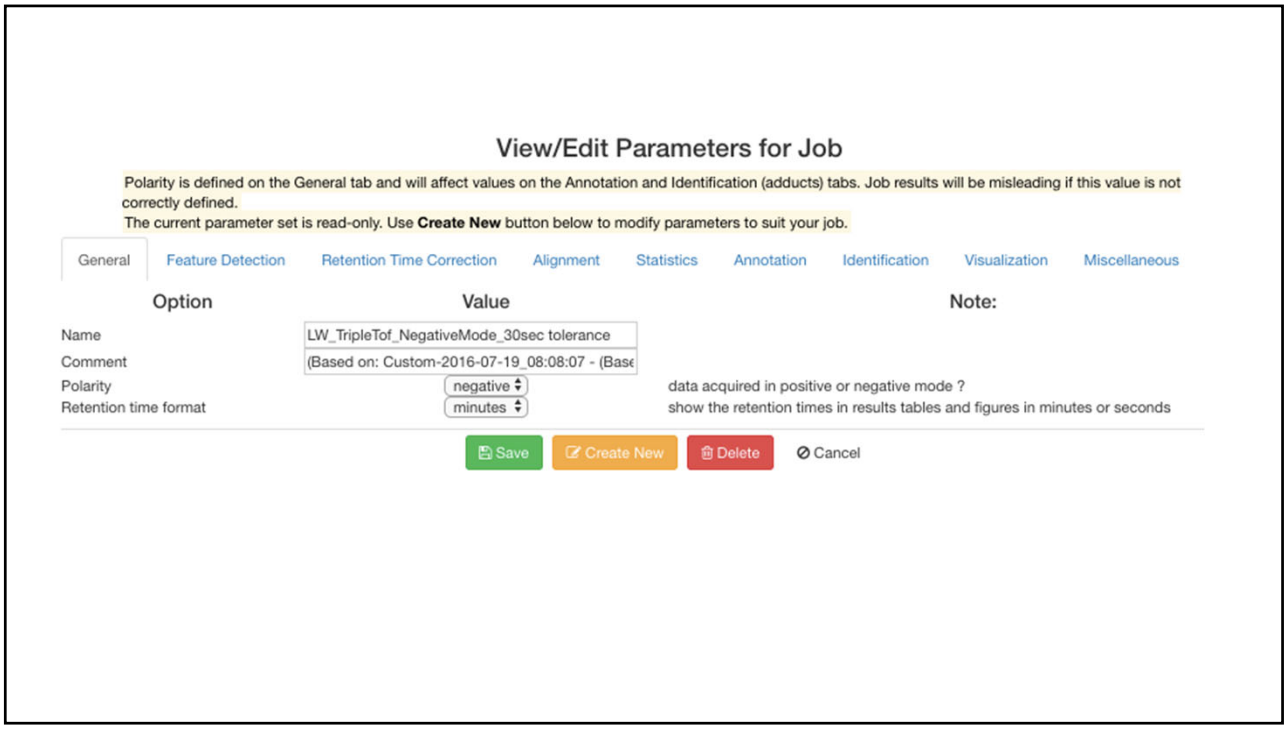
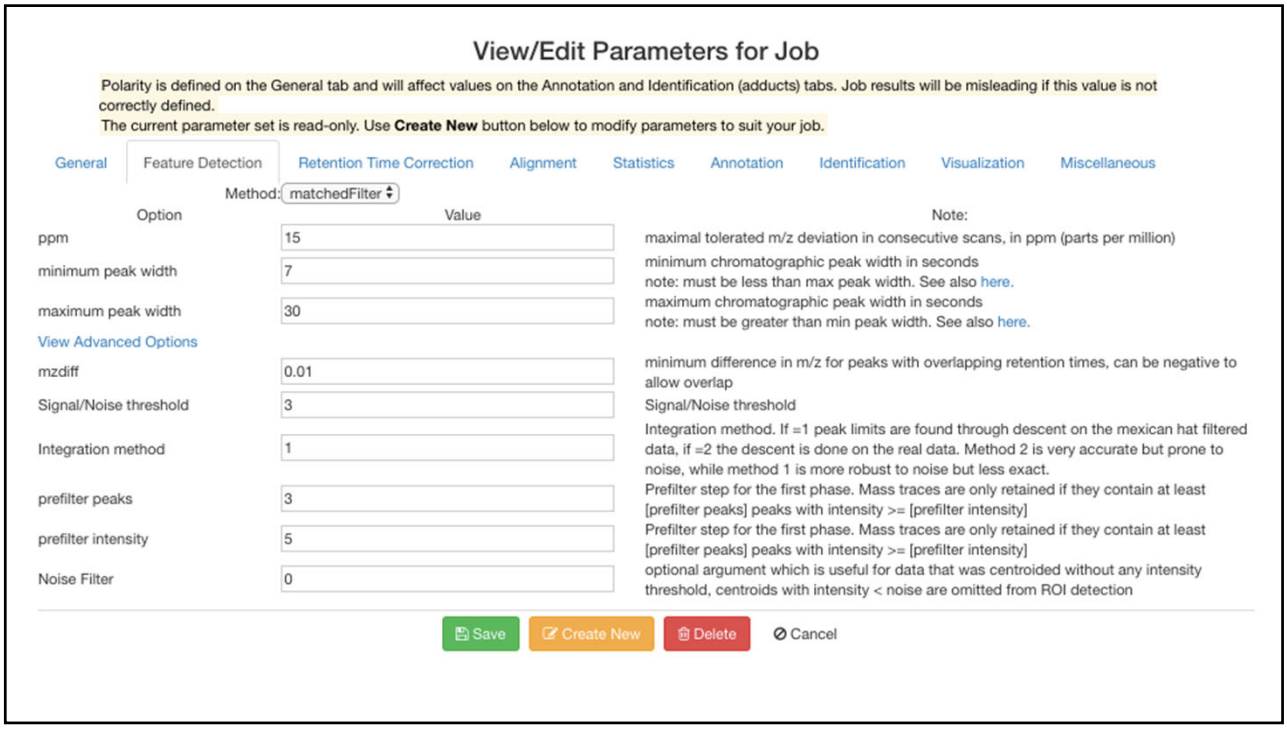
Examining the XCMS output

