Published online 29 September 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 19* ***e126***
doi:10.1093/nar/gkl522

Robust analysis of 5′-transcript ends (5′-RATE): a novel technique for transcriptome analysis and genome annotation

Malali Gowda, Haumeng Li, Joe Alessi[1], Feng Chen[1], Richard Pratt[2] and Guo-Liang Wang*

Department of Plant Pathology, The Ohio State University, Columbus, OH 43210, USA, [1]US DOE Joint Genome Institute, Walnut Creek, CA 94598, USA and [2]Department of Horticulture and Crop Science, Ohio Agricultural Research and Development Center, The Ohio State University, Wooster, OH 44691, USA

Received May 5, 2006; Revised June 21, 2006; Accepted July 7, 2006

ABSTRACT

Complicated cloning procedures and the high cost of sequencing have inhibited the wide application of serial analysis of gene expression and massively parallel signature sequencing for genome-wide transcriptome profiling of complex genomes. Here we describe a new method called robust analysis of 5′-transcript ends (5′-RATE) for rapid and cost-effective isolation of long 5′ transcript ends (~80 bp). It consists of three major steps including 5′-oligocapping of mRNA, NlaIII tag and ditag generation, and pyrosequencing of NlaIII tags. Complicated steps, such as purification and cloning of concatemers, colony picking and plasmid DNA purification, are eliminated and the conventional Sanger sequencing method is replaced with the newly developed pyrosequencing method. Sequence analysis of a maize 5′-RATE library revealed complex alternative transcription start sites and a 5′ poly(A) tail in maize transcripts. Our results demonstrate that 5′-RATE is a simple, fast and cost-effective method for transcriptome analysis and genome annotation of complex genomes.

INTRODUCTION

Rapid sequencing of many complex eukaryotic genomes has provided unprecedented opportunities to understand gene function, genome structure and genome evolution. However, accurate annotation of all expressed genes in the sequenced genomes remains one of the most challenging tasks for genome biologists. Although various computer-based gene prediction methods play a role in genome annotation, experimental data provide essential evidence for the determination of gene structure and function. In the last decade, various sequence-based strategies, such as expressed sequence tags (ESTs) (1), full-length (FL) cDNA (2,3), serial analysis of gene expression (SAGE) (4,5) and massively parallel signature sequencing (MPSS), have been developed for transcriptome studies (6,7). These approaches have contributed valuable resources for gene discovery and genome annotation, but their application in most molecular studies has been limited.

Generally, EST and FL-cDNA sequencing techniques are neither cost-effective nor deep enough to isolate rare transcripts or address transcript variability. Sequencing millions of cDNA clones from various tissues can only sample ~60% of the expressed genes (8). To overcome this limitation, high-throughput and short tag-based approaches such as SAGE (4) and MPSS (6) have been developed. SAGE library construction involves several tedious steps before tags can be cloned into a plasmid vector. The process includes isolation of short tags (14–26 bp) from the 3′ or 5′ ends of transcripts, ditag formation, concatenation and sequencing of SAGE clones. The time-consuming procedure of colony picking and storage, and the high cost for sequencing individual clones in SAGE library construction has prohibited use of this approach in many biological studies (5,9). The MPSS strategy involves *in vitro* cloning of cDNA molecules on the surface of microbeads and non-gel-based sequencing of millions of tags (17–20 bp) (6). MPSS library construction can be performed only by experienced technicians at Solexa, Inc. The multiple-location matching of some 17–21 bp tags from SAGE or MPSS libraries in a sequenced genome is problematic when mapping tags to the EST or genomic sequence. To obtain accurate matches for interested tags in the genome, longer transcripts have to be isolated. This is usually accomplished using techniques such as rapid amplification of cDNA ends (RACE) (10) or generation of longer cDNA fragments using the GLGI method (11,12). These individual gene confirmation assays are tedious and expensive, and they are not practical when many positive tags have been identified.

The Sanger method of DNA sequencing is expensive and laborious (13,14). Currently, several strategies and

*To whom correspondence should be addressed. Tel: +1 614 292 9280; Fax: +1 614 292 4455; Email: wang.620@osu.edu

platforms are under development including sequencing by synthesis (SBS), sequencing by hybridization and nanopore sequencing (14). Pyrosequencing is an SBS method that can sequence thousands of DNA fragments in a few hours. The entire genome of a bacterium was sequenced in 4.5 h with high accuracy (13), compared with the several months required by the Sanger procedure (15). The pyrosequencing technique generates high-quality short sequences (~100 bases), and it has many potentially important applications when combined with tag-based expression profiling methods (14).

In this study, we describe a novel approach called robust analysis of 5′-transcript ends (5′-RATE). The method includes three major steps: 5′-oligocapping of mRNA using the FL-cDNA isolation strategy (16,17) NlaIII tag and ditag formation using the RL-SAGE strategy (5), and tag sequencing by pyrosequencing (13). The 5′-RATE method has simplified transcript tag isolation by eliminating the complicated concatemer cloning procedure. This allows for a quick, efficient and cost-effective method for the identification and characterization of the 5′ signatures of expressed genes. This strategy is flexible because it can also be adapted easily for 3′ end isolation of expressed genes.

We applied the 5′-RATE method to characterize expressed genes from the maize inbred line B73, which is being used for whole genome sequencing. Maize, which has a genome ~80% the size of the human genome, is one of the most important crops, a model for plant genetics, breeding and crop evolution (18). Sequencing of the gene rich region has predicted that the maize genome consists of a large number of genes (59 000 genes) as compared to mammalian genomes (18). Although hundred and thousand of ESTs and full-length cDNA (http://www.maizecdna.org/) have been released to the public, only a limited number of 5′ end signatures have been identified. In this study, we developed the 5′-RATE method using total RNA isolated from B73. Sequence analysis of the 5′-RATE tags revealed the complex nature of alternative transcription start sites (TSSs) and promoter regions. Interestingly, the 5′-RATE method is comprehensive enough to identify poly(A) tails at the 5′ regions of many maize transcripts, which has not been detected so far in any organisms using other expression approaches. These results indicate that 5′-RATE is a powerful profiling method for rapid identification of long transcript ends in complex genomes.

MATERIALS AND METHODS

RNA isolation

Total RNA was isolated from ~2.0 g of leaves of 30-day-old maize plants (inbred line B73) using a Trizol solution (Invitrogen, Carlsbad, CA). The mRNA was purified using a Qiagen (Valencia, CA) kit according to the manufacturer's instructions.

