

# IP4M: An integrated platform for mass spectrometry-based metabolomics data mining

V1.8

July, 2018

# Motivation and aim

- Metabolomics study depends more and more on bioinformatics tools. There is still an unmet demand for a transparent and user-friendly desktop platform to serve all the issues of computational metabolomics.
- An integrated platform for mass spectrometry-based untargeted metabolomics data analysis (IP4M) was developed which covers almost all the steps of metabolomics data mining, including peak identification and annotation, raw data and peak table preprocessing, differential analysis, correlation analysis, cluster and sub-cluster analysis, linear regression analysis, ROC analysis, pathway and enrichment analysis, Venn analysis, and sample size and power analysis.

# Development and running environments

- Programming language:
  - Java, Perl and R
- Software environment:
  - Windows 7 and above
  - Administrator privileges are required
- Hardware environment:
  - CPU > 3.0 GHz; Memory > 8 Gb
- Installation:
  - This is a green desktop software and no registration is required.

# Framework

**A** Integrated Platform for Metabolomics Research - IP4M v1.8.0

- LC-MS Preprocess
- GC-MS Preprocess
  - metaMS.runGC Peak picking using LC-MS
  - eRah Peak deconvolution and peak picking using GC-MS
- Peak Annotation by Public and Custom Databases
- Peak Table Operations
- Statistical Analysis
  - Univariate Statistical Analysis
    - Student T-test between two independent samples
    - Wilcoxon-test between two independent samples
    - Analysis of variance among multiple groups
    - Kruskal-Wallis rank sum test among multiple groups
  - Multivariate Statistical Analysis
    - Principal component analysis
    - (Orthogonal) partial least squares
    - Support vector machine
    - Random forest
    - Biosigner molecular signatures
- Pathway Analysis
- Workflows
  - GC-MS data preprocess workflow
  - LC-MS data preprocess workflow
  - Statistical analysis between two groups
  - Pathway analysis
- Other Tools
  - Merge LECO CSV files to Peak Table
  - GLM on two groups
  - Roc analysis
  - Hierarchical cluster analysis on peak table
  - Plot tree on newick tree format file
  - Plot heatmap with tree on peak table
  - Subcluster analysis on peak table

**B**

Merge tables by compound...					Tool: Retrieve rows from peak t...	Tool: For CytoScape, retrieve pa...	Tool: Plot venn diagram	Output: pwr.t.test												
<b>LC-MS preprocessing</b> <ul style="list-style-type: none"> <li>metaMS. runLC           <ul style="list-style-type: none"> <li>cdf/mzXML</li> <li>raw_pk table</li> </ul> </li> </ul>					<b>metaMS. runGC</b> <ul style="list-style-type: none"> <li>cdf/mzXML</li> <li>pk table</li> <li>mass spectra</li> <li>mass spectra 999norm</li> </ul>	<b>eRah</b> <ul style="list-style-type: none"> <li>cdf/mzXML</li> <li>pk table</li> <li>mass spectra</li> <li>mass spectra 999norm</li> </ul>	<b>Pretreatment</b> <ul style="list-style-type: none"> <li>peak table</li> <li>Outlier processed_pk table</li> <li>zerofill_pk table</li> <li>area_norm_pk table</li> <li>internal_norm_pk table</li> <li>QC norm_pk table</li> </ul>	<b>Other operations</b> <ul style="list-style-type: none"> <li>pk table</li> <li>group</li> <li>statistic_summary</li> <li>transform_pk table</li> <li>average_pk table</li> <li>transposed_matrix</li> <li>merged_matrix</li> </ul>												
					<b>GC-MS preprocessing</b>															
					<b>Peak table operations</b>															
<b>Public and custom library search</b> <ul style="list-style-type: none"> <li>pk table</li> <li>query msp</li> <li>database msp</li> <li>pk_table_anno</li> <li>details_info</li> </ul>					<b>Univariate</b> <ul style="list-style-type: none"> <li>pk table</li> <li>group</li> <li>t_result</li> <li>Wilcoxon_result</li> <li>ANOVA_result</li> <li>K-W_test</li> </ul>	<b>Multivariate</b> <ul style="list-style-type: none"> <li>pk table</li> <li>group</li> <li>sites</li> <li>importance</li> <li>rotation</li> <li>plot</li> <li>classification table</li> <li>calssificaton plot</li> </ul>	<b>Biosigner</b> <ul style="list-style-type: none"> <li>pk table</li> <li>group</li> <li>variable Metadata</li> <li>different_variable</li> <li>information</li> <li>figure_boxplot</li> <li>figure_tier</li> </ul>													
<b>Peak annotation</b>					<b>Statistical analysis</b>															
<b>KEGG pathway and enrichment</b> <ul style="list-style-type: none"> <li>Compounds list</li> <li>Compounds_ID mapping</li> <li>Kegg_enrich_result</li> <li>Compounds_idmapping</li> <li>Kegg_enrich_results</li> <li>Enrich_plot</li> </ul>					<b>Workflow</b> <ul style="list-style-type: none"> <li>LC-MS data preprocessing</li> <li>GC-MS data preprocessing</li> <li>Statistical analysis</li> <li>KEGG pathway and enrichment</li> </ul>	<b>Other tools</b> <ul style="list-style-type: none"> <li>Merge LECO CSV file</li> <li>GLM on two groups</li> <li>ROC analysis</li> <li>Hierarchical cluster</li> <li>Plot tree</li> <li>Plot heatmap with tree</li> <li>Subcluster analysis</li> <li>Correlation and distance analysis</li> <li>Plotting tools</li> <li>Sample size and power analysis</li> </ul>														
<b>Pathway and enrichment</b>					<b>Advanced functions</b>															
<b>C</b> <table border="1"> <thead> <tr> <th>Name</th> <th>Start Time</th> <th>End Time</th> <th>State</th> </tr> </thead> <tbody> <tr> <td>Task: pwr.t.test</td> <td>2018/7/26...</td> <td>2018/7/26...</td> <td>Finished</td> </tr> <tr> <td>Task: eRah</td> <td>2018/7/26...</td> <td>2018/7/26...</td> <td>Terminated</td> </tr> </tbody> </table>					Name	Start Time	End Time	State	Task: pwr.t.test	2018/7/26...	2018/7/26...	Finished	Task: eRah	2018/7/26...	2018/7/26...	Terminated	<b>D</b> <ul style="list-style-type: none"> <li>Outputs.html.files</li> <li>outputs.html</li> <li>pwr.t.test.txt</li> </ul>			
Name	Start Time	End Time	State																	
Task: pwr.t.test	2018/7/26...	2018/7/26...	Finished																	
Task: eRah	2018/7/26...	2018/7/26...	Terminated																	
					<b>Right click</b> <ul style="list-style-type: none"> <li>View log</li> <li>Rerun</li> <li>Delete</li> </ul>															

The interface consists of four parts: a) function navigation window, b) main window for parameters setting and results display, c) task window to show tasks and running status, and d) output window to list resulting files.

