NMR Data Pre-processing

UAB Metabolomics Training Course

June 14-18, 2015

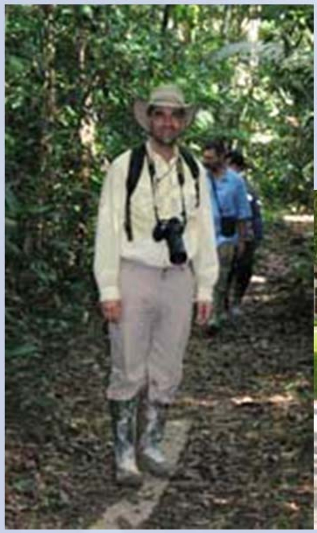
Wimal Pathmasiri, Rodney Snyder

NIH Eastern Regional Comprehensive Metabolomics Resource Core

(RTI RCMRC)

NIH Common Fund Metabolomics Cores

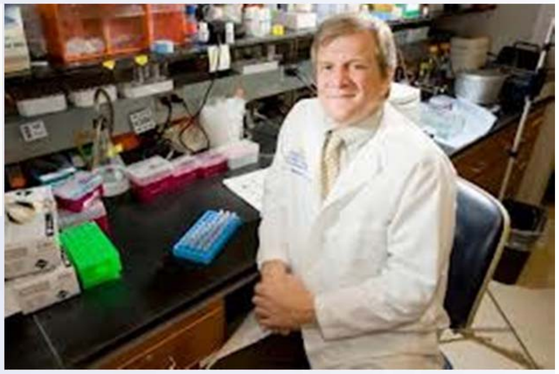
NIH Metabolomics Centers Ramp Up | November 4, 2013 Issue - Vol. 91 Issue 44 | Chemical & Engineering News. by Jyllian Kemsley



