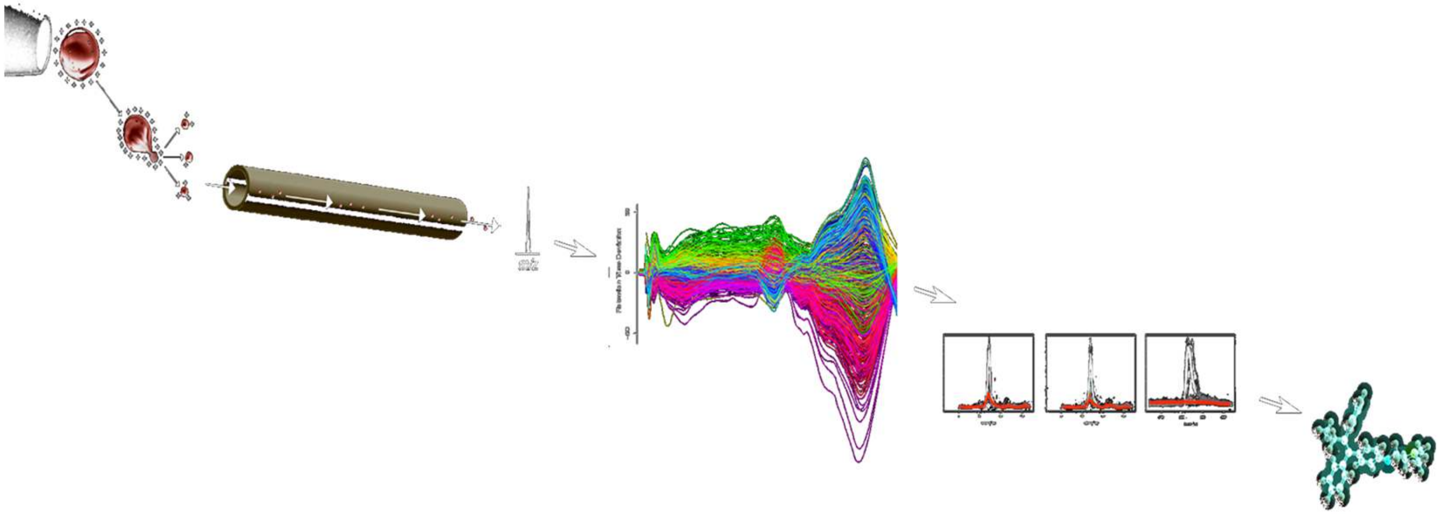
XCMS Online
&
Understanding XCMS algorithms

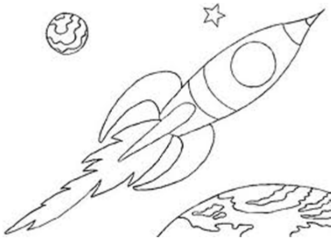


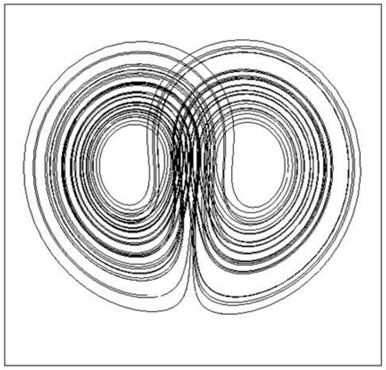
H. Paul Benton PhD

The Siuzdak Laboratory - The Scripps Research Institute

To do this morning

- Learn how to fly a rocket ship





- Get to look at chaos itself

- Shine light into the depths of nature

Ok really what are we going to do

- Learn how to fly a rocket ship
  - Computing hundreds of calculations at the speed of intel – using XCMS Online
- Get to look at chaos itself
  - Its your data not mine !
    - Data can be messy
    - Chaos can being about order metabolism is highly ordered
- Shine light into the depths of nature
  - We get to look at some of the most complex questions at the smallest biological level – metabolites.

Getting started with XCMSOnline

What did we do

- Registered on XCMS Online
  - Confirmed real email address
- Uploaded some data
  - In the old days we had to convert data ourselves – you are all very lucky!
  - XO supports – Agilent .d , Waters .RAW, Bruker .d, AB Sciex .wiff (remember the .wiff.scan files) and open source formats (mzML, mzXML, mzData, netCDF)

Processing Data

Now step by step

- We’ve loaded up two datasets – 2 classes to compare
- Set our parameters and launched a job
  - Looking at the parameters and what they mean.
  - Junk in, junk out. – Biologist
  - Good data in, bad parameter selection, junk out – bioinformticist

