Spring 2022

Network Approaches to the Study of Genomic Variation in Cancer

Hussein Mohsen
Yale University Graduate School of Arts and Sciences, hussein.mohsen5253@gmail.com

Abstract

Network Approaches to the Study of Genomic Variation in Cancer

Hussein Mohsen

2022

Advances in genomic sequencing technologies opened the door for a wider study of cancer etiology. By analyzing datasets with thousands of exomes (or genomes), researchers gained a better understanding of the genomic alterations that confer a selective advantage towards cancerous growth. A predominant narrative in the field has been based on a dichotomy of alterations that confer a strong selective advantage, called cancer drivers, and the bulk of other alterations assumed to have a neutral effect, called passengers. Yet, a series of studies questioned this narrative and assigned potential roles to passengers, be it in terms of facilitating tumorigenesis or countering the effect of drivers. Consequently, the passenger mutational landscape received a higher level of attention in attempt to prioritize the possible effects of its alterations and to identify new therapeutic targets.

In this dissertation, we introduce interpretable network approaches to the study of genomic variation in cancer. We rely on two types of networks, namely functional biological networks and artificial neural nets. In the first chapter, we describe a propagation method that prioritizes 230 infrequently mutated genes with respect to their potential contribution to cancer development. In the second chapter, we further transcend the driver-passenger dichotomy and demonstrate a gradient of cancer relevance across human genes. In the last two chapters, we present methods that simplify neural network models to render them more interpretable with a focus on functional genomic applications in cancer and beyond.

Network Approaches to the Study of Genomic Variation in Cancer

A Dissertation

presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Hussein Mohsen

Dissertation Director: Mark Gerstein

May 2022

Table of Contents

Dedication

To Layla, and to Toni Morrison and James Baldwin, whose radiant minds helped shape the ways I see the world.

Acknowledgement

I wish to thank members of my family, chosen (i.e. friends) and biological, thesis advisor Mark Gerstein, committee members: Kei-Hoi Cheung, Lajos Pusztai, and Sahand Negahban, and collaborators.

Funding

During my doctoral studies, I received the Gruber Science Fellowship, Nicholas Jabr Fellowship, and Franke Fellowship in Science and the Humanities. National Science Foundation grant DBI-1660648 and the AL Williams Professorship Fund provided to thesis advisor Mark Gerstein, and the Breast Cancer Research Foundation investigator award AWDR11559 and the Susan Komen Leadership grant SAC160076 provided to collaborator and committee member Lajos Pusztai, have also supported part of the research presented in this dissertation.

Chapter 1

Introduction

Rapid advancements in nucleotide sequencing since the early 2000's reshaped the genomic study of cancer. During this period, studies that utilize sequences from 100 or more tumors have become common, and reports on cancer etiology often transcend the sequence level to incorporate functional data on the route between genotype and phenotype. In light of the wide spectrum of discovered genomic variation across and within cancer types, research questions proliferated, and the identification of mutational contributions to cancer development took a more central position in biomedical research [1].

Historically, cancer genomics studies predominantly focused on somatic mutations and strictly divided them into two classes: drivers that confer advantageous fitness on cells to develop cancer, and passengers with assumed neutral selective advantage. To identify drivers and passengers, researchers often adopted frequency-based strategies that primarily compare a genomic region's mutational load in tumor versus healthy tissues. At times described as "mountains" spanning the landscape of genomic cancer mutations [2], frequently mutated driver genes increasingly defined the central narrative of cancers' genomic etiologies. Another strategy that confers the "driver" title on genes (or mutations) is based on the functional impact of their mutations [3]. Both strategies have been used to collate consistently updated lists of cancer drivers that often include hundreds of genes [2, 4-10].

Yet, recent large-scale studies, including those led by The Cancer Genome Atlas and the Pan Cancer Analysis of Whole Genomes projects, exposed limitations of existing driver lists. On one front, these studies further highlighted that the intricacies of cancer genomic development considerably transcend acquired somatic point mutations in coding regions. Long structural variants, epigenetic changes, and noncoding mutations and novel somatic-germline associations have all been discovered to play a significant role in the development of the disease [11, 12]. On another front, and despite the efficacy of frequency-based discovery strategies, they have missed a significant subset of infrequently mutated but functionally important candidate genes.

Relatedly, a growing body of research have raised questions about the rigidity of the driver-passenger dichotomy and suggested varying important roles contributed to mutations previously portrayed as neutral. Described as “mini-drivers”, “latent drivers”, or “hitchhikers”, subsets of passenger mutations were assigned functional roles that would facilitate the progression of cancer by altering signaling pathways and optimizing the effect of driver mutations [13-15]. Interestingly, a second set of studies suggested an opposite role of passenger mutations in what was described as a “tug-of-war” between drivers and passengers, where the combined effect of multiple passenger mutations might be either slowing down tumor growth or reducing metastatic progression [11, 16, 17].

The demonstrated effect of passenger mutations opened new venues for the study of genomic variation in cancer. In this context, we present a series of works that primarily aim to assess the functional significance of genomic variation in cancer with a relative

focus on previously unknown driver mutations. These collaborative works integrate multiple layers of biomedical data stemming from the functional genomics matrix, drug response assays, and biomedical literature text mining, and rely on two types of networks according to which the chapters of this document can be classified. In Chapters 2 and 3, we describe novel approaches to prioritizing genomic variation using genomic mutation and functional network data. In Chapters 4 and 5, we present algorithmic approaches to enhancing the interpretability of artificial neural networks trained to fulfill prediction tasks pertaining to cancer genomics and other biomedical applications.

Chapter 2

Network propagation-based prioritization of long tail genes in 17 cancer types

This chapter is based on the work described in Mohsen *et al.* [18].

2.1. Abstract

The diversity of genomic alterations in cancer poses challenges to fully understanding the etiologies of the disease. Recent interest in infrequent mutations, in genes that reside in the "long tail" of the mutational distribution, uncovered new genes with significant implications in cancer development. The study of cancer relevant genes often requires integrative approaches pooling together multiple types of biological data. Network propagation methods demonstrate high efficacy in achieving this integration. Yet, the majority of these methods focus their assessment on detecting known cancer genes or identifying altered subnetworks. In this chapter, we introduce a network propagation approach that entirely focuses on prioritizing long tail genes with potential functional impact on cancer development.

We identify sets of often overlooked, rarely to moderately mutated genes whose biological interactions significantly propel their mutation-frequency-based rank upwards during propagation in 17 cancer types. We call these sets "upward mobility genes" and hypothesize that their significant rank improvement indicates functional importance. We

report new cancer-pathway associations based on upward mobility genes that are not previously identified using driver genes alone, validate their role in cancer cell survival *in vitro* using extensive genome-wide RNAi and CRISPR data repositories, and further conduct *in vitro* functional screenings resulting the validation of 18 previously unreported genes. Our analysis extends the spectrum of cancer relevant genes and identifies novel potential therapeutic targets.

2.2. Background

Rapid developments in sequencing technologies allowed comprehensive cataloguing of somatic mutations in cancer. Early mutation-frequency-based methods identified highly recurrent mutations in different cancer types, many of which were experimentally validated as functionally important in the transformation process and are commonly referred to as cancer driver mutations. However, the biological hypothesis that recurrent mutations in a few driver genes account fully for malignant transformation turned out to be overly simplistic. Recent studies indicate that some cancers do not harbor any known cancer driver mutations, and all cancers carry a large number of rarely recurrent mutations in unique combinations in hundreds of potentially cancer relevant genes [19-25]. These genes are part of a long tail in mutation frequency distributions and referred to as “long tail” genes.

Many long tail mutations demonstrated functional importance in laboratory experiments, but studying them all and assessing their combined impact is a daunting task for experimentalists. This creates a need for new ways to estimate the functional importance

and to prioritize long tail mutations for functional studies. A central theme in finding new associations between genes and diseases relies on the integration of multiple data types derived from gene expression analysis, transcription factor binding, chromatin conformation, or genome sequencing and mechanistic laboratory experiments. Protein-protein interaction (PPI) networks are comprehensive and readily available repositories of biological data that capture interactions between gene products and can be useful to identify novel gene-disease associations or to prioritize genes for functional studies. In this chapter, we rely on a framework that iteratively propagates information signals (i.e. mutation scores or other quantitative metrics) between each network node (i.e. gene product) and its neighbors.

Propagation methods have often leveraged information from genomic variation, biological interactions derived from functional experiments, and pathway associations derived from the biomedical literature. Studies consistently demonstrate the effectiveness of this type of methods in uncovering new gene-disease and gene-drug associations using different network and score types. Nitsch *et al.* [26] is one of the early examples that used differential expression-based scores to suggest genes implicated in disease phenotypes of transgenic mice. A study by Lee *et al.* shortly followed to suggest candidate genes using similar propagation algorithms in Crohn's disease and type 2 diabetes [27]. Other early works that use propagation account for network properties such as degree distributions [28] and topological similarity between genes [29-31] to predict protein function or to suggest new candidate genes.

Cancer has been the focus of numerous network propagation studies. We divide these studies into two broad categories: (A) methods that initially introduced network propagation into the study of cancer, often requiring several data types, and (B) recent methods that utilize genomic variation, often focusing on patient stratification and gene module detection (for a complete list, see [32]).

Köhler *et al.* [33] used random walks and diffusion kernels to highlight the efficacy of propagation in suggesting gene-disease associations in multiple disease families including cancer. The authors made comprehensive suggestions and had to choose a relatively low threshold (0.4) for edge quality filtering to retain a large number of edges given the limitations in PPI data availability in 2008. Shortly afterwards, Vanunu *et al.* [34] introduced PRINCE, a propagation approach that leverages disease similarity information, known disease-gene associations, and PPI networks to infer relationships between complex traits (including prostate cancer) and genes. Propagation-based studies in cancer rapidly cascaded to connect gene sequence variations to gene expression changes using multiple diffusions [35], to generate features used to train machine learning models that predict gene-disease associations in breast cancer, glioblastoma multiforme, and other cancer types [36, 37], or to suggest drug targets in acute myeloid leukemia by estimating gene knockout effects *in silico* [38].

Hofree *et al.* introduced network-based stratification (NBS) [39], an approach that runs propagation over a PPI network to smoothen somatic mutation signals in a cohort of patients before clustering samples into subtypes using non-negative matrix factorization.

Hierarchical HotNet [40] is another approach that detects significantly altered subnetworks in PPI networks. It utilizes propagation and scores derived from somatic mutation profiles as its first step to build a similarity matrix between network nodes, constructs a threshold-based hierarchy of strongly connected components, then selects the most significant hierarchy cutoff according to which mutated subnetworks are returned. Hierarchical HotNet makes better gene selections than its counterparts with respect to simultaneously considering known and candidate cancer genes, and it builds on two earlier versions of HotNet (HotNet [41] and HotNet2 [42]).

These studies have addressed varying biological questions towards a better understanding of cancer, and they have faced limitations with respect to (i) relying on multiple data types that might not be readily available [35, 36], (ii) limited scope of biological analysis that often focused on a single cancer type [35, 38], (iii) suggesting too many [38] or too few [37] candidate genes, or (iv) being focused on finding connected subnetworks, which despite its demonstrated strength as an approach to study cancer at a systems level might miss lone players or understudied genes [35, 40-42]. To address these issues and parallel the emerging focus on long tail genes and non-driver mutations [11, 13-15, 17, 20, 22, 23], we build on the well-established rigor of propagation and introduce a new approach that particularly prioritizes rarely to moderately mutated genes implicated in cancer. Our analysis spans 17 cancer types and relies centrally on two data types: mutation frequency and PPI connectivity data. We hypothesize that a subset of long tail genes, originally with low mutation frequency ranks, can leverage their positionality in PPI networks and the mutational burden within their extended neighborhoods to play an important role in cancer

as signaled by the much higher individual ranks they attain after propagation. These genes are not merely pinpointed based on their high post-propagation ranks, but rather on the strong improvement in their pre- and post-propagation ranking difference that exceeds stringent measures. Hence, we describe these genes throughout this chapter as upward mobility genes (UMGs). To the limits of our knowledge, this is the first propagation approach that focuses entirely on long tail genes.

We efficiently identify a considerable number of UMGs (n = 28-83 per cancer type) and demonstrate their functional importance in cancer on multiple levels. Using somatic mutation data from the TCGA and two comprehensive PPI networks with significant topological differences, STRING v11 and HumanNet v2, we detect UMGs in BRCA, CESC, CHOL, COAD, ESCA, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, READ, STAD, THCA, and UCEC. These genes reveal a significant number of regulatory pathway associations that would be overlooked when relying on known driver genes alone. Further, *in silico* analysis demonstrates that UMGs exert highly significant effect on cancer cell survival *in vitro* with cancer type specificity, and they outperform genes suggested by other network methods with respect to this impact on cancer cell survival. We then validate a previously unreported subset of the identified genes *in vitro* through siRNA knockout experiments. Finally, we perform a analysis of UMGs' positionality in a combined STRING-HumanNet v2 PPI network to classify each UMG as a potential cancer driver, drug target, or both. Together with known drivers, we hope that UMGs will draw a more complete portrait of the disease.

2.3. Results

2.3.1. Overview

First, we generate PPI networks specific to each of 17 cancer types in the TCGA using only genes that are expressed in a given cancer type (Figure 2.1a). We use the STRING and HumanNet v2 networks that have different topologies and information channels for constructing the networks and use only high-quality edges. We then perform propagation over each network, where each sample's somatic mutation profile includes a quantized positive value $\in$ [1,4] for genes with mutations, and 0 otherwise (Figure 2.1b). Next, we perform the Mann Whitney U test to assess the significance of propagation-based rankings by measuring the enrichment of known functionally important COSMIC genes towards higher ranks in post-propagation lists. Results demonstrate high statistical significance across all studied cohorts ($p < 10^{-5}$) demonstrating the validity of the method to identify genes with functional importance. We then calculate the difference in pre- (i.e. raw mutation frequency) and post-propagation ranking for each gene. Genes that move up in the rank order in the post propagation list are called UMGs. We construct a preliminary UMG list for each cancer cohort based on stringent final rank cutoff and upward rank increase (i.e. upward mobility) threshold. In this chapter, genes whose rank significantly improves during propagation and land in a pre-defined top block of post-propagation ranked lists are retained (Figure 2.1c). Using this strategy, our approach focuses on long tail genes and excludes frequently mutated genes (including classical cancer drivers) that occupy high ranks before propagation and therefore cannot meet the upward mobility

threshold. We identify UMGs separately for each of the 17 cancer types. To further filter UMGs for potential functional importance, we remove genes with minimal or no impact on corresponding cancer cell survival after gene knockdown in the Cancer Dependency Map Project (DepMap) [43]. This step eliminates 4-13% of UMGs (Figure 2.1d). We finally analyze the biological and topological properties of the shortlisted UMGs on pan-cancer and cancer type levels (Figure 2.1e).

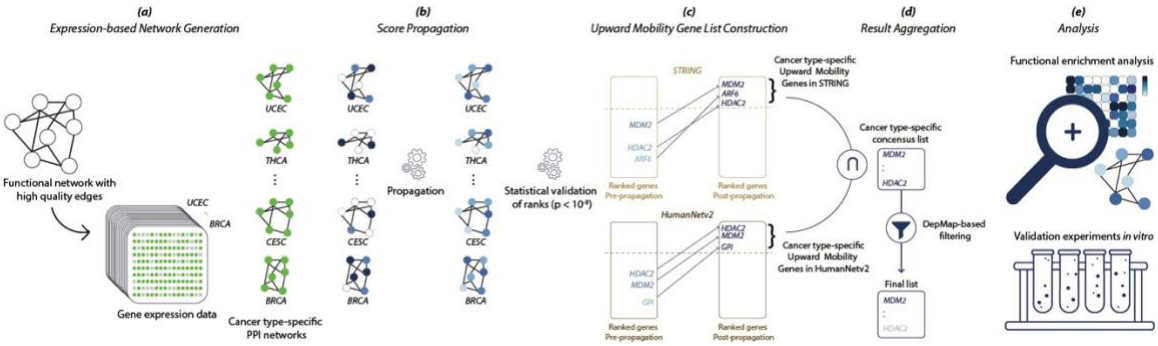


Figure 2.1. *Schematic overview of the UMG identification strategy.*

2.3.2. UMGs across 17 cancer types

We report 230 UMGs across 17 cancer types. UMG lists capture the expected biological heterogeneity of cancer types: 76 genes (33%) are specific to one cancer type, 116 (50.4%) to 2-9 types, and only 38 (16.5%) to 10 or more types. The longest list of UMGs corresponds to CESC (n = 83 genes) and the shortest to CHOL (n = 28). Hierarchical complete linkage clustering of cancer types (right of Figure 2.2) using UMG list membership and DepMap dependency scores of the genes (which reflect their importance

in cell growth) reveals interesting patterns. Similar to results based on driver gene sets identified in [25], subsets of squamous (ESCA, HNSC, and LUSC) and gynecological (BRCA, CESC, and UCEC) cancers cluster together. Close clustering results also correspond to the lung (LUAD and LUSC) and colon and rectum (COAD and READ) as tissues of origin, while others match with the rates of driver mutations across cancer types (i.e. Figure 1D in [25]), particularly (i) STAD and CESC, (ii) KIRP, READ, and COAD, and (iii) LUSC, LUAD, HNSC, ESCA, and LIHC, suggesting similarities between driver and long tail mutational patterns. Interestingly, UMGs specific to a single cancer type (left of Figure 2.2) include a considerable number of genes whose products have similar functions such as *COL4A1* and *COL1A1* that encode different types of collagen (specific to ESCA), and triplets of genes that encode proteins in the 26S proteasome complex (*PSMC1/2/3*, specific to UCEC) and mitogen-activated kinases (*MAPK1* and *MAP2K1/2*, specific to THCA). Functional gene clusters shared among cancer types include *DYNC1LI2/I2/H1* that encode different components of the cytoplasmic dynein 1 complex and *PPP1CC/1CA/2CB/2CA* that encode subunits of protein phosphatase enzymes. The circos plot [44] of Figure 2.2 shows the distribution of UMGs across cancer types, their relative ranks within UMG lists, and their impact on cancer type-specific cell survival.

2.3.3. UMGs reveal known and novel cancer-pathway associations

Biological enrichment analysis of UMGs, separately and in combination with known drivers, confirms some already known functional importance of the UMGs and suggests

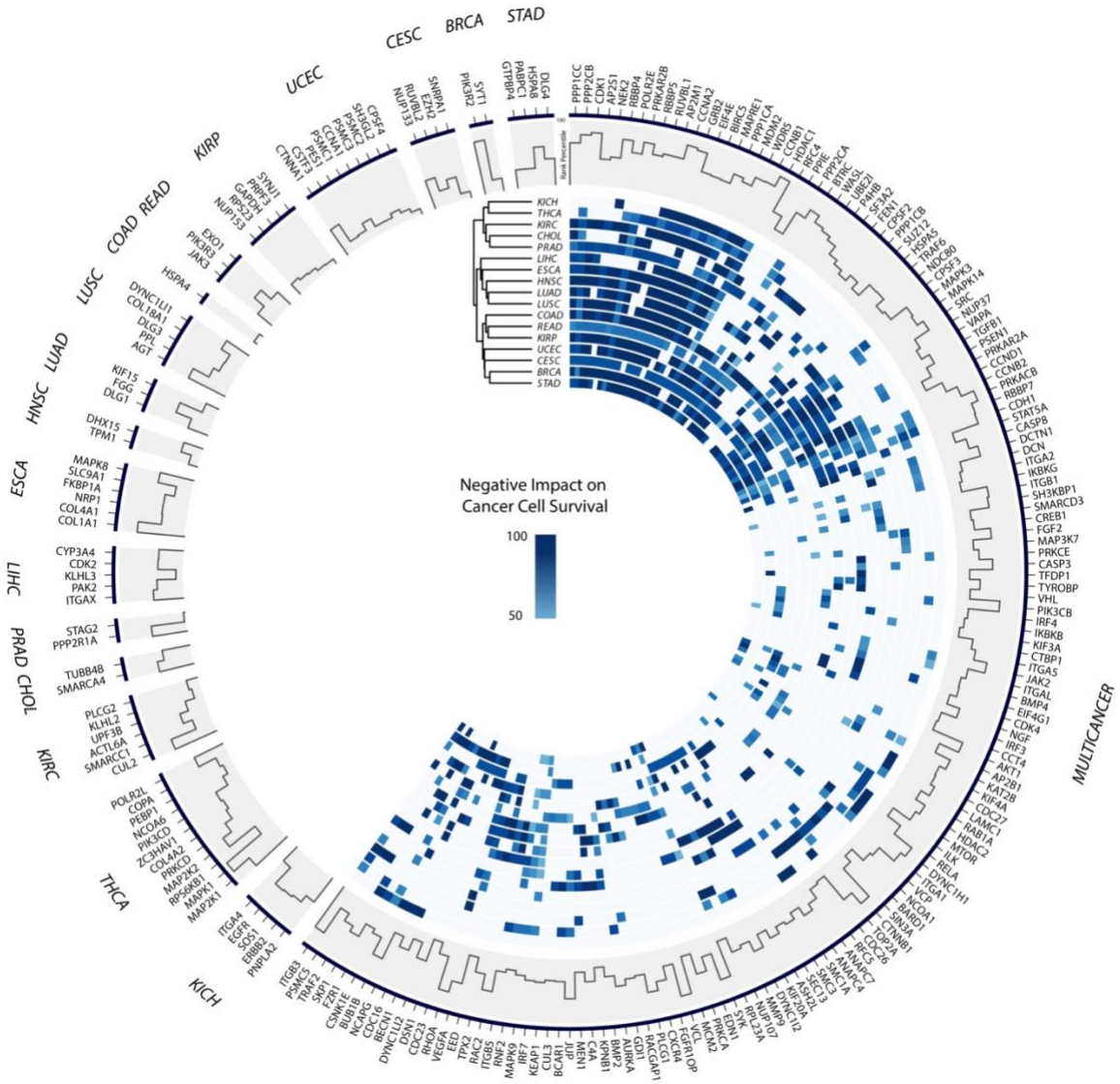


Figure 2.2. **Distribution of UMGs across 17 cancer types.** *Right: genes in 2 or more cancer types. Dendrogram is based on hierarchical clustering of heatmap rows. Each heatmap value corresponds to a percentage-based score of a cancer type's cell lines whose survival is negatively impacted by a gene's knockout. For each value, the maximum percentage across RNAi and CRISPR experiments is selected. Left: cancer type-specific genes. Histogram throughout the plot corresponds to the normalized rank of each UMG in the lists it belongs to.*

new associations between cancer types and biological pathway alterations. UMGs analyzed alone or together with known cancer drivers have statistically significant associations (Benjamini *p-adjusted* < 0.05) with most of the oncogenic pathways (8 out of 10) curated by Sanchez-Vega *et al.* [45] (Figure 2.3a). These results indicate that UMGs are members of known biological pathways and can broaden the study of biological processes that contribute to malignant transformation. This is particularly relevant in cancers where driver gene-based pathway associations revealed only a few relevant pathways (e.g. KICH and CHOL in [25]). Interestingly, the p53 pathway has only a small number of associations with UMGs in contrast to the many associations we detected with the TGF-beta and Hippo signaling pathways. Other known cancer pathways are also altered by UMGs and include Notch, HIF-1 and mTOR. Notably, the number of cancer type-specific pathway associations does not correlate with the size of UMG lists. For example, KICH, which has one of the smallest lists of UMGs (n = 41 genes), has a sizeable set of pathway associations, while CESC with the largest UMG list (n = 83) has considerably fewer associations. These findings suggest greater diversity in altered biological processes that lead to development of KICH compared to CESC.

On the pancancer level, we partitioned enrichment results for all 230 UMGs into 9 major functional clusters based on biological function (Figure 2.3b). Using EnrichmentMap (EM) [46], we built a network of intra- and inter-cluster similarity measured through gene overlap between enrichment entities (i.e. pathways, biological processes and molecular functions; Methods). Connectivity patterns within the EM network provide insights into

the sets of entities and UMGs. Within the clusters, we identified biological entities with high connectivity (red labels, Figure 2.3b). These entities include oncogenic pathways such as PI3K-AKT, RAS, and mTOR, and important biological processes including cell matrix adhesion and chromatin remodeling. Their high connectivity is often driven by a selected subset of UMGs with high frequency in their constituent edges (Table 1.1). Subsets of these frequent UMGs encode subunits of proteins and members of protein complexes with strong association with cancer (e.g. *PIK3R2/R3/CB/CD*'s products in phosphatidylinositol kinases (PI3Ks) [47], and *IKBKB/G*'s products that are regulatory subunits in an inhibitor of the Nuclear Factor Kappa B kinase (NFKB) [48]). Given their significant and wide range of biological functionality, these genes constitute a potential subset of potent drug targets. A similar analysis on KEGG mega-pathways corresponding to diseases and infections revealed another subset of frequent UMGs and demonstrated that UMGs are generally important genes that participate in broader biological processes than cancer alone (Figure 2.3c, Table 1.1). Observed associations include well-studied ones between multiple cancers and Hepatitis C [49], Type II Diabetes Mellitus [50, 51], and HTLV-I infection [52], and new ones such as the potential association with COVID-19 [53].

2.3.4. UMGs impact survival of cancer cells in vitro

To assess the functional importance of UMGs in cancer cell survival *in vitro*, we obtained their cancer type-specific dependency scores from the DepMap project. DepMap reports results on comprehensive genome-wide loss of function screening for all known human genes using RNA interference (RNAi) and CRISPR to estimate tumor cell viability after

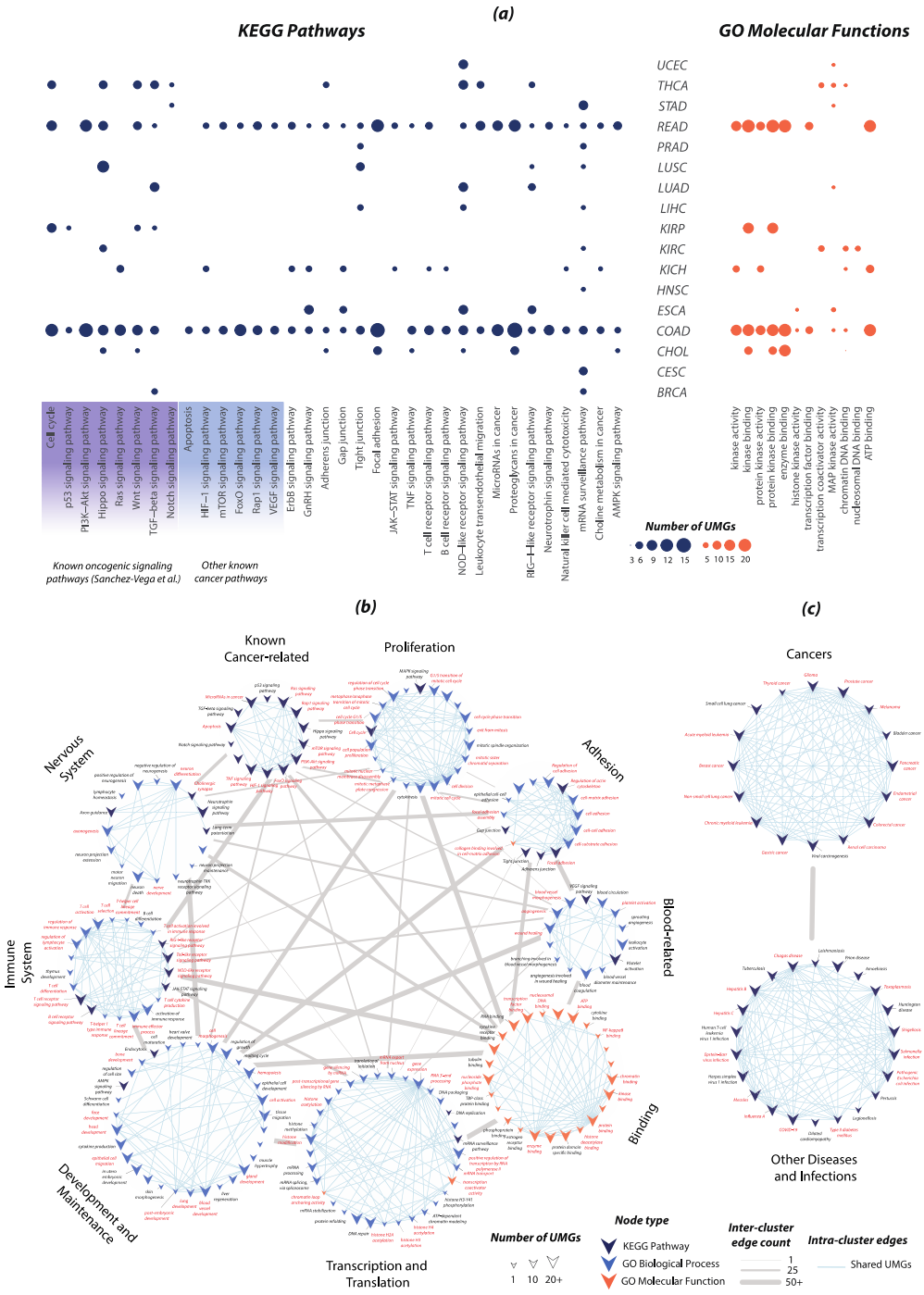


Figure 2.3. Biological enrichment results for UMGs at cancer type and pancancer levels. a *UMGs uncover known and novel associations between cancer types and biological pathways. Enrichment analyses are performed for each cancer type's combined list of UMGs and drivers. Shown results correspond to significant pathway and molecular*

function associations exclusively uncovered by UMGs. **b** *Pancancer analysis of all 230 UMGs allows for the identification of biological pathways, processes and functions strongly associated with UMGs (in red) that suggests potential therapeutic targets.* **c** *Similar analysis to* **b** *on clusters of KEGG mega-pathways uncover disease-disease and disease-infection associations pertaining.*

gene silencing in hundreds of cancer cell lines. The CRISPR dataset includes 990 cell lines, and the RNAi dataset includes 712 cell lines [43]. A dependency score of 0 corresponds to no effect on cell viability, and a negative score corresponds to impaired cell viability after knocking down the gene; the more negative the dependency score, the more important the gene is for cell viability. We used the most recent data release that accounts for batch and off-target effects and therefore provides more accurate estimates of functional impact [54].

