A Network Propagation Approach to Prioritize Long Tail Genes in Cancer

Hussein Mohsen[1,*], Vignesh Gunasekharan[2], Tao Qing[2], Sahand Negahban[3], Zoltan Szallasi[4], Lajos Pusztai[2,*], Mark B. Gerstein[1,5,6,3,*]

[1] Computational Biology & Bioinformatics Program, Yale University, New Haven, CT 06511, USA

[2] Breast Medical Oncology, Yale School of Medicine, New Haven, CT 06511, USA

[3] Department of Statistics & Data Science, Yale University, New Haven, CT 06511, USA

[4] Children's Hospital Informatics Program, Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA

[5] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA

[6] Department of Computer Science, Yale University, New Haven, CT 06511, USA

* Corresponding author

Abstract

Introduction. The diversity of genomic alterations in cancer pose 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 implication in cancer development. The study of these genes often requires integrative approaches with multiple types of biological data. Network propagation methods have demonstrated high efficacy in uncovering genomic patterns underlying cancer using biological interaction networks. Yet, the majority of these analyses have focused their assessment on detecting known cancer genes or identifying altered subnetworks. In this paper, we introduce a network propagation approach that focuses on long tail genes with potential functional impact on cancer development.

Results. 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” (UMGs, 42-81 genes per cancer type) and hypothesize that their significant rank improvement indicates functional importance. We validate UMGs’ role in cancer cell survival *in vitro* using genome-wide RNAi and CRISPR databases and report new cancer-pathway associations based on UMGs that were not previously identified using driver genes alone.

Conclusion. Our analysis extends the spectrum of cancer relevant genes and identifies novel potential therapeutic targets.

1. 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 [1-7]. 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 genes and gene products and can be useful to identify novel gene-disease associations or to prioritize genes for functional studies. In this paper, 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.* [8] 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 [9]. Other early works that use propagation account for network properties such as degree distributions [10] and topological similarity between genes [11-13] 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 [14]).

Köhler *et al.* [15] 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 filtering given the limitations in PPI data availability in 2008. Shortly afterwards, Vanunu *et al.* [16] 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 [17], to generate features used to train machine learning models that predict gene-disease associations in breast cancer, glioblastoma multiforme, and other cancer types [18, 19], or to suggest drug targets in acute myeloid leukemia by estimating gene knockout effects *in silico* [20].

Hofree *et al.* introduced network-based stratification (NBS) [21], 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 [22] 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 [23] and HotNet2 [24])

These studies have made significant strides in understanding several aspects of cancer, but they have faced limitations with respect to (i) relying on multiple data types that might not be readily available [17, 18], (ii) limited scope of biological analysis that often focused on a single cancer type [17, 20], (iii) suggesting too many [20] or too few [19] 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 [17, 22-24]. To parallel recent research on long tail genes and the functional importance of non-driver mutations [2, 4, 5, 25-29], we build on the well-established rigor of the propagation framework and introduce a computationally efficient approach that prioritizes rarely to moderately mutated genes implicated in cancer based on the observation that specific genes make significant upward jumps in their ranks during the propagation of mutation frequency scores over PPI networks. We call these genes upward mobility genes (UMGs).

Using somatic mutation data from the TCGA and two PPI networks with significant topological differences, STRING v11 [30] and HumanNet v2 [31]), we identify lists of UMGs in 17 cancer types, including BRCA, CESC, CHOL, COAD, ESCA, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, READ, STAD, THCA, and UCEC. In silico analysis demonstrates that

UMGs exert highly significant effect on cancer cell survival *in vitro* with cancer type specificity. Further, these genes considerably outperform genes suggested by state-of-the-art network methods with respect to this impact on cancer cell survival, and they could aid in refining the discovery of cancer type-specific previously overlooked regulatory pathways. Additional analysis of UMGs' positionality in a combined STRING-HumanNet v2 PPI network allows for classifying each UMG as a potential cancer driver, drug target, or both. Python implementation of our approach is available for the execution at cohort or single sample level.

2. Results

2.1 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 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 1b). Next, we perform hypergeometric test to assess the significance of propagation-based rankings by measuring the enrichment of known cancer genes above 90th percentile in post-propagation lists. Results demonstrate high statistical significance across all studied cohorts ($p < 10^{-8}$) in a validation of propagation as a tool to rank genes for potential 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 paper, genes whose rank significantly improves during propagation *and* land above the 90th percentile of post-propagation ranked lists are retained (Figure 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) [32]. This step eliminates 6-11% of UMGs (Figure 1d). We finally analyze the biological and topological properties of the shortlisted UMGs on pan-cancer and cancer type levels (Figure 1e).

2.2 UMGs across 17 cancer types

We report 237 UMGs across 17 cancer types. UMG lists capture the expected biological heterogeneity of cancer types: 83 genes (35%) are specific to one cancer type, 113 (47.7%) to 2-9 types, and only 41 (17.3%) to 10 or more types. The longest list of UMGs corresponds to LUSC (n = 80 genes) and the shortest to CHOL (n = 42). Hierarchical complete linkage clustering of cancer types (right of Figure 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 [7], subsets of gastrointenstinal (ESCA and COAD), gynecological (BRCA and UCEC), and squamous (ESCA, LUSC, and HNSC) cancers cluster together. Other clustering results match with the rates of driver mutations across cancer types in [7], particularly the closeness between (i) STAD, READ, and CESC, (ii) HNSC and COAD, (iii) and LIHC and ESCA, suggesting similarities between mutational patterns within driver and long tail genes. Interestingly, UMGs specific to a single cancer type (left of Figure 2) include a considerable number of genes with similar functions: *COL4A1* and *COL1A1* that encode different types of collagen (specific to ESCA), *RPS10* and *RPS23* that encode ribosomal proteins (specific to KIRP), and triplets of genes that encode proteins in the 26S proteasome complex (*PSMC1/2/3*, specific to UCEC), subunits of RNA polymerase (*POLR2C/G/H*, specific to CHOL), or proteins in the SNF/SWI chromatin remodeling complex (*SMARCB1/C1/E1*, specific to KIRC). The circos plot [33] of Figure 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 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 performed genome-wide loss of function screening for all known human genes using RNA interference (RNAi) and CRISPR to estimate tumor cell viability after gene silencing in hundreds of cancer cell lines [32]. 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 provide more accurate estimates of functional impact [34].

