Quantitative Biology 2020, 8(2): 130–142
https://doi.org/10.1007/s40484-019-0185-6

RESEARCH ARTICLE

A pan-cancer integrative pathway analysis of multi-omics data

Henry Linder, Yuping Zhang*

Department of Statistics, University of Connecticut, Storrs, CT 06269, USA
* Correspondence: yuping.zhang@uconn.edu

Received July 4, 2019; Revised September 11, 2019; Accepted September 14, 2019

Background: **Multi-view -omics datasets offer rich opportunities for integrative analysis across genomic, transcriptomic, and epigenetic data platforms. Statistical methods are needed to rigorously implement current research on functional biology, matching the complex dynamics of systems genomic datasets.**
Methods: **We apply imputation for missing data and a structural, graph-theoretic pathway model to a dataset of 22 cancers across 173 signaling pathways. Our pathway model integrates multiple data platforms, and we test for differential activation between cancerous tumor and healthy tissue populations.**
Results: **Our pathway analysis reveals significant disturbance in signaling pathways that are known to relate to oncogenesis. We identify several pathways that suggest new research directions, including the Trk signaling and focal adhesion kinase activation pathways in sarcoma.**
Conclusions: **Our integrative analysis confirms contemporary research findings, which supports the validity of our findings. We implement an interactive data visualization for exploration of the pathway analyses, which is available online for public access.**

Keywords: multi-platform data integration; pathway analysis; imputation; cancer genomics; data visualization

Author summary: Genomic Big Data is now collected across multiple experimental platforms as a matter of course. These data offer multiple unique perspectives of the human genome and its processes, with great potential to improve our understanding of complex diseases such as cancer. In this paper, we apply a statistical model of the biological structure of genetic signaling pathways, which we use to explore functional differences between healthy and tumorous tissue. Our analysis, applied across multiple cancers and hundreds of signaling pathways, is accompanied by an interactive web application for exploratory visualization of our findings.

INTRODUCTION

Multi-view -omics datasets offer unprecedented detail for molecular analysis of a wide range of biological phenomena, particularly the genesis and evolution of complex diseases. Especially because of the complexity of data collection, processing, and multi-step analyses, it is critical to apply methods that are scientifically valid and statistically rigorous. Different data types can provide new insight into unique facets of genomic systems, and deepen understanding of complicated biological functions. However, development of flexible statistical methods for these trends in research and next-generation data platforms has not kept pace with the speed of data collection. Moreover, the viability and validity of a statistical method does not equate with accessibility and interpretability. Expressive tools for functional analysis are necessary to advance our understanding of the systems processes of the cell, and to apply novel treatments for complex diseases with distinctive genomic characteristics.

The variety of approaches available for analysis of -omics data is rich, and may address different dimensions of genomic Big Data. With respect to -omics datasets, basic approaches often focus on a single data type. Gene expression and clinical covariates have been applied to

assess cancer survival outcomes across many different cancer types [1], which leverages the large sample sizes that are available. On the other hand, integrative analyses offer unified approach for joint analysis of a large number of genomic features across multiple data platforms [2]. Linear methods have been applied to gene-level measurements of expression, methylation, and copy number within individual cancers [3,4], emphasizing the aggregate information obtained from distinct data platforms.

Methods for analysis of higher-level biological functionality began by testing for differential expression through application of gene set enrichment analysis (GSEA) among collections of genes [5], disregarding functional relationships. Pathway analysis offers a more systematic, structural approach to modeling genomic processes [6]. Application of GSEA to genes *a priori* known to comprise a signaling pathway can give comparative insight into patterns of differential activity across multiple cancers [7]. But, more sophisticated pathway methods have been introduced and applied. The random walk with restart has been applied to assess both network cohesion and compare the explanatory power of several different pathway databases [8]. The SAFE model [9] identifies local neighborhoods of high enrichment situated within larger networks. A factor model approach, PARADIGM, was applied to a single unified network of multiple signaling pathways [10], and has also been used for comparative analysis of individual pathways across cancer types [11]. Graphical methods have also been developed, for directed as well as undirected Gaussian networks [12].

A natural fusion combines data integration and pathway analysis. This has been applied to comprehensive analysis of individual signaling pathway in a single cancer using gene-level covariates including expression, methylation, and copy number, and somatic mutations to analyze androgen receptor signaling [13]. Pathway analysis has also been applied across multiple cancers and multiple pathways [14], and the Lemon-Tree model [15] used a module network model to integrate data and identify novel pathway components.

The large volume of data available for statistical analysis of -omics data present an immense and unique challenge to presenting, exploring, and interpreting analytic results. Phandango [16] provides an interactive web application for visualization of phylogenetic datasets and analyses. General tools for interactive visualization and analysis have introduced software packages tooled for downstream implementation of -omics data in individual analyses [17,18]. The software SeqPlots [19] implemented novel statistical graphics for visualizing cluster analysis applied them to gene expression data.

In this paper, we performed a fully integrative pathway analysis on a pan-cancer dataset. We constructed a multi-platform -omics cancer dataset, including measurements of gene expression, methylation, and DNA copy number from tumor and healthy tissue samples across 22 cancer types. We imputed missing data using the iterative integrated imputation (I3) procedure [20]. We performed pathway analysis using the EMC-NetGSA model [21], a graph-theoretic statistical framework that integrates the multi-modal genetic, transcriptomic, and epigenetic data in a single, unified linear model. We estimated the EMC-NetGSA model for a set of 173 known signaling pathways, and test for differential activation in tumor and healthy tissues. We identified pathways with strong significance in several cancers. We implement a web-based interactive data visualizations for dynamic exploration of the results of our pathway analysis, accessible to the public.

RESULTS

Overview of dataset, pre-processing, and imputation

For our analysis, we used a multi-platform -omics dataset from The Cancer Genome Atlas (TCGA). This long-running study is funded by the National Cancer Institute (NCI), and aggregates data from many research sites [22]. We considered gene expression, methylation, and copy number data collected from tumor and healthy tissues in 32 cancers. The available sample sizes for each cancer are given in Table 1. The samples exhibit a systematic imbalance between the population sample sizes: the number of tumor sample is consistently higher than the number of healthy samples.

For analysis of signaling pathways, we also obtained the NCI Pathway Interaction Database (PID), a set of 212 known signaling pathways expressed as directed functional relationships between genes [23]. We used the pathways to identify the genes of interest among all -omics features available from TCGA. The PID contains 2393 distinct gene symbols, and the TCGA dataset contains some -omics feature for 2369 of these gene symbols. At the level of the individual data types, our dataset contained 2351 genes with expression observations; 2279 genes with methylation; and 2331 genes with copy number. In total, the data matrix for each cancer type contained 6973 -omics features across the three data platforms.

We used TCGA level-3 data for our analysis. We aggregated methylation and copy number observations at the gene level, and applied basic pre-processing steps, both according to the procedures detailed below in Section of “Methods”. This produced an observation matrix of all -omics features for genes in the PID. The data matrix contained a substantial number of missing values.

