A systems biology approach for unsupervised clustering of high-dimensional data

Diana Diaz[1], Tin Nguyen[1], and Sorin Draghici[1,2,*]

[1] Wayne State University, Computer Science, Detroit MI 48202
[2] Wayne State University, Obstetrics and Gynecology, Detroit MI 48202

Abstract. One main challenge in modern medicine is the discovery of molecular disease subtypes characterized by relevant clinical differences, such as survival. However, clustering high-dimensional expression data is challenging due to noise and the curse of high-dimensionality. This article describes a disease subtyping pipeline that is able to exploit the important information available in pathway databases and clinical variables. The pipeline consists of a new feature selection procedure and existing clustering methods. Our procedure partitions a set of patients using the set of genes in each pathway as clustering features. To select the best features, this procedure estimates the relevance of each pathway and fuses relevant pathways. We show that our pipeline finds subtypes of patients with more distinctive survival profiles than traditional subtyping methods by analyzing a TCGA colon cancer gene expression dataset. Here we demonstrate that our pipeline improves three different clustering methods: k-means, SNF, and hierarchical clustering.

This is an author-version of an article presented at International Workshop on Machine Learning, Optimization and Big Data 2016. The final publication is available at http://www.springerlink.com

1 Introduction

Identifying homogeneous subtypes in complex diseases is crucial for improving prognosis, treatment, and precision medicine [1]. Disease subtyping approaches have been developed to identify clinically relevant subtypes. High-throughput technologies can measure the expression of more than ten thousand genes at a time. Subtyping patients using the whole-genome scale measurement is challenging due to the curse of high-dimensionality. Several clustering methods have been developed [2–5] to handle this type of high-dimensional data. Other approaches, such as iCluster [6], rely on feature selection to reduce the complexity of the problem.

There are many widely used feature selection methods [7–11]. The simplest way to perform unsupervised feature selection for subtyping is by ranking the list of genes and filtering out those with low rankings. For example, genes can be ranked using Fisher score-based methods [8, 9] or t-test based methods [10].

Other methods, such as [11], use general purpose filtering metrics like Information Gain [12], Consistency [13], Chi-Squared [14] and Correlation-Based Feature Selection [15]. These filter-based methods are computationally efficient, but they do not account for dependency between genes or features. To address this, wrapper methods [16, 17] use learning algorithms to find subsets of related features or genes. Even though these methods consider feature dependency, they have a high degree of computational complexity due to repeated training and testing of predictors. This makes them impractical for analyzing high-dimensional data.

Meanwhile, some approaches incorporate to gene-expression-based subtyping other types of data such as clinical variables [18–20] and multi 'omics' data [6, 21, 22]. These types of data are more and more available nowadays. Large public repositories, including the Cancer Genome Atlas (TCGA) (cancergenome.nih.gov), accumulate clinical and multi 'omics' data from thousands of patients. Clinical variables used for subtyping include survival data [18], epidemiological data [19], clinical chemistry evaluations and histopathologic observations [20]. These variables have shown to provide useful information for a better subtyping.

Subtyping patients using gene expression data has additional challenges because genes do not function independently. They function in synchrony to carry on complex biological processes. Knowledge of these processes is usually accumulated in biological pathway databases, such as KEGG [23] and Reactome [24]. Biological pathways are graphical representations of common knowledge about genes and their interactions on biological processes. This valuable information has been used to cluster related genes using gene expression [25–28] and should be used to identify disease subtypes as well. Clinical data and biological knowledge are complementary to gene expression and can leverage disease subtyping.

Here we present a disease subtyping pipeline that includes a new feature selection approach and any existing unsupervised clustering method. To the best of our knowledge, this is the first approach that integrates pathway knowledge and clinical data with gene expression for disease subtyping. Our framework is validated using gene expression and clinical data downloaded from the Cancer Genome Atlas (TCGA) and pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG). Using the features selected with our approach and three different clustering methods (k-means, SNF, and hierarchical clustering), our pipeline is able to identify subtypes that have significantly different survival profiles. This pipeline was developed in R programming language. The source code is available on github (http://datad.github.io/disSuptyper) to ease the reproducibility of the methods presented here [29, 30].

