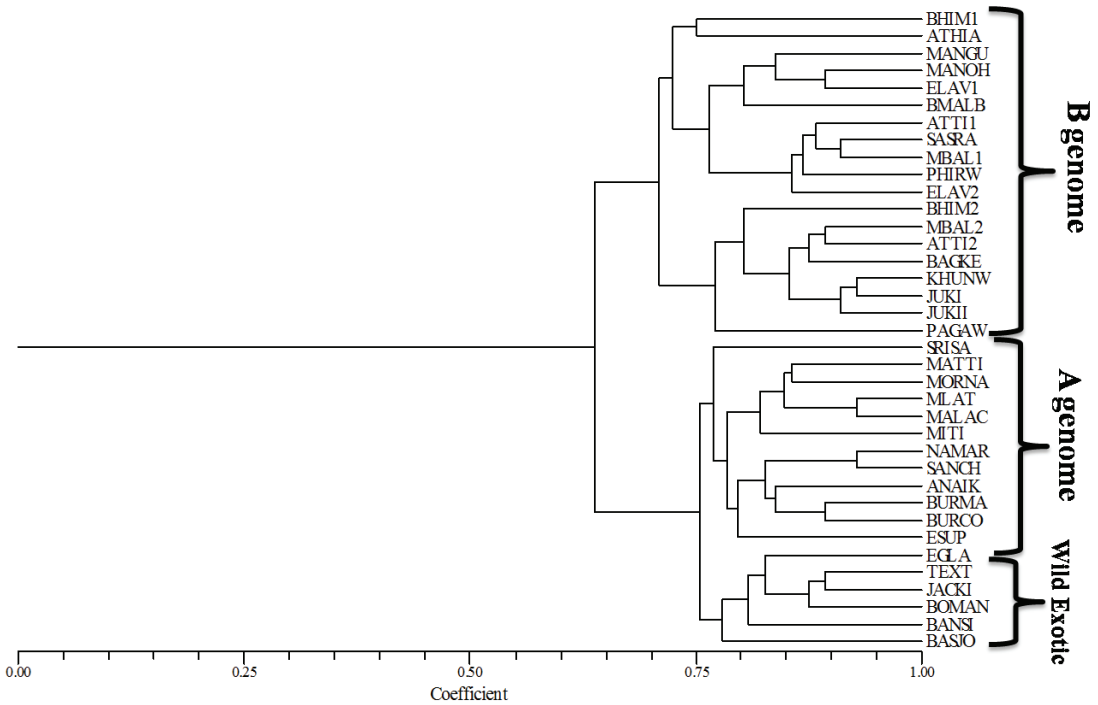

Figure 6. Dendrogram showing the phylogenetic relationship between *Musa* B and other genomes in banana based on defense-related EST-SSR markers derived from B genome whole transcriptome.

and other biochemical processes in *Musa*. This will be a potential resource for future genetic or genomic studies on *Musa* species and is expected to bridge a critical gap existing in banana comparative genomics. Consequently, this will contribute to the evolutionary and functional studies of plant genes and genomes.

Acknowledgments

This study was supported by the Indian Council of Agricultural Research (ICAR), India, under the project NPTC - Functional Genomics. We express our gratitude to the Director of the ICAR-National Research Centre for Banana for the facilities provided for the conduction of this study.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M et al. (2004). UniProt: the universal protein knowledgebase. Nucleic Acids Res 32: D115–D119.

Backiyarani S, Uma S, Varatharj P, Saraswathi MS (2013). Mining of EST-SSR markers of *Musa* and their transferability studies among the members of order the Zingiberales. Appl Biochem Biotechnol 169: 228–238.

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000). Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156: 847–854.

Crowhurst RN, Gleave AP, MacRae EA, Ampomah-Dwamena C, Atkinson RG, Beuning LL, Bulley SM, Chagne D, Marsh KB, Matich AJ et al. (2008). Analysis of expressed sequence tags from Actinidia: applications of a cross species EST database for gene discovery in the areas of flavor, health, color and ripening. BMC Genomics 9: 351–377.

