De novo and supervised endophenotyping using network-guided ensemble learning
Supplementary materials

Simon J. Larsen[1,*], Harald H.H.W. Schmidt[2], and Jan Baumbach[1,3]

[1]Department of Mathematics and Computer Science, University of Southern Denmark
[2]Department of Pharmacology and Personalised Medicine, Faculty of Health, Medicine and Life Science, Maastricht University
[3]Chair of Experimental Bioinformatics, Wissenschaftszentrum Weihenstephan, Technical University of Munich
[*]Corresponding author, e-mail: sjlarsen@imada.sdu.dk

S1 Definitions

S1.1 Jaccard index

The Jaccard index is a statistic used to measure the similarty of two sets. The statistic is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

where $0 \leq J(A,B) \leq 1$. A greater value correspond to a greater agreement between the two sets, where 0 means the sets are completely disjoint and a 1 means the sets are identical.

S1.2 Hypergeometric test for overrepresentation

The hypergeometric test is used to measure the statistical significance of drawing a certain number of special objects from a population constiting of special and non-special object. The statistic is used to measure the statistical significance of enrichment of a gene set among the genes in a gene module. A p-value is computed as

$$P(X \geq k) = \sum_{i=k}^{\min(K,n)} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where n is the size of the gene module, k is the number of genes in the module also belonging to the gene set, K is the number of genes in the gene set and N is the total number of genes in the data set (population the gene module is drawn from).

S1.3 Silhouette index

The silhouette index is an internal cluster validity index. The silhouette value of a data point i is the measures how similar it to other data points in its own cluster, compared to the data points in other clusters. Let $a(i)$ be the average distance between data point i and all other data points in the same cluster. Let $b(i)$ be the minimum average distance from i to the points in a different cluster, minimized over all clusters. Then the silhouette value of data point i is defined

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)}.$$

We then compute a silhouette index for the clustering as the average silhouette value over all data points. The silhouette index ranges from -1 to 1, where a higher value is interpreted as a better clustering.

S1.4 Feature subgraph sampling procedure

Feature subgraphs are extracted using a breadth-first search procedure. The procedure is outlined in Algorithm 1.

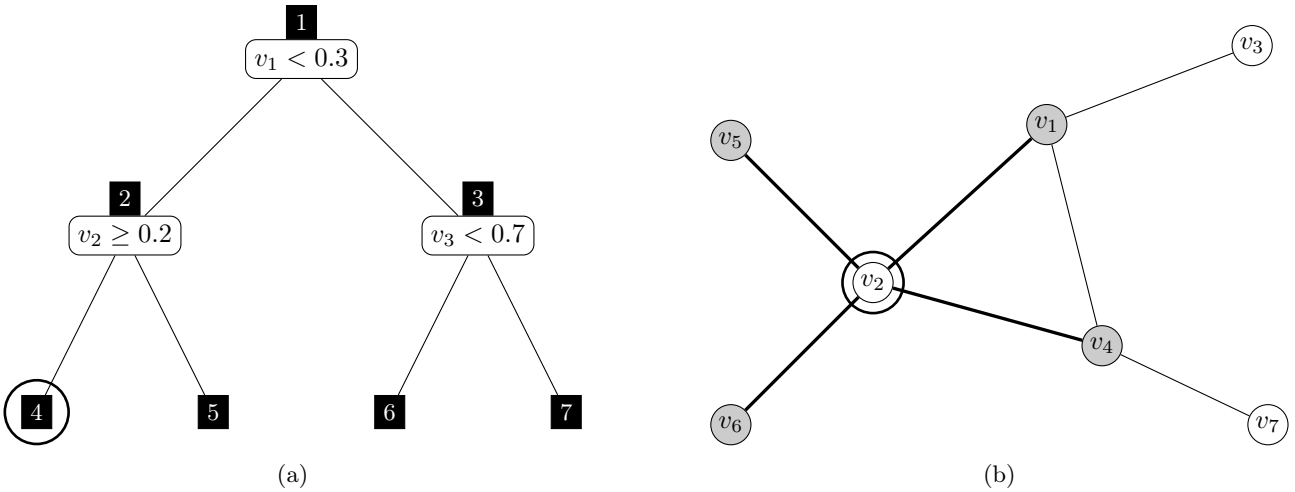
Algorithm 1: BFS feature bagging subgraph procedure

```
Data: G: Feature graph.
Data: m: Size of subgraph to extract.
open ← empty queue
selected ← ∅
v_s ← random vertex ∈ V(G)
enqueue(open,v_s)
marked ← {v_s}
while open ≠ ∅ ∧ |selected| < m do
    u ← dequeue(open)
    neighbors ← shuffle(N(u))
    for v ∈ neighbors do
        if v ∉ marked then
            enqueue(open,v)
            marked ← marked ∪ {v}
        end
    end
end
return induced-subgraph(G,selected)
```

