Computational prediction of cancer-gene function

Pingzhao Hu[*‡§], *Gary Bader*[*‡], *Dennis A. Wigle*[‖] *and Andrew Emili*[*‡]

Abstract | Most cancer genes remain functionally uncharacterized in the physiological context of disease development. High-throughput molecular profiling and interaction studies are increasingly being used to identify clusters of functionally linked gene products related to neoplastic cell processes. However, *in vivo* determination of cancer-gene function is laborious and inefficient, so accurately predicting cancer-gene function is a significant challenge for oncologists and computational biologists alike. How can modern computational and statistical methods be used to reliably deduce the function(s) of poorly characterized cancer genes from the newly available genomic and proteomic datasets? We explore plausible solutions to this important challenge.

Global
A large-scale or genome-wide biological perspective, often with reference to high-throughput experimental datasets.

*Program in Proteomics and Bioinformatics, Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada.
‡Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada.
§Department of Computer Science and Engineering, York University, Toronto, Ontario, Canada.
‖Division of Thoracic Surgery, Department of Surgery, and Department of Biochemistry and Molecular Biology, Mayo Clinic Cancer Center, Rochester, Minnesota, USA.
Correspondence to A.E.
e-mail:
andrew.emili@utoronto.ca
doi:10.1038/nrc2036
Published online
14 December 2006

Decades of molecular genetic research have shown that cancer is a heterogeneous cellular disorder caused by the deregulation of many interacting cellular pathways that converge to generate tumour formation and growth. Functional genomic and proteomic methods, such as global-expression profiling and molecular-interaction screens, are increasingly being used to infer the molecular associations among the many gene products causally implicated in cancer cell growth, survival, progression, metastatic invasiveness and/or therapeutic resistance[1–3]. Although the use of a specific gene(s) for cancer diagnostics does not depend on a causative role in the cancer process[4], the patterns detected from such large-scale genomic and proteomic datasets are already being used experimentally to stratify cancer subtypes and to predict patient survival and treatment response[5]. However, there remains a pressing need to derive even more clinically and biologically relevant information from these molecular associations, particularly with regard to determining the fundamental biological functions of previously uncharacterized genes newly implicated in neoplasia[6,7]. This improved knowledge of cancer-gene function will increase mechanistic understanding of the molecular basis of oncogenesis, and thereby facilitate the identification of new drug targets and the development of more effective molecular diagnostic and prognostic tests.

It has been postulated that a better way to systematically uncover gene function and the higher-level organization of proteins into biological pathways is through the analysis of molecular-interaction networks[8,9]. Although large-scale functional prediction based on this strategy is rapidly gaining popularity among computational biologists for investigating gene action in model organisms such as yeast, worms and flies[10–18], so far there have been only limited attempts to specifically infer the functions of known or candidate cancer genes by computational procedures based on association networks[19], and even fewer cases in which these predictions have been rigorously benchmarked and experimentally validated.

In this Review, we introduce state-of-the-art computational procedures that enable the automated prediction of cancer-gene function on the basis of analyses of the patterns of functional associations of known or predicted cancer-gene products in the context of interaction networks. We discuss the value and limitations of current algorithms and publicly accessible molecular datasets for accurate computational function prediction. Outstanding challenges that must be overcome to increase the relevance of such predictions to a clinical setting are also summarized, making reference to well-established cancer genes[20] to illustrate key concepts. Emphasis is placed on practical tools and resources that are readily accessible to, and useable by, cancer biologists.

Necessity of predicting cancer-gene function

It has been suggested that 5–10% or more of the ~25,000 putative genes encoded in the human genome probably contribute to oncogenesis[21]. But a recent exhaustive census based on an updated list provided by the Sanger Centre has compiled a list of only 354 experimentally validated genes that are causally implicated in neoplasia

At a glance

- Many cancer genes remain functionally uncharacterized. Experimental methods to characterize their functions are inefficient, time consuming and expensive.
- The increasing availability of diverse molecular profiles and functional-interaction data make the prediction of cancer-gene functions possible.
- New computational prediction methods now enable the automated assessment of cancer-gene function.
- The main difficulties are how to simultaneously integrate different high-throughput data sources and dependably assign multiple functions to a cancer gene.
- Trustworthy gene annotations are crucial to achieving the best possible functional predictions for newly discovered or uncharacterized cancer genes.
- Rigorous evaluation of the accuracy of functional predictions generated by computational methods is vital for formulating biologically relevant hypotheses to direct further rounds of experimentation.

development, or only roughly 1% of all predicted human genes[20]. These cancer genes have historically been identified in a step-wise manner by the positional cloning of individual familial susceptibility loci[22], the discovery of viral and mutated forms of cellular proto-oncogenes[23], or by the association of specific chromosome anomalies with gain- or loss-of-function alleles of select genes[24]. This contrasts with the actual burden of disease, where solid tumours of unknown genetic aetiology account for most cancer cases — for instance, lung, colon, breast, prostate and pancreatic tumours lead to ~55% of all cancer mortality in the United States based on statistics from the American Cancer Society[25], equating to 316,305 of 564,830 cancer deaths predicted for 2006.

The rapid accumulation of high-resolution cancer genetic and epigenetic molecular data now promises to enable far more comprehensive and unbiased inference of uncharacterized cancer genes linked to complex tumour traits such as metastasis and angiogenesis[4]. However, despite the demand to uncover the mechanisms that underlie cancer emergence and progression, knowledge of the functions of even the currently accepted collection of well-established cancer genes remains incomplete and skewed (as discussed below). To compound matters, most in-depth mechanistic experimental studies are still typically focused on those few gene products in which rare mutations or polymorphisms have been linked to susceptibility to familial forms of cancer, such as the germ-line inactivation of select tumour-suppressor genes[26]. This contrasts with the emerging sense that the biological processes that underlie stochastic cancer predisposition, initiation and progression are most commonly polygenic, and probably involve combinations of common alleles across many loci, most with weak effects, rather than a few rare alleles with large effects[27].

Given the laborious and expensive nature of traditional experimental approaches, sophisticated computational procedures for systematically predicting the functional roles and relationships of uncharacterized cancer-gene products are increasingly seen as useful for focusing the necessary biological validation[28]. The most convenient and well-known computational method for function prediction is based on the detection of significant sequence similarity to gene products of known function, using such basic bioinformatic software tools as BLAST (basic local alignment search tool)[29]. The assumption is that proteins that are similar in sequence probably have similar biological properties. An important caveat with this simplistic approach is that only those functions tied directly to sequence, such as enzymatic activity, can generally be predicted accurately.

An interesting alternate study[28], typical of the increasing use of more sophisticated computational procedures to deduce functional relatedness, used a hierarchical clustering method to group mRNA expression patterns derived from the microarray-based profiling of different cancers to detect cancer-specific expression-based functional modules, or gene sets that are seemingly involved in one or more related biological processes because they are typically co-expressed in many tumour samples. The resulting groupings were found to be enriched not only for the 'usual suspects', such as genes linked to control of the cell cycle, DNA repair or transcription, but other functional categories like inflammation, immunity and the extracellular matrix, which are now increasingly recognized as important determinants of tumour progression[30]. In essence, the presence of genes with well-defined biological properties in these modules can be used to make reasonable guesses as to the roles of co-clustered genes of unknown function — a concept formalized in computational prediction procedures.

Cancer-gene functional annotation

Importantly, functions must be clearly defined to set the stage for computational prediction[13]. Cancer genes are often thought to be oncogenes, tumour suppressors, stability factors or cancer-progression genes[4]. But biological function has many facets, reflecting the diversity of cellular activities and biochemical properties, ranging from the basic attributes of a protein product (such as an enzyme, like a protein kinase), to the nature of physical and regulatory interactions (such as protein–protein interactions), to membership in a given pathway (such as a signalling cascade). Explicitly defining these functions in a concise manner is often difficult, particularly in the cancer setting, as the labels must reflect the complex networks of gene products that interact dynamically across a wide range of spatial and temporal scales, from subcellular compartments to entire tissues or a whole organism. Perturbation of these pathways, processes or networks, together with the genomic instability typically seen in cancer, adds an additional layer of complexity to functional definitions by changing the natural context of operation for cancer-gene products and even the gene products themselves, such as in chromosomal translocations that create the aberrant gene fusions involved in many haematological malignancies[31].

Many annotation schemas for the representation of cancer-gene-product function have been devised, of which several prominent examples are listed in

Interaction network
A graphical description of a large ensemble of molecular associations, the nodes of which correspond to gene products, and the edges of which reflect direct links or connections between the gene products.

Hierarchical clustering
A statistical method for finding relatively homogeneous clusters of gene products based on some measure of similarity.

Functional module
A set of gene products that together function in a single process.

