19th Annual Conference of the Metabolomics Society
METABOLOMICS 2023
NIAGARA FALLS, CANADA JUNE 18-22

MetaboAnalyst 5.0

Statistical analysis (II)

-- complex experimental designs

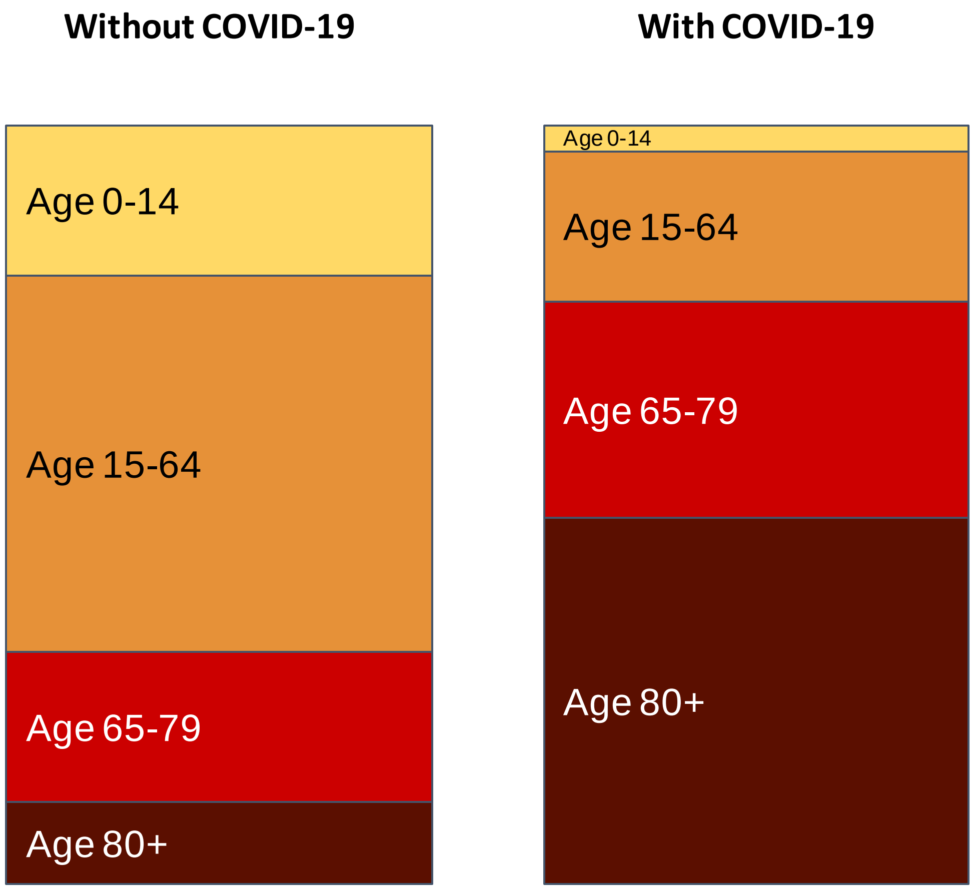
Jessica Ewald, Postdoctoral Fellow
jessica.ewald@mcgill.ca
McGill University, Montreal, QC Canada

Why complex meta-data?

- More studies are collecting samples from the “real world”:
  - Epidemiology
  - Field studies
- These samples have many uncontrolled variables (covariates):
  - Biological: sex, age, disease status
  - Environmental: location, lifestyle, temperature
- Taking covariates into account can increase statistical power and reduce confounding relationships between primary variable of interest and ‘omics data

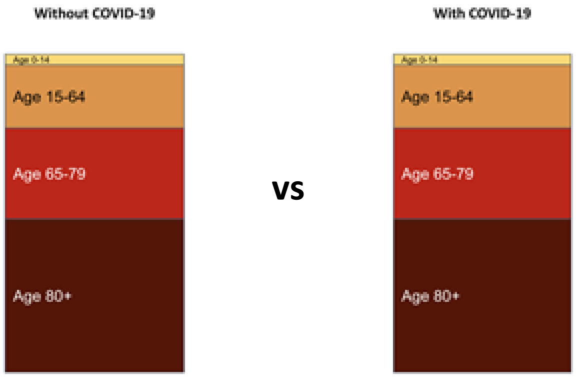
Confounding factors: example

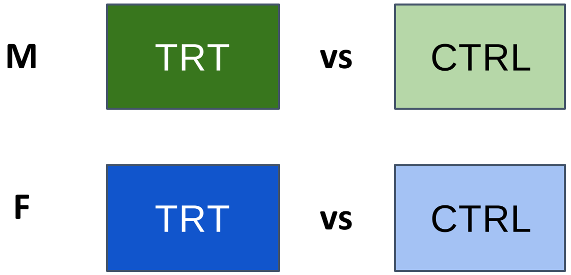
- Objective: identify metabolites associated with COVID-19
- Data: blood samples from subjects with and without COVID-19
  - Not a controlled experiment
  - Early in pandemic, more COVID-19 diagnoses in older subjects
- Problem: difficult to distinguish metabolites related to age vs. related to COVID-19



Strategies for dealing with meta-data

- Experimental design:
  - **Control** group that 'matches' your group of interest
- Data analysis (often limited by sample size):
  - **Stratification** - for factors with few classes, split up the data and analyze separately
  - **Take into account** - statistical Analysis [metadata table] module in MetaboAnalyst