Overview of XCMS

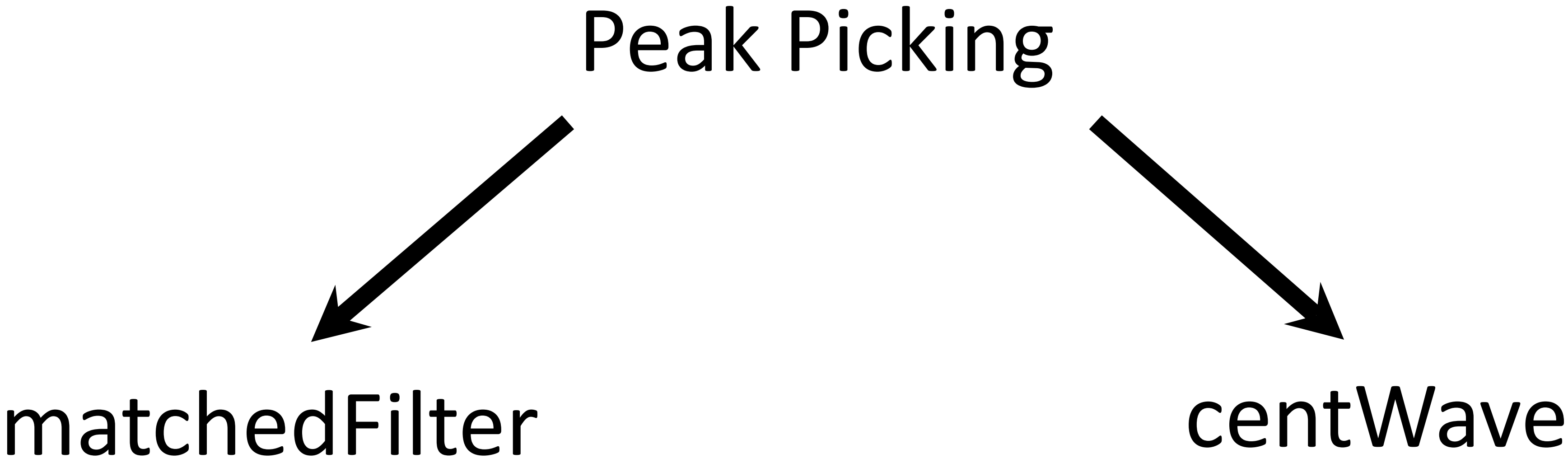
Peak Picking

Grouping similar peaks across replicates

Retention time alignment

Statistical analysis of Peaks Between classes

Peak detection choice



Peak Picking

matchedFilter

- Profile Data
- Low resolution data
- Original algorithm

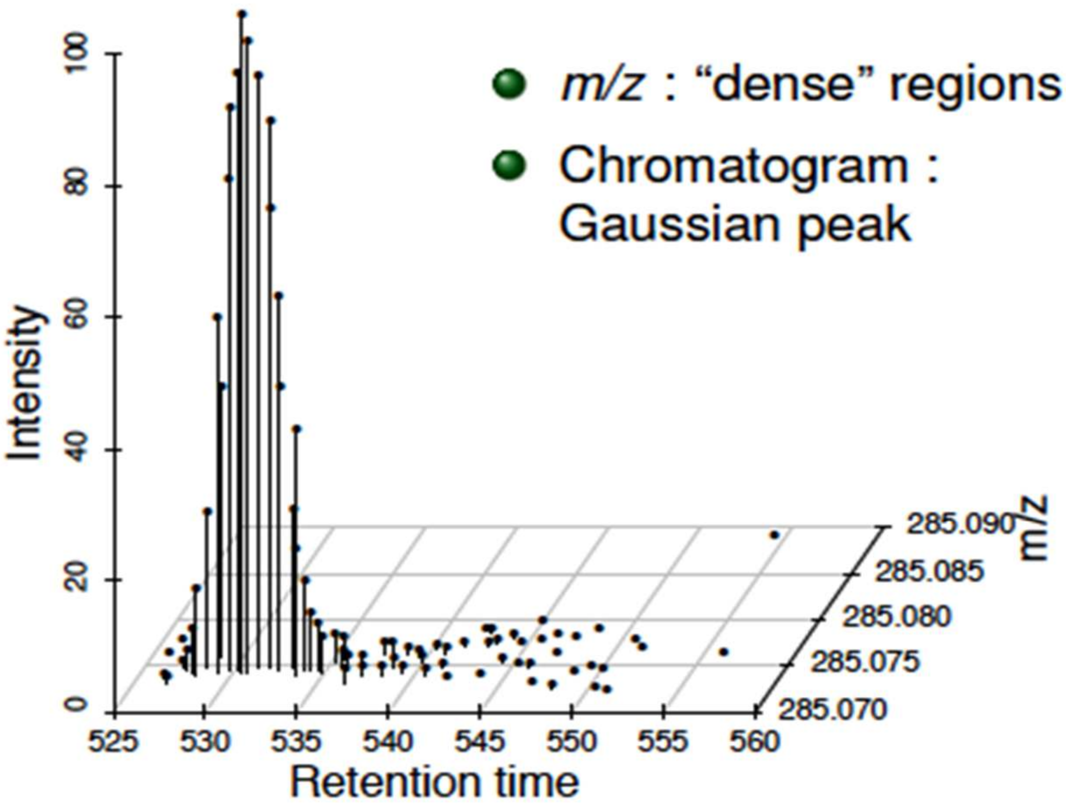
centWave

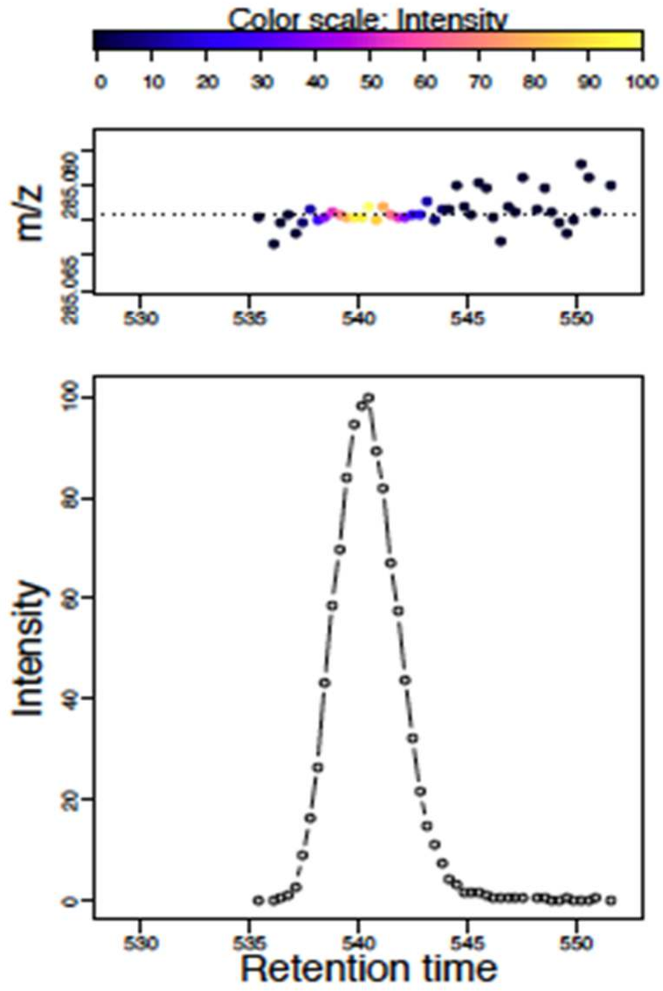
- Centroid data
- High resolution data
- New published algorithm

Rockets are like ions !!

CentWave

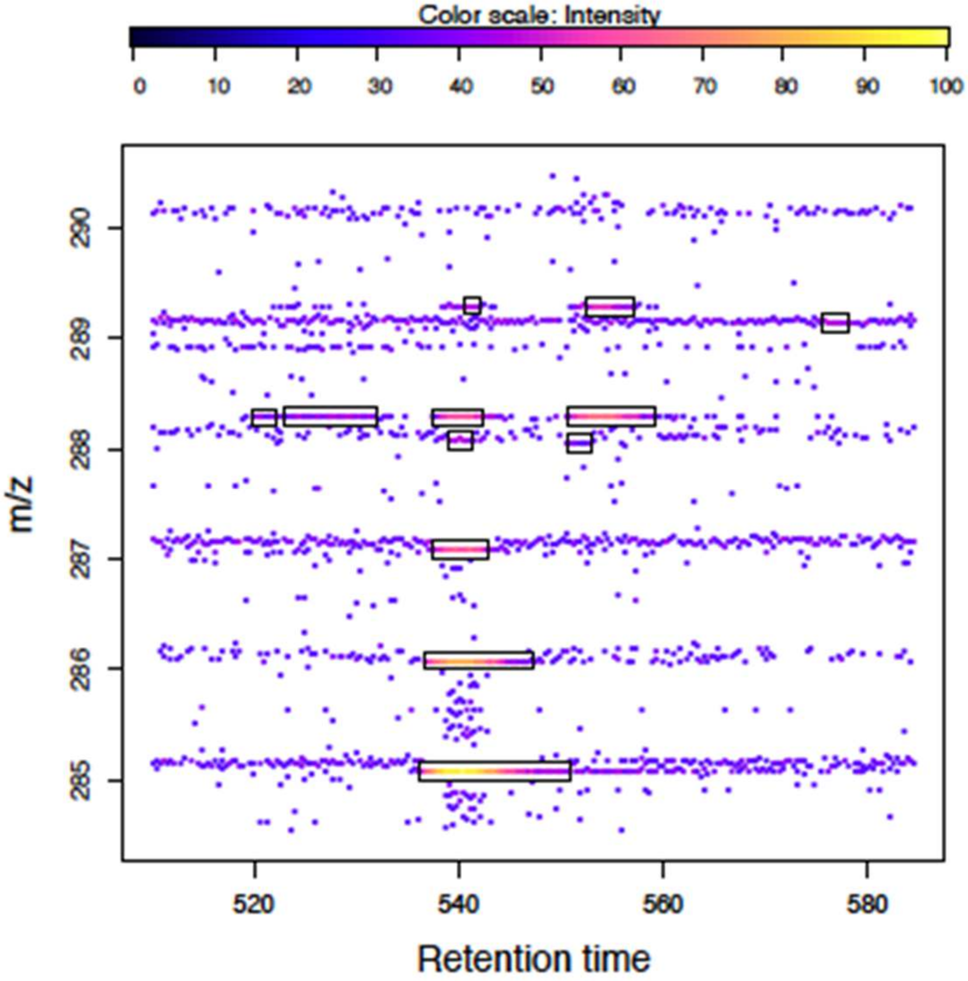
- *m/z* : “dense” regions
- Chromatogram : Gaussian peak



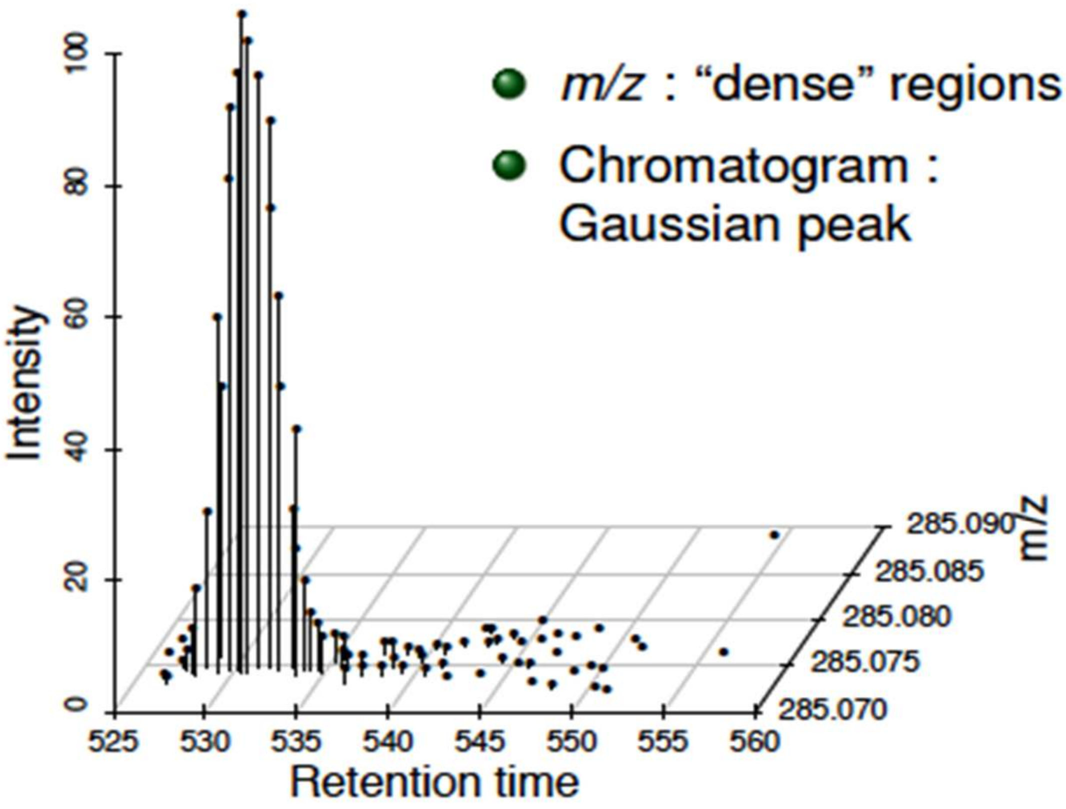


Auto/Dynamic binning

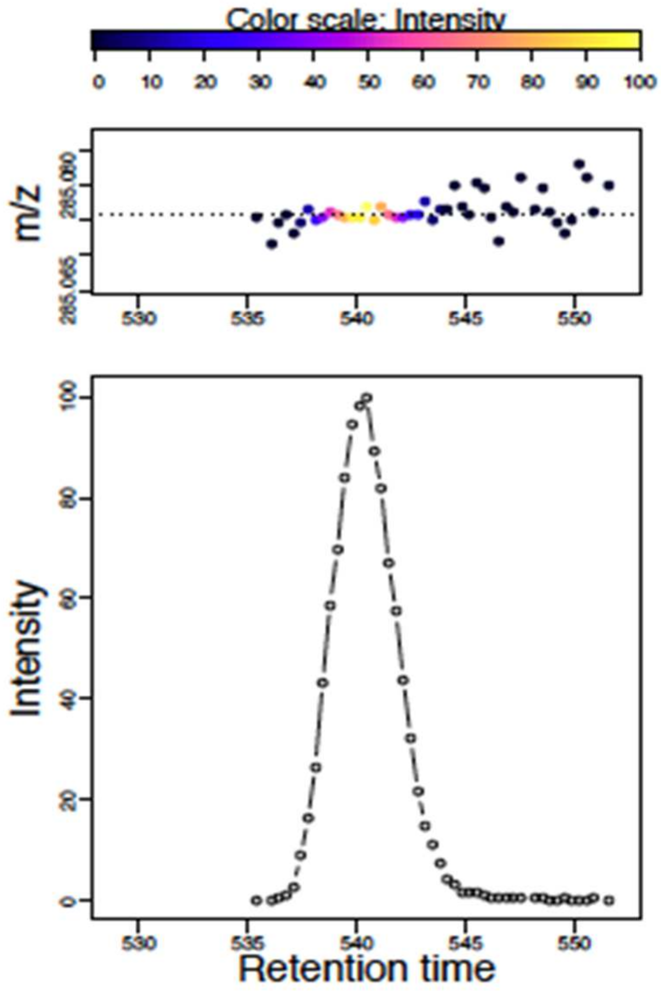
- ROI are found by making a first pass over the data to find areas that conform to expected chromatography and mass spectrometer parameters



