Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets

Arunachalam Vinayagam[a,1], Travis E. Gibson[b], Ho-Joon Lee[c,2], Bahar Yilmazel[d,e], Charles Roesel[d,e,3], Yanhui Hu[a,d], Young Kwon[a], Amitabh Sharma[b,f,g], Yang-Yu Liu[b,f,g,1], Norbert Perrimon[a,h,1], and Albert-László Barabási[f,g,1]

[a]Department of Genetics, Harvard Medical School, Boston, MA 02115; [b]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115; [c]Department of Systems Biology, Harvard Medical School, Boston, MA 02115; [d]Drosophila RNAi Screening Center, Department of Genetics, Harvard Medical School, Boston, MA 02115; [e]Bioinformatics Program, Northeastern University, Boston, MA 02115; [f]Center for Complex Network Research, Department of Physics, Northeastern University, Boston, MA 02115; [g]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115; and [h]Howard Hughes Medical Institute, Harvard Medical School, MA 02115

Contributed by Norbert Perrimon, March 24, 2016 (sent for review November 25, 2015; reviewed by Reka Albert and Tatsuya Akutsu)

The protein–protein interaction (PPI) network is crucial for cellular information processing and decision-making. With suitable inputs, PPI networks drive the cells to diverse functional outcomes such as cell proliferation or cell death. Here, we characterize the structural controllability of a large directed human PPI network comprising 6,339 proteins and 34,813 interactions. This network allows us to classify proteins as "indispensable," "neutral," or "dispensable," which correlates to increasing, no effect, or decreasing the number of driver nodes in the network upon removal of that protein. We find that 21% of the proteins in the PPI network are indispensable. Interestingly, these indispensable proteins are the primary targets of disease-causing mutations, human viruses, and drugs, suggesting that altering a network's control property is critical for the transition between healthy and disease states. Furthermore, analyzing copy number alterations data from 1,547 cancer patients reveals that 56 genes that are frequently amplified or deleted in nine different cancers are indispensable. Among the 56 genes, 46 of them have not been previously associated with cancer. This suggests that controllability analysis is very useful in identifying novel disease genes and potential drug targets.

network biology | controllability | protein–protein interaction network | disease genes | drug targets

The need to control engineered systems has resulted in a mathematically rich set of tools that are widely applied in the design of electric circuits, manufacturing processes, communication systems, aircraft, spacecraft, and robots (1–3). Control theory deals with the design and stability analysis of dynamic systems that receive information via inputs and have outputs available for measurement. Issues of control and regulation are central to the study of biological systems (4, 5), which sense and process both external and internal cues using a network of interacting molecules (6). The dynamic regulation of this molecular network in turn drives the system to various functional states, such as triggering cell proliferation or inducing apoptosis. This feature of specific input signals driving networks from an initial state to a specific functional state suggests that the need to control a biological system plays a potentially important role in the evolution of molecular interaction networks. Note that the term "state" is also used in a control context where the "state space" of a control system is the space of values the "state variables" can attain. For a protein–protein interaction (PPI) network, the state variables are the specific protein concentrations and the state space is all positive real numbers of dimension equal to the total number of proteins in the PPI network.

According to control theory, a dynamic system is controllable if, with a suitable choice of inputs, the system can be driven from any initial state to any desired final state in finite time (2, 7). Previous studies have shown that network components exhibit properties of control systems such as proportional action, feedback control, and feed-forward control (8–12). However, the main challenges that hinder systematic controllability analysis of biological networks are the availability of large-scale biologically relevant networks and efficient tools to analyze their controllability. To address these issues, two resources were integrated in this work: (*i*) a directed human PPI network (13); and (*ii*) an analytical framework to characterize the structural controllability of directed weighted networks (14). The directed human PPI network represents a global snapshot of the information flow in cell signaling. For a given weighted and directed network associated with linear time-invariant dynamics, the analytical framework identifies a minimum set of driver nodes, whose control is sufficient to fully control the dynamics of the whole network (14, 15).

In this work, we classified the proteins (nodes) as "indispensable," "neutral," or "dispensable," based on the change of the minimum number of driver nodes needed to control the PPI network when a specific protein (node) is absent. In addition, we analyzed the role of different node types in the context of human diseases. Using known examples of disease-causing

Significance

Large-scale biological network analyses often use concepts used in social networks analysis (e.g. finding "communities," "hubs," etc.). However, mathematically advanced engineering concepts have only been applied to analyze small and well-characterized networks so far in biology. Here, we applied a sophisticated engineering tool, from control theory, to analyze a large-scale directed human protein–protein interaction network. Our analysis revealed that the proteins that are indispensable, from a network controllability perspective, are also commonly targeted by disease-causing mutations and human viruses or have been identified as drug targets. Furthermore, we used the controllability analysis to prioritize novel cancer genes from cancer genomic datasets. Altogether, we demonstrated an application of network controllability analysis to identify new disease genes and drug targets.

Author contributions: A.V., Y.-Y.L., N.P., and A.-L.B. designed research; A.V., T.E.G., and Y.-Y.L. performed research; H.-J.L., B.Y., C.R., Y.H., Y.K., and A.S. contributed new reagents/analytic tools; A.V., T.E.G., and Y.-Y.L. analyzed data; and A.V., T.E.G., Y.-Y.L., and A.-L.B. wrote the paper.

