*Tree and Forestry Science and Biotechnology* ***©2010 Global Science Books***

# Transcriptome Analysis of Leaves and Roots of *Musa balbisiana* var. 'Pisang Klutuk Wulung'

**Candice Mello Romero Santos[1] • Claudia Fortes Ferreira[2] • Lizangela Pinheiro Cassiano[1] • Marly Catarina Felipe Coelho[1] • Roberto Coiti Togawa[1] • Natalia Florêncio Martins[1] • Manoel Teixeira Souza Júnior[3*]**

[1] Embrapa Genetic Resources & Biotechnology, Pq Estação Biológica, CP 02372. Final W5 Norte, Brasília, DF, CEP 70.770-900, Brasil
[2] Embrapa Cassava & Tropical Fruits, Rua Vitória s/n, CP 007, Cruz das Almas, BA, CEP 44380-000, Brasil
[3] Embrapa LABEX Europe. Plant Research International – Wageningen University & Research Centre, P.O. Box 16, 6700AA, Wageningen, The Netherlands

***Corresponding author*: * msouza@cenargen.embrapa.br**

## ABSTRACT

Banana and plantain (*Musa* spp.) are perennial herbs belonging to the Musaceae family. Almost the totality of cultivated banana genotypes originated from natural intra- and inter-specific hybridization of two diploid species; *Musa acuminata* Colla and *Musa balbisiana* Colla; representing the A and B genomes, respectively. This study reports the construction and characterization of two cDNA libraries from the *M. balbisiana* var. 'Pisang Klutuk Wulung' (BB), also known as PKW. Roots and leaves from PKW hydroponics plants were used to construct these two cDNA libraries. Sequencing of cDNA clones, followed by trimming and CAP3 assembling of high quality ESTs resulted in 2,967 MbAES (*Musa balbisiana* Assembled EST Sequences), with 435 contigs and 2,532 singletons. Approximately 38% of the MbAES came out as "no hits" after Blastx search against several public databases. Assembled EST sequences were classified into 23 different categories according to putative protein functions (KOG), whereas the range of similarity was from 35 to 60%. The classification of the EST clusters into the Gene Ontology (GO) function classes showed a predominance of catalytic activity and binding as molecular function. Among the most populated contigs there were genes related to pathogenicity response, stress response and photosynthesis. A differential digital display between *M. acuminata* data already stored at the DATA*Musa* and *M. balbisiana* assembled EST sequences revealed 10 unknown genes exclusive to the B genome and nine exclusive to the A genome with statistically significant difference.

**Keywords:** PKW, genomics, EST, genome B, transcriptomics

**Abbreviations: EST**, expressed sequence tag; **DDD**, digital differential display; **MbAES** *Musa balbisiana* Assembled EST Sequences; **PKW**, 'Pisang Klutuk Wulung'; **PERL**, practical extraction and reporting language; **RT**, reverse transcriptase; **STMS**, sequence tagged microsatellite site

## INTRODUCTION

Bananas and plantains (*Musa* spp.), which will be called "bananas" or *Musa* from now on, belong to the Musaceae family. It is a monocotyledon with a relative small genome distributed over eleven chromosomes. The *Musa* genus includes some 25 species, which have been divided into four sections: *Australimusa*, *Callimusa*, *Rhodochlamys*, and *Eumusa* (Simmonds and Shepherd 1955). The *Eumusa* section is the most widely geographically represented of these sections and contains the two major species, *M. acuminata* (A genome) and *M. balbisiana* (B genome), which are at the origin of the edible bananas.

Most of the edible bananas, cultivated and highly appreciated worldwide nowadays, derived basically from the several subspecies of *Musa acuminata*, as well as from hybrids with *M. balbisiana*. These edible bananas are mostly triploid, although some diploid and tetraploid varieties are also known and used in several countries (Daniells *et al.* 2001). The A and B genomes of *Musa* differ in size, according to estimates carried out by flow cytometry of nuclei stained by propidium iodide (Lysák *et al.* 1999). The average haploid genome size of *M. balbisiana* is 537 Mbp, while *M. acuminata* showed statistically significant variants among subspecies and clones, ranging from 591 to 615 Mbp. Therefore, the *Musa* B genome is only 15% larger than *Oryza sativa* L. ssp. *indica* (Yu *et al.* 2002), while the A genome is about 30% larger.

According to the FAO (http://faostat.fao.org/), in 2007, bananas and plantains were produced in 126 and 50 countries, respectively. The total area harvested and the total production quantity was, respectively, 4,410,509 hectares and 81,263,358 tonnes for bananas; and, 5,457,065 hectares and 34,444,795 tonnes for plantains. In 2006, only approximately 20% of the banana, and 10% of the plantains, produced worldwide entered the international market, generating about 6 billion US dollars in export value (97% banana, 3% plantain); being an important source of revenue for the developing countries producing this fruit crop. These numbers substantiate the general acceptance that besides being an important commodity for the economy of many developing countries, banana plays a major social and economic role in the tropics and sub-tropics; being a staple food for millions of people worldwide.

