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

V1.8

July, 2018

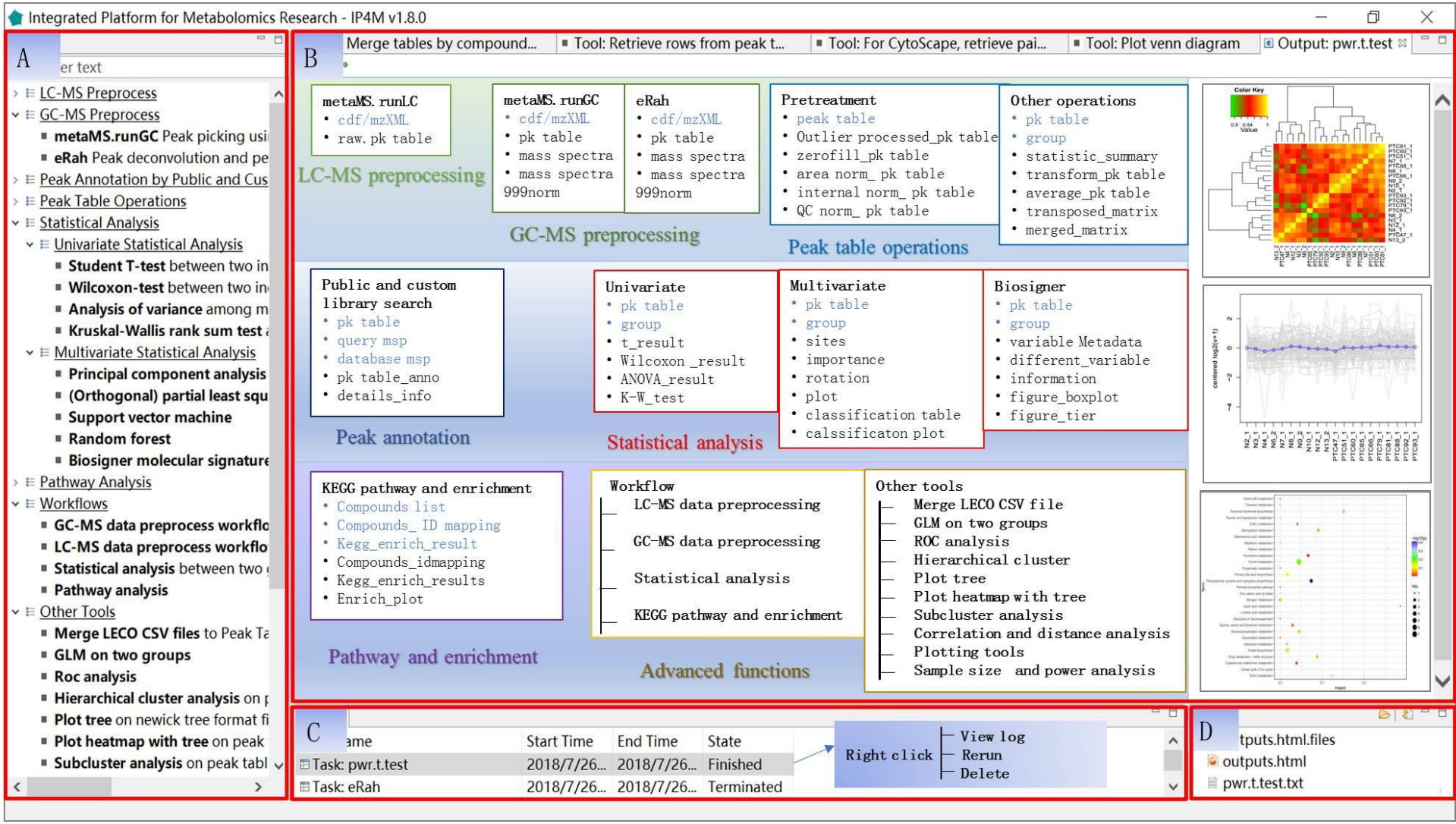
Motivation and aim

- 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



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.

