OPINION

Integrated global profiling of cancer

Samir Hanash

Tumours are complex biological systems. No single type of molecular approach fully elucidates tumour behaviour, necessitating analysis at multiple levels encompassing genomics and proteomics. Integrated data sets are required to fully determine the contributions of genome alterations, host factors and environmental exposures to tumour growth and progression, as well as the consequences of interactions between malignant or premalignant cells and their microenvironment. The sheer amount and heterogeneous nature of data that need to be collected and integrated are daunting, but effort has already begun to address these obstacles.

In the 1980s, at the dawn of the era of molecular medicine, researchers believed that cancer was caused by dysregulation of a few oncogenes or tumour-suppressor genes. The identification of these genes would therefore lead to effective approaches for preventing or treating cancer. Substantial progress has been made in uncovering cancer genes that are altered through point mutations, deletions, amplifications, rearrangements or other events, and as a result effective targeted therapies for certain cancers have been developed. It has become clear, however, that human tumours are more complex and heterogeneous than expected, and are caused by defects in numerous pathways and factors that operate at many levels. For example, a gene can be amplified 100-fold in certain tumours with no demonstrable effect on RNA levels for that gene. Alternatively, protein levels can be increased, decreased or modified with no demonstrable changes in the levels of their corresponding RNAs. It is therefore a challenge to fully understand tumour behaviour, based on a single type of analysis. The factors that determine the consequences of a particular event or alteration can be highly context dependent, and are governed by the spatial and temporal activity of numerous interacting components. The intricate nature of the contributions of many factors ultimately determines the impact that a particular alteration has on the properties of a tumour or a precursor lesion.

There are two basic approaches to address the complexity of cancer. One is to reduce complexity through analysis of experimental models, such as cell lines or animal models, to characterize the fundamental processes of tumour growth and to elucidate the effects of single genes. Another is to integrate large data sets, to yield a model for tumour development and behaviour. Each approach has its own advantages and disadvantages. The first approach has been effective in many respects; for example, the early stages of tumorigenesis have been investigated using mouse models, and transformation and metastasis have been modelled in *Drosophila*[1]. However, in studying animal models of cancer, many factors that are relevant to human cancer are lost. The conclusions reached from these models are therefore not always applicable to human tumours[2]. The second approach, involving integration of large data sets, is challenging in part because only a limited number of samples, such as tumours or preneoplastic tissues, can be analysed in a given study. This makes data interpretation and model development difficult, given the large amount of heterogeneity between human tumours.

Profiling strategies

Improving our understanding of cancer and developing theoretical models will require an increased understanding of the contributions of and interactions between the numerous components that contribute to tumour formation and progression (FIG. 1). Strategies are available to profile changes at various levels, including the genome, transcriptome and proteome (TABLE 1). The host genome can be scanned for inherited variations such as mutations and polymorphisms that might contribute to cancer risk. Tumour cells and their precursors can be assayed for genomic alterations, such as chromosomal deletions or amplification, or changes in DNA methylation status, that promote their proliferation and survival. The cancer-cell transcriptome can be examined for patterns of gene expression, or its proteome analysed to uncover alterations in proteins, that contribute to tumour development or progression and would not be predicted by genome or transcriptome analysis.

A challenge for global profiling is the need to capture all the elements of the individual compartments that are profiled, such as the whole transcriptome or the whole proteome. Although this is possible for the transcriptome, other compartments, such as the proteome and metabolome, have numerous features that are difficult to capture, requiring several different profiling approaches (TABLE 1). For example, it is not possible to assay for protein functional activity, profile protein–protein interactions, and assess protein modifications all with the same platform. In all, there remains a substantial need to improve the breadth, sensitivity and throughput of global-profiling technologies.

In addition to global profiling of DNA, RNA or protein in normal, premalignant and malignant tissues, and in biological fluids, a comprehensive analysis would measure other characteristics from these samples to detect changes in nutritional, metabolic and immune status, as well as to detect environmental exposures. These

