Online Methods

Average normalized difference

Average normalized difference is used to measure patient similarity where an input feature used five or fewer variables. For a single variable *g* (e.g. age), similarity based on normalized difference is defined as:

$$S(a,b,g) = 1 - \frac{abs(a-b)}{max(g) - min(g)}$$

The term with the fraction measures normalized difference, and inverting the value (by subtracting it from 1) results in the similarity metric. For a set of *k* variables $G=\{g_1,g_2,..g_k\}$, where $1<=k<=5$, the similarity *S* between two patients *a* and *b* is defined as the average of normalized differences for each of the variables:

$$S(a,b,G) = \frac{\sum_{i=1}^{k} \frac{abs(a_i - b_i)}{max(g_i) - min(g_i)}}{k}$$

Integrated patient network

The integrated patient network starts by compiling the edges from all feature-selected networks into a single network, such that each pair of patients now has multiple similarity edges. The similarity between two patients in the integrated network is the mean of these pairwise similarities. Visually, the goal is to view more similar patients as being more tightly grouped, and more dissimilar patients as being farther apart. Similarity is therefore converted to dissimilarity, defined as 1-similarity. Weighted shortest path distances are

computed on this resulting dissimilarity network. For visualization, only edges representing the top 20% of distances in the network are included. For the network with a single clinical network, the top 50% of distances are included, to limit the number of patients without edges.

Pathway networks

Pathway definitions were aggregated from HumanCyc[1] (http://humancyc.org), IOB's NetPath[2] (http://www.netpath.org), Reactome[3,4] (http://www.reactome.org), NCI Curated Pathways[5], mSigDB[6] (http://software.broadinstitute.org/gsea/msigdb/), and Panther[7] (http://pantherdb.org/) (downloaded from http://download.baderlab.org/EM_Genesets/January_24_2016/Human/symbol/Human_AllPathways_January_24_2016_symbol.gmt)[8]. Only pathways with 10 to 500 genes were included (1,801 pathways). Pathway-level patient similarity was defined as the Pearson correlation of the expression vectors corresponding to member genes, and the network was sparsified (see next section).

Sparsification of input networks

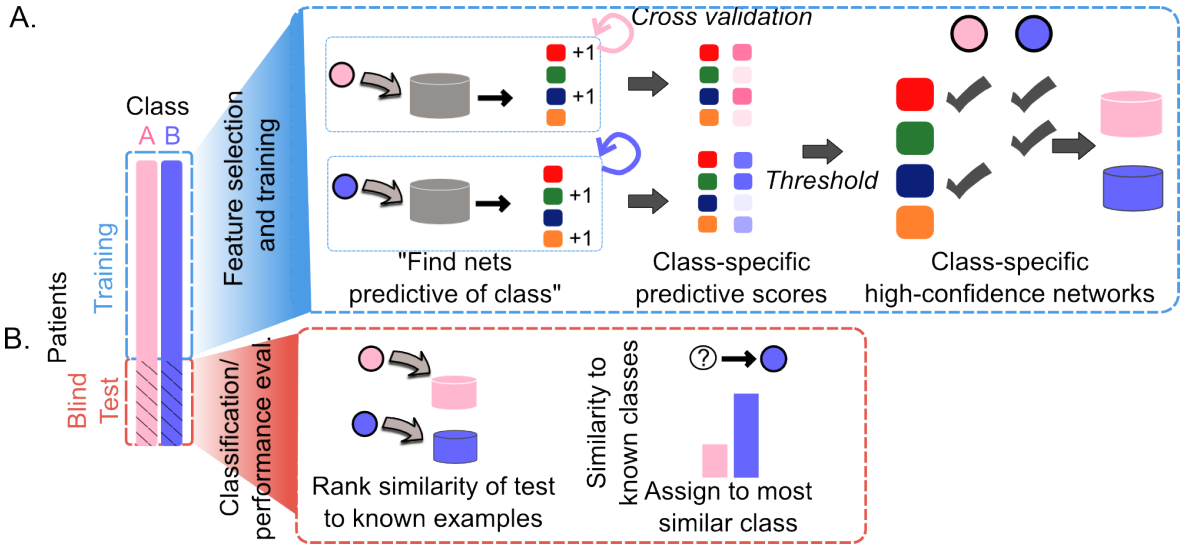
All negative similarities are set to zero. The 50 strongest correlations are kept per patient. In case of ties, all interactions tied with the 50th ranked interaction are retained, for a maximum of 2% of the sample size, or 600 patients. This follows the parameters established in the GeneMANIA algorithm (used here to integrate networks) for gene expression correlation network sparsification[9].

Map of feature-selected networks

The Enrichment Map app (v2.1.1-HOTFIX_1) in Cytoscape 3.5.1[10] was used to generate the map of selected pathways in Figure 2D[8]. A Jaccard overlap threshold of 0.05 was used to prune identical gene sets. AutoAnnotate v1.1.0 was used to cluster similar pathways using MCL clustering with default parameters. The network was visualized in Cytoscape 3.5.1[30].

The weighted shortest path between patient classes (a node set) was computed using Dijkstra's method (*igraph* v1.01[11]); distance was defined as 1-similarity (or edge weight from a patient similarity network). The overall shortest path was defined as the mean pairwise shortest-path for a node set.

