



MetaboAnalyst 5.0

A Web-based Tool for streamlined
Metabolomics Data Analysis

2022.07.12

7. Statistical Analysis with metadata

Statistical analyses [Metadata table] that take these covariates into account can draw more robust conclusions about the relationships between the primary variable and the omics data. Statistical Analysis with metadata become available now.

Highlights:

- This module is designed to perform covariate analysis, including covariate adjustment and identifying major patterns with regard to the two given factors and their interactions.



7.0 Knowledge & Background

Covariate Analysis

- Metadata describes the data, and typically contains details on the experimental condition (i.e., treatments), sample source (i.e., species, tissue), and sample collection (i.e., location, time). Such metadata are critical for data interpretation, allowing researchers to analyze the data with respect to their biological and environmental context, and for data re-use, allowing other researchers to search for, and meaningfully compare and potentially integrate, results from across diverse studies.
- Details on the environmental context and sample source are becoming increasingly important as observational studies that collect omics data from human populations or animals outside of laboratory settings are becoming more common⁴³. In these cases, there is typically a primary variable of interest, such as presence/absence of a certain disease or exposure to a specific chemical, as well as other variables such as age, sex, or other potential factors that co-vary with the primary metadata



7.1 Start Statistical Analysis [Metadata Table]

Input Data Type	Available Modules (click on a module to proceed, or scroll down for more details)					
Raw Spectra (mzML, mzXML or mzData)				LC-MS Spectra Processing		
MS Peaks (peak list or intensity table)			Functional Analysis	Functional Meta-analysis		
Annotated Features (compound list or table)		Enrichment Analysis	Pathway Analysis	Joint-Pathway Analysis	Network Analysis	
Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Power Analysis	Other Utilities

Click here to start



7.2 Data Upload Page

Upload your data and metadata

- Data table contains compound concentrations or peak intensities for multiple samples;
- Metadata table contains various descriptors of these samples and can be discrete or continuous;
- Data and metadata tables must use the same sample IDs;
- For detailed instructions, please visit our [Data Formats](#) page (under **Metadata/Time-series** tab)

Data Type: Concentrations Spectral bins Peak intensities

Study Design: Multiple factors / covariates

Data Format: Samples in columns

Data File:

Metadata File:

Try our test data

Data	Study Design	
<input checked="" type="radio"/> Data Metadata	Multiple factors / covariates	meta-data of 20 healthy and 39 COVID-19 individuals with 10 categorical and 1 numeric.
<input type="radio"/> Data Metadata	Multiple factors / covariates	LC-MS peak intensity data from plasma samples of 175 individuals to study trichloroethylene (TCE) exposure. Nine metadata - 6 categorical and 3 numeric. Please refer to Walker D. et al for more details.
<input type="radio"/> Data Metadata	Time series + one condition	LC-MS peak intensity data collected from <i>Arabidopsis thaliana</i> during a wounding time course (four time points). WT - wild type; MT - <i>dde2-2</i> mutant. Please refer to Meinicke P. et al for more information
<input type="radio"/> Data Metadata	Time series only data	LC-MS peak intensity data collected from only wild type <i>Arabidopsis thaliana</i> during a wounding time course (four time points). Please refer to Meinicke P. et al for more information

Upload both data table and metadata file by clicking the options respectively.

Click "Submit" to continue.

TIP: A separate metadata file is required for Statistical Analysis [metadata table], in addition to a data table containing feature abundance information. The first column of the metadata must be the same sample names used by the data file. The second column contains the primary condition of interest. Other metadata (such as sex, age, or batch) can be included in the remaining columns. Both continuous and discrete variables are acceptable. MetaboAnalyst cannot deal with missing values (NA or empty) in the metadata. Users will be asked to manually fix these values if detected.

7.3 Data integrity check

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.

Samples are in columns and features in rows.

The uploaded data file contains 175 (samples) by 7830 (peaks(mz/rt)) data matrix.

Samples are not paired.

3 groups were detected from primary meta-data factor: TCE_Exp_Category.

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 193204 (14.1%) missing values were detected.

By default, missing values will be replaced by 1/5 of min positive values of their corresponding variables

Click the **Proceed** button if you accept the default practice;

Or click the **Missing Values** button to use other methods.

Edit Groups Missing Values **Proceed**

Click "**Proceed**" to continue.

7.4 MetaData integrity check

Meta-data check

1. Infer the type (categorical or continuous) for each metadata column;
2. For categorical metadata, at least two groups and three replicates per groups are required;
3. For continuous metadata, all values must be numerical;
4. Missing values are not allowed - you can manually add those missing values, or exclude the metadata from further analysis

Metadata processing information:

Checking metadata content passed.

A total of 8 metadata factors were detected: TCE_Exp_Category, TCE_Exp_Conc, Age, Sex, Smoking_Status, Alcohol_Use, BMI, Batch.

The primary metadata factor is: TCE_Exp_Category, which contains 3 groups.

