Edison T. Liu
is the Executive Director of the Genome Institute of Singapore, one of the key research institues established by Singapore Agency for Science, Technology and Research. The Genome Institue of Singapore focuses on transcriptional genomics and gene regulation on a genome-wide scales addressing questions pertinent to human disease.

Keywords: *transcription, cDNAs, expression signatures, gene regulation, drug effects*

Edison T. Liu,
Genome Institute of Singapore,
60 Biopolis Street, #02-01,
Genome Institue of Singapore,
Singapore 128672

Tel: +65 6478 8007
Fax: +65 6478 9051
E-mail: liue@gis.a-star.edu.sg

Expression genomics and drug development: Towards predictive pharmacology

Edison T. Liu
Date received (in revised form): 6th December 2004

Abstract

Expression genomics can be defined as the study of the dynamic transciptome and its regulatory elements. Technologies are available that can assess transcripts on a genome-wide scale over time and across many samples. This comprehensive and dynamic database is being used to decipher signalling pathways and to identify new biomarkers and targets. Biomarkers emerging from these studies have prognostic potential and can be used to predict therapeutic outcome. The multiplex nature of this approach not only telescopes the time to discovery, but also allows for detection of complex interactions. Taken together, these capabilities, if carefully used, can speed drug development, enhance the identification of potent drug combinations and identify patient populations that will benefit from these new drugs.

INTRODUCTION

The enabling pharmacology that underpins modern drug development relies on the understanding of the biochemical and genetic components of a cellular state and how to manipulate this state. Biochemistry brought molecular precision to the process, but the complexity of the configurations associated with disease ultimately limited the usefulness of this one-gene/protein-at-a-time approach: it is too slow and provides insufficient information to cover all the interlocking mechanisms pertinent to a disease process or a pharmacological response. The emergence of mega-throughput sequencing capabilities, computational tools to visualise and analyse this massive body of information and the complete sequencing of the entire human genome and much of the transcriptome provided the final pieces that made genomics a fundamental platform in pharmacology and in drug development. This genomic approach is characterised by the comprehensive and complete assessment of genetic descriptors associated with specific cellular states. In this regard, the dynamic aspect of the genome, the expressed genetic elements, is among the most important determinant of pharmacological output and is therefore an important genomic read-out in pharmaceutical development.

Surprisingly, concepts arising from recent work on a remote topic have had some influence on current thinking in human pharmacology and, perhaps, in drug development. Davidson and colleagues, after identifying the critical genetic components for developmental specification for endomesodermal differentiation in the sea urchin, pursued large-scale perturbation analysis integrating gene knockdown technologies, *cis*-regulatory analysis and classical developmental biology.[1] From this analysis, they defined the genetic network for endomesodermal specification on a genomic scale.[2,3] Their ability to frame the network as a precise functional model in a control diagram makes feasible the hypothetical concept of forward engineering of a complex biological process. This became the focal point for systems biology. When applied to pharmacology, this means that computational predictions for physiological outcomes after a pharmacological challenge may be

achievable. This concept of systems biology has been embraced by some pharmaceutical companies and integrated into their organisational structure (eg integrated biology and systems biology groups at Eli Lilly). Fundamental to this approach is a focus on the dynamic nature of the transcriptome and the function of its protein products. For dynamic processes such as cell biology and pharmacology, the transcripts associated with a condition represent the most important genomic determinants of biological outcome.

The focus of this paper, therefore, is the analysis of the transcriptome relevant to drug development. Although genetic mutations and polymorphisms are also important in the pharmacological framework, these topics have been reviewed previously and will be only mentioned briefly.[4] Specifically, the area that will not be covered here will be the pharmacogenomics of drug metabolism, where polymorphisms of metabolising enzymes alter individual processing of therapeutic agents.

THE TECHNOLOGIES

Current technologies that interrogate the transcriptome and its regulation on a genome-wide scale include EST and full length cDNA sequencing, SAGE, MPSS, microarrays and chromatin immunoprecipitation

The interrogation of the transcriptome, which is the centrepiece of expression genomics and is the focus of this paper, has been enabled by a number of technologies. Expressed sequence tags (ESTs), full-length cDNAs (FLcDNA), serial analysis of gene expression (SAGE), massively parallel signature sequencing (MPSS), expression arrays and chromatin immunoprecipitation (ChIP) are among the technologies developed for this purpose.

Expression genomics, as assessed by microarrays, is both powerful and limiting. The nature of this technology has been well described[5,6] and previously reviewed.[7–9] Its utility is in the massively parallel nature of the gene expression analysis and the ability to assess many samples for comparison. The resultant expression footprints are akin to signatures of cellular states and can be used to distinguish between subtle differences. It demonstrated a hierarchy of transcriptional impact, with cell lineage making a greater impact than biochemical pathways, and, in turn, greater than individual gene effects. Arrays can discern disease classes that standard clinical assays and assessments cannot, and the expression profiles can serve as a sensitive and specific measure of pharmacodynamic response (see below). Surprisingly, the complexity of the data with the simultaneous measurement of thousands of genes, when viewed at a helicopter level, is remarkably robust and is reproducible across cellular systems and array platforms (Box 1).

FLcDNA cloning and sequencing (such as is represented in the Cancer Genome Anatomy Project (CGAP; http://cgap.nci.nih.gov/) is perhaps the most precise in gene annotation and most complete in transcript discovery, but is too costly and slow for use in answering cell biology questions.[10]

Transcript tagging technologies are surrogates for FLcDNA cloning and sequencing. SAGE and MPSS can detect new transcripts in an unbiased manner. SAGE was a technology designed digitally to quantify expression of genes designated by short sequence tags representing cDNAs. SAGE, unlike microarrays, can assess the expression of unknown or unexpected genes and therefore can be used for gene discovery. The basis of this technology is to render each cDNA into a representative short 14–21 base pair (bp) tag, which can be concatemerised for ease of sequencing. MPSS couples a SAGE-like approach with a novel restriction–ligation bar code identification and a bead-based detection system to identify 17–20 bp tags of transcripts in any sample.[11] SAGE and MPSS can assess transcript diversity deeply by sequencing many tags, but has significant 'noise' in the data output: ~30–40 per cent of tags cannot be mapped to a single genomic location. Moreover, these approaches do not assess transcript processing/splicing, and are sufficiently cumbersome to preclude widespread use. Thus, they are

Technologies in expression genomics

Sequencing-based

Expressed sequence tag: Single-pass sequence of cDNA clones, primarily assessing the 3′ region: 300–800 base pairs (bp) per tag.

Full-length cDNA library: Full-length cDNA clones are enriched and the complete sequence of each clone determined. This is the gold standard for transcriptome technologies, in that all aspects of a transcript can be ascertained through this one approach.

Serial analysis of gene expression) (SAGE) and long SAGE: Short tag sequences (14 bp or 21 bp tags per transcript) are extracted, concatenated, cloned and sequenced. This covers 30–50 tags per sequence read. The 14 bp tags are representative of individual transcripts, although significant ambiguities of gene assignment are found. The long SAGE tags encompass up to 21 bp tags and are sufficiently specific to be mapped directly to genome sequences. Coupling 5′ long SAGE and 3′ long SAGE approaches in a single sample allows for the computational reconstruction of the start and end of every transcript. These technologies can only assess gene expression, but in an unbiased fashion.

Massively parallel signature sequencing: The 3′ sequences of individual cDNA clones are ligated to short signature tag sequences (17–20 bp tags per transcript) linked *in vitro* onto individual and addressable microbeads. Parallel sequencing is conducted by a hybridization–ligation-based approach, using fluorescently labeled probes for the individual nucleotides. The depth of sequence information that can be extracted from any one library is significantly greater than the output from a SAGE library.

Microarray-based

Deposition microarrays: Oligonucleotides or polymerase chain reaction (PCR) products from cDNA clones are spotted onto glass slides using robotic dispensers. Each spot contains a specific probe for a gene, and the data output provides only the levels of expression of individual putative transcripts and can only assess which genes are represented on the array. The advantage is the robustness and low cost of the system.

In situ **DNA oligonucleotide synthesised microarrays:** *In situ* synthesised oligonucleotide microarrays employ the photolithographic synthesis of oligonucleotide probes directly on the chip (eg Affymetrix; http://www.affymetrix.com/technology/index.affx). The advantages of this technology are the density of the arrays and the economies of scale.

PCR-based

Chromatin immunoprecipitation: A transcription factor is chemically cross-linked to its bound genomic DNA fragment. The DNA is sheared and the transcription factor–DNA complex is immunoprecipitated with the protein-specific antibody, and enrichment of the bound fragments is detected by quantitative PCR. This technology allows for the direct assessment of binding sites of transcription factors. When coupled with cloning/sequencing or tiled array approaches, the binding sites of any transcription factor can be mapped on a whole-genome scale.

Box 1: Technologies in expression genomics

New technologies combine components of earlier approaches, like chromatin immunoprecitation-tiled assays, or full length cDNA synthesis-SAGE

less useful when applied to cell biological or clinical experiments, where analysis of many individual samples is required.