Oligo-capping at the 5′ regions of mRNA

About 1.0 μg poly(A^+) mRNA was used for 5′-decapping. RNA oligocapping was carried out as described by Suzuki *et al*. (16,17) and Hashimoto *et al*. (19) with minor modifications (see details at http://www.mtir.org/protocols/5p_rate_protocol.pdf). Bacterial alkaline phosphatase was used to remove 5′ phosphate groups from mRNAs, while, subsequently, 5′G-capping was hydrolyzed using tobacco acid pyrophosphatase (http://www.mtir.org/protocols/5p_rate_protocol.pdf). The decapped mRNA was divided into two pools (pools 1 and 2) and ligated with two different synthetic RNA oligos (5′-oligo A and 5-oligo B) (Figure 1; http://www.mtir.org/protocols/5p_rate_protocol.pdf) using T_4 RNA ligase (TaKaRa, New York, NY).

First-strand cDNA synthesis

The decapped mRNA was pre-heated at 70°C for 10 min to prepare for single-strand cDNA synthesis. Pre-heated mRNA

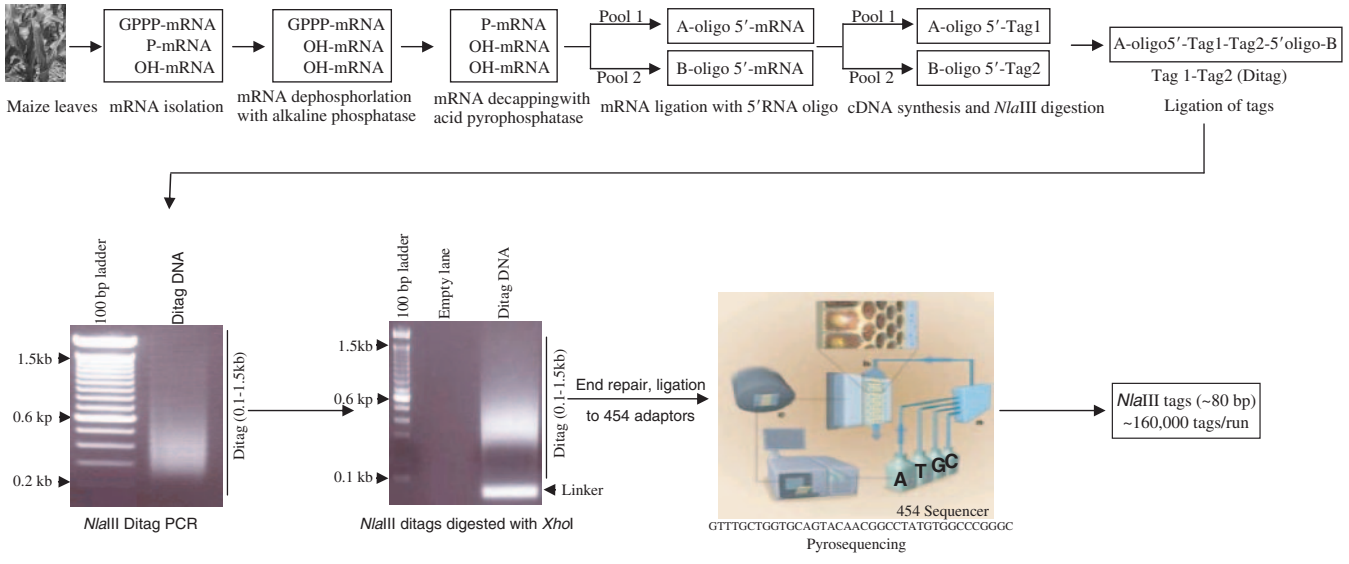

Figure 1. Experimental procedure for the 5′-RATE. The mRNA from maize is treated with bacterial alkaline phosphatase and acid pyrophosphatase to modify the cap structure at the 5′ regions. The 5′ decapped mRNA is divided into pools 1 and 2 and ligated with RNA oligos (**A** and **B**). The cDNA is synthesized and tags are released from the 5′ regions of cDNA using the NlaIII enzyme. Tags from the two pools are self-ligated to generate ditag cassettes. Ditags are amplified using PCR and linkers are removed by XhoI digestion. Ditag fragments are sequenced using the 454 pyrosequencer at DOE Joint Genome Institute (JGI), CA.

was then combined with 10 pmol of random adapter primer and 3.5 μl (200 U/μl) of reverse transcriptase (Superscript; Invitrogen) in 100 μl volume. The RT reaction was incubated at 12°C for 1 h, followed by 42°C for 4 h, according to the procedure described by Hashimoto *et al*. (19). The mRNA was hydrolyzed using 15 μl of 0.1 M NaOH at 65°C for 40 min. The cDNA synthesis was confirmed using actin and ubiquitin primers (see details at http://www.mtir.org/protocols/5p_rate_protocol.pdf).

Double-strand cDNA amplification

The single stranded cDNA (10 μl) was amplified in a 50 μl PCR using 10 pmol of 3′ primer specific to random primer sequences and 10 pmol of biotinylated 5′ primer (primer A for pool 1 and primer B for pool 2) (see details at http://www.mtir.org/protocols/5p_rate_protocol.pdf). About 5 U of *Pfu*Turbo® DNA polymerase (Stratagene Inc., La Jolla, CA) was used for each PCR. A total of 12 PCR cycles were performed at 94°C, 1 min; 58°C, 1 min; 72°C, 1 min with a final extension of 5 min at 72°C.

NlaIII tag and ditag formation

Amplified cDNA was digested with 200 U of NlaIII enzyme for 3 h at 37°C (see details at http://www.mtir.org/protocols/5p_rate_protocol.pdf). Biotinylated PCR fragments were captured using 100 μl of Dynal streptavidin beads. Ditag cassettes were formed by ligating NlaIII tags from pools 1 and 2 with 15 U of T_4 DNA ligase (USB Inc., Cleveland, OH) in 25 μl volume at 16°C overnight. Ditags (1 μl of 1:100 dilution) were amplified in five 50 μl PCRs (22 cycles of 94°C, 5 min; 94°C, 1 min; 60°C, 30 s; 72°C, 1 min; and 72°C, 5 min extension) with 10 pmol of biotinylated forward and reverse primers (see details at http://www.mtir.org/protocols/5p_rate_protocol.pdf) and 5 U of platinum *Taq* DNA polymerase (Invitrogen). Ditag DNA fragments (50 bp to 1.5 kb) were purified from a 3% agarose gel (Figure 1) and linkers corresponding to pools 1 and 2 were removed by digesting with 200 U of XhoI in 300 μl volume for 3 h at 37°C. Ditags were then purified from a 3% agarose gel and dissolved in 50 μl of ultra pure water. The linkers and any undigested ditags were removed using 100 μl of Dynal streptavidin beads. The supernatant was treated twice with phenol:chloroform:isoamyl alcohol (24:24:1), precipitated and dissolved in 10 μl of ultra pure water.

