

Cancer Genetic Network Inference Using Gaussian Graphical Models

Haitao Zhao^{1,2} and Zhong-Hui Duan^{1,2} 

¹Integrated Bioscience Program, The University of Akron, Akron, OH, USA.

²Department of Computer Science, The University of Akron, Akron, OH, USA.

Bioinformatics and Biology Insights
Volume 13: 1–9
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1177932219839402



ABSTRACT: The Cancer Genome Atlas (TCGA) provides a rich resource that can be used to understand how genes interact in cancer cells and has collected RNA-Seq gene expression data for many types of human cancer. However, mining the data to uncover the hidden gene-interaction patterns remains a challenge. Gaussian graphical model (GGM) is often used to learn genetic networks because it defines an undirected graphical structure, revealing the conditional dependences of genes. In this study, we focus on inferring gene interactions in 15 specific types of human cancer using RNA-Seq expression data and GGM with graphical lasso. We take advantage of the corresponding Kyoto Encyclopedia of Genes and Genomes pathway maps to define the subsets of related genes. RNA-Seq expression levels of the subsets of genes in solid cancerous tumor and normal tissues were extracted from TCGA. The gene expression data sets were cleaned and formatted, and the genetic network corresponding to each cancer type was then inferred using GGM with graphical lasso. The inferred networks reveal stable conditional dependences among the genes at the expression level and confirm the essential roles played by the genes that encode proteins involved in the two key signaling pathway phosphoinositide 3-kinase (PI3K)/AKT/mTOR and Ras/Raf/MEK/ERK in human carcinogenesis. These stable dependences elucidate the expression level interactions among the genes that are implicated in many different human cancers. The inferred genetic networks were examined to further identify and characterize a collection of gene interactions that are unique to cancer. The cross-cancer genetic interactions revealed from our study provide another set of knowledge for cancer biologists to propose strong hypotheses, so further biological investigations can be conducted effectively.

KEYWORDS: Computational biology, machine learning, network meta-analysis

RECEIVED: February 22, 2019. **ACCEPTED:** March 4, 2019.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work is partially supported by the Choose Ohio First for Bioinformatics Scholarship Program.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Zhong-Hui Duan, Department of Computer Science, The University of Akron, Akron, OH 44325, USA. Email: duan@uakron.edu

Introduction

Cancer is caused by genetic changes that alter normal cell behavior that leads to uncontrolled cell growth. Studying cancer genomics, identifying cancer-causing genes, and learning genetic networks could provide insights into understanding the biology of cancer and developing targeted drug and treatment of cancer.¹ Several studies have reported genome-wide mutational patterns in different types of cancer and identified the potential cancer-causing gene mutations.^{2–5} The Cancer Genome Atlas (TCGA) project has generated data sets of the key genomic changes in 33 different types of cancer and gives researchers unprecedented access to the cancer genomic data.^{6,7} These data sets, including genomic, transcriptomic, epigenomic, and clinical data, have been made publicly available by TCGA through its data portal. Cross-cancer gene alterations in 21 different cancer types were analyzed and reported.⁸ Despite the advances in cancer genomics, mining the biological significance hidden in the TCGA data sets remains to be a challenge. Many attempts from developing software tools to innovative algorithmic approaches have been made to facilitate the mining process.^{9–13}

Gaussian graphical model (GGM) as an analytics tool is often used to analyze gene interactions based on gene expression levels.^{14–21} For a multivariate random vector having a normal distribution, GGM defines an undirected graph structure through the precision matrix (inverse covariance matrix) which reveals the conditional dependences among variables. A node

in the graph represents a variable, and an edge implies the conditional dependence between the two variables incident to the edge. The expression levels of genes are generally considered to be log-normal because their distributions are typically skewed to the right.²² Gaussian graphical model provides an effective approach for learning genetic regulatory networks from log-transformed gene expression profiles. In the networks inferred from gene expression data sets, a node represents an expressed gene, and an edge denotes conditional dependence between the two expressed genes connected by the edge.

To estimate the inverse covariance matrix based on sample observation data, several algorithms have been reported. The underlying idea of the reported approaches is based on the maximum likelihood estimation. The traditional covariance selection with maximum likelihood estimation aims at identifying zero elements in the inverse covariance matrix; the standard algorithm for covariance selections is greedy forward and backward search.^{14,15} During greedy forward search, the initial set of edges is empty, and then the edges are iteratively added when the hypothesis testing reaches an indicated level α . The cost of this algorithm is extremely high, making it infeasible to estimate high-dimensional graphs.

An inferred genetic network is expected to be sparse and stable, and it should not change much with different sample data sets. In other words, the estimated inverse covariance matrix should only contain non-zero entries for pairs of genes that are highly related. To ensure the sparsity, the least absolute



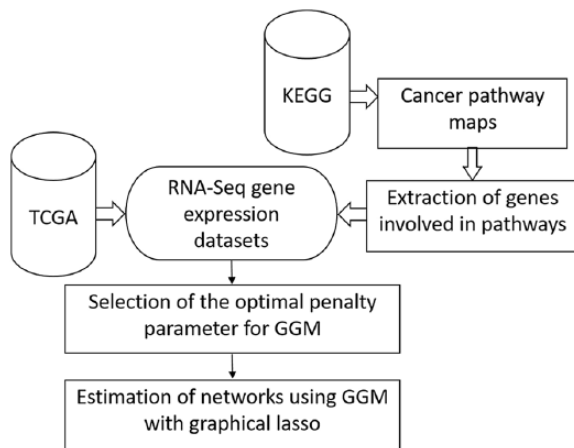


Figure 1. Pipeline of inferring gene-interaction networks.

