From genomics to proteomics: techniques and applications in cancer research

Daniel B. Martin and Peter S. Nelson

New technologies designed to facilitate the comprehensive analyses of genomes, transcriptomes and proteomes in health and disease are poised to exert a dramatic change on the pace of cancer research and to impact significantly on the care of cancer patients. These approaches have already demonstrated the power of molecular medicine in discriminating among disease subtypes that are not recognizable using traditional pathological criteria and in identifying specific genetic events involved in cancer progression. This review outlines the current status of these technologies and highlights recent studies in which they have been applied in the context of carcinogenesis.

A TRENDS Guide to
Cancer Biology

Daniel B. Martin and Peter S. Nelson*
Divisions of Human Biology and Clinical Research, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA.
*e-mail: pnelson@fhcrc.org

The scope of molecular biology has undergone revolutionary changes in the past two decades. In part, this has been catalyzed by the immense task of sequencing the human genome and the requirement for tools that facilitate the rapid and accurate acquisition of raw sequence data and subsequently assist in complex analyses. Technological advances in automation and bioinformatics have spawned a discipline of biology, termed genomics, which can broadly be defined as the generation and analysis of information about genes and genomes. The field is characterized by global, comprehensive studies that are conducted in a systematic fashion. A full understanding of the complex processes that occur during the transition from normal to neoplastic cellular growth would also appear to require such a comprehensive and systematic approach. Thus, the application of genomics to cancer biology holds great potential for identifying the mechanisms that lead to malignancy and for developing therapeutic strategies. Although still in its infancy, genomics has begun to produce the anticipated results through the identification and characterization of individual genes and, recently, patterns of gene expression that distinguish malignant or premalignant cells from their normal counterparts.

The importance of the development of high-throughput DNA sequencing methods to our current ability to perform genome-scale science cannot be overstated. The completion of the human genome sequence by the Human Genome Project (HGP) and Celera was possible only because of improvements in sequencing technology. Automated capillary sequencing using fluorescent nucleotides has allowed researchers to rapidly sequence individual genes of interest and, when multiplexed in assembly-line format, permitted the sequencing of the entire 14.8 billion base-pair human genome over just nine months[1]. This genomic sequence, with the attendant chromosomal mapping data, has greatly enhanced the ability to isolate specific genes involved in heritable cancers, such as those responsible for predisposition to breast cancer, BRCA1 and BRCA2[2,3]. Employing genome-wide scans, investigators have also performed linkage studies of individuals from families at high risk for the development of prostate cancer and, to date, six distinct loci have been identified[4]. A specific gene, *ELAC2*, located within the Hereditary Prostate Cancer gene 2 (*HPC2*) locus on chromosome 17 was recently shown to exhibit a polymorphism that segregates with prostate cancer in two pedigrees[5]. In a case-controlled study, the probability of having prostate cancer

 0962-8924/01/$ – see front matter PII: S0962-8924(01)02123-7

was increased in men who carried a Leu217/Thr541 variant of the *ELAC2* gene (odds ratio = 2.37; 95% CI 1.06–5.29). Genotypes at *HPC2/ELAC2* were estimated to account for 5% of prostate cancer in the study population[6]. Searches for the specific genes within the remaining prostate-cancer-associated loci are under way.

In addition to the raw sequence and mapping data, the two sequencing projects have uncovered millions of single nucleotide polymorphisms (SNPs) representing individual DNA bases that vary between individuals. Although the bulk of these polymorphisms do not affect gene regulation or the function of encoded proteins, they can be used as genetic markers to pinpoint the location of nearby genes that are responsible for a disease phenotype. In addition, identifying the SNPs that do confer susceptibility to disease owing to an alteration in RNA or protein function could have a profound impact on cancer screening and therapy. A recent study described the analysis of a breast-cancer-associated SNP in the 3′ untranslated region of human prohibitin[7]. Prohibitin binds to members of the retinoblastoma tumor-suppressor protein family, and this interaction leads to repression of gene expression. The prohibitin 3′ untranslated region encodes an RNA molecule that arrests cell proliferation when microinjected into normal mammary epithelial cells and breast cancer cell lines, and exhibits tumor-suppressor activity in animals. A SNP was found in the 3′ untranslated region that lacks the antiproliferative function. In a case-control study of breast cancer patients, a strong association between this prohibitin variant allele and cancer was reported in patients who had a first-degree relative with the disease (odds ratio 2.5, $P = 0.005$) and in a subset of women diagnosed with breast cancer before the age of 50 (odds ratio 4.8, $P = 0.003$).

