

# Virtual-SAGE: A New Approach to EST Data Analysis

Valeriy POROYKO, Vladimir CALUGARU, Mark FREDRICKSEN, and Hans J. BOHNERT\*

*Department of Plant Biology, and Department of Crop Sciences, University of Illinois at Urbana/Champaign, 1201 W. Gregory Drive, Urbana, IL 61821, USA*

(Received 9 January 2004; revised 20 February 2004)

## Abstract

We present a computer program, termed V-SAGE (Virtual-SAGE), designed to facilitate the analysis of gene expression profiles by combining elements of SAGE (Serial Analysis of Gene Expression) with high-throughput EST analysis. The program re-iteratively correlates sequence tags adjacent to poly(A) tail sequence strings with a second or several tags located within the cDNA adjacent to the recognition sequences of frequently-cutting endonucleases. By recording tags and their distance, the program generates an expression profile from large numbers of sequences, groups sequences according to tags, and identifies alternatively spliced transcripts as well as transcripts that are characterized by 3'-UTR sequences of different length. We discuss the application of V-SAGE to a collection of corn root segment transcripts.

**Key words:** Virtual-SAGE; alternative splicing; transcript 3'-end variation; high-throughput EST analysis; software

## 1. Introduction

Serial Analysis of Gene Expression (SAGE), described in 1995, is by now an established method of gene expression profiling.<sup>1</sup> The method is based on the principle that tags 10–14 nucleotides in length are sufficiently instructive for the identification of each transcript in a collection of cDNA clones. SAGE provides valid data about the structure of a gene expression profile, and can be used for gene discovery as well. In comparison with other methods of gene profiling, only in-depth EST analysis has the same capacity for providing comprehensive and quantitative expression profiles.

In contrast to SAGE, EST-data analysis necessitates not only substantial efforts devoted to high throughput cDNA sequencing, but requires equally extensive bioinformatics-type work for quality control, contig assembly, often to be followed by reiterative annotations. Also, the complex sequence of operational steps can lead to the misinterpretation of data, especially when low quality DNA sequence data are included into contig assemblies. SAGE has several advantages over EST-based profiling in that it does not require the same scrupulous attention to library construction, that it accelerates analyses, and that it is more cost efficient. One limitation may be identification of individual tags or transcripts with high reliability, mostly because of the

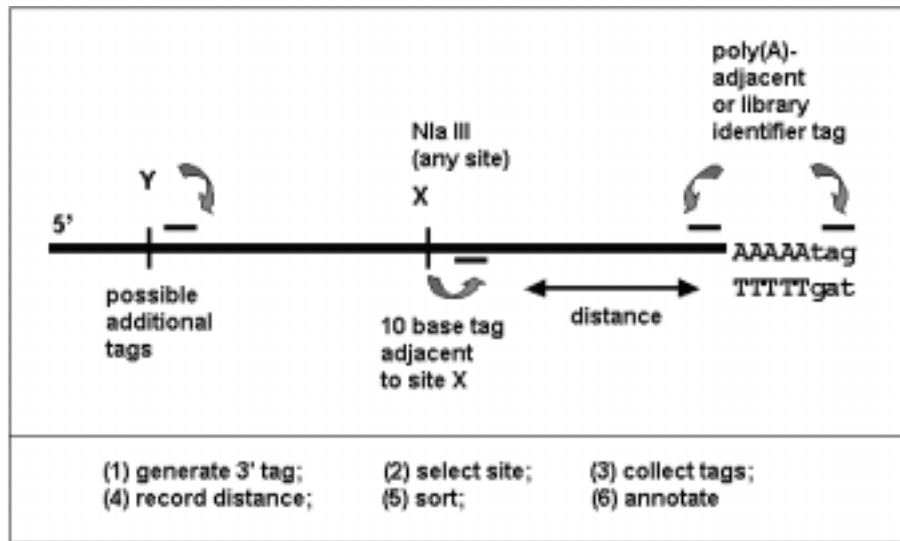
possibility that population polymorphism might exist.<sup>2,3</sup> Another compounding feature is that SAGE shows extreme sensitivity for identifying 3'-end transcript population heterogeneity,<sup>3–6</sup> and it is this quality which we set out to exploit. SAGEmap,<sup>3</sup> a public domain program facilitating the identification of tags in SAGE, does not completely solve the problem of tag-to-gene ambiguity, which is possibly one reason why SAGE is not more widely used. Presently, tag identification is available for 11 species only, including *Arabidopsis thaliana*, *Bos taurus*, *Homo sapiens*, *Medicago truncatula*, *Meleagris gallopavo*, *Mus musculus*, *Rattus norvegicus*, *Sus scrofa*, *Triticum aestivum*, *Pinus taeda*, and *Vitis vinifera* (<http://www.ncbi.nlm.nih.gov/sage/>). We describe a program that allows for the extraction of a combination of user-chosen tags and their clustering to obtain information about the complexity of sequences in a cDNA library, its application to identify 3'-end variation and, to some degree, the identification of alternatively spliced transcripts, in any population of cDNAs irrespective of the state of annotation. The program is exemplified by the analysis of ESTs from a project that determined unigene composition in a corn root collection of cDNA libraries that had been tagged according to their position in the root tip.