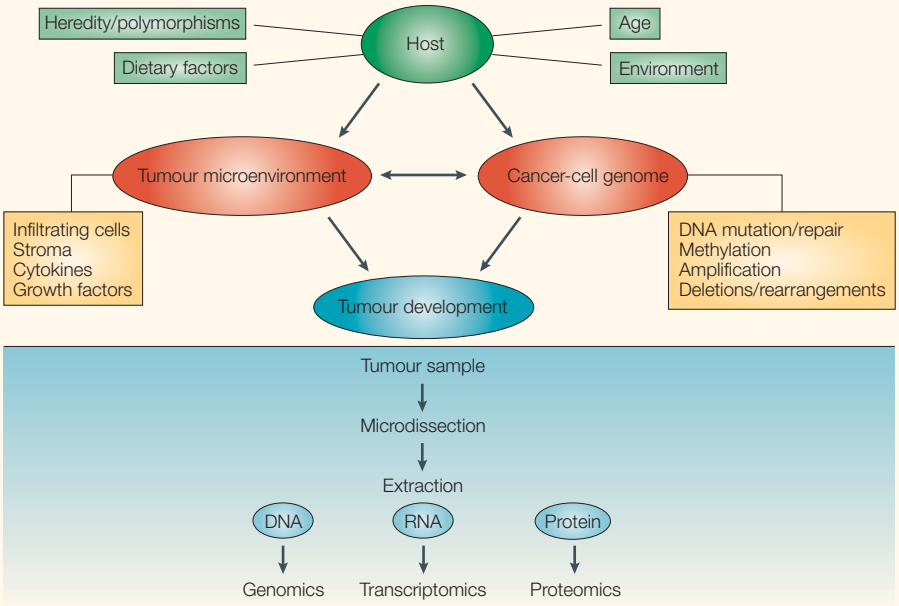

Figure 1 | **Numerous components must be integrated to study the molecular basis of human cancer.** Several host factors contribute to tumorigenesis in humans, including diet, environmental factors, polymorphisms and mutations in susceptibility genes, age and immunity. Cells undergo genomic changes (DNA mutations and repair, methylation, amplification, deletions and rearrangements), leading to tumorigenesis. Tumour development also depends on factors in the microenvironment — some of these are produced locally, whereas others are produced systemically (growth factors, infiltrating cells and cytokines). Reciprocal interactions between the premalignant and malignant cells, stromal cells, extracellular-matrix components, various inflammatory cells and a range of soluble mediators therefore contribute to tumour development and progression. Once tumour samples are obtained, genomic, transcriptomic and proteomic tools can be used to profile specific compartments.

types of data come from metabolic and nutritional profiles, immunohistochemical assays, assays of host immunity to tumour antigens, and patient questionnaires. Such data need to be integrated with molecular profile data.

Integrating data sets

So far, very few cancer studies have attempted to integrate data sets that were obtained by several different profiling techniques. Rather, the few large-scale integrated molecular-profiling efforts undertaken have combined data of a similar nature, notably combining transcriptome data obtained from several sources. Some studies have combined data obtained through two different global-profiling platforms (genomic and transcriptomic, or transcriptomic and proteomic) for the same set of study samples (such as lung tumours). These integrated data sets have also included variables such as clinical and pathological characteristics of the study individuals and their tumours, or mutations in cancer genes such as *TP53* and *RAS*. However limited in scope, these studies illustrate the potential impact of integrating data across numerous data sets in elucidating certain features of cancer[3–8].

Integrating gene-expression data from different sources. Profiling gene expression using DNA arrays has had a tremendous impact on biomedical research. Although the field is still in its infancy, there is increasing emphasis on integration of diverse sets of data. From a cancer research point of view, applications of global profiling of gene expression include uncovering unsuspected associations between genes, or identifying specific clinical features of cancer that result in novel molecular-based disease classifications. For example, DNA microarray analysis has been used to associate specific gene-expression profiles with different clinical outcomes of patients with the same types of tumours (responders versus non-responders[9]), or with cancer subtypes of the same lineage (high-stage versus low-stage tumours). Specific gene-expression signatures have also been associated with tumours of different lineages[10].

Lamb *et al.*[3] performed a study that illustrates the merits of integrating gene-expression data from several sources to develop a mechanistic understanding. They integrated gene-expression data from cell lines and human tumours to uncover a cyclin-dependent kinase (CDK)-independent mechanism of cyclin D1 function. Cyclin D1, which activates CDK, is frequently overexpressed in human tumours, but the mechanisms by which this promotes tumorigenesis has been unclear. Cyclin D1 and a cyclin-D1 mutant that was incapable of activating CDK4 were each ectopically expressed in cultured human mammary epithelial cells.

Table 1 | **Profiling strategies for genome-related components**

Platform	What we can learn	What is detected	Tools used for analysis
Genome	The hereditary components to cancer, as well as genome alterations in somatic cells that lead to cancer	Chromosome structural changes; gene copy-number changes; gene rearrangements; mutations/polymorphisms; methylation changes	DNA sequencing; cytogenetics; CGH; array CGH; SNP analysis; RLGS
Transcriptome	Changes in gene expression that are associated with cancer	Changes in RNA abundance; alterations in alternative splicing	Differential-display analysis; SAGE; DNA microarray analysis; PCR- and non-PCR-based gene-expression assays
Proteome	How proteins are modified or how their levels change in tumours	Protein levels; post-translational modifications; localization; protein–protein interactions; enzymatic activity	Sample-enrichment strategies (fractionation, protein tagging); separation-based profiling (2D gels, MS, LC, LC-MS); non-separation-based strategies (protein microarrays, direct MS analysis); protein-detection strategies (immunohistochemistry, immunofluorescence)

2D, two dimensional; CGH, comparative genomic hybridization; LC, liquid chromatography; MS, mass spectrometry; PCR, polymerase chain reaction; RLGS, restriction landmark genome scanning; SAGE, serial analysis of gene expression; SNP, single nucleotide polymorphism.

