Identification of early-stage lung cancer biomarker from single cell RNA sequencing

Yao Chen

Abstract:

The original research experiment utilized four lung cancer cell lines to obtain differential expressed genes in lung cancer cells. The single cell RNA sequencing raw data was generated from fluidigm C1 microfluidic system and Illumina sequencer.[1] The raw count was count per million (CPM) normalized and we are able to download the normalized count matrix file from NCBI website for reanalysis. The matrix files consist of read count from 40000 genes and 400 cells from each of the 4 lung cancer cell lines. We analyzed the normalized matrix data by the Seurat clustering workflow and GO annotation. At the end, the potential Differential expressed genes (DEGs) from the lung cancer cell lines were identified and filtered based on their average log fold change and p-value, the top 50 DEGs from each cluster were used for GO annotation analysis. We found common DEGs including CXCL1 and CXCL2 which in-depth investigated from the original research and we also find common GO annotation pathways for the DEGs. The Seurat workflow is able to reproduce similar result to the original literature's t-SNE and SC3 clustering workflow. [1]

Introduction:

Lung cancer has high mortality rate among all other cancer, it accounts for more than half of the new cancer diagnosed in north American. For Cancer in general, survival rate is better for early stage than the late stage. Especially now a day, many targeted therapy options to tackle recurrent somatic mutation in cancers. For example, lung cancer patient's life quality can be significantly improved if the patient is candidate for targeted drug like "Iressa" which interrupts signaling through the epidermal growth factor receptor (EGFR) in target cells. Therefore, it is essential to discover new molecular biomarkers for early-stage lung cancers to improve survivorship and quality of life. [2 3 4]

In this project, we attempted to use alternative workflow to investigate the lung cancer molecular biomarker by differential gene expression analysis from single cell RNA sequencing data. The normalized sequencing data was downloaded from NCBI project and analyzed by Seurat workflow and GO annotation. The candidate biomarkers are selected based on their log folder change from different clusters and gene annotation.

Results:

The normalized single cells RNA sequencing read count data was downloaded from the NCBI website. The matrix was analyzed by Seurat package in R. Initially, the Seurat object is created from the matrix data with minimal 100 cells and 1000 genes as parameters to eliminate cells with too many genes drop out or genes expressed from insignificant number of cells. The matrix then subset by removing excessive mitochondrial gene expression and range of number of genes from each cell. The data then log normalized for feature selection based on the cell-to-cell variation in gene expression. The scaled and selected features then undergo linear dimensional reduction method called principal component analysis (PCA). Four clusters are generated based on PC1 and PC2. PC heat map shows good heterogeneity from PC1 to PC6.[6]

