Short Communication
Open Access
CrossMark

Randomized grouping statistical analysis in clinical omics biomarker discovery

Abstract

When using omics technology to study disease biomarkers, the differences are usually screened between disease group and control group. Because omics data is huge but sample size is limited, the differences between two groups may be randomly generated. To this end, we have proposed a randomized grouping statistical analysis strategy, which is suitable for the study of clinical omics disease biomarkers with limited sample size, and to determine whether the differences between two groups are randomly generated.

Keywords: randomized grouping, omics, biomarker

Volume 9 Issue 3 - 2020

Wenshu Meng, Youhe Gao
Department of Biochemistry and Molecular Biology, Beijing Normal University, Gene Engineering Drug and Biotechnology Beijing Key Laboratory, China

Correspondence: Youhe Gao, Beijing Normal University, Xin Jie Kou wai street No.19, Beijing, China, 86-13910861559, Email gaoyohe@bnu.edu.cn

Received: July 22, 2020 | **Published:** August 04, 2020

Introduction

Recent advancements in high-throughput omics based on mass spectrometry have promoted the study of disease biomarkers. In the basic stage of using omics data to study biomarkers, the differences between the disease group and control group are usually as candidate biomarkers for disease. However, whether these differences are caused by randomly generated is a question we should consider. First, it is easy to screen out differences between the two groups that meet the statistical criteria because the omics data is huge. For example, to investigate the urinary proteome differences between the Li and Han ethnic groups, the urine samples from 6 Li and 6 Han ethnic volunteers were analyzed by LC/LC-MS. In total, 1,555 urinary proteins were identified, and twenty-five of the urinary proteins were statistically significantly different.[1] Second, given the large time and economic costs of mass spectrometry identification and analysis, the quantity of samples based on omics research is often restricted. The number of human samples used in many omics clinical studies is relatively small.[2–5] Therefore, the differences between two groups are likely to be randomly generated in omics studies with limited sample size. For example, to distinct early- and advanced-stage lung adenocarcinoma, the proteome from 11 early and 11 advanced tumor samples were determined. In total, 155 proteins differentially expressed between early- and advanced-stage lung adenocarcinoma groups.[2] Since these differential proteins were identified in the proteomic data of limited clinical samples, it may be questioned whether these differences were generated randomly. Therefore, to improve the credibility of differences identified between two groups, we need to invest a larger number of samples in clinical omics studies. However, large numbers of clinical samples are difficult to collect, such as brain tissue samples of Alzheimer's disease, so the results of many studies may be questioned by the limited number of samples. Especially at the discovery stage of disease biomarkers, if the identified differences are randomly generated, it is difficult to succeed in the clinical verification stage, which is one of the reasons why only a few omics biomarkers enter clinical practice. How do we examine the differences between two groups are random when the clinical samples are limited?

In this paper, we proposed the randomized grouping statistical analysis strategy, which is suitable for the study of clinical omics disease biomarkers with limited sample size, and to determine whether differences between the two groups are randomly generated.

Randomized grouping statistical analysis

Randomized grouping statistical analysis is a strict strategy that should be performed when screening differences between two groups using omics data of limited samples. Specifically, screen the differences between the disease group and the control group, divide all samples into two groups randomly, screen the differences in each random combination, and calculate the average number of differences in all combinations. We compare it with the number of differences in the normal group to determine whether these are randomly generated. To realize randomized grouping statistical analysis, we use python to write a simple program. The workflow of randomized grouping statistical analysis is presented in Figure 1.

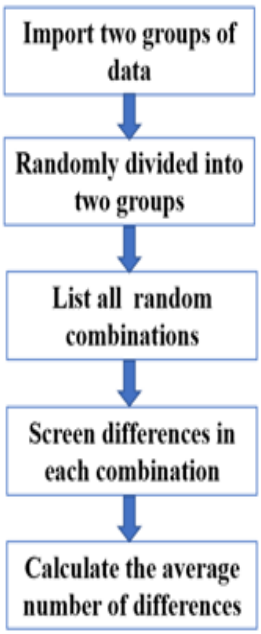

Figure 1 The workflow of randomized grouping statistical analysis program.

