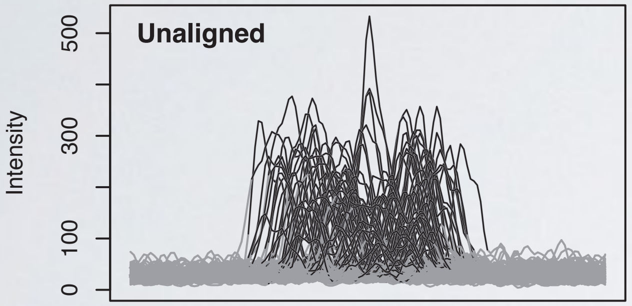
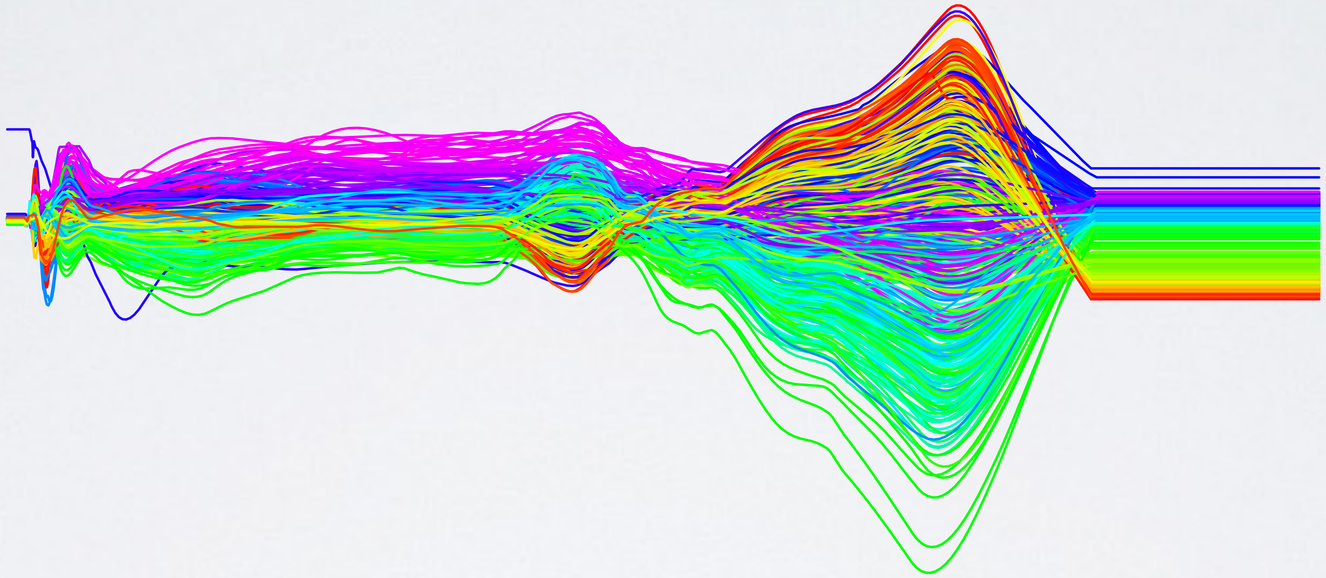
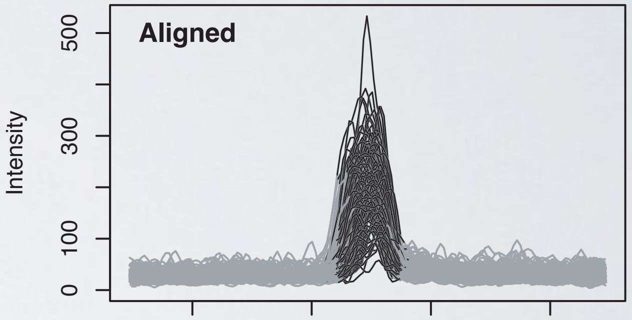
INTRODUCTION TO XCMS





