Supplementary Methods

General. Unless otherwise specified, method development and analysis was performed in R v3.2.5 and used Bioconductor[1].

Pathways. Pathway definitions were aggregated from HumanCyc[2] (http://humancyc.org), IOB's NetPath[3] (http://www.netpath.org), Reactome[4,5] (http://www.reactome.org), NCI Curated Pathways[6], mSigDB[7] (http://software.broadinstitute.org/gsea/msigdb/), and Panther[8] (http://pantherdb.org/) (downloaded from http://download.baderlab.org/EM_Genesets/January_24_2016/Human/symbol/Human_AllPathways_January_24_2016_symbol.gmt)[9]. Only pathways with 10 to 500 genes were included.

Cross-method comparison. Data consisted of TCGA level 3 gene expression data for 348 primary breast tumours (https://tcga-data.nci.nih.gov/docs/publications/brca_2012/); the goal is binary classification of a tumour as being of type "Luminal A" or not. For all methods, 67% of samples were kept as training data. The R package *caret* (v6.0-7)[10] was used to automate the pipeline and tune model parameters. Elastic net implementation is from the R *glmnet* package (v2.0-5)[11], and random forests implementation from the *randomForest* package (v4.6-12)[12]. All three methods - elastic nets, random forests and netDx - used 10-fold cross validation over three re-samplings for feature selection. Code is provided as part of the netDx codebase; see http://netdx.org/index.php/netdx-reviewer-page/

Sparsification of input networks: All negative similarities are set to zero. For a patient, interactions that are not among the 50 strongest correlations, are excluded. In case of ties, all interactions tied with the 50th ranked interactions are retained, for a maximum of 2% of the sample size or 600 patients. This follows the parameters established in the original GeneMANIA algorithm for gene expression correlation network sparsification.

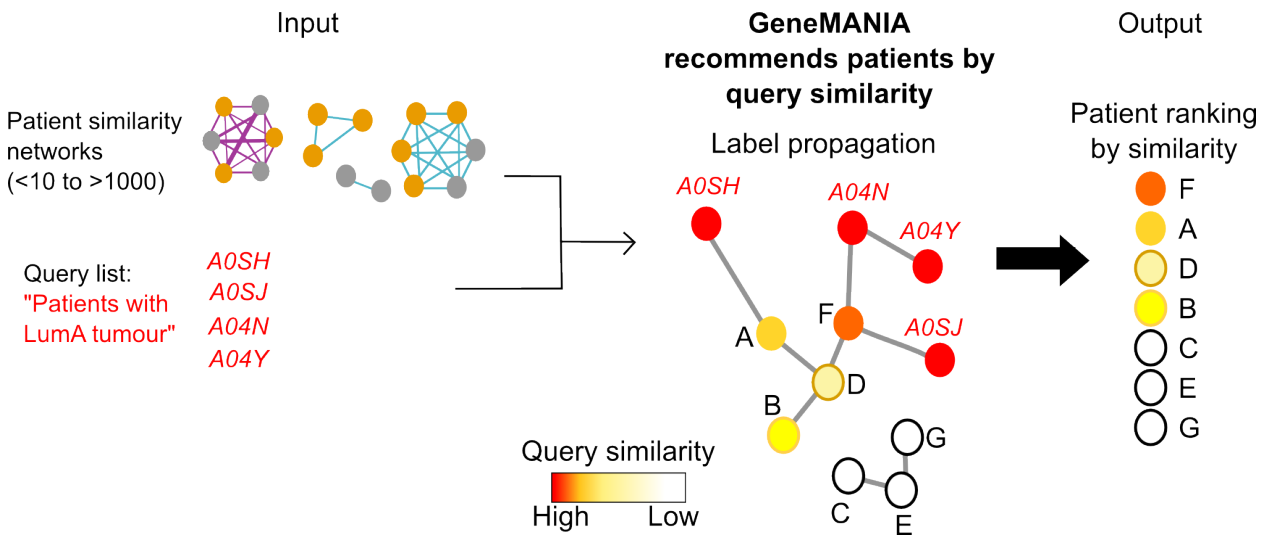
Map of feature-selected networks. The Enrichment Map app (v2.1.1-HOTFIX_1) in Cytoscape 3.4.0 was used to generate the map of selected pathways in Figure 2D[9]. A Jaccard overlap threshold of 0.05 was used to prune identical gene sets. AutoAnnotate v1.1.0 was used to cluster similar pathways using MCL clustering with default parameters. The network was visualized in Cytoscape 3.4.0[13].

