De Novo and Supervised Endophenotyping Using Network-Guided Ensemble Learning

Systems Medicine
Volume 3.1, 2020
DOI: 10.1089/sysm.2019.0008

SYSTEMS MEDICINE
Journal of Medical Systems Biology and Network Medicine

Mary Ann Liebert, Inc. publishers

ORIGINAL RESEARCH Open Access

De Novo and Supervised Endophenotyping Using Network-Guided Ensemble Learning

Simon J. Larsen,[1,*] Harald H.H.W. Schmidt,[2] and Jan Baumbach[1,3]

Abstract

Introduction: Precision medicine requires the accurate identification of genes and pathways that mechanistically define a disease phenotype. Modern omics may deliver this, but has until now yielded only few translational successes. While gene signatures derived from single omics analysis have proven useful for disease diagnosis and prognosis, they often do not explain the underlying mechanism.

Methods: We here present Grand Forest, an ensemble learning method that extends random forests and integrates experimental data with molecular interaction networks to discover relevant endophenotypes and their defining gene modules. Our method covers two application scenarios: a supervised method for finding modules associated with outcome and an unsupervised method for finding *de novo* patient subgroups.

Results: We applied the supervised Grand Forest methodology to five disease-related transcriptome data sets and compared the results with four state-of-the-art methods. Grand Forest consistently found gene modules with greater biomedical relevance, reproducibility, and interaction density, but fewer differentially expressed genes. Using the unsupervised method to discover gene modules from unlabeled data, lung cancer patients could be *de novo* stratified into clinically relevant molecular subgroups. Further analysis revealed that known disease genes were only marginally over-represented among differentially expressed genes, and that our method was driven mainly by network topology.

Conclusion: With Grand Forest, we developed a novel approach to disease module discovery and demonstrated it identifies biologically relevant gene modules and patient subgroups. We conclude that differential expression was not effective for identifying driving genes and that the results were likely confounded by bias in the network data. We caution readers to consider these issues when applying network-based methods to gene expression analysis. Grand Forest is available at https://grandforest.compbio.sdu.dk.

Keywords: gene expression; network analysis; machine learning; module discovery; endophenotyping

Introduction

The increasingly large amounts of functional genomic data currently available in public databases such as the Gene Expression Omnibus (GEO) and through extensive data collection efforts such as The Cancer Genome Atlas have enabled large-scale integrative analyses aiming to discover mutations and expression patterns associated with a specific disease. A key aim in precision medicine has been the identification of molecular subtypes from molecular profiling data. By classifying patients as different subtypes, the aim is to stratify patients into groups with distinct clinical traits, such as expected survival time, risk of disease recurrence, or response to treatment. To this end, significant effort

[1]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark.
[2]Department of Pharmacology and Personalised Medicine, Faculty of Health, Medicine and Life Science, Maastricht University, Maastricht, The Netherlands.
[3]Chair of Experimental Bioinformatics, Wissenschaftszentrum Weihenstephan, Technical University of Munich, Freising, Germany.

*Address correspondence to: Simon J. Larsen, PhD, Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, Odense 5230, Denmark, E-mail: sjlarsen@imada.sdu.dk

has been put into identification of gene signatures—small sets of genes that exhibit a distinct expression or mutation pattern associated with a specific phenotype.[1–4] Despite proving useful for prognosis, different breast cancer signatures have little overlap in genes and have been shown to be inconsistent across data sets.[5] Furthermore, most random gene signatures of 100 or more genes were found to be significantly associated with outcomes in breast cancer, despite having no relationship with the disease itself.[6] This demonstrates a major limitation of gene expression-based analysis: a change in phenotype may lead to gross global changes in the transcriptome, and thus, the genes that are best suited for distinguishing different symptoms or outcomes are not necessarily important for development or progression of the disease itself.

To cope with the inherently noisy and overdetermined nature of molecular profiling data, many researchers have proposed integrating experimental data with secondary data, in the form of biological interaction networks, to produce more stable and biologically meaningful models. This is commonly achieved either through searching for functional enrichment in known pathways[7,8] or finding enriched gene modules in global interaction networks (*de novo* pathways).[9–14] The latter approach is especially promising as it may help uncover previously unknown molecular interactions and mechanisms not currently reported in databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome. Machine learning methods are also increasingly often utilized to develop more sophisticated models from biological data, for instance, in analyzing expression patterns,[15] regulatory network inference,[16] protein–protein interaction (PPI) prediction,[17] and elucidating genotype–phenotype relationships.[18,19]

In this work, we present a novel kind of module discovery method called Grand Forest (graph-guided random forest). Besides the experimental data to be analyzed (e.g., gene expression or methylation data), our method also integrates a network describing the pairwise relationship between the features comprising the data set (e.g., PPIs). When building the decision tree forest, a connected subnetwork is randomly sampled from the full network for each decision tree, and each decision tree is allowed to use only the features contained in its corresponding subnetwork. Furthermore, each tree is built under the constraint that each split variable in the tree must be a neighbor of the variable in the split directly above it in the decision tree. This constraint enforces that the set of variables in each decision tree form a connected subnetwork in the interaction network and that each split always follows a split on an adjacent gene. By estimating feature importance from the trained model, we are then able to extract a highly connected gene set that explains the phenotype. The subnetwork induced by the most important genes is then extracted and returned as result. We introduce two application scenarios: a supervised and an unsupervised analysis workflow (Fig. 1). Our method extends significantly on previous ideas from Dutkowski and Ideker[20] (see Supplementary Text S1 for details).

We first apply the supervised Grand Forest method to whole-genome gene expression data from patients diagnosed with breast cancer, lung cancer, Huntington's disease, ulcerative colitis, and amyotrophic lateral sclerosis (ALS) and show that our method is able to discover subnetwork modules with greater biological relevance than other existing, network-based, disease-gene module detection tools while also being less sensitive to the sampling of the patient population. We then demonstrate that our method can also be applied to unsupervised endophenotyping, applying it to analyze a lung cancer data set. Unlike most other module discovery tools, Grand Forest does not employ statistical hypothesis tests or differential expression analysis to score the individual genes and, as such, does not make any assumptions on the underlying distribution of the expression data. The use of decision trees may also make it possible to discover interaction effects between genes. Furthermore, Grand Forest can be applied directly to both categorical and numerical clinical variables, as well as right-censored survival data. Hence, its supervised version can—in addition to classification—also be utilized for network module-based regression as well as survival analysis, which makes it, to our knowledge, the first such tool available. In addition, it is the first method supporting unsupervised (i.e., *de novo*) stratification of patients into groups while simultaneously extracting subnetworks whose genetic expression explains the difference between the identified groups.