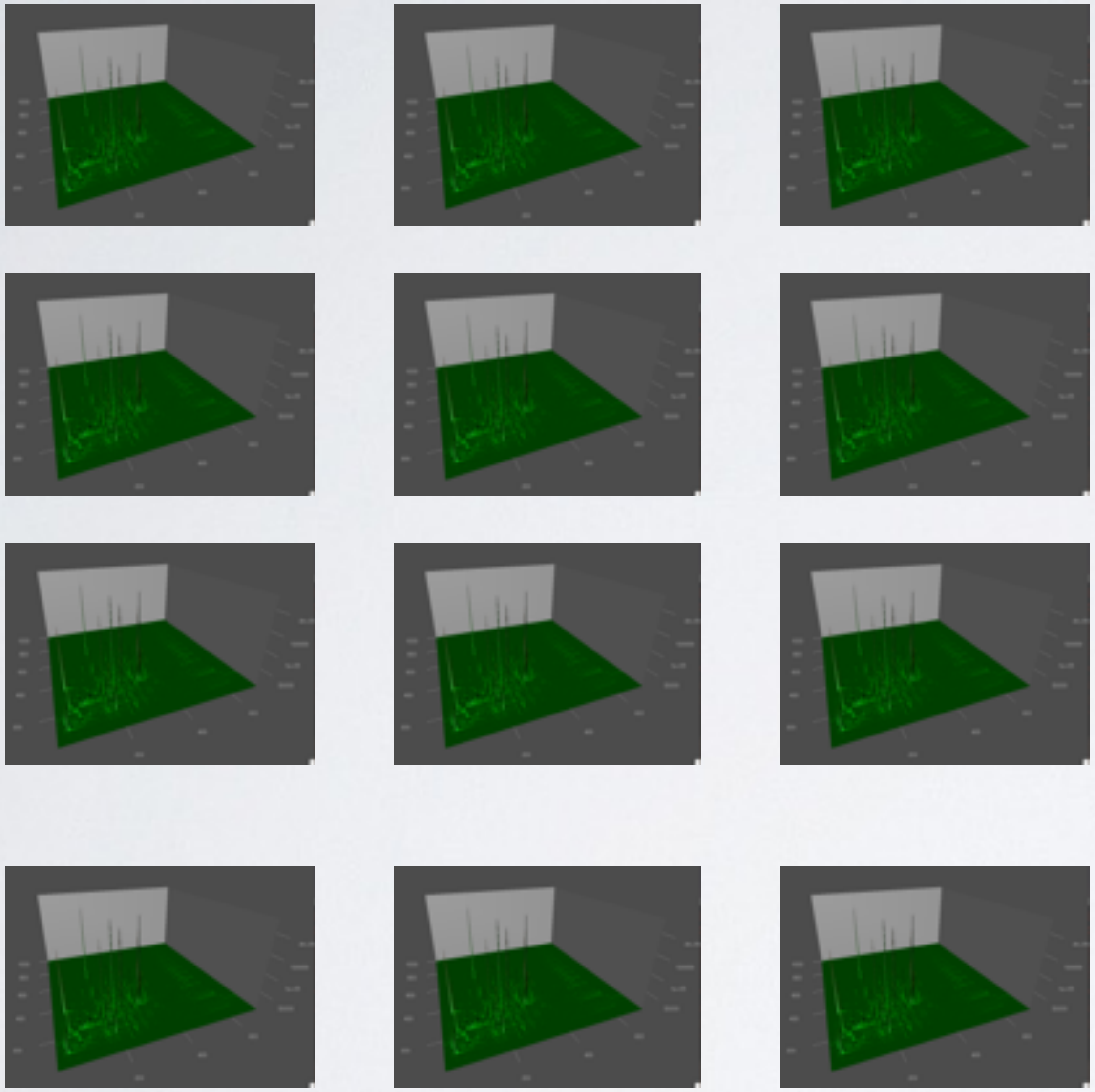


H. Paul Benton

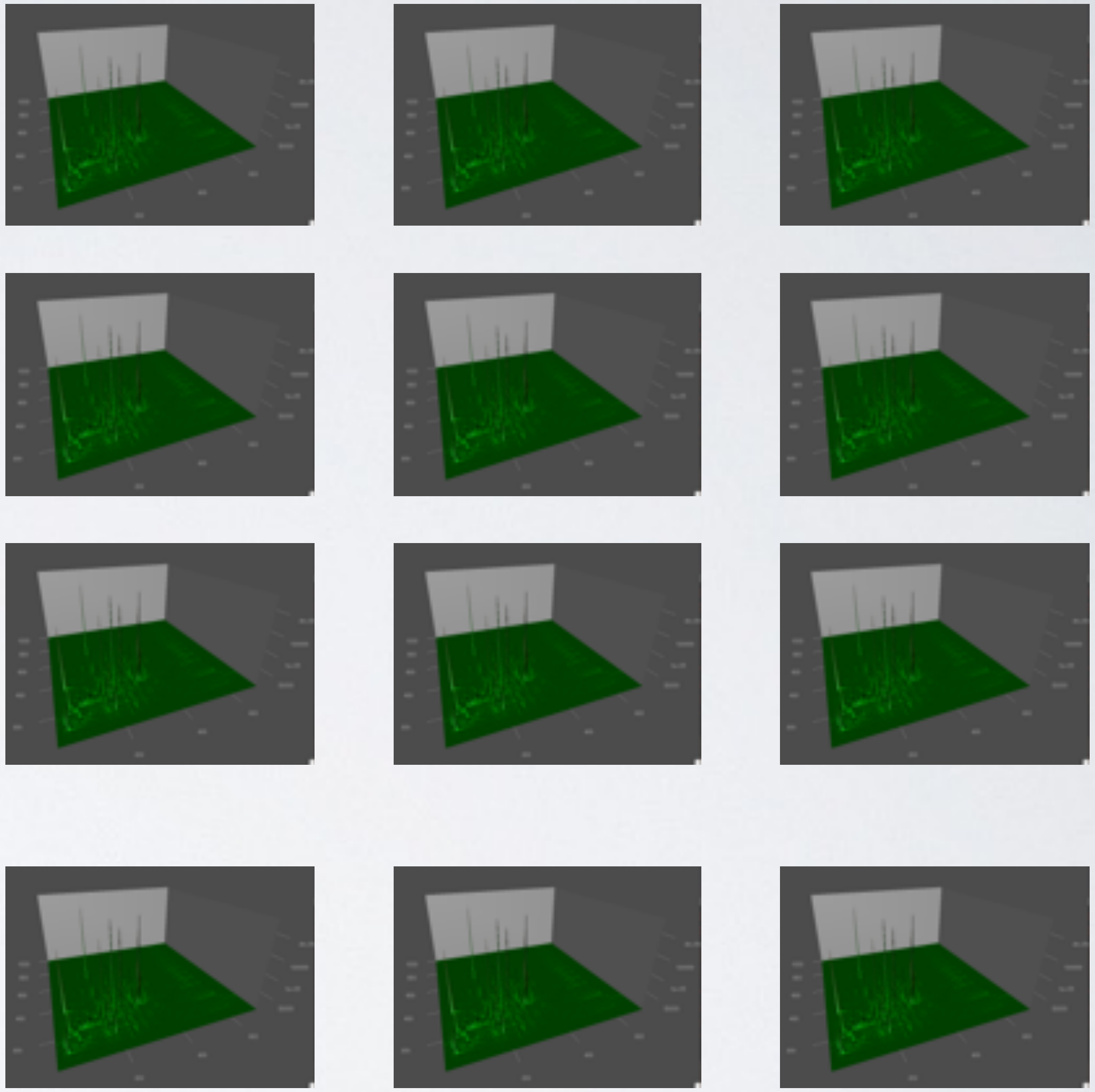
WHAT IS XCMS

- Non-linear alignment tool
- LC-MS data processing platform
    - LC-MS data is several 10-100s (poss 1000s) of samples

