Available online at www.sciencedirect.com
ScienceDirect
Plant Science 173 (2007) 371–380
PLANT SCIENCE
www.elsevier.com/locate/plantsci

Review

Tag-based approaches for deep transcriptome analysis in plants

Miguel E. Vega-Sánchez, Malali Gowda, Guo-Liang Wang *

Department of Plant Pathology, Plant Molecular Biology and Biotechnology Program, The Ohio State University, 2021 Coffey Rd., 201 Kottman Hall, Columbus, OH 43210, United States

Received 16 May 2007; received in revised form 10 July 2007; accepted 16 July 2007
Available online 20 July 2007

Abstract

Serial analysis of gene expression (SAGE) has pioneered the use of short-tag sequences derived from the 3′-ends of cDNAs for transcriptome research. Many new tag-based technologies, capitalizing on the success of SAGE, have been developed in the last decade greatly improving the tag length, cloning efficacy and the depth of transcriptome analysis in targeted genomes. Moreover, the recent introduction of novel 5′ mRNA-end isolation methods has allowed for transcription start site determination and promoter region analyses and thus has facilitated accurate annotation of sequenced genomes. Another deep transcriptome analysis method called massively parallel signature sequencing (MPSS) has recently been applied to the annotation of both *Arabidopsis* and rice genomes, not only unraveling the complexities of transcribed regions in plants, but also leading to the identification of novel small RNA species. As new robust sequencing technologies are being developed, tag-based transcriptome analysis will become more affordable and popular in the plant research community. The present review discusses the current status and future challenges of tag-based technologies for the in-depth analysis of plant transcriptomes.

Keywords: Tag-based methods; Transcriptome research; MPSS; RL-SAGE; 5′-RATE; SuperSAGE

Contents

1. Introduction

The recent completion of the rice (*Oryza sativa*) [1–3] and *Arabidopsis thaliana* [4] full genome sequences constitutes a paramount achievement in plant sciences. With a completed genome comes the next daunting task of identifying, classifying and characterizing both gene-coding units and regulatory elements interspersed in a myriad of repetitive and so-called ‘‘junk DNA’’. The global study of transcription and its regulation is the objective of transcriptome research [5], aiming at identifying all transcripts in a cell or tissue present at a given time.

Long before any complex genome sequence was ever elucidated, expressed sequenced tags (EST) emerged as the first high throughput, tag-based method for the study of gene expression and for genome annotation [6]. EST-based

* Corresponding author. Tel.: +1 614 292 9280; fax: +1 614 292 4455.
E-mail address: wang.620@osu.edu (G.-L. Wang).

0168-9452/$ – see front matter
doi:10.1016/j.plantsci.2007.07.005

approaches constitute one of the most important resources for transcriptome analysis in many species [7], and can probably be considered the starting point for any new genomic project. The more recent generation of full-length (FL) cDNA libraries has led to improved and more accurate genome annotation, as identification of open reading frames as well as 5′ and 3′ untranslated regions is achieved [8,9]. However, EST and FL-cDNA sequencing are limited due to insufficient depth of coverage that renders them ineffective for low-abundance transcript detection [10]. Furthermore, since each clone corresponds to a single transcript, ESTs and FL-cDNAs have not proven economical for library saturation in the long-run, even after considerable drops in sequencing costs [11,12].

Serial analysis of gene expression (SAGE) pioneered the use of short-tag sequencing methods and created an unprecedented array of modifications that have greatly improved the original technique first described in 1995 (Fig. 1) [11,13–15]. Just as with cDNA library construction, SAGE and its derivative technologies start with isolation of total RNA, purification of mRNA and cDNA synthesis. Instead of cloning the cDNA fragments into a plasmid vector, the double stranded cDNA is cleaved with an anchoring enzyme (the restriction endonuclease NlaIII) at the 3′-most end of the molecule, followed by adapter ligation containing a Type IIs restriction endonuclease site such as BsmFI (also known as the tagging enzyme). Type IIs enzymes cut a few base pairs away from their recognition site creating "tags". Multiple tags are ligated together to generate concatemers and then cloned into a plasmid vector for sequencing [13]. Approximately 22–50 tags, each from a distinct transcript, are present per clone, considerably improving the throughput of data generation per sequencing run as compared to EST or FL-cDNA sequencing approaches [11,13]. It has been estimated that SAGE is 26 times more sensitive than EST in detecting rarely expressed transcripts [10]. Another

