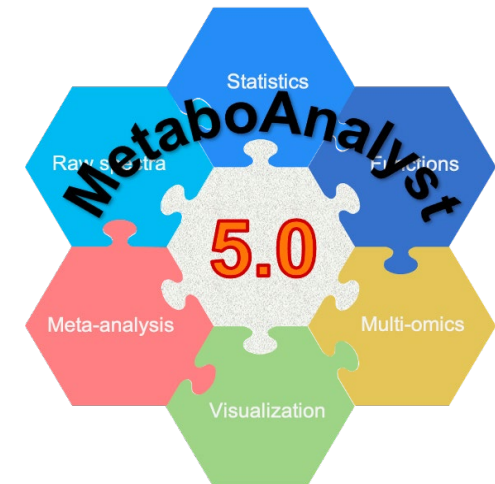# Spectra processing, functional integration and **covariate adjustment** of global metabolomics data using MetaboAnalyst 5.0

Section III: Analysis with complex meta-data

TA: Jessica Ewald

(jessica.ewald@mail.mcgill.ca)

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
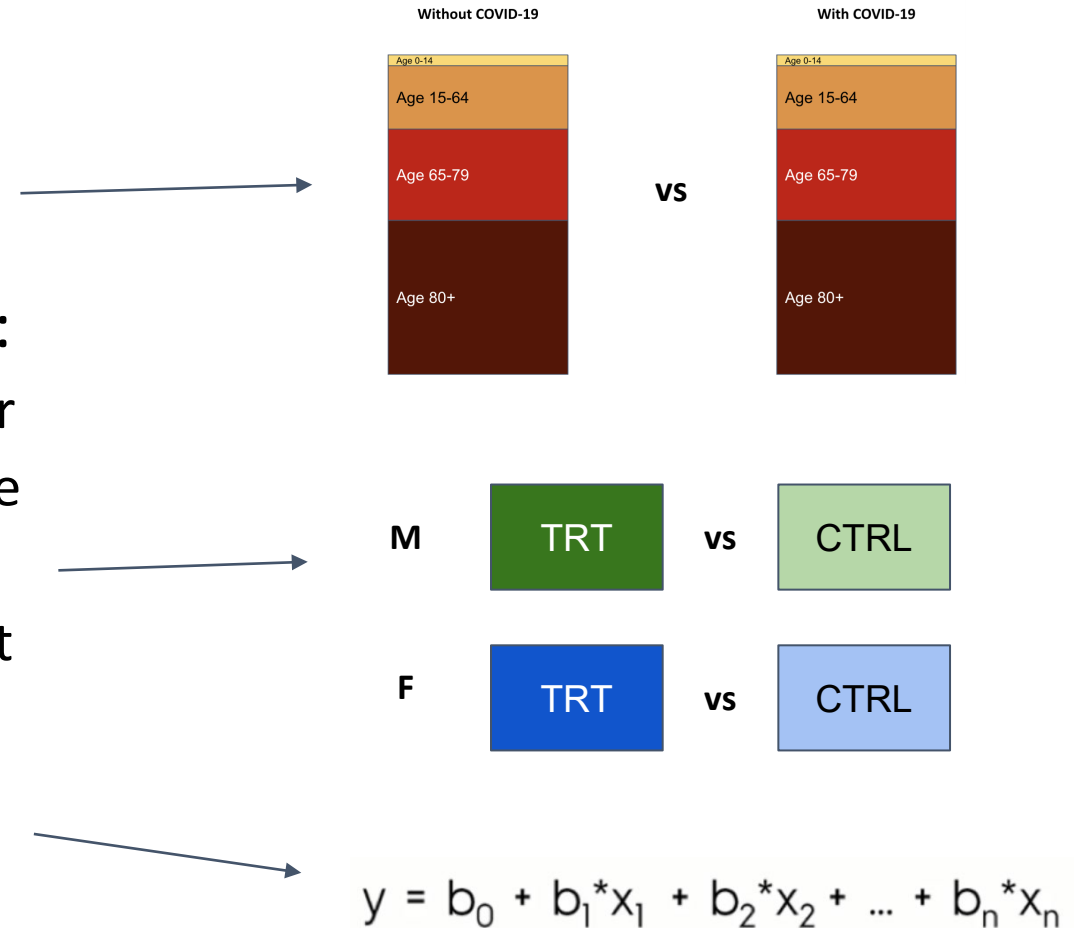
Confounding factors: example

