

Integrative analysis of the cancer transcriptome

Daniel R Rhodes & Arul M Chinnaiyan

DNA microarrays have been widely applied to the study of human cancer, delineating myriad molecular subtypes of cancer, many of which are associated with distinct biological underpinnings, disease progression and treatment response. These primary analyses have begun to decipher the molecular heterogeneity of cancer, but integrative analyses that evaluate cancer transcriptome data in the context of other data sources are often capable of extracting deeper biological insight from the data. Here we discuss several such integrative computational and analytical approaches, including meta-analysis, functional enrichment analysis, interactome analysis, transcriptional network analysis and integrative model system analysis.

The widespread application of DNA microarrays to cancer research is nothing less than astounding. In the short ten-year history of this versatile technology, hundreds of large-scale experiments have been done, generating global quantitative profiles of gene expression in cancer. Known types and subtypes of cancer have been readily distinguished by their gene-expression patterns, and more importantly, new molecular subtypes of cancer have been discovered that are associated with a host of tumor properties, including the propensity to metastasize and sensitivity or resistance to particular therapies. The clinical utility of array-based gene profiles is evidenced by recent studies showing that cancer gene-expression signatures may affect clinical decision-making in breast cancer and lymphoma management^{1,2}. It may not be long before every human cancer sample is profiled with a gene chip to ascertain a molecular diagnosis and prognosis and to define an optimal treatment strategy. Before this becomes a reality, however, careful validation to identify optimal signatures is needed³. Apart from the impact of microarrays on clinical decision-making, cancer microarray profiling is also poised to advance our understanding of cancer biology and, ultimately, aid in the development of new and more effective therapies.

The biologist's initial instinct might be to use cancer microarray data as a prioritized list of candidate genes for experimental work-up, hoping to strike gold and identify a gene important in cancer among a sizeable differential expression profile. Although this approach has identified several genes important in cancer, it is akin to asking what makes an airplane different from an automobile, taking both apart,

making a list of differences in the parts and then focusing on a single part. Opposite to this single-part approach, a new line of attack seeks to examine the cancer profile as a whole, often in the context of other cancer signatures or other types of genomic data. Such integrative approaches are capable of simplifying complex cancer signatures into coordinately regulated modules, transforming one-dimensional cancer signatures into multidimensional interaction networks and extracting regulatory mechanisms encoded in cancer gene expression. Here, we review approaches that glean biological insight from cancer microarray data by applying integrative computational and analytical methodologies.

Before exploring the integrative analyses carried out on cancer transcriptome data, it is useful to provide a brief overview of the cancer profiling field. In the past few years, we have witnessed an explosion of cancer profiling studies. Once a technology available in only a few laboratories, DNA microarrays now seem as pervasive as PCR. Querying the Affymetrix database of publications for all reports relating to human cancer returned 646 primary research articles, 453 of which were published in the last two years. GEO (Gene Expression Omnibus), which has emerged as one of the principal repositories for microarray data^{4,5}, returned 84 data sets when queried for 'Homo sapiens' and 'cancer.' The Oncomine database (<http://www.oncomine.org/>)⁶, which includes data sets that have profiled ten or more human tumor samples (excluding cell line studies), has catalogued 300 primary research articles and amassed 114 data sets, totaling > 8,000 microarray experiments, each profiling a distinct human tissue sample. From this large body of cancer gene-expression research, several themes have emerged. First, cancer types can be reliably distinguished from normal tissue of the same type based on global gene-expression patterns. Second, predominant clinical and pathological subtypes of cancer often have distinct gene-expression profiles. Third, gene-expression signatures of primary tumors can often predict disease recurrence, distant metastasis, survival and treatment response. Fourth, heterogeneous cancers can be subclassified into molecular subtypes on the basis of gene-expression signatures⁷. These results were generated with primary analytical methods such as hierarchical clustering⁸ and statistically based differential expression analysis^{9,10}, usually with careful consideration for multiple-hypothesis testing¹¹.

Although these primary analyses have made great strides in deciphering the complex molecular heterogeneity of cancer, integrative bioinformatics approaches that leverage multiple types of information have begun to show promise in uncovering important biology not apparent from standard analysis methods (*i.e.*, differential expression analysis, hierarchical clustering, etc.). We highlight several such approaches, including meta-analysis for extracting robust profiles from

Daniel R. Rhodes is in the Departments of Pathology and Bioinformatics and the Comprehensive Cancer Center, and Arul M. Chinnaiyan is in the Departments of Pathology, Bioinformatics and Urology and the Comprehensive Cancer Center at the University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. e-mail: arul@umich.edu