UF UNIVERSITY of FLORIDA

mayo

THE UNIVERSITY OF MICHIGAN ARTES SCIENTIA VERITAS 1817

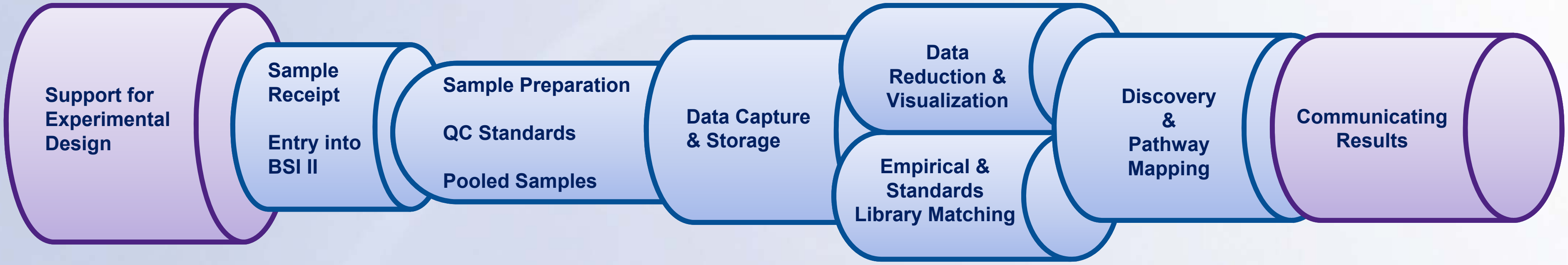


UK

UCDAVIS UNIVERSITY OF CALIFORNIA

RTI INTERNATIONAL

NIH Eastern Regional Comprehensive Metabolomics Resource Core at RTI


Support for Experimental Design

Sample Receipt

Entry into BSI II

Sample Preparation

QC Standards

Pooled Samples

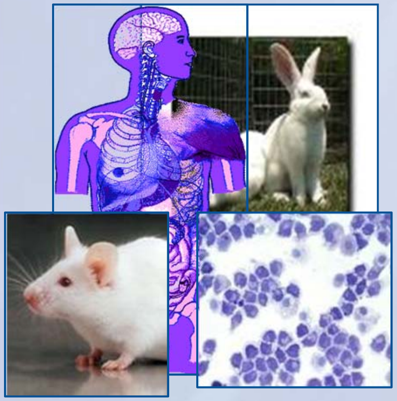
Data Capture & Storage

Data Reduction & Visualization

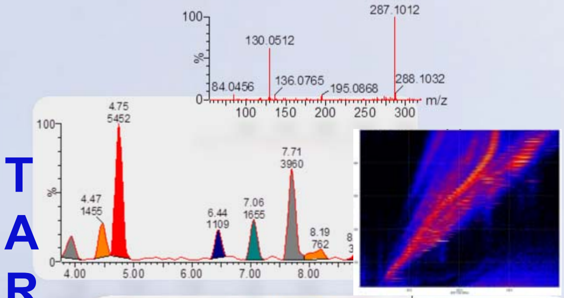
Empirical & Standards Library Matching

Discovery & Pathway Mapping

Communicating Results

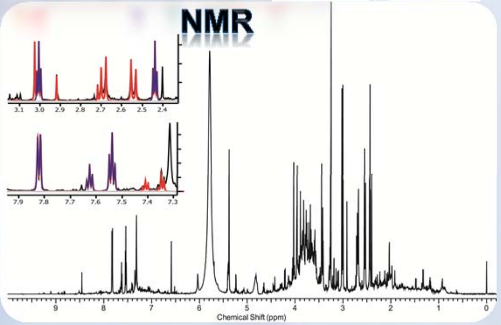


LC-MS, MS/MS and ION-MOBILITY



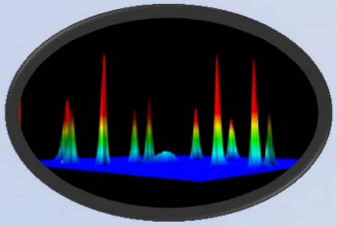
TARGETED

NMR

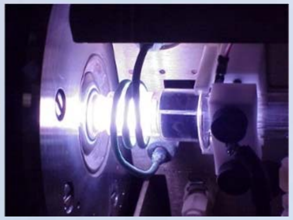


BROAD SPECTRUM

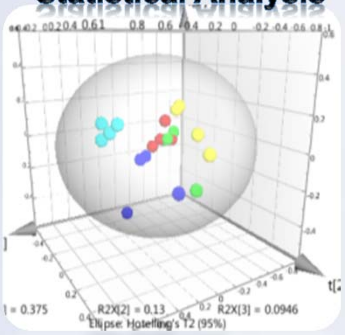
GC-MS



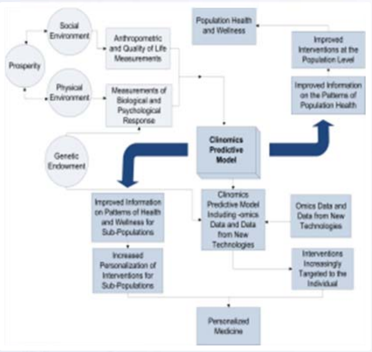
ICP-MS



Multivariate and Statistical Analysis



Predictive Modeling


Social Environment

Prosperity

Physical Environment

Genetic Endowment

Anthropometric and Quality of Life Measurements

Measurements of Biological and Psychological Response

Population Health and Wellness

Improved Interventions at the Population Level

Improved Information on the Patterns of Population Health

Clinomics Predictive Model

Improved Information on Patterns of Health and Wellness for Sub-Populations

Clinomics Predictive Model Including -omics Data and Data from New Technologies

Omics Data and Data from New Technologies

Increased Personalization of Interventions for Sub-Populations

Interventions Increasingly Targeted to the Individual

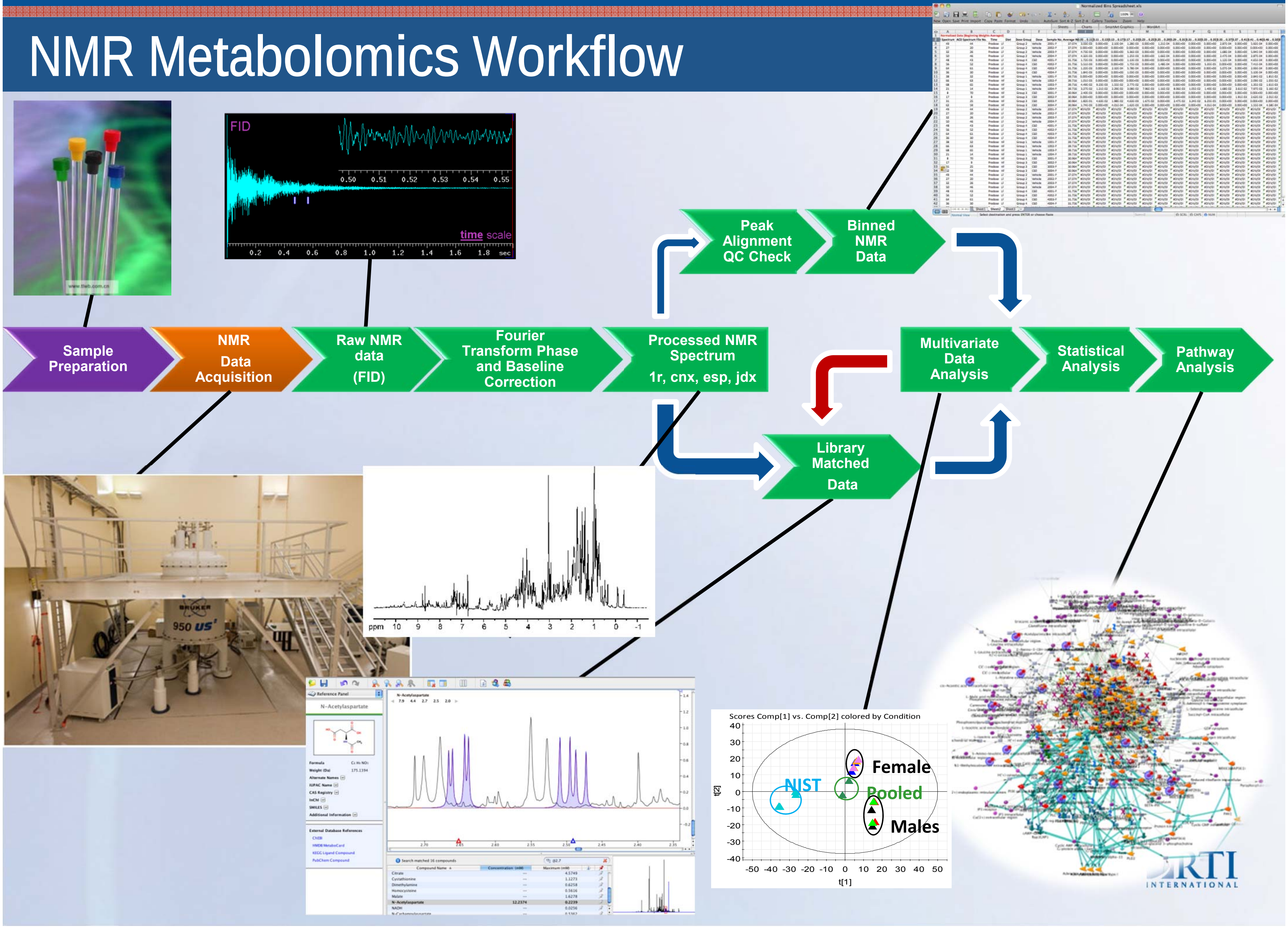
Personalized Medicine

Pathway Mapping Metabolites-Proteins-Genes

DRUG DISCOVERY • ENVIRONMENT • UROLOGY • IMMUNOLOGY • OBESITY • WOMEN'S HEALTH •
RTI INTERNATIONAL

RTI International

NMR Metabolomics Workflow



Sample Preparation

NMR Data Acquisition

Raw NMR data (FID)

Fourier Transform Phase and Baseline Correction

Processed NMR Spectrum 1r, cnx, esp, jdx

Peak Alignment QC Check

Binned NMR Data

Library Matched Data

Multivariate Data Analysis

Statistical Analysis

Pathway Analysis

Reference Panel

N-Acetylaspartate

Formula $C_6H_9NO_5$

Weight (Da) 175.1394

Alternate Names

IUPAC Name

CAS Registry

InChI

SMILES

Additional Information

External Database References

ChEBI

HMDB MetaboCard

KEGG Ligand Compound

PubChem Compound

Search matched 16 compounds

Compound Name	Concentration (mM)	Maximum (mM)
Citrate	---	4.5749
Cystathionine	---	1.1273
Dimethylamine	---	0.6258
Homocysteine	---	0.5616
Malate	---	1.6278
N-Acetylaspartate	12.2374	0.2239
NADH	---	0.0256
N-Carbamoylaspartate	---	0.5362

Scores Comp[1] vs. Comp[2] colored by Condition

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

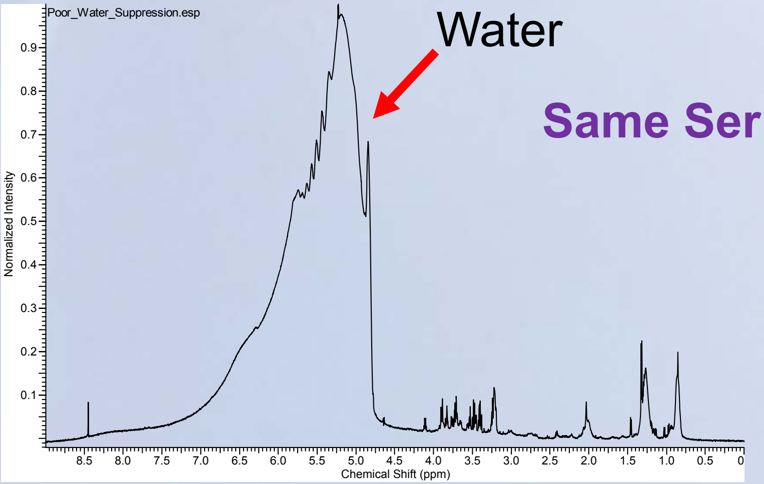
- Quality of metabolomics analysis depends on data quality

- 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)

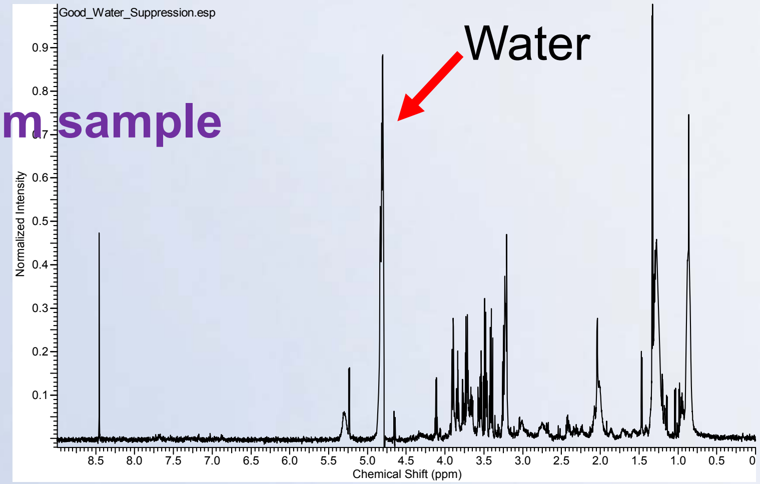
- High quality of data is needed for best results

