Simpati: patient classifier identifies signature pathways based on similarity networks for the disease prediction

Luca Giudice [a] (corresponding author: luca.giudice@univr.it)

[a] Department of Computer Science, University of Verona, Strada le Grazie 15, 37134, Verona, Italy

ABSTRACT

BACKGROUND

Pathway-based patient classification is a supervised learning task which supports the decision-making process of human experts in biomedical applications providing signature pathways associated to a patient class characterized by a specific clinical outcome. The task can potentially imply to simulate the human way of thinking in predicting outcomes by pathways, decipher hidden multivariate relationships between the patient's characteristics and their class, and provide more information than a probability value. However, classifiers are rarely integrated into a routine bioinformatics analysis of high-dimensional biological data because they require a nontrivial hyper-parameter tuning, are difficult to interpret and lack in providing new insights. There is the need of new classifiers which can provide novel perspectives about pathways, be easy to apply with different biological omics and produce new data enabling a further analysis of the patients.

RESULTS

We propose Simpati, a pathway-based patient classifier which combines the concepts of network-based propagation, patient similarity network, cohesive subgroup detection and pathway enrichment. It exploits a propagation algorithm to classify both dense, sparse, and non-homogenous data. It organizes patient's biological features in pathways represented by patient similarity networks for being interpretable, handling missing data and preserving the patient privacy. A network represents patients as nodes and a novel similarity determines how much every pair act co-ordinately in a pathway. Simpati detects signature biological processes based on how much the topological properties of the related networks separate the patient classes. In this step, it includes a novel cohesive subgroup detection algorithm to handle patients not showing the same pathway activity as the other class members. An unknown patient is classified based on how much is similar with known ones. Simpati outperforms state-of-art classifiers on five cancer datasets described with two biological omics, classifies well sparse data, identifies more relevant pathways associated to the patient's diseases than the competitors and has the lowest computational requirements between pathway-based classifiers.

CONCLUSION

Simpati can serve as generic-purpose pathway-based classifier of patient classes. It learns signature pathways that are enriched in a class with respect the reference. It tests the pathways based on how much they can correctly predict the class of unknown patients. It detects signature pathways divided into up or downinvolved based on how much the members of one class are similar. Upinvolved if the activity of the molecules in a pathway is high and similar between the patients considering the signaling cascades derived from the activity of also all the other molecules. The signature pathways are then related to the patients in study with the integration of the Disgnet and Human Protein Atlas databases. Together with the results of a unique pathway analysis, Simpati provides in a vectorial and graphical format the patient similarity networks associated to the biological processes. This allows to find the patient which is core or outlier with respect the other members due to specific molecules and pathways. We provide an R implementation which starts Simpati with one function, a GUI interface for the navigation of the

patient's propagated profiles and a function which offers an ad-hoc visualization of patient similarity networks. The software is available at: https://github.com/LucaGiudice/Simpati

KEYWORDS
Pathway-based classification; Graph theory; Network-based propagation; Patient similarity network; Subgroup cohesive algorithm; Cancer Stage;

1.INTRODUCTION

High-throughput biological data provide valuable information to clinicians for the prognosis and treatment response of patients. They offer quantitative and qualitative evidences to biomedical scientists for developing a study or confirming wet-lab results [1–3]. Pathway-based analysis is a technique for mining these data. It provides an intuitive and comprehensive understanding of the molecular mechanisms related to the patients [4,5]. The pathway space is more robust to noise than the single feature level, summarizes the information of multiple patient's features into the pathway activity (inhibited or activated), reduces the model complexity and maintains predictive accuracy in the face of uncontrolled variation [6–9]. These motivations boosted the development of enrichment tools for the pathway analysis but not of machine learning algorithms. The latter are neither considered in reviews [10–13] nor in bioinformatics best practises [14–16]. Fabris et al. [17] detailed the drawbacks of a supervised classification approach. It lacks a formal statistical basis, is computationally expensive, includes a not trivial hyper-parameter setting and does not well handle neither imbalances classes nor structured feature types as the biological pathways.

Few attempts have been made in this direction. In 2010, Pang et al. [18] proposed a bivariate node-splitting random forest integrating pathways for the survival analysis of cancer patients in microarray studies. In 2018, Hao et al. [19] proposed the first generic-purpose pathway-based deep neural network for the prediction of Glioblastoma patients. The method builds a network model by leveraging prior biological knowledge of pathway databases and predicts considering hierarchical nonlinear relationships between the biological processes and the patient classes in comparison. However, the method requires a non-trivial tuning of hyper-parameters which are difficult to interpret for bioinformaticians or computational biologists without a background in deep learning, demands high computational resources, does not provide a graphical representation to explain why specific pathways have been selected and includes in the results only the classification performances.

In the same year, Pai et al. officially introduces the emerging patient similarity network (PSN) paradigm [20] for the precision medicine. In a PSN, each node is an individual patient and an edge between two patients corresponds to pairwise similarity for a given datum (gender, height, gene expression …). The paradigm brings many advantages. Analysing the similarities to gather new information is conceptually intuitive. A PSN can lead to the identification of patient subgroups or the prediction of a patient's class. Similarity networks can represent any datum, naturally handle missing and heterogenous data, have a history of successes in gene and protein function prediction [21–23] and can preserve the patient privacy by being shared in place of the sensitive raw information (topic which is growing in concern) [24–26].

Pai and Giudice et al. propose netDx [27,28] as patient classifier based on the PSN paradigm. Any available datum (e.g., age, gender, gene expression, …) is converted into a PSN. The method proceeds by performing GeneMania on the similarity networks. GeneMANIA [23] scores each input PSN based on how well it classifies an input set of patients known to be in the same class (i.e., training set). A linear combination of the best PSNs is used to create a composite network on which the unknown patients (i.e., testing patients) are classified based on their similarity with the training ones. Despite researchers' efforts for standing up to the challenge, the method does not accept data in matrix format, requires to define multiple functions in order to set up the model, does not provide a graphical representation of the PSNs used to predict, does not give access to the data processed during the workflow, depends by the quality of the user-selected similarity measure, requires multiple hyper-parameters to manually tune, demands high computational

resources and includes in the results only the classification performances together with the pathway names.

There is the need of new classifiers able to get the benefits of both the methodologies: classification and enrichment. As defined by Fabris et al., from the enrichment side the new method should be computationally light, easy to understand, provide more information than the probability value and not requiring assumptions to satisfy. While, from the classification side, it should be non-parametric, interpretable, consider multivariate interactions between features and patient classes, not requiring hyper-parameters difficult to set by the final user and cope well with both high-class imbalance and structured feature types.

We want to stand up to the challenge by proposing a pathway-based classifier called Simpati. Our main contributions with this method include: 1) the combination of graph-theory concepts as network propagation, cohesive subgroup detection and graph topology with machine learning to handle different biological omics, unbalanced patient classes and outliers, 2) a novel patient similarity measure adapt for any biological data type, 3) a novel concept of enriched pathways, 4) a user-friendly implementation, 5) an integrated prioritization system for the biomarker selection, 6) an integrated literature-based enrichment of the pathways, 7) explorable results which can lead to further findings, 8) two visualization tools for the patient analysis

2. METHODS

2.1 OVERVIEW

Simpati considers the patient's biological profiles (e.g., genes per patients) divided into classes based on a clinical information (e.g., cases versus controls). It prepares the profiles singularly applying guilty-by-association approach to determine how much each biological feature (e.g., gene) is associated and involved with the other ones and so to the overall profile. Higher is the guilty score and more the feature is involved in the patient's biology. Simpati proceeds by building a pathway-specific patient similarity network (psPSN). It determines how much each pair of patients is similarly involved in the pathway. If the members of one class are more similar (i.e., stronger intra-similarities) than the opposite patients and the two classes are not similar (i.e., weak inter-similarities), then Simpati recognizes the psPSN as signature. If the classes are likely to contain outlier patients (i.e., patients not showing the same pathway activity as the rest of the class), then Simpati performs a filtering to keep only the most representative members of each class and re-test the psPSN for being signature. Unknown patients are classified in the best pathways based on their similarities with known patients and on how much they fit in the representative subgroups of the classes (i.e., more they are similar to the representatives of a class and more they fit). As results, Simpati provides the classes of the unknown patients, the tested statistically significant signature pathways divided into up and down involved (new pathway activity paradigm based on similarity of guilty scores), the patient similarity networks in vectorial and graphical format to allow further analysis of the subjects in studies using graph-theory methods, the guilty scores associated to the biological single-level features and all the data produced during the workflow.

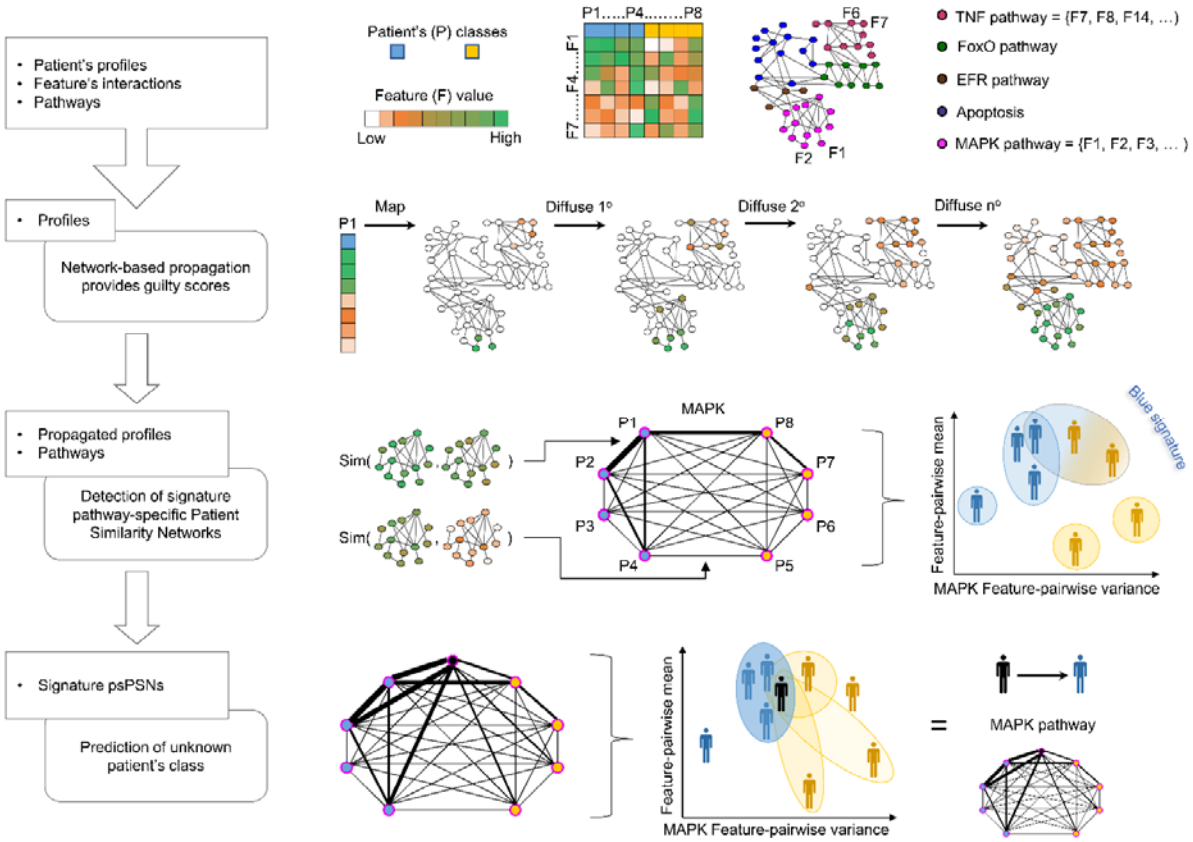


Fig.1. Workflow of Simpati. Patient profiles are divided in two classes and are described by biological features. A feature-feature interaction network together with pathways are further input data required by the software (e.g., gene-gene interaction network and KEGG pathways). All profiles are individually propagated over the network. The profile's values are replaced by scores that reflect the feature's starting information and interactions. Simpati proceeds by creating a patient similarity network for each pathway (psPSN). The pairwise similarity evaluates how much two patients have a similar pathway activity. It evaluates how much the features between two patients are close and high in term of propagation values. Two patients that act on a pathway with the same features and same expression values get the maximum similarity. In the figure, more an interaction is thick and more the two corresponding patients are similar. The psPSN is recognised as signature if one class is cohesive, one is sparse, and the two classes are not similar. In case of signature pathway, an unknown patient is classified based on how much is like the other patients.

2.2 DATA PREPARATION

Simpati works with patient's biological profiles (e.g. gene expression profiles), the classes of the patients (e.g. cases and controls), a list of pathways and an interaction network (e.g. gene-gene interaction network). Simpati is designed to handle multiple biological omics but requires that the type of biological feature (e.g. gene) describing the patients is the same one that composes the pathways and the network which models how the features interact or are associated. In this study, we tested Simpati in the classification of Early versus Late cancer stage patients. In fact, identifying the cancer mechanisms which drive the tumor from early to late stages is challenging [29–31] but it can improve the early cancer diagnosis, lead to develop more precise therapeutic strategies and increase the survival rates [32]. A late-stage cancer spreads to nearby lymph nodes and other organs, the survival rate decreases due to the necessity of more advanced and risky treatment strategies. While, early localized stages are easier to treat and have better survival rates [31,33–35]. For setting up this biological and pathway-based classification challenge, we collected data about Liver hepatocellular carcinoma (LIHC), Stomach adenocarcinoma (STAD), Kidney renal clear cell carcinoma (KIRC), Bladder Urothelial Carcinoma (BLCA), Lung squamous cell carcinoma (LUSC) and Esophageal carcinoma (ESCA) cancers from The Cancer Genome Atlas (TCGA)

using the R packages curatedTCGAData [36] and TCGAutils [37]. We kept only the patients having RNA sequencing (RNAseq) data, somatic mutation data and the histological type as clinical information. We added a new information based on the pathological stage attribute. We applied a binarization and labelled the stage I and II in Early, while the stage III and IV in Late based on the tumor/node/metastasis (TNM) system [38–41]. We proceeded with preparing the biological omics. We followed the workflow defined by Law et al. [42] for the RNAseq. Genes not expressed at a biologically meaningful level have been filtered out to increase the reliability of the mean-variance relationship. We removed the differences between samples due to the depth of sequencing and normalised the data using the trimmed mean of M-values (TMM) [43] method. While somatic mutation data have been converted into a binary data type, where a value equal to one indicates a mutated gene in a patient and zero otherwise. We ended up with two biological omics for five datasets with 14 LIHC (7 Early, 7 Late), 21 STAD (8 Early, 13 Late), 37 KIRC (24 Early, 13 Late), 45 BLCA (8 Early, 37 Late), 75 LUSC (60 Early, 13 Late) and 152 ESCA (91 Early, 61 Late) patients. The first four datasets to simulate wet-lab routine studies and the last two to have more precise classification performances [44]. We then collected the pathways and created the biological interaction network. We retrieved the biological processes from the major databases MSigDB [45], GO [46] and Kegg [47], while we used Biogrid [48] to model the biological feature's interactions. A node represents a gene, and the edges are experimental and manually curated gene-gene interactions (GGi) (564,325 interactions and 26,433 genes).

Formally, given a set of features $FEA = \{F_1, F_2, F_3, \dots, F_A\}$ with $A \in \mathbb{N}$ and of patient's profiles $PAT = \{P_1, P_2, P_3, \dots, P_B\}$ with $B \in \mathbb{N}$ where each element is a vector of feature's values, Simpati requires the concatenation of the patient's profiles in a matrix $M: AxB = (m_{a,b})$ where $m_{a,b}$ is equal to the value of F_a in patient P_b. The set of pathways $PATH = \{PH_1, PH_2, PH_3 \dots, PH_C\}$ where each element is defined by a finite set not exclusive of features (e.g., $PH_1 = \{F_4. F_5, F_6\}$). The biological network is defined as an undirected unweighted graph $GGi = (V, E)$ where $V = \{v_1, v_2, \dots, v_N\}$ is a set of vertices (aka nodes) representing features $f: FEA \rightarrow V$ and $E = \{(v_i, v_j) | v_i, v_j \in V\}$ is a set of pairs of vertices indicating edges (aka interactions, associations or relations) between features.

The adjacency matrix of GGi is the symmetric matrix $W: N \times N = (w_{i,j})$ defined by:

$$w_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Finally, the patient's classes are two vectors containing the indexes of the profiles belonging to them: $\overline{EARLY} = (1,2,3, \dots, D)$ and $\overline{LATE} = (D+1, \dots, B)$ with $D <= B$.

