Supporting Information

Vinayagam et al. 10.1073/pnas.1603992113

SI Text

Directed Human PPI Network. The directed human PPI network was compiled from our previous study (13). Briefly, a Naïve Bayesian classifier was applied to predict potential direction of signal flow between the ith and jth interacting proteins p_i and p_j as $p_i \rightarrow p_j, p_j \rightarrow p_i$ or both. The classifier uses features derived from the shortest PPI paths between membrane receptors and transcription factors and assigns confidence for each predicted edge directions ranging from 0.5 to 1. The weighted and directed edges are then encoded in an $N \times N$ matrix, and A is denoted as the weighted adjacency matrix of the directed graph for the PPI network. The element of A in the ith row and jth column is denoted as a_{ij} and is defined as follows: a_{ij} is in the range [0.5,1] if there is signal flow from protein p_j to p_i otherwise $a_{ij} = 0$.

Controllability Analysis and Node Classification. Recently, we developed a mathematical framework and analytical tools to identify MDS, with size denoted as N_D, whose control is sufficient to ensure the structural controllability of linear dynamics (14) and local structural controllability for nonlinear dynamics (*SI Text*, *Local Structural Controllability*) on any directed weighted network. This is achieved by mapping the structural controllability problem in control theory to the maximum matching problem in graph theory, which can be solved in polynomial time (15). Here, an edge subset M in a directed network or digraph is called a "matching" if no two edges in M share a common starting node or a common ending node. A node is "matched" if it is an ending node of an edge in the matching. Otherwise, it is "unmatched." A matching of maximum cardinality is called a "maximum matching." (In general, there could be many different maximum matchings for a given digraph.) We proved that the unmatched nodes that correspond to any maximum matching can be chosen as driver nodes to control the whole network. Identifying a minimum set of driver nodes is equivalent to choosing an input matrix (often denoted as B) with the minimum number of columns (see *SI Text*, *Local Structural Controllability* and ref. 14 for more details). The detailed construction of the input matrix B is not necessary for the identification of driver nodes. This is only mentioned to connect the notion of a driver node to the theoretical discussions in *SI Text*, *Local Structural Controllability*.

After a node is removed, denote the minimum number of driver nodes of the damaged network as N_D'. In this work, we classified nodes into three categories. (*i*) A node is indispensable if in its absence we have to control more driver nodes (i.e., $N_D' > N_D$). For example, removal of any node in the middle of a directed path will break the path and cause the N_D increase. Hence, all but the start and end nodes of a directed path are indispensable. (*ii*) A node is dispensable if in its absence we have $N_D' < N_D$. For example, removal of one leaf node in a star will decrease N_D by 1. (*iii*) A node is neutral if in its absence $N_D' = N_D$. For example, removal of the central hub in a star will not change N_D at all.

Note that a driver node in any MDS can never be an indispensable node. This can be proven by contradiction. Assume a driver node i is indispensable. According to the minimum input theorem (9), driver nodes are just unmatched nodes with respect to a particular maximum matching. There are two cases; case 1: the driver node i has no downstream neighbors [i.e., $k_{\mathrm{out}}(i) = 0$], then in its absence, $N_D' = N_D - 1$; and case 2: the driver node i has at least one downstream neighbors [i.e., $k_{\mathrm{out}}(i) > 0$]. There are two subcases; case 2.1: if in the maximum matching, one of node i's downstream neighbors (node j) is matched by node i, then in the absence of node i, node j will become unmatched (i.e., a new driver node), rendering $N_D' = N_D$; case 2.2: if none of node i's downstream neighbors are matched by node i, then in the absence of node i, $N_D' = N_D - 1$. In all of the cases, we do not have $N_D' > N_D$, which is in contrast to the definition of indispensable nodes. Hence, driver nodes cannot be indispensable.

Enrichment Analysis. To estimate the significance of overlap between a given node type S and given dataset D, we compute an enrichment z score as

$$z \text{ score} = \frac{(S_D - \text{mean of } R_D)}{\text{SD of } R_D},$$

where S_D is number of proteins from dataset D overlapping with node type S and R_D is the number of proteins from dataset D overlapping with random set of proteins of same size as N. Mean and SD of R_D is computed from 1,000 simulations of random sets. Note that the entire network with 6,339 proteins is used as the background for random sampling. In addition to the z score, we also computed the P value (two-tailed) by comparing the S_D with R_D distribution (modeled as Gaussian distribution). In the case of degree- or literature-controlled random sets, the random sets are sampled such that the average degree or average PubMed records of random sets matches the average of node type S.

Datasets Used for Enrichment Analysis. All of the datasets used for the enrichment analysis in this study are listed in Dataset S2. This includes the source of the data, reference, number of proteins compiled, and overlap with human directed PPI network. The datasets were downloaded from respective databases or publications as mentioned in Dataset S2. The gene or protein IDs from various resources were mapped to Entrez gene IDs. All compiled datasets are available as an integrated table that shows the nodes and the overlap with respective datasets (Dataset S1).