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



Same Serum sample

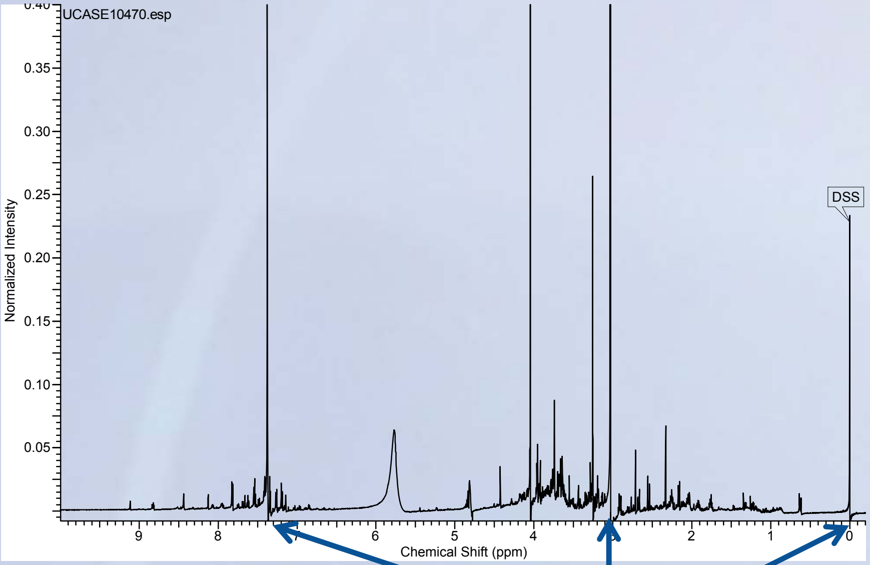


Poor water suppression

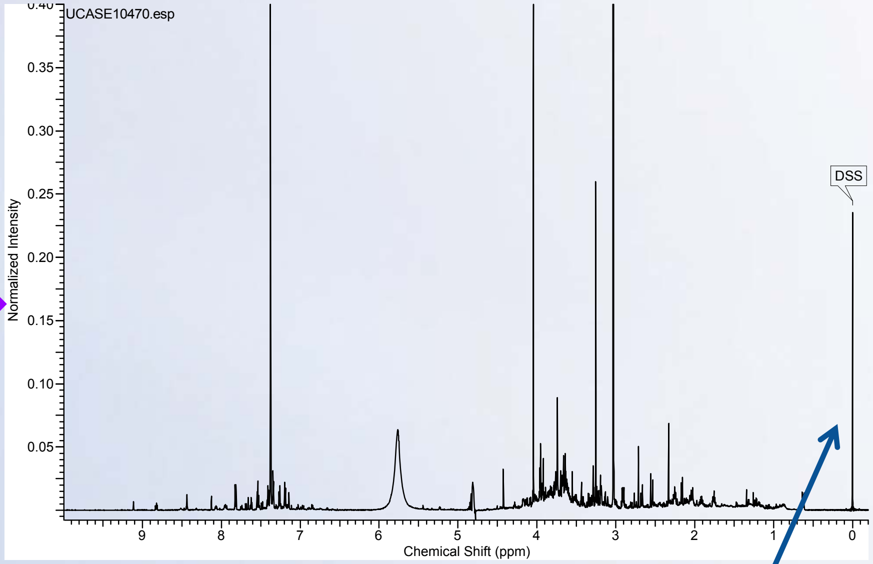
Good water suppression

NMR Pre-processing

Before



After



Phase
Baseline to be corrected

Reference,
Line shape

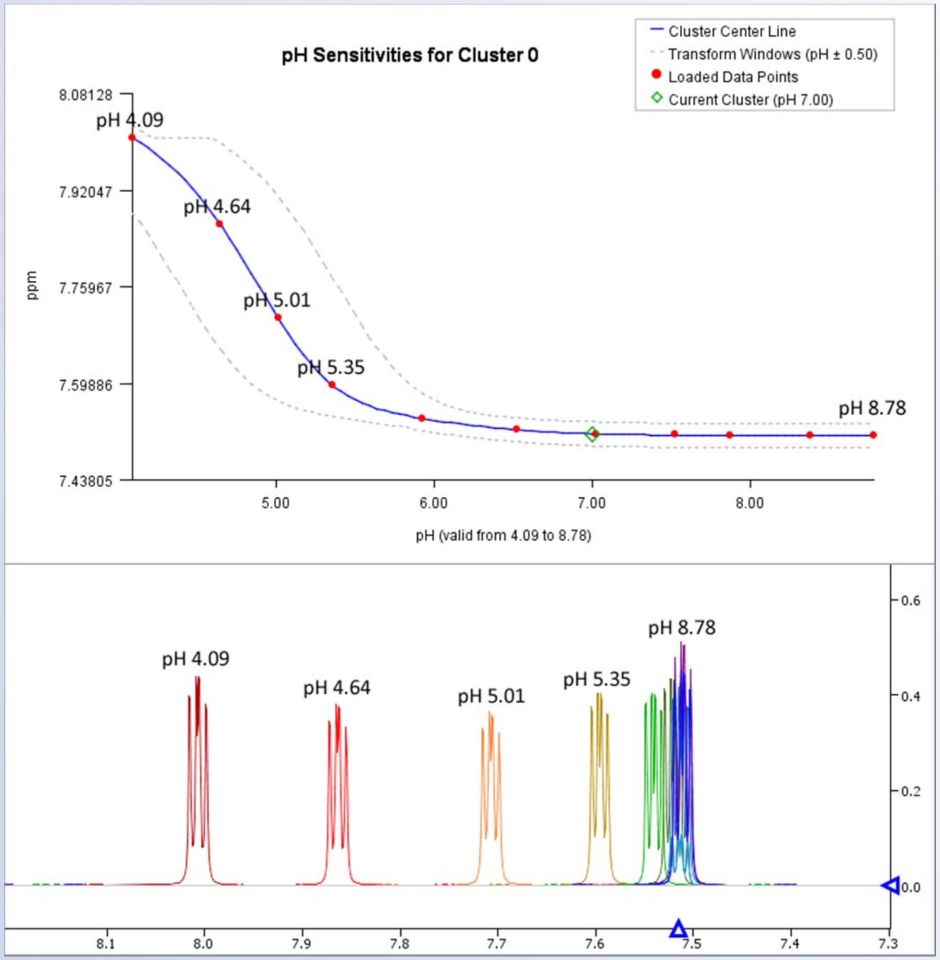
pH Dependence of Chemical Shift

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



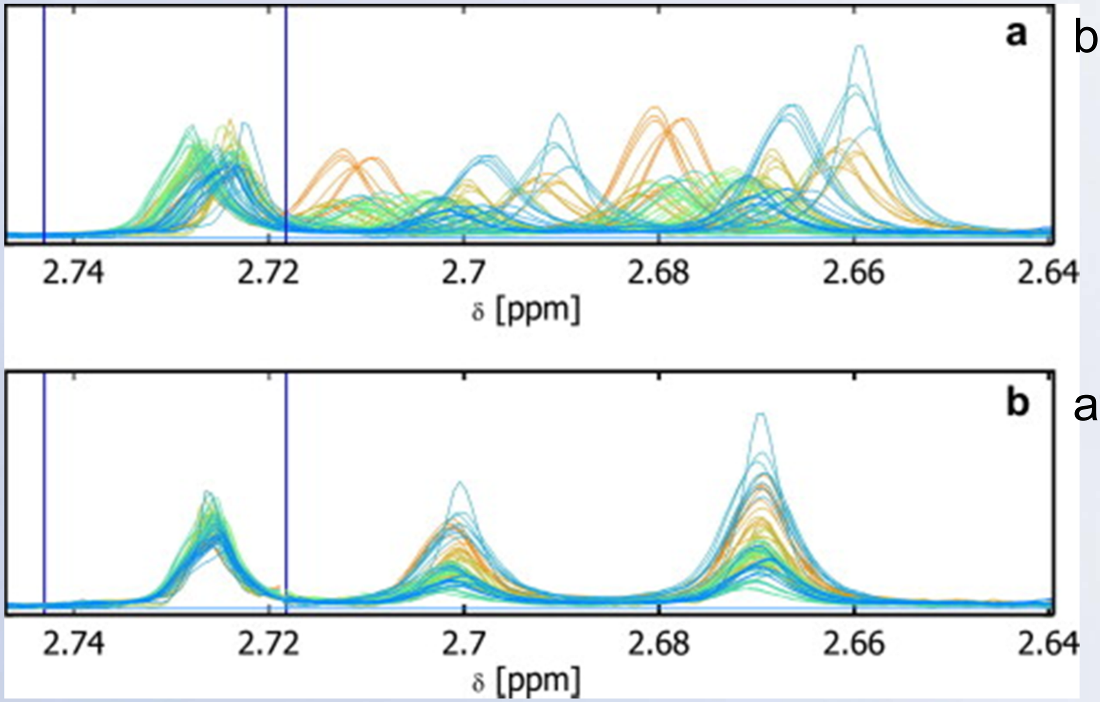
http://www.chenomx.com/software/software.php
Savorani , F. et al, Journal of Magnetic Resonance, Volume 202, Issue 2, 2010, 190 – 202
Vu, T. N. et al., *BMC Bioinformatics* 2011, **12**:405