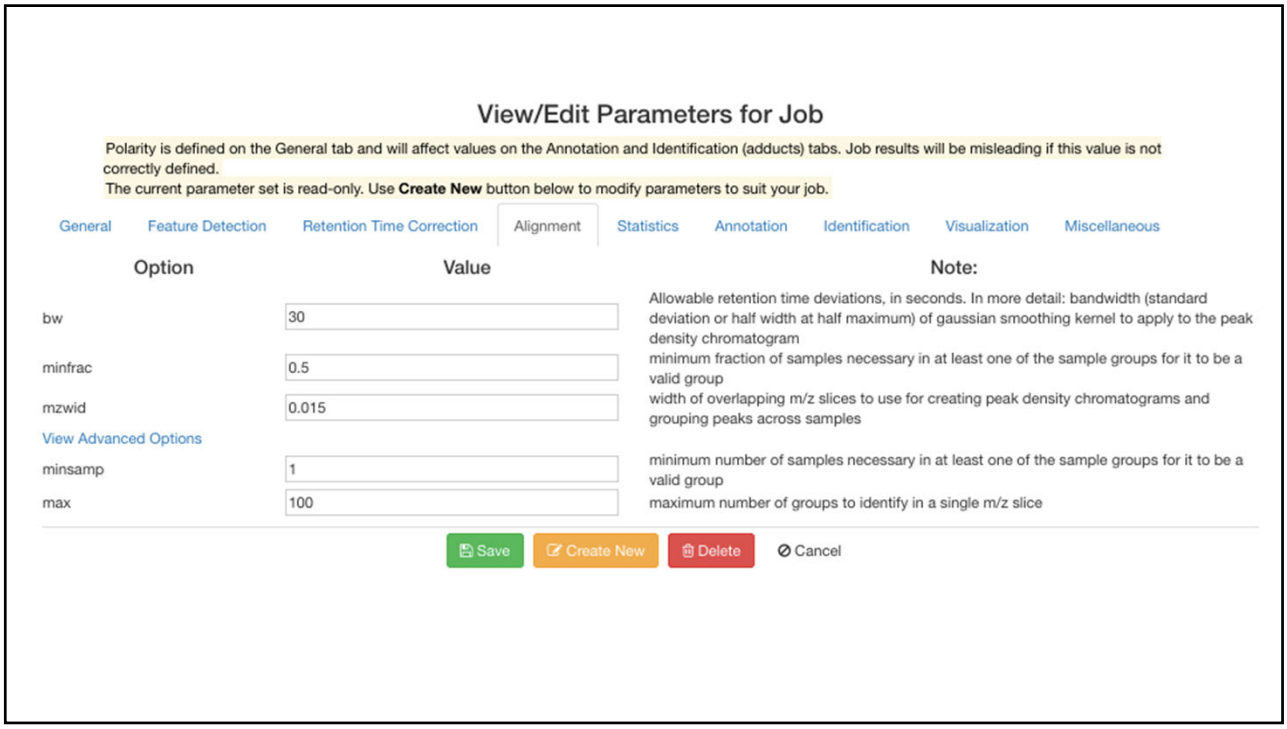
Stephen Barnes, PhD

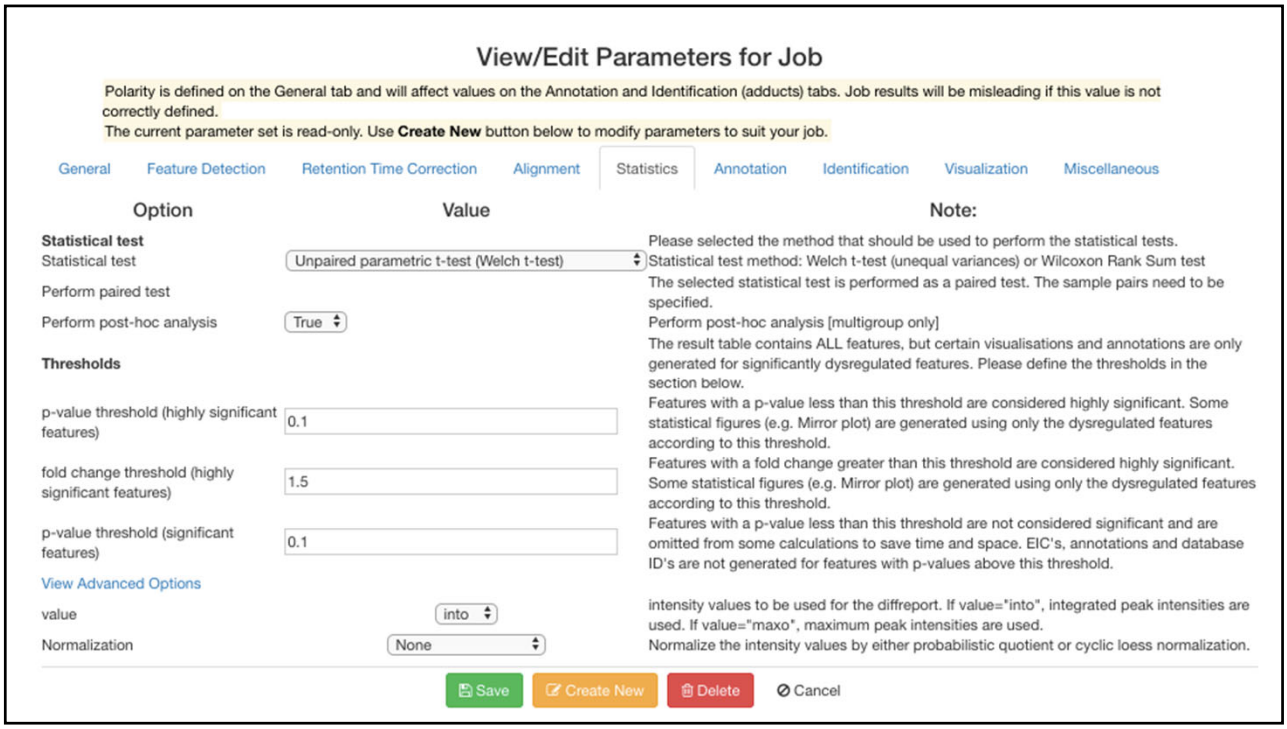
The XCMS parameters I used

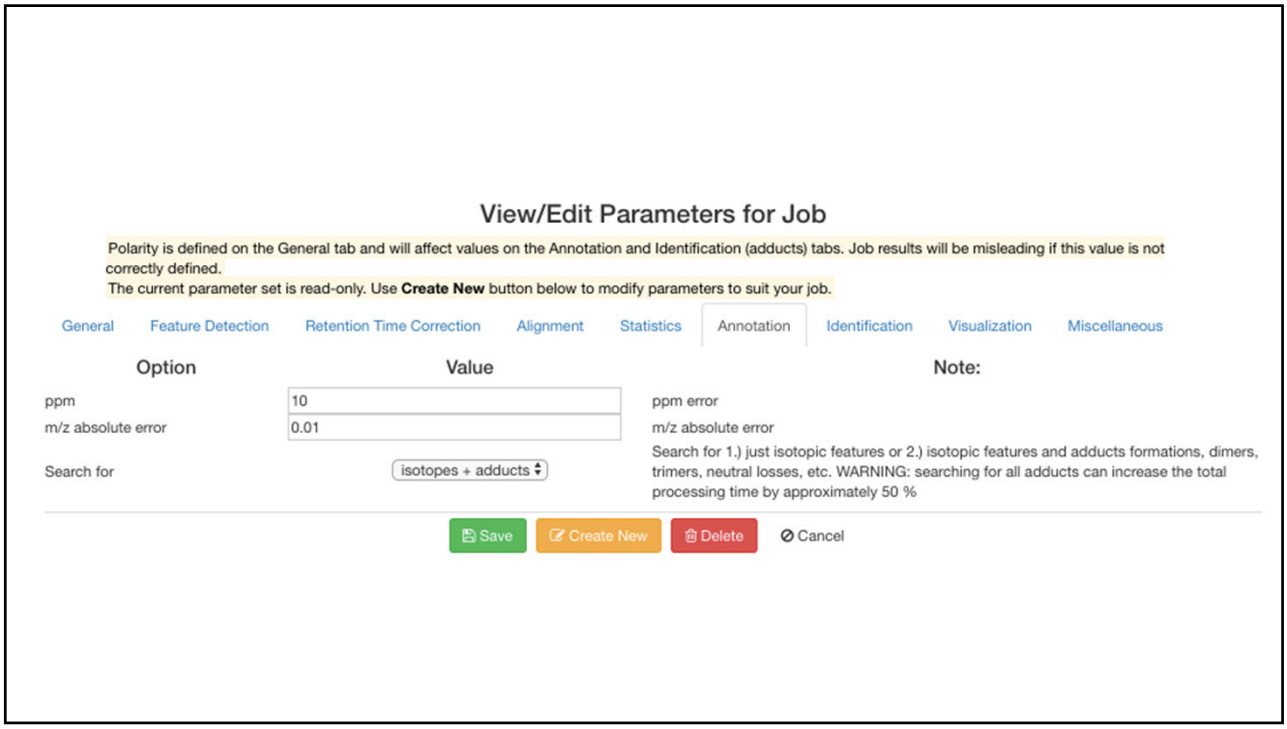

View/Edit Parameters for Job
Polarity is defined on the General tab and will affect values on the Annotation and Identification (adducts) tabs. Job results will be misleading if this value is not correctly defined.
The current parameter set is read-only. Use Create New button below to modify parameters to suit your job.
General
Feature Detection
Retention Time Correction
Alignment
Statistics
Annotation
Identification
Visualization
Miscellaneous
Option
Value
Note:
Name
LW_TripleTof_NegativeMode_30sec tolerance
Comment
(Based on: Custom-2016-07-19_08:08:07 - (Base
Polarity
negative
data acquired in positive or negative mode ?
Retention time format
minutes
show the retention times in results tables and figures in minutes or seconds
Save
Create New
Delete
Cancel

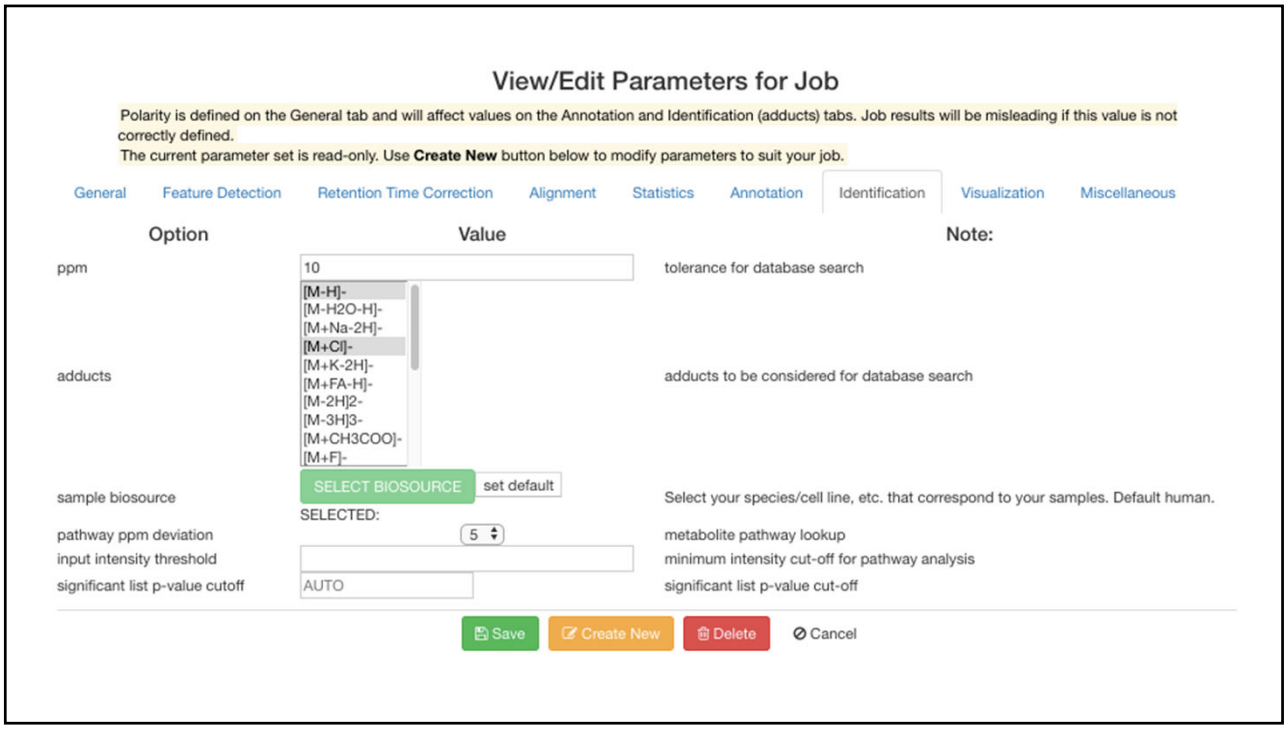

View/Edit Parameters for Job
Polarity is defined on the General tab and will affect values on the Annotation and Identification (adducts) tabs. Job results will be misleading if this value is not correctly defined.
The current parameter set is read-only. Use Create New button below to modify parameters to suit your job.
General
Feature Detection
Retention Time Correction
Alignment
Statistics
Annotation
Identification
Visualization
Miscellaneous
Method:
matchedFilter
Option
Value
Note:
ppm
15
maximal tolerated m/z deviation in consecutive scans, in ppm (parts per million)
minimum peak width
7
minimum chromatographic peak width in seconds
note: must be less than max peak width. See also here.
maximum peak width
30
maximum chromatographic peak width in seconds
note: must be greater than min peak width. See also here.
View Advanced Options
mzdiff
0.01
minimum difference in m/z for peaks with overlapping retention times, can be negative to allow overlap
Signal/Noise threshold
3
Signal/Noise threshold
Integration method
1
Integration method. If =1 peak limits are found through descent on the mexican hat filtered data, if =2 the descent is done on the real data. Method 2 is very accurate but prone to noise, while method 1 is more robust to noise but less exact.
prefilter peaks
3
Prefilter step for the first phase. Mass traces are only retained if they contain at least [prefilter peaks] peaks with intensity >= [prefilter intensity]
prefilter intensity
5
Prefilter step for the first phase. Mass traces are only retained if they contain at least [prefilter peaks] peaks with intensity >= [prefilter intensity]
Noise Filter
0
optional argument which is useful for data that was centroided without any intensity threshold, centroids with intensity < noise are omitted from ROI detection
Save
Create New
Delete
Cancel

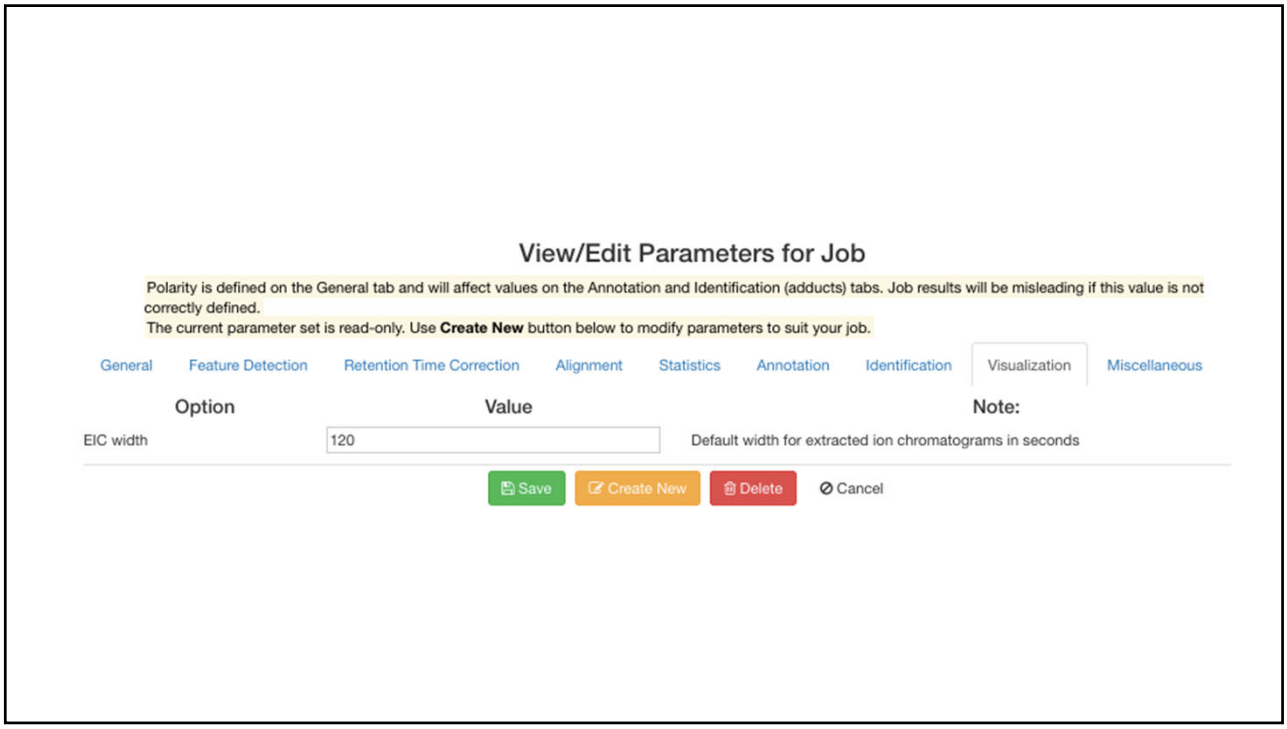

View/Edit Parameters for Job
Polarity is defined on the General tab and will affect values on the Annotation and Identification (adducts) tabs. Job results will be misleading if this value is not correctly defined.
The current parameter set is read-only. Use Create New button below to modify parameters to suit your job.
General
Feature Detection
Retention Time Correction
Alignment
Statistics
Annotation
Identification
Visualization
Miscellaneous
Method: obiwarp
Option
Value
Note:
profStep
1
step size (in m/z) to use for profile generation from the raw data files
Save
Create New
Delete
Cancel