# Inputs and outputs

- Raw data of mzXML and NetCDF formats and other files (peak table, sample information, compound list etc.) of tab-delimited txt format are accepted inputs.
- The free tool ProteoWizard is recommended for format conversion (<http://proteowizard.sourceforge.net/>).
- All the intermediate and final results are exported as .txt (data and tables) or .pdf (figures) files.

# Modules and main functions-1

- LC-MS data preprocessing (XCMS)
- GC-MS data preprocessing (metaMS/eRah)
- Peak annotation
  - 15000+ compounds from various public libraries (HMDB, NIST, Golm, ..... )
  - support custom library
- Peak table operations
  - Outlier pretreatment
  - Missing value imputation (4 algorithms)
  - Normalization (3 algorithms)
  - Data transformation (3 algorithms)
  - Multi-table combination (for more samples, more variables, or both)
  - Basic statistics computation per sample/group/peak
- Statistical analysis
  - Parametric and nonparametric: t, paired t, Mann-Whitney, Wilcoxon, ANOVA, Kruskal-Wallis
  - Multivariate: PCA, OPLS, SVM, RF, and combined biosigner

# Modules and main functions-2

- Pathway and enrichment analysis
  - ~1600 pathways
  - 7000+ metabolite sets
- Workflows (for a quick analysis or batch applications)
  - GC-MS Raw data -> peak table (with annotation and MS information)
  - LC-MS Raw data -> peak table (with annotation)
  - Peak table -> statistical analysis results of all or selected methods
  - Compound name list (or HMDB IDs) -> pathway and enrichment analysis
- Other tools
  - Combination of peaks from multiple.csv files (replace the “SC” tool of Chromatof (LECO))
  - General linear regression model
  - HCA and Heatmap analysis
  - Sub-cluster analysis
  - Correlation and distance analysis (with interface to Cytoscape network)
  - Plotting tools (hierarchical tree, Venn, box, line, bar, scatter, plots)
  - Power and sample size analysis

# Demos

The screenshot illustrates the workflow for LC-MS data preprocessing using the `metaMS.runLC` tool within the IP4M platform.

**Tools Panel:** On the left, under the **LC-MS Preprocess** section, the `metaMS.runLC` tool is selected and highlighted with a red box. A black arrow points from this selection to the main tool interface.

**Tool Interface:** The main window shows the `metaMS.runLC` tool details. It includes a note: "Please select multiple cdf/mzXML files" with a **Browse...** button, and a warning: "Some input files are required netcdf/mzXML format." Below this, it specifies "Settings needed for functions runLC – Conditional Param" set to "RP - reverse-phase chromatography".

**Tasks Table:** A table titled "Tasks" lists the following tasks:

Task Name	Start Time	End Time	State
Task: Internal standard normal...	2018/7/...	2018/7/...	Finished
Task: Internal standard normal...	2018/7/...	2018/7/...	Failed
Task: Total area normalization	2018/7/...	2018/7/...	Finished
Task: Zero Filling	2018/7/...	2018/7/...	Finished
Task: Zero Filling	2018/7/...	2018/7/...	Finished
Task: Zero Filling	2018/7/...	2018/7/...	Finished
Task: Zero Filling	2018/7/...	2018/7/...	Finished

A black arrow points from the bottom of the Tasks table to the **Files** panel.

**Files Panel:** The **Files** panel displays the generated output files: `outputs.html.files`, `detailed_information.txt`, and `identified_pkTable.txt`.

Workflow of usage

Integrated Platform for Metabolomics Research - IP4M v1.8.0

Tools

type filter text

- > LC-MS Preprocess
- > GC-MS Preprocess
- > Peak Annotation by Public and Custom Libraries
- > Peak Table Operations
- > Statistical Analysis
  - Univariate Statistical Analysis
    - Student T-test between two independent or pair
    - Wilcoxon-test between two independent or pair
    - Analysis of variance among more than two group
    - Kruskal-Wallis rank sum test among more than
  - Multivariate Statistical Analysis
    - Principal component analysis
    - (Orthogonal) partial least squares discriminant
    - Support vector machine
    - Random forest
    - Biosigner molecular signature discovery with P
- > Pathway Analysis
- > Workflows
- > Other Tools

Tool: Biosigner molecular signature discovery Log: Biosigner molecular signature discovery

**Task Command Line:**

```
D:\GCMPA\IP4M_v1.8.5_final_version/bin/Perl/bin/perl "D:\GCMPA\IP4M_v1.8.5_final_version/tools/diff/biosigner.pl" -input1 "C:\Users\liangdandan\IP4M_Outputs\metaMS.runLC_2018.7.12_14.49.29\cms_raw_pkTable.txt" -input2 "C:\Users\liangdandan\Desktop\c_ko_groups.txt" -methodVc "all" -bootl "50" -pvalN "0.05" -perm1 "1" -tierC "S" -html "C:\Users\liangdandan\IP4M_Outputs\Biosigner_molecular_signature_discovery_2018.7.23_12.33.58\outputs.html" -output1 "C:\Users\liangdandan\IP4M_Outputs\Biosigner_molecular_signature_discovery_2018.7.23_12.33.58\outputs.html"
```