2.3 NETWORK-BASED PROPAGATION

Simpati starts with protecting the privacy of the patients and enhancing the advantage of using the patient similarity network paradigm in the workflow. It converts the patient's original information in anonymous labels. It then gives to the user the possibility to share data and results with or without the map. When the patient privacy is preserved, Simpati transforms the patient's biological profiles using a network-based propagation algorithm. Each feature gets a new value based on its a priori information (e.g., expression or mutation value) and by its associations with all the other features. In other words, it gets a new value based on how much is found "guilty" of being involved in the overall patient's biology and disease. This is based on three assumptions: the a priori information measures the strength of the link between the feature and the patient (e.g. the expression measures the gene importance and activity), the feature's associations indicate shared molecular or phenotypic characteristics (e.g. interacting genes have similar cellular functions) [49] and a disease is rarely a consequence of an abnormality in a single feature but reflects the perturbations and signaling cascades provoked by multiple dysfunctional molecules [49].

In our application, Simpati maps the a priori values of the genes to their corresponding nodes in the GGi network. It propagates the values through the interactions using the propagation algorithm. Each node, even the one without value, gets a score which reflects its starting information and the

amount given and received from its neighbours. The amount shared between nodes depends by the propagation type and the network topology. Simpati uses the random walk with restart (RWR) algorithm and the row-normalized version of the network. The RWR is a state-of-the-art network-based propagation algorithm [50] and flexible standardization technique [49] which has been successfully applied for the disease characterization [51] and the prioritization of multiple disease-associated biological features as genes [50], pathways [52], miRNAs [53,54], lncRNA [55], proteins [56] and somatic mutations [57]. While, the row normalization guarantees that a node gives the same amount of information equally to all its neighbours independently by their degree [58,59]. This allows to not favour specific nodes against others.

Simpati uses the propagation to get the same continuous numeric datum from any input profile, to include in a single feature's value also the perturbations provoked by the disease-specific activities of all the other patient's molecules and to boost the signal-to-noise ratio [49]. For example, a poorly expressed gene gets a high score if close to strongly expressed genes and a non-mutated gene gets a high score if close to a mutated one. As consequences, this step allows to handle different biological omics (e.g. dense gene expression data, sparse somatic mutation data) [60–63], to use a novel ad-hoc similarity measure and to not let the interpretation and meaning of Simpati's results depending by user-defined parameters.

For each profile $b \in \{1, \dots, B\}$, we define the set of its features represented as vertices with a priori information $SV_b = \{V_i \in V \mid \forall i \in \{1, \dots, N\}\, s.t.\ \exists\, F_i \rightarrow V_i\ and\ m_{i,b}\,! = 0\}$. The RWR algorithm measures the importance of each node v to SV_b. RWR mimics a walker that moves from a current node to a randomly selected adjacent node or goes back to source nodes with a back-probability $\gamma \in (0, 1)$. RWR is described as follows:

$$P_b^{t+1} = (1-\gamma)W' P_b^t + \gamma P_b^0 \quad (2)$$

where P^t is a $N \times 1$ probability vector at time point t of which the i_{th} element represents the probability of the walker being at node $v_i \in V$, P^0 is the $N \times 1$ initial probability vector and defined as follows:

$$P_b^0 = \{\tfrac{1}{|SV|}\ if\ v_i \in SV, 0\ otherwise\} \quad (3)$$

W' is the transition matrix of the graph GGi and $W'_{i,j}$ denotes a probability with which a walker at v_i moves to v_j. Formally, $W'_{i,j}$ is defined based on the row normalization:

$$W'_{i,j} = \frac{W_{i,i}}{\sum_i W_{i,j}} \quad (4)$$

The propagated profile P_b^{t+1} replaces the original profile P_b^0 in the matrix M.

2.4 TRENDING MATCHING SIMILARITY MEASURE

Simpati works with patient's propagated biological profiles. The feature has a score which measures how much is "guilty" of being associated and involved in the patient's disease or generic clinical condition driving its biology with specific alterations and disfunctions. Higher the score and more the feature is involved. Plus, the propagation score is meaningful also between profiles. Lower the propagation score varies between patients and more the feature is assuming the same role. This point from a pathway perspective is important. Two patients may strongly involve one biological process but may act on it from different directions. For example, two patients may have high scores for the EGFR pathway genes (n=79) but have very different values for few ones as EGFR, JAK, IL-6 and GAB1. This may due to the fact that, the disease of one patient is acting on the pathway using exclusively EGFR and JAK [64], while the other is using IL-6 and GAB1 [65,66]. By literature, the direction of regulation of a pathway is disease specific and a biomarker.

We wanted to capture both the aspects of the propagation score in developing the similarity measure to estimate how much two patients were similar, so we designed a novel pairwise similarity measure called Trending Matching (TM) similarity (0 lowest, 1 highest). It is the weighted sum of two components: the mean and the variation of the propagation scores of the features

belonging to a pathway. The first component measures how much the same feature is strongly or poorly involved in the patients. While the second component measures how much the same feature is similarly involved. For example, two patients described by a gene with high but different propagation scores (e.g., 1 and 0.7) are less similar than the pair which has lower but more close values (e.g., 0.8 and 0.7). The components are first determined for each single gene and then are summarized to represent the pathway. More the genes are strongly guilty, more the genes are similarly guilty and more the patients are considered similar in involving the pathway. This also prevents that, one outlier patient with genes strongly associated to the pathway can have a high similarity with another patient when they act on the process differently.

The trending matching similarity measures can be defined as follows; given a pathway $PH_u = \{ F_a \mid a \in \{1, \dots, A\}\}$ and two patient's profiles P_b and P_k, the similarity $TM_u(P_b, P_k)$ is the sum of three components:

$$WJ_u(P_b, P_k) = \frac{\sum_a \min(m_{a,b}, m_{a,k})}{\sum_a max(m_{a,b}, m_{a,k})} \quad (5)$$

$$MG_u(P_b, P_k) = \frac{(\sum_a (m_{a,b} + m_{a,k})/2)}{|PH_u|} \quad (6)$$

$$DIFF_u(P_b, P_k) = 1 - |WJ_u(P_b, P_k) - MG_u(P_b, P_k)| \quad (7)$$

$$TM_u(P_b, P_k) = WJ_u(P_b, P_k) + MG_u(P_b, P_k) + DIFF_u(P_b, P_k) \quad (8)$$

$$TM_u^-(P_b, P_k) = WJ_u(P_b, P_k) + (1 - MG_u(P_b, P_k)) + DIFF_u(P_b, P_k) \quad (9)$$

The TM similarity is designed in two variants. TM_u is designed to capture what we will call upinvolved psPSNs, higher is the second component and higher is the similarity between two patients. The most cohesive class has higher propagation scores for the same genes than the opposite class. TM_u^- is designed to capture downinvolved psPSNs, lower is the second component and highest is the similarity between two patients. In the next sections and paragraphs, we will detail the operations using only TM_u but the method performs every task with both the variants.

2.5 PATIENT SIMILARITY NETWORKS

Simpati aims to predict the class of an unknown patient comparing its propagated biological profile to the ones of the patients who are composing the classes of interest. For accomplishing the task, Simpati simulates a physician's decision process applied to solve the diagnosis and prognosis of a new individual. Creation of a mental database of known patients linked by their similarity (e.g. Lung cancer patients and healthy controls), selection of the features in which patients of the same class are similar between each other but dissimilar from others (e.g. EGFR biomarker with overexpression in Lung cancer patients with respect healthy controls) and assessment of the clinical outcome of the new individual based on its similarities with the database ones.

Doing a parallelism, the mental database is a set of pathway-specific patient similarity networks. For each biological process, Simpati represents the patients as vertices of a network and weighs their interactions based on the Trending Matching similarity measure. Formally, the similarity network is defined as an undirected weighted graph $PSN_u = (PV, PE)$ composed by a set of vertices $PV = \{pv_1, pv_2, \dots, pv_B\}$ representing the patient's profiles and a set of weighted edges $PE = \{(pv_n, pv_m) | pv_n, pv_m \in PV\}$ with $f': PE -> R$ such that $f'(pv_n, pv_m) = TM_{u \in \{1,\dots,C\}}(P_n, P_m)$ representing how much the pair of patients is similar in a specific pathway PH_u. The adjacency matrix corresponding to $PSN_u(PV, PE)$ is $W'': BxB = (w''_{n,m})$ where $w''_{n,m} = TM_u(P_n, P_m)$.

Simpati proceeds by selecting the pathways recognised as signature because represented by a PSN dividing the two classes while characterizing one. The members of one class must be more similar (i.e., stronger intra-similarities) than the opposite patients and the two classes not similar (i.e. weak inter-similarities). For this task, we developed a ranking system which evaluates a psPSN from 0 to 10 (the power of a PSN). Higher is the power and more a class is stronger than the opposite one and less the classes are similar (i.e., mixed together due to strong inter-similarities). First, we obtain three distributions based on the values of the similarities in the

psPSN. The distribution of the intra-similarities possessed by the members of one class, the distribution of the intra-similarities of the opposite patients and the inter-similarities between the members of the two classes. For each distribution, we compute a low and high percentile (e.g., 0.4 and 0.6). Then, we check if the distribution of intra-similarities of one class has the low percentile greater than the high percentiles of the other two distributions. In case the condition is satisfied, we decrease the low percentile, increase the high percentile, and compare again. For example, power 7 is satisfied when the 20 percentile of the intra-similarities of one class is higher than the 80 percentile of the other distributions. The power 9 when the 15 percentile of one class is higher than the 85 percentile of the other distributions. When, a psPSN has at least power 1 is considered signature of the most cohesive patient class.

When the psPSN is built with the TM_u similarity and has a power greater than 1 is considered signature and upinvolved because the members of the most cohesive class are similar due to higher feature's guiltless than the patients of the opposite class. On the contrary, the psPSN built with the TM_u^- similarity is considered downinvolved because the most cohesive class has the lowest feature's propagation values. Biologically speaking, the two classes are acting on the pathway differently, the members of one class are cohesive because their shared clinical condition is requiring and leading a precise alteration of the pathway, while the opposite class shows an heterogenous behaviour and a less need of acting on that cell function. We designed to capture this topological pattern and we do not require that the weak class must be cohesive following the study of Marquand et al. who reported that, assuming that both the classes in comparison are well defined precludes the inference of true diagnostic labels [67]. A clinical population may be composed of subjects belonging to different subtypes of the same disease or the heterogeneity may appear as result of misdiagnosis and comorbidities.

2.6 BEST FRIENDS CONNECTOR ALGORITHM

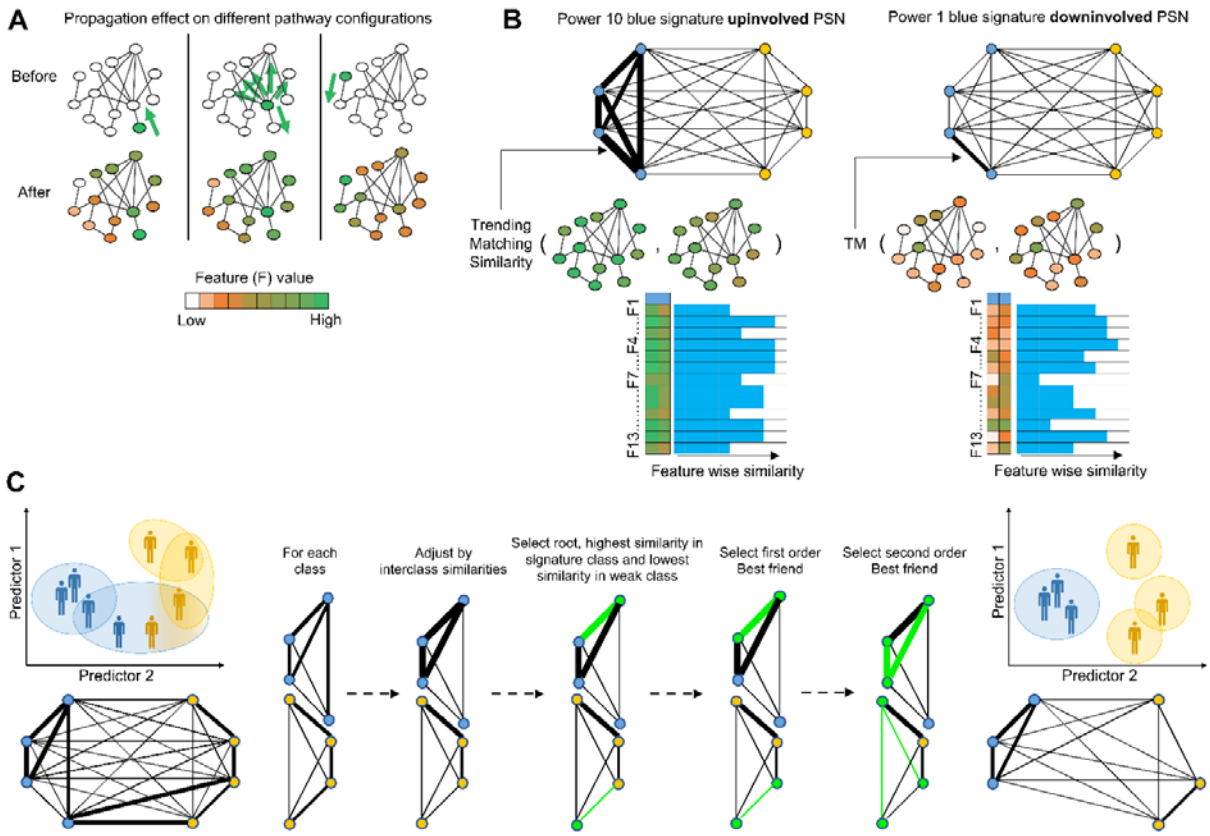
Simpati creates the database of psPSNs and then it aims to find the signatures. On the contrary of the starting situation in which a patient is described by the set of its biological feature's values, now the patient is described by its similarities with both the members of its same class and the non-members. This is extraordinary with respect supervised enrichment (e.g., differential expression analysis and gene set enrichment analysis tools) and machine learning tools which do not normally neither expect nor assume that a patient can relate to the individuals of the opposite class. However, the patient similarity network paradigm intrinsically introduces the presence of outliers as it could be likely to have them in the study of a real clinical cohort [67] or disease-specific class [68,69]. The meaning behind is that the outlier patient shows a pathway activity which is different from the one exhibited by the rest of the members.

Simpati tackles this aspect assuming that if a psPSN is not signature, but it still includes a class more cohesive than the opposite one, then this psPSN is likely to contain outlier patients with low intra-similarities and high inter-similarities. If this case appears, Simpati performs a filtering to get rid of possible outliers which are misleading the topological analysis of the network.

We introduce the Best Friend Connector algorithm (BFC) for identifying the most representative and cohesive subgroup in each class, for removing members that are not similar to the majority and for maximizing the psPSN signature power (i.e., the grade of separability between the classes). At first, it determines which class is the strongest, then it performs the selection. For the strongest class, it finds the subgroup with the strongest intraclass, and weakest interclass similarities. For the weakest class, the subgroup with the weakest interclass similarities. The patients not selected in the subgroups are considered outliers and removed from the psPSN. Simpati keeps count of in how many pathways a patient has been considered outlier and it provides this information as result of the workflow to allow a further analysis of the a priori input data. The psPSN is tested for being signature and then recomposed as it was originally.

The algorithm exploits the concept of first order best friend (1BF), second order best friend (2BF) and outsiders. A patient is a 1BF of another member called root when their similarity is in the root's best ones. A patient is a 2BF when it is 1BF of one root's 1BF and it has the root as 1BF. An outsider is a patient that does not belong to the class which is undergoing the analysis. The

algorithm performs the following operations for each class individually. It adjusts the weights of the intraclass connections. Precisely, it increases the similarity of two patients when both have a weak similarity with outsiders, while decreases in the opposite case. Iteratively, it considers one patient as root, it assesses the average of the intraclass similarities of the subgroup composed by his 1BFs and 2BFs. When each patient has been considered as root, the algorithm retrieves the set of



best friends who got the highest average. This guarantees of avoiding selecting multiple strong subgroups identified by different roots which would not represent the starting class uniquely.

Fig.2. This figure includes graphical details about the phase of data preparation, creation of psPSNs and filtering of outliers based on the Best Friends Connector algorithm. The frame A shows how the result of the propagation changes due to the position of the patient's molecule having a priori information. The propagation models how the pathway is specifically deregulated depending by which molecule is altered by the patient's disease or generically active in a sample's control profile. In the frame B, the Trending Matching similarity captures this aspect of the propagation. Two patients are considered strongly similar if each pathway's feature is involved in the same way because it means that also the pathway is perturbated and deregulated similarly by the disease of the patients. Plus, the frame illustrates the difference between up and downinvolved pathway. In the first case, higher are the propagation scores and higher is the similarity between patients, while in the second case is the opposite. Frame C shows the steps which compose the Best Friends Connector algorithm. For each class, it adjusts the similarity between every pair of patients based on their connections with the members of the opposite class, then it selects first and second order best friends to compose the subgroup around a root. When all the patients have been considered as root once, the subgroup of best friends with the strongest interclass connectivity is chosen as best one.

