ELSEVIER

Journal of Immunological Methods 250 (2001) 45–66

JIM
Journal of Immunological Methods

www.elsevier.nl/locate/jim

Use of serial analysis of gene expression (SAGE) technology

Mikio Yamamoto*, Toru Wakatsuki, Akiyuki Hada, Akihide Ryo[1]

Department of Biochemistry, National Defense Medical College, 3-2 Namiki, Tokorozawa, Saitama 359-8513, Japan

Abstract

Serial analysis of gene expression, or SAGE, is an experimental technique designed to gain a direct and quantitative measure of gene expression. The SAGE method is based on the isolation of unique sequence tags (9–10 bp in length) from individual mRNAs and concatenation of tags serially into long DNA molecules for a lump-sum sequencing. The SAGE method can be applied to the studies exploring virtually any kinds of biological phenomena in which the changes in cellular transcription are responsible. SAGE is a highly competent technology that can not only give a global gene expression profile of a particular type of cell or tissue, but also help us identify a set of specific genes to the cellular conditions by comparing the profiles constructed for a pair of cells that are kept at different conditions. In this review, we present an outline of the original method, several studies achieved by using the method as a major strategic tool, technological difficulties and intrinsic problems that emerged, and improvements and modifications of the method to cope with these drawbacks. We then present our modified SAGE procedure that generates longer sequence tags (14 bp) rather in detail, and the profile (80K profile) derived from HeLa cells that is composed of 80 000 tags obtained from a single library. In addition, a series of smaller profiles (2, 4, 10, 20 and 40K) was made by dividing the 80K profile. When we compared these smaller profiles with respect to tag counts for a number of genes, it became apparent that counts of most gene tags increase stably and constantly as the size of profiles increase, while several genes do not. This may be another problem we have to keep in mind, when the profiles are compared for the identification of 'specific genes'. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Serial analysis of gene expression (SAGE); Genomic sequence; Cellular transcription

1. Introduction

The genomic sequence of a wide variety of organisms, including that of humans, are being elucidated one after another. The genomes of eukaryotic organisms are long and massive, and contain an enormous number of genes. By delicately regulating activities of these genes, each organism can supply required amount of products at an appropriate time that confer functions proper to the organism. It is thus believed that the majority of biological phenomena found in a variety of organisms can be explained by the quantity of gene products. Although the gene function is certainly conducted by its final product, protein, there are a large number of observations that the amount of protein produced is directly dependent on the amount of mRNA that encodes it. This means that, to generally understand the cellular functions under the certain conditions at a certain time, it can be attained by measuring the species and respective numbers of mRNAs at a point of time. However,

*Corresponding author.

E-mail address: yamama13@cc.ndmc.ac.jp (M. Yamamoto).

[1] Present address: Cancer Biology Program, Hematology/Oncology Division, Harvard Institute of Medicine, 1047, BJDMC/Harvard Medical School, 330 Brookline Avenue, Boston, MA 02215, USA.

0022-1759/01/$ – see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0022-1759(01)00305-2

each cell contains more than 10 000 species, copies of each species ranging from less than one to more than 10 000, and, as a total, up to half a million mRNA transcript copies. It was therefore practically impossible to determine them. A feasible tactic was only to identify genes whose expression was influenced by a variety of internal or external factors. These were classical differential colony (plaque) hybridization of cDNA clones (Yamamoto et al., 1983), subtractive hybridization (Kavathas et al., 1984; Hubank and Schatz, 1994) and a more recent differential display method (Liang and Pardee, 1992; Welsh et al., 1992). A large-scale random cDNA sequencing by EST project was very useful for the identification of unknown genes expressed in given cells or tissues (Adams et al., 1991). However, this approach was not designed to quantify expressed genes, since the cDNA library to be sequenced was usually normalized to eliminate recurring transcripts derived from abundant class mRNA sequences for the purpose of expanding the size of the gene collection (Ko, 1990).

The body mapping project was the unique and direct attempt to construct gene expression profiles of a number of cells and tissues by random sequencing of a 3′-directed cDNA library (Okubo et al., 1992). About 300 bp fragments of these 3′-region were called gene signature and each represented a particular mRNA species. By sequencing 1000 or so cDNA clones, they could make a rough pattern of gene expression and identify mRNAs of highly abundant class. However, as an inevitable weakness common to both EST and body mapping projects, they include an inefficient sequencing step, in which one sequencing process yields only one cDNA sequence. Mainly because of this low throughput, the profiles obtained by the body mapping project unavoidably became a long way from what is expected and demanded. Although the more recent methods of hybridization-based analyses (DNA microarray) using immobilized cDNAs (Schena et al., 1995) or oligonucleotides (Lockhart et al., 1996) can potentially examine the expression patterns of a relatively large number of genes, the method can only examine expressed sequences that have already been identified.

In contrast, the SAGE method allows for a quantitative and simultaneous analysis of a large number of transcripts in any particular cells or tissues, without prior knowledge of the genes (Velculescu et al., 1995). As the body mapping procedure does, this method takes advantage of the 3′-portion of mRNA as the gene tag, but of much shorter form (9–10 bp). These tags can be serially connected before cloning into a plasmid vector. Since the resulting plasmid clones contain multiple tags, sequences of several dozens of mRNAs can be obtained by a single sequencing reaction. Rapid and cost-saving sequencing by this original device allows quantification and identification of a large number of cellular transcripts.

In this review, we present the principle and an outline of this powerful high-throughput original method, several studies achieved by using the method as a major strategic tool, technological difficulties and intrinsic problems that emerged, and technical improvements and modifications of the method to cope with these drawbacks. We then present our modified SAGE procedure that generates longer sequence tags (14 bp) in detail, and studies utilizing it.

2. The principle of SAGE and a methodological outline

SAGE is based mainly on two principles, representation of mRNAs (cDNAs) by short sequence tags and concatenation of these tags for cloning to allow the efficient sequencing analysis. Fig. 1 illustrates the scheme of the principle, in which the hypothetical eukaryotic cell that contains seven mRNA molecules composed of four species is depicted. If one wants to elucidate the gene expression profile of this particular cell, they would have to conduct several cDNA sequencing reactions. However, if each mRNA species can be represented by a short unique sequence stretch (such as 9 bp tag), the purpose would be attained by sequencing them, because a sequence stretch as short as 9 bp can distinguish 4^9 (262 144) transcripts, provided a random nucleotide distribution throughout the genome. This ability appears sufficient for the discrimination of all the human transcripts, because the human genome is estimated to encode between 28 642 and 153 478 genes (Pennisi, 2000). However, since current sequencing procedure handles one clone at a time, one

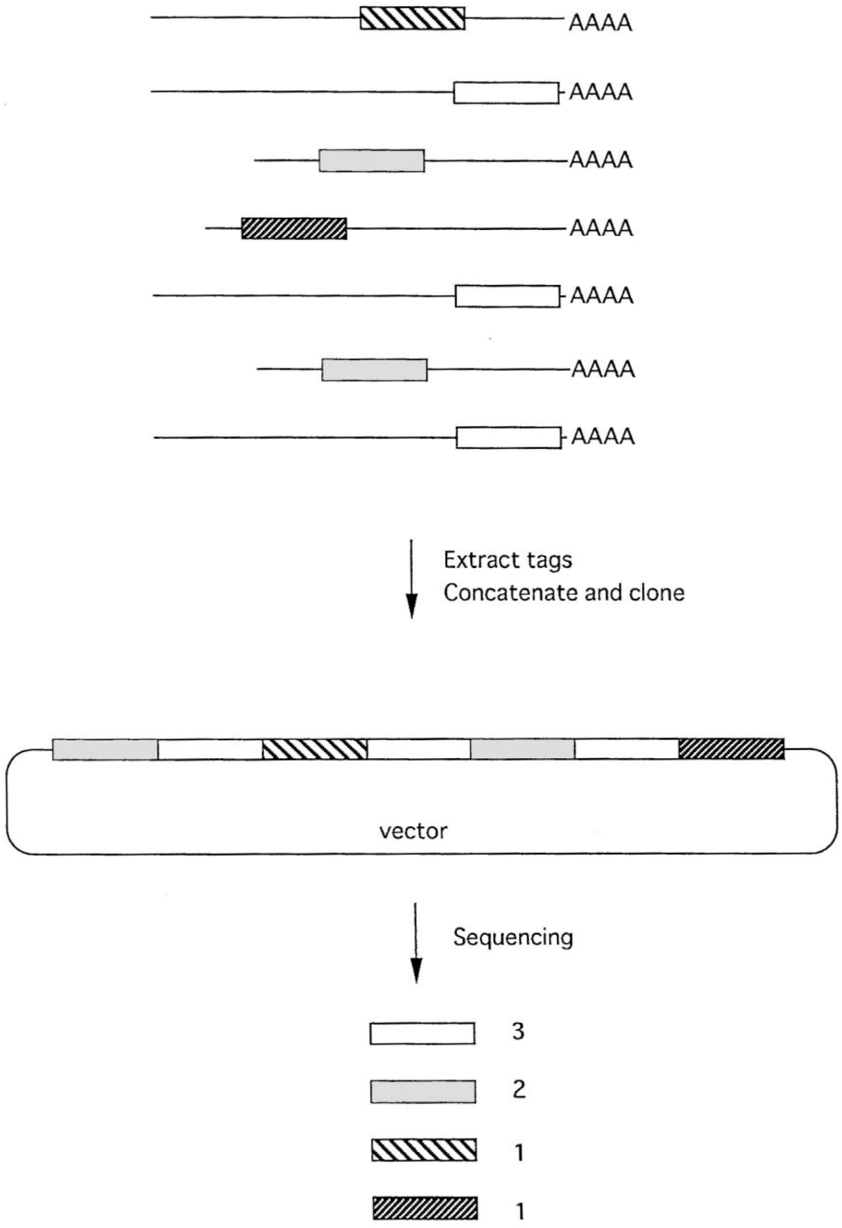

Fig. 1. The principle of SAGE. The hypothetical eukaryotic cell that contain seven mRNA molecules composed of four species is shown as a model. Boxed are tags that are proper to mRNA species.

has to conduct at least seven sequencing reactions for the profiling of this hypothetical cell. There is no particular merit by replacing mRNA with short sequence stretch, and this is the reason why the body mapping project fell into a setback despite its ideological importance. However, if we could connect these tags into a long stretch of DNA molecule, sequencing reaction would be needed only once. Since a currently-used automated DNA sequencer stably gives 5–600 nucleotides for any given clones, one would be able to obtain 50–60 9-bp tag-represented mRNA sequences by a single reaction and run. This is more than enough for the elucidation of gene expression profile of this hypothetical cell.

