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]

[1]Department of Mathematics and Computer Science, University of Southern Denmark
[2]Department of Pharmacology and Personalised Medicine, Faculty of Health, Medicine and Life Science, Maastricht University
[3]Chair of Experimental Bioinformatics, Wissenschaftszentrum Weihenstephan, Technical University of Munich
[*]Corresponding author, e-mail: sjlarsen@imada.sdu.dk

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 don't 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 to four state of the art methods. Grand Forest consistently found gene modules wither 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 overrepresented 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.

Introduction

The increasingly large amounts of functional genomic data currently available in public databases such as the Gene Expression Omnibus and through extensive data collection efforts like The Cancer Genome Atlas, has 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 has been put into the 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 outcome in breast cancer, despite having no relation to 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 the development or progression of the disease itself.

In order 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, in order 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 KEGG and Reactome.

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. protein-protein interactions). 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 (Figure 1). Our method extends significantly on previous ideas from Dutkowski *et al.*[15] (see Supplementary text S2 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 disease-gene module detection tools, while also being less sensitive to the sampling of the patient

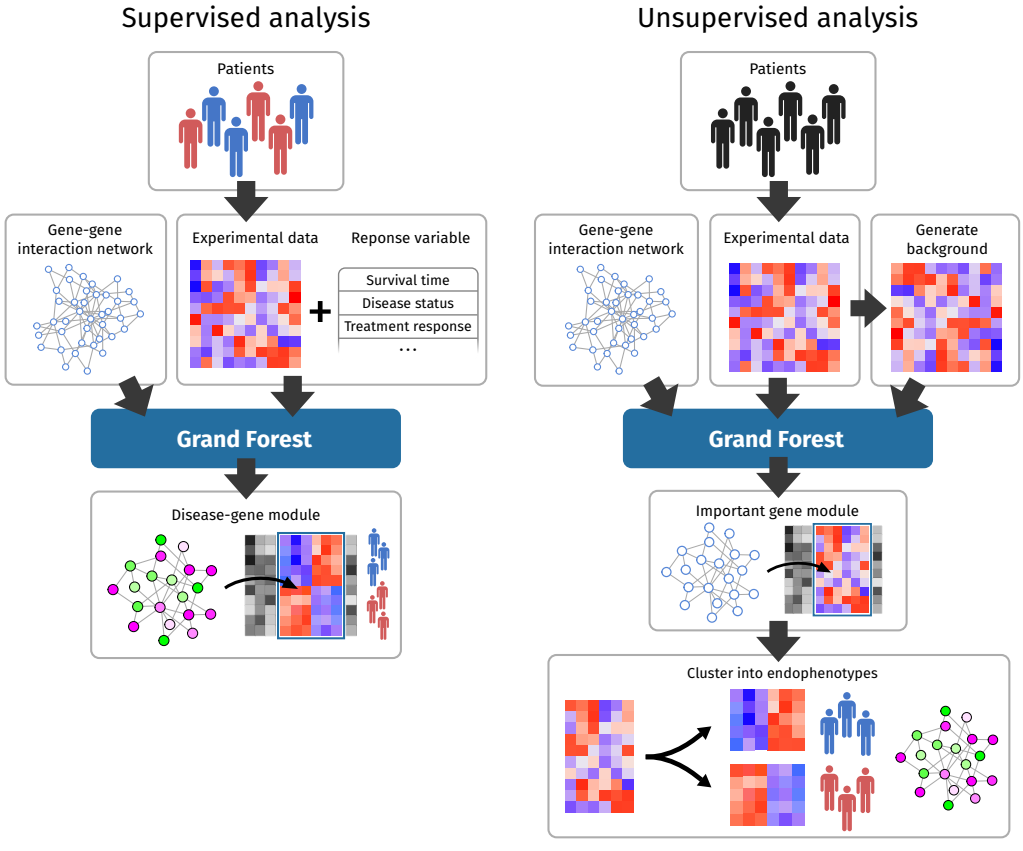

Figure 1: Overview of the supervised and unsupervised Grand Forest workflows. In the supervised workflow expression data is integrated with an interaction network to identify a gene module associated with a response variable, i.e. survival time. In the unsupervised workflow a model is trained to recognize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