We found that on average, UMGs have a higher impact on the survival of cancer cell lines than that of all other genes. We conducted these comparisons in cell lines corresponding to each of the cancer types used for UMG discovery. In all 17 cancer types and in both CRISPR and RNAi experiments, knockout of UMGs yields a stronger negative *in vitro* effect on the survival of more cell lines than that of non-UMGs (Mann-Whitney U test, $p < 5 \times 10^{-3}$).

Next, we compared the functional impact of UMGs to that of genes selected by Hierarchical HotNet (HHotNet) in 3 different settings and nCOP [35], a non-prpagation method that recently demonstrated good performance in uncovering cancer-related genes. HHotNet reported statistically significant results after the integration over both PPIs in 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 of the PPIs. As HHotNet and nCOP do not solely focus on long tail genes, we removed known cancer-specific driver genes in our comparisons—although retaining them had no considerable effect on results.

In terms of mean DepMap scores, all methods' selected gene sets have an average negative impact on cell survival except for a small number of instances. Of the methods we tested for gene selection, the UMGs have the strongest negative impact on cancer cell survival across all cancer types in both CRISPR and RNAi experiments (Figure 3a). Similarly, the median percentage of cell lines negatively impacted by UMGs' knockout is higher than that for genes selected by other methods in 30 out of the 32 cancer type-assay combinations (Figure 3b), with the remaining two demonstrating ties (CRISPR-THCA and CRISPR-READ). 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 < -1.75. We note that similar results were obtained for these comparisons before the optional DepMap filtering step that only removed 6-11% of UMGs.

2.4 UMGs reveal known and novel cancer-pathway associations

Biological enrichment analysis of UMGs, separately and in combination with known drivers, confirms already known functional importance of some of the UMGs and suggests new associations between cancer types and biological pathway alterations. UMGs have statistically significant associations (Benjamini *p-adjusted* < 0.05) with most of the oncogenic pathways curated by Sanchez-Vega *et al.* (8 of 10) [36], alone and also when combined with cancer type - specific drivers (Figure 4a). These results indicate that UMGs are members of known biological pathways and can broaden the 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 [7]). Interestingly, the p53 pathway has only a small number of associations with UMGs in contrast to the many more associations we detected with the cell cycle, TGF-beta and Hippo signaling pathways. Other known cancer pathways are also altered by UMG including apoptosis, HIF-1 and mTOR. Notably, the number of cancer type-specific pathway associations does not correlate with the size of UMG lists. For example, CHOL, which has the smallest number of UMGs (n = 42 genes), has the largest set of pathway associations, while BRCA with a large UMG list has few pathway associations. These findings suggest greater diversity in altered biological processes that lead to development of CHOL compared to BRCA.

On the pancancer level, we partitioned enrichment results for all 237 UMGs into 9 clusters based on biological function (Figure 4b). Using EnrichmentMap (EM) [37], 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; see Methods). Connectivity patterns within the EM network provide insights into the sets of entities and UMGs. Within 6 of the 9 clusters, namely ones with known relation to cancer pathways, proliferation, adhesion, binding, immune response and transcription and translation, we identified biological entities with high connectivity (red labels, Figure 4b). These entities include oncogenic pathways such as PI3K-AKT and mTOR, and important biological processes including matrix cell adhesion and gene silencing. Underlying their high connectivity is a selected subset of UMGs with high frequency in their constituent edges. 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 the ability of long tail gene analysis to uncover disease-disease/infection associations (Figure 4c, Table 1). Observed associations include well-studied ones between multiple cancers and Hepatitis C [38], Type II Diabetes Mellitus [39, 40], and HTLV-I infection [41].

Functional Cluster	**Frequent UMGs**
Known Cancer-related	*AKT1, IKBKB, MAPK1, MAPK3, PIK3CB, PIK3R2*
Proliferation	*BUB1B, CDC23, CDC27*
Adhesion	*CDC42, EGFR, ITGA1, ITGA2, ITGA5, ITGB1, ITGB3, ITGB5, SRC, VCL*
Transcription and Translation	*POLR2C, POLR2E, POLR2G, POLR2H, POLR2I, POLR2L*
Immune System	*IKBKB, IKBKG, MAPK1, MAPK3, NFKB1, RELA*
Cancer Mega-pathways	*AKT1, CCND1, CDK4, EGFR, GRB2, IKBKB, IKBKG, MAPK1, MAPK3, MDM2, NFKB1, PIK3CB, PIK3R2, RAF1, RELA, SOS1*
Other Diseases and Infections Mega-pathways	*AKT1, IKBKB, IKBKG, IRF3, JAK2, MAPK1, MAPK14, MAPK3, MAPK8, MAPK9, NFKB1, PIK3CB, PIK3R2, RELA, TRAF6*

Table 1. Frequent UMGs within EnrichmentMap functional clusters

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

The aim behind identifying UMGs is to expand the narrative of known driver genes underlying cancer in line with many recent studies whose results defy the neutrality of long tail genes and passenger mutations in cancer development [2, 4, 5, 25-29]. In this section, we categorize each UMG as a potential "weak driver" that supplements known drivers during carcinogenesis, a potential drug target whose suppression kills cancer cells, or both, according to its positionality in PPI networks with respect 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 score is high. We study these properties for UMGs in a composite PPI network that merges signals from STRING and HumanNet v2. Figure 5 shows a representative network (BRCA). For convenience in visualization, we include immediate neighborhoods of each gene and UMG-driver edges only.

The first category of UMGs includes ones connected to high scoring drivers (Figure 5 left side, olive and orange edges). By virtue of sharing connections with these potent and frequently mutated drivers, this subset of UMGs likely includes cancer type-specific potential drug targets with little effect on carcinogenesis. This becomes even more relevant for UMGs 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 5 right side, dark blue and purple edges) constitute the second category and are likely to have a supplementary driving force. Interestingly, the third category includes an often small subset with nearly no observed mutations in the cohort (e.g. 5 genes in Figure 5: *NUP37*, *UBE21*, *POLR2E*, *IRF7*, and *EIF4E*). Such genes are likely to be drug targets or false positives limited by the size of the cohort under study. The fourth category includes UMGs with positive initial score and no connections to driver genes (Figure 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 5).