Recently, new technologies have been developed to overcome the limitations in the completeness and speed of discovery of the dynamic transcriptome. Tiling arrays, whereby every nucleotide in the genome is covered, have been used to assess all possible transcribed sequences.[12] The results show that much more of the genome is transcribed than was previously imagined. Moreover, the dynamic binding of transcription factors to the genome can be comprehensively assessed using these tiled arrays in chromatin immunoprecipitation onto chip experiments.[13] Again, the surprise is how often transcription factors bind in places that we did not expect. Specific exon-junction arrays covering more than 10,000 multi-exon genes have revealed significant splice heterogeneity across tissues and cell lines.[14] It was estimated

that 74 per cent of multi-exon human genes are spliced. Wei *et al.*[15] coupled a 5′-long SAGE and a 3′-long SAGE technology to annotate completely the putative start and ends of every transcript. This is a significant boost to FLcDNA library production and clone sequencing because of the speed of sequencing in a SAGE format. It is also an improvement over 3′-directed SAGE approaches because of the increased information content from the 5′ annotation.

With these new technologies, the full identification of the dynamic transcriptome can be realised.

EXPRESSION GENOMICS AND DRUG DISCOVERY

The drug discovery process involves target identification; target validation; lead identification and optimisation; and preclinical studies and toxicology, leading to clinical trials. Genomic approaches have been used at each stage of this pipeline to significant advantage.

Finding the target, uncovering pathways: Somatic mutations

Specific somatic mutations in tumours function as targets for therapeutics or predictors of treatment response

Since cancer can be propelled by activating mutations in signalling molecules, early strategies identified specific somatic mutations amenable to targeting by small molecule therapeutics. Cytogenetic rearrangements directed the first successful example of a therapeutic targeted to an oncogene. The BCR–ABL translocation in chronic myelogenous leukaemia activates the ABL kinase by the generation of a chimeric protein. Gleevec was synthesised as a small molecular inhibitor of the ABL kinase, and proved to be remarkably effective.[16] Point mutations were previously more difficult to assess but also identified markers for therapeutic response or yielded potential targets. One of the first examples was the analysis of RAS mutations (N-RAS, H-RAS and K-RAS). Most studies in epithelial cancers found no association between activating RAS mutations and response to therapy; however, fortuitous observations in acute myelogenous leukaemias (AMLs) suggested that aberrations in RAS rendered AMLs more sensitive to dose-intensive cytosine arabinoside,[17] findings supported by *in vitro* studies.[18,19] The observation that mutations in the receptor tyrosine kinase, FLT3, are common in AML (25–30 per cent) and are associated with poor prognosis[20] led to the development of specific anti-FLT inhibitors with therapeutic effect.[21,22]

Activating BRAF mutations are found in malignant melanomas

Transforming growth factor-beta (TGFbeta) is a growth-suppressing ligand for its cognate receptor, TGFbetaRII, and abrogation of TGFbeta signalling is thought to be involved in epithelial carcinogenesis. It has been observed that a dinucleotide repeat in the coding region of TGFbetaRII is susceptible to mutations in colon cancer patients exhibiting the microsatellite instability phenotype first found in hereditary non-polyposis coli.[23] These mutations inevitably generated termination codons, rendering the receptor inactive. Thus, even without application of genomic approaches, mutations in key signalling genes have pointed to genes and pathways amenable to drug targeting. The strategy, however, involved a gene-by-gene analysis which was reliant on chance findings.

Recently, more genomic approaches have been successfully applied to target discovery in cancer. Characteristic of this approach has been the surveying of entire classes of molecules across a large number of samples in an attempt to discover genetic variances. Significant advances in sequencing technologies that enable large-scale sequencing at low cost have encouraged several groups to embark on exon sequencing of candidate signalling genes. Davies *et al.* from the Sanger Institute sequenced the codon domains of RAS, RAF, MEK and the mitogen-activated protein (MAP) kinase pathway in a large number (>500) of cell lines.[24] This was the initiating effort in a larger programme to sequence the exons of all known genes in cancer cell lines. These authors found somatic missense mutations in the serine kinase, BRAF, in malignant

melanomas and confirmed that 66 per cent of primary malignant melanomas harboured activating mutations in the kinase domain, with a single substitution (V599E) accounting for 80 per cent. This has led to the development of BRAF-specific inhibitors in cancer therapy.[25]

A similar strategy has been pursued in the analysis of protein kinases in cancers. The importance of these candidate targets has been primarily based on the clinical effectiveness of anti-HER-2 and epidermal growth factor receptor (EGFR) therapy. The original focus on these anti-kinase drugs was based on the overexpression of HER-2/ERBB2 and EGFR in primary cancers, but, again, the advances in sequencing technologies now allow for large-scale interrogation of mutations in tumours. The results have been striking. Paez *et al.* examined sequence aberrations in EGFR in human lung cancers and found that an activating mutation is seen in about 21 per cent of adenocarcinomas.[26] These mutations in the kinase domain were surprisingly high in adenocarcinomas of the lung from Japanese women when compared with Caucasians (32 per cent vs 3 per cent). Intriguingly, these mutations appeared to confer a better response to gefitinib, an anti-EGFR small molecule.[26,27] The Sanger Institute group further extended this sequencing strategy to the HER-2/ERBB2 gene in lung cancers and found that 10 per cent of adenocarcinomas and 4 per cent of non-small cell carcinomas of the lung had mutations in the kinase domain.[28] Anti-HER-2/ERBB2 therapeutic antibodies have been ineffective in treating lung cancer, but these target the overexpressed form of the receptor. These results suggest that small molecule inhibitors for HER-2 should be clinically re-evaluated in the specific subset of lung cancer patients whose tumours harbour HER-2/ERBB2 mutations. Intriguingly, several pharmaceutical companies have mixed kinase inhibitors directed at both EGFR and HER-2/ERBB2. Sequence observations in lung cancer suggest that dual kinase inhibition might have a particularly effective role in adenocarcinoma of the lung.

Activating mutations in EGFR predict for response to anti-EGFR small molecule therapy

Underscoring the importance of tyrosine kinase activation in human cancers, Wang *et al.* scanned epithelial cancers for mutations in six protein tyrosine phosphatases (PTPs) and found that 26 per cent of colorectal cancers have mutations in at least one of these PTPs, with a large number being inactivating or attenuating mutations.[29] Whereas protein tyrosine kinases (PTKs) activate pathways through phosphorylating protein substrates in tyrosine residues, PTPs remove these phosphates and therefore act negatively to regulate downstream kinase effects. Inactivating mutations in the PTPs would lead to augmented PTK action in the cancer cells.

Thus, in these examples, genomic approaches have dramatically increased the discovery rate of potential new targets for cancer therapies by screening the coding exons of 'druggable' targets such as protein kinases.

Finding the target, uncovering pathways: Transcriptional footprints

Transcriptional profiling provides a footprint of many biochemical pathways

Mutational analysis is a specific strategy, for identifying a specific target molecule (eg FLT3 or BRAF mutations) for therapeutic intervention; however, the pathways involved are inferred. Analysis of the transcripts has also identified specific protein targets. Recall that the original development of anti-HER-2/ERBB2 and EGFR inhibitors was based on the overexpression of these proto-oncogenes, which was sometimes coupled with amplification in primary tumours. The multiplexed analytical approaches provided by expression microarrays, however, allow for a more sensitive sensor of pathway activity. Although not all biochemical pathways are regulated transcriptionally, genome-wide expression profiling has been successful in deciphering many pathway interactions and in uncovering new pathways associated with biological processes.

Expression assays can be used for gene discovery in cancer biology or to identify class distinctions amongst tumours

Expression signatures can give an indication of the hierarchy of impact of a cellular process or a drug is on or off target

Again, work in lower organisms provided the proof of principle that on-target/off-target determination can be assessed with surprising precision. Hughes *et al.*[30] used *Saccharomyces cerevisiae* as the test organism and a whole genome expression array as the read-out in testing the expression footprint of yeast strains after chemical challenge. These challenges were therapeutic molecules, many of which had no primary applications in microbial conditions. Nevertheless, what was found was that therapeutic molecules affecting specific targets would exhibit little transcriptional change when the target was deleted. Molecules with non-specific effects would register alterations in the expression cassette.

Miller *et al.*,[31] while studying the transcriptional response of rat pituitary cells to thyroid hormone (T3), uncovered an unexpected link to cancer biology. A time course analysis after T3 exposure revealed a coordinated inhibition of the wnt pathway: downregulation of disheveled, T cell factor (TCF) and beta-catenin, and upregulation of axin and adenomatous polyposis coli (*APC*). This was confirmed by the disappearance of beta-catenin protein levels in T3-treated cells. Conversely, we would expect that the inhibition of T3 signalling would result in a net upregulation of *wnt* signalling. In fact, *v-erbA*, a retroviral oncogene encodes the truncated and inactive form of the thyroid hormone receptor that functions as a dominant negative mutant. Based on the array investigation, the net result of *v-erbA* expression would be to augment *wnt* signalling, which is oncogenic.

