FEBS Letters

FEBS PRESS science publishing by scientists

REVIEW ARTICLE

Cancer genetics meets biomolecular mechanism—bridging an age-old gulf

Juan Carlos González-Sánchez[1,2,†], Francesco Raimondi[1,2,†] and Robert B. Russell[1,2]

1 Bioquant, Heidelberg University, Germany
2 Heidelberg University Biochemistry Center (BZH), Germany

Correspondence
R. B. Russell, Bioquant, Heidelberg University, Im Neuenheimer Feld 267, Heidelberg 69120, Germany
Fax: +49 6221 54 51 486
Tel: +49 6221 54 51 362
E-mail: robert.russell@bioquant.uni-heidelberg.de

[†]These authors contributed equally to this work

(Received 13 November 2017, revised 15 January 2018, accepted 19 January 2018, available online 8 February 2018)

doi:10.1002/1873-3468.12988

Edited by Wilhelm Just

Increasingly available genomic sequencing data are exploited to identify genes and variants contributing to diseases, particularly cancer. Traditionally, methods to find such variants have relied heavily on allele frequency and/or familial history, often neglecting to consider any mechanistic understanding of their functional consequences. Thus, while the set of known cancer-related genes has increased, for many, their mechanistic role in the disease is not completely understood. This issue highlights a wide gap between the disciplines of genetics, which largely aims to correlate genetic events with phenotype, and molecular biology, which ultimately aims at a mechanistic understanding of biological processes. Fortunately, new methods and several systematic studies have proved illuminating for many disease genes and variants by integrating sequencing with mechanistic data, including biomolecular structures and interactions. These have provided new interpretations for known mutations and suggested new disease-relevant variants and genes. Here, we review these approaches and discuss particular examples where these have had a profound impact on the understanding of human cancers.

Keywords: bioinformatics; genetics; mechanism; protein interactions; protein structure

Genetic variants: the latest data explosion

The significant development of high-throughput sequencing technologies has boosted gene discovery and helped formulate new genetic variant-disease associations [1,2]. However, establishing precise, causative genetic variation has only been possible for a minority of rare monogenic diseases that are the result of high-penetrance alleles, or particular recurrent cancer mutations [3–5]. Unfortunately, most human illnesses, including those with a wide societal impact, are more complex and cannot be reduced to a single allele or gene.

Although a typical sequencing exercise can identify a large number of variants, determining which are responsible for disease still remains a challenging task [6]. Accordingly, several computational tools have been developed to prioritize candidate pathogenic variants, and these already enhance strategies for early detection, diagnosis and treatment [6,7]. While many of these methods are generic, others target specific diseases, particularly cancers. Methodologically, there are many flavours, but most popular approaches rely on properties like sequence conservation, allele frequency or known/

Abbreviations
CGC, Cancer Gene Census; COSMIC, Catalogue Of Somatic Mutations In Cancer; GAP, GTPase-activating protein; GDI, guanine dissociation inhibitor; GEF, guanine exchange factor; ICGC, International Cancer Genomics Consortium; TCGA, The Cancer Genome Atlas; TSG, tumour suppressor gene; VUS, variant of unknown significance.

Table 1. Classical computational methods for variant/gene prioritization and functional impact assessment. Underline indicates those methods that are specific for cancer. Variant types: missense variants (M), short insertions/deletions (ID), and noncoding variants (NC).

Tool	Variant types	Method overview
SIFT [57]	M	Predicts impact through assessment of phylogenetic conservation.
PROVEAN [58]	M, ID	Predicts impact through assessment of phylogenetic conservation.
PolyPhen-2 [42]	M	Predicts impact based on sequence, phylogenetic and structural features.
PANTHER-PSEP [59]	M	Predicts impact based on position-specific 'evolutionary preservation'.
FATHMM-XF [60]	M, NC	Predicts impact based on a wide set of features including sequence conservation, proximity to genomic features or chromatin accessibility.
MutationTaster2 [61]	M, ID, NC	Integrates data from 1000 Genomes, ClinVar, HGMD and ENCODE to predict variant pathogenicity.
VAAST 2 [62]	M, ID, NC	Employs an aggregative variant association test that combines amino acid substitution, allele frequency and phylogenetic conservation.
CADD [63]	M, ID, NC	Combines diverse annotations (conservation, regulatory information, expression levels) and scores of SIFT and PolyPhen-2.
ANNOVAR [64]	M, ID	Integrates several annotation data sets from the UCSC Genome Browser, including conservation or transcription factor binding site prediction.
REVEL [65]	M	Combines scores of 13 other tools to predict pathogenicity.
fitCons [66]	M, ID, NC	Integrates high-throughput functional genomic data to cluster genomic positions, and then, estimate a probability of fitness for each based on patterns of polymorphism and divergence.
CHASM [67]	M	Predicts impact based on several biochemical, sequence and structural features.
MutationAssessor [43]	M	Predicts impact through assessment of phylogenetic conservation.
MutSig [68]	M, ID	Combines mutation frequency compared to gene-specific background mutation rate, mutation clustering along sequence and functional impact assessed by other predictors or by phylogenetic conservation.
MuSiC [69]	M, ID	Combines several metrics including frequency, clustering and correlation with clinical phenotypes. Allows identification of significantly affected pathways and correlation/mutual exclusion between mutated genes.
OncodriveFM [70]	M, ID	Assesses the accumulation of variants of high functional impact in the gene, which is predicted by SIFT, Polyphen-2 and MutationAssessor.
OncodriveCLUST [71]	M	Searches for mutation clustering along the sequence.
ActiveDriver [72]	M	Considers frequency and clustering of mutations in the context of phosphorylation signalling (phosphosites, kinase domains, etc).
CRAVAT [73]	M, I	Integrates CHASM, VEST and different annotations.
IntOGen [74]	M, I	Web platform for pan-cancer driver identification. Its pipeline integrates OncodriveFM, OncodriveCLUST and MutSig.

predicted protein structure (Table 1; Fig. 1A). While these are often excellent indicators of possible pathogenicity, they do not capture a critically important aspect that could help predict it better: biomolecular mechanism.

A gulf between genetics and molecular biology

Knowledge of disease genes has helped drive the understanding of biological mechanism. For instance, we know a great deal about genes such as *TP53* or *KRAS* because they were discovered early on to play roles in cancer. These were products of an age when the identification of disease variants was relatively slow, as were the complementary experiments to determine gene function. However, when thousands of gene variants can be identified in a single sequencing experiment, it is not feasible to wait for the community of molecular biology to delve deeply into the function of each of them. The consequence is a growing number of disease variants for which geneticists are highly confident that they are causative, but for which there is often little or no reasonable explanation as to how the molecular change relates to the disease phenotype. In addition, many are also in genes that are themselves poorly understood. For example, although there are many disease genes associated with the cilia, it has been less well studied than other parts of the cellular machinery, meaning that many of the protein products coded by these genes are still comparatively obscure [8].

