A how-to/manual for the MeKO database

(Apr 30, 2014 version)

Atsushi Fukushima[1,2*]

[1]RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa 230-0045, Japan

[2]JST, National Bioscience Database Center (NBDC), 5-3, Yonbancho, Chiyoda-ku, Tokyo, 102-0081, Japan

[*]Contact, e-mail: a-fukush@psc.riken.jp

Note: The current analysis page in MeKO does not work well with Mozilla Firefox. We recommend using Google Chrome or Microsoft Internet Explorer.

Introduction

Despite recent intensive research efforts in functional genomics, the functions of only a limited number of Arabidopsis (*Arabidopsis thaliana*) genes have been determined experimentally and improving gene annotation remains a major challenge in plant systems biology (Fukushima et al., 2009). As metabolite profiling can characterize the metabolomic phenotype of a genetic perturbation in the plant metabolism, it provides clues to the function(s) of genes of interest. We evaluated the comprehensive metabolite profiles of 50 Arabidopsis mutants, including a set of characterized and uncharacterized mutants, that resemble wild-type plants. We detected previously hidden alterations in the production of metabolites, particularly in primary metabolism. We expect our data to help improve our understanding of gene function as well as plant growth and development. We developed the MeKO database (http://prime.psc.riken.jp/meko/) to use the dataset as a functional genomics tool. Metabolites whose accumulation differs from that of wild-type plants are conveniently displayed in a metabolic map that includes relational information. Our analyses not only revealed the metabotype of each mutant but the MeKO database can also be used for the generation of testable hypotheses concerning the functions of genes of interest and it is highly useful for improving gene annotation.

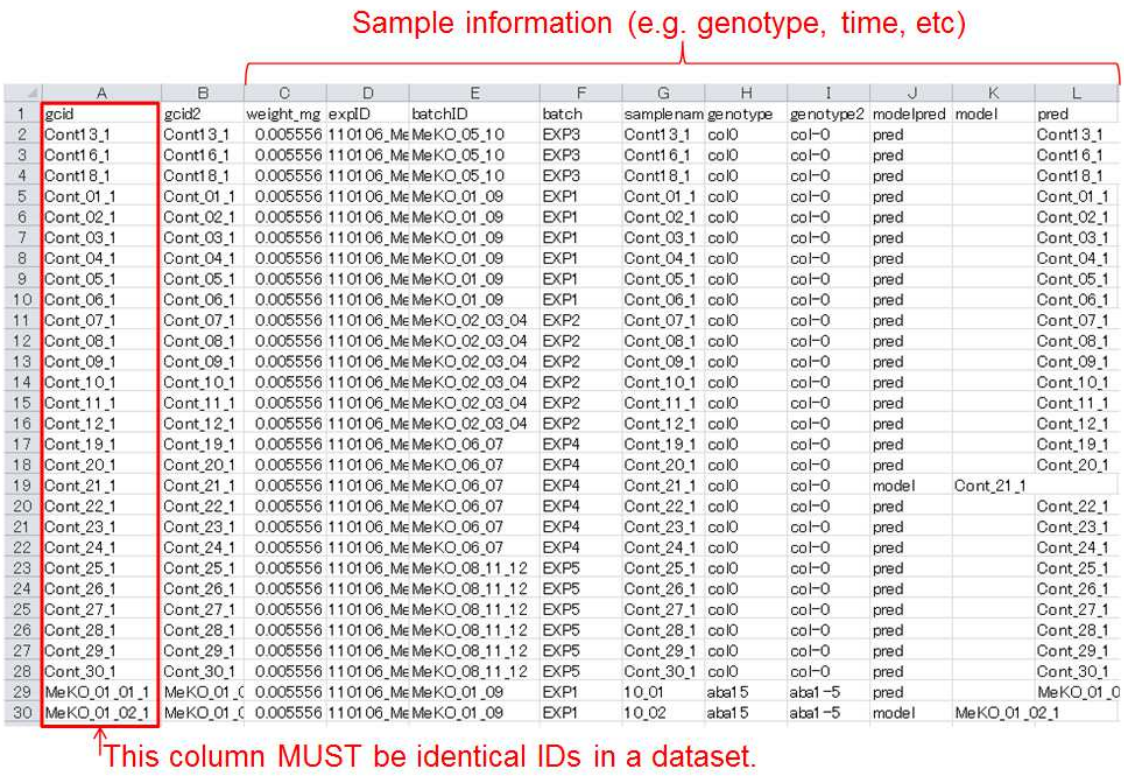
Typical data format

(1) RIKEN GC-MS format

Figure 1 shows a typical data format in MeKO. Our example data are downloadable from here. As we are using R 'shiny' package to implement our analysis tools on MeKO, in our server, all datasets are manipulated as an ***ExpressoinSet*** class in a 'Biobase' package; in brief, as information describing the features or metabolites (corresponding to columns in a data matrix) and the expression/accumulation data (numerical values) as assay data. Information describing the sample phenotypes (hereafter called 'phenodata') is also required (Figure 2). See also the Bioconductor paper by (Gentleman et al., 2004).