**Task Standard Output:**

```
C:\Users\liangdandan\ip4m\tmp\709d7645-a386-40cc-9f5b-1c3090566a87> D:\GCMPA\IP4M_v1.8.5_final_version/bin/Perl/bin/perl "D:\GCMPA\IP4M_v1.8.5_final_version/tools/diff/biosigner.pl" -input1 "C:\Users\liangdandan\IP4M_Outputs\metaMS.runLC_2018.7.12_14.49.29\cms_raw_pkTable.txt" -input2 "C:\Users\liangdandan\Desktop\c_ko_groups.txt" -methodVc "all" -bootl "50" -pvalN "0.05" -perm1 "1" -tierC "S" -html "C:\Users\liangdandan\IP4M_Outputs\Biosigner_molecular_signature_discovery_2018.7.23_12.33.58\outputs.html" -output1 "C:\Users\liangdandan\IP4M_Outputs\Biosigner_molecular_signature_discovery_2018.7.23_12.33.58\variable_results.txt" -output2 "C:\Users\liangdandan\IP4M_Outputs\Biosigner_molecular_signature_discovery_2018.7.23_12.33.58\variable_significant_results.txt" -output3 "C:\Users\liangdandan\IP4M_Outputs\Biosigner_molecular_signature_discovery_2018.7.23_12.33.58\biosigner_summary.txt" -output4 "C:\Users\liangdandan\IP4M_Outputs\Biosigner_molecular_signature_discovery_2018.7.23_12.33.58\figure-tier.pdf" -output5 "C:\Users\liangdandan\IP4M_Outputs\Biosigner_molecular_signature_discovery_2018.7.23_12.33.58\figure-boxplot.pdf"
```

**Task Standard Error Output:**

Warning message:  
 In read.table("\_group\_", header = F, sep = "\t", check.names = FALSE, :  
 incomplete final line found by readTableHeader on '\_group\_'  
 Error in apply(data[, group1], 1, mean) : dim(X)的值必需是正数  
 停止执行  
 Error, died with 256 at D:\GCMPA\IP4M\_v1.8.5\_final\_version/tools/diff/biosigner.pl line 128.

Tasks

Task Name	Start Time	End Time	State
Task: Zero Filling	2018/7/...	2018/7/...	Finished
Task: Outlier Processing	2018/7/...	2018/7/...	Finished
Task: metaMS.runLC	2018/7/...	2018/7/...	Finished
Task: Outlier Processing	2018/7/...	2018/7/...	Finished
Task: Outlier Processing	2018/7/...	2018/7/...	Failed
Task: Biosigner molecular sign...	2018/7/...	2018/7/...	Failed
Task Plot pathway enrichment	2018/7/...	2018/7/...	Finished

Files

View logs, error messages, and R codes of specific task

Integrated Platform for Metabolomics Research - IP4M v1.8.0

Tools

type filter text

- LC-MS Preprocess
  - metaMS.runLC** LC-MS data preprocess using meta
- GC-MS Preprocess
- Peak Annotation by Public and Custom Libraries
- Peak Table Operations
- Statistical Analysis
- Pathway Analysis
- Workflows
- Other Tools

Tool: metaMS.runLC   Output: metaMS.runLC   Output: metaMS.runGC

Output Files:

- [gcms raw\\_pkTable.txt](#)
- [gcms mass spectra.msp](#)
- [gcms mass spectra\\_999norm.msp](#)

Name	Run1L	Run1R	Run1S
Unknown 1	985470112	866679388	1045892292
Unknown 11	94499098	78043352	168482033
Unknown 30	9892283	0	15867485
Unknown 4	332262984	314474394	347787267
Unknown 12	81057561	0	82819384
Unknown 46	0	229043993	403749803
Unknown 13	78381121	0	135909286
Unknown 36	2527342	0	3707291
Unknown 24	22362512	0	28676735
Unknown 9	124028621	105133303	197074532
Unknown 25	18664219	35825326	14111533
Unknown 29	10024783	9197096	12461662

Name: Unknown 1  
DB. idx: -1  
rt: 3.024  
Class: Unknown  
rt. sd: 0.0074  
Num Peaks: 137  
33 0.19; 40 0.45; 41 0.88; 42 3.32; 43 19.34; 44 10.76; 45 44.70; 46 6.51; 47 8.83; 48 1.74; 49 8.12; 50 1.33; 51 0.89; 53 1.15; 54 1.01; 55 2.93; 56 2.86; 57 3.78; 58 6.59; 59 15.86; 60 1.70; 61 1.37; 62 0.91; 63 4.77; 64 0.56; 65 0.77; 66 0.60; 67 0.48; 68 0.54; 69 6.18; 70 8.30; 71 2.86; 72 20.17; 73 313.11; 74 36.50; 75 17.90; 76 4.98; 77 37.95; 78 3.40; 79 2.63; 81 6.21; 84 2.05; 85 2.10; 86 3.53; 87 2.49; 88 3.51; 89 18.20; 90 2.87; 91 8.06; 92 1.81; 93 1.19; 95 2.22; 96 0.82; 99 6.57; 100 20.85; 101 4.04; 102 5.24; 104 4.82; 105 2.95; 107 1.82; 109 0.39; 110 0.69; 112 0.32; 113 0.93; 118 0.67; 121 0.57; 122 1.43; 123 0.54; 126 0.45; 128 0.33; 130 0.45; 131 3.41; 132 0.99; 134 2.10; 135 0.86; 136 0.39; 137 0.33; 138 1.16; 139 0.32; 140 0.28; 141 0.33; 142 4.42; 147 10.40; 148 1.84; 149 1.10; 150 0.68; 151 1.61; 152 0.37; 153 0.29; 154 0.17; 155 0.17; 156 0.17; 158 0.18; 160 0.20; 162 0.16; 164 0.05; 165 0.05; 166 0.03; 167 0.03; 168 0.37; 169 0.89; 172 0.41; 173 0.25; 174 0.35; 176 1.78; 177 0.38; 178 0.42; 179 0.14; 180 0.21; 184 0.27; 187 0.35; 188 80.33; 189 15.32; 190 7.25; 191 2.63; 192 70.30; 193 12.95; 194 6.41; 195 5.12; 196 1.52; 197 0.88; 199 6.55; 200 1.15; 201 0.31; 214 4.54; 215 0.61; 216 0.20; 238 0.79; 239 0.14; 242 12.41; 243 2.76; 244 1.10; 245 0.12; 257 9.69; 258 2.07; 259 0.84; 260 0.06;