Formally, variants are classified as either: positive, if they disrupt gene function; negative, if they do not; or as variants of unknown significance (VUS), when their

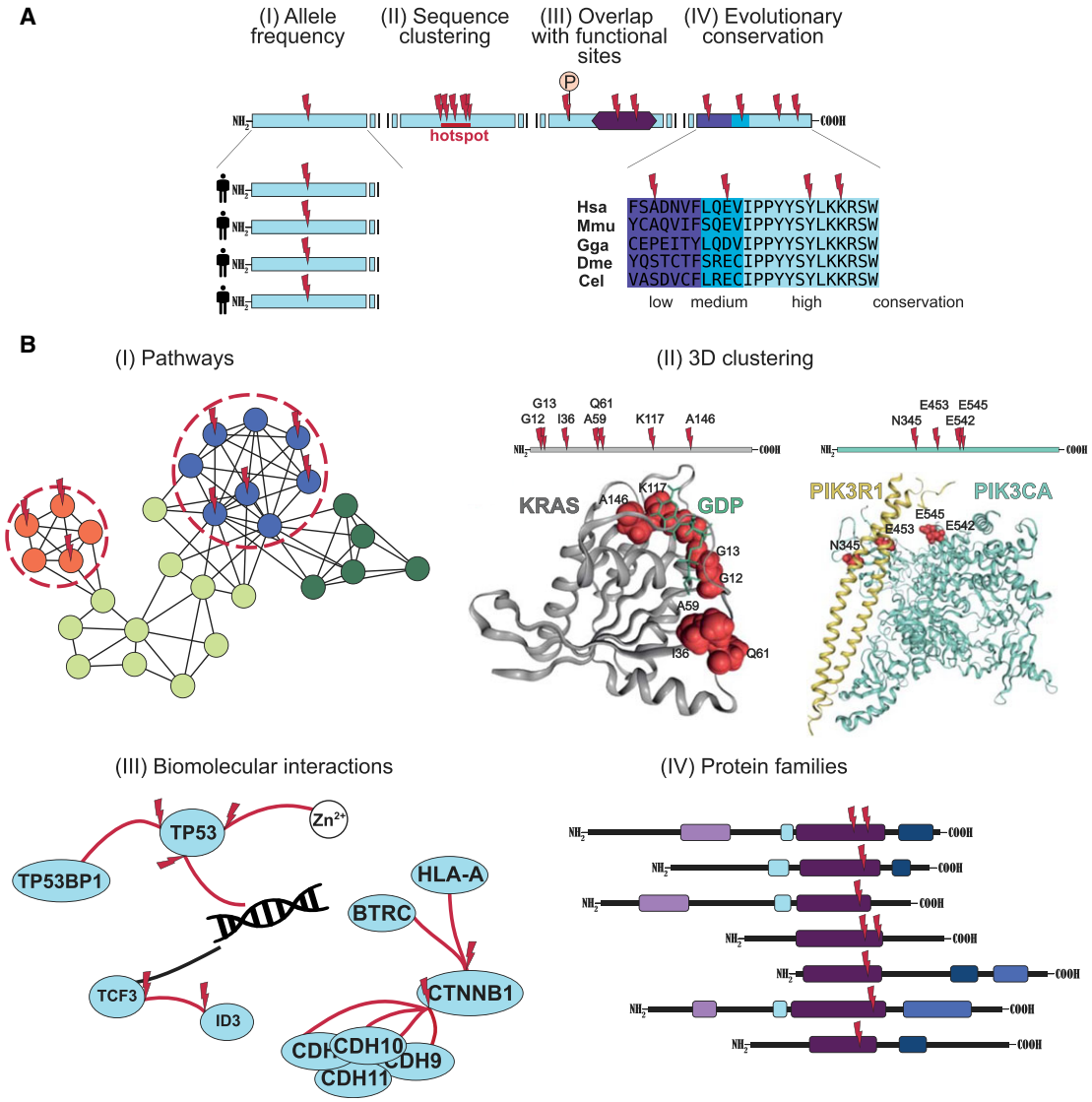

Fig. 1. Biological features exploited by variant prioritization tools. (A) Gene-level features used for disease variant identification are illustrated along different segments of a schematic protein. Red thunderbolt icons represent point mutations. (I) Allele frequency is assessed by comparing sequences from different patients/samples. Below, four additional segments represent the same section of the protein, harbouring the same mutation, in four different patients. (II) Sequence clustering is represented by proximal mutations along the protein sequence defining a hotspot. (III) Overlap with functional sites is illustrated with mutations overlapping with a phosphorylation site (circled 'P') and a protein domain (purple hexagonal box). (IV) Evolutionary conservation of the residues affected by mutations can be assessed by protein sequence alignment with orthologues, as illustrated below. The colours in the alignment indicate a rough conservation level: low (dark blue), medium (cyan) and high (light blue), for which corresponding sections are equally coloured on the protein segment. (B) Mechanistic approaches at multiscale resolution that provide deeper insights into mutation effects. Red thunderbolt icons represent mutations. (I) Pathways commonly affected by mutated genes. In this schematic, same-colour spheres represent genes participating in the same pathway. Red doted circles encompass two commonly perturbed pathways by individual mutations in their genes. (II) 3D clustering of mutations is illustrated in two different oncogenes: *KRAS* and *PIK3CA* (PDB codes: 3GFT and 2RD0). The positions affected by mutations in each protein are displayed along their sequences above. On the structures below, these positions are displayed as red balls that locate spatially close, affecting the same interfaces: with GDP, in the case of *KRAS*, and with *PIK3R1*, in the case *PIK3CA*. (III) Biomolecular interactions schematic illustrating different cases where mutations in the same protein affect different (protein-protein, protein-DNA/RNA, and protein-small molecule) interfaces and thus have different impacts. (IV) Protein families can be used to detect scarce mutations in different proteins that are affecting similar positions within a shared functional domain. Black lines and coloured boxes are schematics of proteins and their different domains. They are aligned by a common domain (dark purple) where all proteins show mutations in an equivalent position.

impact is still unknown [9]. It comes as no surprise that the last class is where the majority of variants uncovered by sequencing land (e.g. in cancer [10]), complicating diagnosis and risk assessment. Thus, there is generally a pressing need to understand whether these variants have a functional impact, even if they are not clinically confirmed. This is particularly true when a VUS is discovered in a gene that already has established causative variants, and where diagnostic or treatment decisions need to be made (e.g. in Alport syndrome [11], or in cancer [12]).

