Portable Proteomics Pipeline (P3)
Spectrum Count Quantification

Daniel Kristiyanto (daniel.kristiyanto@pnnl.gov)

April 10, 2016

Portable Proteomics Pipeline (P3) is a series of protein mass spectrometry data pre-processing pipelines in Docker containers. The packages includes protein identification, filtering, and quantification for both labeled and label-free mass spectrometry data.

This document discuss the `kristiyanto/p3:scquant` container, a spectrum count quantification pipeline for label-free mass spectrometry protein quantification. The pipeline is based on MSnID pipeline [1].

Input

The container takes the following files:

```
p3.config # Configuration file
*.mzid    # mzIdentML files resulted from p3:msgf container or other identification tools.
```

`p3.config` may contain various parameters for p3 related containers. To run spectrum count quantification, `p3.config` must contain the following information:

```
[spectrum_count]
score_treshold = 7.0  # The scoring treshold
error_treshold = 20   # Error Treshold
fdr = 0.01            # False Discovery Rate
iteration = 5000      # Number of iteration
```

Running the Container

to run the container, a docker engine must be installed. A more information about installing Docker engine is available at https://docs.docker.com/engine/installation/. Input files must be mounted to `/root/data` within the container. This can be done by using `-v` switch. For MacOS and Windows users, the folder should be located under `C:\Users` or `/Users/`. More information about volumens in Docker containers is available at http://container-solutions.com/understanding-volumes-docker/

```
# Download/update the container from DockerHub
docker pull kristiyanto/p3:squant
# Run the container
docker run --rm -v /Users/path/files:/root/data kristiyanto/p3:scquant
```

Output

Once the quantification process is completed, `LabelFreeQuant.txt` and `msnset.rda` are generated. `LabelFreeQuant.txt` is a tab delimited file with the quantification results, with each column represented from each of the `mzid` file provided. `msnset.rda` is an msnset object for the result that can be easily imported to R for additional analysis.

Pipeline

`kristiyanto/p3:scquant` is based on R, and it uses MSnID package and pipeline. A more detailed information about the pipeline is available at http://bioconductor.org/packages/release/bioc/vignettes/MSnID/inst/doc/msnid_vignette.pdf.

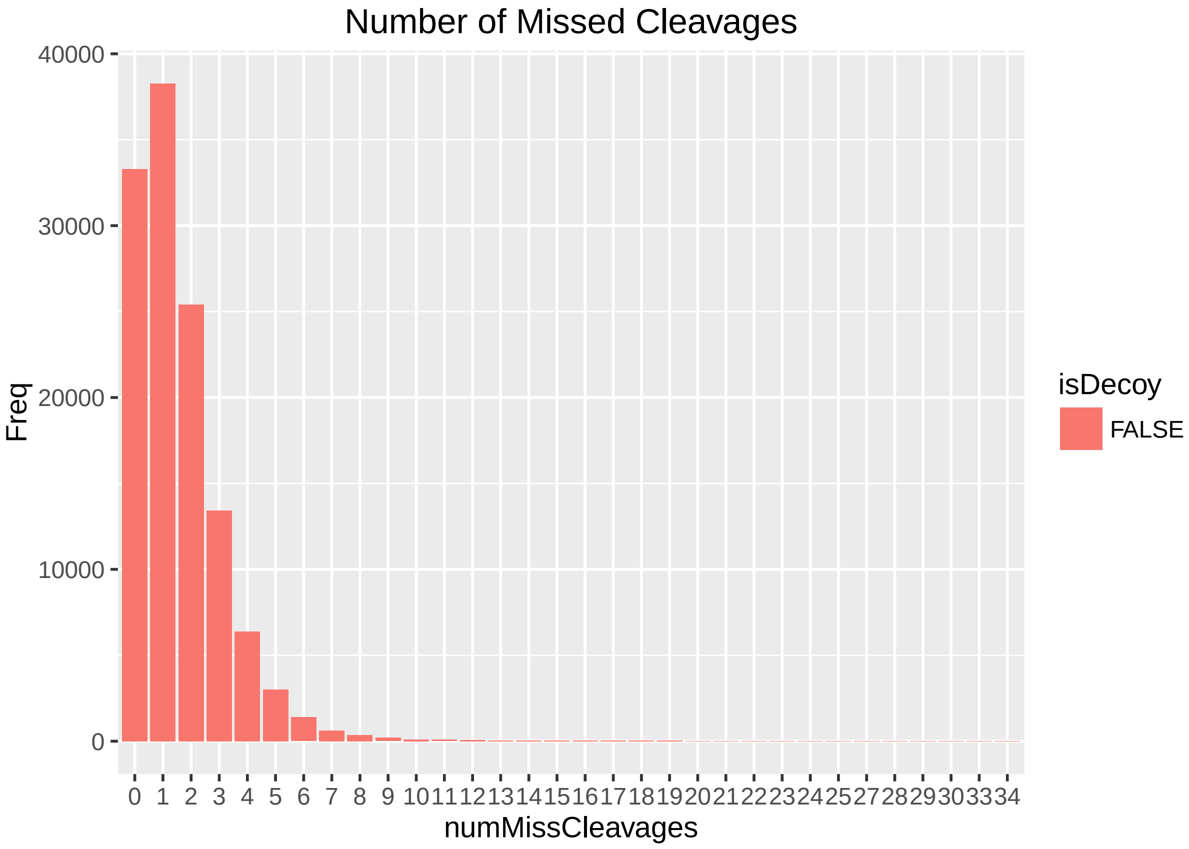
For this documentation, three `mzid` files are processed.

```
# Files
print(mz.files)
```

```
## [1] "Biodiversity_mix_01_28Oct15_Arwen_15-07-13.mzid"
## [2] "Biodiversity_mix_02_28Oct15_Arwen_15-07-13.mzid"
## [3] "Biodiversity_mix_03_28Oct15_Arwen_15-07-13.mzid"
```

```
prop.table(table(mzid$numIrregCleavages))
```

```
##
##            0            1            2
## 0.9668602085 0.0327185017 0.0004212899
```