We found that cancer type-specific mean dependency scores of UMGs is higher (i.e. more negative) than non-UMGs' across all 17 cancer types, and in both CRISPR and RNAi experiments. This indicates that knockout of UMGs consistently yields a stronger negative effect on cancer cell survival than that of non-UMGs (Mann-Whitney U test, $p < 5 \times 10^{-3}$, Methods).

Our UMG detection method is entirely focused on prioritizing long tail genes for functional importance. Most existing network methods focus their assessment on uncovering known cancer genes or are geared towards other goals—such as detecting subnetworks that maximize coverage of mutational profiles or are highly mutated—and therefore may be

Functional Cluster	**Frequent UMGs**
Known Cancer-related	*PIK3R2, PIK3R3, AKT1, IKBKB, MAPK1, MAPK3, PIK3CB, PIK3CD, MAP2K1, MAP2K2*
Proliferation	*CCND1, BUB1B, CDC16, ANAPC4, ANAPC7, CDC23, CDC26, CDC27, CUL3, TGFB1, AURKA, CDK1, CDK2, CDK4, CCNB1, NDC80*
Adhesion	*ITGB1, ITGB5, RHOA, SRC, ITGA2, ITGA4, VCL*
Transcription and Translation	*RUVBL2*
Binding	*SRC, RELA*
Immune System	*TRAF6, MTOR, IRF4, IKBKB, IKBKG*
Cancer Mega-pathways	*CCND1, PIK3R2, PIK3R3, GRB2, EGFR, AKT1, MAPK1, MAPK3, PIK3CB, SOS1, PIK3CD, MAP2K1, MAP2K2*
Other Diseases and Infections Mega-pathways	*CASP3, MAPK14, CASP8, PIK3R2, PIK3R3, TRAF6, AKT1, MAP3K7, IRF3, IKBKB, IKBKG, MAPK1, MAPK3, PIK3CB, RELA, MAPK8, PIK3CD, MAPK9*

Table 1.1. *Frequent UMGs driving high connectivity within EnrichmentMap functional clusters.*

less efficient to prioritize long tail genes. To have a better understanding of the specifications of UMGs, we compared their impact on the survival of cancer cell lines to that of non-driver genes selected by five other methods. Three of these methods are

propagation-based and include FDRNet [55], Hierarchical HotNet (HHotNet) [40]—in 3 different settings, and Zhou *et al.*'s propagation algorithm that resembles random walk with restart—in its original and edge-normalized settings [56]. The other two include nCOP [57], a non-propagation network method that recently demonstrated an ability to uncover non-driver genes across multiple cancer types, and MutSig [58], which identifies genes mutated more often than expected in a given cohort. HHotNet reported statistically significant results after the integration over both PPI networks in only 5 out of the 17 cancer types. Hence, we included two other settings (largest and all subnetworks) where the method was able to report statistically significant results in one network. FDRNet successfully generated results on STRING, and its reported results across cancer types are based on this network (Methods).

Almost all methods' generated gene sets had a knockdown negative impact on cancer cell survival, but UMGs had the strongest impact across cancer types and in both CRISPR and RNAi experiments (Figure 2.4a). The median percentage-based score of cell lines negatively impacted by UMGs' knockout is also consistently higher than that for genes selected by the other methods in 28 out of the 34 cancer type-assay combinations (Figure 2.4b), with the remaining 6 including 4 ties. Notably, a number of UMGs have an extremely strong negative impact on cell survival across cancer types. For instance, PRAD, READ, and THCA sets include genes with mean DepMap CRISPR score < -2 in their cell lines, and all other cancer types except HNSC include genes with score < -1.7. Similar results were also obtained for these comparisons before the optional DepMap filtering step that only removed 4-13% of UMGs. As FDRNet, HHotNet, Zhou *et al.*, nCOP and MutSig do

not solely focus on long tail genes and gene sets generated by these methods include known cancer drivers, we performed the same comparisons after including known cancer-specific drivers from all gene lists, which also produced similar results (Supplementary Figure 2). Concurrently including both subsets of UMGs (pre-DepMap filtering and drivers) produced similar results across cancer types as well (Supplementary Figure 3).

2.3.5 UMGs as "weak drivers" and potential novel drug targets

The aim behind identifying UMGs is to expand the repertoire of cancer relevant genes in line with recent studies whose results defy the neutrality of long tail genes or passenger mutations in carcinogenesis [11, 13-15, 17, 20, 22, 23]. In this section, we categorize each UMG as a potential "weak driver" that may complement known drivers, a candidate drug target whose inhibition could arrest cancer growth, or both, based on positionality in PPI networks relative to currently known drivers.

In the propagation framework we use, two of the most important factors that determine a node's score after convergence are the number of high scoring nodes within its neighborhood and the connectivity of these neighbors. For a node to rank higher, the best case scenario involves having near exclusive connections with multiple neighbors ($k \geq 1$ steps) whose initial scores are high. We examine these properties of each cancer type's UMGs. We use a composite PPI network that merges signals from STRING and HumanNet v2 by including the union of high-quality edges of both networks. Figure 2.5 shows a

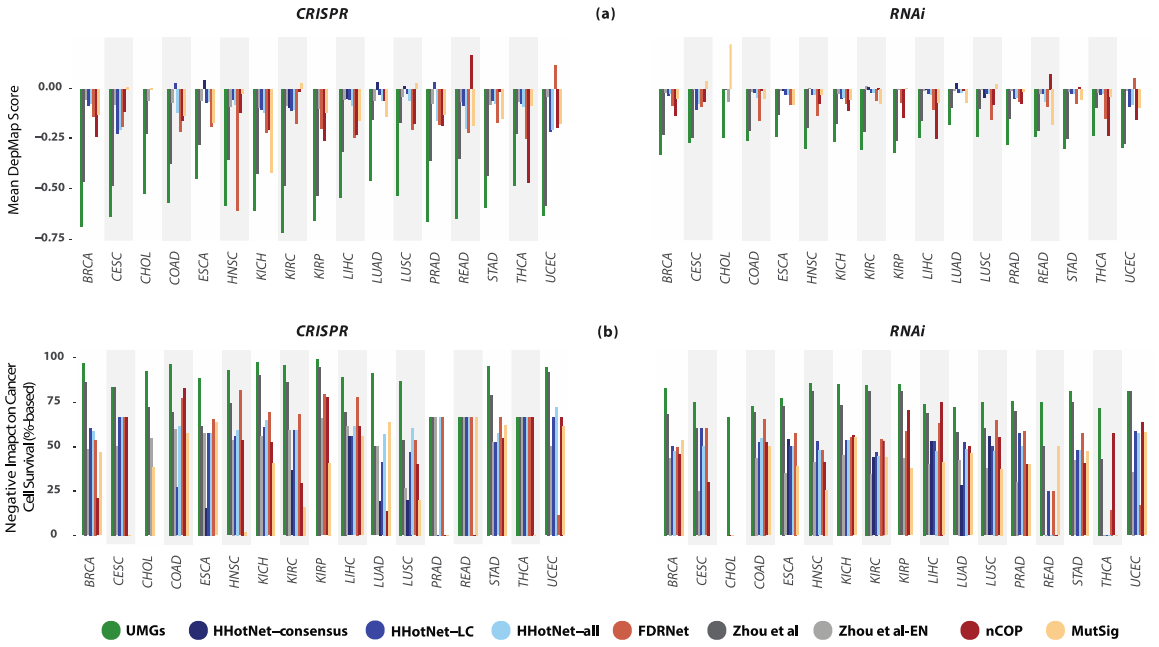


Figure 2.4. Comparisons with other methods. a *UMGs demonstrate considerably stronger (CRISPR- and RNAi-measured) impact on survival of cancer cell lines than other non-driver genes suggested by HHotNet (in 3 settings), FDRNet, Zhou et al. (in its original and edge-normalized settings), nCOP, and MutSig. Higher negative values indicate greater negative effect on cell survival after gene knockdown.* **b** *UMGs' strong impact on the survival of cancer cell lines is significantly broader than that of genes selected by other methods. The median percentage-based score of cancer cell lines negatively impacted by UMGs' knockout is consistently higher with cancer type specificity.*

representative network that corresponds to BRCA, with all others included in the supplement (Supplementary Figures 4-19). For convenience in visualization, we include immediate neighborhoods of each node and the UMG-driver edges only.

The first category of UMGs includes genes connected to high scoring known drivers (Figure 2.5 left side, olive and orange edges). By virtue of sharing connections with these frequently mutated drivers, this subset of UMGs likely includes cancer type-specific potential drug targets. The most promising UMG drug target candidates are those connected to high degree, high scoring drivers (via olive edges). Building on the same reasoning, low scoring drivers might not be the dominating force driving cancer across the majority of samples. UMGs connected to these low scoring drivers (Figure 2.5 right side, dark blue and purple edges) constitute the second category and are considered potential supplementary drivers that enhance the driver function. The third category includes UMGs with nearly no observed mutations in the TCGA cohort (i.e. very low initial score). These UMGs often form a small subset and are likely to be drug targets or false positives limited by the size of the cohort under study. In Figure 2.5 (and Supplementary Figures 4-19), they can be distinguished by their lack of node border (e.g. 6 genes in Figure 2.5: *NUP37*, *UBE21*, *POLR2E*, *IRF7*, *BIRC5*, and *EIF4E*). The fourth category includes UMGs with positive initial score and no connections to driver genes (Figure 2.5, top right grid). These genes' positive scores and connectivity with non-drivers significantly lift their rank during propagation and render them potentially overlooked weak drivers. While most UMGs are designated either potential drug targets or weak drivers, others are connected to multiple types of driver genes and accordingly might be considered for both (e.g. *RBBP5* with multi-colored edges in Figure 2.5). We also point out the well-connectedness of many UMGs, which in part allows them to have enough upward mobility to be detected by our approach. Yet, UMGs tend to have a considerably smaller number of neighbors compared to very well-studied drivers such as *TP53*, *PIK3CA*, and *BRCA1*.

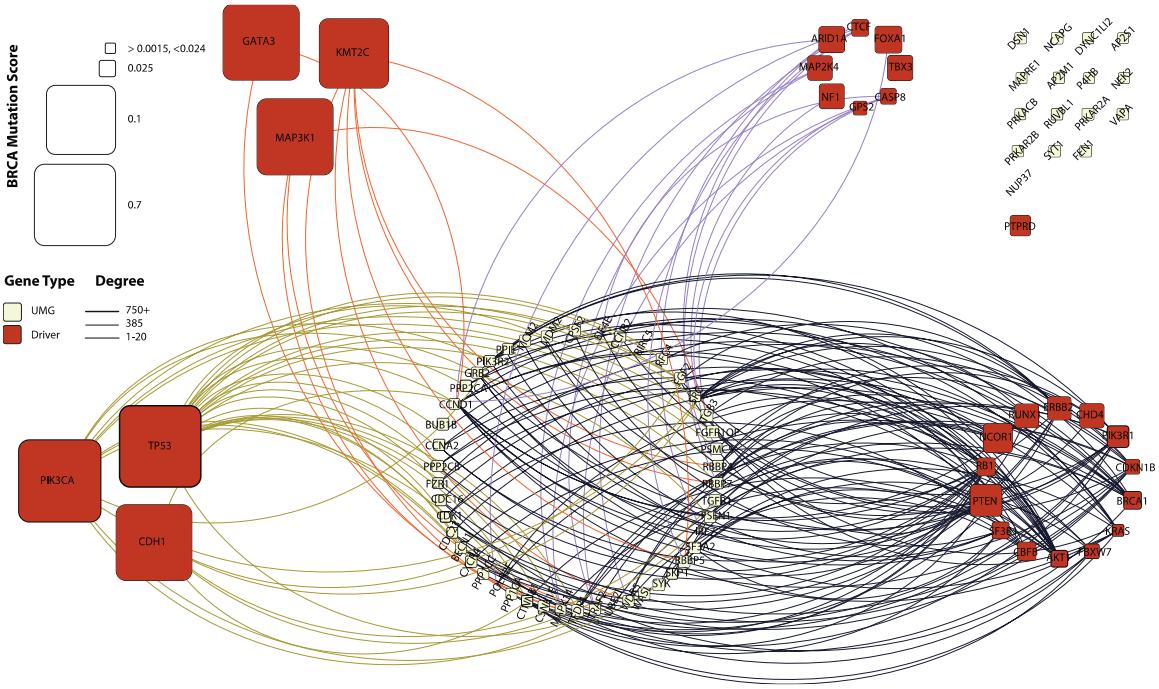


Figure 2.5. PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in breast invasive carcinoma (BRCA) suggest roles of UMGs. *Driver genes are split into categories based on initial mutation score and node degree: (i) high score, high degree (bottom left), (ii) high score, low degree (top left), (iii) low score, low degree (top right) and (iv) low score, high degree (bottom right). UMGs connected to driver subsets (i) and (ii) (olive and orange edges) and ones with no mutation score (e.g. POLR2E) are likely to be drug targets. UMGs connected to (iii) and (iv) and ones without connections to drivers (top right corner, e.g. DSN1) are likely to be "weak drivers."*

2.3.6. UMGs bridge gaps in literature and suggest novel genes

The study of cancer has long been interdisciplinary, often in the realms of various scientific and medical spheres. Disciplinary paradigms evolved over time to produce varying types of associations between genes and cancers. To further estimate the functional importance of UMGs, we manually cross referenced our UMG lists with publications and found that a large percentage of UMGs have been previously reported to play a role in cancer based on functional experiments. This percentage is as high as 85% of UMGs in cancer types like BRCA. Surprisingly, the same percentage drops to only 31% when we used CancerMine to find literature-based associations. CancerMine is an automated tool that applies text mining on existing literature to report drivers, oncogenes, or tumor suppressors across cancers. Similar results were obtained across cancer types.

2.3.7. Screening experiments validate 18 new genes in vitro

We performed a series of siRNA knockdown experiments *in vitro* to validate the DepMap results and to confirm the functional importance of selected UMGs. We selected 29 UMGs that have not been reported in the literature to be tested in gene knockdown experiments in the context of any cancer phenotype (Methods). We used 7 cell lines representing 3 types of cancer, namely H460 and HCC1299 from lung, MDAMB231, MDAMB468, BT549, and HCC187 from breast, and DU145 from prostate cancer.

Experimental results further underscore the efficacy of UMG detection to uncover functionally important long tail genes. The knockdown of 18 out of these 29 UMGs (62%) significantly decreased cell survival in 1 to 5 cell lines exceeding the threshold of 3 standard deviations with respect to negative control samples (Methods). We note that several UMGs demonstrated cell line specificity while others had a more widespread effect (affects 5/7 cell lines). These newly cancer-relevant genes have already known functions in regulating immune response (*AP2M1*, *DCTN1*, *CCT4*, *DYNC1I2*, and *DYNC1LI2*), kinase binding (*DLG3*), cell cycle progression (*SEC13*, *ANAPC7*, *CDC26*, *PSMC3*, *PPP1CC*), DNA repair (*PPIE*, *RFC5*, *POLR2E* and *POLR2L*), cell death (*VAPA*), and mRNA splicing (*SF3A2*). The list also includes *PNPLA2*, which encodes for an enzyme associated with transacetylase activity.

2.4. Discussion

Biological analysis of UMGs demonstrates strong correlations with studies performed on known cancer drivers. It also unlocks a wide range of potential associations between key pathways and cancer types and allows for classifying UMGs based on their centrality to biological functions, which in turn opens the door for a more informed drug targetability. Based on their network positionality, we propose that UMGs include “weak drivers” and cancer type-specific drug targets. Manual curation of literature confirmed that many of our UMGs were previously implicated in cancer biology in various ways, but we also identified previously unstudied potential cancer relevant genes. Yet, results suggest that we have not reached a point of data saturation with respect to analyzing long tail genes. The generation

of new and larger datasets will likely improve UMG prioritization for rare cancer types such as cholangiocarcinoma (CHOL) and chromophobe renal cell carcinoma (KICH). As the functional importance and centrality of known and new cancer relevant genes changes, network propagation results and UMG rankings will likely follow suit. This was already evident in our PPI positionality analysis: with 3 or less known genes identified in KICH and READ in Bailey *et al.* [25] and COSMIC [59], respectively, most of these cancer types' UMGs belong to the third and fourth categories (near-zero mutation scores and no connections with drivers, Supplementary Figures 9 and 16). Another example is CHOL, with its small cohort that brings most UMGs into the third category (no observed mutations, Supplementary Figure 5).

In their current arrangement in the circos plot of Figure 2.2, we also posit that the confidence associated with UMGs increases in a roughly clockwise direction, with the highest confidence to be associated with cancer type-specific genes. The incorporation of additional functional genomics data (e.g. noncoding mutations and methylation data), coupled with improvements in the accuracy of reported PPIs, will strengthen our knowledge on the role of UMGs and long tail genes more broadly. Finally, we note that bridging gaps across disciplines is often essential to biomedical knowledge production. The oncogenic validation of potential drug targets in UMGs also remains central to changing their status from potential to clinically actionable ones.

2.5. Methods

2.5.1. Mutation matrix generation

Variants from the TCGA MC3 somatic mutation dataset (n = 3.6 M) are used to generate initial scores for each of the 17 cancer types. A sample-gene matrix for each cancer type includes mutation counts restricted to splicing and coding exonic variants. Counts are then normalized by gene length, and each resulting non-zero value is finally converted to a discrete integer between 1 and 4 based on its position with respect to 50^{th}, 70^{th} and 90^{th} percentiles in the cancer type-specific normalized mutation frequency distribution. Gene ranks before and after propagation are calculated based on the mean frequency within each cohort.

2.5.2. PPI network processing

We adopt the broad definition of protein-protein interactions that encompasses direct physical interactions alongside indirect functional ones derived from co-expression, gene fusion, text mining, co-essentiality, and pathway membership datasets among others. We perform edge filtering on both PPI networks and retain edges with a confidence score equal to or higher than 0.7 across all information channels in STRING v11 and the top 10% of edges in HumanNet v2. The networks after this filtering have $|V|$ = 17,130 and 11,360 vertices and $|E|$ = 419,772 and 37,150 undirected edges, respectively. We then generate cancer type-specific PPI networks by selecting the largest connected component in each

network and filtering out (proteins of) genes unexpressed in the tumor samples of each cancer type (i.e. genes with FPKM > 15 in > 20% of tumor samples are retained).

2.5.3. Propagation score calculation

To calculate propagation scores, we use an approach that imitates random walk with restart [56]. Briefly, let the PPI network be represented as $G = (V, E)$, where V is the set of gene products and E is the set of edges. Further, let W be the weighted adjacency matrix of G. We choose to normalize W such that $W' = W \cdot D^{-1}$, where D is the diagonal matrix of column sums in W: $D = diag(\sum_{i=1}^{|G|} W_{ij})$, $1 \leq j \leq |G|$.

Let M be a $|G| \times N$ matrix with somatic mutation profiles of $N \geq 1$ samples over genes from which G's nodes originate before transcription. S_{ij} is a positive value for each $g_i \in G$ with mutations in sample $s_j \in S$, and 0 otherwise. Propagation is then executed within each sample until convergence according to the following function:

$$S^{(t+1)} = \alpha W' S^{(t)} + (1 - \alpha) S^{(0)} \quad (1.1)$$

where $S^{(0)} = M$ and $\alpha \in [0.5, 1]$. Convergence of this propagation technique is guaranteed. We summarize the proof noted in [60] below for the sake of completeness.

The function above can be written at convergence as $S = VS + (1 - \alpha) S^{(0)}$, *where* $V = \alpha W'$, which can also be rearranged into $S = (1 - \alpha)(I - V)^{-1} S^{(0)}$. For convergence to a unique, non-negative solution to be guaranteed, $(I - V)^{-1} > 0$ must hold.

Lemma 1. Largest eigenvalue of V < 1. *W'* is a column-stochastic matrix. Per the Perron-Frobenius theorem, its eigenvalues ∈ [-1, +1]. Since $\alpha < 1$, the largest eigenvalue (i.e. spectral radius) of $V < \alpha < 1$.

Lemma 2. $(I - V)^{-1}$ exists, and is non-negative. $(I - V)$ is an M-matrix since its in the form $sI - B$, with $s = 1 > 0$, $s >=$ largest eigenvalue of B (i.e. V) by Lemma 1, and $V > 0$. An M-matrix is inverse positive, hence $(I - V)^{-1} > 0$.

Convergence can also be achieved iteratively [56, 61], which we apply at a maximum of 350 iterations and is more commonly deployed with large PPI matrices for practical considerations. The value of α we pick is 0.8. Other values in the [0.6, 0.8] range have little effect on results.

2.5.4. Upward mobility gene identification

The mobility status of a gene is determined by its rank before and after propagation. A gene's rank is calculated according to its arithmetic average score across samples. For each gene $g_i \in G$,

$$Initial\ score\ IS_i = \frac{1}{N} \sum_{j=1}^{N} S_{ij}^{(0)} \quad \text{and} \tag{1.2}$$

$$Final\ score\ FS_i = \frac{1}{N}\sum_{j=1}^{N} S_{ij}^{(\infty)} \tag{1.3}$$

Let *RIS* and *RFS* be the lists of gene ranks in IS and FS, respectively, i.e. RIS_i = rank of g_i in sorted *IS* and RFS_i = rank in sorted *FS*. The mobility status of g_i, MS_i, is then calculated as the difference between RIS_i and RFS_i as:

$$MS_i = RIS_i - RFS_i \tag{1.4}$$

Since higher scores lead to a higher rank, and a higher rank has a lower value (i.e. rank 1, 2, … |*G*|), genes whose ranks improve because of propagation have positive MS values, and ones with lowered ranks (downward mobility) negative ones.

We then define upward mobility status according to two parameters: mobility β and rank threshold *T*.

$$UMG = \{g_i \mid MS_i \geq \beta\,.|G| \wedge RFS_i \leq T \ \forall\, i \in 1, 2, \ldots |G|\} \tag{1.5}$$

Mobility β value determines the minimum upward jump size a gene needs to make to be considered for UMG status. For instance, a β value of 0.1 in a PPI network with 10,000 nodes requires a gene's position to improve by a minimum of 1,000 ranks. We choose stringent values of β dictated by TCGA cohort size and the variance of each cancer type's

mutational. Cancer types with a high number of samples and/or a high variance of gene mutation frequency receive a value of 0.25 (BRCA, COAD, HNSC, LUAD, LUSC, PRAD, STAD, UCEC), others with moderate variance a value of 0.2 (CESC, KIRC, KIRP, LIHC) and 0.15 (ESCA, READ), and low variance and/or cohort size cancer types a value of 0.05 (CHOL, KICH, THCA). These values ensure that to be considered a UMG, a gene has to jump hundreds to thousands of ranks during propagation depending on the PPI network and cancer type under study. Rank threshold *T* specifies the minimum rank a gene needs to achieve after propagation to be considered a UMG. We choose $T = 1{,}000$ to strictly focus on the top 10-16% of genes (i.e. approximately top 10% in STRING and top 16% in HumanNet v2), a threshold that has proved to be effective in other studies [38].

We further apply two optional selection criteria on the final UMG lists based on (i) each gene's DepMap scores in CRISPR and RNAi experiments and (ii) propagation within multiple PPIs. Per (i), UMG becomes:

$$UMG = \{g_i \mid MS_i \geq \beta \,.|G| \wedge RFS_i \leq T \wedge DM_i \geq p,\ i \in 1, 2, \ldots |G|\}, \tag{1.6}$$

where *p* is the proportion of cancer type-specific cell lines in which a gene's DepMap score is negative (i.e. its knockout has negative impact on cancer cell survival), and DM_i is the maximum value across CRISPR and RNAi experiments. We choose $p = 0.5$ (50%), which ends up eliminating 2-10 out of 30-91 genes per cancer type. Per (ii), integration of lists across *K* PPI networks yields the intersection of lists. In this chapter, to increase confidence

is selected genes, we integrate lists over cancer type-specific STRING and HumanNet v2 networks. Formally,

$$UMG_{Final} = UMG_{G_1} \cap UMG_{G_2} \cap \ldots UMG_{G_K} \tag{1.7}$$

2.5.5. Statistical validation of rankings

To assess the validity of ranking after propagation, we tested if known COSMIC genes are ranked significantly higher than other genes using the one-sided Mann Whitney U statistical test (also known as one-sided Wilcoxon Rank Sum test). Results show a strong enrichment of COSMIC genes towards highly ranked genes for all PPI network-cancer type combinations ($p < 10^{-5}$).

2.5.6. Driver and COSMIC genes

Cancer type-specific driver genes were obtained from Bailey *et al.*'s except for COAD and READ which were combined into a single group in that study. For these two cancer types, we designated tissue-specific COSMIC v90 genes as the driver genes.

2.5.7. UMG vs non-UMG comparisons

In the first set of comparisons, Mann Whitney U one-sided test is used to compare the distribution of a percentage-based score of negatively impacted cell lines by UMGs *vs* non-

UMGs in each cancer type. Each gene's percentage-based score value is equal to the percentage of its negative DepMap scores among *k* cancer type-specific cell lines and the average of these values (to account for distribution of DepMap scores across cell lines). To calculate a more stringent score and reduce false positives, we also assume the presence of at least one cancer cell line with a non-negative DepMap score, which especially accounts for cancer types with a small number of cell lines in the DepMap database. Hence, the score is the sum of each gene's $k + 1$ values mentioned above divided by $k + 2$. Alternative hypothesis for each of the Mann Whitney U tests is $H_1 = \psi(UMG)$ is shifted to the right of $\psi(\overline{UMG})$, where $\psi(X)$ is the percentage-based distribution of negatively impacted cell lines over genes in set *X)*. Cancer type-specific cell lines are selected based on annotations provided in the DepMap dataset. For cancer types not represented among the cell lines in DepMap, we used values across all 750 (CRISPR knockout data) and 712 (RNAi) cell lines. A negative DepMap dependency score indicates decreased cell survival after gene knockout in a particular cell line. For RNAi experiments, we use data with enhanced batch and off-target processing as described in [54].

2.5.8. UMGs vs gene candidates identified by other network methods

Hierarchical HotNet (HHotNet) generates statistically significant results ($p < 0.05$) in only 5 of the 17 cancer types after integrating its results for both PPI networks (HHotNet-consensus): ESCA, KIRC, LIHC, LUAD and LUSC. As a result, we include HHotNet results from two other settings described below. In 13 cancer types, HHotNet generates statistically significant results for one of the two PPI networks, and in two others (PRAD

and READ) significant result with a relaxed threshold ($0.05 < p < 0.1$). We include HHotNet results from both the largest subnetwork (HHotNet-LC) and all subnetworks with more than one node (HHotNet-all) in comparisons. Namely, for 15 cancer types, we choose results from STRING in BRCA, ESCA, HNSC, KICH, KIRC, LIHC, LUAD, LUSC, STAD and THCA and from HumanNet v2 in CESC, COAD, PRAD, READ, and UCEC. In CHOL and KIRP, HHotNet results were not statistically significant for both PPI networks, so we exclude results for this method. In all runs, we execute HHotNet in default settings with 1000 permutations using the second controlled randomization approach suggested in [40]. For FDRNet, we run the method to detect subnetworks for all seed genes and in default settings. We convert MutSig2CV [58] *p*-values across TCGA cohorts to local FDR values using the scripts provided by FDRNet. We use FDRNet results for 16 cancer types over the STRING network as this method was not able to detect any subnetwork over HumanNet v2 for almost all seed genes (664/673, 98%). No FDRNet results could be produced for CHOL. In nCOP, we use lists of rarely mutated genes reported in [57] (Figure 2.4) on the TCGA somatic mutational dataset in 15 of the 17 cancer types studied in our chapter (all except CHOL and ESCA). For Zhou *et al.*'s propagation method, we select the top *k* genes identified post-propagation, where *k* is the equivalent number of UMGs for each cancer type across networks. In its edge normalized setting, we divide each gene's post-propagation score by the same score when propagation $\alpha = 1$ (i.e. ignoring initial scores) before selecting top genes. For MutSig, we select all genes with statistically significant results (FDR < 0.1) across TCGA cohorts. As these methods do not primarily focus on long tail genes, we remove driver genes from these methods' gene lists to ensure balanced comparisons with UMGs. It is worth noting however that including driver genes

or the small percentage of UMGs filtered in the last step of the pipeline did not have a considerable impact on results (Supplementary Figures 1-3).

2.5.9. Enrichment analysis

Enrichment analysis to identify pathways, GO molecular functions, and GO biological processes is performed on g:Profiler [62]. Enrichment results with Benjamini *p*-adjusted < 0.05 are selected for analysis. Network visualization is executed using EnrichmentMap v3.0 on Cytoscape v3.8.2 [63], with a comprehensive subset of results related to cancer shown in Figure 2.3. Frequent terms highlighted in red in Figure 2.3b have ≥ 5 intra-cluster edges and those in Figure 2.3c ≥ 10 edges. Frequent UMGs in Table 1.1 are identified based on their frequent presence in edges between a cluster's nodes according (i.e. presence in ≥ 20 edges in Figure 2.3b clusters and ≥ 30 in those of Figure 2.3c).

2.5.10. PPI analysis

Composite PPI is the union of high-quality edges in STRING v11 and HumanNet v2. Initial score of each gene is the one based on somatic mutations across a cohort as described earlier. Drivers are split according to initial score and degree with thresholds of 150 and 0.075, respectively. Initial scores of < 0.0015 are zeroed to attain lower FPR. Visualization and degree calculation are executed using Cytoscape v3.8.2.

2.5.11. Manual Literature Curation of Functionally Validated UMGs

We manually cross-referenced each UMG with PubMed publications to detect which ones have been earlier reported to play a role in cancer based on functional experiments. We based results on an extensive search using the gene name AND "cancer" as keywords in PubMed. If any gene was the target of a previous functional assay, i.e. was deliberately overexpressed, suppressed, or mutated, and resulted an *in vitro* change in the proliferation or survival of cancer cell lines, it was annotated as functionally validated. Otherwise, the genes is considered not validated.

