Portable Proteomics Pipeline (P3)
MSGF Benchmark

Daniel Kristiyanto (daniel.kristiyanto@pnnl.gov)

May 20, 2016

---

1. Files:

```
## [1] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11.mzML"
```

```
## [1] "ID_003632_9011437E.fasta"
```

2. Identification:

`MSGFPlus.jar` was downloaded from https://omics.pnl.gov/software/ms-gf. Identification was run multiple times (within or without Docker), and resulted the same results.

Identification with different switches resulted different results.

```
# Command Line (Without Switch)
java -Xmx8000M -jar P3/MSGFPlus.jar -s \
TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11.mzML -o \
TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11_no_switch.mzid \
-d ID_003632_9011437E.fasta
```

```
# Command Line (2)
java -Xmx8000M -jar P3/MSGFPlus.jar -s \
TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11.mzML -o \
TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11.mzid \
-d ID_003632_9011437E.fasta  \
-t 10ppm -m 0 -inst 1 -e 1 -ti -1,2 -ntt 2 -tda 1 -minLength 6 \
-maxLength 50 -minCharge 2 -maxCharge 5 -n 1 -thread 7 \
-mod MSGFDB_Mods.txt -minNumPeaks 5 -addFeatures 1
```

```
## [1] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11_no_switch.mzid"
## [2] "TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11.mzid"
```

3. Result from PNNL

```
## [1] 12948
```

```
##    Scan MSGFDB_SpecEValue                                    Peptide
## 1 11612      1.481722e-36 R.IQDLTGLDVSTAEELQVANYGVGGQYEPHFDFAR.K
## 2 14441      8.834771e-36  K.LTPFIIQENLNLALNSASAIGCHVVNIGAEDLR.A
##              Protein
## 1 ref|NP_001017962.1
## 2    ref|NP_005023.2

##       Scan MSGFDB_SpecEValue    Peptide             Protein
## 12947 3823        0.04856372 R.SPAGGG.-     ref|NP_061978.6
## 12948 4099        0.05022644 R.PGGAGM.- XXX_ref|NP_037488.2
```

4. Result from MSGF+ without any switches (sorted by SpecEValue)

```
##          scan number(s) ms-gf:denovoscore ms-gf:specevalue
## X9740.1           10815               277     4.434865e-30
## X8340.1            9275               234     2.157325e-27
## X10711.1          11883               157     1.386474e-25
##                                                                    accession
## X9740.1  ref|NP_001041660.1;ref|NP_001041659.1;ref|NP_001041664.1;ref|NP_001260.1
## X8340.1                        ref|NP_001092.1;ref|NP_001186883.1;ref|NP_001605.1
## X10711.1                                        ref|NP_001003962.1;ref|NP_001740.1
##                                 pepseq
## X9740.1  SHSTEPGLVLTLGQGDVGQLGLGENVMER
## X8340.1          DLYANTVLSGGTTMYPGIADR
## X10711.1      ILGGVISAISEAAAQYNPEPPPPR

##          scan number(s) ms-gf:denovoscore ms-gf:specevalue       accession
## X13036.1          14441               381     4.861374e-08 ref|NP_683515.3
## X10464.1          11612               279     5.246882e-08 ref|NP_000368.1
##                                   pepseq
## X13036.1     KETDLKQIQTLIQGTQTRLKYSQNELEMIKK
## X10464.1 AGISEAQLTDAETSKLIYDFIEDQGGLEAVRQEMR
```

Reading the MZID file manually

```
java -Xmx2000M -XX:+UseConcMarkSweepGC -cp \
../P3/MSGFPlus.jar edu.ucsd.msjava.ui.MzIDToTsv -i \
TCGA_13-1489_42-2590_36-2529_117C_W_PNNL_B2S5_f11_no_switch.mzid \
-showQValue 1 -showDecoy 1 -unroll 1

##   ScanNum                            Peptide            Protein SpecEValue
## 1   10815 R.SHSTEPGLVLTLGQGDVGQLGLGENVMER.K ref|NP_001041660.1   4.43e-30
## 2   10815 R.SHSTEPGLVLTLGQGDVGQLGLGENVMER.K ref|NP_001041659.1   4.43e-30
## 3   10815 R.SHSTEPGLVLTLGQGDVGQLGLGENVMER.K ref|NP_001041664.1   4.43e-30
```