The production of an Expressed Sequence Tag (EST) database has commonly been the initial step in the transcriptomics of any given organism. An EST is a unique DNA sequence derived from a cDNA library, and therefore from a sequence which has been transcribed in some tissue or at some stage of development. It can be derived from a transcribed protein-coding or non-protein-coding nucleotide sequence, and it can be instrumental in gene discovery and sequence determination. Once identified, an EST can be mapped, by a combination of genetic mapping procedures, to a unique locus in the genome and can identify and characterize that gene locus. ESTs are also useful for designing probes for DNA microarrays that can be used to study the pattern of gene expression in one specific scenario of inter-

**Received: 31 January, 2009. Accepted: 15 January, 2010.**

**Original Research Paper**

est, such as infection by a plant pathogen. By identifying the up- or down-regulated genes during that specific moment, and further characterizing them, one can better understand that specific phenomenon, and consequently improve the chances of modifying it via conventional breeding strategies, or via genetic engineering. The production of an EST database is one of the strategies used for the characterization of the banana genome (Santos *et al.* 2005; Roux *et al.* 2008).

These EST databases can be mined for SSR (Simple Sequence Repeats) markers, which are tandem repeated DNA motifs (1-5 bp) being found in coding as well as in non-coding DNA regions of the genome (Kantety *et al.* 2002; Gupta *et al.* 2003; Tamana *et al.* 2005; Iniguez-Luy *et al.* 2008). Through this procedure, SSR markers can be obtained at very low costs, since EST derived SSRs (EST-SSRs) are a free by-product of the currently expanding EST databases (Gupta *et al.* 2003; Thiel *et al.* 2003; Varshney *et al.* 2005).

It is well known that the sequences flanking specific SSR loci in a genome are said to be conserved within a certain species and across species within a genus. These flanking sequences, therefore, have been used to design primers for individual SSR loci; a technique described as Sequence Tagged Microsatellite Site (STMS) analysis (Varshney *et al.* 2005). This approach has been useful for the development of SSRs from ESTs of many species (Kantety *et al.* 2002; Gupta *et al.* 2003; Asp *et al.* 2007; Senthilvel *et al.* 2008; Yang *et al.* 2008; Iniguez-Luy *et al.* 2008). According to Gupta *et al.* (2003), besides being cost-free, EST-SSRs offer other advantages over other genomic DNA-based markers, such as: (1) they can detect variation in the expressed portion of the genome, so that gene tagging should give “perfect” marker-trait associations, and (2) can also be used across a number of related species. Microsatellite markers are extremely important tools for assaying genetic variation and genetic analysis of crop plants, especially due to their reproducibility, co-dominant inheritance, multiallelic nature and good genome coverage; therefore being the marker of choice for breeders (Varshney *et al.* 2002; Gupta *et al.* 2003).

The present work describes the construction of two cDNA libraries from different tissues of *Musa balbisiana* var. ‘Pisang Klutuk Wulung’ (BB), and its consequent EST sequencing, clustering and annotation by assigning putative functions to the transcripts. These sequences were also investigated for abundance of EST-SSR markers.

## MATERIALS AND METHODS

### Plant material

Roots and leaves from *M. balbisiana* var. ‘Pisang Klutuk Wulung’ (BB), also known as PKW (ITC 1063), were collected from hydroponic plants about 25 cm high with abundant root system.

### Total RNA isolation and poly (A) + RNA purification

Total RNA was isolated using the Plant RNA Reagent kit (Invitrogen life technologies, USA), according to the protocols provided by the manufacturer. Total RNA preparations were then submitted to poly (A) + RNA purification using the Micro FastTrack 2.0 mRNA Isolation Kit (Invitrogen life technologies, USA), according to manufacturer’s instructions.

### Construction of cDNA libraries and DNA sequencing

cDNA libraries were constructed with the Creator Smart cDNA library kit and cloned into the pDNR-LIB vector (Clontech Laboratories, Inc., USA). Restriction digest and PCR profiles were also obtained according to the manufacturer’s instructions. The 5′- ends of the cDNA clones were sequenced at Embrapa Genetic Resources and Biotechnology’s DNA sequencing platform and at the Biotechnology Laboratory in the “Centro Avançado de Pesquisa Tecnológica do Agronegócio de Citros Sylvio Moreira (CAPTACSM)”, using the M13 forward primer (5′- TGT AAA ACG ACG GCC AGT - 3′) and Dye Terminator chemistry on automated sequencers.

### Bioinformatics analysis

Trace files were stored at the *Musa ESTs* database - DATA*Musa* (http://genoma.embrapa.br/musa/pt/DATA_musa.html). The raw data was analyzed using the EGassembler web server (http://egassembler.hgc.jp/), which provides an automated as well as a user-customized analysis tool for cleaning, repeat masking, vector trimming, organelle masking, clustering and assembling the of ESTs and genomic fragments (Masoudi-Nejad *et al.* 2006). Sequence comparison using Blastx (Altschul *et al.* 1997) was performed locally. The cutoff Evalue of $<10^{-5}$ was used to define the similar orthologs, and the unigenes set that did not meet this requirement were annotated as unknown. Several databases were used in the annotation step: GenBank nr (Benson *et al.* 2008), MIPS *Arabidopsis thaliana* (Schoof *et al.* 2004) and SwissProt (Gasteiger *et al.* 2001). Predicted protein sequences were aligned with Blastx against KOG - Eukaryotic Orthologous Groups (Tatusov *et al.* 2003) and Gene Ontology (Ashburner *et al.* 2000). The output was parsed by locally developed PERL scripts where the outputs of EST assembly and the result of several alignments were available in a web page to the group of manual curators. Results from BLAST tools against several data bases were used in the manual curation of unigenes as NR, MIPS, KOG and SwissProt. All tools were assembled in the DATA*Musa* database (Souza Júnior *et al.* 2005). *In silico* determination of differential genes from *M. balbisiana* and *M. acuminata* assembled sequences were performed by Audic and Claverie (1997) method and Stekel *et al.* method (2000) incorporated at SisGen (Pappas *et al.* 2008). Audic and Claverie method consists of a probability calculation of the distribution governing the occurrence of the same rare event in duplicate experiments where one event is the observation of a given cDNA sequence tag, and the experiment consists of the random picking and partial sequencing number *N* of cDNA clones (Audic and Claverie 1997). Although the Audic and Claverie method extends the Fisher´s exact test it is still applicable between two datasets considering the sampling size. Therefore, the Stekel method was applied to analyze multiple dataset. The selected libraries from *M. acuminata* were root and leaves of Calcutta 4, as described in Santos *et al.* (2005) and Souza Júnior *et al.* (2005). SSR markers were obtained using the EGassembler software (Masoudi-Nejad *et al.* 2006), which uses a RepeatMaster program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of this pipeline is a very detailed annotation of the di-, tri-, tetra- and pentameric repeats that are present in the query sequence. This comparison of the sequences in the RepeatMaster program is performed by the “Cross-Match”; an efficient implementation of the Smith-Waterman-Gotoh algorithm developed by Phil Green (Bedell *et al.* 2000).

