

TRGN 527: Applied Data Science and Bioinformatics

UNIT I. Introduction and Basic Data Science

Week 4 – Lecture 1 – Case Study Part 1

Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S. | Assistant Professor

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*

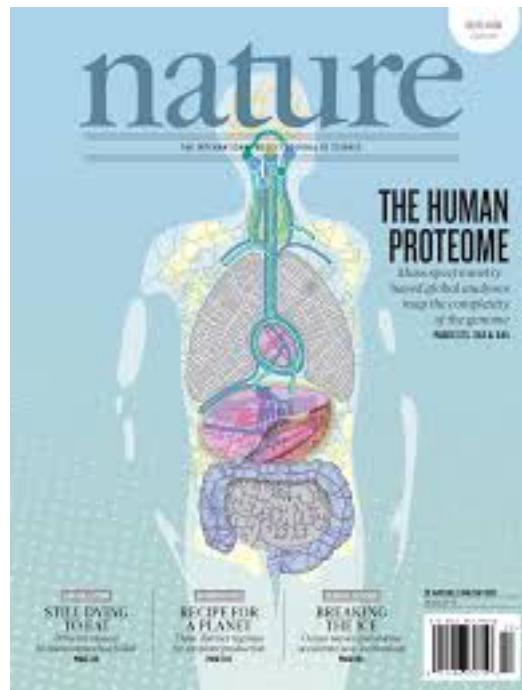
David W. Craig, Ph.D. | Professor and Vice Chair

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Co-Director, Institute of Translational Genomics

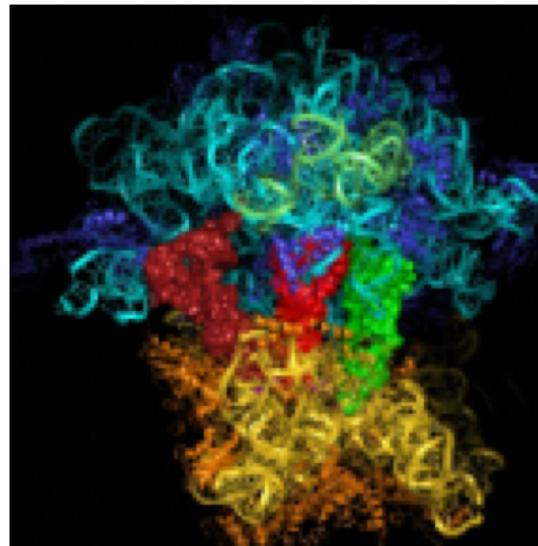
Case Study – Proteomics Analysis



Case Study – Proteomics Analysis

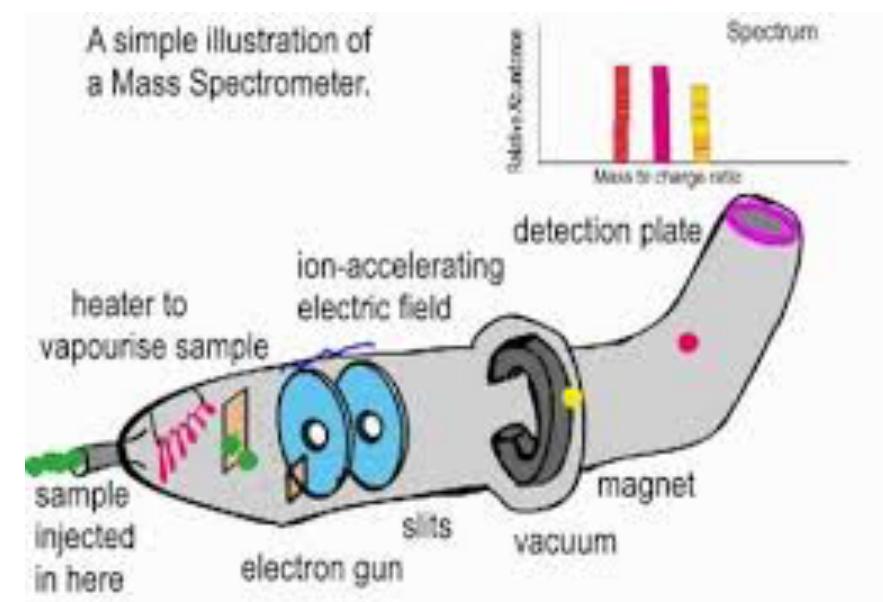
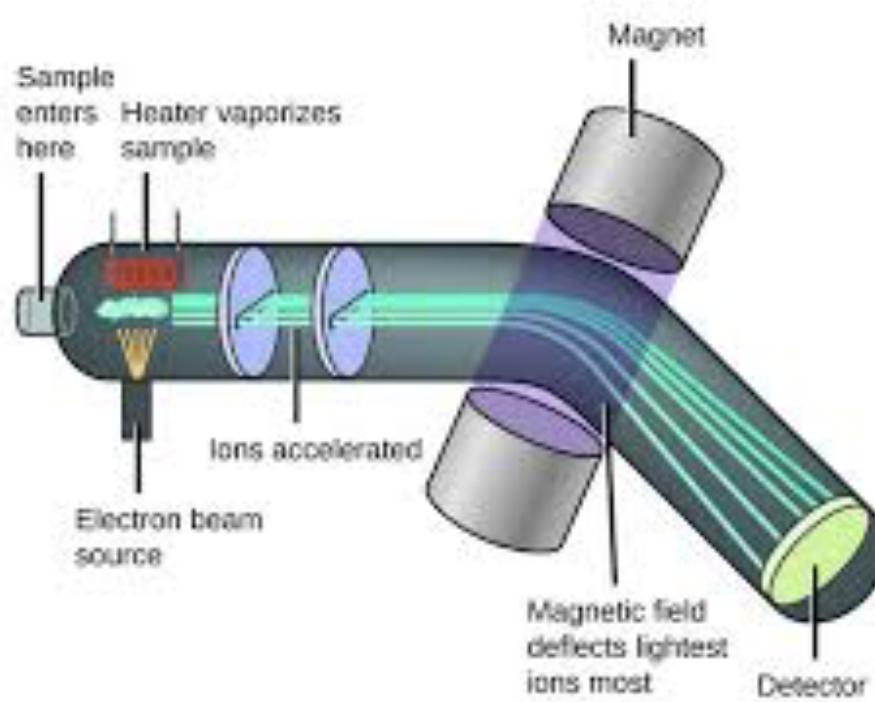
- Introduction

