BIORANKINGS
TECHNICAL REPORT SERIES

Automating the Analysis of Untargeted LC/MS Metabolomics Data

Untargeted metabolomics measures metabolites in samples to find those that correlate with subgroups (e.g., disease or healthy tissue). Current analysis pipelines are time consuming and require analysts to make ad hoc subjective decisions. This means different researchers will analyze data differently and possibly get different results.

Software developed by our group automates the analysis, removes subjective decision making, and runs fast (on a single cpu <1 hour to run 237 samples each ~25MB/~6GB total, and <2 hours with 8 samples each ~5GB/~40GB total). Linear models are added which introduces a larger class of biostatistical models and methods for researchers (e.g., power and sample size calculations, repeated measures analysis, continuous and categorical covariates, dose response models, ANOVA or factorial designs, mixed models, and block designs). These models allow more complex experimental designs and hypothesis testing to be used in metabolomics research.

This Technical Report shows results for three experiments: a 1-way ANOVA comparing Knockout and Wildtype mice, PCA measuring technical and biological variation, and a 2-way ANOVA simultaneously comparing two genotypes and two diets.

Readers who would like to test this software on their data and compare the results generated with what they have found are invited to do so free of charge. Please email us at contact@biorankings.com and we will arrange for this to happen.

TESTING ONE FACTOR (1-WAY ANOVA)

Conventional LC/MS-based metabolomics data measures the retention time (RT), mass-to-charge ratio (m/z), and metabolite intensity. For a possible metabolite, analysts select a maximum intensity from each sample in the RT x m/z region, shift the RT axis to align these maximums, and perform a univariate statistical test across groups[1].

Our new software differs from current methods by using all the data and not reducing the information to a single number per peak per sample, performs analyses on continuous RT regions avoiding the need to subjectively align peaks by RT shifts, and uses multivariate statistics increasing power compared to univariate methods.

Intensity data from m/z $\in$ [327.1,327.2] and RT $\in$ [3300,3550] for 6 wild-type (WT) and 6 knock-out (KO) mice from the *faah* study[2] is used to illustrate some of the features of our software. KO mice appear (black curves) to have more metabolite than in the WT mice (red curves) in part of this *region of interest* (ROI) (top-left plot in Figure 1). Our software automatically identifies all the m/z x RT ROIs in a dataset. Statistical tests comparing intensity levels are run on each ROI independently. The bottom-left plot shows the statistical test result for this ROI. The red line is the statistical test and the blue line the significance threshold – where the red line is above the blue line we reject the null hypothesis of no difference between KO and WT.

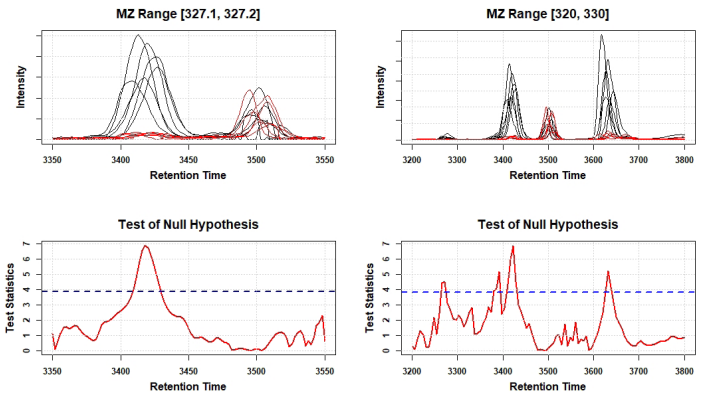

Figure 1

For this analysis, metabolite intensity in approximately RT ∈ [3400,3450] is statistically different. Of interest is where the curves are visually not different shown by the test statistic (red line) below the significance threshold (blue line). Note the peak at RT≈3500 is not statistically different which is confirmed visually by noticing intensity values from both groups overlap.

This analysis applies to any ROI. The top-right plot shows the metabolite intensity for a larger m/z x RT window. Four peaks are found to be statistically significant. The small peak at RT<3300, barely perceptible in the data, is significant, likely because of the increased power of our method. Also note the peak found in the first analysis (smaller ROI window) is split into two peaks suggesting a mixture of two metabolites.

BIOLOGICAL AND TECHNICAL VARIABILITY

Identifying sources of biological and technical variability is an important step in statistical analysis (e.g., analysis-of-variance). Biological variability explains differences due to differences in subjects (e.g., KO versus WT). Technical variability explains differences due to non-biological differences such as sample preparation and machine measurement (e.g., retention time shifts).