Table 1 Sample sizes for cancer and control populations in data for 32 cancers from The Cancer Genome Atlas (TCGA)

Cancer	Code	Sample sizes	
		Cancer	Normal
Adrenocortical carcinoma	ACC	92	5
Bladder urothelial carcinoma*	BLCA	412	36
Breast invasive carcinoma*	BRCA	1096	161
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	308	8
Cholangiocarcinoma*	CHOL	36	15
Colon adenocarcinoma*	COAD	455	92
Lymphoid neoplasm diffuse large B-cell lymphoma	DLBC	48	0
Esophageal carcinoma*	ESCA	185	64
Glioblastoma multiform*	GBM	586	31
Head and neck squamous cell carcinoma*	HNSC	528	82
Kidney chromophobe renal cell carcinoma*	KICH	66	57
Kidney renal clear cell carcinoma*	KIRC	534	427
Kidney renal papillary cell carcinoma*	KIRP	291	87
Brain lower grade glioma	LGG	516	0
Liver hepatocellular carcinoma*	LIHC	377	87
Lung adenocarcinoma*	LUAD	519	179
Lung squamous cell carcinoma*	LUSC	504	240
Mesothelioma	MESO	87	1
Ovarian serous cystadenocarcinoma*	OV	586	130
Pancreatic adenocarcinoma*	PAAD	185	37
Pheochromocytoma and paraganglioma	PCPG	179	5
Prostate adenocarcinoma*	PRAD	500	117
Rectum adenocarcinoma*	READ	167	17
Sarcoma*	SARC	261	22
Skin cutaneous melanoma	SKCM	104	3
Stomach adenocarcinoma*	STAD	478	131
Testicular germ cell tumors	TGCT	150	0
Thyroid carcinoma*	THCA	507	99
Thymoma*	THYM	124	12
Uterine corpus endometrial carcinoma*	UCEC	545	51
Uterine carcinosarcoma	UCS	57	6
Uveal melanoma	UVM	80	0

The cancer "code" is an abbreviation for each cancer. All samples are from tissue in the afflicted, cancerous region. For the cancer population, tissue samples are from tumorous tissue. For the normal population, tissue is from healthy normal tissue. Cancers marked with a star (*) were included in the analysis of pathway disturbance, on the basis of each sample population including more than 10 samples.

The specific features with missing values varied according to each sample in the dataset, depending on the research location at which the data was collected, as well as depending on data quality within each individual sample.

To remedy this missing data, subsets of which may conform to a rectangular submatrix of data but other subsets of which do not, we applied the iterative integrated imputation (I3) procedure [20]. I3 extends an existing matrix-completion method based on the assumption of a low-rank data matrix. Whereas the original method, structured matrix completion (SMC) [24], required the assumption of a missing submatrix, I3 performs SMC separately for each sample that exhibits missing data. At each iteration, we considered one sample and formed the maximal set of other samples of the same cancer type for which the observed features in the sample of interest were also observed. We then imputed the minimal covering submatrix for the missing values in the sample of interest, after appropriate row and column reordering. Full details of the imputation method are given in Section of "Methods".

Integrative pathway analysis

We applied the EMC-NetGSA model [21], discussed in greater detail in Section of "Methods". In brief, we constructed for each signaling pathway in the PID a directed graph, with edges representing functional relationships connecting the corresponding vertices for gene expression. We also introduced graph vertices for gene-level methylation and copy number, with directed edges connecting each to the gene's expression vertex. We estimated association weights for the graph edges within each cancer, and applied the NetGSA hypothesis test [25] to test for differential activity. We tested the entire pathway for significance, as well as the subgraphs corresponding to the three vertices available for each gene.

We considered for pathway analysis the 22 cancers listed in Table 1 that are marked with a star (*). Of the 32 total cancers available from TCGA, these had more than 10 samples in both the tumor and normal populations. After removing the pathways that contained genes for which our data lacks observations of gene expression, we tested 173 of the PID pathways across the 22 paired populations. In order to correct for possible false discoveries due to the multiple testing problem, we used the procedure of Benjamini-Hochberg (BH) [26] to adjust all p-values within each cancer.

Figure 1 shows the $-\log_{10}(p)$ transformation for the p-values in all pathways in all cancers. Many p-values are strongly significant: even after BH-adjustment of the 3806 hypothesis tests we performed, only 88 pathways did not have a significant p-value at the $\alpha = 0.05$ level.

However, it is clear from the figure that there exist associations among the p-values across all cancers in specific pathways. This includes consistent disturbance in

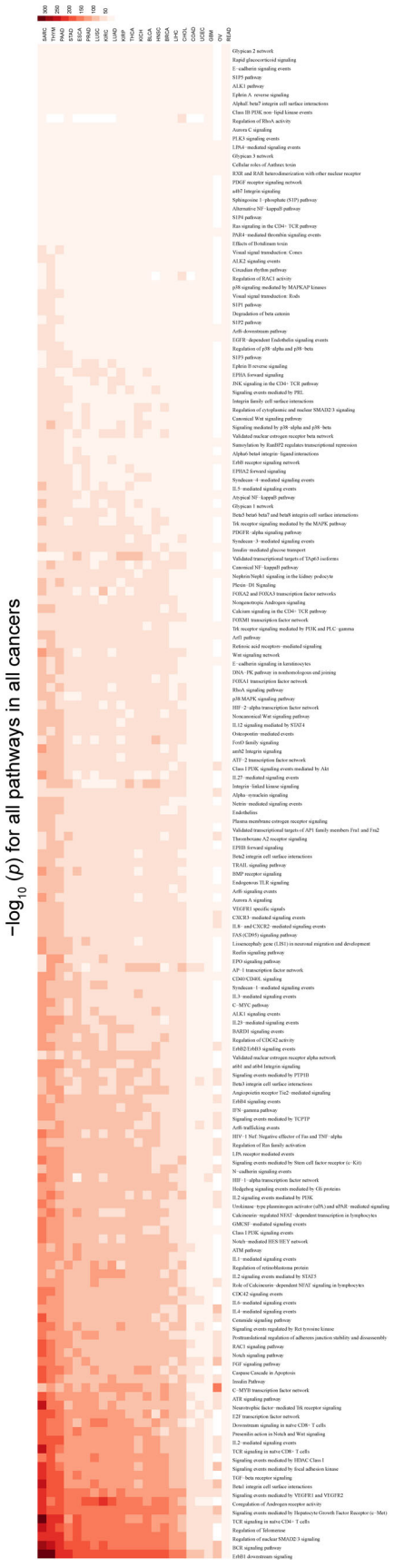

Figure 1. Shown is the $-\log_{10}(p)$ transformation of the BH-adjusted *p*-values obtained for 22 TCGA cancer types among 173 NCI PID pathways. Darker colors indicate stronger significance, in terms of a smaller *p*-value. A white value indicates the pathway was not tested for that cancer. Columns and rows are ranked according to the cumulative significance in each column or row. We observe strong patterns of increased significance across all pathways within individual cancers, such as elevated significance in sarcoma (SARC) and thymoma (THYM) and reduced significance in colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ). We also observe patterns in significance within pathways, across cancers, such as elevated differential activity in B-cell receptor (BCR) signaling, and a lack of disturbance in the glypican 2 and 3 networks.

the pathways B-cell receptor (BCR) signaling, ErbB1 downstream signaling, and transforming growth factor-beta (TGF-beta) receptor signaling, as well as consistent lack of disturbance, as in the glypican 2 and 3 networks, and PDGF receptor signaling. Likewise, we observe systematic elevation of p-values in some cancers, such as sarcoma (SARC) and thymoma (THYM), as well as depressed significance across other cancers, including colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ).