Analysis of Cancer Genomic Datasets. Copy number alteration data for nine cancer types were downloaded from the cBioPortal for Cancer Genomics (version corresponds to April 2013; www.cbioportal.org). Using the GISTIC algorithm (46), the cBioPortal provides putative values of copy number alterations for each cancer patient. The GISTIC score −2, −1, 0, 1, 2 corresponds to deep loss (possibly a homozygous deletion), single-copy loss (heterozygous deletion), diploid, low-level gain, and high-amplification, respectively. The gene expression data for each cancer type were downloaded from the TCGA (version corresponds to April 2013; https://tcga-data.nci.nih.gov/tcga). The tumor-matched datasets (for each participant have been analyzed and compared with normal tissue on the CNA and gene expression level) were used in the analysis. Level 3 TGCA data (expression calls for genes, per sample) was used in our study. The TCGA data were downloaded by using TCGA web interface with filters set as "Data Type: Expression-Genes"; "Data Level: Level 3"; "Tumor/Normal: Tumor-matched."

Next, we filtered for patients with both CNA and expression data available (details are available in Dataset S4). We computed a z score for each gene in a patient to identify whether the amplification or deletion results in expression change for the corresponding gene. Briefly, for each gene the diploid mean and SD of expression values were calculated using the data from patients without any copy number alteration (GISTIC score, 0;

diploid). Using the diploid mean and SD, we computed z score for each gene in a given patient. A gene is defined as amplified if the GISTIC score is ≥ 1 and the z score is ≥ 1.5 and deleted if the GISTIC score is ≤ -1 and the z score is ≤ -1.5. All of the data preprocessing and normalization were performed using Perl and Java scripts developed in house.

Local Structural Controllability. A dynamic system is controllable if, with a suitable choice of inputs, it can be driven from any initial state to any final state in finite time (2). Most complex biological systems are characterized by nonlinear interactions between the components, and often only local properties can be verified. Similarly, it is often easier to obtain local analytical results for controllability of nonlinear systems. Here, we review a sufficient condition for "local controllability" of a nonlinear system about a trim point. A system is "locally controllable" if there exists a neighborhood in the state space such that all initial conditions in that neighborhood are controllable to all other elements in the neighborhood with locally bounded trajectories (47). This definition of controllability can be verified by checking the well-known "Kalman rank condition" used in the controllability analysis of linear systems. The rest of the section is tutorial in fashion so as to illustrate how the adjacency matrix can be used to analyze the local controllability of a PPI network.

Consider a dynamic system governed by a set of ordinary differential equations

$$\frac{dx}{dt}=f(x(t)),$$

where $x=[x_1\ x_2\cdots\ x_n]^T$ is the state vector and t is time. We are interested in determining an $n\times m$ matrix B such that the controlled system

$$\frac{dx}{dt}=f(x(t))+Bu \qquad \textbf{[1]}$$

is locally controllable through the input $u=[u_1\ u_2\ldots\ u_m]^T$. Let z^* be defined as $f(z^*)=0$, $A(z^*)=\frac{\partial f}{\partial x}(z^*)$, and $G(z^*)=[B\ AB\cdots\ A^{n-1}B]$. The matrix $G(z^*)$ is referred to as the Kalman controllability matrix. The dynamics in Eq. **1** are locally controllable around z^* if $\mathbf{G}(z^*)$ is rank n (Theorem 7 in ref. 47; Proposition 11.2 in ref. 48). The local controllability analysis of Eq. **1** about a trim point therefore reduces to the classic Kalman controllability analysis (2) of the linear dynamics

$$\frac{d(z(t)-z^*)}{dt}=A(z(t)-z^*)+Bu. \qquad \textbf{[2]}$$

Recall that the dynamics in Eq. **2** are deemed "structurally controllable" if there exists another pair (A_0,B_0) with the same structure as the pair (A,B) (49). That is, we are not concerned about the particular values in (A,B), just the pattern of the nonzero entries in (A,B). The dynamics in Eq. **1** are deemed "locally structurally controllable" if the linearized dynamics in Eq. **2** are structurally controllable.

For the purposes of this work, the adjacency matrix of the experimentally determined PPI network is used to find the structure of A in ref. 2, then the nodes are classified based upon their impact on the structural controllability (*Methods*).

