Advanced Review

Computational approaches for the identification of cancer genes and pathways

Christos M. Dimitrakopoulos[1,2] and Niko Beerenwinkel[1,2]*

High-throughput DNA sequencing techniques enable large-scale measurement of somatic mutations in tumors. Cancer genomics research aims at identifying all cancer-related genes and solid interpretation of their contribution to cancer initiation and development. However, this venture is characterized by various challenges, such as the high number of neutral passenger mutations and the complexity of the biological networks affected by driver mutations. Based on biological pathway and network information, sophisticated computational methods have been developed to facilitate the detection of cancer driver mutations and pathways. They can be categorized into (1) methods using known pathways from public databases, (2) network-based methods, and (3) methods learning cancer pathways *de novo*. Methods in the first two categories use and integrate different types of data, such as biological pathways, protein interaction networks, and gene expression measurements. The third category consists of *de novo* methods that detect combinatorial patterns of somatic mutations across tumor samples, such as mutual exclusivity and co-occurrence. In this review, we discuss recent advances, current limitations, and future challenges of these approaches for detecting cancer genes and pathways. We also discuss the most important current resources of cancer-related genes. © 2016 The Authors. *WIREs Systems Biology and Medicine* published by Wiley Periodicals, Inc.

How to cite this article:
WIREs Syst Biol Med 2017, 9:e1364. doi: 10.1002/wsbm.1364

INTRODUCTION

Decades of cancer research have demonstrated that cancer is a complex genetic disease caused primarily by somatic mutations in the genome. Somatic mutations can dysregulate specific cellular pathways leading to the acquisition of cellular vulnerabilities that transform a normal cell into an abnormal cancer cell. These properties, often called cancer hallmarks,[1,2] drive cancer initiation and progression. Among others, they include evading growth suppressors, resisting cell death, and other abnormal phenotypes.

Many genetic changes in the genomes of somatic cells initiate and promote tumor growth, and cancer genomics researchers are now aiming at detecting all of these cancer driver mutations. There are several types of somatic mutations varying from single-nucleotide variants (SNVs), small insertions and deletions (indels), to larger copy number aberrations (CNAs, >50 bp)[3] and large genomic rearrangements or structural variants (SVs). Several of these genomic alterations have long been studied using low-throughput approaches, such as targeted gene sequencing, cytogenetic techniques,[4,5] systematic mutagenesis,[6] and genetic linkage analysis.[7] However, these traditional experimental approaches are laborious, time-consuming, and cost-inefficient.

*Correspondence to: niko.beerenwinkel@bsse.ethz.ch

[1] Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

[2] SIB Swiss Institute of Bioinformatics, Basel, Switzerland

Conflict of interest: The authors have declared no conflicts of interest for this article.

Recent high-throughput DNA sequencing techniques have revolutionized cancer genomics, and collaborative projects such as The Cancer Genome Atlas (TCGA)[8] and the International Cancer Genome Consortium (ICGC)[9] publicly released DNA sequences from thousands of tumors. Whole-exome sequencing only targets protein-coding regions (about 2% of the genome) and hence comes at reduced financial, storage, and computing costs for the analysis. This makes large population studies feasible. Whole-genome sequencing, on the other hand, allows examination of somatic mutations in the entire genome, enabling also the coverage of regulatory regions like promoters and enhancers. Both approaches contributed to massive cancer genome sequencing, which recently revealed a variety of mutational patterns such as kataegis,[10] chromothripsis,[11] chromosomal chains,[12] and other complex chromosomal rearrangements.[13] High-throughput sequencing technologies can produce billions of short reads, which need to be aligned to the reference genome in order to detect their genomic location and difference to the normal human genome. During the sequencing and alignment procedures, several errors and artifacts are introduced, including optical polymerase chain reaction (PCR) duplicates, and GC, strand, and alignment bias.[14] Several of the variant callers implemented to detect somatic mutations account for these biases trying to reduce false-positive and false-negative mutation calls. Hence, computational approaches for the systematic detection of cancer-related somatic mutations are increasingly important.

Several cancer mutations can be targeted therapeutically. For example, Sorafenib is used as a multikinase inhibitor primarily in kidney and liver cancer to target extracellular signal-regulated kinases (ERK) and other signaling pathways, which are important for tumor cell proliferation and angiogenesis.[15] Another example is Crizotinib, an oral tyrosine kinase inhibitor that targets anaplastic lymphoma kinase (ALK) in ALK-positive lung cancer patients.[16] The availability of mutation-specific drugs leads inevitably to the emerging field of precision medicine, in which treatment of a cancer patient is guided by his or her individual somatic mutation profile.[14]

The mutational landscape of cancer has been proposed to consist of ‘mountains’ of very frequently mutated genes and of ‘hills’ of significantly, but less frequently mutated genes.[17] Existing drugs mainly target the mountains that are present in a high number of patients with the same cancer type or even across several cancer types. At the same time, the hills are much higher in number and more cancer type-specific. Moreover, intertumor heterogeneity[18] (denoting the large differences in mutations occurring in tumors of the same type) makes it even more complicated to narrow down the mutated genes that drive cancer progression. All together, these issues result in the requirement of generating a large number of personalized treatment options by using specific drugs or combinations of drugs.

For most cancer genes, their involvement in the physiological process of cancer development is yet to be discovered.[19] In addition, we do not know how many cancer driver genes remain to be revealed. It is believed that a plateau is being reached, because in different tumor types, the same mutated cancer driver genes are increasingly rediscovered.[17] The catalogue of somatic mutations in cancer (COSMIC)[20] is currently reporting 572 genes for which mutations have been causally implicated in cancer. However, most variant callers tend to produce long lists of potential somatic mutations, many of which are neutral passenger mutations and only a few are selectively advantageous driver mutations. There is a wide variability of the average number of nonsynonymous mutations that occur in different cancer types, ranging from ~10 in pediatric cancers to several hundred in colorectal cancer. Only around 1–2.5% of these mutations are probably drivers.[17] Hence, additional methods have been developed in order to filter the mutation lists further for drivers. Some of these methods use protein structure and function or spatial clustering of mutations to predict the functional impact of mutations. So far, reviews are mainly focusing on variant callers,[14,21] methods to predict the functional impact of mutations,[22] and problems such as tumor purity estimation.[14]

In this review, we will focus on three different classes of approaches that use methods for predicting cancer genes and pathways in order to narrow down further the number of candidate drivers. The three categories comprise (1) methods using given biological pathways, (2) network-based methods, and (3) methods learning pathways *de novo* by detecting combinatorial patterns of cancer mutations (Figure 1). Methods that use given biological pathways compare a set of cancer-related genes (e.g., mutated genes) to known biological pathways. Network-based methods search for cancer genes and pathways in biological networks that represent the interactions between cellular molecules. Methods learning pathways *de novo* do not use any prior knowledge about the genes (pathways or interactions) and infer cancer genes and pathways based on patterns of co-occurrence or mutual exclusivity between the genetic aberrations. Before discussing each of them separately (Table 1), we summarize the available resources for storing cancer-related genes and drugs as well as important properties of the cancer-related genes such as mutation frequency, expression profiles, and cellular function. Finally, we will discuss current challenges in

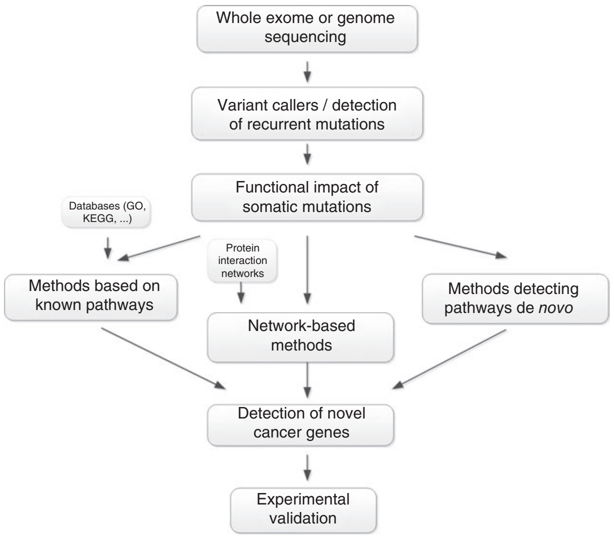
Whole exome or genome sequencing
Variant callers / detection of recurrent mutations
Functional impact of somatic mutations
Databases (GO, KEGG, ...)
Protein interaction networks
Methods based on known pathways
Network-based methods
Methods detecting pathways de *novo*
Detection of novel cancer genes
Experimental validation

FIGURE 1 | Detecting novel cancer genes begins with the sequencing of tumor samples (either whole exome or whole genome). The first step is to detect somatic mutations (single-nucleotide variants, indels, CNVs, structural variants, and gene fusions) from sequencing data using variant callers. The list of variant calls needs to be filtered to remove neutral passenger mutations and to detect candidate cancer driver genes. The simplest ways to perform such filtering is by detecting recurrent variants and by predicting the functional impact of each mutation. Then, methods that are more sophisticated come into play. They can broadly be categorized into three types: (1) methods that use preexisting pathways, (2) methods that are based on existing biological network data, and (3) methods predicting cancer pathways de novo based on their combinatorial patterns of occurrence in a group of tumors. Finally, the discoveries are validated experimentally.

the automated prediction and functional interpretation of cancer genes and pathways.