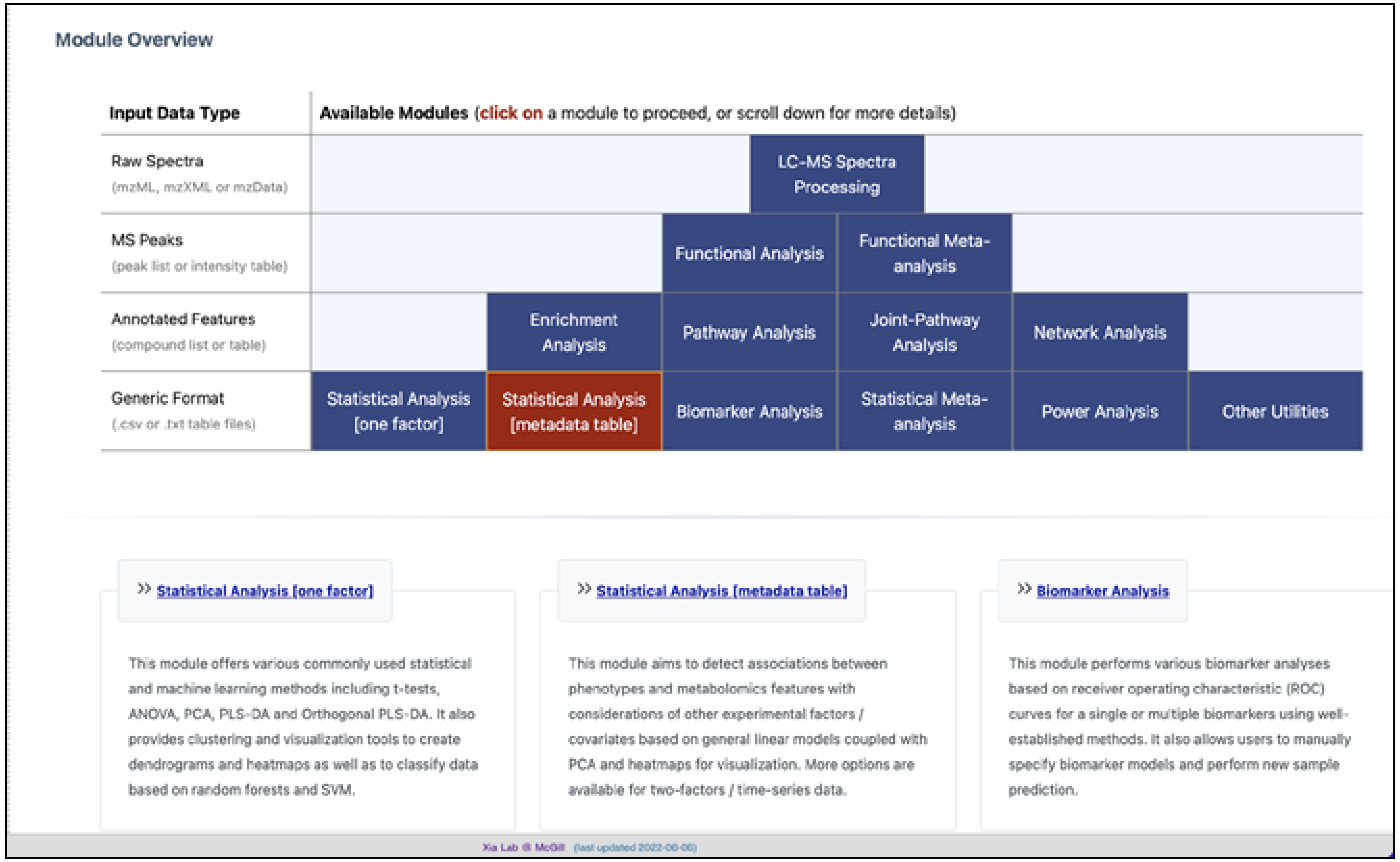




$$y = b_0 + b_1*x_1 + b_2*x_2 + ... + b_n*x_n$$

Complex meta-data with MetaboAnalyst

- We will cover how to:
    - Format meta-data
    - Analyze relationships between meta-data
    - Covariate adjustment with linear model
    - Supervised analysis with Random Forest



Module Overview

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

>> Statistical Analysis [one factor]

This module offers various commonly used statistical and machine learning methods including t-tests, ANOVA, PCA, PLS-DA and Orthogonal PLS-DA. It also provides clustering and visualization tools to create dendrograms and heatmaps as well as to classify data based on random forests and SVM.

>> Statistical Analysis [metadata table]

This module aims to detect associations between phenotypes and metabolomics features with considerations of other experimental factors / covariates based on general linear models coupled with PCA and heatmaps for visualization. More options are available for two-factors / time-series data.

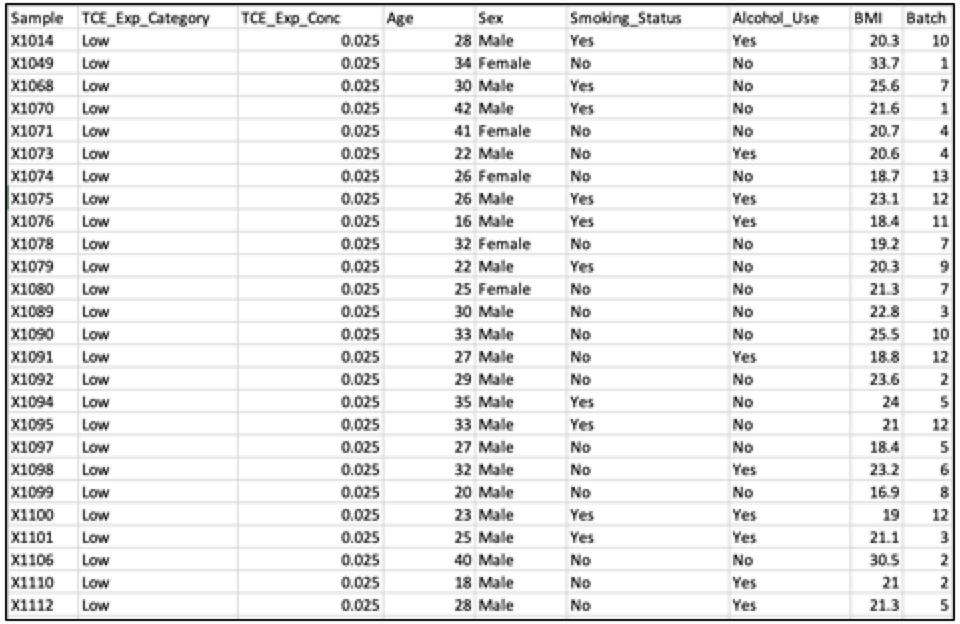
>> Biomarker Analysis

This module performs various biomarker analyses based on receiver operating characteristic (ROC) curves for a single or multiple biomarkers using well-established methods. It also allows users to manually specify biomarker models and perform new sample prediction.

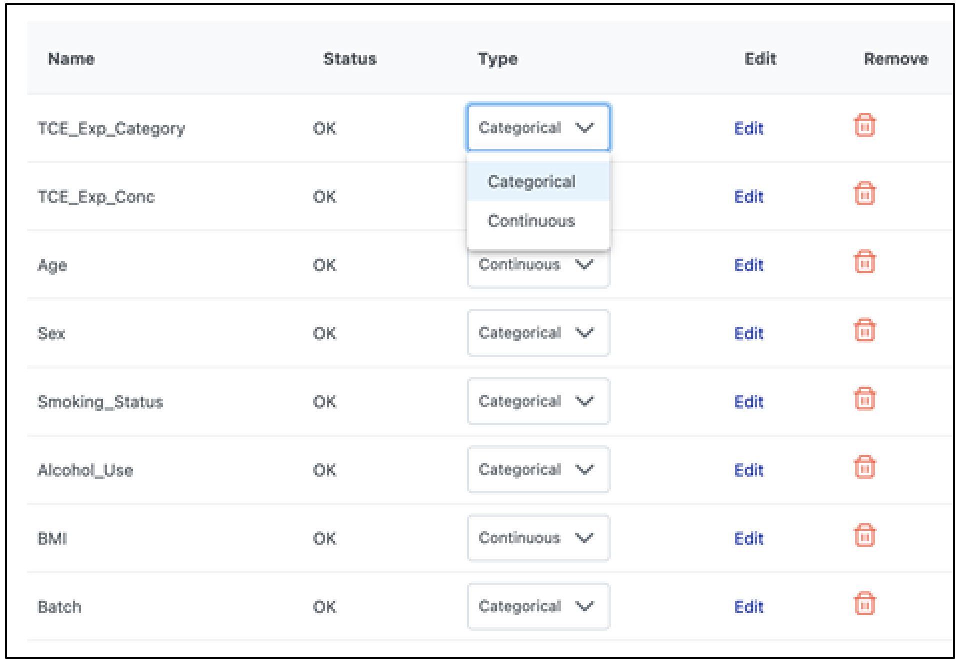
Xia Lab @ McGill (last updated 2022-06-06)

Formatting meta-data

- Essential for downstream analysis
- Categorize as “Categorical” or “Continuous”
    - For categorical, must have at least 2 groups with at least 3 replicates each
- Missing values will be filtered out during analysis
- First meta-data column will be considered primary variable by default



