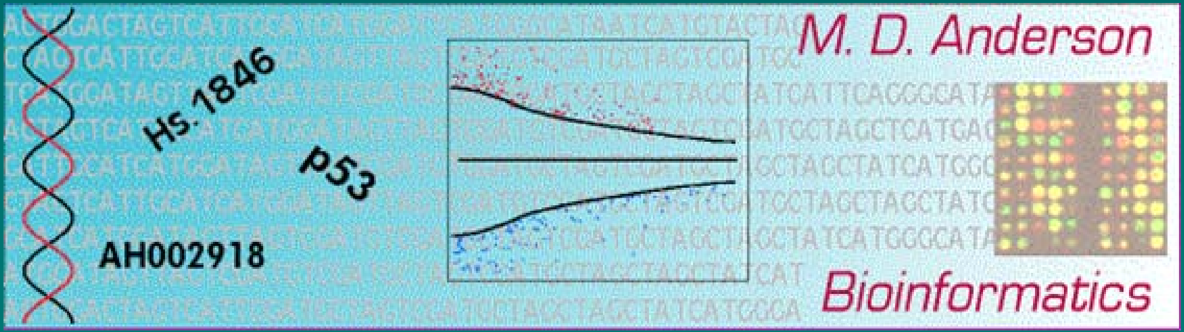
Low-level processing of proteomics spectra

Kevin Coombes

Department of Biostatistics and Applied Mathematics

UT M.D.Anderson Cancer Center



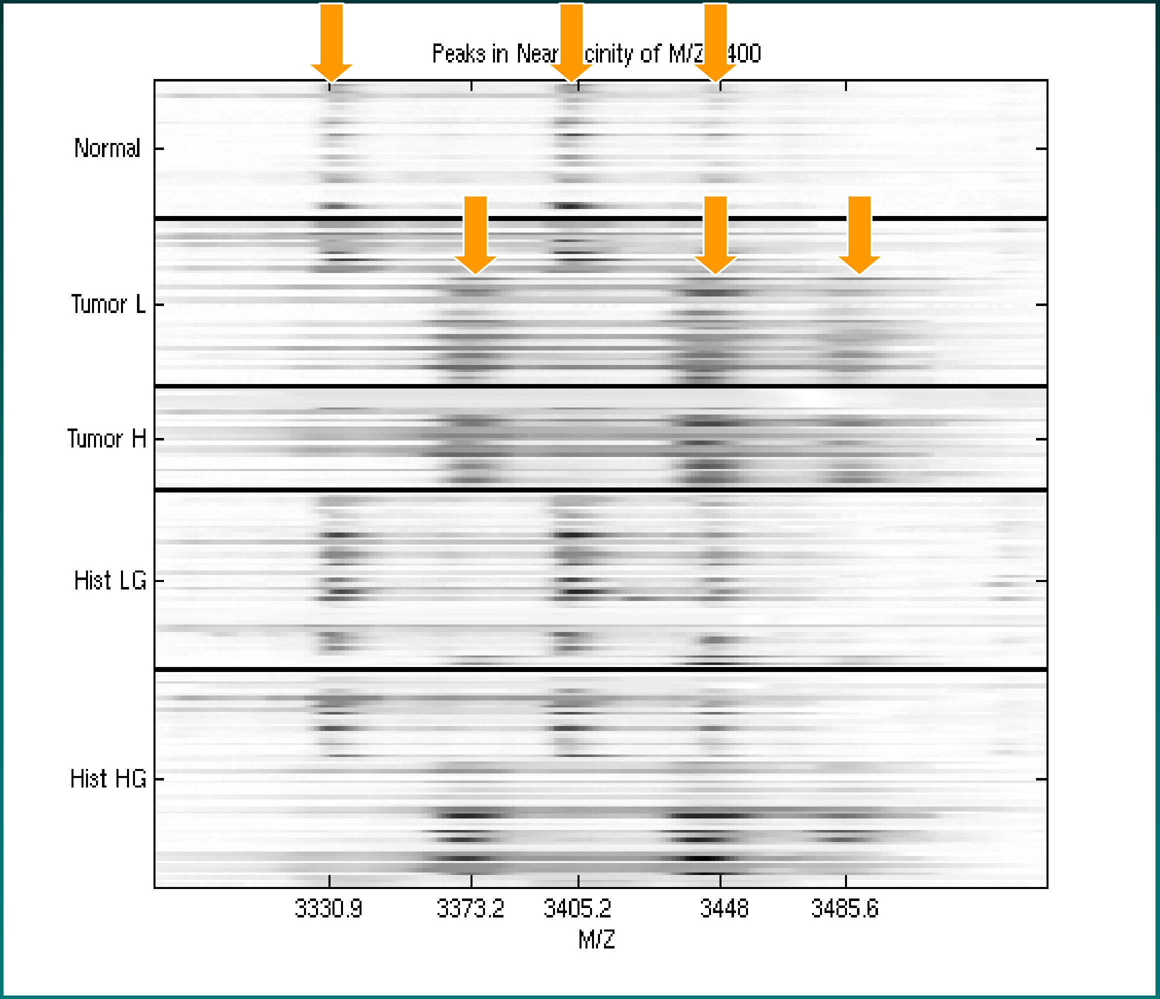
Overview

- Background and motivation
- Description of data set for methodology development and testing
- Wavelet denoising
- Using the mean spectrum
- Simulating spectra
- Open problems

Background and motivation

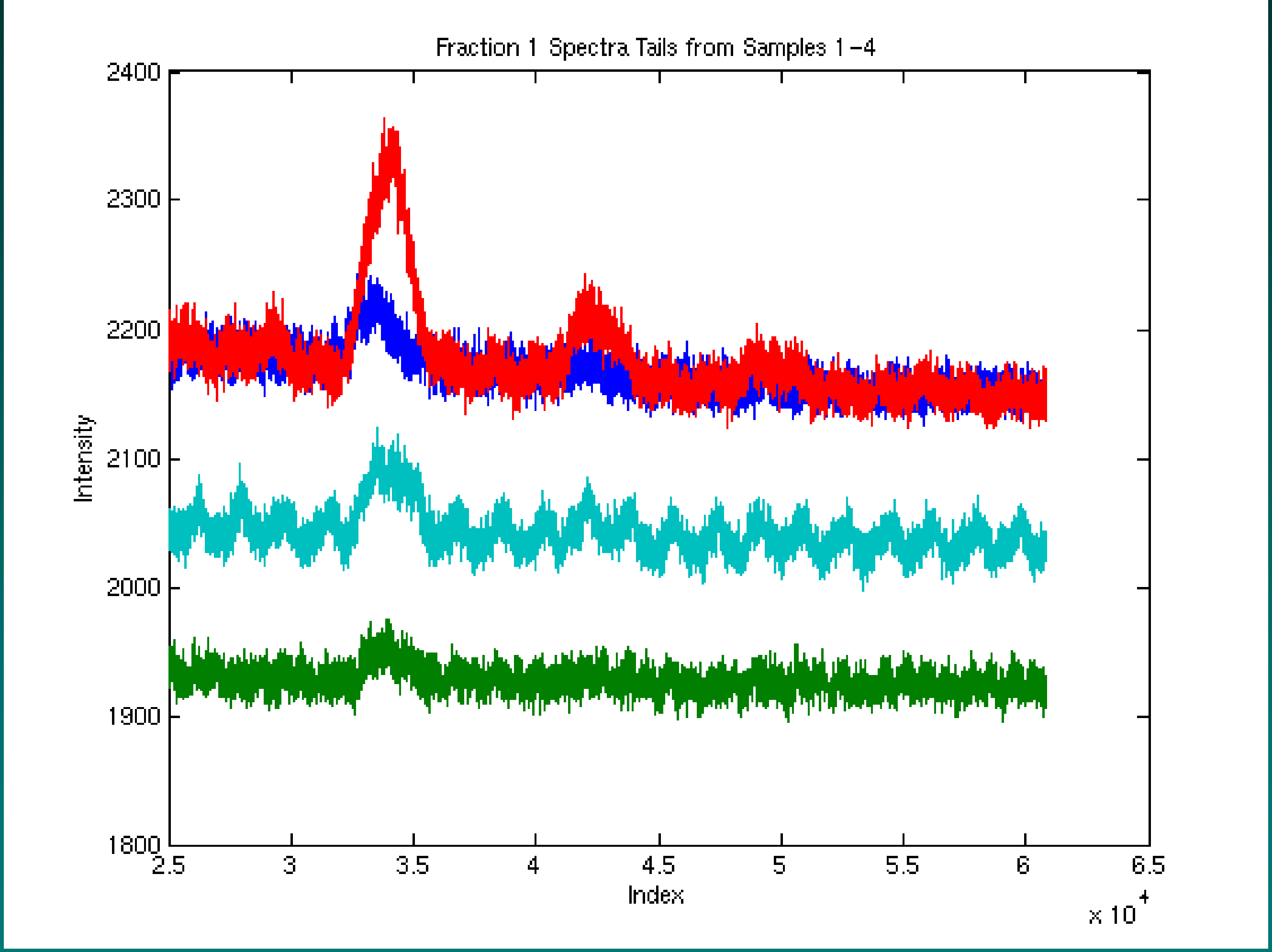
- Mass spectrometry instruments are very sensitive; they see everything
- Artifacts can be introduced into spectra from physical, electrical, or chemical sources
- Low-level processing is an attempt to remove systematic artifacts and isolate the true protein signal