For explaining the pseudocode of the algorithm, we introduce the following mathematical details: Given four points in the Euclidian space $Q(x1, y1), E(x2, y2), R(x3, y3), T(x4, y4)$, a vector of continuous values $\bar{V}$, two indexes of patient's profiles $b, k \in \{1, \dots, B\}$ and a value $th \in [1, \dots, 100]$, we define the formula of the quadrilateral area, the formula of the percentile, the average of the

interclass similarities in two scenarios with respect b and k, and the first and second order best friend with respect b.

The quadrilateral area:

$QA(Q, E, R, T) = (1/2) \cdot \{(x1y2 + x2y3 + x3y4 + x4y1) - (x2y1 + x3y2 + x4y3 + x1y4)\}$. (10)

The percentile formula:

$percentile_value(\bar{V}, th) = \{ \bar{V}(sup(\frac{th}{100}) * |\bar{V}|)$ (11)

The average of the interclass similarities:

if $b, k \in \{EARLY\}$ then $r' = 1 - \sum_{i \in \{LATE\}} \frac{w''_{i,b}}{|LATE|}$ $r'' = 1 - \sum_{i \in \{LATE\}} \frac{w''_{i,k}}{|LATE|}$ (12)

while if $b, k \in \{LATE\}$ then $r' = 1 - \sum_{i \in \{EARLY\}} \frac{w''_{i,b}}{|EARLY|}$ $r'' = 1 - \sum_{i \in \{EARLY\}} \frac{w''_{i,k}}{|EARLY|}$ (13)

and the best friends with respect b:

$1BF_b = \{k \mid k \in \{1, \dots, B\}\ and\ (k, b \in \{EARLY\}\ or\ k, b \in \{LATE\}) AND\ TM_u(P_b, P_k) >= percentile_value([w''_{1,b}, w''_{2,b}, \dots, w''_{B,b}], th)\}$ (14)

$2BF_b = \{k \mid k \in \{1, \dots, B\} AND\ P_k \notin 1BF_b\ AND\ (k, b \in \{EARLY\}\ OR\ k, b \in \{LATE\})\ AND\ TM_u(P_h, P_k) >= percentile_value([w''_{1,h}, w''_{2,h}, \dots, w''_{B,h}], th)\}\ AND\ P_h \in 1BF_b\}$ (15)

$Max_b = \frac{\sum_{k \in \{BFS\}} w''_{k,b}}{|BFS|}\ with\ BFS = \{1BFb\ U\ 2BFb\}$ (16)

The Best Friends Connector algorithm is then shown in form of pseudocode in Fig.3 frame A.

A

```
BFC(PSN_u,th=5){
For each b,k∈{1,…,B} such that b,k∈{EARLY}
    Determine r' and r''; Q=[0,0] E=[0,w''_b,k] R=[0,r'] T=[w''_b,k, r'']:
    w''_b,k = QA(P,Q,T,Z)
For each b,k∈{1,…,B} such that b,k∈{LATE}
    Determine r' and r''; Q=[0,0] E=[0,w''_b,k] R=[0,r'] T=[w''_b,k, r'']:
    w''_b,k = QA(P,Q,T,Z)
For each b∈{1,…,B}
    Determine 1BF_b and 2BF_b, Max=0
    For each b∈{EARLY}
        Determine BFS and Max_b
        If Max_b>Max then Max=Max_b and PV''=BFS
    Max=0
    For each b∈{LATE}
        Determine BFS and Max_b
        If Max_b>Max then Max=Max_b and PV'''=BFS
    PV' ={PV'' U PV'''}
Return PSN'_u(PV',PE')=BFC(PSN_u,x) with PV'⊆PV and a set of
edges PE'⊆PE each one of which is incident with vertices from PV'
only
}
```

B

For each $u \in \{1, \ldots, C\}$

For each $x \in T = (5 * n)_{n=1}^{95}$

M1'=M2'=M1''=M2''=0 and th=95

Determine $1BF_z$ assuming $z \in$ EARLY

$$M1' = \frac{\sum_{k \in \{1BF_z\}} W'''_{k'z}}{|1BF_z|}$$

$PSN'_u(PV', PE') = BFC(PSN_u, x)$ with $PV' \subseteq PV$ and a set of edges $PE' \subseteq PE$ each one of which is incident with vertices from PV' only

If $pv_z \in PV'$ then $M2' = x$

Determine $1BF_z$ assuming $z \in$ LATE

$$M1'' = \frac{\sum_{k \in \{1BF_z\}} W'''_{k'z}}{|1BF_z|}$$

$PSN'_u(PV', PE') = BFC(PSN_u, x)$ with $PV' \subseteq PV$ and a set of edges $PE' \subseteq PE$ each one of which is incident with vertices from PV' only

If $pv_z \in PV'$ then $M2'' = x$

If $M2' > M2''$ AND $M1' > M1''$ then

Add EARLY to the LABELS set

elseIf $M2'' > M2'$ AND $M1'' > M1'$ then

Add LATE to the LABELS set

$$\text{Consensus_class} = \text{argmax}_a |a \in \text{LABELS}|$$

Fig.3. Figure of pseudocode relative to the two algorithms used in the Simpati implementation. The frame A shows the Best Friends Connector algorithm that is applied to a patient similarity network in form of adjacency matrix to filter outlier patients and detect the most cohesive subgroup. The frame B shows the step of class prediction related to an unknown patient performed at the end of the Simpati workflow.

2.7 CLASSIFICATION

Once Simpati created the database of signature psPSNs, it uses them to classify an unknown patient. This for understanding the quality of the selected pathways in characterizing and distinguish the classes in comparison. Simpati performs the operation continuing to follow the physician's decision process. The unknown patient is compared to the ones already annotated in the mental database and assigned to the same class of who is most similar to. However, the only strength of similarity could be misleading. The unknown patient could have the strongest similarity with outlier members of the class. Therefore, we designed Simpati to consider also how much the unknown patient fits in the class.

The method prepares the unknown patient's profile. The profile is replaced with its propagated version, compared to the patients in every signature pathway and added as new node in the corresponding psPSNs. Then, Simpati associates the profile to one of the classes based on the results of two approaches. For the first, it determines the average of the highest values of similarity that the patient has inside each class. The patient would be associated to the class with which has the highest average. While for the second approach, Simpati pretends that the patient belongs to one class and measures how much is far from being considered an outlier. The patient would be associated to the class in which is considered less outlier with respect the other members. In details about this step, the patient is simulated to belong to one class and the BFC algorithm is performed iteratively. At each run, the algorithm decreases the size of the subgroup of patients which retrieves. It stops when the patient does not belong to the best subgroup. Higher is the number of iterations and more the patient is considered having a stronger similarity with the class representatives than the other members which are more likely to be outliers. Simpati uses the

iteration number as distance measure from the “outlier” status. Due to this, the patient would be candidate to be associated to the class in which survived the highest number of iterations.

Simpati associates the patient to the class that has been predicted by both the approaches. In case the results are not concordant, then Simpati does not make the prediction and the pathway together with its PSN are removed from the downstream operations. This step is performed for all signature psPSNs, then Simpati performs the consensus prediction. The patient’s definitive class is the one to which has been most frequently assigned.

Formally this would be, a new patient’s profiles P_z such that $z \notin EARLY\ and\ z \notin LATE$ is added as node in each patient similarity network. Let us define the new $PSN_u(PV, PE)$ composed by the set of nodes $PV = \{pv_1, pv_2, \dots, pv_B, pv_z\}$ representing the patient's profiles and the set of weighted edges $PE = \{(pv_n, pv_m) | pv_n, pv_m \in PV\}\ with\ f': E -> R\ s.t.\ f'(pv_n, pv_m) = TMu(P_n, P_m)$. The class of the new profile is found by a topological analysis of all the psPSNs shown in form of pseudocode in the figure 3 frame B.

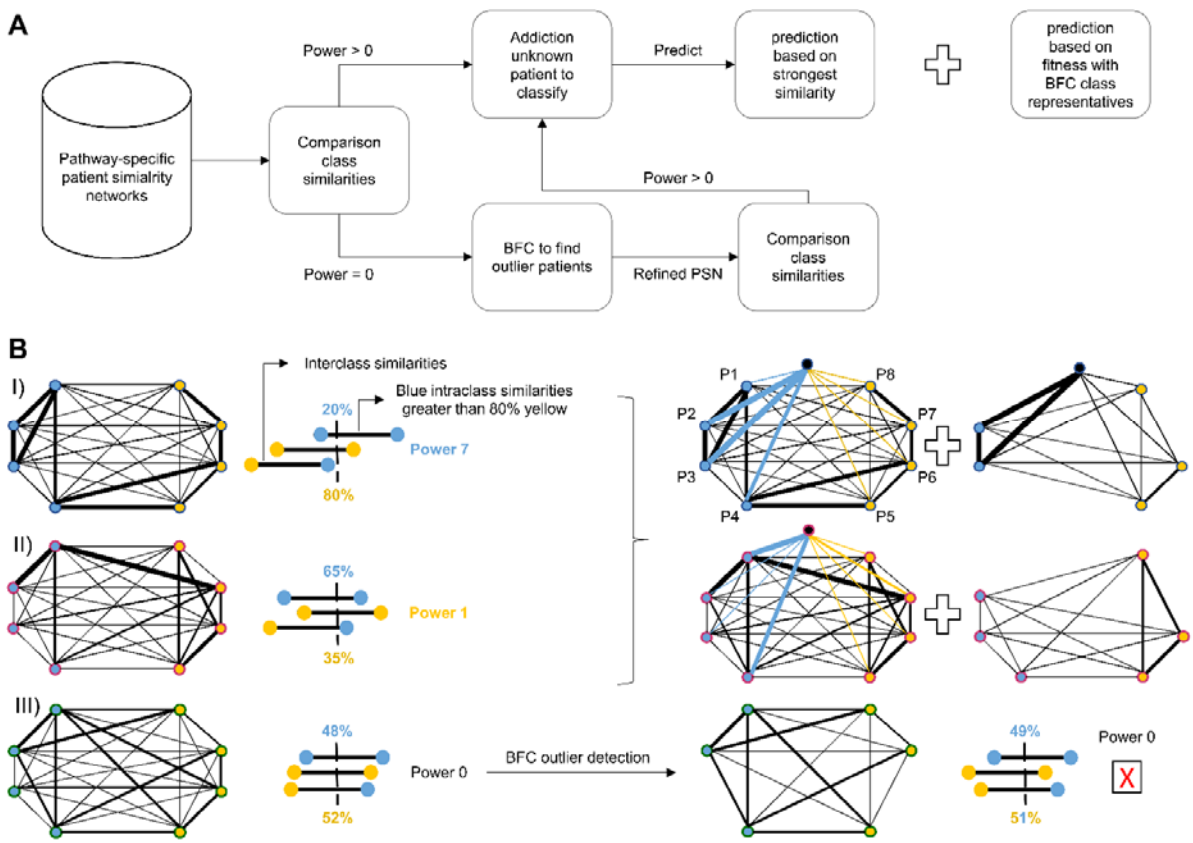


Fig.4. This figure shows the journey of a psPSN. In the frame A with a flow diagram, while in the frame B with different graphical scenarios. First, the psPSNs are created and compose a database. Then, Simpati determines the power of the networks. If a psPSN has a significative power is used for the classification, otherwise the method performs the BFC algorithm trying to detect outlier patients. If there is a cohesive subgroup in the psPSN and the removal of outliers significantly increase the power of the network, then the latter is reintroduced in the main workflow as signature pathway. The networks with sufficient power are used to classify an unknown patient. Simpati measures how much the patient is similar to the members of the classes and how much is distant from being an outlier. For this last step, Simpati runs the BFC algorithm iteratively getting each time a cohesive subgroup stronger than the one of previous iterations, more times the unknown patient is found in the best subgroup and less is considered outlier.

2.8 OUTPUT

Simpati provides the signature pathways used in the classification step, the corresponding PSNs in vectorial format and reports statistics to allow further analysis and considerations. For example, the average of the intra and inter similarities for recognising the most cohesive class, the psPSN power to get the pathways which more separate the classes in comparison and a probability value (p.value). The latter is assessed testing the psPSN to retrieve an equal or higher power than the original one when patients are permutated between classes. This information allows to filter out pathways which have been detected as signature due to random.

Simpati also includes two tools for the visualization of the data produced with the workflow, one tool is an internal function able to produce a compact representation of a psPSN, while one is a graphical user interface (GUI) for the exploration of a patient's propagated biological profile. The function provides a compact representation of a psPSN by reducing the patients which are visible as nodes. This is necessary to allow the user to understand how patients are similar between each other. In fact, more the number of nodes increases and more becomes difficult to follow the edges of the network and so how the patients are similar between each other. To do so, Simpati groups up patients of the same class that are considered similar, chooses a patient to represent each subgroup and filters the original network by keeping only the chosen (i.e., the representative of each subgroup). First, it determines how much every pair of patients are similar in the network. It uses the measure $WJ_u(P_b, P_k)$ which is applied between two patient's profiles composed by their similarity values in the psPSN. It gets a psPSN of the psPSN (aka psPSN^2). Simpati proceeds and iteratively performs the BFC algorithm on the psPSN^2. In this case, the BFC algorithm is not applied to filter outliers but to detect multiple cohesive subgroups. At every iteration, the BFC algorithm detects and removes the most cohesive subgroup composed by the twenty percent of the original members of one class. When the BFC cannot be applied anymore due to the small size of the class, Simpati filters the original psPSN and keeps only the class members which have not been grouped in the last iteration together with the root of each class-specific detected cohesive subgroup. Further about the aesthetic aspects, the size of a node is used to indicate how much the relative patient is similar inside its own class compared to how much is similar with the outsiders. This is assessed with the difference between two PageRank [70] centrality scores, one is measured only with the patient connected by similarity to the members of its class and one with only the outsiders. Higher the difference, higher the size of the node, more central the patient is in its class and less similar to the outsiders. While, the position of the nodes in the plot of the psPSN is determined using the Fruchterman & Reingold's force-directed layout [71]. The network becomes compact, easy to analyse and still representative of how all patients are similar. Thanks to the plot, the user can understand if the psPSN is correctly a signature, see how the similarities are distributed, identify which patient is crucial for the connectivity of its own class and which is instead behaving as outlier.

Complementary to the visualization of the psPSN, we also provide an R shiny graphical user interface (GUI) to allow the exploration of the propagation effect over a patient's profile. This enables the user to understand how the values of the patient's biological features changed and for which reason. For example, a gene with low expression value that has been removed from the Limma analysis due to the function "filterByExpr" in the differential expression analysis, it may get a high propagation score and the user may get interested in understanding the reason. We believe that this can be another useful instrument to make the method and the data more accessible.

2.9 ENRICHMENT

The power, the p.value, the distribution of the similarities are all technical information regarding a psPSN that allow to understand how patients and classes are structured. However, they offer a limited utility in prioritizing the best pathways because they are not related to any biological background of the patients. For this reason, in case the patient's features are genes, we designed Simpati to perform a query in the Disgnet [72] and Human Protein Atlas [73] (HPA). DisGeNET is a database which provides open access to annotated disease associations with genes and variants. While HPA is a unique world-leading effort to map all the human proteins in cells, tissues, and organs in the human body using antibody-based imaging, mass spectrometry-based proteomics, transcriptomics, and systems biology. Simpati requires the semantic type of the patient's disease

(e.g., Neoplastic Process, Congenital Abnormality, etc..) and key words (e.g., TCGA-KIRC: Kidney, Renal, Carcinoma). Then, it gets which published articles have associated the genes of a signature pathway to the semantic type and to the key words. Plus, in case of cancer patients, this integration allows to get which genes are favourable to be prognostic, while in case of non-cancer disease which genes are associated to the user-defined tissue. As indicated by Lin et al. [74] this operation allows to prioritize the signature pathways based on their associations with the patient's clinical outcome and to understand better the validity of Simpati results.

2.10 WORKFLOW OF TESTING

Simpati ability to classify the classes in comparison is tested with a leave one out cross validation (LOO-CV). Given a dataset of patient's biological profiles and the classes associated to them, Simpati iteratively performs the following operations: one patient is considered unknown and compose the testing set, while the remaining patients are considered known and used as training set. The latter is used to build the psPSNs, to find the signature pathways and as ground truth in the classification step. While, the testing patient has the biological profile which class must be predicted. In the end, the predicted classes of the testing patients collected from all the iterations are compared to their real ones for determining the classification performances. Simpati is designed to value its ability to predict based on two measures following netDx design [28]. The first one called AUC-ROC is the area under the curve where the x-axis is the false positive rate (FPR) and the y-axis is true positive rate (TPR). While the second one called AUC-PR is the area under the curve where the x-axis is the recall and y-axis is the precision [75].

