The 2020 12th International Conference on Knowledge and Systems Engineering (KSE)

Disease subtyping using community detection from consensus networks

Hung Nguyen
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
hungnp@nevada.unr.edu

Bang Tran
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
bang.t.s@nevada.unr.edu

Duc Tran
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
duct@nevada.unr.edu

Quang-Huy Nguyen
Department of Computational Biomedicine
Vingroup Big Data Institute
Hanoi, Vietnam
huynguyen96.dnu@gmail.com

Duc-Hau Le
Department of Computational Biomedicine
Vingroup Big Data Institute
Hanoi, Vietnam
hauldhut@gmail.com

Tin Nguyen*
Computer Science & Engineering
University of Nevada, Reno
Reno, USA
tinn@unr.edu

***Abstract*—Cancer is a complex disease including a range of disorders that are activated simultaneously by multiple biological processes on multiple levels. Various genome-wide profiling techniques have been developed to capture the dynamics of these processes at the epigenomic, transcriptomic, and proteomic levels. Integrative analysis of data from these sources has the potential to differentiate cancer subtypes from a holistic perspective that reveals connections that otherwise cannot be detected using observations from a single data type. In this article, we present a novel approach named DSCC (Disease Subtyping using Community detection from Consensus networks) that is able to discover disease subtypes from multi-omics data and is robust against noise. In an extensive analysis using simulation studies and 5,782 real patients belonging to 20 cancer datasets from The Cancer Genome Atlas, we demonstrate that DSCC outperforms state-of-the-art methods by correctly identifying known patient groups and novel subtypes with significantly different survival profiles.**

***Index Terms*—multi-omics integration, cancer subtyping, survival analysis, community detection**

I. Introduction

Despite advances in cancer prognosis and treatment, the probability of a person being diagnosed with prostate or breast cancer has increased twice after 20 years of cancer screening [1], [2]. A large number of patients still fail therapy, resulting in disease progression, recurrence, and overall survival reduction [3]. Concurrently, 30-50% of patients with non-small cell lung cancer (NSCLC) quickly advance to recurrence and die after curative resection [4], suggesting that a particular subgroup of patients should have received more rigorous treatments at initial stages. Moreover, adjuvant and neoadjuvant chemotherapy have proved to be an efficient method for significantly survival improvement of patients with advanced early-stage cancer [5]. These discoveries suggest that a better prognosis method would allow us to manage these diseases better: patients whose cancer is likely to advance rapidly would need more aggressive treatment. However, cancer is widely understood to be a heterogeneous disease. A tumor is a complex ecosystem containing tumor cells, as well as various infiltrating endothelial, hematopoietic, stromal, and other cell types that can influence the function of the tumor as a whole [3]. Due to the diversity of mutations and molecular mechanisms, individual tumor's behavior and response to treatment vary greatly [6]. Therefore, it is important to identify cancer subtypes based on common molecular features and and subgroups of patients [7]–[10]. This can also benefit a wide range of studies related to molecular data from aging, obesity to drug response [11]–[14].

Advanced genome sequencing has demonstrated that cancer within a single patient is a heterogeneous mixture of genetically distinct sub-clones that arise through evolution [15]–[17]. Therefore, recent subtyping methods have shifted toward multi-omics data integration in order to differentiate between subtypes from a holistic perspective that takes into consideration phenomena at different molecular levels (DNA methylation, chromatin openness, microRNA, and other non-coding RNA). These methods can be classified into three main categories: simultaneous data decomposition methods, joint statistical models, and similarity-based approaches. Method in the first category, such as md-modules [18], intNMF [19], and LRAcluster [20], focus on finding a common pattern that exists across multi data types in lower-dimensional representation. However, the subtyping outcomes heavily rely on the assumption that all molecular signals can be linearly and simultaneously reconstructed.

Methods in the second category use statistical approaches in which each data type follows a mixture of distributions and the integration of multiple data types is constructed using a joint statistical model. Methods in this category include BCC [21], MDI [22], iClusterBayes [23], iClusterPlus [24], and iCluster [25], [26]. The main drawback of those methods is that the resulted subtypes strongly depend on the correctness of statistical assumptions on the data. Moreover, the statistical-based methods often require inputs for many parameters and produce results after a long computational time.

