

# **Pre-processing of LC-MS Untargeted Metabolomics Data**

Xiuxia Du

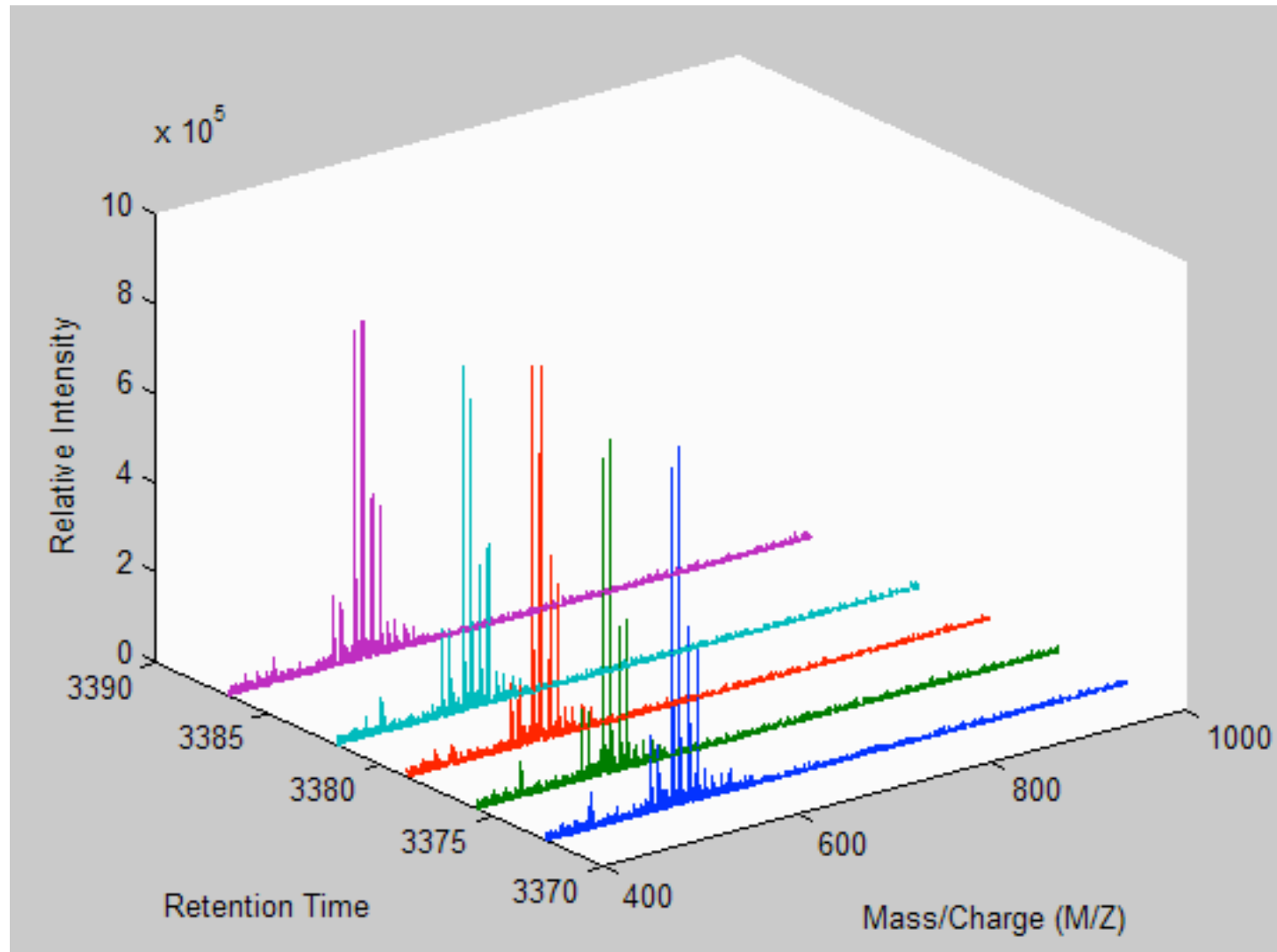
Department of Bioinformatics & Genomics  
University of North Carolina at Charlotte

July 24, 2013

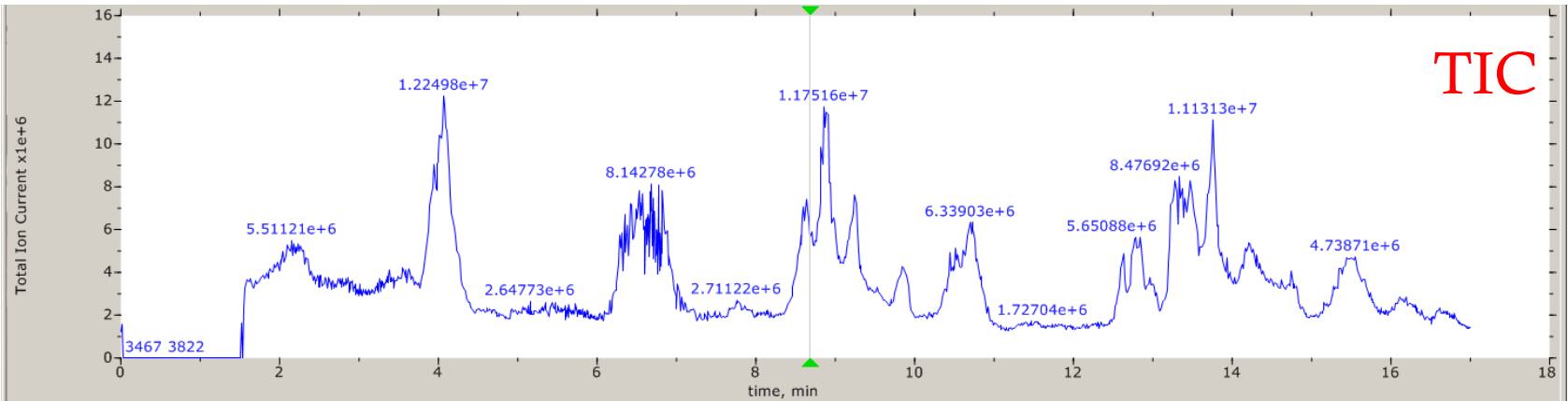
# Outline

- Introduction
- Basic steps of data pre-processing
- Software tools
- Databases

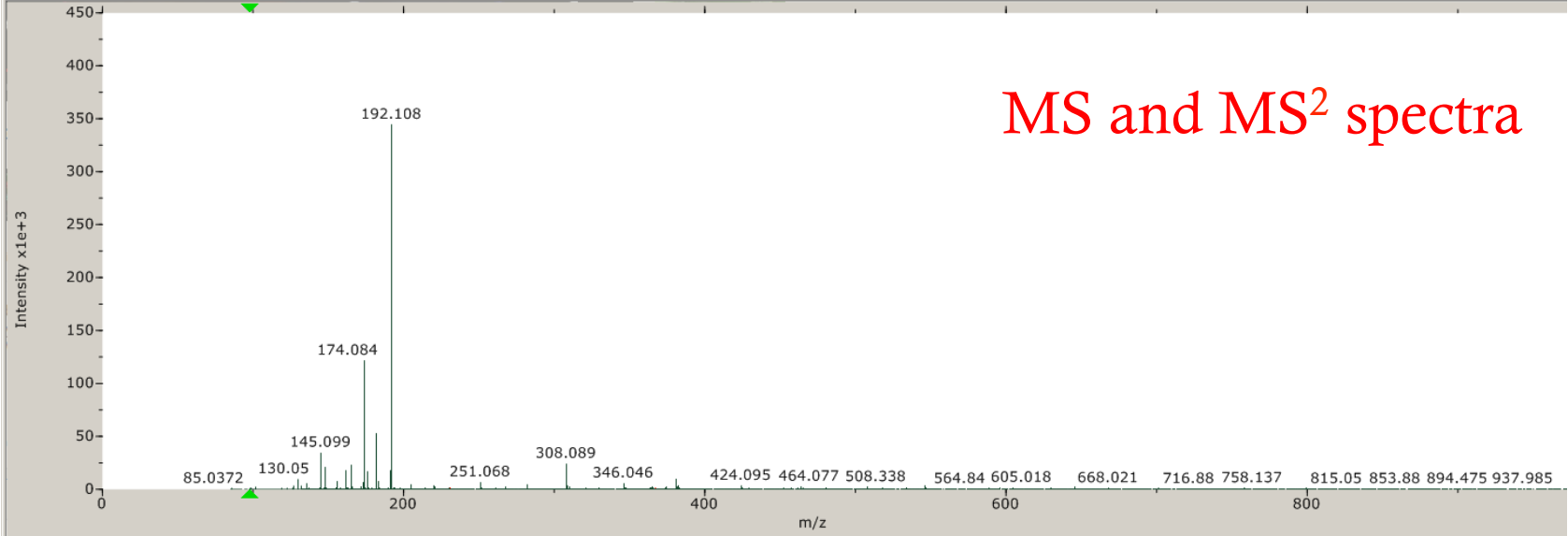
# Raw LC-MS data



# Raw LC-MS data



Scan #1505 (t=8.68 BP=344097.00@m/z=192.11); intensity=867.00@m/z=98.08 Tandem Scan: t=8.69 BP=238.00@m/z=0.00

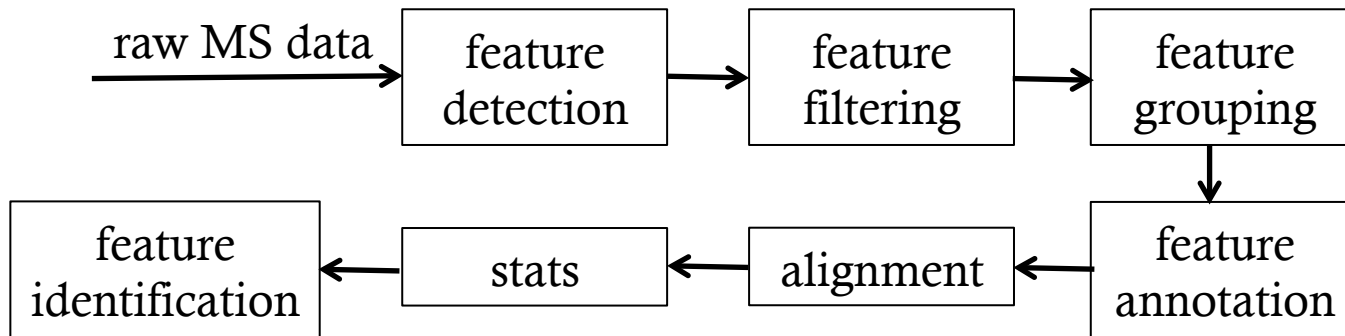


Scan #1505 | Scan #1506 (98.08) | Scan #1507 (231.17) | Scan #1508 (367.08)

