Mining gene expression profiles: expression signatures as cancer phenotypes

Joseph R. Nevins and Anil Potti

Abstract | Many examples highlight the power of gene expression profiles, or signatures, to inform an understanding of biological phenotypes. This is perhaps best seen in the context of cancer, where expression signatures have tremendous power to identify new subtypes and to predict clinical outcomes. Although the ability to interpret the meaning of the individual genes in these signatures remains a challenge, this does not diminish the power of the signature to characterize biological states. The use of these signatures as surrogate phenotypes has been particularly important, linking diverse experimental systems that dissect the complexity of biological systems with the *in vivo* setting in a way that was not previously feasible.

Classification
Use of an unsupervised analysis approach to identify classes of biological states, including clinical states, that often were not previously recognized.

Duke Institute for Genome Sciences and Policy, Duke University Medical Center, Durham, North Carolina 27710, USA.
Correspondence to J.R.N.
e-mail:
nevin001@mc.duke.edu
doi:10.1038/nrg2137
Published online 3 July 2007

The capacity to understand biological complexity is often limited by the ability to define relevant phenotypes. There is perhaps no better example of this challenge than that seen in cancer. The complexity of the oncogenic process, involving the somatic acquisition of large numbers of mutations coupled with variability in the host's genetic constitution, produces a disease of enormous complexity. Indeed, it is probably not an exaggeration to suggest that 100 breast cancer patients represent 100 distinct diseases. The ability to dissect this complexity to understand the unique characteristics of the individual patient is key in developing effective therapeutic strategies.

Traditional methods of characterizing tumours rely on gross visual information (size of the tumour, degree of spread, histological characteristics of the tumour) along with a limited number of biochemical assays (for hormone receptors, growth factor receptors, nuclear antigens, and so on). Although these tools do provide a way to define tumour subgroups with distinct biology, it is abundantly clear that these classifications are imprecise, creating heterogeneous groupings of tumours and patients. Numerous examples show that expression profiles can dissect this heterogeneity, beginning with work that used expression to distinguish acute myeloid and acute lymphoblastic leukaemias[1], and then followed by studies that identified previously unrecognized categories of diffuse large B-cell lymphoma[2]. This capacity to identify otherwise unrecognized biological distinctions lies in the scale of the analysis — a measure of 30,000 or more genes, reflecting the activity of the entire genome. But the scale goes well beyond that of 30,000 data points given the opportunity to identify patterns of gene expression, patterns that can be dynamic in response to both physiological and pathophysiological processes.

The advent of technology to measure gene expression on a genome-wide scale has transformed biology, perhaps more so than the advent of molecular biology in the 1970s. Genome-scale data has turned biology into a 'data science', now amenable to the power of various forms of quantitative and statistical analysis that were previously applied to complex data from finance, meteorology and other fields. Here we focus attention on one particular aspect of this transformation — the ability to develop gene expression signatures that reflect cancer phenotypes and the capacity to use these signatures as surrogates for these phenotypes.

The concept of gene expression signatures

An expression signature is simply a representation of a biological state, in the form of a pattern of gene expression that is unique to that circumstance. Underlying the concept is the realization that virtually any biological condition, whether a developmental state, a cellular response to extracellular ligands or a pathological state, is reflected in changes in gene expression. Although no single gene would have the power to define the biological state, the ability to measure and identify patterns of gene expression provides an opportunity to develop these signatures that reflect biological phenotypes. In a sense, the

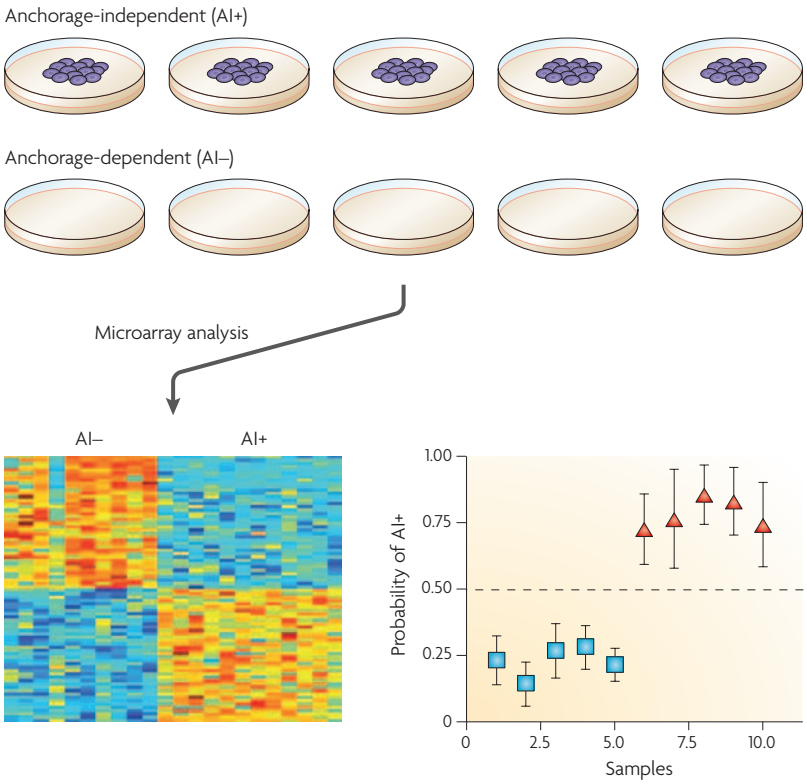

Figure 1 | **Generation of an expression signature.** A collection of cancer cell lines are assayed for growth in soft agar as a measure of anchorage-independent growth. RNA is prepared from the cells prior to the assay and used for DNA microarray analysis. These data are then used for a supervised analysis in which a signature is derived that distinguishes the two cell types (AI+ versus AI–). The extent to which the signature truly reflects the phenotype is assessed by leave-one-out cross validation to measure the predictive capacity of the signature.

expression signature becomes a surrogate representation of the biological phenotype.

