bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Intrinsically disordered protein mutations can drive cancer and their targeted interference extends therapeutic options

Bálint Mészáros[1], Borbála Hajdu-Soltész[1], András Zeke[2], and Zsuzsanna Dosztányi[1,*]
[1]MTA-ELTE Momentum Bioinformatics Research Group, Department of Biochemistry, Eötvös Loránd University, Pázmány Péter stny 1/c, Budapest, H-1117 Hungary
[2]Protein Research Group, Institute of Enzymology, RCNS, HAS, Budapest PO Box 7, H-1518, Hungary
[*]Corresponding author: dosztanyi@caesar.elte.hu

Abstract

Intrinsically disordered regions (IDRs) are important functional modules of several proteins, providing extra layers of regulation as switchable structural elements. Protein disorder has been associated with cancer, but it is unknown whether IDR mutations represent a distinct class of driver events associated with specific molecular mechanisms and system level properties, which would require dedicated targeting strategies. Based on an integrative computational approach, we identified 47 IDRs whose genetic mutations can be directly linked to cancer development. While not as common as alterations of globular domains, IDR mutations contribute to the emergence of the same cancer hallmarks through the modulation of distinct molecular mechanisms, increased interaction potential and specific functional repertoire. We demonstrate that in specific cancer subtypes, IDR mutations represent the key events driving tumorigenesis. However, treatment options for such patients are currently severely limited. We suggest targeting strategies that could enable successful therapeutic intervention for this subclass of cancer drivers, extending our options for personalized therapies.

Keywords

intrinsically disordered regions; protein modules; molecular switches; cancer genomics; driver gene identification; drug targeting options; receptor tyrosine kinases; short linear motifs; protein degradation; enrichment; cancer hallmarks

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Tumorigenic variations at the genome level manifest in changes in protein structure, availability, localization, turnover, or stability. The structural and functional properties of the affected proteins determine their oncogenic or tumor suppressor roles, which, in the case of context-dependent genes, can also depend on tissue type or the stage of tumor progression. Understanding how these tumorigenic roles emerge from specific mutations is essential for subsequent drug development efforts. Genetic variations have been collected for tens of thousands of human cancer incidences and can be accessed via public repositories, such as TCGA (https://cancergenome.nih.gov/), or the COSMIC database, incorporating data from targeted studies as well(Tate et al., 2019). These data revealed that cancer samples are extremely heterogeneous both in terms of the number and type of genetic alterations. However, various patterns start to emerge when these samples are analyzed in combination(Cancer Genome Atlas Research Network et al., 2013), enabling the identification of cancer driving genes that are frequently mutated in specific types of cancer(Lawrence et al., 2014, 2013), highlighting biological processes/pathways that are commonly altered in tumor development(Ali and Sjöblom, 2009; Copeland and Jenkins, 2009) and traits that govern tumorigenic transformation of cells(Hanahan and Weinberg, 2011). Recent analyses estimated the number of driver genes to be in the low to mid-hundreds(Bailey et al., 2018), but this number could increase with growing number of sequenced cancer genomes(Lawrence et al., 2014).

Recent computational methods can not only identify cancer genes, but also highlight specific functional modules that are critical for tumorigenesis(Buljan et al., 2018; Mészáros et al., 2016; Porta-Pardo and Godzik, 2014; Tamborero et al., 2013), identifying oncogenes that are altered via activating mutations, but also the majority of tumor suppressors that are typically deactivated by truncating mutations(Buljan et al., 2018; Mészáros et al., 2016). The positional accumulation of mutations within specific protein regions has been analyzed for structures, domains, or interactions surfaces(Engin et al., 2016; Kamburov et al., 2015; Porta-Pardo et al., 2015; Tokheim et al., 2016; Yang et al., 2015). However, a sizeable portion of human proteins corresponds to protein regions that function without inherent structure(Dyson and Wright, 2005; van der Lee et al., 2014). These intrinsically disordered proteins/regions (IDPs/IDRs) are predicted to represent around 30% of all residues of human proteins(Ward et al., 2004) and are also important for the interpretation of the effect of various disease mutations(Vacic et al., 2012).

The lack of an inherent structure has a profound effect on the way IDPs carry out their functions(Tompa, 2002; van der Lee et al., 2014). IDPs can act as flexible linkers or entropic chains, directly exploiting their conformational heterogeneity. IDPs also often interact with other biomolecules with interaction sites that are usually short and linear. These regions serve as recognition sites for specific protein domains(Davey et al., 2012) or nucleic acids(Staby et al., 2017), be sites of post-translational modifications(Darling and Uversky, 2018), can harbour localization signals(Eisenhaber and Eisenhaber, 2007), or mediate oligomerization modulating protein stability and functionality(Faust et al., 2014). Other types of functional IDRs cover domain sized regions, serving as assembly sites for larger complexes(Cortese et al., 2008). In general, IDPs are core components of interaction networks and fulfill critical roles in regulation and signaling(Wright and Dyson, 2015). In accord with their crucial functions, IDPs are often

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

associated with various diseases(Babu et al., 2011), in particular with cancer. IDRs can be integral parts of both oncogenes – such as β-catenin(Morin et al., 2016) – and tumor suppressors – such as p53(Olivier et al., 2010). The prevalence of protein disorder among cancer-associated proteins was also observed at a more general level(Iakoucheva et al., 2002). A direct link between protein disorder and cancer was suggested in the case of two common forms of generic alterations; chromosomal rearrangements(Hegyi et al., 2009) and copy number variations(Vavouri et al., 2009). In contrast, a recent analysis found that cancer-associated missense mutations had a preference for ordered regions, and suggested that the association between protein disorder and cancer could be indirect(Pajkos et al., 2012). Nevertheless, a direct link between disordered regions and mutations were also suggested either through the abolishment(Uyar et al., 2014) or the creation(Meyer et al., 2018) of IDR-mediated interactions, but only in a few cases. In general, the extent to and the mechanisms through which disordered protein regions directly drive cancer are still largely unexplored.

The identification of functional modules that are directly altered in cancer driver genes can serve with potential targets for pharmaceutical intervention. Most current anticancer drugs are inhibitors designed against enzyme activity (using either competitive or noncompetitive inhibition)(Griffith et al., 2010; Pathania et al., 2018; Scatena et al., 2008). In general, currently successful drug development efforts mainly focus on ordered protein domains, in the framework of structure-based rational drug design(Lounnas et al., 2013). However, IDPs can potentially offer new directions for cancer therapeutics(Kulkarni, 2016). Currently tested approaches include the direct targeting of IDPs by specific small compounds, or blocking the globular interaction partner of IDPs(Metallo, 2010; Neira et al., 2017). However, the direct targeting of IDPs requires radically different molecular strategies, and such approaches have yet to reach maturity.

In this work we analyzed cancer mutations from genome wide screens and targeted studies(Forbes et al., 2016) to identify significantly mutated protein regions(Mészáros et al., 2016), and classify them into ordered and disordered regions integrating experimental structural knowledge and prediction. Automated and high-quality manually curated information was gathered for the collected examples to gain better insights into their functional and network properties, and their roles in tumorigenic processes. We aimed to answer the following questions: What are the characteristic molecular mechanisms, biological processes, and protein-protein interaction network roles associated with proteins mutated at IDRs? And at a more generic level: how fundamental is the contribution of IDPs to tumorigenesis? Are IDP mutations just accessory events, or can they be the main, or even the sole molecular background to the emergence of cancer? How much can we gain by their systematic targeting efforts, which are the cancer types in which IDR mutations should be considered for large therapeutic gain, and finally, how should we select our targeting approaches for specific IDPs?

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

1. Disordered protein modules are targets for tumorigenic mutations

Protein disorder is an integral part of many cancer driver proteins and is preferentially mutated in context-dependent genes

Here we used an integrated approach to define ordered and disordered functional modules in all human proteins, merging experimental annotations and structural predictions (see Data and methods). The majority of human proteins (Supplementary Figure S1) are modular with almost 70% of all proteins incorporating at least one disordered module. Module numbers were assessed separately for the census of cancer driver genes collected from the COSMIC database(Sondka et al., 2018) and the literature(Vogelstein et al., 2013). All census drivers were manually characterized as tumor suppressors, oncogenes and context-dependent cancer genes based on the literature (Supplementary Table S1). Census drivers have a higher average modularity, and contain a higher average number of disordered modules compared to all human proteins (Figure 1A, Supplementary Figure S2), and this modularity is even higher for context-dependent genes. The reported values are likely to be still conservative estimates of the true number of modules in proteins, as individual regions could contain yet uncharacterized tandem modules, especially in the case of disordered regions.

The abundance of IDRs in cancer drivers hint at the importance of disorder for the normal function of these proteins, but is not necessarily reflected in the distribution of cancer mutations. To analyze this, we collected exomic cancer mutations with a clearly localized effect from the COSMIC database (see Data and methods). Figure 1B shows that exomic mutations preferentially target ordered sequence regions in drivers, in line with earlier observations(Pajkos et al., 2012). This preference is most pronounced for oncogenes, for which only less than 10% of known mutations fall into IDRs. In contrast, the fraction of mutations within IDRs shows a three-fold increase in tumor suppressor genes (17%) and reaches 47% in the case of context-dependent genes. The trends remain the same using whole exome sequencing data from TCGA (Supplementary Figure S3, Data and methods).

Several cancer drivers are modulated through disordered regions

In order to explore the possible causal relationship between tumorigenesis and structural properties, we used iSiMPRe(Mészáros et al., 2016) on the pre-filtered mutation data from COSMIC to identify specific regions in proteins that are directly involved in cancer development (see Data and methods). By restricting our analysis to high-confidence cases, we identified 178 ordered and 47 disordered driver regions in 145 proteins from the human proteome (Supplementary Table S2, Figure 1C). The structural status of these regions were confirmed by manual curation. The identified driver regions typically represent compact modules, usually not

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

covering more than 10% or 20% of the sequences in the case of oncogenes and tumor suppressors, respectively (Supplementary Figure S4). While oncogenes and tumor suppressors both incorporate disordered driver regions, these regions are most common in context-dependent drivers (Figure 1C).

According to the 20/20 rule, true oncogenes are recognizable from mutation patterns, having a higher than 20% fraction of missense point mutations in recurring positions (termed the oncogene score(Vogelstein et al., 2013)). In contrast, tumor suppressors have lower oncogene scores, and predominantly contain truncating mutations. Figure 1D shows that the 20/20 rule holds true for the vast majority of the identified region-harboring oncogenes and context-dependent genes. Figure 1D also shows the oncogene scores calculated from the identified regions alone. Despite their short relative lengths (Supplementary Figure S4), the driver regions are the main source of the oncogenic effect in almost all cases. While most drivers contain both ordered and disordered modules, oncogenesis is typically mediated through either ordered or disordered mutated regions. This effectively partitions cancer drivers into ‘ordered drivers’ and ‘disordered drivers’, regardless of the exact structural composition of the full protein. Thus, IDRs are not only essential components of drivers, but the direct modulation of these regions is heavily utilized in the emergence of cancer.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

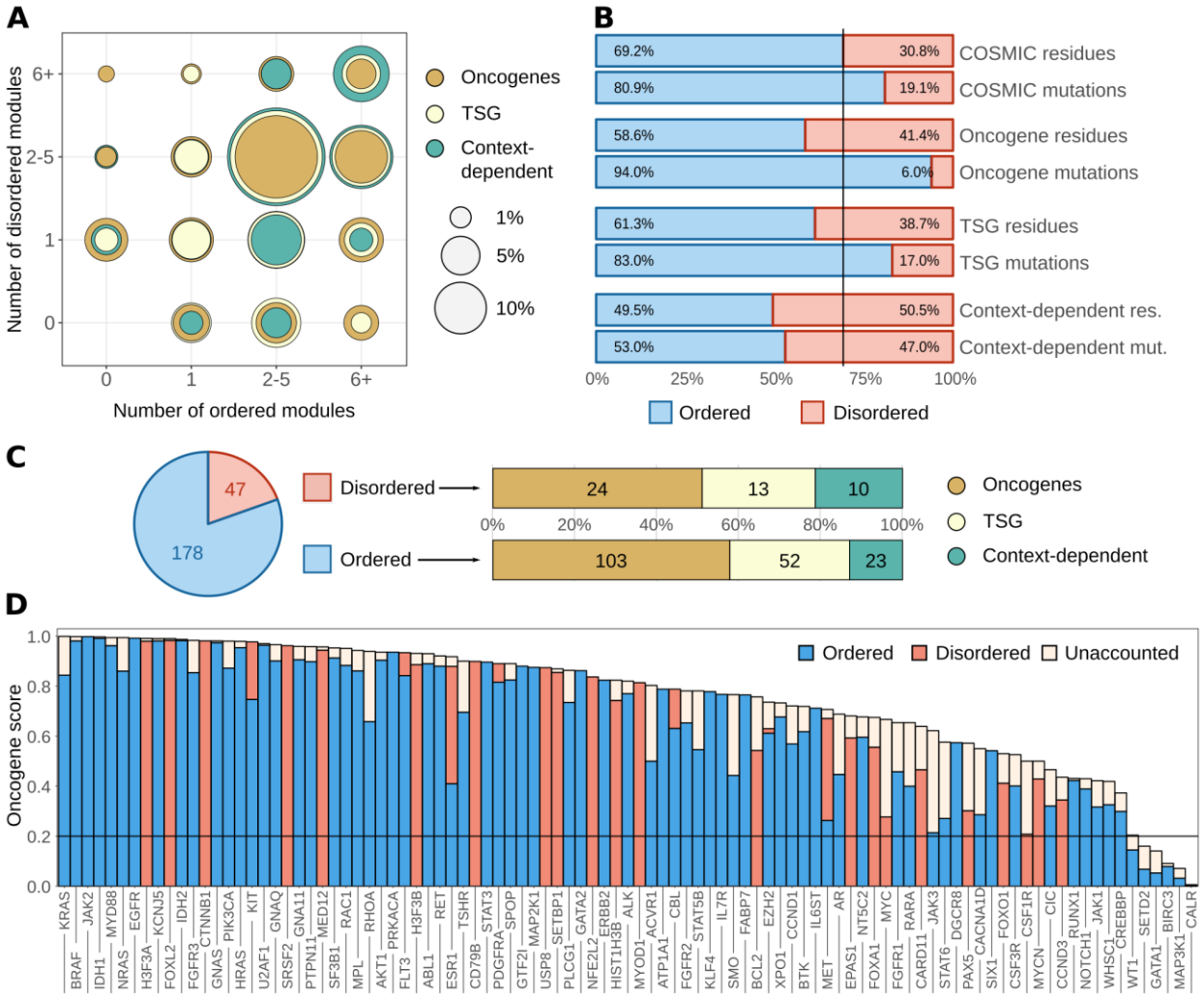


Figure 1. Occurrence of protein disorder at the level of protein classes, modules and in terms of mutated sites. A: The distribution of all census cancer drivers with regard to the structural states of constituent functional modules. B: The distribution of residues and local exomic mutations (missense/inframe indel mutations) from COSMIC in the human proteome and various cancer driver classes. The vertical line marks the ratio of ordered and disordered residues in the human proteome, corresponding to the expected ratio of randomly occurring mutations. C: The distribution of ordered and disordered driver protein regions. D: Oncogene scores of full genes and oncogene scores explained by the identified regions in oncogenes and context-dependent driver genes. 'Unaccounted' corresponds to the fraction of mutations not in the identified, high significance regions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2. Disordered drivers function via distinct molecular mechanisms

Disordered drivers employ distinctive molecular mechanisms of action

Nearly all these proteins have already been identified as potential cancer drivers in the literature, and available structural and functional information can highlight the possible mechanisms of action altered in cancer (Figure 2, Supplementary Table S3), even though the information is often incomplete.