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 to other methods, suggesting that disease-associated genes were selected mainly due to network topology. Furthermore, known disease-associated genes were observed to be only marginally overrepresented 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.[16,17] It is based on the random decision forests method,[18] but extended to use the random subspace method (also known as feature bagging). It has achieved widespread use 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 multi-class problems. Furthermore, it provides several measures for estimating feature importance, making it possible to identify important genes in molecular profiling data.[19,20] We refer to the original manuscripts by Breiman *et al.* for a description of the Random Forest algorithm.[16,17]

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 $|V| = p$ and each vertex $v_i \in V$ corresponds to column i in $\mathbf{X}$. 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. The algorithm is outlined in Figure 2.

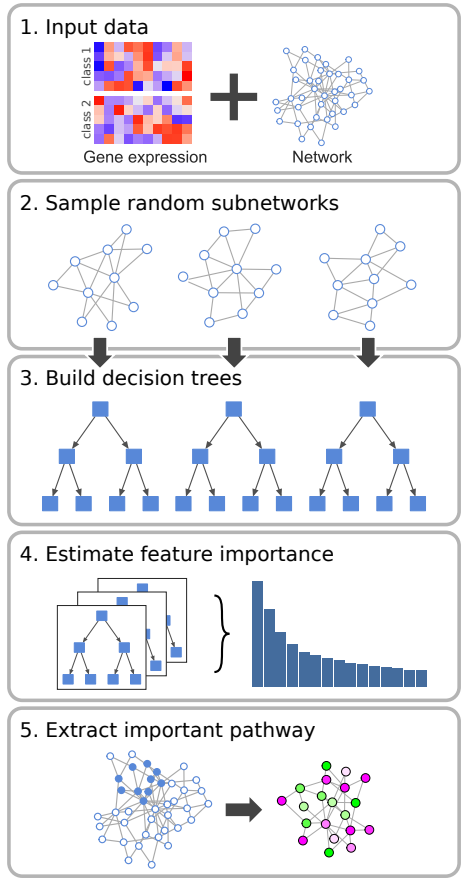

Figure 2: Overview of the Grand Forest algorithm. (1) The method takes as input a matrix of expression values for two or more classes and a network modeling the pairwise interactions between genes. (2) A large number of connected subnetworks are randomly sampled from the full network. (3) A decision 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.

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 S1.4 for a detailed overview.

Split variable selection

When building decision trees, splits are formed by selecting a variable and value to split the partition on, that maximizes some split criteria, e.g. 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 Figure S1). This 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 in order 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 split, conditioned on the splits preceding it.

Implementation

Our implementation of Grand Forest is based on ranger[21] 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,[16] probability forests[22] and survival forests.[23] By extension, Grand Forest has been implemented to support these methods as well. The source code 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 Leo Breiman.[17] The method is based on the following assumption: if the data 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 from 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 Gene Expression Omnibus (GEO). The data sets are available through the following accession IDs: breast cancer (GSE20685, $n = 327$),[24] non-small cell lung cancer (GSE30219, $n = 268$),[25] ulcerative colitis (GSE11223, $n = 202$),[26] Huntington's disease (GSE3790, $n = 54$),[27] and amyotrophic lateral sclerosis (GSE112680, $n = 164$).[28] 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. The Huntington's disease data set contained samples from different brain regions. Only samples from the caudate nucleus were used, 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.* Patients were separated into a control group and a case group (Vonsattel grades 2-4). Samples with Vonsattel grade 0-1 were discarded.