## RESULTS AND DISCUSSION

Total RNA of high quality and in abundance was obtained from both tissues. After purification of mRNAs and production of the first cDNA strain by RT and ds cDNA by PCR, the products of amplification were separated by size and those over 500 bp were cloned in the pDNR-LIB vector. After ligation the vector containing the inserts were transformed into *Escherichia coli* TOP10 (Invitrogen Life Technologies, USA). The transformed bacteria were plated on LBA Medium supplemented with chloramphenicol. Fifteen single colonies were collected from each library for characterization of the average size of insert, as an initial step for validation of the cDNA library, before initiating sequencing. The two cDNA libraries produced contained clones with insert sizes ranging from 0.8 kb to 2.4 kb and an average length of ~1.6 kb. A total of 6,238 reads, 3,166 from leaves and 3,072 from roots, were generated out of these two cDNA libraries.

After submission to the EGassembler pipeline for gene transcripts (EST) fragments, a total of 521 and 428 reads

**Table 1** Annotation results in number of *Musa balbisiana* Assembled EST Sequences (MbAES) per cDNA library, and per combined assembling.

| Library | Sequences generated | Sequences analysed[1] | Singletons[2] | Contigs[3] | Unigenes[4] | Redundancy (%)[5] |
|---|---|---|---|---|---|---|
| Leaves | 3,166 | 2,645 | 1,237 | 200 | 1,437 | 54 |
| Root | 3,072 | 2,644 | 1,427 | 211 | 1,638 | 62 |
| Total | 6,238 | 5,289 | 2,532 | 435 | 2,967 | 56 |

[1] Vector sequences and sequences of low quality were eliminated
[2] The singletons present in each library independent of other libraries
[3] The contigs present in each library independent of other libraries
[4] The unigenes set for each library is the sum of singleton plus contigs for the library.
[5] The redundancy of each library calculated as 1 – (unigene library/number of sequence analyzed).

**Table 2** Most populated *Musa balbisiana* Assembled EST Sequences identified after assembling ESTs from leaf and root of *M. balbisiana* var. 'Pisang Klutuk Wulung' (BB) together.

