Executable Analysis Document Supporting Proteomics Component of the Manusctipt: **“Widespread Abrogation of Triplet Translation Continuity and Stop Codon Function in Euplotes”**

Alexei V. Lobanov, Stephen M. Heaphy, Anton A. Turanov, Maxim V. Gerashchenko, Sandra Pucciarelli, Raghul R. Devaraj, Fang Xie, Vladislav A. Petyuk, Richard D. Smith, Lawrence A. Klobutcher, John F. Atkins, Cristina Miceli, Dolph L. Hatfield, Pavel V. Baranov, Vadim N. Gladyshev

Tue 21 Jun 2016

Contents

1 Introduction 1

2 Post MS/MS Search Analysis Steps 2

2.1 Prerequisites 2

2.1.1 Dowloading Datasets 2

2.1.2 Reading Frameshift Marks 2

2.2 Processing of MS/MS Search Results 3

2.2.1 Trypsin Digest Fractionated by SCX 3

2.2.2 Trypsin Digest Fractionated by HPRP 5

2.2.3 Glu-C Digest Fractionated by HPRP 7

2.3 Compendium of Peptides Covering Frameshift Locations 9

3 Manual Validation 11

4 Session Information 19

1 Introduction

The vignette describes and reproduces all the steps that aimed to confirm frameshifts in the *Euplotes crassus* proteome. The global 8M urea soluble proteome was digested using conventional trypsin protocol and alternatively with Glu-C protease under high pH (7.5) conditions. The latter restricts specificity of Glu-C cleavages to C-terminal of glutamic acid (E). The peptides resulting from trypsin digest were fractionated using two different approaches: with strong cation exchange (SCX) and high pH reverse phase (HPRP) chromatographies. The peptides from Glu-C digest were fractionated using HPRP only.

The datasets were deposited to PRIDE and available by this link http://dx.doi.org/10.6019/PXD004333. Summary of the datasets shown in the table below:

Dataset Prefix	Digestion Enzyme	Fractionation Chromatrography Type
Euplotes_1_SCX	trypsin	SCX
Euplotes_1_HPRP_1	trypsin	HPRP
Euplotes_1_HPRP_2	Glu-C (pH 7.5)	HPRP

Preprocessing of the `raw` files prior MS/MS searches was done in two steps. First, the raw files were processed with

DeconMSn to correct for wrong assignments of monoisotopic peaks. The parameters are as follows:

```
DeconMSN.exe -I35 -G1 -F1 -L6810 -B200 -T5000 -M3 -XCDTA
```

At the second step the peak files were processed with DtaRefinery to perform post-acquisition recalibaration of parent ion mass-to-charge ratios. The peak lists (concatenated dta files in this case) were searched using MS-GF+ tool against 6-frame translated *Euplotes Crassus* genome concatenated with tentatively frameshifted sequences and common contaminants. The 6-frame translated FASTA file, `DtaRefinery` and `MS-GF+` parameter files are available in `extdata` folder of the `EuplotesCrassus.proteome` package.

For example:

```
fpath <- system.file("extdata",
                     "MSGFDB_GluC_StatCysAlk_10ppmParTol.txt",
                     package="EuplotesCrassus.proteome")
cat(readLines(fpath, n=12), sep = '\n')
## #Parent mass tolerance
## #  Examples: 2.5Da or 30ppm
## #  Use comma to set asymmetric values, for example "0.5Da,2.5Da" will set 0.5Da to the left (expMass<the
## PMTolerance=10ppm
##
## #Max Number of Modifications per peptide
## # If this value is large, the search will be slow
## NumMods=3
##
## #Modifications (see below for examples)
## StaticMod=C2H3N1O1,    C,  fix, any,       Carbamidomethyl       # Fixed Carbamidomethyl C (alkylation,
```

2 Post MS/MS Search Analysis Steps

2.1 Prerequisites

2.1.1 Dowloading Datasets

To download the datasets we will take advantage of rpx R package. Note, this step may take awhile depending on the speed of the internet connection (~30 min in my case). However, if they are downloaded the script will use the available datasets instead of downloading them again.

```
library(rpx)
id <- "PXD004333"
px <- PXDataset(id)
repoFiles <- pxfiles(px)
mzids <- grep('*msgfplus.mzid.gz', repoFiles, value=T)
system.time(pxget(px, mzids))
##     user   system  elapsed
##    1.011    5.366 1203.045
```

2.1.2 Reading Frameshift Marks

The FASTA files containing 595 sequences with frameshifts availabe as a part of this package and available as `system.file("extdata", "Euplotes_Crassus_frameshifts.fasta", package="EuplotesCrassus.proteome")`. There is an additional FASTA file with frameshift locations marked with exclamation mark `!`.

```
library(Biostrings)
fasta_clean <- readAAStringSet(
    system.file("extdata",
                "Euplotes_Crassus_frameshifts.fasta",
                package="EuplotesCrassus.proteome"),
    format="fasta", nrec=-1L, skip=0L, use.names=TRUE)
fasta_marks <- readAAStringSet(
    system.file("extdata",
                "Euplotes_Crassus_frameshifts_with_mark.fasta",
                package="EuplotesCrassus.proteome"),
    format="fasta", nrec=-1L, skip=0L, use.names=TRUE)
length(fasta_clean)
## [1] 595
```

2.2 Processing of MS/MS Search Results

2.2.1 Trypsin Digest Fractionated by SCX

For processing of MS/MS identification we will use MSnID R package. First step is to read the LC-MS/MS datasets corresponding to 25 SCX fractions.

```
library(MSnID)
## Warning in fun(libname, pkgname): mzR has been built against a different Rcpp version (0.12.3)
## than is installed on your system (0.12.4). This might lead to errors
## when loading mzR. If you encounter such issues, please send a report,
## including the output of sessionInfo() to the Bioc support forum at
## https://support.bioconductor.org/. For details see also
## https://github.com/sneumann/mzR/wiki/mzR-Rcpp-compiler-linker-issue.
```

```
trypscx <- grep('Euplotes_1_SCX_.*msgfplus.mzid.gz', repoFiles, value=T)
trypscxPrj <- MSnID()
system.time(trypscxPrj <- read_mzIDs(trypscxPrj, trypscx, backend = 'mzR'))
##    user  system elapsed
##  11.491   0.835  32.042
```

Assess the peptide termini for their corresponding cleavage patterns. We will lleave peptides that resuted only from proper trypsin cleavave events. That is we won't allow peptide resulting from irregular clevages.

```
trypscxPrj <- assess_termini(trypscxPrj, validCleavagePattern="[KR]\\.[^P]")
trypscxPrj <- apply_filter(trypscxPrj, "numIrregCleavages == 0")
```

Note, that for this project we are interested only in peptides covering the sites of the frameshifting events. So if a peptide identification can be explained by a regular protein sequence we are not interested in pursuing this identification. The protein/accession names of normal (non-frameshifted) sequences starts with `Contig` or `Contaminant`. If the FASTA entry sequence is a results of the frameshift event if starts with `comp`. Therefore in the code below we retain only peptide-to-spectrum matches that can appear only due to frameshifted sequences.

```
#' Rule on how to split the names.
#' Contig + Contaminants - main piece
#' comp - sequences with frameshifts
trypscxPrj.main <- apply_filter(trypscxPrj, "!grepl('comp', accession)")
trypscxPrj.fmsh <- apply_filter(trypscxPrj, "grepl('comp', accession)")
#' if peptide matches to the main piece we don't care about it
trypscxPrj.fmsh <- apply_filter(trypscxPrj.fmsh,
                                "!(peptide %in% peptides(trypscxPrj.main))")
show(trypscxPrj.fmsh)
## MSnID object
## Working directory: "."
## #Spectrum Files:  25
## #PSMs: 442 at 58 % FDR
## #peptides: 348 at 67 % FDR
## #accessions: 291 at 66 % FDR
```

Setting-up and optimizing filtering options for MS/MS identifications. Since the number of peptides mapping frameshifted sequences is rather low we will loosed up the FDR of the identification up to 5%, however, then follow-up with manual spectra validation.

```
trypscxPrj.fmsh$mme.ppm <- abs(mass_measurement_error(trypscxPrj.fmsh))
trypscxPrj.fmsh$score <- -log10(trypscxPrj.fmsh$`MS.GF.SpecEValue`)
trypscxPrj.fmsh <- apply_filter(trypscxPrj.fmsh, "mme.ppm < 10")

filtr <- MSnIDFilter(trypscxPrj.fmsh)
filtr$mme.ppm <- list(comparison="<", threshold=5.0)
filtr$score <- list(comparison=">", threshold=8.0)
#' pre-optimization with brute-force approach
filtr.grid <- optimize_filter(filtr, trypscxPrj.fmsh, fdr.max=0.05,
                              method="Grid", level="peptide", n.iter=20000)
evaluate_filter(trypscxPrj.fmsh, filtr.grid)
##                  fdr   n
## PSM       0.02970297 104
## peptide   0.03703704  56
## accession 0.04166667  50
```

```
#' fine tune with optimization using simulated annealing technique
filtr.sann <- optimize_filter(filtr.grid, trypscxPrj.fmsh, fdr.max=0.05,
                              method="SANN", level="peptide", n.iter=20000)
evaluate_filter(trypscxPrj.fmsh, filtr.sann)
##                  fdr   n
## PSM       0.02941176 105
## peptide   0.03636364  57
## accession 0.04081633  51
```

```
trypscxPrj.fmsh <- apply_filter(trypscxPrj.fmsh, filtr.sann)
show(trypscxPrj.fmsh)
## MSnID object
## Working directory: "."
## #Spectrum Files:  18
## #PSMs: 105 at 2.9 % FDR
## #peptides: 57 at 3.6 % FDR
## #accessions: 51 at 4.1 % FDR
```

Finally we will extract only those peptides that exactly span the frameshift sites. That is their sequences should be present/identifiable in normal FASTA file, however missing in the file with frameshifts masked with the exclamation mark !.

```
#' extract only those that map frameshift sites
library(dplyr)
pepSeq <- unique(trypscxPrj.fmsh$pepSeq)
pepSeqMapped_to_clean <- pepSeq %>%
    sapply(grep, x=fasta_clean) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
pepSeqMapped_to_with_marks <- pepSeq %>%
    sapply(grep, x=fasta_marks) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
pepSeqFmsh_trypscx <- setdiff(pepSeqMapped_to_clean, pepSeqMapped_to_with_marks)
print(pepSeqFmsh_trypscx)
```

```
## [1] "SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK" "WTPIDLPSEEITFVQGIQTVTGAGDPSMK"
## [3] "ESNHNNDITNKNEIAYILR"             "KKKQEENNLKR"
```