Pyrosequencing of NlaIII tags

The NlaIII ditags were made as blunt ends and were ligated with pyrosequencing adaptors (http://www.mtir.org/protocols/5p_rate_protocol.pdf). Single-stranded ditags were captured on beads and subjected to emulsion PCR (emPCR) to enrich the templates. The enriched beads were loaded on a pico-titer-plate for pyrosequencing according to Margulies *et al*. (13) (see details at http://www.mtir.org/protocols/5p_rate_protocol.pdf). Pyrosequencing was carried out on a pico-titer-plate at the DOE Joint Genome Institute (JGI), CA (http://www.jgi.doe.gov). A limited amount of a nucleotide (A or G or C or T) was added at a time to pause the DNA polymerase reaction. During this process, a pyrophosphate (PPi) was released from each nucleotide incorporation, which in turn was converted into ATP by sulfurylase. The resulting ATP was further catalyzed by luciferase to emit light. The emitted light was detected by a CCD camera and then converted into pyrogram (13) (http://www.454.com/). The signal peaks in the pyrogram were converted into nucleotide sequencing information.

5′-RATE tag extraction

We developed the RATEspy program to extract the NlaIII tags from the 454 raw sequences. For forward tag extraction, the 5′ oligo signature sequence (TCGAGT) was identified. Then a tag was extracted after the signature sequence and before the first NlaIII site (CATG). If no CATG site was found, a tag was extracted until an 'N' was found or up to 80 bp after the signature sequence. For reverse tag extraction, the raw NlaIII sequences were reverse complemented and then the tags were extracted using the same method for the forward tags. Forward and reverse tags were clustered separately to get unique tags using the RATEspy program.

Mapping NlaIII tags

Stand lone local BLAST 2.0 was used to map 5′-RATE tags to the target databases including maize genomic (ftp://ftp.tigr.org/pub/data/MAIZE/gene_enrichment_reads), and maize FL-cDNA sequences (ftp://ftp.genome.arizona.edu/pub/est/maize/seq_dir) separately. The 5′ end sequences of maize FL-cDNAs were used to determine the matching rate of the NlaIII tags. An identity of 90% and an *E*-value = e−5 were used as BLAST search criteria. The BLAST results were processed and analyzed using RATEspy to get matching statistical reports.

Putative promoter identification

About 200 bp of genomic DNA upstream from TSSs were extracted in order to predict maize promoter regions as described by Shahmuradov *et al*. (20). Promoter motifs such as the TATA box and other *cis*-acting elements were predicted using a PlantProm DB program (http://mendel.cs.rhul.ac.uk and http://www.softberry.com) (20).

RESULTS

Improvements in the isolation of 5′ ends and generation of NlaIII tags

The three major steps (5′-oligocapping, formation of NlaIII tags and ditags, and pyrosequencing) involved in 5′-RATE library construction are presented in Figure 1. About 1.0 μg of mRNA was used for the 5′-RATE protocol, as compared to 5.0–25.0 μg mRNA in SAGE (19,21) and FL-cDNA (16,17) library construction protocols. To improve the ligation efficiency, two distinct RNA oligos were ligated overnight to 5′ regions of mRNA pools (1 and 2) as compared to 3–16 h in the original procedures (19). Digestion of synthesized cDNA with NlaIII released much longer 5′ tags (average 250 bp) than those released from the type IIS (22) or type III enzyme (23) digestions during SAGE library construction (Table 1). Ditags (average 500 bp) were generated by ligating tags overnight from the two pools, similar to the ditag ligation in RL-SAGE (5). Only five ditag PCRs were enough to generate a 5′-RATE library as compared to

Table 1. Comparison of 5′-RATE with SAGE and MPSS

Feature	5′-RATE	LongSAGE	SuperSAGE	MPSS
Tagging enzyme	NlaIII (Type II[a])	MmeI (Type IIS[b])	EcoP15I (Type III[c])	BsmFI/MmeI (Type IIS)
Binding sequences	CATG	TCCRAC	CAGCAG	GGGAC/TCCRAC
Cleavage	On the binding site	Away from binding site	Away from the binding site	Away from the binding site
Tag size (bp)	∼80	19–21	25–26	17–20
Method of sequencing	Pyrosequecing	Sanger method	Sanger method	Hybridization
Cloning and colony picking	Not required	Required	Required	Required
Standard kits	Lab made	I-SAGE kit	SAGE kit	Custom library in Solexa, Inc.
Technical difficulties	Simple	Challenging	Challenging	Challenging
Cost/library ($)	Inexpensive (∼9000)	Expensive (∼30 000)	Expensive (∼30 000)	Expensive (∼30 000)
Time requirement	10–15 days	Several months	Several months	Several months

[a]Restriction enzyme consisting of a homodimer that recognition a palindromic sequences and cleave within the recognition site. Only Mg^{2+} is required as a cofactor in this case.
[b]Restriction enzymes consist of monomer which recognize non-palindromic sites and cleave outside the recognition sequence. SAM (*S*-adenosylmethionine and Mg^{2+} are required cofactors for successful cleavage.
[c]Type III restriction enzymes consist of restriction and methylation subunits. Recognition sites are non-palindromic and cleavage is ∼25 bases from the recognition site. ATP and Mg^{2+} are required cofactors for successful cleavage.

20 (5,9) to 1000 PCRs (Catalog no. T5001-01; I-SAGE kit from Invitrogen) in other SAGE methods. The longer RATE ditags (∼500 bp) were easily purified on 2% agarose gel as compared to PAGE for the purification of shorter SAGE tags (5). Purified ditags (∼5.0 μg) were precipitated and shipped to JGI for 454 pyrosequencing. The complete 5′-RATE experimental procedure is available at http://www.mtir.org/protocols/5p_rate_protocol.pdf.