Reviewers: R.A., Pennsylvania State University; T.A., Kyoto University.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: vinu@genetics.med.harvard.edu, yyl@channing.harvard.edu, perrimon@receptor.med.harvard.edu, or alb@neu.edu.

[2]Present address: Department of Laboratory Medicine, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, 062-73, Korea.

[3]Present address: Marine Science Center, Northeastern University, Nahant, MA 01908.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1603992113/-/DCSupplemental.

mutations, virus targets, and drug targets, we identified indispensable nodes that are key players in mediating the transition between healthy and disease states. Our study illustrates the potential application of network controllability analysis as a powerful tool to identify new disease genes.

Results

Characterizing the Controllability of the Directed PPI Network. We applied linear control tools to access local controllability of PPI networks whose dynamics are inherently nonlinear. The experimentally obtained network, however, can be assumed to capture linear affects around homeostasis. Furthermore, given that the tools developed in ref. 14 are for linear dynamics, we are careful to only assume that we can ascertain local controllability around homeostasis. Controllability henceforth referred to local controllability (see *SI Text* for details).

The directed human PPI network consists of 6,339 proteins (nodes) and 34,813 directed edges, where the edge direction corresponds to the hierarchy of signal flow between the interacting proteins and the edge weight corresponds to the confidence of the predicted direction. We applied structural controllability theory to identify a minimum set of driver nodes (i.e., nodes through which we can achieve control of the whole network). Note that the identified minimum driver node set (MDS) is not unique, but its size, denoted as N_D, is uniquely determined by the network topology. We found that the MDS of the directed human PPI network contains 36% of nodes. We also classified the nodes as indispensable, neutral, or dispensable, based on the change of N_D upon their removal. A node is (*i*) indispensable if removing it increases N_D (e.g., node 2 in Fig. 1*A*), (*ii*) neutral if its removal has no effect on N_D (e.g., node 1 in Fig. 1*A*), and (*iii*) dispensable if its removal reduces N_D (e.g., nodes 3 and 4 in Fig. 1*A*). In the directed human PPI network, 21% of nodes are indispensable, 42% are neutral, and the remaining 37% are dispensable (Fig. 1*B*). Interestingly, we found that all of the three node types have a heterogeneous degree distribution, and indispensable nodes tend to have higher in- and out-degrees compared with neutral and dispensable nodes (Fig. 1 *B* and *C*). Similarly, indispensable nodes are associated with more PubMed records (www.ncbi.nlm.nih.gov/pubmed) and Gene Ontology (16) term annotation than neutral and dispensable nodes (Fig. S1 *A* and *B*). However, the correlation between the node-degree and the literature bias is weak (correlation coefficient of 0.37 and 0.41 for in- and out-degree, respectively), suggesting that the higher degree of indispensable nodes is not explained by the literature bias alone (Fig. S1 *C* and *D*).

We characterized indispensable, neutral, and dispensable nodes in the context of essentiality, evolutionary conservation, and regulation at the level of translational and posttranslational modifications (PTMs). Our gene essentiality analysis indicated that indispensable nodes are enriched in essential genes, whereas essential genes are underrepresented among dispensable nodes (Fig. 1*E*, Fig. S1*E*, and Dataset S1). Furthermore, indispensable nodes are evolutionarily conserved from human to yeast compared with the other two node types (Fig. 1*E* and Fig. S1*F*). Next, we analyzed the different node types in the context of cell signaling, which is at the core of cellular information processing. In general, known signaling proteins are enriched as indispensable nodes. However, dissecting different functional classes within signaling proteins reveals that kinases are enriched as indispensable nodes whereas membrane receptors and transcription factors are enriched as neutral nodes (Fig. 1*E* and Fig. S2*A*). Analysis of the protein steady-state abundance in cell lines, as a measure of translational regulation, reveals that indispensable nodes are enriched as high copy number proteins, whereas low-copy number proteins show moderate enrichments for both