Published online 26 May 2005; doi:10.1038/ng1570

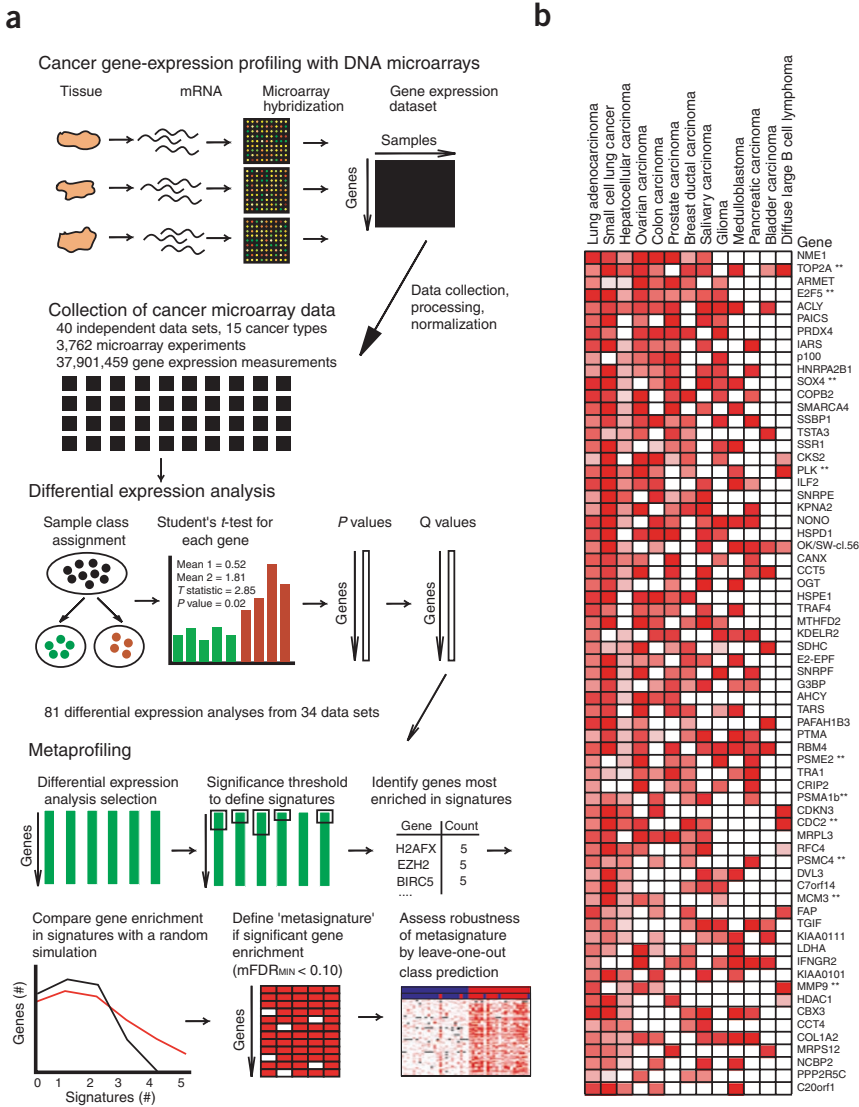


Figure 1 Meta-analysis methods for the integration of cancer signatures. By collecting, standardizing and analyzing several independent cancer signatures simultaneously, meta-analytic approaches can validate and define robust gene signatures and can also define gene signatures representing shared biology across multiple cancer types. **(a)** The meta-analysis method used to identify a 'universal' cancer signature is outlined. **(b)** The universal cancer signature encompasses 67 genes commonly activated in human cancer relative to respective normal tissue types. A red square indicates significant overexpression; the shade of red indicates the degree of statistical significance.

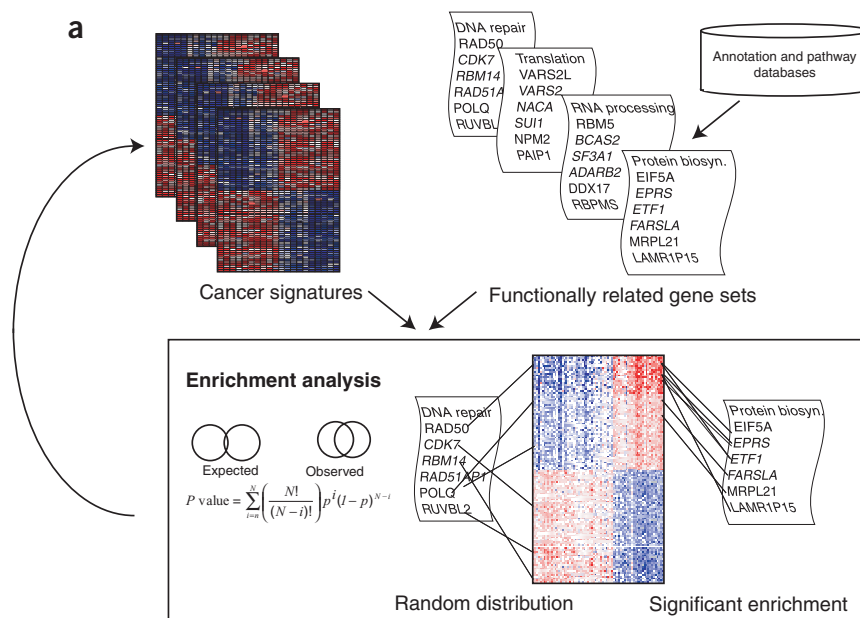
independent data sets, enrichment analysis for identifying coordinately regulated functional gene modules, protein interaction networks for detecting interaction complexes deregulated in cancer, transcriptional networks for inferring regulatory mechanisms in cancer and analyses of model system profiles with human tumor profiles for inferring activity of oncogenic pathways. In a related Perspective, Koller and colleagues discuss computational approaches for defining higher-level functional modules and regulatory models from transcriptome data¹². They detail model-building methods that have been successful in lower organisms and discuss the potential for their application to humans. As diverse fronts of genomic data continue to grow and mature, the need for bioinformatics integration will be imperative to understanding cellular networks and how they are deregulated in human cancer.