2 Method

In this section, we introduce a new feature selection framework for disease subtyping. Figure 1 presents the overall pipeline of our framework. The input includes i) gene expression data, ii) survival data, and iii) biological pathways (see

Figure 1a). The output is a set of selected genes (Figure 1f) for finding subtypes with significantly distinct survival patterns (Figure 1g).

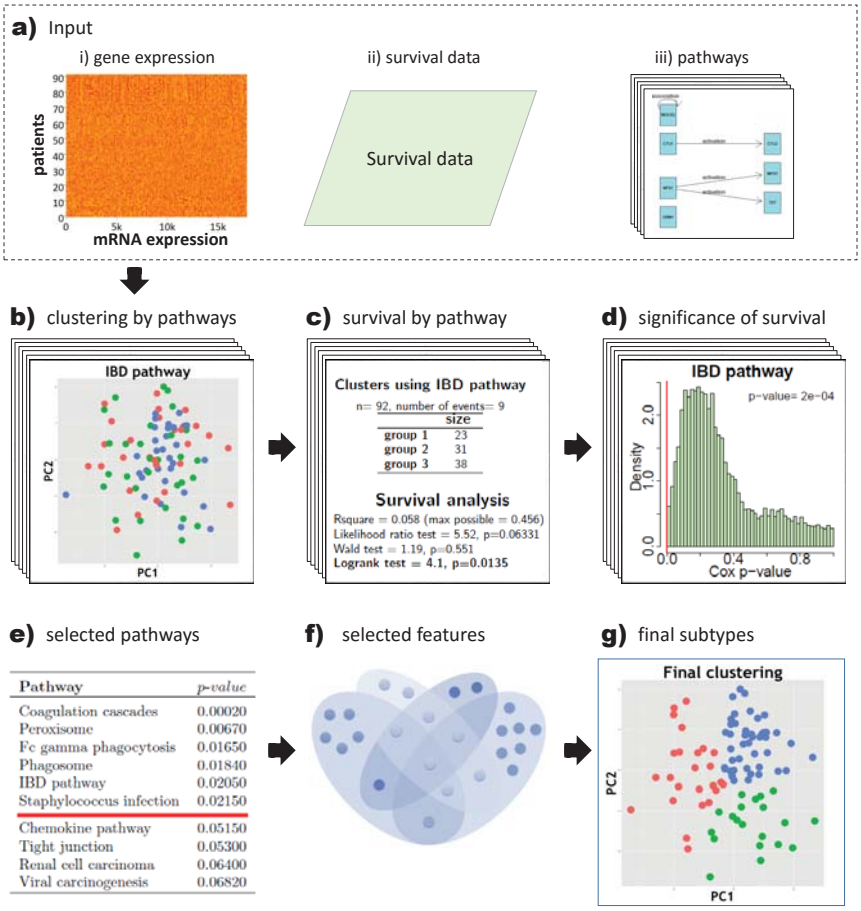
Gene expression data can be represented as a matrix $D \in R^{M \times N}$, where the rows are different patients having the same disease and columns are different features (i.e. genes). M is the number of patients and N is the number of genes. For gene expression data, N can be as large as $20,000$. The survival data include patient's vital status (dead or alive) and follow-up information (time and censored/uncensored). The biological pathways are collected from public pathway databases. In this work, our data analysis are based on KEGG pathways [23], but other databases can also be used.

First, we partition the rows (patients) of gene expression matrix D using the features provided by each pathway in the pathway database (Figure 1b). Formally, let us denote $\mathbf{P}$ as the pathway database which has $n = |\mathbf{P}|$ signaling pathways. We have $\mathbf{P} = \{P_i\}$ where $i \in [1..n]$. For each pathway P_i, we cluster the rows using genes that belong to the pathway P_i as features resulting in a partitioning C_i.

Second, we perform survival analysis on each of the pathway-based clusterings C_i (Figure 1c). We calculate Cox log-rank p-value for the subtypes defined by C_i using the input survival information. This Cox p-value represents how likely the survival curves' difference is observed by chance. So far, we have n Cox p-values, one per pathway.

