Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

A Multivariate Ensemble Approach for Identification of Biomarkers: Application to Breast Cancer ⋆

Gunjan S. Thakur * **Bernie J. Daigle, Jr.** **
Linda R. Petzold *** **Frank J. Doyle III.** ****

* *Institute for Collaborative Biotechnologies,University of California Santa Barbara, Santa Barbara,CA 93106-5080, USA (Tel: 805-893-4325; e-mail: gunjan@engr.ucsb.edu).*
** *Institute for Collaborative Biotechnologies,University of California Santa Barbara, Santa Barbara,CA 93106-5080, USA (Tel: 650-539-8148; e-mail: bdaigle@gmail.com).*
*** *Departments of Computer Science and Mechanical Engineering, University of California Santa Barbara, Santa Barbara,CA 93106-5080, USA (Tel: 805-893-5362; e-mail: petzold@engr.ucsb.edu).*
**** *Department of Chemical Engineering, University of California Santa Barbara, Santa Barbara,CA 93106-5080, USA (Tel: 805-893-8133; e-mail: doyle@engr.ucsb.edu).*

Abstract: Advances in high throughput screening experiments have significantly improved our ability to discover and predict biomarkers for complex diseases. Systems biology approaches have played a critical role in realizing these improvements by providing computational tools for modeling such diseases at the network level. Within these tools, statistical scores such as the two sample t-statistic (t-score) are commonly used to rank genes/features for downstream analyses. In this paper, we propose a new alternative to the t-score —the ensemble sensitivity (ES) metric —which is a multivariate strategy to obtain feature rankings. To validate our method, we employ the COre Module Biomarker Identification with Network Exploration (COMBINER) tool on publicly available breast cancer gene expression data sets. Top candidates obtained by both the t-score and ES method serve as an input to COMBINER, which identifies the candidate biomarkers. Our results, as quantified by the COMBINER-generated area under the ROC curve (AUC), suggest that the ES approach improves the average AUC and identifies biomarkers with $\sim 93\%$ overlap with known cancer-related genes. In addition, the overlap of genes known to be associated with cancer that are identified using the two methods is small. This suggests that our proposed approach captures signals missed by methods relying on the t-score.

1. INTRODUCTION

DNA, proteins, and other small molecules interact within cells in a complex manner to regulate biological function. As one of the primary mechanisms for executing this function, the process of gene expression is responsible for the synthesis and maturation of all gene products in the cell. Consequently, the misregulation of this process is known to be the cause of a broad range of human diseases (Lee and Young [2013]). Fortunately, recent advances in high throughput screening technologies have significantly improved our ability to discover and predict prognostic and diagnostic disease biomarkers at the gene level. A crucial step in this task is the identification of differentially expressed genes (DEGs) between healthy and disease samples. However, this step is not trivial for the following reasons. First, gene products function in the context of interaction networks, where perturbations of a single gene can be propagated thoughout the network. Thus, disease-relevant "driver" DEGs are often accompanied by distantly related "passenger" DEGs. Second, there is a high degree of intrinsic variability between gene expression patterns of differing tissue samples and human patients. Third, even gene expression measurements from the same patient or sample are noisy, due to both intrinsic and extrinsic noise sources present within single cells (Kærn et al. [2005]), the intercellular environment, and the measurement apparatus. For all of these reasons, it is thus a significant challenge to robustly identify the small set of DEGs responsible for a disease phenotype from high throughput data.

An alternative strategy for identifying disease biomarkers using genome-scale data falls under the systems biology paradigm (Ideker et al. [2001], Chuang et al. [2010], Cho et al. [2012]). Specifically, pathway-based approaches utilize systems-level knowledge in the form of known biological pathways to improve biomarker inference. Rather than identifying only DEGs, these approaches (Breslin et al. [2005], Lee et al. [2008], Subramanian et al. [2005]) represent the disease phenotype with a differentially expressed

⋆ We gratefully acknowledge financial support from U.S. Army Research Office (PTSD Grant W911NF-10-2-0111).