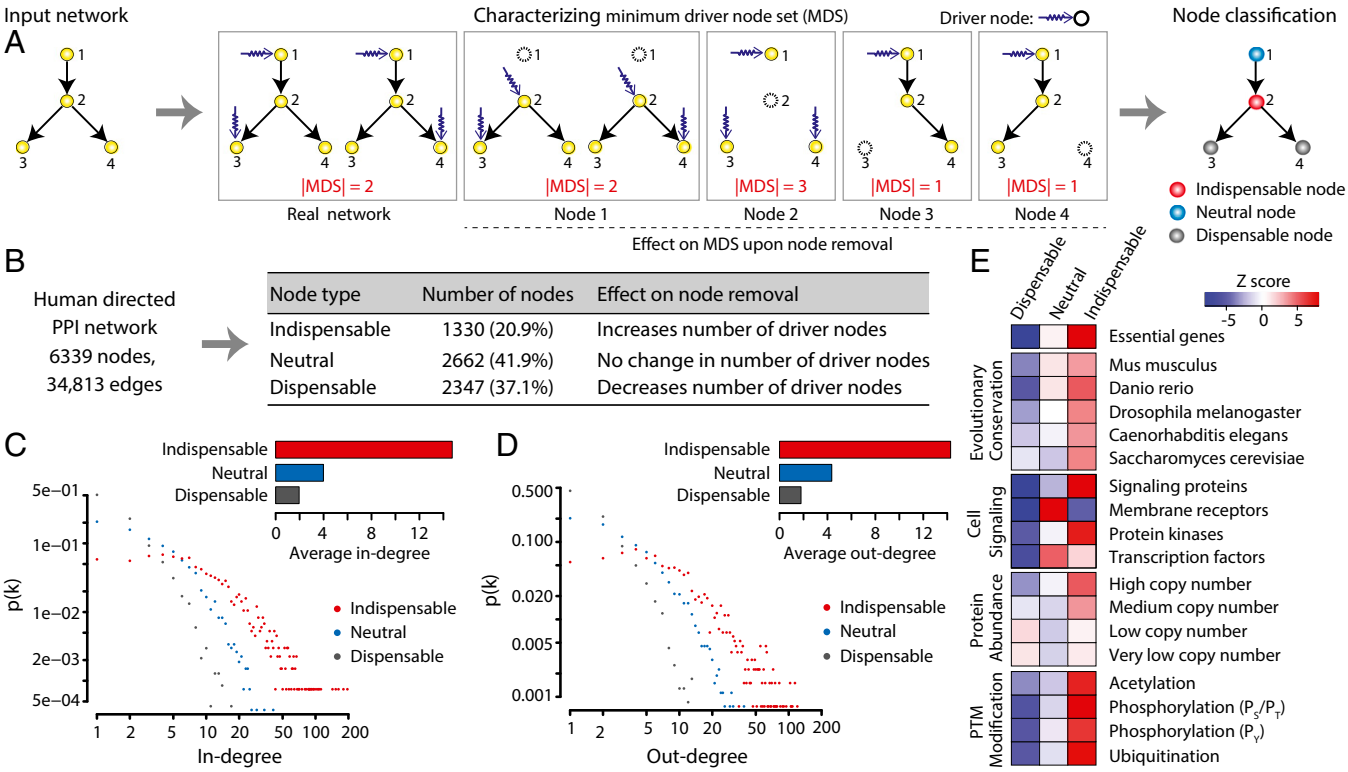

Fig. 1. Characterizing the controllability of human directed PPI network. (*A*) Schematic representation of the node classification using controllability framework. (*B*) Identification of indispensable, neutral, and dispensable nodes in human directed PPI network. (*C*) In-degree distribution and average in-degree for three different node types. (*D*) Out-degree distribution and average out-degree for three different node types. (*E*) Distinct enrichment profiles of indispensable, neutral, and dispensable nodes in the context of essential genes, evolutionary conservation, cell signaling, protein abundance, and PTMs.

indispensable and dispensable nodes (Fig. 1*E* and Fig. S2*B*). Similarly, indispensable nodes are highly regulated through PTM, including acetylation, ubiquitination, and phosphorylation (pS/pT and pY) (Fig. 1*E* and Fig. S2*C*). Altogether, our enrichment analyses revealed distinct functional and regulatory roles for indispensable, neutral, and dispensable nodes.

Understanding Healthy to Disease State Transition Using Network Controllability. We analyzed the node classification in the context of driving the system from healthy to disease condition and vice versa. Specifically, we analyzed the impact of three different transitions: (*i*) healthy to disease transition induced by mutations or other genetic alterations; (*ii*) healthy to infectious transition induced by human viruses; and (*iii*) disease to healthy transition induced by drugs or small molecules. Note that our goal is to determine whether specific node types (indispensable, neutral, or dispensable) are enriched for (*i*) disease-causing mutations, (*ii*) targets of human viruses, and (*iii*) drug targets.

First, we analyzed 445 genes annotated by the Sanger Center as causally implicated in oncogenesis (Cancer Gene Census; cancer.sanger.ac.uk/census) (17). Interestingly, we found that indispensable nodes are highly enriched in cancer genes, whereas neutral nodes showed no enrichment and dispensable nodes are underrepresented (Fig. 2*A*, Cancer I; Fig. S3*A*; and Dataset S2). To ensure that the observed enrichment of indispensable nodes is not attributable to the literature and degree bias, we repeated our analysis using literature- and degree-controlled random sets (*SI Text*). After adjusting for literature and degree bias (Fig. 2*A*, PubMed and Degree; and Dataset S2), indispensable nodes remain significantly enriched as cancer genes. Note that for enrichment analysis below, the degree- and literature-controlled enrichments results were shown in Fig. S3*B*. To further substantiate that indispensable nodes are enriched as cancer genes, we analyzed 3,164 genes predicted as cancer related genes (18) and observed a similar enrichment for indispensable nodes (Fig. 2*B*, Cancer II; and Fig. S3*A*).

Next, we analyzed 1,403 genes annotated by Online Mendelian Inheritance in Man (OMIM) (omim.org) as causal genes for various genetic diseases, aiming to test whether the perturbation of indispensable nodes is a specific feature of cancer or a general feature of human diseases. Our analysis showed that the perturbation of indispensable nodes is a common feature of human diseases (Fig. 2*B*, OMIM; and Fig. S3*A*). Interestingly, however, our analysis of disease genes identified from genome-wide association studies (GWAS) (www.genome.gov/gwastudies) (19) revealed poor enrichment for indispensable nodes (Fig. 2*B*, GWAS; and Fig. S3*A*), most likely reflecting the fact that GWAS identify genomic regions but not specific coding genes that cause the disease (20). Because indispensable nodes are enriched for causal mutations (Fig. 2 *A* and *B*), our resource could help identify causal genes from GWAS.