3. RESULTS

3.1 CLASSIFICATION COMPARISON

We tested Simpati performances to classify patients of five different TCGA cancer types described by two biological omics: transcriptomics with gene expression data and genomics with somatic mutations. The classes assigned to the patients were Early or Late based on their cancer stage. We increased the challenge including the performances of the current published generic-purpose pathway-based classifiers: netDx [28] and PASNet [19]. netDx creates a database of predictive PSNs associated to pathways for each class, builds a consensus similarity network and applies GeneMANIA (state-of-art gene function prediction algorithm) for the prediction of the testing patients. netDx tests its performances with a 10-fold cross-validation which in each run includes another cross-validation for the selection of the most predictive PSN. While, PASNet incorporates biological pathways in a Deep Neural Network. The neural network is composed by a gene layer (an input layer), a pathway layer, a hidden layer that represents hierarchical relationships among biological pathways and an output layer that corresponds to the patient classes. PASNet tests its performances with a stratified 5-fold cross-validation repeated 10 times. The two competitors either support or use the classification evaluation based on the area under the receiver operating characteristic curve and the area under precision-recall curve measures and they differ from canonical supervised machine learning algorithms. For these reasons, we performed the comparison using each method based on how it has been designed and following the vignettes provided by the authors.

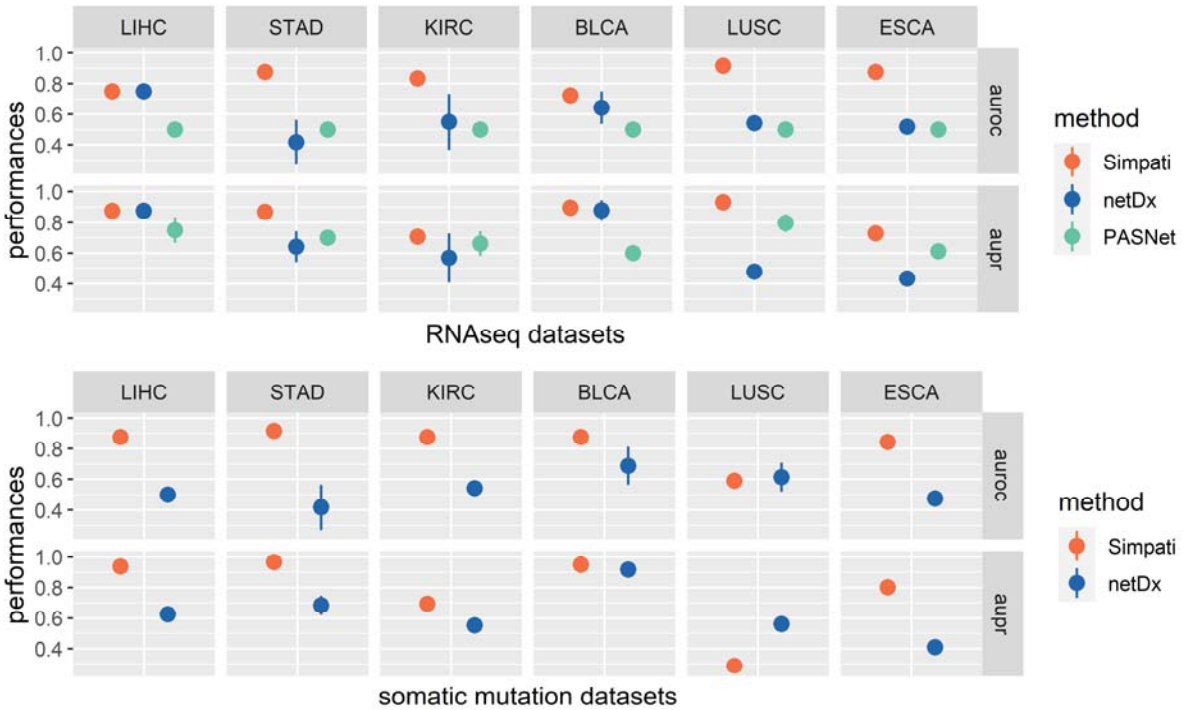


Fig.5. Comparison of the classification performances between the pathway-based classifiers Line plot of median (dot) classification performances with error bars (line). X-axis indicates the datasets. Y-axis indicates the value of area under the roc/pr curve. The same plot is presented twice, one including the performances when the methods classify the RNAseq data, while one the somatic mutations. PASNet does not have performances with somatic mutations because it does not handle sparse biological data. The plot shows that Simpati performs better than the competitors in all the datasets except for LUSC with somatic mutation.

Simpati performs better than the competitors for both the measures with all the omics. Simpati also proves to be more reliable in each dataset with a standard error equal to zero due to its leave one out cross-validation approach. While the performances of the competitors highlight common classification issues; their performances vary a lot probably due to the number of patients, the size of the classes in comparison and the ability of the classifier to naturally handle multiple omics and data types.

3.2 SIMILARITY NETWORK COMPARISON

Simpati and netDx provide pathways and related PSNs as result of the analysis. However, the methods use different techniques for the pathway selection. netDx selects a pathway if the corresponding PSN allows GeneMANIA to correctly predict the classes of the training and testing patients. While Simpati selects a pathway if the corresponding PSN topologically separates the classes in comparison. The resulting pathways and related PSNs should help to characterize the patient classes, explain why they have been used to predict and they should increase the interpretability of the model. We compared the topology of the PSNs selected by the two contenders based on their power.

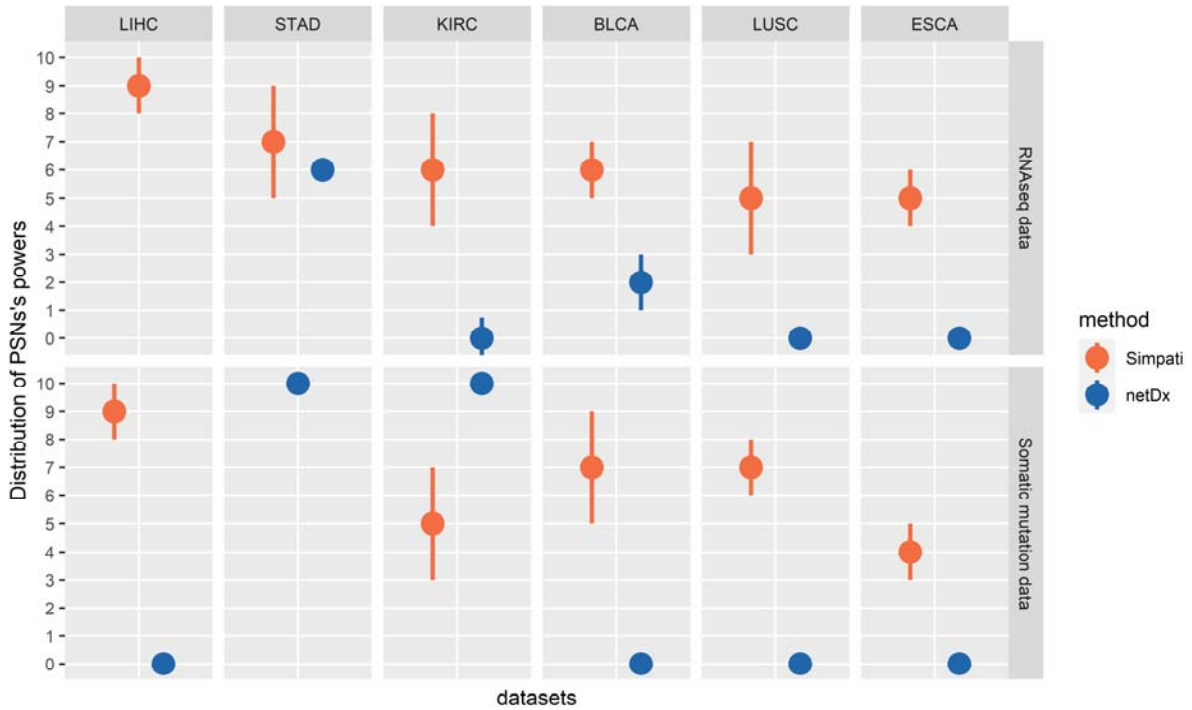


Fig.6. Comparison between the topology of the PSNs retrieved as result by netDx and Simpati with the TCGA datasets. The topology of the PSNs is measured with their power. Each frame of the image is dedicated to the pathways selected for the classification of a specific biological omic. The Y-axis indicates the power of the PSNs retrieved by a specific method. The X-axis indicates the datasets. Specifically, the dot indicates the median of the power of the PSNs resulted by applying a specific method with a specific dataset, while the line ranges based on the standard deviation of the same PSNs' powers. Simpati selects better PSNs except in STAD patients described with somatic mutation profiles.

Simpati provides more pathways with high power than netDx in all the datasets except one. This is probably due to how the selection is done. Simpati discerns PSNs based on their topology and then performs the classification. While netDx evaluates a pathway based on the mere ability of the GeneMANIA algorithm to use its PSN for classifying. This makes the difference in terms of interpretability of the model. From the final user perspective, Simpati's psPSNs together with their visual representation make easier to understand why they have been selected for the classification and can be perceived as more trustable.

3.3 COMPUTATION RESOURCES COMPARISON

The patient similarity network paradigm used by Simpati and netDx brings many advantages both in the feature selection, in the classification phase and in the overall interpretability of the software. However, these pros come with a price which is the software scalability already introduced as challenge by Pai et al. [20]. A PSN is a complete graph that the methods build with all the patients and for every pathway. This means that an increment in the number of patients and in the number of annotated pathways lead the methods to require more computational resources. netDx and Simpati faced this point with different approaches. netDx is implemented in R and Java, uses the disk to save temporary files and applies a sparsification of the PSNs to decrease the number of edges and so the amount of information associated to them. While Simpati is implemented completely in R, natively support parallel computing and handles all the data of the workflow as sparse matrices or vectors. To understand which software handles better this issue, we captured the ram usage and the running time that each method required to classify the TCGA datasets with the same hardware setting (AMD Ryzen Threadripper 3970X 32-Core Processor, 251 Gigabyte

System memory and Linux ubuntu-1804-slurm 5.4.0-72-generic). Simpati resulted to be more efficient in both the running time and the ram used.

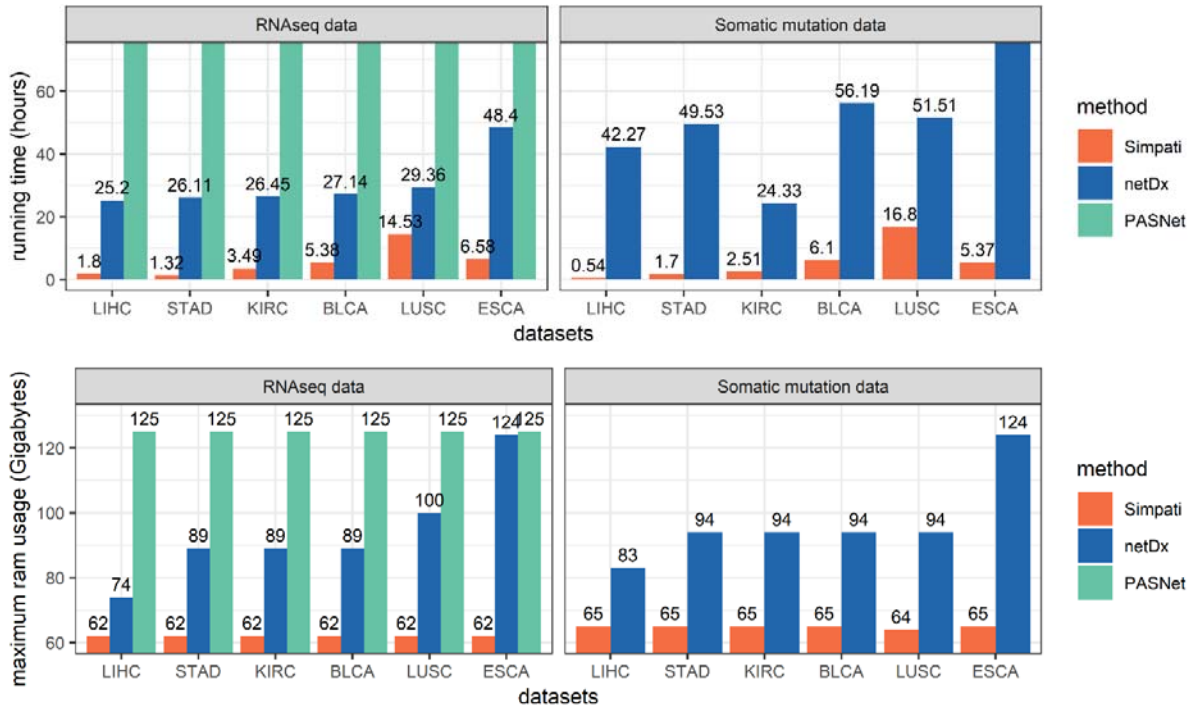


Fig.7. Barplot shows the comparison between the computational resources used by Simpati, netDx and PASNet to classify the TCGA datasets. The measures used for this comparison include the running time in hours and the memory ram in the maximum amount needed by the software in Gigabyte. In fact, the maximum amount is the real obstacle to the correct execution of the software. The X-axis indicates the datasets. The Y-axis indicates the measure. PASNet with the RNAseq data has a running time which exceeds the three days (72 hours). The plot shows how Simpati outperforms the competitors in time and memory for all the datasets.

3.4 ENRICHMENT COMPARISON

Pathway-based classifiers aim to classify correctly unknown patients using the biological pathway information. This means that the prediction of a patient's class passes through the selection of pathways which due to method-specific criteria are considered useful for the task. In a cross-validation setting, the final classification performances indicate how much the classifier is reliable and better than a random predictor. However, they do not represent a measure of how much the pathways are biologically significant. A classifier as Simpati can provide further details about how it used the pathways, why it selected them and the biological interpretation under the filtering criteria but still, this information does not allow to understand if pathways are biologically meaningful. For this reason, we designed Simpati to integrate an enrichment step and we performed this operation also to the results of the other competitors. Precisely, we kept only the resulted significant pathways having at least one publication associated to each of the key words defined per dataset and having at least the 90% of the genes associated to the patient's specific cancer. Then, we compared the numbers of pathways satisfying these constraints.

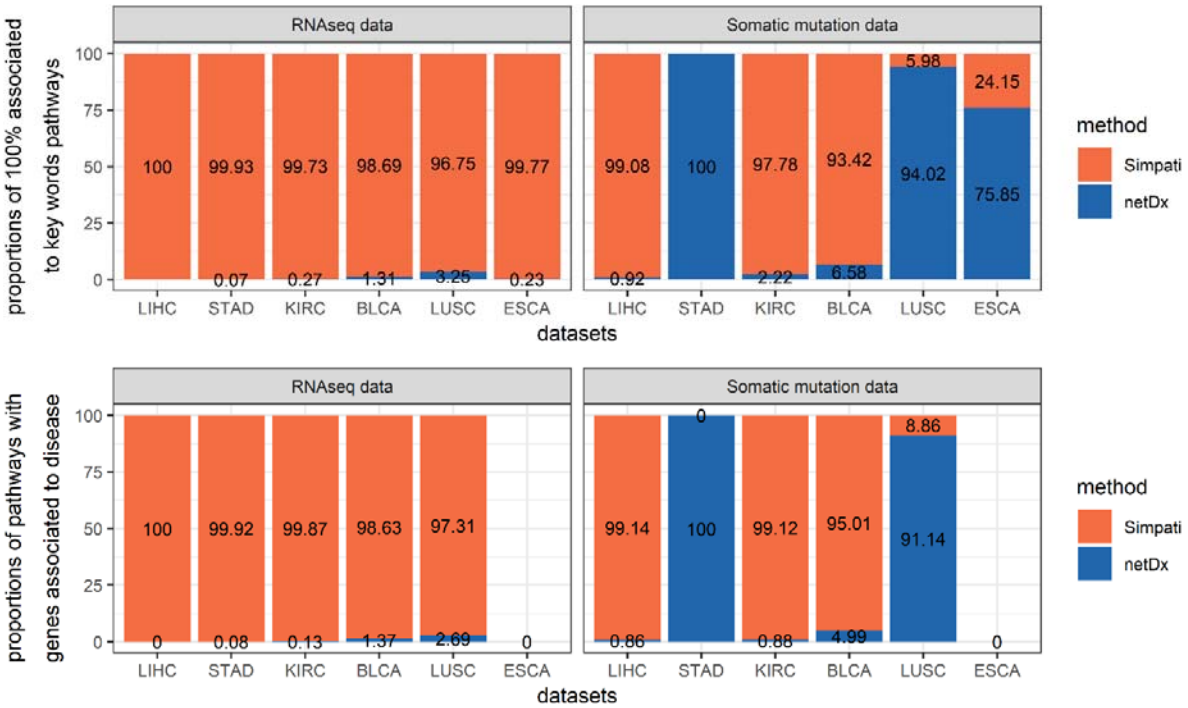


Fig.8. 100% stacked bar chart shows the comparison between the enrichment statistics obtained by querying the resulting pathways of the different classifiers in Disgnet and the Human Protein Atlas with respect the patient's cancer types. Each frame compares the methods based on how their pathways are qualified with a specific measure. The X-axis indicates the datasets in comparison. A bar is divided based on the number of pathways obtained by the two methods and satisfying the criteria indicated by the Y-axis. For example, netDx did not selected any pathway for the classification that satisfied the criteria in the classification of the LIHC patients with RNAseq profiles, while for the somatic mutation profiles Simpati has selected 99% more pathways.