DATABASE SOURCES OF CANCER-RELATED GENES

Several databases exist that contain information about cancer genes and their function. COSMIC[20] currently comprises 572 genes in which mutations have been found and causally implicated in cancer. Among these genes, 484 harbor only somatic mutations, 35 only germline, and 53 harbor both. Vogelstein et al.[17] proposed a list of 54 oncogenes and 71 tumor suppressors by defining the 20/20 rule, which assumes that cancer genes must have either at least 20% inactivating mutations or at least 20% of the mutations must be in recurrent positions and missense. The Atlas of Genetics and Cytogenetics in Oncology and Haematology (AGCOH)[46] contains 1452 cancer genes based on merging the results of several collaborative projects. AGCOH also comprises a set of genes that have not been associated with cancer before and can be used as a negative control set of noncancer genes. The Network of Cancer Genes (NCG)[47] is a web-based repository of systems-level properties of cancer genes and collects information on 518 known (i.e., experimentally supported) and 1053 candidate (i.e., inferred using statistical methods) cancer genes. The Cancer3D database[48] is focusing on the impact of missense somatic mutations on protein structure and helps users analyze distribution patterns of mutations as well as their association with changes in drug activity. It displays mutations from over 14,700 proteins mapped to more than 24,300 protein structures from the Protein Data Bank.[49] However, it should be noted that databases such as COSMIC and NCG review the literature to predict cancer genes, which need further experimental validation. For example, only 120 of the 1053 candidate cancer genes in NCG are supported by cell line experiments that demonstrate the effect of gene silencing or gene overexpression.[47]

Other databases focus on therapeutic agents that may be used against specific cancer alterations. The Cancer Cell Line Encyclopedia[50] is a considerable resource for the systematic translation of SNVs, CNAs, and mRNA expression into therapeutic possibilities by generating genetic, lineage, and gene-expression-based predictors of drug response. Together with Genomics of Drug Sensitivity in Cancer, these databases can aid in determining the genetic factors that lead to the resistance and sensitivity to drugs. The Drug–Gene Interaction database (DGIdb) is the largest database with drug-related information. It combines 15 existing resources that contain information about disease genes, drugs, drug–gene interactions, and potential druggability.[51] Pharmacological data coupled with genomic data provided in these databases can be an important tool for clinicians in the process of accelerating the translation of novel cancer biology discoveries into treatments.

Although the data in the aforementioned databases become more and more accurate and comprehensive, the problem of cancer genome interpretation and respective treatment choice is far from being solved. In the future, we expect that the development of more accurate computational tools and experimental approaches will play an important role in linking cancer genotype and phenotype. To facilitate this process, effective policies and technologies for sharing cancer data are required.[8,9]

METHODS BASED ON KNOWN PATHWAYS

Variant callers[21] and tools predicting the functional impact of mutations[22] are focusing only on single genes, their mutations, and the functional impact in

TABLE 1 | Summary of Methods for Predicting and Interpreting Cancer Genes

Category	Methods	Description
Known pathways	DAVID,[23] FaTiGO,[24] GoSTAT[25]	These methods use a statistical test to assess the significance of the overlap between a gene set and a known pathway.
	GSEA[26]	Tests if the expression levels of the genes in the gene set correlate with a specific phenotype.
	Grossman et al.[27]	Gene Ontology (GO) enrichment analysis by taking into account the GO tree hierarchy. Determines overrepresentation of terms in the context of annotations to the term's parents.
	PathScan[28]	Tests if the mutations of different cancer patients exhibit enrichments in the same pathways.
Network-based	NetBox[29]	Detects network modules in a given list of input genes and accesses the statistical significance of their modularity.
	DriverNet[30]	Identifies driver mutations by their effect on mRNA expression networks.
	Torkamani and Schork[31]	Identifies functionally related gene modules targeted by somatic mutations by reconstructing regulatory interactions.
	NBS[32]	Uses network diffusion to stratify patients based on the observation that their aberrations lie in similar network regions.
	HotNet2[33]	Uses insulated network diffusion to detect mutated subnetworks with statistically significant size. Captures the directionality of interactions.
	TieDIE[34]	Uses network diffusion to link somatic mutations to transcriptional changes.
De novo	RME[35]	Detects gene modules whose members are recurrently mutated and exhibit mutually exclusive patterns.
	Dendrix[36]	Detects driver pathways characterized by high exclusivity and high sample coverage. Requires high coverage of the discovered gene modules instead of each gene separately.
	Multi-Dendrix[37]	Identifies multiple mutually exclusive sets of genes in parallel.
	CoMEt[38]	Identifies multiple sets of mutually exclusive genetic alterations from different subtypes of the same cancer type.
	TiMEx[39]	Models the interplay between the waiting times to alterations and the observation time. Highly sensitive to mutually exclusive occurring low-frequency driver alterations.
	pathTiMEx[40]	Takes into account the evolutionary order constraints among pathways to detect mutually exclusive cancer alterations.
	Sakoparnig et al.[41]	Identifies low-frequency genomic alterations based on mutational dependencies.
	muex[42]	Models the generative process of mutually exclusive patterns in the presence of noise.
Combined	MEMo[43]	Detects network cliques of aberrant genes with mutually exclusive patterns.
	Mutex[44]	Identifies mutually exclusive groups of genes with a common effect on a given signaling network.
	MEMCover[45]	Detects mutually exclusive groups of mutated genes in the same or across different tissues.

their gene products. However, genes do not work isolated but they interact through complex cellular reactions whose normal dynamics are altered in cancer. Based on these interactions, they are organized in groups, often called pathways.

Early pathway-based approaches interpret somatic mutations by comparing them to known pathways from public databases. They compute the overlap between a list of mutated genes and sets of genes with known functional annotations and assess its chance of occurring randomly by using statistical measures such as Fischer's exact test or the hypergeometric test. If the probability of the observed overlap is adequately small under the random null model,

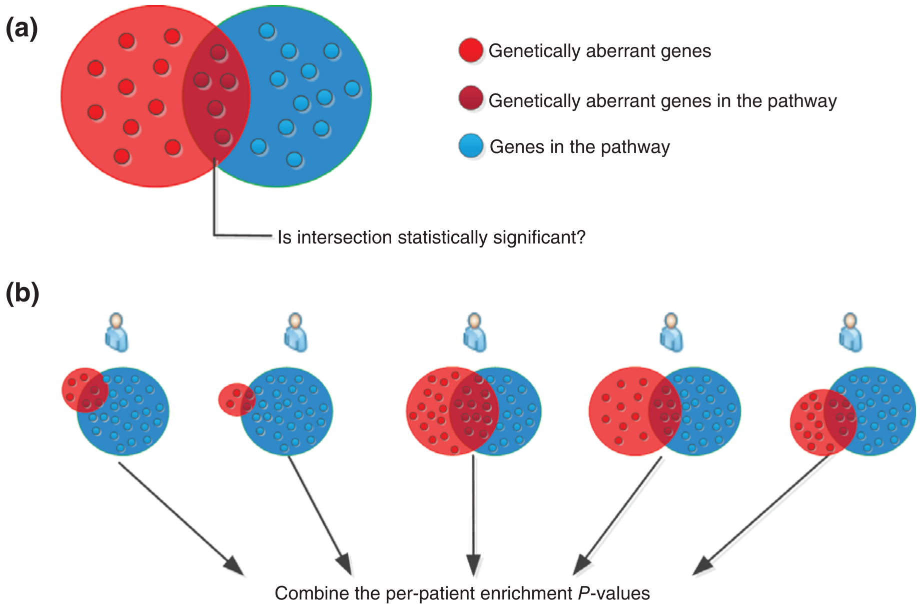
(a)
Genetically aberrant genes
Genetically aberrant genes in the pathway
Genes in the pathway
Is intersection statistically significant?
(b)
Combine the per-patient enrichment *P*-values