We also characterized the network controllability in the context of host–parasite interactions, specifically human–virus interactions. Upon infection, viruses control the host cellular network to use the host resources to replicate and to evade the host immune response. Here, we analyzed the node types targeted by human viruses to drive the network from a healthy state to an infectious state. First, we analyzed the targets of HIV, a member of the lentivirus family that causes AIDS. Putative human genes, identified to have an effect on HIV-1 replication from large-scale functional genomic screens (data compiled from four RNAi datasets) (21–24) tend to be indispensable nodes (Fig. 2*C*, RNAi; and Fig. S3*C*). However, we did not detect a significant enrichment—most likely reflecting the quality of the HIV RNAi screens (25). To analyze direct targets of HIV, we compiled the HIV–human interactome (from recent literature and PPI databases) (26, 27), finding that indispensable nodes are enriched for physical interactions with HIV proteins (Fig. 2*C*, PPIs; and Fig. S3*C*). Analysis of 208 different human–virus networks (26–29) reveals that human viruses commonly target indispensable nodes to control the host network (Fig. 2*C*, Virus targets; and Fig. S3*C*). We noticed that after adjusting for literature bias indispensable nodes remain as viral targets, whereas adjusting for degree bias shows only weak enrichment (Fig. S3*D*). This finding is in agreement with the previous observations that viruses tend to target hubs (30).

Finally, we characterized the network controllability in the context of driving the system from disease to healthy state. Specifically, we analyzed the node types that are targeted by the drugs/small molecules (Fig. 2*D*). By analyzing the targets of drugs approved by the Food and Drug Administration (FDA) (31), we found that indispensable nodes are enriched for drug targets (Fig. 2*D*, FDA targets; and Fig. S3 *E* and *F*). Extending the analysis to the list of proteins that are annotated as druggable (32), i.e., a presence of protein folds that favor interactions with drug-like chemical compounds, showed that the druggable genome list is not significantly enriched for indispensable nodes (Fig. 2*D*, D I; and Fig. S3*E*). Interestingly, analyzing the druggable genome list by excluding FDA-approved drug targets showed underrepresentation of indispensable nodes (Fig. 2*D*, D I; and Fig. S3*E*). This finding suggests a potential application of our analysis to redefine the druggable genome based on the network controllability.

All of the above analyses of disease mutations, viruses, and drugs consistently showed that indispensable nodes are preferred targets. We also analyzed how often indispensable nodes act as driver nodes by using a recently developed approach to identify the role of each node as drivers in the MDSs (33). We found that 378 nodes appear in all MDSs (i.e., they play roles in all of the control configurations), 3,330 nodes are in some but not all MDSs (i.e., they play roles in some control configurations but the network can still be controlled without directly controlling them), and 2,631 nodes do not belong to any MDS (i.e., they play no roles in control) (Dataset S1) (33). Interestingly, we found that indispensable nodes are never driver nodes in any MDS (Fig. S3*G* and Dataset S1). This fact can actually be rigorously proven (*SI Text*). Moreover, perturbing indispensable nodes increases the number of driver

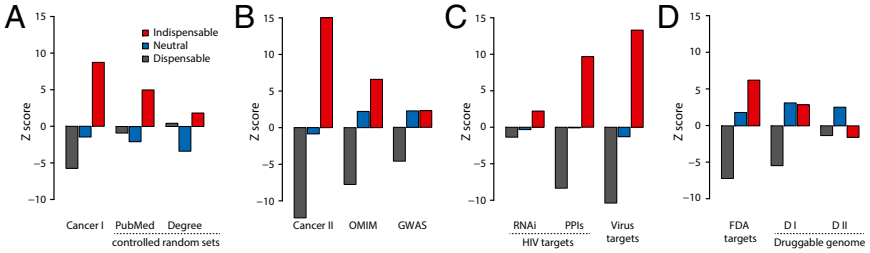

Fig. 2. Characterizing network controllability in transition from healthy to disease state. (*A*) Bar graph showing the enrichment results (*z* scores) of cancer genes compared with the random sets (Cancer I, cancer gene census) and the random sets controlled for literature (PubMed) or degree (Degree) bias. 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. (*B*) Results from enrichment analysis of dataset corresponding to extended list of cancer genes (Cancer II), other human diseases (OMIM), and GWAS. (*C*) Results from enrichment analysis of the targets of HIV identified using RNAi screens (RNAi) and PPI networks (PPIs) and targets of other human virus (208 viruses). (*D*) Enrichment results from targets of FDA-approved drugs and druggable genome. DI, druggable genome; DII, druggable genome excluding FDA-approved targets.

nodes to control, suggesting that, from a controllability perspective, these nodes are fragile points in the network.

We further analyzed indispensable nodes in specific signaling pathways such as receptor tyrosine kinase (RTK) signaling pathways, which are commonly perturbed in cancer (34). Strikingly, 67 out of 170 RTK pathway members are indispensable nodes ($P < 0.0001$), including 51 indispensable nodes targeted by disease mutations, viruses, or drugs (Fig. S4*A* and Dataset S2). Furthermore, we identified 21 indispensable nodes from different signaling pathways that are shared targets of cancer mutations, viruses, as well as drugs (Fig. S4*B* and Dataset S2).