CentWave

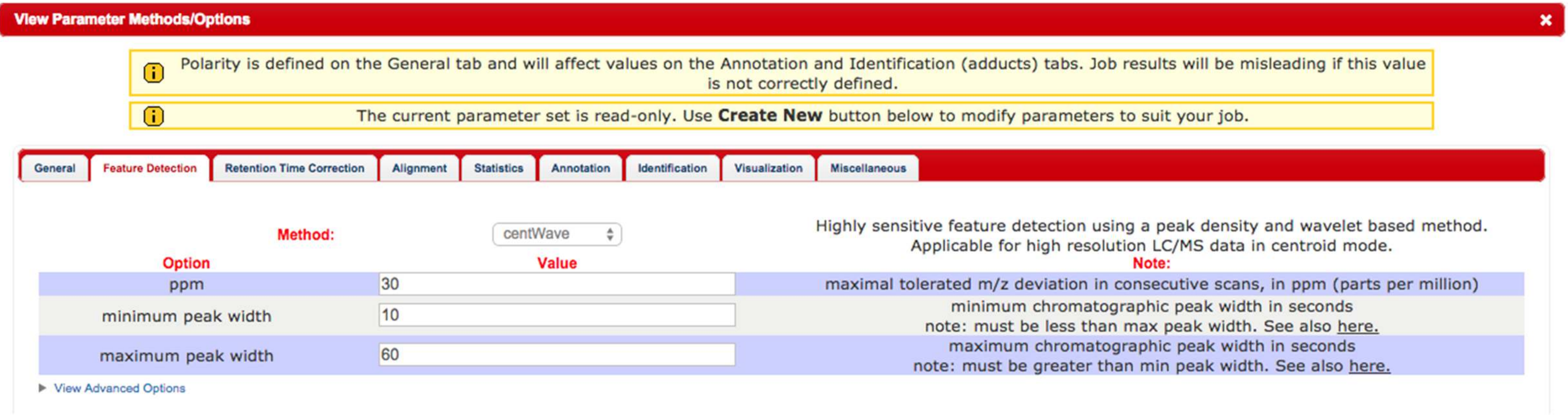


- *m/z* : “dense” regions
- Chromatogram : Gaussian peak



CentWave paramaters

- Peakwidth = How wide is your peak – from a minimum to a maximum in seconds
- Ppm = how much does the peak vary across scans



Polarity is defined on the General tab and will affect values on the Annotation and Identification (adducts) tabs. Job results will be misleading if this value is not correctly defined.

The current parameter set is read-only. Use **Create New** button below to modify parameters to suit your job.

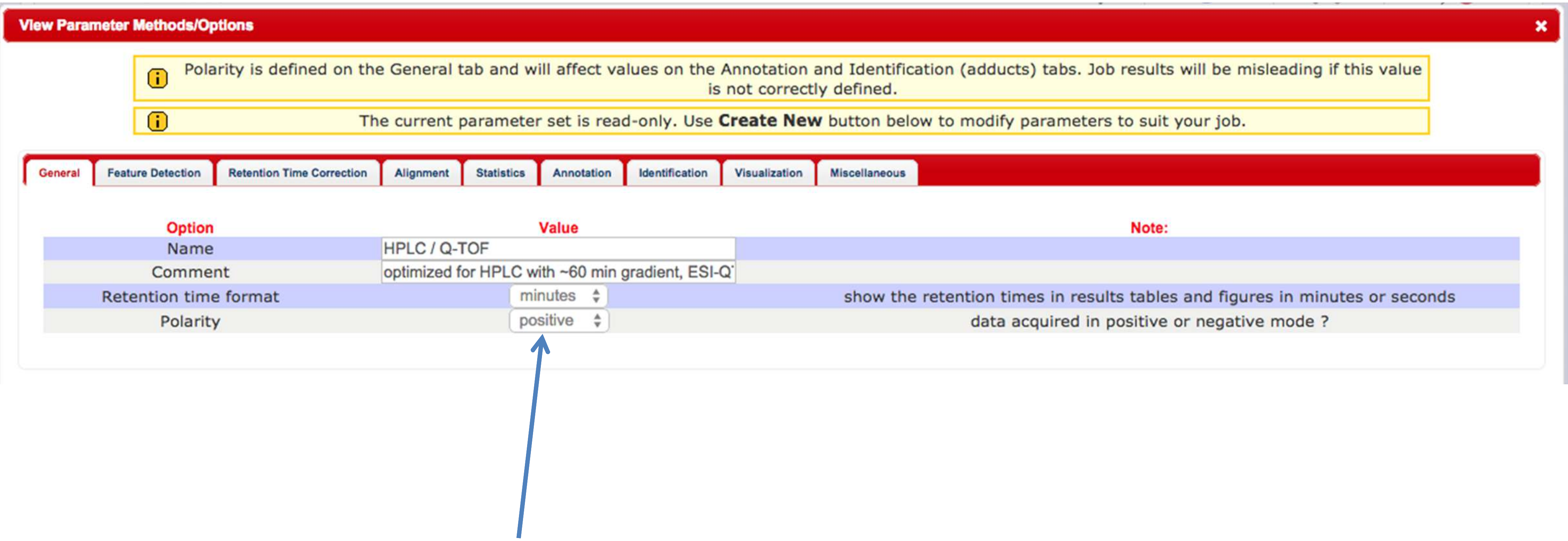
Method: centWave

Highly sensitive feature detection using a peak density and wavelet based method. Applicable for high resolution LC/MS data in centroid mode.

Option	Value	Note:
ppm	30	maximal tolerated m/z deviation in consecutive scans, in ppm (parts per million)
minimum peak width	10	minimum chromatographic peak width in seconds note: must be less than max peak width. See also here.
maximum peak width	60	maximum chromatographic peak width in seconds note: must be greater than min peak width. See also here.

One thing to note

- Choose your polarity correctly!!



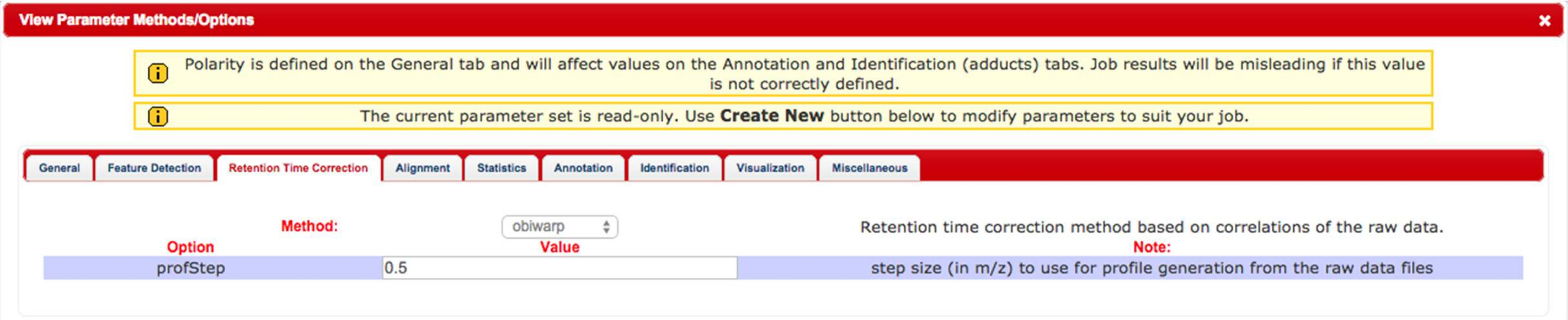
View Parameter Methods/Options

Polarity is defined on the General tab and will affect values on the Annotation and Identification (adducts) tabs. Job results will be misleading if this value is not correctly defined.

The current parameter set is read-only. Use **Create New** button below to modify parameters to suit your job.

Option	Value	Note:
Name	HPLC / Q-TOF	
Comment	optimized for HPLC with ~60 min gradient, ESI-Q	
Retention time format	minutes	show the retention times in results tables and figures in minutes or seconds
Polarity	positive	data acquired in positive or negative mode ?

Retention time alignment



Polarity is defined on the General tab and will affect values on the Annotation and Identification (adducts) tabs. Job results will be misleading if this value is not correctly defined.

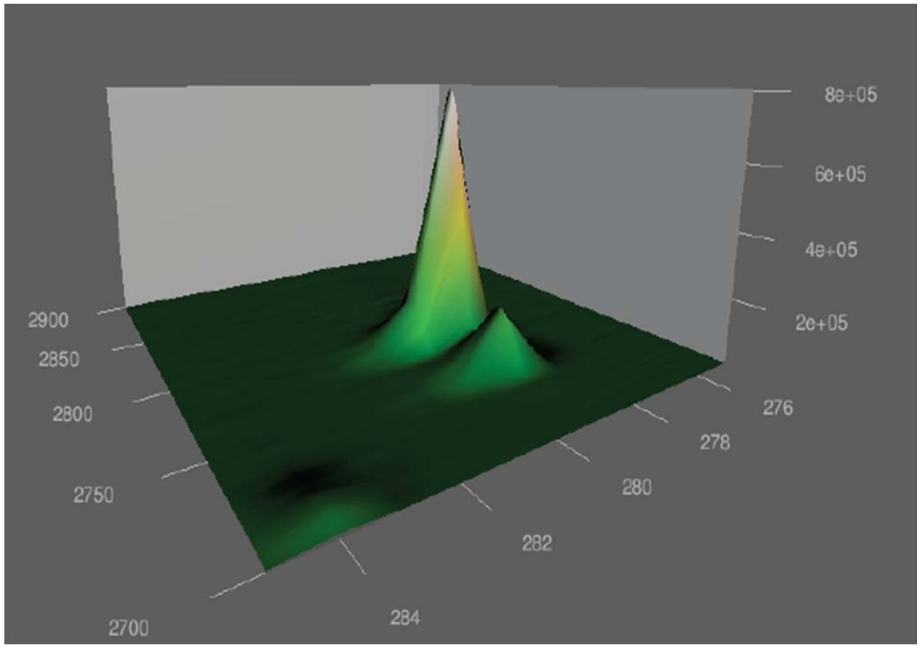
The current parameter set is read-only. Use **Create New** button below to modify parameters to suit your job.