Fig. 2 shows a schematic presentation of SAGE procedure. Briefly, double-stranded cDNA is synthesized from mRNA by means of a biotinylated oligo(dT) primer. The cDNA is then cleaved with a restriction enzyme (called anchoring enzyme, AE in the figure). Any four-base recognizing enzymes may be used, because they cleave every 256 bp (4^4) on average, while the majority of mRNAs are considered to be much longer. Actually, *Nla*III, is the most frequently used enzyme. The 3′-most portion of the cleaved cDNA with a common *Nla*III cohesive end at its 5′-terminus is then recovered by binding to streptavidin-coated beads. After dividing the reaction mixture into two portions, two independent linkers are ligated using *Nla*III cohesive termini to each portion. These linkers are designed to contain type IIS enzyme (usually *Fok*I or *Bsm*FI, and designated as tagging enzyme, TE in the figure) site near (or

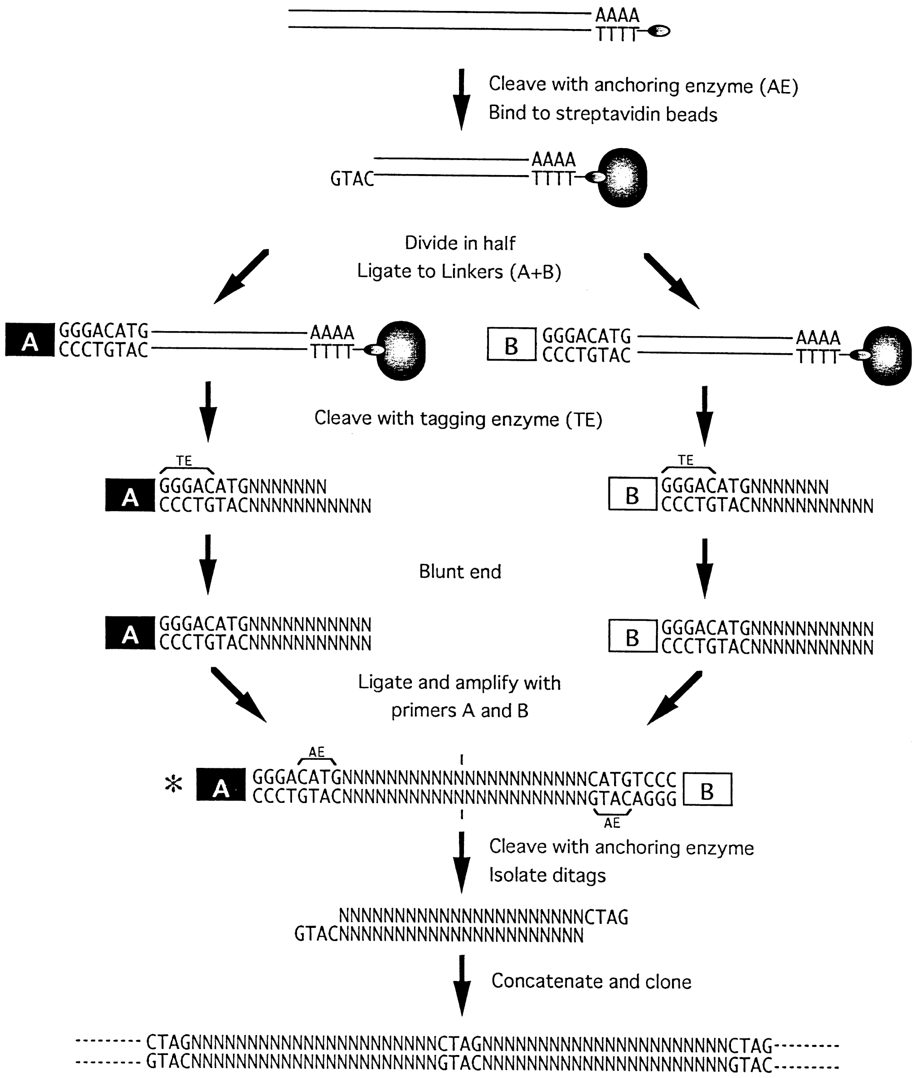
AAAA
TTTT
Cleave with anchoring enzyme (AE)
Bind to streptavidin beads
AAAA
GTAC
TTTT
Divide in half
Ligate to Linkers (A+B)
A
GGGACATG
CCCTGTAC
AAAA
TTTT
B
GGGACATG
CCCTGTAC
AAAA
TTTT
Cleave with tagging enzyme (TE)
TE
A
GGGACATGNNNNNNN
CCCTGTACNNNNNNNNNNN
TE
B
GGGACATGNNNNNNN
CCCTGTACNNNNNNNNNNN
Blunt end
A
GGGACATGNNNNNNNNNNN
CCCTGTACNNNNNNNNNNN
B
GGGACATGNNNNNNNNNNN
CCCTGTACNNNNNNNNNNN
Ligate and amplify with
primers A and B
AE
*
A
GGGACATGNNNNNNNNNNNNNNNNNNNNNNCATGTCCC
CCCTGTACNNNNNNNNNNNNNNNNNNNNNNGTACAGGG
B
AE
Cleave with anchoring enzyme
Isolate ditags
NNNNNNNNNNNNNNNNNNNNNNCTAG
GTACNNNNNNNNNNNNNNNNNNNNNN
Concatenate and clone
--------- CTAGNNNNNNNNNNNNNNNNNNNNNNCTAGNNNNNNNNNNNNNNNNNNNNNNCTAG---------
--------- GTACNNNNNNNNNNNNNNNNNNNNNNGTACNNNNNNNNNNNNNNNNNNNNNNGTAC---------

Fig. 2. Schematic of SAGE procedure. The anchoring enzyme (AE) is *Nla*III and tagging enzyme (TE) is *Bsm*FI. Boxed A and B are independent linkers, whose 3′ portions are designed to contain TE sequence. Transcript-derived tag sequences are denoted by Ns. Blunt end ligation step is denoted as *, and discussed later in the text.

partially overlapping) the 3′-*Nla*III sequence. After the reaction mixtures are digested with type IIS enzyme, released portions are recovered. Resulting staggered ends of the products are then blunt-ended by T4 DNA polymerase. Two portions are mixed again and ligated. Since the 5′-ends of the linkers are blocked by amino group, only the mRNA-derived termini are able to be ligated in a tail-to-tail orientation. The products are PCR-amplified, cleaved by *Nla*III, an anchoring enzyme, and then separated by polyacrylamide gel electrophoresis (PAGE). Ditag fragments flanked both ends with *Nla*III cohesive terminus are isolated and ligated to obtain concatemers. Highly concatenated products are recovered by PAGE, and cloned into a plasmid vector for sequencing.

3. Studies made by the use of SAGE

Since the SAGE procedure has been developed and introduced as a tool for the study of gene expression, a variety of biological phenomena has been analyzed. Total tags analyzed by this method are now close to five million (Fig. 3) (Velculescu, 1999). Representative studies are listed in Table 1, in which highly diverse types of cells and tissues under a variety of physiological and pathological conditions can be noticed. Numbers of total collected tags in each study were variable. No theoretical consideration has been made about how many tags should be collected to construct a reliable gene expression profile.

3.1. Cancer studies

The most preferred subjects were human cancers for a variety of reasons. By comparing the gene expression profiles derived from cancer and normal tissue of interest, a large number of genes were identified as tumor specific. Usually Northern blot hybridization analysis was performed for the confirmation of differential expression of these genes against a number of independently isolated tissue samples of similar nature. About one half of the overrepresented genes identified by SAGE were reproducibly present in these samples, while the behavior of the other half was quite different. This may reflect the heterogeneity among tumors from different individuals.

These genes were mostly derived from either a known gene or a matched expressed sequence tag clone. This is mainly due to the tag's smallness. To overcome the difficulty of using totally unknown

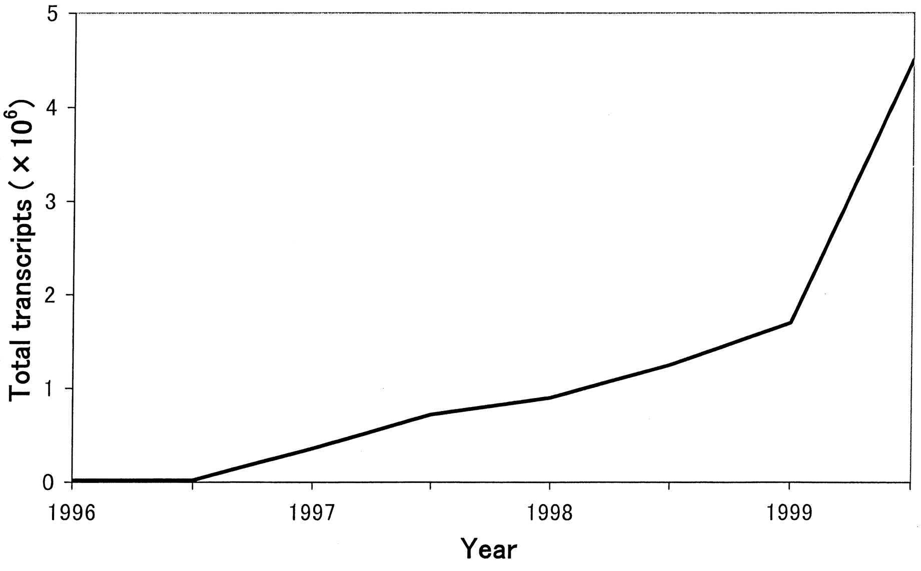

Fig. 3. Cumulative transcripts analyzed by SAGE worldwide (from Velculescu, 1999).

Table 1
Summary of SAGE analysis[a]

Cell, tissue	Total tags sequenced	Unique genes	Reference
Yeast (glucose-grown)	60 633	4665	Velculescu et al., 1997
Normal colon	62 168	14 721	Zhang et al., 1997
Colon tumor	60 878	19 690	Zhang et al., 1997
Colon cell	60 373	17 092	Zhang et al., 1997
Pancreatic tumor	61 592	20 471	Zhang et al., 1997
Pancreas cell	58 695	14 247	Zhang et al., 1997
Colorectal cancer cell	101 694	7202	Polyak et al., 1997
Rat embryonic fibroblast			
REF-Val135 (32′C)	30 386	9950	Madden et al., 1997
REFVal135 (38′C)	30 313	9240	Madden et al., 1997
REF-Val135 (32′C+38′C)	60 629	15 562	Madden et al., 1997
REF-Phe 132 (32′C)	10 519	5119	Madden et al., 1997
Rat mast cell	40 759	11 300	Chen et al., 1998
Lung-1	58 273	15 070	Hibi et al., 1998
Lung-2	59 885	15 667	Hibi et al., 1998
Lung cancer-1	56 817	17 535	Hibi et al., 1998
Lung cancer-2	51 901	16 443	Hibi et al., 1998
Endothelial cell	12 721	5448	de Waard et al., 1999
Skeletal muscle	53 875	12 207	Welle et al., 1999
Reed–Steinberg cell	1055	701	van den Berg et al., 1999a
Monocyte	57 560	35 037	Hashimoto et al., 1999a
Macrophage (monocytes stimulated with GM-CSF)	57 463	(overall)	Hashimoto et al., 1999b
Macrophage (monocyte stimulated with M-CSF)	55 856		Hashimoto et al., 1999b
Dendritic cells	58 540	17 000	Hashimoto et al., 1999a
Kidney	12 154	4800	Virlon et al., 1999
Dentate gyrus	1792	1242	Datson et al., 1999
Rice	10 122	5921	Matsumura et al., 1999
Yeast (oleate-grown)	13 979	1700	Kal et al., 1999
Thyroid	10 994	6099	Pauws et al., 2000
Mesenchymal progenitor	3177	2107	Ji et al., 2000
Liver	30 982	8596	Yamashita et al., 2000
Oocyte	50 000		Neilson et al., 2000

[a] Reports that do not contain appropriate information about numbers of tag or unique gene are not listed in the table. A public SAGE tag database is also available (Lal et al., 1999, http://www.ncbi.nlm.nih.gov/SAGE/).

tags of 13–14 bp, RT-PCR-based recloning method has been devised (see below).

3.2. Immunological studies

As seen in Table 1, only a few SAGE analysis has been directly applied for the study of immunological phenomena. Chen et al. (1998) have reported that the changes in gene expression in the rat mast cells before and after they were stimulated through high affinity receptors for immunoglobulin E (Fc ε RI). Among the diverse genes that had not been previously associated with mast cells were macrophage migration inhibitory factor, receptors for growth hormone-releasing factor and melatonin, and a number of components functioning as the exocytic machinery. Dozens of differentially expressed genes in response to Fc ε RI were also identified. These were the genes for preprorelaxin, mitogen-activated protein kinase kinase 3, the dual specificity protein phosphatase, rVH6, and many others, majority of which have not been identified as stimulation-reactive genes before this analysis. Though these findings were obtained from the rat mast cell line, extension to their normal rat counterparts and to human cells can easily be carried out.