shrinkage and selection operator (lasso) technique was introduced.¹⁶ Furthermore, several approaches were proposed to obtain sparse high-dimensional graphs using lasso with a penalty on the L_1 -norm of the precision matrix. Neighborhood-selection algorithm learns sparse high-dimensional Gaussian graphs by performing neighborhood selection for each node; it estimates each node separately by fitting linear lasso model.¹⁷ The essence of estimation procedure is equivalent to variable selection for Gaussian linear models. The value of a pair of variables i and j in inverse covariance matrix is estimated to be non-zero if both $p(i|j)$ and $p(j|i)$ are non-zero. To improve the performance of the algorithm and the stability of the estimated network, penalized likelihood methods were proposed, which minimize the negative log-likelihood function with L_1 penalty by taking advantage of the efficiency of the maxdet algorithm developed in convex optimization.^{18,23} To handle the computational challenge in high dimension, a model-selection method based on block coordinate descent was proposed.¹⁹ This approach converted the inverse covariance matrix finding problem to be a box-constrained quadratic problem that is then solved using interior point algorithm. The proposed method is equivalent to a recursive linear regression with L_1 penalty. Based on similar idea, a simple, but more efficient graphical lasso algorithm was proposed.²⁰ This graphical lasso algorithm cycles through the variables, fitting a modified lasso regression to each variable and the individual lasso problems are solved by coordinate descent. Despite all the efforts made to improve the computational performances, graphical lasso is still not efficient enough to be directly used to construct genome-scale networks, which contain tens of thousands genes.

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway repository is a collection of pathway maps; these pathway maps represent biologists' knowledge on the molecular interactions, reaction, and relation networks for metabolism, genetic information processing, human diseases, cellular processes, drug development, and so on.²⁴

In this study, we focus on inferring gene interactions in 15 specific types of human cancer using RNA-Seq expression level data and GGM with graphical lasso. We used the subsets of genes outlined in the corresponding KEGG cancer pathways.^{24,25} RNA-Seq expression levels of the subsets of genes in solid cancerous tumor and normal tissue were extracted from TCGA. The gene expression data sets were cleaned and formatted; and then the genetic network corresponding to each cancer type was inferred using GGM with graphical lasso. The inferred genetic networks were compared and examined to further identify a collection of cross-cancer gene interactions.

Materials and Methods

To infer genetic networks and identify cross-cancer gene interactions, genes presented in KEGG pathway maps and their RNA-Seq expression levels were extracted from KEGG and TCGA, respectively. We cleaned these data sets and inferred the genetic networks using GGM with graphical lasso. The pipeline for inferring these networks is illustrated in Figure 1 and the details of each step are described below.

RNA-Seq gene expression levels in 15 solid cancerous tissues as well as normal tissues were extracted from TCGA⁷ using the *R* script, TCGA-Assembler version 2.0.5.²⁶ The RNA-Seq data from primary solid tumor and the corresponding normal samples were retrieved. The expression profiles of 20531 genes were then preprocessed in three steps: (1) removal of 29 putative and retired genes (in the data sets, the IDs of these genes are present, but the symbols are absent); (2) remove the genes whose expression levels are zero across all samples. We note that it is possible that these genes did express but their levels were so low and were not picked up by RNA-Seq technology. There are about 300 such genes for each tissue type; (3) log-transformation of the expression data. We replaced 0 in the data sets by 1 before the transformation.

A total of 13 KEGG cancer pathway maps were downloaded from KEGG pathway repository.²⁵ Genes specific to each pathway were extracted to form the subset of genes for the pathway. The RNA-Seq expression levels in 15 solid cancerous tissues as well as normal tissues of the genes in each subset were extracted from TCGA (data sets not related to any KEGG cancer pathway were not considered).²⁵ Table 1 presents the list of the 15 cancer types/subtypes, their corresponding KEGG pathway IDs, and the number of available TCGA gene expression data sets. The normal gene expression data sets from a specific organ were pooled to form one normal tissue data set for that organ; the pool data sets include one for kidney renal clear cell carcinoma (KIRC) and kidney renal papillary cell carcinoma (KIRP) and one for lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). In addition, the normal data sets for brain lower grade glioma (LGG) and skin cutaneous melanoma (SKCM) are not available, and the normal data set for pancreatic adenocarcinoma (PAAD) was not used because the data set with a sample size of three is too

Table 1. The 15 cancer types and their KEGG and TCGA IDs as well as the number of TCGA data sets.

CANCER TYPE	KEGG ID	KEGG CANCER TYPE	TCGA CANCER TYPE	TCGA SAMPLE SIZE	
				CANCER	NORMAL
BLCA	hsa05219	Bladder cancer	Bladder urothelial carcinoma	408	19
BRCA	hsa05224	Breast cancer	Breast invasive carcinoma	1094	112
COAD	hsa05210	Colorectal cancer	Colon adenocarcinoma	284	40
KIRC	hsa05211	Renal cell carcinoma	Kidney renal clear cell carcinoma	533	72
KIRP	hsa05211	Renal cell carcinoma	Kidney renal papillary cell carcinoma	290	32
LGG	hsa05214	Glioma	Brain lower grade glioma	515	0
LIHC	hsa05225	Hepatocellular carcinoma	Liver hepatocellular carcinoma	371	50
LUAD	hsa05223	Non-small-cell lung cancer	Lung adenocarcinoma	515	59
LUSC	hsa05223	Non-small-cell lung cancer	Lung squamous cell carcinoma	502	50
PAAD	hsa05212	Pancreatic cancer	Pancreatic adenocarcinoma	178	3
PRAD	hsa05215	Prostate cancer	Prostate adenocarcinoma	496	51
SKCM	hsa05218	Melanoma	Skin cutaneous melanoma	102	0
STAD	hsa05226	Gastric cancer	Stomach adenocarcinoma	415	35
THCA	hsa05216	Thyroid cancer	Thyroid carcinoma	504	59
UCEC	hsa05213	Endometrial cancer	Uterine corpus endometrial carcinoma	176	23

Abbreviations: BLCA, bladder carcinoma; BRCA, breast cancer; COAD, colon adenocarcinoma; KEGG, Kyoto Encyclopedia of Genes and Genomes; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, lower grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PAAD, pancreatic adenocarcinoma; PRAD, prostate adenocarcinoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; TCGA: The Cancer Genome Atlas; THCA, thyroid cancer; UCEC, uterine corpus endometrial carcinoma.

small to construct a reliable network. The data sets were then preprocessed and log transformed, resulting in 15 RNA-Seq expression data sets from solid cancerous tissues and 10 data sets from normal tissues.

