Statistical detecting of genes associated with PIK3C2B on lung disease

Jiamin Wei[1], Hongbo Wei[1], Yuxuan Xing[1], Bin Wang[1], Lu Han, Liang Tong[1,a*], Ying Zhou[1,b *]

[1] School of Mathematical Sciences, Heilongjiang University, Harbin 150080, China

Abstract. Statistical gene detection plays an important role in biostatistics and bioinformatics. So far, many gene loci associated with human complex disease have been found by statistical methods. However, it is difficult to find all the mutation genes that are associated with a certain disease. Researchers need to detect more associated genes aiming at a disease so that human will conquer the disease one day. In this paper, we considered a real and big data set and study the detection problem of genes associated with the PIK3C2B gene on lung disease. 168 significant genes associated with the PIK3C2B gene were detected at nominal significance level 0.001 by using statistical multiple testing method. The detected genes will provide some reference to further study the function of the PIK3C2B gene to lung disease for biologists and medical scientists.

1. Introduction

Biology life has made rapid development in recent years, in which statistical methods play an important role [1, 2]. Many people in the world suffer from some complex diseases (e.g., diabetes, cancers, Alzheimer's disease and so on), and these diseases are controlled by some mutation genes. Now it is glad to see that lots of gene loci associated with some important phenotypes/disease have been detected via statistical methods by researchers, and these loci are further validated by medical scientists [3, 4].

From view of biostatistics, researchers prefer to use linkage or association analysis method to conduct gene mapping. And from the perspective of bioinformatics, analyzing gene expression data is an effective way to find the latent genes or proteins that are responsible to a certain disease [5]. Many researchers engage in the analysis of gene expression data of cancers, and made significant progress on conquering this kind stubborn diseases [6, 7].

Lung cancer is a common complex disease and its incidence is higher than many other diseases. Many people die because of this disease every year. Recently more and more researchers try to study the genetic mechanism of lung cancer [8-11]. The PIK3C2B gene is considered to be an important gene that affects lung diseases, especially for lung cancer [12], and there is report that it is also related to some other complex diseases [13, 14]. The Cancer Cell Line Encyclopedia (CCLE) project was launched by the Broad Institute, and the Novartis Institutes for Biomedical Research and its Genomics Institute of the Novartis Research Foundation, and the CCLE provides public access to genomic data, analysis and visualization for more than 1100 cell lines (https://portals.broadinstitute. org/ccle/).

In this paper, we downloaded a data set of gene expression from the CCLE, and mainly analyzed the correlation between the PIK3C2B gene and other genes when acting on trait of lung cancer. We tested the degree of linear correlation between random variables using the Pearson correlation coefficient, which allows us to obtain genes that are positively, negatively, or uncorrelated with PIK3C2B gene. In addition, we constructed a linear relationship between 1 gene and other genes using a linear regression model and performed hypothesis testing. The results can further investigate the pathogenicity of genes, provide some valuable references for the relationship between genes and lung cancer traits, and provide theoretical support for further medical research.

2. Material and methods

The gene expression data that we downloaded from the CCLE is a typical big data, the original data set was studied for lung cancer traits and genotypes, which include 56202 genes and their mRNA expression values on 1019 cell lines, but the original data did not classify the lung cancer types. Among these 1019 cell lines, there are 188 ones that aim at lung tissues of patients with lung cancer.

The expression values of the PIK3C2B gene locate at the 4322th line of the data set. The distribution histogram of the expression values of the PIK3C2B gene on 188 lung tissues is presented in Figure 1. In Figure 1, the horizontal coordinates indicate the PIK gene expression values in 188 lung tissues, and the vertical coordinates indicate the

* Corresponding authors' e-mails: [a]2022012@hlju.edu.cn; [b]yzhou@aliyun.com

probability that the PIK3C2B gene expression values fall within the range of the values. The histogram has some difference with the histogram of the expression values of the PIK3C2B gene on the total 1019 cell lines (see Figure 2, the horizontal coordinate indicates the expression value of the PIK3C2B gene in 1019 lung tissues, and the vertical coordinate indicates the probability that the PIK gene expression value falls within that value.), i.e., the distribution of the former is a little right-biased. Our purpose is to detect the associated expressions of other genes with the PIK3C2B gene that act on lung tissue.

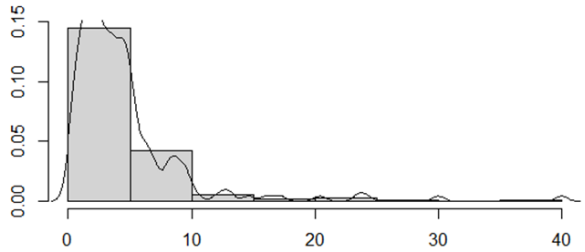

Figure 1. Histogram of the expression values of the PIK3C2B gene on 188 lung tissues.