# Goals of pre-processing

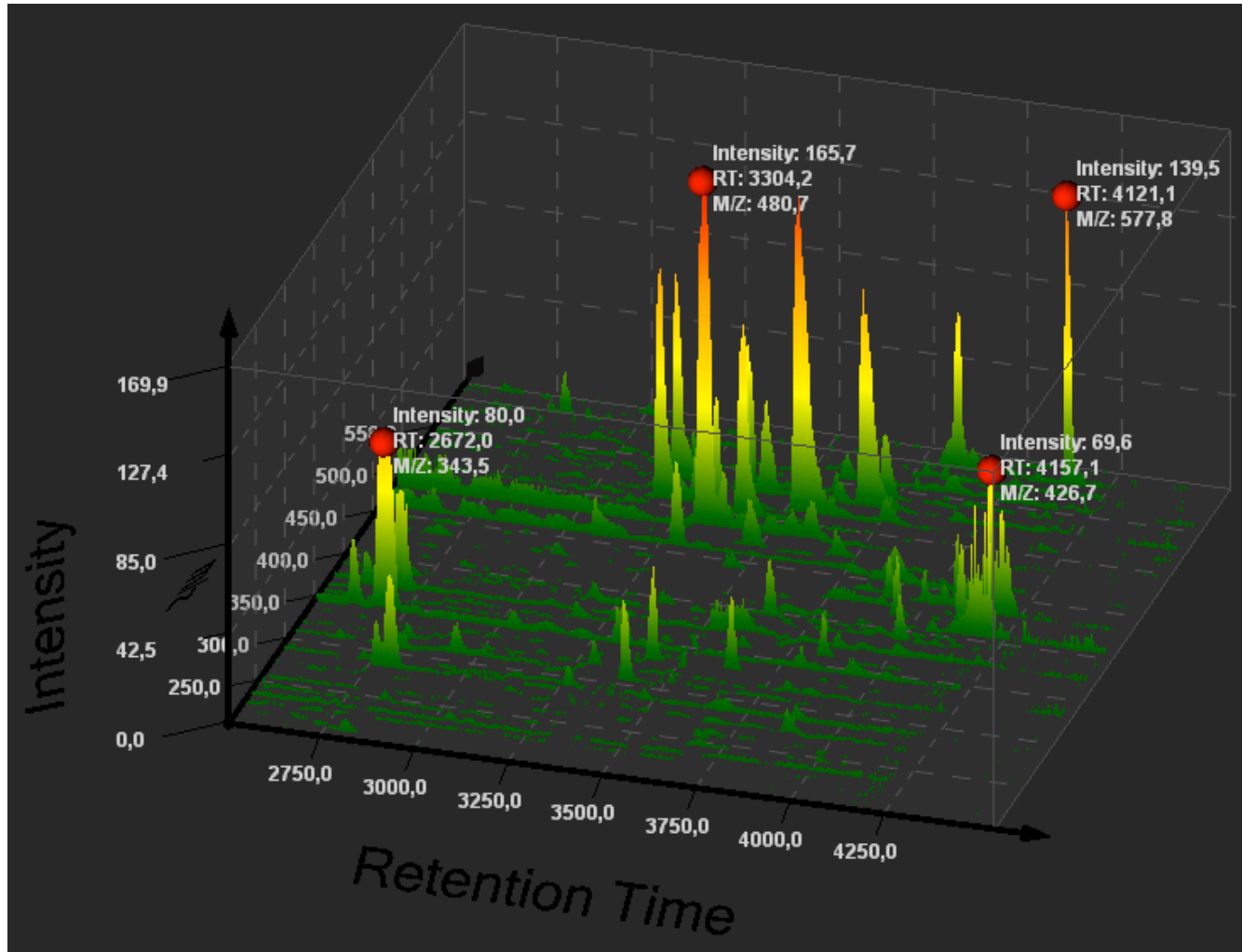
- Extract qualitative and quantitative information of possible metabolites
  - Determine the identity
  - Estimate the relative abundance
- Align samples to correct retention time shifts
- Produce a table of possible metabolites with their quantitative information for subsequent statistical analysis

# Work flow

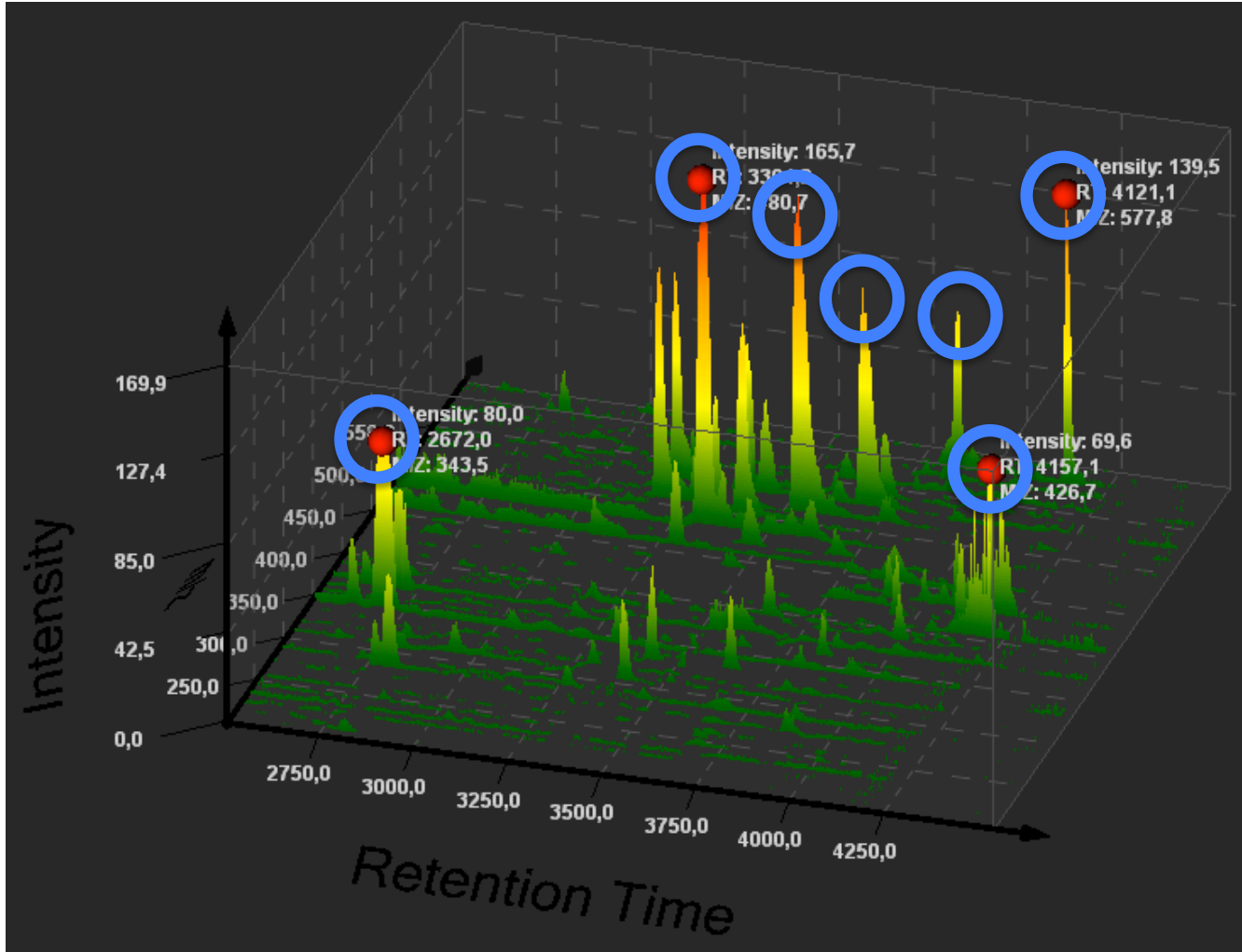


- **Feature:** a 3D signal induced by a single ion species (e.g.  $[M+H]^+$  or  $[M-H]^-$  of a compound)

# Features



# Features

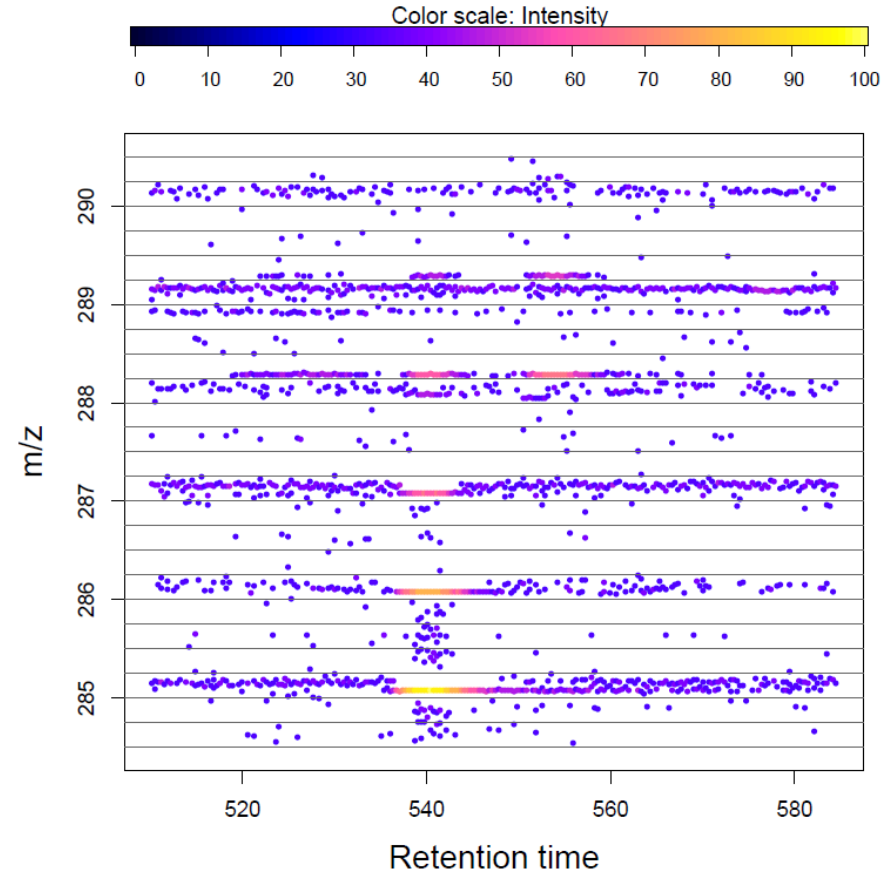
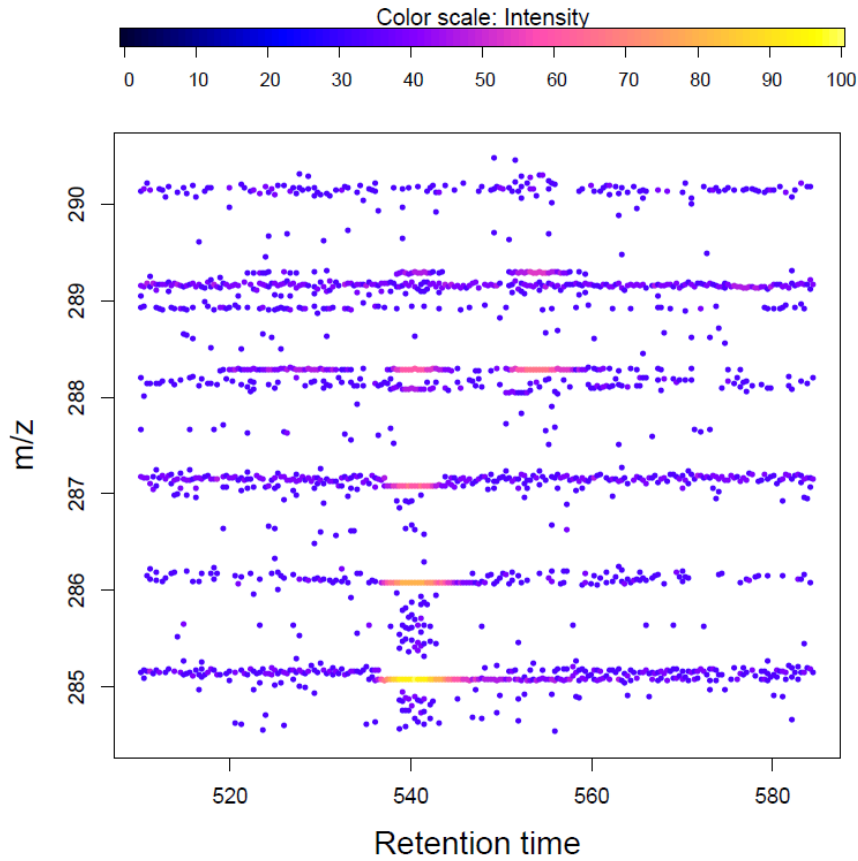




