Computational and Statistical Issues in the Analysis of Metabolic Profiling Data

David M. Rocke
Division of Biostatistics (Medicine)
Department of Applied Science (Engineering)
University of California, Davis

Biostatistics
UC Davis

What is Metabolomic Profiling: A Statistican’s View?

- Measure many analytes from each biological sample
- Need to measure analytes over a very wide range of concentrations in the same sample
- Metabolic profile = estimated concentrations of many analytes
- What parts of the profile are important?

Biostatistics
UC Davis

Metrology

- Comparing two large values of an analyte, one from each sample, we might use the ratio x/y.
- Log(x/y) = log(x) – log(y)
- The log ratio is often better behaved statistically.

- When one or both of x and y is small, the ratio no longer makes sense.
- An increase from 1000 to 1200 is a 20% increase, or a ratio of 1.2
- An increase of 0 to 1000 has no percentage increase or ratio.
- Which is biologically more important?
- Log(0) is not defined.

The Glog Transform

- We use a transformation that looks like the log at high levels, but is defined and well behaved at low levels, even 0 or negative.
- This often helps regularize data.
- $\mathrm{Log}((y-\alpha)+\sqrt{[(y-\alpha)^2+\lambda]})$
- Defined for all y, monotonic, linear at 0, log for high levels

Experimental Design

- No matter how much is measured from one biological sample, the size of the experiment is the number of distinct samples/organisms.
- If you need 100 people in each group to test for one response, you need 100 to test for many.

Technologies

- Mass spectrometry of many varieties
- Many possible separation technologies with many types of detectors
- NMR spectroscopy

Statistics and Bioinformatics

- Data handling and processing of spectra, including baseline estimation
- Transformation to appropriate scales
- Identification of specific compounds or spectral regions that discriminate

- Identification of compounds from spectra
  - Multiple peaks per compound (NMR)
  - Multiple compounds per mass in MS
- Identification by computation and by database searches