Meta-analysis of cancer signatures

With hundreds of cancer signatures published in the literature, several types of integrative analysis can be done. We call analysis of multiple data sets 'meta-analyses,' similar to meta-analyses done in clinical research, in which multiple studies interrogating a common hypothesis are analyzed together. In the realm of microarray data, meta-analysis is complicated by distinct experimental platforms and designs, and so gene-expression measurements are not always directly comparable. Several studies have applied meta-analysis methods to cancer microarray data, both to identify robust gene-expression signatures in a single cancer type and to look for common expression patterns across different types of cancer. We applied a summary statistic-based method coupled with false discovery rate analysis to compare four gene-expression studies that analyzed clinically localized prostate cancer relative to benign prostate tissue and identified a signature of genes commonly activated in prostate cancer across all data sets, irrespective of technological platform¹³. After defining a robust signature for prostate cancer, we carried out pathway analysis with the KEGG database¹⁴ and found that the polyamine biosynthesis pathway was hyperactivated, consistent with the known elevation of polyamines in prostate cancer. Another report identified gene expression-based classes of breast cancer in one data set and then searched for the existence of the classes in two independent data sets¹⁵. Three main biological subtypes held up in the two independent datasets: basal-like, ERBB2-overexpressing and luminal-like, reflecting diversity in cell of origin and transforming mechanisms in breast cancer. Other studies have identified prognostic signatures capable of predicting disease outcome across distinct data sets^{16,17}. As the application of genomic technologies becomes more commonplace, it will be important to define robust signatures that hold up across multiple independent data sets.

Perhaps more interesting from a biological standpoint are meta-analyses that look for common signatures across data sets and cancer types. In a demonstration of such an analysis, Ramaswamy and colleagues identified a gene-expression signature that was differentially expressed in metastatic tumors of diverse origins relative to primary cancers¹⁸. They found that the metastatic signature was also expressed in a subset of the primary tumors analyzed, leading to the hypothesis that the signature might represent a metastatic program that is encoded in primary tumors destined for metastasis. To test the generality of the signature, they attempted to predict time to metastasis in several independent solid tumor data sets including prostate, lung and breast adenocarcinomas, as well as medulloblastomas. In all of the data sets analyzed, the metastatic signature was significantly associated with clinical outcome and metastatic disease, suggesting that metastatic potential is encoded in

Figure 2 Analyzing cancer signatures in the context of functionally related gene sets can identify coordinately regulated functional modules. Various analytical methodologies have been applied to define functional modules or enriched gene sets. **(a)** A method that uses the binomial distribution to calculate the chance probability that a gene set would show a given degree of enrichment in a cancer signature. Gene set enrichment scores were computed for several types of gene sets (Gene Ontology, KEGG, Biocarta) across hundreds of cancer signatures from the Oncomine database. **(b)** Two functional modules enriched in a metastatic breast cancer signature are associated with primary breast tumors that metastasized relative to primary breast tumors that did not. The mitosis module and the Y branching of actin filaments module showed significant enrichment, suggesting that these processes are important in metastatic breast cancer.



primary tumors and is at least in part encoded by a common signature across tumor types. We also carried out meta-analysis in search of a common signature across distinct cancer types¹⁹. We collected 40 independent data sets (>3,700 array experiments) from which we generated 36 cancer signatures representing genes activated in a particular cancer type relative to the normal tissue type from which it arose. A meta-analysis of these signatures identified 67 genes activated in 12 or more signatures. This 67-gene signature could predict cancer versus normal status in most of the cancer signatures tested, as well as in nine independent cancer signatures, of which three represented cancer types not included in the original analysis. These results suggest that a common transcriptional program pervades most types of cancer (Fig. 1). Similarly, another meta-analysis uncovered a molecular signature that can distinguish different types of undifferentiated cancers from their well-differentiated counterparts, suggesting that there are universal pathways to dedifferentiation. This work highlights the ability of meta-analysis to detect collective molecular signatures that may be more generally related to tumorigenesis than are individual cancer type signatures.