- It is crucial to study cellular processes also at the level of proteome and to measure the abundance of the proteins in biological samples.
- The segment of molecular biology that investigates the behavior of proteins is called “proteomics”.
- There are several methods in this field that aim to measure the abundance of proteins in parallel in different cellular systems.
- Mass spectrometry (MS) is one of the essential approaches that is capable of the identification of proteins, sometimes even with quantitative results.



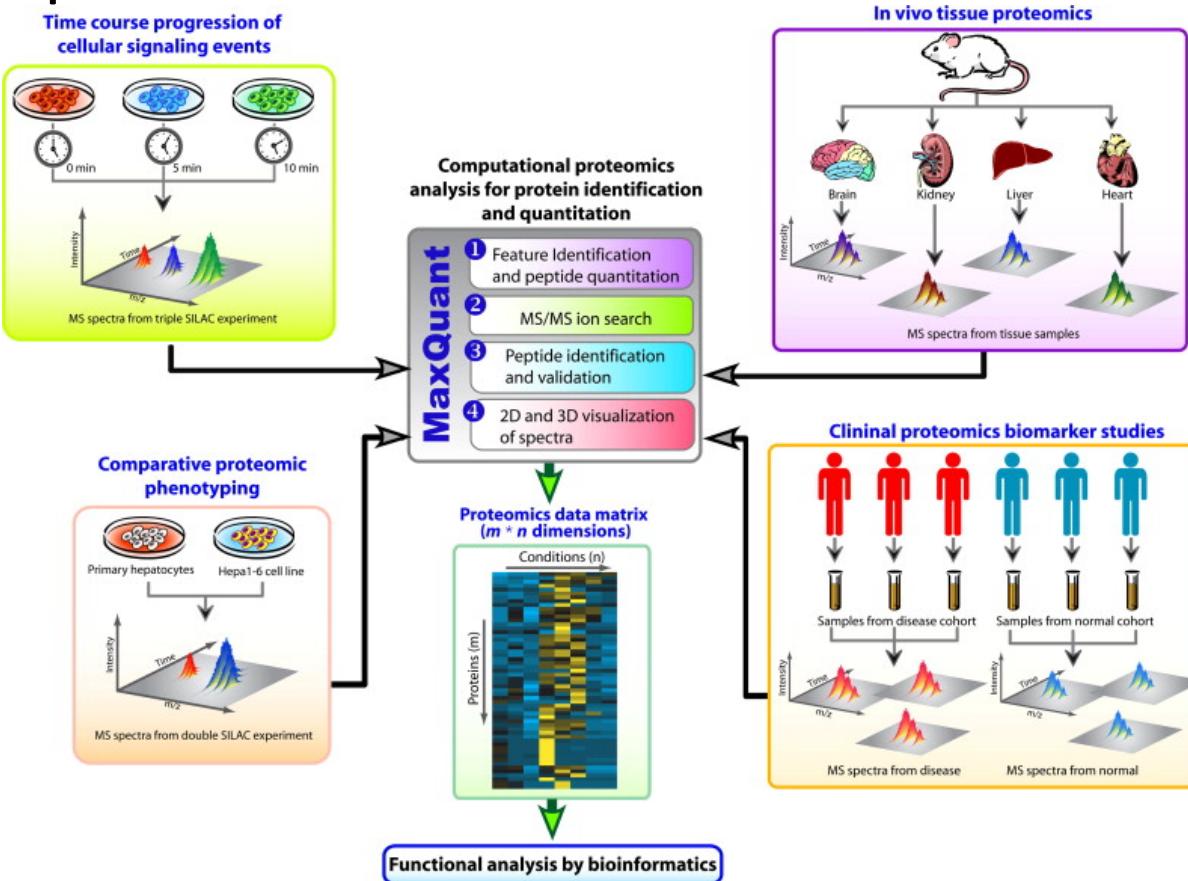
Case Study – Proteomics Analysis

- Proteomics: Mass spectrometry (MS)
- MS is an analytic technique that produces the mass-to-charge ratio spectra of peptide fragment in samples.
- Typically samples contain mainly purified proteins or mixtures representing biological specimens.



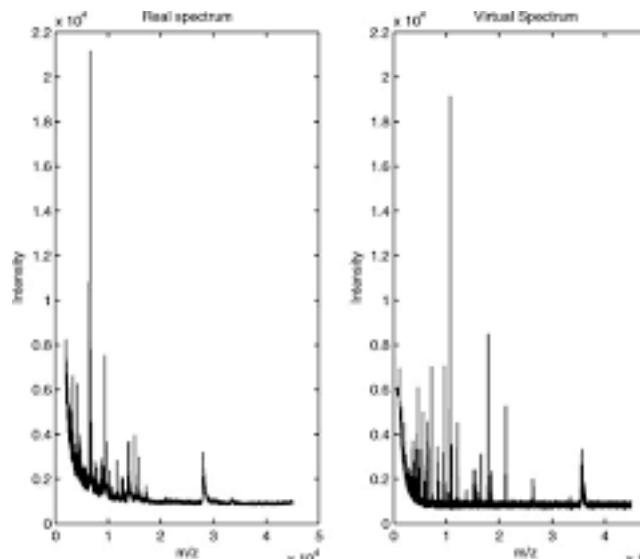
Case Study – Proteomics Analysis