2.5.12. Experiment validation: siRNA screening and annotation

Cell lines from breast (MDAMB231, MDAMB468, BT549, HCC187), lung (H460, HCC1299), and prostate cancers (DU145) were cultured in RPMI medium supplemented with 10% HI-FBS and penicillin/streptomycin (1:100). The siRNA transfection experiments were performed at the Yale Center for Molecular Discovery. Reverse transfections were performed using 384-well tissue-culture treated plates (Corning CLS3764) pre-plated with siRNAs to achieve 20 nM final assay concentration. RNAiMax transfection reagent (Invitrogen) was added to plates according to the manufacturer's recommendations and incubated with siRNAs for 20 minutes. Cells were then seeded at plating densities optimized during assay development (MDAMB468, HCC1187, and BT549 seeded at 4000 cells per well; MDAMB231 and H460 seeded at 1000 cells per well; DU145 and HCC1299 seeded at 500 cells per well) and incubated at 37C. After 72 hours,

CellTiter-Glo (Promega) was used to monitor viability. Each screening plate contained 16 replicates of negative siRNA controls (either siGENOME Smart Pool non-targeting control #1, #2, or #4, Dharmacon) and positive siRNAs controls (siGENOME Smart Pool Human PLK1 or KIF11, Dharmacon). Signal-to-background (S/B), coefficient of variation (CV), and Z prime factor (Z') were calculated for each screening plate using mean and standard deviation values of the positive and negative controls to monitor assay performance. All cell lines were obtained from ATCC and have been thoroughly tested and authenticated by the vendor. The cell lines will be routinely monitored for correct morphology and growth characteristics to confirm cell line identity. For each cell line, test siRNA data were normalized relative to the mean of negative control samples (set as 0% effect) and the mean of positive control samples (set as 100% effect). Three standard deviations of the negative control samples were used as a cutoff to define screen actives.

2.5.13. Availability of data and materials

UMG detection code is available at https://github.com/gersteinlab/UMG [64] and https://doi.org/10.5281/zenodo.5500467 [65]. Results in the chapter are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. MC3 high-quality somatic mutation dataset is obtained from [66]. STRING v11 [67] and HumanNet v2 [68] functional network (FN) are respectively downloaded from https://string-db.org/ and https://www.inetbio.org/humannet. Gene expression data corrected for batch effect and study-specific bias are downloaded from RNAseqDB [69] at https://github.com/mskcc/RNAseqDB. Variant annotations are based on RefSeq hg19

provided via ANNOVAR 2018b [70], and gene length values are provided via the bioMart Bioconductor package [71]. Genetic dependency data from the Cancer Dependency Map [43] (for both CRISPR and RNAi experiments) are downloaded from https://depmap.org/portal/download/, MutSig2CV [58] data across cancer types from http://gdac.broadinstitute.org, COSMIC v90 census gene list from https://cancer.sanger.ac.uk/cosmic, and CancerMine v24 [72] gene lists from http://bionlp.bcgsc.ca/cancermine.

2.6. Conclusion

In this chapter, we describe a new network propagation-based approach that is particularly well suited to estimate the functional importance of rarely mutated long tail genes in cancer. The method is computationally efficient and is based on change in ranking before versus after network propagation. We show that upward mobility genes that attain significant improvements in mutation score-based ranking after propagating through PPI networks are enriched in functionally relevant genes. By virtue of high post-propagation ranks, cancer-related biological function, and significantly strong impact on cancer cell line survival, our approach prioritizes long tail genes across 17 cancer types. To reduce false positivity rate, we integrate results over two major networks, filter out nodes whose genes are unexpressed in each cancer type's tumor samples, and statistically validate rankings and cell survival impact. Computational and *in vitro* analyses further highlight the importance of reported genes and open the door for an expanded spectrum of gene cancer relevance.

Chapter 3

Network distance-based cancer relevance of human genes

This chapter is based on the work described in Qing *et al.* [73].

3.1. Abstract

The evolving narrative underlying the genomic basis of carcinogenesis engenders the need for new approaches to measuring the cancer relevance of human genes. Transcending the driver-passenger dichotomy, we analyze a compendium of human genes and their potential contribution to cancer development based on their positionality in functional biological networks. We categorize genes into 1-, 2-, 3-, and >3-steps removed from the nearest core cancer gene (CCG) in the STRING network and demonstrate that the cancer-related functional contribution of the genes in these different neighborhood categories decreases as their distance from the CCGs increases. Genes closer to CCGs seem to have higher levels of (a) impact on cancer cell survival *in vitro*, (b) negative selection pressure in healthy populations, and (c) somatic mutational burden. These results suggest that the cancer relevance of human genes transcends the driver-passenger binary and might better be portrayed as a wider spectrum of gradient effects.

3.2. Background

The identification of genes whose altered function, or lack thereof, contribute to carcinogenesis has been a central question in cancer research. Since the 1970's, researchers deployed a variety of experimental media—i.e. *in vivo*, *in vitro*, and *in silico*—to prioritize hundreds of genes with respect to their contribution to cancer [25, 59, 74, 75], often describing them as "cancer drivers" to imply therapeutic potential. The majority of somatic mutations, which appear random and are not recurrent in a cancer type, had until recently been described as "passengers" to suggest that they confer no selective advantage on cancer cells. Yet, it has also become increasingly clear that deleterious mutations in driver genes alone cannot lead to initiation and progression of all tumors. Large scale efforts to sequence thousands of whole exomes and whole genomes revealed tumors without previously known driver mutations and identified functionally significant genomic alterations associated with the disease—point mutations, indels, and structural variants [25, 75-77]. Relatedly, recent studies suggested different cancer-related roles of "passengers" [11, 16, 17] or elaborated on the potential role of germline genomic variation underlying predisposition to the disease [78-80]. Altogether, these advances highlighted that the spectrum of cancer-relevant genes is broader than our current models suggest.

From a systems biology perspective, gene products, in their "healthy" state, interact to maintain homeostasis. Deleterious alterations to one or more of the products' functions can be assumed to have cascading effects on their immediate or surrounding neighbors. The

availability of comprehensive lists of interactions, such as the ones represented as a functional network in the STRING database that we use in this chapter, allow for assessing the potential significance of genes based on their proximity to a starting subset known to be associated with a phenotype of interest, herein cancer. We hypothesize that proteins physically associated with, or known to directly interact with, an experimentally or clinically validated core cancer gene (CCG) can also have an impact on cancer biology and denote these genes as "one step removed" from a CCG (Figure 3.1). By extension, we also assume that genes that directly interact with the “one-step removed genes” might also

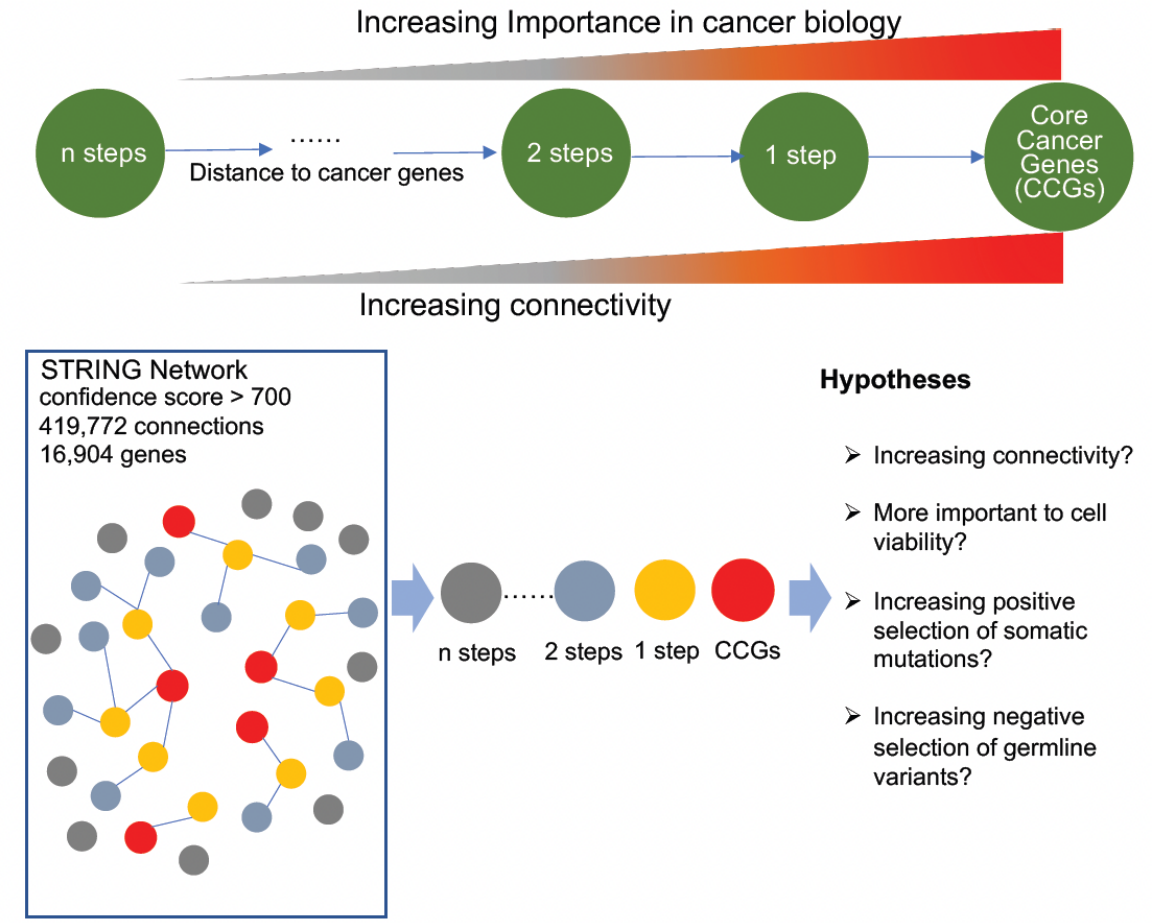


Figure 3.1. Study schema. *Overview of our hypothesis that genes closer to core cancer genes in STRING network are more functional important in cancer development.*

influence cancer biology, although to a lesser extent. Based on this model, one could categorize human genes into one-, two-, three-, and > three-steps removed from the nearest CCG in the STRING network. We perform a series of analyses based on this paradigm to assess the cancer relevance of genes (herein used interchangeably with gene products) in each of these four categories to somatic mutational burden, effect size, germline selection pressure, and *in vitro* cancer cell survival.

3.3. Results

3.3.1. Gene connectivity descreases with distance to CCGs

We select the 486 validated cancer genes of the Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets list as the core genes to resemble the starting point of our analyses. Next, we measure the shortest distance between each of the 16,904 in the high-quality (i.e. edge score >700) STRING functional network and any of the CCGs. Each gene is then assigned one of four categories based on the resulting distance: n=1 for immediate CCG neighbors, n = 2 for 2-step removed ones, and so forth for n = 3 and n > 3. The resulting distribution indicates that 6791, 7742, 1587, and 362 genes are 1-, 2-, 3-, and >3-steps removed from CCGs (Figure 3.2a, 3.2b). These results demonstrate that the majority of genes are in the immediate (i.e. 1-step) or close (i.e. 2-steps) neighborhoods of CCGs, and that the latter set unsurprisingly plays a central role in PPI networks. The cancer relevance of the association with these genes is further supported when considering citations in the cancer literature and connectivity (i.e. degree or number of edges) in the functional network. Based on citations extracted by CancerMine [72], an

automated text mining approach, we demonstrate that 18.2%, 6.1%, 3.8%, and 2.2% of the 1-step, 2-step, 3-step, and >3-step genes are associated with cancer in the literature, respectively (Figure 3.2c). Similarly, with connectivity values range from 1 to 1,435, CGCs show higher connectivity than other genes, and connectivity values decrease with distance from this gene set (Figure 3.2d).

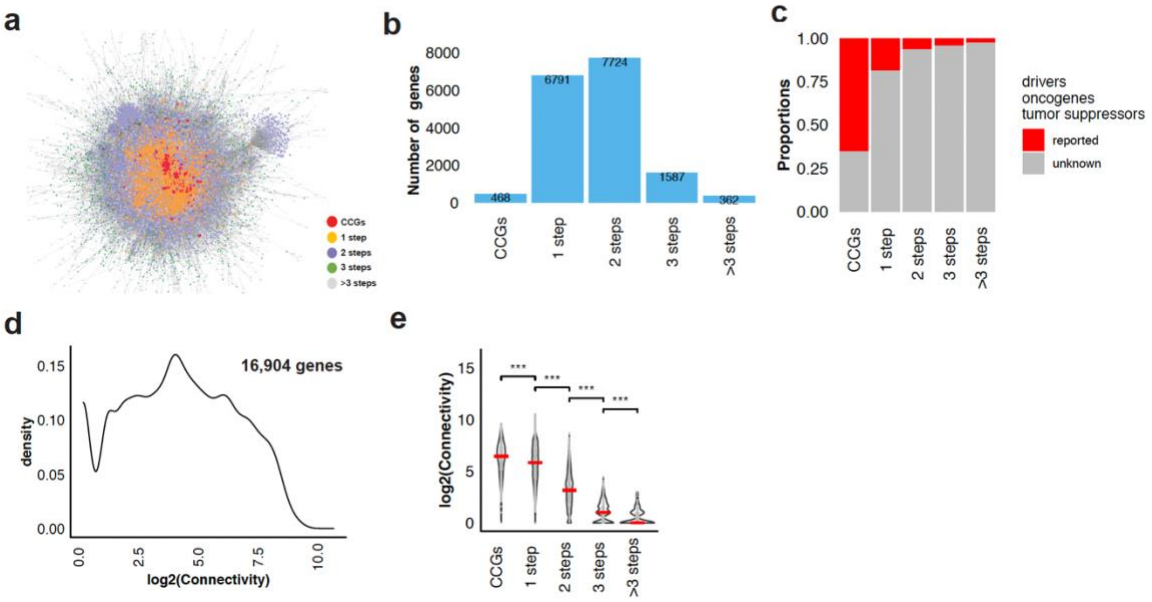


Figure 3.2. Connectedness of cancer genes. a *STRING protein interaction network. Each dot represents a gene, colors indicate distance from core cancer genes. The grey lines show between-gene connections.* **b** *Number of human genes in 4 cancer gene neighborhood categories.* **c** *Proportion of genes implicated in cancer biology in the literature (reported or not in connection with cancer) by neighborhood categories.* **d** *Distribution of log2-transformed connectivity score of 16,904 human genes in STRING.* **e** *The distribution of log2 transformed connectivity score for the cancer genes and 4 neighborhood categories.*

*One-sided Mann–Whitney U test (values of closer neighborhood genes are greater than that of all the genes in the remoter steps) p-values are symbolized by ***, **, * corresponding to p < 0.0001, 0.001, and 0.01, respectively. Red bars correspond to the median of the distributions. CCGs: core cancer genes.*

3.3.2. Impact on cancer cell viability decreases with distance to CCGs

Another aspect of a gene's functional significance is its impact on cell viability. To assess the importance of the four CCG-based gene categories, we leverage the Cancer Dependency Map (DepMap) described in the previous chapter. In summary, DepMap reports the results of genome-wide pooled loss-of-function (CRISPR and short hairpin (sh) RNA interference) screening experiments to provide estimates of *in vitro* gene impact on the viability, i.e. a dependency score, of cancer cell lines [43]. DepMap scores are reported for 712 cell lines in CRISPR and 563 lines in shRNA (i.e. RNAi) experiments. A dependency score of 0 corresponds to no impact on cell survival, and a negative score's magnitude corresponds to the level of impaired cell viability after gene knockdown. We calculated the average dependency scores for each of the four gene categories based on distance to CCGs. For both CRISPR and RNAi experiments, a significant pattern emerged further highlighting the gradient of gene importance relative to distance from CCGs: (i) CCGs and step-1-removed genes had the most negative dependency scores, i.e. stronger impact on cell viability, (ii) step-2 and step-3- removed genes had scores closer to 0, and (iii) >3-step-removed genes had an average of 0 indicating no impact. This pattern suggests that a large number of genes, rather than a selected subset of "core" or driver cancer genes, affect cancer cell survival, and that this effect is proportional to distance from CCGs in the

functional network (Kendall's τ z-statistic = 15.13 in the CRISPR dataset (Figure 3.3a), 19.33 in RNAi (Figure 3.3b), $p < 10^{-5}$).

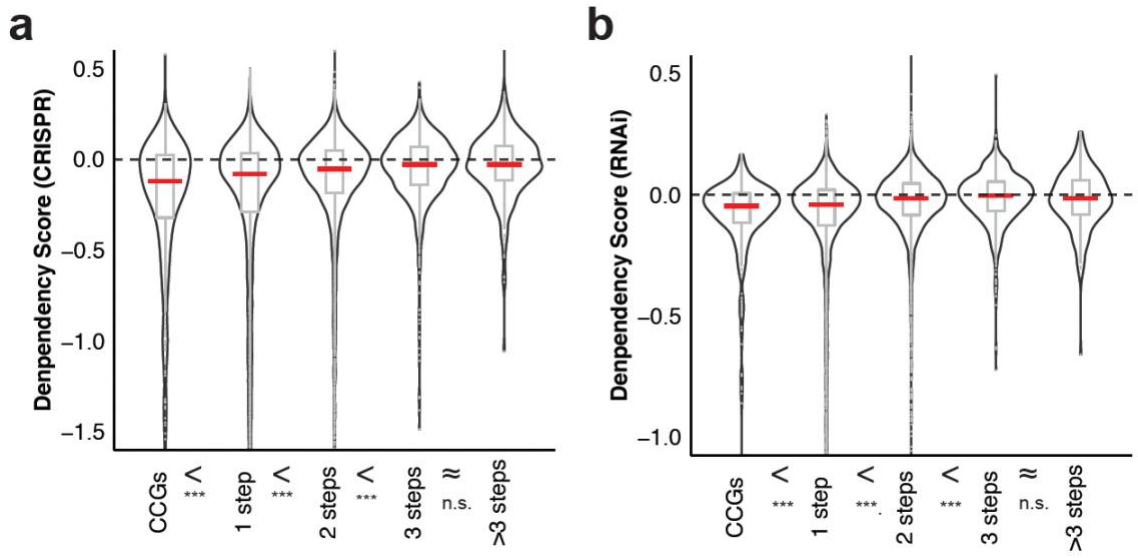


Figure 3.3. Cell viability dependence scores for cancer genes and genes in different cancer gene neighborhood categories. a *Distribution of DepMap CRISPR-based dependency scores.* **b** *Distribution of DepMap RNAi-based dependency scores. Y-axes are dependency scores—the lower the value, the more important the gene is for cell viability. One-sided Mann–Whitney U test (values of closer neighborhood genes are greater than that of all the genes in the remoter steps) p values are symbolized by ***, **, and *, corresponding to p < 0.0001, p < 0.001, and p < 0.01, respectively, and n.s. abbreviating not significant. Red bars correspond to the median of the distributions. CCGs: core cancer genes.*

3.3.3. Positive selection of somatic mutations correlates with distance to CCGs

While results based on the DepMap sheds the light on an important aspect of gene cancer relevance (i.e. loss of function), gain of function, altered protein substrate affinity and altered expression associated with somatic mutations can also be indicative of the significance of distance to CCGs in a functional network. For this purpose, we compare the average prevalence of somatic mutations in CCGs and the four distance-based categories across 32 cancer types in The Cancer Genome Atlas (TCGA). In 21 0f the 32 types, CCGs expectedly exhibit the highest average prevalence as they in part were considered "core" cancer genes because of the functional significance of their associated somatic mutational burden. Prevalence gradually decreases, however, with the increase of the distance from CCGs in the functional network further suggesting the significance of this measure (Kendall's τ z-statistic < -2.96, FDR < 0.018, Figure 3.4a). Interestingly, across all 32 cancers, the trends of somatic mutation prevalence across CCGs and neighborhood categories were negatively correlated with cancer incidence rate, tumor mutation burden, and number of somatic mutation affected genes (Figure 3.5), where results suggest that in common cancers, a larger number of genes may contribution to transformation than in rare tumors. Further, we demonstrate that average cancer effect size of each neighborhood category tend to decrease with distance from CCGs (Figure 3.4b). An effect size in a neighborhood category is an aggregate, herein average, scaled selection coefficient of the advantage that somatic mutations in a category's genes confer on the cancer cell lineage [81]. This in turn suggests that higher proximity to CCGs is associated

with an increased positive selection of somatic mutations that might bolster tumor progression or increase cell fitness.

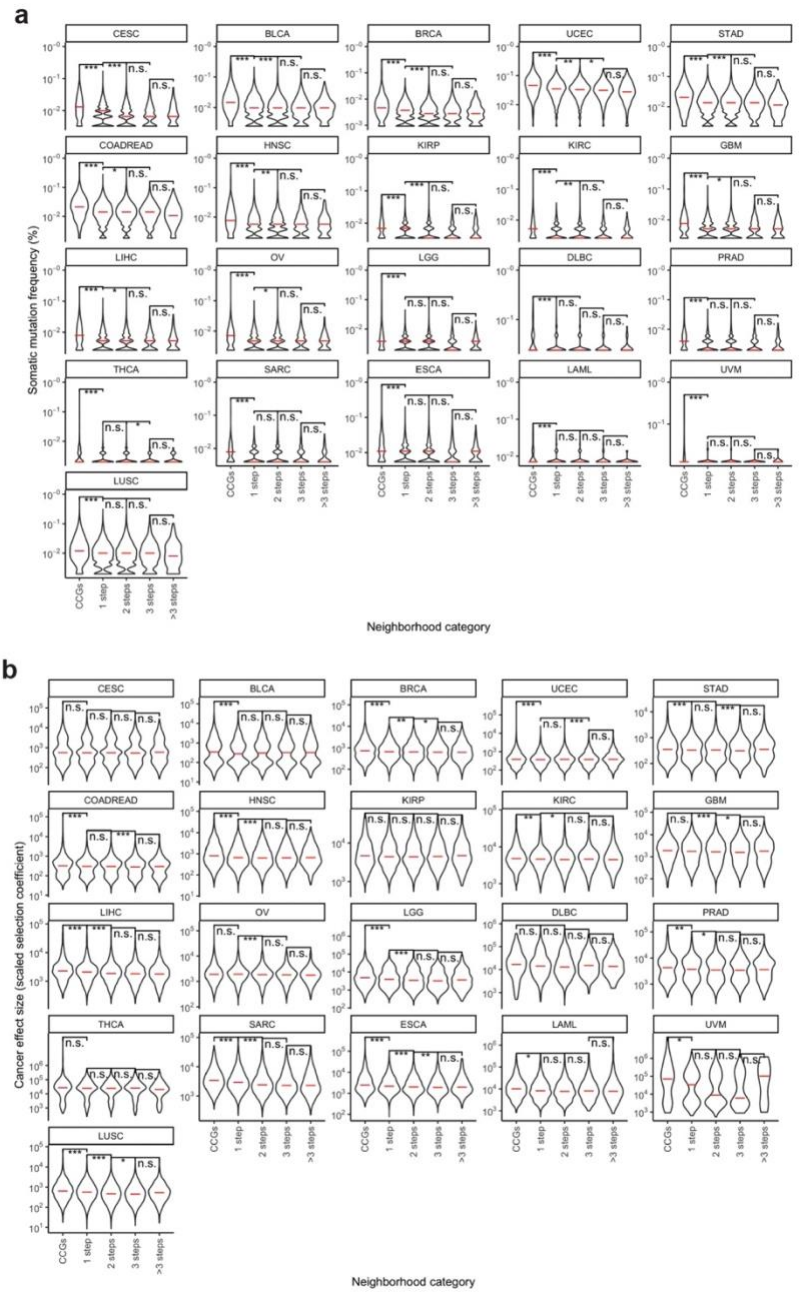


Figure 3.4. Somatic mutation frequencies of genes and cancer effect sizes of variants in genes across CCGs and 4 neighborhood categories in 21 well-sampled TCGA cancer types. a *Somatic mutation frequencies of many TCGA types show decreasing*

somatic mutation frequency for genes with increasing distance from CCGs (FDR < 0.05). **b** *Average cancer gene effect size (scaled selection coefficients) of variants in all genes of 4 neighborhood categories decrease with increasing distance from CCGs. Red bars correspond to the medians of the distributions. One-sided Mann–Whitney U test (values of closer neighborhood genes are greater than that of all the genes in the remoter steps) P values are symbolized by ***, **, and *, corresponding to p < 0.0001, p < 0.001, and p < 0.01, respectively, and n.s. abbreviating not significant. CCGs: core cancer genes.*

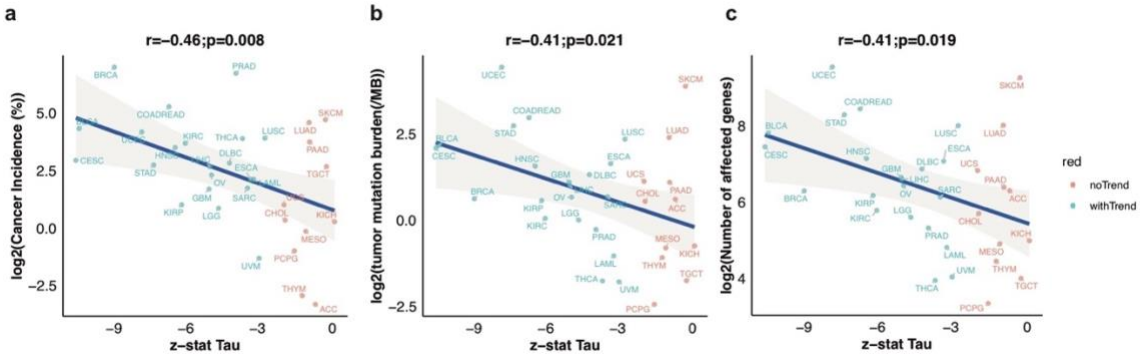


Figure 3.5. Associations between somatic mutation frequency trends and cancer incidence, tumor mutation burden (TMB), and number of mutated genes across all TCGA cancer types. a *Z-statistic τ versus cancer incidence rate.* **b** *Z-statistic τ versus TMB.* **c** *Z-statistic τ versus number of affected genes by somatic mutation. Each dot represents a cancer type and the colors indicate if it showed a significant trend for decreasing average somatic mutation frequency across CCG neighborhood categories with increasing distance from core cancer genes. The Y-axes shows average values within cancer types. The r value represents Pearson correlation coefficient. P-values are estimated based on the Pearson product-moment correlation coefficient.*

3.3.4. Negative selection of germline variants is associated with distance to CCGs

An additional measure of functional importance can be drawn from an evolutionary perspective. Namely, this measure is germline negative selection. Aberrations in genes that play roles in essential processes such as cell division, cell differentiation, cellular metabolism, and cell death can contribute to a wide range of diseases including cancer. Given the central roles these genes play in cell—and an individual's—survival, there has been a strong selection pressure to preserve their sequence and, consequently, function. Indeed, deleterious germline variants in these genes decrease fitness and tend to be rare in human populations [82]. Relatedly, the greater the functional importance, the stronger this negative selection pressure [83], and population-based whole-exome sequencing studies indicate strong negative selection pressure on deleterious germline variants in many cancer-related genes [84].

To investigate the possible association between the negative selection pressure in the germline and distance from CCGs in the functional network, we obtained coefficients of negative selection of heterozygous rare protein-truncating variants (S_h) for 15,998 human genes from Cassa *et al.* [83] and loss-of-function intolerance (pLI) scores for 18,225 genes from Lek *et al.* [85]. The higher the S_h and pLI score, the greater the selection pressure against protein-truncating germline variants of a given gene. We then observed a pattern that confidently supports our hypothesis on proximity-based cancer relevance of genes: the average S_h and pLI scores of each neighborhood category decreases significantly as

distance from CCGs grows (Kendall's τ z-statistic = -27.7 S_h and -29.4 for pLI, $p < 10^{-5}$; Figure 3.6a and 3.6b, respectively).

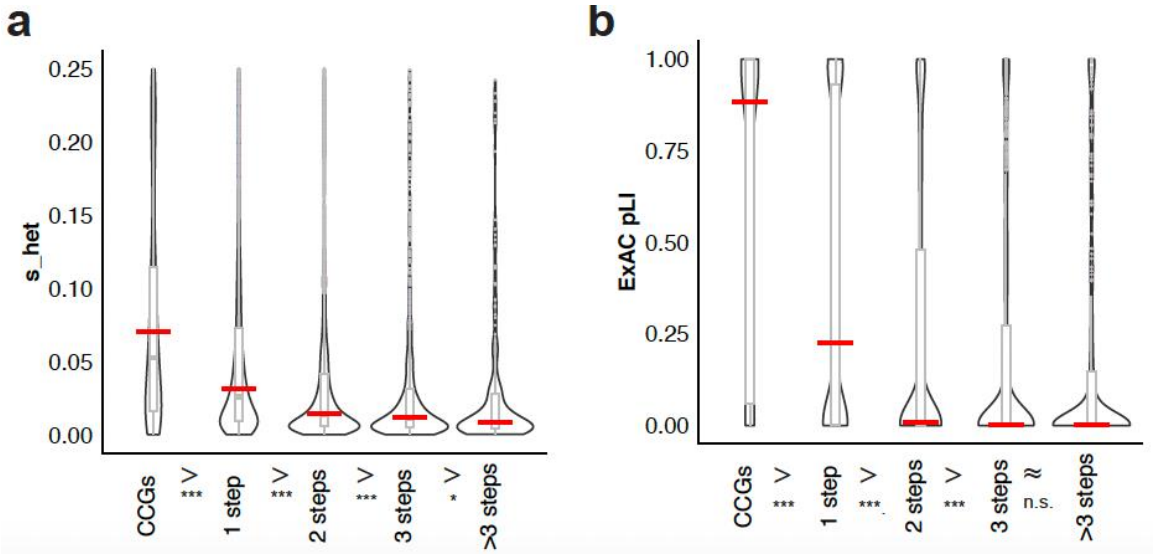


Figure 3.6. Germline selection pressure on genes in different cancer-gene neighborhood categories. a *Selection pressure against protein-truncating variants (PTV): the lower the S_h score, the more tolerant the gene is for a germline PTV.* **b** *Loss-of-function variant intolerance (pLI): the lower the pLI score, the more tolerant the genes is for germline loss of function variants. One-sided Mann–Whitney U test (values of closer neighborhood genes are greater than that of all the genes in the remoter steps) p values are symbolized by ***, **, and * corresponding to $p < 0.0001$, $p < 0.001$, and $p < 0.01$, respectively, and n.s. abbreviating not significant. Red bars correspond to the median of the distributions. CCGs: core cancer genes.*

3.4. Discussion

An increasing body of evidence suggests that a relatively small subset of genes, often described as "drivers," are not enough to explain the development and progression of all tumors [20, 86]. Instead, dysfunctions in a wide array of genes seem to be required for the transformation of a normal cell to a cancer cell, at least in a subset of samples. We comprehensively investigate a spectrum of cancer relevance of most human genes by aggregating a compendium of biomedical datasets ranging from multi-channel functional networks to mutational and evolutionary signals. We hypothesize that the level of proximity to a set of core cancer genes [87] is a reliable, continuous proxy of cancer relevance. To maximize sample size and in turn increase statistical power, we divide genes into four categories: 1-, 2-, 3- and >3-steps away from a core gene in the STRING functional network. Building on multiple experiments, we demonstrate that the aggregate relevance of each category continuously increases with proximity to core genes. Namely, the average somatic mutational burden, negative selection in the germline, impact of cancer cell survival *in vitro*, and literature prevalence of gene categories vary in association with distance to core cancer genes, confirming our hypothesis. Yet, discordant within-category results can emerge on the gene level. The functional importance of most genes, even in the malignant transformation process, is likely tissue specific [88]. However, there is no agreed upon list of tissue-specific cancer genes. Protein function may be affected through multiple mechanisms other than somatic mutations (transcriptional regulation, posttranslational modifications, protein degradation, binding partners, etc.). The importance of any mutation and protein dysfunction is also molecular context dependent, which in a cancer cell with

unstable genome opens opportunities for a large number of potential systems level combinatorial abnormalities [88, 89].