Several of the identified highly mutated disordered regions correspond to **linear motifs**, including sites for protein-protein interactions (e.g. hUBPy [corresponding gene: USP8], forkhead box protein O1 [FOXO1] and ER-α [ESR1]), degron motifs that regulate the degradation of the protein (e.g β-catenin [CTNNB1], cyclin-D3 [CCND3] and CSF-1R [CSF1R]) and localization signals (e.g. $p14^{ARF}$ [CDKN2A] and BAF47 [SMARCB1]). However, other types of disordered functional modules can also be targeted by cancer mutations. **IDRs with autoinhibitory roles** (e.g. modulating the function of adjacent folded domains) are represented by EZH2 [EZH2], a component of the polycomb repressive complex 2. While the primary mutation site in this case is located in the folded SET domain, cancer mutations are also enriched within the disordered C-terminus that normally regulates the substrate binding site on the catalytic domain. Another category corresponds to **regions involved in DNA and RNA binding**. The highly flexible C-terminal segment of the winged helix domain is altered in the case of HNF-3-α [FOXA1], interfering with the high affinity DNA binding. For the splicing factor hSNF5 [SRSF2], mutations affect the RNA binding region (Figure 2).

Larger functional disordered modules, often referred to as **intrinsically disordered domains** (IDDs), can also be the primary sites of cancer mutations. Mutated IDDs exhibit varied structural and sequence features. In pVHL [VHL], the commonly mutated central region adopts a molten globule state in isolation(Sutovsky and Gazit, 2004). The mutated region of APC [APC] incorporates several repeats containing multiple linear motif sites, which are likely to function collectively as part of the β-catenin destruction complex(Aoki and Taketo, 2007). In calreticulin [CALR], cancer mutations alter the C-terminal domain-sized low complexity region, altering Ca^{2+} binding and protein localization(Elf et al., 2016).

Linker IDRs, not diredtly involved in molecular interactions, are also frequent targets of cancer mutations. The juxtamembrane region located between the transmembrane segment and the kinase domain of Kit [KIT] and related kinases, are the main representatives of this category. Similarly, the regulatory linker region connecting the substrate- and the E2 binding domains is one of the dominant sites of mutations in the case of the E3 ubiquitin ligase c-Cbl [CBL].

One of the recurring themes among cancer-related IDP regions is the formation of molecular switches (Supplementary Table S3). The most commonly occurring switching mechanism involves various post-translational modifications (PTMs), including serine or threonine phosphorylation (e.g. cyclin-D3 [CCND3], c-Myc [MYC] and APC [APC]), tyrosine phosphorylation (e.g. c-Cbl [CBL], CD79b [CD79B], and CSF-1R [CSF1R]), methylation (e.g. histone H3s [H3F3A/H3F3B/HIST1H3B]) or acetylation (e.g. ER-α [ESR1]). An additional way of forming

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

molecular switches involves overlapping functional modules (Figure 2). In the case of BAF47 [SMARCB1], the mutated inhibitory sequence is likely to normally mask a nuclear export signal in the autoinhibited state (Craig et al., 2002). In the case of Pax-5 [PAX5], the mutated flexible linker region is also involved in the high affinity binding of the specific DNA binding site(Garvie et al., 2001). Cancer mutations of the bZip domain of C/EBP α [CEBPA] disrupt not only the DNA binding function, but the dimerization domain as well(Paz-Priel and Friedman, 2011). In addition to their linker function, the juxtamembrane regions of kinases are also involved in autoinhibition and trans-phosphorylation, regulating degradation and downstream signaling events(Hubbard, 2004; Li and Hristova, 2010).

Different types of disordered drivers are mutated with specific mutational mechanisms

Mutational patterns are strongly associated with the role of the perturbed functional modules (see Figure 2 and online visualization links in Supplementary Table S3). Short linear motifs are typically mutated in a few key positions, often only affecting a single PTM site that plays a key role in regulating the interaction. IDDs generally show more distributed mutational patterns, in accord with their larger sizes. These regions can also be targeted by truncating mutations, which affect only the specific region, such as in the case of calreticulin [CALR] and APC [APC]. The most common mutation type for linker regions involves in-frame insertions and deletions, which alter the length of the region. These types of mutations are also common for multifunctional modules that have autoinhibitory functions and serve as linkers as well.

The collected examples of disordered regions mutated in cancer cover both oncogenes and tumor suppressors, as well as context-dependent genes. There is a slight tendency for tumor suppressors to be altered via longer functional modules, such as IDDs. Nevertheless, with the exception of linkers in tumor suppressors and IDDs in context-dependent genes, every other combination occurs even within our limited set.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

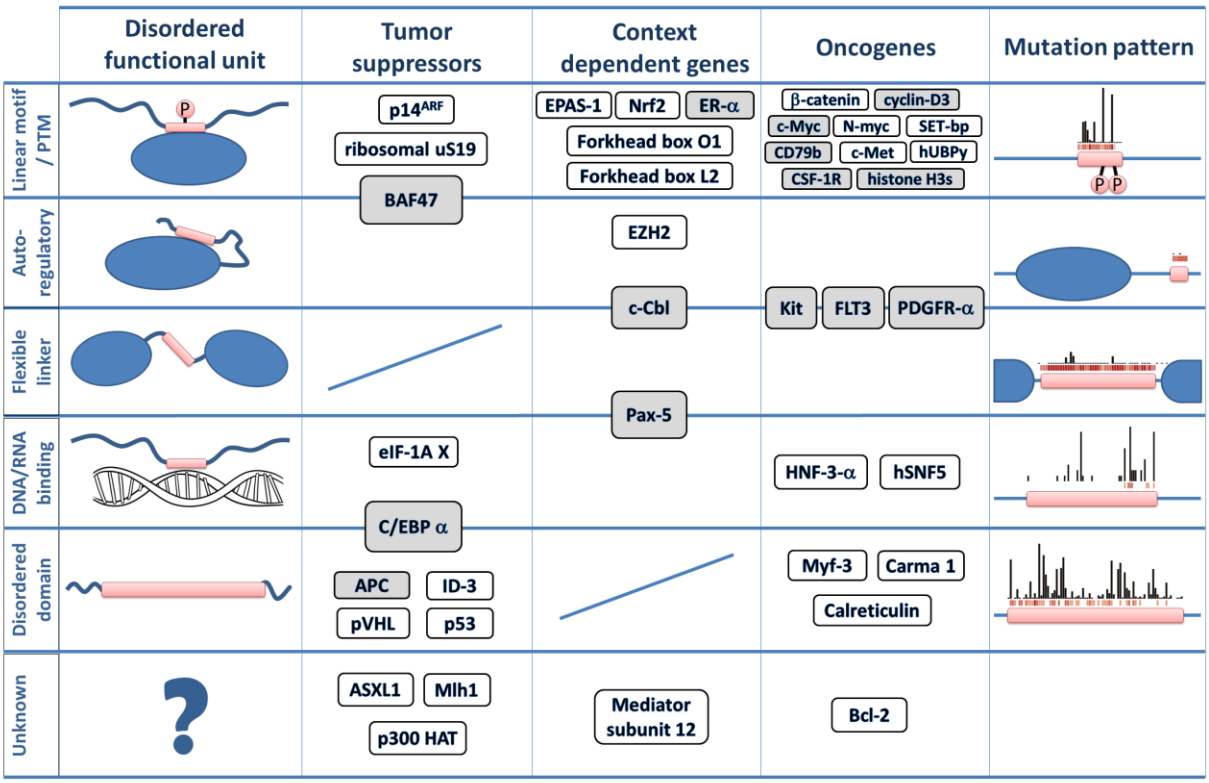


Figure 2. IDP regions mutated in cancer. The classification of identified disordered cancer drivers. Protein names in gray indicate known switching mechanisms either via PTMs or overlapping functions. In protein architecture schematics, blue ovals represent folded domains, blue lines disordered regions and red rectangles represent disordered driver modules. Mutation patterns show the typical distribution of missense mutations (black bars) an indels (red boxes). For detailed mutation profiles for each gene, see online visualization links in Supplementary Table S3.

3. Disordered mutations give rise to cancer hallmarks by targeting central elements of biological networks

Disordered drivers integrate biological processes through their increased interaction capacities

Almost all of the analyzed IDRs are involved in binding to a molecular partner, even some of the linkers owing to their multifunctionality. Therefore, we analyzed known protein-protein interactions of ordered and disordered cancer drivers in more detail (see Data and methods). Our results indicate that both sets of drivers are involved in a large number of interactions, and show increased betweenness values compared to average values of the human proteome, and even compared to the direct interaction partners of cancer drivers (Figure 3A). However, this trend is even more pronounced for disordered drivers. The elevated interaction capacity could also be detected at the level of molecular function annotations using Gene Ontology (see Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

table S4 and Data and methods). Figure 3B shows the average number of types of molecular interaction partners for both disordered and ordered drivers, contrasted to the average of the human proteome. The main interaction partners are similar for both types of drivers, often binding to nucleic acids, homodimerizing, or binding to receptors. However, disordered drivers are able to physically interact with a wider range of molecular partners, and are also able to more efficiently interact with RNA and the effector enzymes of the post-translational modification machinery. This, in particular, can offer a way to more easily integrate and propagate signals through the cell, relying on the spatio-temporal regulation of interactions via previously demonstrated switching mechanisms (Supplementary Table S3).

The high interaction capacity and central position of disordered drivers allows them to participate in several biological processes. The association between any two processes can be assessed by quantifying the overlap between their respective protein sets (see Data and methods). We analyzed the average overlap between various processes using a set of non-redundant human-related terms of the Gene Ontology (Supplementary Table S5). The average overlap of proteins for two randomly chosen processes is 0.15%, showing that as expected, in general, biological processes utilize characteristically different gene/protein sets. Restricting proteins to the identified drivers, and only considering processes connected to at least one of them, the average overlap between processes is increased to 3.00% for ordered drivers and 5.80% for disordered drivers (Figure 3C). This shows that the integration of various biological processes is a distinguishing feature of cancer genes in general, and for disordered drivers in particular; and that IDPs targeted in cancer are efficient integrators of a wide range of processes.

Disordered drivers employ characteristic molecular toolkits

Disordered and ordered drivers can employ different molecular mechanisms in order to fulfill their associated biological processes. To quantify these differences, we assembled a set of molecular toolkits integrating Gene Ontology terms (see Data and methods and Supplementary Table S6). Based on this, we calculated the enrichment of each molecular toolkit for both disordered and ordered drivers, in comparison with the full human proteome, highlighting enriched and possibly driver class-specific toolkits (Figure 3D). Receptor activity is the most enriched function for both types of drivers, owing at least partially to the fact that receptor tyrosine kinases can often be modulated via both ordered domains and IDRs (Figure 1 and Figure 2). In contrast, the next three toolkits enriched for disordered drivers are highly characteristic of them. These are gene expression regulation, the modulation of DNA structural organization, and the degradation of proteins, mainly through the ubiquitin-proteasome system. In addition, RNA processing, translation and folding is also highly characteristic of disordered drivers; and while this toolkit is not highly enriched compared to the human proteome in general, ordered drivers are almost completely devoid of this toolkit.

Modulation of disordered proteins presents an independent tumorigenic mechanism for all hallmarks of cancer

There are ubiquitously displayed features of cancer cells, often described as the ten hallmarks of

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cancer(Hanahan and Weinberg, 2011). In order to quantify the contribution of drivers to each hallmark, we manually curated sets of biological process terms from the Gene Ontology that represent separate hallmarks (see Data and methods and Supplementary Table S7). Enrichment analysis of these terms shows that all hallmarks are significantly over-represented in census cancer drivers compared to the human proteome (Supplementary Figure S5A), serving as a proof-of-concept for the used hallmark quantification scheme. Furthermore, comparing drivers with identified regions to all census cancer drivers shows a further enrichment (Supplementary Figure S5B), indicating that the applied region identification protocol of iSiMPRe is able to pick up on the main tumorigenic signal by pinpointing strong driver genes. Separate enrichment calculations for ordered and disordered drivers show that despite subtle differences in enrichments, in general, all ten hallmarks are over-represented in both driver groups (Figure 3E). This indicates that while the exact molecular mechanisms through which ordered domain and IDR mutations contribute to cancer are highly variable, both types of genetic modulation can give rise to all necessary cellular features of tumorigenic transformation. Hence, IDR mutations provide a mechanism that is sufficient on its own for cancer formation.