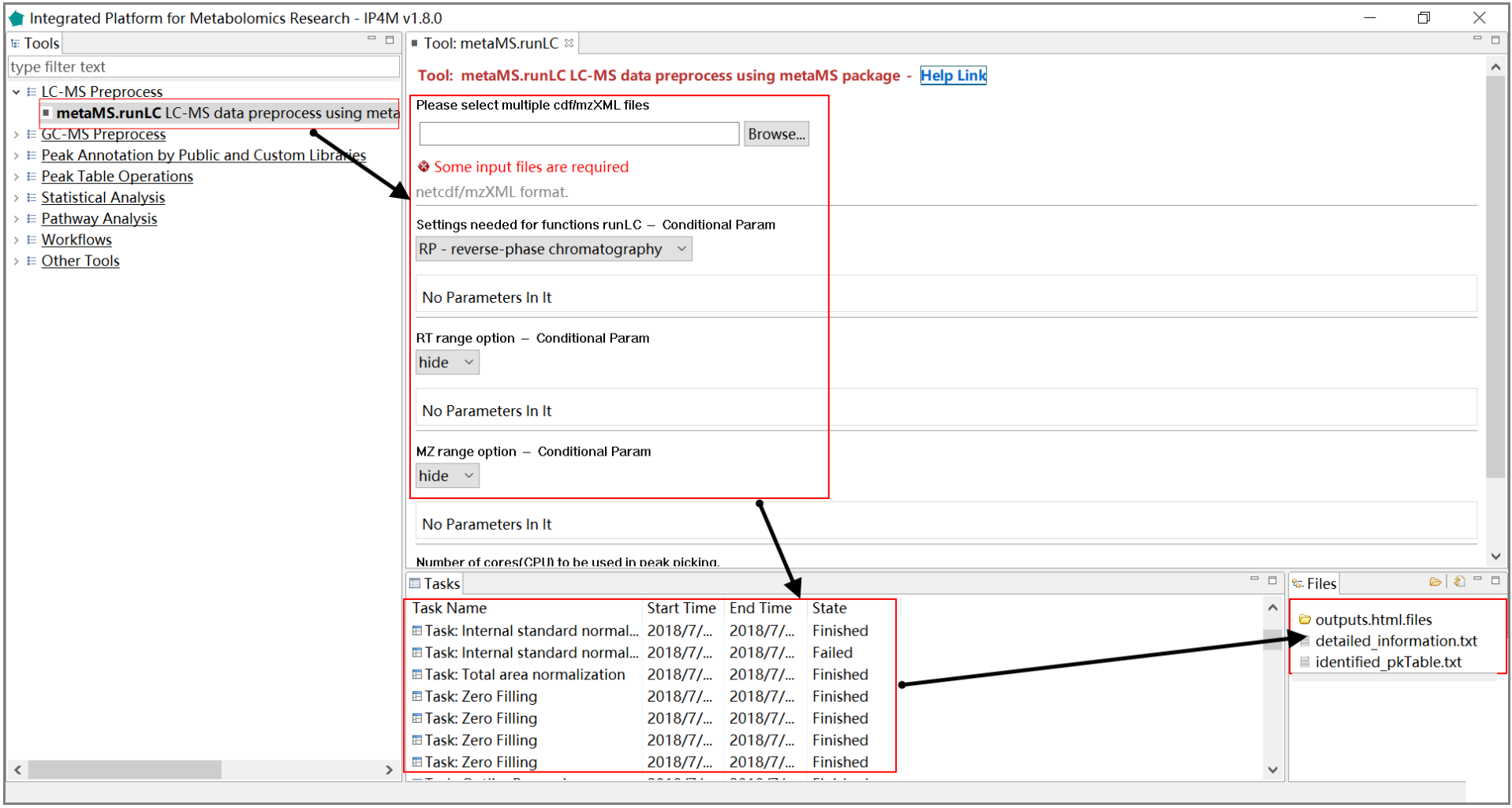
Modules and main functions-1

- 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

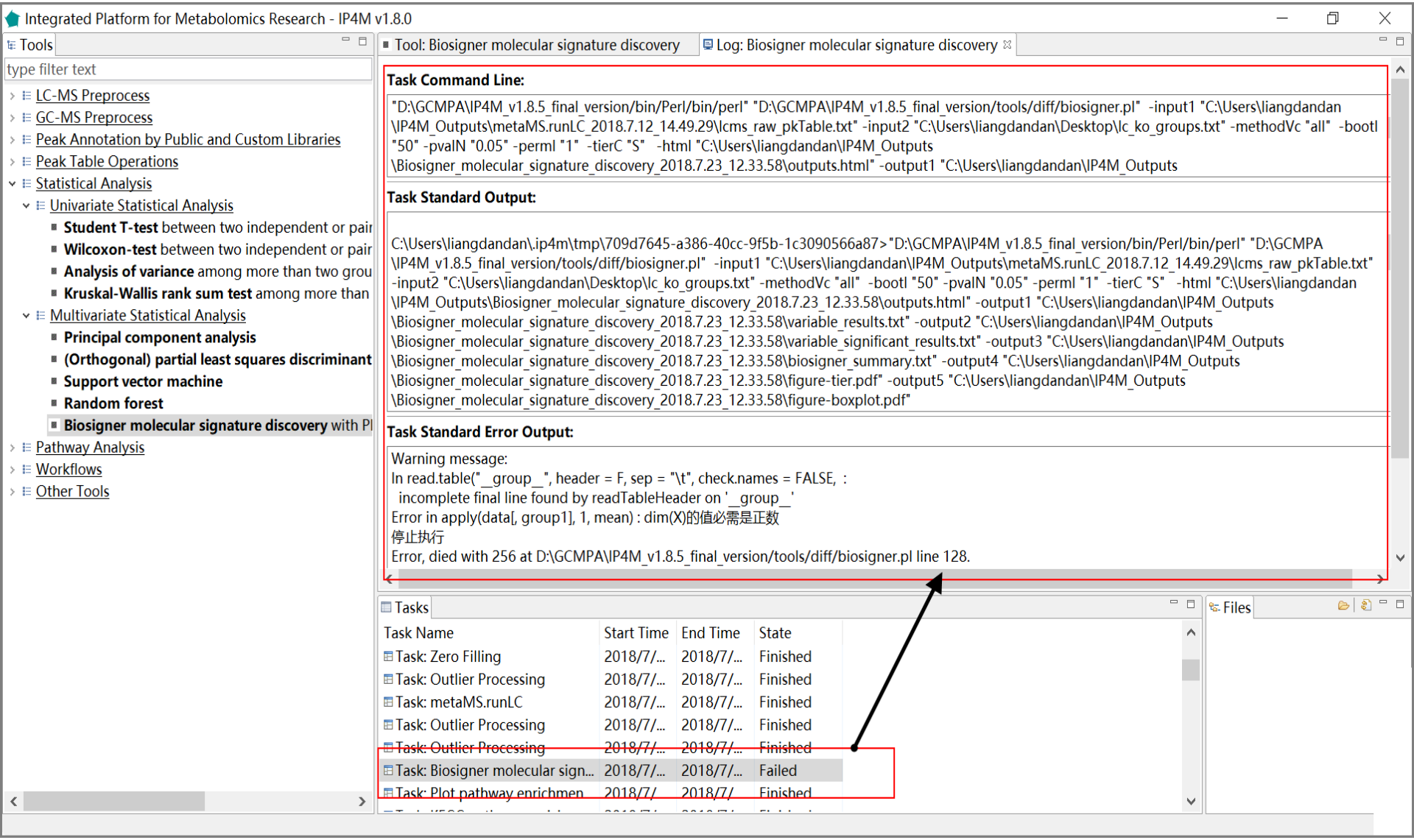
Modules and mian functions-2

- 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