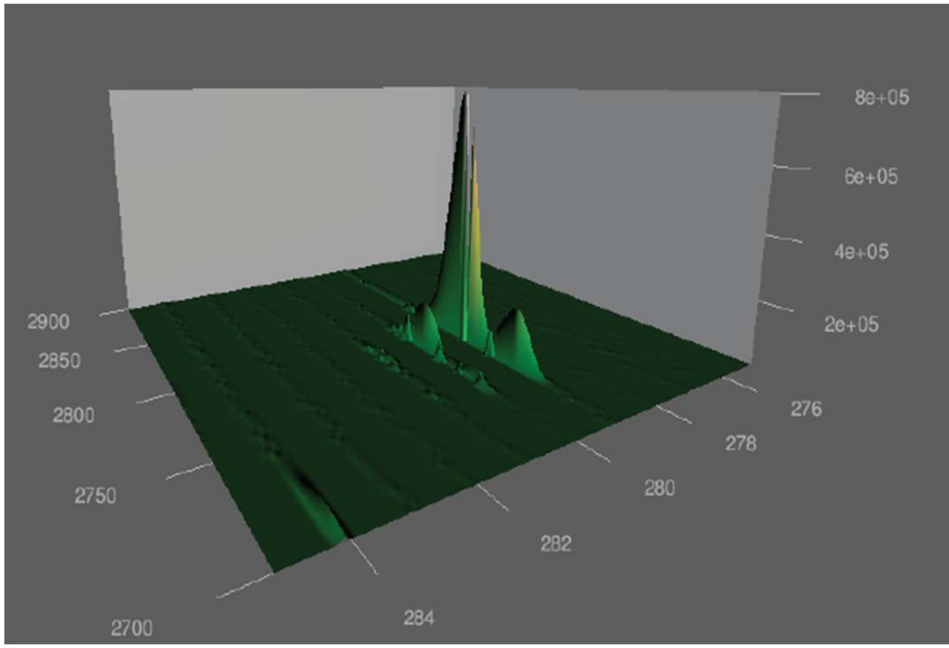
Method: obiwarp — Retention time correction method based on correlations of the raw data.

Option	Value	Note:
profStep	0.5	step size (in m/z) to use for profile generation from the raw data files

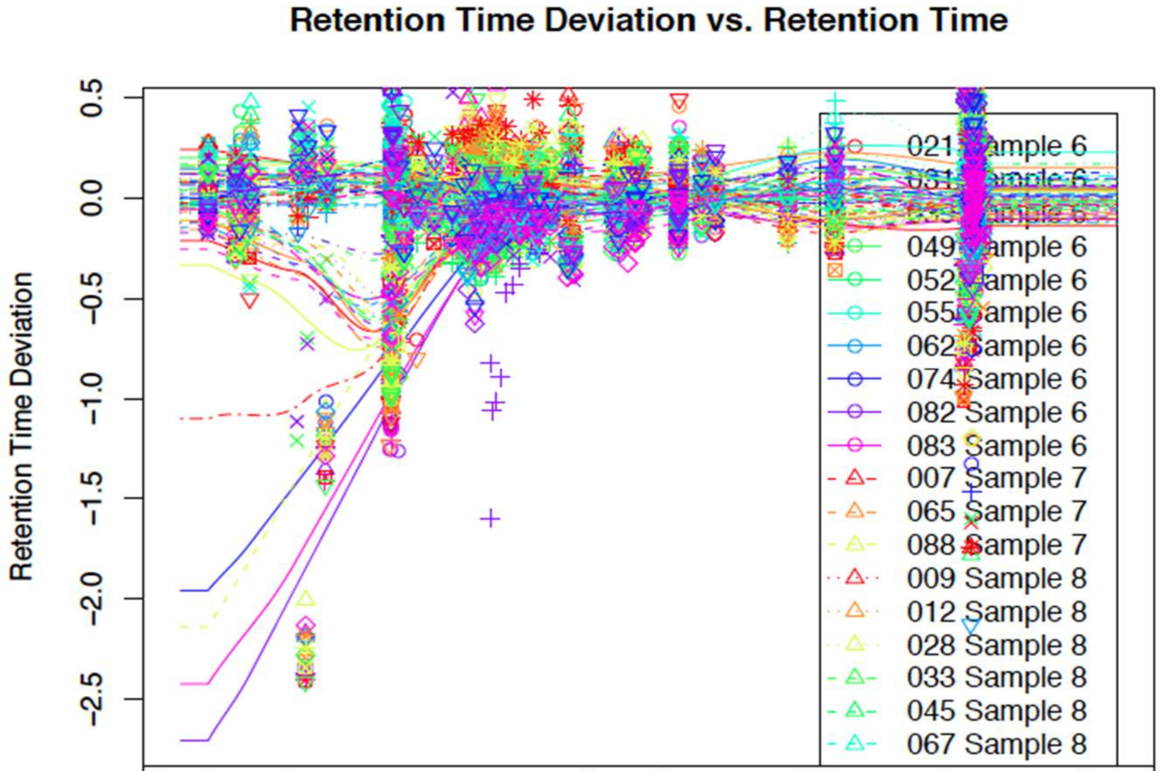
Obiwarp –

A Digital signal processing algorithm. Very good for high drift alignment. Fits data as if each LC-MS 3D landscape was play dough to squeeze these together. Technically this is warping not aligning



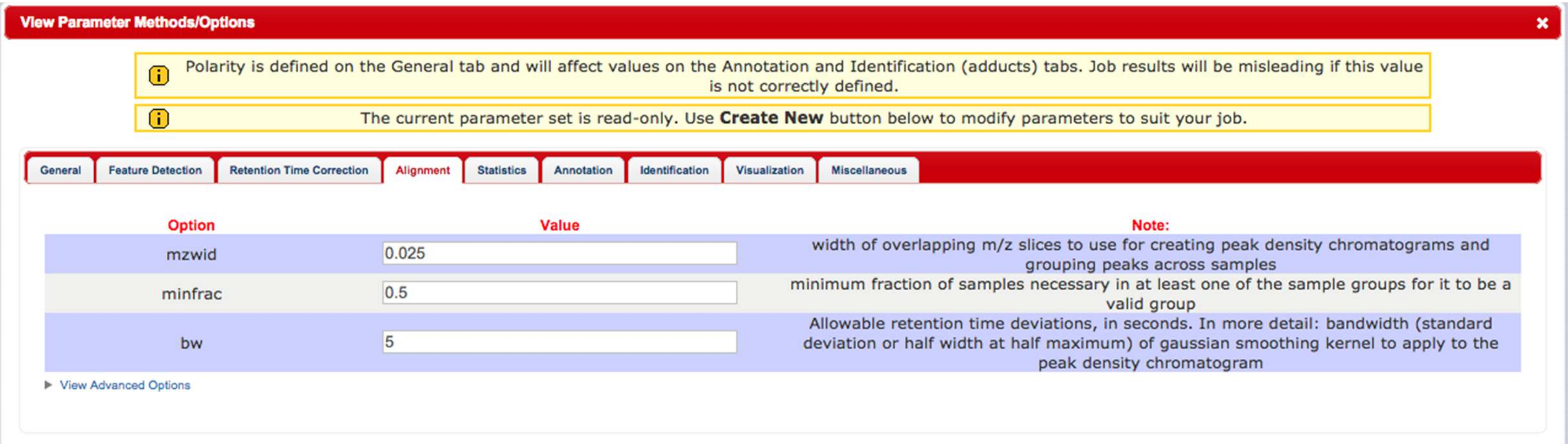


Retention time alignment



- Loess – this is a model to fit the data to using the residuals to correct/align the samples
  - Relies on anchors distributed across the RT

Grouping

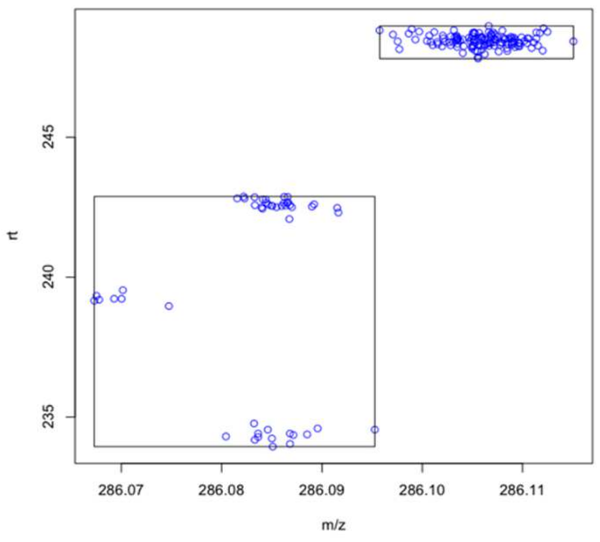


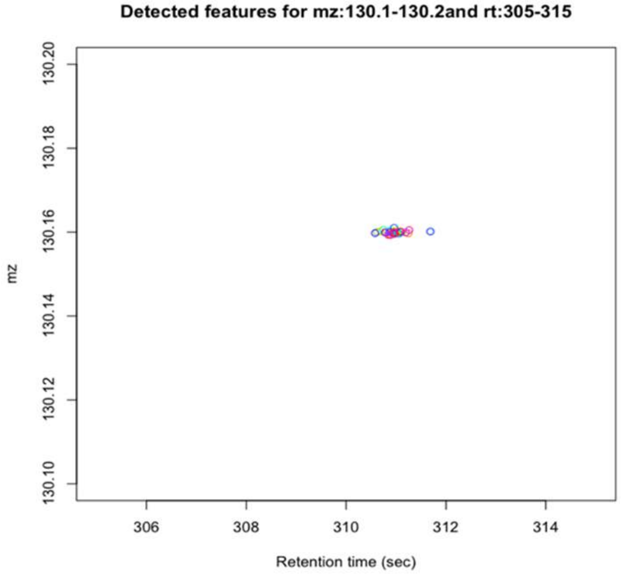
View Parameter Methods/Options

Polarity is defined on the General tab and will affect values on the Annotation and Identification (adducts) tabs. Job results will be misleading if this value is not correctly defined.