- The GOAL of the analysis is to identify (one or more) peptides from the original protein or proteins, because these peptides are used for the qualitative or quantitative determination of the proteins in the specimens, or sometimes, for the identification of the fine modifications (i.e., phosphorylation, glycosylation, etc...) of the original protein.



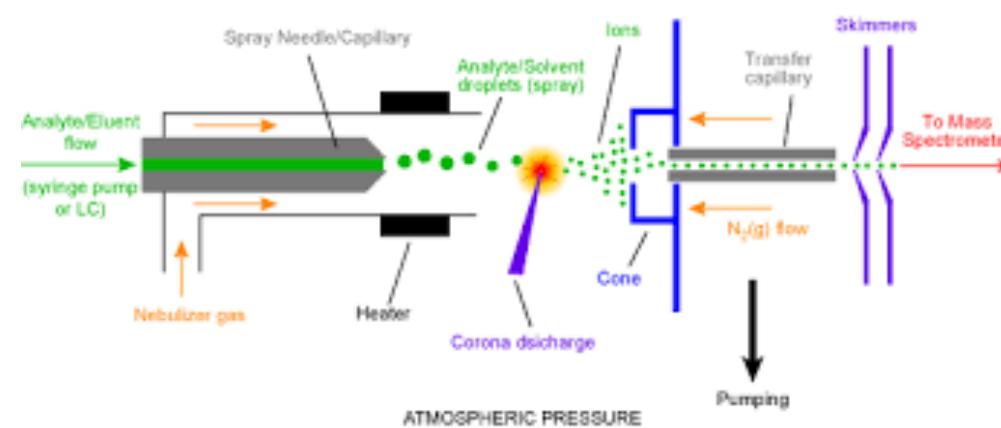
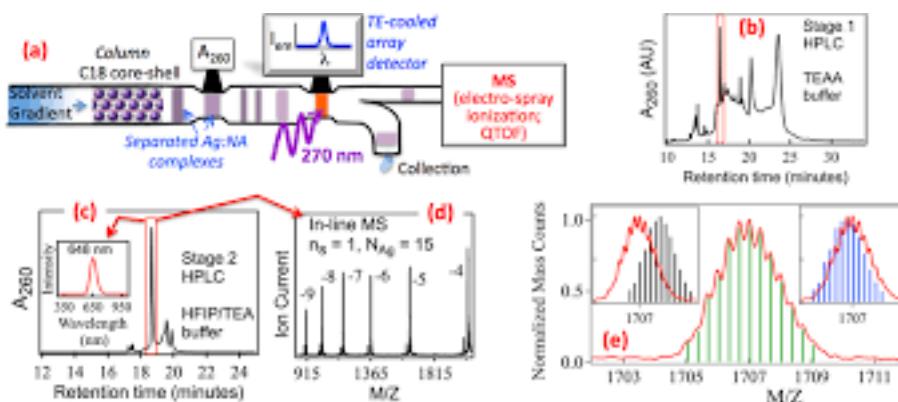
Case Study – Proteomics Analysis

- It is very important to consider the origin of the samples.
 - 1) Proteins are isolated from tissue samples or cell cultures according to the nature of the experiment.
 - 2) As these samples typically contain 5-40 thousand kinds of proteins, it would be hopeless to find differences without further separation.
 - 3) The most often used separation method is the two-dimensional polyacrylamide gel electrophoresis that distributes the proteins along two dimensions based on their isoelectric point and molecular mass.
 - 4) The spots on the gel contain a mixture of proteins that are cut and submitted to the procedure of MS.