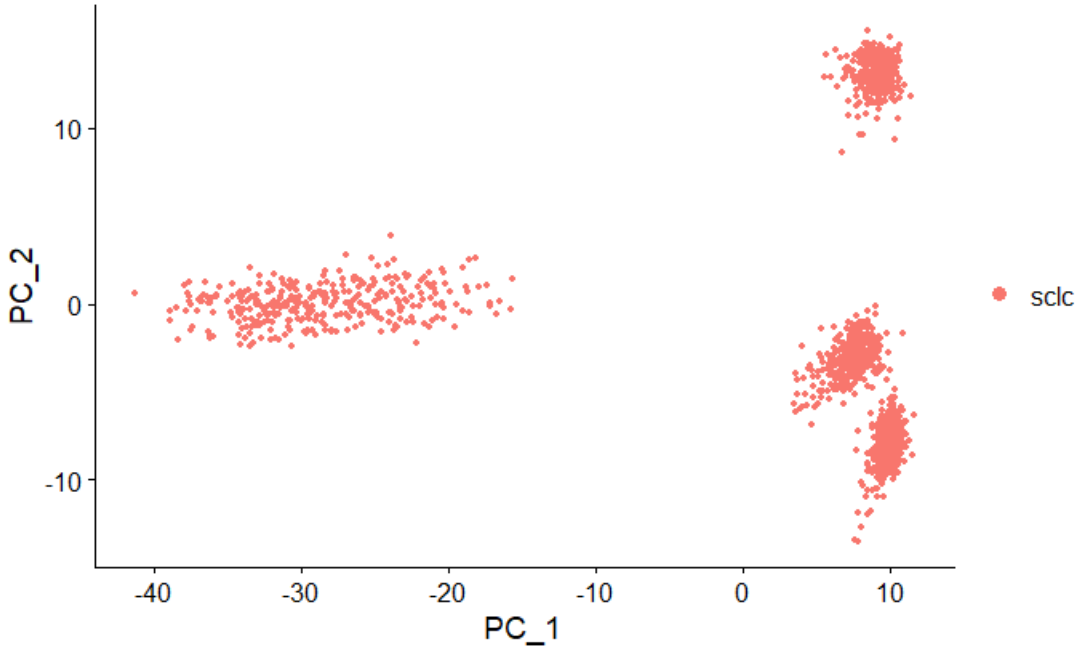

Figure 1. Four clusters shown under principal component 1 and principal component 2

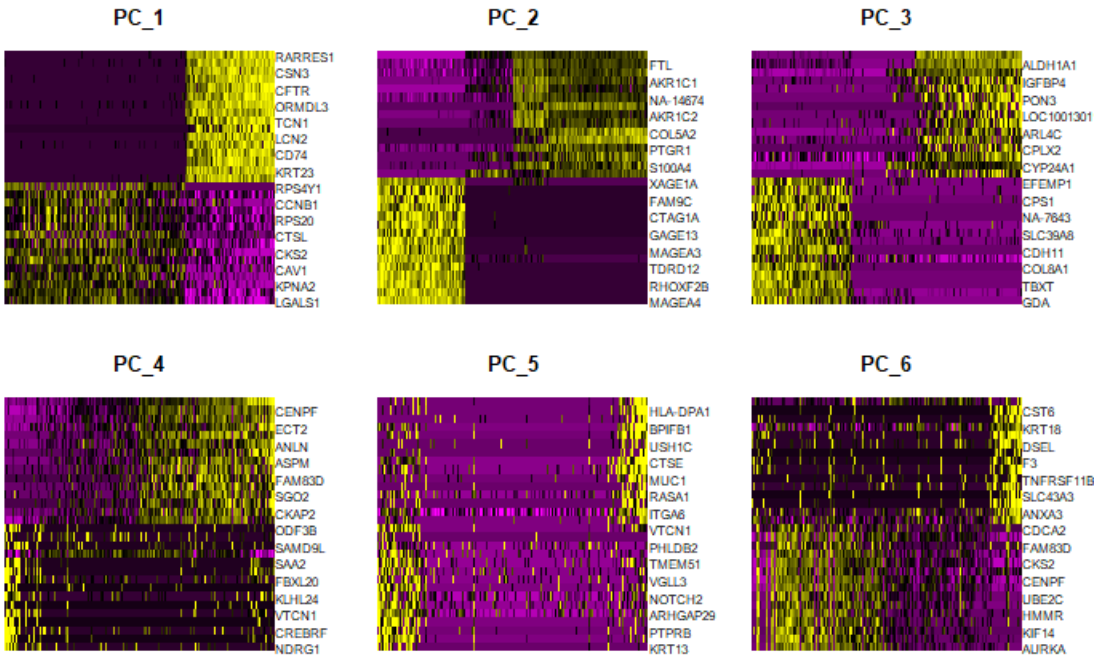

Figure 2. Gene expression dispersion based on each principal component; the dispersion decreases significantly from PC5.