The high significance in some pathways is in accordance with current scientific consensus. BCR signaling regulates B cell activity, and has additional downstream effects on other pathways that promote tumor growth [27]. Burger and Wiestner [28] discussed recent work to target malignancy in B cell signaling. ErbB1 plays a role related to epidermal growth factor (EGFR), and its disturbance is known to relate to a variety of cancers: Roskoski found interactions in non-small cell lung cancer, colon cancer, and breast cancer [29]. They also discussed ErbB1-related targets for clinical treatment of tumors. TGF-beta plays a central role in basic cellular activity [30], and it can suppress or promote tumorigenesis, dependent on the cell type [31]. Fabregat *et al*. [32] discussed treatments that target and inhibit malignant behavior in the TGF-beta pathway, in order to recover normal functioning.

It is also reasonable that we observed co-disturbance of pathways within cancers. Iengar [33] considered genes that consistently exhibit mutation across multiple cancer types, and identified a large number of pathways to which they relate. Leiserson *et al.* [34] directly considered the contrast between driver and passenger mutations in oncogenesis, and focus on finding multiple drivers that coexist within individual cancers.

Among the hypothesis tests shown in Fig. 1, we considered a subset of 14 pairs of cancer and pathway that exhibited stronger significance than would appear typical for the combination of the two. The cancer-pathway pairs we considered are listed in Table 2. We considered 8 pathway disturbances in sarcoma, 3 in thymoma, and 1 pathway disturbance each for kidney renal clear cell carcinoma, ovarian serous cystadeno-carcinoma, and pancreatic adenocarcinoma. In both sarcoma and thymoma, disturbance of T-cell receptor (TCR) signaling in naïve $CD4^+$ T cells was elevated, as was ErbB1 disturbance. The remaining 9 pathways are found in only 1 cancer.

Some of the pathways in Table 2 reflect well-known drivers of the respective cancers. Differential expression of MYB in ovarian cancer has been suspected for a long time [35], and some recent work has given attention to the role of C-MYB activation in other cancers [36]. The latter study found up-regulation in non-small cell lung cancer, a finding we corroborate — albeit to a lesser degree — in lung adenocarcinoma (LUAD), the corresponding TCGA cancer type. The prominent significance of ErbB1 is unsurprising in light of its known role in a wide range of cellular functions.

The significance of Trk signaling mediated by neurotrophic factor is perhaps surprising in sarcoma, a bone cancer. But while research on this type of neural signaling is relatively new, it does support a tumorigenic role for Trk signaling. Increased expression in TrkB and TrkC has been found to correlate with tumor growth in brain tumors [37]. The brain-derived neurotrophic factor

Table 2 The cancer-pathway pairs

Cancer	Pathway
Kidney renal clear cell carcinoma	Coregulation of androgen receptor activity
Ovarian serous cystadenocarcinoma	C-MYB transcription factor network
Pancreatic adenocarcinoma	Regulation of nuclear SMAD2/3 signaling
Sarcoma	BCR signaling pathway
Sarcoma	Beta1 integrin cell surface interactions
Sarcoma	Ceramide signaling pathway
Sarcoma	ErbB1 downstream signaling
Sarcoma	Neurotrophic factor-mediated Trk receptor signaling
Sarcoma	Signaling events mediated by focal adhesion kinase
Sarcoma	TCR signaling in naïve $CD4^+$ T cells
Sarcoma	TCR signaling in naïve $CD8^+$ T cells
Thymoma	BCR signaling pathway
Thymoma	ErbB1 downstream signaling
Thymoma	TCR signaling in naïve $CD4^+$ T cells

Pairs of cancer and pathway for which the residual from a regression of the logit (p)-value was negative, and larger in magnitude than a Bonferroni-adjusted critical value. In sarcoma, we observed 8 disturbed pathways; in thymoma, we observed 2 disturbed pathways; and we observed 1 pathway disturbance in each of kidney renal clear cell carcinoma, ovarian serous cystadenocarcinoma, and pancreatic adenocarcinoma.

(BDNF) was found to bind to the TrkB receptor, and elevated expression had strong downstream effects in other pathways [38]. A Phase I clinical trial was also performed for an inhibitor of Trk in several cancer types including sarcoma [39]. Another study found that inhibition of Trk signaling corresponded to improved clinical outcomes in Ewing sarcoma [40].

Increased adhesion of shed tumor cells in the presence of focal adhesion kinase (FAK) activation has been reported, and it is hypothesized that inhibition of FAK may yield superior patient outcomes in a variety of cancers, including sarcoma [41]. Others observed tumor suppression as a result of FAK inhibition [42], and FAK inhibition-mediated signaling was found to complement and reinforce synergistically the therapeutic effects of Aurora kinase B inhibition in Ewing sarcoma [43].

The role of androgen receptors has been considered in murine tissues, including kidney tissue, which exhibit a different biological effect from prostate and epididymis tissue [44]. That effect suggests a novel, kidney-specific role for the androgen receptor pathway. Increased androgen receptor expression in kidney renal clear cell carcinoma was found to correspond to improved clinical outcomes [45], which supports earlier findings that also found a possible tumor-suppressive role for androgen receptors via circadian regulation in the kidney [46].

Data visualization

The variety of these results hints at the complex and exhaustive detail that is produced as a result of the integrative procedure. The results just detailed were identified on the basis of the p-value of the hypothesis test alone, but the EMC-NetGSA analysis offers many additional outputs beyond a binary decision. First, it is important to understand the network topology of the pathways under consideration, and consideration of the association weights in the weighted graphs. In addition, EMC-NetGSA estimates network-adjusted expression parameters that offer insight into the contribution of individual genes to pathway disturbance, as well as which pathway components have differential activity across populations. Subsets of genes within a pathway, or individual genes themselves, may be of interest for testing differential pathway activity, especially while controlling for the network effects of other genes that are not of interest but may contribute to genomic activity through the signaling pathway network.

To facilitate in-depth exploration of these pathway analysis results, we implemented an interactive web application and data visualization. The software can be accessed online at the website (zhang-lab.shinyapps.io/pathway-analysis-tcga-cancers/). We provide an exploratory tool to interact with the network topology of each signaling pathway from our analysis, for all 22 cancers we considered. We also provide visualizations of the hypothesis test outcomes, for the entire pathway, as well as the subpaths formed by the 3 -omics vertices — expression, methylation, and copy number — for each gene in the dataset. Moreover, we performed a comparative analysis of the results of different -omics integration strategies: in addition to the EMC-NetGSA model, we also considered integrated pathway of expression and methylation (EM-NetGSA); expression and copy number (EC-NetGSA); and expression-only pathway analysis (NetGSA). Figure 2 shows an example pair of plots for the signaling events mediated by focal adhesion kinase pathway in the sarcoma cancer. We display the $-\log_{10}(p)$ -values for significance tests of the full pathway, as well as the integrated subgraphs of -omics features at the level of individual genes, and we also plot the test statistic to show the sign and magnitude of the overall pathway disturbance effect. The plots were generated by the interactive data visualization website, and provide comparative analysis of significance and the direction and magnitude of differential activity across data integration schemes and attributable to different data types.

DISCUSSION

By applying multi-platform pathway analysis to a network model of signaling pathways, we identified significant pathway disturbances in several cancers. We applied an iterative imputation procedure to pre-process missing data in configurations that would otherwise be inaccessible for analysis. We performed a top-to-bottom integrative pathway analysis of the TCGA cancer dataset using observations collected on 3 different data platforms.