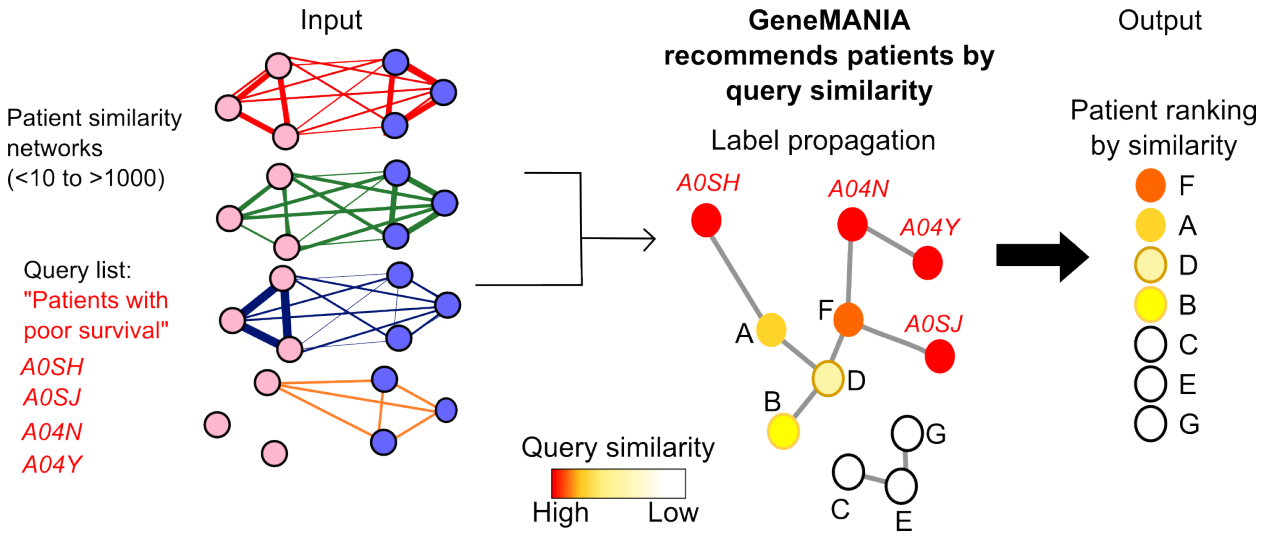
Supplementary Figures



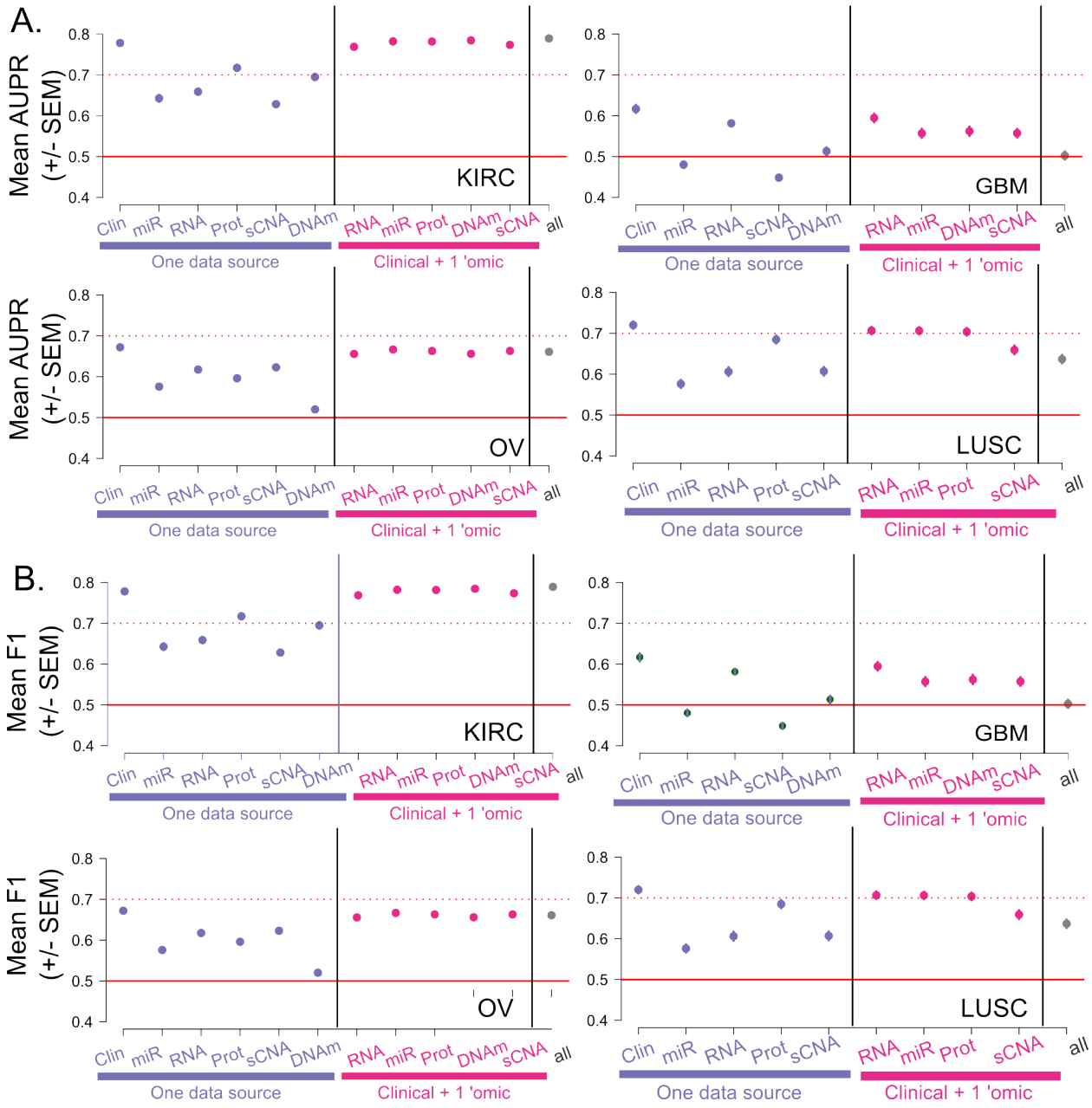
Supplementary Figure 1. Details of the netDx feature selection and patient classification steps.

A. Machine learning is used to identify networks predictive of each patient class. Data are split into training and blind test samples, and feature selection uses only training samples. Cross validation is used to score how frequently a network is predictive of a given class (e.g. high-risk). This step results in network scores, with higher values indicating networks that contribute more to prediction. These scores can be thresholded to identify a set of high-confidence networks for each class of interest (pink and blue cylinders).