S2 Differences between Grand Forest and NGF

Grand Forest significantly extends on the well-known Random Forest algorithm and Network-Guided Forests (NGF) introduced by Dutkowski *et al.* It differs in some key aspects: In NGF, gene importance is estimated using permutation importance on out-of-bag samples. In Grand Forest we instead use the mean decrease in Gini impurity, because we are not concerned with how important a gene is for predictive performance but, rather, how much information it provides at the time of split, conditioned on the splits preceding it. To confirm that Gini impurity was better suited for module discovery we computed modules for the five data set using permutation importance instead and observed that Gini impurity produced significantly better results in all five data sets (Supplementary Figure S4). Another key difference is that Grand Forest enforces that all decision tree splits (except for the root) must be made on a variable that is adjacent to the variable in the split above it. NGF is less strict enforcing only the requirement that the resulting decision tree variables must induce a connected subnetwork in the global interaction network. Finally, Grand Forest supports not only categorical values as response (clinical) variable, but also numerical values and right-censored survival data, and the methodology is extended to support unsupervised analysis as well.

S3 Supplementary figures



(a) (b)

Figure S1: Example of possible split variables available during split selection. (a) A decision tree consisting of three splits. The next split to be performed is marked with a circle (split number 4). (b) The corresponding feature bagging subgraph for the decision tree. The set of possible split variables is the neighborhood of v_2 (gray nodes) because v_2 is the split variable in the parent node of split 4.

