Network-Based Metric Space for Phenotypic Stratification of Samples Using Transcriptome Profiles

Inyoung Sung
Seoul National University
Dohoon Lee
Seoul National University
Sangseon Lee
Seoul National University
Sun Kim (sunkim.bioinfo@snu.ac.kr)
Seoul National University

Research Article

Keywords: gene interaction information, furthermore, The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO), Human Protein Atlas (HPA)

Posted Date: August 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-839818/v1

Network-Based Metric Space for Phenotypic Stratification of Samples Using Transcriptome Profiles

Inyoung Sung[1], **Dohoon Lee**[1,+], **Sangseon Lee**[2,+], **and Sun Kim**[1, 3, 4, 5, *]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
[2]BK21 FOUR Intelligence Computing, Seoul National University, Seoul, Republic of Korea
[3]Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea
[4]Institute of Engineering Research, Seoul National University, Seoul, Republic of Korea
[5]Bioinformatics Institute, Seoul National University, Seoul, Republic of Korea
[*]sunkim.bioinfo@snu.ac.kr
[+]these authors contributed equally to this work

ABSTRACT

With the advancements of high-throughput sequencing technology, several recent studies addressed the clinical/phenotypic stratification of samples by utilizing transcriptome data. However, existing stratification methods lack efficient utilization of gene interaction information, and furthermore, handling more than 20,000 genes causes the curse of high dimensionality that hinders elucidating the linkage between genetic profiles and clinical/phenotypic differences. To overcome these challenges, we propose a network-based two-step computational framework. We first reduce dimensions of transcriptome to a few tens of dimensions by mapping transcriptome to protein interaction network followed by performing network propagation algorithm and clustering analysis. Then, each network is converted into a single numeric metric by utilizing information theoretic quantification of gene expression abnormality, which results in a single sample mapping to a metric space generated by each subnetwork in the form of vectors. The proposed network-based stratification method was used to analyses Pan-Caner dataset and Oryza sativa dataset. Extensive experiments showed that our method generates a metric space that captures data-specific biological functions and improves the stratification performance compared to existing methods. Therefore, the proposed method successfully stratified the samples, addressing the problem in the complex gene space. The proposed method is implemented in Python and available at https://github.com/Sunginyoung/net_stratification.

Introduction

Biological state of an organism can be defined by the various regulatory mechanisms of genes. When regulatory disorder or malfunction occurs in the regulation mechanism within cells, the state of the organism becomes an abnormal state out of normal state. For example, cancer is an abnormal state that results from uncontrolled cell proliferation or evasion of programmed cell death[1–4]. In the case of plants, external stimuli such as drought stress alter the growth or survival by changing major physiological processes such as photosynthesis, water systems and hormone metabolism[5,6]. One of the ways to infer the state of cells is to measure the state of gene expression. With the development of sequencing technology, transcriptome data of biological samples are measured and accumulated in biobanks such as The Cancer Genome Atlas (TCGA) or Gene Expression Omnibus (GEO)[7,8]. Therefore, using huge amount of transcriptome data, it will be possible to measure the dysfunction of the regulatory mechanism according to the degree of abnormality, and furthermore, achievable to stratify the samples in a clinically and phenotypically meaningful direction through the degree of dysfunction[9,10].

Therefore, for stratification of samples, it is necessary to be able to define the distance between samples. However, it is difficult to calculate the difference between the genetic profile and the clinical or phenotype of the samples using transcriptome data for more than 20,000 genes. In the classical method, it is assumed that all genes are on the same line, that is, have equal importance regardless of order, and the gene expression represented as a vector to measure the distance between sample using basic mathematical methods such as Euclidean distance or Pearson's correlation[11,12]. Furthermore, there have several efforts to reduce gene space using prior knowledge, such as hallmark geneset in cancer or organ specific gene signature in the Human Protein Atlas (HPA) or to consider interactions between genes through biological pathways or experimental analysis[13–16]. However, these methods may not reflect the numerous biological mechanisms that can arise from relationship between genes to compare samples by interpreting gene relationships one-dimensionally. On the other hand, there are methods that consider interactions between genes using biological networks such as protein interaction network or gene regulation network[17,18].