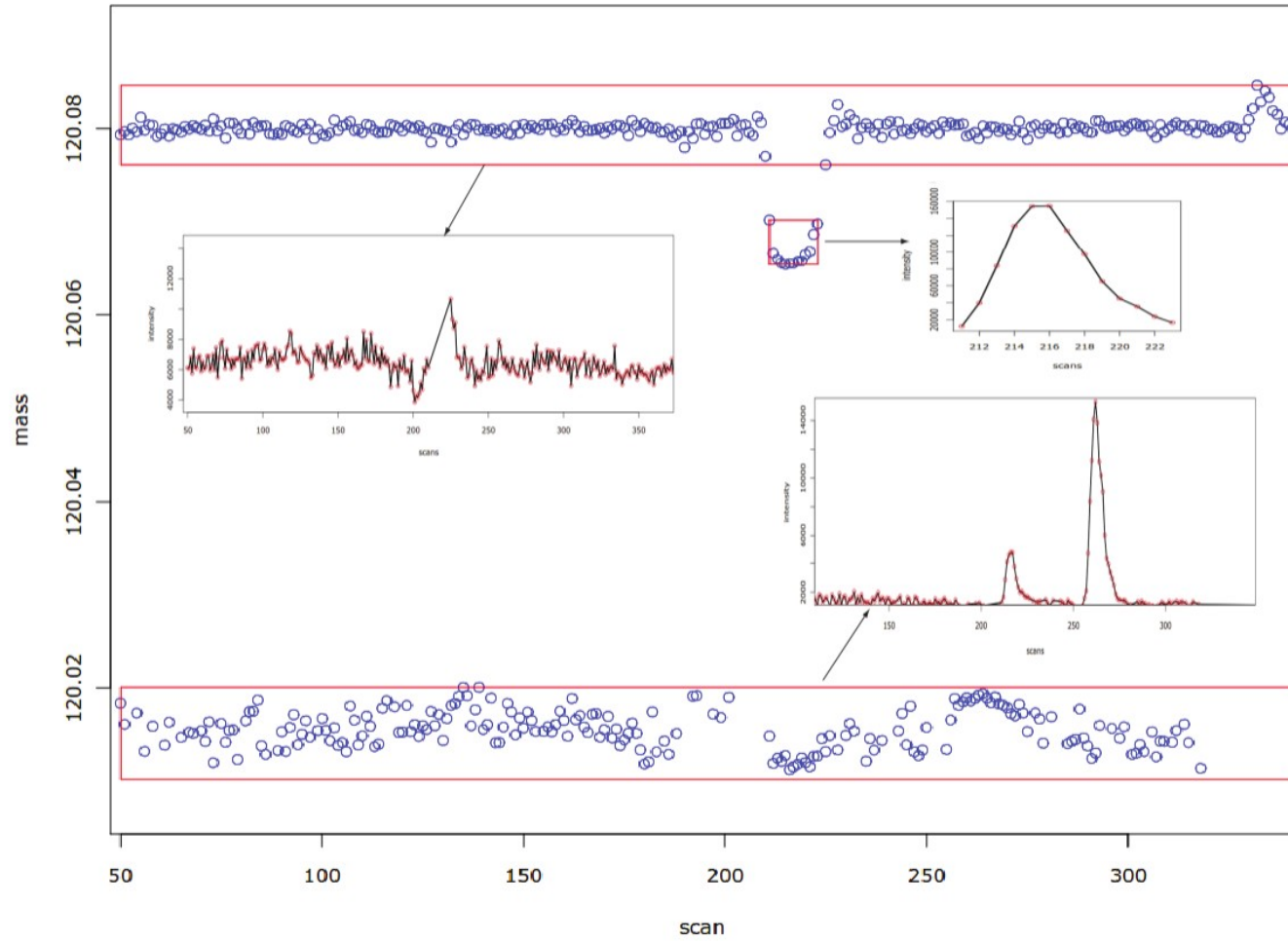
# Feature detection

- 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

# Binning

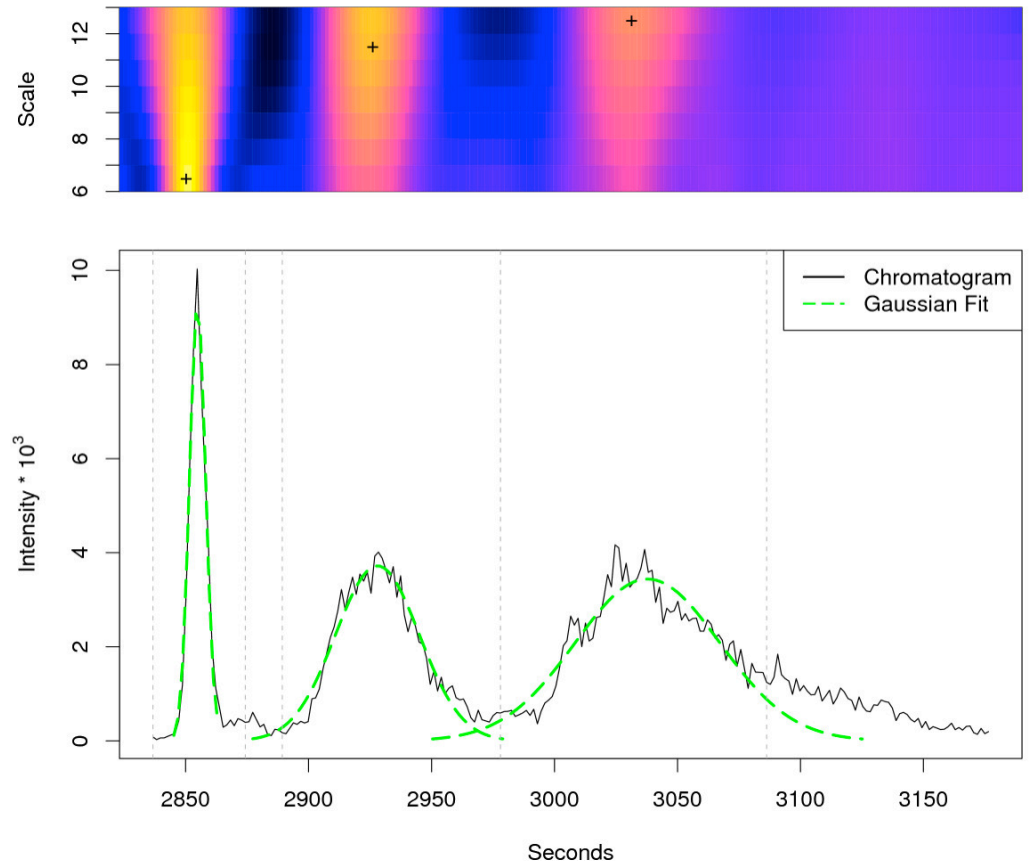
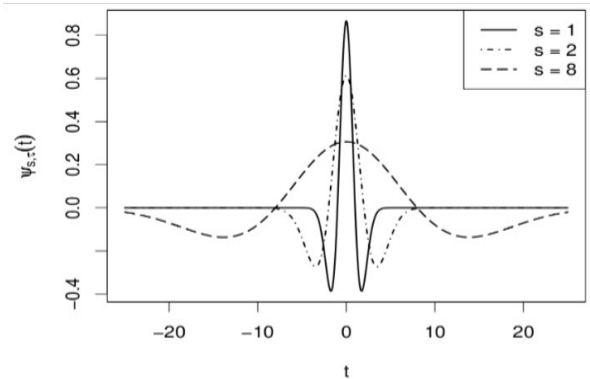