3.5. Conclusion

Our results supplement the emerging trend of transcending the driver-passenger dichotomy of classifying genes in human cancers. We suggest a discretized spectrum of cancer relevance based on aggregate subsets of genes. The ongoing growth of data generation expected to yield hundreds of thousands of whole genomes will likely help to enhance the resolution of this relevance spectrum and identify the potential role each gene might be playing in malignant transformation.

3.6. Methods

3.6.1. Data Sources and Preparation

Protein-protein interactions: STRING (v11.0) is a comprehensive database of protein-protein associations using data from genomic context, high-throughput experiments, conserved co-expression and experimental results as well as text mining of the scientific literature [67]. The data is available through https://string-db.org/.

Cancer Dependency Map (DepMap) data:. DepMap project provides a gene dependency score for the majority of known human genes that represents the effect of gene silencing on cancer cell viability [43]. The data are available at https://depmap.org/.

Somatic mutation data: Somatic mutations of 32 cancer types of 10,208 cancers in TCGA were obtained from the Multi-Center Mutation Calling in Multiple Cancers (MC3) dataset [66] that is available at https://gdc.cancer.gov/about-data/publications/mc3-2017.

S_h and pLI scores: The S_h coefficients were derived from analyses of exome sequence data from 60,706 individuals and measure genome-wide estimates of selection against germline heterozygous protein-truncating variants of a gene using Bayesian estimates [83]. The coefficients are available at http://genetics.bwh.harvard.edu/genescores/. The probability of being loss-of-function (LoF) intolerant (pLI) in the germline score was derived from whole exome sequence data of 60,706 individuals generated as part of the Exome-Aggregation Consortium [85] and is available at https://gnomad.broadinstitute.org/.

CancerMine: A text-mining based, regularly updated database of cancer driver genes, oncogenes and tumor suppressors in different types of cancer [72]. Data are available at http://bionlp.bcgsc.ca/cancermine.

3.6.2. Defining cancer gene neighbors

The shortest distance from one protein to the other in the STRING (v11.0) network was calculated by Dijkstra's algorithm [90] using the NetworkX v1.11 Python package https://networkx.github.io/documentation/networkx-1.11/. We visualized the connection of CCGs to neighboring genes using Cytoscape (v3.7.2) with default settings [63]. To plot the results, we manually set the size of gene nodes to 50, 40, 30, 20 and 10 for CCGs, 1-step, 2-step, 3-steps and >3-step removed genes, respectively.

3.6.3. Somatic mutation analysis

For somatic mutation frequency, we only considered the 2,257,845 nonsynonymous mutations that comprised missense, non-sense, frameshifting, in-frame shifting, or splice-site altering single-nucleotide changes or indels in 32 cancer types. Somatic mutation frequency at gene level was defined as the percent of cases that carried at least one nonsynonymous mutation of the gene within a cancer type. Gene level mutation frequencies were averaged over each gene neighborhood class. Tumor mutation burden (TMB) was calculated for each cancer as the number of somatic mutations, including both nonsynonymous and synonymous, per sequenced megabase. For each cancer type, we averaged TMB across all patients. The total number of genes which were affected by at least one nonsynonymous mutation was also calculated for each cancer and was averaged across all the patients with a given cancer type.

Cancer effect size is the scaled selection coefficient of the mutation, conveying the degree to which the mutation enhances the survival or reproduction of the mutant lineage. Cancer

effect sizes were calculated with cancereffectsizeR 0.1.1.9006 (https://github.com/Townsend-Lab-Yale/cancereffectsizeR) as in Cannataro *et al.* [81] except that the likelihood of the scaled selection coefficient was maximized based on tumor-specific mutation rates, and only COSMIC v3 signatures consistent with Alexandrov *et al*. [91] were used for each tumor type. We calculated average cancer effect size for all somatic mutations in TCGA cancer types effecting all genes in a given neighborhood category.

3.6.4. Statistical analysis

The connectivity score, dependency score, somatic mutation frequency, cancer effect size, S_h and pLI score were compared between different groups of genes (e.g. CCG, 1 step, 2 step, etc..) using the one-sided Mann–Whitney U test with the "base" package of the R-project (www.R-project.org/).

We estimated the statistical significance of the trend of the average dependency score, somatic mutation frequency, S_h and pLI score across the different gene groups (e.g. CCG, 1-step, 2-step, etc.) using Jonckheere Terpstra (JT) trend analysis [92]. *P*-values were calculated using the "JonckheereTerpstraTest" function of "DescTools" packages (31) in the R-project. The number of permutations for the reference distribution was set as 100,000. Z statistic of Kendall's tau (τ) coefficient was estimated to show the increasing (positive value) burden, and the number of affected genes. Pearson correlation coefficient and p-values were also calculated using the "cor.test" function of the "stats" R package.

We separated the 32 TCGA cancer types into two groups: (i) “withTrend” indicating statistically significant decreasing trend of somatic mutation frequency, and (ii) “noTrend” corresponding to cancers with no decreasing trend. We assigned a cancer type to the “withTrend” group if the Jonckheere Terpstra FDR was less than 0.05, otherwise, a cancer type was assigned to the “noTrend” group.

We used Kendall's τ to quantify association between cancer incidence rate, tumor mutation burden, and the number of affected genes. Pearson correlation coefficient and P-value were also calculated using the “cor.test” function of the “stats” package.

Chapter 4

Weight-based neural network interpretability using activation tuning and personalized products

This chapter is based on the work described in Mohsen *et al*. [93].

4.1. Abstract

We introduce approaches to simplifying neural networks and enhancing their interpretability using activation-based neuron tuning and personalized weight matrix products. Inspired by the evolutionary principle of the survival of the fittest, we gradually remove neurons with little to no learning activity during training and hypothesize that their absence renders opaque models more interpretable. Experimental results pertaining to cancer and diabetes hospital readmission appear to support our hypothesis and generate biomedically salient results. Our approaches also allow for interpretations at the sample level, a feature of high importance in multiple fields including personalized medicine.

4.2. Introduction

Wide applicability of neural network models is contingent on our understanding of the underlying dynamics leading to their exemplary performance. In fields like biomedicine, interpretability is a necessary bridge to establish trust between AI and medical scientists

[94]. Artificial neural networks (ANNs) were arguably first used in biomedicine in 2007 [95], but the popularity of these models started to rise in 2014 starting with applications to understand the human RNA splicing code [96]. Due to the recent popularity of deep learning and the availability of large biomedical datasets, neural network interpretation approaches in biomedicine have been limited. Current methods, for instance, use trained networks to estimate the effects of genetic variants by comparing model outputs when provided with reference and mutated sequences as inputs [97], or to suggest new functional sequence regions using learned weights [98-100]. In cancer genomics, Yousefi *et al.* [101] developed a method called Risk Propagation to estimate the contribution of input features to predicted patient survival. Inspired by the backpropagation algorithm that calculates gradients used to update weights during training, Risk Propagation uses the chain rule to calculate the partial derivative of predicted risk (rather than error) with respect to each input feature. Warrell *et al.* [102] presented a greedy approach that traverses paths in trained networks to identify related subsets of inputs, with applications in functional genomics. More recently, Keunzi *et al.* developed hierarchical interpretability approaches to simulate drug response to cancer therapies [103].

Our goal in this chapter is two-fold. First, we aim to further understand how neural networks learn with a focus on learned weights, rather than gradients. Second, we aim to generate biomedical hypotheses at the sample level by scrutinizing the high number of parameters learned during training—i.e. learned weight matrices. To these ends, we introduce two complementary approaches: Activation-based Neuron Tuning (ANT) to discard neurons considered inactive during training, and Personalized Weight Product

(PWP) to interpret the resulting network using products of data and weight paths. While each of ANT and PWP can be deployed as a standalone approach that serve different yet related tasks, we connect them through a bio-inspired hypothesis on the learning process of neural networks that renders ANT a favorable precursor to PWP.

4.3. Activation-based Neuron Tuning

4.3.1. Hypothesis

Activation-based neural tuning is inspired from biological phenomena where only a subset of entities participating in a process endure or contribute to the final outcome. Whether it is a key cellular pathway whose disruption leads to cancer after inactivating only a few genes, or the brain responding to external stimuli using a small fraction of its neurons, prioritizing biomarkers according to their contribution intensity is a recurring theme in biology. Applicability of this hypothesis on neural network training is centered around two ideas inherent to the training process. The first pertains to the stochasticity of training: networks with different weight initializations can yield different learned weights but comparable overall predictive performance, suggesting that neural networks can take multiple "learning routes" to identify patterns in data. The second relates to the comparable predictive performance of networks with different architectures. In supervised learning tasks, network size often reaches a saturation limit where adding neurons does not improve performance.

Our tuning approach trims a network during training to (1) keep only enough neurons to learn target patterns and (2) restrict the "learning route" to untrimmed neurons considered significant by the virtue of receiving concentrated learning flow during training. We hypothesize that discarded neurons could be inducing noise on the learning process. By the end of training, remaining neurons are expected to resemble the "learning bottleneck" of the network, i.e. a small set of neurons that suffice for effective and less noisy learning. This perspective resembles an indirect relation to the "information bottleneck" [104], and from an evolutionary biology angle, it can be seen as a model of Darwin's survival of the fittest. The measure of fitness is based on the level of a neuron's engagement during training measured through an activation function-specific proxy described below.

4.3.2. Neuron Selection Criteria

For weight updates to effectively navigate the loss function's (L) error surface, gradient magnitudes must take values higher than 0 or ε (i.e. small values that often pertain to the saturation problem). To turn off neurons during training, our ANT selection criteria measure the properties of neurons' input distributions (Z's) to rank them according to the magnitude of weight updates. Neurons with inputs concentrated around activation function-specific favorable intervals are prioritized, while others distant from a concentrated target distribution (Φ_t) are permanently turned off. The number of neurons to be removed per tuning step, n, and the number of epochs at which neurons are regularly turned off, k (leading to $m = \lfloor$ Total number of epochs / $k \rfloor$ tuning steps) are pre-defined

parameters that indicate the total number of neurons eliminated from each layer by the end of training (Figure 4.1).

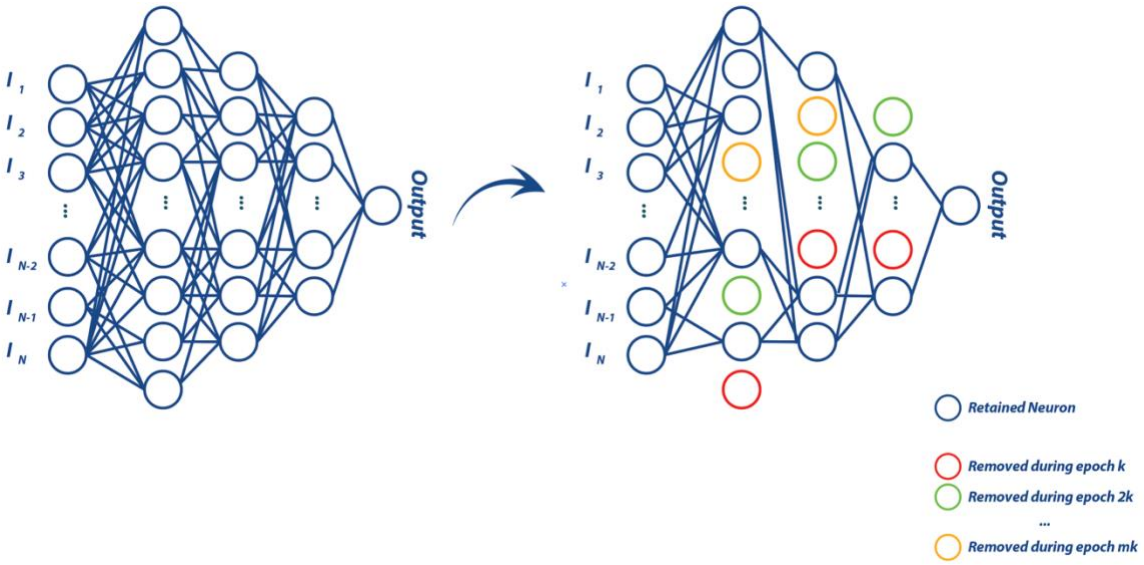


Figure 4.1. *The ANT tuning scheme.*

4.3.3. Calculus Interpretation

In calculus terms, we define neuron activity in terms of its gradients' updates during optimization. By virtue of the chain rule used to calculate gradient values during each backward pass, overall gradients are affected by derivatives of neuron activation functions with respect to their inputs (i.e. middle term of equation (4.1)).

$$\frac{\partial L}{\partial w_{ij}^{l-1}} = \frac{\partial z_j^l}{\partial w_{ij}^{l-1}} \frac{\partial a_j^l}{\partial z_j^l} \frac{\partial L}{\partial a_j^l} \tag{4.1}$$

where L is the loss function, l and $l - 1$ are subsequent layers, i is the source neuron index in layer $l - 1$, j is the destination neuron index in layer l, a^l is the activation function in l, $a^l_j = a^l(z^l_j)$, $z^l_j = w^{l-1}_j a^{l-1} + b_l$, w^{l-1}_j is the weight vector incoming from $l - 1$ to neuron j in l, and b_l is the bias term of layer l. We explain next how we accordingly select neurons to turn off based on input distributions to activation functions. We focus on the cases of ReLU and sigmoid functions and describe a rationale that generalizes to other functions for neuron selection.

Derivative of the sigmoid function $\sigma'(x) \in]0, 0.25]$, with its highest values at $x \in [3; +3]$ (Figure 4.2A). To encourage active updates in a layer's neurons, we select the target distribution for sigmoid to be $\Phi_t^{Sigm.} \sim N(0, 1.5)$, a distribution with high peaked-ness centered around $\mu = 0$ and 2 [-3, +3] (Figure 4.2C). Neurons receiving input distributions (Z) close to $\Phi_t^{Sigm.}$ encourage non-zero and relatively large sigmoid gradient values ($\gg \varepsilon$) resulting active overall neuron gradient updates during backpropagation. In contrast, input distributions furthest from $\Phi_t^{Sigm.}$ lead to recurring 0 and ε-like gradients impeding progress during optimization. ANT uses Kullback–Leibler divergence to measure the difference between the histograms of both distributions Φ and Z over training data points.

A similar rationale is adopted to select target for ReLU, where Φ_t^{ReLU} encourages positive, larger derivatives and discourages 0-valued ones. Generally, the gradient of ReLU is either 0 or 1 depending on the input value passed during the forward pass: positive values lead to a derivative of 1, while negative values lead to a derivative of 0 (Figure 4.2B). We select Φ_t^{ReLU} to be an "inverted power law" distribution representing a considerably higher

density shifted towards positive values (Figure 4.2D). This same goal can drive the selection of target distributions that favor high activity regions of other activation functions.

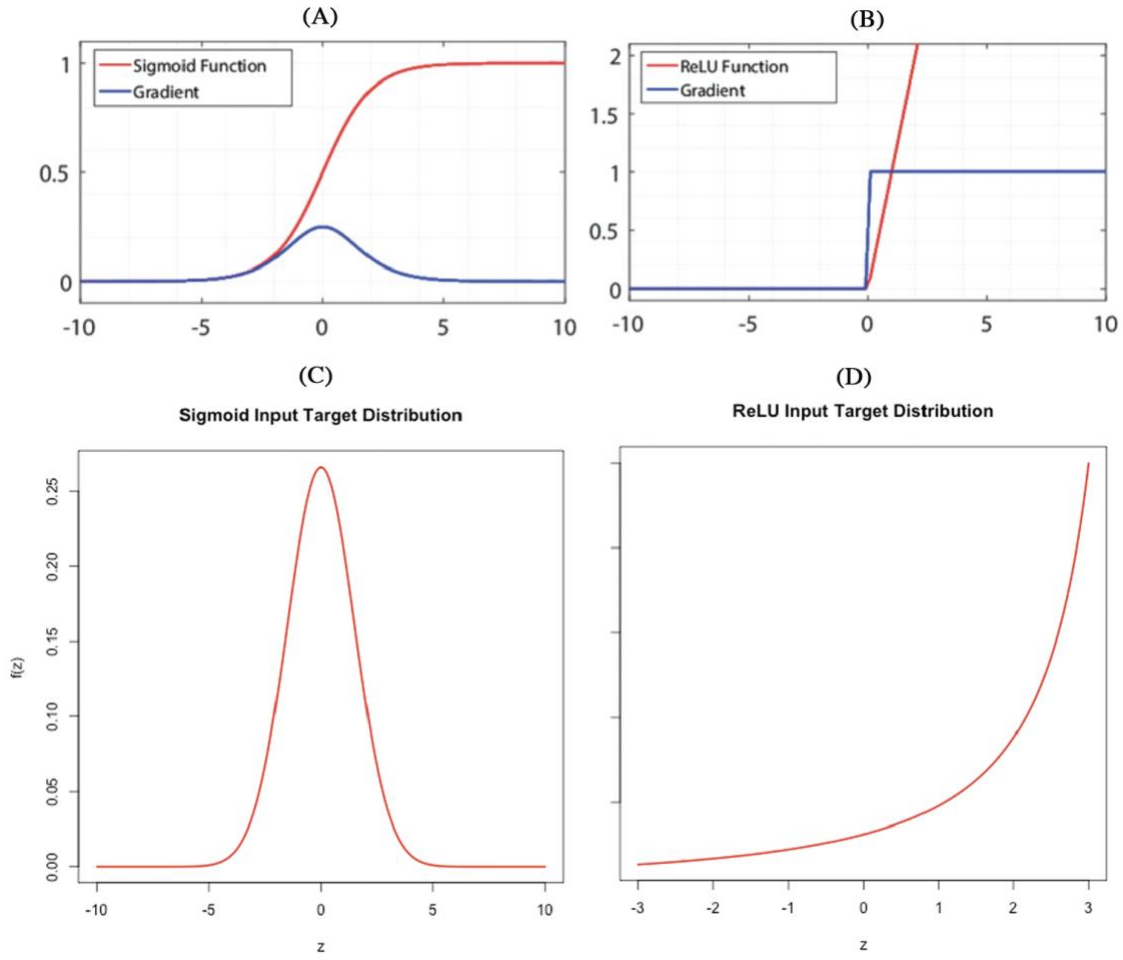


Figure 4.2. Function and gradient curves (A, B) and ANT target distributions (C, D) of sigmoid and ReLU functions, respectively. Subfigures (A) and (B) are adopted from part of Figure 3-5 in [105].

4.3.4. Algorithm

We lay out the steps of ANT in Algorithm 1.

Input: Training data D, Initialized neural network N, Target layers T, tuning step k number of tuned neurons n

Output: Tuned neural network N'

for epoch ← 1 to E epochs **do**

    $SGD(D; N; W)$

    if epoch % k = 0 **then**

        for layer $l \in 2\ T$ **do**

            $S^l \leftarrow D_{KL}(Z_i^l \,||\, \Phi_t)\ \forall$ neuron $\in l$

            $N^l_{tuned} \leftarrow N^l_{tuned} \cup \underset{1..n}{\text{argmax}}\ S^l$

            Remove N^l_{tuned} from the network

        end for

    end if

end for

Algorithm 1. Activation-based Neuron Tuning (ANT) algorithm.

4.4. Personalized Weight Product

The idea of leveraging weight matrix products to interpret trained neural networks was arguably first introduced long before deep learning garnered its recent popularity, namely

with Garson's algorithm [106]. Recent cancer genomics research highlighted the high heterogeneity of cancer subtypes, emphasizing the need for patient- or subgroup-level treatments, a trend that falls under a set of practices that became known as "personalized medicine." [1, 75]. Driven by this and other recent trends in biomedicine, we introduce PWP with an ability to estimate the contribution of input features to prediction on whole set, subset, or individual sample levels. We also leverage biomedical domain knowledge to incorporate the signs of the weights during matrix multiplication to mimic the important directionality of interactions between genes or clinical phenotypes pertaining to disease. Hence, unlike Garson's algorithm that relies on the absolute values of every weight matrix, we use absolute values only in the final step after signed matrices take part in iterative multiplication. PWP is formalized in equation (4.2) below.

$$PWP_I = X. \mid W_1 . W_2 \ldots W_L \mid, \tag{4.2}$$

where I is the set of inputs, L is the number of layers, and X is the dataset based on which input contributions are to be calculated.

5. Results

We evaluate ANT and PWP on two biomedical datasets to predict drug response in acute myeloid leukemia (AML) [107] and hospital readmission of diabetes patients [108, 109], and on the MNIST dataset. The first set of experiments investigates the possibility of

deteriorating performance caused by neuron removal by ANT, while the second studies the performance of PWP as a standalone approach and combined with ANT. Reported results are aggregated over 10 reproducible runs.

5.1. One-Layer and Two-Layer Tuning

While the predictive performance of a trained network is not the central goal of ANT, this performance must not be sacrificed in exchange of higher levels of interpretability. To this end, we compare ANT-tuned networks with baseline models (i.e. without neuron tuning). Each baseline model constitutes 3 layers with its hyperparameters selected using `hyperopt` [110]. We note that a slightly higher performance has been achieved on the MNIST dataset using CNNs, but we focus on the fully-connected neural nets, the target network type of our current approaches.

Results from all predictive tasks demonstrate that ANT maintains high AUC, accuracy, and precision across all three datasets while turning off up to 70% of the first hidden layer's neurons (Figure 4.3A). Similar results are obtained when tuning two hidden layers while shrinking the model by at least 50% (Figure 4.3B).

5.2. Biomedical Interpretation: Cancer Genomics and Clinical Diabetes

In the first task, we perform biological enrichment analysis on the top 100 genes prioritized by PWP vs weight-based Garson's algorithm out of >26,000 input features encompassing

gene expression and genomic variation profiles. Enrichment results returned by the DAVID web service [111] for the top 100 genes prioritized by each approach demonstrate

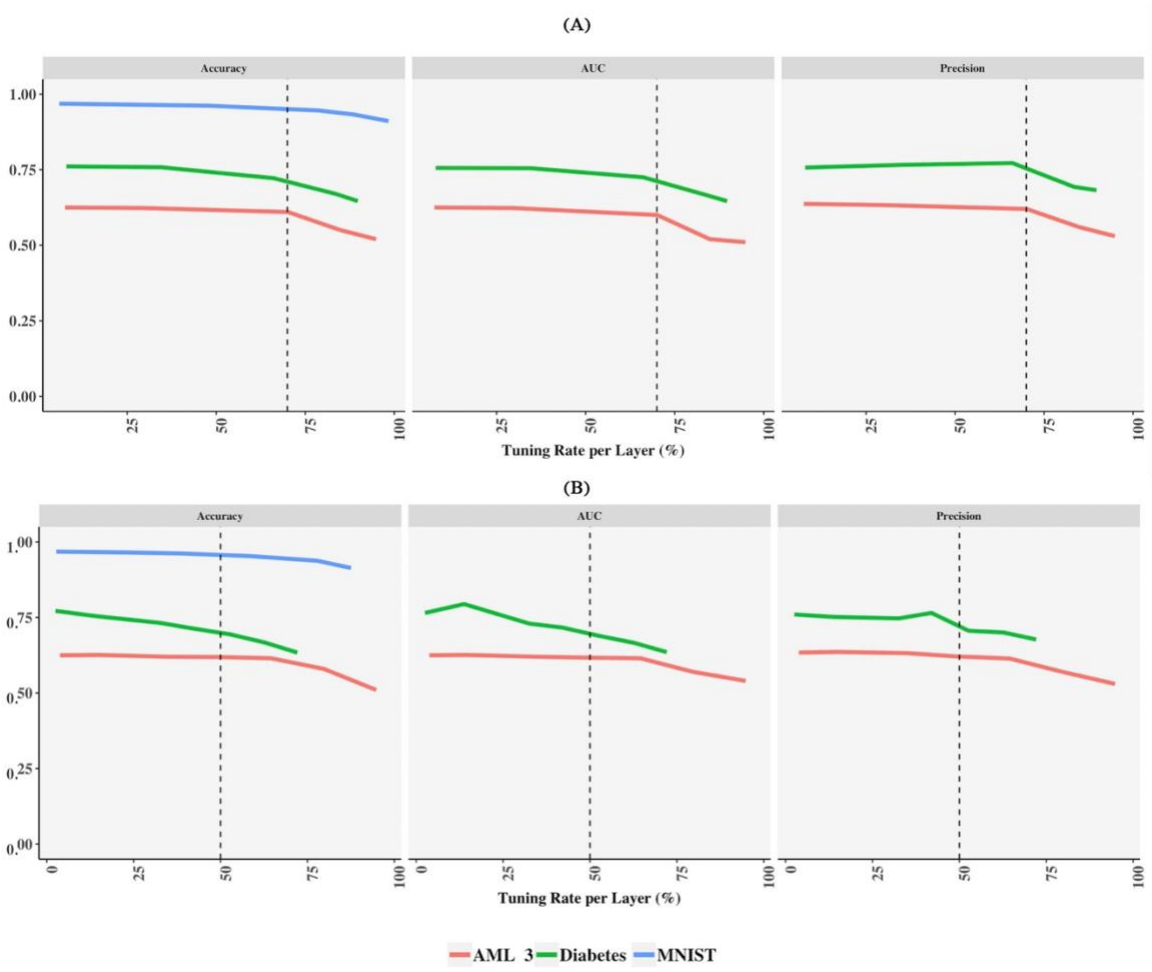


Figure 4.3. *Predictive performance of neural networks across ANT tuning rates applied on (A) one or (B) two hidden layers.*

that PWP identifies significantly more biological entities associated with AML than Garson's: "AML" term count, number of associated publications, and statistically significant (Benjamini p-value < 0.05) chart and clustering annotation records. More interestingly, PWP applied on ANT-tuned models (labeled PWP-ANT) achieves better performance than PWP alone. Annotation of the PWP-ANT prioritized genes, the majority

of which are long tail ones, is also the only one to include AML as a directly reported GAD disease (Figure 4.4A). A similar pattern is observed in the diabetes patient readmission task. PWP-ANT's top 5 features included 3 of the gold standard clinical features curated based on expansive literature review, compared to two features using PWP alone and only one feature by Garson's algorithm (Figure 4.4B). These results highlight the significance of using signed network matrices to capture interactions between features. We also note that PWP variants achieved significantly better results compared to randomly selected genes as another baseline in the AML task.

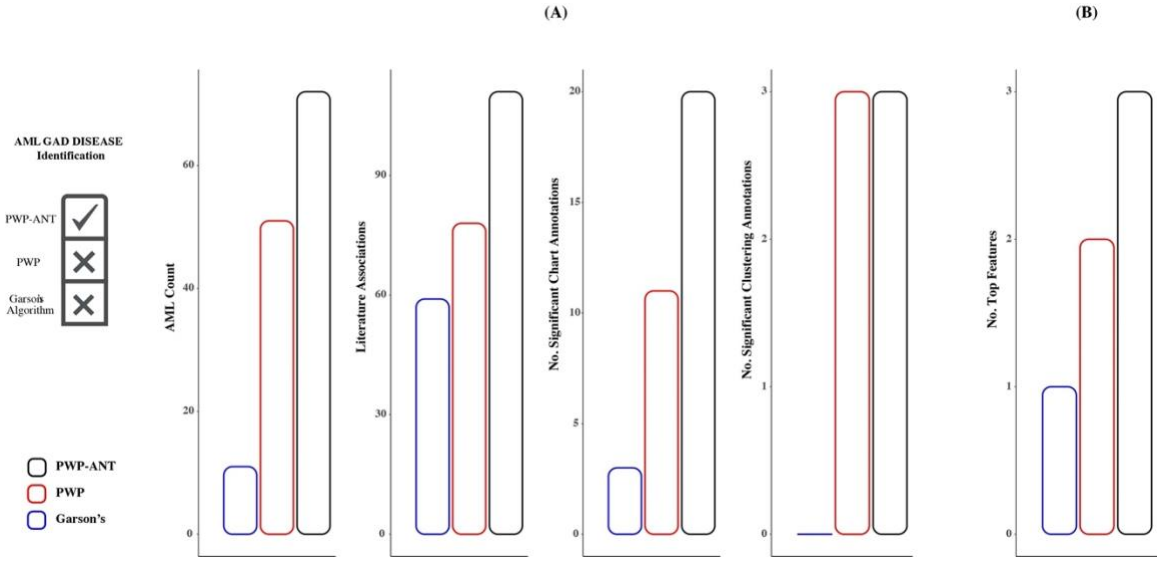


Figure 4.4. AML and Diabetes Results. A *PWP-ANT's top gene list uncovers more biomedical annotations pertaining to AML than that of PWP alone or Garson's algorithm.* **B** *PWP-ANT prioritizes more clinically important features than both approaches*.

