Journal of Cellular Biochemistry Supplements 30/31:286–296 (1998)

High Throughput Analysis of Differential Gene Expression

John P. Carulli,[1†] Michael Artinger,[1*] Pamela M. Swain,[1] Colleen D. Root,[1] Linda Chee[1], Craig Tulig,[2] Jennifer Guerin,[1] Mark Osborne,[1] Gary Stein,[3] Jane Lian,[3] and Peter T. Lomedico[1]

[1]Department of Human Genetics, Genome Therapeutics Corporation, Waltham, Massachusetts 02154
[2]Department of Bioinformatics, Genome Therapeutics Corporation, Waltham, Massachusetts 02154
[3]Department of Cell Biology, University of Massachusetts Medical Center, Worcester, Massachusetts 01655

Abstract Elucidation of the changes in gene expression associated with biological processes is a central problem in biology. Advances in molecular and computational biology have led to the development of powerful, high-thoughput methods for the analysis of differential gene expression. These tools have opened up new opportunities in disciplines ranging from cell and developmental biology to drug development and pharmacogenomics. In this review, the attributes of five commonly used differential gene expression methods are discussed: expressed sequence tag (EST) sequencing, cDNA microarray hybridization, subtractive cloning, differential display, and serial analysis of gene expression (SAGE). The application of EST sequencing and microarray hybridization is illustrated by the discovery of novel genes associated with osteoblast differentiation. The application of subtractive cloning is presented as a tool to identify genes regulated in vivo by the transcription factor *pax-6.* These and other examples illustrate the power of genomics for discovering novel genes that are important in biology and which also represent new targets for drug development. The central theme of the review is that each of the approaches to identifying differentially expressed genes is useful, and that the experimental context and subsequent evaluation of differentially expressed genes are the critical features that determine success. J. Cell. Biochem. Suppls. 30/31:286–296, 1998. © 1998 Wiley-Liss, Inc.

Key words: EST; cDNA microarray; RDA; osteoblast differentiation; pax-6

High throughput analysis of differential gene expression is a powerful tool that can be applied to many areas in molecular cell biology, including differentiation, development, physiology, and pharmacology. In recent years, a variety of techniques have been developed to analyze differential gene expression, including comparative expressed sequence tag (EST) sequencing, differential display, PCR-based subtractive cloning, mRNA hybridization to cDNA or oligonucleotide arrays, and serial analysis of gene expression (SAGE).

A generalized paradigm for the application of differential gene expression methods is illustrated in Figure 1. In basic biology, these methods are typically used to identify genes that are critical for a developmental process, to identify genes that mediate cellular responses to a variety of chemical or physical stimuli, or to understand the molecular events effected by mutations in a gene of interest. Additional applications in biotechnology include identification of molecular markers for various disease processes, identification of potential drug targets, and pharmacogenomics: the elucidation of the molecular events associated with drug treatment. In this review we illustrate the application of these methods to a variety of biological questions, and draw particularly on our experience in EST sequencing, microarray hybridization, and subtractive cloning.

EST SEQUENCING

The concept of EST sequencing first came into public view in 1991 [Adams et al., 1991]. The basic idea is simple: create cDNA libraries from tissues of interest, pick clones randomly from these libraries, and then perform a single sequencing reaction from a large number of

Grant sponsor: NIH; Grant number: 5 R44 HS33803.

†Current address: Department of Molecular Genetics, Biogen, Inc., 14 Cambridge Center, Cambridge, MA 02142

*Correspondence to: Michael Artinger, Department of Human Genetics, Genome Therapeutics Corporation, 100 Beaver St., Waltham, MA 02154.
E-mail: michael.artinger@genomecorp.com

Received 19 August 1998; Accepted 21 August 1998

I. Cells, tissues, or animals differing in:

- Developmental stage
- Genotype
- Disease state
- Drug, hormone, growth factor treatment, etc.

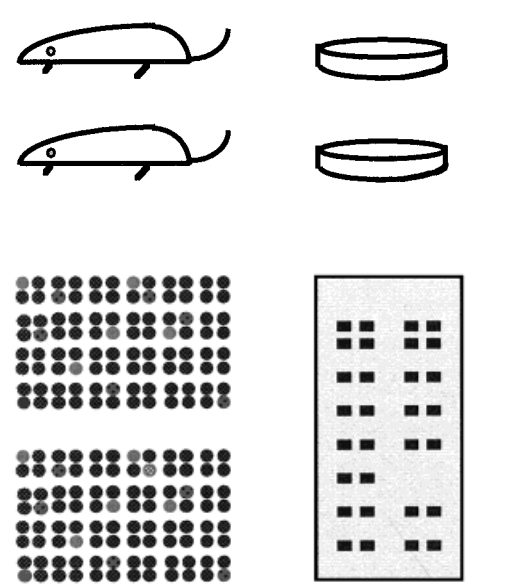
II. Identify differentially expressed genes:

- EST sequencing
- Microarray hybridization
- Subtractive Cloning
- Differential display
- SAGE