Tasks

Task Name

Task: GC-MS peak annotation

Task: metaMS.runGC

Name: Unknown 2

2010/7/1 2010/7/1 Edited

## Outputs of LC-MS preprocessing (peak identification)

Tools

type filter text

- LC-MS Preprocess
  - [metaMS.runLC](#) LC-MS data preprocess using meta
- GC-MS Preprocess
  - [metaMS.runGC](#) Peak picking using metaMS packa
  - [eRah](#) Peak deconvolution and peak picking using e
- Peak Annotation by Public and Custom Libraries
  - [GC-MS peak annotation](#) on msp database file
  - [LC-MS peak annotation](#) on Tab-delimited txt data
- Peak Table Operations
- Statistical Analysis
- Pathway Analysis
- Workflows
- Other Tools

Query	Query_mz	Best_hit	Best_hits_mz	Other_hits	Other_hits_mz			
1	480.0832663	Sorafenib N-oxide	480.081217342	Neoglucobassicin	478.071586314	(6-carboxy-3, 4, 5-trihydroxyoxan-2-yl)[2-(3, 4-dihydroxyphenyl)-3, 5, 6-trihydroxy- <i>H</i> -chromen-7-ylidene]oxidanium	479.082017092	(6-carboxy-3, 4, 5-trihydroxy-2-yl)[3, 5-dihydroxy-2-(3, 4, 5-trihydroxyphenyl)-chromen-7-ylidene]oxide
2	No Hit							
3	158.9643996	12, 13-epoxy-9-alkoxy-10E-octadecenoate	156.952712	Methyl 2-propenyl selenide	135.979122084	Methylselenopyruvate	181.948215886	Dimethylidithiophosphate
4	No Hit							
5	197.9399916	3-Bromosulfolane	197.935012803					
6	216.0677236	1-Isothiocyanato-8-(methylthio)octane	217.095890993	Kinetin	215.080709935	2-[(3, 5-dimethoxyphenyl)(hydroxy)methylidene]amino acetic acid	239.079372523	N-Acetylserotonin
7	214.0902781	2-(4-Methyl-5-thiazolyl)ethyl butanoate	213.082349419	2-(4-Methyl-5-thiazolyl)ethyl isobutyrate	213.082349419	Acetyl vitamin K6	215.094628665	Macaridine
8	No Hit							
9	No Hit							
10	186.9569229	Rhenium	186.955750787	1, 3, 5-Trichloro-2-methoxybenzene	209.940597903	Methoxyflurane	163.960726574	
11	No Hit							
12	118.944347	Methyl methylthio selenide	141.93554271	Sulfate	95.951729178	Hydrogen phosphate	95.961245032	Methaneselenol
13	135.9459395	1, 3, 5-Trithiane	137.963162262	1, 2, 3-Trithiane	137.963162262	3H-1, 2-Dithiole-3-thione	133.931862134	Dimethylarsinic acid
14	215.0932357	Atrazine	215.09377318	3-(3, 4, 5-trimethoxyphenyl)prop-2-enic acid	238.084123551	3-(6, 7-dimethoxy-2H-1, 3-benzodioxol-5-yl)propanal	238.084123551	2H-1, 3-benzodioxol-5-yl; en-1-ol
15	128.9513485	2, 2-dichloro-1, 1-ethanediol	129.9588348	Methylselenic acid	127.9376512			
16	331.0162247	Furosemide	330.007719869	Pyrroloquinoline quinone	330.012415178	cis-Resveratrol 3-sulfate	308.035458806	cis-Resveratrol 4'-sul

## Output Files:

- [identified\\_pkTable.txt](#)
- [identified\\_uniq\\_pkTable.txt](#)
- [detailed\\_information.txt](#)

ID	SH07597A	SH07608B
Sorafenib N-oxide	1814. 270491	914. 2119107
2	1894. 813633	689. 8584025
12, 13-epoxy-9-alkoxy-10E-octadecenoate	45163. 21917	21898. 24251
4	7437. 592726	2586. 034908
3-Bromosulfolane	1709. 85541	1033. 111766
1-Isothiocyanato-8-(methylthio)octane	5273. 157932	3444. 396109
2-(4-Methyl-5-thiazolyl)ethyl butanoate	127267. 4435	92000. 93498
8	1339. 633605	607. 8319578
9	18481. 78677	9878. 253492
Rhenium	25611. 5955	13053. 88925
11	3786. 077065	2811. 874155
Methyl methylthio selenide	2927. 077413	1981. 117768
1, 3, 5-Trithiane	26293. 78656	15871. 88666
Atrazine	12273. 09315	8659. 789807
2, 2-dichloro-1, 1-ethanediol	53137. 55769	37501. 16727
Furosemide	8548. 496021	4205. 645366
1, 3-Dichloropropene	5170. 811548	3778. 113902
Ammonium peroxydisulfate	11685. 96462	5839. 060418
19	17961. 18956	11415. 48618
20	2721. 396709	1675. 759201
2-Methyl-2-cyclopenten-1-one	1421. 038453	732. 6912366

