Tandem arrayed ligation of expressed sequence tags (TALEST): a new method for generating global gene expression profiles

Dominic G. Spinella*, Alexandra K. Bernardino, Amy C. Redding, Patricia Koutz, Yalin Wei, Ellen K. Pratt, Kristi K. Myers, Gary Chappell, Steven Gerken and Stephen J. McConnell

Chugai Biopharmaceuticals, Inc., 6275 Nancy Ridge Drive, San Diego, CA 92121, USA

Received March 1, 1999; Revised and Accepted June 25, 1999

ABSTRACT

We have developed a new and simple method for quantitatively analyzing global gene expression profiles from cells or tissues. The process, called TALEST, or tandem arrayed ligation of expressed sequence tags, employs an oligonucleotide adapter containing a type IIs restriction enzyme site to facilitate the generation of short (16 bp) ESTs of fixed position in the mRNA. These ESTs are flanked by GC-clamped punctuation sequences which render them resistant to thermal denaturation, allowing their concatenation into long arrays and subsequent recognition and analysis by high-throughput DNA sequencing. A major advantage of the TALEST technique is the avoidance of PCR in all stages of the process and hence the attendant sequence-specific amplification biases that are inherent in other gene expression profiling methods such as SAGE, Differential Display, AFLP, etc. which rely on PCR.

INTRODUCTION

In recent years, a massive influx of DNA sequencing data from the Human Genome Project and various EST programs have placed into public and private databases sequence information about a substantial proportion of the 80–100 thousand expressed genes (1) in the human genome. However, mere DNA sequence information is insufficient to allow complete understanding of gene function and regulation. Because only a fraction of the full genetic repertoire is expressed in any given cell at any given time, and because the degree of gene expression can dramatically influence cellular phenotype, such understanding requires tools to monitor global gene expression at both qualitative and quantitative levels.

A number of methods have been developed to assess and quantify gene expression. Classical techniques such as northern blotting and nuclease protection assays are accurate and quantitative, but are inherently too slow and cumbersome to provide the parallel information about thousands of genes required to generate a global gene expression profile. Approaches such as Differential Display (2) and AFLP (3) provide gene expression information at the appropriate scale, but can suffer from a lack of quantitative precision and reproducibility and are also susceptible to quantitative PCR artifacts as will be described below. Hybridization of RNA or cDNA to high-density microarrays corresponding to known genes (for review, see 4) represents perhaps the best method for parallel analysis of global gene expression. However, limitations remain with respect to the number of genes that can be analyzed in a single microarray, and of course the method is applicable only to those genes whose sequence is already available.

Currently, the most direct approach to generating quantitative gene expression profiles remains sequencing of random isolates from cDNA libraries to generate expressed sequence tags (ESTs) (5,6). Within appropriately constructed libraries, frequency distributions of cDNA clones are largely proportional to steady-state transcript levels in the RNA population from which the library was derived, rendering the technique reasonably quantitative (7,8). However, this approach requires a massive DNA sequencing effort which is beyond the capacity of most biomedical research laboratories. Recently, a variation of the random cDNA sequencing approach called serial analysis of gene expression (SAGE) was described by Velculescu *et al.* (9). The SAGE technique involves the use of type IIs restriction endonucleases to generate short but positionally defined sequences from cDNAs which are randomly ligated in a tail-to-tail fashion and amplified by PCR to form ‘di-tags’. These di-tags are then concatenated into arrays which are cloned and analyzed by DNA sequencing. Because each sequencing template contains identifiable tags corresponding to many genes, the potential throughput of SAGE vastly exceeds that of traditional cDNA sequencing efforts, bringing global gene expression profiling within reach of smaller research laboratories.

A drawback to the SAGE technique is its reliance on PCR amplification to generate di-tags which compromises the quantitative aspects of the method. In order for any cDNA sequencing approach to be quantitative, it is critical that the frequency of isolates in the library (whether individual cDNA clones or tags as in the case of SAGE) reflect the frequency of the mRNAs in the parent pool from which the clones or tags were derived. Owing to the exponential amplification of templates by PCR, even very minor variations in amplification efficiency of template

*To whom correspondence should be addressed. Tel: +1 858 535 5913; Fax: +1 858 546 5977; Email: dspinella@chugaibio.com
Present address:
Gary Chappell, PE-Applied Biosystems, 850 Lincoln Centre Dr., Foster City, CA 94404, USA

sequences can give rise to dramatic differences in quantity of PCR product. Among the factors that influence PCR amplification efficiency is the DNA sequence of the amplicon itself. The authors of the SAGE technique recognize this problem and attempt to ameliorate it by eliminating from consideration any repetitive isolation of an identical di-tag sequence. Because the assortment of individual tags into di-tags is a random event, and because the relative frequency of even a high abundance tag is very low, the probability of any two tags randomly associating together more than once is extremely low. Any observed multiple occurrence of the same di-tag sequence in SAGE is therefore presumed to reflect the selective amplification advantage of that particular sequence in the PCR reaction required to produce the di-tags. However, the very observation of these multiple di-tag sequences serves to underscore rather than solve the problem. As we will demonstrate below, any tag sequence that has a relatively high intrinsic amplification efficiency will be over-represented after the SAGE PCR reaction while any tag sequence which is inherently difficult to amplify will be under-represented—regardless of the partner sequence with which it is paired in a di-tag. Hence, in the SAGE technique, the frequency of tag isolation is influenced not only by the starting frequency of the mRNA templates, but also by the intrinsic amplification efficiency of the individual tag sequences.