| Contig Name | Number of Reads | Length (bp) | BlastX Results - SwissProt (expect threshold = 1) | E-value | Length _ Identity _ Frame |
|---|---|---|---|---|---|
| PKW_Leaf&Root_Contig37 | 269 (268 Leaf / 1 Root) | 1036 | Ribulose bisphosphate carboxylase small chain (GI:3914598) | 9.00E-98 | 180 aa _ 165/168 (98%) _ +2 |
| PKW_Leaf&Root_Contig31 | 180 (21 Leaf / 159 Root) | 833 | No Hits | | |
| PKW_Leaf&Root_Contig35 | 106 (7 Leaf / 99 Root) | 938 | Bowman-Birk type proteinase inhibitor DE-4 (GI:124035) | 0.001 | 80 aa _ 20/56 (30%) _ +2 |
| PKW_Leaf&Root_Contig185 | 102 (36 Leaf / 66 Root) | 342 | No Hits | | |
| PKW_Leaf&Root_Contig126 | 95 (85 Leaf / 10 Root) | 731 | Metallothionein-like protein type 3 (GI:2497907) | 9.00E-13 | 65 aa _32/33 (96%) _ +2 |
| PKW_Leaf&Root_Contig414 | 49 (9 Leaf / 40 Root) | 789 | No Hits | | |
| PKW_Leaf&Root_Contig227 | 45 (40 Leaf / 5 Root) | 825 | Photosystem II 10 kDa polypeptide (GI:131399) | 5.00E-48 | 138 aa _ 107/138 (77%) _ +3 |
| PKW_Leaf&Root_Contig106 | 38 (36 Leaf / 2 Root) | 641 | Metallothionein-like protein type 3 (GI:2497907) | 3.00E-23 | 65 aa _ 46/64 (71%) _ +1 |
| PKW_Leaf&Root_Contig359 | 36 (4 Leaf / 32 Root) | 727 | No Hits | | |
| PKW_Leaf&Root_Contig45 | 33 (32 Leaf / 1 Root) | 811 | Photosystem II 10 kDa polypeptide (GI:131399) | 5.00E-48 | 138 aa _ 107/138 (77%) _ +2 |
| PKW_Leaf&Root_Contig339 | 22 (Root) | 716 | No Hits | | |
| PKW_Leaf&Root_Contig211 | 18 (Root) | 820 | Bowman-Birk type proteinase inhibitor DE-4 (GI:124035) | 0.001 | 80 aa _ 20/56 (35%) _ +2 |
| PKW_Leaf&Root_Contig319 | 12 (Root) | 967 | Pathogenesis-related protein 1 (GI:548591) | 6.00E-51 | 158 aa _ 95/161 (59%) _ +1 |
| PKW_Leaf&Root_Contig234 | 10 (Root) | 878 | Bowman-Birk type proteinase inhibitor 2 (GI:51338751) | 2.00E-05 | 79 aa _ 24/61 (39%) _ +3 |
| PKW_Leaf&Root_Contig226 | 9 (Root) | 689 | Bowman-Birk type proteinase inhibitor 2 (GI:51338751) | 3.00E-06 | 79 aa _ 23/66 (34%) _ +1 |
| PKW_Leaf&Root_Contig229 | 9 (Root) | 713 | No Hits | | |
| PKW_Leaf&Root_Contig252 | 9 (Root) | 683 | Hydrophobic protein LTI6B (GI:122169560) | 3.00E-14 | 55 aa _ 47/53 (88%) _ +1 |
| PKW_Leaf&Root_Contig352 | 8 (Root) | 658 | Bowman-Birk type proteinase inhibitor 2 (GI:51338751) | 3.00E-06 | 79 aa _ 23/66 (34%) _ +1 |
| PKW_Leaf&Root_Contig294 | 7 (Root) | 874 | Germin-like protein subfamily 1 member 7 (GI:18203443) | 9.00E-66 | 229 aa _ 128/181 (70%) _ +1 |
| PKW_Leaf&Root_Contig14 | 50 (Leaf) | 291 | No Hits | | |
| PKW_Leaf&Root_Contig117 | 43 (Leaf) | 836 | Glycine cleavage system H protein (GI:152032493) | 1.00E-71 | 164 aa _ 129/164 (78%) _ +2 |
| PKW_Leaf&Root_Contig199 | 23 (Leaf) | 638 | No Hits | | |
| PKW_Leaf&Root_Contig84 | 12 (Leaf) | 545 | Metallothionein-like protein type 3 (GI:2497907) | 9.00E-13 | 65 aa _ 32/33 (96%) _ +1 |
| PKW_Leaf&Root_Contig101 | 12 (Leaf) | 461 | No Hits | | |
| PKW_Leaf&Root_Contig151 | 12 (Leaf) | 822 | 2-Cys peroxiredoxin BAS1 (GI:14916972) | 8.00E-70 | 266 aa _ 148/210 (70%) _ +1 |
| PKW_Leaf&Root_Contig195 | 11 (Leaf) | 671 | Protein SPT2 homolog (GI:82185481) | 0.018 | 800 aa _ 30/103 (29%) _ +1 |
| PKW_Leaf&Root_Contig5 | 10 (Leaf) | 802 | Carbonate dehydratase 1 (GI:1168738) | 3.00E-69 | 330 aa _ 87/123 (70%) _ +2 |
| PKW_Leaf&Root_Contig121 | 8 (Leaf) | 744 | Ribulose bisphosphate carboxylase small chain (GI:3914598) | 2.00E-47 | 180 aa _ 90/94 (95%) _ +2 |
| PKW_Leaf&Root_Contig135 | 8 (Leaf) | 724 | FYVE, RhoGEF and PH domain-containing protein 5 (GI:61213482) | 0.59 | 1462 aa _ 25/72 (35%) _ +1 |

were discarded from each library, leaves and roots, respectively. In total, 2,645 ESTs from leaves generated 2,076,863 bp of high quality sequence, with a GC level of 50.46%; while 2,644 ESTs from roots generated 2,077,247 bp, with a GC level of 49.45%. The average length of the ESTs, based on the high quality sequences from both libraries, was of 785 nucleotides. All 5,289 ESTs were submitted to NCBI (FL646263 to FL651551).

The assembling of these ESTs generated 435 contigs and 2,532 singletons - 1,158 singletons were from leaves and 1,374 from roots. Within the 435 contigs, 132 had exclusively ESTs from leaves, and 139 had exclusively ESTs from roots. A total of 164 contigs were composed of ESTs from both libraries. The 2,967 unigenes (contigs and singletons) were named MbAES (*Musa balbisiana* Assembled EST Sequence). The annotation of each MbAES obtained was performed by a semi-automated process of parsing the BLAST output format into a manually curating HTML page stored at DATA*Musa* database (http://genoma.embrapa.br/musa/index.html/DATA_musa.html). The manual curation considered the results of the BLAST against several databases, the KOG Functional Catalogue and Gene ontology. A summary of the annotation results is shown in **Table 1**. The overall novelty was 47.2% for leaves and 62.3% for roots,

**Table 3** Microsatellites identified in ESTs from *M. balbisiana* var. 'Pisang Klutuk Wulung' (BB). SSR repeat, EST of origin of the SSR, blastX results against SwissProt database.

