Serial Analysis of Gene Expression: Technical Considerations and Applications to Cardiovascular Biology

Willmar D. Patino, Omar Y. Mian, Paul M. Hwang

Abstract—It has been 7 years since serial analysis of gene expression (SAGE) and microarray hybridization techniques were simultaneously introduced to allow the screening of thousands of expressed genes. Both techniques have stood up to the test of time as evidenced by their widespread use, and both have been used for studying cardiovascular diseases. SAGE has been used more extensively to study cancer cells, but it has also been used to examine gene expression in systems as divergent as rice seedlings, yeast, and Caenorhabditis elegans. In this review, a summary of the advances in SAGE technology and its unique attributes and potential applications to the cardiovascular system will be presented.

Key Words: expression ■ gene ■ cardiovascular

The major difference between microarray hybridization and serial analysis of gene expression (SAGE) techniques is that the latter does not require prior knowledge of the sequences to be analyzed, as SAGE is a sequencing-based gene expression profiling technique.1,2 For organisms with poorly characterized genomic and expressed sequences, SAGE can be used to obtain complete transcriptional profiles of expressed genes, albeit unknown genes. But even after the sequencing of the human genome, predicting all the encoded genes continues to be a significant challenge.3,4 A recent adaptation of SAGE, called LongSAGE, allows the derived transcriptome to be used in annotating expressed genes in the genome.5 In this sense, SAGE is a truly global and unbiased gene expression technique.

The choice of SAGE versus microarray hybridization technique depends on several factors, such as the scope of the genetic screen, the number of samples, the amount of starting material, and the availability of resources such as an automated DNA sequencer. Approximately $1.5 \times 10^6$ bases need to be sequenced for a simple two library comparison, but this factor should not be a significant impediment for most laboratories with the current advances in automated sequencing. Nonetheless, large numbers of samples may be analyzed more efficiently by using the microarray technique. For sample-limiting experiments, the polymerase chain reaction (PCR) amplification step of the SAGE technique decreases the amount of RNA needed compared with standard microarray techniques. The amount of RNA needed ranges from 50 to 500 ng of mRNA or 5 to 50 μg of total RNA. SAGE libraries have been made using significantly less material (9 oocytes to 100,000 cells).6–8 Recently, the construction of a SAGE library from a one-cell mRNA equivalent has also been reported although the final quality of this library remains to be demonstrated.9

In studies that have compared SAGE and microarray data, there appear to be a good correlation between the two techniques although the SAGE technique was found to be more quantitatively reproducible in one recent study.10–12 It is also possible to take advantage of both techniques by screening broadly with the SAGE technique, then serially following with a more focused custom-made microarray experiment.10 To date, no software is available for directly comparing...
SAGE to microarray databases, but recently a software capable of using either microarray or SAGE databases for visualizing metabolic pathways at a genome-wide level has been reported.13

The sequencing requirement of SAGE gives it a unique advantage. Its digital database facilitates direct comparisons between SAGE libraries.1 The ability to query over 100 human SAGE libraries representing three million transcripts via the internet and to perform virtual Northern blots represents a powerful resource available to everyone in the scientific community (www.ncbi.nlm.nih.gov/SAGE).1 In contrast, comparing microarray experiments may be more difficult due to a number of random and systematic errors that are difficult to eliminate between different investigators or laboratories.14

Concepts and Protocol

The technique of SAGE uses multiple enzymatic, PCR amplification, purification, and cloning steps, but it rests on two basic principles.15 The first principle is that a short oligonucleotide sequence, defined by a specific restriction endonuclease (anchoring enzyme, AE) at a fixed distance from the poly(A) tail, can uniquely identify mRNA transcripts (Figure). Theoretically, a 10-bp sequence tag can give 4^10 (1,048,576) different sequence combinations, which is overtly sufficient to discriminate all the transcripts derived from the human genome.16,17 The second principle is that the end-to-end concatenation of these short oligonucleotides allows multiple transcript detection per sequencing reaction.