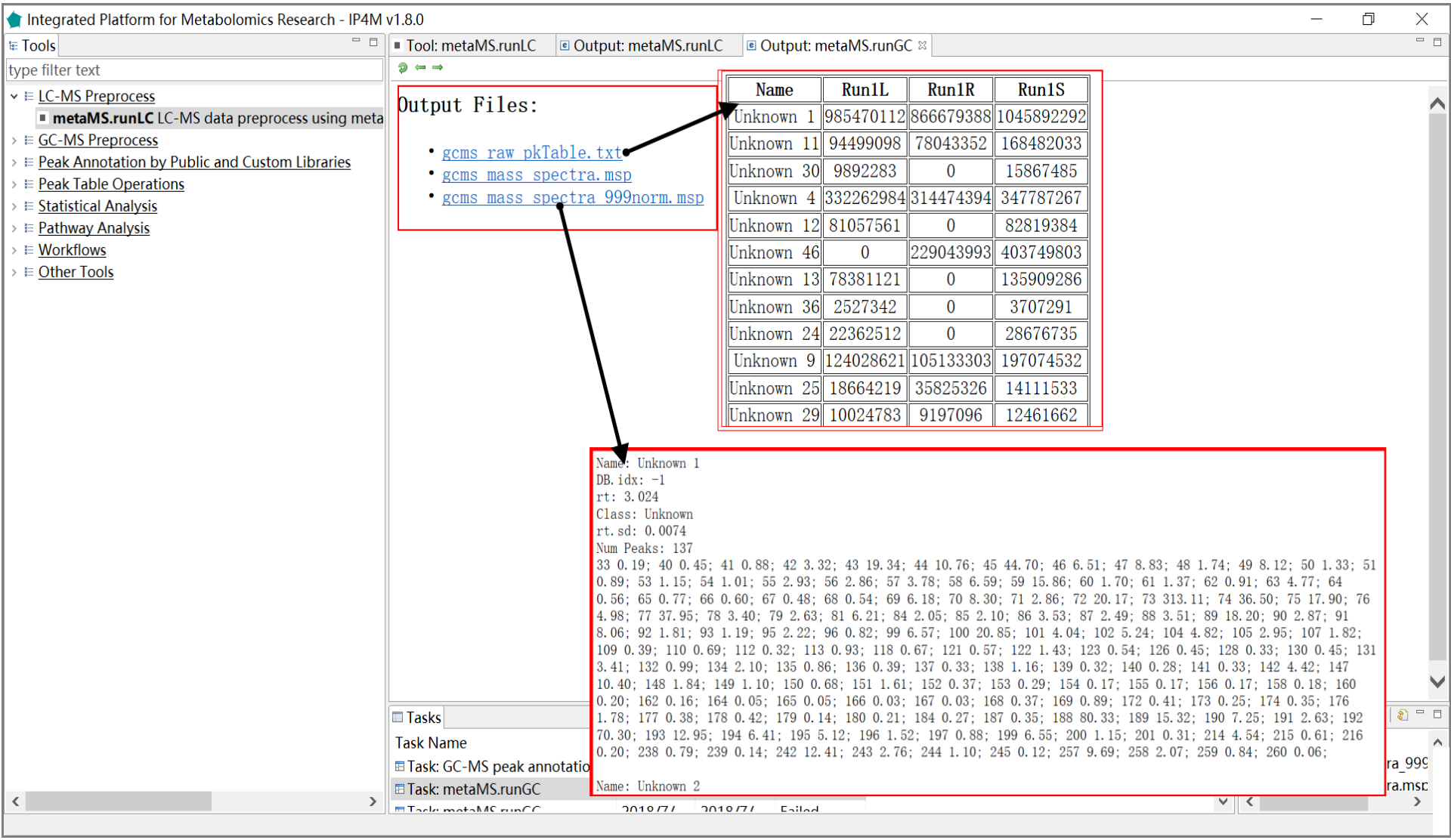
Demos



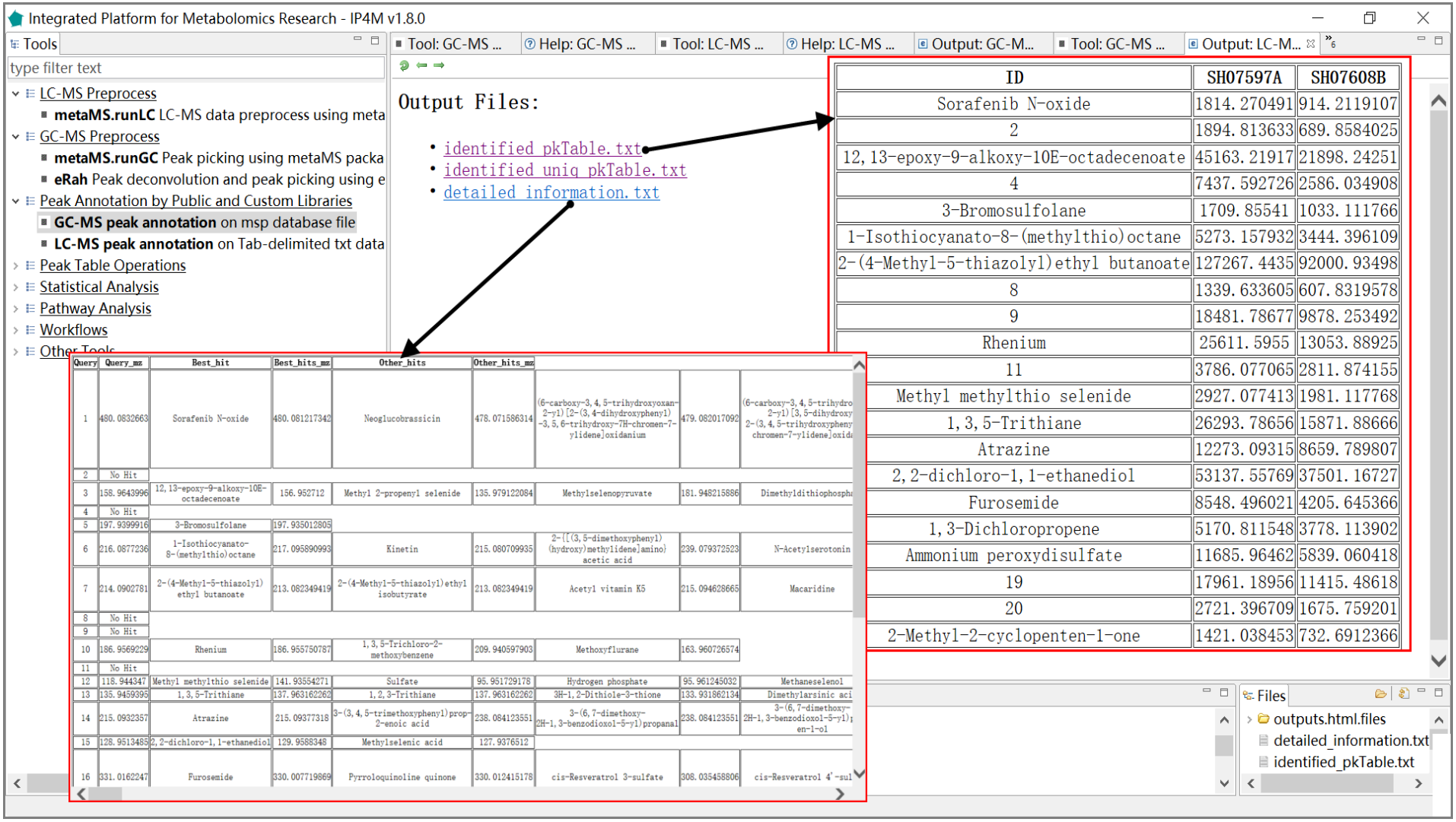
Workflow of usage



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



Outputs of LC-MS preprocessing (peak identification)