The power of the expression signature is twofold. First, the enormous complexity of the expression data that can be sampled provides the opportunity to identify patterns of expression that reflect subtle distinctions in biology. The main limitation lies in the capacity to define the biological state of interest, whether through the generation of distinct experimental states that can then be used to train an analysis of expression, or by taking advantage of existing biological conditions that can be used as the training opportunity. Second, the expression signature is portable in the sense that it can be assayed in varied contexts — an expression signature that is developed in a cell-culture context can be measured in a tumour or a histological section. As such, the expression signature provides a link between otherwise heterologous systems. A cell-culture phenotype, such as pathway activation, is difficult to represent in a diverse sample such as a tumour. By contrast, an expression profile provides a mechanism to link these two states — the profile represents pathway activation in the cell culture that can then be used to interrogate the expression data from a tumour. In a sense, the gene expression signature becomes the common currency to connect the experimental state with the *in vivo* state.

Supervised analysis
An approach to gene expression analysis in which some aspects of the experimental samples are used to drive the analysis to identify a pattern (signature) of gene expression that characterizes the difference in the samples.

Prediction
The use of a signature to predict the state of an unknown sample, either through cross-validation procedures that use the training samples to evaluate the robustness of the signature, or by predicting an independent data set.

Developing a useful signature

Key to the development of useful expression signatures is the experimental condition. In many instances, these signatures are generated *in vitro*, using experimental settings that can be carefully controlled to define a specific biological process. To understand the concept of signatures, consider the cancer cell phenotype of anchorage-independent growth. This has been used historically as an experimental measure to model the invasive properties of tumour cells[3]. To develop an expression signature that is characteristic of this property, one could select a series of cancer cell lines that varied in their capacity to grow in soft agar, a measure of anchorage independence (FIG. 1). RNA is prepared from the various cell lines, both those that exhibit anchorage-independent growth and those that do not, and used for DNA microarray analysis. These gene expression data are then used in a supervised analysis (see below) to identify a pattern of gene expression — a signature — that reflects and, importantly, predicts the phenotype of anchorage-independent growth. The extent to which the profile actually reflects the biological state, rather than what would be observed by chance, can be assessed through statistical measures of predictive ability, typically a leave-one-out cross validation procedure, in which the profile is generated from a set of samples with one held out and then used to predict the state of that one sample. Further validation makes use of independent samples that reflect the phenotype of interest but were not used in the development of the signature. The ability of the signature that is developed from the initial training set to then accurately predict the set of test samples that is used for validation provides an important measure of the signature's robustness.

Importantly, this signature provides an opportunity to interrogate other samples, including tumour samples, to assess the extent to which they exhibit the signature. In this case, one simply uses the tumour sample, or whatever other set of samples one wishes to profile, as a validation case, measuring the probability that the microarray data that is derived from the sample shows the pattern that is defined by the training set that generated the signature. If done across a set of samples, such as a collection of tumour samples, one can generate a spectrum of probabilities of the signature. In the example in FIG. 2, these probabilities have been depicted as a heatmap, where red represents high probability and blue represents low probability. Moreover, if the set of samples is interrogated with multiple signatures, the results can be used for clustering, in much the same way as one would use the primary gene expression data to cluster samples. In this case, the clusters represent patterns of signatures across the samples.

As a consequence of a large number of studies making use of DNA microarray analysis in cancer contexts, a diverse set of expression signatures has emerged with relevance for cancer phenotypes that can be applied in future studies. In many instances, these signatures can be directly used for other applications and studies whereas, in other instances, the data can be 're-utilized' for the development of signatures in specific applications.