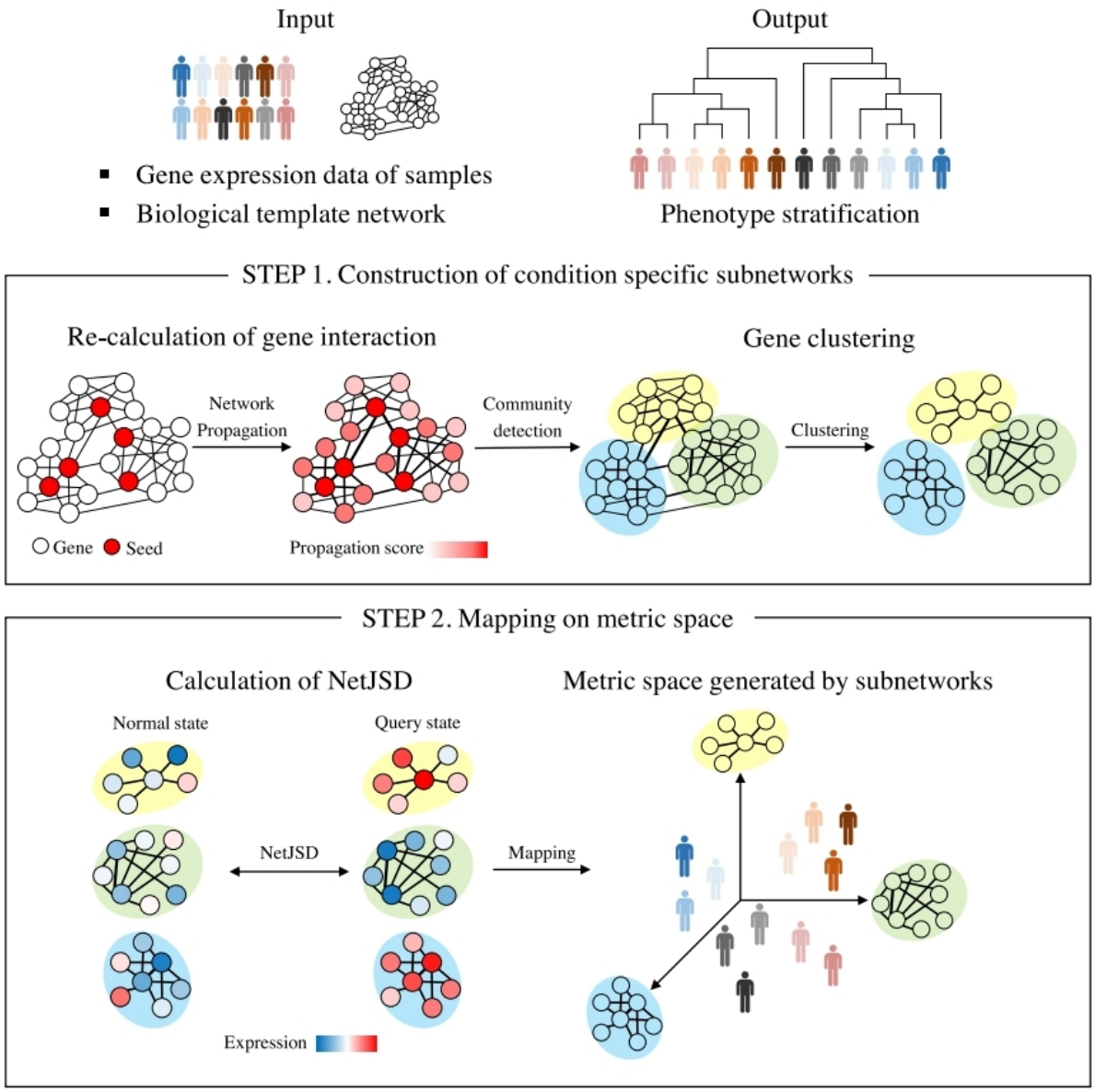

Figure 1. The overview of the proposed method. From gene expression data and biological template network, STEP 1: To generate condition specific subnetworks, gene interactions were re-calculated based on propagation score, through a network propagation algorithm with DEGs as seeds. Then, in a redefined weighted network, subnetworks were created from the results of gene clustering using a community detection algorithm. STEP 2: To measure phenotypic changes using a network, NetJSD was calculated using information theory and network structure in each subnetwork. Then, a single sample is mapped into the metric space coordinated by each subnetwork. Finally, samples are stratified in the metric space in clinically and phenotypically meaningful ways.

However, since most existing studies are using the whole network, these approaches are still insufficient for finding meaningful information unique to a given data set without using prior knowledge within a network where many information is convoluted.

Therefore, we attempted to find meaningful information by reflecting the properties of a given dataset without prior knowledge and to measure the abnormality of sample using information theory, in the biological template network that can consider the interaction of genes. In this paper, we proposed a network-based two-step computational framework for constructing metric space for measuring and stratifying samples of different phenotypes. We first reduced the dimension of 20,000 transcript genes in the template network to tens of dimensions. By mapping the transcript to a public protein interaction network, we performed gene clustering through two network analysis algorithms: network propagation and community detection algorithm, and generated tens of subnetwork from clustering result[19,20]. Then, in each subnetwork, the abnormality

in gene expression in each sample was transformed into a single metric value using network structure and information theory. As a result, a single sample is mapped into the metric space coordinated by each subnetwork in the form of a vector. We used two different public datasets: 14 cancer types from the Pan-Cancer Atlas and drought response over time for three Oryza sativa cultivars from GEO, to evaluated the proposed method. In the pan-cancer dataset, we showed that the proposed method generates subnetworks reflecting the properties of the given data in the pathway enrichment test and metric space coordinated by the subnetworks successfully stratifies the samples in the survival analysis for four clinical endpoints. In particular, we evaluated the phenotypic changes in the sample in a metric space coordinated by the subnetworks and outperformed existing methods in survival analysis for all endpoints. In the Oryza sativa cultivar dataset, we showed phenotypic changes over time and cultivars as drought stress continued. Therefore, the metric space generated by the proposed method efficiently captured clinical or phenotype information of samples and successfully stratified the samples, addressing the problem in the high-dimensional gene space of transcriptome data.

Results and Discussion

The overview of our method is showed as Figure 1. The proposed two-step computational method aimed to generate network-based metric space to measure abnormality compared to the normal state and the stratify the samples using the measured values (see details in the Methods section). This method constructed condition specific subnetworks of given gene expression data and protein-protein interaction (PPI) network using two network analysis algorithms in the first step (Figure 1. STEP 1). For the given gene expression data of normal and query samples, a network propagation algorithm was performed in PPI network using seed genes as different expression genes between normal and query[21]. As a result of propagation, the weight of edge of the PPI network was redefined using the propagation score (or diffusion score) of each gene. And then, gene clustering was performed using a community detection algorithm on the PPI network with redefined edge weight[20]. The given data specific subnetworks were constructed by constraining the network size for the result of gene clustering (here, we constrained the number of nodes and edge weight). With the subnetworks result from the first step, samples are mapped into a metric space generated by each subnetwork utilizing information theoretic quantification values in the second step (Figure 1. STEP 2). For each subnetwork, each sample was measured phenotypic changes from normal samples utilizing network level Jensen-Shannon divergence and network structure level importance. By combining these two changes as NetJSD (Equation 7), the abnormality was defined for each query sample. Finally, all query samples map into a metric space coordinated by subnetworks in the form of vectors.

Descriptions of datasets

The proposed network-based stratification method required gene expression data of normal and query samples and biological template network. For gene expression dataset, we used with two different public datasets: cancer patients from Pan-Cancer Atlas and drought response over time for three Oryza sativa cultivars from GEO. Pan-Cancer Atlas provided expression data profiled by TCGA in matched tumors and adjacent normal samples for 33 cancer types (TCGA RNASeqV2 http://www.cancergenome.nih.gov/). Among them, we used cancers with both normal and tumor patient samples present and with more than 300 tumor patient samples, and as a result, we evaluated 14 cancer types: Breast Invasive Carcinoma (BRCA) Kidney Renal Clear Cell Carcinoma (KIRC), Lung Adenocarcinoma (LUAD), Thyroid Cancer (THCA), Head-Neck Squamous Cell Carcinoma (HNSC), Uterine Corpus Endometrial Carcinoma (UCEC), Lung Squamous Cell Carcinoma (LUSC), Prostate Adenocarcinoma (PRAD), Colon Adenocarcinoma (COAD), Skin Cutaneous Melanoma (SKCM), Stomach Adenocarcinoma (STAD), Urothelial Bladder Carcinoma (BLCA), Liver Hepatocellular Carcinoma (LIHC) and Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC). The clinical data for all patient samples of all caner types were downloaded from the Genomic Data Commons portal (GDC, https://portal.gdc.cancer.gov/), and each sample was provided with clinical endpoints: overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progression-free interval (PFI).

Furthermore, we evaluated Oryza sativa cultivar(commonly known as Asian rice) dataset to showed that the proposed method can stratify samples while representing biological knowledge in addition to cancer patient dataset with clinical information. Oryza sativa cultivar dataset with drought response over time was obtained from GEO under the accession number of GSE142470 and GSE74465. The GSE142470 dataset consists of four subtypes of 20 rice cultivars[22]. The GSE74465 dataset consists of three time points (0h, 1h and 6h) RNA sequencing measured to drought resistance for three rice cultivars: Nipponbare (Oryza sativa ssp. *japonica*, Nip), Nipponbare AP2 transgenic plants (AP2), Vandana (Van)[23].

For biological template network, we used protein-protein interaction (PPI) network where nodes are genes and edges are interactions. From STRING portal (https://string-db.org), we obtained two different species PPI networks: Homo sapiens and Oryza sativa. STRING generated PPI networks with literature-based functional interactions[24]. Also, STRING provided combined scores that probabilistically calculated the interactions between gene for each edge[25], but these edge scores are not used in this paper. Homo sapiens PPI network consisted of 19,354 genes and 5,879,727 edges and Oryza sativa PPI network consisted of 25,106 genes and 4,474,524 edges.

