Pathway and network analysis of cancer genomes

Pau Creixell[1,21,22], Jüri Reimand[2,22], Syed Haider[3], Guanming Wu[3,4], Tatsuhiro Shibata[5], Miguel Vazquez[6], Ville Mustonen[7], Abel Gonzalez-Perez[8], John Pearson[9], Chris Sander[10], Benjamin J Raphael[11], Debora S Marks[12], B F Francis Ouellette[3,13], Alfonso Valencia[6], Gary D Bader[2], Paul C Boutros[3,14,15], Joshua M Stuart[16,17], Rune Linding[1,18], Nuria Lopez-Bigas[8,19] & Lincoln D Stein[3,20] for the Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium

Genomic information on tumors from 50 cancer types cataloged by the International Cancer Genome Consortium (ICGC) shows that only a few well-studied driver genes are frequently mutated, in contrast to many infrequently mutated genes that may also contribute to tumor biology. Hence there has been large interest in developing pathway and network analysis methods that group genes and illuminate the processes involved. We provide an overview of these analysis techniques and show where they guide mechanistic and translational investigations.

As sequencing costs continue to decrease, it is becoming common to assay genomic information from a cohort of cancer patients at the level of single-nucleotide variants (SNVs) and copy-number alterations (CNAs). Other alterations including structural changes, fusion transcripts and epigenetic reprogramming are also studied routinely. These genomic data are associated with rich clinical annotation, and some groups have begun to incorporate sequencing into standard clinical practice[1]. Recent studies have painted a portrait of the mutation landscape for multiple cancers[2] including pancreatic[3], lung[4], breast[5], brain[6] and ovarian[7]. In each case, the distribution of somatic SNVs across the samples typically includes a few altered genes at frequencies higher than 10% and a long 'tail' of many genes mutated at frequencies of 5% or lower[2,8]. Interestingly, some tumor types, including prostate cancer and some pediatric cancers, have relatively few SNVs or CNAs[9]; their biology is presumably driven by other types of somatic variation such as DNA methylation[10]. Driver genes are detected mostly from positive-selection signals found in the mutation patterns of individual genes across tumors[11]. However, this approach will miss less-frequently mutated but functionally important genes that a typical cohort with hundreds of tumor samples is not statistically powered to detect[12]. Recent pan-cancer analyses have detected cancer genes using several thousand samples of different tumor types; however, these studies still remain limited in power because of tissue-specific drivers such as *APC* in colorectal and ovarian cancers, *VHL* in renal cell carcinoma and *ERG* fusion genes in prostate cancers. Alternatively, grouping of genetic alterations using prior knowledge about cellular mechanisms allows investigation of the full complement of mutations in a tumor and the determination of affected oncogenic pathways.

[1]Cellular Signal Integration Group (C-SIG), Technical University of Denmark, Lyngby, Denmark. [2]The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [3]Informatics and Biocomputing Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [4]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA. [5]Division of Cancer Genomics, National Cancer Center, Tokyo, Japan. [6]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre, Madrid, Spain. [7]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [8]Research Unit on Biomedical Informatics, University Pompeu Fabra, Barcelona, Spain. [9]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, University of Queensland, St. Lucia, Brisbane, Queensland, Australia. [10]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [11]Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, USA. [12]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [13]Department of Cell and Systems Biology, University of Toronto, Toronto, Ontario, Canada. [14]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [15]Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada. [16]Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, California, USA. [17]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, Santa Cruz, California, USA. [18]Biotech Research & Innovation Centre (BRIC), University of Copenhagen (UCPH), Copenhagen, Denmark. [19]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain. [20]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [21]Present address: Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [22]These authors contributed equally to this work. Correspondence should be addressed to L.D.S. (lincoln.stein@gmail.com).

RECEIVED 20 JANUARY; ACCEPTED 27 APRIL; PUBLISHED ONLINE 30 JUNE 2015; DOI:10.1038/NMETH.3440

In this Perspective, the phrase 'pathway and network analysis' denotes any analytic technique that benefits from biological pathway or molecular network information to gain insight into a biological system. The fundamental aim is to reduce data involving thousands of altered genes and proteins to a smaller and more interpretable set of altered processes (see recent reviews[13,14]). This process-oriented view helps generate testable hypotheses, identify drug targets, find tumor subtypes with clinically distinct outcomes and identify both cancer-specific and cross-cancer pathways.