View/Edit Parameters for Job
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.
General
Feature Detection
Retention Time Correction
Alignment
Statistics
Annotation
Identification
Visualization
Miscellaneous
Option
Value
Note:
bw
30
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
minfrac
0.5
minimum fraction of samples necessary in at least one of the sample groups for it to be a valid group
mzwid
0.015
width of overlapping m/z slices to use for creating peak density chromatograms and grouping peaks across samples
View Advanced Options
minsamp
1
minimum number of samples necessary in at least one of the sample groups for it to be a valid group
max
100
maximum number of groups to identify in a single m/z slice
Save
Create New
Delete
Cancel


View/Edit Parameters for Job
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.
General
Feature Detection
Retention Time Correction
Alignment
Statistics
Annotation
Identification
Visualization
Miscellaneous
Option
Value
Note:
Statistical test
Statistical test
Unpaired parametric t-test (Welch t-test)
Please selected the method that should be used to perform the statistical tests.
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]
Thresholds
The result table contains ALL features, but certain visualisations and annotations are only generated for significantly dysregulated features. Please define the thresholds in the section below.
p-value threshold (highly significant features)
0.1
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)
0.1
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.
Save
Create New
Delete
Cancel


View/Edit Parameters for Job
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.
General
Feature Detection
Retention Time Correction
Alignment
Statistics
Annotation
Identification
Visualization
Miscellaneous
Option
Value
Note:
ppm
10
ppm error
m/z absolute error
0.01
m/z absolute error
Search for
isotopes + adducts
Search for 1.) just isotopic features or 2.) isotopic features and adducts formations, dimers, trimers, neutral losses, etc. WARNING: searching for all adducts can increase the total processing time by approximately 50 %
Save
Create New
Delete
Cancel


View/Edit Parameters for Job
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.
General
Feature Detection
Retention Time Correction
Alignment
Statistics
Annotation
Identification
Visualization
Miscellaneous
Option
Value
Note:
ppm
10
tolerance for database search
[M-H]-
[M-H2O-H]-
[M+Na-2H]-
[M+Cl]-
[M+K-2H]-
[M+FA-H]-
[M-2H]2-
[M-3H]3-
[M+CH3COO]-
[M+F]-
adducts
adducts to be considered for database search
SELECT BIOSOURCE
set default
sample biosource
SELECTED:
Select your species/cell line, etc. that correspond to your samples. Default human.
pathway ppm deviation
5
metabolite pathway lookup
input intensity threshold
minimum intensity cut-off for pathway analysis
significant list p-value cutoff
AUTO
significant list p-value cut-off
Save
Create New
Delete
Cancel


View/Edit Parameters for Job
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.
General
Feature Detection
Retention Time Correction
Alignment
Statistics
Annotation
Identification
Visualization
Miscellaneous
Option
Value
Note:
EIC width
120
Default width for extracted ion chromatograms in seconds
Save
Create New
Delete
Cancel

View/Edit Parameters for Job

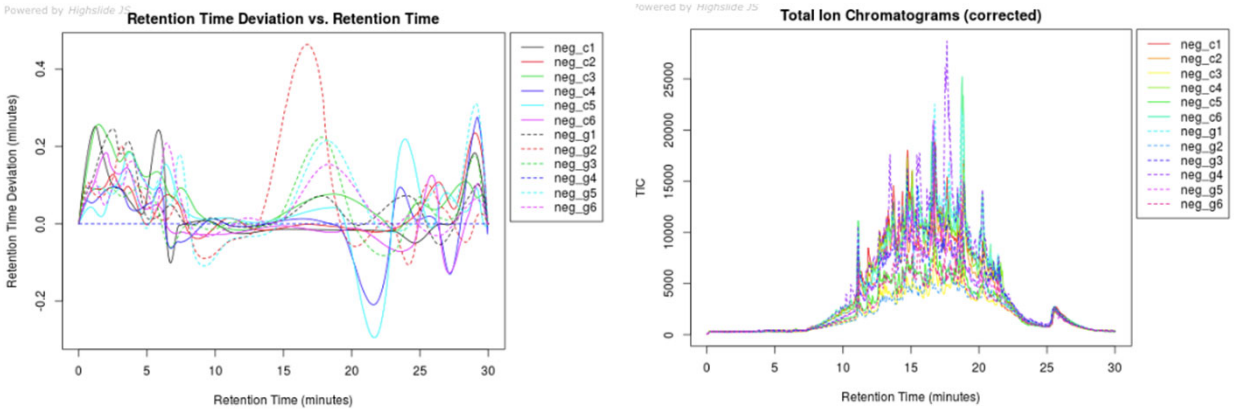
Polarity is defined on the General tab and will affect values on the Annotation and Identification (adducts) tabs. Job results will be misleading if this value is not correctly defined.
The current parameter set is read-only. Use **Create New** button below to modify parameters to suit your job.

General | Feature Detection | Retention Time Correction | Alignment | Statistics | Annotation | Identification | Visualization | Miscellaneous

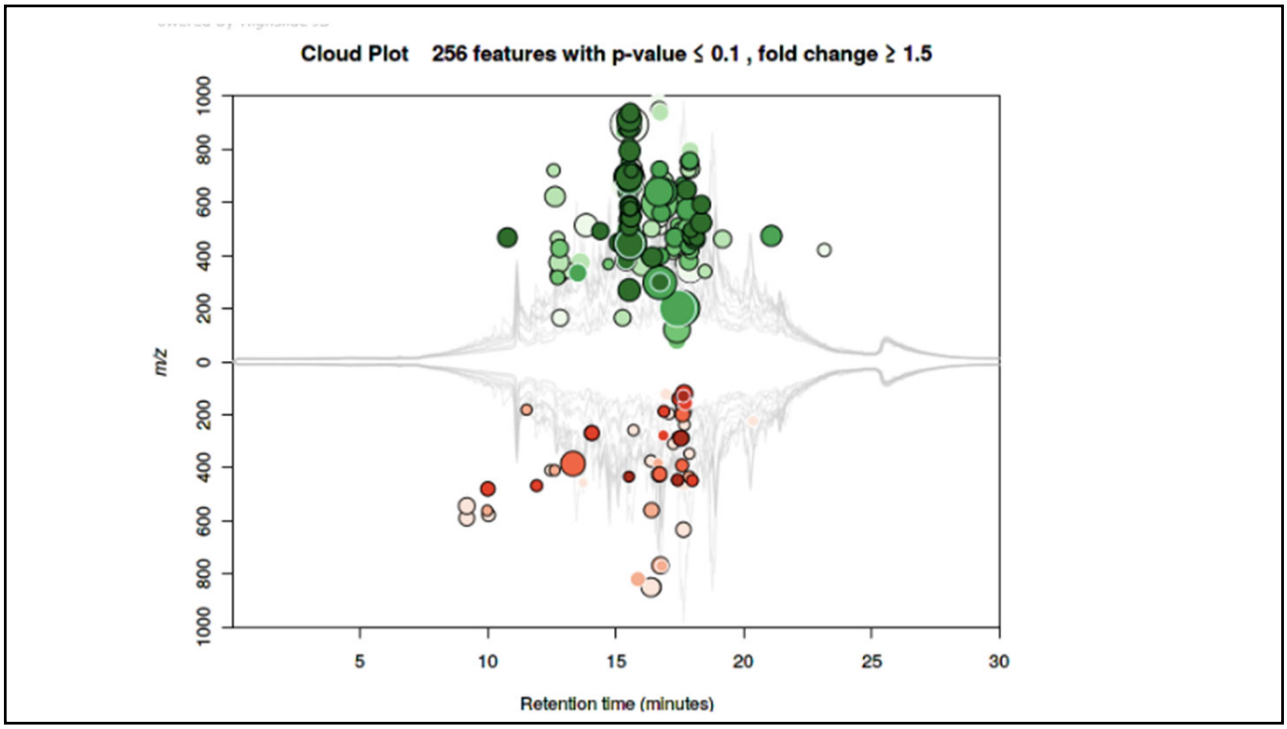
Option	Value	Note:
Correct mass calibration gaps	☐	Correction of mass calibration gaps - subtract LockMass scans from data. Only applicable for Waters instruments !
Bypass file sanity check	☐	All uploaded files are normally checked for different types of errors and inconsistencies. This option disables the file sanity check. Unfortunately necessary for certain types of GC/MS data when used with centWave. Only disable this check if you know what you are doing!