### 1.1. The V-SAGE objective

We have made use of the fact that 3'-end processing signals in plants show considerable irregularity, and cannot easily be predicted or explained by a simple pattern.<sup>8</sup>

Communicated by Satoshi Tabata

\* To whom correspondence should be addressed. Tel. +1-217-265-5475, Fax. +1-217-333-5574, E-mail: bohnert@life.uiuc.edu



**Figure 1.** The V-SAGE algorithm. The scheme identifies the process of tag and distance collection that can be applied to any EST population in which tags can be identified, e.g., adjacent to poly(A) extensions of the sequences.

We set out from an algorithm for tag-to-gene assignments that had originally been described by Lash et al.<sup>3</sup>, designed to search for polyadenylation signals downstream of the 3'-end of predicted ORFs in human genomic DNA. This algorithm cannot straightforwardly be applied to plant genomic DNA sequences, and it excludes much information derived from the generation of SAGE maps. The considerable unpredictability of plant 3'-ends delays SAGE tag mapping for plant species, and suggests that 3'-end ESTs could be used as a valid source for generating SAGE maps in plants. This recognition focused our attention on the advantages that might result from combining two transcript profiling methods, SAGE and EST analysis.

The approach described here, termed 'Virtual SAGE' (V-SAGE), takes the efficiency, speed, and reliability of data mining from classical SAGE, and combines it with the expediency of gene identification that characterizes EST analysis. The concept is based on establishing a correlation between several tags extracted from EST sequence collections at different distances from the poly(A) region (Fig. 1). By extracting tags at the extreme 3'-end and internal tags from the EST sequences, complexity is reduced, clustering of similar transcripts into groups (populations of tags) is possible, and BLAST analysis marks the resulting contigs. Finally, 3'-terminal variants and alternatively spliced transcripts, mostly in the 3'-region of a contig population, within these groups are rapidly identified.

### 1.2. The V-SAGE software structure

The V-SAGE software that has been generated is an *in silico* emulation of the established SAGE protocol. The

input data consists of a FASTA file that contains a string or strings of EST sequences from the library of interest that have been determined from the 3' end. Data processing includes the following steps (Fig. 1):

1. Program identifying the poly(A) region (with a minimum of eight A residues).<sup>2</sup>

2. Extracting the first 10-base tag immediately upstream of the poly(A)<sup>+</sup> region. In libraries that have been tagged to trace the origin of the sequence from specific tissues (e.g., root segments, leaves, flowers) or experimental conditions (e.g., light, drought, cold, pathogen), V-SAGE can extract and collect the tag identity.

3. Collecting the set of 10-base tags, which are immediately 3'-adjacent to the site that is most 3' of a selected restriction endonuclease cleavage site, e.g., for *Nla* III (CATG), and records the distance from this site to the poly(A)<sup>+</sup>-adjacent tag in nucleotides.

4. Assigning a clone name identifier to each pair of tags.

The result of the data processing consists of a set of tags located upstream of the poly(A)<sup>+</sup> region in each EST. This set is a unique identifier for any transcript, which can then be used for further analyses, such as a digital representation of cellular gene expression, for studies of 3'-UTR variability, and also as a map for regular SAGE transcript profiling.

The principle of V-SAGE is applicable to any number of sequence collections and short oligonucleotide strings with certain precautions. The use of restriction endonucleases with 4-bp recognition sites, for example, is determined by the average length of the sequences targeted. In our example, the average length of approximately 500 nucleotides (theoretical optimum: 256 bp, with the experimental data presenting a range from 100 to 600 nu-

cleotides) provided a suitable frequency. This length does not, however, allow for the extraction of tags by restriction sites with 6-bp recognizing sites, which would require a sequence length of at least 1 kb, and preferably higher. When longer sequences are available, splice variants within the 5'-coding regions of transcripts may also be identified by V-SAGE.

The Perl code script representing the V-SAGE algorithm is freely available at <http://www.life.uiuc.edu/bohnert/vsage/VSAGE.htm>. The EST sequences analyzed here were derived from three root cDNA libraries from corn, which had tags added during cDNA library construction to identify transcripts in specific root segments, ([http://rootgenomics.missouri.edu/Plantrootgenomics\\_current/index.htm](http://rootgenomics.missouri.edu/Plantrootgenomics_current/index.htm)). These sequences have been deposited in GenBank, and can also be retrieved from a local databank (<http://www.life.uiuc.edu/Bohnert/>). The contig assembly platform used here was CAP3.<sup>7</sup>

### 1.3. Digital representation of cellular gene expression

The output file that includes the tags was exported to Excel worksheets. The data for tags extracted from the EST sequence collection can then be used for the generation of a list of unique tags. By counting the frequency of appearance of any tag and combination of tags in a sequence collection, expression profiles can be established for the selected population. One advantage is that the clustering is not dependent on prior annotation. Ultimately, the score and structure of tags from different libraries can be compared, or the data may be used for further cluster analysis to reveal any underlying complexity. Data clustering and mining may be done by, for example, the Spotfire Decision Site software, which facilitates finding and categorizing genes and gene expression patterns (<http://www.spotfire.com/>).