Pathways and networks are similar concepts with certain distinctions. Both comprise systems of interacting genes, proteins and other biomolecules that carry out biological functions. Pathways are small-scale systems of well-studied processes where interactions comprise biochemical reactions and events of regulation and signaling. Pathways represent consensus systems based on decades of research and can be visualized in detailed linear diagrams. In contrast, networks comprise genome- or proteome-wide interactions derived from large-scale screens or integrative analyses of multiple data sets. Network interactions are simplified abstractions of complex cellular logic. For instance, physical protein-protein interactions may be represented as directionless edges, and directed edges may stand for inhibitory or activating gene regulation. Networks are noisy and challenging to visualize and interpret; however, they likely contain information not covered in well-defined pathways. A related concept to pathways and networks is a functionally annotated gene set that comprises all genes involved in a particular process or pathway without their interactions. Annotated gene sets of the Gene Ontology and other resources are based on multiple types of evidence and are broader in scope than pathways.

Pathway and network analyses have a number of benefits relative to analyzing genomics data at the level of individual genes. First, these techniques aggregate molecular events across multiple genes in the same pathway or network neighborhood, thus increasing the likelihood that the events will pass a statistical detection threshold and reducing the number of hypotheses tested[15]. Second, the results are often easier to interpret, as genomic alterations are related to familiar concepts such as cell cycle or apoptosis. Third, potential causal mechanisms can be identified—for instance, by predicting a particular microRNA or transcription factor that explains expression differences between tumor samples and controls. Fourth, results obtained from related data sets may become more comparable because pathway information allows interpretation in a common feature space. Finally, the techniques facilitate integration of diverse inputs such as genomic, transcriptomic and proteomic data into a unified view of tumor biology, thereby improving statistical and interpretative power.

Pathway and network analyses have been applied to cancer data sets to find driver genes and pathways[16,17], to identify hidden tumor subtypes distinguished by common patterns of network alteration[18], to propose cancer mechanisms and biomarkers[17,18] and to identify key regulators of cancer-related gene networks[19,20].

The Mutation Consequences and Pathway Analysis (MUCOPA) working group of the ICGC[21] has developed standard operating procedures for the analysis of cancer genome data generated by the ICGC. In a recent review[11] we outlined our recommendations for prioritizing somatic mutations using gene-level statistics, including criteria for the functional impact of mutations and positive selection for mutations in genes within the patient population. Here we describe diverse analytic techniques to prioritize altered gene sets, pathways and networks consisting of multiple interacting genes. Although we focus on somatic SNVs and altered RNA expression, the concepts are generally applicable to other oncogenic alterations such as CNAs, epigenetic changes and genomic rearrangements, though the details of analysis, including data processing and confounding-factor control, can be different for other data types.

Major types of pathway and network analysis techniques

We consider three major approaches to network and pathway analyses to interpret somatic cancer mutations, which we present in order of complexity (**Fig. 1**). The simplest analysis provides a high-level summary of pathways affected in the tumor, whereas more complex methods provide detailed hypotheses about affected cellular mechanisms. We recommend that approaches from each of these classes be applied to cancer genome sequencing projects wherever feasible.

All three approaches require two general resources. The first is a list of oncogenic alterations that affect protein-coding genes. The second is a database of gene sets, pathways or network interactions[22]. Gene-set databases are lists of genes that have been grouped according to common biological properties; a familiar example would be the association of gene products with the Gene Ontology controlled vocabulary of biological processes, molecular functions and cellular locations[23]. Pathway databases represent biological processes as series of biochemical reactions and other physical events (for example, complex formation, phosphorylation events, subcellular localization and conformational changes), whereas network databases use a simpler data model that treats biological processes as sets of bimolecular interactions. A simplified version of the epidermal growth factor (EGF) pathway illustrates the essential difference between pathway and network interaction databases (**Fig. 2**). The first approach, fixed–gene set enrichment analysis, analyzes functionally annotated gene sets that can be extracted from either pathway databases or network interaction databases. Inputs to the second approach, *de novo* network construction and clustering, are provided by network interaction databases. And in the most sophisticated approach, network-based modeling, both types of databases are used.

Approach 1: fixed–gene set enrichment analysis. This approach treats pathways, biological processes and networks simply as gene sets and ignores information about their interactions. Fixed–gene set enrichment analysis identifies genes in pathways (or any other functionally related grouping) that are present in a gene list more frequently than expected by chance. The gene sets are usually collected from curated community databases or the gene annotation tables of the Gene Ontology[23] but may also be experimentally derived (for example, genes upregulated in a cell line exposed to low oxygen levels). Several recommended software tools are available (**Supplementary Table 1**). The simplest input to such analysis is a list of genes that is most differentially expressed or frequently mutated in a data set. A typical analysis workflow consists of two steps: (i) a gene list is defined by filtering experimental data for genes with significant gene-level statistics, and