The SAGE protocol starts with the purification of mRNA bound to solid phase oligo(dT) magnetic beads. The cDNA is synthesized directly on the oligo(dT) bead and then digested with the anchoring enzyme (AE) NlaIII to reveal the 3’-most restriction site anchored to the oligo(dT) bead (Figure). Most SAGE experiments have used the 4-bp recognition site anchoring enzyme NlaIII, predicted to occur every 256 bp and thus present on most mRNA species. However, creating a second SAGE library with a different anchoring enzyme may be useful for detecting transcripts without a NlaIII site and also for reconfirming transcript identity in those with both anchoring restriction sites. This may significantly lessen the work associated with data analysis, but the marginal utility of such an approach remains to be demonstrated.

Next, the sample is equally divided into two separate tubes and ligated to two different linkers, A or B. Both linkers contain the recognition site for BsmFI, a type IIS restriction enzyme that cuts 10-bp 3’ from the anchoring enzyme recognition site. BsmFI generates a unique oligonucleotide known as the SAGE tag, hence called the tagging enzyme (TE). The SAGE tags released from the oligo(dT) beads are then separated, blunted, and ligated to each other to give rise to ditags. The ditags are PCR amplified, released from the linkers, gel purified, serially ligated, cloned, and sequenced using an automated sequencer.
Modifications of SAGE

The 10-bp tag is sufficient to identify a less complex list of expressed genes, but it cannot definitively identify all genes in the unannotated human genome. The “LongSAGE” modification uses a different type IIS restriction endonuclease Mmel as the tagging enzyme that cuts 17-bp 3′ from the anchoring site. The theoretical LongSAGE tag uniqueness probability is greater than 99%, assuming the genome contains 30×10^6 NlaIII-derived tags, enabling direct matching of tags to the unannotated human genome.

A number of other technical modifications have been introduced to optimize the SAGE technique. Improved efficiencies of library construction requiring less mRNA have been reported using “SAGE-Lite” and “MicroSAGE.” The former technique uses a preliminary cDNA PCR step that could potentially introduce an amplification bias, whereas the latter modification has formed the basis of performing the standard enzymatic steps on mRNA attached to a solid phase. Other modifications have improved cDNA synthesis by using more efficient polymerases, minimized contaminants that inhibit ditag formation, release, and concatenation by adding purification steps, and improved the screening of concatemer inserts.

Because the transcripts are anchored to oligo(dT) beads, potential internal poly(A) priming has been addressed in a recent study. Contaminating genomic DNA fragments containing poly(A) stretches can be eliminated by DNase pretreatment but RNA species could create spurious tags. Whether oligo(dT) primers in solution or on solid phase magnetic beads have similar internal priming potential is not clear. In practice, this issue has not been a significant factor for the numerous SAGE projects performed to date.

SAGE Data Analysis and Followup Strategies

The sequence files generated by the automated sequencer are analyzed using the SAGE2000 software. The three steps involved in obtaining a differential gene expression list are as follows: (1) deciphering the SAGE tags from the sequence data files by using the SAGE2000 software for extracting ditags and checking for duplicate ditags; (2) downloading a reference sequence database from the NCBI Web site (SAGEmap, www.ncbi.nlm.nih.gov); and (3) associating the tags to the expressed gene database. The relative transcript abundance can then be calculated by dividing the unique tag count by the total tags sequenced, and the fold change can be determined by the ratio of tags between libraries.

The initial analysis is usually limited to a predefined tag ratio of greater than 5-fold and a value of P≤0.05. The latter is based on a rigorously derived significance test for digital gene expression in which the rates of false-positives associated with different probability values have been computed by Monte-Carlo simulation. Using this test, the rate of false-positives has been demonstrated to have good behavior, validating the obtained confidence intervals. Depending on the preliminary results, the SAGE data can be reanalyzed by varying the P values and the fold-change thresholds.