Cancer	Normal	Tumor	Node	Edge	Network	Group1	Group2	OS	DSS	DFI	PFI
BRCA	73	842	13K	379K	12	322	520	0.0283	0.1274	0.0066	0.0212
KIRC	72	534	13K	380K	12	366	168	1.41e-7	3.64e-11	0.0563	2.1e-5
LUAD	59	517	13K	379K	13	188	329	3.88e-6	0.0008	0.1157	0.0294
THCA	59	513	12K	376K	13	266	247	0.3520	0.7165	0.0015	0.0126
HNSC	44	522	13K	380K	14	265	257	0.3961	0.1745	0.2033	0.0767
UCEC	22	533	13K	380K	12	356	177	0.0095	0.0166	0.2572	0.0284
LUSC	51	502	13K	383K	15	85	417	0.0174	0.0014	0.0319	0.0012
PRAD	52	498	12K	374K	12	232	266	0.6164	0.2385	0.1654	0.0675
COAD	41	451	13K	378K	12	168	283	0.6702	0.1122	0.2400	0.4417
SKCM	1	472	13K	382K	12	255	217	1.3e-6	6.86e-6		0.0316
STAD	35	415	12K	367K	21	111	304	0.0598	0.3716	0.0311	0.0434
BLCA	19	408	13K	379K	12	211	197	0.8555	0.6622	0.2335	0.5869
LIHC	50	374	13K	379K	13	191	183	0.0013	0.0155	0.0078	0.0062
CESC	3	306	14K	382K	15	127	179	0.3657	0.6587	0.0059	0.1582

Table 1. Pan-cancer result. The table shows the number of samples and the results for the metric space for each cancer type in the pan-cancer dataset: the number of normal and patient samples, the total number of nodes and edges in the metric space, the number of subnetworks, the number of patient samples in two groups using K-means and log rank p-value of survival analysis on four clinical endpoints. The bold test represented a significant result (p-value $<= 0.1$) in at least one survival analysis.

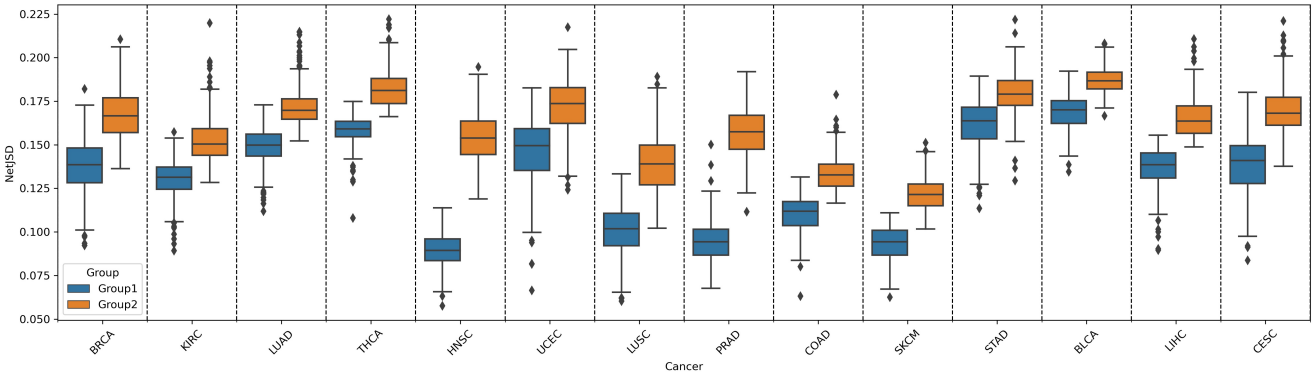

Figure 2. Boxplot of NetJSD on pan-cancer dataset. For each cancer type, the boxplot shows the average value of NetJSD across all subnetworks for the two sample groups in K-means clustering.

Pan-Cancer dataset

Table 1 shows the results for the 14 cancer types in the pan-cancer dataset. The proposed method constructed around subnetworks for all cancer types. In addition, we generated a metric space consisting of approximately 13 thousand nodes and 380 thousand edges for all cancer type, reducing the complex gene space while capturing meaningful gene interactions for a given cancer data in the template STRING network. To investigate the function of all genes that composed a metric space of each cancer type, we performed pathway enrichment test widely used pathway database Kyoto Encyclopedia of Genes and Genomes (KEGG[16]), and as a result, cancer-related pathways such as Pathways in cancer (hsa05200), PI3k-Akt signaling pathway (hsa04151), Pathways of neurodegeneration (hsa05022) and MAPK signaling pathway (has04010) were enriched at the top in all cancer types[26–28] (detailed in Supplementary file 1). In particular, pathway enrichment results showed that pathway specific to cancer type were enriched, for example, Breast cancer (has05224, p-value: 1.81e-23) in BRCA, Gastric cancer (hsa05226, p-value: 5.66e-22) in STAD, Prostate cancer (hsa05215, p-value: 1.17e-21) in PRAD, Small cell lung cancer (hsa05222, p-value: 1.55e-15) and Non-small cell lung cancer (hsa05223, p-value: 2.61e-12) in LUSC and Hepatitis B (hsa05161, p-value: 5.52e-21) and Non-alcoholic fatty liver disease (hsa04932, p-value: 8.0e-18) in LIHC.

For each cancer type, we mapped samples into a metric space coordinated by its subnetworks in the form of a vector of NetJSD values. Then, we divided samples into two groups using a K-means clustering algorithm to evaluate how well the proposed method metric space stratified the samples. As a result of K-means clustering, each cancer type was divided the samples into two groups as shown in 'Group1' and 'Group2' in Table 1 and Figure 2 shows the distribution of each group for the average value of NetJSD across all subnetworks of each cancer type, showing that the average NetJSD between two groups

Subentwork	Number of nodes	Number of edges	Top pathway	P-value
1	4590	69361	PI3K-Akt signaling pathway	6.26E-98
2	2611	114889	Ribosome	3.01E-80
3	1928	21492	Metabolism of xenobiotics by cytochrome P450	6.76E-55
4	1304	34900	Cell cycle	2.26E-74
5	767	42808	Neuroactive ligand-receptor interaction	9.61E-303
6	545	1175	Herpes simplex virus 1 infection	1.96E-263
7	479	4858	Olfactory transduction	0
8	341	3163	Glycosphingolipid biosynthesis	4.78E-16
9	224	3886	Oxidative phosphorylation	1.94E-110
10	113	1175	Arrhythmogenic right ventricular cardiomyopathy	7.53E-05
11	63	1441	Estrogen signaling pathway	6.86E-45
12	47	175	RNA degradation	0.014845

Table 2. BRCA dataset subnetworks. The table show information about the subnetworks generated by the BRCA dataset: network structure, top enriched pathway and its p-value (KEGG).