| SSR | EST read | BlastX Results – SwissProt (expect threshold =1) | e-value | Length_Identity_Frame | SSR Location |
|---|---|---|---|---|---|
| $(AC)_{10}$ | Mbalbisiana_PKW_Root_009_g04_b | No hits | | | ni |
| $(AT)_{11}$ | Mbalbisiana_PKW_Leaf_024_C11_b | Photosystem I reaction center subunit V | 2.00E-52 | 160aa_104/140 (74%)_+1 | 3´UTR |
| $(AGG)_6$ | Mbalbisiana_PKW_Leaf020_E08_b | KH domain-containing, RNA-binding, signal transduction-associated | 0.019 | 443 aa _ 29/78 (37%) _ +2 | ni |
| $(AGC)_7$ | Mbalbisiana_PKW_Leaf_022b_F10_b | Thioredoxin M-type 2, chloroplastic (gi:12643846) | 7.00E-33 | 186 aa _ 55/109 (50%) _ +1 | 5´UTR |
| $(ATCC)_5$ | Mbalbisiana_PKW_Root_014_B12_b | Glycine cleavage system H protein | 6.00E-49 | 165aa_91/134 (67%)_+3 | 3´UTR |
| $(ATTA)_5$ | Mbalbisiana_PKW_Root_018_H12_b | No hits | | | ni |
| $(CT)_{13}$ | Mbalbisiana_PKW_Leaf_008_d04_b | Dehydration-responsive element-binding protein | 4.00E-11 | 236aa_54/148 (36%)_+1 | ni |
| $(CAA)_6$ | Mbalbisiana_PKW_Root_009_h09_b | no hits | | | ni |
| $(CCA)_6$ | Mbalbisiana_PKW_Leaf_023_f08_b | Thylakoid membrane phosphoprotein 14 kDa | 3.00E-34 | 174aa_86/161 (53%)_+2 | ORF |
| $(CTT)_6$ | Mbalbisiana_PKW_Root_015_F09_b | Alpha-1,4-glucan-protein synthase [UDP-forming] | 1.00E-30 | 364aa_17/105 (67%)_+2 | ORF |
| $(CAT)_7$ | Mbalbisiana_PKW_Leaf_007_d04_b | No hits | | | Ni |
| $(CTT)_7$ | Mbalbisiana_PKW_Leaf020_B03_b1 | Photosystem I reaction center subunit psaK, chloroplastic | 9.00E-32 | 129 aa _ 71/97 (73%) _ +2 | ORF |
| $(CTG)_7$ | Mbalbisiana_PKW_Root_030_c11_b | 60S ribosomal protein L12 (gi:6094002) | 4.00E-80 | 166 aa _ 146/166 (87%) _ +1 | 3´UTR |
| $(CTT)_8$ | Mbalbisiana_PKW_Leaf_024_D10_b | Transcription factor CPC | 5.00E-21 | 94aa_51/82 (62%)_+1 | ORF |
| $(CCT)_9$ | Mbalbisiana_PKW_Leaf_013_E04_b | Metallothionein-like protein type 3 | 0.013 | 65aa_19/28 (67%)_+3 | ni |
| $(CTT)_{11}$ | Mbalbisiana_PKW_Leaf_022b_B04_b | Photosystem I reaction center subunit psaK, chloroplastic | 3.00E-33 | 130 aa _ 87/120 (72%) _ +3 | 3´UTR |
| $(CCTG)_5$ | Mbalbisiana_PKW_Leaf_006_e04_b | Transcription factor TFIIIB component B | 9.00E-06 | 594aa_24/46 (52%)_+2 | 3´UTR |
| $(GA)_{11}$ | Mbalbisiana_PKW_Leaf_018_H04_b | Photosystem I reaction center subunit IV | 1.00E-23 | 125aa_51/56 (91%_+2 | 5´UTR |
| $(GCA)_6$ | Mbalbisiana_PKW_Leaf_009_d12_b | No hits | | | ni |
| $(GCC)_6$ | Mbalbisiana_PKW_Leaf_007_a04_b | SEL1 - like repeat-containing protein KIAA0746 | 5.10E-01 | 1,137 aa _ 37/122 (30%) _ +1 | ni |
| $(GAA)_7$ | Mbalbisiana_PKW_Root_016_G03_b | No hits | | | ni |
| $(GAT)_7$ | Mbalbisiana_PKW_Leaf_014_D04_b | Elongation factor 1-delta | 2.00E-49 | 229aa_79/123 (64%)_+1 | ORF |
| $(GGT)_7$ | Mbalbisiana_PKW_Leaf_027_g08_b | Zinc finger protein 521 | 0.006 | 1311aa_17/37 (45%)_-1 | ni |
| $(GAT)_8$ | Mbalbisiana_PKW_Root_028_g03_b | Elongation factor 1-delta 1 (gi:6166140) | 2.00E-51 | 229 aa _ 154/223 (69%) _ +1 | ORF |
| $(GCC)_8$ | Mbalbisiana_PKW_Leaf004_A06_b | No hits | | | ni |
| $(GAA)_{10}$ | Mbalbisiana_PKW_Leaf_012_B12_b | No hits | | | ni |
| $(GGAT)_6$ | Mbalbisiana_PKW_Root_016_G01_b | 60S ribosomal protein L44 (gi:2500380) | 2.00E-52 | 105 aa _ 95/105 (90%) _ +1 | 3´UTR |
| $(TC)_{10}$ | Mbalbisiana_PKW_Leaf_016_B11_b | Lipid transfer-like protein VAS | 8.00E-06 | 151aa_19/50 (38%)_+3 | 5´UTR |
| $(TA)_{11}$ | Mbalbisiana_PKW_Leaf_010_e09_b | Shugoshin | 0.49 | 594aa_24/85 (28%)_+3 | ni |
| $(TC)_{19}$ | Mbalbisiana_PKW_Root_008_g09_b | Fumarylacetoacetase (gi:121962541) | 0.66 | 427 aa _ 16/35 (45%) _ -3 | ni |
| $(TCG)_5$ | Mbalbisiana_PKW_Root_012_F01_b | No hits | | | ni |
| $(TTA)_5$ | Mbalbisiana_PKW_Root020_D05_b | No hits | | | ni |
| $(TCA)_7$ | Mbalbisiana_PKW_Leaf_007_d04_b | No Hits | | | ni |
| $(TCC)_8$ | Mbalbisiana_PKW_Leaf020_G01_b | Elongation factor 1-alpha (gi:6015058) | 2.00E-91 | 449 aa _ 164/178 (92%) _ +3 | 3´UTR |
| $(TGC)_8$ | Mbalbisiana_PKW_Leaf003_A12_b | No Hits | | | ni |
| $(TTAA)_4$ | Mbalbisiana_PKW_Root_018_H12_b | No hits | | | ni |
| $(TGGA)_5$ | Mbalbisiana_PKW_Root_011_D04_b | No Hits | | | ni |
| $(TGCC)_6$ | Mbalbisiana_PKW_Leaf_006_e04_b | Serine - glyoxylate aminotransferase (gi:90185106) | 4.00E-49 | 401 aa _ 97/131 (74%) _ +1 | 5´UTR |