To ensure that the log-transformed gene expression profiles are normal or at least close to normal, we calculated Ryan-Joiner (RJ) statistic measures, representing correlations between the expression levels of genes and the corresponding normal distribution scores. The calculation was performed on one gene at a time; effectively, it was testing how well the log-transformed conditional distribution of a gene correlates to a normal distribution. We observed a significant variability in the RJ statistic measures. The correlation coefficient varies from one gene to another. The weakest is between CDK4 and normal with the correlation coefficient 0.956. The strongest is between MAPK3 and normal with JR score 0.999. Statistically speaking, the expression level of CDK4 is considered not normal despite the RJ score 0.956; the distribution is still slightly skewed to the right (some patients have relative high CDK4 levels). Nevertheless, the statistics indicates that normal distribution provides a fairly good approximation to the distribution of gene expression levels.

Given gene expression data $Y = [y_1, \dots, y_n]^T \in \mathbb{R}^{n \times p}$ for n samples and p genes, the gene expression profile of each

sample, $y_i = [y_i^1, \dots, y_i^p]^T$, is assumed to be independent and follow a Gaussian distribution $N(\mathbf{u}, \Sigma)$, where \mathbf{u} is the mean and Σ is the $p \times p$ covariance matrix. The precision matrix $\Theta = \Sigma^{-1}$ is a positive definite and symmetric matrix and presents a model for an undirected graph $G = (V, E)$ where V is a set of p vertices corresponding to the p genes, and the edge set $E = \{e_{i,j}\}$ describes the conditional dependences among the p genes. $e_{i,j} = 1$ indicates that genes i and j are conditionally dependent, whereas $e_{i,j} = 0$ states the two genes i and j are conditionally independent of each other. Each entry $\theta_{i,j}$ of the precision matrix signifies the strength of the dependence relation. Therefore, learning genetic network is equivalent to estimating the precision matrix Θ , that is, to maximize the log-likelihood with L_1 norm penalty on its precision matrix Θ :

$$\log \det(\Theta) - \text{trace}(S\Theta) - \rho \|\Theta\|_1 \quad (1)$$

where S is sample covariance matrix, ρ is non-negative penalty parameter which controls the sparsity of the inverse covariance matrix Θ and $\|\Theta\|_1 = \sum_i \sum_j |\theta_{i,j}|$ represents the L_1 -norm of Θ .^{19,20} Clearly, the larger the parameter ρ is, the sparser the estimated Θ would be. If $\rho = 0$, this problem is

reduced to the typically maximum likelihood estimation problem, while when $\rho \rightarrow \infty$, $\Theta = 0$ regardless what sample data sets are used in estimation. In this study, the graphical lasso procedure is deployed to estimate the sparse precision matrices corresponding to sparse gene regulatory networks.²⁰ The glasso software version 1.10 was used.²⁷

To apply graphical lasso to infer genetic networks, one important issue is to choose the optimal penalty parameter ρ , which controls the sparsity level of the estimated Θ and ensures its stability. Any network to be learned from experimental data unavoidably could include some irrelevant and unexpected interactions resulting from the intrinsic “noise” in the experimental data. We expect an estimated network robust with respect to different sample data. Therefore, models with certain stability require ρ to be at a level so that the “noisy” edges in the estimated precision matrix are filtered out. Furthermore, genetic networks are typically considered sparse, and therefore, we expect that the estimated networks to be sparse as well, and the edges represent the true dependences of the genes.^{24,28,29} To select an optimal value of the penalty parameter ρ , we implemented and tested the subsampling-based approach.³⁰

Results

We inferred 15 cancer networks and 10 normal networks using TCGA RNA-Seq gene expression data sets and GGM with graphical lasso. During the process of penalty parameter selection, a P -value of .05 was used. The typical optimal value of ρ is found to be around 0.03. The edge/node ratio of KEGG networks ranges from 0.8 to 1.2 with an average of 1.01, reflecting the sparsity of these networks. The initial learned GGM networks are much denser despite the application of the graphical lasso. We believe that the edges with extremely low weights are false positive. There are possibly two types of error coming into play during the modeling process: (1) noise in the RNA-Seq data due to variations in patients and experiments and (2) gene expression profiles are not perfect normal (some distributions, such as the one of CDK4, are only approximately normal). As a result, inaccuracies from GGM also contribute to the noise. To reduce the effects of the noise in the data, to strengthen the network stability, and to highlight the most important dependences between genes, we used a threshold of 0.2 to reduce false positives. The estimated network edges with weights below a threshold of 0.2, that is, the corresponding entries in the precision matrix <0.2 , were removed. The resulting edge/node ratio is approximately 5.81.

We constructed a map of the gene interactions in 15 types of human cancer and a map of the gene interactions in 10 normal human tissue types. These two maps encapsulating all 25 networks are shown in the Supplemental material Tables S1 and S2. Through comparison and analysis of these derived cancer and normal networks, we identified the cross-cancer gene

Table 2. Cross-cancer gene interactions that are present in cancer but absent in all normal networks.

CONSENSUS GENE INTERACTIONS	NUMBER OF SHARED CANCER NETWORKS
BAD-MAP2K2	9
CDKN1A-DDB2	9
PIK3CA-PIK3CB	9
MAP2K1-MAPK1	8
PIK3R1-POLK	8
ARAF-E2F3	7
GRB2-PIK3CD	7
POLK-RB1	7
SOS1-SOS2	7
BAD-SOS1	6
GSK3B-PIK3CA	6
HRAS-SOS2	6
KRAS-MAP2K2	6
MAPK1-NRAS	6
AKT2-BAX	5
BAD-KRAS	5
BAD-MAPK3	5
BAK1-CDKN1A	5
E2F1-RB1	5
E2F3-MAPK3	5
GADD45B-MAP2K2	5
GSK3B-PIK3CB	5
HRAS-SOS1	5
MAPK1-MTOR	5
MAPK1-PIK3CA	5
MTOR-PIK3R3	5
PIK3R1-SOS2	5

interactions that were altered in cancerous tissues. The consensus interactions that are unique to cancer networks are presented in Table 2 and Figures 2 and 3. These interactions are shared by at least five cancer networks but absent in all normal networks. The gene interactions that are mostly unique to cancer networks, that is, they appear in at least five cancer networks but also in one normal network, are shown in Table 3 and Figures 4 and 5. Figure 6 integrates the two networks shown in Figures 2 and 4 to illustrate both sets of strong gene interactions presented in cancer.