2.6 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 support the functional importance of UMGs, we manually cross referenced our UMG lists with publications and found that a large percentage of UMGs have been reported to play a role in cancer based on functional experiments. This percentage is as high as 88% of UMGs in cancer types like BRCA. Surprisingly, the same percentage drops to only 32% when we used CancerMine v24 to find literature-based associations [42]. 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.

3. Discussion

In this paper, we expand the set of propagation methods to parallel the growing interest in long tail genes in cancer. We introduce a computationally efficient approach based on the notion of upward mobility genes that attain significant improvements in mutation score-based ranking after propagating through PPI networks. By virtue of high post-propagation rank and significantly strong impact on cancer cell line survival, our approach prioritizes long tail genes across 17 cancer types. To reduce false positivity rate, it applies expression-based filtering on two major PPIs with different topologies and statistically validate rankings and cell survival impact.

Biological analysis of UMGs provides novel insights and demonstrates strong correlations with studies performed on known cancer drivers. Enrichment analysis results unlock a wide range of potential associations between key pathways and cancer types. A network-based analysis of enrichment results allows for classifying UMGs based on their centrality to biological functions, opening the door for a more informed drug targetability. Another network-based analysis categorizes each UMG as a "weak driver," cancer type-specific drug target. Manual curation of literature further validates UMGs' connection to cancer which could be overlooked by automated literature mining alone.

Results suggest that we have not reached a point of data saturation with respect to analyzing long tail genes yet. The generation of new datasets will likely improve results in rare cancer types such as cholangiocarcinoma (CHOL) and chromophobe renal cell carcinoma (KICH). Novel discoveries on frequently mutated genes in cancer, among which are many drivers, will likely reflect on the study of long tail genes as well. This was particularly evident in the PPI positionality analysis: with 3 or less drivers identified in READ and KICH by Bailey *et al.* [7], most of these cancer types' UMGs belong to the fourth category (non-zero mutation scores and no connections with drivers. Another example is CHOL, with its small cohort that brings most UMGs into the third category (no observed mutations). Finally, we note that bridging gaps across disciplines is 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.

4. Methods

4.1 Somatic mutation data

The results in this paper are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. Variants from the MC3 high quality somatic mutation dataset (n = 3.6 M) [43] are used to generate initial scores for each of the 17 cancer types. Sample-gene matrix for each cancer type includes mutation counts restricted to splicing and coding exonic variants based on RefSeq hg19 annotations by ANNOVAR 2018b [44]. Each count is normalized by gene length values provided by bioMart Bioconductor package [45]. Each non-zero value is then converted to a discrete number in {1, 2, 3, 4} based on its position with respect to 50th, 70th and 90th quantiles 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.

4.2 PPI networks

STRING v11 and HumanNet v2 functional network (FN) are downloaded from https://string-db.org/ and https://www.inetbio.org/humannet/, respectively. We perform edge filtering on both PPIs and retain edges with a confidence score equal to or higher than 0.7 in STRING 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 select the largest connected component in each network and perform gene expression-based filtering to generate cancer type-specific PPI networks.

4.3 Expression-based filtering

Gene expression filtering is performed on TCGA expression data corrected for study-specific biases and batch effects from RNASeqDB [46]. For each cancer type, genes with FPKM > 15 in > 20% of tumor samples are retained in the cancer type specific PPI network.

4.4 Propagation score calculation

To calculate propagation scores, we use an algorithm similar to the one introduced by Zhou *et al.* in [47]. 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 $W' = W . D^{-1}$, where D is the diagonal matrix of columns sums in W: $D = diag(\sum_{i=1}^{|G|} W_{ij}), j \in \{1, 2, \ldots, |G|\}$.

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 according to the following function:

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

where $S^{(0)} = M$ and $\alpha \in [0.5, 1[$. We use the Power Method [49] to iteratively propagate scores and converge when $\| S^{t+1} - S^{t} \| < \epsilon$.

4.5 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 $g_i \in G$,

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

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

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:

$$MS_i = RIS_i - RFS_i$$

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*.

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

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. 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, a threshold that has also been used in other studies [20].

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:

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

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 only 3-10 genes out of 45-90 genes per cancer type. Per (ii), integration of lists across *K* PPI networks yields the intersection of lists. In this paper, to increase confidence is selected genes, we integrate lists over cancer type-specific STRING and HumanNet v2 networks. Formally,

$$UMG_{Final} = UMG_{G_1} \cap UMG_{G_2} \cap \dots \, UMG_{G_K}$$

4.6 Ranking validation

To assess the validity of ranking after propagation we tested if known COSMIC genes are ranked higher than other genes using the hypergeometric statistical test [50] as earlier applied in [20]. Results show strong enrichment of COSMIC genes in the top 1000 genes across all cancer types and PPI networks ($p < 10^{-8}$ across all cancer types).

4.7 Driver and COSMIC genes

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

4.8 UMG *vs* non-UMG comparisons

In the first set of comparisons, Mann Whitney U one-sided test is used to compare the percentage distribution of negatively impacted cell lines by UMGs *vs* non-UMGs in each cancer type. Alternative hypothesis for each test is $H_1 = \psi(UMG)$ is shifted to the right of $\psi(\overline{UMG})$, where $\psi(X)$ is the percentage 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 DEMTER2 data with enhanced batch and off-target processing as described in [34].

4.9 UMGs *vs* non-driver 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 in for both PPI networks, so we exclude their HHotNet results. In all runs, we execute HHotNet in default settings with 1000 permutations using the second controlled randomization approach suggested in [22]. In nCOP, we use lists of rarely mutated genes reported in [35] (Figure 4) on the TCGA somatic mutational dataset in 15 of the 17 cancer types studied in our paper (all except CHOL and ESCA). As HHotNet and nCOP 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.

4.10 Enrichment analysis

Enrichment analysis to identify KEGG Pathways and GO molecular functions and biological processes is performed on DAVID v6.8 [51]. DAVID chart results with Benjamini *p*-adjusted < 0.05 are selected for analysis. Network visualization is executed using EnrichmentMap v3.0 on Cytoscape v3.8.2 [52]. Frequent terms highlighted in red in Figure 4b have ≥ 5 inter-cluster edges and those in Figure 4c ≥ 10 edges. Frequent UMGs in Table 1 are identified based on their presence edges between highlighted nodes according to the same thresholds (i.e. ≥ 5 and ≥ 10).

