Serial analysis of gene expression (SAGE): its modification and application in human disease

S. Abdolhamid Angaji*, Ghazal Aminian, Bahram Khosravi, Masoomeh Taran

Department of Cell and Molecular Biology, Faculty Biological Sciences, Karazmi University, Tehran, Iran

Key words: serial analysis of gene expression (SAGE), microarray, Differential gene expression, human diseases.

Abstract

Among experimental techniques as regards providing gene expression profiles, serial analysis of gene expression (SAGE) yields quantitative, comprehensive data in the cell being studied. Short fragments of genetics information corresponding to those expressed genes of interest form the bedrock of SAGE method, in which these short sequences consisting of 9-10 bp, called SAGE tags, are bound with one another in order for sequencing to be efficient. Unlike the other techniques, requiring no prior knowledge of genes in addition to preparing both the qualitative and quantitative levels of gene expression make SAGE become singled out in fact. Considering different conditions so as to improve the efficiency of this technique, various modifications have been introduced to, including long SAGE, micro SAGE, super SAGE, to name a few. In this review, we have strived to present some new findings as to SAGE technology, along with discussing its practical applications to common diseases, peculiarly genetic disorders.

*Corresponding Author: Abdolhamid Angaji  Angaji@khu.ac.ir

Abbreviation

DGE, Differential gene expression; SAGE, serial analysis of gene expression; TOGA, total gene expression analysis; READS, restriction enzyme analysis of differentially expressed sequences; VDR, vitamin D receptor; CYP A, cyclophilin A.
Introduction
In recent years, several methods of gene expression have been discovered. Differential gene expression (DGE) technologies have been utilized in great many biological challenges as of their development in the early 1990s. DGE applications have greatly promoted the landscape of basic biological research, together with the discovery and development of pharmaceuticals. DGE is broadly divided into two families of technologies: closed architecture systems and open architecture systems (Geen et al., 2001). Open system methods analyze the expression of genes which are unknown or there is only limited previous knowledge of their genome and therefore these methods are useful for recognizing limited number of the expressed genes. With open system methods discovery of many unknown genes is probable with no need for a priori comprehensive knowledge of the transcriptome sprang into existence, where the field of discovery is ‘Open’. GeneCalling, SAGE, TOGA, READS are four examples of open systems in fact (Geen et al., 2001). In contrast to open system, the closed system analyses are quite limited owing to the fact that genes ought to be known since completeness of our knowledge concerning genomes defines their coverage and the more complete knowledge of genome, the greater would be its analysis of the transcriptome. Of closed systems, Oligonucleotide or cDNA array hybridization technologies, and quantitative polymerase chain reaction (qPCR, or TaqMan) are among the most commonly used, also beads array for the detection of gene expression and microarray are examples of closed system methods (Marshall and Hodgson, 1998; Heidet al, 1996). Although closed systems are useful to identify genes in a good way and work well for particular light novel genes, they are limited by the defect of the transcriptome knowledge. For example in the known genome such as human genome, the transcriptome would be more complex because of RNA editing, splice variants and polymorphisms. But if there is complete information, closed systems would provide a desirable knowledge of genome. In open system if a given transcriptome is complex, it won’t change the quality of expression analysis and won’t preclude the discovery, but technically, the capacity of open systems is limited. In situations where complete knowledge of a transcriptome exists, they cannot really compete with closed architecture systems in terms of time and labor intensity. Even an open system that theoretically permits analysis of a whole genome is limited by the sensitivity of the method in detecting rare transcripts and splice variants (Lorkowski and Cullen, 2004). Having been touched upon, open systems might not depend upon existing transcriptome information, however exploiting existing expressed genome databases makes considerably these systems effective according to the both facilitating the analysis process as well as improving the efficiency of known and novel gene identification.

In identifying the new genes which require planned works, both closed and open systems are needed to achieve this aim. The unknown genes discover with open systems, then closed systems use the resulted information and can ended up in such discoveries as has been applied to recent genomes.