A	B	C	D	E	F	G	H	I	J	K
Numbering for matrix	Numbering for unique peaks	IIS_I_A_or_U	I_A_M_or_P	final annotation	RI	Win	Cont13_1	Cont16_1	Cont18_1	Cont_01_1
Mekomodpred01_ID001	Mekomodpred01_cmb_ID001	U			1016.2	252.7 Win001_C01	3270230	4110148	2566931	96531.44
Mekomodpred01_ID002	Mekomodpred01_cmb_ID002	U			1028.9	257.21 Win002_C01	4773480	4928577	4615155	6248591
Mekomodpred01_ID003	Mekomodpred01_cmb_ID003	U			1036	259.7 Win003_C01	1659116	2046839	1431654	17133722
Mekomodpred01_ID004	Mekomodpred01_cmb_ID004	I		Lactic acid, DL- (2TMS)	1046.4	263.36 Win004_C01	8481751	9978936	7831072	1606885
Mekomodpred01_ID005	Mekomodpred01_cmb_ID005	U			1049.5	264.52 Win004_C02	2720221	2956172	1290158	3415789
Mekomodpred01_ID006	Mekomodpred01_cmb_ID006	U			1051.5	265.18 Win004_C03	557997.3	231395.1	804563.1	425612.7
Mekomodpred01_ID007	Mekomodpred01_cmb_ID007	U			1065.6	270.16 Win004_C04	8506.966	24960.98	31451.96	749991.7
Mekomodpred01_ID008	Mekomodpred01_cmb_ID008	U			1080	275.34 Win005_C01	4157717	3264610	1954569	1353434
Mekomodpred01_ID009	Mekomodpred01_cmb_ID009	U			1087.3	277.86 Win005_C02	137099.6	91571.09	131303.3	27908.3
Mekomodpred01_ID010	Mekomodpred01_cmb_ID010	I		Alanine (2TMS)	1094.9	280.55 Win006_C01	24789174	13074553	24043492	25210793
Mekomodpred01_ID011	Mekomodpred01_cmb_ID011	U			1096.6	281.18 Win006_C02	3027172	3472154	2698063	502015.3
Mekomodpred01_ID012	Mekomodpred01_cmb_ID012	U			1098.9	281.95 Win006_C03	911408.6	910153.3	740711	4808484
Mekomodpred01_ID013	Mekomodpred01_cmb_ID013	A1		Undecane	1101.2	282.78 Win006_C04	635.7388	0	1706.031	5234.836
Mekomodpred01_ID014	Mekomodpred01_cmb_ID014	U			1106.5	284.67 Win006_C05	405748.2	423751.1	374907.7	1002811
Mekomodpred01_ID015	Mekomodpred01_cmb_ID015	I		Hydroxylamine (3TMS)	1113	286.99 Win006_C06	41209434	25502060	46563710	66284279
Mekomodpred01_ID016	Mekomodpred01_cmb_ID016	A2		Glycine (2TMS)	1114.7	287.59 Win006_C07	1145346	1252506	1092015	2124538
Mekomodpred01_ID017	Mekomodpred01_cmb_ID017	U			1122	290.15 Win007_C01	33000538	35424420	31798115	215731.3
Mekomodpred01_ID018	Mekomodpred01_cmb_ID018	A2		Dihydrouracil[2TMS]	1133.9	294.43 Win007_C02	1707098	3127041	5207291	4220966
Mekomodpred01_ID019	Mekomodpred01_cmb_ID019	U			1140.4	296.69 Win008_C01	1973255	2279767	1619931	63699.5
Mekomodpred01_ID020	Mekomodpred01_cmb_ID020	U			1147.4	299.15 Win008_C02	54925.99	63209.71	49646.01	20317.66
Mekomodpred01_ID021	Mekomodpred01_cmb_ID021	M1		M000000_A123002-101_CONT_1228.8	1156.2	302.3 Win009_C01	2812570	3078225	2158938	3165168
Mekomodpred01_ID022	Mekomodpred01_cmb_ID022	U			1157	302.6 Win009_C02	1050316	1184370	1040731	493126.4
Mekomodpred01_ID023	Mekomodpred01_cmb_ID023	U			1159	303.3 Win009_C03	19752.95	7511.435	18618.77	10908
Mekomodpred01_ID024	Mekomodpred01_cmb_ID024	U			1161.5	304.19 Win009_C04	456905.9	413820	720016	684514.1
Mekomodpred01_ID025	Mekomodpred01_cmb_ID025	U			1165.5	305.59 Win009_C05	97991.42	58422.75	77047.91	119388.1
Mekomodpred01_ID026	Mekomodpred01_cmb_ID026	U			1172.8	308.21 Win010_C01	93562.05	105111.3	134738.1	88478391
Mekomodpred01_ID027	Mekomodpred01_cmb_ID027	U			1174.5	308.84 Win010_C02	4357970	4948605	4841491	7956125
Mekomodpred01_ID028	Mekomodpred01_cmb_ID028	U			1184.6	312.39 Win010_C03	43063.35	151.0021	336084.5	1179792
Mekomodpred01_ID029	Mekomodpred01_cmb_ID029	U			1186.9	313.16 Win010_C04	138899.9	138328.7	159698.6	1710887
Mekomodpred01_ID030	Mekomodpred01_cmb_ID030	U			1194	315.65 Win011_C01	254507.6	295495.6	201848.1	4314200
Mekomodpred01_ID031	Mekomodpred01_cmb_ID031	A1		Dodecane, n-	1201.8	318.37 Win011_C02	25597.13	36662.58	20827.16	28368.52
Mekomodpred01_ID032	Mekomodpred01_cmb_ID032	U			1204.3	319.23 Win011_C03	43110.44	27034.33	60204.42	37755.4
Mekomodpred01_ID033	Mekomodpred01_cmb_ID033	I		Valine, DL- (2TMS)	1208.4	320.46 Win012_C01	5428727	4609067	4759269	5611424
Mekomodpred01_ID034	Mekomodpred01_cmb_ID034	U			1210.3	321.06 Win012_C02	2298128	2227719	1704673	3254110

Figure 1. Overview of the RIKEN GC-MS format. We assume that the feature data include "Numbering for matrix", "Numbering for unique peaks", "IIS_I_A_or_U", "I_A_M_or_P", "final annotation", "RI", and "Win". "Numbering for matrix" and "Numbering for unique peaks" represent the peak ID in a dataset. "IIS_I_A_or_U" and "I_A_M_or_P" indicate the annotation quality in the metabolite annotation step. "RI" means retention time index. "Win" indicates information from hyphenated data analysis (HDA) (Jonsson et al., 2004) and raw data analysis (RDA). The header must be in the first row and sample identifiers must be the same as in phenodata. This input file must be in tab-delimited text format.



	A	B	C	D	E	F	G	H	I	J	K	L
1	gcid	gcid2	weight_mg	expID	batchID	batch	samplenam	genotype	genotype2	modelpred	model	pred
2	Cont13_1	Cont13_1	0.005556	110106_Me	MeKO_05_10	EXP3	Cont13_1	col0	col-0	pred		Cont13_1
3	Cont16_1	Cont16_1	0.005556	110106_Me	MeKO_05_10	EXP3	Cont16_1	col0	col-0	pred		Cont16_1
4	Cont18_1	Cont18_1	0.005556	110106_Me	MeKO_05_10	EXP3	Cont18_1	col0	col-0	pred		Cont18_1
5	Cont_01_1	Cont_01_1	0.005556	110106_Me	MeKO_01_09	EXP1	Cont_01_1	col0	col-0	pred		Cont_01_1
6	Cont_02_1	Cont_02_1	0.005556	110106_Me	MeKO_01_09	EXP1	Cont_02_1	col0	col-0	pred		Cont_02_1
7	Cont_03_1	Cont_03_1	0.005556	110106_Me	MeKO_01_09	EXP1	Cont_03_1	col0	col-0	pred		Cont_03_1
8	Cont_04_1	Cont_04_1	0.005556	110106_Me	MeKO_01_09	EXP1	Cont_04_1	col0	col-0	pred		Cont_04_1
9	Cont_05_1	Cont_05_1	0.005556	110106_Me	MeKO_01_09	EXP1	Cont_05_1	col0	col-0	pred		Cont_05_1
10	Cont_06_1	Cont_06_1	0.005556	110106_Me	MeKO_01_09	EXP1	Cont_06_1	col0	col-0	pred		Cont_06_1
11	Cont_07_1	Cont_07_1	0.005556	110106_Me	MeKO_02_03_04	EXP2	Cont_07_1	col0	col-0	pred		Cont_07_1
12	Cont_08_1	Cont_08_1	0.005556	110106_Me	MeKO_02_03_04	EXP2	Cont_08_1	col0	col-0	pred		Cont_08_1
13	Cont_09_1	Cont_09_1	0.005556	110106_Me	MeKO_02_03_04	EXP2	Cont_09_1	col0	col-0	pred		Cont_09_1
14	Cont_10_1	Cont_10_1	0.005556	110106_Me	MeKO_02_03_04	EXP2	Cont_10_1	col0	col-0	pred		Cont_10_1
15	Cont_11_1	Cont_11_1	0.005556	110106_Me	MeKO_02_03_04	EXP2	Cont_11_1	col0	col-0	pred		Cont_11_1
16	Cont_12_1	Cont_12_1	0.005556	110106_Me	MeKO_02_03_04	EXP2	Cont_12_1	col0	col-0	pred		Cont_12_1
17	Cont_19_1	Cont_19_1	0.005556	110106_Me	MeKO_06_07	EXP4	Cont_19_1	col0	col-0	pred		Cont_19_1
18	Cont_20_1	Cont_20_1	0.005556	110106_Me	MeKO_06_07	EXP4	Cont_20_1	col0	col-0	pred		Cont_20_1
19	Cont_21_1	Cont_21_1	0.005556	110106_Me	MeKO_06_07	EXP4	Cont_21_1	col0	col-0	model	Cont_21_1	
20	Cont_22_1	Cont_22_1	0.005556	110106_Me	MeKO_06_07	EXP4	Cont_22_1	col0	col-0	pred		Cont_22_1
21	Cont_23_1	Cont_23_1	0.005556	110106_Me	MeKO_06_07	EXP4	Cont_23_1	col0	col-0	pred		Cont_23_1
22	Cont_24_1	Cont_24_1	0.005556	110106_Me	MeKO_06_07	EXP4	Cont_24_1	col0	col-0	pred		Cont_24_1
23	Cont_25_1	Cont_25_1	0.005556	110106_Me	MeKO_08_11_12	EXP5	Cont_25_1	col0	col-0	pred		Cont_25_1
24	Cont_26_1	Cont_26_1	0.005556	110106_Me	MeKO_08_11_12	EXP5	Cont_26_1	col0	col-0	pred		Cont_26_1
25	Cont_27_1	Cont_27_1	0.005556	110106_Me	MeKO_08_11_12	EXP5	Cont_27_1	col0	col-0	pred		Cont_27_1
26	Cont_28_1	Cont_28_1	0.005556	110106_Me	MeKO_08_11_12	EXP5	Cont_28_1	col0	col-0	pred		Cont_28_1
27	Cont_29_1	Cont_29_1	0.005556	110106_Me	MeKO_08_11_12	EXP5	Cont_29_1	col0	col-0	pred		Cont_29_1
28	Cont_30_1	Cont_30_1	0.005556	110106_Me	MeKO_08_11_12	EXP5	Cont_30_1	col0	col-0	pred		Cont_30_1
29	MeKO_01_01_1	MeKO_01_(	0.005556	110106_Me	MeKO_01_09	EXP1	10_01	aba15	aba1-5	pred		MeKO_01_0
30	MeKO_01_02_1	MeKO_01_(	0.005556	110106_Me	MeKO_01_09	EXP1	10_02	aba15	aba1-5	model	MeKO_01_02_1	