One of the more effective means of using expression arrays to dissect specific pathways is to challenge cells experimentally and assess changes over time, or to compare two isogenic cell lines that differ by a gene disruption. This is especially useful when both conditions are employed. Guo *et al.*[32] assessed the serum response of wild-type myc and myc-null cells and found predominant induction of genes involved in protein synthesis. This was consistent with observations that one of the primary roles of myc is to regulate cell size and augment protein synthesis.[33] Aprelikova *et al.*,[34] using BRCA1 wild-type and knockout cells, uncovered a mechanism whereby BRCA1 and p53 specifically regulate the expression of the stress response gene, 14-3-3σ. Kho *et al.*[35] used a 19,000 genetic element array to uncover approximately 230 genes that were p53 dependent, responsive to 5-fluorouracil exposure and were associated with cell death. Further investigation revealed that some of the genes repressed by p53, polo-like kinase (*PLK)* and pituitary transforming gene-1 (*PTTG1*), have a significant role in cell survival after genotoxic stress.

On occasion, this approach yields potentially new markers for medical applications. The engagement of surface receptor tyrosine kinases activates phosphoinositide-3-kinase and one of its downstream effectors, AKT protein kinase B, which in turn can activate molecular targets of rapamycin (*mTOR*). *mTOR* is a key signalling node involved in the initiation of protein synthesis via the ribosomal protein S6 kinase and eukaryotic initiation factor 4E-binding protein pathways, transcription and protein stability. Majumder *et al.*[36] used transgenic mice expressing *AKT1* in the prostate gland (AKT1-Tg) to show that pharmacological inhibition of *mTOR* effectively inhibits the development of prostatic intraepithelial neoplasia. They analysed the expression cassette of the AKT1-Tg prostate glands before and after *mTOR* inhibition and found that *HIF-1a* and many of its downstream targets — including glycolytic enzymes and the glucose transport gene *GLUT1* — were upregulated by AKT and downregulated after exposure to an *mTOR* inhibitor. Since a common cancer imaging approach is positron emission tomography using [18F] fluorodeoxyglucose (FDG), it is speculated that FDG uptake can be an effective clinical measure of the pharmacodynamic efficacy of therapeutic mTOR inhibitors.

Even in the absence of widespread expression differences, expression arrays can uncover important truths. Chen *et al*.[37] examined seven pairs of androgen-sensitive and -resistant xenograft tumours and asked if there are expression differences that could explain the emergence of androgen resistance. Of the 12,559 gene probes, they found that only one gene in their array was consistently altered in the resistant tumours: increased expression of the androgen receptor. Subsequent biochemical and biological experiments confirmed this to be the primary mechanism for androgen resistance.

Different genetic origins of tumours inhibit distinguishable expression signatures

In studying the biology of other nuclear hormone receptors, microarray analysis has enabled unique observations to be made. Some of these observations are pertinent to the discovery of new drugs. Focusing on oestrogen receptor (ER) biology, several groups have identified the universe of genes induced or repressed by oestrogen.[38–40] It was found that the ER response element (ERE) is the primary target for the ER, despite the observations that AP1 and SP1 sites can be the targets for ER regulation.[40] The fact that a large number of genes are oestrogen responsive and that many of these genes are involved in growth regulation is not novel; however, the observation that only a small number of genes (~89) out of 19,000 represented on an array are direct targets of ER suggests that the primary transcriptional targets for ER are limited. The concordance between the array results in cell lines is significant and points to some common transcriptional rationale. When the ER expression cassette is compared with the genes that determine ER status in breast cancers (see below), there is indeed more overlap than expected by chance alone.[41] Thus, despite the noise in the system, the ordered control of gene expression by ER is relatively consistent across several human cell systems. Interestingly, the near ubiquity of ERE in the genome, and the fact that many EREs bind to ER, suggest that there are more factors involved in ER signalling than simply the presence of even a perfect ERE. Using EREs validated by ER binding, it has been found that specific sequences within a 250 bp proximity of the core ERE can significantly improve the prediction of ER binding (paper in preparation). The complexity, yet consistency, of the genomic level read-out provided by these arrays enables the reclassification of selective ER modifiers based on their DNA-binding patterns and gene expression profiles. The on-target/off-target considerations can be parsed with high resolution.

Oncogenic pathways have a greater impact on creating transciptional profiles than individual genes

The logical extension of this concept is that specific oncogenic origins of any cancer may be marked by a specific expression fingerprint. If this is true, then, theoretically, any discernible expression cassette can be used to trace a specific genetic mutation of a tumour. Experimentally, it is not possible to prove this with human tumours; however, Desai *et al*.[42] sought to resolve this using transgenic murine models of mammary cancers. They examined the gene expression patterns of mammary carcinomas from mouse transgenic lines bearing one of six oncotransgenes driving expression to the mammary gland (MMTV-*Ha-ras*; MMTV-*neu*; MMTV-polyoma middle T antigen; WAP-SV40 large T antigen; C3T-SV40 large T antigen (SV40LT); and MMTV-*myc*), and found that the tumours arising from mutant *Ha-ras*, *neu* and polyoma middle T (PyMT) antigen transcriptionally behave as a cluster. The tumours from animals expressing SV40LT antigen clustered together regardless of the promoter used, and those from the *myc* transgenics were represented in another distinct cluster. *ras*, *neu* and PyMT converge on the *ras*-MAPK pathway, and the mechanisms of action for SV40LT and *myc* are distinct. Taken together, this suggests that oncogenic pathways have primacy in defining downstream expression patterns over the effects of individual genes. The mammary cancers from the different oncotransgenes were associated with the

BRCA1 specifically alters expression of genes on chromosome Xp

induction of specific gene cassettes which may be associated with certain phenotypic differences. For example: calcium signalling pathways with the SV40 LT transgenic tumours; ribosomal RNA and *Notch* pathway genes with the *myc* transgenic tumours; and cyclin D1, cdk-2 and E2F with the *ras*-*neu*-PyMT group.

The observation that specific genetic points of origin in tumours give rise to discernable expression profiles has also been seen in human tumours. Jazaeri *et al.*[43] studied the transcript profile of ovarian cancers from carriers of BRCA1 or BRCA2 germline mutations and compared them with sporadic ovarian cancers. Biochemically, BRCA1 and BRCA2 appear to be primarily involved in DNA repair. Moreover, sporadic ovarian cancers harbour no mutations in either BRCA1 or BRCA2. Based on this information, the expectation is that BRCA1 and BRCA2 ovarian cancers would resemble each other, but would be different from sporadic cancers. Surprisingly, however, their results revealed that BRCA1 and BRCA2 ovarian cancers were distinct from one another and that all the sporadic cancers could be divided into those with a BRCA1-like profile and those with a BRCA2-like profile. Thus, tumours with BRCA1 and BRCA2 points of origin may engage different downstream pathways. This was also observed in breast cancers arising in BRCA1 and BRCA2 carriers,[44] although the results were more confounded by the different ER status and grade of the tumours. Interestingly, Jazaeri *et al.*[43] noted that an unusual number of genes on chromosome Xp11, including some known to be involved in maintaining ovarian structures, were consistently upregulated only in BRCA1 mutant tumours. They suggested that BRCA1 might act to repress genes on regions of the X-chromosome. This was corroborated by *in vitro* observations by Ganesan *et al.*,[45] who found that BRCA1 colocalises with non-coding *XIST* RNA that coats inactivated X-chromosomes, and that BRCA1 is needed for X inactivation. More recently, Jazaeri[46] further confirmed that when a BRCA1 construct was introduced into a BRCA1-negative cell, genes on chromosome Xp were specifically downregulated by BRCA1. Taken together, these results show that, despite the biochemical similarities of BRCA1 and BRCA2, in that they are both involved in DNA repair and homologous recombination, their transcriptional footprints are different and may generate different carcinogenic cascades that are definable with high resolution using microarrays. Moreover, this collection of studies shows that the strategic use of array approaches can uncover completely new associations surrounding one key gene.

In another example of pathway assessment in primary cancers, Ferrando *et al.*[47] explored pathway discovery in one of the cancers most tractable for molecular studies: acute leukaemia. Using expression microarrays, they found that expression signatures of T cell leukaemias were consistent with the arrest of these leukaemic cells at specific stages of normal thymocyte development. Specifically, they were able to identify HOX11L2 (TLX3) activation as an important event in T cell leukaemogenesis, conferring a poor response to treatment. This is surprising, given the positive prognostic effect of another HOX gene, HOX11 (TLX1).[48]

CLINICAL STRATIFICATION: PATIENT SELECTION FOR OPTIMAL THERAPY

Expression profiling, prognostic subsets and markers of tumour behaviour

The notion that genetic or biochemical markers in tumours can segregate patients into therapeutic subsets has been well established in clinical pharmacology. There are host genes that reflect the host's tolerance of therapy, as in the case

of thiopurine *S*-methyltransferase activity in determining haematopoietic toxicity in childhood leukaemias.[49,50] An even greater literature exists on the marker configurations of cancer and how they relate to treatment efficacy.[51] Notables are HER-2/neu in breast cancer and bcr/abl in chronic myelogenous leukaemia, both markers for clinical outcome and specific tumour characteristics and, equally importantly, both now targets for specific and effective therapeutics (see above). From these monogenetic beginnings has arisen the current format for discovery: highly multiplex and parallel ascertainment of biomarkers using EST sequencing or expression microarrays.