The 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 cutoff of 62 months in lung cancer and 10.6 years in breast cancer, and 4 years in ALS. Patients that could not be placed in either group due to censoring or lack of follow-up were discarded.

Statistical significance tests

For the survival data sets (breast and lung cancer) 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.[29]

Network data preparation

We collected network data from the Integrated Interactions Database (IID)[30] (version 2017-04). IID integrates experimentally validated protein-protein interactions (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)[31] 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.[32,33] The IID network was chosen over other networks, such as BioGRID and HPRD, constructed solely from manual curation of literature, in order 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 amyotrophic lateral sclerosis. 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 Integrated Interactions Database (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.[34] 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.

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 module discovery tools: BioNet,[12] KeyPathwayMiner (KPM),[13] GXNA[11] and GiGa.[10] These tools were selected based on the 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 approximately 25 and 100 genes (Supplementary Table S5). Statistical significance of enrichment was computed using a hypergeometric overrepresentation test. Grand Forest significantly outperformed all tools on all data sets (Figure 3a). The difference was especially pronounced in the two cancer data sets, where Grand Forest achieved a highly significant enrichment (median p-value 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.

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 to the hallmarks of cancer.[35] If the genes selected by a method are biologically relevant for the proliferation of cancer, we would expect a functional enrichment of pathways related to these cancer hallmarks. Alcaraz *et al.*[36] 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 S4). Due to the great number of genes being compared against, we restricted this analysis to only modules of 75 or more genes (see Supplementary Figure S2 and S3 for all sizes).

