METHOD OF GENERATING A DATABASE

FIELD

The present invention relates to the field of biology, in particular, the fields of molecular biology, genomics, genome annotation, gene expression, transcriptome analysis and diagnostics. In particular, the present invention relates to methods of generating a database or a computer readable medium comprising a database using Serial Analysis of Gene Expression (SAGE).

BACKGROUND

One the most burdensome tasks in the post genome-sequencing era is the accurate and complete annotation of all genes and their products, primarily mRNA transcripts of a sequenced genome. Bioinformatics analyses of fragmentary experimental data have led to widely varying estimates of the number of human genes. Human EST assembly resulted in 89,000 Unigene clusters; *ab initio* genome annotation identified approximately 30,000 genes by two independent studies; and the manually curated RefSeq database has only 17,000 genes identified with stringent evidences. It is apparent that current technologies applied for genome annotation including computational gene prediction, cDNA cloning and sequencing, and other new technologies are inefficient, incomplete, and unconvincing.

Computational methods including homology studies, domain searches, and *ab initio* gene predictions have great limitation and fallibility. Current prediction programs may be fine for many 'internal' exons, but perform particularly poorly on border exons in UTR regions. They need to be trained by more experimental data. The precise annotation of every gene in complex genomes by computation methods alone is still a distant goal.

Although cDNA cloning and sequencing, including EST, full-length, and OFESTES, have generated immense data in EST and full-length cDNA sequences, low abundance and large size transcripts are discriminated against during cloning steps. Library-based cDNA approaches are incomplete for identifying all transcripts due to high redundancy of abundant transcripts and high percentage of truncated clones. A

comprehensive cDNA library approach may be efficient for capturing the first 50-70% of all expressed transcripts, but it soon becomes prohibitively expensive and inefficient for getting the rest, in particular the rare transcripts.

Genome-wide scans by oligonucleotide microarrays provide another strategy that has the potential to help annotate complex genomes. In this approach, oligo probes representing predicted exons are synthesized, micro-arrayed, and subsequently hybridised to mRNA samples. Experimental data generated would provide validations to true exons. This approach is expected to be efficient to examine many different biological stages and environmental conditions for expressed transcripts. However, a major limitation of this method is its ability to convincingly determine the existence of rare genes because the signal detection sensitivity of probe hybridization is limited.

Serial Analysis of Gene Expression (SAGE)

Serial Analysis of Gene Expression (SAGE) represents a unique strategy to identify the existence of transcripts and quantify them by counting a small tag for each transcript molecule in a complex transcriptome. Three principles underlie the SAGE methodology: (1) A short sequence tag (10-14bp) contains sufficient information to uniquely identify a transcript provided that that the tag is obtained from a unique position within each transcript; (2) Sequence tags can be linked together to from long serial molecules that can be cloned and sequenced; and (3) Quantitation of the number of times a particular tag is observed provides the expression level of the corresponding transcript. The unique feature of SAGE is that a 14-bp sequence is enough to be transcript specific and small tags from each transcript can be extracted and concatenated into larger pieces for efficient sequencing analysis. Because of all transcripts are represented by small tags in same size, there is no discrimination in SAGE tag cloning. Essentially all transcripts should be represented in SAGE tags.

SAGE is described in US Patent numbers 5,695,937, 5,866,330 and 6,383,743, and is illustrated in Figures 1A and 1B, and also in Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial Analysis Of Gene Expression. Science 270, 484-487,

as well as Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997). Characterization of the yeast transcriptome. Cell 88. A number of websites devoted to SAGE may also be consulted for teachings on this technique, including www.sagenet.org and http://www.ncbi.nlm.nih. gov/sage (SAGEnet, a public gene expression data repository and online data access and analysis site, see Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF. (2000) SAGEmap: a public gene expression resource. Genome Res 2000 Jul;10(7):1051-60). Other websites describing SAGE and its uses may be found at http://www.google.com/search?sourceid=navclient&ie=UTF-8&oe=UTF-8&q=Serial+Analysis+of+Gene+Expression

In brief, mRNA is obtained from a cell or tissue, and reverse transcribed to obtain cDNA (see Figure 1). The cDNA is then cleaved by a first restriction enzyme (the "Anchoring Enzyme", typically a 4-base cutter), and the 3' end of the cDNA is anchored to a bead. The beads are optionally divided into two pools, and the cDNA attached to the beads is ligated to two sets of adaptors or linkers. Each of these adaptors comprises a defined nucleotide sequence for PCR priming and amplification at its 5' end, as well as a recognition site for a type IIS enzyme (the "Tagging Enzyme", for example BsmFI and FokI), which directs cleavage by the enzyme at a position 3' downstream of the recognition site. The tags are released by cleavage with the relevant Tagging Enzyme, and ligated together end to end to form ditags. The ditags are then amplified using PCR, digested with the Anchoring Enzyme, and ligated together to form concatamers. Sequencing of the concatamers reveals the identity and frequency of the tags, and provides expression data for the various genes which are transcribed in the cell or tissue. Due to the efficiency of sequencing small tags, SAGE has the potential to capture all expressed transcripts.

Despite these promises however, the original SAGE tags are too short for direct mapping to complex genomes. The 14-bp tags are only reliable for mapping onto existing EST or cDNA sequences in databases, or small genomes such as yeast. This shortcoming limits the application of SAGE to use only as an expression profiling tool, not for genome

annotation. To overcome this problem, the developers of the original method managed to make the SAGE tags longer, simply by taking the advantage of a new IIS enzyme MmeI that cleaves DNA 20 base pairs away from its recognition site. The modified method is known as LongSAGE, and is described in WO 02/10438. This modification makes the LongSAGE tags specific enough to be directly mapped onto human chromosome sequences (Table 1 in **Appendix B**). This function is an important addition because new SAGE tags now can be directly marked onto specific chromosome locations for potential new genes or exons identifications, therefore, facilitating genome annotations.

However, despite its advantages, LongSAGE still has limitations to its effective use. Like the original SAGE, LongSAGE tags are extracted randomly depending on where the NlaIII sites are in a transcript sequence, so providing only 'internal' sequence clues for new transcripts. Furthermore, the new SAGE tags identified have to go through very tedious and long process such as 5' and 3' RACE to extend the information about the existence and characteristics of new genes. Finally, only one tag is generated for each expressed sequence, and it is not possible with the prior art methods to obtain further sequence information for the expressed gene readily.

The present invention seeks to solve these and other problems in the prior art techniques for expression analysis.

SUMMARY

We have realised that the prior art SAGE and LongSAGE techniques have disadvantages. In particular, we have recognised that particular disadvantages arise from the fact that the tags obtained in the prior art techniques reflect internal sequences of the expressed genes.

No one to our knowledge has appreciated the usefulness of obtaining sequence tags from the 5' and 3' termini of transcripts. On the other hand, we have realised that such terminal tags have advantages in that they can be used to define the boundary regions of

genes. In addition, the full-length coding sequence of unknown genes may easily be obtained by PCR, once the sequence tags from the 5' and 3' termini of the transcribed region are known. Furthermore, obtaining a 5' terminal transcribed sequence in a tag can provide a handle for promoter identification, as described below.

We therefore provide methods for generating a database or a computer readable medium comprising a database from tags corresponding to 5' and 3' terminal transcribed sequences of expressed genes, i.e., tags corresponding to the 5' and 3' termini of transcripts. We also describe methods of obtaining further sequence information in the form of further sequence tags from expressed genes, which may be conduced in a repetitive or iterative manner.

According to a 1st aspect of the present invention, we provide a method of generating a database comprising a plurality of records, each record comprising an indication whether a gene is expressed by a particular cell, the method comprising steps of: (a) providing a first nucleic acid sequence; (b) linking the first nucleic acid sequence to an linker sequence to form a linked nucleic acid, in which the linker sequence comprises: (i) a first recognition site for a first nucleic acid cleavage enzyme that allows cleavage of the first nucleic acid sequence at a site distant from the first recognition site, and (ii) a second recognition site for a second nucleic acid cleavage enzyme that allows nucleic acid cleavage at a site distant from the second recognition site, said cleavage site located at a position within or about the first recognition site; in which the linked nucleic acid has the structure: 5' - second recognition site - first recognition site - first nucleic acid - 3' (c) cleaving the linked nucleic acid with the first nucleic acid cleavage enzyme to provide: (i) a linked tag comprising the linker sequence linked to a nucleotide sequence tag representative of the first nucleic acid sequence and comprising a terminal portion thereof; and (ii) a second nucleic acid sequence comprising a remainder portion of the first nucleic acid. (d) detecting the presence, sequence or identity of the linked tag or the nucleotide sequence tag to provide an indication of an instance of gene expression; and storing an indication of expression of the gene as a record in the database.

Preferably, the first nucleic acid comprises a complementary deoxyribonucleic acid (cDNA) having a terminus comprising a 5' terminal transcribed sequence of a gene, and in which the linker sequence is linked to said terminus.

Preferably, the first nucleic acid comprises a complementary deoxyribonucleic acid (cDNA) having a terminus comprising a 3' terminal transcribed sequence of a gene, and in which the linker sequence is linked to said terminus.

Preferably, the second nucleic acid cleavage enzyme comprises a restriction endonuclease, preferably a Type IIS restriction endonuclease, whose recognition site is 6 bases or greater, preferably MmeI.

Preferably, the first nucleic acid cleavage enzyme comprises a restriction endonuclease, preferably a Type IIS restriction endonuclease, preferably BseRI.

Preferably, the method comprises sequentially generating a plurality of nucleic acid sequences each comprising a nucleotide sequence tag from a nucleic acid, the method comprising repeating steps (a) to (c) of the 1st aspect of the invention at least once, in which the first nucleic acid sequence of step (a) is provided by the second nucleic acid sequence of step (c)(ii); and storing each indication of an instance of gene expression generated as a record in the database.

There is provided, according to a 2nd aspect of the present invention, a database generation method of generating a database comprising a plurality of records, each record comprising an indication whether a gene is expressed by a particular cell, the method comprising the steps of: (a) providing a first linked tag and a second linked tag, each independently produced by a method according to the 1st aspect of the invention; (b) linking the first linked tag and the second linked tag such that the nucleotide sequence tag portion of one linked tag is linked to the nucleotide sequence tag of the other linked tag to form a ditag, the ditag comprising terminal transcribed sequences from first and second genes; and (c) detecting the presence or identity of the ditag, or at least one nucleotide sequence tag comprised therein, to detect gene expression; and storing the indication of

expression of the gene as a record in the database.

Preferably, each of the first and second linker sequences comprised in the first and second linked tags comprises an amplification primer sequence, and in which the method further comprises a step of amplifying the ditag, preferably by means of the polymerase chain reaction (PCR).

Preferably, the database generation method further comprises the steps of: (d) cleaving the ditag with the or each second nucleic acid cleavage enzyme; (e) linking a plurality of the resultant trimmed ditags to form a concatamer; and (f) obtaining the nucleic acid sequence of at least a portion of the concatamer.

We provide, according to a 3rd aspect of the present invention, a database generation method of generating a database comprising a plurality of records, each record comprising an indication whether a gene is expressed by a particular cell, the method comprising: (a) obtaining an indication of an instance of expression of a gene by a method comprising: (i) providing a complementary deoxyribonucleic acid (cDNA) having a terminus comprising a terminal transcribed sequence of a gene; (ii) linking the cDNA to an linker sequence thereby forming a linked nucleic acid, in which the linker sequence comprises a first recognition site for a first nucleic acid cleavage enzyme, preferably a restriction endonuclease, that allows nucleic acid cleavage at a site distant from the first recognition site; and (iii) cleaving the linked nucleic acid with the first nucleic acid cleavage enzyme to provide a linked tag, in which the linked tag comprises a nucleotide sequence tag representative of a terminal transcribed sequence of the gene; and (iv) detecting the presence or identity of the linked tag or the nucleotide sequence tag to provide an indication of an instance of gene expression; and storing the indication of expression of the gene as a record in the database.

Preferably, the 5' end of the cDNA comprises a sequence corresponding to the 5' terminal transcribed sequence of the gene, and in which the linker sequence is linked to the 5' end of the cDNA.

Preferably, the nucleotide sequence tag comprises a 5' terminal transcribed sequence of the gene, preferably at least the first 16 bases of the transcribed portion of the gene, more preferably the first 20 bases of the transcribed portion of the gene.

Preferably, the 3' end of the cDNA comprises a sequence corresponding to the 3' terminal transcribed sequence of the gene, and in which the linker sequence is linked to the 3' end of the cDNA.

Preferably, the nucleotide sequence tag comprises a 3' terminal transcribed sequence of the gene, preferably at least the last 16 bases of the transcribed portion of the gene, more preferably the last 20 bases of the transcribed portion of the gene.

Preferably, step (a) comprises: (i) deriving a cDNA from an mRNA by reverse transcription with an primer comprising a sequence 5'-NV$(T)^{13}$CCGGCCGG-3', in which N = A, C, G or T and V = A, C or G; (ii) producing a double stranded cDNA therefrom and digesting the double stranded cDNA with FseI to produce a cleaved cDNA comprising 3' $\frac{\text{GGCCGG}}{\text{CC}}$ overhang; (iii) linking the cleaved cDNA to a linker comprising $\frac{\text{pTCGGA}}{\text{GGCCAGCCT}}$; and (iv) cleaving the resulting molecule with MmeI to produce a cDNA lacking a polyA/T tail.

Preferably, the cDNA comprises the 5' terminal transcribed sequence of the gene, or the 3' terminal transcribed sequence of the gene, or both.

Preferably, the cDNA is full length cDNA, preferably comprising substantially all the coding sequence of the gene.

Preferably, the linker sequence further comprises a second recognition site for a second nucleic acid cleavage enzyme, preferably a second restriction endonuclease, that allows nucleic acid cleavage at a site distant from the second recognition site, and in which the second recognition site is located 5' of the first recognition site in the linker sequence.

Preferably, the location of the second recognition site with respect to the first recognition site is such that, when the linked tag is exposed to the second nucleic acid cleavage enzyme, cleavage of the nucleic acid occurs at a position within or about the first recognition site.

Preferably, the first recognition site or the second recognition site, or both, comprises a Type IIS restriction enzyme recognition site.

Preferably, the first restriction enzyme comprises Mme1, or in which the first recognition site comprises a Mme1 recognition site 5'-TCC RAC-3', preferably 5'-TCC GAC-3'.

Preferably, the second restriction enzyme comprises a restriction enzyme capable of recognising a site which is 6 bases or longer, preferably BseR1, or in which the second recognition site comprises a BseRI recognition site 5'-GAGGAG-3'.

Preferably, the linker sequence comprises the sequence 5'-GAGGAGNNNNNNTC▲CG▼AC -3', preferably 5'-GAGGAGCGTCTCTCCGAC-3'.

As a 4th aspect of the present invention, there is provided a database generation method of generating a database comprising a plurality of records, each record comprising an indication whether a gene is expressed by a particular cell, the method comprising steps of: (a) providing a first complementary deoxyribonucleic acid (cDNA) having a terminus comprising a terminal transcribed sequence of a first gene; (b) providing a second complementary deoxyribonucleic acid (cDNA) having a terminus comprising a terminal transcribed sequence of a second gene; (c) linking the first cDNA so produced to a first linker sequence thereby forming a first linked nucleic acid, in which the first linker sequence comprises a first recognition site for a first nucleic acid cleavage enzyme, preferably a first restriction endonuclease, that allows nucleic acid cleavage at a site distant from the first recognition site; (d) linking the second cDNA so produced to a second linker sequence thereby forming a second linked nucleic acid, in which the second linker sequence comprises a second recognition site for a second nucleic acid cleavage enzyme, preferably a second

restriction endonuclease, that allows nucleic acid cleavage at a site distant from the second recognition site; (e) cleaving the first linked nucleic acid with the first nucleic acid cleavage enzyme to provide a first linked tag, in which the first linked tag comprises a first nucleotide sequence tag representative of a terminal transcribed sequence of the first cDNA. (f) cleaving the second linked nucleic acid with the second nucleic acid cleavage enzyme to provide a second linked tag, in which the second linked tag comprises a second nucleotide sequence tag representative of a terminal transcribed sequence of the second cDNA. (g) ligating the first and second tags to form a ditag; and (h) determining the nucleotide sequence of at least one tag of the ditag to detect gene expression; and storing the indication of expression of the gene as a record in the database.

Preferably, the database generation method further comprises storing the database in a computer readable medium.

We provide, according to a 5th aspect of the present invention, a method of generating a computer readable medium comprising a database comprising a plurality of records which indicate whether a gene is expressed by a particular cell, the method comprising generating a database by a method according to any preceding aspect of the invention and storing the database on the computer readable medium.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1A shows an overview of a known SAGE technique. Immobilized cDNA fragments at 3'end on magnetic beads are digested by a 4-bp restriction enzyme NlaIII. On average, NlaIII cleaves along cDNA in every 250 base pairs. Linkers with a recognition site for BsmFI, a Type IIS type restriction enzyme, are ligated to the NlaIII cleaved end of cDNA fragments on the bead. BsmFI cuts the DNA asymmetrically [GTCCCN(14/10)] resulting in the release of a 14 base pair tag representing the corresponding transcript. The extracted tags are then ligated into ditags for PCR amplification followed by removing the linker arms with NlaIII digestion. Ditags with 4-bp anchor sequence are concatenated, cloned, and sequenced. **Figure 1B** is a schematic diagram of SAGE tag concatenate library construction.

Figure 2 is a schematic diagram of the 5'and 3'Terminal SAGE techniques described in this document and its application in genome annotation. 5'and 3'Terminal SAGE tags from full-length transcripts in a transcriptome are extracted using IIS type restriction enzymes. Tags are concatenated, cloned, sequenced, counted, and mapped to chromosome. Mapping of 5' and 3'tags to genome sequences allows precise localization and identification of genes as well as gene activities on chromosomes. Overlapping tags generated by serial rounds of SAGE Walking will immensely increase the tag specificity and confidence for identification of new gene and confirmation of predicted genes. The 5' tags also allow quick identification of large number of promoter sequences.

Figure 3 is a schematic diagram showing the 5'Terminal SAGE and SAGE Walking techniques described in this document. See Detailed Description for details.

Figure 4 is a schematic diagram showing the 3'Terminal SAGE and SAGE Walking techniques described in this document. See Detailed Description for details.