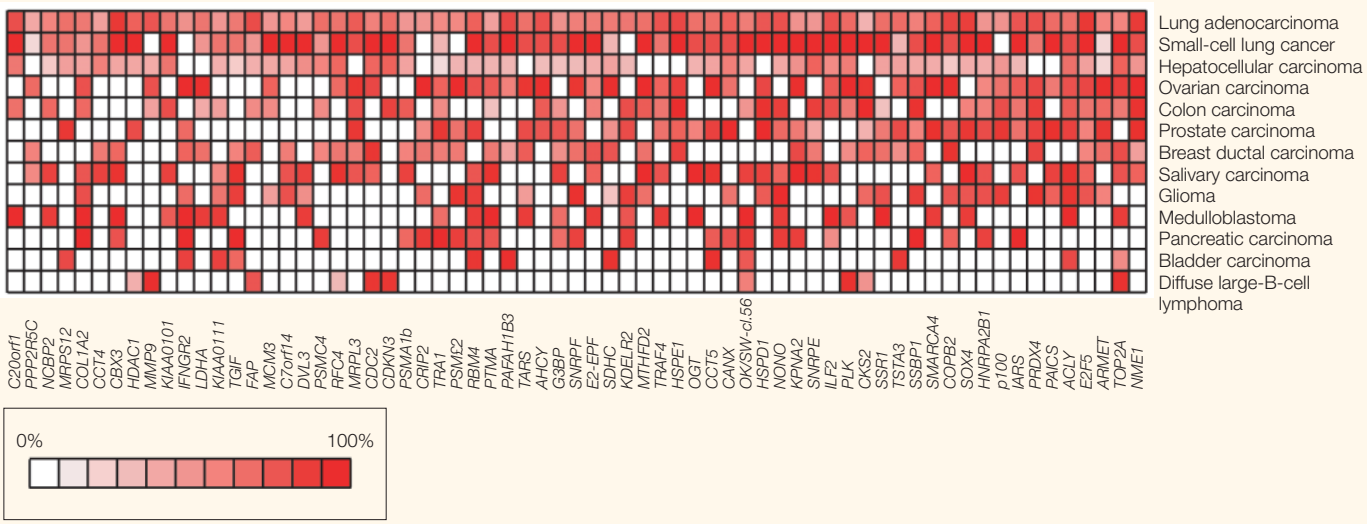

Figure 2 | **Integrated gene-expression profile of neoplastic transformation.** Public sharing of gene-expression data has led to the identification of 67 genes that are commonly overexpressed in tumour samples, relative to normal tissue. This 'meta-signature' analysis compared 'cancer versus normal' gene-expression signatures from 21 independent microarray data sets. Thirteen distinct cancer types were selected for this figure (listed on the right). White boxes signify genes for which no changes in expression were observed between tumour and normal cells. Light and dark red boxes signify genes that were significantly overexpressed in tumour cells, relative to normal tissue. Dark red indicates that the expression level was in the 90th percentile of all samples tested. Figure reproduced with permission from REF. 11 © (2004) National Academy of Sciences.

Twenty-one genes were found to be induced by both wild-type and mutant cyclin D1, indicating that these genes are CDK4 independent. Furthermore, the rapidity with which expression of these genes was induced indicated the direct involvement of a transcription factor. A database of gene-expression profiles from 190 primary human tumours was therefore also analysed, to identify cyclin-D1 target genes. The expression pattern of the set of 21 genes uncovered from *in vitro* studies was correlated with the levels of cyclin D1 in human tumours. A 'data-mining' process was applied to several human tumour gene-expression data sets, to identify genes that had a pattern of expression that matched the patterns of the genes that comprised the cyclin-D1 signature pattern. The transcription factor C/EBPβ was consistently co-expressed with the set of cyclin-D1 target genes. Functional analyses confirmed the involvement of C/EBPβ in the transcriptional regulation of cyclin D1. This study illustrates the types of findings that can be uncovered by integrating different sets of data.

