Multiple approaches to elucidate subtypes of lung cancer patients

Jiyeon Kim[1], Seokbo Shim[2], Kwoneel Kim[2]
[1]Department of Biomedical and Pharmaceutical Sciences, Kyung Hee University, Seoul, Korea
[2]Department of Biology, Kyung Hee University, Seoul, Korea

Introduction

According to the findings of a survey from Institute for Health Metrics and Evaluation (IHME), cancer is the second cause of death worldwide[1]. In 2019, Lung & Bronchus cancer was estimated a second most common new cancer cases through both sex, and most death cancer cases in both males and females[2]. About 80 percent to 85 percent of lung cancers are non-small cell lung cancer(NSCLC)[3]. Lung adenocarcinoma(LUAD) is the most common type of lung cancer seen in non-smokers, and smokers[3]. Therefore, understanding genetic and epigenetic interaction between genes are important to understand lung cancer. In this analysis, we tried to consult most influential genes in lung cancer patients. Our co-researcher team divided the subtypes of the samples by using the NMF(Non-negative Matrix Factorization) clustering method using global protein, phospho and acetyl data. We attempted to explain this sample subtype with different methodologies for the gene or sample.

Method

Construct co-expression regulatory network with multi-omics data

We used expression data of 515 LUAD and 501 LUSC patients from TCGA database. By bioinformatics approaches, co-expression networks are increasingly being used[4]. We used WGCNA(Weighted correlation network analysis), Clust, CC(CrossClustering), K-means, SOMs, HC(Hierarchical Clustering) for clustering co-expression genes as a module of highly correlated expression genes. We divided genes into 51 modules. Then, we collected genomic and epigenomic data from Encyclopedia of DNA Elements(ENCODE) database to understand transcriptional regulatory interactions based on co-expression modules. We modeled specific transcriptional regulation in each module using various epigenome data as described in the previous study[5]. Bayesian networks are a type of probabilistic graphical model that uses Bayesian inference for probability computations[6]. Using Bayesian networks, we constructed 50 statistically proved direct regulatory networks.

