Theor Appl Genet (2005) 111: 1118–1126
DOI 10.1007/s00122-005-0039-7

ORIGINAL PAPER

Bert Coemans · Hideo Matsumura · Ryohei Terauchi
Serge Remy · Rony Swennen · László Sági

SuperSAGE combined with PCR walking allows global gene expression profiling of banana (*Musa acuminata*), a non-model organism

Received: 21 March 2005 / Accepted: 4 July 2005 / Published online: 25 August 2005

Abstract Super-serial analysis of gene expression (SuperSAGE) was used to characterize, for the first time, the global gene expression pattern in banana (*Musa acuminata*). A total of 10,196 tags were generated from leaf tissue, representing 5,292 expressed genes. Forty-nine tags of the top 100 most abundantly expressed transcripts were annotated by homology to cDNA or EST sequences. Typically for leaf tissue, analysis of the transcript profiles showed that the majority of the abundant transcripts are involved in energy production, mainly photosynthesis. However, the most abundant tag was derived from a type 3 metallothionein transcript, which accounted for nearly 3% of total transcripts analysed. Furthermore, the 26-bp long SuperSAGE tags were applied in 3′-rapid amplification of cDNA ends (3′RACE) for the identification of unknown tags. In combination with thermal asymmetric interlaced PCR (TAIL-PCR), this allowed the recovery of a full gene sequence of a novel NADPH:protochlorophyllide oxidoreductase, the key enzyme in chlorophyll biosynthesis. SuperSAGE in conjunction with 3′RACE and TAIL-PCR will be a powerful tool for transcriptomics of non-model, but otherwise important organisms.

Abbreviations MT: metallothionein · RACE: rapid amplification of cDNA ends · SAGE: serial analysis of gene expression · TAIL-PCR: thermal asymmetric interlaced PCR

Communicated by J. S. (Pat) Heslop-Harrison

B. Coemans (✉) · S. Remy · R. Swennen · L. Sági
Laboratory of Tropical Crop Improvement,
Katholieke Universiteit Leuven, Kasteelpark Arenberg 13,
3001 Leuven, Belgium
E-mail: bert.coemans@agr.kuleuven.ac.be
Tel.: +32-16-329605
Fax: +32-16-321993

H. Matsumura · R. Terauchi
Iwate Biotechnology Research Center, Narita 22-174-4,
024-0003 Kitakami, Iwate, Japan

Introduction

Global gene expression profiling or transcriptome mapping techniques such as microarray analysis and serial analysis of gene expression (SAGE) have become powerful tools for the identification of novel candidate genes and the characterization of specific metabolic or regulatory pathways. SAGE allows an accurate, quantitative analysis of gene expression for thousands of genes at a time (Velculescu et al. 1995). Briefly, a short cDNA fragment or tag (13 bp) is extracted from a defined position in each transcript by a series of linker ligations and restriction digestions. Tags are then amplified once by PCR, concatenated, cloned in a vector and sequenced. These tags, in most cases, contain sufficient information for unambiguous annotation to EST sequences. The frequency of each tag directly reflects the abundance of the corresponding mRNA (Velculescu et al. 1995). However, given the fact that annotation of the tags depends largely on the availability of cDNA libraries or EST collections, the use of SAGE in plants has mainly been limited to model organisms such as Arabidopsis and rice (Matsumura et al. 1999, 2003a,b; Chakravarthy et al. 2003; Ekman et al. 2003; Gibbings et al. 2003; Jung et al. 2003; Lee and Lee 2003; Fizames et al. 2004). In addition, various efforts have been made to increase the tag length, thereby enhancing annotation frequency. These improvements include modified SAGE (18-bp tag, Ryo et al. 2000), LongSAGE (21-bp tag, Saha et al. 2002) and SuperSAGE (26-bp tag, Matsumura et al. 2003b).

In non-model organisms, i.e., with limited or no genomic DNA and cDNA/EST sequences available, classical SAGE yielding short tags would not be practical due to the very low chance and reliability of annotating the sequenced tags. However, non-model plant species possess numerous important traits not available for study in model plants, which emphasises the need for high-throughput transcript profiling generally applicable to all crop plants. These traits may

include different organs (e.g., fleshy fruits), special developmental processes (e.g., apomixis or parthenocarpy) and distinct quality traits such as flavour, nutrient or medicinal substances. Certain plant–pathogen interactions with serious economical consequences may also justify a large-scale functional analysis in the target organisms directly.

Bananas and plantains (*Musa* spp.) are typical of such potential target organisms. The most important fruit crop on earth (in terms of production and consumption), bananas provide a staple food for nearly 400 million people in developing countries in the tropics. Almost 90% of the production is used locally for cooking, baking and brewing as well as consumed fresh from a wide range of cultivars grown by subsistence farmers. A few characteristics that make banana interesting for genomics studies are vegetative propagation, parthenocarpy, and its well-characterised fruit physiology. Despite its agricultural importance and small genome size of only 600 Mbp (Lysák et al. 1999), this genus has not yet been investigated extensively at a genomic scale. Only recently, random BAC clones from a wild diploid banana were sequenced (Aert et al. 2004). Our goal, therefore, was to determine whether SAGE could be an efficient means for rapid gene discovery, if the obtained tags are long enough to allow unambiguous annotation and walking in the corresponding genes.

Here we report on the application of SuperSAGE to characterise the banana leaf transcriptome. Our analysis of 10,192 sequenced tags represented 5,292 expressed genes, of which approximately 80% occurred only once. In addition, unknown SuperSAGE tags were successfully annotated by applying 3′-rapid amplification of cDNA ends (3′RACE) alone or in combination with thermal asymmetric interlaced PCR (TAIL-PCR) for 5′ extension. This study clearly illustrates the power of SuperSAGE combined with 3′RACE and TAIL-PCR for transcript profiling and gene discovery in non-model organisms.