group of genes, which belong to one or multiple pathways. By grouping genes with similar functions together, pathway-based approaches effectively reduce noise while also providing a better understanding of the underlying disease process (Cho et al. [2012]). A common strategy used by these methods is to first rank genes in each pathway by a univariate statistical score (e.g. the two-sample t-statistic or "t-score") that summarizes its level of differential expression between healthy and disease samples. Genes in each pathway are then aggregated together based on the significance of their scores and a composite score (also typically univariate) is generated. Finally, a subset of candidate pathway biomarkers with the most significant scores are selected and their predictive performance is evaluated with a machine learning classifier. However, due to the noise sources discussed above, some disease-relevant genes and pathways will exhibit relatively insignificant scores. In addition, certain genes that are known not to be significantly differentially expressed still play important roles by mediating connections between other disease-associated genes or pathways (Ideker et al. [2002]). Biomarkers from both of these classes would be excluded from the first two steps of pathway-based methods.

In this work, we propose a new multivariate ensemble-based algorithm to score genes and pathways for biomarker identification. Our approach employs supervised machine learning to generate importance scores for each gene/pathway (hereafter also referred to as "features") of a given ensemble. We then quantify the average sensitivity of these scores by removing one feature at a time (with replacement) and measuring changes in the remaining importance scores. Features for which, on average, the important scores increase are considered biologically relevant. We use this algorithm along with the COre Module Biomarker Identification with Network ExploRation (COMBINER) tool (Yang et al. [2012]) to robustly identify biomarkers for breast cancer from publicly available expression datasets. Our results show that, in comparison with univariate methods like the t-score, use of the ES algorithm leads to improved true positive rates of candidate biomarkers. In addition, we show that the ES algorithm identifies known cancer related genes that are largely non-overlapping with those discovered by univariate methods, suggesting that our proposed approach captures subtle biological signals that would otherwise be missed.

1.1 A short overview of COMBINER

COMBINER is a computational tool that enables the identification of disease gene and pathway biomarkers and the construction of their associated regulatory network. It takes multiple cohorts of gene expression data as input and identifies "core modules" (i.e. reproducible groups of pathway-associated genes) that best represent the expression differences between healthy and disease samples, thus providing insights into the disease mechanism. Briefly, COMBINER works by first projecting gene expression data from one cohort (the "inference dataset") onto known pathways. Next, a particular aggregation method (e.g. COndition Responsive Genes [CORG] (Lee et al. [2008]), Core Module Inference [CMI] (Yang et al. [2012]), Principal Component Analysis [PCA] (Jolliffe [2005])) is used to identify the active members of each pathway. Within

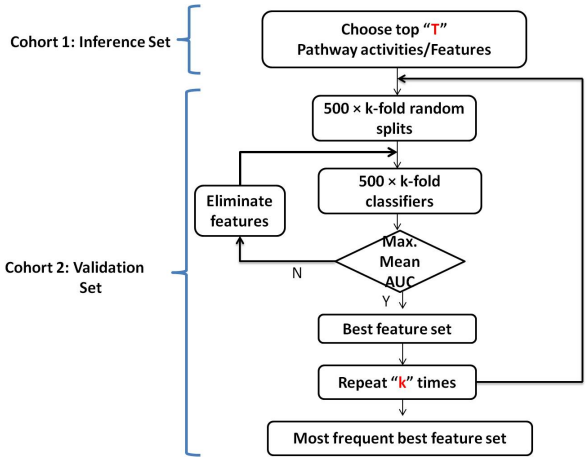

Fig. 1. The flow diagram for COMBINER

each pathway, the expression values of these active members are aggregated into an activity score, after which all activity scores are sorted and the top T "modules" are selected for downstream analysis. Given the inferred candidate modules, corresponding pathway activities are computed using data from the second cohort ("validation dataset"). As its name suggests, this cohort is used to train a supervised classifier and validate the predictive performance of the candidate module biomarkers. Details of the supervised classification is are follows. First, the validation dataset is partitioned into k-folds. One of these k folds is designated as a test set, and the remaining k-1 folds are used to construct the classifier. This partitioning procedure on the validation dataset is performed 500 times and thus we construct a total of 500× k-fold different classifiers. For each classifier, the decision boundary is constructed using linear discriminant analysis (LDA). We quantify the quality of the decision boundary using area under the Receiver Operating Characteristic curve (AUC). In the consensus feature elimination (CFE) module of COMBINER, the lowest ranked features are removed and the set of classifiers are again constructed (see Figure 1). The process is repeated until the maximum mean AUC is found. We note that such an method for computing the maximum AUC is a greedy approach and an optimal solution is thus not guaranteed. Finally, the set of pathways (and their constitutive active genes) corresponding to the maximum mean AUC is considered to comprise the core module biomarkers.