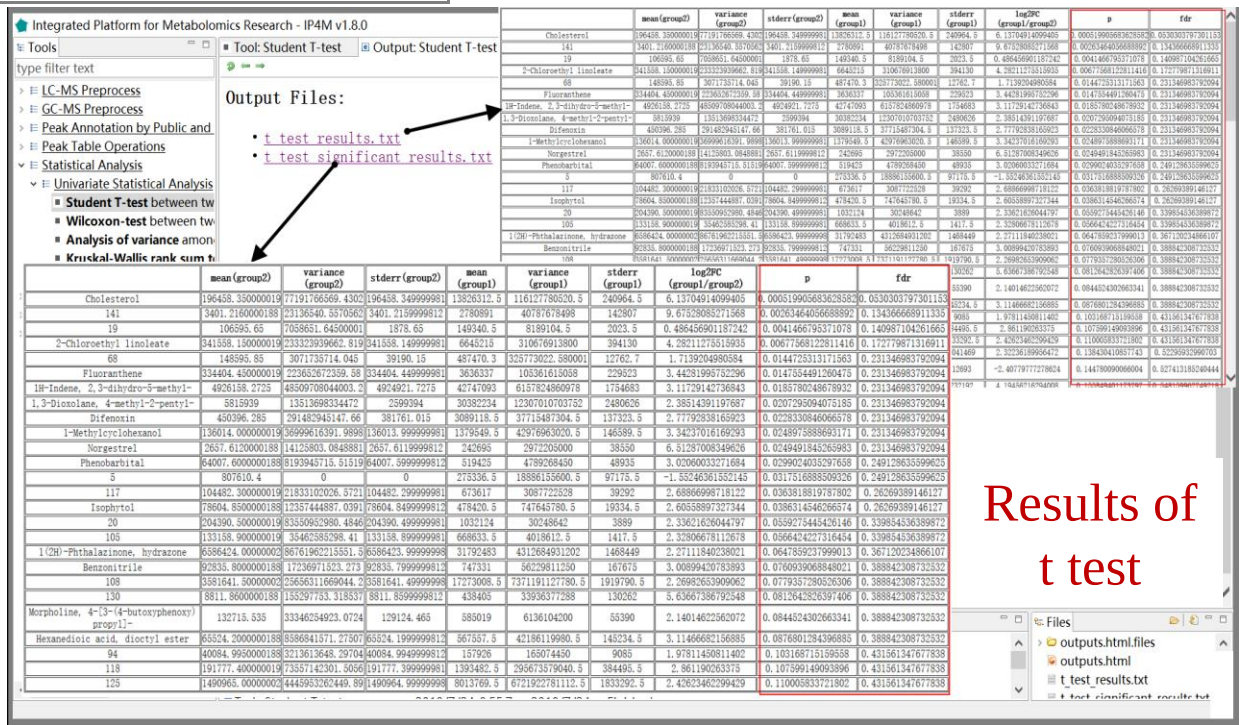
ID	SH07597A	SH07608B
Sorafenib N-oxide	1814.270491	914.2119107
2	1894.813633	689.8584025
12,13-epoxy-9-alkoxy-10E-octadecenoate	45163.21917	21898.24251
4	7437.592726	2586.034908
3-Bromosulfolane	1709.85541	1033.111766
1-Isothiocyanato-8-(methylthio)octane	5273.157932	3444.396109
2-(4-Methyl-5-thiazolyl)ethyl 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

Query	Query_mz	Best_hit	Best_hits_mz	Other_hits	Other_hits_mz			
1	480.0832663	Sorafenib N-oxide	480.081217342	Neoglucobrassicin	478.071586314	(6-carboxy-3,4,5-trihydroxyoxan-2-yl)[2-(3,4-dihydroxyphenyl)-3,5,6-trihydroxy-7H-chromen-7-ylidene]oxidanium	479.082017092	(6-carboxy-3,4,5-trihydro 2-yl)[3,5-dihydroxy 2-(3,4,5-trihydroxypheny chromen-7-ylidene]oxid
2	No Hit							
3	158.9643996	12,13-epoxy-9-alkoxy-10E-octadecenoate	156.952712	Methyl 2-propenyl selenide	135.979122084	Methylselenopyruvate	181.948215886	Dimethyldithiophosph
4	No Hit							
5	197.9399916	3-Bromosulfolane	197.935012805					
6	216.0877236	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 K5	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 aci
14	215.0932357	Atrazine	215.09377318	3-(3,4,5-trimethoxyphenyl)prop-2-enoic acid	238.084123551	3-(6,7-dimethoxy-2H-1,3-benzodioxol-5-yl)propanal	238.084123551	3-(6,7-dimethoxy-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