Figure 2. An example of 'phenodata'. The phenodata file must be in CSV format

(2) GENERIC format

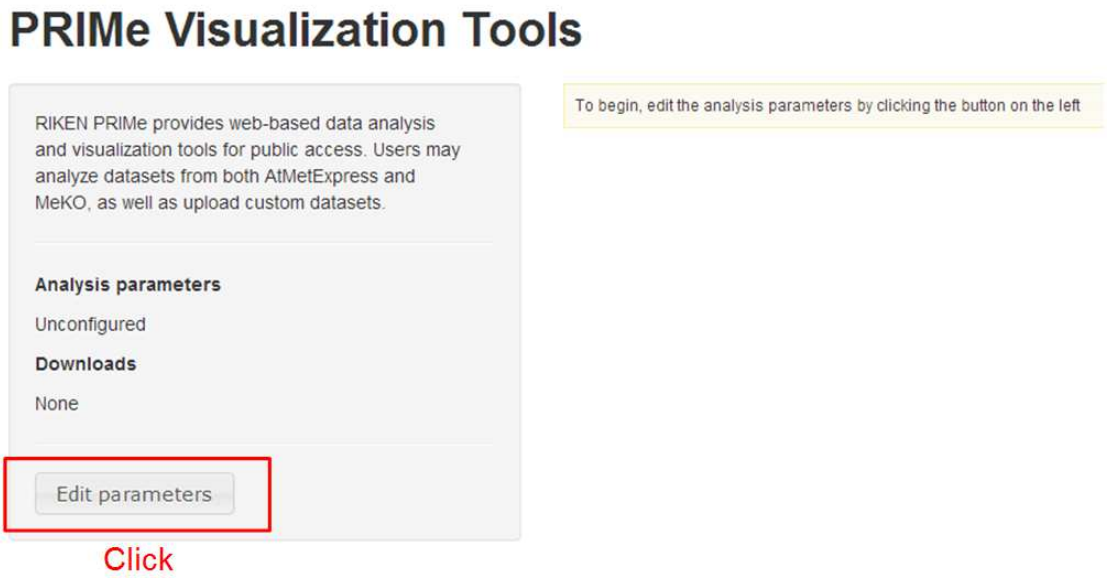
This is a generic omics data format. We assume that this file format contains a sample header in the first row and the metabolite name in the first column.

	A	B	C	D	E	
1	id	MeKO_01_(	MeKO_01_(	MeKO_01_(	MeKO_01_(	Me
2	Valine, DL- (2TMS)	3488298	7145745	5255154	4503673	6
3	Urea (2TMS)	2916023	374585.2	4232899	441553	4
4	Undecane	282.0605	55500.47	16423.15	7795	2
5	Tyrosine, DL- (3TMS)	765290.4	1390258	1294954	946359	1
6	Tryptophan, DL- (2TMS)	281674.1	560526.6	418561.1	391526	4
7	Tricosane, n-	31499.24	13757.1	41493.67	6971	2
8	Tricosane, n-	274.4928	2206.36	541.2611	3665	4
9	Tricosane, n-	401.2358	1527.462	381.3885	1220	3
10	Tricosane, n-	1777.953	885.3993	24.26211	598	2
11	Tocopherol, alpha- (1TMS)	116796.5	138678.2	118111.8	103577	1
12	Threonine, DL- (3TMS)	13800702	28346541	21688872	22270920	27
13	Threonic acid-1,4-lactone (2TMS)	3789018	5113109	5624383	3812101	5
14	Threonic acid (4TMS)	14270067	14708551	21034580	14341799	18
15	Tetradecanoic acid, 1,2,3-C3-, n- (1TI	38894273	61266980	60097925	39806304	54
16	Tetradecanoate_13C3	789028	1630548	1279768	827826	1
17	Sucrose_13C12	199697	351554	293926	210491	
18	Sucrose, D- (8TMS)	57489501	1.04E+08	89299279	72572132	1
19	Sucrose	183554	559096	306749	280377	
20	Succinate_d4_2	3026010	5098308	3527046	3106944	3

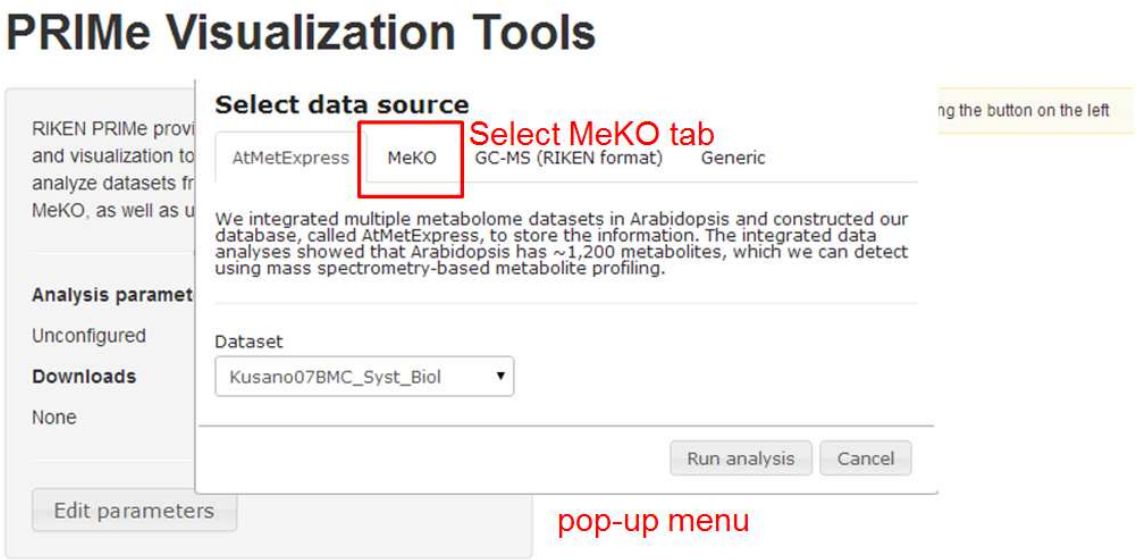
How to use the MeKO dataset

(1) Enter ‘Analysis’ page in MeKO (http://atmetexpress.riken.jp/analysis/)