Table 1 | **Publicly available data sources to examine cancer-gene-product function**

Name	Data type	URL	Source or reference
Gene Ontology (GO)	Gold standard	http://www.geneontology.org	32
MIPS (Munich Information Center for Protein Sequence)	Gold standard and interaction	http://mips.gsf.de	77
Gene Map Annotator and Pathway Profiler (GenMAPP)	Gold standard and interaction	http://www.genmapp.org	78
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Gold standard and interaction	http://www.genome.jp/kegg	79
BioCarta	Gold standard and interaction	http://www.biocarta.com	N/A
Cancer Cell Map	Gold standard and interaction	http://cancer.cellmap.org/cellmap	N/A
Module Map	Gold standard	http://ai.stanford.edu/~erans/cancer	28
SwissProt	Gold standard and interaction	http://www.expasy.org/uniprot/	67
Biomolecular Interaction Network Database (BIND)	Interaction	http://www.bind.ca/Action	80
IntAct	Interaction	http://www.ebi.ac.uk/intact/site/	81
Human Protein Reference Database (HPRD)	Interaction	http://www.hprd.org/	82
Database of Interacting Proteins (DIP)	Interaction	http://dip.doe-mbi.ucla.edu/	83
Online Predicted Human Interaction Database (OPHID)	Interaction	http://ophid.utoronto.ca/ophid	88
Molecular Interaction Network Database (MINT)	Interaction	http://mint.bio.uniroma2.it/mint/	84
Protein–protein interactions (PPI) of cancer proteins	Interaction	http://bmm.cancerresearchuk.org/~pip/bioinformatics/	47
Cancer Gene Census	Cancer genes	http://www.sanger.ac.uk/genetics/CGP/Census	Sanger Institute
Cancer Gene Data Curation Project	Cancer genes	http://ncicb.nci.nih.gov/NCICB/projects/cgdcp	US National Cancer Institute
Cancer Genes Resequencing Resource	Cancer genes	http://cbio.mskcc.org/cancergenes	Memorial Sloan-Kettering Cancer Center
The Tumor Gene Family Databases	Cancer genes	http://condor.bcm.tmc.edu/ermb/tgdb/tgdf.html	Baylor College of Medicine
Oncomine	Cancer profiling	http://www.oncomine.org/main/index.jsp	University of Michigan
Cancer Program Data Sets	Cancer profiling	http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi	Broad Institute
Stanford Microarray Database (SMD)	Cancer profiling	http://genome-www5.stanford.edu	Stanford University
Gene Expression Omnibus (GEO)	Cancer profiling	http://www.ncbi.nlm.nih.gov/geo	National Center for Biotechnology Information, US National Institutes of Health
Cancer Gene Expression Database (CGED)	Cancer profiling	http://cged.hgc.jp/cgi-bin/input.cgi	Osaka University School of Medicine

N/A, not applicable

Directed acyclic graph
A network data structure used to represent a gene-function classification system in the Gene Ontology database, having ordered relationships between nodes (for example, parent and child terms, wherein the graph direction indicates which term is subsumed by the other), and no cycles (no path returns to the same node twice). Nested terms can have several parents.

TABLE 1. The most widely adopted system is the Gene Ontology (GO) database[32], which uses a clearly defined and computationally friendly vocabulary for representing the cellular, biochemical and physiological roles of gene products in a systematic manner. From the perspective of functional computation, GO provides a standardized way to assess whether a set of genes have similar functions, which has led to its increasing popularity for the many function-prediction procedures used in model organism settings[15,33,34]. GO terms are organized in a tree-like structure, starting from more general (for example, cellular metabolism) at the root to the most specific at the leaves (for example, the regulation of DNA recombination) distributed across three main semantic domains — molecular function, biological process and cellular location. As GO terms might have more than one parent, they are technically structured as a network called a directed acyclic graph (FIG. 1). For instance, 'B-cell apoptosis' represents a sub-type of both the term 'apoptosis' and 'B-cell homeostasis'. Therefore, the functional classes are not necessarily independent of one another, and the dependencies are explicitly defined. Additionally, GO enables a single cancer-gene product to be associated with more than one functional term

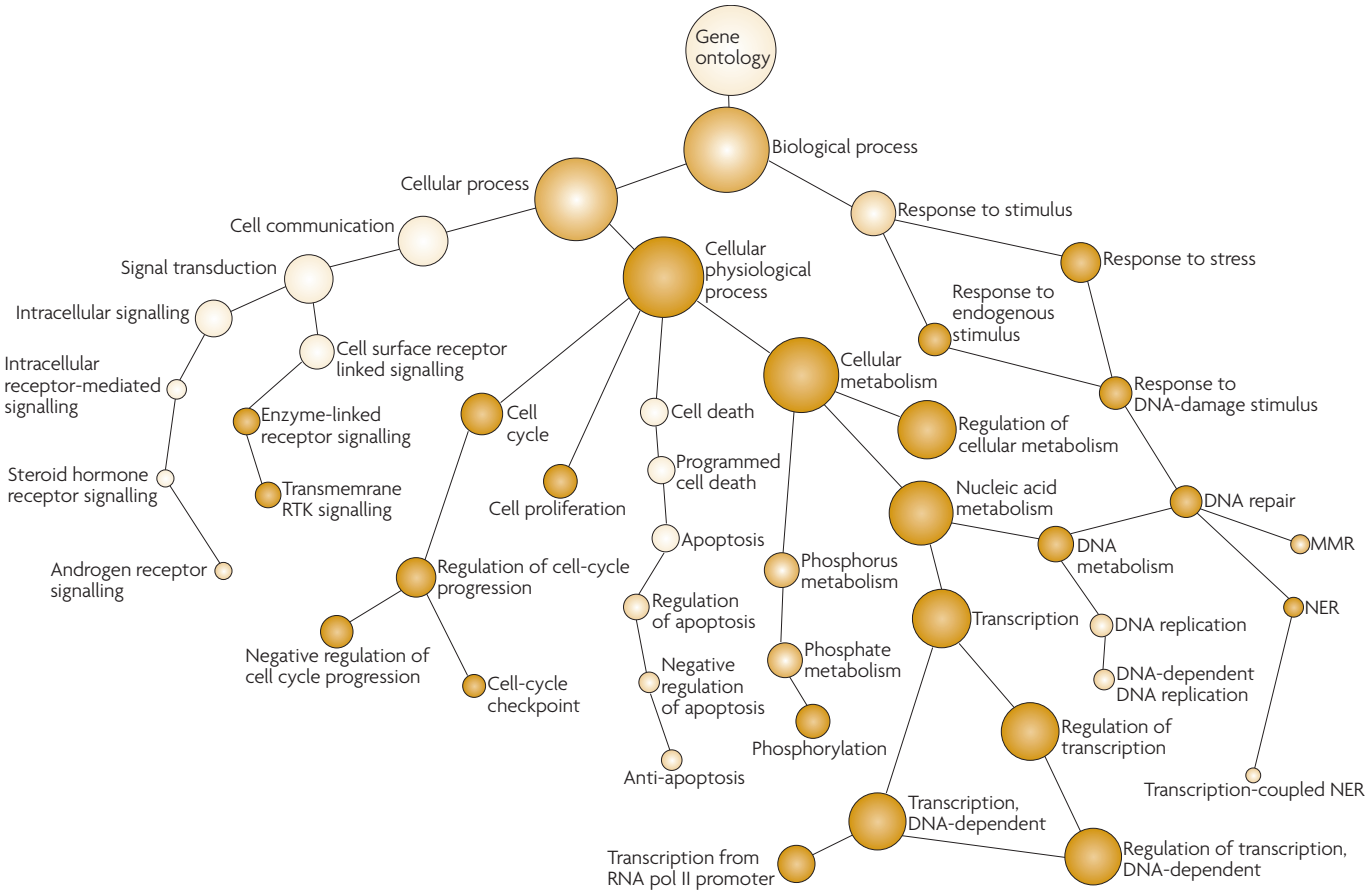

Figure 1 | **Cancer gene annotations.** A graphical representation of a nested Gene Ontology (GO) classification showing the skewed functional annotations (GO terms) assigned to known cancer genes derived largely from familial syndromes or single-gene defect cancers[20]. The size of the terms (circles) is proportional to total gene membership, and the colour shading indicates the degree of statistical significance (darker tones denote decreasing *P* values). The GO provides a computationally accessible, organism-independent means for examining and reporting gene function[15,33,34]. Although expert curators often manually assign terms on the basis of published experimental evidence, most terms are electronically inferred on the basis of sequence similarity to other well-studied gene products or other criteria[32]. Each annotation is assigned an evidence code (from the Gene Ontology website; not shown in the diagram) stating how the annotation is supported, which enables one to assess the reliability of an annotation. If the annotation is based on experimental evidence traceable to an author or publication, it is presumably more reliable than if it was simply inferred through sequence similarity. The GO has over 10 such evidence codes, which are not part of the core ontology. The figure was created using Bingo[89]. NER, nucleotide excision repair; MMR, mismatch repair; RTK, receptor tyrosine kinase.

during the curation process, rather than being restricted to a single functional class. However, as discussed below, this flexibility for allowing multiple-label classifications has not yet been fully exploited by computational prediction methods.