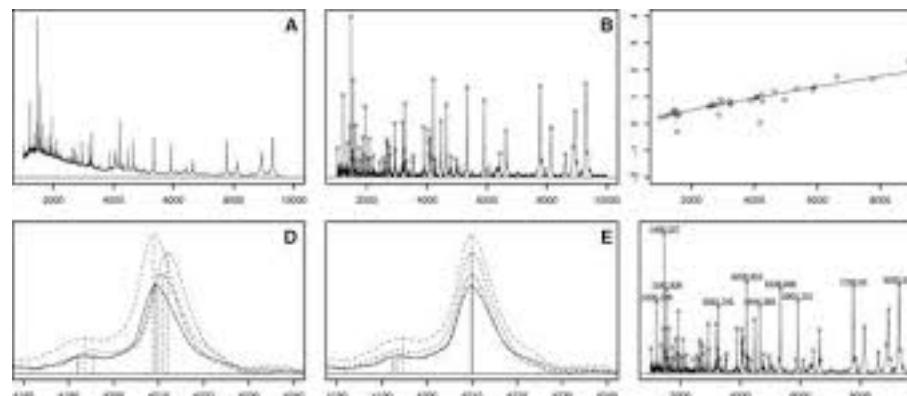
Case Study – Proteomics Analysis

- The proteins isolated from the 2D-gel or the samples loaded to the HPLC columns are digested with a protease enzyme
- MS always detects peptides—the fragments of the original proteins.
- The actual MS follows the preparation of separated peptide mixture.
- The spectrometer itself is built from three parts:
 - Ion source
 - Mass analyzer
 - Detector



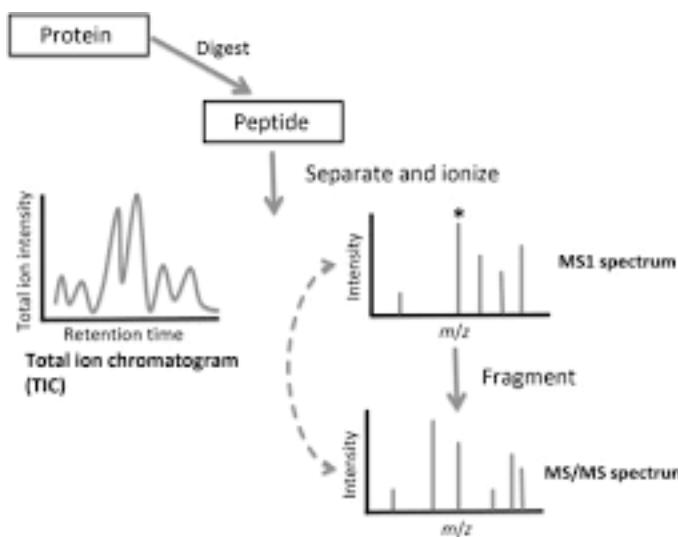
Case Study – Proteomics Analysis

- The Role of Data science/Bioinformatics in this kind of experiment is:
 - To Identify the peaks originating from peptides in the spectrum
 - To identify the source proteins of the peptides
 - Optionally to detect most translational modifications in the proteins
 - To find qualitative and quantitative differences between the biological samples



Case Study – Proteomics Analysis

- Dedicated software is available to analyze MS spectra, usually from the producers of the measurement equipment.
- Independent data analysis tools are available from Bioconductor packages.



Case Study – Proteomics Analysis

- File formats for MS data
 - The diversity formats known from the sequence segment of bioinformatics can also be seen in MS data.
 - The used file formats can be divided into two categories:
 - Files containing raw measurement data.
 - Files carrying processed data.
 - Standard formats for data exchange:
 - ANDI-MS
 - XML-based formats
 - mzR package from Bioconductor contains parsers for the following file formats:
 - netCDF
 - mzXML
 - mzData
 - mzML

