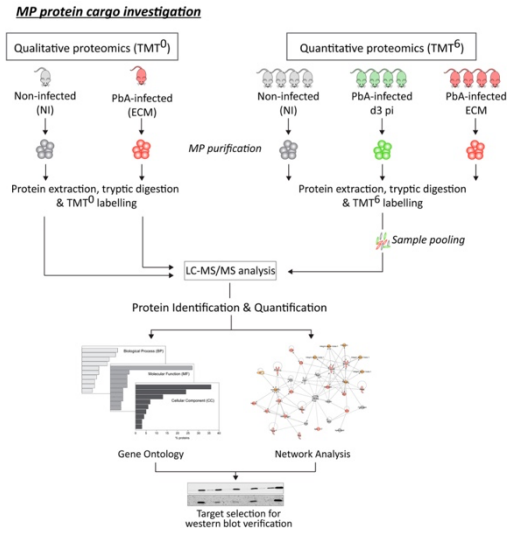
Assignment for SDC Biostatistics

Introduction

The analysis will be based on the results (TMT-6 labelling based protein mass spectrometry) comparing different stages of malaria presented in https://www.ncbi.nlm.nih.gov/pubmed/27917875.



At **TMT^6 quantitative experiments,** two parallel experiments were performed comparing plasma MP proteins from non-infected mice (NI, n = 4), PbA-infected mice at day 3 post-infection (d3 pi, n = 4) and PbA-infected mice at d8 post-infection when all the signs of cerebral malaria are detected (ECM, n = 4).

In order to analysis the data from these two parallel TMT^6 quantitative experiments, the processes are as following steps:

Loading data

First, we need to download the *AllQuantProteinsInAllSamples.csv* and read the file into the R environment (*read.csv* function). The rownames are the protein groups given by accession numbers. The table contains log-transformed protein quantifications to compare between replicates and three conditions.

What does it mean when you have multiple accession numbers?

A: The proteins represented by these accession numbers have similar amino acid sequence. If we zoom into the first row, we have A2BIM8, B5X0G2, P02762, P04938, P04939, P11589. While, they corresponding to Major urinary protein 18(MUP18), MUP17 MUP6, MUP11, MUP3, and MUP2 respectively. The different between A2BIM8 and B5X0G2 is just one amino acid. More detailly, the major urinary protein genes, Mup9, Mup17, and Mup18, all function as lipophilic chemical transporters modulating lipid metabolism(Thoß et al., 2019).

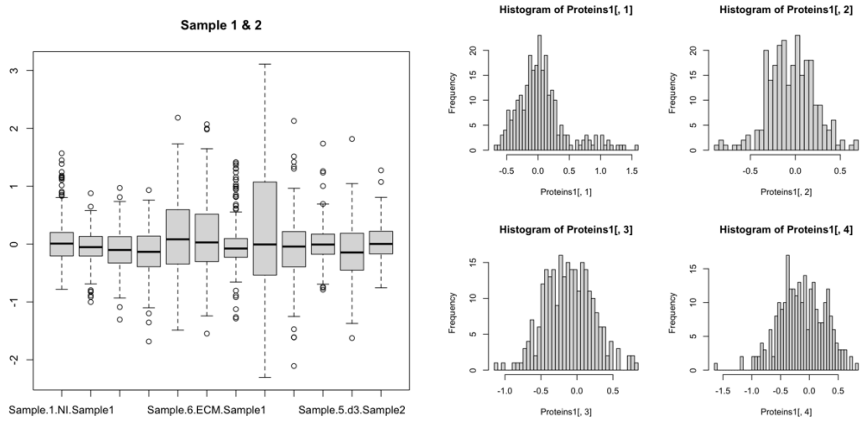
Data preprocessing

1. Viewing the dataset, we can notice that the first 6 columns is the sample1 and least 6 is belong to sample2. So, I divided them into 3 groups: Proteins1(only sample1), Proteins2 (only sample2), Proteins3(all samples). In this way, we can find the most regulated proteins separate and combined simultaneously.