There is something of a historic gap between classical medical genetics, which ultimately relates genotype to macroscopic phenotype, and basic molecular biology, which generally aims to understand the molecular basis for biological phenomena. Any cursory glance at many of the thousands of papers discussing particular disease genes or mutants often shows limited interest in how the (confidently determined) genetic cause relates to the underlying molecular biology. Many also operate, unwittingly, under a rather lazy generalization that any mutation in a particular gene will lead to a similar functional consequence, something that would horrify structural biologists, and a presumption which is challenged by systems biology [13,14]. Moreover, even when mechanism is considered, there is a widespread tendency to stop interpretation once a list of mutated genes is obtained.

Meanwhile, mechanistic data are also on the rise. Decades of advances in structural and cell biology, not to mention the plethora of high-throughput methods to study proteins, nucleic acids, metabolites and their respective interactions, means that we are better placed than ever to understand the roles that gene products play in the wider functional context of the cell [15,16]. However, in practice the degree to which these findings impact on genetics remains much lower than it could be. Fortunately, recent years have seen a variety of high-throughput methods that can exploit existing mechanistic data fast enough to cope with the growth of sequence information emerge (Table 2). By using and integrating diverse sources of mechanistic information (Fig. 1B), they have provided deeper insights into the molecular basis of how genetic variants ultimately cause diseases, particularly in the case of cancer.

Cancer as a paradigm

Cancers are diverse and complex diseases that have been the subject of intense scientific scrutiny by a flurry of recent genomic sequencing projects, such as those coordinated by The Cancer Genome Atlas (TCGA) [17] and the International Cancer Genomics Consortium (ICGC) [18]. Together with advances in molecular oncology, these efforts have charted the complex landscape of genetic factors underpinning the most common cancers, and led to an expansion of the repertoire of biological processes considered as hallmarks of cancer, with metabolism reprogramming and immune system evasion as the latest additions [19]. Consequently, the list of genes confirmed to drive tumour initiation and development has increased, but it has done so faster than any mechanistic understanding. For example, although *ARID2* and *NUTM2A* are established to have

Table 2. Computational methods for variant/gene prioritization that provide mechanistic information. Underline indicates those methods that are specific for cancer. Variant types: missense variants (M) or sets of mutated genes (GS).

Tool	Variant Types	Method Overview
GSEA [39]	GS	Identifies pathways/networks enriched in a set of differentially expressed or mutated genes.
G:Profiler [75]	GS	Identifies pathways, Gene-Ontology or transcription factor binding sites enriched in a set of differentially expressed or mutated genes.
SLEA [76]	GS	Enrichment analysis at the level of samples which allows discovery of different cancer subtypes.
ReactomeFIViz [40]	GS	Platform for pathway and Reactome functional interaction network analysis.
GeneMANIA [77]	GS	Reconstructs networks from gene sets using interaction, pathway (and more) data, with the aim of identifying other functionally related genes.
PARADIGM [41]	GS	Integrates multidimensional cancer data to identify significantly affected pathways.
Mechismo [44]	M	Predicts impact of an amino acid substitution or modification on protein interactions by mapping them on 3D structures and the interactome.
dSysMap [45]	M	Systematic mapping of disease-related missense mutations on the structures of the human interactome.
CLUMPS [47]	M	Detects clustering of mutations on the protein 3D structure and enrichment of those in PPI interfaces.
HotMAPS [49]	M	Detects clustering of mutations on the protein 3D structure, including protein complexes.
3dhotspots [46]	M	Detects clustering of mutations on the protein 3D structure with a focus on rare variants.
MutationAligner [53]	M	Identifies mutation hotspots in protein domains across many cancer types.
Oncodomains [55]	M	Identifies mutation hotspots in protein domains across many cancer types.