Ni, not identified; ORF, Open Reading Frame; UTR, Untranslated Region

and 73.5% for the assembling of both libraries.

A list of the most populated contigs, from the three groups of contigs identified is presented in **Table 2**. As expected, the most populated contig – having 269 reads – was positive for Ribulose bisphosphate carboxylase small chain after submitting its consensus sequence to Blastx against the SwissProt database. As predicted, 268 reads in this contig – PKW_Leaf&Root_Contig37 – derived from leaves and only one from roots (**Table 2**).

Interestingly, none of the ten most populated contigs, which were composed of ESTs from both libraries, had a ratio between numbers of reads from these libraries close to one. In five of them (contig 37, 45, 106, 126, and 227), the number of reads from leaves was much higher than from roots; while in the others (contigs 31, 35, 185, 359, and 414), the number of reads from roots was much higher (**Table 2**).

The KOG analysis reveals a conserved core of largely essential eukaryotic genes related to physiological and cellular processes such as translation, ribosomal structure and biogenesis, and those related to post-translational modifications, protein turnover and chaperones. The categories of *Musa balbisiana* ESTs are shown in **Fig. 1**. The bars indicate the number of unigenes assigned to each category as a percentage of the total number of unigenes for the assembling samples. Comparison of datasets reveals significant differences in the representation of genes within the various functional categories. For root ESTs, the largest category was related to the translation, ribosomal structure and biogenesis. Secondly, the posttranslational modification, protein turnover and chaperone related genes were prominently expressed. The third was the group of intracellular trafficking, secretion and vesicular transport. For the leaf cDNA library, the largest category was the group of genes related to energy production and conversion. An important number of unigenes were grouped in the 'others' with no indication of their corresponding function.

The manual curation of previously annotated genes allowed the identification of several candidates to further characterization and new findings, consistent with previous transcriptome studies. Among the annotated genes the enzyme glycine decarboxylase, from the glycine cleavage sys-

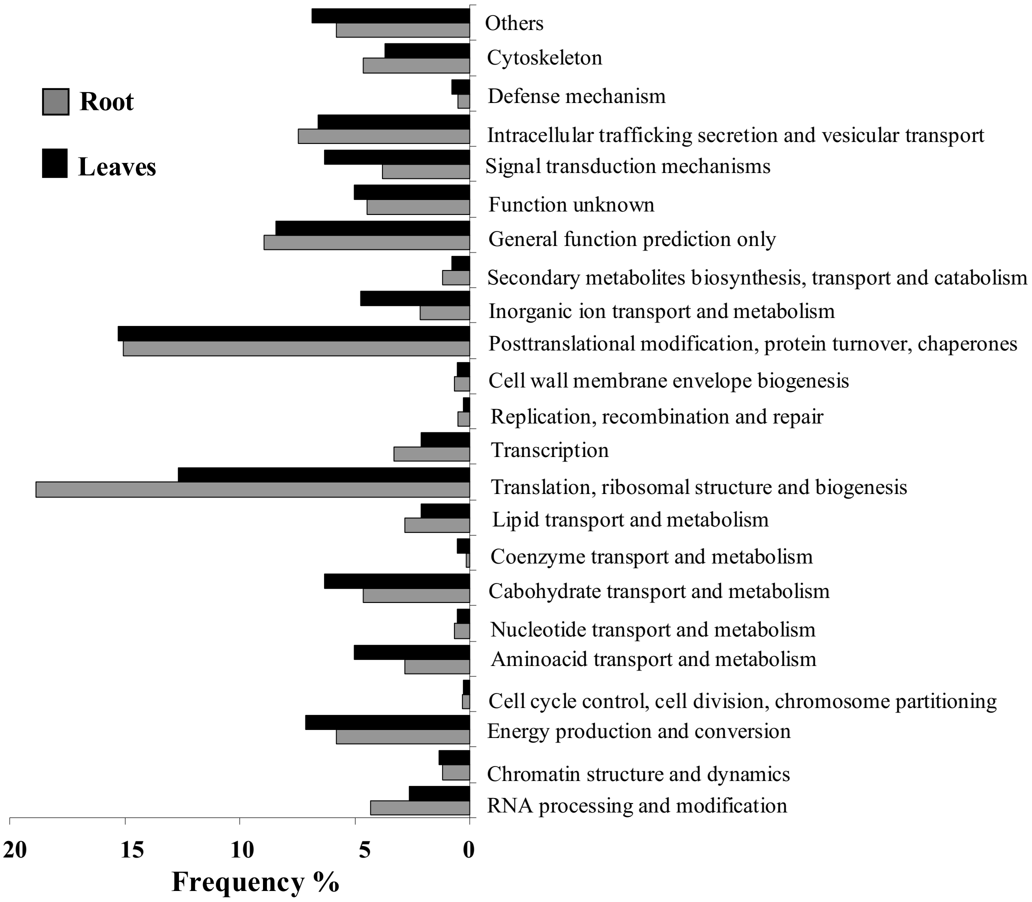

**Fig. 1 KOG categories of *Musa balbisiana* cDNA libraries.**

tem H, was revealed as an interesting candidate as it was previously implicated on acute sensitivity to drought stress in peas (Taylor *et al.* 2002). Another candidate was endo-xyloglucan transferase (EXGT), which is a class of glycosyltransferase that catalyzes transfer of a segment of xyloglucan molecule to another xyloglucan molecule, mediating molecular grafting between matrix polysaccharides in plant cell walls (Okazawa *et al.* 1993).