Expression genomics can define specific classes of tumours associated with clinical behaviour

Some of the best information relating expression profiles and prognosis are in breast cancer, the leukaemias and lymphomas. In primary breast cancer, Perou *et al.*[52] analysed gene expression patterns in normal and malignant human breast tissues from 42 individuals. Their analysis revealed two major subtypes based on the expression characteristics of 496 selected genes. One subgroup was characterised by tumours distinguished by the expression of genes normally expressed by breast luminal cells and were clinically associated with ER-alpha positivity. The second subgroup was characterised by tumours predominantly expressing genes associated with breast basal/myoepithelial cells (keratins 5 [*KRT5*] and 17 [*KRT17*], laminin and fatty-acid binding protein 7, secreted frizzled-related protein 1 [*SFRP1*], the oncogene *c-kit* and lower expression of fibronectin 1 [*FN1*] and mucin 1 [*MUC1*]) and that were mainly ER negative.

Oestrogen receptor positive breast tumours exhibit greater heterogeneity in gene expression patterns and clinical outcome than oestrogen negative tumours

Subsequently, it has been consistently observed that the major factor that appears to affect the expression profile-based class of breast cancer is the ER status. Sorlie *et al.*[53] refined this classification by analysing a larger number of breast tumours analysed by similar microarray cDNA platforms. They observed that the luminal/ER+ tumours could be further subdivided into at least two subgroups with distinct molecular signatures. Luminal subgroup A was characterised by the highest level of ER expression, as well as by a high expression of GATA-binding protein 3, hepatocyte nuclear factor 3α, X-box binding protein 1, trefoil factor 3 and LIV-1. The second subgroup (B+C) showed lower expression of the luminal-specific genes. Importantly, these array-based subgroups could discern prognostic differences. The basal-like/ER$^-$ subgroup had the shortest relapse-free and overall survival, but the luminal subtypes could provide prognostic differences not previously observed: luminal subgroup A showed the best clinical outcome, with an almost 80 per cent survival, whereas the luminal B+C subgroup had approximately half this survival rate.

In analysing a cohort of 99 breast cancer patients, Sotiriou *et al.*[54] also found that the ER status of the tumour was the most important discriminator of expression subtypes. Tumour grade was a distant second, whereas lymph node positivity, tumour size and menopausal status were not associated with clear expression patterns. The larger ER-positive subgroup was again similar in expression profile to the luminal-like subtype described by Sorlie and Perou, and the ER-negative subgroup was more akin to the basal-like subtype. Interestingly, despite the differences in patient populations, treatment and array platforms used, the survival outcomes of the array subgroups were remarkably similar to those in these two comparable studies: the luminal-like subgroup had better relapse-free and breast cancer survival when compared with the basal-like tumours. Three subgroups were found within the luminal-like (predominantly ER-positive) cluster that showed distinct differences in survival. Luminal-1 had the best outcome, with an 80 per cent ten-year relapse-free survival, and was correlated with lower grade tumours and higher expression of *c-kit*, hepatocyte growth factor (*HGF*), insulin-like growth factor-binding protein-3, activating transcription factor-3 (*ATF-3*),

Transcriptional profiling can be used to avoid unnecessary administration of adjuvant therapies in breast cancer

as well as components of the *AP-1* transcriptional factor. There was, however, attenuated expression of cell growth-associated genes. The luminal-2 subgroup had the worst outcome, with a ten-year relapse-free survival of 40 per cent, and was characterised by higher expression of tumour necrosis factor receptor-associated factor 3 (*TRAF3*), *RAD21*, BRCA1-associated protein 1 (*BAP1*), a protein tyrosine phosphatase type IVA member (*PTP4A2*) and lower expression of *CXCR4*, *ATF-3, FGFR1* and *VCAM1*. The luminal-3 subgroup had an intermediate survival outcome of 60 per cent at ten years. Intriguingly, although *ERBB2* overexpression is a known and powerful prognostic factor, there was no clustering of *ERBB2*-overexpressing tumours in any of the luminal subgroups to account for the differences in clinical outcome. All of these subgroupings were confirmed in a 'meta-analysis' of three independent datasets of detailed array and clinical data.[55]

Two reports from the Netherlands Cancer Institute[56,57] used expression arrays to identify marker clusters that could predict for clinical outcomes such as relapse and death from cancer. Using a 25,000 genetic element array, 70 genes were found whose expression profile could segregate untreated, node-negative patients into prognostic groups in terms of relapse and death. Their 70-gene classifier performed better than the standard St Gallen and the National Institutes of Health consensus clinical prognostic indicators. In a follow-up study with 295 node-negative and node-positive patients, many of whom also received adjuvant therapies, this 70-gene classifier could identify those with odds ratios >5 for developing distant metastases.

These studies show that expression profiling can consistently identify groups of patients who will have a poor outcome. Moreover, they have uncovered a hierarchy of biological effects for these markers: the ER status and grade of the tumour appear to have a greater influence on the expression profile of breast cancers than do nodal status or tumour size. The finding that the same expression profiles are associated with specific clinical outcomes, regardless of stage of diagnosis, suggests that the metastatic potential of a tumour is decided early in the course of the disease. Many of these markers are already being developed into prognostic 'arrays' for clinical breast cancer use.[58]

The experimental strategy used in the breast cancer studies discussed delved into the cluster of genes that defined tumour classes and did not focus on the clinical associations of individual genes. Studies on the expression genomics of prostate cancer have been more centred on using expression arrays to identify specific markers of prognosis. Singh *et al.*[59] assessed the expression patterns of 52 prostate cancers and found a set of genes that correlated with a standard measure of differentiation, the Gleason score. Interestingly, a minimal group of five genes (*ITPR3*, *sialyltransferase I*, *PDGFR-beta*, *chromogranin A* and *HoxC6*) were identified that could predict for relapse. Dhanasekaran *et al.*[60] examined a smaller number of primary prostrate cancers and normal prostates. They found a consistent downregulation of PTEN and gelsolin (also affected in mammary/breast cancers), upregulation of c-myc and upregulation of hepsin, a transmembrane serine protease, and pim-1, a serine/threonine kinase. These findings were confirmed for hepsin and pim-1 by using tissue arrays scanning over 700 prostate cancers: low hepsin or high pim-1 expression was correlated with bad outcome. In a further extension of these observations, this same group found that the increased expression of EZH2 was associated with the metastatic state in prostate cancer. EZH2 is the human homologue of the *Drosophila* protein enhancer of zeste (E[z])2, a polycomb protein involved in homeotic gene expression during development. EZH2 is thought to be a chromatin-associated regulator of gene expression and siRNA

The key cytogenetic drivers for expression signatures in acute myclogenous leukemia are *AML1/ETO*, *CBF* β/ *MYH11*; and *PML/RAR* α

knockdown of EZH2 expression inhibited cell proliferation.[61] Thus, expression arrays were used as a screening tool for gene discovery, and the individual marker relevance was validated by exploiting tissue arrays and in detailed *in vitro* experimentation.

Because of the ease of tissue access — and a good understanding of normal and malignant biology — malignant lymphomas and acute leukaemias have been especially informative diseases to study using these genomics approaches. In an elegant series of studies, the St. Jude group investigated the comprehensive expression architecture of over 300 childhood acute lymphoblastic leukaemia (ALL) cases.[62–64] They found that the leukaemias could be classified into distinct subgroups that correlate with histological and cytogenetic abnormalities such as T-ALL, hyperdiploid with >50 chromosomes and *BCR-ABL*, *E2A-PBX*, *TEL-AML1*, and *MLL* rearrangement. In addition, these genome-wide approaches uncovered a new leukaemic subgroup that lacked consistent cytogenetic abnormalities. Where the array-based classification and the cytogenetic assignments disagreed in the assignment of the *TEL-AML1* subclass, more detailed molecular testing always confirmed the array assignment by finding cryptic genetic rearrangements of the TEL transcription factor. Not only did this verify the array-based classification, but also suggested that the important component of the rearrangement is in the TEL locus.