The current parameter set is read-only. Use **Create New** button below to modify parameters to suit your job.

Option	Value	Note:
mzwid	0.025	width of overlapping m/z slices to use for creating peak density chromatograms and grouping peaks across samples
minfrac	0.5	minimum fraction of samples necessary in at least one of the sample groups for it to be a valid group
bw	5	Allowable retention time deviations, in seconds. In more detail: bandwidth (standard deviation or half width at half maximum) of gaussian smoothing kernel to apply to the peak density chromatogram



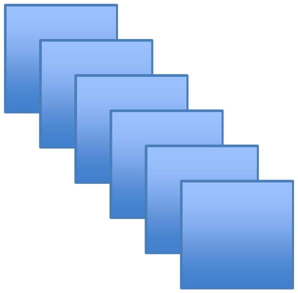


Detected features for mz:130.1-130.2and rt:305-315

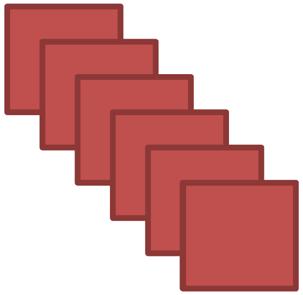
MinFrac !

- More questions on minfrac than any other!

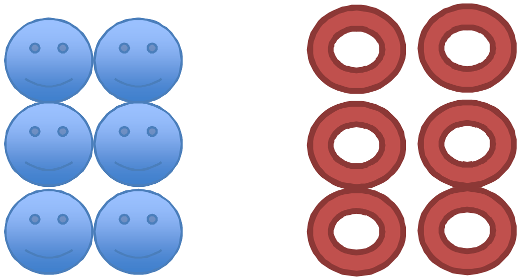
KO – 6 samples



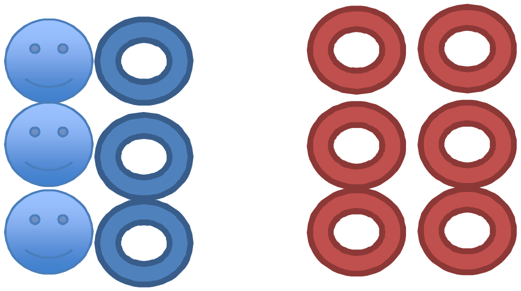
WT – 6 samples



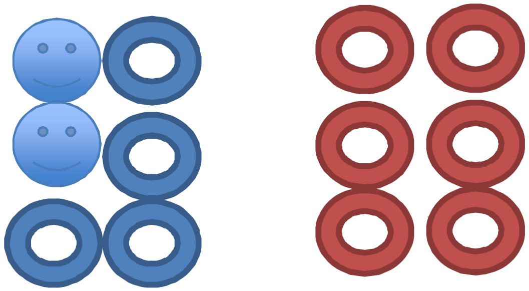
minFrac = 0.5 = 50%



Group become a
valid feature
Perfect biomarker

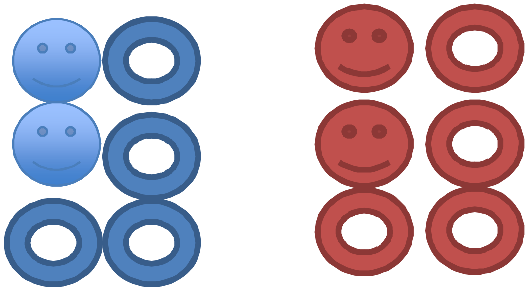


Group become a
valid feature
Just hits 50% - OK

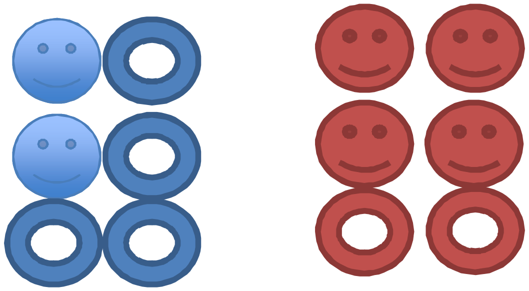


Group is **not** a
valid feature

minFrac test

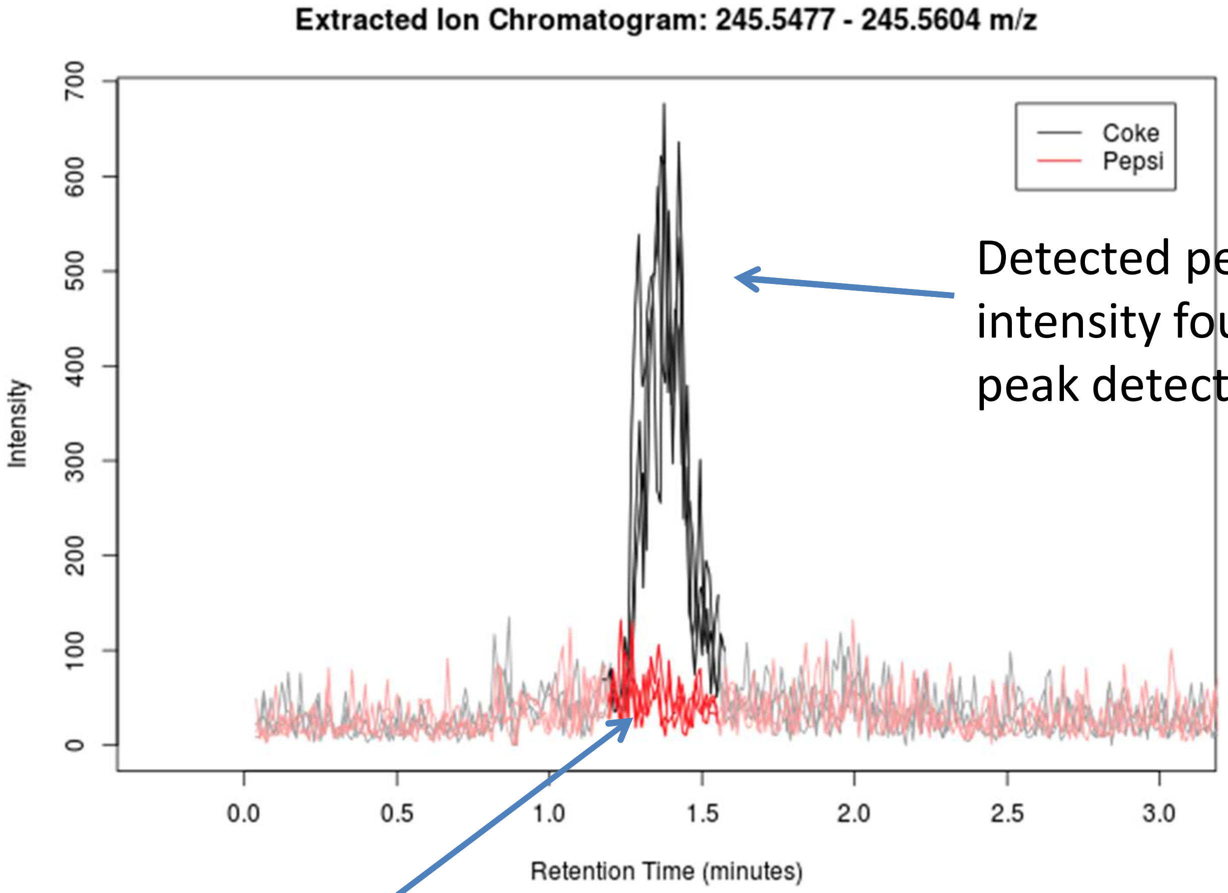


Not a valid feature



A valid feature

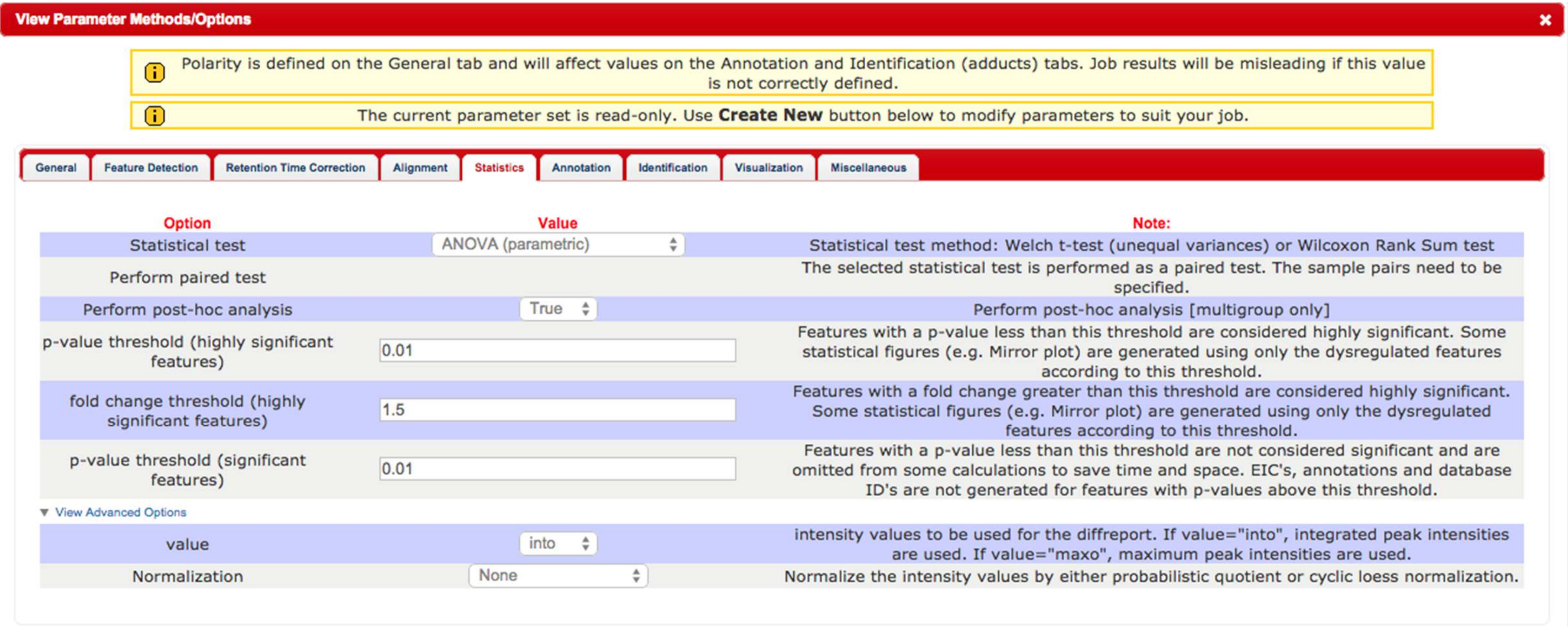
Peak Filling



Detected peak – peak intensity found by peak detector

Peak not detected – intensity filled by fillPeaks

Statistics !! Yea !!