Comparison of the modules reported by each method revealed that Grand Forest generally selected modules with high interaction density, but lower differential expression, compared with other methods, suggesting that disease-associated genes were selected mainly due to network topology. Furthermore, known disease-associated genes were observed to be only

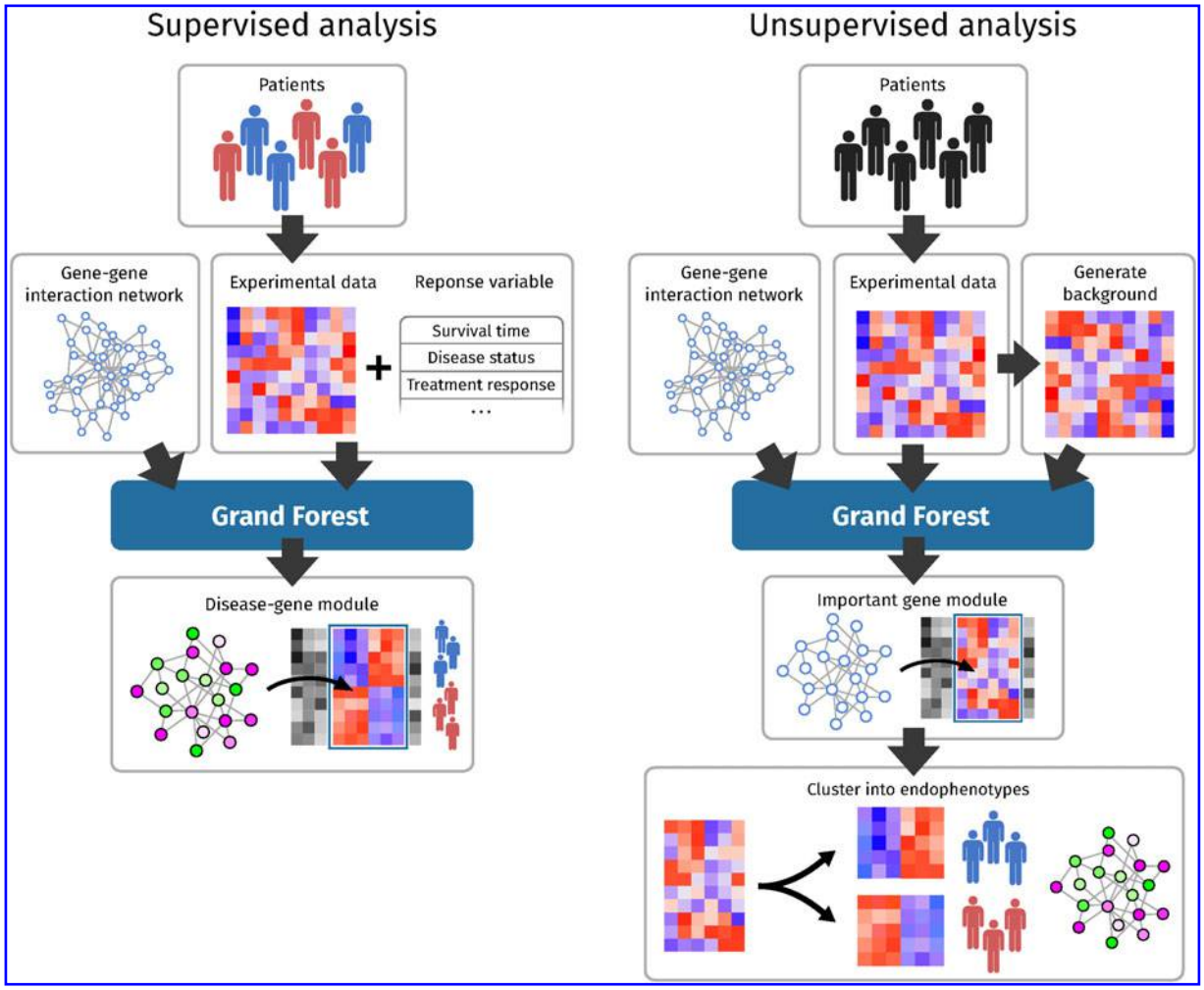

FIG. 1. Overview of the supervised and unsupervised Grand Forest workflows. In the supervised workflow, expression data are integrated with an interaction network to identify a gene module associated with a response variable, that is, survival time. In the unsupervised workflow, a model is trained to recognize unlabeled patients from a generated background distribution. From the trained model, highly informative genes are then selected and used to stratify patients into groups with different endophenotypes.

marginally over-represented among differentially expressed genes. We conclude that one should exercise caution when applying network-based methods for identifying disease gene modules from gene expression data and be aware of the limitations of gene expression analysis as well as possible biases in molecular interaction networks.

Grand Forest is freely available at https://grandforest.compbio.sdu.dk where we provide the source code, a package for the R programming language, and an easy-to-use online analysis platform.

Methods

Graph-guided random forest algorithm

Random forest is an ensemble learning method that works by generating a large ensemble of decision trees.[21,22] It is based on the random decision forests method,[23] but extended to use the random subspace method (also known as feature bagging). It has achieved widespread use in biomedical research as it works well for data sets with many more features than samples, can be applied to data with a mix of continuous and categorical variables, and works for

multiclass problems. Furthermore, it provides several measures for estimating feature importance, making it possible to identify important genes in molecular profiling data.[24,25] We refer to the original articles by Breiman and Cutler for a description of the random forest algorithm.[21,22]

The Grand Forest algorithm works similarly to the random forest algorithm, but differs in the way split variables are selected during decision tree building. Grand Forest takes as input a design matrix $\mathbf{X} = \{x_{i,1}, \ldots, x_{i,p}\}_{i=1}^{n}$, response variables $Y = y_1, \ldots, y_n$, and a simple graph $G = (V, E)$, where each vertex $v_i \in V$ corresponds to the column i in $\mathbf{X}$ and E is a set of two sets of vertices from V. The algorithm builds a forest of decision trees on the training set, using the graph G to guide the feature bagging procedure and split variable selection. The graph is only used during training and does not affect the prediction procedure, which is carried out like it would in the standard random forest algorithm. Each decision tree is trained on a random sample with replacement of the training data. The algorithm is outlined in Figure 2.

Feature bagging

Grand Forest uses the topology of the feature interaction graph G to perform feature bagging. Each decision tree is trained on a subset of m response variables, where each set of variables induces a connected subgraph in G. This subgraph is computed only once for each decision tree and used in all splits in that tree. Each subgraph is generated by first selecting a vertex v_s uniformly at random from all vertices. A subgraph is then grown by performing a breadth-first search traversal starting at v_s until m vertices have been selected or until there are no more vertices to visit. When a new vertex is visited, its neighbors are added to the queue in random order to further randomize the sampling. See Supplementary Text S2.1 for a detailed overview.