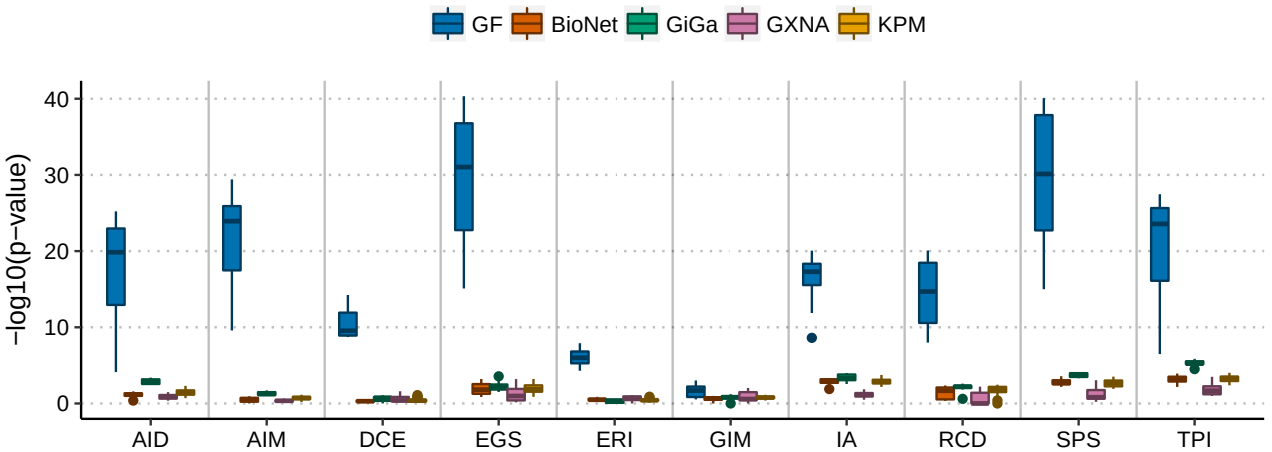


Figure S2: Enrichment of pathways related to the Hallmarks of Cancer for modules extracted from breast cancer. Hallmarks: Avoiding Immune Destruction (AID), Activating Invasion and Metastasis (AIM), Deregulating Cellular Energetics (DCE), Evading Growth Suppressors (EGS), Enabling Replicative Immortality (ERI), Genome Instability and Mutation (GIM), Inducing Angiogenesis (IA), Resisting Cell Death (RCD), Sustaining Proliferative Signaling (SPS), Tumor-promoting Inflammation (TPI). Enrichment was computed using a hypergeometric overrepresentation test.

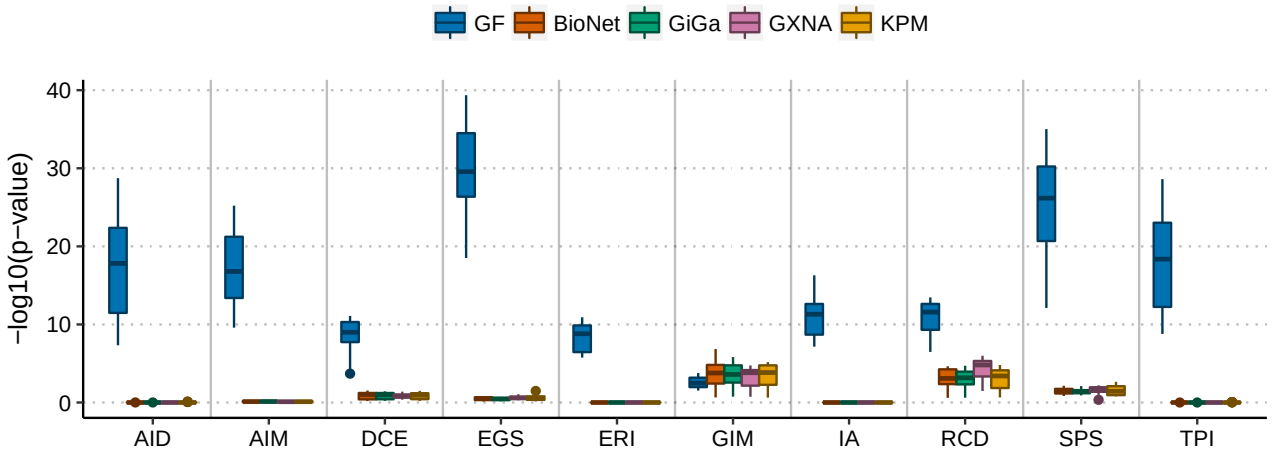


Figure S3: Enrichment of pathways related to the Hallmarks of Cancer for modules extracted from non-small cell lung cancer. Hallmarks: Avoiding Immune Destruction (AID), Activating Invasion and Metastasis (AIM), Deregulating Cellular Energetics (DCE), Evading Growth Suppressors (EGS), Enabling Replicative Immortality (ERI), Genome Instability and Mutation (GIM), Inducing Angiogenesis (IA), Resisting Cell Death (RCD), Sustaining Proliferative Signaling (SPS), Tumor-promoting Inflammation (TPI). Enrichment was computed using a hypergeometric overrepresentation test.