Save | Create New | Delete | Cancel

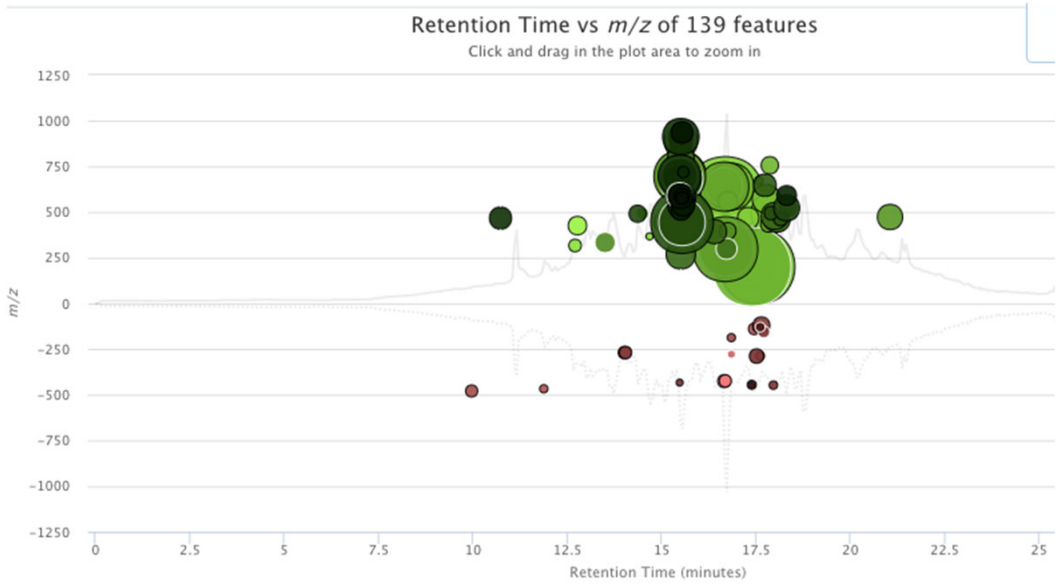
Chromatography of mouse urine samples



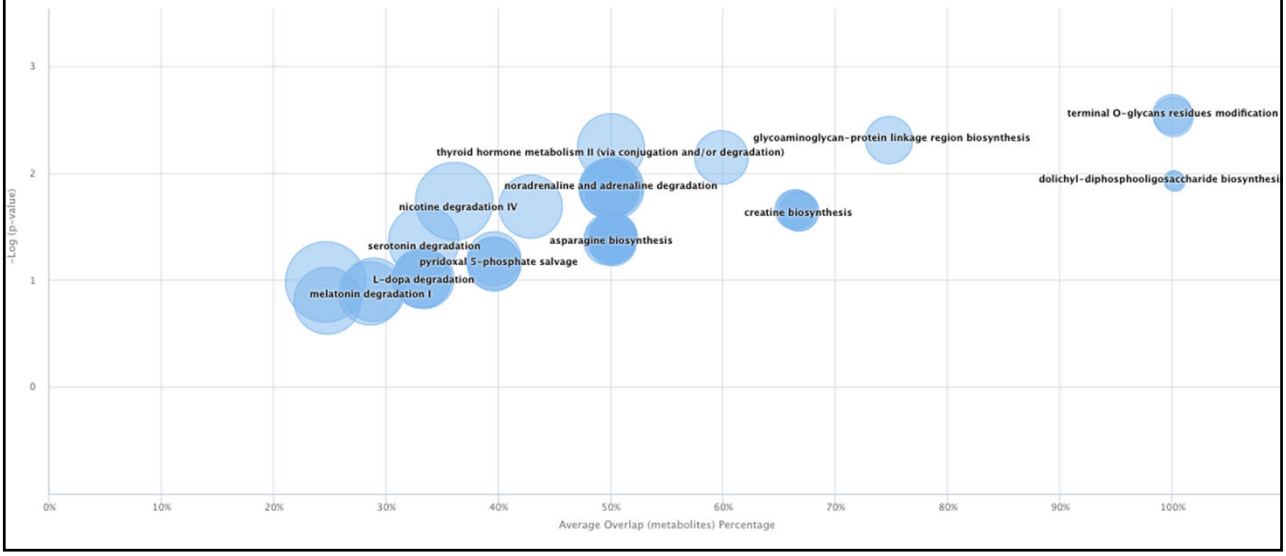
The dotted lines are the genistein-containing samples – they cause a disturbance of retention times between 15-20 min



Interactive Cloud plot (p<0.05)



Pathway analysis



Job#1150303 : TrygveNegMode_Urine_031217

Columns | Hide isotopic peaks | Page 1 of 36

featureidx	fold	pvalue	updown	mzmed	rtmed	maxint
6	2.5	0.00125	UP	565.0197	15.53	3
7	2.7	0.00138	UP	587.0171	15.53	4
8	1.2	0.00143	DOWN	399.8719	25.95	2
9	1.7	0.00177	DOWN	445.2424	17.41	43
10	3.3	0.00179	UP	508.0811	15.52	22
11	4.8	0.00212	UP	591.0061	15.50	8
12	3.3	0.00270	UP	591.0354	18.34	7
13	2.6	0.00280	UP	513.0706	15.55	54
14	2.0	0.00280	UP	718.1110	15.60	4
15	4.8	0.00284	UP	535.0422	15.53	14
16	5.2	0.00296	UP	575.0318	15.56	12
17	3.5	0.00330	UP	590.0077	15.50	20
18	14.0	0.00403	UP	695.5807	15.50	10
19	4.1	0.00410	UP	546.0116	15.56	9
20	3.9	0.00416	UP	467.0841	10.75	12
21	9.2	0.00420	UP	913.1462	15.51	29
22	3.1	0.00448	UP	585.0249	15.53	5
23	2.5	0.00515	UP	592.0140	15.58	3
24	2.9	0.00577	UP	708.1911	15.55	24

Data organized by p-values

Organize data by retention times

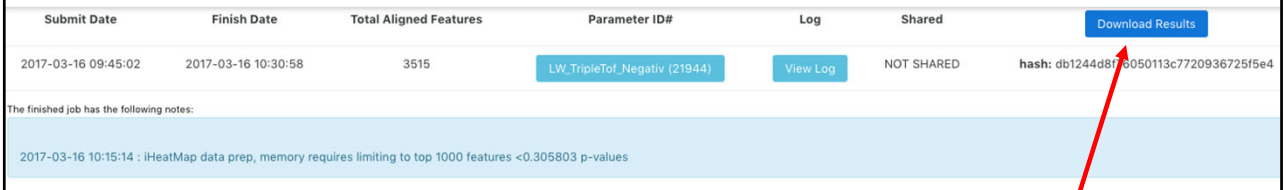
1.1	0.79265	DOWN	647.7817	8.67	4
1.3	0.34537	DOWN	354.4955	8.68	1
1.4	0.28009	DOWN	353.4950	8.71	3
1.4	0.29076	DOWN	662.9952	8.71	7
1.4	0.30109	DOWN	661.9921	8.71	22
1.4	0.26342	DOWN	329.4948	8.71	2
1.4	0.26943	DOWN	331.4926	8.71	4
1.4	0.27764	DOWN	378.4863	8.71	2
1.4	0.28574	DOWN	330.9940	8.72	8
1.4	0.29122	DOWN	330.4925	8.72	24
1.4	0.31924	DOWN	663.9925	8.72	4
1.5	0.42966	DOWN	880.3870	8.77	8

See the ^{13}C-isomers

See the evidence for a doubly charged state

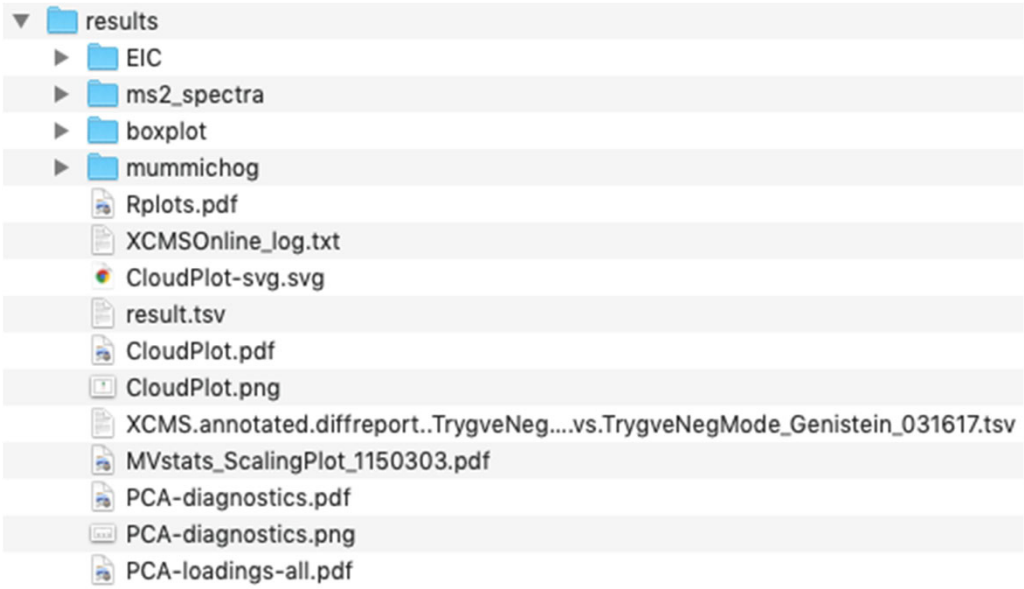
The true *m/z* value is 661.9921. This mass suggests that the compound contains an odd number of nitrogen atoms and it is highly oxygenated and contains sulfate or phosphate groups,