Materials and methods

Plant material and RNA isolation

Plant material was obtained from the international *Musa* germplasm collection of the INIBAP Transit Centre at the Katholieke Universiteit Leuven (Belgium). A single plant of the wild diploid banana, *Musa acuminata* "Tuu Gia" (accession No. ITC.610) was grown under greenhouse conditions until a ten-leaf stage. A sample was taken from the youngest unfurled leaf on February 11 at 2.30 p.m., immersed immediately in liquid nitrogen and stored at −80°C. Total RNA isolation was essentially done as described by Eggermont et al. (1996). Purity and quantity of the isolated total RNA was confirmed and determined by denaturing gel electrophoresis and spectrophotometric analysis, respectively.

Generation of SuperSAGE libraries and data analysis

SuperSAGE library construction was performed as described by Matsumura et al. (2003b). Briefly, 5 µg of mRNA was purified from total RNA with an mRNA Purification Kit (Amersham Biosciences, Piscataway, NJ, USA). The mRNA was subsequently used for double stranded cDNA synthesis (SuperScript Double-Stranded cDNA Synthesis Kit, Invitrogen, Carlsbad, CA, USA) using a biotinylated oligo(dT) primer, followed by digestion with *Nla*III (New England Biolabs, Beverly, USA). The 3′-end fragments of the cDNA were bound to streptavidin-coated magnetic beads (Promega, Madison, WI, USA) and purified. FITC-labeled linkers, harboring a recognition site for the *Eco*P15I endonuclease were ligated to the cDNA. SuperSAGE tags adjacent to the linkers were then released by *Eco*P15I digestion. This type III restriction enzyme recognises the asymmetric hexameric sequence 5′-CAGCAG-3′ and cleaves the DNA 27 bp downstream the recognition site leaving a 5′ overhang of 2 bp. Two pools of linker-tags were blunt-ended by filling-in with KOD DNA polymerase (from *Thermococcus kodakaraensis* strain KOD1, Toyobo, Osaka, Japan) and randomly ligated to each other. PCR with AmpliTaq Gold (Applied Biosystems, Foster City, USA) and biotinylated linker-specific primers were used to amplify the resulting ditags. Linker fragments were released by digestion with *Nla*III and separated from the ditags on polyacrylamide gels. Subsequently, traces of contaminating linker fragments were removed from gel-purified ditags by binding to streptavidin-coated magnetic beads (Promega). Size-fractionated concatemers (>500 bp) were cloned into the pGEM3Z plasmid (Promega) and electroporated into *Escherichia coli* DH10B (Invitrogen). Clones were sequenced with the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) and the RISA384 DNA autosequencer (Shimadzu, Kyoto, Japan). Extraction and counting of the SuperSAGE tags was performed by the SAGE2000 software kindly provided by Dr. K.W. Kinzler (Johns Hopkins University, Baltimore, MD, USA). The data discussed in this publication have been deposited in NCBIs Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo) and are accessible through GEO Series accession number GSE2811.

3′-Rapid amplification of cDNA ends PCR

Extension of SuperSAGE tags was conducted according to Van den Berg et al. (1999) with some modifications. Single-stranded cDNA was synthesised from total RNA by using an oligo(dT) primer with a M13 tail [5′ TAG TTG TAA AAC GAC GGC CAG $(T)_{18}$ 3′] and

Omniscript reverse transcriptase (Qiagen, Hilden, Germany). A 21-bp primer complementary to the M13 tail was used for PCR in combination with primers specific to the 26-bp tags.

Thermal asymmetric interlaced PCR

cDNA or genomic sequences were amplified by TAIL-PCR (Liu et al. 1995). Two or three gene-specific nested primers were designed for each of three consecutive TAIL-PCR walking steps based on sequences obtained from 3′RACE or the previous TAIL-PCR steps (Table 1). These primers were then applied in consecutive PCR reactions in combination with a 128-fold degenerate primer AD2 according to the thermal cycling program described by Liu et al. (1995). For the first TAIL-PCR walking step, these primers were 42S1 and 42T1 for the secondary and tertiary reactions, respectively; for the second TAIL-PCR step, the corresponding primers were 42P2, 42S2 and 42T2 for the primary, secondary, and tertiary reactions, respectively; and finally, for the third TAIL-PCR walking step, the primers 42P3, 42S3, and 42T3 were applied for the three semi-nested PCR reactions (Table 1).

Sequencing and sequence analysis

3′RACE and TAIL-PCR products were excised from agarose gels by using the QIAquick Gel Extraction Kit (Qiagen, Hilden, Germany) and cloned in the pCR4-TOPO® vector (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. Following custom sequencing (VIB Genetic Research Facility, Antwerp, Belgium), analysis was performed by BLASTN (Altschul et al. 1997) against a *Musa* 3′ EST database donated to the Global *Musa* Genomics Consortium by Syngenta and maintained at the MIPS (Munich Information Center for Protein Sequences, Germany), as well as all publicly available sequences at NCBI (National Center for Biotechnology Information, GenBank + RefSeq Nucleotides + EMBL + DDBJ + PDB). A sequence was considered to be homologous with a match in the database if the predicted *E*-value was below a threshold of e^{-25}.