# ROI



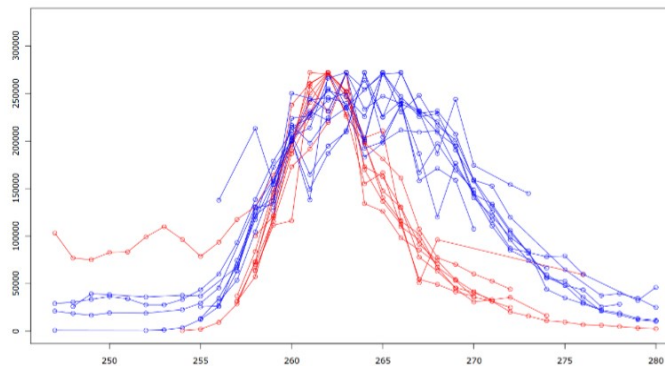
# Detect chromatographic peaks

- Use wavelet transform



# Feature filtering and grouping

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



# Feature annotation

$$m/z = (N * \text{compound mass} + \text{mass shift}) / CS$$

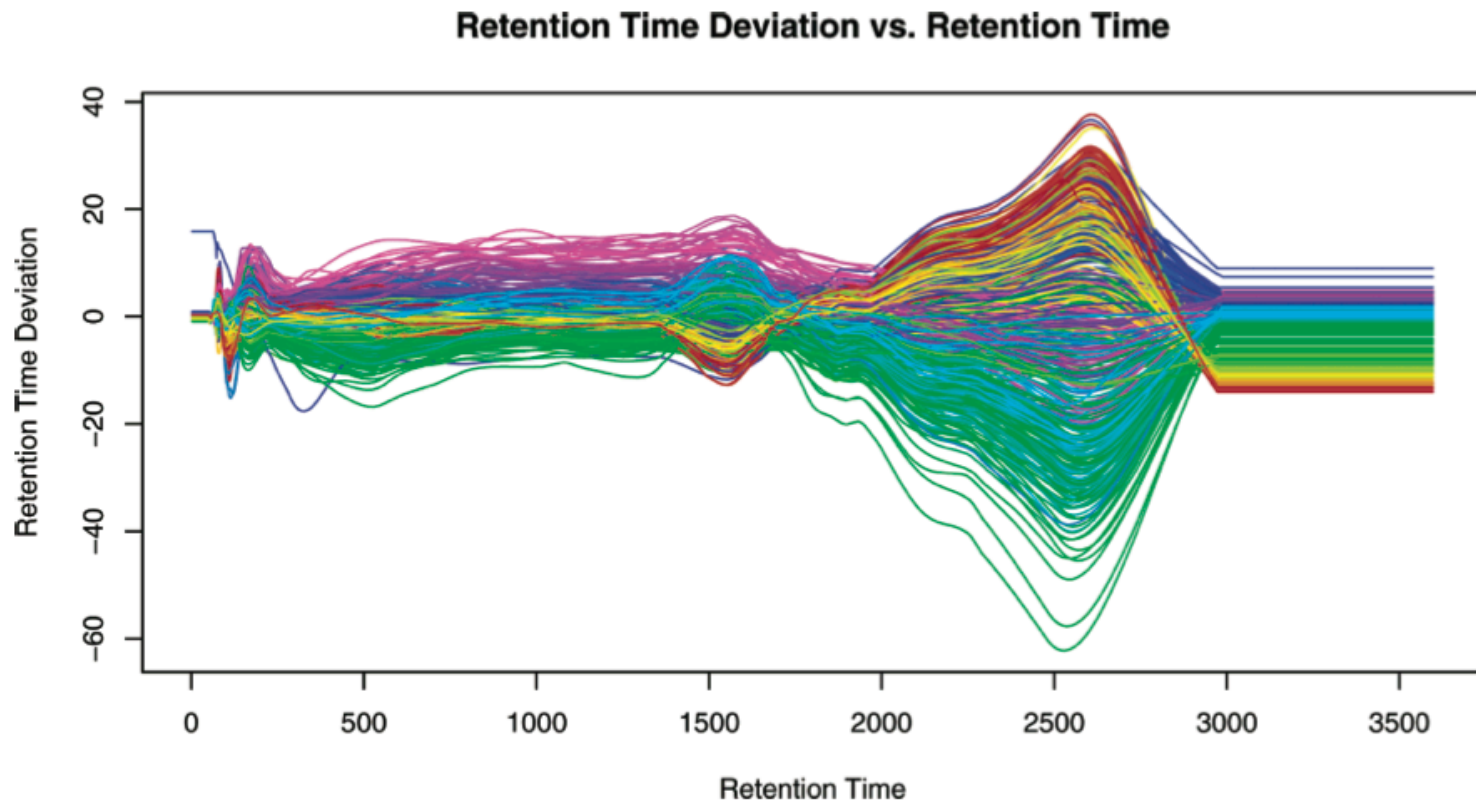
Formula	N	Mass shift
[M+H] <sup>+</sup>	1	1.007276
[M+2H] <sup>+</sup>	1	2.014552
[M+3H] <sup>+</sup>	1	3.021828
[M+Na] <sup>+</sup>	1	22.98977
[M+K] <sup>+</sup>	1	38.963708
[M-C <sub>3</sub> H <sub>9</sub> N] <sup>+</sup>	1	-59.073499
[M+2Na-H] <sup>+</sup>	1	44.96563
[2M+Na] <sup>+</sup>	2	22.98977
[M+H-NH <sub>3</sub> ] <sup>+</sup>	1	-16.01872
[2M+H] <sup>+</sup>	2	1.007276
[M-OH] <sup>+</sup>	1	-17.0028

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

- using the R package CAMERA
- <http://bioconductor.org/packages/devel/bioc/html/CAMERA.html>

# Alignment

- Goal: Correct retention time shift from run to run



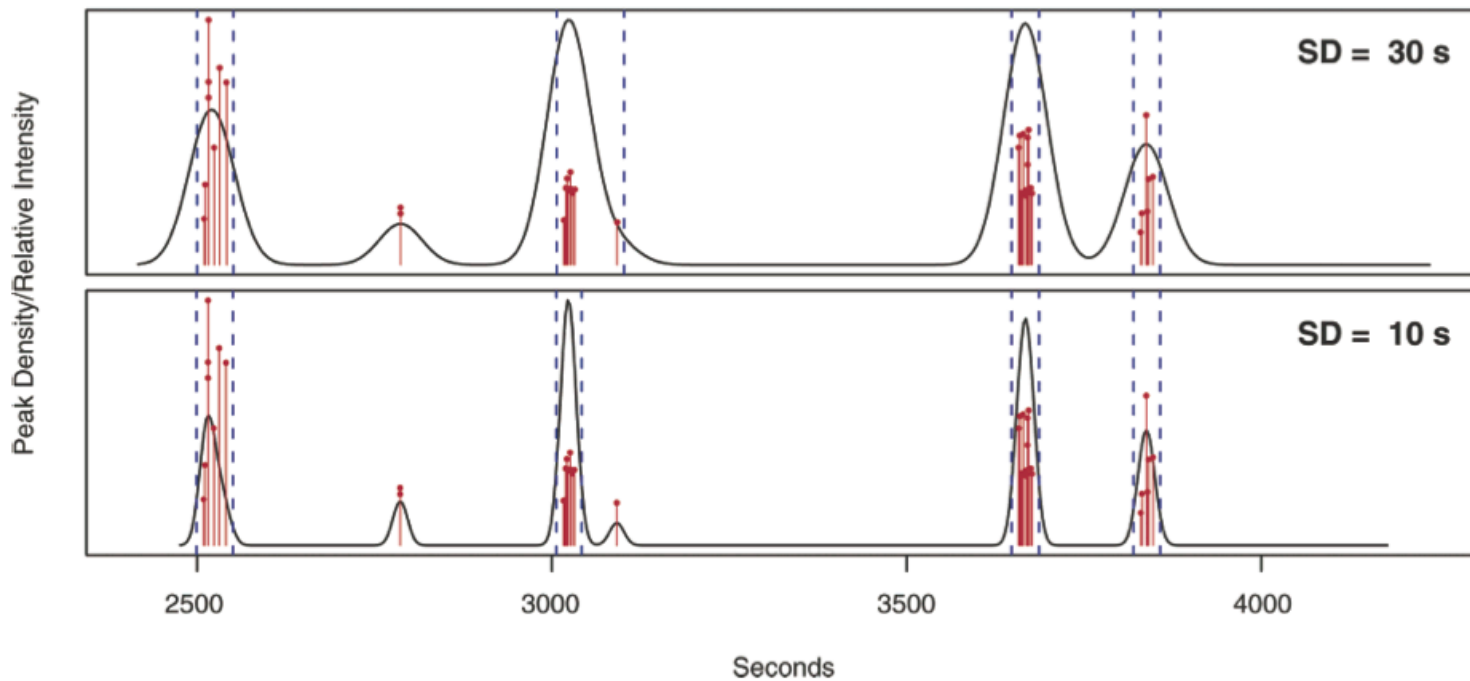
# Alignment

- Alignment method in XCMS
  - Use “well-behaved” peak groups
  - For each well-behaved group, calculate the median retention time and, for every sample, a deviation from that median
  - Within a sample, the deviation generally changes over time in a nonlinear fashion.
  - Those changes are approximated using a local polynomial regression (loess).



# Alignment

Grouping of Peaks in Mass Bin: 337.975 – 338.225 m/z



# Software tools

- Proprietary

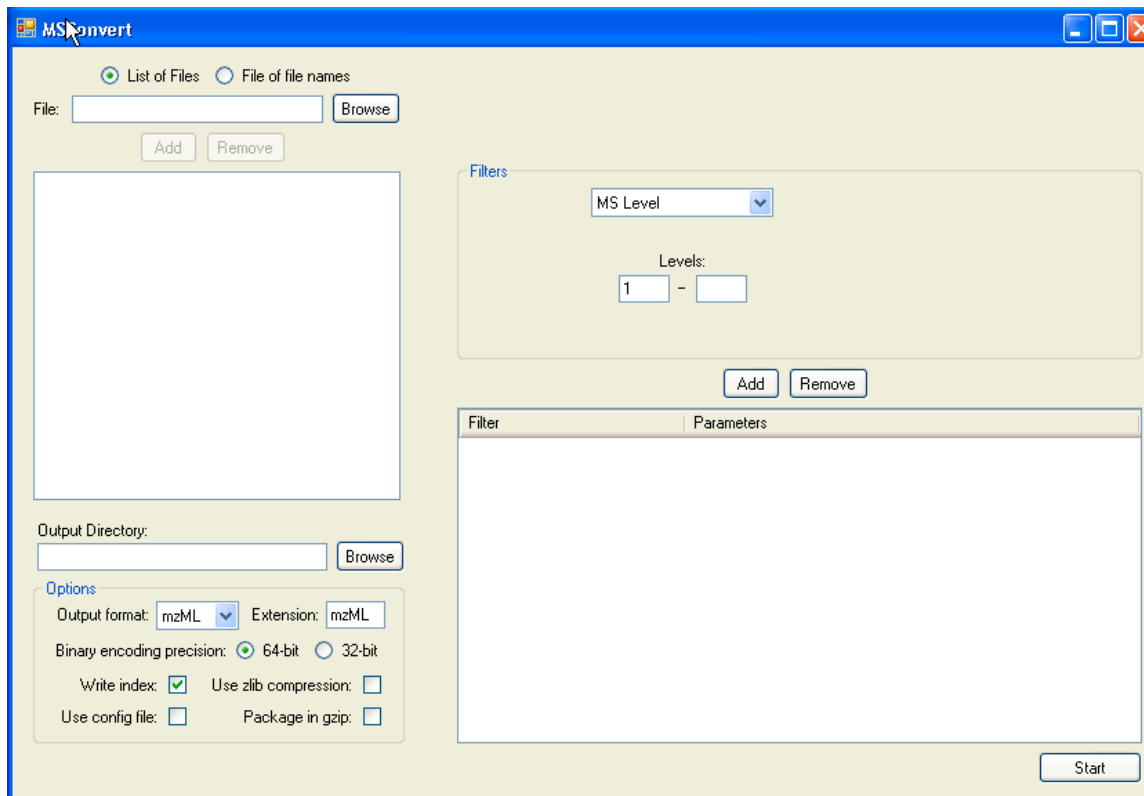
Vendor	Format	Viz	Software
Agilent	.d	MassHunter	Mass Profiler Pro
AB Sciex	.wiff		MarkerView
Waters	.raw	MassLynx	MarkerLynx
Bruker	.d, YEP, BAF, FID	Compass	Metabolic Profiler
Thermo	.raw	Xcalibur	

