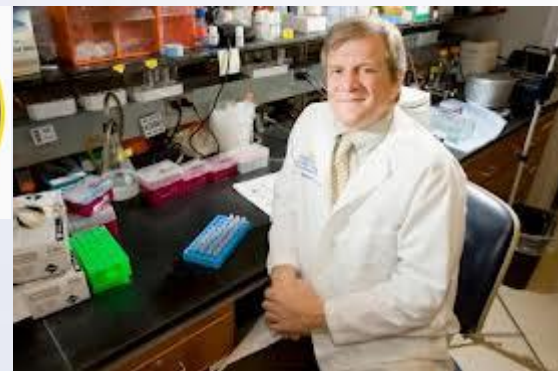


# NMR Data Pre-processing

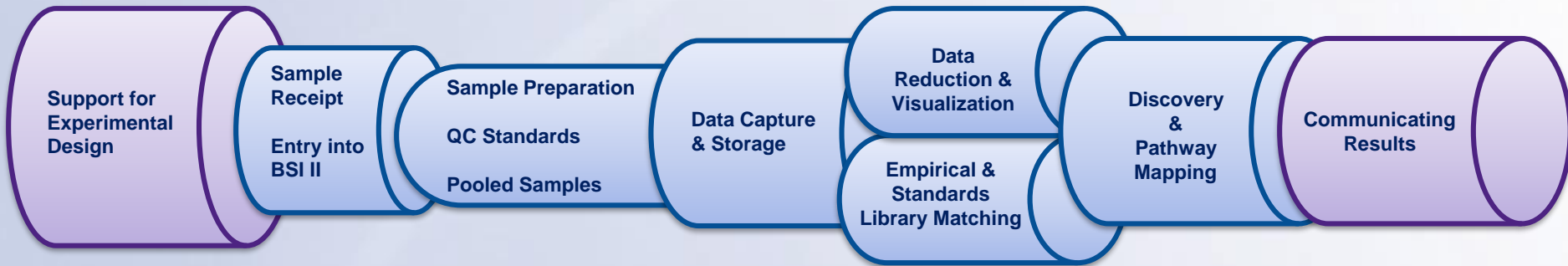
UAB Metabolomics Training Course  
July 17-21, 2017

Wimal Pathmasiri and Delisha Stewart  
NIH Eastern Regional Comprehensive Metabolomics Resource Core  
(ERCMRC)  
Department of Nutrition – Nutrition Research Institute  
University of North Carolina at Chapel Hill

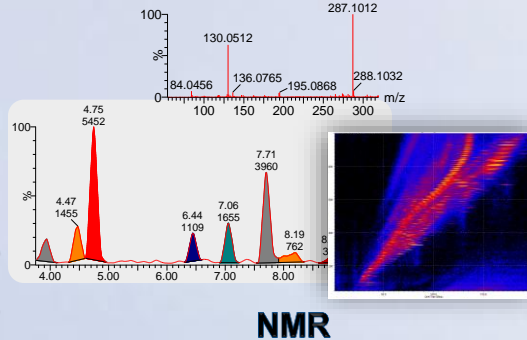
# NIH Common Fund Metabolomics Cores



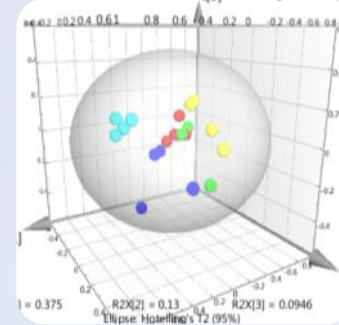
# NIH Eastern Regional Comprehensive Metabolomics Resource Core at NRI



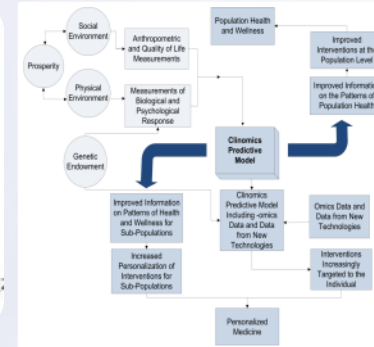
## LC-MS, MS/MS and ION-MOBILITY



## Multivariate and Statistical Analysis



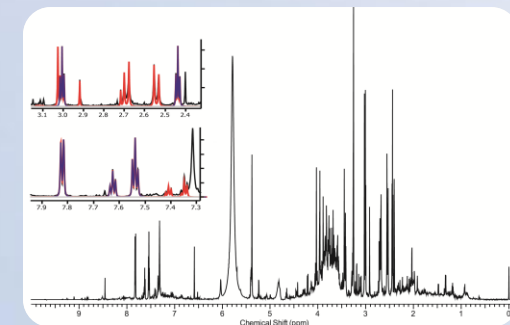
## Predictive Modeling



T  
A  
R  
G  
E  
T  
E  
D

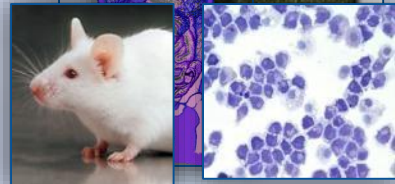
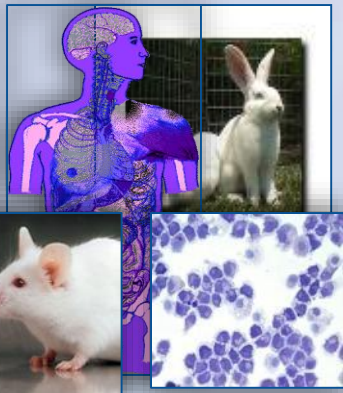
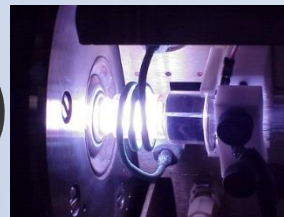
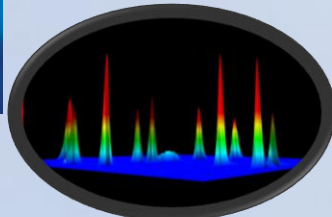
B  
R  
O  
A  
D  
S  
P  
E  
C  
T  
R  
U  
M

## Pathway Mapping Metabolites-Proteins-Genes



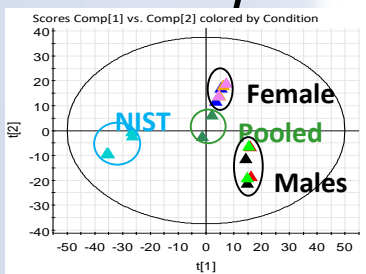
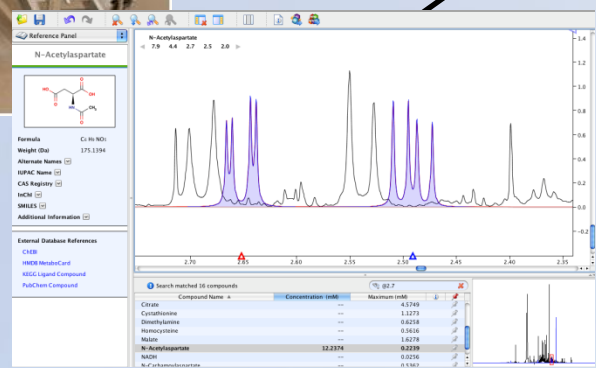
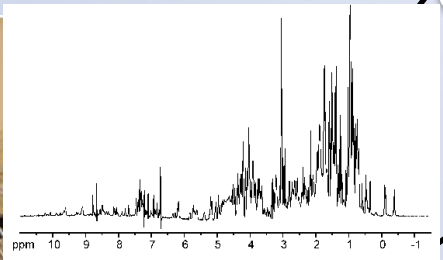
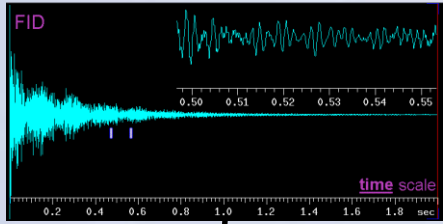
## GC-MS