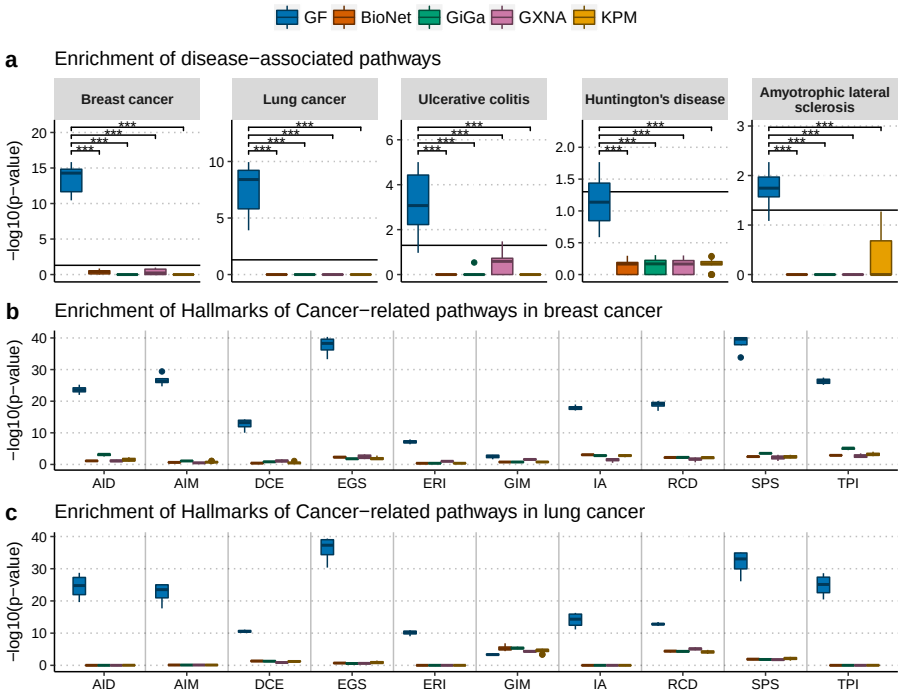

Figure 3: Enrichment of disease-associated KEGG pathways in extracted gene modules. (a) Enrichment of disease-associated KEGG pathways for each data set. (b-c) Enrichment of pathways related to the Hallmarks of Cancer for modules extracted from breast cancer (b) and non-small cell lung cancer (c) data sets. Only modules of at least 75 genes were included. Hallmarks: Avoiding Immune Destruction (AID), Activating Invasion and Metastasis (AIM), Deregulating Cellular Energetics (DCE), Evading Growth Suppressors (EGS), Enabling Replicative Immortality (ERI), Genome Instability and Mutation (GIM), Inducing Angiogenesis (IA), Resisting Cell Death (RCD), Sustaining Proliferative Signaling (SPS), Tumor-promoting Inflammation (TPI). Enrichment was computed using a hypergeometric overrepresentation test. Bold horizontal line indicates $p = 0.05$. ($^{***}p < 0.001$.)

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 (Figure 3b-c). Grand Forest generally achieved a highly significantly degree of enrichment, with p-values below 1e−10 in all but two hallmarks.

Stability of selected genes

In order for the gene modules to be biologically meaningful, the genes in the extracted subnetwork modules should be stable and reproducible, i.e. not varying significantly between samples from the same population. We evaluated how stable the modules produced by our method was compared to 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 approximately 25, 50, 75 and 100 genes. Gene set similarity was measured using the Jaccard index.

We were unable to obtain results for KPM and BioNet. Neither of these methods provide 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 to existing GiGa and GXNA (Figure 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 to KPM and BioNet, it is unlikely that either is more stable given their high sensitivity to the hyperparameters.

Interaction density of selected modules

We evaluated the number of protein-protein interactions between genes in the extracted modules, in order 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 (Figure 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.

