NMR Metabolomics Analysis
February 09, 2018
Wimal Pathmasiri, PhD
Delisha Stewart, PhD
NIH Common Fund Eastern Regional Comprehensive Metabolomics Resource Core (ERCMRC)

Outline of Today’s Training

- Introduction: Wimal Pathmasiri
- NMR Metabolomics: Wimal Pathmasiri
  - Study Design
  - Sample Preparation
  - Data Acquisition
  - Data Pre-processing
  - Statistical Analysis
  - Library Matching
  - Pathway Analysis
- Hands On Exercise: Delisha Stewart
NMR Metabolomics

- **Broad Spectrum**
  - High throughput
  - NMR Binning
  - Multivariate analysis and other statistics
  - Identifying bins important for separating study groups
  - Library matching of bins to metabolites

- **Targeted Metabolomics**
  - Identifying a set of metabolites
  - Quantifying metabolites
  - Multivariate analysis and other statistics

- **Pathway analysis**
  - Use identified metabolites
  - Use other omics data for integrated analysis

NMR Metabolomics Workflow

[Diagram showing the workflow of NMR Metabolomics]
Free Software available for NMR Metabolomics

- **NMR Data Processing**
  - ACD Software for Academics (ACD Labs, Toronto, Canada)

- **Multivariate data analysis**
  - MetaboAnalyst 3.0 (http://www.metaboanalyst.ca)
  - MetATT (http://metatt.metabolomics.ca/MetATT/)
  - MUMA (http://www.biomolnmr.org/software.html)
  - Other R-packages

- **Library matching and Identification**
  - BATMAN (Imperial College), Bayesil (David Wishart lab)
  - Use of databases
    - Birmingham Metabolite library, HMDB, BMRB

- **Pathway analysis**
  - Metaboanalyst, metaP Server, Met-PA, Cytoscape, KEGG, IMPALA

Also available through www.metabolomicsworkbench.org

---

Other Software available for NMR Metabolomics

**COMMERCIAL**

- **NMR Data-preprocessing**
  - ACD Software (ACD Labs, Toronto, Canada)
  - Chenomx NMR Suite 8.1 Professional

- **Multivariate data analysis**
  - SIMCA 14

- **Other statistical analysis**
  - SAS, SPSS

- **Library matching and quantification**
  - Chenomx NMR Suite 8.1 Professional

- **Pathway analysis**
  - GeneGo (MetaCore Module)
  - Ingenuity Pathway Analysis (IPA)
Sample Preparation, Data Acquisition, and Pre-processing

### Study design Considerations
- Factors such as gender, ethnicity, age, BMI (human studies)
- Species, strains, feed, housing (animal studies)

### Sample collection
- Collection vials, anticoagulant use (heparin, citrate, EDTA)

### Sample storage
- -80 °C is optimal, minimize freeze-thaw cycles
- -20 °C is sometimes more practical (i.e. field studies)

### Sample preparation
- Optimize the methods and use them consistently throughout study
- Daily balance and pipette checks

### Use Quality Check (QC) samples
- Pooled QC samples (Phenotypic and combined pooled samples)
- Use matching external pooled QC samples where pool samples cannot be prepared from study samples

### Optimize all procedures and use them consistently throughout the study

---

**Important Steps in Metabolomics Analysis**

- **Study design Considerations**
  - Factors such as gender, ethnicity, age, BMI (human studies)
  - Species, strains, feed, housing (animal studies)
- **Sample collection**
  - Collection vials, anticoagulant use (heparin, citrate, EDTA)
- **Sample storage**
  - -80 °C is optimal, minimize freeze-thaw cycles
  - -20 °C is sometimes more practical (i.e. field studies)
- **Sample preparation**
  - Optimize the methods and use them consistently throughout study
  - Daily balance and pipette checks
- **Use Quality Check (QC) samples**
  - Pooled QC samples (Phenotypic and combined pooled samples)
  - Use matching external pooled QC samples where pool samples cannot be prepared from study samples
- **Optimize all procedures and use them consistently throughout the study**
Check the samples and the Metadata

- Why are these serum samples straw colored?
  - Are these samples actually plasma or urine?