Functional enrichment analysis of cancer signatures

Gene-expression profiling with DNA microarrays often generates large gene-expression signatures characteristic of a particular type or subtype of tissue. Sifting through hundreds or even thousands of differentially expressed genes is a daunting task for even the most schooled molecular biologist. Often, a handful of genes of interest are selected for experimental validation and work-up, and hundreds of others are ignored. In the previous section, we described meta-analysis methods for validating analogous gene-expression signatures and extracting common gene-expression signatures from independent cancer signatures. Here, we describe the use of external functional information for interpreting and summarizing large cancer signatures. Early demonstrations

of cancer molecular profiling used hierarchical clustering and manual annotation of gene clusters to identify subclasses of cases that expressed functionally related gene sets^{7,20}. Although the results obtained from such studies were convincing, the functional analyses were subjective and relied on the experimenters' knowledge of gene families, processes and pathways.

Several methods have been proposed for systematic functional analysis of gene-expression signatures^{21–26}. Most approaches use external annotation databases such as Gene Ontology²⁷, which is a database of controlled vocabulary gene annotations describing the biological processes, molecular functions and cellular localizations of genes. Because Gene Ontology encompasses thousands of functional annotations assigned to a large fraction of the genome (>14,000 unique genes), it serves as an ideal resource for enrichment analysis in cancer signatures. For example, a cancer signature may consist of 1,000 overexpressed genes out of 10,000 measured genes. A statistical test of enrichment might find that 100 of 150 genes annotated to function in 'protein

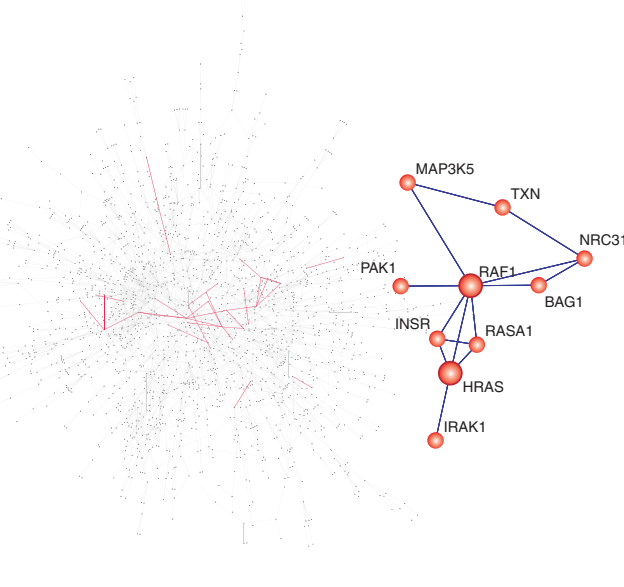
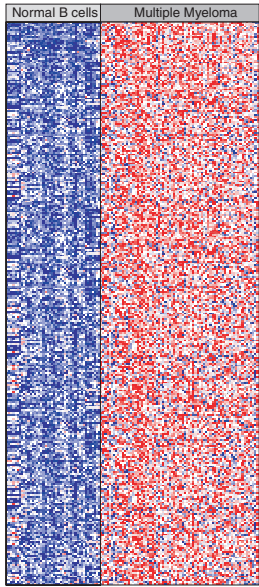


Figure 3 The human protein-protein interaction network as a framework for interpreting cancer gene-expression signatures. On the left is a signature of 300 genes significantly overexpressed in multiple myeloma. The interactions between pairs of proteins overexpressed in multiple myeloma are highlighted (in red) among the known protein interaction network according to the Human Protein Reference Database. A focused interaction subnetwork in which all members are overexpressed was extracted using OncoPrint. This subnetwork suggests that RAF1 overexpression coupled with multifaceted RAF1 activation by RAS, PAK1 and BAG1 may contribute to multiple myeloma tumorigenesis.

biosynthesis' are present in the cancer signature (Fig. 2a). By chance, we would expect only ~15 genes (10%) to be present in the cancer signature; therefore, this enrichment would be statistically significant, and we would conclude that the 'protein biosynthesis' gene-expression module is activated in the cancer signature. This general approach can be applied for all gene sets represented in Gene Ontology and other pathway resources such as KEGG¹⁴, Biocarta (<http://www.biocarta.com/>) and GenMAPP (<http://www.genmapp.org/>).

The power of systematic functional analysis of disease signatures was demonstrated by Mootha *et al.* By applying a method called gene set enrichment analysis, which uses a Kolmogorov-Smirnov test statistic, they found that genes involved in oxidative phosphorylation are coordinately downregulated in diabetic muscle²⁴. This result was particularly notable because in the original analysis, no individual genes showed significant deregulation in diabetic muscle. Only when examining functionally related gene sets were they able to uncover a coordinately deregulated process. Another study used gene set enrichment analysis to characterize three predominant gene expression-based subclasses of diffuse large B-cell lymphoma²⁸. They found enrichment of particular gene sets in each subset (oxidative phosphorylation, B-cell receptor/proliferation and host response), suggesting that distinct biological processes differentiate subclasses of diffuse large B-cell lymphoma.