## ICP-MS



# NMR Metabolomics Workflow

Peak No.	Retention Time (min)	Chemical Name
1	0.25	Acetic acid
2	0.35	Formic acid
3	0.45	Glucose
4	0.55	Fructose
5	0.65	Sucrose
6	0.75	Galactose
7	0.85	Mannose
8	0.95	Starch
9	1.05	Cellulose
10	1.15	Lignin
11	1.25	Chitin
12	1.35	Chitosan
13	1.45	Protein
14	1.55	Lipid
15	1.65	Carbohydrate
16	1.75	Alkaloid
17	1.85	Flavonoid
18	1.95	Terpene
19	2.05	Phenol
20	2.15	Alcohol
21	2.25	Amine
22	2.35	Imine
23	2.45	Enone
24	2.55	Aldehyde
25	2.65	Ketone
26	2.75	Carbonyl
27	2.85	Acid
28	2.95	Base
29	3.05	Salt
30	3.15	Complex



# Data Pre-processing

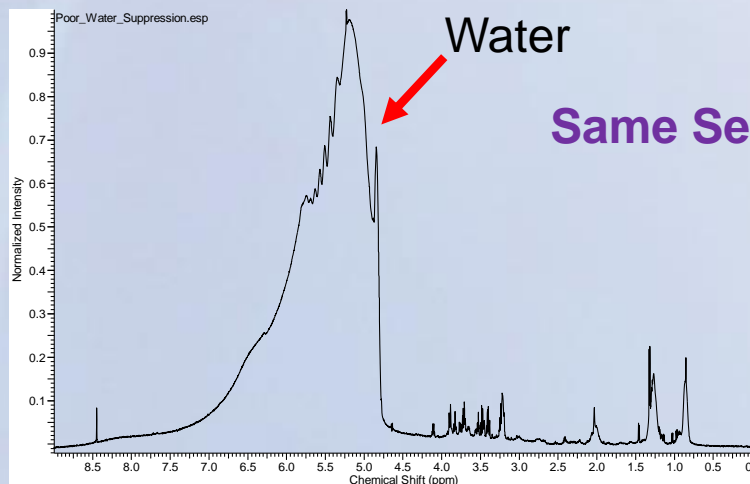
- After NMR data acquisition, the result is a set of spectra for all samples.
- For each spectrum, quality of the spectra should be assessed.
  - Line shape
  - Phase
  - Baseline
- Spectra should be referenced
  - Compounds commonly used: DSS, TSP, Formate
- Variations of pH, ionic strength of samples has effects on chemical shift
  - Peak alignment
  - Bucket integration
- Remove unwanted regions

# Quality Control Steps

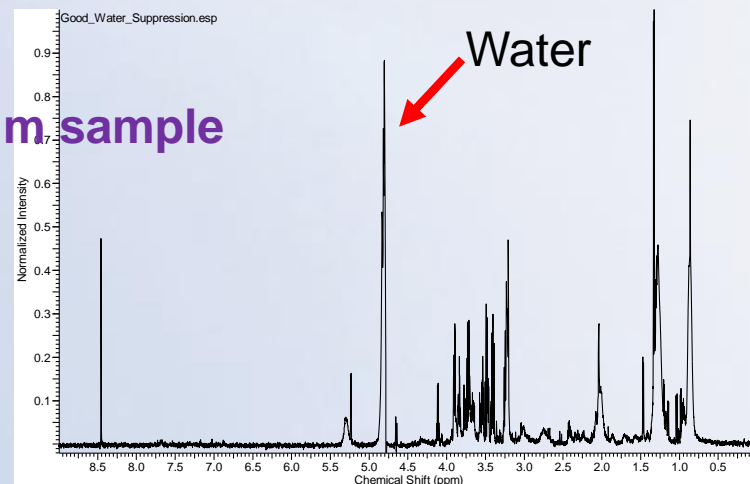
- Quality of metabolomics analysis depends on data quality
- Typical problems
  - Water peak (suppression issues)
  - Baseline (not set at zero and not a flat line)
  - Alignment of peaks (chemical shift, due to pH variation)
  - Variation in concentration (eg. Urine)
- High quality of data is needed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