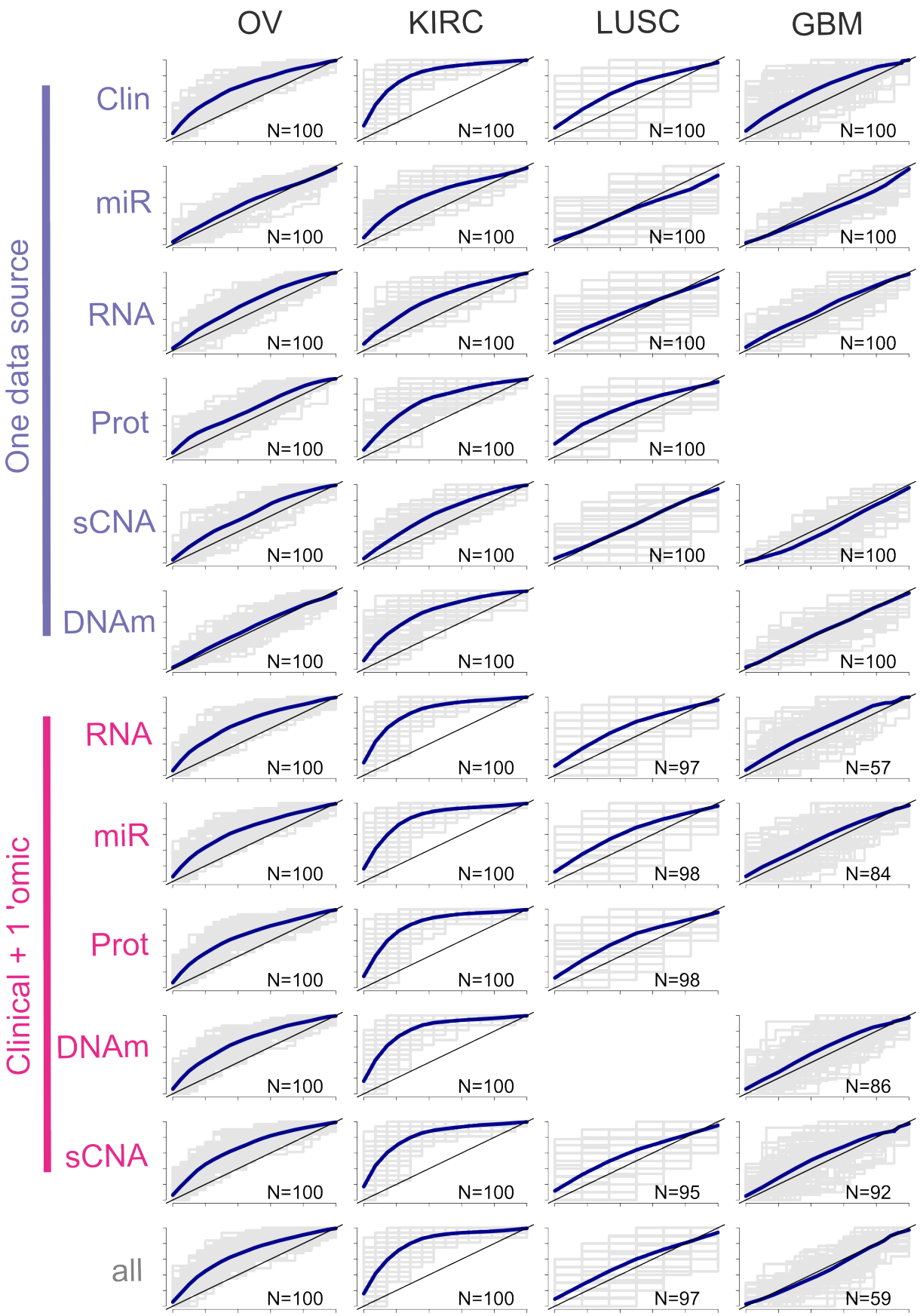
B. Blind test patients are ranked by similarity to known examples from the training set. For this step, only class-specific feature-selected networks are used. Patients are assigned to the class to which they have highest similarity.



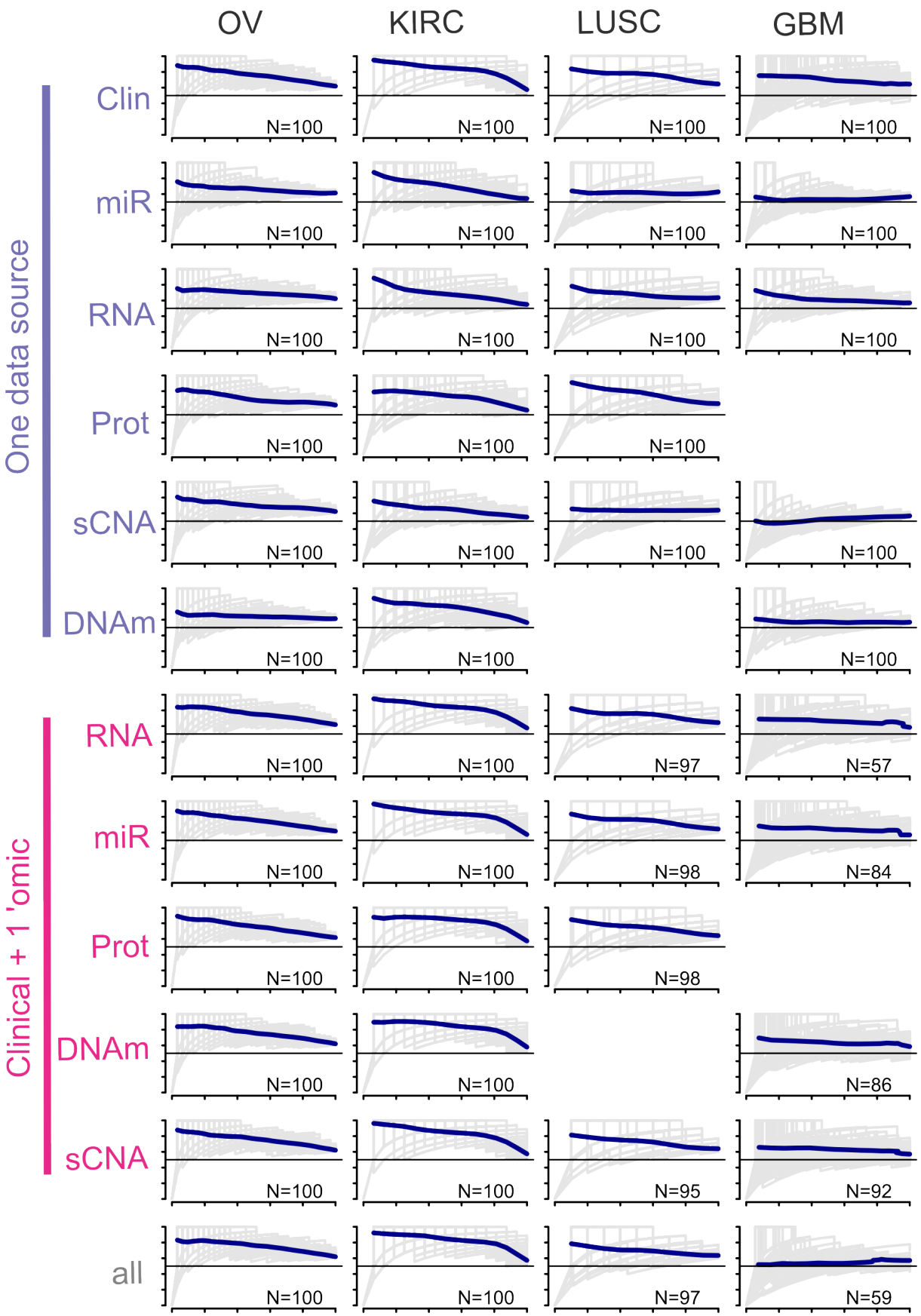
Supplementary Figure 2. Conceptual overview of the GeneMANIA algorithm, used for network integration here. GeneMANIA is a network-based recommender system that ranks all nodes by similarity to an input query (or "positive" nodes). In netDx, the nodes are patients and GeneMANIA uses the set of user-defined patient similarity networks (left). An example application is predicting KIRC poor survival by ranking all patient tumours by similarity to known KIRC poor survivors. The patient ranking is achieved by a two-step process. First, input networks are integrated into a single association network via regularized regression that maximizes connectivity between nodes with the same label and reduces connectivity to other nodes (middle); this step computes network weights or predictive value. Second, label propagation is applied to the integrated network starting with the query nodes (red), thereby ranking patients from most to least similar to the query (right).



