Supplementary: Multi-View Kernel Learning Using Genomic Data for Identification of Disease Genes

Ekta Shah and Pradipta Maji

S1 Description of Data Sets

In the present section, a brief description of the different data sets, used to evaluate the performance of the proposed algorithm, is reported.

S1.1 Gene Expression Data Sets

In the present study, five gene expression data sets have been used to evaluate the performance of the proposed algorithm. A detailed description of the data sets is reported below:

- **GSE25070:** It is the gene expression data retrieved from study of Hinoue et al. [1]. The data set contains the expression profiles of 26 colorectal tumors matched histological to normal adjacent colonic tissue samples. Illumina Ref-8 whole-genome expression BeadChip with 24526 probes corresponding to 18491 genes was used to obtaining the gene expression profiles.
- **GSE24514:** It is the gene expression data of human MSI colorectal cancer and normal colonic mucosa, prepared by Alhopuro et al. [2] using the Affymetrix Human Genome U133A Array. It contains the expression profiles of 34 MSI colorectal cancers and 15 normal colonic mucosas. The Affymetrix array contains the expression data for 22283 probes, which map to 12985 genes.
- **West et al.:** The breast cancer data set contains expression levels of 7129 genes in 49 breast tumor samples [3]. The samples are classified according to their estrogen receptor (ER) status: 25 samples are ER positive while other 24 samples are ER negative.
- **GSE42568:** The gene expression data compares 104 breast cancer biopsies against 17 normal breast tissues, using the Affymetrix Human Genome U133 Plus 2.0 Array. It contains the expression data for 45782 probes, which correspond to 22277 genes [4].
- **GSE44861:** It is a gene expression data from the tissues of colon cancer patients, which is obtained using the Affymetrix HT Human Genome U133A Array. It contains the expression profiles of 111 colonic samples that were obtained from 56 tumors and 55 adjacent noncancerous tissues. It contains the expression profiles for 22277 probes that correspond to 13496 genes [5].

Each microarray data set is pre-processed by standardizing each sample to zero mean and unit variance. The genes that are common to both the PPI network and gene expression data only are further considered. Moreover, the present study is based on the assumption that disease-associated genes are differentially expressed and closely connected to each other. So, the genes with a low variance, low differential expressibility and weak connectivity are discarded from the gene sets obtained above. The genes that have a variance below 0.5% of the maximum variance in the set are discarded from further consideration. In order to remove genes that exhibit poor differential expressibility and low connectivity in the PPI network, their average relevance and average degree of connectivity are considered as thresholds. Hence, the genes that have both their relevance and degree below the thresholds are discarded from further processing. Thus, it leads to the generation of a reduced set of 5485, 5706, 6758 and 5859 genes for GSE25070, GSE24514, GSE42568 and GSE44861 data, respectively. It must be noted that the breast cancer data set by West et al. already has a low cardinality, hence no further processing is performed on it.

S1.2 Disease Gene Lists

In order to evaluate the performance of the proposed algorithm, some cancer-related gene lists corresponding to each of the two forms of cancer are used. A detailed description of the gene lists is reported as follows:

- List A: It contains a set of disease-related genes that have been curated based on their expression profiles. The colon cancer related list contains 531 genes [6], [7] and the breast cancer related list conatins 1380 genes [8].
- List B: The list contains 616 and 749 biologically validated, disease related genes, related to colon and breast cancer, respectively. The genes are enlisted in the *Atlas of Genetics and Cytogenetics in Oncology and*

• *E. Shah and P. Maji are with the Biomedical Imaging and Bioinformatics Lab,Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. E-mail: {ekta_r,pmaji}@isical.ac.in.*

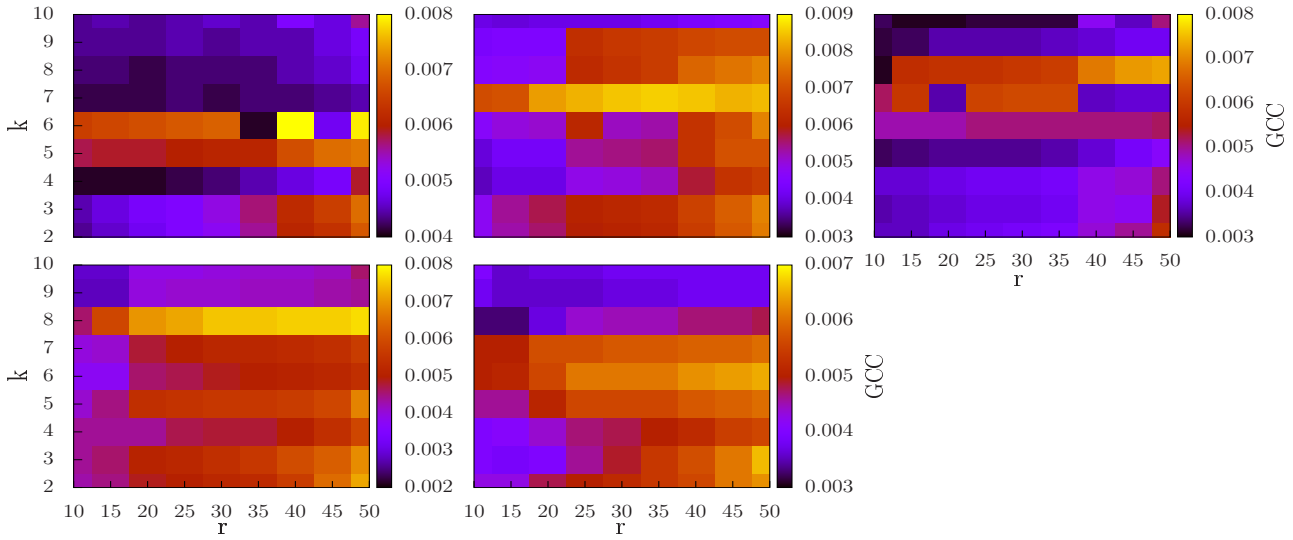

Fig. S1. Variation of GCC for GSE25070, GSE44861, GSE24514 (top row, left to right), West et al. and GSE42568 (bottom row, left to right).

Haematology [9], [10] and the *COSMIC Cancer Gene Census* [11].

- List C: It enlists genes that are extracted from the disease ontology database [12]. The list contains 249 and 239 genes related to colon and breast cancer, respectively.
- List AB: The list is created by merging List A and List B. It contains a total of 1088 and 1431 genes for colon and breast cancer, repsectively.
- List AC: It is formed by merging List A and List C. This list contains 769 and 1554 genes that are associated to colon and breast cancer, repsectively.
- List BC: This gene list is formed by merging two sets of biologically validated gene lists, namely, List B and List C. The gene list corresponding to colon and breast cancer contain 795 and 989 genes, respectively.
- List ABC: The list is formed by combining all the genes found in List A, List B and List C. Herein, 1251 and 1602 genes related to colon and breast cancer, respectively, are listed.