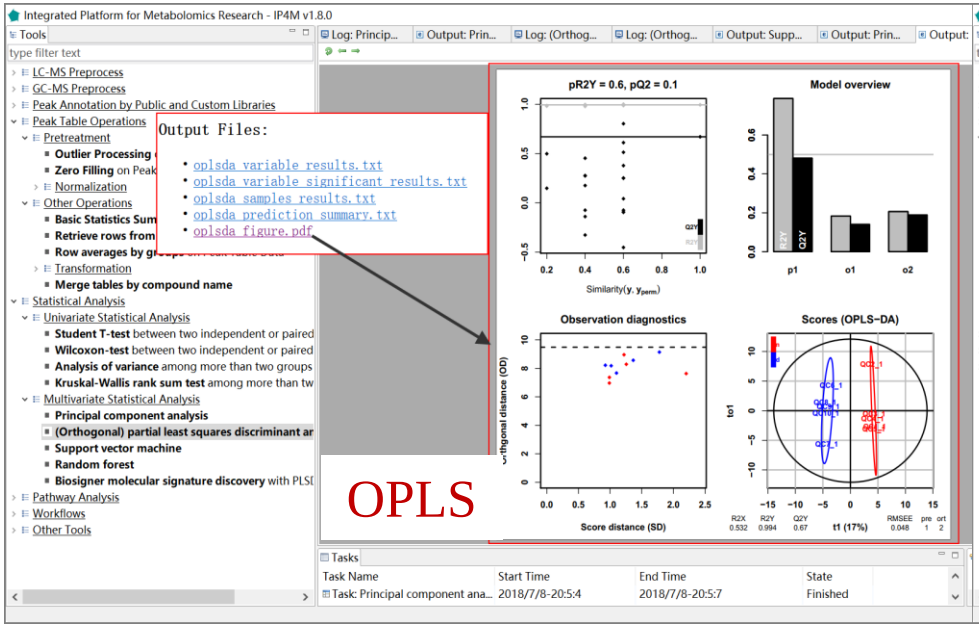
Outputs of GC-MS peak annotation

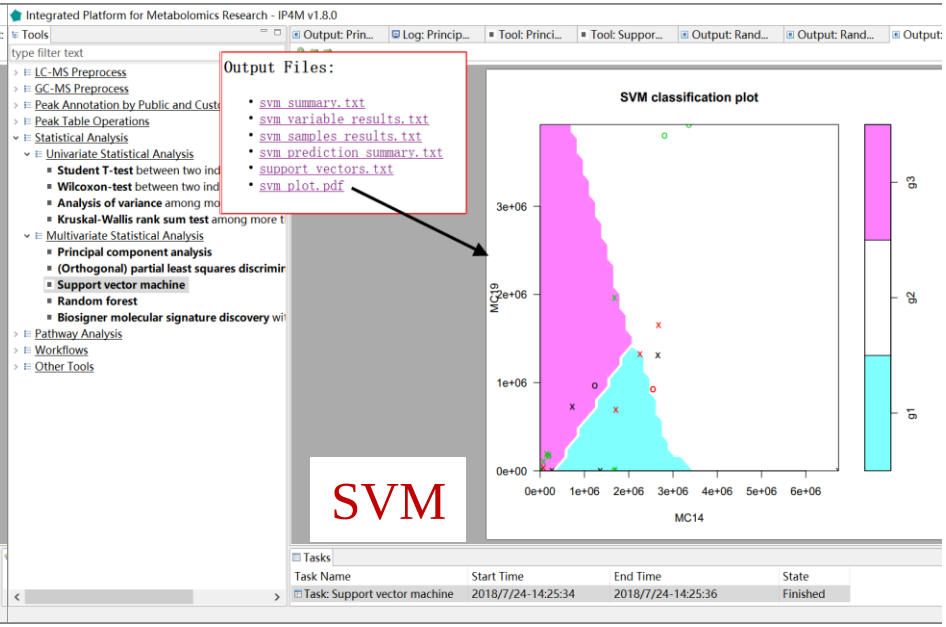
	Glycine-d5_1	N2_1	N3_1	QC1_1
nbr.val	102	102	102	102
nbr.null	0	0	0	0
nbr.na	0	0	0	0
min	3.85417760431785e-12	0.00893943757733268	0.057258514677102	8.31486109559897e-14
max	329.680542925297	112.89922337076	125.740860123838	252.160556248509
range	329.680542925293	112.890283933183	125.683601609161	252.160556248509
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	426.822220522143	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		