FIGURE 2 | Methods that are based on known pathways to identify cancer drivers. (a) Most methods statistically assess the significance of the overlap between a user-defined gene set (red-colored nodes) and a known pathway (blue-colored nodes).[23–25] The user-defined gene set is usually the result of an experiment (e.g., a list of genetically aberrant genes). Dark red-colored nodes correspond to the genes that belong both to the known pathway and to the user-defined gene set. (b) Other methods compute per-patient enrichment scores[28] for a known pathway in the same way, which are then combined into an overall score.

then the list is considered to be enriched for the respective function of the annotated gene set or pathway (Figure 2). Predefined functional annotations can be found in public databases such as KEGG,[52] Gene Ontology (GO),[53] MSigDM,[26] and Reactome.[54] There are also tools that integrate different databases and methodologies to perform this type of analysis such as DAVID,[23] FaTiGO,[24] and GoStat.[25]

A different, widely used approach, called gene set enrichment analysis (GSEA),[26] determines whether members of a certain gene set tend to occur toward the top (or bottom) of the given gene list and in which case the correlation of the genome-wide expression profiles of the genes in the list with a phenotypic class distinction is tested. Here, the phenotypic class distinction is defined by prior biological knowledge, for example, known biochemical pathways or coexpression experiments. Except for gene expression measurements, GSEA can be used with other types of data, such as mutation frequency.[55]

Gene Ontology[53] is the database most widely used in enrichment analysis. It consists of three hierarchically structured ontologies that describe gene products in terms of their associated biological processes, cellular components, and molecular functions (GO terms). The straightforward approach to measure enrichment of GO terms by considering them independently cannot account for the hierarchical structure of GO. Grossman et al.[27] presented an approach for GO enrichment analysis that determines overrepresentation of terms in the context of annotations to the term's parents. Goeman and Mansmann[56] proposed a multiple testing method that preserves the GO graph structure. It requires a user to choose a focus level in the GO graph, which reflects the level of specificity of terms in which the user is most interested. Similar approaches have been developed in Refs 57,58.

Often the sets of mutated genes and the predefined functional pathways have a very general biological characterization or contain a large number of genes. To overcome these limitations, per-patient enrichment analysis methods attempt to detect enriched pathways across all patients, by testing patient-related gene sets (e.g., mutated genes) versus known pathways (Figure 2). PathScan[28] is one such approach, which tests if the mutations of different cancer patients exhibit enrichments in the same pathways. PathScan accounts for variations in gene length and differentiates frequently mutated genes from genes having only a few mutations. Similarly, Boca et al.[59] computed per-patient gene set enrichment scores, which were merged to an overall ranking score. Per-patient gene enrichment analysis has proved to be more interpretable and statistically powerful compared with standard GSEA.

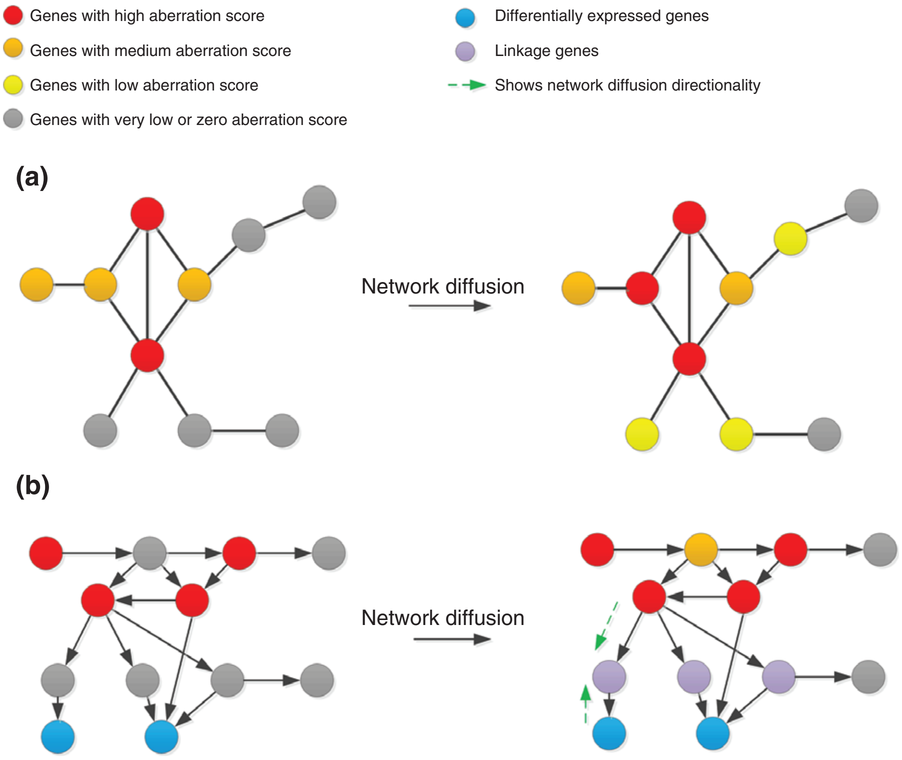
Genes with high aberration score
Genes with medium aberration score
Genes with low aberration score
Genes with very low or zero aberration score
Differentially expressed genes
Linkage genes
Shows network diffusion directionality
(a)
Network diffusion
(b)
Network diffusion

FIGURE 3 | Network-based methods to identify cancer genes. (a) Network-based methods like HotNet2[33] detect cancer driver genes as strongly connected components of aberrant genes in the network. Network diffusion is used to estimate how strongly the aberrant genes are connected in the network. Nodes are initialized with an aberration score that corresponds to the proportion of samples that contain a single-nucleotide variant or copy number aberration in the gene. Using network diffusion the aberration scores are spread in the network until an equilibrium state is reached, where there are no more significant changes in the scores during time. Nodes correspond to genes and edges to gene interactions. Colors correspond to the amount of aberration score that is concentrated at the node before and after network diffusion (red: high, orange: medium, yellow: low). Gray nodes correspond to very low or zero aberration score. (b) TieDIE[34] also uses network diffusion to capture the genes in the network that link genetically aberrant genes to differentially expressed genes at the transcription level, the so-called linker genes. Blue colored nodes correspond to differentially expressed genes. The linker genes are represented with purple color and during network diffusion, they receive flowing aberration scores from both genetically aberrant and differentially expressed genes.

Although pathway-based approaches can be used to evaluate and interpret gene lists, they cannot be used to predict novel cancer pathways because they are based on predefined pathways. Moreover, they ignore crosstalk between different pathways by considering them as distinct groups. Finally, each gene is not equally important for a pathway and the topology of the interactions, which can capture the dependencies between genes of a specific pathway, is usually not taken into account by pathway-based approaches.

NETWORK-BASED METHODS

In contrast to pathway-based approaches that analyze pathways with already well-established functions, network-based approaches use interaction networks to infer novel cancer genes and pathways (Figure 3).

Protein networks can be either undirected (physical protein–protein interactions) or directed (high-level functional interactions). Although most of the current approaches use undirected networks, the use of directed networks is of high importance as they can explicitly reveal the different types of interactions that lead to cancer progression. Reactome[54] contains a human protein functional interaction network[60] based on integrating expert-curated pathways, gene coexpression, protein domain interactions, and other sources. The interactions of the protein interaction networks can be either experimentally verified, which is considered more reliable, or predicted by computational methods. HPRD[61] and BioGRID[62] are examples of databases that contain experimentally verified interactions. Other databases, like STRING[63] for example, contain both experimentally verified and computationally predicted interactions. KEGG[52] and Reactome[54] are considered the most reliable sources of experimentally verified metabolic reactions. iRefWeb[64] has the highest coverage of the human interactome, that is, the complete human protein interaction network, because it integrates interaction data from 10 public databases. It is common to use several networks of different origins independently to assess their effect on the final results of a network-based analysis.