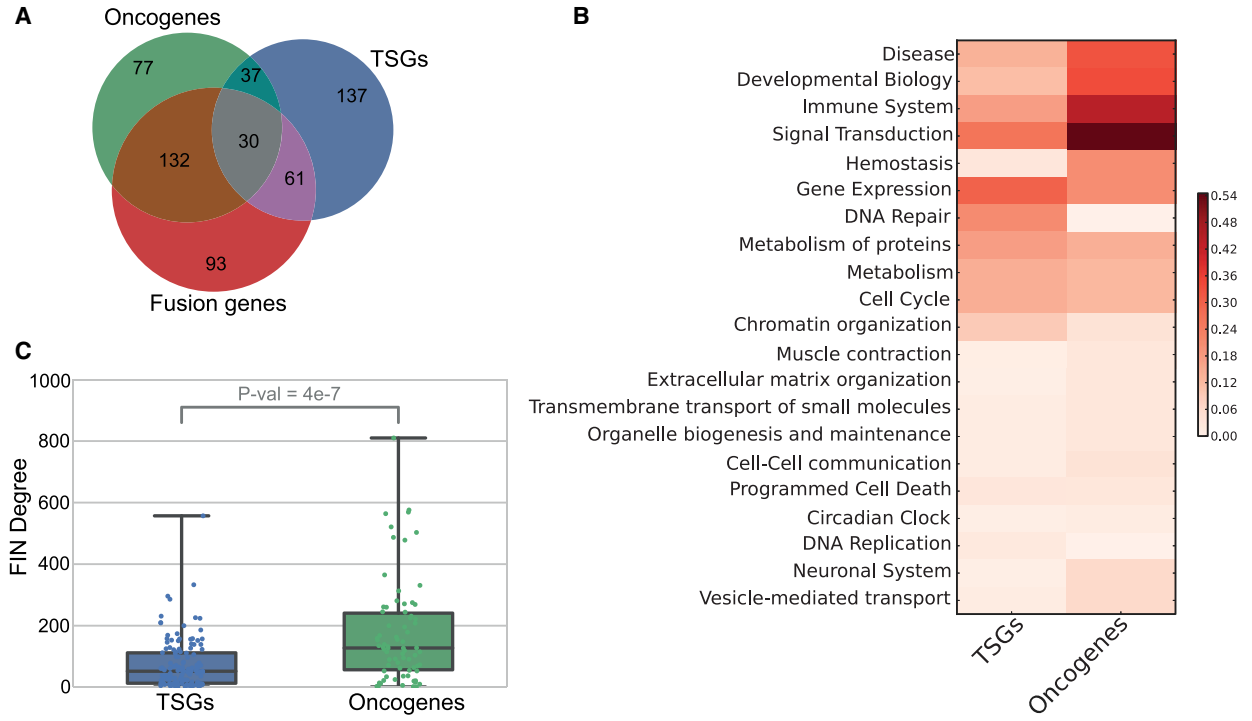

Fig. 2. Classification of Cancer Gene Census (CGC) genes and differences in Functional Interaction Network (FIN) properties. (A) Venn diagram showing the total number of genes in each of the three CGC role in cancer categories and in the overlaps between them. (B) Heatmap depicting the fractions of oncogenes and tumour suppressor genes (TSGs) involved in Reactome top-hierarchy biological processes. Cell colour intensity is proportional to the gene fraction number. (C) Box plot displaying the distribution of the individual values of degree (number of connections) of each oncogene and TSG in the Reactome FIN. The boxes encompass from the first to the third quartiles, while the middle line indicates the median of the distribution. On average, oncogenes have a higher degree than TSGs (raw data from these analyses are available upon request).

roles in cancer [20,21], they lack the level of functional information of heavily studied genes such as *KRAS* or *CTNNB1*. The reality is that, despite the significant efforts, the disciplinary gap is also present between cancer genomics and molecular biology.

Cancer driver genes have been traditionally classified into two groups: oncogenes or tumour suppressors, according to whether they positively or negatively regulate the growth and proliferation of a neoplastic cell. Tumour suppressors are generally characterized by a diverse set of presumed deleterious mutations, which disrupt and inactivate protein function, usually localized without any apparent pattern. In contrast, oncogene hyperactivity is ascribed to either overexpression (*via* fusions, copy number increases or other alterations) or overactivation due to mutations affecting protein function (usually enzyme catalytic activity), which are thus normally at highly specific protein positions [22].

The Cancer Gene Census (CGC) is an ongoing initiative from COSMIC [23] to collect all genes that have mutations confirmed to be involved in cancer [24]. Currently (version 81), it contains 568 genes which are classified according to their role in tumorigenesis, including 137 tumour suppressor genes (TSGs), 77 oncogenes and 93 in a third class defined as fusion genes. However, most CGC genes cannot be described by any of the three classes alone (Fig. 2A). Among them, 67 genes stand out as they are classified as both oncogene and TSG (30 of them also as fusions) when, in theory, the two classes play a role in oncogenesis *via* opposite mechanisms. Indeed, growing evidence is supporting the fact that some genes can function as one or the other depending highly on the cellular context. Perhaps the most striking example of these is *TP53*, which was originally regarded as the canonical TSG, but it is now well-established that several *TP53* mutations are oncogenic [25]. Elsewhere, the small GTPases *RHOA*, *RAC1* and *CDC42*, key signal transducers of a wide range of biological functions, have been traditionally considered oncogenes, and therefore, targeted by antineoplastic agents [26,27]; however, recent cancer genomics studies and *in vivo* experiments have revealed new tumour suppressor properties of

these genes [28]. Similarly, the alpha subunit of the heterotrimeric G_s protein (*GNAS*) also shows a context-dependent role in cancer: while oncogenic mutations have been reported in thyroid, pituitary, pancreatic and gastrointestinal tumours [29], this gene also presents a tumour-suppressive function in skin basal-cell malignancies [30].

The power of biological context

The above examples emphasize the importance of not assuming an immovable role for disease genes and the need to study the molecular and cellular mechanisms, or biological context, in which these genes or variants operate. The simplest way to put a gene in functional context is through biological pathways. A total of 406 (71%) of the CGC genes are found in curated Reactome pathways (reactome.org [31]), of which 174 (30% of the total) participate in at least 10 different pathways. Interestingly, a stark contrast can be observed between some pathway/network properties of oncogenes and TSGs. First, they are involved in different biological processes: oncogenes are mostly found in signal transduction, developmental biology, immune system and haemostasis pathways, while TSGs tend to prevail in gene expression, DNA repair and protein metabolism (Fig. 2B). Moreover, oncogenes participate, on average, in a significantly higher number of pathways (rank sum $P = 0.0009$), and also have significantly higher number of connections to other proteins (or degree; $P = 4 \times 10^{-7}$; Fig. 2C) and a higher betweenness centrality (centrality of the molecule by measuring the extent to which it lies on paths between others; $P = 4 \times 10^{-6}$) when compared to TSGs within the Reactome Functional Interaction Network [32] (data available upon request). While we cannot exclude that this trend might in part be due to more intensive investigations into the specific biological functions of oncogenes, it suggests that they tend to lie at the cross-roads of several signalling pathways and wire a more complex network of interactions. In contrast, TSGs appear to be more specialized, with roles in fewer biological processes.

Biological context can also be illuminating by highlighting instances where multiple genes are effectively equivalent in terms of their roles in disease. Because the reductionist approach still predominates, many studies focus on adding another gene to the set of those responsible for a particular cancer. In the search of these genes, the frequency of mutations (defined as the number of samples or patients having a particular mutation) is important to avoid genes that have rare, possibly artefactual or passenger, mutations. However, mutual exclusivity between mutations in genes that are functionally linked can highlight cases where different mutations might perform equivalent roles in cancer, even when their frequency is comparatively low [33]. For instance, mutations in the transcription factor *TCF3* and its inhibitor *ID3* are largely mutually exclusive in Burkitt Lymphoma, and all of them equally disrupt the interaction between these proteins. Consequently, inhibition is avoided, resulting in an increased expression of *TCF3* targets, which ultimately promotes cell proliferation [34,35]. Similar exclusivity is present between mutations of three neighbouring genes in the Gα12/13 signalling pathway: *P2RY8*, *GNA13* and *RHOA* [36]. As their frequencies can be relatively low, mutations in *P2RY8* or *TCF3* might be overlooked when considering allele frequency alone. Moreover, this process leads to a greater fraction of patients being covered by the particular biological process, and thus, indicates an overall common mechanism of the disease, which can ultimately aid future diagnoses.

Exploiting molecular mechanism to prioritize genetic variants

Standard computational methods for cancer driver identification usually search for positive selection signals within individual genes, such as a mutation frequency higher than the background rate, or a nonrandom distribution of mutations along the sequence (Fig. 1A; Table 1). However, despite the growing number of sequenced cancer genomes, there are many less frequent, but nevertheless functionally important genes or variants that these approaches will often miss. All estimates indicate that the landscape of genomic variation in cancer is in fact dominated by such scarce events [22,37].