Filter

Filtering is done by using the parameter defined in the `p3.config` file. The filtering is including the error calculation (`variable: error_treshold`), scoring (`variable: score_treshold`), and filter optimization (`variable: iteration and fdr`).

```
show(mzid)
```

```
## MSnID object
## Working directory: "."
## #Spectrum Files:  3
## #PSMs: 795177 at 0 % FDR
## #peptides: 122960 at 0 % FDR
## #accessions: 167543 at 0 % FDR
```

```
show(fObj)
```

```
## MSnIDFilter object
## (msmsScore > 7) & (massError < 20)
```

```
evaluate_filter(prj, fObj, level="PSM")
```

```
##     fdr      n
## PSM   0 589592
```

```
evaluate_filter(prj, fObj, level="peptide")
```

```
##         fdr     n
## peptide   0 49166
```

```
evaluate_filter(prj, fObj, level="accession")
```

```
##           fdr      n
## accession   0 102866
```

```
show(prj)
```

```
## MSnID object
## Working directory: "."
## #Spectrum Files:  3
## #PSMs: 622994 at 0 % FDR
## #peptides: 66562 at 0 % FDR
## #accessions: 121791 at 0 % FDR
```

Quantification

Quantification is done by using the MSnBase package [2].

```
msnset  <- as(prj, "MSnSet")
msnset  <- combineFeatures(msnset, fData(msnset)$accession, redundancy.handler="unique",
                           fun="sum",cv=FALSE)
```

```
## Combined 46483 features into 39394 using sum
```

Results

```
exprs.table     <- exprs(msnset)
exprs.table     <- cbind(Protein=row.names(exprs.table), as.data.frame(exprs.table))
head(exprs.table)
```

```
##                                    Protein
## Contaminant_Trypa5     Contaminant_Trypa5
## sp|A0A321|ATPF_COFAR sp|A0A321|ATPF_COFAR
## sp|A0A348|CYF_COFAR   sp|A0A348|CYF_COFAR
## sp|A0A393|YCF1_COFAR sp|A0A393|YCF1_COFAR
## sp|A0A4Z3|A3LT2_RAT   sp|A0A4Z3|A3LT2_RAT
## sp|A0AAS4|SMS1_PIG     sp|A0AAS4|SMS1_PIG
##                      Biodiversity_mix_01_28Oct15_Arwen_15-07-13.mzXML
## Contaminant_Trypa5                                                  2
## sp|A0A321|ATPF_COFAR                                                0
## sp|A0A348|CYF_COFAR                                                 0
## sp|A0A393|YCF1_COFAR                                                1
## sp|A0A4Z3|A3LT2_RAT                                                 0
## sp|A0AAS4|SMS1_PIG                                                  1
##                      Biodiversity_mix_02_28Oct15_Arwen_15-07-13.mzXML
## Contaminant_Trypa5                                                  1
## sp|A0A321|ATPF_COFAR                                                0
## sp|A0A348|CYF_COFAR                                                 1
## sp|A0A393|YCF1_COFAR                                                3
## sp|A0A4Z3|A3LT2_RAT                                                 0
## sp|A0AAS4|SMS1_PIG                                                  0
##                      Biodiversity_mix_03_28Oct15_Arwen_15-07-13.mzXML
## Contaminant_Trypa5                                                  1
## sp|A0A321|ATPF_COFAR                                                1
## sp|A0A348|CYF_COFAR                                                 1
## sp|A0A393|YCF1_COFAR                                                2
## sp|A0A4Z3|A3LT2_RAT                                                 1
## sp|A0AAS4|SMS1_PIG                                                  0
```

References

1. Laurent Gatto VP with contributions from: *MSnID: Utilities for Exploration and Assessment of Confidence of LC-MSn Proteomics Identifications.*

2. Gatto L, Lilley K: **MSnbase - an r/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation**. *Bioinformatics* 2012, **28**:288–289.