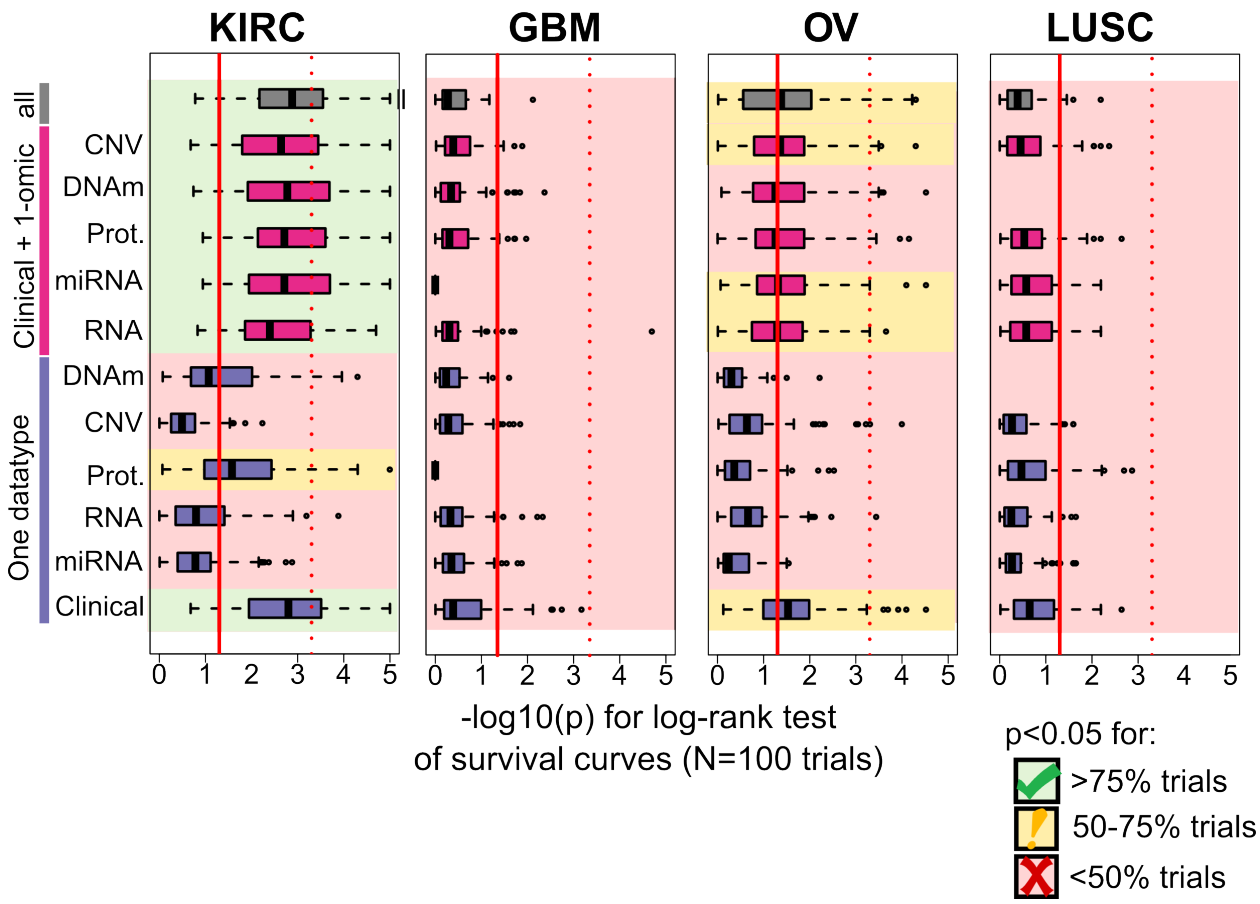
Supplementary Figure 3. AUPR and F1 score for PanCancer binary survival prediction, when each datatype is coded as a single similarity network. (A) shows AUPR across cancer types and data combinations, and (B) shows F1 score across cancer types and data combinations. Each dot shows mean over 100 train/blind test splits, and error bars show standard error of the mean (SEM).


OV
KIRC
LUSC
GBM
One data source
Clin
miR
RNA
Prot
sCNA
DNAm
Clinical + 1 'omic
RNA
miR
Prot
DNAm
sCNA
all
N=100
N=97
N=57
N=98
N=84
N=86
N=95
N=92
N=59