Case Study – Proteomics Analysis

- Installing Packages:

```
1 - ---
2   title: "Rmarkdown_TRGN_599_Week_4_Proteomics_Analysis"
3   author: "Enrique I. Velazquez Villarreal, MD, PhD"
4   date: "1/25/2019"
5   output: html_document
6   ---
7
8   # Checking R working directory
9   ```{R}
10  # Check your current working directory
11  getwd()
12
13 # Set your working directory
14 setwd("/Users/enriquevelazquez/Documents/R_working_directory")
15 ```
16
17 # Install mzR
18
19 # To install this package, start R (version "3.5") and enter:
20 #     if (!requireNamespace("BiocManager", quietly = TRUE))
21 #         install.packages("BiocManager")
22 #     BiocManager::install("mzR", version = "3.8")
23
24 # Install msdata
25
26 #if (!requireNamespace("BiocManager", quietly = TRUE))
27 #    install.packages("BiocManager")
28 #BiocManager::install("msdata", version = "3.8")
29
30 # Install magrittr
31 # install.packages("magrittr")
32
33 # Accessing the raw data of published studies
```

Case Study – Proteomics Analysis

- Let us use the data:

```
35 ~ ``{R}
36 library(mzR)
37 library(msdata)
38 library(magrittr)
39
40 msfile <- system.file("threonine/threonine_i2_e35_pH_tree.mzXML",
41                         package = "msdata")
42
43 my.data <- openMSfile(msfile)
44
45 ````
```

- `openMSfile()` function can determine the file format, and interprets the details in the `my.data` variable.

Case Study – Proteomics Analysis

- It is important that after we have finished working on the data package, it would be closed:

```
49 - `` {R}
50   close(my.data)
51   ``
```

Case Study – Proteomics Analysis

- Accessing the metadata itself:

```
56  
57 # Accessing the measurement - metadata:  
58  
59 ``{r}  
60  
61 runInfo(my.data)  
62  
63 ...  
64  
65 ``{r}  
66 instrumentInfo(my.data)  
67  
68 ...  
69  
70 ``{r}  
71 header(my.data)  
72 ...  
--
```

Case Study – Proteomics Analysis

```
59: ````{r}  
60  
61 runInfo(my.data)  
62  
63 ````
```

```
$scanCount  
[1] 55
```

```
$lowMz  
[1] 50.0036
```

```
$highMz  
[1] 298.673
```

```
$dStartTime  
[1] 0.3485
```

```
$dEndTime  
[1] 390.027
```

```
$msLevels  
[1] 1 2 3 4
```

```
$startTimeStamp  
[1] NA
```

Case Study – Proteomics Analysis

```
65 ~ ` ``{r}
66 instrumentInfo(my.data)
67
68 ```

$manufacturer
[1] "Thermo Scientific"

$model
[1] "LTQ Orbitrap"

$ionisation
[1] "electrospray ionization"

$analyzer
[1] "fourier transform ion cyclotron resonance mass spectrometer"

$detector
[1] "unknown"

$software
[1] "Xcalibur software 2.2 SP1"

$sample
[1] ""

$source
[1] ""
```

Case Study – Proteomics Analysis

seqNum	acquisitionNum	msLevel	polarity	peaksCount	totIonCurrent	retentionTime	basePeakMZ	basePeakIntensity
1	1	1	1	684	341427000	0.3485	120.0660	211860000.0
2	2	2	2	432	160473000	5.8561	120.0660	139169000.0
3	3	3	2	1	340	58862000	12.5000	102.0560
4	4	4	3	1	273	30770400	19.7568	102.0550
5	5	5	3	1	238	5291800	26.6056	56.0497
6	6	6	3	1	140	14294000	34.0324	74.0605
7	7	7	3	1	186	1413070	40.8786	56.0498
8	8	8	4	1	78	1492820	48.3417	56.0497
9	9	9	4	1	278	2354340	55.3885	110.0830
10	10	10	4	1	317	2875910	62.7088	84.0448
11	11	11	4	1	300	2368190	70.0325	110.0850
12	12	12	1	1	798	313900000	77.2653	120.0660
13	13	13	2	1	378	113438000	84.3054	120.0660
14	14	14	2	1	370	63132600	90.9602	102.0560
15	15	15	3	1	273	28086600	98.2067	102.0550
16	16	16	3	1	227	4741460	105.0660	56.0497
17	17	17	3	1	167	15533800	112.4830	74.0604
18	18	18	3	1	148	1159100	119.3430	53.7828
19	19	19	4	1	88	1455980	126.7930	56.0497
20	20	20	4	1	220	1717640	133.8520	110.0850
21	21	21	4	1	337	2904410	141.1620	84.0448
22	22	22	4	1	301	2473480	148.4970	110.0850
23	23	23	1	1	760	298444000	155.7190	120.0660
24	24	24	2	1	448	156749000	162.7680	120.0660
25	25	25	2	1	307	54785800	169.4140	102.0560
26	26	26	3	1	259	28587300	176.6710	102.0560
27	27	27	3	1	193	4729110	183.5190	56.0497
28	28	28	3	1	176	15149600	190.9490	74.0605
29	29	29	3	1	163	1281970	197.7980	64.5802
30	30	30	4	1	87	1544430	205.2600	56.0497
31	31	31	4	1	257	2099510	212.3040	110.0860
32	32	32	4	1	287	2449720	219.6240	84.0449
33	33	33	4	1	253	2073550	226.9490	110.0860
34	34	34	1	1	762	324080000	234.1820	120.0660
35	35	35	2	1	458	150504000	241.2210	120.0660
36	36	36	2	1	316	56265300	247.8760	102.0560
37	37	37	3	1	217	28866800	255.1210	102.0550
38	38	38	3	1	220	5248920	261.9830	56.0497
collisionEnergy ionisationEnergy lowMZ highMZ precursorScanNum precursorMZ precursorCharge								
1	0	0	50.3254	298.6730	0	0.0000	0	