S2 Experimental Results and Discussion

This section reports some additional results to establish the importance of the proposed multi-view kernel learning based approach for the identification of potential disease genes.

S2.1 Selection of Optimum Parameters

From Section 2.2 of the main paper, it is known that the proposed approach depends on three regularization parameters, namely, α, β and γ along with the number of cluster indicator vectors r and number of clusters k. In order to determine an optimal value of the regularization parameters, an extensive analysis has been performed. Thereafter, the value of α, β and γ parameters has been fixed to 0.1, 0.01 and 0.5, respectively, for all the data sets being used.

In the present study, a graph clustering coefficient (GCC) based approach is used to determine an optimal $r^\star$ and $k^\star$. Fig. S1 presents the variation of GCC for different combinations of (r, k) on five omics data sets. Careful analysis of the heat maps reported in Fig. S1 shows that smaller values of the parameter r attain relatively lower GCC values. This establishes that considering too small values of the parameter may lead to the loss of meaningful information, thereby affecting the quality of the learned clusters. Moreover, analysis of the GCC values with respect to the parameter k shows that partitioning the data set into a large number of clusters also has an adverse affect. In the present study, the value of r is varied in the interval $[10, 50]$ in gaps of 5 and the value of k is varied in the range $[2, 10]$, while considering $\alpha = 0.1$, $\beta = 0.01$ and $\gamma = 0.5$. The selected parameters for the GSE25070, GSE24514, GSE44861, West et al. and GSE42568 data sets are (40,6), (35,7), (50,8), (50,8) and (50,3), respectively.

S2.2 Importance of Proposed Weight Updation Criterion

This section reports the comparative performance analysis between fixed and proposed weight updation criteria. The degree of overlapping based comparative analysis using the GSE44861 data set, is reported in Fig. 2 of main paper and Fig. S2. Careful analysis of the graphs, reported in Fig. S2, shows that the proposed weight learning strategy outperforms the fixed network weight based approach in 4 out 7 cases. In the remaining 3 cases, their performance is comparable to each other.

The biological significance analysis is performed using biological process (BP), cellular components (CC) and molecular function (MF) of gene ontology (GO), along with disease ontology (DO) and KEGG pathway enrichment analysis. Their corresponding results are reported in Table 3 of main paper, Table S1 and Table S2. Analysis of Table 3 of main paper and Table S1 shows that the proposed weight updation strategy annotates to disease relevant BP, KEGG and DO terms with significantly lower p-values. Table S2 depicts the comparative study between the two strategies based on the annotated CC and MF terms. Careful analysis

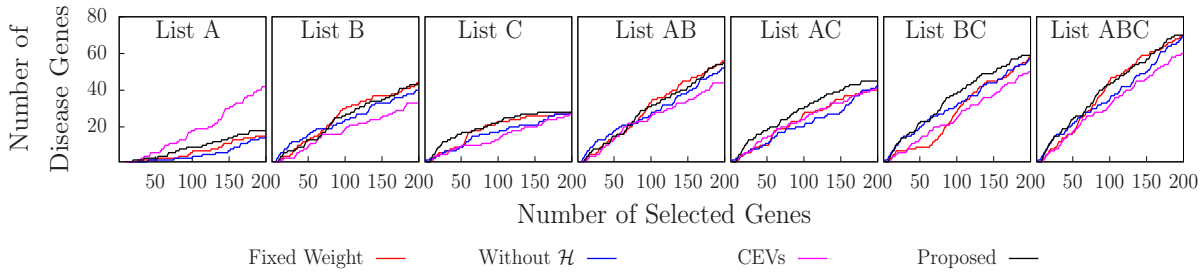

Fig. S2. Degree of overlapping based comparative study to establish the importance of the proposed approach over other strategies on GSE44861.

TABLE S1
Comparative Performance Analysis Between Proposed Approach and Other Strategies Based on Annotated BP, KEGG and DO

Data Sets	Different Approaches	GO: Biological Process		KEGG Pathway		Disease Ontology		WS_GoD
		Term	p-value	Term	p-value	Term	p-value	
GSE44861	Fixed	anatomical structure morphogenesis	8.12E-40	MAPK signaling pathway	2.35E-11	colon cancer	4.09E-08	**0.162**
	Without $\mathcal{H}$	regulation of programmed cell death	2.90E-33	MAPK signaling pathway	1.15E-09	colon cancer	5.23E-09	0.141
	CEVs	tube morphogenesis	6.13E-38	MAPK signaling pathway	1.27E-08	colon cancer	5.16E-10	0.133
	Proposed	BP1	2.26E-46	MAPK signaling pathway	9.48E-15	colon cancer	4.09E-10	0.146

TABLE S2
Comparative Performance Analysis Between Proposed Approach and Other Strategies Based on Annotated MF and CC Terms

Data Sets	Weight Updation	GO: Molecular Function		GO: Cellular Components	
		Term	p-value	Term	p-value
GSE25070	Fixed	transcription factor binding	2.09E-09	transcription regulator complex	8.34E-09
	Without $\mathcal{H}$	DNA-binding transcription activator activity	2.77E-17	chromatin	1.40E-18
	CEVs	endopeptidase activity	2.20E-11	intracellular protein-containing complex	3.20E-14
	Proposed	regulation of kinase activity	6.39E-22	membrane raft	5.20E-11
GSE24514	Fixed	cytokine receptor binding	2.34E-44	cell surface	2.20E-20
	Without $\mathcal{H}$	signaling receptor binding	3.21E-30	cytoplasmic vesicle lumen	1.16E-06
	CEVs	transcription regulator activity	6.27E-37	chromatin	2.50E-31
	Proposed	transcription regulator activity	1.86E-40	chromatin	3.46E-29
West et al.	Fixed	positive regulation of transferase activity	6.38E-24	anchoring junction	1.15E-14
	Without $\mathcal{H}$	DNA-binding transcription activator activity	3.12E-15	chromatin	1.50E-16
	CEVs	DNA-binding transcription activator activity, RNA polymerase II-specific	3.12E-16	chromatin	2.98E-17
	Proposed	cytokine receptor binding	3.09E-33	cell surface	1.55E-25
GSE42568	Fixed	transcription factor binding	2.42E-17	chromatin	1.37E-19
	Without $\mathcal{H}$	transcription factor binding	2.01E-20	chromatin	3.20E-22
	CEVs	transcription factor binding	5.22E-14	transcription regulator complex	2.89E-07
	Proposed	positive regulation of catalytic activity	3.87E-39	anchoring junction	3.02E-13
GSE44861	Fixed	protein kinase activity	1.71E-24	chromatin	1.21E-16
	Without $\mathcal{H}$	transcription factor binding	5.24E-09	chromatin	2.35E-08
	CEVs	transcription factor binding	2.51E-08	chromatin	8.64E-13
	Proposed	transcription factor binding	1.96E-33	chromatin	2.29E-23

of the table shows that proposed weight learning strategy annotates to the terms with a lower p-value in 8 out of 10 cases. The biological importance of the BP, MF, CC and KEGG terms that are annotated by the proposed approach, is established in Section S2.7. Overall, the results reported in Fig. 2 and Table 3 of main paper, along with Tables S1 and S2 demonstrate the importance of learning the optimal view weights in the proposed approach.

