Patient Centric Data Integration for Improved Diagnosis and Risk Prediction

Hanie Samimi, Jelena Tešić, and Anne Hee Hiong Ngu

Department of Computer Science
Texas State University
San Marcos, TX 78666

Abstract. A typical biological study includes analysis of heterogeneous biological databases, e.g., genomics, proteomics, metabolomics, and microarray gene expression. These datasets correlate at the patient-level, e.g., decrease in the workload of a group of genes in body cells increases the work of other group and raises the number of their products. Joint analysis of correlated patient-level data sources improves the final diagnosis. State-of-art biological methods, such as differential expression analysis, do not support heterogeneous data source integration and analysis. Recently, scientists in different computational fields have made significant improvements in classical algorithms for data integration to enable the investigation of different data types at the same level. Applying these methods to biological data gives more insight into associating diseases with heterogeneous groups of patients. In this paper, we improve upon our previous study and propose the use of a combination of a data reduction technique and similarity network analysis (SNF) as a scalable mechanism for integrating new biological data types. We demonstrated our approach by analyzing the risk factors of Acute Myeloid Leukemia (AML) patients when multiple data sources are presented and uncovering new correlations between patients and patient survival time.

1 Introduction

Computational biology has made significant advances in the last decade as more heterogeneous data is collected and analyzed, and new genes associated with the diseases are uncovered. The current state of the art for biological studies tends to focus on one dataset and or to study one specific correlation between gene and disease. The ultimate goal in computational biology is to be able to include and analyze all relevant biological datasets within a given biological study.

In the past decade, scientists were able to gain a detailed view of cellular processes and to advance the modeling of supramolecular assemblies. These advances led to insights into genome regulation processes and motivated the collection and analysis of genomics, proteomics, metabolomics, and microarray gene expression data [21]. All these datasets are related to the patient level, e.g., a decrease in the workload of a group of genes in body cells increases the work of other group and raises the number of their products [8]. Due to the complex

nature of biological systems, any model trained on a single dataset can only offer a one-dimensional projection view of the complex relationship between genomic, clinical, and diagnosis data. No single biological data type can capture the complexity of all the factors relevant to understanding a phenomenon such as a disease or patient-level biological processes [27].

Our work expands on the recently proposed integration method for biological data to analyze DNA methylation and gene expression data [19]. Genes are clustered into correlated groups using hierarchical clustering, and these gene groups are found to have their expression values or methylation levels highly correlate with the survival time of patients. Authors used Acute Myeloid Leukemia (AML) gene expression, DNA methylation, and survival time patients dataset. The method performed superbly in terms of predicting the risk level of patients that had an unknown risk level based on their clinical information. The average accuracy of the method in terms of newly identified patients is greater than 90%, as evaluated on different datasets. However, the proposed method only managed to re-classify a small percentage of the population with unknown risk level based on the ground truth provided by the clinical information. The number of patients whose predicted risk level is undetermined remains high, so the overall effectiveness of the proposed method is hard to evaluate for the broader patient population as the recall is low. The proposed approach does not scale to other datasets such as the messenger RNA (mRNA) and micro RNA (miRNA) datasets provided in the same study due to the expensive prohibitory computation required for integrating multiple datasets.

In patient-centric similarity networks, patients are clustered or classified based on their similarities in genomic and clinical profiles. This precision medicine paradigm has shown to have a high intractability and accuracy, and it helps advance the diagnosis and treatment[17]. The main downside of this state of the art method is their scalability for contemporary genetic cohorts, and the ability to incorporate a wide range of genomics and clinical data. In this paper, we propose to expand and scale multi-source data analysis at patient-level by (a) parallelizing single source dataset processing step that allows for scaling and multi-source integration; (b) adapting graph-based approach for data integration at patient level; and (c)introducing new exploratory result analysis methods that correlate genomics and clinical data. The proposed data analysis pipeline (see Figure 1) is demonstrated with AML cancer patient dataset. The rest of the paper is organized as follows: related work in Section 2, our proposed approach is described in Section 3, Results, and Discussion are in Section 4, and directions of ongoing work are presented in Section 5.