- Why are there more samples in the box than listed on the inventory emailed?
  - The wrong box was pulled from their biorepository and shipped.

- There is only 3 pieces of dry ice in this box!
  - Did they really pack these “precious samples” in a way to risk them thawing?

- Check every label on the samples shipped to verify they match the inventory.
  - Most sample labels will match, but the wrong tubes can get pulled meaning the right samples were not shipped.
  - Sometimes hand-written labels are illegible and will require further communication to verify the sample ID.

- Check the metadata.
  - Did they really send us female controls to compare with male cases?

- Communicate sample and metadata discrepancies/issues immediately.
  - Use of pictures here can be very helpful.

Sample Preparation for Metabolomics Analysis

Current sample preparation practices (in brief)

- **Biofluids**
  - Dilute with D$_2$O/ buffer/ 0.9% Saline
  - Add internal standard (ISTD, eg. Chenomx) solution or formate (for serum).
  - Centrifuge and transfer an aliquot into NMR tube

- **Tissue and Cells**
  - Homogenization performed in ice cold 50/50 acetonitrile/water
  - Supernatant dried down (lyophilized)
  - Reconstituted in D$_2$O and ISTD (eg. Chenomx) solution

- **Pooled QC Samples (Sample Unlimited)**
  - Mix equal volume of study samples to get pooled QC samples
  - 10% QC samples

- **Pooled QC Samples (Sample Limited)**
  - Use independent pool of similar samples
  - 10% QC samples

- **Daily balance and pipette check**

Samples are randomized for preparation and data acquisition
Preparing Pooled QC Samples

- Aliquots from each sample in the study phenotype are pooled (phenotypic pool)
- Equal amount of each phenotypic pools are pooled (Combined phenotypic pool)
- Replicates of pools are prepared
- Pool samples are prepared along with the study samples

Pooled samples should cluster tightly

NMR Data Acquisition

- 1D NMR
  - 1st increment of NOESY
    - noesyprid (Bruker)
  - CPMG (serum or plasma)
    - cpmgr1d (Bruker)
    - To remove broadening of signals due to macromolecules (eg. Proteins and lipids)

- 2D NMR (for structure elucidation)
  - 2D J-Resolved
  - COSY
  - TOCSY
  - HSQC
  - HMBC
Sample Amount in NMR tube

- At least 10% D₂O in the sample
- Optimum volume
  - 550 – 600 uL (5mm tube)
  - 200 uL (3 mm tube)
- Sample gauge is used

For very small sample amounts, a NMR with a microcoil probe is an option.

NMR Data

- A typical 1H NMR Spectrum consists of thousands of sharp lines or signals.
- The intensity of the peak is directly related to the number of protons underlying the peak.
- The position of a particular peak in the X-axis of the NMR spectrum is called the “Chemical Shift” and it is measured in ppm scale.
- The NMR spectrum obtained for the biological sample is referenced using a reference compound such as DSS, TSP, or Formate added to the sample in sample preparation step.
- pH indicator may also be used (for example, Imidazole)

DSS=4,4-dimethyl-4-silapentane-1-sulfonic acid, TSP=Trimethylsilyl propionate
### 1H NMR Spectrum for Alanine

- **Peak position is called “chemical shift”**
- It depends on the chemical environment
- Splitting of peak is dependent on neighboring 1H atom(s)
- Area under peak proportional to the number of 1H atoms underlying it

