

TIMELINE

Analysing differential gene expression in cancer

Peng Liang and Arthur B. Pardee

Analysis of messenger RNA and proteins is widely used to compare patterns of gene expression between cells or tissues of different kinds and under different conditions; for example, between normal and cancer cells. The goal of the individuals who are developing these methods has been to enable faster, simpler, more sensitive and systematic analyses, and over the past few decades techniques have become increasingly more sophisticated. This timeline article reviews the evolution of these technologies as well as strategies for identifying differentially expressed genes in normal and cancer cells. It also highlights their use for the search for target genes of the tumour suppressor p53.

In the past 50 years, we have made remarkable progress in understanding genetics. Although the central genetic dogma is defined as the flow of genetic information from DNA to messenger RNA and then to protein, the complete sequencing of the 3-billion base-pair human genome has shed little light so far on precisely how such unidirectional information flow from tens of thousands of genes is programmed in a cell. Decoding such information from the human genome is likely to be even more challenging and time-consuming than the sequencing of the genome. Although classical genetics has been a powerful tool for dissecting molecular diseases that are affected by the gain or loss of function of a protein encoded by a single gene, such a strategy has proved to be less fruitful for understanding diseases such as

cancer¹ that are controlled by many genes. Compounding the complexity is the fact that many of the so-called oncogenes or tumour-suppressor genes are signalling molecules themselves, each of which functions to control the expression of a subset of downstream genes^{2,3}. So, the analysis of differential gene expression — known as expression genetics or functional genomics — has become one of the most widely used strategies for discovering and understanding the molecular circuitry underlying cancer. Throughout this review, we provide examples of gene discovery in cancer research made by studying differential gene expression mainly at the mRNA level, with a special emphasis on genes that are regulated by the tumour suppressor p53.

Over the past two decades, several methods have been developed to allow comparative studies of gene expression between normal and cancer cells. Starting with simple approaches that used gel electrophoresis to compare protein expression, methods that focus on mRNA analysis have evolved and become increasingly sophisticated, as a result of the inventions of recombinant DNA, DNA sequencing and PCR technologies (see TIMELINE)⁴. To better understand the principles behind some of the main methods, we can group them into six categories.

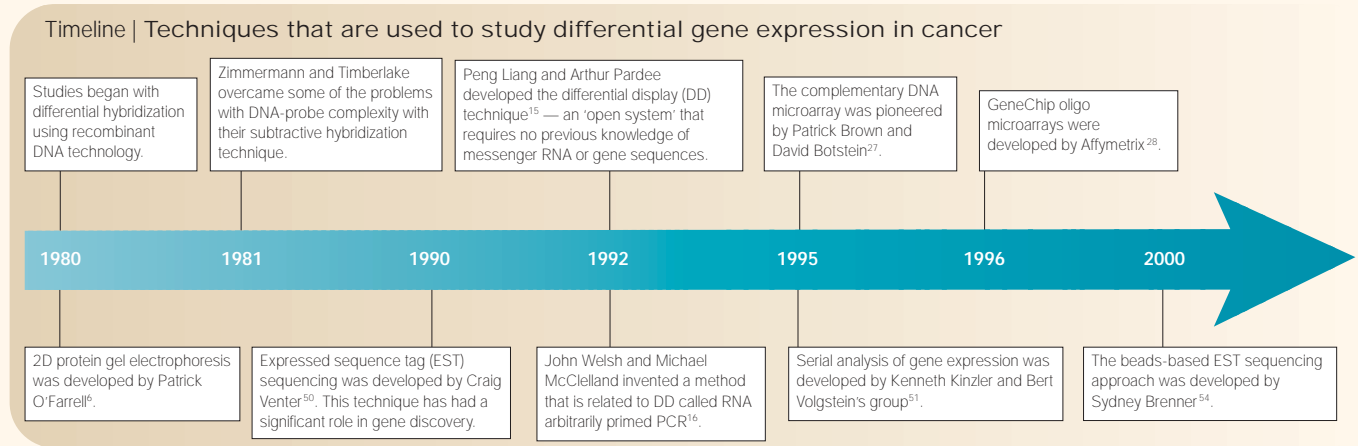
Protein gel electrophoresis
Perhaps the earliest and arguably the most successful example of studying differential gene expression in cancer was the discovery of the p53 tumour-suppressor protein in the late 1970s. The protein was found to

be overexpressed on a one-dimensional protein gel when normal cells were compared with those that were infected with simian virus 40 (SV40) DNA tumour virus⁵. It turned out that the increase in the expression of the p53 protein was caused by the stabilization of the protein through its interaction with SV40 large T antigen.

The development, by Patrick O'Farrell, of two-dimensional (2D) protein gel electrophoresis, which separates proteins by both size and charge, allowed a more complete visualization of cellular protein expression⁶. Later, Robert Croy and Arthur Pardee found that a transformed cell differs from its normal parental cell in expression of only a few of the more than 1,000 proteins that are detected on 2D protein gels. One of these proteins, named p68, was shown to be expressed at a much higher level in carcinogen-transformed murine fibroblasts⁷. This result reinforced the notion that subtle changes in gene expression might hold the key to the understanding of cancer. However, frustration often followed such studies, owing to the inability to recover sufficient amounts of the differentially expressed protein species for further molecular characterization. Also, it became clear that this method was not sensitive enough, detecting only about 2,000 of the estimated 10,000 or more different proteins that are expressed in a cell. Newer techniques for the analysis of protein expression — collectively known as proteomics — have been developed in recent years and are mainly powered by the use of mass spectrometry to greatly improve sensitivity and allow the characterization of small quantities of protein.

Differential hybridization

With the advent of recombinant DNA technology in the late 1970s, studies of comparative gene expression quickly shifted from looking at proteins to the analysis of mRNA expression using complementary DNA. The earliest approach was differential hybridization, in which the pair of mRNA samples to be compared (for example, from normal



versus cancer cells) were radioactively labelled as cDNA probes with ³²P by reverse transcription with oligo-dT primers that anneal to the polyadenylic chains (polyA tails) present at the 3' termini of all eukaryotic mRNAs. The resulting two cDNA probes were then differentially hybridized to duplicate filters, which had on them tens of thousands of plaques from a phage cDNA library⁸. Comparison of the hybridization pattern to cDNA-containing phage plaques between two mRNA probes allowed the identification of genes that were uniquely expressed in one but not the other RNA sample (FIG. 1). Although this strategy has implicated several differentially expressed genes that are involved in the hormone responsiveness of human breast cancer cells⁹ and that are overexpressed during infection by human T-cell leukaemia/lymphoma virus¹⁰, it was soon realized that such a 'reverse northern' approach of using complex cDNA probes (BOX 1) would not be able to detect most genes, which are expressed at a low level⁸. As a result, differential screening quickly gave way to hybridization methods that use cDNA probes with reduced complexity after a 'subtraction' process.