Comprehensive analyses of gene expression: transcripts

Genome-wide screening of cellular gene expression profiles using microarrays of DNA molecules is a recently developed tool that has already proven useful in characterizing human cancers. Combining molecular biology with robotic technology has resulted in the ability to create arrays comprising thousands of distinct spots of DNA, each tens of micrometers in size and each corresponding to a different specific gene. The two most common microarray configurations consist of oligonucleotides[8] (chemically synthesized DNA chains) or cDNAs[9] (DNA fragments obtained from reverse-transcribing messenger RNAs). These arrays are currently capable of detecting and quantitating the transcript abundance of more than 30 000 different target genes simultaneously. The hypothesis underlying this approach is that transcript expression patterns, when compared in sufficient numbers, will allow reproducible descriptions of cell and tissue types (e.g. brain versus muscle) and tissue ‘states’ (e.g. normal versus cancerous) that have scientific applications and predictive clinical power. Additionally, any interesting changes in the expression of individual genes can be followed with more detailed biochemical analyses.

Transcript profiling has already shown great potential to improve our understanding of tumor behavior. Microarray analysis of genes expressed in microscopically indistinguishable tissue specimens from a common subtype of B-cell lymphoma – diffuse large B-cell lymphoma (DLBCL) – identified two molecularly distinct forms of DLBCL that exhibited gene expression patterns characteristic of different stages of B-cell differentiation (Fig. 1)[10]. One type expressed genes characteristic of B-cells found in the germinal center of a lymph node, and the other type expressed genes normally induced upon activation of B-cells in the peripheral blood. Patients with germinal center B-like DLBCL had a significantly better overall survival than those with activated B-like DLBCL. The array-based risk stratification added significantly to the best available prediction method based on patient clinical characteristics and thus essentially identified a previously unknown clinical subtype of lymphoma.

A study of gene expression profiles acquired from breast cancer specimens further illustrates the utility of microarray technology to classify and stratify human cancers. Primary breast tumors from seven carriers of the *BRCA1* mutation, seven carriers of the *BRCA2* mutation and seven patients with sporadic cases of breast cancer were compared using a microarray of 5361 genes. Statistical analyses demonstrated that the gene-expression profile of each tumor clustered with tumors of the same type and differed significantly from tumors of the other types. Thus, these three subtypes of breast cancer could be distinguished rapidly based solely upon their gene expression signatures. In addition, the microarray experiments identified 176 genes that were expressed differentially between tumors with *BRCA1* and *BRCA2* mutations, suggesting that there are significant functional differences between these breast tumor subtypes[11]. This study indicates that tumor expression profiles can be used not only for identifying specific molecular pathways that could be altered in tumorigenesis but also for stratifying tumors into prognostic categories that have implications for treatment decisions.

Beyond genomics: proteomics and cancer biology

Although genomic data and transcript profiling offer tremendous opportunities to identify and understand molecular alterations in cancer, even the complete ‘genetic