Visualization of patient similarity network. Networks that scored 10 out of 10 in the feature selection algorithm were concatenated; edges with similarity less than 0.7 were excluded. Multiple edges between the same pair of patients were resolved by taking the maximum edge weight. The resulting network was visualized in Cytoscape. Nodes were organized by the edge-weighted spring-embedded layout based on the integrated similarity value.

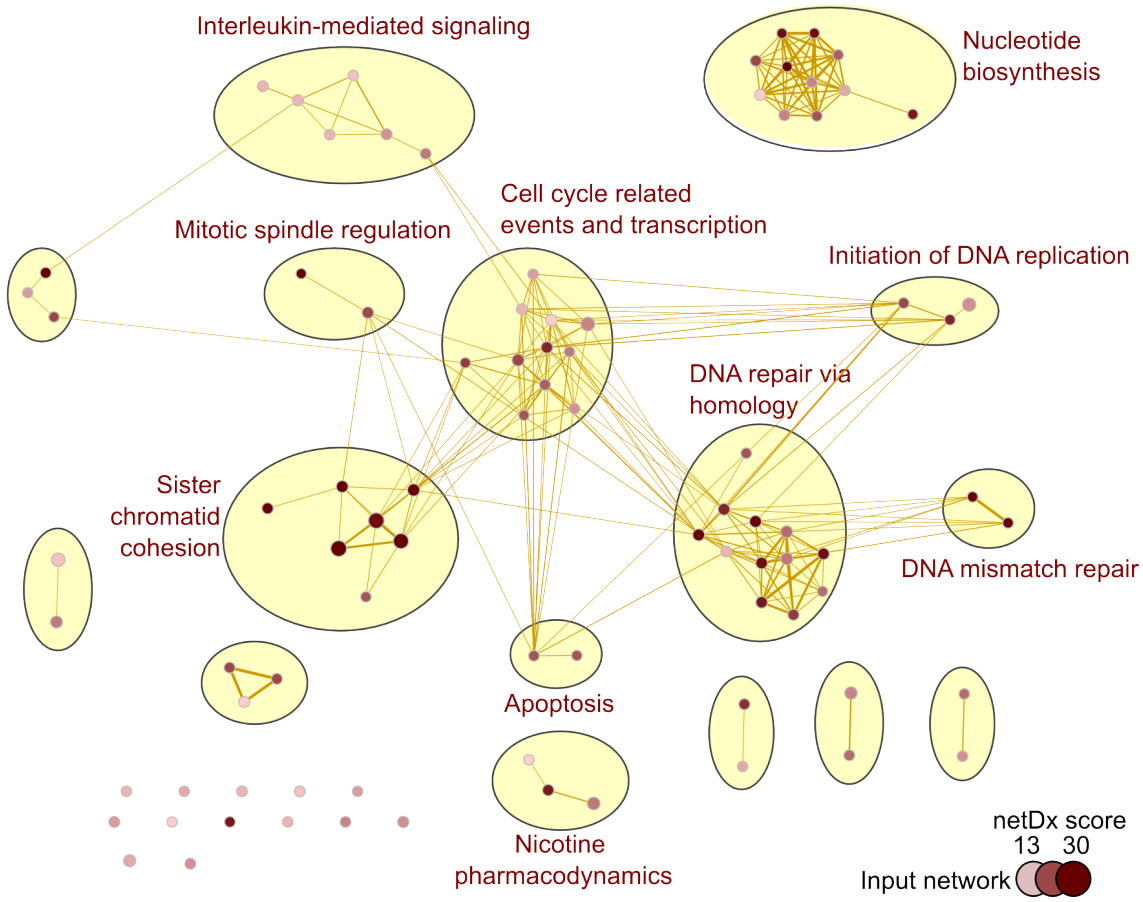
The shortest path between patient classes (a node set) was computed using Dijkstra's method for weighted edges (*igraph* v1.01[14]); distance was defined as 1-similarity (or edge

