GBS 724 class
2-1-16

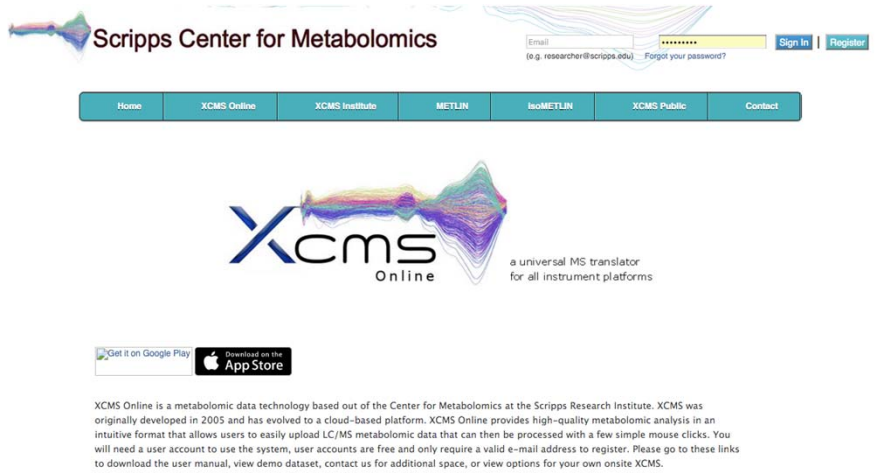
Preparing to analyze LC-MS data

Stephen Barnes, PhD

Synopsis

- LC-MS (and GC-MS) analysis generates a lot of data and requires alignment of individual data sets before statistical analysis can be performed
- We will discuss
  - Uploading data sets
  - Alignment principles (acknowledging the work of Xiuxia Du at UNC-Charlotte and her slides used at the UAB Metabolomics Workshop)
- On Wednesday, Paul Benton from Scripps Research Institute will describe and show you how the online software XCMS works

Uploading Experiments for Metabolomics Analyses using XCMS Online Applications



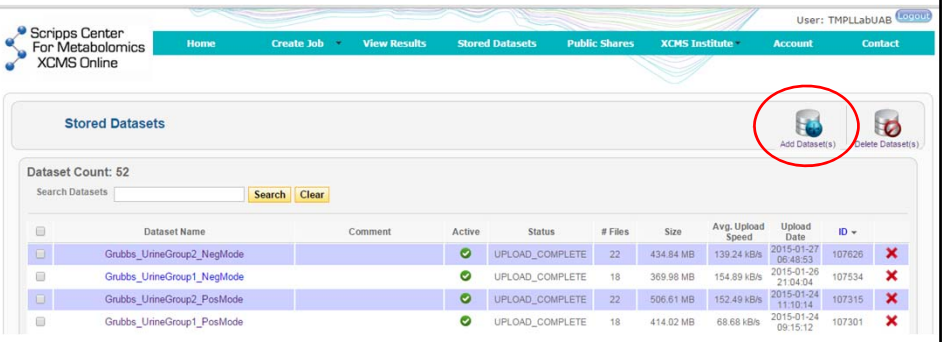
https://xcmsonline.scripps.edu/index.php

Uploading Experiments for Metabolomics Analyses using XCMS Online Applications continued…

Once you have completed your LC-MS/MS analyses of samples in your particular experiment, files can be uploaded directly for most instrumentation types.

Select the option "*Stored Datasets*," on the XCMS toolbar to access datasets already stored, or to upload new datasets for future analyses.

Uploading Experiments for Metabolomics Analyses using XCMS Online Applications continued…

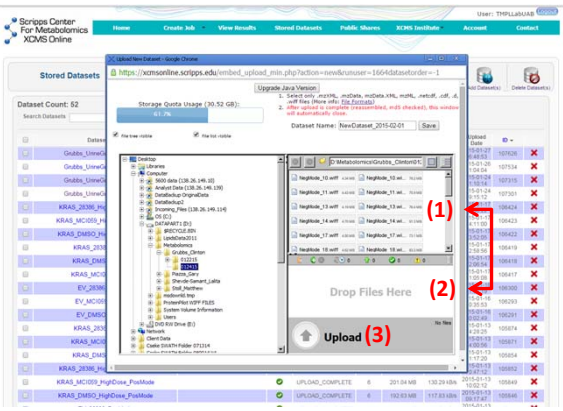


Select the option "*Add Dataset(s)*," on the XCMS toolbar to open the Java directed upload panel. Certain instruments can directly upload raw data while others must be transformed to a universal language such as mzXML.

