Hunting for differentially expressed genes

C.G. Vedoy, M.H. Bengtson
and M.C. Sogayar

Instituto de Química, Universidade de São Paulo,
São Paulo, SP, Brasil

Abstract

Correspondence

M.C. Sogayar
Instituto de Química, USP
Caixa Postal 26077
05599-970 São Paulo, SP
Brasil
Fax: +55-11-818-3820
E-mail: mcsoga@quim.iq.usp.br

Differentially expressed genes are usually identified by comparing steady-state mRNA concentrations. Several methods have been used for this purpose, including differential hybridization, cDNA subtraction, differential display and, more recently, DNA chips. Subtractive hybridization has significantly improved after the polymerase chain reaction was incorporated into the original method and many new protocols have been established. Recently, the availability of the well-known coding sequences for some organisms has greatly facilitated gene expression analysis using high-density microarrays. Here, we describe some of these modifications and discuss the benefits and drawbacks of the various methods corresponding to the main advances in this field.

Introduction

The identification of differentially expressed genes has been used as an experimental approach to understand not only gene function but also the molecular mechanisms underlying several biological processes. This approach has been used in a wide range of studies including cell cycle control in mammalian cells (1-4), signal transduction in *Drosophila* (5), and circadian rhythms (6). To fully describe the differential gene expression of a given biological system, it is important to ensure that most (or all) differentially expressed mRNAs are represented in the cDNA library, i.e., both abundant and rare mRNA transcripts.

Several methods have been used to analyze differential gene expression, namely, differential hybridization, electronic subtraction (including serial analysis of gene expression; SAGE), differential display reverse transcriptional-polymerase chain reaction (DDRT-PCR), cDNA subtraction and, more recently, DNA chips.

Differential hybridization

The general scheme for differential colony hybridization is based on generation of a cDNA library containing the gene sequences of interest. These cDNA clones are transferred to bacterial plates, in an orderly array, and replica plated onto duplicate membrane filters (7). Each filter is then hybridized to two different ^{32}P-labelled cDNA probes, made from polyA$^+$ RNA. The RNA from which the cDNA probes are made are pre-

pared from any two cell populations that are expected to display differences in gene expression. This strategy was successfully utilized to isolate and characterize the first platelet-derived growth factor (PDGF)-regulated genes (1), genes expressed during the G0-G1 transition in mouse cells (2), the early genetic response to growth factors in mouse fibroblasts (3) and glucocorticoid-regulated genes in C6/ST1 rat glioma phenotypic reversion (8). Although the isolation of differentially expressed genes of unknown sequences in several systems was first achieved by differential screening, this procedure is very laborious and time consuming. DDRT-PCR and subtractive hybridization coupled to PCR appeared as less laborious, more rational and promising approaches.

Differential display

DDRT-PCR relies on randomly primed amplification of a sub-fraction of total mRNA from two cell populations, with the amplicons run side by side on sequencing gels, and with the isolation of cDNA fragments which are expressed at different levels under both conditions. Since the introduction of DDRT-PCR in 1992 (9) over 100 reports regarding improvements and/or successful applications have been published. Although DDRT-PCR seems to be technically simple, the road from band on the gel to a positive clone can be treacherous. The primary criticisms are: 1) a high false-positive rate, 2) questioned ability of DDRT-PCR to identify both abundant and rare mRNAs, 3) coding regions of mRNAs are usually not cloned, and 4) the verification process is time consuming and usually requires a fair amount of RNA. Methodological modifications have since been introduced to streamline the techniques. Major efforts have centered on how to eliminate false positives as approached from a variety of angles, ranging from RNA sample preparation, Northern blot confirmation and primer length variation. A detailed review

can be found in Ref. 10.

Subtractive hybridization

There are numerous protocols for subtractive hybridization, but the principle remains the same. The most common methods have employed subtraction based on synthesized cDNAs instead of mRNA. This procedure improves the final efficiency because it minimizes RNA degradation that may occur during the hybridization procedure. In general, cDNAs from the target cells/tissues are hybridized using a vast molar excess of driver cDNA (control cells/tissue) followed by separation of the double-stranded nucleic acid hybrids from the single-stranded cDNAs (corresponding to differentially expressed mRNAs) by hydroxyapatite or streptavidin-biotin interaction and, more recently, by suppression PCR. The resulting subtracted cDNA is then used either as a labelled probe to screen libraries or for the construction of a subtracted cDNA library.