Reporting extra information on the peptide sequences spanning frameshift sites: dataset, scan, charge, score, and mass measurement error.

```
meta_tryp_scx <- trypscxPrj.fmsh %>%
    apply_filter('pepSeq %in% pepSeqFmsh_trypscx') %>%
    psms %>%
    select(spectrumFile,MS.GF.SpecEValue,mme.ppm,spectrumID,chargeState,peptide) %>%
    rename(SpecEValue = MS.GF.SpecEValue, charge = chargeState, `MME (ppm)`=mme.ppm) %>%
    mutate(spectrumFile = sub('_msgfplus.mzid.gz','',spectrumFile))
library(xtable)
print(xtable(meta_tryp_scx, display = c('d','s','e','f','s','d','s')),
      include.rownames=FALSE,
      comment = FALSE,
      size='scriptsize',
      floating = F)
```

spectrumFile	SpecEValue	MME (ppm)	spectrumID	charge	peptide
Euplotes_1_SCX_10_13Nov09_Falcon_09-09-14	3.41e-15	0.30	index=6106	3	K.SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK.V
Euplotes_1_SCX_10_13Nov09_Falcon_09-09-14	3.41e-15	0.30	index=6106	3	K.SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK.V
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	1.53e-21	0.08	index=8908	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	1.07e-20	1.10	index=8896	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	7.29e-19	1.10	index=8897	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	2.17e-15	0.94	index=8895	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_18_13Nov09_Falcon_09-09-15	9.27e-17	0.11	index=5912	2	K.ESNHNNDITNKNEIAYILR.Y
Euplotes_1_SCX_20_13Nov09_Falcon_09-09-15	2.23e-11	0.70	index=10317	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_22_13Nov09_Falcon_09-09-15	4.36e-10	3.76	index=9720	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_23_13Nov09_Falcon_09-09-15	2.47e-09	1.64	index=9440	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_SCX_24_13Nov09_Falcon_09-09-15	3.42e-10	8.85	index=2127	3	R.KKKQEENNLKR.K

2.2.2 Trypsin Digest Fractionated by HPRP

All the processing steps are conceptually the same as in the section above.

```
library(MSnID)
tryphprp <- grep('Euplotes_1_HPRP_1_.*msgfplus.mzid.gz', repoFiles, value=T)
tryphprpPrj <- MSnID()
system.time(tryphprpPrj <- read_mzIDs(tryphprpPrj, tryphprp, backend = 'mzR'))
```

```
##    user  system elapsed
##   8.047   0.767  30.555
```

```
tryphprpPrj <- assess_termini(tryphprpPrj, validCleavagePattern="[KR]\\.[^P]")
tryphprpPrj <- apply_filter(tryphprpPrj, "numIrregCleavages == 0")
```

```
tryphprpPrj.main <- apply_filter(tryphprpPrj, "!grepl('comp', accession)")
tryphprpPrj.fmsh <- apply_filter(tryphprpPrj, "grepl('comp', accession)")
tryphprpPrj.fmsh <- apply_filter(tryphprpPrj.fmsh,
                                "!(peptide %in% peptides(tryphprpPrj.main))")
show(tryphprpPrj.fmsh)
## MSnID object
## Working directory: "."
## #Spectrum Files:  24
## #PSMs: 511 at 49 % FDR
## #peptides: 399 at 62 % FDR
## #accessions: 293 at 78 % FDR
```

```
tryphprpPrj.fmsh$mme.ppm <- abs(mass_measurement_error(tryphprpPrj.fmsh))
tryphprpPrj.fmsh$score <- -log10(tryphprpPrj.fmsh$`MS.GF.SpecEValue`)
tryphprpPrj.fmsh <- apply_filter(tryphprpPrj.fmsh, "mme.ppm < 10")

filtr <- MSnIDFilter(tryphprpPrj.fmsh)
filtr$mme.ppm <- list(comparison="<", threshold=5.0)
filtr$score <- list(comparison=">", threshold=8.0)
filtr.grid <- optimize_filter(filtr, tryphprpPrj.fmsh, fdr.max=0.05,
                              method="Grid", level="peptide", n.iter=20000)
evaluate_filter(tryphprpPrj.fmsh, filtr.grid)
##                  fdr   n
## PSM       0.02631579 195
## peptide   0.04504505 116
## accession 0.07142857  75
```

```
filtr.sann <- optimize_filter(filtr.grid, tryphprpPrj.fmsh, fdr.max=0.05,
                              method="SANN", level="peptide", n.iter=20000)
evaluate_filter(tryphprpPrj.fmsh, filtr.sann)
##                  fdr   n
## PSM       0.02631579 195
## peptide   0.04504505 116
## accession 0.07142857  75
```

```
tryphprpPrj.fmsh <- apply_filter(tryphprpPrj.fmsh, filtr.sann)
show(tryphprpPrj.fmsh)
## MSnID object
## Working directory: "."
## #Spectrum Files:  23
## #PSMs: 195 at 2.6 % FDR
## #peptides: 116 at 4.5 % FDR
## #accessions: 75 at 7.1 % FDR
```

```
library(dplyr)
pepSeq <- unique(tryphprpPrj.fmsh$pepSeq)
pepSeqMapped_to_clean <- pepSeq %>%
    sapply(grep, x=fasta_clean) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
pepSeqMapped_to_with_marks <- pepSeq %>%
    sapply(grep, x=fasta_marks) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
```

```
pepSeqFmsh_tryphprp <- setdiff(pepSeqMapped_to_clean, pepSeqMapped_to_with_marks)
print(pepSeqFmsh_tryphprp)
```

```
##  [1] "FFAAPEK"                         "ELAFLKRAQEIGLEPYNEYHGKKK"
##  [3] "VVQEGNTNVKK"                     "WTPIDLPSEEITFVQGIQTVTGAGDPSMK"
##  [5] "IIQNFQINTVFEDLDEIMQTQVQR"        "KSSKACEEERRKR"
##  [7] "LINDLTNDK"                       "LISELTSEK"
##  [9] "IVENFNK"                         "LSQEHLSYISR"
## [11] "LINDLTNDKANLK"
```

```
meta_tryp_hprp <- tryphprpPrj.fmsh %>%
    apply_filter('pepSeq %in% pepSeqFmsh_tryphprp') %>%
    psms %>%
    select(spectrumFile,MS.GF.SpecEValue,mme.ppm,spectrumID,chargeState,peptide) %>%
    rename(SpecEValue = MS.GF.SpecEValue, charge = chargeState, `MME (ppm)`=mme.ppm) %>%
    mutate(spectrumFile = sub('_msgfplus.mzid.gz','',spectrumFile))
library(xtable)
print(xtable(meta_tryp_hprp, display = c('d','s','e','f','s','d','s')),
      include.rownames=FALSE,
      comment = FALSE,
      size='scriptsize',
      floating = F)
```

spectrumFile	SpecEValue	MME (ppm)	spectrumID	charge	peptide
Euplotes_1_HPRP_1_04_17Nov09_Falcon_09-09-14	7.58e-11	0.08	index=3031	1	R.FFAAPEK.I
Euplotes_1_HPRP_1_04_17Nov09_Falcon_09-09-14	2.44e-09	0.00	index=3046	2	R.FFAAPEK.I
Euplotes_1_HPRP_1_05_17Nov09_Falcon_09-09-14	1.46e-09	5.31	index=8245	3	R.ELAFLKRAQEIGLEPYNEYHGKKK.T
Euplotes_1_HPRP_1_06_17Nov09_Falcon_09-09-14	5.54e-10	2.21	index=759	2	K.VVQEGNTNVKK.L
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	5.93e-22	2.11	index=8644	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	2.18e-21	0.78	index=8638	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	3.05e-21	2.11	index=8646	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	4.19e-16	0.82	index=8639	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.19e-21	0.70	index=8806	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.20e-21	1.57	index=8812	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	5.49e-20	1.64	index=8802	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	4.33e-15	1.53	index=8810	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A
Euplotes_1_HPRP_1_16_22Nov09_Falcon_09-09-14	4.51e-21	0.33	index=10684	2	K.IIQNFQINTVFEDLDEIMQTQVQR.H
Euplotes_1_HPRP_1_16_22Nov09_Falcon_09-09-14	1.36e-11	1.25	index=10678	3	K.IIQNFQINTVFEDLDEIMQTQVQR.H
Euplotes_1_HPRP_1_18_17Nov09_Falcon_09-09-15	5.08e-09	2.64	index=13785	2	K.KSSKACEEERRKR.E
Euplotes_1_HPRP_1_20_17Nov09_Falcon_09-09-15	1.91e-11	0.00	index=3425	1	K.LINDLTNDK.A
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	6.65e-11	1.67	index=3600	2	K.LISELTSEK.S
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	2.55e-10	0.78	index=3602	1	K.LISELTSEK.S
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	1.89e-09	0.49	index=2595	2	K.IVENFNK.I
Euplotes_1_HPRP_1_23_17Nov09_Falcon_09-09-15	3.01e-13	1.01	index=2200	2	K.LSQEHLSYISR.L
Euplotes_1_HPRP_1_24_17Nov09_Falcon_09-09-15	2.45e-16	1.41	index=2709	2	K.LINDLTNDKANLK.D