Case Study – Proteomics Analysis

Case Study – Proteomics Analysis

37	0	0	50.9147	112.8510	36	102.0600	0
38	35	0	50.6186	114.6040	36	102.0600	0
precursorIntensity mergedScan mergedResultScanNum mergedResultStartScanNum mergedResultEndScanNum							
1	0.00000e+00	0	0	0	0	0	0
2	2.10140e+08	0	0	0	0	0	0
3	2.10140e+08	0	0	0	0	0	0
4	1.44188e+06	0	0	0	0	0	0
5	1.44188e+06	0	0	0	0	0	0
6	1.55609e+06	0	0	0	0	0	0
7	1.55609e+06	0	0	0	0	0	0
8	2.02325e+04	0	0	0	0	0	0
9	2.02325e+04	0	0	0	0	0	0
10	1.28766e+04	0	0	0	0	0	0
11	1.28766e+04	0	0	0	0	0	0
12	0.00000e+00	0	0	0	0	0	0
13	2.00932e+08	0	0	0	0	0	0
14	2.00932e+08	0	0	0	0	0	0
15	1.62085e+06	0	0	0	0	0	0
16	1.62085e+06	0	0	0	0	0	0
17	1.66728e+06	0	0	0	0	0	0
18	1.66728e+06	0	0	0	0	0	0
19	1.75168e+04	0	0	0	0	0	0
20	1.75168e+04	0	0	0	0	0	0
21	8.48349e+03	0	0	0	0	0	0
22	8.48349e+03	0	0	0	0	0	0
23	0.00000e+00	0	0	0	0	0	0
24	1.89890e+08	0	0	0	0	0	0
25	1.89890e+08	0	0	0	0	0	0
26	1.47298e+06	0	0	0	0	0	0
27	1.47298e+06	0	0	0	0	0	0
28	1.63855e+06	0	0	0	0	0	0
29	1.63855e+06	0	0	0	0	0	0
30	9.33749e+03	0	0	0	0	0	0
31	9.33749e+03	0	0	0	0	0	0
32	1.09943e+04	0	0	0	0	0	0
33	1.09943e+04	0	0	0	0	0	0
34	0.00000e+00	0	0	0	0	0	0
35	2.06837e+08	0	0	0	0	0	0
36	2.06837e+08	0	0	0	0	0	0
37	1.52646e+06	0	0	0	0	0	0
38	1.52646e+06	0	0	0	0	0	0

Case Study – Proteomics Analysis

```
38      1.52646e+06      0      0      0      0      0
  injectionTime filterString
  1      0      <NA> controllerType=0 controllerNumber=1 scan=1      TRUE      NA
  2      0      <NA> controllerType=0 controllerNumber=1 scan=2      TRUE      NA
  3      0      <NA> controllerType=0 controllerNumber=1 scan=3      TRUE      NA
  4      0      <NA> controllerType=0 controllerNumber=1 scan=4      TRUE      NA
  5      0      <NA> controllerType=0 controllerNumber=1 scan=5      TRUE      NA
  6      0      <NA> controllerType=0 controllerNumber=1 scan=6      TRUE      NA
  7      0      <NA> controllerType=0 controllerNumber=1 scan=7      TRUE      NA
  8      0      <NA> controllerType=0 controllerNumber=1 scan=8      TRUE      NA
  9      0      <NA> controllerType=0 controllerNumber=1 scan=9      TRUE      NA
  10     0      <NA> controllerType=0 controllerNumber=1 scan=10     TRUE      NA
  11     0      <NA> controllerType=0 controllerNumber=1 scan=11     TRUE      NA
  12     0      <NA> controllerType=0 controllerNumber=1 scan=12     TRUE      NA
  13     0      <NA> controllerType=0 controllerNumber=1 scan=13     TRUE      NA
  14     0      <NA> controllerType=0 controllerNumber=1 scan=14     TRUE      NA
  15     0      <NA> controllerType=0 controllerNumber=1 scan=15     TRUE      NA
  16     0      <NA> controllerType=0 controllerNumber=1 scan=16     TRUE      NA
  17     0      <NA> controllerType=0 controllerNumber=1 scan=17     TRUE      NA
  18     0      <NA> controllerType=0 controllerNumber=1 scan=18     TRUE      NA
  19     0      <NA> controllerType=0 controllerNumber=1 scan=19     TRUE      NA
  20     0      <NA> controllerType=0 controllerNumber=1 scan=20     TRUE      NA
  21     0      <NA> controllerType=0 controllerNumber=1 scan=21     TRUE      NA
  22     0      <NA> controllerType=0 controllerNumber=1 scan=22     TRUE      NA
  23     0      <NA> controllerType=0 controllerNumber=1 scan=23     TRUE      NA
  24     0      <NA> controllerType=0 controllerNumber=1 scan=24     TRUE      NA
  25     0      <NA> controllerType=0 controllerNumber=1 scan=25     TRUE      NA
  26     0      <NA> controllerType=0 controllerNumber=1 scan=26     TRUE      NA
  27     0      <NA> controllerType=0 controllerNumber=1 scan=27     TRUE      NA
  28     0      <NA> controllerType=0 controllerNumber=1 scan=28     TRUE      NA
  29     0      <NA> controllerType=0 controllerNumber=1 scan=29     TRUE      NA
  30     0      <NA> controllerType=0 controllerNumber=1 scan=30     TRUE      NA
  31     0      <NA> controllerType=0 controllerNumber=1 scan=31     TRUE      NA
  32     0      <NA> controllerType=0 controllerNumber=1 scan=32     TRUE      NA
  33     0      <NA> controllerType=0 controllerNumber=1 scan=33     TRUE      NA
  34     0      <NA> controllerType=0 controllerNumber=1 scan=34     TRUE      NA
  35     0      <NA> controllerType=0 controllerNumber=1 scan=35     TRUE      NA
  36     0      <NA> controllerType=0 controllerNumber=1 scan=36     TRUE      NA
  37     0      <NA> controllerType=0 controllerNumber=1 scan=37     TRUE      NA
  38     0      <NA> controllerType=0 controllerNumber=1 scan=38     TRUE      NA
[ reached 'max' / getOption("max.print") -- omitted 17 rows ]
```