**Chemical shift of CH**
- 4 lines (quadrat)

**Chemical shift of CH3**
- 2 lines (doublet)

---

### Typical 1H NMR Spectrum of Urine
Typical $^1$H NMR Spectrum of Urine

Data Pre-processing

- After NMR data acquisition, the result is a set of spectra for all samples.
- For each spectrum, quality of the spectra should be assessed.
  - Line shape, Phase, Baseline
- Spectra should be referenced
  - Compounds commonly used: DSS, TSP, Formate
- Variations of pH, ionic strength of samples has effects on chemical shift
  - Peak alignment
  - Binning or Bucket integration
  - High quality data are needed
- Remove unwanted regions
- Normalize data (remove variation in concentration of samples)
Quality Control Steps

- Quality of metabolomics analysis depends on data quality
- Typical problems
  - Water peak (suppression issues)
  - Baseline (not set at zero and not a flat line)
  - Alignment of peaks (chemical shift, due to pH variation)
  - Variation in concentration (eg. Urine)
- High quality of data is needed for best results

Water Suppression Effects and Other Artifacts

- If water is not correctly suppressed or removed there will be effects on normalization
- Need to remove other artifacts
- Remove drug or drug metabolites

![Poor water suppression vs. Good water suppression](image-url)
NMR Pre-processing

Before

After

Phase
Baseline to be corrected

Reference,
Line shape

pH Dependence of Chemical Shift

Chemical shift variability
- pH
- ionic strength
- metal concentration

Methods to overcome this problem
- Use a buffer when preparing samples
- Binning (Bucketing)
  - Fixed binning
  - Intelligent binning
  - Optimized binning

Available data alignment tools
- Recursive Segment-wise Peak Alignment (RSPA)
- Icoshift
- speaq

http://www.chenomx.com/software/software.php
Vu, T. N. et al., BMC Bioinformatics 2011, 12:405
Peak Alignment

Example

One of the Citrate peaks

a) before

b) after


Peak Alignment

Example

speaq

Vu, T. N. et al., BMC Bioinformatics 2011, 12:405
NMR Binning

- A form of quantification that consists of segmenting a spectrum into small areas (bins/buckets) and attaining an integral value for that segment

- Binning attempts to minimize effects from variations in peak positions caused by pH, ionic strength, and other factors.

- Two main types of binning
  - Fixed binning
  - Flexible binning

Peak shift can cause the same peak across multiple samples to fall into different bins

The entire NMR spectrum is split into evenly spaced integral regions with a spectral window of typically 0.04 ppm.

The major drawback of fixed binning is the non-flexibility of the boundaries.

If a peak crosses the border between two bins it can significantly influence your data analysis.

Signals for citrate are split into multiple bins
NMR Binning

Signals for citrate are split into multiple bins

Signals for citrate are properly captured

Fixed Binning

Smart Binning

NMR Binning

Normalized intensity

Chemical Shift (ppm)
**NMR Binning**

- Integrate bins (0.04 ppm bin size)
- Normalize integral of each bin to the total integral of each spectrum
- Merge metadata
- Result is a spreadsheet ready for further multivariate data analysis and other statistical analysis

---

**Sample ID**

**Disease Group**

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Control</th>
<th>Cases</th>
<th>Metal Data</th>
<th>Normalized Binned Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0586</td>
<td>Control</td>
<td>Cases</td>
<td>0.00E+00</td>
<td>1.69E-02</td>
</tr>
<tr>
<td>C2177</td>
<td>Control</td>
<td>Cases</td>
<td>0.00E+00</td>
<td>3.02E-02</td>
</tr>
<tr>
<td>D0909</td>
<td>Control</td>
<td>Cases</td>
<td>0.00E+00</td>
<td>1.08E-02</td>
</tr>
<tr>
<td>D0945</td>
<td>Control</td>
<td>Cases</td>
<td>0.00E+00</td>
<td>4.79E-04</td>
</tr>
<tr>
<td>D1174</td>
<td>Control</td>
<td>Cases</td>
<td>0.00E+00</td>
<td>9.33E-04</td>
</tr>
<tr>
<td>D0729</td>
<td>Control</td>
<td>Cases</td>
<td>0.00E+00</td>
<td>3.43E-03</td>
</tr>
<tr>
<td>D0485</td>
<td>Control</td>
<td>Cases</td>
<td>0.00E+00</td>
<td>1.55E-03</td>
</tr>
<tr>
<td>D0262</td>
<td>Control</td>
<td>Cases</td>
<td>2.39E-05</td>
<td>6.04E-02</td>
</tr>
<tr>
<td>D0379</td>
<td>Control</td>
<td>Cases</td>
<td>2.73E-02</td>
<td>1.81E-03</td>
</tr>
</tbody>
</table>
Data Normalization, Transformation, and Scaling