Polarity is defined on the General tab and will affect values on the Annotation and Identification (adducts) tabs. Job results will be misleading if this value is not correctly defined.

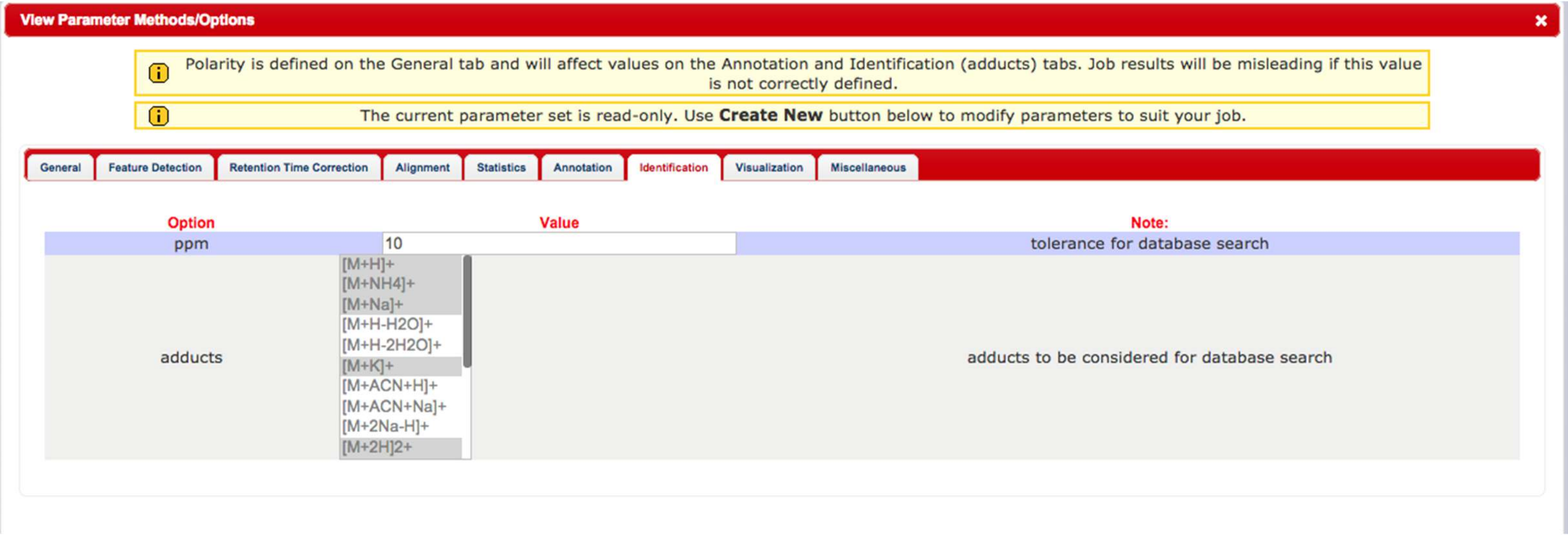
The current parameter set is read-only. Use **Create New** button below to modify parameters to suit your job.

Option	Value	Note:
Statistical test	ANOVA (parametric)	Statistical test method: Welch t-test (unequal variances) or Wilcoxon Rank Sum test
Perform paired test		The selected statistical test is performed as a paired test. The sample pairs need to be specified.
Perform post-hoc analysis	True	Perform post-hoc analysis [multigroup only]
p-value threshold (highly significant features)	0.01	Features with a p-value less than this threshold are considered highly significant. Some statistical figures (e.g. Mirror plot) are generated using only the dysregulated features according to this threshold.
fold change threshold (highly significant features)	1.5	Features with a fold change greater than this threshold are considered highly significant. Some statistical figures (e.g. Mirror plot) are generated using only the dysregulated features according to this threshold.
p-value threshold (significant features)	0.01	Features with a p-value less than this threshold are not considered significant and are omitted from some calculations to save time and space. EIC's, annotations and database ID's are not generated for features with p-values above this threshold.

▼ View Advanced Options

value	into	intensity values to be used for the diffreport. If value="into", integrated peak intensities are used. If value="maxo", maximum peak intensities are used.
Normalization	None	Normalize the intensity values by either probabilistic quotient or cyclic loess normalization.

Adduct selection

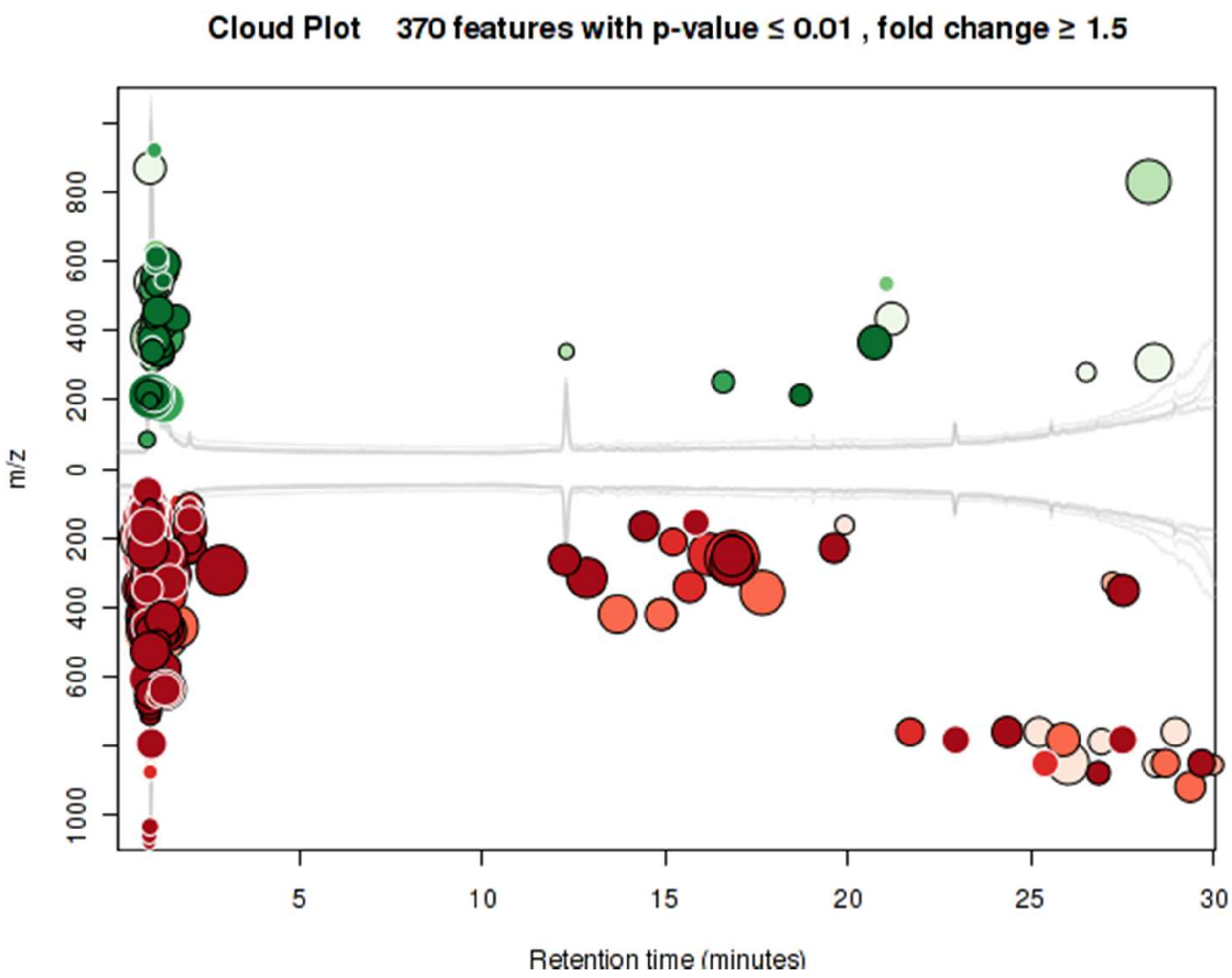


Polarity is defined on the General tab and will affect values on the Annotation and Identification (adducts) tabs. Job results will be misleading if this value is not correctly defined.

The current parameter set is read-only. Use **Create New** button below to modify parameters to suit your job.