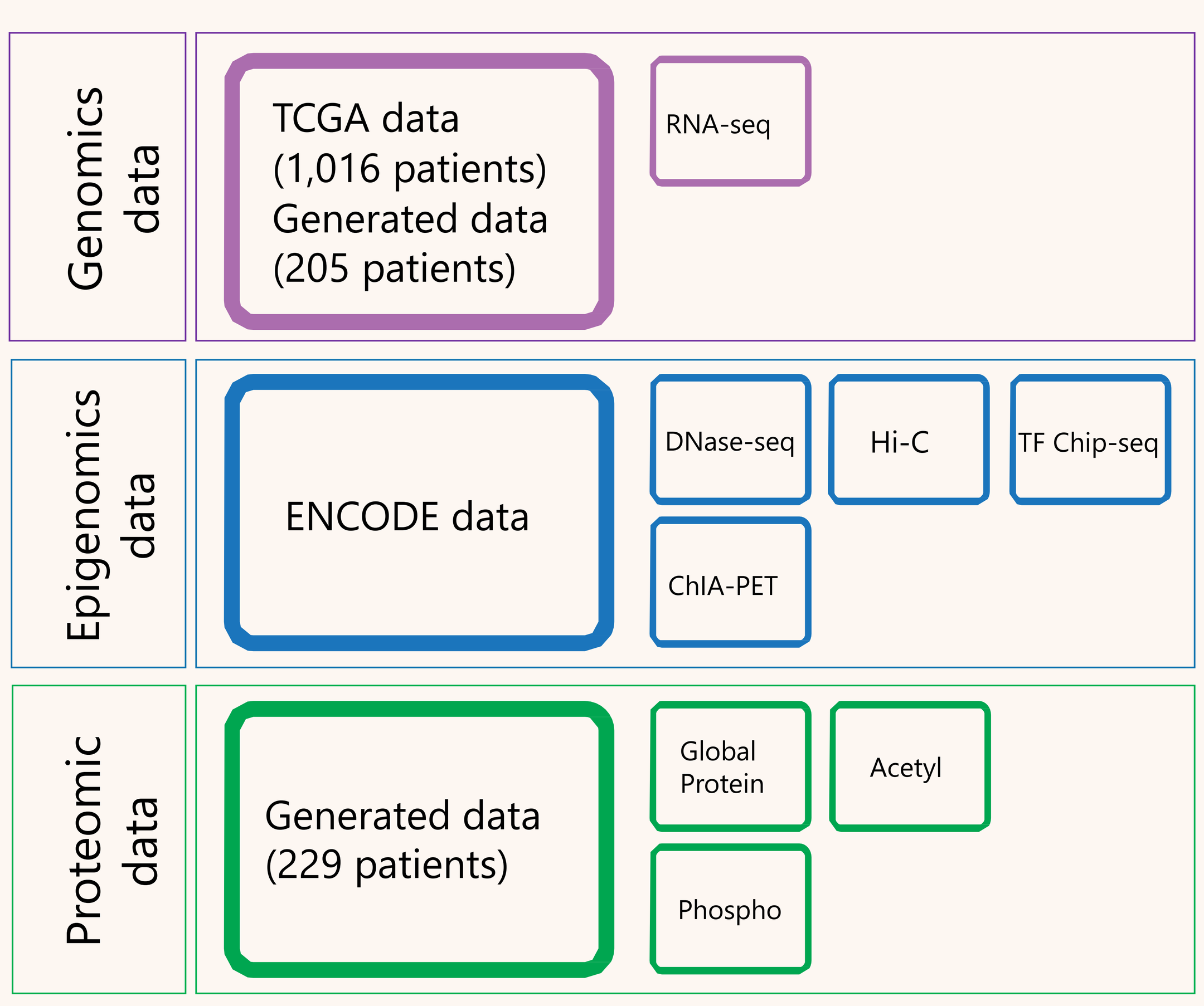

Fig1. Overview of Input data

Immune clustering

We used RNA expression data of 205 Tumor samples and 85 Normal adjacent tissues (NATS) samples generated from Asan Medical Center. We verified NMF subtypes(4 clusters) in each method using RNA expression data as described in the refs[9,10,11,12]. In the Immunogram clustering method, GSVA was performed with gene sets involved in the Cancer-immunity cycle(CLC), followed by K-means clustering(Hot, Cold, Warm)[9]. In other method, samples was clustered by Consensus Clustering(Hot, Cold, Warm, NAT_enriched) using xcell scores converted from RNA expression data[10, 11]. Finally, in the Fges clustering method, ssGSEA was performed for 28 gene sets related to immune invasion to calculate nes for each sample. After that, our data was clustered by Louvain clustering(Immune-enriched, fibrotic(IE/F), Immune-enriched, non-fibrotic(IE), fibrotic(F), immune-depleted(D))[12].
Among the various methods, the xCell method was best distinguished the NMF cluster, so our results were summarized based on the xCell method.