Peak Alignment

Example

icoshift

One of the Citrate peaks



before

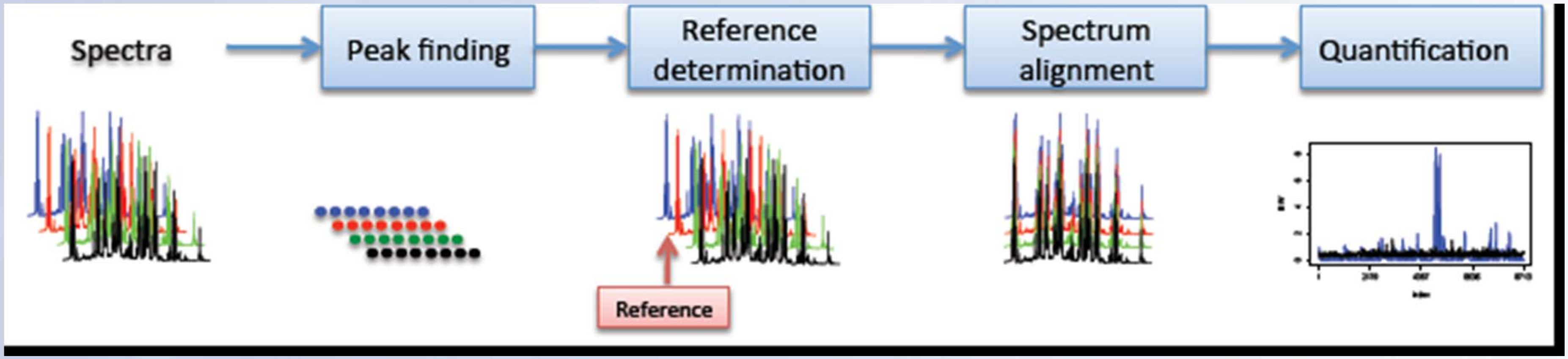
after

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

Peak Alignment

Example

speaq



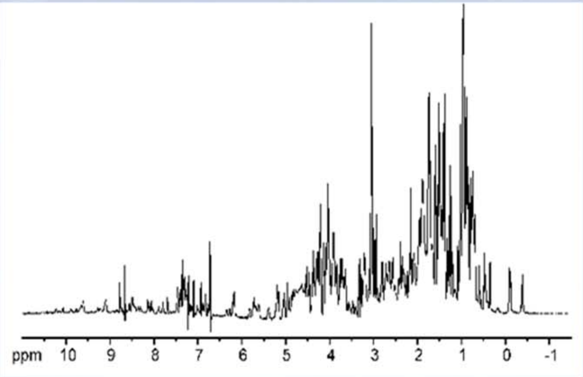
Vu, T. N. et al., *BMC Bioinformatics* 2011, **12**:405

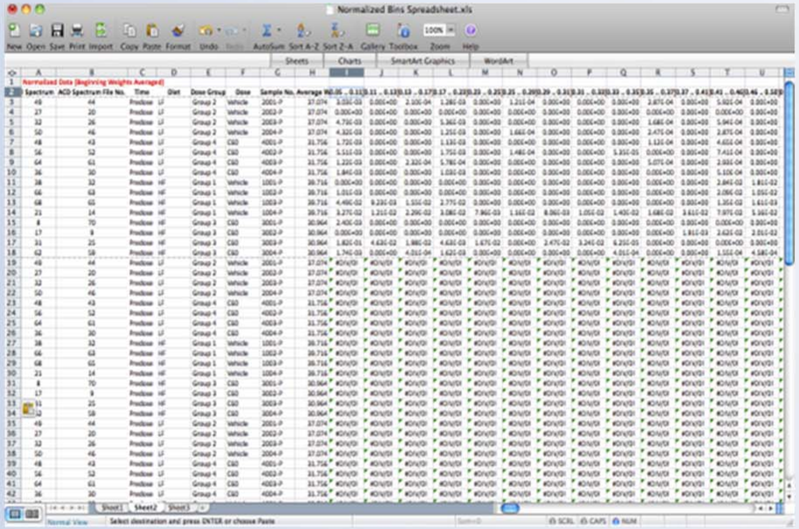
NMR Binning

- 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





Normalized Data (Beginning Weights Averaged)

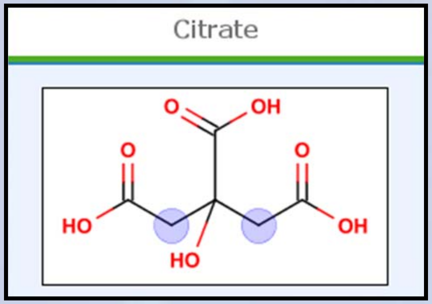
Spectrum	ACD Spectrum File No.	Time	Diet	Dose Group	Dose	Sample No.	Average W	0.05 .. 0.11	0.11 .. 0.13	0.13 .. 0.17
49	44	Predose	LF	Group 2	Vehicle	2001-P	37.074	3.03E-03	0.00E+00	2.10E-04
27	20	Predose	LF	Group 2	Vehicle	2002-P	37.074	0.00E+00	0.00E+00	0.00E+00
32	26	Predose	LF	Group 2	Vehicle	2003-P	37.074	4.73E-03	0.00E+00	0.00E+00
50	46	Predose	LF	Group 2	Vehicle	2004-P	37.074	4.32E-03	0.00E+00	0.00E+00
48	43	Predose	LF	Group 4	C60	4001-P	31.756	1.72E-03	0.00E+00	0.00E+00
56	52	Predose	LF	Group 4	C60	4002-P	31.756	5.51E-03	0.00E+00	0.00E+00
64	61	Predose	LF	Group 4	C60	4003-P	31.756	1.22E-03	0.00E+00	2.32E-04
36	30	Predose	LF	Group 4	C60	4004-P	31.756	1.84E-03	0.00E+00	0.00E+00
38	32	Predose	HF	Group 1	Vehicle	1001-P	38.716	0.00E+00	0.00E+00	0.00E+00
66	63	Predose	HF	Group 1	Vehicle	1002-P	38.716	1.01E-03	0.00E+00	0.00E+00
68	65	Predose	HF	Group 1	Vehicle	1003-P	38.716	4.49E-02	9.23E-03	1.55E-02
21	14	Predose	HF	Group 1	Vehicle	1004-P	38.716	3.27E-02	1.21E-02	2.29E-02
8	70	Predose	HF	Group 3	C60	3001-P	30.964	2.40E-03	0.00E+00	0.00E+00
17	9	Predose	HF	Group 3	C60	3002-P	30.964	0.00E+00	0.00E+00	0.00E+00
31	25	Predose	HF	Group 3	C60	3003-P	30.964	1.82E-01	4.63E-02	1.88E-02
62	58	Predose	HF	Group 3	C60	3004-P	30.964	1.74E-03	0.00E+00	4.01E-04
49	44	Predose	LF	Group 2	Vehicle	2001-P	37.074	#DIV/0!	#DIV/0!	#DIV/0!
27	20	Predose	LF	Group 2	Vehicle	2002-P	37.074	#DIV/0!	#DIV/0!	#DIV/0!
32	26	Predose	LF	Group 2	Vehicle	2003-P	37.074	#DIV/0!	#DIV/0!	#DIV/0!
50	46	Predose	LF	Group 2	Vehicle	2004-P	37.074	#DIV/0!	#DIV/0!	#DIV/0!
48	43	Predose	LF	Group 4	C60	4001-P	31.756	#DIV/0!	#DIV/0!	#DIV/0!
56	52	Predose	LF	Group 4	C60	4002-P	31.756	#DIV/0!	#DIV/0!	#DIV/0!
64	61	Predose	LF	Group 4	C60	4003-P	31.756	#DIV/0!	#DIV/0!	#DIV/0!
36	30	Predose	LF	Group 4	C60	4004-P	31.756	#DIV/0!	#DIV/0!	#DIV/0!
38	32	Predose	HF	Group 1	Vehicle	1001-P	38.716	#DIV/0!	#DIV/0!	#DIV/0!
66	63	Predose	HF	Group 1	Vehicle	1002-P	38.716	#DIV/0!	#DIV/0!	#DIV/0!
68	65	Predose	HF	Group 1	Vehicle	1003-P	38.716	#DIV/0!	#DIV/0!	#DIV/0!
21	14	Predose	HF	Group 1	Vehicle	1004-P	38.716	#DIV/0!	#DIV/0!	#DIV/0!
8	70	Predose	HF	Group 3	C60	3001-P	30.964	#DIV/0!	#DIV/0!	#DIV/0!
17	9	Predose	HF	Group 3	C60	3002-P	30.964	#DIV/0!	#DIV/0!	#DIV/0!
31	25	Predose	HF	Group 3	C60	3003-P	30.964	#DIV/0!	#DIV/0!	#DIV/0!
62	58	Predose	HF	Group 3	C60	3004-P	30.964	#DIV/0!	#DIV/0!	#DIV/0!
49	44	Predose	LF	Group 2	Vehicle	2001-P	37.074	#DIV/0!	#DIV/0!	#DIV/0!
27	20	Predose	LF	Group 2	Vehicle	2002-P	37.074	#DIV/0!	#DIV/0!	#DIV/0!
32	26	Predose	LF	Group 2	Vehicle	2003-P	37.074	#DIV/0!	#DIV/0!	#DIV/0!
50	46	Predose	LF	Group 2	Vehicle	2004-P	37.074	#DIV/0!	#DIV/0!	#DIV/0!
48	43	Predose	LF	Group 4	C60	4001-P	31.756	#DIV/0!	#DIV/0!	#DIV/0!
56	52	Predose	LF	Group 4	C60	4002-P	31.756	#DIV/0!	#DIV/0!	#DIV/0!
64	61	Predose	LF	Group 4	C60	4003-P	31.756	#DIV/0!	#DIV/0!	#DIV/0!
36	30	Predose	LF	Group 4	C60	4004-P	31.756	#DIV/0!	#DIV/0!	#DIV/0!