Table 1 Primer sequences applied for TAIL-PCR

Primer name[a]	Primer sequence (5′ → 3′)
AD2	NGTCGASWGANAWGAA
42S1	TCAGCAGAGCATCAGCTTGT
42T1	AGCATCAGCTTGTTGCTGTA
42P2	ACAGGGGGATGTGTTCTCTG
42S2	TGTTGCAGATCTTGCTGTCC
42T2	TGGTGTTCCTTGTGATCGAG
42P3	TCCAGGTGCATGATGGAGTA
42S3	TTCTCGGCCTTGAGGAAGT
42T3	CTGTGGTCTGGGCTTTCACT

[a]P, S and T refer to primary, secondary and tertiary PCR reactions, respectively; whereas 1, 2 and 3 refer to first, second and third TAIL-PCR walking steps, respectively

Results

Generation of a banana SuperSAGE library

Leaf material collected from a wild diploid banana plant was used to generate a SuperSAGE library. A total of 10,196 SuperSAGE tags were analysed, which corresponded to 5,292 unique tags. Although the most abundant tag was represented more than 300 times, which is close to 3% of the total transcripts, the vast majority of transcripts occurred only once (Fig. 1). Figure 1 also demonstrates that the number of unique tags steeply decreased with increasing tag copy number. For example, 4,409 tags, which correspond to 83.3% of all unique tags, occurred once and constituted 43.2% of the transcripts present in the analysed tissue. In contrast, only 214 tags, representative of 2.1% of all unique tags, had a tag copy number ≥5. These 214 tags, however, represented 40.9% of the transcript pool analysed.

Annotation of SuperSAGE tags

The following criteria were used to confirm positive tag identification: (i) the tag sequences (26 bp) match an EST perfectly, (ii) the tags match the EST in 5′ to 3′ orientation, and (iii) the CATG (*Nla*III) site is the most 3′ terminal. Figure 2 shows that 30% of the 100 most abundant tags perfectly matched entries in the *Musa* 3′ EST database whereas for 51% no match could be found. The remaining 19% of the tags did not perfectly match according to the defined criteria, but were considered as likely to be annotated because only one mismatch or one CATG site further to the 3′ end occurred.

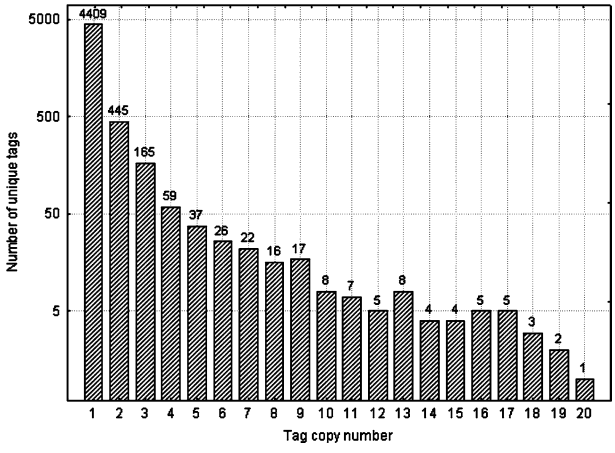


Fig. 1 Number of unique tags in relation to tag copy number. Only tags with a copy number ranging from 1 to 20 (5,248 tags or 99.17% of all unique tags) were plotted on the graph. A logarithmic *Y*-axis makes the presentation of data more clear

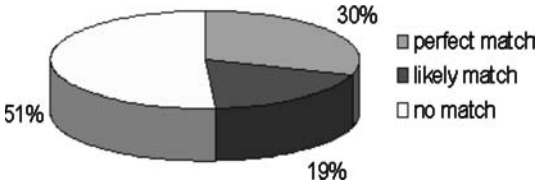


Fig. 2 Annotation of the 100 most abundant transcripts. Analysis was performed by BLASTN (Altschul et al. 1997) search to the *Musa* 3′ EST database. A tag is considered annotated if the following criteria are fulfilled: (1) the tag sequence (26 bp) matches an EST perfectly, (2) the tag matches the EST in 5′ to 3′ orientation, and (3) the CATG (*Nla*III) site is the most 3′ terminal. In case of one mismatch or one CATG site further to the 3′ end, the tag is considered as a "likely match"

Although tags were also searched against all publicly available sequences, this resulted in no increase of annotation frequency. From Figure 3, it can be seen that the annotation frequency increased from below 17% up to 43% with the tag copy number ranging from one to five, respectively.

A significant advantage of SuperSAGE compared to conventional SAGE is the fact that annotation becomes more species-specific due to the longer tag sequence (Matsumura et al. 2003b). As a demonstration for banana, 15 SuperSAGE tags were randomly selected and analysed by BLASTN against the *Musa* 3′ EST database as well as the entire GenBank database. To simulate different versions of SAGE, the length of these 15 tags was reduced from 26 bp (SuperSAGE) to 20 bp (LongSAGE), 18 bp (LongSAGE with blunting treatment) and finally, to 15 bp (conventional SAGE). It can be observed from Table 2 that the average, as well as the maximum number of species that possess accessions

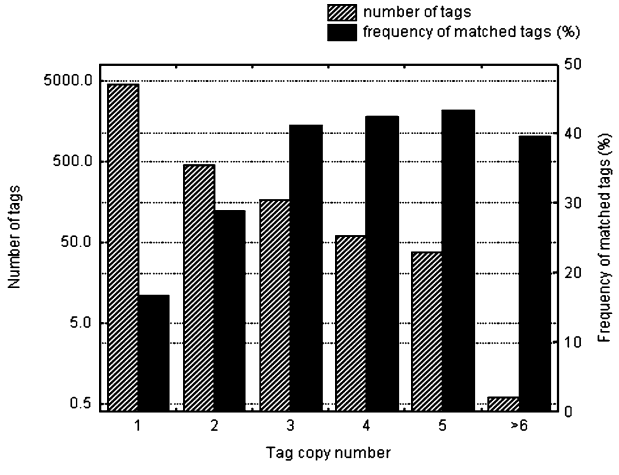


Fig. 3 Number of unique tags and their annotation frequency in relation to tag copy number. A total of 5,292 unique tags were plotted as a function of tag copy number. A logarithmic *Y*-axis makes the presentation of data more clear. Annotation frequencies were calculated as the percentage of tags with a perfect match to an accession in the *Musa* 3′ EST database. Data derived from tags with a copy number of six or higher were taken together as these tags comprised a small number of tags