- Freely available

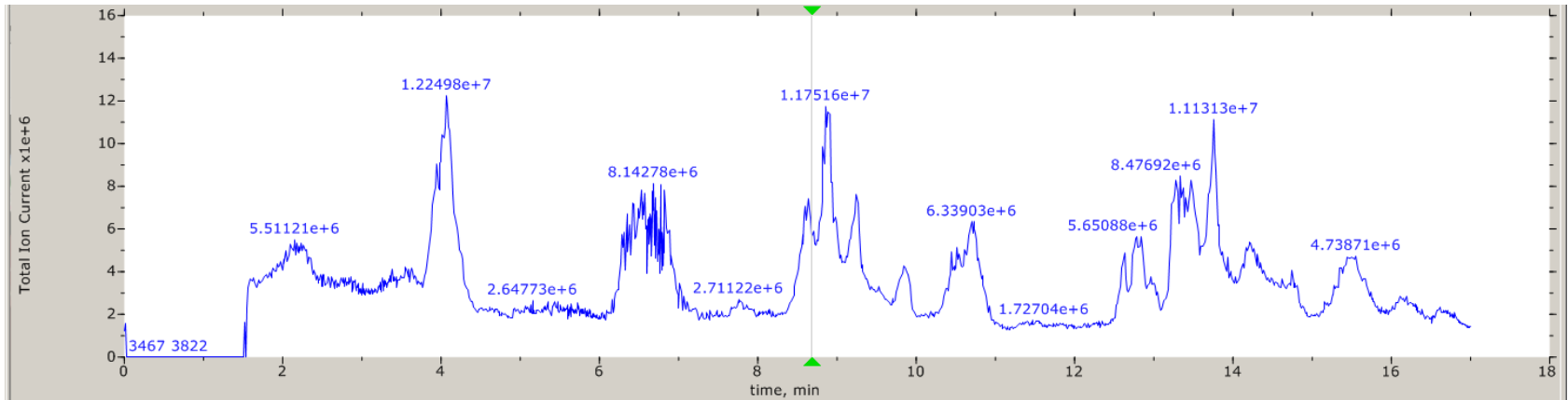
- **XCMS**      -- MAVEN
- MZmine      -- MetAlign

# File format conversion

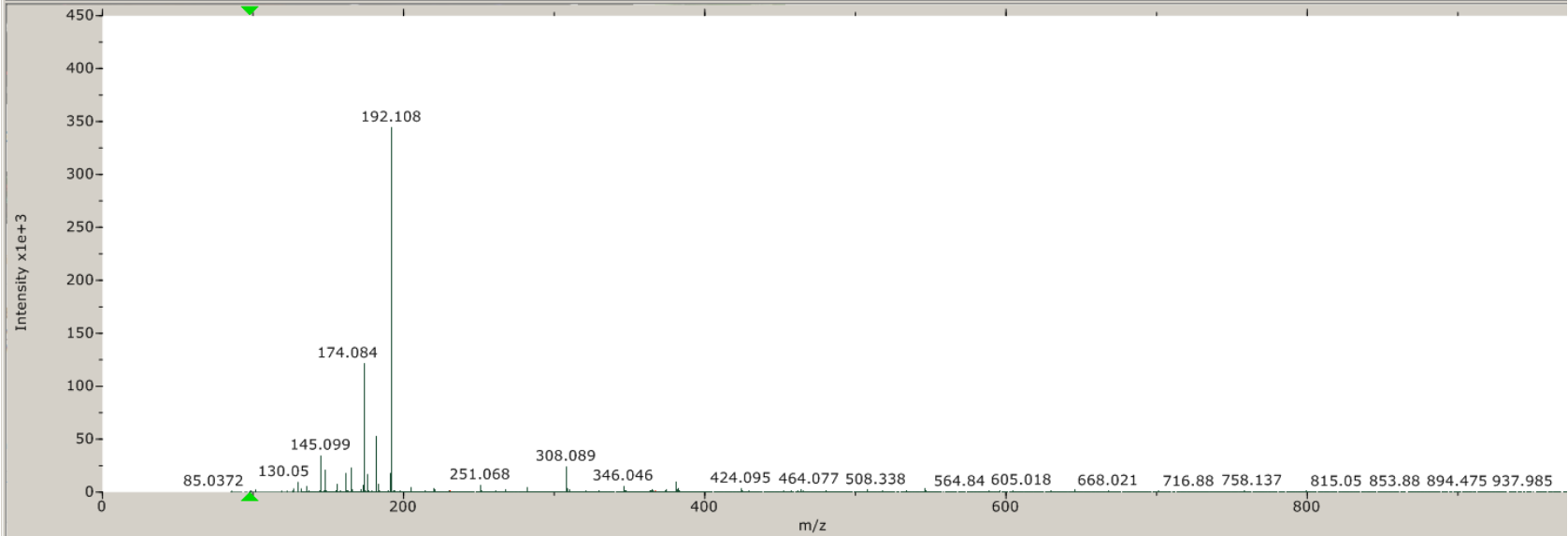
- Convert files to mzXML using MSConvert



# Raw data viz in Incilicos Viewer



Scan #1505 (t=8.68 BP=344097.00@m/z=192.11); intensity=867.00@m/z=98.08 Tandem Scan: t=8.69 BP=238.00@m/z=0.00



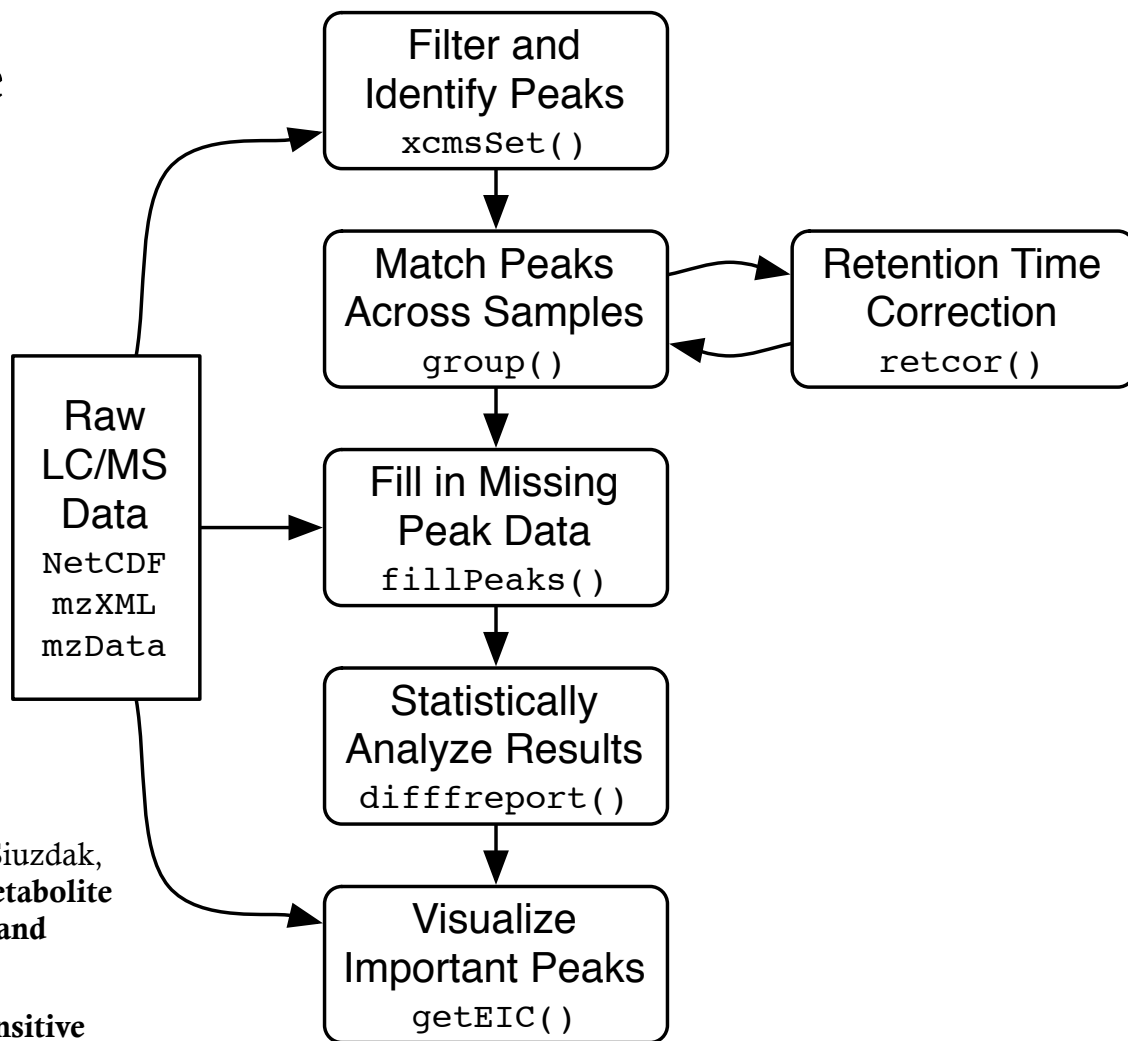
Scan #1505 | Scan #1506 (98.08) | Scan #1507 (231.17) | Scan #1508 (367.08)

Ready

NUM

# XCMS

- Free and open source
- Written in R



Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G., **XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.** *Analytical chemistry* **2006**, 78 (3), 779-87.

Tautenhahn, R.; Bottcher, C.; Neumann, S., **Highly sensitive feature detection for high resolution LC/MS.** *BMC bioinformatics* **2008**, 9, 504.

# XCMS Online

https://xcmsonline.scripps.edu

Scripps Center  
For Metabolomics  
XCMS Online

Home

Public Shares

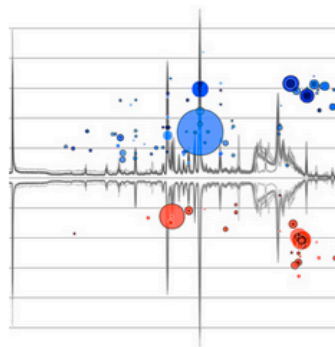
FAQ

Publications

Contact

METLIN

XCMS  
Online



Click here to view the new Interactive Cloud Plot!

XCMS Online is a cloud-based metabolomic data processing platform that provides high-quality metabolomic analysis in a

## New to XCMS Online?

Simple mass spectrometry data processing.