This analysis highlights that Simpati is both able to select, use and provide biologically significant pathways directly associated to the patients that it is classifying and that performs better than the competitors. netDx retrieves always much fewer pathways biologically associated to the tumor of the patient's profiles than Simpati. We have been unable to include PASNet due to its lack in providing the pathways that it considers significant and predictive of the patients classified.

3.5 PATHWAY ANALYSIS

In this last section, we compare the Simpati signature pathways obtained from the classification of KIRC patients with literature findings. We considered the key pathways identified by Pang et al. [76] and Cui et al. [77].

Late stage KIRC regulates the cell adhesion to escape from the immune system. It disturbs the ATP supplement, making cells to adopt anaerobic respiration, producing an ATP deficit and hypoxia. It provokes an immune response from the host organism, regulates the PI3K-Akt signalling pathway, promotes anabolism and inhibits catabolism.

Simpati has been able to find key pathways differentiating the pathological stages of the patients. Specifically, it confirmed the relationship between the late cancer stage and both the upregulation of the cell adhesion, the response of the immune system and the downregulation of the catabolism. For this comparison, we also included the results of the standard gene set enrichment analysis (GSEA) performed thanks to the GAGE method [78]. The two methods found similar key pathways. However, GSEA provided the statistics mean, while Simpati provided the patient similarity network.

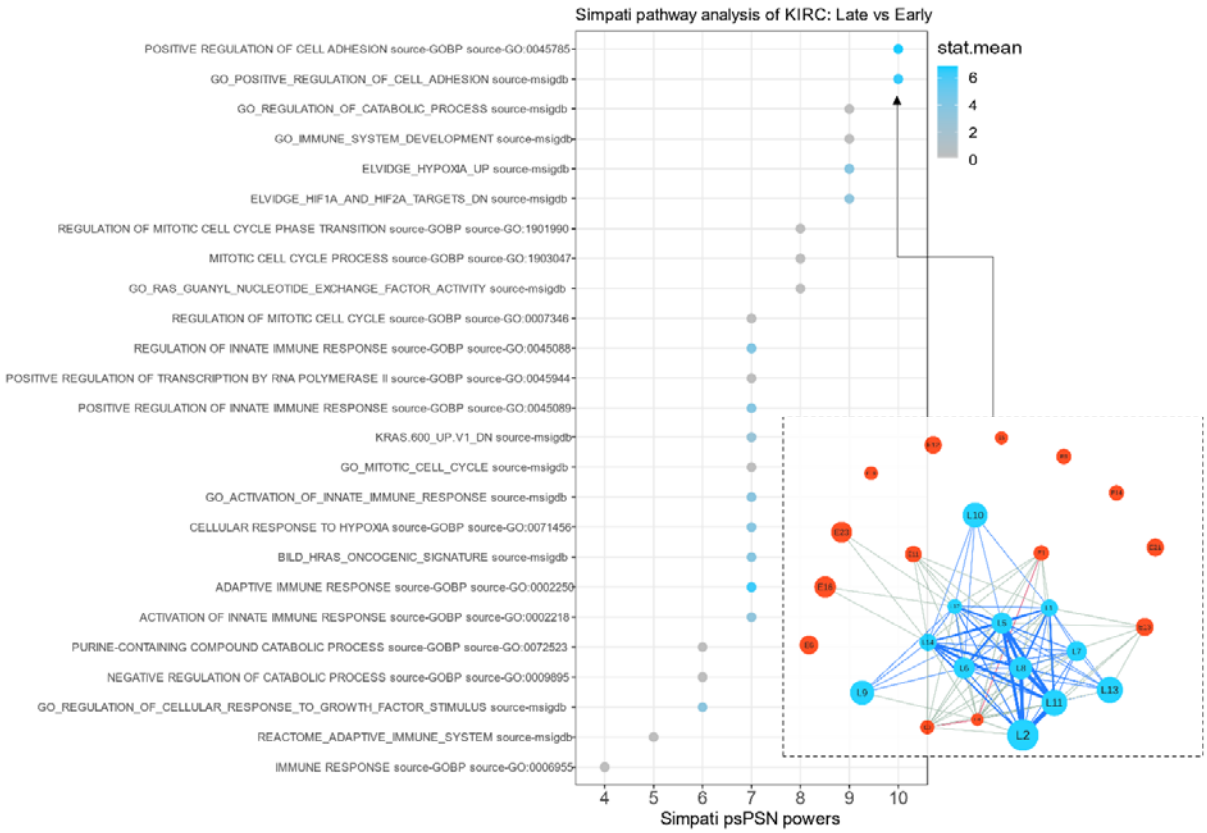


Fig.9 Dotplot showing the signature pathways detected by Simpati and the standard gene set enrichment analysis for the comparison of the KIRC patients between Late and Early stage. In the y-axis, there are the signature pathways and psPSNs resulting by performing Simpati. The x-axis indicates the power of the similarity networks. The legend describes how much the standard GSEA finds the same pathways altered between the two classes. The pathways with zero statistics mean (stat.mean) are not significant for the GSEA. Finally, in the bottom right cornet, the plot of the psPSN related to the positive regulation of the cell adhesion pathway. Blue nodes represent Late KIRC patients, while red nodes represent Early patients. The blue edges are the Late interclass similarities and the grey are the interclass.

4. CONCLUSIONS

We propose the pathway-based classifier called Simpati. The method can be applied to different omics, proved to obtain quality classification performances and detect signature pathways. In other words, it identifies biological processes that distinguish and uniquely characterize the clinical classes of the patients in comparison.

On top of the technical conclusions, we want to suggest Simpati as tool for computational biologists and bioinformaticians that want to get unique insights about their patients or samples. We designed Simpati to simulate a physician's decision process applied to solve the diagnosis and prognosis of a new individual. As a physician, our software processes, stores and learns information related to the patients. All the data used during the classification are then made available for allowing further analysis.

Simpati associates to each single biological feature (e.g. gene, protein, mutation, ...) a propagation score which reflects the overall biology of the patient. A high score indicates that the feature is strongly involved in the patient's biology, while a low score the opposite. The scores can be explored in two ways. They can be considered as values of a standard high-dimensional matrix with patients at the columns and features at the row. They can be visually taken into account in an ad-hoc graphical-user interface. The matrix format allows any statistical analysis with clusterProfiler [79], while the GUI permits to understand how much specific biological features of

interest are important without any programming and statistical knowledge. The information retrieved by analysing the propagation scores can be combined to the results obtained from a differential expression (DE) analysis. For example, a DE list can be filtered to keep only the genes that have a high score in order to reduce the false positive or can be expanded integrating those genes that are DE in term of propagation values.

Simpati models pathways as patient similarity networks. In a psPSN, patients are connected to others based on how much their biology similarly regulate a specific pathway. Like in a social network in which people are connected to others based on their hobbies and how they practice them (e.g., the place, the effort, the time). More two patients involve and regulate similarly a pathway (e.g., with the same genes and with the same expression values) and more they are strongly connected. In case a pathway is found significant and of interest, it can be explored in two ways. The adjacency matrix or the graphical representation of the related psPSN. The matrix format allows any network analysis with NetworkToolbox [80], while the plot permits to have intuitions about how much the patient classes separate and to identify patients that are central or tend to be outlier. The information retrieved by analysing the topology of the psPSNs can be used to verify the clinical information associated to the patients, identify subclasses, and can be combined to the results obtained from a clustering analysis or a non-negative matrix factorization (NMF) [81]. For example, the subclasses of patients that have been identified with an unsupervised technique can be checked in the psPSNs to find in which pathways are mostly similar.

Simpati finds signature pathways to characterize and distinguish the patient classes in comparison. The pathways must satisfy a constraint. The members of one class must be more similar than the opposite patients. Then, it uses the similarities to predict the class of new patients. In this sense, Simpati can be combined to a standard gene set enrichment analysis because detects pathways that satisfy a criterium not taken into account by other tools.

Simpati does not assume that the patient classes are well defined, and it considers the possibility that members of the same class may regulate the same biological process differently. When this is likely to happen with a pathway, Simpati identifies the patients that most represent their own class, uses only them to check the signature condition and the remaining members are considered outliers. For this reason, Simpati is suitable to real case scenarios which often include either patients or samples associated to clinical outcomes due to a priori information by wet lab scientists. The latter check the expected sample classes using a principal component analysis or clustering. However, both the methods are not designed to detect differences at the level of pathways or single biological features which could reveal unique biological aspects of a sample and differentiate it from the rest of its class. For example, in a knock-out study, samples are labelled as knocked-out based on the experiment but this does not necessarily imply that each member of the clinical population shows changes in gene expression levels against the control group. Standard gene set enrichment analysis tools are ineffective due the possible low variation between the classes and possible knock-out samples not showing any change.

5. DISCUSSION

Generic purpose pathway-based classifiers propose themselves as powerful tools for classifying patients and providing biologically meaningful results in form of pathways. The first one has been introduced in the 2010 but, at the time of writing, there are very few software available. they remain very few. This is due to the many challenges that must be faced to produce high-quality and functioning software they inherit. We tried to report and detail all the issues related to the development of this kind of machine learning algorithm. At the same time, we tried to build a software that could have been considered a future example for other researchers. Thanks to the combination of new and popular strategies, Simpati proves that is possible to both obtain satisfying results and tackle common issues in the pathway-based classification.

The preparation of the patient's biological profiles with a transformation technique as the network propagation allows to get the same kind of data and information before the classification. This allows the researchers to develop a workflow which is flexible, consistent, and involving less hyper-parameters. As a matter of fact, we developed the Trending Matching similarity to capture a

specific relationship between the patient's propagated profiles and the scored genes. On the contrary, netDx suggests using the Pearson correlation as default measure directly on the raw profiles but the authors did not provide a biological interpretation of what kind of patient similarity leads to catch and leave the user to the uncontrolled intrinsic disadvantages of the measure [82]. For example, the correlation is biased and leads to incorrect inferences when considers genes that have not been perturbed (e.g., environmentally or genetically) in order to cause a meaningful change in expression level [83].

The selection of the predictive pathways including criteria that can be explained from a biological point of view allows the classifier to not drift away from the patient's biology. For example, both PASNet [19] and netDx [27] select the pathways which perform the best in predicting the training patients. This approach is undeniable well suited for securing the ability to predict an unknown patient. However, it may lead to select pathways which are useful for the algorithm of prediction and meaningless for the patient's biology. On the contrary, Simpati performs a first selection based on criteria that can be biologically interpretated and then it analyses the pathways for the classification. As we proved with our results, a biological selection does not necessarily negatively affect the final classification performances but it indeed changes positively the resulting pathways produced by the classifier.

The analysis of outlier training patients because their biological features are not showing the same activity in a pathway as the rest of the class increases the granularity of the cell process description that the classifier handles and provides as result. This brings multiple advantages. It makes the classifier less sensitive to how much the data have been cleaned, how the patient classes have been defined, and it allows to give a hint about subclasses of patients which are using different pathways at the net of one shared clinical status.

At the same time, it is worth to also mention the price of such strategies. The propagation leads the classifier to require also the network of interactions or associations between the features of the biological omic. The selection of pathways based on criteria which are biologically explainable is not trivial and may makes the classification inconclusive due to no pathway passing the filter. The analysis of the outliers requires either parameters to set up which may be not correct for all the applications or hyper-parameters to determine.

6. AVAILABILITY OF DATA AND MATERIALS

All the work has been made in R programming language, from the data extraction to the enrichment. We provide a github repository with a tutorial about how to replicate all the results of this project: https://github.com/LucaGiudice/supplementary-Simpati .We provide an R package to use Simpati: https://github.com/LucaGiudice/Simpati and to use the GUI: https://github.com/LucaGiudice/propaGUIation

AUTHOR INFORMATION

LG designed and developed the software. LG wrote the article and supervised the project. LG produced the images and created the github repositories.

ACKNOWLEDGEMENTS

I want to thank my best friend Samuele Cancellieri who supported me during the development of this work, who has been the true reader and critic of this manuscript, and that was always there to hear my difficulties and complains. I want to thank Shraddha Pai, author of netDx, who inspired me to develop this software and taught me what to improve and how to do always better. I want to thank Gary Bader who started this adventure of mine by letting me work on patient similarity networks in his research group during my master thesis. I want to thank my girlfriend Lucy who had to deal with all the stress of mine caused by this project.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Giannuzzi D, Giudice L, Marconato L, Ferraresso S, Giugno R, Bertoni F, et al. Integrated analysis of transcriptome, methylome and copy number aberrations data of marginal zone lymphoma and follicular lymphoma in dog. Vet Comp Oncol 2020;18:645–55. https://doi.org/10.1111/vco.12588.

[2] Giudice L, Cascione L, Ferraresso S, Marconato L, Giannuzzi D, Napoli S, et al. Long Non-Coding RNAs as Molecular Signatures for Canine B-Cell Lymphoma Characterization. Noncoding RNA 2019;5. https://doi.org/10.3390/ncrna5030047.

[3] Kolosowska N, Gotkiewicz M, Dhungana H, Giudice L, Giugno R, Box D, et al. Intracerebral overexpression of miR-669c is protective in mouse ischemic stroke model by targeting MyD88 and inducing alternative microglial/macrophage activation. J Neuroinflammation 2020;17:194. https://doi.org/10.1186/s12974-020-01870-w.

[4] Jin L, Zuo X-Y, Su W-Y, Zhao X-L, Yuan M-Q, Han L-Z, et al. Pathway-based analysis tools for complex diseases: a review. Genomics Proteomics Bioinformatics 2014;12:210–20. https://doi.org/10.1016/j.gpb.2014.10.002.

[5] Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. PNAS 2013;110:6388–93. https://doi.org/10.1073/pnas.1219651110.

[6] Segura-Lepe MP, Keun HC, Ebbels TMD. Predictive modelling using pathway scores: robustness and significance of pathway collections. BMC Bioinformatics 2019;20:543. https://doi.org/10.1186/s12859-019-3163-0.

[7] Raghavan N, Amaratunga D, Cabrera J, Nie A, Qin J, McMillian M. On methods for gene function scoring as a means of facilitating the interpretation of microarray results. J Comput Biol 2006;13:798–809. https://doi.org/10.1089/cmb.2006.13.798.

[8] Lee E, Chuang H-Y, Kim J-W, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol 2008;4:e1000217. https://doi.org/10.1371/journal.pcbi.1000217.

[9] Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353–7. https://doi.org/10.1038/nature04296.

[10] Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, et al. Methods and approaches in the topology-based analysis of biological pathways. Front Physiol 2013;4:278. https://doi.org/10.3389/fphys.2013.00278.

[11] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 2012;8:e1002375. https://doi.org/10.1371/journal.pcbi.1002375.

[12] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009;37:1–13. https://doi.org/10.1093/nar/gkn923.

[13] Xie C, Jauhari S, Mora A. Popularity and performance of bioinformatics software: the case of gene set analysis. BMC Bioinformatics 2021;22:191. https://doi.org/10.1186/s12859-021-04124-5.

[14] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17:13. https://doi.org/10.1186/s13059-016-0881-8.