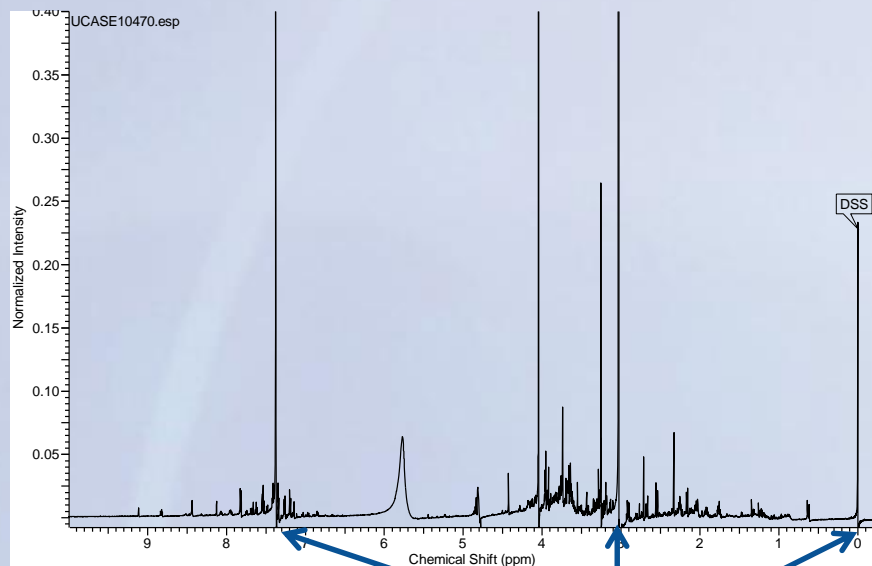


**Good water suppression**

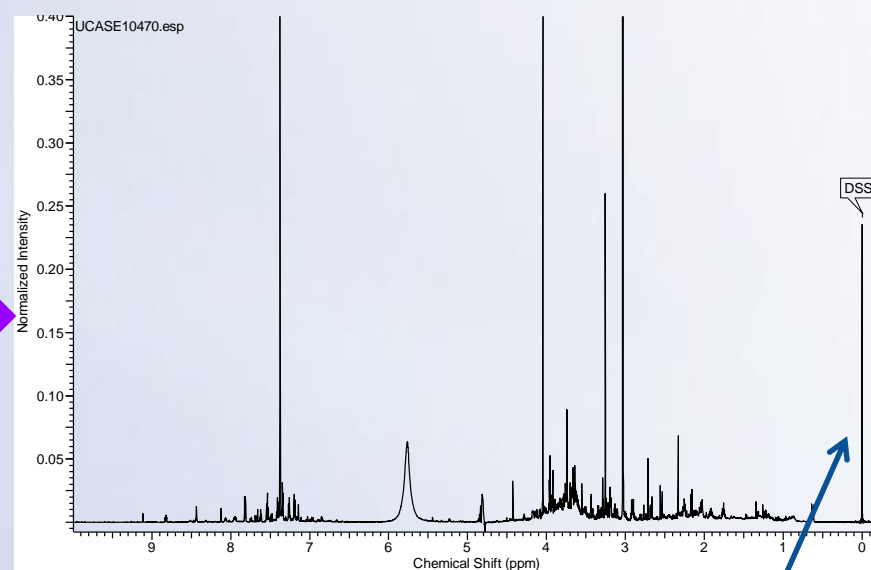
Same Serum sample

# NMR Pre-processing

Before



After



Phase  
Baseline to be  
corrected

Reference,  
Line shape



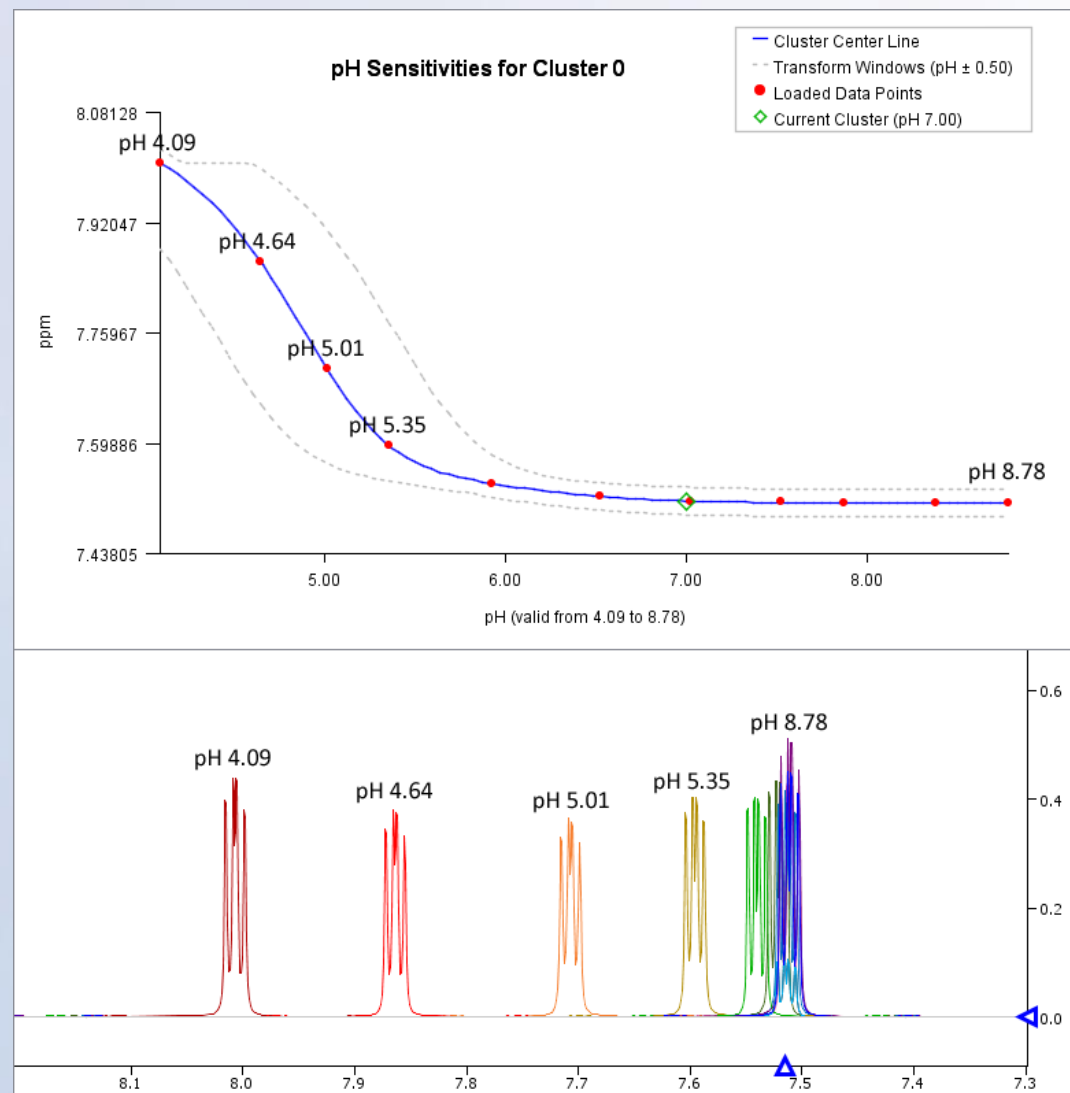
# pH Dependence of Chemical Shift

## 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>

Savorani, F. et al., *Journal of Magnetic Resonance*, Volume 202, Issue 2, 2010, 190 – 202

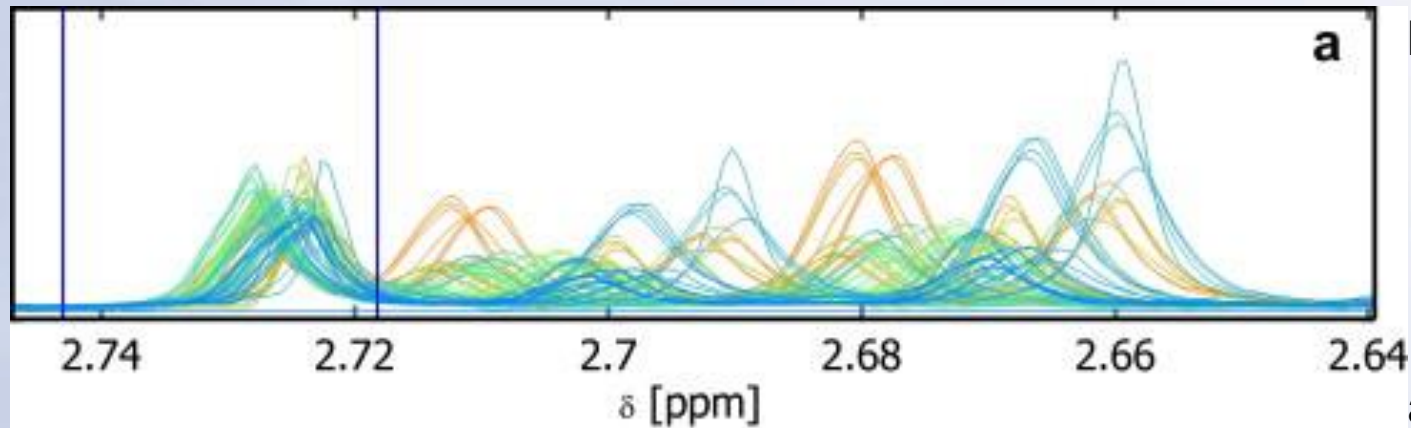
Vu, T. N. et al., *BMC Bioinformatics* 2011, **12**:405