As with the other annotation schemas, the GO can be used to describe many, but not all, of the specific biological properties of known cancer genes. For example, the Wilms tumour protein, which results in renal cancer when mutated, is listed in GO as a DNA-dependent regulator of transcription, located in the nucleus, and having transcription factor activity. This is a relatively informative description of its molecular role as a cancer gene. By contrast, the human homologue of patched (*ptc*), which is mutated in basal cell carcinomas of the skin and medulloblastomas[35], is associated with less informative GO classifiers, such as dorsal-ventral pattern formation, embryonic limb morphogenesis and the regulation of smoothened (*smo*) activity, amongst others. Although these identifiers are correct, and do indicate relevant parts of the function of the gene, it is more difficult to infer the relationship of this gene to cancer from these terms. The question, then, is how exactly one links an uncharacterized gene to a cancer-specific process.

Analysis of the currently available GO annotations indicates the often ambiguous assignments that are made in the cancer setting. For example, although roughly three-quarters of well-established cancer genes[20] have at least one or more GO biological process terms and GO molecular function terms (TABLE 2), most of these genes are only sparsely annotated with simple functional annotations, such as 'cell proliferation', in the top, most generic levels of the GO classification system. Most of the GO terms

Table 2 | **Current Gene Ontology (GO) annotations for cancer genes**

Function	Level in the GO classification system							Unannotated at any level
	3	4	5	6	7	8	9	
Biological process	270* (39)	267 (94)	263 (136)	253 (172)	239 (171)	209 (141)	84 (88)	84
Molecular function	277 (43)	254 (79)	219 (93)	156 (67)	80 (38)	41 (18)	27 (15)	77

Summary of the distribution of 354 known cancer genes and their putative functions (number of GO terms listed in brackets) on the basis of current annotations reported using the nine-level GO classification system. The functions are defined from the most general (lower levels) to the more specific (higher levels). All gene products assigned to lower levels are also subsumed in the higher level terms. For example, the 84 cancer genes annotated with GO terms at the highest resolution (ninth level of annotation) are also counted in levels one to eight. The data for levels one to two are not shown, as these terms are overly general. We also did not include the 'cellular component' branch of the GO schema, but rather only report 'biological process' and 'molecular function', as our focus is on cancer-gene-function prediction and not gene-product localization. Cancer-gene information was obtained from REF. 20. *For example, 270 out of the 354 putative cancer genes were assigned to 39 functional GO terms in level 3 of the GO classification. This analysis is based on Babelomics[90].

have been assigned with a highly skewed distribution, reflecting a marked enrichment to a few select functional categories (FIG. 1). These data indicate the need for more efficient, comprehensive and informative methods for the computational prediction of cancer-gene function, and hint at the challenges ahead (discussed below).

Functional associations

The typical workflow for computationally assigning functions to cancer-gene products in an automated fashion starts with the transfer of functional annotations from established genes to uncharacterized cancer loci on the basis of previous knowledge of different types of functional associations. These range from sequence similarity, to the co-occurrence of the protein products in the same macromolecular complex, to similarity in mRNA or protein-expression patterns[36]. The process is shown schematically in FIG. 2.

Importantly, these functional relationships or gene-product interactions must be used appropriately to avoid erroneous inferences. For example, if the protein sequences of two genes encode protein kinase domains, they are likely to share the same molecular function (that is, they are both kinases), although they cannot then be assumed to participate in the same biological process or pathway. Conversely, if two proteins interact physically, there is no guarantee that the corresponding genes share the same molecular function, even though it is possible they participate together in the same physiological process.

The modular structure of most proteins, especially those that are involved in signalling, and their dynamic roles in many biological processes must also be taken into account[37]. For example, a protein with a kinase domain will immediately be labelled a kinase, but if it also has a SRC homology 3 (SH3) domain, it is likely to be involved in protein binding as an adaptor. Similarly, although the protein products of oncogenes are known to physically interact with dozens of other proteins, only some of these interactions might plausibly occur at any given time in a particular cell type.

Viewed collectively, the functional interactions among the molecular components of a cancer cell can be represented as a network of interacting component gene products. In the schematic representation of an interaction network shown in FIG. 3, nodes or points correspond to genes or proteins, and the edges or lines define functional links between the gene products. Such cancer maps often reflect protein–protein interactions measured by high-throughput experimental platforms[38–41], but can also be used to represent a rich source of other functional associations, such as correlated expression patterns deduced from genome-scale mRNA profiles generated for cancer cell lines and tumours contained in the Gene Expression Omnibus (GEO) and Oncomine databases, shared protein–DNA binding patterns[42] and other shared phenotypes[43]. Large-scale proteomic studies[44] are also increasingly informative about tissue and subcellular specificity.

The integration of these association networks can be helpful for examining the consequences of changes in mRNA levels or in transcriptional regulation. Human co-expression profiles and molecular-interaction networks can also be extended on the basis of the extensive evolutionary conservation of orthologues across more experimentally tractable model species, such as mouse, fly, worm or yeast[45]. Although there are limits to the relevance of model organisms to cancer biology, evolutionary conservation indicates a strong selective advantage, suggesting that these alternate data sources can potentially have analogous conserved functional relationships in human tumour cells[46]. Such an approach was used recently to derive a large probabilistic network involving over 100,000 putative functional interactions, including over 500 protein isoforms encoded by 346 cancer genes, each with evidence-weighted edges[47].

Many of these and other interactions that involve established cancer-gene products collected from the literature are summarized in public-domain databases[48] (TABLE 1). These databases can be readily downloaded and used to generate visually informative exploratory pathway maps (FIG. 3) or, as described below, used for more principled pattern discovery as applied to cancer-gene functional predictions[19]. Surprisingly, however, a recent survey of the functional associations of known cancer genes[20] in several of these databases showed that only 28% currently have extensive functional associations in the Kyoto Encyclopedia of Genes and Genomes (KEGG), while only 59%, 58%, 48% and 26% are listed in the IntAct, Biomolecular Interaction Network Database (BIND), Molecular Interaction Network (MINT), and Database of Interacting Proteins (DIP), respectively. This indicates that roughly half of all established cancer genes still lack functional-association information in the main public functional-association databases.

Gene-function prediction methods

Molecular associations defined by high-throughput experimental platforms serve as the starting point for predictive discovery using computational procedures. Integration of these association networks can also be helpful for elucidating the often complex relationships