Table 2 Summary of a simulated BLASTN search of 15 randomly chosen SuperSAGE tags to the *Musa* 3′ EST database and the GenBank

	Tag length (bp)			
	26	20	18	15
Average no. of species with hit to tag	1.1	1.2	1.8	8.7
Maximum no. of species with hit to tag	2	4	7	17

with a perfect match to the selected tags, significantly decreased with increasing tag length. In addition, tags extracted by SuperSAGE matched to an average of 1.1 species and maximum two species, clearly illustrating the power of SuperSAGE in gene identification.

Functional analysis of SuperSAGE tags

The 50 most abundant tags, together with their abundance and putative function are listed in Table 3. Assignment of a putative function to each tag was performed by BLASTN analysis of the matching *Musa* EST sequence to the NCBI database (Altschul et al. 1997). A type 3 metallothionein (MT) was by far the most abundant transcript and it accounted for nearly 3% of total transcripts analysed. Together with two other MT transcripts (nos. 13 and 18), they constituted more than 4% of total tags analysed. As can be seen in Table 3, the most abundant transcripts were involved in energy production, mainly photosynthesis, such as Rubisco, chlorophyll a/b-binding protein and oxygen-evolving enhancer protein. These transcripts accounted for 20% of the 100 most abundant transcripts and together with MTs they represented 32.2% of these transcripts (Fig. 4).

3′RACE for the identification of unknown tags

As banana is a non-model organism, a substantial number of SuperSAGE tags did not match entries in databases (Fig. 2). Therefore, we attempted using the 26-bp tags as primers for 3′RACE. This was successfully applied to five unknown tags (including nos. 10, 22 and 42 in Table 3) as well as five known tags (including nos. 47 and 48 in Table 3), all randomly selected. The recovered 3′RACE products represented partial cDNA sequences with an average length of 250 bp (data not shown). BLASTN (Altschul et al. 1997) analysis revealed significant homology ($E < e^{-25}$) to known plant genes for only two of the obtained 3′RACE products derived from the unknown tags (nos. 10 and 22, Table 3). One of the 3′RACE products without homology (Fig. 5a, tag no. 42) was further extended to the 5′ end by three consecutive TAIL-PCR walking steps, each consisting of two or three reactions (Fig. 5b, c). The obtained sequences were assembled and subjected to BLASTX (Altschul et al. 1997) revealing

Table 3 Tag sequence, abundance and putative function of the 50 most abundant transcripts

No.	Tag sequence (5′ → 3′)[a]	Abundance (%)	Putative function	*E* score[b]
1	ATGTCAGATGGGATAGGGTTGT	2.952	Metallothionein protein type 3	2e-18
2	TTCGGGTGCACCGACGCTGTGC	1.795	Ribulose-1,5-biphosphate carboxylase/oxygenase small subunit	Gene
3	GCCTCCGACCCCTCAAGCCGAC	1.403	No match	
4	CGGCAAACAGTGGCCGTCGGTG	1.324	No match	
5	GCGGCCGCAACCGCCGGTGGTG	1.304	No match	
6	GCGGCCGCGACCGCCGGTGGTG	1.255	No match	
7	GCGATCTCCTTGTGTGGGATTC	1.000	No match	
8	GCACCTGGGGCTGTTCCTTTGC	0.971	Putative 16 kDa membrane protein	3e-14
9	CTCATCTTCTACTTCGAGGCCG	0.902	Chlorophyll a/b-binding apoprotein CP24	5e-74
10	GCGGCTTCCGACGTCGAGTTCC	0.893	No match (RNA binding protein)[c]	(8e-36)[c]
11	ATTGCTGCGCGTGTTCGTTCTT	0.804	No match	
12	TAGACGATCGATGTTGCTTCCT	0.804	Chlorophyll a/b-binding protein precursor	3e-44
13	AGTGCTGGTGTTGGTCTCTGTG	0.677	Metallothionein protein type2	Gene
14	GGCTCGTGTCACCGGTGTAGCC	0.657	26S ribosomal RNA	3e-64
15	CAAGCTTGAGTATTCTATAGTG	0.539	No match	
16	TGCTCAGACCACCAGCCACTCC	0.539	Tonoplast intrinsic protein 1	2e-73
17	CTGCTGCGTACATCGTTAGAAG	0.520	No match	
18	ACGGCAAGTGCAAGTGCGGCGC	0.461	Metallothionein protein type 3	Gene
19	TACTAGTCTACTCTACTCAACA	0.441	Chlorophyll a/b-binding protein	1e-26
20	TGGTGTGGTCGTGGTGGTGATG	0.422	No match	
21	CCGTGGCCCGTCGCGGTGGTCG	0.412	Unknown protein	3e-23
22	GCGGACACCGGCTACGGCGTCG	0.392	No match (CONSTANS-like protein)[c]	(6e-27)[c]
23	TCCTGGCTGTTCGTCACTAACG	0.383	DNA-binding protein MNB1B (HMG1-like protein)	3e-37
24	TACTACTCTACTCTACTCAACA	0.373	Chlorophyll a/b-binding protein	1e-26
25	TAATTATAAGCAGCAATGGACG	0.353	No match	
26	GTTACCCCGACCTCCAAGAGCT	0.333	Oxygen-evolving enhancer protein	2e-69
27	TGTTGTGTTGTACTGTAATATT	0.284	Chlorophyll a/b-binding protein	9e-79
28	CAGACTGTGTGTACTTGCTGTG	0.275	No match	
29	ATGTCGTTGGATGATCCAATTA	0.275	No match	
30	GGGCTGGTGGTGCTCAAGCTCG	0.265	Blue copper-binding protein II	1e-19
31	TTAGGAGTGCTACCGGATGAAG	0.265	Translationally controlled tumor protein-like protein	4e-31
32	AGCGAGGAGGACCTACTCAACG	0.255	Plastocyanin precursor	2e-43
33	TTCCGTTCATCATTGCAAACAG	0.255	Plastoquinol–plastocyanin reductase	4e-74
34	AGCACCACGAGAAGAAGGATGC	0.235	ABA and stress-inducible protein (Asr1)	3e-42
35	AACGTATCTGTAAATTAATCTG	0.235	No match	
36	ATGCTTGCAGAAACTTGTGGTG	0.226	No match	
37	CCCGGCTTCGGCGGCGGGTGGG	0.226	No match	
38	AACGGCCTGGTGCTGGTCGGCG	0.226	Cytochrome b6f complex subunit	3e-13
39	TTAGCTTTGGCAAGACGCCGTT	0.226	No match	
40	TCTCTCTGTTCGGTGTTGTCGT	0.226	No match	
41	GCGGATCCCGGCTACGGCGTCG	0.226	No match	
42	GCGCTGTAAACAGAGCGTTGTT	0.216	No match (NADPH:protochlorophyllide oxidoreductase)[c]	(e-140)[c]
43	ATAAGGAGGCCATCCATCTCAT	0.206	No match	
44	TGTCTCACGGAGTTATCCATAT	0.206	No match	
45	GCTATGTAATCTGCATCTGCTG	0.206	(*S*)-2-hydroxy-acid oxidase	6e-57
46	ATGTGAGATGTCTTGTTGTATC	0.186	No match	
47	GGTCCCCCATCGTATCCGTGCG	0.186	Class III acidic chitinase	3e-76
48	GGCAACGACGACGGAAAGCACG	0.177	CCR4-associated factor-like protein	4e-57
49	GGGTTAGCCGCTTCTATGGGAT	0.177	No match	
50	TATAGTGGAGATTGAAATGTGC	0.177	Chlorophyll a/b-binding protein 4	1e-40