Separation of single-stranded cDNA by hydroxyapatite has considerable disadvantages. In addition to requiring large amounts of mRNA, the unhybridized mRNA, recovered after chromatography, is very diluted. Moreover, the chromatographic separation procedure requires a high temperature ($60^{o}C$) which presumably increases the probability of mRNA degradation.

A profound modification of cDNA subtraction was obtained by coupling it to amplification by PCR to increase the starting material to be subtracted or to select the resulting subtracted products. This method was first applied by Duguid and Dinauer (11) to identify differentially expressed genes in scrapie infection. An interesting modification of this method (12) utilizes oligo-$(dT)_{30}$-latex particles and PCR. The fine latex particles, with a large surface area, form a milky suspension that can be easily recovered by centrifugation. The poly, A plus RNA can be efficiently annealed to the oligo-

$(dT)_{30}$-latex within a short reaction period and cDNA synthesis is carried out using the annealed mRNA as a template. This allows subtractive hybridization to be carried out in an Eppendorf tube and the unhybridized mRNA to be separated by brief centrifugation at low temperature. The resulting mRNA can be enriched by successive hybridization reactions of unhybridized mRNA to the cDNA-oligo-$(dT)_{30}$-latex in a relatively short period of time and, subsequently, amplified by PCR after conversion to cDNA. This method has been successfully employed in the isolation of cDNA clones that are specific for undifferentiated human embryonal carcinoma cells (12).

A similar approach utilizing the biotin/streptavidin affinity to separate subtracted cDNAs requires no RNA isolation and has been applied to cells removed from cryostat tissue sections of different cell populations (13). This was possible because the reverse transcription-PCR (RT-PCR) technique allows the use of a very small amount of RNA that is reverse-transcribed to cDNA and amplified. The procedure involves homopolymeric A tailing of cDNA synthesized from released RNA using an anchored oligo-dT primer. PCR amplification is then carried out using a biotinylated $(X)_n T_{16}$ primer-adaptor in the presence of biotin-dATP. This biotinylated driver cDNA is twice hybridized, in 50-fold excess, to heterologous target cDNA made with a non-biotinylated primer. Common driver and excess driver cDNA are magnetically removed following the addition of streptavidin-coated magnetospheres which bind to biotinylated strands, leaving behind the enriched target population sequences.

Another important feature added to the subtraction technology arose with the ability to rapidly reduce the number of candidate genes to a few which could be easily characterized. Two techniques with this potential have been described, namely DDRT-PCR and RDA (representational difference analysis), both of which employ PCR to amplify messages to detectable levels, but their mode of operation is fundamentally different.

RDA is a process of subtraction coupled to amplification, originally developed to be used with genomic DNA, as a method capable of revealing the differences between two complex genomes. Differential display amplifies fragments from all represented mRNA species, whereas RDA eliminates those cDNA fragments present in both populations, allowing different cDNAs to stand out. Genomic RDA relies on the generation, by restriction enzyme digestion and PCR amplification, of simplified versions known as representations of the genomes under investigation. In a population of cDNAs derived from some 15,000-50,000 genes in a typical cell, RDA can be directly applied only to the smaller cDNAs while most cDNAs require prior reduction of their complexity before RDA can be applied. This is accomplished by restriction of cDNAs with a four-base cutting enzyme to ensure that the majority of the cDNA species will contain at least one amplifiable fragment, which is sufficient to isolate the difference and identify the gene. Also, elimination of highly abundant sequences is necessary because they can interfere with subtraction and lead to unacceptably high levels of false positives. This becomes particularly important when gene expression is not qualitatively affected after a stimulus, but only varies quantitatively in scale. In this case, it is possible to modify the tester:driver ratio so as to bias the kinetic enrichment in favor of, for example, species that are up-regulated relative to basal levels. A method employing RDA was sucessfully used in studies of recombination activation genes (RAG-1, RAG-2) that are involved in the site-specific V(D)J (variable, diversity, joining) recombinational process which assembles immunoglobulin and T cell receptor genes (14). On the other hand, normalization or equalization protocols can be easily performed to reduce bias in cDNA

sequence representation (15,16).