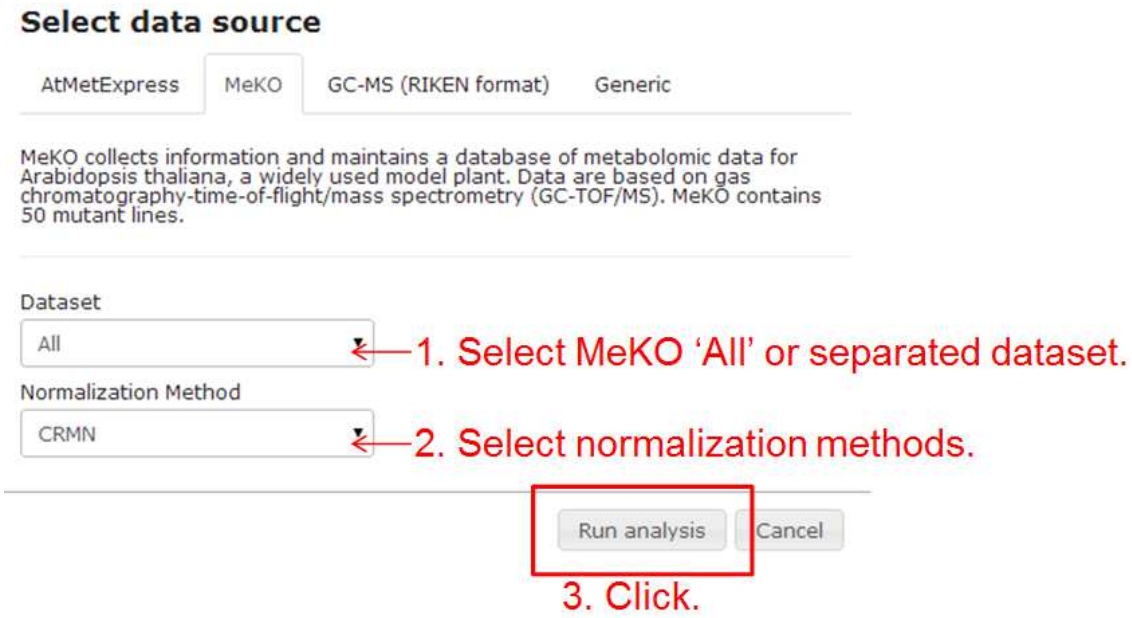
(2) Click the bottom ‘Edit parameters.’



(3) Select the MeKO tab.



(4) Select our dataset and normalization methods.



(5) Users can interactively visualize our MeKO 'All' dataset with different multivariate statistical analyses, normalization methods, distance metrics, and scaling methods (Figure 3). The results are also downloadable.

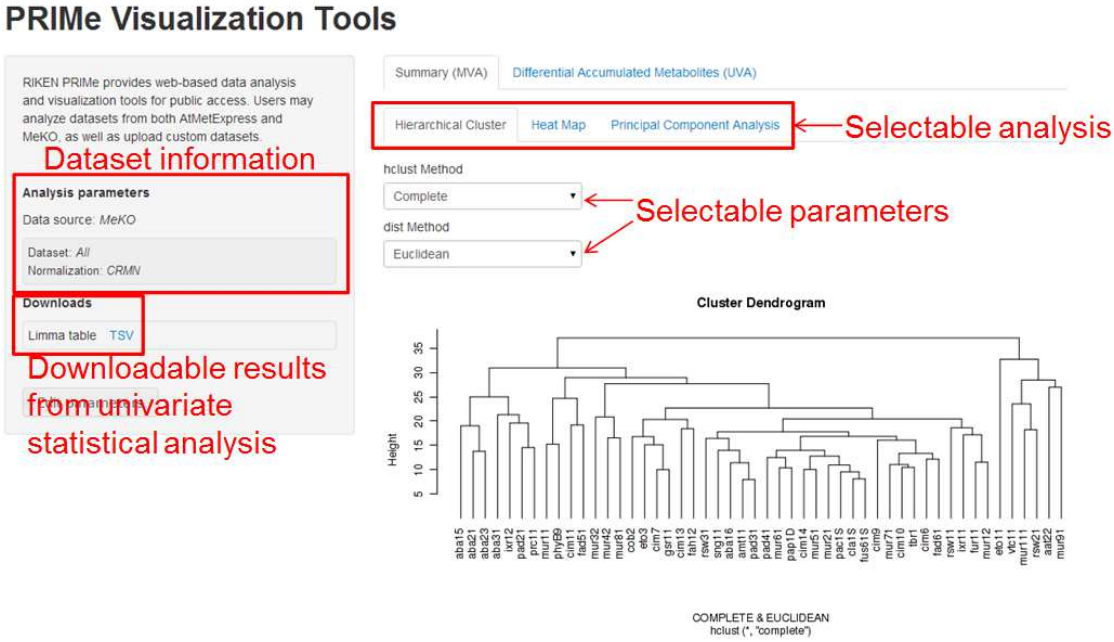
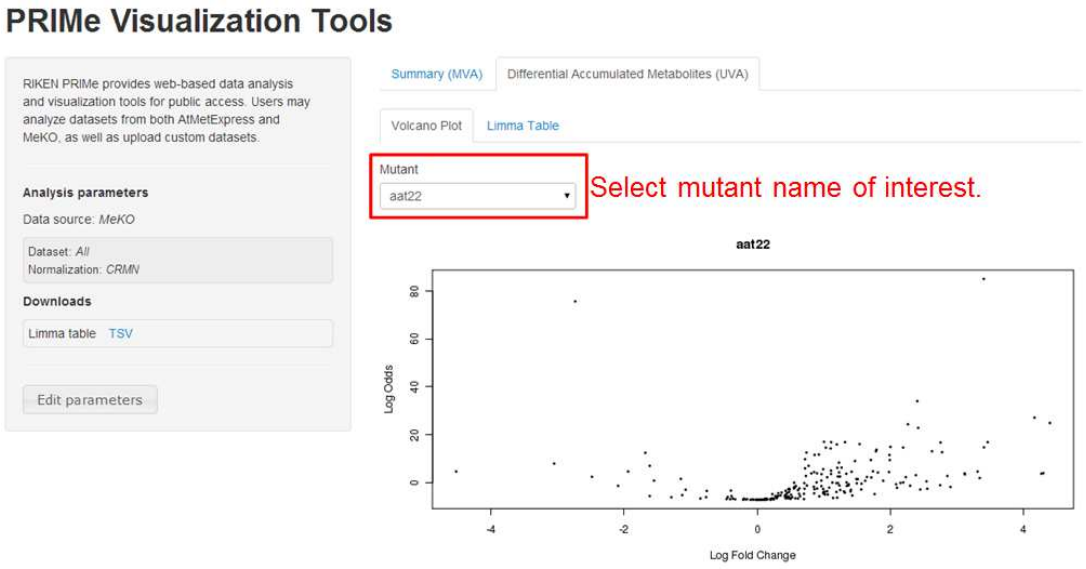
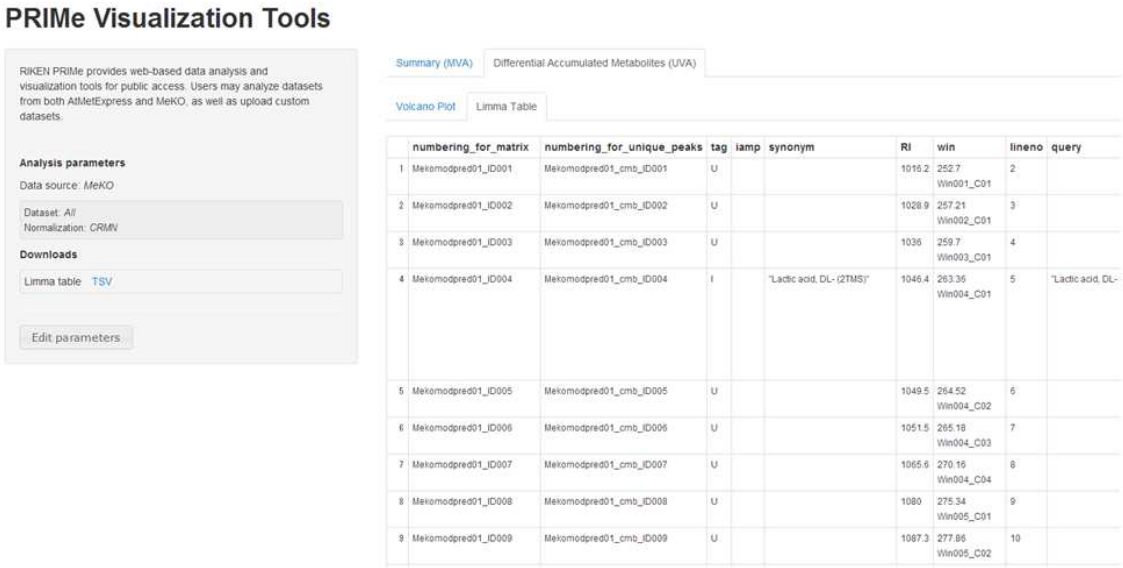