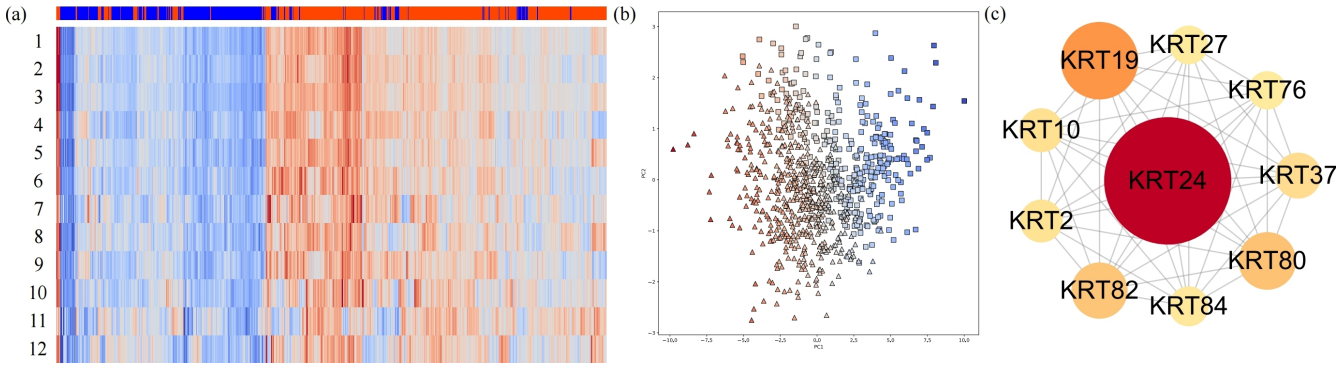

Figure 3. NetJSD on BRCA dataset. (a) Abnormality profiles of BRCA dataset on the proposed metric space. Each sample is represent as a vecor with NetJSD values. Row is subnetworks which are axes of the metric space and column is BRCA patient samples. For each sample, it is denoted by K-means group: Group1 is blue and Group2 is orange. Samples and subnetworks are reorganized by similarity measure using the Euclidean distance and hierarchical clustering. The color scheme is represented by NetJSD values normalized subnetwork-wise. (b) Two-dimensional (2D) embedding space of BRCA dataset. BRCA samples in metric space are mapped in 2D space using Principal Component Analysis (PCA). The x-axis represents Principal Component 1 and the y-axis represents Principal Component 2. The color of the point represents the average value of NetJSDs and the shape of the point represent the K-means groups: Group1 is a blue square, Group2 is an orange triangle. (c) Subsubnetwork of Subnetwork 11. This network is a subsubnetwork for 10 genes with large differences between the two groups in Subnetwork 11, where the estrogen signaling pathway is concentrated: the bigger the difference, the bigger the letters and circles and the redder the color.

is significantly different. Using K-means two groups, we perform survival analysis on four clinical endpoints: OS, DSS, DFI and PFI (Kaplan-Meier curves for each cancer type in Supplementary figure). As a result, we showed that the log rank p-value of the survival analysis between two groups was below 0.1, which was significant result in all cancer types except COAD and BLCA. Interestingly, the endpoint with the lowest p-value for each cancer was different, which can be said to capture the properties of the cancers in the TCGA dataset[29]. Therefore, we showed that the proposed method reduces the complex genetic space by reflecting the properties of a given cancer type and generates a metric space that stratifies cancer patient samples clinically meaningful way. For more detailed interpretation and evaluated the results, we used the BRCA dataset to analyze the function of the subnetworks, the measured phenotypic change for each subnetwork and the sample stratification result compare to the existing methods.

BRCA dataset

The proposed method generated 12 subnetworks with 13,012 gene and 379,546 edges for the BRCA dataset, as shown in Table 2. As a result of the pathway enrichment test using KEGG to investigate the biological function of the subnetworks, the subnetworks were enriched in cancer-related pathways such as PI3k-Akt signaling pathway (hsa04151), Cell cycle (hsa04110),

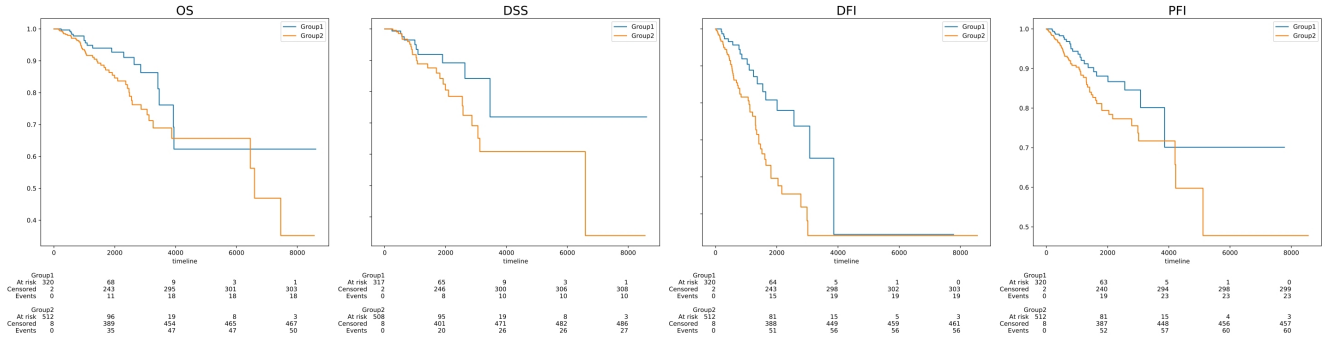

Figure 4. Survival analysis results on four clinical endpoints of the BRCA dataset. BRCA samples divided into two groups using a K-means clustring algorithm and performed survival analysis using four endpoints: OS, DSS, DFI and PFI. The plots show Kaplan-Meier curves of each endpoint (see Table 1 together for the log rank p-value.)

Endpoint	Euclidean distance	tITH	TSD	JSD	NetJSD
OS	0.3828	0.1369	0.0363	0.0346	**0.0283**
DSS	0.4400	0.3261	0.1932	0.2410	**0.1274**
DFI	0.0709	0.1745	0.0887	0.0684	**0.0066**
PFI	0.1676	0.2968	0.1597	0.0923	**0.0212**

Table 3. Performance comparison result. The table shows the p-value of the survival analysis results on four clinical endpoints using the three existing methods (Euclidean distance, tITH and TSD), the network level JSD value without the network structure level importance and the propose method. The bold value represented the best performance for each endpoint.

Olfactory transduction (hsa04740) and Pathwaysin cancer (hsa05200), and in particular, it showed enrichment results in breast cancer-related pathways such as Estrogen signaling pathway (hsa04915) and Breast cancer (has05224) (detailed in Supplementary file 2)[27,28,30–35]. Therefore, we showed that the propose method generated subnetworks that capture the properties of breast cancer patient samples.