Subtractive hybridization

Realizing the problems with cDNA-probe complexity for differential hybridization, in the early 1980s C. Zimmermann and colleagues devised an ingenious approach known as subtractive hybridization to enrich for cDNA probes that represent mRNAs that are uniquely expressed in one cell but not the other¹¹. This method removed most of the cDNAs that represented the genes that are commonly expressed in both cells being compared, and left behind only single-stranded cDNAs that represented a few differentially

expressed genes (FIG. 2). The resulting cDNA probes with reduced complexity were then individually checked by northern blot for confirmation. The discovery of T-cell receptors in the mid 1980s by Mark Davis and colleagues — when they compared the differences in mRNA expression between T and B cells using such strategies — is one of the best examples of gene discovery through the analysis of differential gene expression¹². Dr. Davis recalled in his acceptance speech for his General Motor Cancer Research Award that he would not have achieved the feat had he relied on differential hybridization of dot blots with complex cDNA probes¹³.

This exciting discovery fuelled a flood of biomedical research using gene-expression analysis as a basic strategy to understand a wide variety of biological systems. The discovery of the cyclin-dependent kinase inhibitor **WAF1** (also known as p21) as a target gene of p53 by Bert Vogelstein's group provided another great example of the power of this conceptually simple and logical method¹⁴. Since then, several PCR-based subtractive-hybridization strategies have been developed, including representational difference analysis (RDA) and suppression PCR, which allow a smaller amount of mRNA samples to be analysed.

Differential display

To speed up the identification of differentially expressed genes, we developed differential display (DD) in the early 1990s, with the aim of overcoming the limitations of previous methods¹⁵. John Welsh and Michael McClelland invented a related method known as RNA arbitrarily primed PCR (RAP-PCR) using only primers of random sequence¹⁶. A sensitive method was required so that it could be applied to systems for which scarce biological samples

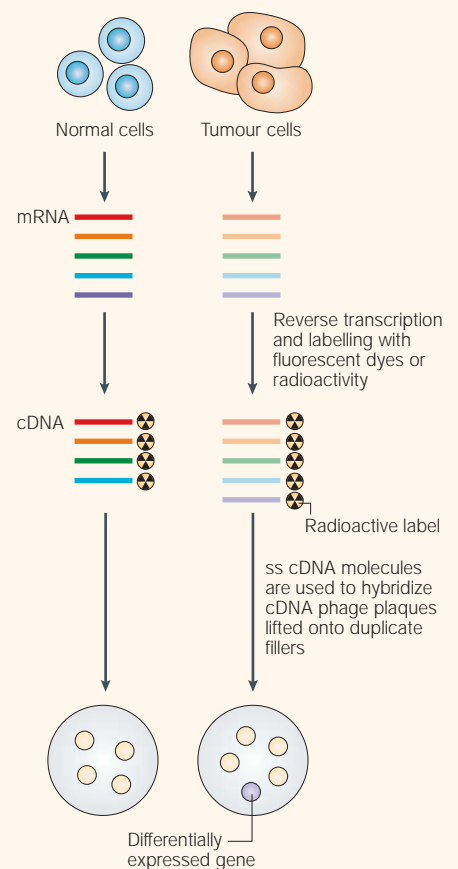


Figure 1 | **Differential hybridization.**

All messenger RNAs that are expressed in the cells being compared are first labelled either radioactively or fluorescently by copying them into single-stranded complementary DNAs (ss cDNAs) using reverse transcriptase. The resulting complex cDNA probes (BOX 1) are then used to hybridize to the cDNA phage plaques that are lifted onto duplicate nylon membranes. The key difference between differential screening and microarrays (FIG. 4) is that phage cDNA plaques contain some degree of redundancy in sequences (for example, a highly expressed gene might be represented by more than one plaque on a filter), whereas microarrays usually strive to minimize such redundancy.

are available, and by which all mRNAs, whether scarce or abundant, can be represented. Also, the method needed to be systematic, so that a complete search of all the expressed genes in a cell was possible. Based on these crucial criteria, differential display was developed by integrating two of the most simple, powerful and commonly used molecular biological methods; namely, PCR and DNA sequencing by gel electrophoresis¹⁵. In essence, DD works by systematically amplifying the 3' termini of eukaryotic mRNA by reverse transcription-PCR using one of the three anchored oligo-dT primers (that is, the run of Ts ending with a C, G or A) in combination with a set of short primers of arbitrary sequences (FIG. 3). Based on the finding that each arbitrary primer would recognize its corresponding mRNA targets with a minimum of seven matching bases, mathematical models have been proposed to predict the relation between the number of arbitrary primers and the coverage of expressed genes in any given eukaryotic cell¹⁷. One of the main advantages of DD is its technical simplicity and accessibility, which makes it and several related methods the most widely used approaches for studying differential gene expression¹⁸. Unlike microarrays, DD does not require any previous knowledge of mRNA or gene sequences, making it an 'open' system that is applicable to any eukaryotic organism. Many oncogene targets have been identified by DD, including genes that are regulated by **RAS**^{19,20}, **v-REL**²¹ and **ERBB**²². One of the RAS target genes was shown to be a new cytokine, now known as interleukin-24 (**IL-24**) (REF. 23). Many important target genes of the p53 tumour suppressor have also been discovered by DD, which will be discussed in more detail below.

Initially, DD suffered from a high rate of 'false positives' (also known as 'noise' in reference to microarrays), but technical improvements as well as care in experimental design²⁴ have greatly reduced the number of false positives to allow truly differentially expressed genes to be steadily identified¹⁸. Despite the great impact of the method on biomedical research, most work involving DD has used limited PCR reactions to take a 'shot-gun' approach that identifies only a few genes at a time, giving DD a low-tech, low-precision and low-throughput image. It is only recently that efforts have been made for the automation of DD technology with fluorescent digital data acquisition and analysis to increase its throughput and accuracy for systematic gene-expression analysis^{25,26}.

Box 1 | cDNA probe complexity

This term is used to specify the number of complementary DNA species (or messenger RNA species) and their relative concentrations within a cDNA probe. For differential screening and microarrays, a cDNA probe is made by reverse transcription of all the mRNAs that are expressed in a cell or within a tissue specimen using an oligo-dT primer, which targets the polyA tails that are present in most eukaryotic mRNAs. In fact, such a cDNA probe is so complex that it consists of as many as 10,000 different species, each ranging from a few to thousands of copies per cell — this approach is used in both differential hybridization (FIG. 1) and microarrays (FIG. 4). For subtractive hybridization, the complexity of a subtracted cDNA probe is greatly reduced by removing most of the commonly expressed mRNAs for a pair of RNA samples being compared (FIG. 2), whereas differential display displays 50–100 mRNAs at a time using different primer pairs (FIG. 3) and serial analysis of gene expression (SAGE) counts 25–75 mRNAs during each sequencing run of SAGE tags (FIG. 5).