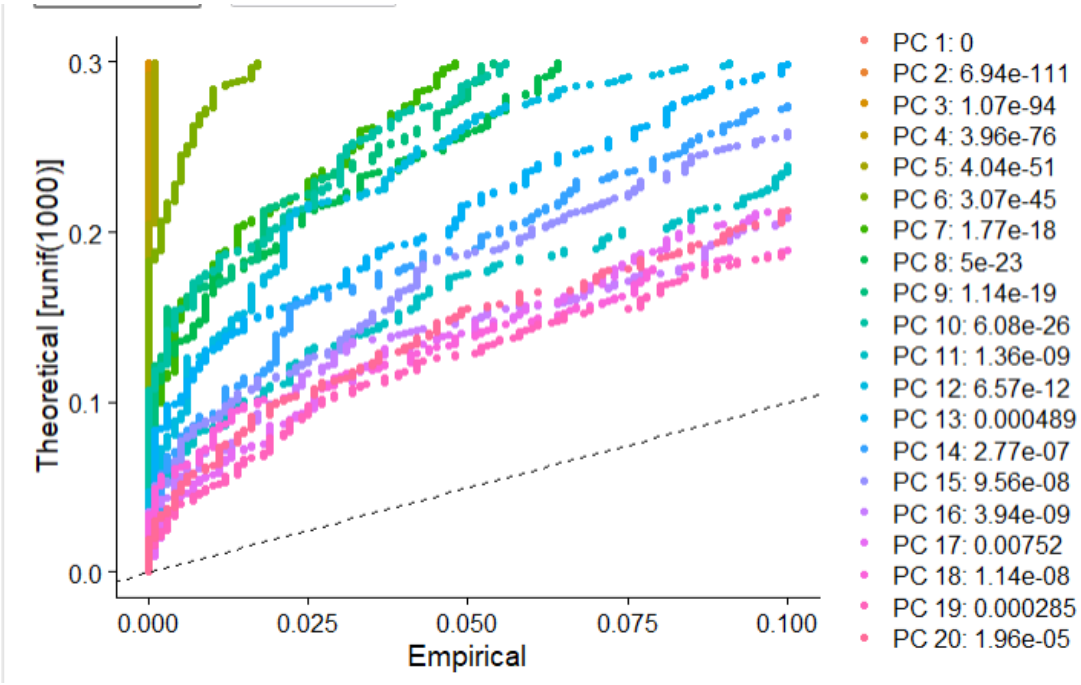

Figure 3. The Jackstrawplot demonstrates good p- values for the first 20 principal components.