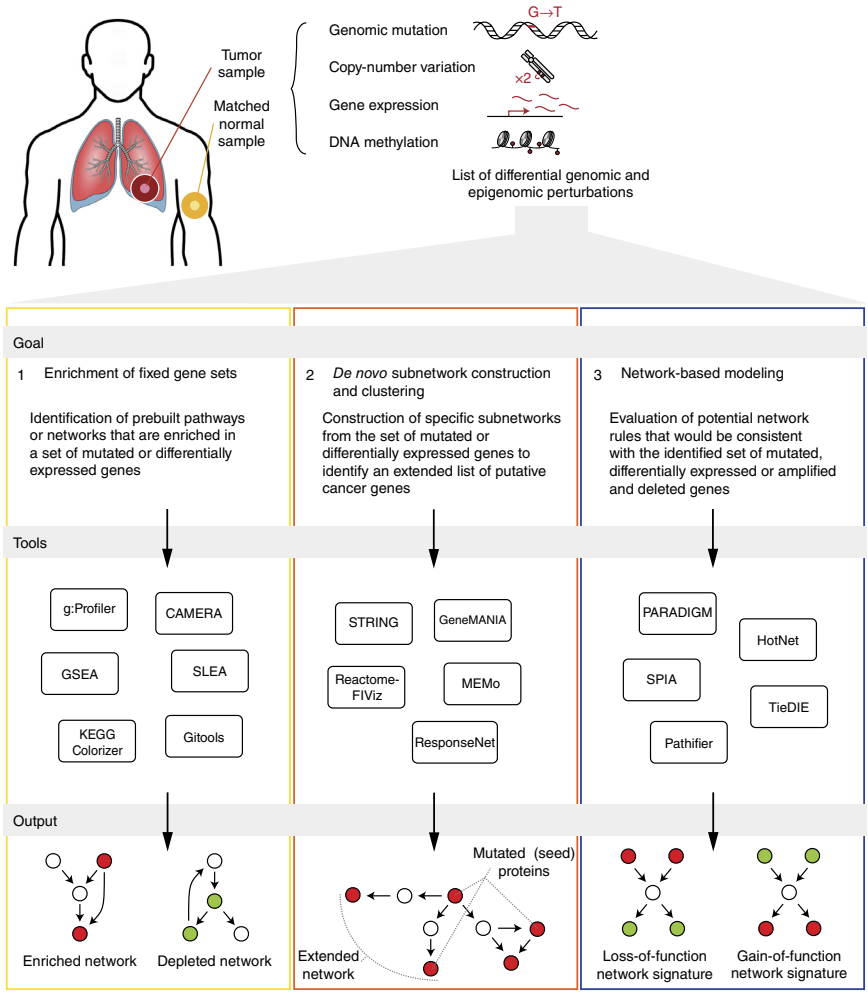

Figure 1 | Major approaches to pathway and network analysis of cancer data. Some commonly used tools are included for each approach. In the network diagrams ("Output"), red circles indicate genes whose activities are increased (first and third columns) or altered by mutations (center column). Green circles are genes whose activities are decreased.

(ii) enrichment analysis is performed to determine processes and pathways over-represented in the gene list.

A hypergeometric distribution (Fisher's exact test) is commonly used to calculate the statistical significance of this over-representation, followed by a correction for multiple testing to estimate the proportion of enriched gene sets that would occur by chance given the number of tested gene sets. The basic form of this test is applied in many tools (**Supplementary Table 1**) including the widely used but no longer updated web service DAVID[24]. However, the key drawback of this approach is that an arbitrary threshold is used to select the input genes, and potentially informative genes below the threshold are excluded. An alternative approach enables interpretation of a ranked list of genes in the experiment (for example, by strength of differential expression) with the assumption that top-ranking genes are more important in terms of biological function. One recommended web service, g:Profiler[25], applies a modified hypergeometric test to analyze increasingly complete ranked lists of input genes and determines a sub-list with the strongest level of enrichment. The GSEA method[26] is designed to work with continuous data and searches for gene sets that are enriched at the top (overexpressed vs. control) or bottom (underexpressed) of a ranked list of all genes. Both methods score each gene set separately and compute additional statistics to estimate *P* values and make multiple-testing corrections with false discovery rate.

Enhancements of these approaches allow enrichment analysis for each tumor sample, thereby enabling the discovery of distinct cancer subtypes from different enrichment patterns. Examples of methods that allow comparisons among samples include sample-level enrichment analysis (SLEA)[27], single-sample GSEA (ssGSEA)[26] and gene-set variation analysis (GSVA)[28].

Rank-based enrichment methods do best when genes are easily ranked but may be suboptimal in scenarios such as cancer mutation analysis in which most genes are difficult to rank owing to low mutation counts. A pathway association analysis may be helpful in case of a two-class experimental design (for example, cases vs. controls). This resembles a genome-wide association analysis and uses pathways and other gene sets instead of genetic markers. For each experimental class and gene set, one counts all samples containing a mutation that may affect that gene set. A series of Fisher's exact tests identify gene sets significantly mutated in cases versus controls, followed by multiple-testing correction.

