บทฟื้นฟูวิชาการ • Review Article

การวิเคราะห์การแสดงออกของยีนด้วยวิธีที่ได้ปริมาณงานสูง

กัลยาณี สวรรยาวิสุทธิ์
ภาควิชาชีวเคมี คณะแพทยศาสตร์ มหาวิทยาลัยขอนแก่น

High Throughput Gene Expression Analysis: a Review

Kanlayanee Sawanyawisuth
Department of Biochemistry, Faculty of Medicine, Khon Kaen University

ปัจจุบันข้อมูลจากการศึกษาจีโนมของมนุษย์เข้ามามีบทบาทในการศึกษาโรคต่างๆ ในทางการแพทย์เป็นอย่างมาก เนื่องจากโรคส่วนใหญ่มีความซับซ้อนในกลไกการเกิดโรค และสัมพันธ์กับยีนจำนวนมาก ดังนั้นการวิเคราะห์การแสดงออกของยีนด้วยวิธีที่ได้ปริมาณงานสูง จึงเป็นเทคนิคที่สามารถให้ข้อมูลการแสดงออกของยีนจำนวนมหาศาลได้ในเวลาเดียวกัน และทำให้เข้าใจกลไกการเกิดโรคในระดับโมเลกุลมากขึ้น การวิเคราะห์การแสดงออกของยีนด้วยวิธีที่ได้ปริมาณงานสูง แบ่งออกเป็น 2 ประเภท ได้แก่ ระบบปิด และระบบเปิด ซึ่งในแต่ละระบบมีหลากหลายเทคนิคที่มีข้อดี และข้อเสียแตกต่างกัน

Currently, human genome study project is broadly involved in most of medical fields. Most diseases are complex process and many are involved in multiple genes. Nowadays, high throughput gene expression analysis facilitates the massive information of gene expression at the same time. It will provide us with better understanding of the diseases at the molecular level. This review aims to describe high throughput gene expression analysis techniques. The analysis has two main categories including closed and open system. Each system has various techniques that have their own advantages and disadvantages.

ศรีนครินทร์เวชสาร 2552; 24(2): 154-8 • Srinagarind Med J 2009; 24(2): 154-8

Introduction

The genomic sequences of a wide variety of organisms, including that of humans, are being elucidated one after another. The genomes of eukaryotic organisms are long, massive, and contain an enormous number of genes. By delicately regulating activities of these genes, each organism can supply required amount of products at an appropriate time that confer functions proper to the organism. It is thus believed that the majority of biological phenomena found in a variety of organisms can be explained by the quantity of gene products.

Although the gene function is certainly conducted by its final protein product, there are a large number of observations that the amount of protein produced is directly dependent on the amount of mRNA that encodes it. This means that to understand the cellular functions under the certain conditions at a certain time; it can be attained by measuring the species and respective numbers of mRNAs at that point of time.

However, each cell contains more than 10,000 species of mRNA transcripts.[1] Therefore, high throughput or large-scale gene expression analysis, is a technology for simultaneously analyzing the expression levels of large numbers of genes, provides the opportunity to study the activity of whole genomes, rather than the activities of single, or a few genes.

Categories of high throughput gene expression analysis

This technology is broadly divided into two families: closed system and open system.

1. Closed system

Closed system is inherently limited by the retrospective nature of the inquiry-one must begin with genes that are known. Coverage of genomes is strictly dependent upon the completeness of the knowledge of that genome, thus severely limiting the applicability to the most well characterized species or systems. Genes that are not represented in a closed system are not assessed. The most common method of closed systems is oligonucleotides or cDNA microarray technologies.[2]

Microarray technology

This technology is designed for the simultaneous measurement of the expression of several thousand genes in a single hybridization procedure.[3] Microarray is manufactured in a reproducible pattern of thousands of DNAs (primarily PCR products or oligonucleotides) attached to a solid support such as glass. Fluorescent-labeled DNA or RNA prepared from mRNA is then hybridized to their complementary DNA contained on the microarray and detected via laser scanning. Differences in labeling intensity are converted into a quantitative output of relative gene expression.