Now the question is whether the features provided by the pathway P_i help to better differentiate the subtypes. We will answer this question by using random sampling technique. Denote $|P_i|$ as the number of genes in the pathway P_i. We randomly select $|P_i|$ genes from the original set of N genes. We partition the patients using this randomly selected set of genes and then compute the Cox p-value. We repeat this random selection $10,000$ times which results in a distribution that has $10,000$ Cox p-values (Figure 1d). This distribution represents the distribution of Cox p-values when randomly selecting $|P_i|$ features for subtyping. In Figure 1d, the vertical red line shows the real Cox p-value calculated from the actual genes in P_i, whereas the green distribution shows the $10,000$ random Cox p-values. Now we compare the Cox p-value obtained from the pathway P_i with the distribution of randomly selected genes. We estimate the probability of obtaining this Cox p-value (using genes in P_i) by computing the ratio of the area to the left of this Cox p-value divided by the total area of the distribution. We denote this probability as p_i. In total, we have n values $\{p_i, i \in [1..n]\}$, one for each pathway. Each of these p-values p_i quantifies how likely it is to observe by chance a Cox log-rank statistic as extreme or more than the one observed. Therefore, this p-value of a pathway P_i represents how likely the features provided by the pathway help to improve the subtyping.

The third step is to choose a set of pathways that certainly help to improve the subtyping. To do this, we adjusted the p-values for multiple comparisons using False Discovery Rate (FDR), we rank the set of pathways and select those that have the corresponding nominal *p-values* less than or equal to the significance threshold of 5%. Let us name the pathways yielding significantly distinct



Pathway	*p-value*
Coagulation cascades	0.00020
Peroxisome	0.00670
Fc gamma phagocytosis	0.01650
Phagosome	0.01840
IBD pathway	0.02050
Staphylococcus infection	0.02150
Chemokine pathway	0.05150
Tight junction	0.05300
Renal cell carcinoma	0.06400
Viral carcinogenesis	0.06820

Fig. 1: New feature selection pipeline for disease subtyping using biological knowledge. (a) The input includes i) gene expression data, ii) survival data, and iii) pathways downloaded from a database. (b) First, we partition the gene expression data using the set of genes in each pathway as features. (c) Second, we perform survival analysis on each resulting partition. (d) Third, we compute the p-value that represents how likely the pathway improves the subtyping. (e) Fourth, we rank the list of pathways by corrected p-value and select pathways that have a nominal p-value less than or equal to the significance threshold 5%. (f) Fifth, we merge the relevant pathways to construct the final set of features. (g) Finally, we subtype the patients using the selected features. The clustering is demonstrated in the first two principal components, but we use all dimensions/genes for clustering. Note: IBD pathway stands for Inflammatory Bowel Disease pathway.

Table 1: List of pathways selected by our approach when using RSS k-means. We first ranked the pathways by FDR adjusted p-value ($p\text{-}value.fdr$), then selected the pathways with a nominal $p\text{-}value \leq 0.05$ as relevant pathways.

Pathway	$p\text{-}value$	$p\text{-}value.fdr$
Complement and coagulation cascades	0.00020	0.03680
AGE-RAGE signaling pathway in diabetic complications	0.00420	0.38640
Peroxisome	0.00670	0.41093
Cytokine-cytokine receptor interaction	0.01040	0.45448
Fc gamma R-mediated phagocytosis	0.01650	0.45448
Phagosome	0.01840	0.45448
Inflammatory bowel disease (IBD)	0.02050	0.45448
Staphylococcus aureus infection	0.02150	0.45448
Leukocyte transendothelial migration	0.02330	0.45448
NF-kappa B signaling pathway	0.03710	0.50048
Renin secretion	0.03850	0.50048
Malaria	0.04780	0.51326
Platelet activation	0.06980	0.54970

survival curves as *relevant pathways*. For example, In Figure 1e, the horizontal red line shows the significance threshold of 5%. In this example, the relevant pathways are *Coagulation cascades*, *Peroxisome*, *Fc gamma phagocytosis*, *Phagosome*, *Inflammatory Bowel Disease (IBD) pathway*, and *Staphylococcus infection*.

Considering all the genes in the relevant pathways as favorable features, we merge these pathways to get a single set of genes (Figure 1f). We use this merged set of genes as the selected features for our final subtyping. In our example, the final selected genes are the genes in the six pathways listed above. We then use these genes to construct the final clustering as shown in Figure 1g.