2 Related Work

No single biological data type can capture the complexity of all the factors relevant to understanding a phenomenon such as a disease. Integration aims to harness heterogeneous data across several dimensions of biological variation without losing important information. The main challenge in biological data

integration is how to optimally combine and interpret data from multiple sources, which are heterogeneous, high dimensions, noisy, dynamics, bias, and incomplete. In general, biological data integration approaches can be grouped based on when in the processing pipeline, the data is integrated, early, intermediate, or late integration [22].

Early data integration approach concatenates individual dataset points prior to analysis. Each data set is individually processed and normalized, and data points are concatenated into a broader feature set for the data analysis step. Multi-array Analysis approach [10] is an example of early concatenation, where sample's CEL files are normalized to different Affymetrix platforms. Early integration proved to be useful for homogeneous data types, even as it increases the dimensionality of the data. Concatenation assumes pseudo-orthogonality of the space, and lots of contextual information is lost using this simple merging technique. The approach does not support data integration for biological data sets of a different abstraction. *Late* data integration focuses on individual data set analysis, as if no other related data sets are available. Differential expression analysis for data sets of different levels is an example of late data integration [18] where findings are combined and merged at the end. Late data integration does not take advantage of the underlying correlation and or complementary information datasets provide. *Intermediate* data integration supports the pre-processing step for each data set first and focuses on finding the correlation between the object of the study and all data sets at the same level before data analysis. Multi-view analysis, dimensionality reduction, and graph integration are examples of intermediate data integration [22]. Each of the proposed methods is unique to a fixed set of data types, and the main challenge for intermediate data integration is its generalizability to different types of biological studies, and its extension to different types of data sets within the same study [19].

Biological data is inherently heterogeneous and noisy, as the signal is collected using multiple sampling techniques on a very small subspace of the population for a particular phenotype. The high-dimensionality and sparsity of the datasets make it extremely challenging to make meaningful insights, even when the number of subjects is high. The genetic correlations can be viewed as inconclusive due to population stratification [11, 26]. There has been a large number of omics data integration approaches, including biochemical pathway-, ontology-, network-, and empirical-correlation-based methods [25]. They all suffer from the same downsides, as an extension of the analysis is hindered by the system's inability to extend to multiple data types and lack of generalization due to noisy nature of collected samples, and inherent high dimensionality curse in data analysis [1, 4].

Lately, the biological data analysis turned to machine learning tools as a way to address these pressing issues: sophisticated dimensionality reduction methods and ensemble classification methods show some improvements over baseline integration techniques but do not generalize well [27]. Big data approaches from social network analysis and computer vision show more promise: incoherent, missing, noisy, small sample high-dimensional data has been extensively stud-

ied in those areas [15]. Network-based integration of multi-omics data approach showed promising results in correlating genes and diagnosis [17]. Graph diffusion-based method for prioritizing cancer genes by integrating diverse molecular data types on a directed functional interaction network is proposed in [6]. The network prioritizes genes by their mediator effect individually and integrates them using the rank aggregation approach.

Similarity Network Fusion (SNF) that integrates mRNA expression data, DNA methylation data, and microRNA expression data for five cancer data sets was proposed in [24]. Integration of these similarity networks follows the construction of a similarity network for patients for each single data type into a single fused similarity network using a nonlinear combination method that iteratively updates each network making it more similar to each other. Machine learning approaches for genomics and clinical data integration bridge the gap between big data processing and interpreting the data [27]. In this paper, we focus on how to scale the integrative data analysis and correlate the genomic and clinical datasets.

3 Approach

Patient-similarity networks offer ease of interpretability and ease of discovery of the underlying correlations between patients genomics and clinical data in precision medicine. Here, we propose an approach to scale the network-based algorithm in terms of various heterogeneous datasets and processing. We demonstrate the scalability and improved analysis capability of the proposed approach when applied to AML genome and clinical patient-level analysis. The processing pipeline included the data preparation, dimensionality reduction, similarity network fusion, and unsupervised/supervised data classification for interpretability analysis, as illustrated in Figure 1 (left). To better comprehend the proposed approach, we compare it to previous work pipeline Figure 1 (right) as outlined in [19].