Example

We have performed randomized grouping statistical analysis of many urine proteome studies. For example, to examine the sensitivity of urine proteome, 9 rats were subcutaneously inoculated with approximately ten Walker-256 tumor cells, and urine proteomes on day 0, 13, and 21 were analyzed.[6] The results showed that 123 and 165 differentially proteins were identified on day 13 and 21 compared with day 0, respectively. It should be noted that no detectable tumor mass was observed in this study, and theoretically there should be no differences in urine proteomes. Were these differential proteins identified randomly generated? So, we conducted randomized grouping statistical analysis on the proteomics data. We imported 18 samples (Number1-18) from 9 rats on days 0 and 13 into the program. First, 18 samples were randomly divided into two groups, with a total of 48620 combinations. Then, the differential proteins in each random combination were screened according to the same criteria, and the average number of differential proteins in all random combinations was calculated. Final, the statistical analysis report of this random grouping was output. The results showed that the average number of differential proteins for all random combinations is 4, which indicates that only 3.25% of the 123 differential proteins we identified are likely to be randomly generated, indicating that the reliability is high. The output report is shown in Figure 2, and the statistical result is presented in Table 1. Similarly, we ran the data on days 21, and the results showed that only 2.42% of the 165 differential proteins identified were likely to be generated randomly, indicating that the reliability is high. The output report is shown in Figure 3, and the statistical result is presented in Table 1. In summary, these differential proteins identified in this study were due to subcutaneously inoculation of ten Walker-256 tumor cells rather than random differences.

Random combinations	Group Ⅰ	Group Ⅱ	Differential proteins
1	1,2,3,4,5,6,7,8,9	10,11,12,13,14,15,16,17,18	123
2	1,2,3,4,5,6,7,8,10	9,11,12,13,14,15,16,17,18	58
3	1,2,3,4,5,6,7,8,11	9,10,12,13,14,15,16,17,18	62
4	1,2,3,4,5,6,7,8,12	9,10,11,13,14,15,16,17,18	66
5	1,2,3,4,5,6,7,8,13	9,10,11,12,14,15,16,17,18	53
6	1,2,3,4,5,6,7,8,14	9,10,11,12,13,15,16,17,18	65
⋮	⋮	⋮	⋮
48615	9,10,11,12,13,15,16,17,18	1,2,3,4,5,6,7,8,14	65
48616	9,10,11,12,14,15,16,17,18	1,2,3,4,5,6,7,8,13	53
48617	9,10,11,13,14,15,16,17,18	1,2,3,4,5,6,7,8,12	66
48618	9,10,12,13,14,15,16,17,18	1,2,3,4,5,6,7,8,11	62
48619	9,11,12,13,14,15,16,17,18	1,2,3,4,5,6,7,8,10	58
48620	10,11,12,13,14,15,16,17,18	1,2,3,4,5,6,7,8,9	123

Figure 2 The output report of randomized grouping of 18 samples on days 0 and days13. Number1-8 represent Rat1-D0-Rat 9-D0; Number 9-18 represent Rat1-D13-Rat9-D13.

Random combinations	Group Ⅰ	Group Ⅱ	Differential proteins
1	1,2,3,4,5,6,7,8,9	19,20,21,22,23,24,25,26,27	165
2	1,2,3,4,5,6,7,8,19	9,20,21,22,23,24,25,26,27	58
3	1,2,3,4,5,6,7,8,20	9,19,21,22,23,24,25,26,27	51
4	1,2,3,4,5,6,7,8,21	9,19,20,22,23,24,25,26,27	69
5	1,2,3,4,5,6,7,8,22	9,19,20,21,23,24,25,26,27	49
6	1,2,3,4,5,6,7,8,23	9,19,20,21,22,24,25,26,27	64
⋮	⋮	⋮	⋮
48615	9,19,20,21,22,24,25,26,27	1,2,3,4,5,6,7,8,23	64
48616	9,19,20,21,23,24,25,26,27	1,2,3,4,5,6,7,8,22	49
48617	9,19,20,22,23,24,25,26,27	1,2,3,4,5,6,7,8,21	69
48618	9,19,21,22,23,24,25,26,27	1,2,3,4,5,6,7,8,20	51
48619	9,20,21,22,23,24,25,26,27	1,2,3,4,5,6,7,8,19	59
48620	19,20,21,22,23,24,25,26,27	1,2,3,4,5,6,7,8,9	165

Figure 3 The output report of randomized grouping of 18 samples on days 0 and days21. Number1-8 represent Rat1-D0-Rat 9-D0; Number 19-27 represent Rat1-D21-Rat9-D21.