Tumour gene-expression patterns are modulated by many extrinsic factors and by the microenvironment — these features could be crucial factors in determining the response to anticancer drugs. The gene-expression profiles of *in vitro* cultures of cancer cells have been compared with those of tumours grown *in vivo*, to determine the effects of the microenvironment on gene expression. In one study[4], two human cancer cell lines (a lung adenocarcinoma and a glioblastoma cell line) were transplanted into immunodeficient mice and allowed to form tumours, and the gene-expression profiles of these tumours were compared with those of cells grown in culture. A bioinformatics approach was used to associate genes into functional classes. The classes of genes that were expressed at higher levels in cells grown *in vitro* were associated with increased cell division and metabolism, reflecting the more favourable environment for cell proliferation. By contrast, *in vivo* tumour growth resulted in upregulation of a significant number of genes involved in extracellular-matrix formation, cell adhesion, cytokine and metalloproteinase activity, and neovascularization. When placed in comparable *in vivo* tissue environments, the lung cancer and the glioblastoma cells expressed different sets of extracellular-matrix- and cell-adhesion-related genes, indicating different mechanisms of extracellular interaction at work in the different tumour types. Importantly, gene products that are typically targeted by cancer therapies, such as tyrosine kinases, showed varied expression patterns when the same cancer cells were grown *in vitro* versus *in vivo*. This provides an indication of why therapeutics that are effective in *in vitro* studies might not always function *in vivo*.

A study that illustrates the merits of data sharing among investigators is a meta-analysis of cancer microarray data[11]. In this study, 40 published cancer microarray data sets comprising gene-expression measurements from over 3,700 tumour samples were collected and analysed. A common transcriptional profile that is activated in most cancer types, relative to corresponding normal tissues, was delineated from some of the data sets, providing a meta-signature of neoplastic transformation (FIG. 2).

Integrating genomic and transcriptomic data. Most tumours show numerous genomic alterations, but it has been a challenge to identify those that are required for different stages of tumour development. As most genome alterations — chromosomal gains and losses, deletions, amplification and methylation — affect the transcriptome, it would be useful to integrate genome profiling with transcriptome profiling. Several approaches are now available to scan the genome for gains and losses. These include fluorescence *in situ* hybridization, comparative genomic hybridization, hybridization of genomic DNA to various types of DNA microarrays, and restriction landmark genome scanning[5–8]. Additionally, oligonucleotide arrays are now available that can be used to detect single-nucleotide polymorphisms and that allow genome-wide loss-of-heterozygosity maps to be developed from tumours, including samples isolated by laser-capture microdissection[12].

Pollack *et al.* profiled DNA copy-number alterations across 6,691 mapped human genes in 44 samples of predominantly advanced, primary breast tumours and 10 breast cancer

cell lines[13]. Parallel DNA microarray-based measurements of mRNA levels allowed assessment of the extent to which variation in gene copy number contributes to variation in gene expression in tumour cells. 62% of highly amplified genes showed increased expression levels. Additionally, DNA copy number correlated with gene expression across a range of DNA copy-number alterations, including deletions. On average, a twofold change in DNA copy number was associated with a corresponding 1.5-fold change in mRNA levels. It was estimated that overall, at least 12% of all the variation in gene expression among the breast tumours analysed was attributable to underlying variation in gene copy number, the remainder presumably attributable to a multitude of other factors.

In another study[14], restriction landmark genomic scanning was used to detect amplified genomic DNA fragments in 47 primary ovarian tumours. This approach uncovered amplification of the *LMYC* oncogene in several tumours. Transcriptome profiling of these tumours using oligonucleotide microarrays demonstrated frequent overexpression of *LMYC* in tumour cells, compared with cells of the normal ovarian surface epithelium — even in tumours without genomic amplification of *LMYC* — indicating that tumours use different mechanisms to upregulate *LMYC* expression. This finding prompted an assessment of the expression status of various members of the *MYC* gene family in ovarian tumours. Interestingly, a pattern was uncovered in which deregulated expression of one of the members of the *MYC* gene family was observed in most of the tumours.

Integrating transcriptome and proteome profiling. There is a need to profile gene expression at the level of the proteome and to correlate changes in gene-expression profiles with changes in proteomic profiles. The two are not always linked — numerous alterations occur in protein levels that are not reflected at the RNA level[15]. Translational control is an important cellular process that is regulated by several genes with tumour-suppressor or oncogenic properties[16]. For example, the proteins encoded by the tumour-suppressor genes tuberous sclerosis 1 (*TSC1*) and *TSC2* form a functional complex that inhibits the phosphorylation of S6 kinase and 4EBP1 — two key regulators of mRNA translation. TSC2 functions as a key regulator of the TOR pathway, which regulates protein synthesis, cell growth and viability in response to changes in cellular energy levels[17].