In order to circumvent this problem, we have developed a cDNA tag-based technique called TALEST (tandem arrayed ligation of expressed sequence tags) that does not rely on PCR amplification to generate tag arrays. The TALEST technique retains all the advantages of SAGE with respect to throughput without introducing the quantitative biases associated with PCR. The technique itself is relatively simple to perform, requiring only common tools and methods used to construct standard cDNA libraries and yet provides an ~30–40-fold increase in throughput relative to random cDNA sequencing approaches to gene expression profiling.

MATERIALS AND METHODS

cDNA synthesis

Five micrograms of poly(A)$^+$ RNA from normal adult human lung (Clontech, Inc., Palo Alto, CA) was primed with 5' biotinylated oligo(dT)$_{25}$ and copied into cDNA using the Superscript II® cDNA synthesis kit (GIBCO Life Technologies, Gaithersburg, MD) according to the manufacturer's directions and then treated with 160 U of *Eco*RI methylase (New England Biolabs, NEB, Tozer, MA). The methylase was inactivated by heating to 65°C for 15 min and the sample was ethanol precipitated with glycogen carrier. The pellet was dissolved in 200 µl of 1× NEB buffer II containing 500 U of *Msp*I and digested for 2 h at 37°C. Fifty µl of 5 M NaCl and 2.5 µl 0.5 M EDTA were then added to the mix. Three hundred µl of streptavidin magnetic beads (Dynal, Inc., Lake Success, NY) were washed twice with 500 µl of magnetic bead binding buffer (MBB: 10 mM Tris–HCl pH 7.4, 1 mM EDTA, 1 M NaCl) and added to the *Msp*I digested cDNA. The reaction tube was rotated overnight at room temperature and the beads washed extensively with several 1-ml volumes of MBB to remove non-bound *Msp*I fragments.

Adapter ligation and library generation

Oligonucleotide adapters were created by mixing ~5 O.D. units each of two synthetic oligonucleotides (sense strand: HO-GCCGAATTCGAAACGGCCGATGTCTTCAGTCGACCT-GTATGGCCCTTAGCATTCGCGCCGTGCAGC; antisense strand: P-CGGCTGCACGGCGCGAATGCTAAGGGCCAT-ACA-GGTCGACTGAAGACATCGGCCGTTTCGAATTCG-GC) in standard buffer (10 mM Tris–HCl pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$) heating them to 95°C, and allowing them to cool slowly to room temperature. Adapter DNA (2.4 nmol) was added to the solid-phase cDNA in a total volume of 100 µl of 1× ligase buffer containing 25 U of T4 ligase (Gibco BRL). The reaction was incubated for 2 h at room temperature followed by 20 min at 65°C to inactivate the enzyme. Beads were again washed extensively in MBB and resuspended in 100 µl of *Bsg*I buffer containing 50 U of *Bsg*I and incubated at 37°C for 2 h with gentle rotation. The tag-containing supernatant was collected and the beads were washed twice with 200 µl of 10 mM Tris–HCl, pH 7.6. The supernatant and washes were combined, brought to 0.3 M NaCl and incubated at 65°C for 20 min to inactivate the *Bsg*I. The reaction was then concentrated to ~30 µl using a Microcon-10 spin filter (Amicon, Beverly, MA).

A second, 16-fold degenerate adapter molecule was prepared by annealing synthetic oligos as described above (sense strand: P-CC-GGTCTAGAGCGGCCGCCTGACCAGGTATGA; antisense strand: Biotin-TGATACCTGGTCGAGCGGCCGCTCTAGAC CGGNN). Two hundred and forty-one pmol of adapter were added to the tag DNA in a total volume of 45 µl of 1× ligase buffer containing 10 U of T4 ligase and incubated for 2 h at room temperature. The ligase was inactivated by incubation at 65°C for 20 min. The reaction was then brought to a final volume of 250 µl in 1× GIBCO React 3 buffer containing 625 U of *Not*I, and incubated overnight at 37°C. The following day, 125 U of *Eco*RI (Gibco-BRL) were added to the reaction mix and incubated for 2 h at 37°C. The reaction was brought to 1 M NaCl and 5 mM EDTA, added to 300 µl of MBB-washed streptavidin magnetic beads, and incubated for 2 h at room temperature to remove any non-ligated adapters.

The supernatant from binding was concentrated to ~30 µl using a Microcon-10 spin filter and loaded on a 10% TAE acrylamide gel to resolve the 93 bp fragment containing TALEST tags from multimers of the adapter. Tag fragments were excised from the SYBR Green-stained gel, eluted overnight in 400 µl of TE, and ethanol precipitated with 15 µg Glyco Blue (Ambion, Austin, TX) as carrier. Tag DNA was quantified by spectrophotometer and ligated overnight at 16°C into the *Not*I and *Eco*RI sites of pTALEST vector (pUC 19 in which endogenous *Msp*I sites were destroyed by site-directed mutagenesis) at a 3:1 insert:vector ratio. The ligation mix was transformed into competent XL-10 Gold® (Stratagene, La Jolla, CA) according to the manufacturer's directions and grown in 2 l of LB containing 100 µg/ml ampicillin.

