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

Taehoon Ha

Abstract

Identifying differentially expressed (DE) metabolites associated with patient characteristic(s) is the primary objective of this study. Previous studies ranked all the metabolites according to a preselected statistic based on a single-model to perform two-group or multiple comparisons. Such methods inevitably cause model misspecification, which increase error originated from bias and decrease efficiency. This study, however, used a Bayesian Model Averaging model for identifying differentially expressed (DE) metabolites associated with one or more patient characteristics. Unlike single model approaches, this method can apply different models to different sets of metabolites so that each model only includes a relevant set of patient characteristics, which describes the metabolites well. Also, it can correct model misspecification by averaging over model spaces formed by all relevant patient characteristics. A two-stage process was taken for BMA analysis. First, patient characteristics were considered unimportant and were filtered out if False Discovery Rate (FDR) is greater than 0.25. Next, BMA analysis was carried out with filtered patient characteristics up to 3 covariates; thereby, 64 model spaces were created. Seven patient characteristics (Age, Menopausal, BMI, Diabetes status, Total Lean Mass, Trunk Fat Mass, and CLS-B) were identified after BMA with model space consisting of single-variable models. Overall, there are 8 DE metabolites associated with diabetes status, and Trunk Fat Mass, 5; Age, 4; CLS-B, 4; Total Lean Mass, 2 in order. There was no identified DE metabolite(s) associated with Menopausal status and BMI.

Introduction

It is common that there is uncertainty from the noise of measurement and small sample size when analyzing observational metabolomics data. However, a recent breakthrough in biotechnology allows us to conduct larger-scale studies, which led to more accurate measurement at a cheaper cost. More complex results can be found in differential gene expression studies with heterogeneous sample characteristics compared to homogenous samples such as identifying differentially expressed genes with multiple features. In other words, it is possible that metabolites could be associated with one or more patient characteristics. In order to get such results, we need to quantify the strength of association between the expression of each metabolite and a set of patient characteristics.

Current methods to identify differentially expressed genes (or metabolites) include t-statistics, F-statistics, and non-parametric methods, or Bayesian approach. Most of them rank the metabolites based on the effect size estimated using the same model, which means it shares the same structure of the model with the same set of covariates across the metabolites. However, this type of approach could result in a problem, especially in high dimensions. For example, different metabolites may be involved in different biological processes. Also, the expression of metabolites may be affected by a different set of patient characteristics. Even we can think of the model could be misspecified for some metabolites.

To overcome model misspecification in the previous studies, Bayesian Model Averaging (BMA) method can be a good solution because it allows applying different models to different sets of metabolites so that each model includes an only relevant set of patient characteristics which describes the metabolites well.

Recently, BMA approaches enables to deal with various problems involving high throughput genetic data such as improving the assessment of candidate gene effects in the genome-wide association studies [Wu et al. (2010); Xu, Craiu and Sun (2011)], and to improve the DE gene detection in settings where the microarray data involved two different distributional assumptions [Sebastiani, Xie, and Ramoni (2006)]. One drawback of these approaches is that they use the Markov chain Monte Carlo (MCMC) simulation when estimating the model parameters, which requires high computational power. In this study, however, it is more computationally efficient compared to previous studies as the MCMC method was not utilized. Only linear regression models were applied to identify DE metabolites associated with one or more patient characteristics.

The following describes two datasets and their preprocessing. One is an observational metabolite expression, and the other includes patient characteristics. Next, we will discuss the methodology: how the BMA approach was used to control for sample heterogeneity and model uncertainty properly. The final section concludes with the application of the BMA approach to observational metabolite expression and patient characteristics datasets.

Data

Two datasets were used to identify DE metabolites associated with one or more patient characteristics. The patient characteristic data set consisting of 80 patients with their 14 characteristics. Patient characteristics include:

- **Age**
- **Menopause**
- **BMI**
- **Adipocyte**
- **Hypertension**
- **Diabetes status (DM)**
- **Dyslipidemia**
- **Steroid**
- **Total % Fat**
- **Total Fat Mass (kg)**
- **Total Lean Mass (kg)**
- **Fat Lean Ratio**
- **Trunk Fat Mass (kg)**
- **CLS-B.**

For better interpretation, Age, BMI, Adipocyte level, Total Percent Fat (kg), Total Fat Mass (kg), Total Lean Mass (kg), Fat Lean Ratio, and Trunk Fat Mass (kg) were dichotomized by the median in the preprocessing step.