SEQUENCE LISTING

Appendix A shows the sequences of the various linkers and primers used in the methods and compositions described here, in particular in the Examples. The sequences shown are:

NotI-dT15 primer used for cDNA synthesis; Biotin-NotI linker; MmeI-BseRI Linker A (48nt); MmeI-BseRI Linker B (48nt); second round MmeI-BseRI Linker A (50nt); second round MmeI-BseRI Linker B (50nt); PCR primer A (29nt, 20nt); PCR primer B (29nt, 20nt); FseI-dT15 primer (for cDNA synthesis).; 5' biotin linker (The biotin oligo is same as the PCR primer A); 1/2FseI-MmeI linker (to introduce MmeI site and remove polyA); Second round MmeI-BseRI Linker A (50nt); Second round MmeI-BseRI Linker B (50nt); PCR primer A (29nt, 20nt); PCR primer B (29nt, 20nt).; NotI-dT20 primer; NotI Linker top; NotI Linker bottom; Linker A top (N5); Linker A top (N6); Linker A bottom; Linker B top (N5); Linker B top (N6); Linker B bottomx; PCR primer A; PCR primer B; GsuI-dT16 primer; MmeI-BseRI Linker A top; MmeI-BseRI Linker A bottom; MmeI-BseRI Linker B top; MmeI-BseRI Linker B bottom; SalI adaptor top; SalI adaptor bottom; PCR primer A; PCR primer B.The methods and compositions described here may suitably employ any one or more of the sequences shown in the Sequence Listing.

DETAILED DESCRIPTION

The methods and compositions described here enable the generation of databases comprising records indicative of gene expression. The methods involve isolation of defined nucleotide sequence tags from termini of nucleic acids. We provide methods for the

detection of gene expression in a particular cell or tissue, or cell extract, for example, including at a particular developmental stage or in a particular disease state and storing such information on a database. Nucleotide sequence tags derived from cDNA can provide information on the gene, allele, or isoform which is expressed, and their incidence in a sample reflects the level of expression of the gene. Our methods therefore enable quantitative and qualitative analysis of gene expression.

In particular, our database generating methods may involve isolating nucleotide sequence tags which reflect the 5' and / or 3' sequences of transcripts corresponding to expressed genes. Such nucleotide sequence tags are therefore representative of a terminal transcribed sequence of a gene. In general, they comprise a terminal transcribed sequence of a gene.

Such tags are referred to for convenience as "Terminal Tags", and the techniques for obtaining such terminal tags will be referred to, again for convenience only, as "Terminal SAGE".

As explained above, the prior art SAGE and LongSAGE techniques have disadvantages associated with the fact that the tags produced using these techniques are essentially "internal" in origin (i.e., they correspond to internal sequences of the expressed genes).This is because, in the prior art SAGE techniques, the "start" of the tag is defined by the 5' most (or 3' most) terminal Tagging Enzyme site on the gene, while its "end" position is defined by the offset between the Tagging Enzyme recognition site and its cutting site. While either the 5' end or the 3' end of the cDNA can be anchored in the known SAGE and LongSAGE techniques, this anchoring, and the fact that the cDNA is treated with the Anchoring Enzyme prior to ligation to the adaptors or linkers, means that the tags generated by the SAGE methods of the prior art essentially correspond to "internal" sequences.

Accordingly, pre-digestion of the cDNA with Anchoring Enzyme means that the position of the tags in relation to the entire sequence is essentially determined by the

position of the most terminal Anchoring Enzyme site in the transcribed sequence. The prior art SAGE methods therefore invariably result in "internal" nucleotide sequence tags.

In contrast, the Terminal SAGE techniques described in this document enable the isolation of nucleotide sequence tags which are "terminal" in origin. In other words, the tags obtained using our techniques comprise the 5' or 3' sequences of the processed nucleic acids and hence the transcribed sequences of the expressed genes. For convenience, we refer to the method for obtaining a 5' terminal transcribed sequence of a gene described as "5' Terminal SAGE", and the method for obtaining a 3' terminal transcribed sequence of a gene "3' Terminal SAGE".

Similarly, the nucleotide sequence tag comprising a 5' terminal transcribed sequence of a gene may conveniently be referred to as a "5' terminal tag", and the nucleotide sequence tag comprising a 3' terminal transcribed sequence of a gene may be conveniently referred to as a "3' terminal tag". However, it will be appreciated that where a nucleic acid which is not part of a gene is being processed using the methods described here, these terms may equally have more general meanings (i.e, as referring to the 5' terminal sequence or 3' terminal sequence, as the case may be, of the nucleic acid in question).

In the methods of Terminal SAGE as described here, a "first nucleic acid sequence" is first provided. This is lined with a "linker sequence" to form a "linked nucleic acid". The linker sequence comprises a "first recognition site" for a "first nucleic acid cleavage enzyme". The first recognition site is such that that allows cleavage of the first nucleic acid sequence at a site distant from the first recognition site. The linker sequence also comprises a "second recognition site" for a "second nucleic acid cleavage enzyme", that allows nucleic acid cleavage at a site distant from the second recognition site. The cleavage site for the second nucleic acid cleavage enzyme is located at a position within or about the first recognition site. The linked nucleic acid preferably has the structure: 5' - second recognition site - first recognition site - first nucleic acid - 3'

The linked nucleic acid is then cleaved with the first nucleic acid cleavage enzyme. The products of this cleavage reaction include a linked tag comprising the linker sequence linked to a nucleotide sequence tag representative of the first nucleic acid sequence and comprising a terminal portion thereof; and a second nucleic acid sequence comprising a remainder portion of the first nucleic acid.

In highly preferred embodiments, the database generation methods described here involve anchoring an entire full-length cDNA and ligating an adaptor/linker to the free end. The adaptor/linker comprises a PCR primer site, as well as two Type IIS restriction enzyme recognition sites. The first site may be referred to as the Tagging Enzyme site (here MmeI). In addition, a site which may be referred to as the "Anchor Enzyme" site (e.g., BseRI) is also provided for removal of the primer sequence following ligation and amplification of the ditags. The amplified ditags are then concatamerised, sequenced, and analysed, as described in further detail below.

It will be appreciated that where the first nucleic acid sequence comprises a cDNA, the nucleotide sequence tag which is produced from the method comprises a terminal transcribed sequence of the relevant gene. Preferably, the cDNA is first processed to remove at least a portion of the 5' untranslated region (UTR) or the 3' UTR as the case may be. Thus, for example, where a tag comprising a 3' terminal transcribed sequence is desired, the cDNA may be processed to remove at least a portion of the polyA/T tail. However, it will be appreciated that removal of the entirety of the 3' UTR such as the polyA/T tail is not strictly necessary, provided the tag is long enough to capture sequence information upstream of the polyA/T tail.

The nucleotide sequence tag which is produced may be detected to determine the identity of the first nucleic acid. Its sequence may also be determined for this. In preferred embodiments, the nucleotide sequence tag which is produced is linked "tail to tail" with another nucleotide sequence tag to produce a "ditag", which is then analysed (for details see below). The ditag may be amplified, for example, by PCR, by introducing appropriate sequences in the linker sequence. Two or more ditags may be joined together to form

concatermers, and the sequence of the concatamers determined for high throughput genomic screening.

"Nucleotide Sequence Tags"

The terms "tag" and "nucleotide sequence tag" as used in this document should be taken as synonymous with each other, and to mean a short nucleotide sequence which is diagnostic of a longer one. By this, we mean that the presence of the larger sequence may be detected by detecting the presence of the tag. Preferably, identification of the tag sequence, or at least a part of it, is sufficient to identify the longer sequence. In highly preferred embodiments, the sequence of the nucleotide sequence tag uniquely identifies a gene.

The tag sequence is typically derived from the longer sequence, and may be a fragment of it. Preferably, the length of the tag is less than 15%, preferably less than 10%, most preferably less than 5% or less of the length of the gene. The tag may be of any length, but typically, the length of the tag is between 5-40 bases, preferably between 10-30 bases, more preferably between 15-20 bases in length.

Although longer sequence tags are preferred, the methods and compositions described here may suitably be used with shorter tags, for example, tags of 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 or 15 bases. However, preferably, the tags have a minimum length of 16 bases, and are preferably 17 bases, 18 bases, 19 bases, or preferably 20 bases or more.

In preferred embodiments, where the nucleotide sequence tag is of a gene, the length of the tag (i.e., the amount of sequence derived from the longer sequence) is sufficient to provide at least a preliminary indication of the identity of the gene. Preferably, the length of the tag is sufficient to provide a unique indication of the identity of a gene, i.e., it is long enough to identify the gene.

As will be apparent from the description below, the length of the tag may be varied by choice of the first nucleic acid cleavage enzyme (i.e., the first Type IIS restriction

endonuclease in preferred embodiments), in particular, the "offset" by which the first nucleic acid cleavage enzyme cuts from its recognition sequence.

Preferably, the tag is diagnostic of a gene, or a transcription or translation product thereof. The gene, etc corresponding to the tag may be easily identified by hybridisation or library screening, or by database comparison. The nucleotide sequence tag may be indicate the presence of a gene in a cell (or organ, tissue or individual) or the expression of the gene. Where a gene is highly expressed gene by a cell, the number of copies of the nucleotide sequence tag corresponding to the gene would be expected to be higher than for a gene whose expression is relatively low. Therefore, the number of copies of the nucleotide sequence tag in any particular sample may be determined to establish the degree or amount of expression of a gene. The sequence of the tag may be derived from a portion of the sequence of the gene, which may be from a non-coding portion of the gene (e.g., an intron, an untranslated region such as a 5' UTR or a 3' UTR) or a coding portion (e.g., an exon), or a combination of the two (e.g., a junction). The sequence of the tag may correspond to a sequence of the coding strand of the gene or mRNA or cDNA, or its complement or opposite strand.

The nucleotide sequence tag may be derived from any part of the longer sequence. Where the longer sequence comprises a gene, the nucleotide sequence tag in preferred embodiments comprises a terminal transcribed sequence (the meaning of this term is explained below). In highly preferred embodiments, the nucleotide sequence tag is derived from a 5' terminal sequence, preferably a 5' terminal transcribed sequence. Alternatively, or in addition, the nucleotide sequence tag is derived from a 3' terminal sequence, preferably a 3' terminal transcribed sequence.

The nucleotide sequence tag may indicate the presence of an expression product of a gene, such as a messenger RNA. The presence or identity of the tag may be determined to establish whether a protein product is present in a cell, etc. Thus, where the "longer sequence" comprises a gene, the nucleotide sequence tag may be used for establishing or quantitating gene expression.

It will be appreciated that the gene which corresponds toe the nucleotide sequence tag may be one which is known to be expressed, whether in the particular cell, tissue or organism, or known to be expressed in general in other types of cells, tissues or organisms. Alternatively, it may be a completely unknown or new gene, or a new transcript corresponding to a known gene, whether previously known to be expressed in the particular cell, etc or not.

Our methods may therefore be used to provide sequence "signatures" corresponding to terminal portions of nucleic acids, as opposed to internal portions. Such terminal sequence information, including terminal transcribed sequence information, is useful for various purposes, as described in detail below.

"Terminal Portion"

By "terminal portion", we mean the sequence of a nucleic acid at or about the 5' or 3' terminus of the nucleic acid. The "terminal portion" may comprise any number of bases. It may comprise the extreme terminal base, or it may comprise sequence just within the terminus (i.e., lacking the extreme terminal base). The "terminal portion" may lack 1, 2, 3, 4 or 5 bases of the extreme terminus. However, in preferred embodiments the terminal portion includes residues at the extreme terminus of the nucleic acid. Thus, in such preferred embodiments, the "terminal portion" comprises, preferably consists of, the first N residues, or the last N residues, of the nucleic acid, where N = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15...29, 30, 31, 32, 33, 34 or 35 or more. In preferred embodiments, terminal portions include the extreme end, i.e., the extreme terminal base, the first or last nucleotide, residue, base or base pair in the particular nucleic acid sequence.

In relation to a coding sequence such as a gene or mRNA, the terminal portion or "terminal transcribed sequence" (and the terms "5' terminal transcribed sequence" and "3' terminal transcribed sequence") should be understood to be defined in relation to the sequence of the gene as expressed in mRNA.

That is to say, the terminal transcribed sequence should preferably include sequences at the beginning or end, as the case may be, of the transcribed portion of a relevant nucleic acid sequence (i.e., a gene, an mRNA, a cDNA or a genomic sequence). A 5' terminal transcribed sequence will therefore typically comprise a portion corresponding to the 5' untranslated region (5' UTR) of a transcript. Similarly, a 3' terminal transcribed sequence will therefore typically comprise a portion corresponding to the 3' untranslated region (3' UTR) of a transcript. Preferably, however, at least a portion of the polyA/T tail is removed from the cDNA prior to it being processed, preferably the entirety of the polyA/T tail is removed from the cDNA. Where this occurs, the 3' terminal transcribed sequence will correspond to the 3' sequence of the mRNA as it is transcribed, before addition of the polyA/T tail.

A full length cDNA comprises four sections: 5' untranslated region, coding sequences (the sequences which translated into protein), 3' untranslated region and poly A tail. Poly A sequences do not exist in the genomic DNA but are added after the mRNA was processed to remove introns. So, by removing the poly A region from the cDNA and obtaining the terminal region of 3' end, it is possible using our techniques to identify the exact 3' border of the transcripts and map back to the corresponding chromosomes.

It will be appreciated that the cDNA may be further manipulated before being processed in the Terminal SAGE methods described here. For example, one or more residues may be removed from the relevant terminus. This embodiment may be useful to remove untranslated sequences, so that the nucleotide sequence tag comprises fewer residues of untranslated sequence. The untranslated sequence may be removed entirely from the relevant terminus, to leave only coding sequence (so that the nucleotide sequence tags generated correspond to 5' or 3' terminal coding sequences).

Accordingly, the "terminal transcribed sequence" preferably includes at least an extreme upstream portion, or at least an extreme downstream portion, of the sequence of an mRNA, cDNA, etc, as the case may be. However, as with the terminal portions discussed above, the "terminal transcribed sequence" may comprise the extreme terminal

base, or it may comprise sequence just within that base (i.e., lacking the extreme terminal base). The "terminal transcribed sequence" may lack 1, 2, 3, 4 or 5 bases of the extreme terminal base. In preferred embodiments, however, the "terminal transcribed sequence" includes the first base, or the last base, as the case may be.

In such preferred embodiments, the 5' terminal transcribed sequence includes the first 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 or 30 nucleotides or bases of the transcribed portion of the gene. The 3' terminal transcribed sequence may include such numbers of nucleotides or bases of the transcribed portion of the gene, preferably not counting the polyA/T tail. In highly preferred embodiments, the terminal transcribed sequence includes sequence corresponding to at least the first or last 16 bases, preferably at least the first or last 20 bases of the transcript.

However, as noted above, the 5' terminal transcribed sequences (for example, a 5' terminal transcribed sequence of a cDNA) may comprise a sequence corresponding to a location at or about the 5' terminus of the transcribed portion, but it preferably includes the first base in the transcript. Similarly, a 3' terminal transcribed sequence, for example of a cDNA, may comprise a sequence corresponding to a location at or about the 3' terminus of the transcribed portion, but it preferably includes the last base, i.e., the last base which was transcribed from the gene.

It is apparent from the above that using our methods it is possible to determine the identities of genes which are expressed by a particular cell type, tissue, organ or individual. This enables a gene expression profile to be compiled for the relevant cell, etc. Transcriptome analysis may also be readily carried out using our methods.

Our methods therefore enable both 5' and 3' terminal transcribed sequences to be obtained with ease.

Furthermore, is will be apparent that the presence of a particular nucleotide sequence tag as a product of the reactions indicates an instance of expression of the

corresponding gene. Therefore, the level of expression of a particular gene by a particular cell, tissue, organ or individual may be established. It is also possible to determine which genes are expressed, and which genes are not expressed, in such cells, etc. Relative levels of gene expression may be compared between genes, and between cells, etc. It will be apparent that our methods may also be used to determine differences in gene expression between different states of a cell, for example, a healthy state and a diseased state. Comparison of gene expression profiles of a cell (or tissue, etc) known to be diseased and a candidate cell may be used to diagnose particular diseases, or susceptibility to such diseases. Differences in gene expression profiles between different developmental states of a cell, between pluripotent and differentiated cells, or between cells in different parts of the cell cycle may also be determined with ease.

Knowledge of 5' and/or 3' terminal sequences of a nucleic acid such as a gene provides many advantages, which have been mentioned in passing, and will be explained in greater detail below.

Other uses of the methods and compositions described here will be apparent to the skilled reader.

The practice of the present invention will employ, unless otherwise indicated, conventional techniques of chemistry, molecular biology, microbiology, recombinant DNA or immunology, which are within the capabilities of a person of ordinary skill in the art. Such techniques are explained in the literature. See, for example, J. Sambrook, E. F. Fritsch, and T. Maniatis, 1989, *Molecular Cloning: A Laboratory Manual*, Second Edition, Books 1-3, Cold Spring Harbor Laboratory Press (this reference is informally known as "Maniatis"); Ausubel, F. M. et al. (1995 and periodic supplements; *Current Protocols in Molecular Biology*, ch. 9, 13, and 16, John Wiley & Sons, New York, N.Y.); B. Roe, J. Crabtree, and A. Kahn, 1996, *DNA Isolation and Sequencing: Essential Techniques*, John Wiley & Sons; J. M. Polak and James O'D. McGee, 1990, *In Situ Hybridization: Principles and Practice*; Oxford University Press; M. J. Gait (Editor), 1984, *Oligonucleotide Synthesis: A Practical Approach*, Irl Press; and, D. M. J. Lilley and

J. E. Dahlberg, 1992, *Methods of Enzymology: DNA Structure Part A: Synthesis and Physical Analysis of DNA* Methods in Enzymology, Academic Press. Each of these general texts is herein incorporated by reference.

TERMINAL SAGE

In a general aspect, the database generation methods described here involve the production of a nucleotide sequence tag, or a number of nucleotide sequence tags, from the terminus of a nucleic acid such as a transcribed gene. In particular, our methods may involve isolating nucleotide sequence tags which reflect the 5' and / or 3' termini of the transcribed sequences of expressed genes.

Our methods generally involve providing a first nucleic acid sequence of interest. Our methods are suitable for use on any nucleic acid sequence, but are most useful for DNA sequences, such as complementary deoxyribonucleic acid (cDNA) sequences. Use of cDNA sequences, typically generated for example by reverse transcription from an mRNA extracted from a cell, or derived from a library, is preferred for the gene expression aspects of the methods and compositions described here.