Tag isolation and concatemer formation

Transformed bacteria were recovered by centrifugation and plasmid DNA isolated using the Mega Plasmid Prep Kit (Qiagen, Inc., Valencia, CA) according to the manufacturer's protocol. One milligram of plasmid DNA was digested with 5000 U *Msp*I in a total volume of 1 ml (20°C, 1 h) and loaded onto several lanes of a 15% TAE acrylamide gel to resolve the 16 bp tags from the plasmid backbone. Tag fragments were excised from the EtBr-stained gel and eluted overnight in 300 mM sodium acetate, pH 5.2 and then ethanol precipitated using 15 µg of Glyco Blue carrier. The tag pellet was resuspended in 20 µl of 1× ligase buffer containing 10 U of T4 ligase, and incubated at room temperature for 1 h. This was followed by an additional 20 min incubation with 7.5 U of Klenow (Boehringer Mannheim, Indianapolis, IN) and 2 mM dNTPs to blunt the concatemers. The reaction was then loaded onto a 0.8% TAE agarose gel containing EtBr and electrophoresed. Concatemers of 600–1200 bp in length were electrophoresed onto DE-81 paper, washed with 450 µl of 20 mM Tris–HCl, pH 7.4, 1 mM EDTA, 100 mM NaCl and eluted with Tris–EDTA containing 1.5 mM NaCl. The eluate was ethanol precipitated, resuspended in 6 µl of water and quantified. The blunt-ended concatemers were then ligated into *Sma*I-cut, alkaline phosphatase-treated pUC19 for 1 h at room temperature in 0.5× ligase buffer and 5 U of T4 ligase at a 40:1 insert:vector ratio. One microliter of the ligation mix was transformed into 100 µl of competent XL-10 Gold® (Stratagene) using the manufacturer's protocol. The reaction mix was plated on LB-Amp plates with IPTG-Xgal and individual white colonies were picked for DNA sequence analysis.

DNA sequencing

Sequencing-grade plasmid templates were prepared using the Qiagen BioRobot 9600 in 96-well plates. Templates were subjected to cycle sequencing using ABI Prism BigDye® terminator chemistry from the M13 reverse primer. Reactions were run on ABI 377 automated sequencers. Extracted data were ported to an Oracle® database using the BioLims® system (PE Informatics, San Jose, CA). Raw sequence data were searched for valid tags using the TALEST software package as described. The software ensures that only unambiguous tag sequences are placed into the database for profile generation (tags with uncalled bases or Ns are binned separately and can be called manually at a later time).

cDNA library generation and hybridization

Five micrograms of the pooled human lung mRNA was primed with oligo(dT) and converted into double-stranded cDNA using a commercial cDNA synthesis kit (GIBCO). *Eco*RI adapters were ligated onto the cDNA and the mixture cloned into *Eco*RI-digested bacteriophage lambda gt10 arms (Promega, Madison, WI) according to the manufacturer's protocol. The library was plated on *Escherichia coli* strain C600 Hfl-. Two replicate nylon filter lifts were prepared from a large plate containing approximately 2700 primary plaques. Each filter was hybridized to a 5' radiolabeled probe corresponding to the TALEST tag for EF-1α (CCGGCTATGCCCCTGT). Filters were hybridized to the probe using the ExpressHyb™ system (Clontech, Palo Alto, CA) according to the manufacturer's protocol.

RESULTS AND DISCUSSION

Effects of PCR on quantitative aspects of SAGE

Given the reliance of the SAGE technique on PCR, we suspected that the quantity of any individual tag after amplification would be influenced by its nucleotide sequence. In order to experimentally test this hypothesis, we synthesized and annealed two pairs of complementary oligonucleotides corresponding to synthetic SAGE amplicons (Fig. 1A). Each amplicon contains a single di-tag of 18-bases in length flanked by anchoring enzyme sequences (CATG) and PCR priming sites exactly as described by Velculescu *et al.* (9) One of the tags in each amplicon is an identical 9-base arbitrary sequence (CTACTAACA). This common tag sequence is paired with either an AT-rich non-palindromic 'tag' sequence (TATAATAAA for amplicon 1) or a GC-rich palindromic sequence (CCCGATCGG for amplicon 2) to form artificial di-tags. The entire double-stranded synthetic amplicon terminates in single-stranded ends compatible with *Bam*HI and *Hin*dIII digested vectors to facilitate cloning. The experiment consisted of ligating these artificial SAGE amplicons into a plasmid vector, and preparing precisely equivalent concentrations of plasmid DNA from each. The plasmid DNA was diluted and subjected to either 15 or 20 cycles of amplification using 6-FAM end-labeled SAGE primers and amplification conditions described (9). As shown in Figure 1B, the intensity of the bands derived the corresponding PCR products is far from identical. After 15 cycles of amplification, there is no visible PCR product from derived from amplicon 2 while the band derived from amplicon 1 is clearly evident. When the cycle number is increased to 20, both PCR products are visible, but peak area analysis (PE Biosystems Prism 377 Genescan™ software) demonstrates that the quantity of product derived from amplicon 1 is more than seven times that of amplicon 2—despite the fact that there was no difference in starting template concentration or PCR primers. Note that the only difference between the two 68-bp SAGE amplicons resides in one of the 9-bp sequences that comprises half of a di-tag. The clear implication is that the amount of PCR product produced after amplification can be dramatically influenced by the sequence of any tag within it. This problem affects not only the SAGE technique, but also virtually any gene expression methodology in which the quantity of a PCR product is measured and used to infer starting template abundance.

Overview of the TALEST technique

In order to avoid the problems introduced by PCR-based profiling techniques, we devised the TALEST method to generate short tag sequences from expressed genes without the use of *in vitro* amplification. Figure 2 depicts a schematic diagram of the TALEST technique. Double-stranded cDNA is prepared from the target mRNA pool by standard methods using a biotinylated oligo(dT) primer. The cDNA is treated with *Eco*RI methylase to protect internal *Eco*RI sites and then digested with a 4-base cutting restriction enzyme (referred to as the punctuating enzyme or PE–typically *Msp*I). The 3'-most fragment is then isolated by affinity capture on streptavidin-coated magnetic particles which are extensively washed to remove upstream fragments. A double-stranded adapter molecule with a 5' overhang compatible with *Msp*I is then ligated to the cDNA. The adapter introduces a type IIs restriction site (*Bsg*I) immediately 5' to the ligated cDNA and contains an *Eco*RI site at its 5' terminus to facilitate later cloning. Digestion of the adapter-ligated, solid-phase cDNA with *Bsg*I releases into the solution phase a linear DNA fragment consisting of the adapter itself and an additional 12/10 nucleotides of unknown cDNA sequence separated from the adapter by the *Msp*I punctuation