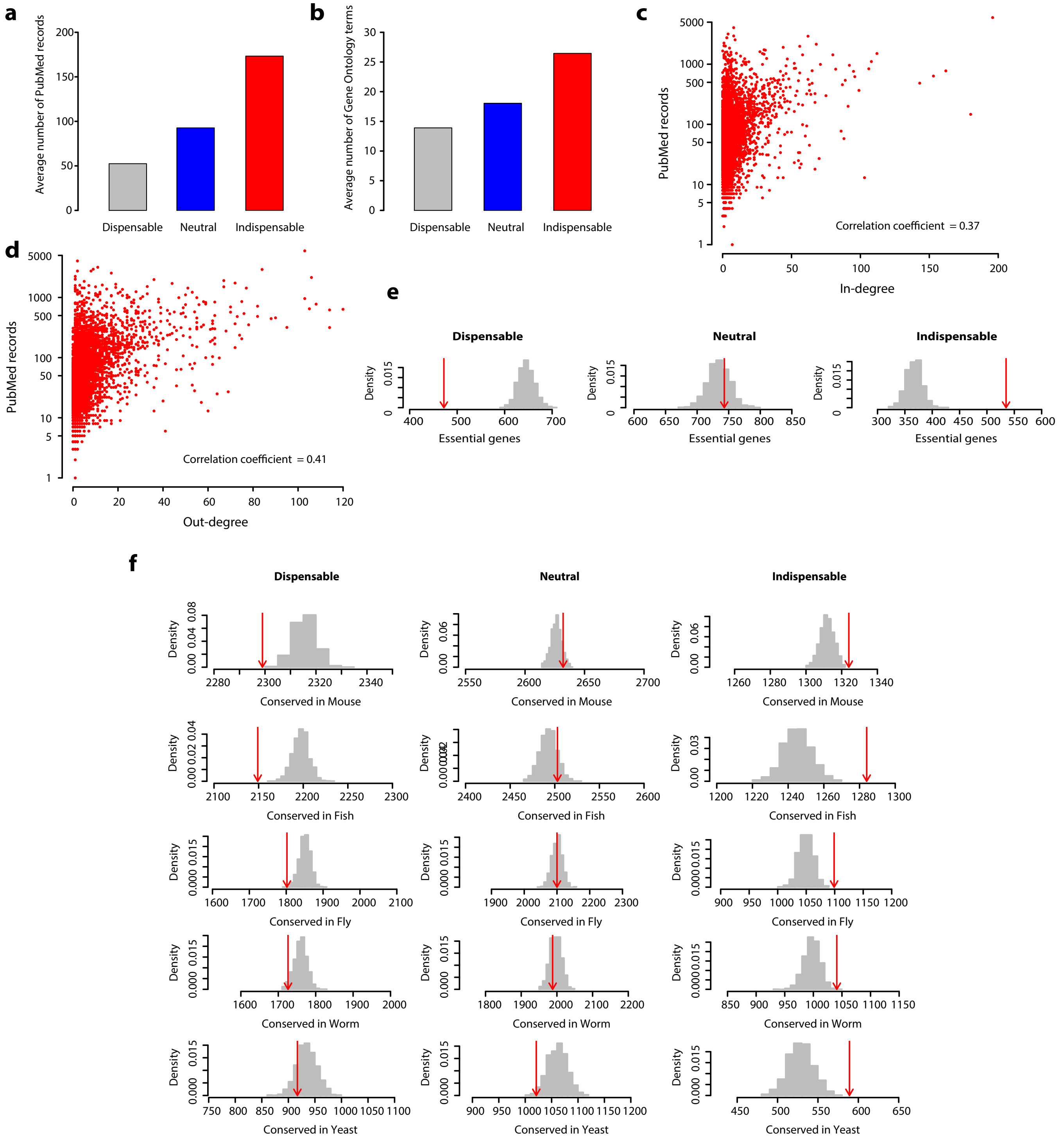


Fig. S1. (*A* and *B*) Literature and annotation bias for the three node types. Bar plots show average PubMed records associated (*A*) and Gene Ontology terms annotated (*B*) for each node type. (*C* and *D*) Correlation of node degree vs. literature bias. The plots show the correlation of in-degree (*C*) and out-degree (*D*) to the number of PubMed records associated with each node in the entire network. (*E*) Enrichment analysis of essential genes. Numbers of essential genes overlapping with dispensable, neutral, and indispensable nodes are shown in red arrows. The essential genes are compiled from the Database of Essential Genes (DEG) (tubic.tju.edu.cn/deg) (50) and Online GEne Essentiality database (OGEE) (ogeedb.embl.de) (51). Numbers of essential genes overlapping with size-controlled random sets are shown in gray bars. (*F*) Enrichment analysis of conserved genes. Numbers of genes conserved in *Mus musculus* (mouse), *Danio rerio* (fish), *Drosophila melanogaster* (fly), *Caenorhabditis elegans* (worm), and *Saccharomyces cerevisiae* (yeast) are shown in red arrows, and their respective size-controlled random set distributions are shown in gray bars. The ortholog mapping was performed using the Drosophila RNAi Screening Center (DRSC) Integrative Ortholog Prediction Tool (DIOPT) (www.flyrnai.org/DIOPT) (52).