III. Analyze genes for identity or similarity to previously known genes.

IV. Perform further functional experiments.

- Expression studies
- Gene knockout
- Antisense
- Phenotype rescue



```
AAAGGAAGTAACCTTGTCCGTCCCTGTCCTCAGACAAACTGGGCAGCCTC
|||||  ||||  |||||||  |||  ||||||||||||||||||||||||||
AAAGGCAGTAGCCTTGTCTGTCTCTGTCCTCAGACAAACTGGGCAGCCTC

TCCTGTGCCTTTTTCTCCATCGCAGCTCTCTGAGTCCCAGAGGCCTGTTG
|||  |||||  ||||||||||  |||||  ||||||  ||||||||||||||
TCCCGTGCC.TTTTCTCCATCACAGCTTCCTGAGTTCCAGAGGCCTGTTG

GGTACTGGAAGCAACAGTTAAACAGGTCACTGCTATGGTGGCTGAGATTG
|||  |||||||    |||||  ||||  ||||||||||||||||||||||
GGTGCTGGAAG...CAGTTGAACAAGTCACTGCTATGGTGGCTGAGATTG
```

I. Each cell or tissue expresses tens of thousands of genes.

II. Differences in phenotype are associated with differences in expression of tens or hundreds of genes.

III. A limited number of genes will be novel in their identity or in the specific experimental setting.

IV. The hypothetical role of individual differentially expressed genes must be experimentally verifie

Fig. 1. A general scheme for the application of high throughput differential gene expression analysis. **Color plate on page 333.**

clones. Each sequencing reaction generates 300 base pairs or so of sequence that represents a unique sequence tag for a particular transcript. An EST sequencing project is technically simple to execute, since it requires only a cDNA library, automated DNA sequencing capabilities, and standard bioinformatics protocols. To generate meaningful amounts of data, however, high throughput template preparation, sequencing, and analysis protocols must be used.

EST sequencing can be accomplished using normalized or nonnormalized cDNA libraries. A normalized cDNA library is one in which each transcript is represented in more or less equal numbers [Patanjali et al., 1991; Soares et al., 1994]. The advantage of using normalized cDNA libraries is that redundant sequencing of highly expressed genes is minimized, and the potential for identification of rare transcripts is maximized [Bonaldo et al., 1996]. An advantage of nonnormalized, nonamplified libraries is that the transcript abundance of the original cell or tissue is accurately reflected in the frequency of clones in the library. Nonnormalized libraries can be used for an EST project to identify highly expressed, unknown genes and to compare the expression of highly expressed genes in different cell or tissue samples [Ji et al., 1997].

We have used EST sequencing to characterize changes in gene expression during differentiation of rat osteoblast cells in culture [reviewed by Stein and Lian, 1993]. These cells pass through three well-defined stages as they differentiate in vitro: proliferation, matrix maturation, and mineralization. Our goal was to identify new genes that are associated with and/or are critical for osteoblast differentiation, and to use these data to better understand the intrinsic and extrinsic factors that influence bone formation. To do this, we prepared nonnormalized, nonamplified cDNA libraries from osteoblast cell cultures at two critical junctures during differentiation: late proliferation (8 days in culture) and early mineralization (17 days in culture). The libraries were plated, and 5,000–6,000 clones from each library were chosen at random for 5' end sequencing. The final data set included 4,919 5' ESTs from the day-8 culture, and 4,362 5' ESTs from the day-17 cell culture. We chose to sequence from the 5' end of each clone to maximize the amount of protein-coding information in each sequence, and therefore to maximize our ability to assign function through homology to other sequences in publicly available databases.

The 9,281 DNA sequences were initially passed through an automated series of sequence analysis steps, including Blastn [Altshcul et al., 1990, 1997] searches against the Genbank and Unigene databases, and Blastx [Altshcul et al., 1990, 1997] searches against a nonredundant protein database derived from Swiss-Prot, Genbank, and Unigene (A. Caruso, personal communication). In addition to similarity searches, the ESTs were clustered using the sequence assembly program PHRAP [Gordon et al., 1998] to help identify previously undescribed genes that were present multiple times in the data set. The outcome of this analysis was a database representing quantitative transcript profiles of proliferating and differentiating osteoblasts. A total number of 2,795 cDNA clusters was identified, which represents the maximum number of genes in the database (since these are 5' ESTs, it is possible that, for long transcripts, the sequences may not overlap and therefore a single gene can be represented by more than one cluster). Approximately 75% of the genes have mammalian orthologues in Genbank or Unigene, 15% appear to have no known orthologue but do have significant similarity to known protein families, and 10% appear to have little or no similarity to any sequences in public databases at either the nucleotide or protein level.

The power of EST sequencing for gene discovery is illustrated by the presence of several moderate to highly expressed genes (0.1% or more) that were novel in the sense that they were not represented in any publicly available databases, or else they were present as ESTs that were not associated with any defined function. This illustrates the main strength of EST sequencing from nonnormalized cDNA libraries: discovery of novel genes for selected cells or tissues is possible, despite the large publicly available databases of gene sequences. However, the number of new genes identified, as well as the statistical significance of the data, is proportional to the number of clones sequenced as well as the complexity of the tissue being analyzed. In addition, the continued efforts of public and private sequencing organizations will likely identify all of the expressed genes in human and other mammalian genomes within the next few years [Adams et al., 1995; Hillier et al., 1996].