Discussion

Gene expression is an extremely complicated process during which multiple genes and/or proteins interact in a coordinated way and directly or indirectly control each other's expression levels. Highlighting the dependences of the coordination helps to elucidate the complicated process. In this study, we identified and analyzed the cross-cancer gene interactions hidden in

the gene expression data sets and inferred 15 cancer networks and 10 normal networks using TCGA RNA-Seq gene expression levels and GGM. We focused on the subsets of genes presented in the 13 corresponding KEGG cancer pathways. We uncovered the gene interactions shared among the cancer networks and analyzed these cross-cancer gene interactions and created a map of the altered gene interactions in various types of cancer.

As illustrated in Table 2, Figures 2 and 3, significant numbers of the cross-cancer interactions are closely related to the signaling molecules on the two critical signaling pathways: phosphoinositide 3-kinase (PI3K)/AKT/mTOR pathway and Ras/Raf/MEK/ERK pathway. The *PI3K* pathway is an intracellular signaling pathway that plays key roles in regulating cell cycle and is linked to many essential cellular processes such as cell proliferation, survival, growth, and motility. The signaling cascade is mediated through serine and/or threonine phosphorylation of a range of downstream molecules. The key proteins involved include PI3K, AKT, GSK3, BAD, BAX, and CDKN1A. It has been widely reported that the PI3K pathway is overactive in many cancers, thus reducing apoptosis and allowing proliferation and uncontrolled cell growth.^{31–35} In addition, the alterations of PI3Ks in cancer were detailed along with the therapeutic efficacy of PI3K inhibitors in the cancer treatment.³⁵

Our results show that genes that encode several signaling proteins on the PI3K pathway, AKT2 and BAX, GRB2 and PIK3CD, GSK3B and PIK3CA, GSK3B and PIK3CB, MAPK1 and PIK3CA, MTOR and PIK3R3, PIK3CA and PIK3CB, PIK3R1 and POLK, and PIK3R1 and SOS2, are conditionally dependent on each other at their expression

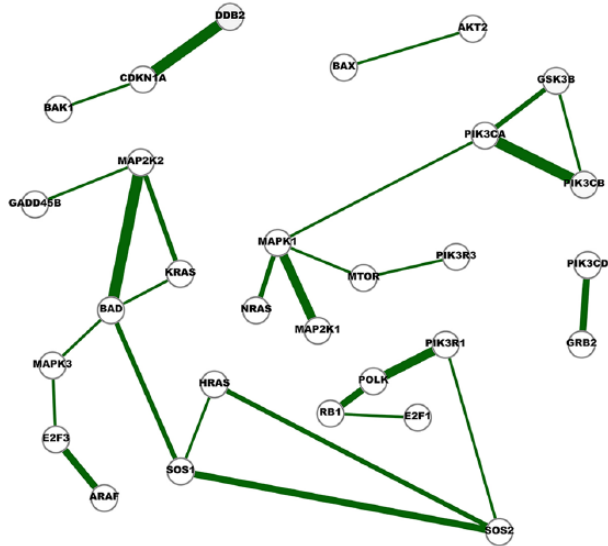


Figure 2. A network of cross-cancer gene interactions that are unique to the inferred cancer networks. A node in the network represents a gene, and an edge indicates the conditional dependence of the two incident genes. The conditional dependence depicts the interaction of the genes at the expression level. The thickness of an edge represents the degree of consensus of the interaction among the cancer networks. The edges in this network are shared by at least five cancer networks but absent in all normal networks.

Cross-cancer gene interactions across various cancer networks (not in any one of normal networks)