Clearly, when so complex a cDNA probe is used, one of the main challenges during microarray analysis is to be certain that a hybridization signal is specific and quantitative to a known gene sequence that is laid on a 'chip'. One simple control experiment could clearly illustrate the potential problems in hybridization with a complex cDNA probe; instead of labelling all the mRNAs by reverse transcription with oligo-dT primers, only one mRNA at a time could be labelled for several genes that are highly or rarely expressed that are represented on an array with a corresponding gene-specific primer (for example, a primer that anneals just upstream of the polyA of a gene of interest). These single gene-specific probes, when hybridized to the microarrays individually, will provide an accurate glimpse of the actual sensitivity and specificity of microarrays, in comparison with complex cDNA probes labelled with an oligo-dT primer. If an experiment fails to detect only one gene specifically at a time on an array that contains tens of thousands of other genes when a gene-specific cDNA probe is used, how can we be certain that all the 'colourful' signals seen on an array or 'chip' truly reflect an accurate and sensitive snapshot of global gene expression within a cell?

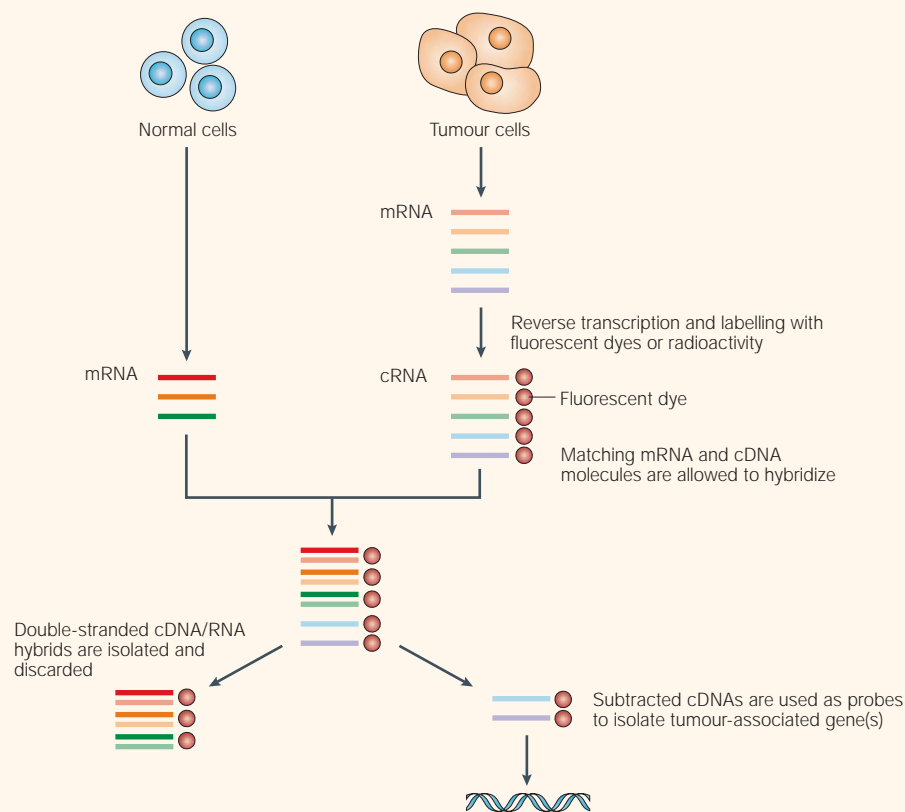


Figure 2 | Subtractive hybridization. All messenger RNAs that are expressed in a cell are first converted to single-stranded complementary DNAs (ss cDNAs), which are then hybridized to an excess of all the mRNAs that are expressed in the other cell type being compared. Genes that are commonly expressed in both cells will form cDNA/mRNA double-stranded duplexes, whereas cDNA that is uniquely expressed in the first cell will be in single-stranded form, which can be separated from most double-stranded cDNA/mRNA species. The resulting 'subtracted cDNAs' are then further characterized to determine if they indeed represent genes that are uniquely expressed in the first cell.

Microarrays

More than a decade after differential hybridization was introduced came two 'hot' technologies — cDNA microarrays, pioneered by Patrick Brown and David Botstein²⁷, and GeneChip arrays (oligo arrays), developed at Affymetrix²⁸ (FIG. 4). These techniques are, in essence, based on the differential-hybridization strategy, in which

cDNA plaques are replaced with spotted cDNAs or oligos, and radioactive labels are replaced with fluorescent ones. The promises of these methods^{29,30} are based on their potential in being able to simultaneously analyse the expression of mRNAs from tens of thousands of genes, which can then be further analysed using computers, in the hope that gene-expression patterns can be transformed

into more easily interpretable biological pathways for the understanding and classification of cancer. Indeed, DNA microarrays have been used to profile gene-expression patterns of almost all of the main cancers — including leukaemia^{31,32}, lymphoma³³, adenocarcinoma of the lung^{34,35}, breast³⁶ and prostate³⁷ — and promise to change the way cancer is diagnosed, classified and treated in the clinic. However, the realization of these potentials will be a considerable challenge, as the differences between studies of the same tumour types can often be more striking than their similarities^{38–41}. The best such examples are two large microarray studies on lung cancer^{34,35}.

One of the greatest advantages of microarrays over other methods, including DD, is that each spot on a microarray contains a known sequence. So, once a signal is detected, the nature of the gene is known. However, the hidden catch of such a benefit is that it also makes array-based methods 'closed' systems that are only able to cover known gene sequences. This might be best exemplified by the fact that more than 80% of the more well-documented p53 tumour-suppressor gene targets were identified by other, open-system-based methods such as DD and serial analysis of gene expression (SAGE), which can identify both known and novel genes (TABLE 1).

The inherent complexity of the cDNA probes that are used in differential-hybridization strategies remains the root cause of the lack of signal sensitivity and specificity for most low-abundance mRNAs^{8,42,43}. Without a doubt, all human genes can eventually be condensed on a single array or 'chip', but uncertainty remains as to whether each of these tens of thousands of cDNA probes will hybridize to only their corresponding target template and to nothing else on the chip. Not surprisingly, more and more researchers are cautious about the accuracy of microarray data, but most studies place the blame only on inadequate bioinformatical and statistical tools for 'data mining' (the analysis of noisy data)^{42–49}, rather than on the fundamental problem of the complexity of cDNA probes (BOX 1). Therefore, as with any other method for the analysis of differential gene expression, data from microarray experiments should be considered with caution, unless each data point can be verified by an independent method such as northern-blot analysis.