Biological networks have been used in many cancer-specific studies[29,65–68] and in order to detect patterns across several cancer types.[30,68–70] NetBox[29] was one of the first network-based approaches that attempted to distinguish driver from passenger mutations in glioblastoma. It maps genetically aberrant genes [SNVs or copy number variations (CNVs)] to a protein interaction network and extracts only the part of the network that includes the aberrant genes. The identified network is partitioned into network modules using the Newman–Girvan algorithm. It was found that, although being dissimilar between patients, the glioblastoma genetic alterations tend to occur within the same network modules. Similarly, several network-based methods measure the impact of genetic variations on transcriptional networks.[30,68,69,71] For example, DriverNet[30] identifies driver mutations by their effect on mRNA expression networks. Some methods use reverse engineering approaches to reconstruct networks of cancer-related interactions,[31,72] while other methods discover altered protein sets in protein interaction networks.[32,33,73] The idea of the former approaches is that they reconstruct network interactions (e.g., regulatory networks) *de novo* by processing the expression levels of genes. For example, ARACNE[72] defines an interaction between two genes by computing the mutual information between their expression profiles. These dependencies have been shown to be useful in identifying direct regulatory interactions. Torkamani and Schork[31] used ARACNE to identify functionally related gene modules targeted by somatic mutations in various cancer types.

Several methods that integrate somatic mutations with interaction networks in order to detect groups of interacting mutated genes have been proposed. The idea behind these methods is that the mutations patients suffering from the same cancer type are divergent in the genes they hit, but the affected genes participate in the same biological processes represented by densely connected subgraphs. Hofree et al.[32] built a patient stratification method that clusters patients with mutations in similar network regions. Using their network-based stratification (NBS) method, the authors identified cancer subtypes that are predictive of clinical outcomes such as patient survival and response to therapy. NBS is less accurate when clustering without network information, demonstrating the importance of biological networks in the procedure. Similarly, Leiserson et al. and Vandin et al.[33,73] performed a TCGA pan-cancer analysis of 12 cancer types by using three different interaction networks. Their method, HotNet2, uses insulated network diffusion to detect strongly connected components of aberrant genes, followed by a statistical test to determine the significance of the number and size of the subnetworks (Figure 3). HotNet2 captures the directionality of interactions and effectively detects rare driver mutations.

Another method that uses graph diffusion to detect cancer genes is TieDIE.[34] TieDIE is using a directed graph diffusion from two different sources, which include genetic aberrations and transcriptional changes. While scores from genetic aberrations are diffused along the directions of the interactions of the network, the scores of the differentially expressed genes are diffused in the opposite direction. Hence, TieDIE can detect linker genes that connect genetic aberrations to transcriptional changes and sheds light on the way that somatic mutations affect the expression levels of other genes that are not genetically altered.

Because of the complex structure of intracellular pathways, drugs targeting specific molecules often fail owing to acquired drug resistance caused by mutations or other molecular alterations. Drug resistance enhances the need for alternative cancer therapies, such as those based on combinations of drugs. DrugComboRanker[74] is the first computational method to predict combinations of already approved drugs *de novo*. It aims to repurpose drugs by combining high-throughput cancer data with a set of approved drugs for several diseases.[75] DrugComboRanker constructs a drug similarity network based on the genomic profiles of the genes they target. Next, it uses a Bayesian nonnegative matrix factorization approach to partition the drug network into drug communities. Given the observation that drugs in the same community share common functionalities, they build a drug recommendation system.

Cheng et al.[76] exploited the fact that kinases are often drug targets and created a human kinome interaction map by merging kinase–substrate, protein–protein, and kinase–drug interactions. Their approach is a useful resource for combining and designing kinase inhibitor drugs. They found that drug-targeted kinases are significantly enriched as central hubs in the human protein interactome. Inhibiting a hub node affects also many other molecules, and it is consequently more likely to lead to side effects. Moreover, the authors suggested that targeting hub kinases could more easily provide the adaptive crosstalk or feedback within cellular networks that leads to drug resistance. Mitsopoulos et al.[77] revealed distinct patterns in the local network topology of drug targets (higher degree than average) and suggested how drugs from other therapeutic areas could be used to create successful drug combinations that overcome drug resistance. Their approach is the first to propose potential drug targets as network neighborhoods of genes and not individual genes. In the future, network models in systems pharmacology,[78] a promising emerging field, are expected to assist in decoding the complex drug–target interactions.

The network-based methods mentioned so far are using large biological networks that integrate different types of interactions and are therefore analyzing the interactome from a global perspective. In this way, the function of a gene is considered constant between different tissues. However, in order to interpret the genetic basis of tumors in their environment we need to use tissue-specific networks that are able to capture the dynamic nature of gene functions. A significant effort toward this direction is the GIANT webserver,[79] which integrates tissue-specific data of various genome-scale experiments using a naïve Bayes classifier in order to predict functional interaction networks for 144 human tissues. It provides NetWAS, a network-based approach that reprioritizes genome-wide association study (GWAS) *P* values by using the tissue-specific functional networks and performs better in predicting gene–disease associations than using the GWAS *P* values alone. A GWAS *P* value (<0.01) indicates whether the occurrence of a single nucleotide polymorphism in the tumor samples of a specific cancer type is significantly high.

Biological networks are the tortuous wiring diagrams of the cell, the disruption of which can lead to cancer-specific phenotypes. Hence, complete understanding of cancer requires complete understanding of biological networks. In addition to the current incompleteness of the human interactome, network-based methods usually do not account for the gene expression changes that a mutation can cause in other genes.

COMBINATORIAL PATTERNS OF CANCER MUTATIONS

Besides network analysis, the detection of combinatorial patterns among mutational events is another promising approach to detect cancer genes and pathways.[80] Here, we discuss two different types of combinatorial patterns, namely mutually exclusive and co-occurring mutations. The idea behind the detection of mutually exclusive mutations is that once a gene involved in a biological process is mutated, the tumor cell acquires a selective advantage, for example, increased proliferation, which promotes clonal expansion. A second hit in another gene of the same process is then much less likely to occur, because it does not confer an additional selective advantage to the cell.[39] This phenomenon results in strong negative correlation among mutations in genes of the same pathway. By contrast, co-occurring mutations provide evidence for positively correlated gene mutations. Here, mutations in two or more genes need to be present simultaneously for the cell to acquire the selective advantage (Figure 4).

The Recurrent and Mutually Exclusive (RME)[35] algorithm detects gene modules whose members are recurrently mutated and exhibit mutually exclusive patterns. It considers only genes that are mutated with frequency ≥10% and therefore misses rare driver mutations. Moreover, mutually exclusive events among rare mutations are more likely to happen by chance and are therefore more difficult to detect. MEMo[43] uses a statistical permutation test that permutes the mutated genes between the samples in order to detect mutually exclusive events between genetic aberrations. The permutation test is performed on groups of genes detected as cliques from a protein interaction network. Hence, MEMo integrates different data sources and combines different methodologies (Tables 1 and 2). Methods that address the limitation due to rare mutations are Dendrix (*De novo* Driver Exclusivity)[36] and Multi-Dendrix.[37] Dendrix identifies driver pathways by their high exclusivity and high sample coverage. Unlike RME, Dendrix requires high coverage of the discovered gene modules instead of each gene separately. Multi-Dendrix simultaneously identifies multiple mutually exclusive sets of genes by an integer linear programming approach. It is much faster than Dendrix on genome-scale data and identified mutually exclusive mutations in well-studied cancer pathways such as p53 and PI3K/AKT signaling.