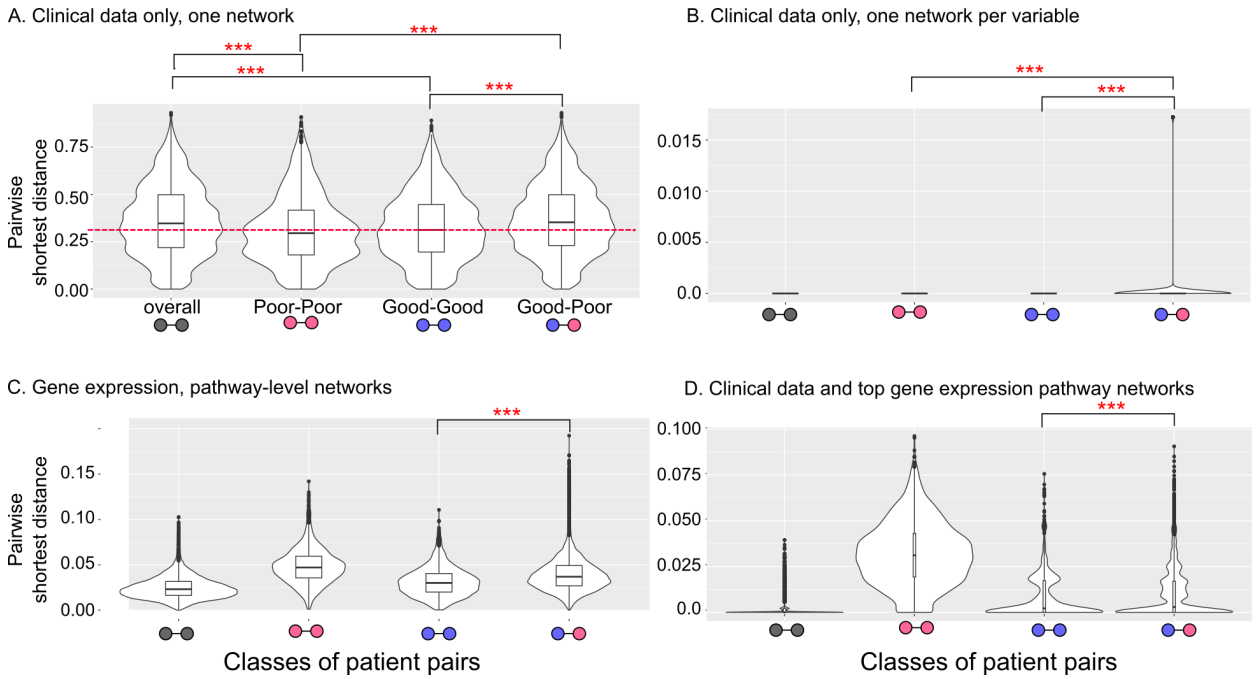
Supplementary Figure 4. ROC curves for PanCancer binary survival discrimination when each source is represented by a single network. Rows are input data types and columns show data for different cancer types. Blue: Mean ROC curve (calculated by averaging rank-ordered x and y-values respectively; Black: random predictor performance. N shows the number of splits with 1+ features selected; splits without selected features are not included. Blank slots mean that data type was not available for the given cancer type.


OV
KIRC
LUSC
GBM
One data source
Clin
miR
RNA
Prot
sCNA
DNAm
Clinical + 1 'omic
RNA
miR
Prot
DNAm
sCNA
all
N=100
N=97
N=57
N=98
N=84
N=86
N=95
N=92
N=59

Supplementary Figure 5. Precision-recall curves for PanCancer binary survival discrimination when each source is represented by a single network. Rows are data types included, and columns show data for different cancer types. Blue: Mean PR curve (calculated by averaging rank-ordered x and y-values respectively; Black: y=0.5. N indicates the number of splits with 1+ feature selected networks; splits without feature selected networks are not included in the plot. Blank slots mean that data type was not available for the given cancer (no protein data for GBM and no methylation data for LUSC).



Supplementary Figure 6. Separation of survival profiles of netDx-predicted good and poor survivors in 4 tumour types (panels). Each panel shows data for an individual tumour type and each boxplot shows log-rank test p-values for 100 train/blind test trials for a given combination of input data. The background for each boxplot is coloured based on whether the condition passed, conditionally failed, or failed the test of having separable survival curves in the two predicted classes (legend). Test status is based on the fraction of the 100 splits to achieve $p < 0.05$ on the log-rank test. Reference lines indicate $p = 0.05$ (solid) and $p = 0.0005$ (Bonferroni-corrected p; dashed).



Supplementary Figure 7. Pairwise weighted shorted path distances for the KIRC integrated patient similarity network, based on different feature designs. Each panel shows data for a single configuration. Within a panel, the boxplot and violin plots show the distribution of pairwise distances; nodes are grouped by the class identity of node pairs.

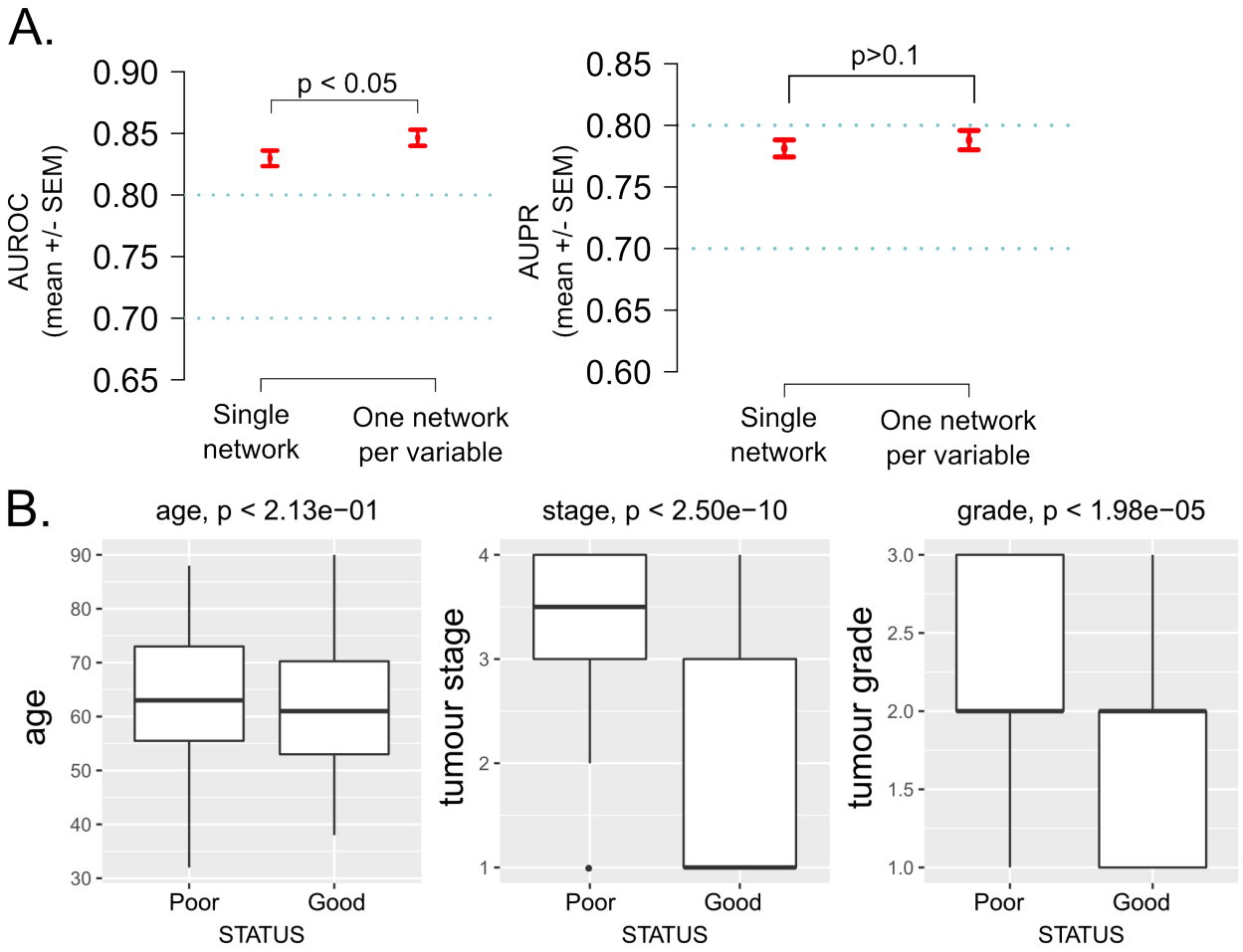
A. Predictor configuration limited to clinical data, where clinical data is represented as a single patient similarity network.