A recent study took systematic functional analysis a step further by analyzing hundreds of gene sets in the context of a compendium of diverse cancer profiles²⁹. They defined 'modules' as functionally related sets of genes that are conditionally activated or repressed across a wide variety of cancer types. This analysis uncovered several modules deregulated in cancer. Some modules were shared across many cancer types, such as the cell cycle module, whereas others were specific to cancer types or subtypes. For example, a growth inhibitory module was repressed in leukemias, and a bone osteoblastic module was activated in breast tumors. The module map generated by this analysis suggests several hypotheses linking biological processes to specific cancer types and subtypes. Careful interpretation and validation of such modules will be required to fully appreciate the value of the approach.

As part of the OncoPrint database, we have also begun to interrogate cancer gene-expression signatures for the enrichment of particular Gene Ontology annotations, KEGG metabolic pathways, Interpro protein families and Biocarta signaling pathways. To date, 512 cancer

signatures computed from 112 independent data sets have been tested for the enrichment of these types of gene sets (D.R.R. *et al.*, unpublished data). Figure 2b highlights examples of enriched functional modules in a metastatic breast cancer signature. In a landmark study, van't Veer and colleagues identified a signature of genes capable of predicting distant metastasis and survival from primary tumor gene-expression profiles³⁰. To identify functional modules active in the primary tumors destined for metastasis, we recapitulated this signature in OncoPrint and carried out enrichment analysis. The most significantly enriched biological process was 'mitosis' (odds ratio = 7.02, $Q = 2.7 \times 10^{-6}$), whereas the most enriched Biocarta pathway was 'Y branching of actin filaments' (odds ratio = 8.14, $Q = 0.037$), suggesting that these modules are important for breast cancer metastasis. The activated 'mitosis' module is not surprising (the authors noted a preponderance of cell cycle genes in the signature), but the 'Y branching of actin filaments' module is less obvious and may point to specific transcriptional machinery that promotes metastatic cell spreading.

Protein interaction networks and cancer signatures

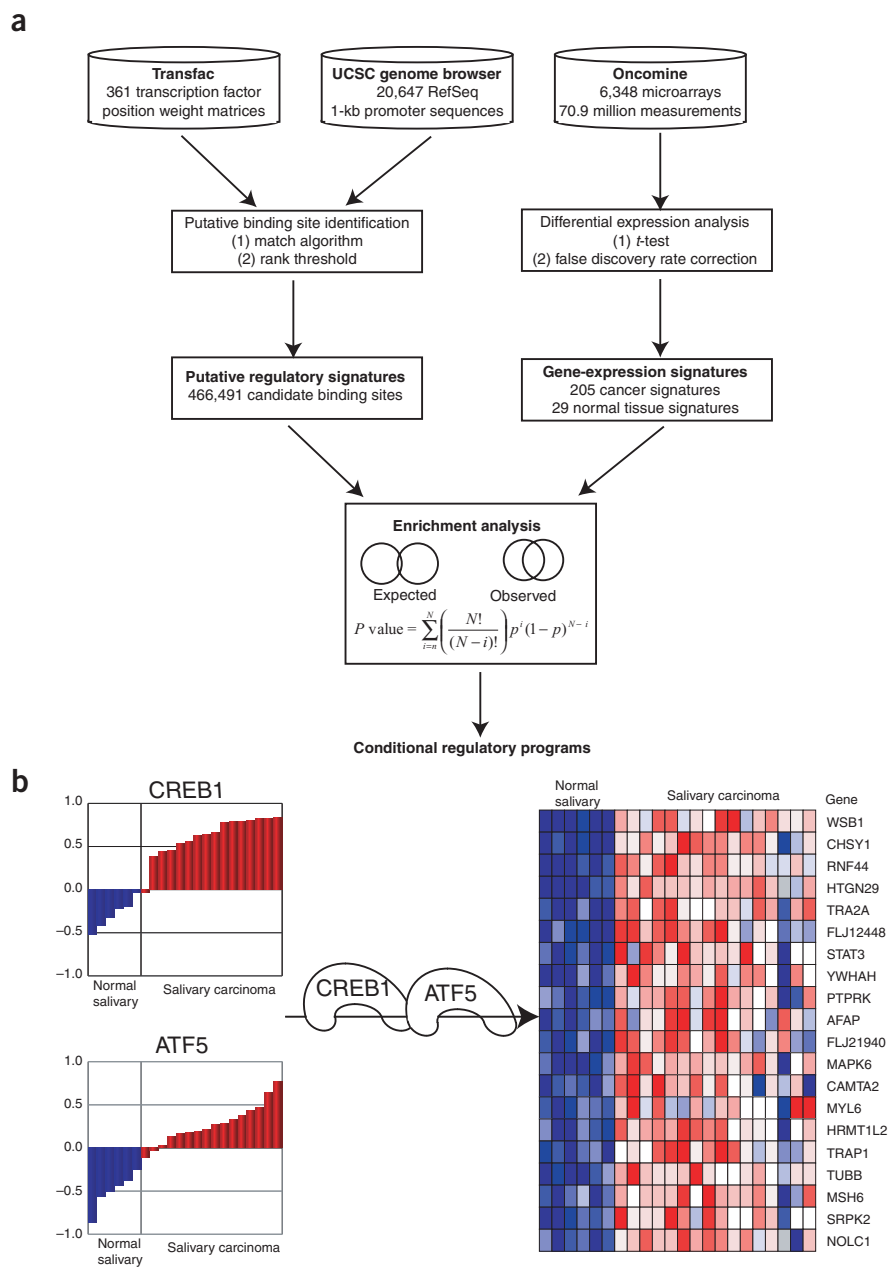
To understand complex biological processes, such as cancer initiation and progression, it is important to consider differential gene expression in the context of complex molecular networks. The study of such networks requires detailed protein-protein interaction ('interactome') maps. Preliminary versions of such maps have been generated by high-throughput methodologies such as global yeast two-hybrid screens, by computational prediction algorithms and, in the case of the Human Protein Reference Database (<http://www.hprd.org/>)³¹, by curating known interactions from the literature. A detailed human interactome network that captures the entire cellular network would be invaluable in interpreting cancer signatures, allowing one to infer activated sub-networks and specific proteins that are most important to a sub-network. Although human interactome maps are still in their infancy, representing only a fraction of the complete interaction network, we and others have begun to evaluate their utility in interpreting complex cancer signatures. Figure 3 shows how a multiple myeloma signature from OncoPrint can be mapped to the human interactome network, as defined by the Human Protein Reference Database, to define a sub-network of coordinate activation (D.R.R. *et al.*, unpublished data). Upon exploring the links in the subnetwork, we speculate that RAF1