Fixed–gene set enrichment analysis generates a list of processes and pathways and provides a bird's-eye view of affected biological systems. However, sometimes many related gene sets are enriched. The key functional themes in these large pathway lists can be identified using tools such as the Enrichment Map[29] app of the Cytoscape network visualization software[30]. Another useful approach is to overlay the original genomics data on a detailed pathway diagram or high-level molecular interaction network. For example, the databases KEGG[31], Reactome[32] and HumanCyc[33] enable diagrams of enriched pathways with colors highlighting the genes of interest. This may help researchers to move beyond asking which pathways are enriched among alterations toward understanding the functional consequences of the altered gene set.

This family of techniques is still evolving. For instance, most enrichment statistics assume that genes in the list occur independently, an assumption that does not hold true for co-regulated genes in gene expression data, overlapping or shared exons in point-mutation data, or colocated paralogous genes with similar functions. The quality and coverage of gene sets can also affect interpretation of fixed–gene set enrichment analysis, as databases report genes and their functions with variable levels of detail and confidence. Combined use of multiple databases, filtering and visualization help overcome these problems. Another issue is that many annotated pathways represent normal physiology that may be altered in disease. New methods have been developed to address these issues: for example, CAMERA[34] corrects gene-set

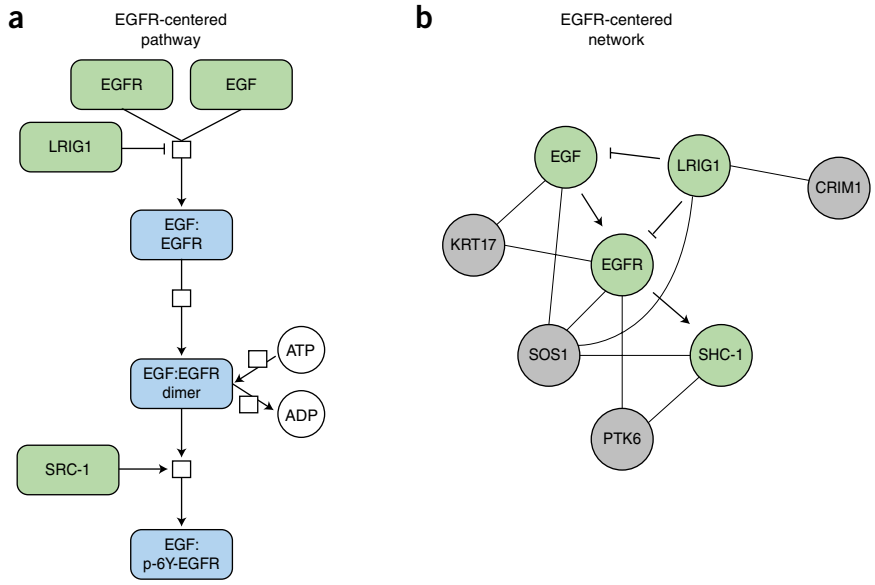

Figure 2 | Pathway and network representations of EGF signaling. (**a**) In the simplified pathway representation, heterogeneous nodes and edges correspond to genes, proteins, small molecules and their regulatory and catalytic relationships. Nodes do not interact directly but participate in reaction events designated by white squares. (**b**) In the network representation, all nodes correspond to the same type of biological entity (gene products). Edges derived from curated pathways are shown as arrows. Additional gene-gene interactions derived from gene coexpression and physical protein interactions are shown as light lines. Green nodes are proteins participating in curated reactions or interactions, blue nodes are complexes and gray nodes are uncurated proteins participating in physical interactions.

enrichment statistics for inter-gene correlations. A more fundamental limitation of this class of algorithms is their ignorance of interactions between genes and proteins, as neither network topology nor dynamics is taken into account. These limitations are addressed by the next two approaches.

Approach 2: *de novo* network construction and clustering. Methods in this section construct cancer gene networks *de novo* by analyzing networks of molecular or functional interactions. These methods begin with a list of mutated or otherwise altered genes as well as one or more databases of gene or protein interactions, such as those compiled by iRefIndex[35], BioGRID[36], IntAct[37], STRING[38] or GeneMANIA[39] (**Supplementary Table 2**). The altered genes and a subset of their neighbors are then extracted from the databases and reconstructed as an interaction network. The resulting network reveals interactions of input genes and helps discover additional related genes by 'guilt by association', highlighting nonmutated genes that likely participate in tumor biology because of their interactions. By clustering and annotating the discovered networks with the enrichment and colorization approaches described above, one may find similarities and differences among distinct tumors that would not be apparent at the gene level.