In such embodiments, the cDNA preferably comprises a terminal transcribed sequence of the relevant gene, that is, either a 5' terminal transcribed sequence, or a 3' terminal transcribed sequence, or both. A number of methods are known for obtaining such cDNA, and are also disclosed in detail below. Where a tag with 5' sequence is desired, "full length" cDNA with 5' terminal transcribed sequence may be provided. To obtain a tag with 3' terminal transcribed sequence, techniques may be employed to remove at least some or all of the polyT tail or optionally some non-coding sequence; alternatively, the length of the tag may be chosen so that it is long enough to span the polyA/T tail and capture at least some 3' terminal transcribed sequence of the cDNA in the tag.

Generation of Linked Nucleic Acids

The first nucleic acid is then linked to a linker sequence without any prior processing, in particular without any prior cleavage step. Accordingly, in contrast to the prior art methods, the first nucleic acid sequence (in preferred embodiments, a cDNA) is not digested by restriction endonucleases prior to linking to the linker sequence. Thus, in our Terminal SAGE methods no pre-digestion with a 4-base cutter (i.e., NlaIII) is carried out before ligation to the adaptor. It is this difference that enables our Terminal SAGE technique to generate sequence tags corresponding to the extreme ends of the DNA, instead of internal sequence tags.

In preferred embodiments of the database generation methods described here, therefore, the cDNA may be attached to the linker as soon as it is made, so long as the terminal transcribed portions are present in the cDNA. This enables sequence information from terminal transcribed portions of the gene or cDNA to be present in the nucleotide sequence tags which are produced by our methods.

The linker sequence is linked to the 5' terminus or the 3' terminus of the first nucleic acid, depending on whether sequence information is to be obtained from the 5' or 3' end. Thus, in preferred embodiments where it is desired to obtain 5' terminal transcribed sequence from a cDNA (i.e., in the method we refer to as "5' Terminal SAGE"), the linker is linked to the 5' terminus of the cDNA. Likewise, to obtain nucleic acid sequence information from the 3' terminal transcribed sequence of a cDNA in preferred embodiments ("3' Terminal SAGE"), the linker is linked to the 3' terminus of the cDNA.

Linkage of the first nucleic acid to the linker sequence produces a construct we refer to as a "linked nucleic acid".

The linkage of the first nucleic acid to the linker sequence to produce the linked nucleic acid may be by any means, but is preferably by ligation (for example, by use of ligases). The first nucleic acid may be immobilised during linkage, for example, on a

microbead. Immobilisation eases the purification of products, but is not strictly necessary. Where immobilisation is carried out, the first nucleic acid is immobilised at the terminus which is not being linked. That is to say, if it is desired to obtain a nucleotide sequence tag from one end of the nucleic acid, the nucleic acid may be immobilised on a bead at the other end. For example, the 5' cap of a transcript (the cDNA) can also be utilized for labeling or binding a capture means for isolation of a nucleotide sequence tag corresponding to the 3' terminal transcribed sequence.

The first nucleic acid may be provided with a capture means for this purpose. The capture means may comprise a binding element, as known in the art, such as biotin, strepavidin, digoxigenin, etc. Any suitable binding element or capture means may be employed. The microbead may be magnetic.

Linker Sequences

The linker sequence at least comprises a first recognition site for a first nucleic acid cleavage enzyme. The first nucleic acid cleavage enzyme is such that it is capable of cleaving nucleic acid at a site remote and preferably downstream from its recognition sequence. For example, where 5' sequence information is desired from a cDNA sequence, the cleavage site of the first nucleic acid cleavage enzyme is preferably 3' of the recognition sequence, relative to the coding strand of the cDNA.

Linker sequences may be constructed using means known in the art, for example, known oligonucleotide synthesis techniques.

The location of the first recognition site for the first nucleic acid is such that the cleavage site is preferably present in the first nucleic acid sequence. For example, the cleavage site may be located in the sequence of the cDNA which is linked to the linker sequence. Such an arrangement enables the capture of some or more sequence information from the cDNA by cleavage of the nucleic acid by the first nucleic acid cleavage enzyme, as described later. Needless to say, the offset between the recognition site and the cleavage

site of the first nucleic acid cleavage enzyme may be manipulated to obtain more or less sequence information, again as described below.

The first nucleic acid cleavage enzyme preferably comprises a restriction enzyme or restriction endonuclease, preferably a Type IIS restriction endonuclease. However, the use of artificial or modified enzymes is envisaged, as described in detail below. The first nucleic acid cleavage enzyme (such as a restriction endonuclease) may sometimes be referred to as a "Tagging Enzyme". The distance (i.e., the number of bases) between the recognition site of the nucleic acid cleavage enzyme and its cleavage site may be varied, but will typically be between 5 and 15 or more (for example 20, 25, 30 etc). The larger this "offset", the more sequence information is captured in the tag (as will be explained below).

In highly preferred embodiments, the first nucleic acid cleavage enzyme comprises a restriction endonuclease, preferably a Type IIS restriction endonuclease, preferably BseRI.

The linker sequence may further comprise a second recognition site for a second nucleic acid cleavage enzyme. The second nucleic acid cleavage enzyme may sometimes be referred to as an "Anchoring Enzyme". The second recognition site enables the second nucleic acid cleavage enzyme to cleave at a site distant from the second recognition site. Such an embodiment is useful where high throughput analysis by means of ditags is desired (see later). The second recognition site where present is located on the linker sequence such that the second nucleic acid cleavage enzyme cleavage site is at a position within or about the first recognition site. Preferably, the cleavage site of the second nucleic acid cleavage enzyme is located within the first recognition site.

Preferably, the second nucleic acid cleavage enzyme comprises a restriction endonuclease, preferably a Type IIS restriction endonuclease, whose recognition site is 6 bases or greater, preferably MmeI. Preferably, the second nucleic acid cleavage enzyme is one which produces overhangs (i.e., is not a "blunt cutter").

In preferred embodiments, the linked nucleic acid comprising the first nucleic acid sequence and the linker sequence has the arrangement 5' - second recognition site - first recognition site - first nucleic acid - 3'. In such embodiments, therefore, the first recognition site is located between the first nucleic acid and the second recognition site.

Generation of Linked Tags and Nucleotide Sequence Tags

The linked nucleic acid is then subject to cleavage by the first restriction enzyme. This produces two products. The first product comprises the linker sequence linked to a terminal portion of the first nucleic acid sequence, and may be referred to as a "linked tag". Such a terminal portion comprises the nucleotide sequence tag, and is representative of the first nucleic acid sequence. Preferably, the first nucleic acid cleavage enzyme is one which generates an overhang, preferably a 3' overhang, and preferably therefore the nucleotide sequence tag comprises a 3' overhang at its 3' end.

The second product of the cleavage of the linked nucleic acid with the first nucleic acid cleavage enzyme comprises the remainder of the first nucleic acid, lacking the terminal portion (see below for a use of this in generating further tags).

It will be apparent that detecting the nucleotide sequence tag enables the determination of the identity of the first nucleic acid sequence. Thus, in the preferred embodiment where the first nucleic acid sequence comprises a cDNA, the nucleotide sequence tag comprises a terminal transcribed portion of the cDNA, and hence is representative of the terminal transcribed sequence of the expressed gene.

TAG ANALYSIS

The nucleotide sequence tags generated may be identified, and in addition may have their sequences determined.

Detection of the nucleotide sequence tag may be undertaken by any means, for example, by hybridisation, pull down, gel analysis, fragment length analysis, antibody

binding, Single Strand Conformation Polymorphism (SSCP) analysis, mass spectrometry, MALDI, etc. Clonal analysis, as described in further detail below, may also be carried out. Quantitation of the amount of sequence tag may be done by any means known in the art, for example, hybridisation using a suitably labelled probe and quantitation of signal emitted by bound probe. Radioactive and non-radioactive methods are envisaged for this.

For example, the tags may be used to screen an appropriate library to obtain further sequence information. The tag itself may be used as a probe, or a oligonucleotide probe produced which comprises or corresponds to the tag sequence. The probe is used to screen for example a cDNA library, and a cDNA clone obtained. The cDNA may be sequenced and compared to databases to determine its identity.

The term "oligonucleotide" as used herein refers to primers or oligomer fragments comprised of two or more deoxyribonucleotides or ribonucleotides, preferably more than three. The exact size will depend on many factors, which in turn depend on the ultimate function or use of the oligonucleotide.

In preferred embodiments, however, the nucleotide sequence tag is detected or quantitated, preferably both, by determining its sequence. This may be achieved by isolating the various nucleotide sequence tags obtained by the procedure, cloning or subcloning them into suitable vectors, and sequencing the cloned tags.

As noted above, the nucleotide sequence tags resulting from a reaction may be analysed or sequenced individually.

For example, clonal sequencing may be carried out to identify the tags produced. Clonal sequencing is similar to limiting dilution techniques used in the cloning of cell lines, and involves dilution of linked tags, ditags or concatamers thereof, and placement into individual receptacles, such that each receptacle contains about one DNA molecule per receptacle. The DNA in each receptacle may then be amplified and sequenced by standard methods known in the art, including mass spectroscopy.

We provide an oligonucleotide composition having at least two defined nucleotide sequence tags, wherein at least one of the sequence tags corresponds to at least one expressed gene, and in which at least one of the nucleotide sequence tags is generated by a Terminal SAGE method as described here. The composition may comprise about 1 to 200 ditags, and preferably about 8 to 20 ditags. Such compositions are useful for the analysis of gene expression by identifying the defined nucleotide sequence tag corresponding to an expressed gene in a cell, tissue or cell extract, for example.

Ditag Analysis

However, in addition to, or as an alternative to, such individual analysis, the tags may be combined or linked into multimeric structures, and the multimeric structures themselves analysed to determine the identity and/or quantity of the component tags, and hence the expressed genes in preferred embodiments.

By the term "multimeric", we intend to include anything that comprises more than one monomer. Hence, multimeric includes dimers, trimers, tetramers, etc.

We therefore provide a method of detecting gene expression, the method comprising the steps of: (a) providing a first linked tag and a second linked tag, each independently produced by a Terminal SAGE method as described; (b) linking the first linked tag and the second linked tag such that the nucleotide sequence tag portion of one linked tag is linked to the nucleotide sequence tag of the other linked tag to form a ditag, the ditag comprising terminal sequences from first and second genes; and (c) detecting the presence or identity of the ditag, or at least one nucleotide sequence tag comprised therein, to detect gene expression.

Thus, in such preferred embodiments, the tags which result from the methods as described above are dimerised to form ditags. By "ditag", we mean a nucleic acid sequence which comprises two linked tags joined in any fashion. The ditag may comprise only sequences from the relevant genes, or it may further include one or more linker sequences, preferably at the ends.

Each ditag therefore represents two defined nucleotide sequences of at least one transcript, representative of at least one gene. Typically, a ditag represents two transcripts from two distinct genes. The presence of a defined nucleotide sequence tag within the ditag is indicative of expression of a gene having a sequence of that tag. As described below, ditags may be amplified before analysis. However, the analysis of ditags, formed prior to any amplification step, provides a means to eliminate potential distortions introduced by amplification, e.g., PCR.

The pairing of tags for the formation of ditags is a random event. The number of different tags is expected to be large, therefore, the probability of any two tags being coupled in the same ditag is small, even for abundant transcripts. Therefore, repeated ditags potentially produced by biased standard amplification and/or cloning methods are readily excluded from analysis.

The dimerisation of the tags into ditags is preferably done in a "tail to tail" fashion, with the portion of the nucleotide sequence tag furthest from the linker sequence being referred to as the "tail". That is to say, two linked tags are linked to each other via their nucleotide sequence tag portions. The "linker" portions are therefore at the termini of the ditags, with the two nucleotide sequence tags in the middle sandwiched therebetween. Accordingly, the ditag preferably has a structure 5' - linker sequence (I) - nucleotide sequence tag (I) - nucleotide sequence tag (II) - linker sequence (II) - 3'.

Cleavage of the linked nucleic acid may give rise to blunt ends or overhanging ends - e.g., 5' or 3' overhangs. In the latter case, the two linked tags are conveniently linked via their overhangs (see Figure 3 for an example of such linking by overhangs). Overhangs enable efficient linkage or ligation. Linkage via blunt ends is also possible, where the cleavage reaction produces such blunt ends. Preferential linkage at such blunt ends may be achieved by using linkers which are dephosphorylated at their 5' ends. Of course, it will be appreciated that overhanging ends may be "polished", or made blunt, if needed. This may be achieved by exposing the nucleic acid to any suitable enzyme which removes the overhangs, such as an enzyme comprising the appropriate exonuclease

activity. Thus, an enzyme comprising 3'-5' exonuclease activity may be used to remove 3' overhangs, and an enzyme comprising 5'-3' exonuclease activity may be used to remove 5' overhangs. For example, the 3'-5' exonuclease activity of Klenow fragment may be used to generate blunt ends from a 3' overhang (as described in Maniatis).

However, it will be appreciated that conversion from overhangs to blunt ends results in loss of sequence information. For example, where 20 base tags are produced with 2 base overhangs, ligation or linkage of such tags will produce 38 base ditags comprising 20 bases of useful information in each tag. In contrast, removal of the overhangs will result in the loss of information (2 bases) at the 3' end of each tag.

Linkage of the two linked tags may be by any means, but is preferably through DNA or RNA ligases. Such ligases, for example, T4 RNA ligase and T4 DNA ligase, are known in the art.

The ditags once generated may be detected by hybridisation, etc, or sequenced directly. Alternatively, and preferably, a number of ditags are linked to each other to form concatamers, and the concatamers sequenced. Such an embodiment is suitable for high throughput analysis of gene expression.

Ditag Amplification

The ditags may be amplified prior to their being detected, sequenced or concatermerised. Amplification of the ditags increases their number, and hence increases the sensitivity of the detection or sequencing of the tags or ditags.

Amplification may be by any means, but is preferably by means of the polymerase chain reaction (PCR). Polymerase chain reaction technologies are described in detail in Dieffenbach CW and GS Dveksler (1995, PCR Primer, a Laboratory Manual, Cold Spring Harbor Press, Plainview NY) and U.S. Pat. No. 4,683,195.

The ditag can be amplified by utilizing primers which specifically hybridize to one strand of each linker. For this purpose, the linker sequence may comprise a suitable amplification primer sequence.

The term "primer" as used herein refers to an oligonucleotide, whether occurring naturally or produced synthetically, which is capable of acting as a point of initiation of synthesis when placed under conditions in which synthesis of primer extension product which is complementary to a nucleic acid strand is induced, i.e., in the presence of nucleotides and an agent for polymerization such as DNA polymerase and at a suitable temperature and pH. The primer is preferably single stranded for maximum efficiency in amplification. Preferably, the primer is an oligodeoxy ribonucleotide. The primer must be sufficiently long to prime the synthesis of extension products in the presence of the agent for polymerization. The exact lengths of the primers will depend on many factors, including temperature and source of primer.

The primers herein are selected to be "substantially" complementary to the different strands of each specific sequence to be amplified. This means that the primers must be sufficiently complementary to hybridize with their respective strands. Therefore, the primer sequence need not reflect the exact sequence of the template. Preferably, the primers are substantially complementary to at least part of the linker sequences.

The amplification primer sequences in the two linked tags may be the same, or different. In preferred embodiments, therefore, the linked tags are generated by the methods described previously, and ligated to form ditags using DNA ligase, for example as shown in Figure 3 or 4. Appropriate PCR primer sequences, dNTPs, polymerase, etc, and a suitable reaction buffer are then added to the ditags, and the mixture subjected to rounds of amplification as known in the art. Alternatively, the ditags can be amplified by cloning in prokaryotic-compatible vectors or by other amplification methods known to those of skill in the art.

Ditag Concatamerisation

As noted above, the ditags may be formed into higher order structures for processing and analysis. For example, concatamers or polymers of the ditags may be formed. Thus, in preferred embodiments, the identity of the nucleotide sequence tag is determined by producing and concatamerising ditags. This allows large scale nucleotide sequence tag analysis to be carried out.

In preferred embodiments, all, or most, of the ditags arising from a reaction are concatermerised. In such an embodiment, the concatamer may be sequenced to determine the sequence of the component tags and ditags, without the necessity of sequencing the tags and ditags individually. The ditags may be linked using ligase, in the same manner as for the ditags.

Appropriate "spacer" sequences may be introduced or incorporated into the concatamer to separate units of sequence which define the individual nucleotide sequence tags from which the concatamers are built up. The spacer sequences may be specifically introduced in the concatamerisation reaction, by the use of appropriate linkers or adaptors, or they may be carried over from previous reactions. In the latter case, the spacer sequences may comprise the linker sequences, or portions thereof. For example, and as illustrated in Figures 3 and 5, the ditags which are formed may be ligated end to end to form concatamers; in such a situation, two nucleotide sequence tags are separated by two linker sequences forming the spacer.

The concatamer may therefore have the structure 5' - linker sequence (I) - nucleotide sequence tag (I) - nucleotide sequence tag (II) - linker sequence (II) - linker sequence (III) - nucleotide sequence tag (IIII) - nucleotide sequence tag (IV) - linker sequence (IV) - etc - 3' or 5'-LNNL-LNNL-LNNL-LNNL-LNNL-LNNL-LNNL-LNNL-LNNL-3', where L is a linker sequence, and N is a nucleotide sequence tag.

The presence of the "spacer" sequences enables the units comprising nucleotide sequence tags to be precisely defined. They may be ignored or stripped out during analysis of the sequence.

Alternatively, or in addition, the ditags may be treated or processed before being linked to form concatamers. Such treatment may comprise stripping or removing at least a portion of the linker sequence from the ditags. In preferred embodiments, the treatment comprises removing the bulk or the majority of the linker sequence from the ditag.

In preferred embodiments, the linker sequence comprises a second recognition site for a second nucleic acid cleavage enzyme as described above, and the linked tag or ditag is treated with the second nucleic acid cleavage enzyme, preferably BseRI. As explained above, the second recognition site is such that it directs nucleic acid cleavage by the second nucleic acid cleavage enzyme at a site remote from the recognition site. The separation between the first recognition site and the second recognition site is such that the second recognition site directs the second nucleic acid cleavage enzyme to cleave the linked tag at or about the first recognition site (see Figures 3 and 5 for illustrative examples).