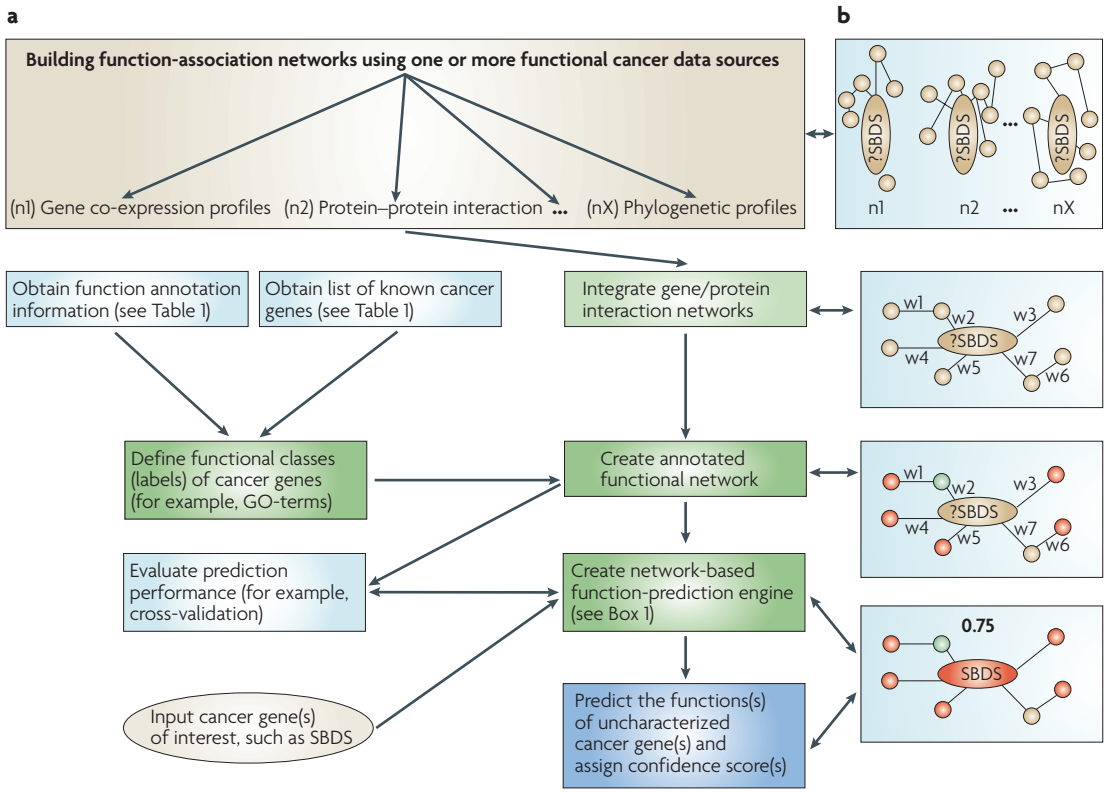

Figure 2 | **Schematic diagram of key steps for automated cancer-gene functional prediction. a** | Cancer-related genomic data sources (for example, protein–protein interactions, gene co-expression, and so on) that correspond to distinct functional-association networks (labelled n1, n2 … nX) are obtained. **b** | These can be integrated into a single network, with the corresponding evidence weights (w1, w2 … wX) defined. Cancer genes with known functions, as indicated by different node colours (for example, red and green), are assigned all relevant public annotations, such as Gene Ontology (GO)-terms or pathway information. An uncharacterized cancer gene of interest (for example, Shwachman–Bodian–Diamond syndrome (*SBDS*)) is then processed and functionally interpreted using an automated function-prediction engine that exploits the functions of its interaction partners or neighbourhood. Overall prediction performance is determined by cross-validation and a confidence score assigned (shown in bold).

between crucial components within core biological processes. Although network-based inferences are ultimately influenced by the reliability of the input functional associations, the predictive power of such network-wide inference procedures can be improved by assigning a confidence score or weight to each of the associations or edges of the interaction network according to the reliability of the information obtained from each of the data sources or data types[17,49].

There are two types of automatic function-prediction paradigms. One directly associates gene products with functional classes on the basis of pattern recognition, which is often accomplished with supervised learning[10,11,15,18] or unsupervised learning methods[43]. Here we focus on another automatic function-prediction paradigm, wherein cancer-gene functions are predicted by analysing the entire set of functional associations recorded between human-gene products in the context of a network (as in FIG. 2). The idea is to use the set of associations in the network to propagate the functional classes from well-characterized protein nodes onto those with limited or no annotation, such as newly discovered candidate oncoproteins of unknown function. Many functional-prediction studies that follow this paradigm are often focused first on sub-grouping or clustering the interaction networks into functional modules[14,34,50–53] on the basis of the pattern or distribution of protein nodes and interaction links, which can be highly suggestive of shared functions[9,54,55]. These modules might be distinct or overlapping. Any unannotated gene products in a given module can be subsequently assigned the most common functional annotation(s) associated with its interacting partners or neighbours. This 'unsupervised' approach often works well if there is extensive coherent annotation available and relatively few uncharacterized proteins per cluster, but there can be difficulty if a module contains many proteins without annotations or with diverse, seemingly unrelated functions.

Alternate computational methods (BOX 1) have been devised to automatically assign functional labels, such as GO terms, to the uncharacterized proteins present in an interaction network in a 'supervised' manner according to the annotations of the broader neighbourhood of interacting gene products. Differentiating from the module-based methods cited above, these newer approaches often exploit both the global and

Supervised learning
A computational procedure to identify sets of gene products that are similar to a reference set of manually-defined examples using a principled-prediction rule or criteria. Any genes of unknown function that are grouped with the set of pre-defined genes are deemed similar in function.

Unsupervised learning
A computational procedure to identify subsets of gene products that are more similar to each other than to others. The function of unknown genes can then be predicted based on the functions of other known genes within a given cluster.

Functional label
The function terms, such as Gene Ontology terms, that are assigned to cancer genes.

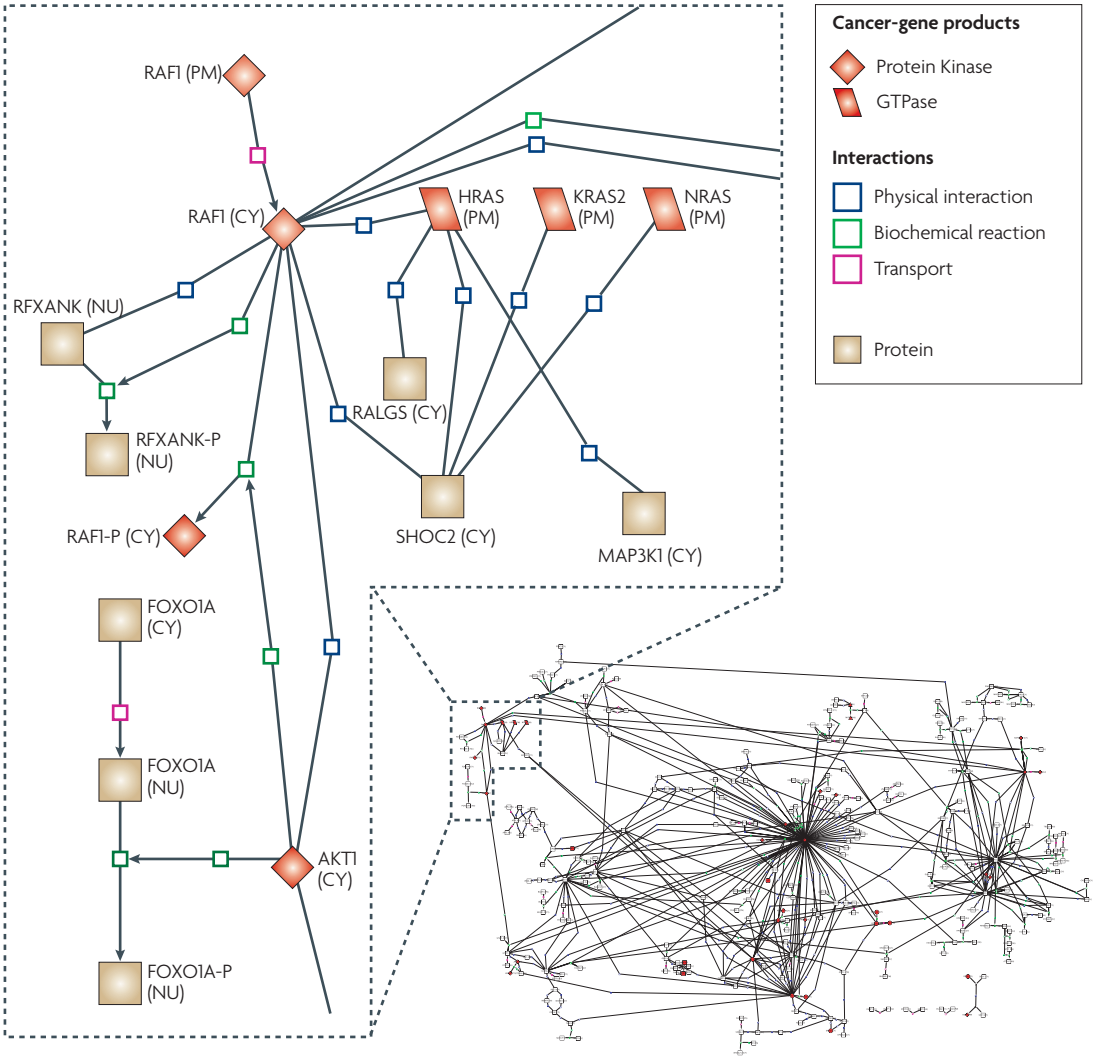

Figure 3 | **Cancer interaction networks.** A graphical representation of an association network of diverse experimentally derived molecular interactions that involve a subset of cancer-related human proteins. The zoom-in (dashed box) shows three different types of functional associations, where known. The network was visualized using Cytoscape; the data are from the Cancer Cell Map web site, using annotations from the Gene Ontology website. CY, cytoplasm; NU, nucleus; PM, plasma membrane.

local properties of network graphs. For example, the algorithm Function Flow[17] can predict functional classes for uncharacterized proteins that not only have direct functional associations (that is, immediate neighbours with annotations) but also indirect links (through the functional labels that are available for indirect partners that, in turn, are associated with any immediate neighbours that lack annotation). The trade off is that additional error or uncertainty can be introduced by assuming functional similarity among more loosely connected gene products that are more than one step apart in an interaction network.

In principle, computational methods potentially enable any type of molecular association to be examined. The inclusion of many categories of functional interactions increases the probability of genuine functional assignments[56]. Several publicly accessible bioinformatics tools that can be used to examine and integrate various genomic and proteomic data sources in order to generate an integrated functional association network are listed in TABLE 3. However, biological interpretation of such network-based functional inferences is often highly sensitive to the quality of the input association data, the clustering procedure used to derive the modules[2,3,57] and, as seen below, the validation procedure used to evaluate the quality of the predictions.