Two recent methods for the detection of mutually exclusive genomic events include Mutex[44] and CoMEt.[38] Mutex detects groups of genes with a common downstream effect on a signaling network, by introducing a mutual exclusivity criterion that avoids large imbalances in the contribution of each gene to the overall mutual exclusivity pattern. To achieve this, every gene is tested for mutual exclusivity against the union of the rest of the group

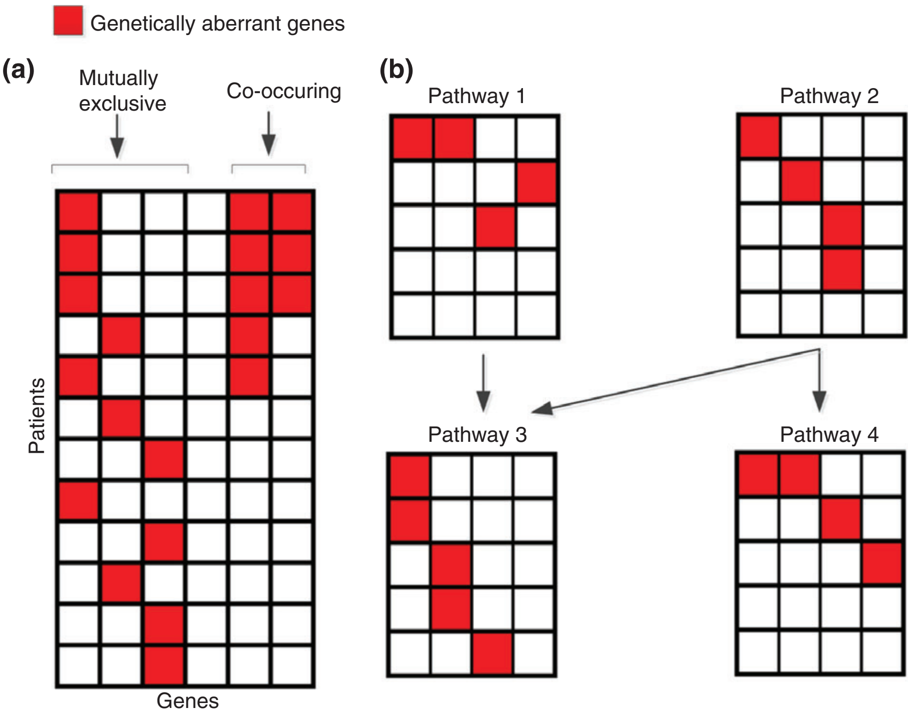

FIGURE 4 | Methods that identify cancer genes de novo by detecting combinatorial patterns of cancer mutations across patients. Red squares correspond to genetically aberrant genes. The rows of the depicted matrix correspond to different patients and the columns to different genes. (a) Mutual exclusivity and co-occurrence, as depicted in the figure, are two combinatorial patterns of mutations and the statistical significance of which can aid in detecting cancer genes de novo. Methods attempt to detect these patterns in the presence of noise; hence, the patterns detected are not perfectly co-occurring or mutually exclusive. (b) pathTiMEx[40] predicts cancer progression at the level of pathways by introducing a probabilistic waiting time model for mutually exclusive cancer alterations. It explicitly accounts for the evolutionary order constraints among pathways, which would otherwise confound the detection of mutually exclusive gene groups (directed arrows).

alterations. CoMEt performs an exact statistical test conditioning on the frequency of each alteration and can therefore more effectively detect rare mutations. It can detect in parallel multiple sets of mutually exclusive alterations, which can overlap, vary in size, and belong to different cancer subtypes. Mutex and CoMEt have been shown to exhibit an improved performance in the prediction of mutually exclusive events compared to Dendrix, Multi-Dendrix, MEMo, and RME.

A unique approach for detecting mutually exclusive patterns of genetic alterations either within the same tissue or across different tissue types is MEMCover.[45] MEMCover uses a random permutation test (similar to MEMo) systematically to detect mutually exclusive patterns that occur in one tissue type, in many tissue types, or between tissue-specific drivers. It then uses interaction data to discover dysregulated pathways that are present across many cancer types. When compared to HotNet2, MEMCover was able to detect a higher number of known cancer genes. Although MEMo, Mutex, and MEMCover are primarily detecting mutually exclusive groups of genetic aberrations, they also use interaction data in their analysis pipeline. Hence, they differ from the other methods covered in this section (Tables 1 and 2).

Statistical models of mutual exclusivity have been devised in order to handle the ubiquitous noise that is present in cancer mutation data as well as deviation from perfect mutual exclusivity (referred to as impurity). Szczurek and Beerenwinkel[42] developed an approach that models the generative process of mutually exclusive patterns and takes into account the presence of noise in the form of false positives and false negatives. TiMEx,[39] another approach for the detection of mutually exclusive events, uses a generative probabilistic model for tumorigenesis that explicitly accounts for the temporal interplay between the waiting times to the occurrence of alterations and the observation time. It permits direct estimation of the probability of mutual exclusivity and is highly sensitive to low-frequency mutually exclusive alterations.

Although mutual exclusivity is a frequently observed pattern in cancer mutations, it is not the

TABLE 2 | Description of the Output of Each Method

Method	Driver Genes or Mutations	Groups of Mutually Exclusive Genetic Alterations	Network Modules
NetBox[29]			√
DriverNet[30]	√		
Torkamani and Schork[31]	√		√
NBS[32]			√
HotNet2[33]			√
TieDIE[34]			√
RME[35]		√	
Dendrix[36]		√	
Multi-Dendrix[37]		√	
CoMEt[38]		√	
TiMEx[39]		√	
pathTiMEx[40]		√	
Sakoparnig et al.[41]	√		
muex[42]		√	
MEMo[43]		√	√
Mutex[44]		√	√
MEMCover[45]		√	√

Only network-based methods and methods based on combinatorial patterns of mutations are included.

only one. There are also patterns of co-occurring mutations in cancer. For example, KRAS and TP53 mutations have been shown to co-occur in lung, pancreas, and large intestine cancers. There are also observations demonstrating contrasting mechanisms in different cancers. For instance, mutations of genes in the Ras and Wnt pathways tend to co-occur in the large intestine but are mutually exclusive in cancers of the pancreas.[80] In contrast to mutually exclusive mutations, two co-occurring mutations can be members of different pathways.[81] In recent years, synthetic lethality has emerged as an attractive therapeutic strategy against cancer.[82] Two genes are synthetically lethal if a combination of mutations in both of these genes, but not in any single gene, leads to cell death. It has been observed that pairs of genes that are frequently aberrant and mutually exclusive in cancer often constitute synthetically lethal genes. Intuitively, both genes are performing essential functions and the cell acquires cancer-related properties by hitting one of them. Mutations in both are not observed, as this event would lead to cell death. Hence, pairs of genes characterized by synthetic lethality can be selectively targeted in order to kill cancer cells that harbor already one mutated gene.

Another approach to identify low-frequency genomic alterations based on mutational dependencies was proposed in Ref 41. This approach is based on the observation that the cumulative probabilities of neutral mutations increase linearly with the total number of mutations in a tumor. By contrast, driver mutations that depend on other mutations occur with probabilities displaying a nonlinear pattern of increase. The idea behind this approach is that the occurrence of drivers is subject to hidden constraints (dependencies between driver mutations), whereas the passenger mutations are independent and selectively neutral. For example, mutation of KRAS tends to occur after mutation or loss of fibroblast activation protein alpha (FAP) in colorectal tumorigenesis.[83] To distinguish independent from dependent mutations, their rate of occurrence is estimated among tumors. Independent mutations tend to occur at a constant rate among tumors, and driver–passenger discrimination is based on detecting deviation from this behavior.

Cancer progression can be regarded as the accumulation of mutations in different genes. However, more robust models can be derived by considering dependencies among altered pathways, mainly because of the large interpatient gene-wise mutational heterogeneity.[78,84] Cristea et al.[40] devised pathTiMEx, an extension TiMEx[39] that introduces a probabilistic waiting time model for mutually exclusive cancer alterations on the level of pathways rather than on the level of genes. The idea is to explicitly account for the evolutionary order constraints among pathways, which would otherwise confound the detection of mutually exclusive gene groups (Figure 4). Raphael and Vandin[85] propose a combinatorial model, corresponding to the specialized situation when the dependency structure among pathways is restricted to be linear.

Combinatorial mutational patterns can suggest candidate cancer genes and pathways in an unbiased fashion without using any prior knowledge. Furthermore, the type of combination can give insight into the functional relationship of the genes.