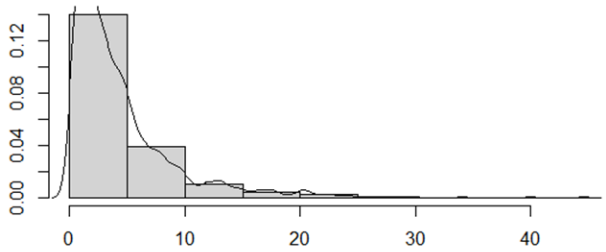

Figure 2. Histogram of the expression values of the PIK3C2B gene on 1019 cell lines.

2.1 Pearson correlation coefficient

Pearson correlation coefficient is widely used to measure the relationship of two random variables [15]. If random variable X has observation valued $X_1, X_2, \ldots, X_n$, and random variable Y has observation valued $Y_1, Y_2, \ldots, Y_n$, then the Pearson correlation coefficient has the following formula

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \quad (1)$$

where $|r| \le 1$. The Pearson correlation coefficient can well describe the linear relationship degree between random variable X and Y.

2.2 Test based on linear regression model

If a response variable is a continuous responsible variable, when considering the relationship of the variable with some other variables, a linear regression model can be constructed, and the hypothesis testing about the model coefficients can be performed to judge which variables have significant effect on the responsible variable. Consider linear regression model

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \quad (2)$$

the model coefficients b_j (j=0,1,…,p) can be estimated by the least square method, if a certain amount of observation values of variable y and x_j (j=1, 2,…,p) have been collected . In order to test H_0: b_j=0 v.s. H_1: Not H_0, the t test or F test can be applied.

3. Analysis results of real data

Aiming at the above-mentioned big data set, after we selected the expression values of all 56202 genes on lung issue, we obtain a 56202 by 188 data set.

Firstly, we calculated 56201 Pearson correlation coefficients of the PIK3C2B gene and the other genes (see the histogram of the Pearson correlation coefficients in Figure 3 and the top big 30 correlation coefficients in Table 1). From this result, we find that most of the correlation coefficients are positive and the biggest correlation coefficient is larger than 0.6, although large part of the values are close to zero, this is to say there are exactly genes that are positively associated with the PIK3C2B gene when acting on the trait of lung cancer. In the 56201 Pearson correlation coefficients, 53586 values are non-zero.

Secondly, we further test the linear relationship of the expression value of the PIK3C2B gene with the expression values of the other genes. Taking the expression value of each gene as responsible variable, and the expression value of the PIK3C2B gene as independent variable, we construct 53586 linear regression models. To find the significant relationship among them, we conduct 53586 hypotheses testing, and make Bonferroni adjustment. The significance level 0.001 is taken and then the one for each test is 0.001/53586 after Bonferroni adjustment. The P values of the t test for all genes are calculated and we found 168 significant results among the 53586 multiple tests. The top 30 significant P values are listed in Table 2. These significant results are obtained from the aspect of statistics, so some of them may be false positive, but we wish the inference results can provide some valuable reference for the biologists and medical scientists who can do further research on the pathogenicity of these detected genes and their relationship on the trait of lung cancer.

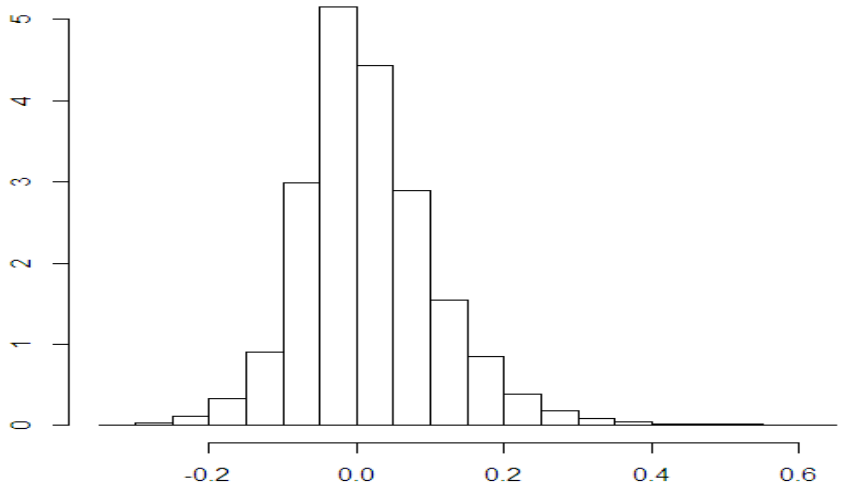

Figure 3. The histogram of coefficients between expression value of the PIK3C2B gene and expression value of the other genes.

Table 1. The top 30 correlation coefficients of expression values of the PIK3C2B and other genes.

No.	Gene	Coefficient	No.	Gene	Coefficient
1	ASCL2	0.621685	16	RP11-171A24.2	0.551137
2	CTA-992D9.6	0.618396	17	RHOU	0.54993
3	RP11-778O17.4	0.617189	18	RBM38	0.548353
4	TRPM5	0.605256	19	IGHV3-65	0.54623
5	RP13-198D9.3	0.589566	20	IGHV1-58	0.543053
6	IL1RAPL2	0.586601	21	NMU	0.540594
7	RP11-2E17.1	0.586347	22	RP11-571M6.17	0.536643
8	IGHV3-72	0.585688	23	IGHV4-59	0.536013
9	MSTN	0.574384	24	C1orf115	0.532988
10	C11orf53	0.568785	25	HLA-DPB2	0.531272
11	Y_RNA	0.563167	26	IGHV1-69	0.528687
12	AC002456.2	0.561312	27	AC004158.3	0.528141
13	IGKV3OR2-268	0.556687	28	PARD3B	0.527567
14	AVIL	0.55569	29	IGHV4-61	0.525359
15	IGHV3-49	0.554808	30	TAS2R62P	0.525193

Table 2. The significant P values of testing expression correlation of the PIK3C2B gene.

