Portable Proteomics Pipeline (P3) Labelled (iTRAQ4) Benchmark

Daniel Kristiyanto (daniel.kristiyanto@pnnl.gov)

May 3, 2016

The purpose of this document is to benchmark the iTRAQ quantification on the Portable Proteomics Pipelines (P3) containers.

Input

1. TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzid : The identification file (product of MSGF+)
2. TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzML : The mass spectometry dataset
3. TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01_ReporterIons.txt : The quantification results used to benchmark (calcullated by PNNL)

Methods

Identification

```
## The identification file
mzids.raw
```

```
## An mzID object
##
## Software used:   MS-GF+ (version: Beta (v10089))
##
## Rawfile:         /root/data/TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzML
##
## Database:        /root/data/ID_003632_9011437E.fasta
##
## Number of scans: 15696
## Number of PSM's: 16978
```

```
head(mzid.flat[order(mzid.flat[,"ms-gf:evalue"]),])[1:3,]
```

```
##                                      spectrumid scan number(s)
```

```
## 1 controllerType=0 controllerNumber=1 scan=17146          17146
## 2 controllerType=0 controllerNumber=1 scan=13662          13662
## 3 controllerType=0 controllerNumber=1 scan=14281          14281
##   acquisitionnum passthreshold rank calculatedmasstocharge
## 1          17146          TRUE    1               943.4951
## 2          13662          TRUE    1               727.1045
## 3          14281          TRUE    1               936.7944
##   experimentalmasstocharge chargestate ms-gf:denovoscore ms-gf:evalue
## 1                 943.8400           3               233 7.599588e-24
## 2                 727.0984           4               235 7.777001e-22
## 3                 936.7979           3               188 3.604979e-20
##   ms-gf:rawscore ms-gf:specevalue assumeddissociationmethod isotopeerror
## 1            227     6.677701e-31                      HCD            1
## 2            229     6.859839e-29                      HCD            0
## 3            180     3.179835e-27                      HCD            0
##   isdecoy post pre end start          accession length
## 1   FALSE    A   K 993   967    ref|NP_001840.3   1019
## 2   FALSE    C   R 494   470    ref|NP_004915.2    911
## 3   FALSE    C   R 475   451 ref|NP_001123476.1    914
##                                                                    description
## 1 collagen alpha-2(VI) chain isoform 2C2 precursor, gi|115527062 [Homo sapiens]
## 2                                  alpha-actinin-4, gi|12025678 [Homo sapiens]
## 3                       alpha-actinin-1 isoform a, gi|194097350 [Homo sapiens]
##                       pepseq modified modification
## 1 QNVVPTVLALGSDVDMDVLTTLSLGDR    FALSE         <NA>
## 2   VEQIAAIAQELNELDYYDSHNVNTR    FALSE         <NA>
## 3   VEQIAAIAQELNELDYYDSPSVNAR    FALSE         <NA>
##                                                idFile
## 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_f01.mzid
## 3 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzid
##                                             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_f01.mzML
## 3 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzML
##              databaseFile
## 1 ID_003632_9011437E.fasta
## 2 ID_003632_9011437E.fasta
## 3 ID_003632_9011437E.fasta
```

Based on this data, peptide QNVVPTVLALGSDVDMDVLTTLSLGDR was listed as a good quality of identification with lowest e-value.

Spectrum Objects

```
msexp.raw
```

```
## Object of class "MSnExp"
##  Object size in memory: 111.6 Mb
## - - - Spectra data - - -
##  MS level(s): 2
##  Number of MS1 acquisitions: 1606
##  Number of MSn scans: 16058
##  Number of precursor ions: 16058
##  14573 unique MZs
##  Precursor MZ's: 300.18 - 1732.7
##  MSn M/Z range: 100 1999.87
##  MSn retention times: 1:2 - 98:59 minutes
## - - - Processing information - - -
## Data loaded: Tue May  3 16:38:00 2016
##  MSnbase version: 1.16.2
## - - - Meta data  - - -
## phenoData
##   rowNames: 1
##   varLabels: sampleNames
##   varMetadata: labelDescription
## Loaded from:
##   TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzML
## protocolData: none
## featureData
##   featureNames: X1.1 X10.1 ... X9999.1 (16058 total)
##   fvarLabels: spectrum
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
```

The spectrum data consists of 1606 percursor ions and with features ranged from X1.1 to x9999.1.

