Genetic Mutations Associated with Diffuse Large B-Cell Lymphoma

Jinghan Qiu[1] and Pingzhang Wang[2]
[1]Rutgers Preparatory School, Somerset, United States
[2]Peking University, China
Email: jqiu20@rutgersprep.org

***Abstract*—Diffuse Large B-Cell Lymphoma (DLBCL) is the most common non-Hodgkin lymphoma (NHL) among adults. [1] A cancer of B cells, DLBCL can arise in any part of the body. [2] Although DLBCL is not well understood and classified right now, substantial credible clinical data took by authoritative organizations are available online. Online clinical data related to DLBCL include data of mutated gene, high sequence, gene expression, copy number, etc. Here, utilizing clinical data, the researcher finds the most frequently mutated genes, and by analyzing the llluminaHiSeq in TCGA DLBCL data set, the researcher finds eleven frequently mutated genes that have significant effect on certain genes' expression once mutated (ARID1A, HIST1H1E, MGA, ATM, SGK1, IRF8, TET2, BTG2, EP300, CHD8, MLL2). If these eleven genes' mutation can be controlled by medical therapies, patients with DLBCL may be treated because of the reducing irregular gene expressions.**

***Index Terms*—genetic mutations, Diffuse Large B-Cell Lymphoma (DLBCL), gene regulation, database**

I. Introduction

"Genetic and Functional Drivers of Diffuse Large B-Cell Lymphoma (DLBCL)" collected by Duke University in Cell 2017 [3] contains data that profiles 1001 patients' gene mutations. Frequently mutant genes can be determined by statistically analyzing this set of data. Genes which encode transcription factors regulate the expression of other genes. If the annotations of a certain gene indicate that it binds chromosome or DNA, the gene is considered participating in gene regulation. If a sample that has a gene that regulates other genes' expression mutated, then its other gene's expression can be significantly changed. In many oncological cases, the overexpression and underexpression of genes take a significant role in the cause of cancers. Data of DNA expression in DLBCL patients are recorded in the TCGA DLBCL data set [4] by high-throughput sequencing. The website SOURCE [5] can get the annotations of all frequently mutant genes found in former set of data. With the annotations of frequently mutant genes, genes that encode for transcription factors can be filtered out, in other words, genes that have the ability to control the expression of other genes, which are the genes of interest in this research. Samples with a gene of interest mutated are in mutant groups. The RNA-seq level of the other genes in a wild type group is compared with the RNA-seq level of mutant group statistically. The results will show the influence of genes of interest's mutation on the other genes' expression in DLBCL.

Manuscript received December 22, 2019; revised March 12, 2019.

II. Methods

A set of data published by Duke University in *Cell 2017* records the gene mutant in 1001 DLBCL patients' tumor samples. Statistical analysis counts the frequency of mutation of each gene in this set of data. Genes with a mutation frequency higher than 5% are further studied.

There are forty genes in total that have a mutation frequency higher than 5% in this data set. According to the annotations of these forty genes in SOURCE, eleven genes (ARID1A, HIST1H1E, MGA, ATM, SGK1, IRF8, TET2, BTG2, EP300, CHD8, MLL2) have DNA binding activities, Chromosome binding activities, or transcription regulation activities. When these genes mutate, it's likely that gene transcription activities that they regulate are impacted.

Since the data set published by Duke University does not contain information on the sample's RNA expression, further analysis of the gene mutations' influence on other gene's expression is based on TCGA Large B-Cell Lymphoma (DLBCL) data set, in which 48 samples' RNA expression data are recorded. Using the somatic mutation (SNPs and small INDELs) in this data set [6], we utilized the mutant samples' codes to represent frequently mutant gene for statistical analysis.