978-1-7281-4510-5/20/$31.00 ©2020 IEEE

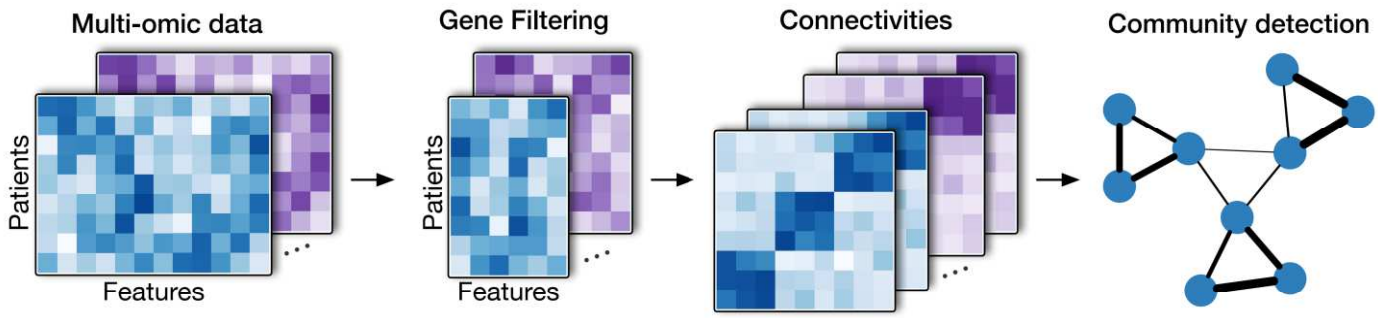

Fig. 1. The overall workflow of DSCC. The method consists of three main steps: i) gene filtering using non-negative matrix factorization, ii) building patients connectivities using k-means with different numbers of clusters, and iii) clustering using community detection.

Methods in the last category follow a similarity-based approach in which patient connectivity for each data is represented in the form of a graph with patients as node and connectivity as edges (indicates how frequently the patients are grouped). A similarity matrix is generated by merging the connectivity from all data types, and a similarity-based algorithm is used to identify subtypes. Methods in this category include SNF [27], NEMO [28], PINS [7], [29], [30], CIMLR [6], and SCFA [31]. SNF integrates multi-omics data sets using a network fusion method by creating a network for each data type and then fuses them into a single similarity network. NEMO computes inter-patient similarity matrices for each data type through a radial basis function kernel and uses spectral clustering to cluster the combined similarity matrix. PINS identifies how often the patients are grouped together when the data are perturbed and clusters strongly connected patients across all data types together. On the other hand, CIMLR combines multiple gaussian kernels (one per data type) to measure the similarity between each pair of patients and uses k-means to subtype the final similarity matrix. SCFA uses an autoencoder to eliminate unimportant genes and factor analysis to project large data in lower dimensions. Next, it uses k-mean clustering to determine cluster assignments from each lower dimension presentation and uses an ensemble meta-clustering algorithm to generate final clusters. Methods in this category are usually computational efficient and can easily support different omic types. However, it may be difficult to interpret the results in term of how original features contributes to the discovered subtypes.

Here we introduce DSCC (Disease Subtyping using Community detection from Consensus networks) that exploits the local relationships between patients from each data type to build a consensus network from patient connectivities. It then uses a community detection technique to discover different groups within patients that have significantly different survival profiles. In an extensive analysis using simulation studies and 5,782 real patients related to 20 cancer datasets from The Cancer Genome Atlas, we demonstrate that DSCC is robust against noise and outperforms state-of-the-art methods in identifying known patient classes and novel subtypes with significantly different survival profiles.

II. Methods

Figure 1 shows the overall workflow of DSCC. The method requires a list of data matrices (mRNA, methylation, miRNA, etc.). In each matrix, rows represent samples/patients, and columns represent genes/features. For each matrix, the method first applies gene filtering using non-negative matrix factorization and then builds connectivities between patients using k-means clustering. Finally, the method applies community detection on the combined connectivity using Louvain modularity [32] to cluster patients.

A. Gene filtering using Non-negative Matrix Factorization

Our hypothesis is that although the total number of features in omics data is large (e.g. ∼20k for mRNA data), only a subset of them truly differentiates among cancer subtypes. Therefore, we first focus on filtering out genes that are not likely to play a major role in subtyping. Figure 2 shows the workflow of our gene filtering approach using 1-factor Non-Negative Matrix Factorization (NNMF). Briefly, Matrix Factorization is a technique that decomposes a matrix into the product of two lower dimensionality matrices:

$$V = W \times H + E$$

where in the context of this article:

- V is a matrix of size $p \times g$ (the original omic data, e.g. gene expression matrix), in which p is the number of patients and g is the number of genes;
- W is a matrix of size $p \times k$, a representation of patients in a latent space with the number of factors is k;
- H is a matrix of size $k \times g$ representing meta-gene matrix in the latent space; and
- E is a matrix of size $p \times g$, the error between the original data and the reconstructed data from W and H.