- 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

Without COVID-19

Age 0-14

Age 15-64

Age 65-79

Age 80+

With COVID-19

Age 0-14

Age 15-64

Age 65-79

Age 80+
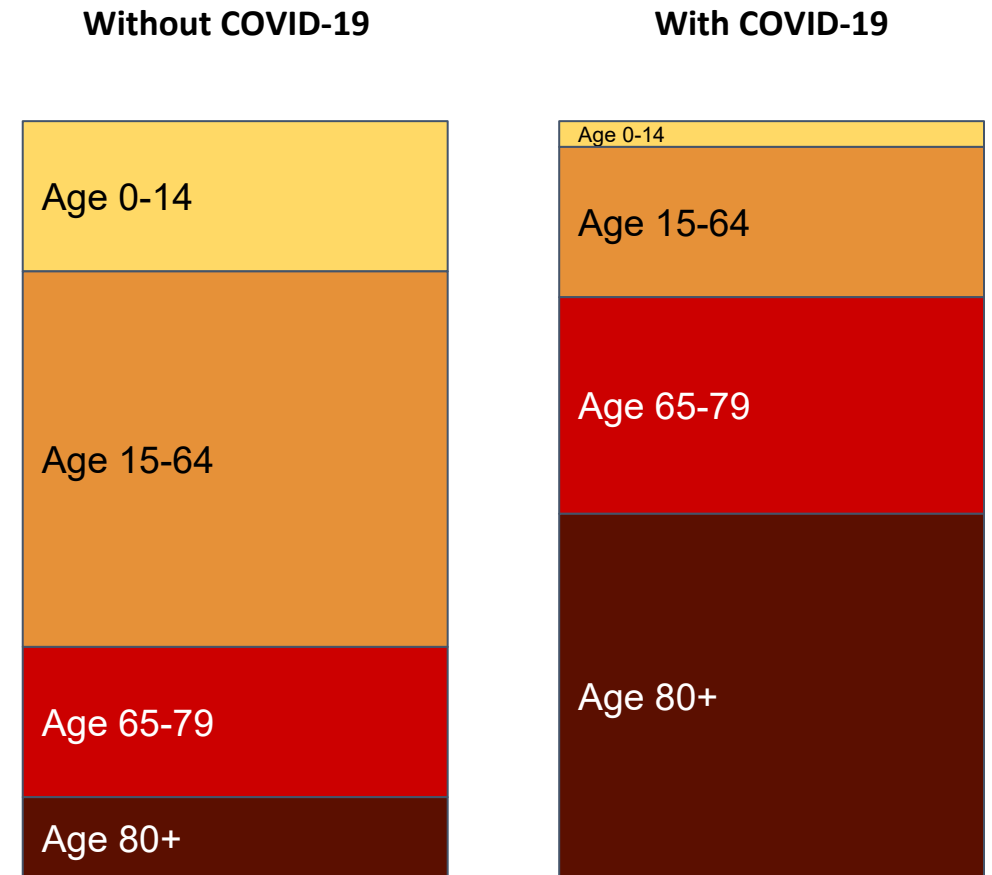
Strategies for dealing with meta-data

- Experimental design:
  - **Control** group that 'matches' your group of interest
- Data analysis (often limited by sample size):
  - **Reduce** the number of factors - use PCA + other tools to choose only factors with most influence on the data
  - **Stratification** - for factors with few classes, split up the data and analyze separately
  - **Take into account** - statistical Analysis [metadata table] module in MetaboAnalyst