Protein, network, and cellular level attributes provide complementary information on disordered drivers

In the previous sections we elucidated the characteristic features of disordered drivers at various levels. These include five major features: the molecular mechanisms (1) at the protein level (Figure 2); the number and type of interaction partners (2), and the molecular toolkits (3) at the network/pathway level (Figure 3); the associated hallmarks (4) (Figure 3), and the tumorigenic character of the protein at the cellular level (5) (Figure 2 and Supplementary Table S3). These features provide information on the drivers at very different levels, but in theory, they might be associated in non-trivial ways. Such associations could be used to establish disease subtypes, or might direct targeting efforts. For example, a hypothetical association between SLiM mutations and the 'Inducing angiogenesis' hallmark would mean, that counteracting SLiM mutations would be a generic approach to solid tumor treatments. However, no strong association was found when calculating the common information content encoded in the five main features (see Data and methods, Supplementary Table S8 and Supplementary Table S9). This indicates that these five aspects of disordered drivers are highly complementary, and potentially all of them need to be considered for determining efficient targeting sites.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

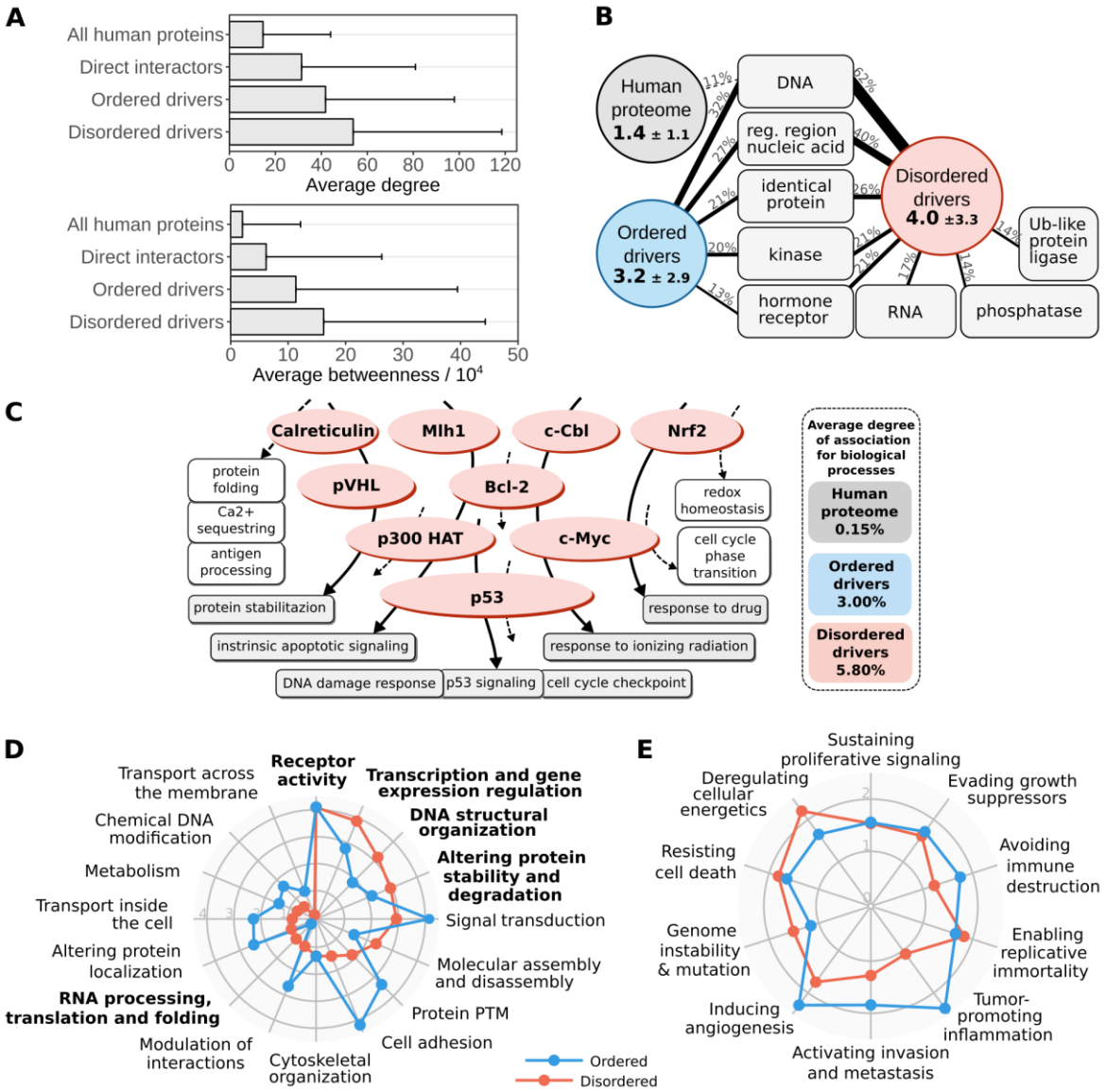


Figure 3. Characteristics of cancer drivers at the network/pathway-, and cellular levels. A: Average degree (top) and betweenness (bottom) of all human proteins, the direct interaction partners of drivers, ordered drivers, and disordered drivers. B: the average occurrence of various types of interaction partners for the whole human proteome (grey circle), ordered drivers (blue circle) and disordered drivers (red circle). Values in circles show the average number of types of interactions together with standard deviations. The most common interaction types are shown in grey boxes with connecting lines showing the fraction of proteins involved in that binding mode. Only interaction types present for at least 1/8th of proteins are shown. C right: average values of overlap between protein sets of various biological processes, considering the full human proteome (grey), ordered drivers (blue) and disordered drivers (red). Left: an example subset of disordered drivers with associated biological processes marked with arrows (dashed and solid arrows marking processes involving only one, or several disordered drivers). Process names in grey represent processes that involve at least two disordered drivers, names in white boxes mark other processes attached to disordered drivers. D: Overrepresentation of molecular toolkits defined based on GO terms for ordered (blue) and disordered (red) drivers, compared to the average of the whole human proteome. E: Overrepresentation of hallmarks of cancer for ordered (blue) and disordered (red) drivers compared to all census drivers.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4. Disordered drivers are important players in cancer at the patient sample level

Cancer incidences can arise through disordered drivers

Using whole-genome sequencing data from TCGA we assessed the role of the identified drivers at the patient level. 10,197 tumor samples, containing over three and a half million genetic variations were considered, to delineate the importance of disordered drivers at the sample level across the 33 cancer types covered in TCGA. In driver region identification we only considered mutations with a local effect (missense mutations and in frame indels), which naturally yielded only a restricted subset of all true drivers. However, in patient level analyses, we also considered other types of genetic alterations of the same gene, in order to get a more complete assessment of the alteration of identified driver regions per cancer type (see Data and methods).

In spite of the incompleteness of the identified set of driver genes, we still found that on average about 80% of samples contains genetic alterations that affect at least one identified ordered or disordered driver region. Thus, the identified regions are able to describe the main players of tumorigenesis at the molecular level (Figure 4A). While at the protein level typically either ordered or disordered regions are modulated (Figure 1E), at the patient level most samples show a mixed structural background, most notably in colorectal cancers (COAD and READ). Some cancer types, however, show distinct preferences for the modulation of a single type of structural element: for thyroid carcinoma (THCA) or thymoma (THYM) the molecular basis is almost always the exclusive mutation of ordered protein regions. At the other extreme, the modulation of disordered regions is enough for tumor formation in a considerable fraction of cases of liver hepatocellular, adrenocortical, and renal cell carcinomas, together with diffuse large B-cell lymphoma (LIHC, ACC, KIRC and DLBC). These results, in line with the previous hallmark analyses, show that IDR mutations can constitute a complete set of tumorigenic alterations. Hence, there are specific subsets of patients that carry predominantly, or exclusively disordered driver mutations in their exome.

Disordered drivers display characteristic cancer-type specificity

Whole genome sequencing data was also used to assess the cancer type specificity of disordered drivers (Figure 4B). Nearly all studied cancer types have at least one disordered driver that is mutated in at least 1% of cases, with the exception of thyroid carcinoma (THCA). As such, disordered mutations play important roles in almost all cancer types, but in a highly heterogeneous way. There are only four disordered drivers that can be considered as generic drivers, being mutated in a high number of cancer types. p53 presents a special case in this regard, as it is the main tumor suppressor gene in humans, and thus is most often affected by gene loss or truncations affecting a large part of the protein. These alterations abolish the function of both the ordered and disordered driver regions at the same time (the DNA-binding domain and

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the tetramerization region). In contrast, the other three generic disordered drivers are predominantly altered via localized mutations in their disordered regions: the degrons of β-catenin and NRF2, and the central region of APC, and hence these are true disordered drivers. While the four generic drivers are commonly mutated in several cancer types, these cases are the exception. The majority of disordered drivers show a high degree of selectivity for tumor types, being mutated only in very specific cancers. This specificity is strongly connected to the tumorigenic roles of disordered drivers (Figure 4C). Considering 1% of patient samples as the cutoff, tumor suppressors are typically implicated in a broad range of cancer types, while oncogenes on average show a high cancer type specificity. Context-dependent disordered drivers are often mutated in only a very restricted set of cancers.

Mutation data collected in TCGA corresponds to relatively broad classes of tumor types, but does not include results of targeted studies corresponding to several rarer cancers or more specific cancer subtypes. However, for several tumor subclasses, including both malignant and benign cases, mutations in a specific disordered driver is the main, or one of the main driver events (Table 1). In the case of these tumor types, targeting disordered regions can have a potentially huge treatment advantage, and in many cases, the counteraction of these IDR mutations may be the only viable therapeutic strategy.

Cancer incidences arising through disordered drivers lack effective drugs

Next, we addressed how well disordered drivers are targetable by current FDA approved drugs, as collected by the OncoKB database(Chakravarty et al., 2017). This database currently contains 83 FDA-approved anticancer drugs, either as part of standard care or efficient off-label use (see Data and methods). These drugs have defined exome mutations that serve as indications for their use. The majority of these drugs target ordered domains, mostly inhibiting kinases. Currently only 7 drugs are attached to disordered region mutations, which correspond to only four sites in FGFR and c-Met. These drugs act indirectly, targeting ordered kinase domains, to counteract the effect of the listed activating disordered mutations.

This represents a clear significant negative treatment option bias against patients whose tumor genomes contain disordered drivers. Considering all mutations in patient samples gathered in TCGA, the fraction of disordered driver mutations actually serves as an indicator of whether there are suitable drugs available. Patients with mostly ordered driver mutations have a roughly 50% chance that an FDA-approved drug can be administered with expected therapeutic effect. This chance drops to 10% for patients with predominantly disordered mutations (Figure 4D). Thus, incidences of cancer arising through disordered driver mutations are currently heavily under-targeted, highlighting the need for efficient targeting strategies for IDP driven cancers.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