# Peak Alignment

Example

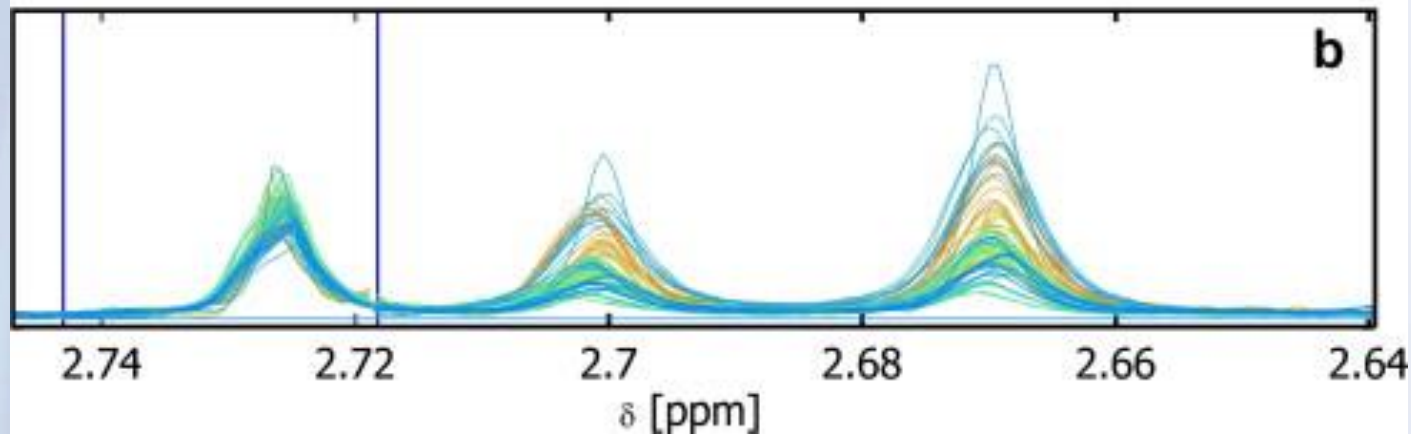
icoshift

One of the Citrate peaks



before

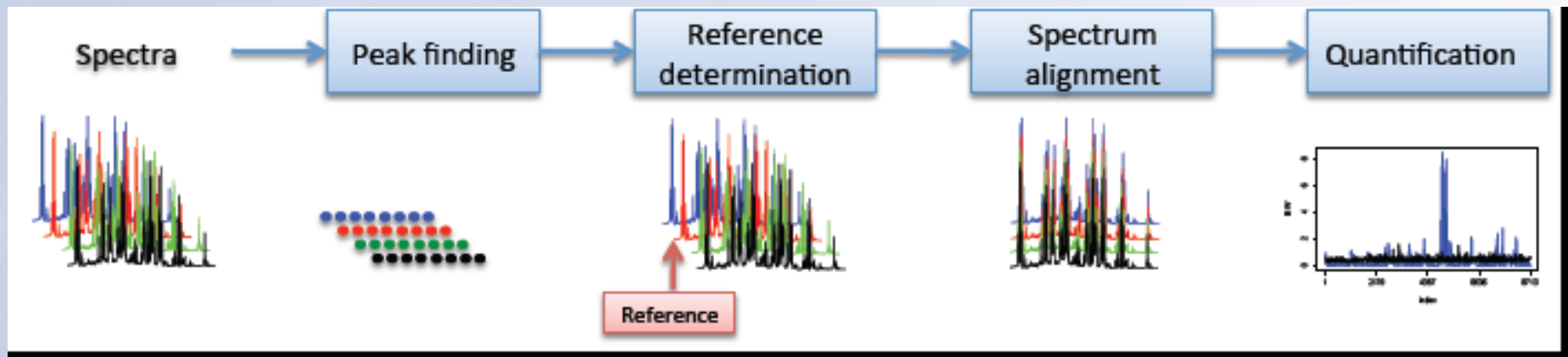
after



# Peak Alignment

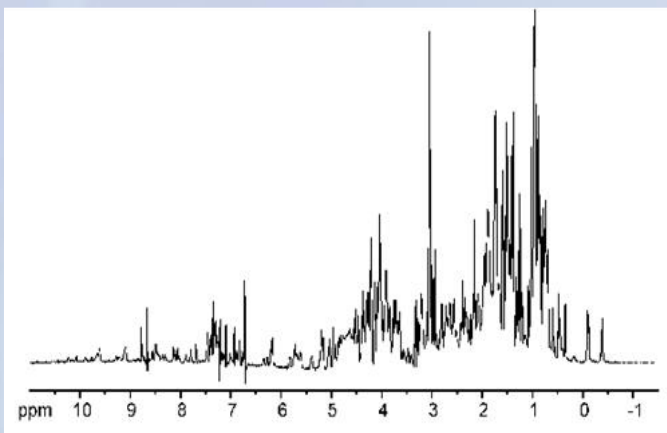
Example

speaq



# 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

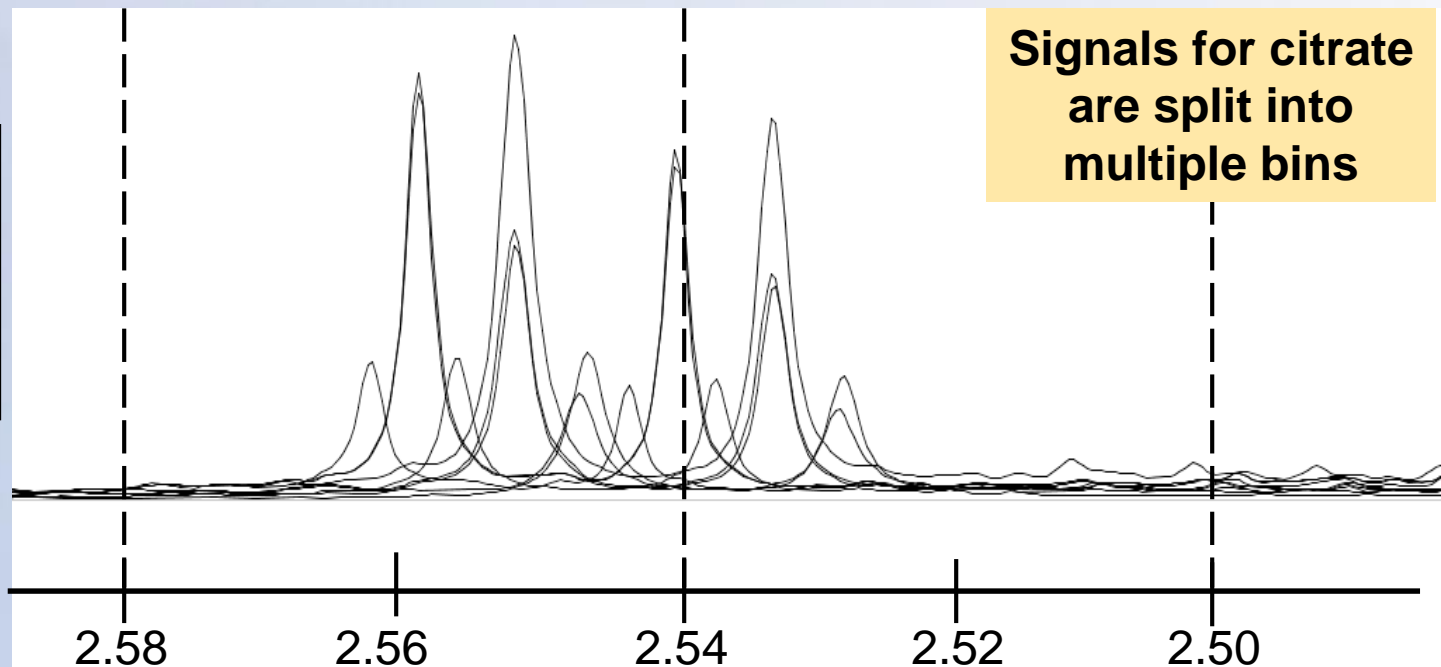


The screenshot shows a spreadsheet with columns for 'Spectrum', 'ACQ Spectrum File No.', 'Time', 'Date', 'Dose Group', 'Dose', 'Sample No.', 'Average', and a series of columns for 'NO. 05', '0.1130.13', '0.1310.13', '0.1730.17', '0.2130.21', '0.2530.25', '0.2930.29', '0.3130.31', '0.3330.33', '0.3530.35', '0.3730.37', '0.4130.41', '0.4330.43', '0.4530.45'. The data rows contain numerical values for each of these columns, representing the results of the binning process for various samples and conditions.