Miscalibration can be misleading



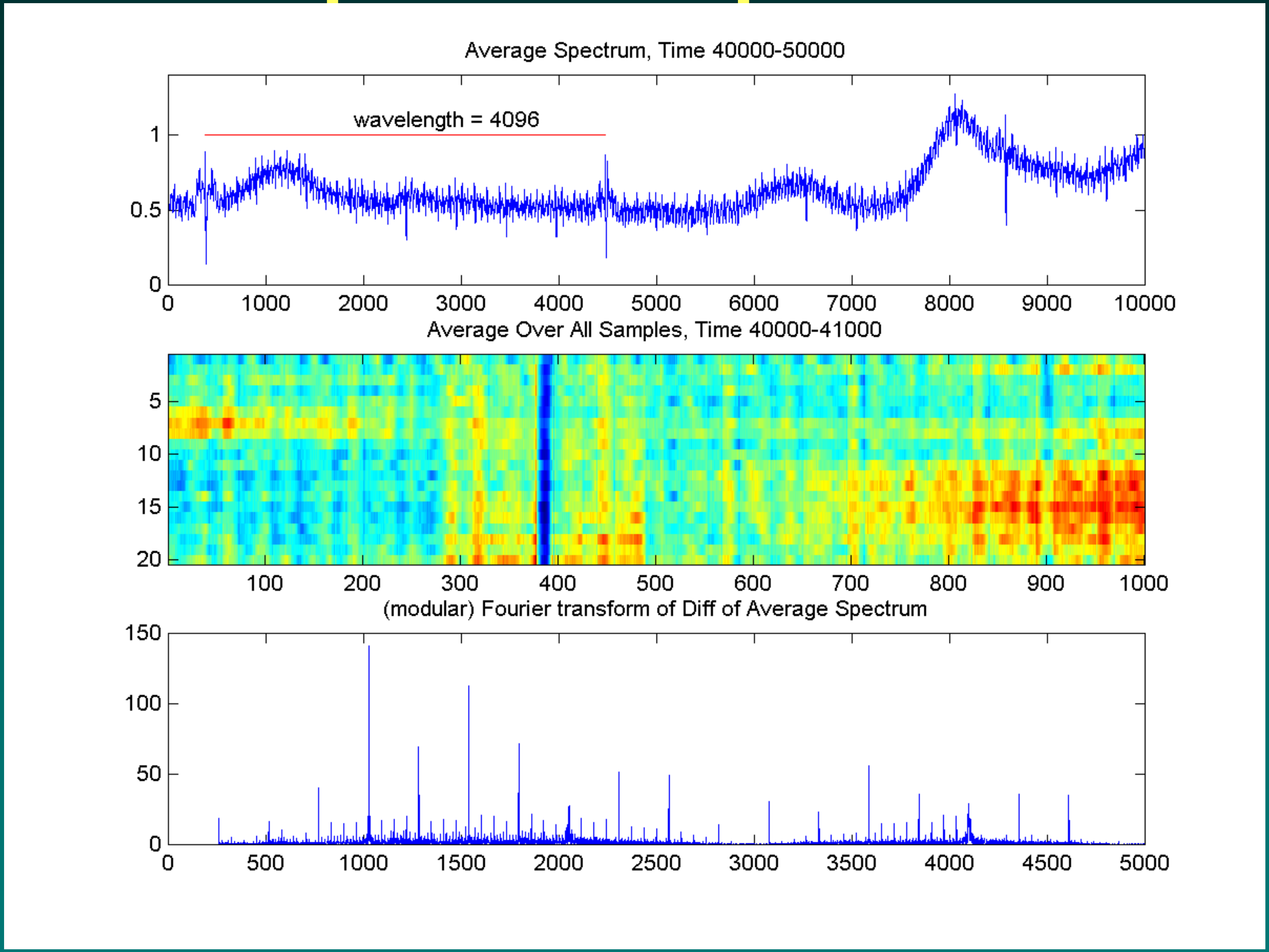
SELDI data from MDACC

Sinusoidal noise can be caused by faulty power supplies or detectors



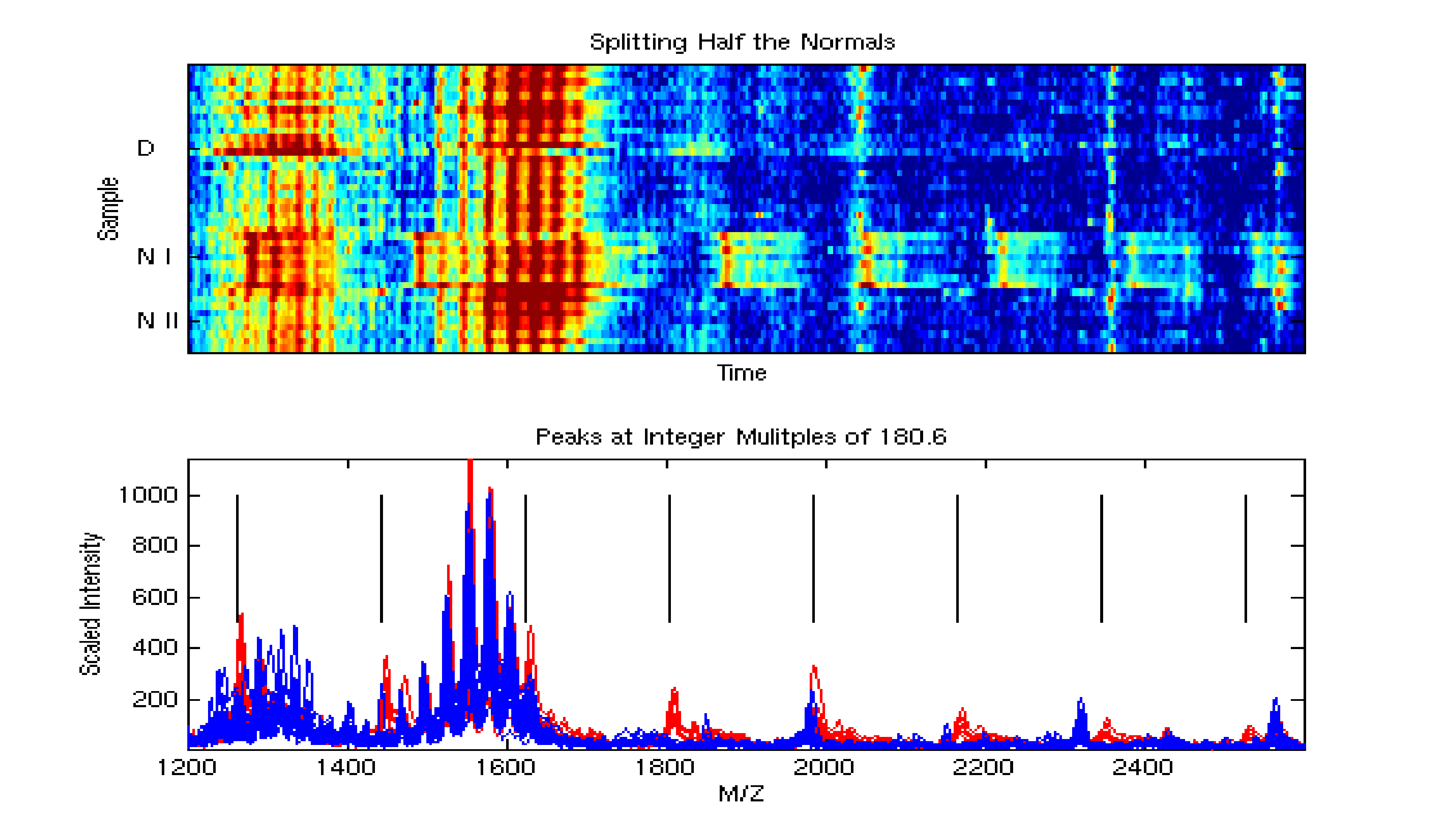
Lung cancer data from Duke Radiology

Computer clock can insert unusual spikes into spectra



Lung cancer data from Duke Radiology

Polymers are unlikely to yield interesting biology

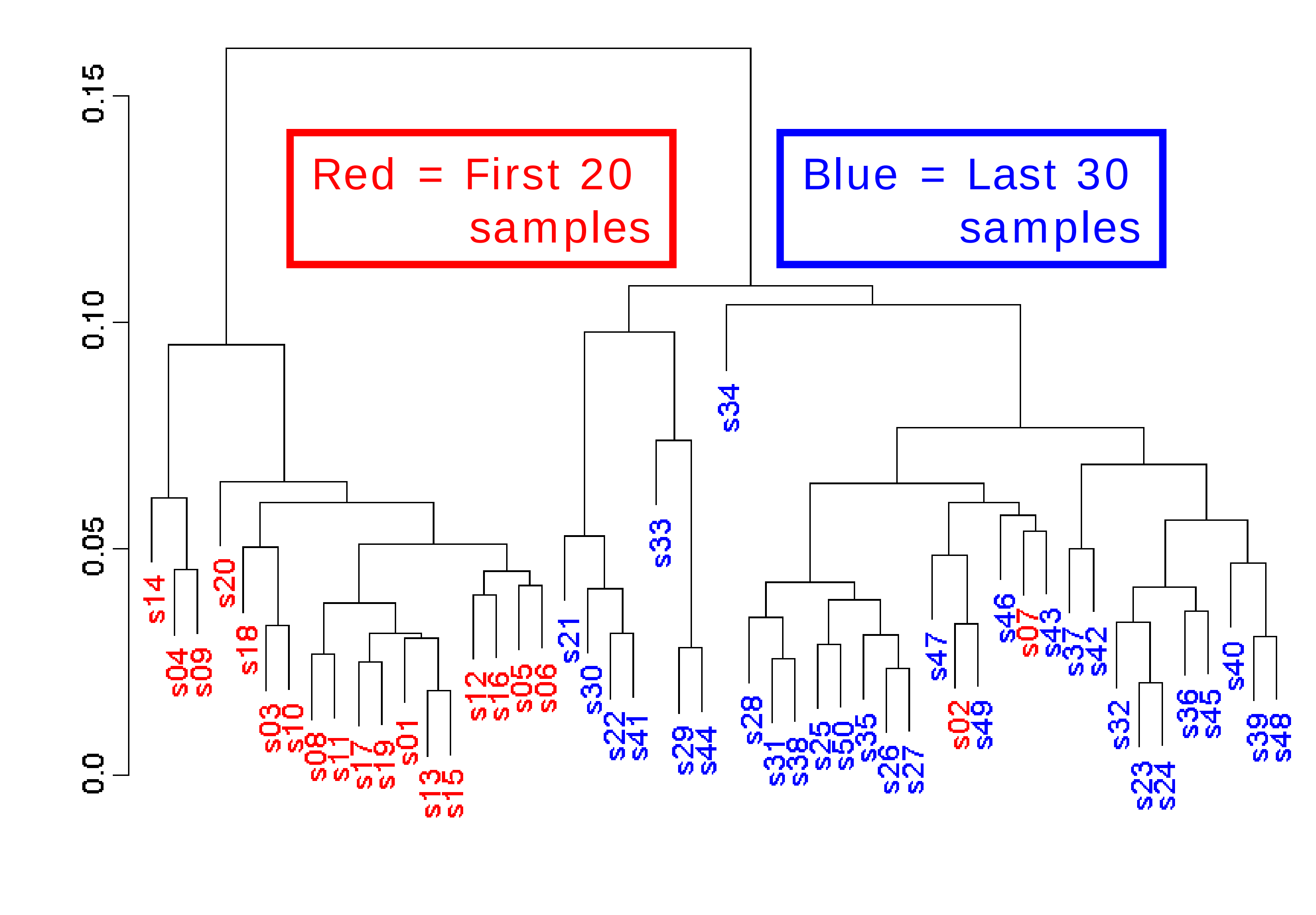