Urine Grubb Group 1



Urine Grubb Group 2



Metabolomics Data Analysis Software

Mass Profiler Pro

Agilent Technologies

XCMS plus

AB SCIEX

Progensis QI

Waters

THE SCIENCE OF WHAT'S POSSIBLE.™

XCMS Online

MZmine 2

apLCMS

mzMatch

GENERAL PRINCIPALS

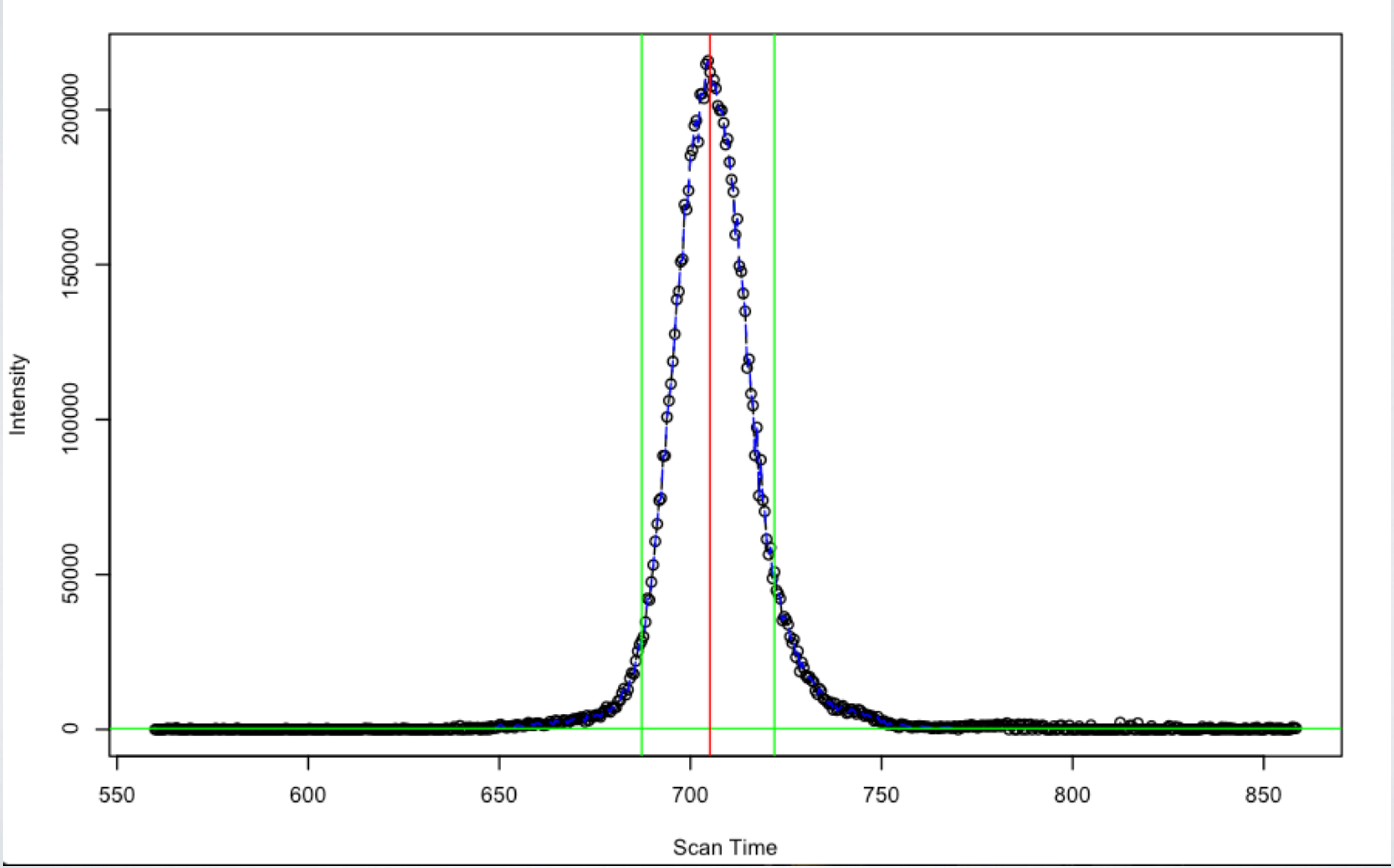
Peak Detection

Grouping
Groups similar Peaks
across replicates

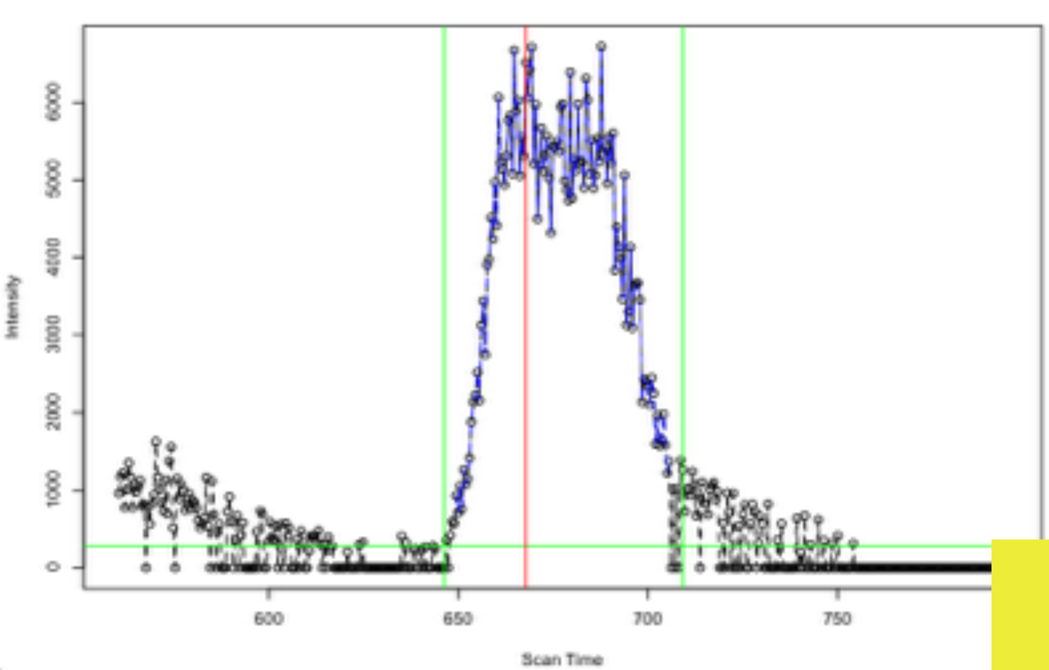
Retention Time
Alignment

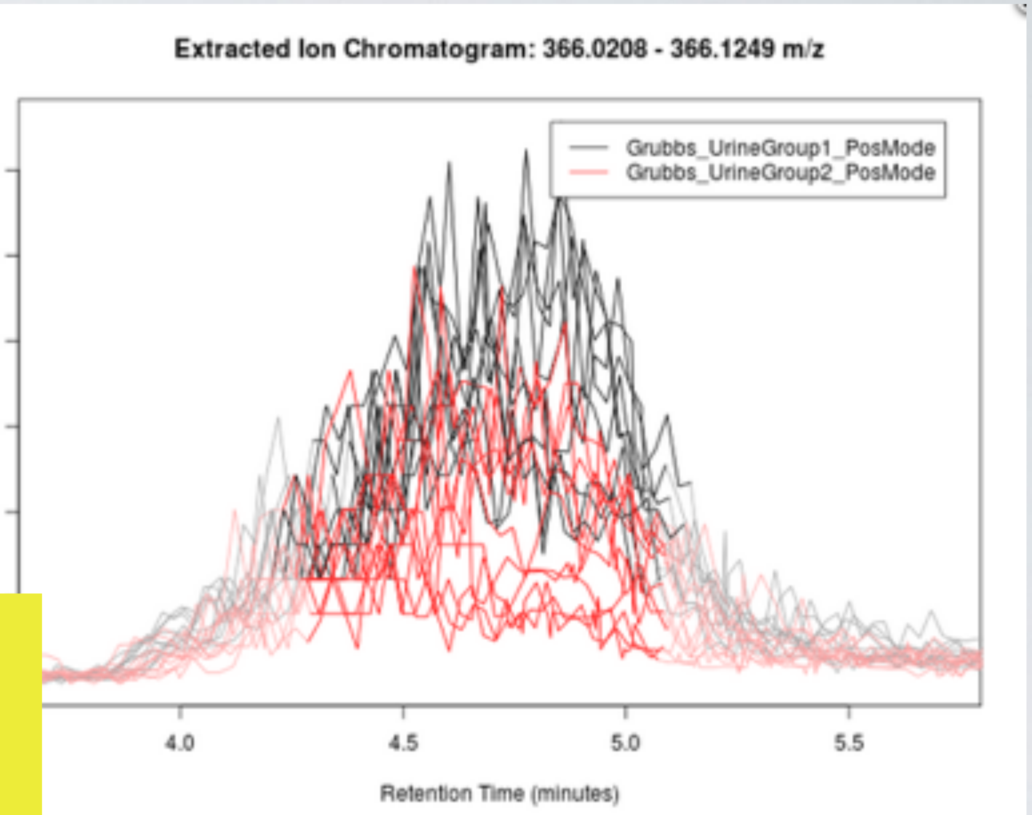
Statistical Analysis
of Classes

EASY PEAK DETECTION

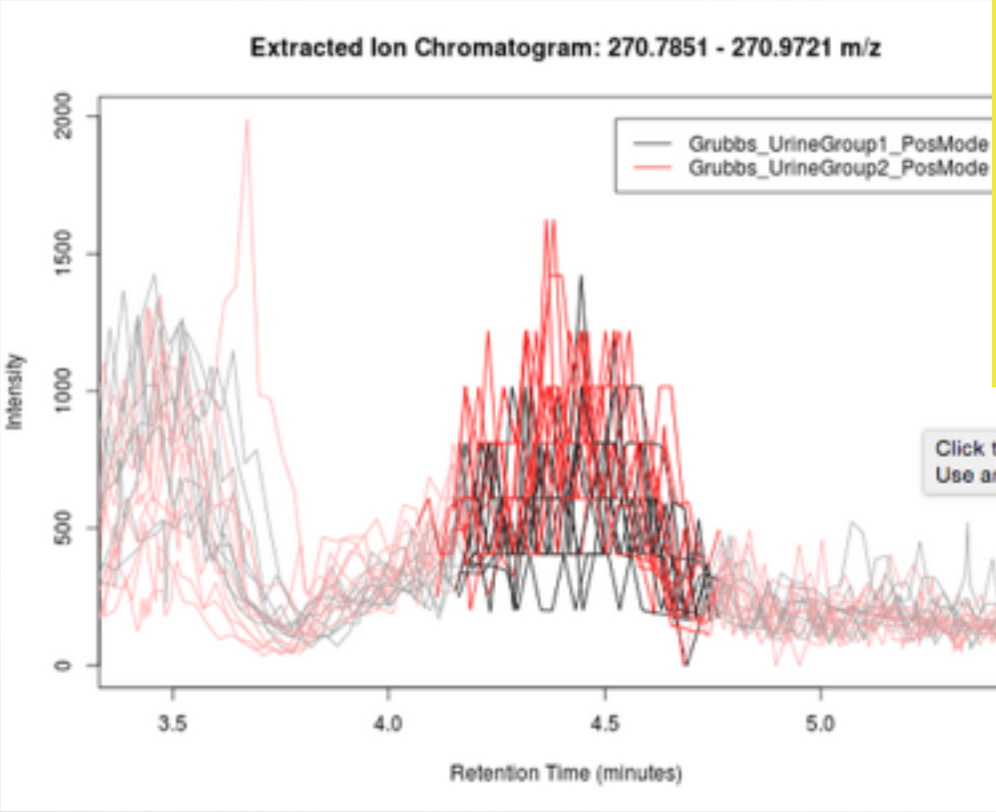


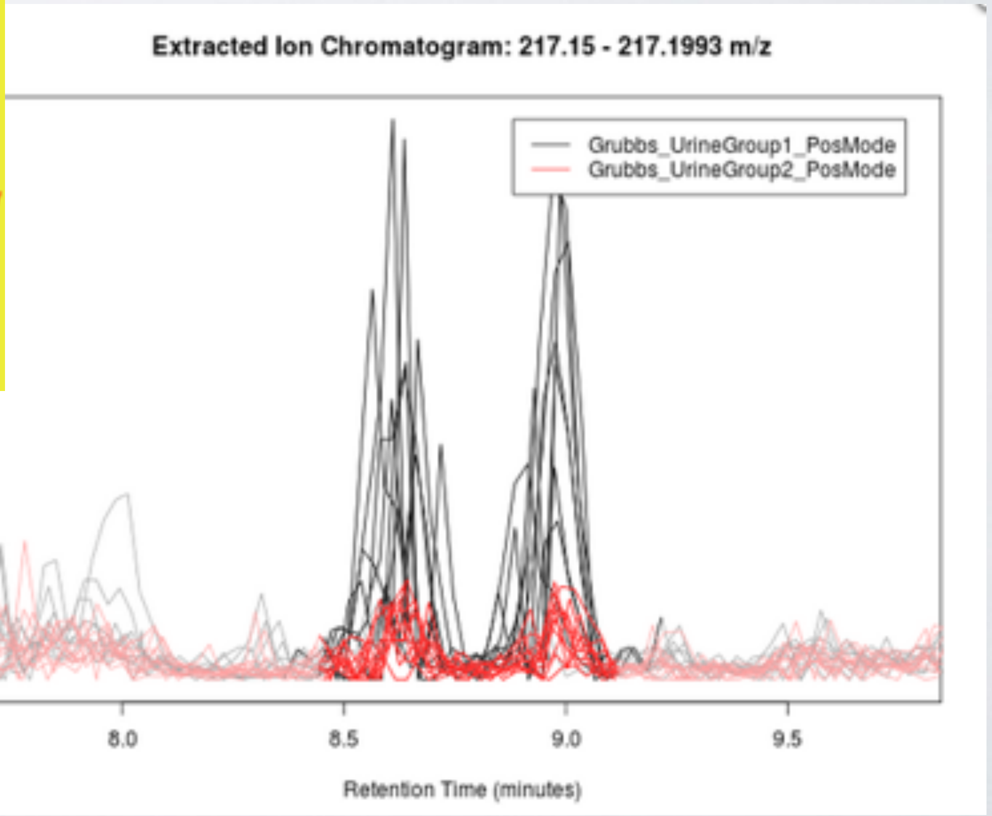
RIGHT?


Intensity
6000
5000
4000
3000
2000
1000
0
600
650
700
750
Scan Time


Extracted Ion Chromatogram: 366.0208 - 366.1249 m/z
Grubbs_UrineGroup1_PosMode
Grubbs_UrineGroup2_PosMode
Intensity
3000
2500
2000
1500
1000
0
4.0
4.5
5.0
5.5
Retention Time (minutes)


KEEP
CALM
AND FOLLOW
THE YELLOW
BRICK ROAD


Extracted Ion Chromatogram: 270.7851 - 270.9721 m/z
Grubbs_UrineGroup1_PosMode
Grubbs_UrineGroup2_PosMode
Intensity
2000
1500
1000
500
0
3.5
4.0
4.5
5.0
Retention Time (minutes)
Click to close i
Use arrow keys


Extracted Ion Chromatogram: 217.15 - 217.1993 m/z
Grubbs_UrineGroup1_PosMode
Grubbs_UrineGroup2_PosMode
Inten
30
2000
1000
0
8.0
8.5
9.0
9.5
Retention Time (minutes)

HAT FITTING

- Different hat for different heads (& faces apparently)
- A hat has to fit well so it must be sized

PEAK DETECTION

- Data comes in two types in MS : centroid & profile
- Generally high resolution or low resolution ~ high mass accuracy or low mass accuracy
- Two main choices in XCMS
    - **MatchedFilter** - profile low res
    - **CentWave** - centroid high res