2. In Protein3 we apply the following function on the table:

```
Proteins3[,1:6] <- Proteins3[,1:6] + matrix(rep(rnorm(nrow(Proteins),0,0.1),6),ncol=6)
```

3. Then, we make three **boxplots** for each group and draw **histogram** for the first four samples where the figures should be informative so we change the number of bins (*breaks=50*).



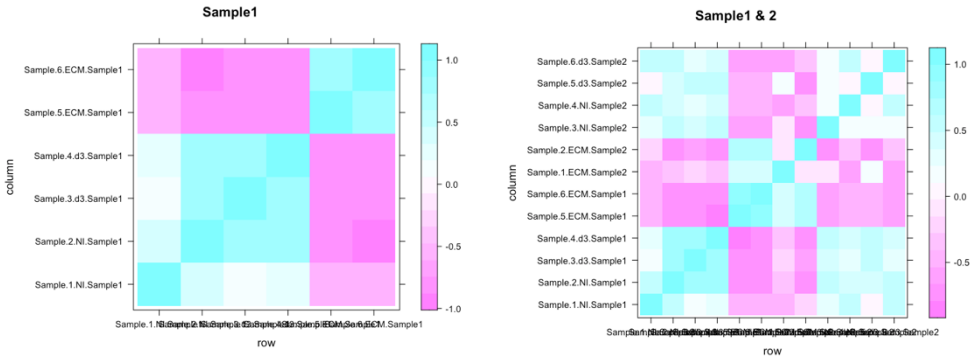
By viewing the boxplot, it is easy to find the ***Sample.2.ECM.Sample2*** has a wide range data, which has a maximum over 3, minimum lower than -2 and the third quartile is approximate to 1. This is also the reason I want to test the group separately; I think we may get different outcomes if we ignore this sample. Here, we further check the data from first four samples, luckily each of them shows a very great distribution.

4. Next step is **normalized** samples (columns) by their median and check their normalization with a boxplot. Normalization typically seeks to produce data with a mean of 0 and a standard deviation of 1, so we transform each data point x_i by $z_i = (x_i - \bar{x})/s$. This will guarantee that the transformed z variable has a mean of 0 and standard deviation of 1, except for (maybe) some numerical issues on a computer. However, perhaps we want our transformed variable to have a median of 0 and median deviation of 1. The above formula will not guarantee that, but we can follow the same idea by subtracting the median of X and then dividing by the median deviation.

5. To check the dataset weather it is complete, we **filtered** these table to guarantee each row have at least 6 values/protein (rowSums(!is.na(NormalizedData3))>5). After filtering Protein1 and Protein2 both remain 256 proteins while Protein3 still have 324 rows.

6. Last, we calculate **Pearson's correlations** (*cor* function) between all samples in each group and plot all of them at once with the *levelplot* function. Here, we set use = "na.or.complete", method = "pearson", because in *cor* function:

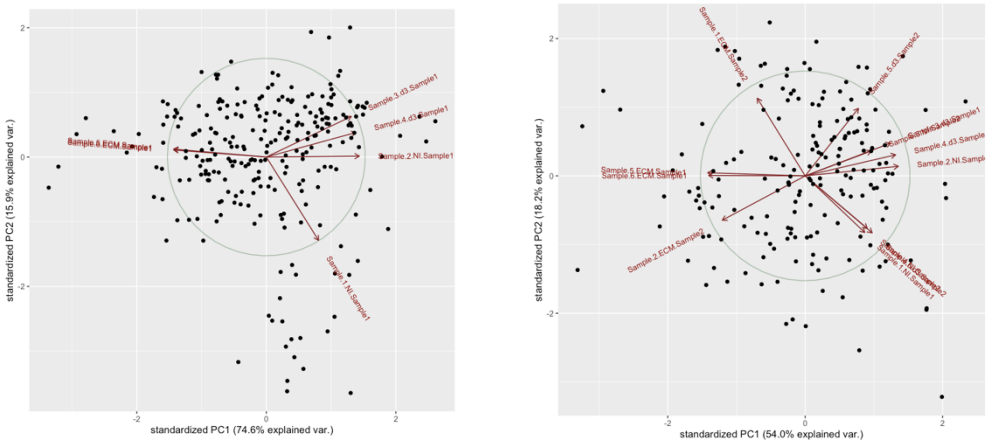
   i) If use is "everything", NAs will propagate conceptually, i.e., a resulting value will be NA whenever one of its contributing observations is NA.
   ii) If use is "all.obs", then the presence of missing observations will produce an error.
   iii) If use is "complete.obs" then missing values are handled by casewise deletion (and if there are no complete cases, that gives an error). "na.or.complete" is the same unless there are no complete cases, that gives NA.