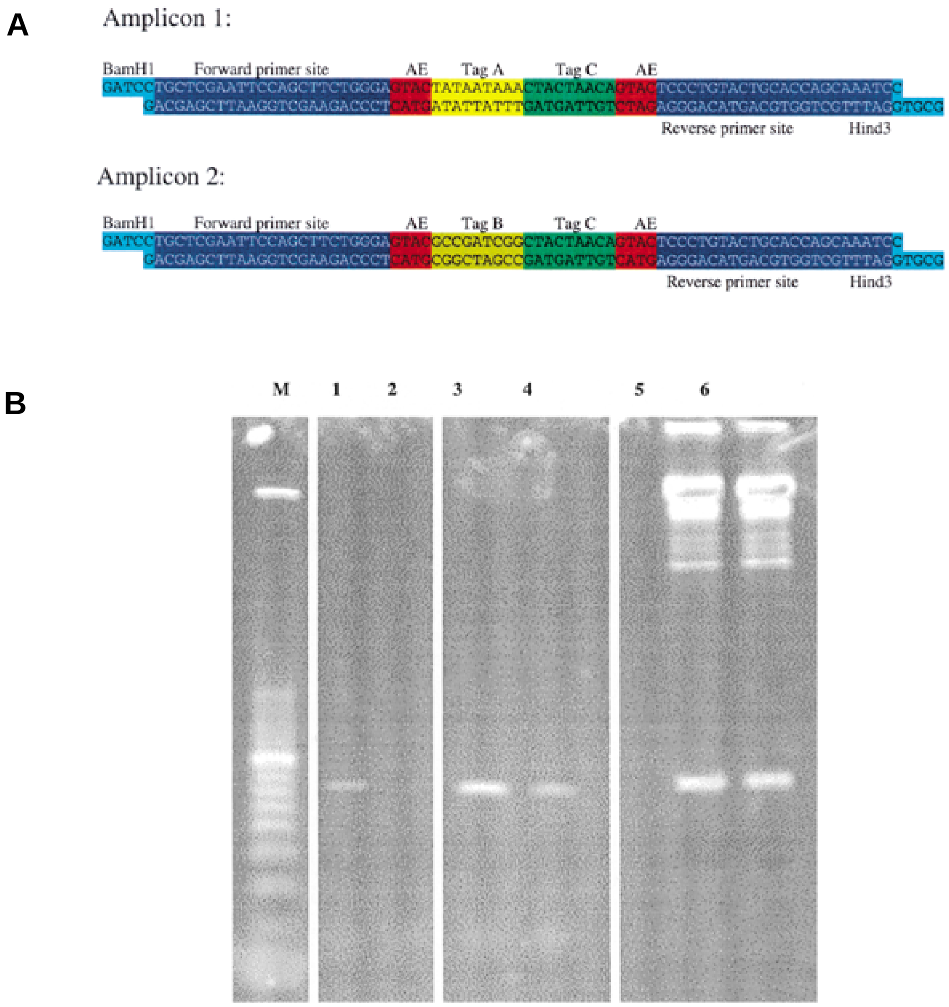
A
Amplicon 1:
BamH1
Forward primer site
AE
Tag A
Tag C
AE
GATCCTGCTCGAATTCCAGCTTCTGGGAGTACTATAATAAACTACTAACAGTACTCCCTGTACTGCACCAGCAAATCC
GACGAGCTTAAGGTCGAAGACCCTCATGATATTATTTGATGATTGTCTAGAGGGACATGACGTGGTCGTTTAGGTGCG
Reverse primer site
Hind3
Amplicon 2:
BamH1
Forward primer site
AE
Tag B
Tag C
AE
GATCCTGCTCGAATTCCAGCTTCTGGGAGTACGCCGATCGGCTACTAACAGTACTCCCTGTACTGCACCAGCAAATCC
GACGAGCTTAAGGTCGAAGACCCTCATGCGGCTAGCCGATGATTGTCATGAGGGACATGACGTGGTCGTTTAGGTGCG
Reverse primer site
Hind3
B
M
1
2
3
4
5
6

Figure 1. Effects of PCR on SAGE quantification of gene expression. (**A**) A pair of synthetic SAGE amplicons was produced by annealing oligonucleotides to create double-stranded templates. The amplicons differ only in one of the 9-base sequences comprising half of a ‘di-tag’. (**B**) The constructs were cloned into *Bam*HI/*Hin*dIII digested plasmid, diluted to precisely 1 pg/µl and subjected to 15 or 20 cycles of amplification using 6-FAM-labeled primers and conditions as described by Velculescu *et al.* (9). Lane M, 100 bp ladder; lane 1, amplicon 1, 15 cycles; lane 2, amplicon 2, 15 cycles; lane 3, amplicon 1, 20 cycles; lane 4, amplicon 2, 20 cycles; lane 5, undiluted amplicon 1 plasmid DNA digested with *Bam*HI/*Hin*dIII; lane 6, undiluted amplicon 2 plasmid DNA digested with *Bam*HI/*Hin*dIII. PCR products were run on a Perkin-Elmer 377 Sequencer and analyzed with Genescan™ software. Peak area integration analysis reveals that amplicon 2 yielded no detectable PCR product at 15 cycles and 7-fold less PCR product after 20 cycles relative to amplicon 1, despite identical template concentrations, primers and PCR conditions.

sequence as shown in Figure 2. This fragment is then ligated to a second biotinylated adapter molecule containing a *Not*I site at its 5' end and a 16-fold degenerate 3' overhang which renders it compatible with all possible cDNA sequences released by *Bsg*I. This adapter introduces a second *Msp*I site to the 3' end of the original DNA fragment such that all the molecules contain a 12-base cDNA-derived ‘tag’ sequence flanked at both ends by *Msp*I ‘punctuation’ sites. The resulting molecule is double-digested with *Eco*RI/*Not*I and the biotinylated *Not*I fragment is removed by affinity capture on streptavidin-coated magnetic particles. The remaining *Eco*RI/*Not*I fragment is isolated by acrylamide gel electrophoresis to resolve it from dimers of the adapter. The result of these manipulations is an *Eco*RI/*Not*I-tailed DNA fragment containing a 12-bp cDNA tag flanked at both ends by the *Msp*I punctuation sequence. The fragment is eluted from the gel and recovered by ethanol precipitation in preparation for cloning into the pTALEST vector. The vector itself is simply pUC 19 in which endogenous *Msp*I sites have been destroyed by site-directed mutagenesis.