GENERAL PRINCIPALS
GROUPING

Peak Detection

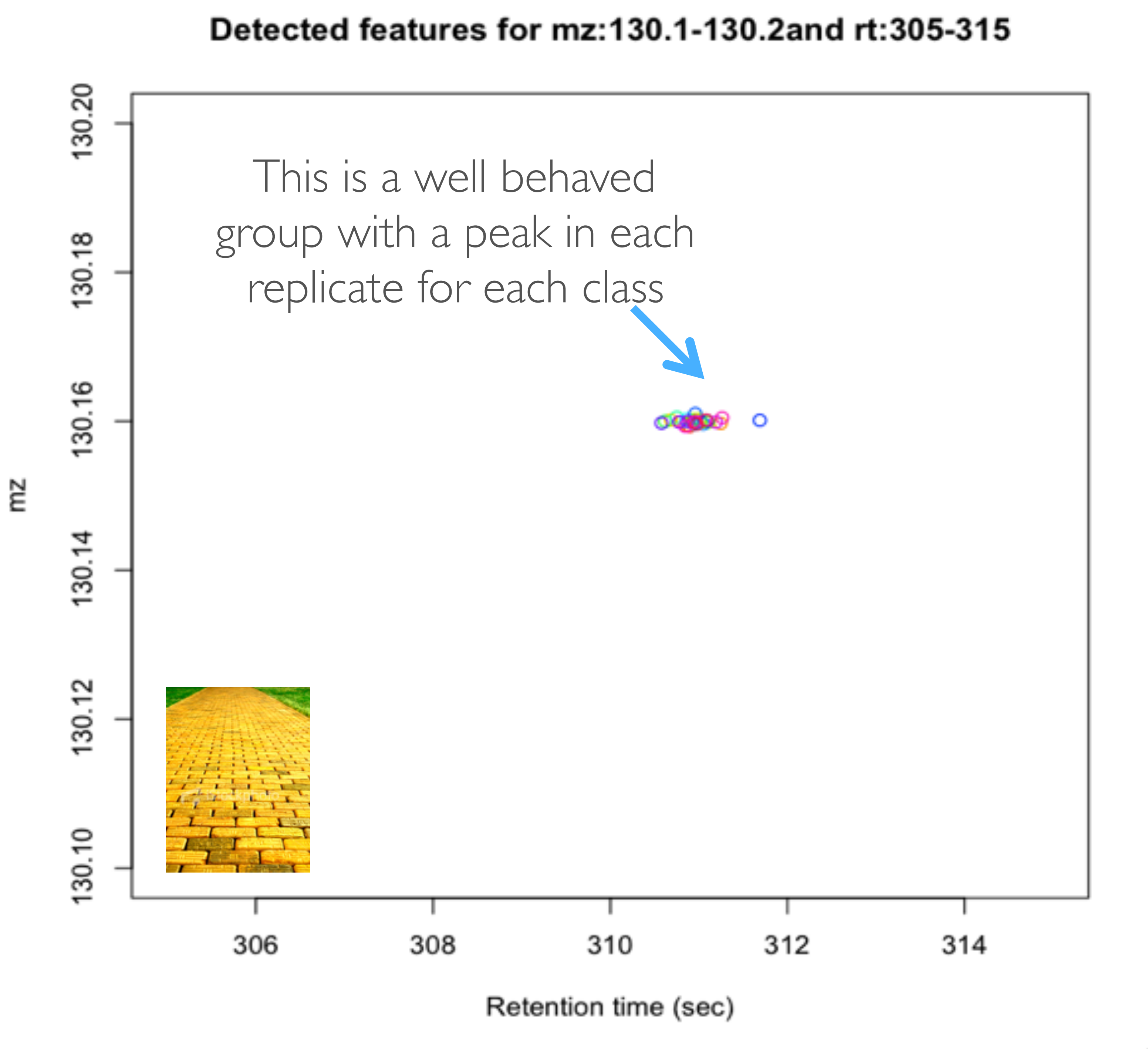
Grouping
Groups similar Peaks across replicates

Retention Time Alignment

Statistical Analysis of Classes

GROUPING

- First time using all of the files
- Looks for closely clustered/**dense** peaks across multiple files.
    - Once peaks are grouped they're know as a group or feature