### 1.4. Large-scale screening for 3'-UTR variants

Another possible application of V-SAGE is described in detail below. This is the search for 3'-UTR variants. Many studies have suggested that the length of a 3'-UTR sequence could play an important role in determining both translational efficiency and the stability of mRNAs.<sup>9,10</sup> For example, a 39-fold higher expression of luciferase has been observed based on the presence of a 19-bp 3'-UTR when comparing poly(A)<sup>+</sup> RNA to the poly(A)<sup>-</sup> variant.<sup>11,12</sup> Increasing the size of the 3'-UTR tail to 156 bp reduced this difference to sevenfold higher. In addition, mRNA half-life of the poly(A)-mRNA increased by a factor of 2.5 when a 156-bp 3'-UTR, compared to a 4-bp 3'-end structure, was present.<sup>11,12</sup> Also, the presence of a long 3'-UTR sequence influenced the translatability of transcripts, possibly by stabilizing their recognition by initiation or elongation factors and/or stability of binding to ribosomes or

polyribosomes.<sup>11</sup> Finally, it has been documented that the structure of 3'-UTR sequences between the stop codon and the polyadenylated string influences the stability of mRNAs during stress situations, for example during heat shock.<sup>13</sup> The output of the V-SAGE software provides necessary information for the identification of 3'-UTR variants and also allows massively parallel, rapid screening of large sets of data.

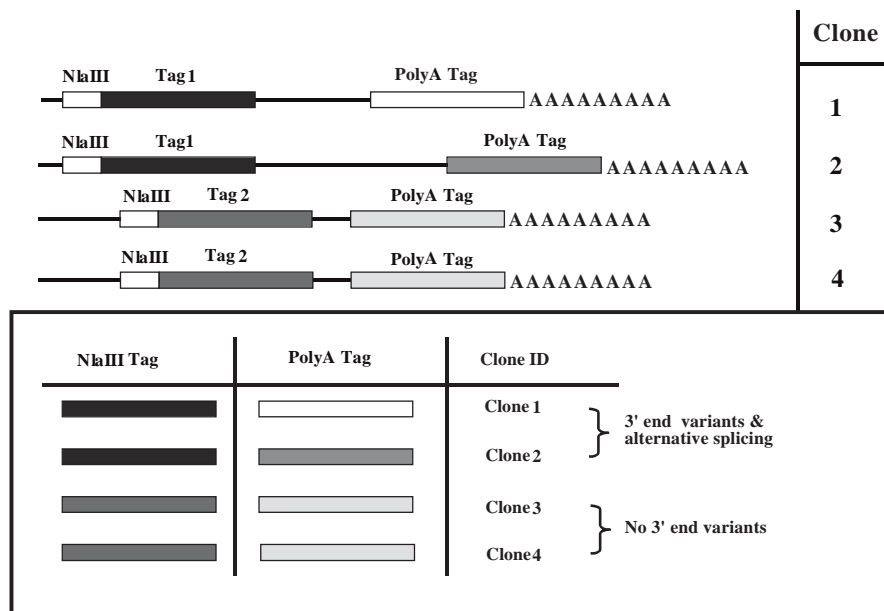
### 1.5. V-SAGE application examples

We have used V-SAGE for the analysis of corn root segments that show distinct growth characteristics during water deficit, compared to well-watered roots.<sup>14</sup> For the analysis presented here, sequences were used that were derived from a normalized corn root cDNA library. The library was constructed by normalizing a combination of four primary cDNA libraries. Each of these four primary libraries represented cDNAs from a segment of the corn root tip: segment 1 contained the cDNAs from the tip 3 mm of the root, segment 2 (3–7 mm), segment 3 (7–12 mm), and segment 4 (12–20 mm). In each primary library (segments 1–4) an oligonucleotide identifier (bar code) was introduced, and the four primary libraries were then combined to generate a normalized cDNA library. In this way, three normalized libraries were generated for three different growth conditions: (i) segments from well-watered roots, (ii) 5-hr drought-stressed root segments, and 48 hr drought-stressed root segments, each containing four segment libraries. A collection of approximately 15,000 ESTs, representing approx. 7,000 unigenes from such root segments, has been deposited in GenBank (available at: [http://rootgenomics.missouri.edu/Plantrootgenomics\\_current/index.htm](http://rootgenomics.missouri.edu/Plantrootgenomics_current/index.htm)).