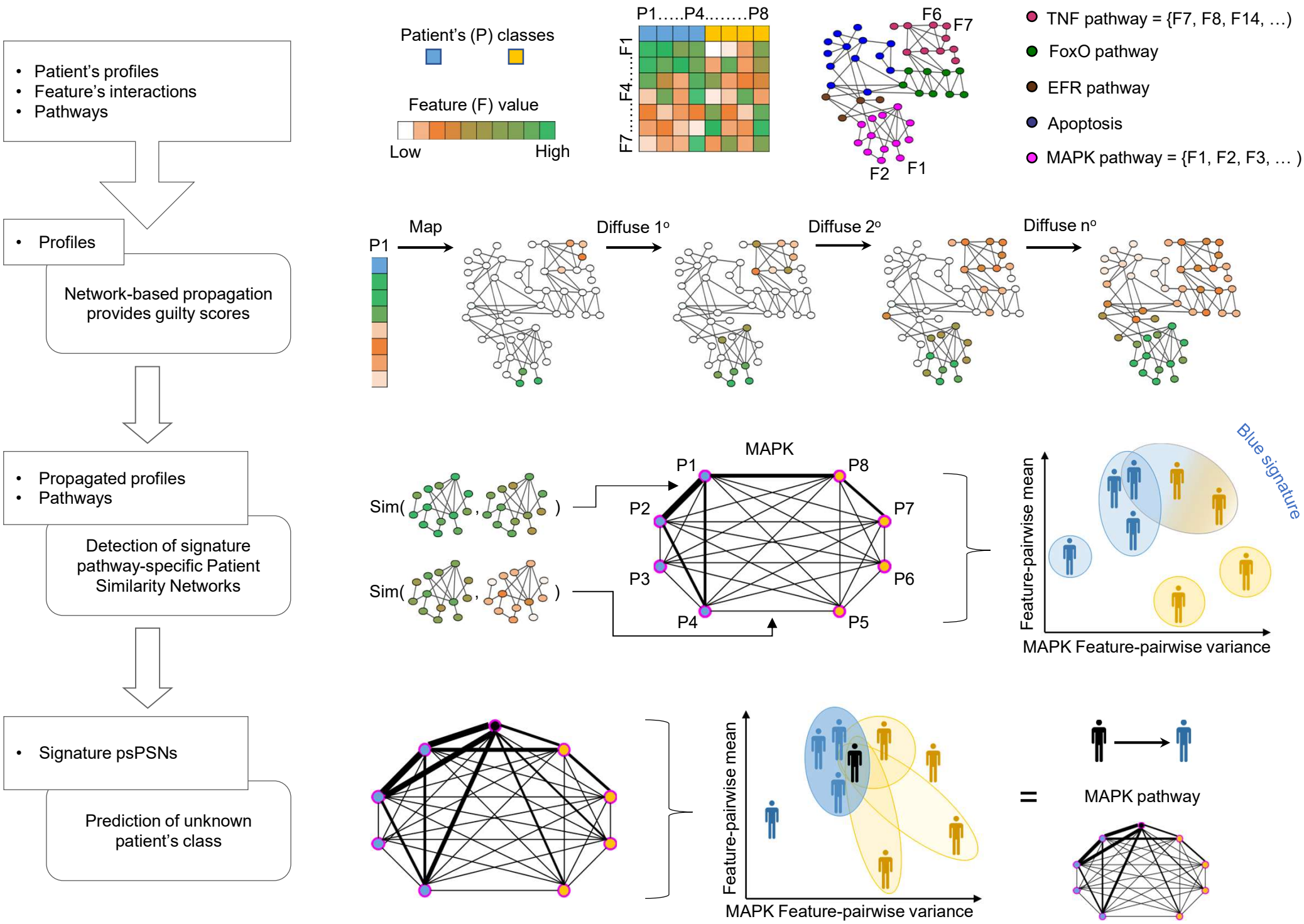
[15] Geraci F, Saha I, Bianchini M. Editorial: RNA-Seq Analysis: Methods, Applications and Challenges. Front Genet 2020;11. https://doi.org/10.3389/fgene.2020.00220.

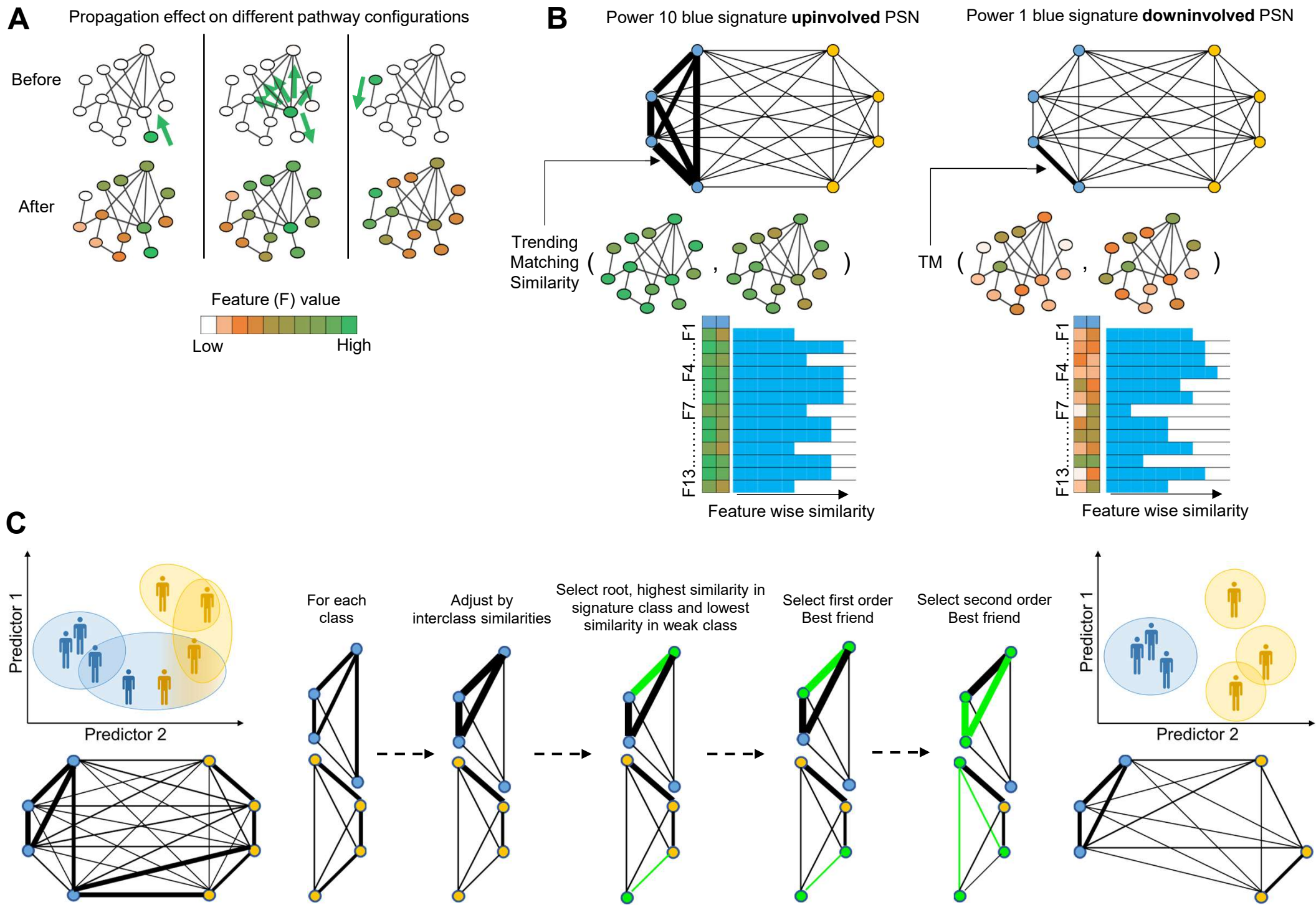
[16] Koch CM, Chiu SF, Akbarpour M, Bharat A, Ridge KM, Bartom ET, et al. A Beginner's Guide to Analysis of RNA Sequencing Data. Am J Respir Cell Mol Biol 2018;59:145–57. https://doi.org/10.1165/rcmb.2017-0430TR.
[17] Fabris F, Palmer D, de Magalhães JP, Freitas AA. Comparing enrichment analysis and machine learning for identifying gene properties that discriminate between gene classes. Brief Bioinform 2020;21:803–14. https://doi.org/10.1093/bib/bbz028.
[18] Pang H, Datta D, Zhao H. Pathway analysis using random forests with bivariate node-split for survival outcomes. Bioinformatics 2010;26:250–8. https://doi.org/10.1093/bioinformatics/btp640.
[19] Hao J, Kim Y, Kim T-K, Kang M. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. BMC Bioinformatics 2018;19:510. https://doi.org/10.1186/s12859-018-2500-z.
[20] Pai S, Bader GD. Patient Similarity Networks for Precision Medicine. J Mol Biol 2018;430:2924–38. https://doi.org/10.1016/j.jmb.2018.05.037.
[21] Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods 2014;11:333–7. https://doi.org/10.1038/nmeth.2810.
[22] Li L, Cheng W-Y, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science Translational Medicine 2015;7:311ra174-311ra174. https://doi.org/10.1126/scitranslmed.aaa9364.
[23] 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 Biology 2008;9:S4. https://doi.org/10.1186/gb-2008-9-s1-s4.
[24] McGuire AL, Fisher R, Cusenza P, Hudson K, Rothstein MA, McGraw D, et al. Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: points to consider. Genetics in Medicine 2008;10:495–9. https://doi.org/10.1097/GIM.0b013e31817a8aaa.
[25] Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying Personal Genomes by Surname Inference. Science 2013;339:321–4. https://doi.org/10.1126/science.1229566.
[26] Chen F, Wang S, Jiang X, Ding S, Lu Y, Kim J, et al. PRINCESS: Privacy-protecting Rare disease International Network Collaboration via Encryption through Software guard extensionS. Bioinformatics 2017;33:871–8. https://doi.org/10.1093/bioinformatics/btw758.
[27] Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD. netDx: interpretable patient classification using integrated patient similarity networks. Mol Syst Biol 2019;15. https://doi.org/10.15252/msb.20188497.
[28] Pai S, Weber P, Isserlin R, Kaka H, Hui S, Shah MA, et al. netDx: Software for building interpretable patient classifiers by multi-'omic data integration using patient similarity networks. F1000Res 2021;9:1239. https://doi.org/10.12688/f1000research.26429.2.
[29] Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. Nature 2018;553:446–54. https://doi.org/10.1038/nature25183.
[30] Lu C, Bera K, Wang X, Prasanna P, Xu J, Janowczyk A, et al. A prognostic model for overall survival of patients with early-stage non-small cell lung cancer: a multicentre, retrospective study. Lancet Digit Health 2020;2:e594–606. https://doi.org/10.1016/s2589-7500(20)30225-9.
[31] Shridhar V, Lee J, Pandita A, Iturria S, Avula R, Staub J, et al. Genetic analysis of early-versus late-stage ovarian tumors. Cancer Res 2001;61:5895–904.
[32] Promoting Cancer Early Diagnosis n.d. https://www.who.int/activities/promoting-cancer-early-diagnosis (accessed May 31, 2021).
[33] Hanna TP, King WD, Thibodeau S, Jalink M, Paulin GA, Harvey-Jones E, et al. Mortality due to cancer treatment delay: systematic review and meta-analysis. BMJ 2020;371:m4087. https://doi.org/10.1136/bmj.m4087.
[34] Raphael MJ, Biagi JJ, Kong W, Mates M, Booth CM, Mackillop WJ. The relationship between time to initiation of adjuvant chemotherapy and survival in breast cancer: a systematic review and meta-analysis. Breast Cancer Res Treat 2016;160:17–28. https://doi.org/10.1007/s10549-016-3960-3.

[35] Russell B, Liedberg F, Khan MS, Nair R, Thurairaja R, Malde S, et al. A Systematic Review and Meta-analysis of Delay in Radical Cystectomy and the Effect on Survival in Bladder Cancer Patients. Eur Urol Oncol 2020;3:239–49. https://doi.org/10.1016/j.euo.2019.09.008.
[36] Multiomic Integration of Public Oncology Databases in Bioconductor - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/33119407/ (accessed May 31, 2021).
[37] Ramos M, Schiffer L, Davis S, Waldron L. TCGAutils: TCGA utility functions for data management. TCGAutils: TCGA Utility Functions for Data Management 2021.
[38] Rosen RD, Sapra A. TNM Classification. StatPearls, Treasure Island (FL): StatPearls Publishing; 2021.
[39] Barclay ME, Abel GA, Greenberg DC, Rous B, Lyratzopoulos G. Socio-demographic variation in stage at diagnosis of breast, bladder, colon, endometrial, lung, melanoma, prostate, rectal, renal and ovarian cancer in England and its population impact. British Journal of Cancer 2021;124:1320–9. https://doi.org/10.1038/s41416-021-01279-z.
[40] Hu Z-D, Zhou Z-R, Qian S. How to analyze tumor stage data in clinical research. J Thorac Dis 2015;7:566–75. https://doi.org/10.3978/j.issn.2072-1439.2015.04.09.
[41] McCormack V, Aggarwal A. Early cancer diagnosis: reaching targets across whole populations amidst setbacks. British Journal of Cancer 2021;124:1181–2. https://doi.org/10.1038/s41416-021-01276-2.
[42] Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Res 2016;5. https://doi.org/10.12688/f1000research.9005.3.
[43] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11:R25. https://doi.org/10.1186/gb-2010-11-3-r25.
[44] Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. Analytica Chimica Acta 2013;760:25–33. https://doi.org/10.1016/j.aca.2012.11.007.
[45] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102.
[46] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9. https://doi.org/10.1038/75556.
[47] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27.
[48] Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci 2021;30:187–200. https://doi.org/10.1002/pro.3978.
[49] Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics 2017;18:551–62. https://doi.org/10.1038/nrg.2017.38.
[50] Le D-H. Random walk with restart: A powerful network propagation algorithm in Bioinformatics field. 2017 4th NAFOSTED Conference on Information and Computer Science, 2017, p. 242–7. https://doi.org/10.1109/NAFOSTED.2017.8108071.
[51] Ideker T, Sharan R. Protein networks in disease. Genome Res 2008;18:644–52. https://doi.org/10.1101/gr.071852.107.
[52] Vandin F, Clay P, Upfal E, Raphael BJ. Discovery of mutated subnetworks associated with clinical data in cancer. Pac Symp Biocomput 2012:55–66.
[53] Le D-H. Network-based ranking methods for prediction of novel disease associated microRNAs. Comput Biol Chem 2015;58:139–48. https://doi.org/10.1016/j.compbiolchem.2015.07.003.
[54] Shi H, Xu J, Zhang G, Xu L, Li C, Wang L, et al. Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. BMC Syst Biol 2013;7:101. https://doi.org/10.1186/1752-0509-7-101.

[55] Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, et al. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. Mol Biosyst 2014;10:2074–81. https://doi.org/10.1039/c3mb70608g.
[56] Le D-H. A novel method for identifying disease associated protein complexes based on functional similarity protein complex networks. Algorithms Mol Biol 2015;10:14. https://doi.org/10.1186/s13015-015-0044-6.
[57] Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods 2013;10:1108–15. https://doi.org/10.1038/nmeth.2651.
[58] Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. ArXiv:160902907 [Cs, Stat] 2017.
[59] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. ArXiv:171010903 [Cs, Stat] 2018.
[60] Di Nanni N, Bersanelli M, Milanesi L, Mosca E. Network Diffusion Promotes the Integrative Analysis of Multiple Omics. Front Genet 2020;11. https://doi.org/10.3389/fgene.2020.00106.
[61] Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet 2017;8. https://doi.org/10.3389/fgene.2017.00084.
[62] Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics 2016;17:S15. https://doi.org/10.1186/s12859-015-0857-9.
[63] Pak M, Jeong D, Moon JH, Ann H, Hur B, Lee S, et al. Network Propagation for the Analysis of Multi-omics Data. In: Yoon B-J, Qian X, editors. Recent Advances in Biological Network Analysis: Comparative Network Analysis and Network Module Detection, Cham: Springer International Publishing; 2021, p. 185–217. https://doi.org/10.1007/978-3-030-57173-3_9.
[64] Andl CD, Mizushima T, Oyama K, Bowser M, Nakagawa H, Rustgi AK. EGFR-induced cell migration is mediated predominantly by the JAK-STAT pathway in primary esophageal keratinocytes. Am J Physiol Gastrointest Liver Physiol 2004;287:G1227-1237. https://doi.org/10.1152/ajpgi.00253.2004.
[65] Badache A, Hynes NE. Interleukin 6 inhibits proliferation and, in cooperation with an epidermal growth factor receptor autocrine loop, increases migration of T47D breast cancer cells. Cancer Res 2001;61:383–91.
[66] Takahashi-Tezuka M, Yoshida Y, Fukada T, Ohtani T, Yamanaka Y, Nishida K, et al. Gab1 acts as an adapter molecule linking the cytokine receptor gp130 to ERK mitogen-activated protein kinase. Mol Cell Biol 1998;18:4109–17. https://doi.org/10.1128/MCB.18.7.4109.
[67] Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. Biological Psychiatry 2016;80:552–61. https://doi.org/10.1016/j.biopsych.2015.12.023.
[68] Kirk R. Personalized medicine and tumour heterogeneity. Nat Rev Clin Oncol 2012;9:250–250. https://doi.org/10.1038/nrclinonc.2012.46.
[69] Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 2012;366:883–92. https://doi.org/10.1056/NEJMoa1113205.
[70] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 1998;30:107–17. https://doi.org/10.1016/S0169-7552(98)00110-X.
[71] Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Software: Practice and Experience 1991;21:1129–64. https://doi.org/10.1002/spe.4380211102.
[72] Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 2017;45:D833–9. https://doi.org/10.1093/nar/gkw943.
[73] Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science 2017;357. https://doi.org/10.1126/science.aan2507.
[74] Lin L, Yang T, Fang L, Yang J, Yang F, Zhao J. Gene gravity-like algorithm for disease gene prediction based on phenotype-specific network. BMC Systems Biology 2017;11:121. https://doi.org/10.1186/s12918-017-0519-9.