4.11 PPI analysis

Composite PPI is the union of high quality edges in STRINGv11 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 zero-fied to attain lower FPR. Visualization and degree calculation is executed using Cytoscape v3.8.2.

Figure Captions

Figure 1. Schematic overview of the Upwardly Mobile Gene identification strategy

Figure 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 the percentage of a cancer type’s cell lines whose survival is negatively 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 mean normalized ranking of each UMG in the lists it belongs to.

Figure 3. 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) and nCOP. Higher negative values indicate greater negative effect on cell survival after gene knockdown. (b) UMGs have a broader impact on the survival of cancer cell lines compared to that of genes selected by HHotNet and nCOP. The median percentage of cell lines negatively impacted by UMGs’ knockout is higher across cancer types.

Figure 4. 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 237 UMGs visualized using EnrichmentMap 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 driven by UMGs.

Figure 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."

References

1. Pon JR, Marra MA: **Driver and passenger mutations in cancer.** *Annu Rev Pathol* 2015, **10:**25-50.
2. 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.
3. Scholl C, Frohling S: **Exploiting rare driver mutations for precision cancer medicine.** *Curr Opin Genet Dev* 2019, **54:**1-6.
4. 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.
5. 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.
6. ICGC-TCGA PCAWG Consortium: **Pan-cancer analysis of whole genomes.** *Nature* 2020, **578:**82-93.
7. 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.
8. 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.

9. 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.

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

11. 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.

12. 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.

13. 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.

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

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

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

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

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

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

24. 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.

25. 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.

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

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

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

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

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

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

42. 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.

43. 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.

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

45. 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.

46. 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.

47. 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.

48. 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.

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

50. 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.

51. 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.

52. 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.

Figure 1

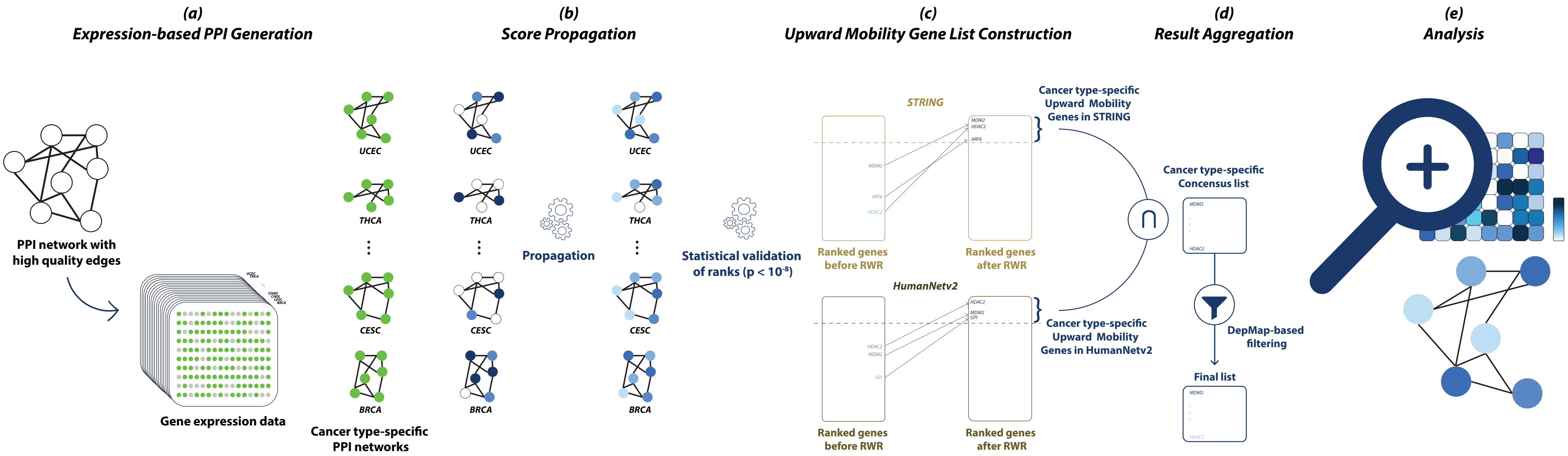

(a)
Expression-based PPI Generation
PPI network with
high quality edges
UCEC
THCA
...
COAD
CHOL
CESC
BRCA
Gene expression data
UCEC
THCA
⋮
CESC
BRCA
Cancer type-specific
PPI networks
(b)
Score Propagation
UCEC
THCA
⋮
CESC
BRCA
Propagation
UCEC
THCA
⋮
CESC
BRCA
Statistical validation
of ranks ($p < 10^{-8}$)
(c)
Upward Mobility Gene List Construction
STRING
MDM2
ARF6
HDAC2
MDM2
HDAC2
ARF6
Ranked genes
before RWR
Ranked genes
after RWR
Cancer type-specific
Upward Mobility
Genes in STRING
HumanNetv2
HDAC2
MDM2
GPI
HDAC2
MDM2
GPI
Ranked genes
before RWR
Ranked genes
after RWR
Cancer type-specific
Upward Mobility
Genes in HumanNetv2
∩
(d)
Result Aggregation
Cancer type-specific
Concensus list
MDM2
.
.
.
HDAC2
DepMap-based
filtering
Final list
MDM2
.
.
.
HDAC2
(e)
Analysis