Expression profiling describes molecular subclasses in non-Hodgkins lymphoma associated with outcome after treatment

In leukemias, expression profiling can identify a hierarchy of impact of specific

More recently, two groups refined these analyses further. Bullinger *et al.*[65] examined 116 AML samples and derived a 133-gene prognosis classifier that could predict survival. Uniquely, this prognosis classifier could predict outcome even in patients with normal karyotypes. Valk *et al.*[66] uncovered expression-defined gene clusters associated with specific cytogenetic abnormalities that could predict cytogenetic configuration in a separate validation set. The results from both studies are surprisingly similar, with a definable hierarchy of impact of specific cytogenetic configurations. t(8;21) — *AML1/ETO*; inv(16) — *CBF*β*/MYH11*; and t(15;17) — *PML/RAR*α defines three separate clusters requiring only a minimal number of genes to define that cluster — for example, *ETO* for t(8;21), *MYH11* for inv(16) and *HGF* for t(15;17); however, all other cytogenetic classes were less distinct.

Taken together, recurrent observations arise from these array studies: expression profiles are consistently associated with specific cancer phenotypes; there is a hierarchy of biological effect for these biomarkers; and the expression profile can be more precise in defining cancer subtypes than single markers alone.

Predicting response to therapy

In primary tumours, progress in identifying expression patterns predictive of therapeutic response have been most advanced in malignant lymphomas and in breast cancer. The predictive potential of arrays in leukaemias, discussed above, is primarily because of the known associations of cytogenetic class with therapeutic response. Alizadeh *et al.*[67] compared the patterns of gene expression of primary diffuse large B cell lymphomas (DLBCL) with those of normal lymphoid cells at different stages of differentiation. They found that DLBCL could be divided into at least two major expression subclasses, one that resembled activated normal B-lymphocytes with a distinctly poor prognosis, and another that resembled germinal centre B-cells with a good prognosis. These studies were extended with the Lymphoma/Leukemia Molecular Profiling Project,[68] which increased the analysis to 240 patients with DLBCL. This later study confirmed the association of an activated B cell-like subgroup with poor prognosis (median survival ~ two years), and of a germinal centre B cell-like subgroup with good prognosis (median survival ~ seven years). They also identified a third subgroup from expression arrays, whose survival profile (median survival ~ two years) was

cytogenetic abnormalities

similar to that of the B cell-like subgroup. In total, five patterns of expression signatures were discerned; the MHC class II signature, germinal centre signature and lymph node signature were associated with a good prognosis, whereas the proliferation signature and genes in 'another' category were associated with a poor prognosis. These results can be interpreted as being predictive of response to standard chemotherapy, since all patients received doxorubicin-based combination chemotherapy. Shipp *et al.*[69] specifically investigated those profiles associated with survival and identified a set of 13 genes that divided 58 patients with DLBCL into a good outcome group (five-year overall survival of 70 per cent) and a bad outcome group (five-year overall survival of 12 per cent). The overlap of some of these genes (*NOR1*, *PDE4B*, *PKC-beta*) with the prognostic gene set of Alizadeh *et al.*[67] raised the possibility that these genes may have a significant role in the biology of human lymphomas. Of note, small molecular inhibitors against protein kinase C-beta (PKC- beta) have been developed and appear to be effective against xenograft epithelial tumours.[70–72] Informed by these array results, clinical trials are ongoing to investigate the efficacy of these inhibitors in lymphomas.

The number of genes whose expression is altered after chemotherapy is an indicator of probability of response

The fact that the expression behaviour of a gene is associated with sensitivity to chemotherapeutic agents *in vitro* has been well documented using the NCI 60 cell lines.[73–75] Kikuchi *et al.*[76] furthered this concept to primary cancers by comparing the expression profiles of primary lung cancers with their *in vitro* behaviour in a cell-based assay for chemotherapeutic sensitivity. They observed that the expression levels of two previously uncharacterised genes were highly correlated with sensitivity to CPT11 and to gemcitabine.

Once again, investigators in the St. Jude group exploited their understanding of the pharmacology of treatments for childhood ALL and used expression arrays in uncovering gene cassettes associated with primary drug resistance. They tested the response of leukaemia cells from 173 patients (on whom they had detailed information on the *in vitro* sensitivity of the cells) to the most common anti-leukaemia therapies using an Affymetrix array platform for patterns of gene expression.[77] The array results were then progressively correlated with the sensitivity profile for each of the four compounds vincristine, asparaginase, prednisolone and daunorubicin. Using specific expression cassettes correlating with the sensitivity of each individual agent, they were able to predict *in vivo* treatment outcome in a validation set of leukaemia patients. Intriguingly, 121 of the 124 treatment-outcome genes had not previously been associated with drug resistance.

An extension of the therapeutic question is whether the response of the tumour cells after chemotherapeutic exposure could predict future therapeutic response/outcome. Sotirou *et al.*[78] addressed this in a clinical setting by examining the expression profiles of ten untreated patients undergoing doxorubicin-based chemotherapy for primary breast cancer before and after the first cycle. Samples were taken using fine needle aspirates. Five had partial remissions (poor responders) and five patients achieved clinical complete remission (good responders). Although very few genes in the pretreatment samples could distinguish good from poor responders, the expression profile of the tumour aspirates 21 days after the first cycle of chemotherapy was predictive of response: the good responders had ten times the number of gene outliers than the poor responders. Thus, temporal changes in gene expression after chemotherapy can be used to quantify pharmocodynamic response. These authors noted upregulation of genes implicated in differentiated states such as laminin, *TIMP1* and *CDK9*, and downregulation of cell proliferation genes such as that encoding minichromosome maintenance protein 2 (*MCM2*). Cheok

et al.[79] took this observation further and assessed the *in vivo* response of lymphoid leukaemic cells in patients undergoing methotrexate and mercaptopurine therapy. They found 124 genes that discriminated between the different treatments — that is, the gene expression changes were treatment specific, and the effect of combination therapy was not the same as the summation of single-agent treatment.

Bani *et al.*[80] expanded these studies by exploring the genome-wide expression changes in a controlled xenograft model of ovarian cancer after paclitaxel administration. The majority of gene expression changes occurred 24 hours after the paclitaxel dose and, again, therapeutic response was correlated with the number of genes that changed. In responding tumours, genes involved in cell-cycle regulation and cell proliferation (*CDC2*, *CDKN1A*, *PLAB* and the gene encoding topoisomerase IIalpha [*TOP2A*]) were coordinately perturbed in the direction of growth cessation. Metabolism genes were uniformly downregulated, as were genes involved in interferon-mediated signalling (*G1P3*, *IFI16*, *IFI27*, *IFITM1* and *ISG15*). Extending this to a larger panel of cell lines, these authors found that the most consistent expression markers of drug sensitivity were the upregulation of CDKN1A (p21/Cip1) and the downregulation of *TOP2A*, directions of gene expression associated with cell-cycle arrest. Thus, in both the human clinical situation, as well as in xenograft models, gene expression evidence of a cessation of proliferation appears to be associated with good therapeutic outcome.

HOST ONCOGENOMICS: GENETICS OF TUMOUR BEHAVIOUR

The genetic make-up of an individual can alter tumour behaviour associated with changes in expression signatures

Classical pharmacogenetics studies investigated the genes that modulated drug effects primarily by altering drug metabolism.[81] These genetic variations would lead to differences in therapeutic response. The aspect that previously has not been well considered, however, is the role of germline genetics on the primary behaviour of a tumour. It was assumed that the virulence of a cancer, for example its metastatic potential, was a result of environmental exposure, or from stochastic events.

Hunter[82] questioned this logic by asking whether by keeping the oncogenic inducer constant, one could assess the effect of distinct host genetic factors on the phenotypes and the expression profiles of the resultant cancer. FVB/NJ animals bearing the MMTV-PyMT oncotransgene were crossed to five different strains. The transgene-bearing F1 mice all developed mammary cancers, but with distinct phenotypes of different tumour latencies, growth rates and metastatic potential. By examining the expression profiles of a large number of these tumours, Qiu *et al.*[83] found that host genetic backgrounds can alter the downstream expression profiles of the cancers, and that the tumours from the different backgrounds could be further grouped together: 1) tumours derived in the LP/J F1 and MOLF/Ei F1 strains, where tumour growth and dissemination are suppressed and latency prolonged; 2) the most aggressive tumours from the FVB/NJ parental strain and I/LnJ F1 genomic backgrounds; and 3) an intermediate virulence phenotype with tumours from NZB/B1NJ F1 crosses. Intriguingly, of the 17 genes whose expression was previously suspected to define a metastasis phenotype in human cancers,[84] 16 murine orthologues in this study also predicted the ability to develop lung metastasis in the PyMT model of mammary cancer .[83] More generally, one advantage of this animal model for breast cancer is that defined normal breast can be used as a reference. Thus, all experiments described here used RNA from the mammary glands of 16-week virgin female animals as a reference. With this design, those genes associated with mammary gland transformation can be consistently identified. Interestingly, a significant number of genes and their

The gene expression signatures of cellular toxicity may be able to predict the toxicity of a new compound

related members found to be perturbed in mouse mammary tumours have also been reported to be altered in human breast cancers.

These findings suggest that the germline genetic make-up of an individual may have a significant role in determining the virulence of a primary cancer. Such virulence factors as tumour grade or invasiveness would have an impact on the choice of, and the responsiveness to, specific therapeutics.