5.3. Personalized Interpretations

A severe limitation of the Garson's algorithm's weight matrix product approach is its estimation of a singular value for each feature's contribution to the output. The data-driven nature of PWP allows it to identify prioritized features on a sample- or subset-levels of interest. To examine the potential of PWP-ANT as an attribution method, we run PWP-ANT on MNIST with three input datasets: (i) all images of all digits, (ii) all images of each digit separately, and (iii) only two images of the same digit. Prioritized pixels varied depending on the subset being considered. On dataset (i), PWP-ANT highlights pixels pertaining to specific features of multiple digits included the set. Interestingly, these pixels are located in discriminative locations that allow for the distinction between similar-looking digits such as the edges in the center of 3 and 8 or 0 and 9 (white rectangles of Figure 4.5A). On subset (ii), prioritized pixels become more specific to the target digit. Each row in Figure 4.5B highlights the same pixels prioritized to cover discriminative features of the target digit (0, 1 or 7 shown as examples). Selected pixels might also demonstrate locations where the digit of interest uniquely has no pixels. For instance, being the only digit without a single pixel in the center, these pixels were highlighted for digit 0 (center top image of 4.5B, orange rectangle). Prioritized pixels become even more specific for subset (iii) as shown in 4.5C. When only two images of the same digit are provided to the method, PWP-ANT uncovers the specific edge pixels of these particular images. We note that no retraining of any baseline or ANT-tuned network was required in these or other experiments, and the specificity of prioritized features is based solely on data provided to PWP as described in equation (2) with minimal computational overhead.

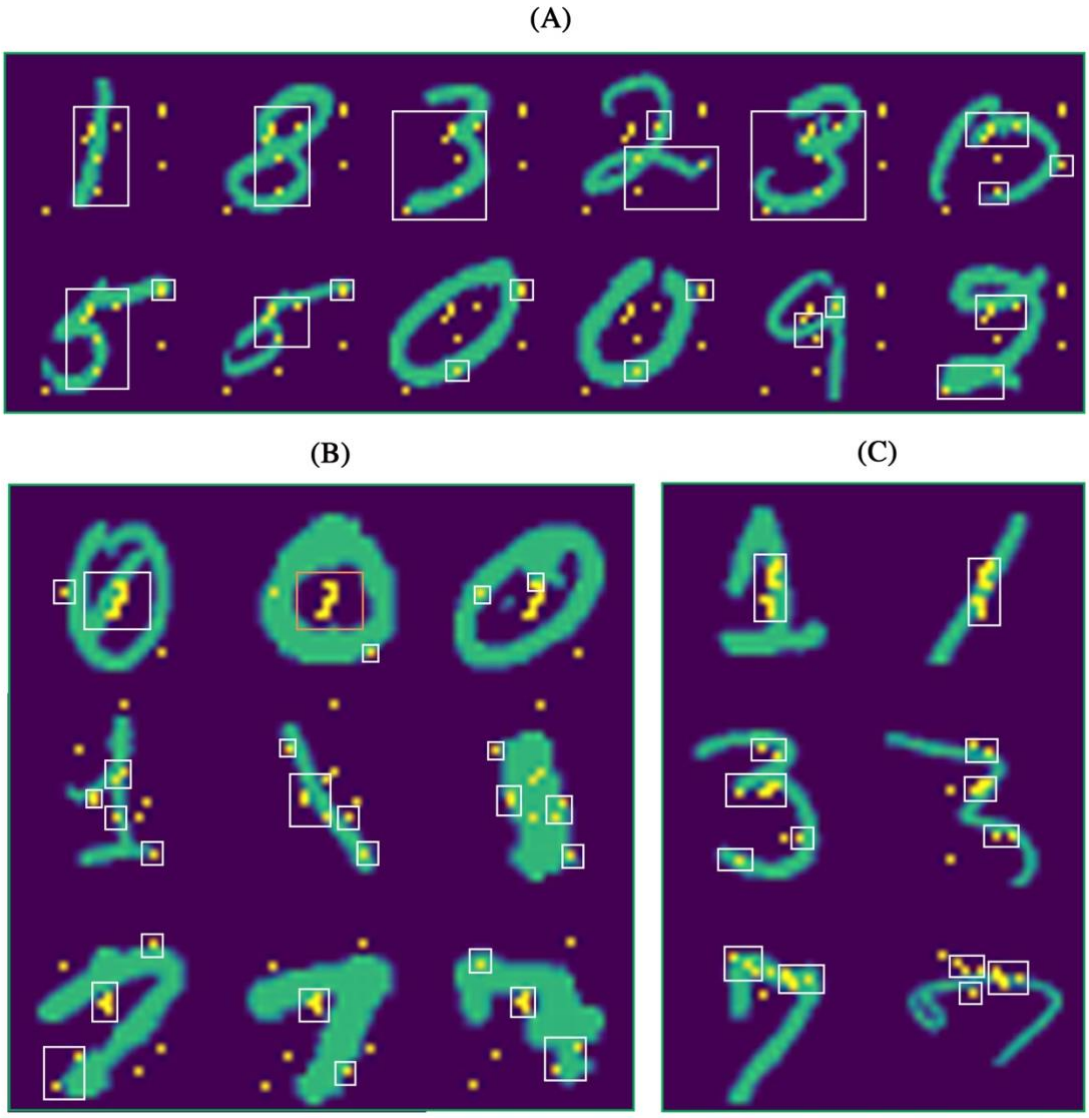


Figure 4. Representative MNIST results. *Highlighted pixels prioritized by PWP-ANT capture the important discriminative features used to distinguish digits in the input in each of three scenarios:* **A** *all validation data including all digits,* **B** *all data of for a single digit, and* **C** *two data points of the same digit.*

6. Potential Future Directions

We introduce efficient approaches to simplifying neural networks and enhancing our understanding of learned parameter values. Driven by biomedical domain knowledge, our results highlight the importance of learned weight signs and the efficacy of adopting a parsimonious perspective in training yielding smaller networks. While we demonstrate the improvement our method introduces to its closest counterpart (i.e. weight-based Garson's algorithm), experiments can be expanded in relation to related work by: (i) comparing ANT's tuning to other methods including the lottery ticket theory [112] and the work in [113], (ii) extending PWP to detect feature interactions in line with weight-based Garson's algorithm-inspired work in [114], or (iii) elaborating on the attributive side of PWP in comparison with other attribution methods that have made significant recent advances with potential opportunities for additional improvement (e.g. [115-117]).

Chapter 5

Compression-based neural network interpretability with applications in functional genomics

This chapter is based on the work described in Warrell, Mohsen and Gerstein [118].

5.1. Abstract

We introduce complimentary approaches that allow for an efficient compression of a trained neural network and the extraction of domain-related patterns learned by the network. The compression approach, named Rank Projection Trees (RPTs), selects a sparse set of predicative network paths forming a DAG based on a branching factor and the properties of learned parameter. Using data from multiple functional genomics tasks, we demonstrate the efficacy of RPTs and provide a measure to score the interpretability potential of resulting compressed networks.

5.2. Introduction

A multitude of definitions of neural network interpretability has been proposed in the literature [119, 120]. Should researchers focus on gradient updates or values of learned parameters after training? Single data point or an entire dataset underlying a predictive task? Design customized loss functions inspired by the underlying question? When it comes to the trained network, should they consider the whole network or, instead, search

for a subset of representative paths that can act as hypothesis generators in the task's domain of interest? To the last question, which addresses what we describe herein as network compression, a number of approaches have been suggested to respond to limited memory or real time execution requirements [121], to provide bounds on the generalization error of a trained model [122, 123], or to establish a link between generalization and regularization [112, 124]. Yet, the relationship between compression and interpretability has not been thoroughly explored, which is the gap we aim to address in the chapter.

While a number of existing approaches suggest approximating a complex model using a predefined target model (e.g. linear models) [125-127], they leave open questions to the user regarding the degree and type of compression. We propose complimentary approaches capable of identifying a considerably compressed, domain-relevant structure in the network's architecture that are not restricted to a target model class. We frame these approaches together as an *interpretability scheme* (Figure 5.1) that a) performs efficient post-hoc network compression, b) interprets the compressed model to extract domain-relevant information, and c) provides (PAC-Bayes based) score estimates for interpretable model selection.

5.3. Rank Projection Trees

We first introduce a compression approach that identifies an interpretable tree grouping in the learned network. While other methods such as the Shapley features [126] can select groupings of interest, they do not explicitly derive these from the network structure, which

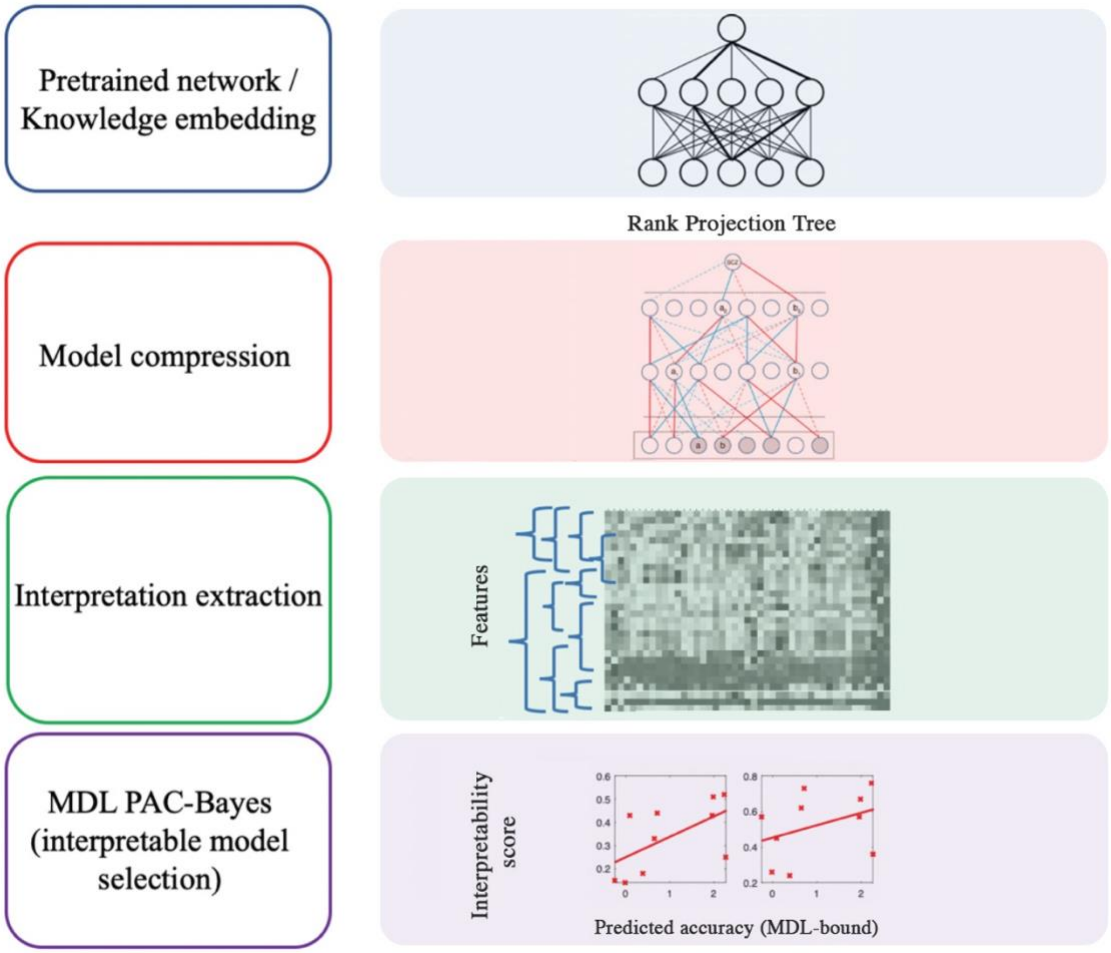


Figure 5.1. *The general framework of the RPT interpretability scheme.*

might lead to more opaque interpretations. The resulting trees of our approach, referred to as Rank Projection Trees (RPTs), are instead meant to identify nested groups of nodes which have potential joint interactions, a common scenario in functional genomics where sets of biomarkers interact to determine a trait's phenotype. The RPT framework provides a multiscale output-to-input interpretation of the learned neural network and is agnostic to the ranking function used to select nodes while building the tree. This in turn provides more flexibility allowing for many node-based scoring functions, including those derived from

interpretability approaches that focus on scoring individual nodes rather than paths (e.g. [116, 128, 129]), to be used in tree construction.

Let N be a neural network with L+1 layers, where $n_{l,i}$ is the i^{th} neuron of layer l such that $0 \leq l \leq L$, and let $W_{l1,l2}$ and β_l be the weight matrix between layers l_1 and $l_2 = l_1$ -1 and the bias vector at layer l, respectively. A *rank projection tree* (RPT, see Figure 5.2) over a given network is fully determined by specifying (a) a half branching factor $B < \frac{N_l}{2}, \forall\, l$, and (b) a ranking function $r_{i,l,m}(j)$, where $l < m \leq L$ are layer indices, i and j are node indices on layers l and m respectively, and the function returns an integer specifying the position of node j in an ordering of the nodes at layer m according to their 'score' with respect to node i and layer l. Semantically, we expect that increased activation of node j towards the top of the ranking will lead to increased activation of node i, while increased activation of nodes towards the bottom will lead to decreased activation of i; hence, any score function of the kind described above (such as the gradient) may be used.

The nodes of the rank projection tree T are lists of "branching indices" of the form [ ], $[b_1], [b_1, b_2], \ldots [b_1, \ldots, b_L]$, where $b_l \in \{1, \ldots, B\} \cup \{-1, \ldots, -B\}, \forall\, l$. The node [ ] is the root of the tree, and the parent function is defined as $Pa([b_1, \ldots, b_{l-1}, b_l]) = [b_1, \ldots, b_{l-1}]$. A node t of T, where t is a list of length l, is then associated with a node in layer l of the neural network via a function ϕ defined recursively as follows:

$$\phi : T \rightarrow N$$

$$\phi([\,]) \;=\; n_{0,1}$$

$$\phi(t = [b_1, \ldots b_l]) \;=\; r^{-1}_{Pa(t),l-1,l}(b_l), \tag{5.1}$$

where $r^{-1}_{i,l,m}(b)$ is a "quasi-inverse" of the ranking function, which returns the node $n_{m,j}$ for which $r_{i,l,m}(j) = b$ if $b > 0$, and $n_{m,j}$ for which $r_{i,l,m}(j) = N_m + b + 1$ if $b < 0$. For node t in T then, which maps to $\phi(t) = n$ at layer l, the mappings of the children of t are set by first ranking layer $l + 1$ of the neural net with respect to n, and assigning the top B and bottom B nodes of this ranking to the children of t, hence *projecting* the full ranking onto a reduced ranking across the $2B$ children.

Each node t in T may be associated with positive and negative subsets, S_t^+ and S_t^-, at a reference layer, which we take to be the input layer L. These are defined as:

$$S_t^+ \;=\; \{\phi(t' = [t, b_{l+1}, b_{l+2}, \ldots, b_L]) | b_{l+1} \cdot\; b_{l+2} \ldots \cdot\; b_L \;>\; 0\}$$

$$S_t^- \;=\; \{\phi(t' = [t, b_{l+1}, b_{l+2}, \ldots, b_L]) | b_{l+1} \cdot\; b_{l+2} \ldots \cdot\; b_L \;<\; 0\} \tag{5.2}$$

Hence, S_t^+ contains all those nodes mapped to by descendants of t at layer L along paths where the product of the branching indices below l is positive, and S_t^- is defined similarly, but where the product of the branching indices is negative. A collection of 'prioritized'

subsets at multiple levels is thus formed by applying Eq. 5.2 as t runs across T. We note that, since multiple nodes in T may map to the same node in N, sets at the same layer may overlap, including positive and negative sets associated with the same node in T. Finally, we may define a prioritization function π (or 'salience map') of the nodes at the reference layer in N, $\pi(n) = f(\phi^{-1}(n))$, where $\phi^{-1}(n) = \{t_1, t_2 \ldots\}$ is the pre-image of n under ϕ, and f may be chosen from a number of possibilities.

5.4. Compression Bounds

We next introduce a PAC-Bayes approach to interpretable model selection. This approach allows for the use of RPT selection for both enhanced comprehension of the model and providing a *minimum description length* (MDL) prior to identify its generalizable structure. For this purpose, we use the following basic form of the PAC-Bayes bound, outlined in [130]:

$$R(N) \leq_{\delta} \psi(N, X) = R(N, X) + \left(\frac{1}{\lambda}\right)\left[\mathrm{KL}(N, \pi) + \log\left(\frac{1}{\delta}\right) + \left(\frac{\lambda^2}{N_X}\right)\right] \tag{5.3}$$

where $R(N)$ is the true risk of network N, $R(N, X)$ is the empirical risk on the observed sample X, π is a prior over networks, and $\mathrm{KL}(N, \pi)$ is the KL-divergence of a delta-distribution at N with π. As introduced in [122], given an encoding scheme, π may be formulated as a minimum description length (MDL) prior:

$$\mathrm{KL}(N,\pi) \leq |\widehat{N}|_c \log 2 - \log\left(m(\left|\widehat{N}\right|_c)\right) \tag{5.4}$$

where $\widehat{N}$ is the code-word for N (which may be a lossy code), $|.|_c$ is the code length, and $m(.)$ is a prior over code lengths, which for convenience may be taken to be uniform. For RPT, codewords may be generated by taking the compressed representation, i.e. sparsified network, and subjecting it to further compression via LZW coding to generate a binary code. Equation 5.3 can thus be used directly as a generalization bound after substituting equation 5.4 for the KL term.

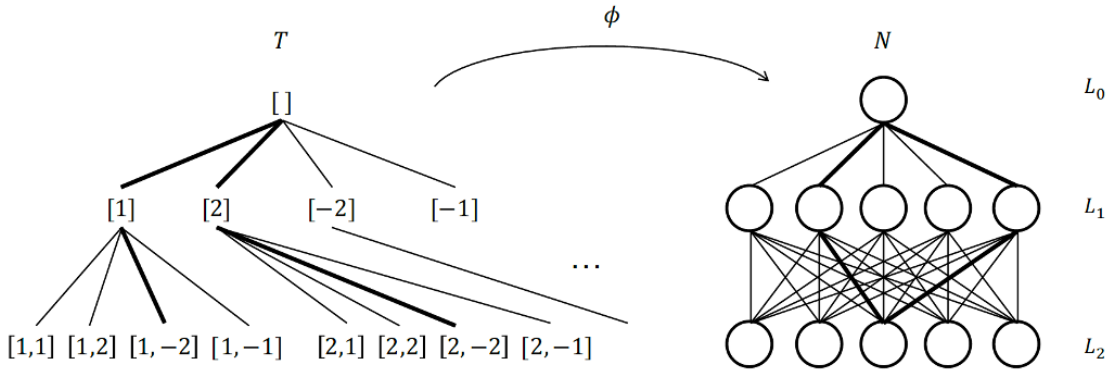


Figure 5.2. Rank projection trees. *The rank projection tree (left, $\mathcal{T}$) is mapped onto a trained neural network (right, $\mathcal{N}$) via the mapping ϕ which depends on an arbitrary ranking function r. The image of $\mathcal{T}$ under ϕ is used to prioritize inputs and sets of inputs in $\mathcal{N}$ in an output dependent fashion.*

As noted previously, we are also interested in combining compression with prior information to produce a modified MDL bound. For this purpose, we introduce the following bound:

Theorem 1 (Modified MDL Bound). Let π_{MDL} be an MDL prior and π_{dep} be a data dependent prior, $\mathcal{N}(.; N_0, \sigma_2)$, where N_0 is a pretrained neural network, and $\mathcal{N}(.; ., \sigma)$ is a Gaussian with symmetric covariance σ. Then, for the weighted prior $\pi = \alpha\pi_{MDL} + (1 - \alpha)\,\pi_{dep}$ and posterior $\rho = \mathcal{N}(.; N, \sigma_3)$ we have:

$$\mathrm{KL}(\rho, \pi) \leq \alpha \left[|\widehat{N}|_c \log 2 - \log\left(m\left(|\widehat{N}|_c \right) \right) + \mathrm{KL}\left(\mathcal{N}(.; N, \sigma_3), \mathcal{N}\left(.; \widehat{N}, \sigma_2\right) \right) \right] + (1 - \alpha)\mathrm{KL}\left(\mathcal{N}(.; N, \sigma_3), \mathcal{N}(.; N_0, \sigma_2) \right) \tag{5.5}$$

The bound from equation 5.5 can be directly substituted into equation 5.3, after training N_0 on hold-out data following [131, 132].

5.5. Results

For our empirical investigation, we use the RPT scheme to interpret networks trained on functional genomics tasks related to cancer and psychiatric genomics.

5.5.1. Predictive Tasks

In the cancer-related task, we use somatic and germline genomic variation data from The PanCancer Analysis of Whole Genomes (PCAWG) study [75]. PCAWG includes a variety of biological data types corresponding to 2,800 samples from the International Cancer Genome Consortium. To train networks for our analysis, rare variants are singled out for Skin Melanoma and Esophageal Adenocarcinoma samples. The predictive task according

to which the neural networks have been trained is the prediction of somatic and germline variation co-occurrence at the gene level for 718 genes of the COSMIC census list fetched on May 08, 2018. Input data included 43 features ranging from germline variant signatures of known cancer genes alongside a set of biological features extracted from multiple data and annotation repositories, namely UCSC Genome Browser [133], Gencode v27 [134], and COSMIC [59]. Each model whose weights have been analyzed by *rank projection trees* has 3 hidden layers. Number of hidden nodes (285-941), optimization algorithm (Adam or Nesterov Adam), and activation functions (Exponential or Rectified Linear Unit) for each network have been determined by automated hyperparamter optimization using the `hyperopt` package [110]. Results are averaged over five neural networks trained on randomly stratified training datasets for each cancer type, with test performance of high precision and recall values ranging between 70% and 83%. To balance training datasets, we deployed the SMOTE oversampling algorithm [135] using the implementation in the `imbalanced-learn` Python package [136].

In the Schizophrenia-related task (SCZ), we use bulk data of the transcriptome, which is heavily affected by the environment, from the PsychENCODE [137, 138] consortium project. We create 10 training and testing partitions (including 640 and 70 samples respectively) of control and schizophrenia subjects, which are balanced 50-50% for controls and cases. We train neural networks with 2 hidden layers to predict a binary case/control indicator, with 100 and 400 nodes at layers 1 and 2 respectively, logistic sigmoid activations, and SGD with early stopping for training. We train separate neural networks using individual gene expression levels as inputs, and mean expression levels

across modules of genes pretrained using WGCNA [139], pre-selecting the top 1% and 15% of genes/modules respectively according to the absolute Pearson correlation between the input and the binary output indicator on each training partition (resulting in 187 genes and 754 modules in each respective model). The test performance of the models averaged across partitions was 73.6% and 66.1% for the gene- and module-based models, respectively.

5.5.2. Prioritization Functions

After training the networks, we first compare the relevance of input genes prioritized by RPT with different prioritization functions based on their literature associations with the disease under study on Google Scholar as a proxy for 'ground truth' ranking. We select a ranking function based on the signed values of learned weights, i.e. for layers l and $m = l + 1$, $r_{i,l,m}(j)$ returns the rank of $n_{m,j}$ in the ordering induced by the weights $W_{l,m}(i,.)$ (signed values, descending), and set $B = 2$. We then calculate the cumulative rank scores c_t over each path in *T* for the prioritization function π, i.e. for leaf node $t = [b_1, \ldots b_L]$, we set $c\_t = (\sum_l |b_l|)(\prod_l \text{sign}(b_l))$. We then set $\pi(n) = f(\phi^{-1}(n)) = g([c_{t_1}, c_{t_2}, \ldots])$ for $t_1, t_2 \ldots \in \phi^{-1}(n)$, where $g(.)$ is one of the functions: $\{\text{sum}, \text{average}, \max, \min\}$ of signed or absolute values, or length. In Table 5.1, we show the scored rankings of the top 20 genes (averaged across networks) according to each $g(.)$ based on the ℓ_1-distance between the two rankings (RPT and literature-based 'ground truth'), normalized by its maximum ($\ell_1([1 \ldots 20], [20 \ldots 1])$). In general, the absolute average and max functions appear to be better indicators of individual gene importance.

Network	**Absolute**				**Signed**				Count
	Sum	**Avg.**	**Max**	**Min**	**Sum**	**Avg.**	**Max**	**Min**	
Skin Mel.	0.628	**0583**	0.632	0.624	0.588	0.597	0.614	0.625	0.614
Es. Adeno.	0.655	0.644	**0.603**	0.661	0.670	0.612	0.642	0.671	0.670
Schiz.	0.642	**0.628**	0.664	0.656	0.710	0.674	0.708	0.686	0.634

Table 5.1. Prioritization function comparison. *The rankings of genes induced by different prioritization functions are compared against citation-based rankings from existing literature. Table shows normalized ℓ1-distances of predicted and citation-based rankings for the top 20 genes, with best performing metrics highlighted. Rows are: Skin Melanoma, Esophageal Adenocarcinoma, and Schizophrenia.*

5.5.3. Multigene Groupings

We then measure RPT's ability to extract multigene groupings of high relevance to disease. To generate RPT groupings, we rely on the networks with the WGCNA modules average expression levels as inputs. Hence, for set S formed from equation 5.2, we take the union of the genes in all modules which are elements of S, where all positive and negative groupings across layers $l = 0\ldots3$ of the schizophrenia networks are extracted. In parallel, we extract similar groupings but based on a randomized ranking function as a control. After a gene-set enrichment analysis to annotate all of the groupings based on the KEGG pathway terms with a q-value < 0.001, a ranking across KEGG terms is calculated

independently for each layer by counting the number of groupings a term is associated with across all models (including duplicate groupings, hence accounting for increased importance of nodes in N mapped to by multiple nodes in T). A literature-based analysis of the top 20 terms demonstrates that the RPT scheme selects more disease-relevant terms than the random tree across all network layers, with the groupings selected at the higher hidden layers (i.e. closer to input layer) associating with more trait-relevant. Literature analysis is based on Google Scholar results, and gene-set enrichment analysis is performed using clusterProfiler [140].

We then extend the analysis to measure the enrichment for "high-confidence trait genes" in the groupings from all layers. In our example, these are genes which can be linked to GWAS hits for schizophrenia by any three of the following four methods: Hi-C interactions; enhancer-target links; eQTL linkages, and isoform-QTL linkages (321 genes; list available in [137]). The enrichment of such genes in each module is scored using a *p*-value from the hyper-geometric test. Figure 5.3 shows that these genes are significantly more enriched in the RPT groupings than randomized ones, with the penultimate layer L_1, again, being the most trait-relevant. We also compare the L_3 distribution with a gradient-based prioritization scheme which ranks the modules at this level according to the absolute magnitude of the gradient of the network output with respect to each input (as in [128]), but report that it is not significantly better than the randomized tree (p=0.78); the RPT is better than the randomized tree across all layers ($p < 10^{-4}$ for L_{1-3}, and p=0.012 for L_0, all *p*-values using 1-tail KS-tests).

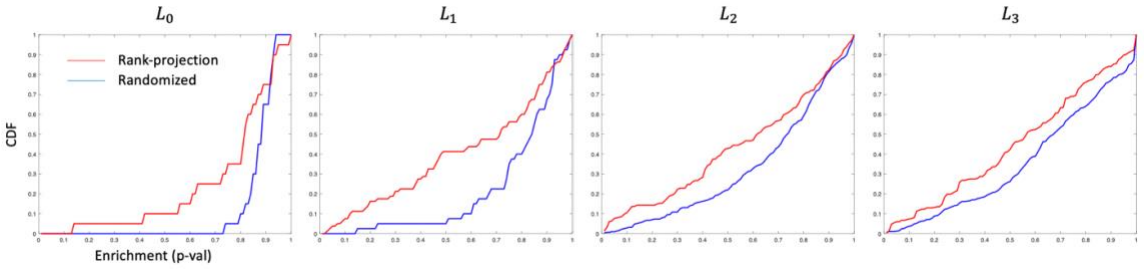


Figure 5.3. Visualization of network interpretations. *Enrichment of high-confidence schizophrenia associated genes in gene groupings found associated with different neural network layers. Enrichment p-values are from the hyper-geometric test, and the empirical cumulative density function (CDF) is plotted on the y-axis.*

5.5.4. Interpretability and Generalization

Finally, we investigate the relationship between generalization and interpretability of neural networks compressed using the RPT scheme, and whether there is a relationship between the strength of the semantic associations and the generalization of the network. We begin by investigating the strength of association between the PAC-Bayes bound (equation 5.3) with the MDL-prior (equation 5.4) in predicting the test error. To compress networks to multiple degrees, we vary the RPT half-branching factor to $B = \{1, 2, \ldots 5\}$ on each of the 10 SCZ data splits. We evaluate the bound in equation 5.3 by compressing the parameters of the resulting networks using LZW compression [33], and use the resulting binary code as the network representation, $\widehat{N}$ in equation 5.4. Figure 5.4 shows plots of the correlation between MDL bounds and empirical test error, where the bound parameter λ is fitted independently for each group of compressed models associated with a given

uncompressed model (corresponding to a fixed data split); hence each plot includes networks associated with $B = \{1,2, \ldots 5\}$. We further standardize each bound, by dividing by the variance per plot, and subtracting the minimum distance with the test error across all models. A significant positive correlation is detected between bounds and test error (sign-test, $p = 0.022$), suggesting a considerable ability to predict test error by the predefined bounds.