RAW data file parameters per major manufacturer

Vendor	Instrument Software	File Format	Converter	Can be uploaded directly	Notes
AB SCIEX	Analyst	.wiff	ProteoWizard (see below)	YES	*.wiff files can be uploaded directly and do not need to be converted manually. Please make sure the .wiff.scan files are uploaded together with the .wiff files.* Note for manual conversion: Conversion to centroid mode only works with most recent ProteoWizard versions (>= 3.0.3569)
Agilent	MassHunter	.d	ProteoWizard (see below)	YES	*.d folders can be uploaded directly and do not need to be converted manually.* Note for manual conversion: A bug related to the conversion of Agilent was fixed recently, please update proteoWizard to newest version (>= 3.0.3782)
Agilent	ChemStation	.D	export from Chemstation as 'AIA'	NO	
Bruker	Compass	.d, YEP, BAF, FID	CompassXport or ProteoWizard (see below)	YES	*.d folders can be uploaded directly and do not need to be converted manually.* 1. to use the latest recalibration for the exported data a setting might need to be enabled in the windows registry (see CompassXport manual) 2. Only CompassXport 3.0.6 can convert data from the newest Bruker instruments at the moment.
Thermo Fisher	Xcalibur	RAW	ProteoWizard (see below)	NO	Conversion of Q-Exactive data to centroid mode works only with most recent ProteoWizard versions (>= 3.0.3631)
Waters	MassLynx	.raw	MassLynx (CDF) ProteoWizard (see below)	YES	*.raw folders will be uploaded directly and do not need to be converted manually.* • If you get error messages like *"Error in xcmsRaw(file, profstep = 0) : Time for scan XXX greater than scan YYY"* you have to add the "sortByScanTime" filter for the file conversion as described here. 1. The exported data does not make use of the latest recalibration. No solution at the moment. 2. Removing the lock mass calibration scans and filling in the resulting gaps in the data (Misc/Correct mass calibration gaps) seems to work only if files were converted to CDF format using MassLynx.

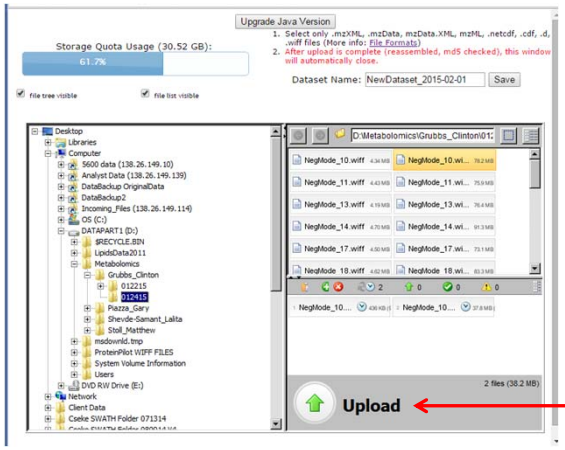
Uploading Experiments for Metabolomics Analyses using XCMS Online Applications continued…



Once the Java pop-up window opens, a standard file architecture for selecting data files is accessible. Select the files you want to add in a particular group from the top right box (1), and drag them to the "*Drop Files Here*" box (2) below.

Wait for the "*Upload*" (3) button to turn green to begin the uploading process to the XCMS server. This might take a few minutes as the server determines if the files are in the correct format.

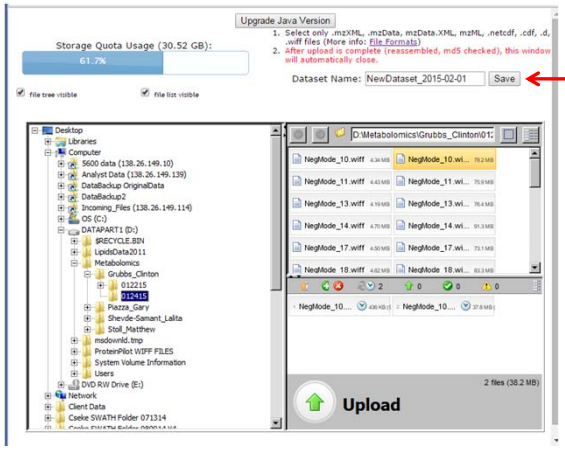
Uploading Experiments for Metabolomics Analyses using XCMS Online Applications continued…



Activate the "Upload" button to begin the process. Depending on the size and number of the files, this process can take up several hours.

If any errors occur in the uploading process, the Exclamation point icon will have a number associated with the amount of files that have issues.