Split variable selection

When building decision trees, splits are formed by selecting a variable and a value to split the partition on, which maximizes some split criteria, for example, the decrease in Gini impurity for classification forests. The first split in each decision tree is selected among all features in the feature subgraph. In subsequent splits, each split variable must be selected only from the variables that are connected to the parent node in the decision tree (Supplementary Fig. S1). This

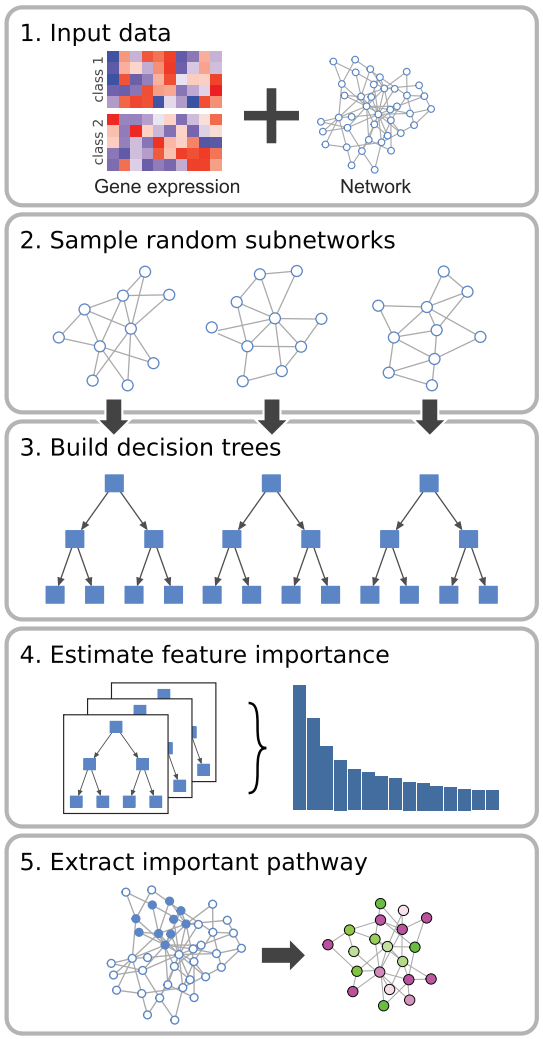

FIG. 2. Overview of the Grand Forest algorithm. (1) The method takes as input a matrix of expression values for two or more classes and a network modeling the pairwise interactions between genes. (2) A large number of connected subnetworks are randomly sampled from the full network. (3) A decision tree is trained for each subnetwork under the constraint that genes in two adjacent splits in the decision tree must also be adjacent in the subnetwork. (4) Global feature importance is computed as mean estimated importance over all decision trees. (5) The subnetwork induced by the most important features is extracted as a disease module.

requirement ensures that the set of variables in each decision tree induces a connected subgraph in the full feature network.

Feature importance

In Grand Forest, we use the mean decrease in Gini impurity to estimate feature importance. Gini impurity was chosen over other methods, such as permutation importance, because we are not concerned with how important a gene is for predictive performance, but rather how much information it provides at the time of the split, conditioned on the splits preceding it.

Implementation

Our implementation of Grand Forest is based on ranger[26] and is written in C++ with bindings to R. The feature graph is only used when selecting features as possible split variables and does not affect the splitting procedure itself. Because of this, it is trivial to generalize the method to other variations of random forest. Besides random forest for classification, ranger also implements regression forests,[21] probability forests,[27] and survival forests.[28] By extension, Grand Forest has been implemented to support these methods as well. The source code is available through GitHub (https://github.com/SimonLarsen/grandforest).

Unsupervised analysis using Grand Forest

The Grand Forest algorithm is used for unsupervised learning using an approach proposed by Breiman.[22] The method is based on the following assumption: if the data item is structured in some way, it should be distinguishable from a randomized version of itself. Given a design matrix $\mathbf{X}$, we compute a synthetic matrix $\mathbf{X}'$ with the same number of rows and columns by randomly sampling values for the corresponding variable in $\mathbf{X}$. Sampling can be done either with or without replacement. In this work, we sampled with replacement. A combined design matrix $\mathbf{X}^*$ is then built by concatenating the rows from $\mathbf{X}$ and $\mathbf{X}'$, and the vector of response variables is defined as $Y = \{y_i\}_{i=1}^{2n}$, where $y_i = 1$ if row i came from $\mathbf{X}$ and $y_i = 0$ otherwise.

A Grand Forest model is trained on the design matrix X^* and response variables Y guided by some graph G. The most important features are selected by ranking all features based on some importance measure and selecting all features above some cutoff. These features are assumed to contain a high amount of information and are thus good for clustering the data set into clusters. The final clustering is performed by clustering the original design matrix $\mathbf{X}$ based only on the top features identified by the Grand Forest model.

Gene expression data preparation

Gene expression data sets were obtained through the GEO. The data sets are available through the following accession IDs: breast cancer (GSE20685, $n = 327$),[29] non-small cell lung cancer (GSE30219, $n = 268$),[30] ulcerative colitis (GSE11223, $n = 202$),[31] Huntington's disease (GSE3790, $n = 54$),[32] and ALS (GSE112680, $n = 164$).[33] All five datasets are from microarrays (Table 1 and Supplementary Table S1). Processed, probe-level expression values were obtained as series matrix files. Probes were mapped to NCBI Entrez gene IDs using the corresponding platform data tables provided through GEO. For genes mapping to multiple probes, the median probe value was used.

The lung cancer data set contained samples from both small cell and non-small cell patients. Only non-small cell cancer samples were used because their molecular pathways (as described in KEGG) are different, and non-small cell was the most common type in the data set (Supplementary Table S2). The Huntington's disease data set contained samples from different brain regions. Only samples from the caudate nucleus were used

Table 1. Data Sets Used in the Evaluation

				Survival			
Cohort	Ref.	Platform	Samples	Median follow-up	Deaths	Cases	Controls
ALS	33	GPL10558	164	2.34 years	31	21[a]	23[a]
Breast cancer	29	GPL570	327	8.1years	83	79[a]	79[a]
Lung cancer	30	GPL570	268	4.67 years	177	125[a]	125[a]
Ulcerative colitis	31	GPL1708	202	N/A	N/A	129	73
Huntington's	32	GPL96	53	N/A	N/A	22	32

Platform column contains the GEO platform ID.
[a]These case/control labels are used only by GXNA due to not supporting noncategorical outcomes.
ALS, amyotrophic lateral sclerosis; GEO, Gene Expression Omnibus; GXNA, Gene eXpression Network Analysis; N/A, not applicable.

because this region was found to have the largest change in gene expression resulting from Huntington's disease in the original publication by Hodges et al.

For ALS, breast cancer, and lung cancer data sets, we used the right-censored survival time as the outcome variable. For the ulcerative colitis (UC) data set, we used the disease status (UC or no UC) as the outcome variable. For the Huntington's disease (HD) data set, we separated the subjects into a control group (no HD) and a case group (Vonsattel grades 2–4). Samples with Vonsattel grade 0–1 were discarded.

The Gene eXpression Network Analysis (GXNA) method requires samples to be stratified into discrete classes. For ulcerative colitis and Huntington's disease, we used the classes described above. For the survival data sets, patients were stratified into high- and low-risk groups of approximately equal size. We used a cut-off of 62 months in lung cancer and 10.6 years in breast cancer and 4 years in ALS. Patients who could not be placed in either group due to censoring or lack of follow-up were discarded. See Table 1 for an overview.