Figure 4 Regulatory programs encoded in cancer gene-expression signatures. **(a)** An enrichment analysis analogous to the functional gene set enrichment analysis aimed at identifying specific transcription factor-binding sites that are common in cancer gene-expression signatures. **(b)** An example cancer regulatory program that was defined by CREB-ATF binding sites enriched among a salivary carcinoma gene-expression signature. CREB1 and ATF5 were specifically overexpressed, suggesting that their overexpression may activate a program of genes with CREB-ATF binding sites in salivary carcinoma.

may be the linchpin, as several members of the network (RAS, PAK1 and BAG1) function to activate RAF1. Thus, by targeting RAF1, as opposed to other members of the network, one may be able to blunt the effects of the entire subnetwork. Although our interpretation is speculative, it highlights the potential for interaction networks in the analysis of cancer signatures. We have developed a tool called HiMAP (Human Interactome Map; <http://www.himap.org/>) that allows one to input a gene-expression signature and identify activated or repressed interactome subnetworks on the basis of known and predicted protein-protein interactions. Other commercial products have similar functions, using their proprietary databases of literature-defined protein interactions and pathways. Companies with such products include Ingenuity (<http://www.ingenuity.com/>), GeneGo (<http://www.genego.com/>) and Ariadne Genomics (<http://www.ariadnegenomics.com/>). When considering the current state of human molecular interaction maps and databases, it is important to note the limitations and pitfalls. First, because these early representations of the cellular network are largely incomplete and probably contain several errors, important pathways and subnetworks may be missed by such analysis. For example, when considering networks generated from the literature, only well-studied proteins and interactions will be represented, and when considering networks generated from high-throughput yeast two-hybrid experiments, only nuclear and cytoplasmic proteins will be represented. Furthermore, protein-protein interactions are probably often context-dependent (*i.e.*, an interaction that occurs in a specific cell line or artificial system may not occur *in vivo*). Therefore, until interaction networks have matured, insights gained from their analysis must be treated as hypotheses requiring careful experimental validation.