Generation and characterization of a maize 5′-RATE library

About 160 000 sequence reads were obtained from a 454 sequence run. Using the RATEspy program, we isolated over 116 000 NlaIII tags with good quality sequences (Table 2). The size of sequenced 5′ NlaIII tags varied from 21 to 150 bp with an average of 80 bp (Table 2 and Figure 2). The RATE procedure improved tag size ∼3- to 5-fold in comparison with SAGE or MPSS procedures (tags size, 14–26 bp; Table 1). To validate the 5′-RATE method, 3259 significant tags (⩾2 copies or more) were matched against the maize genome sequence (ftp://ftp.tigr.org/pub/data/MAIZE/gene_enrichment_reads) and the 5′ regions of maize FL-cDNAs (http://www.maizecdna.org). Among these tags, 44% matched to the 5′ regions of FL-cDNAs and 34% matched to the maize genome sequence at 95% identity (Table 2) which is lower than that of the 3′ RL-SAGE tags (data not shown). The low matching rate is likely because of unfinished genome sequencing, incomplete sampling of FL-cDNAs, and heterogeneity of the TSSs that was observed in our study (see below) and other 5′ LongSAGE libraries (19,21). As expected, >70% of the tags matched to the 5′ region (within 100 bp) of the maize FL-cDNAs, which is similar to the results in other 5′ LongSAGE studies (19,21). Matching analysis with different lengths of 5′-RATE tags showed that as the tag length was increased, the rate of multiple hits to non-redundant nucleotide database at NCBI was decreased (Supplementary Table 1), which was also demonstrated by Matsumura *et al*. (23,24). For example, only the 60 bp tag has a unique match in the NCBI non-redundant nucleotide database for the homolog of the rice acireductone dioxygenase 2 gene (Supplementary Table 1).

Table 2. Features of the maize B73 5′-RATE library

Inbred line	B73
Treatment	None
Growth stage	4-week-old leaves
Total mRNA	1 μg
Template DNA sequenced	Ditag
No. of reads sequenced	160 000
Total cost/library ($)	9000
Average tag size	∼80 bp
Matching of significant tags to genomic DNA	34%
Matching of significant tags to 5′ regions of maize FL-cDNA	44%

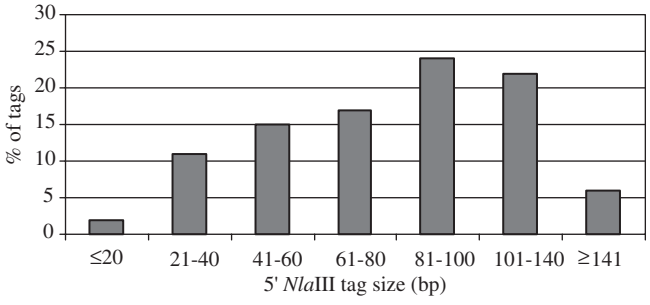

Figure 2. Size distribution of the 5′-RATE tags.

Sequence diversity in the 5′ region of maize transcripts

Preliminary sequence analysis of the 5′-RATE tags revealed that many maize transcripts had alternative TSSs. For example, the gene encoding a jasmonate-induced protein (ID: Q564C9) had 46 different TSSs (Figure 3) and the rubisco small subunit-encoding gene (ID: P05348) had 9 different TSSs (Supplementary Figure 3). In general, the TSS location of different transcripts varied from 1 to 99 nt from the 5′ region of maize FL-cDNAs. The length of the TSSs ranged from 8 to 14 nt (Figure 3 and Supplementary Figures 1–3). The analysis of the TSS data did not reveal any consensus sequences. Among the analyzed transcripts, the rubisco small subunit-encoding gene had the lowest 5′-tag diversity

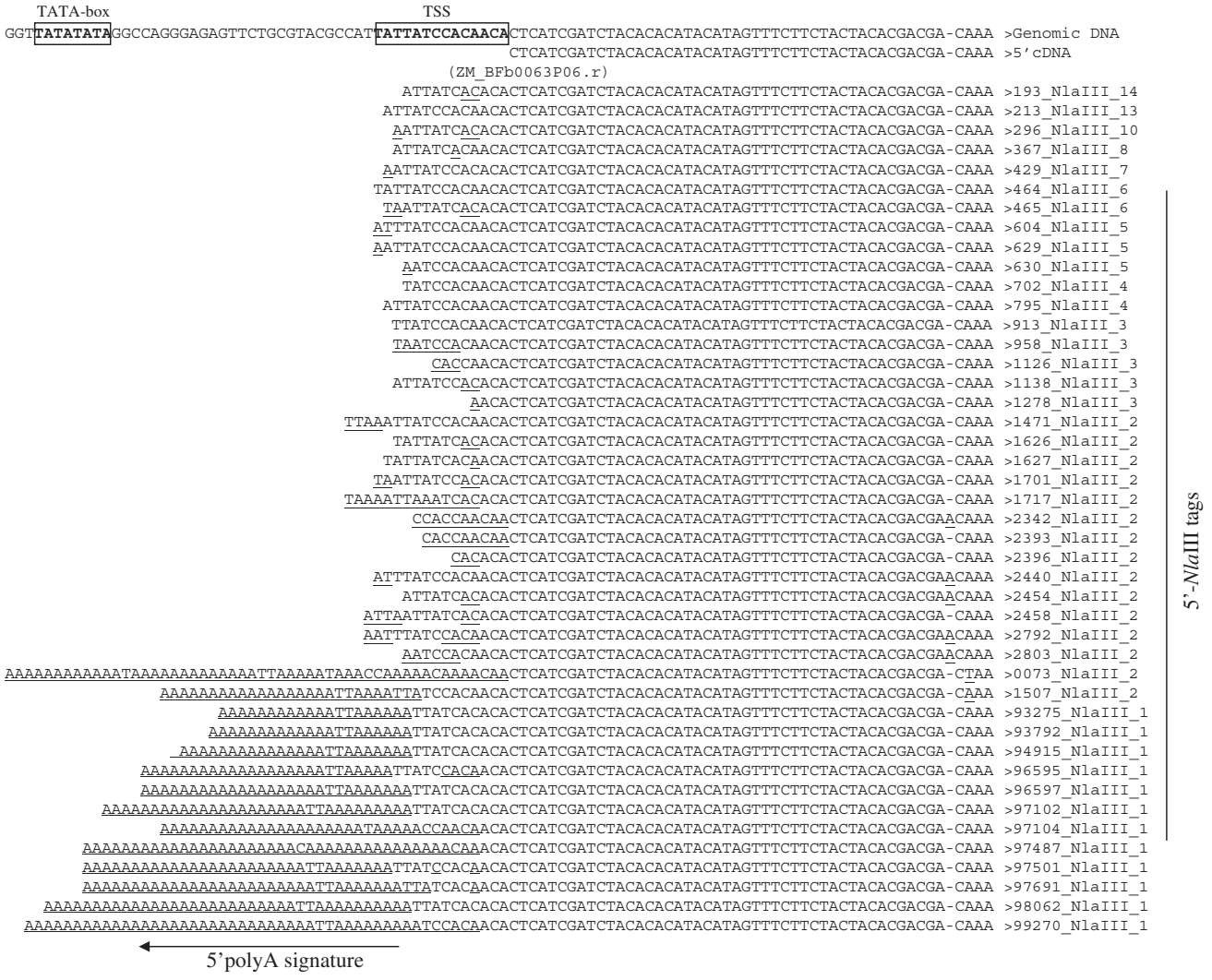

Figure 3. Sequence alignment of the jasmonate-induced gene (ID: Q564C9) with its alternative TSSs and 5′ poly(A) tail. The nucleotides in underlined are non-template sequences.