SAGE analyses were also conducted for human monocytes and their differentiated descendants, macrophages and dendritic cells (Hashimoto et al., 1999a,b). Since human blood monocytes can be differentiated into macrophages and dendritic cells in vitro by culturing monocytes in the presence of granulocyte–monocyte colony-stimulating factor (GM-CSF) or monocyte colony-stimulating factor (M-CSF) (Tushinski et al., 1982; Gasson, 1991; Matsuda et al., 1995; Hashimoto et al., 1996), and GM-CSF, interleukin-4 and tumor necrosis factor-α concomitantly (Sallusto and Lanzavecchia, 1994; Akagawa et al., 1996; Palucka et al., 1998), respectively. The SAGE profiles for these cells were compared to each other. Both GM-CSF-induced and M-CSF-induced macrophages expressed similar sets of genes, indicating their functional similarity. However, differences in gene activity were also noticed, such as in monocyte-derived chemokine, legumain, prostaglandin D synthetase m and lysosomal sialoglycoprotein genes. These genes may provide tools to define macrophage subsets.

From the comparison between monocytes and dendritic cells, many differentially expressed genes were identified. Up-regulation of a number of chemokine genes, such as TARC, MDC, and MCP-4 may explain preferential chemoattraction of Th2-type lymphocytes. TARC overexpression was also prominent in Reed–Steinberg cells (van den Berg et al., 1999b), which are characteristic to Hodgkin's lymphoma. This disease is known to cause a remarkable influx of Th2-like lymphocytes. Many other genes that were differentially expressed were those related to cell structure and cell motility, and numerous unknown genes that showed no database-matching. Since dendritic cells have been considered to be heterogeneous, these genes may help define, if any, subsets. No further analyses were conducted to unknown sequences.

3.3. Yeast

Yeast is widely used to clarify the biochemical and physiologic parameters underlying eukaryotic cellular functions. The entire genome sequence has been determined (Goffeau, 1997) and the number of genes has been estimated to be about 6300. Total mRNA molecules were also been estimated to be 15 000 per cell (Hereford and Rosbach, 1977). For these reasons, yeast was chosen as a model organism to evaluate the power of the SAGE technology. The most extensive SAGE profile was thus made for yeast, total tags of which corresponded to 60 633 representing 4665 genes (Velculescu et al., 1997). Of these tags, 93% matched the yeast genome and the expression levels of each tag were 0.3 to more than 200 per cell. These expressed genes included 76% of the total genes predicted from analysis of the yeast genome. However, strangely enough, several hundred new genes were identified that had not been predicted. These sequences may represent very small genes, since the yeast genome sequence has been annotated for >300 bp ORFs.

Correlation between protein and mRNA abundance has been studied for more than 150 proteins (Gygi et al., 1999). These authors found that the correlation was insufficient to predict protein expression levels from the SAGE tag data. Indeed, for some genes, the protein levels varied by more than 20-fold, while the mRNA levels were of the same value. Conversely, invariant steady-state levels of certain proteins were observed with respective mRNA transcript levels that varied by as much as 30-fold. These results indicate either that there is no direct correlation between protein and mRNA abundance, or that the numbers of SAGE tag does not reflect accurately those of corresponding mRNA. In any case, we have to remind that simple deduction of protein level from SAGE analysis is insufficient.

4. Drawbacks, problems and technical modifications of SAGE

Several problems are pointed out for the SAGE procedure. They are technical difficulties at first, and more serious problems intrinsic to the method secondly.

As technical problems, a disadvantage of the need of relatively high amount of mRNA, relative difficulty to construct tag libraries, and others are pointed out.

On the first point, a couple of reports dealing with it, namely, MicroSAGE (Datson et al., 1999) requires 500–5000-fold less starting input RNA, and is simplified by the incorporation of a 'one-tube' procedure for all steps from RNA isolation to tag release. Using this technique, the authors were able to obtain an expression profile of hippocampal punch from a rat brain slice containing less than ten cells. SAGE-lite, is another similarly-devised protocol which also allows the global analysis of transcription from less than 100 ng of total starting RNA (Peters et al., 1999). SAGE adaptation for downsized extracts was also set up, enabling a 1000-fold reduction of the amount of starting material (Virlon et al., 1999). The potential of this approach was evaluated by studying gene expression in microdissected kidney tubules of about 50 000 cells.

As for the technical difficulty of the procedure; in the original SAGE protocol, major products of PCR are often linker-dimers. To minimize contaminating linker molecules, biotinylated PCR primers were introduced (Powell, 1998). This modification generates biotinylated ditag products at an early stage in the SAGE protocol, thus allowing removal of the unwanted linkers by binding to streptavidin beads used at a later stage.

To eliminate a small average size of cloned concatemers by which the efficiency of tag collection is limited, final ligation step was modified (Kenzelmann and Muhlemann, 1999). A simple introduction of heating step yields cloned concatemers with an average of 67 tags as compared to 22 tags obtained by the original protocol.

A major problem of the SAGE approach is how to further analyze the unknown tags. In the original report (Velculescu et al., 1995) the utilization of a conventional oligonucleotide-based plaque lift method was employed successfully for the isolation and cloning of a number of genes. However, in a practical sense, it is almost impossible to discriminate one-base mismatched sequence within oligonucleotides of only 13–14 bp in length by a rather gross temperature-regulated DNA–DNA hybridization technology, thus resulting in numerous false positives. An RT-PCR-based method was thus developed to analyze the corresponding genes (van den Berg et al., 1999a). This approach utilizes identified tag sequences and oligo-dT as PCR primers. Although this PCR is suffered from two disadvantages, i.e. shortness of 5′ tag-derived specific primer and the common nature of the 3′ primer to all mRNAs, the authors claimed that the method worked well at least for some unknown genes. Similarly, Matsumura et al. (1999) reported a procedure to recover a longer cDNA fragment by PCR using the SAGE tag sequence as a primer, thereby facilitating the analysis of unknown genes identified by tag sequence in SAGE.

As for the problems intrinsic to the SAGE procedure:

(1) The length of gene tag is extremely short (9 or 10 bp). As already discussed above, short tag makes further analysis difficult, especially when tags are derived from unknown genes. Meanwhile, isolation of the unknown gene is often the ultimate goal for most analyses using the SAGE procedure.

The linkers used in the SAGE method are designed to use the type IIS restriction enzyme (mainly *Bsm*FI) for tagging gene sequences from outside the anchoring enzyme site. Therefore, 11 bp may be the longest obtainable tag by this protocol. Furthermore, *Bsm*FI does not always give exact 14 bp tags, but often yield longer or shorter fragments (between 12 and 16 bp from our experience). Especially when cleavage is carried out at lower temperature like at 37°C, instead of the manufacturer's recommended 65°C, smaller fragments tend to appear more frequently. This ambiguity occurred in sequence tagging may generate another problem. Since ditag formation is performed by direct tail-to-tail ligation without any artificial demarcating nucleotides in between, delineation of tag ends may become ambiguous. (How can one discriminate 12+16, 13+15, 14+14 and so forth?)

(2) Since the publication of the SAGE methodology in 1995, only a limited number of laboratories were able to use it successfully in spite of its overwhelming potential, tacitly indicating its intrinsic difficulty of preparing tag libraries. Contamination of large quantities of linker-dimer molecules that arose during a linker ligation step and low efficiency in blunt end ligation are perhaps the main reasons that account for the difficulty. Blunt end ligation (* in Fig. 2) is by itself highly inefficient compared to cohesive end ligation. The more serious problem in blunt end ligation is that the reaction rate varies with the terminal sequence of the DNA by more than 10-fold. This means that ditag formation may occur

unevenly depending upon tag's tail nucleotides, inevitably leading to the generation of bias in the tag distribution in the library. Furthermore, this tail-to-tail ligation does not necessarily generate ditag molecules flanked both sides by linker A and B (for A and B, see Fig. 2), but half of the products would have only A or B for both sides. The latter two types of ditag molecules would easily take a panhandle structure by preferential intramolecular annealing after denaturation during PCR, resulting in low efficiency in amplification. This may be another cause of bias.

(3) Depending upon anchoring enzyme and tagging enzyme used, some fraction of mRNA species would be lost. Although recognition sites for four base cutter are present every 256 bp stretch on average, and the majority of mRNAs should have such sites of any kinds, some species may not. It is hard to estimate the fraction. On the other hand, the recognition sequence is GGGAC for *Bsm*FI, the most frequently-used tagging enzyme. Recognition sites for this enzyme should appear every 500 bp, since GGGAC is equivalent to GTCCC because of its non-palindromic nature. Thus, about 2% of tag species would be lost during a tagging step (10 bp tag×50=500 bp). To minimize loss of mRNA species, it is recommended to construct two independent profiles made by the use of two different combinations of anchoring enzyme and tagging enzyme. This task would be tough, but one should be able to examine the reliability of them and would have a dependable profile.

(4) The fourth problem is a little more serious and is discussed in detail in the SAGEmap (www.ncbi.nlm.nih.gov/SAGE/). There are two problems to be coped with when dealing with SAGE data. The first deals with sequencing error, and the second, with making valid tag to gene assignments. Assuming that there is an average 1% per base sequencing error rate because they are usually only single-pass sequenced, the chance of one or more errors occurring is roughly 10% for ten bases. The error will lower the correct tag counting, but will also either increase the tag count of an already established tag, or will establish and count a tag which does not, in reality, exist. Currently, the data tags counted only once are omitted from analysis, though this may not be an ideal approach. This empirical approach has been used in SAGE tag-count sets in which roughly 250 000 total tags have been sequenced.

In consideration of the second problem, tag to gene assignments, several difficulties are also encountered. A ten base tag is by no means a perfect representation of a gene's entire transcript. There will be instances in which multiple genes share the same tag as observed frequently in the family genes, and instances in which one gene has multiple tags as in the genes having alternate poly A sites. A population polymorphism may also cause a similar problem.