The number of latent factors k has been used as the number of clusters in a number of clustering methods [33], [34]. If a dataset consists of k subtypes, it is expected that genes contributing to differentiating subtypes will have different expression patterns among subtypes and these patterns can be captured in each latent factor. In our method, we use NNMF to filter features that have insignificant contributions to differentiating subtypes rather than directly assign clusters

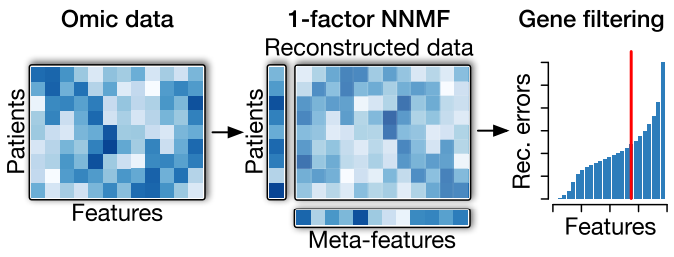

Fig. 2. Gene filtering using non-negative matrix factorization. The original data matrix is decomposed into two vectors representing patients and their features in 1-dimensional latent space. The error of the reconstructed data using these two vectors is used to rank each gene. Only 30% of genes that have the largest error are kept for the next steps.

using data from NNMF. Here, we choose the number of factors $k = 1$. This makes it difficult to fit the model for genes that have significantly different expression patterns on different subtypes. As a result, genes that have a significant contribution to differentiating subtypes will have more errors in the reconstructed data. We then rank the genes by its total absolute error $\sum |E_{.g}|$ and keep only 30% of genes that have the largest error for the next steps.

After filtering unimportant features, the number of remaining features is still on the scale of hundreds or even thousands. It is necessary to perform dimension reduction to reduce the time complexity for network construction. Therefore, we finally use principal component analysis to perform dimension deduction on each filtered data with the number of principal components is 20. This data is then used to generate connectivities between patients in the next step.

B. Consensus network generation and subtyping

To generate the overall connectivities for patients in each data type, we run k-means on the 20-dimension data with different numbers of clusters. The connectivities between patients are defined as a square matrix where both rows and columns represent patients. Its values are 1 when two patients are clustered into the same group and otherwise 0.

In this step, we aim to group a certain number of patients into the same clusters. This can be achieved by adjusting the number of clusters inputted for the k-means algorithm. For example, if the number of patients is p and the number of clusters is k, it is expected that each cluster will have an average of $\frac{p}{k}$ patients, assuming that the clustering yields balanced clusters. We choose the number of clusters k so that each cluster will have the number of members from 2 to 50. Our assumption is that if a group of patients belongs to the same subtypes then they will tend to establish connections regardless of the predefined number of clusters. Also, by using a large different numbers of clusters, we expect that both local and global connections between patients will be established.

It is known that the k-means algorithm often converge at local minima, especially with big numbers of clusters. However, the more time that samples clustered into the same groups, the more chance these samples belong to the same cluster in the final assignments. Therefore, for each number of clusters k, we run k-means $1,000$ times. The final patient connectivity matrix for each data type is the average of connectivities from all runs of all k.

Finally, we create an undirected weighted graph from the average of all patient connectivities across all data types. We apply community detection using the Louvain method to discover communities from the graph as the final clusters. The Louvain algorithm [32] optimizes a modularity quality function in two elementary phases: i) local moving of nodes, and ii) network aggregation. The modularity function measures the edges density within communities compared to those between communities and is computed as follow:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where

- A_{ij} is the edge weight between the two nodes i and j;
- $k_i = \sum_j A_{ij}$;
- $m = \frac{1}{2} \sum_{ij} A_{ij}$;
- c_i is the community that node i is assigned; and
- $\delta(u, v) = 1$ if $u = v$ and 0 otherwise.

First, each node in the graph is assigned to a community. In the nodes moving phase, each node is moved to one of its neighbor communities that yields the largest increase in the quality function. If no increase is gained from all moves, the node remains in its original community. This process repeats until no increase in the quality function occurs. In the network aggregation phase, each community in the first phase becomes a node to form an aggregate network. The two phases are repeated until the modularity quality function converges. The final detected communities for the network are the output clusters for all data types.

III. Results

In this section, we assess the performance of the proposed method using i) simulation studies and ii) 20 real datasets from TCGA. We compare DSCC with other four state-of-the-art methods in cancer subtyping including Consensus Clustering (CC), Similarity Network Fusion (SNF), iClusterBayes (iCB), and Cancer Integration via Multikernel LeaRning (CIMLR). Among the four methods, CC is the only method that does not inherently support multiple data type integration. Therefore, in each analysis, we concatenate all data types for the integrative analysis.