Robustness of Indispensable Node Classification. The false-positive and false-negative interactions are major concerns in PPI networks, especially the false negatives because the current networks are vastly incomplete (35). Hence, we systematically analyzed the robustness of node classification with respect to adding or removing interactions. Specifically, we analyzed the indispensable node classification as a function of removing edges (or network filtering). The network filtering is achieved by using a confidence score assigned to edge directions, where the most stringent filtering resulted in smaller high-confidence–directed networks (20,151 edges and 5,317 nodes). We analyzed the controllability of filtered networks and compared them to the original network. The results show that 90% of the indispensable nodes in the stringent filtered network are indispensable in the original network (Fig. 3*A*, Fig. S5 *A* and *B*, and Dataset S3), suggesting that the indispensable node classification is robust with respect to adding or removing edges in the network.

Next, we analyzed the controllability of networks with perturbations (e.g., edge rewiring or edge-direction flipping). In the case of random rewiring, up to 100% of the edges are rewired (node degrees are preserved), and in the case of direction-flipped networks, up to 100% of the edge directions are reversed. We observed that up to 50% of indispensable nodes in the rewired or direction-flipped network do not agree with the original annotation, showing that indispensability is highly sensitive to the connectivity pattern and edge direction (Fig. 3*B*, Fig. S5 *C–F*, and Dataset S3). Comparing indispensable nodes of the real network to that of the rewired (100% rewiring) and flipped (40% flipping) networks revealed two subtypes (type-I and type-II) of indispensable nodes (Fig. 3*C* and Dataset S3). If a node's indispensability is robust to rewiring or flipping, then we call it a type-I node; if the node's indispensability is sensitive to rewiring or flipping, then we refer to them as type-II nodes. We found that 57% of indispensable nodes are type-I nodes and 43% are type-II.

Degree distribution of the subtypes shows that type-I nodes tend to be hubs, whereas the average degree of type-II nodes is similar to the average degree of the network (Fig. 3*D*). Indeed, type-II nodes cannot be distinguished from the rest of the nodes based on any other network properties analyzed (Fig. S5*G*). Furthermore, type-I nodes show literature and annotation bias compared with type-II nodes (Fig. 3 *E* and *F*). With respect to diseases, both node types show similar enrichment for cancer genes and other human diseases (Fig. 3*G*). In contrast to type-I nodes that tend to be hubs and well-studied genes, type-II nodes are poorly studied and show no special network feature except indispensability, suggesting that control theory brings orthogonal information to traditional network analysis.

Applying Network Controllability Analysis to Mine Cancer Genomic Data. Our finding that indispensable nodes (both type-I and type-II) are more likely to correspond to cancer genes prompted us to systematically survey the perturbation of those genes in cancer. We analyzed data from 1,547 patients obtained from The Cancer Genome Atlas (TCGA) (cancergenome.nih.gov) and cBioPortal for Cancer Genomics (36), representing nine different cancer types (Dataset S4). Specifically, we analyzed the amplification or deletion of type-II indispensable nodes in nine cancer types. Note that the copy number alteration (CNA) data are normalized to the expression levels to identify the amplification or deletion that results in expression level changes (*SI Text*). We ranked all genes based on the number of patients where the gene is amplified or deleted and selected the top 1% as frequently amplified/deleted genes; 56 type-II genes were identified as part of the top 1% of deleted/amplified genes in nine cancer types (Fig. 4*A* and Dataset S4). Strikingly, 10 of 56 type-II genes are known cancer genes, an overlap that is highly significant ($P = 0.00002$) (Fig. 4*B* and Fig. S6*A*). Interestingly, the frequency of deletion and amplification of type-II indispensable nodes is not significantly enriched compared with random sets, an observation that was similar to cancer gene census gene list (Dataset S4). Furthermore, we compared the type-II genes with results from a cell proliferation screen (37) that identified a subset of genes that regulate cell proliferation ("GO" genes induce the proliferation and "STOP" genes suppress the proliferation); 17 of 56 genes represent regulators of cell proliferation (11 GO genes, 8 STOP genes, and 2 genes part of both GO and STOP genes) (Fig. 4*C* and Fig. S6 *B* and *C*). The overlap between type-II genes and GO genes are statistically significant ($P = 0.0003$). Of 56 genes, 10 genes are frequently perturbed in multiple cancer types [e.g., proteasome 26S subunit, non-ATPase, 4 (PSMD4) in four different cancers], and all of them show similar deletion or amplification profile (e.g.,

