Weill Cornell Medicine

Application of a Bayesian Model Averaging Method to Observational Metabolomics Data Analysis

Supervisor: Xi Kathy Zhou, Ph.D

Taehoon Ha

07/15/2019 Mon.

Motivation

To improve the differentially expressed (DE) metabolites identification in high-dimensional setting.

- To identify DE metabolites, there are multiple methods[1)]:
  - e.g. t-statistics, F-statistics, non-parametric methods, and etc.
  - Rank the metabolites based on the effect sizes estimated using the same model.
  - A model with the same structure and same set of covariates, for all metabolites.

- However, this approach may not be appropriate in high-dimensional setting.
  - Different metabolites may be involved in different biological processes.
  - The expression of metabolites may be affected by different sets of covariates.
  - The model may be misspecified for some metabolites.

[1)] Jeffery, Higgins and Culhane (2006)

Objective

- Investigate the association between patient characteristics and metabolites.

- Identify metabolites that are differently expressed in association with one or more patient characteristics.

Dataset: Patient Characteristics

- Consists of 80 patients X 14 patient characteristics.
- Patient characteristics include:

  Age Menopause BMI Adipocyte Hypertension Diabetes Dyslipidemia Steroid

  Total % Fat Total Fat Mass (kg) Total Lean Mass (kg) Fat Lean Ratio Trunk Fat Mass (kg) CLS-B
- Continuous variables were dichotomized by median for interpretation.

Dataset: Metabolites Expression

- Consists of 130 metabolites X 80 patients.

- Collected from a heterogeneous sample.
  - The effects of correlation in covariates need to be properly handled.

Method: Bayesian Model Averaging

Bayesian Model Averaging (BMA) allows us to coherently synthesize the information gathered from probing with different models.

- Based on linear regression models

- Zellner-Siow prior for the model parameters. [Liang et al., JASA 2008, 103 (481):410-423.]
  - Consistent for model selection
  - Bayes factors can be calculated in closed form.

Method: Bayesian Model Averaging

Bayesian Model Averaging (BMA) allows us to coherently synthesize the information gathered from probing with different models.

- An empirical approach to specify prior model probabilities
  - Good calibration of posterior inclusion probabilities
  - More realistic estimation of FDR

- Computationally efficient
  - Does not need Markov chain Monte Carlo (MCMC) simulation.

Application: 2-Stage Process

First, identified important patient characteristics.
Second, conducted BMA analysis using these patient characteristics.

Stage 1: Filter out unimportant patient characteristics

- Method: BMA with model space consisting of single variable models.
- Identify DE metabolites associated with each patient characteristics using FDR cut off 0.5.
- Patient characteristics without associated with metabolites are removed from further analysis

▼

Stage 2: Bayesian Model Averaging up to 3 covariates

- BMA analysis with filtered patient characteristics identified in the first stage.

Result

After first stage, 6 patient characteristics were identified.
Using these characteristics, 42 model spaces were generated.

- There are 7 patient characteristics that have FDR < 0.25.
  - Age, Menopausal, BMI, Diabetes, Total Lean Mass (kg), Trunk Fat Mass (kg), and CLS-B

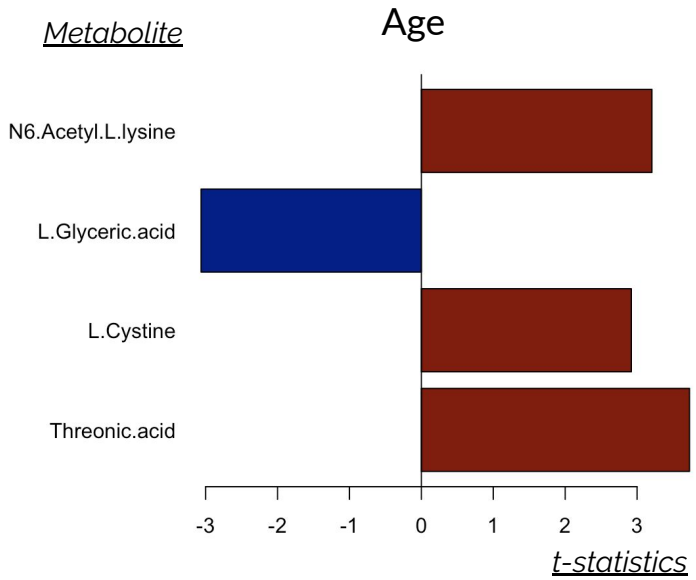
- However, Diabetes (DM) variable wad dropped due to the small sample size.
  - Only 2 patients with diabetes in the dataset.