Without COVID-19

| Age 0-14 |
| Age 15-64 |
| Age 65-79 |
| Age 80+ |

vs

With COVID-19

| Age 0-14 |
| Age 15-64 |
| Age 65-79 |
| Age 80+ |

M   TRT   **vs**   CTRL

F   TRT   **vs**   CTRL

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

Complex meta-data with MetaboAnalyst

- Third module will cover all topics addressed above
  - Format meta-data
  - Analyze relationships between meta-data
  - Covariate adjustment with linear model
  - Supervised analysis with Random Forest
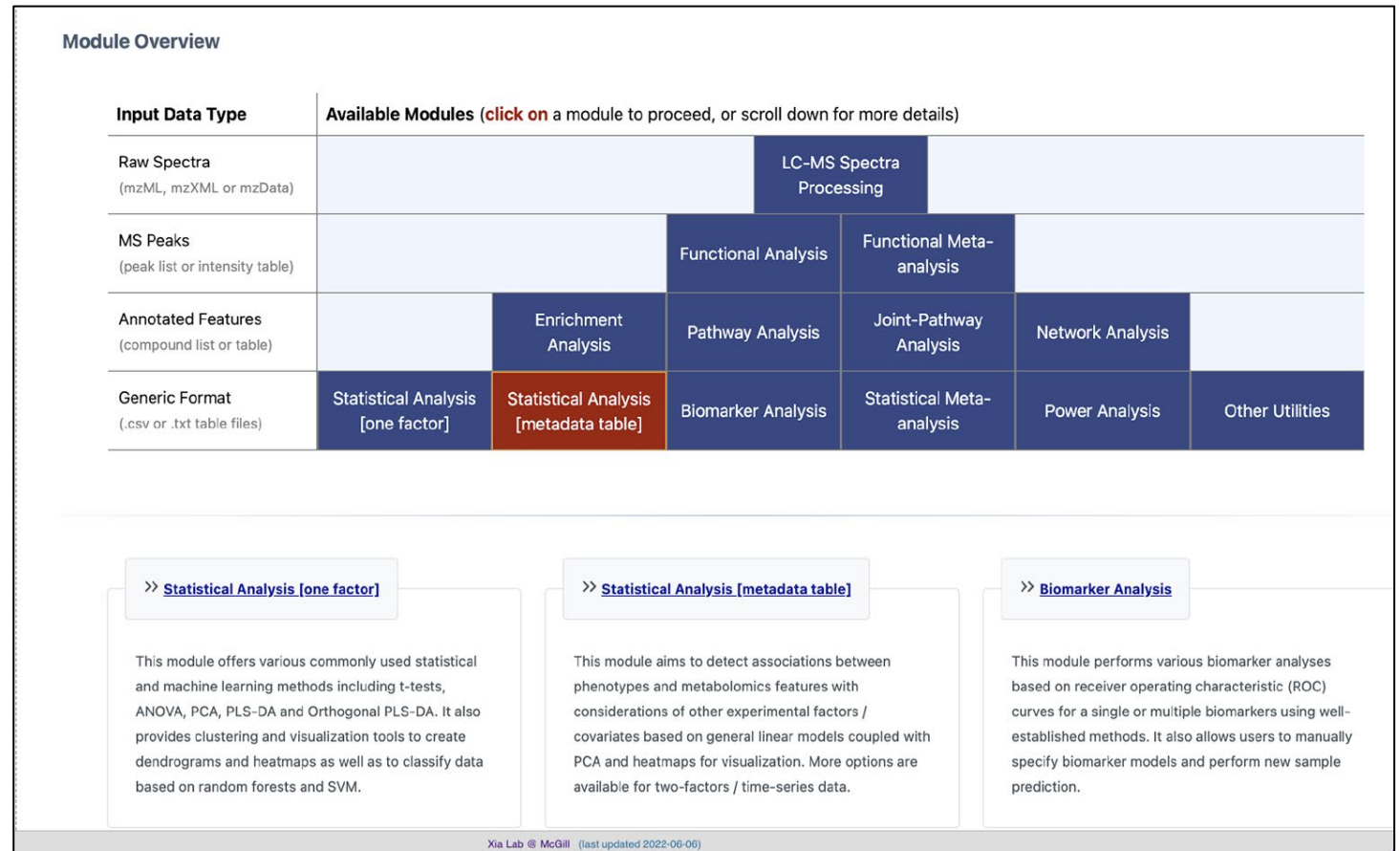
Method overview