- **Age**: Old vs. Young
- **Menopause**: Post vs. Pre
- **BMI**: Normal and Underweight vs. Overweight and Obese
- **Adipocyte**: Small vs. Big
- **Hypertension**: Yes vs. No
- **Diabetes(DM)**: Yes vs. No
- **Dyslipidemia**: Yes vs. No
- **Steroid**: Yes vs. No
- **Total % Fat**: High fat vs. Low fat
- **Total Fat Mass**: High vs. Low fat
- **Total Lean Mass**: High vs. Low fat
- **Fat Lean Ratio**: High vs. Low fat
- **Trunk Fat Mass**: High vs. Low fat
- **CLS-B**: Yes vs. No

Metabolite expression data is composed of 80 patients' 130 metabolites. Since it is collected from a heterogeneous sample, it is crucial that the effects of correlation in covariates need to be handled appropriately in a further step.

Methodology

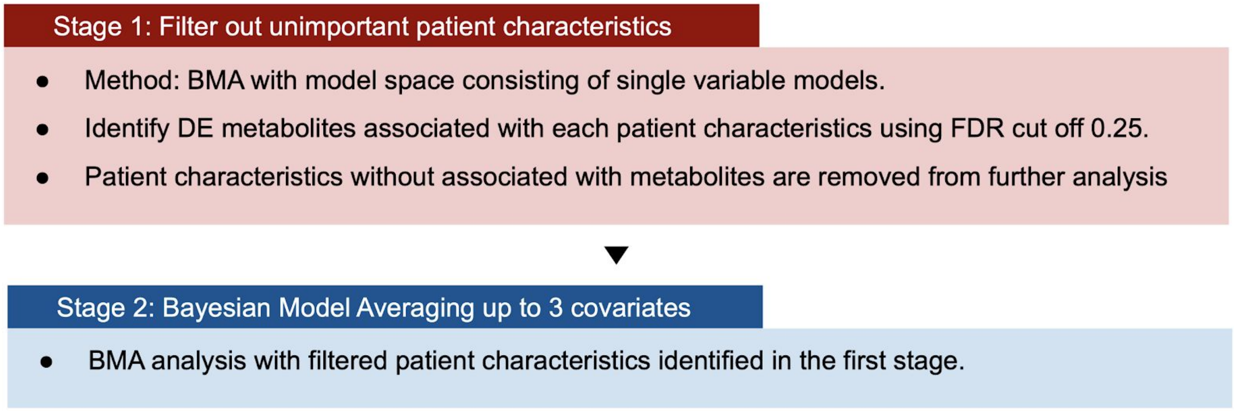
It is difficult to identify appropriate models for different sets of metabolites as model uncertainty to identify a single best model. Compared to the single-model approach, BMA gives us a more flexible and coherent framework to identify DE metabolites associated with a single, or multiple patient characteristics. In detail, the BMA method enables us to apply different models to different sets of metabolites whereby each model contains only the set of covariates relevant to the metabolites it is describing.

Here we propose a BMA approach for observational metabolomics data based on linear regression models. Zellner–Siow prior for model parameters were utilized. The consistency property of this prior is important, as it allows for obtaining a consistent estimate of the distribution of the metabolites in the model space using Bayes factors.

We also used an iterative procedure to generate the prior model probabilities. This allows us to match the estimated distribution of the metabolites within the model space based on posterior model probabilities and the estimate based on the Bayes factors. Through this procedure, we can achieve the efficient computation of the Bayes factors and the posterior inclusion probabilities without using MCMC simulation. Notice that the rank-based approaches showed less variability to a wide range of choice for prior model probabilities, while the accuracy of the FDR directly estimated from the posterior model/inclusion probabilities was relatively sensitive to the choice of prior.

In the process of application of the BMA method, a two-stage process was implemented to identify DE metabolites associated with one or more covariates. As a first step, unimportant patient characteristics were removed by BMA with model space consisting of 14 single-variable models. We identified DE metabolites associated with each patient characteristic. Patient characteristics with FDR < 0.25 were removed for further analysis. Next, BMA analysis was carried out with filtered patient characteristics up to 3 covariates; thereby, 64 model spaces were generated.

[Figure 1] Two-stage approach to identify DE metabolites associated with covariate(s)



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 0.25.
- 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

Seven patient characteristics (Age, Menopausal, BMI, Diabetes status, Total Lean Mass, Trunk Fat Mass, and CLS-B) were identified after BMA with model space consisting of single-variable models.