2. ENSEMBLE SENSITIVITY (ES) ALGORITHM

The proposed algorithm represents an alternative to univariate scoring methods (such as the "t-score") for an ensemble of N features. The method begins by constructing a supervised classifier using the feature ensemble and generating an importance score (i.e., classifier weight) for each feature. We then remove one feature at a time (with replacement) and estimate the change in scores of all other features due to this perturbation. We note that removing a feature from the ensemble is equivalent to projecting the N-dimensional data onto an (N-1)-dimensional subspace. After repeating this procedure N times (once for each feature), we compute the average changes in importance scores for all features. We consider those features with the

largest average changes to be biologically important in the disease or process being studied.

Mathematical details of the ES algorithm are as follows. Let us map the set of N features to a set of integers $\{1, 2, ..., N\}$. Define $\mathbf{S}^{unpert}$ as the vector of importance scores S_i of each feature in the ensemble and $\mathbf{S}^{pert}$ as the importance scores of $N-1$ features obtained by removing the j^{th} feature. For computational ease (see algorithm below), we increase the dimension of $\mathbf{S}^{pert}$ by unity by inserting $\mathbf{S}^{unpert}(j)$ after the $(j-1)^{th}$ component. The ES algorithm is given below. Upon completion of the algorithm, we sort the average score vector $\mathbf{S}_{avg}$ and choose a subset of features with the largest scores for downstream analysis. We note that estimation of feature importances can be further improved by also considering the effects of removing two or more features at a time. However, the combinatorial complexity of such an operation grows rapidly, and the computational cost quickly becomes prohibitive. Thus, as a first approximation, we consider only the removal of one feature at a time.

Algorithm 1 ES algorithm

$\mathbf{S}^{unpert} = S_{i \in T} \;\; T = \{1, 2, ..., N\}$
$\mathbf{S}_{avg} = \mathbf{0}_{N \times 1}$
for j= 1 :N **do**
  $\mathbf{S}^{pert} = S_{i \in \{T - \{j\}\}}$
  $\mathbf{S}^{pert}(j) = \mathbf{S}^{unpert}(j)$
  $\mathbf{\Delta} = \mathbf{S}^{pert} - \mathbf{S}^{unpert}$
  $\mathbf{S}_{avg} = \mathbf{S}_{avg} + \mathbf{\Delta}$
end for
$\mathbf{S}_{avg} = \frac{\mathbf{S}_{avg}}{N}$

3. SIMULATION RESULTS

Here, we demonstrate the applicability of the ES algorithm on three cohorts of publicly available breast cancer gene expression data, using COMBINER as an analysis tool. The human tissue samples for the cohorts were collected in three different countries, namely the Netherlands (Van De Vijver et al. [2002]), the United States (USA) (Wang et al. [2005]) and Belgium (Desmedt et al. [2007]). Each patient assayed in these datasets were monitored for the development of metastasis after five years of surgery. Out of 295, 286 and 198 patients in the Netherlands, USA and Belgium datasets, respectively, 78, 107 and 35 patients experienced metastasis. Given that data from each of the cohorts were collected using a slightly different microarray platform, only genes common to all three data sets were used in the analysis. We imputed missing data using the Bioconductor *impute* package, which implements a k-nearest neighbors imputation. We obtained biological pathway information using the MSigDB v4.0 Canonical Pathways subset (Subramanian et al. [2005], Liberzon et al. [2011]). For the purposes of this comparison, we used the CORG method (Lee et al. [2008]) to compute activity scores for each known pathway. Briefly, the CORG method first "z-transforms" the vector of expression data for each gene by subtracting the mean and dividing by the standard deviation. Next, expression vectors of genes belonging to each pathway are ranked in descending order of their two-sample t-statistics (t-scores) computed between the healthy and disease samples. Finally, the "pathway activity" vector is computed by averaging the expression vectors together starting at the top of the ranked list in a greedy fashion, stopping when the pathway activity t-score reaches a local maximum. Thus, the CORG method identifies the set of genes in each pathway (candidate "core module") that provides the (relative) maximum discriminating power for the disease phenotype. In this work, we sort the candidate core modules by their pathway activity t-scores and select the top 100 as the input to COMBINER.