Files
> outputs.html.files
detailed_information.txt
identified_pkTable.txt

## Outputs of GC-MS peak annotation

	Glycine-d5_1	N2_1	N3_1	QC1_1
nbr. val	102	102	102	102
nbr. null	0	0	0	0
nbr. na	0	0	0	0
min	3.85417760431785e-12	0.00893943757733268	0.057258514677102	8.31486109559897e-14
max	329.680542925297	112.89922337076	125.740860123838	252.160556248509
range	329.680542925293	112.890283933183	125.683601609161	252.160556248509
sum	1000	1000	1000	1000
median	3.85417760431785e-12	1.10299535500165	1.36297406812425	1.08348644707193
mean	9.80392156862746	9.80392156862746	9.80392156862745	9.80392156862745
SE. mean	3.78909734127497	2.04561276241289	2.09451749617876	2.90974935184046
CI. mean. 0.95	7.51654986910382	4.05794545684011	4.15495929340276	5.77216000007589
var	1464.44038348902	426.82220522143	447.474361263493	863.597411634667
std. dev	38.2680073101412	20.6596761959655	21.1535897961432	29.3870279483085
coef. var	3.9033367456344	2.10728697198848	2.15766615920661	2.99747685072747
skewness	6.43059173977852	2.8854557982957	3.25217946337447	6.08756886871145
skew. 2SE	13.4492419508164	6.03477793958065	6.80176105720342	12.7318277882741
kurtosis	47.2488232127589	8.28380925870972	11.2179513653653	43.9662007414141
kurt. 2SE	49.858367667054	8.74		
normtest. W	0.273608347703713	0.52		
normtest. p	3.47635486569766e-20	1.2623		

## Basic statistics computation

Integrated Platform for Metabolomics Research - IP4M v1.8.0

Tools □ Tool: Student T-test □ Output: Student T-test

type filter text

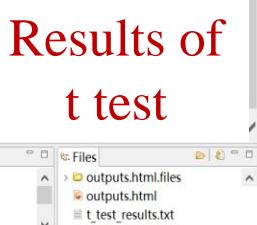
- LC-MS Preprocess
- ESI-MS Preprocess
- Peak Annotation by Public and
- Peak Table Operations
- Statistical Analysis
- Univariate Statistical Analysis
- Student T-test between tw
- Wilcoxon-test between tw
- Analysis of variance amon
- Kruskal-Wallis rank sum t

Output Files:

- t test results.txt
- t test significant results.txt

mean (group2)	variance (group2)	stderr (group2)	mean (group1)	variance (group1)	stderr (group1)	logFC (group/group2)	p	fd
Cholesterol	196458.350000019	77191766569.4302	196458.349999981	13826312.5	1161277520.5	240964.5	1.3710914099405	0.000519956836358320
141	3401.216000018	2316548.6570563	3401.2159999912	2780981.4078765498	142807.5	42807.5	4.74830.5	4.74830.5
19	10695.65	705851.64500001	1878.65	149340.5	918010.5	2023.5	0.48456691187242	0.04416679371073
2-Chloroethyl limoleate	241588.150000019	23323919682.819	341588.149999981	6645218.3	31061691380	394130	4.2821175315193	0.00671278151693
Fluoranthene	334406.450000019	22632672539.58	334406.449999981	3856357.0	105361615058	22923	3.4428199575296	3.4428199575296
1H-Indene, 2,3-dihydro-5-methyl-	4926158.273	48509704000.2	4924921.273	4747093.4018	1754883	2.05780.4	2.05780.4	2.05780.4
1,3-Dioxolane, 4-methyl-2-pentyl-	5815939	1513689334472	2599398	30382234	1230701070352	2480626	2.38514391197887	2.38514391197887
Difenoxin	450396.285	2914829545147.66	381761.05	3089118.5	137326.5	137326.5	4.7474576.5	4.7474576.5
1-Methylchloroethanol	136014.000000019	36699616391.9986	136013.999999981	1375948.5	429769302.5	436859.5	3.7715828161692	3.7715828161692
Norgestrel	2657.610000018	1412880.084881	2657.6119999912	242695	297220500	38550	3.3427370.161023	3.3427370.161023
Phenobarbital	64607.600000019	6193945718.519	64607.6199999912	519453	478926450	49393	6.51287005349628	6.51287005349628
5	807610.4	0	0	2735322.58002	12762.7	12762.7	-1.352468312143	-1.352468312143
117	104482.000000019	2183102026.5721	104482.000000019	3653167.3	108736120499584	300715.5	3.027361120499584	3.027361120499584
Iophytol	78604.850000018	12357444887.0391	78604.849999981	47840.5	47764578.5	19334.5	2.6526939148127	2.6526939148127
20	20430.500000018	3355095290.946	20430.499999981	63761	3024862	3889	2.386715604479	2.386715604479
1(20)-Phthalazine, hydrazone	656842.000000019	17239811523.150	656842.000000019	1723981.5	4019812.5	4019812.5	0.026345273375	0.026345273375
Benzonitrile	3581641.000000019	2356831166904.2	3581641.000000019	371191790.5	431268491202	1409449	2.7111162033901	2.7111162033901
108	8811.860000018	155297733.318537	8811.859999981	343805	3393371288	130262	5.6366738672548	5.6366738672548
Morpholine, 4-(3'-butoxyphenoxy)propyl-	12715.535	3346255829.0724	12715.535	4858010	613610200	5539	2.1401462562072	2.1401462562072
Hexanedioic acid, diethyl ester	65524.000000018	6558641571.3707	65524.000000018	65575.5	42186119980.5	142324.5	0.08842030823502	0.08842030823502