[Table 1] Summary table of the number of DE metabolites associated with each covariate

Patient Characteristics	Number of DE Metabolites Associated
DM: Diabetes Status	8
Trunk Fat Mass	5
Age	4
CLS-B	4
Total Lean Mass	2
Menopausal Status	0
BMI	0

After the second stage, the result showed that diabetes status has eight associated DE metabolites, which is the most among filtered patient characteristics. There are 5 DE metabolites in association with Trunk Fat Mass; 4 for Age; 4 for CLS-B; 2 for Total Lean Mass in order. No DE metabolite was identified in association with Menopausal status and BMI. We will discuss the results BMA analysis of each patient characteristic.

Age

[Figure 2] demonstrated that there are 4 DE metabolites associated with age covariate. L.Glyceric acid is the only down-regulated DE metabolite, while the other 3 are up-regulated. In particular, [Figure 3] reported that L.Glyceric acid is more abundant in younger patients, whereas other 3 down-regulated DE metabolites are more abundant in older patients.

DM: Diabetes status

[Figure 4] tells us that there are 7 up-regulated and 1 down-regulated DE metabolites in association with diabetes status. According to [Figure 5], we can identify that Creatinine is relatively more abundant in patients without diabetes, but other metabolites are highly abundant in patients with diabetes. However, there are some limitations to interpret this result due to the small sample size; the dataset includes only 2 patients who have diabetes.

Total Lean Mass

No up-regulated, and 2 down-regulated metabolites are identified to be differentially expressed in association with Total Lean Mass in [Figure 6]. Both Uridine and Hexanoylcarnitine look more abundant in patients with low fat as compared to patients with high fat in Total Lean Mass.

Trunk Fat Mass

In total, 5 DE metabolites were identified that there is an association with Trunk Fat Mass: 3 up-regulated and 2 down-regulated. In particular, [Figure 9] says that the first two DE metabolites, Guanidinoacetic acid, and L.Asparagine, are more plentiful in patients with low fat, whereas Glycerol3, L.Lactic.Acid and Propionylcamitine seem more prevalent in patients with high fat.

CLS-B

Only 4 down-regulated metabolites were identified to be differentially expressed in association with CLS-B. All four DE metabolites are more abundant in patients with CLS-B as compared to the patients without CLS-B. [Figure 10 and 11]

Conclusion

Suggested methods in the previous studies have the same structure of the model with the same set of covariates across the metabolites. This could result in an increased error due to bias and for some metabolites and reduced efficiency, especially in a high-dimensional setting. We proposed to use the BMA approach to improve our ability to identify DE metabolites. This approach utilizes the Zellner-Siow prior for model parameters. The consistency property of this prior is important as it allows for obtaining a consistent estimate of the distribution of the genes in the model space using Bayes factors. These prior choices allow the efficient computation of the Bayes factors and the posterior inclusion probabilities that does not depend on a MCMC simulation.

After the two stages process, we could identify the association between metabolites and patient characteristics. In particular, [Figure 12] describes the comprehensive relationship between metabolites and patient characteristics by using FDR heatmap. Some metabolites are solely associated with specific patient characteristics, while some of the other metabolites are in association with not only one single patient characteristics, but also with other patient characteristics. For instance, the metabolites associated with diabetes status (DM) do not have other association with other patient characteristics. However, some of the metabolites associated with CLS-B are also associated with BMI and Trunk Fat Mass.

Discussion

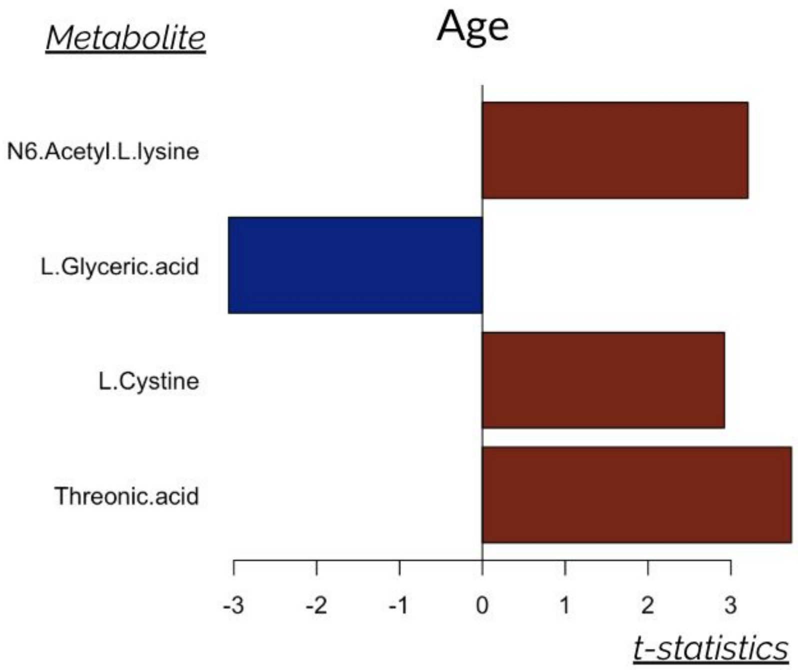
There are a couple of limitations in this study. First, in the dataset, there are only two patients with diabetes. Attention is in need when using the empirical prior since the small sample size property of the Zellner-Siow prior is less certain. Larger sample size with enough patients within each level will reduce the bias and increase the efficiency in the future study.

