netDx: Patient classification using integrated patient similarity networks

Authors:
Shraddha Pai[1,5], Shirley Hui[1], Ruth Isserlin[1], Hussam Kaka[1], Gary D. Bader*[1,2,3,4]

Affiliations:
1. The Donnelly Centre, University of Toronto, Toronto, Canada
2. Department of Molecular Genetics, University of Toronto, Toronto, Canada
3. Department of Computer Science, University of Toronto, Toronto, Canada
4. The Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada
5. Affiliate Scientist, The Centre for Addiction and Mental Health, Toronto , Canada

* gary.bader@utoronto.ca

Abstract

Patient classification has widespread biomedical and clinical applications, including diagnosis, prognosis, disease subtyping and treatment response prediction. A general purpose and clinically relevant prediction algorithm should be accurate, generalizable, be able to integrate diverse data types (e.g. clinical, genomic, metabolomic, imaging), handle sparse data and be intuitive to interpret. We describe netDx, a supervised patient classification framework based on patient similarity networks that meets the above criteria. netDx models input data as patient networks and uses the GeneMANIA machine learning algorithm for network integration and feature selection. We demonstrate the utility of netDx by integrating gene expression and copy number variants to classify breast cancer tumours as being of the Luminal A class (N=348 tumours). In a simplified comparison using gene expression, netDx performed as well as or better than established state of the art machine learning methods, achieving a mean accuracy of 89% (2% s.d.) in classifying Luminal A. netDx uses pathway features to aid biological interpretability and results can be visualized as an integrated patient similarity network to aid clinical interpretation. Upon publication, netDx will be made publicly available via github.

Introduction

The goal of precision medicine is to build quantitative models that guide clinical decision-making by predicting disease risk and response to treatment using data measured for an individual. Within the next five years, several countries will have general-purpose cohort databases with 10,000 to >1 million patients, with linked genetics, electronic health records, metabolite status, and detailed clinical phenotyping; examples of projects underway include the UK BioBank[1], the US NIH Precision Medicine Initiative (www.whitehouse.gov/precision-medicine), and the Million Veteran Program (http://www.research.va.gov/MVP/). Additionally, specific human disease research projects are moving towards profiling of multiple data types at the population level, including genetic and genomic assays, brain imaging, behavioural testing and clinical history from integrated electronic medical records[2-4] (e.g. the Cancer Genome Atlas, http://cancergenome.nih.gov/). Computational methods to integrate these diverse patient data types for analysis and prediction will aid understanding of disease architecture and ideally provide actionable clinical guidance.

Statistical models that predict common disease risk are in routine clinical use in the fields of cardiovascular health, metabolic disorders, and certain cancers[5-8]. Most of these models were developed in the pre-genomic era, and use a combination of clinical history and metabolite state (e.g. PSA test[9] for prostate cancer or risk categories for diabetes[7]). Models that integrate genetic information are still rare; a notable exception is the BOADICEA model that predicts breast cancer risk by including medical and family history, and BRCA1/2 status[10,11]. Established clinical risk prediction models typically use generalized linear regression or survival analysis, in which all individual measures are presented as terms (or features) of a single equation. Standard methods of this type have limitations analyzing the large data from genomic assays. Machine learning methods can handle large data, but are often treated as black boxes that take substantial effort to understand how specific features

are useful for prediction. Many existing methods also do not natively handle missing data, and require the user to address this by data pruning or imputation. Additionally, each data type may capture a characteristic "view" of patient similarity through its correlation structure, with different data types separating patients into different classes. A simple combination of heterogeneous data sources into a single model ignores this data type specific correlation and may lose important patient similarity structure[12].