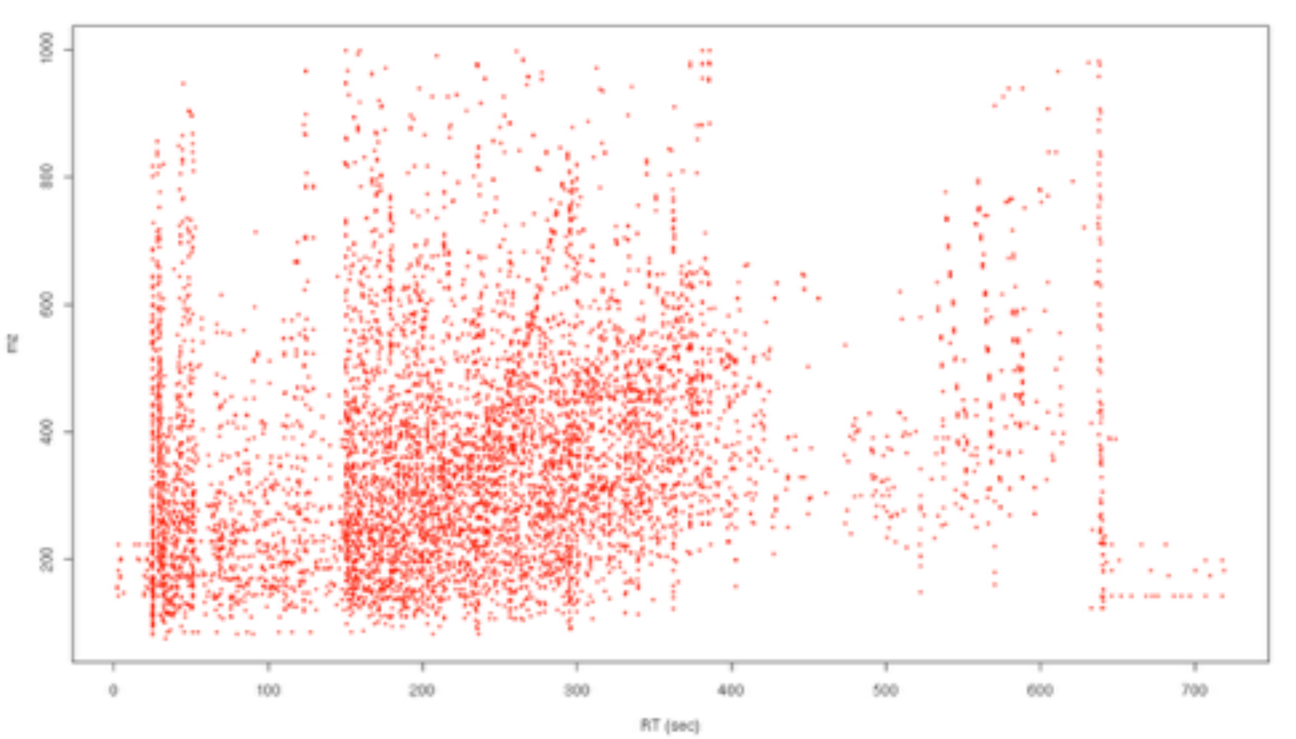


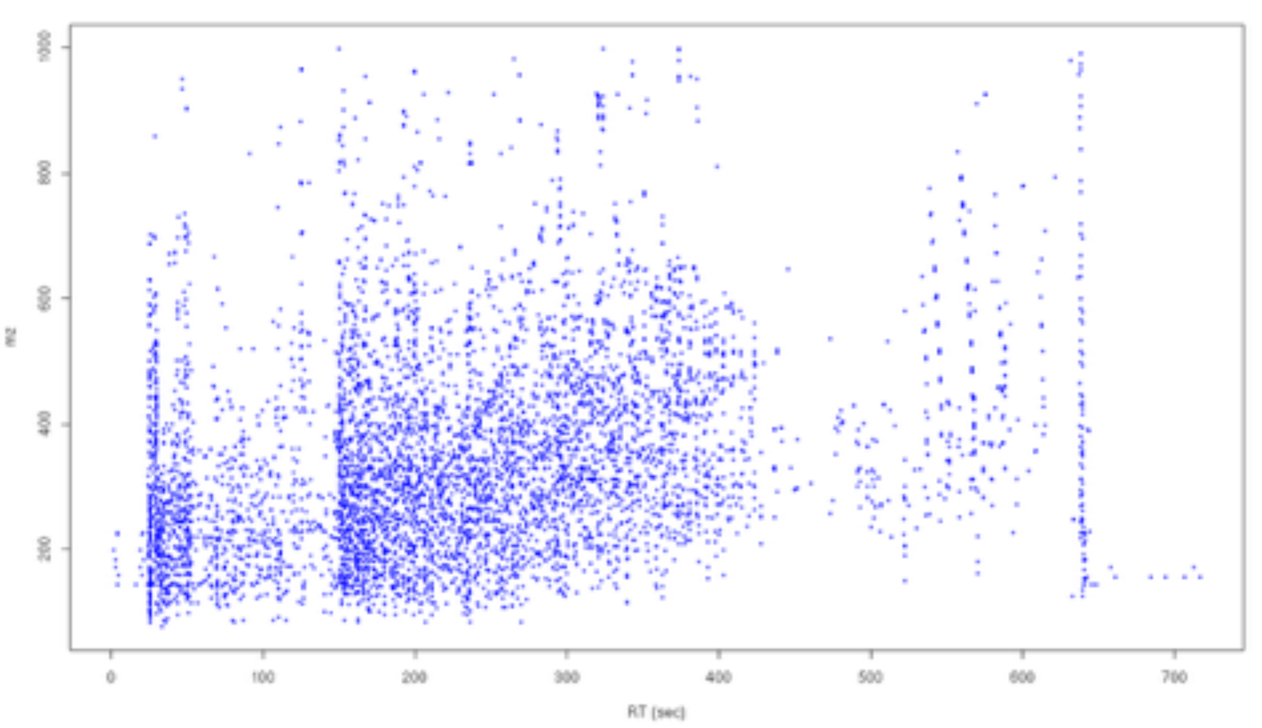
Detected features for mz:130.1-130.2and rt:305-315

This is a well behaved group with a peak in each replicate for each class

GROUPING = NEAREST

- Based on mzMine grouping/alignment algorithm
  - Uses nearest neighbor estimation.





GENERAL PRINCIPALS

Peak Detection

Grouping
Groups similar Peaks
across replicates

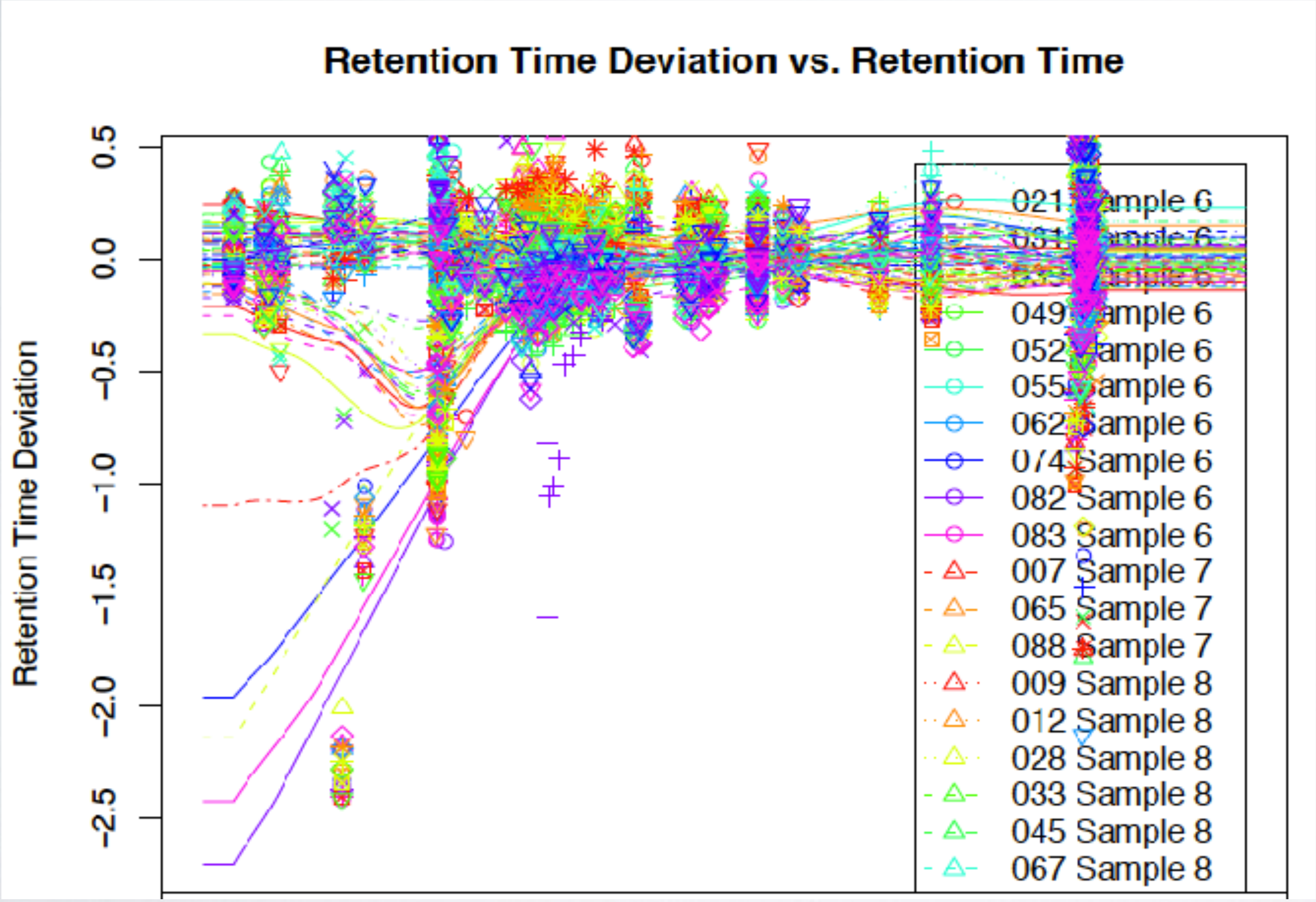
**Retention Time
Alignment**

Statistical Analysis
of Classes

RETENTION TIME ALIGNMENT

- XCMS finds 'well behaved groups'
  - These include group that have missing peaks, extra peaks or perfect groups (parameters)
    - Missing < n/2 !!
  - Median found for each group
  - Local regression used for each sample to find the deviation profile

