Review TRENDS in Genetics Vol.23 No.1

Full text provided by www.sciencedirect.com
ScienceDirect

Understanding SAGE data

San Ming Wang

Center for Functional Genomics, ENH Research Institute, Robert H. Lurie Comprehensive Cancer Center, Northwestern University, 1001 University Place, Evanston, IL 60201, USA

Serial analysis of gene expression (SAGE) is a method for identifying and quantifying transcripts from eukaryotic genomes. Since its invention, SAGE has been widely applied to analyzing gene expression in many biological and medical studies. Vast amounts of SAGE data have been collected and more than a thousand SAGE-related studies have been published since the mid-1990s. The principle of SAGE has been developed to address specific issues such as determination of normal gene structure and identification of abnormal genome structural changes. This review focuses on the general features of SAGE data, including the specificity of SAGE tags with respect to their original transcripts, the quantitative nature of SAGE data for differentially expressed genes, the reproducibility, the comparability of SAGE with microarray and the future potential of SAGE. Understanding these basic features should aid the proper interpretation of SAGE data to address biological and medical questions.

Introduction

Perhaps the best-known way to analyze gene expression is by microarray. This method relies, however, on knowledge of the sequence of either the whole genome or individual cDNAs [i.e. full-length, partial or expressed sequence tags (ESTs)] for probe design. Serial analysis of gene expression (SAGE) is a method invented in 1995 for transcriptome study [1] (Box 1). Similar to microarrays, SAGE provides quantitative information on gene expression but, unlike microarrays, SAGE detects unknown transcripts because it does not require prior knowledge of what is present in the sample under analysis.

Recently, conventional SAGE has been developed further into different methods, such as LongSAGE, SuperSAGE, cap analysis of gene expression (CAGE), gene identification signature (GIS) and others, for specific purposes (Box 2). SAGE has been widely applied to biological and medical studies. So far, more than 64 million SAGE tags have been collected from many species, of which nearly 30 million are from the humans (Table 1). Understanding SAGE is crucial for translating the vast amount of data being collected into biological and medical means. By taking the standard human 14-bp SAGE tags and other types of tag as examples, here I discuss the basic features of SAGE data.

What information does a typical SAGE data set provide?

A typical SAGE data set collected from a SAGE library generally provides the following information:

(i) It collects 50 000 to 100 000 SAGE tags, which represent 20 000 to 40 000 unique SAGE tags. The unique SAGE tags provide qualitative information showing the transcripts detected by SAGE tags; the copy number of each unique SAGE tag provides quantitative information showing the abundance of the transcripts detected by SAGE tags.

(ii) Several hundreds of the unique SAGE tags have more than 100 copies and several thousand of the unique SAGE tags have 2–100 copies. The remaining SAGE tags, which account for 70–80% of the total SAGE data set, have a single copy.

(iii) Typically, 50–70% of the SAGE tags match known transcripts or genes, whereas 30–50% of SAGE tags have no match to known transcripts or genes.

How specific is a SAGE tag for its original transcript?

A standard SAGE tag is the 14-bp sequence located immediately after the last defined restriction site in the 3′ part of the detected transcript. This restriction site is used to release SAGE tags from cDNA templates. The restriction site that is used the most is the *Nla*III site CATG.

Of the standard 14-bp SAGE tags matched to known transcripts, the specificity is relatively high in simple genomes. For example, 96.4% of mapped *Drosophila* SAGE tags are unique to known *Drosophila* transcripts or genes [2]. The specificity of the tags decreases, however, in more complicated genomes. For human SAGE tags, a third of the mapped SAGE tags are shared by different transcripts or genes [3,4]. Attempts to improve the specificity have been made using highly qualified transcript sequences as the references for SAGE tag mapping [2,5]. The increase in mapping specificity achieved by this approach might be artificial, however, because the actual composition of the transcriptome is far more complicated than these highly qualified sequences.

Attempts have also been made by increasing the tag length; for example, the LongSAGE approach increases the length of the SAGE tag from 14 to 21 bp [6], and the SuperSAGE approach increases the tag length from 14 to 26 bp [7]. Although longer tags indeed increase proportionally both the specificity of a SAGE tag to represent a single transcript or gene and the specificity of a SAGE tag to map uniquely in the genome, they do not solve the problem of specificity completely. In addition, a major drawback of increasing the tag length is that the mapping rate of LongSAGE tags is markedly decreased. For example, of the 632 813 human LongSAGE tags, only 137 333 (22%) can be mapped to the SAGEmap database (http://www.ncbi.

Corresponding author: Wang, S.M. (swang1@northwestern.edu).
Available online 15 November 2006.

 0168-9525/$ – see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2006.11.001

Box 1. SAGE

Serial analysis of gene expression (SAGE) is a sequencing-based method for gene expression profiling that facilitates the global and quantitative characterization of a transcriptome (Figure I). By using a type II restriction enzyme, *Bsm*FI, as the tag-releasing enzyme, SAGE extracts 14-bp fragments or 'tags' after the last restriction site (mostly the *Nla*III site CATG) in cDNA templates. These tags are ligated together ('concatemerization') into a longer fragment ('concatemer'), which is then cloned for DNA sequencing. The tags detected in a sample represent the parent transcripts, and the frequency of detection of each tag represents the quantity of the transcript detected.

The gene origin of a SAGE tag is determined by positively matching the SAGE tag to sequences in a pre-constructed reference database that contains virtual tags extracted from known transcript sequences. Negative mapping of a tag to sequences in the reference database suggests that the SAGE tag has detected a novel transcript. Sequencing short tags facilitates higher sensitivity than other technologies for transcript detection. A detailed description of the different aspects of SAGE is given in Ref. [45].