Pathways

The use of prior contextual knowledge about molecular and cellular mechanisms enables the grouping of different genetic alterations that are functionally related, disregarding, to a certain degree, their frequency. Similarly, one can identify affected common pathways involved in the development of the disease (Fig. 1B), which allows for the investigation of a wider set of mutations in a known biological context. Accordingly, several network- and pathway-based methods have been developed and applied to cancer genomics data to prioritize significantly affected genes and the biological processes in which they participate [38]. These approaches range from gene enrichment (e.g. GSEA [39]), to network analysis (e.g. ReactomeFIViz [40]) and pathway

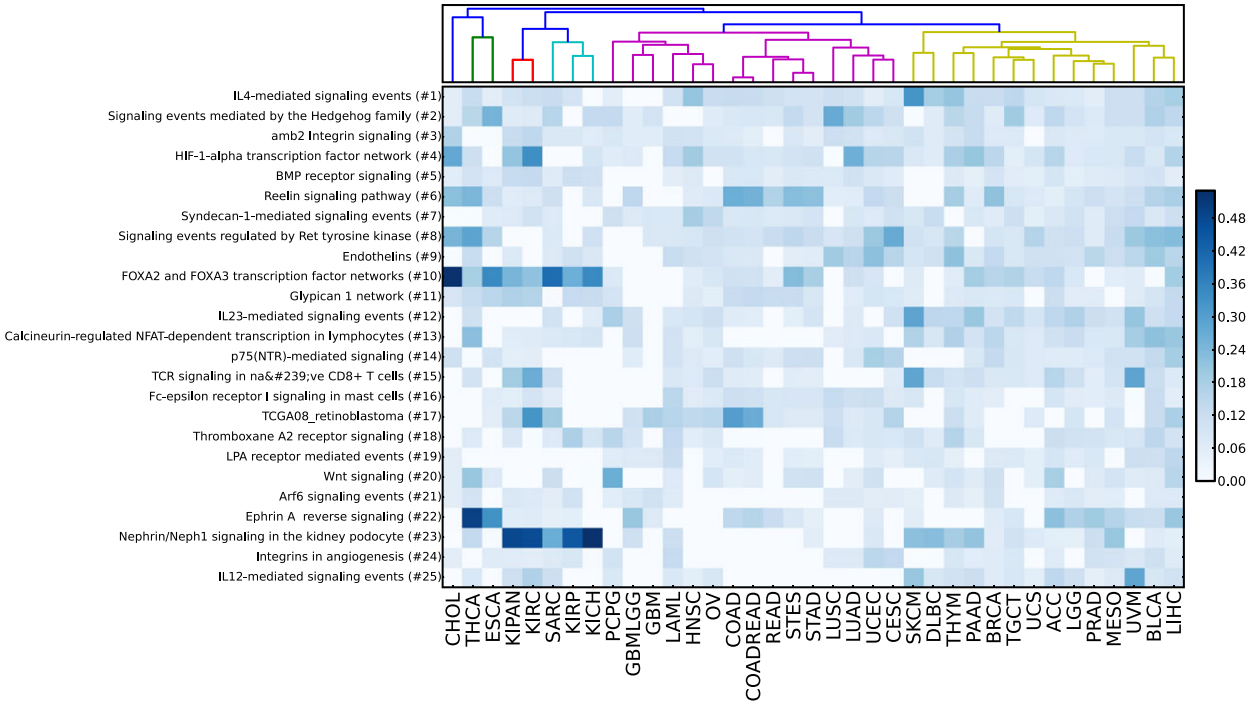

Fig. 3. Most commonly perturbed pathways in different cancers. Heatmap cells are coloured according to the significance ratio of the top 25 pathways (rows) most commonly altered in the PARADIGM analysis across TCGA cancer types (gdac.broadinstitute.org) (columns) on a blue scale. Pathways are ranked according to the number of cancer types where they are predicted to be significantly perturbed (i.e. significance ratio > 0.05). Cancer types are clustered by complete-linkage hierarchical clustering, based on distance calculation of pathways significance ratio values.

modelling (e.g. PARADIGM [41]) (Table 2). For example, the PARADIGM framework integrates multi-omics information—mainly gene expression and copy number variation data from TCGA—with curated pathways. It finds that while some pathways are in general equally perturbed in most cancers (e.g. IL4-mediated signalling or Hedgehog signalling), a few are specifically altered only in particular cancers, like nephrin/Neph1 signalling in kidney tumours (KIRC, KICH and KIRP) (Fig. 3).

Structures and interactions

The last few years have also seen the increased use of biomolecular structures to interpret the functional consequences of protein variation from cancer genomics studies. Indeed, structural information was already part of some of the original methods to predict mutation pathogenicity (e.g. PolyPhen [42]), and there are now several methods that allow one to identify when genetic variants affect biomolecular interfaces (e.g. MutationAssessor [43]). The need to study, often thousands of variants across hundreds of genes, has prompted the development of sophisticated platforms that systematically integrate three-dimensional structures and protein interactions with variant data, to predict their impact and identify putative mechanisms (Table 2). These include Mechismo [44], which allows queries of any putative protein change in the context of biomolecular interactions, or dSysMap [45], which is a systematic map of known disease mutations in a mechanistic context.

Other methods evaluate the 3D clustering of cancer missense mutations in protein structures or interaction interfaces, based on the assumption that spatially proximal mutations—which would not be revealed by positional frequency alone—might imply similar functional perturbations and phenotypes (Fig. 1B; Table 2). Systematic analyses using these approaches have been able to identify potential driver mutations and hotspot regions in both known (e.g. *MAP2K1* and *RAC1* [46]) and new candidate cancer genes, such as the kinetochore complex component *NUF2* [47]. Notably, both oncogenes and TSGs were shown to display 3D clustering patterns, contradicting the classic view that this is a hallmark feature of oncogenes [47,48]; however, a recent study has additionally reported some differences between the features of their hotspot regions, which likely reflect the distinct nature of gain-of-function and loss-of-function mutations [49]. As protein function is directly dependent on structure, the detection of three-dimensional hotspots

offers an improved capability to capture functional effects of variants over just considering linear sequence.

Some studies have systematically combined information of biomolecular structures with interaction data to detect enrichment of somatic missense variants at protein–protein [50,51] or protein–chemical/nucleic acid interfaces [51]. These analyses reveal significant differences in the way interactions of important driver genes are perturbed in different cancer types or even in different patients, which sometimes correlates with cancer severity. They also highlight that not all mutations affecting a gene are equivalent, but that their consequence can depend on which of several possible interfaces they affect (Fig. 1B). For instance, when analysed across all cancers, fold-disrupting mutations of the zinc binding site in *TP53* appear to be associated with poorer survival, in contrast to those more specific to DNA or regulator protein binding [51]. They also showed that oncodrivers harbour the bulk of mutations, while those affecting their direct interaction partners are unusual: for example, *TP53* is highly mutated, but mutations in its most common interaction partners (e.g. *TP53BP1*, *TP53BP2*) are comparatively rare [51].