mean (group2)	variance (group2)	stderr (group2)	mean (group1)	variance (group1)	stderr (group1)	logFC (group/group2)	p	fd
Cholesterol	196458.350000019	77191766569.4302	196458.349999981	13826312.5	1161277520.5	240964.5	1.3710914099405	0.000519956836358320
141	3401.216000018	2316548.6570563	3401.2159999912	2780981.4078765498	142807.5	42807.5	4.74830.5	4.74830.5
19	10695.65	705851.64500001	1878.65	149340.5	918010.5	2023.5	0.48456691187242	0.04416679371073
2-Chloroethyl limoleate	241588.150000019	23323919682.819	341588.149999981	6645218.3	31061691380	394130	4.2821175315193	0.00671278151693
Fluoranthene	334406.450000019	22632672539.58	334406.449999981	3856357.0	105361615058	22923	3.4428199575296	3.4428199575296
1H-Indene, 2,3-dihydro-5-methyl-	4926158.273	48509704000.2	4924921.273	4747093.4018	1754883	2.05780.4	2.05780.4	2.05780.4
1,3-Dioxolane, 4-methyl-2-pentyl-	5815939	1513689334472	2599398	30382234	1230701070352	2480626	2.38514391197887	2.38514391197887
Difenoxin	450396.285	2914829545147.66	381761.05	3089118.5	137326.5	137326.5	4.7474576.5	4.7474576.5
1-Methylchloroethanol	136014.000000019	36699616391.9986	136013.999999981	1375948.5	429769302.5	436859.5	3.7715828161692	3.7715828161692
Norgestrel	2657.610000018	1412880.084881	2657.6119999912	242695	297220500	38550	6.51287005349628	6.51287005349628
Phenobarbital	64607.600000019	6193945718.519	64607.6199999912	519453	478926450	49393	3.0260030327168	3.0260030327168
5	807610.4	0	0	2735322.58002	12762.7	12762.7	0.027361120499584	0.027361120499584
117	104482.000000019	2183102026.5721	104482.000000019	3653167.3	108736120499584	300715.5	3.027361120499584	3.027361120499584
Iophytol	78604.850000018	12357444887.0391	78604.849999981	47840.5	47764578.5	19334.5	2.6526939148127	2.6526939148127
20	20430.500000018	3355095290.946	20430.499999981	63761	3024862	3889	2.386715604479	2.386715604479
1(20)-Phthalazine, hydrazone	656842.000000019	17239811523.150	656842.000000019	1723981.5	4019812.5	4019812.5	0.020735780526306	0.020735780526306
108	3581641.000000019	2356831166904.2	3581641.000000019	371191790.5	2.69682653909602	3.0884203073352	3.7730981273352	3.7730981273352
130	8811.860000018	155297733.318537	8811.859999981	343805	3393371288	130262	0.38812432639074	0.38812432639074
Morpholine, 4-(3'-butoxyphenoxy)propyl-	12715.535	3346255829.0724	12715.535	4858010	613610200	5539	2.1401462562072	2.1401462562072
Hexanedioic acid, diethyl ester	65524.000000018	6558641571.3707	65524.000000018	65575.5	42186119980.5	142324.5	0.08842030823502	0.08842030823502
94	40064.999000018	321613648.29704	40064.999999981	157926	168074450	9055	1.9781149011402	1.9781149011402
118	191777.400000018	73557142301.5067	191777.399999981	1393482.5	20461579640	3844965.5	2.8611962035375	2.8611962035375
125	1469965.00000002	444593326449.8	1469964.999999981	8013769.5	671219271112.5	1833292.5	2.42623462299429	2.42623462299429



# Outputs of multivariate machine learning analysis

**OPLS**

**SVM**

**RF**

**Biosinger**

Integrated Platform for Metabolomics Research - IP4M v1.8.0

Output: Random forest   Output: Biosigner mole...   Log: Biosigner molecu...   Output: Biosigner molecu...   Output: Biosigner molecu...   Output: Compounds ID m...

type filter text

LC-MS Preprocess   GC-MS Preprocess   Peak Annotation by Pub   Peak Table Operations   Statistical Analysis   Univariate Statistical A   Student T-test betw...   Wilcoxon-test betw...   Analysis of variance   Kruskal-Wallis rank   Multivariate Statistical   Principal compone...   (Orthogonal) partic...   Support vector ma...   Random forest   Biosigner molecule...   Pathway Analysis   Workflows   Other Tools

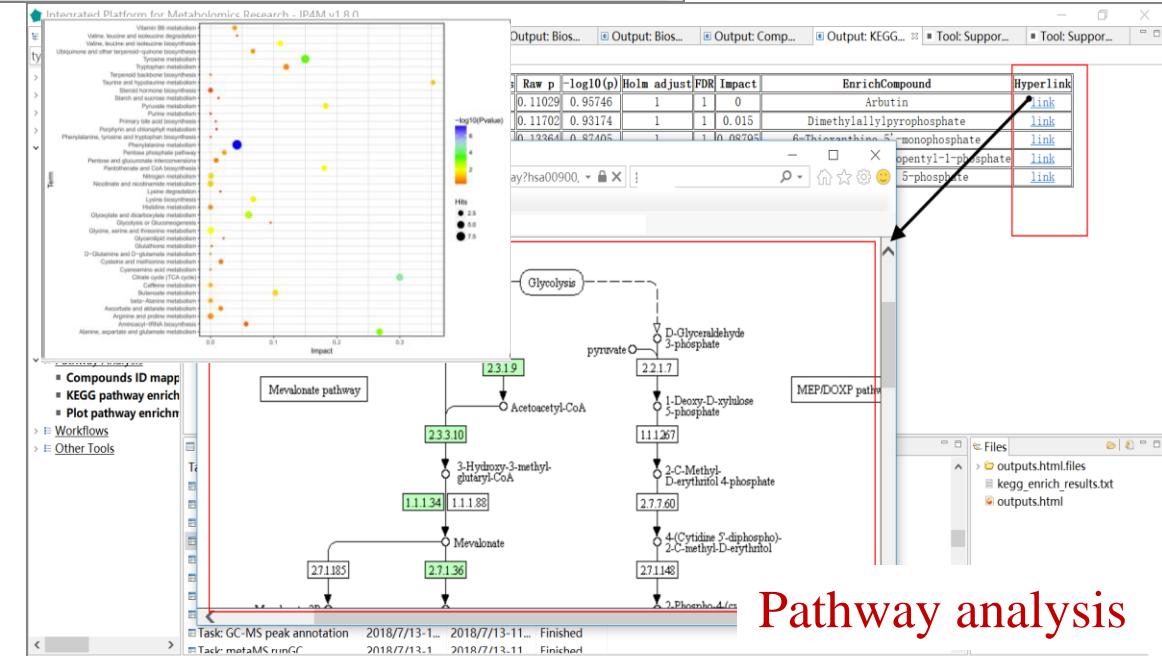
Query   Match   HMDB   PubChem   KEGG   chemical\_formula   average\_molecular\_weight   super\_class   pathways