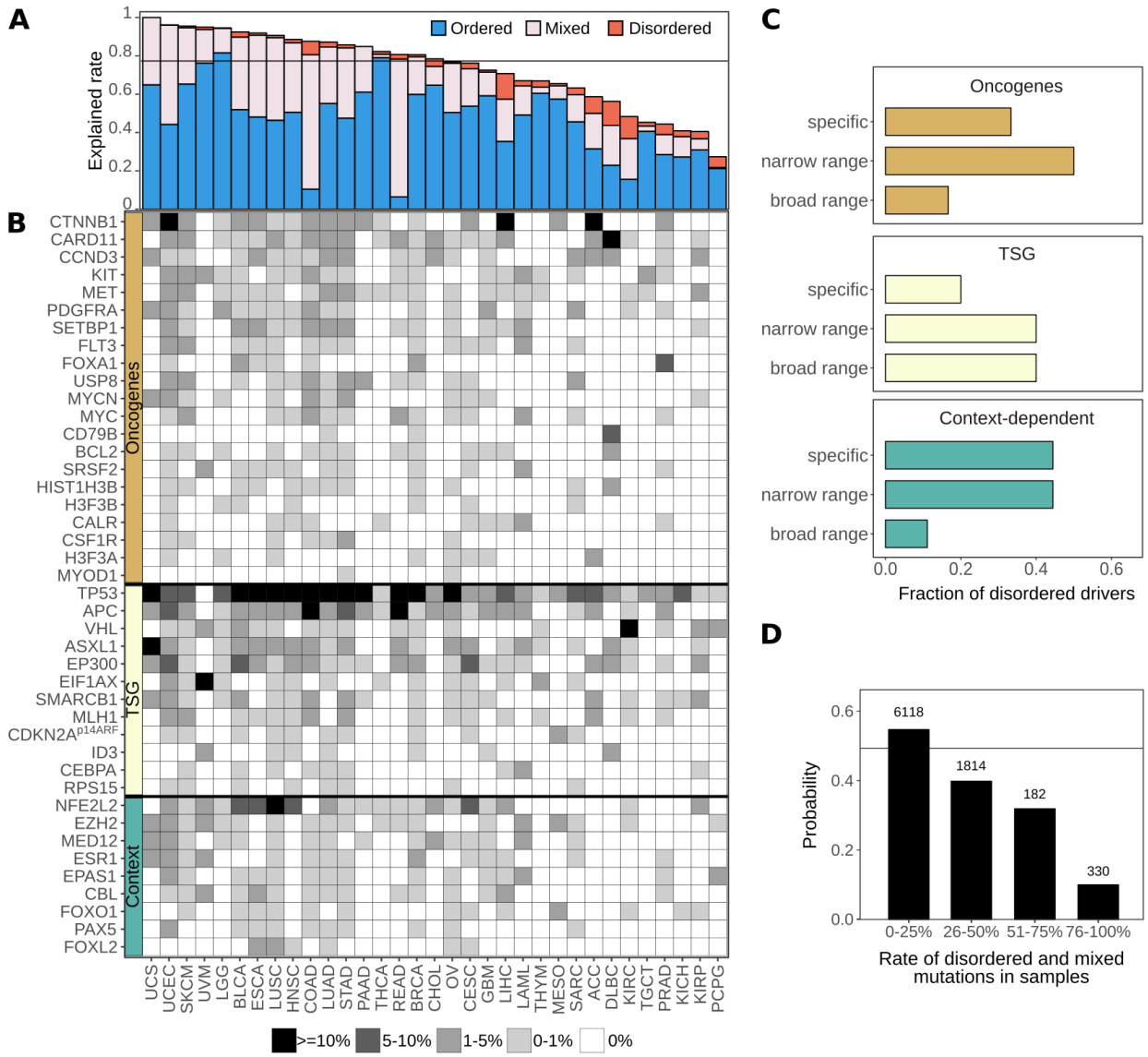


Figure 4: Therapeutic options for targeting disordered drivers. A: Fraction of samples that contain altered driver genes per cancer types. Samples can contain mutations affecting only ordered drivers (blue), only disordered drivers (red) or both (mixed - gray). B: Percentage of cancer samples, grouped by cancer types, containing genetic alterations that target the identified disordered driver regions. C: The distribution of disordered drivers from the three classes of cancer genes categorized into specific, narrow and broad range, based on the frequency of samples they are mutated in (see Data and methods). D: The probability of having an available FDA approved drug for at least one mutation-affected gene for patients, as a function of the ratio of affected disordered genes compared to all mutated genes in the sample. The horizontal black line represents the total fraction of targetable samples (0.49) from 8,444 samples.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tumor type (name)	Gene implicated	Malignancy	Incidence	Reference
Diffuse large B-cell lymphoma (ABC subtype)	CARD11	malignant	9.6-10.8% (7/73, 4/37)	(Compagno et al., 2009; Lenz et al., 2008)
Burkitt lymphoma	CCND3	malignant	14.6% (6/41)	(Schmitz et al., 2012)
Diffuse large B-cell lymphoma (ABC subtype)	CCND3	malignant	10.7% (3/28)	(Schmitz et al., 2012)
Diffuse large B-cell lymphoma (PCNS subtype)	CD79B	malignant	31.6% (6/19)	(Zheng et al., 2017)
Acute myeloid leukaemia	CEBPA	malignant	15% (16/104)	(Lin et al., 2005)
Myelodysplasia and acute myeloblastic leukemia	CSF1R	malignant	12.7% (14/110)	(Ridge et al., 1990)
Endometrioid endometrial carcinoma (low-grade)	CTNNB1	malignant	87.0% (47/54)	(Liu et al., 2014)
Ovarian endometrioid carcinomas (low-grade)	CTNNB1	malignant	53.3% (16/30)	(McConechy et al., 2014)
Hepatocellular carcinoma (HBV/HCV related)	CTNNB1	malignant	26% (32/122)	(Pezzuto et al., 2016)
Desmoid tumor	CTNNB1	benign	73% (106/145)	(Mullen et al., 2013)
Juvenile nasopharyngeal angiofibroma	CTNNB1	benign	75% (12/16)	(Mishra et al., 2016)
Paraganglioma	EPAS1	possibly malignant	17% (7/41)	(Comino-Méndez et al., 2013)
Adult granulosa cell tumors of the ovary	FOXL2	malignant	93-97% (52/56, 86/89)	(Jamieson et al., 2010; Shah et al., 2009)
Pediatric anaplastic astrocytoma / glioblastoma	H3F3A	malignant	17.9-27.1% (5/28, 35/129)	(Gielen et al., 2013)
Giant cell tumor of bone (stromal cell)	H3F3A	benign	92% (49/53)	(Behjati et al., 2013)
Chondroblastoma (stromal cell)	H3F3B	benign	95% (73/77)	(Behjati et al., 2013)
GIST	KIT	malignant	47% (57/121)	(Xu et al., 2014)
Extrauterine leiomyoma and leiomyosarcoma	MED12	(possibly) malignant	19% (6/32)	(Ravegnini et al., 2013)
Phyllodes tumor of breast	MED12	possibly malignant	49% (41/83)	(Laé et al., 2016)
Uterine leiomyoma	MED12	benign	70% (159/225)	(Mäkinen et al., 2011)
Rhabdomyosarcoma	MYOD1	malignant	20% (10/49)	(Rekhi et al., 2016)
Esophageal squamous cell carcinoma	NFE2L2	malignant	9.6% (47/490)	(Du et al., 2017)
B-cell progenitor acute lymphoblastic leukemia	PAX5	malignant	34-39% (40/117, 94/242)	(Familiades et al., 2009; Mullighan et al., 2007)
Chronic myelomonocytic leukemia	SETBP1	malignant	25% (14/56)	(Ouyang et al., 2017)

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Atypical Chronic Myeloid Leukemia	SETBP1	malignant	24.3% (17/70)	(Piazza et al., 2013)
Chronic myelomonocytic leukaemia	SRSF2	malignant	47% (129/275)	(Meggendorfer et al., 2012)
Pituitary adenoma	USP8	possibly malignant	14% (6/42)	(Ballmann et al., 2018)

Table 1: Cancer types with mutation incidence rates around or above 10% in the disordered driver gene of interest per total patients studied.

5. Targeting strategies can be developed for disordered drivers by considering protein features

Disordered drivers can be targeted by indirect strategies

To date, the most viable targeting approaches are enzyme inhibition(Griffith et al., 2010; Pathania et al., 2018; Scatena et al., 2008), which yielded several FDA-approved anti-cancer drugs, and protein-protein interaction (PPI) disruption, which yielded several compounds with great therapeutic promise(Ivanov et al., 2013). However, drugs directly targeting disordered regions are still lacking. Here we examined the strategy of indirect targeting, ie. compensating for disordered driver mutations with small molecule/peptide based inhibitors against ordered domains, developed based on existing strategies. Through a few selected examples, we analyzed how the discussed features of driver proteins can help to establish successful target selection (Figure 5), while the cancer type distribution analyses of chapter 4 provide cancer subtypes, where these targeting options provide a significant therapeutic gain.

As a general rule, the targeted protein domain should be as close to the mutated IDR as possible. Therefore, the first consideration should be the modular architecture of disordered drivers, and the identification of ordered domains for inhibition in the same protein. This approach works for drivers with an oncogenic effect (excluding tumor suppressors), harboring an activating disordered mutation. It also assumes an appropriate molecular mechanism (mostly excluding disordered domains), and finally the presence of an ordered module. For these cases, the next decisive information is whether the ordered module has an enzymatic activity. For enzymes, the relevant targeting approach is likely the use of enzyme inhibitors, such as in the case of Kit. Targeting the Kit kinase domain yielded several FDA approved drugs against GIST, melanoma and thymic tumors(Abbaspour Babaei et al., 2016), one of which has proven to be effective for the studied IDR mutations in a clinical setup as well(Groisberg and Subbiah, 2017). For domains without catalytic activity, PPI inhibition can be used to block the upregulated interactions that are responsible for the oncogenic activity, such as in the case of β-catenin. This approach produced several compounds that have shown promise in vitro and in mouse xenograft models of colorectal cancer(Shin et al., 2017).

While this selection approach can yield viable targets in a nearly automated fashion, the use of extra information about the driver proteins can offer better solutions. For example, mutations in

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CSF-1R upregulate receptor activity, therefore, kinase inhibitors against CSF-1R could counteract over-activation, similarly to Kit inhibitors. However, the overactivity of CSF-1R still depends on the binding of CSF-1, and - uncharacteristically for RTKs - does not bypass upstream regulation. As a result, aberrant signaling can be shut down with more easily accessible extracellular anti-CSF-1 antibodies. Similarly, blocking the cyclin D3:Cdk4/6 interaction would counteract cyclin D3 upregulation arising from degron mutations. However, taking into account that the direct interaction partner Cdk4/6 has enzymatic activity, enzyme inhibitors against Cdk can be more efficient. This substitute inhibitor approach yielded ribociclib, an FDA approved drug for breast cancer incidences with functional Rb checkpoint(Barroso-Sousa et al., 2016).

When there is no immediate target for enzyme/PPI inhibition, suitable target domains can be located more distantly by analyzing the molecular toolkits the driver is involved in. In this case, the primary requisites of efficient targeting (oncogenic role, the presence of an ordered module, and the preference for enzymes) can be applied to proteins in the same toolkit. For example, histone H3.3 mutations - abolishing recognition by EZH2 methyltransferase - cannot be targeted by same-protein domains, as ordered histone dimers have no suppressible activity, and histone SLiM mutations do not have an activation effect. However, targeting the UTX [KDM6A] histone demethylase, which is also involved in the DNA structural modification toolkit, provides an option to counteract the effects of impaired histone methylation. As histone H3.3 G34 mutations disrupt EZH2 methyltransferase activity, resulting in aberrantly low histone methylation, UTX is also a candidate target to counteract disordered mutations resulting in EZH2 repression. This approach has shown promise both in vitro and in vivo against AML(Li et al., 2018).

In all cases, the corresponding hallmarks of cancer can give an indication of the achievable cellular effect of the applied inhibition (Supplementary Table S3). Both Kit and CSF-1R inhibition is expected to shut down proliferative signaling and to be effective against metastasis, which is a usual scenario for GIST, melanoma and breast cancer. Counteracting β-catenin mutations is expected to be effective against replicative immortality and angiogenesis, which is crucial for solid tumors, such as colorectal cancers. The main expected effect of Cdk inhibition is shutting down proliferative signaling, which is essential for breast cancer. Histone mutations are mainly connected to enabling replicative immortality, and targeting that hallmark (as opposed to angiogenesis) can be effective even in the case of non-solid tumors, like AML. All five presented target proteins in Figure 5 offer points of targeting for hallmarks that are essential for the associated cancer (sub)types, where their targeting can potentially have a huge advantage in a high fraction of know cases (with incidence rates in the 10.7-97% range, see Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

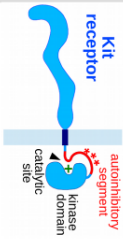
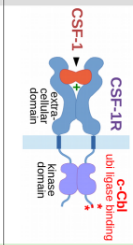
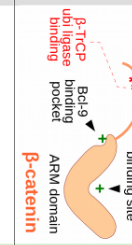
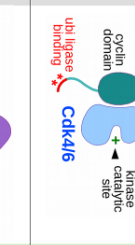
Protein name	Targeting scheme	Tumorigenic character	Molecular mechanism	Ordered module?	Enzyme activity	Additional information	Target site	Current therapeutics
Kit		oncogene	linker/auto-regulatory (activating)	yes	yes	-	**** **FDA approved, tested for IDP mutations** **kinase inhibitor acting on the same protein**	Regorafenib (GIST) Imatinib (GIST/melanoma) Sunitinib (GIST/Thymic tumor)
CSF-1R		oncogene	degron SLiM (activating)	yes	yes	does not bypass upstream regulation	** **Phase Ib/II trial** **antibody inhibitor of the extracellular ligand** (alternative: kinase inhibitor acting on the same protein)	Lacnotuzumab (melanoma, endometrial, pancreatic, or triple-negative breast cancer)
β-catenin		oncogene	degron SLiM (activating)	yes	no	ordered domain contains separate binding regions for different partners	* **Xenograft/animal model** **interaction inhibitors acting on the same protein**	HI-B1 (colon cancer) PMID:29033371
cyclin D3		oncogene	degron SLiM (activating)	yes	no	main interacting partner has catalytic activity	*** **FDA approved, different indication** **kinase inhibitor acting on direct interaction partner** (alternative: interaction inhibitors acting on the same protein)	Ribociclib (breast cancer)
histone H3.3		oncogene	PTM SLiM (non-activating)	no	no	involved in **DNA structural modification** toolkit	* **Xenograft/animal model** **enzyme inhibitor acting on a downstream component of the same pathway**	GSK-J4 (AML) PMID:29594337

Figure 5: Potential points for therapeutic intervention for IDP mutations. Green boxes mark protein features that are compatible with enzyme inhibition approaches. Fulfilling the first three criteria, and failing the fourth one (enzymatic activity) marks PPI inhibition as the most likely approach. In the targeting schematics, mutation hotspots are marked by red text and red asterisks. Potential therapeutic intervention points using inhibitors against ordered domains are labelled by green crosses. Current therapeutic options are classified according to their usability in clinical settings. **** - FDA approved drug, proven effective for IDP mutations. *** - FDA approved drug, proven effective for other, related ordered domain mutations. ** - candidate drug in clinical trial, * - candidate efficient drug in animal model.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

The identification of cancer driver genes and the elucidation of their mechanisms of action is the first step in developing efficient therapeutics. Successful identification of drivers is possible using highly customized bioinformatics pipelines that analyze genetic alterations distilled from genomic screening of tumor samples. A combined large-scale method using TCGA full exome sequencing data recently provided a catalogue of hundreds of driver genes(Bailey et al., 2018). While the method covers several main players behind common cancers, it misses many driver genes that were identified mainly in targeted efforts in cancer subtypes, which are not covered by large scale screens, mainly due to lower incidence numbers. In addition, several of the applied methods are focused on structural description of tumorigenesis at the protein level, biasing the results to preferentially cover functional protein driver regions that have well defined structures.

To get an unbiased assessment of the role of protein disorder in cancer, we used a single, structurally unbiased method on mutation data from both full exome and targeted sequencing efforts, and annotated the resulting driver set with high-quality, manual structure assignations. This resulted in a fully annotated catalogue of disordered drivers, containing 42 driver proteins containing 47 disordered driver regions. We only considered high-confidence regions to reduce noise, and to gain a largely method-independent set of drivers. Results of prediction methods in general depend on the architecture of the predictor; however, in the high-confidence regime iSiMPRe gives balanced results, largely in agreement with concurrent benchmarked methods(Porta-Pardo et al., 2017). While several of the identified driver IDRs are novel, in general, disordered driver regions have been present across several releases of the COSMIC datasets (Figure 6), starting with a few well-characterized examples, such as β-catenin and RTKs. The steady increase of the number of these examples with the continuing growth of cancer mutations, forecasts the emergence of further disorder-driven genes among strong drivers in the future. Furthermore, these examples can be complemented by additional disordered regions that are altered by more complex genetic mechanisms in cancer, such as specific frameshift mutations in NOTCH1(Wang et al., 2011), chromosomal translocations in BCR(Ballerini et al., 2012), or copy number variations in $p14^{ARF}$ (Lesueur et al., 2008). While our current collection is incomplete, roughly 30% of the contained proteins are not covered by recent pan-cancer identification efforts(Bailey et al., 2018), representing a major extension of drivers (Supplementary Table S3). In addition, even this limited set allowed us to understand some of the basic properties and common themes of how IDPs contribute to cancer development through their distinct structural and functional properties.