B. Limited to clinical data, where each clinical variable is represented as its own patient similarity network.

C. Limited to gene expression data, where input features are defined at the pathway level.

D. Clinical data combined with gene-expression-based pathways that were feature-selected in C.

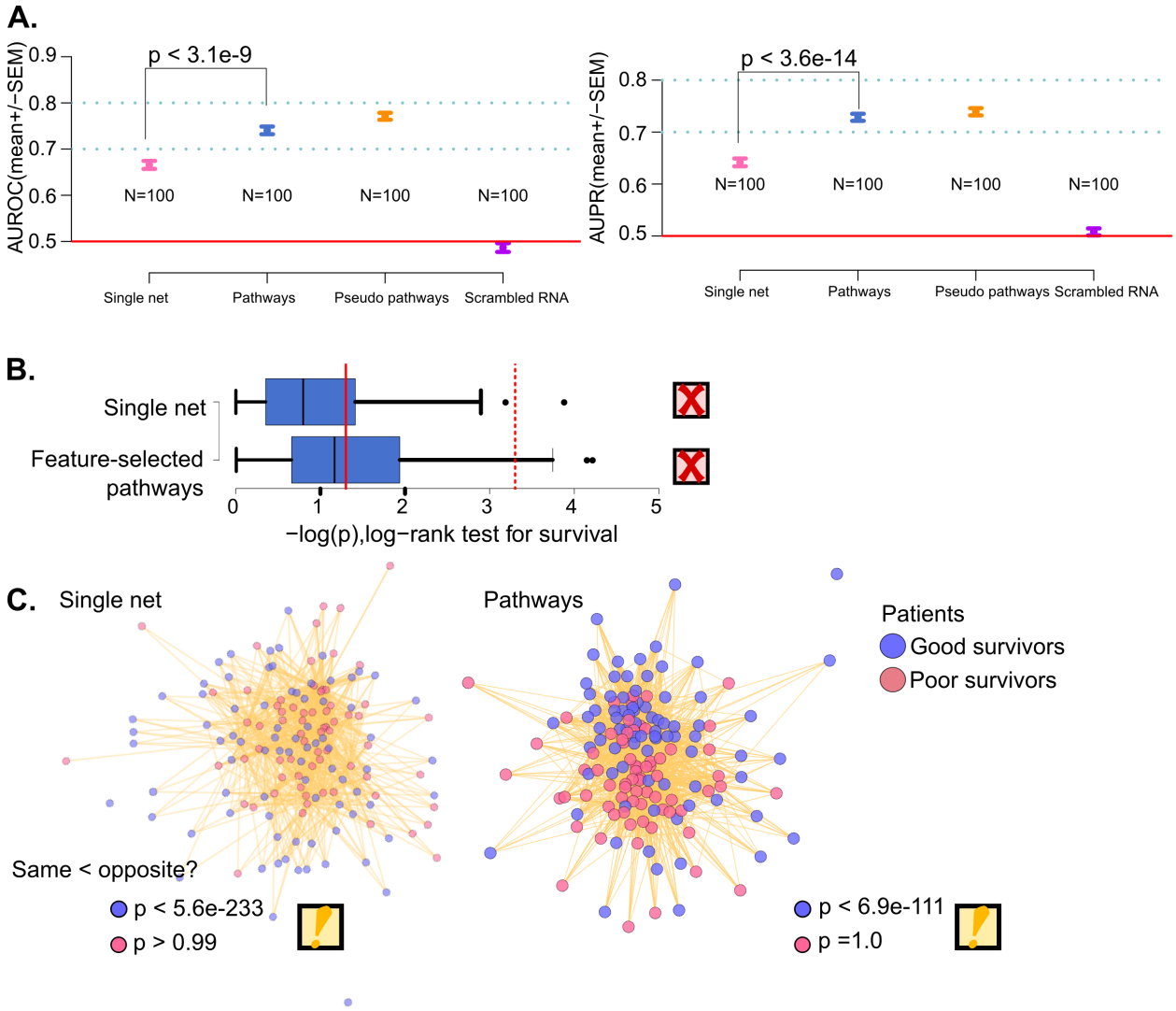
*** one-sided WMW (same class < different classes or same class < overall), $p < 0.001$



Supplementary Figure 8. Effect of individual clinical variables on KIRC binary survival prediction.

A. Area under the ROC curve and precision-recall curve of a netDx predictor where clinical data is grouped as a single input feature, compared to one where each variable is its own input feature. Shown is the mean over 100 train/blind test splits; error bars show SEM.