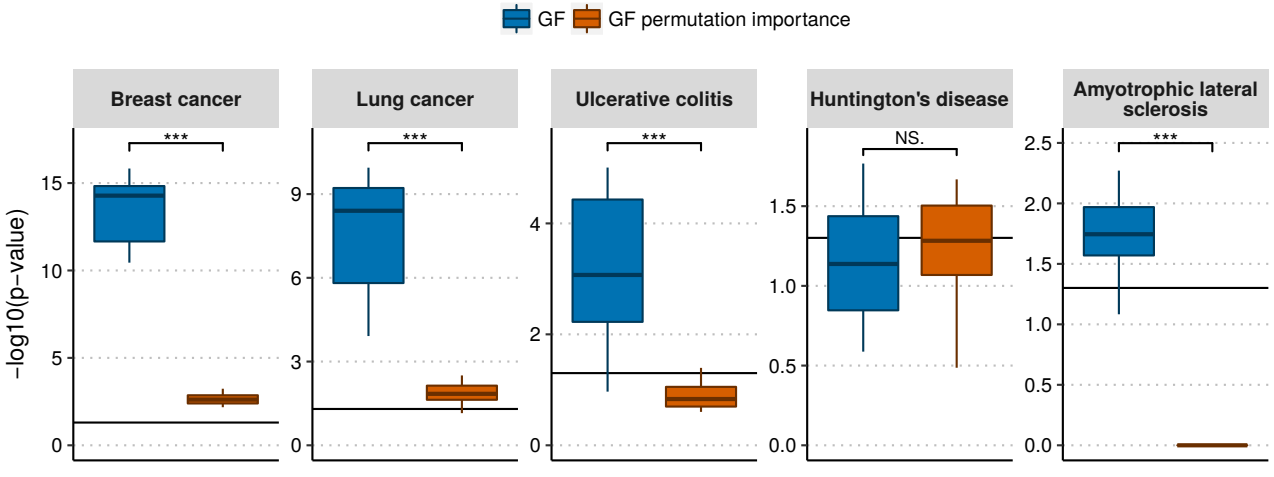


Figure S4: Enrichment of disease-related KEGG pathways for modules selected with Grand Forest using Gini impurity and permutation importance for estimating gene importance. ($^{***}p < 0.001$, NS. = not significant.)

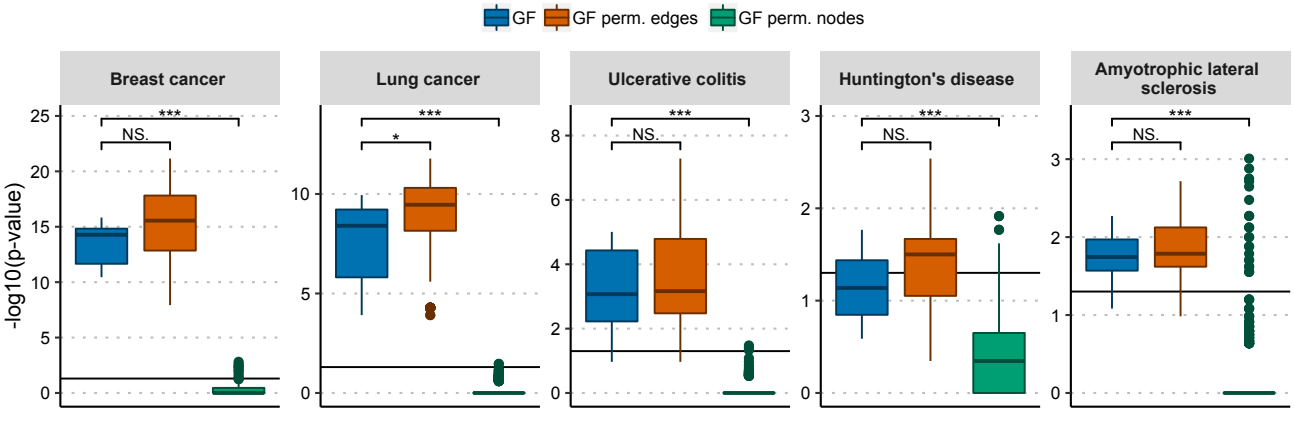


Figure S5: Enrichment of disease-related KEGG pathways for gene modules extracted with Grand Forest using unperturbed IID network, IID network by random edge rewiring and IID network perturbed by randomly rearranging node labels. The perturbed experiment was repeated 30 times. ($^{***}p < 0.001$, NS. = not significant.)