Expressed sequence tags and SAGE. In the early 1990s, when the human genome project was taking shape, Craig Venter realized that

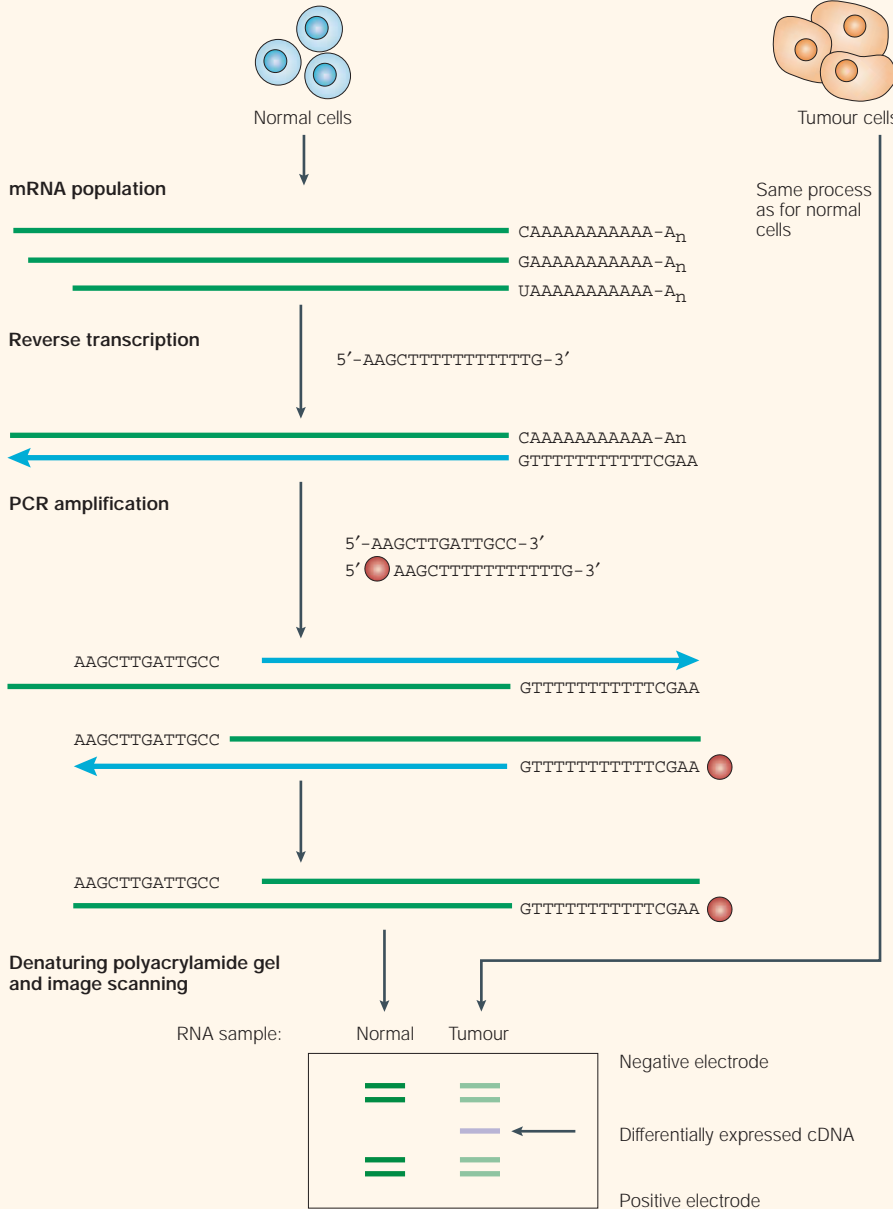


Figure 3 | Differential display. In this approach the 3' termini of eukaryotic messenger RNA are systematically amplified by reverse transcription-PCR using one of the three anchored oligo-dT primers (in this example, 5'-AAGCTTTTTTTTTTTG -3') in combination with a set of short primers of arbitrary sequences (in this example, 5'-AAGCTTGATTGCC-3'). The length of the arbitrary primers is designed such that each will recognize about 50–100 mRNAs under a given PCR condition. As a result, mRNA 3' tails, defined by any given pair of anchored primer and arbitrary primer, are amplified. By changing primer sequences, different subsets of mRNA can be analysed and displayed by denaturing polyacrylamide gel electrophoresis. Side-by-side comparisons of such complementary DNA patterns between or among relevant RNA samples indicate differences in gene expression. Differentially expressed cDNA bands can be retrieved and sequenced for further molecular characterization.

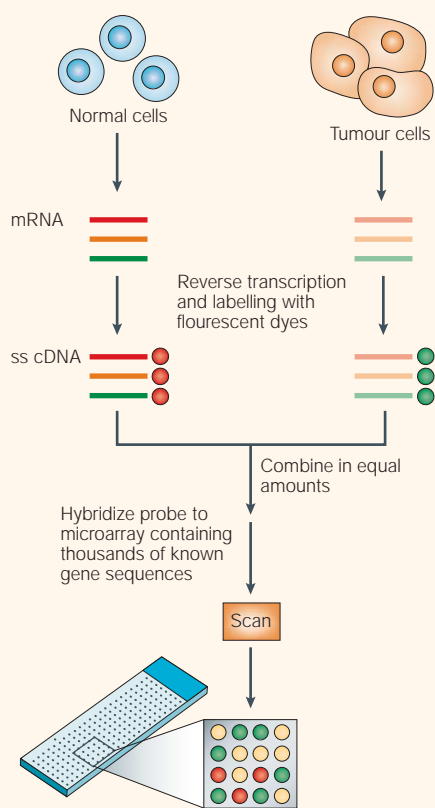


Figure 4 | DNA microarrays. All messenger RNAs that are expressed in the cells being compared are first labelled fluorescently by copying them into complementary DNAs using a reverse transcriptase. The resulting complex cDNA probes (**BOX 1**) are then used to hybridize to the cDNA templates or gene-specific oligos either directly synthesized or spotted on a glass surface to determine the expression of thousands of genes simultaneously. Red and green spots represent cDNAs that are only expressed in normal or tumours cells, respectively. Yellow fluorescence indicates cDNAs that are expressed in both samples.

sequencing expressed genes (cDNA), instead of the whole genome (in which up to 98% of DNA sequences are non-coding 'junk' sequences), would be a more sensible approach. His strategy, by a single run of sequencing of the 3' ends of randomly picked cDNA clones from a cDNA library, generated a comprehensive collection of such expressed sequence tags (ESTs)⁵⁰. The EST sequencing not only resulted in the discovery of many novel genes but also provided information on the relative abundance in expression of each gene based on the number of times a corresponding cDNA sequence was represented in a cDNA library from either normal or tumour cells. The EST sequencing strategy has had a key role in gene discovery and cataloging for

the National Institutes of Health (NIH) Cancer Genome Anatomy Program, which provides a convenient source of cDNA clones for functional studies of genes that have been identified by methods for comparative analysis of gene expression. Because of the high cost and labour-intensive nature of comprehensive EST sequencing, the method itself is rarely used directly to identify differentially expressed genes.

In the mid 1990s, SAGE was developed by Kenneth Kinzler and Bert Vogelstein's group⁵¹. Unlike the original EST sequencing strategy, in which cDNA clones were randomly picked from cDNA libraries, SAGE technology measures the level of gene expression based on the frequency of occurrence of the 3' signature SAGE tags of 10–14 bases in length that might be unique to each transcript (FIG. 5). Because of the minimal sequence information that is necessary to define an expressed gene or mRNA, a dozen or more SAGE tags from different genes can be obtained and sequenced at a time, which greatly speeds up the EST counting process. Like DD, SAGE is an open-system-based gene-discovery tool. However, because of the limited sequence information that is contained in 10–14-base SAGE tags, adequate gene assignment using SAGE methods still requires an extensive bioinformatic support, such as an extensive collection of cDNA sequences for the biological system under investigation. The NIH Cancer Genome Project (CGAP) now maintains a comprehensive SAGE database for various normal and cancer cell lines and tissues, with a web interface known as SAGE Genie^{52,53}, which allows a quick glimpse of the expression pattern for a gene of interest.