S2.3 Importance of Learning Unified Kernel

The present section aims to establish the importance of learning the unified kernel $\mathcal{H}$, and using its information to update the consensus kernel $\mathcal{S}$. The present section optimizes the objective function, reported in Section 3.3 of main paper, to learn the consensus kernel $\mathcal{S}$ using a weighted combination of individual views $\mathcal{H}_i$ without learning the unified kernel $\mathcal{H}$. Fig. 2 of main paper and Fig. S2 report the degree of overlap based comparative analysis between the two approaches. The graphs reported in Fig. S2 show that the proposed approach attains a better result in all the seven cases.

Table 3 of main paper, Table S1 and Table S2 report a comparative study between the two approaches, using the annotated BP, CC, MF, KEGG and DO terms. Careful analysis of Table 3 of main paper and Table S1 shows that the proposed approach annotates to disease associated BP, KEGG and DO terms with significantly lower p-values using all the five data sets. Table S2 depicts the CC and MF based analysis between the two approaches. Analysis of the table shows that proposed approach annotates to CC and MF terms with the lowest p-value in 8 out of 10 cases. This establishes the importance of learning the unified kernel $\mathcal{H}$

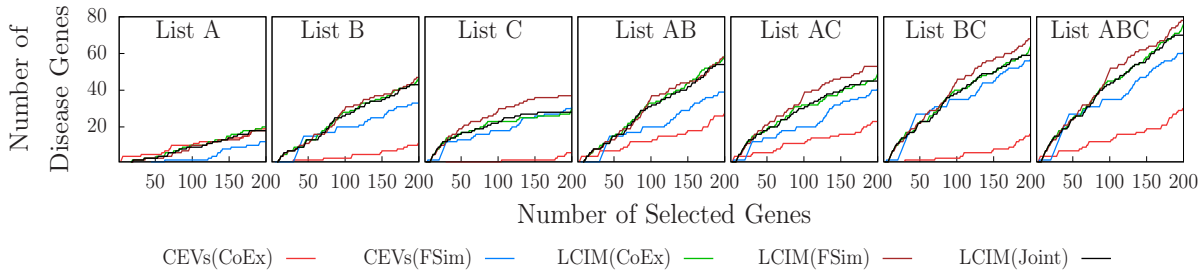

Fig. S3. Performance of the learned joint and individual cluster indicator matrices, along with the computed eigenvectors of the individual views on GSE44861.

TABLE S3
Comparative Study to Establish the Importance of Joint and Individual Clustering Approach Based on BP, KEGG and DO

Data Sets	Cluster Structure		GO: Biological Process		KEGG Pathway		Disease Ontology		WS_GoD
			Term	p-value	Term	p-value	Term	p-value	
GSE44861	CEVs	CoEx	cellular response to organic cyclic compound	7.49E-08	Sulfur metabolism	9.31E-06	colon cancer	4.56E-02	0.061
		FSim	BP1	1.92E-44	MAPK signaling pathway	3.90E-12	colorectal cancer	5.23E-09	**0.194**
	LCIM	CoEx	cellular response to organic substance	1.59E-44	MAPK signaling pathway	4.21E-13	colon cancer	2.27E-11	0.157
		FSim	cellular response to organic substance	6.98E-41	MAPK signaling pathway	3.50E-19	colon cancer	1.72E-16	0.187
	LCIM(Joint)		BP1	2.26E-46	MAPK signaling pathway	9.48E-15	colon cancer	4.09E-10	0.146

BP1: positive regulation of nitrogen compound metabolic process

in the proposed framework.

S2.4 Importance of Computed Cluster Indicator Matrix

The current section aims to establish the efficiency of the learned clustering solution over the one obtained by computing the eigenvectors (CEVs) of the learned kernel. Fig. 2 of main paper depicts the degree of overlap based comparative analysis between the two approaches for the GSE25070, GSE24514, West et al. and GSE42568 data sets. A similar analysis for the GSE44861 data set, is reported in Fig. S2. The graphs show that the proposed approach for learning the clustering solution outperforms the latter in 6 out of 7 cases.

Table 3 of main paper and Table S1 reports the BP, KEGG and DO based analysis of the two approaches for learning the cluster indicator matrix. Table S2 reports a comparative study between the two approaches based on the annotated MF and CC terms. Analysis of the table shows that the proposed approach annotates to disease-associated terms with the lowest p-value in 8 out of 10 cases. It has also been observed that the reduced set of effective genes $\mathcal{R}$ extracted using the computed eigenvectors based approach contain more than 70% of the complete gene set. On the other hand, the proposed approach generates a relatively smaller and meaningful set of effective genes. Overall, the results reported in Table 3 of main paper and Tables S1 and S2 establish the efficiency of the learned cluster indicator matrix based approach in curating a reduced set of potential biomarkers.

S2.5 Importance of Joint and Individual Clustering

The present section aims to compare the performance of the computed eigenvectors based clustering solution of the individual views with that of the learned joint and individual cluster indicator matrix. Fig. 3 of main paper and Fig. S3 reports the degree of overlapping based comparative study between these approaches. The biological analysis of the gene sets curated using the different approaches is reported in Table 4 of main paper, along with Tables S3 and S4.

Fig. S3 depicts the comparative study using the GSE44861 data set. Analysis of the graphs reported in the figure shows that the learned joint clustering solution outperforms the individual views in 5 out of 7 cases. BP, KEGG and DO based analysis of the two approaches is reported in Table 4 of main paper and Table S3. Careful analysis of the latter table shows that the learned joint clustering solution annotates to the terms with lowest p-value in all three cases. Moreover, CC and MF based analysis of the two approaches show that the learned joint clustering solution annotates to disease relevant terms with the lowest p-value in 5 out of 10 cases. It must also be noted that the functional similarity network cannot efficiently partition the network and extract a reduced set of effective genes. On the other hand, the proposed approach extracts a relatively smaller and meaningful set of effective genes, which is further used to curate a set of potential disease genes.

The present section also aims to evaluate the performance of the learned joint and individual cluster structures. Careful analysis of Fig. S3 shows that the performance of the learned clustering solutions is comparable to each other in all seven cases for the GSE44861 data. Similar results are observed in Table S3. Careful analysis of the annotated MF and CC terms, reported in Table S4, show that the learned joint clustering solution attains lowest p-values in 8 out 10 cases. Overall, the comparable performance of the learned joint and individual cluster structures validate the basic assumption of multi-view kernel learning and also establish the efficiency of the proposed approach in learning the consensus cluster structure.