To evaluate the ES method, we first select the top 300 candidate core modules and use the algorithm described above to choose a subset of size T=100 features. More specifically, we generate the importance score for each candidate core module using adaptive boosting of an ensemble of 150 tree-based classifiers with a learning rate of 1 (using the AdaBoostM1 implementation in the bioinformatics tool box of MATLAB). This use of adaptive boosting increases the weight of samples that are more difficult to classify in subsequent trees, leading to a more accurate overall classifier. We note that in general, adaptive boosting guarantees that the overall performance of an ensemble classifier is better than that of the individual weak learner (Freund and Schapire [1995], Polikar [2006]). Ultimately, each candidate core module receives an importance score quantifying its contribution to an accurate classifier.

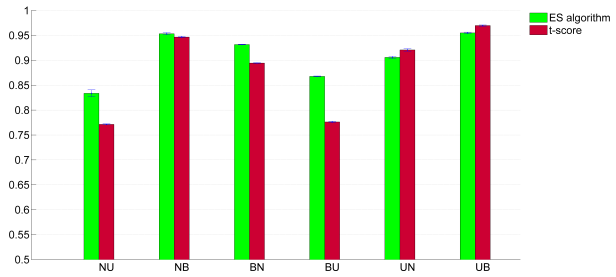

Fig. 2. **Comparison of the area under the ROC curve (AUC) for the t-score and ES algorithm methods.** Given the three data cohorts (Netherlands(N), USA(U) and Belgium(B)), a total of six inference/validation pairs can be made (shown on the x-axis). Using the inference dataset, we compute the pathway activities (using the CORG method) and select the top 300 candidates. Next, a subset of 100 candidates are selected based only on t-scores or using the EA algorithm. For these candidate core modules, we recompute the pathway activities using the validation dataset and construct a total of 500 LDA classifiers. Here we present the classifier accuracy in term of AUC $\pm$ standard deviation, obtained over 100 runs. On average, using core modules selected by the ES algorithm improves the AUC of the classifier by approximately 3%.

3.1 Comparison of classification accuracy

Given three cohorts of breast cancer gene expression data available, six possible inference/validation pairs are possible: $\{\{N, U\}, \{N, B\}, \{B, N\}, \{B, U\}, \{U, N\}, \{U, B\}\}$. In

Cancer gene Data bases	Threshold=100 t-score % of N=110	Threshold=100 ES % of N=182	Threshold=120 t-score % of N=42	Threshold=120 ES % of N=57	Threshold=220 t-score % of N=21	Threshold=220 ES % of N=15
NetPath	40.74	47.80	56.10	50.88	71.43	73.33
Atlas	50.93	43.96	48.78	47.37	52.38	33.33
Census	4.63	7.69	4.88	8.77	4.76	13.33
CANgene	1.85	1.10	2.44	0	4.76	0
G2SBC	68.52	54.94	60.98	45.61	47.62	40.00
COSMIC	8.33	13.19	7.32	5.26	9.52	13.33
KEGG	14.81	20.33	24.39	24.56	42.86	46.67
NCG	11.11	11.54	12.19	10.53	19.05	20.00

Table 1. **Enrichment rates of known cancer genes within biomarkers identified using COMBINER with t-scores only, and COMBINER with the ES algorithm**. Candidate biomarker genes are those that belong to the 95% most frequently chosen "core modules" over 100 runs of COMBINER for a given pair of inference-validation datasets. A common subset of genes occurring more often than the threshold of $\{100, 120, \text{or } 220\}$ times for every inference-validation pair is considered to comprise the predicted biomarker genes. N represents the total number of predicted biomarkers.