[a]Tags are presented as 22-bp sequences excluding the common *Nla*III site (CATG) at the 5′ end
[b]*E* score as calculated by BLASTN analysis (Altschul et al. 1997) of the *Musa* EST sequence to GenBank
[c]Extended by 3′RACE (and TAIL-PCR) and annotated in the GenBank

extensive similarity ($E < e^{-140}$) to NADPH:protochlorophyllide oxidoreductase genes of several dicot (*Cucumis sativus*, *Pisum sativum*, *Arabidopsis thaliana*) and monocot (*Oryza sativa*, *Hordeum vulgare*) species over the full length of the gene, indicating that the entire translated sequence was isolated. Furthermore, 68 bp upstream of the translation start codon a putative TATA box was identified in the full sequence (Fig. 5c). Comparison of the amplified banana genomic and cDNA sequences of the coding region revealed the presence of four introns. Figure 6 shows the exon–intron structure of NADPH:protochlorophyllide oxidoreductase genes in banana (GenBank accession AY862405, 395 aa), *A. thaliana* (GenBank accession AB013387, 405

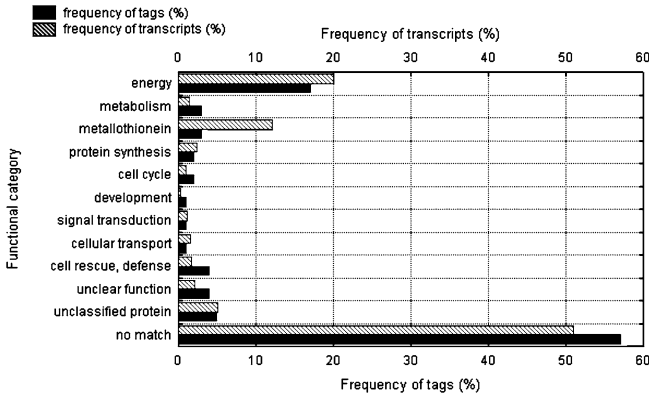


Fig. 4 Functional classification of the 100 most abundant transcripts. The frequency of transcripts was calculated by summing the tag abundance of each transcript for a given category to the total abundance of the 100 most abundant tags. The frequency of tags was calculated by counting the number of tags for a given category and expressed as a percentage of the total. Functional categories are based on the criteria of the MIPS (Munich Information Center for Protein Sequences, ftp://ftpmips.gsf.de/catalogue/funcat-2.0_scheme)