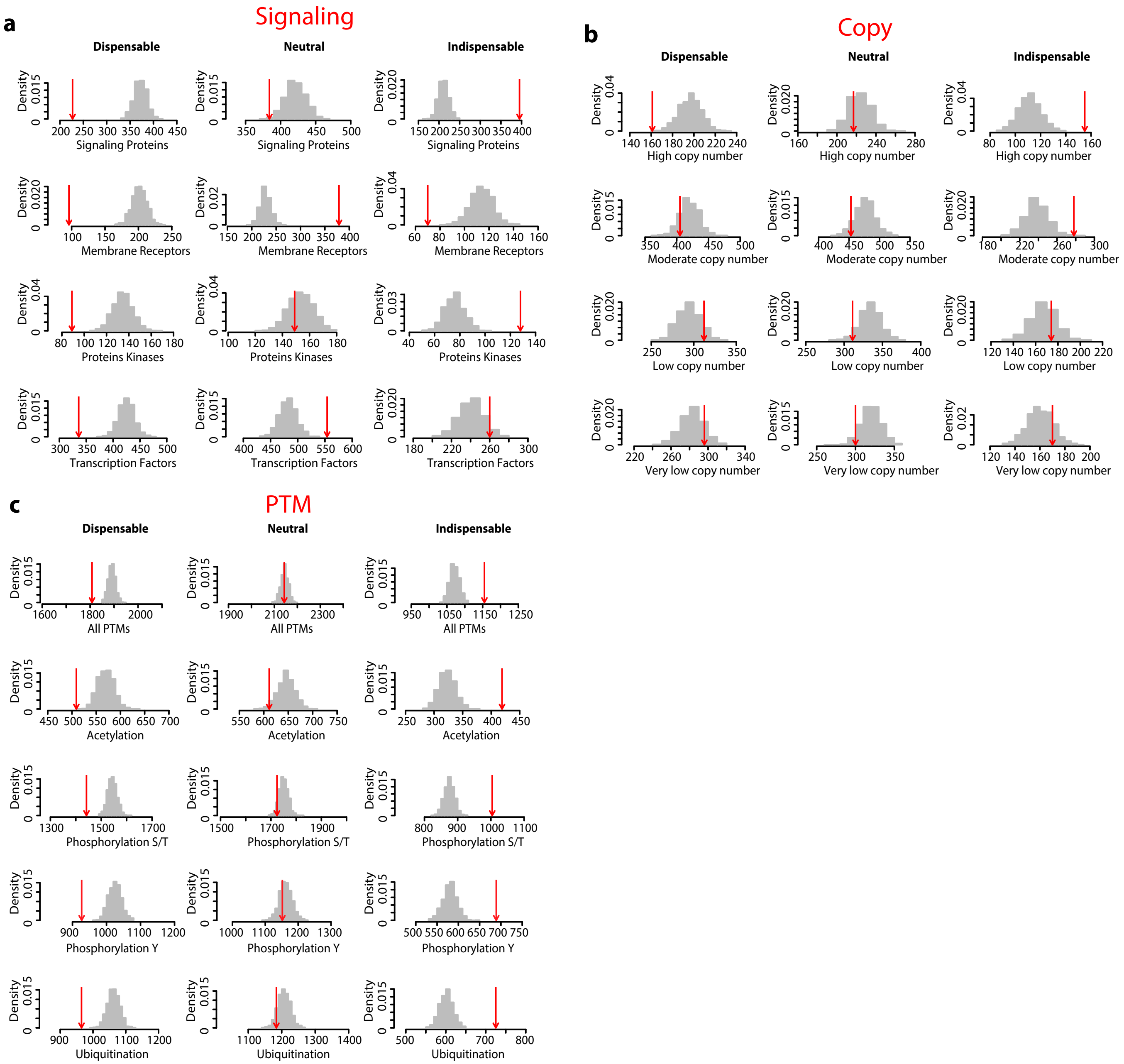


Fig. S2. (*A*) Enrichment analysis of signaling proteins. Numbers of nodes overlapping with signaling proteins (annotated with signaling pathways in Cell Signaling Technology database www.cellsignal.com/common/content/content.jsp?id=science-pathways) (53), receptors (54), protein kinases (55, 56) (kinase.com/kinbase/index.html), and transcription factors (57) are shown in red arrows and, their respective size-controlled random set distributions in gray bars. (*B*) Enrichment analysis of protein abundance. Numbers of nodes overlapping with high copy numbers (>100,000 copies) (*A*), moderate copy numbers (5000–100,000 copies), low copy numbers (500–5,000 copies) (*C*), and very low copy numbers (<500 copies) are shown in red arrows, and their respective size-controlled random set distributions in gray bars. The copy number dataset was obtained from Beck et el. (58). (*C*) Enrichment analysis of protein PTMs. Numbers of nodes overlapping with any PTM [Acetylation, Tyrosine Phosphorylation (Phosphorylation Y), Serine/Threonine Phosphorylation (S/T), or Ubiquitination], Acetylation, Tyrosine Phosphorylation, Serine/Threonine Phosphorylation, and Ubiquitination datasets are shown in red arrows and their respective size-controlled random set distributions in gray bars. The PTM dataset was obtained from Woodsmith et al. (59).