Statistical significance tests

For the survival data sets (breast and lung cancers), the statistical significance of each gene was computed using a Cox proportional hazards regression model. For the regression (Alzheimer's disease) and classification data sets (ulcerative colitis and Huntington's disease), significance was estimated using a linear model with the R/Bioconductor package limma.[34]

Network data preparation

We collected network data from the Integrated Interactions Database (IID)[35] (version 2017-04). IID integrates experimentally validated PPIs from multiple major databases, such as BioGRID, IntAct, and HPRD, as well as interactions from orthologs and computational prediction. Gene identifiers were mapped from UniProt IDs to Entrez gene IDs using the human genome-wide annotation package in Bioconductor (version 3.4.1).[36] After removing self-loops and duplicated edges, the resulting network contained 17,487 genes and 891,969 interactions. Biological networks generated by aggregating interactions from literature are associated with study bias arising from disease-related genes being studied more often.[37,38] The IID network was chosen over other networks, such as BioGRID and HPRD, constructed solely from manual curation of literature, to minimize the effect of study bias on the results.

Results

Enrichment of known pathways

To evaluate the results produced by Grand Forest, we extracted gene modules from gene expression data from patients diagnosed with breast cancer, non-small cell lung cancer, ulcerative colitis, Huntington's disease, and ALS. As response variables, we used the overall survival time in breast cancer, lung cancer, and ALS and disease status (case vs. control) in ulcerative colitis and Huntington's disease, respectively. The interaction network was constructed from experimentally validated and computationally predicted interactions obtained from the IID (see Methods).

We evaluated the biological relevance of the extracted gene modules by investigating how congruent the genes in the extracted modules were with published curated molecular pathways related to the phenotype of each data set. Reference gene sets were extracted from KEGG.[39] For breast cancer, lung cancer, Huntington's disease, and ALS, we extracted the disease-specific pathway for each disease. Because KEGG has not published a specific pathway for ulcerative colitis, we instead aggregated all genes from the three pathways indicated as associated with UC: inflammatory bowel disease, cytokine–cytokine receptor interaction, and the Jak-STAT signaling pathway (Supplementary Table S3).

To evaluate Grand Forest's ability to find meaningful gene modules, we compared our results for all five data sets against the results obtained using four, state-of-the-art, network-based module discovery tools: BioNet,[12] KeyPathwayMiner (KPM),[13] GXNA,[11] and GiGa.[10] These tools were selected based on results of a recent evaluation by Batra et al.[14]

For each method, we extracted gene modules in each data set over a range of parameters chosen such that they generate modules in a range between $\sim$25 and 100 genes (Supplementary Table S4). Statistical significance of enrichment was computed using a hypergeometric overrepresentation test (Supplementary Text S2.2). Grand Forest significantly outperformed all tools on all data sets (Fig. 3a). The difference was especially pronounced in the two cancer data sets, where Grand Forest achieved a highly significant enrichment (median p-values of 6.13e−15 and 4.64e−9, respectively), while the other methods found little or no overlap with the associated pathways. Grand Forest performed worst on the Huntington's data (median p-value 0.074). All other tools delivered insignificant results (i.e., median $p > 0.05$) on all data sets.

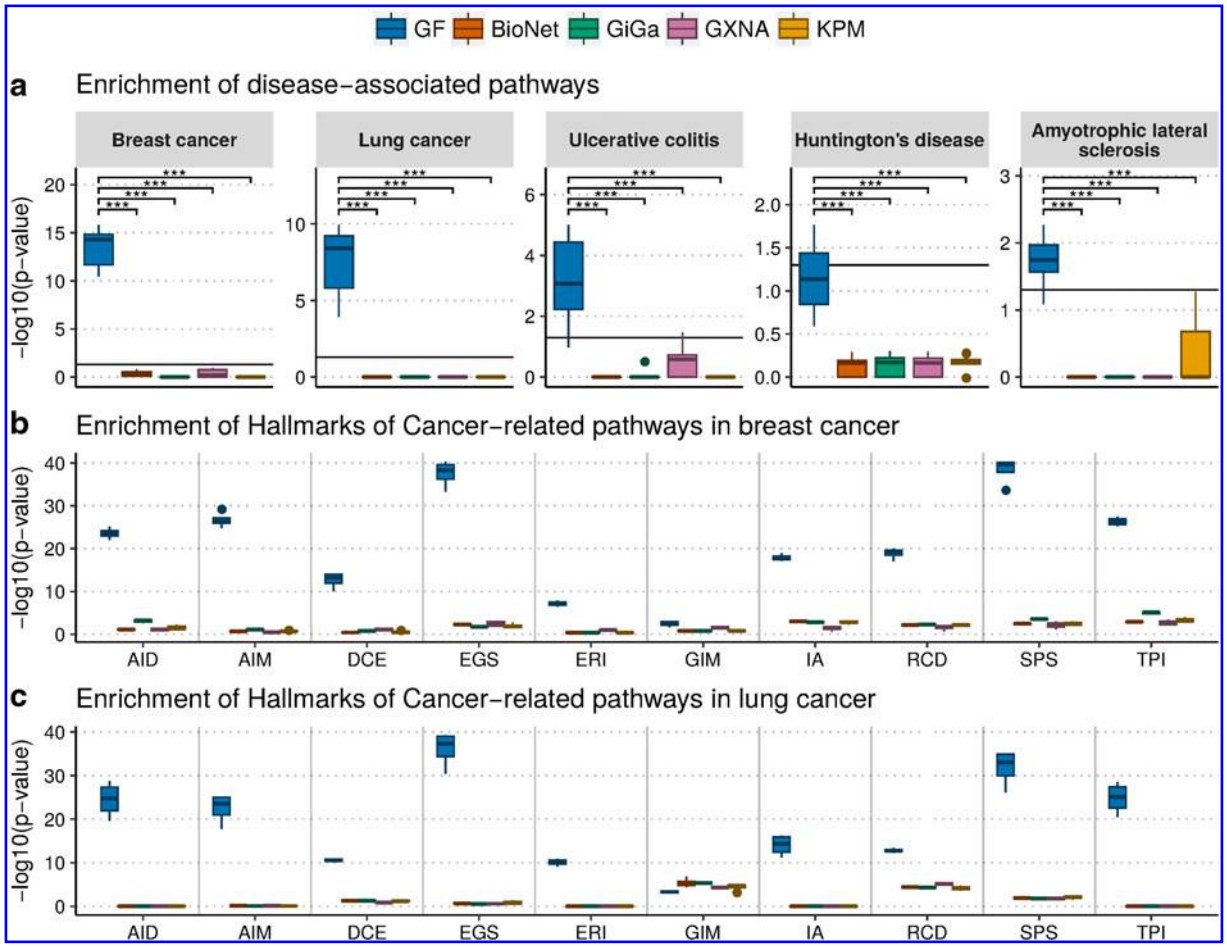

FIG. 3. Enrichment of disease-associated KEGG pathways in extracted gene modules. **(a)** Enrichment of disease-associated KEGG pathways for each data set. **(b, c)** Enrichment of pathways related to the hallmarks of cancer for modules extracted from breast cancer **(b)** and non-small cell lung cancer **(c)** data sets. Only modules of at least 75 genes were included. Hallmarks: AID, avoiding immune destruction; AIM, activating invasion and metastasis; DCE, deregulating cellular energetics; EGS, evading growth suppressors; ERI, enabling replicative immortality; GIM, genome instability and mutation; IA, inducing angiogenesis; RCD, resisting cell death; SPS, sustaining proliferative signaling; and TPI, tumor-promoting inflammation. Enrichment was computed using a hypergeometric overrepresentation test. The bold horizontal line indicates $p = 0.05$ ($^{***}p < 0.001$). KEGG, Kyoto Encyclopedia of Genes and Genomes.