Figure 2

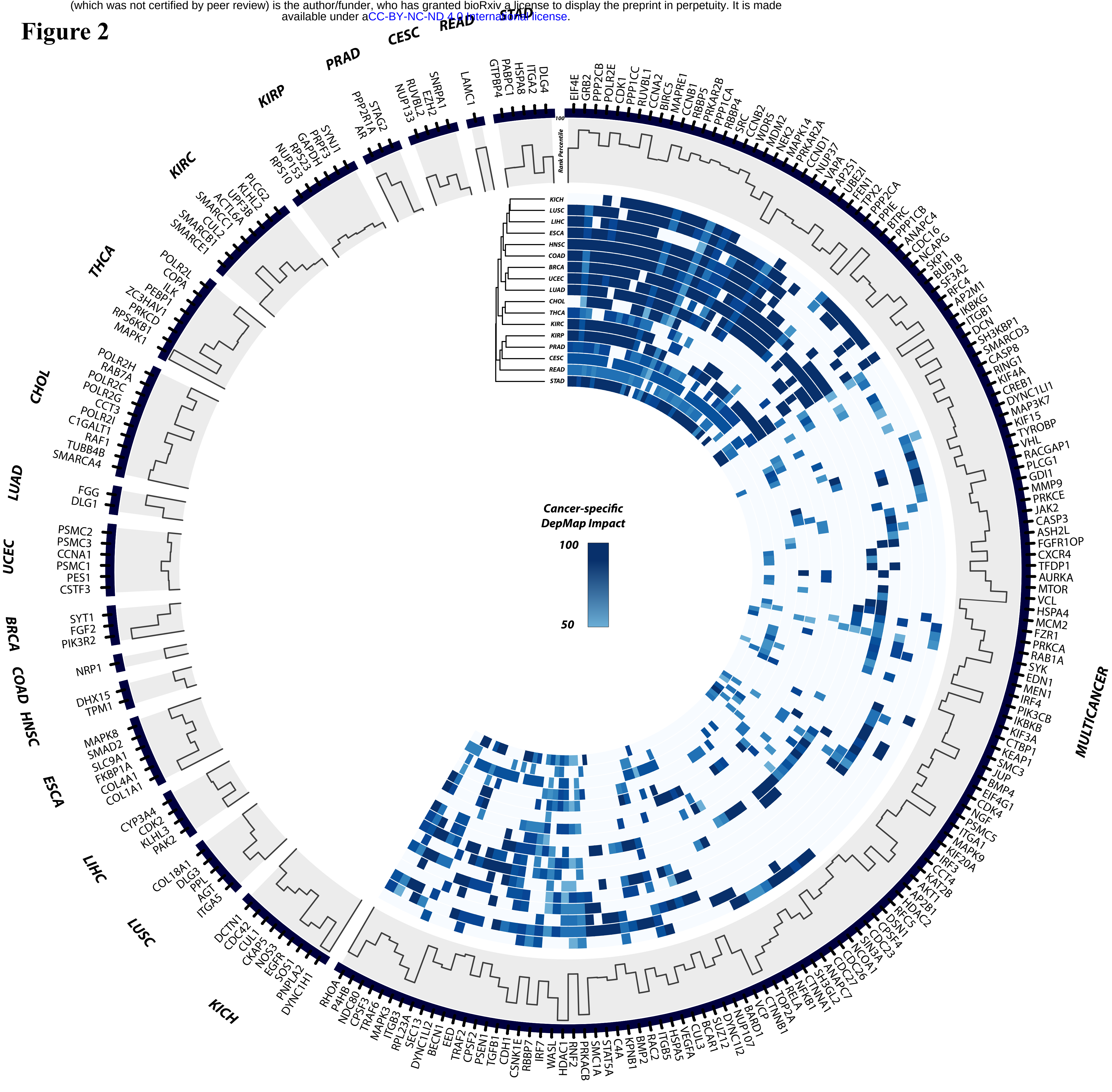

MULTICANCER
EIF4E
GRB2
PPP2CB
POLR2E
CDK1
PPP1CC
RUVBL1
CCNA2
BIRC5
MAPRE1
CCNB1
RBBP5
PRKAR2B
PPP1CA
RBBP4
SRC
CCNB2
WDR5
MDM2
NEK2
MAPK14
PRKAR2A
CCND1
NUP37
VAPA
AP2S1
UBE2I
FEN1
TPX2
PPP2CA
PPIE
BTRC
PPP1CB
ANAPC4
CDC16
NCAPG
SKP1
BUB1B
SF3A2
RFC4
AP2M1
IKBKG
ITGB1
DCN
SH3KBP1
SMARCD3
CASP8
RING1
KIF4A
CREB1
DYNC1LI1
MAP3K7
KIF15
TYROBP
VHL
RACGAP1
PLCG1
GDI1
MMP9
PRKCE
JAK2
CASP3
ASH2L
FGFR1OP
CXCR4
TFDP1
AURKA
MTOR
VCL
HSPA4
MCM2
FZR1
PRKCA
RAB1A
SYK
EDN1
MEN1
IRF4
PIK3CB
IKBKB
KIF3A
CTBP1
KEAP1
SMC3
JUP
BMP4
EIF4G1
CDK4
NGF
PSMC5
ITGA1
MAPK9
KIF20A
IRF3
CCT4
KAT2B
AKT1
AP2B1
HDAC2
RFC5
DSN1
CPSF4
CDC23
SIN3A
NCOA1
CDC26
CDC27
ANAPC7
SH3GL2
CTNNA1
NFKB1
RELA
TOP2A
CTNNB1
VCP
BARD1
NUP107
DYNC1I2
SUZ12
BCAR1
CUL3
VEGFA
HSPA5
ITGB5
RAC2
BMP2
KPNB1
C4A
STAT5A
SMC1A
PRKACB
RNF2
HDAC1
WASL
IRF7
RBBP7
CSNK1E
CDH1
TGFB1
PSEN1
CPSF2
TRAF2
EED
BECN1
DYNC1LI2
SEC13
RPL23A
ITGB3
MAPK3
TRAF6
CPSF3
NDC80
P4HB
RHOA
KICH
DYNC1H1
PNPLA2
SOS1
EGFR
NOS3
CKAP5
CUL1
CDC42
DCTN1
LUSC
ITGA5
AGT
PPL
DLG3
COL18A1
LIHC
PAK2
KLHL3
CDK2
CYP3A4
ESCA
COL1A1
COL4A1
FKBP1A
SLC9A1
SMAD2
MAPK8
HNSC
TPM1
DHX15
COAD
NRP1
BRCA
PIK3R2
FGF2
SYT1
UCEC
CSTF3
PES1
PSMC1
CCNA1
PSMC3
PSMC2
LUAD
DLG1
FGG
CHOL
SMARCA4
TUBB4B
RAF1
C1GALT1
POLR2I
CCT3
POLR2G
POLR2C
RAB7A
POLR2H
THCA
MAPK1
RPS6KB1
PRKCD
ZC3HAV1
PEBP1
ILK
COPA
POLR2L
KIRC
SMARCE1
SMARCB1
CUL2
SMARCC1
ACTL6A
UPF3B
KLHL2
PLCG2
KIRP
RPS10
NUP153
RPS23
GAPDH
PRPF3
SYNJ1
PRAD
AR
PPP2R1A
STAG2
CESC
NUP133
RUVBL2
EZH2
SNRPA1
READ
LAMC1
STAD
GTPBP4
PABPC1
HSPA8
ITGA2
DLG4
100
Rank Percentile
KICH
LUSC
LIHC
ESCA
HNSC
COAD
BRCA
UCEC
LUAD
CHOL
THCA
KIRC
KIRP
PRAD
CESC
READ
STAD
Cancer-specific
DepMap Impact
100
50