Sample	TCE_Exp_Category	TCE_Exp_Conc	Age	Sex	Smoking_Status	Alcohol_Use	BMI	Batch
X1014	Low	0.025	28	Male	Yes	Yes	20.3	10
X1049	Low	0.025	34	Female	No	No	33.7	1
X1068	Low	0.025	30	Male	Yes	No	25.6	7
X1070	Low	0.025	42	Male	Yes	No	21.6	1
X1071	Low	0.025	41	Female	No	No	20.7	4
X1073	Low	0.025	22	Male	No	Yes	20.6	4
X1074	Low	0.025	26	Female	No	No	18.7	13
X1075	Low	0.025	26	Male	Yes	Yes	23.1	12
X1076	Low	0.025	16	Male	Yes	Yes	18.4	11
X1078	Low	0.025	32	Female	No	No	19.2	7
X1079	Low	0.025	22	Male	Yes	No	20.3	9
X1080	Low	0.025	25	Female	No	No	21.3	7
X1089	Low	0.025	30	Male	No	No	22.8	3
X1090	Low	0.025	33	Male	No	No	25.5	10
X1091	Low	0.025	27	Male	No	Yes	18.8	12
X1092	Low	0.025	29	Male	No	No	23.6	2
X1094	Low	0.025	35	Male	Yes	No	24	5
X1095	Low	0.025	33	Male	Yes	No	21	12
X1097	Low	0.025	27	Male	No	No	18.4	5
X1098	Low	0.025	32	Male	No	Yes	23.2	6
X1099	Low	0.025	20	Male	No	No	16.9	8
X1100	Low	0.025	23	Male	Yes	Yes	19	12
X1101	Low	0.025	25	Male	Yes	Yes	21.1	3
X1106	Low	0.025	40	Male	No	No	30.5	2
X1110	Low	0.025	18	Male	No	Yes	21	2
X1112	Low	0.025	28	Male	No	Yes	21.3	5



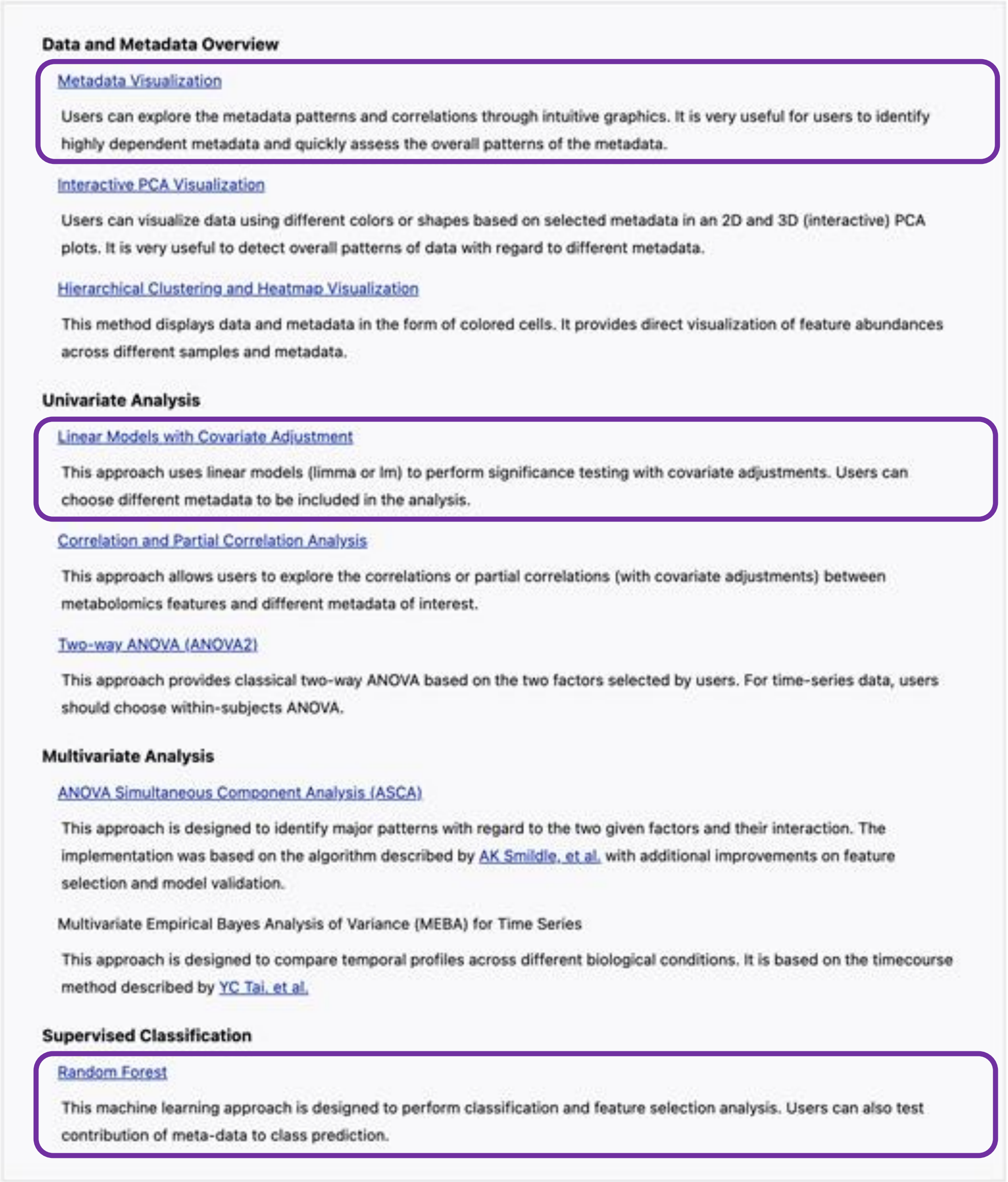
Name	Status	Type	Edit	Remove
TCE_Exp_Category	OK	Categorical	Edit	🗑
TCE_Exp_Conc	OK		Edit	🗑
Age	OK	Continuous	Edit	🗑
Sex	OK	Categorical	Edit	🗑
Smoking_Status	OK	Categorical	Edit	🗑
Alcohol_Use	OK	Categorical	Edit	🗑
BMI	OK	Continuous	Edit	🗑
Batch	OK	Categorical	Edit	🗑

Method overview

Overview to understand structure of data & metadata

Simple yet effective univariate statistical analysis

Advanced multivariate statistics & machine learning



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 an 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 (limma 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.