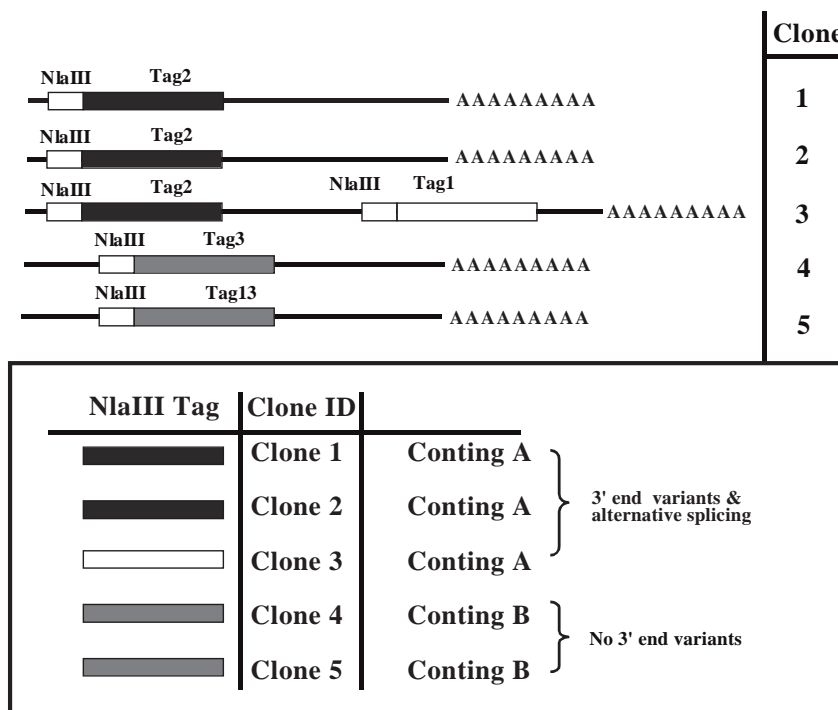
One example, schematically depicted in Fig. 2, monitors 3'-UTR variability and provides a paradigm for the application of V-SAGE as an analytical tool. Different poly(A)<sup>+</sup> tags that become associated with the same or a different *Nla* III-associated tag distinguish either different transcripts or copies of the same transcript with differing 3'-ends.

Another illustration of 3'-UTR variability is described in Fig. 3. In this scenario, the 3'-end of the sequence for one transcript is longer and contains an additional *Nla* III site. Such a situation would confuse a simple search algorithm and make recognition difficult, but a combination of V-SAGE and contig-assembly software will uncover such variants. While any contig assembly platform may be used, we have been using CAP3 (settings: 50 bp for sequence overlap and 95% identity for the overlapping region) with subsequent BLAST annotation of the obtained contigs. In the example chosen, a 3'-end difference is characterized by a divergence in the *Nla* III-tag structure for the same annotated contig.

V-SAGE is also suitable to detect splice variants,



**Figure 2.** Schematic presentation of V-SAGE results. Identical *Nla* III-tags are correlated with the same or different poly(A)<sup>+</sup>-tags that identify 3'-UTR variability and/or alternatively spliced transcripts.



**Figure 3.** Schematic presentation of complex variation of 3'-ends of cDNAs. The example identifies 3'-end variation as a longer 3'-UTR region. Different *Nla* III tags identify different polyadenylation sites for one gene.

for example in the region between a stop codon and the poly(A)<sup>+</sup>-tail, which has, for some genes, been shown to regulate expression to an appropriate level in the adjustment of protein expression to external conditions.<sup>12</sup> Splice variants are reliably recognized by V-SAGE through a comparison of the distance between

the *Nla* III and poly(A)<sup>+</sup>-tag for any pair of tags.

Several cDNAs encoding functionally unknown glycine-rich proteins in *Zea mays* can illustrate the theoretical scenarios of 3' UTR variability introduced before. Figure 4A shows alignment of two clones, ZMRWS48\_0B20-006-C01.S1 and ZMRWS05\_0A20-007-