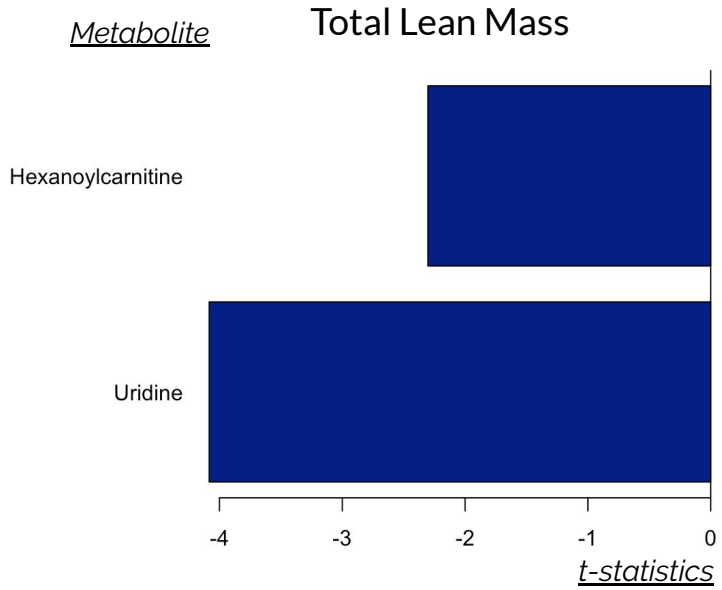
- 64 model spaces were introduced.

Null Model (~ 1)	Univariate (~ 1 + Var1)	Two-variable (~ 1 + Var1 + Var2)	Three-variable (~1 + Var1 + Var2 + Var3)	Total
${}_6C_0 = 1$	${}_6C_1 = 6$	${}_6C_2 = 15$	${}_6C_3 = 20$	42

Result: Up-down Barplot

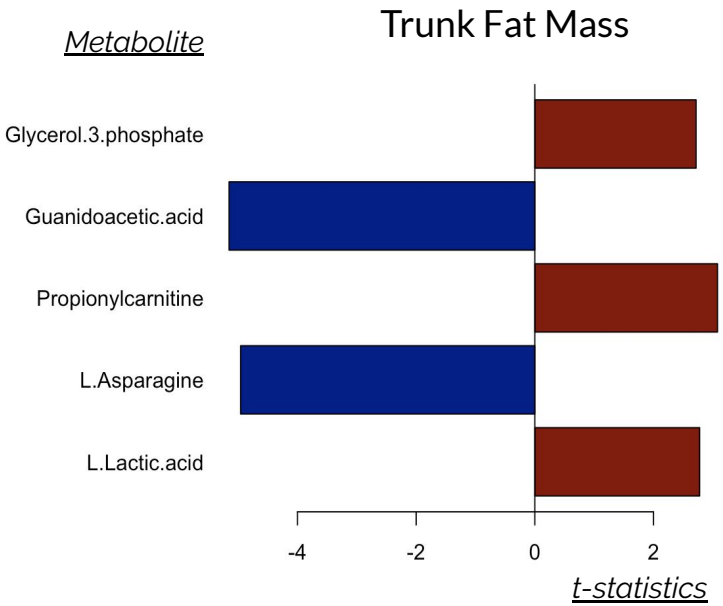
3 up- and 1 down-regulated metabolites are identified to be differentially expressed in association with Age. Only 2 down-regulated with Total Lean Mass.