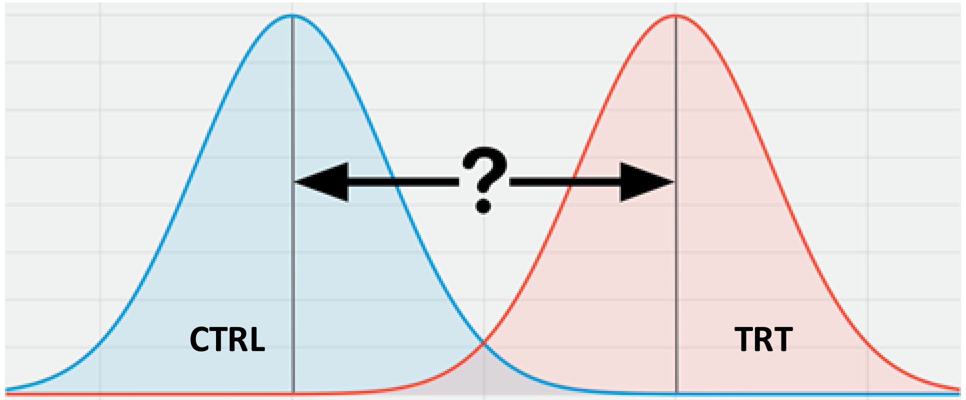
Supervised Classification

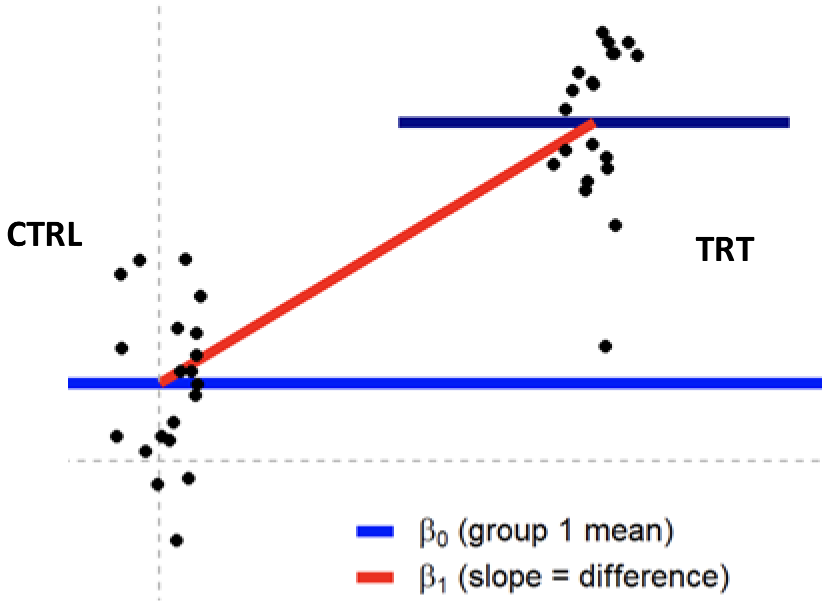
Random Forest

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

T-test vs. linear regression

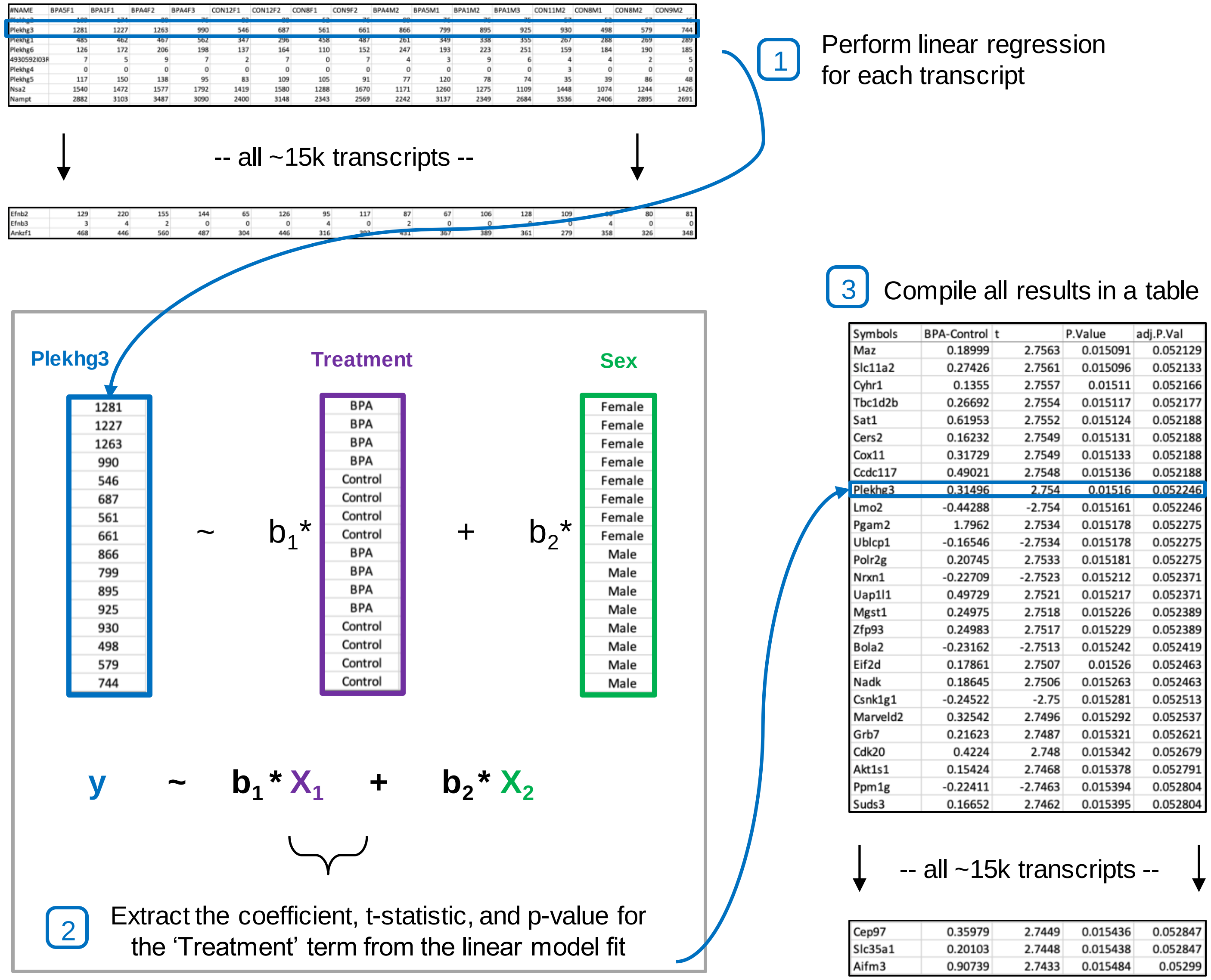
- You can do t-test with linear regression:
- $y = B_0 + B_1 * x$
  - y: level of metabolite A
  - x: variable of interest





More complex experimental design

- Treatment (Control/BPA)
- Sex (male/female)
- Features associated with ***treatment*** while considering ***sex***
- Treatment = primary
- Sex = covariate
- Can include many covariates
- Primary/covariate can be continuous or discrete



#NAME	BPA5F1	BPA1F1	BPA4F2	BPA4F3	CON12F1	CON12F2	CON8F1	CON9F2	BPA4M2	BPA5M1	BPA1M2	BPA1M3	CON11M2	CON8M1	CON8M2	CON9M2
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Plekhg3	1281	1227	1263	990	546	687	561	661	866	799	895	925	930	498	579	744
Plekhg1	485	462	467	562	347	296	458	487	261	349	338	355	267	288	269	289
Plekhg6	126	172	206	198	137	164	110	152	247	193	223	251	159	184	190	185
4930592I03R	7	5	9	7	2	7	0	7	4	3	9	6	4	4	2	5
Plekhg4	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
Plekhg5	117	150	138	95	83	109	105	91	77	120	78	74	35	39	86	48
Nsa2	1540	1472	1577	1792	1419	1580	1288	1670	1171	1260	1275	1109	1448	1074	1244	1426
Nampt	2882	3103	3487	3090	2400	3148	2343	2569	2242	3137	2349	2684	3536	2406	2895	2691

-- all ~15k transcripts --

Efnb2	129	220	155	144	65	126	95	117	87	67	106	128	109	[illegible]	80	81
Efnb3	3	4	2	0	0	0	4	0	2	0	0	0	0	4	0	0
Ankzf1	468	446	560	487	304	446	316	[illegible]	431	367	389	361	279	358	326	348

1 Perform linear regression for each transcript

Plekhg3	Treatment	Sex
1281	BPA	Female
1227	BPA	Female
1263	BPA	Female
990	BPA	Female
546	Control	Female
687	Control	Female
561	Control	Female
661	Control	Female
866	BPA	Male
799	BPA	Male
895	BPA	Male
925	BPA	Male
930	Control	Male
498	Control	Male
579	Control	Male
744	Control	Male