Enrichment of pathways related to cancer hallmarks

To provide further validation of our results in the breast and lung cancer data sets, we investigated how strongly associated the extracted gene modules were with the hallmarks of cancer.[40] If the genes selected by a method are biologically relevant for proliferation of cancer, we would expect a functional enrichment of pathways related to these cancer hallmarks. Alcaraz et al.[41] compiled a set of KEGG pathways related to each hallmark. Based on their findings, we compiled the relevant genes for each hallmark as the union of all genes in the pathways related to that hallmark (Supplementary Table S5). Due to the great number of genes being compared against, we restricted this analysis to only modules of 75 or more genes (see Supplementary Figs. S2 and S3 for all sizes).

Grand Forest achieved a significantly higher enrichment in both breast cancer and lung cancer in all but one hallmark, namely genome instability and mutation, where Grand Forest was outperformed by the other

tools in lung cancer (Fig. 3b, c). Grand Forest generally achieved a highly significant degree of enrichment, with *p*-values below 1e−10 in all but two hallmarks.

Stability of selected genes

For the gene modules to be biologically meaningful, the genes in the extracted subnetwork modules should be stable and reproducible, that is, not varying significantly between samples from the same population. We evaluated how stable the modules produced by our method were compared with other methods, by repeatedly removing 20% of the patients, selected randomly, and measuring their pairwise similarity between all repetitions with the same parameters. Parameters were chosen to produce modules of ∼25, 50, 75, and 100 genes. Gene set similarity was measured using the Jaccard index (Supplementary Text S2.3).

We were unable to obtain results for KPM and BioNet. Neither method provides a way to enforce a specific module size, and the size instead depends on the chosen hyperparameters. For both methods, the size of the extracted modules varied significantly between repetitions using the same parameters, often by several orders of magnitude, making it infeasible to obtain appropriately sized modules for each repetition.

Overall, Grand Forest produced more stable results compared with existing GiGa and GXNA (Fig. 4a) in all data sets except for Huntington's disease. The difference in performance was most significant in breast cancer and ALS, where Grand Forest was stable even for small gene sets, while the other methods produced little overlap between repetitions. We observed that for the other methods, stability generally decreased with smaller module sizes; however, this effect was less pronounced in Grand Forest. These results suggest that the modules produced by Grand Forest are less sensitive to the sampling of the patient population. While we cannot compare it with KPM and BioNet, it is unlikely that either is more stable given their high sensitivity to hyperparameters.

Interaction density of selected modules

We evaluated the number of PPIs between genes in the extracted modules to better understand why the results produced by Grand Forest differed so much from other methods. For each module, we extracted the subnetwork induced by the constituent genes and counted the number of conserved edges. We observed that Grand Forest selected significantly more dense modules than the other methods for breast cancer, ulcerative colitis, and ALS, but similarly dense modules for lung cancer and Huntington's disease (Fig. 4b). All methods selected highly dense modules for lung cancer and sparse modules for Huntington's disease.

The observed difference in density may, in part, explain why the genes selected by Grand Forest were more congruent with published molecular pathways. This suggests that a large part of the power comes from the network rather than the gene expression data. However, given that all methods produced highly dense modules in lung cancer, even though only Grand Forest achieved a significant level of enrichment, this does not fully explain the difference in performance.

Statistical significance of selected genes

To shed further light on the source of the signal in data sets, we evaluated how many of the genes in extracted modules were significantly differentially expressed. For each module, we counted how many genes were significantly associated with the outcome (nominal $p < 0.05$).

We observed that Grand Forest generally selected fewer significant genes compared with other methods (Fig. 4c). This contrasted greatly with the other methods, where all modules consisted almost exclusively of significant genes. This is not surprising given that the other methods are designed to explicitly maximize this property in some way, either by maximizing the number of significant genes or by maximizing some aggregate significance measure. Interestingly, in the four data sets where Grand Forest selected fewer significant genes, namely in breast cancer, lung cancer, ulcerative colitis, and ALS, the difference in performance wrt. enrichment of KEGG pathways was greatest. This difference was especially pronounced in breast cancer where Grand Forest only selected around 25–35% significant genes while achieving a highly significant degree of enrichment. We also evaluated how many of the genes in the associated KEGG pathways were statistically significant. We observed that in all data sets, a large fraction of genes were in fact not significantly associated with the outcome, and the fraction of significant genes in the reference pathways was overall not significantly larger than among all genes (Fig. 4d and Supplementary Fig. S4). Only in Huntington's disease was a majority of genes significant (66%). These results suggest that a statistically significant association of expression with a phenotype is not necessarily adequate to determine which genes are important for development or progression of a disease.

