김연재

2018250103

BSMS222 Biostatistics

Submission date 11/22

김미향 <Visualise and investigate protein phosphorylation in non-smoking females by tumor stage>

[Summary]

Following the differentially regulated proteins according to three proteomic subtypes, correlation of protein phosphorylation and tumor stage was visualized. Depending on the first Figure, 5 proteins with the highest phosphorylation (P15088, P23946, P55083, Q9BU40, P08311)and lowest phosphorylation(O75940, O95171, O15541, Q9UKY7, Q16625) were chosen respectively. The result of plotting each high and low phosphorylation protein shows both the highest and lowest phosphorylated proteins were in tumor stage I, which can be seen by the density of the color compared with other tumor stages. 3 out of the 5 highest phosphorylated proteins (P15088, P23946, P08311) in tumor stage IA had a function of angiotensin maturation. In conclusion, as the tumor stage, IA and IB are early stage, a protein with angiotensin maturation would be highly phosphorylated in order for cancer cells to develop.

[Peer opinion]

First, it is thought that there was a lot of effort in preprocessing to create a heat map using this large data. in addition, I think it was a good choice for patients in order to change their order according to the stage. However, it would have been nice to find the median value of patients with something in common and check the amount of phosphorylation of proteins. Rather than obtaining the amount of phosphorylation of all proteins, it would have tended to classify proteins in relation to pathway, mechanism, and system of interest. One of the words the writer said was, "both highest and lowest phosphorylation of proteins exist in tumor stage IA" but I couldn't figure out what I found out through this. When using a heat map, it would have been easier to see if the scale color was used as a change in the darkness of one color. Because the log2T/N value has both - and + values, but in the end, since it is a compared value, there is no problem in expressing the degree of phosphorylation even if - is set as white and - is made darker as +. In Figure 2, there is a part I wonder whether it is intended or a mistake. In Figure 1, the larger the value, the red it is expressed, but in Figure 2, the darker it is, the blue it is expressed. I don't know why the scale color was reversed in one paper. If I wanted to distinguish it, I think I should have used other colors such as purple and yellow at all.

김민서 <Investigating the difference between Taiwan cohort(Chen et al., 2020) and CPTAC cohort patients' gene expression on transcriptome and proteome level.>

[Summary]

According to Chen's research, Taiwanese LUAD patients' samples have four unique characteristics such as a high proportion of never-smokers, EGFR mutations, females, early stage patients. This time by using CPTAC project data, which is the most recent and multi-national LUAD data, we would like to check whether the four characteristics of the Taiwan cohort appear noticeable. And the result shows that those 4 properties were responsible for the gene expression differences between cohorts' implicit differentiating characteristics. The 7 genes' expression also seemed to differ due to some patient properties, including gender, smoking status, country of origin, and LUAD stage.

[Peer Opinion]

I think it's very good to preprocess TW data and most recent research data using merge. In addition, the process of making it easy to see and aligning it with one huge data frame would have been very complicated and difficult, but it was successfully performed. There are a few things that I'm bummed about is, figures contain too much information. Not only is a large amount of information contained in a small size, but also plots of other topics are gathered in one figure. It would not have been difficult to grasp the information if each result had been attached together while showing the process of making a plot in code. Now, there is no more limit to figure attachment, so I hope the writer can distinguish more figures you want to show. It seems that A1 ~ D1 wanted to show that there was a difference in the aspects of the samples in the two cohorts. If so, I think it would have been better not to count the number of samples, but to compare the proportion of patients in each group. There is a question in the results from Figure1. In general, the graph shows the result count value in the form of a bar-plot, and it seems that the result value of TW is about 2,492. As far as I know, the number of cohort patients used in Chen is 103, but if so, is the remaining sample of 2300 to 2400 people classified as Taiwan patients in multi-nation? I may have completely misunderstood the writer's code, but if so, can it be said that the sample of multi-nation patients was evenly distributed?

김민영 <MAPK Pathway Relation with Tumor Stages in EGFR Activating Mutations Focused on Druggable sites>

[Summary]

Three druggable phospho-proteins of MAPK cascade that can be targeted by a known inhibitor and can potentially act as a biomarker was identified. Since the increased phosphorylation of the MAPK pathway correlates with EGFR activating mutations, the possibility of MAPK druggable targets acting as a biomarker for non-smoking female Taiwan cohorts with EGFR mutations is highly anticipated. And the result indicated that MAPK1_pT190 as a significant drug target and biomarker for tumor stages between stage 1A/1B and stage 3A. NSCLC defines stage 1 as cancer that has not yet metastasized, and stage 3 indicates that cancer has metastasized to the lymph nodes near the chest, so the transition from the early stage to the late stage may have significant implications.

[Peer Opinion]

The introduction shows what the writer is interested in and specific goals to observe have been set. In the first figure, the tendency was roughly examined by observing the phosphorylation of MAPK druggable sites for each stage of cancer. At this time, it was good to clearly point out the limitations of the data by specifying that the graph was not drawn because there were not many samples.

After that, the flow that led to visualization once again through Figure 2 was smooth by selecting a topic that could be compared in Figure 1. Figure 2 used a boxplot, and personally, making a boxplot horizontally as it is now seems to have been an excellent choice. It is regrettable that the figures are not clearly drawn on the plot and covered, but I think it can be easily supplemented.

Additional suggestions here are as follows. I wonder if this value varies depending on the type of mutation in EGFR. And now it's for women who are nonsmokers, so I hope you can check what will happen if you put male in the sample together. Finally, how about visualization that samples the world to see if the range of utilization can be further increased in the study that wants to measure the stage through markers.