Splitting Half the Normals

Peaks at Integer Mulitples of 180.6

Lung cancer data from Duke Radiology

Differences in the sample collection protocol can dominate the results



Red = First 20 samples

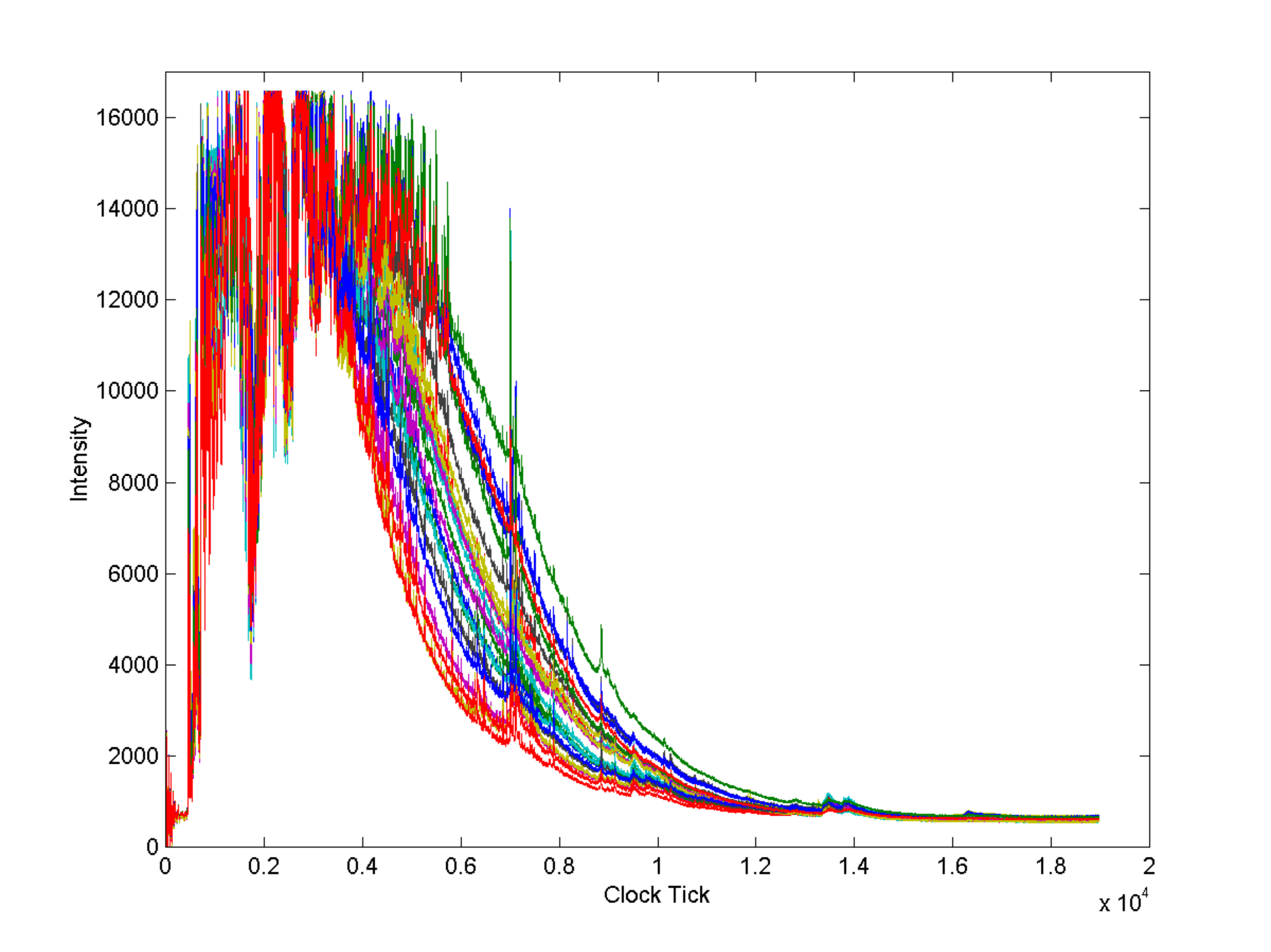
Blue = Last 30 samples

MALDI data from MDACC

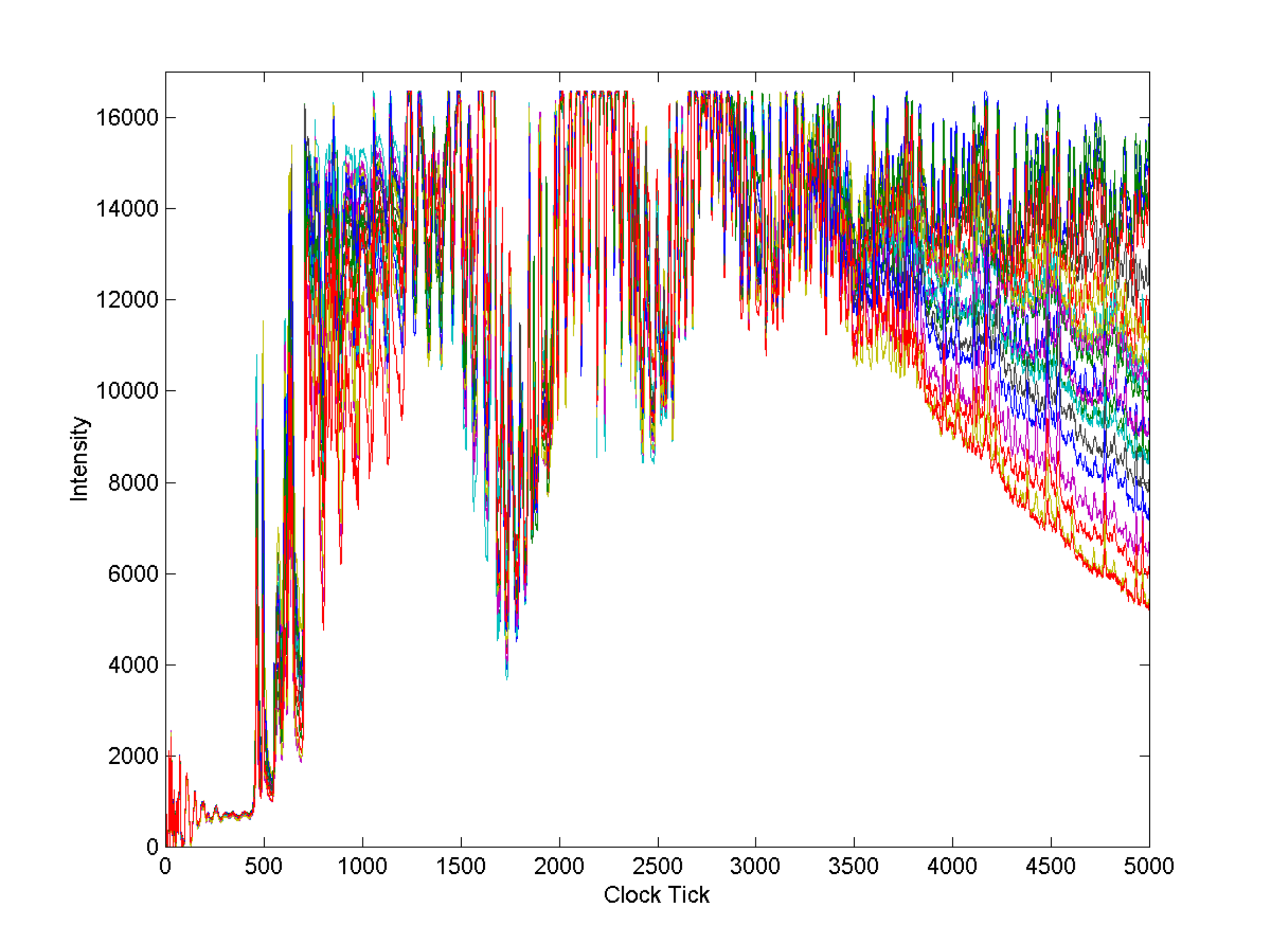
Data set for developing and testing methods for low-level processing

