Supplementary Information for

Patient stratification reveals the molecular basis of disease comorbidities

Beatriz Urda-García[1,2], Jon Sánchez-Valle[1,*], Rosalba Lepore[1,3] and Alfonso Valencia[1,4,*]

*Correspondence to: jon.sanchez@bsc.es and alfonso.valencia@bsc.es

This PDF file includes:

Supplementary Notes

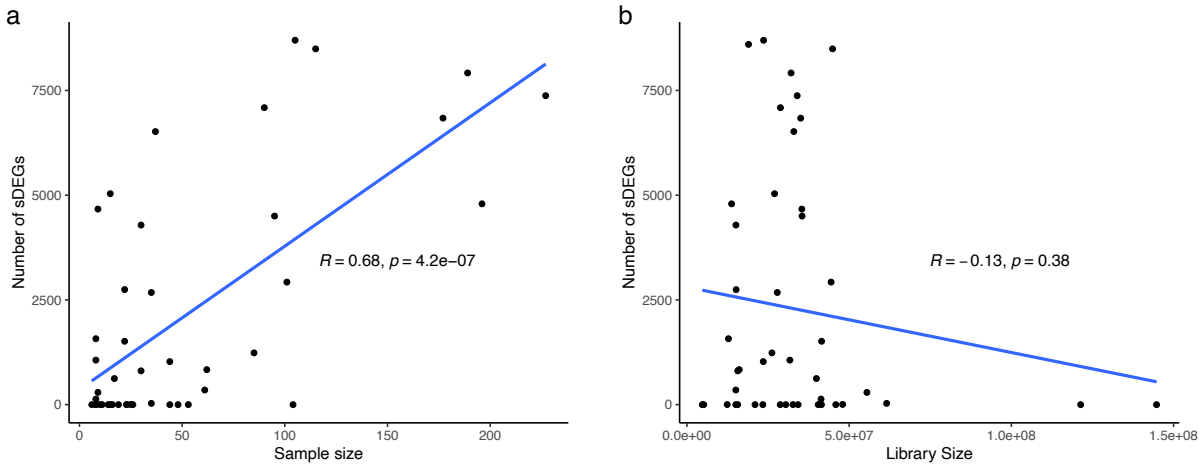
Supplementary Figures 1 to 12

Supplementary Tables 1 to 11

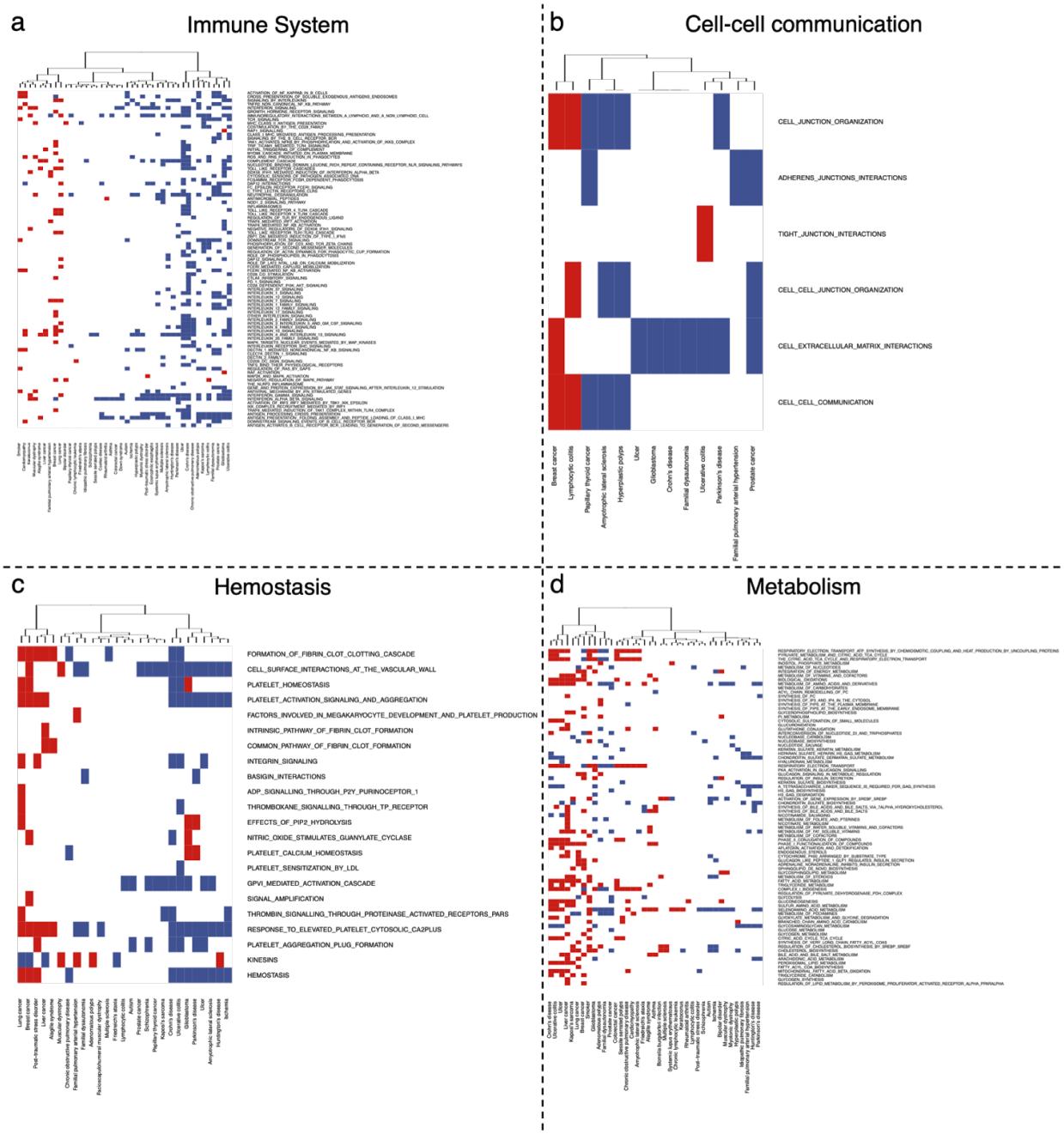
Supplementary Notes

1 - Computing the overlap of the Disease Similarity Network (DSN) with the epidemiology

We computed the overlap between the positive and negative interactions in the DSN and the ones based on medical records by Hidalgo *et al*[1]. To do so, we transformed the disease names in the DSN into ICD9 codes. Then, we computed the overlaps following the same methodology described in methods. In the cases in which several disease names referred to the same ICD9 code, only the interactions shared by all the diseases that correspond to that code were considered. The results of these overlaps can be found in the Supplementary Table 11 and are consistent with the ones obtained by defining diseases at the ICD9 level from the beginning; the positive interactions present high and significant overlap with the epidemiology and the overlap with the negative interactions is not significant.