Functional-association network
An interaction network in which gene products are linked if they have experimentally measured or predicted functional associations.

How does one evaluate the predictions?

Effective computational inference relies on the availability of relevant, reliable and verifiable (that is, traceable) functional annotations and molecular associations derived from the literature. False predictions arise when physiologically irrelevant associations are made between functionally unrelated genes. Usually these stem from artefacts in the input datasets, which can arise even after extensive pre-filtering to increase data quality. Therefore,

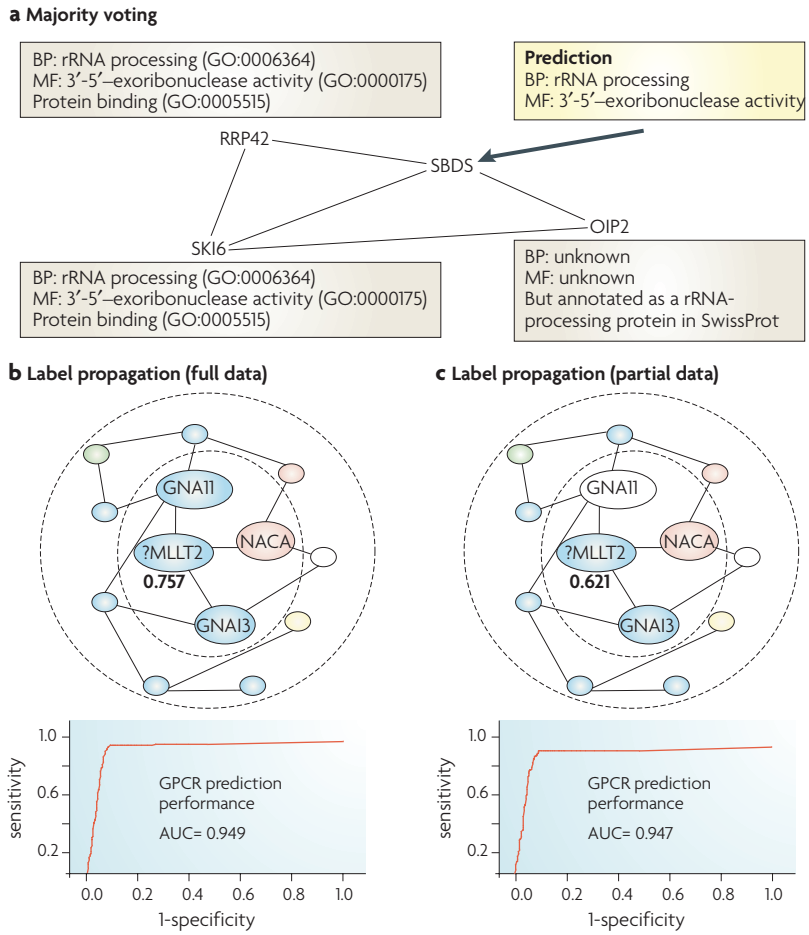

Figure 4 | **Example interaction networks and functional predictions for uncharacterized cancer genes. a** | Shows a network of functional associations with medium–high confidence (≥40%) obtained from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database for a small cluster of gene products associated with the functionally uncharacterized cancer-gene product SBDS. The biological process (BP) and molecular function (MF) of RRP42 and SKI6 have been defined using Gene Ontology (GO) terms, whereas the annotation for OIP2 was obtained from the UniProt Knowledgebase. The logical computational inference (majority voting) is that SBDS is probably a ribosomal RNA (rRNA)-processing factor (BP) with 3′-5′-exoribonuclease activity (MF). Predictive performance was not evaluated as the network is too small for class validation. **b** and **c** | Shows two closely related sub-networks of functional associations centred on the cancer gene *MLLT2*, one based on real data, on the left, and the other one slightly modified to assume that GNA11 has no GO functional annotations (see text for details). Different GO functions are indicated with distinct node colours (transcription (pink), G-protein coupled receptor (GPCR; blue), unknown (white) and other functions (green and yellow). The prediction confidence scores for the GPCR function assigned to MLLT2 in each analysis is shown in bold. The performance of the automated functional predictions of GPCR activity by label propagation for the two networks, as assessed by 10-fold cross-validation, is also shown in the accompanying receiver operating characteristic (ROC) plots. AUC, area under the curve.

Gold standard
A reference gene set used for labelling learning data, both for building prediction models and for creating test data to evaluate classifier performance.

one needs to carefully benchmark the predictions and the datasets on which they were based.

One successful evaluation strategy is to rely on external, independent data sets to function as reference gold standards for assessing both the primary data quality and the effectiveness or performance of the computational procedure used[58,59]. Although many different annotation resources are suitable for use as gold standards, the GO and KEGG stand out because of their extensive systematic curation of gene products. Although the GO and KEGG curators pursue meticulous categorization, the annotation process is often incomplete and, for diverse reasons, potentially biased and error-prone[59,60]. For example, although many cancer genes are pleiotropic, it is probable that not all of their discrete functions are equally well-defined, experimentally recognized or even cancer-relevant, resulting in incomplete or misleading labels for functional prediction.

Although potentially more accurate information regarding biological relationships might be available, such as high-resolution protein structures (for example, from X-ray crystallography), computational approaches for predicting cancer-gene function can still produce errors. The suitability of a gold-standard gene set for stringently assessing functional predictions is determined by three factors: first, the projected coverage or overlap relative to the target gene set under consideration, which should ideally be large enough to offer a statistically reliable evaluation using the available interaction data; second, the informativeness of the reference sets, which should have few irrelevant examples in a gold-standard reference set; and third, the resolution or specificity of the annotations, which must be consistent with the functional categories or resolution that one wishes to draw. For example, the fact that two proteins co-localize to the inner mitochondrial membrane supports the possibility of a functional interaction, but the association is too general to draw a detailed functional inference.

Different methods have been developed to evaluate the performance of computational predictors[58,59,61]. The standard procedure is cross-validation. A common concern with cross-validation is the dreaded over-fitting problem, which can be mitigated by careful study design[62]. Receiver operating characteristic (ROC) curves are usually drawn by plotting sensitivity versus specificity, or precision (or positive predictive value) versus recall, to evaluate the performance of computational methods in the cross-validation procedure. Such values are computed and plotted over a range of thresholds of discriminant values. Each threshold gives one pair of sensitivity and specificity values and, therefore, one point on the curve. Such analyses can be summarized with a single statistic — the area under the curve (AUC) — which provides a quantitative indication of how well a particular functional classifier performs.

Function prediction in practice

To show the potential informativeness of computational inferences, we applied the computational methods shown in BOX 1 to assign tentative functions to two uncharacterized cancer genes: Shwachman–Bodian–Diamond syndrome (*SBDS*) and *MLLT2* (also known as *AFF1*).

The *SBDS* gene is mutated in one of the most common forms of Shwachman–Diamond syndrome[63]. Patients with Shwachman–Diamond syndrome have exocrine pancreatic insufficiency, poor food absorption (malabsorption), low white blood cell counts (neutropaenia) and frequently develop leukaemia. This gene is

Box 1 | **Network-based functional prediction methods**

Several of the network-based computational methods suitable for the automated computational prediction of cancer-gene function that have been introduced include:

Majority voting[55]
The annotated functions of all direct neighbours (interacting partners) of a given gene or protein in a network are ordered in a list, from the most to least frequent. The function of an associated uncharacterized gene product(s) is then predicted to be the top *k* (a value defined by the user) or fewer functions in this list. This method is simple and fast, but takes only limited advantage of the overall network topology or any relationship among annotations.

Markov random fields[61,75]
Probability analysis methods that integrate many functional-association networks and compute a single probability value that a given gene product has a certain function given the functions of all other interacting proteins in the different networks. Although the approach provides a unified framework, current versions cannot solve the pleiotropy problem directly.

Label propagation[17,71]
Evaluates every node or gene product in an association or interaction network, not just the adjoining nearest direct neighbours, as a relevant source of function labels for functional annotation. There are different variants of this approach. One is to propagate or 'flow' a function through all the paths that connect the gene products[17]. A score is assigned that corresponds to the amount of flow to each node for a given function; another is based on Gaussian random field theory[71], a type of nearest-neighbour approach, in which the nearest labelled genes are computed in terms of a random walk on the network. The advantage of this method is that it takes into account both the global and local topology of the network; however, current incarnations of this approach are unable to assign multiple functions to a cancer gene simultaneously.