We performed the integrative pathway analysis on 22 cancers from TCGA and 173 pathways published by the NCI PID. We tested for differential pathway activation for each cancer, and identified 14 cancer-pathway pairs that exhibited robust, high significance across 11 different pathways in 5 cancers. We discussed current research to support the disturbance of these pathways. The concordance of results of our pathway analysis and ongoing biological research studies indicates that our pathway analysis across the set of TCGA cancers identifies real biological phenomena. Two particular findings that warrant further investigation relate to Trk signaling and FAK inhibition in sarcoma.

We also introduced an interactive web application for data visualization of the results of the pathway analysis. Available to the public, this web-based software provides a simple interface to explore the large number of model outputs, including estimated signaling pathway topology and weights, significance test results, and parameter estimates.

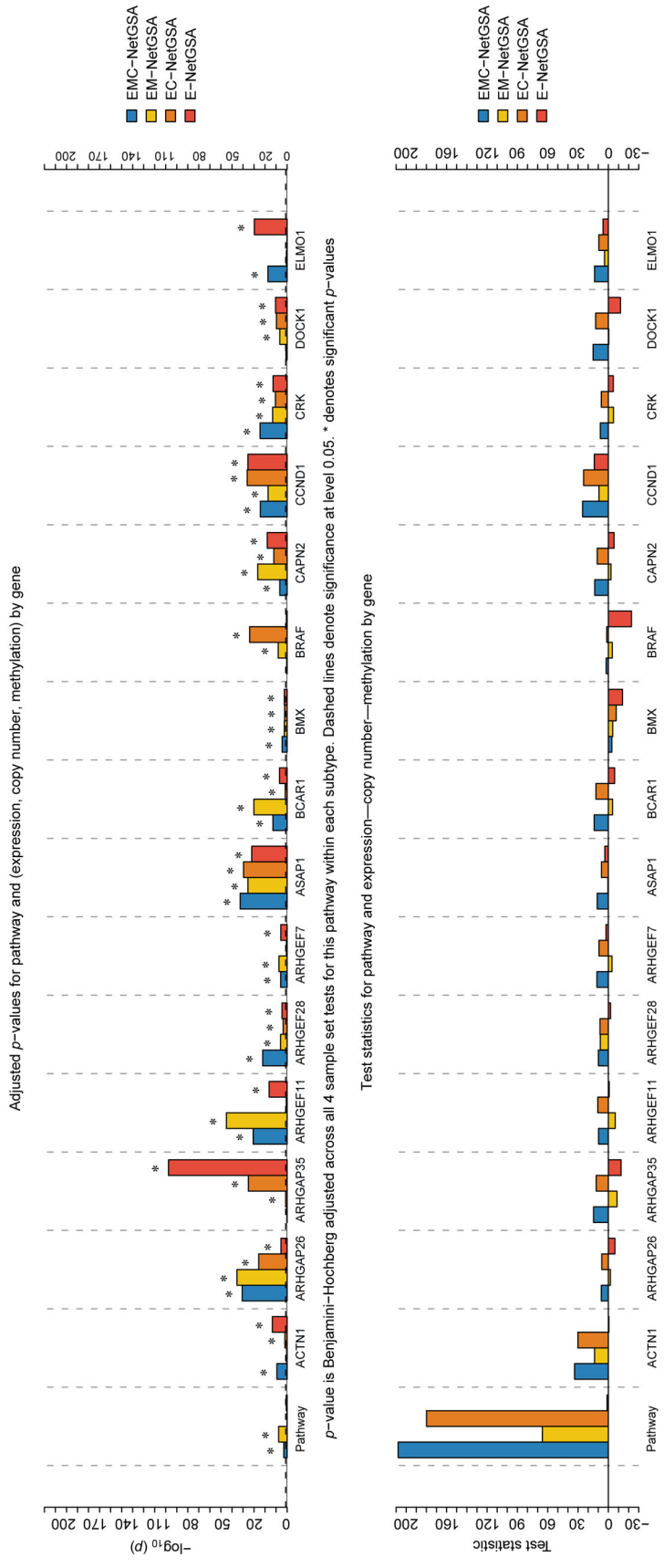

Figure 2. **$-\log_{10}(p)$-values (top) and signed test statistics (bottom) for the signaling events mediated by focal adhesion kinase pathway in sarcoma cancer tumors.** These plots were generated procedurally using an interactive data visualization application, publicly accessible online. Shown are the results of hypothesis tests for the full pathway, as well as the -omics features for each individual gene, tested as separate subpathways. The pathway analysis was performed on a fully-integrated dataset consisting of expression, methylation, and copy number, using the EMC-NetGSA model; expression and methylation (EM-NetGSA); expression and copy number (EC-NetGSA); and expression alone (NetGSA).

The TCGA project offers additional data types such as point mutations, non-coding RNAs, and protein microarray data. These additional data types present opportunities for further work, based on their structural relationships with the expression, methylation, and copy number data. These data and the functional units they represent may be of intrinsic value, as entities of scientific focus in their own right. They may also be used to refine estimates that relate to the gene-level features, to increase statistical power and precision.

The accessibility and interpretability of our pathway analysis improves substantially through a visualization interface. It is possible to quickly examine a large number of plots, within a cancer and by pathway. It also offers access to the data necessary to drill down into the effect of data integration within a pathway, and understanding the role of individual -omics features. As an example, the pathway analysis in Fig. 2 indicates that integration of copy number into the pathway analysis does not increase significance, relative to an expression-only analysis, despite a large increase in the magnitude of the test statistic. On the other hand, integration of methylation using EM-NetGSA yields a smaller test statistic, in magnitude, but a much higher degree of statistical significance. And, full integration using the final EMC-NetGSA model sustains the significance of the pathway disturbance. This provides a fuller picture of the biological components that comprise the larger biological process. Moreover, our data visualization was easy to construct, using the R language, and easily deployed for open access, demonstrating the potential for widespread use of similar tools to analyze and assess a large volume of analytic outputs.

METHODS

Dataset and pre-processing

Prior to statistical analysis, we assembled the -omics dataset from TCGA. We downloaded multi-platform data for 32 cancers. The data consisted of measurements of gene expression, methylation, and copy number variation (CNV). We downloaded the TCGA data from the NCI Genomic Data Commons (GDC), an online interface to access TCGA data [47]. We used the TCGA-Assembler software, version 2.0.0 [48,49], with which we downloaded the level-3 TCGA data.

We also downloaded the NCI Pathway Interaction Database (PID), a set of known signaling pathways. The pathway information specifies known, directed functional relationships between genes. We used the pathways to identify the -omics features to include in our dataset for imputation and, later, for analysis of pathway disturbance and differential activity. We downloaded network information for 212 signaling pathways from the NCI PID, using the graphite R package [50].

Prior to imputation, we performed pre-processing, transformations, and aggregation of the -omics features to the gene level. Gene expression data was measured on the RNASeqV2 platform. We used normalized data supplied by TCGA, which converts the raw read counts to fragments per kilobase of transcript per million mapped reads upper quartile (FPKM-UQ). We applied a $\log_2$ transformation to the normalized read counts. Methylation beta values were collected using the HumanMethylation450 BeadChip platform, by CpG site. We grouped the sites according to the corresponding gene, and took the average beta value across all sites for that gene. We used copy number variation with germline copy number variants removed, and took the average across all DNA regions for a given gene.