Select destination and press ENTER or choose Paste

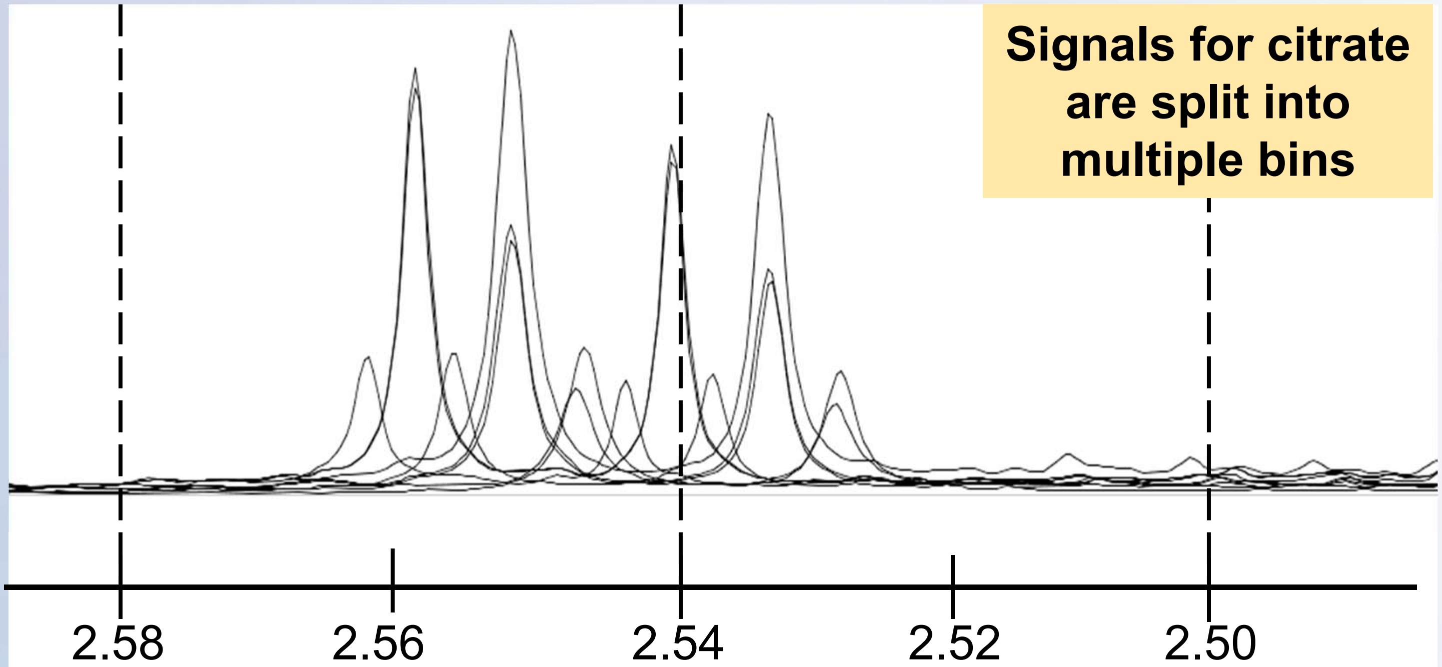
NMR Binning

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



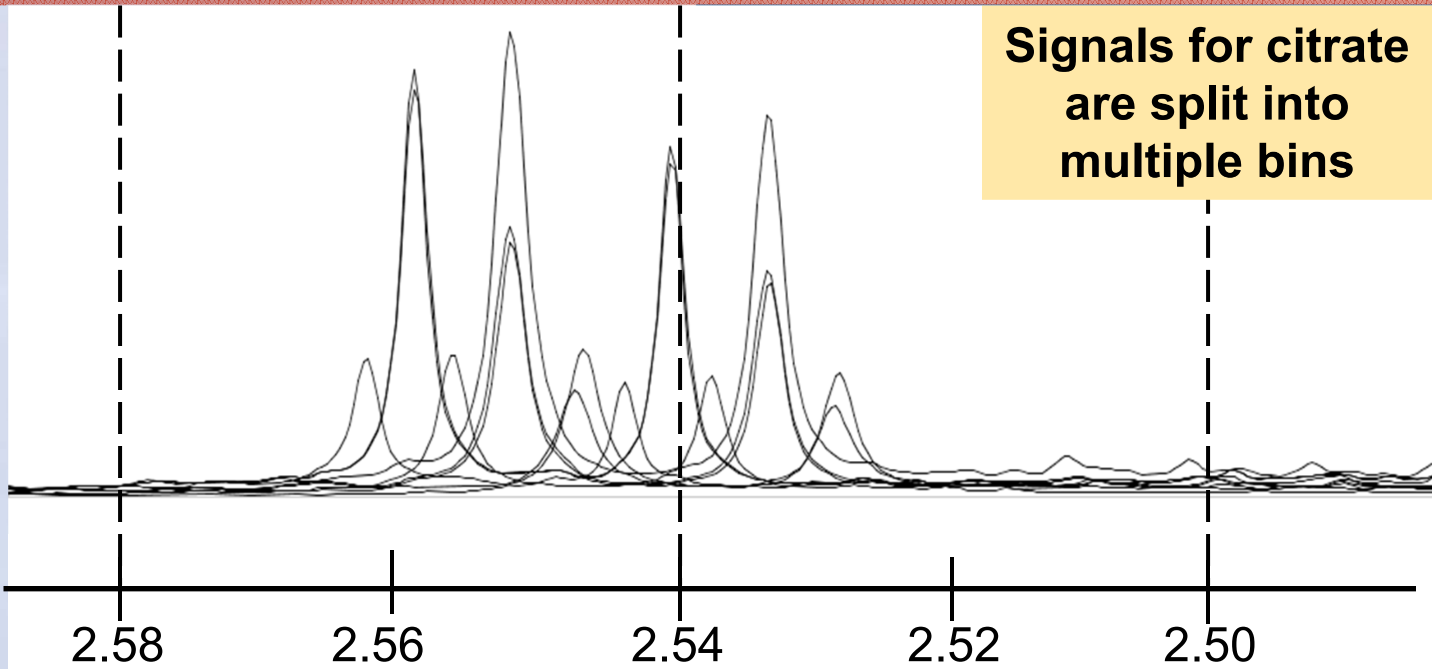
Fixed Binning



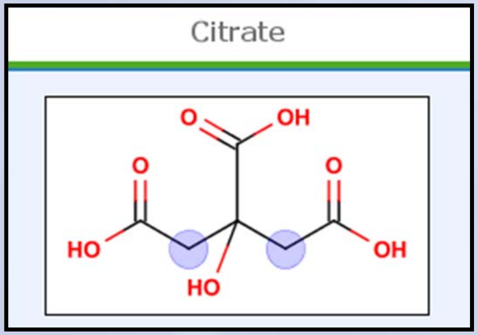
Signals for citrate are split into multiple bins