# NMR 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

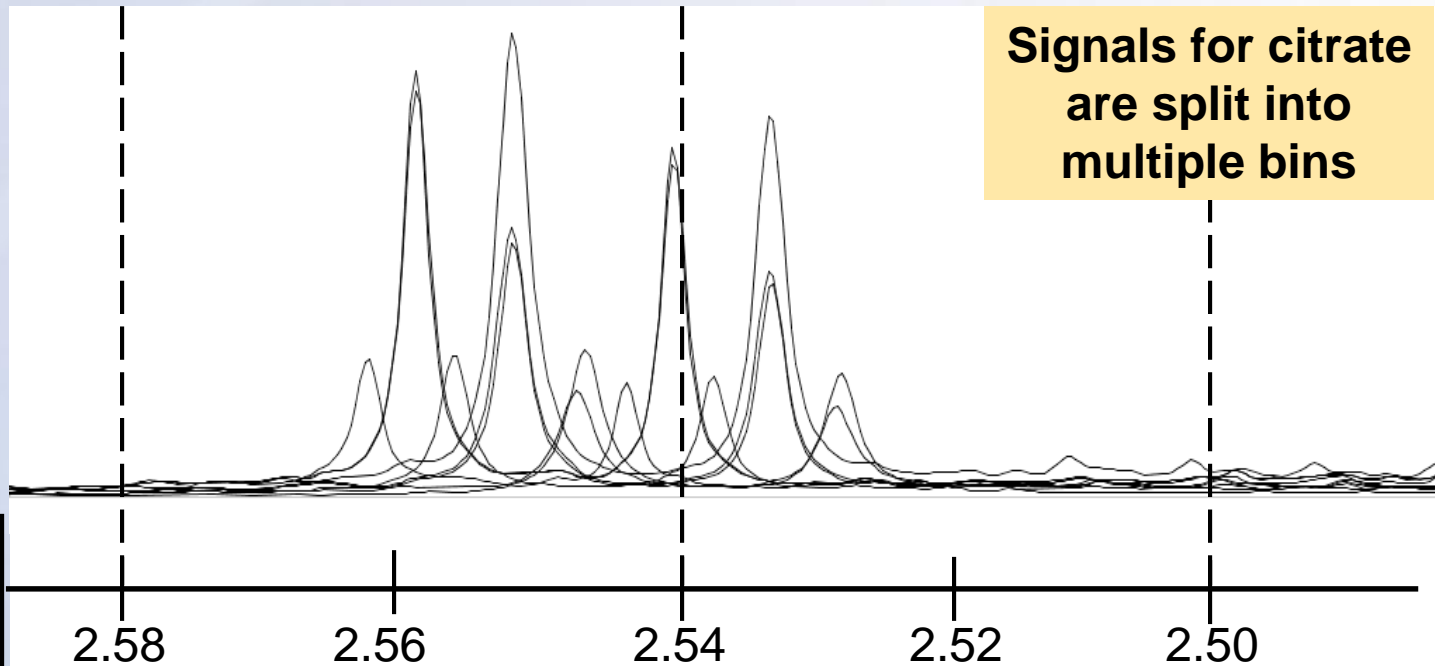


**Signals for citrate are split into multiple bins**

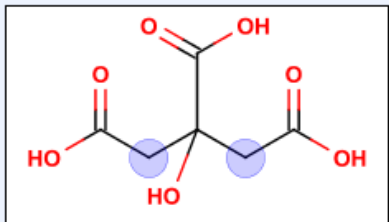
**Fixed Binning**

# NMR Binning

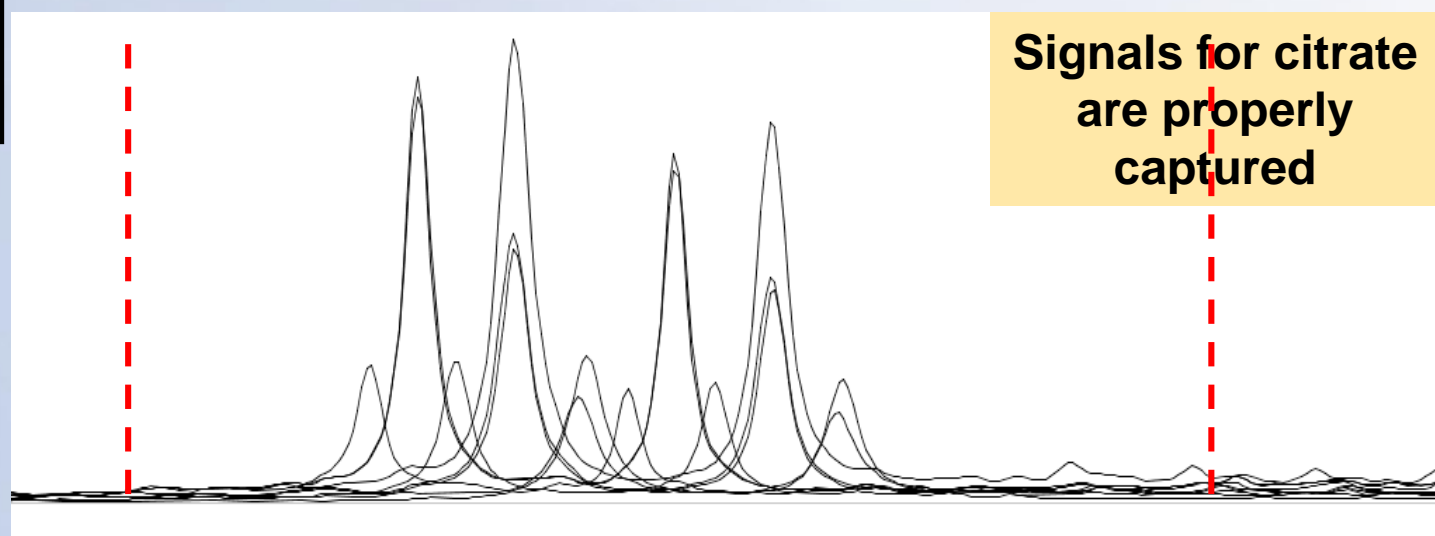
**Fixed Binning**



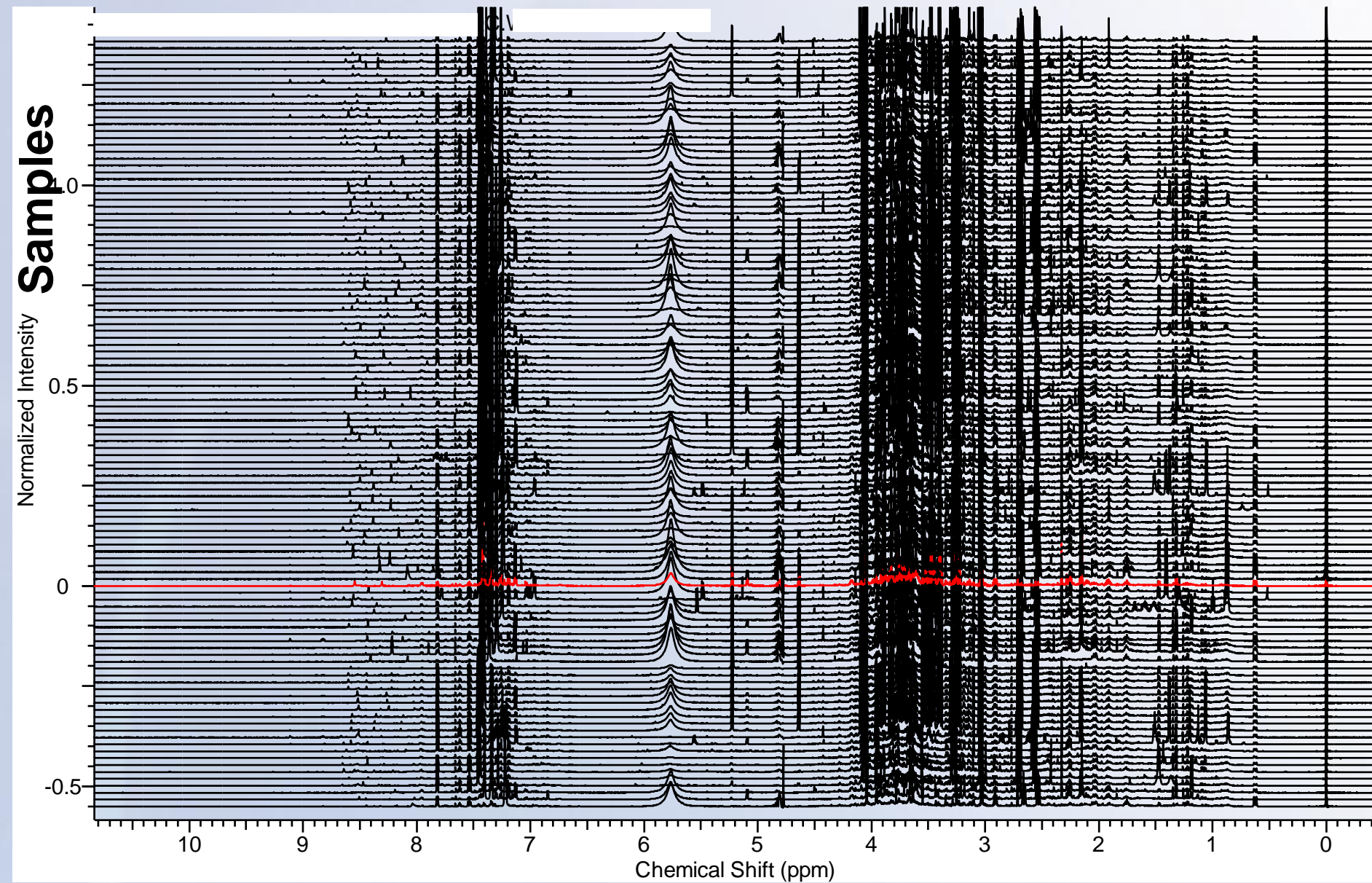
Citrate

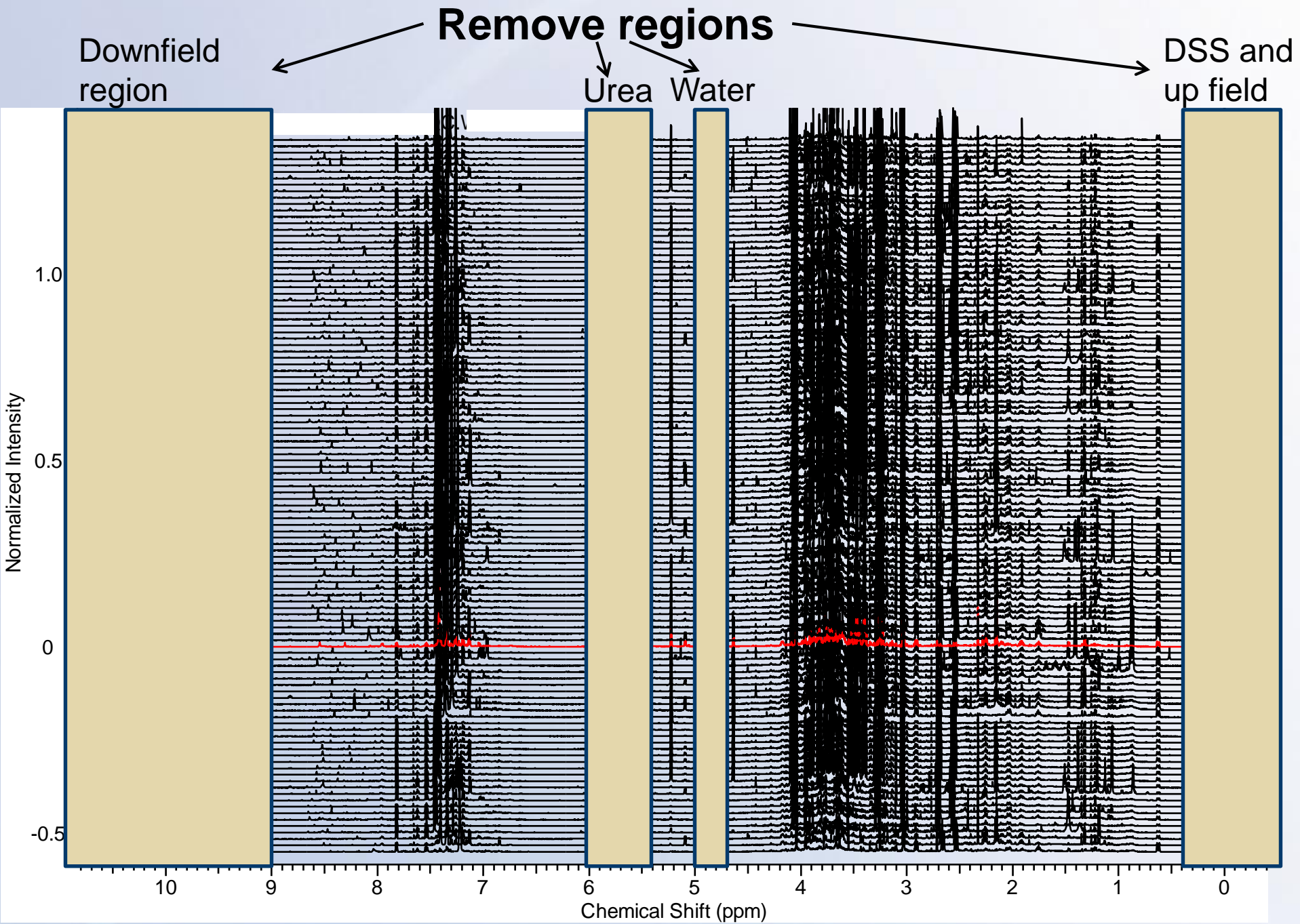


**Smart Binning**



# NMR Binning







# Binned Data

- 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	[0.40 .. 0.46]	[0.46 .. 0.52]	[0.52 .. 0.54]	[0.54 .. 0.57]	[0.57 .. 0.60]	[0.60 .. 0.66]	[0.66 .. 0.68]	[0.68 .. 0.71]	[0.71 .. 0.75]
C0559	Cases	7.60E-05	0.00E+00	7.32E-02	8.48E-02	3.20E-02	1.84E+00	1.31E-01	3.60E-01	3.67E-01
C0629	Cases	0.00E+00	1.78E-02	0.00E+00	2.18E-02	0.00E+00	1.08E+01	0.00E+00	0.00E+00	3.02E-02
C0640	Cases	3.44E-04	0.00E+00	1.83E-03	1.86E-04	0.00E+00	4.51E+00	0.00E+00	0.00E+00	0.00E+00
C0835	Cases	6.41E-04	0.00E+00	6.44E-03	0.00E+00	3.96E-03	3.28E+00	0.00E+00	5.12E-03	1.75E-02
D0613	Cases	6.63E-03	0.00E+00	0.00E+00	1.06E-02	0.00E+00	5.79E+00	0.00E+00	6.36E-02	3.02E-01
D0762	Cases	0.00E+00	0.00E+00	1.79E-02	1.98E-02	0.00E+00	9.37E+00	0.00E+00	0.00E+00	1.74E-02
D1113	Cases	3.14E-03	2.42E-03	8.02E-02	1.04E-01	5.32E-03	3.74E+00	0.00E+00	2.02E-02	1.84E-01
D1158	Cases	0.00E+00	3.71E-03	2.35E-02	4.83E-02	0.00E+00	5.02E+00	0.00E+00	1.91E-02	0.00E+00
D2090	Cases	0.00E+00	0.00E+00	2.45E-03	9.98E-04	0.00E+00	5.76E+00	0.00E+00	1.24E-02	1.04E-02