The collected disordered drivers, in agreement with the known versatility of IDRs, can function in multiple ways (Figure 2). They can form short linear motifs that mediate interactions with specific globular protein partners, but can also act as auto-regulatory regions, RNA/DNA binding regions, disordered domains, and flexible linkers(Piovesan et al., 2017; van der Lee et al., 2014). As common to many IDPs, they can form complex molecular switches and increase the interaction capacity of proteins(Dosztányi et al., 2006; Van Roey et al., 2012), which provides means for their mutations to have a wide modulatory effect. While these typical IDP functions cover most known cases, a small subset of disordered drivers have currently unknown perturbed molecular functions, due to the lack of available structural data. However, as each molecular function is

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

reflected in charactersitic distribution of mutations, the analysis of mutation patterns provide a way of assessing the underlying mechanistic features. For example the extremely localized patch of dominantly missense mutations in MLH1 is characteristic of linear motifs, while the dominance of inframe indels in MED12 probably indicate a linker/autoinhibitory function, or a short binding region, possibly mediating interaction with cyclin C(Turunen et al., 2014).

While disordered drivers function through distinct molecular mechanisms compared to ordered drivers, these differences diminish at the level of protein function. Most biological processes that give rise to the hallmarks of cancer can be altered both via ordered or disordered regions. Nevertheless, four key process groups, including the alteration of gene expression regulation, DNA organization, protein degradation, and RNA processing, translation and folding, are more likely to be modulated through disordered mutations (Figure 3). Yet, despite these preferences, all ten hallmarks of cancer can arise through both ordered and disordered mutations alone, highlighting that IDP mutations are a sufficient means for tumorigenesis. This is in line with our current results, showing that disordered protein region mutations are in fact sufficient for tumorigenesis. For several cancer types, a subset of patients carry only disordered driver mutations in their exome (Figure 4), which shows that – at least in these cases – there is not only a correlation between the presence of protein disorder and cancer, but there is a causal relationship.

While disordered mutations are the sole genetic events behind many disease instances, currently available cancer drugs are heavily biased against treatment for these cases. As a result, patients with mostly disordered mutations have considerably worse chances of having effective treatment options. IDRs would be especially important candidates for therapeutic interventions for several cancer types, such as liver, adrenocortical, and kidney cancers, together with diffuse large B-cell lymphoma. For some rarer cancer subtypes, and several benign but locally invasive neoplasms, counteracting IDP mutations seem to be the most efficient targeting option, and may be the only viable option for several subclasses of cancers (Table 1).

The successful identification of disordered drivers and corresponding tumor types provides the first step in successful targeting. In general, the direct targeting of IDPs requires fundamentally different approaches, and the development of such methods is being viewed as a new direction for cancer therapeutics(Kulkarni, 2016). Considering that most identified disordered drivers mediate an unusually high number of molecular interactions, the therapeutic modulation of these interactions can have therapeutic benefits. Compounds, such as nutlins, that block interactions between IDPs and their ordered domain partners have been extensively studied(Tisato et al., 2017). Other direct IDP targeting approaches under development focus on the inherent structural properties of disordered functional regions(Neira et al., 2017). While these approaches hold great promise, they are yet to reach clinical relevance(Metallo, 2010), and current drugs almost exclusively act on globular domains, leaving a sizeable portion of cancer patients without treatment options. In such cases, currently the only viable strategy is indirect targeting, using inhibitors against the catalytic activity or the interactions of an ordered domain, which can compensate for the alteration of the disordered region. The detailed understanding of the molecular architecture of the disordered driver, and the network it is embedded into, can offer ways for successful target selection in such cases.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The molecular, the network/pathway, and the cellular level information pertaining to the effect of disordered driver mutations provide complementary information for building generic target selection strategies (Figure 5). The described examples show that these strategies can be built using a decision tree-like approach for the identification of a suitable ordered protein module either inside the driver protein, or in the same pathway. The targeting options of Kit serve as proof-of-principle for our target selection approach. Imatinib is an FDA-approved kinase inhibitor designed against the ordered kinase domain of Kit mutated in GIST. Lately, Imatinib has been shown to be also effective against disordered exon 11 deletions and insertions, corresponding to a more aggressive subtype of GIST(Groisberg and Subbiah, 2017). In treatment regimens spanning three years, Imatinib was able to yield similar survival rates for GIST cases with both ordered and disordered Kit mutations. In this case, a drug developed based on classical structural considerations was proven to be highly effective against specific IDR mutations. Such a relationship can be accurately identified by our proposed indirect target selection protocol.

In general, the systematic analysis of disordered regions mutated in cancer can open up new, major treatment options – even solely with new approaches to target selection but without necessitating radically new drug development methods. This could be especially important in already somewhat targetable tumors, and can provide targets in rarer cancers that were previously deemed untargetable (e.g. glioblastoma, endometrioid carcinoma, or chronic myelomonocytic leukaemia), or in the case of less invasive or benign tumors that pose surgical difficulties (e.g. uterine leiomyomas in fertile women or hypophyseal adenomas) (Table 1). As a next step, combination therapy targeting several disordered drivers - such as c-Myc/ID-3/Forkhead box protein O1 and cyclin D3, which are all heavily mutated in Burkitt lymphoma(Schmitz et al., 2012) - may be the rational option instead of the currently available aggressive chemotherapy. Disordered drivers typically show high cancer type specificity, and their mutations often correspond to specific cancer subtypes (see Figure 4 and Table 1). Furthermore, most cancer types can have heterogeneous mutation patterns at the patient level (Figure 4). Thus, in the context of personalized medicine, targeted IDR sequencing can provide the necessary information for indirect target selection, which in turn can provide the means for new therapeutic interventions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

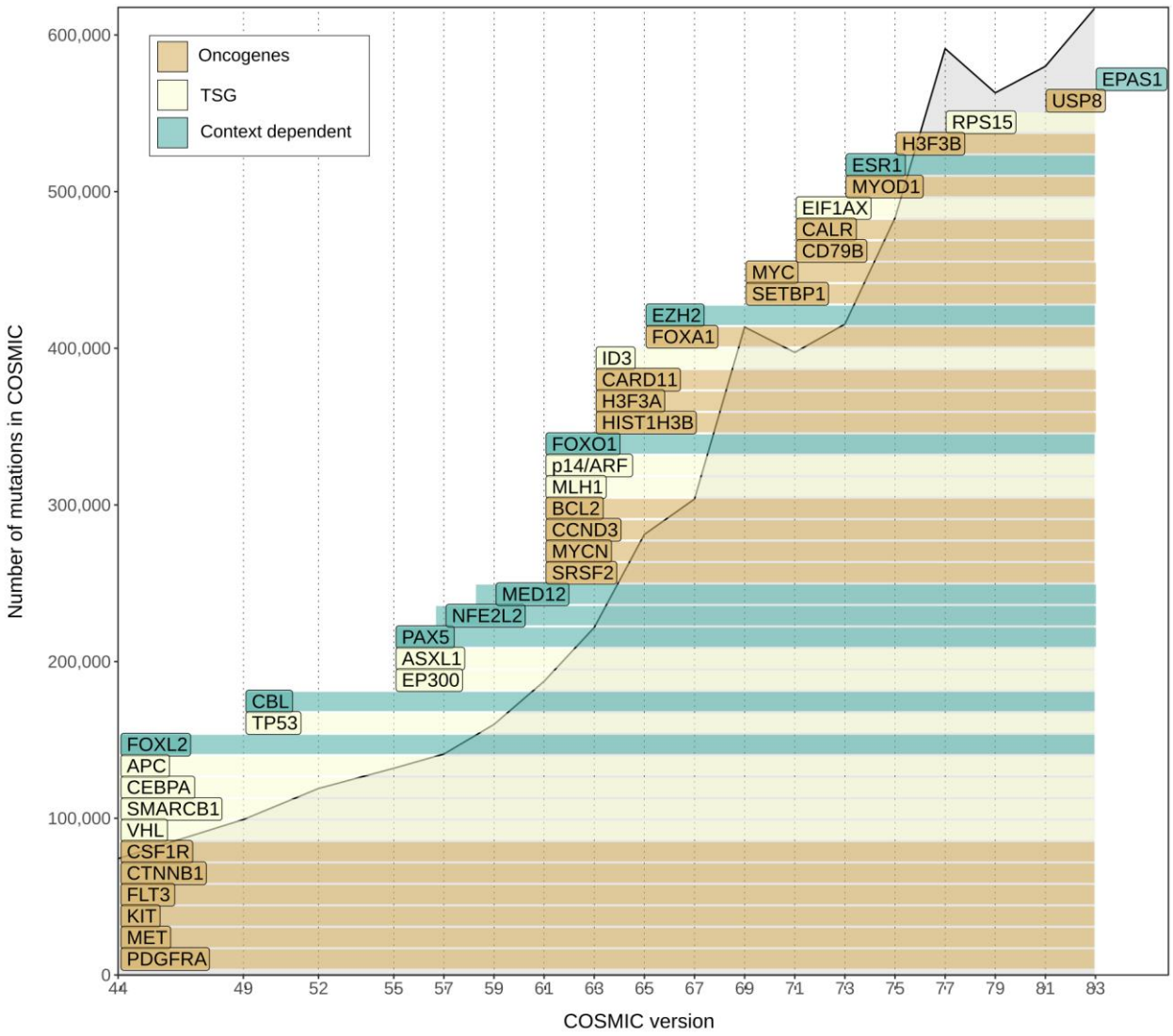


Figure 6: Growth of the list of disordered cancer drivers across various versions of the COSMIC database.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

This work was supported by the “Lendület” grant from the Hungarian Academy of Sciences (LP2014- 18) (Z.D.), OTKA grants (K108798 and K129164) (Z.D), the EMBO|EuropaBio fellowship 7544 (B.M.), and the grant PD-120973 of the National Research, Development and Innovation office of Hungary (A.Z). The authors thank Mark Adamsbaum and Drs Toby J. Gibson, Péter Tompa and László Buday for the critical reading of and their constructive comments on the manuscript.

Author Contributions

B.M. contributed to conceptualization, development of methodology and software, formal analysis, investigation of the findings, developing resources, data curation, writing the manuscript, visualization of data, project administration, and acquisition of funding. B.H.S. contributed to developing software, formal analysis, investigation of the findings, writing the manuscript, and visualization of data. A.Z. contributed to conceptualization, investigation of the findings, writing the manuscript, visualization of data, and acquisition of funding. Z.D. contributed to conceptualization, development of methodology and software, investigation of the findings, developing resources, data curation, writing the manuscript, supervision, project administration, and acquisition of funding.

Declaration of Competing Interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data and methods

Definition of modules in human proteins

Protein regions corresponding to Pfam entities(Finn et al., 2016), binding regions in DIBS(Schad et al., 2018) and MFIB(Fichó et al., 2017), regions in DisProt(Piovesan et al., 2017) and IDEAL(Fukuchi et al., 2014), and regions with structures in the PDB(Berman et al., 2000), were considered to be separate functional modules. Pfam entities were classified structurally based on instances overlapping with either DIBS, MFIB, DisProt or IDEAL entries (annotated as disordered), or with monomeric single domain protein chains in the PDB (annotated as ordered). Pfam entities with no instances overlapping with any protein regions with a clear structural designation, were annotated using predictions, together with protein residues not covered by known structural modules. Such protein regions were defined as ordered or disordered using predictions from IUPred(Dosztányi et al., 2005; Mészáros et al., 2018) and ANCHOR(Dosztanyi et al., 2009; Mészáros et al., 2009). Residues predicted to be disordered or to be part of a disordered binding region, together with their 10 residue flanking regions were considered to form disordered modules. Regions shorter than 10 residues were discarded.

Census cancer driver genes

Known cancer driver genes were collected from the 16/01/2018 version of COSMIC census database(Futreal et al., 2004), complemented with examples from the literature(Vogelstein et al., 2013). Genes were categorized as oncogenes, tumor suppressor genes, or context-dependent genes according to annotations in the COSMIC census, in dedicated databases(Zhao et al., 2016) or the literature. The full list of census drivers is given in Supplementary Table S1.

Mutation data from COSMIC and TCGA

Cancer mutations were retrieved from the v83 version of COSMIC(Forbes et al., 2016) and the v6.0 version of TCGA. Mutations used from both databases include missense mutations, and in-frame insertions and deletions. Mutations were filtered similarly to the procedure described in(Mészáros et al., 2016). Mutations from samples with over 100 mutations were discarded to avoid the inclusion of hypermutated samples. Samples including a large number of mutations in pseudogenes or mutations indicated as possible sequencing/assembly errors in(Buljan et al., 2018) were also discarded. Samples were compared to each other in a pairwise fashion and samples sharing a large fraction of mutations were clustered together. In all analyses, only mutations from the sample with the highest number of mutations were kept. Mutations falling into positions of known common polymorphisms(Smigielski et al., 2000) were filtered. The final set of COSMIC mutations used as an input to region identification consists of 599,137 missense mutations, 4,189 insertions and 12,670 deletions from 253,568 samples. The final set of TCGA mutations used as an input to region identification consists of 274,109 missense mutations, 2,775 insertions and 2,900 deletions from 7,058 samples.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Genetic alterations considered in TCGA whole sample analysis

For sample/patient level analyses (Figure 4), somatic mutations, copy number alterations and expression levels were downloaded for all 33 TCGA cancer types via the GDC Transfer Tool. Fusion data were downloaded from the Tumor Fusion Gene Data Portal. In total, we used 10,921 TCGA samples for which at least one indication was found among the four types of data. A driver region was considered to be affected in a cancer sample when there were either (i) missense point mutations and/or in-frame indels within the region, (ii) nonsense or frameshift mutations anywhere in that gene, (iii) when the gene was over- or under-expressed in the sample, or (iv) when the gene was involved in a fusion event. We tagged the gene as overexpressed, if the elevated expression level was accompanied by an increase in the copy number. Similarly, we required copy number reduction, along with decreased expression, level for considering underexpression. We found (i) missense mutations and/or in-frame indels falling in any region in 6,427 samples, (ii) driver genes with truncating mutations in 4,263 samples, (iii) over- and under-expressed driver genes in 1,304 and 317 samples, respectively, and (iv) driver genes involved in fusion events in 778 samples. Taken together, there were 8,444 samples (77.3%) with driver genes affected by at least one alteration.