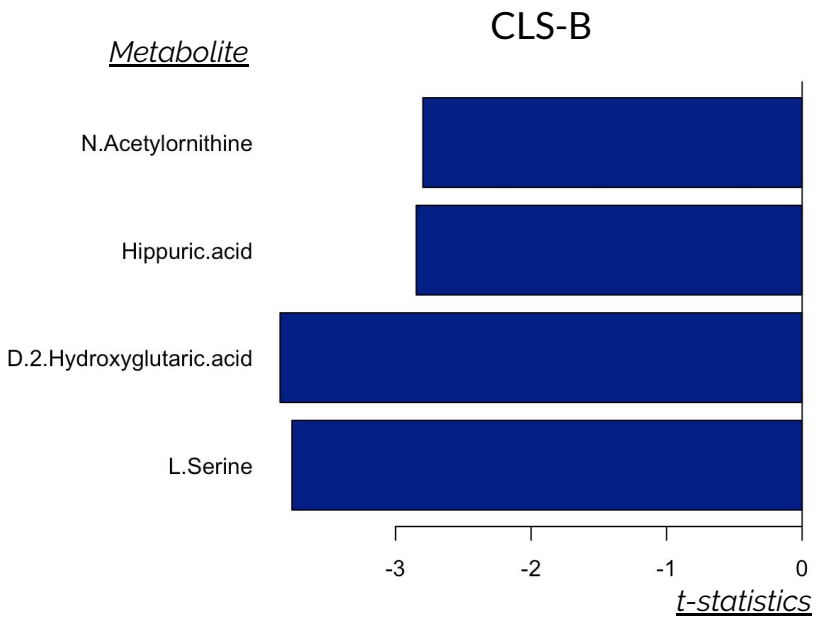




Result: Up-down Barplot

3 up-regulated and 2 down-regulated metabolites are identified to be differentially expressed in association with Trunk Fat Mass. 4 down-regulated with CLS-B.