Case Study – Proteomics Analysis

- Investigating the number of actual spectra in the set:

```
75 # Investigating the number of actual spectra in the set:  
76 ``{r}  
77 runInfo(my.data)$scanCount  
78 ``  
  
[1] 55
```

Case Study – Proteomics Analysis

- Investigating the ion source for the experiment:

```
80 # Investigating the ion source for the experiment
81 ````{r}
82 instrumentInfo(my.data)$ionisation
83 ````

[1] "electrospray ionization"
```

Case Study – Proteomics Analysis

- The data itself can be accessed by using the peaks() function:

```
85 # Accessing the data itself using the peaks() function.  
86 ``{r}  
87 peaks(my.data,1)  
88 ```
```

```
      [,1]      [,2]  
[1,] 50.32539 1.185034e+04  
[2,] 50.65150 1.201500e+04  
[3,] 50.91941 1.068692e+04  
[4,] 51.02643 1.053706e+04  
[5,] 51.14503 1.140725e+04  
[6,] 51.16061 1.060811e+04  
[7,] 51.23557 1.214280e+04  
[8,] 51.28455 1.084886e+04  
[9,] 51.32047 9.873077e+03  
[10,] 51.66521 1.217426e+04  
[11,] 51.81630 1.007943e+04  
[12,] 52.61617 1.240245e+04  
[13,] 52.61874 1.033738e+04  
[14,] 52.61882 9.897575e+03  
[15,] 52.65551 1.015991e+04  
[16,] 52.82106 1.264292e+04  
[17,] 53.28766 1.157443e+04  
[18,] 53.46381 1.192924e+04  
[19,] 54.06857 1.218992e+04  
[20,] 54.64257 1.271298e+04  
[21,] 55.02473 1.083895e+04  
[22,] 55.16191 1.197379e+04  
[23,] 55.39608 1.318504e+04  
[24,] 55.69463 1.107968e+04  
[25,] 55.69698 1.182305e+04  
[26,] 55.73802 1.141378e+04  
[27,] 55.99680 1.449282e+04  
[28,] 56.01848 1.053534e+04  
[29,] 56.04968 2.106181e+04  
[30,] 56.07721 1.168999e+04  
[31,] 56.37001 1.271762e+04  
[32,] 56.49357 1.116523e+04  
[33,] 56.59671 1.161248e+04  
[34,] 57.03291 1.253886e+04  
[35,] 57.44817 1.237266e+04  
[36,] 57.56149 1.232441e+04  
[37,] 57.57074 1.251690e+04  
[38,] 57.63020 1.122068e+04  
[39,] 57.80737 1.249781e+04  
[40,] 57.87675 1.230323e+04
```