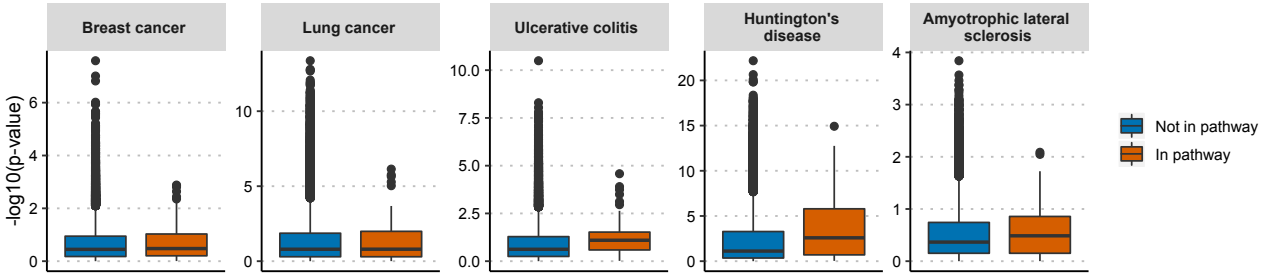


Figure S6: Comparison of p-values for genes in disease-associated pathways and other genes for each gene expression data set.

S4 Supplementary tables

Type	Count
Adenocarcinoma (ADC)	85
Basaloid (BAS)	39
Carcinoid (CARCI)	24
Large cell carcinoma (LCC)	3
Large cell neuroendocrine (LCNE)	56
Squamous cell carcinoma (SQC)	61
Total	268

Table S1: Distribution of histological types in non-small cell lung cancer data set (GSE30219).

Disease	Pathways
Breast cancer	hsa05224
Non-small cell lung cancer	hsa05223
Ulcerative colitis	hsa05321, hsa04060
Huntington's disease	hsa05016
Amyotrophic lateral sclerosis	hsa05014

Table S2: KEGG pathways used as disease-associated gene sets in enrichment analysis.

ID	Name
1	Sustaining Proliferative Signaling
2	Enabling Replicative Immortality
3	Inducing Angiogenesis
4	Evading Growth Suppressors
5	Resisting Cell Death
6	Deregulating Cellular Energetics
7	Genome Instability and Mutation
8	Avoiding Immune Destruction
9	Activating Invasion and Metastasis
10	Tumor-promoting Inflammation

Table S3: List of cancer hallmarks and their IDs in Table S4.

KEGG ID	Pathway name	Hallmark IDs
hsa03410	Base excision repair	7
hsa03420	Nucleotide excision repair	7
hsa03430	Mismatch repair	7
hsa03440	Homologous recombination	7
hsa03450	Non-homologous end-joining	7
hsa04010	MAPK signaling pathway	1
hsa04012	ErbB signaling pathway	1
hsa04070	Phosphatidylinositol signaling system	1
hsa04150	mTOR signaling pathway	1
hsa04310	Wnt signaling pathway	2
hsa04330	Notch signaling pathway	3
hsa04350	TGF-beta signaling pathway	4,8,10
hsa04370	VEGF signaling pathway	3
hsa04110	Cell cycle	1
hsa04115	p53 signaling pathway	4,5,6,7
hsa04210	Apoptosis	5
hsa04510	Focal adhesion	4,9
hsa04520	Adherens junction	9
hsa04640	Hematopoietic cell lineage	8,10
hsa04610	Complement and coagulation cascades	8,10
hsa04620	Toll-like receptor signaling pathway	8,10
hsa04621	NOD-like receptor signaling pathway	8,10
hsa04622	RIG-I-like receptor signaling pathway	8,10
hsa04623	Cytosolic DNA-sensing pathway	8,10
hsa04650	Natural killer cell mediated cytotoxicity	8,10
hsa04612	Antigen processing and presentation	8,10
hsa04660	T cell receptor signaling pathway	8,10
hsa04662	B cell receptor signaling pathway	8,10
hsa04664	Fc epsilon RI signaling pathway	8,10
hsa04666	Fc gamma R-mediated phagocytosis	8,10
hsa04670	Leukocyte transendothelial migration	8,10
hsa04672	Intestinal immune network for IgA production	8,10
hsa04062	Chemokine signaling pathway	8,10
hsa00030	Pentose phosphate pathway	6
hsa04512	ECM-receptor interaction	9
hsa04060	Cytokine-Cytokine receptor interaction	1,10
hsa04024	cAMP signaling pathway	1
hsa04151	PI3K-Akt signaling pathway	1,4
hsa04630	Jak-STAT signaling pathway	9,10
hsa03320	PPAR signaling pathway	1
hsa04611	Patelet Activation	8
hsa00010	Glycolysis / Gluconeogenesis	6
hsa00190	Oxidative phosphorylation	6
hsa00020	Citriate cycle (TCA cycle)	6
hsa00260	Glycine serine and threonine metabolism	6
hsa00471	D-Glutamine and D-Glutamate metabolsim	6
hsa00330	Arginine and proline metabolsim	6
hsa04066	HIF-1 signaling pathway	6
hsa00250	Alanine, aspartate and glutamate matabolism	6
hsa00564	Glycerophopholipid metabolism	4
hsa04810	Regulation of actin cytoskeleton	1,3,10
hsa05230	Central Carbon metabolism in cancer	6
hsa05231	Choline metabolism in cancer	1,3,5,10
hsa04064	NF-kappa B signaling pathway	8,10