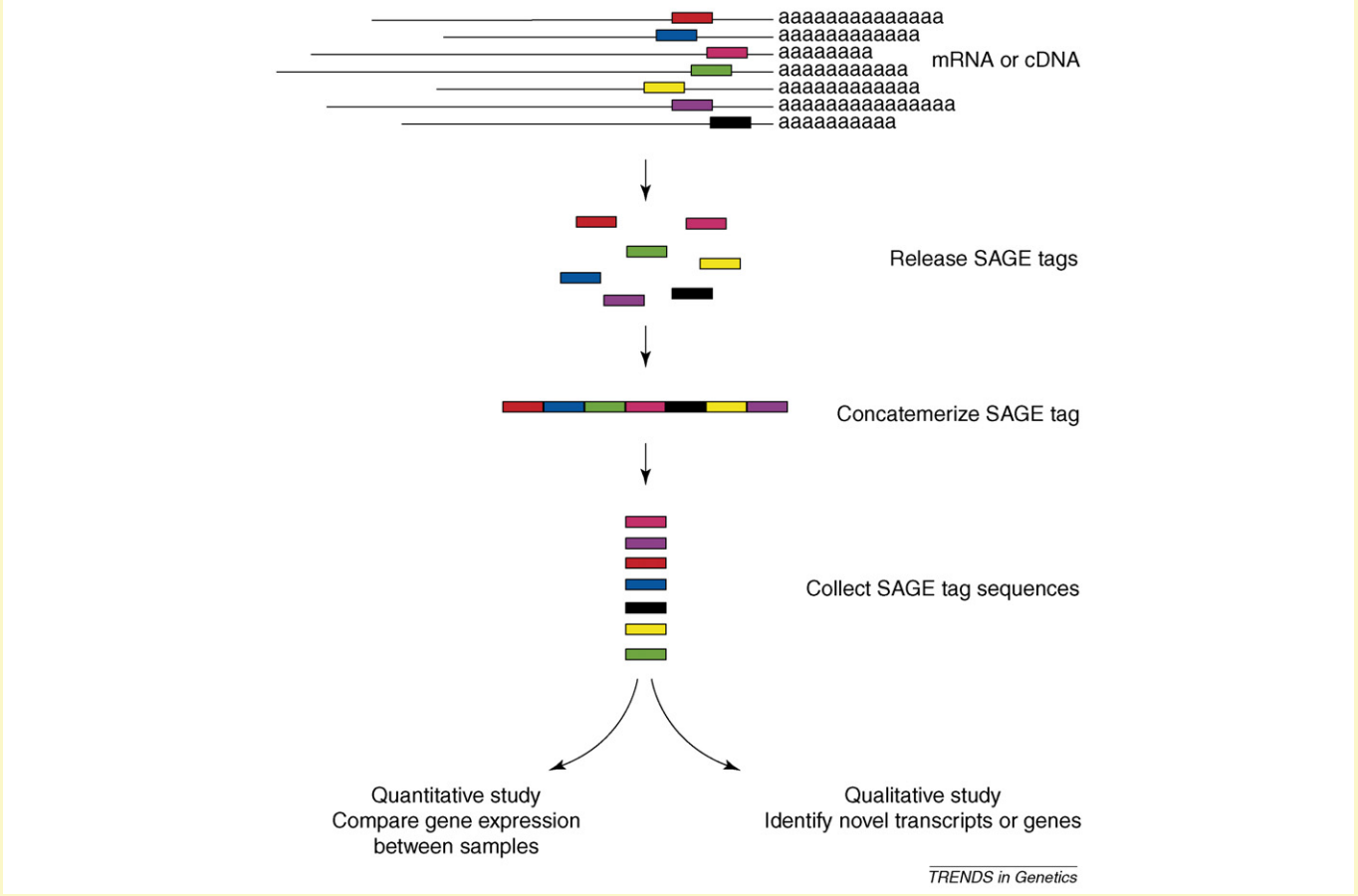

Figure I. Schematic representation of SAGE process.

nlm.nih.gov/SAGE) [5]. This finding is probably related to the increased probability of incorporating base errors or single nucleotide polymorphisms in a long tag as compared with a short tag. In addition, longer tags also increase the sequencing cost and decrease the throughput capacity of SAGE [8]. Thus, users need to consider these factors when choosing the type of tag for their own studies. Although SAGE tags have lower specificity for their original transcripts than do longer sequences such as ESTs or full-length cDNAs, their advantage is a gain in sensitivity over other approaches for transcript detection.

How reliable do the genes assigned to SAGE tags by reference databases?

Unlike longer sequences such as ESTs, the limited length of the SAGE tag does not enable it to be used directly to search known transcript sequences for gene identification. Instead, SAGE relies on a pre-constructed SAGE reference database to determine the gene origin of SAGE tags. The reference database contains virtual SAGE tags extracted after the last CATG site in the sequence of known transcripts or genes of a particular species. If a SAGE tag matches a virtual SAGE tag, the transcript or gene that contributed the virtual reference tag is assigned to the SAGE tag.

Several SAGE reference databases, such as SAGEmap [5] and SAGE Genie (http://cgap.nci.nih.gov/SAGE) [9], have been constructed and are used widely to determine the gene origin of SAGE tags. Although the gene origin can be assigned reliably for many SAGE tags through these reference databases, the accuracy of the gene origin determined for some SAGE tags cannot be guaranteed. For example, one study has estimated that 20% of the genes assigned to SAGE tags could be incorrect [10]. The following factors might influence the reliability of reference databases.

Box 2. Further developments of SAGE

The standard SAGE technique collects 14-bp tags and has been widely used in transcriptome studies. The principle of standard SAGE has been adapted to several new platforms for studying different topics.