2.2.3 Glu-C Digest Fractionated by HPRP

All the processing steps are conceptually the same as in the section above. The only substantial diffence is the specification of the enzyme digestion rule.

```
library(MSnID)
gluchprp <- grep('Euplotes_1_HPRP_2_.*msgfplus.mzid.gz', repoFiles, value=T)
gluchprpPrj <- MSnID()
system.time(gluchprpPrj <- read_mzIDs(gluchprpPrj, gluchprp, backend = 'mzR'))
```

```
##    user  system elapsed
##   5.866   0.698  28.562
```

```
gluchprpPrj <- assess_termini(gluchprpPrj, validCleavagePattern="E\\.[^P]")
gluchprpPrj <- apply_filter(gluchprpPrj, "numIrregCleavages == 0")
```

```
gluchprpPrj.main <- apply_filter(gluchprpPrj, "!grepl('comp', accession)")
gluchprpPrj.fmsh <- apply_filter(gluchprpPrj, "grepl('comp', accession)")
gluchprpPrj.fmsh <- apply_filter(gluchprpPrj.fmsh,
                                "!(peptide %in% peptides(gluchprpPrj.main))")
show(gluchprpPrj.fmsh)
## MSnID object
## Working directory: "."
## #Spectrum Files:  24
## #PSMs: 555 at 67 % FDR
## #peptides: 440 at 80 % FDR
## #accessions: 297 at 89 % FDR
```

```
gluchprpPrj.fmsh$mme.ppm <- abs(mass_measurement_error(gluchprpPrj.fmsh))
gluchprpPrj.fmsh$score <- -log10(gluchprpPrj.fmsh$`MS.GF.SpecEValue`)
gluchprpPrj.fmsh <- apply_filter(gluchprpPrj.fmsh, "mme.ppm < 10")

filtr <- MSnIDFilter(gluchprpPrj.fmsh)
filtr$mme.ppm <- list(comparison="<", threshold=5.0)
filtr$score <- list(comparison=">", threshold=8.0)
filtr.grid <- optimize_filter(filtr, gluchprpPrj.fmsh, fdr.max=0.05,
                              method="Grid", level="peptide", n.iter=20000)
evaluate_filter(gluchprpPrj.fmsh, filtr.grid)
##                  fdr  n
## PSM       0.02222222 46
## peptide   0.03448276 30
## accession 0.05000000 21
```

```
filtr.sann <- optimize_filter(filtr.grid, gluchprpPrj.fmsh, fdr.max=0.05,
                              method="SANN", level="peptide", n.iter=20000)
evaluate_filter(gluchprpPrj.fmsh, filtr.sann)
##                  fdr  n
## PSM       0.02222222 46
## peptide   0.03448276 30
## accession 0.05000000 21
```

```
gluchprpPrj.fmsh <- apply_filter(gluchprpPrj.fmsh, filtr.sann)
show(gluchprpPrj.fmsh)
## MSnID object
## Working directory: "."
## #Spectrum Files:  18
## #PSMs: 46 at 2.2 % FDR
## #peptides: 30 at 3.4 % FDR
## #accessions: 21 at 5 % FDR
```

```
library(dplyr)
pepSeq <- unique(gluchprpPrj.fmsh$pepSeq)
pepSeqMapped_to_clean <- pepSeq %>%
    sapply(grep, x=fasta_clean) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
pepSeqMapped_to_with_marks <- pepSeq %>%
    sapply(grep, x=fasta_marks) %>%
    sapply(length) %>%
    subset(.>0) %>%
    names
pepSeqFmsh_gluchprp <- setdiff(pepSeqMapped_to_clean, pepSeqMapped_to_with_marks)
print(pepSeqFmsh_gluchprp)
```

```
## [1] "NFNKITGKEQEEEE"                    "SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE"
## [3] "NLDNEKLINDLTNDKANLKDIVFDLMFE"      "NKIRFFAAPEKIFE"
## [5] "MQDEEILKSIEESKLEQEQEEEKKNE"        "VYLGLMEEYE"
```

```
meta_gluc_hprp <- gluchprpPrj.fmsh %>%
    apply_filter('pepSeq %in% pepSeqFmsh_gluchprp') %>%
    psms %>%
    select(spectrumFile,MS.GF.SpecEValue,mme.ppm,spectrumID,chargeState,peptide) %>%
    rename(SpecEValue = MS.GF.SpecEValue, charge = chargeState, `MME (ppm)`=mme.ppm) %>%
    mutate(spectrumFile = sub('_msgfplus.mzid.gz','',spectrumFile))
library(xtable)
print(xtable(meta_gluc_hprp, display = c('d','s','e','f','s','d','s')),
      include.rownames=FALSE,
      comment = FALSE,
      size='scriptsize',
      floating = F)
```

spectrumFile	SpecEValue	MME (ppm)	spectrumID	charge	peptide
Euplotes_1_HPRP_2_06_22Nov09_Falcon_09-09-15	6.80e-07	2.95	index=13369	2	E.NFNKITGKEQEEEE.Y
Euplotes_1_HPRP_2_08_25Nov09_Falcon_09-09-15	3.78e-17	0.19	index=9982	3	E.SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE.K
Euplotes_1_HPRP_2_08_25Nov09_Falcon_09-09-15	3.33e-07	0.57	index=9974	4	E.SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE.K
Euplotes_1_HPRP_2_09_17Nov09_Falcon_09-09-17	5.74e-16	0.44	index=10771	3	E.NLDNEKLINDLTNDKANLKDIVFDLMFE.K
Euplotes_1_HPRP_2_09_17Nov09_Falcon_09-09-17	5.03e-07	1.11	index=10770	4	E.NLDNEKLINDLTNDKANLKDIVFDLMFE.K
Euplotes_1_HPRP_2_12_17Nov09_Falcon_09-09-17	2.09e-09	0.43	index=3933	3	E.NKIRFFAAPEKIFE.T
Euplotes_1_HPRP_2_12_17Nov09_Falcon_09-09-17	1.62e-07	0.07	index=3930	2	E.NKIRFFAAPEKIFE.T
Euplotes_1_HPRP_2_15_17Nov09_Falcon_09-09-17	2.83e-07	1.61	index=1758	2	E.MQDEEILKSIEESKLEQEQEEEKKNE.E
Euplotes_1_HPRP_2_21_22Nov09_Falcon_09-09-17	2.17e-07	0.10	index=6671	1	E.VYLGLMEEYE.A
Euplotes_1_HPRP_2_22_22Nov09_Falcon_09-09-17	2.12e-08	0.88	index=6753	1	E.VYLGLMEEYE.A

2.3 Compendium of Peptides Covering Frameshift Locations

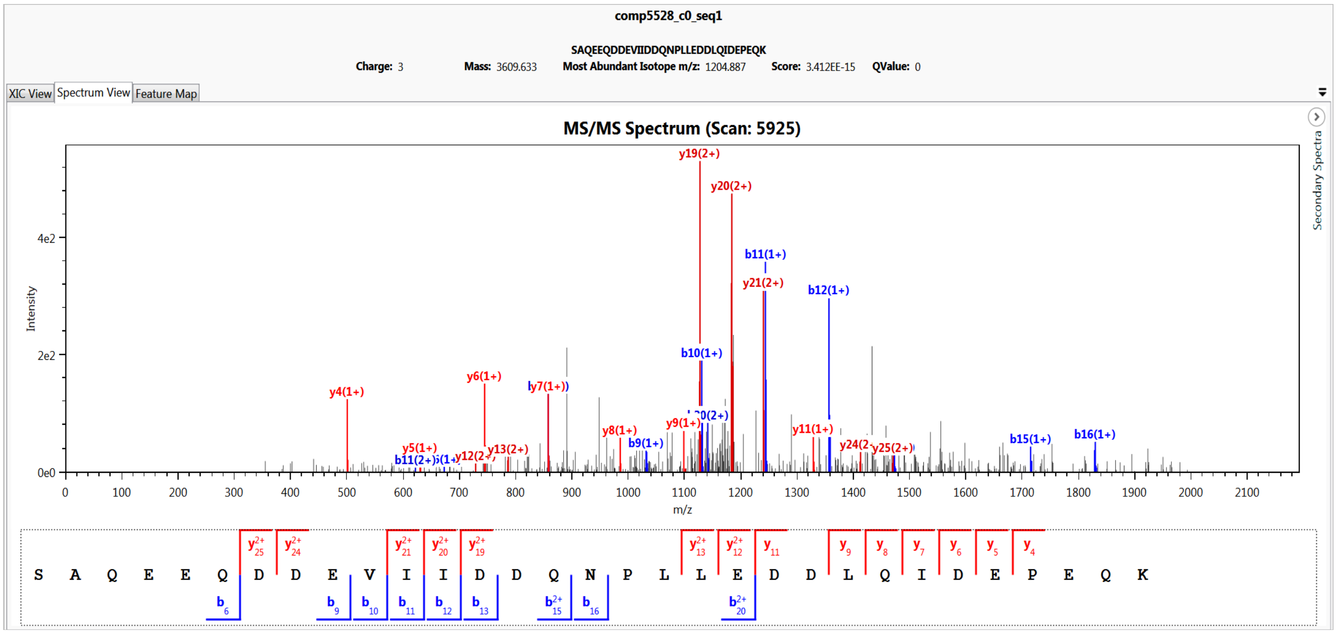
Final set of peptides and corresponding references to LC-MS/MS datasets and spectra. Overall, **4**, **11**, and **6** unique peptide sequences spanning the frameshift sites were identified in `trypsin/SCX`, `trypsin/HPRP`, and 'Glu-C/HPRP‘ experiments, respectively.