Transcriptional targets and cancer signatures

The previous sections dealt with integrative bioinformatics approaches for characterizing and interpreting cancer gene-expression signatures. Because most cancer signatures are obtained at a specific time, providing only a snapshot of gene expression, it is difficult to understand and differentiate cause and effect from a gene-expression signature



(*i.e.*, does gene A activate gene B? Does gene B activate gene A? Or are both genes A and B activated by gene C?). Integrative approaches have been developed to begin to infer causality and define directional pathways activated in cancer signatures. In the next two sections, we examine approaches that leverage transcription factor-binding site data and approaches that combine model system oncogene signatures with *in vivo* human cancer signatures to infer oncogene activation and downstream targets. Similar to protein interactome networks, global transcriptional networks have the potential to improve the interpretation of cancer signatures. For example, if the targets of all transcription factors (and multicomponent transcription factor modules) were known, then one could easily infer which transcription factors must be activated in a tumor to yield the observed cancer signature. Reducing a complex cancer signature to a small number of activated transcriptional programs might shed light on neoplastic mechanisms and suggest points of therapeutic intervention. Several

approaches have been devised to identify consensus binding sites in the promoters of coexpressed genes, but this approach is difficult in higher eukaryotes, which often have complex regulatory structures. Here, we instead discuss approaches that use known binding sites for characterized transcription factors.

New technologies such as chromatin immunoprecipitation coupled with promoter microarrays (ChIP-chip) allow for genome-wide identification of transcription factor-binding sites^{32–34}, but only a few human transcription factors have been profiled to date. Sequence-based methods for defining putative transcription factor-binding sites are probably less accurate than experimentally defined binding sites, but hundreds of consensus binding sequences have been defined³⁵, making it feasible to carry out large-scale integrative analysis of binding-site profiles and cancer signatures. Elkon *et al.* demonstrated this approach first in humans, identifying known transcription factor-binding sites enriched among promoters of genes expressed in various phases of the cell cycle³⁶. We applied a similar approach to the Oncomine compendium of cancer signatures to identify regulatory programs activated in cancer (Fig. 4). In this example, 265 gene-expression signatures were compared with 361 sequenced-derived transcription factor-binding site profiles. We identified more than 300 cases in which a specific transcription factor profile was enriched in a cancer signature, suggesting that the transcription factor may be responsible for the observed gene expression³⁷. As a validation exercise, we tested several normal tissue signatures and identified enriched transcription-factor profiles that corresponded to transcription factors with known tissue specificity. Several proposed cancer regulatory programs suggest the importance of E2F, Myc, Ets-1, Rel and ATF in regulating cancer gene expression. **Figure 4b** details a candidate regulatory program that links the CREB-ATF binding sites to salivary carcinoma. We anticipate that as experimental transcriptional network data grow, this type of analysis will become more powerful, facilitating comprehensive deconvolution of the cancer transcriptome.

Model systems and cancer signatures

Several oncogenes are known to have causative roles in tumorigenesis³⁸. To begin to understand the mechanisms by which oncogenes cause cancer, studies have used gene-expression profiling to identify downstream targets of oncogenic pathways in cell-culture systems. Integrating results from such experiments with *in vivo* cancer signatures holds the potential both to infer activity of specific oncogenic pathways *in vivo* and to identify relevant effectors of oncogenic pathways. In a demonstration of the former application, Huang *et al.* developed distinct *in vitro* oncogenic signatures for Myc, Ras and E2F1-3 (ref. 39). These signatures were able to predict Myc and Ras state in mammary tumors that developed in transgenic mice expressing either Myc or Ras, suggesting that specific oncogenic events are encoded in global gene-expression profiles. In an effort to elucidate the direct transcriptional effects of oncogene Cyclin D1, Lamb *et al.* carried out an integrative analysis of an *in vitro*-defined Cyclin D1 gene-expression signature with a panel of human tumor expression profiles with variable expression of Cyclin D1 (ref. 40). They found that the *in vitro*-defined targets of Cyclin D1 were correlated with Cyclin D1 levels *in vivo*, suggesting that direct transcriptional regulation by Cyclin D1 may be important in tumorigenesis. They also went on to search for other genes associated with the Cyclin D1 signature *in vivo* and found that CEBP/B had a strong association with the signature; in fact, Cyclin D1 regulates transcription through CEBP/B.

In another report, Sweet-Cordero *et al.* defined a KRAS signature by comparing lung tumors generated from a spontaneous KRAS mutation mouse model to normal mouse lung⁴¹. After mapping the mouse

data to human lung tumor data, they found that the mouse signature shared significant similarity with human lung adenocarcinoma but not with other lung cancer types. Next, they looked for evidence of the KRAS signature in human tumors carrying activating KRAS mutations relative to wild-type tumors. Although no individual genes were significantly associated with KRAS mutation status in human tumors, the mouse KRAS signature was significantly enriched among genes rank-ordered by differential expression in human tumors with a KRAS mutation. The authors concluded that a KRAS mutation signature does exist in human tumors, but that the genetically driven mouse model signature was needed to act as a filter to extract it. In summary, integrating model system gene-expression profiling, in which a cancer gene can be overtly activated or repressed, with *in vivo* human cancer signatures holds great promise both for identifying cancers in which particular oncogenic pathways are active and for delineating important oncogenic effectors.