Query	Match	HMDB	PubChem	KEGG	chemical_formula	average_molecular_weight	super_class	pathways
3-hydroxy-2-(hydroxymethyl)-2-[(sulfoxy)methyl]propanoic acid	NA	NA	NA	NA	NA	NA	NA	NA
Bis(2-furanylmethyl)disulfide	Bis(2-furanylmethyl)disulfide	HMDB29988	20499		C10H10O2S2	226.315	Organic compounds	NA
5,7-Dimethoxyisoflavone	5,7-Dimethoxyisoflavone	HMDB30698	6710704		C17H14O4	282.2907	Phenylpropanoids and polyketides	NA
L-Glutamic acid 5-phosphate	L-Glutamic acid 5-phosphate	HMDB01228	193475	C03287	C5H10N07P	227.1092	Organic compounds	NA
Octafluorocyclobutane	Octafluorocyclobutane	HMDB31292	8263		C4F8	200.03	Organic compounds	NA
6-Thioxanthine 5'-monophosphate	6-Thioxanthine 5'-monophosphate	HMDB60418	3081384	C16618	C10H13N4O8PS	380.271	Organic compounds	NA
Arbutin	Arbutin	HMDB29943	346	C06186	C12H16O7	272.2512	Organic compounds	NA
Isopropyl beta-D-glucoside	Isopropyl beta-D-glucoside	HMDB32705	15613266		C9H18O6	222.2356	Organic compounds	NA
Phenytoin	Phenytoin	HMDB14397	1775	C07443	C15H12N2O2	252.268	Organoheterocyclic compounds	Fosphenytoin (Antiarrhythmic Pathway)   Fosphenytoin (Antiar Pathway)   Phenytoin (Antiarr Pathway)
N-lactoyl-Phenylalanine	NA	NA	NA	NA	NA	NA	NA	NA
L-prolyl-L-proline	NA	NA	NA	NA	NA	NA	NA	NA
1-Methoxypyrene	NA	NA	NA	NA	NA	NA	NA	NA
Dimethylallylpyrophosphate	Dimethylallylpyrophosphate	HMDB01120	647	C00235	C5H12O7P2	246.0921	Organic compounds	Alendronate pathway   Atorvastatin Pathway   Cerivastatin Pathway   Cholesteratin ester disease   Chondrodysplasia Punctata   Dominant (CDPX) Desmosterolemiosis   Fluvastatin Pathway   Hyper-IgD syndrome   Hypercholesterolemia   Lovastatin Pathway   Mevalonate pathway

## Compound ID mapping

Time   State  
3/7/23-14...   Finished

Files  
outputs.html.files  
compounds\_idmapping.txt  
outputs.html



Integrated Platform for Metabolomics Research - IP4M v1.8.0

Tools

- LC-MS Preprocess
- GC-MS Preprocess
- Peak Annotation by Public and Custom
- Peak Table Operations
- Statistical Analysis
- Pathway Analysis
- Workflows
  - GC-MS data preprocess workflow
  - LC-MS data preprocess workflow
  - Statistical analysis between two groups
  - Pathway analysis
- Other Tools

Tool: LC-MS data preprocess workflow Output: LC-MS data preprocess workflow

**Output Files:**

- 1. lcms raw pkTable.txt
- 2. identified\_pkTable.txt
- 2. identified\_uniq\_pkTable.txt
- 2. detailed\_information.txt
- 3. outliers\_processed\_pkTable.txt
- 4. zero\_filled\_pkTable.txt
- 5. total\_area\_norm\_pkTable.txt
- 6. log2\_transformed\_pkTable.txt

Tasks

Task Name	Start Time	End Time	State
Task: LC-MS data preprocess ...	2018/7/24...	2018/7/24...	Finished
Task: LC-MS data preprocess ...	2018/7/24...	2018/7/24...	Finished
Task: GC-MS data preprocess ...	2018/7/24...	2018/7/24...	Finished
Task: Biosigner molecular sign... 2018/7/24...	2018/7/24...	2018/7/24...	Finished

ko15      ko16

Oxybuprocaine	2. 92396425012336e-005	0. 0450543728218289
Rosoxacin	0. 109335983283099	0. 0454074223891323
1, 2, 4, 5, 7-Pentathiocane	2. 92396425012336e-005	0. 02180796719104
1-Stearoylglycerophosphoglycerol	0. 284237991428169	0. 928704683641863
Bikojic acid	0. 00643794035994368	0. 266170748105035
Furazolidone	0. 0471102240063989	0. 050740476479346
Tetracosatetraenoyl carnitine	0. 0201552869045265	0. 819580982653954
Deserpidine	0. 291188603563351	3. 9990675487824
Thiamylal	0. 00625212355835191	0. 0996460776609158
LC352	0. 00222151045122858	0. 137465097063102
Citicoline	0. 631904789552198	0. 039145856971904
LysoPE (0:0/22:5(4Z, 7Z, 10Z, 13Z, 16Z))	0. 776451709495551	2. 43571330069476e-005
Divanillyltetrahydrofuran ferulate	0. 0111626635195886	0. 054192345895557
Ganosporeic acid A		
Dynorphin B (10-13)		
1, 2, 3-Tris(1-ethoxyethoxy)propane		
Muricin E		

## Outputs of LC-MS preprocessing workflow

Integrated Platform for Metabolomics Research - IP4M v1.8.0

Tools

- LC-MS Preprocess
- GC-MS Preprocess
- Peak Annotation by Public and Custom
- Peak Table Operations
- Statistical Analysis
- Pathway Analysis
- Workflows
  - GC-MS data preprocess workflow
  - LC-MS data preprocess workflow
  - Statistical analysis between two groups
  - Pathway analysis
- Other Tools