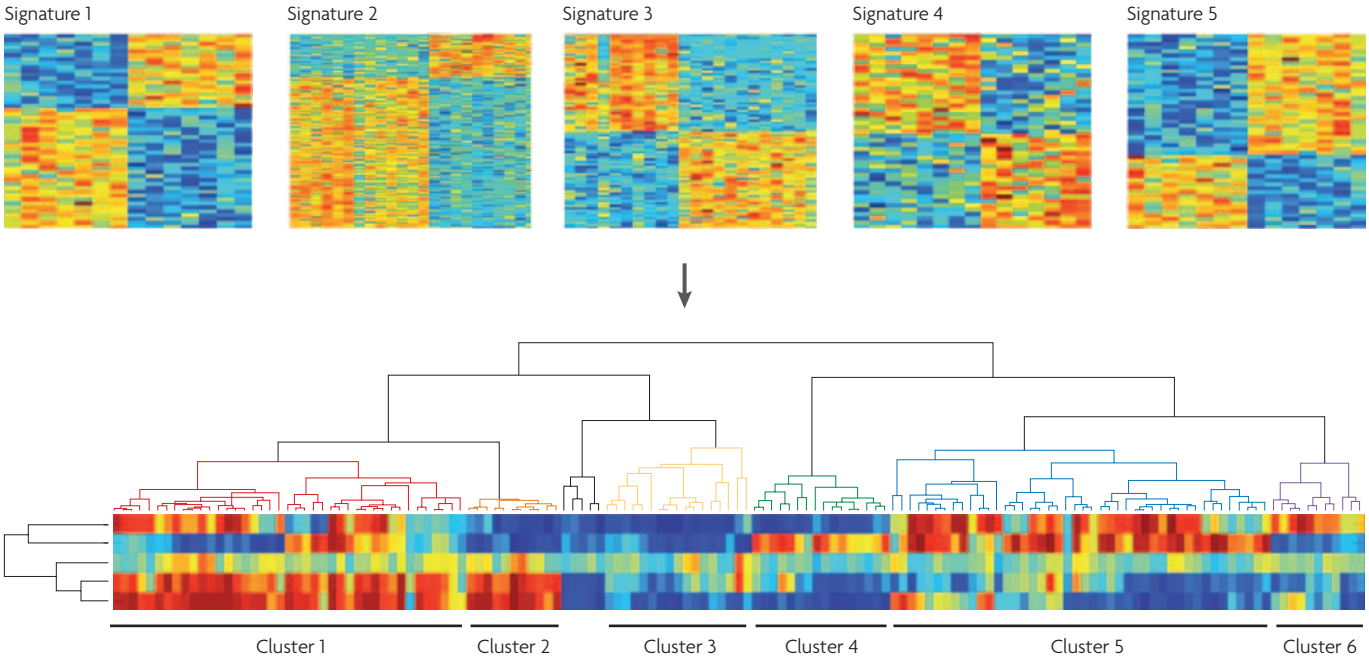

Figure 2 | **Identification of patterns of signatures.** In this example, a series of five signatures are developed to represent various biological states. These are then used to predict the status of the signature in a collection of tumour samples. The predictions are expressed as a probability measure reflecting the extent to which the tumour sample exhibits the signature. This probability is then displayed as a heatmap with red representing high probability and blue representing low probability. The samples are then clustered hierarchically on the basis of the predictions, revealing patterns of signatures across the samples.

An excellent resource for identifying the available opportunities is the Oncomine cancer-profiling database, developed at the University of Michigan[4–6], Ann Arbor, USA.

Strategies for gene-signature development

Two basic strategies have been described for the analysis of DNA microarray data[1]. One involves the discovery of structure in a given data set without regard for prior knowledge of the underlying biology — that is, no assumptions are made about what mechanisms might underlie a given gene expression profile. This approach, often referred to as 'unsupervised analysis', simply uses the gene expression data to find structure in the data that can then be used to infer biologically meaningful structure. As pioneered by Brown, Botstein and colleagues[7], this approach can be an effective tool in classifying biological samples into categories that were not previously known to exist. The power of this approach can be seen in the work of Perou and colleagues, who have used expression patterns to define clinically significant subtypes of human breast cancer[8,9].

Conceptually, unsupervised analysis is perhaps the simplest of the tools that are applied to the analysis of gene expression data to find patterns or clusters of similarly expressed genes that can then be studied further. There are now numerous examples in which this approach has been used to analyse gene expression data, often uncovering biological complexity that was not previously appreciated, including the identification of previously unrecognized tumour subtypes[1,2] or the refinement of tumour classification[10–14].

Although most studies that use expression data to dissect cancer phenotypes have relied on mRNA expression information, it is now clear that the expression of microRNAs (miRNAs) can be used to identify cancer-relevant signatures[15–19]. Although the data content of miRNA profiles is far less than that of mRNA profiles, the regulatory nature of miRNAs suggests that this class of gene products is enriched in information content. For a recent discussion of the role of miRNAs in cancer, including the identification of cancer-relevant signatures, see REF. 20.

By contrast, 'supervised analysis' strategies do consider existing information and, indeed, use it to guide the analysis of the gene expression data. This approach has been particularly useful in the identification of gene expression patterns that relate to clinically relevant phenotypes such as the ability to predict the potential for recurrence of disease[21–33]. The power of the supervised analysis lies in the ability to specifically drive the analysis to the phenotype of interest, taking advantage of the relevant information as a guide. Moreover, when the underlying biology that is relevant to the phenotype is uncertain, this might be the only approach to finding those gene expression patterns that relate to the phenotype. Equally important, if the clinical outcome reflects multiple components of the subtypes that are defined by unsupervised analysis, then it is only when the analysis is driven by the actual clinical outcome that this complexity can be addressed.

Unsupervised analysis
A form of gene expression analysis that involves discovery of empirical structure (patterns) in a given data set without regard to prior knowledge of the underlying biology. Gene expression patterns that are discovered in this manner then organize the biological samples.

Much of the work using genome-scale gene expression data has focused on biological discovery, identifying genes in the profiles that can lead to a better understanding of the biological processes and, perhaps, new therapeutic targets. It is logical to assume that if one can identify an expression profile that defines a unique new subtype of breast cancer, or one that predicts which tumour is most likely to metastasize, it should be possible to identify genes in these profiles that could represent new drug targets. Nevertheless, gaining an understanding of the underlying biology from the various gene expression profiles remains a challenge. Although one can often identify function associated with some of the genes in a signature, the challenge is to put this information into perspective with regard to the entire genomic profile. To appreciate this point, consider the development of a 70-gene profile that predicts breast cancer recurrence[29,30]. An examination of the genes in the profile reveals many that have annotations such as cell cycle, invasion, metastasis, angiogenesis and signal transduction — all obviously relevant for a cancer-recurrence phenotype. Nevertheless, they represent general categorizations and reflect only a small number of the genes from the overall profile. So, by focusing on these 'recognizable' genes, we ignore many of the genes in the profile. As such, it is entirely possible that the biological and clinical deductions regarding the few recognizable genes is taken out of context. In addition, it turns out that from the same data set multiple 70-gene profiles that predict breast cancer recurrence can be identified, thus expanding the number of informative genes and emphasizing the complexity of the biology that underlies this clinical phenotype[34].

An approach that was developed to go beyond the analysis of patterns on a gene-by-gene basis, known as gene-set enrichment analysis (GSEA), uses statistical measures of enrichment of annotated gene sets within expression profiles[35]. The value of GSEA and another, similar approach, 'gene module map'[36], is to attempt to examine the true context by looking at representations of gene sets that might better reflect the underlying biology. In other words, the power lies in the ability to tap into a wide array of gene sets, including signalling pathways. Although this is a step towards putting genes into a functional context, appreciating the true meaning of the individual genes or sets of genes in a signature remains a challenge.

Using cancer-relevant expression signatures

The concept of the expression signature as representing a distinct and well defined experimental state that can then be connected to an otherwise unrelated biological system opens the way for a myriad of applications to inform and understand biological complexity.