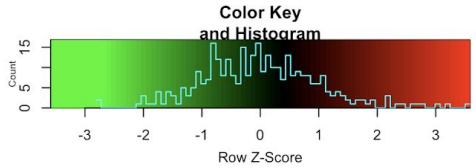


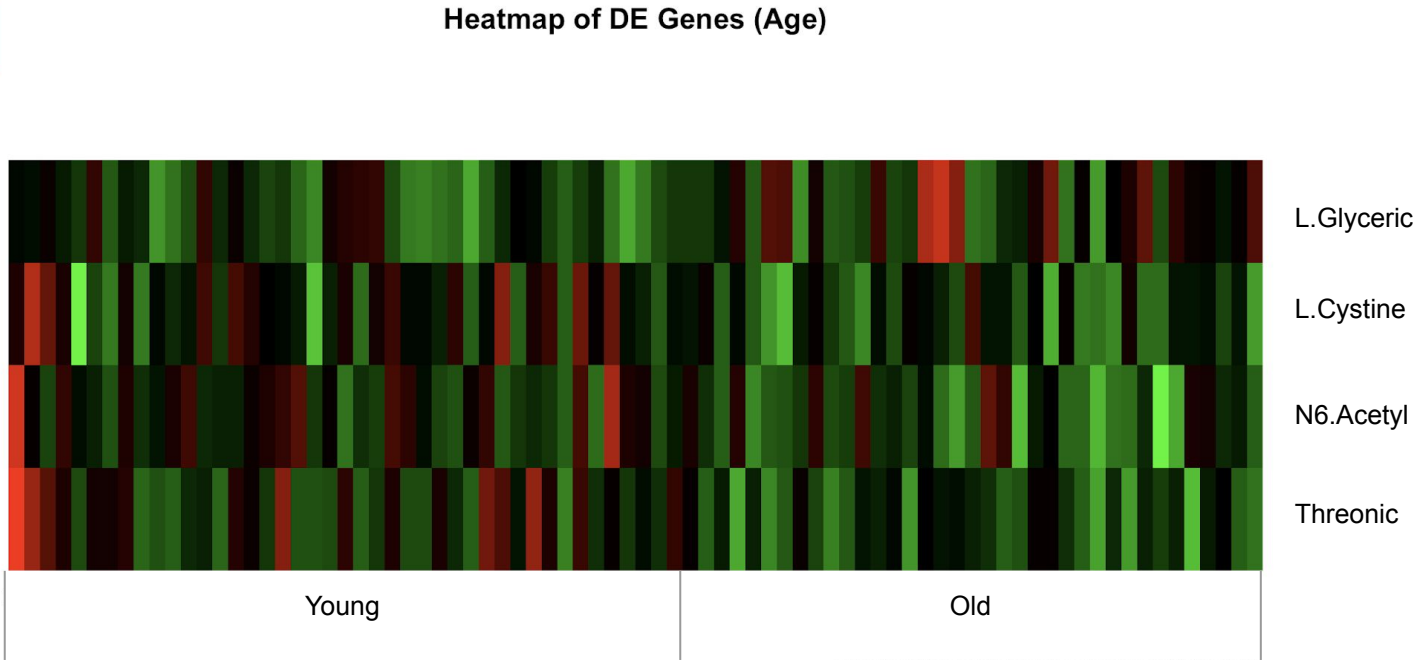
Result

Patient Characteristic	Number of DE Metabolites Associated with Patient Characteristic
Trunk Fat Mass (kg)	5
Age	4
CLSB	4
Total Lean Mass (kg)	2
Menopausal Status	0
BMI	0

Result: Heatmap - Age

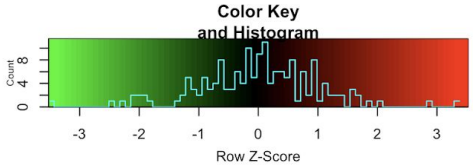
L.Cystine, N6.Acetyl, Threonic are more abundant in older patients, while L.Glyceric is more abundant in younger patients.