According to these figures, we can find higher correlations between replicates and d3~NI samples.

PCA analysis

PCA forms the basis of multivariate data analysis based on projection methods. The most important use of PCA is to represent a multivariate data table as smaller set of variables (summary indices) in order to observe trends, jumps, clusters and outliers. This overview may uncover the relationships between observations and variables, and among the variables.



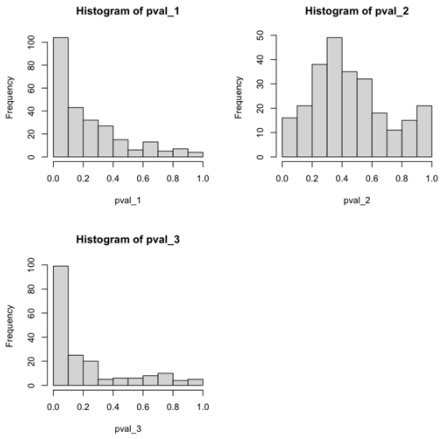
(Left is from sample1; right contains 1&2 samples)

By using the PCA analysis of these samples, we can find the data samples can be separated into two groups: the ECM samples and NI/d3 samples. We can also find in the sample1 data is highly explained by PC1 with 74.6%, and for the whole sample sets the PC1 can still explained 54%.

T-test

We further apply a t-test for each protein to see whether it is significantly changed between the conditions *NI* and *ECM* (use *for* loop and *apply* t.test(conditionA, conditionB)$p.value). Here, we also get p-values for 3 groups for comparative analysis.

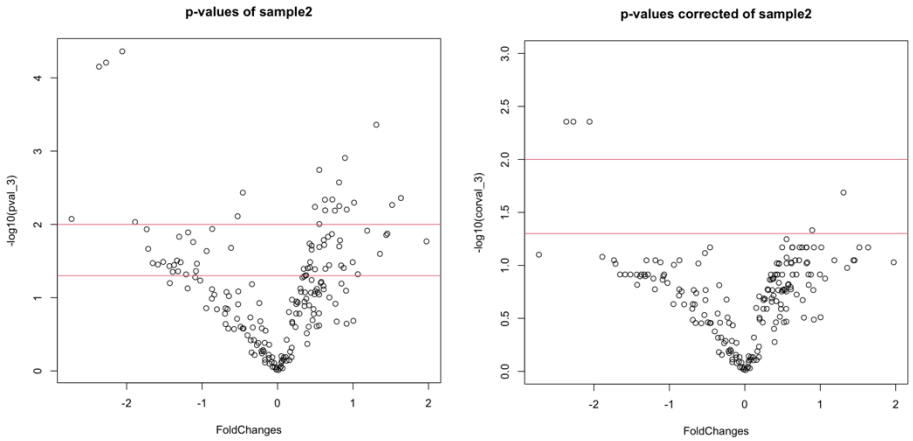
Then, we plot the distribution of p-values. We can notice the sample1 have lots of differential expression proteins since they have lots of p-value below 0.1, we have 57 and 12 proteins show a p-value below 0.05 and 0.01 respectively. If we look at sample2, it is easy to notice only a few p-values (5 below 0.05) is smaller which indicate the proteins do not show a lot of differential expression.