Identification and categorization of driver regions in cancer-associated proteins

Driver regions were identified using iSiMPRe(Mészáros et al., 2016) with the filtered mutations from COSMIC and TCGA, separately. Then, regions obtained from COSMIC and TCGA mutations were merged, and p-values for significance were kept from the dataset with the higher significance. Only regions with p-values lower than 10^{-6} were kept. Regions were initially assigned ordered or disordered status based on the structural annotation of the corresponding functional unit, incorporating experimental data as well as predictions. (see Definition of modules in human proteins). The final ordered/disordered status of the identified regions was based on manual assertion taking into account information from the literature, if available (Supplementary Table S2). For the disordered regions, the level of supporting information for the disordered region is also included (Supplementary Table S3).

Protein-protein interaction network analysis

Binary protein-protein interactions for the human proteome were downloaded from the IntAct database(Orchard et al., 2014) on 06/05/2018. Data were filtered for human-human interactions only, where interaction partners were identified by UniProt accessions. Interactions from spoke expansions were excluded. Interactions were kept in an undirected way. (Values for disordered drivers are quoted in Supplementary Table S3).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Gene Ontology annotations

Gene Ontology terms (GO(Ashburner et al., 2000; The Gene Ontology Consortium, 2017)) were used to quantify interaction capabilities, involvement in various biological processes, molecular toolkits, and hallmarks of cancer. In each case a separate collection of GO terms (termed GO Slim) was compiled. Each GO Slim features a manual selection of GO terms that are independent from each other, meaning that they are neither child or parent terms of each other. Terms were assigned a level showing the fewest number of successive parent terms that include the root term of the ontology namespace (considered to be level 0).

GO term enrichments in a set of proteins were calculated by first obtaining expected values. Expected mean occurrence values for GO terms together with standard deviations were calculated by assessing randomly selected protein sets from the background (the full human proteome) 1,000 times. The enrichment in the studied set is expressed as the difference from the expected mean in standard deviation units.

GO Slim for assessing interaction capacity: terms from levels 1-4 from the molecular_function namespace were filtered for ancestry and only the more specific terms were kept. I.e. terms from levels 1-3 were only included if they have no child terms. Only terms describing interactions containing the keyword ‘binding’ were kept. Individual terms are shown in Supplementary Table S4.

GO for the assessment of process overlaps: terms from levels 1-4 from the biological_process namespace were filtered for ancestry and only the most specific terms were kept. Only those terms were considered that were attached to at least one protein from the set studied (full human proteome, ordered drivers, or disordered drivers). Individual terms are shown in Supplementary Table S5.

GO for molecular toolkits: biological_process terms attached to proteins with identified regions were filtered for ancestry. The resulting set was manually filtered, yielding 93 terms, which were manually grouped into 16 toolkits. Enrichments for toolkits were calculated as the ratio of the sum of expected and observed values for individual terms. Individual terms and enrichments for each toolkit are shown in Supplementary Table S6.

GO for hallmarks of cancer: Terms were chosen from the biological_process namespace via manual curation using the GO annotations of known cancer genes as a starting point. Terms were only kept if they showed a significant ($p<0.01$) enrichment on proteins in the full census cancer driver set compared to randomly selected human proteins. Individual terms and enrichments for each hallmark are shown in Supplementary Table S7.

Current anticancer drugs from OncoKB

The list of anticancer drugs and their targets were downloaded from OncoKB(Chakravarty et al., 2017). This list contains 247 indications, which consist of 83 drugs targeting 45 genes with varied alterations at different levels. We filtered for levels 1, 2 and 3 only (FDA-approved, Standard care,

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and Clinical evidence, respectively). From the different alterations we selected point mutations, indels, truncating mutations, fusions and amplifications, irrespective of the listed cancer types.

There were 28 targetable positions in 10 genes with 24 associated drugs. There were only 4 disordered positions targeted by 7 drugs (AZD-4575, BGJ-398, JNJ-42756493 and Debio1347 targeting FGFR residues 370, 371 and 373; and Cabozantinib, Capmatinib and Crizotinib targeting MET in position 1010). Notably, 5 of the 28 positions (EGFR-747, KIT-654, KIT-670, MET-1010 and MTOR-2014) were not mutated in TCGA. Targetable insertion and deletion events were only assigned to EGFR exon 19. Only PITCH1 had targetable truncating mutations. There were 13 actionable fusion events. For most cases only one fusion partner was defined, and there were only two cases for which both fusion partners were stated (BCR-ABL1, PCM1-JAK2).

Comparison between protein features

The shared information content/association between various protein features were assessed using Jaccard indices. All of the following considered features were converted into binary descriptor vectors (see Supplementary Table S3): molecular functions (protein level - 6 descriptors), number of PPI partners (3 descriptors - network level), involvement in molecular toolkits (16 descriptors - network level), involvement in hallmarks of cancer (10 descriptors - cell level), and tumorigenic character (2 descriptors - cell level). The Jaccard indices calculated between all possible pairs of descriptors are given in Supplementary Table S8, and the average Jaccard indices between all possible pairs of features are given in Supplementary Table S9.

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

Abbaspour Babaei M, Kamalidehghan B, Saleem M, Huri HZ, Ahmadipour F. 2016. Receptor tyrosine kinase (c-Kit) inhibitors: a potential therapeutic target in cancer cells. *Drug Des Devel Ther* **10**:2443–2459.

Ali MA, Sjöblom T. 2009. Molecular pathways in tumor progression: from discovery to functional understanding. *Mol Biosyst* **5**:902–908.

Aoki K, Taketo MM. 2007. Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene. *J Cell Sci* **120**:3327–3335.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**:25–29.

Babu MM, van der Lee R, de Groot NS, Gsponer J. 2011. Intrinsically disordered proteins: regulation and disease. *Curr Opin Struct Biol* **21**:432–440.

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Ng PK-S, Jeong KJ, Cao S, Wang Z, Gao J, Gao Q, Wang F, Liu EM, Mularoni L, Rubio-Perez C, Nagarajan N, Cortés-Ciriano I, Zhou DC, Liang W-W, Hess JM, Yellapantula VD, Tamborero D, Gonzalez-Perez A, Suphavilai C, Ko JY, Khurana E, Park PJ, Van Allen EM, Liang H, MC3 Working Group, Cancer Genome Atlas Research Network, Lawrence MS, Godzik A, Lopez-Bigas N, Stuart J, Wheeler D, Getz G, Chen K, Lazar AJ, Mills GB, Karchin R, Ding L. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**:371–385.e18.

Ballerini P, Struski S, Cresson C, Prade N, Toujani S, Deswarte C, Dobbelstein S, Petit A, Lapillonne H, Gautier E-F, Demur C, Lippert E, Pages P, Mansat-De Mas V, Donadieu J, Huguet F, Dastugue N, Broccardo C, Perot C, Delabesse E. 2012. RET fusion genes are associated with chronic myelomonocytic leukemia and enhance monocytic differentiation. *Leukemia* **26**:2384–2389.

Ballmann C, Thiel A, Korah HE, Reis A-C, Saeger W, Stepanow S, Köhrer K, Reifenberger G, Knobbe-Thomsen CB, Knappe UJ, Scholl UI. 2018. Mutations in Pituitary Cushing Adenomas-Targeted Analysis by Next-Generation Sequencing. *J Endocr Soc* **2**:266–278.

Barroso-Sousa R, Shapiro GI, Tolaney SM. 2016. Clinical Development of the CDK4/6 Inhibitors Ribociclib and Abemaciclib in Breast Cancer. *Breast Care* **11**:167–173.

Behjati S, Tarpey PS, Presneau N, Scheipl S, Pillay N, Van Loo P, Wedge DC, Cooke SL, Gundem G, Davies H, Nik-Zainal S, Martin S, McLaren S, Goodie V, Robinson B, Butler A, Teague JW, Halai D, Khatri B, Myklebost O, Baumhoer D, Jundt G, Hamoudi R, Tirabosco R, Amary MF, Futreal PA, Stratton MR, Campbell PJ, Flanagan AM. 2013. Distinct H3F3A and H3F3B driver mutations define chondroblastoma and giant cell tumor of bone. *Nat Genet* **45**:1479–1482.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* **28**:235–242.

Buljan M, Blattmann P, Aebersold R, Boutros M. 2018. Systematic characterization of pan-cancer mutation clusters. *Mol Syst Biol* **14**:e7974.

Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**:1113–1120.

Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, Chang MT, Chandarlapaty S, Traina TA, Paik PK, Ho AL, Hantash FM, Grupe A, Baxi SS, Callahan MK, Snyder A, Chi P, Danila D, Gounder M, Harding JJ, Hellmann MD, Iyer G, Janjigian Y, Kaley T, Levine DA, Lowery M, Omuro A, Postow MA, Rathkopf D, Shoushtari AN, Shukla N, Voss M, Paraiso E, Zehir A, Berger MF, Taylor BS, Saltz LB, Riely GJ, Ladanyi M, Hyman DM, Baselga J, Sabbatini P, Solit DB, Schultz N. 2017. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**. doi:10.1200/PO.17.00011

Comino-Méndez I, de Cubas AA, Bernal C, Álvarez-Escolá C, Sánchez-Malo C, Ramírez-Tortosa CL, Pedrinaci S, Rapizzi E, Ercolino T, Bernini G, Bacca A, Letón R, Pita G, Alonso MR, Leandro-García LJ, Gómez-Graña A, Inglada-Pérez L, Mancikova V, Rodríguez-Antona C, Mannelli M, Robledo M, Cascón A. 2013. Tumoral EPAS1 (HIF2A) mutations explain sporadic pheochromocytoma and paraganglioma in the absence of erythrocytosis. *Hum Mol Genet* **22**:2169–2176.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Compagno M, Lim WK, Grunn A, Nandula SV, Brahmachary M, Shen Q, Bertoni F, Ponzoni M, Scandurra M, Califano A, Bhagat G, Chadburn A, Dalla-Favera R, Pasqualucci L. 2009. Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma. *Nature* **459**:717–721.

Copeland NG, Jenkins NA. 2009. Deciphering the genetic landscape of cancer--from genes to pathways. *Trends Genet* **25**:455–462.

Cortese MS, Uversky VN, Dunker AK. 2008. Intrinsic disorder in scaffold proteins: getting more from less. *Prog Biophys Mol Biol* **98**:85–106.

Craig E, Zhang Z-K, Davies KP, Kalpana GV. 2002. A masked NES in INI1/hSNF5 mediates hCRM1-dependent nuclear export: implications for tumorigenesis. *EMBO J* **21**:31–42.

Darling AL, Uversky VN. 2018. Intrinsic Disorder and Posttranslational Modifications: The Darker Side of the Biological Dark Matter. *Front Genet* **9**:158.

Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ. 2012. Attributes of short linear motifs. *Mol Biosyst* **8**:268–281.

Dosztányi Z, Chen J, Dunker AK, Simon I, Tompa P. 2006. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res* **5**:2985–2995.

Dosztányi Z, Csizmók V, Tompa P, Simon I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**:827–839.

Dosztanyi Z, Meszaros B, Simon I. 2009. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**:2745–2746.

Du P, Huang P, Huang X, Li X, Feng Z, Li F, Liang S, Song Y, Stenvang J, Brünner N, Yang H, Ou Y, Gao Q, Li L. 2017. Comprehensive genomic analysis of Oesophageal Squamous Cell Carcinoma reveals clinical relevance. *Sci Rep* **7**:15324.

Dyson HJ, Wright PE. 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* **6**:197–208.

Eisenhaber B, Eisenhaber F. 2007. Posttranslational modifications and subcellular localization signals: indicators of sequence regions without inherent 3D structure? *Curr Protein Pept Sci* **8**:197–203.

Elf S, Abdelfattah NS, Chen E, Perales-Patón J, Rosen EA, Ko A, Peisker F, Florescu N, Giannini S, Wolach O, Morgan EA, Tothova Z, Losman J-A, Schneider RK, Al-Shahrour F, Mullally A. 2016. Mutant Calreticulin Requires Both Its Mutant C-terminus and the Thrombopoietin Receptor for Oncogenic Transformation. *Cancer Discov* **6**:368–381.

Engin HB, Kreisberg JF, Carter H. 2016. Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLoS One* **11**:e0152929.

Familiades J, Bousquet M, Lafage-Pochitaloff M, Béné M-C, Beldjord K, De Vos J, Dastugue N, Coyaud E, Struski S, Quelen C, Prade-Houdellier N, Dobbelstein S, Cayuela J-M, Soulier J, Grardel N, Preudhomme C, Cavé H, Blanchet O, Lhéritier V, Delannoy A, Chalandon Y, Ifrah N, Pigneux A, Brousset P, Macintyre EA, Huguet F, Dombret H, Broccardo C, Delabesse E. 2009. PAX5 mutations occur frequently in adult B-cell progenitor acute lymphoblastic leukemia and PAX5 haploinsufficiency is associated with BCR-ABL1 and TCF3-PBX1 fusion genes: a GRAALL study. *Leukemia* **23**:1989–1998.

Faust O, Bigman L, Friedler A. 2014. A role of disordered domains in regulating protein oligomerization and stability. *Chem Commun* **50**:10797–10800.

Fichó E, Reményi I, Simon I, Mészáros B. 2017. MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics* **33**:3682–3684.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**:D279–85.

Forbes SA, Beare D, Bindal N, Bamford S, Ward S, Cole CG, Jia M, Kok C, Boutselakis H, De T, Sondka Z, Ponting L, Stefancsik R, Harsha B, Tate J, Dawson E, Thompson S, Jubb H, Campbell PJ. 2016. COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. *Curr Protoc Hum Genet* **91**:10.11.1–10.11.37.