In this review we discuss the SAGE as an open system which requires no prior knowledge on expressed sequence. It in fact allows the identification of novel genes by using the SAGE tag as a specific polymerase chain reaction primer to amplify the unknown cDNA (van den Berg, 1999).

Indeed, Velculesco, Vogelstein, and Kinzler devised, serial analysis of gene expression (SAGE) method in 1995 at Johns Hopkins University (Velculescu, 1995). Higher organisms such as humans possess a long, massive genome constituting an enormous number of genes whose sequence determination appears to be an attainable objective these days. A highly efficacious technology, SAGE, provides a global expression profile of a specific cell’s or tissue’s genes, Furthermore, it is useful in discerning particular genes according to cellular conditions through the comparison of gene expression profiles generated for a pair of cells stored in different conditions (Bartlett, 2001; Polyak et al, 1997; Yamamoto et al, 2001).
Three Founding Principles
Having derived from a specific location within transcripts, SAGE tags, 10-11 base pair (bp), contiguous, though unique sequences, form the first stage in SAGE so as to identify mRNAs. Not only SAGE tags utilized in gene identification, but they are of invaluable help to measure the relative abundance of their cognate transcripts in the mRNA population grounded on their occurrence rate.

Throughput limitation with respect to sequencing-based approaches seems to be a serious challenge ought to be met. SAGE at the second step, hence, uses serial processing where 20-25 tags are analyzed on each lane of an automated DNA sequence while pursuing the parallel process.

Finally, SAGE similar to other contemporary expression analysis methods, uses polymerase chain reaction (PCR) amplification, yet this technique is unique by means of recognizing, along with eliminating PCR-mediated amplification bias from the profile. While convoluted natural mixtures of mRNA are analyzed, transcript representation is individually maintained (Neilson et al, 2000).

SAGE Methodological Outline
SAGE, basically, consists of several steps which are described briefly as follows; (Aldaz, 2003; Patino et al, 2002; Porteret al, 2006; Sharma et al, 2007). To synthesize a double-stranded cDNA, mRNA and a biotinylatedoligo (dT) primer is required. Then the cDNA ought to be cleaved by dint of a restriction enzyme (also called anchoring enzyme). A wide range of four-base recognizing enzymes can be utilized since they can cut every 256 bp (4^4) on cDNA as mRNAs are mostly regarded as much longer indeed. Among restriction enzymes, Nla III is used most frequently as an anchoring enzyme, which cleaves the cDNA. The 3’-terminus of the cleaved cDNA is then bound to streptavidin-coated beads for recovery, which after dividing the reaction mixtures into two portions, are ligated to two independent linkers containing recognition sites for class II restriction endonuclease (also called tagging enzyme, usuallyFok1 and BsmFI).

The tagging enzyme cuts the cDNA at a constant number up to 20bp away from their asymmetric recognition sites, resulting in a release of a short piece of cDNA(tag) plus the linker from the beads. The 3’-ends of the released tags attached to linkers are, at the next stage, blunt ended by dint of T4 DNA polymerase. Next, the two pools of the released tags mixed again to become ligated to each other. The mRNA-derived termini are solely able to get ligated in a tail-to-tail orientation to make ditagsas the linkers at the 5’-ends are blocked by amino groups. Ditag fragments flanking the both termini with NlsIII cohesive ends, after isolation, are ligated so as to acquire concatemers, which afterwards are recovered by polyacrylamide gel electrophoresis (PAGE) to get cloned into plasmid vectors (Figure 1).

Fig. 1. Schematic SAGE protocol outline. The anchoring enzymes (AE) is NlaIIIand tagging enzyme (TE) is BsmFI. Linkers A and B are independent and their 3’ end contain TE sequence.

Advantages
Unlike microarray-based expression profiling, the SAGE provides the detection capacity of novel transcripts from the previously uncharacterized genes, which makes it stand out for diagnostic use indeed. What makes it become way noticeable compared to other platforms is owing to the fact that SAGE can reveal which genes are expressed and their level of expression rather than merely quantifying the expression level of predetermined and presently incomplete, set of genes as carried out by closed system gene expression profiling platforms such as
microarray. These distinguishing attributes enable SAGE to be used as a primary discovery engine that can characterize human diseases at the molecular level while clarifying potential targets and markers for therapeutic and diagnostic purposes, respectively.