We note that this feature selection procedure can be used in conjunction with any clustering method. In our experimental studies, we used three clustering methods that belong to different clustering models. The first method is the classical k-means. It is well-known that k-means does not always converge to a global optimal point, it depends on the initialization. To overcome this problem, we ran k-means several times and chose the partitioning that has the smallest residual sum of squares (RSS). In the rest of the manuscript, we refer to this as "RSS k-means". The second method is Similarity Network Fusion (SNF) [4], which is based on spectral clustering. The third one is the traditional hierarchical clustering using cosine similarity as the distance function. We will show that our framework helps to improve the subtyping using any of the three mentioned clustering methods.

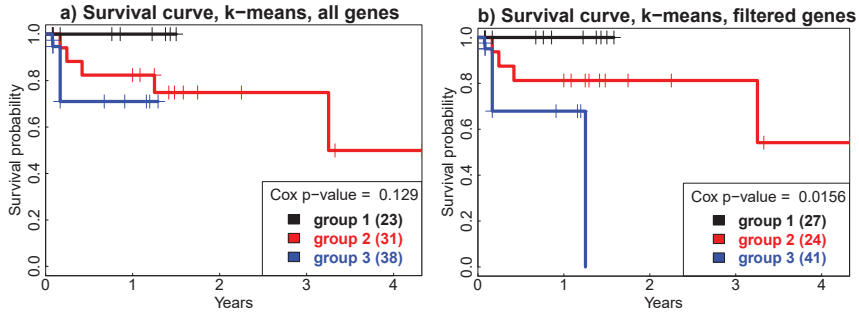


Fig. 2: Kaplan-Meier survival analysis of the obtained subtypes using RSS k-means algorithm. a) Survival curves using all genes. b) Survival curves using selected genes.

3 Results

In this section,we assess the performance of our feature selection for disease subtyping framework using gene expression data (Agilent G4502A-07 platform level 3) generated by the Cancer Genome Atlas (TCGA) (cancergenome.nih.gov). We selected the samples that have miRNA and methylation measurements as were selected in SNF [22]. A copy of the dataset is available in the github repository (http://datad.github.io/disSuptyper). The number of patients is $M = 92$, and the number of genes is $N = 17,814$. For all the performed clusterings, we set the number of clusters as $k = 3$ according to prior knowledge of the number of subtypes of colon cancer [4]. When running our method, we used 184 pathways from the KEGG pathway database [23].

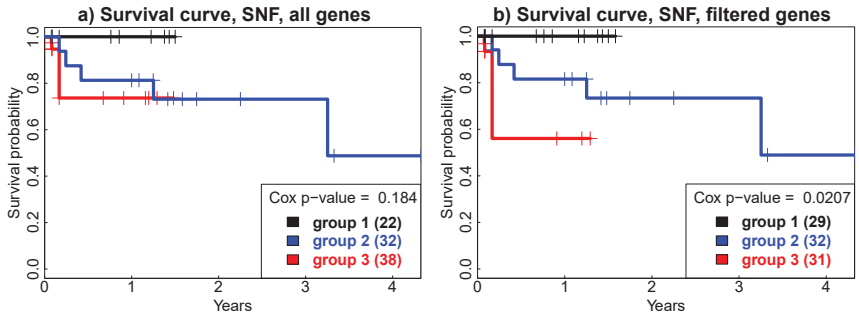
As described in Section 2, our framework can be used in conjunction with any unsupervised clustering algorithm. Here we test it using three clustering methods: RSS k-means, SNF [22], hierarchical clustering [2]. For all clustering methods, we first clustered the patients using all the measured genes, then clustered the patients using only the genes selected by our technique. To contrast the difference between the three traditional clustering methods and our pipeline results, we performed survival analysis for all the cases using Kaplan-Meier analysis and Cox p-value.

3.1 Subtyping using k-means

We clustered the patients from the TCGA colon adenocarcinoma dataset using our pipeline in conjunction with RSS k-means. We used the 184 signaling pathways from the KEGG database [23]. For each pathway P_i, we partitioned the patients using the genes in the pathway P_i as features to get a clustering C_i.