Fukuchi S, Amemiya T, Sakamoto S, Nobe Y, Hosoda K, Kado Y, Murakami SD, Koike R, Hiroaki H, Ota M. 2014. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res* **42**:D320–5.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

census of human cancer genes. *Nat Rev Cancer* **4**:177–183.
Garvie CW, Hagman J, Wolberger C. 2001. Structural studies of Ets-1/Pax5 complex formation on DNA. *Mol Cell* **8**:1267–1276.
Gielen GH, Gessi M, Hammes J, Kramm CM, Waha A, Pietsch T. 2013. H3F3A K27M mutation in pediatric CNS tumors: a marker for diffuse high-grade astrocytomas. *Am J Clin Pathol* **139**:345–349.
Griffith D, Parker JP, Marmion CJ. 2010. Enzyme inhibition as a key target for the development of novel metal-based anti-cancer therapeutics. *Anticancer Agents Med Chem* **10**:354–370.
Groisberg R, Subbiah V. 2017. The big, the bad, and the exon 11: adjuvant imatinib for all gastro-intestinal stromal tumors or just the ugly? *Transl Gastroenterol Hepatol* **2**:81.
Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell* **144**:646–674.
Hegyi H, Buday L, Tompa P. 2009. Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins. *PLoS Comput Biol* **5**:e1000552.
Hubbard SR. 2004. Juxtamembrane autoinhibition in receptor tyrosine kinases. *Nat Rev Mol Cell Biol* **5**:464–471.
Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK. 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* **323**:573–584.
Ivanov AA, Khuri FR, Fu H. 2013. Targeting protein-protein interactions as an anticancer strategy. *Trends Pharmacol Sci* **34**:393–400.
Jamieson S, Butzow R, Andersson N, Alexiadis M, Unkila-Kallio L, Heikinheimo M, Fuller PJ, Anttonen M. 2010. The FOXL2 C134W mutation is characteristic of adult granulosa cell tumors of the ovary. *Mod Pathol* **23**:1477–1485.
Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, Lander ES, Getz G. 2015. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci U S A* **112**:E5486–95.
Kulkarni P. 2016. Intrinsically disordered proteins and prostate cancer: pouring new wine in an old bottle. *Asian J Androl* **18**:659–661.
Laé M, Gardrat S, Rondeau S, Richardot C, Caly M, Chemlali W, Vacher S, Couturier J, Mariani O, Terrier P, Bièche I. 2016. MED12 mutations in breast phyllodes tumors: evidence of temporal tumoral heterogeneity and identification of associated critical signaling pathways. *Oncotarget* **7**:84428–84438.
Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**:495–501.
Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortés ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau D-A, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CWM, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**:214–218.
Lenz G, Davis RE, Ngo VN, Lam L, George TC, Wright GW, Dave SS, Zhao H, Xu W, Rosenwald A, Ott G, Muller-Hermelink HK, Gascoyne RD, Connors JM, Rimsza LM, Campo E, Jaffe ES, Delabie J, Smeland EB, Fisher RI, Chan WC, Staudt LM. 2008. Oncogenic CARD11 mutations in human diffuse large B cell lymphoma. *Science* **319**:1676–1679.
Lesueur F, de Lichy M, Barrois M, Durand G, Bombled J, Avril M-F, Chompret A, Boitier F, Lenoir GM, French Familial Melanoma Study Group, Bressac-de Paillerets B, Baccard M, Bachollet B, Berthet P, Bonadona V, Bonnetblanc J-M, Caron O, Chevrant-Breton J, Cuny J-F, Dalle S, Delaunay M, Demange L, De Quatrebarbes J, Doré J-F, Frénay M, Fricker J-P, Gauthier-Villars M, Gesta P, Giraud S, Gorry P, Grange F, Green A, Huiart L, Janin N, Joly P, Kérob D, Lasset C, Leroux D, Limacher J-M, Longy M, Mansard S, Marrou K, Martin-Denavit T, Mateus C, Maubec E, Olivier-Faivre L, Orlandini V, Pujol P, Sassolas B, Stoppa-Lyonnet D, Thomas L, Vabres P, Venat L, Wierzbicka E, Zattara H. 2008. The contribution of large genomic deletions at the CDKN2A locus to the burden of familial melanoma. *Br J Cancer* **99**:364–370.
Li E, Hristova K. 2010. Receptor tyrosine kinase transmembrane domains: Function, dimer structure and

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

dimerization energetics. *Cell Adh Migr* **4**:249–254.

Lin L-I, Chen C-Y, Lin D-T, Tsay W, Tang J-L, Yeh Y-C, Shen H-L, Su F-H, Yao M, Huang S-Y, Tien H-F. 2005. Characterization of CEBPA mutations in acute myeloid leukemia: most patients with CEBPA mutations have biallelic mutations and show a distinct immunophenotype of the leukemic cells. *Clin Cancer Res* **11**:1372–1379.

Liu Y, Patel L, Mills GB, Lu KH, Sood AK, Ding L, Kucherlapati R, Mardis ER, Levine DA, Shmulevich I, Broaddus RR, Zhang W. 2014. Clinical significance of CTNNB1 mutation and Wnt pathway activation in endometrioid endometrial carcinoma. *J Natl Cancer Inst* **106**. doi:10.1093/jnci/dju245

Li Y, Zhang M, Sheng M, Zhang P, Chen Z, Xing W, Bai J, Cheng T, Yang F-C, Zhou Y. 2018. Therapeutic potential of GSK-J4, a histone demethylase KDM6B/JMJD3 inhibitor, for acute myeloid leukemia. *J Cancer Res Clin Oncol* **144**:1065–1077.

Lounnas V, Ritschel T, Kelder J, McGuire R, Bywater RP, Foloppe N. 2013. Current progress in Structure-Based Rational Drug Design marks a new mindset in drug discovery. *Comput Struct Biotechnol J* **5**:e201302011.

Mäkinen N, Mehine M, Tolvanen J, Kaasinen E, Li Y, Lehtonen HJ, Gentile M, Yan J, Enge M, Taipale M, Aavikko M, Katainen R, Virolainen E, Böhling T, Koski TA, Launonen V, Sjöberg J, Taipale J, Vahteristo P, Aaltonen LA. 2011. MED12, the mediator complex subunit 12 gene, is mutated at high frequency in uterine leiomyomas. *Science* **334**:252–255.

McConechy MK, Ding J, Senz J, Yang W, Melnyk N, Tone AA, Prentice LM, Wiegand KC, McAlpine JN, Shah SP, Lee C-H, Goodfellow PJ, Gilks CB, Huntsman DG. 2014. Ovarian and endometrial endometrioid carcinomas have distinct CTNNB1 and PTEN mutation profiles. *Mod Pathol* **27**:128–134.

Meggendorfer M, Roller A, Haferlach T, Eder C, Dicker F, Grossmann V, Kohlmann A, Alpermann T, Yoshida K, Ogawa S, Koeffler HP, Kern W, Haferlach C, Schnittger S. 2012. SRSF2 mutations in 275 cases with chronic myelomonocytic leukemia (CMML). *Blood* **120**:3080–3088.

Mészáros B, Erdos G, Dosztányi Z. 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* **46**:W329–W337.

Mészáros B, Simon I, Dosztányi Z. 2009. Prediction of Protein Binding Regions in Disordered Proteins. *PLoS Comput Biol* **5**:e1000376.

Mészáros B, Zeke A, Reményi A, Simon I, Dosztányi Z. 2016. Systematic analysis of somatic mutations driving cancer: uncovering functional protein regions in disease development. *Biol Direct* **11**:23.

Metallo SJ. 2010. Intrinsically disordered proteins are potential drug targets. *Curr Opin Chem Biol* **14**:481–488.

Meyer K, Kirchner M, Uyar B, Cheng J-Y, Russo G, Hernandez-Miranda LR, Szymborska A, Zauber H, Rudolph I-M, Willnow TE, Akalin A, Haucke V, Gerhardt H, Birchmeier C, Kühn R, Krauss M, Diecke S, Pascual JM, Selbach M. 2018. Mutations in Disordered Regions Can Cause Disease by Creating Dileucine Motifs. *Cell* **175**:239–253.e17.

Mishra A, Singh V, Verma V, Pandey S, Trivedi R, Singh HP, Kumar S, Dwivedi RC, Mishra SC. 2016. Current status and clinical association of beta-catenin with juvenile nasopharyngeal angiofibroma. *J Laryngol Otol* **130**:907–913.

Morin PJ, Kinzler KW, Sparks AB. 2016. β-Catenin Mutations: Insights into the APC Pathway and the Power of Genetics. *Cancer Res* **76**:5587–5589.

Mullen JT, DeLaney TF, Rosenberg AE, Le L, Iafrate AJ, Kobayashi W, Szymonifka J, Yeap BY, Chen Y-L, Harmon DC, Choy E, Yoon SS, Raskin KA, Hornicek FJ, Nielsen GP. 2013. β-Catenin mutation status and outcomes in sporadic desmoid tumors. *Oncologist* **18**:1043–1049.

Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, Girtman K, Mathew S, Ma J, Pounds SB, Su X, Pui C-H, Relling MV, Evans WE, Shurtleff SA, Downing JR. 2007. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**:758–764.

Neira JL, Bintz J, Arruebo M, Rizzuti B, Bonacci T, Vega S, Lanas A, Velázquez-Campoy A, Iovanna JL, Abián O. 2017. Identification of a Drug Targeting an Intrinsically Disordered Protein Involved in Pancreatic Adenocarcinoma. *Sci Rep* **7**:39732.

Olivier M, Hollstein M, Hainaut P. 2010. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb Perspect Biol* **2**:a001008.

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H. 2014. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**:D358–63.

Ouyang Y, Qiao C, Chen Y, Zhang S-J. 2017. Clinical significance of CSF3R, SRSF2 and SETBP1 mutations in chronic neutrophilic leukemia and chronic myelomonocytic leukemia. *Oncotarget* **8**:20834–20841.

Pajkos M, Mészáros B, Simon I, Dosztányi Z. 2012. Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. *Mol Biosyst* **8**:296–307.

Pathania S, Bhatia R, Baldi A, Singh R, Rawal RK. 2018. Drug metabolizing enzymes and their inhibitors' role in cancer resistance. *Biomed Pharmacother* **105**:53–65.

Paz-Priel I, Friedman A. 2011. C/EBPα dysregulation in AML and ALL. *Crit Rev Oncog* **16**:93–102.

Pezzuto F, Izzo F, Buonaguro L, Annunziata C, Tatangelo F, Botti G, Buonaguro FM, Tornesello ML. 2016. Tumor specific mutations in TERT promoter and CTNNB1 gene in hepatitis B and hepatitis C related hepatocellular carcinoma. *Oncotarget* **7**:54253–54262.

Piazza R, Valletta S, Winkelmann N, Redaelli S, Spinelli R, Pirola A, Antolini L, Mologni L, Donadoni C, Papaemmanuil E, Schnittger S, Kim D-W, Boultwood J, Rossi F, Gaipa G, De Martini GP, di Celle PF, Jang HG, Fantin V, Bignell GR, Magistroni V, Haferlach T, Pogliani EM, Campbell PJ, Chase AJ, Tapper WJ, Cross NCP, Gambacorti-Passerini C. 2013. Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nat Genet* **45**:18–24.

Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidović R, Dosztányi Z, Elofsson A, Gasparini A, Hatos A, Kajava AV, Kalmar L, Leonardi E, Lazar T, Macedo-Ribeiro S, Macossay-Castillo M, Meszaros A, Minervini G, Murvai N, Pujols J, Roche DB, Salladini E, Schad E, Schramm A, Szabo B, Tantos A, Tonello F, Tsirigos KD, Veljković N, Ventura S, Vranken W, Warholm P, Uversky VN, Dunker AK, Longhi S, Tompa P, Tosatto SCE. 2017. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res* **45**:D1123–D1124.

Porta-Pardo E, Garcia-Alonso L, Hrabe T, Dopazo J, Godzik A. 2015. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput Biol* **11**:e1004518.

Porta-Pardo E, Godzik A. 2014. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30**:3109–3114.

Porta-Pardo E, Kamburov A, Tamborero D, Pons T, Grases D, Valencia A, Lopez-Bigas N, Getz G, Godzik A. 2017. Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nat Methods* **14**:782–788.

Ravegnini G, Mariño-Enriquez A, Slater J, Eilers G, Wang Y, Zhu M, Nucci MR, George S, Angelini S, Raut CP, Fletcher JA. 2013. MED12 mutations in leiomyosarcoma and extrauterine leiomyoma. *Mod Pathol* **26**:743–749.

Rekhi B, Upadhyay P, Ramteke MP, Dutt A. 2016. MYOD1 (L122R) mutations are associated with spindle cell and sclerosing rhabdomyosarcomas with aggressive clinical outcomes. *Mod Pathol* **29**:1532–1540.

Ridge SA, Worwood M, Oscier D, Jacobs A, Padua RA. 1990. FMS mutations in myelodysplastic, leukemic, and normal subjects. *Proc Natl Acad Sci U S A* **87**:1377–1380.

Scatena R, Bottoni P, Pontoglio A, Mastrototaro L, Giardina B. 2008. Glycolytic enzyme inhibitors in cancer treatment. *Expert Opin Investig Drugs* **17**:1533–1545.

Schad E, Fichó E, Pancsa R, Simon I, Dosztányi Z, Mészáros B. 2018. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* **34**:535–537.

Schmitz R, Young RM, Ceribelli M, Jhavar S, Xiao W, Zhang M, Wright G, Shaffer AL, Hodson DJ, Buras E, Liu X, Powell J, Yang Y, Xu W, Zhao H, Kohlhammer H, Rosenwald A, Kluin P, Müller-Hermelink HK, Ott G, Gascoyne RD, Connors JM, Rimsza LM, Campo E, Jaffe ES, Delabie J, Smeland EB, Ogwang MD, Reynolds SJ, Fisher RI, Braziel RM, Tubbs RR, Cook JR, Weisenburger DD, Chan WC, Pittaluga S, Wilson W, Waldmann TA, Rowe M, Mbulaiteye SM, Rickinson AB, Staudt LM. 2012. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature* **490**:116–120.

Shah SP, Köbel M, Senz J, Morin RD, Clarke BA, Wiegand KC, Leung G, Zayed A, Mehl E, Kalloger SE, Sun M, Giuliany R, Yorida E, Jones S, Varhol R, Swenerton KD, Miller D, Clement PB, Crane C, Madore J, Provencher D, Leung P, DeFazio A, Khattra J, Turashvili G, Zhao Y, Zeng T, Glover JNM, Vanderhyden B, Zhao C, Parkinson CA, Jimenez-Linan M, Bowtell DDL, Mes-Masson A-M, Brenton

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

JD, Aparicio SA, Boyd N, Hirst M, Gilks CB, Marra M, Huntsman DG. 2009. Mutation of FOXL2 in granulosa-cell tumors of the ovary. *N Engl J Med* **360**:2719–2729.