Beads-based EST sequencing is a recent method that was developed by Sydney Brenner and marketed by Lynx Therapeutics. The method, also known as massively parallel signature sequencing (MPSS) combines non-gel-based signature sequencing with *in vitro* cloning of millions of templates on separate 5-micron-diameter microbeads⁵⁴. MPSS captures and identifies transcript sequences and analyses the level of expression of expressed genes in a sample by counting the number of individual mRNA molecules that represent each gene. Individual mRNAs are identified through the generation of a 17–20-base signature sequence, immediately adjacent to the 3' end of the 3'-most Sau3A restriction site in cDNA sequences. (For a more detailed description of this methodology, see **Beads-based EST sequencing** in online links box)

A p53 profile

Cancer research has taken a great leap forward in terms of the number of cancer-related genes that have been identified by the different methods for analysing differential gene expression. To provide a glimpse of the progress that has been made over the past decade in the understanding of cancer through the study of differential gene expression, we summarize the identification of key p53 target genes that have been functionally characterized, as well as the methods that were used for their identification.

As mentioned earlier, p53 was discovered more than 20 years ago using one-dimensional protein gel electrophoresis. However, it was not until a decade later that Bert Vogelstein's group showed that p53 is a nuclear-DNA-binding protein and that it functions as a transcription factor⁵⁵. Since then, the search for p53 target genes has intensified. The p21 cyclin-dependent kinase 2 inhibitor was the first prototype p53 target to be identified using the subtractive-hybridization technique¹⁴. Proof that p21 alone was not sufficient to mediate the tumour-suppressor function of p53 came when mice with a p21 knockout failed to show a tumour-prone phenotype⁵⁶. This revelation fuelled the search for additional p53 target genes, especially those that might mediate the apoptotic function of p53. Increasing numbers of such candidate p53 target genes are being identified⁵⁷ (TABLE 1), but, so far, no single gene has been shown to be the *bona fide* p53 target based on either human or mouse genetic evidence³. An emerging thought has been that p53 might control a network of tissue-specific genes, rather than a single target, which, together, carry out the full p53 tumour-suppressor functions³. Most of the newer potential p53 target genes listed in TABLE 1 have been partially characterized and their expression confirmed, at least, by the 'gold' standard — northern-blot analysis. Functionally, these p53 target genes fall mainly into four categories: cell-cycle or growth regulators (such as p21); proapoptotic factors (such as BAX); DNA-damage repair proteins (such as GADD45 and TP53TG1); and negative regulators of p53 (such as MDM2 and PIRH2). About half of these better understood p53 target genes were identified by DD, and the other half by SAGE and subtractive hybridization (TABLE 1). Using a human gene chip from Affymetrix that contained 6,000 genes, Arnold Levine's group identified 107 upregulated and 54 downregulated p53 target genes⁵⁸. This result extrapolates to at least 500 upregulated and 260 downregulated p53

Table 1 | Key potential p53 target genes identified by different technologies and further characterized

Gene	Definition/function	Method*	References
<i>MDM2</i>	p53 negative regulator	Candidate	64
<i>WAF1</i>	CDK2 inhibitor	SH, SAGE	14
<i>14-3-3 σ</i>	Growth inhibition	SAGE	57
<i>GADD45</i>	DNA repair	SH	57
<i>BAX</i>	Apoptosis	Candidate	65
Cyclin G	Cell-cycle regulator	DD	66
<i>IGFBP3</i>	IGF binding protein, growth inhibition	SH	67
<i>PIG3</i>	NADPH-quinone oxidoreductase	SAGE	68
<i>KILLER/DR5</i>	Apoptosis	SH, SAGE	69
<i>E124/PIG8</i>	Novel gene, apoptosis	DD, SAGE	68,70
<i>PAG608</i>	Novel zinc-finger protein, apoptosis	DD	71
<i>DDA3</i>	Novel gene, growth inhibition	DD	72
<i>TP53TG1</i>	Novel gene, DNA damage	DD	73
<i>TP53TG3</i>	Novel gene, cell-cycle checkpoint	DD	74
<i>p53R2</i>	Ribonucleotide reductase	DD	75
<i>PERP</i>	Novel gene, pro-apoptotic	SH	76
<i>PIR121</i>	Novel gene, RNA binding	Array	77
<i>NOXA</i>	Novel gene, pro-apoptotic BH3 protein	DD	78
<i>PIDD</i>	Novel gene, death-domain protein	DD	79
<i>p53AIP1</i>	Novel gene, apoptosis, p53 phosphorylation	DD	80
<i>p53DINP1</i>	Novel gene, apoptosis, p53 phosphorylation	DD	81
<i>PUMA</i>	Novel gene, pro-apoptotic BH3-protein	SAGE, Array	82,83
<i>PIRH2</i>	Ubiquitin ligase, p53 negative regulator	DD	84
<i>PAC1</i>	Protein phosphatase, pro-apoptotic	Array	85
<i>FAS/APO1</i>	Cell-death receptor	Candidate	86
<i>APAF1</i>	Apoptosis	Array	87
<i>PTEN</i>	Tumour suppressor	Candidate	88
<i>BID</i>	Apoptosis	Array	89

*Differential display (DD), DNA-microarray (Array), serial analysis of gene expression (SAGE), subtractive hybridization (SH) and candidate gene approach (Candidate), which evaluates an individual gene of an investigator's choosing. *APAF1*, apoptotic protease activating factor; *BAX*, BCL2-associated X protein; *BID*, BH3 interacting domain death agonist; *CDK2*, cyclin-dependent kinase 2; *DDA3*, differential display and activated by p53; *GADD45*, growth arrest and DNA-damage-inducible; IGF, insulin-like growth factor; *IGFBP3*, insulin-like growth factor binding protein 3; *PAG608*, p53-activated gene 608; *PTEN*, phosphatase and tensin homologue.

target genes, given a minimum estimate of 30,000 genes in the human genome. Another recent cDNA microarray screening for p53 target genes that used more than 33,000 unique human cDNAs or ESTs implicated more than 1,500 potential p53 targets⁵⁹. It is interesting to note that although many candidate p53 target genes have been isolated by microarrays, few of these genes have been confirmed by northern-blot analysis or further characterized functionally, which could take some time to sort out. Without further confirmation and functional characterization of these genes, it is not yet clear whether the large number of p53 target genes that have been identified by arrays are in fact 'noise' (false positives) of the method, or whether the screenings by other methods were not comprehensive enough.