- One pooled sample of nipple aspirate fluid, divided into aliquots
- Three 8-spot Ciphergen chips
- On each of four days, apply sample to two spots on each chip
- Produces 24 replicate spectra
- Note: We performed the experiment with WCX2 and IMAC3 chips, and scanned each spot at two different intensities

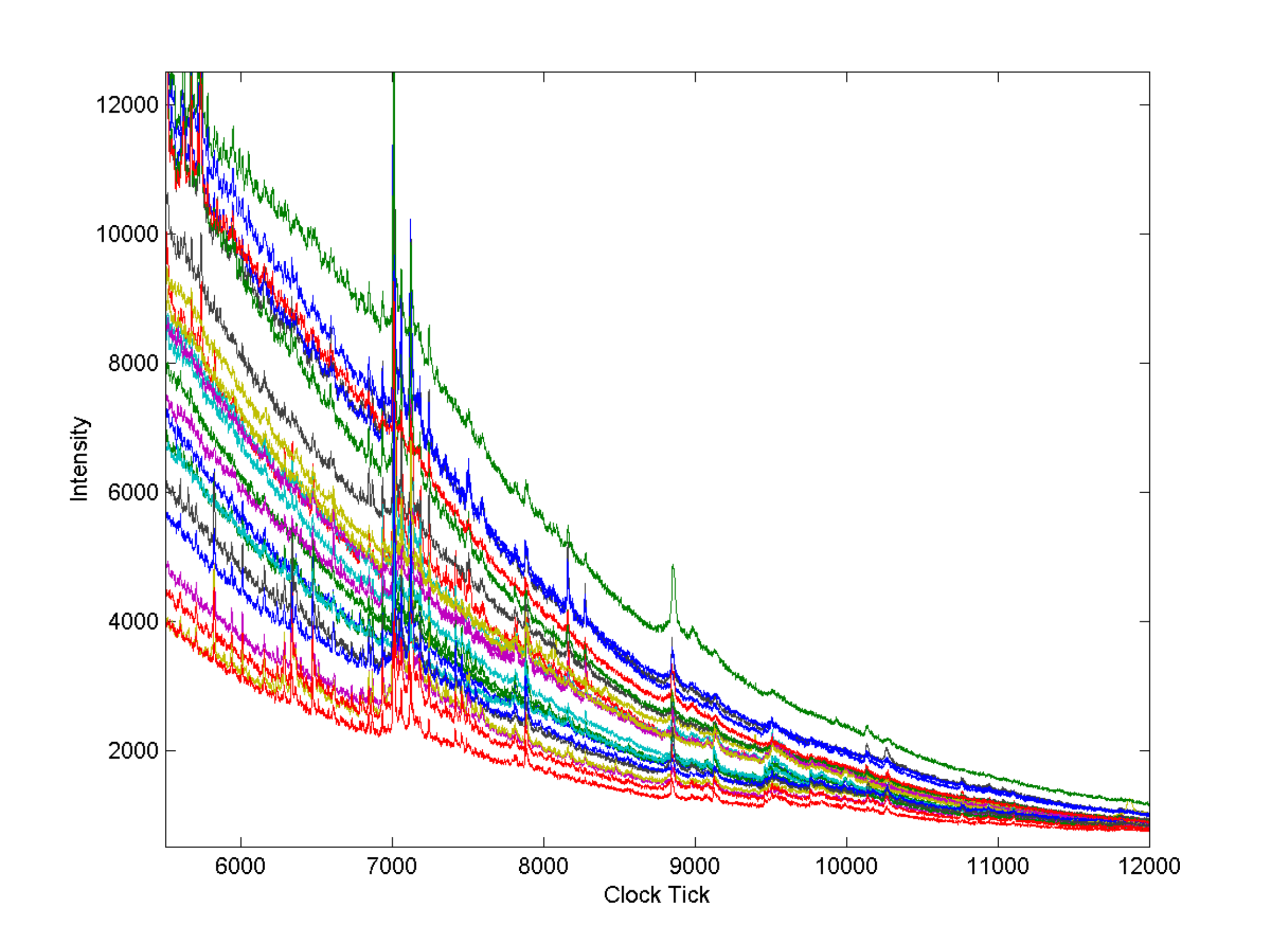
Twenty-four spectra from the same sample on WCX2 chips, low mass range



Saturation occurs frequently in the early portion of the spectrum



Individual spectra have different baseline curves, but reproducible peaks



Low-level processing

- View each spectrum as composed of three components
  - True peak signal
  - “Exponential” baseline
  - White noise
- Primary goal of low-level processing is to disentangle these three components

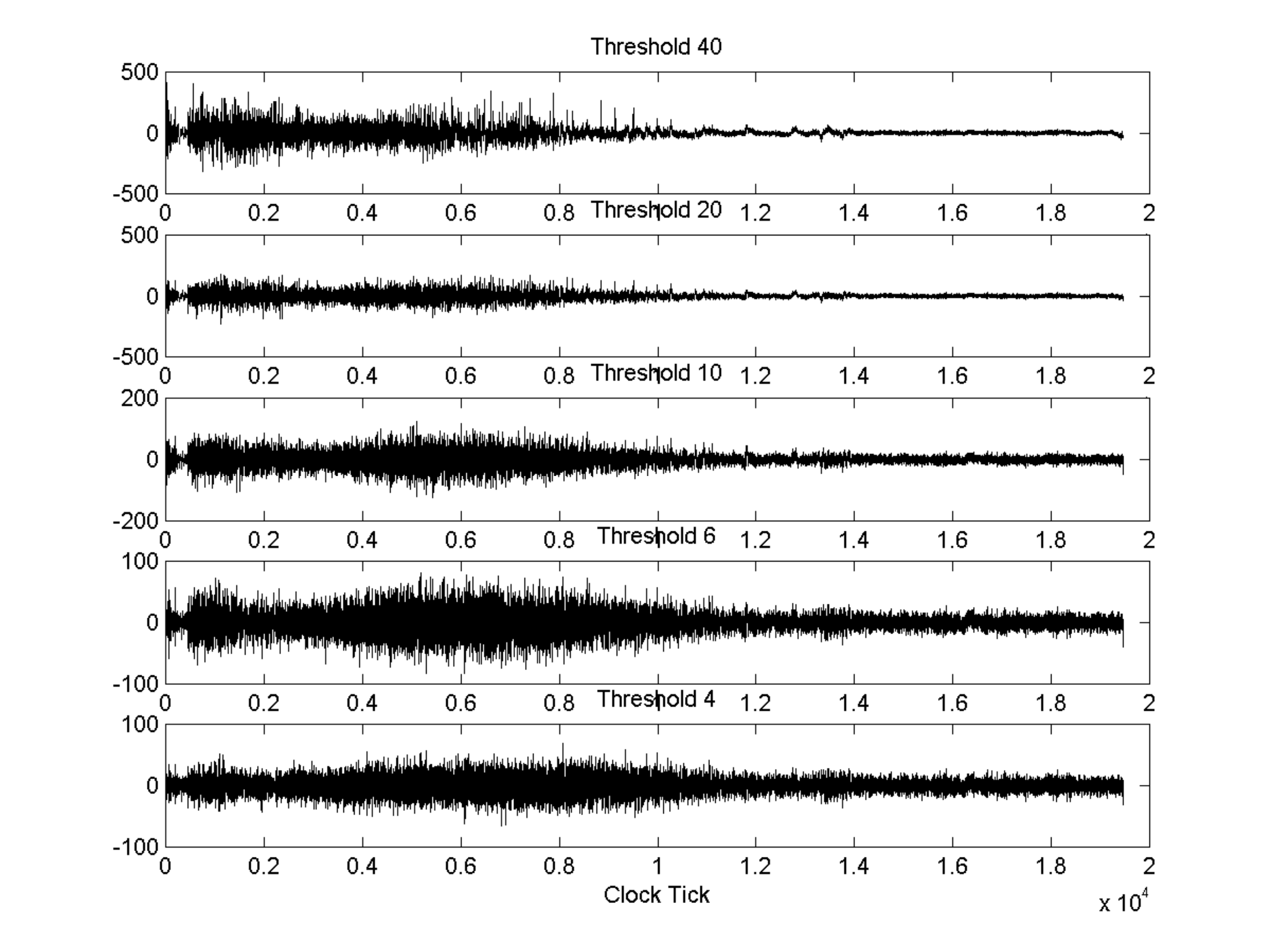
Wavelet denoising

- Idea: use the undecimated discrete wavelet transform to isolate the white noise component
  - Undecimated implies “shift-invariant”, so the results don’t depend on where you start processing the signal
  - Established tool in image processing and other scientific fields
  - Code freely available in the Rice Wavelet Toolbox (http://www-dsp.rice.edu/software/rwt.shtml)