For each gene mentioned above (ARID1A, HIST1H1E, MGA, ATM, SGK1, IRF8, TET2, BTG2, EP300, CHD8, MLL2), the mutant samples and their corresponding llluminaHiSeq data [7] are grouped into the "mutant group," and the normal RNA expressions and their corresponding llluminaHiSeq data are grouped into the "wild group." To compare the wild groups' expression levels of all genes and those of the mutant groups, we apply a T test. If the p value of the comparison is lower than 0.05, then it's very likely the mutant gene affects that the compared gene's expression. If the p value of the comparison is lower than 0.01, then it's very likely the mutant gene influences that the compared gene's expression significantly. We pick genes with a p value less than 0.01 to choose the genes that are most influenced by the mutant gene.

doi: 10.18178/ijpmbs.9.2.87-90

To further select the genes whose expression is highly affected by the mutant gene, we calculate the mean of each somatic gene's HiSeq data in both the wild type group and the mutant group. The genes have a difference higher than 1 or lower than -1 in mean are highly influenced in their expression because of the mutant gene. These genes are selected to be presented in heatmaps.

III. Results

Table I shows 40 genes which has a mutation rate more than five percent in DLBCL samples. We selected these 40 genes for further data filtration.

TABLE I. Genes with a Mutation Frequency more than 5%

Gene Name	Frequency of mutation (%)
MLL2	24.5754246
MYD88	17.1828172
PIM1	14.985015
CREBBP	11.4885115
BCL2	10.5894106
HIST1H1E	10.4895105
SPEN	10.0899101
ARID1A	9.49050949
TP53	9.39060939
SOCS1	9.19080919
CARD11	8.89110889
ARID1B	8.29170829
GNA13	8.29170829
SETD1B	8.29170829
TNFRSF14	7.89210789
DUSP2	7.69230769
SMARCA4	7.59240759
MGA	7.49250749
NOTCH2	7.29270729
ATM	7.19280719
SGK1	6.79320679
ZNF608	6.79320679
KLHL6	6.69330669
BIRC6	6.39360639
IRF8	6.39360639
MTOR	6.39360639
SETD2	6.19380619
TET2	6.19380619
EZH2	6.09390609
PIK3CD	6.09390609
ZNF292	5.89410589
BTG2	5.79420579
EP300	5.69430569
B2M	5.59440559
MEF2B	5.49450549
MLL3	5.49450549
STAT6	5.29470529
KLHL14	5.19480519
TBL1XR1	5.19480519
CD70	5.09490509

Using SOURCE as a reference, we selected eleven genes-- ARID1A, HIST1H1E, MGA, ATM, SGK1, IRF8, TET2, BTG2, EP300, CHD8, MLL2-- from the genes listed above. These eleven genes all have activities related to gene transcription, in other words, the abilities to control the expression level of other genes.

After filtration using methods described in methods, we determined that genes which correspond to a p value less than 0.01 and have a difference higher than 1 or lower than -1 in mean expression level between the wild-type and mutant-type groups are influenced by the mutant gene of interest on a significant level. In a heatmap for a specific gene of interest (one of the eleven translation-related genes listed above), they are represented in rows.

IV. Conclusion

Above are genes that frequently mutate in DLBCL samples which have a mutation frequency larger than 5%. There are eleven frequently mutant genes among them that highly influence other somatic genes' We constructed heatmap to represent the gene expression levels of all 48 samples in the dataset. A gap in the heatmap separates the samples in the mutant type group from samples in the wild type group. There exists difference in color between the left side of the gap and the right side of the gap, representing that their expressions in the wild type group are different from that of the mutant group. An obvious color difference indicates that when the gene of interest (which is the heatmap's title) is mutated, the somatic genes presented using color code in the heatmap are affected more than significantly. Among the eleven genes which influence other genes' expression significantly, we can observe a relative evidence visual difference in most genes' expression levels for genes ARID1A, MGA, IRF8, TET2, CHD8, whose heatmaps (Fig. 1-Fig. 5) are presented below.