These deeper insights into mechanism have also called into question the binary TSG/oncogene paradigm. For example, in Burkitt Lymphoma, several seemingly functional and disease-relevant variants within *RHOA* appear to be neither clearly oncogenic nor tumour suppressive in nature [44,52]. Several of these changes lie at the interface with multiple regulators of *RHOA* function (GAPs, GEFs or GDIs), and seem to tinker with their interaction affinities, thus probably shifting *RHOA* towards particular pathways

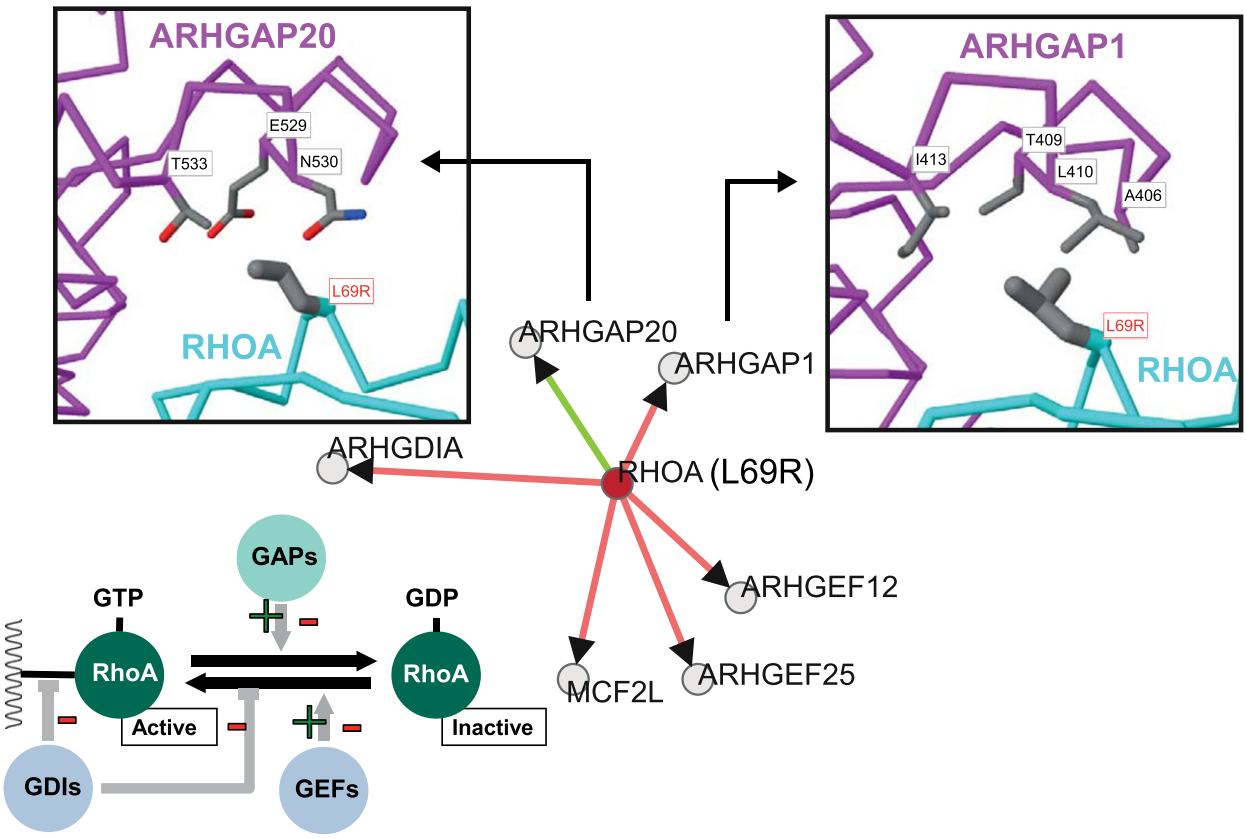

Fig. 4. Differential effects of the *RHOA* p.L69R mutation on interaction partners. The central network shows *RHOA* interacting with key GTPase-activating proteins (GAPs) *ARHGAP1* and *ARHGAP20*, Guanine exchange factors (GEFs) *ARHGEF1*, *ARHGEF25* and *MCF2L* and a Guanine dissociation inhibitor (GDI) *ARHGDI1*. Red and green lines indicate predicted destabilizing and enhancing interactions respectively (determined using Mechismo). The panels on top left and right show the location of the Leucine in *RHOA* in the context of the *ARHGAP20* and *ARHGAP1* structures. Note that the Leucine in the *ARHGAP1* structure lies in a hydrophobic pocket, suggesting that the mutation to a charged Arginine is predicted to disfavour the interaction, in contrast to the polar/negatively charged pocket in *ARHGAP20* that is predicted to favour the change to Arginine. The bottom left of the figure shows a schematic of how GAPs, GEFs and GDIs modulate the balance between inactive and active *RHOA*.

(Fig. 4). In this context, the enzyme appears to be neither over- nor underactivated, but altered subtly somehow to optimize tumour growth.

Protein families

Protein families and evolutionary relationships have been exploited to detect sparse mutations affecting equivalent, functionally conserved amino acids within a shared protein domain (Fig. 1B; Table 2). For example, several rare uncharacterized mutations in different cancer types, in the genes *EGFR* (L861), *EPHA2* (V763), *FGFR1* (D647) and *PDGFRA* (D842), lie at positions of the kinase domain that are analogous to well-known mutation hotspots, like those in *KIT* (D816), *FLT3 (*D835) and *BRAF* (V600), suggesting similar molecular mechanisms and downstream consequences, and thus, potential therapeutic actions [53]. Interestingly, domain-centric approaches have revealed clusters of mutations at the domain level (i.e. domain hotspots) in both oncogenes and TSGs, further narrowing the separation between them [54,55].

Overall, it is clear that leveraging mechanistic data, in the form of pathways, interactions, networks and three-dimensional structures, provide deeper insights for known cancer variants and genes, as well as the means to unearth those previously concealed by their low frequency. This is bound to have a profound impact on the understanding of the molecular processes that result in cancer initiation and progression.