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 Smildle, 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.

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
- No missing values
- 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	Yes	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 ˅ Categorical Continuous	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	🗑

Relationships between metadata

- Must understand relationships between predictors
  - Know which to include in model (sample size)
  - Guide interpretation of the results
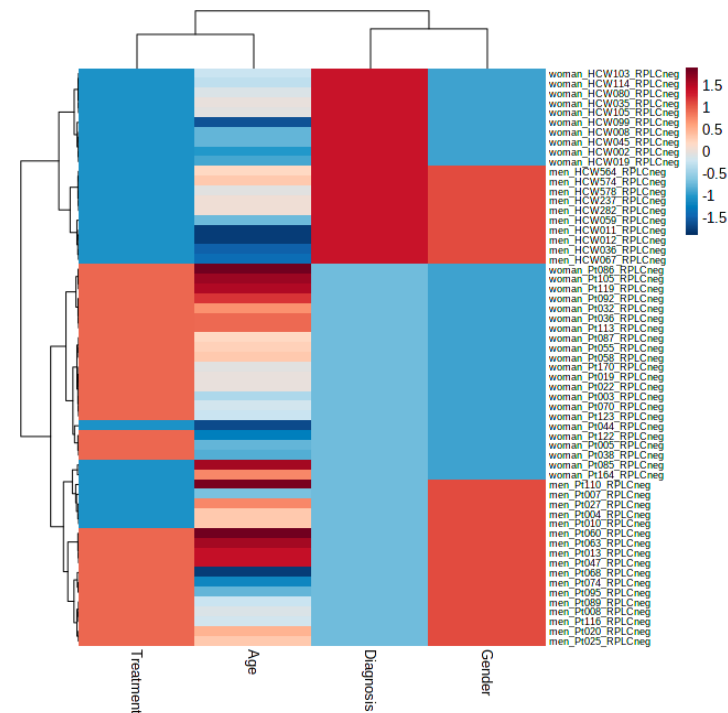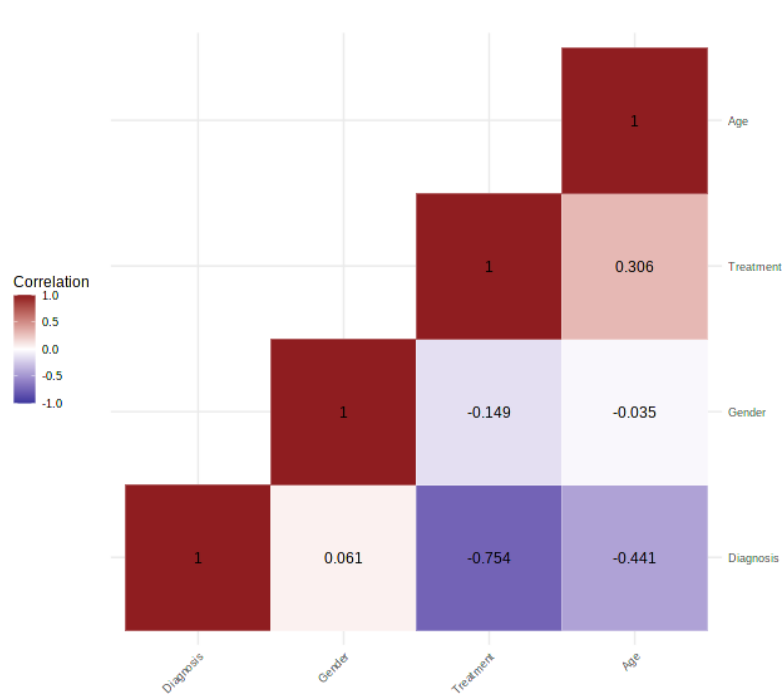
T-test vs. linear regression