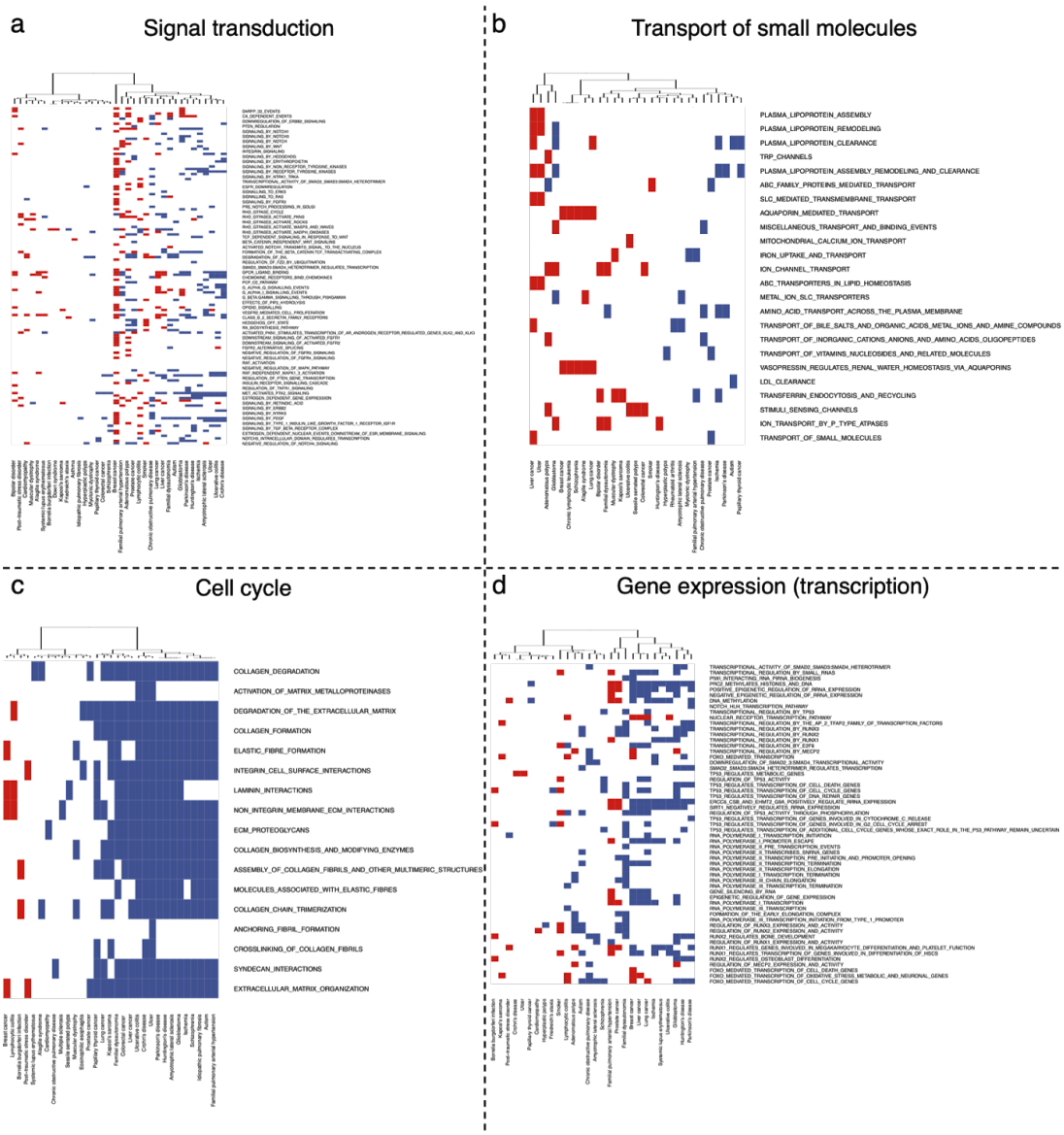


Supplementary Fig. 1. Correlation between the number of significantly differentially expressed genes (sDEGs) and the sample size and library size. (**a**) Correlation between the number of sDEGs and the sample size in our disease set. (**b**) Correlation between the number of sDEGs and the average library size in our disease set.



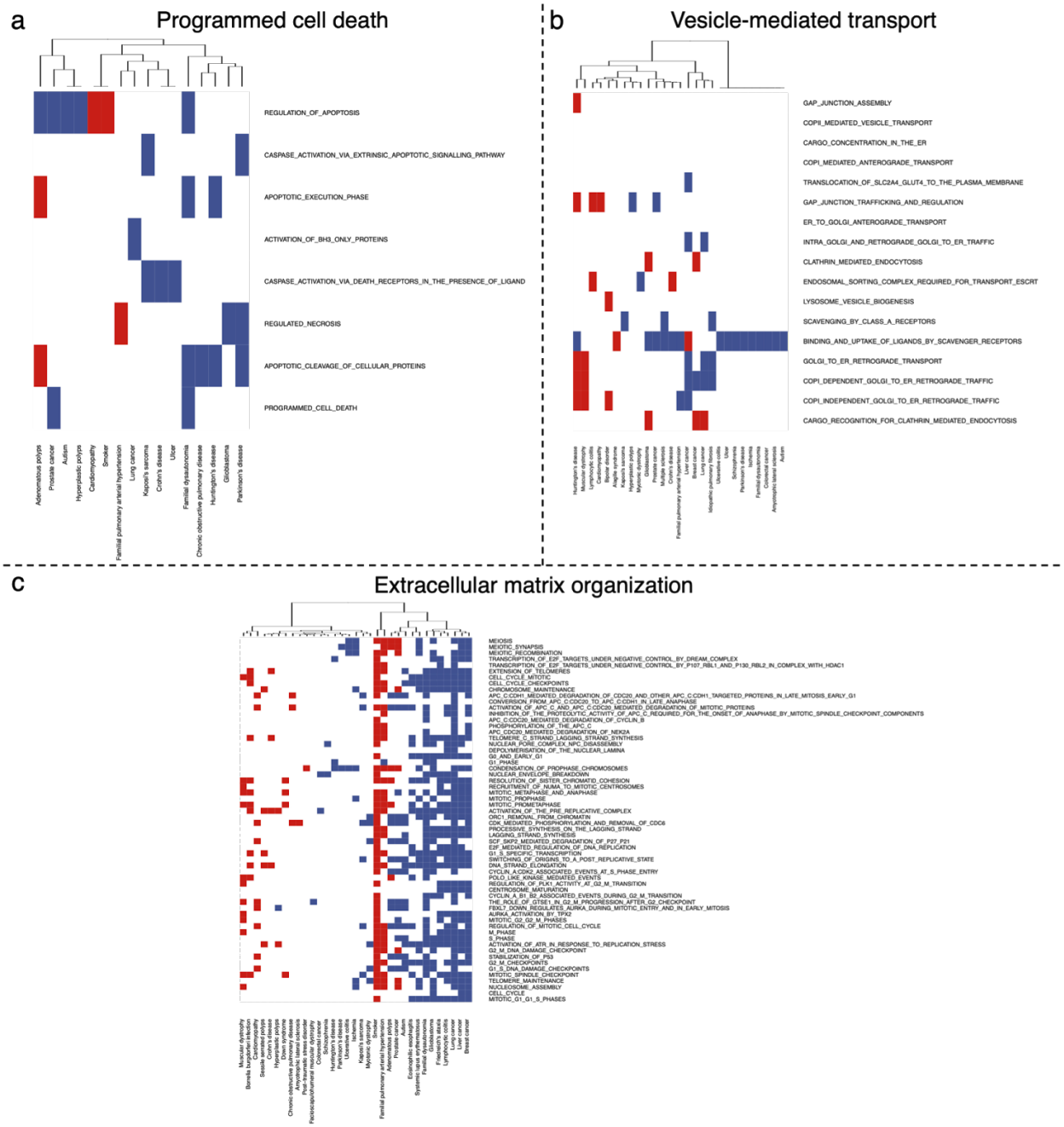
Supplementary Fig. 2. Immune system, cell-cell communication, hemostasis, and metabolism Reactome pathways significantly enriched in human diseases. For each disease, Reactome pathways significantly up- and down-regulated were identified using GSEA method[2] (FDR < 0.05). Ward2 algorithm was applied to cluster diseases based on the Euclidean distance of the binarized Normalized Effect Size. The heatmap shows the dysregulated pathways of the pathway category (rows) in the diseases (columns), where up- and down-regulated pathways are blue and red colored