We note that our method is completely unsupervised learning, in which besides input is multi-omics data, no additional knowledge is provided for our clustering method. To make it fair with all other methods, we let each method detect the true number of clusters from the input data and use that number to generate the final cluster assignments.

A. Simulation study

To generate data for the simulation study, we use three different models to simulate different types of omics data

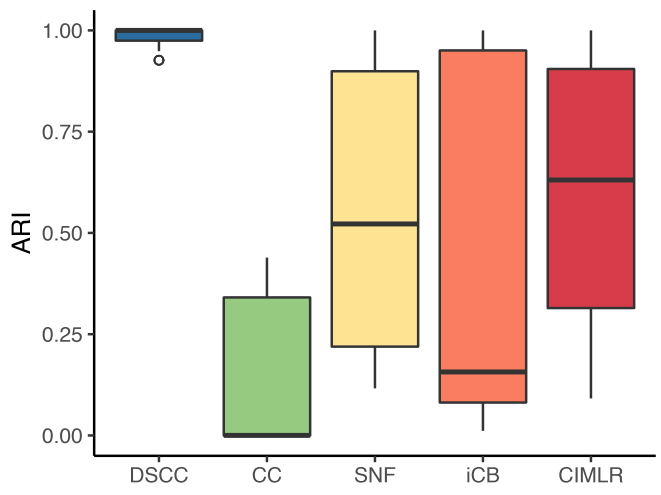

Fig. 3. ARI values of clusters produced by DSCC, Consensus Clustering (CC), Similarity Network Fusion (SNF), iClusterBayes (iCB) and Cancer Integration via Multikernel LeaRning (CIMLR) using 20 simulated data.

including Gaussian, Beta-like, and Binary model. These simulation models are inspired by Pierre-Jean et. al. [35].

We simulate three different data types using the three models. Each data type has 100 samples, 10,000 features, and is splitted into five groups in which each group has 50 differential features to distinguish between clusters. The parameters for each model is as follow: for Gaussian model, $\mu = 2$ and $\sigma = 1$; for Beta-like model, $\mu_1 = -2$, $\sigma_1 = 0.5$, $\mu_2 = -2$, and $\sigma_2 = 0.5$; and for Binary model, we set the probabilty $p = 0.6$ for a value to be 1.

We also simulated noise in each data type, in which we define the based-noise for each model as follow: for Beta-like model, we add noise using normal distribution with $\mu = 0$ and $\sigma = 1$; for Beta-like model, we also add noise using normal distribution with $\mu = 0$ and $\sigma = 0.1$; and for Binary model, we add randomly add 1 value to the data with probability $p = 0.1$. With this based level of noise, clusters in all data types are well separated. We finally simulate in total 20 datasets in which each dataset consists of three data types from the three distributions with different noise levels. The noise level is adjusted by increasing σ and p in the noise added to the data from 10% to 200%.

Since the true cluster assignments are known, we use Adjusted Rand Index (ARI) [36] to assess the performance of the methods. Briefly, ARI measures the similarity between two cluster assignments with correction for chance. ARI values range from -1 to 1 where $ARI = 1$ indicates a perfect match between two cluster assignments, $ARI = 0$ indicates the agreements are expected to be the same with random cluster assignments, and negative ARI indicates that the agreement is less than what is expected from a random result.

Figure 3 shows the distribution of ARIs for the 20 simulated datasets for the five methods. CC produces clusters with the lowest ARI values since this method fails to detect the true number most of the time. SNF, iCB, and CIMLR can reach ARI values of 1 when the level of noise $< 30\%$. However, when the noise level increases more, their performance drastically decreases. While iCB can still detect the true number of clusters when the noise level increase, SNF and CIMLR fail to do so when the noise level is $> 100\%$. On the other hand, it is clear that DSCC can easily maintain the ARI values close to 1 in all datasets. The performance of DSCC is slightly affected only when the noise level is $> 150\%$.

B. Performance on TCGA data

To better assess the performance of DSCC, we compare DSCC and CC, SNF, iCB, and CIMLR on 20 TCGA datasets. Since the true subtypes are not available for any of the datasets, we use Cox Proportional-Hazards Model [37] to validate the subtypes produced by each method. The p-values from this regression model represent the association between the survival time of patients with the subtype they are assigned. Table I show the Cox p-values of subtypes produced by the five methods on the 20 datasets.

TABLE I
COX P-VALUES OF SUBTYPES IDENTIFIED BY DSCC, CC, SNF, ICLUSTERBAYES (ICB), AND CIMLR FOR 20 TCGA DATASETS. CELLS IN YELLOW INDICATE SIGNIFICANT P-VALUES (< 0.05). CELLS IN GREEN INDICATE THE MOST SIGNIFICANT P-VALUE FOR EACH DATASET.