Tool: Statistical analysis Help: Statistical analysis

**Output Files:**

- 1. pkTable\_summary.txt
- 2. t-test\_results.txt
- 2. t-test\_significant\_results.txt
- 3. wilcoxon\_test\_results.txt
- 3. wilcoxon\_test\_significant\_results.txt
- 4. aov\_results.txt
- 4. aov\_significant\_results.txt
- 5. kruskal\_wallace\_results.txt
- 5. kruskal\_wallace\_significant\_results.txt
- 6. pca\_scores.txt
- 6. pca\_importance.txt
- 6. pca\_rotation.txt
- 6. pca\_plot.pdf
- 7. colpsda\_variable\_results.txt
- 7. colpsda\_variable\_significant\_results.txt
- 7. colpsda\_samples\_results.txt
- 7. colpsda\_prediction\_summary.txt
- 7. colpsda\_figure.pdf
- 8. svm\_summary.txt
- 8. svm\_variable\_results.txt
- 8. svm\_predictions\_results.txt
- 8. support\_vectors.txt
- 8. svm\_plot.pdf
- 9. rf\_summary.txt
- 9. rf\_variable\_results.txt
- 9. rf\_samples\_results.txt
- 9. rf\_prediction\_summary.txt
- 9. rf\_error\_rates.pdf
- 9. rf\_predictions\_margin\_plot.xlsx
- 10. biosigner\_summary.txt
- 10. biosigner\_variable\_significant\_results.txt

Model overview

SVM classification plot

The Margin of Predictions Plot

Significant features from 'S' groups:

- plida randomforest
- MC2
- MC10
- MC46
- MC1
- MC2
- MC4
- MC5
- MC6
- MC16
- MC13
- MC21
- MC22
- MC27
- MC31
- MC34
- MC40
- MC41
- MC42
- MC43
- MC44
- MC45
- MC48

Accuracy:

plida randomforest	svm
Full: 0.926	0.999, 0.957
AS: 0.765	0.989, 0.982
S: N/A	0.976, 0.982

Time End Time State

7/24... 2018/7/24... Finished

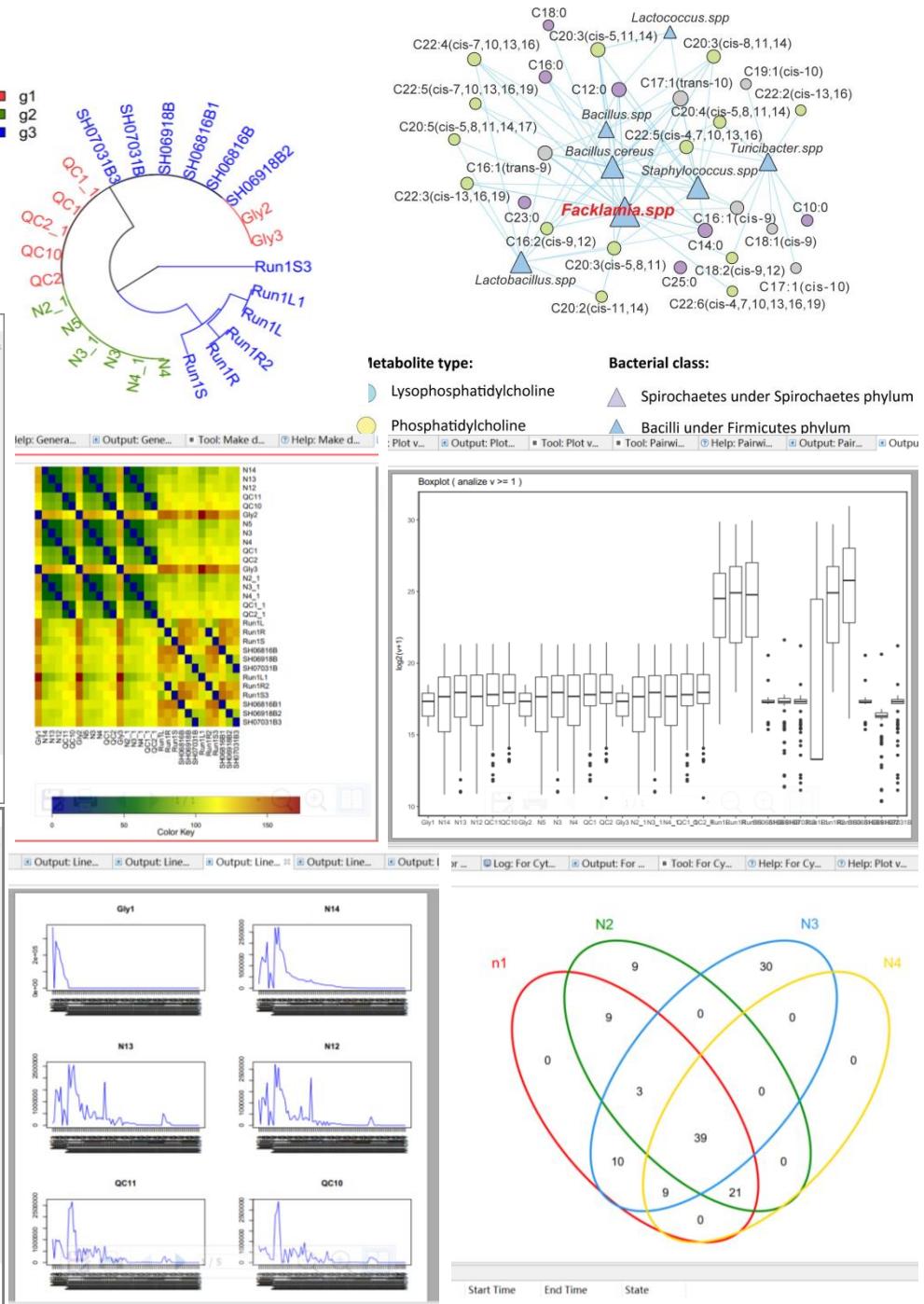
Files

outputs.html.files

1\_pkTable\_summary.txt

## Outputs of statistical analysis workflow

# Results of “Other tools”



- **Conclusions:** Compared with other multi-function platforms, the strengths of IP4M are the GC-MS peak identification, many simple but useful tools, and rich knowledge base. However, it is limited in integration with other omics data. IP4M can be further extended to an online platform and NMR data preprocessing module is warranted to be incorporated. Nevertheless, it is still an attractive alternative to existing platforms.
- **Availability:** IP4M is freely available at  
<https://github.com/IP4M>
- **Contact:** [wjia@cc.hawaii.edu](mailto:wjia@cc.hawaii.edu); [chentianlu@sjtu.edu.cn](mailto:chentianlu@sjtu.edu.cn)