- Generation of longer 3′ cDNA from SAGE tags for gene identification (GLGI) [13] aims to increase the specificity of SAGE tags. It uses the SAGE tags as the sense primer to extend the SAGE tags into the 3′ end of the cDNA, which is then amplified by PCR together with a universal antisense primer at the 3′ end of cDNA. Through this process, a SAGE tag is converted into a 3′ EST with up to hundreds of bases.
- LongSAGE [6] aims to increase the specificity of SAGE tags for transcript identification and for mapping SAGE tags to the genome. LongSAGE is a modification of standard SAGE. By using a different type II restriction enzyme, *Mme*I, as the tag-releasing enzyme, LongSAGE collects tags of 21 bp. The longer length of these SAGE tags provides higher specificity for the original transcripts and increases the unique mapping frequency of the tags to the genome. LongSAGE has a crucial role in adapting the SAGE principle to studies of genomic DNA.
- Cap analysis gene expression (CAGE) [21,46] aims to identify transcriptional initiation sites and promoters. It collects 21-bp SAGE tags from the 5′ ends of cap-purified cDNAs. The information can be used to identify transcriptional initiation sites and to locate gene promoters in the genome. CAGE has been used in mouse and human transcriptome studies [47].
- Gene identification signature (GIS) [22,48] aims to identify the gene boundary. It collects 20-bp LongSAGE tags from both the 5′ and the 3′ end of the same transcript to determine the 5′ and 3′ end of the transcripts detected in the genome. GIS has been applied to human and mouse transcriptome studies [47,48].
- SuperSAGE [7] aims to increase the specificity of SAGE tags and to use the tags directly as a microarray probe. By using a type III restriction endonuclease, *Eco*P15I, for tag releasing, SuperSAGE collects 26-bp tags. SuperSAGE tags increase the specificity of SAGE tags for transcript identification and genome mapping, and can be applied directly as probes for microarray design [23]. This approach is particularly valuable for studying gene expression in the genomes that have no EST or genomic sequence information. SuperSAGE has been used in plant SAGE studies [23].
- Digital karyotyping [49] aims to analyze genome structure. It adapts the LongSAGE protocol to collect 21-bp tags from genomic DNA. Digital karyotyping has been used to identify amplification and deletion in several types of cancer [49,50].
- Paired-end ditag [51] aims to identify protein-binding sites in the genome. It uses the GIS principle to collect tags from both ends of the DNA templates bound by specific proteins isolated through immunoprecipitation. Paired-end ditag has been applied to the identification of p53-binding sites in the human genome [51].

(i) Biological factors. Transcript isoforms from the same gene origin, such as alternatively spliced transcripts, might contribute different SAGE tags, and single nucleotide polymorphisms located at the SAGE tag can give rise to different tags for the same gene transcript from different individuals [11].

(ii) Experimental factors. The process of SAGE tag collection involves many steps. Each step could introduce experimental artifacts. For example, incomplete digestion of cDNA templates could result in the generation of different SAGE tags for the same transcript.

(iii) Transcript redundancy. The transcript sequences deposited in the expression databases are highly redundant. To construct a SAGE reference database, multiple transcripts expressed from the same gene need to be clustered or grouped. This grouping is mainly achieved through sequencing alignment among the transcript sequences or between the transcript sequences and the genome sequence. However, this process is by no means easy; for example, transcripts from different genes might be clustered together owing to high sequence similarities, whereas different transcripts from the same gene might be separated into distinct clusters owing to high sequence differences.

(iv) Incomplete SAGE reference databases. Many known transcript sequences are not included in the SAGE reference database. For example, of the 7.7 million human ESTs collected in the Database of ESTs (dbEST; http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html), 1.3 million are not included in the human UniGene Database (UniGene Build 189; http://www.ncbi.nlm.nih.gov/UniGene), which is the basis for constructing the SAGEmap reference database. As discussed above, the use of better quality sequences, such as well-annotated mRNA sequences, to construct the reference database can provide higher specificity for gene mapping with SAGE tags. Such an approach, however, eliminates even more known transcripts for mapping. Although some of the sequences excluded might have poor quality or lack orientation information, many of them could represent transcripts that contributed SAGE tags. As a result, eliminating these sequences might lead to the classification of experimental SAGE tags as unknown tags even though their parent transcripts have been identified.

(v) Low-specificity of SAGE tags for their original transcripts. The same SAGE tag can be shared by more than one gene [3,4]. Identifying the correct gene out of multiple genes for a SAGE tag solely on the basis of the short tag sequence is akin to gambling. Increasing the tag length improves the specificity but, as discussed above, also has some side-effects.

In view of these factors, the transcript or gene assigned should be verified by another approach. Of the limited methods available, 3′ RACE (rapid amplification of cDNA ends) [12] and GLGI (generation of long cDNA from SAGE tags for gene identification; Box 2) [13] can convert a SAGE tag into a 3′ EST by extending the tag to the 3′ end of the corresponding cDNA template. The increased length provides higher specificity to determine the gene origin of the SAGE tag. To identify the correct gene out of multiple genes that share the same SAGE tag sequence, a microarray-based reference database has also been developed that uses microarray-detected transcripts as the reference to annotate SAGE tags collected from the same tissue type (www.basic.northwestern.edu/SAGE) [4].

What is the origin of unknown SAGE tags?

In a given SAGE data set, up to half of the SAGE tags will have no match to known transcripts or genes [4]. The origin of these unmapped SAGE tags remains debatable and the following two explanations have been proposed.

First, the unmapped tags could largely result from an accumulation of sequencing error products [14,15]. Considering the many steps involved in SAGE tag collection,

Table 1. Major types of SAGE data holding in NCBI and RIKEN SAGE databases

Organism	Library	Types of tag (bp)			Total
		Standard (14)	LongSAGE (21)	5′ CAGE (21)	
Homo sapiens	350	16 138 558			
	41[a]			10 165 217	29 921 847
	29		3 618 072		
Mus musculus	202		12 830 383		
	145[a]			11 567 973	28 123 257
	112	3 724 901			
Caenorhabditis elegans	12	1 464 148			1 928 482
	5		464 334		
Rattus norvegicus	24	893 339			893 339
Oryza sativa	4		805 823		805 823
Bos taurus	11	609 110			609 110
Drosophila melanogaster	5	489 140			489 140
Magnaporthe grisea	2		484 021		484 021
Zea mays	1		232 948		368 519
Zea mays	1	135 571			
Arabidopsis thaliana	7	248 659			248 659
Gallus gallus	2		129 568		156 243
	2	26 675			
Pinus taeda	2	150 885			150 885
Medicago truncatula	3	131 599			131 599
Bombyx mori	2	126 557			126 557
Meleagris gallopavo	2	95 325			95 325
Sus scrofa	5	84 780			84 780
Drosophila pseudoobscura	2	52 040			52 040
Palaemonetes pugio	4		37 153		37 153
Danio rerio	1	27 486			27 486
Leishmania donovani	1	20 299			20 299
Lentinula edodes	7	19 586			19 586
Total (%)	984	24 438 658 (38)	18 602 302 (29)	21 733 190 (34)	64 774 150 (100)