$\sim \quad b_1* \quad + \quad b_2*$

$$y \sim \underbrace{b_1 * X_1} + b_2 * X_2$$

2 Extract the coefficient, t-statistic, and p-value for the 'Treatment' term from the linear model fit

3 Compile all results in a table

Symbols	BPA-Control	t	P.Value	adj.P.Val
Maz	0.18999	2.7563	0.015091	0.052129
Slc11a2	0.27426	2.7561	0.015096	0.052133
Cyhr1	0.1355	2.7557	0.01511	0.052166
Tbc1d2b	0.26692	2.7554	0.015117	0.052177
Sat1	0.61953	2.7552	0.015124	0.052188
Cers2	0.16232	2.7549	0.015131	0.052188
Cox11	0.31729	2.7549	0.015133	0.052188
Ccdc117	0.49021	2.7548	0.015136	0.052188
Plekhg3	0.31496	2.754	0.01516	0.052246
Lmo2	-0.44288	-2.754	0.015161	0.052246
Pgam2	1.7962	2.7534	0.015178	0.052275
Ublcp1	-0.16546	-2.7534	0.015178	0.052275
Polr2g	0.20745	2.7533	0.015181	0.052275
Nrxn1	-0.22709	-2.7523	0.015212	0.052371
Uap1l1	0.49729	2.7521	0.015217	0.052371
Mgst1	0.24975	2.7518	0.015226	0.052389
Zfp93	0.24983	2.7517	0.015229	0.052389
Bola2	-0.23162	-2.7513	0.015242	0.052419
Eif2d	0.17861	2.7507	0.01526	0.052463
Nadk	0.18645	2.7506	0.015263	0.052463
Csnk1g1	-0.24522	-2.75	0.015281	0.052513
Marveld2	0.32542	2.7496	0.015292	0.052537
Grb7	0.21623	2.7487	0.015321	0.052621
Cdk20	0.4224	2.748	0.015342	0.052679
Akt1s1	0.15424	2.7468	0.015378	0.052791
Ppm1g	-0.22411	-2.7463	0.015394	0.052804
Suds3	0.16652	2.7462	0.015395	0.052804

-- all ~15k transcripts --

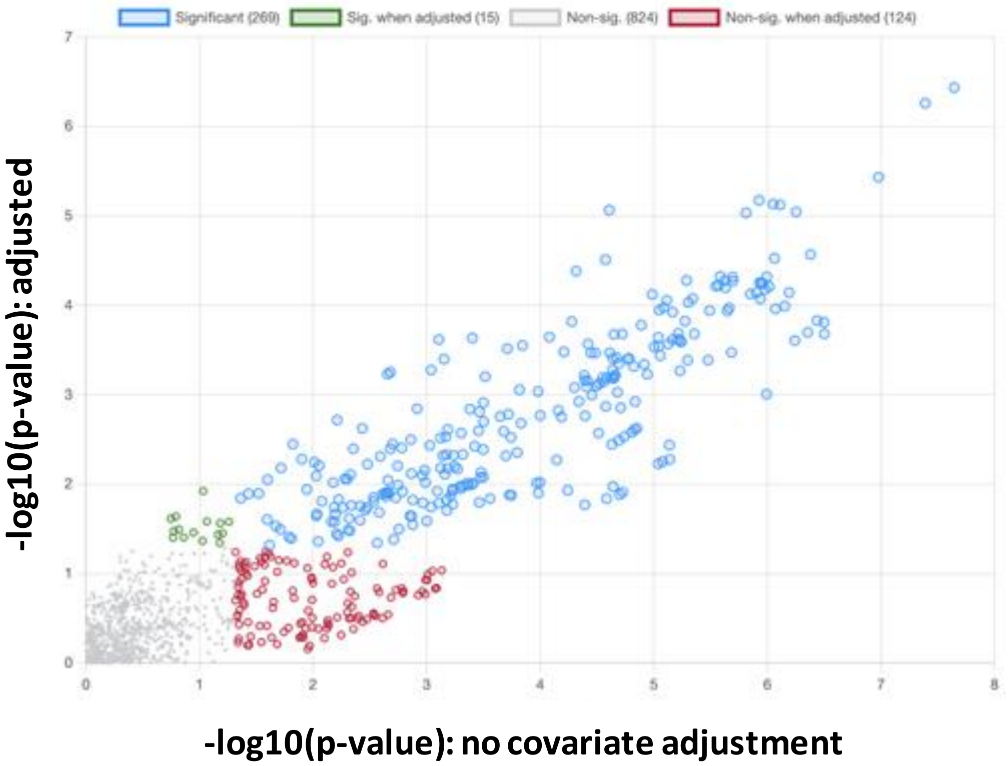
Cep97	0.35979	2.7449	0.015436	0.052847
Slc35a1	0.20103	2.7448	0.015438	0.052847
Aifm3	0.90739	2.7433	0.015484	0.05299

Interpretation of coefficient results

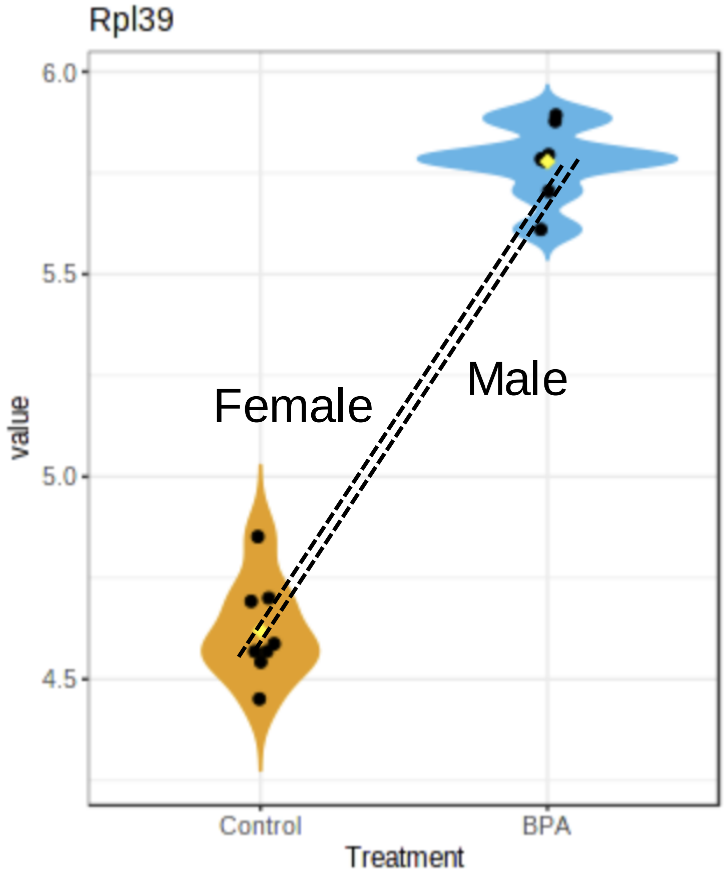
Multiple linear regression example:

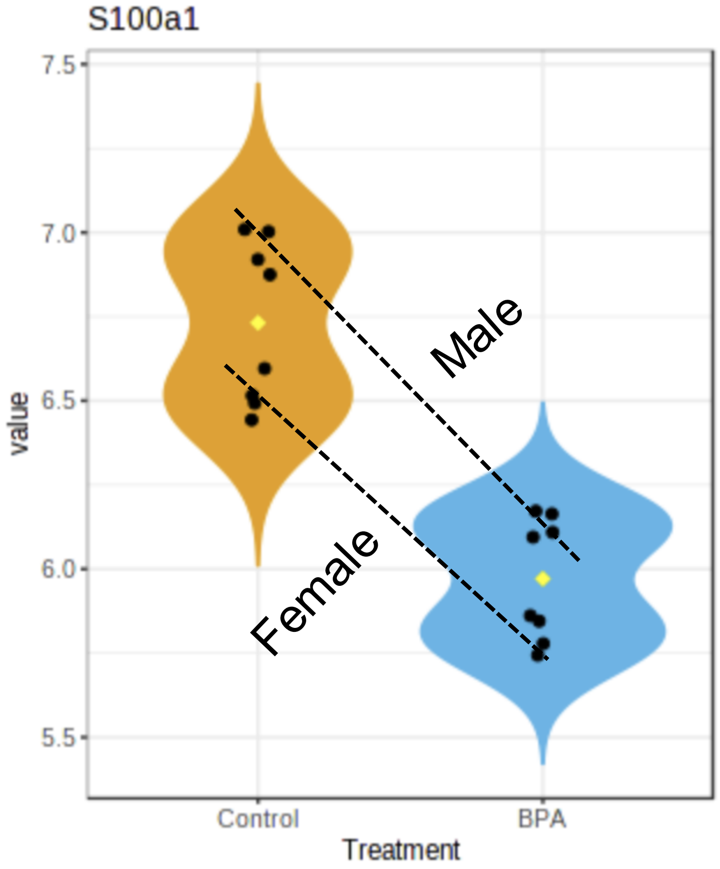
- $y = B_0 + B_1 * x_{treatment} + B_2 * x_{sex}$