Uploading Experiments for Metabolomics Analyses using XCMS Online Applications continued…



After the upload is complete, give a descriptive name to your data set by typing in the *"DataSet Name"* box and clicking Save.

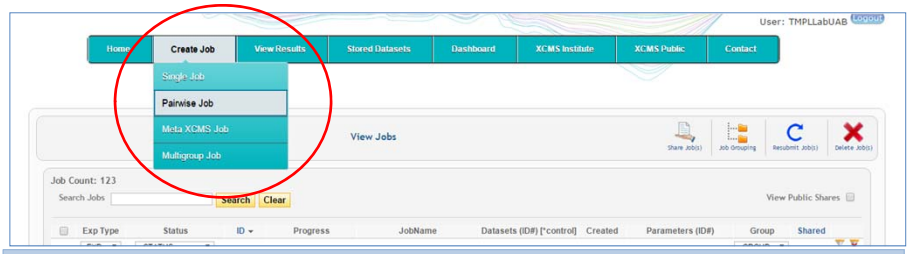
Since you will be comparing a minimum of two data sets, you will repeat the process again with the additional group(s) you wish to evaluate.

Tips for naming files for upload to XCMS Online

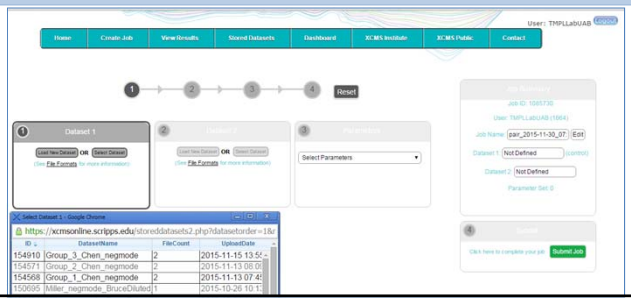
- Before uploading, it is a good idea to create separate file folders on your hard drive to better organize your data into the groups you want to examine.
- Add enough description to discriminate between different samples and sample set names.
- Adding the ionization mode in the name is preferable, i.e. PosMode or NegMode.
- Eliminate open spaces in the data file name by using "_" (undescore) notation. Open spaces can cause upload errors in XCMS Online.

Example of DataSet Name: Grubbs_UrineGroup2_NegMode

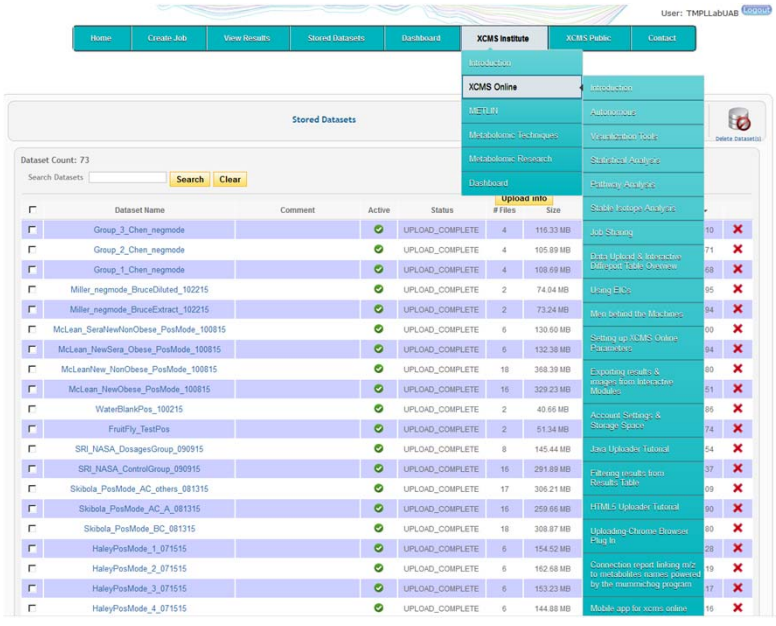
Starting a new XCMS database search…



Select the type of "Job" or test for your newly uploaded data sets. From here, you can select the files you wish to evaluate.