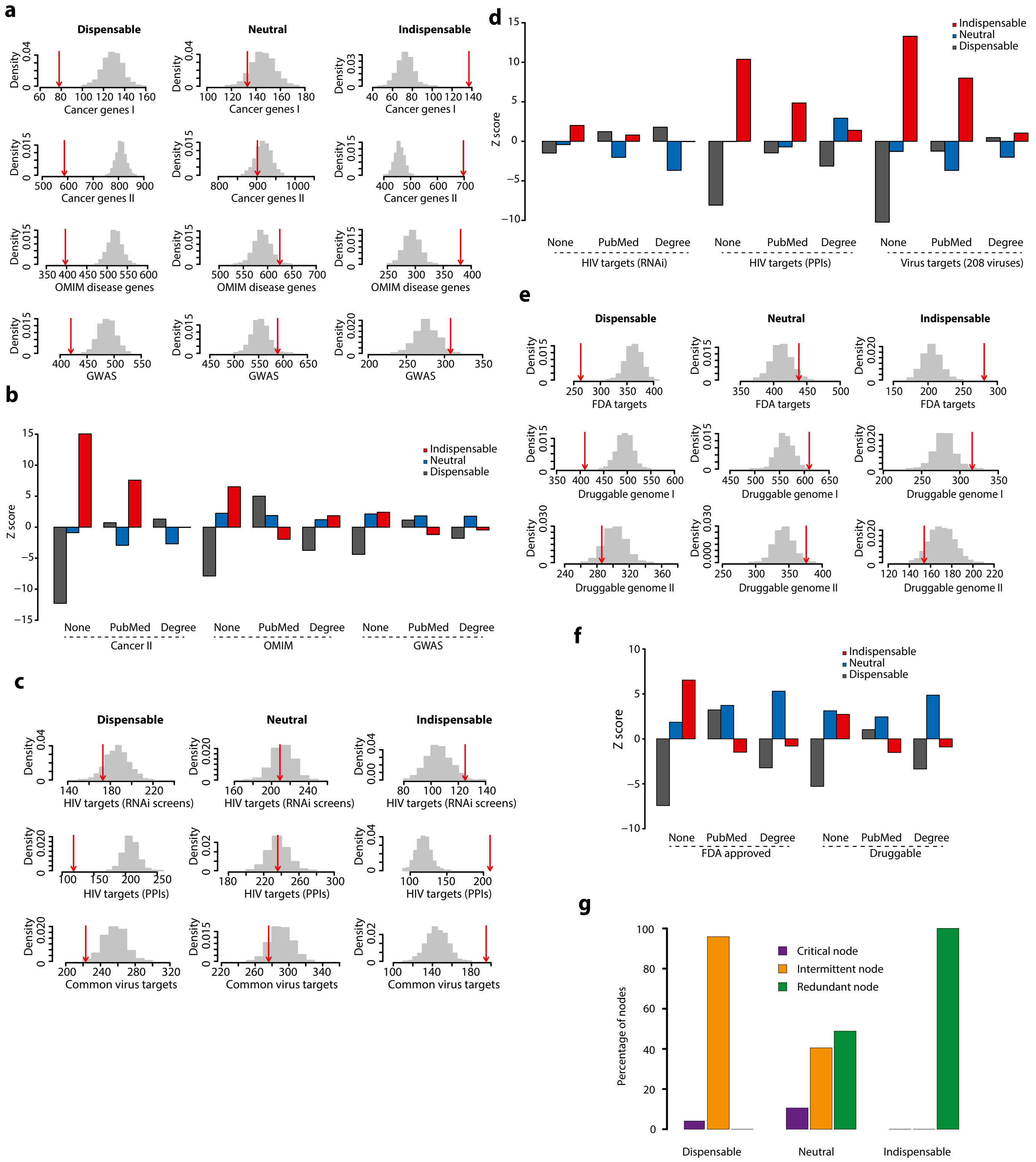
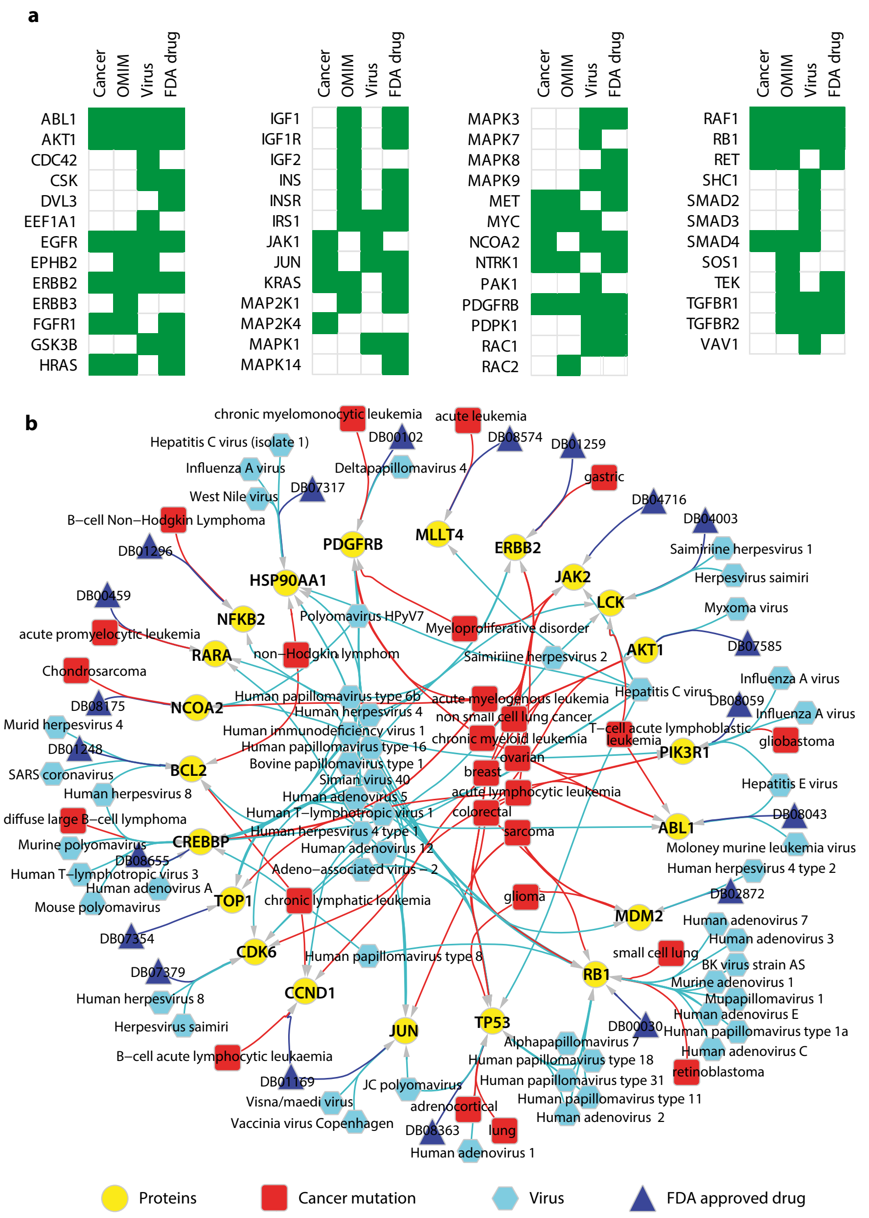