NMR Binning

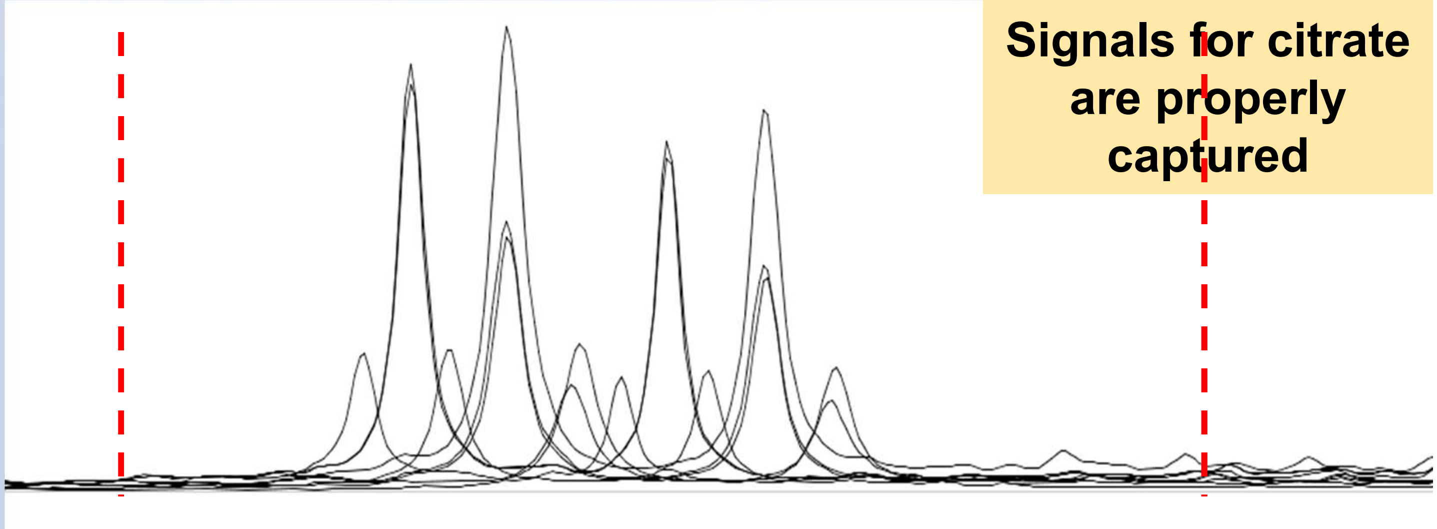
Fixed Binning



Signals for citrate are split into multiple bins

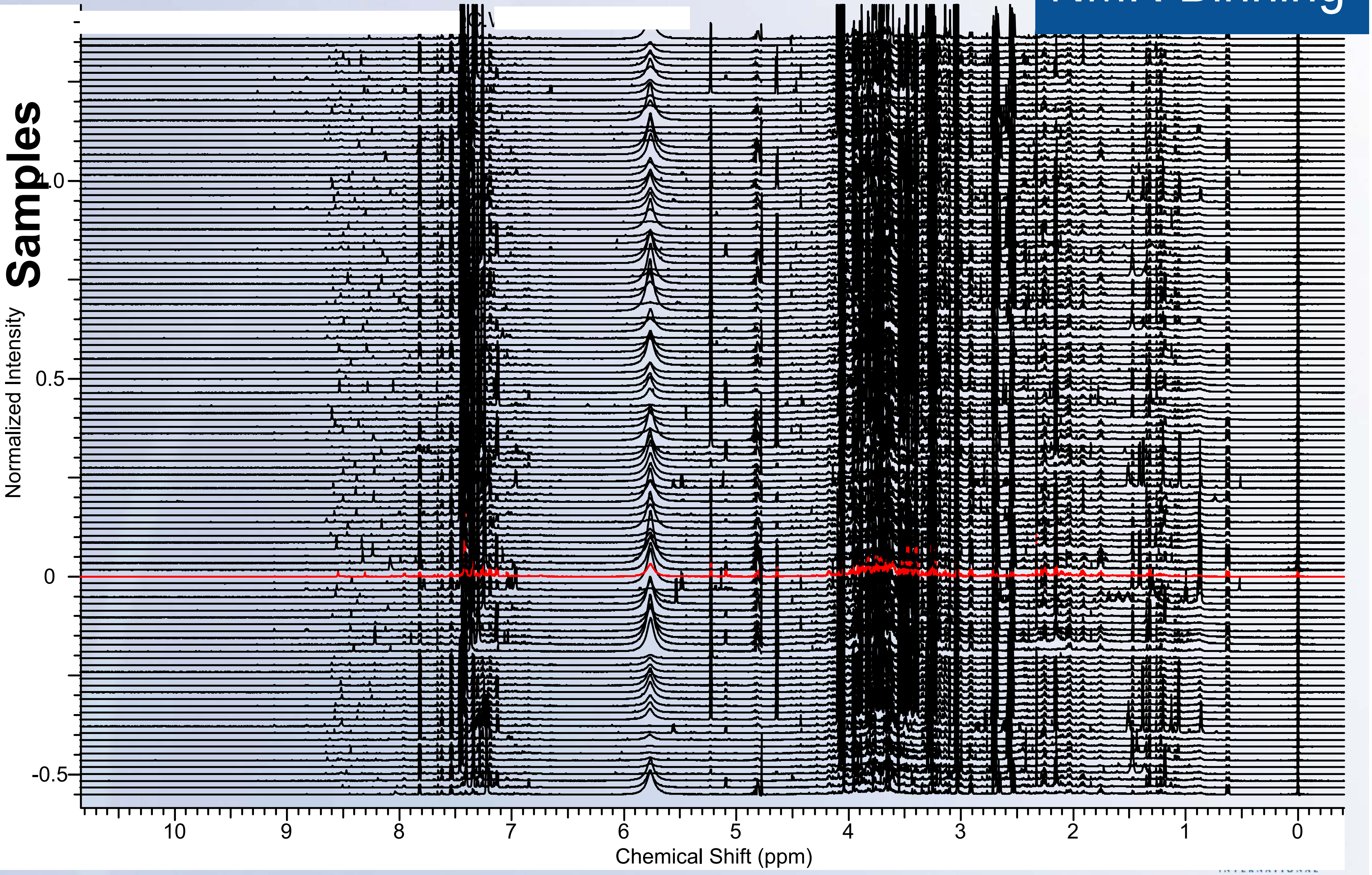

Citrate

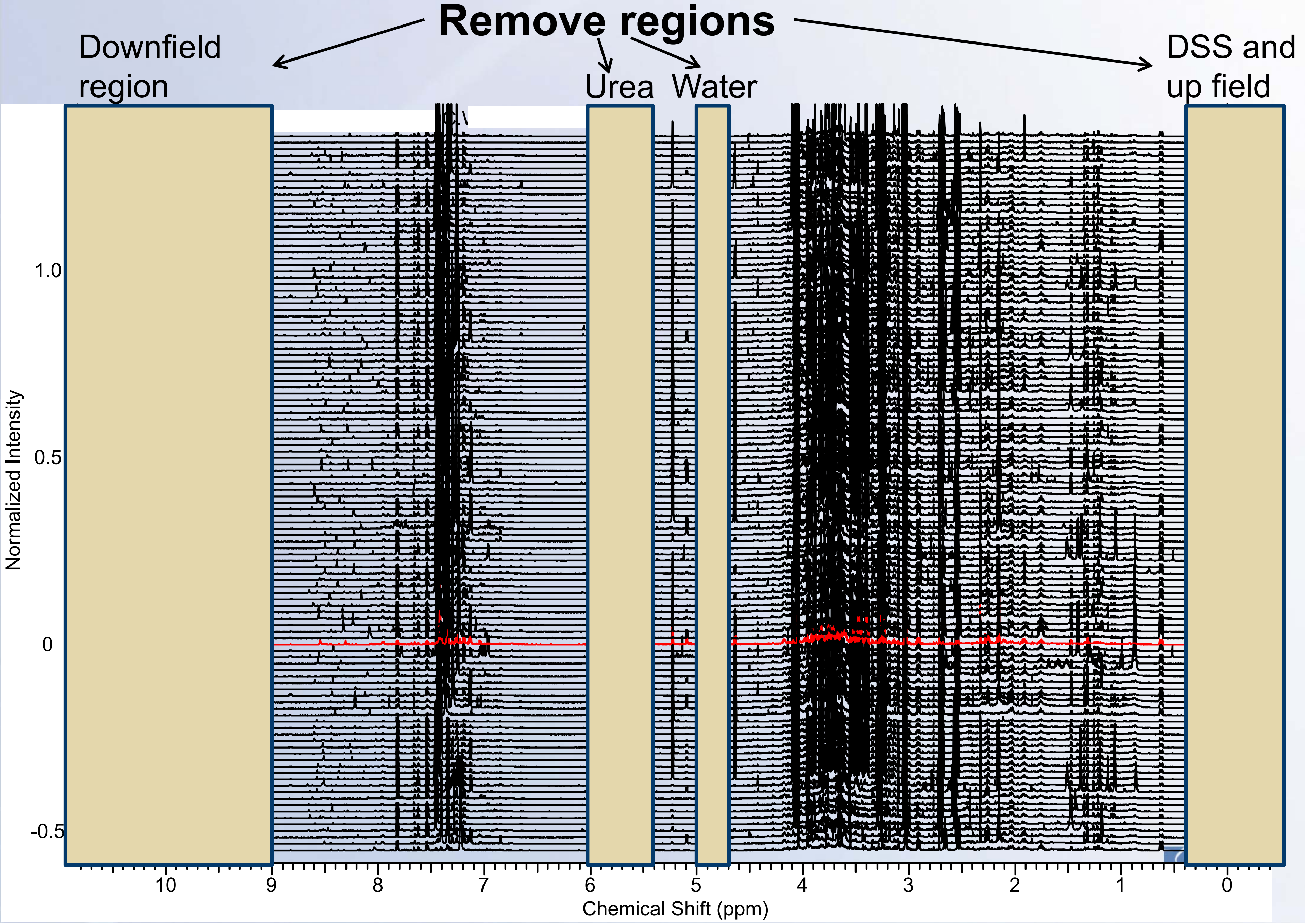
Smart Binning



Signals for citrate are properly captured

NMR Binning


Samples
Normalized Intensity
1.0
0.5
0
-0.5
10
9
8
7
6
5
4
3
2
1
0
Chemical Shift (ppm)