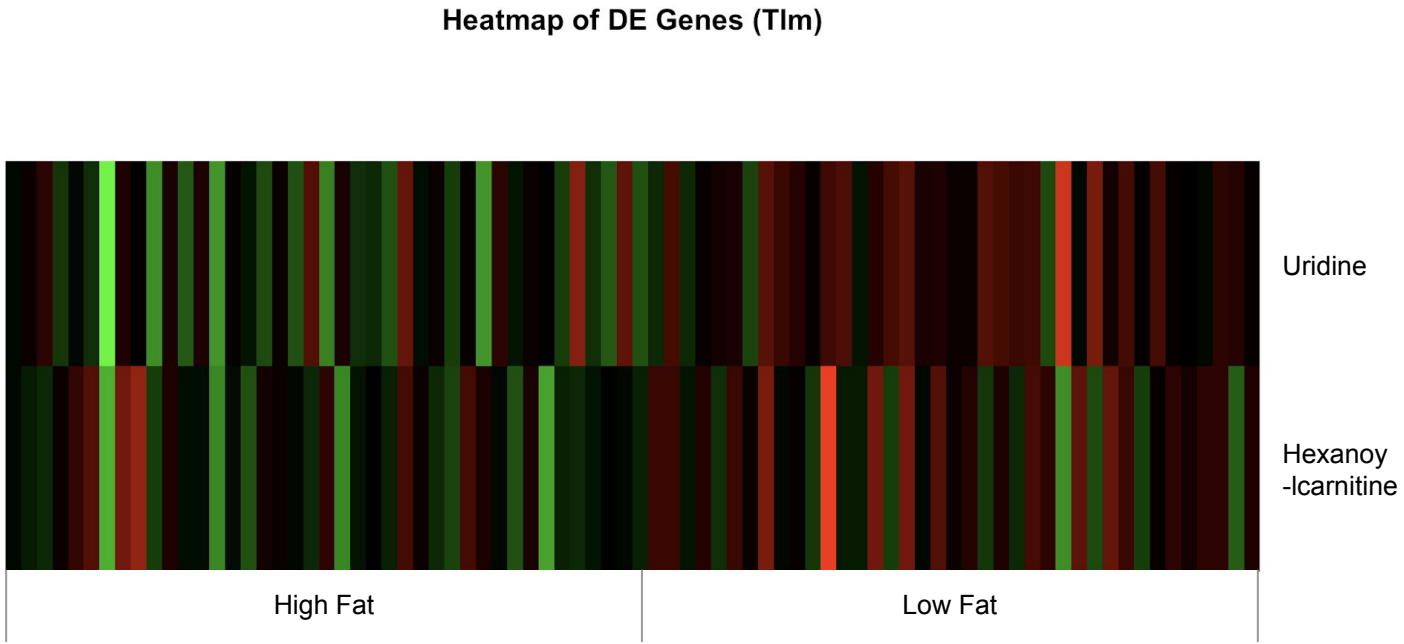




Result: Heatmap - Total Lean Mass

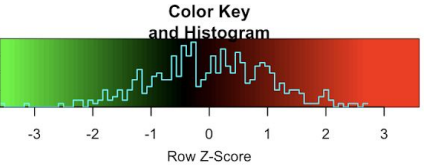
Both metabolites look more abundant in patients with low fat as compared to patients with high fat in Total Lean Mass.

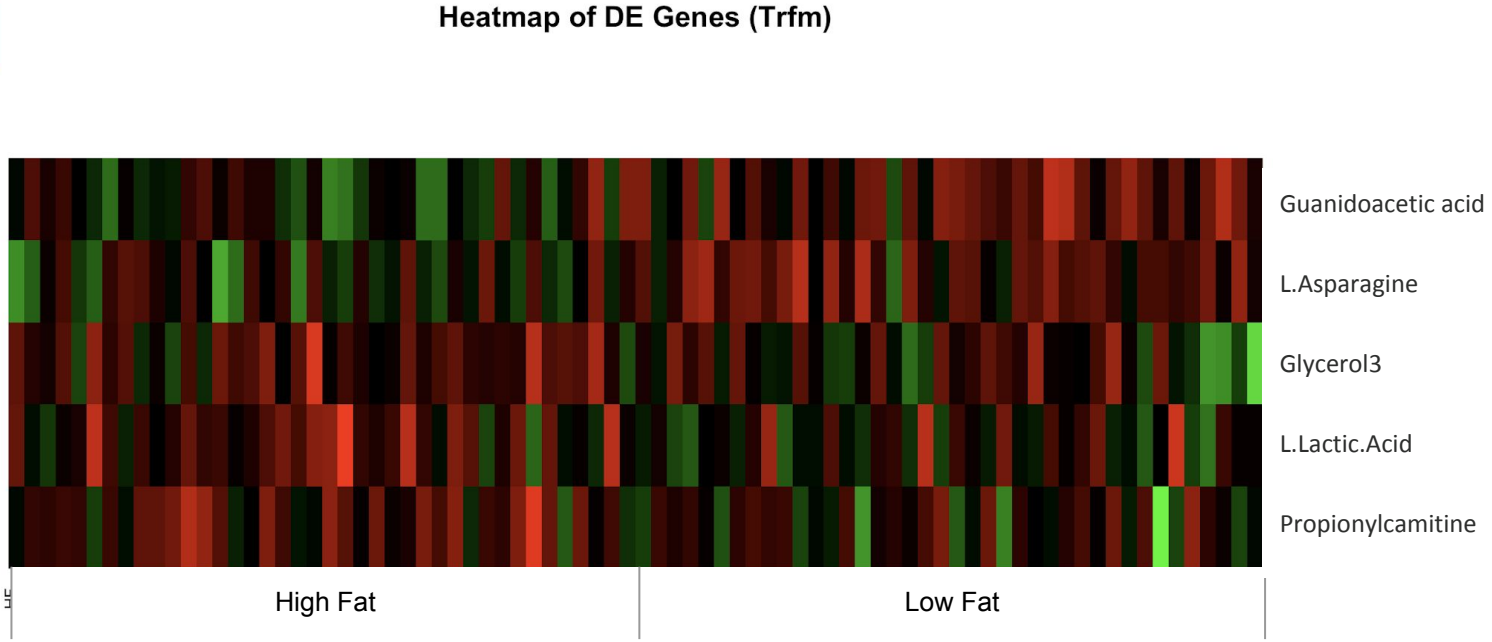




Result: Heatmap - Trunk Fat Mass

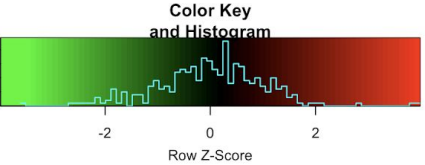
First 2 metabolites look more abundant in patients with low fat, while last 3 metabolites are more abundant in patients with high fat in Trunk Fat Mass.