TCE_Exp_Conc, Age, BMI meta-data factors are assigned to be continuous and remaining are categorical.

For categorical metadata, at least **two** groups with **three replicates** per group are required.

Please double check if these auto-assigned metadata types are correct.

You can manually update the metadata using the table below.

Update your metadata using the table below

- Update metadata type: categorical option for experimental groups (i.e. control vs diseased), continuous for numerical measures (i.e. age);
- Edit metadata content: click **Edit** to modify underlying groups to address those that do not meet requirements.
- Modify metadata name: click on corresponding cell on the main table to modify name
- Specify group order of categorical metadata: click **Edit** and go to **Order** tab to specify the order which the underlying metadata groups are (i.e. low, medium, high)
- Exclude metadata that do not pass sanity check.

Name	Status	Type	Edit	Remove
TCE_Exp_Category	OK	Categorical	Edit	
TCE_Exp_Conc	OK	Continuous	Edit	
Age	OK	Continuous	Edit	
	OK	Categorical	Edit	
	OK	Categorical	Edit	

You can change the levels' order or name of the metadata.

Edit metadata

Edit (sample-level) | **Order (factor-level)** | Edit (factor-level)

Available

- Low
- Moderate
- High

Update Cancel

Click **"Edit"** to edit the metadata information.

Remember to make sure the types of the factors are specified correctly.

Click **"Proceed"** button from the bottom to continue.

7.5 MetaData integrity check

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 [Lackstadi, 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 interquartile 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. The maximum allowed number of variables is 5000. [For power analysis, the max number is 2500](#) to improve power and to control computing time. Over that, the IQR filter will still be applied to keep only top maximum features, even if you choose None option.

Filtering features if their RSDs are > % in QC samples

None (less than 5000 features)

Interquartile 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

Click "**Proceed**" to continue.

Normalization Overview:

Show R Commands

The normalization procedures are grouped into three categories. You can use one or combine them to achieve better results.

- Sample normalization is for general-purpose adjustment for systematic differences among samples;
- Data transformation applies a mathematical transformation on individual values themselves. A simple mathematical approach is used to decrease the square root ([FAQ #18](#));
- Data scaling adjusts each variable/feature by a scaling factor computed based on the dispersion of the variable.

OK
You can click **View Result** button to view the effect, or **Proceed** button to analysis page!

Sample normalization

- None
- Sample-specific normalization (i.e. weight, volume) [Specify](#)
- Normalization by sum
- Normalization by median
- Normalization by a reference sample (PQN) [Specify](#)
- Normalization by a pooled sample from group (group PQN) [Specify](#)
- Normalization by reference feature [Specify](#)
- Quantile normalization (suggested only for > 1000 features)

Data transformation

- None
- Log transformation (base 10)
- Square root transformation (square root of data values)
- Cube root transformation (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

Click "**Proceed**" to continue.



7.6 Selection of Analyzing Items

Select an analysis path to explore :

Data and Metadata Overview

[Metadata Visualization](#)

Users can explore the metadata patterns and correlations through intuitive graphics. It is very useful for users to identify highly dependent metadata and quickly assess the overall patterns of the metadata.

[Interactive PCA Visualization](#)

Users can visualize data using different colors or shapes based on selected metadata in a 2D and 3D (interactive) PCA plots. It is very useful to detect overall patterns of data with regard to different metadata.

[Hierarchical Clustering and Heatmap Visualization](#)

This method displays data and metadata in the form of colored cells. It provides direct visualization of feature abundances across different samples and metadata.

Univariate Analysis

[Linear Models with Covariate Adjustment](#)

This approach uses linear models (lmma or lm) to perform significance testing with covariate adjustments. Users can choose different metadata to be included in the analysis.

[Correlation and Partial Correlation Analysis](#)

This approach allows users to explore the correlations or partial correlations (with covariate adjustments) between metabolomics features and different metadata of interest.

[Two-way ANOVA \(ANOVA2\)](#)

This approach provides classical two-way ANOVA based on the two factors selected by users. For time-series data, users should choose within-subjects ANOVA.

Multivariate Analysis

[ANOVA Simultaneous Component Analysis \(ASCA\)](#)

This approach is designed to identify major patterns with regard to the two given factors and their interaction. The implementation was based on the algorithm described by [AK Smilde, et al.](#) with additional improvements on feature selection and model validation.

Multivariate Empirical Bayes Analysis of Variance (MEBA) for Time Series

This approach is designed to compare temporal profiles across different biological conditions. It is based on the timecourse method described by [YC Tai, et al.](#)

Supervised Classification

[Random Forest](#)

This machine learning approach is designed to perform classification and feature selection analysis. Users can also test contribution of meta-data to class prediction.