F

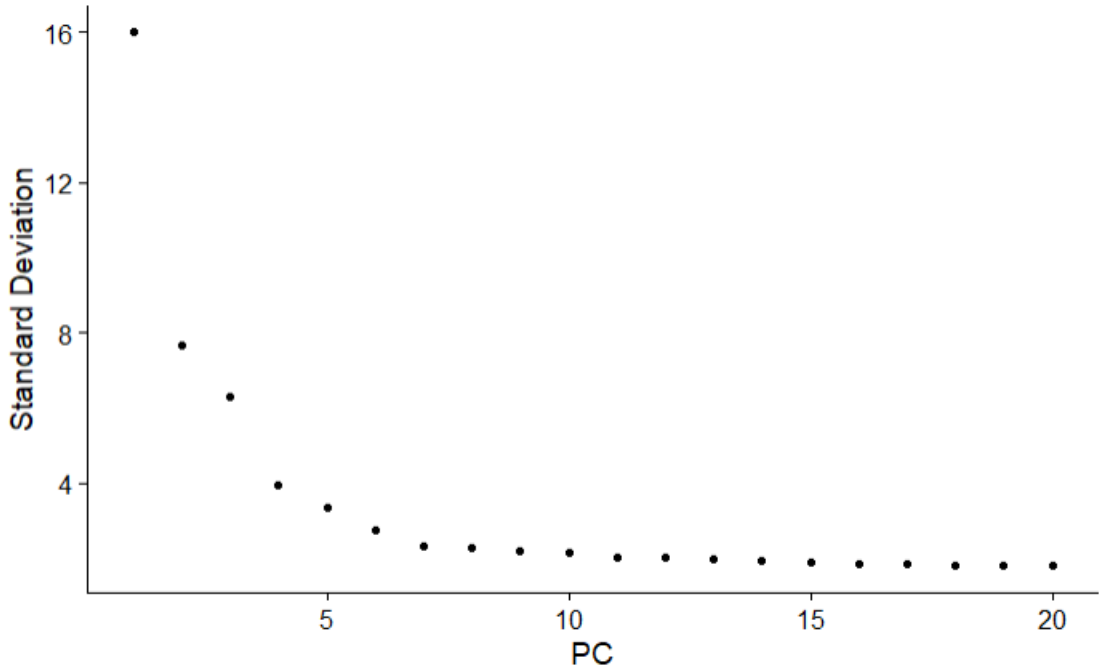

Figure 4. The elbow plot suggests that most of the good signals are coming from PC 1 to PC 6.

The cells are clustered based on Euclidean distance in PCA space and Louvain algorithm. The non-linear dimensional reduction method was used to visualize the clusters of cells. [6]

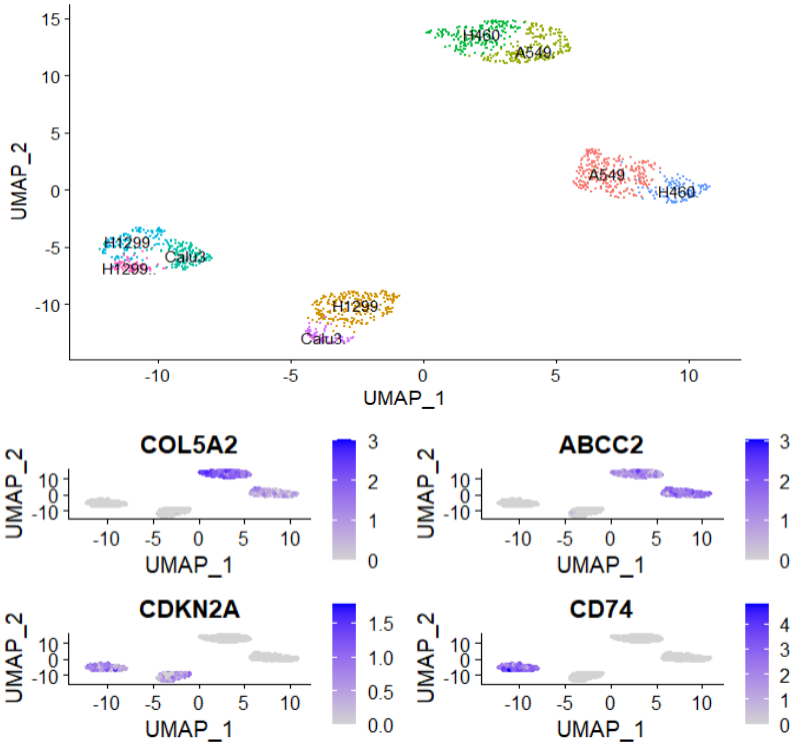

Figure 5. Umap plot based on the non-linear dimensional reduction technique, the 4 cell lines are assigned to the clusters based on their cell specific markers COL5A2, ABCC2, CDKN2A and CD74. [1]

The differential expressed markers are generated by comparing a single cluster to all other cells for all clusters. Only the markers with log fold change greater than 0.8 and minimum 50% percentage of cells from each group are selected. At the end top 50 markers based on their log fold change from each cluster are identified as good biomarkers for in-depth investigation. [6]