5. Detailed descriptions of modified SAGE procedure

Since the SAGE methodology has been published, 15 or so laboratories applied it for studies of a variety of cells and tissues. All these studies handled 9–10 bp tags as substitutes of mRNAs, and restriction enzymes used are the same (*Nla*III–*Bsm*FI) as in the original method, except for one (Virlon et al., 1999) that used *Sau*3AI as an anchoring enzyme. To increase tag length, we searched for a restriction enzyme file, and tried to construct the tag library with a combination of *Rsa*I and *Bsm*FI, that would generate 14 bp tags (Ryo et al., 2000). Together with GTAC (*Rsa*I site sequence), 18 bp stretch should be conveniently used for further study of unknown genes. Fig. 4 shows the schematic representation of the procedure. Usually 30–50 μg of total RNA was used to synthesize double-stranded (ds) cDNA with the cDNA synthesizing kit (Takara, Tokyo, Japan) and oligotex dT30-latex beads (Takara) (Ryo et al., 1998). After washing with TE (10 mM Tris–HCl, pH 7.5, 1 mM EDTA), ds cDNA was digested with *Rsa*I (NEB, Beverly, MA). The 3′ portion of cDNA was collected by centrifugation and then treated with T4 DNA polymerase (Takara) in the presence of dATP, dCTP and dGTP (or dGTP only) (200 μM each) to generate a 5′ single A protrusion. Linker A (5′-TACAGGATACGCCATGGGAC-3′, 5′-pTCCCATGGCGTATCCTGTA-3′), designed to have a 5′ single T overhang, was then ligated to the cDNA with T4 DNA ligase (Takara) similar to the A-T cloning procedure (Marchuk et al., 1991). The

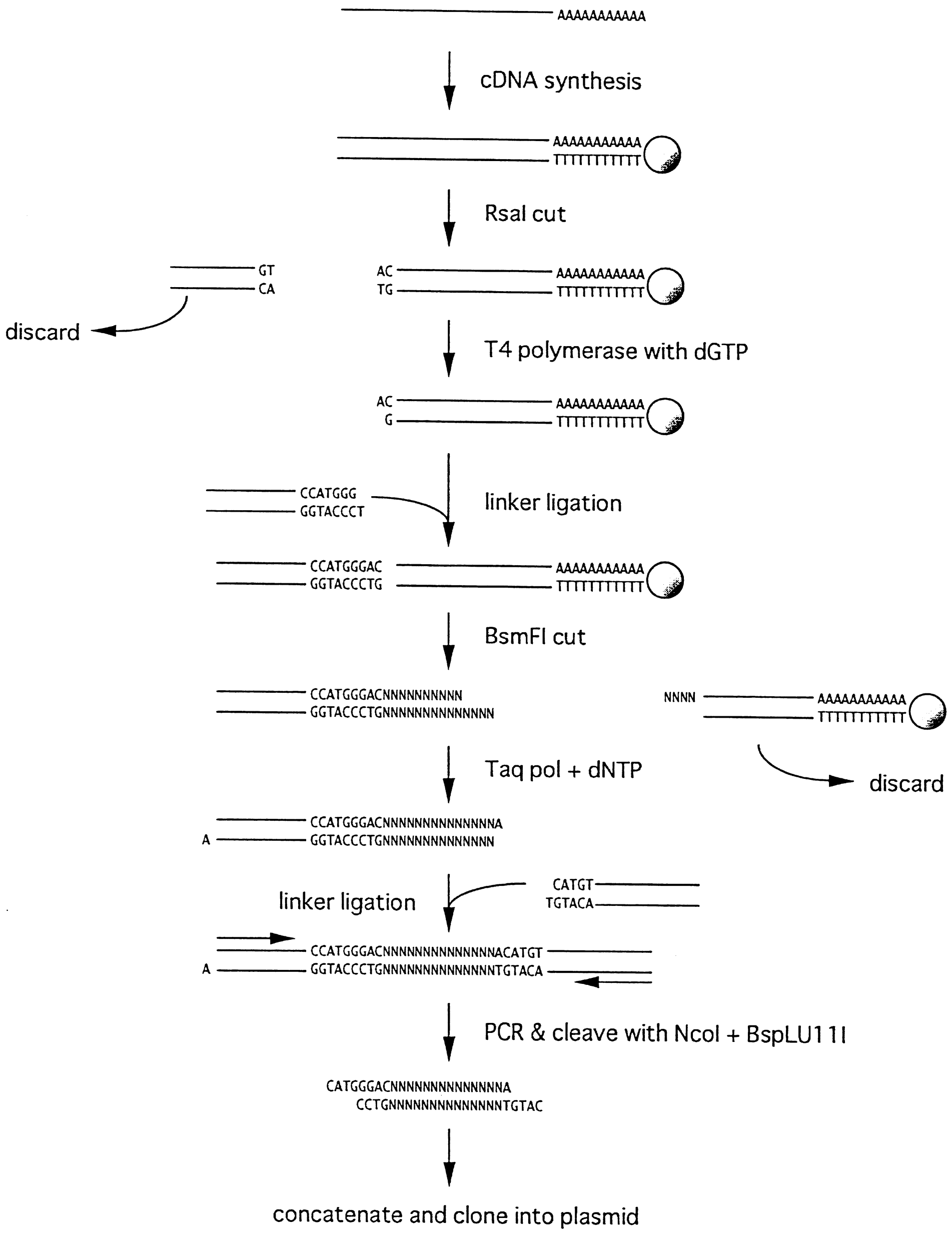

Fig. 4. Schematic representation of modified SAGE. By the use of *Rsa*I and *Bsm*FI as AE and TE, respectively, 14 bp monotag sequences can be obtained. Two linker ligation reactions are conducted by a cohesive but non-palindromic manner.

bound cDNA was digested with the tagging enzyme *Bsm*FI (NEB), whose site was designed to be generated at the linker-cDNA junction. The supernatant fraction was then subjected to phenol–chloroform extraction followed by ethanol precipitation. Subsequently, cDNAs having 3′ recessive ends were treated with Taq DNA polymerase (Takara) in the presence of four dNTPs to fill in and generate a single A protrusion at their 3′ ends. Linker B (5′-TAGTCAGTTGCGACACATGT-3′, 5′-pCATGTG-TCGCAACTGACTA-3′), designed to have a 3′ single T cohesive end, was then ligated to cDNA in an A-T ligation manner. The ligation products were subjected to PCR amplification in 4×100 μl reaction consisting of an initial denaturation step at 95°C for 2 min followed by 10–15 cycles of 95°C for 30 s, 58°C for 30 s, and 72°C for 30 s, with a final extension step at 72°C for 4 min using 5′-TACAGGATACGCCATGGGAC-3′ and 5′-TAGTCAGTTGCGACACATGT-3′ as primers. The PCR products were separated with 6% PAGE and 54-bp band was recovered by electroelution. When the amount is not sufficient for the following procedure, second PCR may be performed with this isolated material under the same reaction conditions as in the first PCR. The 54-bp DNA was double-digested with *Nco*I (NEB) and *Bsp*LU11I (Roche Molecular Biochemicals, Tokyo, Japan) and separated by 12% PAGE. The 23-bp band was recovered by electroelution. Since these restriction enzymes generate the same cohesive ends, 23-bp fragments were concatenated with T4 DNA ligase. The products were separated again with 4.5% PAGE and the >500-bp regions were excised and recovered by electroelution. Purified concatemers were cloned into a *Nco*I-digested pUC-based plasmid vector. Recombinant plasmids were sequenced with an ABI 377 automated sequencer (PE Applied Biosystems). Tag extraction and further analyses were performed with PROGENEX software (Fujiyakuhin Co. Ltd., Saitama, Japan).

The method described took advantage of non-palindromic, but cohesive termini generated on cDNA tags for the ligation of both 5′- and 3′-linkers, allowing high efficiency in tag-linker ligation but low possibility in linker-dimer production. These are beneficial for subsequent PCR reaction not only to obtain good yield of specific products but also to avoid contaminating by-products that interfere with the band isolation from the gel. In addition, the biases derived from PCR should be considerably low because each 14-bp tag is flanked on both sides by two distinct linkers. There should be no tag sandwiched by the same linker.

6. Gene expression profile in HeLa cell by modified SAGE

Thus we made two tag libraries independently from the same RNA preparation of HeLa cells. After small-scale sequencing to confirm that the mRNA expression patterns were similar at least in abundant classes, they were mixed for normalization. Total numbers of independent clones were 104 000. Sixteen clones were randomly chosen, plasmids isolated, digested with an appropriate pair of restriction enzymes, and separated on an agarose gel. An average insert size was approximately 1.2 kbp, that is, equivalent to about 50 tags, indicating that overall numbers of tags were 5.2×10^6, sufficient for a cDNA library. A large-scale sequencing was thus performed for this library.

A total of 80 000 tags were cataloged and corresponded to 12 976 unique genes. Fig. 5 shows the increase in gene representation as the number of tags sequenced increases, indicating that new transcripts are still being identified after 80 000 tags were sequenced. However, since the rate of increase became to near constant at 50 000 or more total tag counts, i.e. about 1200 new species per 10 000 tags sequenced, a considerable portion of the increase might be derived from sequencing error, as discussed above.

Fig. 6 shows a plot of the number of gene species and the frequency of each tag. As the histogram shows, frequencies of gene species appeared generally continuous from low to high abundance classes. This is distinct from that derived from reassociation kinetics (a Rot analysis) (Bishop et al., 1974). Only 80 gene species (0.77% of the total unique genes) appeared >100 times, 586 gene species 10–99 times and 9174 gene species only once.

Anchoring enzyme used in our procedure was *Rsa*I, a 4-bp-recognizing enzyme. Since the recognition site appears every 256 bps on average, many mRNAs should have multiple cutting sites within the genes. To see whether the tag sequences obtained corresponded, in fact, to the 3′-most site of mRNAs, four representative genes were analyzed (Fig. 7). Among them, ribosomal protein L13A gene has four *Rsa*I sites and 8370 tags were derived indeed from the 3′-most one, while 31 (11, null and 20) were from upstream sites. Similar results were also obtained from other three genes, namely, 807 vs. 21, 257 vs. 1, and 190 vs. 3 for elongation factor 1-α, GAPDH and transketolase, respectively. Although tags from unexpected sites might be derived either at the step of cDNA synthesis or *Rsa*I digestion, it is also possible that the mRNA itself has errors at the

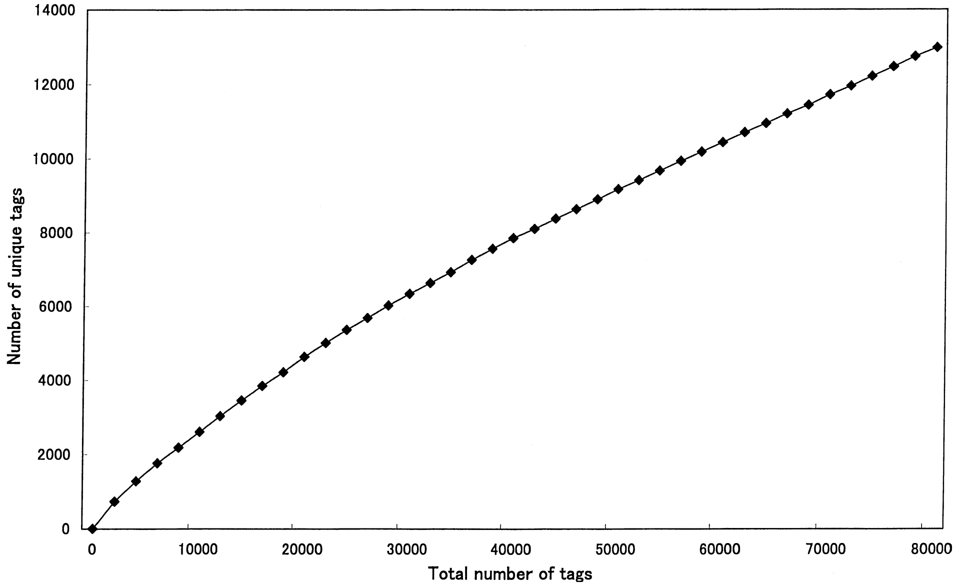

Fig. 5. Cumulative total gene representation in HeLa cells. Sequenced cDNA tag accumulation was monitored for unique genes using the PROGENEX software package (Fujuyakuhin, Saitama, Japan).