5. Results with switches

```
##          scan number(s) ms-gf:denovoscore ms-gf:specevalue
## X11197.1          12418               321     6.104450e-34
## X9050.1           10056               258     4.553317e-33
## X9060.1           10067               288     9.340615e-33
##                                                                    accession
## X11197.1                                                      ref|NP_000683.3
## X9050.1  ref|NP_004376.2;ref|NP_001157569.1;ref|NP_001157570.1;ref|NP_001119808.1
## X9060.1  ref|NP_004376.2;ref|NP_001157569.1;ref|NP_001157570.1;ref|NP_001119808.1
##                                pepseq
## X11197.1 VYLASLETLDNGKPFQESYALDLDEVIK
## X9050.1     CLCLPSYVGALCEQDTETCDYGWHK
## X9060.1     CLCLPSYVGALCEQDTETCDYGWHK
```

```
##          scan number(s) ms-gf:denovoscore ms-gf:specevalue       accession
## X13036.1          14441               345     2.196462e-06 ref|NP_060250.2
## X10464.1          11612               279     2.280634e-06 ref|NP_000487.1
##                                   pepseq
## X13036.1 FILPNVSTPVSDAFKTQMELLQAGLSRTPTR
## X10464.1  GDYKDSSDFGAPHPQVQSVRRIRDMQWHQR
```

Reading the MZID file manually

```
##   ScanNum
## 1   12418
## 2   10056
## 3   10056
##                                                                  Peptide
## 1                  R.+144.102VYLASLETLDNGK+144.102PFQESYALDLDEVIK+144.102.V
## 2 R.+144.102C+57.021LC+57.021LPSYVGALC+57.021EQDTETC+57.021DYGWHK+144.102.F
## 3 R.+144.102C+57.021LC+57.021LPSYVGALC+57.021EQDTETC+57.021DYGWHK+144.102.F
##              Protein SpecEValue
## 1    ref|NP_000683.3   6.10e-34
## 2    ref|NP_004376.2   4.55e-33
## 3 ref|NP_001157569.1   4.55e-33
```

6. Filtering

a Spec-Evalue 10^{-10}

From PNNL

```
## [1] "==Head=="
```

```
##    Scan MSGFDB_SpecEValue                              Peptide
## 1 11612      1.481722e-36 IQDLTGLDVSTAEELQVANYGVGGQYEPHFDFAR
## 2 14441      8.834771e-36  LTPFIIQENLNLALNSASAIGCHVVNIGAEDLR
##              Protein
## 1 ref|NP_001017962.1
## 2    ref|NP_005023.2
```

```
## [1] "==Tail=="
```

```
##       Scan MSGFDB_SpecEValue       Peptide         Protein
## 4634  7324      9.821299e-11 VTGTQPITCTWMK ref|NP_444253.3
## 4635 12688      9.998422e-11      SFLFQLLK ref|NP_004926.1
```

There are 4635 rows remained.

From Docker

```
## [1] "==Head=="
```

```
##          scan number(s) ms-gf:specevalue                        pepseq
## X11197.1          12418     6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK
## X9050.1           10056     4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK
##                                                                    accession
## X11197.1                                                      ref|NP_000683.3
## X9050.1  ref|NP_004376.2;ref|NP_001157569.1;ref|NP_001157570.1;ref|NP_001119808.1
```

```
## [1] "==Tail=="
```

```
##         scan number(s) ms-gf:specevalue        pepseq
## X3022.1           3426     9.690560e-11       MKPEFEK
## X6566.1           7324     9.821299e-11 VTGTQPITCTWMK
##                                                    accession
## X3022.1                                          ref|NP_006801.1
## X6566.1 ref|NP_444253.3;ref|NP_444256.3;ref|NP_444255.3;ref|NP_444254.3
```

There are 1665 rows remained.

Differences

Data from Docker is matched with from PNNL using scan number as ID. ms-gf:specevalue, pepseq, accession are values from Docker, MSGFDB_SpecEValue, Peptide, Protein are the corresponding values from PNNL. There are 1681 of rows in total with 13 of NULL/unmapped values.