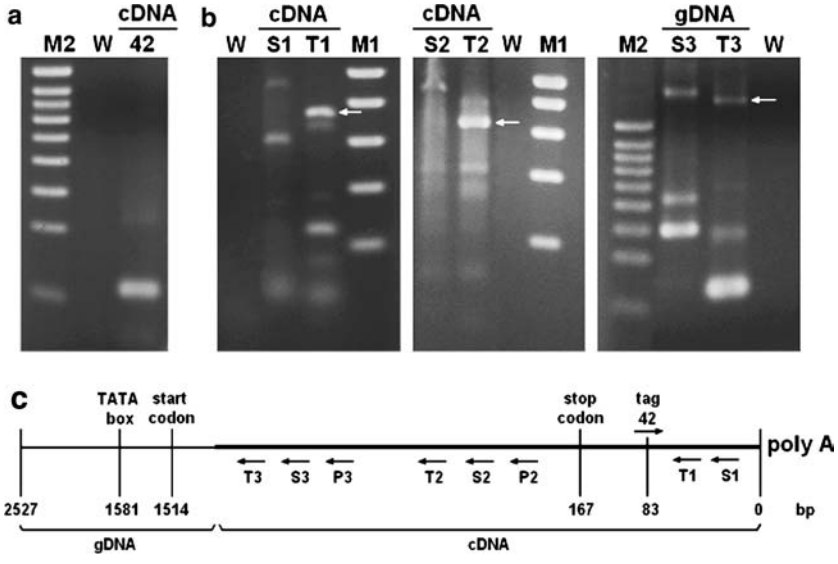
Fig. 5 **a** 3′RACE for the extension of tag 42 using a primer specific to the 26-bp tag. The size of the obtained PCR product was 120 bp. **b** Extension of the 3′RACE product to the 5′ end by three consecutive TAIL-PCR steps (*left*, *middle* and *right* agarose gels, respectively). Secondary (*S*) and tertiary (*T*) TAIL-PCR products are represented for each TAIL-PCR walking step (*1*, *2* and *3* for first, second and third walking steps, respectively). The first and second TAIL-PCR steps were performed on cDNA, whereas the third one was performed on genomic DNA. PCR products marked with arrows were excised from the gel and sequenced (730, 620 and 1,360 bp product for first, second and third TAIL-PCR steps, respectively). *W* refers to water control, whereas *M1* and *M2* are the molecular weight markers SmartLadder (200 bp) and SmartLadder SF (100 bp), respectively (Eurogentec, Seraing, Belgium). **c** Schematic overview of the full sequence amplified by 3′RACE and TAIL-PCR (not to scale). Positions of primers used for the 3′RACE and the three consecutive TAIL-PCR walking steps are indicated (*P*, *S* and *T* refer to primary, secondary and tertiary PCR reactions, respectively; whereas *1*, *2* and *3* refer to first, second and third walking steps, respectively)

aa) and *O. sativa* (Genbank accession AE017109, 402 aa). Although the total length of the exons is similar in these three species, the fourth intron is absent in *O. sativa* resulting in a fused exon at the C-terminus. Furthermore, each of the introns in *O. sativa* is much longer and the splicing positions are different when compared to the banana gene, the structure of the latter being more similar to *Arabidopsis* (Fig. 6).

Discussion

In this study, SuperSAGE was used to characterise the banana leaf transcriptome. To our knowledge, this is the first global analysis of gene expression in banana, which revealed 5,292 unique tags, i.e., 51.9% of the 10,196 tags sequenced in total. These findings are similar to a SuperSAGE study in rice leaves, where sequencing 12,119 tags identified 7,546 (62.3%) unique tags (Matsumura et al. 2003b). More than 80% of all tags sequenced occurred only once indicating that sequencing more concatemers could identify many more low abundant transcripts. Indeed, a SAGE study in yeast, a much less complex organism, revealed that the number of unique transcripts reached a plateau only at around 60,000 sequenced tags (Velculescu et al. 1997). Another explanation for the high proportion of unique single tags could be that these may have arisen as artefacts resulting from errors during PCR amplification. However, the latter should be minimized taking into account that a high-fidelity DNA polymerase was used for PCR amplification. The abundance profile observed in our SAGE library clearly shows that the number of unique tags steeply decreases with increasing tag copy number, which is in accordance with other SAGE studies in plants (Lorenz and Dean 2002; Gibbings et al. 2003; Lee and Lee 2003).

Identification of about 50% of the 100 most abundant tags was not possible, as the number of banana



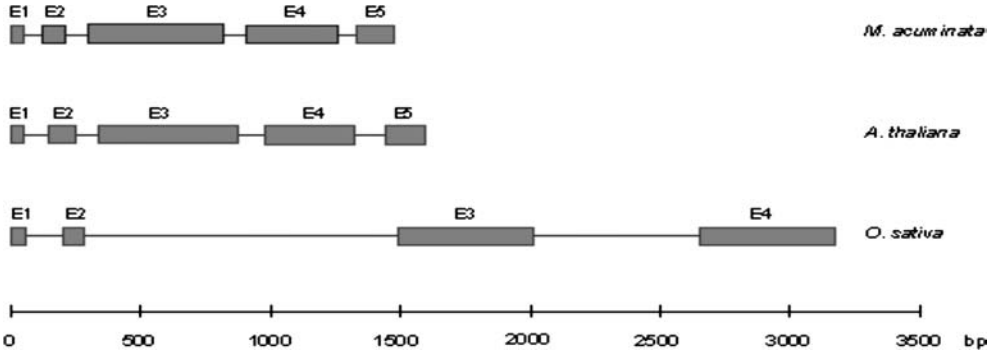


Fig. 6 Overview of the exon–intron structure of NADPH:protochlorophyllide oxidoreductase genes in *Musa acuminata* (accession AY862405, 395 aa), *Arabidopsis thaliana* (accession AB013387, 405 aa) and *Oryza sativa* (accession AE017109, 402 aa). Exons and introns are represented by *shaded rectangulars* and *lines*, respectively

cDNA/EST sequences available is still limited. The annotation frequency becomes even lower for low abundant tags as cDNA/EST libraries consist mainly of highly abundant transcripts.

As shown by BLASTN analysis on randomly chosen tags, SuperSAGE (26-bp tag) is much more species-specific in comparison with conventional SAGE as well as LongSAGE with blunting treatment (Table 2). Although the difference between SuperSAGE and LongSAGE (21-bp tag without blunting treatment) seems relatively small, the latter introduces bias, as blunting is required for random ligation of tags (Matsumura et al. 2003b). Thus, SuperSAGE provides significantly increased species-specificity compared to other SAGE protocols available at the moment.