Data and Metadata Overview

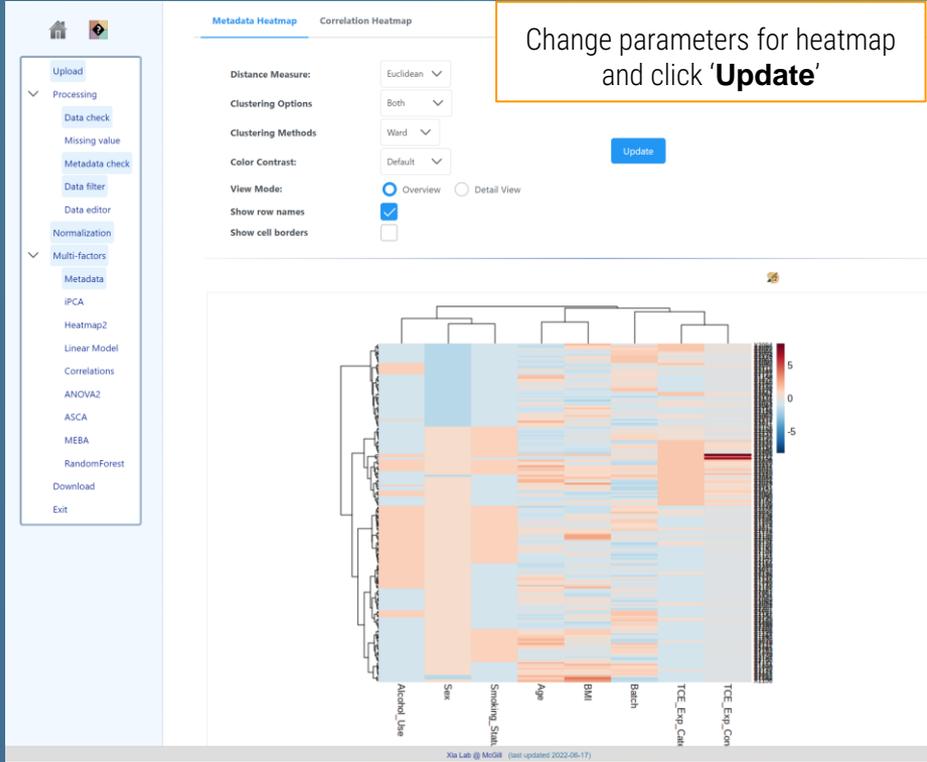
Univariate Analysis

Multivariate Analysis

Supervised Classification



7.6 Metadata Visualization



Change parameters for heatmap and click **'Update'**

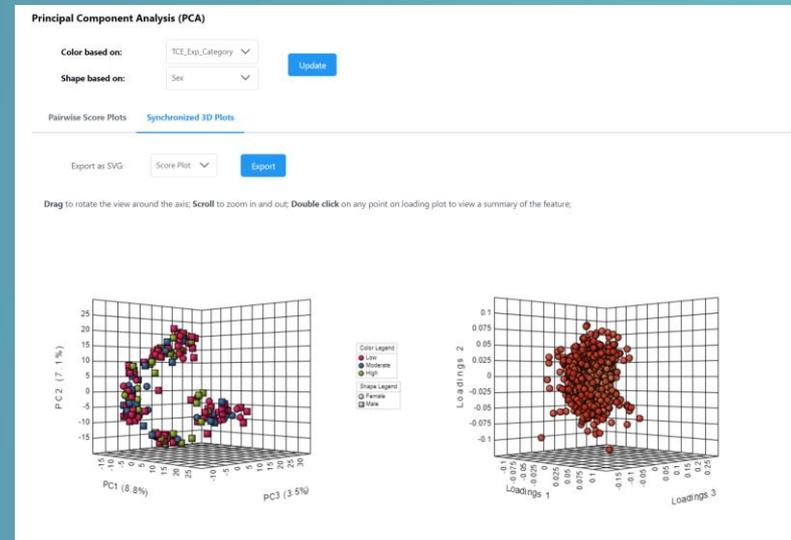
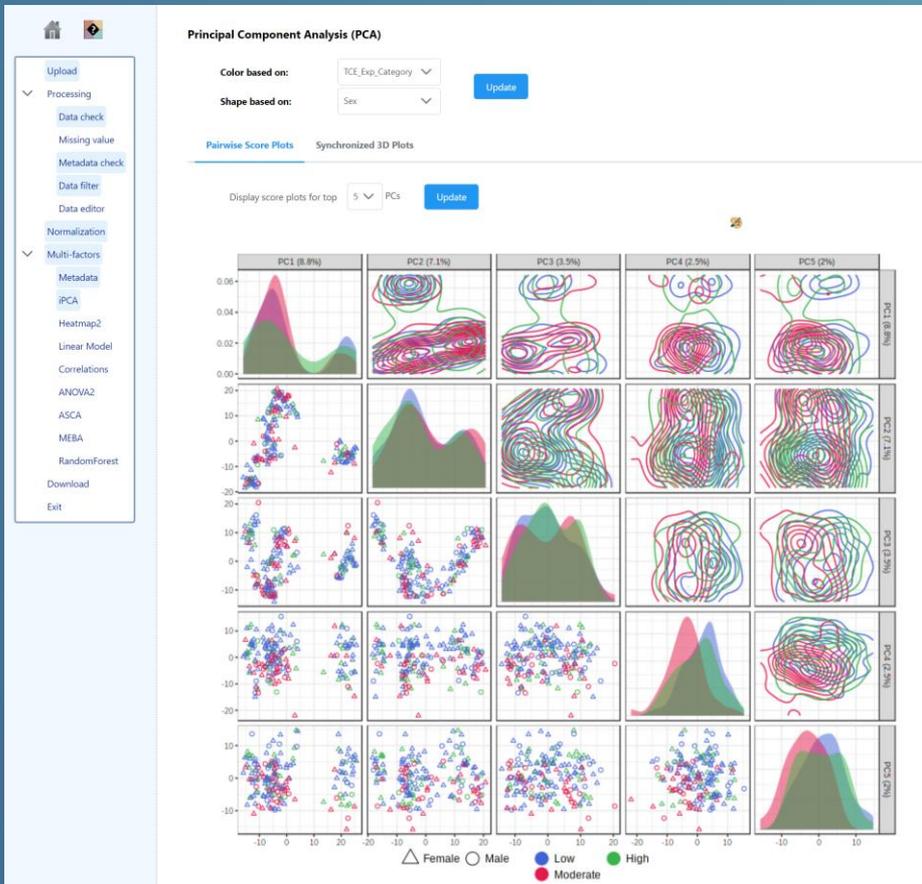
Heatmap of metadata