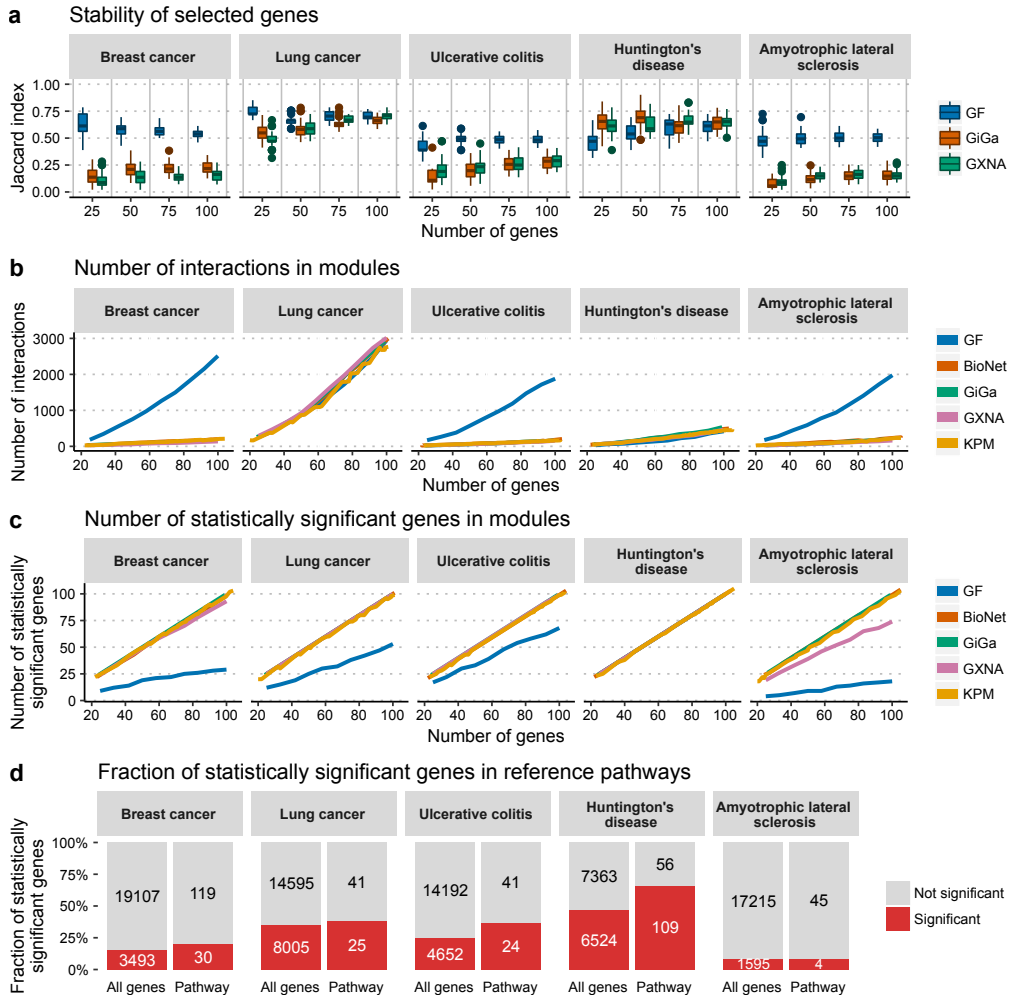

Figure 4: Properties of genes selected by each method and frequency of statistically significant genes. (a) Stability of genes selected my each method for different module sizes. Gene modules were computed over 10 repetitions, sampling 80% of patients randomly. Stability was computed between all pairs 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 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 outcome.

Statistical significance of selected genes

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

We observed that Grand Forest generally selected fewer significant genes compared to the other methods (Figure 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 outcome, and the fraction of significant genes in the reference pathways was overall not significantly larger than among all genes (Figure 4d, Supplementary Figure S6). In fact, only in Huntington's disease was a majority of genes significant (66%). These results suggest that a statistically significant association of expression to a phenotype is not necessarily adequate to determine which genes are important for the development or progression of a disease.

De novo endophenotyping of lung adenocarcinoma

Grand Forest can also be applied to unlabeled data in order to discover modules of highly interacting genes that stratify patients into distinct clusters, e.g. 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 comprised of the 20 most important genes (Figure 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* ($p = 2.6\text{e}{-5}$). Furthermore, we found four known oncogenes *MYC*,[37,38] *JUN*,[39,40] *EGFR*[41,42] and *CTNNB1*,[43] and two important tumor suppressor genes *BRCA1*[44] and *TP53*.[45] In order 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 the randomly chosen initialization, the clustering was repeated 20 times and the result with the greatest Silhouette index was chosen. The clustering of the patients was significantly associated with overall survival (log-rank, $p = 0.0079$) (Figure 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 ($p = 0.014$) (Figure 5d).