Figure 3

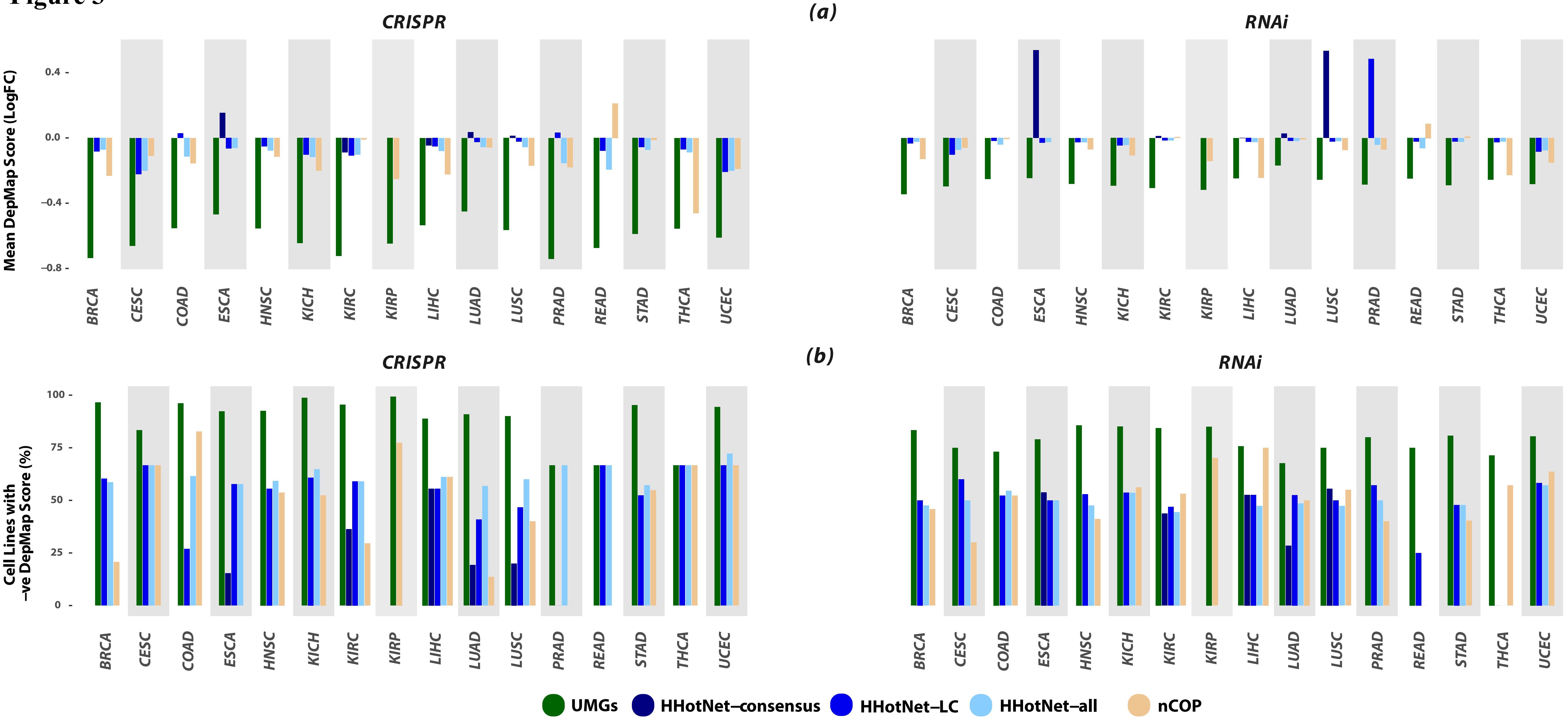

(a)
CRISPR
RNAi
Mean DepMap Score (LogFC)
0.4
0.0
−0.4
−0.8
BRCA
CESC
COAD
ESCA
HNSC
KICH
KIRC
KIRP
LIHC
LUAD
LUSC
PRAD
READ
STAD
THCA
UCEC
(b)
CRISPR
RNAi
Cell Lines with
−ve DepMap Score (%)
100
75
50
25
0
UMGs
HHotNet−consensus
HHotNet−LC
HHotNet−all
nCOP