Change correlation method and Click **'Update'**

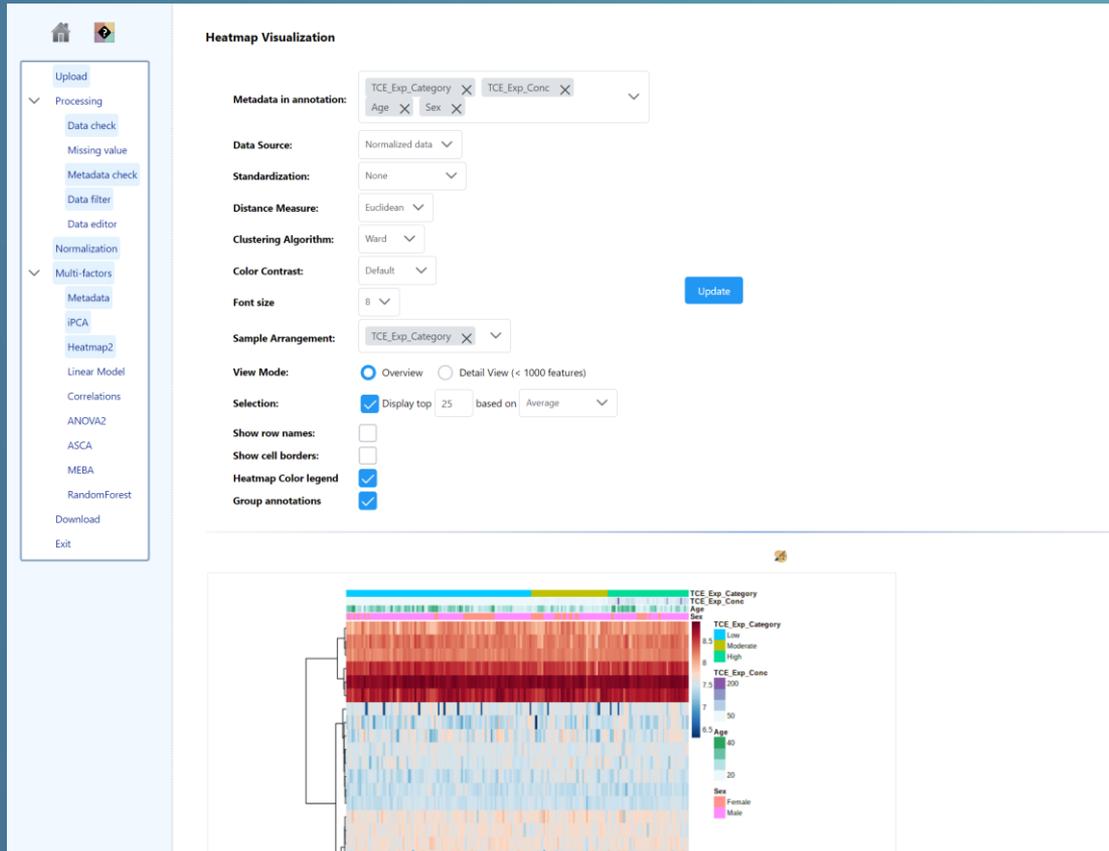
Correlation analysis of covariates

7.7 Interactive PCA



TIP: Principal Component Analysis (PCA) on the data with multiple factors. For the score plot, the PCA are shown as scatter plot and contour plot. The scatter plot shows the details of all samples, while contour plot shows the distribution center of all samples. The density plot on the diagonal display the density of different groups at the corresponding PC based on the primary factor. A pair of dynamic 3D PCA can be explored and rotated easily.