E0004	Cases	1.72E-03	0.00E+00	6.85E-02	3.05E-02	0.00E+00	1.47E+00	6.90E-02	3.61E-01	4.08E-01
E0195	Cases	0.00E+00	1.69E-03	5.57E-02	6.29E-02	0.00E+00	2.77E+00	1.34E-01	2.04E-01	4.56E-01
E0225	Cases	1.25E-03	0.00E+00	4.40E-03	1.69E-02	0.00E+00	9.17E+00	0.00E+00	1.08E-02	2.30E-02
E0309	Cases	4.11E-03	0.00E+00	2.23E-02	7.54E-03	3.08E-03	3.54E+00	0.00E+00	3.28E-02	9.09E-01
E0487	Cases	1.72E-03	0.00E+00	0.00E+00	1.00E-02	0.00E+00	4.00E+00	0.00E+00	1.36E-02	0.00E+00
F0036	Cases	1.66E-02	0.00E+00	0.00E+00	2.06E-02	0.00E+00	1.22E+01	1.04E-02	0.00E+00	5.97E-01
F0108	Cases	0.00E+00	2.31E-03	6.30E-03	1.11E-02	0.00E+00	7.17E+00	0.00E+00	1.65E-02	2.21E-01
A0233	Control	0.00E+00	1.86E-02	0.00E+00	1.82E-02	0.00E+00	1.61E+01	0.00E+00	2.91E-03	0.00E+00
A0490	Control	0.00E+00	0.00E+00	2.99E-03	3.60E-02	0.00E+00	2.97E+00	0.00E+00	4.00E-02	5.46E-01
A2003	Control	0.00E+00	0.00E+00	3.45E-02	2.20E-02	0.00E+00	1.80E+00	0.00E+00	0.00E+00	0.00E+00
C0586	Control	0.00E+00	1.69E-02	0.00E+00	6.64E-03	0.00E+00	1.92E+01	0.00E+00	6.51E-02	0.00E+00
C2177	Control	0.00E+00	0.00E+00	3.02E-02	3.59E-02	0.00E+00	2.35E+00	0.00E+00	3.19E-02	1.49E-01
D0177	Control	9.21E-03	0.00E+00	1.69E-02	1.47E-02	0.00E+00	2.43E+00	0.00E+00	4.46E-02	0.00E+00
D0729	Control	0.00E+00	1.88E-03	5.58E-02	7.87E-02	2.92E-02	3.16E+00	6.59E-02	2.80E-01	4.30E-01
D0909	Control	0.00E+00	1.08E-03	0.00E+00	5.69E-03	0.00E+00	2.49E+00	0.00E+00	1.01E-02	1.87E-01
D0945	Control	0.00E+00	4.79E-04	7.00E-03	0.00E+00	4.19E-03	3.99E+00	0.00E+00	1.11E-03	3.96E-02