A high proportion of the most abundant tags were derived from genes required for energy production, mainly photosynthesis, reflecting the characteristics of leaf tissue. Similar genes and comparable frequencies were also observed in rice (Gibbings et al. 2003) and *Arabidopsis* (Robinson et al. 2004). Nevertheless, the most abundantly expressed gene was a type 3 MT, accounting for nearly 3% of total transcripts analysed. A similarly high expression of MT genes was also observed in previous SAGE studies in rice seedlings (Matsumura et al. 1999) and leaves (Gibbings et al. 2003), which indicates that SuperSAGE does not result in altered transcript profiles. The exact function of plant MTs remains unclear, but the strength and diversity in MT gene responses and in sites of expression indicates fundamental importance and/or highly diverse functions. Indeed, experimental data accumulated so far suggest that plant MTs might be involved in various biological processes such as defence reaction to plant pathogens, apoptosis, development and heavy-metal metabolism (Cobbett and Goldsbrough 2002). Two other MT transcripts were detected among the top 20 of highest expressed genes (Table 3). Interestingly, MT transcript no. 13 codes for a type 2 MT indicating that SuperSAGE, due to its higher information content compared to conventional SAGE, can discriminate among different members of a gene family. On the other hand, transcript no. 18 was annotated to the same type 3 MT gene as the most abundant transcript denoting that two different tags might derive from one single gene. Since tag no. 18 was counted 47 times among the sequenced tags, it cannot be an artefact. This observation indicates that multiple transcripts could be generated from a single gene in banana by alternative splicing, as recently observed by SAGE in *Arabidopsis* (Robinson et al. 2004) and mammals (Pauws et al. 2001).

In addition to providing quantitative information on the abundance of expressed genes, a substantial advantage of SuperSAGE over other profiling techniques is that it can be used for the identification of unknown transcripts. For *Musa acuminata*, a large number of tags identified by SuperSAGE do not match entries in any database. However, identification of these tags is necessary to fully exploit the information gained by SAGE. Velculescu et al. (1995) proposed cDNA library screening with SAGE tags as hybridisation probes. This approach is labor-intensive and depends on multiple cDNA libraries, which may not be available in non-model species. Therefore, several authors have applied PCR techniques for the extension of unknown SAGE tags (Matsumura et al. 1999; Van den Berg et al. 1999; Chen et al. 2000). In most cases, however, it is difficult to obtain specific 3′RACE products, as the conventional SAGE tag is too short as a primer. The 26-bp SuperSAGE tags, however, enable to design more optimal and species-specific primers (as discussed above) for specific PCR amplification. Using 26-bp tags as primers, 3′RACE was successfully accomplished on five known and five unknown tags, identifying partial cDNA sequences with significant homology to known plant genes. Furthermore, TAIL-PCR was used for the 5′ extension of a 3′RACE product lacking homology to known sequences. This allowed the recovery of a full gene sequence with significant homology to a NADPH:protochlorophyllide oxidoreductase. The enzyme encoded by this gene, hitherto unknown in banana, catalyses the photoreduction of protochlorophyllide to chlorophyllide, the key regulatory step in chlorophyll biosynthesis (Raskin and Schwartz 2002). The identification of this abundant transcript confirms once again that SuperSAGE is likely to reflect faithfully the characteristic gene expression patterns in leaf tissue.

The absence of homology in some 3′RACE products can be ascribed to the fact that they are still located in

the 3′ untranslated region. As this region is less conserved, annotation remains a difficult task unless cDNA libraries and large numbers of ESTs are produced. One possible solution could be combining 3′SAGE with 5′SAGE described by Hashimoto et al. (2004) and Wei et al. (2004). Since 5′SAGE extracts tags at the 5′ end of each transcript, it is possible to generate tags at both gene boundaries. Recently, Ng et al. (2005) succeeded in physically linking the 5′ and 3′ tags for each transcript, making it feasible to apply these tags directly as PCR primers to amplify full-length transcripts on a large scale.

In conclusion, we have used SuperSAGE to perform a first transcriptome analysis of banana leaves. Furthermore, the 26-bp SuperSAGE tags were successfully applied as primers in 3′RACE, thus allowing the identification of unknown transcripts. This approach is especially promising for the annotation of differentially expressed tags, e.g., for the analysis of plant–pathogen interactions. Hence, SuperSAGE combined with 3'RACE and TAIL-PCR provides a powerful tool for functional genomics in non-model organisms for which supporting sequence resources are less extensive.

Acknowledgements The authors are indebted to Dr. S. Reich and Dr. D. Krüger (Humboldt University, Berlin, Germany) for providing the *Eco*P15I endonuclease and Dr. K.W. Kinzler (Johns Hopkins University, Baltimore, USA) for making the SAGE2000 software available. We are also grateful to Ms. I. Van den houwe (INIBAP Transit Center, Katholieke Universiteit Leuven, Belgium) for providing the plant material. Many thanks are due to Mr. I. Op De Beeck, Mr. C.O. Dimkpa and Ms. E. Thiry for technical assistance. Access to the Syngenta *Musa* 3′ EST database, donated by Syngenta to the International Network for the Improvement of Banana and Plantain (INIBAP) for use within the framework of the Global Musa Genomics Consortium is acknowledged. The research conducted at the Katholieke Universiteit Leuven was supported via an agreement with INIBAP by a grant of the Belgian Directorate-General for Development Cooperation (DGDC).