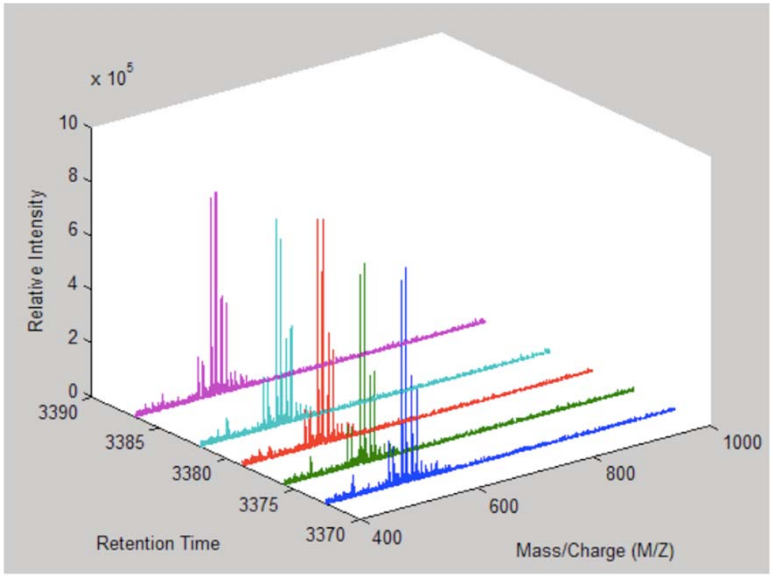
XCMS Online Tutorial Videos



What does a LC-MS data set consist of?

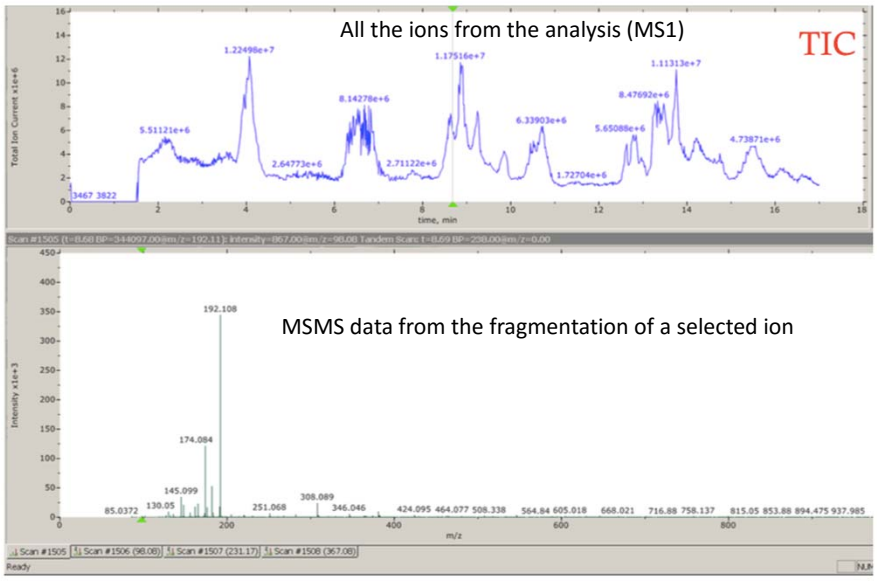
- **Using a Q-TOF instrument, during the LC run for example it acquires data on a 2-second duty cycle**
  - **0-100 msec**
    - **High resolution/mass accuracy MS spectrum**
  - **100-2000 msec**
    - **A succession of selected MSMS spectra**
    - **If each MSMS spectrum is collected for 100 sec, then 19 precursor ions can be selected in the duty cycle**
    - **The precursor ions are selected from the MS spectrum observed in the current duty cycle**
    - **Once an ion has been selected for MSMS it can be placed on a "don't observe" list for say 90 sec**