2. Open system

Open system is defined by the fact that no comprehensive knowledge of the transcriptome is required. Open system has a natural advantage over closed system in that the source of the transcriptome and its inherent complexity such as, alternative splices and RNA editing, etc., is immaterial and is not a barrier to discovery.

Closed system and open system are complementary. Once novelty has been identified, whether it is absolute novelty or known genes in less well-characterized systems, these genes can then be used in directed ways in the closed system. These two systems have in common that the output of analysis results in gene lists, which require a well-planned strategy for annotation and classification by functional role hierarchies, as has been applied to already completed genomes.[4]

Examples of open system are differential display (DD), restriction enzyme analysis of differentially expressed sequences (READS) and serial analysis of gene expression (SAGE).

2.1 Differential Display (DD)

The earliest technology for transcript profiling is differential display (DD).[5] DD is an expression analysis method whereby mRNA from each sample is converted to cDNA, cDNA is PCR-amplified using a combination of random primers and anchored oligo-dT primers, and then run on a gel. Each mRNA is represented as a single band and differentially expressed bands are excised, cloned, and sequenced to reveal identity. A particular difficulty with DD is a high rate of false positives.[6] However, DD is still the most widely used method for expression analysis because it can be performed in any laboratory equipped with standard molecular biology reagents and instrumentation, and in its most basic form, the need for advanced bioinformatics is minimal.

2.2 Restriction Enzyme Analysis of Differentially Expressed Sequences (READS)

READS[7], generates only one restriction tag for each gene fragment. By the use of a unique Y-shaped adapter design, only the most 3'- end fragment including part of the poly-A tail is amplified. The cDNA pool which is divided into subpopulations based on a set of primers and multiple iterations using several different single restriction enzymes are employed to ensure good coverage of the expressed genome. Differentially expressed genes are identified by differences in gel band intensities. Excision, cloning, and DNA sequencing determine the identity of the differentially expressed fragments.[8]

2.3 Serial Analysis of Gene Expression (SAGE)

SAGE is a comprehensive profiling method that allows for global, unbiased and quantitative characterization of transcriptomes.[9] The SAGE method is based on the isolation of unique sequence tags (10-11 bp in length) from individual mRNAs and concatenation of tags serially into long DNA molecules for sequencing. SAGE provides a statistical description of the mRNA population present in a cell without prior selection of the genes to be studied, and this constitutes a major advantage. A second major advantage is that the information generated is digital in format, and can