Davey MW, Gudimella R, Harikrishna JA, Sin LW, 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: 683–703.

Deschamps S, Llaca V, May GD (2012). Genotyping-by-sequencing in plants. Biology 1: 460–483.

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488: 213–217.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. (2011). Full length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29: 644–652.

Hobert O (2010). The impact of whole genome sequencing on model system genetics: get ready for the ride. Genetics 184: 317–319.

Jones JDG, Dangl JL (2006). The plant immune system. Nature 444: 323–329.

Lescot M, Piffanelli P, Ciampi AY, Ruiz M, Blanc G, Leebens-Mack J, da Silva FR, Santos CM, D'Hont A, Garsmeur O et al. (2008). Insights into the Musa genome: Syntenic relationships to rice and between *Musa* species. BMC Genomics 9: 58.

Li CY, Deng GM, Yang J, Viljoen A, Jin Y, Kuang RB, Zuo CW, Lv ZC, Yang QS, Sheng O et al. (2012). Transcriptome profiling of resistant and susceptible Cavendish banana roots following inoculation with *Fusarium oxysporum* f. sp. *cubense* tropical race 4. BMC Genomics 13: 374–385.

Liu M, Qiao G, Jiang J, Yang H, Xie L, Xie J, Zhuo R (2012). Transcriptome sequencing and *de novo* analysis for Ma bamboo (*Dendrocalamus latiflorus* Munro) using the Illumina platform. PLoS ONE 7: e46766.

Liu Q, Zhu A, Chai L, Zhou W, Yu K, Ding J, Xu J, Deng X (2009). Transcriptome analysis of a spontaneous mutant in sweet orange [*Citrus sinensis* (L.) Osbeck] during fruit development. J Exp Bot 60: 801–813.

Luo H, Sun C, Li Y, Wu Q, Song J, Wang D, Jia X, Li R, Chen S (2010). Analysis of expressed sequence tags from the *Huperzia serrata* leaf for gene discovery in the areas of secondary metabolites biosynthesis and development regulation. Physiol Plant 139: 1–12.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al. (2005). Genome sequencing in micro fabricated high-density picolitre reactors. Nature 437: 376–380.

Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV (2009). Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics 10: 219–235.

Newcomb RD, Crowhurst RN, Gleave AP, Rikkerink EH, Allan AC, Beuning LL, Bowen JH, Gera E, Jamieson KR, Janssen BJ et al. (2006). Analyses of expressed sequence tags from apple. Plant Physiol 141: 147–166.

Nürnberger T, Kemmerling B (2009). PAMP-triggered basal immunity in plants. Adv Bot Res 51: 1–38.

Nwakanma DC, Pillay M, Okoli BE, Tenkouano A (2003). Sectional relationships in the genus Musa L. inferred from the PCR-RFLP of organelle DNA sequences. Theor Appl Genet 107: 850–856.

Ong WD, Voo LY, Kumar VS (2012). De novo assembly, characterization and functional annotation of pineapple fruit transcriptome through massively parallel sequencing. PLoS ONE 7: e46937.

Ortiz R, Swennen R (2014). From cross breeding to biotechnology-facilitated improvement of banana and plantain. Biotechnol Adv 32: 158–169.

Passos MA, de Oliveira Cruz V, Emediato FL, de Camargo Teixeira C, Souza MT Jr, Matsumoto T, Rennó Azevedo VC, Ferreira CF, Amorim EP, de Alencar Figueiredo LF et al. (2012). Development of expressed sequence tag and EST-SSR marker resources for *Musa acuminata*. AoB PLANTS 2012: pls030.

Passos MAN, de Cruz VO, Emediato FL, de Teixeira CC, Azevedo VC, Brasileiro AC, Amorim EP, Ferreira CF, Martins NF, Togawa RC et al. (2013). Analysis of the leaf transcriptome of *Musa acuminata* during interaction with *Mycosphaerella musicola*: gene assembly, annotation and marker development. BMC Genomics 14: 78.

Peng JH, Lapitan N L (2005). Characterization of EST-derived microsatellites in the wheat genome and development of eSSR markers. Funct Integr Genomics 5: 8–96.