Figure 4

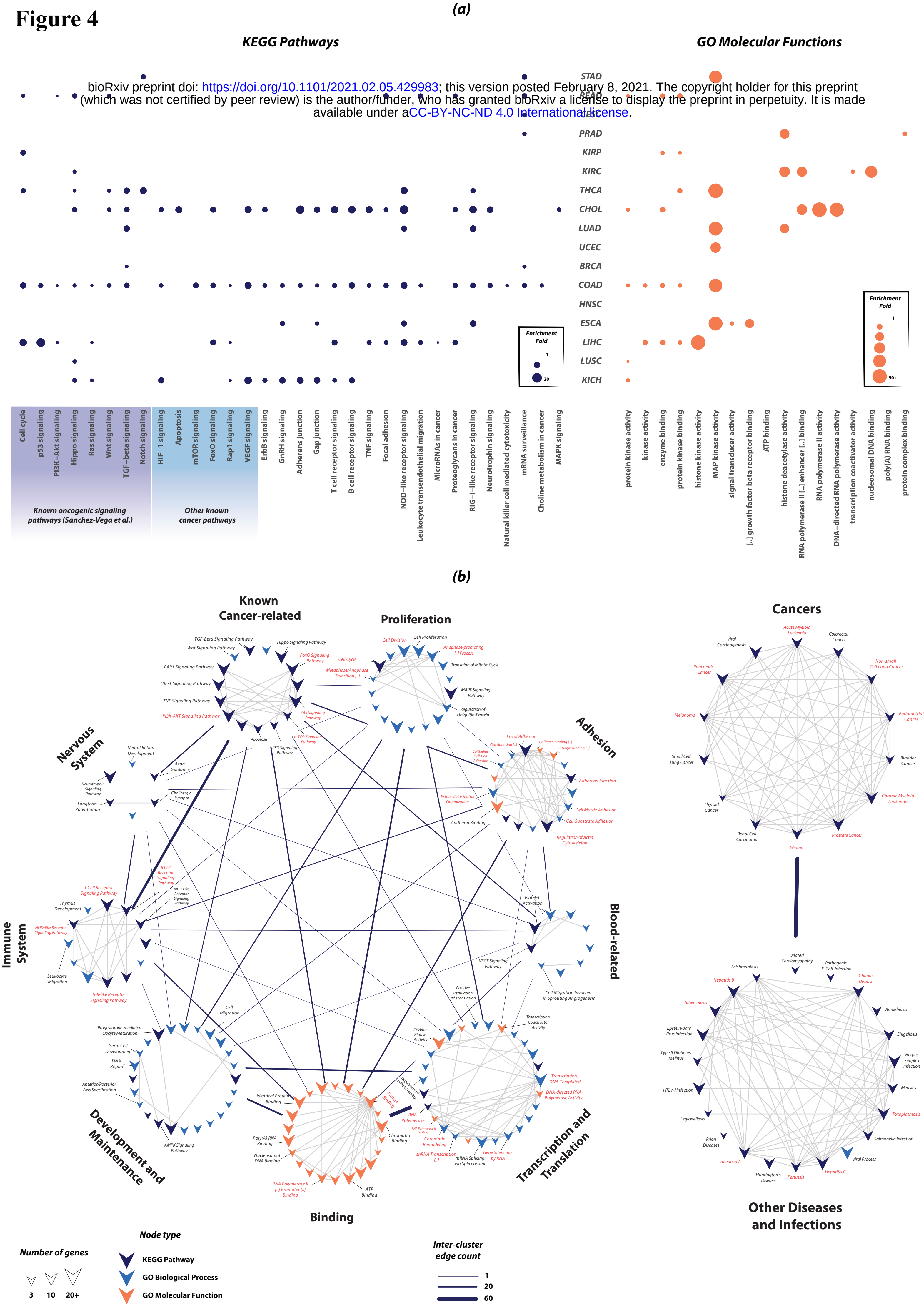
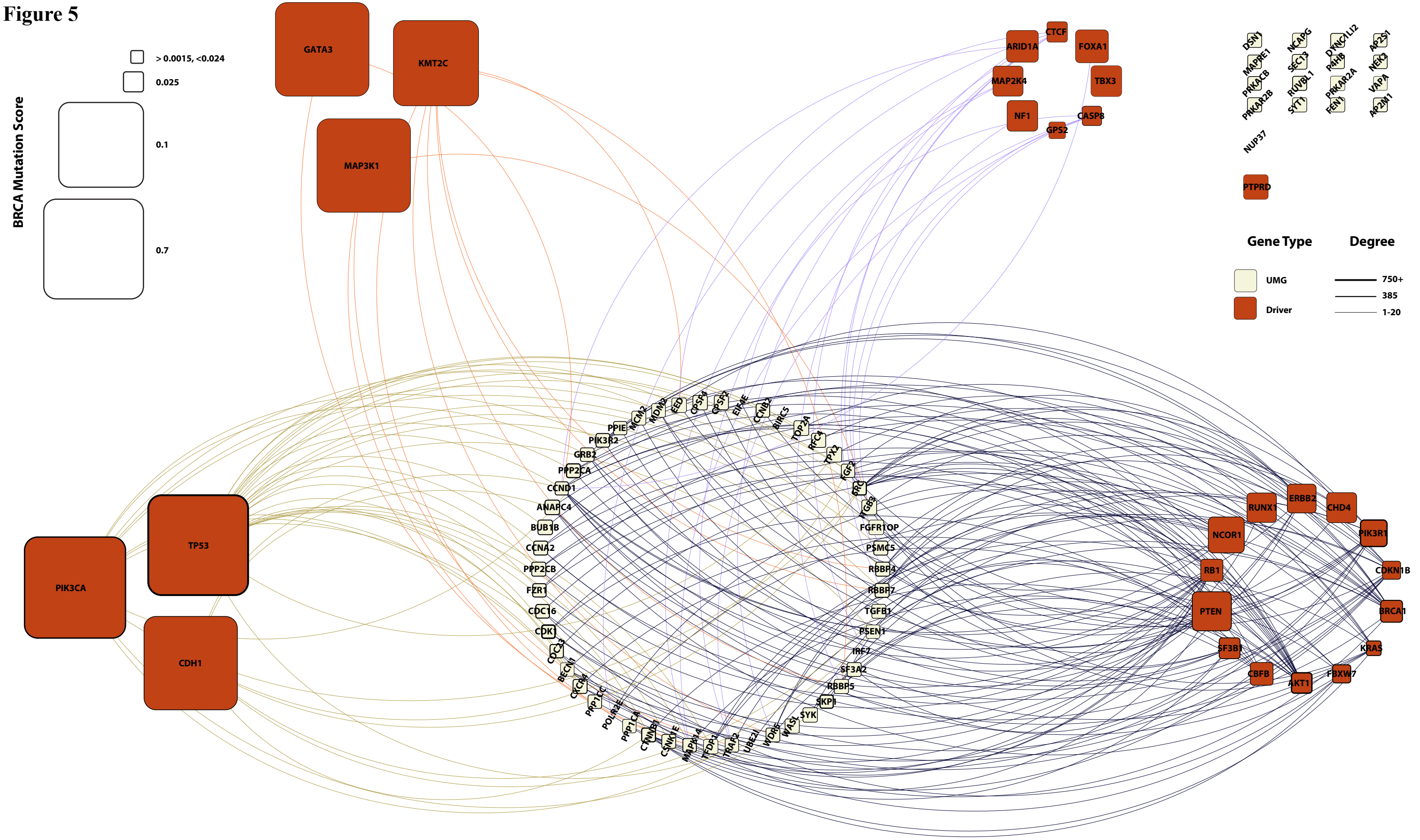