- You can do t-test with linear regression:
- $y = B_0 + B_1 * x$
  - y: level of metabolite A
  - x: variable of interest
  - Categorical variables expressed using 'dummy variables'
  - Null hypothesis: $B_1 = 0$
  - Alternative hypothesis: $B_1 > 0$
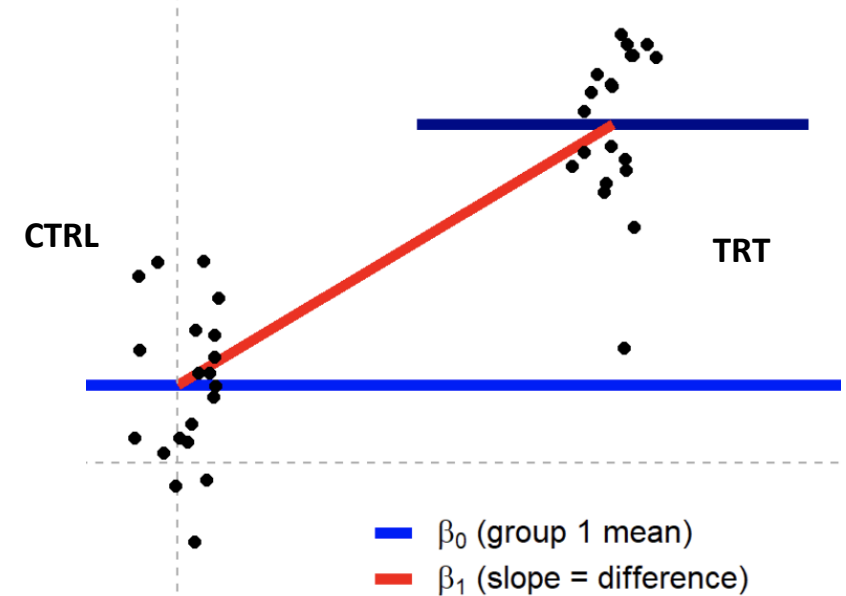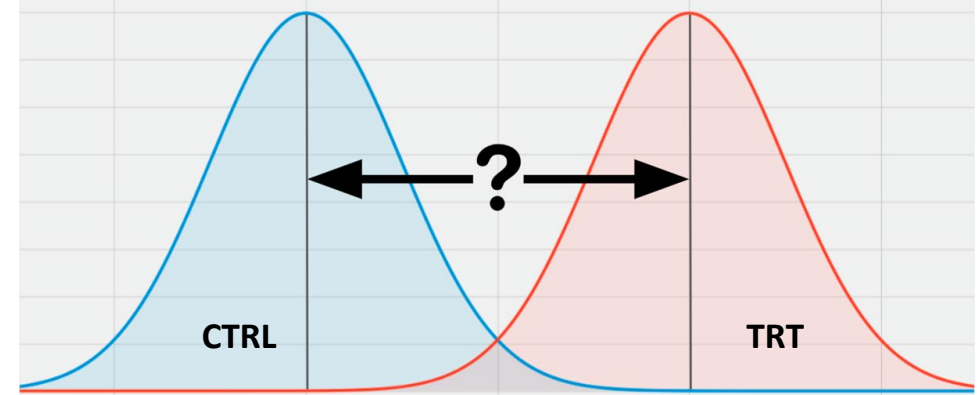
Linear regression is flexible