Figure 3. Visualization of MeKO 'All' data by using Hierarchical Cluster Analysis (HCA).

(6) Users can visualize the results of univariate statistical analysis with limma (Smyth, 2004) as a volcano plot and a simple text table.

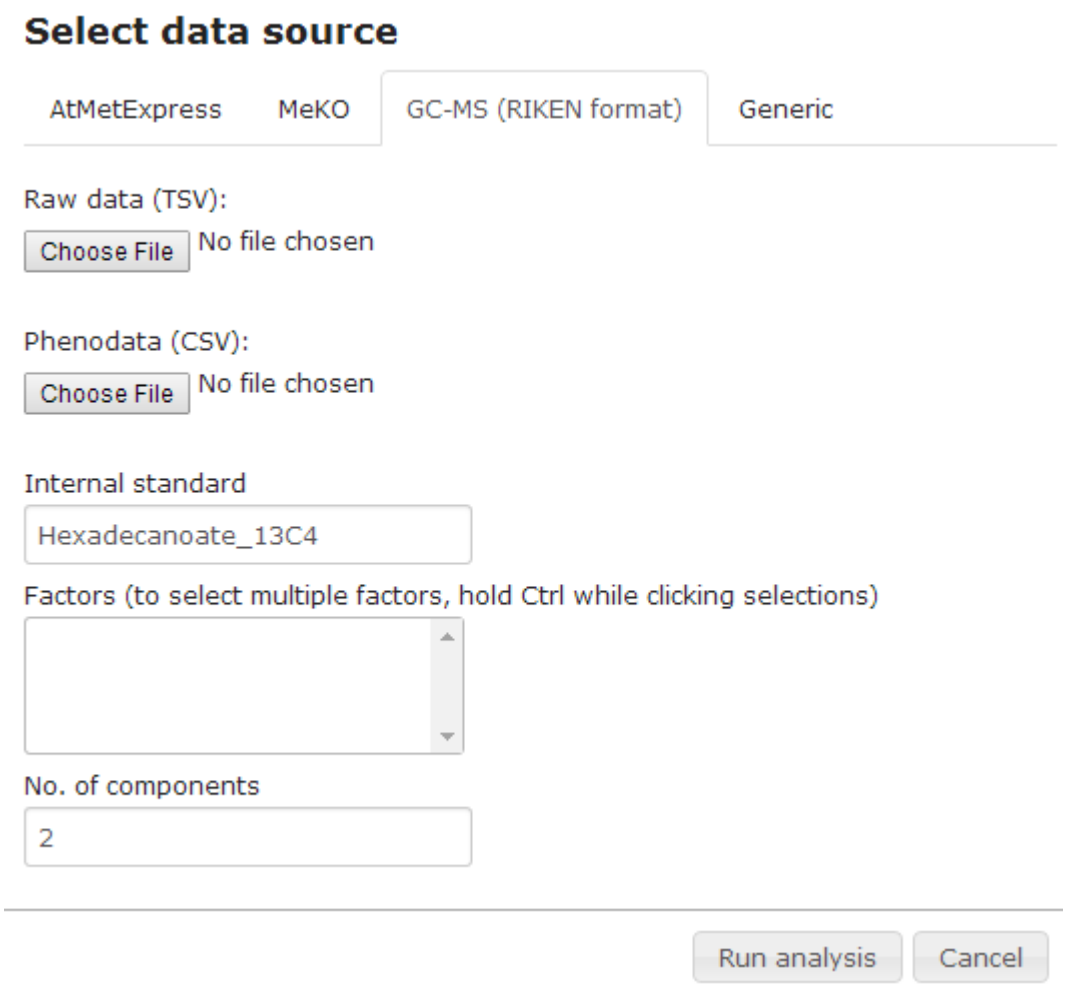




How to analyze your own dataset

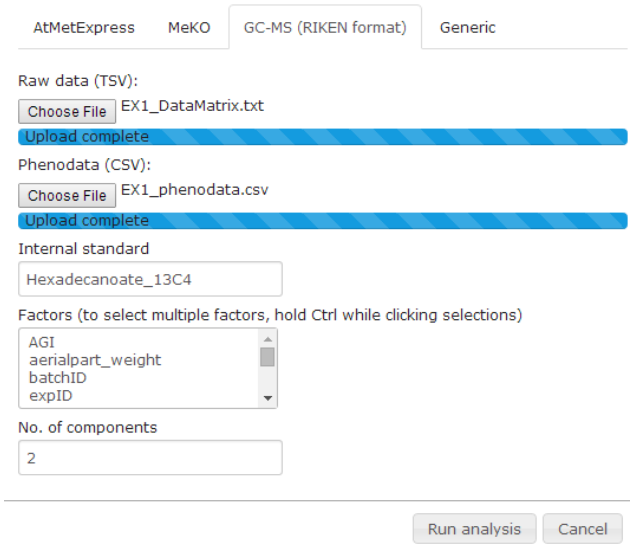
Visualization of user data with our interactive analysis tools [e.g., hierarchical cluster analysis (HCA) and a pseudo-color heatmap] includes multiple steps involving (i) dataset name, (ii) downloadable results, (iii) editing parameters, (iv) importing user data, (v) selectable parameters for multivariate analysis, and (vi) visualization.

(1) Enter ‘Analysis’ page in MeKO (http://atmetexpress.riken.jp/analysis/).
(2) Click the bottom ‘Edit parameters.’
(3) Select ‘GC-MS (RIKEN format)’.

