# cblb_workflow

*Farhad Shakeri*

*6/25/2018*

## Overlook

Generally, the standard workflow in proteomics is as foloows:

- **Pre-processing the raw data**
  - Reformatting & cleanning
  - Protein counts, uniqueness, repeated measurements
  - Dealing with missing values
  - Quality control

- **Data Summarization**
  The raw data are in `Feature` level. A `Feature` is defined as the combination of `Protein + Peptide + Charge`. For significance analysis, the abundance should be rolled-up (or summarized) to `Protein` level. There are various methods to do that like simple averaging, linear model, sum and etc. I use ***Tukey's median polish***, which is a robust averaging method, resistant against outliers. The median polish is applied to every `Protein` over all `Channels` and `Mixtures`. Note that the `Fractions` (or *Runs*) belonging to each *biological* `Mixture` should be combined prior to this step. In practice, for each `Protein` we have a matrix with `Feature Abundances` vs. `Channels` and `Mixtures`.

- **Significance Analysis**
  Finding differentially abundant proteins across conditions. Here I use moderated *t-test* from `Limma` Package. The outcome of this step is a list of *p values* and *log-fold-changes*, which will be used to select the top hit proteins and generate volcano plots.

## Workflow

- **Pre-processing the raw data**

1. Removing shared PSMs between protein-groups.
   The assumption is that each `PSM` should belong to only one `Protein`. The column `#.Protein.Groups` from *Proteome Discoverer* output file provides information on the number of `Proteins` to which belongs a specific `PSM`. As the first filter, I only use the rows in which the `#.Protein.Groups` is equal to 1.

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

1. Removing shared peptides between protein-groups. Using only unique peptides.
   > Raw data: #Protein 6068 > Raw data: #Protein 6068

## Pre-processing steps

1. Removing shared peptides between protein-groups. Using only unique peptides.

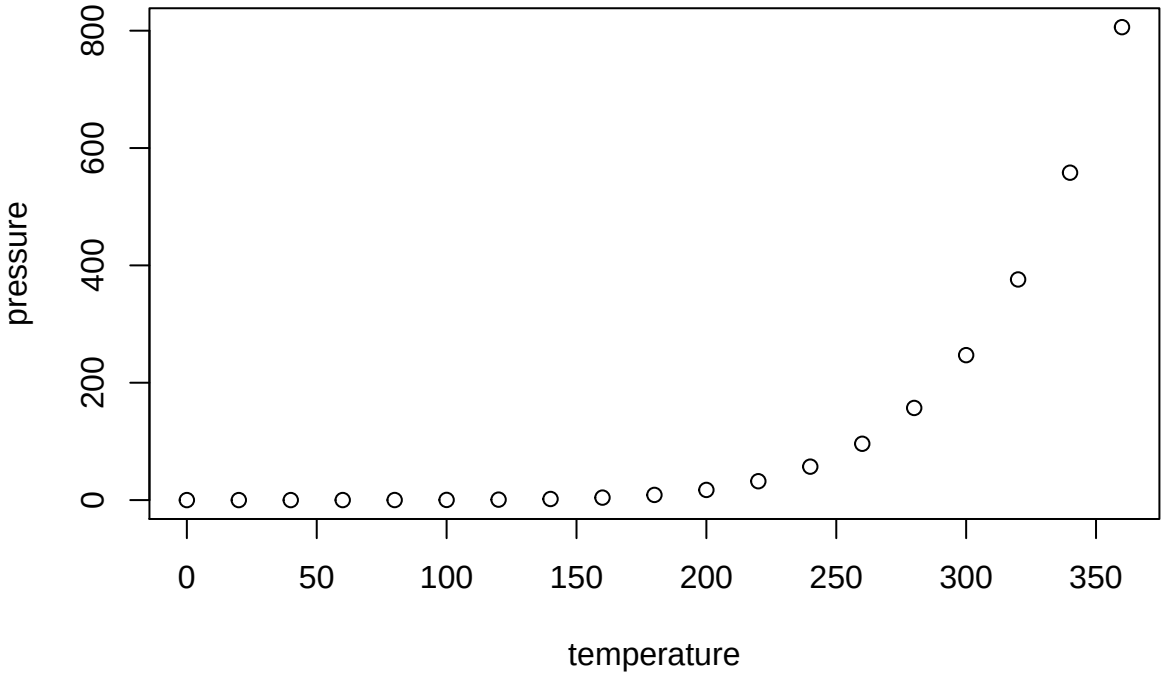
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

## Including Plots

You can also embed plots, for example:




Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.