Discussion and conclusion

Here 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

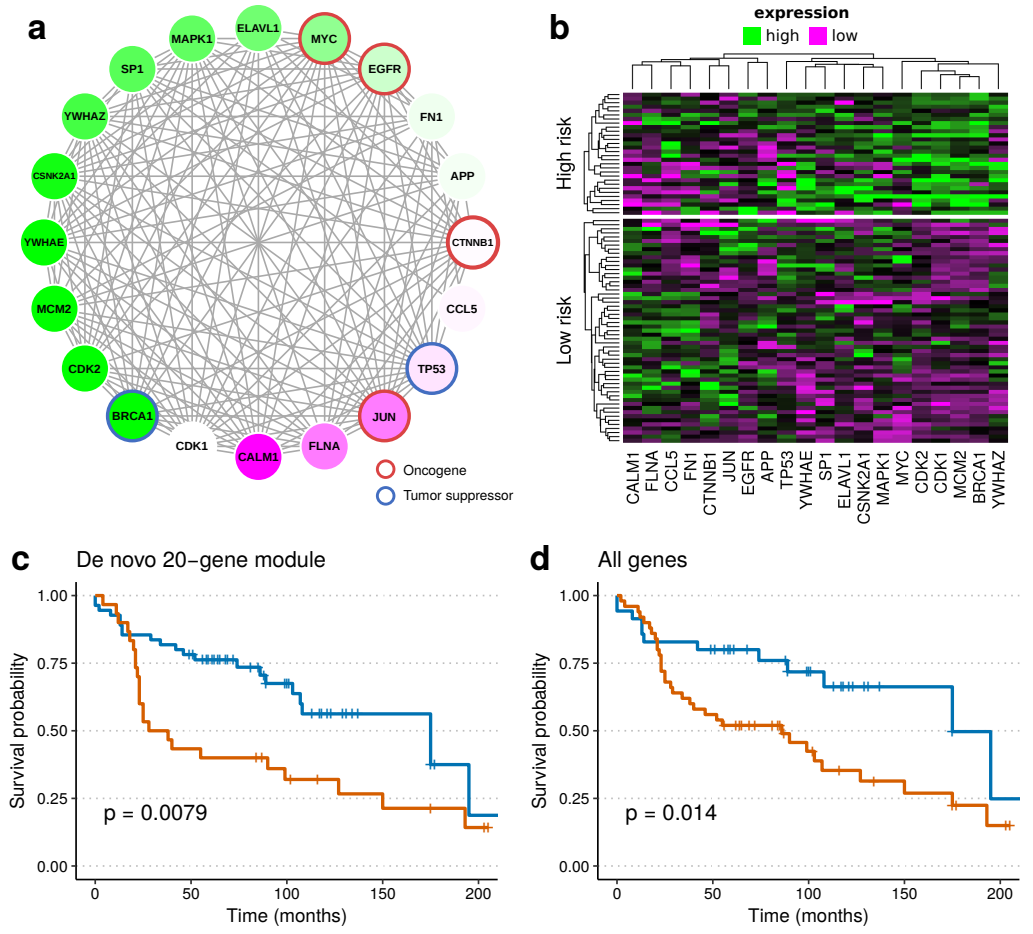

Figure 5: *De novo* endophenotyping of lung adenocarcinoma based on a 20-gene subnetwork module extracted with Grand Forest. (a) Subnetwork induced by the genes in the module. Nodes are colored according to difference between mean expression in high risk and low risk group. (b) Heatmap of gene expression 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 20-gene module (c) and all genes (d).

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 to 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 to KeyPathwayMiner 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.[36,46,47] 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 Figure 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.

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 reponse to the disease rather than the underlying cause. Furthermore, a likely reason for why Grand Forest selects highly dense modules is that 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 protein-protein interaction 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.

In order 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: an supervised work and an unsupervised workflow, mirroring the types of analyses carried out in this paper. A set of commonly used genetic interaction networks are provided, but users can also upload custom network data. A gene set enrichment analysis is provided for both