PHARMACOTOXICITY

Genetic polymorphisms in metabolising genes have long been implicated in modulating drug toxicity and will not be discussed here. Concerning gene expression, a number of studies have already shown that specific toxic agents can give specific gene expression signatures.[30,85–87] These signatures can therefore be used to address whether a drug might have off-target effects or might induce end-organ toxicity. As the database of expression signatures improves, it is conceivable that new drugs can be categorised in terms of how they can be clustered together. In this manner, expression signatures can be used to redefine chemical 'space'. Thus, drugs with a common toxicity profile might cluster together in the transcriptional signature they produce in liver cells.[88–90] Kier *et al.*[91] used limited microarrays with rat-specific toxicologically relevant genes to generate profiles and outcomes for 89 compounds. Gene expression specific to dose and time for many compounds could be clustered together, suggesting similar mechanisms of liver toxicity. Moreover, these authors observed that the expression signature at 24 hours was found to correlate well with organ toxicity seen at 72 hours. Such an approach has been piloted in assessing susceptibility to radiation-associated toxicity using the expression response in lymphoblastoid cell lines as surrogate tissues.[92] As might be expected, companies have used this strategy potentially to select chemical leads with less toxicity (Gene Logic, Gaithersburg, MD. http://www.genelogic.com/solutions/toxexpress/).

Previously, the limiting factor for the application of expression microarrays in toxicology studies was the absence of a rich EST database for the species commonly used for toxicological testing: rats and dogs. Although this is, in large part, being solved, the widespread development has lagged behind work in mice and in humans. This, coupled with the fact that many of the data generated are proprietary, has limited the expansion of expression genomic approaches in toxicology testing in drug development.

INTEGRATED GENOMIC STRATEGY FOR PHARMACEUTICAL DEVELOPMENT

It has been long acknowledged that the drug discovery process needs to be improved, and that genomics might provide such a solution; however, the reality is that the sequencing of the human genome has not been able to give new targets directly, but is a strong enabling platform for innovations in drug discovery. This paper provides evidence that transcriptional genomics is a technology that is very appropriate for advanced pharmacological applications. The author's conceptual and operational recommendations are as follows:

1. *Genomic information, when interchangeable increases its value as the size of the database and the quality of its annotation expands*. Thus, the standard approach of each research unit (discovery, therapeutic areas, bioinformatics, toxicology) holding its own expression data and other genomic data is a wasted opportunity, just as much as internet connectivity limited to a single research unit is a wasted resource. All genomic data, including expression, sequence, single nucleotide polymorphism and comparative genomic data, should be

Information integration will be key in making predictive pharmacology a reality

placed in a common database using standards such that meta-analyses can be performed on the entire dataset. This opens up the possibility that different and completely unexpected and new indications for drugs in development can be discovered through computational means. Although institutions defend the integrative nature of their databases, most commonly the integration is available only to the bioinformatics experts or is actually piecemeal. What is needed is a platform whereby the biologists and the chemists involved in the development of a specific lead compound can directly access all pertinent information, rendered in a useable manner to bench scientists.

2. *Genomic information is best used in an iterative fashion with wet lab biology*. The standard operational concept that large genomics studies can be performed by one isolated unit within a company and then 'handed off' to the therapeutic areas for exploitation is not viable, and is very costly. Instead, an incremental and repetitive process of hypothesis generation through genomic approaches, followed by wet lab — biological validation and experimentation — is the best strategy.
3. The wealth of validated markers arising from these genomics approaches, which can be very precise in assessing cellular and clinical states, suggests that *biomarkers will play an increasingly important role in cellular and preclinical screens and in Phase I dose-finding exercises*. Given that multiple markers can provide a more refined estimate of a clinical state, simply asserting that the purported target of a therapeutic is downregulated by an inhibitor may be insufficient to prove specificity. The author projects that large panels of markers will be used, rather than single tests. These markers will be used in the full research and developmental cycle: in screening compounds, in preclinical target ascertainment, in toxicological assessements, in patient stratification and selection, and in clinical monitoring. Moreover, the coupling of marker development with the therapeutic will be increasingly important (eg Herceptin/trastuzumab therapeutic and HercepTest or PathVysion diagnostic). This suggests that there should be renewed interest in the biomarkers sector.
4. *Rational combinations of targeted therapeutics must be one of the future goals of the pharmaceutical industry*. The concept of tailored therapeutics specifically styled for the individual patient requires combining treatments that target individual clinical 'signatures'. Given the combinatorial possibilities, some conceptual framework will be needed to guide the development of drug combinations. The author believes that assessment of global gene expression changes can provide the information to project the optimal combinations for a cellular phenotype. The first applications might be in synthetic ligands and inhibitors for the nuclear hormone receptors where the expression read-out is the most proximate response to drug intervention.
5. *The author believes that the precise and complete mapping of transcriptional regulatory networks is feasible in certain conditions*. Such a map will provide the necessary framework to test whether pharmacological output can be predicted much in the same way that small molecular structures that dock enzyme pockets can be estimated. Such an approach can be considered as predictive pharmacology.
6. *The same factors that heighten the effectiveness of expression genomics in the drug discovery process are also the factors that can curtail the development process*. The breadth and comprehensiveness of genomic data are both advantageous and disadvantageous. If, in the course

of an array experiment, an oncogene has been uncovered that is consistently upregulated by a drug to treat diabetes, does this mean that even without biological plausibility, the development of the compound should be stalled or even aborted? Certainly, there is a concern that in the toxicological analysis of expression array data, any expression change that raises the remotest doubt as to the safety of a compound would be viewed negatively. Given the broad nature of the expression response to any agent, there is always a possibility of a single gene acting in what might be considered as an adverse manner, despite the absence of any biological significance. How regulatory agencies will deal with this information is unclear. Equally concerning, however, is how decision makers within a pharmaceutical company will treat such information. Would small ambiguities among this mass of information lead them to abort a project simply because of speculative toxicities?

Overall, the outlook is good. The speed, precision and predictive power of such genome-wide approaches will surely accelerate the discovery and validation of new targets and new drugs. The limiting factor is archaic organisational structures that cannot adapt to such fluid and voluminous data flow. This, therefore, will be our foremost challenge.

References

1. Davidson, E. H., Rast, J. P., Oliveri, P. *et al.* (2002), 'A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo', *Dev. Biol.*, Vol. 246, pp. 162–190.
2. Davidson, E. H., Rast, J. P., Oliveri, P. *et al.* (2002), 'A genomic regulatory network for development', *Science*, Vol. 295, pp. 1669–1678.
3. Bolouri, H. and Davidson, E. H. (2003), 'Transcriptional regulatory cascades in development: Initial rates, not steady state, determine network kinetics', *Proc. Natl. Acad. Sci. USA*, Vol.100, pp. 9371–9376.
4. Futreal, P. A., Coin, L., Marshall, M. *et al.* (2004), 'A census of human cancer genes', *Nat. Rev.*, Vol. 4, pp. 177–183.
5. Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R. and Lockhart, D. J. (1999), 'High density synthetic oligonucleotide arrays', *Nat. Genet.,* Vol. 21, pp. 20–24.
6. Eisen, M. B. and Brown, P. O. (1999), 'DNA arrays for analysis of gene expression', in Weissman, S. M. (ed.), 'Methods in Enzymology: cDNA Preparation and Characterization', Vol. 303, Academic Press, San Diego, CA, pp.179–205.
7. Harrington, C. A., Rosenow, C. and Retief, J. (2000), 'Monitoring gene expression using DNA microarrays', *Curr. Opin. Microbiol.*, Vol. 3, pp. 285–291.
8. Miller, L. D., Long, P. M., Wong, L. *et al.* (2002), 'Optimal gene expression analysis by microarrays', Cancer Cell, Vol. 2, pp. 353–361.
9. Stears, R. L., Martinsky, T. and Schena, M. (2003), 'Trends in microarray analysis', *Nat. Med.*, Vol. 9, pp. 140–145.
10. Ruan, Y., Le Ber, P., Ng, H. H. and Liu, E. T. (2004), 'Interrogating the transcriptome', *Trends Biotechnol.*, Vol. 22, pp. 23–30.
11. Brenner, S. and Livak, K. J. (1989), 'DNA fingerprinting by sampled sequencing', *Proc. Natl. Acad. Sci. USA*, Vol. 86, pp. 8902–8906.
12. Kapranov, P., Cawley, S. E., Drenkow, J. *et al.* (2002), 'Large-scale transcriptional activity in chromosomes 21 and 22', *Science*, Vol. 296, pp. 916–919.
13. Cawley, S., Bekiranov, S., Ng, H. H. *et al.* (2004), 'Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs', *Cell*, Vol. 116, pp. 499–509.
14. Johnson, J. M., Castle, J., Garrett-Engele, P. *et al.* (2003), 'Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays', *Science*, Vol. 302, pp. 2141–2144.
15. Wei, C. L., Ng, P., Chiu, K. P. *et al.* (2004), '5′ Long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation', *Proc. Natl. Acad. Sci. USA*, Vol. 101, pp. 11701–11706.
16. Druker, B. J. (2004), 'Imatinib as a paradigm of targeted therapies', *Adv. Cancer Res.*, Vol. 91, pp. 1–30.
17. Liu, E. T., Sandler, D., Neubauer, A. *et al.* (1991), 'Clinical and etiologic importance of mutant ras genes in adult acute myeloid leukemia (AML)', *Blood*, Vol. 78 (Suppl. 1), p. 1340.
18. Koo, H. M., McWilliams, M. J., Alvord, W.

