Portable Proteomics Pipeline (P3)
Labelled (iTRAQ4) Quantification

Daniel Kristiyanto (daniel.kristiyanto@pnnl.gov)

April 12, 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:itraquant` container, a spectrum count quantification pipeline for label-free mass spectrometry protein quantification. The pipeline is based on MSNbase pipeline [4].

Input

The container takes the following files:

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

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

```
[itraq4]
evalue_treshold = 1e-10 # Evalue treshold. Features with evalue higher than
                # this value will be discarded.
                # Check MSNbase: removeNoID documentation for more detailed information.
pNA    = 0      # 0 to 1. Ratio of NA allowed for a feature.
                # 0 indicates that features with any missing value will be discarded.
                # Check MSNbase: filterNA documentation for more detailed information.
quant_method = sum      # Quantification method.
                # Check MSNbase: quantify documentation for more detailed information.
combine_by = mean       # Function used for feature aggregation.
                # Check MSNbase: combineFeatures for more detailed documentation for more detai
```

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:itraquant
# Run the container
docker run --rm -v /Users/path/files:/root/data kristiyanto/p3:itraquant
```

Output

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

Pipeline

`kristiyanto/p3:scquant` is based on R, and it uses MSNbase [4] 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 paired of experiment files are processed.

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

```
## [1] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzid"
## [2] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f02.mzid"
## [3] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f03.mzid"
```

```
print(mzml.files)
```

```
## [1] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzML"
## [2] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f02.mzML"
## [3] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f03.mzML"
```

Identification

Identification is performed by using `addIdentificationData()` function from MSnBase package.

```
idSummary(msexp.id)
```

```
##                                              spectrumFile
## 1 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzML
## 2 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f02.mzML
## 3 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f03.mzML
##                                                    idFile coverage
## 1 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzid    0.977
## 2 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f02.mzid    0.976
## 3 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f03.mzid    0.978
```

```
fData(msexp.id)[1:3,]
```

```
##      spectrum scan number(s) passthreshold rank calculatedmasstocharge
## X1.1        1             48          TRUE    1               329.2057
## X1.2        2             48          TRUE    1               323.6872
## X1.3        3             48          TRUE    1               379.2012
##      experimentalmasstocharge chargestate ms-gf:denovoscore ms-gf:evalue
## X1.1                 329.4545           4               117     5.595577
## X1.2                 323.9427           4               118     1.038507
## X1.3                 379.4494           4                93     4.714499
##      ms-gf:rawscore ms-gf:specevalue assumeddissociationmethod
## X1.1             33     5.147427e-07                       HCD
## X1.2             48     9.608537e-08                       HCD
## X1.3              9     4.315219e-07                       HCD
##      isotopeerror isdecoy post pre  end start       accession length
## X1.1            1   FALSE    R   R   44    34 ref|NP_775799.2    301
## X1.2            1   FALSE    H   R 1120  1111 ref|NP_075463.2   3013
## X1.3            1   FALSE    S   R  223   212 ref|NP_057257.1    338
##                                                             description
## X1.1       hypothetical protein LOC161502, gi|148747373 [Homo sapiens]
## X1.2                protein furry homolog, gi|117606355 [Homo sapiens]
## X1.3 hemK methyltransferase family member 1, gi|7705409 [Homo sapiens]
##            pepseq modified modification
## X1.1  DKGKLLIQRSR    FALSE         <NA>
## X1.2   RFLFPQQSLR    FALSE         <NA>
## X1.3 IWIIHLDMTSER    FALSE         <NA>
##                                                      idFile
## X1.1 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzid
## X1.2 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f02.mzid
## X1.3 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f03.mzid
##                  databaseFile nprot npep.prot npsm.prot npsm.pep
## X1.1 ID_003632_9011437E.fasta     1         1         2        1
## X1.2 ID_003632_9011437E.fasta     1         2         7        2
## X1.3 ID_003632_9011437E.fasta     1         1         2        1
```

Filter

Filtering is done by: 1) Removing unidentified features, and features with e-value higher than `evalue_treshold` value described in `p3.config`. 2) features that are asssigned by multiple protein groups.

```
# Prior Filtering
length(msexp.id)
```

```
## [1] 47545
```

```
print(evalue_treshold)
```

```
## [1] 1e-10
```

```
k                 <- (fData(msexp.id)$'ms-gf:evalue'< evalue_treshold)
k[is.na(k)]       <- FALSE
print(sum(k)) # Number of kept features
```

```
## [1] 48
```

```
msexp.filter1     <- removeNoId(msexp.id, keep=k)
# Subsquent evalue filter
length(msexp.filter1)
```

```
## [1] 48
```

```
msexp.filter2     <- removeMultipleAssignment(msexp.filter1)
# Subsquent multiple assigment filter
length(msexp.filter2)
```