Table S4: List of pathways related to the hallmarks of cancer.

Grand Forest		
GSE11223	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE112680	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE20685	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE30219	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE3790	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
BioNet		
GSE11223	fdr	0.0002, 0.0002536, 0.0003071, 0.0003607, 0.00039, 0.0004143, 0.0004679, 0.0005214, 0.000575, 0.0006821, 0.0006286, 0.0007357, 0.0007893, 0.0008429, 0.0008964, 0.00095
GSE112680	fdr	0.29, 0.2943, 0.2986, 0.3029, 0.3071, 0.3114, 0.3157, 0.32, 0.3243, 0.3286, 0.3329, 0.3371, 0.338, 0.34, 0.3414, 0.344, 0.3457, 0.35
GSE20685	fdr	0.0091, 0.0103, 0.01286, 0.01414, 0.0158, 0.018, 0.0206, 0.0226, 0.0235, 0.0255
GSE30219	fdr	5.72e-09, 1e-08, 1.78e-08, 2e-08, 3.286e-08, 4.2e-08, 6.07e-08, 7.46e-08, 1.025e-07, 1.58e-07
GSE3790	fdr	2e-15, 4e-15, 8e-15, 1e-14, 2e-14, 4.471e-14, 8.743e-14, 1.301e-13, 1.729e-13, 2.156e-13, 2.583e-13, 3.01e-13, 3.437e-13, 3.864e-13, 4.291e-13, 4.719e-13, 5.146e-13, 5.573e-13
GiGa		
GSE11223	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE112680	size	27, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE20685	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE30219	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE3790	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GXNA		
GSE11223	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE112680	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE20685	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE30219	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
GSE3790	size	25, 33, 42, 50, 58, 67, 75, 83, 92, 100
KPM		
GSE11223	k	1, 2, 3
	p-value	2e-06, 2e-06, 3.107e-06, 2e-06, 3.107e-06, 4.214e-06, 3.107e-06, 5.321e-06, 4.214e-06, 6.429e-06,
	cutoff	4.214e-06, 5.321e-06, 7.536e-06, 6.429e-06, 5.321e-06, 6.429e-06, 7.536e-06, 8.643e-06, 9.75e-06, 7.536e-06, 8.643e-06, 1.086e-05, 9.75e-06, 8.643e-06, 1.196e-05, 9.75e-06, 1.086e-05, 1.307e-05, 1.086e-05, 1.196e-05, 1.418e-05, 1.196e-05, 1.307e-05, 1.529e-05, 1.307e-05, 1.418e-05, 1.639e-05, 1.529e-05, 1.75e-05, 1.418e-05, 1.529e-05, 1.639e-05
GSE112680	k	1, 2, 3
	p-value	0.001857, 0.001429, 0.002286, 0.001857, 0.002714, 0.003143, 0.002286, 0.002714, 0.003143, 0.002286,