This data then paired with the identification file using `addIdentificationData` tool from MSnBase, resulting the table below:

```
### The spectra objects enrinched with identification file
msexp.id
```

```
## Object of class "MSnExp"
##  Object size in memory: 118.25 Mb
## - - - Spectra data - - -
##  MS level(s): 2
##  Number of MS1 acquisitions: 1606
##  Number of MSn scans: 16058
##  Number of precursor ions: 16058
##  14573 unique MZs
```

```
##  Precursor MZ's: 300.18 - 1732.7
##  MSn M/Z range: 100 1999.87
##  MSn retention times: 1:2 - 98:59 minutes
## - - - Processing information - - -
## Data loaded: Tue May  3 16:38:00 2016
##  MSnbase version: 1.16.2
## - - - Meta data  - - -
## phenoData
##   rowNames: 1
##   varLabels: sampleNames
##   varMetadata: labelDescription
## Loaded from:
##   TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzML
## protocolData: none
## featureData
##   featureNames: X1.1 X10.1 ... X9999.1 (16058 total)
##   fvarLabels: spectrum scan number(s) ... npsm.pep (30 total)
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
```

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

```
##                                                spectrumFile
## 1 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzML
##                                                      idFile coverage
## 1 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzid    0.977
```

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

Table focusing on the peptide QNVVPTVLALGSDVDMDVLTTLSLGDR:

```
zoomData <- fData(msexp.id)[fData(msexp.id)$pepseq==pep & !is.na(fData(msexp.id)$pepseq),]
zoomData
```

```
##          spectrum scan number(s) passthreshold rank calculatedmasstocharge
## X15545.1     6164          17146          TRUE    1               943.4951
##          experimentalmasstocharge chargestate ms-gf:denovoscore
## X15545.1                   943.84           3               233
##          ms-gf:evalue ms-gf:rawscore ms-gf:specevalue
## X15545.1 7.599588e-24            227     6.677701e-31
##          assumeddissociationmethod isotopeerror isdecoy post pre end start
## X15545.1                       HCD            1   FALSE    A   K 993   967
##                accession length
## X15545.1 ref|NP_001840.3   1019
##                                                                        description
## X15545.1 collagen alpha-2(VI) chain isoform 2C2 precursor, gi|115527062 [Homo sapiens]
```

```
##                                pepseq modified modification
## X15545.1 QNVVPTVLALGSDVDMDVLTTLSLGDR    FALSE         <NA>
##                                                     idFile
## X15545.1 TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f01.mzid
##                       databaseFile nprot npep.prot npsm.prot npsm.pep
## X15545.1 ID_003632_9011437E.fasta     1         1         2        1
```

From this table, we know that this peptide was identified from features X15545.1.

Using the `quantify` function from MSnBase, with parameters `mehod=max`, `verbose=FALSE`, `strict=FALSE`, resulting:

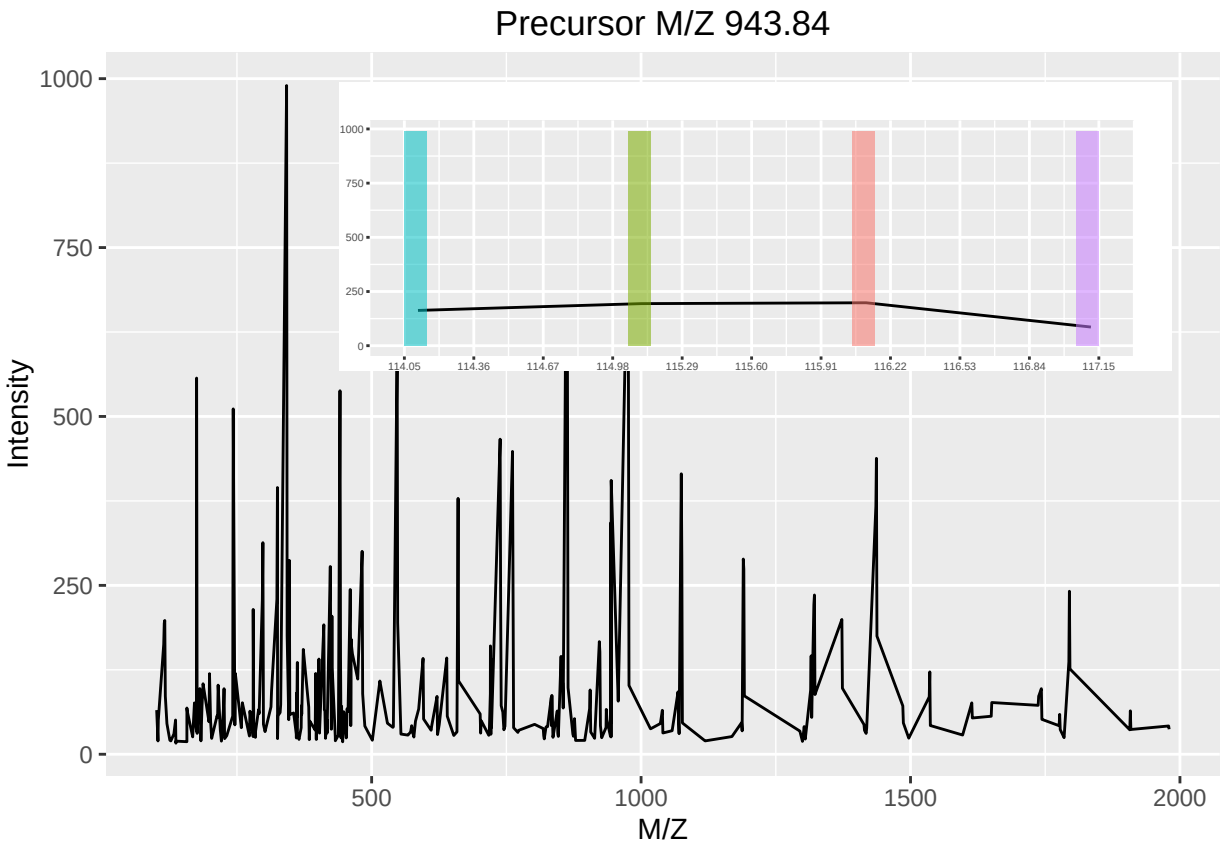
```
## Features
zoomData[,cols]
```

```
##                accession                          pepseq ms-gf:evalue
## X15545.1 ref|NP_001840.3 QNVVPTVLALGSDVDMDVLTTLSLGDR 7.599588e-24
```

```
exprs(qnt)[row.names(exprs(qnt))==row.names(zoomData),]
```

```
## iTRAQ4.114 iTRAQ4.115 iTRAQ4.116 iTRAQ4.117
##  162.69118  194.13533  197.98782   86.02748
```

```
plot(msexp.id[["X15545.1"]],reporters=iTRAQ4, full=T)
```