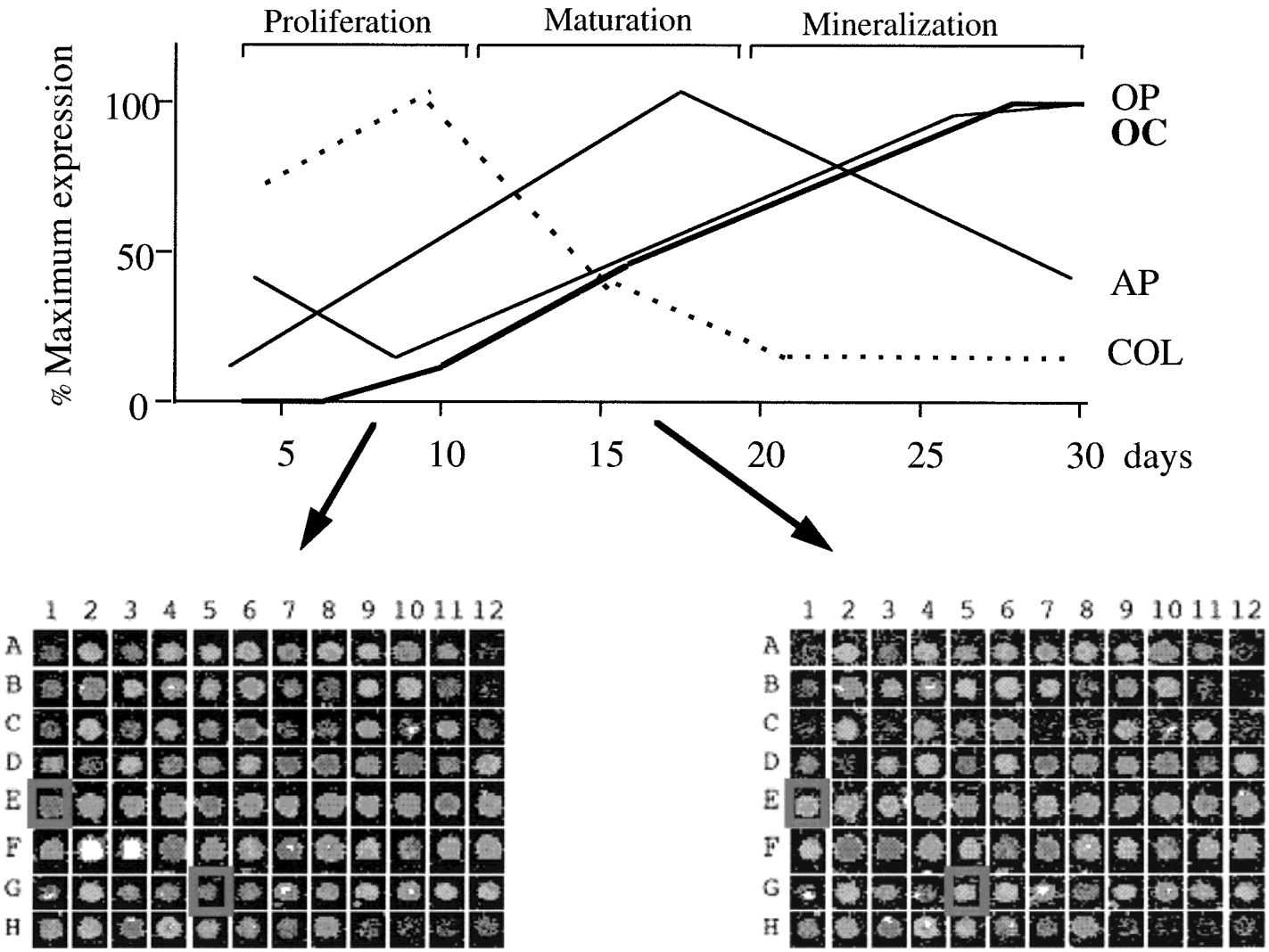


Fig. 2. Use of cDNA microarray analysis to identify changes in gene expression during osteoblast differentiation. **Top:** Changes in expression of several osteoblast markers during cellular differentiation. **Bottom:** Results of differential cDNA array hybridization of mRNA from 8-day **(left)** and 17-day **(right)** osteoblast cultures. The cDNA array was prepared using clones identified by EST sequencing from osteoblast cDNA libraries. The probes were prepared by isolating polyA+ RNA from cells after 8 days or 17 days of culture, and reverse transcribing the RNA, using an oligo-dT primer with incorporation of Cy3 or Cy5 dCTP. The probes were simultaneously hybridized to the array, and the results from each probe were analyzed by CCD camera. Here, the hybridization intenstity is represented in pseudocolor: genes which are not highly expressed are represented in blue, moderately expressed genes are in yellow, and highly expressed genes are in red. The clone at position E1 represents a previously undescribed gene that is significantly downregulated during differentiation, while the G5 clone represents a previously undescribed gene that is significantly upregulated. **Color plate on page 334.**

CDNA MICROARRAYS

To further analyze the expression of genes identified in the osteoblast EST project, we arrayed a set of 960 clones for cDNA microarray analysis [Schena et al., 1995]. The cDNAs were arrayed, in duplicate, on glass slides at a density of >1,000 clones/cm^2. Differential hybridization was then performed using RNA derived from two different cell or tissue samples. Each polyA+ RNA sample was labeled by reverse transcription, with incorporation of fluorescently labeled (either Cy3 or Cy5) dCTP. The two labeled cDNA pools, representing all of the RNAs expressed in each cell or tissue sample, were then simultaneously hybridized to the microarray. The intensity of the hybridization was read by CCD camera, and the relative expression level of each gene was represented by the intensity of the hybridization signal (Fig. 2).