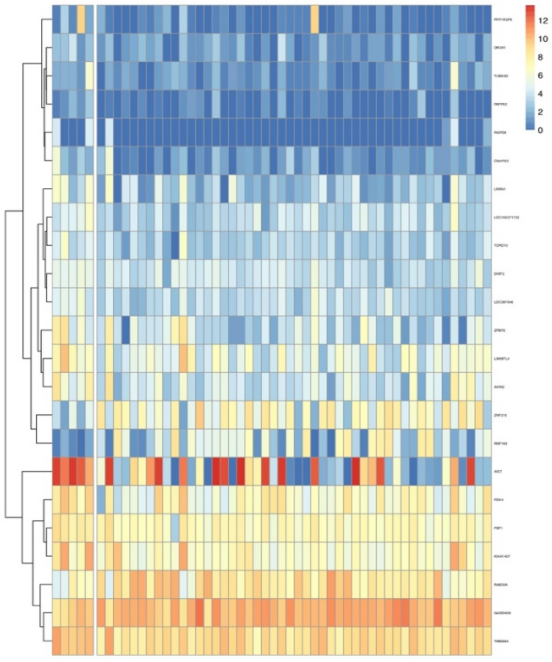

Figure 1. ARID1A.

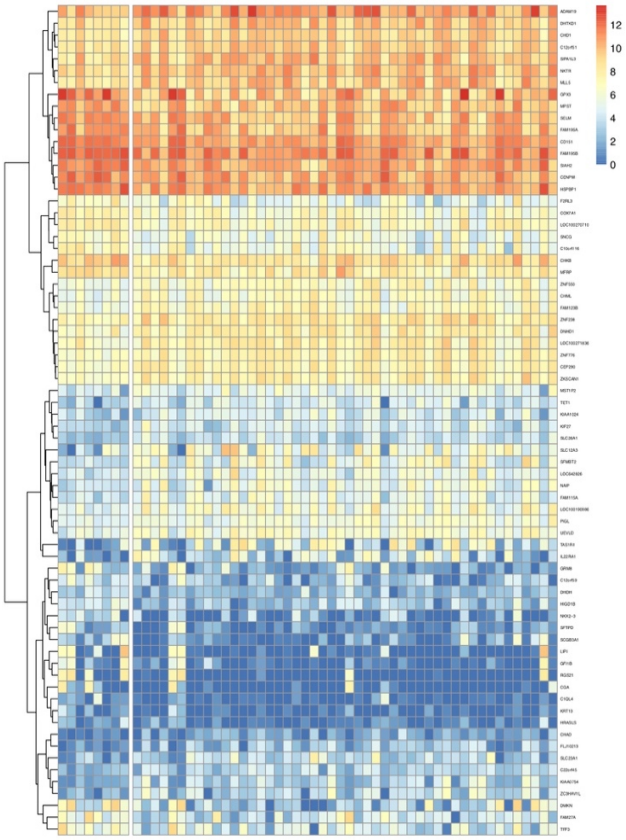

Figure 2. TET2.