Using subnetwork generated BRCA dataset, we mapped samples into a metric space in the form of a vector with length 12 and then, we divided samples into two groups using K-means clustering algorithm. Figure 3 shows the results of BRCA dataset mapped into a metric space. Figure 3 (a) shows the NetJSD results for each sample measured in each subnetwork and K-means groups of each sample (Group 1 is blue and Group 2 is orange) and Figure 3 (b) is the result of mapping samples NetJSD values into a two-dimensional space using Principal Component Analysis (PCA, Group1 is a blue square, Group2 is an orange triangle and the color is the average NetJSD value of the entire subnetwork). And Figure 3 (c) shows the subsubnetwork for top genes with the large differences between the two groups (the bigger the difference, the bigger the letters and circles and the redder the color).We showed that the NetJSD values between two groups differ significantly and samples from the same group have similar NetJSD trends across most of the subnetworks, as shown in Figure 2 and Figure 3 (b). In particular, when the difference between the two groups was measure, it showed an interesting result that the top genes belongs to all Subnetwork 11, which is enrich in the Estrogen signaling pathway (Figure 3 (c)). Furthermore, we performed survival analysis on four clinical endpoints and produced a Kaplan-Meier (KM) plot to investigate the BRCA dataset stratification in a metric space (Figure 4). As a result, we showed that the log rank p-value is significant on all endpoints except DSS and the higher the NetJSD values (that is, Group2 with the larger NetJSD) the worse the prognosis. Therefore, these results showed that a BRCA metric space successfully represents abnormalities in cancer patient samples and, consequently, well stratified samples for clinical prognosis.

We performed comparison analysis to evaluate how well samples are stratified in a clinically meaningful way. We compared our method with classical mathematical method (Euclidean distance) and two existing methods (tITH[18] and TSD[13]). In addition, we compared the values of the network level JSD without using network structure information (called JSD) to evaluated the importance of the combination of two proposed changes. Each comparison method was calculated each samples abnormality in the following way. Euclidean distance is a classical and basic distance method of measuring the length of a linear segment between two points. From Euclidean distance, we measured the distance of the sample from the normal sample by setting each sample as a point in the gene space generated using 15,477 genes in STRING network. tITH is a method of measuring the heterogeneity of the query sample compared to the normal sample using information theory in the network. From tITH, we measured the distance of the sample from the normal sample using STRING with 15,477 genes and 387,144 links. And TSD is a method of measuring the distance between normal and query using information theory and rank correlation for tissue specific

Subentwork	Number of genes	Number of edges	Top pathway	P-value
1	433	5,219	Spliceosome	4.74E-129
2	400	1,158	Phosphatidylinositol signaling system	2.66E-15
3	302	9,676	Ribosome	2.66E-15
4	274	6,745	Ribosome biogenesis in eukaryotes	8.55E-73
5	244	614	Metabolic pathways	4.47E-54
6	293	4,766	Proteasome	5.31E-46
7	169	390	Metabolic pathways	2.53E-85
8	207	900	Oxidative phosphorylation	1.15E-66
9	208	1,353	Ubiquitin mediated proteolysis	9.02E-10
10	199	3,861	Ubiquitin mediated proteolysis	7.09E-80
11	136	489	RNA transport	5.25E-38

Table 4. Oryza sativa cultivars subnetworks. The table shows information about the subenetworks generated by the Oryza sativa cultivar datast: network structure, top enriched pathway and its p-value (PlantGSEA).

signature genes provided by HPA. From TSD, we measured the distance of the sample from the normal sample using 156 breast specific signature genes in HPA. To compare the patient stratification results using the phenotypic change from each method, we divided sample into two groups using K-means clustering and performed log rank test to analyze the survival of the two groups. Table 3 shows the results of the log rank p-value at the four clinical endpoints of each method (Kaplan-Meier curves for each comparison method in Supplementary figure). As a results, the proposed method outperformed the existing method in all endpoints. In addition, combining two changes, network level JSD and network structure level importance, showed better results than Network JSD with only one change. All methods showed low p-value results in OS or DFI and TSD showed the best results among comparison methods. Through the result of TSD using the prior knowledge provided by HPA, it can be said that selecting important genes according to the given dataset in the complex gene space shows better result. In addition, when comparing the JSD and NetJSD results with tITH result, it can be argued that reducing the gene space by generating subnetworks to fit the given dataset shows better performance than using the entire template network. As a results, our method significantly reduced the gene space without using prior knowledge and was able to stratify samples more than other methods in the generated metric space.

Oryza sativa dataset

In addition to the cancer patient dataset with clinical information, we evaluated changes in plants caused by stressful environment. As the drought stress continued in three Oryza sativa cultivars, we showed how each cultivar changes over time in the standard rice conditions and what difference exist between cultivars. The proposed method generated 11 subnetworks with 2,865 genes and 35,171 edges for the Oryza sativa dataset, as shown in Table 4. To investigate the function of the reduced gene space by the proposed method, we performed pathway enrichment test using plant organism gene set enrichment website PlantGSEA[36]. All genes constituting the metric space were enriched in metabolic pathways such as Biosynthesis of secondary metabolites (osa01110) and Glycerophospholipid metabolism (osa00564)[37,38] and each subnetwork was enriched in the pathways shown in the Table 4 (detailed in Supplementary file 3). Interestingly, the subnetworks showed highly enrichment in pathways known as drought resistance or survival maintenance such as the photosynthetic pathway (osa00195), the phosphatidylinositol signaling pathway (osa04070), and the oxidative phosphorylation pathway (osa00190)[23,39,40]. As a result, we showed that the proposed method generate Oryza sativa subnetworks composed of genes that respond to the drought stress and are involved in biological processes for sustaining life.

Figure 5 shows the result of nine Oryza sativa samples mapped to a metric space. Figure 5 (a) shows the NetJSD result of each sample measured in the subnetworks and Figure 5 (b) shows the values measured in each gene of consisting the subnetwork. And Figure 5 (c) showed the NetJSD values at 0h and 6h for each cultivar in the photosynthetic pathway, one of the pathway where the relationship between rice and drought resistance has been most studied[41,42]. We interpreted the results for the Oryza sativa dataset from two perspectives: time-wise and cultivar-wise. For time-wise perspective, most of the cultivars showed an increase in NetJSD values as the drought continued. In particular, the difference between 0h when drought stress start and 6h when the drought stress was sufficiently sustained was showed that 6h were on average 1.5 times greater than 0h in all cultivar in all subnetworks. These results were also confirmed in the photosynthetic pathway. In Figure 5 (c), the values of each gene in the pathway increased over time. And in particular, it showed that genes such as Os01g0869800 (chlorophyll a/b binding protein) and Os07g0513000 (ATP synthase subunit gamma, chloroplastic) that changes significantly over time, i.e., showed large abnormal expression compared to the normal rice state, were associated with drought stress[43–45]. Therefore, it can be showed that each cultivar represents in increasing direction in the proposed metric space for sustained drought stress and