3.1 Data Preparation and Dimensionality Reduction

Setup: The Genome Cancer Atlas (TCGA) data portal provides different types of genomic and clinical data related to a type of cancer in the form of a project [9], and this is the dataset of choice for this work. All scripts for the experiments were written in R and running on a machine with Ubuntu 16.04 operating system, 32 Gigabytes of RAM and 500 Gigabytes of the hard disk.

Data Preparation: Some of the gene expression or DNA methylation values are not available for more than 50% of patients analyzed. We normalize the dataset to accommodate missing values. DNA methylation data is cleaned and normalized using RnBeads R packages [2]. Next, we omit the DNA methylation level of those loci that had low Spearman correlation with the survival time of death

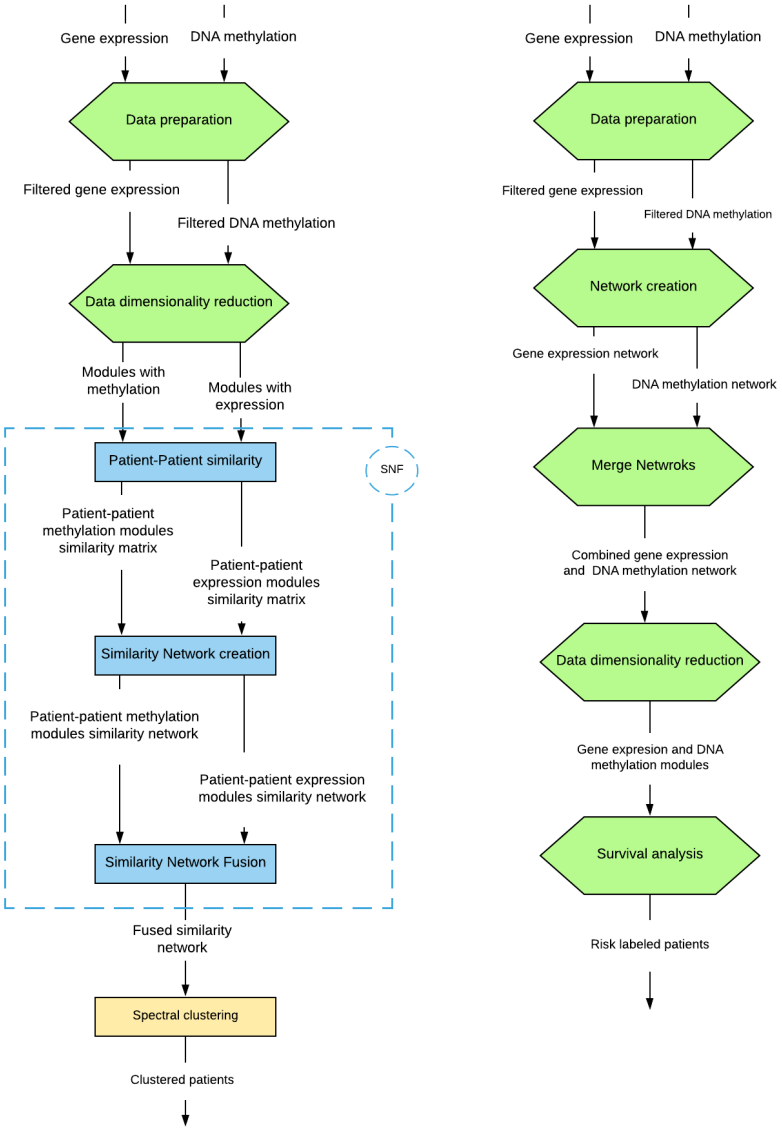

Fig. 1: Flowchart of proposed methods. The left one is related to the newly proposed method while the right one shows the steps that have been taken in the previously proposed method.

patients. After that, we group remaining genes based on their similar biological features and behavior. For this step, we analyze datasets separately, and construct two different Pearson similarity matrices, one for gene expression and

one for DNA methylation data, and construct one network per each of the data types. The nodes in both networks are genes, and edges are the Pearson correlation between every two genes.