spectrumFile	SpecEValue	MME (ppm)	spectrumID	charge	peptide	experiment
Euplotes_1_SCX_10_13Nov09_Falcon_09-09-14	3.41e-15	0.30	index=6106	3	K.SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK.V	trypsin/SCX
Euplotes_1_SCX_10_13Nov09_Falcon_09-09-14	3.41e-15	0.30	index=6106	3	K.SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK.V	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	1.53e-21	0.08	index=8908	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	1.07e-20	1.10	index=8896	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	7.29e-19	1.10	index=8897	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_12_13Nov09_Falcon_09-09-14	2.17e-15	0.94	index=8895	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_18_13Nov09_Falcon_09-09-15	9.27e-17	0.11	index=5912	2	K.ESNHNNDITNKNEIAYILR.Y	trypsin/SCX
Euplotes_1_SCX_20_13Nov09_Falcon_09-09-15	2.23e-11	0.70	index=10317	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_22_13Nov09_Falcon_09-09-15	4.36e-10	3.76	index=9720	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_23_13Nov09_Falcon_09-09-15	2.47e-09	1.64	index=9440	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/SCX
Euplotes_1_SCX_24_13Nov09_Falcon_09-09-15	3.42e-10	8.85	index=2127	3	R.KKKQEENNLKR.K	trypsin/SCX
Euplotes_1_HPRP_1_04_17Nov09_Falcon_09-09-14	7.58e-11	0.08	index=3031	1	R.FFAAPEK.I	trypsin/HPRP
Euplotes_1_HPRP_1_04_17Nov09_Falcon_09-09-14	2.44e-09	0.00	index=3046	2	R.FFAAPEK.I	trypsin/HPRP
Euplotes_1_HPRP_1_05_17Nov09_Falcon_09-09-14	1.46e-09	5.31	index=8245	3	R.ELAFLKRAQEIGLEPYNEYHGKKK.T	trypsin/HPRP
Euplotes_1_HPRP_1_06_17Nov09_Falcon_09-09-14	5.54e-10	2.21	index=759	2	K.VVQEGNTNVKK.L	trypsin/HPRP
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	5.93e-22	2.11	index=8644	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	2.18e-21	0.78	index=8638	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	3.05e-21	2.11	index=8646	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_08_17Nov09_Falcon_09-09-14	4.19e-16	0.82	index=8639	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.19e-21	0.70	index=8806	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	1.20e-21	1.57	index=8812	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	5.49e-20	1.64	index=8802	2	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_09_17Nov09_Falcon_09-09-14	4.33e-15	1.53	index=8810	3	R.WTPIDLPSEEITFVQGIQTVTGAGDPSMK.A	trypsin/HPRP
Euplotes_1_HPRP_1_16_22Nov09_Falcon_09-09-14	4.51e-21	0.33	index=10684	2	K.IIQNFQINTVFEDLDEIMQTQVQR.H	trypsin/HPRP
Euplotes_1_HPRP_1_16_22Nov09_Falcon_09-09-14	1.36e-11	1.25	index=10678	3	K.IIQNFQINTVFEDLDEIMQTQVQR.H	trypsin/HPRP
Euplotes_1_HPRP_1_18_17Nov09_Falcon_09-09-15	5.08e-09	2.64	index=13785	2	K.KSSKACEEERRKR.E	trypsin/HPRP
Euplotes_1_HPRP_1_20_17Nov09_Falcon_09-09-15	1.91e-11	0.00	index=3425	1	K.LINDLTNDK.A	trypsin/HPRP
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	6.65e-11	1.67	index=3600	2	K.LISELTSEK.S	trypsin/HPRP
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	2.55e-10	0.78	index=3602	1	K.LISELTSEK.S	trypsin/HPRP
Euplotes_1_HPRP_1_22_17Nov09_Falcon_09-09-15	1.89e-09	0.49	index=2595	2	K.IVENFNK.I	trypsin/HPRP
Euplotes_1_HPRP_1_23_17Nov09_Falcon_09-09-15	3.01e-13	1.01	index=2200	2	K.LSQEHLSYISR.L	trypsin/HPRP
Euplotes_1_HPRP_1_24_17Nov09_Falcon_09-09-15	2.45e-16	1.41	index=2709	2	K.LINDLTNDKANLK.D	trypsin/HPRP
Euplotes_1_HPRP_2_06_22Nov09_Falcon_09-09-15	6.80e-07	2.95	index=13369	2	E.NFNKITGKEQEEEE.Y	Glu-C/HPRP
Euplotes_1_HPRP_2_08_25Nov09_Falcon_09-09-15	3.78e-17	0.19	index=9982	3	E.SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE.K	Glu-C/HPRP
Euplotes_1_HPRP_2_08_25Nov09_Falcon_09-09-15	3.33e-07	0.57	index=9974	4	E.SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE.K	Glu-C/HPRP
Euplotes_1_HPRP_2_09_17Nov09_Falcon_09-09-17	5.74e-16	0.44	index=10771	3	E.NLDNEKLINDLTNDKANLKDIVFDLMFE.K	Glu-C/HPRP
Euplotes_1_HPRP_2_09_17Nov09_Falcon_09-09-17	5.03e-07	1.11	index=10770	4	E.NLDNEKLINDLTNDKANLKDIVFDLMFE.K	Glu-C/HPRP
Euplotes_1_HPRP_2_12_17Nov09_Falcon_09-09-17	2.09e-09	0.43	index=3933	3	E.NKIRFFAAPEKIFE.T	Glu-C/HPRP
Euplotes_1_HPRP_2_12_17Nov09_Falcon_09-09-17	1.62e-07	0.07	index=3930	2	E.NKIRFFAAPEKIFE.T	Glu-C/HPRP
Euplotes_1_HPRP_2_15_17Nov09_Falcon_09-09-17	2.83e-07	1.61	index=1758	2	E.MQDEEILKSIEESKLEQEQEEEKKNE.E	Glu-C/HPRP
Euplotes_1_HPRP_2_21_22Nov09_Falcon_09-09-17	2.17e-07	0.10	index=6671	1	E.VYLGLMEEYE.A	Glu-C/HPRP
Euplotes_1_HPRP_2_22_22Nov09_Falcon_09-09-17	2.12e-08	0.88	index=6753	1	E.VYLGLMEEYE.A	Glu-C/HPRP

3 Manual Validation

Manual valiation was perfomed by LCMSSpectator. The spectra that have passed the consensus opinion of 5 independed experts are shown below. Necessary raw and mzIdenML files to reproduce the analysis are available at http://dx.doi.org/10.6019/PXD004333. Note, the `MS/MS scan number` is not the same identifier as `spectrumID` in the table above.

SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK

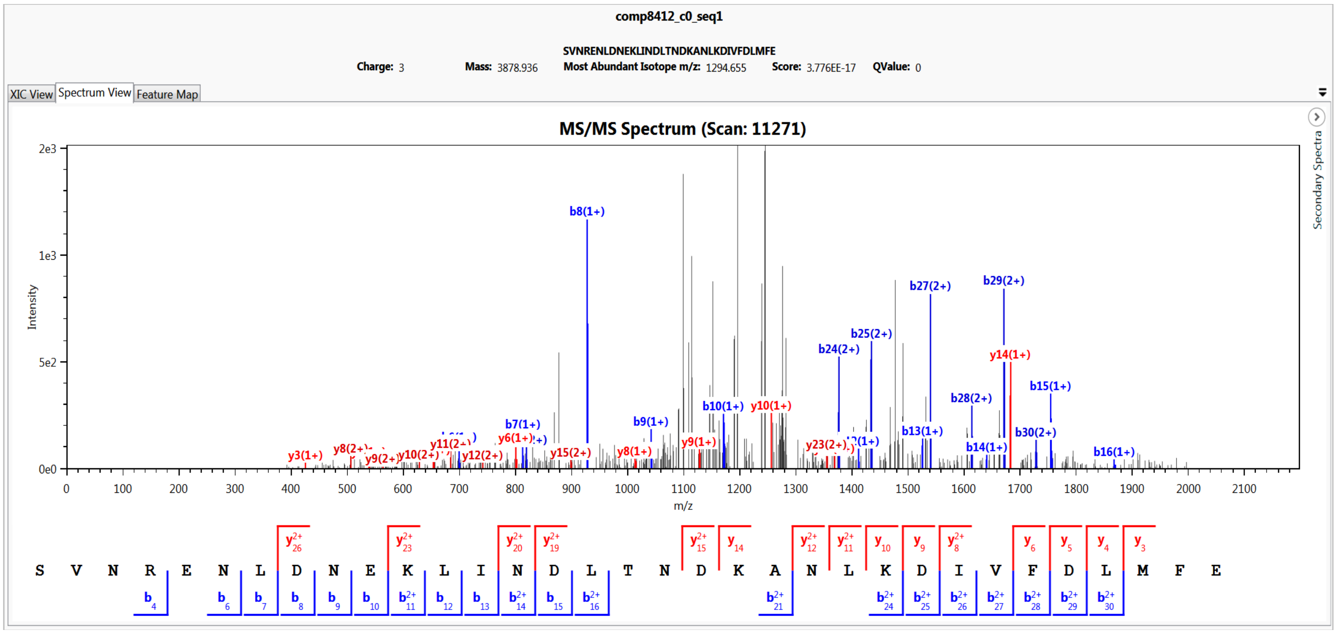

comp5528_c0_seq1

SAQEEQDDEVIIDDQNPLLEDDLQIDEPEQK

Charge: 3 Mass: 3609.633 Most Abundant Isotope m/z: 1204.887 Score: 3.412EE-15 QValue: 0

XIC View
Spectrum View
Feature Map
MS/MS Spectrum (Scan: 5925)
Secondary Spectra
Intensity
m/z
4e2
2e2
0e0
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
y19(2+)
y20(2+)
b11(1+)
y21(2+)
b12(1+)
b10(1+)
y6(1+)
y7(1+)
y4(1+)
y8(1+)
b9(1+)
y9(1+)
y11(1+)
y24(2+)
y25(2+)
b15(1+)
b16(1+)
y5(1+)
b11(2+)
y12(2+)
y13(2+)
S A Q E E Q D D E V I I D D Q N P L L E D D L Q I D E P E Q K
y^{2+}_{25} y^{2+}_{24} y^{2+}_{21} y^{2+}_{20} y^{2+}_{19} y^{2+}_{13} y^{2+}_{12} y_{11} y_9 y_8 y_7 y_6 y_5 y_4
b_6 b_9 b_{10} b_{11} b_{12} b_{13} b^{2+}_{15} b_{16} b^{2+}_{20}

SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE

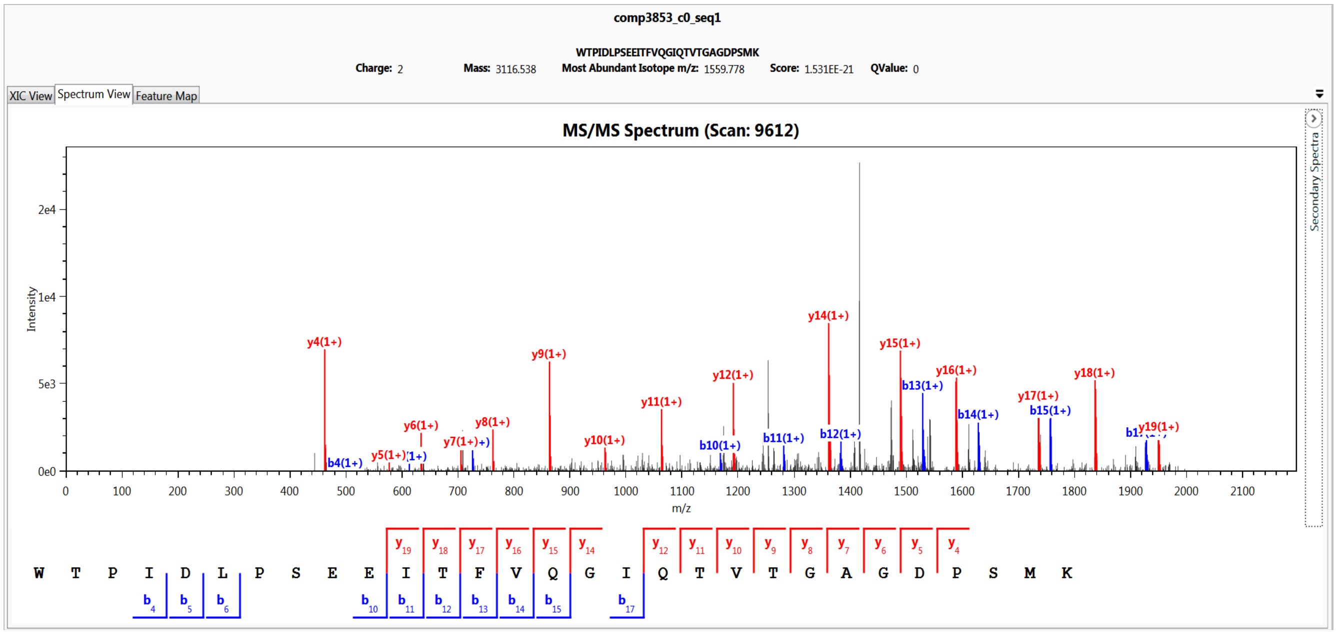

comp8412_c0_seq1

SVNRENLDNEKLINDLTNDKANLKDIVFDLMFE

Charge: 3 Mass: 3878.936 Most Abundant Isotope m/z: 1294.655 Score: 3.776EE-17 QValue: 0