Perspectives

Gaps between disciplines can be difficult to bridge, but the challenge to connect cancer genomics and molecular biology is ideally suited to bioinformatics groups specializing in the interrogation of mechanistic data. Great progress has already been made in the development of tools, for example, to study the structural consequences of individual sequence variations, or to assess gene-ontology or pathway enrichment. However, at the origin of this gap are deep-rooted differences in perspective. The average geneticist is less inclined to consider biomolecular mechanism just as the average molecular biologist often shows limited interest for information from genetics. The truth is that this gap has not been of such profound concern in the past decades, principally because there was so much to do on both sides that did not particularly involve the other. However, the need to prioritize the exponentially growing number of genetic variants makes it imperative to link them to mechanism, bringing both disciplines together.

Certainly the case needs to be made for whether mechanism can make a substantial difference to genetics, and there are already many studies demonstrating that a systematic coupling of genetic information to mechanistic biology can deliver promising results (e.g. [56]). This is particularly true given the aims of personalized medicine, whereby thousands of previously uncharacterized variants will be found within an individual genome, and will need to be quickly assessed in terms of their likely impact on phenotype. Equally, structural and molecular biologists could profit from a more systematic interrogation of existing genetic variant data in terms of how it might impact on molecules of interest, and moreover, to link, where possible, any findings to human disease.

Computational tools will need first to systematically integrate data on genetic variation, protein structure, interactions, biological pathways and annotated, or predicted, protein features. They will then, most importantly, need to provide the means to answer questions of direct relevance to human geneticists, molecular biologists, clinicians and the entire spectrum of scientists who can profit from these game changing advances in the era of high-throughput sequencing for personalized medicine.

Acknowledgements

The group is supported by the Cell Networks Excellence initiative of the Germany Research Foundation (DFG). F.R. was supported by a Alexander von Humboldt Foundation post-doctoral fellowship and a grant from the Michael J Fox Foundation.

References

1 1000 Genomes Project Consortium; Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA *et al.* (2015) A global reference for human genetic variation. *Nature* **526**, 68–74.

2 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291.

3 Boycott KM, Vanstone MR, Bulman DE and MacKenzie AE (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* **14**, 681–691.

4 Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T *et al.* (2015) The genetic

basis of mendelian phenotypes: discoveries, challenges, and opportunities. *Am J Hum Genet* **97**, 199–215.
5 Amberger JSS, Bocchini CAA, Schiettecatte F, Scott AFF and Hamosh A (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**, D789–D798.
6 Eilbeck K, Quinlan A and Yandell M (2017) Settling the score: variant prioritization and Mendelian disease. *Nat Publ Gr* **18**, 599–612.
7 Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J and Trajanoski Z (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform* **15**, 256–278.
8 Boldt K, van Reeuwijk J, Lu Q, Koutroumpas K, Nguyen TM, Texier Y, van Beersum SE, Horn N, Willer JR, Mans DA *et al.* (2016) An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. *Nat Commun* **7**, 11491.
9 Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV *et al.* (2008) Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat* **29**, 1282–1291.
10 Hayes DN and Kim WY (2015) The next steps in next-gen sequencing of cancer genomes. *J Clin Invest* **125**, 462–468.
11 Fernandez-Rosado F, Campos A, Alvarez-Cubero MJ, Ruiz A and Entrala-Bernal C (2015) Improved genetic counseling in Alport syndrome by new variants of *COL4A5* gene. *Nephrology* **20**, 502–505.
12 Schram AM, Reales D, Galle J, Cambria R, Durany R, Feldman D, Sherman E, Rosenberg J, D'Andrea G, Baxi S *et al.* (2017) Oncologist use and perception of large panel next-generation tumor sequencing. *Ann Oncol* **28**, 2298–2304.
13 Zhong Q, Simonis N, Li QR, Charloteaux B, Heuze F, Klitgord N, Tam S, Yu H, Venkatesan K, Mou D *et al.* (2009) Edgetic perturbation models of human inherited disorders. *Mol Syst Biol* **5**, 321.
14 Creixell P, Schoof EM, Simpson CD, Longden J, Miller CJ, Lou HJ, Perryman L, Cox TR, Zivanovic N, Palmeri A *et al.* (2015) Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell* **163**, 202–217.
15 Kim Y-A and Przytycka TM (2013) Bridging the gap between genotype and phenotype via network approaches. *Front Genet* **3**, 227.
16 Lappalainen T (2015) Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res* **25**, 1427–1431.
17 The Cancer Genome Atlas Research; Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C and Stuart JM (2013) The cancer genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120.
18 The International Cancer Genome Consortium; Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I *et al.* (2010) International network of cancer genome projects. *Nature* **464**, 993–998.
19 Hanahan D and Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* **144**, 646–674.
20 Zhao H, Wang J, Han Y, Huang Z, Ying J, Bi X, Zhao J, Fang Y, Zhou H, Zhou J *et al.* (2011) ARID2: a new tumor suppressor gene in hepatocellular carcinoma. *Oncotarget* **2**, 886–891.
21 Sugita S, Arai Y, Aoyama T, Asanuma H, Mukai W, Hama N, Emori M, Shibata T and Hasegawa T (2017) NUTM2A-CIC fusion small round cell sarcoma: a genetically distinct variant of CIC-rearranged sarcoma. *Hum Pathol* **65**, 225–230.
22 Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr and Kinzler KW (2013) Cancer genome landscapes. *Science* **339**, 1546–1558.
23 Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* **45**, D777–D783.
24 Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N and Stratton MR (2004) A census of human cancer genes. *Nat Rev Cancer* **4**, 177–183.
25 Soussi T and Wiman KG (2015) TP53: an oncogene in disguise. *Cell Death Differ* **22**, 1239–1249.
26 Shang X, Marchioni F, Evelyn CR, Sipes N, Zhou X, Seibel W, Wortman M and Zheng Y (2012) Small-molecule inhibitors targeting G-protein – coupled Rho guanine nucleotide exchange factors. *Proc Natl Acad Sci USA* **110**, 1–6.
27 Diviani D, Raimondi F, Del Vescovo CD, Dreyer E, Reggi E, Osman H, Ruggieri L, Gonano C, Cavin S, Box CL *et al.* (2016) Small-molecule protein-protein interaction inhibitor of oncogenic rho signaling. *Cell Chem Biol* **23**, 1135–1146.
28 Zandvakili I, Lin Y, Morris JC and Zheng Y (2017) Rho GTPases: anti- or pro-neoplastic targets? *Oncogene* **36**, 3213–3222.
29 O'Hayre M, Vázquez-Prado J, Kufareva I, Stawiski EW, Handel TM, Seshagiri S and Gutkind JS (2013) The emerging mutational landscape of G proteins and G-protein-coupled receptors in cancer. *Nat Rev Cancer* **13**, 412–424.
30 Iglesias-Bartolome R, Torres D, Marone R, Feng X, Martin D, Simaan M, Chen M, Weinstein LS, Taylor SS, Molinolo AA *et al.* (2015) Inactivation of a G