Data Dimensionality Reduction: Clustering genes using these two networks results in groups of genes which are highly correlated with each other based on either their methylation levels or gene expression values. We have used the *Blockwisemodule* function from the WGCNA R package. This function performs hierarchical clustering using average linkage. The output from the hierarchical clustering is correlated clusters of genes which are called gene modules. For each of the modules, we computed a weighted average value using Principal Component Analysis (PCA). PCA is a data reduction method that summarizes the information of a high dimensional dataset into a few vectors in the direction with the highest variance [13]. The first principal component is the direction that the projected data has the most significant variance. We used the coefficients of the first principal components to compute a weighted average for the gene expression and DNA methylation modules separately. As an output of the data preparation step, each of the similar gene groups is represented by a single feature vector per dataset level (gene expression values or DNA methylation levels). Note that this approach scales to different types of input dataset, as each dataset is prepared independently using the proposed method. The output of data preparation and dimensionality reduction is two gene modules: module with DNA methylation and module with gene expression.

3.2 Patient-centric Network Fusion and Analysis

To integrate two gene modules produced in Section 3.1, we apply network fusion paradigm at the patient level, that consists of four generic steps:

1. *C*ompute The similarity between patients for each data type: similarity metric is constructed for each datatype using pairwise correlation of gene module, as described in Section 3.1 between two patients.
2. *C*onstruct A patient network per data type: a graph where patients are nodes and edges represent patients pairwise similarities calculated for that dataset.
3. *M*erge Two constructed patient similarity networks using Similarity Network Fusion approach [24]. Nodes of both networks are the same (patients), and edges' weights are iteratively updated to converge to a joint network.
4. Produce A final single patient-centric fused network where nodes are the patients and edges define the similarity value between the patients based on integrated data types.

*C*ompute and *C*onstruct steps in the approach are per dataset. For multiple datasets, patient-centric network constructions can be executed concurrently, scaling data preparation and dimensionality reduction phase of the pipeline. *M*erge and *P*roduce steps can use any network fusion and metric normalization

method. Network-based fusion approaches can capture local and general similarity of multiple datasets provided in a study [6]. As the proof-of-concept, we employ Similarity Network Fusion analysis from [24].

Spectral Clustering step is a data analysis approach of the patient-centric fused network output [23]. Most stable clustering is given by the value of K that maximizes the eigen-gap (difference between consecutive eigenvalues), and we use this method to find the optimal high value of K for spectral clustering, as proposed in [14]. We have experimented with clustering results and visualize them for data analysis purpose, as shown in Section 4.

In our previous work, network was created from two datasets using early integration and it does not scale to new data types, see *Network Creation* module on the left in Figure 1(right), [19]. This proposed work proceeds with data preparation and data dimensionality reduction for each data type separately, see *Data dimensionality reduction* in Figure 1. Previous work merged the network based on the linear combination of similarity scores, and then applied dimensionality reduction approach to obtain relevant gene modules per patient for survival analysis, see *Merge Networks*, *Data dimensionality reduction*, and *Survival analysis* modules in Figure 1(right), [19]. Proposed approach is more generic as it supports similarity network fusion of any number of data type modules, see SNF group in Figure 1(left).

4 Results for the AML Study

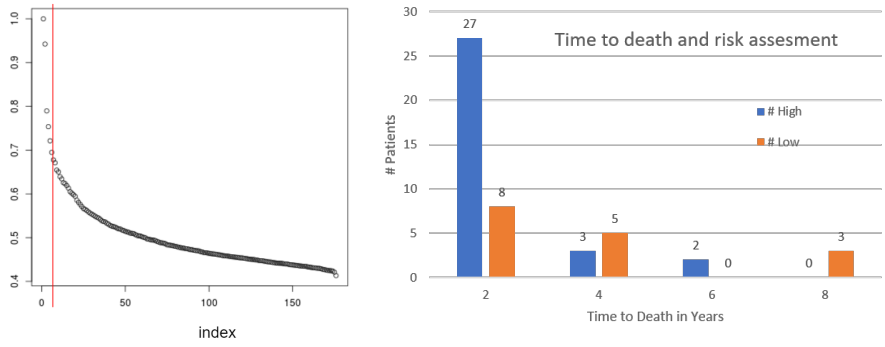

Fig. 2: AML Study Data Analysis: Eigenvalues of fused similarity matrix (left) and correlation of risk assessment and time of survival (right)