respectively. Diseases with 0 dysregulated pathways are not shown. (**a**) Immune system, (**b**) Cell-cell communication, (**c**) Hemostasis, and (**d**) Metabolism Reactome pathways significantly enriched in human diseases.



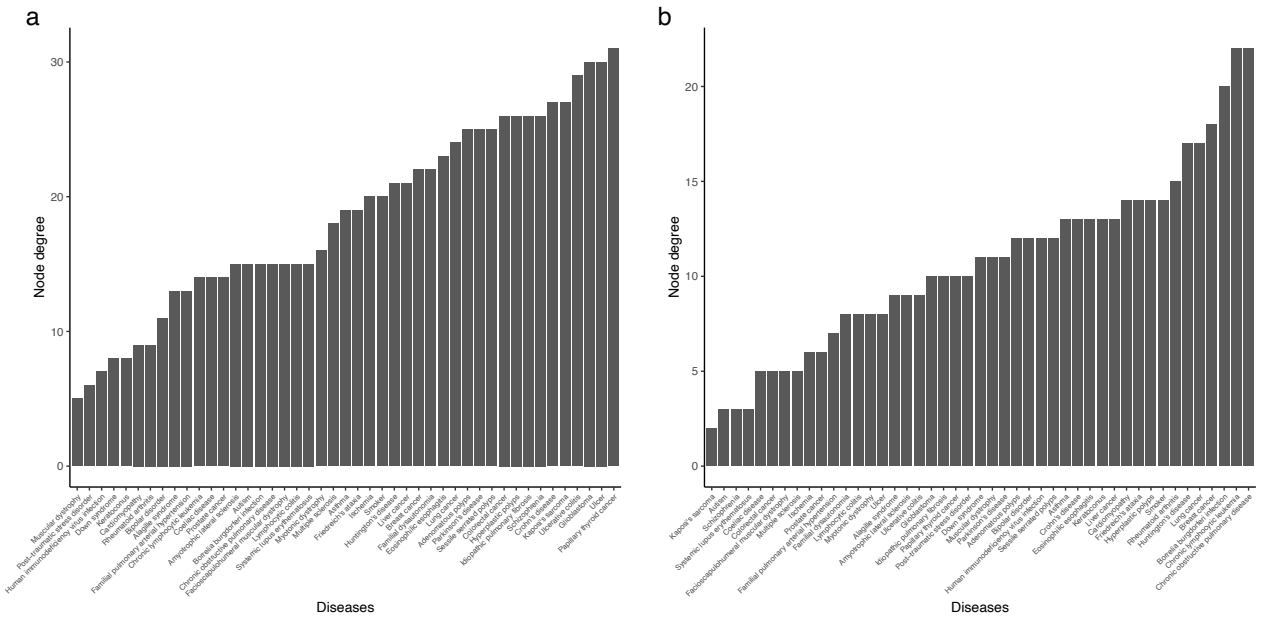
Supplementary Fig. 3. Signal transduction, transport of small molecules, cell cycle, and gene expression Reactome pathways significantly enriched in human diseases. For each disease, Reactome pathways significantly up- and down-regulated were identified using GSEA[2] method (FDR < 0.05). Ward2 algorithm was applied to cluster diseases based on the Euclidean distance of the binarized Normalized Effect Size. The heatmap shows the dysregulated pathways of the pathway category (rows) in the diseases (columns), where up- and down-regulated pathways are blue and red colored respectively. Diseases with 0 dysregulated pathways are not shown. (**a**) Signal

transduction, (**b**) Transport of small molecules, (**c**) Cell cycle, and (**d**) Gene expression Reactome pathways significantly enriched in human diseases.

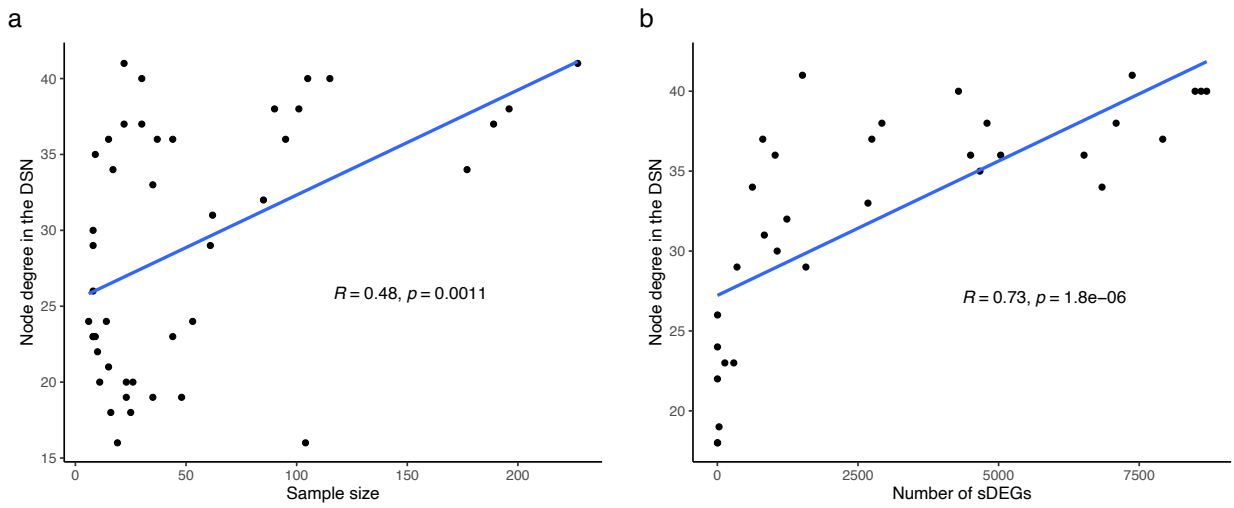


Supplementary Fig. 4. Programmed cell death, vesicle-mediated transport, and extracellular matrix organization Reactome pathways significantly enriched in human diseases. For each disease, Reactome pathways significantly up- and down-regulated were identified using GSEA[2] method (FDR < 0.05). Ward2 algorithm was applied to cluster diseases based on the Euclidean distance of the binarized Normalized Effect Size. The heatmap shows the dysregulated pathways of the pathway category (rows) in the diseases (columns), where up- and down-regulated pathways are blue and red colored respectively. Diseases with 0 dysregulated pathways are not shown. (**a**)