Fig. S3. (*A*) Enrichment analysis of disease genes. Numbers of nodes overlapping with genes causally associated with cancer (Cancer genes I, cancer gene census) (Cancer Gene Census; cancer.sanger.ac.uk/census) (17) and a list of predicted cancer genes (Cancer genes II, extended list of cancer genes) (18), annotated as disease genes in the OMIM database (omim.org) and associated with disease in GWAS (www.genome.gov/gwastudies), are shown in red arrows, and their respective size-controlled random set distributions in gray bars. (*B*) Enrichment analysis of disease genes using literature- and degree-controlled random sets. In the case of degree- or literature-controlled random sets, the random sets are sampled such that the average degree or average PubMed records of random sets matches the average of node type N. (*C*) Enrichment analysis of virus targets. Numbers of nodes overlapping with genes identified to have an adverse effect on HIV-1 replication when knocked down (RNAi screens) (21–24), human proteins that directly interact with HIV proteins (HIV targets PPI) (26, 27), and human proteins that are known to physically interact with proteins from 208 viruses (common virus targets) (26–29) are shown in red arrows, and their respective size-controlled random set distributions in gray bars. (*D*) Enrichment analysis of virus targets using literature- and degree-controlled random sets. Random sets are generated as explained in *B*. (*E*) Enrichment analysis of drug targets. Numbers of nodes overlapping with proteins that are targeted by FDA-approved drugs (31), proteins with domains or folds that could bind to drug-like molecules (druggable genome I) (32), and a subset of druggable genome I excluding the FDA-approved drug targets (druggable genome II) are shown in red arrows, and their respective size-controlled random set distributions in gray bars. (*F*) Enrichment analysis of drug targets using literature- and degree-controlled random sets. Random sets are generated as explained in *B*. (*G*) Characterizing indispensable, neutral, and dispensable nodes based on their roles as driver nodes. The recently developed approach is used to classify a node as critical, intermittent, or redundant if it acts as a driver node in all, some, or none of the control configurations, respectively (33). The bar graph compares the indispensable, neutral, and dispensable nodes against the critical, intermittent, and redundant node classification.



	Cancer	OMIM	Virus	FDA drug
ABL1	🟩	🟩	🟩	🟩
AKT1	🟩	🟩	🟩	🟩
CDC42			🟩	
CSK			🟩	🟩
DVL3				🟩
EEF1A1			🟩	
EGFR	🟩	🟩	🟩	🟩
EPHB2		🟩	🟩	
ERBB2	🟩	🟩	🟩	🟩
ERBB3		🟩		
FGFR1	🟩	🟩		🟩
GSK3B			🟩	🟩
HRAS	🟩	🟩		🟩

	Cancer	OMIM	Virus	FDA drug
IGF1		🟩		🟩
IGF1R		🟩		🟩
IGF2		🟩		
INS		🟩		🟩
INSR		🟩		🟩
IRS1		🟩	🟩	🟩
JAK1	🟩		🟩	
JUN	🟩		🟩	🟩
KRAS	🟩	🟩		🟩
MAP2K1		🟩		🟩
MAP2K4	🟩			
MAPK1			🟩	🟩
MAPK14				🟩

	Cancer	OMIM	Virus	FDA drug
MAPK3			🟩	🟩
MAPK7			🟩	
MAPK8				🟩
MAPK9			🟩	🟩
MET	🟩	🟩		🟩
MYC	🟩	🟩	🟩	
NCOA2	🟩		🟩	🟩
NTRK1	🟩	🟩		🟩
PAK1			🟩	
PDGFRB	🟩	🟩	🟩	🟩
PDPK1			🟩	🟩
RAC1			🟩	🟩
RAC2		🟩		