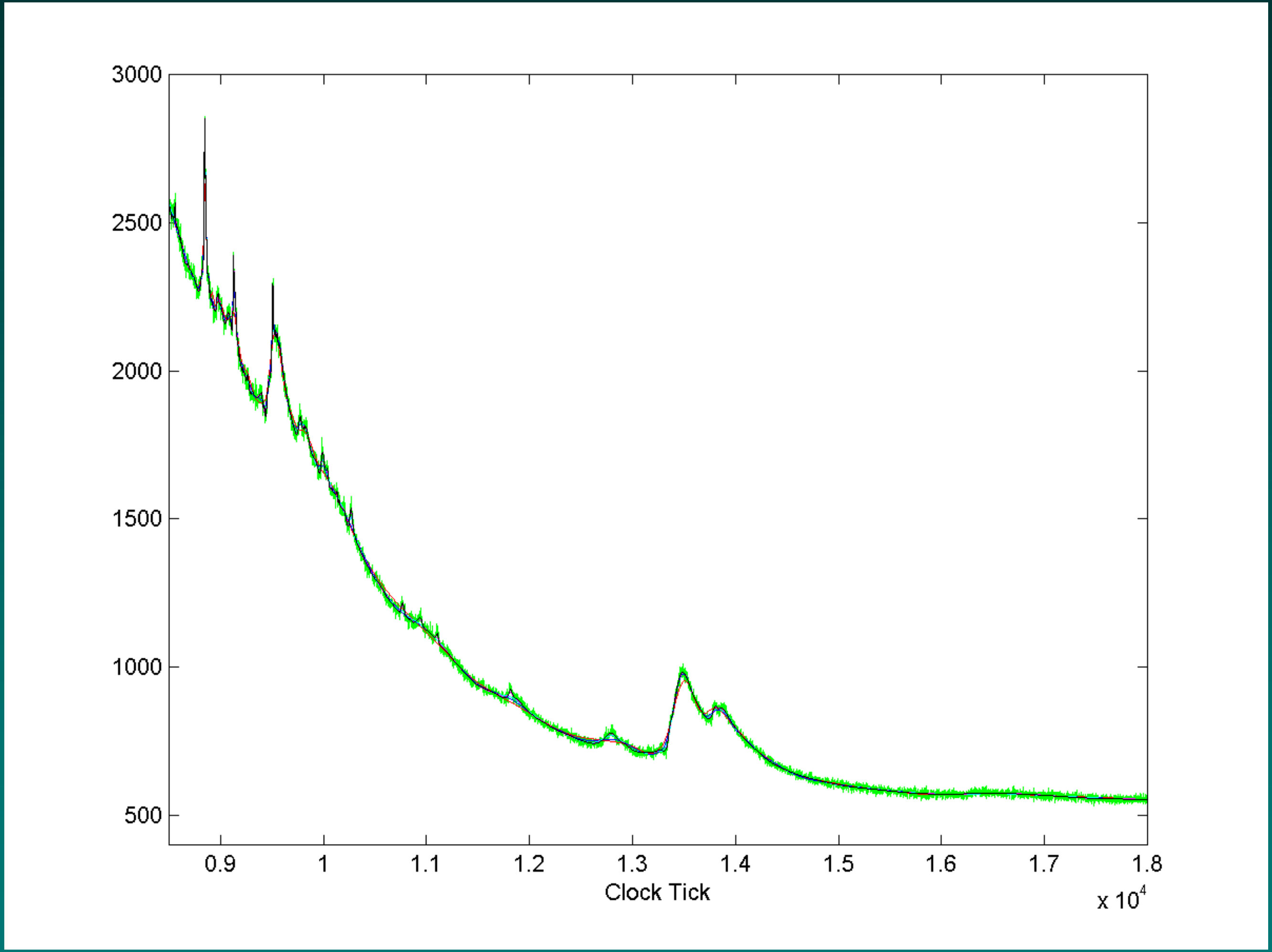
Underlying principle of denoising spectra using wavelets

- Idea:
  - Transform from the time domain to the wavelet domain
  - Discard wavelet coefficients below some threshold
  - Transform back.
- Noise should be equally distributed over all wavelet coefficients at low levels.
- True signal should be represented in a few wavelet coefficients at high levels.

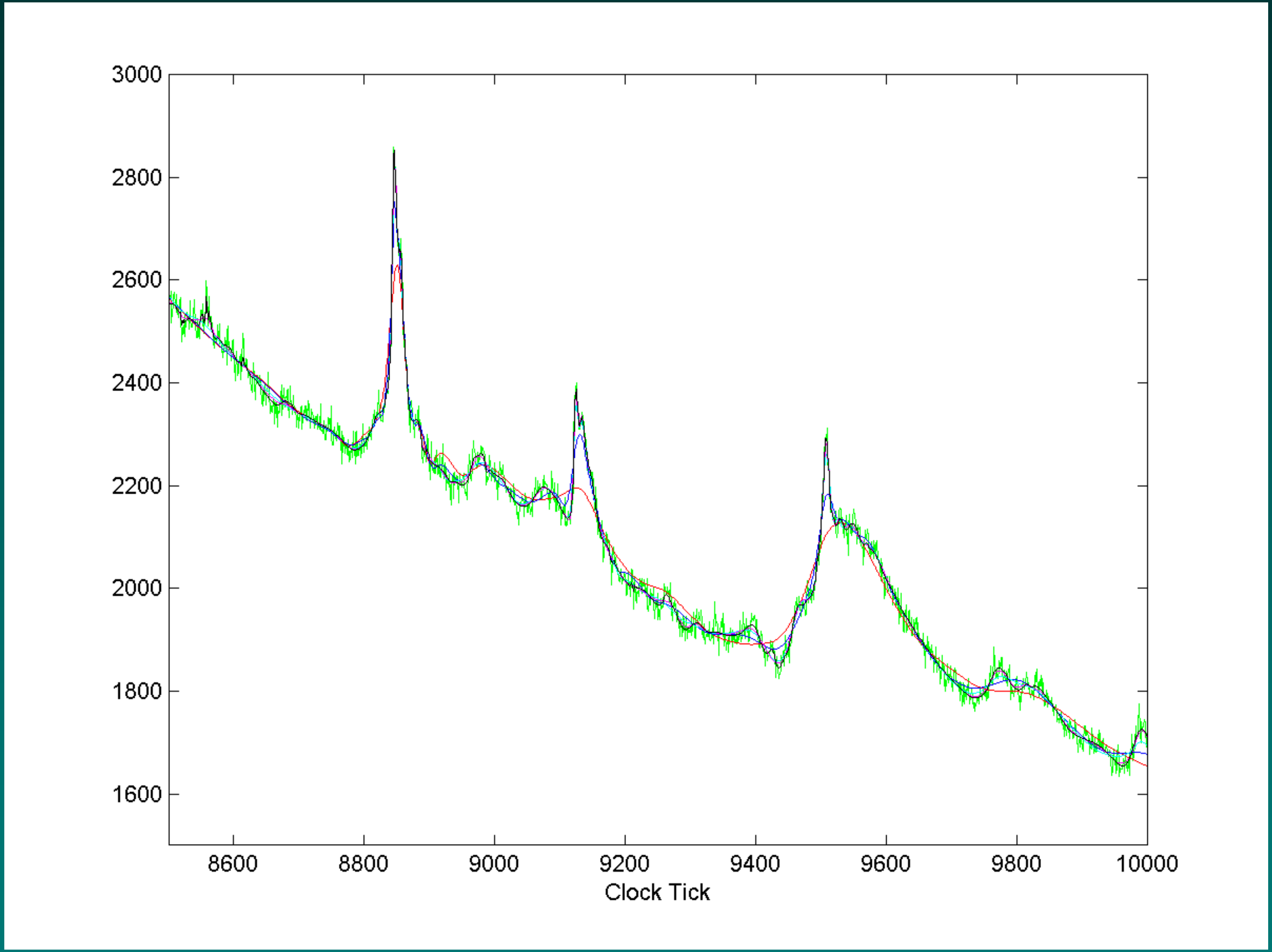
As the threshold increases, more true signal is included in the noise



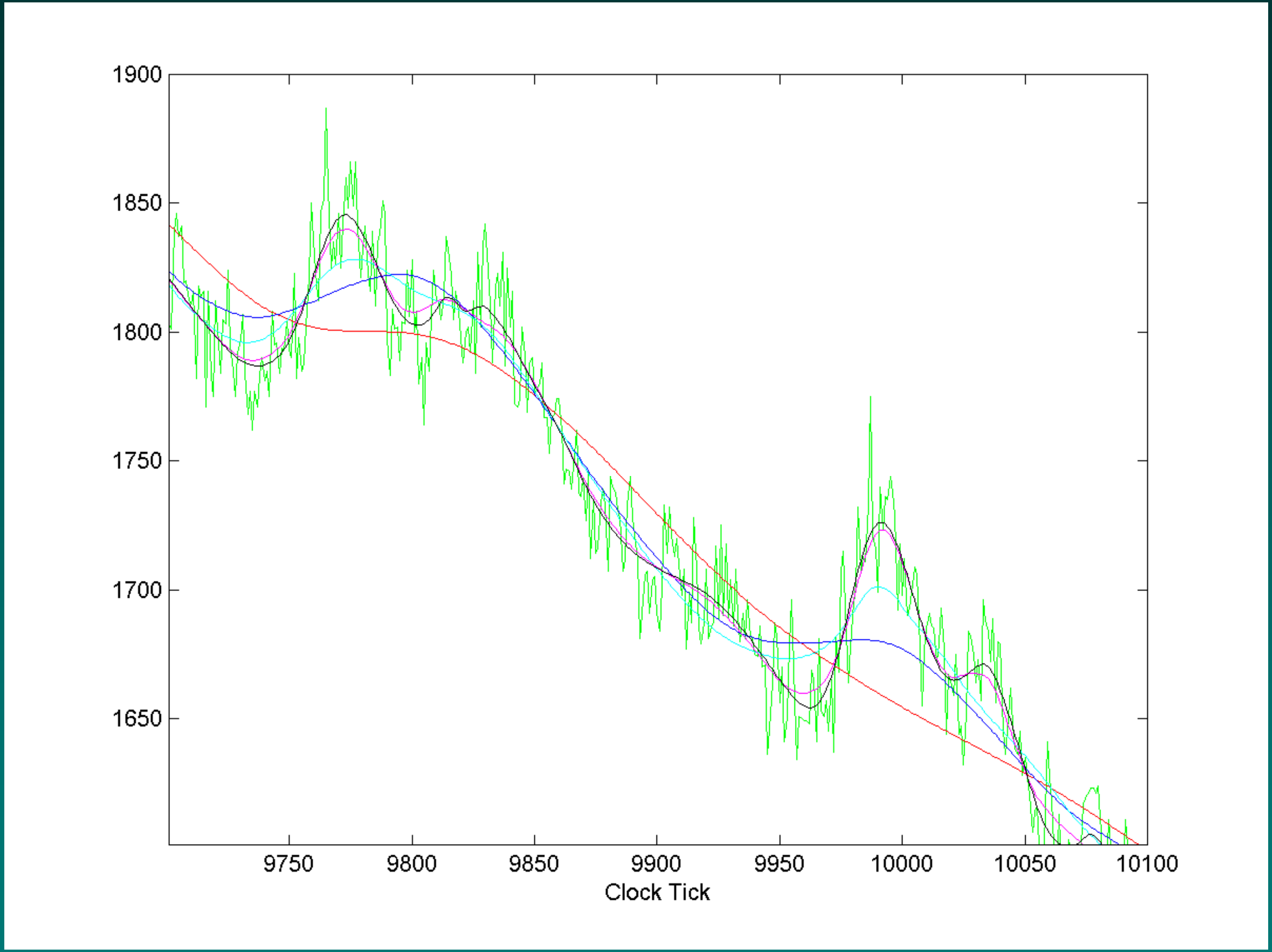
Long-range view of raw spectrum with wavelet denoised overlays