Acute myeloid leukemia (AML) is a type of blood and bone marrow cancer in which the bone marrow makes abnormal myeloblasts (a type of white blood cell), red blood cells, or platelets. This type of cancer usually gets worse quickly

if it is not treated, and accounts for 1.2% of cancer deaths in the United States [12]. Acute myeloid leukemia is a category of diseases with a common aggressive clinical presentation but with prognosis and management that is dependent upon the underlying genetic characteristics of the neoplasm. AML treatments have consistently improved, yet the treatment remains unchanged for the past three decades with the majority of patients eventually relapsing and dying of the disease [20]. Cytogenetics and mutation testing remain a critical prognostic tool for post-induction treatment as it identifies a subgroup of AML patients characterized by poor response to induction chemotherapy and poor long-term survival after treatment with consolidation chemotherapy. There are effective high-precision predictors for high risk and low risk patients that aid the course of treatment for AML [20]. Current methods for labeling risk level in AML patients leave out most of the diagnosed population, as most of the people with AML diagnosis do not fall into high or low risk group [19]. Treatment for such patients is unknown and risky. Here, we focus on the patient-level analysis of the people with AML diagnosis to improve the patient's risk level classification.

Experiments are run using the TCGA-LAML project's data that has several biological data types of 200 AML patients [9]. Each type of data in TCGA-LAML project is collected using several types of platforms. The clinical data of patients contain several different information such as their vital status, gender, days to death, days to last follow-up, race, etc. We used days to the death of dead patients and days to the last follow-up of alive patients as their survival values. The risk levels assessment of patient are based on their mutation and cytogenetic abnormalities and broadly grouped as low, medium, and high.

In the data preparation step, we have used Gene Expression Quantification values of the HTSeq-FPKM workflow. Each gene is mapped from at most 4 loci with a probability of 95% resulting in 19911 genes. Loci that had low correlation (¡0.1) with the survival time of the dead AML patients are excluded, bringing down the number of genes to 6637. Next, applying the Dimensionality reduction step resulted in 39 groups of genes. The second dataset was DNA methylation Beta values collected using Illumina Human Methylation 450 platform. Dimensionality reduction grouped DNA methylation data into 37 groups of genes. Two patients-patients similarity networks for 176 patients were created using the average value of the gene expression modules and DNA methylation modules. For unknown values in networks, we used the K Nearest Neighbour (KNN)[5] method for 10 iterations. Similarity Network Fusion approach was used for merging these two networks. The result is a *s*ingle fused network where the nodes are patients, and the edges are the combined similarity value based on both gene expression and DNA methylation values.

Spectral clustering step for data analysis results is shown in Figure 2. As shown in the left of Figure 2, values of eigenvalues accelerate up to $K = 7$. Thus, we limit our experimentation on optimal K values with the limit of 7 to evaluate the difference in the performance. Using the fused similarity network, we classified the patients into sub-classes $K \in [2, 7]$ using the spectral clustering function from the SNF R package [3]. The results are illustrated in Figure 3.