The fact that the sequence data is derived from a non-normalized cDNA library allows us to infer gene expression levels based on the number of reads contained in the contigs and, moreover, the comparison of expression in two closely related genomes. Therefore, digital methods based on the generation of sequence tags allow the significant expression level identification of candidate genes stored in computer databases. The Audic and Claverie (1997) method is a rigorous significance test that detects differentially expressed genes from relevant cDNA libraries.

The Digital Differential Display (DDD) applied for cDNA libraries from *Musa balbisiana* and *M. acuminata* through both statistical methods Audic and Claverie (1997) and Stekel (2000) revealed that almost totality of the expressed genes is common for both dataset, but it also revealed a few specificities. Among the common genes are the housekeeping genes, as well as the abundant ribulose-1,5-bisphosphate carboxylase/oxygenase and Rubisco activase. On the other hand, the comparison between multiple datasets, showed that Metallothionein – like proteins and annexins are among the highly expressed genes preferentially expressed in roots from *M. acuminata*. Annexins have been studied in relation to a large number of physiological responses. Plant annexins have been found to be a multi-functional gene family whose members play roles in the cellular response to gravity (Clark *et al.* 2006), diurnal cycles (Hoshino *et al.* 2004) and in imparting tolerance to various abiotic stress such as salt, drought, and high- and low-temperature conditions (Cantero *et al.* 2006). Plant annexins are ubiquitous, soluble proteins capable of $Ca^{2+}$-dependent and $Ca^{2+}$-independent binding to endomembranes and the plasma membrane. The *in vitro* properties of annexins and their known, dynamic distribution patterns, suggest that they play a key role as regulators or effectors of plant growth and stress signaling (Mortimer *et al.* 2008).

All 5.289 EST were analyzed for their potential use in developing SSR markers (EST-SSRs). EST-SSRs were present in about 1% of the total ESTs in the *M. balbisiana* var ‘Pisang Klutuk Wulung’ genome (**Table 3**). The (GA)n repeats was the most abundant class (44%), which is in agreement with Gupta *et al.* (2003), who reports these repeats as the most abundant in plants. For the trimeric repeats, the (CTT/GAA) motif was the most abundant (24%) and the relative abundance of di, tri and tetra nucleotide repeat motifs in the entire EST collection was 30, 56 and 14%, respectively. In order to validate and use these EST-SSRs derived markers, the next step is to develop primers from the flanking sequences, and evaluate the potential of these markers in genetic diversity studies, marker assisted breeding, map construction and their transferability within this crop species.

Analysis of the leaf and root transcriptome of *M. balbisiana* has proven to be a useful starting point for many other studies, both in terms of specific genes found to be expressed as well as in allowing a glimpse into which genes are related to plant defense. This work is a contribution to the Global *Musa* Genomics Consortium - GMGC (http://www.musagenomics.org/), and offers the first considerable amount of *M. balbisiana* EST sequences for further studies.

The *Musa* database provided is an interesting resource for recognizing genes previously described as involved in a wide range of physiological phenomena and as well as those that are involved in stress response in related species as *M. acuminata* and *M. balbisiana*. On the other hand, the transcriptome studied revealed a majority group of unknown genes in which their biological roles are yet to be determined.

## ACKNOWLEDGEMENTS

The authors would like to acknowledge the support from the National Council for Scientific and Technological Development (CNPq grant 506165/2004-3) and the Generation Challenge Program (Project Number SP2-15 / Commissioned Research 2005).

## 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 Research* **25**, 3389-3402

**Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K Dwight, SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G** (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25 (1)**, 25-29

**Asp T, Frei UK, Didion T, Nielsen KK, Lubberstedt T** (2007) Frequency, type and distribution of EST-SSRs from three genotypes of *Lolium perenne*, and their conservation across orthologous sequences of *Festuca arundinacea*, *Brachypodium distachyon* and *Oryza sativa*. *BMC Plant Biology* **7**, 36

**Audic S, Claverie JM** (1997) The significance of digital gene expression profiles. *Genome Research* **7**, 986-995

**Bedell JA, Korf I, Gish W** (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16 (11)**, 1010-1041

**Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL** (2008) GenBank. *Nucleic Acids Research* **36**, D25-D30

**Cantero A, Barthakur S, Bushart TJ, Chou S, Morgan RO, Fernández MP, Clark GB, Roux SJ** (2006) Expression profiling of the *Arabidopsis* annexin gene family during germination, de-etiolation and abiotic stress. *Plant Physiology and Biochemistry* **44**, 13-24

**Clark GB, Rafati DS, Bolton RJ, Dauwalder M, Roux SJ** (2000) Redistribution of annexin in gravistimulated pea plumules. *Plant Physiology and Biochemistry* **38**, 937-947