The patient similarity network framework can overcome many of these challenges and excels at integrating heterogeneous data, handling missing information, and model interpretability. In this framework, each input patient data feature is represented as a pairwise patient similarity network (PSN) (Figure 1). Each PSN node is an individual patient and an edge between two patients corresponds to pairwise similarity for a given feature. PSNs can be constructed based on any available data as long as a similarity measure is available (e.g. Pearson correlation for gene expression data). Patients that are highly similar to one another in one or more PSNs can be grouped (unsupervised classification/clustering/subtyping), and those of unknown status can be classified based on their relative similarity to patients with known labels (supervised classification). Patient similarity networks (PSN) have been used for unsupervised class discovery in cancer and type 2 diabetes[4,12]. We describe a PSN-based approach, called netDx, for supervised patient classification. Conceptually, this patient similarity-based classification is like that used in routine clinical diagnosis, which often involves a physician relating a patient to a mental database of similar patients they have seen. As demonstrated below, our netDx PSN framework is accurate, supports heterogeneous and missing data, and incorporates feature selection. Additionally, our use of biological pathway based features supports improved accuracy and generalization, aids interpretability of genome oriented patient data and identifies disease-altered physiological processes. To our knowledge, netDx is the first reported supervised patient network-based classifier.

Methods

The overall netDx workflow is shown in Figure 1.

Input data design. Each patient similarity network (PSN) is a feature, similar to a variable in a regression model (we use the terms "input networks" and "features" interchangeably). A PSN can be generated from any kind of patient data, so long as a measure of pairwise patient similarity can be defined. For example, gene expression similarity can be measured using Pearson correlation for the genes of interest, while patient age similarity could be measured by normalized age difference. A simple design is to define one similarity network per data type, such as a single network based on correlating the expression of all genes in the human genome, or a network based on similarity of responses to a multi-question clinical questionnaire. Using unit measurements is more interpretable as individual data types (e.g. individual gene expression levels or questionnaire answers) can be identified as important for classification. However, this approach can easily lead to too many features generated (e.g. millions of SNPs), which increases risk of overfitting and leads to large computational resource requirements. To address this for 'omics data (e.g. genomics, proteomics, metabolomics), we group measurements into biological pathways, which we assume capture relevant aspects of cellular and physiological processes underlying disease

and normal phenotypes. This biological process-based design generates ~1,000 networks from gene expression data, with one network per pathway. Feature selection identifies pathways that have good discriminative value, which helps improve our understanding of the underlying mechanisms. This idea can be extended to non-genomic data; functional brain imaging data could be grouped by brain regions co-activated during a task of interest, while responses to behavioural assessments could be grouped by measurement of the same latent variable.

We have considered two broad categories of patient network types. The first is a dense network where the same measures are obtained on all or almost all patients (Figure 1A). Examples include standardized assays or tests, such as gene expression assays that provide gene-level measures. The second is a sparse network in which features are rare, thus there is little opportunity to relate patients based on them. Examples include rare genomic events such as *de novo* DNA copy number variants (Figure 1B). This sparseness is addressed by aggregating sparse counts into larger counts by grouping units, such as grouping a set of copy number variants affecting a set of genes in a biological pathway.

Network integration and similarity-based ranking. netDx uses the GeneMANIA[13,14] multiple association network integration algorithm to integrate all input networks into a single composite patient similarity (or association) network (Figure 1C; Supplementary Figure 1). When provided with a set of query patients – for example, all patients with a tumour of a given class – GeneMANIA first reduces redundant networks, and then weights each network based on how informative it is for discriminating the set of query patients from all other patients. A weighted linear combination of all networks is used to create the composite network. GeneMANIA then uses label propagation on this integrated network to rank all patients (network nodes) by similarity to the query set (Supplementary Figure 1). Top ranked patients (according to a threshold) can be classified as being part of the query patient set. For example, if the query patients are all of a specific class, top ranked patients can be classified as part of this class.

Feature selection to identify predictive networks. Feature selection identifies the input networks with the highest general predictive power. For a given query, the network weight computed by GeneMANIA is a measure of the value of that network in the classification problem. netDx is trained based on samples from the class of interest; this is achieved by a set of GeneMANIA queries using a cross-validation based approach designed to reduce the chances of over-fitting (Figure 1C, top panel). The score for a given network is the frequency with which it is assigned a positive weight in the cross-validation procedure. The classifier's sensitivity and specificity can then be controlled by thresholding this score; a network with a higher score achieves greater specificity and lower sensitivity. This feature selection step results in a subset of input networks that can be integrated by GeneMANIA to produce a predictor for the patient class of interest. Data can be optionally resampled during training to improve test classifier generalization, though this is computationally intensive.