Iterative integrated imputation

Within each cancer, we constructed a $q \times N$ data matrix $\mathbf{X}$ of gene expression, methylation, and copy number observations. q gives the number of genomic features corresponding to any gene in the PID pathways, and N is the number of samples available for the cancer. We further distinguished samples as either tumorous or healthy tissue. Denote the number of tumor samples by N_1 and the number of healthy samples by N_2, which are such that $N = N_1 + N_2$. We found the sample sizes were not balanced, *i.e.*, $N_1 \geqslant N_2$, so we applied integration to the full data matrix.

For each sample, we observed a vector $\mathbf{x}_i$ of q genomic observations on one of the three platforms, $i = 1,..., N$. This gives the column-wise block structure $\mathbf{X} = (\mathbf{x}_1,..., \mathbf{x}_N)$. The data matrix obtained from TCGA contains a large number of missing values, which can be categorized as one of three types. Some features were present across all samples in the dataset. Other features were missing in a block-wise fashion, typically in subjects collected at the same experimental site, where a certain data platform may not have been collected. Finally, individual samples were missing values at-random, due to quality problems in the data. The patterns of these last features cannot be rearranged in such a way that the missing values form a rectangular submatrix.

To address the second type of missing data, Cai *et al*. [24] introduced the Structured Matrix Completion (SMC) method, which they applied to a TCGA dataset to impute block-missing data. Their method used a low-rank assumption to perform imputation by way of an approximation to the singular value decomposition (SVD). The applicability of this method is limited, however, due to the values that are missing at-random. To address the non-block missing data, Linder and Zhang

[20] proposed the iterative integrated imputation (I3) method. They adapted the SMC method for at-random missing data in individual columns of $\mathbf{X}$. The method produces imputed values that are exactly those of the original SMC, when applied to the special case of rectangular missing data. They demonstrated superior performance relative to the naive application of SMC to the minimal covering submatrix for the missing data in $\mathbf{X}$, and we used this method for imputation in our pathway analysis.

We note that I3 and SMC rely upon the SVD, which exhibits the property of invariance under row and column permutations. Therefore, we may rearrange the rows and columns of $\mathbf{X}$ without loss of generality. We partitioned the columns of $\mathbf{X}$ into two groups: (1) samples with no missing values; (2) samples with any missing values. Thus, we may consider $\mathbf{X}$ as a block matrix, $\mathbf{X} = (\mathbf{X}_1 \;\; \mathbf{X}_2)$ where $\mathbf{X}_1$ contains no missing values, and $\mathbf{X}_2$ contains at least one missing value in each column. It is possible the matrix $\mathbf{X}_1$ has zero columns, that is, that every sample in the data matrix contains at least one missing value. Another attractive feature of the I3 method is its ability to impute data even in this edge scenario.

Denote the dimension of $\mathbf{X}_1$ by $q \times n_1$, and the dimension of $\mathbf{X}_2$ by $q \times n_2$, where $N = n_1 + n_2$. Using I3, we iterated over the n_2 columns of $\mathbf{X}_2$, performing a separate imputation for each column.

Consider the column $\mathbf{x}_i$ of $\mathbf{X}_2$. We may form a new matrix by $\chi_i' = (\mathbf{X}_1 \;\; \mathbf{x}_i)$, which has dimension $q \times (n_1 + 1)$, and by the invariance of SVD, we may rearrange the columns and rows such that the q_i missing values in $\mathbf{x}_i$ are located in the bottom-right corner χ_i', so that χ_i' is a matrix with n_1 columns with no missing values, and a $q_i \times 1$ rectangular submatrix of missing values.

Denote the number of features that are not missing in x_i by $p_i = q - q_i$. Among the $n_2 - 1$ columns that remained in $\mathbf{X}_2$ after removing $\mathbf{x}_i$, we located the k_i columns that were not missing any values in the p_i features that are observed in $\mathbf{x}_i$, where $0 \leqslant k_i \leqslant n_2 - 1$. Denote the index set for these k_i samples by $\omega_i = \{\omega_{i1},...,\omega_{ik_i}\}$. That is, the index ω_{ij} corresponds to a column index in $\mathbf{X}_2$, where $\omega_{ij} \neq i$ and $j = 1,..., k_i$. Finally, we formed the matrix

$$\chi_i = (\mathbf{X}_1 \;\; \mathbf{x}_i \;\; \mathbf{x}_{\omega_{i1}} \cdots \mathbf{x}_{\omega_{ik_i}}) \tag{1}$$

which had dimension $q \times (n_1 + k_i + 1)$. Without loss of generality, we suppose the only missing values in χ_i are located in the bottom-right $q_i \times (k_i + 1)$ submatrix. It may be that some of the observations in this submatrix are observed, but by construction, the first column of the submatrix is entirely missing, and corresponds to the missing values in $\mathbf{x}_i$.

We imputed this entire submatrix using SMC, which produced an imputed submatrix. In particular, we used the imputed values found in the first column of this submatrix to form the imputed vector $\hat{\mathbf{x}}_i$ which contained no missing values. We applied this procedure iteratively to each of the n_2 columns of $\mathbf{X}_2$, so that at the completion of the I3 procedure, we formed the imputed matrices $\hat{\mathbf{X}}_2$ and $\hat{\mathbf{X}} = (\mathbf{X}_1 \;\; \hat{\mathbf{X}}_2)$.

Intuitively, at each step of the iteration, we imputed the minimal covering submatrix of the missing values in a column of $\mathbf{x}_i$ that had the maximal number of columns drawn from $\mathbf{X}_2$ for which all observed values in $\mathbf{x}_i$ were also observed. Although any individual imputation of χ_i may have ignored missing values in other columns $\mathbf{x}_j$, $j \neq i$, those sample vectors were imputed at a separate iteration, using all available information for that sample. In the event that the missing data conforms to a rectangular shape, we would perform n_2 iterative steps using all columns of $\mathbf{X}_2$ in each iteration, which produces exactly the result of application of SMC to that submatrix.

EMC-NetGSA pathway model

Having imputed the missing values in $\mathbf{X}$, we next applied pathway analysis to the data matrix. Our analysis proceeds from a graphical representation of a signaling pathway. We expanded the set of graph vertices to include not only elements that correspond to genes, but also vertices representing gene methylation and DNA copy number variation (CNV). We aggregated methylation and CNV data at the level of the gene. We integrated these secondary data platforms with gene expression with directed edges leading from the methylation and copy number vertices, into the gene expression vertex.

Suppose a known signaling pathway, specified as directed functional relationships between genes. Denote the signaling pathway of interest by a graph $\mathcal{G}_{\mathrm{E}}$, defined through a set of p vertices $\mathcal{V}_{\mathrm{E}}$ and a set of n_{E} directed edges $\mathcal{E}_{\mathrm{E}}$. The vertices in $\mathcal{V}_{\mathrm{E}}$ correspond to the genes in the signaling pathway, and the edges in $\mathcal{E}_{\mathrm{E}}$ correspond to the network topology of the signaling pathway. The network topology is given by the known functional interactions between genes vertices in $\mathcal{E}_{\mathrm{E}}$.