The ability of an expression signature to connect two states, where the expression signature is the intermediary, is exemplified in the recent studies of Golub and colleagues, which describe a 'Connectivity Map'[37]. In this case, expression profiles have been generated from a well-characterized cancer cell line that has been treated with many small molecule compounds so as to create a library of signatures of a drug response. This library of signatures is then used as a database that can be queried with expression information for other biological contexts, thereby linking otherwise disparate physiological events. In this context, the expression signature is represented as a group of gene identities, not by the actual properties of expression that are defined in the experimental setting. The power of this approach lies in its independence of the methodology for determining expression — that is, differences between assay platform or methods of measuring actual expression. Instead, a signature is simply a list of genes obtained from one experiment that is then compared to other gene lists to find matches, quantitatively assessing the degree of similarity between the gene lists.

A conceptually similar, although mechanistically distinct, approach to the Connectivity Map makes use of quantitative aspects of the expression profile to examine other samples for similar properties. In essence, the task is to define a profile or signature experimentally and then ask if the same pattern can be recognized in another sample. Although this approach requires that expression data are largely similar, because one is examining the extent to which the actual expression properties that are characteristic of the profile are present, this approach has the power of quantitative assessment — generally, a probability that the signature (phenotype) is seen in the test sample. Importantly, not only does the quantitation provide a means to estimate relative contributions of signatures across samples, or the relative contributions of multiple signatures in one sample, but it also provides the opportunity to identify patterns of signatures, not unlike the ability to find patterns of gene expression. Below, we provide several examples to show how the signature concept can be applied.

Identification of new therapeutic opportunities from the Connectivity Map. A search of the Connectivity Map database for profiles that coincided with a gene expression signature of glucocorticoid sensitivity or resistance in acute lymphoblastic leukaemia (ALL) cells revealed a link with a profile of the mTOR inhibitor rapamycin[38]. Further experimental assays provided evidence that rapamycin could induce glucocorticoid sensitivity in lymphoid tumour cells and sensitization to apoptosis through the modulation of the antiapoptotic gene *MCL1*. These findings provide a potential path towards a new therapeutic strategy involving rapamycin and glucocorticoid.

In another example, screens were carried out for compounds that inhibit a therapeutically relevant expression signature, in this case, androgen receptor (AR) signalling. The results were of limited use because the screen yielded compounds such as celastrol and gedunin, the molecular targets of which were unknown. However, comparison of these signatures for inhibition of AR signalling against the Connectivity Map library provided connections to heat shock protein 90 (HSP90) inhibitors, suggesting HSP90 inhibition as a strategy for controlling AR activity[39].

Signatures that predict pathway activation in cancer. One example of the use of signatures to define biology can be seen in the generation of signatures of oncogenic pathway deregulation. Again, defined experimental conditions are used to create circumstances that represent 'pathway off' or 'pathway on'; expression profiles that accurately reflect this process are then developed[40–46]. In one example, samples of quiescent cells versus RAS-, MYC-, SRC-, E2F3-, or β-catenin-expressing cells (all of which contribute to a cell proliferation response) were used to develop signatures that predict activation of these pathways[45]. The fact that these signatures do indeed reflect pathway activation is evident because they can be used to accurately predict the molecular basis of various mouse models of cancer. Specifically, tumours arising from mice that are transgenic for mouse mammary tumor virus (MMTV)–MYC or MMTV–HRAS, and tumours arising from the loss of retinoblastoma 1 (RB), were correctly identified with the appropriate pathway signature[42,43,45].

These results demonstrate the ability of these pathway profiles to predict tumours that arise from deregulation of the corresponding pathways, and provide evidence that this approach can be used as a basis for characterization of the status of the pathways in various tumour contexts. Importantly, the assay of gene expression profiles provides a measure of the consequence of the oncogenic process, irrespective of how the pathway might have been altered. As an example, one might revisit the classic studies of Fearon and Vogelstein that identified the accumulation of genetic alterations as a function of colon carcinoma initiation and progression from a premalignant adenomatous stage to a more aggressive carcinoma[47]. Using pathway signatures as representations of the consequence of the accumulated mutations, one might anticipate a somewhat different picture from that seen by gene mutation alone, with signatures reflecting the consequence of these mutations irrespective of how they might have been activated. For instance, the RAS pathway might be active either as a result of a mutation of RAS, a mutation of a growth receptor or an alteration in one of the RAS effector activities. Moreover, the signature-based pathway analysis approach provides an opportunity to integrate these signatures to discover patterns of pathway activation, adding further value to the analysis. Indeed, predictions of pathway status in a series of lung cancer samples, followed by clustering of the samples on the basis of the oncogenic pathway signatures, has revealed distinct patterns in which subgroups of tumours are identified using pathway patterns[45]. Furthermore, survival analysis based on the expression patterns reveals that the ability to integrate pathway analysis by identifying patterns of pathway deregulation does provide a way of better categorizing lung cancer patients.

Pathway predictors have also been shown to have the potential to direct the use of therapeutics that target a component of that pathway. In particular, a measure of probability of pathway activation in a panel of cancer cell lines, on the basis of the probability that is predicted from the pathway signature, reveals a relationship between the state of pathway deregulation and sensitivity to the respective therapeutic[45,48]. This relationship is further emphasized by studies showing that cells that harbour a mutation in *BRAF* (a proto-oncogene encoding serine/threonine-protein kinase) are sensitive to inhibition of MEK (mitogen-activated protein kinase kinase 1)[49]. Taken together, the results would suggest a potential strategy where the prediction of pathway activation, using a collection of signatures, could serve to guide the use of targeted therapeutic agents that otherwise lack a biomarker to direct their use.

Signatures that predict drug sensitivity. Although the strategy of the Connectivity Map is to identify signatures of drug effects, an alternative approach is to develop signatures that predict drug sensitivity. The ability to predict patient response to various cancer therapies, including commonly used cytotoxic chemotherapies, is key to achieving the goal of personalized medicine whereby patients are accurately matched to the therapy or therapies that best address the biology of their cancer. The logical approach is to make use of clinical studies in which patients are treated with a given drug, their response is measured and gene expression data are generated from samples that were collected before treatment in order to develop a profile that can predict response. Indeed, this strategy has already been used successfully in several studies[48,50,51]. The main limitation to this approach is that clinical studies are lengthy and focused on a particular regimen.