Class prediction using selected features. For every class label, an integrated network database is built using networks with high feature selection scores for that class. This

database must comprise all training and test samples in the data set. For each class, a GeneMANIA query is run against the network database, using training samples for that class as the query. This step results in a class-specific ranking for all test patients (Figure 1C, bottom panel). The patient is then assigned to the class with the highest normalized rank. This process is equivalent to labelling the patient with the class to which the patient is most similar.

Integrating heterogeneous data types and feature selection: Classifying a breast tumour as "Luminal A" class

We used netDx as a binary classifier for the Luminal A breast cancer subtype (or class) and performed feature selection. Luminal A breast tumours are low-grade, with a majority (~87%) being clinically positive for estrogen receptor and negative for HER2[15]. These features increase the chances that this tumour type responds to hormone therapy and chemotherapy. Gene expression and CNV coordinates were downloaded for 348 primary breast tumours from TCGA (Level 3 data; https://tcga-data.nci.nih.gov/docs/publications/brca_2012/)[15]. The workflow for building a classifier is shown in Figure 2A-C. Data were split 70:30 into training and test groups respectively; for feature selection, input networks were constructed using only training samples (N=104 LumA, 130 other). Patient similarity was defined at the level of cellular pathways collected from curated pathway databases (Supplementary Methods). Individual networks were based on similarity by gene expression data or by shared CNV occurrence in genes of each pathway (Figure 2A). Networks based on gene expression used pathway-level Pearson correlation as a measure of similarity (N=1,801 networks); networks were sparsified to retain only the strongest connections (Supplementary Methods). CNV-based networks used a binary measure of shared overlap of a pathway (N=1,622 pathways). Unlike the networks derived from gene expression, these networks do not each contain all the patients in the data set; rather, each network contains only those patients with CNVs overlapping genes of the same pathway. Feature selection was separately performed for each class (Figure 2B). In each case, ten-fold cross validation was used to score the predictive value of networks for the class of interest (positive labels). For each fold of cross validation, GeneMANIA was run with a different 9/10th of positive training samples as each input query. We ensured that each sample was part of the non-query set exactly once. Each time a network was returned in the GeneMANIA network ranking table, its score was increased by one; a network could therefore score a maximum of ten for the ten folds of cross validation. In this design, feature selected networks for a given class were those that score >= 9/10.

Out of a total of 3,423 networks (1,801 based on gene expression and 1,622 on CNV), 57 networks were feature selected for the LumA class; eight of these were CNV-based networks. Feature selected networks included previously reported themes of cell cycle regulation[16,17], mitotic spindle checkpoint[16,18], DNA damage repair[18], and pyrimidine metabolism[17] (Figure 2D). Following feature selection, a patient similarity network database was constructed for each class (Figure 2B); each database contained all samples in the dataset. A single query was run against each database, using all training samples for the corresponding class as query. This resulted in a class-specific ranking for each sample and test samples were assigned to the class with the higher rank (Figure 2C). This predictor

has a 92% accuracy in classifying LumA tumours, with a positive predictive value of 84% (Table 1); the overall predictor accuracy is 88.8%. As a comparison, Paquet and Hallett[19] recently report an accuracy of 86.6% in Luminal A classification with a predictor that uses relative gene expression of pairs of genes within the same sample. Luminal A samples are closer to each other and separate from other samples in the resulting integrated network (Figure 2E); the average weighted shortest path for LumA-LumA nodes is 0.138 (SD=0.051), which is shorter than the LumA-other distance of 0.266 (SD 0.075) (other-other=0.162 (SD=0.055); all-all=0.157 (SD=0.047)).