We considered the graph adjacency matrix of $\mathcal{G}_{\mathrm{E}}$, a conventional representation of a graph specified in a compact, discrete form. Denote the unweighted graph adjacency matrix of $\mathcal{G}_{\mathrm{E}}$ by $\mathbf{A}_{\mathrm{E}}^{*}$, a $p \times p$ matrix where the element α_{jk} is equal to 1 if vertex j is conditionally dependent upon vertex k. In terms of the signaling pathway, we may consider α_{jk} to be an indicator function, where

$$\alpha_{jk} = 1(\exists \text{ a directed pathway edge from gene } k \text{ to gene } j) \tag{2}$$

The edges in $\mathcal{E}_{\mathrm{E}}$ specify known biological interactions between genes, which operate in a causal manner.

Symmetric to this interpretation of the gene expression vertices, Zhang *et al*. [21] introduced the EMC-NetGSA model for integration of methylation and copy number with expression values. We applied that integration method to obtain a pathway graph that includes -omics observations as well as the basic gene expression values. First, we specified a new set of $2p$ vertices $\mathcal{V}_{\mathrm{MC}}$, containing one methylation vertex and one CNV vertex for each gene in $\mathcal{G}_{\mathrm{E}}$. Denoting the collection of all gene expression, methylation, and copy number vertices by $\mathcal{V} \equiv \mathcal{V}_{\mathrm{E}} \cup \mathcal{V}_{\mathrm{MC}}$, we also introduced a set of graph edges $\mathcal{E}_{\mathrm{MC}}$ between the vertices in $\mathcal{V}$. $\mathcal{E}_{\mathrm{MC}}$ contained $2p$ edges, one leading from each methylation and copy number vertex to the corresponding gene expression vertex in $\mathcal{V}_{\mathrm{E}}$.

Taking $\mathcal{E} \equiv \mathcal{E}_{\mathrm{E}} \cup \mathcal{E}_{\mathrm{MC}}$, we obtained a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consisting of $3p$ vertices — three for each gene — and $n_{\mathrm{E}} + 2p$ edges. We may write the full unweighted graph adjacency matrix by

$$\mathbf{A}^* = \begin{pmatrix} \mathbf{A}^*_{\mathrm{E}} & \mathbf{I}_p & \mathbf{I}_p \\ \mathbf{O}_{2p\times p} & \mathbf{O}_{2p\times p} & \mathbf{O}_{2p\times p} \end{pmatrix} \tag{3}$$

where $\mathbf{O}g\times g$ is a $g \times g$ matrix of zeros. Thus defined, $\mathbf{A}^*$ is a $3p\times 3p$ matrix composed of the graph adjacency matrix for the signaling pathway, $\mathbf{A}^*_{\mathrm{E}}$; identity matrices along its right hand margin, and zeros elsewhere. The identity matrices express the directed edges leading from the methylation and copy number vertices to the gene expression vertices. The zeros reflect that we did not include any graph relationships between the secondary -omics features, nor did we include any feedback mechanism by which gene expression influences methylation or copy number activation.

We considered a population of N samples, and we denote the observation vector of $3p$ -omics features by $\mathbf{y}_i$, $i = 1,..., N$. $\mathbf{y}_i$ consisted of three subvectors of p elements each, *i.e.*, $\mathbf{y}_i = (\mathbf{y}'_{i1}, \mathbf{y}'_{i2}, \mathbf{y}'_{i3})'$. Here, $\mathbf{y}_{i1}$ contains gene expression elements; $\mathbf{y}_{i2}$ contains methylation elements; and $\mathbf{y}_{i3}$ contains copy number elements.

Graphical statistical models are often characterized in terms of the inverse of the covariance matrix for a multivariate normal distribution. In that Gaussian setting, the elements of the resultant matrix specify conditional dependence between graph vertices, and we say that vertex j is conditionally dependent on vertex k when the element α_{jk} of $\mathbf{A}^*$ is nonzero. This coincides with the presence of a directed edge from vertex k to vertex j.

Conditional dependence is formalized using the partial correlation coefficient. Denoting two random variables by X and Y, we also consider a set $\mathcal{Z}$ of additional covariates. Denote the linear projection onto the elements of $\mathcal{Z}$ by $P_{\mathcal{Z}}$. The orthogonal complement of X with respect to $\mathcal{Z}$ is then $X_{X\setminus\mathcal{Z}} = X - P_{\mathcal{Z}}X$. In the linear models setting, $X_{X\setminus\mathcal{Z}}$ is simply the residual of a regression of X on $\mathcal{Z}$. Applying an identical procedure to Y, we obtain these orthogonal complements, and the partial correlation coefficient is then given by $\rho_{XY} = \mathrm{corr}\,(X_{X\setminus\mathcal{Z}}, Y_{Y\setminus\mathcal{Z}})$ [51].

Returning to the signaling pathway, for a vertex j conditionally dependent on the vertex k, we performed two linear regressions, one of each vertex on the remaining $(3p-2)$ vertices. Then, we calculated the Pearson correlation coefficient r_{jk} between the residuals of these two regressions [52]. Finally, we formed a weighted adjacency $\mathbf{A}$ matrix from the elements of $\mathbf{A}^*$ and $\{ r_{jk} \mid \mathbb{1}(\alpha_{jk} = 1)$, where the element a_{jk} of $\mathbf{A}$ corresponding to α_{jk} in $\mathbf{A}^*$ is given by $a_{\mathrm{jk}} = r_{jk}\alpha_{jk}$. Although in principle we may calculate the values of r_{jk} for all j, $k = 1,..., 3p$, in practice it is only necessary to do so for the $n_{\mathrm{E}} + 2p$ edges given by $\mathcal{E}$.

Shojaie and Michailidis [25] introduced a transformation of $\mathbf{A}$, denoted by $\mathbf{\Lambda}$ and called the influence matrix, that expresses the cumulative network effect of each graph vertex on the other vertices in the matrix. They originally demonstrated that for directed acyclic graphs (DAGs), the transformation has the analytic form $\mathbf{\Lambda} = (\mathbf{I} - \mathbf{A})^{-1}$, where $\mathbf{I}$ is an identity matrix with the same dimension as $\mathbf{A}$. Shojaie and Michailidis [53] extended the transformation in two ways. First, they demonstrated that it holds for any substochastic graph, which coincides with graphs with an adjacency matrix that has eigenvalues all smaller than 1 in magnitude. Second, they derived a limit result to induce substochasticity in an arbitrary directed graph. This permits approximation of the influence matrix for graphs that are not substochastic.

Accompanying the influence matrix, Shojaie and Michailidis [25] introduced the NetGSA framework for significance testing of differential activation in signaling pathways. Zhang *et al*. [21] applied that same framework to the integrated graph, yielding the EMC-NetGSA model. The essential components of the NetGSA framework are a mixed-effects linear regression model for the response vectors $\mathbf{y}_i$. Adjusted appropriately for the EMC-NetGSA framework, the model has the following form:

$$\mathbf{y}_i = \mathbf{\Lambda}\beta + \mathbf{\Lambda}_{\gamma i} + \epsilon_i, \;\; i = 1,...,N \tag{4}$$

$$\epsilon_i \sim N_{3p}(\mathbf{0}_{3p}, \sigma^2_\epsilon \mathbf{I}_{3p}) \tag{5}$$

$$\gamma_i \sim N_{3p}(\mathbf{0}_{3p}, \sigma^2_\gamma \mathbf{I}_{3p}) \tag{6}$$

In addition to the observation vectors $\mathbf{y}_i$ and the influence matrix $\mathbf{\Lambda}$, derived from the weighted adjacency matrix $\mathbf{A}$, the model consists of β, a vector of network-adjusted mean parameters; γ_i, a sample-level random effect; and ϵ_i, a sample-level random error. The term γ_i serves to model the correlation in the observations within each sample.