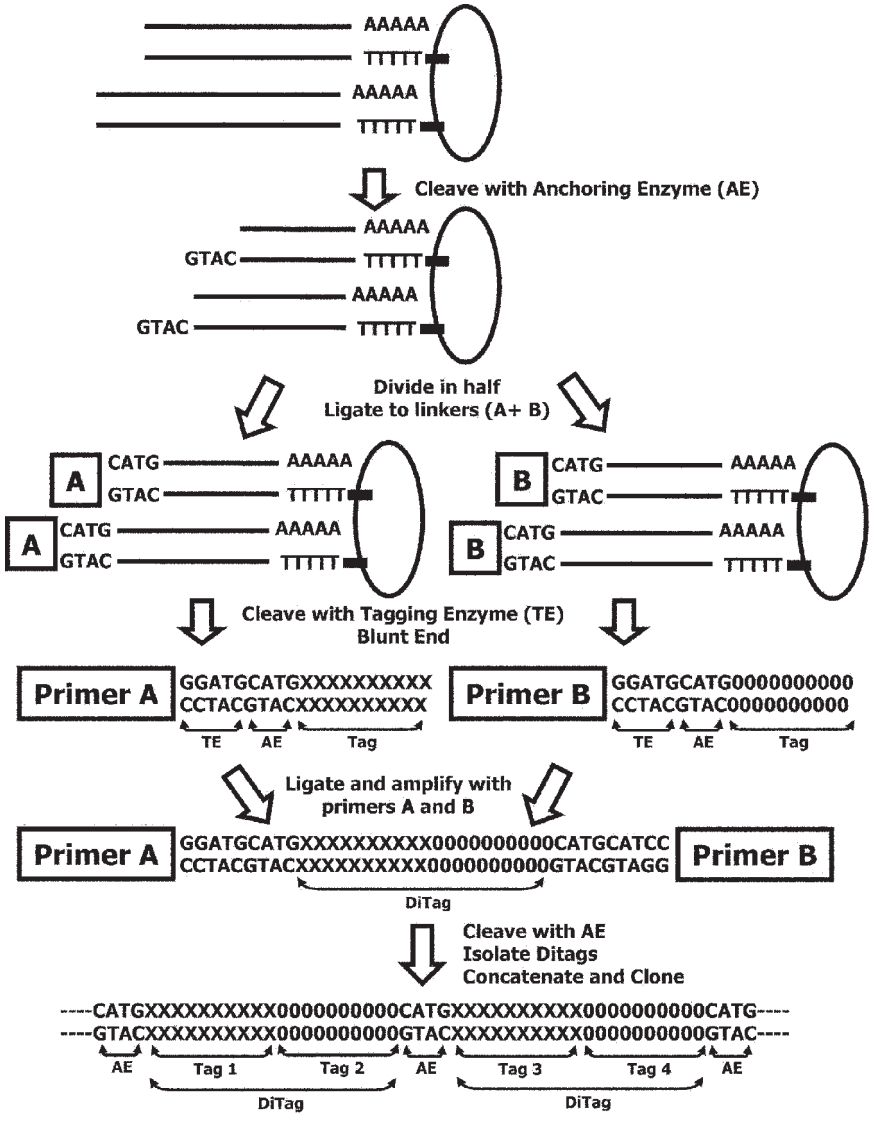

Figure 1 The process of SAGE library construction (modified from St. Croix B et al., 2000)[13]

be directly compared with data generated from any other laboratory or with data available in public databases such as the Cancer Genome Anatomy Project.[10]

SAGE is based on the following two principles:

1. A short (10-11 bp) oligonucleotide fragment, or SAGE tag, is sufficient to uniquely identify a specific transcript. More recently a 17 + 4 bp restriction site tag has been developed and called 'long SAGE'.[11] A 10-bp oligonucleotide sequence has 4^{10} different potential combinations. Based on SAGE and expressed sequence tags (ESTs) data, approximately 80,000 to 100,000 transcripts are estimated to derive from the approximately 35,000 unique genes encoded by the human genome.[12] Therefore, a 10-bp sequence tag obtained from a defined position in cDNA is sufficient to uniquely identify most human transcripts. This prediction is based on statistical calculations. In practice, multiple independent genes occasionally share the same SAGE tag, and multiple SAGE tags occasionally are derived from a single gene because of alternative 3'-end processing.

2. Concatenation of sequence tags allows the serial analysis of transcripts, significantly increasing the efficiency of sequence-based analysis. Concatemers of SAGE tags subcloned into a vector serve as excellent templates for automated sequencing. Consequently, a single sequencing

reaction can provide information on as many as 30 to 35 different genes. Even so, a typical SAGE experiment requires the sequencing of approximately 3,000 concatamer clones, which is a significant cost and throughput limitation.

The generation of a SAGE library involves sequential enzymatic steps. The process is depicted in Figure 1 and described briefly.[13] Double-stranded cDNA is generated from mRNA isolated from the cells or tissues of interest and immobilized on oligo $(dT)_{25}$-coated magnetic beads. The cDNA is then cleaved with a frequent cutting restriction enzyme to ensure that every cDNA is cleaved at least once. NlaIII is the most frequently used enzyme and cuts DNA at an average of every 256 bp. The most 3'-end of the cDNA (up to the most 3' NlaIII site) is then collected on the beads and ligated to a linker. This linker has a recognition site for a type IIS restriction enzyme (BsmfI) and a PCR primer site. Type IIS restriction enzymes cut DNA into a certain number of bases away from the recognition sequence; therefore, a short fragment of the cDNA (SAGE tag) remains attached to the linker but is cleaved from the beads. These tags are blunt-ended, ligated to each other to form ditags, and used as templates for PCR amplification. Cleavage of this PCR product with NlaIII releases the ditags, which are isolated, concatenated, subcloned into an appropriate vector, and sequenced. The analysis of the SAGE data is performed using SAGE software (Johns Hopkins University, Baltimore, MD), which extracts the tags from the sequence and determines their abundance and identity.