XIC View
Spectrum View
Feature Map
MS/MS Spectrum (Scan: 11271)
Secondary Spectra
Intensity
m/z
2e3
1e3
5e2
0e0
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
b8(1+)
b27(2+)
b29(2+)
b25(2+)
b24(2+)
y14(1+)
b15(1+)
b28(2+)
b10(1+)
y10(1+)
b7(1+)
b9(1+)
y6(1+)
b13(1+)
b30(2+)
b14(1+)
b16(1+)
y3(1+)
y8(2+)
y9(2+)
y10(2+)
y11(2+)
y12(2+)
y15(2+)
y8(1+)
y9(1+)
y23(2+)
S V N R E N L D N E K L I N D L T N D K A N L K D I V F D L M F E
y^{2+}_{26} y^{2+}_{23} y^{2+}_{20} y^{2+}_{19} y^{2+}_{15} y_{14} y^{2+}_{12} y^{2+}_{11} y_{10} y_9 y^{2+}_8 y_6 y_5 y_4 y_3
b_4 b_6 b_7 b_8 b_9 b_{10} b^{2+}_{11} b_{12} b_{13} b^{2+}_{14} b_{15} b^{2+}_{16} b^{2+}_{21} b^{2+}_{24} b^{2+}_{25} b^{2+}_{26} b^{2+}_{27} b^{2+}_{28} b^{2+}_{29} b^{2+}_{30}

WTPIDLPSEEITFVQGIQTVTGAGDPSMK

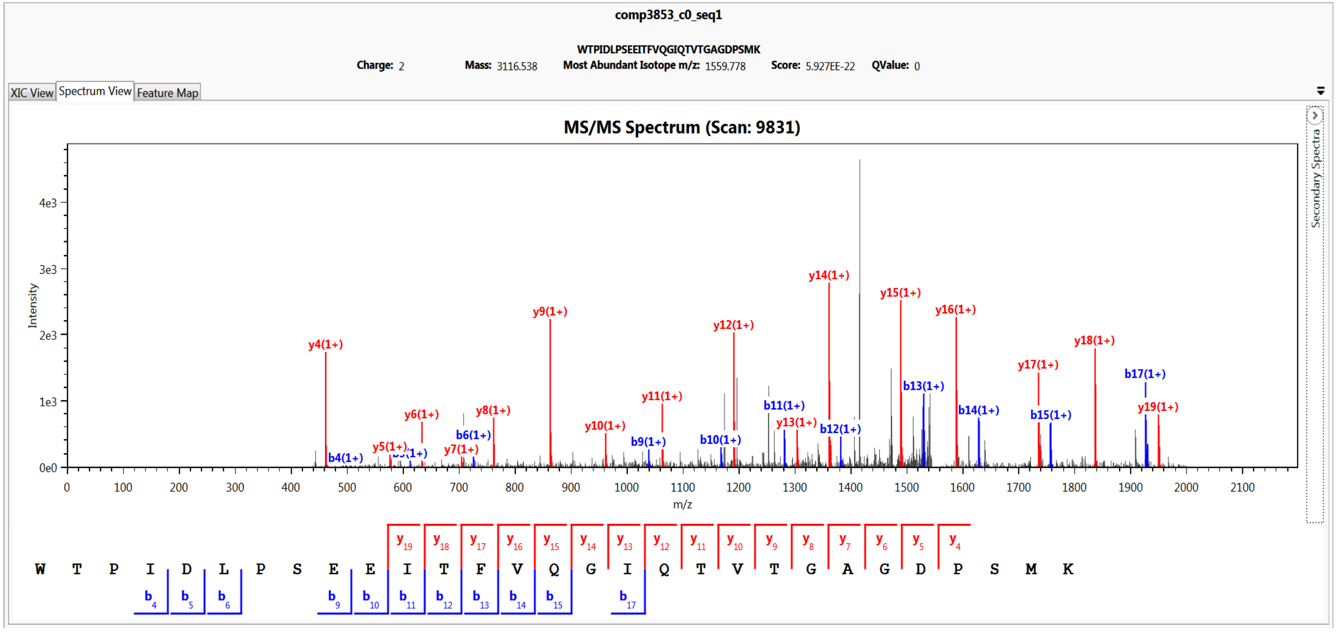


comp3853_c0_seq1

WTPIDLPSEEITFVQGIQTVTGAGDPSMK

Charge: 2 Mass: 3116.538 Most Abundant Isotope m/z: 1559.778 Score: 1.531EE-21 QValue: 0

MS/MS Spectrum (Scan: 9612)



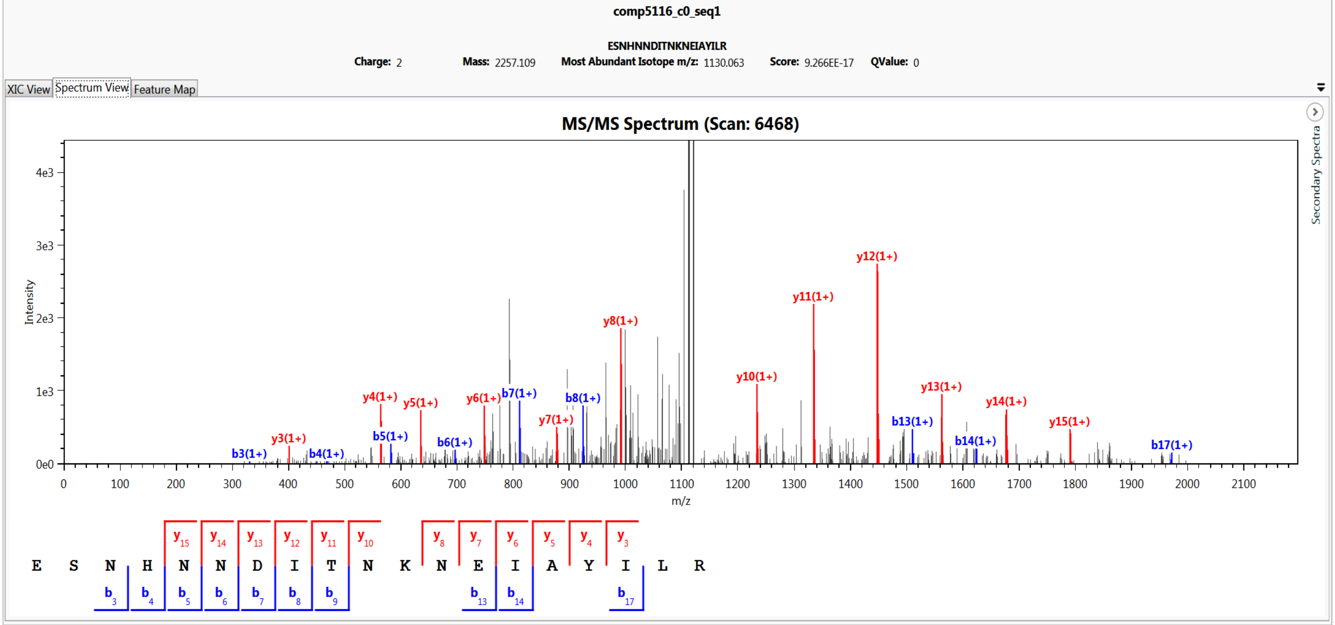
comp3853_c0_seq1

WTPIDLPSEEITFVQGIQTVTGAGDPSMK

Charge: 2 Mass: 3116.538 Most Abundant Isotope m/z: 1559.778 Score: 5.927EE-22 QValue: 0

MS/MS Spectrum (Scan: 9831)

ESNHNNDITNKNEIAYILR

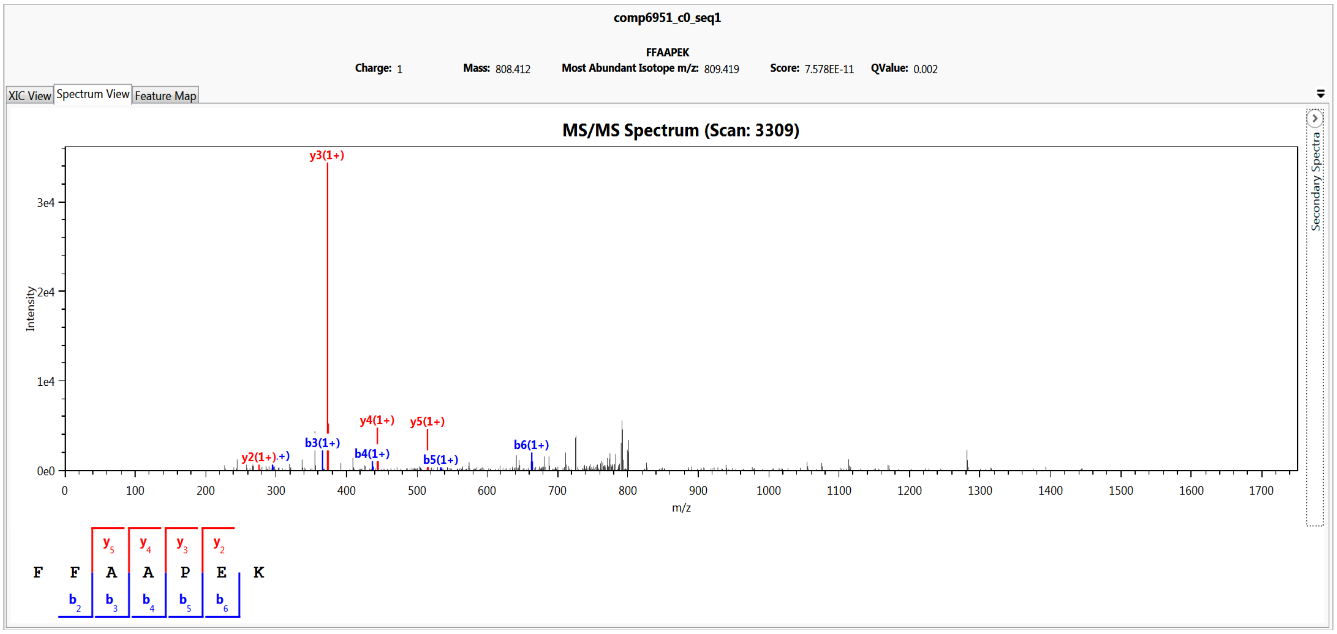


comp5116_c0_seq1

ESNHNNDITNKNEIAYILR

Charge: 2 **Mass:** 2257.109 **Most Abundant Isotope m/z:** 1130.063 **Score:** 9.266EE-17 **QValue:** 0