Raw LC-MS data



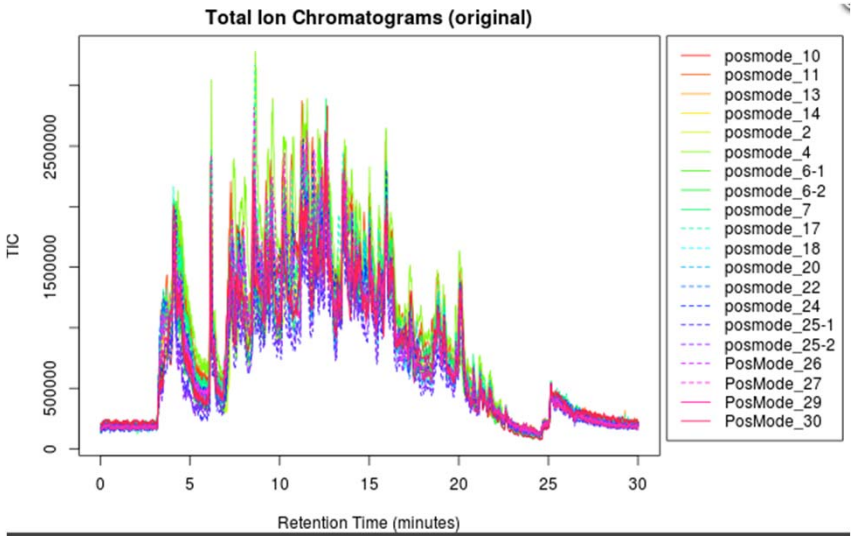
Xiuxia Du

Raw LC-MS data



Xiuxia Du

Overlay of data from multiple samples



Goals of pre-processing

- Extract qualitative and quantitative information of possible metabolites
  - Determine the identity
  - Estimate the relative abundance
- Align samples to correct retention time shifts
- Produce a table of possible metabolites with their quantitative information for subsequent statistical analysis

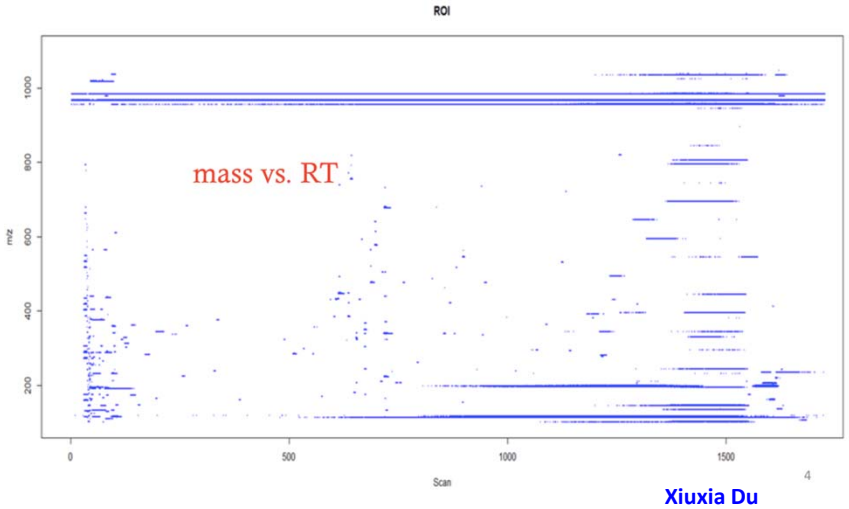
Xiuxia Du

What ions are observed in LC-MS data?

- **Ions coming from the biological system being studied**
- **Ions from compounds introduced into the extract during storage and extraction**
- **Ions from the solvent used for the chromatography**
- **Ions from the column material**
- **Ions from the previous sample that was run**

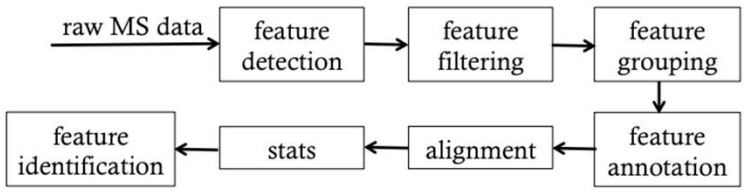
Global look at the ions being eluted

Raw LC-MS data



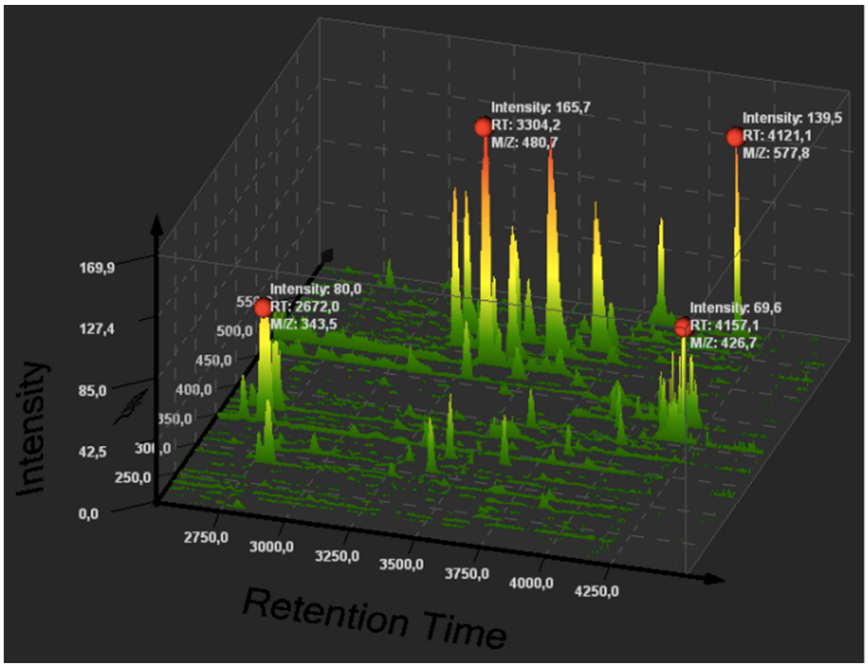
Xiuxia Du

Work flow



- Feature: a 3D signal induced by a single ion species (e.g. $[M+H]^+$ or $[M-H]^-$ of a compound)