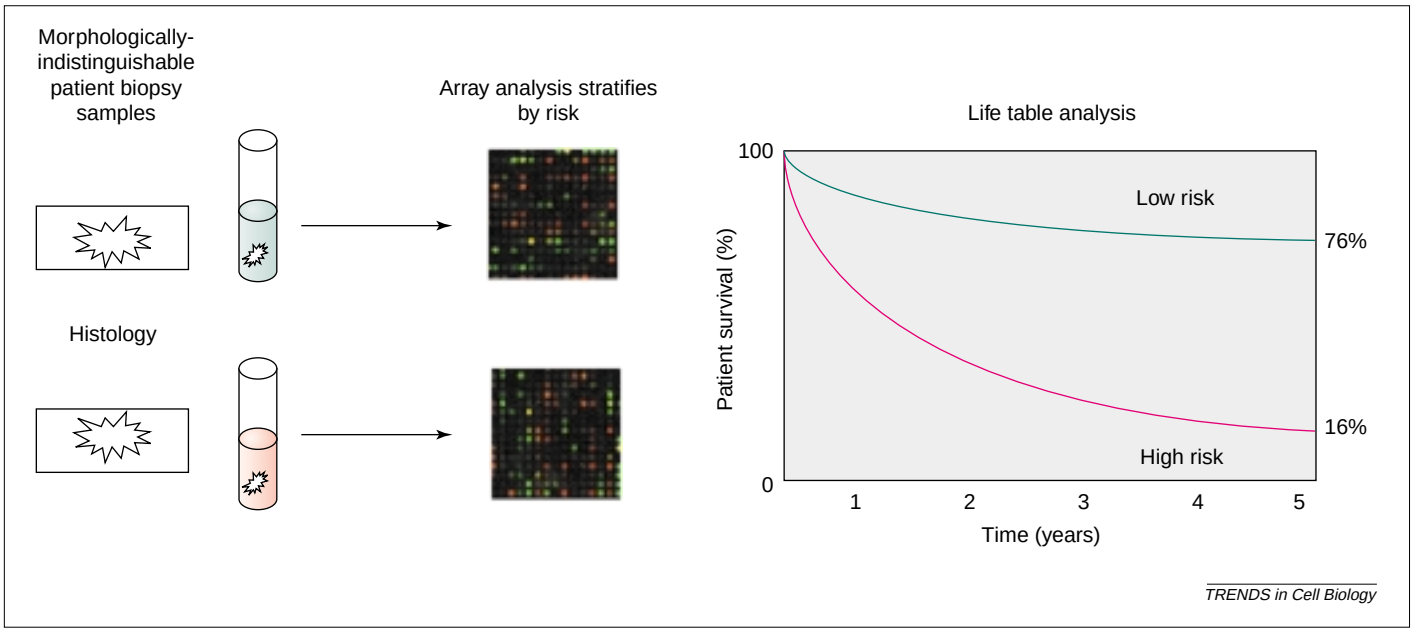

Figure 1. Risk stratification by tumor gene expression profiling

DNA microarray analysis stratification of histologically-similar cancer samples into two molecular subtypes. When compared with patient outcomes, these categories associate with significant differences in overall survival. Ultimately, tumor gene expression differences determined using microarray profiling could lead to the recommendation that some patients receive early aggressive therapy, whereas others require a more conservative approach.

blueprint' has serious limitations. Apart from the obvious fact that cellular functions are carried out by proteins, not by DNA and RNA, there are numerous protein modifications that are not apparent from the nucleic acid or amino acid sequence. These include differential RNA splicing and post-translational modifications such as phosphorylation and glycosylation. The genomic sequence does not specify which proteins interact, how interactions occur or where in the cell a protein localizes under various conditions. Transcript abundance levels do not always correlate with protein abundance levels and one cannot tell from the genomic sequence whether a gene is ever translated into protein or rather functions as an RNA. Recent genetic triumphs have been paralleled by a surge in interest in the comprehensive study of proteins and protein systems. This field has been dubbed 'proteomics', a word derived from 'proteome', which means the complete set of proteins expressed by the genome. From a biomedical standpoint, the field of proteomics has great potential because the bulk of pharmacological interventions and diagnostic tests are directed at proteins rather than genes. The inherent advantage afforded to proteomics over genomics is that the identified protein is itself the biological end-product.