The future of integrative molecular analysis of cancer

Though far from comprehensive, this report attempts to focus on conceptual highlights in integrative cancer biology. The pace at which the cancer profiling field is growing, coupled with advancements in other high-throughput molecular approaches such as SNP arrays, array comparative genome hybridization, promoter arrays, proteomics and 'metabolomics,' ensures that integrative approaches will be crucial for extracting maximum biological insight from the collective cancer genomics data set. In order for such integration to occur, it will be important to define standards for communicating genomic profiles across diverse experimental systems. We propose that although the types of genomic data that are analyzed vary widely, a common language, quantitative gene lists, might unify disparate genomic data. When we consider the integrative analyses highlighted here, we find lists of genes activated in cancer; lists of genes that function in particular processes, pathways and interaction networks; lists of genes targeted by transcription factors; and lists of genes activated by oncogenes in model systems. Other types of gene lists not discussed include genes expressed in normal human cell types, genes expressed at various stages of development, genes located on specific chromosome arms or in specific bands, genes deregulated by drug treatment and genes conferring sensitivity or resistance to drug treatment. Analogous to how a new DNA sequence can be effortlessly compared to all DNA sequences in GenBank, we propose that the same such resource should exist for gene lists or 'molecular concepts.' A 'molecular concepts map' would compute the degree of overlap among all concepts and allow experimenters to import new concepts and study their relationships with the global concept map. Beginning with a concept of interest (*e.g.*, a new cancer signature), one might find that the signature shares similarity with an *in vitro* oncogene signature, a transcription factor-binding site profile and a drug-treatment profile, suggesting that the oncogene and transcription factor are active in the new cancer signature and that a specific drug treatment might reverse the signature. Such an effort would require reporting standards similar to those required for microarray data⁴², able repositories and journal requirements for deposition before publication. An initiative to collect and standardize all quantitative gene lists already published would be of great use and might serve as the backbone to a larger community-wide contributory effort. As evidenced by the Affymetrix publication list, the growing GEO and ArrayExpress microarray repositories and the Oncomine database of cancer signatures, the limitation is not in the quantity or quality of genomic profiles, but rather in the lack of integrative bioinformatics solutions for sharing and analyzing such profiles.

ACKNOWLEDGMENTS

We thank the OncoPrint team for their efforts on this project. This work is partially supported by grants from the American Cancer Society and the National Institutes of Health and by the University of Michigan Cancer Center Support Grant. D.R.R. is a Fellow of the Medical Scientist Training Program and is supported by the Cancer Biology Training Program. A.M.C. is a Pew Biomedical Scholar.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at <http://www.nature.com/naturegenetics/>

- 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).
- 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).
- Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488–492 (2005).
- Parkinson, H. *et al.* ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **33**, D553–D555 (2005).
- Barrett, T. *et al.* NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res.* **33**, D562–D566 (2005).
- Rhodes, D.R. *et al.* ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1–6 (2004).
- Perou, C.M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
- 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).
- Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
- Golub, T.R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
- Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
- Segal, E., Friedman, N., Kaminski, N., Regev, A. & Koller, D. From signatures to models: understanding cancer using microarrays. *Nat. Genet.* **37** Suppl., 38–45 (2005).
- Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. & Chinnaiyan, A.M. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.* **62**, 4427–4433 (2002).
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
- 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).
- Shen, R., Ghosh, D. & Chinnaiyan, A.M. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics* **5**, 94 (2004).
- Jiang, H. *et al.* Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* **5**, 81 (2004).
- Ramaswamy, S., Ross, K.N., Lander, E.S. & Golub, T.R. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* **33**, 49–54 (2003).
- 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).
- Alizadeh, A.A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
- Draghici, S. *et al.* Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.* **31**, 3775–3781 (2003).
- 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).
- Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**, 578–580 (2004).
- Mootha, V.K. *et al.* PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
- Volinia, S. *et al.* GOAL: automated Gene Ontology analysis of expression profiles. *Nucleic Acids Res.* **32**, W492–W499 (2004).
- Zeeberg, B.R. *et al.* GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**, R28 (2003).
- Harris, M.A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).
- Monti, S. *et al.* Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood* **105**, 1851–1861 (2005).
- Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* **36**, 1090–1098 (2004).
- van't Veer, L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
- Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371 (2003).
- Odom, D.T. *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378–1381 (2004).
- Li, Z. *et al.* A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl. Acad. Sci. USA* **100**, 8164–8169 (2003).
- Weinmann, A.S., Yan, P.S., Oberley, M.J., Huang, T.H. & Farnham, P.J. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.* **16**, 235–244 (2002).
- Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
- Elkon, R., Linhart, C., Sharan, R., Shamir, R. & Shiloh, Y. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.* **13**, 773–780 (2003).
- Rhodes, D.R. *et al.* Mining for regulatory programs in the cancer transcriptome. *Nat. Genet.* (in the press).
- Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
- Huang, E. *et al.* Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat. Genet.* **34**, 226–230 (2003).
- 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).
- Sweet-Cordero, A. *et al.* An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat. Genet.* **37**, 48–55 (2005).
- Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).