#	Dataset	DSCC	CC	SNF	iCB	CIMLR
1	ACC	6.0e-05	8.7e-04	4.3e-05	5.4e-03	1.3e-01
2	BLCA	7.2e-05	1.1e-01	1.1e-01	2.1e-01	4.4e-01
3	BRCA	1.7e-03	1.0e-02	1.2e-01	2.7e-02	5.2e-03
4	CESC	1.6e-02	2.2e-01	5.1e-01	2.0e-02	1.9e-01
5	CHOL	5.9e-01	7.9e-02	5.7e-01	7.0e-01	3.4e-01
6	COAD	2.6e-01	5.5e-01	1.3e-01	4.2e-01	2.6e-01
7	COADREAD	6.6e-01	7.2e-01	6.6e-01	8.0e-01	3.3e-01
8	DLBC	8.8e-01	5.1e-01	7.5e-01	1.9e-01	7.4e-01
9	ESCA	3.1e-01	8.1e-01	3.9e-01	1.9e-01	5.6e-01
10	GBM	5.0e-03	7.5e-01	2.1e-02	2.6e-01	5.4e-02
11	GBMLGG	2.6e-16	4.9e-04	4.8e-14	8.0e-02	3.7e-10
12	HNSC	1.5e-03	5.1e-01	3.7e-01	7.8e-02	4.0e-01
13	KICH	5.1e-01	9.3e-01	7.0e-01	1.4e-01	4.6e-01
14	KIPAN	6.3e-19	5.3e-08	2.1e-07	1.4e-01	9.8e-05
15	KIRC	1.7e-03	8.3e-01	6.9e-01	2.1e-01	2.9e-01
16	KIRP	7.0e-03	2.2e-02	5.3e-03	4.9e-02	1.9e-02
17	LAML	3.6e-04	2.0e-01	1.7e-03	8.7e-03	8.7e-01
18	LGG	2.4e-19	1.3e-06	1.6e-14	2.3e-05	7.1e-15
19	LIHC	3.2e-04	8.2e-01	3.3e-01	2.0e-01	1.3e-01
20	LUAD	7.5e-03	7.6e-01	5.0e-01	2.2e-02	3.7e-01
	#Sigificant	14	6	7	7	5

Among 20 datasets, there are 7 datasets (CHOL, COAD, COADREAD, DLBC, ESCA, KIRC and LIHC) for which none of the five methods is able to discover subtypes with significant survival differences. In the remaining 14 datasets, DSCC identifies subtypes with significantly different survival profiles on all 14 datasets. That number for CC, SNF, iCB, and CIMLR is 5, 7, 5, and 5 respectively. Moreover, DSCC has the most significant p-values for 12 out of 14 datasets.

To further investigate the effect of data integration on the clustering results using DSCC, we also perform subtyping analysis for each data type and gather p-values from the produced clusters. Figure 4 shows the distribution of $-$log10(p-value) by each data type and also integrated data (mRNA, Methylation, and miRNA combine together). Among

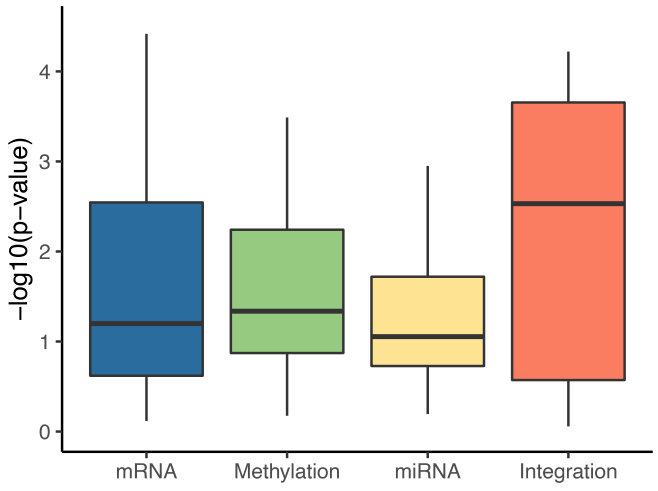

Fig. 4. Cox p-values of subtypes identified by DSCC on each single data type and on integrated data. Overall, data integration yield better results compared to those of single data type only.

20 datasets, the Cox p-values obtained from integrated data has the median $-\log10$(p-value) of 2.5, compared to 1.2, 1.3, and 1.0 from mRNA, Methylation and miRNA respectively. With a significant threshold of p-value $= 0.05$ or $-\log10$(p-value) $= 1.3$, subtyping using integrated data shows that it can identify subtypes with significant differences in survival profiles while subtyping using single data type fail to do so.