Finally, the computed eigenvectors based clustering solution of the individual views is compared with that of

TABLE S4
Comparative Study to Establish the Importance of Joint and Individual Clustering Approach Based on MF and CC

Different Data Sets	Cluster Structure		GO: Molecular Function		GO: Cellular Components	
			Term	p-value	Term	p-value
GSE25070	CEVs	CoEx	3′-5′-exodeoxyribonuclease activity	4.54E-03	preribosome, small subunit precursor	8.93E-03
		FSim	regulation of kinase activity	1.32E-12	chromatin	1.20E-36
	LCIM	CoEx	regulation of kinase activity	9.76E-23	membrane raft	2.91E-09
		FSim	regulation of kinase activity	5.03E-21	membrane raft	6.56E-11
	LCIM (Joint)		regulation of kinase activity	6.39E-22	membrane raft	5.20E-11
GSE24514	CEVs	CoEx	catalytic activity, acting on DNA	2.11E-09	chromosomal region	1.46E-22
		FSim	positive regulation of molecular function	3.70E-59	chromatin	1.12E-17
	LCIM	CoEx	transcription regulator activity	6.97E-37	chromatin	4.22E-24
		FSim	mRNA binding	6.40E-08	intracellular protein-containing complex	1.57E-12
	LCIM(Joint)		transcription regulator activity	1.86E-40	chromatin	3.46E-29
West et al	CEVs	CoEx	protein kinase binding	7.93E-08	focal adhesion	2.19E-06
		FSim	cytokine receptor binding	2.51E-45	cell surface	7.79E-39
	LCIM	CoEx	cytokine receptor binding	2.65E-03	somatodendritic compartment	9.43E-14
		FSim	DNA-binding transcription activator activity,RNA polymerase II-specific	5.55E-19	somatodendritic compartment	5.32E-11
	LCIM(Joint)		cytokine receptor binding	3.09E-33	cell surface	1.55E-25
GSE42568	CEVs	CoEx	bile acid binding	2.18E-04	membrane microdomain	3.71E-05
		FSim	positive regulation of catalytic activity	1.76E-38	anchoring junction	1.86E-10
	LCIM	CoEx	positive regulation of catalytic activity	3.49E-38	anchoring junction	6.97E-12
		FSim	positive regulation of catalytic activity	9.15E-21	chromatin	2.23E-23
	LCIM(Joint)		positive regulation of catalytic activity	3.87E-39	anchoring junction	3.02E-13
GSE44861	CEVs	CoEx	sodium ion transmembrane transporter activity	3.78E-05	apical part of cell	9.81E-10
		FSim	transcription factor binding	6.21E-15	membrane microdomain	5.73E-26
	LCIM	CoEx	transcription factor binding	1.96E-33	chromatin	1.45E-22
		FSim	transcription factor binding	4.13E-27	chromatin	1.14E-18
	LCIM(Joint)		transcription factor binding	1.96E-33	chromatin	2.29E-23

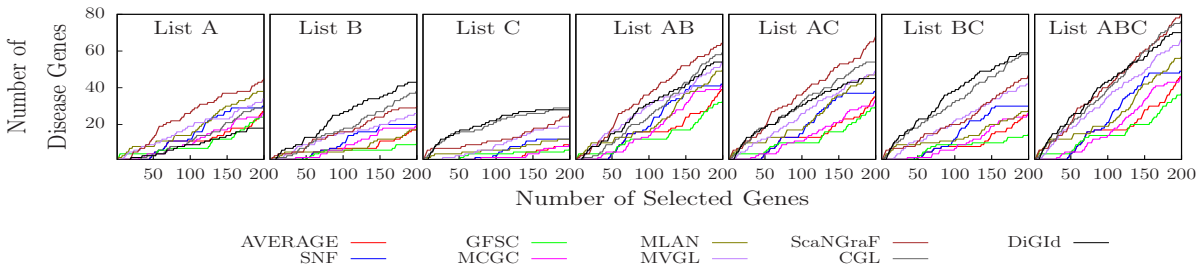

Fig. S4. Degree of overlap based comparative analysis between proposed and existing approaches of multi-view kernel learning on GSE44861.

the learned individual cluster structures. Careful analysis of the graphs reported in Fig. S3 shows that the learned clustering solution of the individual views exhibits a better performance that the computed eigenvectors based solution, in 5 out of 7 cases. Table 4 of main paper and Table S4 show that the genes selected using the learned clustering solution of individual views annotates to disease-relevant GO, KEGG and DO terms with significantly low p-values. A significant improvement in performance of the co-expression network is also observed. Careful analysis of Table S4 shows that the genes curated using the learned individual cluster structure annotate to CC and MF terms that are closely associated to the disease in 5 out of 10 cases. Overall, it can be inferred that the learned joint cluster indicator matrix performs better than the eigenvectors of the individual views. Moreover, the results reported in Fig. 3 of main paper, Fig. S3, Table 4 of main paper, Table S3 and Table S4, establish that the proposed approach efficiently deals with the incompleteness associated with the individual views and thereby, curates a set of potential disease genes.

S2.6 Comparative Study with Existing Approaches

In the current section, a detailed comparative study to demonstrate the effectiveness of the proposed DiGID algorithm is reported. A two-fold comparative study is undertaken to establish (i) the efficiency of the proposed multi-view kernel technique over existing kernel learning approaches in the domain of biomarker discovery and (ii) the potential of the DiGId over existing algorithms for disease gene selection.

S2.6.1 Multi-View Kernel Learning Techniques

Extensive analysis of the proposed and existing algorithms for multi-view kernel learning are reported in Fig. 4 of main paper, Fig. S4, Table 5 of main paper, Table S5 and Table S6. Fig S4 depicts the degree of overlap based analysis of the GSE44861 data set. It can be seen that the proposed DiGId algorithm attains better result in 3 out of 7 cases. Its performance is comparable to that of CGL in 2 of the remaining 4 cases. From Table 5 of main paper, Table S5 and Table S6, it can easily be inferred that the proposed approach annotates to disease-related GO, KEGG and DO terms with a significantly lower p-value using all the five data sets.