Option	Value	Note:
ppm	10	tolerance for database search
adducts	[✓ [M+H]+] [✓ [M+NH4]+] [✓ [M+Na]+] [M+H-H2O]+ [M+H-2H2O]+ [✓ [M+K]+] [M+ACN+H]+ [M+ACN+Na]+ [M+2Na-H]+ [✓ [M+2H]2+]	adducts to be considered for database search

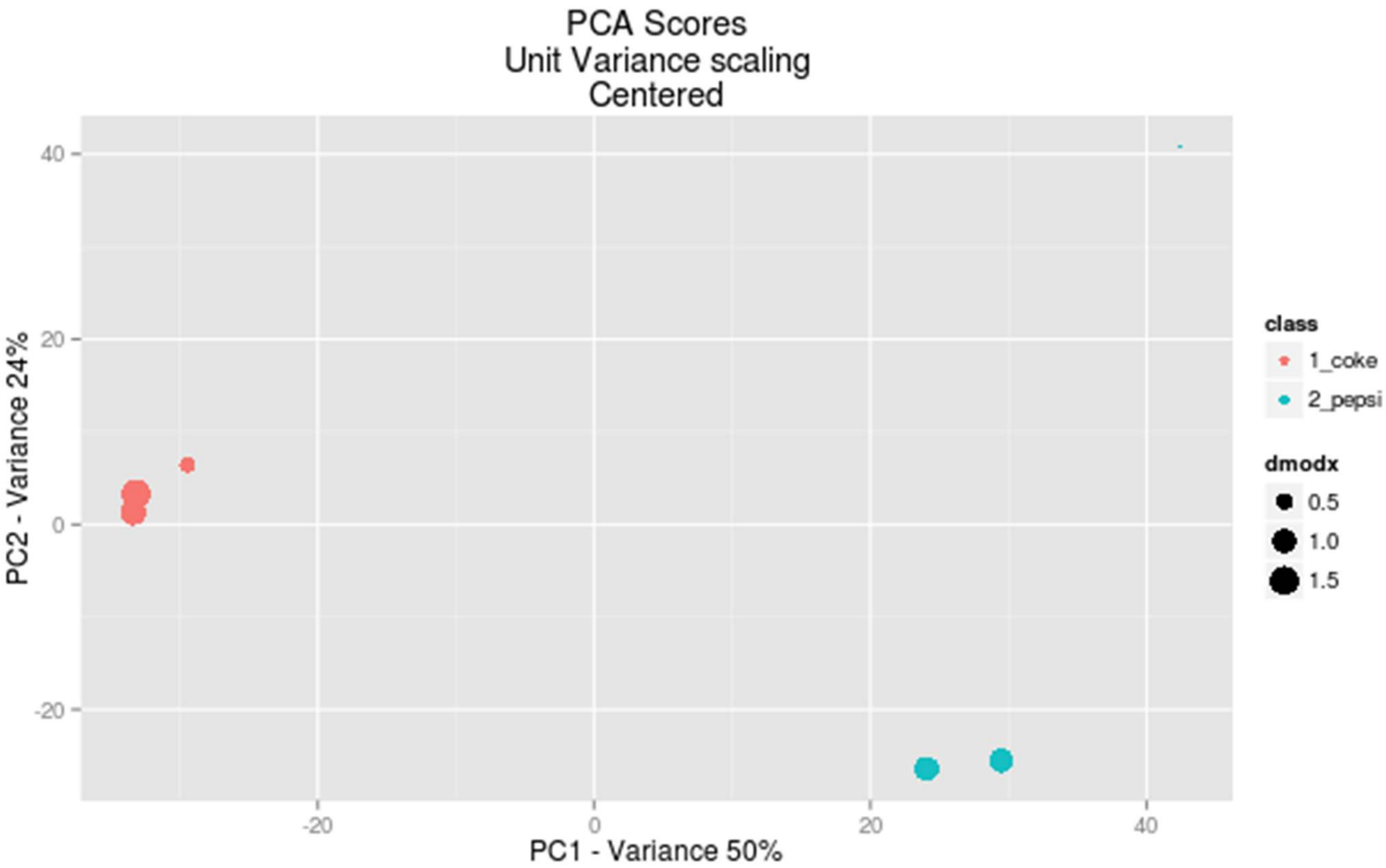
Cloud plot

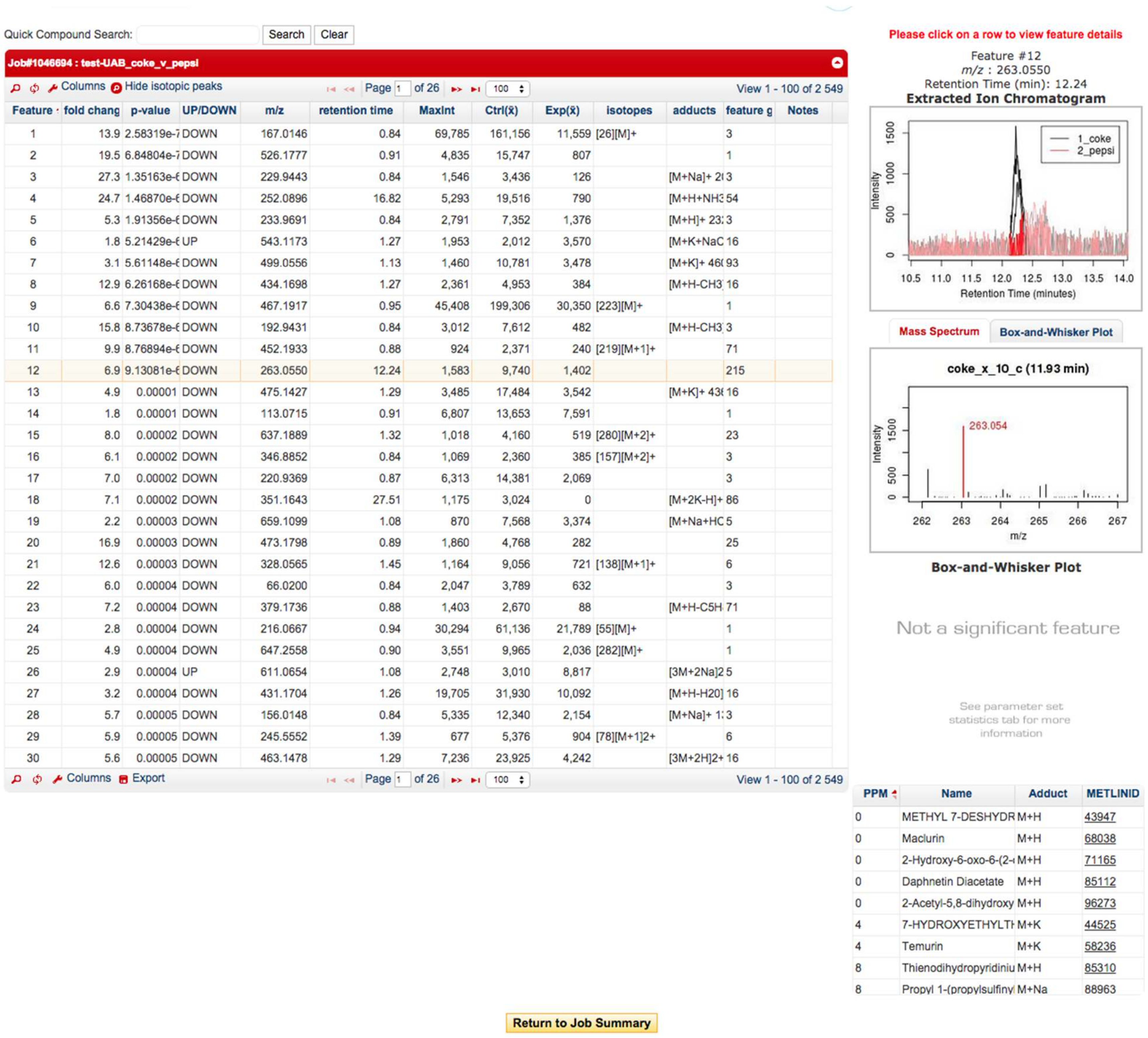


Size = fold change
Colour = signficance (lower p-value)

Black or white ring = metlin hits

Static PCA


PCA Scores
Unit Variance scaling
Centered
PC2 - Variance 24%
40
20
0
-20
-20
0
20
40
PC1 - Variance 50%
class
1_coke
2_pepsi
dmodx
0.5
1.0
1.5



Job#1046694 : test-UAB_coke_v_pepsi

Columns Hide isotopic peaks Page 1 of 26 100 View 1 - 100 of 2 549

Feature ·	fold chang	p-value	UP/DOWN	m/z	retention time	MaxInt	Ctrl(x̄)	Exp(x̄)	isotopes	adducts	feature g	Notes
1	13.9	2.58319e-7	DOWN	167.0146	0.84	69,785	161,156	11,559	[26][M]+		3	
2	19.5	6.84804e-7	DOWN	526.1777	0.91	4,835	15,747	807			1	
3	27.3	1.35163e-6	DOWN	229.9443	0.84	1,546	3,436	126		[M+Na]+ 2(	3	
4	24.7	1.46870e-6	DOWN	252.0896	16.82	5,293	19,516	790		[M+H+NH3	54	
5	5.3	1.91356e-6	DOWN	233.9691	0.84	2,791	7,352	1,376		[M+H]+ 23:	3	
6	1.8	5.21429e-6	UP	543.1173	1.27	1,953	2,012	3,570		[M+K+NaC	16	
7	3.1	5.61148e-6	DOWN	499.0556	1.13	1,460	10,781	3,478		[M+K]+ 46(	93	
8	12.9	6.26168e-6	DOWN	434.1698	1.27	2,361	4,953	384		[M+H-CH3	16	
9	6.6	7.30438e-6	DOWN	467.1917	0.95	45,408	199,306	30,350	[223][M]+		1	
10	15.8	8.73678e-6	DOWN	192.9431	0.84	3,012	7,612	482		[M+H-CH3	3	
11	9.9	8.76894e-6	DOWN	452.1933	0.88	924	2,371	240	[219][M+1]+		71	
12	6.9	9.13081e-6	DOWN	263.0550	12.24	1,583	9,740	1,402			215	
13	4.9	0.00001	DOWN	475.1427	1.29	3,485	17,484	3,542		[M+K]+ 43(	16	
14	1.8	0.00001	DOWN	113.0715	0.91	6,807	13,653	7,591			1	
15	8.0	0.00002	DOWN	637.1889	1.32	1,018	4,160	519	[280][M+2]+		23	
16	6.1	0.00002	DOWN	346.8852	0.84	1,069	2,360	385	[157][M+2]+		3	
17	7.0	0.00002	DOWN	220.9369	0.87	6,313	14,381	2,069			3	
18	7.1	0.00002	DOWN	351.1643	27.51	1,175	3,024	0		[M+2K-H]+	86	
19	2.2	0.00003	DOWN	659.1099	1.08	870	7,568	3,374		[M+Na+HC	5	
20	16.9	0.00003	DOWN	473.1798	0.89	1,860	4,768	282			25	
21	12.6	0.00003	DOWN	328.0565	1.45	1,164	9,056	721	[138][M+1]+		6	
22	6.0	0.00004	DOWN	66.0200	0.84	2,047	3,789	632			3	
23	7.2	0.00004	DOWN	379.1736	0.88	1,403	2,670	88		[M+H-C5H	71	
24	2.8	0.00004	DOWN	216.0667	0.94	30,294	61,136	21,789	[55][M]+		1	
25	4.9	0.00004	DOWN	647.2558	0.90	3,551	9,965	2,036	[282][M]+		1	
26	2.9	0.00004	UP	611.0654	1.08	2,748	3,010	8,817		[3M+2Na]2	5	
27	3.2	0.00004	DOWN	431.1704	1.26	19,705	31,930	10,092		[M+H-H20]	16	
28	5.7	0.00005	DOWN	156.0148	0.84	5,335	12,340	2,154		[M+Na]+ 1:	3	
29	5.9	0.00005	DOWN	245.5552	1.39	677	5,376	904	[78][M+1]2+		6	
30	5.6	0.00005	DOWN	463.1478	1.29	7,236	23,925	4,242		[3M+2H]2+	16	