## A

```

1
ZMRWS48_0B20-006-C01.S1 catgAACGGCAAGGAGCTCGACGGCCGTAACATCACCGTTAACAGGCCAGTCCCGTGG
ZMRWS48_0B10-005-F03.S1 catgAACGGCAAGGAGCTCGACGGCCGTAACATCACCGTTAACAGGCCAGTCCCGTGG
ZMRWS05_0A20-007-C07.S1 catgAACGGCAAGGAGCTCGACGGCCGTAACATCACCGTTAACAGGCCAGTCCCGTGG

ZMRWS48_0B20-006-C01.S1 CGGTGGCGGTGGCGGCGGTGGCTACGGCGGCGGTGCGGCGGCGGCGGCTATGGTGGCGG
ZMRWS48_0B10-005-F03.S1 CGGTGGCGGTGGCGGCGGTGGCTACGGCGGCGGTGCGGCGGCGGCGGCTATGGTGGCGG
ZMRWS05_0A20-007-C07.S1 CGGTGGCGGTGGCGGCGGTGGCTACGGCGGCGGTGCGGCGGCGGCGGCTATGGTGGCGG

ZMRWS48_0B20-006-C01.S1 GCGCCGTGACGCGGTTATGGCGGCGGTGGCGGCTACGGCGGTGCGGCGGAGGGTGGTGG
ZMRWS48_0B10-005-F03.S1 GCGCCGTGACGCGGTTATGGCGGCGG-----
ZMRWS05_0A20-007-C07.S1 GCGCCGTGACGCGGTTATGGCGGCGGTGGCGGCTACGGCGGTGCGGCGGAGGGTGGTGG
                                     G G G G Y G G R R E G G G

ZMRWS48_0B20-006-C01.S1 CGGCGGCTACGGAGGCGGTGGCGGCTACGGCGGTGCGGCGGAGGGTGGTGGTGGCGGCTA
ZMRWS48_0B10-005-F03.S1 -----
ZMRWS05_0A20-007-C07.S1 CGGCGGCTACGGAGGCGGTGGCGGCTACGGCGGTGCGGCGGAGGGTGGTGGTGGCGGCTA
                                     G G Y G G G G G Y G G R R E G G G G G Y

ZMRWS48_0B20-006-C01.S1 CGGCGGCGGCGGCGGCGGTGGAGGGACTGATGTTGGGCCCATCTGGCTTCGGCCGAG
ZMRWS48_0B10-005-F03.S1 -----CGGCGGCTGGAGGGACTGATGTTGGGCCCATCTGGCTTCGGCCGAG
ZMRWS05_0A20-007-C07.S1 CGGCGGCGGCGGCGGCGGTGGAGGGACTGATGTTGGGCCCATCTGGCTTCGGCCGAG
                                     G G G G G G W R D STOP

ZMRWS48_0B20-006-C01.S1 TTATCTTATCTATCTATAGTATCGTGTACCGTTCGCTTCTGTCACCGTGTAGTGTCCG
ZMRWS48_0B10-005-F03.S1 TTATCTTATCTATCTATAGTATCGTGTACCGTTCGCTTCTGTCACCGTGTAGTGTCCG
ZMRWS05_0A20-007-C07.S1 TTATCTTATCTATCTATAGTATCGTGTACCGTTCGCTTCTGTCACCGTGTAGTGTCCG

ZMRWS48_0B20-006-C01.S1 TTCTACCTTTGGATTAGGTGTTGGTACCCCTGTTGTTCCCTTTGTTTGTCTCCGCTATGA
ZMRWS48_0B10-005-F03.S1 TTCTACCTTTGGATTAGGTGTTGGTACCCCTGTTGTTCCCTTTGTTTGTCTCCGCTATGA
ZMRWS05_0A20-007-C07.S1 TTCTACCTTTGGATTAGGTGTTGGTACCCCTGTTGTTCCCTTTGTTTGTCTCCGCTATGA

ZMRWS48_0B20-006-C01.S1 AACGAGACGAGAGAAGAATGAGCAA-----AAAAAAAAAAAAAAAA
ZMRWS48_0B10-005-F03.S1 AACGAGACGAGAGAAGAATGAGCAAGTTTTTTGTTTCGAGCTAAAAAAAAAAAAAAAA
ZMRWS05_0A20-007-C07.S1 AACGAGACGAGAGAAGAATGAGCAAGTTTTTTGTTTCGAGCTAAAAAAAAAAAAAAAA

```

## B

```

3
ZMRWS48_0B10-016-H05.S3 catgAACGGCAAGGAGCTGGACGGCCGAACATCACCGTCAACGAGGCCAGTCCCGCGG
ZMRWS48_0A20-016-E10.S1 catgAACGGCAAGGAGCTGGACGGCCGAACATCACCGTCAACGAGGCCAGTCCCGCGG
1

ZMRWS48_0B10-016-H05.S3 CGGCCGTGGAGCGGCGGCGGTGGTACGGTGGTGGCCGTGGAGGCGGCGGCTACGGCGG
ZMRWS48_0A20-016-E10.S1 CGGCCGTGGAGCGGCGGCGGTGGTACGGTGGTGGCCGTGGAGGCGGCGGCTACGGCGG

ZMRWS48_0B10-016-H05.S3 TGGCGGGCGCCGTGATGGCGGCGGCGGCTACGGCGGTGGCGGCGGCTACGGTGGCGGCGG
ZMRWS48_0A20-016-E10.S1 TGGCGGGCGCCGTGATGGCGGCGGCGGCTACGGCGGTGGCGGCGGCTACGGTGGCGGCGG

ZMRWS48_0B10-016-H05.S3 CGGCTACGGTGGTGGTGGCGGCGGCTACGGCGGTGGCAACCGTGGCGGCGGCTACGGCAA
ZMRWS48_0A20-016-E10.S1 CGGCTACGGTGGTGGTGGCGGCGGCTACGGCGGTGGCAACCGTGGCGGCGGCTACGGCAA
                                     STOP
ZMRWS48_0B10-016-H05.S3 CTCCGACGGGAAGTGGAGGAACTCGACGGTGGGGCCCGCGGCCAAGTTATCTGTGTCG
ZMRWS48_0A20-016-E10.S1 CTCCGACGGGAAGTGGAGGAACTCGACGGTGGGGCCCGCGGCCAAGTTATCTGTGTCG

ZMRWS48_0B10-016-H05.S3 CTGCCGTGATGTTTACCCTAGTCCAGAGGGTTTATCTTCTGTTCTGTTGTTGTTGTT
ZMRWS48_0A20-016-E10.S1 CTGCCGTGATGTTTACCCTAGTCCAGAGGGTTTATCTTCTGTTCTGTTGTTGTTGTT

ZMRWS48_0B10-016-H05.S3 GCCCATCTGTGTTTTTGGATTGCAAGGTCGCTCTGTGTGTCAGTTGTTAGTGTGTTGTTATC
ZMRWS48_0A20-016-E10.S1 GCCCATCTGTGTTTTTGGATTGCAAGGTCGCTCTGTGTGTCAGTTGTTAGTGTGTTGTTATC
                                     2 1
ZMRWS48_0B10-016-H05.S3 CTCGGCTCCAGCAGACCcatgCATCAAACAGcatgGACTGCGGATCGATGGATGCTGTTA
ZMRWS48_0A20-016-E10.S1 -----

ZMRWS48_0B10-016-H05.S3 CCCCCTCAGGCTTTATTCTAAGTTAATCTTAAGGAAAAAAAAAAAAAAAA
ZMRWS48_0A20-016-E10.S1 -----