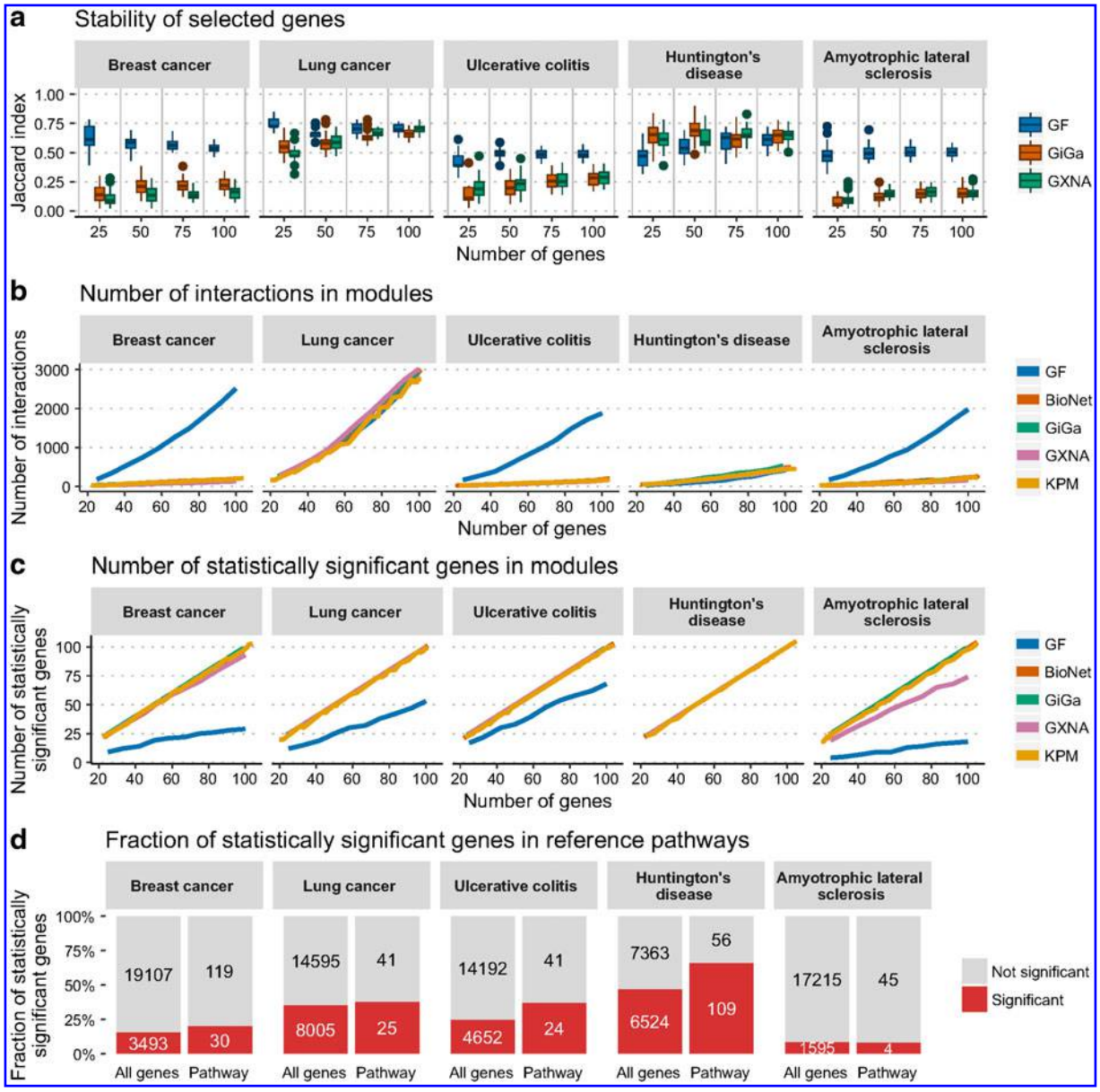

FIG. 4. Properties of genes selected by each method and frequency of statistically significant genes. **(a)** Stability of genes selected by each method for different module sizes. Gene modules were computed over 10 repetitions, sampling 80% of patients randomly. Stability was computed between all pairs of modules of same size using the Jaccard index. **(b)** Number of interactions in the induced subnetwork of each module. **(c)** Number of genes in modules selected by each method that were significantly differentially expressed (nominal $p < 0.05$) with respect to the outcome. **(d)** Fraction of genes in reference pathways that were differentially expressed (nominal $p < 0.05$) in the corresponding gene expression data set with respect to the outcome.

De novo endophenotyping of lung adenocarcinoma

Grand Forest can also be applied to unlabeled data to discover modules of highly interacting genes that stratify patients into distinct clusters, for example, molecular subtypes or endophenotypes. Feature importance was estimated without any clinical variables by modeling the problem as an unsupervised as supervised learning problem (see Methods). We then extracted the gene module comprising the 20 most important genes (Fig. 5a).

The selected genes induced a highly dense subnetwork in the interaction network. Among the 20 genes were three genes found in the KEGG non-small cell lung cancer pathway: *TP53*, *EGFR*, and *MAPK1* ($p = 2.6e-5$). Furthermore, we found four known oncogenes, *MYC*,[42,43] *JUN*,[44,45] *EGFR*,[46,47] and *CTNNB1*,[48] and two important