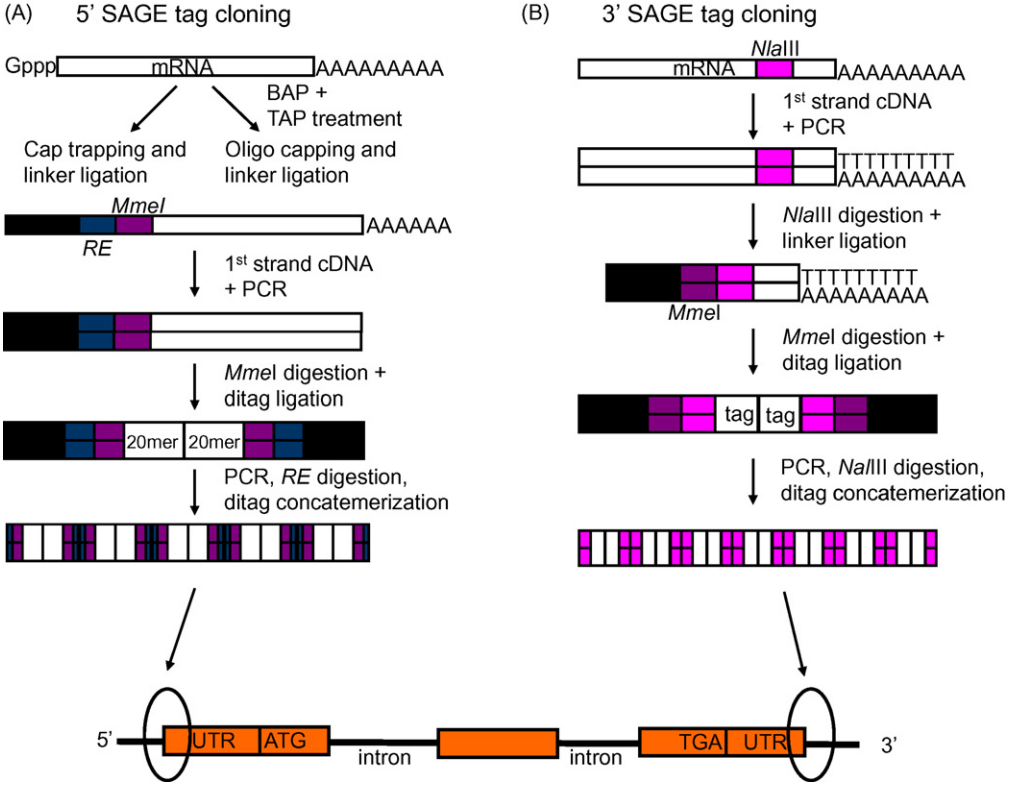

Fig. 1. Overview of 5′ and 3′-end SAGE methods. (A) Overview of 5′ end-derived methods. Full-length cDNAs are captured by using either the cap-trapping (CAGE, 5′LongSAGE) or oligo-capping methods (5′-end SAGE, 5′-RATE). Cap-trapping involves the chemical modification of the cap structure of the mRNA by adding a biotin tag and purification of biotinylated molecules using streptavidin beads. Oligo-capping entails first decapping of mRNA with tobacco acid pyrophosphatase (TAP), 5′-dephosphorylation of uncapped mRNA with bacterial alkaline phosphatase (BAP) and finally, ligation of an adapter (see below). Two pools are usually formed at this point (not shown) and adapters containing primer sequences (in black), a restriction endonuclease site (in blue; XhoI, for 5′-end SAGE, XmaJI for CAGE), and the tagging enzyme (MmeI, in violet) are ligated to the 5′ region of mRNAs. Double stranded cDNA is created by RT-PCR and digested with the tagging enzyme to release the tags; pools are combined, and tags are ligated to form ditags. Ditags are amplified by PCR (using primers complementary to the adapter sequences), followed by digestion with the restriction endonuclease to remove adapters. Ditags are ligated to form concatemers; concatemers are cloned in a plasmid vector and sequenced. CAGE follows slight modifications to this standard procedure; 5′-RATE uses NlaIII as tagging enzyme, and obviates the need for concatemer formation by using 454 sequencing. Identified tags generally match to the transcription start site regions of full-length cDNAs. (B) Overview of 3′-end-derived methods. PolyA mRNA is captured using oligo dT, converted to double stranded cDNA, and then digested with the anchoring enzyme (in most cases NlaIII-showed here in pink). Two pools are created (not shown). Adapters containing the tagging enzyme (BsmFI for SAGE; MmeI (shown here in violet) for LongSAGE, RL-SAGE and 3′LongSAGE; EcoP15I for SuperSAGE) and primer sequences (in black) are ligated to the NlaIII-digested cDNAs. Tags are released by digestion with the tagging enzyme, and at this point tags from both pools are combined for ligation of tags to form ditags. Ditags are amplified by PCR using specific primers complementary to the adapter sequences. Adapters are removed by NlaIII digestion, and ditags are ligated to form longer molecules (concatemers) that are cloned into a plasmid vector. Clones are sequenced, and tags are extracted for further analysis. Most tags match to 3′ regions of cDNAs.