Comparison to other model-building methods

Using the TCGA breast cancer data, we compared the performance of netDx to two established machine-learning algorithms, elastic net regression and random forests (Supplementary Methods). To simplify the comparison, only gene expression data were used. For comparison to pathway-level networks in netDx, features for the other two methods were tested in two ways: the model was either limited to genes present in pathways (N=7,948 features) or each feature was defined as the mean expression value of all genes in a pathway (N=1,802 features). To reduce overfitting, all methods used three-fold resampling (see discussion of predictor design strategies); biological themes identified through this approach are similar to those found by the simpler predictor version presented above (Supplementary Figure 2). netDx performs at par with these methods (504 train/test splits; mean accuracy 89+/- 2%, range 81-95%; mean PPV 84+/-3%SD; range 75-94%) (Table 2). Thus, netDx matches the performance of state of the art methods, is easy to use because of provided software and examples, and automatically provides an interpretable model based on patient similarities and biological pathways.

Computational requirements

Supplementary Table 1 shows computational resources and timing for running the breast cancer predictor on a workstation and a laptop. Running a version with no resampling – one round of 10-fold cross validation – takes 40 minutes on a workstation (Intel Xeon CPU 2.9GHz, 8 cores, 128 Gb RAM). Running the predictor with three-way resampling takes under two hours. By default, all intermediate data is stored, to permit a detailed examination of the process of classifier creation. The memory requirement for running netDx scales with the number of GeneMANIA algorithm queries that run in parallel. Despite sparsification, a single GeneMANIA algorithm query with 1,801 gene expression-based networks requires ~7Gb RAM. In practice, we have found that a compute node running 8 queries in parallel requires between 30 and 50Gb RAM. These memory requirements preclude running predictors with networks with gene-level features (~17K dense networks) in typical cases. Thus, some feature selection is necessary to limit input networks to potentially informative genes; this list may include differentially-expressed genes, most variable genes, or "eigengenes"/hub genes that represent a correlated set of genes[20]. On a modern laptop, running the simplified version of the breast cancer predictor takes 1 hour 40 minutes (Supplementary Table 1). Thus running netDx on a laptop is feasible (see discussion below).

Discussion

We present a novel machine learning application for clinical sample classification based on patient similarity networks. The method performs at a state of the art level on test data, but is more intuitive and interpretable due to the extensive use of feature selection and biological pathway features for genomic data.

While the applications we describe here use genomic data as input, networks can be derived from other types of patient data including clinical, genetic, and brain imaging. A requirement is the definition of a suitable similarity measure from a data type of interest. Optimal feature design is an open problem and is likely to depend on the data, and classification task. New patient data types or similarity measures can be evaluated before generating patient networks in the netDx framework. For example, exploratory data analysis, such as unsupervised clustering approaches, may be used to ascertain if classes separate when using a candidate measure of similarity. Domain expertise must guide the predictor and feature selection building process by assessing the relative merits of various similarity measures and by identifying data groupings grounded in prior knowledge; examples include pathways to group genes, brain regions co-activated in a functional imaging assay, and questionnaire scores that measure the same physiological variable or cognitive ability. Features with clear clinical or mechanistic interpretation are more likely to inform the understanding of the classes, in addition to providing predictive value. However, given enough compute power, it is also possible to use a "brute force" approach to assess the predictive value of all possible candidate networks. Another consideration is the use of appropriate controls to validate initial findings, to ensure that the structure of networks derived from novel data types – for example, sparse networks – does not adversely affect the algorithm. For instance, injecting simulated random similarity networks as input features would serve as a negative control, as such networks should not be feature selected.

As with any machine learning method, principles of good predictor design are applicable when using netDx. A common problem in predictor building is overfitting, where a predictor fits irrelevant variation (or "noise") in the training data and is therefore less generalizable to new input. The ideal scenario, in which a dataset is large enough to accommodate a training set representative of the population being studied, and which can provide sufficiently large sample size for validation, is currently rare in patient genomics. Instead, strategies such as partitioning of data into training and test samples, and resampling are used to mitigate overfitting (Supplementary Figure 3). At a minimum, the sample size should be large enough to be partitioned into a training set and a "blind" test set (Supplementary Figure 3A). Feature selection should be performed using only the training set; the test samples are then used to validate the predictor. In the applications presented here, we use a 70:30 split for training and test data; we find this split provides a reasonable balance for providing sufficient samples to build the predictor and separately to assess performance. A second strategy is resampling, where the network score is computed on multiple subsets of training data, which are then aggregated to provide the final set of scores. The netDx algorithm uses resampling in the cross validation used to assign network scores (Figure 1C top panel, Supplementary Figure 3B). In N-fold cross validation, N