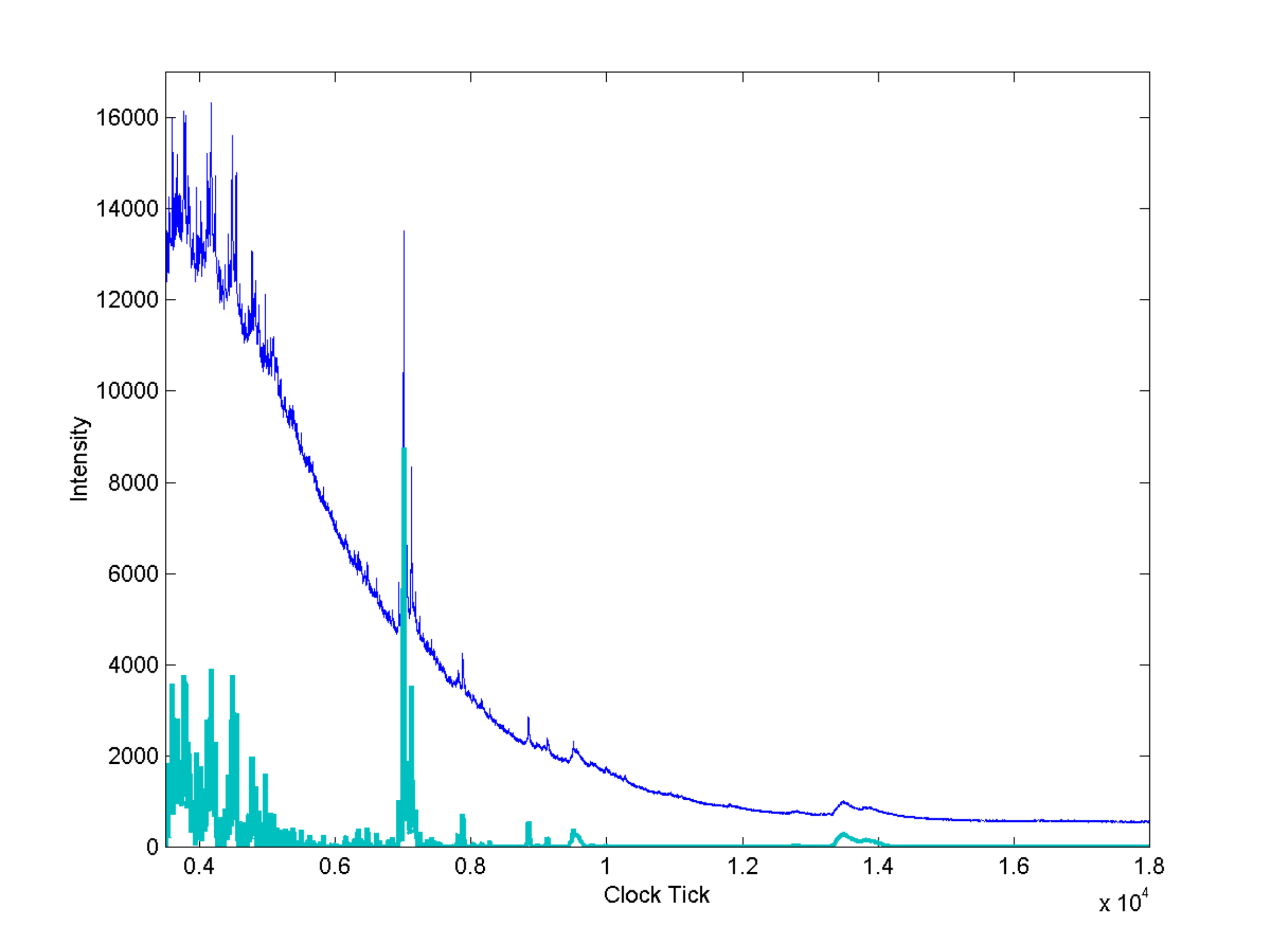
Closer view shows that high thresholds over-smooth the spectra



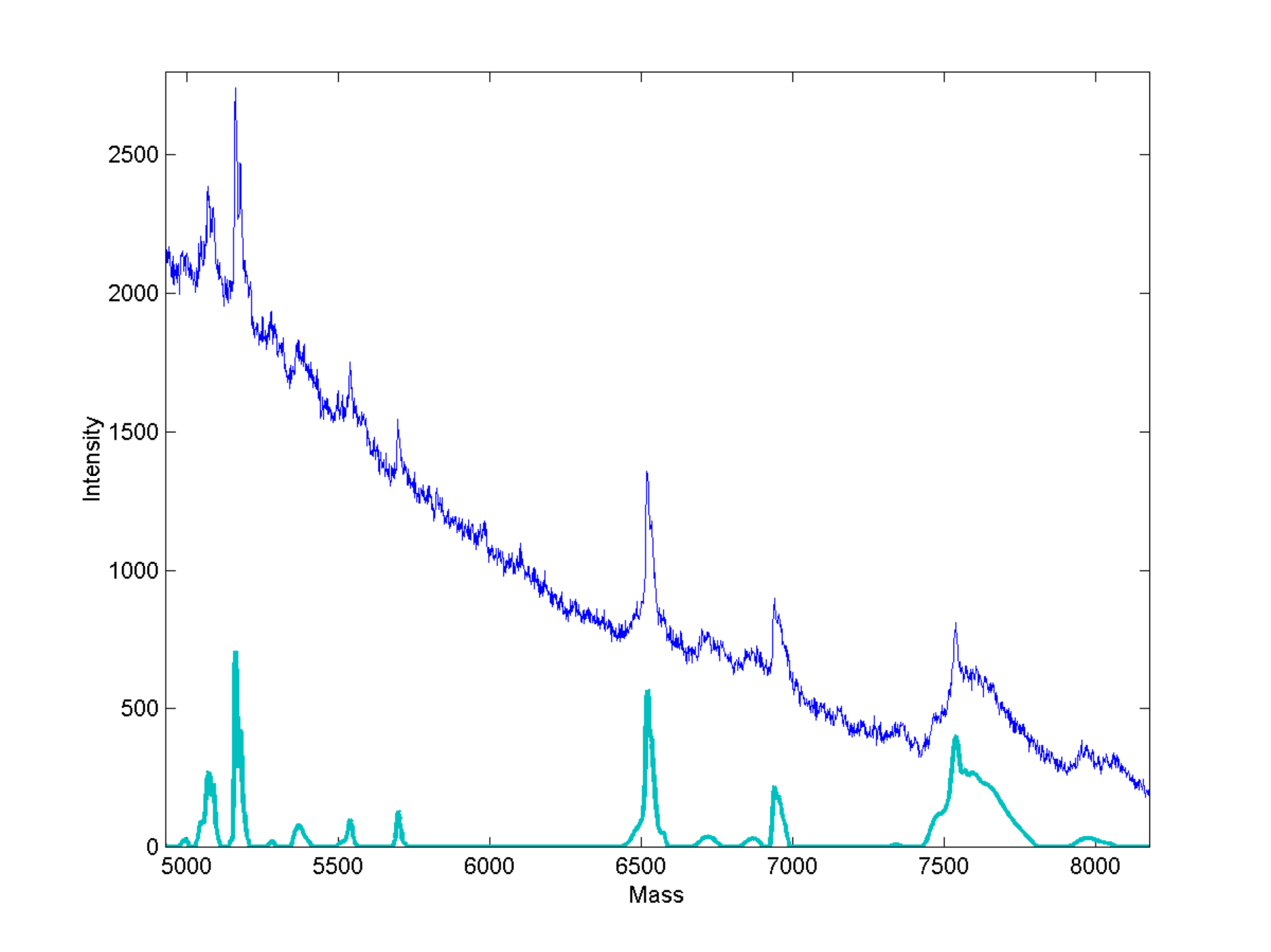
Close-up view shows that smoothing decreases noise and preserves peaks