A typical microarray experiment requires anywhere from 0.5–2.0 µg of polyA RNA. In our experience, the most consistent labeling results are from experiments in which the RNA is labeled in a single round of reverse transcription using an oligo dT primer. Experiments with more complex reverse transcription primers, or protocols that use PCR amplification of the cDNA, have typically resulted in inconsistent amplification of the independent messages in the cell or tissue sample (Root and Carulli, unpublished results). However, linear amplification protocols have been developed that appear to amplify all transcripts uniformly [Lockhart et al., 1996].

Figure 2 shows 96 of the 960 clones that were analyzed in an experiment using RNA from the day-8 and day-17 osteoblast cell cultures described above. The same 96 clones were analyzed with the day-8 probe (Fig. 2, lower left) and the day-17 probe (Fig. 2, lower right). A number of clones showed significant up- or downregulation in this experiment. For example, the clone at position E1 was significantly downregulated during differentiation. This gene encodes a novel protein with *sushi* repeats that are present in a number of mammalian genes [Meindl et al., 1995]. The clone at position G5 corresponds to a gene that is significantly upregulated during osteoblast differentiation. This is a previously undescribed gene that is represented in public databases only as an EST. The G5 clone illustrates the much higher sensitivity of microarray hybridization relative to EST sequencing: the clone is represented only once in the two cDNA libraries that were sequenced, but is clearly differentially expressed, as illustrated by the microarray results.

A significant advantage of cDNA microarray analysis is the ability to analyze the same set of genes under a range of experimental conditions. For example, Heller et al. [1997] created an array of 96 genes known to be involved in inflammatory processes. To identify genes specifically involved in rheumatoid arthritis, they probed the array with RNA from cultured macrophages, chondrocytes, and synoviocytes, as well as arthritic tissue samples. These experiments demonstrated for the first time the involvement of several genes, including interleukin 3 and Groα, in rheumatoid arthritis. In another example, Gray et al. [1998] used an oligonucleotide array (described in more detail below) to monitor the response of virtually all of the genes in the yeast genome to a variety of protein kinase inhibitors. They were able to identify genes that responded uniquely to a specific compound, as well as genes that responded similarly to a range of compounds. The ability to perform multiple assays on the same array provides a powerful approach to streamlining the search for gene(s) with specific characteristics of interest.

In addition to arrays of cDNA clones, arrays of oligonucleotides are also used to study differential gene expression [Lockhart et al., 1996]. In an oligonucleotide array, the genes of interest are represented by a series of 20 nucleotide oligomers that are unique to each gene. Labeled cDNA for each sample is prepared as described above, and hybridization signals are detected from specific sets of oligos that represent different genes. Potential advantages of the oligonucleotide array include enhanced specificity and sensitivity through the parallel analysis of "perfect match" oligos and "mismatch" oligos for each gene [Lockhart et al., 1996]. The hybridization conditions can be adjusted to distinguish a perfect heteroduplex from a single base mismatch, thus allowing subtraction of nonspecific hybridization signals from specific hybridization signals. A disadvantage of oligonucleotide arrays relative to cDNA arrays is the limitation of the technology to genes of known sequence. This limitation is likely to disappear within the next several years, when the sequences of all

human genes and the genes of many common experimental organisms are determined.

While there are many advantages to the cDNA or oligonucleotide array approach to analyzing differential gene expression, disadvantages include the requirement of a good set of clones or oligonucleotides for known genes to array, and the relatively large amount of RNA that is required to prepare the probes. The large RNA requirement can make it difficult to analyze small clinical samples.

SUBTRACTIVE CLONING

Subtractive cloning methods have been in use for many years, but newer methods based on PCR are rapid and easy to execute, and can be used with minute amounts of starting material [Hubank and Schatz, 1994; Diatchenko et al., 1996]. Subtractive cloning offers an inexpensive and flexible alternative to EST sequencing and cDNA array hybridization, and can be performed in any laboratory equipped with basic molecular biology and bioinformatics tools. We have used subtractive cloning to address a number of biological questions, and we illustrate the utility of this method with an example of our effort to identify genes regulated by a specific transcription factor by characterizing differences in gene expression in normal and mutant mouse tissue.

To analyze differences in gene expression between normal and mutant tissue, we used a PCR-based subtractive cloning method adapted from representational difference analysis (RDA), which we refer to as RT-RDA (reverse transcription RDA) [Lisitsyn et al., 1993; Hubank and Schatz, 1994]. In this approach, double-stranded cDNA is created from the two-cell or tissue populations of interest, linkers are ligated to the ends of the cDNA fragments, and the cDNA pools are then amplified by PCR. The cDNA pool from which unique clones are desired is designated the "tester," and the cDNA pool that is used to subtract away shared sequences is designated as the "driver." Following initial PCR amplification, the linkers are removed from both cDNA pools, and unique linkers are ligated to the tester sample. The tester is then hybridized to a vast excess of driver DNA, and sequences that are unique to the tester cDNA pool are amplified by PCR.