Examples of recommended network-construction algorithms include GeneMANIA[39], ReactomeFIViz[40], STRING[38], ResponseNet[41], NetBox[42], MEMo[43] and EnrichNet[44] (**Supplementary Table 2**). GeneMANIA is an interactive web service and a Cytoscape app that uses a diverse set of interaction databases. It suggests genes that are related to those in the experimental data set using network analysis. ReactomeFIViz (previously called Reactome FI Plugin) runs in Cytoscape and features a number of algorithms for clustering and annotating sets of interacting genes and for relating these clusters to tumor phenotype and patient clinical characteristics. For example, ReactomeFIViz identified prognostic biomarkers in breast and ovarian cancer[45]. NetBox is conceptually similar to ReactomeFIViz and reports functional network modules by identifying clusters of altered genes on a background network derived from databases. MEMo studies mutual exclusivity of cancer alterations in groups of genes across tumor samples to discover subnetworks of synthetic lethality and other functional groupings. It nominates sets of oncogenic alterations that have a particularly strong selective effect, potentially pointing to therapeutic combinations where mutual exclusivity reflects synthetic lethality.

A key use of networks is to search for alteration patterns in interacting genes that correlate with clinical information[46]. The HyperModules method[47] identifies subnetworks with cancer mutations that are maximally correlated with clinical characteristics such as patient survival, tumor stage or relapse. This tool can also be used to study tumor subtypes by extracting subnetworks whose mutations are significantly enriched in a particular subtype. HyperModules was applied to the kinase-signaling network in ovarian cancer and revealed network modules with mutations in phosphorylation sites and kinase domains that significantly correlated with patient survival[48].

A drawback of *de novo* network construction and clustering techniques is their use of a simplified data model that discards much information known about biological networks. For example, an alteration may act at the DNA level by deleting a portion of a gene, at the transcriptional level by disrupting a promoter, or at the protein level by altering a catalytic site. The activating effect of a mutation in a transmembrane receptor can be masked by inactivation of a downstream effector of the same signaling pathway. These subtleties are not easily captured in a binary interaction network. In addition, the molecular interactions in databases are derived from specific experiments, such as yeast two-hybrid assays, that may or may not matter for cancer biology. Thus, it is advisable to consult the literature establishing the network interactions when forming hypotheses on the basis of patterns observed in interaction networks; several text-mining tools are available to automate this task[49].

Approach 3: network-based modeling. The approaches discussed in this section infer how network states are disrupted in cancer. Network-based modeling approaches use qualitative and quantitative measurements to infer the activities and interactions of various genetic components in pathway or networks. These methods relate the activities of some components with their influences and consequences on other components. Such modeling approaches have been applied to infer the mechanisms of NRAS signaling in melanoma[50] to map transcriptional regulatory

networks in physiologically normal and diseased states[19,20,51–53], to build maps of phosphorylation networks[54] and to identify cancer drivers[16]. Below, we briefly describe several network-modeling algorithms that are available as user-installable software packages and have been applied to cancer (**Supplementary Table 3**).

The HotNet[55] tool treats the gene network as a metallic lattice and uses the physics of heat diffusion to model the effects of gene alterations. Each gene in the query 'heats up' its local region of the network, and the effect is then metaphorically propagated along metallic wires defined by gene-gene linkages, leading to 'hot' (highly relevant) network neighborhoods. This approach mitigates some of the ascertainment biases in curated gene interaction networks. For example, because tumor suppressor *TP53* is exceedingly well studied, it is an artificially inflated hub of known linkages to other genes; but because of *TP53*'s high degree of connectedness, heat diffuses away from it rapidly, reducing its overall influence. The related method TieDIE[56] extends the network-diffusion concept to integrated analysis of multiple types of genomic alterations.

The Pathifier method[57] transforms gene-level information to network-level information by quantifying molecular activities on a continuous sample-by-sample curve in the multidimensional space of gene expression values. It ranks cancer samples along a gradient of clinical or biological attributes such as tumor aggressiveness or patient survival. The method generates hypotheses and identifies testable markers to predict clinical outcomes.

Signaling pathway impact analysis (SPIA)[58] applies a recursive algorithm similar to that used by Google to rank search results. SPIA scores a gene product as highly impactful if it points to other impactful gene products in the network diagram. By ranking the effects hierarchically, SPIA distinguishes primary changes in gene activity and secondary effects of the regulatory network.

Several methods use information theoretical principles to reconstruct regulatory networks from gene expression data. Application of these methods to cancer genomics has led to insights into tumor biology and identification of actionable drug treatments. ARACNE applies mutual information to discover regulatory networks of transcription factors and target genes[59], whereas MARINa interrogates these networks to identify master regulators[19,20]. For example, application of these tools to the reconstruction of the gene regulatory network in glioblastoma and follow-up experimental validations revealed the transcription factors C/EBPβ and STAT3 to be master regulators of mesenchymal transformation[20].