7.8 Hierarchical Clustering and Heatmap Visualization



TIP: Clustering of the data can be performed with different methods or based on different factors. Edit the parameters and Click 'Update' to view the heatmap



7.9 Covariate Adjustment with Linear Models

Show R Commands

Linear models with covariate adjustments

The underlying method is based on [limma](#) for its high-performance implementation. Some data may include some form of blocking in the study design, which can be modeled as either fixed or random effects. Please note that although you can specify a blocking factor (to be modeled as random effects), we in general recommend **keeping this option unspecified** (the default). Using fixed effect model not only is computationally more efficient, but also gives results that are more consistent with the interpretation of differences. Please refer to the excellent book by [Paul D. Allison \(2009\)](#) for more technical discussions.

Primary metadata: TCE_Exp_Category

Covariates (control for): Age X Sex X Batch X

Blocking factor: -- Unspecified --

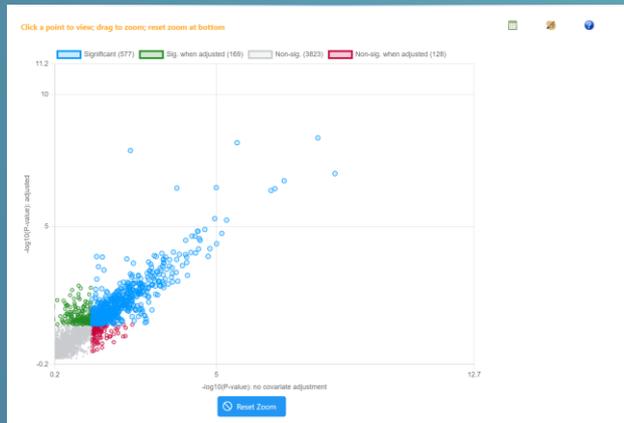
P-value cutoff: 0.05

Reference group: Low

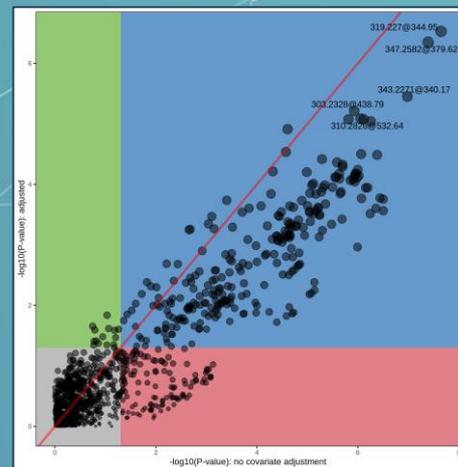
Submit

Select appropriate variables then click **'Submit'**

TIP: Find metabolites are significantly associated with your variable of interest ("Primary metadata"), after adjusting for covariates ("Include metadata"). After selecting the appropriate variables, click **"Submit"**.



Click the paint icon to generate a high resolution results figure



7.10 Correlation Analysis

Correlation and Partial Correlation Analysis

Correlation analysis can be performed for a given feature and metadata of metadata of interest.

- When the covariate is "none" (default), regular correlation analysis will be performed; otherwise, partial correlation analysis will be performed.
- For binary metadata, the point biserial correlation will be used, for continuous metadata, users can choose Pearson or Spearman correlation.

Target of interest

a metadata: TCE_Exp_Category

a feature: []

Covariates of interest

[]

Correlation measure

Pearson r

Submit

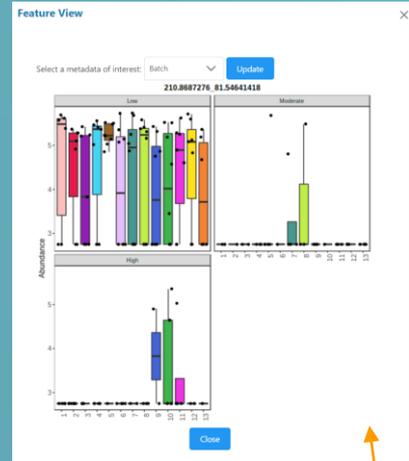
Top 25 metadata correlated with the TCE_Exp_Category

Metadata ID	Correlation Coefficient
249_06092_61_73262	0.45
206_1002143_102_72	0.42
408_6799592_56_535	0.40
190_911905_49_0505	0.38
401_3442333_363_10	0.35
230_8099491_55_186	0.32
346_7266821_56_424	0.30
311_2203343_386_23	0.28
400_3405201_362_98	0.25
380_9504047_58_130	0.22
122_9241151_50_097	0.20
290_7647632_55_201	0.18
296_2733348_449_70	0.15
415_8947902_54_119	0.12
288_7686712_55_287	0.10
170_8537487_54_748	0.08
310_9537414_46_609	0.05
264_8338287_78_718	0.02
282_9111679_106_17	0.00
266_8301722_79_036	-0.02
224_8456068_88_394	-0.05
208_8719473_81_130	-0.08
206_874961_81_2597	-0.10
222_8486453_80_724	-0.12
210_8687276_81_546	-0.15

Select appropriate variables and parameters then click 'Submit'

Click table spread icon to view complete correlation results

Top 25 most correlated peaks are plotted



Feature Details Table

Click a feature name to edit its name and then click the next column to save the change. Click the view link to visualize a graphical summary of the distribution. The bar plots on the left show the original values (lines: +, -). The box and whisker plots on the right summarize the normalized values. Note: positive integer numbers are represented as 99999, and negative integer numbers as -99999.