Quantification

This quantification then compared to the quantification data resulted from PNNL, by matching the scan number.

```
zoomData$`scan number(s)`
```

```
## [1] 17146
```

```
msexp.sp[msexp.sp$ScanNumber==zoomData$`scan number(s)`,]
```

```
##       Dataset ScanNumber Collision.Mode ParentIonMZ BasePeakIntensity
## 15545  322257      17146            hcd      943.84            989.79
##       BasePeakMZ ReporterIonIntensityMax Ion_114 Ion_115 Ion_116 Ion_117
## 15545   342.1776                  198.61  170.97  192.86  198.61   83.31
##       Weighted.Avg.Pct.Intensity.Correction
## 15545                                     2
```

Comparasion and error

The differences between the calculation by MSnBase and by PNNL:

```
zoomData$`scan number(s)`
```

```
## [1] 17146
```

```
from.msnbase  <- exprs(qnt)[row.names(zoomData),]
from.pnnl     <- msexp.sp[msexp.sp$ScanNumber==zoomData$`scan number(s)`,c("Ion_114", "Ion_115"
show(from.msnbase)
```

```
## iTRAQ4.114 iTRAQ4.115 iTRAQ4.116 iTRAQ4.117
##  162.69118  194.13533  197.98782   86.02748
```

```
show(from.pnnl)
```

```
##       Ion_114 Ion_115 Ion_116 Ion_117
## 15545  170.97  192.86  198.61   83.31
```

```
calc.error    <- from.msnbase - from.pnnl
calc.error
```

```
##         Ion_114 Ion_115    Ion_116  Ion_117
## 15545 -8.278823 1.27533 -0.6221765 2.717481
```

Repeating the same process for all of peptides with e-value < 50.

```
# Original data
dim(qnt.id)          # From MSXML File e-value < 50
```

```
## [1] 14279    47
```

```
dim(msexp.sp)        # From PNNL (all)
```

```
## [1] 16058    12
```

```
qnt.merged <- merge(qnt.id, msexp.sp, by.x = "scan number(s)", by.y = "ScanNumber", all.x = T,
dim(qnt.merged)      # Merged Data (left Join)
```

```
## [1] 14279    58
```

```
head(qnt.merged[,c(cols.sp, cols.p3)])
```

```
##   Ion_114 Ion_115 Ion_116 Ion_117 iTRAQ4.114 iTRAQ4.115 iTRAQ4.116
## 1 2909.30 3358.80 7211.01 3741.59  2769.9133  3491.8975  7006.5454
## 2 1554.30 2242.89 1474.75 1375.35  1488.8009  2207.5073  1546.3977
## 3  339.53  141.71 1515.66  556.20   318.2612   196.8605  1431.3364
## 4  325.58  584.60 1859.63  842.26   314.1523   615.4215  1785.3804
## 5  711.88  665.36 1191.37  745.25   674.6411   692.6157  1169.3190
## 6  982.32  245.76  296.35  440.15   917.4916   294.1249   307.1602
##   iTRAQ4.117
## 1  3781.3374
## 2  1338.0596
## 3   581.7185
## 4   861.6777
## 5   742.1422
## 6   419.8419
```