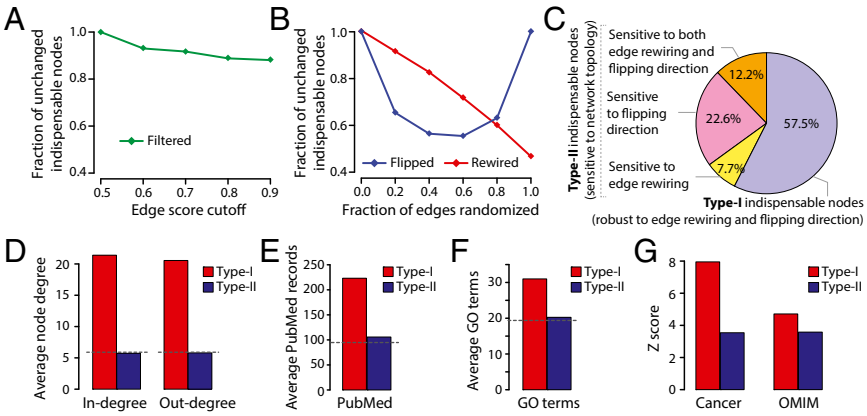

Fig. 3. Perturbation of network connectivity reveals two subtypes of indispensable nodes (type-I and type-II). (*A*) Plot showing the fraction of indispensable nodes in filtered networks that overlaps with the real network. The network filtering achieved using edge confidence score. (*B*) Fraction of indispensable nodes in rewired or direction-flipped overlap with the real network. (*C*) Identification of type-I and type-II indispensable nodes. The average node degree (*D*), PubMed record association (*E*), and Gene Ontology (GO) term annotations (*F*) for type-I and type-II indispensable nodes. (*G*) Enrichment of type-I and type-II indispensable nodes as cancer genes and OMIM disease genes.

PSMD4 amplified in all four cancers) (Fig. 4*D*). Almost half of the genes (23 genes) are poorly studied, with less than 50 associated PubMed records; for instance, small G protein signaling modulator 2 (SGSM2) is associated with only eight PubMed records (Fig. 4*D*). These contextual evidences, along with the indispensability, suggest that these 46 type-II nodes could be potential cancer genes.

Database of Directed PPI Network with Predicted Controllability. We created the DirectedPPI database (www.flyrnai.org/DirectedPPI) to navigate the directed human PPI network with predicted controllability. Users can enter a gene or upload a list of genes and our tool generates a network with directed edges connecting the input list. Our tool also accepts gene list with values (e.g., mutation frequency, *P* values from GWAS, or expression changes). Three different node types (indispensable, neutral, and dispensable) are distinguished with node shape and color and for these nodes all of the properties analyzed in this article are displayed. This tool will be useful to analyze disease datasets and other high-throughput datasets to identify indispensable nodes and their interconnections.

Discussion

Studying the controllability of a complex biological network is rather difficult, because of the fact that we typically do not know the true functional form of the underlying dynamics. However, most biological systems operate near homeostasis, so local properties are indeed what we want to ascertain. Here, we showed that application of linear control tools to study the local structural controllability of inherent nonlinear biological networks provides meaningful predictions. Furthermore, we demonstrated that local controllability tools help identifies known human diseases genes and this can be used to identify novel disease genes and drug targets.

Our analysis of directed human PPI network identifies 36% of the nodes as driver nodes, which is similar to what has been observed in metabolic networks (~30%) (14). The node classification based on network controllability shows distinct biological properties in the context of essentiality, conservation and regulation. Specifically, we found that indispensable nodes are well conserved, highly regulated at the level of translational and PTMs, and important for the transition between healthy and disease states. Interestingly, this enrichment pattern is partially shared by the nodes in the minimum dominating sets that are located in strategically important positions in controlling the network (38–40). Furthermore, identification of the indispensable nodes as primary targets of diseases causing mutations, viruses, and drugs revealed a potential application of this framework to identify novel disease genes and potential drug targets.

Interestingly, disease-causing mutations, viruses, and drugs target fragile points (indispensable nodes) that determine the number of driver nodes rather than the driver nodes themselves, suggesting that network controllability is crucial in transitioning between healthy and disease states. Although network topology-based properties such as hubs and modules are commonly used to identify disease genes (41–44), the controllability perspective provides a complementary network analysis framework for network medicine. In particular, type-II nodes that are not distinguishable from existing network properties and without publications bias were still identified by our controllability framework as nodes of special interest. We envision that in the future, improving the quality and the completeness of interactome maps and integrating dynamics of network components would hugely impact our understanding of biological networks both in the context of biological function and human disease.

Methods

Datasets and Enrichment Analysis. All of the datasets used for the enrichment analysis in this study are listed in Dataset S2. Details on the directed human PPI network, random networks, and datasets used for enrichment analysis can be found in *SI Text*.

Controllability Analysis and Node Classification. 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*) on any directed weighted network, we can map the structural controllability problem in control theory to the maximum matching problem in graph theory, which can be solved in polynomial time (15).

After a node is removed, denote the minimum number of driver nodes of the damaged network as N_D'. We can classify nodes into three categories: (*i*) a node is indispensable if in its absence we have to control more driver nodes (i.e., $N_D' > N_D$); (*ii*) a node is dispensable if in its absence we have $N_D' < N_D$); and (*iii*) a node is neutral if in its absence $N_D' = N_D$. Note that indispensable nodes are never driver nodes in any control configurations or MDSs, which can be proven by contradiction (*SI Text*). More information on node classification and local structural controllability can be found in *SI Text*.