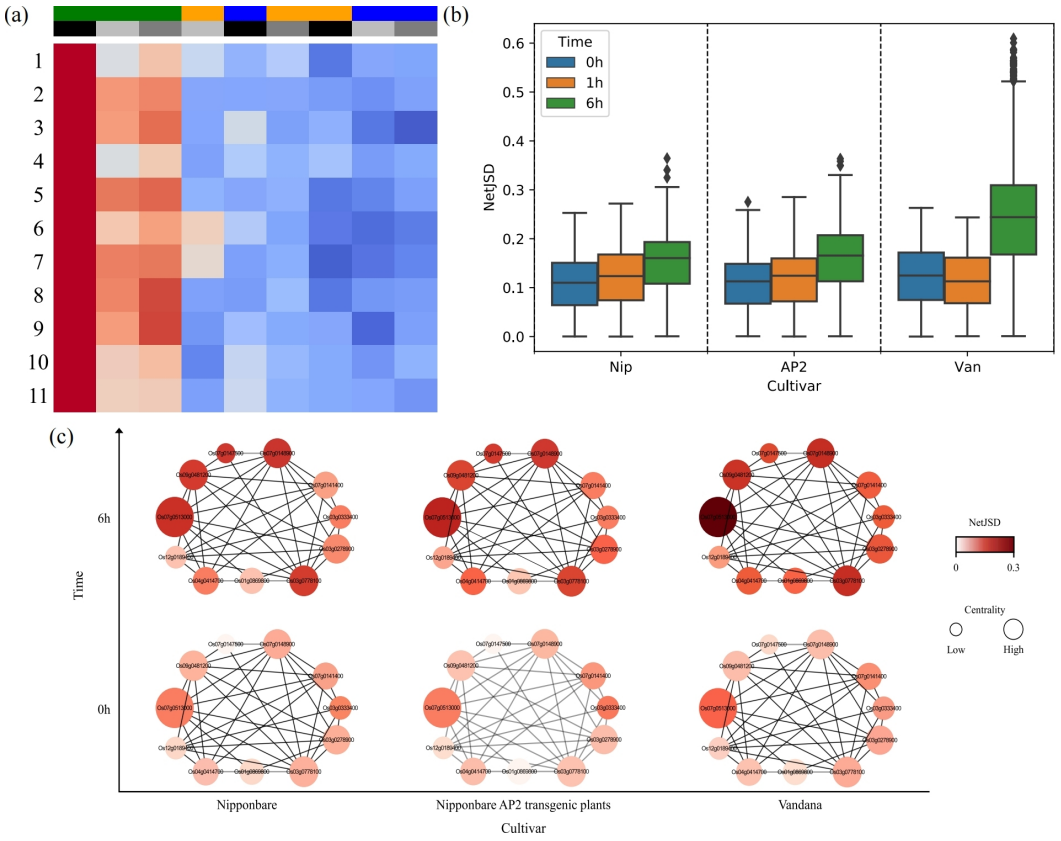

Figure 5. The results of Oryza sativa metric space. (a) Abnormality profiles of Oryza sativa dataset on the proposed metric space. Each sample on the Oryza sativa metric space is presented as a vector with NetJSD values. Row is subnetworks which are axes of the space and columns is Oryza sativa samples. For each sample, it is denoted by cultivar (Nip is silver, AP2 is grey and Van is black) and time point (0h is blue, 1h is orange and 6h is green) and each subnetwork is denoted by in Table 4. (b) Box plots of the abnormality of each gene in the metric space at each time in each cultivar. The x-axis is color-coded (upper left) with the over of time for each cultivar, and the y-axis is the abnormality value for all genes in the metric space. (c) Photosynthetic pathway in Oryza sativa metric space. The x-axis represents cultivar type, and the y-axis represents time points 0h and 6h. The color of genes in the network represents NetJSD. The size of genes represents network centrality

life sustaining efforts over time.

For cultivar-wise perspective, it can be showed that NetJSD increased in the order of Nip, AP2 and Van in most subnetworks. This result can be argued that the NetJSD in each subnetwork reflects the phenotypic properties of the rice cultivar well over time. Since Nip is a subspecies of Oryza sativa japonica and AP2 is a drought-resistant transgenic plant of Nip, AP2 will show stronger resistance to drought than Nip[46]. On the other hand, Van is a drought-resistant improved Indian highland cultivar developed in tropical japonica/aus cross, which is a strong drought-tolerant Southeast Asian rice cultivar, and is expected to survive with stronger resistance to drought than other cultivars[47–50]. Therefore, these results showed that the Oryza sativa metric space successfully represented the degree of activation of life sustaining mechanisms and the drought stress response for each cultivar, and as a result, the proposed method could stratify the three Oryza sativa cultivars over time.

Conclusion

In this paper, we proposed a novel network-based stratification method that constructs the metric space for measuring phenotype changes quantitatively using transcriptome data. To effectively handle the space of more than 20,000 genes, we developed a two-step computational framework. This method performed gene clustering in biological networks, through this result, generated condition specific subnetworks that reflect the properties of the given data. Then, for each subnetwork, we measured the distance between patients or samples using network structures and information theory to quantify the degree of abnormality compared to the normal state. As a result of experiments on Pan-caner dataset and GEO rice dataset, we showed that the proposed metric space consists of the meaningful biological functions that were specifically to the given data and successfully

stratifies the transcriptome data samples. In addition, it showed outperformed the conventional transcriptome data-based distance measurement methods in survival analysis. In particular, we showed that the results of combining two changes, network level JSD and network structure level importance, which measure the changes both gene itself and surrounding gens in the network can capture the actual phenotypic changes in the biological samples. Therefore, the proposed metric space adroitly captured the clinical or phenotype information of samples and successfully stratified the transcriptome data samples in clinically or phenotypically meaningful ways. Moreover, the proposed computational framework has potential applications as it proposes a method to measure the difference between samples no matter what kind of data comes in. As a future work, it will be much more possible to create a linear metric space for easier interpretation and comparison of samples. We are exploring several computational methods of this, including a density-based clustering method that could generate a linear ordering of data points while performing clustering analysis.

Methods

In this paper, we proposed a network-based two-step computational method: constructing condition specific subnetworks (STEP 1) and mapping samples into a metric space (STEP 2) (Figure 1).

Construction of condition specific subnetworks

Public biological network is a powerful resource that help to understand cellular mechanisms through proteins and their interaction. Among them, STRING protein-protein interaction (PPI) network is a well-known protein interaction networks that comprehensively reflects functional and physical interactions of proteins using public available resources[24]. However, since STRING predicted the interaction with an approximate probability (called confidence score in STRING) that predicted link exists between two enzymes in the same metabolic map in the KEGG, there is a limit to measuring and interpreting phenotypic change specifically for given data using the whole network[51]. Therefore, we constructed give gene expression data specific subnetworks using two network analysis algorithms: network propagation algorithm and community detection algorithm.

Re-defining of gene interactions using network propagation

In the public biological network, STRING PPI network, we used largest connected component network with a threshold of edge weight as 0.7 to use a template network with high confidence and low false positive[52]. To re-define the genetic relationships, that is, edge weight in the template network by reflecting the properties of the given gene expression data, we used a network propagation algorithm[53]. Network propagation is a method that can show the impact across the network through random and repetitive walks with nodes connected to the seed nodes, starting with given seed nodes. Here, we used genes with large expression difference between the normal and query samples as seed nodes. For initial probability distribution $\mathbf{P}_0$ and adjacency matrix $\mathbf{W}$, the probability distribution as follows:

$$\mathbf{P}_{t+1}^T = (1-\alpha)\,\mathbf{W}\,\mathbf{P}_t^T + \alpha\,\mathbf{P}_0^T \tag{1}$$

where t is the propagation step and α is restart rate that describes the trade-off between prior information and network smoothing. In this paper, we used the random walk with restart (RWR)-based method for prioritizing candidate gens using global network distance measurement and random walk analysis to define the similarity of network[21]. As result of the RWR-based method, we obtained a propagation or diffusion score of each gene. And then, we constructed a network with given data specific edge weight by re-defining the interaction of two nodes as the product of the propagation scores of the two interacted nodes.