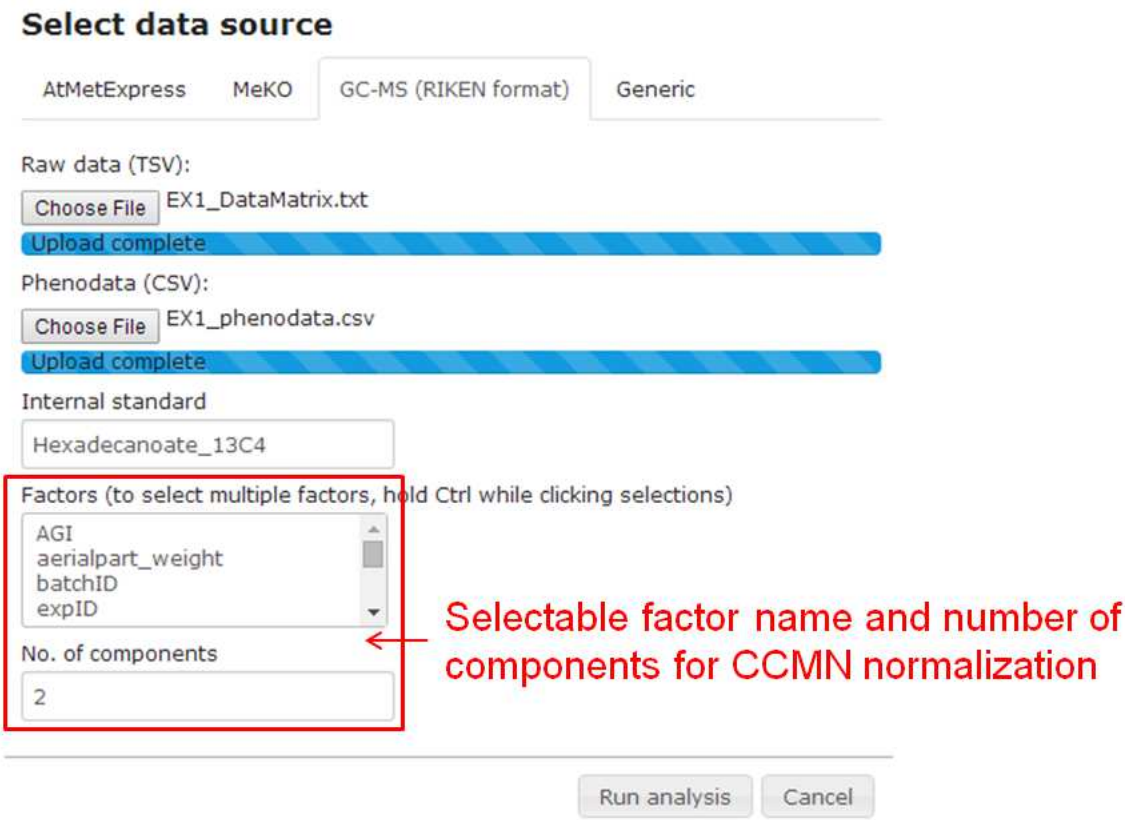


(4) Select raw data (assay data) in RIKEN GC-MS format (tab-delimited) and phenodata (see Section “Typical data format”).

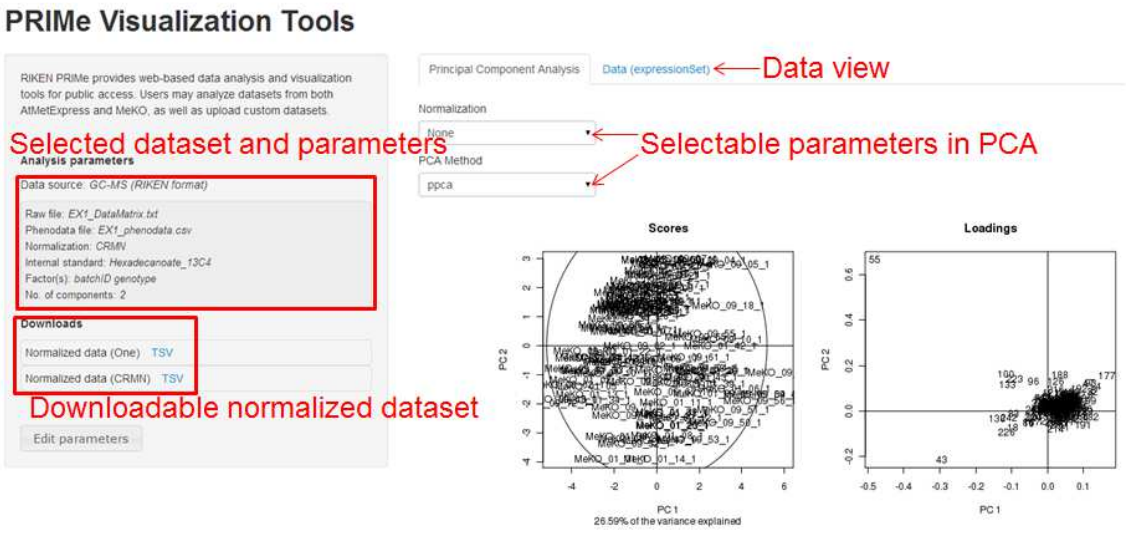
Next, we illustrate the import step using our example data EX1_DataMatrix.txt and EX1_phenodata.csv (http://atmetexpress.riken.jp/data/exdata.zip).



(5) Before selecting "run analysis", users should specify the name of the metabolite/compound to use as an internal standard for one normalization and experimental factors for CCMN normalization (Redestig et al., 2009). In this case, users should select ‘genotype’ and ‘batchID’ as the experimental factors for CCMN normalization and specify the number of components (=2).



(6) Run and visualize user's data.