(33.33%) (Supplementary Figure 3). The transcript with the highest tag diversity (90%) was the gene encoding an intermediate filament C2 protein (ID: Q9NG13) with 2–75 non-template derived nucleotides (Supplementary Figure 2). Alternative TSSs were found for a lot of genes analyzed in this study, which is similar with the findings in the FL-cDNA libraries in *Arabidopsis* (25).

When the 5′-RATE tags matched to the maize genomic sequences, TATA- boxes were found at 30–40 bp upstream from the TSSs (Figure 3 and Supplementary Figures 1 and 3), while similar results were also reported in other plants (20,25) and animals (17–19). This result demonstrated that the majority of the 5′-RATE tags may be true 5′ sequences from the TSSs and the isolation of these sequences will facilitate the identification of their putative promoters.

Surprisingly, we found poly(A) tails at the 5′ ends of maize transcripts (Figure 3 and Supplementary Figures 1 and 2). Over 8% of the maize transcripts consisted in both G-capping and poly(A) signature in the 5′ regions. The length of the 5′ poly(A) tail varied from 20 to 150 adenine residues in most transcripts (Figure 3 and Supplementary Figures 1 and 2). Except for the rubisco small subunit gene (Supplementary Figure 3), we found poly(A) tails for other highly expressed genes (Figure 3 and Supplementary Figures 1 and 2). To investigate if any FL-cDNAs had 5′ poly(A) tails, we searched plant, animal and viral cDNA databases. Similarly, several FL-cDNAs with 5′ poly(A) tail were found in plants (maize, rice and *Arabidopsis)*, animals (human, mouse and *Drosophila*) and viruses (vaccine and cowpox virus) (Figure 4). Interestingly, the translation initiation codon (ATG) was followed immediately after the 5′ poly(A) tail in the FL-cDNAs from *Arabidopsis*, rice, human, mouse, *Drosophila* and virus, but this feature was not observed in maize FL-cDNAs (Figure 4).

DISCUSSION

The mRNA sequence provides crucial information for localizing expressed genes in a sequenced genome including maize, which has a relatively large genome size. Although millions of dollars have been invested in the identification of FL-cDNAs and ESTs in many organisms, these sequences

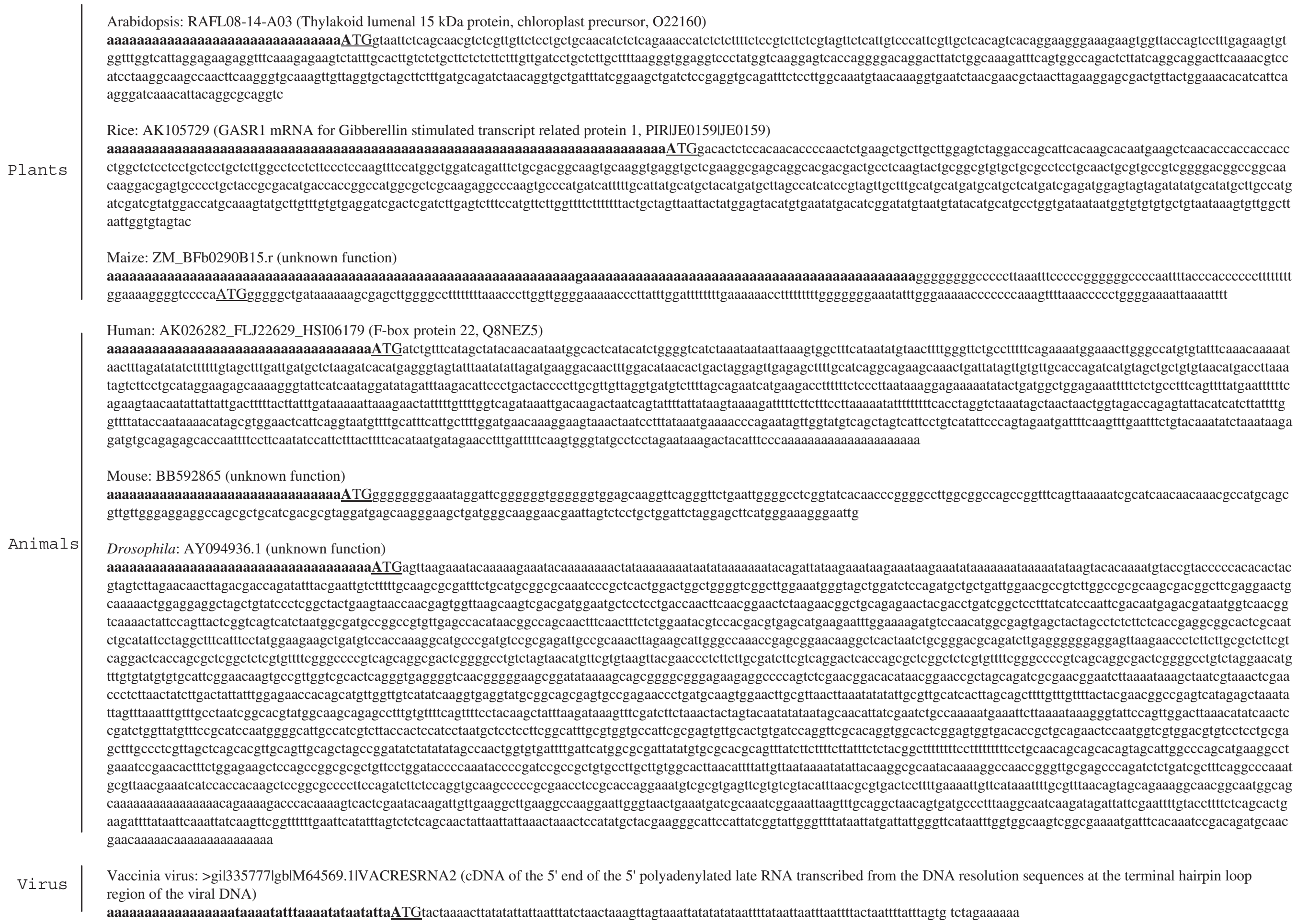

Figure 4. FL-cDNA sequences with 5′ poly(A) tail from plants and animals. The 5′ poly(A) sequences are shown in boldface letters and translation initiation codon (ATG) is shown in capital letters.