We were specifically interested in identifying genes that are directly or indirectly regulated by the transcription factor *Pax-6*. *Pax-6* was first identified as the gene responsible for the autosomal-dominant phenotypes *aniridia* in humans and *small eye* (*sey*) in mice [reviewed by Hanson and Van Heyningen, 1995]. In the homozygous state, null mutations in pax-6 affect the development of the eyes, nose, forebrain, neural tube, and pancreas [Glaser et al., 1992; Artinger and Carulli, unpublished results]. To identify genes regulated by pax-6, we used RT-RDA to clone genes that were present in wild-type mouse tissues, but absent from tissue in *sey* homozygous mice (Fig. 3). Following three rounds of subtraction, the difference products were cloned and 274 clones were chosen for single-pass sequencing. Of these clones, 181 were derived from a total of 20 previously described genes, while 93 were derived from ESTs or novel genes. Among the previously described genes, approximately one third are known to be involved in neural development. The relationship between pax-6 and the genes cloned by RDA is being further explored by a combination of genetic, molecular, and embryological studies.

Several other investigators have also used RT-RDA to identify genes downstream from transcription factors [Buckbinder et al., 1994; Iwama et al., 1998]. To bias the selection for genes that are directly rather than indirectly regulated by the transcription factor, the transcription factor can be placed under the control of an inducible promoter and transfected into cultured cells, and RT-RDA can be performed using the induced cell population as the tester and uninduced cells as the driver [Buckbinder et al., 1994]. Another modification that has been introduced to simplify the difference products from RT-RDA is the introduction of specific genes or tissues into the driver population to minimize the recovery of previously described or contaminating cDNA sequences [Iwama et al., 1998]. In addition to analyzing the effects of mutations or misexpression of genes, RT-RDA has been used to clone genes that are involved in normal development [Wada et al., 1997] as well as genes whose expression changes in response to stimuli such as light [Morris et al., 1998].

The primary limitation of RT-RDA and similar methods is that they are not always comprehensive. The cDNAs identified are typically those that differ significantly in expression level between the cell populations, and subtle quantitative differences are often missed. In addition,