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
1	1.454740e-69	3.4430410	1.000	0.191	1.936405e-65	H1299..	FDCSP
2	2.552884e-49	3.3741355	0.984	0.919	3.398144e-45	H1299.	TOP2A
3	5.643754e-163	3.1858590	1.000	0.215	7.512401e-159	A549	KRT81
4	2.067476e-68	3.0695565	1.000	0.195	2.752017e-64	H1299..	CD74
5	1.368580e-66	3.0585248	1.000	0.192	1.821717e-62	H1299..	CSN3
6	3.679825e-68	3.0104941	1.000	0.197	4.898215e-64	H1299..	TCN1
7	3.764384e-68	2.9754504	1.000	0.183	5.010772e-64	H1299..	HLA-DRA
8	5.739810e-125	2.8948920	0.993	0.153	7.640261e-121	Calu3	CFTR
9	9.367175e-78	2.8457091	0.993	0.416	1.246865e-73	Calu3	IGFBP3
10	4.651343e-104	2.7724398	1.000	0.184	6.191402e-100	Calu3	RARRES1
1	1.112610e-121	-4.760354	0.000	0.938	1.480995e-117	H1299	S100A6
2	8.688089e-32	-4.565204	0.000	0.815	1.156471e-27	Calu3.	S100A6
3	7.986257e-109	-3.979160	0.068	0.898	1.063051e-104	H1299	AKR1C2
4	2.351482e-56	-3.899456	0.124	0.659	3.130057e-52	H1299	CXCL1
5	2.731459e-88	-3.867188	0.008	0.786	3.635845e-84	H1299	AKR1B10
6	1.913003e-112	-3.862281	0.064	0.914	2.546398e-108	H1299	SPP1
7	2.357508e-29	-3.820873	0.042	0.790	3.138080e-25	Calu3.	AKR1C2
8	3.546902e-31	-3.758365	0.000	0.806	4.721281e-27	Calu3.	SPP1
9	1.862391e-16	-3.736415	0.083	0.591	2.479029e-12	Calu3.	CXCL1
10	5.462337e-121	-3.711521	0.004	0.935	7.270916e-117	H1299	AKR1C3

Figure 6. Top 10 up regulated and down regulated biomarkers from the 450 selected biomarkers based on their log fold change.

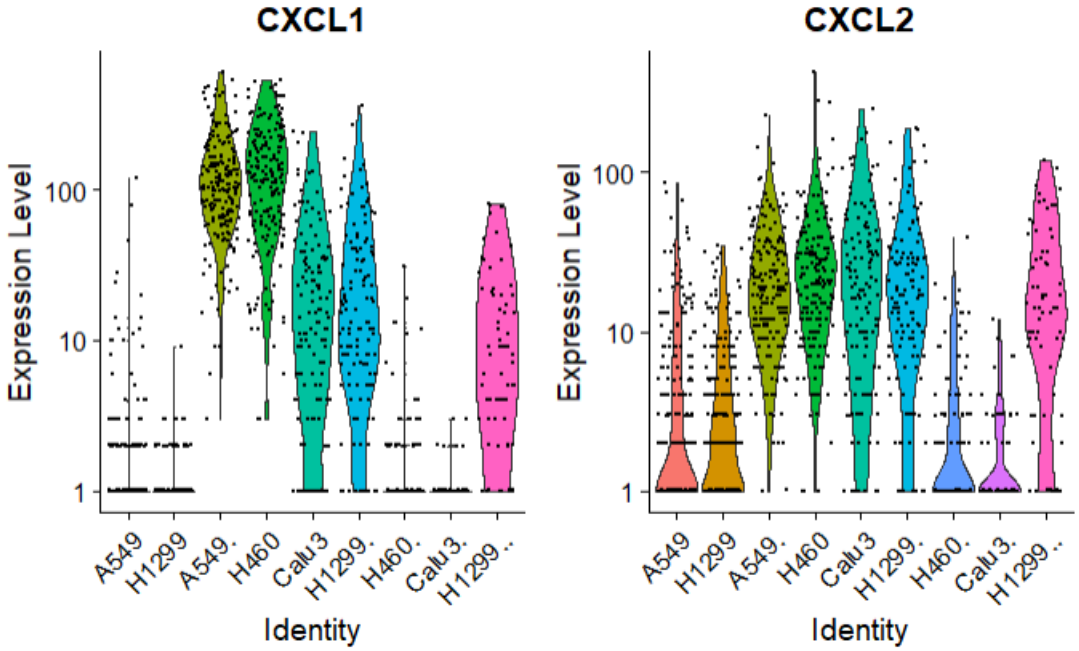

Figure 7. Violin plot to demonstrate the chemokine biomarkers identified from the literature. CXCL2 show good differential expression in all cell lines which suggest this could be a robust biomarker.