TABLE S5
Comparative Performance Analysis between Proposed and Existing Kernel Learning Algorithms Based on BP, KEGG and DO

Data Sets	Different Algorithms	GO: Biological Process		KEGG Pathway		Disease Ontology		WS_GoD
		Term	p-value	Term	p-value	Term	p-value	
GSE44861	AVERAGE	regulation of smooth muscle cell migration	1.33E-06	HIF-1 signaling pathway	2.51E-03	colorectal cancer	1.58E-08	0.000
	SNF	extracellular matrix organization	1.04E-08	Prostate cancer	1.09E-03	colon cancer	3.52E-03	0.000
	GFSC	flavonoid metabolic process	9.56E-05	Retinol metabolism	2.28E-04	musculoskeletal	1.12E-05	0.000
	MCGC	alcohol metabolic process	2.52E-06	Sulfur metabolism	8.33E-05	cystic fibrosis	7.17E-05	0.000
	MLAN	regulation of hormone levels	3.97E-09	Chemical carcinogenesis	4.44E-03	colon cancer	7.62E-05	0.000
	MVGL	tube morphogenesis	3.75E-12	MAPK signaling pathway	1.57E-03	colon cancer	1.51E-04	0.000
	ScaNGraF	tube development	4.27E-19	MAPK signaling pathway	2.82E-04	colon cancer	9.70E-10	**0.184**
	CGL	response to endogenous stimulus	4.56E-36	MAPK signaling pathway	1.98E-07	colon cancer	6.25E-11	0.148
	DiGId	positive regulation of nitrogen	2.26E-46	MAPK signaling pathway	9.48E-15	colon cancer	4.09E-10	0.146

TABLE S6
Comparative Performance Analysis between Proposed and Existing Kernel Learning Algorithms Based on MF and CC

Data Sets	Different Algorithms	GO: Molecular Function		KEGG Pathway	
		Term	p-value	Term	p-value
GSE25070	AVERAGE	extracellular matrix structural constituent	2.78E-05	caveola	1.59E-02
	SNF	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	3.54E-05	*	*
	GFSC	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	1.96E-05	H4 histone acetyltransferase complex	1.22E-02
	MCGC	cytokine receptor binding	3.88E-04	histone acetyltransferase complex	1.04E-03
	MLAN	NAD-retinol dehydrogenase activity	5.64E-03	microvillus	1.64E-03
	MVGL	glycosaminoglycan binding	2.48E-04	gamma-tubulin complex	7.77E-03
	ScaNGraF	signaling receptor activator activity	1.72E-13	vacuolar lumen	1.08E-03
	CGL	transcription factor binding	6.57E-10	SWI/SNF superfamily-type complex	3.55E-04
	DiGId	regulation of kinase activity	6.39E-22	membrane raft	5.20E-11
GSE24514	AVERAGE	carbonate dehydratase activity	2.72E-03	costamere	1.49E-08
	SNF	double-stranded RNA binding	4.41E-03	costamere	2.82E-03
	GFSC	regulation of cyclin-dependent protein serine/threonine kinase activity	4.40E-08	condensed chromosome, centromeric region	6.88E-20
	MCGC	positive regulation of cysteine-type endopeptidase activity involved in apoptotic process	5.02E-08	melanosome	8.60E-09
	MLAN	cytokine receptor binding	4.43E-05	cyclin-dependent protein kinase holoenzyme complex	3.35E-04
	MVGL	chemokine activity	6.54E-09	condensed chromosome	5.88E-05
	ScaNGraF	positive regulation of transferase activity	2.89E-11	membrane microdomain	1.29E-09
	CGL	chromatin binding	1.96E-07	mitotic spindle	4.51E-03
	DiGId	transcription regulator activity	1.86E-40	chromatin	3.46E-29
West et al	AVERAGE	positive regulation of kinase activity	2.51E-09	dendrite	4.35E-08
	SNF	transcription factor binding	7.69E-11	cuticular plate	3.40E-03
	GFSC	regulation of protein serine/threonine kinase activity	8.62E-06	cytoplasmic stress granule	4.76E-03
	MCGC	cytokine receptor binding	1.72E-38	cell surface	5.77E-38
	MLAN	cytokine receptor binding	1.14E-02	membrane microdomain	5.08E-07
	MVGL	regulation of binding	1.97E-04	cuticular plate	5.37E-03
	ScaNGraF	cytokine receptor binding	1.66E-18	membrane raft	9.25E-16
	CGL	cytokine receptor binding	2.20E-05	eukaryotic translation initiation factor 4F complex	6.83E-04
	DiGId	cytokine receptor binding	3.09E-33	cell surface	1.55E-25
GSE42568	AVERAGE	NAD binding	2.72E-04	focal adhesion	8.50E-08
	SNF	calcium-dependent protein kinase C activity	7.01E-11	focal adhesion	1.41E-06
	GFSC	nucleocytoplasmic carrier activity	1.50E-16	chromosome, centromeric region	2.28E-21
	MCGC	aminoacylase activity	3.69E-03	caveola	1.33E-04
	MLAN	chromatin binding	3.79E-13	nuclear protein-containing complex	1.19E-20
	MVGL	cyclin-dependent protein serine/threonine kinase activity	1.73E-04	spindle	6.92E-06
	ScaNGraF	regulation of protein kinase activity	5.64E-11	membrane raft	2.26E-06
	CGL	regulation of transferase activity	4.63E-12	membrane raft	1.50E-07
	DiGId	positive regulation of catalytic activity	3.87E-39	anchoring junction	3.02E-13
GSE44861	AVERAGE	sodium:proton antiporter activity	3.55E-03	microvillus membrane	3.60E-03
	SNF	extracellular matrix structural constituent	2.11E-06	collagen-containing extracellular matrix	5.17E-09
	GFSC	calcium-dependent protein binding	3.55E-03	*	*
	MCGC	monovalent cation:proton antiporter activity	7.13E-04	microvillus	7.04E-05
	MLAN	aryl sulfotransferase activity	7.83E-04	endoplasmic reticulum chaperone complex	8.25E-04
	MVGL	extracellular matrix structural constituent	3.11E-05	extracellular matrix	4.42E-10
	ScaNGraF	positive regulation of transferase activity	1.07E-09	collagen-containing extracellular matrix	2.14E-07
	CGL	positive regulation of molecular function	2.32E-24	membrane raft	9.03E-13
	DiGId	transcription factor binding	1.96E-33	chromatin	2.29E-23

S2.6.2 Comparison with Existing Gene Selection Algorithms

The biological significance analysis of the genes curated using the proposed DiGId algorithm and some existing gene selection algorithms is reported in Fig. 5 of main paper, Fig. S5 Table 6 of main paper, Table S7 and Table S8. Analysis of Fig. S5 shows that the proposed DiGId algorithm outperforms the existing algorithms in 2 out of 7 cases. In 2 of the

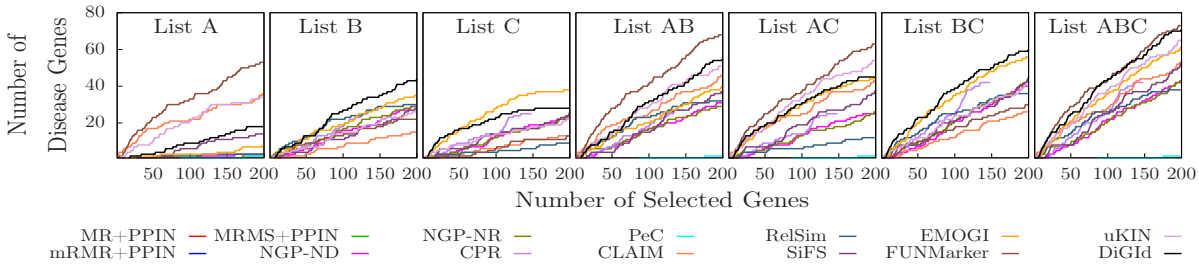

Fig. S5. Degree of overlap based comparative analysis between proposed and existing gene selection algorithms on GSE44861.