The NetGSA framework provides a statistical hypothesis test for differential activation of -omics features in a

signaling pathway. To begin, we classified samples according to two populations, treatment and control. In the context of a complex disease like the 32 TCGA cancers we considered in our analysis, the treatment population represents samples of tumor tissue, whereas the control population represents the normal tissue samples collected from the corresponding anatomical site.

For a given sample i, denote the sample population by a class label $c_i \in \{\mathrm{T}, \mathrm{C}\}$, $i = 1,..., N$. Denote the number of treatment (tumor) samples by N_1 and the number of control (normal) samples by N_2, where $N = N_1 + N_2$. We then parameterize population-specific weighted adjacency matrices $\mathbf{A}_\mathrm{T}$, $\mathbf{A}_\mathrm{C}$, influence matrices $\mathbf{\Lambda}_\mathrm{T}$, $\mathbf{\Lambda}_\mathrm{C}$, and network-adjusted mean parameters β_T, β_T. Thus, the NetGSA model given above becomes

$$\mathbf{y}_i = \mathbf{\Lambda}_{c_i}\ \beta_{c_i} + \Lambda_{c_i}\ \gamma_i + \epsilon_i \tag{7}$$

For analysis of differential activation in a subset of genes of interest, controlling for the network effect of all features in the -omics pathway, we specified an indicator vector $\mathbf{b}$ with $3p$ elements, each equal to zero or one. The elements correspond to the genomic features in the observation vectors $\mathbf{y}_i$. We tested the differential activity of the elements corresponding to values of one in $\mathbf{b}$. The NetGSA network contrast is given by $\ell = (-\mathbf{b}\cdot\mathbf{b}\mathbf{\Lambda}_\mathrm{C}, \mathbf{b}\cdot\mathbf{b}\mathbf{\Lambda}_\mathrm{T})$. As discussed by Shojaie and Michailidis [25], ℓ facilitates significance testing for pathway disturbance while controlling for the cumulative network effects of the full pathway. The test statistic is given by $T \propto \ell\,\beta$, where $\beta = (\beta'_\mathrm{C}, \beta'_\mathrm{T})'$, which approximately follows a Student's t distribution. We estimated the degrees of freedom for T using the Satterthwaite approximation.

Even after imputation of the missing values in the original data matrix $\mathbf{X}$, it may be that individual -omics features that correspond to genes in the PID pathways remain missing — namely, missing across all samples. In these instances, we could not impute the missing data, and the impact on the pathway analysis depended on the data type of the missing feature. In the case of missing gene expression features, we were unable to perform any analysis of signaling pathways in which it is found, and we ignored these pathways. On the other hand, when the absent feature measured methylation or copy number, we simply removed the corresponding vertex and integrative edge from $\mathcal{V}$ and $\mathcal{E}$, respectively, following the argument found in Zhang *et al*. [21]. Therefore, final analysis of a pathway could contain fewer than $3p$ genomic features, but always contained p gene expression features.

Software for interactive data visualization

We pre-rendered the results of the pathway analysis by analyzing all 173 pathways in all 22 cancers, and saving the output. We built the data visualization using several R packages: for interactivity we used shiny, which enables implementation of rich-content, dynamic websites rendered with HTML, CSS, and Javascript, but coded using the R language [54]. For graph visualization, we used the package igraph [55], a cross-platform software package for analysis and display of graphs, and visnetwork [56], an interactive extension of the igraph functionality with integration with the shiny software.

ACKNOWLEDGEMENTS

Yuping Zhang acknowledges Faculty Research Excellence Program Award from University of Connecticut.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Henry Linder and Yuping Zhang declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Chandrashekar, D.S., Bashel, B., Akshaya, S., Balasubramanya, H., Creighton, C.J., Ponce-Rodriguez, I., Chakravarthi, B. and Varambally, S. (2017) UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia, 19, 649–658
2. Zhang, Y., Ouyang, Z. and Zhao, H. (2017) A statistical framework for data integration through graphical models with application to cancer genomics. Ann. Appl. Stat., 11, 161–184
3. Cancer Genome Atlas Research Network (2017) Integrated genomic and molecular characterization of cervical cancer. Nature, 543, 378–384
4. Shen, R., Olshen, A. B. and Ladanyi, M. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics, 25, 2906–2912
5. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci., 102,15545–15550
6. Yan, J., Risacher, S. L., Shen, L. and Saykin, A. J. (2017) Network approaches to systems biology analysis of complex disease: integrative methods for multi -omics data. Brief. Bioinform., 19, 1370–1381
7. Ge, Z., Leighton, J. S., Wang, Y., Peng, X., Chen, Z., Chen, H., Sun, Y., Yao, F., Li, J., Zhang, H., *et al.* (2018) Integrated genomic analysis of the ubiquitin pathway across cancer types. Cell Reports, 23, 213–226.e3
8. Huang, J. K., Carlin, D.E., Yu, M. K., Zhang, W., Kreisberg, J. F., Tamayo, P. and Ideker, T. (2018) Systematic evaluation of molecular networks for discovery of disease genes. Cell Syst., 6, 484–495