Gathering cumulative data, which can be continued at any time, is the other merit in the application of SAGE enabling a more detailed picture of the expression pattern (Hermeking, 2003; Madden et al., 2000).

Disadvantages
Requiring a great deal of money and time are two major drawbacks of SAGE since thousands of polymerase chain reactions should be performed, for example virtually 10,000 reactions are needed for a comparison of two different states, in which each reaction is represented by 50,000 SAGE tags (Hermeking, 2003).

Both SAGE and microarray analysis have limits when applied to the identification of tumor markers. One obvious limit is the fact that differences in mRNA levels detected by these methods may not always be reflected at the protein level. Because proteins represent the targets used for either diagnostic or therapeutic approaches in most cases, confirmation of differential expression at the protein level is required for each candidate transcript. Automated proteome analysis using mass spectrometry has significantly progressed in recent years, suggesting that it may soon be used widely to identify proteins that are specifically expressed in tumors. However, at present, a comprehensive picture of human gene expression can only be achieved with RNA-based techniques (Hermeking, 2003).

To solve such problems, various SAGE modifications have thus far been applied.

**MiniSAGE**
This method was devised by Ye et al. in 2000 followed by an additional PCR amplification, requires merely 1 µg of total RNA, which means hundreds of times less than the amount needed for the original SAGE protocol (Lorkowski and Cullen, 2004).

**MicroSAGE and SAGE-lite**
The other modified methods requiring far much less starting input of RNA are MicroSAGE, together with SAGE-lite with 500-5000fold less input for the former (Table 1), and less than 100 ng for the latter in fact (Peters et al., 1999). As regards microSAGE, a ‘single tube’ procedure is utilized for all the steps through SAGE until the tags are released by digestion with the tagging enzyme. MicroSAGE, furthermore, requires fewer cycles of PCR to generate sufficient ditags, denoting the less input to profile gene expression. Therefore, concerning heterogeneous tissues, microdissection should be applied with microSAGE in order to build region-specific expression profiles. Moreover, profiling in cases of tissue scarcity, biopsy, and tumor metastasis are of the other usages of this method.

<table>
<thead>
<tr>
<th>Differences</th>
<th>SAGE</th>
<th>microSAGE</th>
<th>miniSAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount of incoming materials</td>
<td>2.5 to 5 µg of RNA</td>
<td>1-5 ng of RNA</td>
<td>1 µg of RNA</td>
</tr>
<tr>
<td>Capture of cDNA</td>
<td>streptavidin-coated magnetic beads</td>
<td>magnetic streptavidin-coated PCR tube</td>
<td>streptavidin-coated magnetic beads</td>
</tr>
<tr>
<td>The number of tube reaction</td>
<td>Successive reaction performed in multiple tubes</td>
<td>Single tube reaction</td>
<td>Single tube reaction</td>
</tr>
<tr>
<td>Number of PCR cycle</td>
<td>25-28 of PCR cycle</td>
<td>a limited number of re-PCR cycle does not require the additional PCR amplifications</td>
<td></td>
</tr>
</tbody>
</table>

**LongSAGE**
LongSAGE makes longer tags (21bp) since it utilizes type II restriction enzyme MmeI. The tags show their equivalent gene expression; therefore the longer the tags are, the more dependable the statistical analysis of gene expression is(Fig 2). Having said that,
however, there are more single nucleotide polymorphisms (SNPs) present in longer tags, which make a huge decline in the uniformity of tags in SAGE set. In addition, using longer tags costs heavily more, thereby circumscribing its applications (Porter et al., 2006).

**SuperSAGE**

The major difference between original SAGE and SuperSAGE is the fact that restriction enzymes which are used are different, BsmF1 and EcoP15I respectively. EcoP15I generates 26bp tags from transcripts (Fig. 2), where these tags in fact make superSAGE become more accurate, conferring the ability to profile plant gene expression, along with identification and quantification of new and known genes (Meisel et al., 1992; Matsumura, 2006).