	cutoff	0.002714, 0.003571, 0.003143, 0.003571, 0.004, 0.003571, 0.004, 0.004429, 0.004, 0.004429, 0.004857, 0.004429, 0.004857, 0.005286, 0.004857, 0.005286, 0.005714, 0.005286, 0.005714, 0.006143, 0.005714, 0.006143
GSE20685	k	1, 2, 3
	p-value	6e-05, 6e-05, 6e-05, 9.5e-05, 0.00013, 9.5e-05, 0.000165, 9.5e-05, 0.00013, 0.0002, 0.000235, 0.00013,
	cutoff	0.000165, 0.00027, 0.000165, 0.0002, 0.000305, 0.0002, 0.000235, 0.00034, 0.000375, 0.00027, 0.000235, 0.00041, 0.000305, 0.000445, 0.00048, 0.000515, 0.00027, 0.00034, 0.000305, 0.000375, 0.00034, 0.00041, 0.00055, 0.000445
GSE30219	k	1, 2, 3
	p-value	1e-11, 1e-11, 1e-11, 3.4e-11, 3.4e-11, 3.4e-11, 6.3e-11, 6.3e-11, 6.3e-11, 1e-10, 1e-10, 1e-10, 1.5e-10,
	cutoff	1.5e-10, 1.5e-10, 2.94e-10, 2.94e-10, 2.94e-10, 4.36e-10, 4.36e-10, 4.36e-10, 5.78e-10, 5.78e-10, 5.78e-10, 8.6e-10, 1e-09, 8.6e-10, 8.6e-10, 1e-09, 1e-09
GSE3790	k	1, 2, 3
	p-value	2e-17, 2e-17, 2e-17, 1e-16, 1e-16, 1e-16, 5e-16, 5e-16, 5e-16, 1.233e-15, 1.233e-15, 1.233e-15,
	cutoff	2.446e-15, 2.446e-15, 2.446e-15, 3.659e-15, 3.659e-15, 4.871e-15, 3.659e-15, 4.871e-15, 6.084e-15, 4.871e-15, 6.084e-15, 6.084e-15, 7.297e-15, 7.297e-15, 8.51e-15, 9.723e-15, 7.297e-15, 8.51e-15, 1.094e-14, 9.723e-15, 8.51e-15, 1.215e-14, 1.336e-14, 1.094e-14, 9.723e-15, 1.457e-14, 1.215e-14, 1.336e-14, 1.579e-14, 1.094e-14, 1.457e-14, 1.215e-14, 1.336e-14, 1.579e-14, 1.7e-14, 1.457e-14, 1.579e-14, 1.7e-14

Table S5: Program parameters for each module discovery method. First column is data set, second column is parameter name and third column is parameter value.

	Breast cancer	Lung cancer	Ulcerative colitis	Huntington's disease	Amyotrophic lateral sclerosis
GF	6.128e-15	4.639e-09	0.0008697	0.07429	0.01806
BioNet	0.3709	1	1	0.6837	1
GiGa	1	1	1	0.6796	1
GXNA	0.6792	1	0.2607	0.6835	1
KPM	1	1	1	0.6761	1

Table S6: Table of median p-values for enrichment of disease-associated KEGG pathways for extracted gene modules. Enrichment was computed using a hypergeometric overrepresentation test.

	Breast cancer	Lung cancer	Ulcerative colitis	Huntington's disease	Amyotrophic lateral sclerosis
GF	4.636e-12	1.81e-05	0.02034	0.09969	0.02699
BioNet	0.5099	1	1	0.7603	1
GiGa	1	1	0.929	0.7614	1
GXNA	0.6046	1	0.4984	0.7643	1
KPM	1	1	1	0.7126	0.6858

Table S7: Table of mean p-values for enrichment of disease-associated KEGG pathways for extracted gene modules. Enrichment was computed using a hypergeometric overrepresentation test.