```

**Figure 4.** Alignment of five EST clones for carboxy terminal portions of a glycine-rich, putative RNA-binding protein from *Zea mays* illustrating categories of 3'-end variation. Panel A. Clones ZMRWS48\_0B20-006-C01.S1 (gi:37390782) and ZMRWS05\_0A20-007-C07.S1 (gi: 37376307) share the same *Nla* III tag followed by different poly(A)<sup>+</sup>-tags; ZMRWS48\_0B10-005-F03.S1 (gi:37388008) and ZMRWS05\_0A20-007-C07.S1 (gi:37376307) show identical *Nla* III and poly(A)<sup>+</sup>-tags but different lengths while maintaining the reading frame and represent alternatively spliced transcripts. Panel B. Clones ZMRWS48\_0A20-016-E10.S1 (gi:37387163) and ZMRWS48\_0B10-016-H05.S3 (gi:37389764) show different *Nla* III tags because different polyadenylation sites were used. In the longer clone, ZMRWS48\_0B10-016-H05.S3, two additional *Nla* III sites are present, indicated by numbers. The contig assembling software places the two clones into the same contig and highlights the 3'-end variation. The sequences are deposited at <http://genome.mnet.missouri.edu/Roots/FileProcess/index.html> and are freely accessible.

**Table 1.** Most frequent CAP3 contigs and V-SAGE tags from a corn root tip cDNA library.

CAP3 contig	total	root segment				no barcode	annotation
		S1	S2	S3	S4		
Contig185	49	0	0	0	0	49	gi 27777630 gb AAO23335.1  O-methyltransferase [Secale cereale]
Contig15	43	1	13	9	11	9	gi 20257667 gb AAM15999.1  glycine-rich RNA binding protein [Zea mays]
Contig1,500	13	1	2	3	5	2	gi 554565 gb AAA72758.1  glutathione S-transferase
Contig266	12	2	6	4	0	0	gi 37532730 ref NP_920667.1  Profilin A [Oryza sativa]
Contig1,348	12	0	5	6	1	0	gi 38345484 emb CAE01698.2  OSJNBa0010H02.22 [Oryza sativa]
Contig166	11	4	1	4	1	1	gi 2498077 sp P93554 NDK1_SACOF nucleoside diphosphate kinase I (NDK I)
Contig328	10	0	6	3	0	1	gi 2226329 gb AAC31615.1  physical impedance induced protein [Zea mays]
Contig1,606	9	2	4	2	0	1	gi 7489428 pir T06199 probable lipid transfer protein [Hordeum vulgare]
Contig431	8	6	0	1	0	1	gi 7440759 pir T02039 acidic ribosomal protein P1a [Zea mays]
Contig228	8	5	0	1	0	2	gi 4582787 emb CAB40376.1  adenosine kinase [Zea mays]
<b>V-SAGE tag</b>							
GACTGCGGAT	37	3	13	8	10	3	gi 20257667 gb AAM15999.1  glycine-rich RNA binding protein [Zea mays]
CTTGCTTGTA	13	1	1	2	7	2	gi 554565 gb AAA72758.1  glutathione S-transferase
AACGGCAAGG	12	4	5	2	1	0	gi 20257675 gb AAM16003.1  glycine-rich RNA binding protein [Zea mays]
TTGATGAGCA	12	0	5	6	1	0	gi 38345484 emb CAE01698.2  OSJNBa0010H02.22 [Oryza sativa]
GTGGTCGAGA	11	2	5	4	0	0	gi 37532730 ref NP_920667.1  Profilin A [Oryza sativa]
TCACCATTTG	10	4	1	3	1	1	gi 2498077 sp P93554 NDK1_SACOF nucleoside diphosphate kinase I (NDK I)
GCCGCGTGCA	9	2	3	2	0	2	gi 7489428 pir T06199 probable lipid transfer protein [Hordeum vulgare]
TCGGCTCTT	9	0	6	3	0	0	gi 2226329 gb AAC31615.1  physical impedance induced protein [Zea mays]
GCGTGGCTAC	8	0	0	1	7	0	gi 37531954 ref NP_920279.1  type-1 pathogenesis-related protein [Oryza sativa]
AACTCAAACC	7	6	0	0	0	1	gi 7440759 pir T02039 acidic ribosomal protein P1a [Zea mays]