No.	Gene	P value	No.	Gene	P value
1	ASCL2	1.72E-21	16	RP11-171A24.2	2.50E-16
2	CTA-992D9.6	3.21E-21	17	RHOU	2.99E-16
3	RP11-778O17.4	4.03E-21	18	RBM38	3.77E-16
4	TRPM5	3.58E-20	19	IGHV3-65	5.15E-16
5	RP13-198D9.3	5.53E-19	20	IGHV1-58	8.19E-16
6	IL1RAPL2	9.12E-19	21	NMU	1.17E-15
7	RP11-2E17.1	9.51E-19	22	RP11-571M6.17	2.05E-15
8	IGHV3-72	1.06E-18	23	IGHV4-59	2.25E-15
9	MSTN	6.80E-18	24	C1orf115	3.44E-15
10	C11orf53	1.66E-17	25	HLA-DPB2	4.38E-15
11	Y_RNA	4.01E-17	26	IGHV1-69	6.27E-15
12	AC002456.2	5.34E-17	27	AC004158.3	6.76E-15
13	IGKV3OR2-268	1.08E-16	28	PARD3B	7.31E-15
14	AVIL	1.26E-16	29	IGHV4-61	9.90E-15
15	IGHV3-49	1.44E-16	30	TAS2R62P	1.01E-14

4. Conclusion and discussion

In this paper, we performed extensive statistical inference on the relationship of a large amount of gene expression variables based on a real data set. From the Pearson correlation coefficients and multiple hypotheses testing, we found 168 significant results that show the related genes with the PIK3C2B gene acting on the trait of lung cancer. If these results can be further researched by biologists and medical scientists, some valuable proof for treating the complex disease of lung cancer may be found.

Of course, some other correlation coefficients may be used in our analysis, and more complex models can be built to analyze this data set, so that some complementary results would be obtained. In addition, we can also study the effect on lung cancer traits when there is an interaction between genes and genes (GXG). Further research will be made in our future study.

Acknowledgments

The authors would like to thank the editors and referees for valuable comments on this paper. This research was supported by the Heilongjiang Province Statistical Science Research Project (2022B08), the National Natural Science Foundation of China (Grant No. 12071114) and 2022 Heilongjiang University Graduate Student Innovative Research Project (YJSCX2022-251HLJU).

References

1 Carvalho C M, Chang J, Lucas J E, Nevins J R, Wang Q and West M 2008 High-Dimensional sparse factor modeling: applications in gene expression genomics *J. Am. Stat. Assoc.* 103(484):1438-56

2 Telesca D, Müller P, Parmigiani G and Freedman R S 2012 Modeling dependent gene expression *Ann Stat.* 6(2): 542-60

3 Lian H 2008 MOST: detecting cancer differential gene expression *Biostatistics* 9(3):411-8

4 Rudin C M, et al 2012 Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer *Nature Genetics* 44: 1111-6

5 Bhardwaj N and Lu H 2005 Correlation between gene expression profiles and protein–protein interactions within and across genomes *Bioinformatics* 21(11): 2730–8

6 Ding L, et al 2008 Somatic mutations affect key pathways in lung adenocarcinoma *Nature* 455(7216): 1069-75

7 George J, et al 2015 Comprehensive genomic profiles of small cell lung cancer *Nature* 524(7563): 47–53

8 Liu P, et al 2012 Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing *Carcinogenesis* (7): 1270-6

9 Shi J, et al 2016 Somatic genomics and clinical features of lung adenocarcinoma: A retrospective study *PLoS medicine* 13(12): e1002162

10 McMillan E A, et al 2018 Chemistry-first approach for nomination of personalized treatment in lung cancer *Cell* 173(4):864-78

11 Nahar R, et al 2018 Elucidating the genomic architecture of Asian EGFR-mutant lung adenocarcinoma through multi-region exome sequencing *Nature communications* 9(1): 216

12 Kind M, Klukowska-Rötzler J, Berezowska S, Arcaro A and Charles R 2017 Questioning the role of selected somatic PIK3C2B mutations in squamous non-small cell lung cancer oncogenesis *PLoS One* 12(10): e0187308

13 Wei S Q, et al 2015 Silencing of PIK3C2B inhibits the proliferation and invasion of human prostate cancer PC-3 cells *Cancer* 35(10): 1063-9

14 Sabha N, et al 2016 PIK3C2B inhibition improves function and prolongs survival in myotubular myopathy animal models *Journal of Clinical Investigation* 126(9): 3613

15 Pearson K 1895 Note on regression and inheritance in the case of two parents *Proc. R. Soc. Lond.* 58: 240-2