SAGE has been used for the analysis of various cancer types with the aims of deciphering pathways involved in tumorigenesis and identifying novel diagnostic tools, prognostic markers, and potential therapeutic targets.[14, 15] SAGE is one of the techniques used in the National Cancer Institute-funded Cancer Genome Anatomy Project (CGAP). A database with archived SAGE tag counts and online query tools was created and is now the largest source of public SAGE data.[16-19]

Conclusions

High throughput gene expression analysis is a powerful technique that describes all gene expressions of the cells in the certain condition simultaneously. This technique divides to closed and open system. Since no complete information of genes is necessary for open system, it therefore has an advantage over closed system. This technology is useful for better understanding of diseases that are the complex process and have multiple genes involved such as cancer. Researchers reported the differential gene expression profiles of various cancer types and their normal counter parts. These provide the knowledge to discover the molecular mechanism of diseases and lead to the improvement of diagnosis and treatment strategies.

References

1. Yamamoto M, Wakatsuki T, Hada A, Ryo A. Use of serial analysis of gene expression (SAGE) technology. J Immunol Methods 2001; 250:45-66.
2. Marshall A, Hodgson J. DNA chips: an array of possibilities. Nat Biotechnol 1998; 16:27-31.
3. Southern EM, Case-Green SC, Elder JK, Johnson M, Mir KU, Wang L, et al. Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids. Nucleic Acids Res 1994; 22:1368-73.
4. Rubin RB, Merchant M. A rapid protein profiling system that speeds study of cancer and other diseases. Am Clin Lab 2000; 19:28-9.
5. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992; 257:967-71.
6. Martin KJ, Pardee AB. Identifying expressed genes. Proc Natl Acad Sci USA 2000; 97:3789-91.
7. Prashar Y, Weissman SM. Analysis of differential gene expression by display of 3' end restriction fragments of cDNAs. Proc Natl Acad Sci USA 1996; 93:659-63.
8. Prashar Y, Weissman SM. READS: a method for display of 3'-end fragments of restriction enzyme-digested cDNAs for analysis of differential gene expression. Methods Enzymol 1999; 303:258-72.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995; 270:484-7.
10. SAGE Genie. The Cancer Genome Anatomy Project. Available from: http://cgap.nci.nih.gov/SAGE [cite : January 14, 2009]

11. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, et al. Using the transcriptome to annotate the genome. Nat Biotechnol 2002; 20:508-12.
12. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001; 291:1304-51.
13. St. Croix B, Velculescu VE, Zhang L, Zhou W, Traverso G, Vogelstein B, et al. MicroSAGE Detailed Protocol. Baltimore, MD: Johns Hopkins Oncology Center, Howard Hughes Medical Institute; 2000.
14. Parle-McDermott A, McWilliam P, Tighe O, Dunican D, Croke DT. Serial analysis of gene expression identifies putative metastasis-associated transcripts in colon tumour cell lines. Br J Cancer 2000; 83:725-8.
15. Porter DA, Krop IE, Nasser S, Sgroi D, Kaelin CM, Marks JR, et al. A SAGE (serial analysis of gene expression) view of breast tumor progression. Cancer Res 2001; 61:5697-702.
16. Serial Analysis of Gene Expression Tag to Gene Mapping. The National Center for Biotechnology Information. Available from: http://www.ncbi.nlm.nih.gov/SAGE/ [cite: January 14, 2009]
17. Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, et al. An anatomy of normal and malignant gene expression. Proc Natl Acad Sci USA 2002; 99:11287-92.
18. Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, et al. A public database for gene expression in human cancers. Cancer Res 1999; 59:5403-7.
19. Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, et al. SAGEmap: a public gene expression resource. Genome Res 2000; 10:1051-60.