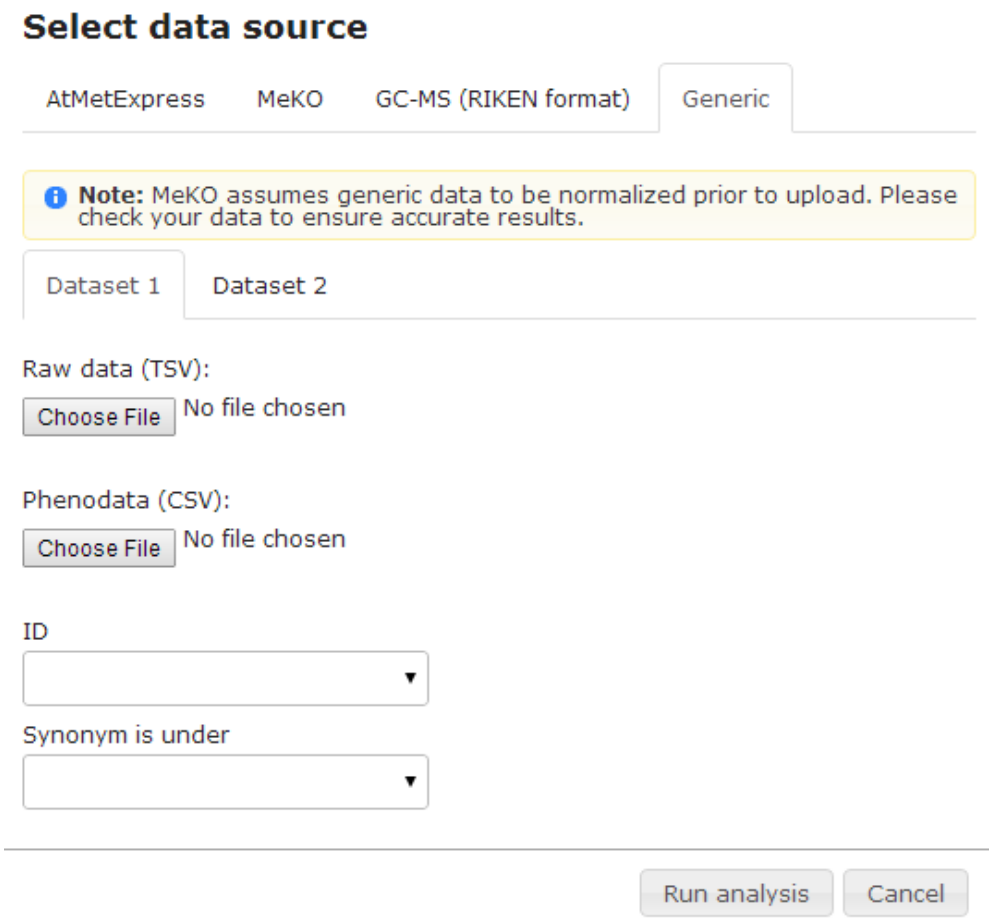


How to concatenate/summarize two datasets from one user

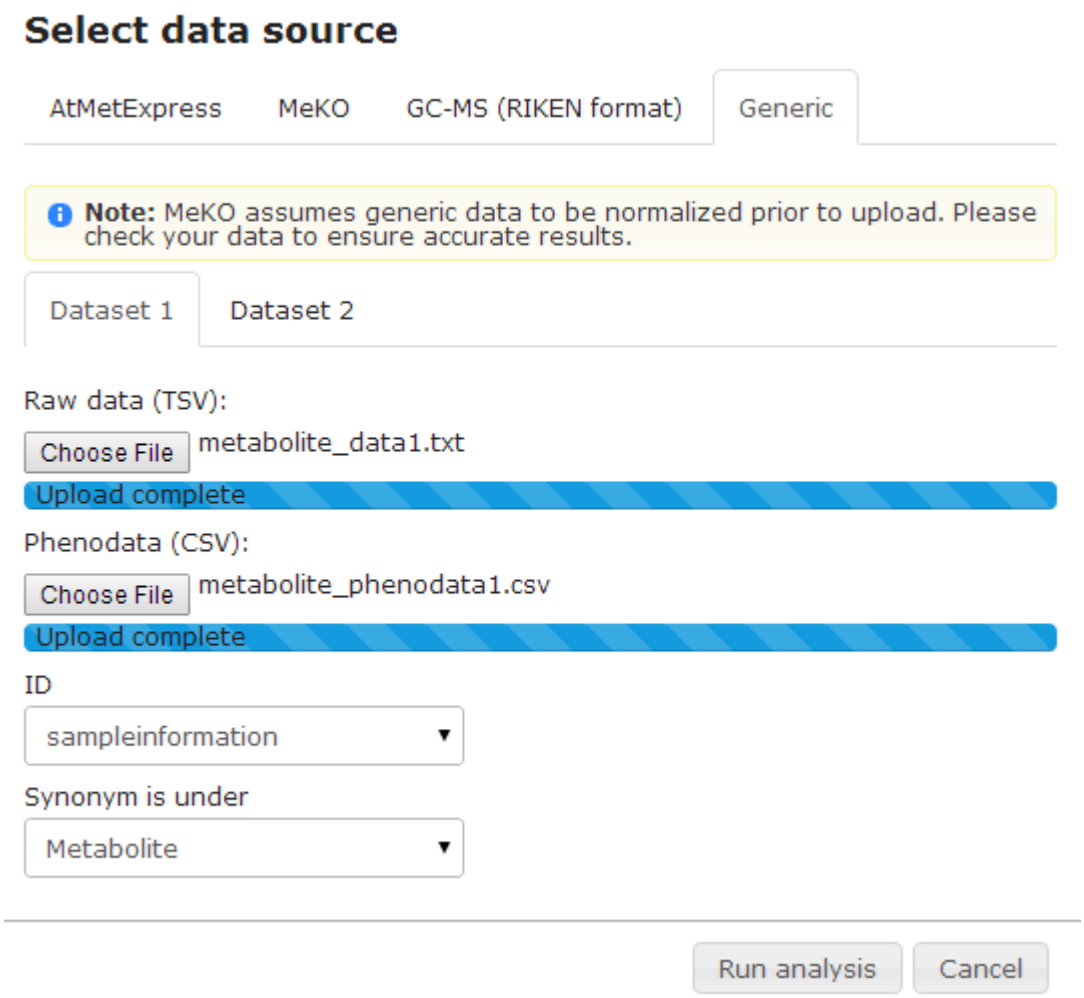
(1) Enter 'Analysis' page in MeKO (http://atmetexpress.riken.jp/analysis/).

(2) Click the bottom 'Edit parameters.'

(3) Select 'Generic'.

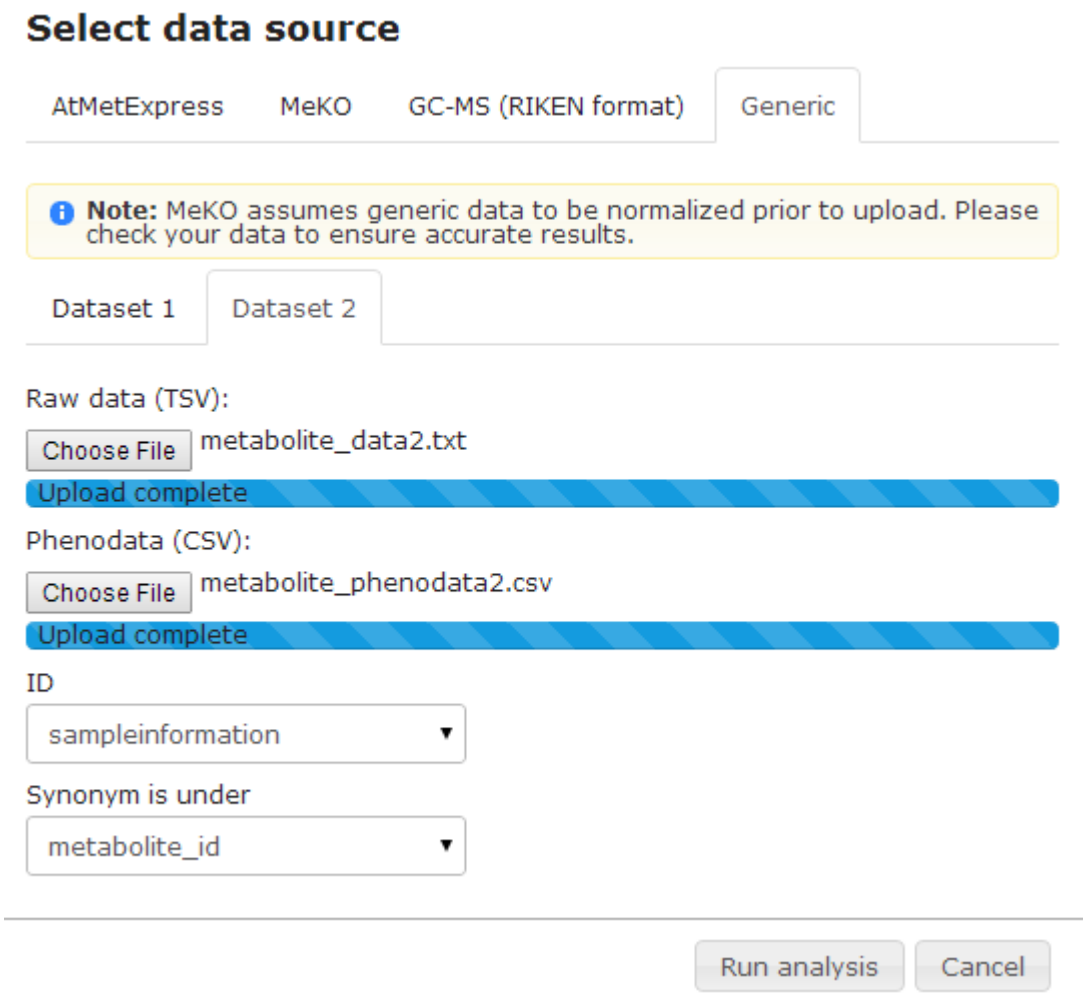


(4) Select a tab (Dataset1 or Dataset 2) and choose raw data (tab-delimited) and phenodata (CSV format). We will use our example datasets (EX2 files) (http://atmetexpress.riken.jp/data/exdata.zip).