9. Baryshnikova, A. (2016) Systematic functional annotation and visualization of biological networks. Cell Syst., 2, 412–421
10. Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D. and Stuart, J. M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. Bioinformatics, 26, i237–i245
11. Campbell, J. D., Yau, C., Bowlby, R., Liu, Y., Brennan, K., Fan, H., Taylor, A. M., Wang, C., Walter, V., Akbani, R., *et al.*, (2018) Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. Cell Reports, 23, 194–212.e6
12. Ma, J., Shojaie, A. and Michailidis, G. (2016) Network-based pathway enrichment analysis with incomplete network information. Bioinformatics, 32, 3165–3174
13. Robinson, D., Van Allen, E. M., Wu, Y. M., Schultz, N., Lonigro, R. J., Mosquera, J. M., Montgomery, B., Taplin, M. E., Pritchard, C. C., Attard, G., *et al.* (2015) Integrative clinical genomics of advanced prostate cancer. Cell, 161, 1215–1228
14. Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., Dimitriadoy, S., Liu, D. L., Kantheti, H. S., Saghafinia, S., *et al.* (2018) Oncogenic signaling pathways in the cancer genome atlas. Cell, 173, 321–337.e10
15. Bonnet, E., Calzone, L. and Michoel, T. (2015) Integrative multi - omics module network inference with Lemon-Tree. PLOS Comput. Biol., 11, e1003983
16. Hadfield, J., Croucher, N. J., Goater, R. J., Abudahab, K., Aanensen, D. M. and Harris, S. R. (2017) Phandango: an interactive viewer for bacterial population genomics. Bioinformatics, 34, 292–293
17. Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag
18. Yin, T., Cook, D. and Lawrence, M. (2012) ggbio: an R package for extending the grammar of graphics for genomic data. Genome Biol., 13, R77
19. Stempor, P. and Ahringer, J. (2016) SeqPlots–Interactive software for exploratory data analyses, pattern discovery and visualization in genomics. Wellcome Open Res., 1, 14
20. Linder, H. and Zhang, Y. (2019) Iterative integrated imputation for missing data and pathway models with applications to breast cancer subtypes. Comm. Statis. Appl. Meth., 26, 411–430
21. Zhang, Y., Linder, H. M.Shojaie, Ouyang, A., Shen, Z., Baggerly, R., Baladandayuthapani, K.A. and Zhao, V. H. (2017) Dissecting pathway disturbances using network topology and multi-platform genomics data. Stat. Biosci., 10, 1–21
22. Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. (Pozn.), 19, A68–A77
23. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T. and Buetow, K. H. (2008) Pid: the pathway interaction database. Nucleic acids research, 37 (suppl), D674–D679
24. Cai, T., Cai, T. T. and Zhang, A. (2016) Structured matrix completion with applications to genomic data integration. J. Am. Stat. Assoc., 111, 621–633
25. Shojaie, A. and Michailidis, G. (2009) Analysis of gene sets based on the underlying regulatory network. J. Comput. Biol., 16, 407–426
26. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B, 57, 289–300
27. Fowler, N. and Davis, E. (2013) Targeting B-cell receptor signaling: changing the paradigm. Hematology, 553–560
28. Burger, J. A. and Wiestner, A. (2018) Targeting B cell receptor signalling in cancer: preclinical and clinical advances. Nat. Rev. Cancer, 18, 148–167
29. Roskoski, R. Jr. (2014) The ErbB/HER family of protein-tyrosine kinases and cancer. Pharmacol. Res., 79, 34–74
30. Jakowlew, S. B. (2006) Transforming growth factor-β in cancer and metastasis. Cancer Metastasis Rev., 25, 435–457
31. Massagué, J. (2008) TGFbeta in Cancer. Cell, 134, 215–230
32. Fabregat, I., Fernando, J., Mainez, J. and Sancho, P. (2014) TGF-beta signaling in cancer treatment. Curr. Pharm. Des., 20, 2934–2947
33. Iengar, P. (2018) Identifying pathways affected by cancer mutations. Genomics, 110, 318–328
34. Leiserson, M. D. M., Blokh, D., Sharan, R. and Raphael. B. J., (2013) Simultaneous identification of multiple driver pathways in cancer. PLOS Comput. Biol., 9, e1003054
35. Barletta, C., Lazzaro, D., Prosperi Porta, R., Testa, U., Grignani, F., Ragusa, R. M., Leone, R., Patella, A., Carenza, L. and Peschle, C. (1992) C-MYB activation and the pathogenesis of ovarian cancer. Eur. J. Gynaecol. Oncol., 13, 53–59
36. Jin, Y., Zhu, H., Cai, W., Fan, X., Wang, Y., Niu, Y., Song, F. and Bu, Y. (2017) B-myb is up-regulated and promotes cell growth and motility in non-small cell lung cancer. Int. J. Mol. Sci., 18, 860
37. Lawn, S., Krishna, N., Pisklakova, A., Qu, X., Fenstermacher, D. A., Fournier, M., Vrionis, F. D., Tran, N., Chan, J. A., Kenchappa, R. S., *et al.* (2015) Neurotrophin signaling via TrkB and TrkC receptors promotes the growth of brain tumor-initiating cells. J. Biol. Chem., 290, 3814–3824
38. Meng, L., Liu, B., Ji, R., Jiang, X., Yan, X. and Xin, Y. (2019) Targeting the BDNF/TrkB pathway for the treatment of tumors. Oncol Lett, 17, 2031–2039
39. Drilon, A., Siena, S., Ou, S. I., Patel, M., Ahn, M. J., Lee, J., Bauer, T. M., Farago, A. F., Wheler, J. J., Liu, S. V., *et al.* (2017) Safety and antitumor activity of the multitargeted pan-trk, ros1, and alk inhibitor entrectinib: combined results from two phase i trials (alka-372-001 and startrk-1). Cancer Discov., 7, 400–409
40. Heinen, T. E., Dos Santos, R. P., da Rocha, A., Dos Santos, M. P., Lopez, P. L., Silva Filho, M. A., Souza, B. K., Rivero, L. F., Becker, R. G., Gregianin, L. J., *et al.* (2016) Trk inhibition reduces cell proliferation and potentiates the effects of chemotherapeutic agents in Ewing sarcoma. Oncotarget, 7, 34860–34880
41. Perry, B. C., Wang, S. and Basson, M. D. (2010) Extracellular pressure stimulates adhesion of sarcoma cells via activation of focal adhesion kinase and akt. Am. J. Surg., 200, 610–614
42. Crompton, B. D., Carlton, A. L., Thorner, A. R., Christie, A. L., Du, J., Calicchio, M. L., Rivera, M. N., Fleming, M. D., Kohl, N. E., Kung, A. L., *et al.* (2013) High-throughput tyrosine kinase activity profiling identifies FAK as a candidate therapeutic target in

Ewing sarcoma. Cancer Res., 73, 2873–2883
43. Wang, S., Hwang, E. E., Guha, R., O'Neill, A. F., Melong, N., Veinotte, C. J., Conway, A.S., Wuerthele, K., Shen, M., McKnight, C. *et al.* (2019) High-throughput chemical screening identifies focal adhesion kinase and aurora kinase B inhibition as a synergistic treatment combination in ewing sarcoma. Clin. Cancer Res.,77
44. Pihlajamaa, P., Sahu, B., Lyly, L., Aittomäki, V., Hautaniemi, S. and Jänne, O. A. (2014) Tissue-specific pioneer factors associate with androgen receptor cistromes and transcription programs. EMBO J., 33, 312–326
45. Foersch, S., Schindeldecker, M., Keith, M., Tagscherer, K. E., Fernandez, A., Stenzel, P. J., Pahernik, S., Hohenfellner, M., Schirmacher, P., Roth, W., *et al.* (2017) Prognostic relevance of androgen receptor expression in renal cell carcinomas. Oncotarget, 8, 78545–78555
46. Zhao, H., Leppert, J. T. and Peehl, D. M. (2016) A protective role for androgen receptor in clear cell renal cell carcinoma based on mining tcga data. PLoS One, 11, e0146505
47. Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A. and Staudt, L. M. (2016) Toward a shared vision for cancer genomic data. N. Engl. J. Med., 375, 1109–1112
48. Zhu, Y., Qiu, P. and Ji, Y. (2014) TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat. Methods, 11, 599–600
49. Wei, L., Jin, Z., Yang, S., Xu, Y., Zhu, Y. and Ji, Y. (2018) TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. Bioinformatics, 34, 1615–1617
50. Sales, G., Calura, E. and Romualdi, C. (2018) graphite: GRAPH Interaction from pathway Topological Environment. R package version 1.26.1
51. Krämer, N., Schäfer, J. and Boulesteix, A.-L. (2009) Regularized estimation of large-scale gene association networks using graphical Gaussian models. BMC Bioinformatics, 10, 384
52. Kim, S. (2015) ppcor: an r package for a fast calculation to semi-partial correlation coefficients. Commun. Stat. Appl. Methods, 22, 665–674
53. Shojaie, A. and Michailidis, G. (2010) Network enrichment analysis in complex experiments. Stat. Appl. Genet. Mol. Biol., 9, e22
54. Chang, W., Cheng, J., Allaire, J. J., Xie, Y. H. and McPherson, J. (2018) shiny: Web Application Framework for R. R package version 1.2.0
55. Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. InterJournal, Complex Syst., 1695
56. Almende B. V., Thieurmel, B. and Robert, T. (2018) visNetwork: Network Visualization using vis.js Library. R package version 2.0.4