C07.S1, by the same *Nla* III-adjacent tags followed by different poly(A)<sup>+</sup>-tags. These sequences represent clones with different 3'-UTR lengths. A third aligned sequence (Fig. 4A) identifies a sequence that is 108 bp shorter in ZMRWS48\_0B10-005-F03.S1 but otherwise identical, including the identical poly(A)<sup>+</sup>- and *Nla* III-adjacent tags found in ZMRWS48\_0B10-005-F03.S1 and ZMRWS05\_0A20-007-C07.S1. It represents an alternatively spliced version of the same gene (accession number: X12564; GI: 22312) in which the splice variation can be identified. Figure 4B displays the appearance of additional *Nla* III-tags due to a longer 3'-UTR sequence present in clone ZMRWS48\_0B10-016-H05.S3 in comparison to ZMRWS48\_0A20-016-E10.S1, an alteration that was revealed in the process of contig assembling. All five sequences identify a glycine-rich RNA-binding protein, homologous to *Arabidopsis* AtGRP7, demonstrating differential 3'-end formation and alternative splicing. Alternative splicing of proteins in this class has been observed before.<sup>15,16</sup>

V-SAGE can be used for routine characterizations of EST collections, as it is documented here in the annotation of sequences from a cDNA library and in EST data mining. The output by V-SAGE is compared to that by the common approach of CAP3-based contig assembly. The example discussed here used 4749 sequences from the normalized (well-watered conditions) corn root cDNA libraries ([http://rootgenomics.missouri.edu/Plantrootgenomics\\_current/index.htm](http://rootgenomics.missouri.edu/Plantrootgenomics_current/index.htm)). The CAP3 program assembled the sequences into a set of 3261 unique transcripts. Out of the 4749 sequences, V-SAGE extracted 4197 sequences that con-

tained poly(A)-tails and produced a set of 3086 unique tag groups. While CAP3 does not depend on the presence of poly(A) tails, bar-coding and the presence of poly(A) tails provides additional information. Table 1 provides an example for the versatility of V-SAGE in comparison with CAP3. It compares the distribution of the most abundant transcripts identified by the two programs in the corn root segment EST collection. An *O*-methyltransferase was identified by CAP3 (49 copies), not through V-SAGE tags but due to low-quality 3'-end of these transcripts. V-SAGE depends on such additional information in the form of tags that are introduced during library construction or by relying on common features of transcripts. The case of a glycine-rich RNA-binding protein provides just one example of the power of the V-SAGE process. V-SAGE tags separate two variants of this gene (see Fig. 4) that are not recognized or distinguished by CAP3. A short form (tag: AACGGCAAGG) and a long transcript (tag: GACTGCGGAT) indicate different 3'-end formation. The long form was more abundant in corn root, and the segment tags indicated that the transcript with a longer 3'-end increased from S1 to S4. This difference highlights development, with the longer form transcript increasing as the root tissue matures.

V-SAGE tags provide the opportunity for clustering in the absence of complete annotations and can thus provide focus to analyses. Clustering of root segment contigs and EST abundance by either V-SAGE or CAP3, which requires prior analysis, exemplifies this statement. The number of transcripts representing each contig or tag group in each of the four segments was divided by the

total number contigs/tag groups found in the segment. Clustering was done according to UPGMA (unweighted average). CAP3 and V-SAGE produced identical patterns. The similarity measure used Euclidian distance, and average values for the ordering function. Segments S1 and S2 appear to be most similar, S3 more similar to S1/S2, with segment S4 most dissimilar. This pattern reflects the maturation process in the root tip.

The V-SAGE program, which can be downloaded from (<http://www.life.uiuc.edu/bohnert/vsage/VSAGE.htm>), accommodates changes in the combination of tags that may be searched. V-SAGE provides further flexibility by the ability to incorporate, in high-throughput fashion, additional EST sequences into an existing species-specific collection even in the absence of annotations. Also, a comparison with the growing collection of certified full-length cDNAs with newly emerging transcript sequences can provide a global picture and information about the complexity underlying any transcript profile with respect to the frequency of alternatively spliced sequences in plant transcriptomes. Finally, V-SAGE allows a point in sequence analysis to be addressed that has not yet received much attention. The formation of transcript 3'-ends in plants shows high variability, and the selection of polyadenylation sites in transcript termination seems to be influenced by the environment or experimental conditions used. The possibility that this variability influences transcript utilization by ribosomes and/or transcript longevity has been pointed out before.<sup>17–23</sup> V-SAGE provides a way to assess this variability. The program can be applied to analyze the growing number of transcript collections that have been generated with plant material grown under different conditions.