In preferred embodiments, therefore, cleavage with the second nucleic acid cleavage enzyme removes all but 2-4 bases of the linker sequence from the ditag. Preferably, cleavage with the second nucleic acid cleavage enzyme produces a two base overhang linked to one strand of the nucleotide sequence tag. Preferably, the remnant of the linker sequence comprises a 4 base sequence comprising a 5' overhang of two bases. Thus, preferably, treatment with the second nucleic acid cleavage enzyme results in a "trimmed" ditag comprising the structure:

ACXXXXXXXXXXXXXXXXXXXXZZZZZZZZZZZZZZZZZZZGTCG
GCTGXXXXXXXXXXXXXXXXXXXZZZZZZZZZZZZZZZZZZZZCA

Where Xs and Ys represent sequences (the length of which may vary) from two nucleotide sequence tags.

In this preferred embodiment, the concatamer may be easily formed by joining the trimmed ditags by their CG/ GC overhangs. The concatamer may be isolated and cloned into a suitable vector for handling, propagation and sequencing.

The trimmed ditag preferably comprises between 12 to 120 base pairs, preferably between 18 to 46 base pairs, preferably 40 base pairs.

Cloning and Sequencing

The linked tags, ditags, or concatamers once generated may be detected by hybridisation, etc, or sequenced directly. They may then be sequenced.

Among the standard procedures for cloning the nucleotide sequence tags is insertion of the tags into vectors such as plasmids or phage. The linked tag, ditag or concatamers of ditags produced by the method may be cloned into recombinant vectors for further analysis, e.g., sequence analysis, plaque/plasmid hybridization using the tags as probes, by methods known to those of skill in the art.

The term "recombinant vector" refers to a plasmid, virus or other vehicle known in the art that has been manipulated by insertion or incorporation of the linked tag, ditag or concatamer genetic sequences. Such vectors contain a promoter sequence which facilitates the efficient transcription of the a marker genetic sequence for example. The vector typically contains an origin of replication, a promoter, as well as specific genes which allow phenotypic selection of the transformed cells. Vectors suitable for use include for example, pBlueScript (Stratagene, La Jolla, Calif.); pBC, pSL301 (Invitrogen) and other similar vectors known to those of skill in the art. Preferably, the linked tags, ditags or concatamers thereof are ligated into a vector for sequencing purposes.

Vectors in which the linked tags, ditags or concatamers are cloned can be transferred into a suitable host cell. "Host cells" are cells in which a vector can be propagated and its DNA expressed. The term also includes any progeny of the subject host cell. It is understood that all progeny may not be identical to the parental cell since there

may be mutations that occur during replication. However, such progeny are included when the term "host cell" is used. Methods of stable transfer, meaning that the foreign DNA is continuously maintained in the host, are known in the art.

Transformation of a host cell with a vector containing ditag(s) may be carried out by conventional techniques as are well known to those skilled in the art. Where the host is prokaryotic, such as E. Coli, competent cells which are capable of DNA uptake can be prepared from cells harvested after exponential growth phase and subsequently treated by the CaCl.sub.2 method using procedures well known in the art. Alternatively, MgCl.sub.2 or RbCl can be used. Transformation can also be performed by electroporation or other commonly used methods in the art.

The linked tags, ditags etc present in a particular clone can be sequenced by standard methods (see for example, Current Protocols in Molecular Biology, supra, Unit 7) either manually or using automated methods.

DATABASE COMPARISONS

One of the differences between the 5' and 3' Terminal SAGE methods described here, as compared to the prior art methods, is that the tags are extracted from the terminal 5' ends and 3' ends. In contrast, the prior art methods extract tags which are internal to the sequence; these tags are those that flank anchoring sites such as NlaIII or DpnII.

Accordingly, the methods described here are particularly useful for genome analysis. Such genome analysis is advantageously undertaken through database comparisons. The generation of 5' terminal tags using the methods described here enable the generation of sequence information which represents the starting of genes or transcripts. Likewise, generation of 3' terminal tags provide sequence information representing the end point of genes or transcripts. Such terminal tags may be compared against databases of known genes, expressed sequence tags, genomic databases, etc, to identify the nature of the gene which is expressed. With this direct mapping, expression

frequency of known genes located on genome sequences can be measured by the number of appearances of the corresponding 5' and 3' tags. Thus, the higher the number of tags corresponding to a particular gene is present, the higher its level of expression in the particular sample.

Hence, in a preferred embodiment, database comparison is carried out by: making the 5' and 3' tags pairs that most likely represent the staring points and end points of genes or transcripts and mapping the tag pairs directly to genome sequences. In addition, alternative transcription start sites and polyadenalytion sites can be identified and quantified. Furthermore, previously uncharacterised genes on chromosomal sequences can be readily identified and quantified by this direct mapping of tag pairs.

The sequences of the tags, ditags, and/or concatamers obtained may be compared with suitable sequence databases, such as GenBank, to determine their identity and/or relationships. Furthermore, it will be appreciated that comparison may be made to databases such as chromosomal DNA sequence databases and cDNA databases.

We therefore provide a method of identifying a first mRNA molecule or a first cDNA molecule reverse transcribed from the first mRNA molecule with a known sequence in a database, comprising the step of: matching a first nucleotide sequence which is located at a defined position at the 5' or 3' terminus of a transcribed sequence within the first mRNA or first cDNA molecule to a second nucleotide sequence in a database consisting of mRNA and/or cDNA sequences which occur at the defined position in their respective mRNA or cDNA molecules, whereby the first nucleotide sequence is identified with the known sequence in the database.

We also provide for a method of identifying a first mRNA molecule or a first cDNA molecule reverse transcribed from the first mRNA molecule with a known sequence in a database, comprising the step of: matching a first nucleotide sequence which is located at a defined position in a first mRNA or first cDNA molecule, wherein the defined position is at the 5' or 3 terminus of a transcribed sequence in the first mRNA or

first cDNA molecule, to a second nucleotide sequence in a database; determining that the second nucleotide sequence in the database is located at the defined position in its respective mRNA or cDNA molecule, whereby the first nucleotide sequence is identified with the known sequence in the database.

We further provide a method of identifying a cDNA molecule which is not represented in a database, comprising the steps of: comparing a first nucleotide sequence which has a predefined position at the 5' or 3' terminus of a transcribed sequence within a messenger RNA, or a cDNA molecule reverse transcribed from the messenger RNA to a database of nucleotide sequences; if no nucleotide sequences are found in the database which both match the first nucleotide sequence and occur at the defined position in an mRNA or cDNA, then hybridizing an oligonucleotide comprising the first nucleotide sequence to a cDNA clone in a library; and determining that the first nucleotide sequence is located at the defined position in the cDNA clone, whereby the cDNA molecule is identified which was not present in the database.

It will be appreciated that it is not necessary to merely compare a single nucleotide sequence against the database, and that it is possible (and perhaps more efficient) to compare pairs of nucleotide sequences against the database. Such pairs preferably comprise a nucleotide sequence corresponding to the 5' terminus of a transcribed sequence (i.e., a "5' terminal tag"), and a nucleotide sequence corresponding to the 3' terminus of a transcribed sequence (i.e., a "3' terminal tag"). In preferred embodiments, the pair comprises a 5' terminal tag and a 3' terminal tag of the same transcribed sequence (e.g., the same gene). Where there is uncertainty, the pair comprises terminal tags which are most likely to represent the ends of the relevant sequence. Thus, the pair in such preferred embodiments contains the information from the starting point and the ending point of a gene, transcript, etc. The pair therefore is sufficient to represent a unit of transcription, a full-length mRNA, or a gene.

Comparison of such pairs against the database may confirm the presence of a known corresponding gene or cDNA having the corresponding termini within the

database, indicating that that known sequence within the database is expressed "in real life" in the sample (i.e., tissue, cell, etc) in question. However, it may be the case that a match is not found in the database; this would indicate that a sequence having the relevant termini was not previously known to have existed or be expressed. Such sequences may represent new genes, or previously unknown transcripts of known genes (having for example different transcription starting sites or different polyadenalytion sites, or both).

Identification of such unknown genes or novel transcripts will aid in genomic mapping and annotation. It will be appreciated that database comparisons using the nucleotide sequence tags generated by our methods, particularly in pairs as described above, will (in cases of new genes or new transcripts) immediately provide boundary sequences of the particular new genes or transcripts, as their 5' termini and 3' termini will be known. Full-length cDNA sequences of the newly identified genes can be cloned by using conventional means. Full-length sequences of the new transcripts may similarly be isolated. It will be appreciated that primers designed from the tag sequences would enable the cloning of the full length sequence from PCR of a genomic or cDNA library or RT-PCR from mRNA.

We also provide a method to verify computationally predicted genes based on genome sequences, comprising the steps of: mapping the pair of 5' and 3' terminal tag sequences directly to the genome sequences; identifying the tag sequences that are mapped to the predicted regions. The match of the tag sequences to the predicted gene sequences will provide evidence of the existence of the genes.

In general, the nucleotide sequence tags, linked tags, ditags and concatamers may be compared against sequence databases in the same manner as described in the prior art SAGE and LongSAGE techniques. Reference is therefore made to descriptions of these techniques, in particular, US Patent Numbers 695,937, 5,866,330 and 6,383,743 as well sa WO 02/10438, hereby incorporated by reference. The following paragraphs in this section are adapted from US 6,383,743.

It is envisioned that the identification of differentially expressed genes using the Terminal SAGE method as described here can be used in combination with other genomics techniques. For example, individual nucleotide sequence tags or linked tags, and preferably ditags, can be hybridized with oligonucleotides immobilized on a solid support (e.g., nitrocellulose filter, glass slide, silicon chip). Such techniques include "parallel sequence analysis" as described below.

Briefly, parallel sequence analysis is performed after ditag preparation, wherein the oligonucleotide sequences to which the ditags are hybridized are preferably unlabeled and the ditag is preferably detectably labeled. Alternatively, the oligonucleotide can be labeled rather than the ditag. The ditags can be detectably labeled, for example, with a radioisotope, a fluorescent compound, a bioluminescent compound, a chemiluminescent compound, a metal chelator, or an enzyme. Those of ordinary skill in the art will know of other suitable labels for binding to the ditag, or will be able to ascertain such, using routine experimentation. For example, PCR can be performed with labeled (e.g., fluorescein tagged) primers. Preferably, the ditag contains a fluorescent end label.

The labeled or unlabeled ditags are separated into single-stranded molecules which are preferably serially diluted and added to a solid support (e.g., a silicon chip as described by Fodor, et al., Science, 251:767, 1991) containing oligonucleotides representing, for example, every possible permutation of a 10-mer (e.g., in each grid of a chip). The solid support is then used to determine differential expression of the tags contained within that support (e.g., on a grid on a chip) by hybridization of the oligonucleotides on the solid support with tags produced from cells under different conditions (e.g., different stage of development, growth of cells in the absence and presence of a growth factor, normal versus transformed cells, comparison of different tissue expression, etc). In the case of fluoresceinated end labeled ditags, analysis of fluorescence is indicative of hybridization to a particular 10-mer. When the immobilized oligonucleotide is fluoresceinated for example, a loss of fluorescence due to quenching (by the proximity of the hybridized ditag to the labeled oligo) is observed and is analyzed for the pattern of gene expression.

The concept of deriving a defined tag from a sequence is useful in matching tags of samples to a sequence database. In the preferred embodiment, a computer method may be used to match a sample sequence with known sequences.

In one embodiment, a sequence tag for a sample is compared to corresponding information in a sequence database to identify known sequences that match the sample sequence. One or more tags can be determined for each sequence in the sequence database as the N base pairs adjacent to each anchoring enzyme site within the sequence. However, in the preferred embodiment, only the first anchoring enzyme site from the 3' end is used to determine a tag. In the preferred embodiment, the adjacent base pairs defining a tag are on the 3' side of the anchoring enzyme site, and N is preferably 9.

A linear search through such a database may be used. However, in the preferred embodiment, a sequence tag from a sample is converted to a unique numeric representation by converting each base pair (A, C, G, or T) of an N-base tag to a number or "tag code" (e.g., A=0, C=1, G=2, T=3, or any other suitable mapping). A tag is determined for each sequence of a sequence database as described above, and the tag is converted to a tag code in a similar manner. In the preferred embodiment, a set of tag codes for a sequence database is stored in a pointer file. The tag code for a sample sequence is compared to the tag codes in the pointer file to determine the location in the sequence database of the sequence corresponding to the sample tag code. (Multiple corresponding sequences may exist if the sequence database has redundancies).

Figure 6 of US 6,383,743 is a block diagram of a tag code database access system. A sequence database 10 (e.g., the Human Genome Sequence Database) is processed as described above, such that each sequence has a tag code determined and stored in a pointer file 12. A sample tag code X for a sample is determined as described above, and stored within a memory location 14 of a computer. The sample tag code X is compared to the pointer file 12 for a matching sequence tag code. If a match is found, a pointer associated with the matching sequence tag code is used to access the corresponding sequence in the sequence database 10.

The pointer file 12 may be in any of several formats. In one format, each entry of the pointer file 12 comprises a tag code and a pointer to a corresponding record in the sequence database 12. The sample tag code X can be compared to sequence tag codes in a linear search. Alternatively, the sequence tag codes can be sorted and a binary search used. As another alternative, the sequence tag codes can be structured in a hierarchical tree structure (e.g., a B-tree), or as a singly or doubly linked list, or in any other conveniently searchable data structure or format.

In the preferred embodiment, each entry of the pointer file 12 comprises only a pointer to a corresponding record in the sequence database 10. In building the pointer file 12, each sequence tag code is assigned to an entry position in the pointer file 12 corresponding to the value of the tag code. For example, if a sequence tag code was "1043", a pointer to the corresponding record in the sequence database 10 would be stored in entry #1043 of the pointer file 12. The value of a sample tag code X can be used to directly address the location in the pointer file 12 that corresponds to the sample tag code X, and thus rapidly access the pointer stored in that location in order to address the sequence database 10.

Because only four values are needed to represent all possible base pairs, using binary coded decimal (BCD) numbers for tag codes in conjunction with the preferred pointer file 12 structure leads to a "sparse" pointer file 12 that wastes memory or storage space. Accordingly, preferably each tag code is transformed to number base 4 (i.e., 2 bits per code digit), in known fashion, resulting in a compact pointer file 12 structure. For example, for tag sequence "AGCT", with A=00.sub.2, C=01.sub.2, G=10.sub.2, T=11.sub.2, the base four representation in binary would be "00011011".

In contrast, the BCD representation would be "00000000 00000001 00000010 000000011". Of course, it should be understood that other mappings of base pairs to codes would provide equivalent function.

The concept of deriving a defined tag from a sample sequence is also useful in comparing different samples for similarity. In the preferred embodiment, a computer method is used to match sequence tags from different samples. For example, in comparing materials having a large number of sequences (e.g., tissue), the frequency of occurrence of the various tags in a first sample can be mapped out as tag codes stored in a distribution or histogram-type data structure. For example, a table structured similar to pointer file 12 in Figure 4 of US6,383,743 may be used where each entry comprises a frequency of occurrence valve. Thereafter, the various tags in a second sample can be generated, converted to tag codes, and compared to the table by directly addressing table entries with the tag code. A count can be kept of the number of matches found, as well as the location of the matches, for output in text or graphic form on an output device, and/or for storage in a data storage system for later use.

The tag comparison aspects may be implemented in hardware or software, or a combination of both. Preferably, these aspects are implemented in computer programs executing on a programmable computer comprising a processor, a data storage system (including volatile and non-volatile memory and/or storage elements), at least one input device, and at least one output device. Data input through one or more input devices for temporary or permanent storage in the data storage system includes sequences, and may include previously generated tags and tag codes for known and/or unknown sequences. Program code is applied to the input data to perform the functions described above and generate output information. The output information is applied to one or more output devices, in known fashion.

Each such computer program is preferably stored on a storage media or device (e.g., ROM or magnetic diskette) readable by a general or special purpose programmable computer, for configuring and operating the computer when the storage media or device is read by the computer to perform the procedures described herein. The inventive system may also be considered to be implemented as a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a

computer to operate in a specific and predefined manner to perform the functions described herein.

SAGE WALKING

As in the prior art SAGE techniques, the method described above enables the production of a single nucleotide sequence tag for each first nucleic acid. That is to say, in preferred embodiments, only a single "signature" is obtained from each cDNA. However, it is possible to provide a plurality of nucleotide sequence tags using further steps.

Accordingly, we provide a method of generating a plurality of nucleic acid sequences, each comprising a nucleic acid sequence tag from a nucleic acid. The further nucleotide sequence tags are derived from sequences internal to the nucleotide sequence tag produced initially. The method may be repeated as many times as desired, to produce as many different nucleotide sequence tags as needed from a particular nucleic acid sequence.

The method of generating multiple nucleotide sequence tags makes use of the second product of the cleavage of the linked nucleic acid with the first nucleic acid cleavage enzyme, i.e., a "second nucleic acid". It will be recalled that the second nucleic acid is derived from the first nucleic acid by cleavage to remove an initial nucleotide sequence tag. The second nucleic acid therefore comprises a sequence of the first nucleic acid 3' of the first nucleotide sequence tag.

In order to obtain a further nucleotide sequence tag from the nucleic acid, a linker sequence is attached to the second nucleic acid, and the resulting structure subjected to cleavage by the first nucleic acid cleavage enzyme to release a further nucleotide sequence tag. The linker sequence may comprise the same or similar structure as described previously. In short, the process of providing a further tag merely feeds the second nucleic acid produced by cleavage into the first step of the process already described. It will be appreciated that each round of this process produces an additional nucleotide sequence tag,

and that as many rounds of this process may be carried out as necessary, to produce as many nucleotide sequence tags as necessary. Each further iteration of the method produces a further internal sequence tag, going deeper and deeper into the nucleic acid.

Accordingly, we provide a method of generating a plurality of nucleic acid sequences, comprising the Terminal SAGE method as described in this document, with one or more rounds of further steps (a) to (c) as set out in Claim 1.

An embodiment of technique of SAGE Walking is described below, without limitation, with reference to Figure 5.

5' TERMINAL SAGE

The Terminal SAGE technique has been described generally in the description above. The 5' Terminal SAGE technique described in general and in detail in this section enables the production of nucleotide sequence tags corresponding to the 5' terminal sequence of a relevant nucleic acid. In preferred embodiments, the method enables the production of a nucleotide sequence tag comprising sequence corresponding to a 5' terminal transcribed sequence of a gene, mRNA or cDNA.

We therefore provide a method of providing an indication of an instance of expression of a gene, the method comprising the steps of: (a) providing a complementary deoxyribonucleic acid (cDNA) having a terminus preferably comprising a 5' terminal transcribed sequence of a gene; (b) linking the cDNA to an linker sequence thereby forming a linked nucleic acid, in which the linker sequence comprises a first recognition site for a first nucleic acid cleavage enzyme, preferably a restriction endonuclease, that allows nucleic acid cleavage at a site distant from the first recognition site; and (c) cleaving the linked nucleic acid with the first nucleic acid cleavage enzyme to provide a linked tag, in which the linked tag comprises a nucleotide sequence tag representative of a 5' terminal transcribed sequence of the gene; and (d) detecting the presence or identity of

the linked tag or the nucleotide sequence tag to provide an indication of an instance of gene expression.