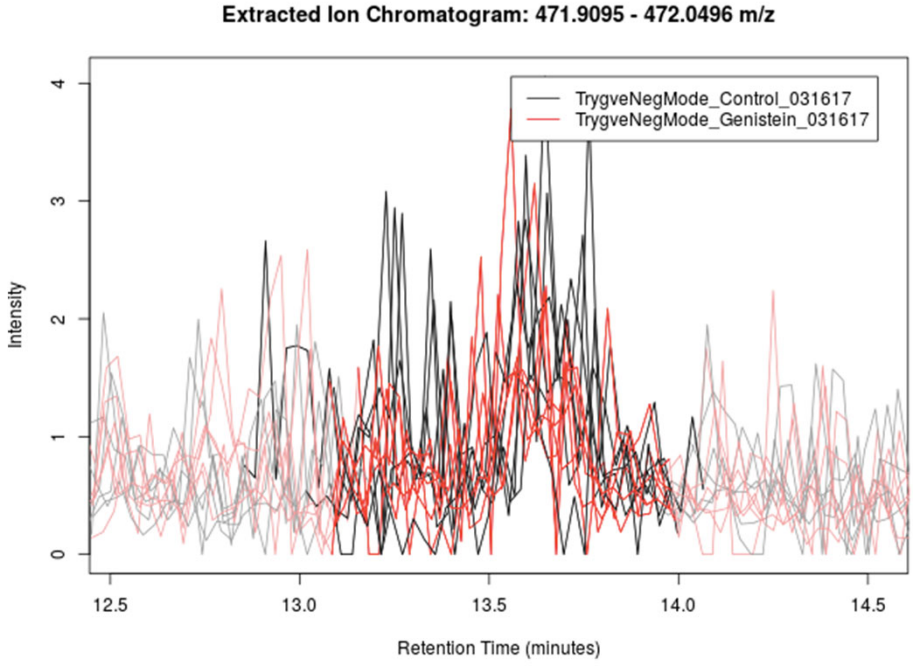
Let's download all the XCMS data

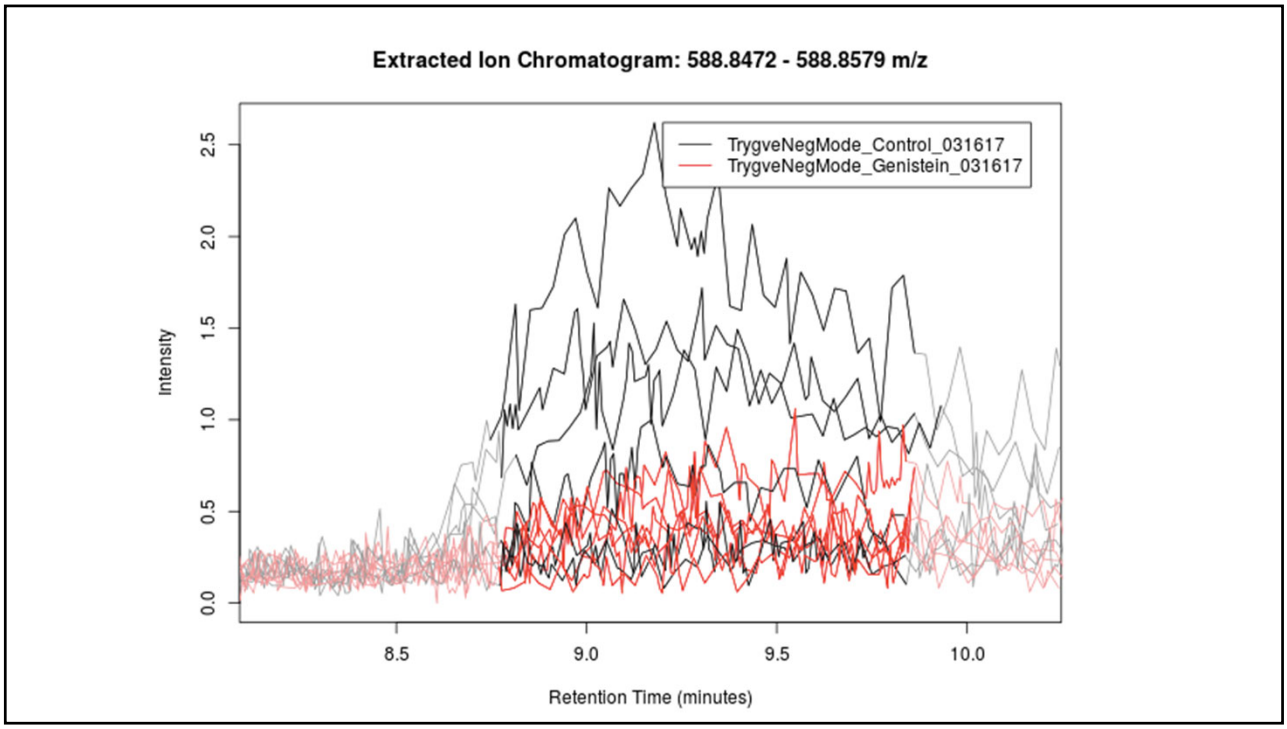


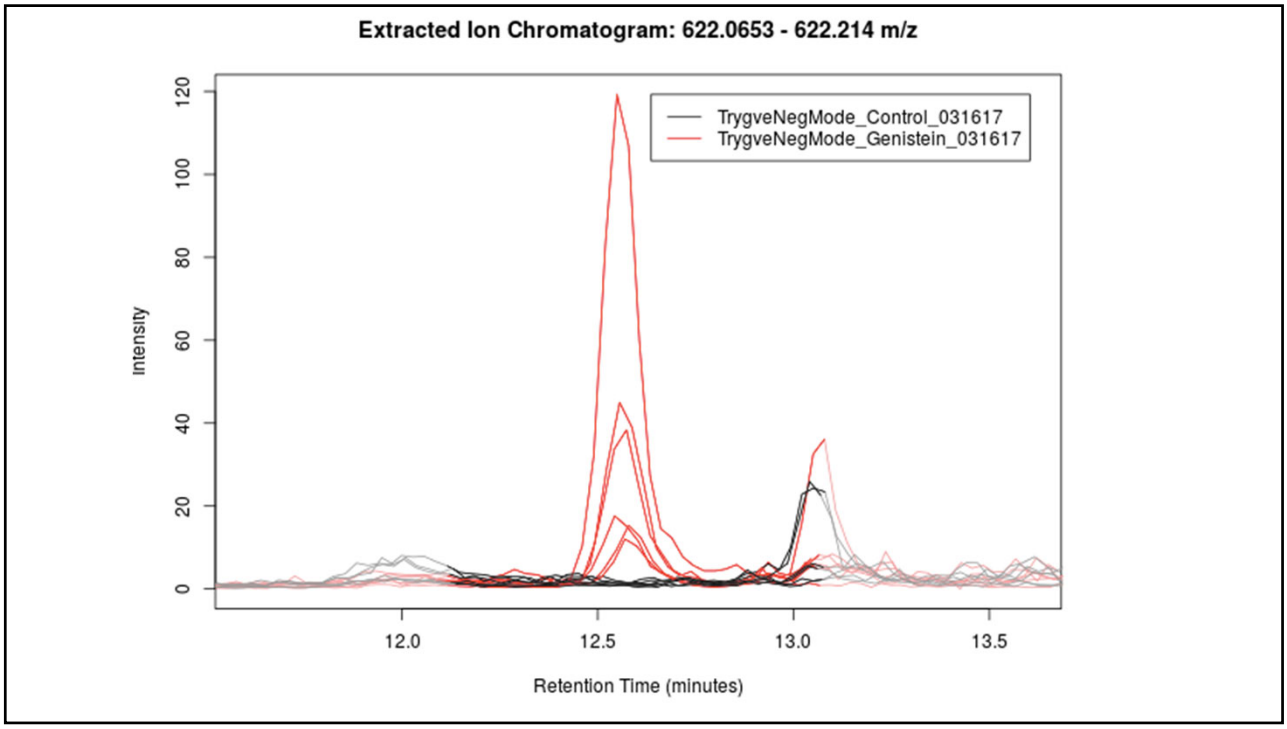
This should produce a ~50 MB zip file – unzip it!

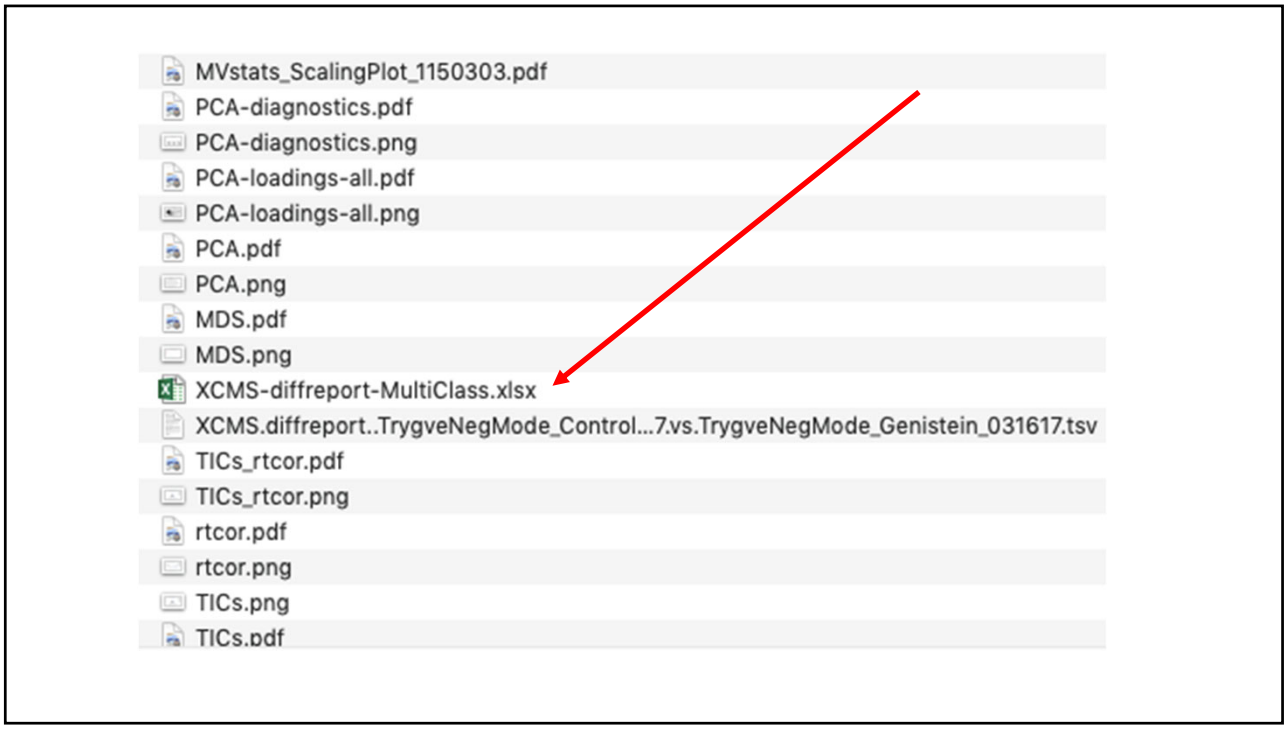
Then open the Results folder


results
EIC
ms2_spectra
boxplot
mummichog
Rplots.pdf
XCMSOnline_log.txt
CloudPlot-svg.svg
result.tsv
CloudPlot.pdf
CloudPlot.png
XCMS.annotated.diffreport..TrygveNeg....vs.TrygveNegMode_Genistein_031617.tsv
MVstats_ScalingPlot_1150303.pdf
PCA-diagnostics.pdf
PCA-diagnostics.png
PCA-loadings-all.pdf


Extracted Ion Chromatogram: 471.9095 - 472.0496 m/z
TrygveNegMode_Control_031617
TrygveNegMode_Genistein_031617
Intensity
0
1
2
3
4
12.5
13.0
13.5
14.0
14.5
Retention Time (minutes)


Extracted Ion Chromatogram: 588.8472 - 588.8579 m/z
TrygveNegMode_Control_031617
TrygveNegMode_Genistein_031617
Intensity
0.0
0.5
1.0
1.5
2.0
2.5
8.5
9.0
9.5
10.0
Retention Time (minutes)


Extracted Ion Chromatogram: 622.0653 - 622.214 m/z
TrygveNegMode_Control_031617
TrygveNegMode_Genistein_031617
Intensity
0
20
40
60
80
100
120
12.0
12.5
13.0
13.5
Retention Time (minutes)