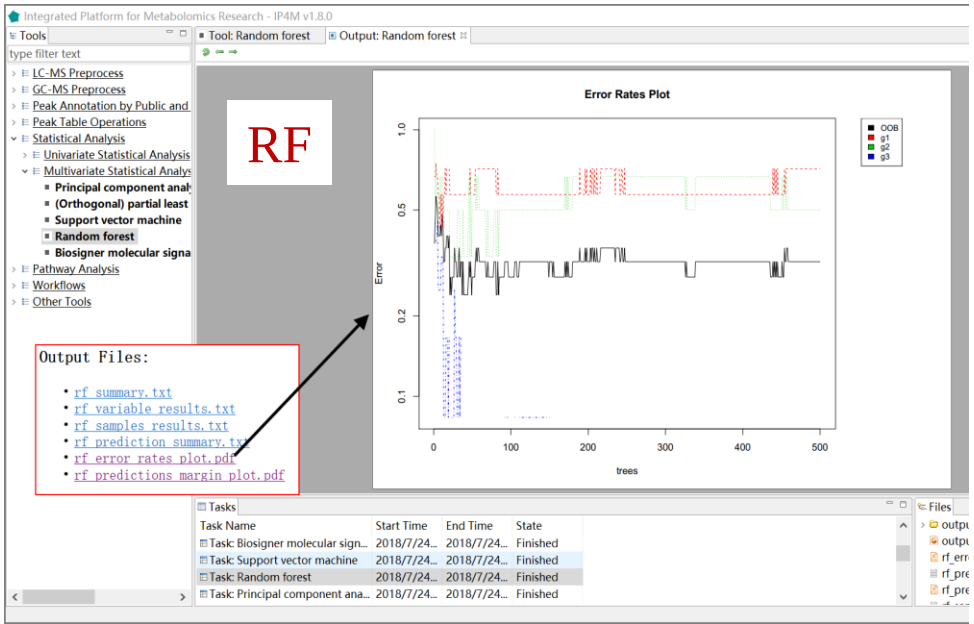
Basic statistics computation

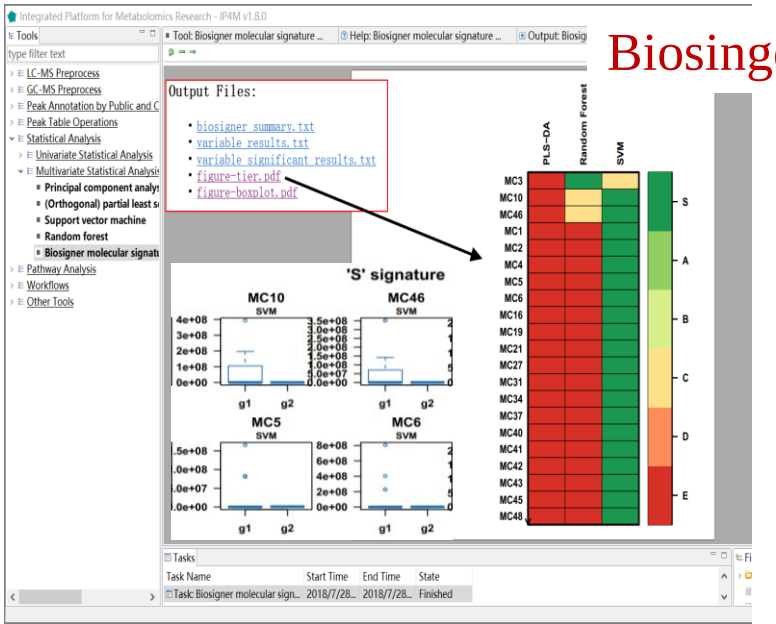


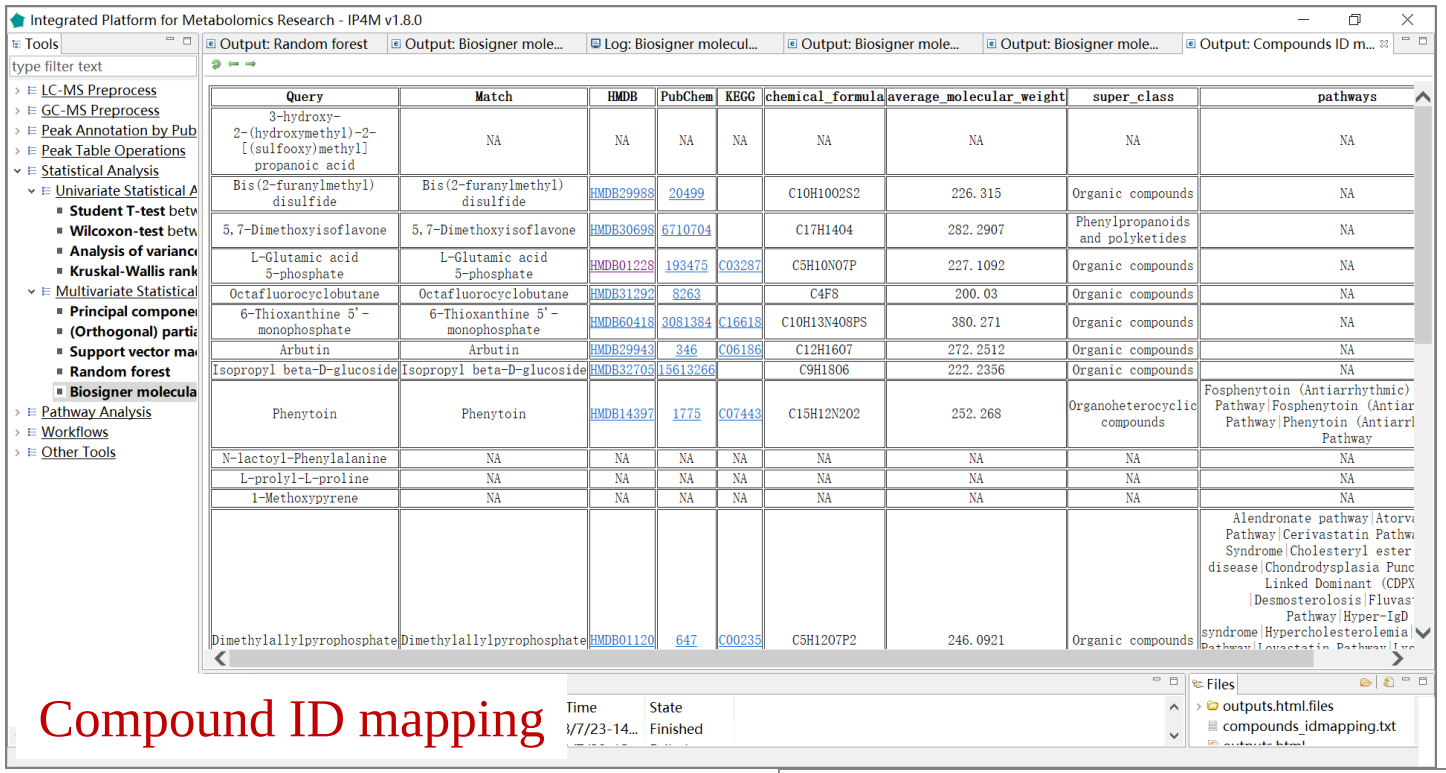
Outputs of multivariate machine learning analysis