D1174	Control	0.00E+00	9.33E-04	0.00E+00	3.43E-03	1.30E-02	7.21E+00	6.53E-03	0.00E+00	1.66E-02
D2054	Control	1.55E-03	0.00E+00	0.00E+00	1.22E-02	0.00E+00	2.07E+00	0.00E+00	1.28E-02	3.90E-01
D2062	Control	2.39E-05	0.00E+00	6.04E-02	2.99E-02	0.00E+00	4.94E+00	0.00E+00	9.95E-03	0.00E+00
D2079	Control	2.73E-02	0.00E+00	1.81E-03	1.17E-02	0.00E+00	3.38E+01	7.87E-02	0.00E+00	5.91E+00

Metadata

Normalized binned data

# Data Normalization, Transformation and Scaling

# Data 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

I Centering

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$$

II Autoscaling

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$$

Range scaling

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{(x_{i_{\max}} - x_{i_{\min}})}$$

Pareto scaling

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$$

Vast scaling

$$\tilde{x}_{ij} = \frac{(x_{ij} - \bar{x}_i)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$$

Level scaling

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$$

III Log transformation

$$\tilde{x}_{ij} = {}^{10}\log(x_{ij})$$

$$\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$$

Power transformation

$$\tilde{x}_{ij} = \sqrt{(x_{ij})}$$

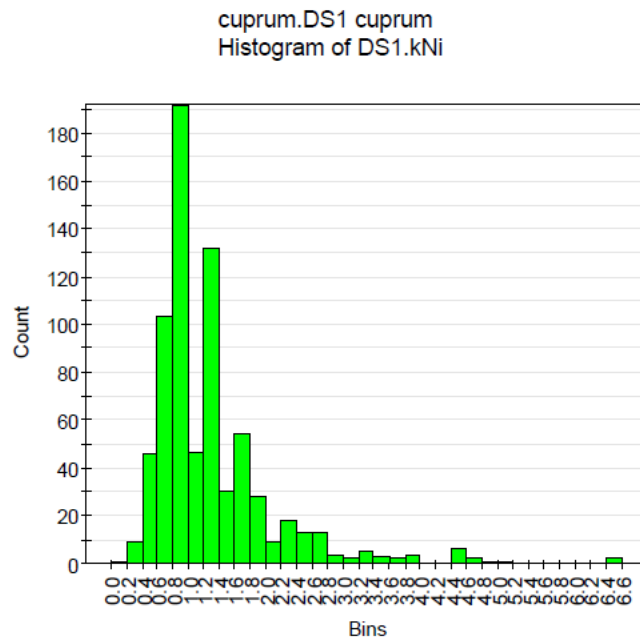
$$\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$$

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

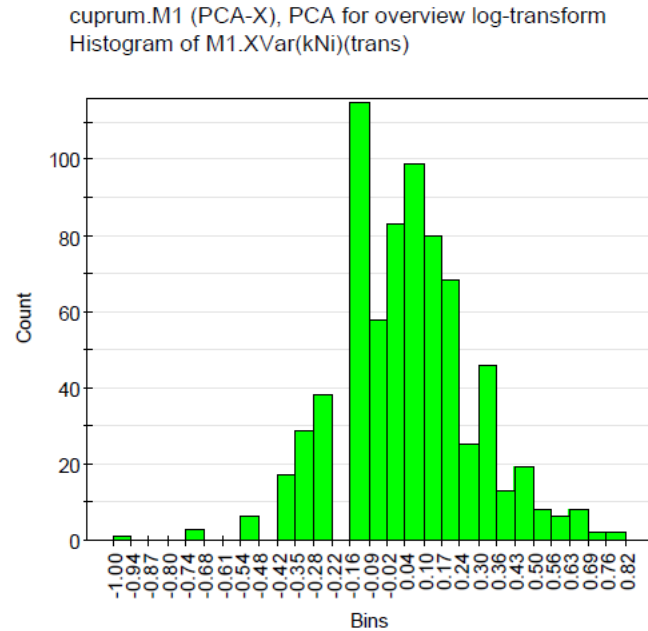
Van den Berg et al 1006, BMC Genomics, 7, 142