Normalization

- Normalization reduces the sample to sample variability due to differences in sample concentrations—particularly important when the matrix is urine
  - Normalization to total intensity is the most common method
    - For each sample, divide the individual bin integral by the total integrated intensity
  - Other Methods
    - Normalize to a peak that is always present in the same concentration, for example, normalizing to creatinine
    - Probabilistic quotient normalization
    - Quantile and cubic spline normalization
### Centering, Scaling, and Transformations

<table>
<thead>
<tr>
<th>Method</th>
<th>Transformation Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I. Centering</strong></td>
<td>( \bar{x}<em>{ij} = x</em>{ij} - \bar{x}_i )</td>
</tr>
<tr>
<td><strong>II. Autoscaling</strong></td>
<td>( \bar{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}_i}{s_i} )</td>
</tr>
<tr>
<td>Range scaling</td>
<td>( \bar{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}<em>i}{x</em>{\text{max}} - x_{\text{min}}} )</td>
</tr>
<tr>
<td>Pareto scaling</td>
<td>( \bar{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}_i}{\sqrt{s_i}} )</td>
</tr>
<tr>
<td>Vast scaling</td>
<td>( \bar{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}_i}{\bar{s}_i} )</td>
</tr>
<tr>
<td>Level scaling</td>
<td>( \bar{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}_i}{\bar{x}_i} )</td>
</tr>
<tr>
<td><strong>III. Log transformation</strong></td>
<td>( \bar{x}<em>{ij} = 10 \log(x</em>{ij}) )</td>
</tr>
<tr>
<td></td>
<td>( \bar{x}<em>{ij} = \log(x</em>{ij}) )</td>
</tr>
<tr>
<td>Power transformation</td>
<td>( \bar{x}<em>{ij} = \sqrt{x</em>{ij}} )</td>
</tr>
</tbody>
</table>

Analysis results vary depending on the scaling/ transformation methods used.

Van den Berg et al 1006, BMC Genomics, 7, 142

---

### Data Transformation

- **Before transformation**
  - Skew distribution
  ![Histogram of non-transformed data](image1.png)

- **After log-transformation**
  - More close to normal distribution
  ![Histogram of log-transformed data](image2.png)

Susan Wicklund, Multivariate data analysis for omics, Sept 2-3 2008, Umetrics training
Scaling

- Unit variance (autoscaling) divides the bin intensity by the standard deviation
  - May increase your baseline noise
  - Dimensionless value after scaling

- Pareto scaling divides the bin intensity by the square root of the standard deviation
  - Not dimensionless after scaling

- For NMR data, centering with pareto scaling is commonly used

Multivariate Data Analysis and Other Statistical Analyses

- Mean centered and scaled data
- Non-supervised analysis
  - Principal component analysis (PCA)
- Supervised analysis
  - PLS-DA and OPLS-DA
- Loadings plots and VIP Plots to identify discriminatory bins
- p-Value, fold change
Library Matching Pathway Analysis

Chenomx Library

- Over 320 metabolites
- pH sensitive library of 1H NMR Spectra
- Customizable
Chemical Shift and pH Dependence

Source: http://www.chenomx.com/software/

NMR Spectrum of Urine with Chenomx Library Fit of Metabolites
Fitting of Metabolites