Random Networks. To compare the real network with its randomized counterparts, we performed two types of randomization: (*i*) edge rewiring:

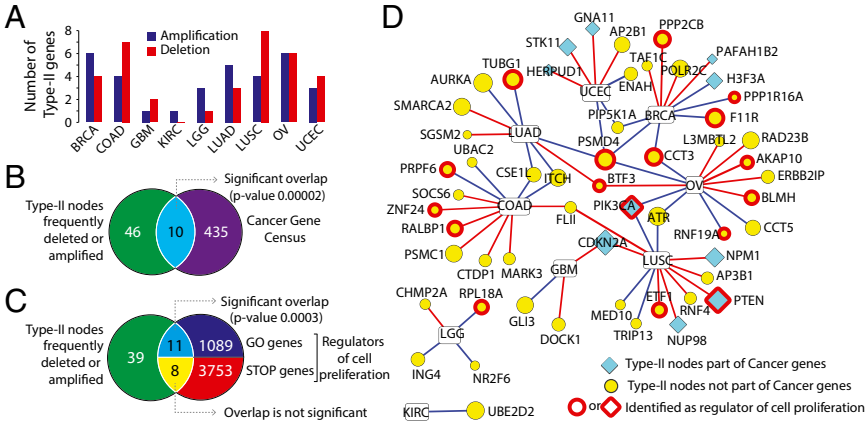

Fig. 4. Applying network controllability to mine cancer genomic data. (*A*) Type-II genes frequently amplified or deleted in cancer patients (part of top 1% genes). The bar plot shows number of type-II genes deleted/amplified in brain lower grade glioma (LGG), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), and glioblastoma multiforme (GBM) cancers. (*B*) Overlap between frequently deleted/amplified type-II genes and known cancer genes. (*C*) Overlap between frequently deleted/amplified type-II genes and regulators of cell proliferation (STOP genes reduces cell proliferation and GO genes increases cell proliferation). The *P* values show the significance of overlap calculated based on 1000 random sets. (*D*) Network representation of 56 type-II genes frequently deleted (red edge) or amplified (blue edge) in nine different cancer types. The node size corresponds to the number of PubMed records associated with the gene.

we randomly choose a p fraction of edges to rewire, using the degree-preserving random rewiring algorithm (45); and (*ii*) edge flipping: we randomly choose a p fraction of edges to flip their directions. We tune p from 0 up to 1, resulting in a series of randomized networks.

Analysis of Cancer Genomic Datasets. Copy number alteration data for nine cancer types were downloaded from the cBioPortal for Cancer Genomics (www.cbioportal.org). Gene expression data for each cancer type were downloaded from TCGA (https://tcga-data.nci.nih.gov/tcga). 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. A gene is defined as amplified if the Genomic Identification of Significant Targets in Cancer (GISTIC) score is ≥ 1 and the z score is ≥ 1.5 and deleted if the GISTIC score (46) is ≤ -1 and the z score is ≤ -1.5 (*SI Text*).

ACKNOWLEDGMENTS. We thank M. W. Kirschner, S. E. Mohr, I. T. Flockhart, S. Rajagopal, and Bingbo Wang for helpful suggestions. This work was supported by National Institutes of Health (NIH) Grants P01-CA120964, Centers of Excellence of Genomic Science (CEGS) 1P50HG004233, and 1R01HL118455-01A1 and John Templeton Foundation Awards PFI-777 and 51977. N.P. is supported by the Howard Hughes Medical Institute.