The global study of proteins has many unique difficulties that set it apart from comprehensive studies of genes and transcripts. First, the behavior of proteins is determined by the tertiary structure of the molecule. Any assay based on protein binding depends on maintaining the native conformation of the protein. This puts constraints on the systems used to capture protein targets in affinity-based assays. Second, the detection of low-abundance proteins poses a particular challenge, especially given that the dynamic ranges of proteins in biological systems can reach parts per million or greater. An amplification system analogous to the polymerase chain reaction has yet to be developed for protein studies. In addition, the behavior of proteins might or might not be governed quantitatively. Protein regulation is often based not on synthesis and degradation but on reversible modifications – for example, phosphorylation. Adding to the difficulty is that RNA splicing can produce splice variants that are highly homologous but which differ in function. Nonetheless, protein science has advanced to the point that some of these hurdles can be overcome.

Until recently, the study of global protein expression was performed nearly exclusively using two-dimensional gel electrophoresis (2DE). This technique allows the display of thousands of proteins as spots on a rectangular gel. Although 2DE gel maps have been made for many cancer cell lines and human bodily fluids[12], the technique is somewhat cumbersome, labor intensive, insensitive and not suitable for high-throughput applications. As a result of these limitations, new approaches for performing large-scale protein studies have been developed, including mass spectrometry (MS), yeast two-hybrid systems and protein arrays. At this juncture, the application of proteomics to the study of cancer biology should be considered an emerging discipline. Most reports have focused on proof-of-principle experiments and the introduction of new technologies. The impact of global qualitative

or quantitative protein analyses has yet to be demonstrated. However, the potential applications of proteomics-based approaches mandate that the technologies in early phases of development be applied to important problems in the field of cancer biology. Therefore, most of the following discussion will focus on the fundamentals of emerging proteome technologies and will consider their potential applications.

MS has emerged as one of the best ways to study proteins. It measures the charge-to-mass ratio of an ionized molecule, either a protein or peptide. The methods used to ionize the molecule of interest are either electrospray ionization, whereby a voltage placed on a fine needle causes a mist of fine droplets of charged particles, or matrix-assisted laser desorption/ionization (MALDI), where the protein/peptide of interest is crystallized within a matrix with an absorption peak at a specific wavelength to allow energy from a laser to excite the matrix and ionize the protein. When combined with a time of flight (TOF) spectrometer, MALDI has a very high sensitivity, requiring only femtomoles of sample for a good mass spectrum. Both of these techniques are amenable to automation and high-throughput analysis.

Until recently, the mass spectrometer was used mainly for protein identification through mass mapping. A purified protein or a spot from a gel with a given mass is digested into peptides by an enzyme known to cleave only after certain amino acids. The mass of each of these peptides is measured in the spectrometer. This list, or 'fingerprint', of peptide masses is characteristic of a specific protein. The protein can be identified unambiguously by comparing its peptide mass fingerprint with fingerprints produced by 'virtual' or *in silico* digests of every protein in the database with an identical mass.

A more powerful approach is to use the mass spectrometer in phases, this time for identifying peptides. In this technique, the spectrometer isolates a single peptide and applies energy to break the peptide randomly at each peptide bond. The fragments created are then measured. The spectrum of the fragment masses uniquely identifies the parent peptide, just as in the example above with a protein. An analogous search algorithm is applied, but this time every sequence of consecutive amino acids from the database with an equivalent mass is collected and fragmented virtually. Despite hundreds of thousands of potential peptides, the chance of a false assignment is low because the probability of matching every amino acid is low.

The field of proteomics is currently being propelled forward by the potential of MS. The potential applications have grown dramatically as reliable instruments have become increasingly affordable. The current generation of machines can perform the selection and fragmentation of peptides at speeds of hundreds an hour. The software algorithms have benefited from improvements in computational power and can identify unique peptides from hundreds of thousands of candidates in just a few seconds. By using chromatography as a separation technique, coupled directly to the inlet of the mass spectrometer, a steady stream of peptides from the digest of a complex