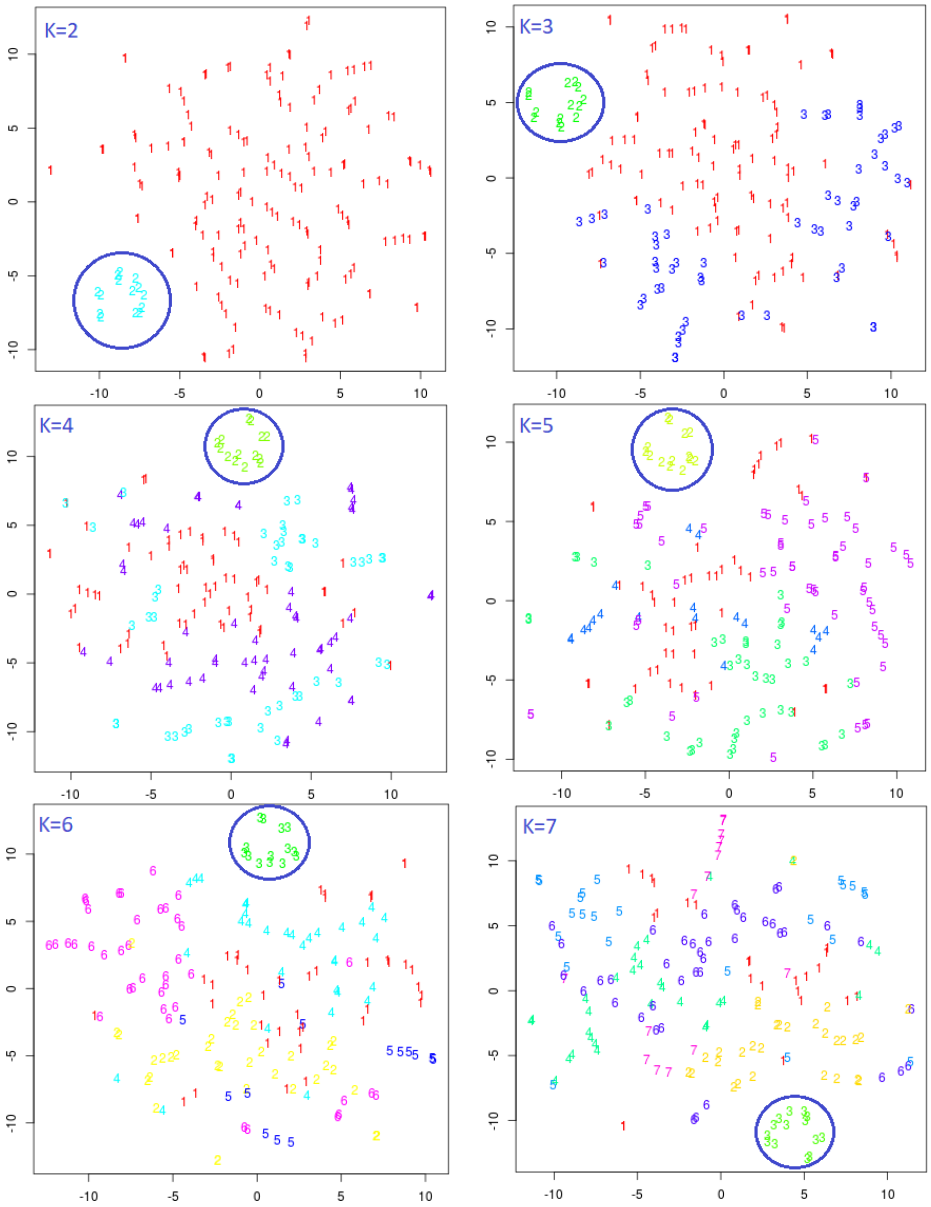

Fig. 3: Clustering visualization for various number K of clusters.

The most interesting finding in this data analysis is that a group of 14 patients is consistently grouped regardless of the value of K, circled in Figure 3. What is interesting about this cluster is that 11 out of 14 patients in this cluster are labeled as low risk based on their mutations, and the rest of them are labeled as high risk in clinical data. When we analyzed the clinical data, this specific group of patients had a similar survival time, even though their assigned risk measure and predicted the time of survival was different.

This type of patient-centric analysis shows that risk assignment needs to take more data into account, as these patients strongly correlate according to multiple datasets. Figure 2(right) shows the study correlation of survival time (in years) with the assigned risk factor. It shows that the labeling based on cytogenetic mutations only is not reliable, and more patient-centric studies like ours need to be incorporated in assigning the risk factor [16].

5 Discussion and Conclusion

Patient-centric analysis has emerged as the most effective paradigm to offer an interpretable diagnosis for precision medicine [17]. This strategy can be viewed as a standard medical diagnosis for big data era, has excellent performance, is interpretable, and can preserve patient privacy. In this paper, we have proposed a path forward to scalable patient-centric genomic and clinic data analysis, as illustrated in Figure 1. We have demonstrated the usability of our approach, and new correlation discovery using patient-centric datasets from TCGA-LAML project [9]. Our current work includes the testing of the proposed pipeline by adding additional data types such as mRNA and miRNA as well as investigating the use of a recently published labeling method, ELN 2017 to examine how much the classification of risks in cancer patients can be further improved.

Our goal is to advance patient-centric data analysis by exploiting links between genomic and clinical data. Clustering, as a data analysis tool, suffers from sensitivity to noisy examples, and classification techniques such as Support Vector Machine can do a better job in capturing complex data correlations. Our next step is to incorporate network fused distance matrix as a distance measure for classification method and assess its practical implications [7].