Applications to Cardiovascular Biology

The SAGE technique has been extensively used for the genetic analyses of various types of cancers consistent with its conception in an oncology laboratory. It has been used to create a Tumor Gene Index, an archived database of SAGE tags from many different types of cancers or tissues, on the Cancer Genome Anatomy Project (CGAP) Web site (cgap.nci.nih.gov). From these comprehensive transcriptional analyses of various cell types under different conditions or treatments, new genetic insights about cancer cells have emerged. These include finding closer transcriptional similarities than anticipated between normal and cancer cells of the same tissue origin, cancer stage-specific differences in transcriptional profile and transcriptional reprogramming in cells placed outside their native environment.

Recently, there has been a literal explosion of gene expression studies in cardiovascular biology, with many using the more readily available microarray technique. There are a number of large cardiovascular genomics projects actively in progress across the United States using microarray technology, such as the CardioGenomics (www.cardiogenomics.org) and PhysGen (www.brc.mcw.edu) projects. In contrast, the number of studies using SAGE in the cardiovascular system is rather limited. The Table lists some of the representative studies on the major cardiovascular cellular elements using the SAGE technique. Most of these
SAGE Application to Cardiovascular Systems and Diseases

<table>
<thead>
<tr>
<th>Cell Types and Conditions</th>
<th>Total Tags/Unique Tags</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUVECs treated with oxidized LDL&lt;sup&gt;36&lt;/sup&gt;</td>
<td>6148–6573/5448</td>
<td>~1% of the genes with at least 5-fold change; most were known genes: IL-8, MCP-1, VCAM-1, PAI-1, Gro-α, Gro-β, E-selectin, and activin βA</td>
</tr>
<tr>
<td>Human monocytes induced by GM-CSF or M-CSF&lt;sup&gt;36&lt;/sup&gt;</td>
<td>170879/35037</td>
<td>Differentiation from monocytes to macrophages revealed high expression of lipid metabolism genes; possible link to atherosclerosis</td>
</tr>
<tr>
<td>Human tumor vs normal endothelial cells&lt;sup&gt;33&lt;/sup&gt;</td>
<td>193000/32500</td>
<td>79 transcripts were differentially expressed, including 46 elevated in tumor endothelial cells; novel neovascularization genes were found</td>
</tr>
<tr>
<td>Hyperlipidemic ApoE3-Leiden mouse liver compared to WT&lt;sup&gt;36&lt;/sup&gt;</td>
<td>18861–19219/9427</td>
<td>44% of tags corresponded to known genes; lipid metabolism genes elevated in hyperlipidemic mice</td>
</tr>
<tr>
<td>Cultured human cardiac cells under hypoxia or with Ad2/HIF-1α/VP16 transduction&lt;sup&gt;42&lt;/sup&gt;</td>
<td>20986–30579/7013–9943</td>
<td>Hypoxia regulated genes associated with transcription, protein turnover, glycolysis, and cell survival; partial overlap of the hypoxia gene profile with Ad2/HIF-1α/VP16 transduction</td>
</tr>
<tr>
<td>Adult mouse heart transcriptome&lt;sup&gt;12&lt;/sup&gt;</td>
<td>88860/23941</td>
<td>Over 66% of the tags matched known genes; 18.7% of total were mitochondrial transcripts. Direct comparison to microarray data showed that SAGE was more quantitative.</td>
</tr>
</tbody>
</table>

Ad2/HIF-1α/VP16 indicates adenovirus vector with the hypoxia-inducible factor 1α fused to the viral transcriptional activator VP-16; COX1, cytochrome C oxidase polypeptide 1; HUVECs, human umbilical vein endothelial cells; and WT, wild type.

Studies have been descriptive in nature, and it remains to be seen how some of the observed candidate genes will be used to elucidate basic mechanisms of disease pathogenesis or applied for diagnostic and therapeutic purposes.