RETENTION TIME ALIGNMENT - LOESS



Retention Time Deviation vs. Retention Time

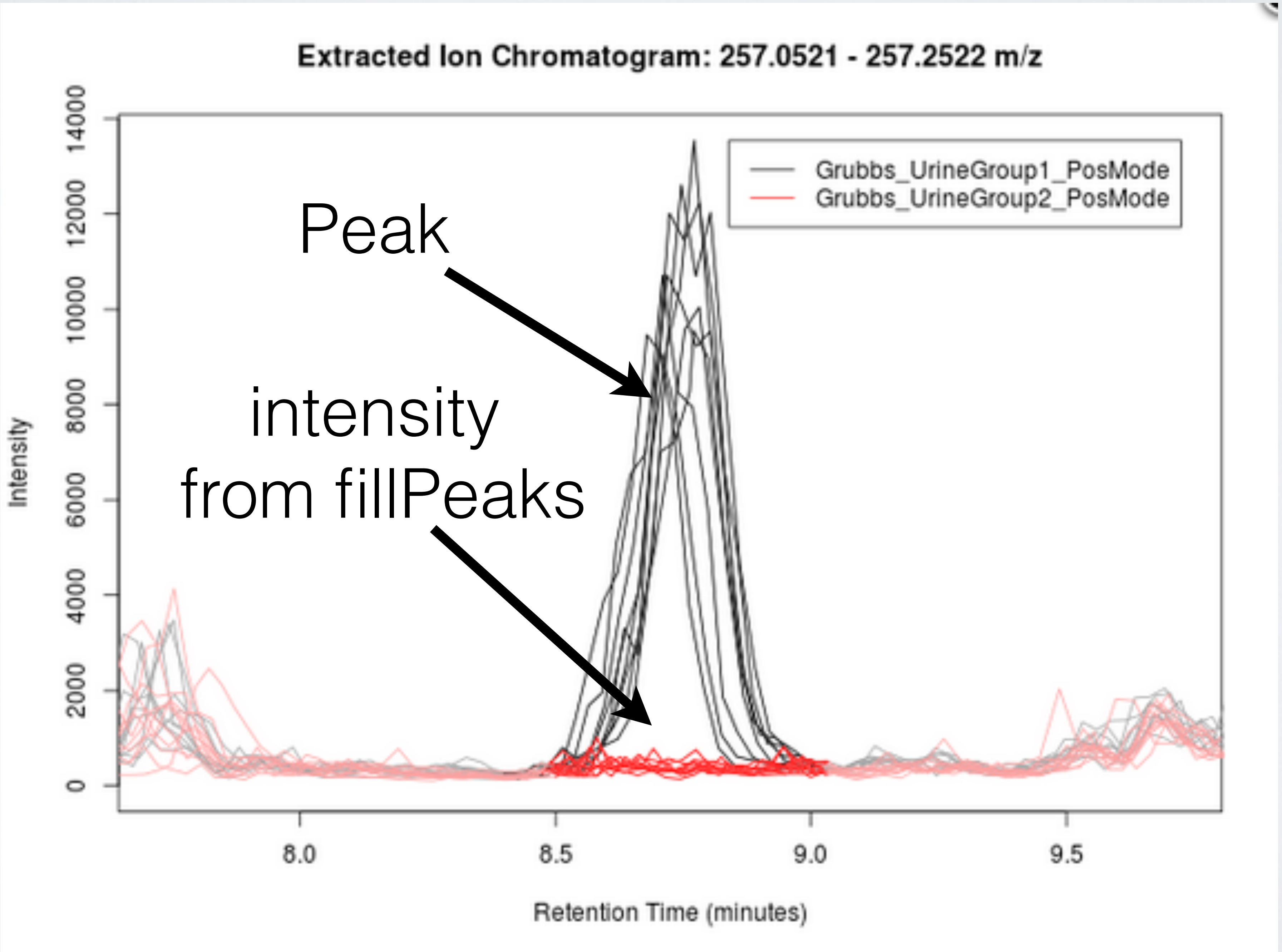
median rt of each
‘well behaved’
group
vs
rt of each file

A good spread of anchors/‘well behaved peak groups’

- After peaks are aligned we must regroup the peaks into features again.
  - Alignment has shifted the peaks and grouping information is lost
  - The New groups ‘will be’ single molecular features (hopefully)