Fitting Taurine
Fitting Fucose

Additive Fit
Additive Fit

Chenomx Helps Resolving Ambiguity in Highly Overlapped Regions
Once we have performed a metabolomics analysis:

- We find some important metabolites that are responsible for the separation of study groups.
- The next questions are
  - What does it mean?
  - How do you correlate these finding to your study questions?
  - Does it explain any findings that are meaningful for your study hypotheses?
  - Does it generate a new hypothesis?
- How do you answer these questions?
  - Next step is to interpret results and perform metabolic pathway analysis

Interpreting Results and Pathway Analysis

- There are a number of freely available software
  - meta-P Server, Metaboanalyst, Met-PA, web based KEGG Pathways, Cytoscape.
  - GeneGo, Ingenuity Pathway Analysis (Commercial)
- Another way of interpreting metabolomics results is to use traditional biochemistry text books.
- The input for pathway analysis is typically a list of metabolites (with any fold change or p-value information)
- Genomics, transcriptomics, and/or proteomics data can be integrated
- Once these pathways are identified, you may perform a targeted metabolomics analysis to validate the findings from global analysis.
Study Example

Day 0 serum- Predicting Day 28 Response to Vaccine

PLS-DA
Day 0 – High Responders (Black) vs Low Responders (Red)

Subset of Metabolites that Influence the Separation of Subjects at Day 0
(VIP ≥ 1 or p-value ≤ 0.1)

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoleucine**</td>
<td>Creatinine**</td>
</tr>
<tr>
<td>Leucine**</td>
<td>Cysteine**</td>
</tr>
<tr>
<td>Valine</td>
<td>Histidine</td>
</tr>
<tr>
<td>3-Methyl-2-oxo-isovalerate</td>
<td>Choline</td>
</tr>
<tr>
<td>3-Hydroxybutyrate</td>
<td>Glucose</td>
</tr>
<tr>
<td>Lactate</td>
<td>Betaine</td>
</tr>
<tr>
<td>Alanine*</td>
<td>TMAO</td>
</tr>
<tr>
<td>Acetate**</td>
<td>Glycine</td>
</tr>
<tr>
<td>Proline*</td>
<td>Glycerol</td>
</tr>
<tr>
<td>Glutamate**</td>
<td>Serine</td>
</tr>
<tr>
<td>Glutamine**</td>
<td>Creatine</td>
</tr>
<tr>
<td>Pyruvate</td>
<td>Tyrosine*</td>
</tr>
<tr>
<td>2-Oxoisocaproate</td>
<td>Histidine</td>
</tr>
<tr>
<td>Methylguanidine**</td>
<td>Tryptophan</td>
</tr>
<tr>
<td>Formate</td>
<td>Phenylalanine</td>
</tr>
</tbody>
</table>

*p-value < 0.05, **p-value ≤ 0.1

Preliminary results
Day 0 High vs Low Responders

GeneGo Network Analysis

- Metabolites that linked in the pathways

Preliminary results

Literature

NMR data acquisition is performed by using methods cited in Beckonert et al. (2007), Nature Protocols, 2 (11), 2692-2703.


Vu, T. N. et al., BMC Bioinformatics 2011, 12:405
NMR Metabolomics
Hands On Exercise

NMR Hands On Exercise: Study Design

- Drug Induced Liver Injury (DILI) Study using Rat Model

- 3 Study groups and 2 time points
  - Vehicle Control (time matched)
  - Low Dose (“no effect” level, Day 01 and Day 14)
  - High Dose (Day 01 and Day 14)

- 24h Urine collected

- Samples prepared by mixing an aliquot of urine with Phosphate buffer + Chenomx ISTD (DSS, $D_2O$, and Imidazole)
  - DSS (Chemical shift and line shape reference)
  - Imidazole (pH reference)
NMR Binned Data