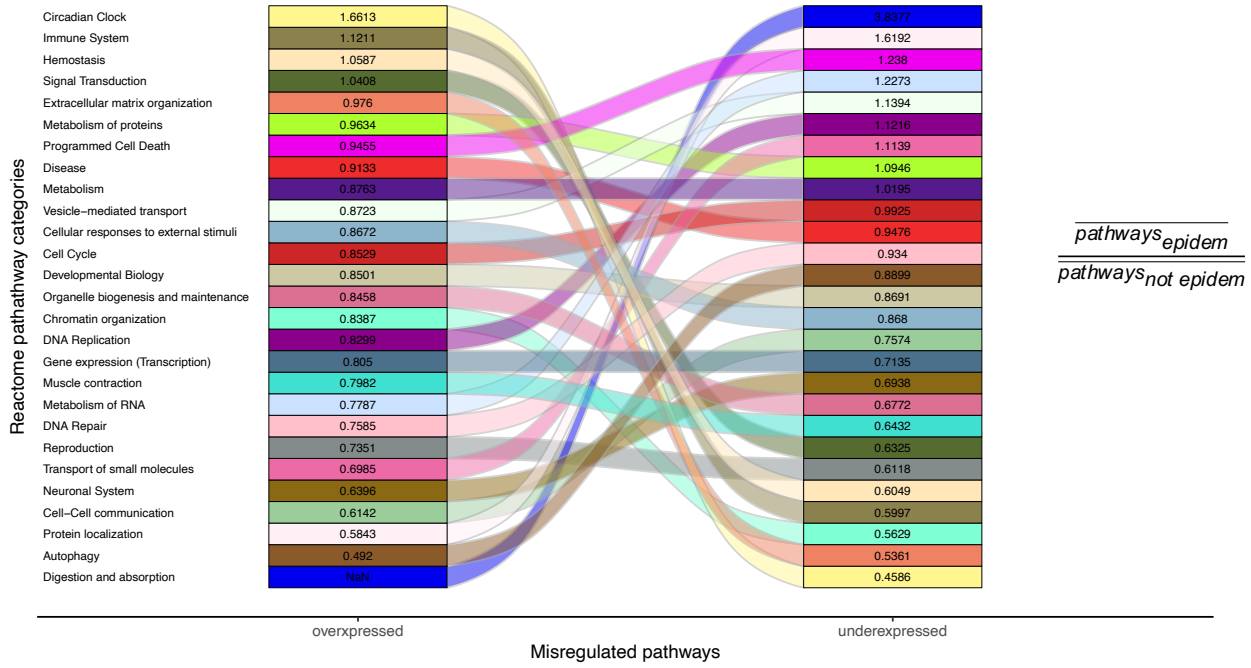
Programmed cell death, (**b**) Vesicle-mediated transport, and (**c**) Extracellular matrix organization Reactome pathways significantly enriched in human diseases.



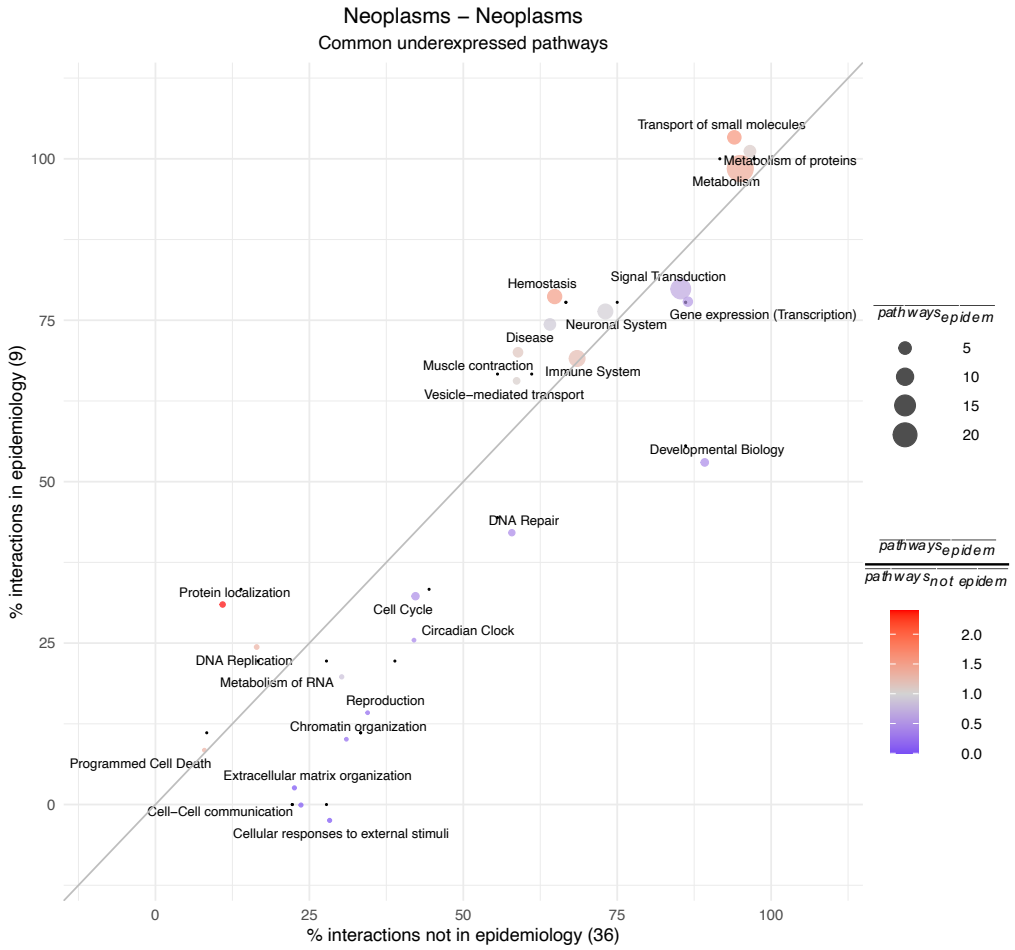
Supplementary Fig. 5. Disease similarity network's nodes' degree considering the positive and negative interactions. (**a**) Bar plot of the diseases' node degree in the disease similarity network with positive interactions. Diseases are sorted by degree. (**b**) Bar plot of the diseases' degree within the disease similarity network with negative interactions. Diseases are sorted by degree.