TABLE S7
Comparative Performance Analysis between Proposed and Existing Gene Selection Algorithms Based on BP, KEGG and DO

Data Sets	Different Algorithms	GO: Biological Process		KEGG Pathway		Disease Ontology		WS_GoD
		Term	p-value	Term	p-value	Term	p-value	
GSE44861	MR+PPIN	*	*	*	*	mumps	6.98E-03	0.000
	mRMR+PPIN	*	*	*	*	Barrett's esophagus	7.10E-03	0.000
	MRMS+PPIN	*	*	*	*	myeloid leukemia	1.55E-02	0.000
	NGP_ND	*	*	*	*	colon cancer	6.27E-08	0.129
	NGP_NR	*	*	*	*	colon cancer	6.58E-10	0.146
	PeC	protein auto-ADP-ribosylation	2.25E-05	*	*	*	*	0.000
	CPR	response to endogenous stimulus	1.00E-41	MAPK signaling pathway	6.87E-18	colon cancer	7.42E-15	0.148
	CLAIM	response to copper ion	3.96E-06	Mineral absorption	3.35E-02	colonic benign neoplasm	6.96E-02	0.044
	RelSim	mitotic cell cycle process	1.18E-81	Cell cycle	4.21E-29	colon cancer	1.11E-01	0.051
	SiFS	cellular response to organic substance	2.43E-56	MAPK signaling pathway	9.83E-10	colorectal cancer	7.57E-11	0.137
	EMOGI	BP1	1.34E-52	MAPK signaling pathway	5.13E-17	colon cancer	5.05E-20	**0.192**
	FUNMarker	hormone activity	1.48E-05	Prostate cancer	3.60E-06	colon carcinoma	3.99E-02	0.064
	uKIN	response to hormone	5.87E-13	MAPK signaling pathway	1.08E-02	colon cancer	3.66E-08	0.112
	DiGId	BP1	2.26E-46	MAPK signaling pathway	9.48E-15	colon cancer	4.09E-10	0.146

remaining 5 cases, they exhibit a comparable performance. Extensive analysis of Table 6 of main paper, Table S7 and Table S8 shows that the proposed approach annotates to disease-associated terms with significantly lower p-values. However, it must also be noted that existing algorithms, like RelSim, SiFS, and NGP also annotate to disease associated terms, but with relatively higher p-values. Thus, it can be inferred that the proposed DiGId algorithm for disease gene curation performs significantly better than majority of the existing gene selection algorithms.

S2.7 Biological Significance Analysis of Annotated Terms

The present section summarizes various BP, MF and DO terms, and KEGG pathways annotated by the proposed approach and briefly describes their association to the disease under study.

S2.7.1 Annotated BP Terms

The biological significance analysis of the BP terms annotated by the proposed approach is listed as follows:

1) **response to oxygen-containing compound**: The presence of oxygen containing compounds has been shown to cause protein dysfunction and DNA damage, leading to gene mutations and cell death. It also activates signaling pathways, such as NF-κB and p38 MAPK, to affect cell proliferation, differentiation, and apoptosis [13].
2) **Positive regulation of nitrogen compound metabolic process**: The term indicates that the curated gene set increases the rate of reactions that regulate the metabolism of nitrogen containing compounds. From [14], it is known that upregulated amino acids metabolism facilitates the abnormal proliferation of cancer cells and its survival.
3) **Response to cytokine**: The role of cytokines in growth and progression of breast cancer cells has been established in [15]. The study has shown that cytokines on breast cancer cell lines are actively involved in decreased cell-cell association with increased cellular motility, inhibition of cellular proliferation and several morphological changes like cell elongation and decrease in inter-cellular adhesion.
4) **Intracellular signal transduction**: The intracellular signals interact and co-ordinate with complex cellular pathways. They regulate several nuclear and cytosolic processes, which include cell diferentiation, proliferation and oncogenic transformation. The curated genes activate several antibodies, cytokines, growth factors, and ions from the extracellular region that trigger several signaling pathways. The interconnected signaling pathways may induce the breast cancer cells to proliferate and survive [16].

S2.7.2 Annotated MF Terms

The analysis of the MF terms annotated by the proposed algorithm is described next.

1) **Regulation of kinase activity**: Protein kinases are actively involved in various cellular processes which include metabolism, cell proliferation, and mediation of growth factors. They also play a key

TABLE S8
Comparative Performance Analysis between Proposed and Existing Gene Selection Algorithms Based on CC and MF