- Three (3) Spreadsheets provided
  1. UAB_RFA_Metaboanalyst.csv
  2. UAB_RFA_Metaboanalyst_D14_NoPools.csv
  3. UAB_RFA_Metaboanalyst_D14_Vehicle_vs_HighDose.csv

- Spreadsheets 2-3 were derived from the initial spreadsheet no. 1 (for easy upload into Metaboanalyst in the subsequent analyses)

Metaboanalyst

Please go to the webpage: http://wwMw.metaboanalyst.ca/MetaboAnalyst/
MetaboAnalyst: Functional Modules

MetaboAnalyst: Data Upload
MetaboAnalyst: Data Integrity Check

Data Integrity Check:
1. Checking the class labels - at least three variables are required in each class.
2. If the samples are paired, then the pair space must conform to the same format.
3. The data file can also contain internal content related to values.
4. The presence of missing values or features with constant values (i.e., 0 or 1)

Data processing information:
- Checking data comment, unpaired:
- Samples are in rows and features in columns.
- The uploaded file is in comma-separated values (csv) format.
- The uploaded file contains 20 samples by 21 (species) times data matrix.
- 7 groups where each group contains 4 samples.
- Samples are not paired.
- All data types are numeric.
- A total of 0 (zero) missing values were detected.
- To include the variables with missing values for output values.
- Check: Missing value imputation to be used in further methods

MetaboAnalyst: Data Filtering

Data Filtering:
The objective of the data filtering is to identify and remove variables that are unlikely to be observed when analyzing the data for meaningful information and provide a filtering process. The data can be filtered at an individual level with a standard deviation. This data filtering recommends the use of integrated microarrays between organs, serum (serum), and several different sets of variables. Non-mean values are included in the filtering process when they are not significant. The following statistical tests are applied during data filtering:
- Leave one variable out at a time (leave-one-out)
- Leave one variable out at a time (leave-one-out)
- Leave one variable out at a time (leave-one-out)
- Leave one variable out at a time (leave-one-out)

Please note, in order to ensure the computed values in the output, the final option is only for a maximum of 2000 features. However, if you choose the top, the rest will still be applied. In addition, the maximum allowed number of variables is 2000. However, 2000 variables were filtered filtering with the top 1000 will be used in the subsequent analysis.

[Images of the MetaboAnalyst interfaces for data integrity check and data filtering]
MetaboAnalyst: Statistical Analysis

Principal Component Analysis (PCA): All Samples
Clustering of Pooled QC Samples

PCA Scores and Loadings Plots: Day 01 and Day 14
Day 14: Vehicle, Low Dose, and High Dose Groups

- We will compare high dose vs vehicle
  - 2. UAB_RFA_Metaboanalyst_D14_NoPools.csv
- Perform PCA
- Perform PLS-DA
- Heat map

Please go back to the start page and upload the data

PCA Scores and Loadings Plots: Day 14

Vehicle, Low Dose, and High Dose groups
Vehicle, Low Dose, and High Dose groups

PLS-DA Scores and Loadings Plots: Day 14

Heat Map: Day 14 Samples
Comparison of Day 14 High Dose and Vehicle

- We will compare high dose vs vehicle
  - 3. UAB_RFA_Metaboanalyst_D14_Vehicle_vs_HighDose.csv

- Perform PCA

- Perform PLS-DA

- VIP Plot

- Heat map

---

PCA Scores and Loadings Plots: Day 14 High Dose vs Vehicle
PLS-DA Scores and Loadings Plots: Day 14 High Dose vs Vehicle

VIP Plot of PLS-DA: Day 14 High Dose vs Vehicle

VIP = Variable Influence on Projections
OPLS-DA Scores Plot and S-Plot: Day 14 High Dose vs Vehicle

Heat Map: Day 14 High Dose vs Vehicle
Contributors through the Years

If you have any questions, please e-mail me

wimal_pathmasiri@unc.edu

Useful link:
Metabolomics Workbench
http://www.metabolomicsworkbench.org/

Thank You!