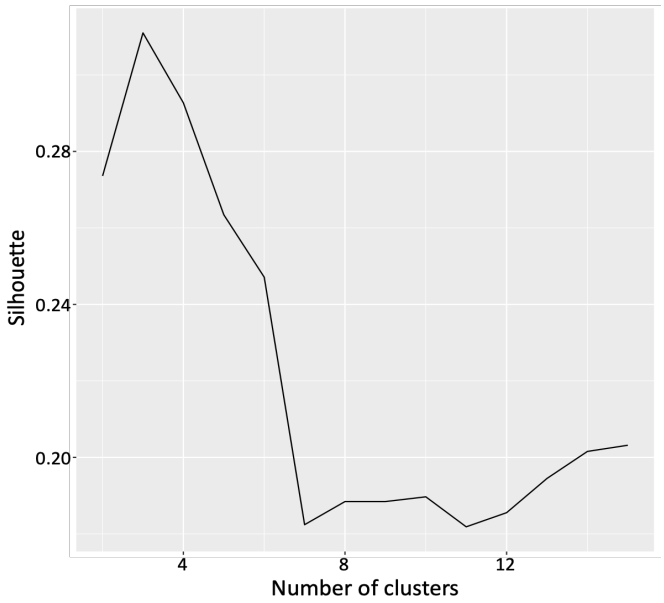
Supplementary Fig. 6. Correlation between the diseases' node degree in the DSN with sample size and the number of sDEGs. (**a**) Correlation between the node degree of the diseases in the Disease Similarity Network (DSN) and their sample size. (**b**) Correlation between the node degree of the diseases in the Disease Similarity Network (DSN) and their number of significantly differentially expressed genes (sDEGs).



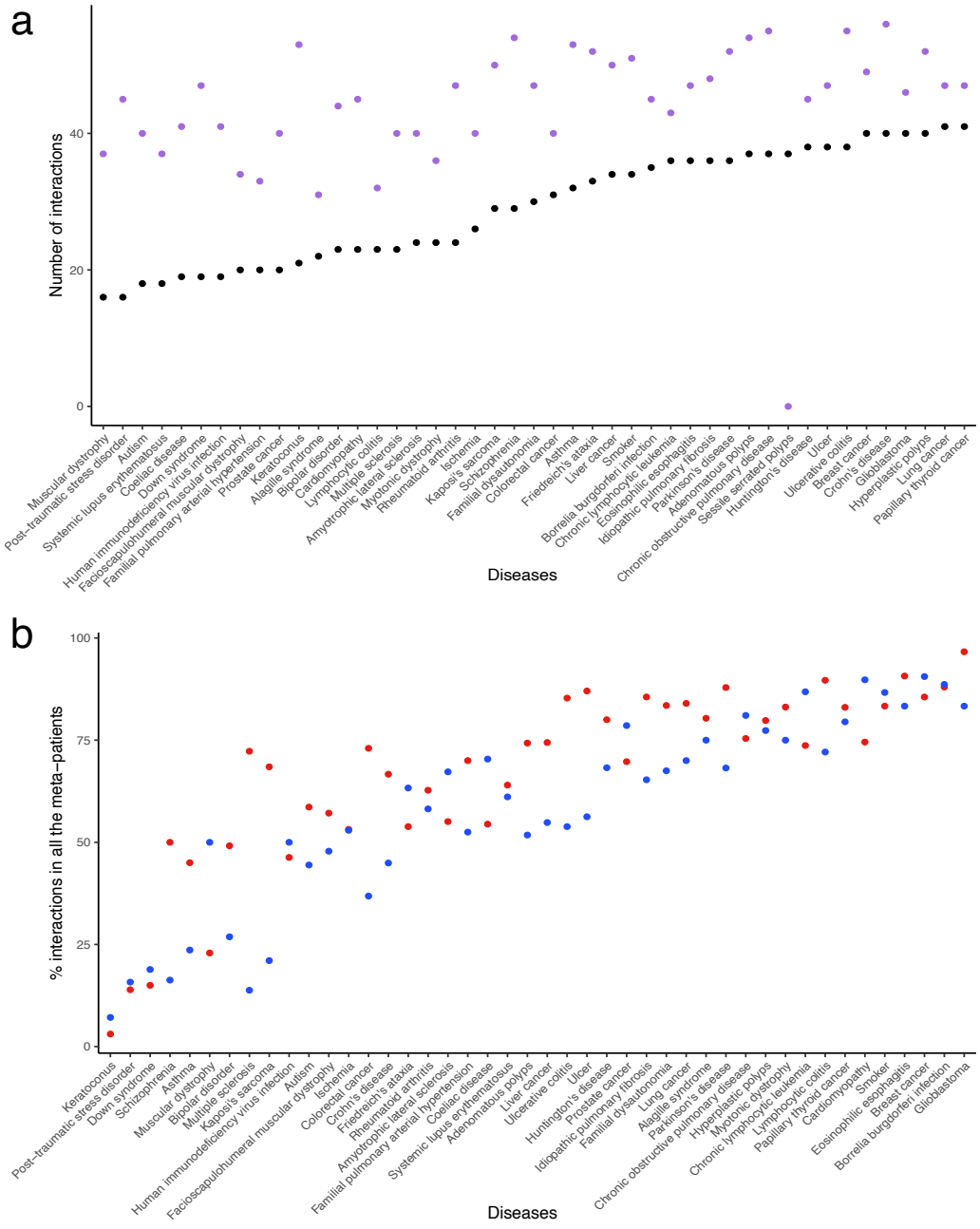
Supplementary Fig. 7. Sankey plot of the pathway categories commonly over and underexpressed in epidemiological versus non-epidemiological interactions. Each color corresponds to a Reactome pathway category. Pathway categories are sorted by the ratio between the mean number of shared pathways in epidemiological versus non-epidemiological interactions (for over and underexpressed).



Supplementary Fig. 8. Underexpressed pathways behind epidemiological and non-epidemiological interactions between neoplasms. Percentage of epidemiological (EIs) versus non-epidemiological interactions (NEIs) between neoplasms that share underexpressed pathways. Each point represents a Reactome pathway category. The size of the points corresponds to the mean number of shared pathways in the EIs. The color corresponds to the ratio of the mean number of shared underexpressed pathways in EIs versus NEIs (e.g. red indicates that epidemiological interactions share more pathways than non-epidemiological interactions). The number of EIs and NEIs is indicated between parentheses in the y and x axis labels, respectively.