	name	fold	log2fold	tstat	pvalue	qvalue	updown	mzmed	mzmin	mzmax	rtmed	rtmin	rtmax
2	M576T16	2.140863	1.098192	6.471971	0.000187	1	UP	576.0311	576.0295	576.0343	15.553	15.504	15.572
3	M588T16	2.503567	1.323985	5.924246	0.000274	1	UP	588.0089	588.0058	588.0124	15.539	15.452	15.553
4	M586T16	4.399275	2.137266	7.286641	0.000405	1	UP	586.0128	586.0107	586.0144	15.529	15.504	15.572
5	M936T16	3.705296	1.889589	7.288022	0.000524	1	UP	936.1249	936.1182	936.1251	15.553	15.520	15.585
6	M544T16	4.175901	2.062087	6.676605	0.000975	1	UP	544.0380	544.0277	544.0401	15.582	15.176	15.634
7	M565T16	2.499698	1.321754	5.367571	0.001252	1	UP	565.0197	565.0178	565.0230	15.530	15.199	15.546
8	M587T16	2.684865	1.42485	5.027966	0.001384	1	UP	587.0171	587.0103	587.0213	15.530	15.199	16.130
9	M400T26	1.206201	-0.27047	-4.35801	0.001426	1	DOWN	399.8719	399.8648	399.8725	25.948	25.922	25.989
10	M445T17	1.717576	-0.78037	-4.22913	0.001766	1	DOWN	445.2424	445.2296	445.2437	17.406	17.333	17.507
11	M508T16	3.277203	1.712465	5.257903	0.001788	1	UP	508.0811	508.0762	508.0856	15.515	15.199	15.531
12	M591T16	4.838643	2.274602	5.525247	0.002122	1	UP	591.0061	591.0034	591.0077	15.504	15.199	15.580
13	M591T18	3.32535	1.733506	4.647917	0.002695	1	UP	591.0354	591.0306	591.0429	18.338	18.219	18.428
14	M513T16_	2.55776	1.354881	4.319248	0.002796	1	UP	513.0706	513.0693	513.0715	15.546	15.199	15.555
15	M718T16	1.978588	0.984471	4.341172	0.002799	1	UP	718.1110	718.1045	718.1116	15.598	15.530	15.693
16	M535T16	4.812779	2.26687	5.262474	0.002836	1	UP	535.0422	535.0415	535.0425	15.525	15.477	15.660
17	M575T16	5.225725	2.385631	5.278349	0.002963	1	UP	575.0318	575.0301	575.0334	15.557	15.530	15.740
18	M590T16	3.467603	1.793939	4.749198	0.003298	1	UP	590.0077	590.0032	590.0143	15.504	15.199	15.546
19	M696T16	13.97159	3.804424	4.93442	0.00403	1	UP	695.5807	695.5769	695.5827	15.502	15.176	15.546
20	M546T16_	4.050259	2.018014	4.91106	0.004103	1	UP	546.0116	546.0095	546.0118	15.557	15.553	15.987
21	M467T11	3.852899	1.945944	4.526655	0.00416	1	UP	467.0841	467.0819	467.0859	10.748	10.672	10.771
22	M913T16_	9.23821	3.207613	4.951031	0.004201	1	UP	913.1462	913.1407	913.1483	15.512	15.199	15.580
23	M585T16	3.085587	1.625545	4.322821	0.004476	1	UP	585.0249	585.0209	585.0259	15.530	15.526	15.546
24	M592T16	2.466057	1.302206	4.206206	0.005148	1	UP	592.0140	592.0085	592.0160	15.578	15.504	15.598
25	M708T16	2.894059	1.533094	4.110799	0.005769	1	UP	708.1911	708.1893	708.1943	15.546	15.222	15.687
26	M294T16	1.261512	-0.33515	-3.70092	0.005784	1	DOWN	294.0246	294.0218	294.0272	16.005	15.564	16.086
27	M679T15	3.994443	1.997994	4.312709	0.006407	1	UP	679.1211	679.1176	679.1290	15.445	15.176	15.477
28	M879T16	3.697832	1.88668	4.052701	0.006754	1	UP	879.2968	879.2888	879.2979	15.510	15.477	15.531
29	M130T18_	1.823658	-0.86684	-3.58867	0.006937	1	DOWN	129.6658	129.6619	129.6722	17.631	17.262	17.659

npeaks	TrygveNegMode_Cor	TrygveNegMode_G	maxint	mean1	sd1	mean2	sd2
3	0	3	4.043847	14.132	3.082039	30.25468	5.26651
4	0	4	4.623051	15.15886	5.0965	37.95122	7.92692
4	0	4	5.232004	11.68306	3.586474	51.397	12.85954
3	0	3	5.582684	6.414646	1.209805	23.76816	5.70563
4	0	4	6.972446	16.56229	2.71291	69.16247	19.10613
3	0	3	3.483185	15.10651	3.905988	37.76171	9.572455
5	0	5	3.817084	15.12652	5.406434	40.61266	11.17728
3	3	0	1.823584	36.72821	2.513168	30.44949	2.477544
5	5	0	43.07945	1065.08	175.5487	620.1065	188.694
4	0	4	22.43803	89.88402	30.49938	294.5682	90.34667
5	0	5	8.250369	13.32621	4.285147	64.48076	22.26965
6	0	6	7.166083	9.577612	4.460669	31.84891	10.85649
3	0	3	53.52679	176.1517	73.97377	450.5538	136.9098
3	0	3	3.779149	25.41962	6.561103	50.29496	12.4079
4	0	4	14.30418	20.55474	5.776563	98.92541	36.01842
3	0	3	12.45094	14.49662	3.536242	75.75536	28.20716
4	0	4	20.02889	54.31329	19.99746	188.3369	66.16949
6	0	6	10.47886	3.564399	2.685634	49.80032	22.79425
3	0	3	8.708616	20.21044	3.665667	81.8575	30.52842
4	0	4	12.21151	44.5073	19.81326	171.4821	65.79065
4	0	4	29.41563	23.51683	5.62778	217.2534	95.68452
3	0	3	5.043276	12.86396	5.150709	39.69287	14.30323
3	0	3	3.472651	12.3396	3.54305	30.43015	9.921402
3	0	3	23.68497	67.72808	25.59442	196.0091	72.02609
5	3	2	3.291979	55.90273	6.573893	44.31409	3.951403
3	0	3	30.82887	37.81171	12.57286	151.0367	63.06738
4	0	4	54.27961	84.28005	41.2275	311.6535	131.0967
7	6	1	1.468472	7.28249	3.171427	1.885698	1.873846
4	0	4	6.4067	2.280975	0.997202	20.41197	10.3555

npeaks is the number of times a feature is detected

Only seen in the genistein group

neg_c1	neg_c2	neg_c3	neg_c4	neg_c5	neg_c6	neg_g1	neg_g2	neg_g3	neg_g4	neg_g5	neg_g6
14.17836	10.46409	19.65348	12.44152	14.60724	13.44734	21.12261	32.53664	28.48849	33.43121	36.29671	29.6524
19.69169	16.42004	14.48966	19.0051	15.78477	5.561922	29.35966	37.8945	51.1383	30.30381	38.25088	40.76021
12.90475	9.75161	12.25851	15.87307	13.71597	5.594468	29.69663	63.24869	58.9191	54.08966	42.50846	59.91945
8.179828	6.945014	6.399402	6.69312	5.668885	4.601626	18.63101	25.13095	31.28403	24.93339	15.62974	26.99983
16.0627	16.19549	16.67228	21.01522	16.9211	12.50694	67.44165	74.30632	63.5489	104.0736	50.54532	55.05907
20.63061	10.56439	16.51313	13.40581	17.9272	11.59795	21.95191	45.73435	45.36661	33.54841	34.54489	45.42412
14.29446	17.00035	16.81882	15.13024	21.98356	5.531669	26.91533	51.94236	33.21791	56.00836	37.05377	38.53821
35.73576	36.63504	33.76354	38.36882	40.77931	35.0868	27.59385	31.44155	28.93661	28.41511	32.66187	33.64795
1193.963	1034.867	760.5618	1122.823	1260.644	1017.622	912.1075	373.8555	622.5087	728.6583	605.2199	478.289
132.0667	92.11987	49.56156	116.027	67.68868	81.84038	262.2396	287.1675	276.5834	473.0071	222.7697	245.6418
13.94604	13.32463	13.59107	14.82712	18.70167	5.566715	40.07138	103.2311	62.00579	66.88835	45.75761	68.93025
6.790679	9.121331	14.40239	14.56422	9.594867	2.992197	16.02655	26.46855	45.77398	42.37089	29.96324	30.49028
293.9329	171.0641	87.70381	210.79	109.9279	183.4916	309.0992	681.6804	361.0186	473.3554	363.279	514.8903
30.58506	30.26517	21.6269	31.95002	22.70482	15.38577	68.84389	37.5695	59.62803	52.31969	38.17052	45.23811
28.15973	18.02209	17.24745	26.80342	19.74938	13.34637	57.9166	139.0121	83.77499	133.0427	61.25313	118.5529
13.69304	19.38434	15.95985	15.41081	13.92451	8.607188	43.16344	118.959	65.90559	71.21568	56.1423	99.14615
72.54606	47.81551	29.65324	80.27969	36.44743	59.13778	83.96173	280.4942	192.1203	227.2418	157.09	189.1136
8.627719	2.457436	1.339852	4.458987	2.214647	2.287751	31.98686	51.16246	60.9011	87.00934	22.39295	45.3492
25.68848	15.77104	18.63453	23.10637	20.41252	17.64967	45.23223	71.2599	92.98628	135.2482	66.7983	79.62007
71.2904	42.35299	22.42719	64.01633	25.4334	41.52351	110.8404	186.8562	175.153	285.6262	103.1588	167.258
28.7406	24.1518	12.79132	26.84819	22.98928	25.57977	134.3484	240.6717	216.2402	386.3223	118.9402	206.9976
20.23427	15.69036	11.06857	13.8287	11.52182	4.840029	24.33643	58.92589	36.86988	27.60374	35.15641	55.26485
15.86213	12.61597	15.85998	11.34527	12.08377	6.27048	14.30224	23.15816	32.07043	38.74999	40.06171	34.23838
91.50297	77.2206	30.56778	90.99144	42.06881	74.01687	299.3855	98.09142	207.5446	248.605	174.3181	148.1098
55.1268	56.84738	43.44796	62.42395	59.18612	58.38419	48.68429	46.64461	46.14431	44.36588	42.60669	37.43875
54.30951	41.03856	26.18081	50.33062	28.79996	26.2108	119.6739	162.433	143.8028	270.1109	92.39981	117.7998