FDR & volcano plot

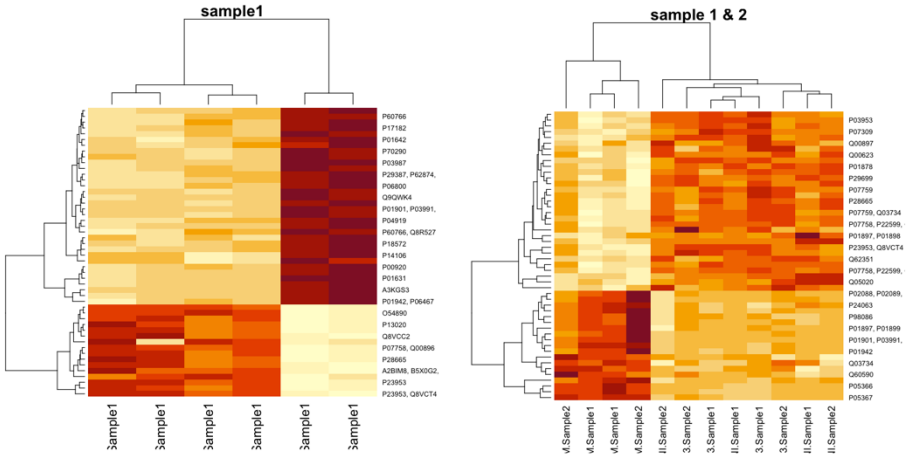
Here, we showed the false discovery rates (p-values corrected by Benjamini-Hochberg) in volcano plot (they are shown as *-log10(FDRs)*). The calculation of the fold changes is given by (FoldChanges <- rowMeans(group1,na.rm = T) - rowMeans(group2,na.rm = T)).

For the Protein3(contain all samples), there 70 and 22 proteins' p-value are below 0.05 and 0.01 respectively, after corrected by BH it still remains 5 and 3 proteins' p-value are below 0.05 and 0.01 respectively. So, these proteins can be regard as regulated proteins, but we cannot find any protein in Protein1 and Protein2 after corrected by BH. I think when correct the p-value with small dataset the correction have a larger impact on the p-values, so we can not find any in sample1.



Heatmap

In this step, we apply hierarchical clustering of the 50 most "regulated" proteins using the heatmap function.



In these heatmaps, they show a clear distinct between ECM and NI/d3 samples, where the red groups are representing the ECM samples.

As we expected, the top 50 proteins in each group are quite different:

Protein1: 'P01631', 'P01897', 'P06330', 'Q9QWK4', 'P00920', 'P03987', 'P05367', 'P08226', 'P17182', 'P29387', 'A2BIM8', 'A3KGS3', 'O35963', 'O54890', 'P01642', 'P01872', 'P01887', 'P01897', 'P01901', 'P01942', 'P01942', 'P02088', 'P02089', 'P06800', 'P14106', 'P15508', 'P18572', 'P23953', 'P28665', 'P60766', 'P62880', 'P70290', 'Q61735', 'Q8VCC2', 'P01898', 'P04919', 'P07758', 'P07758', 'P13020', 'P13634', 'P17156', 'P60766', 'Q06770', 'Q921I1', 'P23953', 'Q00724', 'P23953', 'P05366', 'P98086', 'P01630'.

Protein2: 'A2BIM8', 'A2BIM8', 'A2BIM8', 'A6X935', 'B2RSH2', 'C0HKE1', 'E9PV24', 'E9Q414', 'O35963', 'O54890', 'P01629', 'P01635', 'P01636', 'P01654', 'P01658', 'P01664', 'P01756', 'P01796', 'P01796', 'P01806', 'P01837', 'P01863', 'P01863', 'P01867', 'P01868', 'P01872', 'P01878', 'P01887', 'P01895', 'P01897', 'P01897', 'P01897', 'P01897', 'P01898', 'P01901', 'P01901', 'P01942', 'P01942', 'P02088', 'P02088', 'P02089', 'P02535', 'P03953', 'P03991', 'P03995', 'P04104', 'P04104', 'P04104', 'P04186', 'P04918'.