The recombinant plasmids are transformed into competent *E.coli* in order to generate a tag library. Plasmid DNA is prepared from the library and digested with *Msp*I to release the tags. Each tag is a DNA fragment consisting of a 12-bp sequence derived from the cDNA flanked at both ends by GC

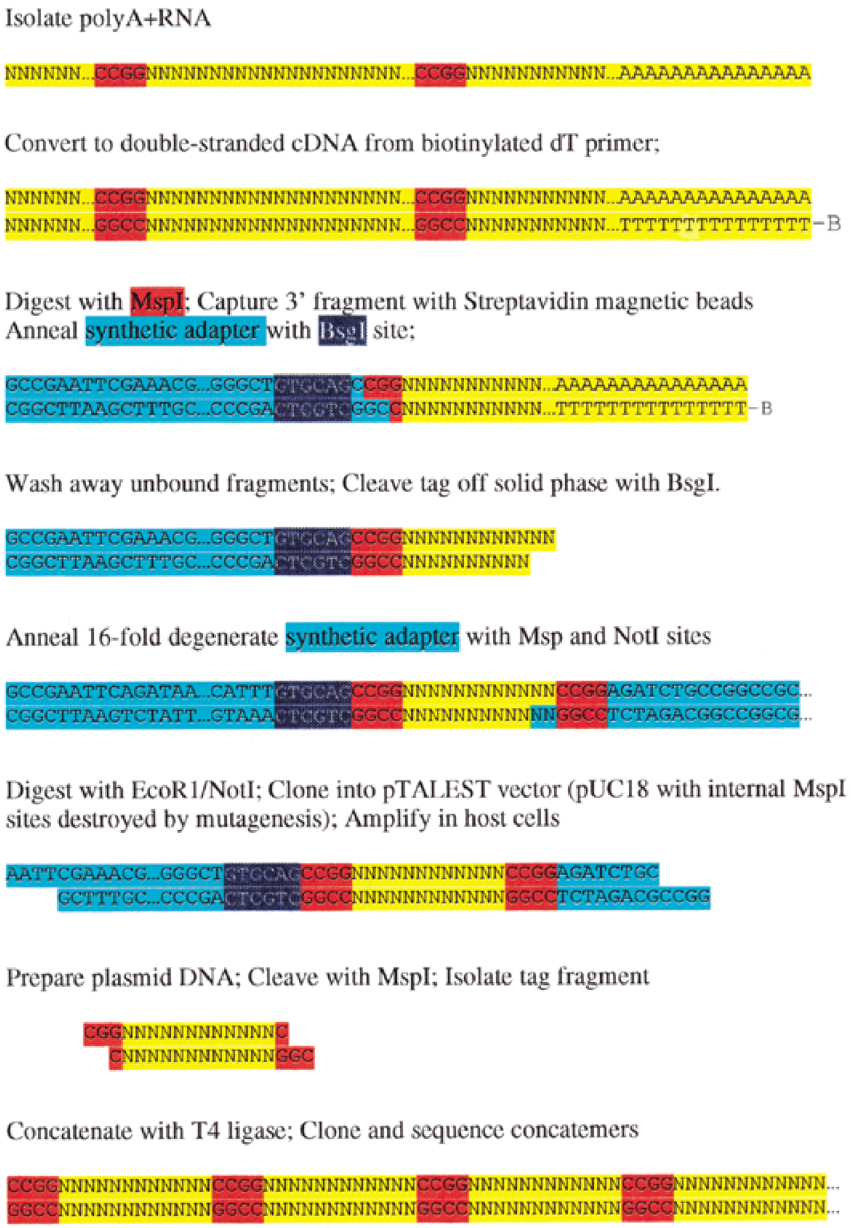
Isolate polyA+RNA
NNNNNN...CCGGNNNNNNNNNNNNNNNNNNNNN...CCGGNNNNNNNNNNN...AAAAAAAAAAAAAAA
Convert to double-stranded cDNA from biotinylated dT primer;
NNNNNN...CCGGNNNNNNNNNNNNNNNNNNNNN...CCGGNNNNNNNNNNN...AAAAAAAAAAAAAAA
NNNNNN...GGCCNNNNNNNNNNNNNNNNNNNNN...GGCCNNNNNNNNNNN...TTTTTTTTTTTTTTT-B
Digest with MspI; Capture 3' fragment with Streptavidin magnetic beads
Anneal synthetic adapter with BsgI site;
GCCGAATTCGAAACG...GGGCTGTGCAGCCGGNNNNNNNNNNN...AAAAAAAAAAAAAAA
CGGCTTAAGCTTTGC...CCCGACTCGTCGGCCNNNNNNNNNNN...TTTTTTTTTTTTTTT-B
Wash away unbound fragments; Cleave tag off solid phase with BsgI.
GCCGAATTCGAAACG...GGGCTGTGCAGCCGGNNNNNNNNNNNN
CGGCTTAAGCTTTGC...CCCGACTCGTCGGCCNNNNNNNNNN
Anneal 16-fold degenerate synthetic adapter with Msp and NotI sites
GCCGAATTCAGATAA...CATTTGTGCAGCCGGNNNNNNNNNNNNCCGGAGATCTGCCGGCCGC...
CGGCTTAAGTCTATT...GTAAACTCGTCGGCCNNNNNNNNNNNNNGGCCTCTAGACGGCCGGCG...
Digest with EcoR1/NotI; Clone into pTALEST vector (pUC18 with internal MspI sites destroyed by mutagenesis); Amplify in host cells
AATTCGAAACG...GGGCTGTGCAGCCGGNNNNNNNNNNNNCCGGAGATCTGC
GCTTTGC...CCCGACTCGTCGGCCNNNNNNNNNNNNGGCCTCTAGACGCCGG
Prepare plasmid DNA; Cleave with MspI; Isolate tag fragment
CGGNNNNNNNNNNNNC
CNNNNNNNNNNNNGGC
Concatenate with T4 ligase; Clone and sequence concatemers
CCGGNNNNNNNNNNNNCCGGNNNNNNNNNNNNCCGGNNNNNNNNNNNNCCGGNNNNNNNNNNNN...
GGCCNNNNNNNNNNNNGGCCNNNNNNNNNNNNGGCCNNNNNNNNNNNNGGCCNNNNNNNNNNNN...