GeneMANIA algorithm queries are run on different training data samplings; network scores are the frequency with which a given network is identified as predictive. Resampling ensures that some samples are not sampled more often than others, by excluding each sample from exactly one iteration. While here we use 10-fold cross validation (i.e. N=10), N can be increased or decreased depending on training set size; the goal is to ensure that a reasonable number of samples are included in a given query, so that the contribution of any single query to the network score is non-trivial. Finally, datasets with large sample sizes can build a more robust predictor by running this inner cross-validation loop for multiple resamples of the overall training set (Supplementary Figure 3C). When added, cross validation scores from different resamplings provide a more fine-grained measure of predictive power. They promote networks that score moderately in any given resampling but that are robust to resampling, over networks with highly fluctuating predictive power. This approach may also be useful for sparse patient networks, such as those based on CNV deletions and duplications, which increase the chance of overfitting.

We plan to further develop netDx to include other data types – such as clinical, genetic, brain imaging – and to optimize similarity measure computation (e.g. using NBS[21]). We also propose storing and sharing of patient similarity networks, useful as features for netDx and other PSN methods in the NDEx network exchange system[22]. Future versions should automatically handle covariates. We currently propose considering covariates by use of a logistic regression, subsequent to netDx prediction, that includes terms for netDx predictor values and for covariates in a single equation.

We propose that netDx will be clinically useful because of its interpretive value. In the future, clinical researchers could run the predictor via a web-based interface using their patient data, while the actual computation could be run on compute servers that can connect to data from patient databases, such as electronic medical records. Results could be provided in the form of reports that indicate model performance, and that plot the organization of high-ranking features, similar to the enrichment map in Figure 2D. While netDx was developed for biomedical data, the algorithm is generic enough to be applicable to any domain requiring supervised classification by integration of multiple data types; examples of possible applications include agriculture (networks of similarity between crop strains) and microbial pathogenicity (similarity between microbes).

The netDx method is implemented as an open-source R software package and will be made available upon publication at http://netdx.org, with worked examples.

References

1. Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
2. Calkins, M.E. et al. The Philadelphia Neurodevelopmental Cohort: constructing a deep phenotyping collaborative. *J Child Psychol Psychiatry* **56**, 1356-69 (2015).
3. Collins, F.S. & Varmus, H. A new initiative on precision medicine. *N Engl J Med* **372**, 793-5 (2015).
4. Li, L. et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med* **7**, 311ra174 (2015).
5. Wilson, P.W. et al. Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837-47 (1998).
6. Lee, A.J. et al. BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. *Br J Cancer* **110**, 535-45 (2014).
7. Schmidt, M.I. et al. Identifying individuals at high risk for diabetes: The Atherosclerosis Risk in Communities study. *Diabetes Care* **28**, 2013-8 (2005).
8. Gail, M.H., Anderson, W.F., Garcia-Closas, M. & Sherman, M.E. Absolute risk models for subtypes of breast cancer. *J Natl Cancer Inst* **99**, 1657-9 (2007).
9. van Vugt, H.A. et al. Compliance with biopsy recommendations of a prostate cancer risk calculator. *BJU Int* **109**, 1480-8 (2012).
10. Antoniou, A.C., Pharoah, P.P., Smith, P. & Easton, D.F. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *Br J Cancer* **91**, 1580-90 (2004).
11. Mavaddat, N., Rebbeck, T.R., Lakhani, S.R., Easton, D.F. & Antoniou, A.C. Incorporating tumour pathology information into breast cancer risk prediction algorithms. *Breast Cancer Res* **12**, R28 (2010).
12. Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* **11**, 333-7 (2014).
13. 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).
14. Zuberi, K. et al. GeneMANIA prediction server 2013 update. *Nucleic Acids Res* **41**, W115-22 (2013).
15. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).
16. Ciriello, G. et al. The molecular diversity of Luminal A breast tumors. *Breast Cancer Res Treat* **141**, 409-20 (2013).
17. Tian, F., Wang, Y., Seiler, M. & Hu, Z. Functional characterization of breast cancer using pathway profiles. *BMC Med Genomics* **7**, 45 (2014).
18. Peng, G. et al. Genome-wide transcriptome profiling of homologous recombination DNA repair. *Nat Commun* **5**, 3361 (2014).
19. Paquet, E.R. & Hallett, M.T. Absolute assignment of breast cancer intrinsic molecular subtype. *J Natl Cancer Inst* **107**, 357 (2015).
20. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

21. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat Methods* **10**, 1108-15 (2013).
22. Pratt, D. et al. NDEx, the Network Data Exchange. *Cell Syst* **1**, 302-305 (2015).

Acknowledgements
We thank Quaid Morris and Daniele Merico for thoughtful discussions and feedback during development of this work.

Author contributions
All authors contributed to netDx method development. S.P. analyzed the datasets in this manuscript and wrote the netDx software package. S.H., H.K and R.I. developed initial versions of netDx. S.P. and G.D.B. wrote the paper.

Competing financial interests
The authors declare no competing financial interests.

Materials and Correspondence
Contact S.P. for inquiries regarding the netDx software package (shraddha.pai@utoronto.ca). Contact G.D.B. for all general inquiries (gary.bader@utoronto.ca)

Figures

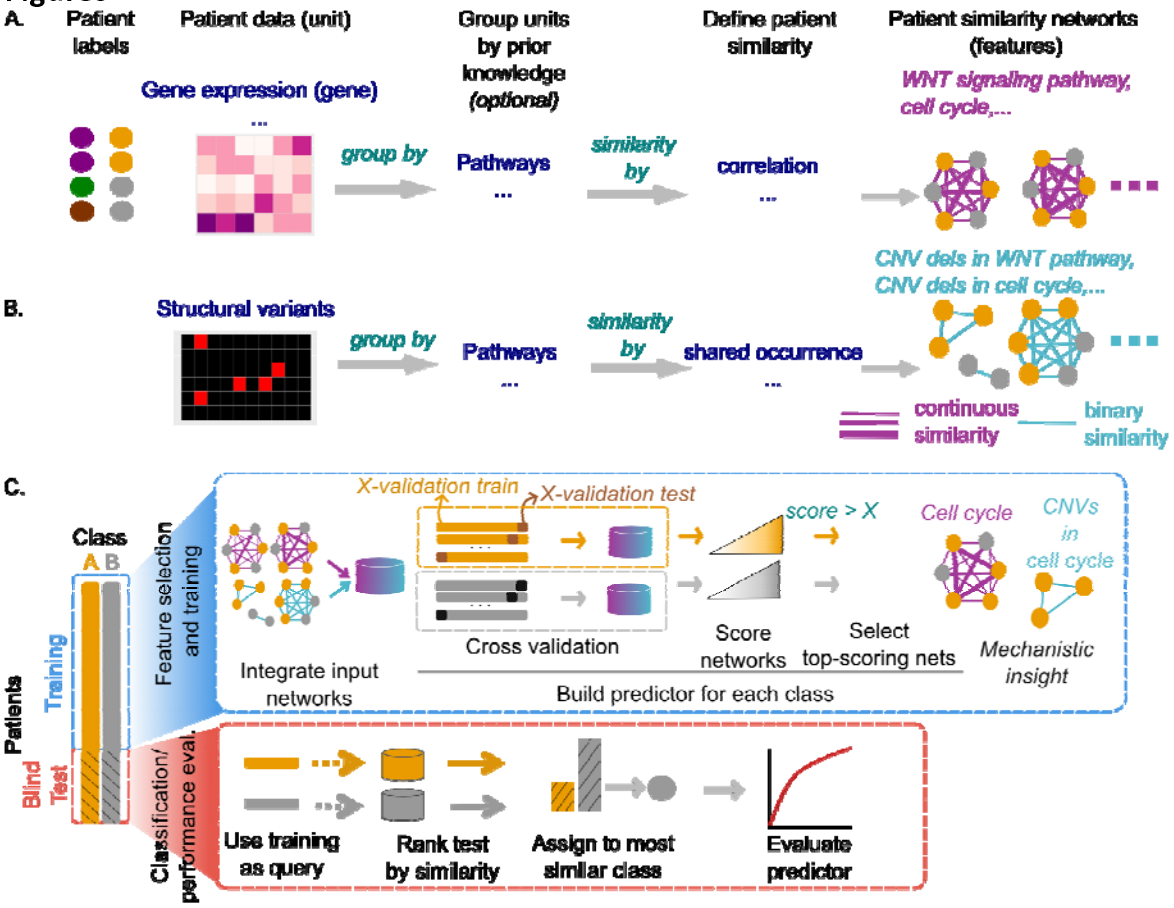

Figure 1. The netDx workflow.

A. netDx converts patient data into a set of patient similarity networks (PSN), with patients as nodes and a user-provided similarity measure as weighted edges. The user can optionally group data using prior knowledge. The method generates one network for each input data or grouping.

B. For sparse data such as genomic structural variants, similarity may be defined as the shared occurrence within a set of genomic regions, such as a pathway. This aggregates sparse counts into larger counts by group. The resulting patient similarity networks are binary and include varying numbers of patients. Steps A and B can be repeated for an arbitrary number of heterogeneous data sources.

C. netDx predictor workflow. Patient data is partitioned into a training and a blind test set. Feature selection (top) is performed once per class. The GeneMANIA algorithm is used to integrate input patient networks and to rank patients by similarity to a query input. For each class, networks are scored by the frequency with which they are identified as being predictive by the GeneMANIA algorithm. A threshold is then used to subset feature-selected networks. Such networks can provide mechanistic insight into the class. Classification of test samples (bottom) serves to evaluate the predictor. Prediction uses a PSN database containing feature-selected networks that includes both training and test samples. Test patients are ranked for class similarity to training samples, and patients are assigned to the highest-ranking class.