Protein3: 'P01901', 'P01901', 'P01942', 'P42703', 'P04104', 'P06909', 'A2BIM8', 'P01654', 'P01895', 'P02088', 'P02535', 'P03995', 'P04104', 'P07759', 'P0DP26', 'P11276', 'P11835', 'P22599', 'P28665', 'P01867', 'P01796', 'P11679', 'O35963', 'P01635', 'P01756', 'P04186', 'P05064', 'P07146', 'P08226', 'P08730', 'P26041', 'P04104', 'P08752', 'P14211', 'P24063', 'P04918', 'P04919', 'P04945', 'P22599', 'P22599', 'P23953', 'P01837', 'B2RSH2', 'P21614', 'P13634', 'P01796', 'P01863', 'P01868', 'P01872', 'P01878'.

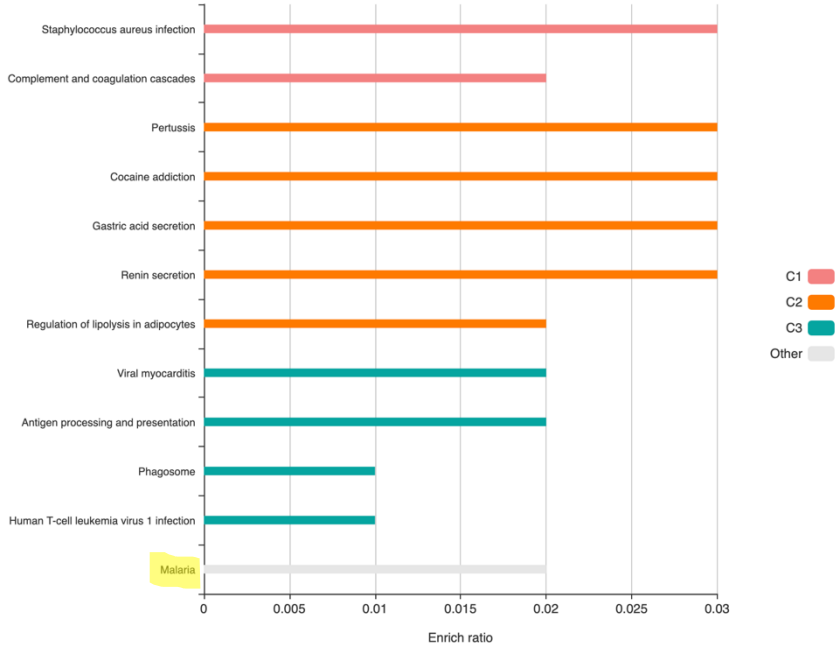
After filter the repeated names, we can find there are only 6 proteins present in all three samples. There are: 'A2BIM8', 'O35963', 'P01872', 'P01901', 'P01942', 'P02088'.

GO analysis

We select the most regulated proteins from all samples and copy their accession numbers into the tool Kobas.

Enriched terms visualized in barplot

Each row represents an enriched function, and the length of the bar represents the enrich ratio，which is calculated as "input gene number"/ "background gene number". The color of the bar is the same as the color in the circular network in above, which represents different clusters. For each cluster, if there are more than 5 terms, top 5 with the highest enrich ratio will be displayed.



The top-10 most represented GO terms are: extracellular space, keratin filament, intermediate filament, acute-phase response, extracellular region, intermediate filament cytoskeleton, complement component C3b binding, protease binding, integrin binding, hemoglobin binding. The GO terms showed that the proteins identified from murine plasma MP have important active roles, being highly significantly associated with binding and regulatory activities.

The top-10 most represented KEGG terms are: Estrogen signaling pathway, Pertussis, Human immunodeficiency virus 1 infection, Leukocyte transendothelial migration, Human cytomegalovirus infection, Gastric acid secretion, Staphylococcus aureus infection, Renin secretion, Complement and coagulation cascades, Rap1 signaling pathway.

Reference

[1] Thoß, M., Luzynski, K.C., Enk, V.M., Razzazi-Fazeli, E., Kwak, J., Ortner, I., and Penn, D.J. (2019). Regulation of volatile and non-volatile pheromone attractants depends upon male social status. Scientific Reports *9*, 489. 10.1038/s41598-018-36887-y.