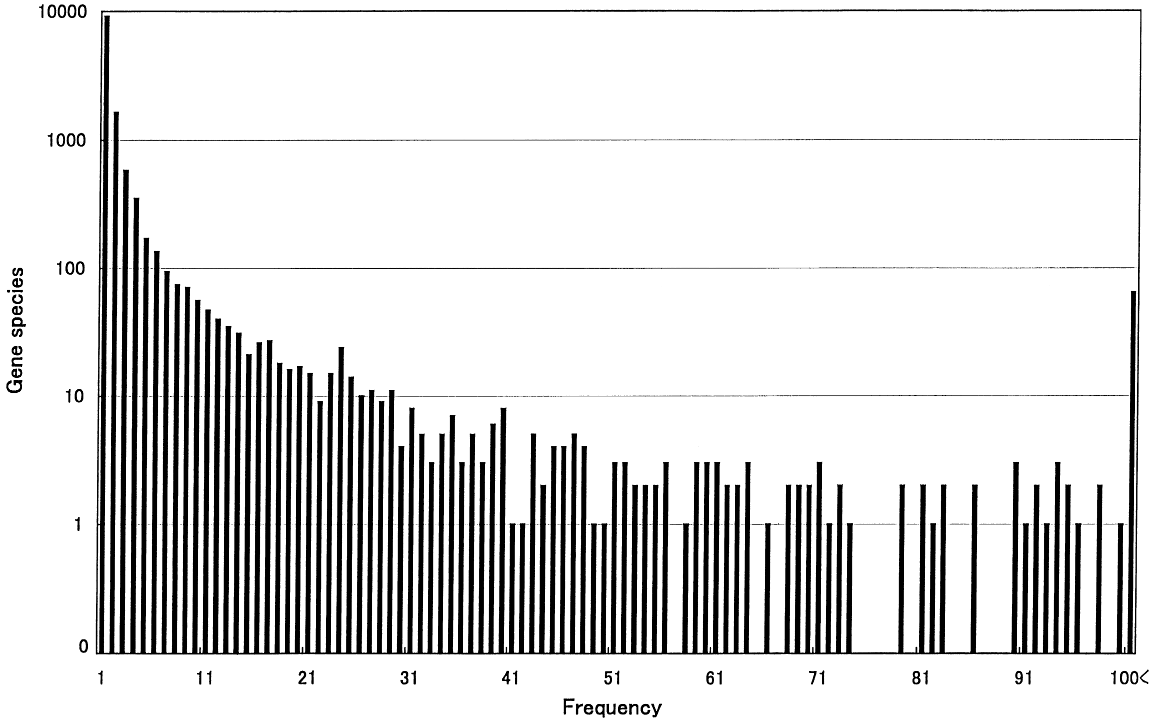

Fig. 6. Histogram of gene expression in HeLa cells. The number of gene species and the frequency of each tag are plotted.

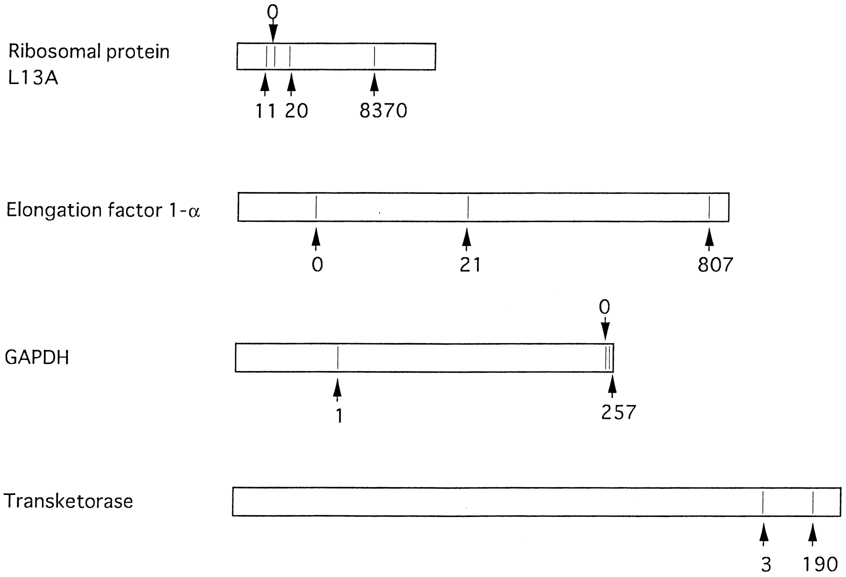

Fig. 7. *Rsa*I restriction sites in some of the abundant genes and tag abundances. Numbers with arrow are tag counts at corresponding restriction sites in a sample of 80 000 total tags.

site. In any case we should be careful for data analysis, by recognizing such a possibility of occurrence of unexpected tags. Frequency may be up to 2–3% for a given gene.

Table 2 lists the 48 most abundantly expressed genes in the HeLa cell, all of which were expressed at greater than 150 times among 80 000 tags sequenced. Most of these genes corresponded to well characterized protein genes involved in protein synthesis, mitochondrial, cytoskeletal, nuclear, and so on. A remarkably high abundance in genes related to protein synthesis and mitochondrial functions might reflect the active state of growth and energy metabolism of this particular type of cell. Among them, genes related to protein synthesis, cytoskeletal and membrane components are shown in Table 3. An unexpected and rather embarrassing observation was the extensive differences in mRNA abundance among ribosomal proteins, which are thought to be present in near equal amounts in each particle. Tag count of the most frequent gene (L13A) is more than a hundred times as abundant as those of infrequent ones. Similar results have been seen in all SAGE analyses, and in the study that did not relied on SAGE protocol (Okubo et al., 1992), though the magnitude was much moderate. These observations may reflect the different turnover rate of each ribosomal subunit protein even though the ribosomal particles appeared highly stable. Another explanation may be that, taking into account the different compositions of the ribosomal components as observed in other SAGE analyses compared with our results, eukaryotic ribosomes may be composed of a diverse set of heterogeneous complexes, not an assembly of homogeneous particles.

Rearranging expressed genes according to their functional categories may give a rough idea for the abundance of a particular gene of interest. For example, if an orphan receptor is to be compared before and after some kind of treatment, 20 000 or so total tags may have to be collected for both states, since the majority of membrane protein gene appears <100 times among 80 000 tags (<25 tags in 20 000) as seen in Table 3.

7. Comparisons of profiles constructed from various numbers of cDNA tags

We then divided this profile composed of 80 000 tags into smaller size profiles (subprofiles), and compared profiles with the same number of tags each

Table 2
Abundantly expressed genes in HeLa cells[a]

Tag (gtac . . .)	Gene product	GenBank accession no./locus	Times detected
CAGGCAGTGACAGC	Ribosomal protein L13A	HS23KDHBP	8370
TACACGCGCCTGGG	Ribosomal protein S17	HUMRPS17	5062
TGGCCGCCATGAGG	Ribosomal protein L36	AF077043	4731
CTGCTGGTGGGGCT	Acidic ribosomal phosphoprotein P2	HUMPPARP2	3473
TGCCGATTGAAGCC	Cytochrome *c* oxidase 2	7440-7453	2669
TTCGAGTCTGCGTT	Cytochrome *c* oxidase 3	9174-9187	2116
TGACAACCTCAGCT	Ribosomal protein S15a	HSRPS15A	1326
CAGCAGCAAGGCAG	MHC protein homologous to chicken B complex protein	HUMMHBA123	1002
TGTGGCGCTCCGTG	H3.3 histone, class B	HUMHISH3B	817
TTTTTAATGGAAAC	Elongation factor 1-alpha	HSEF1AC	807
CTGCAGGCCTCCTA	Ferritin L chain	HUMFERL	690
CTGAAACTGCCGCC	Neuronal tissue-enriched acidic protein (NAP22)	NAP22	521
AAGCTGCTGGAGCC	Ribosomal protein S16	HUMSRAA	520
GTGCCGCGGAAATG	Ribosomal protein S21	HUMRPS21X	497
CTGCTCTCAAARRG	Ribosomal protein L7	HUMRPL7Y	468
CAGGCAGCAGCTCC	EST	AI19 1773	460
CTGCCTCACAGTGG	EST	AA502482	445
CTGGCCATCTTGGG	Ribophorin II	HSRIBIIR	427
TGCCCTCTGCTGGG	Mitochondrial ribosomal protein S12	Y11681	401
CAGGGCCCGCTGTG	EST	AA186619	389
TGACCTCGTCTGTC	Profilin	HUMPROF	378
CCAGTGATCCCCAC	Cytochrome oxidase subunit Vib.	HSCYTVIB	327
TATCCCTATGAGGC	NADH dehydrogenase 4	10974-10987	300
CGAGCAAATGCCAG	PolyA binding protein	HSPOLYAB	291
TACACGCGCCTGGC	EST	AA554166	270
CTGCGTCGAGCTCT	Full length insert cDNA cloneZC45B12	HUMZC45B 12	266
TAGGGGCCCGGATC	Intermediate conductance calcium-activated potassium channel (hKCa4)	AF033021	264
CCTGTGCTCAACCA	Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	HUMGAPDHG	257
CCATGCCCCCTGCC	EST	AI373009	255
CAGGAGGOOTTOCT	EST	AA442510	251
CCCCTCCCCAGTCT	None		236
TGGCCGCAGCAAAG	EST	AA303712	231
TATACTTCGCAACA	EST	AI748937	221
AGCCACTGAGGGTC	None		216
CGGATGCTGGCAGG	EST	AA157980	211
TGGGGCCTGTGTGG	EST	HSPD02590	205
TGGCCCTCGGTGCT	EST	AI250777	203
CCAATGACGGTTGC	Ribosomal protein S23	AB007158	196
CGCCCCGACCTGCG	Ribosomal protein L28	HSU14969	195
TGAGAGGAGGGGTA	Transketolase (TKT)	HSU55017	190
TCGTGCGCCTCGCT	Thymosin beta-4	HUMTHYB4	188
ACTGACTTGAGACC	Beta-actin	HSAC07	180
AGGCAGTGACAGCC	EST	HSFIH165	173
CAGGGCAAGAAGCC	Gamma-interferon-inducible protein (IP-30)	HUMIIP	169
TTGGCCTCGCTGAT	Proteasome subunit p40/Mov34 protein	HUMP40MOV	158
TGCCCCCCGCTCAT	EST	AA446291	152
TGACCATCAGTGTC	EST	W65292	152
CGGGCTGGCCTGTG	Cysteine proteinase inhibitor precursor cystatin C	HSCYSTCR	152

[a] Genes appeared more than 150 times as cDNA tags in a sample of 80 000 total tags are listed in order of abundance. Tags originated from mitochondrial genome (X93334) are denoted as the numbers of the locus instead of accession number.