In high dimensions, the expression of metabolites may be affected by not only a different set of patient characteristics but also interactions. One problem is that, as the number of patient characteristics increases, the number of interactions exponentially increases and requires a much higher computational power. However, it is important to include the interactions to evaluate effects more precisely. Hence, in future studies, it is crucial to find a computationally efficient way to deal with a large number of factors and their interaction effects when identifying the DE metabolites. In addition to adding the interactions, the study can be further developed by analyzing the difference between sub-groups.

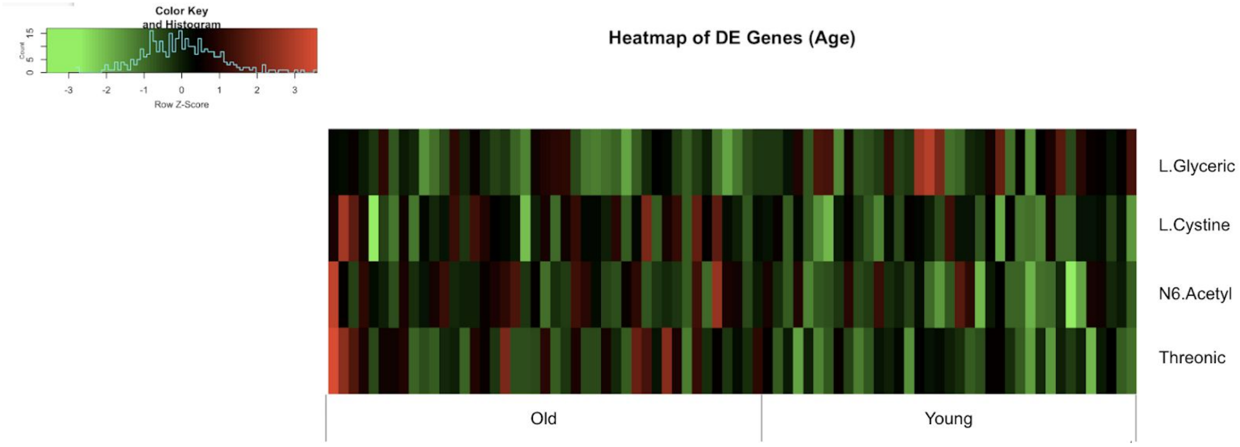
Appendix

Age

[Figure 2] Up-down barlot: Age

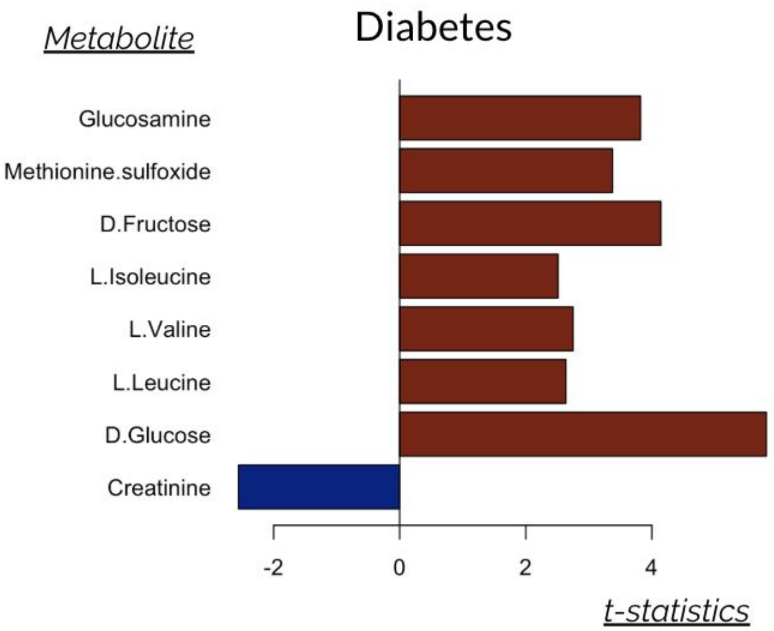


[Figure 3] Heatmap: Age

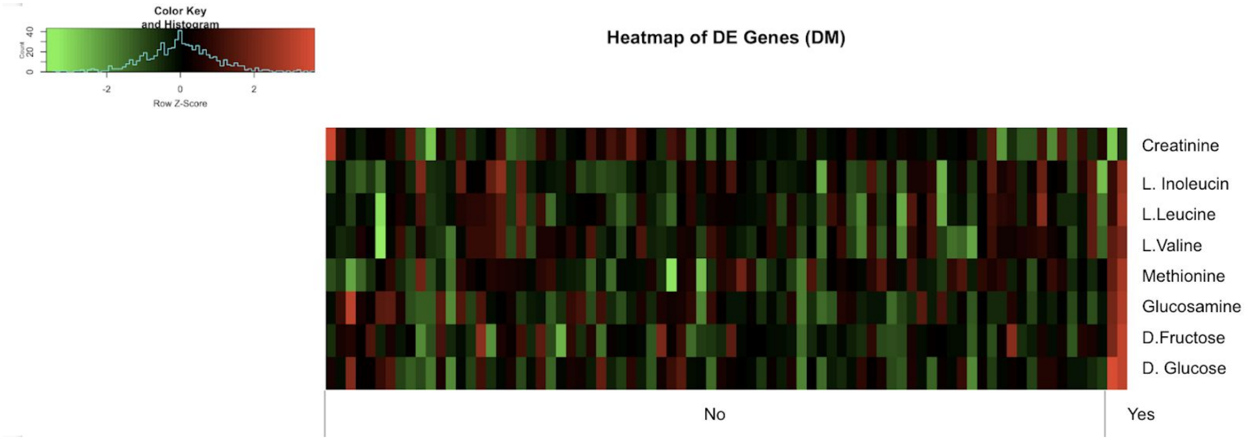


DM: Diabetes status

[Figure 4] Up-down barlot: Diabetes

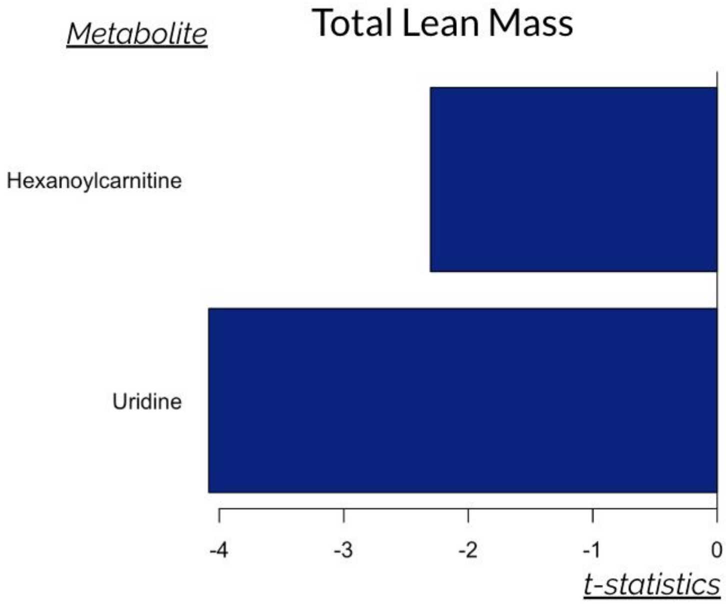


[Figure 5] Heatmap: Diabetes status (DM)