- Linear regression is more flexible than the classical t-test:
  - Predictor variables (x) can be continuous or categorical
  - You can have **multiple predictor variables**

- Multiple linear regression:
  - $y = B_0 + B_1 * x_{diagnosis} + B_2 * x_{age}$
  - Coefficient estimates: relationship between x_i and y *with all other variables held constant*
  - We can generate a t-stat for any coefficient using the same formula from before: $t_i = (B_i$ coefficient estimate) / (standard error of B_i coefficient estimate)

- Coefficients are unstable when the predictors are highly correlated

Interpretation of coefficient results

Multiple linear regression example:
- $y = B_0 + B_1*x_{diagnosis} + B_2*x_{age}$

- By including B_2*x_{age} in the model, we account for effects of age

- Extract B_1 from the model:
  - B_1 value = magnitude & direction of relationship between metabolite 'y' and $x_{diagnosis}$
  - B_1 p-value = statistical significance of relationship

-log10(p-value): no covariate adjustment
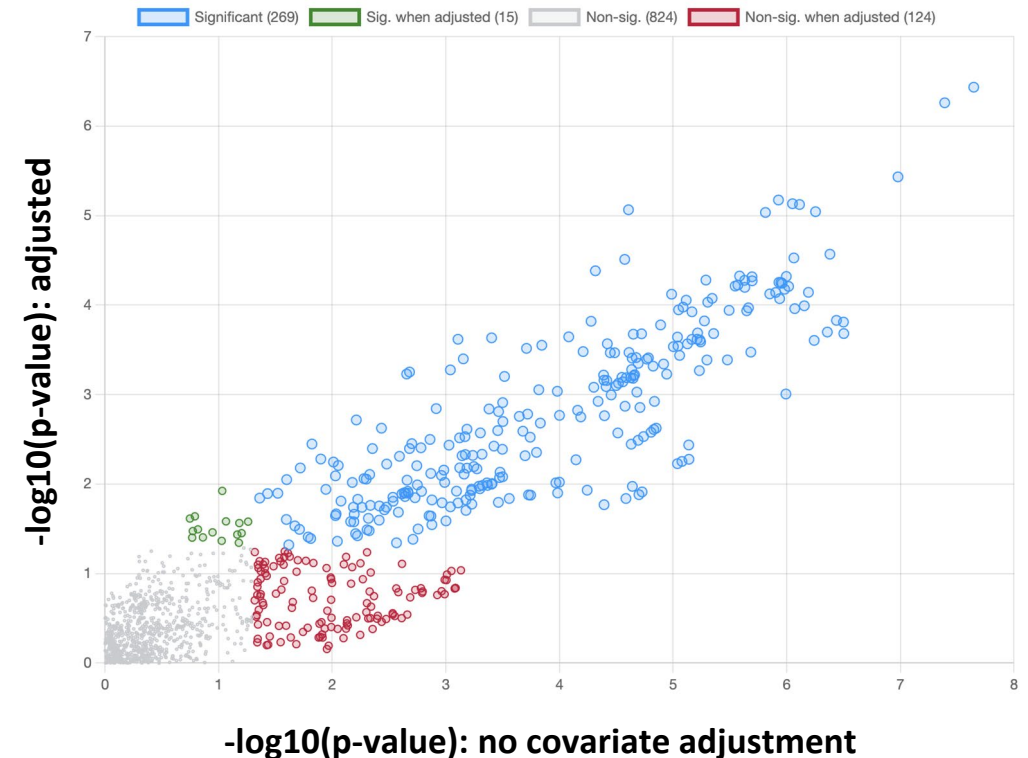
Fixed vs. random effects