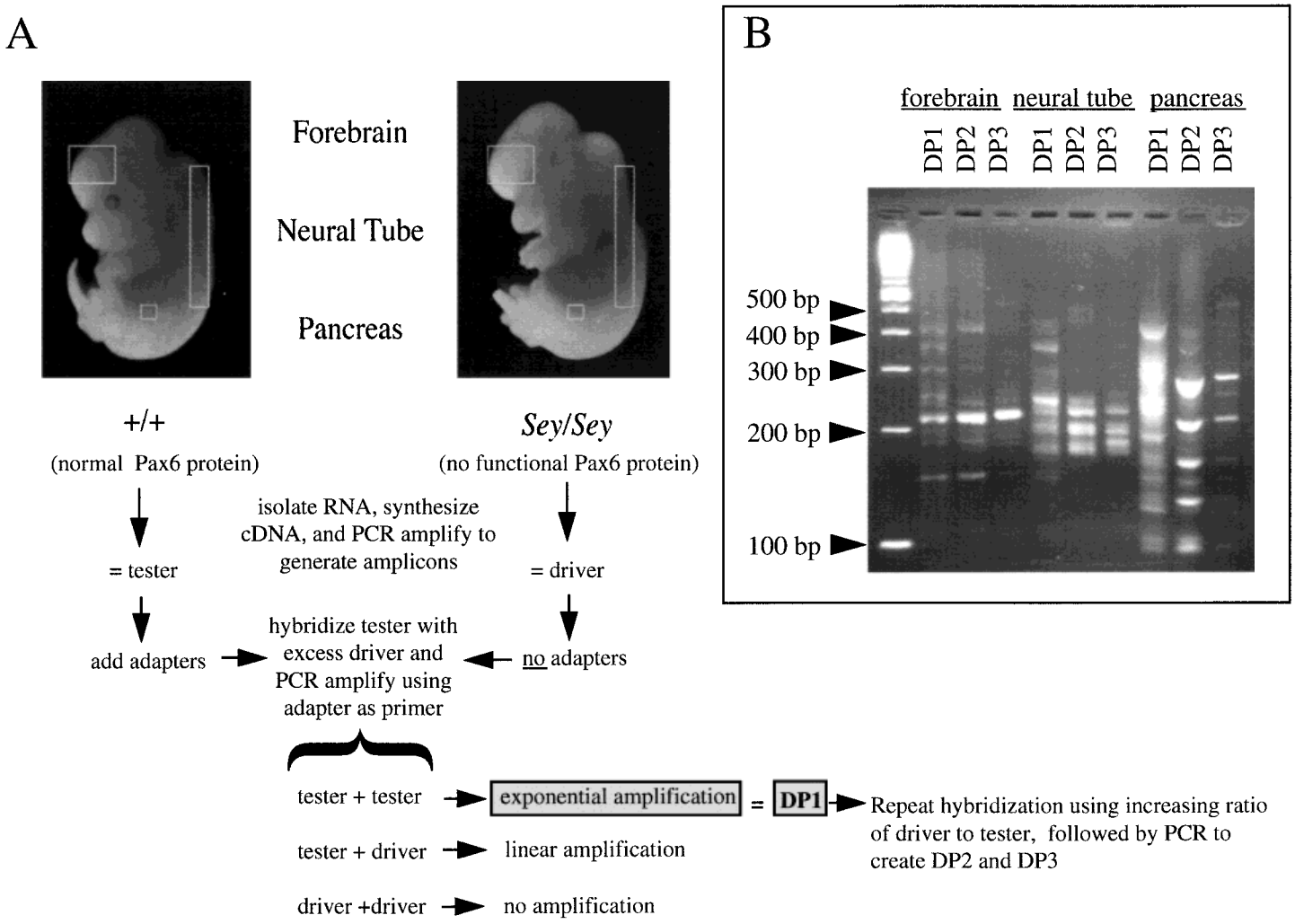


Fig. 3. Representational difference analysis of normal and mutant mouse tissue. **A:** The forebrain, neural tube, and pancreas were dissected from wild-type mice and homozygous small eye (*sey/sey*) mice which lack functional pax-6 protein. RT-RDA was performed using RNA from the wild-type tissue as tester, and RNA from the *sey/sey* mice as driver. **B:** Agarose gel analysis of the difference products after one (DP1), two (DP2), and three (DP3) rounds of RDA. Note that the cDNA pools become progressively less complex after multiple rounds of RDA. **Color plate on page 335.**

each experiment is a pairwise comparison, and since the subtractions are based on a series of sensitive biochemical reactions it is difficult to directly compare a series of RNA samples.

DIFFERENTIAL DISPLAY

Another PCR-based differential cloning method that is extremely popular is differential display or RNA fingerprinting [Liang and Pardee, 1992; Welsh et al., 1992]. In classical differential display, reverse transcription is primed with either an oligo-dT or an arbitrary primer, then an arbitrary primer (10 bases is a common length) is used in conjunction with the reverse transcription primer to amplify cDNA fragments, and the cDNA fragments are separated on a polyacrylamide gel. Differences in gene expression are visualized by the presence or absence of bands on the gel, and quantitative differences in gene expression are identified by differences in the intensity of bands. Adaptation of differential display methods for fluorescent DNA sequencing machines has enhanced the ability to quantify differences in gene expression [Kato, 1995]. Differential display is relatively simple to execute, and is efficient for analyzing small amounts of RNA. As little as 5 or 10 ng of total RNA can be used to perform the experiments, and many samples can be analyzed on a single gel. Dozens of genes involved in a number of critical biological processes have been cloned using this approach. Some recent examples include: PTI-1, an oncogene associated with prostate cancer [Shen et al., 1997]; Smad6 and Smad7, two novel MAD family members that mediate the response to mechanical stress in vascular endothelium [Topper et al., 1997]; and DD7A5–7, a novel seven-transmembrane hormone receptor involved in liver hematopoiesis [Lin et al., 1997].

A limitation of the classical differential display approach is that false-positive results are often generated during PCR, or in the process of cloning the differentially expressed PCR products. A variety of methods have been developed to discriminate true from false positives, but these typically rely on the availability of relatively large amounts of RNA. A modification of differential display based on analysis of 3' end restriction fragments has been developed [Kato, 1995; Prashar and Weissman, 1996], and is claimed to result in fewer false-positive signals. In this method, double-stranded cDNA is prepared and digested with a restriction enzyme that has a four-base recognition site. Linkers are then ligated to the restriction fragments, and the entire pool of transcripts is amplified by PCR. Differences in gene expression are visualized by gel electrophoresis of the 3' end fragments. This approach offers several advantages relative to classical differential display: by using a series of restriction enzymes, every gene in the cell can be analyzed; furthermore, the migration of the bands in the gel is determined by the location of the 3'-most restriction site for the enzyme that is used, thus allowing the identification of known genes in the sample simply be measuring the size of the restriction fragment.

SERIAL ANALYSIS OF GENE EXPRESSION

The last of the methods we will address is serial analysis of gene expression (SAGE). SAGE is a DNA sequence-based method that is essentially an accelerated version of EST sequencing [Velculescu et al., 1995]. In this method, a unique sequence tag of 13 or more bases is generated for each transcript in the cell or tissue of interest. This is accomplished by preparing double-stranded cDNA, digesting it with a restriction enzyme that has a four-base recognition site, and then ligating linkers and amplifying the cDNA pool. The unique feature of this method is that the linkers encode a recognition site for a type II restriction enzyme (such as *Bsm*I). These enzymes digest DNA at a site 20 nucleotides away from the recognition site. When the cDNA pools are digested with this second enzyme, the result is a 13–20-base-pair cDNA fragment that is uniquely defined by the original four-base cutter and the adjacent DNA sequence. These sequence tags are then ligated in a defined series of steps. The sum of all of these steps is a library of clones where each clone includes short, unique tags for 20 or more genes.

Transcript profiles are created by sequencing each SAGE library. Since each sequencing reaction yields information for 20 or more genes, it is possible to generate data points for tens of thousands of transcripts in a modest sequencing effort. The relative abundance of each gene is determined by counting or clustering sequence tags. For most genes this short sequence tag is sufficient to provide a unique identifier. For known genes, the identity can be determined by standard database searches. For previously undescribed genes, the SAGE tag

can be used to obtain a cDNA clone by PCR or hybridization-based methods.