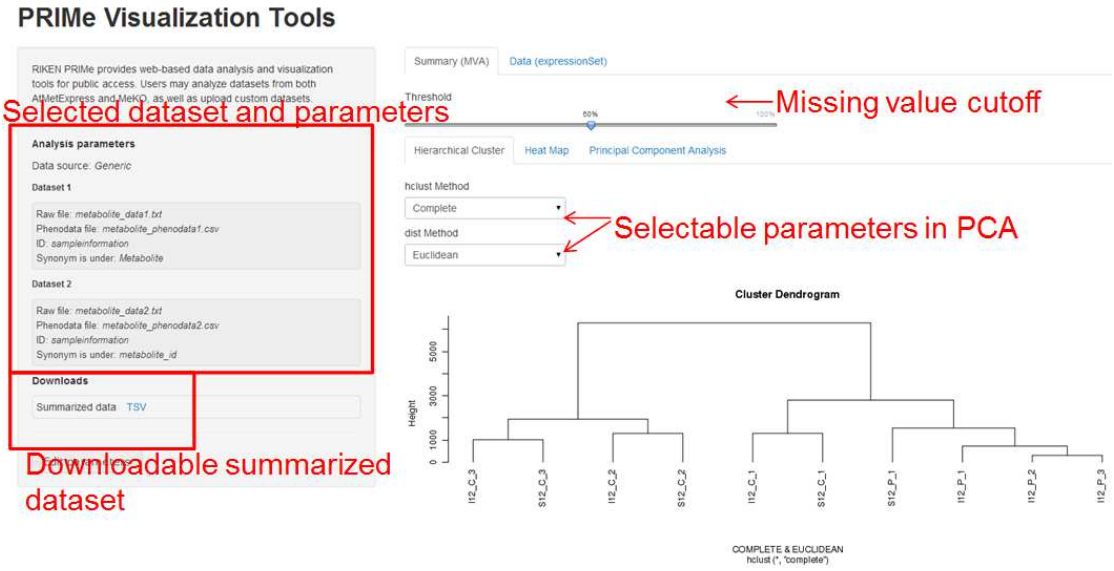
Our concatenation/summarization steps require a unique identifier for this process. In this case, users should select ‘sampleinformation.’ Please note the name of the metabolite identifier in the dataset. In the case of Dataset 1, ‘Metabolite’ is the metabolite identifier; it is ‘metabolite_id’ for Dataset 2.

(5) Select another tab and choose files.

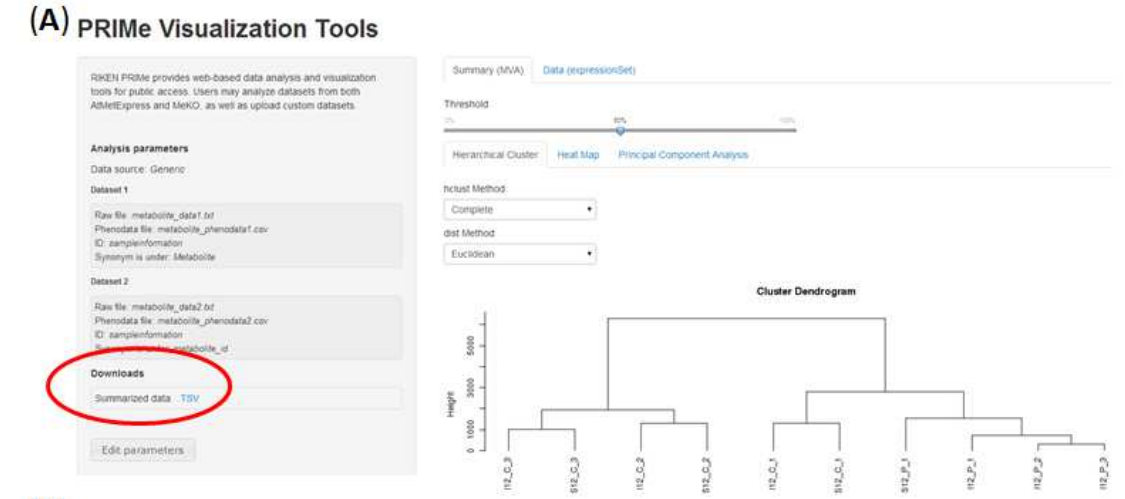


(6) Run analysis

In the current situation, users can concatenate and summarize metabolites that are detected on two different analytical platforms according to Kusano, Redestig et al. (2011).



(7) Look at downloadable summarized dataset.



(B)

	A	B	C	D	E	F	G	H	I	J	
1	final annot	fData::linen	fData::quer	fData::know	fData::prefe	fData::sour	fData::action	I12_C_1	I12_C_2	I12_C_3	I12
2	IAA	Data1_2 X_	IAA	TRUE	indole-3-a	Data1	COR_OK:PC1	61.99	51.98	52.19	
3	IAAsp	Data1_3	IAAsp	TRUE	0	Data1		27.08	22.47	52.25	
4	IAIle	Data1_4	IAIle	TRUE	0	Data1		0	0	0	
5	IALeu	Data1_5	IALeu	TRUE	0	Data1		6.21	5.11	7.2	
6	ABA	Data1_6 X_	ABA	TRUE	abscisic ac	Data1	COR_OK:PC1	155.7142	162.3663	149.9324	12
7	GA3	Data1_8	GA3	TRUE	0	Data1		6.39	6.29	5.93	
8	GA4	Data1_9 X_	GA4	TRUE	gibberellin	Data1	NA_NG:DROPPED	5.81	1.53	0.87	
9	GA19	Data1_10	GA19	TRUE	0	Data1		31.63	25.11	30.01	
10	GA20	Data1_11	GA20	TRUE	0	Data1		1.84	1.99	2.09	
11	GA44	Data1_12	GA44	TRUE	0	Data1		16.16	11.01	14.21	
12	GA53	Data1_13	GA53	TRUE	0	Data1		19.42	14.42	18.6	
13	tZ	Data2_2	tZ	TRUE	trans-zeati	Data2	COR_NG:KEPT	2.078	1.65	1.198556	
14	tZR	Data2_3	tZR	TRUE	trans-zeati	Data2		22.08	16.354	0	
15	tZRPs	Data2_4	tZRPs	TRUE	0	Data2		15.22	15.116	11.49	
16	cZ	Data2_5	cZ	TRUE	trans-zeati	Data2	COR_NG:KEPT	0.108	0.088	0.086667	

Figure 4. Summarized dataset. Users can upload their own data to our server in GENERIC format (Typically the rows and columns represent metabolite names and samples, respectively). Automatically integrated/summarized data are downloadable using the GENERIC tab (A). Users can see the summarized data matrix of their data using principal component analysis (B). Highly correlated metabolites between datasets are replaced by principal component 1 (e.g., IAA and ABA with red rectangles), while weakly correlated metabolites are not replaced in the same manner (e.g., two cytokinins, tZ and cZ with dashed rectangles).

Acknowledgments

We thank Ms. Ursula Petralia for editorial assistance and Dr. Tetsuya Sakurai and Mr. Yutaka Yamada (RIKEN Center for Sustainable Resource Science) for computational assistance.

References

Fukushima A, Kusano M, Redestig H, Arita M, Saito K (2009) Integrated omics approaches in plant systems biology. Curr Opin Chem Biol **13:** 532-538

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5:** R80

Jonsson P, Gullberg J, Nordstrom A, Kusano M, Kowalczyk M, Sjostrom M, Moritz T (2004) A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. Anal Chem **76:** 1738-1745

Redestig H, Fukushima A, Stenlund H, Moritz T, Arita M, Saito K, Kusano M (2009) Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. Anal Chem **81:** 7974-7980

Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol **3:** Article 3