- Fixed effects = covariates
  - Simply variables included in the regression model
  - All values for future known (ie. age, sex, tissue)
- Random effects = Blocking factor
  - Accounted for with multi-level modeling
  - Values could be different in future samples (ie. batch, new patients)
- Model with random effects typically perform better for prediction, but compromise statistical power & is computationally intensive
- Most analysis with MetaboAnalyst for exploration
- We suggest using fixed effects for simplicity & computational efficiency in most cases

Primary metadata:	Diagnosis ∨
Covariates (control for):	Gender ✕ ∨
Blocking factor:	-- Unspecified -- ∨

ASCA overview

- ANOVA-simultaneous component analysis
- Designed for complex, multivariate data (ie. metabolomics) with multiple experimental factors (ie. diagnosis, sex)

variability related to diagnosis

inter-individual variability

$$x_{hki_h} = \mu + \boxed{\alpha_k} + \boxed{(\alpha\beta)_{hk}} + \boxed{(\alpha\beta\gamma)\, hki_h}$$

generic residuals

variability related to interaction of diagnosis & sex

- Advantages: can model interaction terms; conveniently groups results
- Disadvantage: limited to **two categorical** meta-data
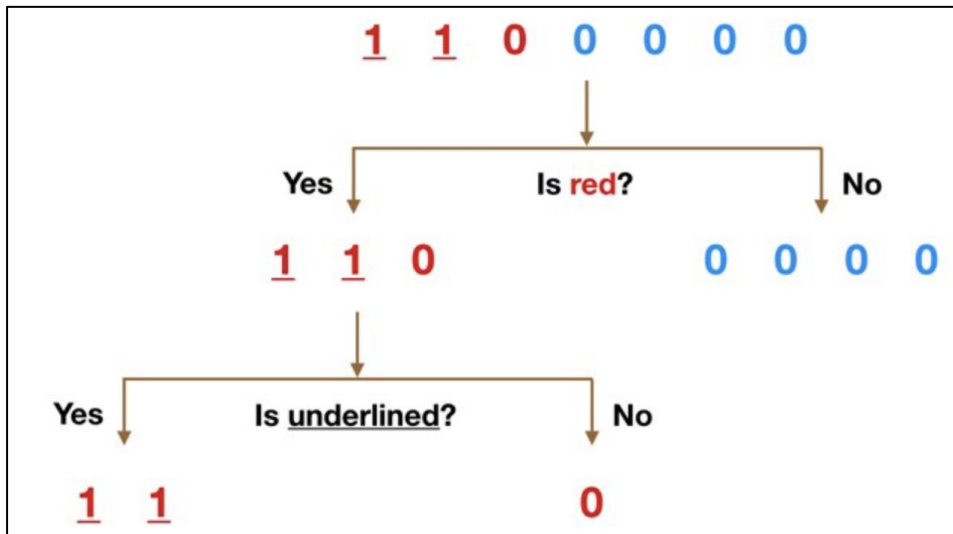
Random Forest Classification

- RF builds a model from predictors to classify observations
  - Decision trees can naturally take into considerations of different types of meta-data as well as their potential interactions
    - Patient as long-COVID vs. fast recovery based on metabolites & clinical meta-data

Example tree
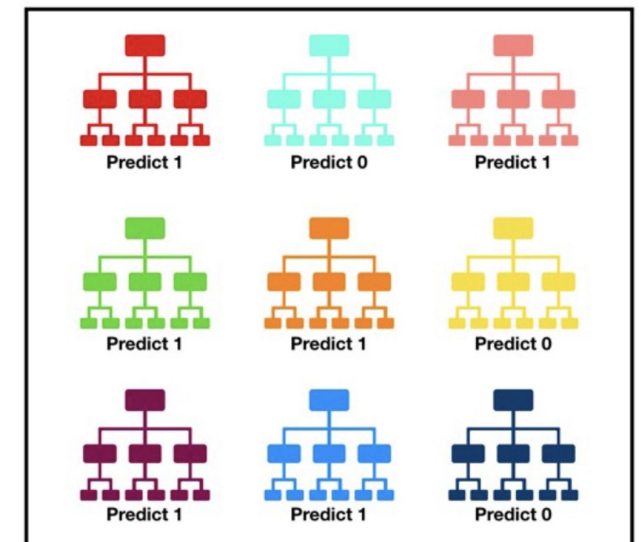

Random Forest Algorithm:
Randomly select:
- Observations (with replacement)
- Predictors