The advantages of SAGE over many other methods include the high throughput that can be achieved, and the ability to accumulate and compare SAGE tag data from a variety of samples. The disadvantages are related to the technical difficulty in generating good SAGE libraries and in analyzing the data. Preparation of a SAGE library requires up to 5.0 µg of high-quality polyA+ RNA, and the quality of the sequence tags is dependent on a series of biochemical reactions: any inefficiencies, mispriming, or incomplete reactions in the cDNA synthesis or restriction digestion steps can result in artifacts that are very misleading. In addition, highly specialized bioinformatics tools are required to analyze SAGE data [Velculescu et al., 1995]. For example, the unique sequence tags for each gene must be extracted from complex sequences of the SAGE library. In addition, customized target databases must be created that include only the 3' end restriction fragments for each four-base cutter used to generate the SAGE libraries [Velculescu et al., 1995].

DISCUSSION

A variety of methods for high throughput analysis of differential gene expression have been developed over the past several years. If these methods are used properly, they offer the opportunity to understand biological processes at a level of molecular detail that was not possible even a few years ago. However, the high throughput nature of these experiments is a double-edged sword: if an experiment is poorly designed, or if the biological materials are compromised, the result is a large body of data that is difficult and time-consuming to analyze. In addition, some approaches are better suited than others for addressing specific biological questions. Table 1 lists some of the attributes of each method discussed in this review, and can be used as a set of guidelines to choose the method that best matches the technical capabili-

TABLE I. Attributes of Five Different Methods for High Throughput Analysis of Differential Gene Expression

	Minimum RNA requirements	Throughput	Sequencing requirements	Cloning requirements	Bioinformatics requirements
EST sequencing	1.0–5.0 µg polyA RNA	Low	High	Full-length cloning may be required for novel genes of interest	Target databases: standard Search protocols: standard Volume: high
Microarray hybridization	1.0 µg or more polyA RNA	High	Low	Full-length cloning may be required for novel genes of interest	Image analysis required Target databases: standard Search protocols: standard Volume: low
RT-RDA	10–100 ng polyA RNA	Medium	Low	Full-length cloning required for novel genes of interest	Target databases: standard Search protocols: standard Volume: low
Differential display	10–100 ng polyA RNA	High	Medium	Full-length cloning required for novel genes of interest	Target databases: standard Search protocols: standard Volume: low
SAGE	1.0–5.0 µg polyA RNA	High	High	Full-length cloning required for novel genes of interest	Target databases: specialized Search protocols: specialized Volume: high

ties of the laboratory and that also is suited for the biological samples available.

As with any experiment, the most important criterion for utilizing high throughput methods for differential gene expression is the hypothesis being tested. If one takes any two samples of cells or tissue, differences in gene expression can be identified. The problem is to identify those genes that are critical for the process at hand, whether that process is normal differentiation of cells or tissue, or if it is response of cells or tissue to a specific stimulus. The experiment should be as simple as possible, and the cells or tissues analyzed should be as similar as possible, so that only the genes involved in the critical step(s) are identified during the analysis. It is also extremely important to avoid sample contamination in differential gene expression studies. Tissue samples should be carefully dissected and should be as fresh as possible. The presence of highly expressed transcripts in a small amount of contaminating tissue can generate signals on a microarray or bands on a differential display gel that have nothing to do with the question at hand.

Even in carefully designed experiments, a frequent problem in differential gene expression studies is that too many genes are identified. If several hundred genes differ between the samples, how does one choose the most important ones to follow up on? A number of solutions exist for this problem. One is to make the best use of available bioinformatics tools. In our experience with microarray analysis, more than 6% (60 of 960) of the genes in our array were downregulated during osteoblast differentiation. However, if the genes encoding ribosomal proteins and protein synthesis-related proteins were eliminated from consideration, then the number of genes to consider was fewer than 10 in total. Another approach to narrowing the number of genes to follow up on is to analyze multiple samples or experimental treatments, and follow up only those genes that behave consistently in all of the samples. For example, if the goal is to identify genes that are important for progression of a particular cancer, then one should compare as many tumor samples as possible to the normal controls, and focus further energy only on those genes that are consistently up- or downregulated.

A number of methods for identifying differentially expressed genes are now available, and they are rapidly becoming standard tools for developmental biologists, cell biologists, geneticists, and drug developers. While these methods are very powerful, it is important to remember that identification of differentially expressed genes is just one tool for understanding a biological process. When a minimal set of interesting genes is identified, it is important to test the hypothesis that the gene is truly involved in the process of interest by using the appropriate cell or animal model.

ACKNOWLEDGMENTS

The authors thank their colleagues at Genome Therapeutics for maintaining a DNA sequencing facility suitable for collecting EST data. The microarray experiments cited in this paper were performed in collaboration with Synteni, Inc. (Fremont, CA). The pax-6 data were obtained with the support of NIH grant 5 R44 HS33803 to J.P.C.

REFERENCES

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao CR, Merril CR, Wu A, Olde B, Moreno RF, et al. (1991): Complementary DNA sequencing: expressed sequence tags and the human genome project. Science 252:1651–1656.

Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, et al. (1995): Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature [Suppl] 377:3–173.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990): Basic local alignment search tool. J Mol Biol 215:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997): Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Bonaldo MF, Lennon G, Soares MB. (1996): Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res 6:791–806.

Buckbinder L, Talbott R, Seizinger BR, Kley N. (1994): Gene regulation by temperature-sensitive p53 mutants: identification of p53 response genes. Proc Natl Acad Sci USA 91:10640–10644.