Preferably, the cDNA is linked to an linker sequence at its 5' end relative to the coding strand, i.e., 5'-linker sequence- cDNA -3', to form the linked nucleic acid. Preferably, the first recognition site for the first nucleic acid cleavage enzyme, preferably a first restriction endonuclease, allows nucleic acid cleavage at a site 3' of the first recognition site, relative to the coding strand. Preferably, the nucleic acid cleavage site cleaves within the sequence of the cDNA.

In order to provide a first nucleic acid sequence which is a cDNA in preferred embodiments for 5' Terminal SAGE, it is necessary to provide cDNA with 5' terminal transcribed portions. In general, any method which is capable of producing "full length" cDNA is suitable for this purpose. Size fractionation may be used to filter out smaller non-full length species.

For example, full length cDNA may be synthesised by the oligo-capping method described in Maruyama and Sugano, Gene 138, 171-174, 1994, Suzuki et al., Gene 200, 149-156, 1997 using appropriate oligo-dT primers. The oligo capping method replaces the cap structure of a mRNA with an oligoribonucleotide (r-oligo) to label the 5' end of eukaryotic mRNAs. The cap is removed with tobacco acid pyrophosphatase (TAP) and r-oligos ligated to decapped mRNAs with T4 RNA ligase. This reaction is made cap-specific by removing 5'-phosphates of non-capped RNAs with alkaline phosphatase prior to TAP treatment. Unlike the conventional methods that label the 5' end of cDNAs, this method specifically labels the capped end of the mRNAs with a synthetic r-oligo prior to first-strand cDNA synthesis. The 5' end of the mRNA is identified quite simply by reverse transcription-polymerase chain reaction (RT-PCR).

Although the oligo-capping method is preferred, a number of other methods of making or enriching full length cDNA with 5' sequences are known in the art, and any of these may be employed for the purposes described here.

A number of libraries of full length cDNA are known in the art, and these may be used as well in the methods and compositions described here. Full length cDNA libraries from various organisms and tissue types may be obtained commercially from a number of manufacturers, such as ResGen (Invitrogen, see http://www.resgen.com/products/PFL.php3). Furthermore, a database of full length sequences may be found for example, at the Institute of Medical Science, University of Tokyo (http://cdna.ims.u-tokyo.ac.jp/)

In the following detailed description, which should not be taken as limiting, reference is made to Figure 3. Reference is also made to the detailed protocols set out in Examples 1 and 2.

To add the MmeI linkers to the starting point of transcripts, we have to have cDNA fragments that are 5' intact. In order to do so, the established cap-trapper method may be employed for full-length cDNA selection to ensure most of the substance cDNA fragments have the intact 5' end. The cap-trapper method is described in detail in Genomics 37 327-336., Biotechniques 32 (5): 984-985 and Biotechniques 30 (6): 1250-1254.

Other techniques to enable full length cDNA to be obtained are also suitable. For example, full length cDNA may be obtained by using eIF-4E cap-binding protein to identify and bind to the 5' mRNA cap (Mol. Cell. Biol. 15 : 3363-3371). Furthermore, methods based on RNA terminal tagging with RNA ligase may also be used (Methods Enzymol. 65: 65-74).

As we make full-length cDNA, we use NotI-dT15VN primer to initiate the cDNA synthesis so introduce a NotI site at the end of the polyA/T tail. After full-length cDNA are selected, the 3' end will be digested by NotI followed by a ligation of the biotin-NotI linker, so the 3'end biotinylated cDNA are able immobilized on streptavidin coated magnetic beads. Since NotI is an 8-bp cuter, the possibility of internal cDNA digestion is rare.

The immobilized cDNA population is then divided in half for the addition of MmeI-BseRI linker A and MmeI-BseRI linker B. These linkers contain the same sequence for MmeI and BseRI sites, but have different sequences for PCR priming sites. Here MmeI is a IIS type enzyme that cut DNA 20-bp down stream from its recognition site (TCCGACN20/18). MmeI is used for releasing tags, called the "Tag Enzyme" (TE). BseRI is another IIS type enzyme that cut DNA 10-bp apart from its recognition site (GAGGAGN10/8). BseRI will be used for removing the PCR arms and leave a minimal anchor sequences with the tags, called the "Anchor Enzyme" (AE).

MmeI digestion releases linked tags comprising a unique sequence signature corresponding to the 5' portion of the transcribed gene.

After MmeI digestion, the released tags from the two populations may be combined and ligated. Every two tags are joined together face to face (tail to tail) with two base pairs overlapping to form a ditag flanked with linkers at both ends. The ditags are subsequently purified and amplified by PCR with biotinylated primer A and B. Ditags with different linkers will be more efficiently amplified in PCR reaction than the one with same linker sequences.

The next technical trick is how to remove the long arms from the amplified ditags, keep the anchor sequence minimal, and make the tags with cohesive ends and ligatable to each other. We place a BseRI site next to the MmeI site apart by 6 base pairs in the linker sequences. That will allow BseRI cut into the MmeI site (GAGGAGNNNNNNTC CG AC) and create a CG-3' overhang. The probability of having a BseRI site within the 20-bp tag sequences of all transcripts is very rare (P = 0.00488). After purification, the trimmed ditags are ligated to each other resulting concatenated ditags separated each other with an anchor sequence CGAC. Because each 20-bp tag has two base pairs overlapping in a ditag unit and a 4-bp anchor sequence per ditag in concatenated clones that means information of every 21 bp sequence in the concatenated clones contains one tag representing one transcript molecule. To tag one million transcript molecules, we only need to generate 21000 concatenated clones (1000

bp) for sequencing at both ends of the clones (42000 seq. reads) assuming an average 500 bp quality sequences per read.

3' TERMINAL SAGE

The 3' Terminal SAGE technique described in general and in detail in this section enables the production of nucleotide sequence tags corresponding to the 5' terminal sequence of a relevant nucleic acid. In preferred embodiments, the method enables the production of a nucleotide sequence tag comprising sequence corresponding to a 3' transcribed sequence of a gene, mRNA or cDNA.

We therefore provide a method of providing an indication of an instance of expression of a gene, the method comprising the steps of: (a) providing a complementary deoxyribonucleic acid (cDNA) having a terminus comprising a 3' terminal transcribed sequence of a gene; (b) linking the cDNA to an linker sequence thereby forming a linked nucleic acid, in which the linker sequence comprises a first recognition site for a first nucleic acid cleavage enzyme, preferably a restriction endonuclease, that allows nucleic acid cleavage at a site distant from the first recognition site; and (c) cleaving the linked nucleic acid with the first nucleic acid cleavage enzyme to provide a linked tag, in which the linked tag comprises a nucleotide sequence tag representative of a 3' terminal transcribed sequence of the gene; and (d) detecting the presence or identity of the linked tag or the nucleotide sequence tag to provide an indication of an instance of gene expression.

In order to provide a first nucleic acid sequence which is a cDNA in preferred embodiments for 3' Terminal SAGE, it is necessary to provide cDNA comprising a terminus with 3' terminal transcribed portions. Preferably, at least some of the polyA/T tail at the 3' terminus of the cDNA can be removed. However, it will be appreciated that it is not strictly necessary to remove the whole of the poly A/T tail from the cDNA, provided that at least some sequence from the transcribed portion is captured in the nucleotide sequence tag. Any polyT sequence which may be in the nucleotide sequence tag may be

ignored for analysis. Provided the nucleotide sequence tag is long enough, it will capture some transcribed sequence from the cDNA sufficient to identify the relevant gene.

Where it is desired to remove at least some of the polyA/T portion, any method which is capable of producing a cDNA without (or substantially without) a poly dT tail may suitably be employed for this purpose. Such techniques are discussed in detail in Shibata et al (2001), Removal of polyA tails from full-length cDNA libraries for high efficiency sequencing, Biotechniques, 31(5), 1042-1049, 2001. As noted above, it is not necessary for the entirety of the polyA/T tail to be removed, so long as the nucleotide sequence tag is long enough to contain useful sequence information. Thus, for example, the Shibata technique may produce cDNA with an average polyA tail length of 4 or fewer, but this does not matter provided the nucleotide sequence tag is long enough to capture 3' transcribed sequence of the gene.

In a particularly preferred embodiment, however, all of the polyA/T tail removed from the cDNA. Thus, in such embodiments, a cDNA having a terminus comprising a 3' terminal transcribed sequence of a gene may in particular be provided by the following steps: (i) deriving a cDNA from an mRNA by reverse transcription with an primer comprising a sequence 5'-NV(T)^{13}CCGGCCGG-3', in which N = A, C, G or T and V = A, C or G; (ii) producing a double stranded cDNA therefrom and digesting the double stranded cDNA with FseI to produce a cleaved cDNA comprising $3'\frac{GGCCGG}{CC}$ overhang; (iii) linking the cleaved cDNA to a linker comprising $\frac{pTCGGA}{GGCCAGCCT}$; and (iv) cleaving the resulting molecule with MmeI to produce a cDNA lacking a polyA/T tail.

In the following detailed description of 3' Terminal SAGE, reference is made to Figure 4. This detailed description should not be taken as limiting. Reference is also made to the detailed protocols set out in Examples 3 and 4.

To add the MmeI linkers to the 3'end of cDNA fragments, we have to first remove the polyA/T tails precisely at their starting point. In order to do so, we use the biotin-FseI-dT13VN primer to initiate cDNA synthesis. The two anchor nucleotides (V = A, C, G; and N = A, C, G, T) will guide the primer to anneal at the first 13 As in the polyA tail of mRNA. Thus all cDNA fragments will have a 13-bp polyA/T tail followed by an FseI site and a biotin at the end. Double strand cDNA will be first immobilized on magnetic beads, and then digested by MmeI enzyme to clean up any possible internal MmeI sites in cDNA. After polishing the ends, cDNA fragments are released from the beads by FseI digestion and create a cohesive 3'end. FseI is an 8-bp cuter like NotI, so the internal FseI sites in cDNA are rare. A 1/2FseI-MmeI linker is first ligated to the cohesive 3' end, and then a biotin linker is added to the 5'end of cDNA. After adding the biotin-linker, the cDNA fragments are once again immobilized on magnetic beads, but this time are by the 5'ends, and with an external MmeI site located 18 base pairs (6 residuals from the FseI site after digestion and ligation and a 13-polyA/T tail) apart from the 3' extremity. A MmeI digestion will precisely remove the polyA/T tail and create a two base pair overhang 3'end.

The polyA/T-less cDNA is then divided by half into two populations, one half is to add the MmeI-BseRI linker A, and the other half is to add the MmeI-BseRI linker B. At this point, an external MmeI site is precisely introduced at the 3' termini of transcripts. All steps in the rest of the procedure including releasing SAGE tags, ligating ditags, amplifying, and concatenating are all exactly the same as in the 5' Terminal SAGE procedure.

NUCLEIC ACID CLEAVAGE ENZYMES

The techniques described here generally utilise nucleic acid cleaving or nicking enzymes, such as DNA cleavage enzymes, which recognise and cleave specific nucleic acid sequences. Such enzymes are preferably restriction enzymes (restriction endonucleases). As used in this document, the terms "restriction endonucleases" and "restriction enzymes" refer to bacterial enzymes which bind to a specific double-stranded

DNA sequence termed a recognition site or recognition nucleotide sequence, and cut double-stranded DNA at or near the specific recognition site.

Preferably, the nucleic acid cleavage enzymes cleave the nucleic acid at a position outside of the recognition sequence. The cleavage position is therefore "offset" from the recognition position, and the enzymes may for convenience be referred to as "offset cleavage enzymes". In particular, use of "Type IIS restriction enzymes" (described in further detail below) are preferred.

As described elsewhere, a linker comprising a recognition sequence for such an offset cleavage enzyme is joined to a cDNA molecule as a first step in providing a tag.

Although the use of Type IIS restriction enzymes is preferred in the methods and compositions described here, it will be appreciated that derivitised or modified or engineered versions of such enzymes may equally be used, provided they have the properties of sequence recognition and nucleic acid cleavage (or nicking) at sites distant from the cognate recognition sequences. Thus, the methods and compositions described here employ enzymes which cleave nucleic acids at offset positions. All that is required for the 5' and 3' Terminal SAGE techniques is that the linker comprise a recognition site for an enzyme that allows nucleic cleavage at a site in the nucleic acid distant from the recognition site for the enzyme. Engineering methods, for example those described in Kim YG, Shi Y, Berg JM, Chandrasegaran S. (1997) Gene 1997 Dec 5;203(1):43-9, Smith J, Berg JM, Chandrasegaran S. (1999) Nucleic Acids Res 1999 Jan 15;27(2):674-81 and Kim YG, Smith J, Durgesha M, Chandrasegaran S. (1998) Biol Chem 1998 Apr-May;379(4-5):489-95, may be used to enhance the cleavage efficiency, modify the recognition sequence, or to increase, decrease, or otherwise modify the offset between the recognition site and the cleavage site. Zinc finger binding domains may be used for engineering chimeric restriction enzymes, as described in Kim,Y.-G., Cha,J. and Chandrasegaran,S. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 1156-1160, Huang,B., Schaeffer,C.J., Li,Q. and Tsai,M.-D. (1996) *J. Prot. Chem.*, **15**, 481-489 and Kim,Y.-G., Shi,Y., Berg,J.M. and Chandrasegaran,S. (1997) *Gene*, **203**, 43-49.

In particular, we envisage the use of chimeric enzymes comprising sequence recognition domains and nucleic acid cleavage domains. The recognition domains and the nucleic acid cleavage domains may be linked and separated by a linker of variable length, to allow cleavage to occur remotely from the recognition sequence.

The recognition domains are capable of recognising and binding to specific nucleic acid sequences, and may comprise, for example, zinc finger domains designed according to the rules established in WO 98/53057, WO 98/53060, WO 98/53058, WO 98/53059. Methods of engineering, as well as rational and rule based design of zinc fingers, are also generally known in the art. The nucleic acid cleavage domains are capable of nicking or cleaving single or double stranded nucleic acids, and may be derived from known or existing cleavage domains found, for example, in restriction nucleases, DNAses or RNAses. The recognition domain and the cleavage domain in the chimeric enzyme may be separated by a linker to allow cleavage remote from the recognition site. Such linkers may comprise structured or flexible linkers as known in the art, and described for example in WO 00/44568, which is hereby incorporated by reference. The length of the linkers may be modified according to the offet desired, i.e., a longer linker will be predicated where a longer tag is required.

Indeed, it may be possible to employ completely artificial enzymes not derived from restriction enzymes which have this property. Such derivitised, modified or engineered restriction enzymes, as well as artificial enzymes, having these properties, should be understood to be encompassed for use in the methods and compositions described here.

TYPE IIS RESTRICTION ENZYMES

Specifically, the methods and compositions described here employ restriction enzymes which do not cleave within their recognition sites, but instead cleave at sequences of defined distance from their recognition sequence. Such restriction enzymes are typically referred to as "Type IIS" restriction enzymes.

Type IIS restriction endonucleases cleave at a defined distance up to 20 bp away from their asymmetric recognition sites (Szybalski, W., Gene, 40:169, 1985). Examples of type IIS restriction endonucleases include BsmFI and FokI. Other similar enzymes will be known to those of skill in the art (see, Current Protocols in Molecular Biology, supra).

Type IIS restrictions suitable for use in the methods and compositions described here are shown in **Table 2** below. Sequence representations in the table use the standard abbreviations (*Eur. J. Biochem.* 150: 1-5, 1985) to represent ambiguity. Recognition sequences are written from 5' to 3', only one strand being given. If the point of cleavage has been determined, the precise site is marked with ^. The character _ is sometimes used to mark the cut site on the complementary strand. Accordingly, it can be seen that the choice of the specific Type IIS restriction endonuclease will determine the offset, and hence the length of the specific sequence tag generated by the methods and compositions described here. It will be apparent that that the longer the offset and hence the tag, the more specific the sequence tag generated will be.

In preferred embodiments, the restriction endonuclease used is MmeI, which has a recognition sequence of TCCRACNNNNNNNNNNNNNNNNN_NN^, and cuts at a position 20/18 bases from its recognition sequence, resulting a 3' protruding overhang of 2 nucleotides (2nt). Accordingly, we envisage the use of a linker comprising a TCCRAC sequence, more preferably a TCCGAC sequence. Cleavage of such a linker joined at its 3' end to a nucleic acid sequence, for example a full length cDNA, will produce a tag comprising 20 bases on one strand, and 18 bases on the other. The strand comprising 20 bases may be referred to as the "top" strand, while the strand comprising 18 bases may be referred to as the "bottom" strand.

Type IIS enzymes are of special interest in molecular biology. These enzymes recognize asymmetric base sequences and cleave DNA at a specified position up to 20 base pairs outside of the recognition site. Blunting the ends of their digestion products, followed by ligation, does not destroy their recognition sites. This property is useful in several applications, including the generation of deletions of increasing length (Hasan, N.,

et al., A novel multistep method for generating precise unidirectional deletions using BspMI, a class-IIS restriction enzyme, Gene, 50, 55-62, 1986) and mapping the sequence specificity of DNA modification (Posfai, G. and Szybalski, W., A simple method for locating methylated bases in DNA, as applied to detect asymmetric methylation by M.FokIA, Gene, 69, 147-151, 1988) and PCR product cloning. Due to the asymmetric nature of their recognition sequences, type IIS R-M systems comprise two methylases, one for each strand, sometimes each methylating a different base (Bitinaite, J., et al., Alw26I, Eco31I and Esp3I - type IIs methyltransferases modifying cytosine and adenine in complementary strands of the target DNA, Nucleic Acids Res., 20, 4981-4985, 1992).

Exemplary Type IIS restriction enzymes are described in the restriction enzyme database REBASE (Roberts RJ, Macelis D. (2001) REBASE--restriction enzymes and methylases. Nucleic Acids Res. 2001 Jan 1;29(1):268-9 and Roberts,R.J. and Macelis,D. (1999) *Nucleic Acids Res.*, **27**, 312–313), which may be accessed at rebase.neb.com.