Xiuxia Du



Xiuxia Du

Feature detection

- Achieves suppression of noise, metabolite ID and quantification
- Two steps
  - Separation of mass traces
    - Binning
    - Region of interest (ROI)
  - Detection of chromatographic features
- Binning
  - Partition the mass-vs-RT map into bins of fixed width
  - Difficult to estimate optimal bin width

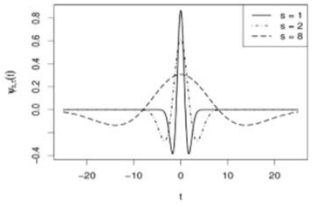
    Too small → split features

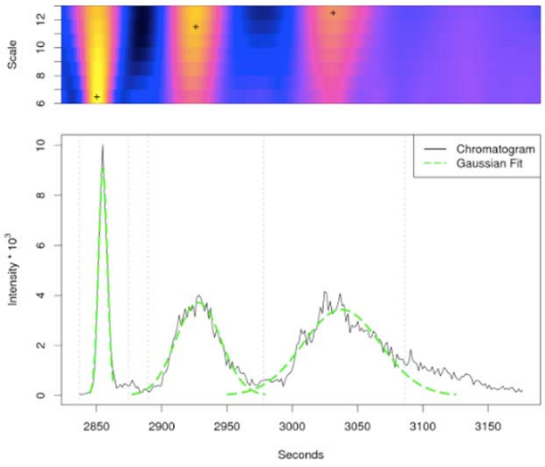
    Too wide → possible feature merging

Xiuxia Du

Detect chromatographic peaks

- Use wavelet transform

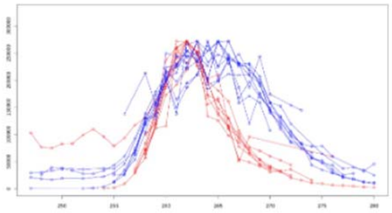




Xiuxia Du

Feature filtering and grouping

- Feature quality measures
  - S/N
  - Feature width
  - Abundance
- Feature grouping
  - Similarity measure: normalized dot product



Xiuxia Du

Feature annotation

m/z = (N*compound mass+mass shift)/CS

Formula	N	Mass shift
[M+H]+	1	1.007276
[M+2H]+	1	2.014552
[M+3H]+	1	3.021828
[M+Na]+	1	22.98977
[M+K]+	1	38.963708
[M-C_3H_9N]+	1	-59.073499
[M+2Na-H]+	1	44.96563
[2M+Na]+	2	22.98977
[M+H-NH_3]+	1	-16.01872
[2M+H]+	2	1.007276
[M-OH]+	1	-17.0028

id	mz	rt	isotopes	adduct	pc
65	176.04	280.09			4
76	136.05	280.43	[14][M+1]1+		5
77	135.05	280.43	[14][M]1+		5
74	153.06	280.43		[M+H]+ 152.05437	5
75	175.04	280.43		[M+Na]+ 152.05437	5
73	197.02	280.76		[M+2Na-H]+ 152.05437	5
78	377.74	286.15			6
79	732.5	286.49			6
83	488.32	286.82		[M+Na]+ 465.33205	7
82	466.34	286.82		[M+H]+ 465.33205	7
...					

Xiuxia Du

Alignment

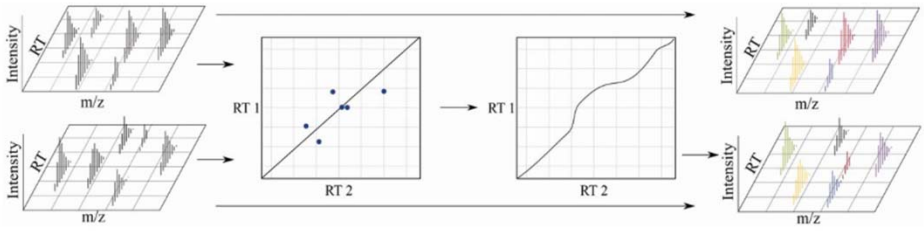
- Goal: Correct retention time shift from run to run

- Approaches
  - Warping
  - Direct match

Xiuxia Du

Alignment approach: warping

- Principle: Models systematic RT shift
- Limitation: the warping functions required for alignment are incapable of capturing component-level variation.