References

1. A. Alyass, M. Turcotte, and D. Meyre. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics*, 8(1):33, Jun 2015.
2. Y. Assenov, F. Müller, P. Lutsik, J. Walter, T. Lengauer, and C. Bock. Comprehensive analysis of dna methylation data with rnbeads. *Nature methods*, 11(11):1138, 2014.
3. F. D. M. F. Z. T. M. B. B. H.-K. A. G. Bo Wang, Aziz Mezlini. Snftool: Similarity network fusion. https://CRAN.R-project.org/package=SNFtool. Published:2018-04-24.

4. G. G. T. B. K. A. M. G. A. C. C. Chen Meng, Oana A. Zeleznik. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4):6280–641, 2016.
5. P. Cunningham and S. J. Delany. k-nearest neighbour classifiers. *Multiple Classifier Systems*, 34(8):1–17, 2007.
6. C. Dimitrakopoulos, S. K. Hindupur, L. Hfliger, J. Behr, H. Montazeri, M. N. Hall, and N. Beerenwinkel. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics*, 34:2441–2448, 2018.
7. B. Haasdonk and C. Bahlmann. Learning with distance substitution kernels. In *Pattern Recognition*, pages 220–227, Berlin, Heidelberg, 2004.
8. V. A. Huynh-Thu and G. Sanguinetti. *Gene Regulatory Network Inference: An Introductory Survey*, pages 1–23. Springer New York, 2019.
9. N. C. Institute. Tcga-laml. https://portal.gdc.cancer.gov/projects/TCGA-LAML. Accessed: 2019-05-30.
10. R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of affymetrix genechip probe level data. *Nucleic acids research*, 31(4):e15–e15, 2003.
11. P. R. Jansen, K. Watanabe, S. Stringer, N. Skene, J. Bryois, A. R. Hammerschlag, C. A. de Leeuw, J. S. Benjamins, A. B. Muoz-Manchado, M. Nagel, J. E. Savage, H. Tiemeier, T. White, T. 23andMe Research Team, J. Y. Tung, D. A. Hinds, V. Vacic, X. Wang, P. F. Sullivan, S. van der Sluis, T. J. C. Polderman, A. B. Smit, J. Hjerling-Leffler, E. J. W. V. Someren, and D. Posthuma. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nature Genetics*, 51:394–403, 02 2019.
12. A. Jemal, A. Thomas, T. Murray, M. Thun, et al. Cancer statistics, 2002. *Ca-A Cancer Journal for Clinicians*, 52(1):23–47, 2002.
13. I. Jolliffe. *Principal component analysis*. Springer, 2011.
14. T. M. Kodinariya and P. R. Makwana. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.
15. V. Marx. Machine learning, practically speaking. *Nature Methods*, 16:463–467, 2019.
16. M. Moarii and E. Papaemmanuil. Classification and risk assessment in aml: integrating cytogenetics and molecular profiling. *Hematology Am Soc Hematol Educ Program*, Dec 2017.
17. S. Pai and G. D. Bader. Patient similarity networks for precision medicine. *Journal of Molecular Biology*, 430(18, Part A):2924 – 2938, 2018. Theory and Application of Network Biology Toward Precision Medicine.
18. M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.
19. H. Samimi. Identification of gene sets that predict acute myeloid leukemia prognosis using integrative gene network analysis. Master's thesis, Texas State University, 8 2018. txi:b4789711.
20. J. N. Saultz and R. Garzon. Acute myeloid leukemia: A concise review. *Journal of Clinical Medicine*, March 2016.
21. E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan. Computational solutions to large-scale data management and analysis. *Nature reviews genetics*, 11(9):647, 2010.
22. A. Serra, M. Fratello, D. Greco, and R. Tagliaferri. Data integration in genomics and systems biology. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 1272–1279. IEEE, 2016.

23. U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.
24. B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333, 2014.
25. K. Wanichthanarak, J. F. Fahrmann, and D. Grapov. Genomic, proteomic, and metabolomic data integration strategies. *Biomarker Insights*, 10s4, 2015.
26. D. T. N. E. J. Y. T. Youna Hu, Alena Shmygelska and D. A. Hinds. Gwas of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. *Nature Communications*, 7, 2016.
27. M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91, 2019.