Normalization not only increases the discriminating power of differential cloning strategies but also provides access to the functionally important class of poorly expressed sequences. Sequences are generally normalized by submitting thermally denatured cDNAs to a self-reassociation reaction and separating the abundant, re-annealed sequences from the rare single stranded species. Additional methods to normalize cDNA libraries have been described (16). Since it is

Figure 1 - Schematic diagram of the cDNA subtraction procedure using the PCR suppression effect (18). Boxes represent the outer and inner portions of adaptors 1 and 2. Solid lines represent the RsaI-digested tester or driver cDNA.
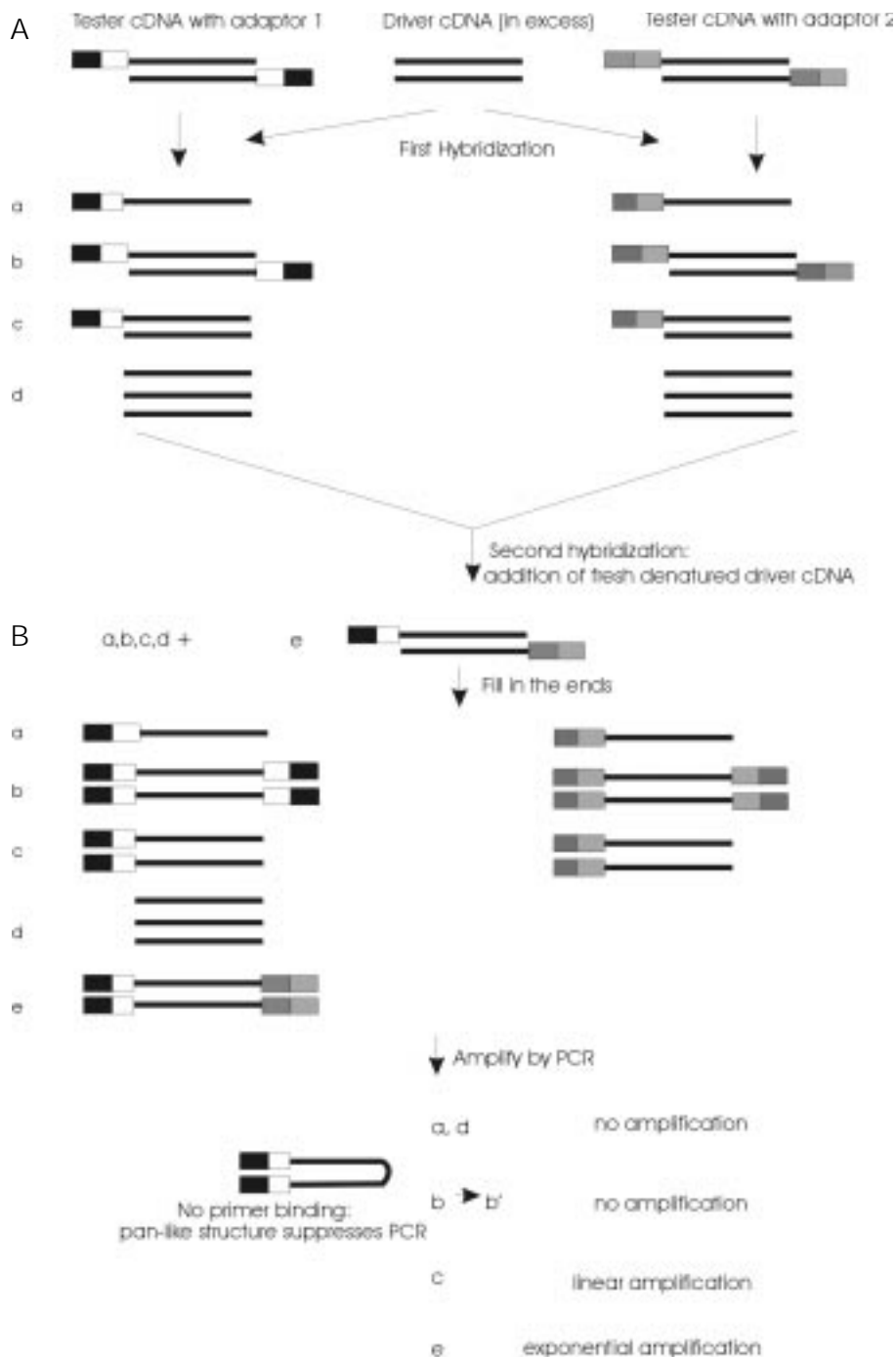
well known that physical methods for the separation of single- and double-stranded DNA are both cumbersome and unreliable, novel approaches which use molecular selection by magnetic beads have been used to eliminate redundant sequences in the normalization procedure (17). However, normalization prior to subtraction is only acceptable when target molecules are entirely absent from the driver cDNA population.