Given the distinct regulation of RNA and protein levels, integration of data pertaining to RNA and protein products that are encoded by the same genes can tell us a lot about tumour function. Nishizuka *et al.* analysed gene-expression patterns of 60 human cancer cell lines (NCI-60) used by the National Cancer Institute to screen compounds for anticancer activity, and measured levels of 52 cancer-related proteins in these cells[18]. Clustered image maps of protein levels uncovered two markers that could be used to distinguish colon from ovarian adenocarcinomas. Integration of protein and mRNA data led to the interesting observation that the levels of structural proteins were highly correlated with the levels of their corresponding mRNAs in the NCI-60 cell lines, whereas the levels of non-structural proteins were poorly correlated with those of their corresponding mRNAs.

Gene-expression and proteomic data sets from lung tumours have also been compared and integrated, along with serum samples from the same patients[19–21]. To determine whether gene-expression profiles could be used in prognosis, mRNA profiles in tumours from 86 newly diagnosed patients, including 67 with early-stage and 19 with advanced-stage lung adenocarcinoma, were measured by oligonucleotide microarray analysis[19]. A gene-expression index, based on expression of the genes that correlated with survival of the 86 patients, was able to identify low-risk and high-risk groups among the patients with stage-I lung adenocarcinomas. The index included many novel genes that were not previously associated with survival in lung adenocarcinoma. A large number of genes, such as the *CRK* oncogene, showed a graded pattern of expression among the tumours. A small number of genes, such as *ERBB2*, were only overexpressed in a small number of tumours, but were also correlated with poor outcome.

In parallel, proteomic studies were undertaken to identify proteins associated with patient outcome[20]. A leave-one-out cross-validation procedure that analysed proteins associated with patient outcome — which were identified by Cox modelling — indicated that specific protein profiles can be used to predict the likelihood of survival in patients with stage-I tumours. Integration of RNA and protein data from the same tumours, and from an independent study, showed that 11 of 27 mRNAs associated with survival were represented in the profile of survival-associated proteins. Interestingly, combined analysis of protein and mRNA data revealed that 11 components of the glycolysis pathway were associated with poor outcome, either at the protein or RNA levels. Phosphoglycerate kinase 1 expression was associated with reduced patient survival time, based on both RNA and protein studies, and also based on immunohistochemistry analysis using tissue microarrays in an independent validation set of 117 lung tumours. The relative abundance of this protein in tumours led to the assessment of its levels in the sera of patients with lung cancer, revealing a correlation between increased serum levels of phosphoglycerate kinase 1 and poor outcome.

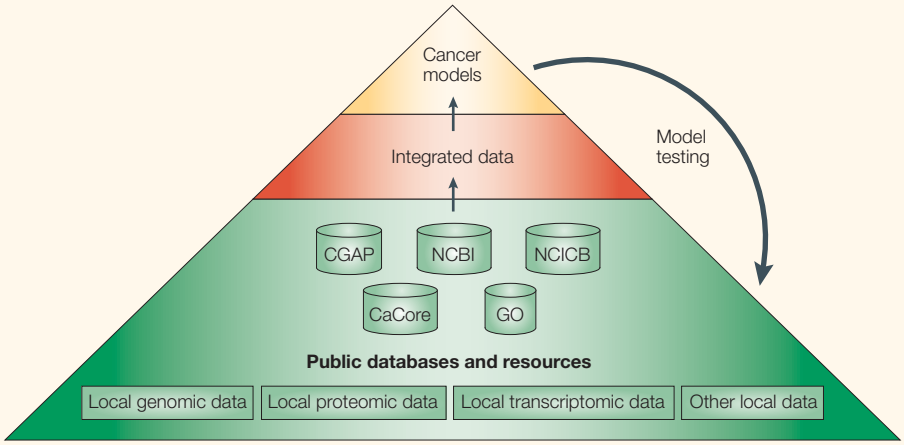

Figure 3 | **Path from data collection and integration to hypothesis testing.** Data produced by one research group can be combined with data in public databases such as the Cancer Genome Anatomy Project (CGAP) and further processed through resources available through various web sites — for example, the National Center for Biotechnology Information (NCBI), National Cancer Institute Center for Bioinformatics (NCICB), CaCore and Gene Ontology (GO) web sites — to yield integrated data sets (for further information on these web sites, see the online links box). This type of 'data mining' using statistical and informatics tools can lead to models for tumour behaviours such as metastasis, recurrence or response to therapy. Models can then be tested experimentally and/or through collection and analysis of additional data sets, and then refined.