GenMultCut[73,76]
Investigates the global topological structure of an interaction network and assigns functions to proteins by minimizing the number of times different annotations are associated with neighbouring proteins. Although this approach can integrate many networks, it does not always use the local proximity of interacting gene products in the network efficiently for functional prediction, and is limited to assigning one function (one label) to a node at a time.

detectably expressed in nearly all tissues, but its molecular and biological functions are currently unknown[64]. The SBDS protein lacks homology to functionally characterized proteins[65], and there are no annotations listed in either the GO, KEGG or the Cancer Cell Map databases. Moreover, there are no reported interactions of SBDS with any other protein in an extensive human protein–protein interaction network constructed by computational methods[47]. Therefore, at first glance, this seems an intractable problem for automated functional prediction.

Yet one can infer some reasonable functional associations for SBDS using publicly accessible automated computational tools (TABLE 3). For instance, the Search Tool for the Retrieval of Interacting Genes/Proteins[66] (STRING) database predicts a conserved genomic association of *SBDS* with three other genes (*SKI6*, *RRP42* and *OIP2*) on the basis of the similarity of their respective genomic contexts (FIG. 4a). These associations can be assigned a confidence score, which in this case corresponds to the probability of finding the linked gene products within the same biological pathway. This confidence score is derived by benchmarking the performance of the STRING prediction engine against a common reference set of trusted functional associations, in this case as defined by the KEGG database.

Cross-validation
A statistical method for evaluating a classifier model. The input-association data is randomly partitioned into at least two or more subsets such that the analysis is initially performed on a single subset (learning set), whereas the other subset(s) (test set) is retained for subsequent use in testing and validating the initial analysis. This splitting can be done many times independently to better assess the accuracy of the classifier.

The SBDS-interacting gene products RRP42 and SKI6 share GO terms linking them to ribosomal RNA (rRNA) processing. Although *OIP2* is currently not annotated in GO, it is listed as an rRNA-processing factor in the UniProt Knowledgebase[67]. Using the simple computational procedure known as majority voting (BOX 1), one can readily predict that SBDS is probably a rRNA processing factor. Indirect experimental evidence is consistent with this prediction. For example, Wu *et al.*[43] identified a potential defect in rRNA processing in a yeast mutant strain that lacked the putative *SBDS* orthologue *YLR022C*. Savchenko *et al.*[68] further studied the SBDS sequence homologue YLR022C in *Saccharomyces cerevisiae*, and found a physical association with over 20 proteins involved in ribosome biosynthesis. In addition, Austin *et al.*[65] have reported that the SBDS protein is particularly concentrated within the human nucleolus, the site of ribosome biogenesis.

Although the relationship of defective non-coding RNA processing to the emergence of cancer is not apparent, and indeed is not pre-eminent among known cancer genes (FIG. 1), it should be noted that gene functions predicted by computation often fall into this broad functional category[15]. This might, in turn, reflect the preponderance of RNA-processing factors encoded by the human genome, underappreciated biases in functional genomic databases[59], or an overall higher coherence of the molecular signatures shown by members of this functional class[69].

As a second example, *MLLT2* has a pivotal role in leukaemogenesis in infancy[70]. Although this gene product has no biological process annotations listed in GO, we were able to derive a sub-network of functional associations centred around MLLT2 (FIG. 4b) using publicly available information[47] as follows: first we obtained the physical interaction partners of all gene products directly connected to MLLT2 (shown in the inner circle), which included GNA11, GNAI3 and NACA. Next, we obtained the interaction partners of these gene products, which are therefore only indirectly linked to MLLT2 (shown in the outer circle). This broader sub-network included a total of 2,314 gene products. Strikingly, 104 of these factors, including the immediate MLLT2 neighbours GNA11 and GNAI3, have previously been linked to G-protein coupled receptor (GPCR) signalling (over representation).

Using the procedure of majority voting, one can readily assign MLLT2 as a cancer-related GPCR given the predominant function of two out of three of its immediate neighbours. However, as the other interacting gene product, NACA, has only been linked to transcription, it might be better to examine the broader neighbourhood of MLLT2 interactions. Therefore, we built a function-prediction engine based on label propagation[71] (BOX 1), which surveys the entire interaction network for functional coherence. Prediction performance was then assessed by 10-fold cross validation. Reassuringly, MLLT2 was again connected to G-protein signalling with high confidence (FIG. 4b).

As a final stringent test, we re-examined the usefulness of the label propagation method by considering

Table 3 | **User-friendly freeware computational tools for network integration and function analysis**

Tool name	URL	Reference
STRING	http://string.embl.de	66
GRIFn	http://avis.princeton.edu/GRIFn	59
DAVID	http://niaid.abcc.ncifcrf.gov/	85
AVID	http://wbe.mit.edu/biology/keating/AVID	86
PLEX	http://bioinformatics.icmb.utexas.edu/plex	87

AVID, Annotation Via Integration of Data; DAVID, Database for Annotation, Visualization, and Integrated Discovery; PLEX, Protein Link Explorer ; STRING, Search Tool for the Retrieval of Interacting Genes/Proteins.

Over-fitting
The phenomenon in which a model has too many free parameters relative to the amount of data, which results in the learning of not only the true functional associations, but also noise and other spurious correlations. A model which has been over-fitted will not make good predictions on fresh (previously unseen) data — that is, the classifier will not generalize well.

Receiver operating characteristic
ROC curves are usually drawn by plotting sensitivity versus specificity or positive predictive value versus recall to evaluate the performance of computational methods in the cross-validation procedure.

Sensitivity
Also called recall. A measure of the ability of a classifier to assign all appropriate genes present in the test dataset the correct relevant functional label. Sensitivity is the proportion of all known members of a functional category for which there is a positive assignment, as determined by the number of true positives divided by the sum of true positives and false negatives. (Contrast with specificity.)

Specificity
An operating characteristic of a functional-prediction procedure that measures the ability of a classifier to exclude the presence of a label when it is truly not warranted. Specificity is defined as the number of true negatives divided by the sum of true negatives and false positives. (Contrast with sensitivity and recall.)

Precision
Also called 'positive predictive value'. The proportion of gene products with a predicted function that truly have the assigned biological attributes, as determined by the number of true positives divided by the sum of true positives and false positives.

an assumed case wherein the interaction partner GNA11 is deemed to have no functional annotations in GO (FIG. 4c) although all other information is preserved. Conceptually, in this scenario, majority voting would arbitrarily link MLLT2 to either transcription or GPCR terms with equal probability. However, the label propagation method was able to assign MLLT2 to G-protein signalling (FIG. 4c). Therefore, one can achieve a robust convergence to the most likely function of the cancer gene product even with incomplete data. Consistent with this prediction, a recent study of cancer-gene-expression patterns[28] has implicated a GPCR-related functional module in the emergence of acute leukaemia.

What are the roadblocks to doing it better?

In a sense, our current level of understanding of cancer in molecular terms dictates the effectiveness of the discovery process. If nothing is known about a specific class of gene function, computational approaches will be essentially useless at filling in the gap. If a little is known, computation might be helpful for refining hypotheses, but will frequently be inaccurate to the point of being deceptive much of the time. However, when a fair amount of relevant information is available, computation can be very good at assigning outstanding pertinent functions, or at least deriving testable ideas about the organization of cancer genes across many functional layers. Presently, in cancer biology, we are somewhere between 'too little' and 'a lot' in terms of comprehension; therefore, there is considerable opportunity, and risk, from a computational standpoint.

Two other troublesome practical issues still make it difficult to apply computational methods to annotate cancer genes in as effective and comprehensive a manner as might be desired. The first relates to the choice of gold standards. A shortcoming of the natural language-based representations, like the GO, is that it is difficult to finely tune annotation terms for genes to reflect the true complexity of biological functions and relationships, such as regulatory relationships in signalling cascades. An example of this shortcoming is apparent in TABLE 2, which indicates that most of the existing GO annotations for known cancer genes have been made using very general, high-level terms, which limits their value. Moreover, there is a significant functional bias in the current GO database, as certain biological processes have been studied experimentally in far greater depth than others. Genes associated with these processes are much more likely to be correctly detected by computational prediction procedures[59].

The second issue is a computational problem that relates, in turn, to the fact that cancers emerge in a multi-step process. Tumours form from pre-malignant cells that harbour lesions in multiple interacting pathways[72], with progressive alterations ultimately resulting in the uncontrolled proliferation of a single clone. As individual cancer genes can be pleiotropic across the various stages of disease progression, instances of cancer genes in the training sets must each be associated with a specific set of gold-standard annotations. To compound matters, it is unlikely that all of the relevant functions and pleiotropic overlaps have been equally well explored by researchers, even for well-studied oncogenes and tumour suppressors. Previous studies that predicted gene function on a genome-wide basis[17,61,73] did not directly consider this multi-function prediction problem. Most computational procedures only support one functional assignment per uncharacterized gene at a time, and fail to exploit the correlation structure that connects functional classes during the prediction process. It has been shown that the prediction accuracy of supervised classification methods, such as support vector machines, can be greatly improved by taking into account the hierarchical structure of annotations like those in the GO database[18]. Therefore, it will be interesting to explore more general computational approaches in which the relationships defined by the hierarchical structure of functional classification schemas, such as the GO database, and functional association networks can be considered directly.