In an alternative strategy, the same goal is accomplished by using cancer cell lines grown in culture. In one such case, a panel of cancer cell lines were treated with dasatinib, a multitargeted kinase inhibitor, and sensitivity to the drug was measured. In parallel, expression data generated from the same panel of cell lines was used to develop a signature to predict sensitivity to the drug[52]. Interestingly, the cells that were defined in this manner as sensitive largely coincided with those showing evidence of SRC pathway activation, as measured by a SRC pathway signature[45]. A second example involves the use of panels of lung cancer cell lines to develop gene expression signatures that predict sensitivity to the epidermal growth factor receptor (EGFR) inhibitors gefitnib[53] or erlotinib[54]. Of these two, the erlotinib-sensitivity signature could more effectively predict sensitivity, possibly suggesting that a signature of sensitivity to EGFR-directed therapy captures the consequence of many events that contribute to the phenomenon of geftinib and erlotinib sensitivity.

An analogous strategy has taken advantage of existing drug-sensitivity data derived from the NCI-60 cancer cell line panel. This data set contains a wealth of drug-sensitivity measures coupled with baseline Affymetrix gene expression data. Cells that were identified as sensitive and resistant to various commonly used cytotoxic chemotherapies have been used to generate expression profiles that could predict this sensitivity[55,56]. Importantly, multiple predictors of chemosensitivity have been shown to also predict response to the drugs in patient studies[56].

NCI-60 cancer cell line panel
A collection of 59 human cancer cell lines derived from tumours of diverse origin and used for extensive analysis of drug sensitivity, using over 100,000 compounds. More recently, various genomic data sets have been generated using these cell lines, including chromosomal copy number and gene expression.

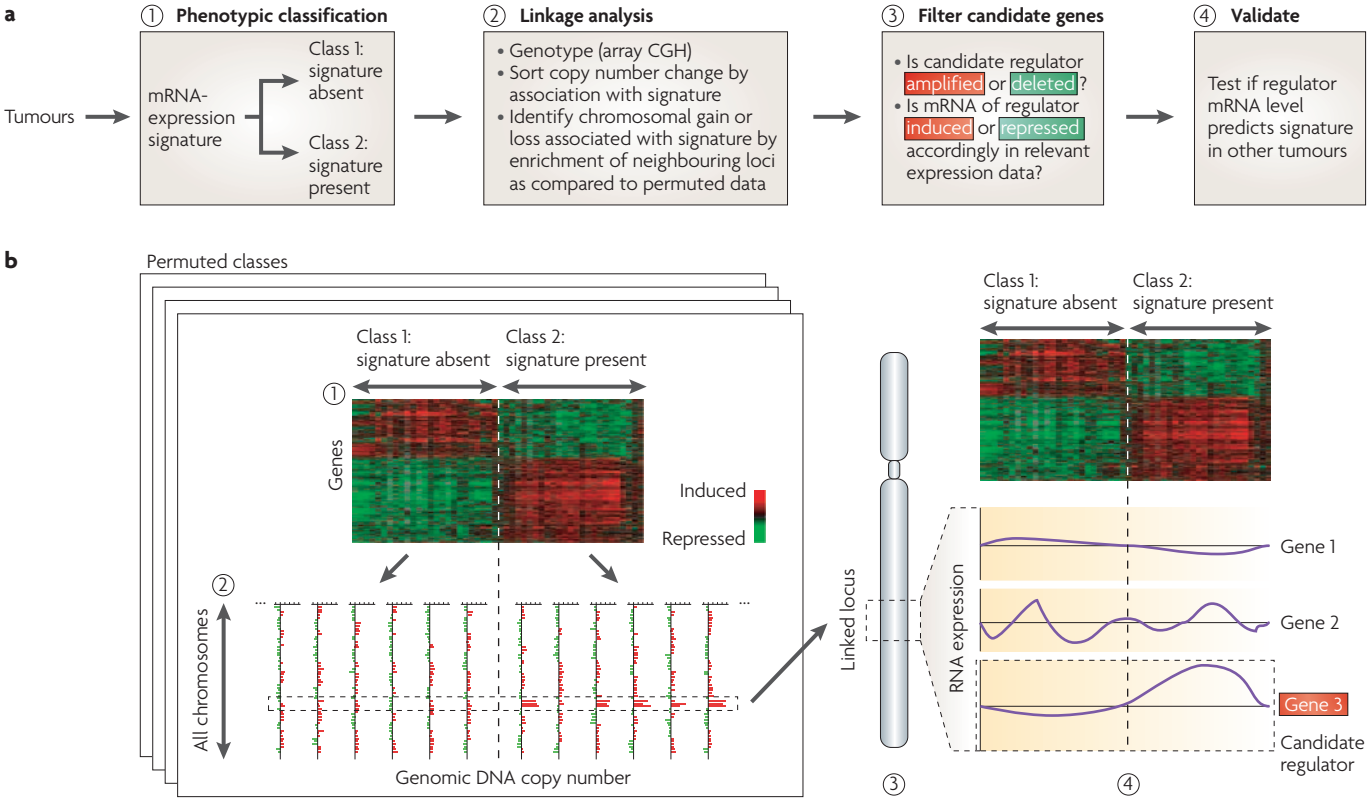

Figure 3 | **Stepwise linkage analysis of microarray signatures (SLAMS). a** | The use of an expression signature as a phenotype in a genetic linkage strategy to identify loci that show amplification in relation to the signature is shown. Samples are initially divided into two groups on the basis of their expression signature (**b**). Then, genome copy data is used to associate gene copy number changes with expression signature. Candidate genes are filtered on the basis of actual expression compared to the signature. Finally, candidate genes are further validated, on the basis of expression, in additional sample sets. Reproduced with permission from *Nature Genetics* REF. 59 © (2006) Macmillan Publishers Ltd.