IV. Conclusion

In this article, we developed a novel method, DSCC, for disease subtyping and data integration. DSCC is robust against noise and can efficiently identify cancer subtypes with significantly different survival profiles. We validated our method using 20 simulated datasets and 20 real datasets from TCGA with a total of $5,782$ patients. Our simulation study shows that DSCC can work well with data that have different distributions. It can precisely detect the true number of clusters and is robust against noise. Our evaluation on real data shows that DSCC is able to discover subtypes with significantly different survival profiles while many other state-of-the-art fail to do so. It also shows that subtyping using data integration produces better subtypes compared to subtyping using only a single data type. The developed method is flexible and can be applied in a wide range of applications. For future work, we will combine DSCC with other methods developed in the context of biological networks [38]–[46], single-cell [47]–[49], genomics and epigenomics [50]–[58], and drug development [59].

V. Sofware and Data availability

Datasets from The Cancer Genome Atlas were downloaded from http://firebrowse.org/. All source code for analyses in this manuscript is available upon request.

VI. Acknowledgements

This work was partially supported by NASA under grant number 80NSSC19M0170, by NIH NIGMS under grant number GM103440, and by NSF under grant number 2001385. Any opinions, findings, and 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] L. J. Esserman, I. M. Thompson, B. Reid, P. Nelson, D. F. Ransohoff, H. G. Welch, S. Hwang, D. A. Berry, K. W. Kinzler, W. C. Black, M. Bissell, H. Parnes, and S. Srivastava, "Addressing overdiagnosis and overtreatment in cancer: a prescription for change," *The Lancet Oncology*, vol. 15, no. 6, pp. e234–e242, 2014.

[2] L. Esserman, Y. Shieh, and I. Thompson, "Rethinking screening for breast cancer and prostate cancer," *Journal of the American Medical Association*, vol. 302, no. 15, pp. 1685–1692, 2009.

[3] A. Kreso and J. E. Dick, "Evolution of the cancer stem cell model," *Cell Stem Cell*, vol. 14, no. 3, pp. 275–291, 2014.

[4] H. Uramoto and F. Tanaka, "Recurrence after surgery in patients with NSCLC," *Translational Lung Cancer Research*, vol. 3, no. 4, pp. 242–249, 2014.

[5] C. Booth and F. Shepherd, "Adjuvant chemotherapy for resected non-small cell lung cancer," *Journal of Thoracic Oncology*, vol. 1, no. 2, pp. 180–187, 2006.

[6] D. Ramazzotti, A. Lal, B. Wang, S. Batzoglou, and A. Sidow, "Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival," *Nature Communications*, vol. 9, no. 1, p. 4453, 2018.

[7] H. Nguyen, S. Shrestha, S. Draghici, and T. Nguyen, "PINSPlus: A tool for tumor subtype discovery in integrated genomic data," *Bioinformatics*, vol. 35, no. 16, pp. 2843–2846, 2019.

[8] T. Nguyen, R. Tagett, D. Diaz, and S. Draghici, "A novel approach for data integration and disease subtyping," *Genome Research*, pp. gr–215 129, 2017.

[9] E. A. Collisson, P. Bailey, D. K. Chang, and A. V. Biankin, "Molecular subtypes of pancreatic cancer," *Nature Reviews Gastroenterology & Hepatology*, vol. 16, no. 4, pp. 207–220, 2019.

[10] R. Dienstmann, L. Vermeulen, J. Guinney, S. Kopetz, S. Tejpar, and J. Tabernero, "Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer," *Nature Reviews Cancer*, vol. 17, pp. 79–92, 2017.

[11] D.-H. Le and V.-H. Pham, "Drug response prediction by globally capturing drug and cell line information in a heterogeneous network," *Journal of Molecular Biology*, vol. 430, no. 18, Part A, pp. 2993 – 3004, 2018.

[12] G. T. T. Nguyen and D.-H. Le, "A matrix completion method for drug response prediction in personalized medicine," in *Proceedings of the Ninth International Symposium on Information and Communication Technology*, ser. SoICT 2018. New York, NY, USA: Association for Computing Machinery, 2018, pp. 410–415.

[13] D.-H. Le and D. Nguyen-Ngoc, "Multi-task regression learning for prediction of response against a panel of anti-cancer drugs in personalized medicine," in *2018 1st International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2018, pp. 1–5.

[14] Q.-H. Nguyen, , H. Nguyen, T. Nguyen, and D.-H. Le, "Multi-omics analysis detects novel prognostic subgroups of breast cancer," *Frontiers in Genetics*, vol. 11, p. 1265, 2020.

[15] R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton, "The causes and consequences of genetic heterogeneity in cancer evolution," *Nature*, vol. 501, no. 7467, pp. 338–345, 2013.