are estimated to cover a maximum of only 60% of the transcriptome (8). The SAGE technology has provided an unprecedented high-throughput and high-efficiency approach to identify uncharacterized transcripts. For example, the Cancer Genome Anatomy Project (CGAP) has adopted the SAGE method for the analysis of different cancerous cell types that has produced >7 million transcript tags from 171 libraries (26). Many of the tags are novel transcripts since they are not present in any FL-cDNA or EST collections (11). However, the current SAGE methods require complicated tag, ditag and concatemer cloning, tedious colony picking and expensive clone sequencing. Similarly, the MPSS library construction and sequencing procedures are more complicated and can only be performed by Solexa Inc. Owing to these limitations, SAGE and MPSS methods have not been widely used for transcriptome analysis. The 5′-RATE procedure reported here is simple and fast, and has eliminated tag cloning and colony picking procedures. One can now make a 5′-RATE library within 2 weeks. In contrast, several months are required for SAGE and MPSS library construction. At the current sequencing cost, a 5′-RATE library with 160 000 tags costs about $9000, which is two to three times cheaper than that of the SAGE and MPSS methods. A unique feature of pyrosequencing is that several small samples can be run on a single 454 chip. This provides the possibility to perform replications and multiplexing of multiple samples from the same or different organisms in a single experiment. If only 40 000 tag sequences are needed for a 5′-RATE library, then the cost will be less than $2500. The 5′-RATE method is highly flexible and can, therefore, be adopted to characterize tags from the 3′ end of the transcripts. A comparison between 5′-RATE with other tag-based methods is summarized in Table 1.

Two unique features of the 5′-RATE method are noteworthy. First, most of the tags in a 5′-RATE library are derived from the 5′ end sequence of transcripts. Since the majority of the SAGE or MPSS tags isolated so far are from the 3′ region of transcripts, identification of the TSS sequence in the 5′ ends is essential for the characterization of complete transcription units. Only few methods for 5′ end isolation have been reported so far, including CAGE (27), 5′LongSAGE (19,21) and PET (28) methods. For example, among 15 448 5′ tags identified in humans, 86–96% of the 5′ LongSAGE tags were assigned within −500 to +200 nt of the mRNA start sites (21). In the maize 5′-RATE library, >70% of the tags matched to 5′ regions (within 100 bp) of the

maize FL-cDNAs. As more FL-cDNA sequences from the maize full cDNA project (http://www.maizecdna.org) become available to the public, we expect a higher matching rate will be obtained from our 5′-RATE tags. The second unique feature of the 5′-RATE method is that the tag length ranges from 21 to 150 bp with an average of 80 bp. This will circumvent the multi-location matching of some of the SAGE (14 bp), LongSAGE tags (21 bp), SuperSAGE (26 bp) and MPSS tags (17–21 bp). Highly homologous gene family members should be more easily distinguished by using the long 5′-RATE tags rather than SAGE/LongSAGE/SuperSAGE or MPSS tags (Supplementary Table 1).

Limited work has been done towards the identification of TSSs in plants as compared to animals. The recent study by Alexandrov *et al*. (25) identified alternative TSSs in 30–50% of genes in the *Arabidopsis* using FL-cDNAs. Similarly, the 5′-RATE method described here has demonstrated that many maize genes produced alternative TSSs. Interestingly, substitutions, deletions and additions were also identified in the maize TSS regions, similar to the results observed in animal genomes (19,21). We also demonstrated that the promoter signature-like TATA boxes are localized at 30–40 bp upstream of the TSS region in maize, which is consistent with other plants (20) and animals (17,19). These results suggest that the 5′-RATE sequence data are an excellent genomics resource for the identification of TSSs, promoter regions and 5′-untranslated regions. The addition of these sequence data should ultimately increase the accuracy of genome annotation.

Newly generated mRNA transcripts, called heterogeneous nuclear RNAs, are further modified by the addition of 5′ cap structures (guanosine nucleotide via 5′–5′ triphosphate triphosphate linkage) and 3′ poly(A) tails (150–200 adenines) in eukaryotes (29). Unexpectedly, sequences obtained using the 5′-RATE method were revealed poly(A) tails (20–150 bp) at the 5′ ends of maize transcripts, which has not been reported previously in any other organism. The size of poly(A) tails at the 5′ end identified from this study is similar to that of 3′ poly(A) tail. The longer 5′ adenylation might increase the half-life of transcripts and may also regulate the translation and stability, which was reported for the 3′ poly(A) tail (30,31). These results motivated us to analyze further FL-cDNA sequences obtained from the biotinylated CAP trapper procedure in rice (3), *Arabidopsis* (2), mouse (32) and also from the oligocapping procedure in maize (http://www.maizecdna.org), human (33) and *Drosophila* (34). Surprisingly, several FL-cDNAs with 5′ poly(A) tails from plants (maize, rice and *Arabidopsis*) and animals (human, mouse and *Drosophila*) were unknowingly deposited in the databases. These results supported that 5′ G-capping and 5′ poly(A) tail structures in the transcripts are widely present in plants and animals. The chance that the 5′poly(A) poly(A) tails are experimental artifacts of the oligocapping method is low because the cDNA clones with a 5′ poly(A) tail in *Arabidopsis* (2) and mouse (32) FL-cDNA*s* were identified by the biotinylation CAP trapper method.

So far only one gene, called late gene or 11 kDa protein (M64569) in poxvirus (vaccine and cowpox), was reported to contain 5′ poly(A) sequences that are not complementary to the viral DNA template (35–42). Until now, there are no reports either on the 5′ poly(A) tail identification or the mechanism of 5′ polyadenylylation in any eukaryotic organisms. However, the recent *in vitro* experiment showed that the poly(A) tail addition to the 5′ regions of mRNA enhances the translation rate (43). They also reported that translation inhibition is possible with an excess of mRNA containing poly(A) tails (43). Our results confirmed the presence of non-template encoded poly(A) sequences at the 5′ regions of mRNAs in maize. The poly(A) tags identified by 5′-RATE method were also G-capped at the first nucleotide as similar to the poxviral late mRNAs (40). We believe that G-capping and poly(A) addition to the 5′ ends might be coupled to each other during mRNA processing. Interestingly, a novel translation initiation codon (ATG) was created due to the presence of 5′ poly(A) tails in the eukaryotic transcripts, as shown in Figure 4. We speculate that 5′ polyadenylation of transcripts might generate a novel protein diversity in eukaryotes. Although the possibility that the poly(A) tails are the artifacts of oligo-capping method is low, we will experimentally confirm the presence of the 5′ poly(A) tails in selected transcripts, and determine how a poly(A) tail is added to the 5′ region and its role in transcript stability and function in the near future.