These are the peaks areas of features in the dataset

Let's do some data filtering

Copy the entire sheet and copy it to a new sheet

Sort the data according to Rtmed

Delete those rows with RT > 5 min and <25 min

3260	M496T25	1.13935713	0.18822003	1.28839883	0.2326318	1	UP	495.9715	495.9689	495.9738	25.000	24.959	25.010
3261	M431T25	1.02187715	0.03122177	0.30512102	0.76722602	1	UP	430.9727	430.9708	430.9741	25.002	24.959	25.058
3262	M385T25	1.00314395	0.00452865	0.03238222	0.97528182	1	UP	384.9347	384.9331	384.9361	25.003	24.934	25.097
3263	M495T25	1.00598704	0.00861172	0.07204178	0.94399072	1	UP	494.9683	494.9659	494.9699	25.008	24.959	25.064
3264	M249T25	1.02399144	-0.0342037	-0.3310653	0.74779385	1	DOWN	248.9613	248.9600	248.9622	25.040	24.963	25.129
3265	M519T25	1.00369664	0.00532329	0.04125286	0.96791297	1	UP	519.2764	519.2725	519.2778	25.176	25.095	25.230
	A	B	C	D	E	F	G	H	I	J	K	L	M
3492	M191T26	1.31704937	-0.3973094	-0.9442564	0.3689252	1	DOWN	191.0211	191.0194	191.0217	26.141	25.973	26.193
3493	M237T26	1.10136006	0.1392862	0.74119801	0.48022559	1	UP	236.7609	236.7586	236.7649	26.152	25.989	26.193
3494	M233T26	1.12846327	0.17435947	0.77205164	0.45822872	1	UP	232.7648	232.7633	232.7657	26.153	26.058	26.198
3495	M235T26_1	1.03210321	0.04558724	0.25440049	0.80575091	1	UP	234.7621	234.7614	234.7633	26.156	26.020	26.192
3496	M231T26_1	1.14337733	0.19330159	0.97320886	0.35643421	1	UP	230.7675	230.7667	230.7681	26.166	26.027	26.198
3497	M267T26	1.08758325	-0.1211258	-1.2906024	0.23459315	1	DOWN	266.9280	266.9232	266.9300	26.184	26.065	26.296
3498	M114T26_1	1.04389879	-0.0619818	-0.6177682	0.55056772	1	DOWN	113.5474	113.5456	113.5561	26.192	25.978	26.289
3499	M291T26	1.1196964	-0.1631076	-0.8605603	0.4096874	1	DOWN	290.9425	290.9382	290.9437	26.208	26.066	26.296
3500	M155T26_2	1.04750638	-0.066959	-1.1551683	0.2765622	1	DOWN	154.9758	154.9741	154.9764	26.208	26.078	26.386
3501	M268T26	1.05535058	-0.0777223	-1.0020458	0.35644544	1	DOWN	267.9044	267.9024	267.9056	26.229	25.892	26.376
3502	M447T26	1.56275105	-0.644088	-1.3606862	0.21251774	1	DOWN	447.1333	447.1318	447.1347	26.271	26.146	26.327
3503	M354T26_2	1.14563806	-0.1961513	-0.9475269	0.37021976	1	DOWN	353.9051	353.8996	353.9069	26.278	26.098	26.528
3504	M290T26	1.14800962	-0.1991347	-0.5631364	0.587771	1	DOWN	289.9370	289.9359	289.9377	26.283	26.139	26.384
3505	M273T26	1.03151981	-0.0447715	-0.6897177	0.50614595	1	DOWN	272.9430	272.9416	272.9441	26.312	26.219	26.528
3506	M211T26	1.00347262	-0.0050013	-0.0237544	0.98154184	1	DOWN	210.8950	210.8898	210.8979	26.344	26.298	26.482
3507	M384T26	1.15806423	-0.2117153	-0.9280129	0.37907712	1	DOWN	383.8822	383.8744	383.8838	26.359	26.264	26.614
3508	M133T26	1.05007905	-0.0704979	-0.9539953	0.3674681	1	DOWN	132.8695	132.8677	132.8700	26.373	25.721	26.510
3509	M137T26	1.05629851	-0.0790176	-0.7575922	0.4666127	1	DOWN	136.8663	136.8641	136.8687	26.378	25.670	26.483
3510	M135T26	1.04187023	-0.0591756	-0.8356325	0.42641599	1	DOWN	134.8670	134.8660	134.8675	26.390	25.721	26.575
3511	M257T27	1.08504035	-0.1177487	-0.7529343	0.47580298	1	DOWN	256.9371	256.9360	256.9378	26.563	26.421	26.815
3512	M289T27	1.07155784	-0.0997097	-1.0213582	0.3333099	1	DOWN	288.9097	288.9086	288.9103	26.626	26.483	26.722
3513	M269T27	1.12412293	-0.1687998	-2.0245118	0.07378117	1	DOWN	268.9276	268.9266	268.9282	26.641	26.463	26.731
3514	M205T27	1.07996919	-0.1109902	-1.8455011	0.09520189	1	DOWN	204.9580	204.9564	204.9585	26.685	26.599	26.873
3515	M159T27	1.00698086	0.01003626	0.10659347	0.91733639	1	UP	158.9528	158.9509	158.9531	26.713	26.629	26.815
3516	M340T27	1.19051121	-0.2515812	-0.7089022	0.49700184	1	DOWN	339.9193	339.9179	339.9199	26.725	26.641	26.940
3517	M255T27	1.1106063	-0.1513475	-3.0632443	0.02146366	1	DOWN	254.9397	254.9386	254.9404	26.925	26.838	27.061

Create .csv files for each sample

A	B	C
mzmed	Rtmed	neg_c1
105.0217	5.186	76.412322
333.9268	6.372	135.734375
224.0226	6.494	10.1311084
147.0316	6.821	23.3679915
167.0227	7.152	16.3293078
424.0473	7.464	80.6544717
164.0579	7.498	6.5289731
85.0321	7.792	12.4028658
174.0145	7.798	29.2833668
129.0213	7.803	131.087296
173.0107	7.803	391.377962
416.9733	8.005	102.454985
956.7816	8.047	3.88426323
955.7747	8.050	8.40612396
260.0211	8.270	32.5742871
878.4950	8.417	4.38984792
495.6045	8.425	2.61264864
498.5982	8.473	2.39346769
194.0469	8.517	102.530871
939.3694	8.613	81.6713453
627.7495	8.614	16.3475148
592.7294	8.670	9.13564246
647.7817	8.672	13.7556701

Copy the *m/z*, RT and sample column data from the filtered Excel file

Save this file in .csv format. Note the name you give the file must not have a space in it

To create the next file, copy just the neg_c2 column (the *m/z* and RT don't change) and save as before as a .csv file

mzmed	Rtmed	neg_c2
105.0217	5.186	59.7180471
333.9268	6.372	166.739963
224.0226	6.494	9.60804319
147.0316	6.821	11.583249
167.0227	7.152	13.9878682
424.0473	7.464	35.5184099
164.0579	7.498	4.7443522
85.0321	7.792	7.17078563
174.0145	7.798	19.3911935
129.0213	7.803	91.1811272
173.0107	7.803	275.436633
416.9733	8.005	38.1917168
956.7816	8.047	3.14255134
955.7747	8.050	4.85878663
260.0211	8.270	24.4317841
878.4950	8.417	4.11934488
495.6045	8.425	2.2000076
498.5982	8.473	3.39449255
194.0469	8.517	127.061463
939.3694	8.613	69.0206465
627.7495	8.614	15.7615689
592.7294	8.670	10.1267181
647.7817	8.672	10.737514

Preparing the .zip file for uploading

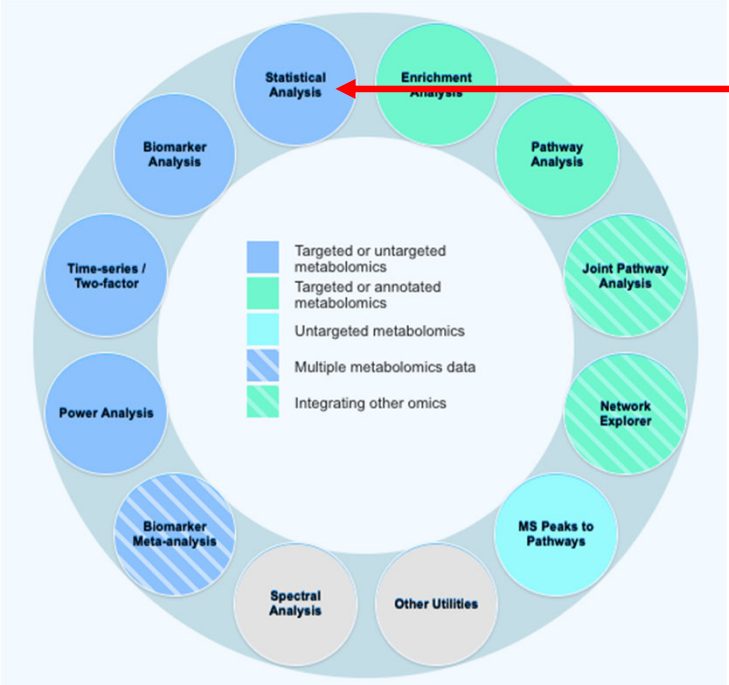
- **Take all the .csv files for the control group and put them in a folder labelled "control" – note, whatever you call the folder it must not have a space in it.**
- **Now repeat the exercise for the genistein group and make the .csv files and put them in a folder labelled "genistein".**
- **Highlight the two folders and convert them to a .zip file and rename it – again, no spaces**
- **Now we're ready for Metaboanalyst (http://www.metaboanalyst.ca)**

MetaboAnalyst - statistical, functional and integrative analysis of metabolomics data