Future directions

We now have an arsenal of gene-expression analysis technologies at our disposal to study differential gene expression. New methods, such as beads-based EST sequencing, are continually being invented and tested⁴. Among these, however, two fundamentally different approaches and schools of thought are diverging. The traditional reductionist approach with hypothesis-driven research that focuses on one gene at a time is now being challenged by high-technology, hypothesis-generating genomic approaches. Two Nobel laureates have been so concerned about this genomic approach — which requires huge resources and generates enormous amounts of data that are hard to analyse⁶⁰ — that one, Sydney Brenner, dubbed it 'Sillycon valley fever'⁶¹, and the other, Walter Gilbert, called the 'omics' era the death of molecular biology⁶².

No matter which gene-discovery methods are used for hunting down the mechanisms of cancer, ultimately it will be the functional characterization of each gene — identified by genetic, cell-biological and biochemical methods — that will shed light on its true relevance. However, these processes can be painstakingly long and hard. Remember that the p53 tumour-suppressor gene was discovered more than two decades ago and has been worked on since by tens of thousands of laboratories throughout the world. For nearly 10 years after its discovery, p53 was believed to be a culprit oncogene. Only later was it vindicated as a tumour-suppressor gene and the 'guardian of the genome'. Although much progress has been made in understanding the function of p53 as a transcription factor that controls the cell-cycle progression and

apoptosis in response to DNA damage, the exact molecular nature of how p53 acts as this crucial tumour suppressor still remains elusive. The recently developed RNA interference techniques should help to speed up gene characterizations.

Any gene-expression profiling studies, whether they are carried out by DD, SAGE, DNA microarrays or proteomics, should be considered prone to error without further confirmation of the expression of each gene by methods that are independent of the original study — in other words, guilt should not be assumed without a true association being corroborated. As a judiciary analogy, guilt by association, racial profiling and the assumption of guilt until innocence is proven are all flawed legal practices. Similarly, we believe that identifying the culprit genes for cancer through the analysis of differential gene expression should be held to the same standard. Like a crime suspect, each gene should be treated as innocent until proven guilty by further biological and functional studies. The profiling of only candidate gene sequences followed by 'cluster analysis', without any subsequent confirmation, should be considered to be as prone to error as racial profiling.

In the preface to a method book on protein purification, Nobel Laureate Arthur Kornberg quoted an admonition of Efraim Racker — "Don't waste clean thinking on dirty enzymes" — to illustrate the importance of good biochemical practice in enzymology⁶³. A similar doctrine — "Don't waste clear thinking on dirty data" — will certainly continue to help produce a better quality of science and contribute to steady progress in the field of gene-expression analysis of cancer. Let's move beyond gene listing and get down to the biology.

Peng Liang is at the Vanderbilt–Ingram Cancer Center, Department of Cancer Biology, School of Medicine, Vanderbilt University, Nashville, Tennessee 37232, USA.

Arthur B. Pardee is at the Department of Adult Oncology, Dana–Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA.

**Correspondence to A.B.P. or P.L.
e-mail: Arthur_Pardee@dfci.harvard.edu;
peng.liang@vanderbilt.edu**

doi: 10.1038/nrc1214

- Knudson, A. G. Two genetic hits (more or less) to cancer. *Nature Rev. Cancer* **1**, 157–162 (2001).
- Sager, R. Expression genetics in cancer: shifting the focus from DNA to RNA. *Proc. Natl Acad. Sci. USA* **94**, 952–955 (1997)
- Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408**, 307–310 (2000).
- Lorkowski, S. & Cullen, P. (eds.) *Analyzing Gene Expression. A Handbook of Methods: Possibilities and Pitfalls* (Wiley–VCH, Weinheim, 2002).

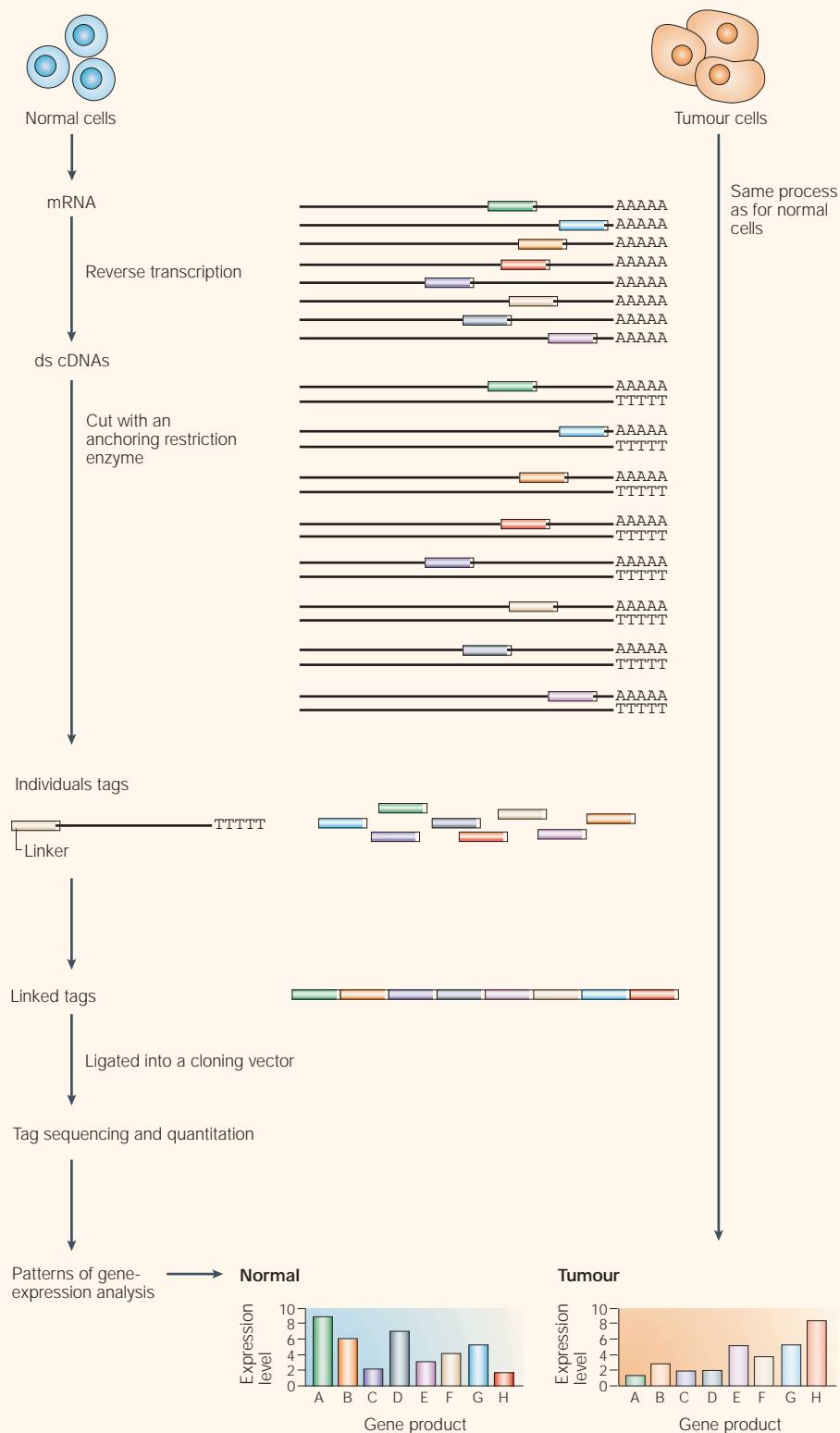


Figure 5 | SAGE. (Serial analysis of gene expression). All messenger RNAs that are expressed in a cell or from a tissue specimen are first reverse transcribed with an oligo-dT primer and converted to double-stranded cDNAs (ds cDNAs), which are then cut with an anchoring restriction enzyme. The most 3' fragment of each cDNA is ligated to a linker that contains a type IIS restriction site that allows a tagging restriction enzyme to cut the tagged cDNA 10–14 bases downstream. The resulting transcript-specific SAGE tags that are released are concatemerized (25–75 tags) and ligated into a cloning vector for sequencing. After sequencing many such concatemerized SAGE tags the level of gene expression can be estimated by computer analysis that is based on the occurrence of each SAGE tag having a unique sequence.