It is worthwhile to report here that the presence of the 5′ poly(A) tail in maize transcripts has caused a major problem in our 5′ LongSAGE library construction (M. Gowda and G.L. Wang, unpublished data). Initially, we optimized the 5′ LongSAGE method (19) using the same RNA from maize that was used for the 5′-RATE method. Owing to the occurrence of long poly(A) signatures at the 5′ regions of maize transcripts, we failed to obtain enough concatemer clones for sequencing. The homopolymeric tracks of A/T in the plasmid might inhibit the replication and gene expression processes as shown in *Escherichia coli* (44,45). Similar problems have also been reported during cDNA library generation (46). In addition, sequencing of 5′ LongSAGE clones with poly(A) sequences was not successful (M. Gowda and G.L. Wang, unpublished data). Therefore, it is impossible to generate long concatemer inserts and obtain good sequencing results from a maize 5′ LongSAGE library.

In summary, 5′-RATE has the following advantages over existing tag-based methods: (i) it is simple because the difficult steps for purifying and cloning concatemers in *E.coli* are eliminated, allowing the technique to be used in most molecular labs with no specialized equipment, (ii) it is fast because colony picking and DNA purification for Sanger sequencing are eliminated, and a 454 sequencing run can be finished in a few hours; (iii) it is more comprehensive because 5′-RATE tags are more informative for genome and EST matching due to the generation of longer tag length (average 80 bp) compared to 21 bp RL-SAGE/LongSAGE tags or 17–21 bp MPSS tags; (iv) it is cost effective as it costs about $9000 for 160 000 5′-RATE tags in comparison to about $30 000 for LongSAGE tags; (v) the 5′-RATE tags will also have potential applications in subsequent biological experiments. For example, the 5′-RATE tag sequences can be used as templates for RNAi-based gene silencing, for probe designing of oligo-chips, or primer designing for RT–PCR assays. The 5′-RATE method could be further improved with the following two approaches. First, average tag size can be increased to >100 bp if other novel sequencing methods are used in ditag sequencing (13,14). Second,

a small fraction of transcripts might be missed during 5′-RATE library construction due to the absence of the NlaIII site on the transcripts. This can be overcome by making an additional 5′-RATE library using different tagging enzymes such as DpnII, Taq1, MseI or Sau3AI.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by Ohio Agricultural Research and Development Center (OARDC) Research Enhancement Grant Program and the Plant Genome Research Program of the National Science Foundation (#0321437). We appreciate the sequencing support from the DOE's Joint Genome Institute. Funding to pay the Open Access publication charges for this article was provided by the NSF-PGRP grant.

Conflict of interest statement. None declared.

REFERENCES

1. Adams,M.D., Kerlavage,A.R., Fleischmann,R.D., Fuldner,R.A., Bult,C.J., Lee,N.H., Kirkness,E.F., Weinstock,K.G., Gocayne,J.D., White,O. *et al*. (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, **377**, 3–174.
2. Seki,M., Narusaka,M., Kamiya,A., Ishida,J., Satou,M., Sakurai,T., Nakajima,M., Enju,A., Akiyama,K., Oono,Y. *et al*. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science*, **296**, 141–145.
3. Kikuchi,S., Satoh,K., Nagata,T., Kawagashira,N., Doi,K., Kishimoto,N., Yazaki,J., Ishikawa,M., Yamada,H., Ooka,H. *et al*. (2003) Collection, mapping, and annotation of 28 000 full-length cDNA clones from *Japonica* rice. *Science*, **301**, 376–379.
4. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
5. Gowda,M., Jantasuriyarat,C., Dean,R.A. and Wang,G.L. (2004) Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. *Plant Physiol*., **134**, 890–897.
6. Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M. *et al*. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol*., **18**, 630–634.
7. Meyers,B.C., Galbraith,D.W., Nelson,T. and Agrawal,V. (2004) Methods for transcriptional profiling in plants. Be fruitful and replicate. *Plant Physiol*., **135**, 637–652.
8. Sun,M., Zhou,G., Lee,S., Chen,J., Shi,R.Z. and Wang,S.M. (2004) SAGE is far more sensitive than EST for detecting low-abundance transcripts. *BMC Genomics*, **5**, 1.
9. Gowda,M. and Wang,G.L. (2005) Robust-LongSAGE (RL-SAGE): an improved LongSAGE method for high-throughput transcriptome analysis. In Nielsen,KL (ed.), *Methods in Molecular Biology: Serial analysis of Gene Expression*. Humana Press, NJ, USA, (in press).
10. Wang,A., Pierce,A., Judson-Kremer,K., Gaddis,S., Aldaz,C.M., Johnson,D.G. and MacLeod,M.C. (1999) Rapid analysis of gene expression (RAGE) facilitates universal expression profiling. *Nucleic Acids Res*., **27**, 4609–4618.
11. Chen,J., Sun,M., Lee,S., Zhou,G., Rowley,J.D. and Wang,S.M. (2002) Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl Acad. Sci. USA*, **99**, 12257–12262.
12. Silva,A.P.M., Chen,J., Carraro,D.M., Wang,S.M. and Camargo,A.A. (2004) Generation of longer 3′ cDNA fragments from massive parallel signature sequencing tags. *Nucleic Acids Res*., **32**, e94.
13. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
14. Metzker,M.L. (2005) Emerging technologies in DNA sequencing. *Genome Res*., **15**, 1767–1776.
15. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al*. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.
16. Suzuki,Y., Yoshitomo-Nakagawa,K., Maruyama,K., Suyama,A. and Sugano,S. (1997) Construction and characterization of a full length-enriched and a 5′-end-enriched cDNA library. *Gene*, **200**, 149-156.
17. Suzuki,Y., Taira,H., Tsunoda,T., Mizushima-Sugano,J., Sese,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Morishita,S. *et al*. (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. *EMBO Rep*., **2**, 388–393.
18. Messing,J., Bharti,A.K., Karlowski,W.M., Gundlach,H., Kim,H.R., Yu,Y., Wei,F., Fuks,G., Soderlund,C.A., Mayer,K.F. *et al*. (2004) Sequence composition and genome organization of maize. *Proc. Natl Acad. Sci. USA*, **101**, 4349–4354.
19. Hashimoto,S., Suzuki,Y., Kasai,Y., Morohoshi,K., Yamada,T., Sese,J., Morishita,S., Sugano,S. and Matsushima,K. (2004) 5′-End SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol*., **22**, 1146–1149.
20. Shahmuradov,I.A., Gammerman,A.J., Hancock,J.M., Bramley,P.M. and Solovyev,V.V. (2003) PlantProm: a database of plant promoter sequences. *Nucleic Acids Res*., **31**, 114–117.
21. Wei,C.L., Ng,P., Chiu,K.P., Wong,C.H., Ang,C.C., Lipovich,L., Liu,E.T. and 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.
22. Saha,S., Sparks,A.B., Rago,C., Akmaev,V., Wang,C.J., Vogelstein,B., Kinzler,K.W. and Velculescu,V.E. (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol*., **20**, 508–512.
23. Matsumura,H., Reich,S., Ito,A., Saitoh,H., Kamoun,S., Winter,P., Kahl,G., Reuter,M., Kruger,D.H. and Terauchi,R. (2003) Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc. Natl Acad. Sci. USA*, **100**, 15718–15723.
24. Matsumura,H., Ito,A., Saitoh,H., Winter,P., Kahl,G., Reuter,M., Kruger,D.H. and Terauchi,R. (2005) SuperSAGE. *Cell Microbiol*., **7**, 11–18.
25. Alexandrov,N.N., Troukhan,M.E., Brover,V.V., Tatarinova,T., Flavell,R.B. and Feldmann,K.A. (2006) Features of Arabidopsis genes and genome discovered using full-length cDNAs. *Plant Mol. Biol*., **60**, 69–85.
26. Pylouster,J., Senamaud-Beaufort,C. and Saison-Behmoaras,T.E. (2005) WEBSAGE: a web tool for visual analysis of differentially expressed human SAGE tags. *Nucleic Acids Res*., **33**, W693–W695.
27. Shiraki,T., Kondo,S., Katayama,S., Waki,K., Kasukawa,T., Kawaji,H., Kodzius,R., Watahiki,A., Nakamura,M., Arakawa,T. *et al*. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA*, **100**, 15776–15781.
28. Ng,P., Wei,C.L., Sung,W.K., Chiu,K.P., Lipovich,L., Ang,C.C., Gupta,S., Shahab,A., Ridwan,A., Wong,C.H. *et al*. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods*, **2**, 105–111.
29. Brown,C.E. and Sachs,A.B. (1998) Poly(A) tail length control in *Saccharomyces cerevisiae* occurs by message-specific deadenylation. *Mol. Cell. Biol*., **18**, 6548–6559.
30. Jacobson,A. (1996) Poly(A) metabolism and translation: the closed-loop model. In Hershey,J.W.B., Mathews,M.B. and Sonenberg,N. (eds), *Translational Control*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, Vol 30, pp. 451–480.
31. Jacobson,A. and Peltz,S.W. (1996) Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. *Annu. Rev. Biochem*., **65**, 693–739.
32. Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. *et al*. (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
33. Ota,T., Suzuki,Y., Nishikawa,T., Otsuki,T., Sugiyama,T., Irie,R., Wakamatsu,A., Hayashi,K., Sato,H., Nagai,K. *et al*. (2004) Complete