After this step, we got a total of 184 clusterings, one per pathway. Also for each pathway, we constructed the empirical distribution and then estimated

Fig. 3: Kaplan-Meier survival analysis of the obtained subtypes using SNF. a) Survival curves using all genes. b) Survival curves using the selected genes.



the *p-value* of how likely the pathway helps to improve disease subtyping. The *p-values* of relevant pathways are shown in Table 1. The horizontal red line represents the significance cutoff at 5%. There are 12 relevant pathways. We then merged the relevant pathways to get a single set of genes that we used as clustering features. This final set of features consists of 851 genes when using RSS k-means algorithm. Finally, we performed RSS k-means clustering using these 851 genes.

Figure 2 shows the survival analysis of the resultant clusterings. Figure 2a shows the resultant clustering when using RSS k-means for all 17,814 genes. The Cox p-value of this clustering is 0.129, which is not significant. Figure 2b shows the resultant clustering using the 851 selected genes. The resultant Cox p-value is 0.0156, which is approximately ten times lower than using all genes.

3.2 Subtyping using SNF

Similar to the assessment performed for k-means, we clustered the patients from the TCGA colon adenocarcinoma dataset using our pipeline in conjunction with SNF. To perform SNF clustering, we ran the SNFtool Bioconductor package with the parameters suggested by the authors [4]. We used the same input (KEGG pathways), settings (three clusters), and process previously described.

After this step, we obtained 184 clusterings, one per pathway. Then for each pathway, we constructed the empirical distribution and estimated the *p-value* of how likely the pathway helps to improve disease subtyping. The estimated *p-values* are shown in Table 2. The horizontal red line represents the significance threshold of 5%. There are 10 relevant pathways. We merged these relevant pathways to get a single set of genes that we used as our final set of selected features. This feature set contains 764 genes for SNF method. Finally, we performed SNF clustering using these 764 genes.

Figure 3 shows the survival analysis of the resultant clusterings. Figure 3a shows the clustering when using SNF for all 17,814 genes. The Cox p-value of

Table 2: List of pathways that contain relevant genes obtained with our approach when using SNF. We first ranked the pathways by $p\text{-}value.fdr$, then selected the pathways with a nominal $p\text{-}value \leq 0.05$.

Pathway	$p\text{-}value$	$p\text{-}value.fdr$
HTLV-I infection	0.00400	0.37765
Endocrine and other factor-regulated calcium reabsorption	0.00680	0.37765
Complement and coagulation cascades	0.00800	0.37765
Aldosterone-regulated sodium reabsorption	0.00830	0.37765
AMPK signaling pathway	0.01410	0.51324
Phagosome	0.02150	0.54196
Fc epsilon RI signaling pathway	0.02290	0.54196
Cytosolic DNA-sensing pathway	0.02680	0.54196
Peroxisome	0.03900	0.61320
Leishmaniasis	0.04300	0.61320
Non-alcoholic fatty liver disease (NAFLD)	0.05400	0.66544

this clustering is 0.1836, which is not significant (this resultant is identical to the result reported in [4]). Figure 3b shows the resultant clustering when using the 764 selected genes. The Cox p-value is 0.0207, which is approximately ten times lower than using all genes. Despite this meaningful improvement, none of the pathways has a corrected $p\text{-}value.fdr \leq 0.05$. This shows a lack of statistical power on our approach and an opportunity for improvement.

3.3 Subtyping using hierarchical clustering

Alike the assessment performed previously, we clustered the colon adenocarcinoma patients using our pipeline in conjunction with Hierarchical Clustering (HC) [2]. We used the 184 signaling pathways from KEGG [23]. The estimated $p\text{-}values$ of the relevant pathways obtained with HC are shown in Table 3. The horizontal red line represents the significance threshold of 5%. We merged these three relevant pathways to get our final set of selected features. This feature set contains 195 genes for HC. Finally, we performed hierarchical clustering using the selected genes only.