Challenges

The studies presented above, although relatively simple from the point of view of extent of integration of heterogeneous data sets, illustrate the merits of an integrated approach to tumour profiling. However, collecting and integrating sets of data that are quite diverse represents a substantial undertaking that necessitates resources not available to most investigators. Experimental data must be processed and stored in a manner that is compatible with integration with other external, scattered data sources. Further complications stem from the substantial variation in the nomenclature used to identify the same object and to designate its attributes. For example, the protein encoded by a gene can be designated differently from the gene itself. Annotation with controlled vocabularies is required to achieve comparability across data sets. Even with adequate resources, the data generated is not always sufficiently reliable for a meaningful integrated analysis. For example, for genes that are expressed at very low levels, mRNA and protein levels can show a lack of correlation simply because of the limited sensitivity of the measurements.

Another serious challenge to studying cancer pathogenesis is the effectiveness of developing models capable of accounting for all the data collected with different high-throughput approaches. Although researchers have attempted for many years to devise mathematical models for many aspects of cancer, such as for tumour growth[22], tumour drug delivery[23] or gene–environment interactions[24], it is challenging to develop models that integrate the numerous pathways and factors that operate at various levels during tumour growth. Development of a model that would be able to predict the consequences of a particular mutation for tumorigenesis is more difficult than predicting the consequences of a mutation for a simple system, such as for a cultured microorganism.

Models of human cancer are also impaired by the substantial lack of homogeneity among study populations and, most importantly, by the inability to manipulate components of the system. Furthermore, numerous members of the 'parts list' that is required to construct any model can not be measured or manipulated, or might not even have been identified. Initially, models might therefore represent approximations that generate hypotheses to be tested through further experimentation. Further experiments could then yield additional data to allow a more robust model to be developed (FIG. 3). For example, the finding that upregulation of glycolytic enzymes correlates with poor outcome in patients with lung cancer led to the finding that increased activity of the transcription factor hypoxia-inducible factor-1α (HIF1α), which is known to regulate expression of glycolytic-pathway genes, was also correlated with poor survival in patients with lung cancer[25]. HIF1α has since been associated with numerous tumour types.

Resources

An expanding array of resources, in the form of databases and tools, is available to allow experimental global profiling data and other types of data to be integrated. Fortunately, a large amount of data has become available on gene expression in normal and cancer cells through initiatives such as the Cancer Genome Anatomy Project and the Director's Challenge initiative, funded by the National Cancer Institute (NCI). There are also numerous other relevant data repositories. So an investigator who finds that a specific gene is upregulated in a certain tumour type would be able to learn more about the expression pattern of this gene in other tumour types, as well as in normal tissues, through various gene-expression databases (BOX 1).

There are now numerous resources available for mining data from various global-profiling techniques. One of the first publicly available web databases of pathway information is the Kyoto Encyclopedia of Genes and Genomes (KEGG)[26]. Over 150 pathways are represented with emphasis on well-defined metabolic pathways. The KEGG pathway reference diagrams can be readily integrated with genomic and proteomic data. GenMAPP (Gene MicroArray Pathway Profiler) is a freely available program for viewing and analyzing expression data on 'microarray pathway profiles' (MAPPs) representing biological pathways or any other functional grouping of genes[27]. Over 50 MAPP files depicting various biological pathways and gene families are available. GenMAPP includes gene annotation information as described by the Gene Ontology (GO) Consortium[28]. The GenMAPP program identifies GO terms that seem to be over-represented in a data set, providing clues to relevant biological processes. Transpath is an online web database on signal transduction and gene-regulatory pathways that lists over 15,000 protein–protein interactions involving several thousand genes[29]. The Kinase Pathway Database[30] uses a natural language processing algorithm to automatically extract protein interaction information from the literature.

Other resources include public databases of protein–protein interactions, namely the Biomolecular Interaction Database (BIND)[31] and the Database of Interacting Proteins[32]. However, the organism most represented in these databases is *Saccharomyces cerevisiae*, for which substantial protein–protein interaction data have been generated. (For further information on the resources discussed above and in the following section, see the online links box.)

Box 1 | Some of the resources available for 'data mining' in cancer research

In addition to maintaining the GenBank nucleic-acid sequence database, the National Center for Biotechnology Information (NCBI) provides data analysis and retrieval resources for the data in GenBank and other biological data made available through the NCBI web site[35]. Relevant NCBI resources include the Cancer Chromosome Aberration Project, Entrez Genomes and related tools, the Map Viewer, Model Maker, Evidence Viewer, the Clusters of Orthologous Groups (COGs) database, SAGEmap, Gene Expression Omnibus (GEO) and the Molecular Modeling Database (MMDB). There are also available custom implementations of the BLAST program that are optimized to search specialized data sets. The National Cancer Institute, through its Center for Bioinformatics, provides informatics infrastructure support to advance translational cancer research. The centre provides open access to large and diverse data sets that result from NCI-funded initiatives. It also provides a resource that integrates such data with outside data and provides facilities for data management and distribution. The resource, designated CaCore, consists of a series of component technologies and services[36]. Enterprise Vocabulary Services provide controlled vocabulary, dictionary and thesaurus services. The Cancer Data Standards Repository provides a meta-data registry for common data elements. Cancer Bioinformatics Infrastructure Objects (caBIO) implements an object-oriented model of the biomedical domain and provides Java, Simple Object Access Protocol and HTTP-XML application programming interfaces.