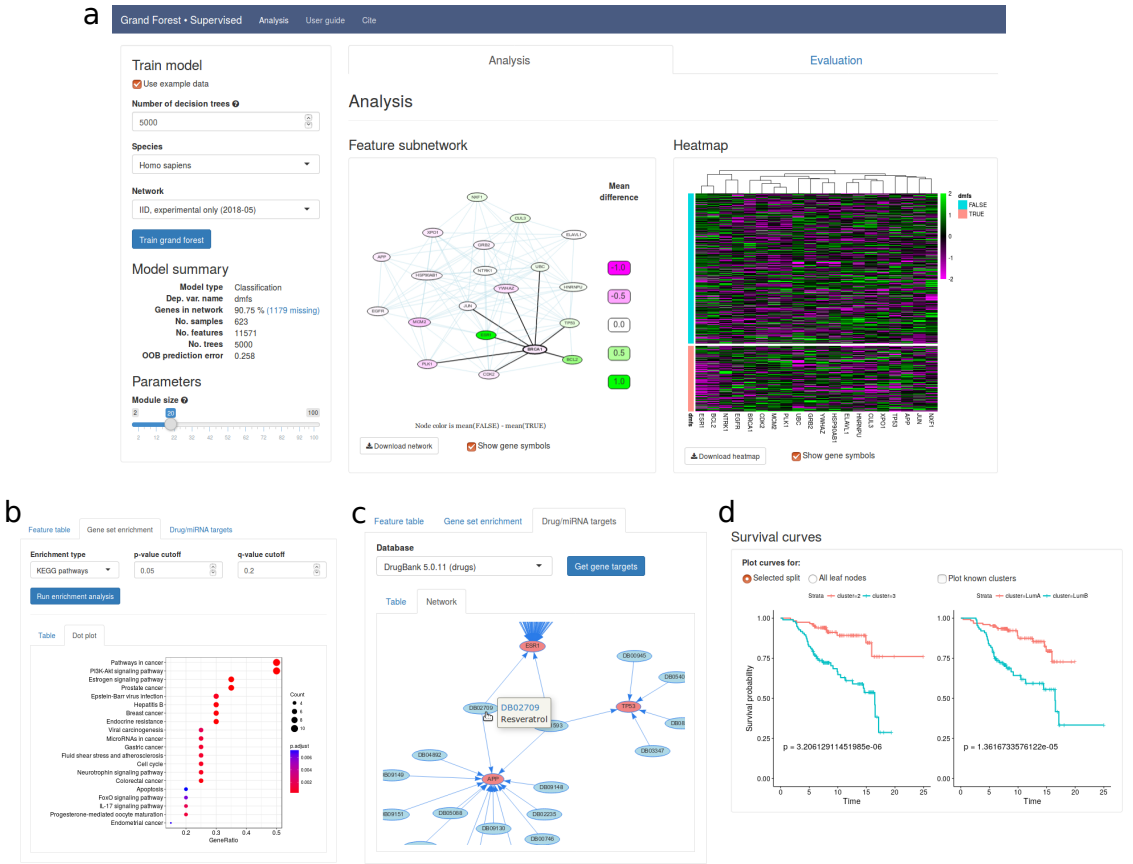

Figure 6: Overview of key features of Grand Forest web server. (a) Main panel of the supervised analysis. Data and parameters are controlled from the left sidebar. An extracted 20-gene subnetwork module is visualized as a network and a heat map annotated with the response variable. (b) Dot plot showing 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 to an existing stratification.

workflows, to enable searching for overrepresented 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 interactions networks. We also introduce the first network-based *de novo* endophenotyping methodology, allowing analysis on unlabelled 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 in order to make our methodology easily available to researchers.

Acknowledgements

SJL and JB are grateful for financial support from JB's VILLUM Young Investigator grant nr. 13154; HHHWS for an ERC AdG 294683-RadMed; JB and HHHWS for support from H2020 grant REPO-TRIAL nr. 777111.

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. Network, TCGAR. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511:543–550.
4. Guinney, J, Dienstmann, R, and Tejpar, S. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21:1350–1356.
5. Vliet, MH van, Reyal, F, Horlings, HM, Vijver, MJ van de, Reinders, MJT, and Wessels, LFA. 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, and Detours, V. Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLoS Comput Biol 2011;7. Ed. by Rigoutsos, I:e1002240.
7. Curtis, RK, Orei, M, and 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, and Siegel, AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 2002;18:S233–S240.
10. Breitling, R, Amtmann, A, and Herzyk, P. Graph-based iterative Group Analysis enhances microarray interpretation. BMC Bioinf 2004;5:100.
11. Nacu, S, Critchley-Thorne, R, Lee, P, and Holmes, S. Gene expression network analysis and applications to immunology. Bioinformatics 2007;23:850–858.
12. Beisser, D, Klau, GW, Dandekar, T, Muller, T, and Dittrich, MT. 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. Dutkowski, J and Ideker, T. Protein Networks as Logic Functions in Development and Cancer. PLoS Comput Biol 2011;7. Ed. by Altman, RB:e1002180.
16. Breiman, L. Random Forests. Machine Learning 2001;45:5–32.