- By including $B_2 * x_{sex}$ in the model, we account for effects of sex
- Extract B_1 from the model:
  - B_1 value = magnitude & direction of relationship between metabolite ‘y’ and $x_{treatment}$
  - B_1 p-value = statistical significance of relationship



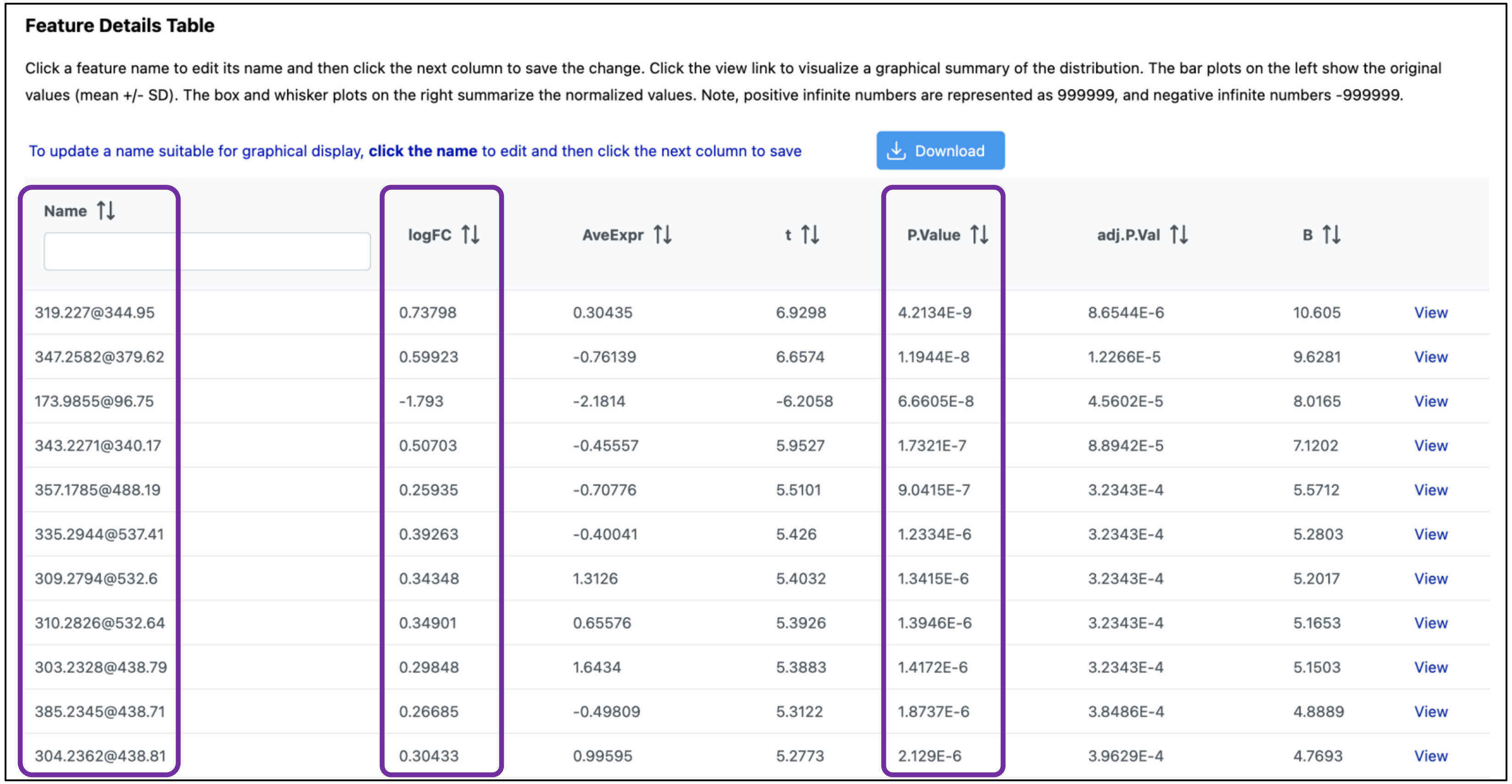
Impact of controlling for sex





Linear Model	**Rpl39 Treatment P-value**	**S100a1 Treatment P-value**
Y ~ **Treatment**	3.7×10^{-12}	2.9×10^{-6}
Y ~ **Treatment** + Sex	9.9×10^{-12}	2.1×10^{-11}

Functional Analysis



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 -999999.

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

Name ⇅	logFC ⇅	AveExpr ⇅	t ⇅	P.Value ⇅	adj.P.Val ⇅	B ⇅	
319.227@344.95	0.73798	0.30435	6.9298	4.2134E-9	8.6544E-6	10.605	View
347.2582@379.62	0.59923	-0.76139	6.6574	1.1944E-8	1.2266E-5	9.6281	View
173.9855@96.75	-1.793	-2.1814	-6.2058	6.6605E-8	4.5602E-5	8.0165	View
343.2271@340.17	0.50703	-0.45557	5.9527	1.7321E-7	8.8942E-5	7.1202	View
357.1785@488.19	0.25935	-0.70776	5.5101	9.0415E-7	3.2343E-4	5.5712	View
335.2944@537.41	0.39263	-0.40041	5.426	1.2334E-6	3.2343E-4	5.2803	View
309.2794@532.6	0.34348	1.3126	5.4032	1.3415E-6	3.2343E-4	5.2017	View
310.2826@532.64	0.34901	0.65576	5.3926	1.3946E-6	3.2343E-4	5.1653	View
303.2328@438.79	0.29848	1.6434	5.3883	1.4172E-6	3.2343E-4	5.1503	View
385.2345@438.71	0.26685	-0.49809	5.3122	1.8737E-6	3.8486E-4	4.8889	View
304.2362@438.81	0.30433	0.99595	5.2773	2.129E-6	3.9629E-4	4.7693	View