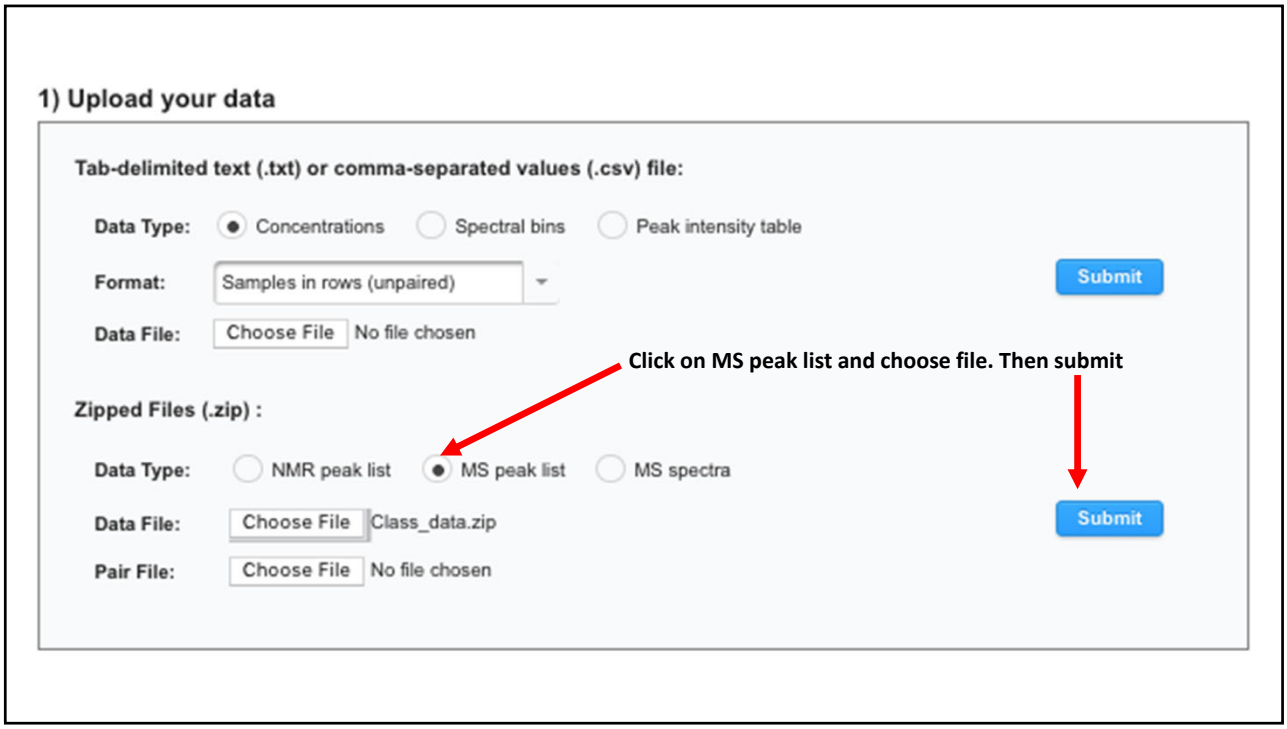
Welcome >> click here to start <<

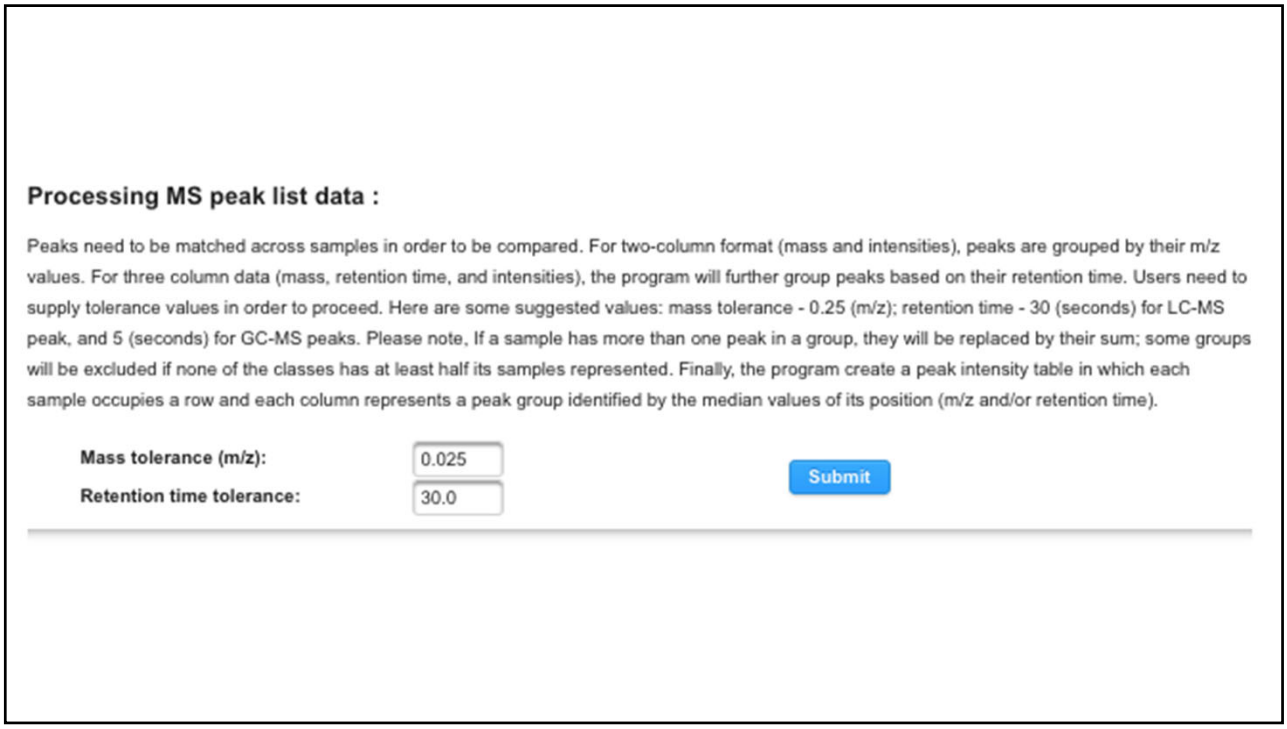
News & Updates

- Check out OmicsNet for multi-omics data integration via 3D network visual analytics;
- Check out MicrobiomeAnalyst for comprehensive analysis of microbiome data;
- Check out NetworkAnalyst for comprehensive gene expression & network analysis;
- Enhanced KEGG pathway generation to address the occasional failure issue during peak time (01/24/2019); NEW
- Fixed the issue for compound view in Pathway Analysis module (01/07/2019); NEW
- Enhanced support for parsing data input and data editing (01/02/2019); NEW
- Updated the Joint Pathway Analysis module to be consistent with Pathway Analysis (12/21/2018); NEW
- Fixed the issue with data filtering (12/18/2018); NEW
- Enhanced error handling for the mummichog analysis module (11/15/2018); NEW
- Upgraded to HTTPs for more secure communications (11/05/2018); NEW
- Enhanced graphics for scores and loadings plots in chemometrics methods (10/06/2018);
- Enhanced visualization for SMPDB pathways (09/28/2018);
- Fixed the performance issue with Data Editor (09/21/2018);
- Minor interface enhancement based on user feedback (09/11/2018);
- Fixed the issue with mummichog p value computing (08/03/2018);



Choose the Statistical Analysis option


1) Upload your data
Tab-delimited text (.txt) or comma-separated values (.csv) file:
Data Type: Concentrations Spectral bins Peak intensity table
Format: Samples in rows (unpaired)
Submit
Data File: Choose File No file chosen
Click on MS peak list and choose file. Then submit
Zipped Files (.zip) :
Data Type: NMR peak list MS peak list MS spectra
Data File: Choose File Class_data.zip
Submit
Pair File: Choose File No file chosen


Processing MS peak list data :
Peaks need to be matched across samples in order to be compared. For two-column format (mass and intensities), peaks are grouped by their m/z values. For three column data (mass, retention time, and intensities), the program will further group peaks based on their retention time. Users need to supply tolerance values in order to proceed. Here are some suggested values: mass tolerance - 0.25 (m/z); retention time - 30 (seconds) for LC-MS peak, and 5 (seconds) for GC-MS peaks. Please note, If a sample has more than one peak in a group, they will be replaced by their sum; some groups will be excluded if none of the classes has at least half its samples represented. Finally, the program create a peak intensity table in which each sample occupies a row and each column represents a peak group identified by the median values of its position (m/z and/or retention time).
Mass tolerance (m/z): 0.025
Retention time tolerance: 30.0
Submit

Processing MS peak list data :

Peaks need to be matched across samples in order to be compared. For two-column format (mass and intensities), peaks are grouped by their m/z values. For three column data (mass, retention time, and intensities), the program will further group peaks based on their retention time. Users need to supply tolerance values in order to proceed. Here are some suggested values: mass tolerance - 0.25 (m/z); retention time - 30 (seconds) for LC-MS peak, and 5 (seconds) for GC-MS peaks. Please note, If a sample has more than one peak in a group, they will be replaced by their sum; some groups will be excluded if none of the classes has at least half its samples represented. Finally, the program create a peak intensity table in which each sample occupies a row and each column represents a peak group identified by the median values of its position (m/z and/or retention time).

Mass tolerance (m/z): 0.001
Retention time tolerance: 0.005

Submit

Processing MS peak list data :

Peaks need to be matched across samples in order to be compared. For two-column format (mass and intensities), peaks are grouped by their m/z values. For three column data (mass, retention time, and intensities), the program will further group peaks based on their retention time. Users need to supply tolerance values in order to proceed. Here are some suggested values: mass tolerance - 0.25 (m/z); retention time - 30 (seconds) for LC-MS peak, and 5 (seconds) for GC-MS peaks. Please note, If a sample has more than one peak in a group, they will be replaced by their sum; some groups will be excluded if none of the classes has at least half its samples represented. Finally, the program create a peak intensity table in which each sample occupies a row and each column represents a peak group identified by the median values of its position (m/z and/or retention time).

Mass tolerance (m/z): 0.001
Retention time tolerance: 0.005

Submit

MS peak processing information

The uploaded files are peak lists and intensities data.
A total of 12 samples were found.
These samples contain a total of 39216 peaks,
with an average of 3268 peaks per sample
A total of 3268 peak groups were formed.
Peaks of the same group were summed if they are from one sample.
Peaks appearing in less than half of all samples in each group were ignored.

Next

Data Integrity Check:

1. Checking the class labels - at least three replicates are required in each class.
2. If the samples are paired, the pair labels must conform to the specified format.
3. The data (except class labels) must not contain non-numeric values.
4. The presence of missing values or features with constant values (i.e. all zeros)

Data processing information:

Checking data content ...passed

The uploaded files are peak lists and intensities data.

A total of 12 samples were found.

These samples contain a total of 39216 peaks,

with an average of 3268 peaks per sample

Samples are not paired.

2 groups were detected in samples.

Only English letters, numbers, underscore, hyphen and forward slash (/) are allowed.

Other special characters or punctuations (if any) will be stripped off.

All data values are numeric.

A total of 0 (0%) missing values were detected.

By default, these values will be replaced by a small value.

Click **Skip** button if you accept the default practice

Or click **Missing value imputation** to use other methods

Missing value estimation | Skip

Data Filtering:

The purpose of the data filtering is to identify and remove variables that are unlikely to be of use when modeling the data. No phenotype information are used in the filtering process, so the result can be used with any downstream analysis. This step is strongly recommended for untargeted metabolomics datasets (i.e. spectral binning data, peak lists) with large number of variables, many of them are from baseline noises. Filtering can usually improve the results. For details, please refer to the paper by Hackstadt, et al.

Non-informative variables can be characterized in three groups: 1) variables of **very small values** (close to baseline or detection limit) - these variables can be detected using mean or median; 2) variables that are **near-constant values** throughout the experiment conditions (housekeeping or homeostasis) - these variables can be detected using standard deviation (SD); or the robust estimate such as interquantile range (IQR); and 3) variables that show **low repeatability** - this can be measured using QC samples using the relative standard deviation(RSD = SD/mean). Features with high percent RSD should be removed from the subsequent analysis (the suggested threshold is 20% for LC-MS and 30% for GC-MS). For data filtering based on the first two categories, the following empirical rules are applied during data filtering:

- **Less than 250 variables**: 5% will be filtered;
- **Between 250 - 500 variables**: 10% will be filtered;
- **Between 500 - 1000 variables**: 25% will be filtered;
- **Over 1000 variables**: 40% will be filtered;

Please note, in order to reduce the computational burden to the server, the **None** option is only for less than 5000 features. Over that, if you choose None, the IQR filter will still be applied. In addition, the maximum allowed number of variables is 10000.

☐ Filtering features if their RSDs are > 25 % in QC samples

- ◉ None (less than 5000 features)
- ○ Interquantile range (IQR)
- ○ Standard deviation (SD)
- ○ Median absolute deviation (MAD)
- ○ Relative standard deviation (RSD = SD/mean)
- ○ Non-parametric relative standard deviation (MAD/median)
- ○ Mean intensity value
- ○ Median intensity value

Submit | Proceed

Data transformation

- (•) **None**
- ( ) **Log transformation** (generalized logarithm transformation or glog)
- ( ) **Cube root transformation** (takes the cube root of data values)

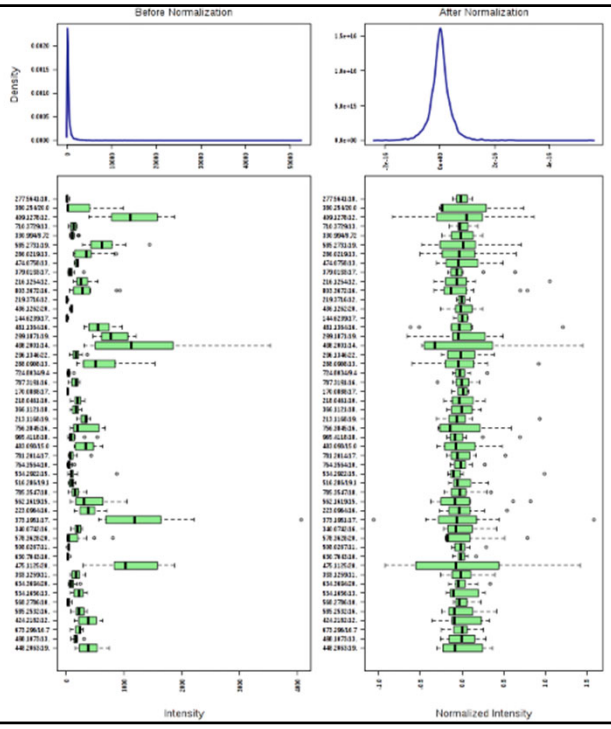
Data scaling

- ( ) **None**
- ( ) **Mean centering** (mean-centered only)
- ( ) **Auto scaling** (mean-centered and divided by the standard deviation of each variable)
- (•) **Pareto scaling** (mean-centered and divided by the square root of the standard deviation of each variable)
- ( ) **Range scaling** (mean-centered and divided by the range of each variable)

Normalize | View Result | Proceed

Mean centering

$= X_i - \text{mean}(X_{ij})$



Homework for Friday's class

- **Group 1**
  - Read and analyze a 2011 Nature paper on the discovery of trimethylamine N-oxide (TMAO) – I'll send it to you separately
  - Break it down to address (1) why the experiment was done, (2) the approach used, (3) how they identified/validated TMAO and (4) how it had a microbial origin
- **Group 2**
  - Since the publication of this paper, there have been 51 further papers on TMAO – I did a PubMed search and again I'll send it to you
  - Divide the 51 papers into 4 groups
  - Describe the significance of work in each group