Threshold	No. predicted Biomarkers t-score	No. predicted Biomarkers ES	No. of known CG t-score	No. of known CG ES	Known CG by t-score and ES algorithm	Known CG by only t-score	Known CG by only ES algorithm
100	110	182	84 (76.36%)	142 (78.02%)	51	33	91
120	42	57	34 (80.95%)	48 (84.21%)	12	22	36
220	21	15	19 (90.48%)	14 (93.33%)	2	17	12

Table 2. **Comparison of the statistics of the breast cancer biomarkers predicted by t-score ranking and ES algorithm ranking**. CG represents cancerous genes.

Figure 2 we compare the maximum mean AUC obtained for the two different sets of top 100 core modules obtained using either the t-score or ES algorithm rankings. Compared to the t-score ranking, the AUC achieved with the ES algorithm is better in four out of six cases. The means of the maximum average AUC over all inference/validation datasets is 0.9078 and 0.8798 for the ES algorithm and the t-score ranking, respectively.

3.2 Comparison of the enriched cancer-related genes

The high average AUC indicates that the proportion of true positive cancer-related genes among all biomarkers should be large. To obtain the set of predicted biomarker genes, we perform the following four steps. First, we select those core modules that were identified by COMBINER in at least 95/100 runs for each inference/validation pair. Second, we collect all the genes in these core modules for each inference/validation pair. Third, we identify the subset of genes common to all of the possible inference/validation pairs. Finally, we define an integer cutoff and select those genes in the above subset that occur more frequently than that cutoff.

We used the following databases to perform the cancer gene enrichment analysis: (1) NethPath (Kandasamy et al. [2010]) (all cancers), (2) Atlas of Cancer Genes (Huret et al. [2000]) (all cancers), (3) Census Genes (Futreal et al. [2004]) (all cancer), (4) CANgenes (Sjöblom et al. [2006])(breast cancer), (5) G2SBC (Mosca et al. [2010]) (breast cancer), (6) KEGG Pathways of Cancer (Kanehisa and Goto [2000]) (all cancers) and (7) Network of Cancer Gene (D'Antonio et al. [2012]) (all cancer). Table 1 shows the enrichment in our predicted biomarkers of known cancer-related genes from various databases. We consider three different cutoffs for selecting biomarkers: 100, 120 and 220. Since a given gene may belong to multiple pathways, the total number of occurrences for a given gene may be higher than the total number of runs. Our results indicate that as the cutoff is increased, the likelihood that a predicted biomarker is a true positive also increases. The table provides the percentage of the total predicted biomarkers that overlap with different cancer-related gene databases. In Table 2 we provide the statistics of the overall true positive predictions made by COMBINER using different set of input features. For a threshold of 220, we found that approximately 93 percent (14/15) of the biomarkers identified using the ES algorithm overlap with known cancer-related genes. This equates to a 3 percent improvement in the total percentage overlap versus the t-score ranking method. Importantly, we also note that the two ranking methods lead to a small overlap of predicted biomarkers, suggesting that the two methods capture complementary signals from the expression data. Further investigation confirms this hypothesis, as we discovered that the ES algorithm selects pathway activities that have relatively lower t-scores. Figure 3 shows these results for the six different inference/validation data cohorts.

4. CONCLUSION

In summary, we have proposed the ensemble sensitivity (ES) algorithm, a new strategy for selecting features from a given ensemble. By computing the average change in feature importance scores due to the removal of each individual feature, the ES algorithm provides a multivariate