	BLCA	BRCA	COAD	KIRC	KIRP	LGG	LHC	LUAD	LUSC	PAAD	PRAD	SKCM	STAD	THCA	UCEC
BAD-MAP2K2	0	0	0.6288	0.9592	1.765	0	0.5335	0.7837	0.2684	0	1.603	1.328	0	0	0.6993
CDKN1A-DDB2	0	0.723	0.5915	0	0	0.2955	0.5555	0.4055	0	0	0	0.235	0.5088	0.5348	0.3446
PIK3CA-PIK3CB	0	0.323	0.6924	0.3093	0.3572	0	0.5037	0.6874	0	0	0.2649	0.8458	0.9734	0	0
MAP2K1-MAPK1	0.6481	0.4776	0.4885	0	0	0.6024	0.2505	0	0	0	0.7773	0	0	0.3396	0.4559
PIK3R1-POLK	0	0.4892	0.6066	0	0	0	0.525	0.7849	0.2488	0.4196	0	0.9163	0	0	0.2868
ARAF-E2F3	0.3513	0.5084	0	0	0	0	0.4644	0.4581	0	0	0.2559	0.2974	0.228	0	0
GRB2-PIK3CD	0	0.2563	0.4838	0.7489	0.4169	0.3763	0	0.3191	0	0	0	0	0.4483	0	0
POLK-RB1	0	0.4888	0	0	0	0.2559	0.2698	0.3084	0.2992	0.316	0	0.4115	0	0	0
SOS1-SOS2	0	0	0.2242	0	0	0.606	0.5391	0.5201	0	0	0.5254	0	0.7808	0	0.4269
BAD-SOS1	0	0	0.9694	0.6614	0	0	0	0.7239	0.5113	0	0.3625	0	0	0	0.337
GSK3B-PIK3CA	0	0.5287	0.3691	0	0	0	0.7672	0	0	0	0.2713	0	0.6633	0	0.7464
HRAS-SOS2	0	0.4265	0.5085	0	0	0.2221	0.3191	0	0.4083	0	0.219	0	0	0	0
KRAS-MAP2K2	0.3543	0.32	0	0.2762	0.5682	0	0	0	0	0	0.662	0.4121	0	0	0
MAPK1-NRAS	0.5193	0	0.2541	0.2632	0	0	0.4352	0	0	0	0.2961	0	0.2068	0	0
AKT2-BAX	0	0.4852	0	0	0	0.5071	0	0.5789	0	0	0	0.3393	0.6169	0	0
BAD-KRAS	0	0	0	0.573	0	0	0.5103	0.3167	0	0	0.2239	0.699	0	0	0
BAD-MAPK3	0	0	0.34	0.2244	0	0	0	0.3103	0	0	0.259	0.3356	0	0	0
BAK1-CDKN1A	0	0	0.4098	0	0	0	0	0.2566	0.4779	0.3644	0	0	0.3484	0	0
E2F1-RB1	0	0.4477	0	0	0	0.2122	0	0.2568	0.3337	0.2131	0	0	0	0	0
E2F3-MAPK3	0	0.3514	0	0	0	0.2199	0	0.4263	0	0	0	0.6734	0.5894	0	0
GADD45B-MAP2K2	0	0.4808	0	0	0	0	0.213	0.3139	0	0	0	0.6197	0	0	0.3266
GSK3B-PIK3CB	0	1.57	0.7294	0	0	0	1.299	0	0	0	0	0	1.344	0	0.3003
HRAS-SOS1	0	0.4492	0.4401	0	0	0.8782	0	0	0	0	0.2576	0	0	0	0.3505
MAPK1-MTOR	0	0.3212	0.4159	0	0	0.2655	0	0	0	0.3667	0	0	0.4616	0	0
MAPK1-PIK3CA	0	0	0	0	0	0	0.8231	0.2101	0.3102	0.2887	0.822	0	0	0	0
MTOR-PIK3R3	0	0.346	0.4684	0	0	0.3471	0	0	0	0.5222	0	0	0.4336	0	0
PIK3R1-SOS2	0	0.2531	0	0	0	0.2578	0.4252	0.4492	0	0	0.3201	0	0	0	0

Figure 3. The map of consensus gene interactions that are appeared in at least 5 of the 15 cancer networks but not present in any normal network. This map depicts a portion of the precision matrices, indicating the conditional dependence between a pair of genes in a specific cancer. The higher the value in the map is, the stronger the conditional dependence (interaction) of the pair of genes.

Table 3. Cross-cancer interactions mostly unique to the inferred cancer networks (appear in at least five cancer networks but also in one normal network).

CONSENSUS GENE INTERACTIONS	NUMBER OF SHARED CANCER NETWORKS
BRAF-PIK3CA	13
BAX-DDB2	9
MAP2K2-PIK3R2	9
BAD-HRAS	8
CDK4-E2F1	7
BAX-MAP2K2	6
BAX-PIK3CA	6
BRAF-HRAS	6
PIK3R1-PIK3R3	6
BAX-CDKN1A	5
HRAS-KRAS	5
HRAS-POLK	5
MAP2K2-PIK3CA	5
PIK3CA-SOS2	5
TGFB1-TGFB3	5

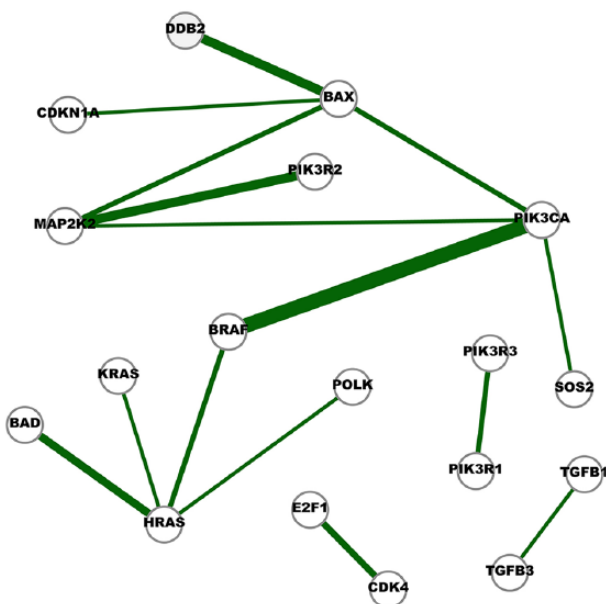


Figure 4. A network of cross-cancer interactions that are mostly unique to the inferred cancer networks.

A node in the network represents a gene and an edge indicates the conditional dependence of the two incident genes. The conditional dependence depicts the interaction of the genes at the expression level. The thickness of an edge represent the degree of consensus of the interaction among the cancer networks. Edges in this network represent the interactions identified in at least five cancer networks but also appeared in one normal network.

levels in at least five different types of human cancer, and these interactions are unique to cancers and do not appear to be significant in any normal organ tissue (Table 2, Figures 2 and 3). DNA polymerase kappa (POLK) encodes a specialized DNA polymerase that catalyzes translesion DNA synthesis, which allows DNA replication in the presence of DNA damages.³⁶ Although few reports find that POLK is linked to PIK3R1 (a gene that provides instructions for making a regulatory subunit of PI3K), our results reveal that the expression level of POLK is conditionally dependent on that of PIK3R1 in eight different cancers, and the dependence is particularly strong in SKCM (Figure 3). One possible connection between PI3Ks and POLK is through the PI3K downstream transcription factor CREB, which is reported to be a regulator of *POLK* promoter activity.³⁶ Furthermore, our results indicate that the expression levels of POLK and the well-known tumor suppressor gene RB1 are conditionally linked to seven cancers (Figure 3).