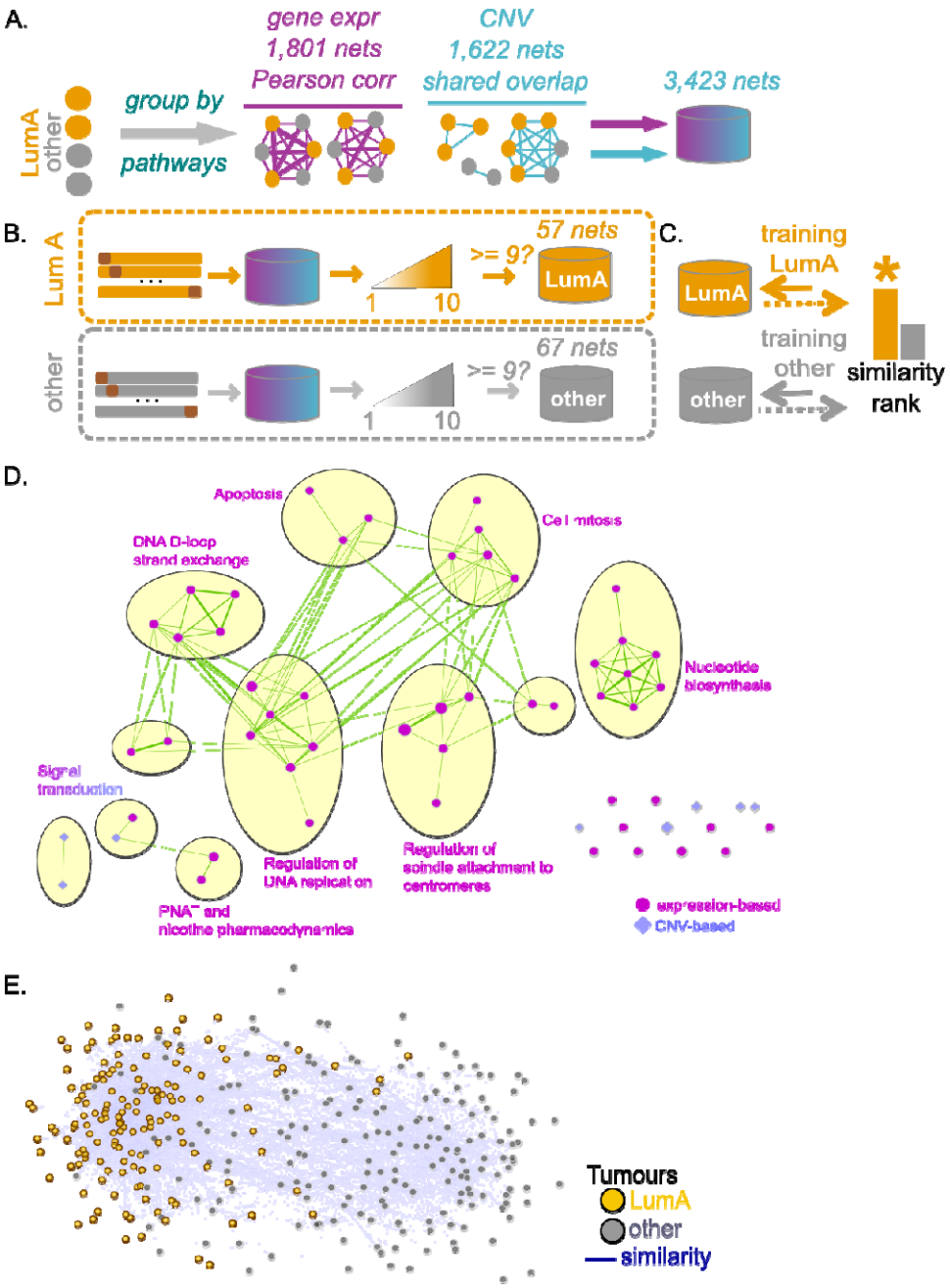

Figure 2. Predicting Luminal A breast tumour status by integrating gene expression and copy number variant data.

A. Feature selection uses samples in the training set (70% per class). Similarity networks are constructed for each data type, and then integrated into a single PSN database.

B. Feature selection is conducted separately for the positive (LumA) and negative (other) class. Networks with a score of nine or ten out of ten pass feature selection. Following this step, a single PSN database is constructed per class; this database contains both training and test samples.

C. Patient classification occurs by assigning a patient to the highest-ranking class.

D. Networks that pass feature selection for the "LumA" class show themes related to DNA repair and cell cycle regulation. A pathway enrichment map is displayed, where nodes indicate pathways and edges connect pathways with overlapping genes. Node colour and cluster label indicates type of member networks, and large nodes correspond to larger gene-sets. See supplementary methods for details.
E. Patient similarity network with integrated networks scoring ten using netDx. Nodes indicate individual Luminal A (orange) or other (grey) tumours; edges show maximum patient similarity between two nodes. For visualization, only the top 50 edges per node were retained.