	Cancer	OMIM	Virus	FDA drug
RAF1	🟩	🟩	🟩	🟩
RB1	🟩	🟩	🟩	🟩
RET	🟩	🟩		🟩
SHC1			🟩	
SMAD2			🟩	
SMAD3			🟩	
SMAD4	🟩	🟩	🟩	
SOS1		🟩		
TEK		🟩		🟩
TGFBR1		🟩	🟩	🟩
TGFBR2		🟩	🟩	🟩
VAV1			🟩	

Fig. S4. (*A*) Members of receptor tyrosine signaling pathways that are predicted as indispensable nodes and targeted by cancer mutations, OMIM disease, viruses, or FDA-approved drugs. RTK pathway members are as defined by the SignaLink database (60). (*B*) Indispensable nodes that are targeted by all three inputs (cancer mutation, viruses, and drugs). The labels of FDA drug nodes correspond to DrugBank IDs. The network was generated using Cytoscape (61).

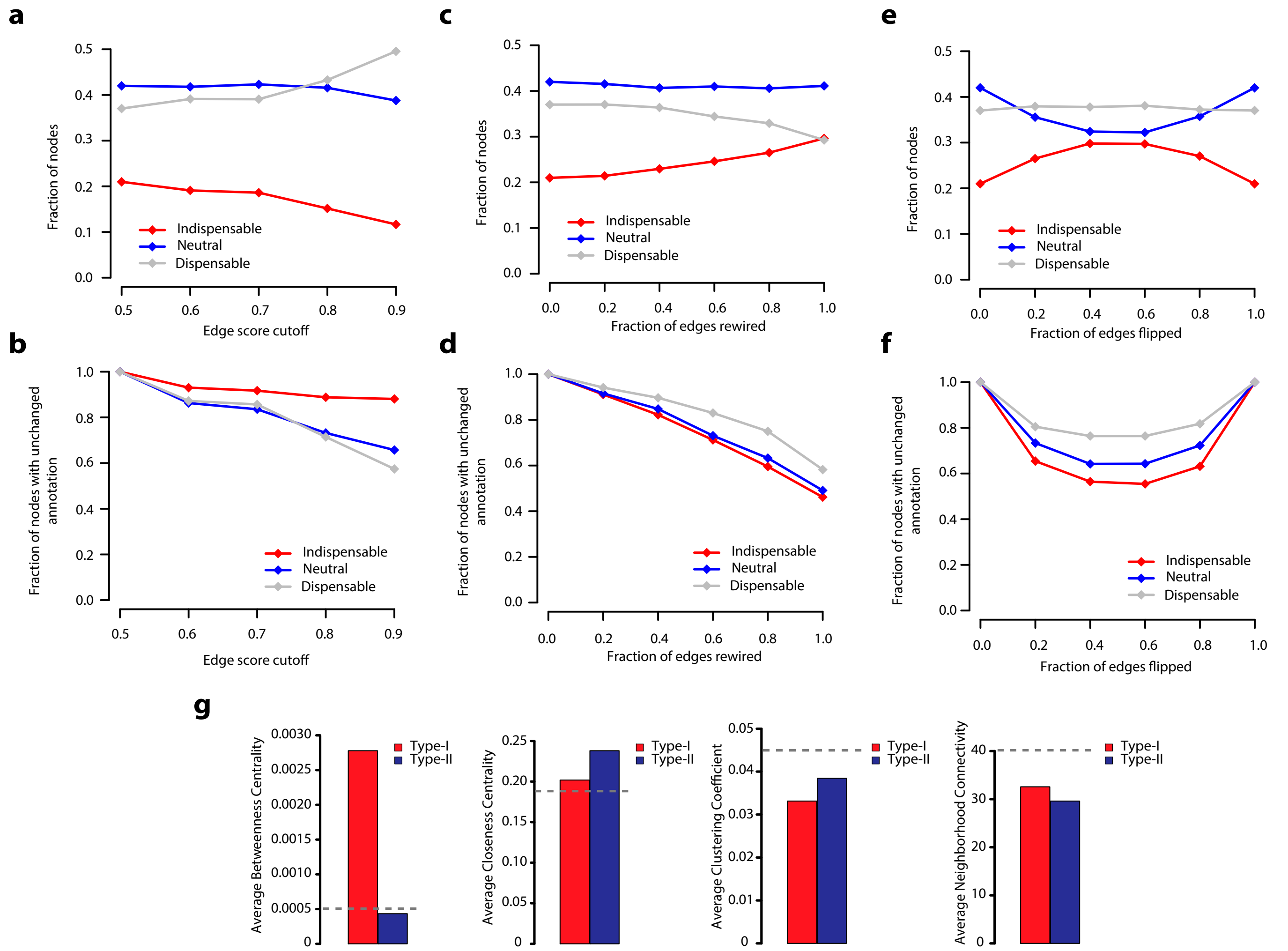


Fig. S5. (*A* and *B*) Robustness of node classification. (*A*) The fraction of indispensable, neutral, and dispensable nodes is plotted as a function of edge filtering (filtering using edge score). (*B*) The fraction of nodes in the filtered network sharing the same node classification as the real network (unchanged annotation) is plotted as a function of edge filtering. (*C–F*) Analysis of node classification in perturbed networks. (*C*) The fraction of indispensable, neutral, and dispensable nodes is plotted as a function of fraction of edges rewired. (*D*) The fraction of nodes in the rewired network sharing the same node classification as the real network (unchanged annotation) is plotted as a function of edge rewired. (*E*) Same as *C*, but the *x* axis corresponds to a fraction of flipped-edge directions. (*F*) Same as *D*, but the *x* axis corresponds to a fraction of flipped-edge directions. (*G*) Comparison of network properties of type-I and type-II indispensable nodes. The network betweenness centrality, closeness centrality, clustering coefficient, and neighborhood connectivity values are calculated using the NetworkAnalyzer Cytoscape plugin (62). The gray dotted line shows the average value of the network.