Other resources include GoMiner, developed by Zeeberg *et al.*[37]. GoMiner is a resource package that organizes lists of genes, such as under- and overexpressed genes from a range of microarray experiments[28]. GoMiner provides quantitative and statistical output files and visualization graph structures. Genes that are displayed in GoMiner are linked to the main public bioinformatics resources. (For further information on the resources discussed above, see the online links box.)

Future directions

Clearly, additional resources are needed to facilitate integration of diverse data sets. The NCI plans to deploy an integrating biomedical informatics infrastructure called the Cancer Biomedical Informatics Grid (CaBIG), which will be developed in partnership with the cancer-research community. Around 50 cancer centres have joined this NCI-led project. The goals of CaBIG are to integrate data from diverse sources and to support interoperable analytic tools. The open-source, open-access grid will allow different research groups to search the expanding collection of cancer research data together with locally generated data. A similar and related effort is also underway in the United Kingdom, where the National Cancer Research Institute, which represents government, philanthropic and private-sector organizations that fund cancer research, has set up a unit to develop cancer research informatics. This will facilitate integration of data generated by laboratories across different organizations.

Apart from informatics considerations, tumour-profiling technologies would benefit from miniaturization of assays and increases in throughput and sensitivity, given the limited availability of tumour tissues. For example, the availability of proteome-scale capture agents would facilitate the use of microarrays in proteomic profiling, in a manner similar to transcriptome profiling. The availability of technologies for global profiling using formalin-fixed tissue would also be beneficial.

Understanding cancer as a complex disease, through systems-biology or systems-pathology approaches, requires teams of investigators from diverse fields such as biomedicine, chemistry, engineering, informatics and computational modelling. Soon, data obtained from molecular imaging studies might also be integrated. The continued development of sensitive molecular-imaging-based assays that do not require tissue samples will be valuable for monitoring molecular and cellular processes in both animal models of cancer and in humans[33]. Integration of molecular imaging with other molecular approaches to tissue analysis could add a spatial and a temporal perspective to our understanding of tumour development and progression.

The need for multidisciplinary research into cancer and other diseases has been recognized by the National Institutes of Health (NIH) with the implementation of the 'NIH roadmap'[34]. A systems-biology approach to cancer that incorporates different genome-scale global-profiling technologies is expected to lead to the development of computational models of gene regulation in cancer and important cancer-related cell processes, such as differentiation, proliferation, transformation and metastasis. This will lead to molecular-based classifications of cancer that transcend organ and tissue types — these should supercede classifications based on histopathology or based on the expression patterns of genes with unknown functional significance. New and important features of tumorigenesis and tumour progression will be uncovered in this manner, leading to more effective screening strategies and therapeutic targets.

Samir Hanash is at the Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M5-C800, PO Box 19024, Seattle, Washington 98109, USA.
e-mail: shanash@fhcrc.org