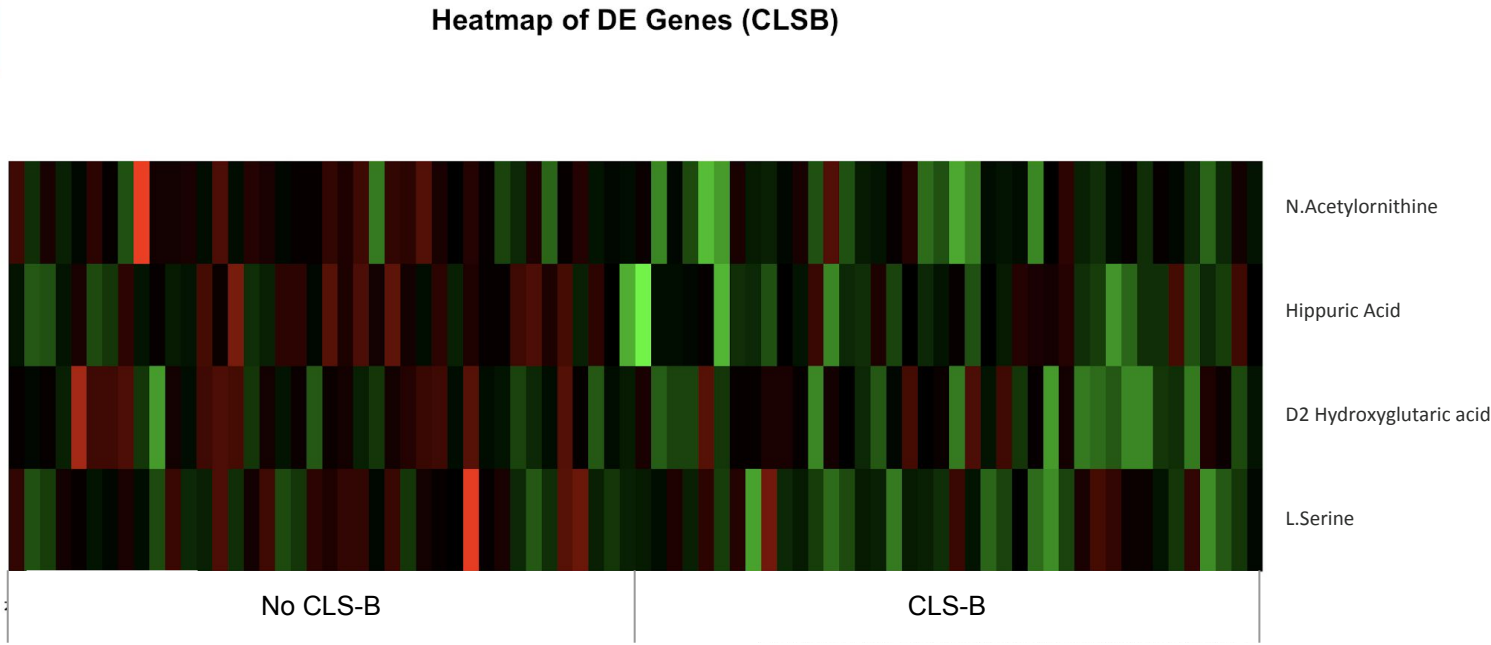




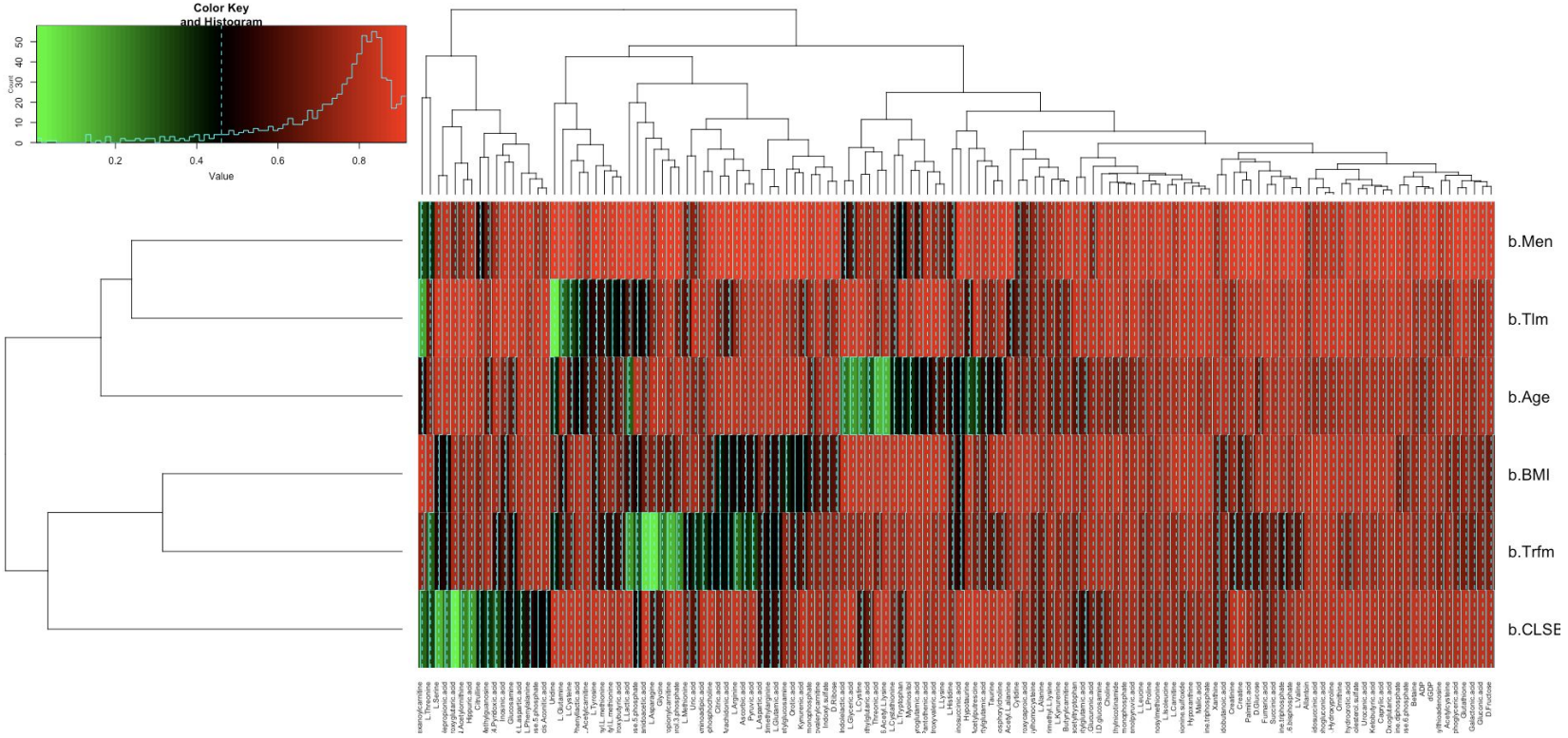
Result: Heatmap - CLS-B

All four metabolites are more abundant in patients with CLS-B as compared to the patients without CLS-B.





Result: Heatmap using FDR



Summary

- Single model approaches could result in model misspecification
  - Could lead to increased error due to bias for some metabolites and reduced efficiency.
  - Especially in high-dimensional data

- BMA was used to provide a flexible and coherent framework for identifying DE metabolites associated with a single or multiple patient characteristics.
  - Averaging over model space formed by all relevant covariates.

Discussion

- Not enough sample size for each level in some variables.
  - e.g. There are only two patients that have diabetes (DM).

- Further research needed to consider the interactions
  - Interested in metabolites associated with interaction of patient characteristics.

- Further research needed to conduct a sub-group analysis.

Weill Cornell Medicine

End of Document