FUTURE DIRECTIONS AND CHALLENGES

Data Integration and Combination of Methodologies

The apposition of several data sources through data integration allows for improved prediction and interpretation of cancer genome data. For example, MEMo, MEMCover, and Mutex are pipelines that integrate different data sources and combine various methodologies. Specifically, they perform both network and mutual exclusivity analyses (Tables 1 and 2). MEMo detects cliques of genetically aberrant genes in a protein–protein interaction network and then assesses their tendency of being mutually exclusive. On the other hand, MEMCover detects mutually exclusive patterns of mutations that are present across many tissue types in its first step and subsequently uses interaction data to assess the chance of the detected mutually exclusive groups to be new pan-cancer dysregulated subnetworks. The combination of different methodologies is particularly useful to produce interpretable lists of candidate cancer genes or gene clusters. For example, the genes in Memo-derived modules are both interacting and exhibit mutually exclusive patterns of mutations, facilitating their biological interpretation and design of validation experiments. Another example of data integration is the integration of different types of somatic mutations (SNVs, CNVs, SVs, etc.) which all should ideally be taken into account in order to cover the widest range of driver mutations.[32,33,35] There are also methods that integrate different types of sequencing[86] and omics data,[34,87] but more work is required to fully integrate the various types of data such as DNA methylation and chromatin modifications. The type of output of two different methods can also vary greatly (Table 2). For example, DriverNet[30] outputs a ranked list of driver mutations, whereas muex[42] outputs groups of genes characterized by mutually exclusive mutations.

Other studies explored the dependencies between different data sources. Reimand et al.[88] studied the dependency between the cancer genome and protein phosphosites. By performing a pan-cancer analysis, they found that phosphorylation-related SNVs (pSNVs) occur in ~90% of tumors, show increased conservation and functional mutation impact compared with other protein-coding mutations, and are enriched in cancer genes and pathways. Jacobsen et al.[89] studied the dependency between miRNA–mRNA networks and genomic mutations resulting in a method that prioritizes cancer-related miRNA–target interactions across 11 TCGA cancer types. As expected, miRNAs with recurrent target relationships were frequently regulated by genetic and epigenetic alterations. Akavia et al.[90] integrated CNAs with gene expression data in a Bayesian framework to detect driver genes and pathways in melanoma. They detected cases where a CNA is correlated with the expression of a group of genes forming a gene module, assuming that the copy number produces changes in the expression level of the gene it affects as well as in the expression levels of the genes in the module.

Overall, the integration of various types of cancer-related data, such as interaction networks, mRNA expression, phosphoproteome abundance, genetic aberrations, and microRNAs, can shed light on the underlying molecular mechanisms of cancer.

Cancer Subtypes and Pan-Cancer Analysis

Efficient computational methods are needed to detect the different subtypes of specific cancers.[32] The classification of a patient in predefined subtypes based on his or her individual tumor molecular profiles facilitates the detection of cancer drivers, as mixing data from different cancer subtypes can complicate the identification of cancer drivers when studying a specific cancer type. However, several TCGA pan-cancer analyses[91,92] have demonstrated that processing data from different cancers at the same time can improve the prediction of certain subtypes. The TCGA pan-cancer initiative[8] is a significant effort to compare different cancer types and measure similarities across them. By detecting similar patterns in the genomic profiles of different cancers, we can make hypotheses for experimental evaluation about how cancer therapies that are successful in one group of patients may be applied to other groups. Several interesting bioinformatics approaches for pan-cancer analysis have recently been proposed.[89,93–97] Patient stratification on pan-cancer data can reveal tissue-specific or tissue-independent molecular subtypes of cancer. On the other hand, studying properties of cancer genes across all samples of the same cancer type (e.g., recurrent mutations) can be inefficient. For example, a mutation may be detected as significantly recurrent in a subset of samples, but insignificantly recurrent over all samples.

Bias Toward Prior Knowledge

One important bias in the methods that predict cancer genes is the direct or indirect incorporation of prior knowledge. In general, we need methods that do not depend on prior knowledge but are focusing on predicting cancer drivers *de novo*. Pathway-based approaches suffer more from this problem compared with network-based approaches. The latter use the entire knowledge about molecular interactions, whereas pathway-based approaches are based on specific prespecified sets of genes. Pathway-based approaches use reference databases, which contain several pathways that have been extensively studied in cancer such as Wnt and MAPK signaling. As a result, these pathways will always show up in the top of the list of enriched pathways. Although less than pathway-based approaches, network-based approaches also suffer from certain biases. For instance, many proteins or genes have been extensively studied and hence have a higher number of connections in the protein networks. Hub genes or genes associated with well-studied cancer pathways are more likely to be correctly detected by network-based methods. Although sensitivity toward well-studied cancer genes is a benchmark for the method's performance, rigorous correction for their increased connectivity can greatly aid in the discovery of novel targets.

Several techniques for assessing the robustness of cancer gene and pathway prediction have been proposed. Ciriello et al.[98] assessed the robustness of different classes of genetic aberrations in stratifying patients by removing different percentages of samples and reclassifying the reduced datasets. To analyze the sensitivity of their variant prioritization method with regard to *a priori* gene or disease association biases, Sifrim et al.[99] stratified the positive testing set of disease-causing variants by year of publication (2000–2012) while training the model only on data published before 2000. For network-based approaches, Brohée and van Helden[100] assessed the robustness of methods for the computational prediction of protein complexes from protein interaction networks by gradually adding noise (insert random edges) to the network or removing parts of it (randomly delete existing edges). It is also common to compare the performance with a certain network to that obtained from randomizing the network, for example, by randomly reassigning edges while keeping the degree distribution of the original network fixed. By contrast, methods detecting combinatorial patterns of mutations learn pathways *de novo* and do not suffer from bias toward prior knowledge of interactions or pathways.[36] Handling noise in the detection of mutually exclusive events[39,42] improves the sensitivity of these methods. Their robustness is typically addressed in the framework of the underlying statistical model by assessing the uncertainty of parameter estimates and model stability.

Another complication is that proteins are involved in multiple cellular functions. As a result, they can also participate in several stages of cancer development and progression. Hence, pathways should be regarded as dynamic structures because they have no clear boundary and there is considerable crosstalk between them. This phenomenon has been partially covered by the use of protein networks, which describe the interactome as one entity and not as a collection of distinct parts. However, methods that are taking into account the type of interactions (e.g., activating or inhibiting) and do not consider only undirected physical protein–protein interaction networks are of high importance[33,34,101] because the way cancer genes interact can shed light on the molecular mechanistic basics of cancer.

Finally, network-based approaches usually use protein–protein interaction networks and consequently they are restricted to mutations that affect protein-coding regions of the genome. Pathway-based approaches are similarly limited by the characterization of the pathway members, which are usually gene products such as proteins or posttranslational modified proteins. By contrast, methods predicting cancer pathways *de novo* by using the combinatorial patterns of mutations are not limited by known pathways or networks and can therefore handle mutations in nonprotein-coding regions of the genome, such as mutations in intergenic or splice-site regions.

Drawbacks of Protein Interaction Networks

Protein interaction networks are suffering from false negatives (missing interactions) and false positives (false interactions).[102] Indeed, network biology is still far from completing the human interactome. There are around 130,000 estimated interactions among human proteins, most of which remain to be discovered.[103] Consequently, several methods have been designed to assess the quality of currently available human interactome maps.[104] For instance, ClusterONE,[105] a method for detecting protein complexes by clustering protein interaction networks, penalizes for false negatives and achieves a higher accuracy than other relevant methods. Moreover, experimental techniques, such as yeast-two-hybrid (Y2H), and computational techniques for predicting protein interactions suffer from high rates of false positives.[106] In the future, much computational effort is needed to complete the human interactome and increase its confidence.[107–109]

In this review, we have discussed methods for detecting genes that drive cancer progression either alone or in collaboration with other genes. The network-based approaches and the approaches detecting combinatorial patterns of mutations detect groups of interconnected genes but without considering the series of changes triggered by these mutations. Gene expression changes due to genetic alterations can cause further changes in the expression levels of other genes via gene interactions. Ruffalo et al.[101] proposed a method to investigate cancer genes that work in cooperation with mutated or differentially expressed genes, but are not aberrant themselves, neither on the DNA nor on the RNA level. These genes were detected by measuring their network proximity to mutated or differentially expressed genes. TieDIE is another approach[34] going beyond networks of mutated genes. It uses network diffusion to link somatic mutations to transcriptional changes (Figure 3). In the future, the development of similar approaches will help understanding somatic mutations as dynamically changing molecular states in the cell that drive cancer progression.