References

Aert R, Sági L, Volckaert G (2004) Gene content and density in banana (*Musa acuminata*) as revealed by genomic sequencing of BAC clones. Theor Appl Genet 109:129–139

Altschul SF, Madden TL, Schäffer 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

Chakravarthy S, Tuori RP, D'Ascenzo MD, Fobert PR, Després C, Martin GB (2003) The tomato transcription factor Pti4 regulates defense-related gene expression via GCC box and non-GCC box *cis* elements. Plant Cell 15:3033–3050

Chen JJ, Rowley JD, Wang SM (2000) Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. Proc Natl Acad Sci USA 97:349–353

Cobbett C, Goldsbrough P (2002) Phytochelatins and metallothioneins: roles in heavy metal detoxification and homeostasis. Annu Rev Plant Biol 53:159–182

Eggermont K, Goderis I, Broekaert W (1996) High-throughput RNA extraction from samples based on homogenization by reciprocal shaking in the presence of sand and glass beads. Plant Mol Biol Rep 14:273–279

Ekman DR, Lorenz WW, Przybyla AE, Wolfe NL, Dean JFD (2003) SAGE analysis of transcriptome responses in *Arabidopsis* roots exposed to 2,4,6-trinitrotoluene. Plant Physiol 133:1397–1406

Fizames C, Muños S, Cazettes C, Nacry P, Boucherez J, Gaymard F, Piquemal D, Delorme V, Commes T, Doumas P, Cooke R, Marti J, Sentenac H, Gojon A (2004) The *Arabidopsis* root transcriptome by serial analysis of gene expression. Gene identification using the genome sequence. Plant Physiol 134:67–80

Gibbings JG, Cook BP, Dufault MR, Madden SL, Khuri S, Turnbull CJ, Dunwell M (2003) Global transcript analysis of rice leaf and seed using SAGE technology. Plant Biotechnol J 1:271–285

Hashimoto SI, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, Sese J, Morishita S, Sugano S, Matsushima K (2004) 5′end SAGE for the analysis of transcriptional start sites. Nat Biotechnol 22:1146–1149

Jung S, Lee J, Lee D (2003) Use of SAGE technology to reveal changes in gene expression in *Arabidopsis* leaves undergoing cold stress. Plant Mol Biol 52:553–567

Lee JY, Lee DH (2003) Use of serial analysis of gene expression technology to reveal changes in gene expression in *Arabidopsis* pollen undergoing cold stress. Plant Physiol 132:517–529

Liu Y, Mitsukawa N, Oosumi T, Whittier RF (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. Plant J 8:457–463

Lorenz WW, Dean JFD (2002) SAGE profiling and demonstration of differential gene expression along the axial developmental gradient of lignifying xylem in loblolly pine (*Pinus taeda*). Tree Physiol 22:301–310

Lysák M, Dolezelova M, Horry JP, Swennen R, Dolezel J (1999) Flow cytometric analysis of nuclear DNA content in *Musa*. Theor Appl Genet 98:1344–1350

Matsumura H, Nirasawa S, Terauchi R (1999) Transcript profiling in rice (*Oryza sativa* L.) seedlings using serial analysis of gene expression (SAGE). Plant J 20:719–726

Matsumura H, Nirasawa S, Kiba A, Urasaki N, Saitoh H, Ito M, Kawai-Yamada M, Uchimiya H, Terauchi R (2003a) Overexpression of Bax inhibitor suppresses the fungal elicitor-induced cell death in rice (*Oryza sativa* L.) cells. Plant J 33:425–434

Matsumura H, Reich S, Ito A, Saitoh H, Kamoun S, Winter P, Kahl G, Reuter M, Krüger DH, Terauchi R (2003b) Gene expression analysis of plant host-pathogen interactions by SuperSAGE. Proc Natl Acad Sci USA 100:15718–15723

Ng P, Wei C, Sung W, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, Liu ET, Ruan Y (2005) Gene identification signature (GIS) analysis for transcriptome characterisation and genome annotation. Nat Methods 2:105–111

Pauws E, Van Kampen AHC, Van de Graaf SAR, De Vijlder JJM, Ris-Stalpers C (2001) Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. Nucleic Acids Res 29:1690–1694

Raskin VI, Schwartz A (2002) The charge-transfer complex between protochlorophyllide and NADPH: an intermediate in protochlorophyllide photoreduction. Photosynth Res 74:181–186

Robinson SJ, Cram DJ, Lewis CT, Parkin IAP (2004) Maximizing the efficacy of SAGE analysis identifies novel transcripts in *Arabidopsis*. Plant Physiol 136:3223–3233

Ryo A, Kondoh N, Wakatsuki T, Hada A, Yamamoto N, Yamamoto M, Yamamoto N (2000) A modified serial analysis of gene expression that generates longer sequence tags by non-palindromic cohesive linker ligation. Anal Biochem 277:160–162

Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE (2002) Using the transcriptome to annotate the genome. Nat Biotechnol 20:508–512

Spano AJ, He Z, Michel H, Hunt DF, Timko MP (1992) Molecular cloning, nuclear gene structure, and developmental expression of NADPH:protochlorophyllide oxidoreductase in pea (*Pisum sativum* L.). Plant Mol Biol 18:967–972

Van den Berg A, Van der Leij J, Poppema S (1999) Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags. Nucleic Acids Res 27:17

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Hieter P, Vogelstein B, Kinzler KW (1997) Characterization of the yeast transcriptome. Cell 88:243–251

Wei CL, Ng P, Chiu KP, Wong CH, Ang CC, Lipovich L, Liu ET, Ruan Y (2004) 5′ Long serial analysis of gene expression (LongSAGE) and 3′ longSAGE for transcriptome characterization and genome annotation. Proc Natl Acad Sci USA 101:11701–11706