[Register >](#)

## Current Users:

E-mail:

(e.g. researcher@scripps.edu)

Password:

[Sign In](#)

[Forgot your password?](#)

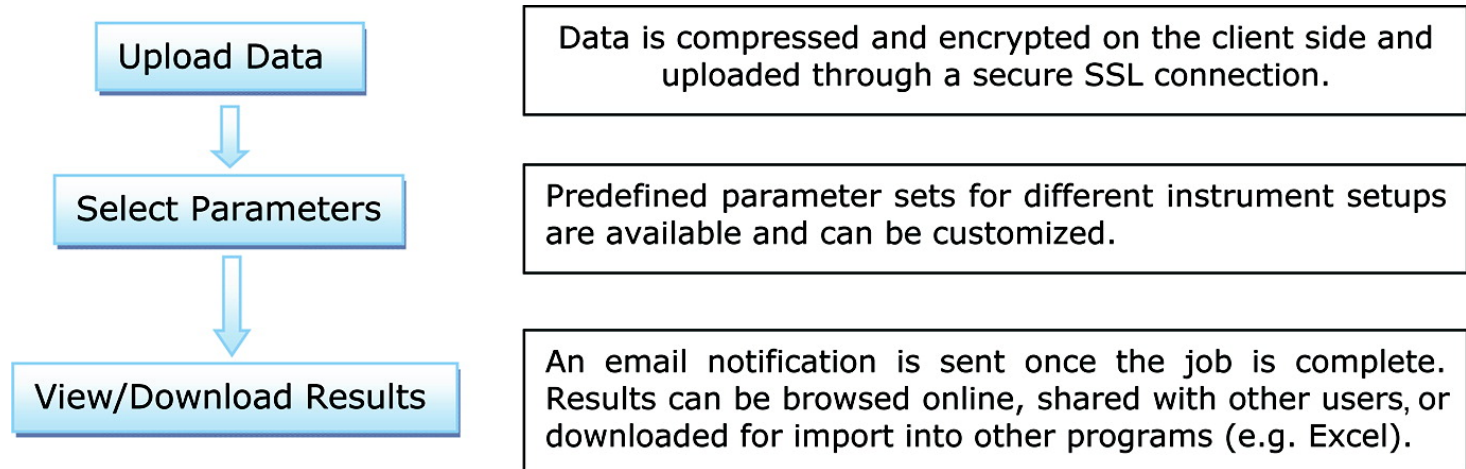
[DOWNLOAD USER MANUAL](#)

Demo login (read-only) for testing:  
Demo Dataset that can be used for testing

Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G., **XCMS Online: a web-based platform to process untargeted metabolomic data.** *Analytical chemistry* **2012**, *84* (11), 5035-9.

# XCMS Online

- Workflow

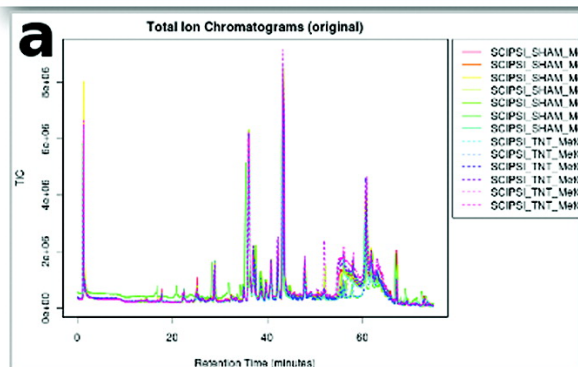


# XCMS Online

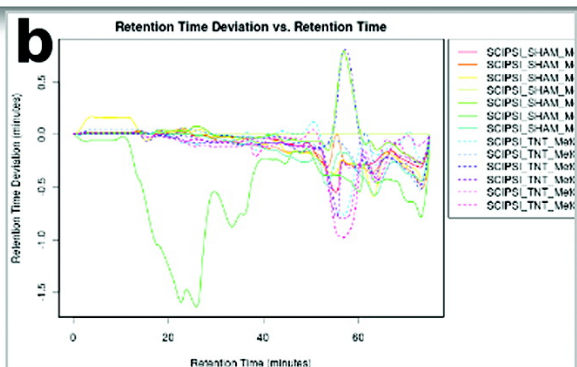
[Home](#)
[Create Job](#)
[View Results](#)
[Stored Datasets](#)
[FAQ](#)
[Account](#)
[Contact](#)
[Logout](#)

<b>Job ID:</b>	109981	<b>Job Name:</b>	SHAM vs. TNT	<b>Create Date:</b>	2012-04-06 17:14:17
<b>Parameter (ID#):</b>	<a href="#">HPLC / Q-TOF (2583)</a>	<b>Log:</b>	<a href="#">View Log</a>	<b>Finish Date:</b>	2012-04-06 17:28:03
<b>Status:</b>	job complete	<b>Total Aligned Features:</b>	8514		
<b>Datasets Used:</b>	<a href="#">SHAM(ID#2831)</a> [control] <a href="#">TNT(ID#2830)</a>				

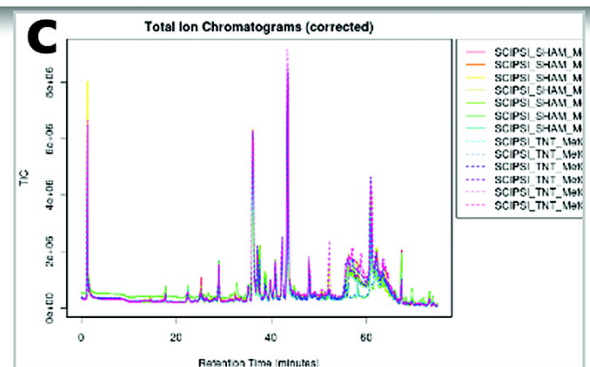
[BROWSE RESULT TABLE](#)



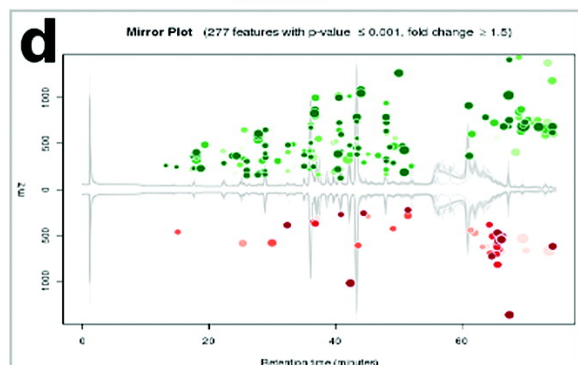
[PNG PDF](#)



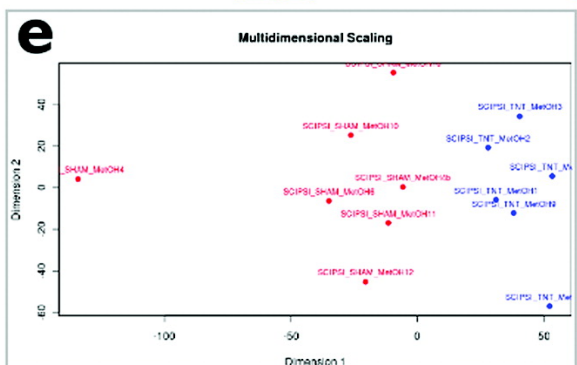
[PNG PDF](#)



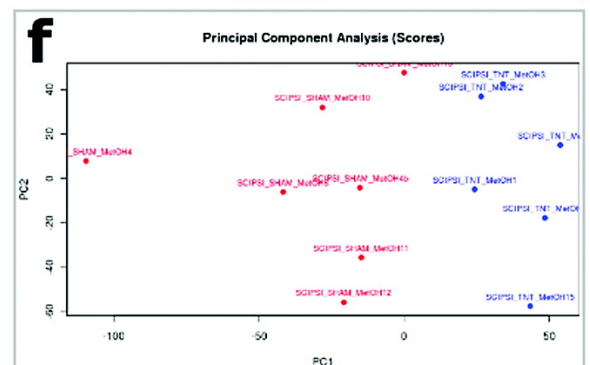
[PNG PDF](#)



[PNG PDF](#)



[PNG PDF](#)



[PNG PDF](#)



# XCMS Online

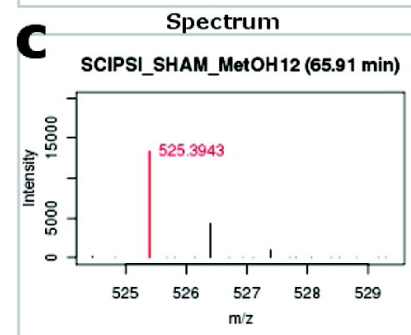
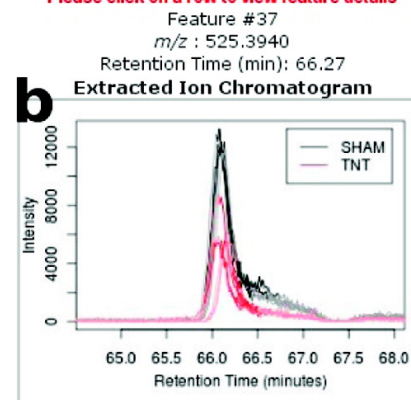
Job#102778 : SHAM vs. TNT