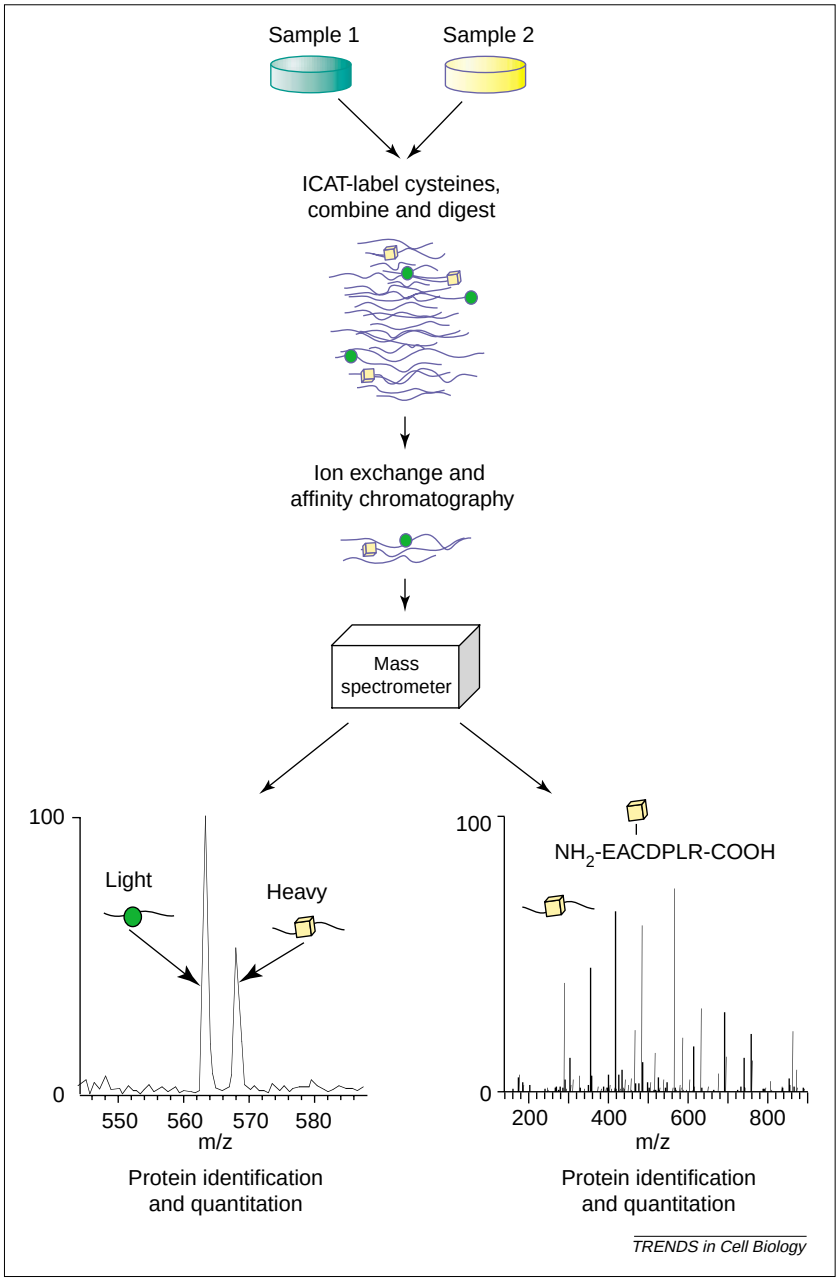

Figure 2. The ICAT analysis strategy

Proteins from a complex mixture are labeled with either the heavy or light isotope-coded affinity tag (ICAT) linker. The mixture is combined and proteolyzed. The ICAT incorporates a biotin molecule, which enables affinity purification following ion exchange chromatography. The isotopically distinct peptides are identified by the mass spectrometer and the ratio between the peptides gives the ratio between the parent proteins in samples 1 and 2. The spectrum of the fragmented peptide (lower right) is unique to that peptide alone and enables unambiguous identification.

mixture of proteins can be delivered continuously into the mass spectrometer for identification. The protein mixture can be as simple as a purified T-cell receptor or as complex as the membrane fraction of cancer cells. The output from the computer search is a list of all peptides, and thus, a list of all proteins present[13,14].