Case Study – Proteomics Analysis

```
90 ````{r}
91 peaks(my.data)
92 ````

[[1]]
      [,1]      [,2]
[1,] 50.32539 1.185034e+04
[2,] 50.65150 1.201500e+04
[3,] 50.91941 1.068692e+04
[4,] 51.02643 1.053706e+04
[5,] 51.14503 1.140725e+04
[6,] 51.16061 1.060811e+04
[7,] 51.23557 1.214280e+04
[8,] 51.28455 1.084886e+04
[9,] 51.32047 9.873077e+03
[10,] 51.66521 1.217426e+04
[11,] 51.81630 1.007943e+04
[12,] 52.61617 1.240245e+04
[13,] 52.61874 1.033738e+04
[14,] 52.61882 9.897575e+03
[15,] 52.65551 1.015991e+04
[16,] 52.82106 1.264292e+04
[17,] 53.28766 1.157443e+04
[18,] 53.46381 1.192924e+04
[19,] 54.06857 1.218992e+04
[20,] 54.64257 1.271298e+04
[21,] 55.02473 1.083895e+04
[22,] 55.16191 1.197379e+04
[23,] 55.39608 1.318504e+04
[24,] 55.69463 1.107968e+04
[25,] 55.69698 1.182305e+04
[26,] 55.73802 1.141378e+04
[27,] 55.99680 1.449282e+04
[28,] 56.01848 1.053534e+04
[29,] 56.04968 2.106181e+04
[30,] 56.07721 1.168999e+04
[31,] 56.37001 1.271762e+04
[32,] 56.49357 1.116523e+04
[33,] 56.59671 1.161248e+04
[34,] 57.03291 1.253886e+04
[35,] 57.44817 1.237266e+04
[36,] 57.56149 1.232441e+04
[37,] 57.57074 1.251690e+04
[38,] 57.63020 1.122068e+04
[39,] 57.80737 1.249781e+04

[[2]]
      [,1]      [,2]
[1,] 50.44595 7.130229e+03
[2,] 50.63268 6.751621e+03
[3,] 51.19802 7.744562e+03
[4,] 51.30608 7.552050e+03
[5,] 52.19186 7.821688e+03
[6,] 52.27118 8.311925e+03
[7,] 52.30317 7.661416e+03
[8,] 52.33964 7.897834e+03
[9,] 52.54125 8.598500e+03
[10,] 52.55791 7.412984e+03
[11,] 52.71907 7.608189e+03
[12,] 52.84742 8.373299e+03
[13,] 52.90979 8.570965e+03
[14,] 53.12151 7.852208e+03
[15,] 53.89261 7.181609e+03
[16,] 54.12863 7.886996e+03
[17,] 54.25117 8.021970e+03
[18,] 54.25359 6.982733e+03
[19,] 54.31281 8.117685e+03
[20,] 54.57345 6.943318e+03
[21,] 54.81168 7.215416e+03
[22,] 54.92728 7.199123e+03
[23,] 56.04983 1.157029e+04
[24,] 56.19535 7.446009e+03
[25,] 56.46938 7.716136e+03
[26,] 56.74446 9.424041e+03
[27,] 56.87177 7.570626e+03
[28,] 57.27736 7.771037e+03
[29,] 57.48447 8.067558e+03
[30,] 57.54039 7.521494e+03
[31,] 57.84854 8.147912e+03
[32,] 57.90528 7.368232e+03
[33,] 58.21653 9.569272e+03
[34,] 58.56902 8.423535e+03
[35,] 58.58501 8.095901e+03
[36,] 58.66974 8.602247e+03
[37,] 58.86484 7.274165e+03
[38,] 59.64755 7.135571e+03
[39,] 59.82327 8.698951e+03

[[3]]
      [,1]      [,2]
[1,] 50.06576 6794.749
[2,] 50.59673 7079.568
[3,] 51.19412 7668.185
[4,] 51.27377 8077.836
[5,] 51.46772 7213.749
[6,] 51.65720 8776.008
[7,] 51.85004 7182.117
[8,] 52.56889 6912.155
[9,] 52.67318 7243.696
[10,] 52.82322 7927.134
[11,] 52.97269 7437.494
[12,] 53.31343 9220.775
[13,] 53.49500 6592.752
[14,] 53.60342 7440.715
[15,] 53.77227 7104.609
[16,] 54.30957 6826.388
[17,] 54.54459 7425.452
[18,] 54.79221 8127.149
[19,] 54.85666 8938.672
[20,] 54.89362 7470.742
[21,] 55.12350 6944.706
[22,] 55.59793 9447.896
[23,] 56.01242 7472.028
[24,] 56.04912 10787.321
[25,] 56.04974 821494.125
[26,] 56.05036 18899.715
[27,] 56.31257 7571.338
[28,] 56.99893 8132.565
[29,] 57.25984 7353.479
[30,] 57.90335 8372.167
[31,] 58.13220 7369.997
[32,] 58.13994 7681.747
[33,] 58.24904 7391.480
[34,] 58.85498 7444.690
[35,] 58.96836 7671.243
[36,] 59.00388 8874.579
[37,] 59.02112 7104.393
[38,] 59.08067 9936.309
[39,] 60.35535 8014.486
[40,] 61.06535 6873.027
[41,] 61.14135 7170.367
```