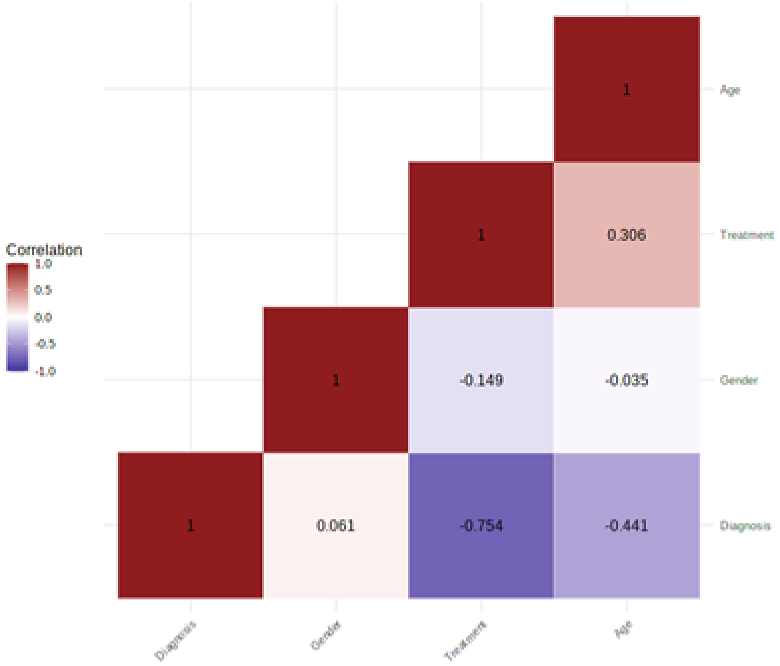
Use for functional analysis

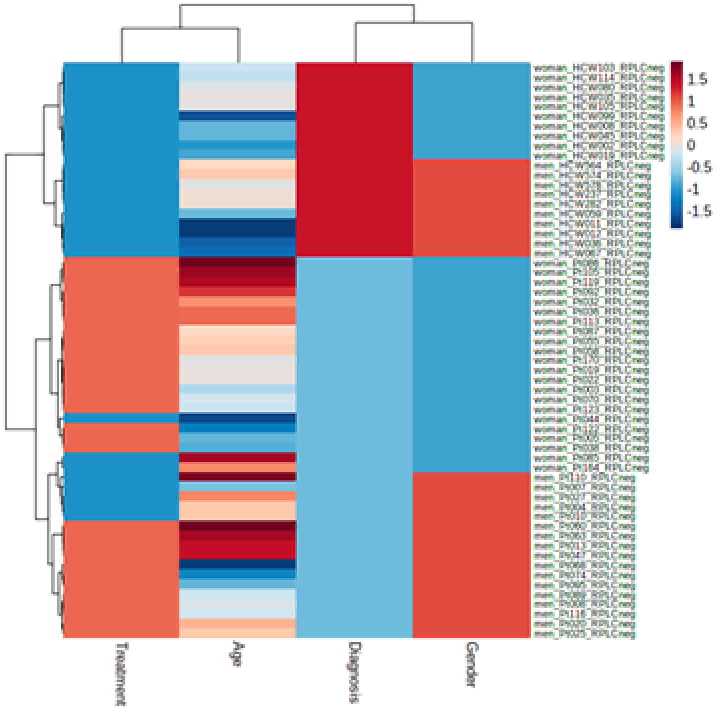
Important considerations

- Adding more variables into the model decreases statistical power
- Be strategic:
    - Do not include variables that have no impact on your data
    - Do not include correlated metadata
- Assess this with PCA, linear models, and metadata overview
    - We will go over this in the demo

Relationships between metadata

- Must understand relationships between predictors
  - Know which to include in model (sample size)
  - Guide interpretation of the results





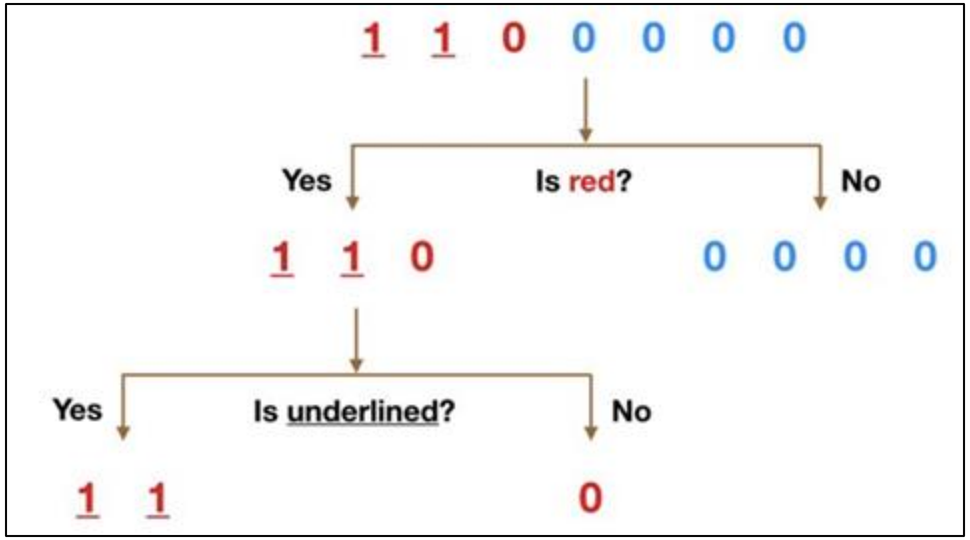
Non-linear Strategies

- All methods so far rely on linear relationships
  - Very stable, robust against overfitting
  - Only option with small sample size
- Many non-linear relationships in complex data
- Random forest is one approach
  - Uses decision trees

Example Decision Tree

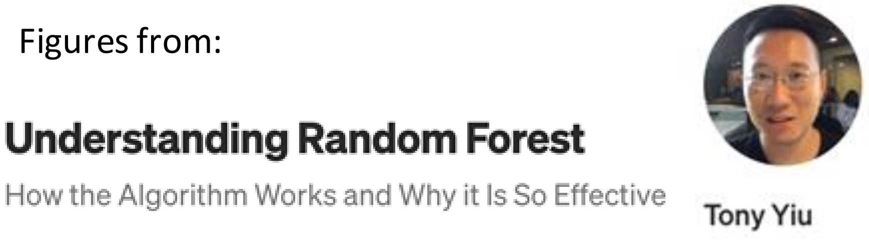
Predict whether digit is a “1” or a “0”
Two variables:

- Color (red, blue)
- Underlined (yes, no)



Random Forest Classification

Figures from:



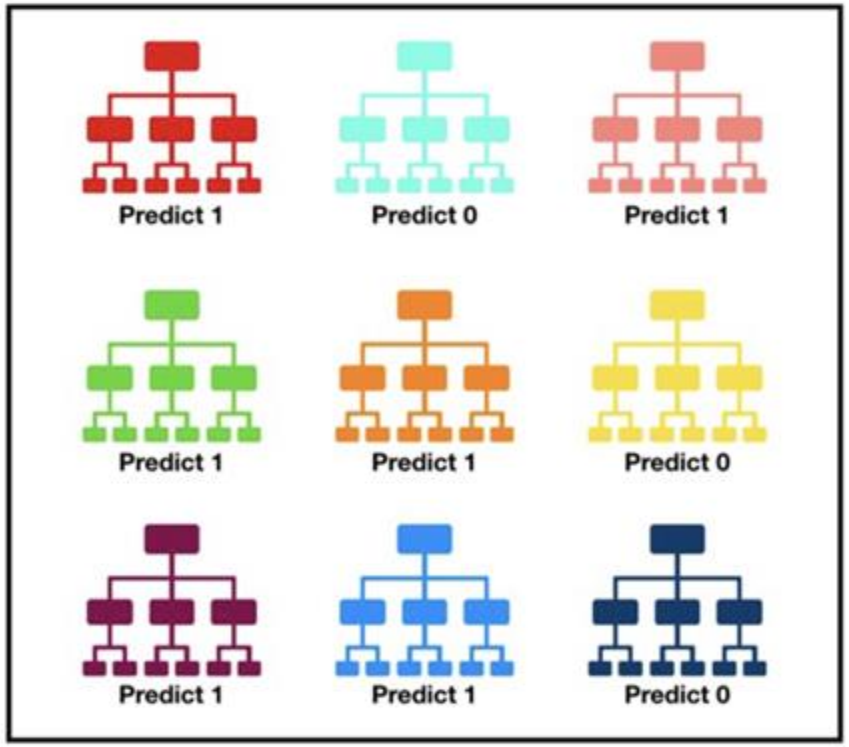
Understanding Random Forest

How the Algorithm Works and Why it Is So Effective

Tony Yiu

- In Random Forest, we build 1000s of trees
- For each tree, randomly select:
  - Random subset of metabolites + metadata
  - Random subset of samples
- Then, build a tree
- Use the rules from each tree to predict the class label for all samples (ie. COVID, healthy)
- Final prediction: majority wins