Remove regions
Downfield region
Urea
Water
DSS and up field
Normalized Intensity
1.0
0.5
0
-0.5
10
9
8
7
6
5
4
3
2
1
0
Chemical Shift (ppm)

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	[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

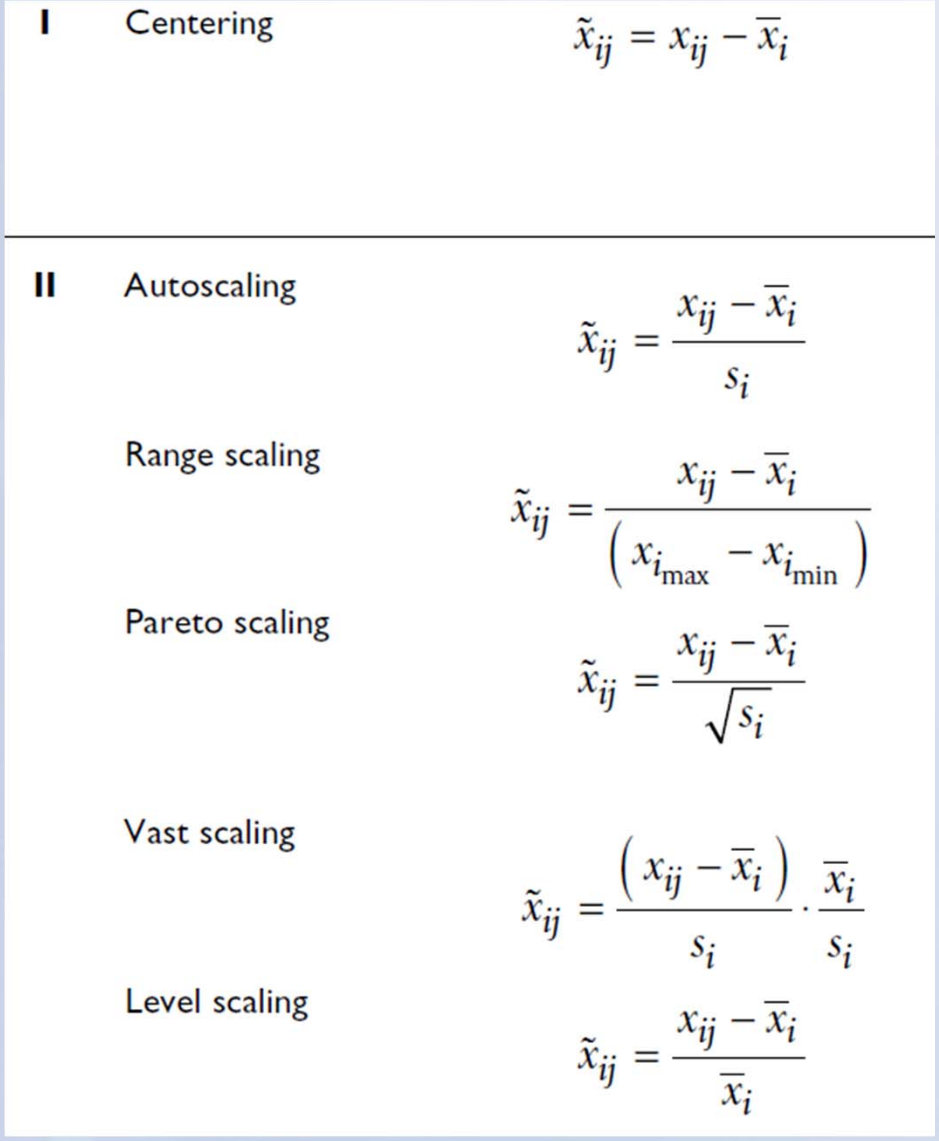
Normalized binned data

Data Normalization, Transformation, and Scaling

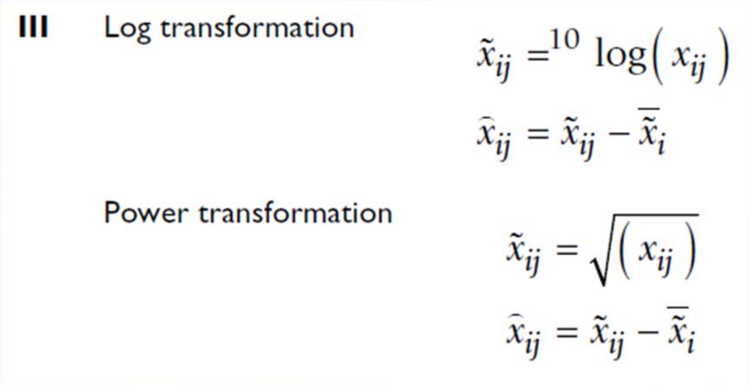
Data Normalization

- 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	$\tilde{x}_{ij} = \dfrac{x_{ij} - \bar{x}_i}{s_i}$
	Range scaling	$\tilde{x}_{ij} = \dfrac{x_{ij} - \bar{x}_i}{\left(x_{i_{\max}} - x_{i_{\min}} \right)}$
	Pareto scaling	$\tilde{x}_{ij} = \dfrac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$
	Vast scaling	$\tilde{x}_{ij} = \dfrac{\left(x_{ij} - \bar{x}_i \right)}{s_i} \cdot \dfrac{\bar{x}_i}{s_i}$
	Level scaling	$\tilde{x}_{ij} = \dfrac{x_{ij} - \bar{x}_i}{\bar{x}_i}$