weight from a patient similarity network). The overall shortest path was defined as the mean pairwise shortest-path for a node set.

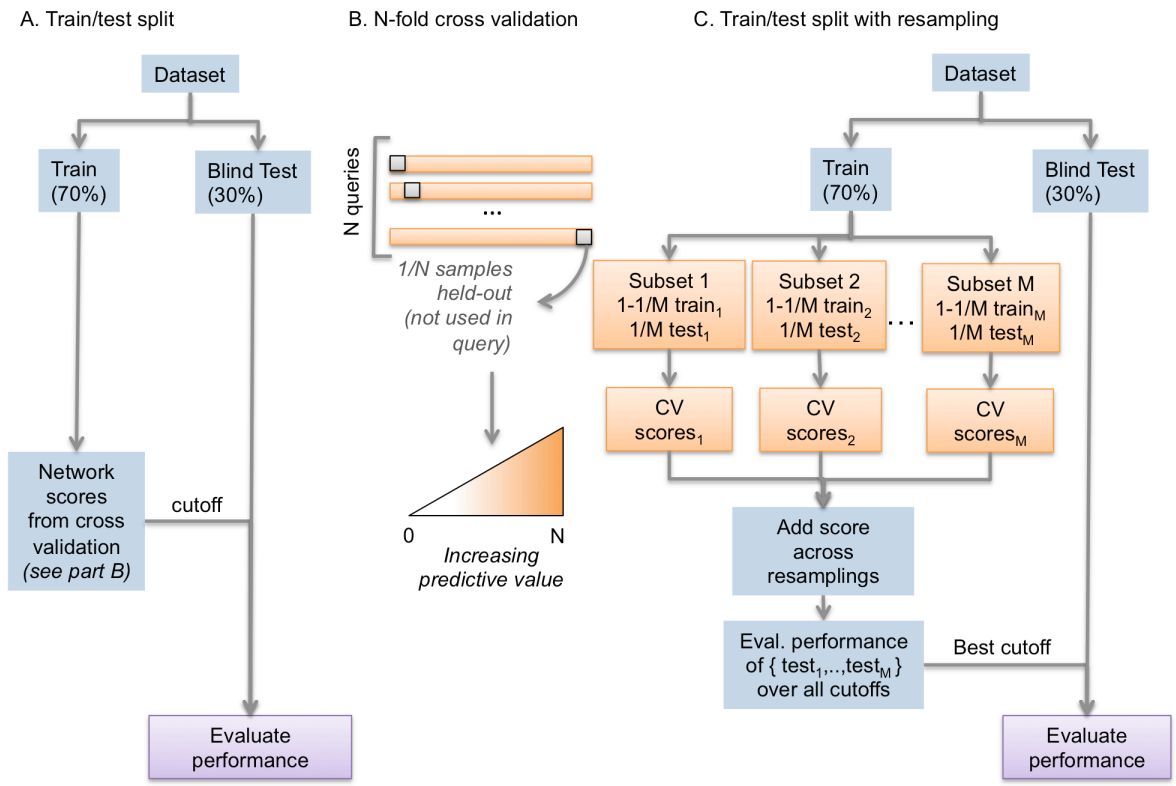
Supplementary Figures



Supplementary Figure 1. Conceptual overview of the GeneMANIA algorithm. GeneMANIA is a network-based recommender system that ranks all nodes in its database by similarity to an input query (or "positive" nodes). In the netDx application, the nodes are patients and the GeneMANIA database is comprised of a set of user-defined similarity networks derived from patient data (left). An example application is predicting Luminal A breast cancer type, by ranking all patient tumours by similarity to known Luminal A tumours. 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 2. Networks predictive of LumA status for netDx predictor with 3-way resampling of the training data. The predictor was built using gene expression from 348 primary breast tumours from the TCGA project. Each node represents a pathway (input network), and fill intensity increases with increasing netDx score. See supplementary methods for details on deriving the pathway enrichment map.



Supplementary Figure 3. Variations in predictor design to reduce overfitting

A. The data set is split into a training set and a "blind" test set. The netDx predictor is fit to the training set, and a score of predictive power is computed for each input network. Performance of the predictor is then evaluated on the test set, which is "blind" because it was not used to build the predictor. The performance is evaluated at various thresholds, to identify an optimal threshold.

B. In the netDx algorithm, cross validation is used to score the predictive value of an input network. For a given fold size (e.g. N=10 for 10-fold cross validation), the training set is resampled N times such that each sample is excluded from exactly one set; resampling occurs without replacement. Each resampled set serves as a GeneMANIA algorithm query, and networks identified as predictive for that query have their score incremented by one. Therefore increasing N correspondingly increases the range of network scores so that finer distinctions in predictive power may be made.

C. Another level of resampling can be introduced at the level of training samples provided for cross validation. After the initial test/train split, cross validation is computed for different random subsamples of the training set; similar to B, subsampling ensures that each sample is held out from exactly one set. The final score of a network is the cumulative score across the various iterations of feature selection. A predictor built with 3-way resampling (M=3) and 10-fold cross