To update a name suitable for graphical display, click the name to edit and then click the next column to save. [Download](#)

Name	correlation	t-stat	p-value	FDR	View
210_8687276_81_54641418	0.51681	-7.9402	2.45E-10	4.5308E-10	View
222_8486453_80_72495653	-0.51462	-7.8944	3.209E-10	4.5308E-10	View
208_8719473_81_25979073	-0.51318	-7.8643	3.8315E-10	4.5308E-10	View
208_8719473_81_25943335	-0.51312	-7.8603	3.8358E-10	4.5308E-10	View
224_8456068_88_38429634	-0.50899	-7.7777	6.371E-10	5.985E-10	View
246_8501722_78_93923212	-0.48335	-7.2621	1.248E-09	9.7803E-09	View
382_9111679_106_17845377	-0.46882	-6.9773	6.1396E-09	6.1189E-08	View
264_8301722_79_71823487	-0.42705	-6.2381	3.5198E-09	2.0671E-08	View
249_86876_01_72026261	0.34829	4.887	2.3229E-09	0.0017123	View
206_1002143_102_7247637	0.33916	6.7421	4.403E-09	0.0020981	View
408_6799592_56_53581003	0.32262	6.6437	6.7451E-09	0.0028802	View
190_911905_49_05059372	0.30876	4.385	1.8276E-05	0.01185	View
401_3442333_363_1064875	0.30584	4.2221	3.8028E-05	0.014101	View

Click 'View' for comparison of abundance

7.11 Two-way ANOVA (ANOVA2)

Type of ANOVA:

- For two-factor independent samples: Two-way (between subjects) ANOVA
 - Type I and Type III ANOVA are supported, the main difference is whether the means are weighted (type I) or unweighted (type III).
 - For unbalanced data, there may be significant difference between type I and type III.
- For time-series only: One-way repeated measures (within subjects) ANOVA
 - The data must have a balanced design (equal sample sizes)
- For time-series + one experimental factor: Two-way repeated measures (within subjects) ANOVA
 - By default, the ANOVA model will consider the interaction (**Time + Treatment**).
 - If your data does not have sufficient residual degrees of freedom to estimate the interactions, you may want to leave out the interaction term (**Time + Treatment**)

Select two metadata:

Choose ANOVA Type:

 (for two-way independent ANOVA)

Consider interactions:

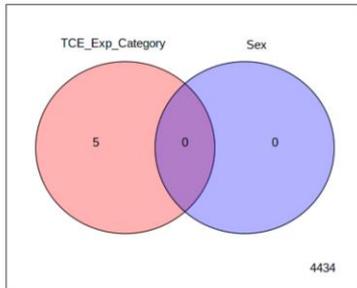
 (for two-way repeated ANOVA)

Adjusted p-value cutoff:

Multiple testing correction:

Submit

Two-way ANOVA (between subjects)



Select appropriate variables and parameters then click **'Submit'** to update the Venn Diagram below. Click table spread icon to view more details.

Feature Details Table

Click a feature name to edit its name and then click the next column to save the change. Click the view link to visualize a graphical summary of the distribution. The bar plots on the left show the original values (mean +/- SD). The box and whisker plots on the right summarize the normalized values. Note, positive infinite numbers are represented as 999999, and negative infinite numbers as -999999.

To update a name suitable for graphical display, **click the name** to edit and then click the next column to save

Download

Name	TCE_Exp_Category(F.val)	TCE_Exp_Category(raw.p)	TCE_Exp_Category(adj.p)	Sex(F.val)	Sex(raw.p)	Sex(adj.p)	
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
206.1002143_102.7247657	11.111	2.9278E-5	0.12997	9.3933E-4	0.97559	0.99741	View
542.9252101_47.92005778	10.221	6.4518E-5	0.13762	0.17591	0.67544	0.98551	View
249.06092_61.73262631	9.806	9.3499E-5	0.13762	0.40834	0.52368	0.9823	View
155.0907026_252.6532109	9.3677	1.3861E-4	0.13762	4.5373	0.03461	0.84297	View
400.3405201_362.9808888	9.2435	1.5501E-4	0.13762	0.15893	0.69065	0.98551	View

<< < 1 > >> 20

7.12 ANOVA Simultaneous Component Analysis (ASCA)

Customize the parameters and click **'Submit'**.