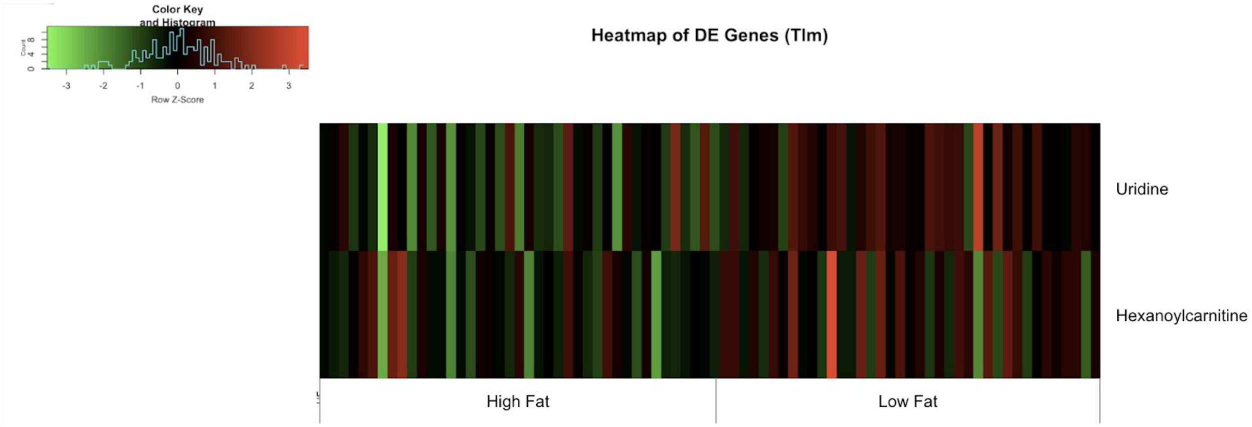


Total Lean Mass

[Figure 6] Up-down barlot: Total Lean Mass

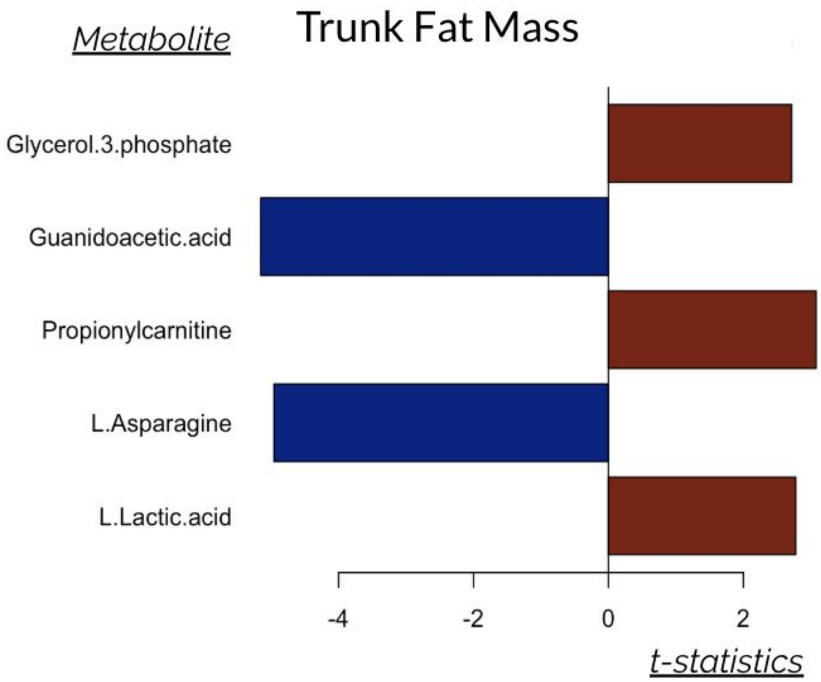


[Figure 7] Heatmap: Total Lean Mass

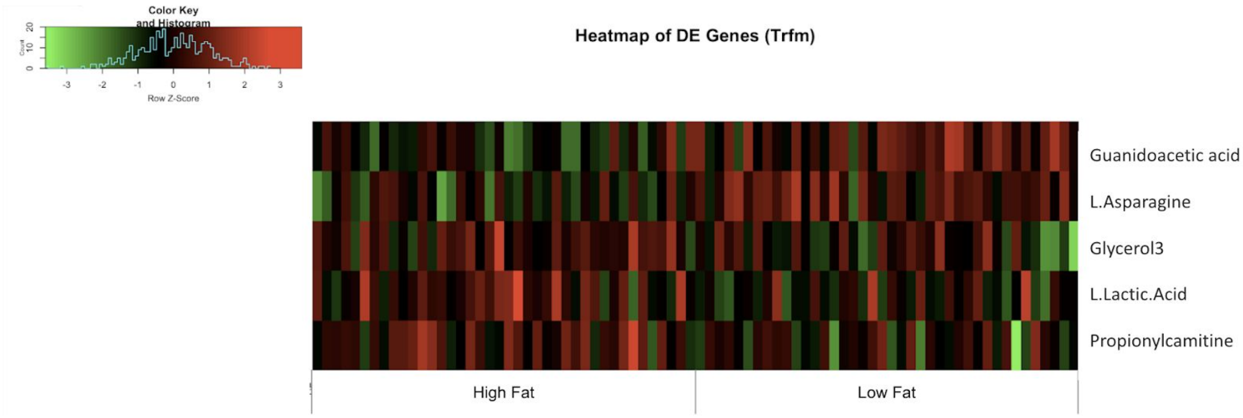


Trunk Fat Mass

[Figure 8] Up-down barlot: Trunk Fat Mass

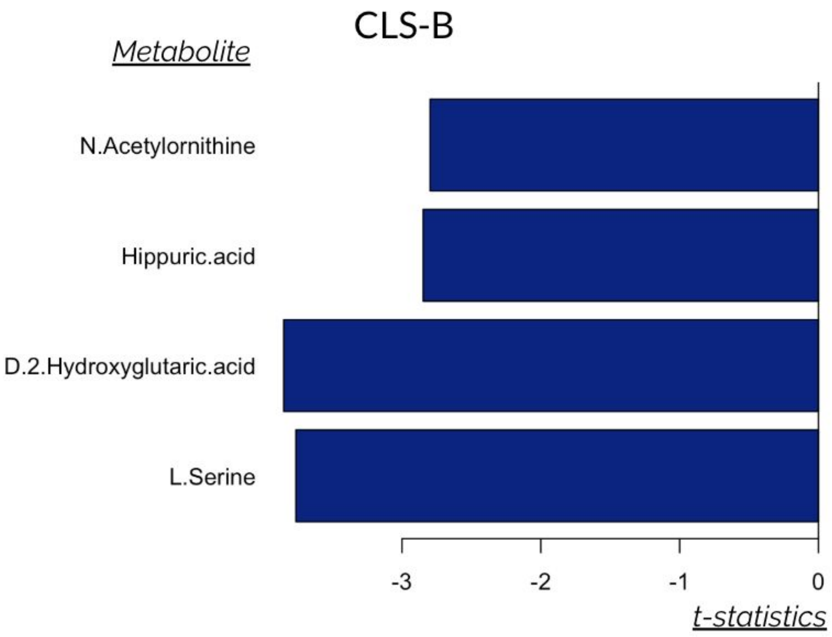


[Figure 9] Heatmap: Trunk Fat Mass

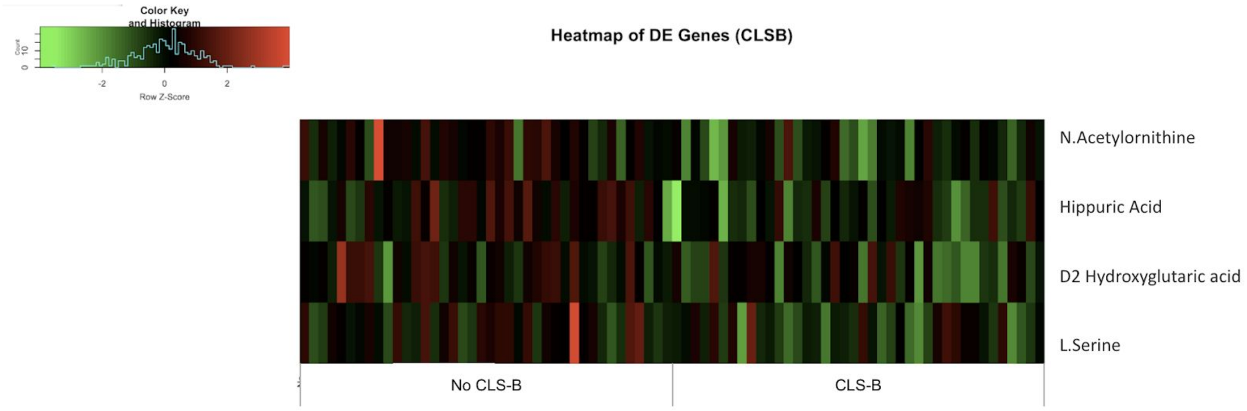


CLS-B

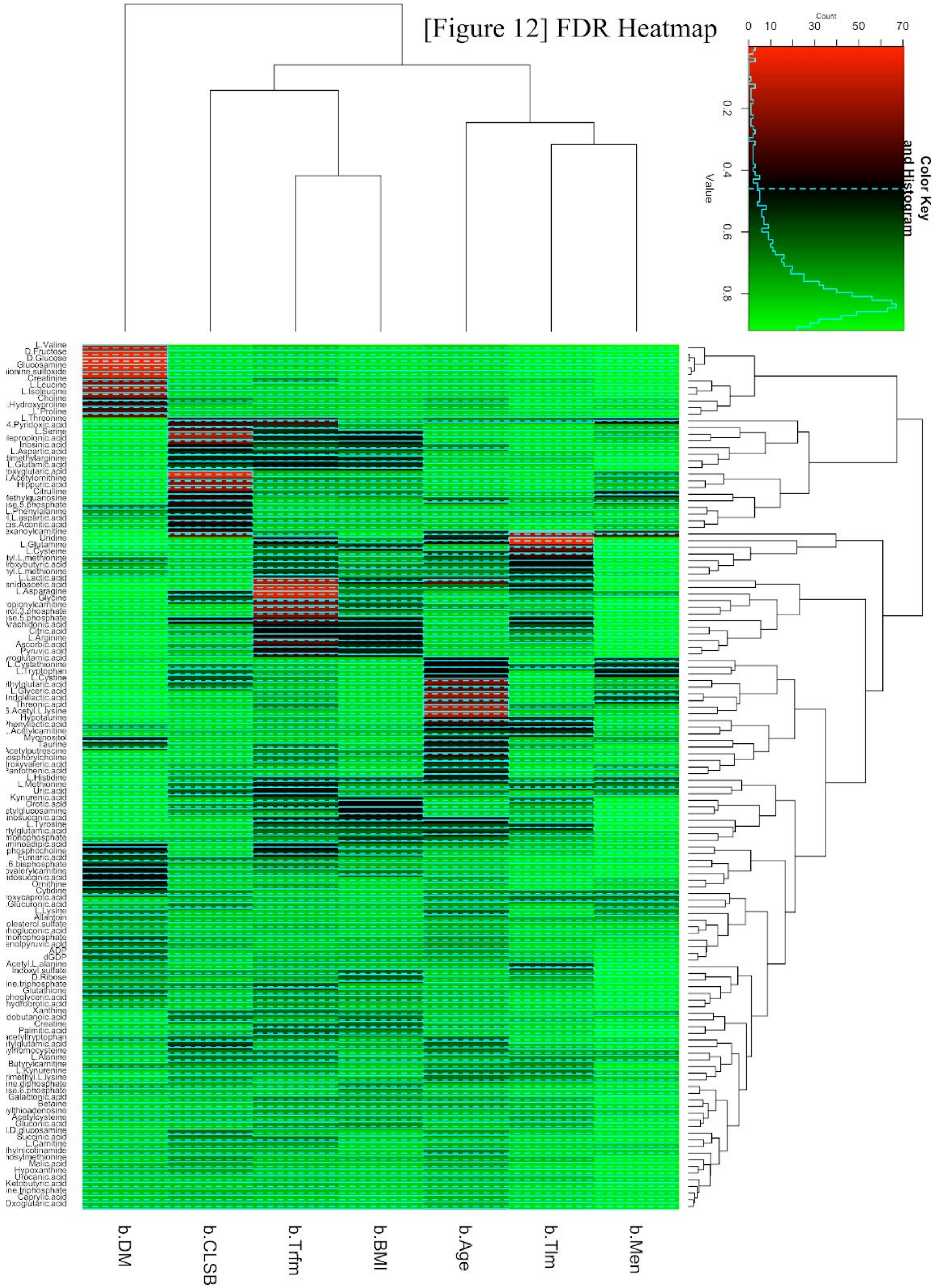
[Figure 10] Up-down barlot: CLS-B



[Figure 11] Heatmap: CLS-B



[Figure 12] FDR Heatmap



References