Comprehensive analyses of gene expression: proteins

Utilizing the power of MS, a high-throughput technique has been developed that facilitates direct qualitative and quantitative comparisons of complex protein mixtures[15]. This method, somewhat analogous to the microarray approach for assessing differential gene expression between two cell states, employs a chemical group or label made in two different isotopic forms: heavy and light. These labels, called isotope-coded affinity tags (ICAT), couple to all the cysteine residues in a protein mixture. The heavy reagent is added to one sample (cancer cells) and the light to another (normal cells). The samples are combined, the proteins are digested with specific proteases to generate peptide mixtures and the mixtures are analyzed in a mass spectrometer (Fig. 2). The isotopic substitutions do not affect the behavior of peptides during separations; thus peptides derived from the two different samples enter the spectrometer at the same time. The mass spectrometer measures the relative abundance of the heavy and light peptide forms in each sample and identifies each peptide by generating and analyzing the peptide fingerprints. In this manner, a global view of protein abundance in cells or tissues in two different states can be determined.

Proof-of-principle experiments using yeast as a model system have shown that the ICAT approach is robust, reproducible and amenable to high-throughput automation[16]. Application of the ICAT method complements transcript-profiling approaches by providing direct information detailing cellular protein alterations that occur during pathological processes.

Other techniques that borrow from the methods developed for comprehensive transcript expression studies incorporate arrays of molecules designed to capture and assay proteins. These methods represent a heterogeneous group of techniques that can assess protein–protein interactions, protein modifications and tissue profiling. The protein arrays have in common the placement of a small amount of target material on a solid support, and they use a variety of techniques to identify enzymatic or protein-binding activity. Two recent reports indicate that the array format can be used for assaying thousands of individual proteins simultaneously. Proteins of interest were attached covalently to a glass surface and washed with a mixture of fluorescently labeled proteins, which included those with highly specific protein–protein interactions. The method for immobilizing the targets maintained selective attachments between the bound proteins and known interactors, resulting in the discrimination of distinct proteins from the complex mixture. Proteins known to be enzymatic substrates were immobilized in a similar fashion, and specific protein interactions allowed enzymatic activity to be measured when the appropriate enzyme was present. Further developments of this approach could allow for the rapid screening of enzyme substrates or identify novel proteins that interact with specific ligands[17].

A second approach utilized a robotic device to print hundreds of specific antibody and antigen solutions on the surface of derivatized microscope slides[18]. Proteins were labeled by the covalent linkage of a fluorescent dye and placed on these arrays along with a fluorescent internal standard for each protein assayed. The fluorescence at each spot was quantitated against the standard. Of the 115 antibody–antigen pairs, 50% of the arrayed antigens and 20% of the arrayed antibodies provided specific and accurate measurements of their cognate ligands at or below concentrations of 0.34 $\mu g\ ml^{-1}$ and 1.6 $\mu g\ ml^{-1}$, respectively. Some of the antibody–antigen pairs allowed detection of the cognate ligands at concentrations below 1 $ng\ ml^{-1}$, a sensitivity sufficient for the measurement of clinically important proteins in patient body fluid samples.

In a variation of the protein chip theme, Ciphergen Biosystems has developed a chimera of array technology and mass spectrometry for profiling complex protein mixtures[19]. In this method, a conventional affinity substrate such as an ion-exchange resin is affixed to a solid support called a ProteinChip. A complex protein mixture such as a whole cell lysate or body fluid is added and the chip is washed to remove unbound proteins. Bound proteins are ionized directly from the slide by a high-power laser and drawn into a mass spectrometer for analysis[20]. The profile of protein masses ranging from 8 to 50 kDa can provide a signature representative of the particular protein mixture. The technique has the advantage of being reproducible between samples, adaptable to small quantities of tissue (such as those available from tumor specimens) and scaleable for high-throughput applications. A practical example of the ProteinChip technology is described in an exploratory study of urinary protein profiles for the identification of transitional cell carcinoma of the bladder[21]. This line of research could, ultimately, also lead to the identification of biomarkers or patterns of protein expression that might be used prognostically for monitoring disease progression or treatment response.