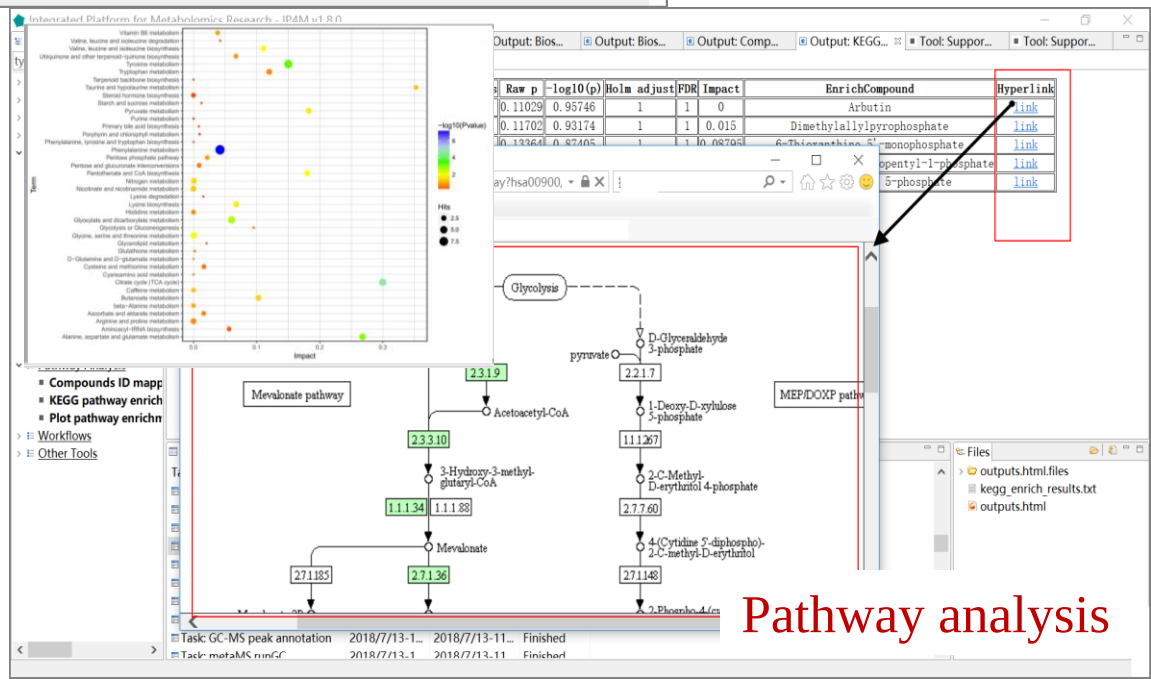




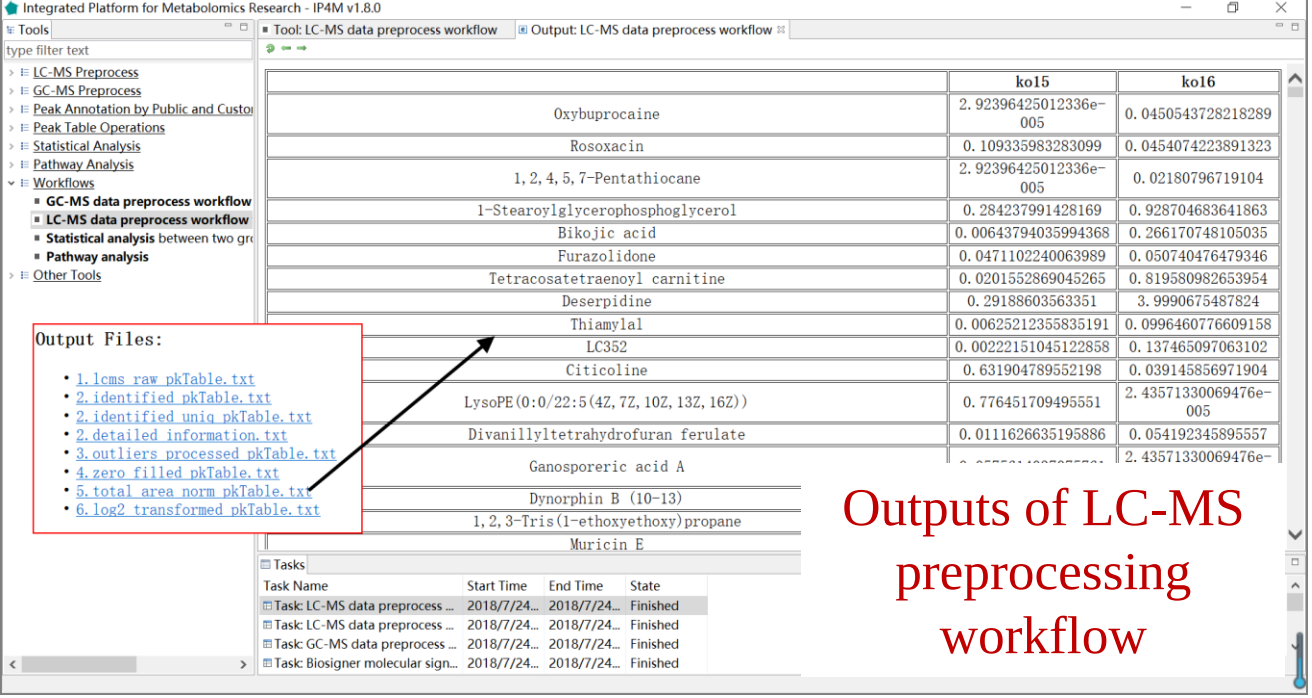




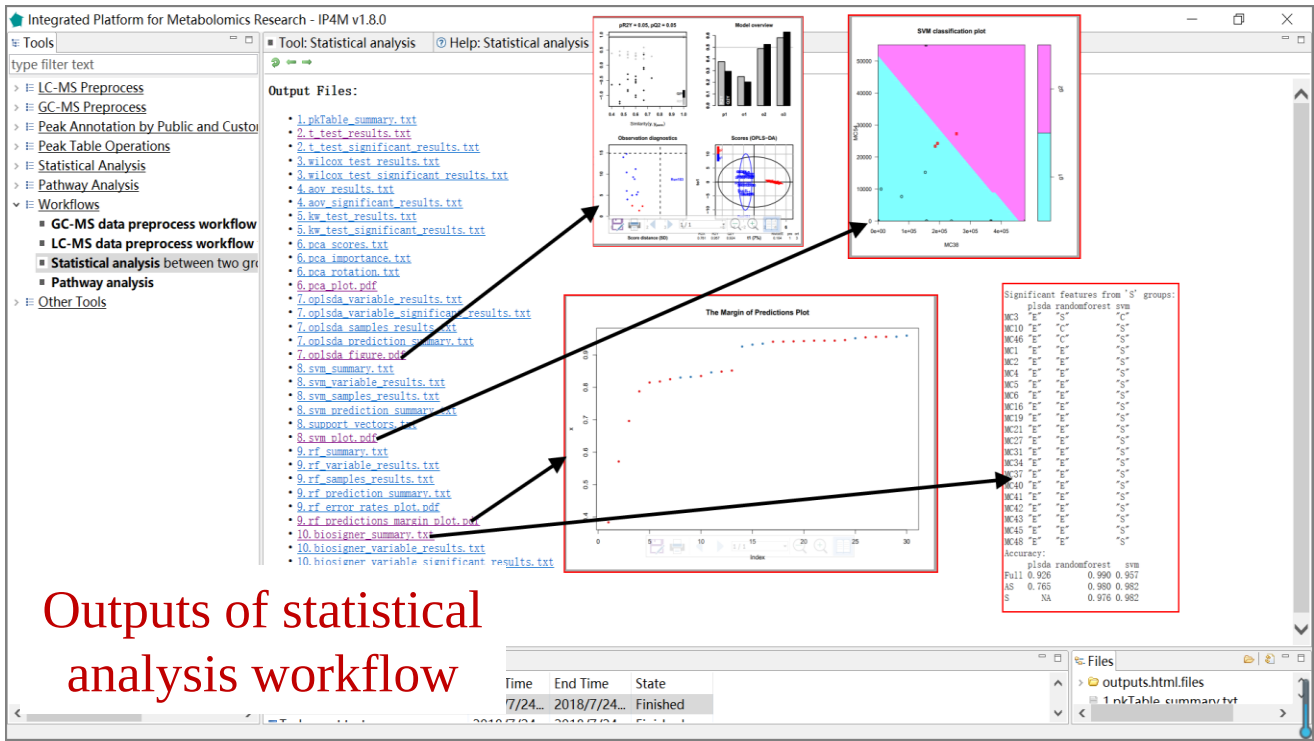
Compound ID mapping



Pathway analysis

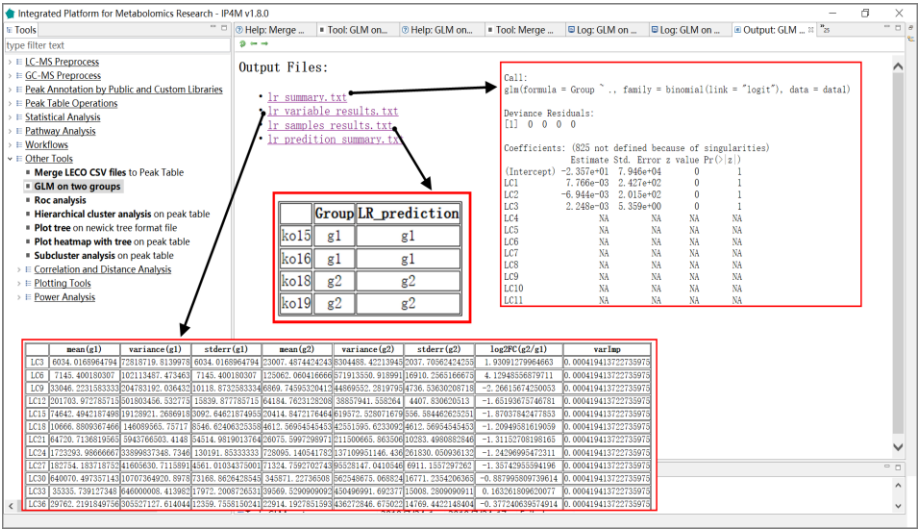


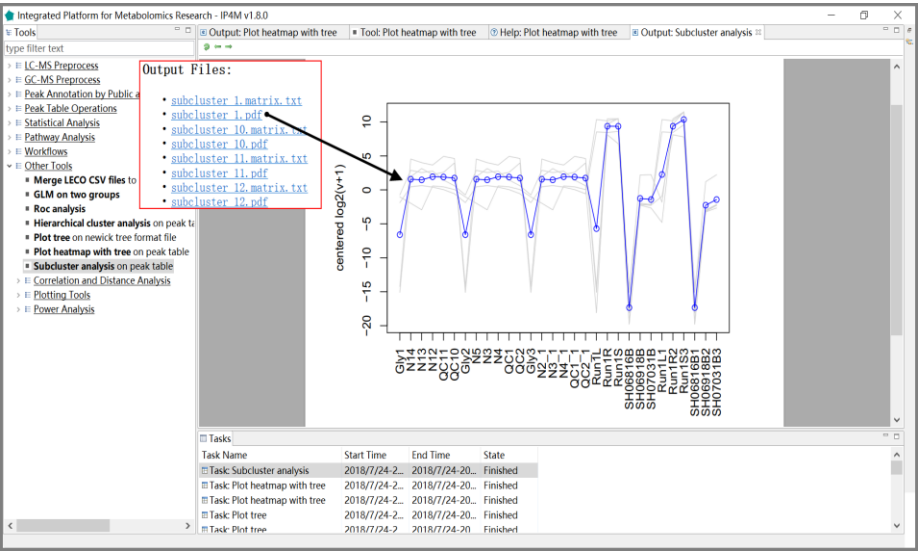
Outputs of LC-MS preprocessing workflow

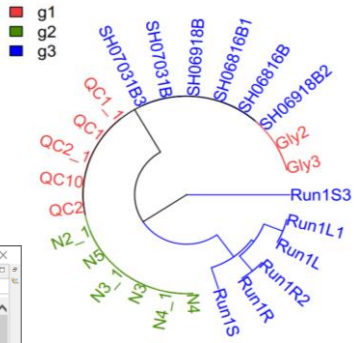


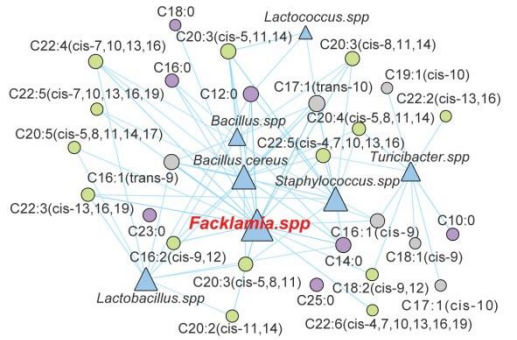
Outputs of statistical analysis workflow