```
## [1] 28
```

```
fData(msexp.filter2)[1:3,]
```

```
##          spectrum scan number(s) passthreshold rank calculatedmasstocharge
## X10257.2      866          11331          TRUE    1               886.9222
## X11009.2     3374          12159          TRUE    1               855.1025
## X11424.2     4757          12615          TRUE    1               914.4767
##          experimentalmasstocharge chargestate ms-gf:denovoscore
## X10257.2                 887.4300           2               182
## X11009.2                 855.4400           3               176
## X11424.2                 914.4795           2               137
##          ms-gf:evalue ms-gf:rawscore ms-gf:specevalue
## X10257.2 5.779398e-12            177     5.227042e-19
## X11009.2 9.363158e-17            159     8.292638e-24
## X11424.2 2.077475e-13            129     1.872784e-20
##          assumeddissociationmethod isotopeerror isdecoy post pre end start
## X10257.2                       HCD            1   FALSE    E   R 326   312
## X11009.2                       HCD            1   FALSE    N   K 125   103
## X11424.2                       HCD            0   FALSE    G   K 145   130
##                accession length
## X10257.2 ref|NP_006816.2    602
## X11009.2 ref|NP_005902.1    395
## X11424.2 ref|NP_000916.2    359
##
## X10257.2                                                    cytoskeleton-associated protein
## X11009.2                                          S-adenosylmethionine synthase isoform type
## X11424.2 pyruvate dehydrogenase E1 component subunit beta, mitochondrial isoform 1 precursor
##                             pepseq modified     modification
## X10257.2         STLQTMESDIYTEVR    FALSE             <NA>
## X11009.2 TCNVLVALEQQSPDIAQGVHLDR     TRUE 57.021463735 (2)
## X11424.2        TYYMSGGLQPVPIVFR    FALSE             <NA>
##                                                          idFile
## X10257.2 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f02.mzid
## X11009.2 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f02.mzid
## X11424.2 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f02.mzid
##                      databaseFile nprot npep.prot npsm.prot npsm.pep
## X10257.2 ID_003632_9011437E.fasta     1         1         3        1
## X11009.2 ID_003632_9011437E.fasta     1         1         1        1
## X11424.2 ID_003632_9011437E.fasta     1         1         1        1
```

Quantification

Quantification is done by using the MSnBase package [4], with method specified on the `p3.config` file.

```
head(exprs(qnt))
```

```
##          iTRAQ4.114 iTRAQ4.115 iTRAQ4.116 iTRAQ4.117
## X10257.2         NA         NA   97508.10         NA
## X11009.2   25410.31   25410.31   25410.31   25410.31
## X11424.2   27149.91   27149.91   27149.91   27149.91
## X12133.1   43594.16   43594.16   43594.16   43594.16
## X12133.3   24688.00   24688.00   24688.00   24688.00
## X12294.2   12488.43   12488.43   12488.43   12488.43
```

```
print(pNA)
```

```
## [1] 0
```

```
qnt.filtered      <- filterNA(qnt, pNA = pNA)
head(exprs(qnt.filtered))
```

```
##           iTRAQ4.114 iTRAQ4.115 iTRAQ4.116 iTRAQ4.117
## X11009.2    25410.31   25410.31   25410.31   25410.31
## X11424.2    27149.91   27149.91   27149.91   27149.91
## X12133.1    43594.16   43594.16   43594.16   43594.16
## X12133.3    24688.00   24688.00   24688.00   24688.00
## X12294.2    12488.43   12488.43   12488.43   12488.43
## X12353.1    13771.04   13771.04   13771.04   13771.04
```

Results

subsequently, the features are aggregated by accession ID, with the value calculated using the function denoted on the `p3.config`.

```
length(exprs(qnt.filtered))
```

```
## [1] 92
```

```
result  <- combineFeatures(qnt.filtered, groupBy = fData(qnt.filtered)$accession, fun=combine_
```

```
## Combined 23 features into 17 using mean
```

```
length(exprs(result))
```

```
## [1] 68
```

```
head(exprs(result))
```

```
##                     iTRAQ4.114 iTRAQ4.115 iTRAQ4.116 iTRAQ4.117
## ref|NP_000843.1      43594.156  43594.156  43594.156  43594.156
## ref|NP_000916.2      27149.912  27149.912  27149.912  27149.912
## ref|NP_001073027.1   47457.660  47457.660  47457.660  47457.660
## ref|NP_001348.2      29536.990  29536.990  29536.990  29536.990
## ref|NP_001531.1      11439.398  11439.398  11439.398  11439.398
## ref|NP_001677.2       7962.812   7962.812   7962.812   7962.812
```

References

1. Domon B, Aebersold R: **Mass spectrometry and protein analysis**. *science* 2006, **312**:212–217.

2. Deutsch EW: **Mass spectrometer output file format mzML**. *Proteome Bioinformatics* 2010:319–331.

3. Gatto L, Gibb S: **MSnbase: Labelled and label-free mS2 data pre-processing, visualisation and quantification.** 2016.

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