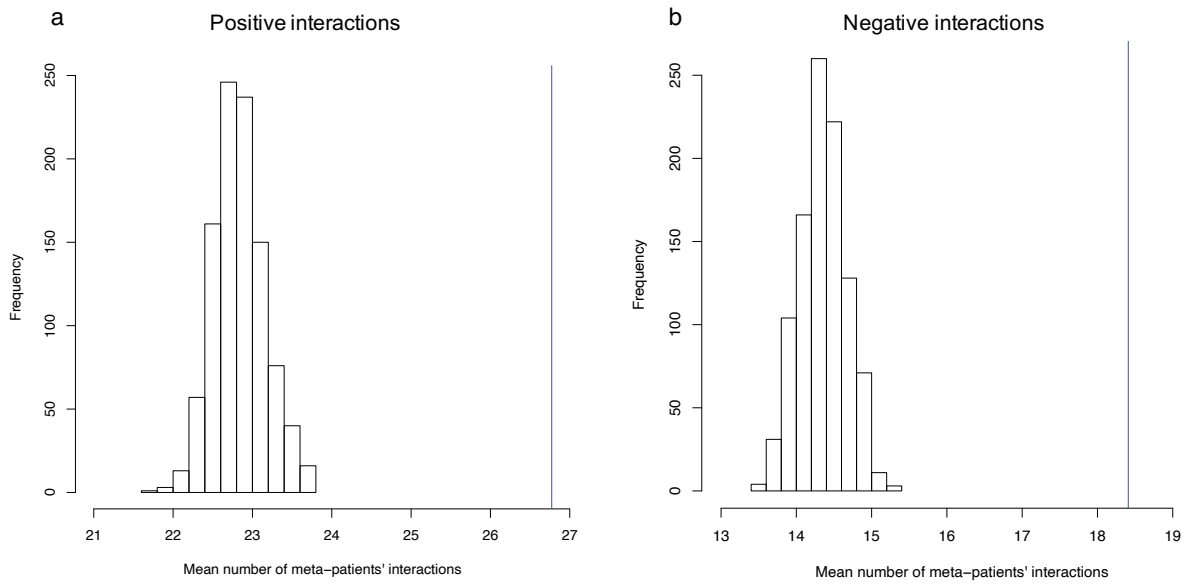


Supplementary Fig. 9. Silhouette values in breast cancer patients' clustering. Silhouette values obtained by applying PAM algorithm to cluster breast cancer patients using a number between 2 and 15.

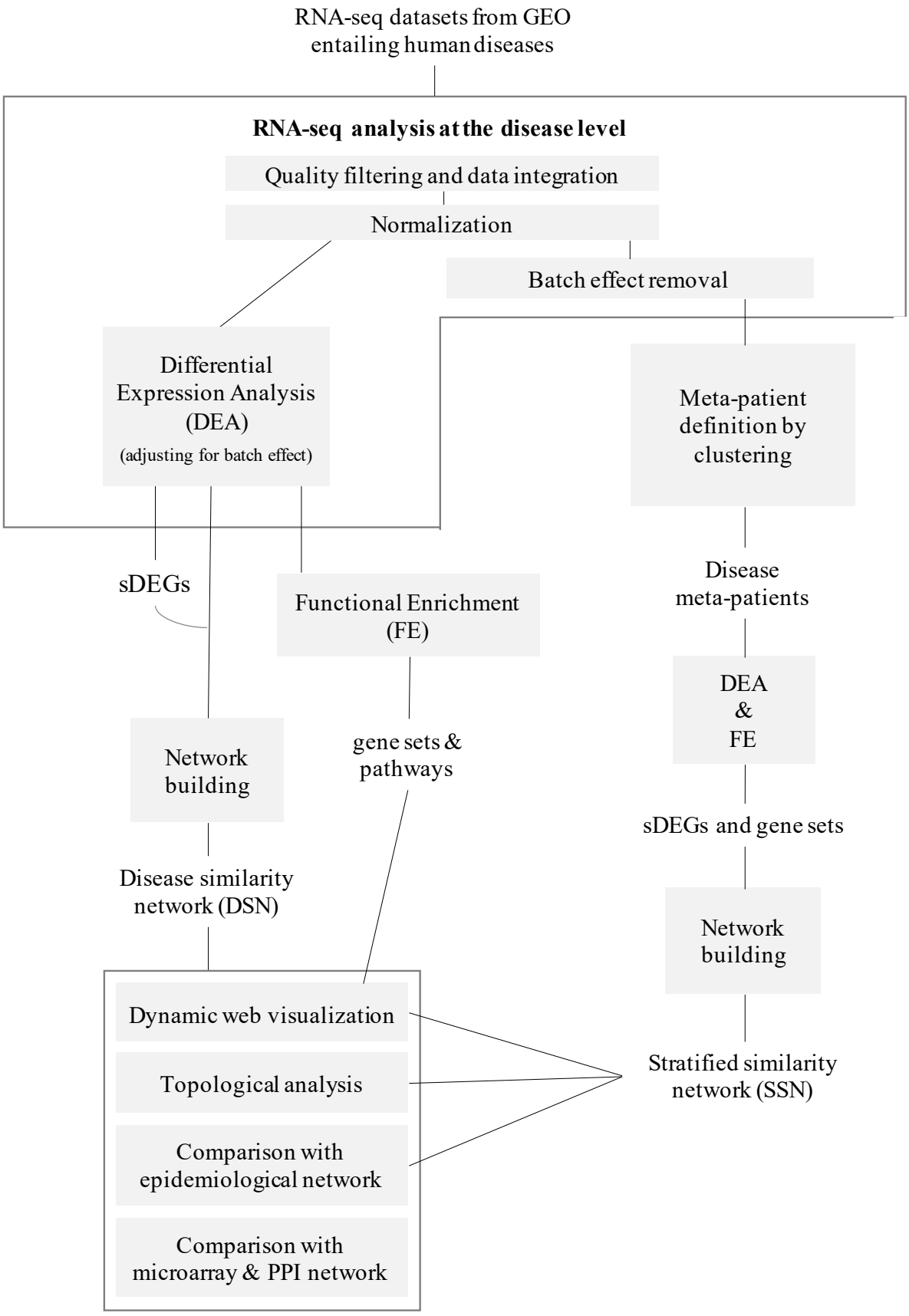


Supplementary Fig. 10. Comparison of the total number of interactions at the disease and meta-patient level and heterogeneity of the interactions at the meta-patient level. (**a**) Comparison of the total number of interactions for a given disease at the disease and meta-patient level. For each disease (columns), it is represented the total number of interactions with unique diseases at the disease level (black) and the meta-patient level (purple). Diseases are sorted based on the number of interactions at the disease level. Meta-patients were considered to be linked to a given disease if they were connected to the disease itself or one of the disease meta-patients.

Interactions between meta-patients or diseases from the same disease were discarded for the figure. (**b**) Percentage of interactions that are observed for all the diseases' meta-patients. For each disease (columns), it is represented the percentage of positive (red) and negative (blue) interactions with unique diseases that are present in all its meta-patients (rows). Diseases are sorted by the mean of their percentage considering the positive and negative interactions. Meta-patients were considered to be linked to a given disease if they were connected to the disease itself or one of the disease meta-patients. Interactions between meta-patients or diseases from the same disease were discarded for the figure.



Supplementary Fig. 11. Histograms of the mean number of meta-patients' interactions. Histogram of the mean number of meta-patients' (**a**) positive and (**b**) negative interactions obtained by randomizations (See Methods). The blue line represents the number of interactions derived from the SSN.