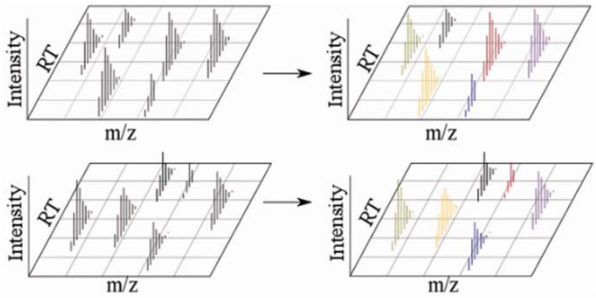


Smith, R.; Ventura, D.; Prince, J. T., LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in bioinformatics* **2013**.

Xiuxia Du

Alignment approach: direct matching

- Principle: analytes are matched directly based on factors such as elution time, charge state, and isotopic envelope characteristics.



Smith, R.; Ventura, D.; Prince, J. T., LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in bioinformatics* **2013**.

Xiuxia Du

Correspondence

- Limitations of warping
  - Warping models systematic shifting.
  - Warping functions are monotonic and cannot capture component-level variation.
  - Warping incorrectly assumes that elution order is preserved across runs.
- The problem should be casted as a correspondence problem.
  - Mapping of identically sourced features across all runs

Smith, R.; Ventura, D.; Prince, J. T., LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in bioinformatics* **2013**.

Xiuxia Du

Result of pre-processing

DB	Name	Mass	RT	platform	IN1	IN2	IN3	IN4	IN5	IN6
HMDB	1-Phenylethylamin	122.09745	24.97845	ES-	0.12862	0.1421305	0.1301326	0.1247924	0.1200045	0.1053275
HMDB	2-Ethylacrylic acid	101.06421	17.811575	ES-	0.0332025	0.0174262	0.0158166	0.0179326	0.0143742	0.0064953
HMDB	Canavanine	177.09653	10.338581	ES-	0.0141136	0.0134146	0.0182777	0.0193855	0.0245958	0.0011908
HMDB	Diketogulonic acid	193.03069	4.7050639	ES-	0.0209463	0.0203901	0.0165056	0.0189088	0.0137482	0.017231
HMDB	Iso-Valeraldehyde	87.080171	11.164359	ES-	0.6558109	0.2742277	0.2651933	0.3093793	0.2101024	0.0541026
in-house	3,4-Dehydro-Dprol	114.04431	3.5491023	ES-	0.2900544	0.287811	0.2290651	0.2754269	0.2314117	0.2061301
in-house	4-hydroxy-proline	132.05326	3.5958634	ES-	0.5584389	0.7353401	0.5273908	0.4412898	0.5074794	0.5423602
in-house	Malic acid	133.01996	3.9406386	ES-	0.0555016	0.0461576	0.0290383	0.0390783	0.0380952	0.0308288
in-house	2,3,4-Trihydroxybu	135.04472	3.5763487	ES+	0.0223984	0.0146371	0.0150894	0.0097238	0.0116862	0.0116129
in-house	2,3-Diaminopropic	105.07016	3.3202935	ES+	0.024859	0.0207034	0.0225235	0.0201288	0.0226763	0.0226569
in-house	4-Methy2-oxovaler	129.07306	16.624045	ES+	0.1341287	0.2458095	0.2138968	0.2383272	0.1646037	0.2156238
in-house	5-Aminopentanoic	116.0542	3.9125471	ES+	0.015214	0.0157145	0.0152048	0.0139855	0.0148445	0.0151512
in-house	Acetylcarnitine	204.12263	3.8790521	ES+	0.503742	0.4063954	0.3690539	0.3346704	0.1894332	0.267591
HMDB	11-beta-hydroxyan	483.25453	21.64161	ES+	0.0352862	0.0143528	0.0117155	0.0149876	0.0110671	0.003493
HMDB	13-Hydroperoxylin	313.23515	21.000715	ES+	0.012489	0.0124697	0.0117186	0.0120185	0.0129048	0.0116153
HMDB	17-Hydroxylinolen	295.22749	19.925457	ES+	0.0141132	0.0156397	0.0151444	0.0142477	0.0153367	0.015173
HMDB	2,4-Diaminobutyri	119.0844	3.8790898	ES+	0.0636478	0.0838566	0.0635174	0.067999	0.0942851	0.0625007
HMDB	2,6 dimethylheptar	302.23203	18.02586	ES+	0.0031349	0.0042189	0.0027814	0.0082044	0.002749	0.0032303
HMDB	2-Ethylhydracrylic	119.07199	15.226531	ES+	0.0236145	0.0239315	0.0242947	0.0237831	0.0239368	0.0242611
HMDB	2-Ketohexanoic ac	131.07027	3.7353582	ES+	0.0038071	0.0051703	0.0041894	0.0056894	0.0057567	0.0036369