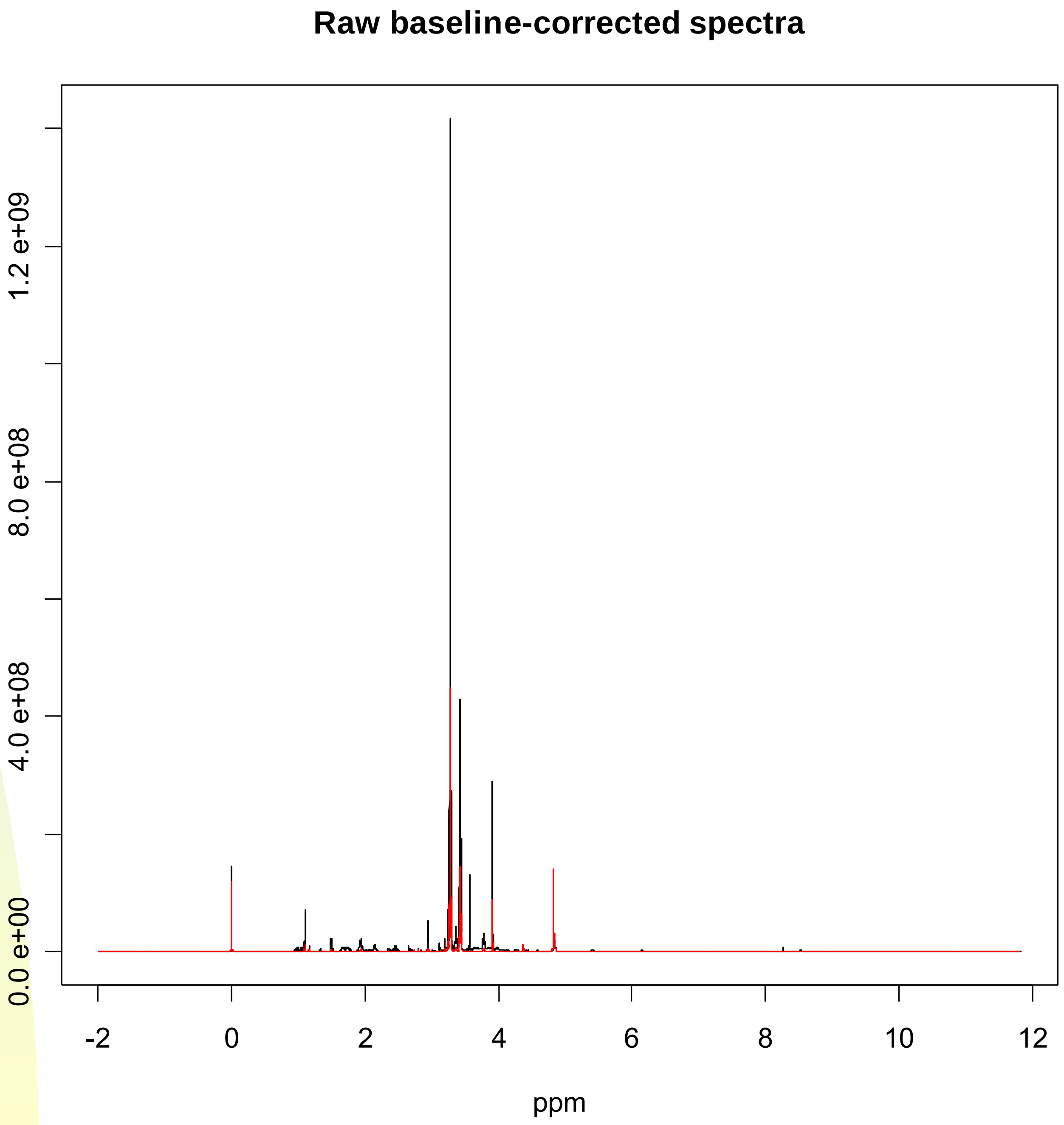
Multivariate Methods

- Methods such as PLS can identify complex biomarkers.
- Investigate other methods of dimension reduction and classification
  - PCA, PLS, SOM, stepwise selection
  - LDA, QDA, PLS, NN, Logistic Regression

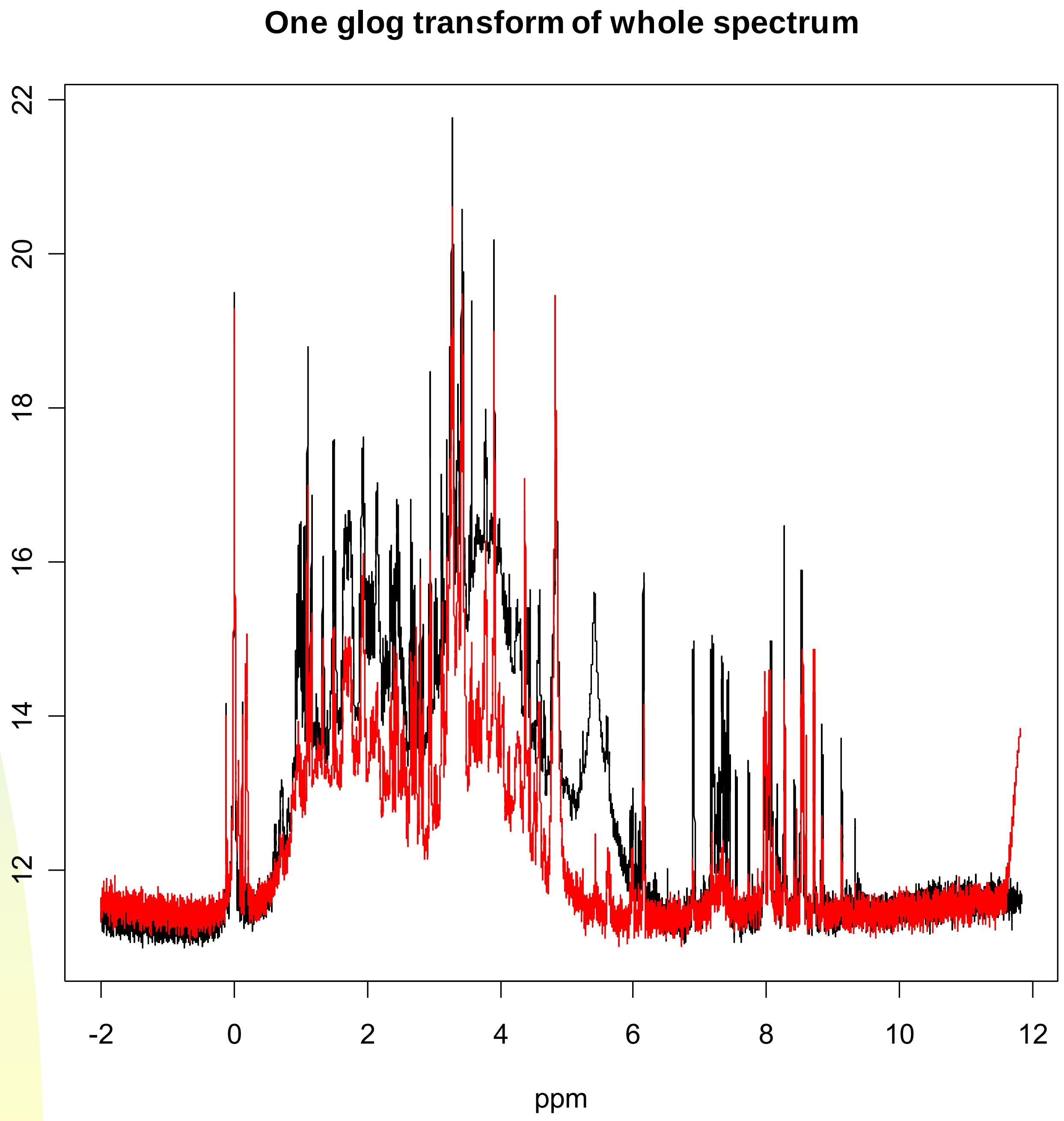
- Critical to get the statistical model and initial data processing right.
- Many methodologies assume stability of variance
- Those that don’t assume it are often more effective when this stability exists
- Carefully chosen data transformations can accomplish this