FFAAPEK

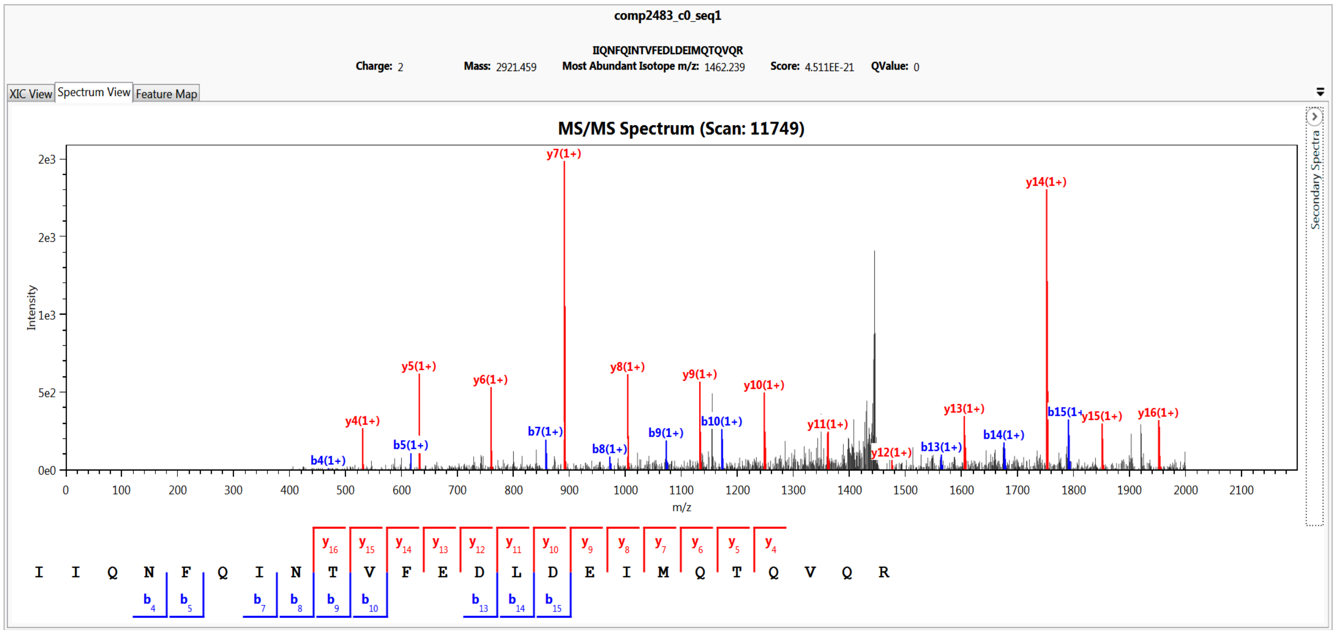


comp6951_c0_seq1

FFAAPEK

Charge: 1 **Mass:** 808.412 **Most Abundant Isotope m/z:** 809.419 **Score:** 7.578EE-11 **QValue:** 0.002

IIQNFQINTVFEDLDEIMQTQVQR

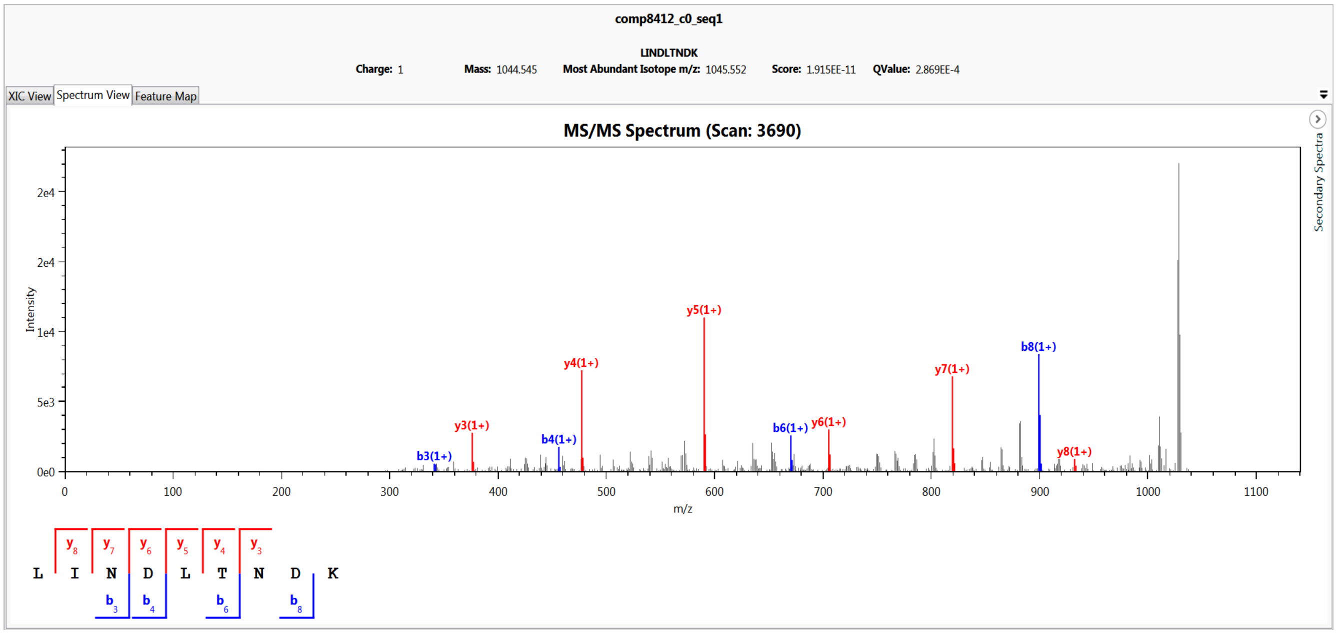


comp2483_c0_seq1

IIQNFQINTVFEDLDEIMQTQVQR

Charge: 2 **Mass:** 2921.459 **Most Abundant Isotope m/z:** 1462.239 **Score:** 4.511EE-21 **QValue:** 0

LINDLTNDK

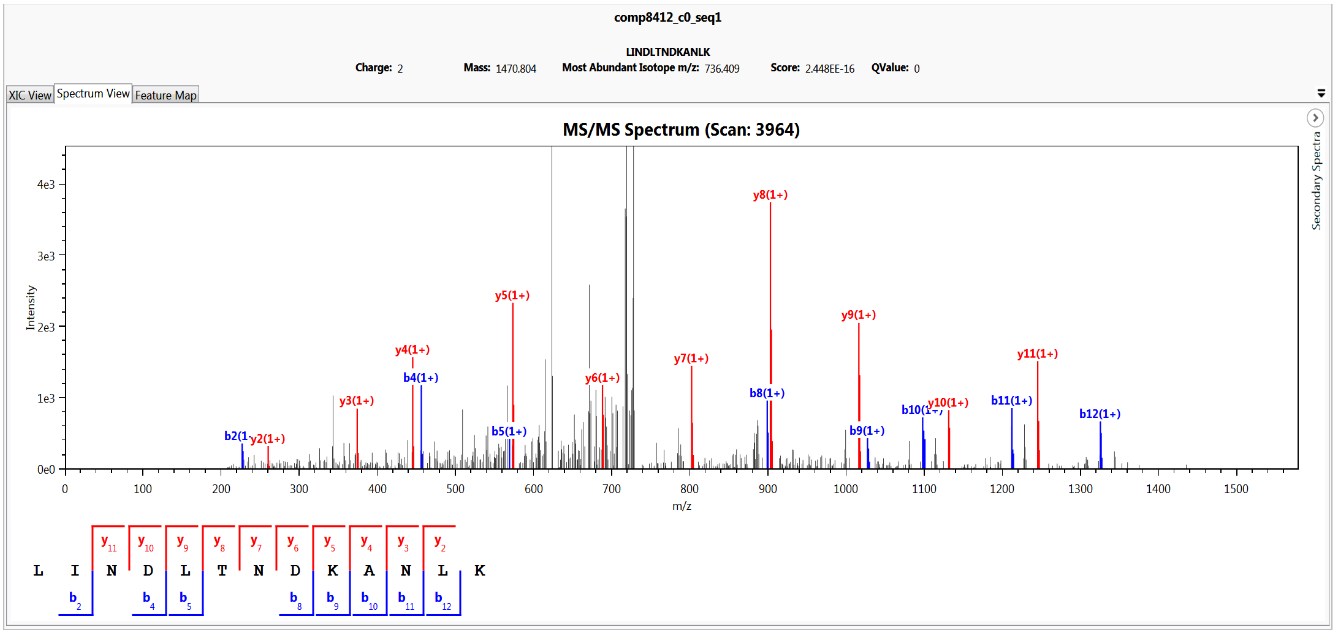


comp8412_c0_seq1

LINDLTNDK

Charge: 1 **Mass:** 1044.545 **Most Abundant Isotope m/z:** 1045.552 **Score:** 1.915EE-11 **QValue:** 2.869EE-4

LINDLTNDKANLK

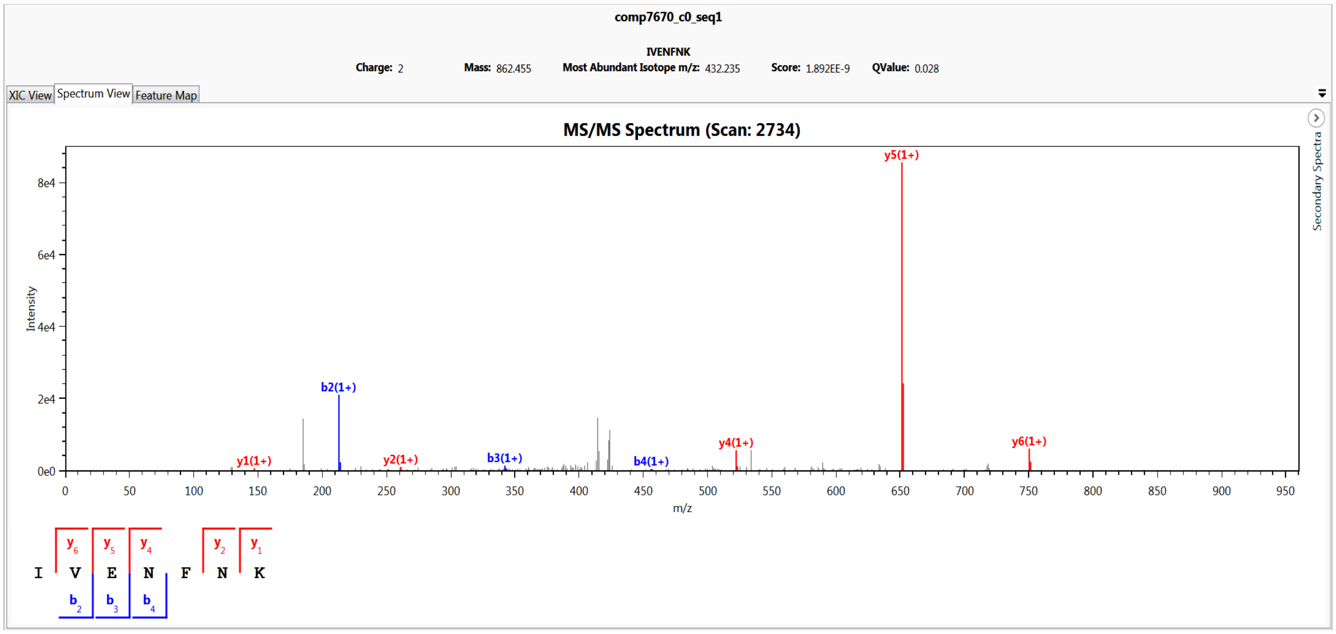


comp8412_c0_seq1

LINDLTNDKANLK

Charge: 2 **Mass:** 1470.804 **Most Abundant Isotope m/z:** 736.409 **Score:** 2.448EE-16 **QValue:** 0