Table 3
Categorized genes related to protein synthesis, membrane and cytoskeleton

Protein synthesis	
Ribosomal protein L13A	8370
Ribosomal protein S17	5062
Ribosomal protein L36	4731
Acidic ribosomal phosphoprotein	3473
Ribosomal protein S15a	1326
Elongation factor 1-alpha	807
Ribosomal protein S16	520
Ribosomal protein S21	497
Ribosomal protein L7	468
Ribosomal II	427
PolyA binding protein	291
Ribosomal protein S23	196
Ribosomal protein L28	195
Ribosomal protein L30	149
Ribosomal protein L35	141
Ribosomal protein L10	121
Ribosomal protein L14	110
Ribosomal protein L8	103
Ribosomal protein L17	100
Acidic ribosomal phosphoprotein P0	94
Ribosomal protein L5	92
Ribosomal protein S7	92
Ribosomal protein S11	90
Eukaryotic initiation factor 4 gamma	86
Translation initiation factor eIF-3 p110 subunit	86
Signal recognition particle subunit 14	63
Ribosomal protein L6	62
Ribosomal protein L27	54
Eukaryotic initiation factor 4 gamma	53
Ribosomal protein L5	49
Ribosomal protein L31	40
Ribosomal protein S10 homologue	35
Putative ribosomal protein L23	32
Ribosomal protein S13	31
Signal recognition particle subunit	27
Elongation factor-1-gamma	26
Ribosomal protein L8	23
Ribosomal protein S29	22
Membrane protein	
MHC protein homologous to chicken B complex	1002
Calcium activated potassium channel (hKCa4)	264
Benzodiazapine receptor (peripheral) (BZRP)	124
Beta-2-microglobulin (B2 M)	109
26-kDa cell surface protein TAPA-1 (CD81)	79
Immunoglobulin receptor alpha chain	73
Leukemia virus receptor 1 (GLVR1)	71
Cyclic nucleotide grated channel 2 (HCN2)	47
Leptin receptor splice variant form 12.1	45
Putative receptor protein (PMI)	37
Peripheral myelin protein 22	36
MHC class I HLA-C-alpha-2 chain	29
Leucocyte antigen CD97	25
Tumor necrosis factor receptor	20
Cytoskeleton	
Beta-actin	
Lamin B2 (LAMB2) gene and ppv1 gene sequence102	180
Vimentin	95
Gamma-tubulin	83
Signal recognition particle subunit 14	63
Cytoplasmic dynein light chain 1 (hdlc1)	47
Beta-tubulin class III isotype (beta-3)	37
Beta-catenin	34
Beta-tubulin	30
Lamin A	27
Beta-actin	23
Zyxin	23

other. The purpose of such comparisons was to examine the reproducibility in tag appearance with respect to final abundance of each tag sequence. Original profile with 80 000 tags (designated as 80K profile) was thus divided randomly into 40 portions. Since it is difficult to divide randomly the once-completed 80K profile, 2K profiles were actually constructed one by one from the beginning of tag collection. By randomly combining these subprofiles, a series of pair profiles with 2000, 4000, 10 000, 20 000 and 40 000 tags was constructed and designated as 2K-1, 2K-2, 4K-1, 4K-2, 10K-1, 10K-2 and so on. Profiles 4K-1 and 4K-2 include 2K-1 and 2K-2, respectively, and profiles 10K-1 and 10K-2 include 4K-1 and 4K-2, respectively, and so on.

The results of comparisons are tabulated (Table 4) with respect to selected gene sequences of variable final abundance classes. By the comparison between 2K-1 and 2K-2, practically no difference in tag counts was observed for the highly abundant genes, e.g. 196 vs. 209 for ribosomal protein L13A gene, 95 vs. 100 for ribosomal phosphoprotein P2, and so on. Among the genes in less abundant class, however, there were genes that showed considerable difference, such as the sequence TACACGCGCCTGGC (EST, bold in the table) that revealed 16 and 7 in 2K-1 and 2K-2 profile, respectively. This sequence showed a big difference in 4K profiles as well, i.e. 29 vs. 11. Similar counts for this gene were observed at last in 40K profiles, as 119 vs. 151.

Table 4
Comparison of gene expression frequencies between two gene profiles consisted of the identical number of tags[a]

Tag sequence (gtac . . . .)	Gene product	Profiles										
		2K-1	2K-2	4K-1	4K-2	10K-1	10K-2	20K-1	20K-2	40K-1	40K-2	80K
Unique genes		861	824	1430	1435	2845	2619	4612	4760	7950	7571	12 976
CAGGCAGTGACAGC	Ribosomal protein L13A	196	209	413	452	1033	1117	2095	2248	3890	4480	8370
TACACGCGCCTGGG	Ribosomal protein S 17	134	122	294	258	709	666	1341	1352	2361	2701	5062
TGGCCGCCATGAGG	Ribosomal protein L36	122	130	238	246	586	607	1167	1269	2204	2527	4731
CTGCTGGTGGGGCT	Acidic ribosomal phosphoprotein P2	95	100	193	188	463	470	928	912	1676	1797	6473
TCCTGATTGAAGCC	Cytochrome *c* oxidase 2	61	72	122	141	313	335	662	652	1332	1337	2669
TATCCCTATGAGGC	NADH dehydrogenase 4	9	8	15	15	30	44	60	80	155	145	300
CGAGCAAATGCCAG	PolyA binding protein	6	4	12	13	37	31	66	66	159	132	291
TACACGCGCCTGGC	EST	**16**	**7**	**29**	**11**	**72**	**28**	**118**	**62**	119	151	270
CTGCGTCGAGCTCT	Full length insert cDNA clone ZC45B12	3	8	11	14	27	32	52	73	123	143	266
TAGGGGCCCGGATC	Intermediate conductance calcium-activated potassium channel (hKCa 4)	5	5	11	14	27	66	65	137	127	264	
TGTGTCTTCCTGTC	EST	2	3	5	4	**16**	**7**	28	24	72	43	115
GACCTACGCACACG	None	**3**	**1**	6	5	14	13	26	28	55	27	112
TAGGGATCATGTGT	EST	2	1	3	2	15	12	28	23	66	46	112
TGCTGCTGCTGCTG	Ribosomal protein L14	1		4	6	10	9	29	23	57	53	110
CAGGCAGTGCAGCC	None	7	5	11	6	23	13	36	28	36	73	109
TGCACTGTGCGCTG	Putative cyclin G1 interacting protein		**3**	**1**	**5**	**4**	**8**	14	16	25	30	55
TGGGCACCACCTCT	None	1	1	2	3	5	4	12	6	33	21	54
TCTGTGGATCTCCC	Ribosomal protein L27(RPL27)	2	3	2	4	9	9	15	16	30	24	54
CACCCCTGCTGTTG	Eukaryotic initiation factor 4 gamma (eIF-4 gamma)		1	3	3	8	8	14	14	33	20	53
TGGGCAGCTGGTGG	RNA helicase (Myc-regulated deadbox protein)	1	4	2	4	6	8	12	17	28	25	53
CAGGATGCCACCGC	Fragment encoding beta-tubulin	1		1	2	3	4	7	8	12	18	30
CAGGCTGGCGTCTT	CLN3 protein (CLN3)	**3**	**1**	4	3	6	3	8	4	17	13	30
CTGCAGCCTCCTAC	EST	**3**		**4**		**11**		**16**	**2**	16	14	30
TGATCCTGTGTGAG	erbB3 binding protein EBP1	1	1	2	1	**3**	**1**	8	5	18	12	30
CAGGACACCCCCGGG	AP-3 complex delta subunit		2	1	2	4	4	7	7	15	14	29
CGCTCCAAACTCAT	BBC1				1	2	4	6	5	14	6	20
CTGGCATCTTGGGC	EST	1	1	1	1	4	2	8	4	8	12	20
CGGGATTCCTTGCC	ADP-ribosylation factor (ARF 3)	2	1	2	4	2	4	5	4	11	9	20
GCAGCACCCCTGCC	Farnesyl pyrophosphate synthetase		2	1	2	33	5	7	12	8	20	
CACCACACCCAGCT	SIRP-beta1	1		1	1	1	1	**5**	**2**	6	4	10
CACCATGCCTGGCT	Dioxin-inducible cytochrome P450 (CYP1B1)							**1**	**4**	**3**	**7**	10
CAGGGAAGAGACCT	NAD$^+$-specific isocitrate dehydrogenase beta subunit precursor		1		2		2	2	3	5	5	10
CAGCCGGAGCCCA	None					1	1	2	2	4	6	10
CACCTACACCCAAC	Protein trafficking protein (S31iii152)		1		1		1		1	6	4	10
CGCGCCGGCTTCCA	None	1		1		1		**1**	**3**	2	3	5
CCAAATCTGCTTCC	*H. sapiens* mitochondrial DNA for D-loop									**5**		5
CCAAGTCTTACGTT	Inducible poly(A)-binding protein								1	**4**	**1**	5
CCTGCAGTGTTGAT	Uridine diphosphoglucose pyrophosphorylase	1		1		1		2		**4**	**1**	5
CTGTGCCAAGCCTA	Calcium-dependent protease (small subunit)						1		2	3	2	5

[a] Bold in the table indicate tag sequences, counts of which are considerably different between two subprofiles of the same size.

As expected, in comparisons of small size profiles like in 2K and 4K pairs, only the sequences appeared more than three times showed similar values. The genes showed this number are the highest 89 and 176 genes, and 10.8 and 12.3% of expressed tag species, respectively. If the profile size becomes greater, the number of genes that can be compared with considerable reliability increases. However, even in 20K or 40K profiles, there were a lot of genes of low abundance class, the appearance of which varies very much. These results show that the genes that can be compared with considerable reliability might depend on both the size of profile and the gene sequence. Chen et al. (1998) made a statistical analysis on the probability of representing a significant difference between the populations, such as between 2K-1 and 2K-2 in this study, or between any paired expression profiles of similar nature. Although statistical approaches might be useful to obtain the probability of differential expression, it would not tell about an individual gene whether it is really differentially expressed or not by simple mathematical manipulation of final profiles.

Accordingly, when the profiles derived from different cellular states are compared, it is strongly recommended to make subprofiles from the beginning of tag extraction, and pursue the gene sequences that show stably high appearance in one biological status but stably low appearance in the other status. The sequences that showed significant difference in expression by such comparison should become candidates for further analyses.

8. Application of the modified method

The modified method was applied to analyze gene expression pattern in the mouse microglial cell line (Inoue et al., 1999), and to systematically search differences in gene expression between hepatocellular carcinoma (HCC) and adjacent normal liver tissue (Kondoh et al., 1999) and the human T cell line MOLT-4 infected with or without HIV-1 (Ryo et al., 1999).

8.1. Microglia

Microglia are ubiquitously distributed in the central nervous system and constitute 5–20% of the neuroglial population in the mature brain (Lawson et al., 1990). The ontogeny of microglia has been a controversial topic for a long time (Theele and Streit, 1993). However, it has been shown that they are a kind of phagocytic cell and arise either from the neuroepithelium or mesodermal tissue (Fedoroff and Hao, 1991; Theele and Streit, 1993). Although the function of microglia has been suggested immunocompetent in the brain (Streit et al., 1988), no comprehensive study has been performed on gene expression prescribing for microglial phenotypes. We therefore applied our modified SAGE method to an immortalized mouse microglial cell line and constructed a gene expression profile composed of 10 386 tags representing 6013 unique transcripts (Inoue et al., 1999).

Among the diverse transcripts that had not been detected previously in microglia were those for cytokines such as endothelial monocyte-activating polypeptide I (EMAP I), and for cell surface antigens, including adhesion molecules such as CD9, CD53, CD107a, CD147, CD162 and mast cell high affinity IgE receptor. These adhesion molecules and others may contribute to microglial migration to the central nervous system (Imai et al., 1997). In addition, we detected transcripts that are characteristic to hematopoietic cells or mesodermal structures, such as E3 protein, Al, EN-7, B94 and ufo. Furthermore, the profile contained a transcript, Hn1, that is important in hematopoietic cells and neurological development (Tang et al., 1997), suggesting the probable neural differentiation of microglia from the hematopoietic system in development.