[a]From Riken (http://fantom3.gsc.riken.jp/); all others are from the NCBI (http://www.ncbi.nlm.nih.gov/projects/geo).

and in particular the errors introduced by single-pass DNA sequencing, many SAGE tags will contain base errors and thus will not be able to map to their known transcripts. This problem is particularly serious for SAGE tags with lower copy numbers. Therefore, these unmapped tags should be eliminated from further analysis.

Second, the unmapped tags could largely comprise true tags representing unknown low-abundance transcripts detected by SAGE owing to its high sensitivity [16]. This explanation is based on the known prevalence of low-abundance transcripts and on the belief that error tags exist but in small numbers owing to the fact that many SAGE data have been collected by the state-of-the-art Big-Dye sequencing system operated in large genome centers or genome industries, where sequence quality control such as Phred20 are routinely used to guarantee the high quality of sequences. Therefore, the unmapped tags should be reserved for further study.

Although the first explanation provides high confidence for the remaining tags, it ignores the existence of unknown transcripts present in lower abundance. The second explanation provides higher coverage of the transcriptome by taking advantage of the high sensitivity of SAGE to detect the total transcriptome in finer detail. It is supported by data showing that a moderate number of novel transcripts have been identified from unmapped SAGE tags [16], and it also fits well with data showing that unknown transcripts are widely present in many model genomes [17–19]. A key to solve the dispute is to distinguish between the true and the error SAGE tags. The following approaches could be helpful, although in themselves they are not trivial.

(i) Using genome mapping as the standard. Although useful, this approach does not definitively determine whether tags are true or erroneous. The human genome sequences (HG17) contain 26 201 271 CATG sites, which contribute 957 056 unique 14-bp genomic tags and 19 618 123 unique 21-bp genomic long tags. The specificity of the 14-bp tag is low: on average, there are 27 locations per tag in the genome; therefore, genome mapping cannot be used for the 14-bp SAGE tags. The specificity of the 21-bp genomic tag is higher: on average, there are 1.3 locations per tag in the genome. However, the mapping rate of LongSAGE tags to the genome is low: of the 632 813 LongSAGE tags in the database (http://www.ncbi.nlm.nih.gov/geo/), only 224 867 (36.5%) can be perfectly mapped to the genome [4]. In addition to the issue of SAGE tag sequences, the high degree of variation between different individual genomes might also influence SAGE tag mapping. The reference human genome sequences originated from a few individual genomes – namely, 66.3% from one individual, 18.6% from seven individuals and the remainder from different resources [20] – whereas the SAGE tags collected in experiments are from various individuals whose genomes could be substantially different from those that contributed to the reference human genome sequences. The lower mapping frequency also exists for other types of tag. About 40% of mouse 5′ CAGE tags, 86% of mouse GIS 5′ tags and 50% of mouse GIS 3′ tags cannot be mapped directly to the mouse genome [21,22]. Therefore,

positive mapping of a tag to the genome might indicate that a SAGE tag is real, but negative mapping is not a reason to reject a SAGE tag as erroneous.

(ii) Using the copy number of SAGE tags as the cut-off. Typically, single-copy SAGE tags account for 70–80% of the total number of tags in a given SAGE data set, and most of these single-copy tags could represent unknown lower abundance transcripts. Like throwing the baby out of the bath water, removing single-copy SAGE tags loses the information for those unknown transcripts. The remaining higher-copy SAGE tags are mainly those representing well-known transcripts and genes and will provide data similar to those obtained from microarrays. With this approach, the high sensitivity of SAGE for detecting unknown transcripts, of which most are at low abundance, is diminished.

(iii) Using high standards to control sequence quality. Most SAGE sequences are collected by single-pass sequencing reactions using an automatic DNA sequencer. Although standard sequencing quality controls, such as Phred20, are routinely applied in sequencing facilities, these controls do not prevent the errors generated before the sequencing reaction stage, such as those introduced during PCR. Ideally, it would be helpful to identify error sequences by applying the approach used in genome sequencing – namely, to provide multiple coverage of the genome in order to exclude error bases. Practically, however, it is very difficult to use this approach in transcriptome studies. For a given genome, the size is known and therefore the fold coverage can be pre-determined; for most genomes, however, the size of the transcriptome is unknown and is probably much larger than its corresponding genome. Thus, the cost required to provide multiple coverage of the transcriptome would be prohibitive.

(iv) Experimental verification. This could be the best approach to determine the gene origin of SAGE tags. However, SAGE tags have only a short length that is not suitable as a probe for confirmation studies using hybridization-based approaches (although the Super-SAGE tag can be an exception owing to its longer 26-bp length [23]). The SAGE tag is also not suitable for direct confirmation by PCR because it provides only one primer and no other sequences for designing the other primer (although with two tags available from the same transcript, the GIS $5' \rightarrow 3'$ tags might be suitable for this approach [22]). In addition, the number of candidate SAGE tags in a given SAGE data set can range from hundreds to thousands; thus, a high-throughput approach will be needed. However, few technologies currently exist that are specifically designed for large-scale SAGE tag confirmation.

Quantitative issues

Does the copy number of SAGE tags reflect the quantity of the detected transcripts?

The quantitative distribution of a given SAGE data set shows three classes of transcript: high, intermediate and low abundance. This pattern fits well with that of RNA reassociation data [24]. It is unclear, however, whether the copy number of each SAGE tag accurately reflects the absolute quantity of the transcripts detected in the sample. The SAGE process involves many PCR amplifications, cloning and colony propagations. These treatments could result in a quantitative bias for different tags. Indeed, studies show that SAGE data are bias to a high G + C content [25,26].