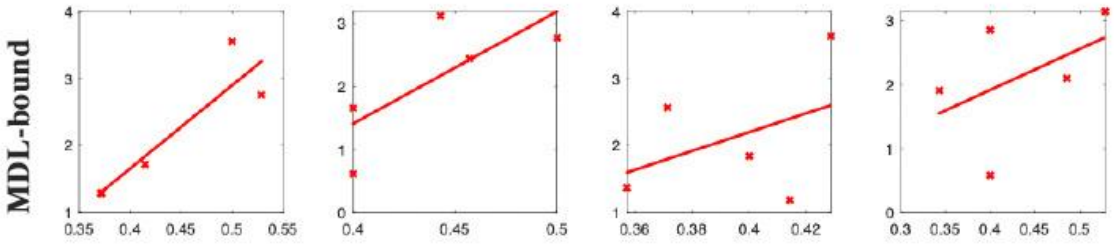


Figure 5.4. *Plots of test error vs. MDL bounds for groups of models derived from the RPT scheme.*

We then measure the correlation between the predicted accuracy of each model, defined as $1 -$ MDL-bound, with the KS test statistic, calculated as in Sec. 5.5.2 for the gene groupings derived per-layer from the compressed networks (based on the enrichment of high-confidence SCZ genes from [137]), to investigate the relationship between generalization and interpretability by measuring. Figure 5.5 shows significantly skewed positive correlations across all models ($p = 0.0072$, 1-sample t-test).

Finally, we combine MDL and data-dependent components to investigate the modified MDL bound. We replicate the generalization test from Figure 5.4 using this bound, while fitting λ and α for each group of models. This achieves an improved mean correlation of $r = 0.955$ for the top 3 models, which is again significant (permutation test, $p < 0.05$).

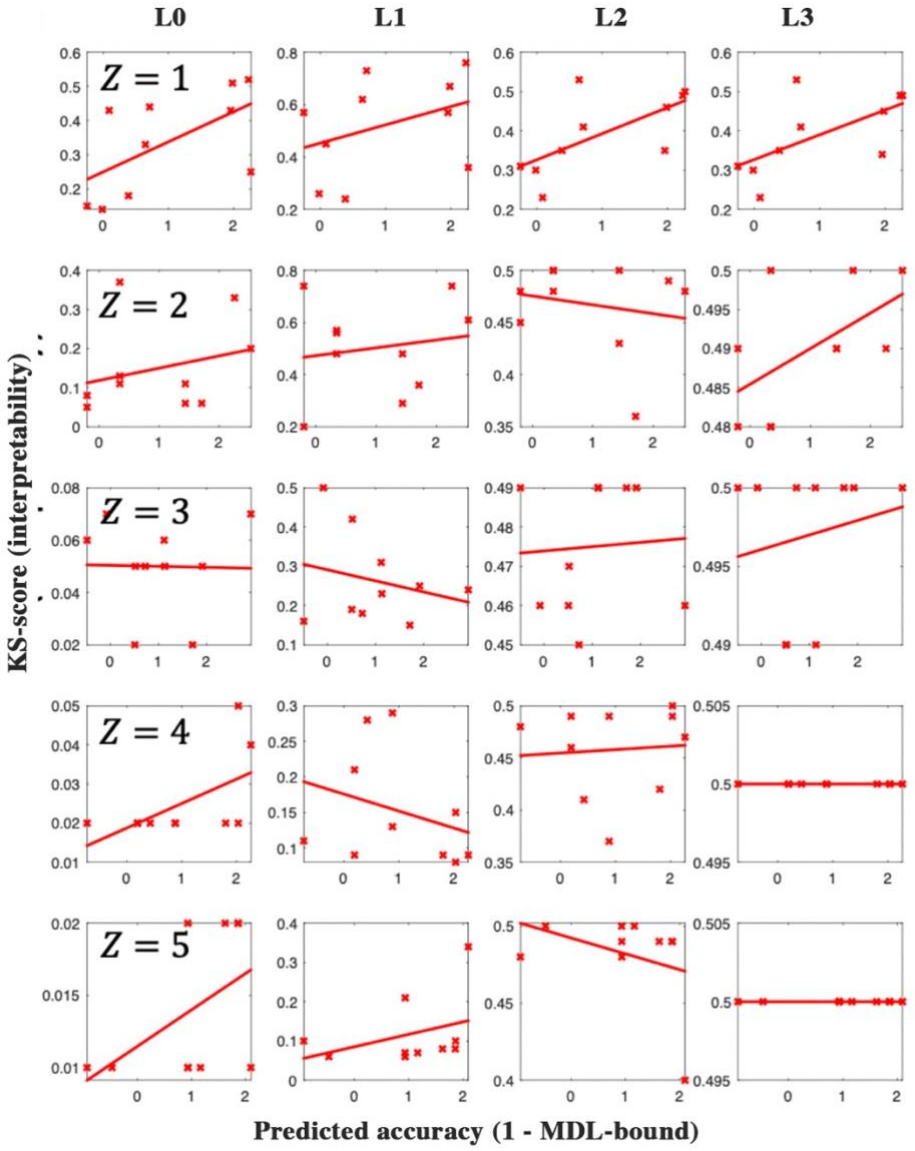


Figure 5.5. *The relationship between interpretability and predicted accuracy (defined as 1 − MDL-bound) per network layer (L) and compression strength (Z) for the RPT scheme.*

Table 5.2 further compares the mean KS-bound correlation across models (as in Figure 5.5) of the MDL and modified-MDL bound, showing only a marginal increase in

correlation. The table shows that both bounds achieve a significantly stronger correlation with the semantic KS-scores than the observed test error. A possible reason for this is the use of the training error in the first term of the bound, which is typically more stable than the test error, given the larger number of data points (640 vs. 70 in the SCZ datasets).

Bound	**Test error**	**MDL**	**MDL + prior**
Mean *r*	0.142	0.217	0.218
p-val	0.025	0.007	0.007

Table 5.2. Comparing MDL and Modified MDL bounds for predicting semantic enrichment. *Table shows mean Pearson correlation and 1-sample t-test p-values for models using the RPT scheme, when the KS-semantic-enrichment is correlated with the quantities shown.*

5.6. Conclusion

We introduced a post-hoc network compression scheme, namely Rank Projection Trees (RPTs), that enhances the interpretability of neural network models trained using functional genomics data. We showed that RPTs are able to pick out biologically important genes using networks trained to predict epistatic interactions of germline and somatic mutations in cancer and risk for Schizophrenia. Finally, we established a link between network generalization and interpretation quality and describe MDL PAC-Bayes bounds that can be used as proxy scores for interpretable model selection, hence supporting the general applicability of compression-based interpretability schemes.

Chapter 6

Conclusion

Rapid technological advances across computational and biomedical disciplines brought to the fore unprecedented challenges and perspectives. The available number of sequenced cancer exomes (and genomes) allowed for a plethora of approaches to test long held hypotheses about the nature of cancer and its heterogeneity across and within tumor types. In parallel, new experimental assays expanded the functional genomics landscape resulting a multitude of angles through which cancer genomics could be studied, among which are the ones underlying the works described in this document: patterns of—altered and "normal"—gene product interactions, cellular response to therapeutic regimens, and the potential cancer relevance of each human gene.

One narrative that dominated the study of genomic variation underlying carcinogenesis focused on a binary description of mutations: recurrent drivers in selected genes that confer selective advantage on cancer cells, and passengers with assumed neutral effect. The previous chapters investigate a central question in light of aforementioned technological attainments: do patterns of genomic variation transcend this binary divide into a spectrum of gene effects? To answer this question, we relied on two types of networks.

In Chapters 2 and 3, we leveraged functional networks that summarize different types of associations between gene products, somatic and germline genomic variation data, and

recently available data on cancer cell viability to prioritize human genes with respect to their relative contributions to cancer development. In Chapters 4 and 5, we turned to artificial neural networks, a class of complex machine learning methods with growing popularity and notable performance. In these two chapters, we suggested three approaches to simplify models and render them more interpretable in order to establish trust in their predictions. On the applications side, we demonstrated the efficacy of the interpretability approaches and prioritized input features, including at the single sample level, in multiple biomedical tasks pertaining to functional genomics.

The nexus of computational methods and cancer theory is expected to garner more attention in the years to come. Expectedly, utilizing new tools to address questions pertaining to cancer genomics has begotten new, fine-scale questions. The growing availability of a population-scale biobanks around the world will likely open more venues to study the spectrum of gene relevance at the—coding and noncoding—DNA level with higher resolution. Recent experimental assays will further enrich the functional lens through which we study biological processes underlying physiology and disease, and the cancer relevance spectrum will likely be among the top targets of research. Together with the immense amount of labor put towards computational method development and the growing investment in cloud technologies, it is expected that our understanding of cancer, or cancers to be more precise, will reach higher levels. It is in this context that the work described in this document has been pursued over the last several years, and it is with the hope that its chapters constitute a contribution on this long road.

References

1. Pon JR, Marra MA: **Driver and Passenger Mutations in Cancer.** *Annual Review of Pathology: Mechanisms of Disease* 2015, **10:**25-50.
2. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW: **Cancer genome landscapes.** *Science* 2013, **339:**1546-1558.
3. Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, Vazquez M, Fink JL, Kassahn KS, Pearson JV, et al: **Computational approaches to identify functional genetic variants in cancer genomes.** *Nat Methods* 2013, **10:**723-729.
4. Forbes SA, Beare D, Bindal N, Bamford S, Ward S, Cole CG, Jia M, Kok C, Boutselakis H, De T, et al: **COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer.** *Curr Protoc Hum Genet* 2016, **91:**10 11 11-10 11 37.
5. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, Schnall-Levin M, White J, Sanford EM, An P, et al: **Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing.** *Nature Biotechnology* 2013, **31:**1023-1031.
6. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G: **Discovery and saturation analysis of cancer genes across 21 tumour types.** *Nature* 2014, **505:**495-501.

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

8. Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, Yakovleva A, Palmieri T, Ciccarelli FD: **The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens.** *Genome Biol* 2019, **20:**1.

9. Rahman N: **Realizing the promise of cancer predisposition genes.** *Nature* 2014, **505:**302-308.

10. Pritchard CC, Salipante SJ, Koehler K, Smith C, Scroggins S, Wood B, Wu D, Lee MK, Dintzis S, Adey A, et al: **Validation and implementation of targeted capture and sequencing for the detection of actionable mutation, copy number variation, and gene rearrangement in clinical cancer specimens.** *J Mol Diagn* 2014, **16:**56-67.

11. McFarland CD, Yaglom JA, Wojtkowiak JW, Scott JG, Morse DL, Sherman MY, Mirny LA: **The Damaging Effect of Passenger Mutations on Cancer Progression.** *Cancer Res* 2017, **77:**4763-4772.

12. Lu C, Xie M, Wendl MC, Wang J, McLellan MD, Leiserson MD, Huang KL, Wyczalkowski MA, Jayasinghe R, Banerjee T, et al: **Patterns and functional implications of rare germline variants across 12 cancer types.** *Nat Commun* 2015, **6:**10086.

13. Nussinov R, Tsai CJ: **'Latent drivers' expand the cancer mutational landscape.** *Curr Opin Struct Biol* 2015, **32:**25-32.

14. Castro-Giner F, Ratcliffe P, Tomlinson I: **The mini-driver model of polygenic cancer evolution.** *Nat Rev Cancer* 2015, **15:**680-685.

15. Kumar S, Warrell J, Li S, McGillivray PD, Meyerson W, Salichos L, Harmanci A, Martinez-Fundichely A, Chan CWY, Nielsen MM, et al: **Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences.** *Cell* 2020, **180:**915-927 e916.

16. McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA: **Impact of deleterious passenger mutations on cancer progression.** *Proc Natl Acad Sci U S A* 2013, **110:**2910-2915.

17. McFarland CD, Mirny LA, Korolev KS: **Tug-of-war between driver and passenger mutations in cancer and other adaptive processes.** *Proc Natl Acad Sci U S A* 2014, **111:**15138-15143.

18. Mohsen H, Gunasekharan V, Qing T, Seay M, Surovtseva Y, Negahban S, Szallasi Z, Pusztai L, Gerstein MB: **Network propagation-based prioritization of long tail genes in 17 cancer types.** *Genome Biol* 2021, **22:**287.

19. Pon JR, Marra MA: **Driver and passenger mutations in cancer.** *Annu Rev Pathol* 2015, **10:**25-50.

20. Loganathan SK, Schleicher K, Malik A, Quevedo R, Langille E, Teng K, Oh RH, Rathod B, Tsai R, Samavarchi-Tehrani P, et al: **Rare driver mutations in head and neck squamous cell carcinomas converge on NOTCH signaling.** *Science* 2020, **367:**1264-1269.

21. Scholl C, Frohling S: **Exploiting rare driver mutations for precision cancer medicine.** *Curr Opin Genet Dev* 2019, **54:**1-6.

22. Armenia J, Wankowicz SAM, Liu D, Gao J, Kundra R, Reznik E, Chatila WK, Chakravarty D, Han GC, Coleman I, et al: **The long tail of oncogenic drivers in prostate cancer.** *Nat Genet* 2018, **50:**645-651.

23. Elman JS, Ni TK, Mengwasser KE, Jin D, Wronski A, Elledge SJ, Kuperwasser C: **Identification of FUBP1 as a Long Tail Cancer Driver and Widespread Regulator of Tumor Suppressor and Oncogene Alternative Splicing.** *Cell Rep* 2019, **28:**3435-3449 e3435.

24. Consortium I-TP-CAoWG: **Pan-cancer analysis of whole genomes.** *Nature* 2020, **578:**82-93.

25. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al: **Comprehensive Characterization of Cancer Driver Genes and Mutations.** *Cell* 2018, **173:**371-385 e318.

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

27. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM: **Prioritizing candidate disease genes by network-based boosting of genome-wide association data.** *Genome Res* 2011, **21:**1109-1121.

28. Erten S, Bebek G, Ewing RM, Koyuturk M: **DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization.** *BioData Min* 2011, **4:**19.

29. Erten S, Bebek G, Koyuturk M: **Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks.** *J Comput Biol* 2011, **18:**1561-1574.

30. Cao M, Zhang H, Park J, Daniels NM, Crovella ME, Cowen LJ, Hescott B: **Going the distance for protein function prediction: a new distance metric for protein interaction networks.** *PLoS One* 2013, **8:**e76339.

31. Cao M, Pietras CM, Feng X, Doroschak KJ, Schaffner T, Park J, Zhang H, Cowen LJ, Hescott BJ: **New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence.** *Bioinformatics* 2014, **30:**i219-227.

32. Cowen L, Ideker T, Raphael BJ, Sharan R: **Network propagation: a universal amplifier of genetic associations.** *Nat Rev Genet* 2017, **18:**551-562.

33. Köhler S, Bauer S, Horn D, Robinson PN: **Walking the interactome for prioritization of candidate disease genes.** *Am J Hum Genet* 2008, **82:**949-958.

34. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R: **Associating genes and protein complexes with disease via network propagation.** *PLoS Comput Biol* 2010, **6:**e1000641.

35. Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, Stuart JM: **Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE).** *Bioinformatics* 2013, **29:**2757-2764.

36. Singh-Blom UM, Natarajan N, Tewari A, Woods JO, Dhillon IS, Marcotte EM: **Prediction and validation of gene-disease associations using methods inspired by social network analyses.** *PLoS One* 2013, **8:**e58977.

37. Ruffalo M, Koyuturk M, Sharan R: **Network-Based Integration of Disparate Omic Data To Identify "Silent Players" in Cancer.** *PLoS Comput Biol* 2015, **11:**e1004595.

38. Shnaps O, Perry E, Silverbush D, Sharan R: **Inference of Personalized Drug Targets Via Network Propagation.** *Pac Symp Biocomput* 2016, **21:**156-167.

39. Hofree M, Shen JP, Carter H, Gross A, Ideker T: **Network-based stratification of tumor mutations.** *Nat Methods* 2013, **10:**1108-1115.

40. Reyna MA, Leiserson MDM, Raphael BJ: **Hierarchical HotNet: identifying hierarchies of altered subnetworks.** *Bioinformatics* 2018, **34:**i972-i980.

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

42. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, Papoutsaki A, Kim Y, Niu B, McLellan M, et al: **Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes.** *Nat Genet* 2015, **47:**106-114.

43. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, et al: **Defining a Cancer Dependency Map.** *Cell* 2017, **170:**564-576 e516.

44. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res* 2009, **19:**1639-1645.

45. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, et al: **Oncogenic Signaling Pathways in The Cancer Genome Atlas.** *Cell* 2018, **173:**321-337 e310.

46. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, Wadi L, Meyer M, Wong J, Xu C, et al: **Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap.** *Nat Protoc* 2019, **14:**482-517.

47. Herrero-Gonzalez S, Di Cristofano A: **New routes to old places: PIK3R1 and PIK3R2 join PIK3CA and PTEN as endometrial cancer genes.** *Cancer Discov* 2011, **1:**106-107.

48. Pan F, Zhang J, Tang B, Jing L, Qiu B, Zha Z: **The novel circ_0028171/miR-218-5p/IKBKB axis promotes osteosarcoma cancer progression.** *Cancer Cell Int* 2020, **20:**484.

49. Torres HA, Shigle TL, Hammoudi N, Link JT, Samaniego F, Kaseb A, Mallet V: **The oncologic burden of hepatitis C virus infection: A clinical perspective.** *CA Cancer J Clin* 2017, **67:**411-431.

50. Haggstrom C, Van Hemelrijck M, Zethelius B, Robinson D, Grundmark B, Holmberg L, Gudbjornsdottir S, Garmo H, Stattin P: **Prospective study of Type 2 diabetes mellitus, anti-diabetic drugs and risk of prostate cancer.** *Int J Cancer* 2017, **140:**611-617.

51. Shlomai G, Neel B, LeRoith D, Gallagher EJ: **Type 2 Diabetes Mellitus and Cancer: The Role of Pharmacotherapy.** *J Clin Oncol* 2016, **34:**4261-4269.

52. Tagaya Y, Gallo RC: **The Exceptional Oncogenicity of HTLV-1.** *Front Microbiol* 2017, **8:**1425.

53. Kuderer NM, Choueiri TK, Shah DP, Shyr Y, Rubinstein SM, Rivera DR, Shete S, Hsu C-Y, Desai A, de Lima Lopes G, Jr., et al: **Clinical impact of COVID-19 on patients with cancer (CCC19): a cohort study.** *The Lancet* 2020, **395:**1907-1918.

54. McFarland JM, Ho ZV, Kugener G, Dempster JM, Montgomery PG, Bryan JG, Krill-Burger JM, Green TM, Vazquez F, Boehm JS, et al: **Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration.** *Nat Commun* 2018, **9:**4610.

55. Yang L, Chen E, Goodison S, Sun Y: **An efficient and effective method to identify significantly perturbed subnetworks in cancer.** *Nat Comput Sci* 2021, **1:**79-88.

56. Zhou DY, Bousquet O, Lal TN, Weston J, Scholkopf B: **Learning with local and global consistency.** *Advances in Neural Information Processing Systems 16* 2004, **16:**321-328.

57. Hristov BH, Singh M: **Network-Based Coverage of Mutational Profiles Reveals Cancer Genes.** *Cell Syst* 2017, **5:**221-229 e224.

58. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al: **Mutational heterogeneity in cancer and the search for new cancer-associated genes.** *Nature* 2013, **499:**214-218.

59. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al: **COSMIC: the Catalogue Of Somatic Mutations In Cancer.** *Nucleic Acids Res* 2019, **47:**D941-D947.

60. Ramakrishnan SR, Vogel C, Kwon T, Penalva LO, Marcotte EM, Miranker DP: **Mining gene functional networks to improve mass-spectrometry-based protein identification.** *Bioinformatics* 2009, **25:**2955-2961.

61. Langville ANaM, Carl D.: **Deeper Inside PageRank.** *Internet Mathematics* 2003, **1**.

62. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J: **g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update).** *Nucleic Acids Res* 2019, **47:**W191-W198.

63. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.

64. Mohsen H, Gunasekharan V, Qing T, Seay M, Surovtseva Y, Negahban S, Szallasi Z, Pusztai L, Gerstein MB: **Network propagation-based prioritization of long tail genes in 17 cancer types** pp. https://github.com/gersteinlab/UMG. GitHub:https://github.com/gersteinlab/UMG.

65. Mohsen H, Gunasekharan V, Qing T, Seay M, Surovtseva Y, Negahban S, Szallasi Z, Pusztai L, Gerstein MB: **Network propagation-based prioritization of long tail genes in 17 cancer types** pp. https://doi.org/10.5281/zenodo.5500467:https://doi.org/10.5281/zenodo.5500467.

66. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan M, et al: **Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines.** *Cell Syst* 2018, **6:**271-281 e277.

67. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al: **STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.** *Nucleic Acids Res* 2019, **47:**D607-D613.

68. Hwang S, Kim CY, Yang S, Kim E, Hart T, Marcotte EM, Lee I: **HumanNet v2: human gene networks for disease research.** *Nucleic Acids Res* 2019, **47:**D573-D580.

69. Wang Q, Armenia J, Zhang C, Penson AV, Reznik E, Zhang L, Minet T, Ochoa A, Gross BE, Iacobuzio-Donahue CA, et al: **Unifying cancer and normal RNA sequencing data from different sources.** *Sci Data* 2018, **5:**180061.

70. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res* 2010, **38:**e164.

71. Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nat Protoc* 2009, **4:**1184-1191.

72. Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM: **CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer.** *Nat Methods* 2019, **16:**505-507.

73. Qing T, Mohsen H, Cannataro VL, Marczyk M, Rozenblit M, Foldi J, Murray MF, Townsend JP, Kluger Y, Gerstein M, Pusztai L: **Cancer relevance of human genes.** *bioRxiv* 2021**:**2021.2002.2004.429823.

74. Martin GS: **The hunting of the Src.** *Nat Rev Mol Cell Biol* 2001, **2:**467-475.

75. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, Perry MD, Nahal-Bose HK, Ouellette BFF, Li CH, et al: **Pan-cancer analysis of whole genomes.** *Nature* 2020, **578:**82-93.

76. Iranzo J, Martincorena I, Koonin EV: **Cancer-mutation network and the number and specificity of driver mutations.** *Proc Natl Acad Sci U S A* 2018, **115:**E6010-E6019.

77. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R: **Evaluating the evaluation of cancer driver genes.** *Proc Natl Acad Sci U S A* 2016, **113:**14330-14335.

78. Agarwal D, Nowak C, Zhang NR, Pusztai L, Hatzis C: **Functional germline variants as potential co-oncogenes.** *NPJ Breast Cancer* 2017, **3:**46.

79. Cannataro VL, Townsend JP: **Neutral Theory and the Somatic Evolution of Cancer.** *Mol Biol Evol* 2018, **35:**1308-1315.

80. Qing T, Mohsen H, Marczyk M, Ye Y, O'Meara T, Zhao H, Townsend JP, Gerstein M, Hatzis C, Kluger Y, Pusztai L: **Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden.** *Nature Communications* 2020, **11:**2438.

81. Cannataro VL, Gaffney SG, Townsend JP: **Effect Sizes of Somatic Mutations in Cancer.** *J Natl Cancer Inst* 2018, **110:**1171-1177.

82. Hughes AL: **Near neutrality: leading edge of the neutral theory of molecular evolution.** *Ann N Y Acad Sci* 2008, **1133:**162-179.

83. Cassa CA, Weghorn D, Balick DJ, Jordan DM, Nusinow D, Samocha KE, O'Donnell-Luria A, MacArthur DG, Daly MJ, Beier DR, Sunyaev SR: **Estimating the selective effects of heterozygous protein-truncating variants from human exome data.** *Nat Genet* 2017, **49:**806-810.

84. Van Hout CV, Tachmazidou I, Backman JD, Hoffman JD, Liu D, Pandey AK, Gonzaga-Jauregui C, Khalid S, Ye B, Banerjee N, et al: **Exome sequencing and characterization of 49,960 individuals in the UK Biobank.** *Nature* 2020, **586:**749-756.

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

86. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, Cagan A, Murai K, Mahbubani K, Stratton MR, et al: **Somatic mutant clones colonize the human esophagus with age.** *Science* 2018, **362:**911-917.

87. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN, et al: **Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology.** *J Mol Diagn* 2015, **17:**251-264.

88. Haigis KM, Cichowski K, Elledge SJ: **Tissue-specificity in cancer: The rule, not the exception.** *Science* 2019, **363:**1150-1151.

89. Schneider G, Schmidt-Supprian M, Rad R, Saur D: **Tissue-specific tumorigenesis: context matters.** *Nat Rev Cancer* 2017, **17:**239-253.

90. Dijkstra EW: **A note on two problems in connexion with graphs.** *Numerische Mathematik* 1959, **1:**269-271.

91. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, et al: **The repertoire of mutational signatures in human cancer.** *Nature* 2020, **578:**94-101.

92. Jonckheere AR: **A DISTRIBUTION-FREE k-SAMPLE TEST AGAINST ORDERED ALTERNATIVES.** *Biometrika* 1954, **41:**133-145.

93. Mohsen H, Warrell J, Min MR, Negahban S, Gerstein M: **Weight-based Neural Network Interpretability using Activation Tuning and Personalized Products.** *Machine Learning for Computational Biology Workshop (MLCB'20)* 2020.

94. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow P-M, Zietz M, Hoffman MM, et al: **Opportunities and obstacles for deep learning in biology and medicine.** *Journal of The Royal Society Interface* 2018, **15:**20170387.

95. Hochreiter S, Heusel M, Obermayer K: **Fast model-based protein homology detection without alignment.** *Bioinformatics* 2007, **23:**1728-1736.

96. Leung MK, Xiong HY, Lee LJ, Frey BJ: **Deep learning of the tissue-regulated splicing code.** *Bioinformatics* 2014, **30:**i121-129.

97. Zhou J, Troyanskaya OG: **Predicting effects of noncoding variants with deep learning-based sequence model.** *Nat Methods* 2015, **12:**931-934.

98. Alipanahi B, Delong A, Weirauch MT, Frey BJ: **Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.** *Nat Biotechnol* 2015, **33:**831-838.

99. Ploenzke MS, Irizarry RA: **Interpretable Convolution Methods for Learning Genomic Sequence Motifs.** *bioRxiv* 2018**:**411934.

100. Umarov RK, Solovyev VV: **Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks.** *PLOS ONE* 2017, **12:**e0171410.

101. Yousefi S, Amrollahi F, Amgad M, Dong C, Lewis JE, Song C, Gutman DA, Halani SH, Velazquez Vega JE, Brat DJ, Cooper LAD: **Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models.** *Scientific Reports* 2017, **7:**11707.

102. Warrell J, Mohsen H, Gerstein M: **Rank Projection Trees for Multilevel Neural Network Interpretation,.** *NeurIPS Machine Learning for Health Workshop (NeurIPS'18 ML4H)* 2018.

103. Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, Ma J, Ideker T: **Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells.** *Cancer Cell* 2020, **38:**672-684.e676.

104. Tishby N, Zaslavsky N: **Deep Learning and the Information Bottleneck Principle.** pp. arXiv:1503.02406; 2015:arXiv:1503.02406.

105. Roffo G: **Ranking to Learn and Learning to Rank: On the Role of Ranking in Pattern Recognition Applications.** *arXiv e-prints* 2017**:**arXiv:1706.05933.

106. Garson GD: **Interpreting neural-network connection weights.** *AI Expert* 1991, **6:**46–51.

107. Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL, Long N, Schultz AR, Traer E, Abel M, et al: **Functional genomic landscape of acute myeloid leukaemia.** *Nature* 2018, **562:**526-531.

108. Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, Clore JN: **Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records.** *BioMed Research International* 2014, **2014:**781670.

109. Goudjerkan T, Jayabalan M: **Predicting 30-Day Hospital Readmission for Diabetes Patients using Multilayer Perceptron.** *International Journal of Advanced Computer Science and Applications* 2019, **10**.

110. Bergstra J, Komer B, Eliasmith C, Yamins D, Cox DD: **Hyperopt: a Python library for model selection and hyperparameter optimization.** *Computational Science & Discovery* 2015, **8:**014008.

111. Jiao X, Sherman BT, Huang da W, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID-WS: a stateful web service to facilitate gene/protein list analysis.** *Bioinformatics* 2012, **28:**1805-1806.

112. Frankle J, Carbin M: **The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks.** 2019.

113. Morcos AS, Barrett DGT, Rabinowitz NC, Botvinick M: **On the importance of single directions for generalization.** In *International Conference on Learning Representations*; 2018.

114. Tsang M, Cheng D, Liu Y: **Detecting Statistical Interactions from Neural Network Weights.** *International Conference on Learning Representation* 2018.

115. Weinberger E, Janizek JD, Lee S-I: **Learning Deep Attribution Priors Based On Prior Knowledge.** In *Neural Information Processing Systems*, vol. 33. pp. 14034-14045; 2020:14034-14045.

116. Shrikumar A, Greenside P, Kundaje A: **Learning important features through propagating activation differences.** In *International Conference on Machine Learning*. pp. 3145-3153; 2017:3145-3153.

117. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller MA: **Striving for Simplicity: The All Convolutional Net.** In *International Conference on Learning Representations*; 2015.

118. Warrell J, Mohsen H, Gerstein M: **Rank Projection Trees for Multilevel Neural Network Interpretation.** *NeurIPS Machine Learning for Health Workshop (NeurIPS'18 ML4H)* 2018.

119. Wells L, Bednarz T: **Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends.** *Frontiers in Artificial Intelligence* 2021, **4**.

120. Zhang Y, Tiňo P, Leonardis A, Tang K: **A Survey on Neural Network Interpretability.** *arXiv e-prints* 2020**:**arXiv:2012.14261.

121. Han S, Mao H, Dally WJ: **Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding.** *arXiv e-prints* 2015**:**arXiv:1510.00149.

122. Zhou W, Veitch V, Austern M, Adams RP, Orbanz P: **Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach.** In *ICLR*. 2019

123. Daniely A, Granot E: **Generalization bounds for neural networks via approximate description length.** In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc.; 2019: Article 1165

124. Gal Y, Ghahramani Z: **Dropout as a Bayesian approximation: representing model uncertainty in deep learning.** In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. pp. 1050–1059. New York, NY, USA: JMLR.org; 2016:1050–1059.

125. Ribeiro MT, Singh S, Guestrin C: **"Why Should I Trust You?": Explaining the Predictions of Any Classifier.** In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 1135–1144. San Francisco, California, USA: Association for Computing Machinery; 2016:1135–1144.

126. Lundberg SM, Lee S-I: **A unified approach to interpreting model predictions.** In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. pp. 4768–4777. Long Beach, California, USA: Curran Associates Inc.; 2017:4768–4777.