From the literature, it was reported that CXCL2 upregulated increased in KRAS mutated lung cancer cells may due to immune escape process. [5]

All the final 450 selected biomarkers are annotated by GO enrichment based on biological process, cellular component, and molecular function. We do find a few common biological processes to the original literature, daunorubicin metabolic process and doxorubicin metabolic process are also found in the literature's gene annotation. [1]

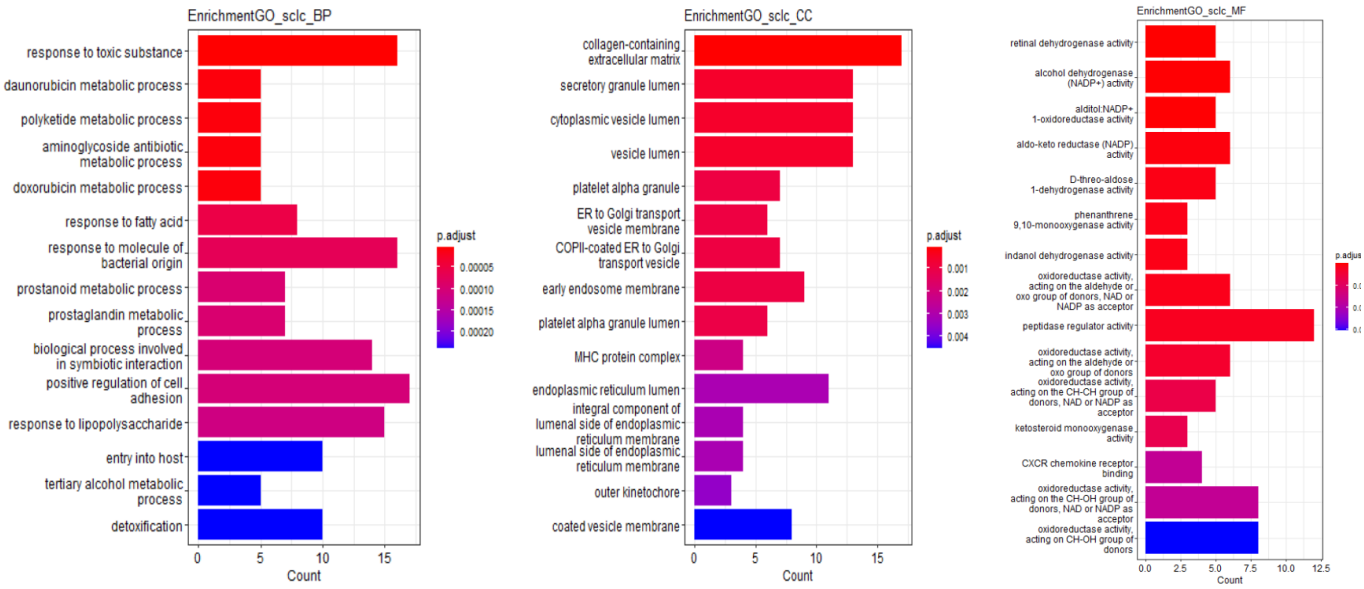

Figure 8. GO annotation based on biological process, cellular component and molecular functions for all the 450 selected biomarkers.

