# Report2

The method of analysis of variance given in the paper is ANOVA, where it is stated that ANOVA is done for each gene at any two time points, and then the genes are screened for FDR < 0.05. However, the problem is that there is only one data for any one gene at any one time point, which means that there are only two values that can be used as a source of data for ANOVA. This is obviously unreasonable.

The method I use here is the DESeq2 method in the R package, which assumes that the data conform to a negative binomial distribution. This assumption is consistent with the facts in two ways. The first point is that, like the Poisson distribution, the negative binomial distribution is a counting distribution, which is consistent with the fact that the number of protein expressions is calculated. The second point is that, for the Poisson distribution, the mean and variance are equal, whereas the distribution of the real data deviates from the Poisson distribution, and the variance is significantly larger than the mean. In addition, unlike the normal distribution test, which requires only one number to calculate the variance, the test in DESeq2 is usually the WALD test. The threshold condition I set here is that the log value of the expression multiple is greater than 1 and padj is less than 0.05. Unlike the commonly used p-value, padj is a correction for p-value and is based on multiple hypothesis testing to better control the false positive rate. Finally, 535 groups of genes were selected.

For comparison with the results of the article, an FDR 0.01 screen was performed here for the ANOVA pvalue of each protein. The final result was 1,794 proteins. This is somewhat different from the 1,712 proteins screened at FDR 0.05 in the paper. I then cross-referenced the filtered proteins in the paper with the ANOVA pvalue filtered proteins to obtain a Jaccard Index of 0.493. Overall, the different filtering strategies resulted in different proteins being filtered.