Columns Show isotopic peaks Page 1 of 64 100 View 1 - 100 of 6 302

Feature	name	fold chang	p-value	m/z	retention time	MaxInt	SHAM(x)	TNT(x)	isotopes	adducts	feature g	Note
a	M365T24_1	3.5	1.01128e-1	365.1374	23.77	2,577	14,191	50,192			329	
	M151T26	3.0	1.14569e-1	151.1124	26.03	8,940	42,005	127,894		[2M+2Na+K]3+ 18	196	
14	M260T44	2.1	1.53363e-1	260.2593	44.48	1,575	23,331	11,245		[M+2Na-H]+ 215.2	236	
15	M649T29	2.5	1.97112e-1	649.4621	28.99	1,142	4,092	10,050	[329][M]+	[2M+2Na-H]+ 302	13	
17	M755T68	4.8	2.40615e-1	755.4356	67.52	377	694	3,339			21	
18	M480T42	2.0	3.33177e-1	480.3685	42.32	545	3,024	5,910		[M+NH4]+ 462.33	18	
19	M405T18	4.3	3.43724e-1	405.0974	18.07	324	439	1,883		[M+K]+ 366.134	362	
20	M204T29	1.5	3.93833e-1	204.1417	28.94	1,244	11,150	16,728			13	
21	M746T68	3.4	4.41636e-1	746.4956	67.54	1,416	3,978	13,443		[M+Na+K]2+ 1431	21	
22	M428T51	10.8	4.50028e-1	428.3552	50.92	2,269	2,834	30,584			76	
23	M734T70	3.1	5.12129e-1	733.5041	69.98	1,144	3,565	11,173		[M+Na+NaCOOH]	392	
26	M207T26	3.0	5.81910e-1	207.1755	26.04	9,652	47,259	142,503	[1148][M]+	[M+Na]+ 184.19 [	196	
28	M393T29	1.8	6.38174e-1	393.2241	28.95	2,084	14,675	25,803	[319][M]+	[M+Na+NaCOOH]	13	
29	M623T74	2.7	6.49237e-1	622.5545	74.37	450	4,037	1,485		[M+H-CH4]+ 637.4	71	
32	M187T51	8.8	8.13620e-1	187.0770	50.92	2,907	4,640	40,658		[M+H]+ 186.066	76	
34	M1269T50	7.1	8.93163e-1	1,269.4214	50.11	439	1,172	8,334	[1093][M]+		170	
35	M477T51_1	2.7	8.98636e-1	477.3203	50.63	1,070	5,084	13,874		[M+Na+NaCOOH]	84	
37	M525T66	2.0	0.00001	525.3940	66.27	13,284	169,432	82,654	[741][M]+	[M+Na]+ 502.405	58	
38	M707T70	2.9	0.00001	707.4880	69.74	1,725	6,729	19,389		[M+Na+NaCOOH]	33	
39	M452T36	3.3	0.00001	452.3369	36.27	376	1,222	4,003			195	
40	M354T36_1	2.1	0.00001	353.5067	36.18	28,080	194,104	402,649	[46][M]+	[M+Na]+ 330.509	2	
41	M289T49	2.2	0.00001	289.2519	49.18	7,821	112,389	245,613	[923][M]+	[M+Na]+ 266.263	95	
42	M1371T68	3.8	0.00001	1,370.9600	67.57	299	2,994	783	[480][M]+	[M+K]+ 1332	21	
43	M225T52	2.1	0.00001	225.2578	51.51	2,000	31,119	14,636		[M+H-HCOOH]+ 2	76	
44	M381T43	2.0	0.00002	381.2995	43.47	632,018	8,412,930	16,961,253	[11][M]+	[M+Na]+ 358.311	1	
45	M185T36_1	1.4	0.00002	185.1204	35.98	5,451	51,360	73,140			35	
46	M1024T42	3.9	0.00002	1,024.4198	42.44	289	3,813	976			417	
47	M1028T67	15.0	0.00002	1,027.5880	67.44	332	208	3,133			212	
48	M261T13_2	1.6	0.00002	261.1114	13.24	2,833	14,162	22,965			174	
50	M220T43_2	1.4	0.00002	220.1488	43.28	3,410	40,940	55,979			234	

Columns Export Page 1 of 64 100 View 1 - 100 of 6 302

Please click on a row to view feature details



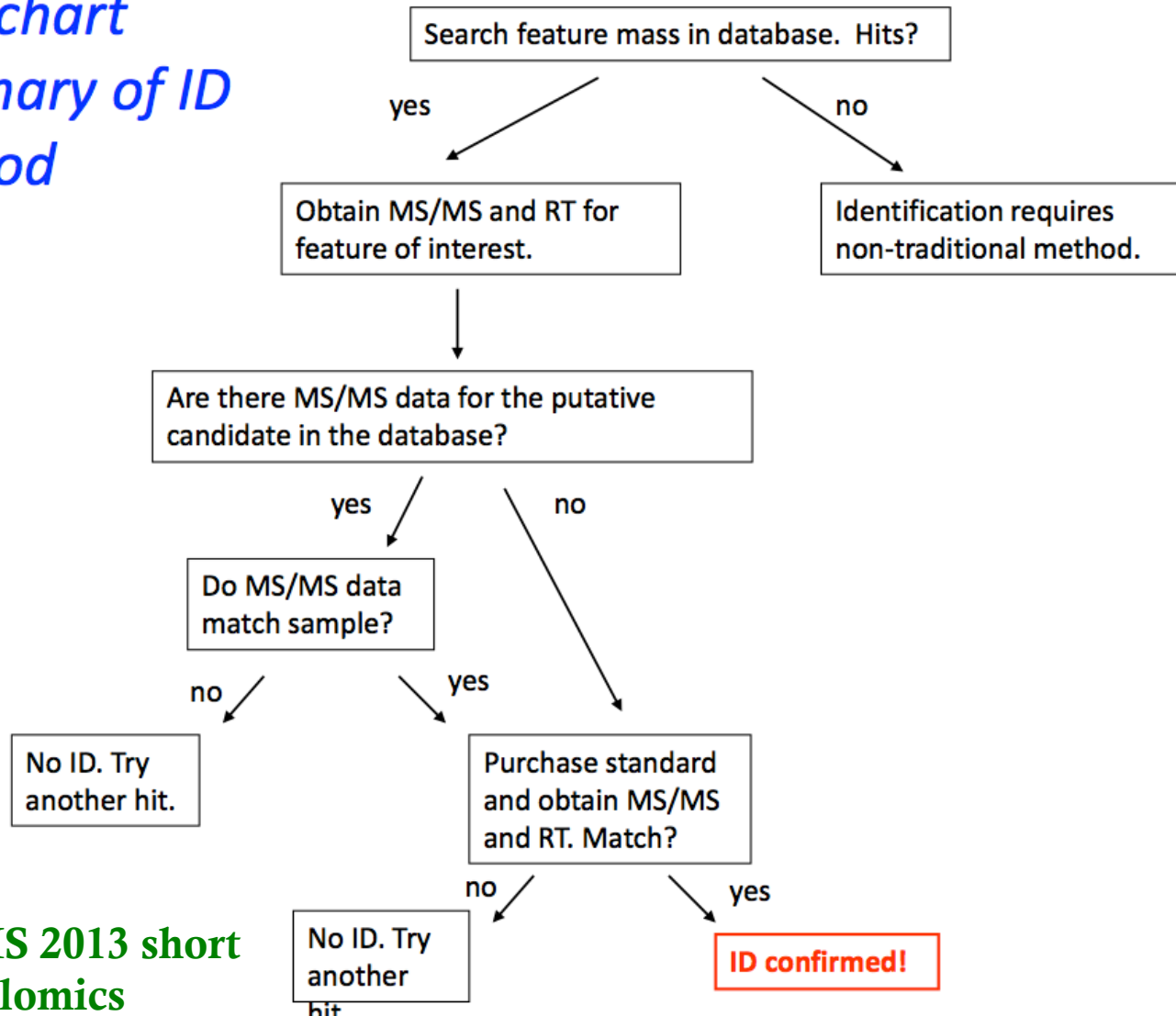
PPM	Name	Adduct	METLINID
5	1 $\alpha$ -hydroxy-2 $\beta$ -	M+Na	<a href="#">42451</a>
5	1 $\alpha$ ,25-dihydroxy-	M+Na	<a href="#">42452</a>
5	(20S)-1 $\alpha$ ,25-di-	M+Na	<a href="#">42453</a>
5	(20S)-1 $\alpha$ ,25-di-	M+Na	<a href="#">42454</a>
5	26,27-diethyl-1-	M+Na	<a href="#">42455</a>
5	1 $\alpha$ ,25-Dihydro-	M+Na	<a href="#">42539</a>

# Feature identification

- **Current bottleneck of metabolomics**
- Use three pieces of information
  - molecular mass
  - retention time
  - MS/MS spectra
- Match with metabolites in databases

# Feature identification

*Flow chart  
summary of ID  
method*



# Databases

- Spectral databases
  - NIST
  - HMDB
  - Metlin
  - Fiehn GC-MS Database
  - BMRB
  - MMCD
  - MassBank
  - Golm Metabolome Database
- Compound databases
  - PubChem
  - ChEBI
  - ChemSpider
  - KEGG Glycan
  - LIPID MAPS
- Pathway databases
  - KEGG
  - MetaCyc
  - HumanCyc
  - BioCyc
  - Reactome
- Drug databases
  - DrugBank
  - PharmGKB
  - SuperTarget
  - Therapeutic Target DB
  - STITCH

# HMDB

- <http://www.hmdb.ca>
- HMDB is a freely available electronic database containing detailed information about small molecule metabolite found in the human body.

# HMDB: metabolite statistics

Description	Count
Total Number of Metabolite Synonyms	333,434
Total Number of Non-redundant Metabolites with Spectra	1,343
Total Number of Metabolites Having Associated Proteins (Enzymes and Transporters)	22,139
Total Number of Metabolite in the Human Metabolome Library (HML)	1,027
Total Number of Metabolites with Synthesis Records	1,622
Total Number of Detected and Quantified Metabolites	5,044
Total Number of Detected and Not Quantified Metabolites	15,846
Total Number of Expected Metabolites	19,674
Total Number of Metabolites	40,564
Total Number of Endogenous Metabolites	29,331
Total Number of Food Metabolites	32,513
Total Number of Microbial Metabolites	171
Total Number of Drug Metabolites	1,921
Total Number of Plant Metabolites	149
Total Number of Toxin/Pollutant Metabolites	164
Total Number of Cosmetic Metabolites	17

# HMDB: metabolite statistics

Description	Count
Total Number of Metabolite Synonyms	333,434
Total Number of Non-redundant Metabolites with Spectra	1,343
Total Number of Metabolites Having Associated Proteins (Enzymes and Transporters)	22,139
Total Number of Metabolite in the Human Metabolome Library (HML)	1,027
Total Number of Metabolites with Synthesis Records	1,622
Total Number of Detected and Quantified Metabolites	5,044
Total Number of Detected and Not Quantified Metabolites	15,846
Total Number of Expected Metabolites	19,674
Total Number of Metabolites	40,564
Total Number of Endogenous Metabolites	29,331
Total Number of Food Metabolites	32,513
Total Number of Microbial Metabolites	171
Total Number of Drug Metabolites	1,921
Total Number of Plant Metabolites	149
Total Number of Toxin/Pollutant Metabolites	164
Total Number of Cosmetic Metabolites	17

# HMDB: spectra statistics

Total Number Compounds with GC-EI-TOF MS/MS Spectra	808
Total Number Compounds with GC-MS Spectra	1,220
Total Number Compounds with LC-APPI-QQ MS/MS Spectra	13
Total Number Compounds with LC-ESI-IT MS/MS Spectra	158
Total Number Compounds with LC-ESI-ITFT MS/MS Spectra	1,058
Total Number Compounds with LC-ESI-ITTOF MS/MS Spectra	2
Total Number Compounds with LC-ESI-QIT MS/MS Spectra	22
Total Number Compounds with LC-ESI-QQ MS/MS Spectra	2,177
Total Number Compounds with LC-ESI-QTOF MS/MS Spectra	529





# HMDB: MS search

for



## Mass Spectrum Search

Query Masses (Da)

*Enter one mass per line ( maximum 150 query masses per request )*

Molecular Weight Tolerance  $\pm$   Da

Molecular Species

[Example](#)