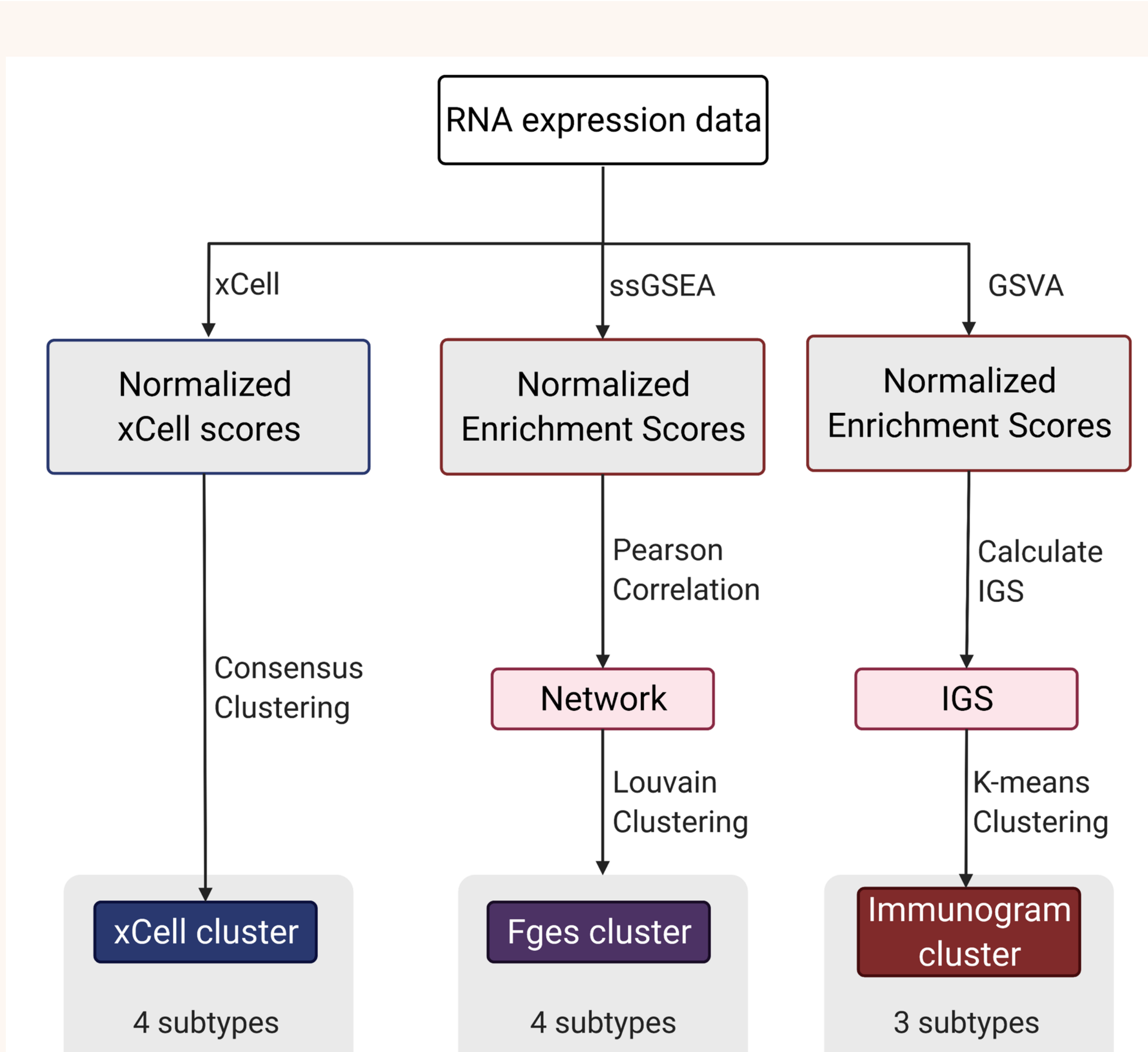

Fig2. Overview of Input data

Result

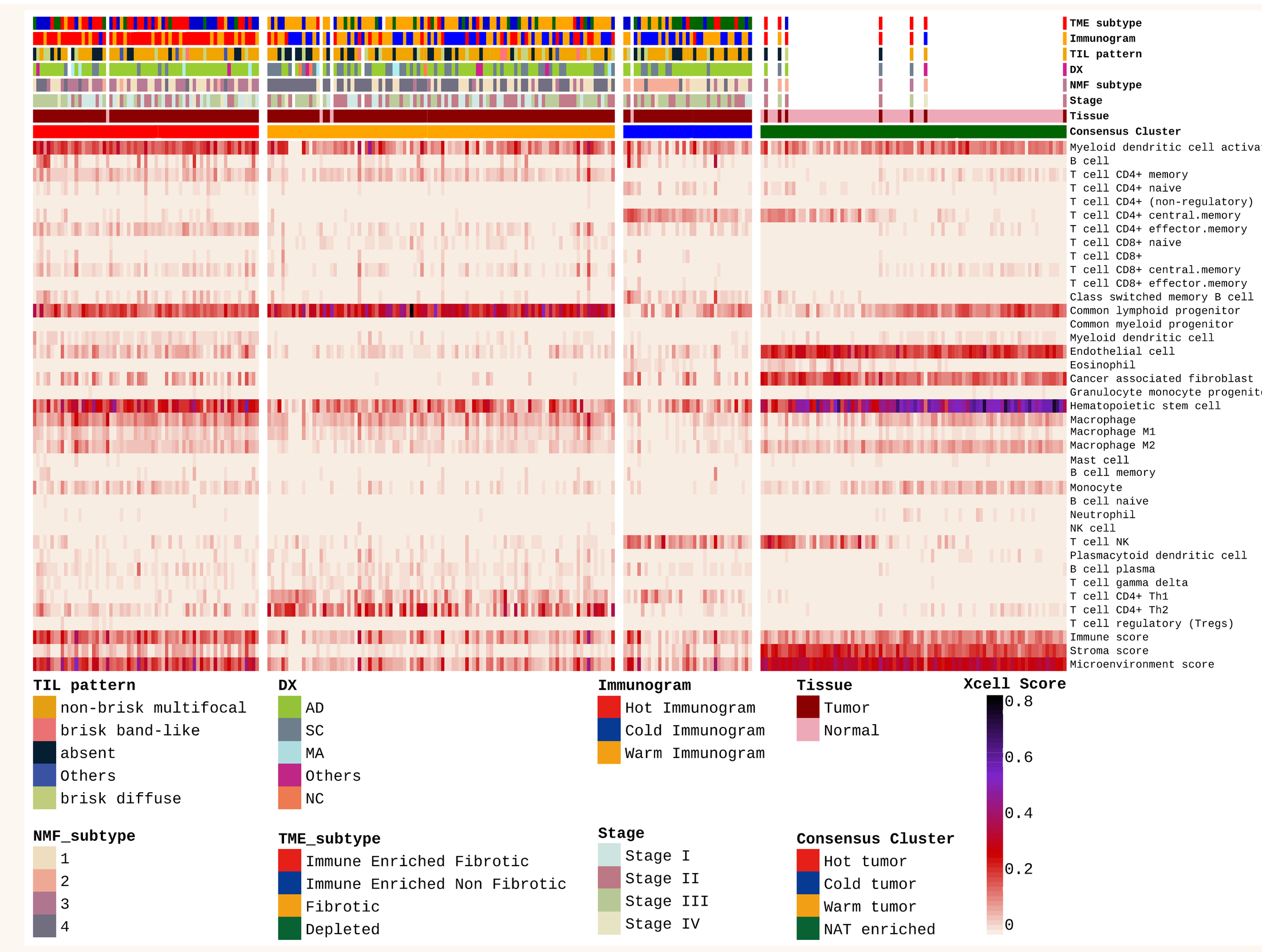

Fig3. Immune clustering results

The survival probability was higher in NMF subtype 3 and 4 clusters than in other clusters.

Immune cells such as CD8+ T cells are associated with prolonged survival of cancer patients and increased efficacy of immunotherapy[12].
According to the result of survival analysis of NMF subtypes, the survival probability was higher in NMF subtype 3 and 4 than other two subtypes.
And according to our results(Fig3), NMF 3 subtype has high scores in T cells CD8+ and enriched in hot tumor immunogram cluster.