Apparently, the best approach is to apply normalization during the subtraction procedure, as proposed by Gurskaya et al. (18). This method, illustrated in Figure 1, utilizes the suppression PCR effect (19), allowing the development of a high-efficiency subtraction procedure that avoids laborious and ineffective physical separation methods (18). This technology uses an adaptor primer which is shorter in length than the adaptor and is capable of hybridizing to the outer primer-binding site. If any PCR products are generated containing the double-stranded adaptor sequences at the both ends, the individual DNA strands will form pan-like structures following every denaturation step due to the presence of inverted terminal repeats. These structures are more stable than the primer-template hybrid and, therefore, will suppress exponential amplification. The use of this technology has allowed the isolation of transcripts activated upon induction of Jurkat cells by phytohemagglutinin and phorbol 12-myristate 13-acetate (18) and of glucocorticoid-regulated genes in C6/ST1 rat gliomas transformed to normal phenotypic reversion (Vedoy CG and Sogayar MC, unpublished results).

DNA microarrays (DNA chips)

Currently, DNA chips constitute the most promising and revolutionary technique ever developed to study differential gene expression. The basic idea is remarkably simple and elegant consisting of the arrangement of different DNA sequences (ESTs or deoxyo-ligonucleotides) in an organized array on a small glass surface.

The two mRNA populations that are to be compared are first converted to cDNA, tagged with different fluorochromes (green and red, for example), denatured and then simultaneously hybridized to the immobilized DNA samples. Upon hybridization, the so-called DNA chip (glass plate) is scanned at the appropriate wavelengths following excitation of the fluorochromes. Comparison of the images generated by the two wavelengths allows the identification of the differentially expressed sequences.

Due to the very small area occupied by the array, the volume of the hybridization reaction can be reduced, with consequent probe concentration and high sensitivity. According to some authors (20), one in 1.5-3.0 x 10^5 molecules can be detected. Moreover, the use of glass instead of porous membranes significantly reduces the background.

Basically, there are 2 kinds of arrays, according to the nature of the DNA: a) cDNA fragments (ESTs) and b) *in situ* synthesized deoxy-oligonucleotides.

In the first category, cDNA fragments are amplified by PCR and robotically spotted onto glass slides coated with polylysine by direct contact with tweezers, capillaries or pins. This method was originally developed by Brown and colleagues (21-23). The robot construction and hybridization protocols can be found at: (http://cmgm.stanford.edu/pbrown). The great advantages of this method are its feasibility and the relatively low cost of the spotting robot (approximately US$ 25,000). In addition, since the DNA fragments are larger than the chemically synthesized oligonucleotides, higher specificity is attained in hybridization. However, a large number of cDNAs have to be available, amplified, purified and quantitated before they can be spotted. In addition, coating of the glass slide with positively charged polylysine, for instance, may alter the conformation of the DNA spotted onto the slide, de-

creasing its affinity for the DNA to be hybridized in solution. Chemically synthesized oligonucleotides can also be robotically spotted, as an alternative to cDNA fragments.