# HMDB: MS search result

Download Results As CSV

MS search for 200.1 m/z *Delta = abs(query mass - adduct mass)*

Show 10 entries

Search:

Compound	Name	Adduct	Adduct MW (Da)	Compound MW (Da)	Delta
<a href="#">HMDB30454</a>	Macrophorin A	M+H+K	200.100247	360.23006	0.000247
<a href="#">HMDB31848</a>	Butylated hydroxyanisole	M+H+K	200.100247	360.23006	0.000247
<a href="#">HMDB03689</a>	Neuroprotectin D1	M+H+K	200.100247	360.23006	0.000247
<a href="#">HMDB04038</a>	Resolvin D5	M+H+K	200.100247	360.23006	0.000247
<a href="#">HMDB39510</a>	7-Hydroxy-2',3',4'-trimethoxyisoflavan	M+2ACN+2H	200.09936	316.131074	0.00064
<a href="#">HMDB38323</a>	Verimol B	M+2ACN+2H	200.09936	316.131074	0.00064
<a href="#">HMDB40534</a>	5'-Hydroxy-3',4',7-trimethoxyflavan	M+2ACN+2H	200.09936	316.131074	0.00064
<a href="#">HMDB38780</a>	7-Hydroxy-2',4',5'-trimethoxyisoflavan	M+2ACN+2H	200.09936	316.131074	0.00064
<a href="#">HMDB39607</a>	Sorgolactone	M+2ACN+2H	200.09936	316.131074	0.00064
<a href="#">HMDB38324</a>	Verimol A	M+2ACN+2H	200.09936	316.131074	0.00064

Showing 1 to 10 of 962 entries

◀ Previous Next ▶

# Metlin



- All data acquired on Agilent Q-TOF
- Search by MS, MS/MS, fragment, and neutral loss

## Statistics

- # Metabolites: 76,420
- # High Resolution MS/MS Spectra: 56,612
- # Metabolites w/ High Resolution MS/MS: 11,208

[example](#) | [details...](#)

## Functionality

- **Single & Batch**  
Precursor Ion ( $m/z$ ) searching
- **Single & Multiple**  
Fragment Ion ( $m/z$ ) searching
- **Neutral Loss** searching
- **De Novo** Fragment Characterization

# Metlin search

## METLIN: Metabolite Search Simple

[Simple](#) | [Advanced](#) | [Batch](#) | [Fragment](#) | [Neutral Loss](#) | [MS/MS Spectrum Match](#) | [Unknowns](#)

Mass:  Tolerance ( $\pm$ )  Da

Charge:	Neutral	M+H
	Positive	M+NH <sub>4</sub>
	Negative	M+Na
		M+H-2H <sub>2</sub> O
		M+H-H <sub>2</sub> O
		M+K
		M+ACN+H
		M+ACN+Na
		M+2Na-H
		M+2H
		M+3H
		M+H+Na
		M+2H+Na

M+2Na
M+2Na+H
M+Li
M+CH <sub>3</sub> OH+H

\*To select multiple Adducts:

 - Hit Ctrl + Adducts

 - Hit Command + Adducts

Select: **all** | **none**

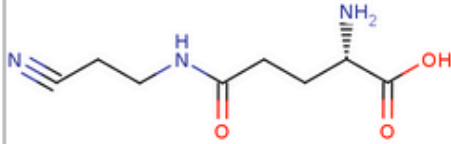
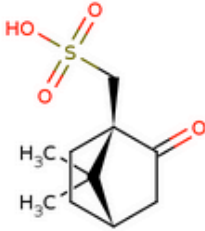
# Metlin search result

## METLIN Metabolites

Mass 200.1 with 0.1 Da mass accuracy

Change Query

### Total: 77 Metabolites







METLIN ID	MASS	$\Delta$ ppm	NAME	MS/MS	STRUCTURE
63620	[M+H] <sup>+</sup> <u>m/z</u> 200.1030 M 199.0957	14	<b><math>\gamma</math>-Glutamyl-<math>\beta</math>-aminopropionitrile</b> <i>Formula: C<sub>8</sub>H<sub>13</sub>N<sub>3</sub>O<sub>3</sub></i> <i>CAS: 3144-16-9</i>	NO	 <p>The structure shows a nitrile group (C≡N) at the end of a propionyl chain, which is linked via an amide bond to a glutamyl chain. The glutamyl chain has a primary amine group (NH<sub>2</sub>) and a carboxylic acid group (COOH) at the end.</p>
2955	[M+H] <sup>+</sup> <u>m/z</u> 200.1043 M 199.0970	21	<b>D-Camphorsulfonic acid</b> <i>Formula: C<sub>10</sub>H<sub>16</sub>O<sub>4</sub>S</i> <i>CAS: 3144-16-9</i>	NO	 <p>The structure shows a camphor skeleton (a bicyclic system with two methyl groups and a ketone group) with a sulfonic acid group (SO<sub>3</sub>H) attached to the bridgehead carbon.</p>

# MassBank

- <http://www.massbank.jp>
- Public repository of mass spectral data
- Data from different platforms

# MassBank: data sources

Last updated May 31, 2013 | Total Number of Spectra : 1 39,293 new

Research Groups (Contact Name)	Prefix of ID	Analysis Equipment (Analysis Method)	Number of Spectra	Number of Compounds
01. <a href="#">IAB, Keio U</a> (  Dr. Tomoyoshi Soga)	KOX	LC-ESI-QTOF (MS2)	*2 839	672
	KO	LC-ESI-QQ (MS2)	4,265	
		LC-ESI-IT (MS2,MS3,MS4)	515	
02. <a href="#">PSC, RIKEN</a> (  Dr. Masanori Arita)	PR	GC-EI-TOF (MS)	241	653
		LC-ESI-QTOF (MS,MS2)	1,371	
		LC-ESI-QQ (MS2)	87	
		CE-ESI-TOF (MS)	20	
03. <a href="#">Nihon Waters K.K.</a> (  Dr. Katsutoshi Nagase)	WA	LC-ESI-Q (MS)	2,719	575
		LC-ESI-QQ (MS2)	273	
04. <a href="#">Grad Sch Pharm Sci, Kyoto U &amp; Res Inst Prod Dev</a> (  Dr. Naoshige Akimoto &  Dr. Takashi Maoka)	CA	FAB-EBEB (MS2)	173	174
		EI-EBEB (MS)	12	
05. <a href="#">College Life Health Sci, Chubu U</a> (  Dr. Ryo Taguchi)	UT	LC-ESI-QIT (MS2)	378	478
		LC-ESI-ITFT (MS2,MS3)	2,121	
		LC-ESI-QTOF (MS)	64	



# MassBank: database services



## Database Service

→ Spectrum Search

→ Quick Search

→ Peak Search

→ Substructure Search

→ Metabolite Prediction

→ Spectral Browser

→ Batch Service

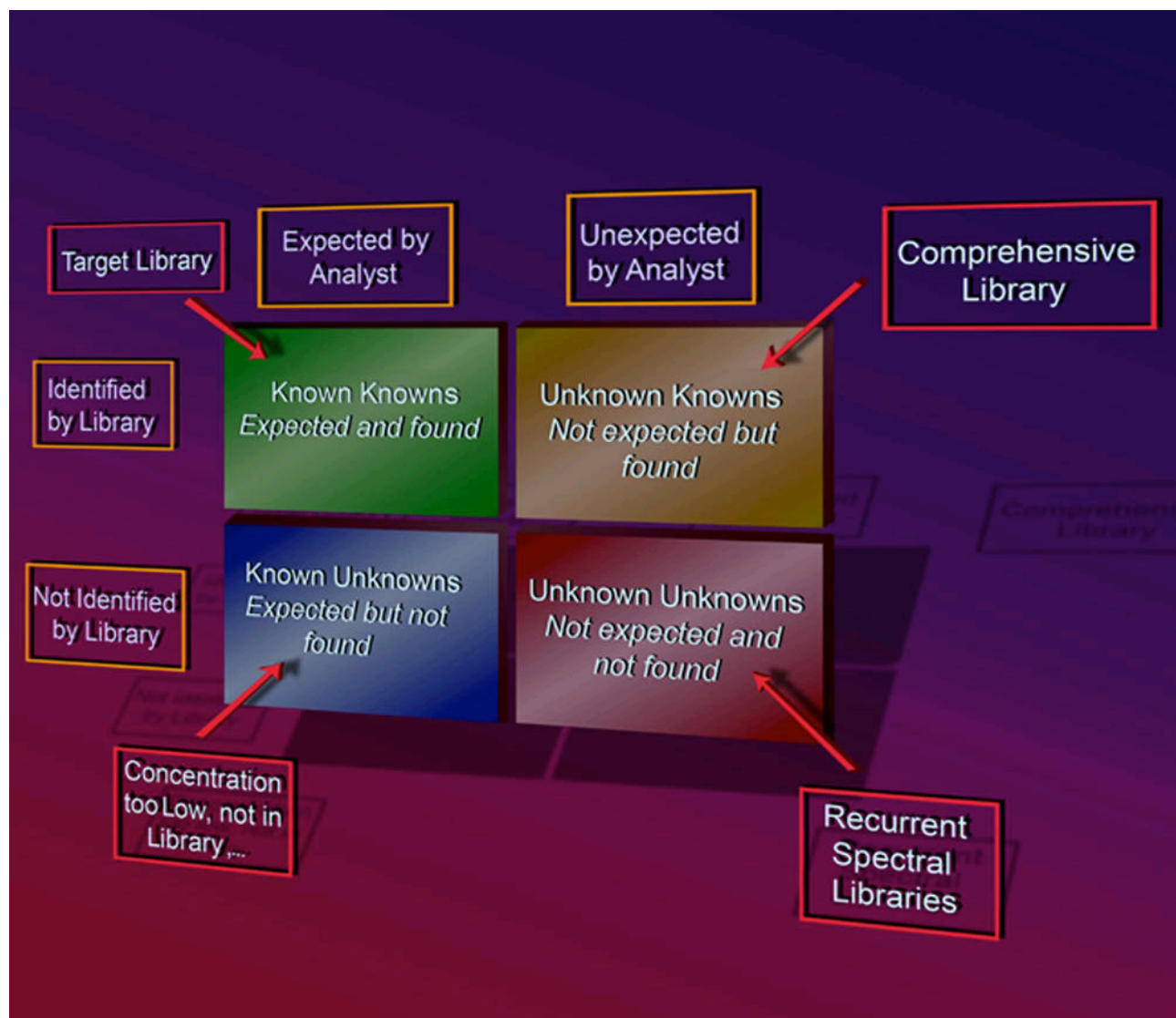
→ Browse Page

→ Record Index

# Identification, again



“...there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know.”



Stein, S., Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Analytical chemistry* **2012**, *84* (17), 7274-82.

# Acknowledgement

- Wenchao Zhang, Ph.D., previously a postdoc in the Du Lab

**Thank you!**