NMR Spectroscopy for Metabolomic Profiling

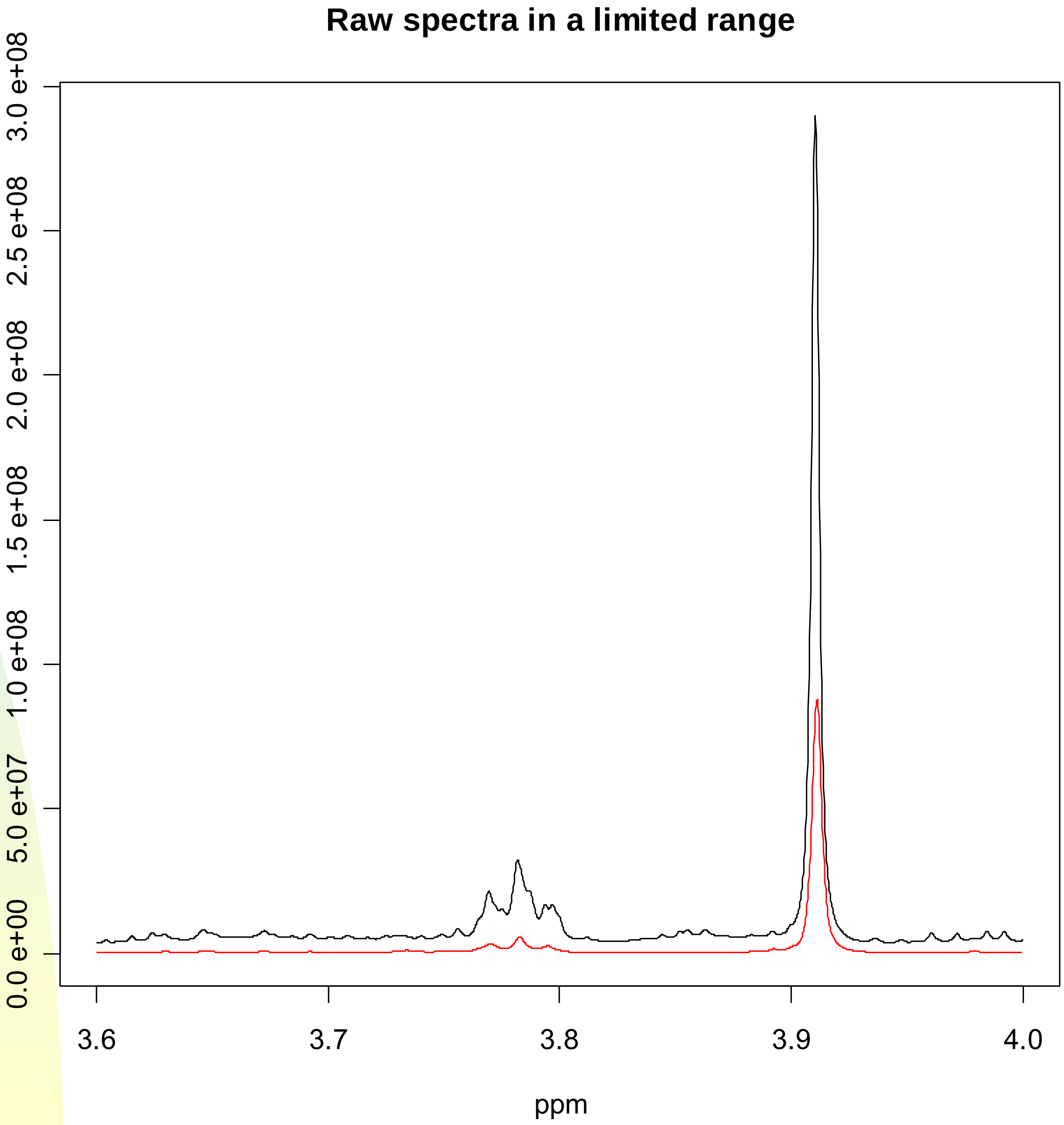
- Many computational and statistical challenges.
  - Baseline correction
  - Peak shifting
  - Multiple peaks per compound
- We are currently exploring methods of analysis for this tool.