Outlook for the cancer biologist

Although most confirmed cancer genes[20] have been deduced from knowledge about familial syndromes or single-gene defects, an emerging paradigm shift is changing the way biologists approach the study of cancer. It is now generally understood that to more fully grasp the mechanistic basis of the cancer phenotype, the community needs to elucidate the molecular properties and inter-relationships of all cancer-gene products, including those involved in metastasis and the development of therapeutic resistance[72]. In this sense, cancer is the ultimate systems biology problem[74]. Most solid tumours and their progression result from many molecular changes, all of which contribute

Discriminant value
A relative measure of confidence that the cancer gene is in the functional category in question.

Genomic context
Similarity among the evolutionary attributes of gene products, such as the propensity of functionally linked gene products to co-occur across the genomes of several species, to be involved in gene-fusion events, or to be conserved in close chromosomal proximity.

Multi-function prediction
A computational procedure wherein a cancer gene product is assigned to at least two or more functional classes.

Correlation structure
A statistical measure of the relationships observed between all pair-wise functional classes examined.

Support vector machine
A popular learning algorithm that performs binary or multi-class supervised classification tasks.

to disease development. Different cells, pathways and even host immune responses all have decisive roles. Therefore, the natural context of an oncoprotein must be taken into account to correctly predict cancer-gene function. By the same token, one must also consider the function of entire pathways and networks so that we can build the modular architecture of a functional-association network and decode its role in the evolutionary process.

In this Review, we have discussed the opportunities and challenges for cancer-gene annotation and classification currently faced by cancer biologists who wish to apply computational approaches to assess cancer-gene function. An important argument for network-based computation is that no one person can envision all the inputs into the molecular equation that leads to tumour formation, bringing modelling to the forefront as a key tool in the future of cancer biology. Although the mathematics behind many of the prediction algorithms is often complicated, computation is not black magic. Many of the computational tools being introduced today are increasingly easy to use by non-computational scientists, and are therefore more accessible on a pragmatic level. Because the interested oncologist needs to make a sizeable investment to become practically familiar with the emerging technology, collaborative interactions with experts in computation will probably remain the norm for the next few years. However, it remains equally crucial for the cancer community to be vigilant about the quality of basic cancer data resources, such as gene-expression profiles and protein–protein interaction datasets, with the aim of generating more reliable quantitative measurements of cancer-gene products in a systematic genome-wide manner.

The emerging systems biology of cancer involves dealing with the complex nature of the disease in an integrative analytical framework that incorporates data and outputs testable hypotheses that extend and refine our understanding of the architecture of cancer pathways and the identity of uncharacterized genes with causal roles in either the initiation, maintenance or spreading of the disease. Although the road map of how we get there is still uncertain, promising new computational frameworks are now on the horizon[12,17,58,61,73], which will potentially permit the evaluation of even broader sources of cancer-related information, such as genetic linkage and/or association information, single nucleotide polymorphisms (SNPs) and gene copy numbers, thereby facilitating an even deeper understanding of the complex genomic–genetic–phenotypic networks that ultimately dictate the altered behaviour of malignant cells. Realizing the power of computational prediction to delve more deeply into the fundamental biology of cancers should improve pharmaceutical drug development and create a more rational and predictive approach to the application of therapeutic strategies.

1. Hanash, S. Integrated global profiling of cancer. *Nature Rev. Cancer* **4**, 638–644 (2004).
2. Rhodes, D. R. & Chinnaiyan, A. M. Integrative analysis of the cancer transcriptome. *Nature Genet.* **37** (Suppl.), S31–S37 (2005).
3. Segal, E., Friedman, N., Kaminski, N., Regev, A. & Koller, D. From signatures to models: understanding cancer using microarrays. *Nature Genet.* **37**, S38–S45 (2005).
4. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nature Med.* **10**, 789–799 (2004).
5. van't Veer, L. J. *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
6. Kastan, M. B. & Bartek, J. Cell-cycle checkpoints and cancer. *Nature* **432**, 316–323 (2004).
7. Roberts, R. J. Identifying protein function — a call for community action. *PLoS Biology* **2**, E42 (2004).
8. Alm, E. & Arkin, A. P. Biological networks. *Curr. Opin. Struct. Biol.* **13**, 193–202 (2003).
9. Barabasi, A. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Rev. Genet.* **5**, 101–113 (2004).
The authors review current network tools that can be used to understand the cell's functional organization and evolution.
10. Mateos, A. *et al*. Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res.* **12**, 1703–1715 (2002).
11. Pavlidis, P., Weston, J., Cai, J. & Noble, W. S. Learning gene functional classifications from multiple data types. *J. Comp. Biol.* **9**, 401–411 (2002).
12. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & Botstein, D. A Bayesian framework for combining heterogeneous data source for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci USA* **100**, 8348–8353 (2003).
The authors present an effective computational method to integrate different functional-association data sets for gene-function prediction.
13. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein–protein interactions. *Genome Res.* **12**, 37–46 (2002).
14. Lee, L., Date, S. V., Adai, A. T. & Marcotte, E. M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
15. Zhang, W. *et al*. The functional landscape of mouse gene expression. *J. Biol.* **3**, 21 (2004).
16. Lanckriet, G. R. G., Deng, M., Gristianini, N., Jordan, M. I. & Noble, W. S. Kernel-based data fusion and its application to protein function prediction in yeast. *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, 300–311 (2004).
17. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. & Singh, M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* **21** (suppl. 1), i302–i310 (2005).
The authors present one of the most efficient network-based label-propagation methods to make gene-function predictions using functional-association data.
18. Barutcuoglu, Z., Schapire, R. E. & Troyanskaya, O. G. Hierarchical multi-label prediction of gene function. *Bioinformatics* **22**, 830–836 (2006).
19. Vidal, M. Interactome modeling. *FEBS Lett.* **579**, 1834–1838 (2005).
20. Futreal, P. A. *et al*. A census of human cancer genes. *Nature Rev. Cancer* **4**, 177–183 (2004).
21. Strausberg, R. L., Simpson, A. J. & Wooster, R. Sequence-based cancer genomics: progress, lessons and opportunities. *Nature Rev. Genet.* **4**, 409–418 (2003).
22. Koenig, M. *et al*. Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the *DMD* gene in normal and affected individuals. *Cell* **50**, 509–517 (1987).
23. Tannock, I. F., Hill, R. P., Bristow, R. G. & Harrington, L. *The basic science of oncology* 4th ed. (McGraw Hill Companies Inc., New York, 2005).
24. Clark, J. *et al*. Genome-wide screening for complete genetic loss in prostate cancer by comparative hybridization onto cDNA microarrays. *Oncogene* **22**, 1247–1252 (2003).
25. American Cancer Society. Cancer Facts and Figures 2006. *American Cancer Society* [online], http://www.cancer.org/downloads/STT/CAFF2006PWSecured.pdf
26. Balmain, A., Gray, J. & Ponder, B. The genetics and genomics of cancer. *Nature Genet.* **33** (Suppl.), 238–244 (2003).
27. Demant, P. Cancer susceptibility in the mouse: genetics, biology and implications for human cancer. *Nature Rev. Genet.* **4**, 721–734 (2003).
28. 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).
The authors develop a strategy to identify functional modules that are common among, or unique to, different types of tumours. The set of genes in each module can also be treated as a gold standard for cancer-gene-function prediction.
29. Altschul, S. F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* **25**, 3389–3402 (1997).
30. Wiseman, B. S. & Werb, Z. Stromal effects on mammary gland development and breast cancer. *Science* **296**, 1046–1049 (2002).
31. Sawyers, C. L. Chronic myeloid leukemia. *N. Engl. J. Med.* **340**, 1330–1340 (1999).
32. Harris, M. A. *et al*. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32** (Database issue), D258–D261 (2004).
33. Chen, Y. & Xu, D. Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **32**, 6414–6424 (2004).
34. Wu, H., Su, Z., Mao, F., Olman, V. & Xu, Y. Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic Acids Res.* **33**, 2822–2837 (2005).
35. Ronald, L. *et al*. Human homolog of patched, a candidate gene for the basal cell nevus syndrome. *Science* **272**, 1668–1671 (1996).
36. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article 17 (2005).
37. Pawson, T. & Nash, P. Assembly of cell regulatory systems through protein interaction domains. *Science* **300**, 445–452 (2003).
38. Barrios-Rodiles, M. *et al*. High-throughput mapping of a dynamic signaling network in mammalian cells. *Science* **307**, 1621–1625 (2005).
39. Bouwmeester, T. *et al*. A physical and functional map of the human TNF-α/NF-κB signal transduction pathway. *Nature Cell Biol.* **6**, 97–105 (2004).