[75] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 2005;38:404–15. https://doi.org/10.1016/j.jbi.2005.02.008.
[76] Pang S, Sun Y, Wu L, Yang L, Zhao Y-L, Wang Z, et al. Reconstruction of kidney renal clear cell carcinoma evolution across pathological stages. Sci Rep 2018;8:3339. https://doi.org/10.1038/s41598-018-20321-4.
[77] Cui H, Shan H, Miao MZ, Jiang Z, Meng Y, Chen R, et al. Identification of the key genes and pathways involved in the tumorigenesis and prognosis of kidney renal clear cell carcinoma. Sci Rep 2020;10:4271. https://doi.org/10.1038/s41598-020-61162-4.
[78] Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics 2009;10:161. https://doi.org/10.1186/1471-2105-10-161.
[79] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation 2021:100141. https://doi.org/10.1016/j.xinn.2021.100141.
[80] Christensen A P. NetworkToolbox: Methods and Measures for Brain, Cognitive, and Psychometric Network Analysis in R. The R Journal 2019;10:422. https://doi.org/10.32614/RJ-2018-065.
[81] Frigyesi A, Höglund M. Non-Negative Matrix Factorization for the Analysis of Complex Gene Expression Data: Identification of Clinically Relevant Tumor Subtypes. Cancer Inform 2008;6:275–92.
[82] Saccenti E, Hendriks MHWB, Smilde AK. Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and correction under different error models. Sci Rep 2020;10:438. https://doi.org/10.1038/s41598-019-57247-4.
[83] Powers S, DeJongh M, Best AA, Tintle NL. Cautions about the reliability of pairwise gene correlations based on expression data. Front Microbiol 2015;6. https://doi.org/10.3389/fmicb.2015.00650.


Patient's profiles
Feature's interactions
Pathways
Profiles
Network-based propagation provides guilty scores
Propagated profiles
Pathways
Detection of signature pathway-specific Patient Similarity Networks
Signature psPSNs
Prediction of unknown patient's class
Patient's (P) classes
Feature (F) value
Low
High
P1.....P4.....P8
F1
F4.....
F7.....
F6
F7
F2
F1
TNF pathway = {F7, F8, F14, …)
FoxO pathway
EFR pathway
Apoptosis
MAPK pathway = {F1, F2, F3, …)
P1
Map
Diffuse 1º
Diffuse 2º
Diffuse nº
Sim(
)
MAPK
P1
P2
P3
P4
P5
P6
P7
P8
Blue signature
Feature-pairwise mean
MAPK Feature-pairwise variance
Feature-pairwise mean
MAPK Feature-pairwise variance
=
MAPK pathway


A
Propagation effect on different pathway configurations
Before
After
Feature (F) value
Low
High
B
Power 10 blue signature upinvolved PSN
Power 1 blue signature downinvolved PSN
Trending Matching Similarity
TM
F1
F4
F7
F13
Feature wise similarity
Feature wise similarity
C
Predictor 1
Predictor 2
For each class
Adjust by interclass similarities
Select root, highest similarity in signature class and lowest similarity in weak class
Select first order Best friend
Select second order Best friend
Predictor 1
Predictor 2

A

BFC(PSN_u,th=5){

For each b,k∈{1,…,B} such that b,k∈{EARLY}

Determine r' and r''; Q=[0,0] E=[0,$w''_{b,k}$] R=[0,r'] T=[$w''_{b,k}$, r'']; $w''_{b,k}$ = QA(P,Q,T,Z)

For each b,k∈{1,…,B} such that b,k∈{LATE}

Determine r' and r''; Q=[0,0] E=[0,$w''_{b,k}$] R=[0,r'] T=[$w''_{b,k}$, r'']: $w''_{b,k}$ = QA(P,Q,T,Z)

For each b∈{1,…,B}

Determine $1BF_b$ and $2BF_b$; Max=0

For each b∈{EARLY}

Determine BFS and Max_b

If Max_b>Max then Max=Max_b and PV''=BFS

Max=0

For each b∈{LATE}

Determine BFS and Max_b

If Max_b>Max then Max=Max_b and PV'''=BFS

PV' ={PV'' U PV'''}

Return PSN'_u(PV',PE')=BFC(PSN_u,x) with PV'⊆PV and a set of edges PE'⊆PE each one of which is incident with vertices from PV′ only

}

B

For each $u \in \{1, \ldots, C\}$

For each $x \in T = (5 * n)_{n=1}^{95}$

M1'=M2'=M1''=M2''=0 and th=95

Determine $1BF_z$ assuming z ∈ EARLY

$$M1' = \frac{\sum_{k \in \{1BF_z\}} W'''_{k'z}}{|1BF_z}$$

$PSN'_u(PV', PE') = BFC(PSN_u, x)$ with $PV' \subseteq PV$ and a set of edges $PE' \subseteq PE$ each one of which is incident with vertices from PV′ only

If $pv_z \in PV'$ then $M2' = x$

Determine $1BF_z$ assuming z ∈ LATE

$$M1'' = \frac{\sum_{k \in \{1BF_z\}} W'''_{k'z}}{|1BF_z}$$

$PSN'_u(PV', PE') = BFC(PSN_u, x)$ with $PV' \subseteq PV$ and a set of edges $PE' \subseteq PE$ each one of which is incident with vertices from PV′ only

If $pv_z \in PV'$ then $M2'' = x$

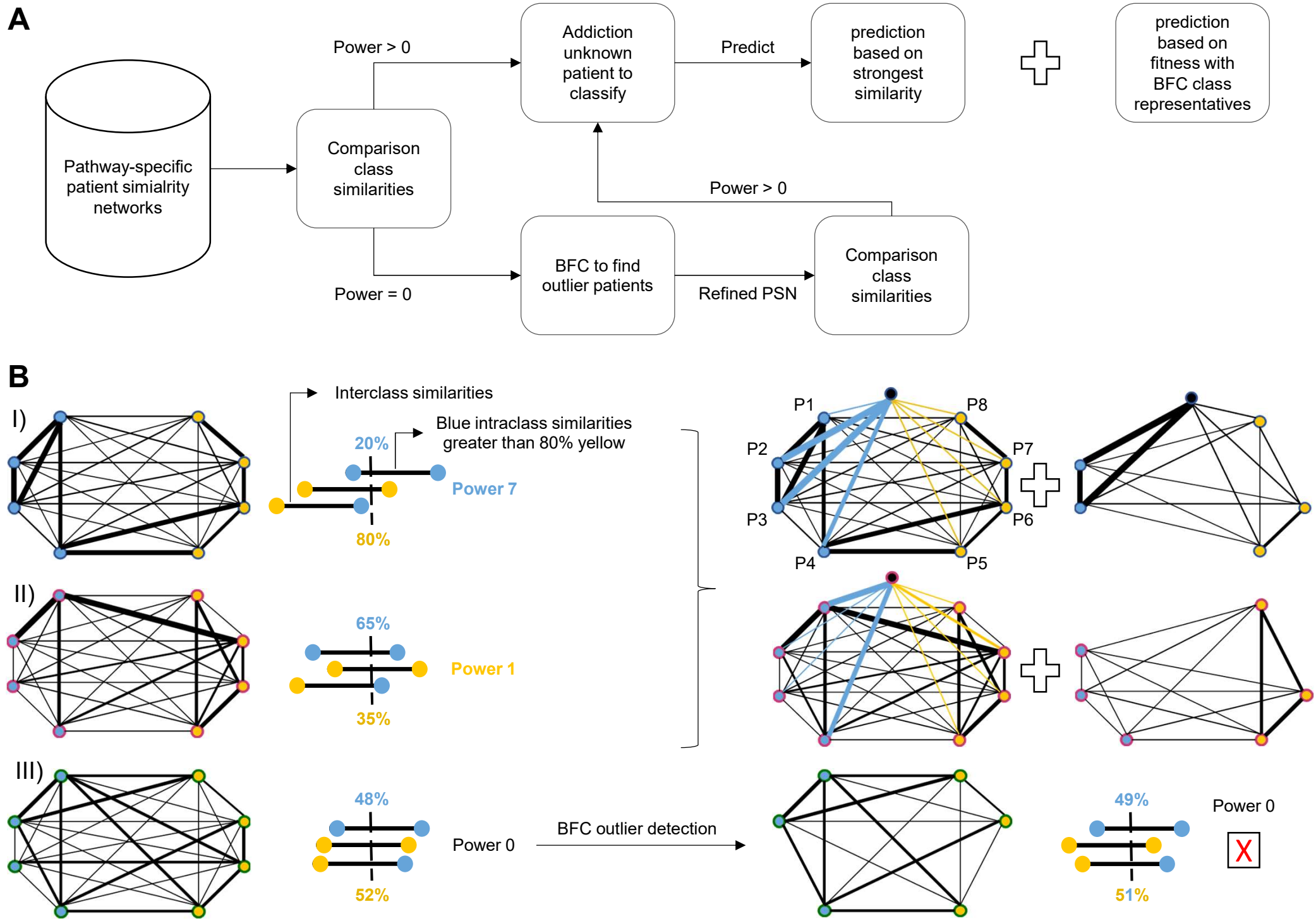
If $M2' > M2''$ AND $M1' > M1''$ then

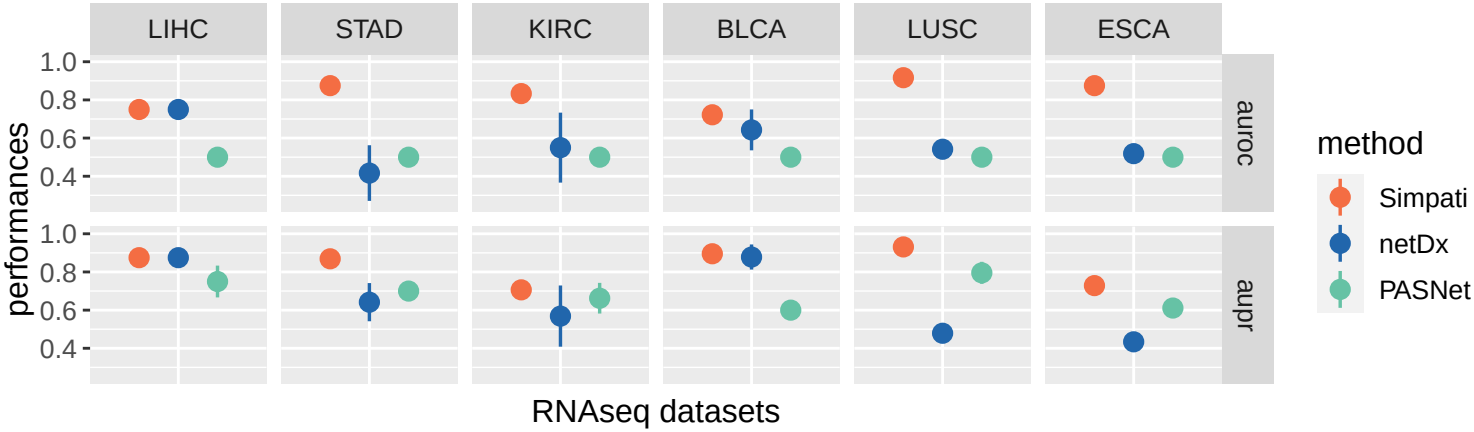
Add EARLY to the LABELS set

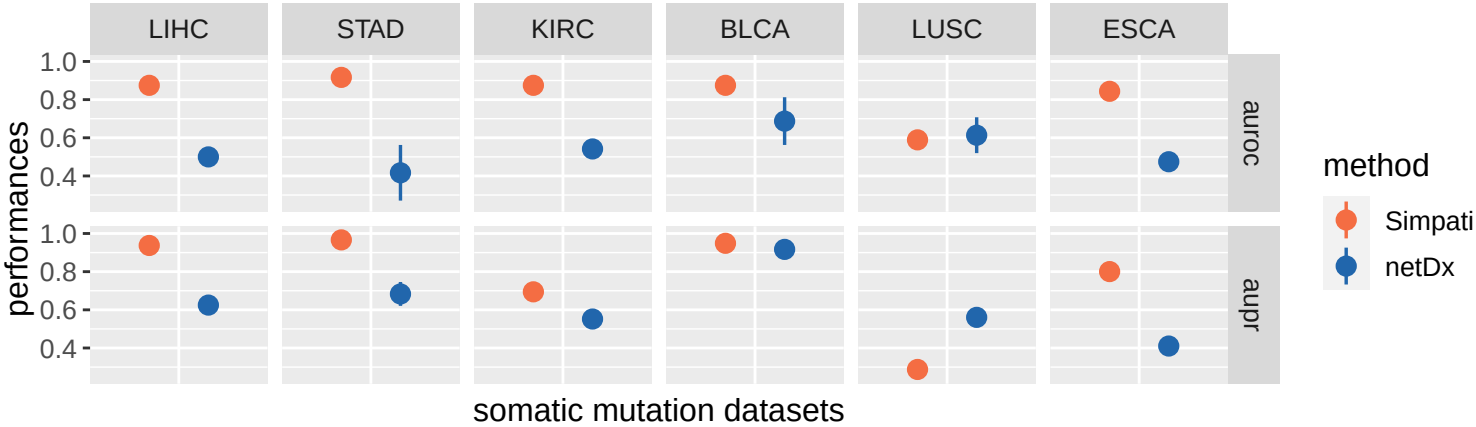
elseIf $M2'' > M2'$ AND $M1'' > M1'$ then

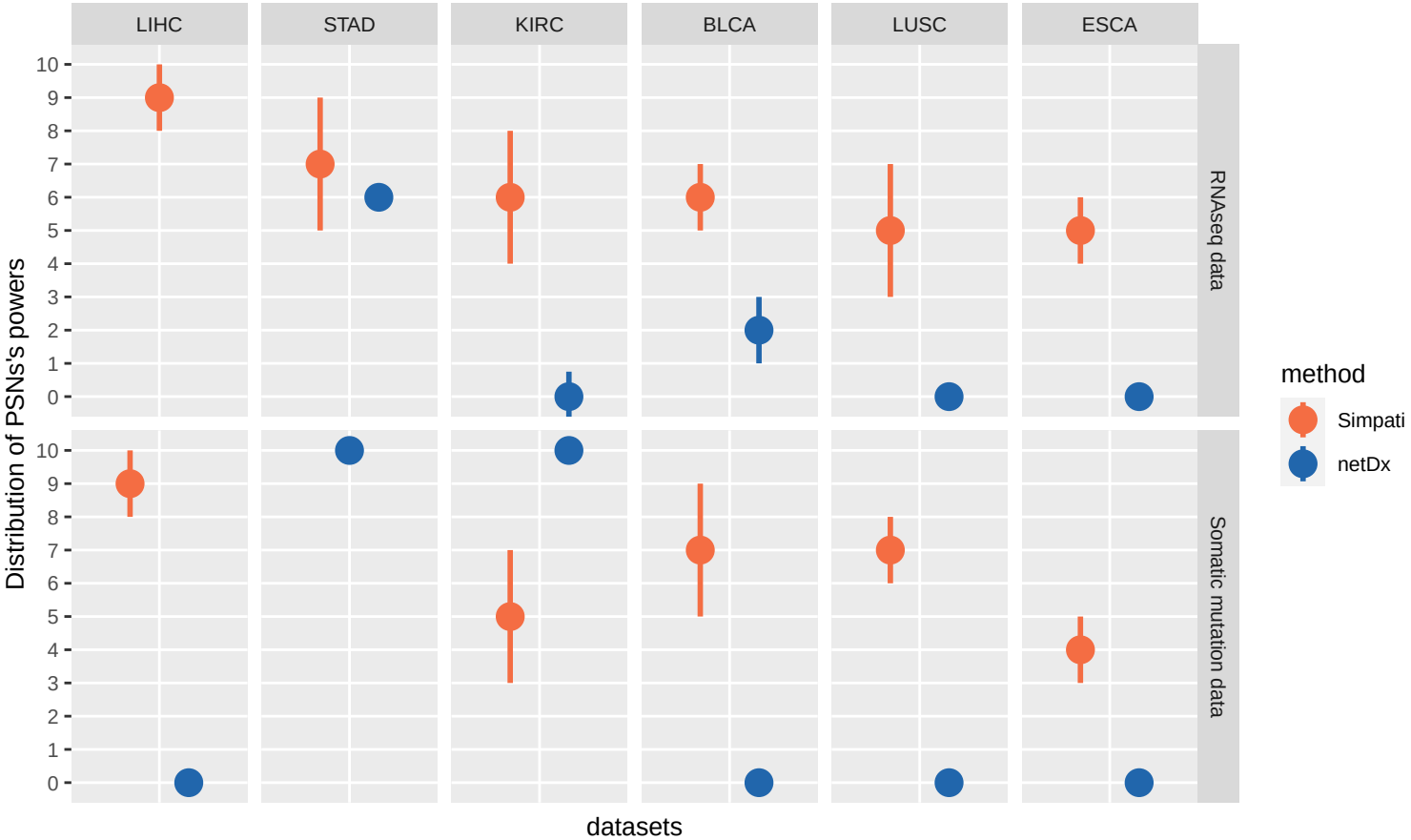
Add LATE to the LABELS set

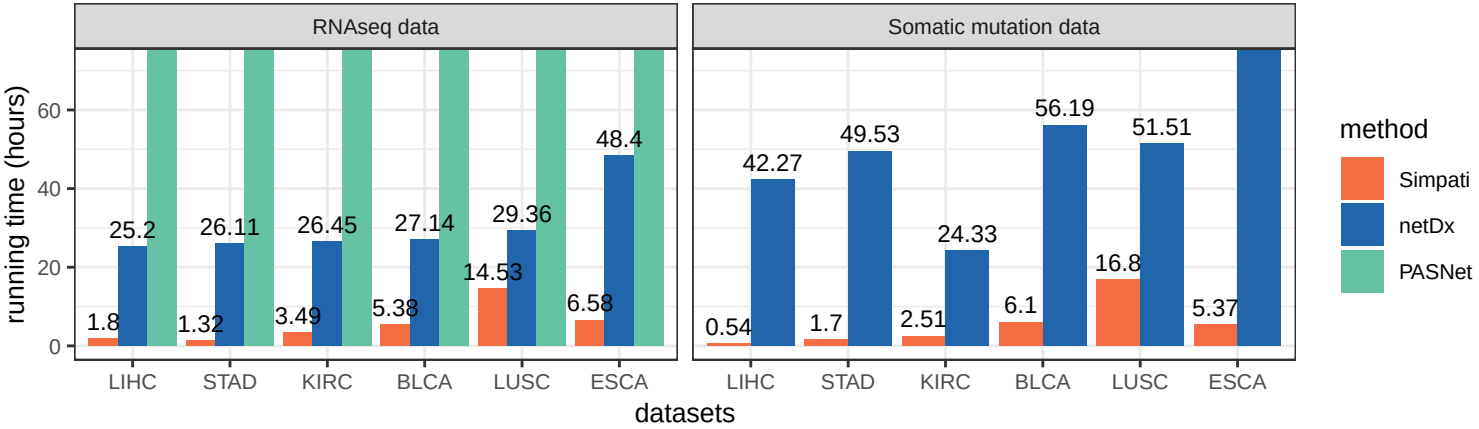
$$\text{Consensus_class} = \text{argmax}_a |a \in \text{LABELS}|$$