[16] M. Greaves, "Darwin and evolutionary tales in leukemia," *ASH Education Program Book*, vol. 2009, no. 1, pp. 3–12, 2009.

[17] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, no. 7381, pp. 306–313, 2012.

[18] S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic Acids Research*, vol. 40, no. 19, pp. 9379–9391, 2012.

[19] P. Chalise and B. L. Fridley, "Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm," *PLOS ONE*, vol. 12, no. 5, p. e0176278, 2017.

[20] D. Wu, D. Wang, M. Q. Zhang, and J. Gu, "Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification," *BMC Genomics*, vol. 16, no. 1, p. 1022, 2015.

[21] E. F. Lock and D. B. Dunson, "Bayesian consensus clustering," *Bioinformatics*, vol. 29, no. 20, pp. 2610–2616, 2013.

[22] P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild, "Bayesian correlated clustering to integrate multiple datasets," *Bioinformatics*, vol. 28, no. 24, pp. 3290–3297, 2012.

[23] Q. Mo, R. Shen, C. Guo, M. Vannucci, K. S. Chan, and S. G. Hilsenbeck, "A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data," *Biostatistics*, vol. 19, no. 1, pp. 71–86, 2018.

[24] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, N. Schultz, C. Sander, R. S. Powers, M. Ladanyi, and R. Shen, "Pattern discovery and cancer gene identification in integrated cancer genomic data," *Proceedings of the National Academy of Sciences*, vol. 110, no. 11, pp. 4245–4250, 2013.

[25] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.

[26] R. Shen, Q. Mo, N. Schultz, V. E. Seshan, A. B. Olshen, J. Huse, M. Ladanyi, and C. Sander, "Integrative subtype discovery in glioblastoma using iCluster," *PLoS ONE*, vol. 7, no. 4, p. e35236, 2012.

[27] 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*, vol. 11, no. 3, pp. 333–337, 2014.

[28] N. Rappoport and R. Shamir, "NEMO: Cancer subtyping by integration of partial multi-omic data," *Bioinformatics*, vol. 35, no. 18, pp. 3348–3356, 2019.

[29] T. Nguyen, R. Tagett, D. Diaz, and S. Draghici, "A novel approach for data integration and disease subtyping," *Genome Research*, vol. 27, no. 12, pp. 2025–2039, 2017.

[30] S. Arslanturk, S. Draghici, and T. Nguyen, "Integrated cancer subtyping using heterogeneous genome-scale molecular datasets," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 25. World Scientific, 2020, p. 551.

[31] D. Tran, H. Nguyen, U. Le, G. Bebis, H. N. Luu, and T. Nguyen, "A novel method for cancer subtyping and risk prediction using consensus factor analysis," *Frontiers in Oncology*, vol. 10, p. 1052, 2020.

[32] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[33] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[34] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[35] M. Pierre-Jean, J.-F. Deleuze, E. Le Floch, and F. Mauger, "Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration," *Briefings in Bioinformatics*, 2019, bbz138.

[36] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[37] D. R. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972.

[38] T. Nguyen, A. Shafi, T.-M. Nguyen, A. G. Schissler, and S. Draghici, "NBIA: a network-based integrative analysis framework–applied to pathway analysis," *Nature Scientific Reports*, vol. 10, p. 4188, 2020.

[39] T.-M. Nguyen, A. Shafi, T. Nguyen, and S. Draghici, "Identifying significantly impacted pathways: a comprehensive review and assessment," *Genome Biology*, vol. 20, no. 1, p. 203, 2019.

[40] H. Nguyen, S. Shrestha, D. Tran, A. Shafi, S. Draghici, and T. Nguyen, "A comprehensive survey of tools and software for active subnetwork identification," *Frontiers in Genetics*, vol. 10, p. 155, 2019.

[41] E. Cruz, H. Nguyen, T. Nguyen, and I. Wallace, "Functional analysis tools for post-translational modification: a post-translational modification database for analysis of proteins and metabolic pathways," *The Plant Journal*, vol. 99, no. 5, pp. 1003–1013, 2019.

[42] T. Nguyen, C. Mitrea, and S. Draghici, "Network-based approaches for pathway level analysis," *Current Protocols in Bioinformatics*, vol. 61, no. 1, pp. 8–25, 2018.

[43] T. Nguyen, C. Mitrea, R. Tagett, and S. Draghici, "DANUBE: Data-driven meta-ANalysis using UnBiased Empirical distributions - applied to biological pathway analysis," *Proceedings of the IEEE*, vol. 105, no. 3, pp. 496–515, 2017.