REBASE is a comprehensive database of information about restriction enzymes and related proteins. It contains published and unpublished references, recognition and cleavage sites, isoschizomers, commercial availability, methylation sensitivity, crystal and sequence data. DNA methyltransferases, homing endonucleases, nicking enzymes, specificity subunits and control proteins are also included. Most recently, putative DNA methyltransferases and restriction enzymes, as predicted from analysis of genomic sequences, are also listed. The data is distributed via Email, ftp (ftp.neb.com), and the Web (http://rebase.neb.com).

Enzymes	**Recognition Sequence**	**Isoschizomers**
AarI	CACCTGCNNNN^NNNN_	-
AceIII	CAGCTCNNNNNNN^NNNN_	-
AloI	_NNNNN^NNNNNNNGAACNNNNNNTCC NNNNNNN_NNNNN^	-
BaeI	_NNNNN^NNNNNNNNNNACNNNNGTAY CNNNNNNN_NNNNN^	-
Bbr7I	GAAGACNNNNNNN^NNNN_	-
BbvI	GCAGCNNNNNNNN^NNNN_	AlwXI BseKI BseXI Bsp423I Bst12I Bst71I BstV1I
BbvII	GAAGACNN^NNNN_	BbsI Bbv16II BpiI BpuAI Bsc91I BspBS31I BspIS4I BspTS514I

		BstBS32I BstTS5I BstV2I
BccI	CCATCNNNN^N_	-
Bce83I	CTTGAGNNNNNNNNNNNNN_NN^	BpuEI
BceAI	ACGGCNNNNNNNNNNNN^NN_	-
BcefI	ACGGCNNNNNNNNNNNN^N_	-
BcgI	_NN^NNNNNNNNNNCGANNNNNNTGCN NNNNNNNNN_NN^	-
BciVI	GTATCCNNNNN_N^	BfuI
BfiI	ACTGGGNNNN_N^	BmrI
BinI	GGATCNNNN^N_	AclWI AlwI BspPI BstH9I Bst31TI EacI
BplI	_NNNNN^NNNNNNNNGAGNNNNNCTCN NNNNNNN_NNNNN^	-
BsaXI	_NNN^NNNNNNNNNACNNNNNCTCCNN NNNNN_NNN^	-
BscAI	GCATCNNNN^NN_	Bst19I
BseMII	CTCAGNNNNNNNN_NN^	-
BseRI	GAGGAGNNNNNNNN_NN^	-
BsgI	GTGCAGNNNNNNNNNNNNNN_NN^	-
BsmI	GAATG_CN^	Asp26HI Asp27HI Asp35HI Asp36HI Asp40HI Asp50HI BsaMI BscCI Mva1269I PctI
BsmAI	GTCTCN^NNNN_	Alw26I BsoMAI
BsmFI	GGGACNNNNNNNNNN^NNNN_	BspLU11III BstOZ616I
Bsp24I	_NNNNN^NNNNNNNNGACNNNNNNTGG NNNNNNN_NNNNN^	-
BspCNI	CTCAGNNNNNNN_NN^	-
BspMI	ACCTGCNNNN^NNNN_	Acc36I BfuAI BveI
BsrI	ACTG_GN^	Bse1I BseNI BsrSI Bst11I Tsp1I
BsrDI	GCAATG_NN^	Bse3DI BseMI
BstF5I	GGATG_NN^	BseGI
BtsI	GCAGTG_NN^	-
CjeI	_NNNNNN^NNNNNNNNCCANNNNNNGT NNNNNNNNN_NNNNNN^	-
CjePI	_NNNNNN^NNNNNNNCCANNNNNNNTC NNNNNNNN_NNNNNN^	-
EciI	GGCGGANNNNNNNNN_NN^	-
Eco31I	GGTCTCN^NNNN_	Bli736I BsaI Bso31I BspTNI EcoA4I EcoO44I
Eco57I	CTGAAGNNNNNNNNNNNNNN_NN^	AcuI BspKT5I
Eco57MI	CTGRAGNNNNNNNNNNNNNN_NN^	-
Esp3I	CGTCTCN^NNNN_	BsmBI BstGZ53I
FalI	_NNNNN^NNNNNNNNAAGNNNNNCTTN NNNNNNN_NNNNN^	-
FauI	CCCGCNNNN^NN_	Bme585I BstFZ438I SmuI
FokI	GGATGNNNNNNNNN^NNNN_	BstPZ418I

GsuI	CTGGAGNNNNNNNNNNNNN_NN^	BpmI
HaeIV	_NNNNNN^NNNNNNNGAYNNNNNRTCN NNNNNNNN_NNNNN^	-
HgaI	GACGCNNNNN^NNNNN_	-
Hin4I	_NNNNN^NNNNNNNNGAYNNNNNVTCN NNNNNNN_NNNNN^	-
HphI	GGTGANNNNNNN_N^	AsuHPI
HpyAV	CCTTCNNNNN_N^	-
Ksp632I	CTCTTCN^NNN_	Bco5I Bco116I BcoKI BseZI Bst6I Bsu6I Eam1104I EarI
MboII	GAAGANNNNNNN_N^	NcuI
MlyI	GAGTCNNNNN^	SchI
MmeI	TCCRACNNNNNNNNNNNNNNNNN_NN ^	-
MnlI	CCTCNNNNNN_N^	-
PleI	GAGTCNNNN^N_	PpsI
PpiI	_NNNNN^NNNNNNNGAACNNNNNCTCN NNNNNNN_NNNNN^	-
PsrI	_NNNNN^NNNNNNNGAACNNNNNNTAC NNNNNNN_NNNNN^	-
RleAI	CCCACANNNNNNNNN_NNN^	-
SapI	GCTCTTCN^NNN_	VpaK32I
SfaNI	GCATCNNNNN^NNNN_	BspST5I LweI PhaI
SspD5I	GGTGANNNNNNNN^	-
Sth132I	CCCGNNNN^NNNN_	-
StsI	GGATGNNNNNNNNNN^NNNN_	-
TaqII	GACCGANNNNNNNNN_NN^, CACCCANNNNNNNNN_NN^	-
TspDTI	ATGAANNNNNNNNN_NN^	-
TspGWI	ACGGANNNNNNNNN_NN^	-
TspRI	_NNCASTGNN^	-
Tth111II	CAARCANNNNNNNNN_NN^	-

Table 2: Type II restriction enzymes cleaving outside their recognition sequences (Type IIS restriction enzymes).

The invention is described further, for the purpose of illustration only, in the following examples.

USES

Our methods have specific advantages over prior art techniques as they enable the generation of terminal transcribed sequences from cDNAs. Knowledge of such terminal transcribed sequences enables the discrimination between genes which are similar over their internal regions. In addition, the 5' and 3' terminal transcribed sequences may be used as "launch pads" for the identification of upstream or downstream (as the case may be) non-coding, untranscribed, untranslated or control regions. For example, the promoters or other control regions (such as enhancers) of a particular gene may be obtained easily once its 5' terminal transcribed sequence is known, by means of "walking" techniques, or by searching sequence databases.

Promoters may be identified by a number of methods. For example, it is known that eukaryotic gene promoters are usually locates upstream of the transcription starting site and possess a consensus sequence such as CAAT box or TATA box. The 5' Terminal SAGE method described here essentially identifies the 5' transcription start site; sequences upstream of the transcription start site may be obtained by chromosomal walking, or preferably by examining a genomic database using standard bioinformatics tools to search for promoter consensus regions based on the presence of consensus sequences. The identity of putative promoters may be confirmed by making constructs with suitable reporters and examining expression of reporter, by means well known in the art.

Furthermore, once the 5' and the 3' terminal transcribed sequences are determined, it is a simple matter to obtain a full length clone (a cDNA clone or a genomic clone) from an mRNA or cDNA or genomic library. Appropriate primers corresponding to these sequences may be synthesised and used as PCR primers in a polymerase chain reaction to amplify the full length clone.

Specific uses of the terminal tags produced according to the methods and compositions described here are described. I particular, we describe specific uses of the

methods described here for transcriptome characterisation, genome annotation, novel gene discoveries, and promoter identification.

It will be apparent that the terminal transcribed sequences derived from cDNAs may be used for many other purposes, and that the skilled person will be aware of such uses. Accordingly, the following uses should not be taken to be limiting.

Transcriptome Analysis

Due to the current limitations, the original SAGE technique has not been applied broadly and deeply comparing to other competing technologies such as microarrays. Many SAGE experiments done in the prior art are simply not deep enough to be statistically significant, particularly to rare transcripts. The Terminal SAGE techniques, on the other hand, may be used to collect a large number of tag pairs of 5'and 3' Terminal SAGE for various genomes such as human and zebrafish.

Transcriptome Analysis of Zebrafish and Other Organisms

The 5'and 3' Terminal SAGE methods may be used in conjunction with large scale sequencing efforts, for example, zebrafish full-length cDNA cloning and sequencing from a mixed adult fish library and an embryonic library, to obtain a comprehensive picture of zebrafish transcriptomes at embryonic and adult stages. The combination of these two approaches has the potential to provide a comprehensive disclosure of zebrafish transcriptomes and large number of full-length transcript clones and sequences.

The experimental data will be invaluable for thorough annotation of the upcoming zebrafish genome sequences.

Using the Terminal SAGE techniques, and obtaining one million 5'and 3'tag-pairs from zebrafish, it will only be necessary to process fewer than 40,000 Terminal SAGE clones (i.e., clones of concatamers) with fewer than 70,000 sequence reads on the assumption that each SAGE concatamer clone is about 2kp; each clone will be sequenced from both ends; and each sequence read will provide 800bp in average.

This approach may be readily applied to other organisms, such as mouse, humans, etc.

Human Stem Cell Transcriptomes

More importantly, the new Terminal SAGE methods may be applied to stem cell research projects.

A deep and comprehensive characterization of stem cell transcriptomes has never been done, though some limited attempts have been reported using the original SAGE techniques. However, the data for novel genes involved in stem cell constitution is not convincing.

Use of the Terminal SAGE methods as described here to collect large number of tags will allow for the first time a complete characterization of stem cell transcriptomes, and will make huge contribution to human genome annotation. Comparisons may be made between embryonic stem cells (ES cells) and a committed type of stem cell for deep 5'and 3' sequence analyses.

We believe that it is possible to obtain a near complete list of genes expressed and their activities in stem cells by generating one million tag-pairs of 5'and 3' Terminal SAGE for each of the ES cell and the other cell. Compared to microarray analysis, this new approach will have a much better chance to convincingly identify key genes that are differentially expressed at low levels between these cells.

Any new transcript identified by this method may be easily isolated in full-length by RT-PCR using the 5'and 3' tag information.

Large Scale Promoter Identification

Applying the 5' Terminal SAGE techniques to cells such as zebrafish and human stem cell will allow the identification of the vast majority of genes at their transcript initiation sites.

The experimental data will enable the identification and extraction of promoter sequences for all of the 5' Terminal SAGE identified genes, and establish a large promoter database for human and zebrafish. These large set of expression-based promoter databases will provide a solid knowledge base for mining promoter functions, training computational programs for promoter predictions, and extending our understanding of comprehensive gene regulatory networking mechanisms.

EXAMPLES

Example 1. 5' Terminal SAGE Operational Protocol

1.1 Full length cDNA synthesis

NotI-oligo dT primers (14 μg) and 20 μg of polyA$^+$ RNA are ethanol precipitated and resuspended into 10 μl of ddH_2O. Heat up at 65^{o}C for 10 min and left at 42^{o}C for 1 min.

In a separate tube, mix the following components: 5X first strand synthesis buffer 30 μl, 0.1M DTT 11 μl, 10mM dNTP 9 μl, saturated trehalose 15 μl, 4.9M sorbitol 50 μl and Superscript(™) II reverse transcriptase 15 μl. Mix with RNA from above and incubate at 40^{o}C for 4 min, 50^{o}C for 2 min and 56^{o}C for 60 min. 2 μl of proteinase K (20 mg/ml) is added and the reaction is incubated at 45^{o}C for 15 min followed by phenol/chloroform extraction and ethanol precipitation.

The RNA/cDNA heter-duplex is resuspended into 44.5 μl of ddH_2O. 3 μl of 1.1 M NaOAc pH 4.5 and 2.5 μl of 100mM $NaIO_4$ are added to oxidize the diol structures of the

mRNA. The 50 μl reaction is incubated on ice in the dark for 45 min followed by adding 0.5 μl of 10% SDS, 11 μl of 5 M NaCl and 61 μl of isopropanol.

The precipitated RNA/DNA is biotinylated in 50 μl of ddH_2O by added the 5μL 1M NaOAc (pH6.1), 5μL 10% (w/v) SDS and 150μL 10mM long-arm biotin hydrazide. Leave at RT/dark/O/N. Add: 5μL 5M NaCl, 75μL 1M RNase-free NaOAc (pH6.1), 750μL 100% EtOH or 200 μL of 100% Isopropanol . -80°C/30'+*. Centrifuge at 14krpm/4°C/30'. Wash pellet w/ 70% (v/v) EtOH/30%, DEPC-treated ddH_20, centrifuge at 14krpm/4°C/10'. Remove XS liquid, air-dry pellet. Resuspend pellet in 70μL DEPC-ddH_20, then add: 10μL 10x RNaseI buffer, **25** U RNaseI / ug of starting mRNA. 37°C/30'. Add 2.5μL of 40mg/mL Yeast tRNA and ½ volume of 5M NaCl to stop the reaction.

While the biotinylated RNA-DNA heteroduplex is ppting, prepare the Streptavidin-labelled Dynabeads(™):

Pipet 500μL of M-280 Streptavidin beads into an RNase-free Eppendorf tube. Place on magnet, wait @ least 30", and remove sup. Resuspend beads in 500μL 1x binding buffer (2M NaCl, 50 mM EDTA, pH 8.0). Place on magnet, wait @ least 30", and remove sup. Repeat the 1x binding buffer wash for 3 times. Resuspend beads in 500μL 1x binding buffer w/ 100μg of Yeast tRNA. 30'+/4°C/mix occasionally. Place on magnet, wait @ least 30", and remove sup. Wash with 1x binding buffer for 3 times. Resuspend in 100μL 1x binding buffer. Mix beads and DNA-RNA heteroduplex (V_t=200μL), binding at 2 M NaCl. 30'/RT/rotating. Place on magnet, wait @ least 30", and remove sup. Wash 2x w/400μL of 1x binding buffer. Place on magnet, wait @ least 30", and remove sup. Wash w/400μL of 0.4%(w/v) SDS plus 50μg/mL Yeast tRNA. Place on magnet, wait @ least 30", and remove sup. Wash w/400μL of 1x wash buffer (10mM Tris-HCl pH7.5, 0.2mM EDTA, 10mM NaCl & 20%(v/v) Glycerol, 40 μg/mL Yeast tRNA). Place on magnet, wait @ least 30", and remove sup. Wash w/400μL of 50μg/mL Yeast tRNA. Place on magnet, wait @ least 30", and remove sup.

Release the first strand cDNA by alkali hydrolysis of RNA. Add: 50μL 50mM NaOH and 5 mM EDTA (pH8.0) 10'/RT/rotating. Transfer the sup. to another tube containing 50μL 1M Tris-Cl (pH7.5). Repeat the lysis procedure for 2 more times. The final volume is 300 μL.

1.2 Single Strand Linker A & B ligation

Pecipitate the single strand first cDNA with glycogen and divide the cDNA into 2 tubes, A and B. Add the following reagents to the each corresponding tube on ice.

Contents	Tube A	Tube B
cDNA	5μL	5μL
Linker A (N5)	1.6μg	−
Linker A (N6)	0.4μg	−
Linker B (N5)	1.6μg	
Linker B (N6)	−	0.4μg
Soln II (Takara kit)	10μl	10μl
Soln I (Takara ligation kit)	20μl	20μl
Total volume	40μl	

16°C/o/n. and then 70°C 10' to inactivate the ligase. Increase the volume to 200μl and phenol/chloroform extraction. Sephacryl(™) -300 to remove the excess linkers and precipitate the cDNA with EtoH and glycogen.

Resuspend the pellet with 60μl of ddH_20, add 8μl 10XExtaq buffer, 8μl 2.5mM dNTP 4μl ExTaq enzymes. 65°C 5' → 68°C 30' → 72°C 10'.

Precipitate the cDNA with EtoH and glycogen. Resuspended with ddH_20.

1.3 Binding cDNA to Magnetic Beads

The double strand cDNAs are digested with NotI at 37°C for 1 hour in a volume of 50μl. The enzyme is then inactivated with proteinase K and extracted with phenol/chloroform and then ethanol precipitated.

The samples are mixed with 200 ng of NotI linker adaptor. Ligation is carried out in total volume of 10µl at 16°C for overnight followed by 70°C for 10' to inactivate enzyme. The excess adaptor is removed by Sephacryl -300.

Dynabeads(™) M280 streptoavidin beads are used to bind the Biotinylated cDNA fragments according to the manufacturer's recommendation.

1.4 Digesting the cDNA with Tagging Enzyme MmeI

Place the two tubes (A and B) on magnetic stand, remove the supernatant (wash buffer from Dynabeads(™)).

Add the following reagents to each tube.

ddH_20	86µl
MmeI	2µl
10X buffer	10µl
(1 :10) SAM	2µl
Total volume	100µl. Incubate at 37°C for 1 hour with mixing.

Remove the supernatant which contains the tags. Phenol/chloroform and precipitate.

Optionally, a Klenow fill-in reaction may be carried out to produce blunt ends. If carried out, the pellet is resuspended and the following carried out.

Add the following components to each tube containing 10µl of sample from the above.

10X Klenow buffer	5µl
100XBSA	1µl
dNTP (10mM each)	2.5µl
ddH_20	30.5µl
Klenow polymerase	1µl
Total volume	50µl

Mix well and incubate for 30' at 37°C.

Pool the ditags and phenol/chloroform followed by EtOH precipitation.

The Klenow fill in step is optional, and may result in less sequence information being obtained (see elsewhere in this document for a detailed discussion).

Example 2. Ligating tags to create ditags

Resuspend the precipitated pellet from the above in 1.5µl and add 1.5µl of the following reagents to form the ditags.

3mM Tris-HCl, pH 7.5	1.25µl
10X Ligation buffer	0.75µl
ddH_20	0.75µl
T4 ligase	1µl

Incubate at 16°C for overnight.

2.1 PCR Amplifying and Gel Purifying the 138 bp Ditags

Set up 200 to 300 reactions as following (in 96 PCR plates).