Evaluating the degree of bias is difficult when other technical platforms are used because each platform itself has an inherent bias. For example, microarrays have reproducibility problems [27]. We therefore have to accept with caution the copy number of SAGE tags as a quantitative measure for the transcripts detected. However, the quantitative issue mainly affects the use of SAGE for identifying the candidate genes between different samples: for transcript discovery using SAGE, qualitative rather than quantitative issues are more important.

How reliable is the differentially expressed gene identified by SAGE?

One of the principal applications of SAGE is to identify differentially expressed genes in samples from different physiological or pathological conditions (see websites http:cgap.nci.nih.gov/SAGE and http://www.SAGEnet.org/pubs/index.html). Many statistical methods have been applied to identify such candidate SAGE tags between different samples, including Poisson approximation [28], Bayesian method [29], and the Chi-square test [30]. The user-friendly IDEG6 website provides major statistical programs specifically designed for online SAGE data analysis (http://telethon.bio.unipd.it/bioinfo/IDEG6_form/) [31]. Each statistical method has its advantages for SAGE data analysis, but the candidate SAGE tags identified by different statistical methods for the same SAGE data set might differ.

To determine which statistical result to believe, biological factors must be considered. For example, SAGE tags with high copy numbers are frequently present at significantly different levels between samples. However, these SAGE tags represent mainly housekeeping genes and the statistical significance of their differential expression between different samples might have less biological interest. By contrast, SAGE tags representing many functional important genes might differ by marginal amounts between samples. Whether such statistically insignificant changes have any biological meaning needs to be considered. This issue is relevant not only to SAGE data but also to data collected by other platforms.

How reproducible is SAGE analysis?

Two aspects of reproducibility occur: qualitative reproducibility, whereby the same tag is detected repeatedly in different experiments; and quantitative reproducibility, whereby the same tag is detected at a similar frequency in different experiments. In comparison to the progress achieved in applying SAGE to various studies, limited efforts have been directed towards addressing the issue of SAGE reproducibility. A possible explanation might be related to the high cost of sequencing SAGE tags.

Table 2. Reproducibility of SAGE data collected from the same tissue or cell types

Library[a]		Human		Mouse	
		Colon	MCF7 cells	Testis	Visual cortex
1	GEO ID	GSM728	GSM752	GSM34768	GSM34030
	Total tag	50 179	60 725	51 879	109 125
	Unique tags	17 913	17 213	18 848	37 337
2	GEO ID	GSM729	GSM753	GSM45170	GSM34004
	Total tag	49 593	60 162	120 136	109 066
	Unique tags	16 569	17 821	44 998	35 029
Overlap (%)		5210 (30)	5210 (42)	9021(42)	10 275 (28)
R		0.92	0.94	0.68	0.92
R^2		0.84	0.88	0.47	0.84

[a]The SAGE libraries were from the NCBI (http://www.ncbi.nlm.nih.gov/geo/).

There are no published reports showing a cross-comparison of SAGE data from different laboratories for the same RNA sample. Several individual laboratories have tested reproducibility within their own laboratory by repeating the SAGE process on the same RNA sample (technical replicates) [32–34]. These studies are relatively consistent, showing that ~30–40% of SAGE tags are reproducibly detected. The SAGE tags that are reproducibly detected tend to be those with higher copy numbers; the reproducibility is poor for SAGE tags with low copy numbers.

Similar results have been found for the comparison of SAGE tags collected from the same tissue type but not the same RNA sample (biological replicates; Table 2). It has been proposed that 300 000 tags must be collected to detect transcripts with more than three copies per cell at 92% chance, with the assumption that there are 300 000 transcripts per cell [35]. Most SAGE studies collect 50 000 to 100 000 tags per SAGE library. At this level of tag collection, the frequency of detecting a particular SAGE tag will depend on the copy number of that SAGE tag in the total tag population. SAGE tags with higher copy numbers have greater chances of being detected; by contrast, SAGE tags with low copy numbers are detected more randomly. This explains the low overlapping frequency for low-copy SAGE tags. Increasing the total number of tags collected is the only way to increase the overlapping frequency and therefore the reproducibility. Fortunately, most of the functional SAGE studies aim to identify candidate genes in the system studied at

Box 3. Comparison of SAGE and microarray for gene expression studies

Several factors must be considered when comparing SAGE and microarray data (Table I).

- The different regions of transcripts detected by SAGE and microarray. SAGE specifically targets the 3′ region of the detected transcript, and the presence of the restriction site for releasing the SAGE tag from the template is the determining factor. Microarray targets various regions of the detected transcript and the base composition is the top consideration to provide high specificity of hybridization.
- The different carriers for transcript detection. SAGE collects sequence information for the detected transcripts and for the copy number of the quantification; microarray relies on fluorescent signals for the detected transcripts and on signal intensity for the quantification.
- The bias present in both platforms. For example, a SAGE tag can contain sequencing error and quantification bias; microarray can contain labeling bias and noise signals from nonspecific hybridization.
- The inconsistency in both the SAGE and the microarray platforms. As discussed in the text, the reproducibility of SAGE for low-abundance transcripts is poor. Similarly, microarray has low sensitivity for detecting low-abundance transcripts owing to the low signal-to-noise ratio. The inconsistency of interplatform comparisons in microarray analyses (e.g. long versus short oligonucleotides, oligonucleotides versus spotted cDNAs) also leads to problems [27] when SAGE data are compared with data from different microarray platforms.

A computational study comparing the data from five common tissue types collected by Affymetrix oligonucleotide chip, EST-based cDNA chip and SAGE [33] concluded that Affymetrix data could be combined with either SAGE or EST–chip to increase the confidence for the genes detected by each system. Unifying the data sets from all three systems into one was not optimal, however, owing to the diversity in each data set.