FILLPEAKS

- Going back to each file to find any intensity that wasn't peak picked



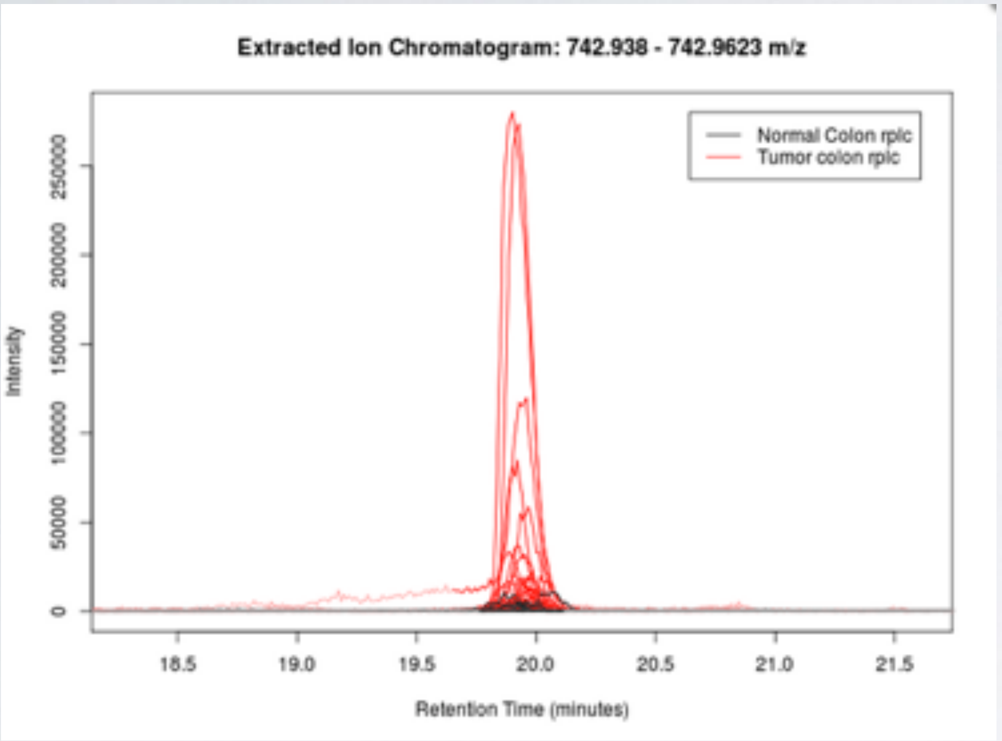
Peak

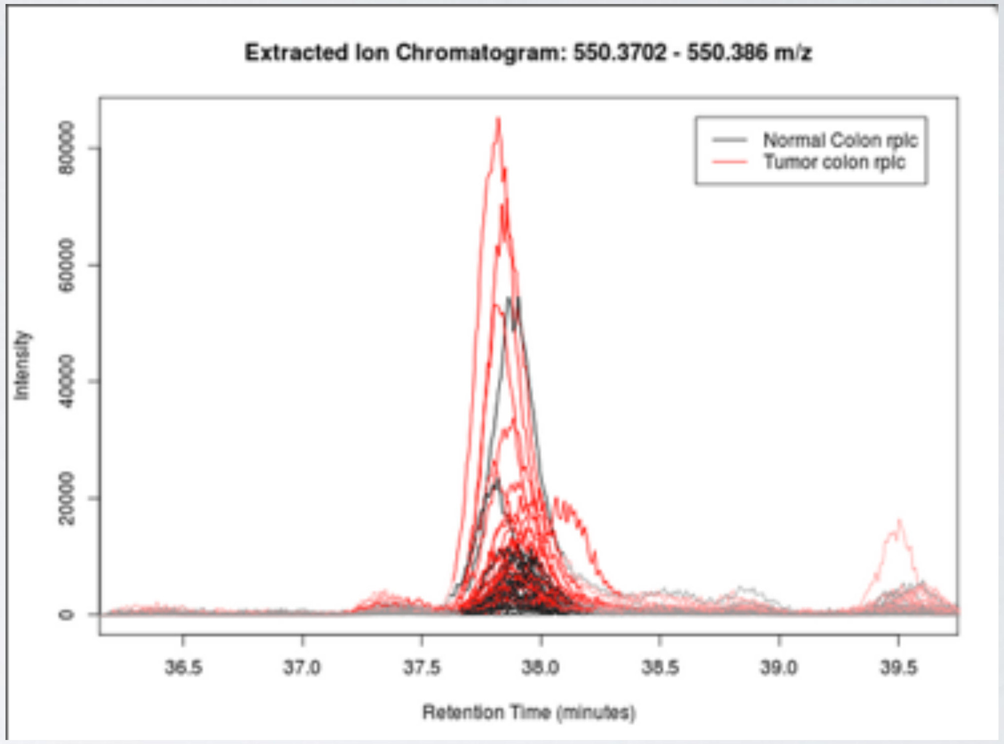
intensity from fillPeaks

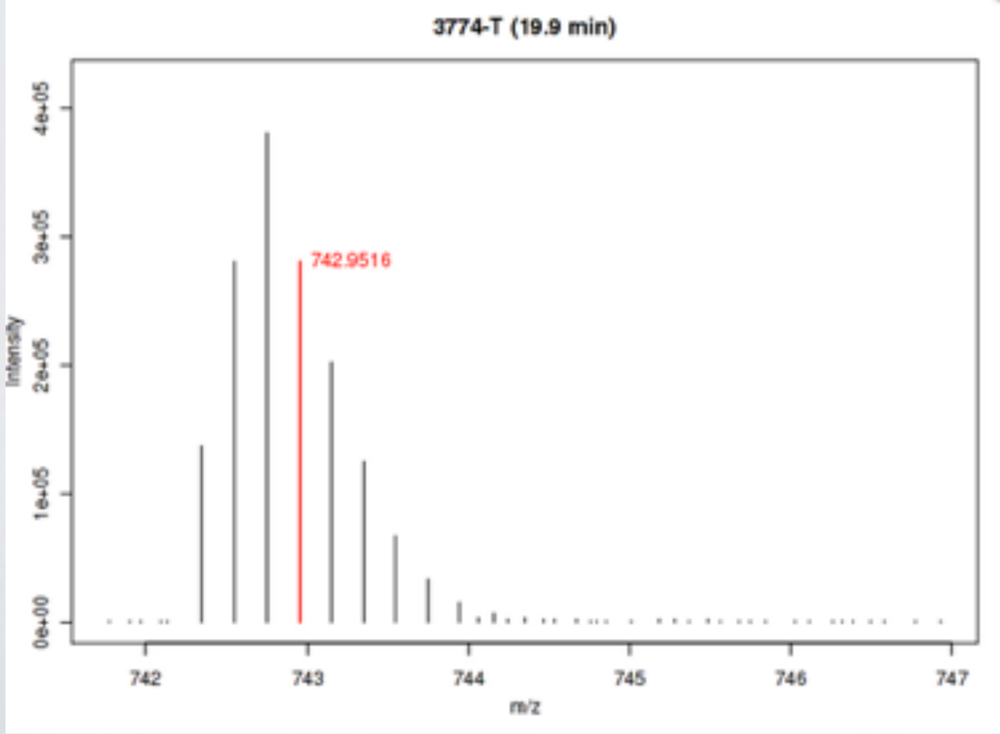
FINALLY !!

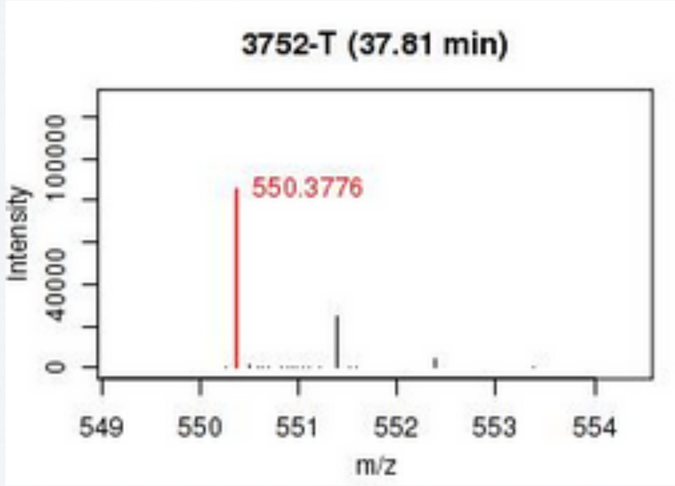
- We have all of our data corrected in a form we can use.
  - Lets look at some data processing:
    - heatmaps
    - PCA
    - Some Stats

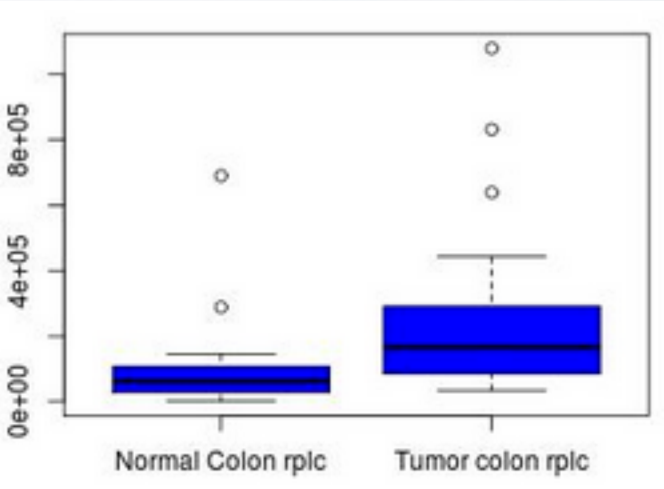
WAIT !!!!


Extracted Ion Chromatogram: 742.938 - 742.9623 m/z
Normal Colon rplc
Tumor colon rplc
Intensity
0
50000
100000
150000
200000
250000
18.5
19.0
19.5
20.0
20.5
21.0
21.5
Retention Time (minutes)


Extracted Ion Chromatogram: 550.3702 - 550.386 m/z
Normal Colon rplc
Tumor colon rplc
Intensity
0
20000
40000
60000
80000
36.5
37.0
37.5
38.0
38.5
39.0
39.5
Retention Time (minutes)


3774-T (19.9 min)
742.9516
Intensity
0e+00
1e+05
2e+05
3e+05
4e+05
742
743
744
745
746
747
m/z


3752-T (37.81 min)
550.3776
Intensity
0
40000
100000
549
550
551
552
553
554
m/z


0e+00
4e+05
8e+05
Normal Colon rplc
Tumor colon rplc

SUMMARY

- XCMS processes LC-MS data and is complex
- XCMS processes LC-MS data and uses some simple algorithms. There are multiple algorithm for different jobs/data types.

QUESTIONS?