Experimental Validation

Although the results of computational methods are important to understand cancer, their integration with experimental approaches is always necessary to produce valid and interpretable outcomes. Creixell et al.[110] created ReKINect, a method that attempts to explain systematically how somatic mutations perturb signaling networks. Somatic mutations can create new phosphorylation sites, destroy existing ones, or rewire interactions between kinases and substrates. A substrate mutation may rewire the interaction with its upstream kinase and a kinase mutation may rewire the interaction with its downstream substrate. ReKINect combines quantitative proteomic and phosphoproteomic mass spectrometry with exome sequencing data to identify direct proteomic evidence of the destruction of phosphorylation sites. In another study, Akavia et al. knocked down genes they detected by CONEXIC[90] using shRNAs. By knocking down these genes, they observed changes in cancer-related cell properties (e.g., proliferation and growth).

In general, the results of computational methods is the first step to create compact lists of candidate cancer genes, which is then cost-effective and time-efficient to validate experimentally. Given that findings of computational methods always need supporting experimental evidence, both experimental and computational techniques are important for detecting and understanding cancer pathways. Except for postcomputational validation, experimental techniques are frequently used to generate the input data for computational methods. For example, Y2H experiments can construct a protein interaction network,[111] knockout siRNA experiments delineate pathway-specific interactions,[112] and mass spectrometry experiments determine abundance measurements for protein expression.[113]

CONCLUSION

Cancer genomics is still in its infancy. The mutational landscape of primary and metastatic tumors has yet to be completed. The unprecedented generation of large amounts of cancer genetic and epigenetic molecular data holds the promise for better prediction of novel cancer genes and more comprehensive reasoning on how they are linked to complex tumor traits like metastasis formation and tissue invasion. In this review, we have described state-of-the-art approaches for the prediction and characterization of cancer pathways, and we have speculated on future strategies to construct better performing methods. As the number of available tumor samples is increasing, we expect that these methods will greatly help in interpreting the complex heterogeneity of tumors, improve our ability to distinguish driver from passenger mutations, and delineate the pathways that are dysregulated in cancer and can be targeted by drugs. Moreover, single-cell sequencing data are expected to provide a new source of valuable information for decoding tumor heterogeneity.

ACKNOWLEDGMENTS

This work was supported by the SIB PhD Fellowship programme and in part by ERC Synergy Grant 609883.

REFERENCES

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011, 144:646–674.
2. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000, 100:57–70.
3. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. *Nat Rev Genet* 2015, 16:172–183.
4. Clark J, Edwards S, Feber A, Flohr P, John M, Giddings I, Crossland S, Stratton MR, Wooster R, Campbell C, et al. Genome-wide screening for complete genetic loss in prostate cancer by comparative hybridization onto cDNA microarrays. *Oncogene* 2003, 22:1247–1252.
5. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009, 458:719–724.
6. Touw IP, Erkeland SJ. Retroviral insertion mutagenesis in mice as a comparative oncogenomics tool to identify disease genes in human leukemia. *Mol Ther* 2007, 15:13–19.
7. Friedman LS, Ostermeyer EA, Szabo CI, Dowd P, Lynch ED, Rowell SE, King MC. Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. *Nat Genet* 1994, 8:399–404.
8. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 2013, 45:1113–1120.
9. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* 2010, 464:993–998.
10. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* 2012, 149:979–993.
11. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 2011, 144:27–40.
12. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, et al. The genomic complexity of primary human prostate cancer. *Nature* 2011, 470:214–220.
13. Holland AJ, Cleveland DW. Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. *Nat Med* 2012, 18:1630–1638.
14. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med* 2014, 6:1–17.
15. Pinter M, Sieghart W, Graziadei I, Vogel W, Maieron A, Koeningsberg R, Weissmann A, Kornek G, Plank C, Peck-Radosavljevic M. Sorafenib in unresectable hepatocellular carcinoma from mild to advanced stage liver cirrhosis. *Oncologist* 2009, 14:70–76.
16. Shaw AT, Kim D, Nakagawa K, Seto T, Crinó L, Ahn M, De Pas T, Besse B, Solomon BJ, Blackhall F, et al. Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *N Engl J Med* 2013, 368:2385–2394.
17. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. *Science* 2013, 339:1546–1558.
18. Cusnir M, Cavalcante L. Inter-tumor heterogeneity. *Hum Vaccin Immunother* 2012, 8:1143–1145.
19. Garraway LA, Lander ES. Lessons from the cancer genome. *Cell* 2013, 153:17–37.
20. Forbes SA, Bharma G, Bamford S, Dawson S, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR. COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res* 2010, 38:D652–D657.
21. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* 2014, 15:256–278. doi:10.1093/bib/bbs086.
22. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* 2013, 14:1–13.
23. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009, 4:D44–D57.
24. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics* 2004, 20:578–580.
25. Beissbarth T, Speed TP. GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* 2004, 20:D1464–D1465.
26. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005, 102:D15545–D15550.
27. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of gene-

ontology annotations with parent child analysis. *Bioinformatics* 2007, 23:3024–3031.

28. Wendl MC, Wallis JW, Lin L, Kandoth C, Mardis ER, Wilson RK, Ding L. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics* 2011, 27: D1595–D1602.
29. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. *PLoS One* 2010, 5:e8918.
30. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, Huntsman DG, Caldas C, Aparicio SA, Shah SP. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol* 2012, 13:1–14.
31. Torkamani A, Schork NJ. Identification of rare cancer driver mutations by network reconstruction. *Genome Res* 2009, 19:1570–1578.
32. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nat Methods* 2013, 10:1108–1115.
33. Leiserson MDM, Vandin F, Wu H, Dobson JR, Eldridge JV, Thomas Jl, Papoutsaki A, Kim Y, Niu B, McLellan M, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* 2015, 47:106–114.
34. Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM. Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). *Bioinformatics* 2013, 29:2757–2764.
35. Miller CA, Settle SH, Sulman EP, Aldape KD, Milosavljevic A. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med Genomics* 2011, 4:34.
36. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. *Genome Res* 2012, 22:375–385.
37. Leiserson MDM, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol* 2013, 9: e1003054.
38. Leiserson MD, Wu H-T, Vandin F, Raphael BJ. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol* 2015, 16:160.
39. Constantinescu S, Szczurek E, Mohammadi P, Rahnenführer J, Beerenwinkel N. TiMEx: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics* 2016, 32:968–975.
40. Cristea S, Kuipers J, Beerenwinkel N. pathTiMEx: joint inference of mutually exclusive cancer pathways and their dependencies in tumor progression. In: Singh, M, ed. *Research in Computational Molecular Biology*. New York, NY: Springer International Publishing; 2016, 65–82.
41. Sakoparnig T, Fried P, Beerenwinkel N. Identification of constrained cancer driver genes based on mutation timing. *PLoS Comput Biol* 2015, 11:e1004027.
42. Szczurek E, Beerenwinkel N. Modeling mutual exclusivity of cancer mutations. *PLoS Comput Biol* 2014, 10:e1003503.
43. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res* 2012, 22:398–406.
44. Babur Ö, Gönen M, Aksoy BA, Schultz N, Ciriello G, Sander C, Demir E. Systematic identification of cancer driving signalling pathways based on mutual exclusivity of genomic alterations. *Genome Biol* 2015, 16:45.
45. Kim Y-A, Cho D-Y, Dao P, Przytycka TM. MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics* 2015, 31:i284–i292.
46. Huret JL, Minor SL, Dorkeld F, Dessen P, Bernheim A. Atlas of genetics and cytogenetics in oncology and haematology, an interactive database. *Nucleic Acids Res* 2000, 28:349–351.
47. An O, Dall'Olio GM, Mourikis TP, Ciccarelli FD. NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. *Nucleic Acids Res* 2016, 44: D992–D999.
48. Porta-Pardo E, Hrabe T, Godzik A. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res* 2015, 43:D968–D973.
49. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Prlic A, Quesada M, et al. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res* 2013, 41:D475–D482.
50. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012, 483:603–607.
51. Griffith M, Griffith OL, Coffman AC, Weible JV, McMichael JF, Spies NC, Koval J, Das I, Callaway MB, Eldred JM, et al. DGIdb: mining the druggable genome. *Nat Methods* 2013, 10:1209–1210.
52. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000, 28:27–30.
53. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25:25–29.

54. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, et al. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 2009, 37:D619–D622.

55. Lin J, Gan CM, Zhang X, Jones S, Sjöblom T, Wood LD, Parsons DW, Papadopoulos N, Kinzler KW, Vogelstein B, et al. A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res* 2007, 17:D1304–D1318.

56. Goeman JJ, Mansmann U. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* 2008, 24:537–544.

57. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 2006, 22:1600–1607.

58. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. *Bioinformatics* 2007, 23:257–258.

59. Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene set analysis for cancer mutation data. *Genome Biol* 2010, 11:R112.

60. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol* 2010, 11:R53.

61. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicheria D, Raju R, Shafreen B, Venugopal A, et al. Human Protein Reference Database—2009 update. *Nucleic Acids Res* 2009, 37:D767–D772.

62. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006, 34: D535–D539.

63. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2013, 41:D808–D815.

64. Turner B, Razick S, Turinsky AL, Vlasblom J, Crowdy EK, Cho E, Morrison K, Donaldson IM, Wodak SJ. iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)* 2010, 2010:baq023.

65. Chen JL, Li J, Stadler WM, Lussier YA. Protein-network modeling of prostate cancer gene signatures reveals essential pathways in disease recurrence. *J Am Med Inform Assoc* 2011, 18:392–402.

66. Bredel M, Scholtens DM, Harsh GR, Bredel C, Chandler JP, Renfrow JJ, Yadav AK, Vogel H, Scheck AC, Tibshirani R. A network model of a cooperative genetic landscape in brain tumors. *JAMA* 2009, 302:261–275.

67. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007, 3:140.

68. Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BKH, Sia Y, Huang SK, Hoon DSB, Liu ET, Hillmer A, et al. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res* 2015, 43: e44. doi:10.1093/nar/gku1393.

69. Nitsch D, Gonçalves JP, Ojeda F, de Moor B, Moreau Y. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics* 2010, 11:460.

70. Babaei S, Hulsman M, Reinders M, de Ridder J. Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion. *BMC Bioinformatics* 2013, 14:29.

71. Chen Y, Zhu J, Lum P, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, et al. Variations in DNA elucidate molecular networks that cause disease. *Nature* 2008, 452:429–435.

72. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 2006, 7(suppl 1):S7.

73. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* 2011, 18:507–522.

74. Huang L, Li F, Sheng J, Xia X, Ma J, Zhan M, Wong STC. DrugComboRanker: drug combination discovery based on target network analysis. *Bioinformatics* 2014, 30:i228–i236.

75. Huang R, Southall N, Wang Y, Yasgar A, Shinn P, Jadhav A, Nguyen D, Austin CP. The NCGC Pharmaceutical Collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci Transl Med* 2011, 3:80ps16.

76. Cheng F, Jia P, Wang Q, Zhao Z. Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy. *Oncotarget* 2014, 5:3697–3710.

77. Mitsopoulos C, Schierz AC, Workman P, Al-Lazikani B. Distinctive behaviors of druggable proteins in cellular networks. *PLoS Comput Biol* 2015, 11:e1004597.

78. Cheng YK, Beroukhim R, Levine RL, Mellingho IK, Holland EC, Michor F. A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis. *PLoS Comput Biol* 2012, 8:e1002337.

79. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, et al. Understanding

multicellular function and disease with human tissue-specific networks. *Nat Genet* 2015, 47:569–576.

80. Yeang C-H, McCormick F, Levine A. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J* 2008, 22:2605–2622.

81. Zhang J, Wu L-Y, Zhang X-S, Zhang S. Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics* 2014, 15:271.

82. Srihari S, Singla J, Wong L, Ragan MA. Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer. *Biol Direct* 2015, 10:57.

83. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990, 61:759–767.

84. Gerstung M, Eriksson N, Lin J, Vogelstein B, Beerenwinkel N. The temporal order of genetic and pathway alterations in tumorigenesis. *PLoS One* 2011, 6:e27136.

85. Raphael BJ, Vandin F. Simultaneous inference of cancer pathways and tumor progression from cross-sectional mutation data. *J. Comput. Biol.* 2015, 22:510–527.

86. McPherson A, Wu C, Wyatt AW, Shah S, Collins C, Sahinalp SC. nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res* 2012, 22:2250–2261.

87. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 2010, 26:i237–i245.

88. Reimand J, Wagih O, Bader GD. The mutational landscape of phosphorylation signaling in cancer. *Sci Rep* 2013, 3:2651.

89. Jacobsen A, Silber J, Harinath G, Huse JT, Schultz N, Sander C. Analysis of microRNA-target interactions across diverse cancer types. *Nat Struct Mol Biol* 2013, 20:1325–1332.

90. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D. An integrated approach to uncover drivers of cancer. *Cell* 2010, 143:1005–1017.

91. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 2010, 17:98–110.

92. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Li D, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep3* 2013, 2650. doi:10.1038/srep02650.

93. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, et al. Mutational landscape and significance across 12 major cancer types. *Nature* 2013, 502:333–339.

94. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* 2013, 10:1081–1082.

95. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, Yang J, Broom BM, Verhaak RGW, Kane DW. TCPA: a resource for cancer functional proteomics data. *Nat Methods* 2013, 10:1046–1047.

96. Cline MS, Craft B, Swatloski T, Goldman M, Ma S, Haussler D, Zhu J. Exploring TCGA pan-cancer data at the UCSC cancer genomics browser. *Sci Rep3* 2013, 2652. doi:10.1038/srep02652.

97. Omberg L, Ellrott K, Yuan Y, Kandoth C, Wong C, Kellen MR, Friend SH, Stuart J, Liang H, Margolin AA. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat Genet* 2013, 45:1121–1126.

98. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz M, Sander C. Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* 2013, 45:1127–1133.

99. Sifrim A, Popovic D, Tranchevent L, Ardeshirdavani A, Sakai R, Konings P, Vermeesch JR, Aerts J, De Moor B, Moreau Y. eXtasy: variant prioritization by genomic data fusion. *Nat Methods* 2013, 10:1083–1084.

100. Brohée S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 2006, 7:488.

101. Ruffalo M, Koyutürk M, Sharan R. Network-based integration of disparate omic data to identify 'silent players' in cancer. *PLoS Comput Biol* 2015, 11: e1004595.

102. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011, 12:56–68.

103. Schwartz AS, Yu J, Gardenour KR, Finley RL Jr, Ideker T. Cost-effective strategies for completing the interactome. *Nat Methods* 2009, 6:55–61.

104. Venkatesan K, Rual J, Vazquez A, Stelzi U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh K, et al. An empirical framework for binary interactome mapping. *Nat Methods* 2009, 6:83–90.

105. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* 2012, 9:471–472.

106. Deane CM, Salwiński Ł, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* 2002, 1:349–356.

107. Mahdavi MA, Lin Y-H. False positive reduction in protein-protein interaction predictions using gene ontology annotations. *BMC Bioinformatics* 2007, 8:262.

108. Nguyen TN, Goodrich JA. Protein-protein interaction assays: eliminating false positive interactions. *Nat Methods* 2006, 3:135–139.

109. Mehla J, Caufield JH, Uetz P. Mapping protein–protein interactions using yeast two-hybrid assays. *Cold Spring Harb Protoc* 2015, 2015:442–452. doi:10.1101/pdb.prot086157.

110. Creixell P, Schoof EM, Simpson CD, Longden J, Miller CJ, Lou HJ, Perryman L, Cox TR, Zivanovic N, Palmeri A, et al. Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell* 2015, 163:202–217.

111. Vidal M, Cusick ME, Barabási A-L. Interactome networks and human disease. *Cell* 2011, 144:986–998.

112. Kondo S, Perrimon N. A genome-wide RNAi screen identifies core components of the G2-M DNA damage checkpoint. *Sci Signal* 2011, 4:rs1.

113. Boldt K, van Reeuwijk J, Gloeckner CJ, Ueffing M, Roepman R. Tandem affinity purification of ciliopathy-associated protein complexes. *Methods Cell Biol* 2009, 91:143–160.