(a)
KEGG Pathways
GO Molecular Functions
STAD
READ
COAD
PRAD
KIRP
KIRC
THCA
CHOL
LUAD
UCEC
BRCA
COAD
HNSC
ESCA
LIHC
LUSC
KICH
Enrichment Fold
1
20
50+
Cell cycle
p53 signaling
PI3K–Akt signaling
Hippo signaling
Ras signaling
Wnt signaling
TGF–beta signaling
Notch signaling
HIF–1 signaling
Apoptosis
mTOR signaling
FoxO signaling
Rap1 signaling
VEGF signaling
ErbB signaling
GnRH signaling
Adherens junction
Gap junction
T cell receptor signaling
B cell receptor signaling
TNF signaling
Focal adhesion
NOD–like receptor signaling
Leukocyte transendothelial migration
MicroRNAs in cancer
Proteoglycans in cancer
RIG–I–like receptor signaling
Neurotrophin signaling
Natural killer cell mediated cytotoxicity
mRNA surveillance
Choline metabolism in cancer
MAPK signaling
Known oncogenic signaling pathways (Sanchez-Vega et al.)
Other known cancer pathways
protein kinase activity
kinase activity
enzyme binding
protein kinase binding
histone kinase activity
MAP kinase activity
signal transducer activity
[..] growth factor beta receptor binding
ATP binding
histone deacetylase activity
RNA polymerase II [..] enhancer [..] binding
RNA polymerase II activity
DNA–directed RNA polymerase activity
transcription coactivator activity
nucleosomal DNA binding
poly(A) RNA binding
protein complex binding
(b)
Known Cancer-related
Proliferation
Cancers
Adhesion
Nervous System
Blood-related
Immune System
Development and Maintenance
Binding
Transcription and Translation
Other Diseases and Infections
TGF-Beta Signaling Pathway
Wnt Signaling Pathway
Hippo Signaling Pathway
FoxO Signaling Pathway
Cell Cycle
RAP1 Signaling Pathway
HIF-1 Signaling Pathway
TNF Signaling Pathway
PI3K-AKT Signaling Pathway
RAS Signaling Pathway
mTOR Signaling Pathway
Apoptosis
P53 Signaling Pathway
Cell Division
Cell Proliferation
Anaphase-promoting [..] Process
Metaphase/Anaphase Transition [..]
Transition of Mitotic Cycle
MAPK Signaling Pathway
Regulation of Ubiquitin-Protein
Focal Adhesion
Cell Adhesion [..]
Collagen Binding [..]
Integrin Binding [..]
Epithelial Cell-Cell Adhesion
Adherens Junction
Extracellular Matrix Organization
Cell Matrix Adhesion
Cell-Substrate Adhesion
Cadherin Binding
Regulation of Actin Cytoskeleton
Neural Retina Development
Axon Guidance
Neurotrophin Signaling Pathway
Cholinergic Synapse
Longterm Potentiation
B Cell Receptor Signaling Pathway
RIG-I-Like Receptor Signaling Pathway
T Cell Receptor Signaling Pathway
Thymus Development
NOD-like Receptor Signaling Pathway
Leukocyte Migration
Toll-like Receptor Signaling Pathway
Platelet Activation
VEGF Signaling Pathway
Positive Regulation of Translation
Cell Migration Involved in Sprouting Angiogenesis
Transcription Coactivator Activity
Protein Kinase Activity
Cell Migration
Progesterone-mediated Oocyte Maturation
Germ Cell Development
DNA Repair
Anterior/Posterior Axis Specification
AMPK Signaling Pathway
Identical Protein Binding
Poly(A) RNA Binding
Nucleosomal DNA Binding
RNA Polymerase II [..] Promoter [..] Binding
ATP Binding
Regulation of mRNA Stability
Protein Binding
RNA Polymerase
RNA Polymerase II Activity
Chromatin Binding
Chromatin Remodeling
snRNA Transcription [..]
mRNA Splicing, via Spliceosome
Gene Silencing by RNA
Transcription, DNA-Templated
DNA-directed RNA Polymerase Activity
Acute Myeloid Leukemia
Viral Carcinogenesis
Colorectal Cancer
Pancreatic Cancer
Non-small Cell Lung Cancer
Melanoma
Endometrial Cancer
Small Cell Lung Cancer
Bladder Cancer
Thyroid Cancer
Chronic Myeloid Leukemia
Renal Cell Carcinoma
Prostate Cancer
Glioma
Dilated Cardiomyopathy
Leishmaniasis
Pathogenic E. Coli Infection
Hepatitis B
Chagas Disease
Tuberculosis
Amoebiasis
Epstein-Barr Virus Infection
Shigellosis
Type II Diabetes Mellitus
Herpes Simplex Infection
HTLV-I Infection
Measles
Legionellosis
Toxoplasmosis
Prion Diseases
Salmonella Infection
Influenza A
Viral Process
Huntington's Disease
Pertussis
Hepatitis C
Node type
KEGG Pathway
GO Biological Process
GO Molecular Function
Number of genes
3
10
20+
Inter-cluster edge count
1
20
60

Figure 5



BRCA Mutation Score
> 0.0015, <0.024
0.025
0.1
0.7
GATA3
KMT2C
MAP3K1
ARID1A
CTCF
FOXA1
MAP2K4
TBX3
NF1
CASP8
GPS2
DSN1
NCAPG
DYNC1LI2
AP2S1
MAPRE1
SEC13
P4HB
NEK2
PRKACB
RUVBL1
PRKAR2A
VAPA
PRKAR2B
SYT1
FEN1
AP2M1
NUP37
PTPRD
Gene Type
Degree
UMG
Driver
750+
385
1-20
PIK3CA
TP53
CDH1
PPIE
MCM2
MDM2
EED
CPSF4
CPSF2
EIF4E
CCNB2
BIRC5
TOP2A
RFC4
TPX2
FGF2
SRC
ITGB3
FGFR1OP
PSMC5
RBBP4
RBBP7
TGFB1
PSEN1
IRF7
SF3A2
RBBP5
SKP1
SYK
WASL
WDR5
UBE2I
TRAF2
TFDP1
MAPK14
CSNK1E
CTNNB1
PPP1CA
POLR2E
PPP1CC
CXCR4
BECN1
CDC23
CDK1
CDC16
FZR1
PPP2CB
CCNA2
BUB1B
ANAPC4
CCND1
PPP2CA
GRB2
PIK3R2
RUNX1
ERBB2
CHD4
NCOR1
PIK3R1
RB1
CDKN1B
PTEN
BRCA1
SF3B1
KRAS
CBFB
AKT1
FBXW7