**Studies made by the use of SAGE**

Significant studies performed by SAGE were summarized in Table 2.

<table>
<thead>
<tr>
<th>Cell, tissue</th>
<th>Total tags</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retina of eye</td>
<td>267,332</td>
<td>Dror Sharon et al., 2001.</td>
</tr>
<tr>
<td>VGP and MGP Melanoma Cell Lines</td>
<td>29,150 and 37,710 in two VGP melanoma SAGE libraries (DB1; DB2) respectively. 19,908 and 961 in two MGP melanoma SAGE libraries (DB3; DB4) respectively.</td>
<td>Smith et al., 2004.</td>
</tr>
<tr>
<td>Lobular carcinoma</td>
<td>67,834</td>
<td>Honda, et al., 2008.</td>
</tr>
<tr>
<td>Endometrial epithelial and stromal cells</td>
<td>19,493 tags in Normal endometrium 80,783 tags in endom1 77,846 tags in endom2 76,100 tags in FbEM-1</td>
<td>Honda, et al., 2008.</td>
</tr>
<tr>
<td>Normal and neoplastic plasma cells</td>
<td>29,918 SAGE tags from normal and 10,340 tags from multiple myeloma tumor</td>
<td>Felix et al., 2009.</td>
</tr>
<tr>
<td>Colorectal Tumor Cells</td>
<td>26,060 unique tags</td>
<td>Meeh, et al., 2009.</td>
</tr>
<tr>
<td>Human breast tissue</td>
<td>14,430</td>
<td>Sakamoto et al., 2010.</td>
</tr>
<tr>
<td>Human breast tissue</td>
<td>Different SAGE-Seq libraries with different total tags</td>
<td>Wu et al., 2010.</td>
</tr>
<tr>
<td>Human breast tissue</td>
<td>2,931,124</td>
<td>Romanuik et al., 2010.</td>
</tr>
<tr>
<td>Striatum</td>
<td>32,216 Uniquestrial tags</td>
<td>Mazarei et al., 2010.</td>
</tr>
<tr>
<td>IPF and normal lung tissue</td>
<td>A total of 954,932 transcript tags were sequenced of which 168,272 were unique</td>
<td>Boon et al., 2011.</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>A total of 100,666 tags were generated including 45,560 unique tags</td>
<td>Fan et al., 2011.</td>
</tr>
<tr>
<td>The analysis of normal and adenoma libraries</td>
<td>21,030 and 22,329 unique SAGE tags in normal and adenoma libraries identified respectively</td>
<td>De Lima et al., 2012.</td>
</tr>
<tr>
<td>Peripheral Blood Cells</td>
<td>11 million tags</td>
<td>Mastrokalias et al., 2012.</td>
</tr>
<tr>
<td>Liver flake-associated Cholangiocarcinoma (CCA)</td>
<td>Three libraries: CCA-K1 (60,319 tags), CCA-K2D Sawanyawisuth et al., (46,853 tags) and normal liver (66,308 tags)</td>
<td>Sawanyawisuth et al., 2012.</td>
</tr>
<tr>
<td>Human Neural Tube Fragments</td>
<td>269,043</td>
<td>Krupp et al., 2012.</td>
</tr>
<tr>
<td>Breast Carcinoma</td>
<td>66,128 tags in the breast ductal carcinoma library and 50,512 tags in normal epithelium</td>
<td>Soloviev et al., 2013.</td>
</tr>
</tbody>
</table>

**SAGE in human studies**

SAGE technology has been used to collect many transcripts from a tissue sample, to compare the profile of mRNA expression in different types of cancer cells and to identify genes that are under the new regulation in response to exposure to external stimuli. It also has been widely used in a number of human studies. The following parts are some examples of these applications:

**Ocular studies**

A new study describes SAGE libraries constructed from the human peripheral retina, macula, and retinal pigment epithelium (RPE). They compared the
expression profile of the retinal periphery versus the macula of the same eye and the expression profile of the peripheral retina from two individuals (ages 44 and 88 years) then evaluated the potential value of the library as a resource for identifying candidate genes for hereditary retinal diseases (Sharon et al., 2002).