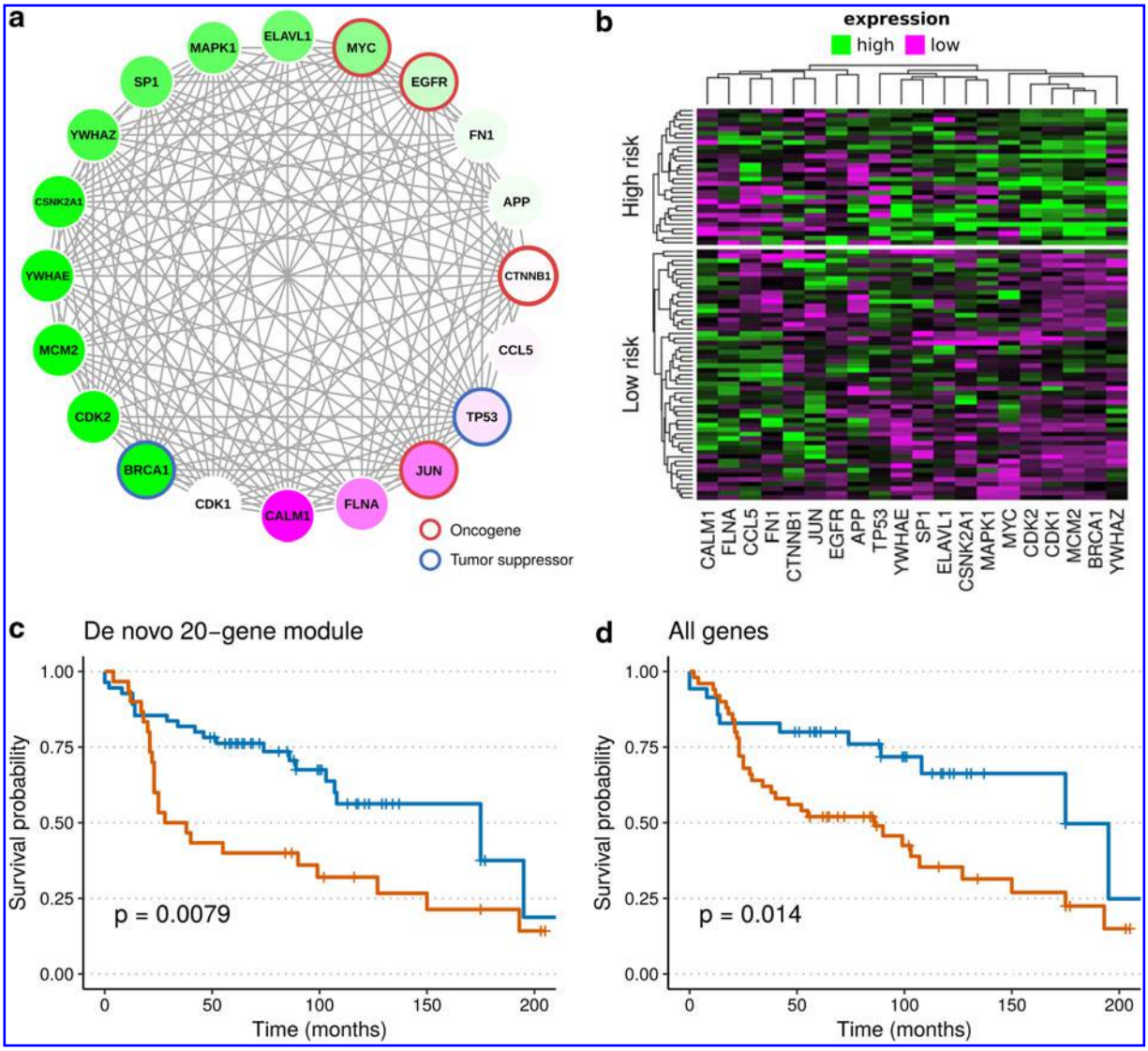

FIG. 5. *De novo* endophenotyping of lung adenocarcinoma based on a 20-gene subnetwork module extracted with Grand Forest. **(a)** Subnetwork induced by the genes in the module. Nodes are colored according to the difference between mean expression in high-risk and low-risk groups. **(b)** Heatmap of gene expression for patients clustered into two clusters using only genes in the module. Expression values are mean centered and scaled with standard deviation. **(c, d)** Overall survival for patients when clustered using the 20-gene module **(c)** and all genes **(d)**.

FIG. 6. Overview of key features of the Grand Forest web server. **(a)** Main panel of the supervised analysis. Data and parameters are controlled from the left sidebar. An extracted 20-gene subnetwork module is visualized as a network and a heatmap annotated with the response variable. **(b)** Dot plot showing KEGG pathways significantly enriched with genes in the extracted module. **(c)** A drug–target network of drugs (blue nodes) targeting the genes (red nodes) in the module. **(d)** In the unsupervised analysis, if survival times are included, survival curves can be visualized and compared with an existing stratification.

tumor suppressor genes, *BRCA1*[49] and *TP53*.[50] To evaluate the clinical relevance of the selected genes, we extracted all adenocarcinoma samples from the lung cancer data set ($n = 85$) and clustered them into two groups with k-means clustering, using only the expression of these 20 genes. Because k-means is dependent on randomly chosen initialization, clustering was repeated 20 times and the result with the greatest Silhouette index was chosen (Supplementary Text S2.4). The clustering of patients was significantly associated with overall survival (log-rank, $p = 0.0079$) (Fig. 5b, c). For comparison, we also clustered the patients on all genes using the same procedure. When using all genes, the resulting stratification was less associated with overall survival ($p = 0.014$) (Fig. 5d).

Discussion and Conclusion

In this study, we introduced Grand Forest, a novel graph-guided set of ensemble learning methods based on the well-known random forest strategy to allow for network-guided supervised and *de novo* endophenotyping. Our tool and the implemented approaches differ significantly from conventional module discovery and patient stratification tools. Grand Forest does not expect the data to follow a specific distribution, and it does not rely on statistical significance tests and differential expression analyses, but instead aims to explain the phenotype using an ensemble of decision trees. When compared with traditional module discovery tools across gene expression data from five diseases, our method achieved a significantly higher degree of enrichment of relevant molecular pathways. Results also showed that Grand Forest was less sensitive to the sampling of the patient population than GXNA and GiGa, but a comparison with KPM and BioNet was not possible. By virtue of being based on decision trees, our method is also invariant to scaling and robust to outliers.

We observed that despite selecting fewer genes with a statistically significant association with the clinical variable, Grand Forest extracted modules that were more congruent with KEGG molecular pathways related to the disease. However, it appeared that the solutions computed with Grand Forest were largely driven by interaction density rather than expression patterns associated with disease outcome. This was further demonstrated by the fact that a large fraction of the genes in the reference gene sets were not statistically significant. This demonstrates that a large fold change, or otherwise significant association with outcome, is not sufficient to identify important causal driver genes for a disease.

A commonly raised concern with network-based methods is that a similar performance can often be achieved using random networks instead.[41,51,52] We evaluated the performance of Grand Forest on the five data sets using two randomized network models, one generated by randomly rewiring edge pairs while preserving node degree and one generated by rearranging the node labels in the network. We observed that rearranging node labels resulted in significantly worse performance wrt. enrichment of relevant pathways (Supplementary Fig. S5) and generally did not achieve a significant level of enrichment. However, rewiring edges did not significantly affect enrichment, which confirms our method is heavily reliant on node degree. Permuting the outcome variables also only had a modest effect on the degree of enrichment for Grand Forest while (in some cases) even increasing enrichment for GiGa and GXNA (Supplementary Fig. S6). We furthermore investigated the effect of false negatives and false positives in the interaction network by repeating the experiment after removing 25% of edges or adding 25% more random edges in the network, respectively (Supplementary Figs. S7 and S8). Neither of these two perturbations significantly affected the results.

Taken together, our results point to a central problem with module discovery from gene expression data: methods relying primarily on gene expression will in many cases not identify disease-driving genes, while methods relying primarily on network structure are likely to select disease drivers due to bias in the network alone. This may, in part, be because gene expression is too far downstream and, as such, expression changes may often correspond to the cellular response to the disease rather than the underlying cause. Furthermore, it is likely that Grand Forest selects highly dense modules because the difference in patient phenotype translates to large-scale changes in the transcriptome, which makes it trivial for the algorithm to build a set of genes that explain the outcome well. Therefore, the algorithm will often choose hub genes since they are easier to reach in a graph traversal. This observation is also in line with previous results on random gene expression signatures in breast cancer.[6] Due to the incompleteness and noisy nature of current PPI networks, it is uncertain whether disease-associated genes have a large number of reported interactions due to research bias or if genes with many interactions are often associated with disease due to

being involved in many key cellular mechanisms. With this in mind, we advise researchers to use caution when applying network-guided methods for discovering disease genes and modules.

To make our method easily available to researchers, we developed an easy-to-use web server for carrying out analyses using Grand Forest. The web server allows users to upload a gene expression data set and analyze their data using two different workflows: a supervised workflow and an unsupervised workflow, mirroring the types of analyses carried out in this article. A set of commonly used genetic interaction networks is provided, but users can also upload custom network data. A gene set enrichment analysis is provided for both workflows to enable searching for over-represented Gene Ontology terms, pathways, and disease associations among the extract genes. Furthermore, users are also able to extract and visualize networks of drugs and miRNAs targeting the genes in a module to search for potentially druggable targets. See Figure 6 for a graphical overview.

In summary, with Grand Forest, we introduce a new method for disease-gene module discovery by integrating genomic profiling data with molecular interaction networks. We also introduce the first network-based, *de novo* endophenotyping methodology, allowing analysis of unlabeled data. We show that Grand Forest identifies gene modules closely associated with known disease genes, but that these results are highly driven by network topology and likely confounded by inherent bias in the underlying network. Finally, we provide a comprehensive web server to make our methodology easily available to researchers.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

S.J.L. and J.B. are grateful for financial support from J.B.'s VILLUM Young Investigator grant no. 13154; H.H.H.W.S. for an ERC AdG 294683-RadMed; and J.B. and H.H.H.W.S. for support from the H2020 grant REPO-TRIAL no. 777111.