Citation: Meng W, Gao Y. Randomized grouping statistical analysis in clinical omics biomarker discovery. *MOJ Proteomics Bioinform.* 2020;9(3):73–75. DOI: 10.15406/mojpb.2020.09.00283

Table 1 The results of randomized grouping of urine proteome at different time points of subcutaneous inoculation of approximately ten Walker-256 tumor cells in rats

Criteria for screening differential proteins	Time point	Total number of random combinations	Differential proteins	Average number of differential proteins with all random combinations	Percentage
$FC \geq 1.5$ *or* $\leq 0,67; P < 0.01$	$\frac{D13(n=9)}{D21(n=9)}$	$C_{18}^{9} = 48620$	123	4	3.25%
			165	4	2.42%

Discussion

In many clinical omics biomarker studies, grouping method is used to screen differences between the disease group and the healthy group, and these differences are believed to be caused by disease. However, when the omics data is huge but the clinical sample size is limited, it is easy to screen the differences between the two groups. Whether these differences are caused by randomly generated is a question we should consider. In addition, many complex diseases have different phenotypes, such as psychiatric conditions, and both health and disease groups may be heterogeneous. Therefore, we need to prove whether the direct grouping method is feasible for these complex diseases. Randomized grouping statistical analysis is a non-negligible strategy for studying disease biomarkers using omics data, especially in clinical studies with limited sample size. We compared the number of differences in normal grouping with the average number of differences in random grouping, to determine whether these are randomly generated.

If there is no difference in random grouping or the number of differences accounts for a small proportion of the number of differences in the normal group, it means that these differences between the two groups are caused by the disease, not randomly. If the number of differences in random grouping accounts for a large proportion of the difference in normal grouping, it means that the differences between the two groups may be generated at random and have little relationship with the disease itself. We need to consider whether the grouping method of the two sets of samples is appropriate. It should be noted that after screening the differences between the disease group and the control group, we will search for the existing biological evidence for these differences to further determine whether they were identified as candidate biomarkers.

Conclusion

In conclusion, we provide a randomized statistical analysis strategy that should be considered in clinical omics biomarker studies, and advocate that more researchers try to use this method in future studies.

Acknowledgments

This research was supported by Super Computing Center of Beijing Normal University.

Conflicts of interest

The authors declarse that they have no conflict of interest.

References

1. Zhang F, Li X, Ni Y, et al. Preliminary study of the urinary proteome in Li and Han ethnic individuals from Hainan. *Sci China Life Sci*. 2020;63(1):125–137.
2. Kelemen O, Pla I, Sanchez A, et al. Proteomic analysis enables distinction of early- versus advanced-stage lung adenocarcinomas. *Clin Transl Med*. 2020;10(2):e106.
3. Koch M, Mitulovic G, Hanzal E, et al. Urinary proteomic pattern in female stress urinary incontinence: a pilot study. *Int Urogynecol J*. 2016;27(11):1729–1734.
4. Yang S, Chen L, Chan DW, et al. Protein signatures of molecular pathways in non-small cell lung carcinoma (NSCLC): comparison of glycoproteomics and global proteomics. *Clin Proteomics*. 2017;14:31.
5. Suganya V, Geetha A, Sujatha S. Urine proteome analysis to evaluate protein biomarkers in children with autism. *Clin Chim Acta*. 2015;450:210–219.
6. Wei J, Meng W, Gao Y. Urine proteome changes in rats subcutaneously inoculated with approximately ten tumor cells. *Peer J*. 2019;7:e7717.

Citation: Meng W, Gao Y. Randomized grouping statistical analysis in clinical omics biomarker discovery. *MOJ Proteomics Bioinform.* 2020;9(3):73–75. DOI: 10.15406/mojpb.2020.09.00283