**Daniells J, Jenny C, Karamura D, Tomekpe K** (2001) *Musa*logue: a catalogue of *Musa* germplasm. In: Arnaud E, Sharrock S (Eds) *Diversity in the Genus Musa*, International Network for the Improvement of Banana and Plantain, Montpellier, France

**Food & Agriculture Organization – FAO** Available online: http://faostat.fao.org/

**Gasteiger E, Jung E, Bairoch A** (2001) SWISS-PROT: Connecting biological knowledge via a protein database. *Current Issues Molecular Bio*logy **3**, 47-55

**Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS** (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Molecular Genetics and Genomics* **270**, 315-323

**Hoshino D, Hayashi A, Temmei Y, Kanzawa N, Tsuchiya T** (2004) Biochemical and immunohistochemical characterization of *Mimosa* annexin. *Planta* **219**, 867-875

**Iniguez-Luy FL, Voort AV, Osborn TC** (2008) Development of a set of public SSR markers derived from genomic sequence of a rapid cycling *Brassica oleracea* L. genotype. *Theoretical and Applied Genetics* **117 (6)**, 977-985

**Kantety RV, Rota ML, Matthews DE, Sorrells ME** (2002) Data mining for simple sequence repeats in expressed sequence tags form barley, maize, rice, sorghum and wheat. *Plant Molecular Biology* **48**, 501-510

**Lysak M, Dolezelova M, Horry J, Swennen R, Dolezel J** (1999) Flow cytometric analysis of nuclear DNA content in *Musa*. *Theoretical and Applied Genetics* **98**, 1344-1350

**Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S** (2006) EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. *Nucleic Acids Research* **34**, W459-462

**Mortimer JC, Laohavisit A, Macpherson N, Webb A, Brownlee C, Battey NH, Davies JM** (2008) Annexins: multifunctional components of growth and adaptation. *Journal of Experimental Botany* **59 (3)**, 533-544

**Okazawa K, Sato Y, Nakagawa T, Asada K, Kato I, Tomita E, Nishitani K** (1993) Molecular cloning and cDNA sequencing of endoxyloglucan transferase, a novel class of glycosyltransferase that mediates molecular grafting between matrix polysaccharides in plant cell walls. *Journal of Biological Chemistry* **268**, 25364-25368

**Pappas GJ Jr., Miranda RP, Martins NF, Togawa RC, Costa MMC** (2008) SisGen: A CORBA-based data management program for DNA sequencing projects. *Lecture Notes in Computer Science* **5109**, 116-123

**Roux N, Baurens FC, Doležel J, Hribová E, Heslop-Harrison JS, Town C, Sasaki T, Matsumoto T, Aert R, Remy S, Souza M, Lagoda P** (2008) Genomics of banana and plantain (*Musa* spp.), Major staple crops in the tropics. In: Moore P, Ming R (Eds) *Genomics of Tropical Crop Plants. Plant Genetics and Genomics: Crops and Models* (Vol 1), Springer, New York, pp 83-111

**Santos CM, Martins NF, Hörberg HM, de Almeida ER, Coelho MC, Togawa RC, da Silva FR, Caetano AR, Miller RN, Souza Jr. MT** (2005) Analysis of expressed sequence tags from *Musa acuminata* ssp. *burmannicoides* var. Calcutta 4 (AA) leaves submitted to temperature stresses. *Theoretical and Applied Genet*ics **110 (8)**, 1517-1522

**Schoof H, Ernst R, Nazarov V, Pfeifer L, Mewes HW, Mayer KFX** (2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Research* **32**, D373-D376

**Senthilvel S, Jayashree B, Mahalakshmi V, Kummar PS, Nakka S, Nepolean T, Hash CT** (2008) Development and mapping of simple sequence repeat markers for pearl millet from data mining of expressed sequence tags. *BMC Plant Biology* **8**, 119

**Simmonds NW, Shepherd K** (1955) The taxonomy and origins of the cultivated bananas. *Journal of the Linnean Society of Botany (London)* **55**, 302-312

**Souza Junior MT, Santos CR, Martins NF, Silva FR, Togawa RC, Cassiano LAP, Almeida AER, Coelho MCF, Caetano AR, Ciampi AY, Costa MM, Piffanelli P, Miller RNG** (2005) Data*Musa* - Banco de dados de genômica de *Musa acuminata*. *Boletim de Pesquisa e Desenvolvimento Embrapa Recursos Genéticos e Biotecnologia* **107**, 1-24. Available online: www.cenargen.embrapa.br/publica/trabalhos/bp109.pdf

**Stekel DJ, Git Y, Falciani F** (2000) The comparison of gene expression from multiple cDNA libraries. *Genome Research* **10**, 2055-2061

**Tamana A, Khan AU** (2005) Mapping and analysis of simple sequence repeats in the *Arabidopsis thaliana* genome. *Bioinformation* **1 (2)**, 64-68

**Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA** (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41

**Taylor NL, Day DA, Millar AH** (2002) Environmental stress causes oxidative damage to plant mitochondria leading to inhibition of glycine decarboxylase. *Journal of Biological Chemistry* **277 (45)**, 42663-42668

**Thiel T, Michalek W, Varshney RK, Graner A** (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics* **106**, 411-422

**Varshney RK, Graner A, Sorrells ME** (2005) Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology* **23 (1)**, 49-56

**Yang C-J, Wang J, Mu L-Q, Li S-C, Liu G-J, Hu C-Q** (2008) Development of an EST-SSR marker in *Panax ginseng*. *Chinese Journal of Agricultural Biotechnology* **5 (2)**, 175-181

**Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H** (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79-92