G. and Vande Woude, G. F. (1999), 'Ras oncogene-induced sensitization to 1-beta-D-arabinofuranosylcytosine', *Cancer Res.*, Vol. 59, pp. 6057–6062.

19. Koo, H. M., Gray-Goodrich, M., Kohlhagen, G. *et al.* (1999), 'The ras oncogene-mediated sensitization of human cells to topoisomerase II inhibitor-induced apoptosis', *J. Natl. Cancer Inst.*, Vol. 91, pp. 236–244.
20. Sattler, M., Scheijen, B., Weisberg, E. and Griffin, J. D. (2003), 'Mutated tyrosine kinases as therapeutic targets in myeloid leukemias', *Adv. Exp. Med. Biol.*, Vol. 532, pp. 121–140.
21. O'Farrell, A. M., Yuen, H. A., Smolich, B. *et al.* (2004), 'Effects of SU5416, a small molecule tyrosine kinase receptor inhibitor, on FLT3 expression and phosphorylation in patients with refractory acute myeloid leukaemia', *Leuk. Res.*, Vol. 28, pp. 679–689.
22. O'Farrell, A. M., Abrams, T. J., Yuen, H. A. *et al.* (2003), 'SU11248 is a novel FLT3 tyrosine kinase inhibitor with potent activity *in vitro* and *in vivo*', *Blood*, Vol. 101, pp. 3597–3605.
23. Parsons, R., Myeroff, L. L., Liu, B. *et al.* (1995), 'Microsatellite instability and mutations of the transforming growth factor betatype II receptor gene in colorectal cancer', *Cancer Res.*, Vol. 55, pp. 5548–5550.
24. Davies, H., Bignell, G. R., Cox, C. *et al.* (2002), 'Mutations of the BRAF gene in human cancer', *Nature*, Vol. 417, pp. 949–954.
25. Wilhelm, S. M., Carter, C., Tang, L. *et al.* (2004), 'BAY 43-9006 exhibits broad spectrum oral antitumor activity and targets the RAF/MEK/ERK pathway and receptor tyrosine kinases involved in tumor progression and angiogenesis', *Cancer Res.*, Vol. 64, pp. 7099–7109.
26. Paez, J. G., Janne, P. A., Lee, J. C., *et al.* (2004), 'EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy', *Science*, Vol. 304, pp. 1497–1500.
27. Herbst, R. S., Fukuoka, M. and Baselga, J. (2004), 'Timeline: Gefitinib — a novel targeted approach to treating cancer', *Nat. Rev. Cancer*, Vol. 4, pp. 956–965.
28. Stephens, P., Hunter, C., Bignell, G. *et al.* (2004), 'Lung cancer: Intragenic ERBB2 kinase mutations in tumours', *Nature*, Vol. 431, pp. 525–526.
29. Wang, Z., Shen, D., Parsons, D. W. *et al.* (2004), 'Mutational analysis of the tyrosine phosphatome in colorectal cancers', *Science*, Vol. 304, pp. 1164–1166.
30. Hughes, T. R., Marton, M. J., Jones, A. R. *et al.* (2000), 'Functional discovery via a compendium of expression profiles', *Cell*, Vol. 102, pp. 109–126.
31. Miller, L. D., Park, K. S., Guo, Q. M. *et al.* (2001), 'Silencing of Wnt signaling and activation of multiple metabolic pathways in response to thyroid hormone-stimulated cell proliferation', *Mol. Cell. Biol.*, Vol. 21, pp. 6626–6639.
32. Guo, Q. M., Malek, R. L., Kim, S. *et al.* (2000), 'Identification of *c-myc* responsive genes using rat cDNA microarray', *Cancer Res.*, Vol. 60, pp. 5922–5928.
33. Iritani, B. M. and Eisenman, R. N. (1999), 'c-Myc enhances protein synthesis and cell size during B lymphocyte development', *Proc. Natl. Acad. Sci. USA*, Vol. 96, pp. 13180–13185.
34. Aprelikova, O., Pace, A. J., Fang, B. *et al.* (2001), 'BRCA1 is a selective co-activator of 14-3-3 sigma gene transcription in mouse embryonic stem cells', *J. Biol. Chem.*, Vol. 276, pp. 25647–25650.
35. Kho, P. S., Wang, Z., Zhuang, L. *et al.* (2004), 'p53-regulated transcriptional program associated with genotoxic stress-induced apoptosis', *J. Biol. Chem.*, Vol. 279, pp. 21183-21192.
36. Majumder, P. K., Febbo, P. G., Bikoff, R. *et al.* (2004), 'mTOR inhibition reverses Akt-dependent prostate intraepithelial neoplasia through regulation of apoptotic and HIF-1-dependent pathways', *Nat. Med.*, Vol. 10, pp. 594–601.
37. Chen, C. D., Welsbie, D. S., Tran, C. *et al.* (2004), 'Molecular determinants of resistance to antiandrogen therapy', *Nat. Med.*, Vol. 10, pp. 33–39.
38. Frasor, J., Danes, J. M., Komm, B. *et al.* (2003), 'Profiling of estrogen up- and downregulated gene expression in human breast cancer cells: Insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype', *Endocrinology,* Vol. 144, pp. 4562–4574.
39. Coser, K. R., Chesnes, J., Hur, J. *et al.* (2003), 'Global analysis of ligand sensitivity of estrogen inducible and suppressible genes in MCF7/BUS breast cancer cells by DNA microarray', *Proc. Natl. Acad. Sci. USA,* Vol. 100, pp. 13994–13999.
40. Lin, C. Y., Strom, A., Vega, V. B. *et al.* (2004), 'Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells', *Genome Biol.*, Vol. 5, p. R66.
41. Bajic, V. B., Tan, S. L., Chong, A. *et al.* (2003), 'Dragon ERE Finder version 2: A tool for accurate detection and analysis of estrogen response elements in vertebrate genomes', *Nucleic Acids Res.*, Vol. 31, pp. 3605–3607.
42. Desai, K. V., Xiao, N., Wang, W. *et al.* (2003), 'Initiating oncogenic event determines gene-expression patterns of human breast cancer models', *Proc. Natl. Acad. Sci. USA*, Vol. 99, pp. 6967–6972.

43. Jazaeri, A. A., Yee, C. J., Sotiriou, C. *et al.* (2002), 'Gene expression profiles of BRCA1-linked, BRCA2-linked, and sporadic ovarian cancers', *J. Natl. Cancer Inst.*, Vol. 94, pp. 990–1000.

44. Hedenfalk, I., Duggan, D., Chen, Y. *et al.* (2001), 'Gene-expression profiles in hereditary breast cancer', *N. Engl. J. Med.*, Vol. 344, pp. 539–548.

45. Ganesan, S., Silver, D. P., Greenberg, R. A. *et al.* (2002), 'BRCA1 supports XIST RNA concentration on the inactive X chromosome', *Cell*, Vol. 111, pp. 393–405.

46. Jazaeri, A. A., Chandramouli, G. V., Aprelikova, O. *et al.* (2004), 'BRCA1-mediated repression of select X chromosome genes', *J. Transl. Med.*, Vol. 2, p. 32.

47. Ferrando, A. A., Neuberg, D. S., Staunton, J. *et al.* (2002), 'Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukaemia', *Cancer Cell*, Vol. 1, pp. 75–87.

48. Ferrando, A. A., Neuberg, D. S., Dodge, R. K. *et al.* (2004), 'Prognostic importance of TLX1 (HOX11) oncogene expression in adults with T-cell acute lymphoblastic leukaemia', *Lancet*, Vol. 363, pp. 535–536.

49. Relling, M. V., Hancock, M. L., Rivera, G. K. *et al.* (1999), ' Mercaptopurine therapy intolerance and heterozygosity at the thiopurine S-methyltransferase gene locus', *J. Natl. Cancer Inst.*, Vol. 91, pp. 2001–2008.

50. Krynetski, E. and Evans, W. E. (2003), 'Drug methylation in cancer therapy: Lessons from the TPMT polymorphism', *Oncogene*, Vol. 22, pp. 7403–7413.

51. Liu, E. T. (2001), 'Molecular Markers in cancer therapy: Evolution of a concept', in Mendelsohn, J., Howley, P., Israel, M. and Liotta, L. (eds), 'Molecular Basis of Cancer', 2nd edn, W. B. Saunders, Philadelphia.

52. Perou, C. M., Sorlie, T., Eisen, M. B. *et al.* (2000), 'Molecular portraits of human breast tumours', *Nature*, Vol. 406, pp. 747–752.