Ras/Raf/MEK/ERK pathway is another key intracellular signaling pathway.^{37–42} Dysregulation of this pathway is a common event in cancer as RAS family, small guanosine triphosphatases (GTPases), is often the most frequently mutated oncogene in human cancer.⁴² Our study shows many identified cross-cancer gene interactions that are unique to cancer are linked to this pathway, including ARAF-E2F3, BAD-KRAS, BAD-MAP2K2, BAD-MAPK3, BAD-SOS1, E2F3-MAPK3, GADD45B-MAP2K2, HRAS-SOS1, HRAS-SOS2, KRAS-MAP2K2, MAP2K1-MAPK1, MAPK1-MTOR, MAPK1-NRAS, MAPK1-PIK3CA, and SOS1-SOS2. Ras/Raf/MEK/ERK signaling cascade transmits signals from upstream receptors such as epidermal growth factor receptor (EGFR) to transcription factors such as CREB and E2F3, which in turn regulate gene expression and prevent apoptosis. Key molecules involved in this cascade are GRAB2, SOS, RAS, RAF, MEK, and MAPK. This study reveals how the expression levels of the genes that encode these proteins conditionally depend on each other and on other genes such as BAD, a pro-apoptotic member of the Bcl-2 gene family involved in initiating apoptosis, and E2F3, a transcription factor which is key to cellular proliferation and differentiation.^{42–45} Figure 3 shows that the gene expression levels of MAP2K2 is conditionally dependent on the levels of BAD in nine cancer networks, and the conditional dependence is particularly strong in three different cancers, KIRC, KIRP, and prostate adenocarcinoma (PRAD). Transcription factor E2Fs are key regulators of cell cycle progression and share a critical role in tissue homeostasis. Our results reiterate the critical roles MAPK signaling pathway plays in cell proliferation, differentiation, cell movement, and apoptosis.

Furthermore, this study elucidates several cross-talks between the PI3K/AKT/mTOR pathway and Ras/Raf/MEK/ERK pathway, particularly through the gene that encodes BRAF, a proto-oncogene playing major roles in human

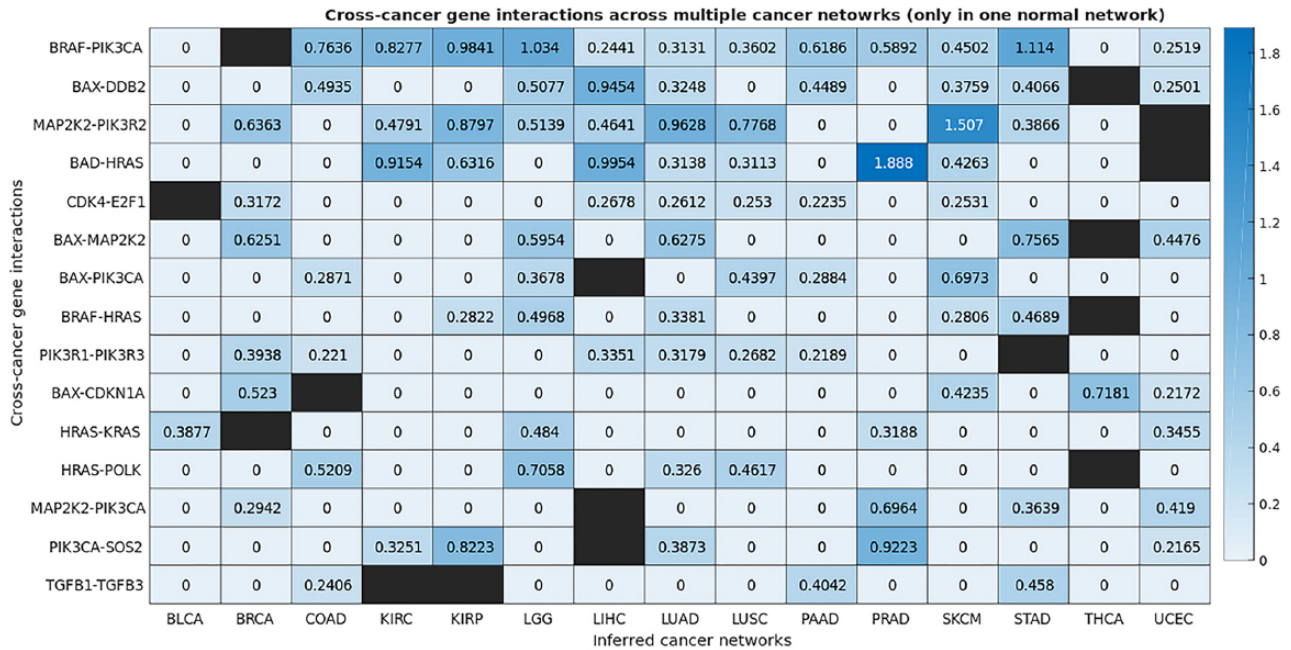


Figure 5. The map of consensus gene interactions that are appeared in at least 5 of the 15 cancer networks; these interactions are also present in one normal network. The higher the value in the map is, the stronger the conditional dependence (interaction) of the pair of genes. A black square denotes the interaction is also appeared in the normal network of the corresponding organ.

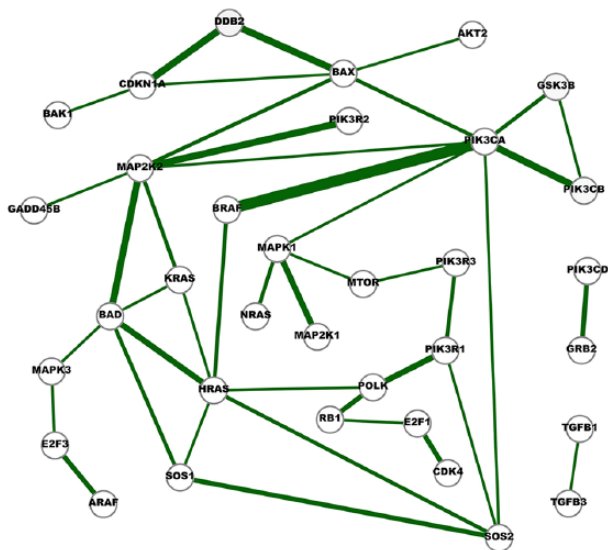


Figure 6. A network of strong cross-cancer interactions. A node in the network represents a gene and an edge indicates the conditional dependence of the two incident genes. The conditional dependence depicts the interaction of the genes at the expression level. The thickness of an edge represents the degree of consensus of the interaction among the cancer networks. Edges in this network represent the interactions identified in at least five cancer networks, and they are either absent in any normal network or appear in just one normal network.