Figure 2. Schematic representation of TALEST. Target mRNA is converted to *Eco*RI methylated, double-stranded cDNA using a biotinylated dT primer, digested with *Msp*I, and annealed to adapters containing a *Bsg*I site. 3' fragments are captured on streptavidin magnetic beads and fragments are released into solution by digestion with *Bsg*I. A second adapter is annealed to the fragment and the product is digested with *Eco*RI, and cloned into the pTALEST vector to generate a tag library. Plasmid DNA is isolated from the tag library and digested with *Msp*I to release free TALEST tags. Tags are concatenated with T4 ligase, size selected, and cloned for sequence analysis.

'clamp' sequences which prevents the melting of tags at ambient temperatures and attendant bias against AT-rich sequences (confirmed by the normal distribution of GC-content in the tags around the expected mean of 6, as well as the presence of numerous tags containing only As and Ts) in addition to single-stranded MspI 5' overhangs. The tag fragments are purified away from the plasmid backbone by PAGE and ligated together to form concatemers. Concatemers of sufficient minimal length are isolated by agarose gel electrophoresis and cloned into standard pUC 19 in preparation for DNA sequence analysis. Each array consists of 20–60 12-base tag sequences separated from each other and from the plasmid backbone by the defined 4-base punctuating enzyme sequence, CCGG.

Both the TALEST and SAGE techniques rely on the ligation of a DNA adapter molecule containing a type IIs restriction enzyme site which allows the generation of fragments containing several nucleotides of positionally defined cDNA-derived sequence (essentially small ESTs). The major difference is that TALEST does not utilize *in vitro* amplified 'di-tags' but, as with traditional EST approaches, relies on simple bacterial amplification of individual clones in a library to generate sufficient material for analysis.

Informatics in support of TALEST

The generation of a representative gene expression profile using TALEST requires analysis of tens to hundreds of thousands of tags. In order to facilitate the tag analysis process, we have developed a software package which is built on the BioLims™ DNA sequencing database package of PE-Applied Biosystems (Foster City, CA). The BioLims system ports data generated by ABI 377 Automated DNA Sequencers into a UNIX relational database (Oracle®) environment as individual projects. The web browser-based TALEST software searches for valid tags in the sequences by looking for pairs of punctuating enzyme sequences interspersed by a defined number of nucleotides (12 bp when *Bsg*I is used). These individual tag sequences are then parsed and sorted into frequency distributions. Because any individual tag sequence can ligate into an array in either sense or antisense orientation, tag sequences which are reverse complements of each other must be binned together. The software accomplishes this by establishing the convention that all tag sequences are compared with their reverse complements as they are found and only the alphabetically primary sequence is placed into the growing frequency distribution—regardless of the actual orientation in which the tag occurs in the array. The software also allows comparison of frequency distributions derived from different TALEST profiles, highlighting only those tags corresponding to some user-defined degree of frequency difference or chi-square threshold. This feature allows one to utilize TALEST to perform what is essentially an *in silico* differential display.

A TALEST tag provides 12 bp of identifying sequence information for each gene (16 bases including the 4-base PE sequence). There are 4^{12} or more than 16 million possible 12mer tag sequences which of course far exceeds the number of expressed genes in the human genome. This number however does not exceed the number of 12mers in the genome. Hence, in order for a TALEST tag to uniquely mark a gene, positional information must also be taken into account. The chemistry of the protocol dictates that the location of a TALEST tag within a gene is always immediately proximal to the 3' terminal punctuating enzyme sequence. In order to identify tagged genes, the tag output file is reformatted into a GCG (Genetics Computer Group, Madison, WI) pattern file. The reformatted file is imported into the GCG Sequence analysis package and each tag is searched (in both orientations) against a human expressed gene database using the FINDPATTERNS algorithm. The actual database was generated from the latest release of The Institute for Genome Research's (TIGR) Expressed Gene Anatomy Database (10) (EGAD) of approximately 7200 full-length human transcripts. The sequences in the EGAD database were converted into 16 base tags corresponding to the 3'-most CCGG and 12 bases of 3' flanking sequence. TALEST tags can also be searched against standard EST databases such as dbest or Unigene; however, in most cases positional information is not available for ESTs and hence the confidence in any match is correspondingly lowered.