Data Sets	Different Algorithms	GO: Molecular Function		GO: Cellular Components	
		Term	p-value	Term	p-value
GSE25070	MR+PPIN	RNA polymerase II-specific DNA-binding transcription factor binding	9.68E-17	apicolateral plasma membrane	8.05E-07
	mRMR+PPIN	RNA polymerase II-specific DNA-binding transcription factor binding	5.83E-20	euchromatin	1.05E-05
	MRMS+PPIN	RNA polymerase II-specific DNA-binding transcription factor binding	3.90E-12	extrinsic component of cytoplasmic side of plasma membrane	5.28E-08
	NGP_ND	regulation of kinase activity	1.21E-22	membrane raft	7.42E-09
	NGP_NR	regulation of kinase activity	5.95E-17	transcription regulator complex	2.86E-13
	PeC	activation of GTPase activity	1.67E-02	nuclear inner membrane	1.97E-02
	CPR	regulation of kinase activity	6.13E-21	membrane raft	7.18E-09
	CLAIM	myosin II binding	3.41E-04	kinetochore microtubule	8.09E-04
	RelSim	regulation of kinase activity	1.56E-15	membrane raft	2.44E-08
	SiFS	regulation of kinase activity	5.11E-14	membrane raft	4.80E-04
	EMOGI	regulation of kinase activity	3.67E-21	membrane raft	1.20E-15
	FUNMarker	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	1.03E-04	myelin sheath	1.19E-02
	uKIN	regulation of protein kinase activity	9.03E-10	transcription factor AP-1 complex	5.29E-04
	DiGId	regulation of kinase activity	6.39E-22	membrane raft	5.20E-11
GSE24514	MR+PPIN	*	*	*	*
	mRMR+PPIN	SNAP receptor activity	2.32E-05	SNARE complex	6.41E-05
	MRMS+PPIN	*	*	*	*
	NGP_ND	enzyme binding	1.60E-49	transcription regulator complex	5.33E-16
	NGP_NR	enzyme binding	1.94E-36	transcription regulator complex	8.35E-19
	PeC	protein ADP-ribosylase activity	6.62E-05	*	*
	CPR	transferase activity, transferring phosphorus-containing groups	1.58E-31	catalytic complex	1.96E-28
	CLAIM	structural constituent of nuclear pore	1.42E-03	chromosome, centromeric region	8.10E-06
	RelSim	G protein-coupled receptor binding	2.00E-23	cyclin-dependent protein kinase holoenzyme complex	3.77E-07
	SiFS	positive regulation of molecular function	2.18E-29	transcription regulator complex	2.14E-07
	EMOGI	*	*	*	*
	FUNMarker	catalytic activity, acting on DNA	2.06E-12	chromosomal region	5.66E-16
	uKIN	catalytic activity, acting on DNA	2.71E-12	chromosomal region	6.29E-19
	DiGId	transcription regulator activity	1.86E-40	chromatin	3.46E-29
West et al	MR+PPIN	*	*	clathrin-coated endocytic vesicle membrane	5.60E-04
	mRMR+PPIN	*	*	*	*
	MRMS+PPIN	fatty acid binding	4.93E-05	endocytic vesicle lumen	1.80E-03
	NGP_ND	phospholipase activity	1.18E-05	focal adhesion	1.20E-09
	NGP_NR	cytokine receptor binding	3.17E-02	focal adhesion	7.40E-09
	PeC	fatty acid binding	5.33E-02	*	*
	CPR	DNA replication origin binding	2.04E-05	CMG complex	9.24E-07
	CLAIM	regulation of ubiquitin protein ligase activity	9.78E-04	ficolin-1-rich granule lumen	4.11E-04
	RelSim	cytokine receptor binding	7.73E-08	receptor complex	1.93E-10
	SiFS	cytokine receptor binding	1.70E-06	chromatin	2.63E-23
	EMOGI	structural constituent of ribosome	2.65E-35	cytosolic ribosome	2.84E-43
	FUNMarker	transcription factor binding	6.40E-08	RNA polymerase II transcription regulator complex	3.17E-05
	uKIN	transcription coregulator binding	8.59E-06	basement membrane	5.30E-04
	DiGId	cytokine receptor binding	3.09E-33	cell surface	1.55E-25
GSE42568	MR+PPIN	chromatin binding	2.38E-10	chromosome, centromeric region	7.02E-29
	mRMR+PPIN	chromatin DNA binding	5.15E-10	chromosomal region	1.98E-20
	MRMS+PPIN	chromatin binding	5.31E-09	chromosomal region	1.60E-21
	NGP_ND	calcium-dependent phospholipase A2 activity	4.77E-05	chromosomal region	5.14E-07
	NGP_NR	signaling receptor activator activity	1.67E-08	transferase complex, transferring phosphorus-containing groups	4.13E-06
	PeC	*	*	microtubule end	5.81E-03
	CPR	G protein-coupled peptide receptor activity	3.45E-08	integral component of presynaptic membrane	1.73E-02
	CLAIM	regulation of ubiquitin protein ligase activity	9.78E-04	ficolin-1-rich granule lumen	4.11E-04
	RelSim	*	*	*	*
	SiFS	*	*	*	*
	EMOGI	positive regulation of catalytic activity	3.60E-26	transcription regulator complex	3.76E-16
	FUNMarker	natriuretic peptide receptor activity	7.75E-04	cell-cell contact zone	3.03E-05
	uKIN	positive regulation of catalytic activity	6.71E-16	membrane raft	8.30E-08
	DiGId	positive regulation of catalytic activity	3.87E-39	anchoring junction	3.02E-13
GSE44861	MR+PPIN	RNA helicase activity	1.92E-08	*	*
	mRMR+PPIN	*	*	*	*
	MRMS+PPIN	*	*	*	*
	NGP_ND	RNA helicase activity	1.92E-08	anchoring junction	4.65E-16
	NGP_NR	*	*	*	*
	PeC	protein ADP-ribosylase activity	6.47E-05	*	*
	CPR	transcription factor binding	5.04E-25	catalytic complex	6.36E-25
	CLAIM	CXCR chemokine receptor binding	2.60E-02	complex of collagen trimers	1.18E-02
	RelSim	catalytic activity, acting on DNA	2.24E-15	chromosomal region	5.31E-47
	SiFS	transcription factor binding	2.25E-13	membrane microdomain	7.16E-14
	EMOGI	transcription factor binding	1.95E-19	transcription regulator complex	4.06E-16
	FUNMarker	hormone activity	2.67E-06	apical part of cell	3.90E-08
	uKIN	regulation of protein serine/threonine kinase activity	2.28E-06	condensed chromosome	8.95E-05
	DiGId	transcription factor binding	1.96E-33	chromatin	2.29E-23

role in activating different signal transduction pathways, which in turn promote growth and progression of colorectal cancer [17].

2) **Transcription factor binding**: The genes selected using the GSE44861 data set play active role in the dysregulation of transcription factors, which are known to be actively involved in the growth and progression of colorectal cancer cells. Binding of the transcription factors to their receptors is important in order to suppress the proliferative and pro-apoptotic growth in the tumors. Thus, dysregulation of the binding mechanism has been shown to adversely affect normal cell functioning [18].
3) **Transcription regulator activity**: The term indicates that the curated genes play an active role in regulating the rate of gene transcription. Some transcription factors, like CNOT3 are generally expressed in colorectal cancer cells and are known to regulate transcription and mRNA stability [19].
4) **Cytokine receptor binding**: The role of cytokines in growth and progression of breast cancer cells has been established in [15]. The study has shown that cytokines on breast cancer cell lines are actively involved in decreased cell-cell association with increased cellular motility, inhibition of cellular proliferation and several morphological changes like cell elongation and decrease in intercellular adhesion.
5) **Positive regulation of catalytic activity**: The function is inolved in regulating the activity levels of enzymes.

S2.7.3 Annotated CC Terms

The biological significance analysis of the CC terms annotated by the proposed approach is listed as follows:

1) **Membrane raft**: From [20], it is known that the different kinase activities are regulated and controlled in various compartments of the plasma membrane.
2) **Chromatin**: From [19], it is known that gene transcription activities are performed using the chromatin fibres, which are present in the nucleus of a cell.
3) **Cell surface**: It is known to be the activation site of cytokine receptors [21].
4) **Anchoring junction**: It forms the tight connection between cells and extracellular matrix. The junctions aid in signal transduction, cell proliferation and metastasis [22].

S2.7.4 Annotated KEGG Pathways

The analysis of the KEGG pathways annotated by the proposed algorithm is described as follows:

1) **MAPK signaling pathway**: It is known to be involved in cell proliferation, regulation of intestinal tissue differentiation and cell apoptosis. It is also known to be involved in control of growth signals, cell survival, and invasion in cancer [23].
2) **Transcriptional misregulation in cancer**: This indicates that the proposed approach curates transcription factor encoding genes, which are misregulated by chromosomal translocation and inversion that further drive the transcription of oncogenes [24].
3) **JAK-STAT signaling pathway**: It is known to be intensely involved in tumorigenesis, maintenance and metastasis of breast cancer. It is known to be associated with cell growth, survival, invasion and motility through regulation of genes like BCL2, P16, VEGF, MMPs, etc [25].
4) **NF-kappa B signaling pathway**: The NF-kappa B activity is an important marker for breast tumors. It is activated by the chemotherapy and radiotherapy approaches and its signalling pathway regulates more than 500 genes that control inflammation, cellular transformation, survival, proliferation, angiogenesis, invasion, and metastasis to pulmonary and brain sites. It is known to regulate the expression and function of growth stimulating cytokines and also upregulate the expression of Cyclin D1, CDK2, c-Myc that drives cell cycle progression and cause uncontrolled cell proliferation [26].