validation results in network scores ranging from 0 to 30. Each cutoff is evaluated on the test data within each resampling, and that performance is averaged. The cutoff with the best mean performance is then used to validate the model on the blind test.

Supplementary Tables

Predictor design	Input size	Runtime, work-station	Runtime, laptop	Data
Simple predictor: no resampling (network score out of 10)	348 patients in total 3,423 nets, 1-15,454 edges (mean=7,121) Of these 1,622 are sparser CNV-based nets, 1-5,671 edges (mean=186)	40 min	1h 40 min	1Gb
Complex predictor: three-way resampling (network score out of 30)	348 patients in total 1801 nets, 8,332-9,680 edges (mean=8,789)	1h 28 min	*Not timed*	1.7Gb data (2.7Mb of result files)

Supplementary Table 1. Computational resources required to run the breast cancer Luminal A predictor and runtimes. The runtime shown includes loading of data, building of input networks, feature selection, ranking of test samples, and performance evaluation. For laptop timings, an early 2014 MacBook Air with 1.7GHz Intel Core i7 (4 cores) with 8GB RAM was used. For workstation timings, an 2.90GHz Intel Xeon CPU (8 cores) with 128GB RAM was used.

Supplementary References

1. Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80 (2004).
2. Romero, P. et al. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* **6**, R2 (2005).
3. Kandasamy, K. et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol* **11**, R3 (2010).
4. Croft, D. et al. The Reactome pathway knowledgebase. *Nucleic Acids Res* **42**, D472-7 (2014).
5. Fabregat, A. et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481-7 (2016).
6. Schaefer, C.F. et al. PID: the Pathway Interaction Database. *Nucleic Acids Res* **37**, D674-9 (2009).
7. 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).
8. Mi, H. et al. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* **33**, D284-8 (2005).
9. 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).
10. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**(2008).
11. Friedman J, H.T., Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1-22 (2010).
12. Liaw A., W.M. Classification and Regression by randomForest. *R News* **2**, 18-22 (2002).
13. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).
14. Csardi G., N.T. The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2006).