Shin SH, Lim DY, Reddy K, Malakhova M, Liu F, Wang T, Song M, Chen H, Bae KB, Ryu J, Liu K, Lee M-H, Bode AM, Dong Z. 2017. A Small Molecule Inhibitor of the β-Catenin-TCF4 Interaction Suppresses Colorectal Cancer Growth In Vitro and In Vivo. *EBioMedicine* **25**:22–31.

Smigielski EM, Sirotkin K, Ward M, Sherry ST. 2000. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* **28**:352–355.

Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. 2018. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**:696–705.

Staby L, O'Shea C, Willemoës M, Theisen F, Kragelund BB, Skriver K. 2017. Eukaryotic transcription factors: paradigms of protein intrinsic disorder. *Biochem J* **474**:2509–2532.

Sutovsky H, Gazit E. 2004. The von Hippel-Lindau tumor suppressor protein is a molten globule under native conditions: implications for its physiological activities. *J Biol Chem* **279**:17190–17196.

Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. 2013. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**:2238–2244.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA. 2019. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**:D941–D947.

The Gene Ontology Consortium. 2017. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res* **45**:D331–D338.

Tisato V, Voltan R, Gonelli A, Secchiero P, Zauli G. 2017. MDM2/X inhibitors under clinical evaluation: perspectives for the management of hematological malignancies and pediatric cancer. *J Hematol Oncol* **10**:133.

Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, Masica DL, Karchin R. 2016. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res* **76**:3719–3731.

Tompa P. 2002. Intrinsically unstructured proteins. *Trends Biochem Sci* **27**:527–533.

Turunen M, Spaeth JM, Keskitalo S, Park MJ, Kivioja T, Clark AD, Mäkinen N, Gao F, Palin K, Nurkkala H, Vähärautio A, Aavikko M, Kämpjärvi K, Vahteristo P, Kim CA, Aaltonen LA, Varjosalo M, Taipale J, Boyer TG. 2014. Uterine leiomyoma-linked MED12 mutations disrupt mediator-associated CDK activity. *Cell Rep* **7**:654–660.

Uyar B, Weatheritt RJ, Dinkel H, Davey NE, Gibson TJ. 2014. Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? *Mol Biosyst* **10**:2626–2642.

Vacic V, Markwick PRL, Oldfield CJ, Zhao X, Haynes C, Uversky VN, Iakoucheva LM. 2012. Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Comput Biol* **8**:e1002709.

van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM. 2014. Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**:6589–6631.

Van Roey K, Gibson TJ, Davey NE. 2012. Motif switches: decision-making in cell regulation. *Curr Opin Struct Biol* **22**:378–385.

Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B. 2009. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* **138**:198–208.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. 2013. Cancer genome landscapes. *Science* **339**:1546–1558.

Wang NJ, Sanborn Z, Arnett KL, Bayston LJ, Liao W, Proby CM, Leigh IM, Collisson EA, Gordon PB, Jakkula L, Pennypacker S, Zou Y, Sharma M, North JP, Vemula SS, Mauro TM, Neuhaus IM, Leboit PE, Hur JS, Park K, Huh N, Kwok P-Y, Arron ST, Massion PP, Bale AE, Haussler D, Cleaver JE, Gray JW, Spellman PT, South AP, Aster JC, Blacklow SC, Cho RJ. 2011. Loss-of-function mutations in Notch receptors in cutaneous and lung squamous cell carcinoma. *Proc Natl Acad Sci U S A* **108**:17761–17766.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**:635–645.

Wright PE, Dyson HJ. 2015. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev*

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Mol Cell Biol **16**:18–29.

Xu Z, Huo X, Tang C, Ye H, Nandakumar V, Lou F, Zhang D, Jiang S, Sun H, Dong H, Zhang G, Liu Z, Dong Z, Guo B, Yan H, Yan C, Wang L, Su Z, Li Y, Gu D, Zhang X, Wu X, Wei X, Hong L, Zhang Y, Yang J, Gong Y, Tang C, Jones L, Huang XF, Chen S-Y, Chen J. 2014. Frequent KIT mutations in human gastrointestinal stromal tumors. *Sci Rep* **4**:5907.

Yang F, Petsalaki E, Rolland T, Hill DE, Vidal M, Roth FP. 2015. Protein domain-level landscape of cancer-type-specific somatic mutations. *PLoS Comput Biol* **11**:e1004147.

Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. 2016. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res* **44**:D1023–31.

Zheng M, Perry AM, Bierman P, Loberiza F Jr, Nasr MR, Szwajcer D, Del Bigio MR, Smith LM, Zhang W, Greiner TC. 2017. Frequency of MYD88 and CD79B mutations, and MGMT methylation in primary central nervous system diffuse large B-cell lymphoma. *Neuropathology* **37**:509–516.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary material

Supplementary table legends

Table S1. List of cancer driver genes. Genes were taken from the COSMIC census list as of 21/03/2018, and were complemented with known cancer genes from the literature.

Table S2. List of regions identified using iSiMPRe, based on both COSMIC and TCGA mutations.

Table S3. Identified disordered driver genes with all annotations.

Table S4. Gene Ontology terms used in the quantification of interaction capabilities.

Table S5. Gene Ontology terms used in the quantification of biological process overlaps.

Table S6. Gene Ontology terms used in the quantification of molecular toolkits used by cancer driver genes.

Table S7. Gene Ontology terms used in the quantification of hallmarks of cancer.

Table S8. Jaccard indices of the similarity between all studies protein features.

Table S9. Averaged Jaccard indices of the similarity between various merged features.

Supplementary figures

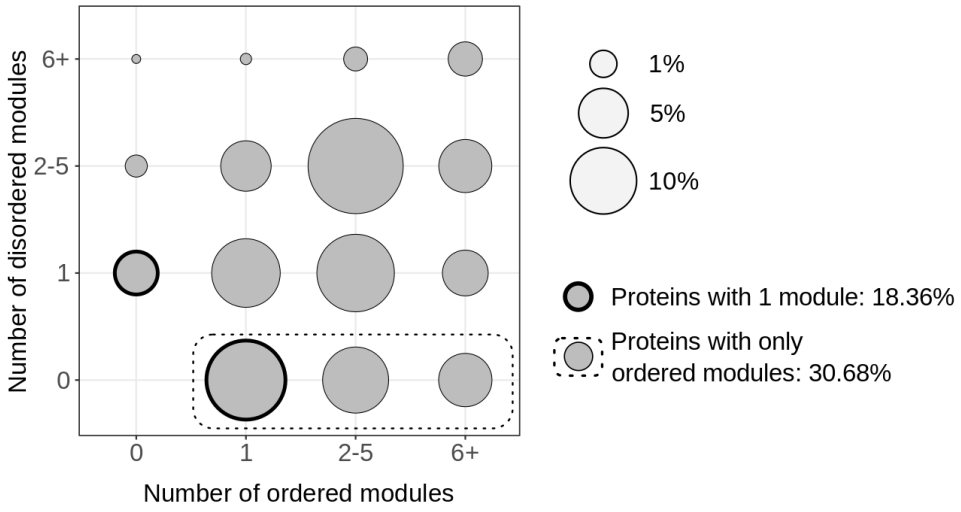


Figure S1. The distribution of all human proteins with regard to the structural states of constituent functional modules.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

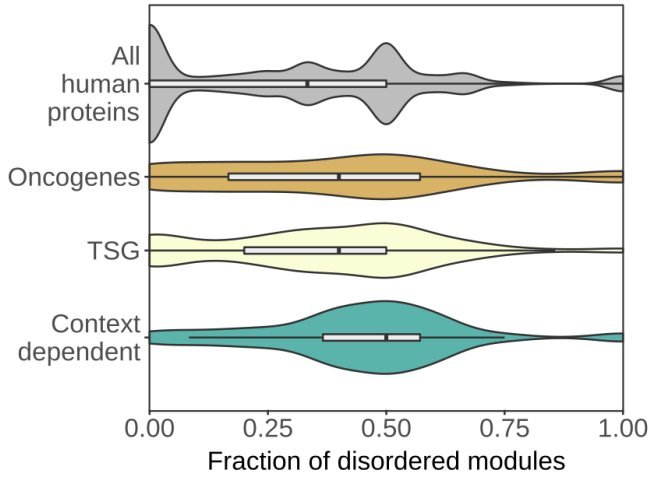


Figure S2. The ratio of disordered modules in all human proteins and the three major groups of cancer drivers.

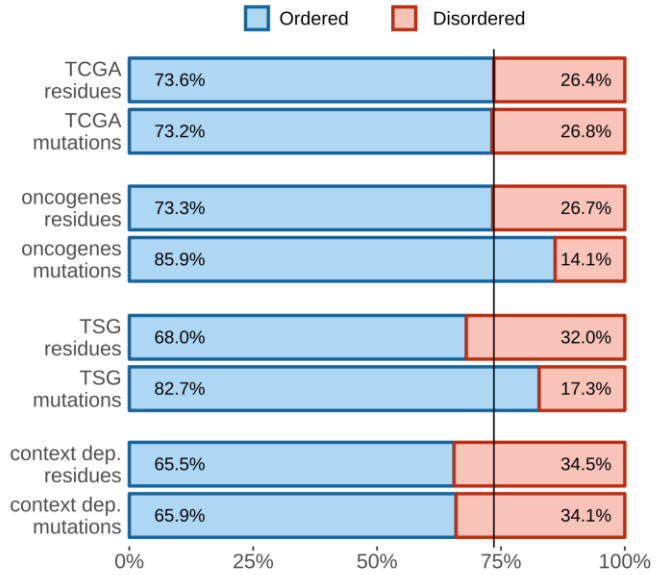


Figure S3. The distribution of residues and TCGA cancer driver mutations in the human proteome and various cancer driver classes. The vertical line marks the ratio of ordered and disordered residues in the human proteome (73.6%), corresponding to the expected ratio of randomly occurring mutations.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.29.069245; this version posted April 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

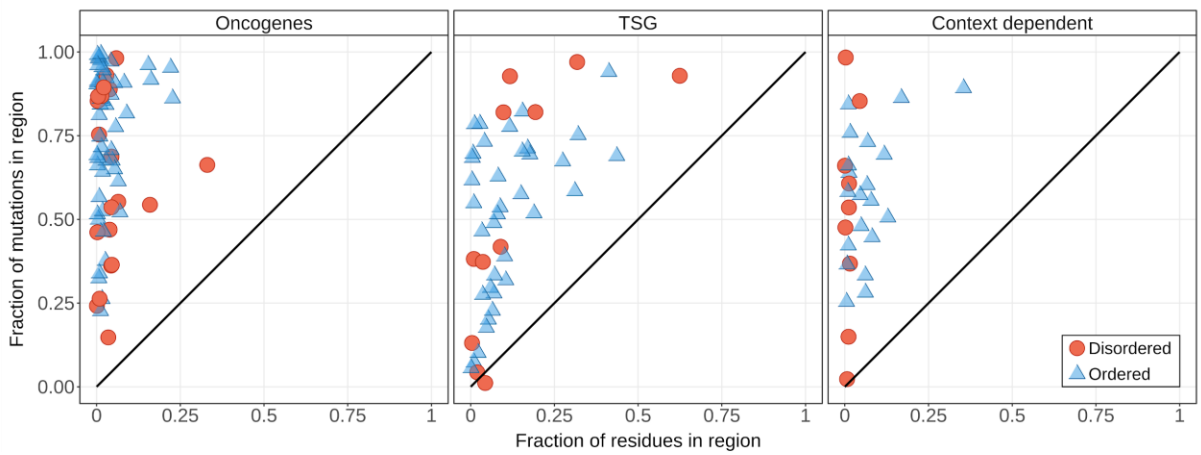


Figure S4. Identified regions are compact functional units. The fraction of mutations residing in the identified regions vs the fraction of residues in the regions for oncogenes, tumor suppressor genes and context-dependent genes. Regions in oncogenes encompass less than 10% of the sequence. Tumor suppressor genes (TSGs) are most often modulated via non-localized truncating mutations, affecting larger regions or the whole of the protein. However, the clustering of localized point mutations covering an extended region of the protein is a frequent hallmark of TSGs as well, making them largely identifiable using point mutations alone. In accord, several TSGs also harbor identified driver regions. While these regions are larger in size compared to that of oncogenes, they still cover only less than 20% of the sequence in most cases. In accordance with their dual nature, regions found inside context-dependent proteins lie between regions in oncogenes and TSGs in terms of length.

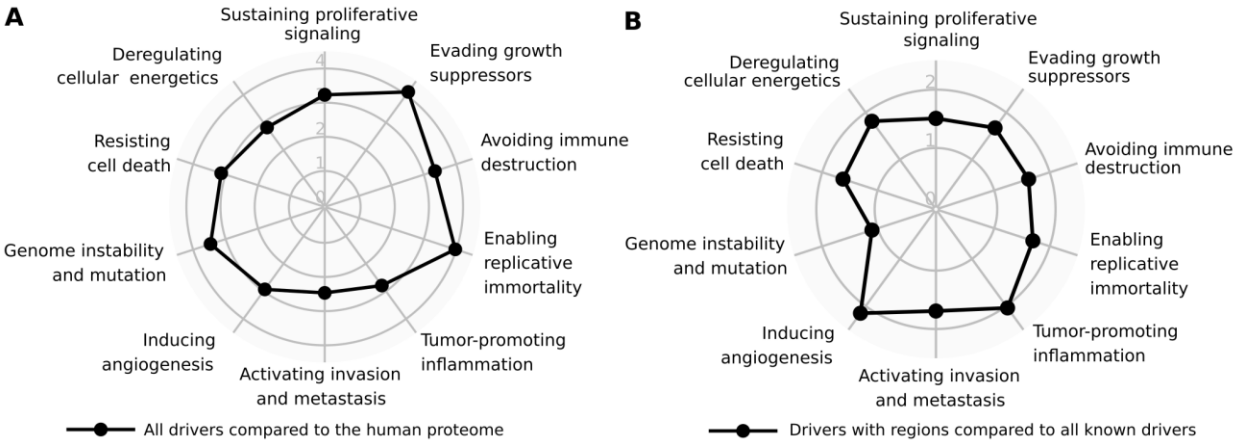


Figure S5. A. Overrepresentation of hallmarks of cancer for all driver genes compared to the human proteome. B. Overrepresentation of hallmarks of cancer for driver genes with identified regions compared to all driver genes.