Gene clustering using community detection

To perform gene clustering by reflecting the properties of given data to a large-scale network, we used the Louvain method on a network with re-defining edge weights a network propagation algorithm[20]. In the Louvain method, the modularity of the weighted graph is defined as: for gene v and w in network node set $\mathbf{N}$,

$$Q = \frac{1}{2m}\sum_{v,w\in\mathbf{N}}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\delta(c_v, c_w) \tag{2}$$

where m is the sum of all edge weights in the network, A_{vw} is an edge weight between genes v and w, k_v is the sum of edge weights with neighbors of the node v, δ is Kronecker delta function and c_v is the communities of the genes. Using the above modularity definition, the Louvain method first found small communities by optimizing modularity locally on the entire node, then grouped each small community into one node and repeated this method to perform community detection. As a result of a community detection algorithm, we obtained clusters of genes that maximizes modularity. Finally, we defined given data specific subnetworks by constraining the network size, here we constrained the number of nodes and edge weight, and through this result, the complex gene space is reduced to few tens of subnetwork dimension.

Map into a metric space

For each subnetwork, we measured the phenotypic changes of each sample compare to normal sample from two perspectives: network level Jensen–Shannon divergence (JSD) and network structure level importance. By combining these two measured values, each sample is mapped into a metric space generated by subnetworks in the form of vector.

Calculation of network level JSD

In information theory, JSD is a method of measuring the difference between two probability distributions, with symmetrical and finite values of Kullback–Leibler divergence (KLD). Motivated by the[18], we measured how different the influence of genes on neighboring genes in network was compared to normal, based on information theory. For each gene in the network, the probability distribution of the gene v is defined using the expression value of the neighbors $\mathbf{N}_v$:

$$P(v) = \left[p_w = \frac{e_w}{\sum_{\mathbf{N}_v} e} \; where \; w \in \mathbf{N}_v \right] \tag{3}$$

where e_w is expression value of gene w in $\mathbf{N}_v$. Then, the KLD of gene v between normal sample distribution $P^N(v) = \left[p^N(w) where w \in \mathbf{N}_v \right]$ and query sample distribution $P^Q(v) = \left[p^Q(w) where | w \in \mathbf{N}_v \right]$ is defined as:

$$KLD_v\left(P^N \parallel P^Q\right) = KLD_v\left(P^N(v) \parallel P^Q(v)\right) = \sum_{w \in \mathbf{N}_v} p^N(w) \log\left(\frac{p^N(w)}{p^Q(w)}\right). \tag{4}$$

Utilizing the KLD defined in the network, we defined the network level JSD of gene v between normal and query as follows:

$$JSD_v\left(P^N \parallel P^Q\right) = \frac{1}{2} KLD_v\left(P^N \parallel P^M\right) + \frac{1}{2} KLD_v\left(P^Q \parallel P^M\right) \tag{5}$$

where $P^M = \frac{1}{2}\left(P^N + P^Q\right)$.

Calculation of network structure level importance

When defining the phenotypic change of a query sample, the probabilistic change based on information theory calculated the change of the same value for gene set with different expression but the same proportion. To compensate for this, we measured the importance of each gene in the network structure. Closeness centrality of node is a value that evaluate the degree of the centrality of the node in the network and is defined as the reciprocal for the sum of shortest path lengths between node and all other nodes in the network:

$$CC_v = \frac{|\mathbf{N}| - 1}{\sum_{w \in \mathbf{N}/\{v\}} d(v,w)} \tag{6}$$

where $\mathbf{N}$ is set of nodes in network, $|\mathbf{N}|$ is the number of nodes and $d(v,w)$ is shortest pathway between v and w. We used the centrality obtained by each gene in the network as the importance of the gene in a network structure.

Calculation of NetJSD

In the subnetworks constructed utilizing the properties of given data in the template network, we defined the phenotypic change in each gene of the subnetwork as the product of the two level changes, network level JSD and network structure level importance. And then, the average value of the change value of all genes in the subnetwork was defined as the phenotypic change of sample in the subnetwork NetJSD:

$$NetJSD = \frac{1}{|\mathbf{N}|} \sum_{v \in \mathbf{N}} CC_v \times JSD_v\left(P^N \parallel P^Q\right). \tag{7}$$

As a result, we computed NetJSD on all subnetworks and then mapped a single sample S into a metric space generated by subnetworks in the form of following vector:

$$S = [NetJSD_1, \cdots, NetJSD_M] \tag{8}$$

where M is the number of subnetworks.

References

1. Sager, R. Expression genetics in cancer: shifting the focus from dna to rna. *Proc. Natl. Acad. Sci.* **94**, 952–955 (1997).
2. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).
3. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683–692 (2007).
4. Cairns, R. A., Harris, I. S. & Mak, T. W. Regulation of cancer cell metabolism. *Nat. Rev. Cancer* **11**, 85–95 (2011).
5. Bhargava, S. & Sawant, K. Drought stress adaptation: metabolic adjustment and regulation of gene expression. *Plant breeding* **132**, 21–32 (2013).
6. Singh, D. & Laxmi, A. Transcriptional regulation of drought response: a tortuous network of transcriptional factors. *Front. plant science* **6**, 895 (2015).
7. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat. genetics* **45**, 1113–1120 (2013).
8. Clough, E. & Barrett, T. The gene expression omnibus database. *Stat. genomics* 93–110 (2016).
9. Bertucci, F. *et al.* Gene expression profiles of poor-prognosis primary breast cancer correlate with survival. *Hum. molecular genetics* **11**, 863–872 (2002).
10. Zhong, X., Yang, H., Zhao, S., Shyr, Y. & Li, B. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. *BMC genomics* **16**, 1–8 (2015).
11. Pereira, V., Waxman, D. & Eyre-Walker, A. A problem with the correlation coefficient as a measure of gene expression divergence. *Genetics* **183**, 1597–1600 (2009).
12. Li, W. V., Chen, Y. & Li, J. J. Trom: A testing-based method for finding transcriptomic similarity of biological samples. *Stat. biosciences* **9**, 105–136 (2017).
13. Manatakis, D. V., VanDevender, A. & Manolakos, E. S. An information-theoretic approach for measuring the distance of organ tissue samples using their transcriptomic signatures. *Bioinformatics* (2020).
14. Kim, S.-K. *et al.* Identification of a molecular signature of prognostic subtypes in diffuse-type gastric cancer. *Gastric Cancer* **23**, 473–482 (2020).
15. Rapin, N. *et al.* Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in aml patients. *Blood* **123**, 894–904 (2014).
16. Kanehisa, M. & Goto, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30 (2000).
17. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. methods* **10**, 1108–1115 (2013).
18. Park, Y., Lim, S., Nam, J.-W. & Kim, S. Measuring intratumor heterogeneity by network entropy using rna-seq data. *Sci. reports* **6**, 1–12 (2016).
19. Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. genetics* **47**, 106–114 (2015).
20. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. statistical mechanics: theory experiment* **2008**, P10008 (2008).
21. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *The Am. J. Hum. Genet.* **82**, 949–958 (2008).
22. Zhao, L. *et al.* Integrative analysis of reference epigenomes in 20 rice varieties. *Nat. communications* **11**, 1–16 (2020).
23. Ahn, H. *et al.* Transcriptional network analysis reveals drought resistance mechanisms of ap2/erf transgenic rice. *Front. plant science* **8**, 1044 (2017).
24. Szklarczyk, D. *et al.* String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research* **47**, D607–D613 (2019).
25. Von Mering, C. *et al.* String: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic acids research* **33**, D433–D437 (2005).
26. Plun-Favreau, H., Lewis, P. A., Hardy, J., Martins, L. M. & Wood, N. W. Cancer and neurodegeneration: between the devil and the deep blue sea. *PLoS genetics* **6**, e1001257 (2010).
27. Sever, R. & Brugge, J. S. Signal transduction in cancer. *Cold Spring Harb. perspectives medicine* **5**, a006098 (2015).