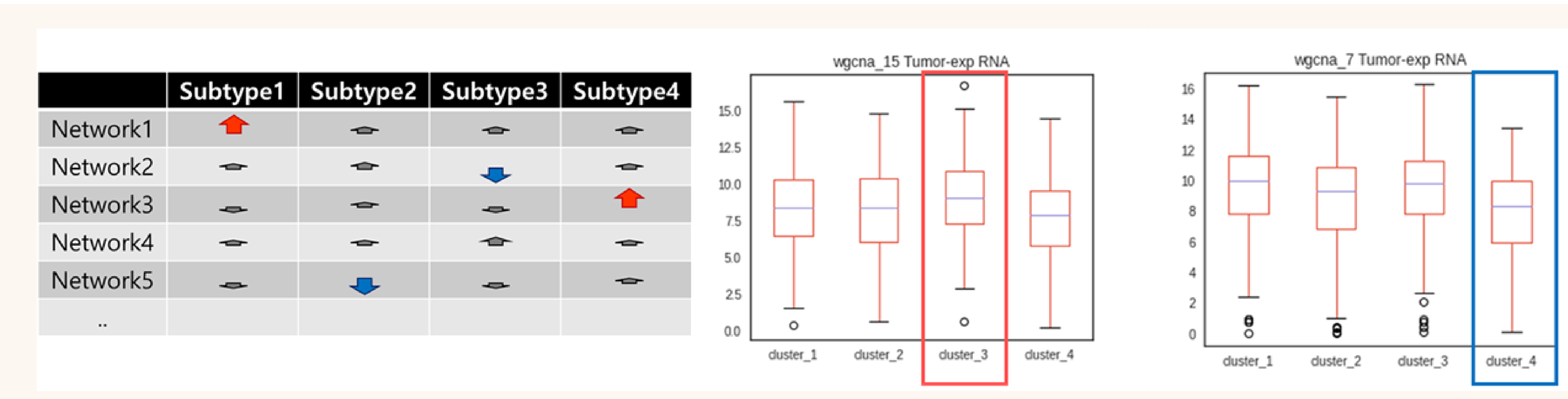

Fig4. Schematic diagram(left) and networks over-/under-expressed in specific subtype(right)

Network mapping to NMF subtypes

Schematic diagram of discovering specific subtype-specific overexpression and under expression transcriptional regulatory networks(left). And result of discovery of networks overexpressed in subtype3 and under expressed in subtype 4. As a result of performing pathway analysis on a network specifically overexpressed in patient subtype 3, cell adhesion and migration were enriched, which was a pathway that was enriched even when analyzed based on multi-omics data. In addition, CDH5, one of the core regulators of the transcriptional regulatory network, is known to be involved in adhesion, and SYNPO protein, which is a one of key factors in determining the sample subtypes, is a transcriptional target of CHD5 and also interacts with CHD5 proteins.

Future work

Our plan is to predict clinical traits(i.e., overall-survival, metastasis) with multi-omics data. Finding networks that outstanding describes the characteristics of samples may contribute to a progressing understanding of the correlation between alterations at the multi-omics level and the patients' clinical traits. We are also attempting to create a model that predicts patient's prognosis.
The next goal is to create model that can predict prognosis of patients when a certain treatment is given.
In addition, since the prognosis for treatment is different for each patient, it is also aimed to find a biomarker that can predict the prognosis for each treatment.

Reference

[1] Hannah Ritchie (2018) - "Causes of Death". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/causes-of-death' [Online Resource]
[2] Cancer statistics, 2019. Rebecca L. Siegel, Kimberly D. Miller, Ahmedin Jemal. ACS Journals. Vol 69, Issue 1 (2019)
[3] What is Non-Small Cell Lung Cancer? American Cancer Society website. https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/what-is-non-small-cell-lung-cancer.html. Revised May 16, 2016. Accessed June 10, (2019).
[4] WGCNA: an R package for weighted correlation network analysis. Peter Langfelder, Steve Horvath. *BMC bioinformatics* 9, 559 (2008)
[5] Global transcription network incorporating distal regulator binding reveals selective cooperation of cancer drivers and rick genes. Kwoneel Kim, et.al. Nucleic Acids Research. Vol 43, No12. 5716-5729 (2015)
[6] "Introduction to Bayesian Networks". Devin Soni. Towards data science website. https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e. (2018)
[7] Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Juri Reimand. et. al. nature protocols. 14, 482-517 (2019)
[8] Alvarez, M., Shen, Y., Giorgi, F. et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nat Genet 48, 838–847 (2016). https://doi.org/10.1038/ng.3593
[9] Ying Hu, Huaibo Sun, et. al. An Immunogram for an Individualized Assessment of the Antitumor Immune Response in Patients with Hepatocellular Carcinoma. Frontiers in Oncology. 2020 July 31; doi: 10.3389
[10] Satpathy S, et.al.; Clinical Proteomic Tumor Analysis Consortium. A proteogenomic portrait of lung squamous cell carcinoma. Cell. 2021 Aug 5;184(16):4348-4371.e40. doi: 10.1016/j.cell.2021.07.016.
[11] Bagaev A, et. al. Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. Cancer Cell. 2021 Jun 14;39(6):845-865.e7. doi: 10.1016/j.ccell.2021.04.014. Epub 2021 May 20. PMID: 34019806.
[12] Bruni D, Angell H, Galon J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. Nat Rev Cancer. 2020;20:662-80