# Data Transformation

- Before transformation
  - skew distribution



- After log-transformation
  - More close to normal distribution

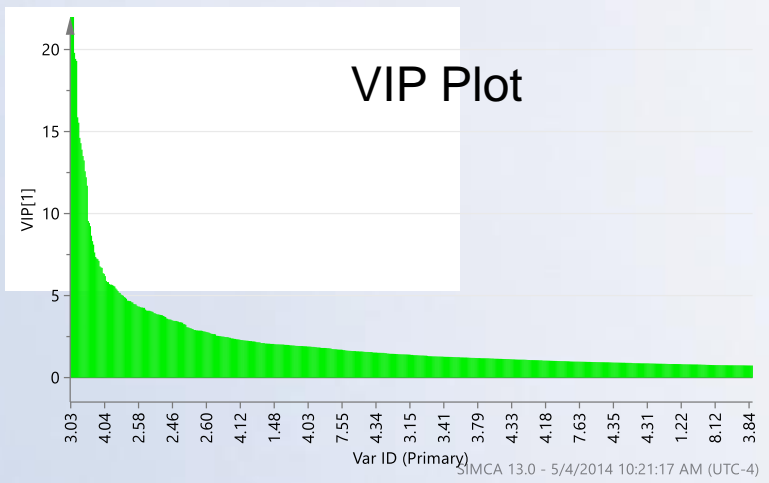
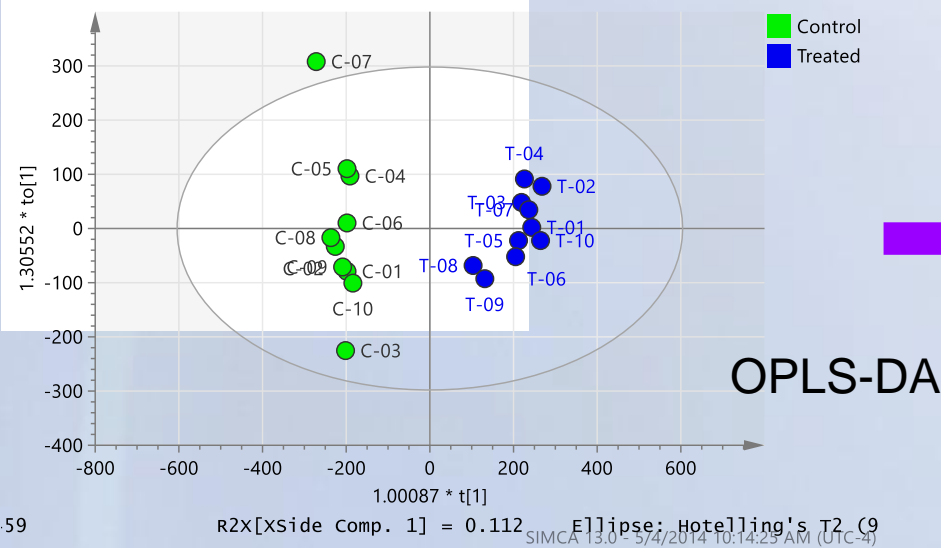
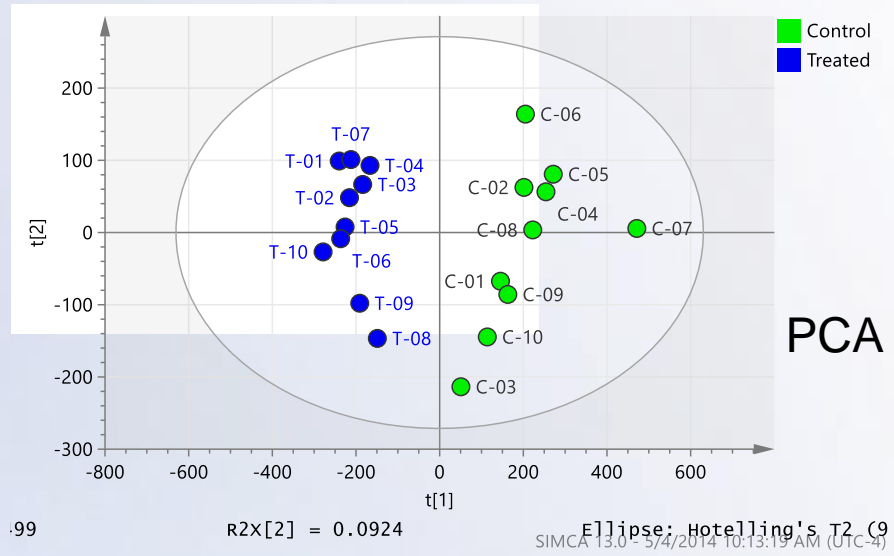


# Scaling

- 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



# ERCMRC at UNC Chapel Hill



Yuanyuan Li  
LC-MS/MS  
LC-TOF-MS



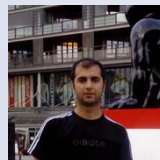
Wimal Pathmasiri  
NMR & GC-TOF-MS



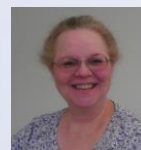
Delisha Stewart  
NMR and LC-TOF-MS



Maria Moreno  
NMR and LC-MS/MS



Reza Ghanbari  
Postdoctoral Fellow



Rose Ewald  
Graduate Studies



Susan Sumner  
PI, ERCMRC



Susan McRitchie  
Program Coordinator  
Data Analysis

## NCRC



Nick Gillitt  
Dole  
700 MHz NMR  
6500 Sciex LC-MS



Colin Kaye  
NCSU  
6500 Sciex  
Triple Quad



UNC-G  
Q-Exactive



Debby Reed  
GC-MS  
GC-TOF-MS



Stephen Orena  
LC-MS/MS



Martin Kohlmeier  
Training



Tim Fennell  
Director,  
Analytical Chemistry &  
Pharmaceutics



Yan Lan Yueh  
LC-MS

## RTI



Jessica Gooding  
LC-MS



Rod Snyder  
NMR and LC-MS

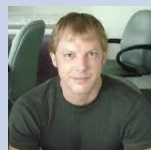


Courtney Whitaker  
LC-MS

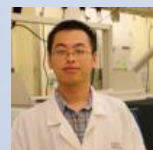
Scott Watson  
Neurotransmitter  
LC/MS

## DHMRI

Jason Winnike  
NMR  
2D-GC-TOF-MS

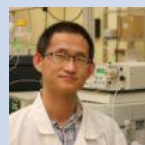


David Kirchner  
LC-MS/MS

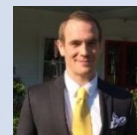


Huadong Chen  
LC-MS  
LC-TOF-MS

Huiyuan Chen  
GC-MS  
GC-TOF-MS



## UNC Charlotte Bioinformatics



Owen Myers



XiuXia Du

Aleksandr  
Smirnov