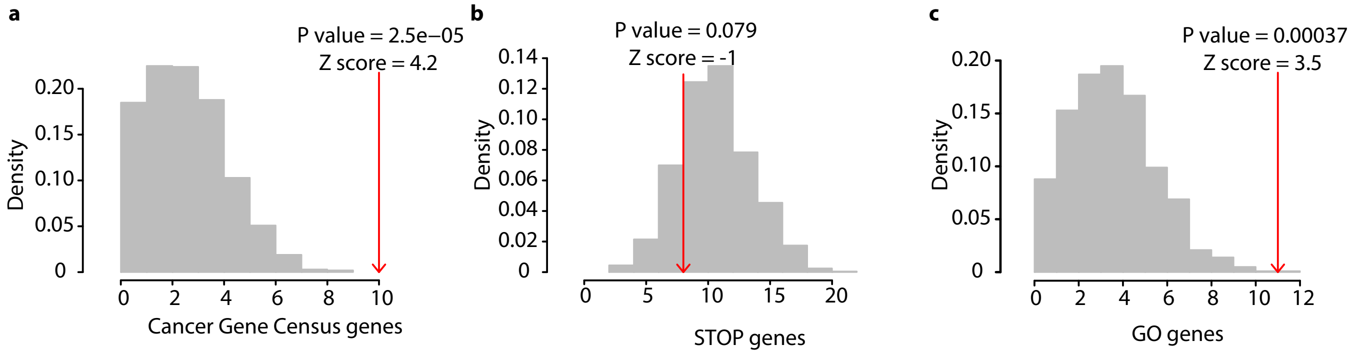


Fig. S6. Enrichment analysis of type-II indispensable nodes frequently amplified/deleted in cancer. Numbers of type-II indispensable nodes (frequently amplified/deleted in cancer) overlapping with genes causally associated with cancer (Cancer Gene Census) (17) (*A*), negative regulators of cell proliferation (STOP genes) (37) (*B*), and positive regulators of cell proliferation (GO genes) (37) (*C*) are shown in red arrows, and their respective size-controlled random set distributions in gray bars.

Dataset S1. Node classification and datasets used in this study

Dataset S1

Node classification: Node classification based on network controllability. The table includes node classifications (indispensable, neutral, or dispensable), their role as a driver node (critical, intermittent, or redundant), in-degree, out-degree, and all other node properties analyzed in this study. Datasets: datasets used in this study for enrichment analysis. Name, source, reference, number of genes/proteins compiled from the dataset, and overlap with the human-directed PPI network are listed.

Dataset S2. Enrichment analysis results and dataset overlaps

Dataset S2

Enrichment: results of enrichment analysis corresponding to Figs. 1*E* and 2 *A*–*D*. The table includes the number of overlapping genes, random mean, random SD, *z* score, and *P* value for indispensable, neutral, and dispensable nodes. Enrichment-controlled: results of degree- and literature-controlled enrichment analysis, corresponding to Figs. 1*E* and 2 *A*–*D*. The table is similar to Enrichment but uses degree- and literature-controlled random sets. Dataset Overlap: indispensable nodes in RTK signaling pathways and the nodes' overlap with cancer mutations, OMIM disease, virus targets, and drug targets (corresponding to Fig. S4*A*). Indispensable Targets: indispensable nodes targeted by all three inputs of cancer mutations, virus targets, and drug targets (corresponds to Fig. S4*B*).

Dataset S3. Results from the analysis of edge-filtered network and subtype of indispensable nodes

Dataset S3

Edge filtering: results from the analysis of edge-filtered network. Edge rewiring: results from the analysis of rewired and direction-flipped networks. Subtypes: list of indispensable nodes with subtype classification (type-I and type-II). The table incudes Entrez gene ID; gene symbol; subtype annotation; results from edge-filtered, rewired, and direction-flipped networks; and overlap with the regulators of cell proliferation.

Dataset S4. Perturbed indispensable nodes in cancer genomic data

Dataset S4

TCGA dataset: the gene expression and CNAs from cancer genomic studies. Gene expression data (level 3 expression calls for genes, per sample) from TCGA and copy number alteration from cBioPortal. The overlapping samples refer to patient samples with both level 3 gene expression and CNA data available. Type-II nodes: most frequently amplified or deleted (top 1%) type-II indispensable nodes in nine different cancer types (corresponds to Fig. 4*E*). Enrichment: enrichment analysis of Cancer Gene Census, type-I and type-II indispensable nodes in human cancer.