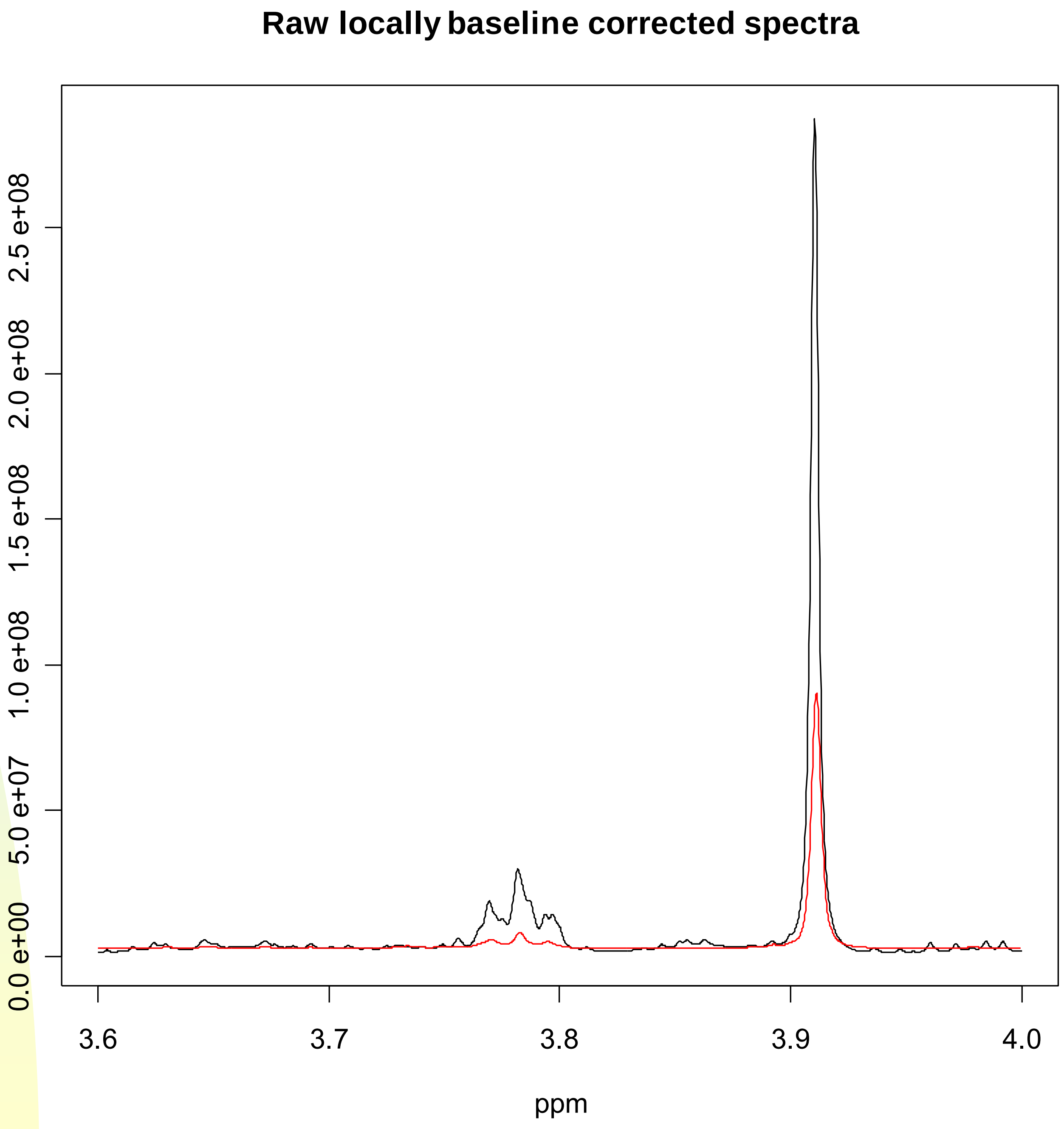
Biostatistics
UC Davis



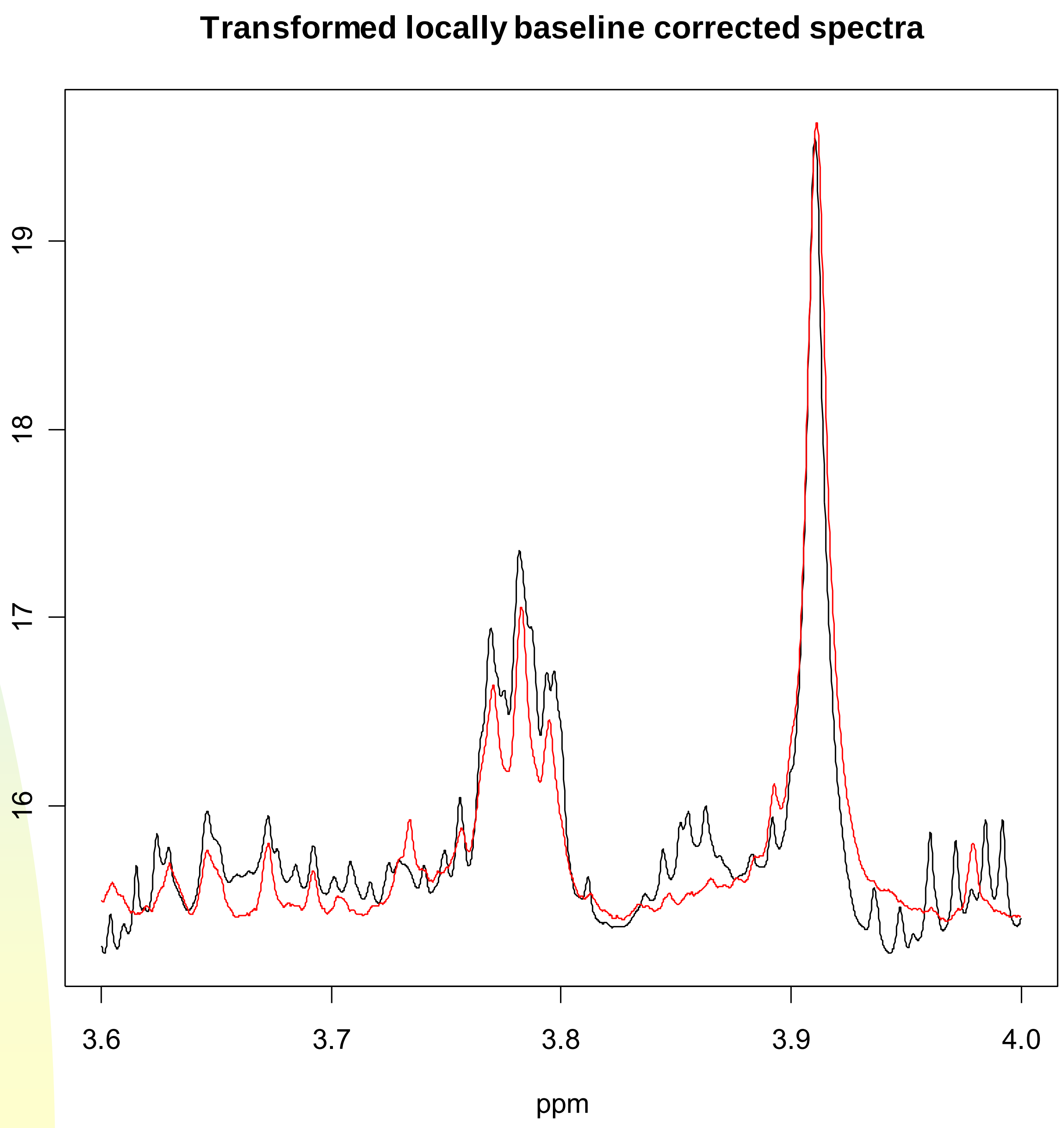
Biostatistics
UC Davis



Biostatistics
UC Davis



Biostatistics
UC Davis



Biostatistics
UC Davis

Tentative Conclusions about NMR Spectroscopy

- The baseline needs to be estimated in an adaptive but statistically principled fashion.
- The data need to be transformed to approximate stability.
- Normalization after transformation is likely to be necessary.
- We need to use this powerful tool to identify biomarkers not easily identifiable otherwise.

Conclusions

- Data handling and statistical design and analysis are both important enabling technologies.
- Similar issues will be apparent in any spectroscopic technology.
- Collaborations between many disciplines will be needed to advance metabolic profiling.