Supplementary Fig. 12. Pipeline of the analysis. Schema of the pipeline used in the study. First, we collected RNA-seq datasets from the Gene Expression Omnibus (GEO) entailing human

diseases. Then, we conducted the RNA-seq analysis at the disease level. To achieve that, we performed quality filtering of the data, we integrated the data and normalized it. Next, we applied differential expression analyses (DEA), from which we obtained the significantly differentially expressed genes (sDEGs) for each disease. Then, we performed functional enrichment (FE), obtaining the significantly differentially expressed and variable gene sets and pathways for each disease. Moreover, we built a Disease Similarity Network (DSN) based on the similarity between the differential gene expression profiles of the diseases. Besides, we obtained disease meta-patients by applying clustering algorithms to the normalized and batch effect corrected counts for each disease. Then, we perform DEA and FE on the meta-patients and built a Stratified Similarity Network (SSN) including the meta-patients in the DSN. Finally, we analyzed the topological properties of the DSN and the SSN and compared both networks with the epidemiological network from Hidalgo *et al.*[1]. The DSN was also compared with other disease-disease networks based on molecular information (microarray[3] and protein-protein interaction[4] data). Finally, we developed a web application in which the networks and their underlying molecular mechanisms can be easily inspected.

Supplementary Table 1. Number of genes that are significantly differentially expressed per disease. Table containing the number of significantly differentially expressed genes (sDEGs) (columns) per disease (rows). Numbers of sDEGs (total, up and down) are shown (see Methods).

	sDEGs		
Disease	**Total**	**Up**	**Down**
Adenomatous polyps	807	515	292
Alagille syndrome	1	0	1
Amyotrophic lateral sclerosis	2	2	0
Asthma	1236	647	589
Autism	5	3	2
Bipolar disorder	0	0	0
Borrelia burgdorferi infection	4668	2296	2372
Breast cancer	8493	4196	4297
Cardiomyopathy	133	57	76
Chronic lymphocytic leukemia	1027	513	514
Chronic obstructive pulmonary disease	7918	3903	4015
Coeliac disease	0	0	0
Colorectal cancer	835	316	519
Crohn's disease	8600	3879	4721
Downs syndrome	30	29	1
Eosinophilic esophagitis	5036	2448	2588
Facioscapulohumeral dystrophy	0	0	0
Familial dysautonomia	1063	355	708
Familial pulmonary arterial hypertension	0	0	0
Friedreich's ataxia	2677	1331	1346
Glioblastoma	8699	4585	4114
HIV	0	0	0
Huntington's disease	2925	1603	1322
Hyperplastic polyps	4287	2231	2056
Idiopathic pulmonary fibrosis	6519	3210	3309
Ischemia	3	2	1
Kaposi's Sarcoma	1574	728	846
Keratoconus	0	0	0
Liver cancer	6839	3712	3127
Lung cancer	7375	3691	3684
Lymphocytic colitis	0	0	0
Multiple sclerosis	292	29	263
Muscular Dystrophy	0	0	0
Myotonic Dystrophy	0	0	0
Parkinson's disease	4502	2187	2315
Posttraumatic stress disorder	0	0	0
Prostate cancer	0	0	0
Rheumatoid arthritis	0	0	0
Systemic lupus erythematosus	1	0	1
Schizophrenia	349	175	174
Sessile serrated polyposis	2746	1578	1168
Smoker	623	509	114
Thyroid cancer papillary	1513	622	891
Ulcer	4794	2375	2419
Ulcerative colitis	7089	3341	3748

Supplementary Table 2. Disease Similarity Network (DSN) properties. Table containing the general characteristics of the generated disease-disease networks (columns). Properties are provided for the DSN at the disease and ICD9 level, and for the epidemiological network from Hidalgo *et al.*[1], considering all the interactions together, and splitting them into positive and negative interactions (rows) (see Methods). It shows the number of nodes, the number of possible and detected interactions, percentage of positive and negative interactions, number and size of the connected components and mean degree of the networks (see Methods).

Type	Characteristics	DSN	DSN (ICD9)	Epidemiology
All	Number of nodes	45	41	41
	Possible	990	820	780
	Significant	658	545	331
	CC	1	1	1
	Biggest CC	45	41	40
	Mean degree	29.24	26.58	16.55
Positive	Detected	417	347	331
	Percentage	63.37	63.67	100
	CC	1	1	1
	Biggest CC	45	41	40
	Mean degree	18.53	16.93	16.55
Negative	Detected	241	198	
	Percentage	36.63	36.33	
	CC	1	1	
	Biggest CC	45	41	
	Mean degree	10.71	9.66	

Supplementary Table 3. Overlap of the ICD9-based DSN with the epidemiology. Table containing the overlap of the ICD9-based DSN with the epidemiological network from Hidalgo *et al.*[1] (See Methods). It shows the number of nodes, number of overlapping interactions, percentage of overlap and p-value with respect to the network and the epidemiology for the positive and the negative interactions (rows).

Type	Characteristics	DSN
Positive	Number of nodes in ICD9	41
	Overlap	152
	Perc. overlap from DSN	43.80
	p-value	0.004
	Perc. overlap from epidemiology	46.20
	p-value	0.0018
Negative	Number of nodes in ICD9	41
	Overlap	73
	Perc. overlap from DSN	36.87
	p-value	0.86
	Perc. overlap from epidemiology	22.18
	p-value	0.867

Supplementary Table 4. Topological properties of the Disease Similarity Network (DSN), Stratified Similarity Network (SSN) and the epidemiological network. Table containing general and topological properties of the DSN, SSN and the epidemiological network by Hidalgo *et al*.[1] (columns). Properties are provided for all the interactions, the positive and the negative ones, when applicable (rows). It shows the number and size of the connected components, mean degree, density, mean transitivity, diameter (longest shortest path) and mean distance (mean length of the shortest paths).

		DSN	**SSN**	**Epidemiology**
All	CC	1	1	1
	Size CC	45	161	995
	Mean degree	29.244	112.012	209.893
	Density	0.665	0.700	0.211
	Mean transitivity	0.705	0.746	0.519
	Diameter	0.162	0.141	4.856
	Mean distance	1.335	1.300	1.805
Positive	CC	1	1	1
	Size CC	45	161	995
	Mean degree	18.533	72.907	209.893
	Density	0.421	0.456	0.211
	Mean transitivity	0.564	0.611	0.519
	Diameter	0.225	0.193	4.856
	Mean distance	1.587	1.545	1.805
Negative	CC	1	1	-
	Size CC	45	161	-
	Mean degree	10.711	39.106	-
	Density	0.243	0.244	-
	Mean transitivity	0.156	0.151	-
	Diameter	0.252	0.201	-
	Mean distance	1.825	1.772	-

Supplementary Table 5. Topological properties of the Disease Similarity Network (DSN) based on ICD9 and the comparable epidemiological subnetwork. Table containing general topological properties of the DSN based on ICD9 and the epidemiological network by Hidalgo *et al.*[1] (columns). It shows the number and size of the connected components, mean degree, density, mean transitivity, diameter (longest shortest path), mean distance (mean length of the shortest paths), mean closeness, mean betweenness, mean degeneracy and disease category assortativity (rows). The assortativity was computed labelling the nodes with their corresponding disease category. When applicable, a paired t-test was used to obtain the significance of the differences between the mean topological values of the two networks (considering as the null hypothesis that the means are equal).