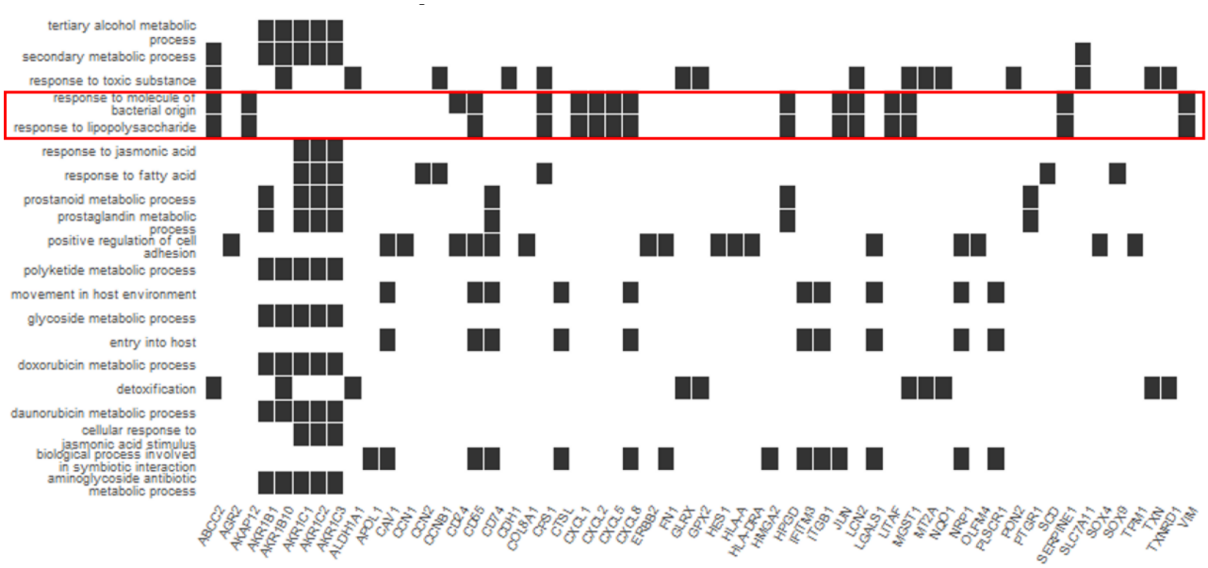

Figure 9. The GO annotation heatmap demonstrate the CXCL gene family play a important role in response to molecule of bacterial origin and response to lipopolysaccharide. [1]

Discussion:

Compared to the original literature that published this dataset we got similar result from the PCA, UMAP clustering, differential expressed genes and GO annotation. The original literature did a lot more in-depth investigation of the CXCL gene family and more comprehensive clustering technique such as 3-D clustering technique. Our limitation of investigation is restricted to bioinformatic analysis, as we don't have healthy and lung cancer patient samples for in-depth biomarker investigation.

The CXCL gene family in-depth investigation from the literature indicative imbalance sensitivity of this biomarker. As the marker did not demonstrate good differential expression in female patients. Even small portion of the male patients did not exhibit good expression in early-stage lung cancer. This suggest sole biomarker is not sufficient for early-stage lung cancer detection. More differential expression gene from the same or other potential biological pathway, cellular component and molecular function should be investigated before clinical implementation.

For in-depth early-stage molecular marker detection method, I will suggest to use NanoString technology to avoid suboptimal sample quality and quantity interference. Ultimately, we want to calculate and use correlation coefficient of each biomarker to construct a score system to determine if the patient is having lung cancer. The test result based on multiple informatic markers will be more reliable.

In terms of clinical implementation, this can also be implemented by Nanostring which is an oligo hybridization-based technology. Since we are detecting small log fold change of the gene expression, we want to completely avoid PCR's preferential amplification which can skew the quantification of gene expression.

References:

1. Kim, J., Xu, Z. & Marignani, P.A. Single-cell RNA sequencing for the identification of early-stage lung cancer biomarkers from circulating blood. *npj Genom. Med.* 6, 87 (2021).
2. World Health Organization. WHO Report on Cancer: Setting Priorities, Investing Wisely and Providing Care for All (World Health Organization, 2020).
3. Brenner, D. R. et al. Projected estimates of cancer in Canada in 2020. CMAJ 192, E199–E205 (2020).
4. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. CA Cancer J. Clin. 70, 7–30 (2020).
5. Arbour, K. C. & Riely, G. J. Systemic therapy for locally advanced and metastatic non-small cell lung cancer: a review. JAMA 322, 764–774 (2019).
6. Hao and Hao et al. Integrated analysis of multimodal single-cell data. Cell (2021)