B. Individual clinical variables stratified by patient survival outcome. P-values shown are from two-sided WMW tests.

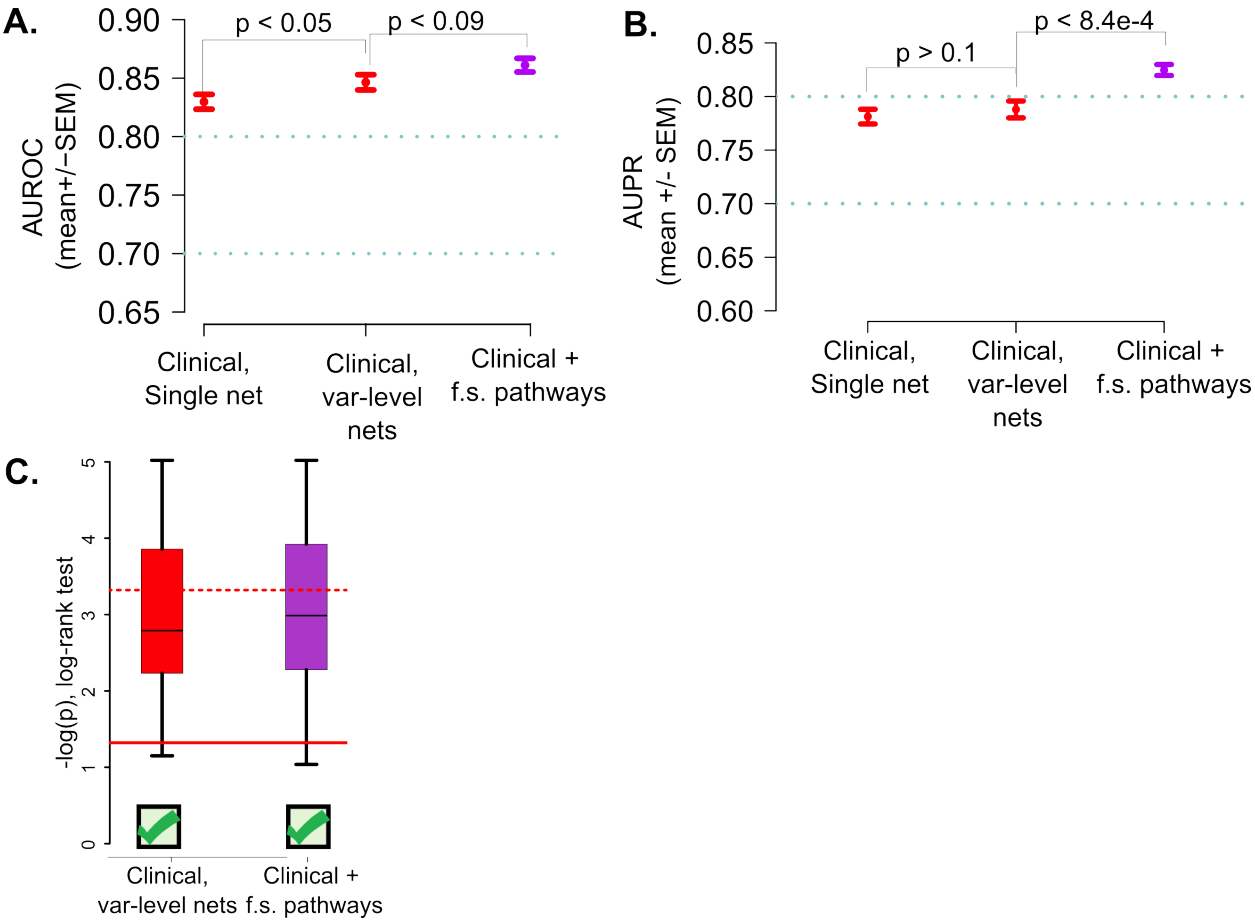


Supplementary Figure 9. Effect of feature design and pathway-level features for mRNA on predicting KIRC survival

A. netDx predictor performance based on whether gene expression was organized as a single patient similarity net, into pathway-based features, random gene groupings (pseudo pathways) size-matched with annotated pathways, or if the gene expression matrix was scrambled to simulate a random data source. Shown are mean+/-SEM over 100 train/test splits. P-values from one-sided WMW tests.

B. Performance of log-rank survival test for netDx-predicted low and high survival. Each boxplot shows the statistical significance (-log10(p)) for a given predictor configuration, over 100 test sets. Legend identical to Figure 3. The two distributions have different medians (one-sided WMW $p < 0.003$).

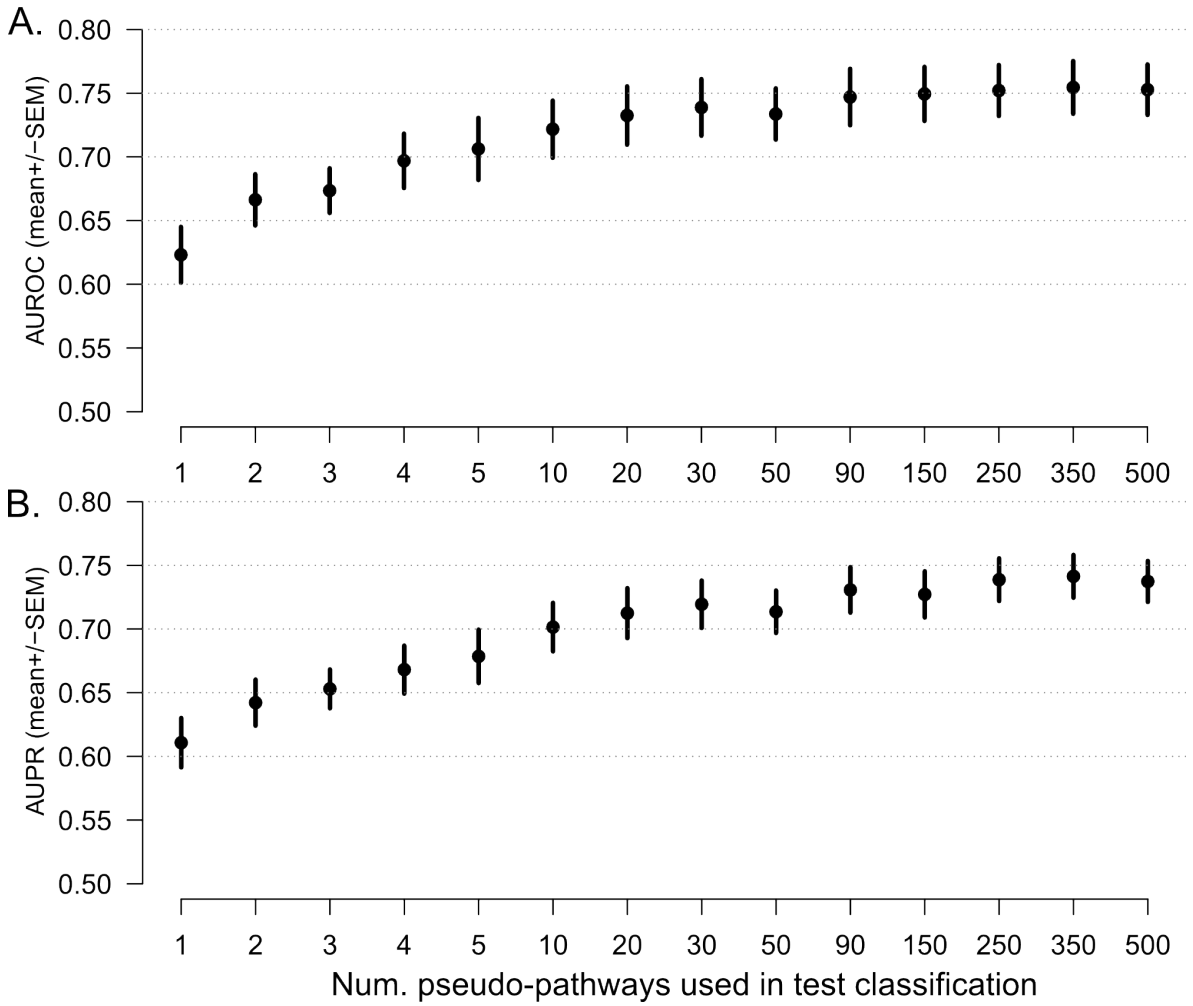
C. Integrated patient similarity network for RNA as a single feature (left) as compared to pathway-level features (right). Each node is a patient, coloured by survival type (red: poor survival; blue: good survival). An edge indicates the mean similarity between the two patients across all feature-selected networks (union for good and poor survival). The network was pruned to retained the largest 20% of distances, and a spring-embedded layout was applied in Cytoscape.