Aside from the 13 unmatched values, 0 unmacthed spec-evalue, 0 of unmatched peptide identification.

```
## [1] "==Head=="
```

```
##      scan number(s) ms-gf:specevalue                       pepseq
## 1596          12418     6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK
## 1445          10056     4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK
##                                                                 accession
## 1596                                                      ref|NP_000683.3
## 1445 ref|NP_004376.2;ref|NP_001157569.1;ref|NP_001157570.1;ref|NP_001119808.1
##      MSGFDB_SpecEValue                       Peptide         Protein
## 1596      6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK ref|NP_000683.3
## 1445      4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK ref|NP_004376.2
```

```
## [1] "==Unmatched=="
```

```
##      scan number(s) ms-gf:specevalue
## 1098           7622     4.957978e-18
## 1569          12009     3.005655e-16
## 1114           7714     9.284183e-16
##                                  pepseq
## 1098                      EQQEAIEHIDEVQNEIDR
## 1569 SGINPFDSQEAKPYKNDPEILAMTNLVSAYQNNDITEFEK
## 1114                            TVVTGIEMFHK
##                              accession MSGFDB_SpecEValue Peptide Protein
## 1098 ref|NP_001116293.1;ref|NP_003002.2               NA    <NA>    <NA>
## 1569 ref|NP_001137359.1;ref|NP_004227.1               NA    <NA>    <NA>
## 1114                    ref|NP_003312.3               NA    <NA>    <NA>
```

```
## [1] "==Tail=="
```

```
##      scan number(s) ms-gf:specevalue       pepseq
## 242            3426     9.690560e-11       MKPEFEK
## 1037           7324     9.821299e-11 VTGTQPITCTWMK
##                                                    accession
## 242                                             ref|NP_006801.1
## 1037 ref|NP_444253.3;ref|NP_444256.3;ref|NP_444255.3;ref|NP_444254.3
##      MSGFDB_SpecEValue       Peptide         Protein
## 242       9.690560e-11       MKPEFEK ref|NP_006801.1
## 1037      9.821299e-11 VTGTQPITCTWMK ref|NP_444253.3
```

b. Spec-Evalue 10^{-15}

From PNNL

```
## [1] "==Head=="
```

```
##    Scan MSGFDB_SpecEValue                          Peptide
## 1 11612      1.481722e-36 IQDLTGLDVSTAEELQVANYGVGGQYEPHFDFAR
```

```
## 2 14441      8.834771e-36  LTPFIIQENLNLALNSASAIGCHVVNIGAEDLR
##              Protein
## 1 ref|NP_001017962.1
## 2    ref|NP_005023.2
```

```
## [1] "==Tail=="
```

```
##       Scan MSGFDB_SpecEValue      Peptide         Protein
## 2510 13805      9.825662e-16 VLSEIFSPMLFR ref|NP_852664.1
## 2511  8246      9.964967e-16  TMFSLDTYSTK ref|NP_004360.2
```

There are 2511 rows remained.

From Docker

```
## [1] "==Head=="
```

```
##          scan number(s) ms-gf:specevalue                      pepseq
## X11197.1          12418     6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK
## X9050.1           10056     4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK
##                                                                 accession
## X11197.1                                                  ref|NP_000683.3
## X9050.1  ref|NP_004376.2;ref|NP_001157569.1;ref|NP_001157570.1;ref|NP_001119808.1
```

```
## [1] "==Tail=="
```

```
##         scan number(s) ms-gf:specevalue          pepseq
## X5748.1           6424     9.527297e-16      QDCETFGMVVK
## X7929.1           8823     9.711903e-16 ITNQVIYLNPPIEECR
##
## X5748.1 ref|NP_958923.1;ref|NP_958848.1;ref|NP_057572.5;ref|NP_001137259.1;ref|NP_958847.1;r
## X7929.1
```

There are 767 rows remained.

Differences:

Data from Docker is matched with from PNNL using scan number as ID. ms-gf:specevalue, pepseq, accession are values from Docker, MSGFDB_SpecEValue, Peptide, Protein are the corresponding values from PNNL. There are 771 of rows in total with 3 of NULL/unmapped values.

Aside from the 3 unmatched values, 0 unmacthed spec-evalue, 0 of unmatched peptide identification.

```
## [1] "==Head=="
```

```
##     scan number(s) ms-gf:specevalue                       pepseq
## 719          12418     6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK
## 633          10056     4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK
##                                                                 accession
## 719                                                       ref|NP_000683.3
## 633 ref|NP_004376.2;ref|NP_001157569.1;ref|NP_001157570.1;ref|NP_001119808.1
##     MSGFDB_SpecEValue                      Peptide         Protein
## 719      6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK ref|NP_000683.3
## 633      4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK ref|NP_004376.2
```

```
## [1] "==Unmatched=="
```

```
##     scan number(s) ms-gf:specevalue
## 449           7622     4.957978e-18
## 706          12009     3.005655e-16
## 457           7714     9.284183e-16
##                                  pepseq
## 449                      EQQEAIEHIDEVQNEIDR
## 706 SGINPFDSQEAKPYKNDPEILAMTNLVSAYQNNDITEFEK
## 457                            TVVTGIEMFHK
##                             accession MSGFDB_SpecEValue Peptide Protein
## 449 ref|NP_001116293.1;ref|NP_003002.2               NA    <NA>    <NA>
## 706 ref|NP_001137359.1;ref|NP_004227.1               NA    <NA>    <NA>
## 457                   ref|NP_003312.3               NA    <NA>    <NA>
```

```
## [1] "==Tail=="
```

```
##     scan number(s) ms-gf:specevalue          pepseq
## 296           6424     9.527297e-16      QDCETFGMVVK
## 537           8823     9.711903e-16 ITNQVIYLNPPIEECR
##
## 296 ref|NP_958923.1;ref|NP_958848.1;ref|NP_057572.5;ref|NP_001137259.1;ref|NP_958847.1;ref|
## 537
##     MSGFDB_SpecEValue          Peptide         Protein
## 296      9.527297e-16      QDCETFGMVVK ref|NP_958923.1
## 537      9.711903e-16 ITNQVIYLNPPIEECR ref|NP_001367.2
```