10XBV buffer	5µl
DMSO	3µl
dNTPmix	7.5µl
PCR primer A (175ng/µl)	2µl
PCR primer B (175ng/µl)	2µl
ddH_20	29µl
Platinum Taq enzyme	0.5µl
Template (1: 230 dilution)	1µl
Total volume	50µl

Cycling condition: 95°C 2' 1 cycle. 95°C 30", 55°C 1', 70°C 1'. 27 cycles. 70°C 5'.

Run a diagnostic 4% agarose gel to analyze the PCR product and purify the 138 bp ditags by 12% polyacrylamide gel electrophoresis.

2.2 Digesting the Ditags with Anchoring Enzyme BseRI and Purifying the 46 Bp Ditag

Digest the 138bp ditags with BseRI to yield 46bp ditag.

138bp ditags	42μl
10X NEB buffer II	15μl
BseRI	12μl
ddH_20	81μl
Total volume	150μl

And incubate at 37°C for 2 to 3 hours. Phenol/chloroform and EtoH precipitate. Run a diagnostic 4% agarose gel to check the efficiency of the digestion and purify the 46 bp ditags by 12% polyacrylamide gel electrophoresis.

2.3 Ligating the 46 bp Ditag to Form Concatamers

Set up the ligation reaction using the gel purified ditag in a total volume of 10μl and incubate 16 C for 2 to 3 hours. Gel purify the concatamers from 8% polyacrylamide, select the DNA size range around 2kbp.

2.4 Cloning Concatamers into Vector

In a separate reaction, digest the plasmid vector with HhaI and ligate the purified concatamers into the linearized vector following standard molecular biology protocol. Transform the ligation into high efficiency competent cells.

The concatamers of 5' Terminal SAGE tags may then be sequenced using conventional means.

Example 3. 3' Terminal SAGE Operational Protocol

3.1 cDNA synthesis

GsuI-dT16 primers (1μg) and 5μg of polyA$^+$ RNA are mixed in final volume of 7μl. Heat up at 70^{o}C for 10 min and leave on ice.

Synthesize the cDNA according to the manufacturer's recommendation (Invitrogen Superscript$^{(TM)}$ cDNA synthesis system) but using the Biotinylated SalI adaptor. The excess adaptor is removed by Sephacryl$^{(TM)}$ -300.

3.2 Binding cDNA to Magnetic Beads

Dynabeads$^{(TM)}$ M280 streptoavidin beads are used to bind the Biotinylated cDNA fragments according to the manufacturer's recommendation.

3.3 Digesting the cDNA with GsuI to Remove PolyA Tail

The double strand cDNA on beads are digested with GsuI at 30 ^{o}C for 1 hour. After the digestion is completed, inactivating the enzyme by washing the tube twice with buffer containing 1%SDS. Four more wash with wash buffer (5mM Tris-HCl pH 7.5, 0.5mM EDTA. 1mM NaCl and 200μg/μl BSA). The last wash use 1X ligation buffer.

3.4 Ligating MmeI-BseRI Adaptors to cDNA

Place the tube on magnetic stand to remove the supernatant (the ligation buffer). Add the following reagents to the beads on ice.

	Tube A	Tube B
cDNA beads	beads	beads
MmeI-BseRI adaptor A (20ng/μl)	1.5μl	−
MmeI-BseRI adaptor B (20ng/μl)	−1.5μl	
TE	14μl	14μl
10X ligation buffer	2μl	2μl

Heat the tubes for 2' at 50 °C and cool to room temp. for 15' and chill the samples on ice. Add 2.5µl T4 DNA ligase and mix. Incubate at 16 °C for 2 hours. Mix occasionally. Wash the tubes with wash buffer to remove the excess adaptors.

3.5 Cleaving with Tagging Enzyme MmeI

Place the two tubes (A and B) on magnetic stand, remove the supernatant (wash buffer). Add the following reagents to each tube.

ddH_20	86µl
MmeI	2µl
10X buffer	10µl
(1 :10) SAM	2µl
Total volume	100µl.

Incubate at 37°C for 1 hour with mixing. Remove the supernatant which contains the tags. Phenol/chloroform and precipitate.

Optionally, a Klenow fill-in reaction may be carried out to produce blunt ends. If carried out, the pellet is resuspended and the following carried out.

Add the following components to each tube containing 10µl of sample from the above section.

10X Klenow buffer	5µl
100XBSA	1µl
dNTP (10mM each)	2.5µl
ddH_20	30.5µl
Klenow polymerase	1µl
Total volume	50µl

Mix well and incubate for 30' at 37°C. Pool the ditags and phenol/chloroform followed by EtOH precipitation.

The Klenow fill in step is optional, and may result in less sequence information being obtained (see elsewhere in this document for a detailed discussion).

Example 4. Ligating Tags to Create Ditags

Resuspend the pellet from the above in 1.5μl water and add 1.5μl of the following reagents to form the ditags.

3mM Tris-HCl, pH 7.5	1.25μl
10X Ligation buffer	0.75μl
ddH_20	0.75μl
T4 ligase	1μl

Incubate at 16°C for overnight.

4.1 PCR Amplifying and Gel Purifying the 138 Bp Ditags

Set up 200 to 300 reactions as following (in 96 PCR plates).

10XBV buffer	5μl
DMSO	3μl
DNTPmix	7.5μl
PCR primer A (175ng/μl)	2μl
PCR primer B (175ng/μl)	2μl
ddH_20	29μl
Platinum Taq enzyme	0.5μl
Template (1: 230 dilution)	1μl
Total volume	50μl

Cycling condition: 95°C 2' 1 cycle. 95°C 30", 55°C 1', 70°C 1'. 27 cycles. 70°C 5'.

Run a diagnostic 4% agarose gel to analyze the PCR product and purify the 138 bp ditags by 12% polyacrylamide gel electrophoresis.

4.2 Digesting the Ditags with Anchoring Enzyme BseRIi and Purifying the 46 Bp Ditag

Digest the 138bp ditags with BseRI to yield 46 bp ditag.

138bp ditags	42μl
10X NEB buffer II	15μl
BseRI	12μl

Total volume	150µl

And incubate at 37°C for 2 to 3 hours. Phenol/chloroform and EtoH precipitate. Run a diagnostic 4% agarose gel to check the efficiency of the digestion and purify the 46 bp ditags by 12% polyacrylamide gel electrophoresis.

4.3 Ligating the 46 Bp Ditag to Form Concatamers

Set up the ligation reaction using the gel purified ditag in a total volume of 10µl and incubate 16 C for 2 to 3 hours. Gel purify the concatamers from 8% polyacrylamide, select the DNA size range from 1 to 1.5kbp, or .

4.4 Cloning Concatamers Into Vector

In a separate reaction, digest the plasmid vector with HhaI and ligate the purified concatamers into the linearized vector following standard molecular biology protocol. Transform the ligation into high efficiency competent cells.

Example 5. SAGE Walking Operational Protocol

Even 20-bp tags may still seem to be not long enough to be specific for mapping on chromosomes. Sequence variations in tags caused by sequencing errors or point mutations can make the mapping to genome sequence ambiguous. Longer tags can certainly add more confidence to include certain level of mismatches. However, the longest cleavage distance of IIS enzyme is 20-bp by MmeI. To overcome this constraint, we can actually walk along the transcript sequence by reintroducing the MmeI site to the same cDNA that already released the first tag, and perform another round of SAGE tag extraction, amplification, and concatenation. The second SAGE will have 2-bp overlap with the first tags on the same transcripts. These tiled two tags of one transcript add up total 38-bp tag information that will be sufficient to tolerate variations due to SNP on either transcript sequence or on the chromosome sequence in genome sequence database, or simply sequence errors. We can even add a third or further rounds of SAGE walk if necessary.

5' SAGE Walk Operational Protocol

Follow the 5' terminal tag protocol to step: **Digesting the cDNA with tagging enzyme MmeI.** After the digestion is completed, inactivate the enzyme by washing the tube twice with buffer containing 1%SDS. Four more washes with wash buffer (5mM Tris-HCl pH 7.5, 0.5mM EDTA. 1mM NaCl and 200μg/μl BSA). For the last wash use 1X ligation buffer.

Ligating MmeI-BseRI Adaptors to cDNA.

Place the tube on magnetic stand to remove the supernatant (the ligation buffer). Add the following reagents to the beads on ice.

Contents	Tube A	Tube B
cDNA beads	beads	beads
MmeI-BseRI adaptor A (20ng/μl)	1.5μl	
MmeI-BseRI adaptor B (20ng/μl)		1.5μl
TE	14μl	14μl
10X ligation buffer	2μl	2μl

Heat the tubes for 2' at 50 °C and cool to room temp. for 15' and chill the samples on ice. Add 2.5μl T4 DNA ligase and mix. Incubate at 16 °C for 2 hours. Mix occasionally. Wash the tubes with wash buffer to remove the excess adaptors.

Cleaving with Tagging Enzyme MmeI

Place the two tubes (A and B) on magnetic stand, remove the supernatant (wash buffer). Add the following reagents to each tube: ddH20 86μl, MmeI 2μl, 10X buffer10μl (1 :10) SAM2μl. Total volume 100μl. Incubate at 37°C for 1 hour with mixing.

Remove the supernatant which contains the tags. Phenol/chloroform and precipitate.

Ligating Tags to Create Ditags

Resuspend the pellet in 1.5μl, mix both A and B tubes and add 1.5μl of the following reagents to form the ditags: 3mM Tris-HCl, pH 7.5 1.25μl, 10X Ligation buffer 0.75μl, ddH_20 0.75μl, T4 ligase 1μl. Incubate at 16°C for overnight.

PCR Amplifying and Gel Purifying the 138 Bp Ditags

Set up 200 to 300 reactions as following (in 96 PCR plates):

10XBV buffer	5μl
DMSO	3μl
dNTPmix	7.5μl
PCR primer A (175ng/μl)	2μl
PCR primer B (175ng/μl)	2μl
ddH_20	29μl
Platinum Taq enzyme	0.5μl
Template (1: 230 dilution)	1μl
Total volume	50μl

Cycling condition: 95°C 2' 1 cycle. 95°C 30", 55°C 1', 70°C 1'. 27 cycles. 70°C 5'. Run a diagnostic 4% agarose gel to analyze the PCR product and purify the 138 bp ditags by 12% polyacrylamide gel electrophoresis.

Digesting the Ditags with Anchoring Enzyme Bseri and Purifying the 48 Bp Ditag

Digest the 138bp ditags with BseRI to yield 48bp ditag.

138bp ditags	42μl
10X NEB buffer II	15μl
BseRI	12μl
ddH_20	81μl
Total volume	150μl

And incubate at 37°C for 2 to 3 hours. Phenol/chloroform and EtoH precipitate. Run a diagnostic 4% agarose gel to check the efficiency of the digestion and purify the 48 bp ditags by 12% polyacrylamide gel electrophoresis. Ligate the 48 bp ditag to form concatamers.

Ligating the 48 Bp Ditag to Form Concatamers

Set up the ligation reaction using the gel purified ditag in a total volume of 10μl and incubate 16 C for 2 to 3 hours. Gel purify the concatamers from 8% polyacrylamide, select the DNA size range from 1 to 1.5kbp.

Cloning concatamers into Vector

In a separate reaction, digest the plasmid vector with HhaI and ligate the purified concatamers into the linearized vector following standard molecular biology protocol. Transform the ligation into high efficiency competent cells.

To keep on walking into the 3' end of the cDNA, just take the beads from step: **Cleavage the cDNA with tagging enzyme MmeI** and repeat the whole process again.

3' SAGE Walk Operational Protocol

Follow the 3' terminal tag protocol to step: **Cleaving with Tagging enzyme MmeI.**

After the digestion is completed, inactivate the enzyme by washing the tube twice with buffer containing 1%SDS. Four more washes with wash buffer (5mM Tris-HCl pH 7.5, 0.5mM EDTA. 1mM NaCl and 200μg/μl BSA). For the last wash use 1X ligation buffer.

Ligating MmeI-BseRI Adaptors to cDNA

Place the tube on magnetic stand to remove the supernatant (the ligation buffer). Add the following reagents to the beads on ice.

Contents	Tube A	Tube B
cDNA beads	beads	beads
MmeI-BseRI adaptor A (20ng/μl)	1.5μl	
MmeI-BseRI adaptor B (20ng/μl)		1.5μl
TE	14μl	14μl

10X ligation buffer	2µl	2µl

Heat the tubes for 2' at 50 °C and cool to room temp. for 15' and chill the samples on ice. Add 2.5µl T4 DNA ligase and mix. Incubate at 16 °C for 2 hours. Mix occasionally. Wash the tubes with wash buffer to remove the excess adaptors.

Cleaving with Tagging enzyme MmeI

Place the two tubes (A and B) on magnetic stand, remove the supernatant (wash buffer). Add the following reagents to each tube.

ddH_20	86µl
MmeI	2µl
10X buffer	10µl
(1 :10) SAM	2µl
Total volume	100µl.

Incubate at 37°C for 1 hour with mixing. Remove the supernatant which contains the tags. Phenol/chloroform and precipitate. Resuspend the pellet.

Ligating Tags to Create Ditags

Pool the A and B tubes. Resuspend the pellet in 1.5µl and add 1.5µl of the following reagents to form the ditags.

3mM Tris-HCl, pH 7.5	1.25µl
10X Ligation buffer	0.75µl
ddH_20	0.75µl
T4 ligase	1µl

Incubate at 16°C for overnight.

PCR Amplifying and Gel Purifying the 138 Bp Ditags

Set up 200 to 300 reactions as following (in 96 PCR plates).

10XBV buffer	5µl
DMSO	3µl
dNTPmix	7.5µl

PCR primer A (175ng/μl)	2μl
PCR primer B (175ng/μl)	2μl
ddH_20	29μl
Platinum Taq enzyme	0.5μl
Template (1: 230 dilution)	1μl
Total volume	50μl

Cycling condition: 95°C 2' 1 cycle. 95°C 30", 55°C 1', 70°C 1'. 27 cycles.70°C 5'. Run a diagnostic 4% agarose gel to analyze the PCR product and purify the 138 bp ditags by 12% polyacrylamide gel electrophoresis.

Digesting the Ditags with Anchoring Enzyme Bseri and Purifying the 48 Bp Ditag

Digest the 138bp ditags with BseRI to yield 48bp ditag.

138bp ditags	42μl
10X NEB buffer II	15μl
BseRI	12μl
ddH_20	81μl
Total volume	150μl

And incubate at 37°C for 2 to 3 hours. Phenol/chloroform and EtoH precipitate. Run a diagnostic 4% agarose gel to check the efficiency of the digestion and purify the 48 bp ditags by 12% polyacrylamide gel electrophoresis.

Ligating the 48 Bp Ditag to Form Concatamers

Set up the ligation reaction using the gel purified ditag in a total volume of 10μl and incubate 16 C for 2 to 3 hours. Gel purify the concatamers from 8% polyacrylamide, select the DNA size range from 1 to 1.5kbp.

Cloning Concatamers Into Vector

In a separate reaction, digest the plasmid vector with HhaI and ligate the purified concatamers into the linearized vector following standard molecular biology protocol. Transform the ligation into high efficiency competent cells.

To keep on walking toward the 5' end of the cDNA, just take the beads from step: **Cleavage the cDNA with tagging enzyme MmeI** and repeat the whole process again.

Example 6. DNA Sequencing and Data Analysis

DNA templates of the concatamer clones are prepared and subject for standard sequencing analysis. A typical DNA sequence provides about 800bp nucleotide sequences and contains about 17-18 ditags, or 30-40 tags.

With a concatamer clone of larger than 2000bp of insert, sequencing analysis is performed from both ends to generate up to 80 tags per run. Therefore, a modest sequencing effort of 12,500 clones generates over one million tags, and 25,000 clones sequences (50,000 reads) generates one million pairs of 5' and 3' LongSAGE tags.

After sequencing of generating the draw sequence data, the ditag sequences are first extracted then the single tags. Both 5' and 3' tags are compared together to the corresponding genome sequence database using BLAST. Pairs of 5' and 3' tags of known genes and putative genes are clustered based on genome and cDNA sequences.