Columns Export Page 1 of 26 100 View 1 - 100 of 2 549

Feature #12
m/z : 263.0550
Retention Time (min): 12.24

Box-and-Whisker Plot

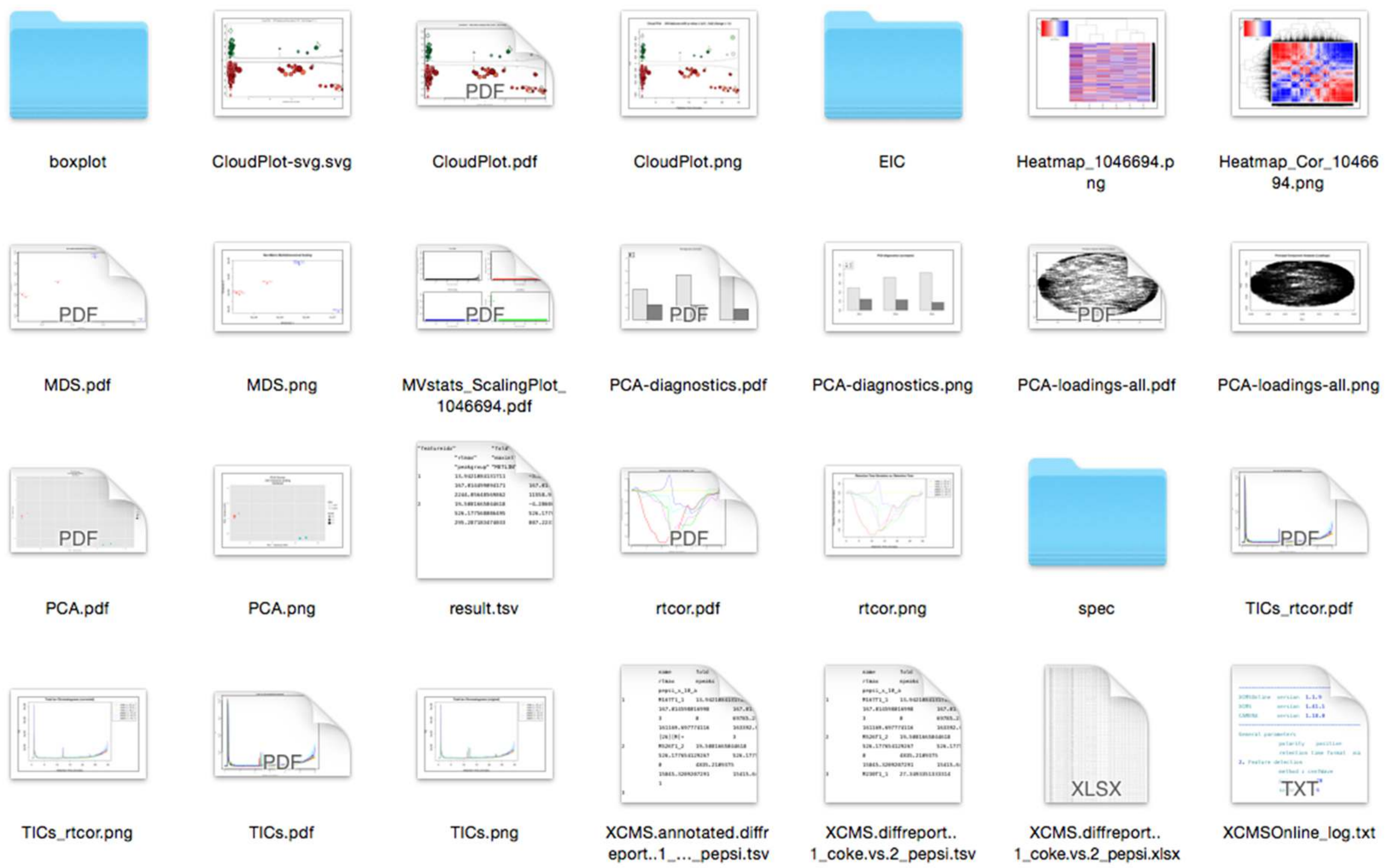
Not a significant feature

See parameter set statistics tab for more information

PPM	Name	Adduct	METLINID
0	METHYL 7-DESHYDR	M+H	43947
0	Maclurin	M+H	68038
0	2-Hydroxy-6-oxo-6-(2-	M+H	71165
0	Daphnetin Diacetate	M+H	85112
0	2-Acetyl-5,8-dihydroxy	M+H	96273
4	7-HYDROXYETHYLTH	M+K	44525
4	Temurin	M+K	58236
8	Thienodihydropyridiniu	M+H	85310
8	Propyl 1-(propylsulfinyl	M+Na	88963

Results.zip download file

- This has all of the plots and information from the processed job.
  - Static PCA
  - Static heat map
  - Static cloud plots
  - Scaling plot – Good for looking at scaling for PCA (trend implicates heteroscedastic noise)



Contents of results.zip file
XCMS.diffreport. And XCMS.annotated.diffreport are the data tables with all the intensity values associated with them not results.tsv

Thank you

Prof. Gary Siuzdak

Questions?

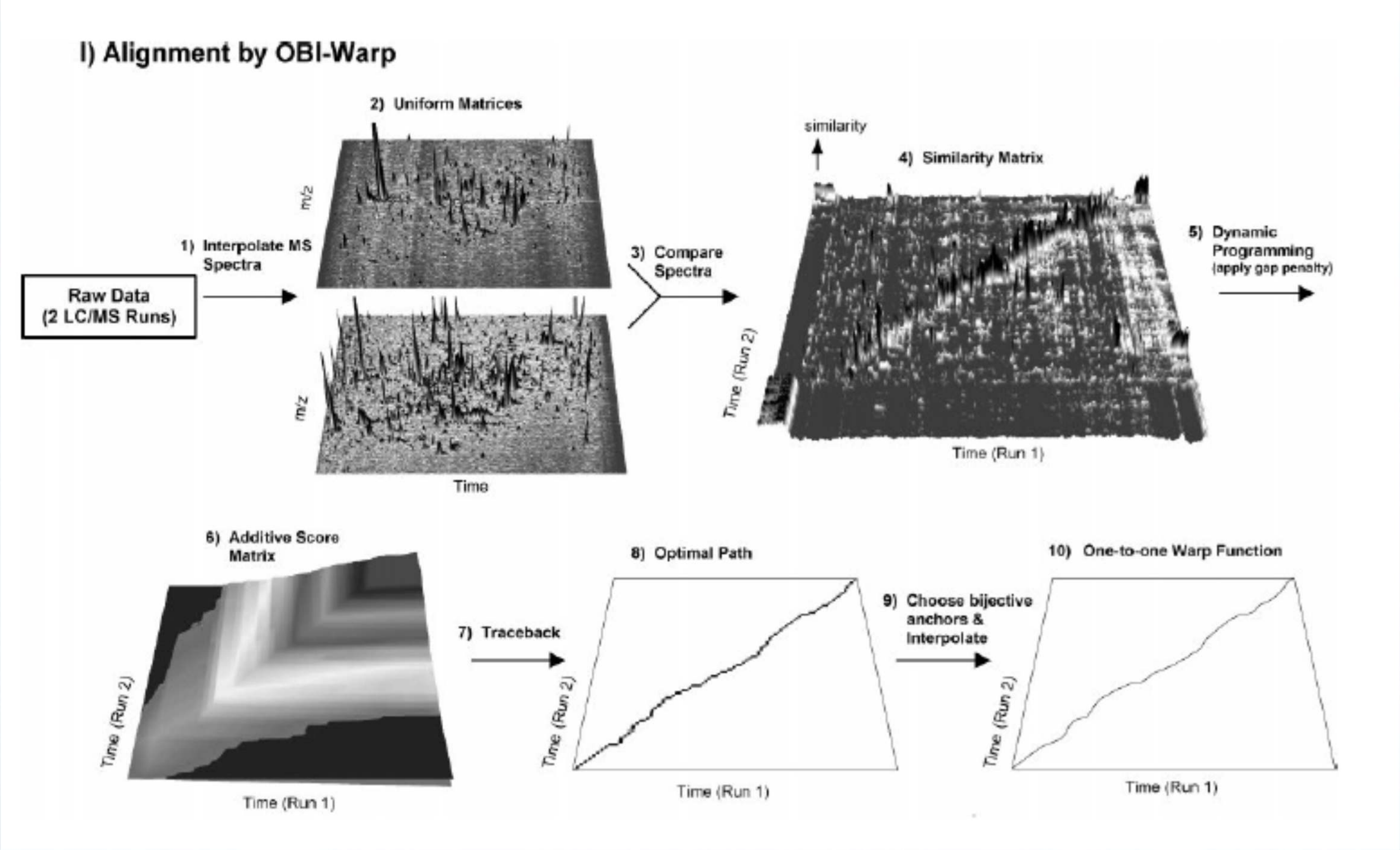
Duane Rinehart

Comments?

Dr. Bill Webb

Thoughts?

OBI-WARP METHOD



I) Alignment by OBI-Warp