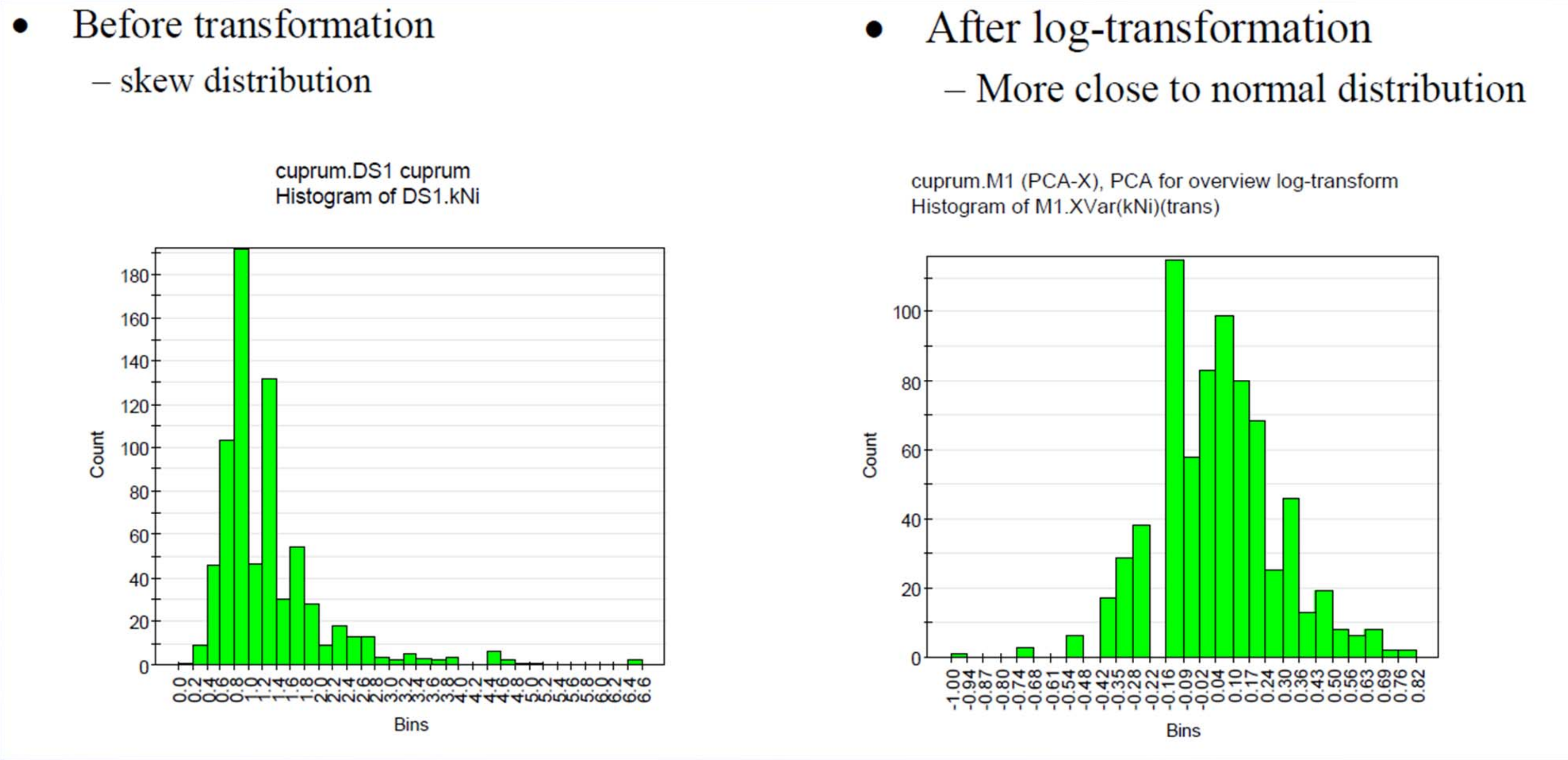


III	Log transformation	$\tilde{x}_{ij} =^{10} \log\left(x_{ij} \right)$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$
	Power transformation	$\tilde{x}_{ij} = \sqrt{\left(x_{ij} \right)}$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$

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

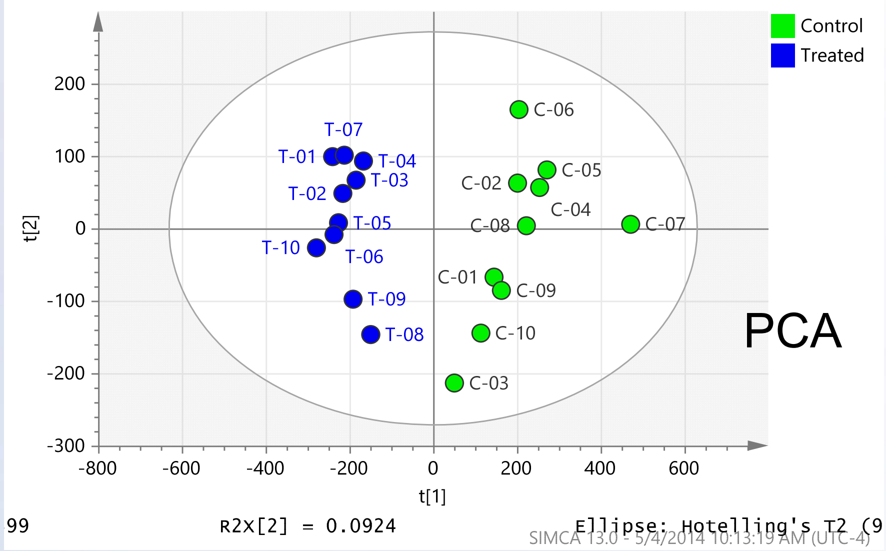
Susan Wicklund, Multivariate data analysis for omics, Sept 2-3 2008, Umetrics training

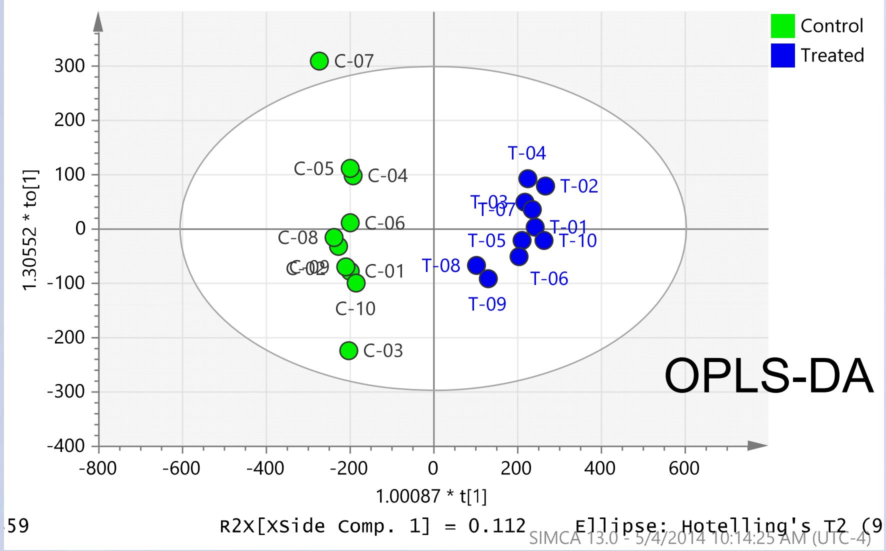
Scaling

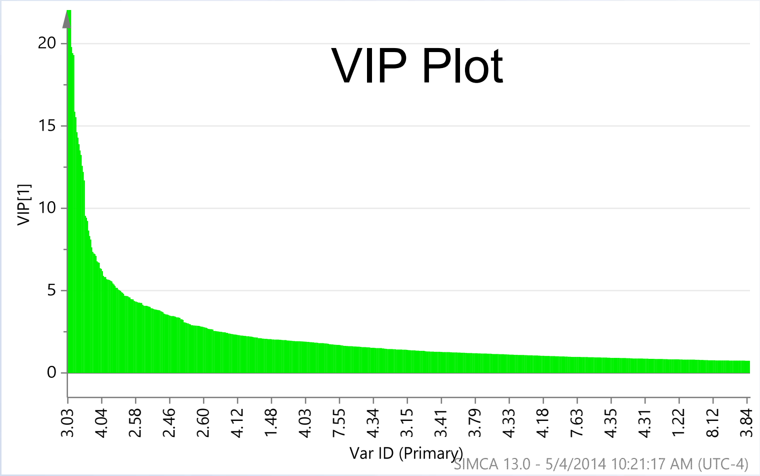
- Unit variance (autoscaling) divides the bin intensity by the standard deviation
  - May increase your baseline noise
  - Dimensionless value after scaling
- Pareto scaling divides the bin intensity by the square root of the standard deviation
  - Not dimensionless after scaling
- For NMR data, centering 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







ACKNOWLEDGEMENTS

Director RTI RCMRC
Susan Sumner, PhD

Program Coordinator
Jason Burgess, PhD

NIH Scientific Officer
David Balshaw, PhD, NIH/NIEHS

Internship Program
Stella Lam, BS

Feasibility Studies
Susan Sumner, PhD
Susan McRitchie, MS
Executive Committee

Website
Roger Austin, MS

Biochemistry and Molecular Biology
Timothy Fennell, PhD
Ninell Mortensen, PhD
Delisha Stewart, PhD

Biorepository
Brian Thomas, PhD
Mike McCleary, BS

Interns
Tammy Cavallo
Aastha Ghimire
Zachery Acuff

LC-MS Metabolomics
Suraj Dhungana, PhD
Brian Thomas, PhD
James Carlson, MS
Alex Kovach, BS
Rodney Snyder, MS
Moses Darko, BS

GC-MS Metabolomics
Wimal Pathmasiri, PhD
Jocelin Deese-Spruill, BS
Keith Levine, PhD
James Harrington, PhD
William Studabaker, PhD

NMR Metabolomics
Wimal Pathmasiri, PhD
Kelly Mercier, PhD
Rodney Snyder, MS
Tammy Cavallo, BS
Kevin Knagge, PhD, DHMRI
Jason Winnike, PhD, DHMRI

Statistics, Bioinformatics, and Computing
Susan McRitchie, MS
Robert Clark, PhD
Andrew Novokhatny, BS

Advisors: Imperial College, UK
Jeremy Nicholson, PhD
Elaine Holmes, PhD
Ian Wilson, PhD