Zhou, Xi Kathy, Liu, Fei, and Dannenberg, Andrew J. "A Bayesian Model Averaging Approach for Observational Gene Expression Studies." The Annals of Applied Statistics 6.2 (2012): 497-520. Web.

Annest, A., Bumgarner, R.E, Raftery, A.E & Yeung, K.Y. (2009). Iterative bayesian model averaging: a method for the application of survival analysis to high‐dimensional microarray data. BMC Bioinf., 10(1), 1–72.

Suvitaival T, Rogers S, Kaski S. Stronger findings for metabolomics through Bayesian modeling of multiple peaks and compound correlations. Bioinformatics. 2014;30(17):i461–i467. doi:10.1093/bioinformatics/btu455

Kaplan D, Chen J. Bayesian Model Averaging for Propensity Score Analysis. Multivariate Behav Res. 2014;49(6):505–517. doi:10.1080/00273171.2014.928492

C ONLON , E. M., S ONG , J. J., AND L IU , J. S. (2006). Bayesian models for pooling microarray studies with multiple sources of replications. BMC Bioinformatics, 7, 247.

E FRON , B. (2008). Microarrays, empirical Bayes and the two-groups model. Statistical Science. 23(1):122.

I SHWARAN , H., AND R AO , J. S. (2003). Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection. Journal of the American Statistical Association. 98(462), 438–455