8.2. Hepatocellular carcinoma

Surgically resected HCC and adjacent noncancerous liver tissue were subjected for the expression profile construction. A total of 50 515 and 50 472 tags were analyzed for HCC and the normal liver, respectively (Kondoh et al., 1999). Comparison of these profiles revealed that about 150 transcripts were expressed at significantly different levels (10-fold or greater). Many of these transcripts were examined by conventional Northern blotting using paired RNA samples from five independent HCC patients. Consistently higher levels of UDPglucuronosyltransferase (UGT2B4), ribosomal phosphoprotein P0 (rpP0), dek, vitronectin, galectin 4 (Gal-4) and insulin-like growth factor binding protein (IGFBP) 1 mRNAs combined with a lower level of retinoic acid-induced gene E (RIG-E) mRNA were observed. Expression of some of these genes appeared to be correlated with histological grading of tumors. The examination using HCC cell lines HuH-7 and HepG2 under different growth conditions suggested that the expression of dek mRNA was growth-associated. In contrast, the expression of Gal-4, UGT2B4, IGFBP-1, and RIG-E mRNAs was regulated in a cell density-dependent manner. It was also demonstrated that sodium butyrate, an inducer of differentiation, up-regulated and down-regulated RIG-E and dek mRNAs, respectively, in a dose dependent manner in HuH-7 cells, supporting in part above-mentioned pathological observations. These transcripts are differentially regulated depending on cell–cell contact, serum growth factors, growth and differentiation status, and/or other mechanisms in premalignant and malignant liver cells. Functional

Table 5
Highly differentially expressed genes in HIV-1-infected T cells

SAGE tag	Gene description	Accession no.	H/M[a]
0. HIV transcripts			
TGGGTCTCTCTGGT	HIV-1 transcript	Z11530	92/0
CTGAGGTGTGACTG	HIV-1 transcript	Z11530	9/0
CGTCAGCGTCATTG	HIV-1 transcript	Z11530	9/0
CCACAGACCCCAAC	HIV-1 transcript	Z11530	7/0
1. Cell activation and signaling pathway			
CAGCAGGCAGAGCC	Mitogen-activated protein kinase kinase 3 (MAPKK3)	D87 116	37/7
TCAGGAGGCTGAGG	Tumor necrosis factor receptor 75 kDa	S63368	7/1
AGGCGCTAATTGTT	Argininosuccinate synthetase (AS)	X01630	15/0
CAGAGGATGGTGAG	Fibroblast growth factor receptor (FGFR)	M60485	12/1
GCACCCGCTGGGCA	Lymphocyte activation antigen 4F2 large subunit	J03569	9/1
TAGCTGTGTGTTCT	Ca channel B3 subunit (CAL Bet 3)	L27584	6/0
AGACGGTGTGGGGG	Leukosialin (CD43)	M61827	5/0
2. Transcription factors			
TGAGACAGGGTGCT	Helix–loop–helix zipper protein	M77476	21/4
TGTGGGCTGTGCTG	Transcription factor ETR101	M62831	13/1
CAGGGCCATGCAGG	Basic-leucine zipper transcription factor MafG	U84249	11/1
TGAGATGTGGCTGG	Zinc finger protein (ZNF139)	U09848	6/0
CCCTCTGACCCACC	Ets domain protein ERF	U15655	5/0
AGCTCCGGACTCTT	GATA-3 enhancer-binding protein	M69106	5/0
3. INE-induced genes			
GGCCTCAAGCCCCT	Interferon-induced 17/15 kDa protein	M13755	33/6
CAGGGCAAGAAGCC	Interferon-inducible protein (IP-30)	J03909	12/1
AATGCTGCTGCCTT	Putative interferon-related protein (SM15)	U09585	6/0
4. Miscellaneous genes			
CAGTGTGTGTTGAT	EST 1	W86328	20/2
TGTCCATCTGCCTG	Rapamycin and FK506 binding protein	M75099	10/0
AGCCCCAGATGGGA	HIV-1 promoter region chimeric mRNA	U19178	5/0
CCTGTGTTTTACCT	Breast tumor autoantigen	U24576	5/0
TTCGCCGAGAGGGT	Cysteine-rich heart protein (CRHP)	U09770	5/0
CCGCCCATGAACCC	Moesin–ezrin–radixin-like protein	L11353	5/0

[a] H/M values indicate the frequency with which each tag appeared in the profiles from HIV (H)- and Mock (M)-infected MOLT-4 cells. The frequency of each tags was calculated within a total population of 71 462 tags and 71 147 tags sequenced from HIV-1- and mock-infected MOLT-4 cells, respectively.

analysis of these gene products should provide a wealth of information to further understand liver carcinogenesis.

8.3. HIV infected MOLT-4

HIV-1 infection alters cellular physiological states leading to disturbance of immune responses, T cell growth arrest and death, together with many other secondary effects, in which a variety of gene activities might be involved (Pantaleo and Fauci, 1996). Studies of HIV-1 pathogenesis have recently been expanded to define the changes in gene expression occurring in infected cells in association with HIV-1 replication and apoptosis (Hashimoto et al., 1997; Kaplan and Sieg, 1998; Scheuring et al., 1998). However, most of these studies have focused on only a limited number of biological parameters.

Table 6
Down-regulated genes in HIV-1-infected T cells

SAGE tag	Gene description	Accession no.	H/M[a]
1. Mitochondria and antioxidants			
TATACTTCACAACA	Mitochondrion cytochrome *b*	U09500	7/50
CCAGTGATCCCCAC	Cytochrome *c* oxidase subunit Vib (COXVib)	X54473	1/30
TGTCTCTCTCCTTG	ATP synthase b subunit	M27132	2/19
CACTGCTAATAAAT	Cytochrome *c* oxidase subunit IV (COX IV)	M21575	2/19
ACATAAGTTATTTC	ADP/ATP carrier protein	J02683	1/17
CAGGAAAGAGGATA	Mitochondrial aspartate aminotransferase	M22632	0/16
CTGGATGAAGCATA	Glutathione *S*-transferase homolog	U90313	2/16
ACAGACGAGCATGG	Thiol-specific antioxidant	X82321	0/12
TGAGACCTAGAGTC	ADP/ATP translocase	J03592	0/11
TCCTATGCAATATT	ADP-ribosylation factor 1	M84326	1/11
AAACCCACGTTTTG	Mitochondrial 75 kDa iron sulphur protein	X61100	0/9
TTTGCTCCATTGTT	150 kDa oxygen-regulated protein (ORP150)	U65785	0/8
2. Actin-related factors			
TGACCTCGTCTGTC	Proflilin	J03191	15/66
TCTGGTGAGTCACC	GTP-binding protein (rhoC)	L25080	2/32
GGGAGTTTCTTGGT	Arp2/3 protein complex subunit p34-Arc (ARC34)	AF006085	1/7
CATACATGAGTTAT	Actin-related protein Arp2 (ARP2)	AF006082	0/6
ACAATCATTTAATA	Rho GDP-dissociation inhibitor 2	X69549	0/5
3. Translational factors			
CCCTGGCCGTGTGT	Eukaryotic initiation factor 4A1	D13748	7/50
TCCAGAGGAGTGTG	Nucleolar protein hNop56	Y12065	2/17
CCAAGTCTTACGTT	Inducible poly(A)-binding protein	U33818	0/10
TGCTTCCAAGCAGC	DEAD box protein family	X70649	1/8
TCCTGTTTGGAAGT	Translation initiation factor nuk34	X79538	0/7
TACGTGAAACTGAA	Nuclear RNA helicase	Z37166	1/6
ACTTGCTGGTCTAG	Translation initiation factor 3	U94855	0/5
4. Miscellaneous genes			
CGATCCTGAGACCT	Ornithine decarboxylase (ODC1)	M16650	2/24
GCAAAGAGAACCAG	Cyclin AICDK2-associated p19 (Skpl)	U33760	1/17
TAACTTTCCTTCAT	Interleukin 2 receptor (IL2RG) g chain	D11086	2/12
TATAAGTAGTTGGT	Prothymosin a	M14630	1/12
GTGCTAACAGGCTC	Transferrin receptor	X01060	1/11
CCTGGGGAATCAAC	EST 2	AA825204	1/10
TGTCTGGCTTGGAT	RanGTP binding protein 5	Y08890	1/8
TTGGTAAGAGGGAG	Down syndrome critical region protein (DSCR1)	U28833	0/6
TTCGAATTTGAGTT	TGF-b receptor interacting protein 1	U36764	1/5

[a] Conditions are as described in Table 2.

To further analyze the cellular events and the pathogenesis occurring after HIV-1 infection, it is essential to survey an overall differential gene expression pattern. The gene expression profile of the HIV-1 infection state was thus analyzed in the human T cell line MOLT-4 (Ryo et al., 1999). A total of 142 603 tags representing 43 581 unique transcripts were sequenced and identified for HIV-1-infected and uninfected T cells. Comparison of the profiles revealed that 53 cellular genes were differentially expressed upon HIV-1 infection. Table 5 lists up-regulated genes. Among these, four transcripts were derived from virus genome itself and detected only in the profile from the infected T cells. The remaining 22 genes were tentatively categorized into four groups, namely; (1) genes related to cell activation and signaling pathways, (2) transcription factors, (3) interferon-induced genes and (4) miscellaneous genes. Overall, the up-regulated genes were mainly comprised of genes that accelerate HIV-1 replication.

On the other hand, 31 down-regulated genes were also classified into four groups as in Table 6; (1) mitochondrial proteins and antioxidants, (2) actin-related factors, (3) translational factors, and (4) miscellaneous genes. These genes were involved in anti-apoptotic cell defense, and regulation of basic cellular functions.

Although the study was performed using an immortalized T cell line and a laboratory clone of HIV-1 for the purpose of simplicity, this is the first systematic demonstration of changes in gene expression accompanied with HIV-1 infection. For the more practical understanding of the HIV-1–host relationship, similar analyses would definitely be needed using blood samples from the virus-infected individuals at various stages of infection.

9. Conclusions

SAGE is a general and powerful method that allows one, not only to obtain global gene expression profile of any kinds of eukaryotic cells, but also to identify genes specifically expressed in various physiological, developmental, and pathological states by simply comparing numbers of gene tags catalogued in the profile. However, it had several disadvantages. Current efforts for methodological improvement have not yet lead to a perfect solution. If we use this high-throughput method after a full understanding of these drawbacks, it will promise to provide much useful information on studies exploring virtually any kinds of biological phenomena in which the changes in cellular transcription are responsible.

Acknowledgements

This work was supported in part by Fujiyakuhin Co. Ltd. and the Epilepsy Research Foundation.

References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, C.R., Merril, C.R., Wu, A., Olde, B., Moreno, R.F. et al., 1991. Complementary DNA sequencing: expressed sequence tags and the human genome project. Science 252, 1651–1656.

Akagawa, K.S., Takasuka, N., Nozaki, Y., Komuro, I., Azuma, M., Ueda, M., Naito, M., Takahashi, K., 1996. Generation of CD^{1+} $RelB^{+}$ dendritic cells and tartrate-resistant acid phosphatase-positive osteoclast-like multinucleated giant cells from human monocytes. Blood 88, 4029–4039.

Bishop, J.O., Morton, J.G., Rosbach, M., Richardson, M., 1974. Three abundance classes in HeLa cell messenger RNA. Nature 250, 199–204.

Chen, H., Centola, M., Altschul, S.F., Metzger, H., 1998. Characterization of gene expression in resting and activated mast cells. J. Exp. Med. 188, 1657–1668.

de Waard, V., van den Berg, B.M., Veken, J., Schultz-Heienbrok, R., Pannekoek, H., van Zonneveld, A.J., 1999. Serial analysis of gene expression to assess the endothelial cell response to an atherogenic stimulus. Gene 226, 1–8.

Datson, N.A., van der Perk-de Jong, J., van den Berg, M.P., de Kloet, E.R., Vreugdenhil, E., 1999. MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. Nucleic Acids Res. 27, 1300–1307.

Fedoroff, S., Hao, C., 1991. Origin of microglia and their regulation by astroglia. Adv. Exp. Med. Biol. 296, 135–142.

Gasson, J.C., 1991. Molecular physiology of granulocyte-macrophage colony-stimulating factor. Blood 77, 1131–1145.

Goffeau, A.E.A., 1997. The yeast genome directory. Nature 387S, 5.

Gygi, P., Rochon, Y., Franza, B.R., Aebersold, R., 1999. Correlation between protein and mRNA abundance in yeast. Mol. Cell. Biol. 19, 1720–1730.