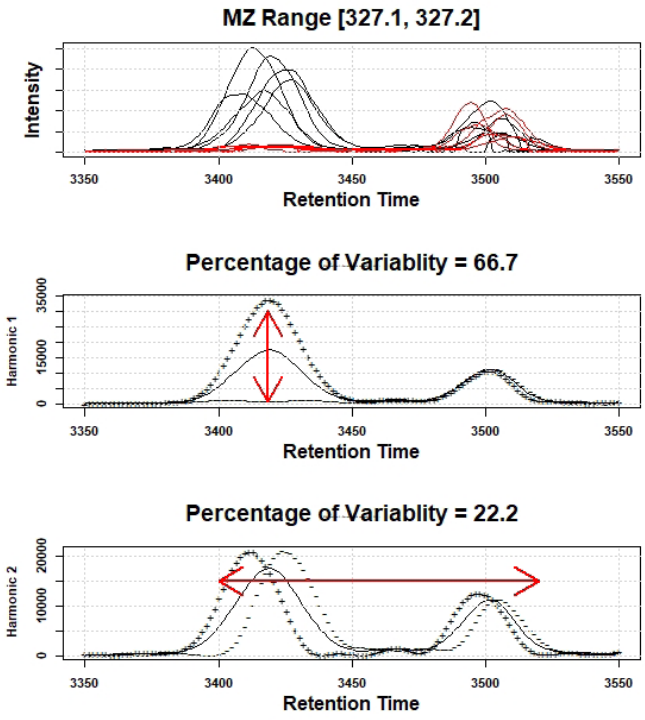

Figure 2

Sources of variation was analyzed in the m/z ∈ [327.1,327.2] and RT ∈ [3300,3550] ROI. Figure 2 shows the raw data (top plot) and two components of variability (middle and bottom plot) explaining 66.7% and 22.2% of the variability, respectively. In the two variability plots the mean metabolite intensity is shown by the solid black lines. The variation range is indicated by the two lines above and below the mean. The vertical arrow in the second plot (66.7% of the variability) suggests biological variation – there is more metabolite in KO mice – and/or technical variation – KO mice samples were run under different conditions than WT mice. The horizontal arrows in the third plot (22.2% of the variability) suggests retention time shifts and provides an automatic rule for RT shifting.

TESTING TWO FACTORS (2-WAY ANOVA)

Factorial experiments are one of the most common designs used in science. These allow one or multiple factors to be studied efficiently.

A 2-way factorial experiment was run with 32 mice to find metabolite differences for two genotypes – KO and WT – and diet interventions – feeding and fasting[3]. The number of samples per factor combination are:

Genotype/Diet	Fasting	Feeding	Total
KO	8	8	16
WT	8	8	16
Total	16	16	32

This design allows main effects to be compared with 16 Fasting versus 16 Feeding mice, and 16 KO versus 16 WT mice. It also allows interaction effects to be measured with 8 mice in each of KO/Fasting, KO/Feeding, WT/Fasting, and WT/Feeding groups.

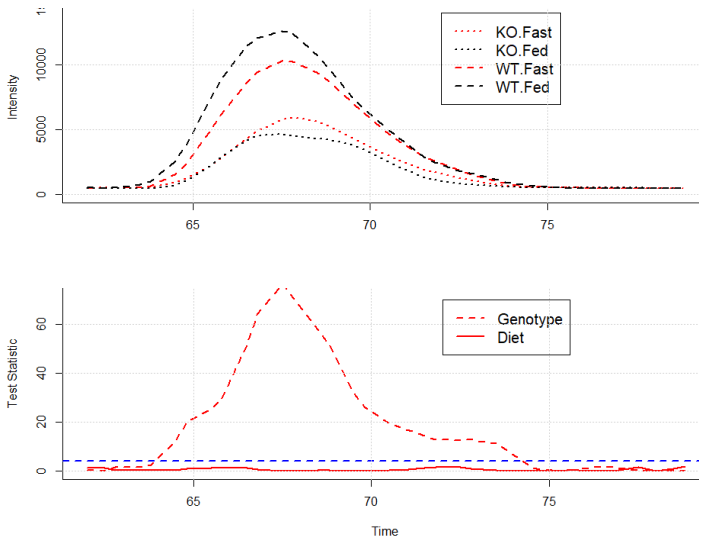

Figure 3

The mean metabolite intensities for the 4 groups are shown in the top plot of Figure 3. A visual comparison shows

genotype is separated (dotted lines are KO, dashed lines are WT) indicating increased metabolite intensity in KO compared to WT. A visual comparison of diet indicates they are not separated (red lines are fasting, black lines are feeding).

The visual analysis is confirmed statistically in the right plot. Genotype (dashed red line) rises above the significance threshold along RT ∈ [64,74] and diet (solid red line) is not significant anywhere (falls below the significance threshold everywhere along RT). The interaction term was also not significant but not shown here to simplify the display.

DISCUSSION

Analysis of untargeted metabolomics data is time consuming and requires subjective decision making about what is a peak and how should data be shifted along the RT axis. This slows down research and means different researchers can obtain different results – both a major roadblock in the use of this technology in science.

In the first year of our Phase II SBIR (R44GM131487) we have developed software that removes these roadblocks:

- The software is automatic, objective, and unbiased for reading in raw untargeted metabolomics, finding all regions of interest where metabolites are found, and apply statistical methods to test for metabolite level differences.
- The software is fast (on a single-cpu <1 hour to run 237 samples each ~25MB/~6GB total, and <2 hours to run 8 samples each ~5GB/~40GB total). Software can also be run in parallel on AWS Cloud.
- Statistical linear models have been added to the software introducing a large class of statistical methods for researchers. Examples are:
  - One-way and multi-way factorial design to test main effects and interactions.
  - Dose response to correlate metabolite levels with dose level.
  - Repeated measures to account for within- and between-subject variance.
  - Regression modeling with continuous and categorical predictors.
  - Power and sample size calculation for experimental design.
  - Mixed models and random block designs.
- No data pre-processing or peak alignment is required.
- Data visualizations and spreadsheets are returned to prioritize follow-up studies.
- Research on annotating metabolites has begun.

BioRankings has partnered with Domino Data Lab (https://www.dominodatalab.com/) as a Solutions Provider (https://www.dominodatalab.com/partners/) and are hosting the software on their AWS cloud platform. Users can access the software through APIs to analyze data in AWS S3 Buckets.

Readers who would like to test this software for free on their data and compare the results generated with what they have found are invited to email us at contact@biorankings.com and we will arrange for this to happen.

REFERENCES

[1] https://pubs.acs.org/doi/10.1021/ac051437y

[2] https://pubs.acs.org/doi/10.1021/bi0480335

[3] Data provided by Patti Lab, Washington Univ.