Supplementary Material

Supplementary Data

References

1. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747–752.
2. Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001;98:10869–10874.
3. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–550.
4. Guinney J, Dienstmann, R, Tejpar, S. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21:1350–1356.
5. Vliet MH van, Reyal F, Horlings HM, et al. Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. BMC Genomics. 2008;9:375.
6. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011;7:e1002240.
7. Curtis RK, Orei M, Vidal-Puig A. Pathways to the analysis of microarray data. Trends Biotechnol. 2005;23:429–435.
8. Werner T. Bioinformatics applications for pathway analysis of microarray data. Curr Opin Biotechnol. 2008;19:50–54.
9. Ideker T, Ozier O, Schwikowski B, et al. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002; 18:S233–S240.
10. Breitling R, Amtmann A, Herzyk P. Graph-based iterative Group Analysis enhances microarray interpretation. BMC Bioinformatics. 2004;5:100.
11. Nacu S, Critchley-Thorne R, Lee P, et al. Gene expression network analysis and applications to immunology. Bioinformatics. 2007;23:850–858.
12. Beisser D, Klau GW, Dandekar T, et al. BioNet: an R-Package for the functional analysis of biological networks. Bioinformatics. 2010;26: 1129–1130.
13. Alcaraz N, Pauling J, Batra R, et al. KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape. BMC Syst Biol. 2014;8.
14. Batra R, Alcaraz N, Gitzhofer K, et al. On the performance of de novo pathway enrichment. NPJ Syst Biol Appl. 2017;3.
15. Pirooznia M, Yang JY, Yang MQ, et al. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. 2008;9:S13.
16. Marbach D, Costello JC, Küffner R, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9:796–804.
17. Zhang QC, Petrey D, Deng L, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature. 2012;490: 556–560.
18. Riemenschneider M, Cashin, KY, Budeus B, et al. Genotypic prediction of co-receptor tropism of HIV-1 subtypes A and C. Sci Rep. 2016;6:24883.
19. Ma J, Yu MK, Fong S, et al. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods. 2018;15:290–298.
20. Dutkowski J, Ideker T. Protein networks as logic functions in development and cancer. PLoS Comput Biol. 2011;7:e1002180.
21. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
22. Breiman L, Cutler A. Setting up, using, and understanding random forests. University of California, Department of Statistics, Berkeley, CA. 2003.
23. Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE Computer Society Press, Montreal, Canada. 1995. DOI: 10.1109/icdar.1995.598994.
24. Daz-Uriarte R, de Andrés SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;7:3.
25. Kursa MB. Robustness of random forest-based gene selection methods. BMC Bioinformatics. 2014;15:8.
26. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw. 2017;77.
27. Malley JD, Kruppa J, Dasgupta A, et al. Probability machines. Methods Inf Med. 2011;51:74–81.
28. Ishwaran H, Kogalur UB, Blackstone EH, et al. Random survival forests. Ann Appl Stat. 2008;2:841–860.
29. Kao KJ, Chang KM, Hsu HC, et al. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer. 2011;11:143.
30. Rousseaux S, Debernardi A, Jacquiau B, et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. Sci Transl Med. 2013;5:186ra66.
31. Noble CL, Abbas AR, Cornelius J, et al. Regional variation in gene expression in the healthy colon is dysregulated in ulcerative colitis. Gut. 2008;57:1398–1405.

32. Hodges A, Strand AD, Aragaki AK, et al. Regional and cellular gene expression changes in human Huntington's disease brain. Hum Mol Genet. 2006;15:965–977.
33. Rheenen W van, Diekstra FP, Harschnitz O, et al. Whole blood transcriptome analysis in amyotrophic lateral sclerosis: a biomarker study. PLoS One. 2018;13:e0198874.
34. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
35. Kotlyar M, Pastrello C, Sheahan N, et al. Integrated interactions database: tissue-specific view of the human and model organism interactomes. Nucleic Acids Res. 2016;44:D536–D541.
36. Carlson M. org.Hs.eg.db: genome wide annotation for human. R package version 3.4.1. 2017.
37. Gillis J, Ballouz S, Pavlidis P. Bias tradeoffs in the creation and analysis of proteinprotein interaction networks. J Proteomics. 2014;100:44–54.
38. Schaefer MH, Serrano L, Andrade-Navarro MA. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. Front Genet. 2015;6:260.
39. Kanehisa M, Furumichi M, Tanabe M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2016;45:D353–D361.
40. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–674.
41. Alcaraz N, List M, Batra R, et al. De novo pathway-based biomarker identification. Nucleic Acids Res. 2017;45:e151.
42. Rapp UR, Korn C, Ceteci F, et al. Myc is a metastasis gene for non-small-cell lung cancer. PLoS One. 2009;4:e6029.
43. Gabay M, Li Y, Felsher DW. MYC activation is a hallmark of cancer initiation and maintenance. Cold Spring Harb Perspect Med. 2014;4:a014241.
44. Szabo E, Riffe ME, Steinberg SM, et al. Altered cJUN expression: an early event in human lung carcinogenesis. Cancer Res. 1996;56:305–315.
45. Vogt PK. Fortuitous convergences: the beginnings of JUN. Nat Rev Cancer. 2002;2:465–469.
46. Lynch TJ, Bell DW, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of nonsmall-cell lung cancer to gefitinib. N Engl J Med. 2004;350:2129–2139.
47. Gandara DR, Hammerman PS, Sos ML, et al. Squamous cell lung cancer: from tumor genomics to cancer therapeutics. Clin Cancer Res. 2015;21:2236–2243.
48. Morin PJ. β-catenin signaling and cancer. Bioessays. 1999;21:1021–1030.
49. Yoshida K, Miki Y. Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. Cancer Sci. 2004;95:866–871.
50. Surget S, Khoury MP, Bourdon JC. Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. Onco Targets Ther. 2013;7:57–68.
51. Staiger C, Cadot S, Kooter R, et al. A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. PLoS One. 2012;7:e34796.
52. Staiger C, Cadot S, Györffy B, et al. Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. Front Genet. 2013;4:289.

Cite this article as: Larsen SJ, Schmidt HHHW, Baumbach J (2020) *De novo* and supervised endophenotyping using network-guided ensemble learning, *Systems Medicine* 3:1, 8–21, DOI: 10.1089/sysm.2019.0008.

Abbreviations Used

ALS = amyotrophic lateral sclerosis
BioGRID = Biological General Repository for Interaction Datasets
GEO = Gene Expression Omnibus
GXNA = Gene eXpression Network Analysis
HPRD = Human Protein Reference Database
IID = Integrated Interactions Database
KEGG = Kyoto Encyclopedia of Genes and Genomes
KPM = KeyPathwayMiner
PPIs = protein–protein interactions
UC = ulcerative colitis