Signature of chromosomal instability that predicts cancer outcome. Using the measured state of aneuploidy in a series of tumour samples, recent work has developed an expression signature that reflects chromosomal instability[57]. Given the variability in how aneuploidy might be measured, the signature provides a quantitative estimate of this trait that can then be applied to other cancer contexts, including an evaluation of the extent to which this phenotype predicts clinical outcome. Indeed, the signature was found to be elevated in metastatic specimens compared with primary tumours, providing a means to measure the role of chromosomal instability in determining malignant potential, possibly in a variety of tumours.

Using signatures as phenotypes for genetic discovery. In a recent study, an expression signature was used to identify genomic alterations that are tightly linked with this 'surrogate phenotype'. The signature was the so-called wound-response signature, developed experimentally on the basis of serum stimulation of fibroblasts in culture[58]. The authors analysed a series of breast tumours for evidence of the wound-healing phenotype, based on the expression signature, and looked for association with DNA copy number changes[59] (FIG. 3). The method of analysis, termed stepwise linkage analysis of microarray signatures (SLAMS), is derived from the concept of genetic association in which a genetic locus that is responsible for driving a phenotype is identified on the basis of co-segregation with individuals (in this case, tumour samples) that exhibit the phenotype (in this case, an expression signature). The authors showed that the wound-response signature, which is also a poor-prognosis expression pattern of 512 genes in breast cancer, is associated with coordinate amplifications of MYC and CSN5 (also known as JAB1 or COPS5). In short, the authors used a gene expression signature as a relevant breast cancer phenotype — or subphenotype — to carry out a genetic association study to identify those genomic alterations that are tightly linked to this phenotype.

The recent identification of the *MITF* gene as a melanoma oncogene made use of what is, in effect, a reversal of the SLAMS strategy. In particular, SNP copy number data from the NCI-60 panel was used to identify a melanoma-specific copy number gain on chromosome 3. This information was then used to identify expression patterns coincident with the SNP profile, leading to the identification of *MITF* [60].

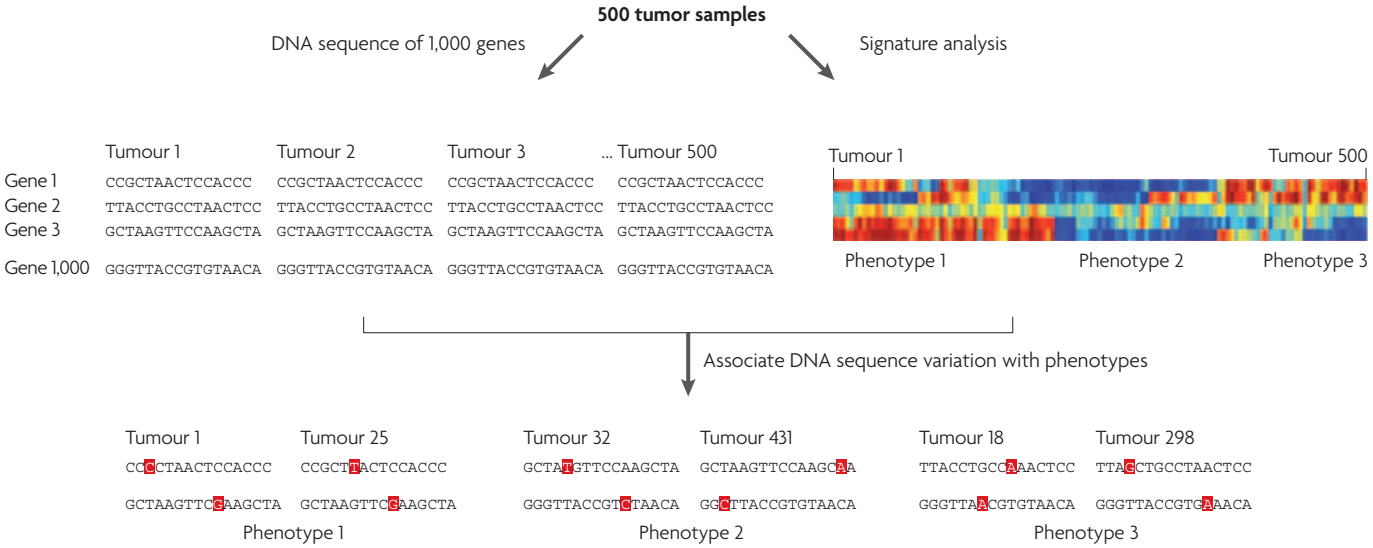

Figure 4 | **Use of signature patterns to identify complex cancer-relevant phenotypes to guide sequence analysis.** Sequence analysis of 1,000 selected genes in a collection of 500 tumour samples is shown. In parallel, the same tumour samples are used for gene expression analysis and signature profiling to produce clusters of signature profiles that reflect biological phenotypes. The patterns of expression signatures are then used as 'phenotypes' to guide the analysis of the DNA sequence data in a manner that is analogous to the concept of SLAMS (stepwise linkage analysis of microarray signatures; see FIG. 3). The goal is to identify DNA sequence variation in specific tumours that coincide with signature-based phenotypes.

Challenges of using expression signatures

Although the various initial studies that are summarized here provide strong evidence for the power of the concept of using expression signatures to define cancer phenotypes, it must be recognized that there are significant challenges inherent in this approach that reflect variation in both the biology and the technical approaches.

The molecular heterogeneity of cancer, resulting from the acquisition of multiple genetic alterations that contribute to the development of the tumour, underlies the heterogeneity of carcinogenesis. Indeed, this heterogeneity might well contribute to the observation that one can often find multiple expression profiles that predict a common phenotype such as breast cancer recurrence[34,61]. Technical variation can also be a factor that contributes to variation in profiles that are identified in distinct studies. Platform differences and batch effects between studies have a significant impact on the ability to consistently apply expression signatures to independent validations[62], potentially limiting the usefulness of the signatures. The impact of this technical variation is reduced by the Connectivity Map strategy, because it does not depend on actual expression data but rather just the identification of genes.