The cone photoreceptor contribution to all transcription in the retina was found to be similar in the macula versus the retinal periphery, whereas the rod contribution was greater in the periphery versus the macula. Genes encoding structural proteins for axons were found to be expressed at higher levels in the macula versus the retinal periphery, probably reflecting the large proportion of ganglion cells in the central retina. In comparison with the younger eye, the peripheral retina of the older eye had a substantially higher proportion of mRNAs from genes encoding proteins involved in iron metabolism or protection against oxidative damage and a substantially lower proportion of mRNAs from genes encoding proteins involved in rod photo transduction. The RPE library had numerous previously unencountered tags, suggesting that this cell type has a large, idiosyncratic repertoire of expressed genes (Sharon et al., 2002).

**Fig. 2.** Simplified scheme of SAGE, LongSAGE and SuperSAGE.

Comparison of these libraries with 100 reported nonocular SAGE libraries revealed 89 retina-specific or enriched genes expressed at substantial levels, of which 14 are known to cause a retinal disease and 53 are RPE-specific genes. We expect that these libraries will serve as a resource for understanding the relative expression levels of genes in the retina and the RPE and for identifying additional disease genes (Sharon et al., 2002).

**Down syndrome (DS)**

Down syndrome is a congenital abnormality caused by the presence of an additional chromosome 21. The most important of Down syndrome abnormalities are mental retardation, heightened risk of Alzheimer’s disease, increased occurrence of leukemia, immunity and heart defects, and muscle hypotonia (Epstein et al., 1991). It is assumed that a dosage imbalance of human chromosome 21 (HSA21) bring about the Down syndrome, also, Mao et al. in their work illustrated that there was an up regulation of HSA21 genes in DS fetal brain (Mao et al., 2003). By the means of SAGE technology, it is possible to create an expression profile of presumed genes involved with the generation and development of DS.

In an interesting study, Sommer and co-workers identified dysregulated genes in lymphocytes from children with Down syndrome (Sommer et al., 2008). Two SAGE libraries were constructed using pooled RNA of normal and Down syndrome children. Comparison between DS and normal profiles revealed that most of the transcripts were expressed at similar levels and functional classes of abundant genes were equally represented. Among the 242 significantly differentially expressed SAGE tags, several transcripts down-regulated in DS code for proteins involved in T-cell and B-cell receptor signaling. These results indicate that trisomy 21 induces a modest dysregulation of disomic genes that may be related to the immunological perturbations seen in DS (Malagó et al., 2005).

In the other work, gene expression profile of human Down syndrome leukocytes is prepared with SAGE to identify the differences in the gene expression patterns of Down syndrome and normal leukocytes. The generated data can be compared with that of other SAGE libraries constructed from different Down syndrome tissues in order to study the differences specific to temporal and spatial gene expression. The resulted library is currently being
used to analyze the expression profile of Down syndrome leukemic patients in an attempt to clarify the relationship between trisomy 21 and the increased risk of leukemia observed in Down syndrome and is useful in understanding the molecular mechanism of Down syndrome pathology (Malagó et al, 2005).

Parkinson disease (PD)
Parkinson disease (PD) is a common neurodegenerative disease that leads to shaking, difficulty with walking and movement and muscle rigidity and weakness. Some mutations like α-synuclein and GIGYF2 have been elucidated in PD families which resulted that this disease is a polygenic disorder (Polymeropoulos et al, 1997; Lautier et al, 2008).