17. Breiman, L and Cutler, A. Setting up, using, and understanding random forests. University of California, Department of Statistics. 2003.

18. Ho, TK. Random decision forests. In: *Proc Int Conf Doc Anal Recognit*. IEEE Comput. Soc. Press, 1995. DOI: 10.1109/icdar.1995.598994.

19. Daz-Uriarte, R and Andrés, SA de. Gene selection and classification of microarray data using random forest. BMC Bioinf 2006;7:3.

20. Kursa, MB. Robustness of Random Forest-based gene selection methods. BMC Bioinf 2014;15:8.

21. Wright, MN and Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw 2017;77.

22. Malley, JD, Kruppa, J, Dasgupta, A, Malley, KG, and Ziegler, A. Probability Machines. Methods Inf Med 2011;51:74–81.

23. Ishwaran, H, Kogalur, UB, Blackstone, EH, and Lauer, MS. Random survival forests. Ann Appl Stat 2008;2:841–860.

24. Kao, KJ, Chang, KM, Hsu, HC, and Huang, AT. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer 2011;11.

25. 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;5:186ra66–186ra66.

26. 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.

27. 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.

28. Rheenen, W van, Diekstra, FP, Harschnitz, O, et al. Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. PLoS One 2018;13. Ed. by Raoul, C:e0198874.

29. 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–e47.

30. Kotlyar, M, Pastrello, C, Sheahan, N, and Jurisica, I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. Nucleic Acids Res 2016;44:D536–D541.

31. Carlson, M. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.4.1. 2017.

32. Gillis, J, Ballouz, S, and Pavlidis, P. Bias tradeoffs in the creation and analysis of proteinprotein interaction networks. J Proteomics 2014;100:44–54.

33. Schaefer, MH, Serrano, L, and Andrade-Navarro, MA. Correcting for the study bias associated with proteinprotein interaction measurements reveals differences between protein degree distributions from different cancer types. Front Genet 2015;6.

34. Kanehisa, M, Furumichi, M, Tanabe, M, Sato, Y, and Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2016;45:D353–D361.

35. Hanahan, D and Weinberg, RA. Hallmarks of Cancer: The Next Generation. Cell 2011;144:646–674.

36. Alcaraz, N, List, M, Batra, R, Vandin, F, Ditzel, HJ, and Baumbach, J. De novo pathway-based biomarker identification. Nucleic Acids Res 2017;45:e151–e151.

37. Rapp, UR, Korn, C, Ceteci, F, et al. Myc Is a Metastasis Gene for Non-Small-Cell Lung Cancer. PLoS One 2009;4. Ed. by Aziz, SA:e6029.

38. Gabay, M, Li, Y, and Felsher, DW. MYC Activation Is a Hallmark of Cancer Initiation and Maintenance. Cold Spring Harb Perspect Med 2014;4:a014241–a014241.

39. Szabo, E, Riffe, ME, Steinberg, SM, Birrer, MJ, and Linnoila, RI. Altered cJUN expression: an early event in human lung carcinogenesis. Cancer Res 1996;56:305–315.

40. Vogt, PK. Fortuitous convergences: the beginnings of JUN. Nat Rev Cancer 2002;2:465–469.

41. 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.

42. Gandara, DR, Hammerman, PS, Sos, ML, Lara, PN, and Hirsch, FR. Squamous Cell Lung Cancer: From Tumor Genomics to Cancer Therapeutics. Clin Cancer Res 2015;21:2236–2243.

43. Morin, PJ. β-catenin signaling and cancer. Bioessays 1999;21:1021–1030.

44. Yoshida, K and 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.

45. Bourdon, JC, Surget, S, and Khoury, MP. Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. Onco Targets and Ther 2013:57.

46. 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. Ed. by Dopazo, J:e34796.

47. Staiger, C, Cadot, S, Györffy, B, Wessels, L, and Klau, G. Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. Front Genet 2013;4:289.