5. Linzer, D. I. & Levine, A. J. Characterization of a 54K dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected embryonal carcinoma cells. *Cell* **17**, 43–52 (1979).
6. O'Farrell, P. H. High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021 (1975).
7. Croy, R. G. & Pardee, A. B. Enhanced synthesis and stabilization of Mr 68,000 protein in transformed BALB/c-3T3 cells: candidate for restriction point control of cell growth. *Proc. Natl Acad. Sci. USA* **80**, 4699–8703 (1983).
8. Sargent, T. D. Isolation of differentially expressed genes. *Methods Enzymol.* **152**, 423–432 (1987).
9. Masiakowski, P. *et al.* Cloning of cDNA sequences of hormone-regulated genes from the MCF-7 human breast cancer cell line. *Nucl. Acids Res.* **10**, 7895–7903 (1982).
10. Manzari, V. *et al.* Abundant transcription of a cellular gene in T cells infected with human T-cell leukemia-lymphoma virus. *Proc. Natl Acad. Sci. USA* **80**, 11–15 (1983).
11. Zimmermann, C. R. *et al.* Molecular cloning and selection of genes regulated in *Aspergillus* development. *Cell* **21**, 709–715 (1980).
12. Hedrick, S. M. *et al.* Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. *Nature* **308**, 149–153 (1984).
13. Davis, M. *On the Trail of T-Cell Receptors. Accomplishment in Cancer Research*. 87–94 (General Motor Cancer Research Foundation, 1996).
14. El-Deiry, W. S. *et al.* WAF1, a potential mediator of p53 tumor suppression. *Cell* **75**, 817–825 (1993).
15. Liang, P. & Pardee, A. B. Differential display of eukaryotic mRNA by means of the polymerase chain reaction. *Science* **257**, 967–971 (1992).
16. Welsh, J. *et al.* Arbitrarily primed PCR fingerprinting of RNA. *Nucl. Acids Res.* **20**, 4965–4970 (1992).
17. Liang, P. *et al.* Analysis of altered gene expression by differential display. *Methods Enzymol.* **254**, 304–321 (1995).
18. Liang, P. A decade of differential display. *Biotechniques* **33**, 338–346 (2002).
19. McCarthy, S. A. *et al.* Rapid induction of heparin-binding epidermal growth factor/diphtheria toxin receptor expression by *Raf* and *Ras* oncogenes. *Genes Dev.* **9**, 1953–1964 (1995).
20. Zhang, R. *et al.* Identification of a novel ligand-receptor pair constitutively activated by *Ras* oncogenes. *J. Biol. Chem.* **275**, 24436–24443 (2000).
21. You, M. *et al.* ch-IAP1, a member of the inhibitor-of-apoptosis protein family, is a mediator of the antiapoptotic activity of the v-Rel oncoprotein. *Mol. Cell Biol.* **17**, 7328–7341 (1997).
22. Park, B.-W. *et al.* Induction of the Tat-binding protein 1 gene accompanies the disabling of oncogenic *erbB* receptor tyrosine kinases. *Proc. Natl Acad. Sci. USA* **96**, 6434–6438 (1999).
23. Wang, M. *et al.* Interleukin-24 (Mob-5/Mda-7) signals through two heterodimeric receptors, IL-22R1/IL-20R2 and IL-20R1/IL-20R2. *J. Biol. Chem.* **277**, 7341–7347 (2002).
24. Liang, P. Factors ensuring successful use of differential display. *Methods* **16**, 361–364 (1998).
25. Shimkets, R. A. *et al.* Gene expression analysis by transcript profiling coupled to a gene database query. *Nature Biotechnol.* **17**, 798–803 (1999).
26. Cho, Y. *et al.* Multi-color fluorescent differential display. *Biotechniques* **30**, 562–572 (2001).
27. Schena, M. *et al.* Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
28. Chee, M. *et al.* Accessing genetic information with high-density DNA arrays. *Science* **274**, 610–614 (1996).
29. Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA* **98**, 15149–15154 (2001).
30. Chung, C. H., Bernard, P. S. & Perou, C. M. Molecular portraits and the family tree of cancer. *Nature Genet.* **32**, S533–S540 (2002).
31. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
32. Yeoh, E. J. *et al.* Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **2**, 133–143 (2002).
33. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
34. Garber, M. E. *et al.* Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl Acad. Sci. USA* **98**, 13784–13789 (2001).
35. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA* **98**, 13790–13795 (2001).
36. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).
37. Dhanasekaran, S. M. *et al.* Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822–826 (2001).
38. Wooster, R. Cancer classification with DNA microarrays: is less more? *Trends Genet.* **16**, 327–329 (2000).
39. Petricoin, E. F. *et al.* Medical applications of microarray technologies: a regulatory science perspective. *Nature Genet.* **32**, S474–S479 (2002).
40. Goodman, N. Microarrays: hazardous to your science. *Genome Technol.* April, 42–45, (2003).
41. Ring, B. Z. & Ross D. T. Microarrays and molecular markers for tumor classification. *Genome Biol.* **3**, 2005 (2002).
42. Kuo, W. P. *et al.* Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405–412 (2002).
43. Kothapalli, R. *et al.* Microarray results: how accurate are they? *BMC Bioinformatics* **3**, 22 (2002).
44. Goryachev, A. B., Macgregor, P. F. & Edwards, A. M. Unfolding of microarray data. *J. Comput. Biol.* **8**, 443–461 (2001).
45. King, H. C. & Sinha, A. A. Gene expression profile analysis by DNA microarrays: promise and pitfalls. *JAMA* **286**, 2280–2288 (2001).
46. Shedden, K. & Cooper, S. Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc. Natl Acad. Sci. USA* **99**, 4379–4384 (2002).
47. Cooper, S. Cell cycle analysis and microarrays. *Trends Genet.* **18**, 289–290 (2002).
48. Jenssen T.-K. *et al.* Analysis of repeatability in spotted cDNA microarrays. *Nucl. Acids Res.* **30**, 3235–3244 (2002).
49. Quackenbush, J. Microarray data normalization and transformation. *Nature Genet.* **32**, S496–S501 (2002).
50. Adams M. D. *et al.* Sequence identification of 2,375 human brain genes. *Nature* **355**, 632–634 (1992).