		DSN (common ICD9)	**Epidemiology (common ICD9)**	**t-test p-value**
Positive	CC	1	1	-
	Size CC	40	40	-
	N interactions	327	329	-
	Mean degree	16.35	16.45	0.953
	Density	0.419	0.422	-
	Mean distance	1.588	1.665	-
	Diameter	3	4	-
	Mean transitivity	0.556	0.664	0.001
	Mean closeness	0.0163	0.016	0.423
	Mean betweenness	11.475	12.975	0.605
	Mean degeneracy	10.825	10.95	0.845
	Dis. Categ. Assortativity	0.035	-0.036	-

Supplementary Table 6. Overlap of the microbiome, miRNA and PPI-based disease-disease networks with the epidemiology. Table containing the number of nodes, edges and overlaps of the positive interactions in the microbiome[5], miRNA[6] and PPI-based[4] disease-disease networks (columns). First, the disease names are transformed into ICD9 codes and the table shows the number of unique nodes, edges, common nodes with the epidemiological network, overlapping interactions, percentage and p-value of the overlaps with the epidemiological network from Hidalgo *et al.*[1] (rows). It also shows the above after selecting only the ICD9 codes present in the DSN (see Methods).

Type	Characteristics	Microbiome	miRNA	PPI
Positive	Number of nodes	33	63	289
	Number of edges	112	414	1383
ICD9 transformed	Unique nodes	23	46	136
	Number of edges	87	324	536
	ICD9 in epidemiology	23	46	136
	Overlap	61	129	400
	Perc. overlap from molecular network	70.11	39.81	74.63
	p-value	0.225	0.265	0
	Perc. overlap from epidemiology	36.09	36.03	8.71
	p-value	0.067	0.412	0
ICD9 in the DSN	Unique nodes	6	13	19
	Number of edges	8	40	20
	Overlap	7	9	15
	Perc. overlap from molecular network	87.5	22.5	75
	p-value	1	0.961	0.0031
	Perc. overlap from epidemiology	63.64	56.25	18.52
	p-value	1	0.96	0.0045

Supplementary Table 7. Comparison of the Disease Similarity Network (DSN, RNA-seq) with the Disease Molecular Similarity Network (microarrays) and the disease network derived from PPI. Table containing the number of nodes and edges as well as the number of common nodes and number of edges with common nodes in the DSN and the microarrays network by Sánchez-Valle *et al.*[3] or the network based on PPI by Menche *et al.*[4]. For the comparison, all the network nodes were transformed into ICD9 codes.

	Comparison with microarrays		**Comparison with PPI**	
	DSN (ICD9)	Microarrays (ICD9)	DSN (ICD9)	PPI (ICD9)
Number of nodes (ICD9)	41	92	41	289
Number of edges	545	2155	545	1383
Common ICD9	27	27	19	19
Number of edges with common ICD9	251	134	73	20

Supplementary Table 8. Overlap between the Disease Similarity Network (DSN, RNAseq) with the Disease Molecular Similarity Network (microarrays) and the disease network derived from PPI. Table containing the overlap between the DSN based on ICD9 codes with the microarrays network from Sánchez-Valle *et al.*[3] and the PPI-based network derived from Menche *et al*.[4] (columns). The overlapping is provided for all the interactions, the positive and negative ones when applicable (rows) (see Methods). It shows the overlap (number of common interactions), the percentage (percentage of interactions from the molecular networks captured by the DSN) and significance of the overlap (see Methods).

Type	Characteristics	DSN (ICD9) vs microarrays	DSN (ICD9) vs PPI
All	Overlap	63	6
	Percentage overlap	47.015%	30%
	p-value	0.0272	0.942
Positive	Overlap	45	6
	Percentage overlap	65.217%	30%
	p-value	0.002	0.942
Negative	Overlap	18	-
	Percentage overlap	27.692%	-
	p-value	0.6243	-

Supplementary Table 9. Breast cancer patients' classification using PAM and Ward2 algorithms. Table containing the breast cancer patient's distribution in the clusters obtained with PAM and Ward2 algorithms and their correspondence with disease subtypes (see Methods). Obtained clusters are represented in columns. Rows correspond to molecular disease subtypes: triple negative (TN), estrogen receptor positive and negative (ER+ and ER-, respectively). Ward's second cluster is divided into its two branches.

	Cluster 1	Cluster 2	Cluster 3	
Triple negative (TN)	16	2		
Estrogen + (ER+)		27	3	**PAM**
Estrogen (ER-)		3	7	

	Cluster 1	Cluster 2		
		Branch 1	*Branch 2*	
Triple negative (TN)	16	2		**Ward2**
Estrogen + (ER+)		24	5	
Estrogen - (ER-)		1	8	

Supplementary Table 10. Overlap of the interactions at the meta-patient level with the epidemiology. Table containing the overlap of the interactions between meta-patients and diseases with the epidemiological network from Hidalgo *et al.*[1] (See Methods). It shows the number of overlapping interactions, percentage of overlap and p-value with respect to the network and the epidemiology for the positive and the negative interactions (rows).

Type	Characteristics	Meta-patient level
Positive	Overlap	211
	Perc. overlap from DSN	43.06
	p-value	0.0253
	Perc. overlap from epidemiology	64.13
	p-value	0.0187
Negative	Overlap	135
	Perc. overlap from DSN	41.28
	p-value	0.8082
	Perc. overlap from epidemiology	40.79
	p-value	0.8035

Supplementary Table 11. Overlap of the DSN with the epidemiology using the alternative approach. Table containing the overlap of the DSN with the epidemiological network from Hidalgo *et al.*[1] by grouping the diseases that correspond to the same ICD9 code (See Supplementary Notes). It shows the number of nodes, overlapping interactions, percentage of overlap and p-value with respect to the network and the epidemiology for the positive and the negative interactions (rows).

Type	Characteristics	DSN
Positive	Number of nodes in ICD9	41
	Overlap	129
	Perc. overlap from DSN	42.43
	p-value	0.0236
	Perc. overlap from epidemiology	39.21
	p-value	0.0082
Negative	Number of nodes in ICD9	41
	Overlap	58
	Perc. overlap from DSN	36.48
	p-value	0.8891
	Perc. overlap from epidemiology	17.63
	p-value	0.9229

References

1. Hidalgo, C. A., Blumm, N., Barabási, A. L. & Christakis, N. A. A Dynamic Network Approach for the Study of Human Phenotypes. *PLoS Comput. Biol.* **5**, e1000353 (2009).

2. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

3. Sánchez-Valle, J. *et al.* Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships. *Nat. Commun.* **11**, 2854 (2020).

4. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science.* **347**, 841 (2015).

5. Ma, W. *et al.* An analysis of human microbe-disease associations. *Brief. Bioinform.* **18**, 85–97 (2017).

6. Lu, M. *et al.* An analysis of human microRNA and disease associations. *PLoS One* **3**, e3420 (2008).