Xiuxia Du

Running XCMS yourself

- Besides the XCMSonline version, you can run XCMS locally on your own computer
- Check the attached set of instructions to download R and Rstudio
- Run XCMS in Rstudio
  - Check with Steve Barnes for a script to run XCMS

Instructions for running XCMS

- Launch Rstudio and enter the following lines plus <return>

```
> installed.packages()
> source("http://bioconductor.org/biocLite.R")
> biocLite("xcms", dep=T)
> biocLite("CAMERA")
> library(multtest) # multiple hypothesis testing
> library(xcms)
> library(faahKO)
> cdfpath <- system.file("cdf", package = "faahKO")
> list.files(cdfpath, recursive = TRUE)
```

Instructions for running XCMS(2)

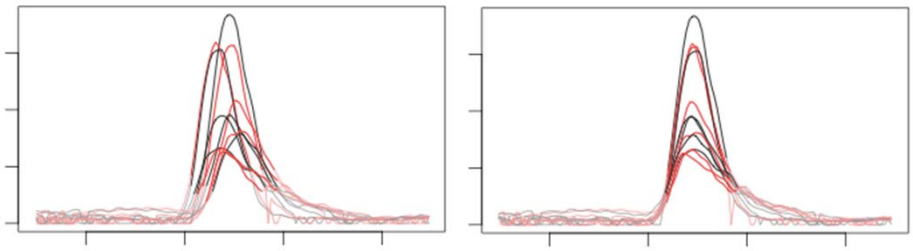
```
> cdffiles <- list.files(cdfpath, recursive = TRUE, full.names = TRUE)
> xset <- xcmsSet(cdffiles)
> xset
> xset <- group(xset)
> xset2 <- retcor(xset, family = "symmetric", plottype = "mdevden")
> xset2 <- group(xset2, bw = 10)
> xset3 <- fillPeaks(xset2)
> xset3
```

Instructions for running XCMS(3)

```
> reporttab <- diffreport(xset3, "WT", "KO", "example", 10, metlin = 0.15, h=480, w=640)
> reporttab[1:4,]
> gt <- groups(xset3)
> colnames(gt)
> groupidx1 <- which(gt[,"rtmed"] > 2600 & gt[,"rtmed"] < 2700 & gt[,"npeaks"] == 12)
.> groupidx2 <- which(gt[,"rtmed"] > 3600 & gt[,"rtmed"] < 3700 & gt[,"npeaks"] == 12)
> eiccor <- getEIC(xset3, groupidx = c(groupidx1, groupidx2))
> eicraw <- getEIC(xset3, groupidx = c(groupidx1, groupidx2), rt = "raw")
```

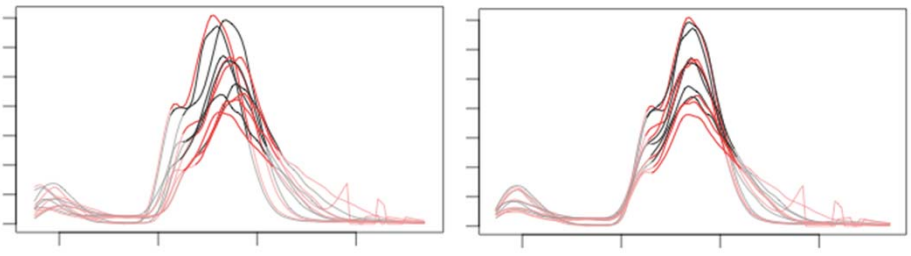
Plotting the data in XCMS

```
> plot(eicraw, xset3, groupidx = 1)
> plot(eiccor, xset3, groupidx = 1)
```



Plotting the data in XCMS

```
> plot(eicraw, xset3, groupidx = 2)
> plot(eiccor, xset3, groupidx = 2)
```



Plotting the data in XCMS

```
> plot(eicraw, xset3, groupidx = 1)
> plot(eiccor, xset3, groupidx = 1)
```