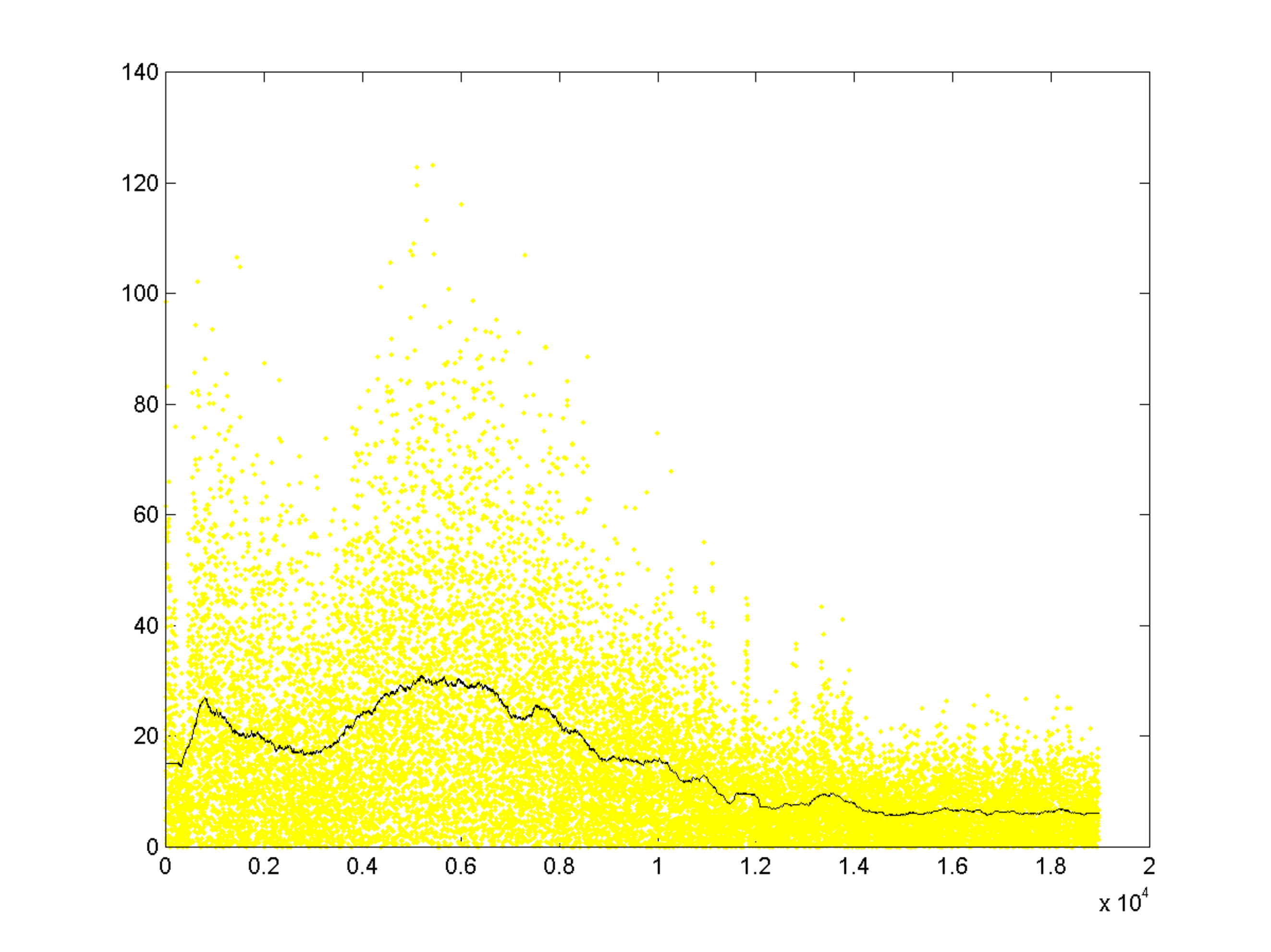
Baseline correction removes the exponential trend



Peaks are easily isolated after denoising and baseline correction



A median filter computes a running estimate of the noise



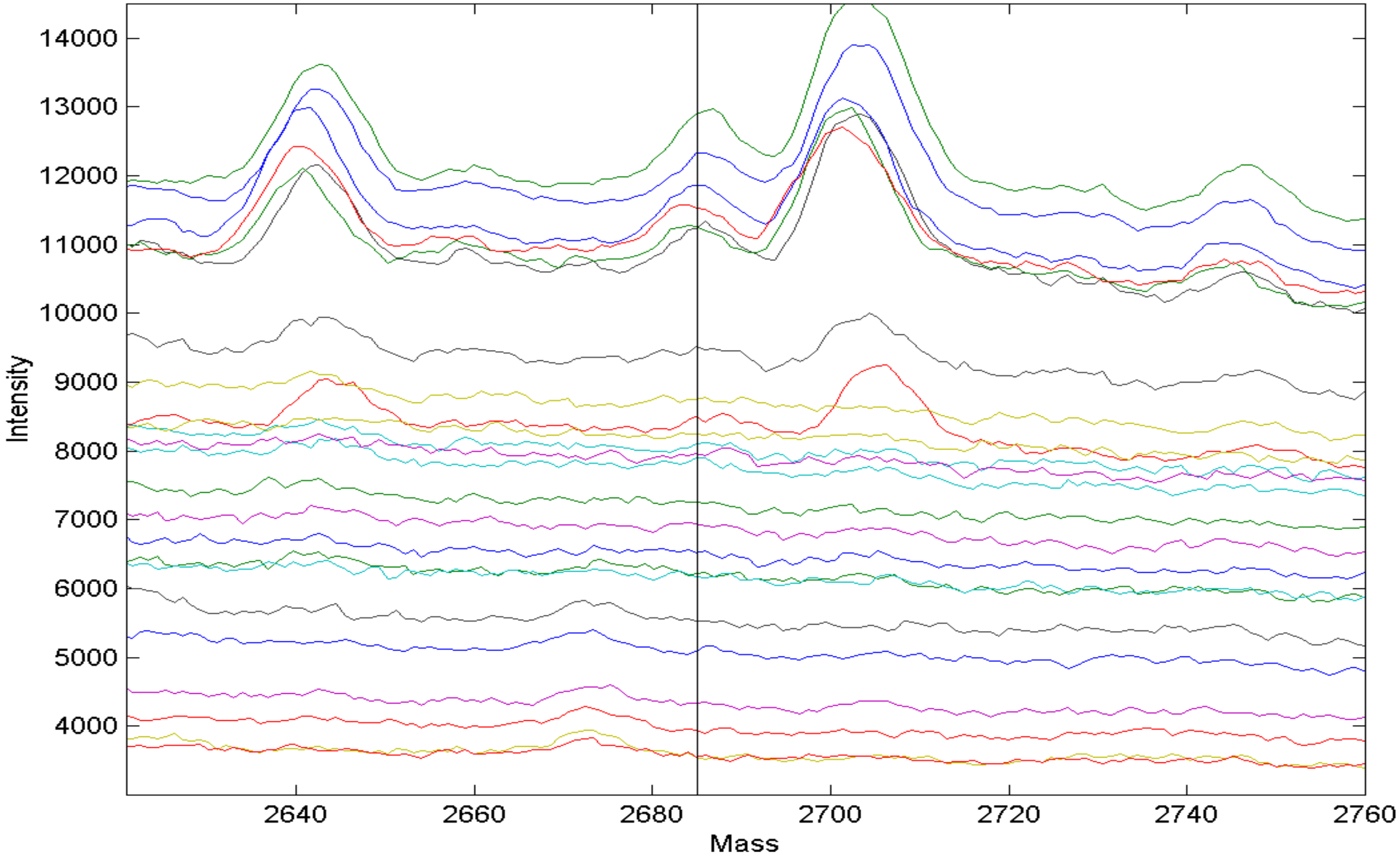
Review of the method

- Denoise using wavelets
- Baseline correct using a monotone minimum
- Normalize to total ion current (usually in a restricted mass range)
- Locate peaks as local maxima after denoising and baseline correction
- Quantify peaks as height at local maximum
- Estimate S/N as height divided by median-smoothed wavelet noise
- Match peaks across spectra (based on clock tick separation or relative mass accuracy)