[44] T. Nguyen and S. Draghici, *BLMA: A package for bi-level meta-analysis*, Bioconductor, 2017, r package.

[45] T. Nguyen, D. Diaz, R. Tagett, and S. Draghici, "Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data," *Nature Scientific Reports*, vol. 6, p. 29251, 2016.

[46] T. Nguyen, R. Tagett, M. Donato, C. Mitrea, and S. Draghici, "A novel bi-level meta-analysis approach-applied to biological pathway analysis," *Bioinformatics*, vol. 32, no. 3, pp. 409–416, 2016.

[47] D. Tran, H. Nguyen, B. Tran, C. La Vecchia, H. N. Luu, and T. Nguyen, "Fast and precise single-cell data analysis using hierarchical autoencoder," *bioRxiv*, p. 799817, 2019.

[48] B. Tran, D. Tran, H. Nguyen, N. S. Vo, and T. Nguyen, "Ria: a novel regression-based imputation approach for single-cell rna sequencing," in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2019, pp. 1–9.

[49] J. Tanevski, T. Nguyen, B. Truong, N. Karaiskos, M. E. Ahsen, X. Zhang, C. Shu, K. Xu, X. Liang, Y. Hu, H. V. Pham, L. Xiaomei, T. D. Le, A. L. Tarca, G. Bhatti, R. Romero, N. Karathanasis, P. Loher, Y. Chen, Z. Ouyang, D. Mao, Y. Zhang, M. Zand, J. Ruan, C. Hafemeister, P. Qiu, D. Tran, T. Nguyen, A. Gabor, T. Yu, J. Guinney, E. Glaab, R. Krause, P. Banda, DREAM SCTC Consortium, G. Stolovitzky, N. Rajewsky, J. Saez-Rodriguez, and P. Meyer, "Gene selection for optimal prediction of cell position in tissues from single-cell transcriptomics data," *Life Science Alliance*, vol. 3, no. 11, 2020.

[50] A. Shafi, T. Nguyen, A. Peyvandipour, and S. Draghici, "GSMA: an approach to identify robust global and test Gene Signatures using Meta-Analysis," *Bioinformatics*, vol. 36, no. 2, pp. 487–495, 2019.

[51] J. C. Stansfield, D. Tran, T. Nguyen, and M. G. Dozmorov, "R tutorial: Detection of differentially interacting chromatin regions from multiple Hi-C datasets." *Current Protocols in Bioinformatics*, vol. 66, no. 1, pp. e76–e76, 2019.

[52] H. Nguyen, S. J. Louis, and T. Nguyen, "MGKA: A genetic algorithm-based clustering technique for genomic data," in *2019 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2019, pp. 103–110.

[53] A. Shafi, T. Nguyen, A. Peyvandipour, H. Nguyen, and S. Draghici, "A multi-cohort and multi-omics meta-analysis framework to identify network-based gene signatures," *Frontiers in Genetics*, vol. 10, p. 159, 2019.

[54] Y. Yan, T. Nguyen, B. Bryant, and F. C. Harris Jr, "Robust fuzzy cluster ensemble on cancer gene expression data," in *Proceedings of 11th International Conference*, vol. 60, 2019, pp. 120–128.

[55] B. Marks, N. Hees, H. Nguyen, and T. Nguyen, "MIA: A Multi-cohort Integrated Analysis for biomarker identification," in *Proceedings of the 9th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2018.

[56] A. Shafi, C. Mitrea, T. Nguyen, and S. Draghici, "A survey of the approaches for identifying differential methylation using bisulfite sequencing data," *Briefings in Bioinformatics*, vol. 19, no. 5, pp. 737–753, 2018.

[57] D. Diaz, M. Donato, T. Nguyen, and S. Draghici, "MicroRNA-augmented pathways (mirAP) and their applications to pathway analysis and disease subtyping," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 22. New Jersey: World Scientific, 2016, pp. 390–401.

[58] D. Diaz, T. Nguyen, and S. Draghici, "A systems biology approach for unsupervised clustering of high-dimensional data," in *The Second International Workshop on Machine Learning, Optimization and Big Data*, 2016.

[59] M. Menden, D. Wang, Y. Guan, M. Mason, B. Szalai, K. Bulusu, T. Yu, J. Kang, M. Jeon, R. Wolfinger, T. Nguyen, M. Zaslavskiy, S. Jang, Z. Ghazoui, M. Ahsen, R. Vogel, E. Neto, T. Norman, E. Tang, M. Garnett, G. Veroli, C. Zwaan, S. Fawell, G. Stolovitzky, J. Guinney, J. Dry, and J. Saez-Rodriguez, "Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen," *Nature Communications*, vol. 10, no. 1, p. 2674, 2019.