doi:10.1038/nrc1414

1. Tapon, N. Modeling transformation and metastasis in *Drosophila*. *Cancer Cell* **4**, 333–335 (2003).
2. Rangarajan, A. & Weinberg, R. A. Comparative biology of mouse versus human cells: modelling human cancer in mice. *Nature Rev. Cancer* **3**, 952–359 (2003).
3. Lamb, J. *et al.* A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell* **114**, 323–334 (2003).
4. Creighton, C. *et al.* Profiling of pathway-specific changes in gene expression following growth of human cancer cell lines transplanted into mice. *Genome Biol.* **4**, R46 (2003).
5. Albertson, D. G., Collins, C., McCormick, F. & Gray, J. W. Chromosome aberrations in solid tumors. *Nature Genet.* **34**, 369–376 (2003).
6. Albertson, D. G. & Pinkel, D. Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.* **2**, R145–R152 (2003).
7. Shi, H. *et al.* Triple analysis of the cancer epigenome: an integrated microarray system for assessing gene expression, DNA methylation, and histone acetylation. *Cancer Res.* **63**, 2164–2171 (2003).
8. Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C. & Vertino, P. M. Predicting aberrant CpG island methylation. *Proc. Natl Acad. Sci. USA* **100**, 12253–12258 (2003).
9. Ntzani, E. E. & Ioannidis, J. P. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* **362**, 1439–1444 (2003).
10. Simon, R. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br. J.Cancer* **89**, 1599–1604 (2003).
11. Rhodes, D. R. *et al.* Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA* **101**, 9309–9314 (2004).
12. Lieberfarb, M. E. & Lin, M. Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP. *Cancer Res.* **63**, 4781–4785 (2003).
13. Pollack, J. R. *et al.* Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA* **99**, 12963–12968 (2002).
14. Wu, R. *et al.* Amplification and overexpression of the *L-MYC* proto-oncogene in ovarian carcinomas. *Am. J. Pathol.* **162**, 1603–1610 (2003).
15. Hanash, S. Disease proteomics. *Nature* **422**, 226–232 (2003).
16. Ruggero, D. & Pandolfi, P. P. Does the ribosome translate cancer? *Nature Rev. Cancer* **3**, 179–192 (2003).
17. Inoki, K., Zhu, T. & Guan, K. L. TSC2 mediates cellular energy response to control cell growth and survival. *Cell* **115**, 577–590 (2003).
18. Nishizuka, S. & Charboneau, L. Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proc. Natl Acad. Sci. USA* **100**, 14229–14234 (2003).
19. Beer, D. G. *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinomas. *Nature Med.* **8**, 816–824 (2002).
20. Chen, G. *et al.* Protein profiles associated with survival in lung adenocarcinoma. *Proc. Natl Acad. Sci. USA* **100**, 13537–13542 (2003).
21. Brichory, F. M. *et al.* An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. *Proc. Natl Acad. Sci. USA* **98**, 9824–9829 (2001).
22. Albert, P. S. & Shih, J. H. Modeling tumor growth with random onset. *Biometrics* **59**, 897–906 (2003).
23. Telford, J. J., Saltzman, J. R., Kuntz, K. M. & Syngal, S. Impact of preoperative staging and chemoradiation versus postoperative chemoradiation on outcome in patients with rectal cancer: a decision analysis. *J. Natl Cancer Inst.* **96**, 191–201 (2004).
24. Merlino, G. & Noonan, F. P. Modeling gene-environment interactions in malignant melanoma. *Trends Mol. Med.* **9**, 102–108 (2003).
25. Semenza, G. L. Targeting HIF-1 for cancer therapy. *Nature Rev. Cancer* **3**, 721–732 (2003).
26. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
27. Doniger, S. W. *et al.* MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* **4**, R7 (2003).
28. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
29. Schacherer, F. *et al.* The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics* **17**, 1053–1057 (2001).
30. Koike, A., Kobayashi, Y. & Takagi, T. Kinase pathway database: An integrated protein-kinase and NLP-Based protein-interaction resources. *Genome Res.* **13**, 1231–1243 (2003).
31. Bader, G. D., Betel, D. & Hogue, C. W. BIND: The biomolecular interation network database. *Nucleic Acids Res.* **31**, 248–250 (2003).
32. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305 (2002).
33. Gelovani-Tjuvajev, J. & Blasberg, R. *In vivo* imaging of molecular-genetic targets for cancer therapy. *Cancer Cell* **4**, 327–333 (2003).
34. Zerhouni, E. The NIH Roadmap. *Science* **302**, 63–72 (2003).
35. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* **32** (Database issue), D35–D40 (2004).
36. Covitz, P. A. *et al.* caCORE: A common infrastructure for cancer informatics. *Bioinformatics* **19**, 2404–2412 (2003).
37. Zeeberg, B. R. *et al.* GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**, R28 (2003).

Competing interests statement
The author declares no competing financial interests.

Online links

DATABASES
The following terms in this article are linked online to:
Cancer.gov: http://cancer.gov/
breast cancer | lung cancer | ovarian cancer
Entrez Gene:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
CDK4 | *CRK* | cyclin D1 | *ERBB2* | *LMYC* | phosphoglycerate kinase 1 | *TP53* | *TSC1* | *TSC2*

FURTHER INFORMATION
ArrayExpress: http://www.ebi.ac.uk/arrayexpress
Biocarta: http://www.biocarta.com
Biomolecular Interaction Database:
http://www.blueprint.org/bind/bind.php
CaCORE: http://ncicb.nci.nih.gov/core
Cancer Biomedical Informatics Grid:
http://cabig.nci.nih.gov
Cancer Genome Anatomy Project: http://cgap.nci.nih.gov/
Cytoscape: http://www.cytoscape.org
Database of Interacting Proteins:
http://dip.doe-mbi.ucla.edu/