Tables

Class	# total	# train	# selected networks (# CNV)	accuracy	PPV
LumA	154	103	57 (8)	47/51 (92%)	47/56 (84%)
other	194	129	67 (25)	56/65 (86%)	56/60 (93%)

Table 1. Statistics for netDx binary classification of breast tumour as LumA or not, by integration of gene expression and CNV data.

	Limit to genes in pathway (N=7,948 features)		Average expression in pathway (N=1,802 features)	
	Accuracy	Positive Predictive Value	Accuracy	Positive Predictive Value
Elastic net	90	84	91	86
Random forest	92	92	90	84
netDx: No resampling (score out of 10)			88	85
netDx: three-way resampling (score out of 30)			91.5	88
netDx: three-way resampling (score out of 30; 504 train/test splits)			**Mean 88.9% (81-95%)**	**Mean 84% (75-94%)**

Table 2. Comparison of netDx to elastic net and random forests machine learning methods. The task involved binary classification of breast tumours as being of class "Luminal A" or not, using only gene expression data. Features for other methods were either limited to genes in one or more pathways (green), or were at the level of individual pathways (blue; mean gene expression). For comparison, netDx was run either with a single round of 10-fold cross validation (no resampling), with three rounds of resampling and a particular train/test split, or with three rounds of resampling and 504 train/test splits.