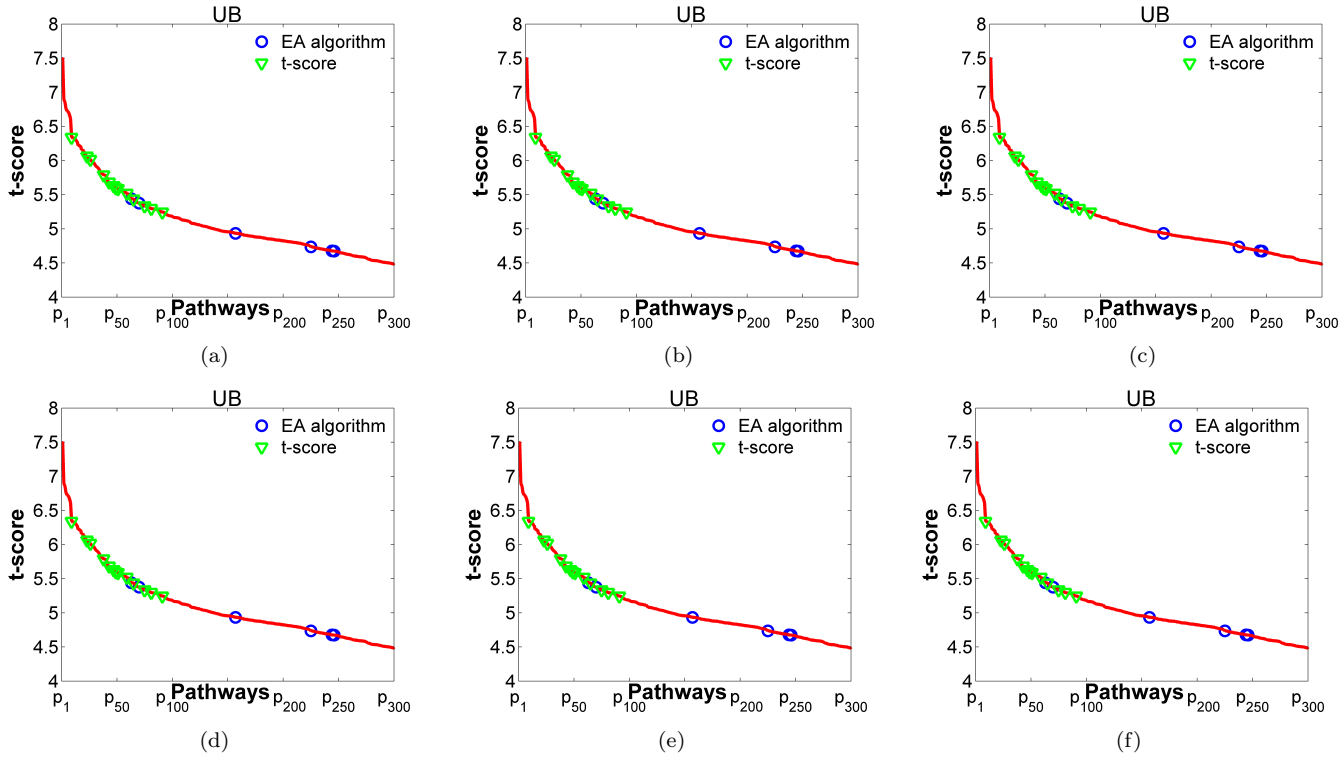

Fig. 3. **Comparison of pathway activity selection: t-score ranking method versus the ES algorithm**. The x-axis represents a set of pathway activities sorted by t-scores for the given inference set and the y-axis represents their corresponding scores. The selection process for the core modules begins with the top 100 candidates given by the t-score method or the ES algorithm. Those modules that are poorly ranked by the LDA classifier are recursively removed. The candidate core modules leading to the maximum area under the ROC curveare considered important. Across 100 different runs of COMBINER for a given inference-validation dataset pair, we plot only those core modules that were selected at least 95 times. (a-f) Inference on set {{N},{N},{B},{B},{U},{U}} and their corresponding validation datasets {{U},{B},{N},{U},{N},{B}}.

scoring scheme that is not offered by more commonly-used univariate methods. We compared the ES algorithm to the univariate t-score ranking method by combining both approaches individually with the biomarker-predicting tool COMBINER. Based on an analysis of three human breast cancer gene expression datasets, we find that the ES algorithm improves the average AUC achieved. In addition, the biomarkers identified by the two methods have minimal overlap, suggesting that the two methods are able to capture two different types of biological signals. Specifically, the ES algorithm appears to capture features that have relatively lower statistical discriminative ability but still have important biological function. Further investigation into this direction is currently underway.