carcinogenesis,⁴⁶ and the ones that code the downstream kinase mTOR and apoptosis regulator BAX.^{47,48}

The almost unique consensus network shown in Figures 4 and 5 supports the observations made from Figures 2 and 3. The PI3K-pathway-related interactions presented in Figure 4 include BAX-PIK3CA, BRAF-PIK3CA, MAP2K2-PIK3CA, MAP2K2-PIK3R2, PIK3CA-SOS2, and PIK3R1-PIK3R3,

whereas BAX-MAP2K2, BAD-HRAS, BRAF-HRAS, HRAS-KRAS, HRAS-POLK, MAP2K2-PIK3CA, and MAP2K2-PIK3R2 are associated with the MEK/ERK pathway.

Figure 6 integrates the two networks presented in Figures 2 and 4. Six network motifs (small cliques) can be observed in the network, including BAD-HRAS-KRAS, BAD-HRAS-SOS1, BAD-KRAS-MAP2K2, BAX-DDB2-CDKN1A, BAX-MAP2K2-PIK3CA, and GSK3B-PIK3CA-PIK3CB. These six cliques reveal the close dependences among the genes. GSK3B-PIK3CA-PIK3CB is linked with PI3K pathway, BAD-HRAS-SOS1, BAD-KRAS-MAP2K2 and HRAS-SOS1-SOS2, are associated with MEK/ERK pathway, and BAX-MAP2K2-PIK3CA is linked with both pathways. In addition, our study identifies a network motif that is closely linked with cell apoptosis and cell cycle arrest, BAX-CDKN1A-DDB2. It has been reported all three molecules in this clique are regulated by the most well-known tumor suppressor gene p53. The cyclin-dependent kinase inhibitor 1 (p21), encoded by the CDKN1A gene, is a critical player in cell cycle arrests at various checkpoints after DNA damage.^{49–51} The coordination of CDKN1A and DDB2 in the cellular response to ultraviolet (UV) radiation has also been reported.⁵² Our study shows the expression levels of the apoptosis regulator BAX and the two cell cycle arrest regulator CDKN1A and DDB2 are conditionally dependent of each other in cancer.

Conclusions

This study provides rich insights for identifying and analyzing the cross-cancer gene interactions hidden in their expression

levels as well as how the interactions are connected in networks. A total of 15 cancer gene interaction networks and 10 normal networks were constructed using TCGA RNA-Seq gene expression data and GGM with graphical lasso. The inferred networks reveal conditional dependences among the genes, and the weights of edges indicate the strength of the dependences. These networks confirm the essential roles played by genes that encode proteins involved in the two key signaling pathway PI3K/AKT/mTOR and Ras/Raf/MEK/ERK in human carcinogenesis. The stable conditional dependences presented in the networks elucidate the expression level interactions among the genes that are implicated in many different human cancers.

The genetic networks constructed in this study are based on RNA-Seq expression data and reveal conditional dependences among the genes at the expression level. We note that these dependences may not indicate the proteins encoded by these genes have direct interactions; in some cases, they belong to different cellular components. In signaling pathways such as PI3K/AKT/mTOR and Ras/Raf/MEK/ERK, the most important event is phosphorylation cascade. One enzyme phosphorylates another may not signify the expression level dependences between the two genes that code the two enzymes. Nevertheless, the expression level dependences of signaling protein-coding genes reveal, to a certain degree, the dynamic nature of phosphorylation cascades because, as concluded by a recent study, on the bulk level and for approximate steady-state conditions, protein levels are largely determined by their transcript concentrations.⁵³ The cross-cancer gene interactions highlighted in the results derived from the expression levels provide another set of knowledge for cancer biologists to propose strong hypotheses so further biological investigations can be conducted effectively.

Acknowledgements

The authors thank the reviewers of the manuscript for their insightful thoughts and comments which have strengthened their report.


Author Contributions

This study was conceptualized by HZ and ZHD. HZ conducted data acquisition, model development, and data collection. HZ and ZHD performed data analysis and completed the writing of the manuscript. All authors read and approved the final version of the manuscript.

Supplemental Material

Supplemental material for this article is available online.