The error/differences (results from MsnBase - results from PNNL):

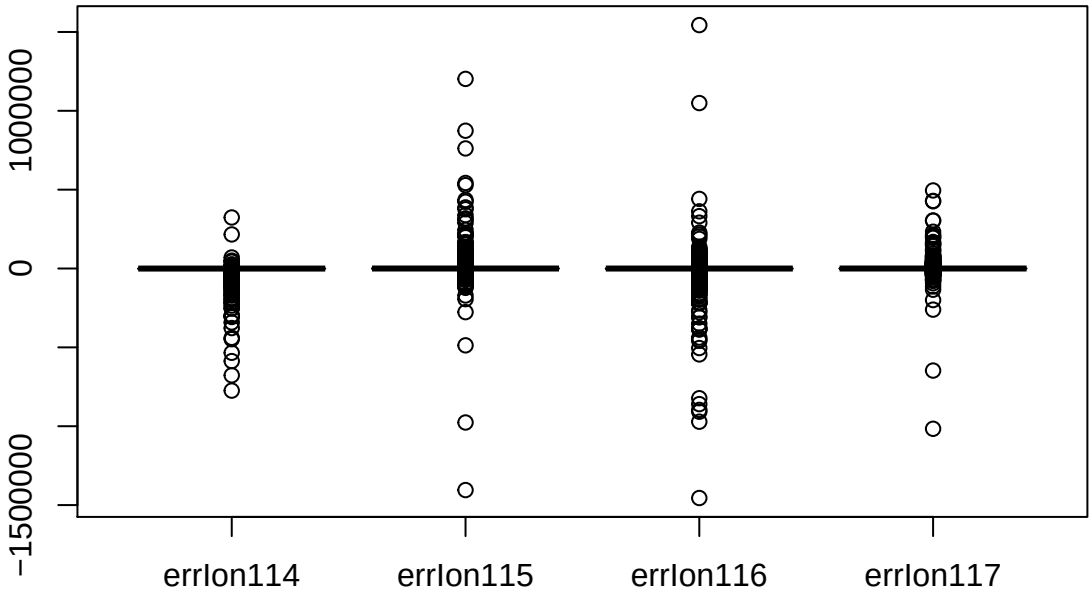
```
q.Error <- qnt.merged[,c("scan number(s)", "pepseq", "ms-gf:evalue")]
q.Error <- cbind(q.Error, "errIon114"= qnt.merged$iTRAQ4.114 - qnt.merged$Ion_114, "errIon115"=

head(q.Error[order(q.Error$`ms-gf:evalue`),])
```

```
##       scan number(s)                      pepseq ms-gf:evalue  errIon114
## 13885          17146  QNVVPTVLALGSDVDMDVLTTLSLGDR 7.599588e-24  -8.278823
## 11008          13662    VEQIAAIAQELNELDYYDSHNVNTR 7.777001e-22 -54.140093
## 11527          14281    VEQIAAIAQELNELDYYDSPSVNAR 3.604979e-20 -61.631328
## 11420          14149        FDGALNVDLTEFQTNLVPYPR 5.992245e-20         NA
```

```
## 10984          13635    VEQIAAIAQELNELDYYDSHNVNTR 6.314509e-19 -24.019705
## 12183          15079 NQAFIEMNTEEAANTMVNYYTSVTPVLR 1.025382e-16 -11.439651
##         errIon115   errIon116 errIon117
## 13885    1.275330  -0.6221765  2.717481
## 11008   34.327627 -82.2730078 70.171985
## 11527    1.552725 -43.6900098  4.359814
## 11420   13.006519 -25.1102393        NA
## 10984   -5.989380 -61.1232422 29.692783
## 12183 -14.804238   9.3017407 -4.080314
```

```
boxplot(q.Error[,c("errIon114","errIon115","errIon116","errIon117")])
```




```
summary(q.Error[,c("errIon114","errIon115","errIon116","errIon117")])
```

```
##    errIon114           errIon115            errIon116
##  Min.   :-774308.0   Min.   :-1404736.0   Min.   :-1455144.0
##  1st Qu.:  -1134.6   1st Qu.:     -46.4   1st Qu.:    -848.9
##  Median :   -295.0   Median :      42.1   Median :    -123.8
##  Mean   :  -2260.1   Mean   :     912.4   Mean   :   -1690.2
##  3rd Qu.:    -65.1   3rd Qu.:     415.4   3rd Qu.:      10.7
##  Max.   : 323633.0   Max.   : 1202377.5   Max.   : 1543492.2
##  NA's   :1238        NA's   :774          NA's   :537
##    errIon117
##  Min.   :-1016143.0
##  1st Qu.:     -19.2
##  Median :      56.6
##  Mean   :     656.0
##  3rd Qu.:     422.2
##  Max.   :  494918.5
##  NA's   :1040
```

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.