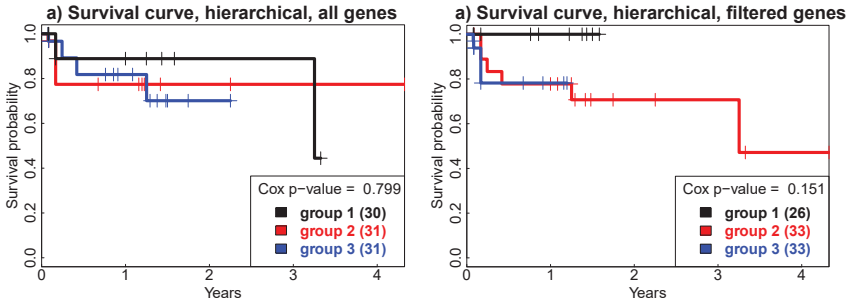
Figure 4 shows the survival analysis of the resultant clusterings. Figure 4a shows the clustering when using HC for all $17,814$ genes. The Cox p-value of this clustering is 0.799 which is not significant. Figure 4b shows the resultant clustering when using the 195 selected genes. The Cox p-value is 0.151 which is lower than using all genes, but it is still not significant. The subtypes obtained with hierarchical clustering do not separate the patients in clinically meaningful subtypes in any of the cases (neither using all genes nor filtered genes).

Given that our approach requires resampling for computing the $p\text{-}values$ p_i, this pipeline is more time consuming than traditional approaches. For the computational experiments presented here, we generated $10,000$ random samplings

Table 3: List of pathways selected by our approach when using hierarchical clustering. We first ranked the pathways by FDR adjusted p-value ($p\text{-}value.fdr$), then selected the pathways with a nominal $p\text{-}value \leq 0.05$ as relevant pathways.

Pathway	$p\text{-}value$	$p\text{-}value.fdr$
Cytosolic DNA-sensing pathway	0.01140	0.63874
Peroxisome	0.01200	0.63874
Fc epsilon RI signaling pathway	0.04090	0.63874
Complement and coagulation cascades	0.12390	0.80770

Fig. 4: Kaplan-Meier survival analysis of the obtained subtypes using hierarchical clustering (HC). a) Survival curves using traditional HC. b) Survival curves using HC in our pipeline.



and clusterings per each pathway (184 pathways in total). Our pipeline took several hours to subtype the set of patients (about 8 hours for k-means, 17 hours for SNF, and 46 hours for hierarchical clustering) while running any traditional clustering method takes only some minutes (less than 6 minutes). We ran these experiments on a typical desktop workstation with a 2.6 GHz Intel Core i5, 8GB of RAM, on a single thread, and the OS X 10.11 operative system.

4 Conclusions

In this article, we describe a framework to combine gene expression data, survival data, and biological knowledge available in pathway databases for a better disease subtyping. The performance of the new approach was demonstrated on the colon adenocarcinoma data downloaded from TCGA. The described framework was tested in conjunction with k-means, Similarity Network Fusion (SNF) and hierarchical clustering. For these clustering algorithms, our approach greatly improves the subtyping. In all cases, the Cox p-value is folds lower when using the selected features. Cox p-value improved from 0.129 to 0.0156 for k-means, from 0.184 to 0.0207 for SNF, and from 0.799 to 0.151 for hierarchical clustering.

Our contribution is two-folds. First, this framework introduces a way to exploit the additional information available in biological databases. Although the framework was demonstrated on KEGG pathways, it can exploit information available in other databases, such as functional modules available in Gene Ontology database or protein-protein interactions available in the STRING database. Second, this framework is the first one that integrates clinical data, biological pathways, and gene expression data for disease subtyping. For future work, we plan to use other clinical variables besides survival information and integrate multiple datatypes, such as microRNA, for a more comprehensive analysis [31]. Additionally, we plan to analyze the performance of feature selection methods from other contexts into the context of disease subtyping.

Acknowledgments. This study used data generated by the TCGA Research Network; we thank donors and research groups for sharing these valuable data. This research was supported in part by the following grants: NIH R01 DK089167, R42 GM087013 and NSF DBI-0965741, and by the Robert J. Sokol Endowment in Systems Biology. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