28. Sanchez-Vega, F. *et al.* Oncogenic signaling pathways in the cancer genome atlas. *Cell* **173**, 321–337 (2018).

29. Liu, J. *et al.* An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416 (2018).

30. Zhang, J., Le, T. D., Liu, L. & Li, J. Inferring mirna sponge co-regulation of protein-protein interactions in human breast cancer. *BMC bioinformatics* **18**, 1–12 (2017).

31. Slattery, M. L. *et al.* The pi3k/akt signaling pathway: associations of mirnas with dysregulated gene expression in colorectal cancer. *Mol. carcinogenesis* **57**, 243–261 (2018).

32. Fan, S. *et al.* Brca1 inhibition of estrogen receptor signaling in transfected cells. *Science* **284**, 1354–1356 (1999).

33. Matthews, J. & Gustafsson, J.-Å. Estrogen signaling: a subtle balance between erα and erβ. *Mol. interventions* **3**, 281 (2003).

34. Caldon, C. E., Daly, R. J., Sutherland, R. L. & Musgrove, E. A. Cell cycle control in breast cancer cells. *J. cellular biochemistry* **97**, 261–274 (2006).

35. Jones, L. *et al.* Activation of estrogen signaling pathways collaborates with loss of brca1 to promote development of er α-negative and er α-positive mammary preneoplasia and cancer. *Oncogene* **27**, 794–802 (2008).

36. Yi, X., Du, Z. & Su, Z. Plantgsea: a gene set enrichment analysis toolkit for plant community. *Nucleic acids research* **41**, W98–W103 (2013).

37. Do, P. T. *et al.* Dissecting rice polyamine metabolism under controlled long-term drought stress. *PLoS One* **8**, e60325 (2013).

38. Tamiru, M. *et al.* A cytochrome p450, osdss1, is involved in growth and drought stress responses in rice (oryza sativa l.). *Plant molecular biology* **88**, 85–99 (2015).

39. Farooq, M. *et al.* Physiological role of exogenously applied glycinebetaine to improve drought tolerance in fine grain aromatic rice (oryza sativa l.). *J. Agron. Crop. Sci.* **194**, 325–333 (2008).

40. Faraji, S., Chari, G. & Najafi-Zarrini, H. Phosphatidylinositol pathway-associated genes adjust the rice growth and stress signaling: A global assay of the 5ptase family in the oryza sativa genome. *Plant Gene* **23**, 100244 (2020).

41. Cornic, G. & Massacci, A. Leaf photosynthesis under drought stress. In *Photosynthesis and the Environment*, 347–366 (Springer, 1996).

42. Reddy, A. R., Chaitanya, K. V. & Vivekanandan, M. Drought-induced responses of photosynthesis and antioxidant metabolism in higher plants. *J. plant physiology* **161**, 1189–1202 (2004).

43. Xu, Y.-H. *et al.* Light-harvesting chlorophyll a/b-binding proteins are required for stomatal response to abscisic acid in arabidopsis. *J. experimental botany* **63**, 1095–1106 (2012).

44. Ereful, N. C. *et al.* Rna-seq reveals differentially expressed genes between two indica inbred rice genotypes associated with drought-yield qtls. *Agronomy* **10**, 621 (2020).

45. Bashyal, B. M., Parmar, P., Zaidi, N. W. & Aggarwal, R. Molecular programming of drought challenged trichoderma harzianum bioprimed rice (oryza sativa l.). *Front. Microbiol.* **12**, 573 (2021).

46. Oh, S.-J. *et al.* Overexpression of the transcription factor ap37 in rice improves grain yield under drought conditions. *Plant Physiol.* **150**, 1368–1379 (2009).

47. Kumar, A. *et al.* Breeding high-yielding drought-tolerant rice: genetic variations and conventional and molecular approaches. *J. experimental botany* **65**, 6265–6278 (2014).

48. Henry, A., Wehler, R., Grondin, A., Franke, R. & Quintana, M. Environmental and physiological effects on grouping of drought-tolerant and susceptible rice varieties related to rice (oryza sativa) root hydraulics under drought. *Annals botany* **118**, 711–724 (2016).

49. Singh, V., Singh, R. *et al.* Rainfed rice: a sourcebook of best practices and strategies in eastern india. (2000).

50. Carrillo, M. G. C., Goodwin, P. H., Leach, J. E., Leung, H. & Cruz, C. M. V. Phylogenomic relationships of rice oxalate oxidases to the cupin superfamily and their association with disease resistance qtl. *Rice* **2**, 67–79 (2009).

51. Wodak, S. J., Pu, S., Vlasblom, J. & Seéraphin, B. Challenges and rewards of interaction proteomics. *Mol. & cellular proteomics* **8**, 3–18 (2009).

52. Bozhilova, L. V., Whitmore, A. V., Wray, J., Reinert, G. & Deane, C. M. Measuring rank robustness in scored protein interaction networks. *BMC bioinformatics* **20**, 1–14 (2019).

53. Can, T., Çamoğlu, O. & Singh, A. K. Analysis of protein-protein interaction networks using random walks. In *Proceedings of the 5th international workshop on Bioinformatics*, 61–68 (2005).

Acknowledgements

This research was supported by the Collaborative Genome Program for Fostering New Post-Genome Industry of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) (NRF-2014M3C9A3063541); a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI15C3224); the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science & ICT(NRF-2019M3E5D307337511); the Ministry of Food and Drug Safety (DY0002258224).

Author contributions statement

I. S., D. L. and S. K. designed the study, I. S. and S. L. conducted the experiment and all authors analyzed the results and wrote the manuscript. All authors reviewed the manuscript.

Additional information

Competing interests The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfigure.zip
- Supplementaryfile1.xlsx
- Supplementaryfile2.xlsx
- Supplementaryfile3.xlsx