IVENFNK

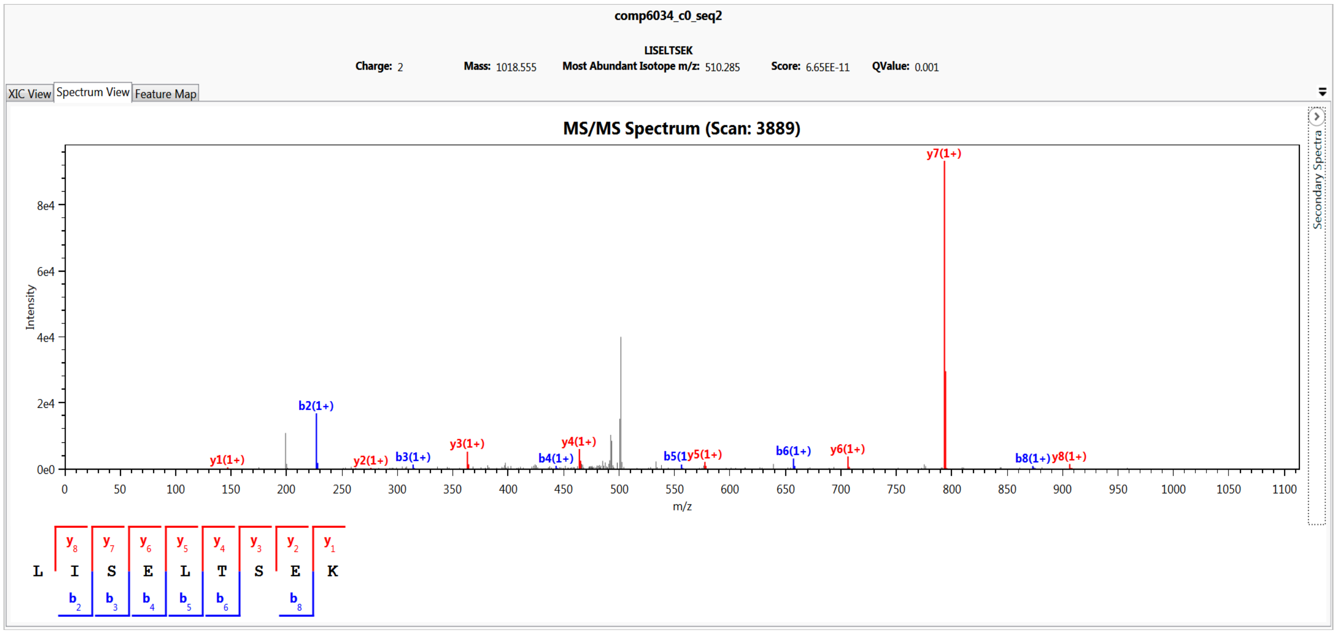


comp7670_c0_seq1

IVENFNK

Charge: 2 **Mass:** 862.455 **Most Abundant Isotope m/z:** 432.235 **Score:** 1.892EE-9 **QValue:** 0.028

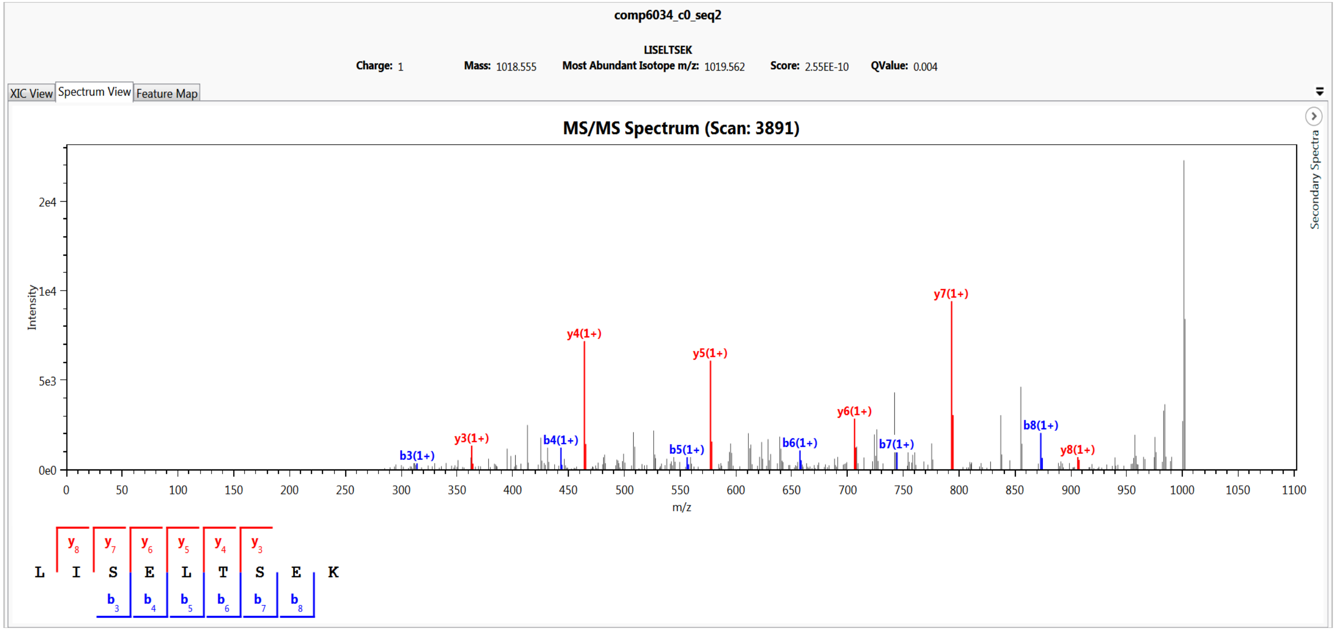
LISELTSEK



comp6034_c0_seq2

LISELTSEK

Charge: 2 **Mass:** 1018.555 **Most Abundant Isotope m/z:** 510.285 **Score:** 6.65EE-11 **QValue:** 0.001

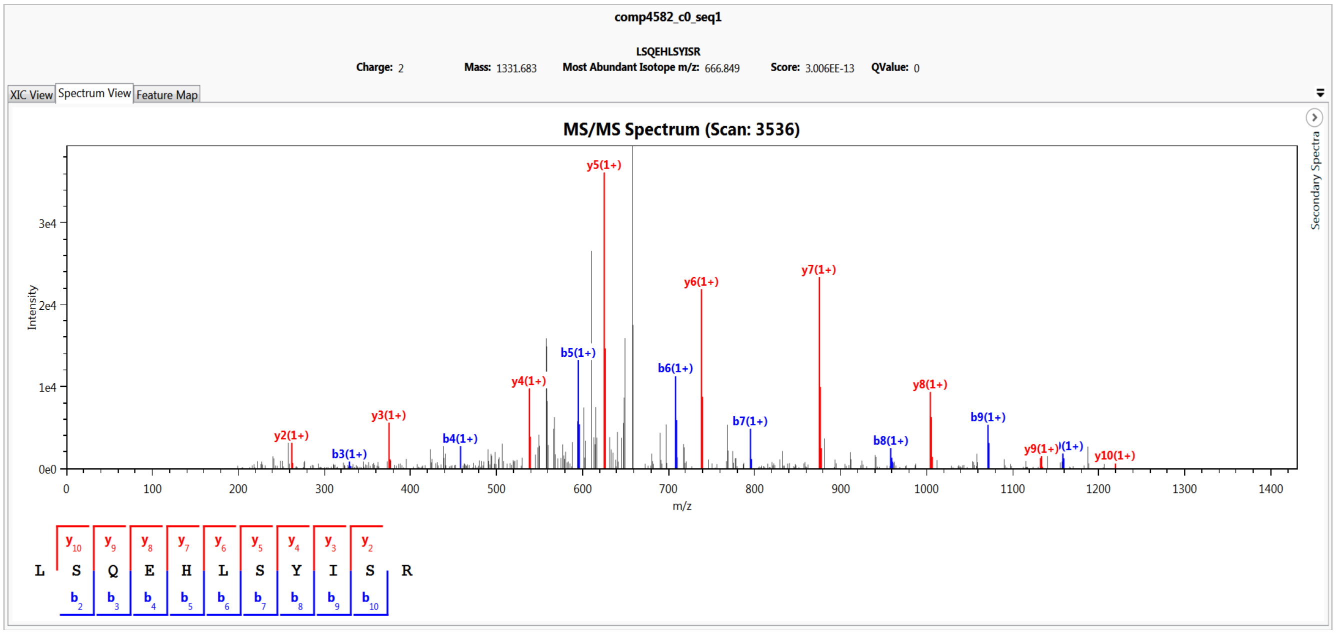


comp6034_c0_seq2

LISELTSEK

Charge: 1 **Mass:** 1018.555 **Most Abundant Isotope m/z:** 1019.562 **Score:** 2.55EE-10 **QValue:** 0.004