The second category comprises 2 different methods, i.e., photolithography and piezoelectric printing. In the former, photolabile protecting groups are used in oligonucleotide 3'OH terminals. In the first step, the glass slide containing the OH-bearing spacer group is illuminated through a lithographic mask, which allows de-protection of predetermined regions on the glass slide. Upon losing the photolabile group, the compound's hydroxyl group is free to react with the first type of protected deoxynucleotide. Thus, by successively varying the photolithographic masks, and subsequent reaction of free OH groups with different nucleotides, it is possible to synthesize thousands of different nucleotides of up to 30 mer at known loca-

tions on the glass slide (24). In these arrays, each oligonucleotide has an almost perfect copy physically adjacent to it, differing in only one base. This method, developed by Affymetrics (http://www.affymetrix.com), has the advantage of eliminating the need to deal with thousands of PCR products that have to be purified, quantitated and properly stored. In addition, synthesis can be directed from data bases, and therefore it is possible to direct it so as to differentiate members of the same gene family. This allows the highest density of oligonucleotides, but the photolithographic masks are very expensive and difficult to generate. In view of the costly process involved, synthesis of these arrays is limited to the industry.

Figure 2 illustrates some approaches used to generate microarrays.

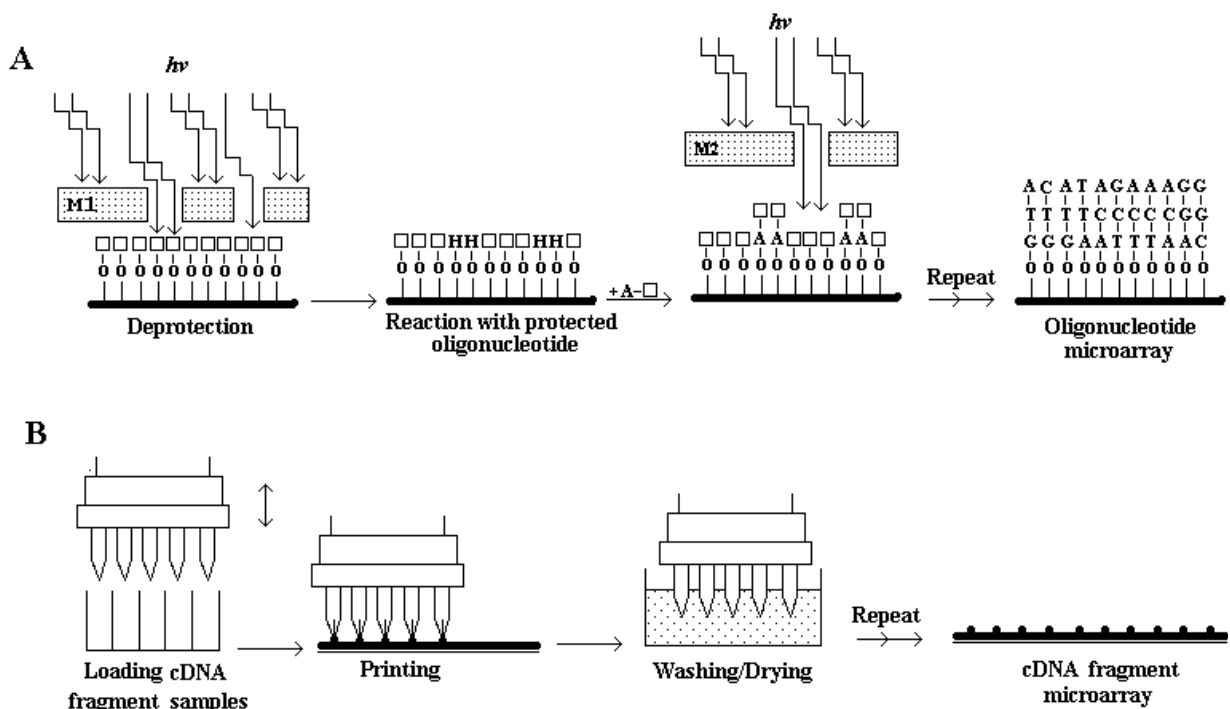Another method of *in situ* synthesis is the piezoelectric printing technique utilized by