Supplementary Figure 10. Integrating expression of feature-selected pathways with clinical data.

A, B. Predictor performance based on whether clinical data was organized as a single patient similarity net, into variable-level features, or integrated with pathway-level features from gene expression data, using pathways that had been previously feature-selected. Shown are mean AUROC (A) or AUPR (B) +/-SEM over 100 train/test splits. P-values from one-sided WMW tests.

C. Performance of log-rank survival test for netDx-predicted low and high survival. Each boxplot shows the statistical significance (-log10(p)) for a given predictor configuration, over 100 test sets. Legend identical to Figure 3. The difference between the two is not statistically significant (one sided WMW p > 0.2)



Supplementary Figure 11. Predictor performance increases with increasing number of selected features. Binary survival was being predicted for renal clear cell carcinoma (KIRC), using solely gene expression data. "Pseudo pathways" were generated by randomly-sampling exactly 10 genes, out of a universe of genes not annotated as being in pathways (11,326 genes). The x-axis shows the effect of increasing the number of pseudo pathways provided as input to test sample classification (artificially feature selected). Performance is shown as mean AUROC (top) or AUPR (bottom) across 25 train/test splits. Error bars show SEM.

Supplementary Tables

	KIRC	OV	LUSC	GBM
Clinical	0.78	0.67	0.72	0.62
sCNA	0.63	0.62	0.61	0.45
Mir	0.64	0.58	0.58	0.48
RNA	0.66	0.62	0.61	0.58
Prot	0.72	0.6	0.68	NA
DNAm	0.69	0.52	NA	0.51
clinical + RNA	0.77	0.66	0.7	0.57
clinical + prot	0.78	0.66	0.7	NA
clinical + mir	0.78	0.67	0.7	0.56
clinical + DNAm	0.78	0.66	NA	0.56
clinical + CNV	0.77	0.66	0.65	0.56
all	0.79	0.66	0.63	0.52

Supplementary Table 1. Average F1 for binarized survival prediction data for kidney, ovarian, lung and brain cancers. In each case, the value shown is the average of F1 across 100 train/blind test splits.

Pathway	Max score in >=70% trials
Reactions specific to the complex N-glycan synthesis pathway	10
Defects in cobalamin (B12) metabolism	9
Defects in vitamin and cofactor metabolism	9
Metabolism of folate and pterines	9
Platelet adhesion to exposed collagen	9
Thyroxine biosynthesis	9
Activation of the pre-replicative complex *	8
FGFR2 ligand binding and activation *	8
Regulation of pyruvate dehydrogenase (PDH) complex *	8
RORA activates gene expression	8
TAK1 activates NFkB by phosphorylation and activation of IKKs complex *	8
BioCarta STEM pathway *	7
Calnexin calreticulin cycle	7
Glypican 1 network	7
Hedgehog off state	7
Retinol biosynthesis *	7
VEGF and VEGFR signaling networks	7
Androgen biosynthesis	6
Hedgehog ligand biogenesis	6
Regulation of cholesterol biosynthesis by SREBP (SREBF)	6
Synthesis of PC *	6
Metabolism of water-soluble vitamins and cofactors	5
The NLRP3 inflammasome *	5
Removal of aminoterminal propeptides from gamma-3carboxylated proteins	4
Gamma-carboxylation, transport, and amino-terminal cleavage of proteins	3

Supplementary Table 2. Scores for pathway-level networks for predicting good survival in renal clear cell carcinoma (KIRC) survival. Score shown is the best achieved by a given network for over 70% of the 100 trials. Only networks scoring a max of 3 or more out of 10 in over 70% trials are shown here. Asterisks indicate singleton nodes omitted from the Enrichment Map in Figure 4A.

Pathway name	Max score in >=70% trials
Abacavir transport and metabolism	10
Bile salt and organic anion SLC transporters	10
Metabolism of water-soluble vitamins and cofactors	10
Platelet Adhesion to exposed collagen	10
Regulation of IFNA signaling	10
Thyroxine biosynthesis	10
Aquaporin-mediated transport	9
Calnexin calreticulin cycle *	9
Class C 3 (Metabotropic glutamate pheromone receptors) *	9
Metabolism of folate and pterines	9
POU5F1 (OCT4), SOX2, NANOG activate genes related to proliferation *	9
Vasopressin regulates renal water homeostasis via Aquaporins	9
Vitamin B5 (pantothenate) metabolism	9
Androgen biosynthesis	8
Regulation of gene expression by Hypoxia-inducible Factor *	8
Hormone ligand-binding receptors	4

Supplementary Table 3. Scores for pathway-level networks for predicting poor survival in renal clear cell carcinoma (KIRC) survival. Score shown is the best achieved by a given network for over 70% of the 100 trials. Only networks scoring a max of 3 or more out of 10 in over 70% trials are shown here. Asterisks indicate singleton nodes omitted from the Enrichment Map in Figure 4A.

Additional References

1. Romero, P. et al. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* **6**, R2 (2005).
2. Kandasamy, K. et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol* **11**, R3 (2010).
3. Croft, D. et al. The Reactome pathway knowledgebase. *Nucleic Acids Res* **42**, D472-7 (2014).
4. Fabregat, A. et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481-7 (2016).
5. Schaefer, C.F. et al. PID: the Pathway Interaction Database. *Nucleic Acids Res* **37**, D674-9 (2009).
6. 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**, 15545-50 (2005).
7. Mi, H. et al. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* **33**, D284-8 (2005).
8. Merico, D., Isserlin, R. & Bader, G.D. Visualizing gene-set enrichment results using the Cytoscape plug-in enrichment map. *Methods Mol Biol* **781**, 257-77 (2011).
9. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol* **9 Suppl 1**, S4 (2008).
10. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).
11. Csardi G., N.T. The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2006).