Hashimoto, F., Oyaizu, N., Karyaneraman, V.S., Pahwa, S., 1997. Modulation of Bcl-2 protein by CD4 cross-linking: a possible mechanism for lymphocyte apoptosis in human immunodeficiency virus infection and for rescue of apoptosis by interleukin-2. Blood 90, 745–753.

Hashimoto, S., Suzuki, T., Dong, H.Y., Nagai, S., Yamazaki, N., Matsushima, K., 1999a. Serial analysis of gene expression in human monocyte-derived dendritic cells. Blood 94, 845–852.

Hashimoto, S., Suzuki, T., Dong, H.Y., Yamazaki, N., Matsushima, K., 1999b. Serial analysis of gene expression in human monocytes and macrophages. Blood 94, 837–844.

Hashimoto, S., Yamada, M., Yanai, N., Kawashima, T., Motoyoshi, K., 1996. Phenotypic change and proliferation of murine Kupffer cells by colony-stimulating factors. J. Interferon Cytokine Res. 16, 237–243.

Hereford, L.M., Rosbach, M., 1977. Number and distribution of polyadenylated RNA sequences in yeast. Cell 10, 453–462.

Hibi, K., Liu, Q., Beaudry, G.A., Madden, S.L., Westra, W.H., Wehage, S.L., Yang, S.C., Heitmiller, R.F., Bertelsen, A.H., Sidransky, D., Jen, J., 1998. Serial analysis of gene expression in non-small cell lung cancer. Cancer Res. 58, 5690–5694.

Hubank, M., Schatz, D.G., 1994. Identifying differences in mRNA expression by representational difference analysis of cDNA. Nucleic Acids Res. 22, 5640–5648.

Imai, F., Sawada, M., Suzuki, H., Kiya, N., Hayakawa, M., Nagatsu, T., Marunouchi, T., Kanno, T., 1997. Migration activity of microglia and macrophages into rat brain. Neurosci. Lett. 237, 49–52.

Inoue, H., Sawada, M., Ryo, A., Tanahashi, H., Wakatsuki, T., Hada, A., Kondoh, N., Nakagaki, K., Takahashi, K., Suzumura, A., Yamamoto, M., Tabira, T., 1999. Serial analysis of gene expression in a microglial cell line. Glia 28, 265–271.

Ji, X., Chen, D., Xu, C., Harris, S.E., Mundy, G.R., Yoneda, T., 2000. Patterns of gene expression associated with BMP-2-induced osteoblast and adipocyte differentiation of mesenchymal progenitor cell 3T3-F442A. J. Bone Miner. Metab. 18, 132–139.

Kal, A.J., van Zonneveld, A.J., Benes, V., van den Berg, M., Koerkamp, M.G., Albermann, K., Strack, N., Ruijter, J.M., Richter, A., Dujon, B., Ansorge, W., Tabak, H.F., 1999. Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. Mol. Biol. Cell 10, 1859–1872.

Kaplan, D., Sieg, S., 1998. Role of the Fas/Fas ligand apoptotic pathway in human immunodeficiency virus type 1 disease. J. Virol. 72, 6279–6282.

Kavathas, P., Sukhatme, V.P., Herzenberg, L.A., Parnes, J.R., 1984. Isolation of the gene encoding the human T-lymphocyte differentiation antigen Leu-2 (T8) by gene transfer and cDNA subtraction. Proc. Natl. Acad. Sci. USA 81, 7688–7692.

Kenzelmann, M., Muhlemann, K., 1999. Substantially enhanced cloning efficiency of SAGE (Serial Analysis of Gene Expression) by adding a heating step to the original protocol. Nucleic Acids Res. 27, 917–918.

Ko, M.S., 1990. An equalized cDNA library by the reassociation of short double-stranded cDNAs. Nucleic Acids Res. 19, 5705–5711.

Kondoh, N., Wakatsuki, T., Ryo, A., Hada, A., Aihara, T., Horiuchi, S., Goseki, N., Matsubara, O., Takenaka, K., Shichita, M., Tanaka, K., Shuda, M., Yamamoto, M., 1999. Identification and characterization of genes associated with human hepatocellular carcinogenesis. Cancer Res. 59, 4990–4996.

Lal, A., Lash, A.E., Altschul, S.F., Velculescu, V., Zhang, L., McLendon, R.E., Marina, M.A., Prange, C., Morin, P.J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K.W., Strausberg, R.L., Riggins, G.J., 1999. A public database for gene expression in human cancers. Cancer Res. 59, 5403–5407.

Lawson, L.J., Perry, V.H., Dri, P., Gordon, S., 1990. Heterogeneity in the distribution and morphology of microglia in the normal adult mouse brain. Neuroscience 39, 151–170.

Liang, P., Pardee, A.B., 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257, 967–971.

Madden, S.L., Galella, E.A., Zhu, J., Bertelsen, A.H., Beaudry, G.A., 1997. SAGE transcript profiles for p53-dependent growth regulation. Oncogene 15, 1079–1085.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, E.L., 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnol. 14, 1675–1680.

Marchuk, D., Drumm, M., Saulino, A., Collins, F.S., 1991. Construction of T-vectors, a rapid and general system for direct cloning of unmodified PCR products. Nucleic Acids Res. 19, 1154.

Matsuda, S., Akagawa, K.S., Honda, M., Yokota, Y., Takebe, Y., Takemori, T., 1995. Suppression of HIV replication in human monocyte-derived macrophages induced by granulocyte/macrophage colony-stimulating factor. AIDS Res. Hum. Retroviruses 11, 1031–1038.

Matsumura, H., Nirasawa, S., Terauchi, R., 1999. Technical advance: transcript profiling in rice (*Oryza sativa* L.) seedlings using serial analysis of gene expression. Plant J. 20, 719–726.

Neilson, L., Andalibi, A., Kang, D., Coutifaris, C., Strauss III, J.F., Stanton, J.-A.L., Green, D.P.L., 2000. Molecular phenotype of the human oocyte by PCR-SAGE. Genomics 63, 13–24.

Okubo, K., Hon, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., Matsubara, K., 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat. Genet. 2, 173–179.

Palucka, K.A., Taquet, N., Sanchez-Chapuis, F., Gluckman, J.C., 1998. Dendritic cells as the terminal stage of monocyte differentiation. J. Immunol. 160, 4587–4595.

Pantaleo, G., Fauci, T., 1996. Immunopathogenesis of HIV infection. Annu. Rev. Microbiol. 50, 825–854.

Pauws, E., Moreno, J.C., Tijssen, M., Baas, F., de Vijder, J.J., Ris-Stalpers, C., 2000. Serial analysis of gene expression as a tool to assess the human thyroid expression profile and to identify novel thyroidal genes. J. Clin. Endocrinol. Metab. 85, 1923–1927.

Pennisi, E., 2000. And the gene number is . . . ? Science 288, 1146–1147.

Peters, D.G., Kassam, A.B., Yonas, H., O'Hare, E.H., Ferrell, R.E., Brufsky, A.M., 1999. Comprehensive transcript analysis in small quantities of mRNA by SAGE-Lite. Nucleic Acids Res. 27, 39.

Polyak, K., Xia, Y., Zweier, J.L., Kinzler, K.W., Vogelstein, B.,

1997. A model for p53 induced apoptosis. Nature 389, 300–305.

Powell, J., 1998. Enhanced concatemer cloning: a modification to the SAGE (Serial Analysis of Gene Expression) technique. Nucleic Acids Res. 26, 3445–3446.

Ryo, A., Kondoh, N., Wakatsuki, T., Hada, A., Yamamoto, N., Yamamoto, M., 1998. A method for analyzing the qualitative and quantitative aspects of gene expression: a transcriptional profile revealed for HeLa cells. Nucleic Acids Res. 26, 2586–2592.

Ryo, A., Kondoh, N., Wakatsuki, T., Hada, A., Yamamoto, N., Yamamoto, M., 2000. A modified serial analysis of gene expression that generates longer sequence tags by nonpalindromic cohesive linker ligation. Anal. Biochem. 277, 160–162.

Ryo, A., Suzuki, Y., Ichiyama, K., Wakatsuki, T., Kondoh, N., Hada, A., Yamamoto, M., Yamamoto, N., 1999. Serial analysis of gene expression in HIV-1-infected T cell lines. FEBS Lett. 462, 182–186.

Sallusto, F., Lanzavecchia, A., 1994. Efficient presentation of soluble antigen by cultured human dendritic cells is maintained by granulocyte/macrophage colony-stimulating factor α. J. Exp. Med. 179, 1109–1118.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270, 467–470.

Scheuring, U.J., Corbeil, J., Mosier, D.E., Theofilopoulos, A.N., 1998. Early modification of host cell gene expression induced by HIV-1. AIDS 12, 563–570.

Streit, W.J., Graeber, M.H., Kreutzberg, G.W., 1988. Functional plasticity of microglia: a review. Glia 1, 301–307.

Tang, W., Lai, Y.H., Han, X.D., Wong, P.M.C., Peters, L.L., Chui, D.H.K., 1997. Murine Hn1 on chromosome 11 expressed in hematopoietic and brain tissues. Mamm. Genome 8, 695–696.

Theele, D.P., Streit, W.J., 1993. A chronicle of microglial ontogeny. Glia 7, 5–8.

Tushinski, R.J., Oliver, I.T., Guilbert, L.J., Tynan, P.W., Warner, J.R., Stanley, E.R., 1982. Survival of mononuclear phagocytes depends on a lineage-specific growth factor that the differentiated cells selectively destroy. Cell 28, 71–81.

van den Berg, A., van der Leij, J., Poppema, S., 1999a. Serial analysis of gene expression: rapid RT-PCR analysis of unknown SAGE tags. Nucleic Acids Res. 27, e17.

van den Berg, A., Visser, L., Poppema, S., 1999b. High expression of the CC chemokine TARC in Reed-Steinberg cells. A possible explanation for the characteristic T-cell infiltrate in Hodgkin's lymphoma. Am. J. Pathol. 154, 1685–1691.

Velculescu, V.E., 1999. Tantalizing Transcriptomes-SAGE and its use in global gene expression analysis. Science 286, 1491–1492.

Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W., 1995. Serial analysis of gene expression. Science 270, 484–487.

Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Hieter, P., Vogelstein, B., Kinzler, K.W., 1997. Characterization of the yeast transcriptome. Cell 88, 243–251.

Virlon, B., Cheval, L., Buhler, J.M., Billon, E., Doucet, A., Elalouf, J.M., 1999. Serial microanalysis of renal transcriptomes. Proc. Natl. Acad. Sci. USA 96, 15286–15291.

Welle, S., Bhatt, K., Thornton, C.A., 1999. Inventory of high-abundance mRNAs in skeletal muscle of normal men. Genome Res. 9, 506–513.

Welsh, J., Chada, K., Dalal, S.S., Cheng, R., McClelland, M., 1992. Arbitrarily primed PCR fingerprinting of RNA. Nucleic Acids Res. 20, 4965–4970.

Yamamoto, M., Maehara, Y., Takahashi, K., Endo, H., 1983. Cloning of sequences expressed specifically in tumors of rat. Proc. Natl. Acad. Sci. USA 80, 7524–7527.

Yamashita, T., Hashimoto, S., Kaneko, S., Nagai, S., Toyoda, N., Suzuki, T., Kobayashi, K., Matsushima, K., 2000. Comprehensive gene expression profile of a normal human liver. Biochem. Biophys. Res. Commun. 269, 110–116.

Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B., Kinzler, K.W., 1997. Gene expression profiles in normal and cancer cells. Science 276, 1268–1272.