Perrier X, De Langhe E, Donohue M, Lentfer C, Vrydaghs L, Bakry F, Carreel F, Hippolyte I, Horry JP, Jenny C et al. (2011). Multidisciplinary perspectives on banana (Musa spp.) domestication. P Natl Acad Sci USA 108: 11311–11318.

Petit RJ, Deguilloux MF, Chat J, Grivet D, Garnier-Géré P, Vendramin GG (2005). Standardizing for microsatellite length in comparisons of genetic diversity. Mol Ecol 14: 885–890.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2102). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.

Ravi I, Uma S (2011). Phenotyping bananas and plantains for adaptation to drought. In: Monneveux P, Ribaut JM, editors. Drought Phenotyping in Crops: From Theory to Practice. Texcoco, Mexico: CGIAR Generation Challenge Programme/ CIMMYT.

Ravi I, Uma S, Mayil Vaganan M, Mustaffa MM (2013). Phenotyping bananas for drought resistance. Front Physiol 4: 9.

Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ (2000). Analysis of SSRs derived from grape ESTs. Theor Appl Genet 100: 723–726.

Simmonds NW (1962a). Where our bananas come from. New Scientist (Reed Business Information) 16: 36–39.

Simmonds NW (1962b). The Evolution of the Bananas. London, UK: Longman Group Ltd.

Simmonds NW (1973). Bananas. 2nd ed. London, UK: Longmans, Green & Co.

Sotto RC, Rabara RC (2000). Morphological diversity of *Musa balbisiana* Colla in the Philippines. Infomusa 9: 28–30.

Swarupa V, Ravishankar KV, Rekha A (2013). Characterization of tolerance to *Fusarium oxysporum* f.sp., *cubense* infection in banana using suppression subtractive hybridization and gene expression analysis. Physiol Mol Plant Pathol 83: 1–7.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11: 1441–1452.

Uma S, Saraswathi MS, Durai P, Sathiamoorthy S (2006a). Diversity and distribution of section Rhodochlamys (Genus *Musa*, Musaceae) in India and breeding potential for banana improvement programmes. Plant Genetic Resource Newsletter 146: 17–23.

Uma S, Siva SA, Saraswathi MS, Manickavasagam M, Durai P, Selvarajan R, Sathiamoorthy S (2006b). Variation and intraspecific relationship in Indian wild *Musa balbisiana* (BB) population an evidenced by random amplified polymorphic DNA. Genet Resources Crop Evol 53: 349–355.

Varshney RK, Graner A, Sorrells ME (2005). Genic microsatellite markers in plants: features and applications. Trends Biotechnol 23: 48–55.

Varshney RK, Grosse I, Hahnel U, Siefken R, Prasad M, Stein N, Langridge P, Altschmied L, Graner A (2006). Genetic mapping and BAC assignment of EST-derived SSR markers shows non uniform distribution of genes in the barley genome. Theor Appl Genet 113: 239–250.

Wang Y, Chung SJ, Song WO, Chun OK (2011). Estimation of daily proanthocyanidin intake and major food sources in the U.S. diet. J Nutr 141: 447–452.

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.

Wong C, Kiew R, Argent G, Set O, Lee SK, Gon TY (2002). Assessment of validity in the sections in *Musa* (Musaceae) using AFLP. Ann Bot 90: 231–238.

You FM, Huo N, Gu YQ, Luo MC, Ma Y, Hane D, Lazo GR, Dvorak J, Anderson OD (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer designing. BMC Bioinformatics 9: 253–266.

Yu JK, La Rota M, Kantety RV, Sorrells ME (2004). EST derived SSR markers for comparative mapping in wheat and rice. Mol Genet Genomics 271: 742–751.

Zhang P, Dreher K, Karthikeyan A, Chi A, Pujar A, Caspi R, Karp P, Kirkup V, Latendresse M, Lee C et al. (2010). Creation of a genome-wide metabolic pathway database for Populus trichocarpa using a new approach for reconstruction and curation of metabolic pathways for plants. Plant Physiol 153: 1479–1491.