In order for an expressed gene to be 'tagged' by TALEST, the gene must contain the 4-base punctuating restriction enzyme site. Since these occur, on average, once every 256 bp, the frequency of cDNAs which lack any given random 4-base sequence is very low. For example, the probability that any particular random 4-base sequence is completely absent in a 1 kb cDNA (containing 996 overlapping tetramers) is $(255/256)^{996} \cong 2\%$. However, the CG dinucleotide in the *Msp*I PE site used in this version of TALEST is under-represented in mammalian genomes (11), and hence this probability is somewhat increased. In order to empirically estimate the frequency of 'untagged' genes, we searched the entire TIGR database of 7223 full-length mRNA transcripts for the subset which lack an *Msp*I (CCGG) site and found 1096 such sequences (15.1%). These genes are invisible to the TALEST technique unless the process is repeated with a different punctuating enzyme. Nevertheless, *Msp*I remains the punctuating enzyme of choice due to its introduction of GC-clamp sequences to the ends of the TALEST tags as well as the high re-ligation efficiency of fragments produced by this enzyme.

A potential concern with any profiling technique which relies upon restriction enzyme digestion of cDNA (TALEST, SAGE, AFLP, etc.) derives from the fact that the initial reverse transcriptase reaction often terminates prematurely and therefore does not always produce full-length copies of the starting mRNA template (particularly when the template is long). This implies that the greater the distance of the 3'-most PE site from the oligo(dT) priming site in the mRNA, the less the probability that it will be contained in any given cDNA copy. Hence, all other things being equal, tags located at sites far distal to the 3' end of their gene would be expected to be somewhat under-represented relative to tags located more proximal to the 3' end of their gene.

Another concern about the use of short tag sequences to identify genes is the risk that a given tag sequence will not be uniquely associated with a single gene. In order to assess this risk, we returned to the EGAD database to provide a representative sample. Of the 6110 genes in the database that contain a *Msp*I restriction site, there were 356 cases (5.8%) in which two or more distinct genes would be expected to generate the same tag sequence. The vast majority of these represent cases in which a prospective TALEST tag fails to discriminate between alternative splice variants of the same gene or between individual members of highly related multi-gene families. There were only 28 instances (0.45%) in which a single tag appears to identify two or more unrelated genes, suggesting that a TALEST tag will uniquely mark a single gene or gene family at >99.5% probability.

Additional statistical considerations

A complete gene expression profile, whether generated by TALEST or traditional EST approaches, requires that the primary library be large enough to be 'representative', i.e., contain cDNA copies of every mRNA in the starting pool. The number of clones that must be obtained in order to ensure the representation of any particular cDNA is given by the formula $N = \ln(1 - P)/\ln(1 - n)$ where N is the number of clones required; P is the probability that the given clone is represented (usually set arbitrarily at 0.99) and n is the fractional proportion of the given sequence in the total mRNA pool (for review see 12). For extremely rare mRNAs, present at only a single copy per cell, the fractional proportion of the total mRNA is $\sim 5 \times 10^{-6}$ and the size of the library required to ensure its presence at 99% probability is $\sim 9.2 \times 10^{5}$. Because TALEST tag constructs are relatively short, their cloning efficiency is quite high and with modern ultra-competent cells we have been able to generate libraries of this size with as little as 2 µg of starting mRNA. TALEST libraries can be produced from smaller quantities of mRNA but

Table 1. Abundant TALEST tags in normal human lung mRNA

Tag sequence	Percent	Database match
GAACTCCTGCCA	1.6068	pulmonary surfactant protein C
AACCTCCGCCTC	0.5962	HLA-E heavy chain
CCCCAAGAGTCA	0.5425	pulmonary surfactant protein A2
AACCTCTGCCTC	0.5217	interleukin 8 receptor beta chain
TGCATGTAGAAA	0.4501	heat shock protein,70 kDa
ACCTGGATTACT	0.4322	Mt. 16S rRNA[a]
GTTTTCCTCTTA	0.3488	Mt. NADH dehydrogenase, subunit 4[a]
AAGAAAGCTTGC	0.3428	pulmonary surfactant protein B
CCCTTCCTGGTC	0.3279	globin, alpha 2
AGGCTGGGCTGG	0.3100	ferritin, light polypeptide
CGTAGTATACCC	0.2862	Mt. Cytochrome C oxidase, subunit 2[a]
AACCTCCACCTC	0.2802	No match
CTGTCACGGCTC	0.2534	immunoglobulin lambda light chain
CAGGAGCCCAGC	0.2504	immunoglobulin alpha heavy chain
GAGCCACTGCGC	0.2266	hypothetical protein related to rat synaptogamin
GAGCCACCGCGC	0.2236	No match
ACAGAAGTCAGC	0.2206	alpha-2-macroglobulin
GACCGCGAGAGC	0.2176	No match
CTGCCAGCCCTA	0.2117	No match
ACAGGGGCATAG	0.2117	translation elongation factor, alpha
AAAACAGCTGGA	0.1878	No match
GTGCGGCTCCAC	0.1818	Wilm’s tumor-related protein
CACATGGCTAGG	0.1789	No match
GAGATCGCGGAC	0.1789	translationally-controlled tumor protein
AGAACCTGAGAC	0.1759	MHC, classII, DR, invariant region
ACGGGCGCCTCC	0.1699	MHC, class I, B
CACCAAGAAGCA	0.1669	ribosomal protein S11
CTGACAGCCCTA	0.1669	No match
GAGCCACCGTGC	0.1610	hereditary haemochromatosis protein HLA-H
GAGCCACCATGC	0.1580	adenosine receptor A2b

[a] Identified by BLAST against human mitochondrial transcripts; mitochondrial sequences are not present in the TIGR EGAD database.

with the attendant risk that tags corresponding to extremely rare messages may not be represented.