Concluding remarks

The tremendous quantities of data generated by technologies capable of comprehensive genome and proteome analyses will only increase as the methods are further automated. To date, genome and proteome research has depended upon, and greatly benefited from, significant advances in bioinformatics. MS-based proteomics is now possible only because of the simultaneous exponential growth both in computing power and in the genetic databases that permit large comparative searches to be performed rapidly and with high accuracy. The ability to assimilate and take full advantage of these complex data sets is closely tied to the further development of software tools that can apply filtering algorithms and recognize patterns that associate expression profiles with cellular characteristics, pharmacological interactions and epidemiological information. Taking full advantage of the information encoded in the genomes and proteomes of tumor cells could provide predictive models for how the cellular constituents interact to produce cellular phenotypes. In a recent proof-of-principle experiment, a physical interaction network was constructed to model the interplay of metabolic pathways in yeast through the integration of array-based genomics and ICAT-based proteomics techniques[16]. This paper represents a foray into systems biology – the study and characterization of relationships and interactions between intra- and intercellular pathways and networks.

Genomics and proteomics approaches will certainly advance our understanding of basic mechanisms that are altered in the complex processes leading to carcinogenesis. New diagnostics and therapeutics will also be discovered using methods that provide global views of cellular function. However, the greatest potential for genomics and proteomics in the care of cancer patients could lie in the personalization of diagnosis and treatment. Malignant disease can now be described in such precise genetic and proteomic detail that unique and individual variations could indicate predisposition to specific carcinogen effects, forecast potential responses to chemoprevention, specify effective drug dosing and predict therapeutic outcomes. Despite the many ethical concerns, the judicious use of personalized genetic information acquired through comprehensive analyses of individual genomes and proteomes should also reduce the morbidity and expense associated with administering ineffective therapies and enable more rational designs for clinical trials.

References

1 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

2 Miki, Y. *et al.* (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 266, 66–71

3 Wooster, R. *et al.* (1995) Identification of the breast cancer susceptibility gene BRCA2. *Nature* 378, 789–792

4 Ostrander, E.A. and Stanford, J.L. (2000) Genetics of prostate cancer: Too many loci, too few genes. *Am. J. Hum. Genet.* 67, 1367–1375

5 Tavtigian, S.V. *et al.* (2001) A candidate prostate cancer susceptibility gene at chromosome 17p. *Nat Genet* 27, 172–180

6 Rebbeck, T.R. *et al.* (2000) Association of HPC2/ELAC2 genotypes and prostate cancer. *Am. J. Hum. Genet.* 67, 1014–1019

7 Jupe, E.R. *et al.* (2001) Single nucleotide polymorphism in prohibitin 39 untranslated region and breast-cancer susceptibility. *Lancet* 357, 1588–1589

8 Chee, M. *et al.* (1996) Accessing genetic information with high-density DNA arrays. *Science* 274, 610–614

9 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470

10 Alizadeh, A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511

11 Hedenfalk, I. *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *New Engl. J. Med.* 344, 539–548

12 Kennedy, S. (2001) Proteomic profiling from human samples: The body fluid alternative. *Toxicol. Lett.* 120, 379–384

13 Link, A.J. *et al.* (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 17, 676–682

14 Washburn, M.P. *et al.* (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 19, 242–247

15 Gygi, S.P. *et al.* (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 994–999

16 Ideker, T. *et al.* (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929–934

17 MacBeath, G. and Schreiber, S.L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science* 289, 1760–1763

18 Haab, B.B. *et al.* Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol.* (in press)

19 Fung, E.T. *et al.* (2001) Protein biochips for differential profiling. *Curr. Opin. Biotechnol.* 12, 65–69

20 Merchant, M. and Weinberger, S.R. (2000) Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis* 21, 1164–1177

21 Vlahou, A. *et al.* (2001) Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. *Am. J. Pathol.* 158, 1491–1502