ORCID iD

Zhong-Hui Duan  <https://orcid.org/0000-0001-6561-0991>

REFERENCES

1. Garraway L, Lander E. Lessons from the cancer genome. *Cell*. 2013;153:17–37.

2. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500:415–421.
3. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*. 2013;3:246–259.
4. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep*. 2013;3:2650.
5. Abaan OD, Polley EC, Davis SR, et al. The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res*. 2013;73:4372–4382.
6. Weinstein JN, Collisson EA, Mills GB, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*. 2013;45:1113–1120.
7. Genomic data commons data portal. Website: <https://portal.gdc.cancer.gov/>. Accessed February 1, 2018.
8. Peng L, Bian XW, Li DK, et al. Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. *Sci Rep*. 2015;5:13413.
9. Chandran UR, Medvedeva OP, Barmada MM, et al. TCGA expedition: a data acquisition and management system for TCGA data. *PLoS ONE*. 2016;11:e0165395.
10. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. 2013;6:p11.
11. Koch A, De Meyer T, Jeschke J, Van Criekinge W. MEXPRESS: visualizing expression, DNA methylation and clinical TCGA data. *BMC Genomics*. 2015;16:636.
12. Zhang Q, Burdette JE, Wang JP. Integrative network analysis of TCGA data for ovarian cancer. *BMC Syst Biol*. 2014;8:1338.
13. Deeter A, Dalman M, Haddad J, Duan ZH. Inferring gene and protein interactions using PubMed citations and consensus Bayesian networks. *PLoS ONE*. 2017;12:e0186004.
14. Lauritzen S. *Graphical Models*. Oxford, UK: Oxford University Press; 1996.
15. Edwards D. *Introduction to Graphical Modelling*. 2nd ed. New York, NY: Springer; 2000.
16. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc Ser B*. 1996;58:267–288.
17. Meinshausen N, Bühlmann P. High-dimensional graphs and variables selection with the lasso. *Ann Stat*. 2006;34:1436–1462.
18. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*. 2007;94:19–35.
19. Banerjee O, Ghaoui LE, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J Mach Learn Res*. 2008;9:485–516.
20. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*. 2008;9:432–441.
21. Zhang L, Kim S. Learning gene networks under SNP perturbations using eQTL datasets. *PLoS Comput Biol*. 2014;10:e1003420.
22. Limpert E, Stahel WA, Abbt M. Log-normal distributions across the sciences: keys and clues. *Bioscience*. 2001;51:341–352.
23. Vandenberghe L, Boyd S, Wu SP. Determinant maximization with linear matrix inequality constraints. *SIAM J Matrix Anal Appl*. 1998;19:499–533.
24. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. 2016;44:D457–D462.
25. KEGG: Kyoto Encyclopedia of Genes and Genomes. Website: <https://www.genome.jp/kegg/>. Accessed February 1, 2018.
26. Wei L, Jin Z, Yang S, Xu Y, Zhu Y, Ji Y. TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*. 2018;34:1615–1617.
27. glasso: graphical lasso: estimation of Gaussian graphical models. Website: <https://cran.r-project.org/web/packages/glasso/index.html>.
28. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet*. 2004;5:101–113.
29. Leclerc RD. Survival of the sparsest: robust gene networks are parsimonious. *Mol Syst Biol*. 2008;4:213.
30. Liu H, Roeder K, Wasserman L. *Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models*, Vol. 2. New York, NY: Curran Associates, Inc.; 2010:1432–1440.
31. Luo J, Manning BD, Cantley LC. Targeting the PI3K-Akt pathway in human cancer: rationale and promise. *Cancer Cell*. 2003;4:257–262.
32. Liu P, Cheng H, Roberts TM, Zhao JJ. Targeting the phosphoinositide 3-kinase pathway in cancer. *Nat Rev Drug Discov*. 2009;8:627–644.
33. Janku F, Yap TA, Meric-Bernstam F. Targeting the PI3K pathway in cancer: are we making headway? *Nat Rev Clin Oncol*. 2018;15:273–291.
34. Samuels Y, Wang Z, Bardelli A, et al. High frequency of mutations of the PIK3CA gene in human cancers. *Science*. 2004;304:554.

35. Thorpe LM, Yuzugullu H, Zhao JJ. PI3K in cancer: divergent roles of isoforms, modes of activation, and therapeutic targeting. *Nat Rev Cancer*. 2015;15:7–24.
36. Barnes R, Eckert K. Maintenance of genome integrity: how mammalian cells orchestrate genome duplication by coordinating replicative and specialized DNA polymerases. *Genes (Basel)*. 2017;8:19.
37. Chang L, Karin M. Mammalian MAP kinase signalling cascades. *Nature*. 2001;410:37–40.
38. Molina JR, Adjei AA. The Ras/Raf/MAPK pathway. *J Thoracic Oncol*. 2006;1:7–9.
39. Roberts PJ, Der CJ. Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene*. 2007;26:3291–3310.
40. Pylayeva-Gupta Y, Grabocka E, Bar-Sagi D. RAS oncogenes: weaving a tumorigenic web. *Nat Rev Cancer*. 2011;11:761–774.
41. Caunt CJ, Sale MJ, Smith PD, Cook SJ. MEK1 and MEK2 inhibitors and cancer therapy: the long and winding road. *Nat Rev Cancer*. 2015;15:577–592.
42. Bonni A, Brunet A, West AE, Datta SR, Takasu MA, Greenberg ME. Cell survival promoted by the Ras-MAPK signaling pathway by transcription-dependent and -independent mechanisms. *Science*. 1999;286:1358–1362.
43. Humbert PO, Verona R, Trimarchi JM, Rogers C, Dandapani S, Lees JA. E2f3 is critical for normal cellular proliferation. *Genes Dev*. 2000;14:690–703.
44. Attwooll C, Denchi EL, Helin K. The E2F family: specific functions and overlapping interests. *EMBO J*. 2004;23:4709–4716.
45. Iglesias-Ara A, Zenarruzabeitia O, Buelta L, Merino J, Zubiaga AM. E2F1 and E2F2 prevent replicative stress and subsequent p53-dependent organ involution. *Cell Death Differ*. 2015;22:1577–1589.
46. Davies H, Bignell GR, Cox C, et al. Mutations of the BRAF gene in human cancer. *Nature*. 2002;417:949–954.
47. Carracedo A, Ma L, Teruya-Feldstein J, et al. Inhibition of mTORC1 leads to MAPK pathway activation through a PI3K-dependent feedback loop in human cancer. *J Clin Invest*. 2008;118:3065–3074.
48. Tsuruta F, Masuyama N, Gotoh Y. The phosphatidylinositol 3-Kinase (PI3K)-Akt pathway suppresses Bax translocation to mitochondria. *J Biol Chem*. 2002;277:14040–14047.
49. Waldman T, Kinzler KW, Vogelstein B. p21 is necessary for the p53-mediated G1 arrest in human cancer cells. *Cancer Res*. 1995;55:5187–5190.
50. Bunz F, Dutriaux A, Lengauer C, et al. Requirement for p53 and p21 to sustain G2 arrest after DNA damage. *Science*. 1998;282:1497–1501.
51. Löhr K, Möritz C, Contente A, Döbelstein M. p21/CDKN1A mediates negative regulation of transcription by p53. *J Biol Chem*. 2003;278:32507–32516.
52. Li H, Zhang X-P, Liu F. Coordination between p21 and DDB2 in the cellular response to UV radiation. *PLoS ONE*. 2013;8:e80111.
53. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. *Cell*. 2016;165:535–550.