LSQEHLSYISR

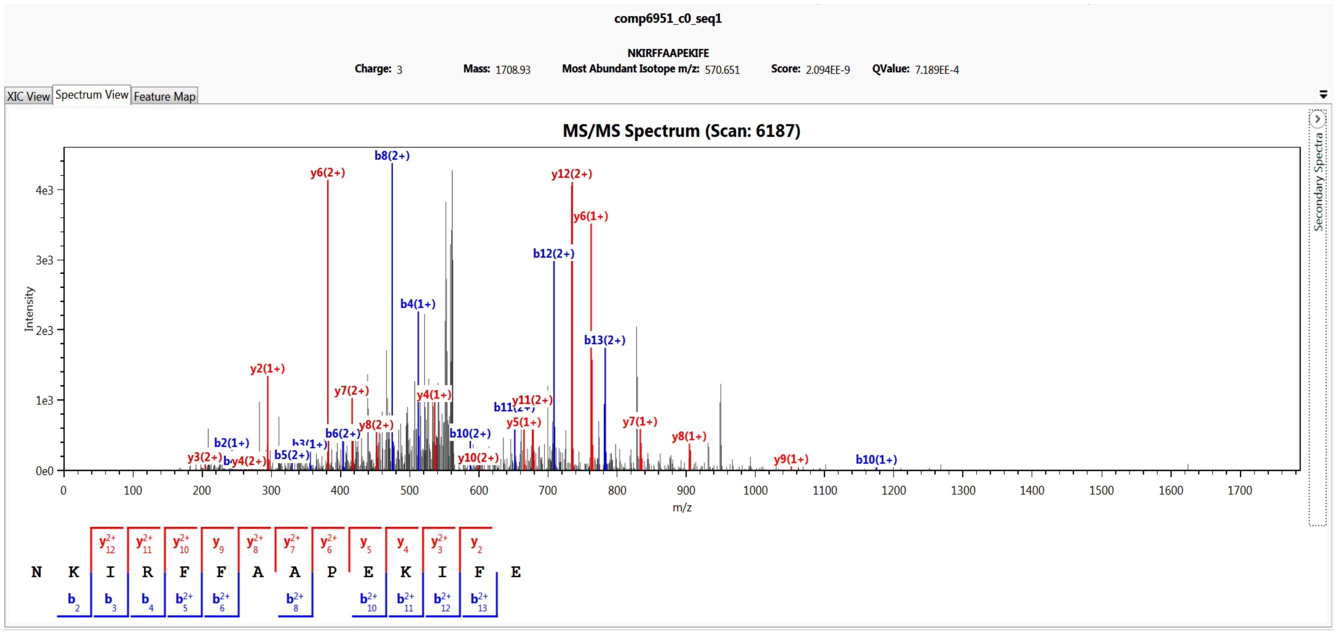


comp4582_c0_seq1

LSQEHLSYISR

Charge: 2 **Mass:** 1331.683 **Most Abundant Isotope m/z:** 666.849 **Score:** 3.006EE-13 **QValue:** 0

NKIRFFAAPEKIFE

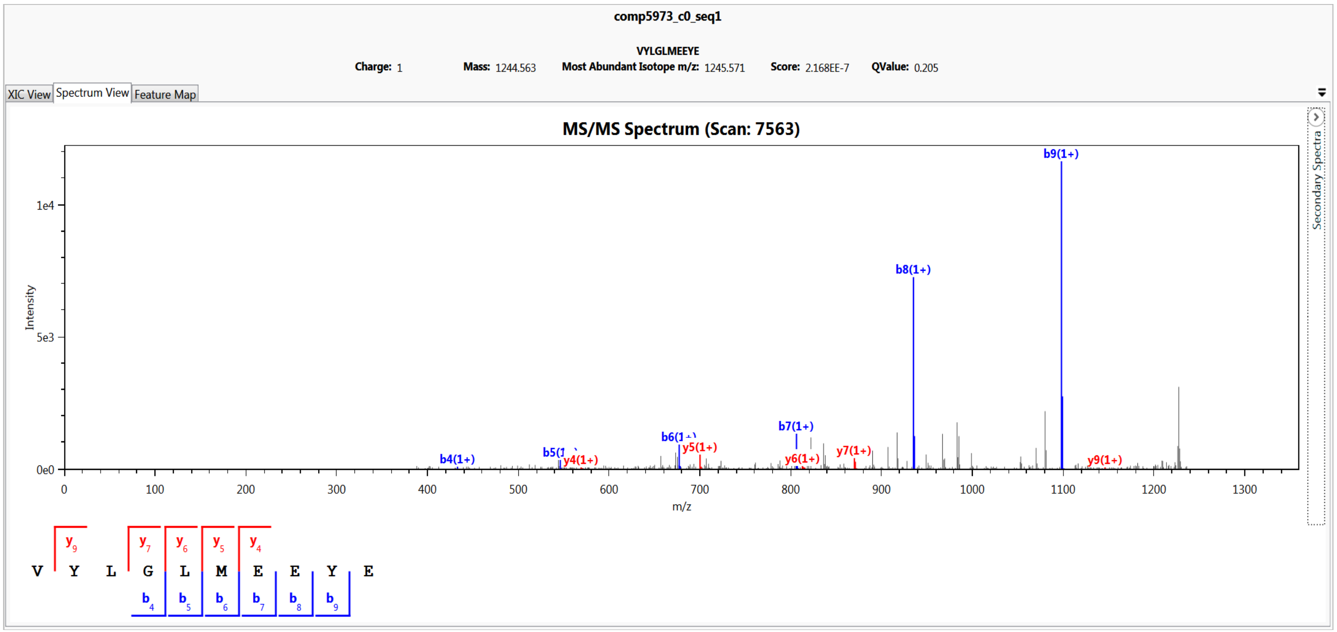


comp6951_c0_seq1

NKIRFFAAPEKIFE

Charge: 3 **Mass:** 1708.93 **Most Abundant Isotope m/z:** 570.651 **Score:** 2.094EE-9 **QValue:** 7.189EE-4

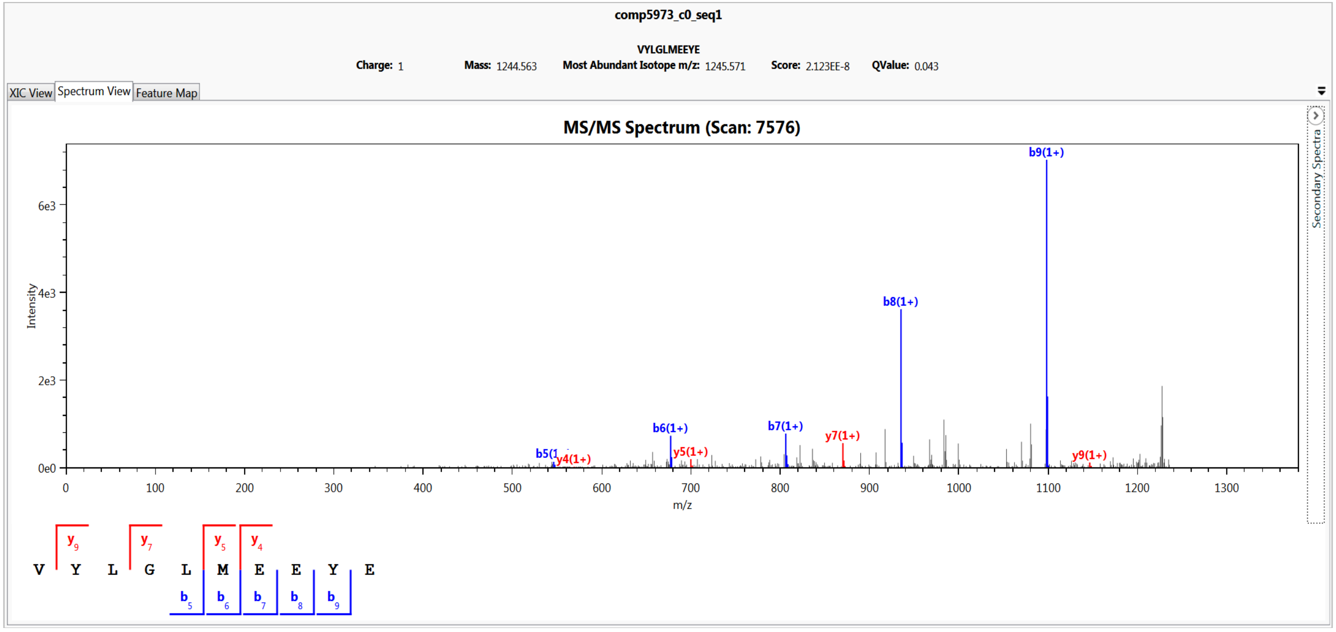
VYLGLMEEYE



comp5973_c0_seq1

VYLGLMEEYE

Charge: 1 **Mass:** 1244.563 **Most Abundant Isotope m/z:** 1245.571 **Score:** 2.168EE-7 **QValue:** 0.205



comp5973_c0_seq1

VYLGLMEEYE

Charge: 1 **Mass:** 1244.563 **Most Abundant Isotope m/z:** 1245.571 **Score:** 2.123EE-8 **QValue:** 0.043

4 Session Information

All software and respective versions used in this document, as returned by sessionInfo() are detailed below.

- R version 3.2.4 (2016-03-10), `x86_64-apple-darwin13.4.0`
- Locale: `C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils
- Other packages: BiocGenerics 0.16.1, BiocStyle 1.8.0, Biostrings 2.38.4, IRanges 2.4.8, MSnID 1.7.3, Rcpp 0.12.4, S4Vectors 0.8.11, XVector 0.10.0, dplyr 0.4.3.9000, knitr 1.12.3, rpx 1.6.0, xtable 1.8-2
- Loaded via a namespace (and not attached): Biobase 2.30.0, BiocInstaller 1.20.1, BiocParallel 1.4.3, DBI 0.3.1, MALDIquant 1.14, MSnbase 1.18.1, ProtGenerics 1.2.1, R.cache 0.12.0, R.methodsS3 1.7.1, R.oo 1.20.0, R.utils 2.3.0, R6 2.1.2, RCurl 1.95-4.8, XML 3.98-1.4, affy 1.48.0, affyio 1.40.0, assertthat 0.1, bitops 1.0-6, chron 2.3-47, codetools 0.2-14, colorspace 1.2-6, compiler 3.2.4, data.table 1.9.6, digest 0.6.9, doParallel 1.0.10, evaluate 0.8.3, foreach 1.4.3, formatR 1.3, futile.logger 1.4.1, futile.options 1.0.0, ggplot2 2.1.0, grid 3.2.4, gtable 0.2.0, highr 0.5.1, htmltools 0.3.5, impute 1.44.0, iterators 1.0.8, lambda.r 1.1.7, lattice 0.20-33, lazyeval 0.1.10, limma 3.26.9, magrittr 1.5, munsell 0.4.3, mzID 1.8.0, mzR 2.4.1, pcaMethods 1.60.0, plyr 1.8.3, preprocessCore 1.32.0, reshape2 1.4.1, rmarkdown 0.9.5, scales 0.4.0, stringi 1.1.1, stringr 1.0.0, tools 3.2.4, vsn 3.38.0, yaml 2.1.13, zlibbioc 1.16.0