For the comparison of disease-related data generated by SAGE and microarray, the situation becomes more complicated. For example, in our SAGE studies on acute myeloid leukemia, we confirmed only 6 out of 48 disease-related genes detected by different microarray platforms [52]. In addition to the factors discussed above, the heterogeneity of the clinical samples used for the studies might also contribute to the inconsistency. A study has shown that using Gene Ontology terms for comparison provides a better chance of identifying the coexpressed genes detected by SAGE, cDNA microarray and oligonucleotide microarray [53].

Table I. Comparison of SAGE and microarray for gene expression studies

Features	SAGE	Microarray
Detects known transcripts	Yes	Yes
Detects unknown transcripts	Yes	No
Detects alternatively spliced transcripts	Yes	Yes or no
Detects antisense transcripts	Yes	Yes or no
Quantification	Absolute measure	Relative measure
Sensitivity	High	Moderate
Specificity	Moderate	High
Reproducibility	Good for higher abundance transcripts	Good for data from intra-platform comparison
Comparability each other	Good for higher abundance transcripts	Good for higher abundance transcripts
Direct cost	5–10 times higher than array	5–10 times lower than SAGE

a statistically significant level. These genes, in most cases, are represented by high-copy SAGE tags, which have higher reproducibility. For SAGE tags with lower copy numbers, SAGE studies are mainly used to discover new transcripts and genes. For this purpose, the issue of reproducibility has less concern.

Comparability between SAGE and microarrays

SAGE is an open system that detects both known and unknown transcripts and genes. Microarray is a closed system that detects known transcripts and genes. A comparison between SAGE and microarray is therefore restricted to known transcripts and genes, which account for only about half of the SAGE data.

Many studies have compared the gene expression profiles obtained with SAGE and microarrays [8,32,36–42] (Box 3). The general conclusion is that comparison of the same RNA samples between these two platforms results in a modest to high correlation for transcripts at higher abundance. The correlation decreases for low-abundance transcripts. Both the SAGE and the microarray platforms have strong and weak points. Data from the two systems can be complementary to each other.

Can SAGE identify the full contents of the transcriptome?

Although SAGE is probably the most sensitive method for detecting transcripts at the genome level, no given SAGE data sets currently show saturated tag collection in any tissue sample analyzed. In two human LongSAGE tag sets, for example, 305 546 and 401 432 LongSAGE tags were collected in total from human brain and human embryonic stem (ES) cells, respectively, representing the highest tag collection among all single tissue or cell types analyzed by SAGE. Plotting SAGE tag collection versus unique SAGE tag identification shows that, in both data sets, the first 50 000 or so tags identified contain a higher number of unique tags. After this point, the number of unique tag detected remains at a relatively constant level until the collection of 300 000 tags in brain RNA and 400 000 tags in ES cell RNA – that is, it decreases from 4000 to 3000 unique tags in brain, and from 3000 to 2000 unique tags in ES cells, per 10 000 SAGE tags collected (Figure 1). The higher number of unique tags in brain RNA confirms that this organ is a transcript-rich tissue, as revealed by EST studies.

The relatively constant level of unique tag detection in both SAGE data sets implies that these two studies are far