40. Stelzl, U. *et al*. A human protein–protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
41. Rual, J. F. *et al*. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178 (2005).
42. Boyer, L. A. *et al*. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
43. Wu, L. F. *et al*. Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nature Genet.* **31**, 255–265 (2002).
44. Kislinger, T. *et al*. Global survey of organ and organelle selective protein expression in mouse: integrated proteomic, genomic and bioinformatic analysis. *Cell* **125**, 173–186 (2006).
45. Bandyopadhyay, S., Sharan, R. & Ideker, T. Systematic identification of functional orthologs based on protein network comparison. *Genome Res.* **16**, 428–435 (2006).
46. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
47. Jonsson, P. F. & Bates, P. A. Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**, 2291–2297 (2006).
The authors show that human proteins translated from known cancer genes have a protein–protein interaction network topology that is different from that of proteins not documented as being mutated in cancer.
48. Bader, G. D., Cary, M. P. & Sander, C. Pathguide: a pathway resource list. *Nucleic Acids Res.* **34** (Database issue), D504–D506 (2006).
49. Chua, H. N., Sung, W. & Wong, L. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics* **22**, 1623–1630 (2006).
50. Brun, C., Herrmann, C. & Guenoche, A. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics* **5**, 95 (2004).
51. Pereira-Leal, J. B., Enright, A. J. & Quzounis, C. A. Detection of functional modules from protein interaction networks. *Proteins* **54**, 49–57 (2004).
52. Farutin, V. *et al*. Edge-count probabilities for the identification of local protein communities and their organization. *Proteins* **62**, 800–818 (2006).
53. Adamcsek, B. *et al*. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* **22**, 1021–1023 (2006).
54. Aittokallio, T. & Schwikowski, B. Graph-based methods for analyzing networks in cell biology. *Brief. Bioinformatics* **7**, 243–255 (2006).
55. Schwikowski, B., Uetz, P. & Fields, S. A network of protein–protein interactions in yeast. *Nature Biotechnol.* **18**, 1257–1261 (2000).
56. Tsuda, K. & Noble, W. S. Learning kernels from biological networks by maximizing entropy. *Bioinformatics* **20** (Suppl. 1), I326–I333 (2004).
57. Goldstein, D. R., Ghosh, D. & Conlon, E. M. Statistical issues in the clustering of gene expression data. *Statistica Sinica* **12**, 219–240 (2002).
58. Jansen, R. & Gerstein, M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. Microbiol.* **7**, 535–545 (2004).
The authors discuss how to define protein functions and select gold standards for protein-function prediction using functional-association data.
59. Myers, C. L., Barrett, D. R., Hibbs, M. A., Huttenhower, C. & Troyanskaya, O. G. Finding function: evaluation methods for functional genomic data. *BMC Genomics* **7**, 187 (2006).
The authors discuss the deficiencies of current computational methods to infer functions from functional-association data, and outline new approaches to deal with these problems.
60. Devos, D. & Valencia, A. Intrinsic errors in genome annotation. *Trends Genet.* **17**, 429–431 (2001).
61. Letovsky, S. & Kasif, S. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* **19** (Suppl. 1), i197–i204 (2003).
62. Tsuda, K., Uda, S., Kin, T. & Asai, K. Minimizing the cross validation error to mix kernel matrices of heterogeneous biological data. *Neural Process. Lett.* **19**, 63–72 (2004).
63. Boocock, G. R. *et al*. Mutations in *SBDS* are associated with Shwachman–Diamond syndrome. *Nature Genet.* **33**, 97–101 (2003).
64. Woloszynek, J. R. *et al*. Mutations of the *SBDS* gene are present in most patients with Shwachman–Diamond syndrome. *Blood* **104**, 3588–3590 (2004).
65. Austin, K. M., Leary, R. J. & Shimamura, A. The Shwachman–Diamond SBDS protein localizes to the nucleolus. *Blood* **106**, 1253–1258 (2005).
66. von Mering, C. *et al*. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261 (2003).
67. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
68. Savchenko, A. *et al*. The Shwachman–Bodian–Diamond syndrome protein family is involved in RNA metabolism. *J. Biol. Chem.* **280**, 19213–19220 (2005).
69. Martinez, N. *et al*. The molecular signature of mantle cell lymphoma reveals multiple signals favoring cell survival. *Cancer Res.* **63**, 8226–8232 (2003).
70. Yamamoto, S. *et al*. High frequency of fusion transcripts of exon 11 and exon 4/5 in *AF-4* gene is observed in cord blood, as well as leukemic cells from infant leukemia patients with t(4;11)(q21;q23). *Leukemia* **12**, 1398–1403 (1998).
71. Zhu, X., Ghahramani, Z. & Lafferty, J. Semi-supervised learning using Gaussian fields and harmonic functions. *Proc. Twentieth Int. Conf. Machine Learning* **20**, 912–919 (2003).
72. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
73. Karaoz, U. *et al*. Whole-genome annotation by using evidence integration in functional – linkage networks. *Proc. Natl Acad. Sci. USA* **101**, 2883–2893 (2004).
74. Khalil, I. G. & Hill, C. Systems biology for cancer. *Curr. Opin. Oncol.* **17**, 44–48 (2005).
75. Deng, M. & Chen, T. S. & Sun,F. An integrated probabilistic model for functional prediction of proteins. *Proc. Seventh Ann. Int. Conf. Res. Comp. Mol. Biol. (RECOMB), Berlin, Germany*, 95–103 (2003).
76. Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. Global protein function prediction from protein-protein interaction networks. *Nature Biotechnol.* **21**, 697–700 (2003).
77. Mewes, H. W. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34 (2002).
78. Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C. & Conklin, B. R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genet.* **31**, 19–20 (2002).
79. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32** (Database issue), D277 – D280 (2004).
80. Bader, G. D., Betel, D. & Hogue, C. W. BIND: the biomolecular interaction network database. *Nucleic Acids Res.* **31**, 248–250 (2003).
81. Hermjakob, H. *et al*. IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32** (Database issue), D452–D455 (2004).
82. 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).
83. Xenarios, I. *et al*. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305 (2002).
84. Zanzoni, A. *et al*. MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140 (2002).
85. Dennis, G. Jr *et al*. DAVID: database for annotation, visualization, and Integrated discovery. *Genome Biol.* **4**, R60 (2003).
86. Jiang, T. & Keating, A. E. AVID: an integrative framework for discovering functional relationships among proteins. *BMC Bioinformatics* **6**, 136 (2005).
87. Date, S. V. & Marcotte, E. M. Protein function prediction using the protein link explorer (PLEX). *Bioinformatics* **21**, 2558–2559 (2005).
88. Brown, K.R. & Jurisica, I. Online predicted human interaction database. *Bioinformatics* **21**, 2076–2082 (2005).
89. Maere, S., Heymans, K. & Kuiper, M. BINGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448–3449 (2005).
90. Al-Sharour, F., Minguez, P., Vaquerizas, J.M., Conde, L. & Dopazo, J. Babelomics: a suite of web/tools for functional annotation and analysis of groups of genes in high-thoughout experiments, *Nucleic Acids Res.* **33**, W460–W464 (2005).

Acknowledgements
We thank H. Jiang, Q. Morris and B. Noble for their critical feedback and thoughtful suggestions, R. Isserlin for skillful preparation of the GO-tree analysis and M. Maris for expert computational support. This work was supported in part by funds from Genome Canada and the Ontario Genomics Institute to A.E.

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
MLL2 | *OIP2* | patched | *RRP42* | *SBDS* | *SKI6* | smoothened

FURTHER INFORMATION
Andrew Emili's homepage: http://www.utoronto.ca/emililab
Gary Bader's homepage: http://baderlab.org
Sanger Centre: http://www.sanger.ac.uk/genetics/CGP/Census
Gene Expression Omnibus: http://www.ncbi.nlm.nih.gov/projects/geo
Oncomine: http://www.oncomine.org/main/index.jsp
Kyoto Encyclopedia of Genes and Genomes: http://www.genome.jp/kegg
IntAct: http://www.ebi.ac.uk/intact/site/
Biomolecular Interaction Network Database: http://bond.unleashedinformatics.com/Action?
Search Tool for the Retrieval of Interacting Genes/Proteins: http://string.embl.de
Molecular Interaction Network: http://mint.bio.uniroma2.it/mint/Welcome.do
Database of Interacting Proteins: http://dip.doe-mbi.ucla.edu/
Gene Ontology website: http://www.geneontology.org/GO.evidence.shtml
Cytoscape: http://cytoscape.org
Cancer Cell Map web site: http://cancer.cellmap.org
UniProt Knowledgebase: www.expasy.org/uniprot/
Access to this interactive links box is free online.