REFERENCES

T. Breslin, M. Krogh, C. Peterson, and C. Troein. Signal transduction pathway profiling of individual tumor samples. *BMC Bioinformatics*, 6:163, 2005.

D.Y. Cho, Y.A. Kim, and T.M. Przytycka. Network biology approach to complex diseases. *PLoS Computational Biology*, 8(12):e1002820, 2012.

H.Y. Chuang, M. Hofree, and T. Ideker. A decade of systems biology. *Annual Review of Cell and Developmental Biology*, 26:721, 2010.

M. D'Antonio, V. Pendino, S. Sinha, and F.D. Ciccarelli. Network of cancer genes (ncg 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Research*, 40(D1):D978–D983, 2012.

C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M.S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A.L. Harris, J.G.M Klijn, J.A. Foekens, F. Cardoso, M.J. Piccart, M. Buyse, and C. Sotiriou. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical Cancer Research*, 13 (11):3207–3214, 2007.

Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory*, pages 23–37. Springer, 1995.

P.A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M.R. Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3): 177–183, 2004.

J.L. Huret, S. Le Minor, F. Dorkeld, P. Dessen, and A. Bernheim. Atlas of genetics and cytogenetics in oncology and haematology, an interactive database. *Nucleic Acids Research*, 28(1):349–351, 2000.

T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics*, 2(1):343–372, 2001.

T. Ideker, O. Ozier, B. Schwikowski, and A.F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl 1):S233–S240, 2002.

I. Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2005.

M. Kærn, T.C. Elston, W.J. Blake, and J.J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464, 2005.

K. Kandasamy, S.S. Mohan, R. Raju, S. Keerthikumar, G.S.S. Kumar, A.K. Venugopal, D. Telikicherla, J.D. Navarro, S. Mathivanan, C. Pecquet, S.K. Gollapudi, S.G. Tattikota, S. Mohan, H. Padhukasahasram, Y. Subbannayya, R. Goel, H.K.C. Jacob, J. Zhong, R. Sekhar, V. Nanjappa, L. Balakrishnan, R. Subbaiah, Y.L. Ramachandra, B.A. Rahiman, T.S.K Prasad, J.X Lin, J.C.D Houtman, S. Desiderio, J.C. Renauld, S.N. Constantinescu, O. Ohara, T. Hirano1, M. Kubo, S. Singh, P. Khatri, S. Draghici, G.D. Bader, C. Sander, W. J. Leonard, and A. Pandey. Netpath: a public resource of curated signal transduction pathways. *Genome Biology*, 11(1):R3, 2010.

M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

E. Lee, H.Y. Chuang, J.W. Kim, T. Ideker, and D. Lee. Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, 4(11):e1000217, 2008.

T.I. Lee and R.A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, 2013.

A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J.P. Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27 (12):1739–1740, 2011.

E. Mosca, R. Alfieri, I. Merelli, F. Viti, A. Calabria, and L. Milanesi. A multilevel data integration resource for breast cancer study. *BMC Systems Biology*, 4(1):76, 2010.

R. Polikar. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6(3):21–45, 2006.

T. Sjöblom, S. Jones, L.D. Wood, D.W. Parsons, J. Lin, T.D. Barber, D. Mandelker, R.J. Leary, J. Ptak, N. Silliman, S. Steve Szabo1, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J.K.V. Willson, A.F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B.H. Park, K.E Bachman, N. Papadopoulos, B. Vogelstein, K.W. Kinzler, and V.E Velculescu1. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314(5797):268–274, 2006.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

M.J. Van De Vijver, Y.D. He, L.J. van't Veer, H. Dai, A.A.M. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. Van der Velde, H. Bartelink, S. Rodenhuis, E.T. Rutgers, S.H. Freind, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

Y. Wang, J.G.M. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang, D. Talantov, M. Timmermans, M.E. Meijer-van Gelder, J. Yu, T. Jatkoe, E.M. Berns, D. Atkins, and J.A. Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, 2005.

R. Yang, B.J. Daigle Jr., L.R. Petzold, and F.J. Doyle III. Core module biomarker identification with network exploration for breast cancer metastasis. *BMC Bioinformatics*, 13(1):12, 2012.