As mentioned earlier, one of the major strengths of SAGE is the electronic nature of the database, allowing direct comparisons of libraries in silico by different investigators. For example, a normal human heart SAGE library is available on the CGAP Web site for gene expression queries, and a normal adult mouse heart SAGE library gene expression profile has recently been reported.<sup>12</sup> Therefore, if both heart SAGE library data were available on an internet platform similar to the CGAP Web site, it may be possible for investigators to determine species similarities or differences in heart gene expression profiles. Because there is no such SAGE cardiovascular Web site, in some instances individual authors have made their SAGE tags available for download and analysis.<sup>36</sup>

There are a number of areas in cardiovascular biology where the SAGE technique may be useful. These areas include stem cell biology, cardiovascular development, angiogenesis, atherosclerosis, and lipid regulation. Some exploratory SAGE studies have already been reported for human hematopoietic stem cells, hyperlipidemic ApoE3-Leiden mice, and endothelial cells exposed to atherogenic stimulus.<sup>37,38</sup> In the future, the SAGE technique could assist in finding new targets of important transcriptional factors such as Nkx2-5 in cardiogenesis, where the number of cells may be limiting. With the burgeoning population of congestive heart failure (CHF) patients, more insights are needed into our basic understanding of the pathogenetic mechanisms of CHF. Potentially, SAGE libraries could be made from human endomyocardial biopsy specimens, but tissue heterogeneity may undermine the gene expression signals. It may be more informative to study the temporal changes in gene expression using controlled animal models of CHF where more tissue material is available for processing. A refined candidate gene list could then be used in the diagnosis and prognosis of larger numbers of patient samples in a microarray format.

What other avenues exist for identifying novel human cardiovascular disease genes using gene expression techniques? A recently reported SAGE study examining photoreceptor-specific genes in the mouse retina may represent one such approach.<sup>39</sup> The combination of two different genetic data sets, familial disease gene chromosomal loci and transcripts expressed in the diseased tissue, may assist in defining the genes or regions of the chromosome for mutational analysis. This type of genetic analysis may be amenable to familial cardiovascular syndromes, such as arrhythmogenic right ventricular dysplasia, where a number of different chromosomal loci have been implicated as containing the candidate disease gene.<sup>41</sup>

Cardiovascular investigators are now in an exciting but challenging phase of research in the post-genomic period that requires the integration of various types of experimental approaches. There are several different gene expression profiling methods available for helping elucidate the pathogenesis of cardiovascular disorders, and there are numerous factors to consider before selecting the appropriate techniques. SAGE is a powerful technique that has not been utilized to its potential in studying cardiovascular biology. However, there are a number of successful SAGE projects that may serve as a guide for those interested in cardiovascular problems. Careful planning of the SAGE library construction and developing a predefined screening assay for the analysis of the expression data are essential for taking full advantage of the gene expression experiment. The establishment of an electronic database of cardiovascular SAGE libraries may facilitate efficient data mining by researchers with new insights in the future. Ultimately, the true measure of success of cardiovascular gene expression studies will be the impact of their discovery on helping focus research on the development of new diagnostic, prognostic, and therapeutic tools for the prevention and treatment of cardiovascular diseases in patients.
Acknowledgments

The authors are supported by the NHLBI-NIH Intramural Program. The corresponding author wishes to thank V.E. Velculescu and K.W. Kinzler for many helpful discussions regarding the SAGE technique. We also wish to thank T. Finkel for his guidance in writing this review.

References

Serial Analysis of Gene Expression: Technical Considerations and Applications to Cardiovascular Biology
Willmar D. Patino, Omar Y. Mian and Paul M. Hwang

Circ Res. 2002;91:565-569
doi: 10.1161/01.RES.0000036018.76903.18
Circulation Research is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2002 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7330. Online ISSN: 1524-4571

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circres.ahajournals.org/content/91/7/565

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation Research can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation Research is online at:
http://circres.ahajournals.org/subscriptions/