Build tree for each
Aggregate predictions across all trees and majority decision wins
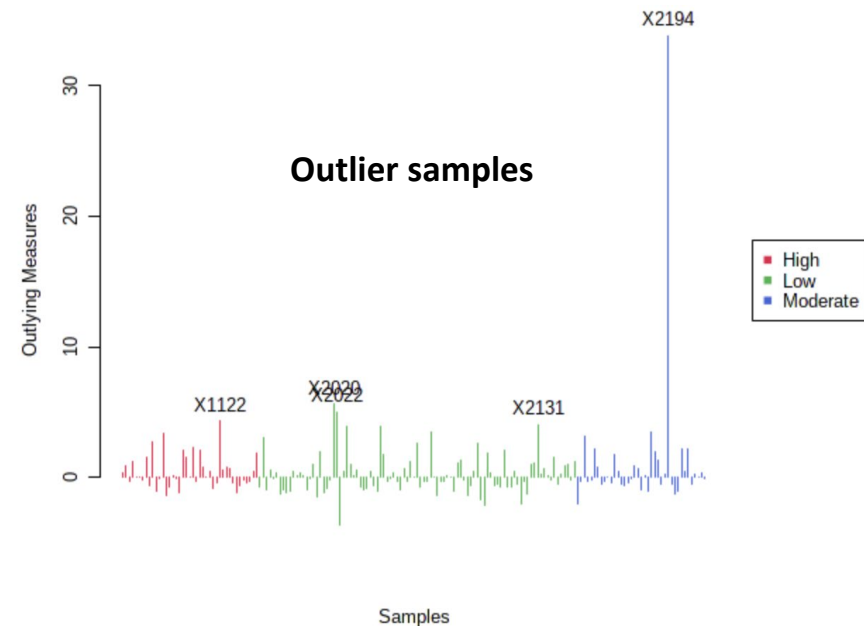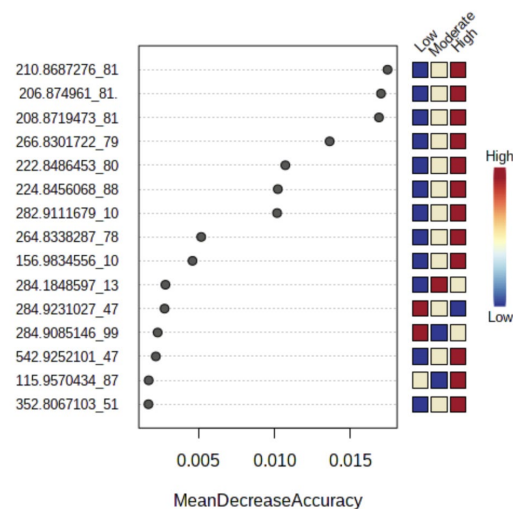
Example forest of trees

Tally: Six 1s and Three 0s
Prediction: 1

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

Tutorials

- **Publication: https://www.nature.com/articles/s41596-022-00710-w**

- **Or our manuscript: https://www.dropbox.com/s/7184c4dheeiiz2p/NP-MetaboAnalyst-2022.pdf?dl=0**

- ☐ **Stage 4: Analyzing metabolomics data with complex metadata**

Questions?

- **https://www.omicsforum.ca/search?q=%23statistics-metadata**
- **If your question is not covered, please create a new topic – we will try to answer them in the coming days**