127. Lage I, Ross AS, Kim B, Gershman SJ, Doshi-Velez F: **Human-in-the-loop interpretability prior.** In *Proceedings of the 32nd International Conference on*

Neural Information Processing Systems. pp. 10180–10189. Montréal, Canada: Curran Associates Inc.; 2018:10180–10189.

128. Simonyan K, Vedaldi A, Zisserman A: **Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.** pp. arXiv:1312.6034; 2013:arXiv:1312.6034.

129. Sundararajan M, Taly A, Yan Q: **Gradients of Counterfactuals.** pp. arXiv:1611.02639; 2016:arXiv:1611.02639.

130. Alquier P, Ridgway J, Chopin N: **On the properties of variational approximations of Gibbs posteriors.** *J Mach Learn Res* 2016, **17:**8374–8414.

131. Bernhard S, John P, Thomas H: **Tighter PAC-Bayes Bounds.** In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference.* MIT Press; 2007: 9-16

132. Parrado-Hernández E, Ambroladze A, Shawe-Taylor J, Sun S: **PAC-bayes bounds with data dependent priors.** *J Mach Learn Res* 2012, **13:**3507–3531.

133. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.

134. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22:**1760-1774.

135. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: **SMOTE: synthetic minority over-sampling technique.** *J Artif Int Res* 2002, **16:**321–357.

136. Lemaître G, Nogueira F, Aridas CK: **Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning.** *J Mach Learn Res* 2017, **18:**559–563.

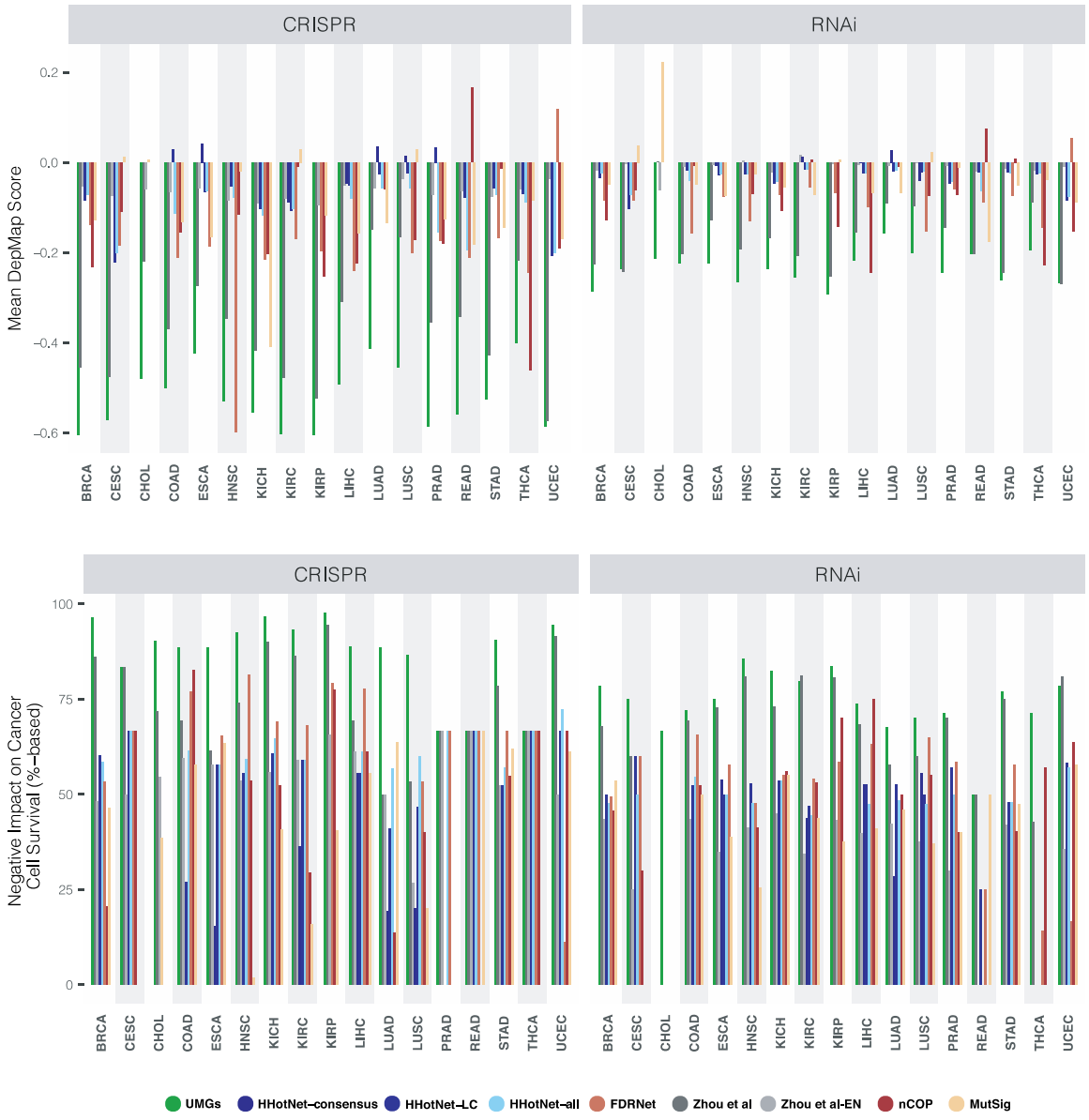
137. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, Clarke D, Gu M, Emani P, Yang YT, et al: **Comprehensive functional genomic resource and integrative model for the human brain.** *Science* 2018, **362:**eaat8464.

138. Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, Crawford GE, Jaffe AE, Pinto D, Dracheva S, Geschwind DH, et al: **The PsychENCODE project.** *Nature Neuroscience* 2015, **18:**1707-1712.

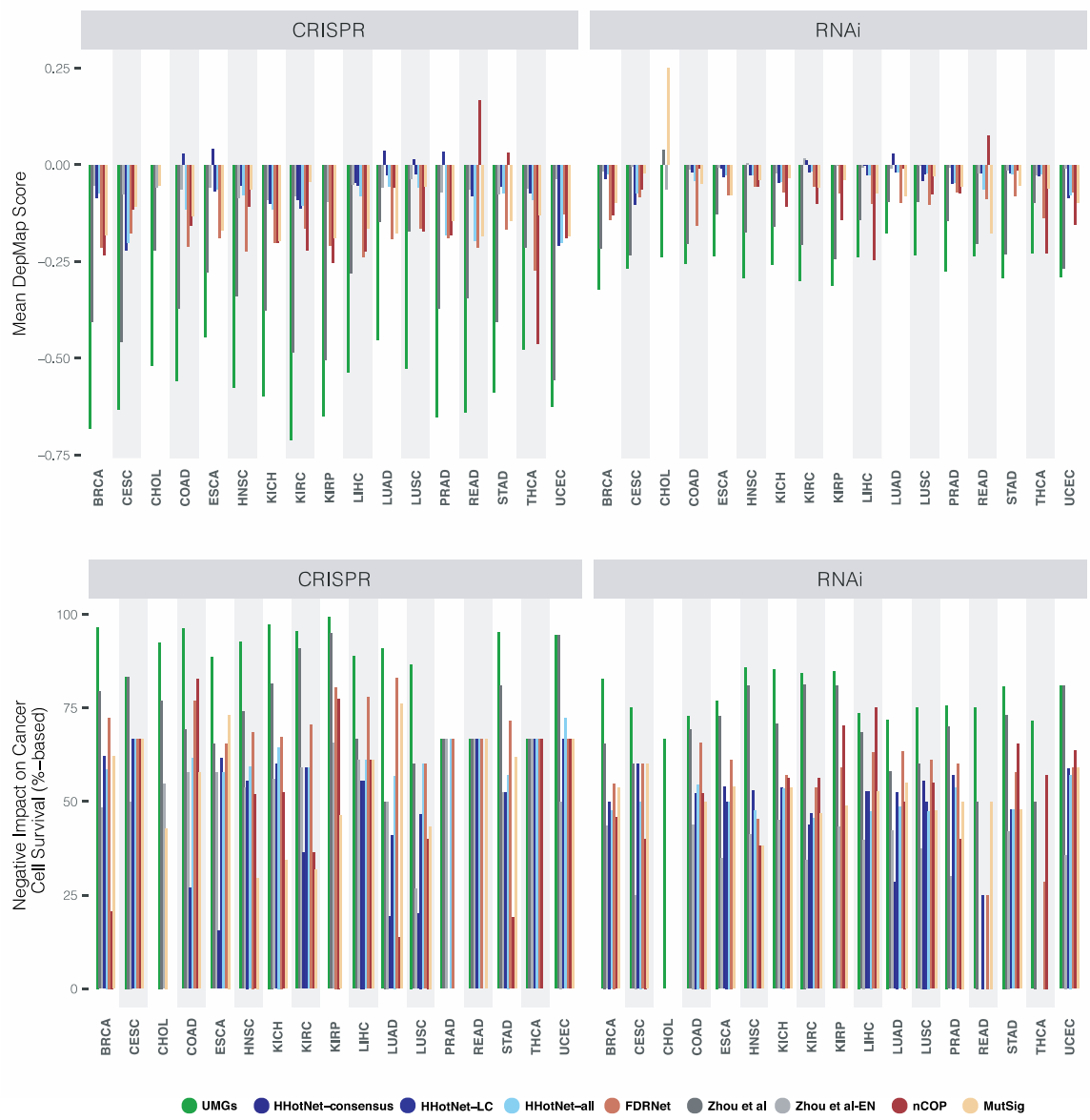
139. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC Bioinformatics* 2008, **9:**559.

140. Yu G, Wang LG, Han Y, He QY: **clusterProfiler: an R package for comparing biological themes among gene clusters.** *Omics* 2012, **16:**284-287.

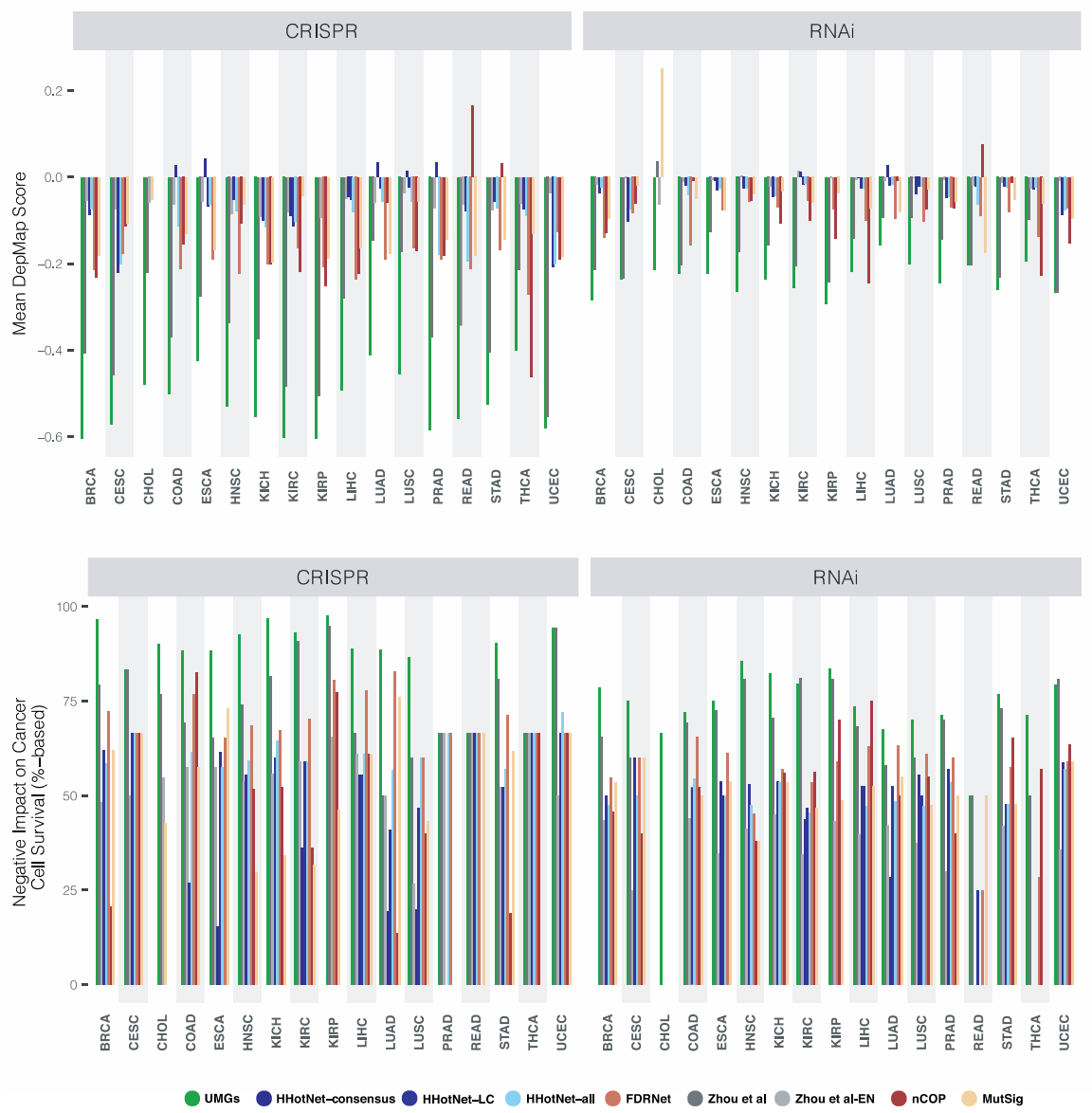
Supplementary Figures



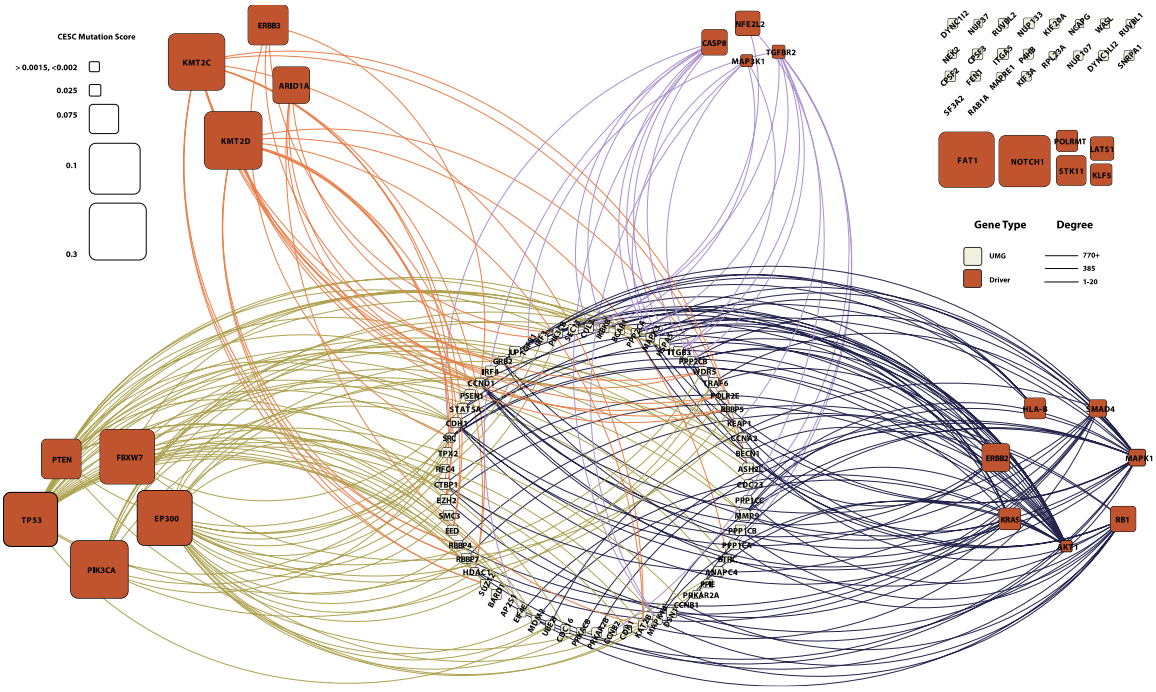
Supplementary Figure 1. *Impact on cancer cell line survival of UMG lists before the DepMap filtering step compared to other methods' lists. Extension to Figure 2.4.*



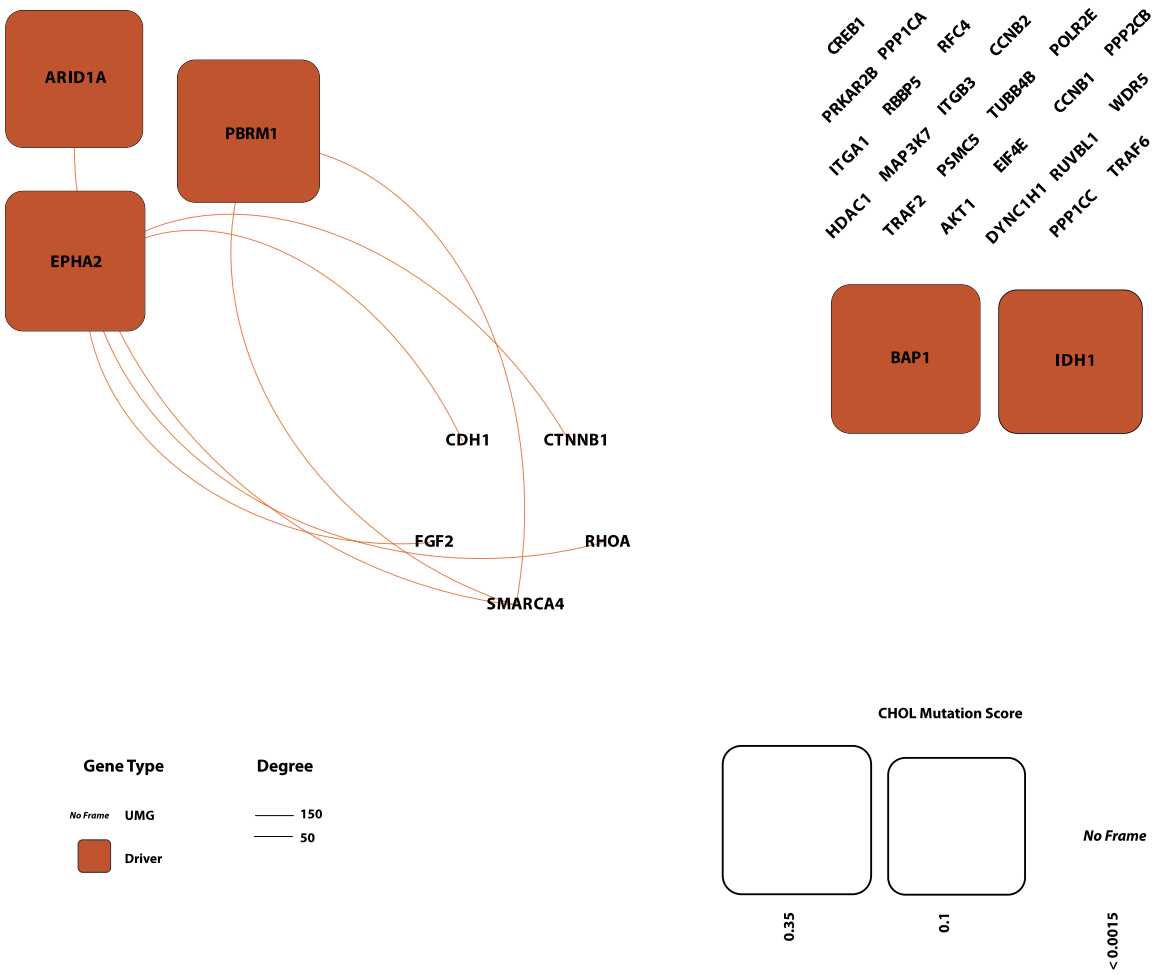
Supplementary Figure 2. *Impact on cancer cell line survival of UMG lists compared to other methods' lists, all including known driver genes. Extension to Figure 2.4.*



Supplementary Figure 3. *Impact on cancer cell line survival of UMG lists before the DepMap filtering step compared to other methods' lists, all including known driver genes. Extension to Figure 4.*

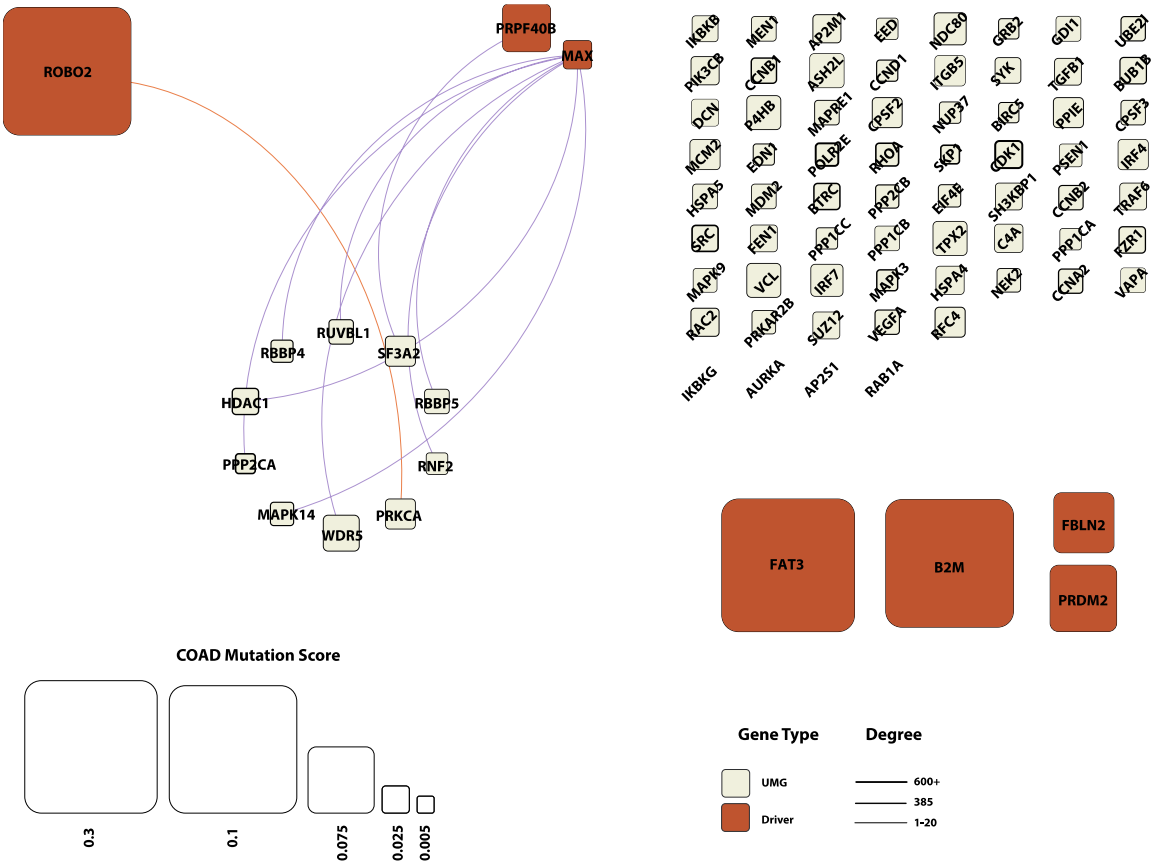


Supplementary Figure 4. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in CESC. Extension to Figure 2.5.*

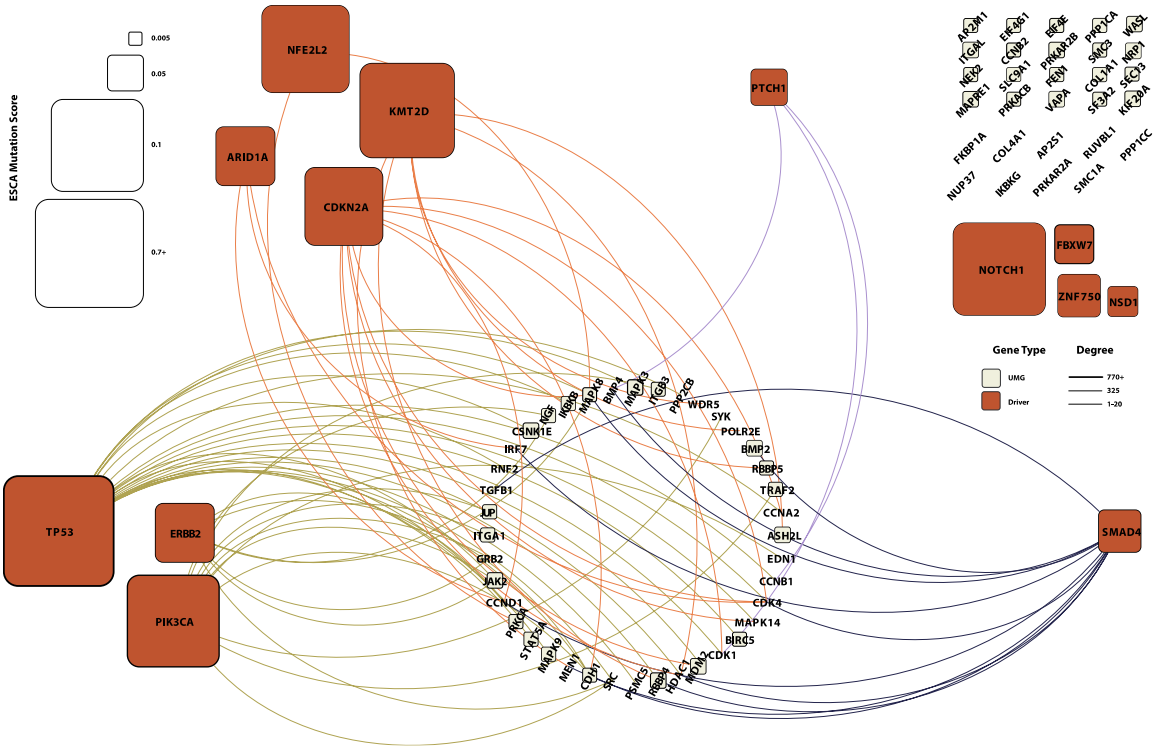


Supplementary Figure 5. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in CHOL. Extension to Figure 2.5.*

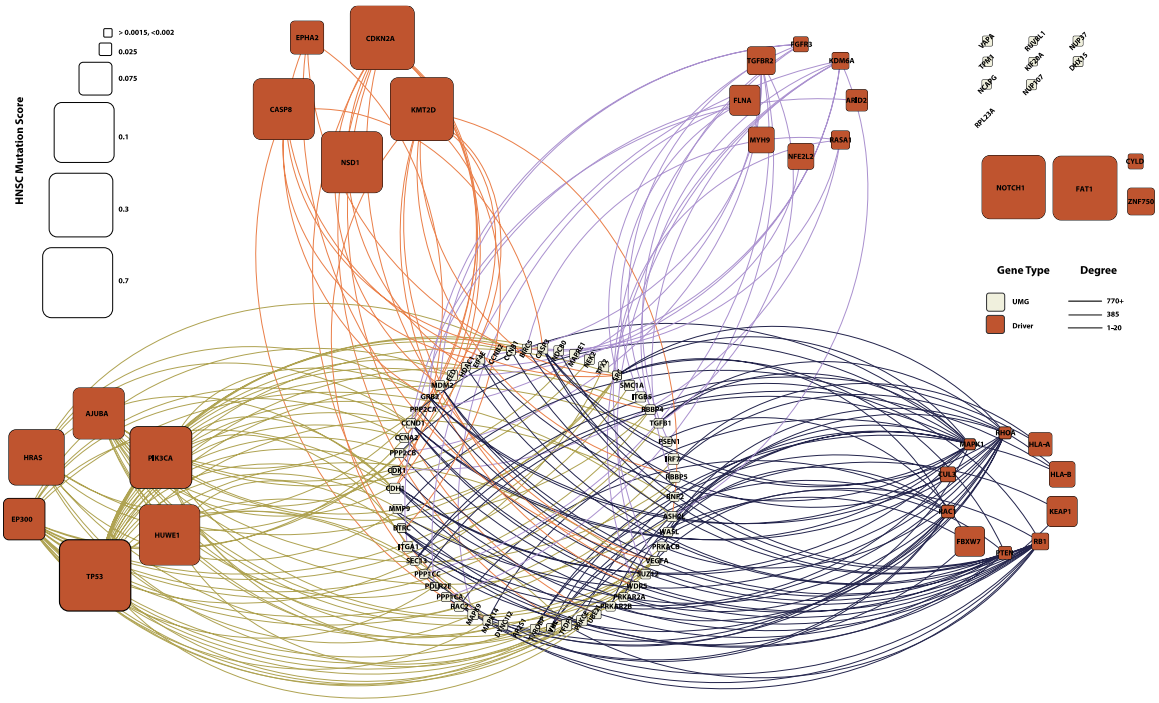
.