ANOVA Simultaneous Component Analysis (ASCA)

ASCA is a direct generalization of analysis of variance (ANOVA) for univariate data to the multivariate case based on [AK Smilde et al.](#), with additional model validation. ASCA is commonly used to model two main effects with one interaction effect. The key parameters are the number of components.

- First select two metadata of interest and click Submit button using the default parameters. Based on the Scree plots (first tab), you can update the component numbers.
- Use 0 if you don't want to model the interaction or residual;
- The max. number of components cannot exceed the total number as shown in the corresponding Scree Plots

Select two metadata:

(categorical only)

TCE_Exp_Category X
Sex X

Model A (Model.a):

Model B (Model.b):

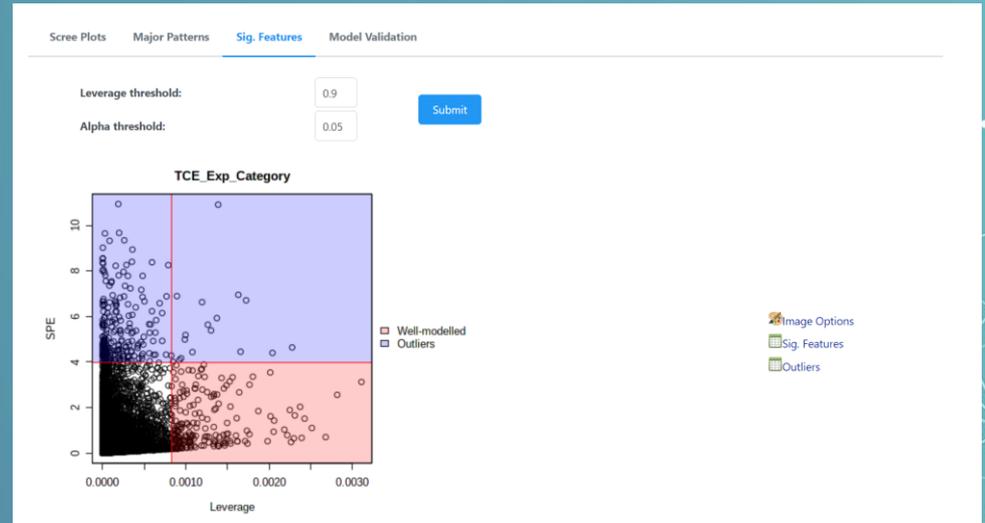
Interaction Model (Model.ab):

Residual Model (Model.res):

Submit

View results from different tabs below the parameters panel.

TIP: To fully understand the mechanism of ASCA and interpret the results, please refer to the [FAQs](#) section of MetaboAnalyst.



7.15 Multivariate empirical Bayes (MEBA) time-series analysis

Time-course profiles:

The table below shows all the features ranked by the corresponding statistics. You can click any feature name to view its corresponding time-course profile.

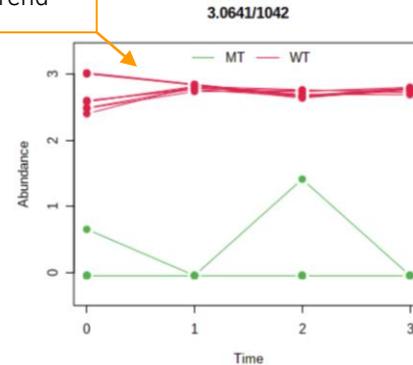
- For time-course only data, we are interested in features which **change** over time.
- For time-course + one experimental factor data, we are interested in features which **respond differently in different experimental conditions** over time.
- For time-course + one experimental factor data, the statistics will be Hotelling-T2 if there are only 2 biological conditions or MB-Statistics if there are more than 2 biological conditions.

Specify metadata: Phenotype Time Update

Customize the metadata included and click '**Update**'.

Name	Hotelling-T2
3.0641/1042	37533.7531
4.0289/873	28526.24962
1.944/1044	23123.73006
3.4428/1019	13978.74526
2.7865/1027	13056.14885
3.108/897	10969.97689
3.0263/811	9867.84225
1.8763/813	9816.35836
3.2347/1012	8767.92088
1.8927/853	7619.41511
1.6114/265	7231.37491
2.7863/1062	6396.38121
4.2231/1096	5607.72909
3.4112/898	5554.81068
4.0271/805	5270.14221
3.0321/869	5266.47452
2.6882/827	5012.98133
3.4422/926	4569.28863
3.1578/841	4288.4332

Select/Click feature name to view the time-series trend



7.14 Supervised Classification with Random Forest

Random Forest

Classification

Var. Importance

Outlier Detection

Primary metadata:

TCE_Exp_Category

Choose metadata for predictors:

Sex

Randomness:

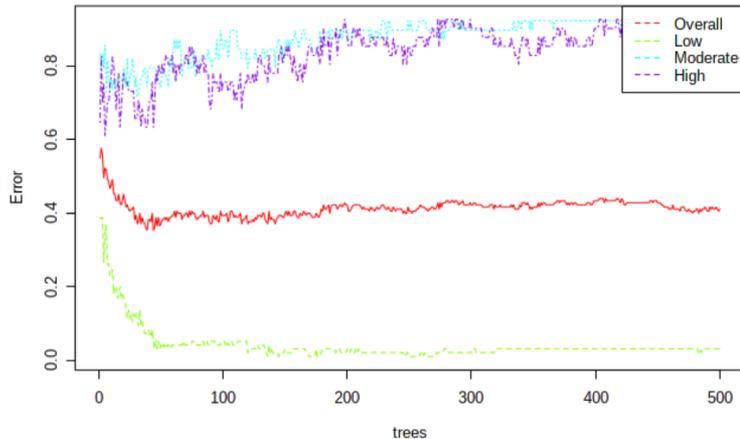
On

Leave defaults and click
'Update'

Update

The model does an o.k. job
classifying samples as either
"Low" or "Moderate / High".

Random Forest classification



The OOB error is 0.411

	Low	Moderate	High	class.error
Low	92	1	2.0	0.0316
Moderate	33	5	1.0	0.872
High	33	2	6.0	0.854

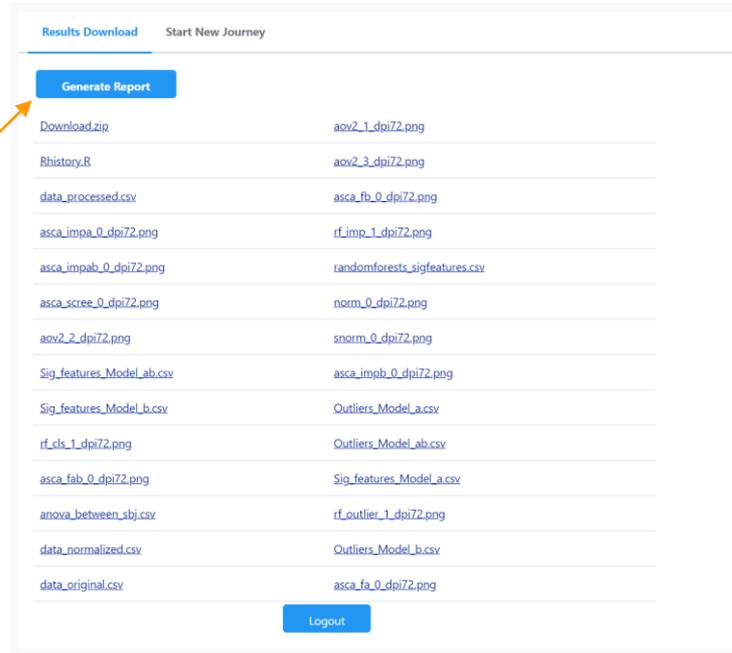
TIP: To build a predictive model with only metabolites features, leave the default settings. If you'd like to include certain covariates in the model, add them using the "Choose metadata for predictors" dropdown.

Look at the results overview

6.4 Result Downloading

Download Results & Start New Journey

Please download the results (tables and images) from the **Results Download** tab below. The **Download.zip** contains all the files in your home directory. You can also generate a **PDF analysis report** using the button. Finally, you can continue to explore other compatible modules using the **Start New Journey** tab.



The screenshot shows a web interface with two tabs: 'Results Download' (active) and 'Start New Journey'. A blue button labeled 'Generate Report' is positioned above a list of files. An orange arrow points from a text box on the left to this button. The file list consists of two columns of links. At the bottom of the list is a blue 'Logout' button.

File Name	File Name
Download.zip	aov2_1_dpi72.png
Rhistory.R	aov2_3_dpi72.png
data_processed.csv	asca_fb_0_dpi72.png
asca_impaa_0_dpi72.png	rf_imp_1_dpi72.png
asca_impab_0_dpi72.png	randomforests_sigfeatures.csv
asca_scree_0_dpi72.png	norm_0_dpi72.png
aov2_2_dpi72.png	snorm_0_dpi72.png
Sig_features_Model_ab.csv	asca_impb_0_dpi72.png
Sig_features_Model_b.csv	Outliers_Model_a.csv
rf_cls_1_dpi72.png	Outliers_Model_ab.csv
asca_fab_0_dpi72.png	Sig_features_Model_a.csv
anova_between_sbjs.csv	rf_outlier_1_dpi72.png
data_normalized.csv	Outliers_Model_b.csv
data_original.csv	asca_fa_0_dpi72.png

Click the **“Generate Report”** to download a pdf report summarizing your analysis.

Thanks

*If you have any questions please read through the FAQs or contact us at
[Zhiqiang.pang\[at\]xialab.ca](mailto:Zhiqiang.pang[at]xialab.ca) or [jeff.xia\[at\]xialab.ca](mailto:jeff.xia[at]xialab.ca)*