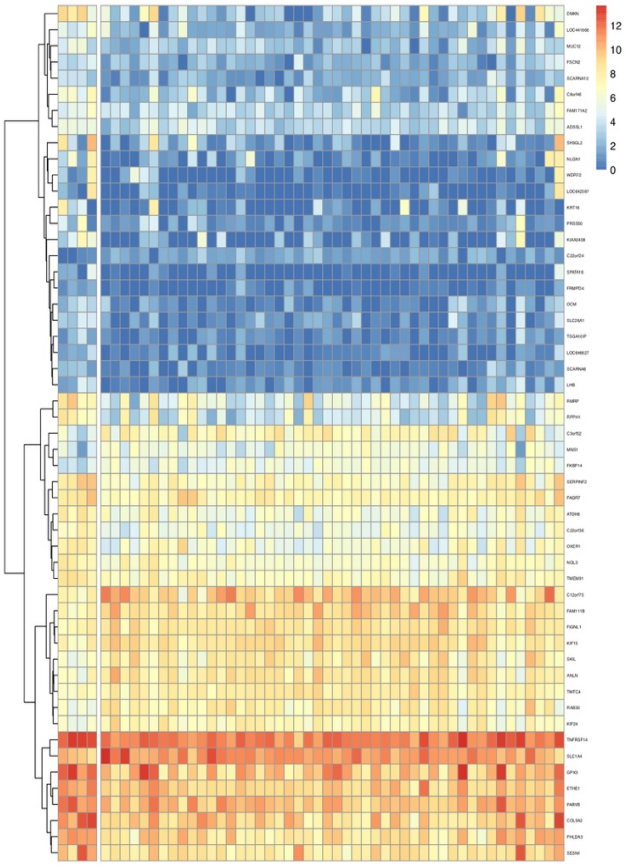

Figure 3. MGA.

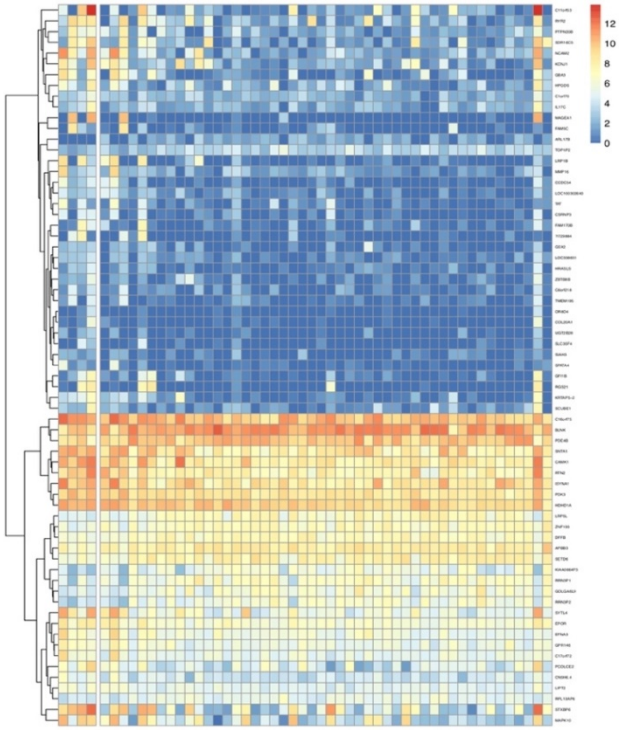

Figure 4. CHD8.

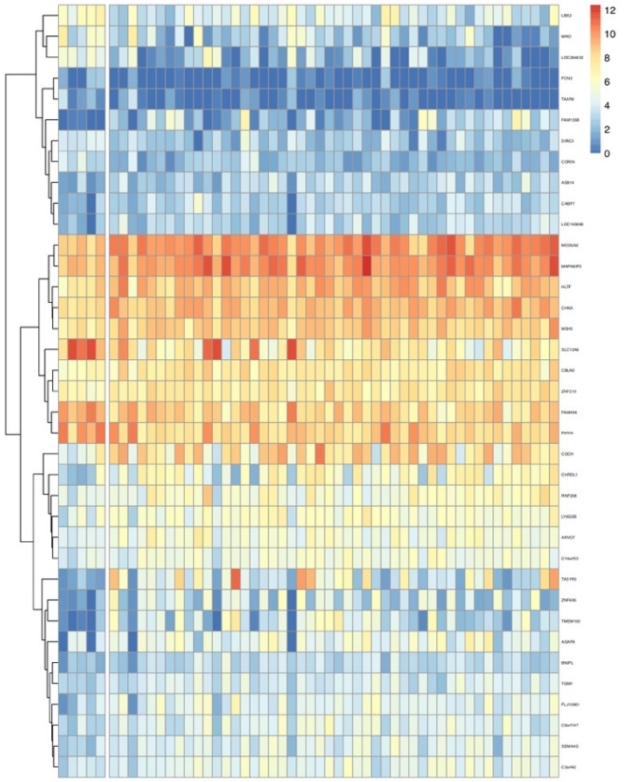

Figure 5. IRF8.