Figure 2 - Examples of two approaches to generate microarrays. A, Photolithography: a glass slide having protected spacer groups is selectively deprotected by shining light through photolithographic masks (M1 and M2). The activated groups react with a protected base (A-□, in the example). Repeated deprotection and reaction with different masks result in high-density oligonucleotide microarrays. B, Array of cDNA fragments: the fragment samples are loaded into pins by capillary action and printed on a glass surface covered with polylysine. The pins are washed, dried and used to load more samples.

ink jet printers, in which the printing head moves along a glass surface, spotting droplets of one type of deoxynucleotide triphosphate. Upon reaction, washing and deprotection, droplets of another type of nucleotide triphosphate are added, until up to 50-mer oligonucleotides are synthesized (25). Currently, this technique is not as potent as photolithography or microarrays but it is certainly very promising.

DNA chips have been widely used. Expression of cytokines induced by phorbol ester in murine 2D6 helper T cell has been studied by photolithography (26). Significant induction of gamma interferon and alterations in IL-3, IL10, granulocyte macrophage-colony-stimulating factor (GM-CSF) and tumor necrosis factor (TNF)-alpha were found. However, as expected, no alterations were found in the expression levels of household genes like beta actin and GAPDH. Calibration experiments pointed to a dynamic range of 1:300,000 to 1:300.

Spotted saccharomyces DNA chips have been used to reveal the genes related to glucose depletion in the anaerobic to aerobic transition (21). At least two-fold induction was found for 710 genes and approximately 2-fold repression was detected for 1,030 genes. In addition, 183 genes were induced 4-fold and 203 genes were repressed 4-fold. Half of the differentially expressed genes had no known function and more than 400 had no apparent homology with known genes. A correlation of 0.87 was found when 2 different microarrays were used and differences between duplicates were lower than a factor of 2 for 95% of the genes (21).

Genes differentially expressed in *S. cerevisae* growing in minimum versus rich medium were sought using DNA chips generated by photolithography (20). In rich medium, 36 RNAs were found to be more abundant, 16 of them by a factor of 10. In minimum medium, more than 140 RNAs were found to be more abundant, by at least 5-fold. Fifty-seven of the 140 were at least 10

times more abundant. The detection specificity was estimated to be 1:150,000 to 1:300,000. Hybridization of the same RNA with 2 different microarrays resulted in more than a 2-fold difference in 14 of a total of 6,200. In 2 independent experiments, 74 RNAs showed differences greater than 2-fold and 6 (less than 0.1% of the total) showed differences of at least 3-fold.

Some companies are concentrating on achieving DNA chips containing ESTs corresponding to all genes expressed by a given organism. This would allow identification of genes that are expressed in different cell types, physiological conditions and/or treatment conditions, in this organism. DNA chips with up to 40,000 ESTs are already available.

Concluding remarks

In the few instances in which the genome coding sequences are well known, the search for differentially expressed genes is greatly facilitated. However, in spite of the efforts put into several genome projects, the genomes of most organisms have yet to be elucidated.

One major approach to gaining insight into the differentially expressed sequences is to construct cDNA libraries using differential hybridization or cDNA subtraction. The quality of these libraries has significantly improved with the introduction of cDNA fractionation and normalization by Soares and colleagues (15,16). These authors have generated a set of normalized cDNA libraries with improved representation of larger (full length) cDNAs that have been widely distributed for sequencing and mapping, constituting the integrated molecular analysis of genomes and their expression (IMAGE) consortium (27).

DNA chips constitute the method of choice when prior knowledge of expressed DNA sequences is available. Perhaps, the main advantage of DNA chips is to eliminate

the inefficient process of examining all cDNAs/mRNAs expressed in order to find those that change each time a new comparison is desired. At any rate, both DNA chips and other methods (differential hybridization, cDNA subraction, DDRT-PCR) involve confirmation and functional characterization of isolated sequences. Although still somewhat conceptual, DNA chips should provide a more versatile tool to understand the alterations of gene expression and the molecular basis of several diseases.