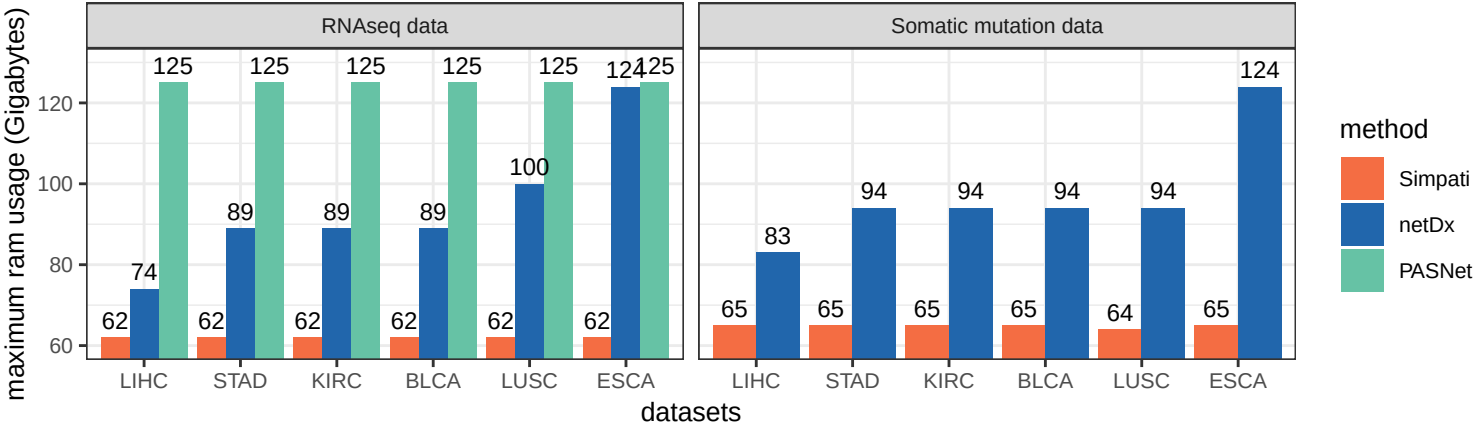

A
Pathway-specific patient simialrity networks
Comparison class similarities
Power > 0
Addiction unknown patient to classify
Predict
prediction based on strongest similarity
prediction based on fitness with BFC class representatives
Power = 0
BFC to find outlier patients
Refined PSN
Comparison class similarities
Power > 0
B
I)
Interclass similarities
Blue intraclass similarities greater than 80% yellow
20%
80%
Power 7
P1
P2
P3
P4
P5
P6
P7
P8
II)
65%
35%
Power 1
III)
48%
52%
Power 0
BFC outlier detection
49%
51%
Power 0
X

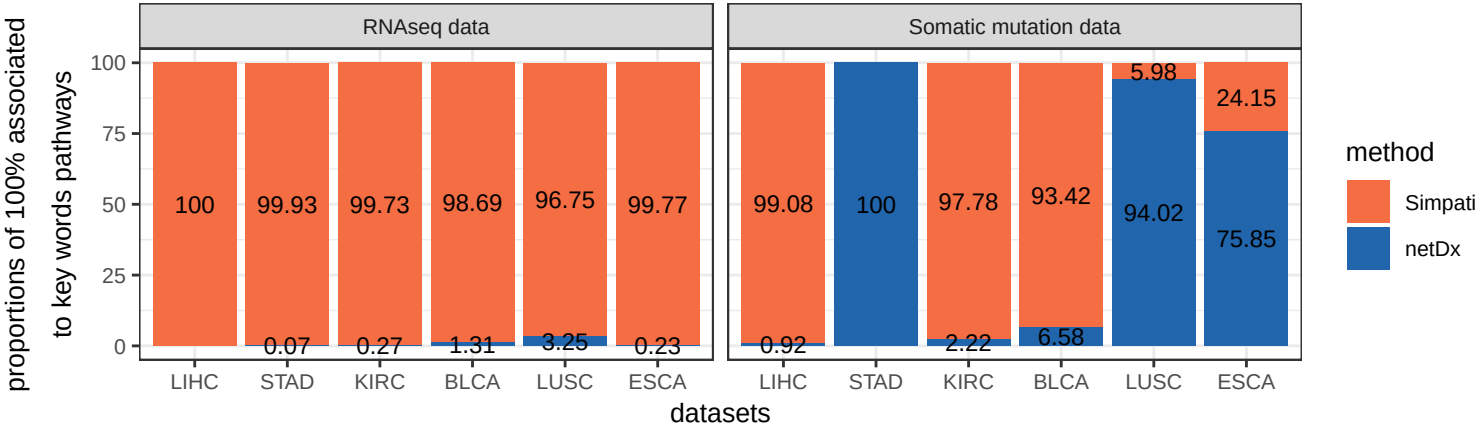

LIHC
STAD
KIRC
BLCA
LUSC
ESCA
auroc
aupr
performances
1.0
0.8
0.6
0.4
RNAseq datasets
method
Simpati
netDx
PASNet

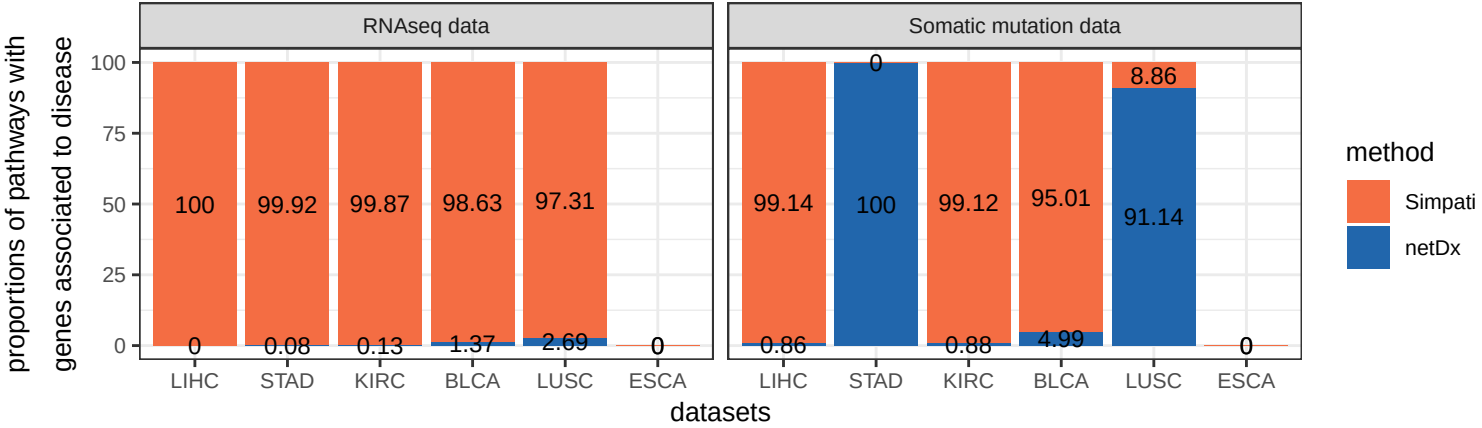

LIHC
STAD
KIRC
BLCA
LUSC
ESCA
auroc
aupr
performances
1.0
0.8
0.6
0.4
somatic mutation datasets
method
Simpati
netDx

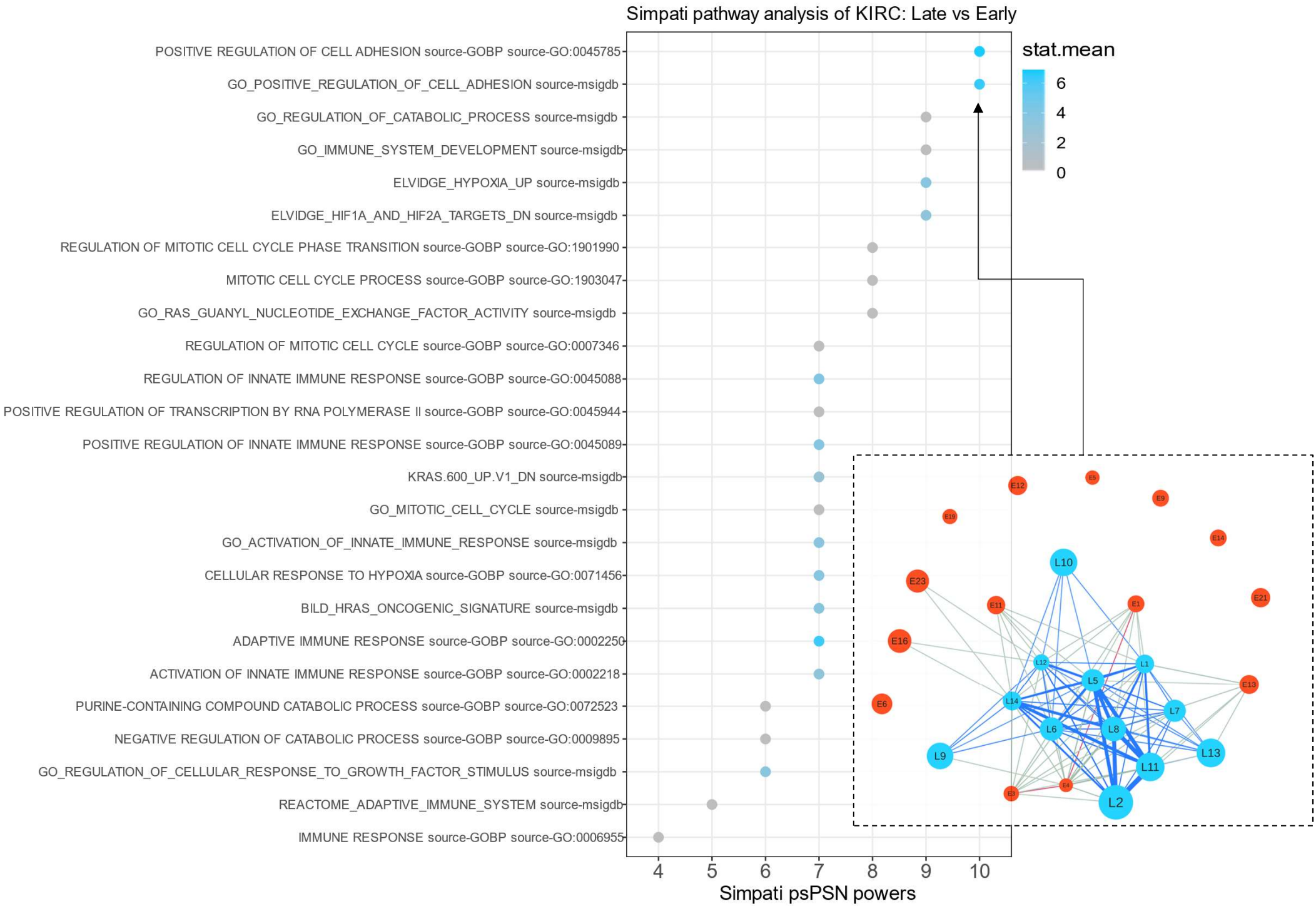

LIHC
STAD
KIRC
BLCA
LUSC
ESCA
RNAseq data
Somatic mutation data
Distribution of PSNs's powers
datasets
method
Simpati
netDx


RNAseq data
Somatic mutation data
running time (hours)
0
20
40
60
1.8
25.2
1.32
26.11
3.49
26.45
5.38
27.14
14.53
29.36
6.58
48.4
0.54
42.27
1.7
49.53
2.51
24.33
6.1
56.19
16.8
51.51
5.37
LIHC
STAD
KIRC
BLCA
LUSC
ESCA
datasets
method
Simpati
netDx
PASNet


RNAseq data
Somatic mutation data
maximum ram usage (Gigabytes)
60
80
100
120
62
74
125
62
89
125
62
89
125
62
89
125
62
100
125
62
124
125
65
83
124
65
94
65
94
65
94
64
94
65
LIHC
STAD
KIRC
BLCA
LUSC
ESCA
datasets
method
Simpati
netDx
PASNet


RNAseq data
Somatic mutation data
proportions of 100% associated to key words pathways
100
75
50
25
0
100
99.93
99.73
98.69
96.75
99.77
0.07
0.27
1.31
3.25
0.23
99.08
100
97.78
93.42
5.98
24.15
94.02
75.85
0.92
2.22
6.58
LIHC
STAD
KIRC
BLCA
LUSC
ESCA
datasets
method
Simpati
netDx


RNAseq data
Somatic mutation data
proportions of pathways with genes associated to disease
100
75
50
25
0
100
99.92
99.87
98.63
97.31
0
0.08
0.13
1.37
2.69
0
99.14
100
99.12
95.01
8.86
91.14
0
0.86
0.88
4.99
0
LIHC
STAD
KIRC
BLCA
LUSC
ESCA
datasets
method
Simpati
netDx


Simpati pathway analysis of KIRC: Late vs Early
POSITIVE REGULATION OF CELL ADHESION source-GOBP source-GO:0045785
GO_POSITIVE_REGULATION_OF_CELL_ADHESION source-msigdb
GO_REGULATION_OF_CATABOLIC_PROCESS source-msigdb
GO_IMMUNE_SYSTEM_DEVELOPMENT source-msigdb
ELVIDGE_HYPOXIA_UP source-msigdb
ELVIDGE_HIF1A_AND_HIF2A_TARGETS_DN source-msigdb
REGULATION OF MITOTIC CELL CYCLE PHASE TRANSITION source-GOBP source-GO:1901990
MITOTIC CELL CYCLE PROCESS source-GOBP source-GO:1903047
GO_RAS_GUANYL_NUCLEOTIDE_EXCHANGE_FACTOR_ACTIVITY source-msigdb
REGULATION OF MITOTIC CELL CYCLE source-GOBP source-GO:0007346
REGULATION OF INNATE IMMUNE RESPONSE source-GOBP source-GO:0045088
POSITIVE REGULATION OF TRANSCRIPTION BY RNA POLYMERASE II source-GOBP source-GO:0045944
POSITIVE REGULATION OF INNATE IMMUNE RESPONSE source-GOBP source-GO:0045089
KRAS.600_UP.V1_DN source-msigdb
GO_MITOTIC_CELL_CYCLE source-msigdb
GO_ACTIVATION_OF_INNATE_IMMUNE_RESPONSE source-msigdb
CELLULAR RESPONSE TO HYPOXIA source-GOBP source-GO:0071456
BILD_HRAS_ONCOGENIC_SIGNATURE source-msigdb
ADAPTIVE IMMUNE RESPONSE source-GOBP source-GO:0002250
ACTIVATION OF INNATE IMMUNE RESPONSE source-GOBP source-GO:0002218
PURINE-CONTAINING COMPOUND CATABOLIC PROCESS source-GOBP source-GO:0072523
NEGATIVE REGULATION OF CATABOLIC PROCESS source-GOBP source-GO:0009895
GO_REGULATION_OF_CELLULAR_RESPONSE_TO_GROWTH_FACTOR_STIMULUS source-msigdb
REACTOME_ADAPTIVE_IMMUNE_SYSTEM source-msigdb
IMMUNE RESPONSE source-GOBP source-GO:0006955
4
5
6
7
8
9
10
Simpati psPSN powers
stat.mean
6
4
2
0
E12
E5
E9
E10
E14
L10
E23
E11
E1
E21
E16
L12
L1
L5
E13
E6
L14
L7
L6
L8
L9
L13
L11
E3
E4
L2