Although the use of expression signatures is conceptually simple and has clearly been shown to represent important biology in many instances, as with any approach, these methodologies must be further validated in larger studies that assess the extent to which a signature is robust across various studies. Furthermore, attention to appropriate statistical design in these studies, as well as demonstration of validation in multiple independent settings, will also be crucial to demonstrating the utility of using expression signatures in clinical practice to better understand cancer biology and to improve patient outcomes.

Future directions

The development of a library of expression-based signatures that can serve as surrogate phenotypes for various biological states represents a powerful tool for dissecting the complexity of biological processes. Linking biological states with specific genetic alterations and the capacity to utilize the information should guide new therapeutic strategies. The ability to integrate these signatures into a coherent view of biological pathways and systems will provide an opportunity to address the complexities of cellular signalling events, particularly the interplay of various key pathways.

Taking advantage of complexity by integration of signatures. The use of expression signatures as cancer phenotypes has already been shown to have significant value in dissecting complex biological states. The challenge for the future will be to take these analyses to a higher level, integrating the various signatures, or phenotypes, into a more complex view of the biological states, including the structure and function of cellular signalling events. Moreover, the power of 'combined signatures' will probably be improved by the integration of new signatures that represent different biological events such as pathway activation or chemosensitivity as opposed to what is achieved by just combining prognostic signatures. Indeed, recent examples have highlighted the potential power of integrating diverse signatures with clinical information to yield more robust predictive models in breast cancer[32,61].

Using signatures to guide the interpretation of DNA sequence variation. The Cancer Genome Atlas initiative is a recently initiated large-scale project with the ambitious goal of characterizing the DNA sequence variation that is unique to three human cancers. Perhaps the most significant challenge for the overall Cancer Genome Atlas initiative will be the ability to interpret the DNA sequence information from a collection of tumour samples — to identify those sequence variations that are significant with respect to a cancer phenotype. This is already illustrated by the initial observations of the cancer genome sequencing efforts: the identification of large numbers of mutations makes it difficult to interpret the role of these alterations in defining the phenotype[62–65]. We believe that one major limitation in this work is the lack of more precise phenotypes that are linked to the sequence data. Cancer is a complex disease and biological process with tremendous variation and heterogeneity, even within a single cancer type. Therefore, an ability to better describe individual phenotypes that might result from a unique collection of DNA sequence alterations offers the potential for better linking the sequence data with clinical phenotypes.

A potentially powerful way of addressing this issue would be to use the multitude of cancer-relevant phenotypes that are generated from expression signatures together with the concept of SLAMS, as described by Chang and colleagues[59]. In principle, this concept of linking a phenotype, as defined by a gene expression signature, to a genetic alteration can be extended to the analysis of DNA sequence variation identified across cancer genomes (FIG. 4). Various cancer-relevant signatures could be used as cancer subphenotypes to drive the analysis of DNA sequence variation in a manner that is analogous to the work of Chang and colleagues. These signatures could be derived from many that already exist, reflecting diverse aspects of cancer biology such as oncogenic signalling pathway activation, clinical outcomes and response to therapy. Alternatively, they could represent newly designed signatures that reflect other relevant aspects of cancer biology. Importantly, the ability to identify patterns of signatures, as shown by the example in FIG. 4, provides an opportunity to go even further in the use of this information to derive cancer-relevant phenotypes.

Conclusions

The ability to use expression signatures to dissect the complexity of cancer phenotypes and ultimately to discover the mechanisms that underlie cancer have been amply demonstrated by now. The development of expression signatures that is summarized here is just the beginning — we envision that many more representations of important experimental and clinical biology can be utilized for the development of these powerful tools. Equally important will be an ability to understand the biological processes that are reflected in the signatures, placing the expression and function of genes in a broader context of pathways and networks. But, even before this important goal is achieved, the expression signatures have significant value as phenotypes, irrespective of an understanding of the specific genes that constitute the profile.

1. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
The paradigm for use of both unsupervised and supervised methods of gene expression analysis to define new classes of tumours and predict these classes in new samples.
2. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
3. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
4. Rhodes, D. R. *et al.* Mining for regulatory programs in the cancer transcriptome. *Nature Genet.* **37**, 579–583 (2005).
5. Rhodes, D. R. & Chinnaiyan, A. M. Integrative analysis of the cancer transcriptome. *Nature Genet.* **37**, S31–S37 (2005).
6. 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).
7. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
8. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA* **100**, 8418–8423 (2003).
9. Perou, C. M. *et al.* Molecular portraits of human breast tumors. *Nature* **406**, 747–752 (2000).
10. Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA* **98**, 15149–15154 (2001).
11. Lapointe, J. *et al.* Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA* **101**, 811–816 (2004).
12. Hayes, D. N. *et al.* Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J. Clin. Oncol.* **24**, 5079–5090 (2006).
13. Dave, S. S. *et al.* Molecular diagnosis of Burkitt's lymphoma. *N. Engl. J. Med.* **354**, 2431–2442 (2006).
14. Dave, S. S. *et al.* Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N. Engl. J. Med.* **351**, 2159–2169 (2004).
15. He, L. *et al.* A microRNA polycistron as a potential human oncogene. *Nature* **435**, 828–833 (2005).
16. Yanaihara, N. *et al.* Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell* **9**, 189–198 (2006).
17. Lu, J. *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435**, 834–838 (2005).
18. Iorio, M. V. *et al.* MicroRNA gene expression deregulation in human breast cancer. *Cancer Res.* **65**, 7065–7070 (2005).
19. Volinia, S. *et al.* A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc. Natl Acad. Sci. USA* **103**, 2257–2261 (2006).
20. Calin, G. A. & Croce, C. M. MicroRNA signatures in human cancers. *Nature Rev. Cancer* **6**, 857–866 (2006).
21. Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. A molecular signature of metastasis in primary solid tumors. *Nature Genet.* **33**, 59–54 (2003).
22. Berchuck, A. *et al.* Patterns of gene expression that characterize long term survival in advanced serous ovarian cancers. *Clin. Cancer Res.* **11**, 3686–3696 (2005).
23. Beer, D. G. *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med.* **8**, 816–824 (2002).
24. Potti, A. *et al.* A genomic strategy to refine prognosis in non-small cell lung carcinoma. *N. Engl. J. Med.* **355**, 570–580 (2006).
25. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
26. Huang, E. *et al.* Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–1596 (2003).
27. Shipp, M. A. *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Med.* **8**, 68–74 (2002).
28. Pomeroy, S. L. *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442 (2002).
29. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
30. van'T Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
The initial example of a gene expression profile developed to refine and improve clinical prognosis.
31. Singh, D. *et al.* Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–209 (2002).
32. Pittman, J. *et al.* Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc. Natl Acad. Sci. USA* **101**, 8431–8436 (2004).
33. West, M. *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA* **98**, 11462–11467 (2001).
34. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171–178 (2005).
35. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
A description of GSEA methodology as a tool to identify biological context in gene expression profiles.

36. Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nature Genet.* **36**, 1090–1098 (2004).
37. Lamb, J. *et al.* The Connectivity Map: using gene expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006). **A description of a novel approach to connect two biological states, using gene expression as the intermediary.**
38. Wei, G. *et al.* Gene expression based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell* **10**, 331–342 (2006).
39. Hieronymus, H. *et al.* Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* **10**, 321–330 (2006).
40. Desai, K. V. *et al.* Initiating oncogenic event determines gene-expression patterns of human breast cancer models. *Proc. Natl Acad. Sci. USA* **99**, 6967–6972 (2002).
41. Ferrando, A. A. *et al.* Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* **1**, 75–87 (2002).
42. Huang, E. *et al.* Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature Genet.* **34**, 226–230 (2003).
43. Black, E. P. *et al.* Distinct gene expression phenotypes of cells lacking RB and RB family members. *Cancer Res.* **63**, 3716–3723 (2003).
44. Sweet-Cordero, A. *et al.* An oncogenic KRAS2 expression signature identified by cross-species gene expression analysis. *Nature Genet.* **37**, 48–54 (2005).
45. Bild, A. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357 (2006). **The development of gene expression signatures that reflect the activation or deregulation of various oncogenic signalling pathways, together with the utilization of these signatures to predict sensitivity to drugs that target the pathways.**
46. 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).
47. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **17**, 671–674 (1990). **The first comprehensive description of genetic events involved in colon carcinogenesis.**
48. Dressman, H. K. *et al.* An integrated genomic-based approach to individualized treatment of patients with advanced stage ovarian cancer. *J. Clin. Oncol.* **25**, 517–525 (2007).
49. Solit, D. B. *et al.* BRAF mutation predicts sensitivity to MEK inhibition. *Nature* **439**, 274–275 (2006).
50. Chang, J. C. *et al.* Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* **362**, 362–369 (2003).
51. Ayers, M. *et al.* Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. *J. Clin. Oncol.* **22**, 2284–2293 (2004).
52. Huang, F. *et al.* Identification of candidate molecular markers predicting sensitivity in solid tumors to dasatinib: rationale for patient selection. *Cancer Res.* **67**, 2226–2238 (2007).
53. Coldren, C. D. *et al.* Baseline gene expression predicts sensitivity to gefitinib in non-small cell lung cancer cell lines. *Mol. Cancer Res.* **4**, 521–528 (2006).
54. Balko, J. M. *et al.* Gene expression patterns that predict sensitivity to epidermal growth factor receptor tyrosine kinase inhibitors in lung cancer cell lines and human lung tumors. *BMC Genomics* **7**, 289 (2006).
55. Staunton, J. E. *et al.* Chemosensitivity prediction by transcriptional profiling. *Proc. Natl Acad. Sci. USA* **98**, 10787–19792 (2001).
56. Potti, A. *et al.* A genomic strategy to guide the use of chemotherapeutic drugs in solid tumors. *Nature Med.* **12**, 1294–1300 (2006).
57. Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nature Genet.* **38**, 973–974 (2006).
58. Chang, H. Y. *et al.* Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol.* **2**, 206–214 (2004).
59. Adler, A. S. *et al.* Genetic regulators of large-scale transcriptional signatures in cancer. *Nature Genet.* **38**, 421–430 (2006). **A demonstration of the use of an expression signature as a cancer phenotype in a genetic association study to identify chromosomal alterations that associate with the phenotype.**
60. Garraway, L. A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–122 (2005).
61. Fan, C. *et al.* Different gene expression based predictors for breast cancer patients are concordant. *N. Engl. J. Med.* **355**, 560–569 (2006).
62. The MicroArray Quality Control Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnol.* **24**, 1151–1161 (2006).
63. Chang, H. Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl Acad. Sci. USA* **102**, 3738–3743 (2005).
64. Davies, H. *et al.* Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res.* **65**, 7591–7595 (2005).
65. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
66. Parsons, D. W. *et al.* Colorectal cancer: mutations in a signalling pathway. *Nature* **436**, 792 (2005).
67. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).

Acknowledgements
We thank A. Bild for many helpful discussions and K. Culler for assistance with the preparation of the manuscript.

Competing interests statement
The authors declare no competing financial interests.

DATABASES
The following terms in this article are linked online to:
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
BRAF | *MCL1* | *MITF*
UniProtKB: http://ca.expasy.org/sprot
AR | β-catenin | CSN5 | EGFR | E2F3 | MYC | SRC

FURTHER INFORMATION
Oncomine: http://www.oncomine.org/main
Cancer Genome Atlas: http: //cancergenome.nih.gov
Access to this links box is available online.