53. Sorlie, T., Perou, C. M., Tibshirani, R. *et al.* (2001), 'Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications', *Proc. Natl. Acad. Sci. USA*, Vol. 98, pp. 10869–10874.

54. Sotiriou, C., Neo, S. Y., McShane, L. M. *et al.* (2003), 'Breast cancer classification and prognosis based on gene expression profiles from a population-based study', *Proc. Natl. Acad. Sci. USA*, Vol. 100, pp. 10393–10398.

55. Sorlie, T., Tibshirani, R., Parker, J. *et al.* (2003), ' Repeated observation of breast tumor subtypes in independent gene expression data sets', *Proc. Natl. Acad. Sci. USA*, Vol. 100, pp. 8418–8423.

56. van 't Veer, L. J., Dai, H., van de Vijver, M. J. *et al.* (2002), 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature*, Vol. 415, pp. 530–536.

57. van de Vijver, M. J., He, Y. D., van't Veer, L. J. *et al.* (2002), 'A gene-expression signature as a predictor of survival in breast cancer', *N. Engl. J. Med.*, Vol. 347, pp. 1999–2009.

58. Garber, K. (2004), 'Gene expression tests foretell breast cancer's future', *Science*, Vol. 303, pp. 1754–1755.

59. Singh, D., Febbo, P. G., Ross, K. *et al.* (2002), 'Gene expression correlates of clinical prostate cancer behaviour', *Cancer Cell*, Vol. 1, pp. 203–209.

60. Dhanasekaran, S. M., Barrette, T. R., Ghosh, D. *et al.* (2001), 'Delineation of prognostic biomarkers in prostate cancer', *Nature*, Vol. 412, pp. 822–826.

61. Varambally, S., Dhanasekaran, S. M., Zhou, M. *et al.* (2002), 'The polycomb group protein EZH2 is involved in progression of prostate cancer', *Nature*, Vol. 419, pp. 624–629.

62. Staudt, L. M. (2002), 'It's ALL in the diagnosis', *Cancer Cell*, Vol. 1, pp. 109–110.

63. Yeoh, E. J., Ross, M. E., Shurtleff, S. A. *et al.* (2002), 'Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling', *Cancer Cell*, Vol. 1, pp. 133–143.

64. Ferrando, A. A., Neuberg, D. S., Staunton, J. *et al.* (2002), 'Look AT. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukaemia', *Cancer Cell*, Vol. 1, pp. 75–87.

65. Bullinger, L., Dohner, K., Bair, E. *et al.* (2004), 'Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukaemia', *N. Engl. J. Med.*, Vol. 350, pp. 1605–1616.

66. Valk, P. J., Verhaak, R. G., Beijen, M. A. *et al.* (2004), 'Prognostically useful gene-expression profiles in acute myeloid leukaemia', *N. Engl. J. Med.*, Vol. 350, pp. 1617–1628.

67. Alizadeh, A. A., Eisen, M. B., Davis, R. E. *et al.* (2000), 'Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling', *Nature*, Vol. 403, pp. 503–511.

68. Rosenwald, A., Wright, G., Chan, W. C. *et al.* (2002), 'Lymphoma/leukemia molecular profiling project. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma', *N. Engl. J. Med.*, Vol. 346, pp. 1937–1947.

69. Shipp, M. A., Ross, K. N., Tamayo, P. *et al.* (2002), 'Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning', *Nat. Med.*, Vol. 8, pp. 68–74.

70. Wheeler, G. D. (2003), 'Ruboxistaurin (Eli Lilly)' I Drugs, Vol. 6, pp. 159–163.

71. Teicher, B. A., Menon, K., Alvarez, E. *et al.* (2002), 'Antiangiogenic and antitumor effects of a protein kinase Cbeta inhibitor in human breast cancer and ovarian cancer xenografts', *Invest. New Drugs*, Vol. 20, pp. 241–251.
72. Teicher, B. A., Alvarez, E., Menon, K. *et al.* (2002), 'Antiangiogenic effects of a protein kinase C beta-selective small molecule', *Cancer Chemother. Pharmacol.*, Vol. 49, pp. 69–77.
73. Scherf, U., Ross, D. T., Waltham, M., *et al.* (2000), 'A gene expression database for the molecular pharmacology of cancer', *Nat. Genet.*, Vol. 24, pp. 236–244.
74. Ross, D. T., Scherf, U., Eisen, M. B. *et al.* (2000), 'Systematic variation in gene expression patterns in human cancer cell lines', *Nat. Genet.*, Vol. 24, pp. 227–235.
75. Staunton, J. E., Slonim, D. K., Coller, H. A. *et al.* (2001), 'Chemosensitivity prediction by transcriptional profiling', *Proc. Natl. Acad. Sci. USA*, Vol. 98, pp. 10787–10792.
76. Kikuchi, T., Daigo, Y., Katagiri, T. *et al.* (2003), 'Expression profiles of non-small cell lung cancers on cDNA microarrays: Identification of genes for prediction of lymph-node metastasis and sensitivity to anti-cancer drugs', *Oncogene*, Vol. 22, pp. 2192–2205.
77. Holleman, A., Cheok, M. H., den Boer, M. L. *et al.* (2004), 'Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment', *N. Engl. J. Med.*, Vol. 351, pp. 533–542.
78. Sotiriou, C., Powles, T. J., Dowsett, M. *et al.* (2002), 'Gene expression profiles derived from fine needle aspiration correlate with response to systemic chemotherapy in breast cancer', *Breast Cancer Res.*, Vol. 4, p. R3.
79. Cheok, M. H., Yang, W., Pui, C. H. *et al.* (2003), 'Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells', *Nat. Genet.*, Vol. 34, pp. 85–90.
80. Bani, M. R., Nicoletti, M. I., Alkharouf, N. W. *et al.* (2004), 'Gene expression correlating with response to paclitaxel in ovarian carcinoma xenografts', *Mol. Cancer Ther.*, Vol. 3, pp. 111–121.
81. Evans, W. E. and Relling, M. V. (2004), 'Moving towards individualized medicine with pharmacogenomics', *Nature*, Vol. 429, pp. 464–468.
82. Hunter, K. W. (2003), 'Allelic diversity in the host genetic background may be an important determinant in tumor metastatic dissemination', *Cancer Lett.*, Vol. 200, pp. 97–105.
83. Qiu, T. H., Chandramouli, G. V., Hunter, K. W. *et al.* (2004), 'Global expression profiling identifies signatures of tumor virulence in MMTV-PyMT-transgenic mice: Correlation to human disease', *Cancer Res.*, Vol. 64, pp. 5973–5981.
84. Ramaswamy, S., Ross, K. N., Lander, E. S. and Golub, T. R. (2003), 'A molecular signature of metastasis in primary solid tumors', *Nat. Genet.*, Vol. 33, pp. 49–54.
85. Chuang, Y. Y., Chen, Y., Gadisetti, C. *et al.* (2002), 'Gene expression after treatment with hydrogen peroxide, menadione, or t-butyl hydroperoxide in breast cancer cells', *Cancer Res.*, Vol. 62, pp. 6246–6254.
86. de Longueville, F., Bertholet, V. and Remacle, J. (2004), 'DNA microarrays as a tool in toxicogenomics', *Comb. Chem. High Throughput Screen*, Vol. 7, pp. 207–211.
87. Suter, L., Babiss, L. E., Wheeldon, E. B. (2004), 'Toxicogenomics in predictive toxicology in drug development', *Chem. Biol.,* Vol. 11, pp. 161–171.
88. Burczynski, M. E., McMillian, M., Ciervo, J. *et al.* (2000), 'Toxicogenomics-based discrimination of toxic mechanism in HepG2 human hepatoma cells', *Toxicol. Sci.*, Vol. 58, pp. 399–415.
89. Jessen, B. A., Mullins, J. S., De Peyster, A. and Stevens, G. J. (2003), 'Assessment of hepatocytes and liver slices as *in vitro* test systems to predict *in vivo* gene expression', *Toxicol. Sci.*, Vol. 75, pp. 208–222.
90. de Longueville, F., Atienzar, F. A., Marcq, L. *et al.* (2003), 'Use of a low-density microarray for studying gene expression patterns induced by hepatotoxicants on primary cultures of rat hepatocytes', *Toxicol. Sci.*, Vol. 75, pp. 378–392.
91. Kier, L. D., Neft, R., Tang, L. *et al.* (2004), 'Applications of microarrays with toxicologically relevant genes (tox genes) for the evaluation of chemical toxicants in Sprague Dawley rats *in vivo* and human hepatocytes *in vitro*', *Mutat. Res.*, Vol. 549, pp. 101–113.
92. Rieger, K. E., Hong, W. J., Tusher, V. G. *et al.* (2004), 'Toxicity from radiation therapy associated with abnormal transcriptional responses to DNA damage', *Proc. Natl. Acad. Sci. USA*, Vol. 101, pp. 6635–6640.