Diatchenko L, Lau Y-FC, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert PD. (1996): Suppression subtractive hybridization: a method for generating differntially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci USA 93:6025–6030.

Glaser T, Walton DS, Maas RL. (1992): Genomic structure, evolutionary conservation and *aniridia* mutations in the human *Pax-6* gene. Nat Genet 2:232–239.

Gordon D, Abajian C, Green P. (1998): Consed: a graphical tool for sequence finishing. Genome Res 8:195–202.

Gray NS, Wodicka L, Thunnissen AWH, Norman TC, Kwon S, Espinoza FH, Morgan DO, Barnes G, LeClerc S, Meijer L, Kim S-H, Lockhart DJ, Schultz PG. (1998): Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. Science 281:533–538.

Hanson I, Van Heyningen V. (1995): Pax-6: more than meets the eye. Trends Genet 11:268–272.

Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE, Davis RW. (1997): Discovery and analysis of inflammatory disease related genes using cDNA microarrays. Proc Natl Acad Sci USA 94:2150–2155.

Hillier L, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, Dubuque T, Favello A, Gish W, Hawkins M, Hultman M, Kucaba T, Lacy M, Le M, Le N, Mardis E, Moore B, Morris M, Parsons J, Prange C, Rifkin L, Rohlfing T, Schellenberg K, Soares MB, Tan F, Thierry-Meg J, Trevaskis E, Underwood K, Wohldman P, Waterston R, Wilson R, Marra M. (1996): Generation and analysis of 280,000 human expressed sequence tags. Genome Res 6:807–828.

Hubank M, Schatz DG. (1994): Identifying differences in mRNA expression by representational difference analysis of cDNA. Nucleic Acids Res 22:5640–5648.

Iwama A, Zhang P, Darlington GJ, McKercher SR, Maki R, Tenen DG. (1998): Use of RDA analysis of knockout mice to identify myeloid genes regulated in vivo by PU.1 and C/EBPα. Nucleic Acids Res 15:3034–3043.

Ji H, Liu YE, Jia T, Wang M, Liu J, Xiao G, Joseph BK, Rosen C, Shi YE. (1997): Identification of a breast cancer-specific gene, BCSG1, by direct differential cDNA sequencing. Cancer Res 57:759–764.

Kato K. (1995): Description of the entire mRNA population by a 3' end cDNA fragment generated by class IIS restriction enzymes. Nucleic Acids Res 18:3685–3690.

Liang P, Pardee AB. (1992): Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257:967–970.

Lin HH, Stubbs LJ, Mucenski ML. (1997): Identification and characterization of a seven transmembrane hormone receptor using differential display. Genomics 41:301–308.

Lisitsyn N, Lysitsyn N, Wigler M. (1993): Cloning the differences between two complex genomes. Science 259: 946–951.

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. Nat Biotechnol 14:1675–1680.

Meindl A, Carvalho MR, Herrmann K, Lorenz B, Achatz H, Lorenz B, Apfelstedt-Sylla E, Wittwer B, Ross M, Meitinger T. (1995): A gene (SRPX) encoding a *sushi*-repeat-containing protein is deleted in patients with X-linked retinitis pigmentosa. Hum Mol Genet 4:2339–2346.

Morris ME, Viswanathan N, Kuhlman S, Davis FC, Weitz CJ. (1998): A screen for genes induced in the suprachiasmatic nucleus by light. Science 279:1544–1547.

Patanjali SR, Parimoo S, Weissman SM. (1991): Construction of a uniform-abundance (normalized) cDNA library. Proc Natl Acad Sci USA 88:1943–1947.

Prashar Y, Weissman SM. (1996): Analysis of differential gene expression by display of 3' end fragments. Proc Natl Acad Sci USA 93:659–663.

Schena M, Shalon D, Davis RW, Brown PO. (1995): Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470.

Shen R, Su ZZ, Olsson CA, Fisher PB. (1997): Identification of the human prostatic carcinoma oncogene PTI-1 by rapid expression cloning and differential RNA display. Proc Natl Acad Sci USA 92:6778–6782.

Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstradiatis A. (1994): Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci USA 91: 9228–9232.

Stein GD, Lian JB. (1993): Molecular mechanisms mediating proliferation-differentiation interrelationships during progressive development of the osteoblast phenotype Endocrine Rev 14(4):424–442.

Topper JN, Cai J, Qiu Y, Anderson KR, Xu YY, Deeds JD, Feeley R, Gimeno CJ, Woold EA, Tayber O, Mays GG, Sampson BA, Schoen FJ, Gimbrone MA, Falb D. (1997): Vascular MADs: two novel MAD-related genes selectively inducible by flow in vascular endothelium. Proc Natl Acad Sci USA 94:9314–9319.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. (1995): Serial analysis of gene expression. Science 270:484–488.

Wada J, Kumar A, Ota K, Wallner EI, Batlle DC, Kanwar YS. (1997): Representational difference analysis of cDNA of genes expressed in embryonic kidney. Kidney Int 51: 1629–1638.

Welsh J, Chada K, Dalal SS, Cheng R, McClelland M. (1992): Arbitrarily primed PCR fingerprinting of RNA. Nuclei Acids Res 20:4965–4970.