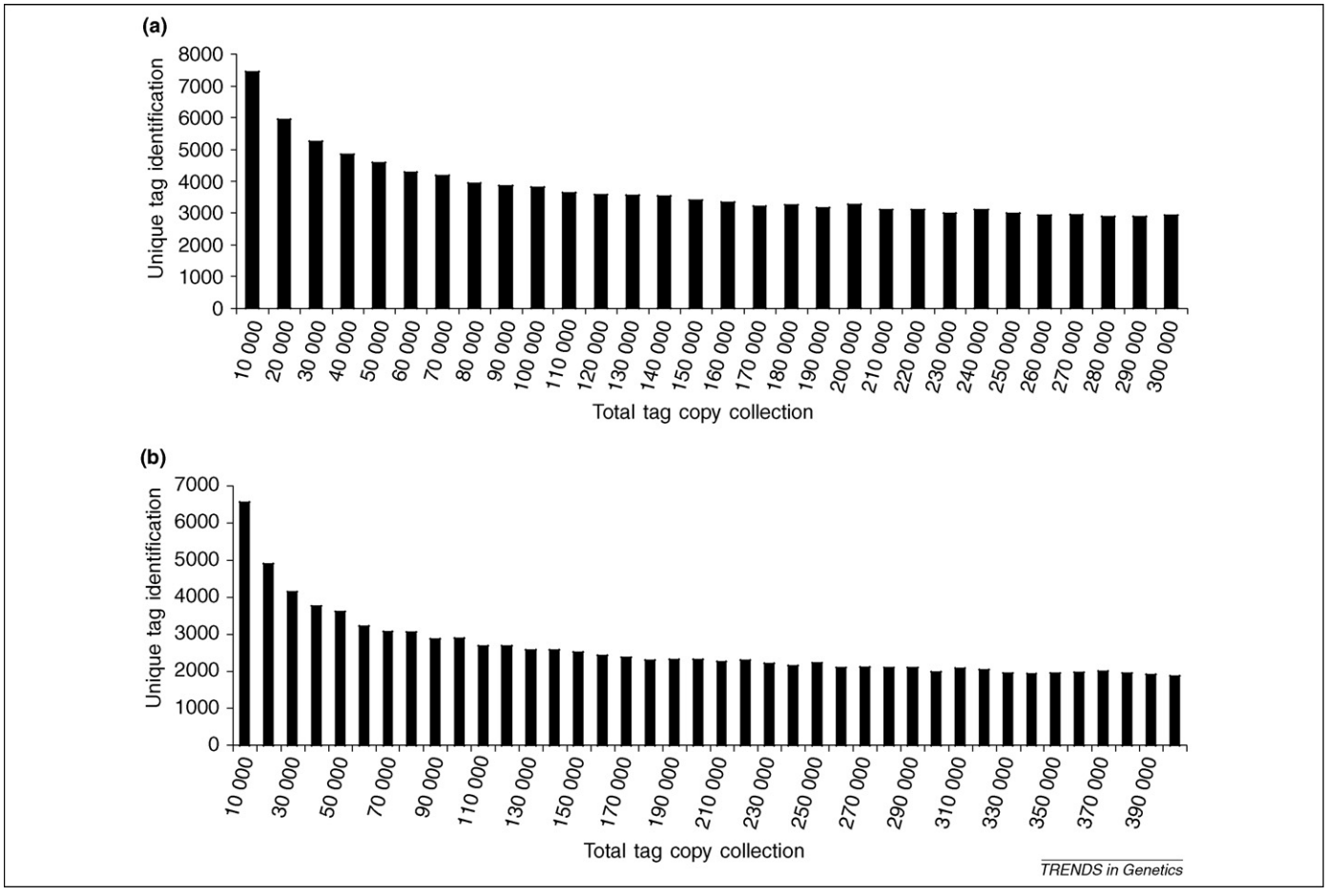

Figure 1. Unsaturated detection of transcripts. LongSAGE (21-bp) data sets from human fetal brain (305 546) and ES cells (401 432) in the GEO database (NCBI GEO GSM31935, GSM31945) were used for this analysis. The SAGE tags were randomly divided into subsets of 10 000 tags and the final subset containing less than 10 000 tags was omitted. The number of unique tags detected in each subset was determined. After the first subset, only the unique tags detected in each subset (i.e., those not previously detected) were plotted. **(a)** LongSAGE tags collected from human fetal brain RNA. **(b)** LongSAGE tags collected from human ES cell RNA. The number of unique tags identified remains at a relatively constant level, indicating that tag identification did not reach saturation.

from reaching the saturation stage of transcript detection in the two samples. Furthermore, the transcripts detected by these two sets of SAGE data seem to reflect, in large part, the high- and intermediate-abundance transcripts shown by the RNA reassociation studies [24,43]. The low-abundance transcript population does not seem to have been detected significantly, because it would have a much lower proportion of unique tags within the total number of SAGE tags collected. Because most SAGE studies collect 50 000 to 100 000 tags per sample under the scope of tag collection, identification of the total transcript content in a given tissue or cell type is impossible. Interestingly, the 300 000–400 000 tags collected from the two above studies match the long-term assumption that there are 300 000 transcripts per cell [43]; however, the fact that the unsaturated transcript detection implies that the actual number of transcripts present in a cell must be higher than this number.

The unsaturated detection of transcripts is not due to SAGE itself but to the restricted scale of DNA sequencing. A given SAGE library probably contains most tags collected from a sample, as judged by the fact that many single-copy tags do not overlap between different SAGE libraries (a certain portion of these tags are those from sequencing error), an indication that the number of the SAGE tags in the library is much larger than the currently detected one. Restricted by the cost-efficiency of current sequencing systems, however, it is difficult to detect comprehensively tags with low copy numbers. On the basis of US $3 for sequencing 700 bp per 50 14-bp tags by the current Big-Dye sequencing system, the cost for collecting 100 000 14-bp tags per sample would be $6,000 and that for collecting 1 million 14-bp tags per sample would be $60 000, which is unaffordable for most academic laboratories. New sequencing technologies that markedly increase the throughput capacity at a low cost should improve this situation. For example, a recently developed 454 sequencing system can collect 200 000 sequence reads of up to 100 bp per sequence per run per sample at cost of about $10 000 [44]. Sequencing a concatenated SAGE library with the 454 system could detect about seven tags per sequence or 1.4 million tags per sample. When such data become available, it will be interesting to see whether it provides high quality tag sequences and whether we are still far from, are closer, or have reached the goal of the exhaustive detection of most transcripts in a cell.

Conclusions and future directions

The transcriptome might turn out to be more complex than the genome. Each technological platform currently used to study the transcriptome has its advantages and disadvantages. For example, hybridization-based microarrays provide high-throughput capacity for known but not unknown transcripts, amplification-based PCR and real-time PCR provide high sensitivity but low-throughput and are limited to known transcripts. For sequencing-based approaches, two main factors dictate the extent of transcript detection. One is the length of the tag sequenced for each transcript, the other is the total number of DNA sequences collected for the total transcript population. Full-length cDNA and EST collections provide high specificity for the transcripts detected but have limited sensitivity for detecting low-abundance transcripts. By decreasing the length sequenced per transcript to a minimum, SAGE converts the transcriptome complex into the simplest form. This simplicity enables SAGE to have high sensitivity albeit low specificity for transcript analysis. Until sequencing technologies reach a stage that is capable of covering the full scope of transcriptome and genome at an affordable cost and with both qualitative and quantitative information, SAGE will continue to have a unique role in transcriptome and genome studies.

Acknowledgements

I thank Y.C. Kim and X. Ge for data collection. I apologize for not citing many recent outstanding publications owing to space restrictions. My studies have been supported by the National Institutes of Health, the US Department of Defense, the Daniel F. and Ada L. Rice Foundation, and Mazza Foundation.