[alpha]s-PKA tumour suppressor pathway in skin stem cells initiates basal-cell carcinogenesis. *Nat Cell Biol* **17**, 793–803.
31 Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S *et al.* (2016) The reactome pathway knowledgebase. *Nucleic Acids Res* **44**, D481–D487.
32 Wu G, Feng X and Stein L (2010) A human functional protein interaction network and its application to cancer data analysis. *Genome Biol* **11**, R53.
33 Canisius S, Martens JWM and Wessels LFA (2016) A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol* **17**, 261.
34 Richter J, Schlesner M, Hoffmann S, Kreuz M, Leich E, Burkhardt B, Rosolowski M, Ammerpohl O, Wagener R, Bernhart SH *et al.* (2012) Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *NatGenet* **44**, 1316–1320.
35 Schmitz R, Young RM, Ceribelli M, Jhavar S, Xiao W, Zhang M, Wright G, Shaffer AL, Hodson DJ, Buras E *et al.* (2012) Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature* **490**, 116–120.
36 Muppidi JR, Schmitz R, Green JA, Xiao W, Larsen AB, Braun SE, An J, Xu Y, Rosenwald A, Ott G *et al.* (2014) Loss of signalling via Gα13 in germinal centre B-cell-derived lymphoma. *Nature* **516**, 254–258.
37 Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G *et al.* (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501.
38 Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, Mustonen V, Gonzalez-Perez A, Pearson J, Sander C *et al.* (2015) Pathway and network analysis of cancer genomes. *Nat Methods* **2**, 1–6.
39 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550.
40 Wu G, Dawson E, Duong A, Haw R and Stein L (2014) ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Research* **3**, 146.
41 Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D and Stuart JM (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, 237–245.
42 Adzhubei I, Jordan DM and Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **7** (Suppl 76), 20.
43 Reva B, Antipin Y and Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39**, e118.
44 Betts MJ, Lu Q, Jiang Y, Drusko A, Wichmann O, Utz M, Valtierra-Gutiérrez IA, Schlesner M, Jaeger N, Jones DT *et al.* (2015) Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. *Nucleic Acids Res* **43**, e10.
45 Mosca R, Tenorio-Laranga J, Olivella R, Alcalde V, Céol A, Soler-López M and Aloy P (2015) dSysMap: exploring the edgetic role of disease mutations. *Nat Methods* **12**, 167–168.
46 Gao J, Chang MT, Johnsen HC, Gao SP, Sylvester BE, Sumer SO, Zhang H, Solit DB, Taylor BS, Schultz N *et al.* (2017) 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med* **9**, 4.
47 Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, Lander ES and Getz G (2015) Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci USA* **112**, E5486–E5495.
48 Fujimoto A, Okada Y, Boroevich KA, Tsunoda T, Taniguchi H and Nakagawa H (2016) Systematic analysis of mutation distribution in three dimensional protein structures identifies cancer driver genes. *Sci Rep* **6**, 26483.
49 Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, Masica DL and Karchin R (2016) Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res* **76**, 3719–3731.
50 Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J and Godzik A (2015) A Pan-Cancer catalogue of cancer driver protein interaction interfaces. *PLoS Comput Biol* **11**, e1004518.
51 Raimondi F, Singh G, Betts MJ, Apic G, Vukotic R, Andreone P, Stein L and Russell RB (2016) Insights into cancer severity from biomolecular interaction mechanisms. *Sci Rep* **6**, 34490.
52 Rohde M, Richter J, Schlesner M, Betts MJ, Claviez A, Bonn BR, Zimmermann M, Damm-Welk C, Russell RB, Borkhardt A *et al.* (2014) Recurrent RHOA mutations in pediatric Burkitt lymphoma treated according to the NHL-BFM protocols. *Genes Chromosom Cancer* **53**, 911–916.
53 Miller ML, Reznik E, Gauthier NP, Aksoy BA, Korkut A, Gao J, Ciriello G, Schultz N and Sander C (2015) Pan-cancer analysis of mutation hotspots in protein domains. *Cell Syst* **1**, 197–209.
54 Yang F, Petsalaki E, Rolland T, Hill DE, Vidal M and Roth FP (2015) Protein domain-level landscape of

cancer-type-specific somatic mutations. *PLoS Comput Biol* **11**, e1004147.
55 Peterson TA, Gauran IIM, Park J, Park DH and Kann MG (2017) Oncodomains: a protein domain-centric framework for analyzing rare variants in tumor samples. *PLoS Comput Biol* **13**, e1005428.
56 Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, Peng J, Weile J, Karras GI, Wang Y *et al*. (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660.
57 Kumar P, Henikoff S and Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073–1082.
58 Choi Y and Chan AP (2015) PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747.
59 Tang H and Thomas PD (2016) PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics* **32**, 2230–2232.
60 Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR and Campbell C (2017) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511–513.
61 Schwarz JM, Cooper DN, Schuelke M and Seelow D (2014) MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361–362.
62 Hu H, Huff CD, Moore B, Flygare S, Reese MG and Yandell M (2013) VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol* **37**, 622–634.
63 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM and Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315.
64 Wang K, Li M and Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164.
65 Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D *et al*. (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* **99**, 877–885.
66 Gulko B, Hubisz MJ, Gronau I and Siepel A (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* **47**, 276–283.
67 Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B and Karchin R (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* **69**, 6660–6667.
68 Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA *et al*. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218.
69 Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER *et al*. (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22**, 1589–1598.
70 Gonzalez-Perez A and Lopez-Bigas N (2012) Functional impact bias reveals cancer drivers. *Nucleic Acids Res* **40**, e169.
71 Tamborero D, Gonzalez-Perez A and Lopez-Bigas N (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244.
72 Reimand J and Bader GD (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol* **9**, 637.
73 Masica DL, Douville C, Tokheim C, Bhattacharya R, Kim R, Moad K, Ryan MC and Karchin R (2017) CRAVAT 4: cancer-related analysis of variants toolkit. *Cancer Res* **77**, e35–e38.
74 Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A and Lopez-Bigas N (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* **10**, 1081–1082.
75 Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H and Vilo J (2016) g:profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res* **44**, W83–W89.
76 Gundem G and Lopez-Bigas N (2012) Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types. *Genome Med* **4**, 28.
77 Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD and Morris Q (2013) GeneMANIA prediction server 2013 update. *Nucleic Acids Res* **41** (Web Server issue), W115–W122.