V. Discussion

The mutations of ARID1A, HIST1H1E, MGA, ATM, SGK1, IRF8, TET2, BTG2, EP300, CHD8, MLL2 ,which are the most frequently mutant genes in DLBCL, indeed

significantly influence other gene's expression in DLBCL. Therefore, these eleven genes might be the targets of medical treatment for DLBCL.

Further research on these eleven genes' pathways should be done to determine the mechanisms behind the statistically significant impact that these gene have on other genes' expression levels. Study of such pathways may reveal a general pattern in a certain gene's mutation and provide insight to mutant genes and their impact in cancers besides DLBCL. Meanwhile, when filtrating impacted genes, p-level selection could be tightened by using a $p<0.001$ as a standard of genes that are impacted in a large extent.

This statistic method in analyzing related gene mutations in DLBCL could also be applied to other cancers or diseases. The central idea of this method that the correlation between numeric gene expression levels and gene mutations remains when it's utilized in other data analysis associated with genetic data. Finding the most frequently mutant genes from a relatively large population helps us to locate genes that may play a significant role in a certain disease. Next, a search for the genes' activities provide information regarding the possible impact of the frequently mutated genes. In this research, we chose to analyze gene transcription and general trend in gene expression, so we selected genes that encode transcription factors or chromosome binding proteins. This target could be adjusted in other cases when the direction of study changes. For instance, genes that encode repressors might be selected to study the abnormally high expression level of many genes in a certain disease. Finally, dividing patients to wild-type group and mutant-type group, a statistical comparison could be made between the RNA-seq level in these two groups. In other cases, RNA-seq level might not be the object of statistical analysis; instead, other genetic data, such as DNA methylation or copy number, could be analyzed according to the specific requirements of the study.

Besides DLBCL, the exact same method could be employed to other cancers. A comparison between the results would be indicative if there are overlaps of resulted genes that mutate frequently and control the other genes' expression level significantly.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Jinghan Qiu conducted the research; Dr. Pingzhang Wang is Jinghan Qiu's advisor and joined to revise the paper.

Acknowledgment

The author wish to thank TCGA Datasets and cBioPortal, which provide the datasets that were analyzed in this research.

References

[1] The Non-Hodgkin's Lymphoma Classification Project, "A clinical evaluation of the international lymphoma study group classification of non-Hodgkin's lymphoma," *Blood*, vol. 89, no. 11, pp. 3909-3918, 1997.

[2] R. N. Mitchell, *et al.*, *Pocket Companion to Robbins & Cotran, Pathologic Basis of Disease*, Elsevier, 2016, p. 607.

[3] Duke University. (2017). *Diffuse Large B-Cell Lymphoma (Duke, Cell 2017)*. [Online]. Available: www.cbioportal.org/study?id=dlbcl_duke_2017&tab=summary

[4] *GDC TCGA Large B-Cell Lymphoma (DLBC)*. [Online]. Available: https://xenabrowser.net/datapages/?cohort=GDC TCGA Large B-cell Lymphoma (DLBC)&removeHub=https://xena.treehouse.gi.ucsc.edu:443

[5] Princeton University. *SOURCE Search.* [Online]. Available: https://source-search.princeton.edu/

[6] Baylor College of Medicine Human Genome Sequencing Center, TCGA Large B-Cell Lymphoma (DLBC), 2017, TCGA, 2017-09-08, somatic mutation (SNPs and small INDELs).

[7] University of North Carolina TCGA genome characterization center, TCGA Large B-Cell Lymphoma (DLBC), 2017, TCGA, 2017-10-13, gene expression RNAseq.

Jinghan Qiu worked as a Volunteering Research Assistant in Johns Hopkins School of Medicine during the summer of 2019. Her research on Conditions that Influence Nanoparticle Size and Yield in Nanoparticle Preparation will be presented at the 10th International Conference on Biomedical Engineering and Technology.