sequencing and characterization of 21 243 full-length human cDNAs. *Nature Genet*., **36**, 40–45.
34. Rubin,G.M., Hong,L., Brokstein,P., Evans-Holm,M., Frise,E., Stapleton,M. and Harvey,D.A. (2000) A *Drosophila* complementary DNA resource. *Science*, **287**, 2222–2224.
35. Bertholet,C., Van Meir,E., ten Heggeler-Bordier,B. and Wittek,R. (1987) Vaccinia virus produces late mRNAs by discontinuous synthesis. *Cell*, **50**, 153–162.
36. Schwer,B., Visca,P., Vos,J.C. and Stunnenberg,H.G. (1987) Discontinuous transcription or RNA processing of vaccinia virus late messengers results in a 5′ poly(A) leader. *Cell*, **50**, 163–169.
37. Patel,D.D. and Pickup,D.J. (1987) Messenger RNAs of a strongly-expressed late gene of cowpox virus contains a 5′-terminal poly(A) leader. *EMBO J*., **6**, 3787–3794.
38. Wright,C.F. and Moss,B. (1987) *In vitro* synthesis of vaccinia virus late mRNA containing a 5′ poly(A) leader sequence. *Proc. Natl Acad. Sci. USA*, **84**, 8883–8887.
39. Schwer,B. and Stunnenberg,H.G. (1988) Vaccinia virus late transcripts generated *in vitro* have a poly(A) head. *EMBO J*., **7**, 1183–1190.
40. Ahn,B.Y. and Moss,B. (1989) Capped poly(A) leader of variable lengths at the 5′ ends of vaccinia virus late mRNAs. *J. Virol*., **63**, 226–232.
41. Ahn,B.Y., Jones,E.V. and Moss,B. (1990) Identification of the vaccinia virus gene encoding an 18-kilodalton subunit of RNA polymerase and demonstration of a 5′ poly(A) leader on its early transcript. *J. Virol*., **64**, 3019–3024.
42. Ink,B.S. and Pickup,D.J. (1990) Vaccinia virus directs the synthesis of early mRNAs containing 5′ poly(A) sequences. *Proc. Natl Acad. Sci. USA*, **87**, 1536–1540.
43. Gudkov,A.T., Ozerova,M.V., Shiryaev,V.M. and Spirin,A.S. (2005) 5′-Poly(A) sequence as an effective leader for translation in eukaryotic cell-free systems. *Biotechnol. Bioeng*., **91**, 468–473.
44. Haugel-Nielsen,J., Hajnsdorf,E. and Regnier,P. (1996) The rpsO mRNA of *Escherichia coli* is polyadenylated at multiple sites resulting from endonucleolytic processing and exonucleolytic degradation. *EMBO J*., **15**, 3144–3152.
45. Dechering,K.J., Cuelenaere,K., Konings,R.N. and Leunissen,J.A. (1998) Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res*., **26**, 4056–4062.
46. Shibata,Y., Carninci,P., Sato,K., Hayatsu,N., Shiraki,T., Ishii,Y., Arakawa,T., Hara,A., Ohsato,N., Izawa,M. *et al*. (2001) Removal of polyA tails from full-length cDNA libraries for high-efficiency sequencing. *Biotechniques*, **1044**, 1048–1049.