51. Velculescu, V. E. *et al.* Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
52. Boon, K. *et al.* An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA* **99**, 11287–11292 (2002).
53. Liang, P. SAGE Genie: a suite with panoramic view of gene expression. *Proc. Natl Acad. Sci. USA* **99**, 11547–11548 (2002).
54. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnol.* **18**, 630–634 (2000).
55. Kern, S. E. *et al.* Oncogenic forms of p53 inhibit p53-regulated gene expression. *Science* **256**, 827–830 (1992).
56. Deng, C. *et al.* Mice lacking p21^{CIP1}/WAF1 undergo normal development, but are defective in G1 checkpoint control. *Cell* **82**, 675–684 (1995).
57. El-Deiry, W. S. Regulation of p53 downstream genes. *Semin. Cancer Biol.* **8**, 345–357 (1998).
58. Zhao, R. *et al.* Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev.* **14**, 981–993 (2000).
59. Wang, L. *et al.* Analyses of p53 target genes in the human genome by bioinformatic and microarray approaches. *J. Biol. Chem.* **276**, 43604–43610 (2001).
60. Gibbs, W. W. Shrinking to enormity: DNA microarrays are reshaping basic biology but scientists fear they may soon drown in the data. *Sci. Am.* **284**, 33–34 (2001).
61. Brenner, S. Sillycon valley fever. *Curr. Biol.* **9**, R671 (1999).
62. Gilbert, W. Life after the helix. *Nature* **421**, 315–316 (2003).
63. Kornberg, A. Why purify enzyme? *Methods Enzymol.* **182**, 1–5 (1990).
64. Wu, X. *et al.* The p53-mdm-2 autoregulatory feedback loop. *Genes Dev.* **7**, 1126–1132 (1993).
65. Miyashita, T. & Reed, J. C. Tumor suppressor p53 is a direct transcriptional activator of the human bax gene. *Cell* **80**, 293–299 (1995).
66. Okamoto, K. & Beach, D. Cyclin G is a transcriptional target of the p53 tumor suppressor protein. *EMBO J.* **13**, 4816–4822 (1994).
67. Buckbinder, L. *et al.* Induction of the growth inhibitor IGF-binding protein 3 by p53. *Nature* **377**, 646–649 (1995).
68. Polyak, K. *et al.* A model for p53-induced apoptosis. *Nature* **389**, 300–305 (1997).
69. Wu, G. S. *et al.* KILLER/DR5 is a DNA damage-inducible p53-regulated death receptor gene. *Nature Genet.* **17**, 141–143 (1997).
70. Gu, Z. *et al.* ei24, a p53 response gene involved in growth suppression and apoptosis. *Mol. Cell Biol.* **20**, 233–241 (2000).
71. Israeli, D. *et al.* A novel p53-inducible gene, *PAG608*, encodes a nuclear zinc finger protein whose overexpression promotes apoptosis. *EMBO J.* **16**, 4384–4392 (1997).
72. Lo, P. K. *et al.* Identification of a novel mouse p53 target gene *DDA3*. *Oncogene* **18**, 7765–7774 (1999).
73. Takei, Y. *et al.* Isolation of a novel *TP53* target gene from a colon cancer cell line carrying a highly regulated wild-type *TP53* expression system. *Genes Chromosom. Cancer* **23**, 1–9 (1998).
74. Ng, C. *et al.* Isolation and characterization of a novel *TP53*-inducible gene, *TP53TG3*. *Genes Chromosom. Cancer* **26**, 329–335 (1999).
75. Tanaka, H. *et al.* A ribonucleotide reductase gene involved in a p53-dependent cell-cycle checkpoint for DNA damage. *Nature* **404**, 42–49 (2000).
76. Altardi, L. *et al.* *PERP*, an apoptosis-associated target of p53, is a novel member of the PMP-22/gas3 family. *Genes Dev.* **14**, 704–718 (2000).
77. Saller, E. *et al.* Increased apoptosis induction by 121F mutant p53. *EMBO J.* **18**, 4424–4437 (1999).
78. Oda, E. *et al.* Noxa, a BH3-only member of the Bcl-2 family and candidate mediator of p53-induced apoptosis. *Science* **288**, 1053–1058 (2000).
79. Lin, Y., Ma, W. & Benichou, S. Pidd, a new death-domain-containing protein is induced by p53 and promotes apoptosis. *Nature Genet.* **26**, 124–127 (2000).
80. Oda, E. *et al.* p53AIP1, a potential mediator of p53-dependent apoptosis, and its regulation by Ser-46-phosphorylated p53. *Cell* **102**, 849–862 (2000).
81. Okamura, S. *et al.* p53^{DINP1}, a p53-inducible gene, regulates p53-dependent apoptosis. *Mol. Cell* **8**, 85–94 (2001).
82. Yu, J. *et al.* PUMA induces the rapid apoptosis of colorectal cancer cells. *Mol. Cell* **7**, 673–682 (2001).
83. Nakano, K. & Vonsden, K. H. *PUMA*, a novel proapoptotic gene, is induced by p53. *Mol. Cell* **7**, 683–694 (2001).
84. Leng, R. P. *et al.* Pirh2, a p53-induced ubiquitin-protein ligase, promotes p53 degradation. *Cell* **112**, 779–791 (2003).
85. Yin, Y. *et al.* PAC1 phosphatase is a transcription target of p53 in signalling apoptosis and growth suppression. *Nature* **422**, 527–531 (2003).
86. Owen-Schaub, L. B. *et al.* Wild-type human p53 and a temperature-sensitive mutant induce Fas/APO-1 expression. *Mol. Cell Biol.* **15**, 3032–3040 (1995).
87. Kannan, K. *et al.* DNA microarray analysis of genes involved in p53 mediated apoptosis: activation of *Apaf-1*. *Oncogene* **20**, 3449–3455 (2001).
88. Stambolic, V. *et al.* Regulation of *PEN* transcription by p53. *Mol. Cell* **8**, 317–325 (2001).

Online links

DATABASES

The following terms in this article are linked online to:

Cancer.gov: <http://cancer.gov/>
breast cancer | leukaemia | lung cancer | lymphoma | prostate cancer
LocusLink: <http://www.ncbi.nlm.nih.gov/LocusLink/>
BAX | ERBB | GAD45 | IL-24 | MDM2 | p53 | p68 | PIRH2 | RAS | v-REL | WAF1

FURTHER INFORMATION

Access to this interactive links box is free online.

Beads-based EST sequencing: <http://www.lynxgen.com>

cDNA microarray: <http://cmgm.stanford.edu/pbrown/>

Differential display: <http://www.differentialdisplay.com>

GeneChip array: <http://www.affymetrix.com>

International Symposia on Differential Gene Expression:

<http://medschool.mc.vanderbilt.edu/GeneXP/>

SAGE: <http://www.sagenet.org>

SAGE Genie: <http://cgap.nci.nih.gov/SAGE>