c. Spec-Evalue 10^{-20}

From PNNL

```
## [1] "==Head=="
```

```
##    Scan MSGFDB_SpecEValue                              Peptide
## 1 11612      1.481722e-36 IQDLTGLDVSTAEELQVANYGVGGQYEPHFDFAR
```

```
## 2 14441      8.834771e-36  LTPFIIQENLNLALNSASAIGCHVVNIGAEDLR
##              Protein
## 1 ref|NP_001017962.1
## 2    ref|NP_005023.2

## [1] "==Tail=="

##      Scan MSGFDB_SpecEValue               Peptide         Protein
## 935 13489      9.813857e-21 YETVIMPVFGIATPFHIATIK ref|NP_009123.1
## 936 13601      9.996277e-21  IAAGLPMAGIPFLTTDLTYR ref|NP_115649.1
```

There are 936 rows remained.

From Docker

```
## [1] "==Head=="

##          scan number(s) ms-gf:specevalue                      pepseq
## X11197.1          12418     6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK
## X9050.1           10056     4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK
##                                                                 accession
## X11197.1                                                 ref|NP_000683.3
## X9050.1  ref|NP_004376.2;ref|NP_001157569.1;ref|NP_001157570.1;ref|NP_001119808.1

## [1] "==Tail=="

##         scan number(s) ms-gf:specevalue             pepseq
## X7456.1           8303     8.950565e-21    TPALVFEHVNNTDFK
## X7761.1           8638     9.647618e-21 LILEQMQKDPQALSEHLK
##                              accession
## X7456.1 ref|NP_808227.1;ref|NP_001886.1
## X7761.1                 ref|NP_006810.1
```

There are 223 rows remained.

Differences

Data from Docker is matched with from PNNL using scan number as ID. ms-gf:specevalue, pepseq, accession are values from Docker, MSGFDB_SpecEValue, Peptide, Protein are the corresponding values from PNNL. There are 223 of rows in total with 0 of NULL/unmapped values.

Aside from the 0 unmatched values, 0 unmacthed spec-evalue, 0 of unmatched peptide identification.

```
## [1] "==Head=="
```

```
##     scan number(s) ms-gf:specevalue                      pepseq
## 199          12418     6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK
## 163          10056     4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK
##                                                                 accession
## 199                                                        ref|NP_000683.3
## 163 ref|NP_004376.2;ref|NP_001157569.1;ref|NP_001157570.1;ref|NP_001119808.1
##     MSGFDB_SpecEValue                      Peptide         Protein
## 199      6.104450e-34 VYLASLETLDNGKPFQESYALDLDEVIK ref|NP_000683.3
## 163      4.553317e-33    CLCLPSYVGALCEQDTETCDYGWHK ref|NP_004376.2
```

```
## [1] "==Unmatched=="
```

```
## [1] scan number(s)    ms-gf:specevalue  pepseq            accession
## [5] MSGFDB_SpecEValue Peptide           Protein
## <0 rows> (or 0-length row.names)
```

```
## [1] "==Tail=="
```

```
##     scan number(s) ms-gf:specevalue             pepseq
## 110           8303     8.950565e-21    TPALVFEHVNNTDFK
## 116           8638     9.647618e-21 LILEQMQKDPQALSEHLK
##                            accession MSGFDB_SpecEValue            Peptide
## 110 ref|NP_808227.1;ref|NP_001886.1      6.808255e-21    TPALVFEHVNNTDFK
## 116                 ref|NP_006810.1      7.600654e-21 LILEQMQKDPQALSEHLK
##             Protein
## 110 ref|NP_808227.1
## 116 ref|NP_006810.1
```