Results of “Other tools”









Metabolite type:

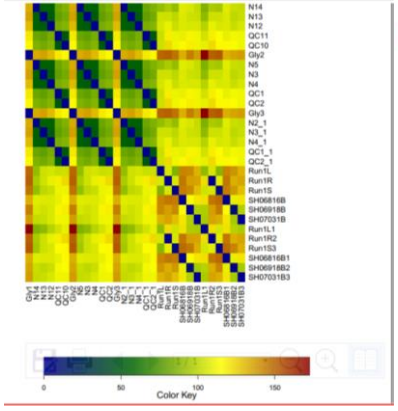
Lysophosphatidylcholine

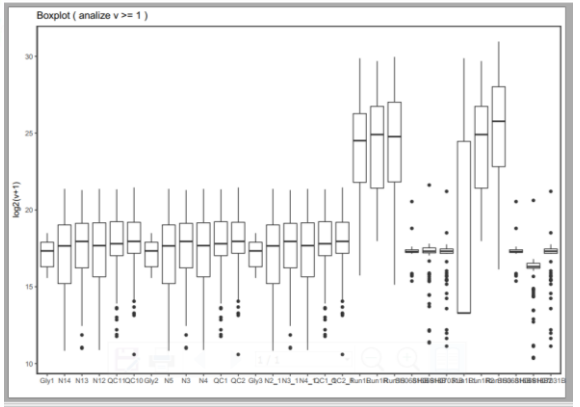
Phosphatidylcholine

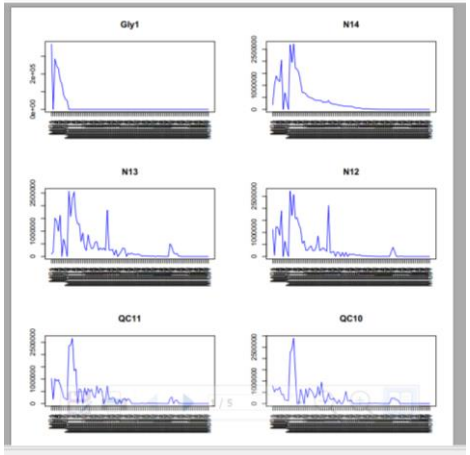
Bacterial class:

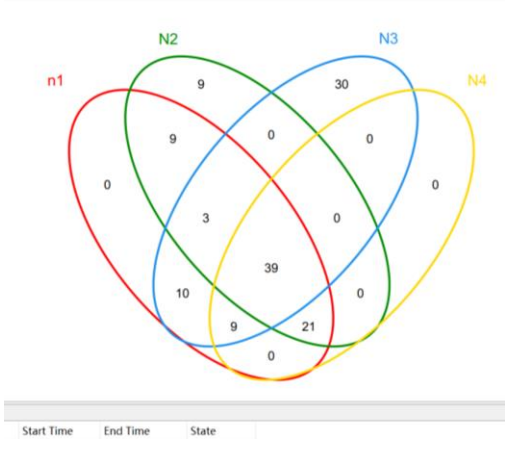
Spirochaetes under Spirochaetes phylum

Bacilli under Firmicutes phylum









- **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; chentianlu@sjtu.edu.cn