References

1 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science* 270, 484–487
2 Pleasance, E.D. *et al.* (2003) Assessment of SAGE in transcript identification. *Genome Res.* 13, 1203–1215
3 Lee, S. *et al.* (2002) Correct identification of genes from serial analysis of gene expression tag sequences. *Genomics* 79, 598–602
4 Ge, X. *et al.* (2006) Annotating nonspecific SAGE tags with microarray data. *Genomics* 87, 173–180
5 Lash, A.E. *et al.* (2000) SAGEmap: a public gene expression resource. *Genome Res.* 10, 1051–1060
6 Saha, S. *et al.* (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.* 20, 508–512
7 Matsumura, H. *et al.* (2003) Gene expression analysis of plant host–pathogen interactions by SuperSAGE. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15718–15723
8 Lu, J. *et al.* (2004) A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. *Genomics* 84, 631–636
9 Boon, K. *et al.* (2002) An anatomy of normal and malignant gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 99, 11287–11292
10 Unneberg, P. *et al.* (2003) Transcript identification by analysis of short sequence tags – influence of tag length, restriction site and transcript database. *Nucleic Acids Res.* 31, 2217–2226
11 Silva, A.P. *et al.* (2004) The impact of SNPs on the interpretation of SAGE and MPSS experimental data. *Nucleic Acids Res.* 32, 6104–6110
12 Aksoy, I.A. *et al.* (1994) Human liver estrogen sulfotransferase: identification by cDNA cloning and expression. *Biochem. Biophys. Res. Commun.* 200, 1621–1629
13 Chen, J. *et al.* (2002) High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3′ complementary DNAs. *Genes Chromosomes Cancer* 33, 252–261
14 Velculescu, V.E. *et al.* (1999) Analysis of human transcriptomes. *Nat. Genet.* 23, 387–388
15 Siddiqui, A.S. *et al.* (2005) A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc. Natl. Acad. Sci. U. S. A.* 102, 18485–18490
16 Chen, J. *et al.* (2002) Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12257–12262
17 Bertone, P. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science* 306, 2242–2246
18 Cheng, J. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* 308, 1149–1154
19 Lee, S. *et al.* (2005) Detecting novel low-abundant transcripts in *Drosophila*. *RNA* 11, 939–946
20 Feuk, L. *et al.* (2006) Structural variation in the human genome. *Nat. Rev. Genet.* 7, 85–97
21 Shiraki, T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15776–15781

22 Wei, C.L. *et al.* (2004) 5′ Long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. *Proc. Natl. Acad. Sci. U. S. A.* 101, 11701–11706
23 Matsumura, H. *et al.* (2006) SuperSAGE array: the direct use of 26-base-pair transcript tags in oligonucleotide arrays. *Nat. Methods.* 3, 469–474
24 Bishop, J.O. *et al.* (1974) Three abundance classes in HeLa cell messenger RNA. *Nature* 250, 199–204
25 Margulies, E.H. *et al.* (2001) Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res.* 29, E60-0
26 Asim, S. *et al.* (2006) Sequence biases in large scale gene expression profiling data. *Nucleic Acids Res.* 34, e83
27 Draghici, S. *et al.* (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* 22, 101–109
28 Madden, S.L. *et al.* (1997) SAGE transcript profiles for p53-dependent growth regulation. *Oncogene* 15, 1079–1085
29 Claverie, J.M. and Audic, S. (1996) The statistical significance of nucleotide position–weight matrix matches. *Comput. Appl. Biosci.* 12, 431–439
30 Man, M.Z. *et al.* (2000) POWER-SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics* 16, 953–959
31 Romualdi, C. *et al.* (2003) IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments. *Physiol. Genomics* 12, 159–162
32 Ishii, M. *et al.* (2000) Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* 68, 136–143
33 Trendelenburg, G. *et al.* (2002) Serial analysis of gene expression identifies metallothionein-II as major neuroprotective gene in mouse focal cerebral ischemia. *J. Neurosci.* 22, 5879–5888
34 Dinel, S. *et al.* (2005) Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome. *Nucleic Acids Res.* 33, e26
35 Zhang, L. *et al.* (1997) Gene expression profiles in normal and cancer cells. *Science* 276, 1268–1272
36 Nacht, M. *et al.* (1999) Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. *Cancer Res.* 59, 5464–5470
37 Evans, S.J. *et al.* (2002) Evaluation of Affymetrix gene chip sensitivity in rat hippocampal tissue using SAGE analysis. *Eur. J. Neurosci.* 16, 409–413
38 Kim, H.L. (2003) Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from $CD34^+$ cells. *Exp. Mol. Med.* 35, 460–466
39 Haverty, P.M. *et al.* (2004) Limited agreement among three global gene expression methods highlights the requirement for non-global validation. *Bioinformatics* 20, 3431–3441
40 Ibrahim, A.F. *et al.* (2005) comparative analysis of transcript abundance using SAGE and Affymetrix arrays. *Funct. Integr. Genomics* 5, 163–174
41 Georgantas, R.W., III *et al.* (2004) Microarray and serial analysis of gene expression analyses identify known and novel transcripts overexpressed in hematopoietic stem cells. *Cancer Res.* 64, 4434–4441
42 van Ruissen, F. *et al.* (2005) Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. *BMC Genomics* 6, 91
43 Hastie, N.D. *et al.* (1976) The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* 9, 761–774
44 Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380
45 Wang, S.M., ed. (2005) *SAGE: Current Technologies and Applications*. Horizon Scientific Press
46 Kodzius, R. *et al.* (2006) CAGE: cap analysis of gene expression. *Nat. Methods* 3, 211–222
47 Carninci, P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563
48 Ng, P. *et al.* (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* 2, 105–111
49 Wang, T.L. *et al.* (2002) Digital karyotyping. *Proc. Natl. Acad. Sci. U. S. A.* 99, 16156–16161
50 Park, J.T. *et al.* (2006) *Notch3* gene amplification in ovarian cancer. *Cancer Res.* 66, 6312–6318
51 Wei, C.L. *et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell.* 124, 207–219
52 Lee, S. *et al.* (2006) Gene expression profiles in acute myeloid leukemia with common translocations using SAGE. *Proc. Natl. Acad. Sci. U. S. A.* 103, 1030–1035
53 Griffith, O.L. *et al.* (2005) Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. *Genomics* 86, 476–488

DEVELOPMENTAL BIOLOGY

Special Issue on the Sea Urchin Genome: Implications and Insights
Edited by Eric Davidson, Dave McClay and Richard Hynes
Volume 300, Number 1, December 1, 2006

***Developmental Biology* is available on ScienceDirect,**
www.sciencedirect.com/science/journal/00121606

Reproduction of material from Elsevier articles

Interested in reproducing part or all of an article published by Elsevier, or one of our article figures? If so, please contact our *Global Rights Department* with details of how and where the requested material will be used. To submit a permission request online, please contact:

Elsevier
Global Rights Department
PO Box 800
Oxford OX5 1DX, UK
Phone: +44 (0)1865 843 830
Fax: +44 (0)1865 853 333
permissions@elsevier.com

Alternatively, please visit:

www.elsevier.com/locate/permissions