References

1. Cochran BH, Reffel AC & Stiles CD (1983). Molecular cloning of gene sequences regulated by platelet-derived growth factor. Cell, 33: 939-947.
2. Lau LF & Nathans D (1985). Identification of a set of genes expressed during G0/G1 transition of cultured mouse cells. EMBO Journal, 4: 3145-3151.
3. Almendral JM, Somer D, MacDonald-Bravo H, Burckhardt J, Perera J & Bravo R (1988). Complexity of the early genetic response to growth factors in mouse fibroblasts. Molecular and Cellular Biology, 8: 2140-2148.
4. el-Deiry WS (1993). WAF1, a potential mediator of p53 tumor suppressor. Cell, 75: 817-825.
5. Smith DP, Shieh BH & Zuker CS (1990). Isolation and structure of an arrestin gene from Drosophila. Proceedings of the National Academy of Sciences, USA, 87: 1003-1007.
6. Lorus JJ, Denome SA & Dunlap JC (1989). Molecular cloning of genes under control of circadian clock in Neurospora. Science, 243: 385-388.
7. Cochran BH, Zumstein P, Zullo J, Rollins B, Mercola M & Stiles CD (1987). Differential colony hybridization: Molecular cloning from a zero data base. Methods in Enzymology, 147: 64-85.
8. Valentini SR & Armelin MCS (1996). Cloning of glucocorticoid-regulated genes in C6/ST1 rat glioma phenotypic reversion. Journal of Endocrinology, 147: 11-17.
9. Liang P & Pardee AB (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science, 257: 967-971.
10. Liang P & Pardee AB (1995). Recent advances in differential display. Current Opinion in Immunology, 7: 274-280.
11. Duguid JR & Dinauer MC (1990). Library subtraction of in vitro cDNA libraries to identify differentially expressed genes in scrapie infection. Nucleic Acids Research, 18: 2789-2792.
12. Hara E, Kato T, Nakada S, Sekiya S & Oda K (1991). Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. Nucleic Acids Research, 19: 7097-7104.
13. Luqmani YA & Lymboura M (1994). Subtraction hybridization cloning of RNA amplified from different cell populations microdissected from cryostat tissue sections. Analytical Biochemistry, 222: 102-109.
14. Hubank M & Schatz DG (1994). Identification differences in mRNA expression by representational difference analysis of cDNA. Nucleic Acids Research, 22: 5640-5648.
15. Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L & Efstratiadis A (1994). Construction and characterization of a normalized cDNA library. Proceedings of the National Academy of Sciences, USA, 91: 9228-9232.
16. Bonaldo MF, Lennon G & Soares MB (1996). Normalization and subtraction: two approaches to facilitate gene discovery. Genome Research, 6: 791-806.
17. Coche T & Dewez M (1994). Reducing bias in cDNA sequence representation by molecular selection. Nucleic Acids Research, 22: 4545-4546.
18. Gurskaya NG, Diatchenko L, Chenchik A, Siebert PD, Khaspekov GL, Lukyanov KA, Vagner LL, Ermolaeva OD, Lukyanov SA & Sverdlov E (1996). Equalizing cDNA subtraction based on selective suppression of polymerase chain reaction: cloning of Jurkat cell transcripts induced by phytohemagglutinin and phorbol 12-myristate 13-acetate. Analytical Biochemistry, 240: 90-97.
19. Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA & Lukyanov SA (1995). An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Research, 23: 1087-1088.
20. Wodicka L, Dong H, Mittmann M, Ho M & Lockhart DJ (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. Nature Biotechnology, 15: 1359-1367.
21. DeRisi JL, Iyer VR & Brown PO (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, 278: 680-686.
22. Schena M, Shalon D, Davis RW & Brown PO (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270: 467-470.
23. Shalon D, Smith SJ & Brown PO (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Research, 6: 639-645.
24. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP & Fodor SPA (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proceedings of the National Academy of Sciences, USA, 91: 5022-5026.
25. Marshal A & Hodgson J (1998). DNA chips: an array of possibilities. Nature Biotechnology, 16: 27-31.
26. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H & Brown EL (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology, 14: 1675-1680.
27. Lennon G, Auffray C, Polymeropoulos M & Soares MB (1996). The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. Genomics, 33: 151-152.