**Acknowledgements:** We thank Dr. Xiaoqiu Huang for providing the CAP3 program, Lihua Jiang for help with the Perl script, and Amber Kroeger for annotations. Work by Amber Kroeger has been supported by an NSF-REU program grant. Supported by UIUC institutional grants, and the U.S. National Science Foundation, plant genome program (DBI-0211842) is gratefully acknowledged.

## References

- Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. 1995, Serial analysis of gene expression, *Science*, **270**, 484–487.
- Kannbley, U., Kapinya, K., Dirnagl, U., and Trendelenburg, G. 2003, Improved protocol for SAGE tag-to-gene allocation, *Biotechniques*, **34**, 1212–1219.
- Lash, A. E., Tolstoshev, C. M., Wagner, L., et al. 2000, SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.
- Zhou, G., Chen, J., Lee, S., Clark, T., Rowley, J. D., and Wang, S. M. 2001, The pattern of gene expression in human CD34(+) stem/progenitor cells. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 13966–13971.
- Lorenz, W. W. and Dean, J. F. D. 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.
- Stanton, J. L., Bascand, M., Fisher, L., Quinn, M., Macgregor, A., and Green, D. P. 2002, Gene expression profiling of human GV oocytes: an analysis of a profile obtained by serial analysis of gene expression (SAGE). *J. Reprod. Immunol.*, **53**, 193–201.
- Huang, X. and Madan, A. 1999, CAP3: A DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
- Graber, J. H., Cantor, C. R., Mohr, S. C., and Smith, T. F. 1999, In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 14055–14060.
- Kawaguchi, R. and Bailey-Serres, J. 2002, Regulation of translational initiation in plants. *Curr. Opin. Plant Biol.*, **5**, 460–465.
- Mazumder, B., Seshadri, V., and Fox, P. L. 2003, Translational control by the 3'-UTR: the ends specify the means. *Trends Biochem. Sci.*, **28**, 91–98.
- Tanguay, R. L. and Gallie, D. R. 1996a, The effect of the length of the 3'-untranslated region on expression in plants. *FEBS Lett.*, **394**, 285–288.
- Tanguay, R. L. and Gallie, D. R. 1996b, Translational efficiency is regulated by the length of the 3' untranslated region. *Mol. Cell Biol.*, **16**, 146–156.
- Schwerin, M., Maak, S., Hagendorf, A., von Lengerken, G., and Seyfert, H. M. 2002, A 3'-UTR variant of the inducible porcine hsp70.2 gene affects mRNA stability. *Biochim. Biophys. Acta*, **1578**, 90–94.
- Sharp, R. E. and LeNoble, M. E. 2002, ABA, ethylene and the control of shoot and root growth under water stress. *J. Exp. Bot.*, **53**, 33–37.
- Hirose, T., Sugita, M., and Sugiura, M. 1993, cDNA structure, expression and nucleic acid-binding properties of three RNA-binding proteins in tobacco: occurrence of tissue-specific alternative splicing. *Nucleic Acids Res.*, **21**, 3981–3987.
- Staiger, D., Zecca, L., Wiczorek Kirk, D. A., Apel, K., and Eckstein, L. 2003, The circadian clock regulated RNA-binding protein AtGRP7 autoregulates its expression by influencing alternative splicing of its own pre-mRNA. *Plant J.*, **33**, 361–371.
- Raz, R., Jose, M., Moya, A., Martinez-Izquierdo, J. A., and Puigdomenech, P. 1993, Different mechanisms generating sequence variability are revealed in distinct regions of the hydroxyproline-rich glycoprotein gene from maize and related species. *Mol. Gen. Genet.*, **233**, 252–259.
- Breiteneder, H., Michalowski, C. B., and Bohnert, H. J. 1994, Environmental stress-mediated differential 3' end formation of chloroplast RNA-binding protein transcripts. *Plant Mol. Biol.*, **26**, 833–849.
- Rothnie, H. M., Reid, J., Hohn, T. 1994, The contribution of AAUAAA and the upstream element UUUGUA to the efficiency of mRNA 3'-end formation in plants. *EMBO J.*, **13**, 2200–2210.
- Kopriva, S., Cossu, R., and Bauwe, H. 1995, Alternative splicing results in two different transcripts for H-protein of the glycine cleavage system in the C4 species *Flaveria trinervia*. *Plant J.*, **8**, 435–441.

21. Sunako, T., Sakuraba, W., Senda, M., Akada, S., Ishikawa, R., Niizeki, M., and Harada, T. 1999, An allele of the ripening-specific 1-aminocyclopropane-1-carboxylic acid synthase gene (ACS1) in apple fruit with a long storage life. *Plant Physiol.*, **119**, 1297–1304.
22. Rasori, A., Ruperti, B., Bonghi, C., Tonutti, P., and Ramina, A. 2002, Characterization of two putative ethylene receptor genes expressed during peach fruit development and abscission. *J. Exp. Bot.*, **53**, 2333–2339.
23. Magnotta, S. M., Gogarten, J. 2002, Multi-site polyadenylation and transcriptional response to stress of a vacuolar type H<sup>+</sup>-ATPase subunit A gene in *Arabidopsis thaliana*. *BMC Plant Biol.*, **2**, 3.