References

1. Saria, S., Goldenberg, A.: Subtyping: What It is and Its Role in Precision Medicine. IEEE Intelligent Systems **30**(4) (2015) 70–75
2. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences **95**(25) (1998) 14863–14868
3. Kim, E.Y., Kim, S.Y., Ashlock, D., Nam, D.: MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. BMC Bioinformatics **10** (2009) 260
4. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.: Similarity network fusion for aggregating data types on a genomic scale. Nature Methods **11**(3) (2014) 333–337
5. Hsu, J.J., Finkelstein, D.M., Schoenfeld, D.A.: Outcome-Driven Cluster Analysis with Application to Microarray Data. PLOS ONE **10**(11) (November 2015) e0141874
6. Shai, R., Shi, T., Kremen, T.J., Horvath, S., Liau, L.M., Cloughesy, T.F., Mischel, P.S., Nelson, S.F.: Gene expression profiling identifies molecular subtypes of gliomas. Oncogene **22**(31) (2003) 4918–4923
7. Hira, Z.M., Gillies, D.F., Hira, Z.M., Gillies, D.F.: A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data, A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Advances in Bioinformatics, Advances in Bioinformatics **2015, 2015** (June 2015) e198363
8. Huang, G.T., Cunningham, K.I., Benos, P.V., Chennubhotla, C.S.: Spectral clustering strategies for heterogeneous disease expression data. Pacific Symposium on Biocomputing (2013) 212–223

9. Pyatnitskiy, M., Mazo, I., Shkrob, M., Schwartz, E., Kotelnikova, E.: Clustering Gene Expression Regulators: New Approach to Disease Subtyping. PLoS ONE **9**(1) (2014)
10. Li, T., Zhang, C., Ogihara, M.: A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics **20**(15) (2004) 2429–2437
11. Hernández-Torruco, J., Canul-Reich, J., Frausto-Solís, J., Méndez-Castillo, J.J.: Feature Selection for Better Identification of Subtypes of Guillain-Barré Syndrome. Computational and Mathematical Methods in Medicine, Computational and Mathematical Methods in Medicine **2014, 2014** (September 2014) e432109
12. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Comput. Surv. **34**(1) (March 2002) 1–47
13. Liu, Y., Schumann, M.: Data mining feature selection for credit scoring models. Journal of the Operational Research Society **56**(9) (April 2005) 1099–1108
14. Zheng, Z., Wu, X., Srihari, R.: Feature Selection for Text Categorization on Imbalanced Data. SIGKDD Explor. Newsl. **6**(1) (June 2004) 80–89
15. Hall, M.A.: Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato (1999)
16. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics **7** (2006) 3
17. Sharma, A., Imoto, S., Miyano, S., Sharma, V.: Null space based feature selection method for gene expression data. International Journal of Machine Learning and Cybernetics **3**(4) (2011) 269–276
18. Bair, E., Tibshirani, R.: Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. PLOS Biol **2**(4) (April 2004) e108
19. Paoli, S., Jurman, G., Albanese, D., Merler, S., Furlanello, C.: Integrating gene expression profiling and clinical data. International Journal of Approximate Reasoning **47**(1) (January 2008) 58–69
20. Bushel, P.R., Wolfinger, R.D., Gibson, G.: Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. BMC Systems Biology **1** (2007) 15
21. Chalise, P., Koestler, D.C., Bimali, M., Yu, Q., Fridley, B.L.: Integrative clustering methods for high-dimensional molecular data. Translational cancer research **3**(3) (June 2014) 202–216
22. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.: Similarity network fusion for aggregating data types on a genomic scale. Nature Methods **11**(3) (2014) 333–337
23. Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research **28**(1) (2000) 27–30
24. Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., D'Eustachio, P.: The Reactome pathway knowledgebase. Nucleic Acids Research **42**(D1) (2014) D472–D477
25. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T.: Co-clustering of biological networks and gene expression data. Bioinformatics **18**(suppl 1) (July 2002) S145–S154
26. Huang, D., Pan, W.: Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics **22**(10) (May 2006) 1259–1268
27. Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., Vert, J.P.: Classification of microarray data using gene networks. BMC Bioinformatics **8** (2007) 35

28. Pok, G., Liu, J.C.S., Ryu, K.H.: Effective feature selection framework for cluster analysis of microarray data. Bioinformation **4**(8) (February 2010) 385–389
29. Prlić, A., Procter, J.B.: Ten Simple Rules for the Open Development of Scientific Software. PLOS Comput Biol **8**(12) (December 2012) e1002802
30. Carey, V.J., Stodden, V.: Reproducible Research Concepts and Tools for Cancer Bioinformatics. In Ochs, M.F., Casagrande, J.T., Davuluri, R.V., eds.: Biomedical Informatics for Cancer Research. Springer US (2010) 149–175
31. Diaz, D., Draghici, S.: mirIntegrator: Integrating miRNAs into signaling pathways. (2015) R package.