To investigation about the polygenic property of PD, by combining genomic convergence with the data generating capacity of SAGE, it demonstrated that a total number of 50 transcripts were found to be directly associated with genomic convergence peaks, thus establishing these genes as high priority candidates for PD risk (Noureddine et al, 2005). In recent studies several SAGE transcripts have been discovered which have association with PD and include the ring finger protein-11 (RNF11) gene, the glial fibrillary acidic protein (GFAP) gene and the chemokine (C C motif) ligand- 2 (CCL2) gene .The study by Noureddenne et al. also found that with the SAGE technique, all UNIGENE sets in linkage regions are detectable and also SAGE was useful in detecting a novel missense polymorphism (A5390G, Ile304Met) in the NADH dehydrogenase subunit 3 (ND3) gene which are seen in the substantia nigra of Parkinsonian brains. This polymorphism make an NlaIII site at the 3’ end of the ND3 gene, then it is easy to utilize in SAGE tag isolation(Anderson et al, 2007; Aponso et al, 2008; Reale et al, 2009).

The recent findings that RNF11, GFAP and CCL2 are associated with PD, illustrate the capacity of SAGE to create candidate genes for finding disease-association elements and use in functional surveys (Sommer et al, 2008; Malagó et al, 2005).

Cancer study
Cancer is a board range of diseases that has unregulated cell growth. In cancer cells grow with no control and divide for several times and invading to the other parts of body and forming malignant tumors. There are more than 200 different types of cancer which affect humans.

Cholangiocarcinoma (CCA) is the most common liver cancer where it’s occurrence rate is so high (Shinet al, 2010). The progression of cancer cells toward malignancy and metastasis, including different events such as changes in gene expression can be studied by SAGE. In CCA tumor different alteration of gene expression happened. These include of TMSB10, GAL3, VDR, CYP450 and some others. Among these genes, vitamin D receptor (VDR) is important because vitamin D3 [1,25(OH)2D3], has recently been used for prevention of human tumors and it has antitumor activity (Lamprecht and Lipkin, 2003; Kumagai et al, 2003). Sawanyawisuth and his co-workers constructed SAGE libraries of CCA and then compared to the publicly available SAGE library of normal liver tissue by using the SAGE DGED online tool on the CGAP website they identified 509 gene by SAGE that have different expression in CCA, 142 of these were up-regulated and 367 down-regulated compare to normal liver tissues (Sawanyawisuth et al, 2012). In vitro, vitamin D3 induces cell cycle arrest, differentiation, and apoptosis of cancer cells (Liu et al, 1996; Danielsson et al, 1999; Diazet al, 2000; Pålmer et al, 2001). The VDR belongs to the steroid/thyroid hormone nuclear receptor superfamily. VDR up-regulation has been shown in primary tumor tissues of breast, colon and pancreatic cancers (Friedrich et al, 1998). Accuracy of the VDR expression pattern corresponds to SAGE analysis. VDR was rarely expressed in normal bile duct epithelia but highly expressed in 74% of CCA tissues. The survival rate of CCA patients with positive VDR expression in tumor tissue was significantly better than that of patients with negative expression of VDR. Indeed, 1,25(OH)2D3 therapy, in the CCA cell lines with high expression of VDR significantly decrease cell proliferation in a dose-
dependent manner. The effect was not shown in lower VDR expressing CCA cell lines (Seubwai et al., 2007). These conclusion are according to those seen in carcinomas of gastric, breast and colon (Albrechtsson et al., 2003, Banerjee and Chatterjee, 2003; Pelczynska et al., 2005). These findings suggest that supplementation of 1,25 (OH)2D3 or its analogs may be a potential strategy for long-term control of tumor development and progression in CCA patients. These data encourage further investigation of 1,25 (OH)2D3 or its analogues as therapeutic agents in the treatment of CCA patients (Seubwai et al., 2010).

References


Dror SH, Seth B, Constance LC, Thaddeus PD. 2001. Profile of the genes expressed in the human peripheral retina, macula, and retinal pigment
http://dx.doi.org/10.1073/pnas.012582799.

http://dx.doi.org/10.1101/gr.6.10.986.


http://dx.doi.org/10.1002/bdra.23040.

http://dx.doi.org/10.1093/jnci/95.12.896.

http://dx.doi.org/10.1038/nrc1144.