1. Isidori A (1995) *Nonlinear Control Systems* (Springer, Berlin, New York), 3rd Ed.
2. Kalman RE (1963) Mathematical description of linear dynamical systems. *J Soc Indust Appl Math Ser A Control* 1(2):152–192.
3. Slotine JJE, Li W (1991) *Applied Nonlinear Control* (Prentice Hall, Englewood Cliffs, NJ).
4. Iglesias PA, Ingalls BP (2010) *Control Theory and Systems Biology* (MIT Press, Cambridge, MA).
5. Del Vecchio D, Murray RM (2015) *Biomolecular Feedback Systems* (Princeton Univ Press, Princeton).
6. Balázsi G, van Oudenaarden A, Collins JJ (2011) Cellular decision making and biological noise: From microbes to mammals. *Cell* 144(6):910–925.
7. Luenberger DG (1979) *Introduction to Dynamic Systems: Theory, Models, and Applications* (Wiley, New York).
8. Csete ME, Doyle JC (2002) Reverse engineering of biological complexity. *Science* 295(5560):1664–1669.
9. Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci USA* 100(21):11980–11985.
10. Callura JM, Cantor CR, Collins JJ (2012) Genetic switchboard for synthetic biology applications. *Proc Natl Acad Sci USA* 109(15):5850–5855.
11. Nepusz T, Vicsek T (2012) Controlling edge dynamics in complex networks. *Nat Phys* 8(7):568–573.
12. Kiel C, Yus E, Serrano L (2010) Engineering signal transduction pathways. *Cell* 140(1):33–47.
13. Vinayagam A, et al. (2011) A directed protein interaction network for investigating intracellular signal transduction. *Sci Signal* 4(189):rs8.
14. Liu YY, Slotine JJ, Barabási AL (2011) Controllability of complex networks. *Nature* 473(7346):167–173.
15. Hopcroft JE, Karp RM (1973) An n^{5/2} algorithm for maximum matchings in bipartite graphs. *SIAM J Comput* 2:225–231.
16. Ashburner M, et al.; The Gene Ontology Consortium (2000) Gene ontology: Tool for the unification of biology. *Nat Genet* 25(1):25–29.
17. Futreal PA, et al. (2004) A census of human cancer genes. *Nat Rev Cancer* 4(3):177–183.
18. Higgins ME, Claremont M, Major JE, Sander C, Lash AE (2007) CancerGenes: A gene selection resource for cancer genome projects. *Nucleic Acids Res* 35(Database issue): D721–D726.
19. Hindorff LA, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106(23):9362–9367.
20. McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: Potential next steps on a genetic journey. *Hum Mol Genet* 17(R2):R156–R165.
21. Brass AL, et al. (2008) Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 319(5865):921–926.
22. König R, et al. (2008) Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* 135(1):49–60.
23. Yeung ML, Houzet L, Yedavalli VS, Jeang KT (2009) A genome-wide short hairpin RNA screening of jurkat T-cells for human proteins contributing to productive HIV-1 replication. *J Biol Chem* 284(29):19463–19473.
24. Zhou H, et al. (2008) Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe* 4(5):495–504.
25. Goff SP (2008) Knockdown screens to knockout HIV-1. *Cell* 135(3):417–420.
26. Jäger S, et al. (2011) Global landscape of HIV-human protein complexes. *Nature* 481(7381):365–370.
27. Zanzoni A, et al. (2002) MINT: A Molecular INTeraction database. *FEBS Lett* 513(1): 135–140.
28. Navratil V, et al. (2009) VirHostNet: A knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Res* 37(Database issue):D661–D668.
29. Rozenblatt-Rosen O, et al. (2012) Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature* 487(7408):491–495.
30. Calderwood MA, et al. (2007) Epstein-Barr virus and virus human protein interaction maps. *Proc Natl Acad Sci USA* 104(18):7606–7611.
31. Knox C, et al. (2011) DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* 39(Database issue):D1035–D1041.
32. Hopkins AL, Groom CR (2002) The druggable genome. *Nat Rev Drug Discov* 1(9):727–730.
33. Jia T, et al. (2013) Emergence of bimodality in controlling complex networks. *Nat Commun* 4:2002.
34. Blume-Jensen P, Hunter T (2001) Oncogenic kinase signalling. *Nature* 411(6835):355–365.
35. Venkatesan K, et al. (2009) An empirical framework for binary interactome mapping. *Nat Methods* 6(1):83–90.
36. Gao J, et al. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 6(269):pl1.
37. Solimini NL, et al. (2012) Recurrent hemizygous deletions in cancers may optimize proliferative potential. *Science* 337(6090):104–109.
38. Wuchty S (2014) Controllability in protein interaction networks. *Proc Natl Acad Sci USA* 111(19):7156–7160.
39. Nacher JC, Akutsu T (2012) Dominating scale-free networks with variable scaling exponent: Heterogeneous networks are not difficult to control. *New J Phys* 14(7):073005.
40. Nacher JC, Akutsu T (2014) Analysis of critical and redundant nodes in controlling directed and undirected complex networks using dominating sets. *J Complex Netw* 2(4):394–412.
41. Rambaldi D, Giorgi FM, Capuani F, Ciliberto A, Ciccarelli FD (2008) Low duplicability and network fragility of cancer genes. *Trends Genet* 24(9):427–430.
42. Barabási AL, Gulbahce N, Loscalzo J (2011) Network medicine: A network-based approach to human disease. *Nat Rev Genet* 12(1):56–68.
43. Taylor IW, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* 27(2):199–204.
44. Jin Y, et al. (2012) A systems approach identifies HIPK2 as a key regulator of kidney fibrosis. *Nat Med* 18(4):580–588.
45. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. *Science* 296(5569):910–913.
46. Beroukhim R, et al. (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc Natl Acad Sci USA* 104(50): 20007–20012.
47. Sontag ED (1998) *Mathematical Control Theory: Deterministic Finite Dimensional Systems* (Springer, New York), 2nd Ed.
48. Sastry S (1999) *Nonlinear System: Analysis, Stability, and Control* (Springer, New York).
49. Lin C-T (1974) Structural controllability. *IEEE Trans Automat Contr* 19(3):201–208.
50. Luo H, Lin Y, Gao F, Zhang CT, Zhang R (2014) DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res* 42(Database issue):D574–D580.
51. Chen WH, Minguez P, Lercher MJ, Bork P (2012) OGEE: An online gene essentiality database. *Nucleic Acids Res* 40(Database issue):D901–D906.
52. Hu Y, et al. (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics* 12:357.
53. Hornbeck PV, et al. (2012) PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 40(Database issue):D261–D270.
54. Ben-Shlomo I, Yu Hsu S, Rauch R, Kowalski HW, Hsueh AJ (2003) Signaling receptome: A genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. *Sci STKE* 2003(187):RE9.
55. Park J, et al. (2005) Building a human kinase gene repository: Bioinformatics, molecular cloning, and functional validation. *Proc Natl Acad Sci USA* 102(23):8114–8119.
56. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* 298(5600):1912–1934.
57. Messina DN, Glasscock J, Gish W, Lovett M (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res* 14(10B):2041–2047.
58. Beck M, et al. (2011) The quantitative proteome of a human cell line. *Mol Syst Biol* 7:549.
59. Woodsmith J, Kamburov A, Stelzl U (2013) Dual coordination of post translational modifications in human protein networks. *PLOS Comput Biol* 9(3):e1002933.
60. Fazekas D, et al. (2013) SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Syst Biol* 7:7.
61. Shannon P, et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 13(11):2498–2504.
62. Doncheva NT, Assenov Y, Domingues FS, Albrecht M (2012) Topological analysis and interactive visualization of biological networks and protein structures. *Nat Protoc* 7(4):670–685.