Other methods integrate gene expression and CNA data to identify cancer driver genes and downstream regulatory networks. For example, CONEXIC assumes that copy-number gains and losses alter gene expression[16]. It employs a Bayesian network algorithm to find significantly altered genes regulating modules of differentially expressed genes. The approach was applied to predict and experimentally validate multiple cancer driver genes in melanoma and glioblastoma[16,17].

Several approaches have been developed to fit gene interactions to the data rather than taking the interactions as prior knowledge. Thus, interactions are not interpreted as direct physical interactions but as measures of influence between network nodes. Functions of discrete logic were used to connect gene products through 'gates'[60] and to infer functions best capturing the observed dynamics in the data. This was extended to fuzzy logic[61] that relaxes the rules of gene interactions and allows for biological noise and uncertainty. Similar approaches were developed for partial least-squares regression models[62] in which parameters are fit to dependent variables typically reflecting a cellular phenotype. These approaches were applied to interpret drug response in triple-negative breast cancer and to suggest effective therapeutic treatments[63]. The DataRail package[64] allows users to experiment with multiple similar model-fitting methods for gene networks.

Probabilistic graphical models (PGMs) have been applied to cancer network analysis. PGMs are widely used in machine learning and statistics for modeling complex dependencies among multiple variables. PathOlogist[65] analyzes pathways from curated databases to derive a set of network interactions. It then uses the inhibitory and excitatory regulatory connections in each pathway-derived network model to determine (i) whether a given cancer gene expression data set is consistent with the model and (ii) whether the pathway-derived network's components are activated. Thus, a collection of known gene interactions with details of co-regulation helps interpret gene expression data. This family of algorithms was applied to develop predictors of drug sensitivity in cancer cell lines[66].

PARADIGM[67] extends the PGM framework of PathOlogist by formally modeling the 'central dogma' of gene transcription to RNA, followed by RNA translation to protein and post-translational events, together representing pathway and network effects of alterations at the DNA, RNA and protein levels. This method uses factor graphs to assign weights to each molecular interaction and to integrate the effects of multiple simultaneous alterations (for example, copy-number changes, simple somatic mutations and expression changes). The tool provides predicted pathway activity scores by integrating all observed variations to assess whether the activities of each pathway are increased, decreased or unaffected. The algorithm was used to identify new tumor subtypes on the basis of shared pathway-activation patterns[18]. An extension called PARADIGM-Shift infers whether somatic mutations are neutral, loss-of-function or gain-of-function mutations[68]. This method has detected several well-known examples of pathway alterations, including loss-of-function events in *TP53* in breast cancer and gain-of-function events in oncogene *NFE2L2* in lung squamous cell tumors. More recent PGM approaches include the application of dynamic Bayesian networks to consider tumorigenesis as a temporally evolving system. The inferred network of breast cancer cell lines contributes an important proof of concept in this area[69].

Higher-resolution modeling of cellular wiring in cancer requires quantitative data that are not yet readily available from patient tissue samples. Established cell lines, organoids and xenograft models will enable collection of more data for integrative analysis. Time courses and perturbation experiments on such cancer models will contribute key data points that will help parameterize more realistic models such as systems of differential equations. Large interacting systems of differential equations such as full cell models[70] also show promise but are in their infancy in their application to cancer.

Challenges and future perspectives

Pathway and network analysis can effectively uncover biological systems perturbed in tumor cells. However, our knowledge of pathways and networks both in normal cells and, more acutely, in cancer cells is far from complete. Many approaches, particularly the techniques

of network-based modeling, require accurate, detailed and comprehensive pathway descriptions with regulatory relationships, orthogonal data (DNA, RNA, protein) and extensive quantitative data. Even among protein-coding genes, high-resolution data are available for only well-studied biological processes and are scarce for pathways involving many noncoding genomic elements. This argues for an expanded effort to develop pathway databases and systematically reconstruct regulatory and signaling networks.

A second challenge is the computationally expensive modeling of biological networks that can consume weeks of CPU time, particularly for permutation-based estimates of statistical significance. This problem will only grow as reference pathways and networks and experimental data sets increase in size. As cancer genomics data become available for progressively larger patient cohorts, fundamental computer science research is needed to optimize these algorithms to scale to thousands of samples[71].

A third challenge arises from the abundant interdependencies in complex biological systems. It is well established that the role of a mutation, such as its functional impact or its role in suppressing or enabling a tumor, is not static. Instead it depends on cell state and the presence of other mutations[72] and could have effects on multiple cellular processes. The establishment of annotation standards that can encapsulate such dependencies also represents a major challenge for the field.