REFERENCES

[1] T. Hinoue et al., "Genome-Scale Analysis of Aberrant DNA Methylation in Colorectal Cancer," *Genome Research*, vol. 22, no. 2, pp. 271–282, 2012.
[2] P. Alhopuro et al., "Candidate Driver Genes in Microsatellite-Unstable Colorectal Cancer," *International Journal of Cancer*, vol. 130, no. 7, pp. 1558–1566, 2012.
[3] West et al., "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles," *Proceedings of the National Academy of Sciences, USA*, vol. 98, no. 20, pp. 11 462–11 467, 2001.
[4] Clarke et al., "Correlating Transcriptional Networks to Breast Cancer Survival: A Large-Scale Coexpression Analysis," *Carcinogenesis*, vol. 34, no. 10, pp. 2300–2308, 2013.
[5] B. M. Ryan et al., "Germline Variation in NCF4, an Innate Immunity Gene, is Associated With an Increased Risk of Colorectal Cancer," *International Journal of Cancer*, vol. 134, no. 6, pp. 1399–1407, 2014.
[6] Bellver et al., "Transcriptome Profile of Human Colorectal Adenomas," *Mol Cancer Res*, vol. 5, no. 12, pp. 1263–1275, 2007.
[7] Nagaraj et al., "A Boolean-Based Systems Biology Approach to Predict Novel Genes Associated with Cancer: Application to Colorectal Cancer," *BMC Systems Biology*, vol. 5, no. 1, p. 35, 2011.
[8] Nacht et al., "Combining Serial Analysis of Gene Expression and Array Technologies to Identify Genes Differentially Expressed in Breast Cancer," *Cancer Research*, vol. 59, no. 21, pp. 5464–5470, 1999.
[9] V. H. Koelzer, H. Dawson, I. Zlobec, and A. Lugli, "Colon: Colorectal adenocarcinoma," *Atlas of Genetics and Cytogenetics in Oncology and Haematology*, vol. 17, no. 5, pp. 348–363, 2013.
[10] M. L. Carcangiu, P. Casalini, and S. MMénardnard, "Breast Tumors: An Overview," *Atlas of Genetics and Cytogenetics in Oncology and Haematology*, vol. 9, no. 4, pp. 335–341, 2005.
[11] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, "The COSMIC Cancer Gene Census: Describing Genetic Dysfunction Across All Human Cancers," *Nature Reviews Cancer*, vol. 18, p. 696705, 2018.
[12] G. Yu, L.-G. Wang, G.-R. Yan, and Q.-Y. He, "DOSE: An R/Bioconductor Package for Disease Ontology Semantic and Enrichment Analysis," *Bioinformatics*, vol. 31, no. 4, pp. 608–609, 2015.
[13] Z. Wang et al., "Oxidative Stress and Carbonyl Lesions in Ulcerative Colitis and Associated Colorectal Cancer," *Oxidative Medicine and Cellular Longevity*, vol. 2016, 2016.
[14] S. Bayram, S. Fürst, M. Forbes, and S. Kempa, "Analysing Central Metabolism in Ultra-high Resolution: At The Crossroads of Carbon and Nitrogen," *Molecular Metabolism*, vol. 33, pp. 38–47, 2020.
[15] A. M. Douglas, G. A. Goss, R. L. Sutherland, D. J. Hilton, M. C. Berndt, N. A. Nicola, and C. G. Begley, "Expression and Function of Members of the Cytokine Receptor Superfamily on Breast Cancer Cells," *Oncogene*, vol. 14, p. 661669, 1997.

[16] X. Chen, J. Gu, A. F. Neuwald, L. Hilakivi-Clarke, R. Clarke, and J. Xuan, "Identifying Intracellular Signaling Modules and Exploring Pathways Associated with Breast Cancer Recurrence," *Scientific Reports*, vol. 11, no. 385, 2021.
[17] M. García-Aranda and M. Redondo, "Targeting Receptor Kinases in Colorectal Cancer," *Cancers*, vol. 11, no. 4, 2019.
[18] P. Laissue, "The Forkhead-Box Family of Transcription Factors: Key Molecular Players in Colorectal Cancer Pathogenesis," *Molecular Cancer*, vol. 15, 2019.
[19] P. Cejas, A. Cavazza, C. Yandava, V. Moreno, D. Horst, J. Moreno-Rubio, E. Burgos, M. Mendiola, L. Taing, A. Goel, J. Feliu, and R. A. Shivdasani, "Transcriptional Regulator CNOT3 Defines an Aggressive Colorectal Cancer Subtype," *Cancer Research*, vol. 77, no. 3, pp. 766–779, 2017.
[20] J. Seong, M. Ouyang, T. Kim, J. Sun, P.-C. Wen, S. Lu, Y. Zhuo, N. M. Llewellyn, D. D. Schlaepfer, J.-L. Guan, S. Chien, and Y. Wang, "Detection of Focal Adhesion Kinase Activation at Membrane Microdomains by Fluorescence Resonance Energy Transfer," *Nature Communications*, vol. 406, no. 2, 2011.
[21] S. E. Broughton, T. R. Hercus, A. F. Lopez, and M. W. Parker, "Cytokine Receptor Activation at the Cell Surface," *Current Opinion in Structural Biology*, vol. 22, no. 3, pp. 350–359, 2012.
[22] D. Kyuno, A. Takasawa, S. Kikuchi, I. Takemasa, M. Osanai, and T. Kojima, "Role of Tight Junctions in the Epithelial-to-Mesenchymal Transition of Cancer Clls," *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1863, no. 3, p. 183503, 2021.
[23] J. Y. Fang and B. C. Richardson, "The MAPK Signalling Pathways and Colorectal Cancer," *The Lancet Oncology*, vol. 6, p. 322327, 2005.
[24] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.
[25] F. Shao, X. Pang, and G. H. Baeg, "Targeting the JAK/STAT Signaling Pathway for Breast Cancer," *Current Medicinal Chemistry*, vol. 28, no. 25, pp. 5137 – 5151, 2021.
[26] W. Wang, S. A. Nag, and R. Zhang, "Targeting the NFκB Signaling Pathways for Breast Cancer Prevention and Therapy," *Current Medicinal Chemistry*, vol. 22, no. 2, pp. 264–289, 2015.