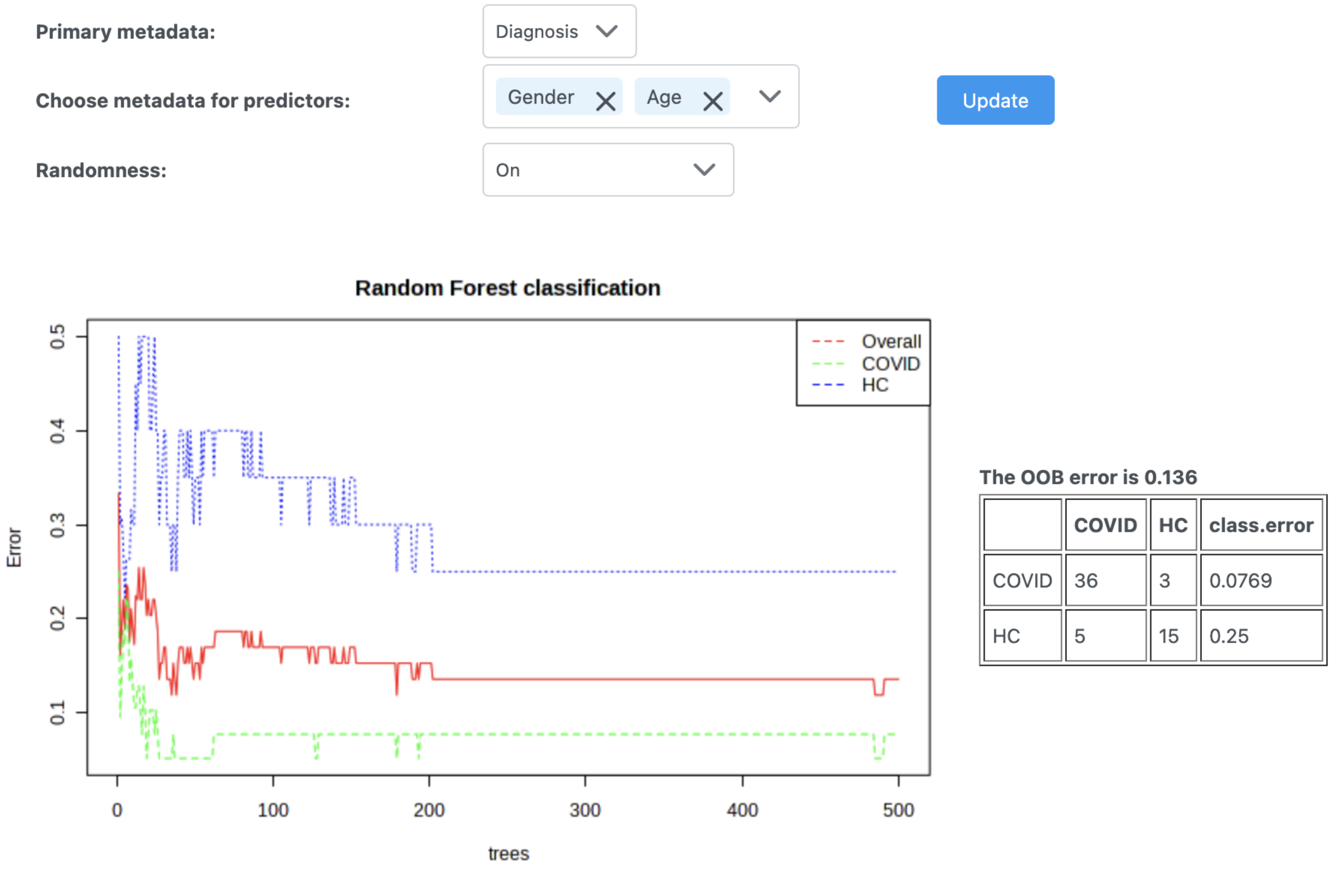
Example forest of trees



Tally: Six 1s and Three 0s

Prediction: 1

Example Classification Results



Primary metadata: Diagnosis

Choose metadata for predictors: Gender, Age

Randomness: On

Random Forest classification

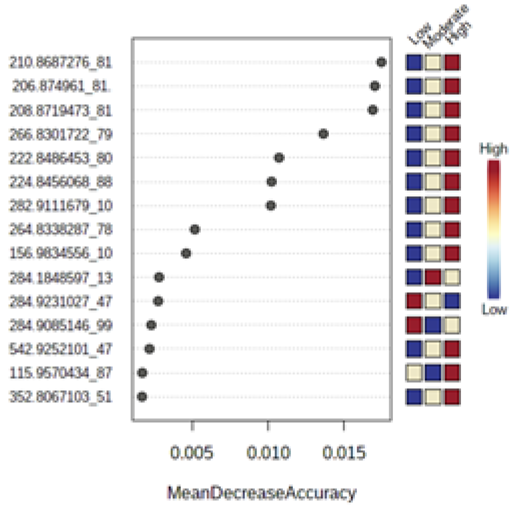
The OOB error is 0.136

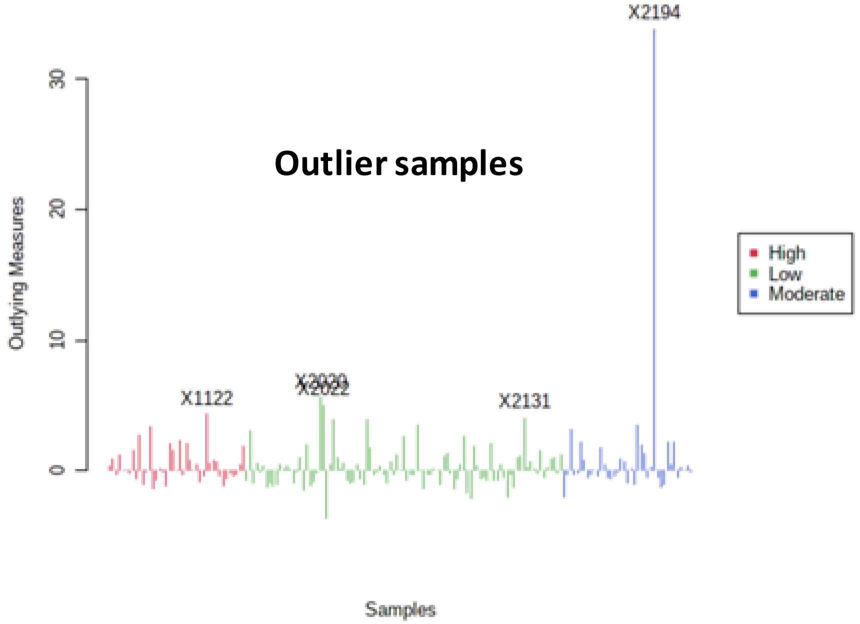
	COVID	HC	class.error
COVID	36	3	0.0769
HC	5	15	0.25

Using Random Forest in MetaboAnalyst

- Using model in real life requires extensive validation & careful design
- Most MetaboAnalyst users: use as form of exploratory statistics
- Understand which variables have high predictive power
  - Var. Importance tab
- Identify potential outlier observations
  - “Outlier Detection” tab

Most important variables





Outlier samples

Hands-On Demo