A final challenge is the evaluation of pathway and network methods in patient care. With a sufficient battery of pathway-specific therapeutics, one can envision the selection of therapies based on networks constructed from the molecular alterations present in individual tumors. It will be a major statistical challenge to devise adaptive clinical trials that leverage such information[73]. The difficulties of communicating genomic information to clinicians and patients will certainly be exacerbated by the complexity of network-level alterations[74].

Our understanding of cancer biology through the lens of pathway and network analyses is nascent, but it holds the potential to transform our thinking on disease etiology and treatment.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We gratefully acknowledge the assistance of J. Jennings during preparation of this manuscript. J.M.S. acknowledges support from the US National Cancer Institute (R01-CA180778 and U24-CA143858), Stand Up To Cancer, the Prostate Cancer Foundation and the Movember Foundation. P.C. is currently funded by a Ludwig Fund Postdoctoral Fellowship. P.C.B. and L.D.S. were supported by the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. P.C.B. was also supported by a Terry Fox Research Institute New Investigator Award and a Canadian Institutes of Health Research New Investigator Award. L.D.S. and G.W. acknowledge support from the US National Institutes of Health (NIH) and National Human Genome Research Institute (P41 HG003751). G.D.B. is supported by NRNB (NIH, National Institute of General Medical Sciences grant number P41 GM103504).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Newman, W.G. & Black, G.C. Delivery of a clinical genomics service. *Genes (Basel)* **5**, 1001–1017 (2014).
2. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
3. Biankin, A.V. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491**, 399–405 (2012).
4. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120 (2012).
5. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409 (2012).
6. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
7. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
8. Garraway, L.A. & Lander, E.S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
This review discusses the advances and findings in cancer genome sequencing as well as current challenges of the field, including the long 'tail' of infrequently mutated genes and the need for functional validation of cancer mutations.
9. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
10. Mack, S.C. *et al.* Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature* **506**, 445–450 (2014).
11. Gonzalez-Perez, A. *et al.* Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* **10**, 723–729 (2013).
This review from the ICGC-MUCOPA working group discusses methods and recommendations to distinguish functional cancer mutations and to predict cancer driver genes.
12. Leiserson, M.D.M., Blokh, D., Sharan, R. & Raphael, B.J. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.* **9**, e1003054 (2013).
13. Pe'er, D. & Hacohen, N. Principles and strategies for developing network models in cancer. *Cell* **144**, 864–873 (2011).
14. Califano, A., Butte, A.J., Friend, S., Ideker, T. & Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* **44**, 841–847 (2012).
15. Chi, Y.Y., Gribbin, M.J., Johnson, J.L. & Muller, K.E. Power calculation for overall hypothesis testing with high-dimensional commensurate outcomes. *Stat. Med.* **33**, 812–827 (2014).
16. Akavia, U.D. *et al.* An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005–1017 (2010).
This is one of the first studies to integrate molecular data at different network levels to pinpoint tumor dependencies.
17. Danussi, C. *et al.* RHPN2 drives mesenchymal transformation in malignant glioma by triggering RhoA activation. *Cancer Res.* **73**, 5140–5150 (2013).
18. Hoadley, K.A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
19. Sonabend, A.M. *et al.* The transcriptional regulatory network of proneural glioma determines the genetic alterations selected during tumor progression. *Cancer Res.* **74**, 1440–1451 (2014).
20. Carro, M.S. *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**, 318–325 (2010).
21. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
22. Bader, G.D., Cary, M.P. & Sander, C. Pathguide: a pathway resource list. *Nucleic Acids Res.* **34**, D504–D506 (2006).
23. The Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.* **38**, D331–D335 (2010).
24. Huang, D.W. *et al.* DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* **35**, W169–W175 (2007).
25. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193–W200 (2007).
g:Profiler is a frequently updated web server for conducting fixed–gene set enrichment analysis of plain and ranked gene lists.
26. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
27. Gundem, G. & Lopez-Bigas, N. Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types. *Genome Med.* **4**, 28 (2012).

28. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
29. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G.D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984 (2010).
30. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
31. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
32. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
33. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42**, D459–D471 (2014).
34. Wu, D. & Smyth, G.K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* **40**, e133 (2012).
35. Razick, S., Magklaras, G. & Donaldson, I.M. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9**, 405 (2008).
36. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* **41**, D816–D823 (2013).
37. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* **40**, D841–D846 (2012).
38. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).
39. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214–W220 (2010).
GeneMANIA is a web server for integrative analysis of gene lists in the context of molecular interaction networks.
40. Wu, G., Dawson, E., Duong, A., Haw, R. & Stein, L. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Res.* **3**, 146 (2014).
ReactomeFIViz is a Cytoscape app with multiple algorithms for network-based clustering and analysis of the Reactome functional interaction network.
41. Lan, A. *et al.* ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Res.* **39**, W424–W429 (2011).
42. Cerami, E., Demir, E., Schultz, N., Taylor, B.S. & Sander, C. Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE* **5**, e8918 (2010).
43. Ciriello, G., Cera, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**, 398–406 (2012).
44. Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* **28**, i451–i457 (2012).
45. Wu, G. & Stein, L. A network module-based method for identifying cancer prognostic signatures. *Genome Biol.* **13**, R112 (2012).
46. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007).
47. Leung, A., Bader, G.D. & Reimand, J. HyperModules: identifying clinically and phenotypically significant network modules with disease mutations for biomarker discovery. *Bioinformatics* **30**, 2230–2232 (2014).
48. Reimand, J. & Bader, G.D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9**, 637 (2013).
49. Krallinger, M. *et al.* The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics* **12**, S3 (2011).
50. Kwong, L.N. *et al.* Oncogenic NRAS signaling differentially regulates survival and proliferation in melanoma. *Nat. Med.* **18**, 1503–1510 (2012).
51. Schadt, E.E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**, 710–717 (2005).
52. Aytes, A. *et al.* Cross-species analysis of genome-wide regulatory networks identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy. *Cancer Cell* **25**, 638–651 (2014).
53. Piovan, E. *et al.* Direct reversal of glucocorticoid resistance by AKT inhibition in acute lymphoblastic leukemia. *Cancer Cell* **24**, 766–776 (2013).
54. Bandyopadhyay, S. *et al.* A human MAP kinase interactome. *Nat. Methods* **7**, 801–805 (2010).
55. Vandin, F., Upfal, E. & Raphael, B.J. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* **18**, 507–522 (2011).
The HotNet algorithm uses a heat-diffusion model to analyze molecular interaction networks and detect significantly mutated modules in cancer.
56. Paull, E.O. *et al.* Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* **29**, 2757–2764 (2013).
57. Drier, Y., Sheffer, M. & Domany, E. Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci. USA* **110**, 6388–6393 (2013).
58. Tarca, A.L. *et al.* A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75–82 (2009).
59. Margolin, A.A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7** (suppl. 1), S7 (2006).
60. Morris, M.K., Saez-Rodriguez, J., Sorger, P.K. & Lauffenburger, D.A. Logic-based models for the analysis of cell signaling networks. *Biochemistry* **49**, 3216–3224 (2010).
61. Morris, M.K., Saez-Rodriguez, J., Clarke, D.C., Sorger, P.K. & Lauffenburger, D.A. Training signaling pathway maps to biochemical data with constrained fuzzy logic: quantitative analysis of liver cell responses to inflammatory stimuli. *PLoS Comput. Biol.* **7**, e1001099 (2011).
62. Janes, K.A. *et al.* A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science* **310**, 1646–1653 (2005).
63. Lee, M.J. *et al.* Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell* **149**, 780–794 (2012).
This study utilized integrative network analysis to identify key rewiring cellular events that informed a combination-based therapeutic strategy for specific tumors.
64. Saez-Rodriguez, J. *et al.* Flexible informatics for linking experimental data to mathematical models via DataRail. *Bioinformatics* **24**, 840–847 (2008).
65. Greenblum, S.I., Efroni, S., Schaefer, C.F. & Buetow, K.H. The PathOlogist: an automated tool for pathway-centric analysis. *BMC Bioinformatics* **12**, 133 (2011).
66. Brubaker, D. *et al.* Drug Intervention Response Predictions with PARADIGM (DIRPP) identifies drug resistant cancer cell lines and pathway mechanisms of resistance. *Pac. Symp. Biocomput.* doi:10.1142/9789814583220_0013 (2014).
67. Vaske, C.J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010).
The PARADIGM algorithm predicts the impact of oncogenic alterations on downstream pathway and network activity by modeling the 'central dogma' of gene expression.
68. Ng, S. *et al.* PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* **28**, i640–i646 (2012).
69. Hill, S.M. *et al.* Bayesian inference of signaling network topology in a cancer cell line. *Bioinformatics* **28**, 2804–2810 (2012).
70. Sanghvi, J.C. *et al.* Accelerated discovery via a whole-cell model. *Nat. Methods* **10**, 1192–1195 (2013).
71. Dittrich, M.T., Klau, G.W., Rosenwald, A., Dandekar, T. & Müller, T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* **24**, i223–i231 (2008).
72. Wu, M., Pastor-Pareja, J.C. & Xu, T. Interaction between Ras^{V12} and scribbled clones induces tumour growth and invasion. *Nature* **463**, 545–548 (2010).
This paper demonstrated the importance of cooperation between mutations in cancer in the RAS signaling pathway.
73. Berry, D.A. Adaptive clinical trials: the promise and the caution. *J. Clin. Oncol.* **29**, 606–609 (2011).
74. Green, R.C. *et al.* ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* **15**, 565–574 (2013).