REFERENCES

APPENDIX A: SEQUENCES

5' TERMINAL SAGE LINKERS AND PRIMERS

NotI-dT15 primer for cDNA synthesis

5'-GACTAGTTCTAGATCGCGAGCGGCCGCCC(T)$_{15}$VN
NotI

Biotin-NotI linker

5'-biotin-GACTAGTTCTAGATCGCGAGC
CTGATCAAGATCTAGCGCTCGCCGGp-3

MmeI-BseRI Linker A (48nt)

BseRI MmeI
5'-TTT GGA TTT GCT GGT GCA GTA CAA CTA GGC GAGGAGCGTCTCTCCGAC
a-CCT AAA CGA CCA CGT CAT GTT GAT CCG CTCCTCGCAGAGAGGCTG-P
BsmBI

MmeI-BseRI Linker B (48nt)

BseRI MmeI
5'-TTT CTG CTC GAA TTC AAG CTT CTA ACG TAC GAGGAGCGTCTCTCCGAC
a-GAC GAG CTT AAG TTC GAA GAT TGC ATG CTCCTCGCAGAGAGGCTG-P
BsmBI

Second round MmeI-BseRI Linker A (50nt)

BseRI MmeI
5'-TTT GGA TTT GCT GGT GCA GTA CAA CTA GGC GAGGAGCGTCTCTCCGACNN
a-CCT AAA CGA CCA CGT CAT GTT GAT CCG CTCCTCGCAGAGAGGCTG-P
BsmBI

Second round MmeI-BseRI Linker B (50nt)

BseRI MmeI
5'-TTT CTG CTC GAA TTC AAG CTT CTA ACG TAC GAGGAGCGTCTCTCCGACNN
a-GAC GAG CTT AAG TTC GAA GAT TGC ATG CTCCTCGCAGAGAGGCTG-P
BsmBI

PCR primer A (29nt, 20nt)

biotin-5'-GTG CTC GTG GGA TTT GCT GGT GCA GTA CA

PCR primer B (29nt, 20nt)

biotin-5'-GAG CTC GTG CTG CTC GAA TTC AAG CTT CT

3' TERMINAL SAGE LINKERS AND PRIMERS

FseI-dT15 primer (for cDNA synthesis)

biotin-GACTAGTTCTAGATCGCGAGGCCGGCC(T)$_{13}$VN
FseI

5' biotin linker

(The biotin oligo is same as the PCR primer A)

5'-biotin-GTGCTCGTGGGATTTGCTGGTGCAGTACA

CACGAGCACCCTAAAGCACCACGTCATGT-p-3'

1/2FseI-MmeI linker (to introduce MmeI site and remove polyA)

pTCGGATCGCGATCTAGAACTAGTC
3'-GGCCAGCCTAGCGCTAGATCTTGATCAGTTT-5'
MmeI

Second round MmeI-BseRI Linker A (50nt)

BseRI MmeI
5'-TTT GGA TTT GCT GGT GCA GTA CAA CTA GGC GAGGAGCGTCTCTCCGACNN
a-CCT AAA CGA CCA CGT CAT GTT GAT CCG CTCCTCGCAGAGAGGCTG-P
BsmBI

Second round MmeI-BseRI Linker B (50nt)

BseRI MmeI
5'-TTT CTG CTC GAA TTC AAG CTT CTA ACG TAC GAGGAGCGTCTCTCCGACNN
a-GAC GAG CTT AAG TTC GAA GAT TGC ATG CTCCTCGCAGAGAGGCTG-P
BsmBI

PCR primer A (29nt, 20nt)
biotin-5'-GTG CTC GTG GGA TTT GCT GGT GCA GTA CA

PCR primer B (29nt, 20nt)
biotin-5'-GAG CTC GTG CTG CTC GAA TTC AAG CTT CT

EXAMPLES 1 & 2 PRIMER / LINKER SEQUENCES

NotI-dT20 primer 40 mers
GAGAGAGAGAGCGGCCGCTTTTTTTTTTTTTTTTTTTVN

NotI Linker top 25 mers
5'phosphate-GGCCGCTCGCGATCTAGAACTAGTC

NotI Linker bottom 21mers
5'biotin-GACTAGTTCTAGATCGCGAGC

Linker A top (N5) 54mers
TTTGGATTTGCTGGTGCAGTACAACTAGGCGAGGAGCGTCTCTCCGACGNNNNN

Linker A top (N6) 54mers
TTTGGATTTGCTGGTGCAGTACAACTAGGCGAGGAGCGTCTCTCCGACNNNNNN

Linker A bottom 45mers
5'phosphate-GTCGGAGAGACGCTCCTCGCCTAGTTGTACTGCACCAGCAAATCC

Linker B top (N5) 54mers
TTTCTGCTCGAATTCAAGCTTCTAACGTACGAGGAGCGTCTCTCCGACGNNNNN

Linker B top (N6) 54mers
TTTCTGCTCGAATTCAAGCTTCTAACGTACGAGGAGCGTCTCTCCGACNNNNNN

Linker B bottom 45mers

5'phosphate-GTCGGAGAGACGCTCCTCGTACGTTAGAAGCTTGAATTCGAGCAG

PCR primer A 29mers
GTGCTCGTGGGATTTGCTGGTGCAGTACA

PCR primer B 29mers
GAGCTCGTGCTGCTCGAATTCAAGCTTCT

EXAMPLES 3 & 4 PRIMER / LINKER SEQUENCES

GsuI-dT16 primer 34 mers
GAGAGAGAGACTGGAGTTTTTTTTTTTTTTTVN

MmeI-BseRI Linker A top 50 mers
5'-TTTGGATTTGCTGGTGCAGTACAACTAGGCGAGGAGCGTCTCTCCGACTT

MmeI-BseRI Linker A bottom 45 mers
5'phosphate-GTCGGAGAGACGCTCCTCGCCTAGTTGTACTGCACCAGCAAATCC

MmeI-BseRI Linker B top 50 mers
5'-TTTCTGCTCGAATTCAAGCTTCTAACGTACGAGGAGCGTCTCTCCGACTT

MmeI-BseRI Linker B bottom 45mers
5'phosphate-GTCGGAGAGACGCTCCTCGTACGTTAGAAGCTTGAATTCGAGCAG

SalI adaptor top 16mers
5'biotin-TCGACCCACGCGTCCG

SalI adaptor bottom 16mers
5'phosphate-CGGACGCGTGGGTCGA

PCR primer A 29mers
GTGCTCGTGGGATTTGCTGGTGCAGTACA

PCR primer B 29mers
GAGCTCGTGCTGCTCGAATTCAAGCTTCT

EXAMPLE 5 SAGE WALKING PRIMER / LINKER SEQUENCES

5' SAGE Walking Primer / Linker sequences

Linker A top 50mers
TTTGGATTTGCTGGTGCAGTACAACTAGGCGAGGAGCGTCTCTCCGACNN

Linker A bottom 45mers
5'phosphate-GTCGGAGAGACGCTCCTCGCCTAGTTGTACTGCACCAGCAAATCC

Linker B top 50mers
TTTCTGCTCGAATTCAAGCTTCTAACGTACGAGGAGCGTCTCTCCGACNN

Linker B bottom 45mers

5'phosphate-GTCGGAGAGACGCTCCTCGTACGTTAGAAGCTTGAATTCGAGCAG
PCR primer A 29mers
GTGCTCGTGGGATTTGCTGGTGCAGTACA

PCR primer B 29mers
GAGCTCGTGCTGCTCGAATTCAAGCTTCT

3' SAGE Walking Primer / Linker sequences

MmeI-BseRI Linker A top 50 mers
5'-TTTGGATTTGCTGGTGCAGTACAACTAGGCGAGGAGCGTCTCTCCGACNN

MmeI-BseRI Linker A bottom 45 mers
5'phosphate-GTCGGAGAGACGCTCCTCGCCTAGTTGTACTGCACCAGCAAATCC

MmeI-BseRI Linker B top 50 mers
5'-TTTCTGCTCGAATTCAAGCTTCTAACGTACGAGGAGCGTCTCTCCGACNN

MmeI-BseRI Linker B bottom 45mers
5'phosphate-GTCGGAGAGACGCTCCTCGTACGTTAGAAGCTTGAATTCGAGCAG
PCR primer A 29mers
GTGCTCGTGGGATTTGCTGGTGCAGTACA

PCR primer B 29mers
GAGCTCGTGCTGCTCGAATTCAAGCTTCT

APPENDIX B: TABLE 1

Tag length n	Tag Complexity, C	Tag uniqueness to transcriptome, Ut	Tag uniqueness to human genome, Ug
4	256	1562	11718750
6	4096	146	732422
10	1048576	0.95	2861
14	268435456	0.0052	11
16	4294967296	0.00037	0.7
18	68719476736	0.000026	0.04
20	1099511627776	0.0000018	0.003

Table 1. Theoretical 5'tag complexity and uniqueness probability of matching to transcriptome and genome

C, the complexity of tags. $C = 4^n$, where n is the tag length. Short tags have less complexity and, therefore, are more frequent in a genome. The numbers indicate, in average, the size of sequence base pairs for a specific tag to occur once. For example a 6-bp tag will occur once in a genome every 4096 bp, while a 10-bp tag will appear only once in a one million bp sequence.

Ut, tag uniqueness to transcripts of a genome, is determined under the assumption that there are 100,000 distinct transcripts in a complex genome like human, mouse, or Zebrafish, and the 5'sequences of transcripts are random. Ut = n*100,000/C, indicates the number of transcripts that might share a same unique 5'tag.

Ug, the tag uniqueness to human genome. Ug is determined as the genome size (3,000,000,000 bp) divided by tag complexity. Ug = Gs/C, Gs is the genome size in bp. The numbers indicate the potential matches of a particular tag in the genome.

CLAIMS

1. A method of generating a database comprising a plurality of records, each record comprising an indication whether a gene is expressed by a particular cell, the method comprising steps of:

(a) providing a first nucleic acid sequence;

(b) linking the first nucleic acid sequence to an linker sequence to form a linked nucleic acid, in which the linker sequence comprises:

(i) a first recognition site for a first nucleic acid cleavage enzyme that allows cleavage of the first nucleic acid sequence at a site distant from the first recognition site, and

(ii) a second recognition site for a second nucleic acid cleavage enzyme that allows nucleic acid cleavage at a site distant from the second recognition site, said cleavage site located at a position within or about the first recognition site;

in which the linked nucleic acid has the structure: 5' - second recognition site - first recognition site - first nucleic acid - 3'

(c) cleaving the linked nucleic acid with the first nucleic acid cleavage enzyme to provide:

(i) a linked tag comprising the linker sequence linked to a nucleotide sequence tag representative of the first nucleic acid sequence and comprising a terminal portion thereof; and

(ii) a second nucleic acid sequence comprising a remainder portion of the first nucleic acid.

(d) detecting the presence, sequence or identity of the linked tag or the nucleotide sequence tag to provide an indication of an instance of gene expression; and

storing an indication of expression of the gene as a record in the database.

2. A database generation method according to Claim 1, in which the first nucleic acid comprises a complementary deoxyribonucleic acid (cDNA) having a terminus comprising

a 5' terminal transcribed sequence of a gene, and in which the linker sequence is linked to said terminus.

3. A database generation method according to Claim 1, in which the first nucleic acid comprises a complementary deoxyribonucleic acid (cDNA) having a terminus comprising a 3' terminal transcribed sequence of a gene, and in which the linker sequence is linked to said terminus.

4. A database generation method according to Claim 1, 2 or 3, in which the second nucleic acid cleavage enzyme comprises a restriction endonuclease, preferably a Type IIS restriction endonuclease, whose recognition site is 6 bases or greater, preferably MmeI.

5. A database generation method according to any preceding claim, in which the first nucleic acid cleavage enzyme comprises a restriction endonuclease, preferably a Type IIS restriction endonuclease, preferably BseRI.

6. A database generation method according to any preceding claim, in which the method comprises sequentially generating a plurality of nucleic acid sequences each comprising a nucleotide sequence tag from a nucleic acid, the method comprising repeating steps (a) to (c) of any of Claims 1 to 5 at least once, in which the first nucleic acid sequence of step (a) is provided by the second nucleic acid sequence of step (c)(ii); and storing each indication of an instance of gene expression generated as a record in the database.

7. A database generation method of generating a database comprising a plurality of records, each record comprising an indication whether a gene is expressed by a particular cell, the method comprising the steps of:

(a) providing a first linked tag and a second linked tag, each independently produced by a method according to any preceding claim;

(b) linking the first linked tag and the second linked tag such that the nucleotide sequence tag portion of one linked tag is linked to the nucleotide sequence tag of the other linked tag to form a ditag, the ditag comprising terminal transcribed sequences from first and second genes; and

(c) detecting the presence or identity of the ditag, or at least one nucleotide sequence tag comprised therein, to detect gene expression; and

storing the indication of expression of the gene as a record in the database.

8. A database generation method according to Claim 7, in which each of the first and second linker sequences comprised in the first and second linked tags comprises an amplification primer sequence, and in which the method further comprises a step of amplifying the ditag, preferably by means of the polymerase chain reaction (PCR).

9. A database generation method according to Claim 7 or 8, further comprising the steps of:

(d) cleaving the ditag with the or each second nucleic acid cleavage enzyme;

(e) linking a plurality of the resultant trimmed ditags to form a concatamer; and

(f) obtaining the nucleic acid sequence of at least a portion of the concatamer.

10. A database generation method of generating a database comprising a plurality of records, each record comprising an indication whether a gene is expressed by a particular cell, the method comprising:

(a) obtaining an indication of an instance of expression of a gene by a method comprising:

(i) providing a complementary deoxyribonucleic acid (cDNA) having a terminus comprising a terminal transcribed sequence of a gene;

(ii) linking the cDNA to an linker sequence thereby forming a linked nucleic acid, in which the linker sequence comprises a first recognition site for a first nucleic acid cleavage enzyme, preferably a restriction endonuclease, that allows nucleic acid cleavage at a site distant from the first recognition site; and

(iii) cleaving the linked nucleic acid with the first nucleic acid cleavage enzyme to provide a linked tag, in which the linked tag comprises a nucleotide sequence tag representative of a terminal transcribed sequence of the gene; and

(iv) detecting the presence or identity of the linked tag or the nucleotide sequence tag to provide an indication of an instance of gene expression; and

storing the indication of expression of the gene as a record in the database.

11. A database generation method according to Claim 10, in which the 5' end of the cDNA comprises a sequence corresponding to the 5' terminal transcribed sequence of the gene, and in which the linker sequence is linked to the 5' end of the cDNA.

12. A database generation method according to Claim 10 or 11, in which the nucleotide sequence tag comprises a 5' terminal transcribed sequence of the gene, preferably at least the first 16 bases of the transcribed portion of the gene, more preferably the first 20 bases of the transcribed portion of the gene.

13. A database generation method according to Claim 10, in which the 3' end of the cDNA comprises a sequence corresponding to the 3' terminal transcribed sequence of the gene, and in which the linker sequence is linked to the 3' end of the cDNA.

14. A database generation method according to Claim 10 or 13, in which the nucleotide sequence tag comprises a 3' terminal transcribed sequence of the gene, preferably at least the last 16 bases of the transcribed portion of the gene, more preferably the last 20 bases of the transcribed portion of the gene.

15. A database generation method according to Claim 10, 13 or 14, in which step (a) comprises:

(i) deriving a cDNA from an mRNA by reverse transcription with an primer comprising a sequence 5'-NV(T)^{13}CCGGCCGG-3', in which N = A, C, G or T and V = A, C or G;

(ii) producing a double stranded cDNA therefrom and digesting the double stranded cDNA with FseI to produce a cleaved cDNA comprising

3' $\frac{\text{GGCCGG}}{\text{CC}}$ overhang;

(iii) linking the cleaved cDNA to a linker comprising $\frac{\text{pTCGGA}}{\text{GGCCAGCCT}}$; and

(iv) cleaving the resulting molecule with MmeI to produce a cDNA lacking a polyA/T tail.

16. A database generation method according to any of Claims 10 to 15, in which the cDNA comprises the 5' terminal transcribed sequence of the gene, or the 3' terminal transcribed sequence of the gene, or both.

17. A database generation method according to any of Claims 10 to 16, in which the cDNA is full length cDNA, preferably comprising substantially all the coding sequence of the gene.

18. A database generation method according to any of Claims 10 to 17, in which the linker sequence further comprises a second recognition site for a second nucleic acid cleavage enzyme, preferably a second restriction endonuclease, that allows nucleic acid cleavage at a site distant from the second recognition site, and in which the second recognition site is located 5' of the first recognition site in the linker sequence.

19. A database generation method according to Claim 18, in which the location of the second recognition site with respect to the first recognition site is such that, when the linked tag is exposed to the second nucleic acid cleavage enzyme, cleavage of the nucleic acid occurs at a position within or about the first recognition site.

20. A database generation method according to any of Claims 10 to 19, in which the first recognition site or the second recognition site, or both, comprises a Type IIS restriction enzyme recognition site.

21. A database generation method according to any of Claims 10 to 20, in which the first restriction enzyme comprises Mme1, or in which the first recognition site comprises a Mme1 recognition site 5'-TCC RAC-3', preferably 5'-TCC GAC-3'.

22. A database generation method according to Claim 20 or 21, in which the second restriction enzyme comprises a restriction enzyme capable of recognising a site which is 6 bases or longer, preferably BseR1, or in which the second recognition site comprises a BseRI recognition site 5'-GAGGAG-3'.

23. A database generation method according to any of Claims 11 to 22, in which the linker sequence comprises the sequence 5'- GAGGAGNNNNNNTC▲CG▼AC -3', preferably 5'-GAGGAGCGTCTCTCCGAC-3'.

24. A database generation method of generating a database comprising a plurality of records, each record comprising an indication whether a gene is expressed by a particular cell, the method comprising steps of:

(a) providing a first complementary deoxyribonucleic acid (cDNA) having a terminus comprising a terminal transcribed sequence of a first gene;

(b) providing a second complementary deoxyribonucleic acid (cDNA) having a terminus comprising a terminal transcribed sequence of a second gene;

(c) linking the first cDNA so produced to a first linker sequence thereby forming a first linked nucleic acid, in which the first linker sequence comprises a first recognition site for a first nucleic acid cleavage enzyme, preferably a first restriction endonuclease, that allows nucleic acid cleavage at a site distant from the first recognition site;

(d) linking the second cDNA so produced to a second linker sequence thereby forming a second linked nucleic acid, in which the second linker sequence comprises a second recognition site for a second nucleic acid cleavage enzyme, preferably a second restriction endonuclease, that allows nucleic acid cleavage at a site distant from the second recognition site;

(e) cleaving the first linked nucleic acid with the first nucleic acid cleavage enzyme to provide a first linked tag, in which the first linked tag comprises a first nucleotide sequence tag representative of a terminal transcribed sequence of the first cDNA.

(f) cleaving the second linked nucleic acid with the second nucleic acid cleavage enzyme to provide a second linked tag, in which the second linked tag comprises a second nucleotide sequence tag representative of a terminal transcribed sequence of the second cDNA.

(g) ligating the first and second tags to form a ditag; and

(h) determining the nucleotide sequence of at least one tag of the ditag to detect gene expression; and

storing the indication of expression of the gene as a record in the database.

25. A database generation method according to any preceding claim, further comprising storing the database in a computer readable medium.

26. A method of generating a computer readable medium comprising a database comprising a plurality of records which indicate whether a gene is expressed by a particular cell, the method comprising generating a database by a method according to any preceding claim and storing the database on the computer readable medium

27. A method of generating a database substantially as hereinbefore described with reference to and as shown in Figures 2 to 4 of the accompanying drawings.

28. A method of generating a computer readable medium substantially as hereinbefore described with reference to and as shown in Figures 2 to 4 of the accompanying drawings.

ABSTRACT

METHOD OF GENERATING A DATABASE

We describe a method of generating a database comprising a plurality of records, each record comprising an indication whether a gene is expressed by a particular cell, the method comprising steps of: (a) providing a first nucleic acid sequence; (b) linking the first nucleic acid sequence to an linker sequence to form a linked nucleic acid, in which the linker sequence comprises: (i) a first recognition site for a first nucleic acid cleavage enzyme that allows cleavage of the first nucleic acid sequence at a site distant from the first recognition site, and (ii) a second recognition site for a second nucleic acid cleavage enzyme that allows nucleic acid cleavage at a site distant from the second recognition site, said cleavage site located at a position within or about the first recognition site; in which the linked nucleic acid has the structure: 5' – second recognition site – first recognition site – first nucleic acid – 3' (c) cleaving the linked nucleic acid with the first nucleic acid cleavage enzyme to provide (i) a linked tag comprising the linker sequence linked to a nucleotide sequence tag representative of the first nucleic acid sequence and comprising a terminal portion thereof: and (ii) a second nucleic acid sequence comprising a remainder portion of the first nucleic acid; (d) detecting the presence, sequence or identity of the linked tag or the nucleotide sequence tag to provide an indication of an instance of gene expression; and storing an indication of expression of the gene as a record in the database.

Figure 3