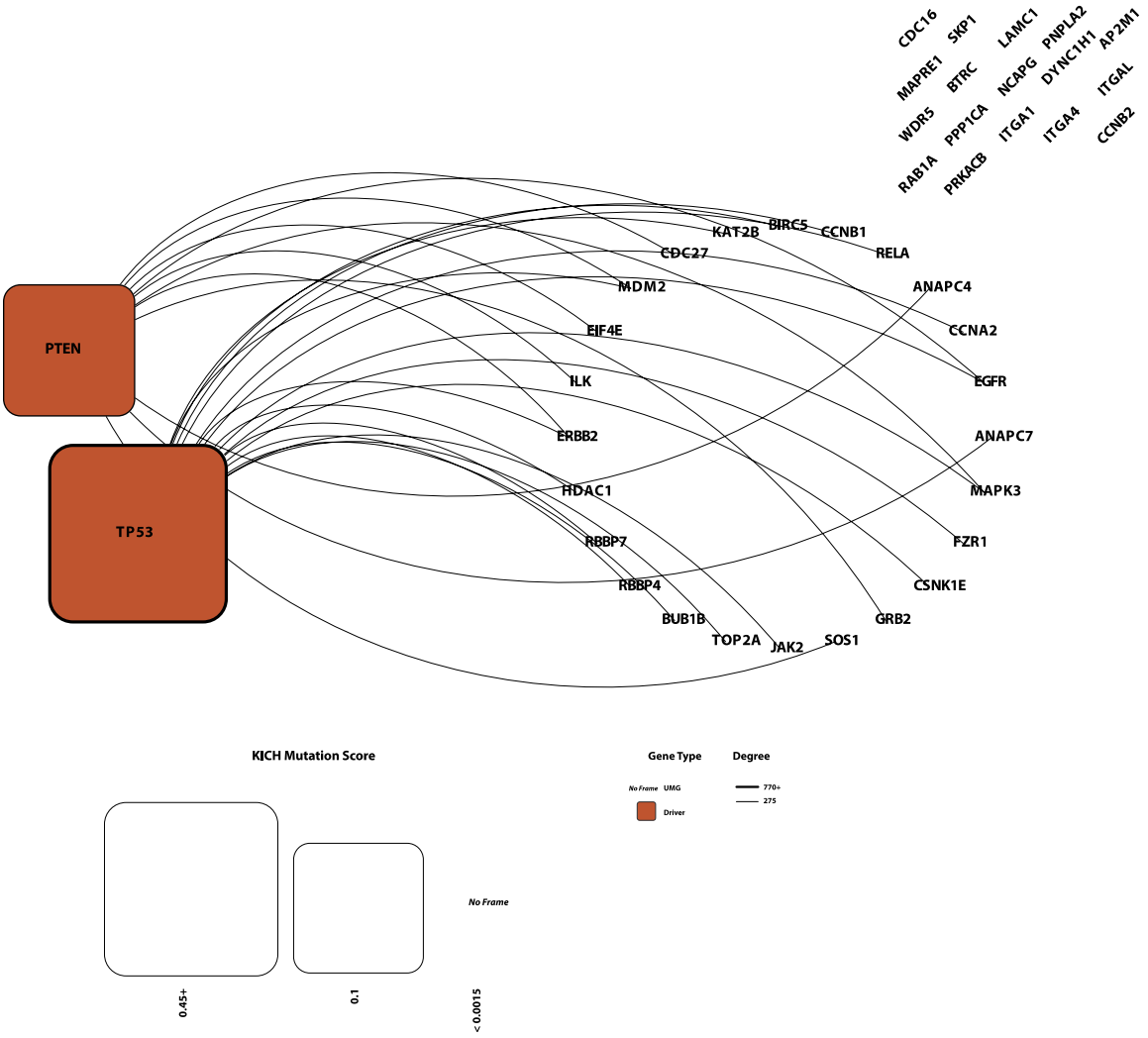
Supplementary Figure 6. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in COAD. Extension to Figure 2.5.*



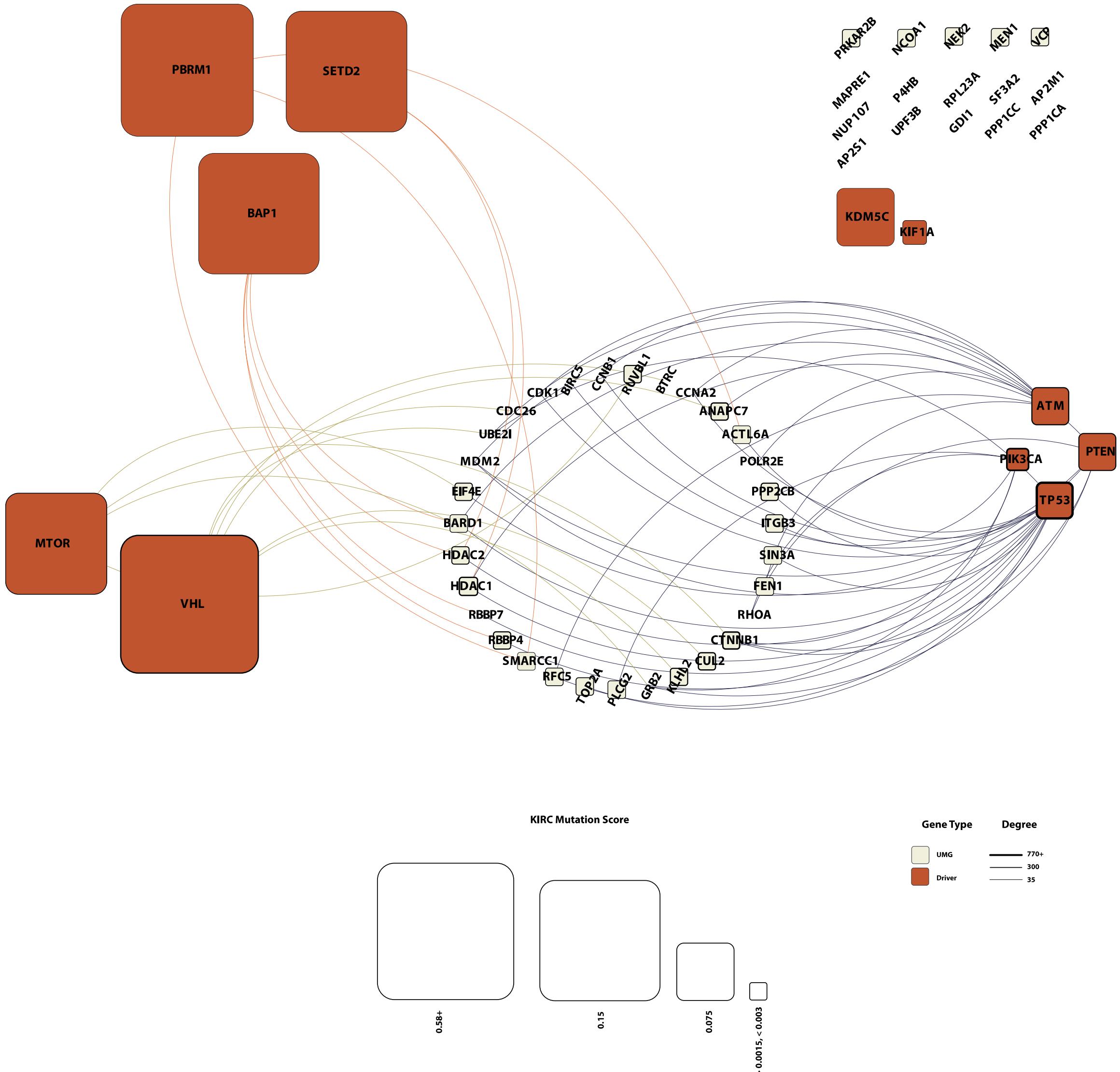
Supplementary Figure 7. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in ESCA. Extension to Figure 2.5.*



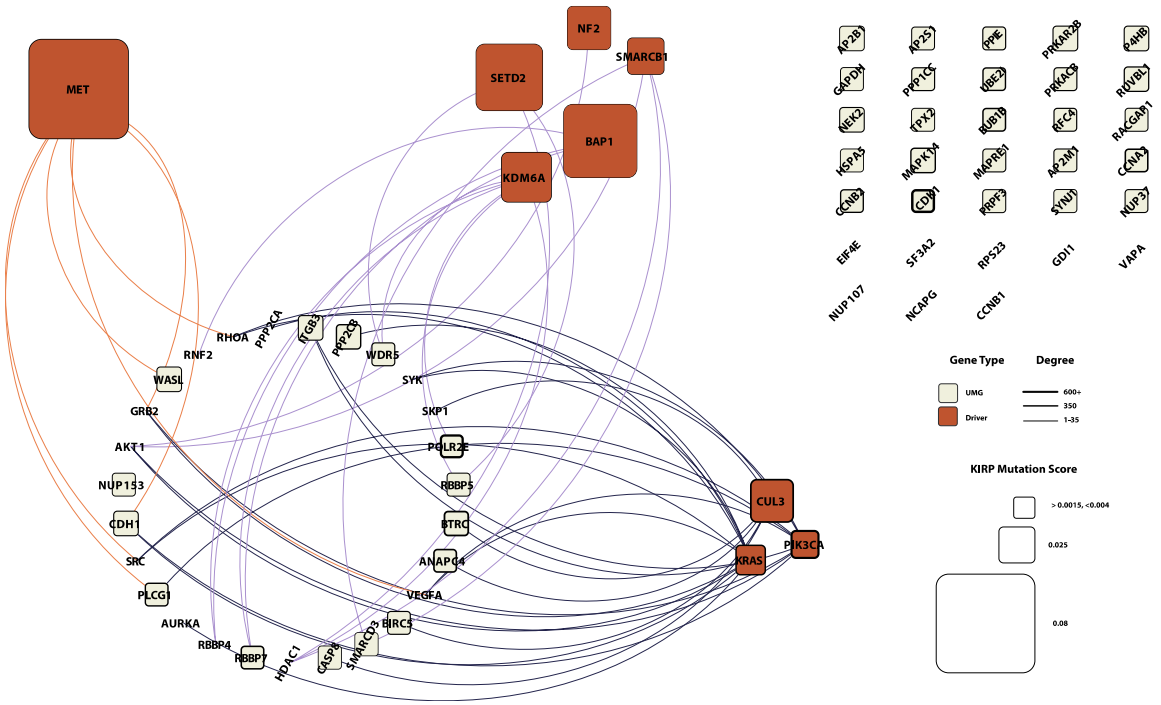
Supplementary Figure 8. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in HNSC. Extension to Figure 2.5.*



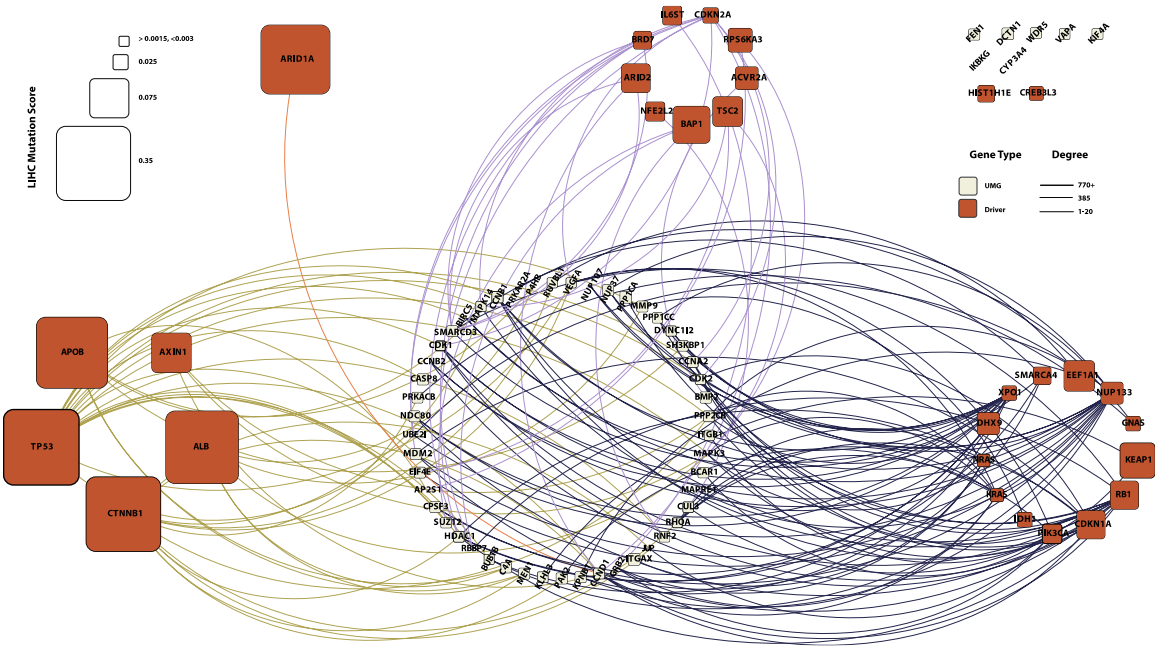
Supplementary Figure 9. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in KICH. Extension to Figure 2.5.*



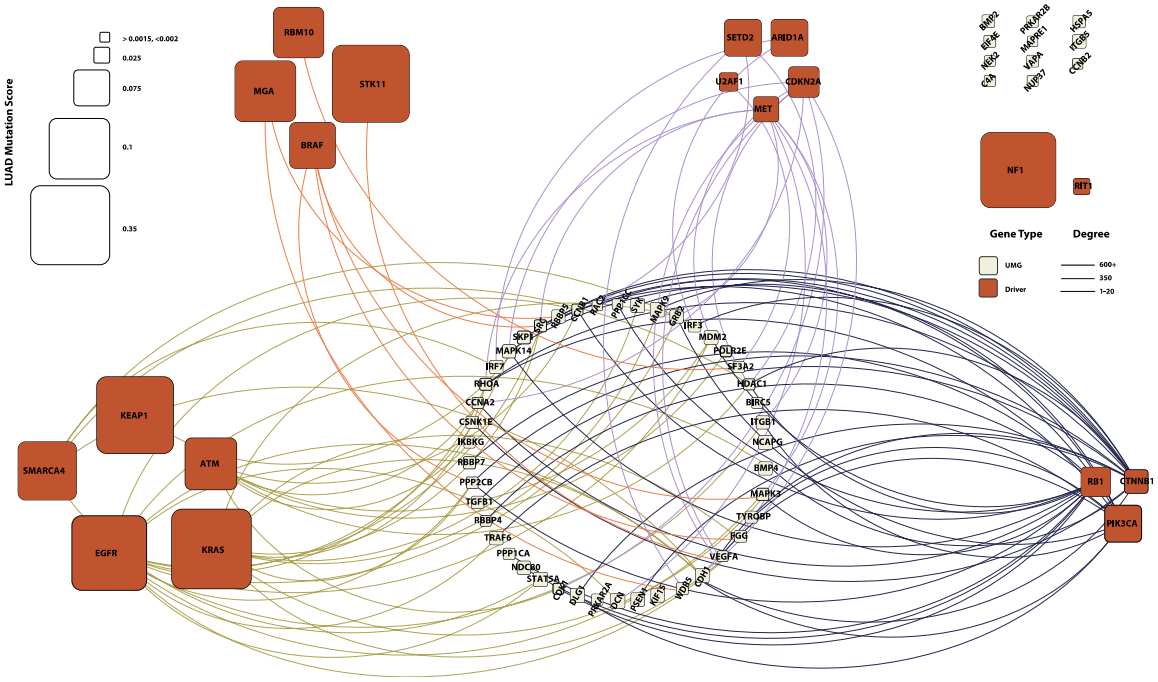
Supplementary Figure 10. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in KIRC. Extension to Figure 2.5.*



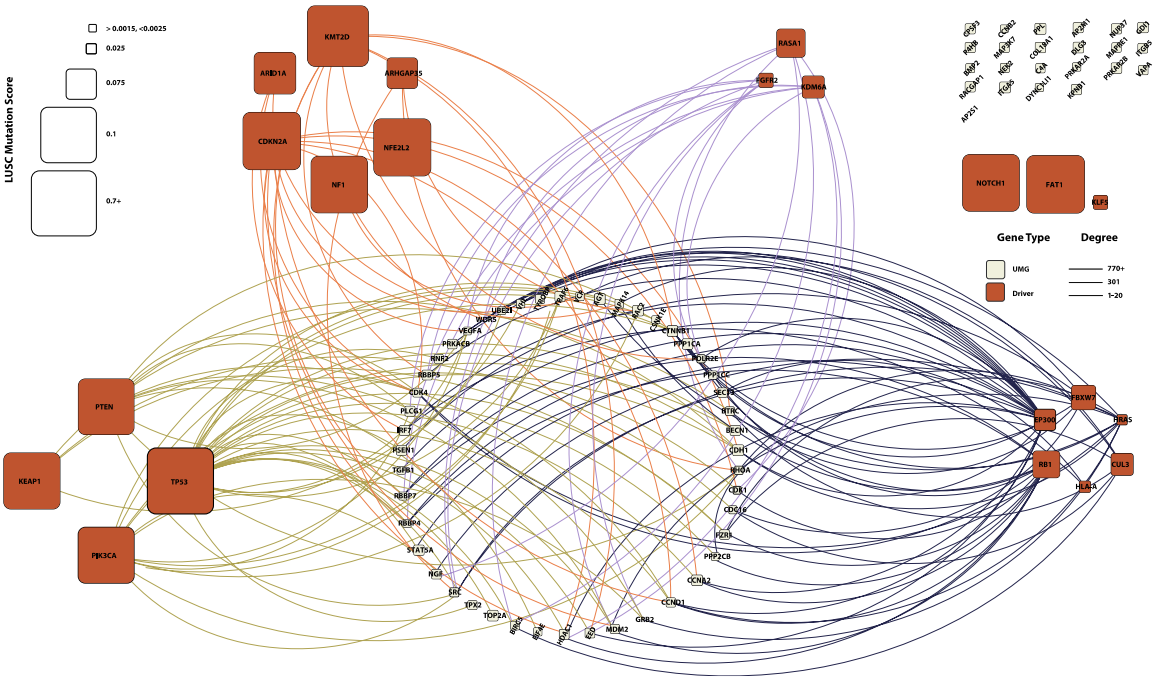
Supplementary Figure 11. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in KIRP. Extension to Figure 2.5.*



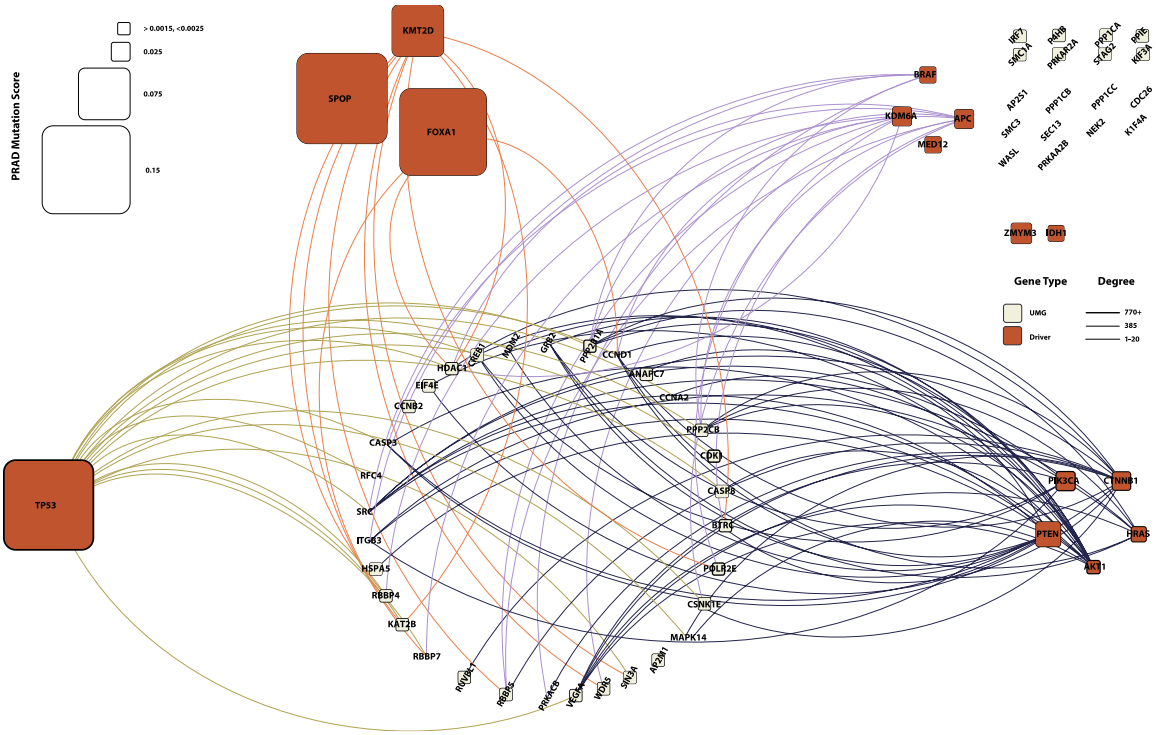
Supplementary Figure 12. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in LIHC. Extension to Figure 2.5.*



Supplementary Figure 13. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in LUAD. Extension to Figure 2.5.*

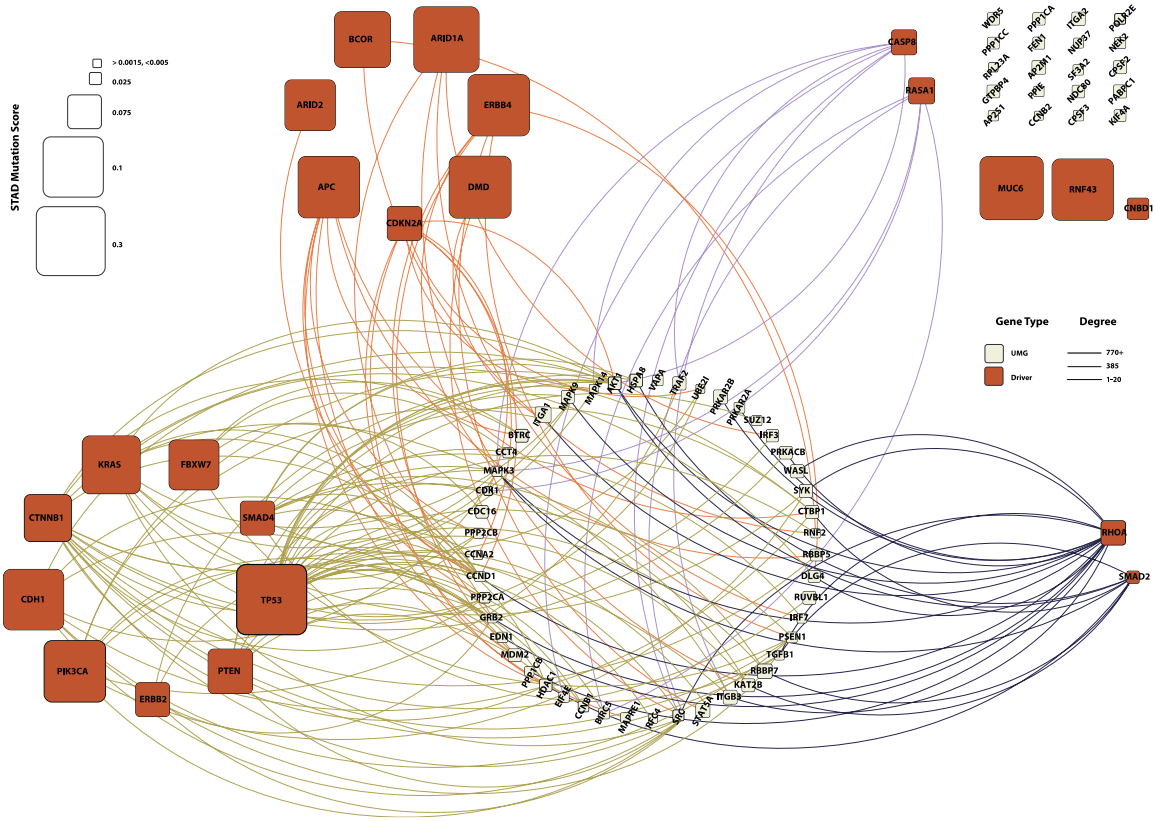


Supplementary Figure 14. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in LUSC. Extension to Figure 2.5.*

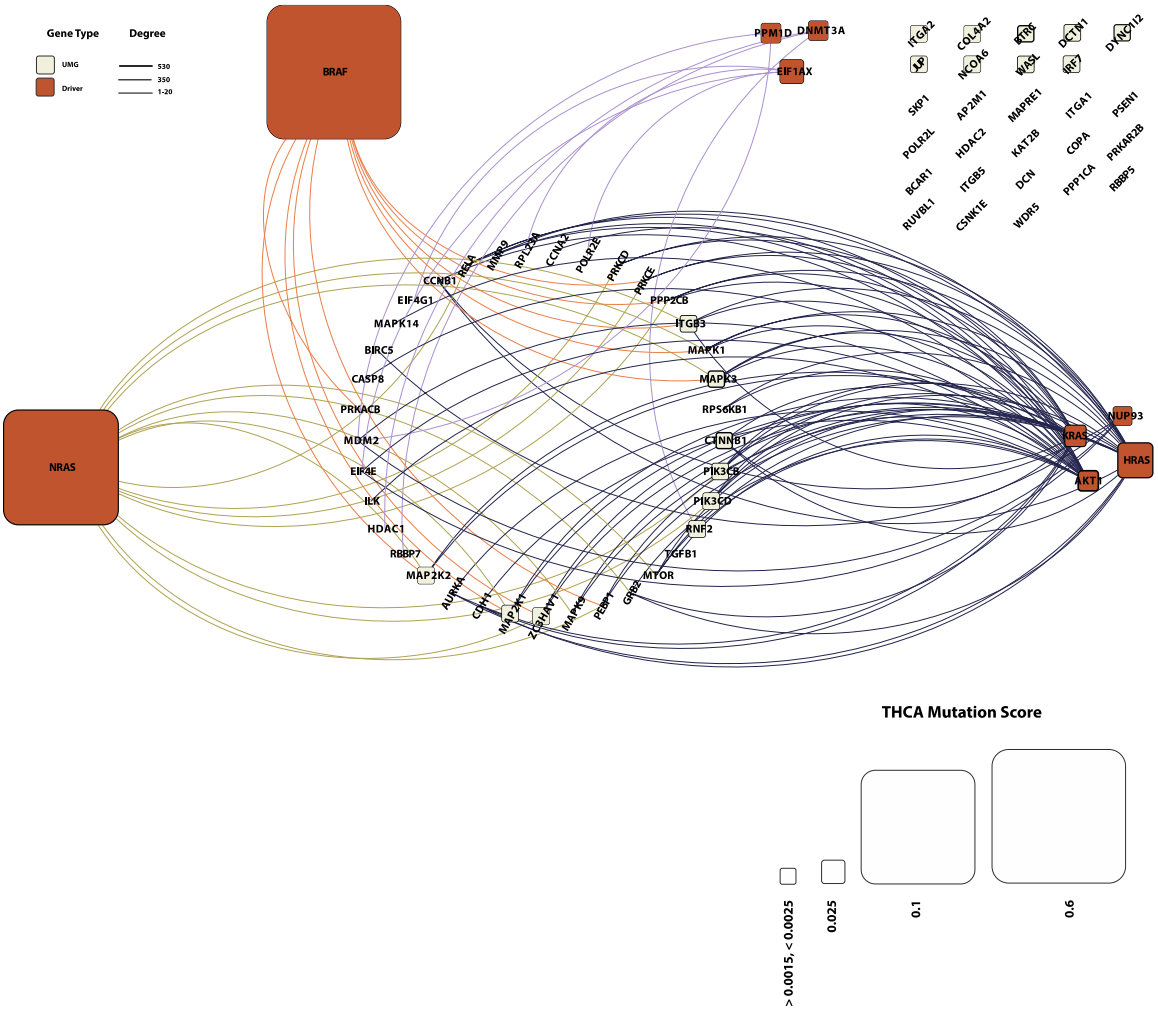


Supplementary Figure 15. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in PRAD. Extension to Figure 2.5.*

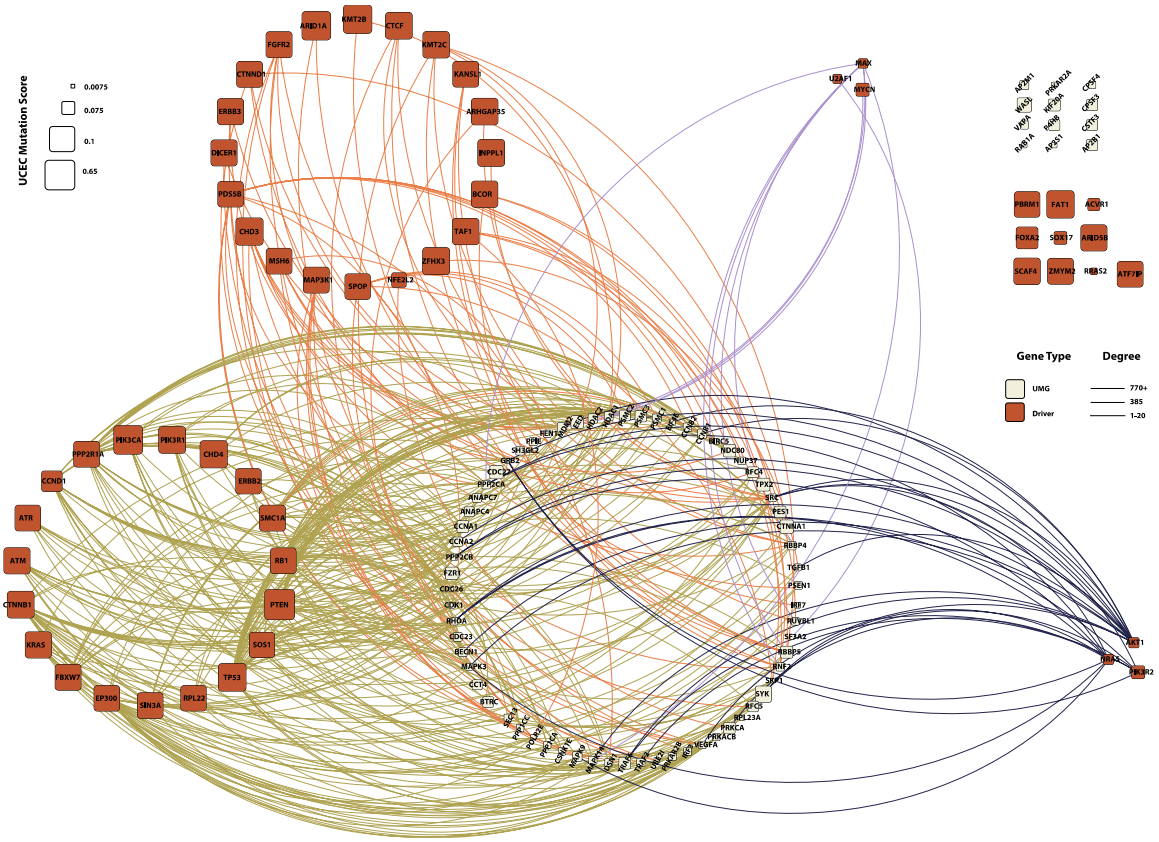
Supplementary Figure 16. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in READ. Extension to Figure 2.5.*



Supplementary Figure 17. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in STAD. Extension to Figure 2.5.*



Supplementary Figure 18. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in THCA. Extension to Figure 2.5.*



Supplementary Figure 19. *PPI network analysis of the relationships between UMGs (white nodes) and known driver genes (red) in UCEC. Extension to Figure 2.5.*