An additional question concerning any expression profiling technique is its accuracy as measured by reproducibility, i.e., if two independent samplings of equivalent size are performed on the same mRNA pool, are the results the same? We have attempted to address this question by generating two independent TALEST profiles on a single mRNA sample derived from a human breast tumor cell line (manuscript in preparation). We then analyzed those tags found at a frequency of at least 0.01 in any library by performing a 2 × 2 chi-square contingency analysis for each tag. None of the variations in tag frequencies between the two profiles were significantly different at a *p* value of 0.99, indicating that observed differences in frequency were simply due to random sampling variation.

TALEST profile of mRNA derived from human lung tissue

In order to test and validate the TALEST technique, we prepared a tag library from 5 µg of commercially obtained mRNA (Clontech, Palo Alto, CA) derived from pooled normal human lung tissue. A primary TALEST library of 1.2 million clones was produced and tags were isolated and concatenated into arrays as described. Arrays of ~600 bp or greater were purified by agarose gel electrophoresis and cloned into pUC19. Plasmid

DNA was prepared from about 1100 independent colonies and subjected to automated DNA sequence analysis. The number of identifiable tags in each array ranged from 12 to 63 (with an average of 32 tags per array). A frequency distribution of tags was generated by the TALEST software and searched against our customized EGAD database to generate an expression profile. The profile contains 32 468 unambiguous (no uncalled bases) tag sequences representing 17 084 independent sequences (13 359 singlets; 3725 multiple isolates). Of these, 1370 (8%) had exact matches to our database and could therefore be definitively associated with known genes. Table 1 contains a list of the 30 most abundant tag sequences in the profile and their putative identification where a database match was found. As might be expected, many of the abundant tags corresponding to identifiable genes in this profile have been previously described as being highly expressed either in the lung (13) or in most human tissues (14).

In order evaluate the quantitative precision of the TALEST approach, we produced from remaining mRNA a standard oligo(dT)-primed cDNA library of about 200 000 primary plaques in lambda-gt10 (15). A series of replicate nitrocellulose filter lifts were prepared from plates containing about 2700 plaques. These filters were probed with an end-labeled oligonucleotide corresponding to the tag for translation elongation factor 1-α (EF-1α), a highly represented tag sequence in the TALEST profile of the same mRNA. Of the 2760 plaques screened, seven hybridized to the EF-1α tag probe for a frequency of 0.25%. By contrast, the EF-1α tag was detected 71 times (out of 32 468 total tags) for a frequency of 0.22% indicating that the TALEST gene expression data provide frequency estimates which are not significantly different from those derived from traditional EST analyses. Sequence analysis of the hybridizing clones confirmed their identity as EF-1α.

In conclusion, the TALEST technique provides a rapid and accurate means to quantitatively analyze the gene expression profile of a tissue of interest and to compare profiles between tissues. Because it does not employ PCR, quantitative biases resulting from differences in amplification efficiency of individual sequences are avoided. The method can also be useful in new gene discovery because the 16-bp tag sequences generated by TALEST can serve as hybridization probes to facilitate the isolation of interesting tagged genes whose function is not yet known.

REFERENCES

1. Fields,C., Adams,M.D., White,O. and Venter,J.C. (1994) *Nature Genet.*, **7**, 345–346.
2. Liang,P. and Pardee,A.B. (1992) *Science*, **257**, 967–971.
3. Vos,P., Hogers,R., Bleeker,M., Reijans,M., van de Lee,T., Hornes,M., Frijters,A., Pot,J., Peleman,J., Kuiper,M. *et al.* (1995) *Nucleic Acids Res.*, **23**, 4407–4414.
4. Ramsay,G. (1998) *Nature Biotechnol.*, **16**, 40–44.
5. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F. *et al.* (1991) *Science*, **252**, 1651–1656.
6. Adams,M.D., Dubnick,M., Kerlavage,A.R., Moreno,R., Kelley,J.M., Utterback,T.R., Nagle,J.W., Fields,C. *et al.* (1992) *Nature*, **355**, 632–634.
7. Okubo,K., Hori,N., Matoba,R., Niiyama,T., Fukushima,A., Kojima,Y. and Matsubara,K. (1992) *Nature Genet.*, **2**, 173–179.
8. Lee,N.H., Weinstock,K.G., Kirkness,E.F., Earle-Hughes,J.A., Fuldner,R.A., Marmaros,S., Glodek,A., Gocayne,J.D., Adams,M.D., Kerlavage,A.R., Fraser,C.M. and Venter,J.C. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 8303–8307.
9. Veculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) *Science*, **270**, 484–487.
10. Korenberg,J.R., Chen,X.N., Adams,M.D. and Venter,J.C. (1995) *Genomics*, **29**, 364–370. Database available at http://www.tigr.org.
11. Normore,W.M., Shapiro,H.S. and Setlow,P. (1976) In Fastman,G.D. (ed.) *CRC Handbook of Biochemistry and Molecular Biology*. CRC Press, London.
12. Jendrisak,J., Young,R.A. and Engel,J.D. (1987) In Berger,S.R. and Kimmel,A.R. (eds), *Methods in Enzymology 152*, *Guide to Molecular Cloning Techniques*. Academic Press, San Diego, CA
13. Itoh,K., Okubo,K., Yosii,J., Yokouchi,H. and Matsubara,K. (1994) *DNA Res.*, **1**, 279–287.
14. Adams,M.D., Kerlavage,A.R., Fleischmann,R.D., Fuldner,R.A., Bult,C.J., Lee,N.H., Kirkness,E.F., Weinstock,K.G., Gocayne,J.D., White,O. *et al.* (1995) *Nature*, **377**, 3–174.
15. Huynh,T.V., Young,R.A. and Davis R.W. (1984) In Glover,D. (ed.) *DNA Cloning: A Practical Approach*. IRL Press, Oxford, UK.