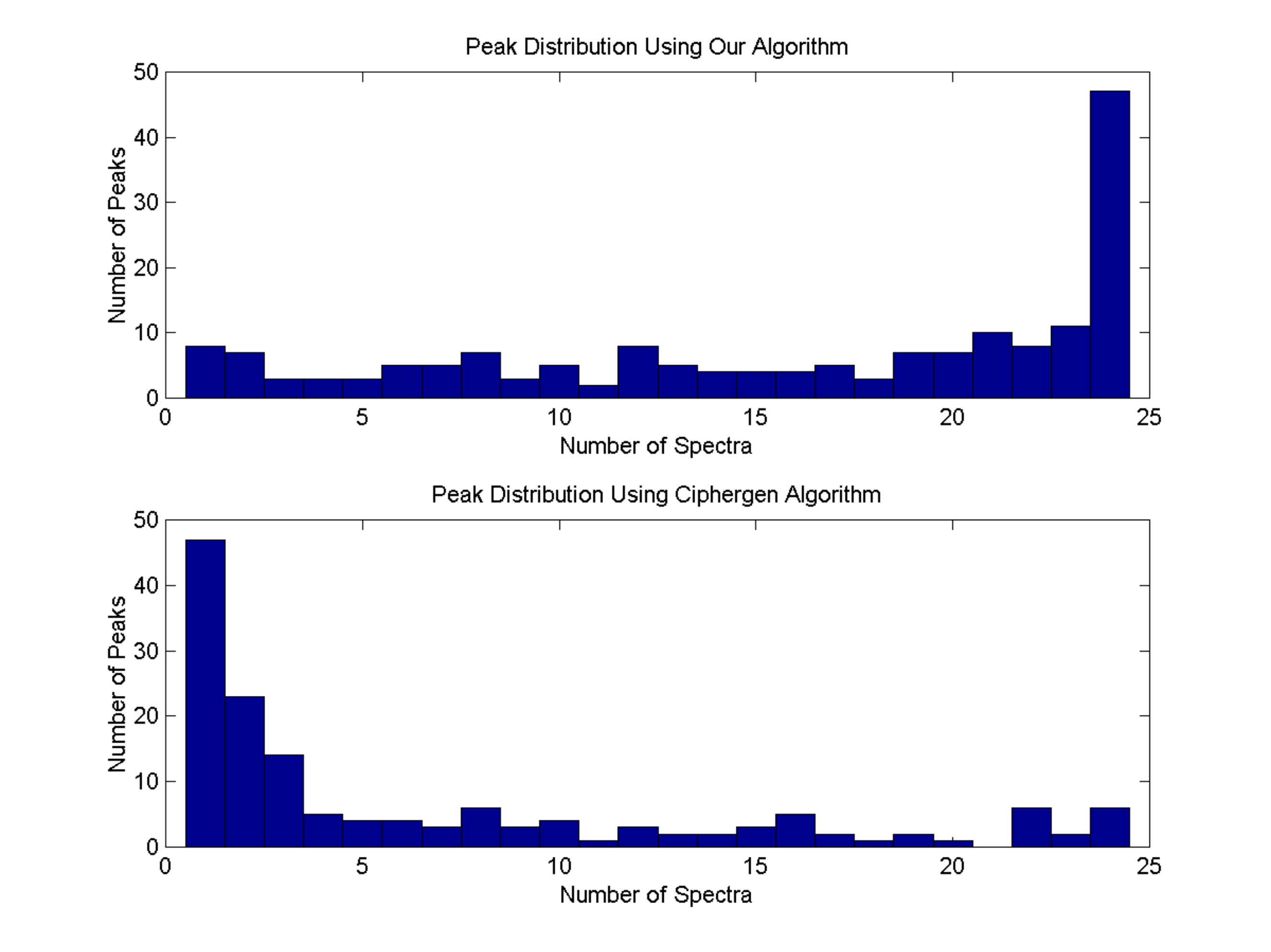
Results on the 24 replicate spectra

- On average, each of the 24 replicate spectra contained 96 peaks with S/N > 10 and 158 peaks with S/N > 2
- Match peaks if separated by 7 clock ticks or by 0.3% mass and find a total of 174 peaks that occur at least once with S/N > 10
- 47 peaks were found in all 24 spectra
- Logarithmic height of peaks found in at least 3 spectra had median CV = 11%

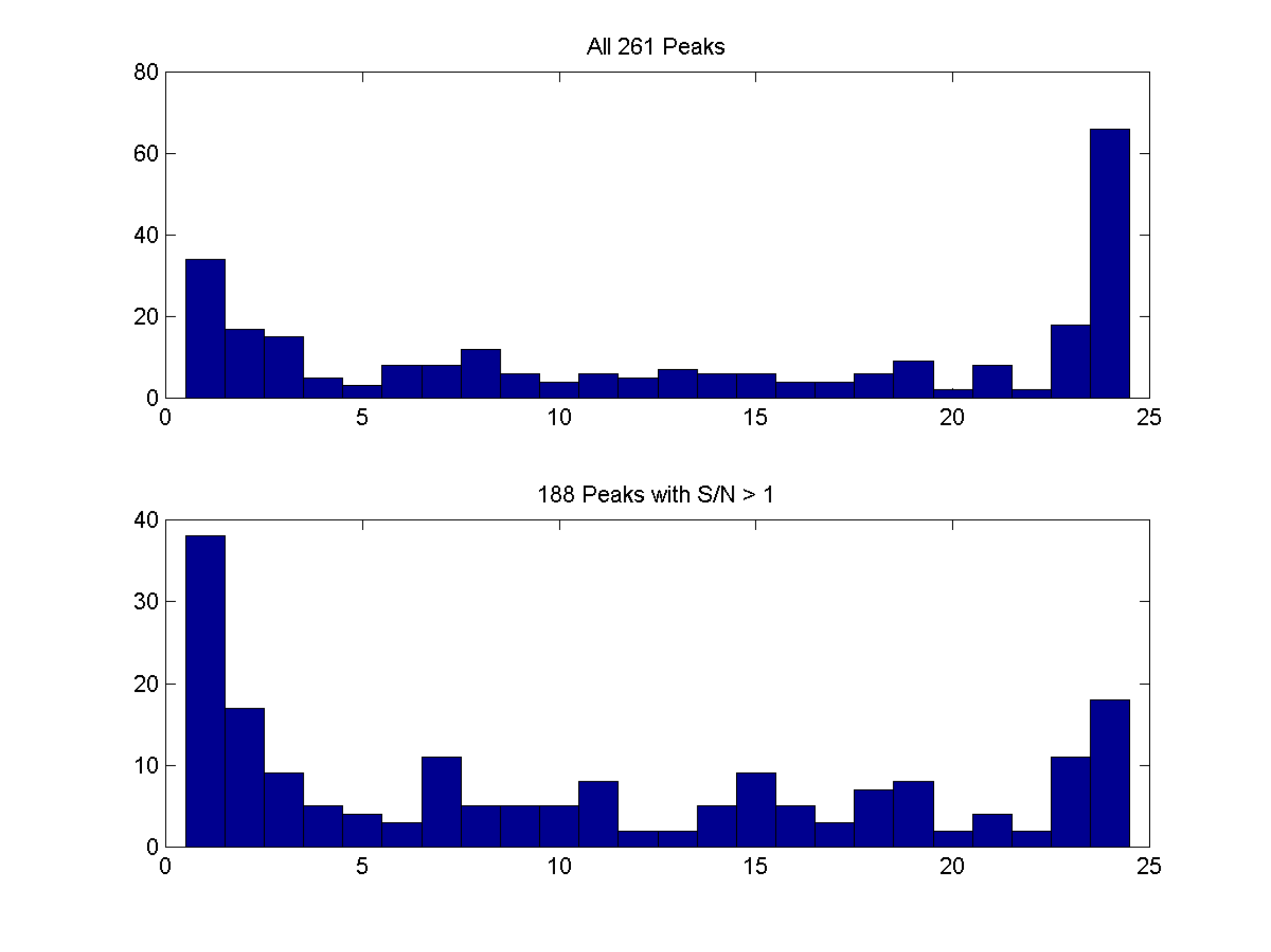
Peaks found 10 times reflect differences in technology, not in statistical processing



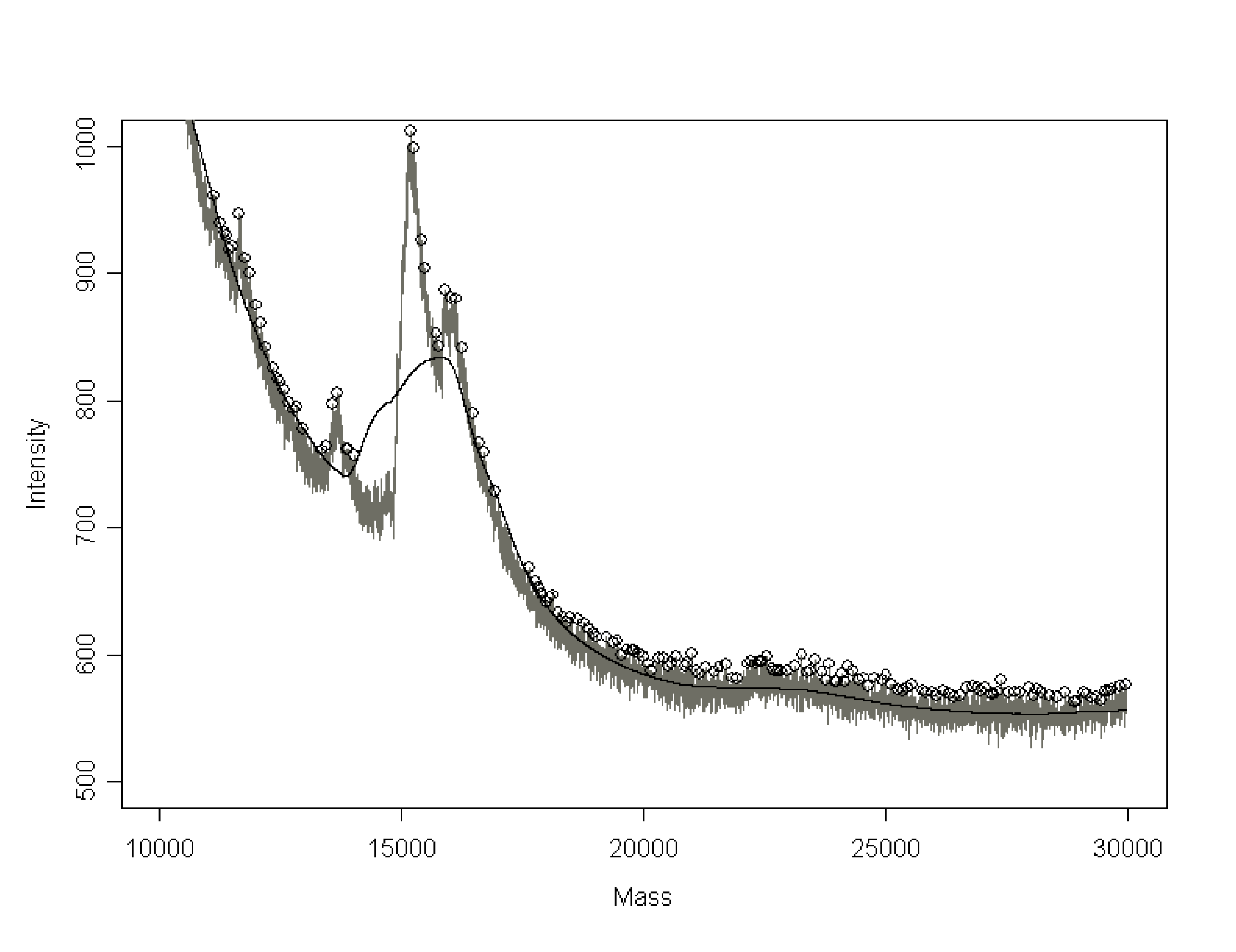
Our method find peaks more reproducibly than Ciphergen



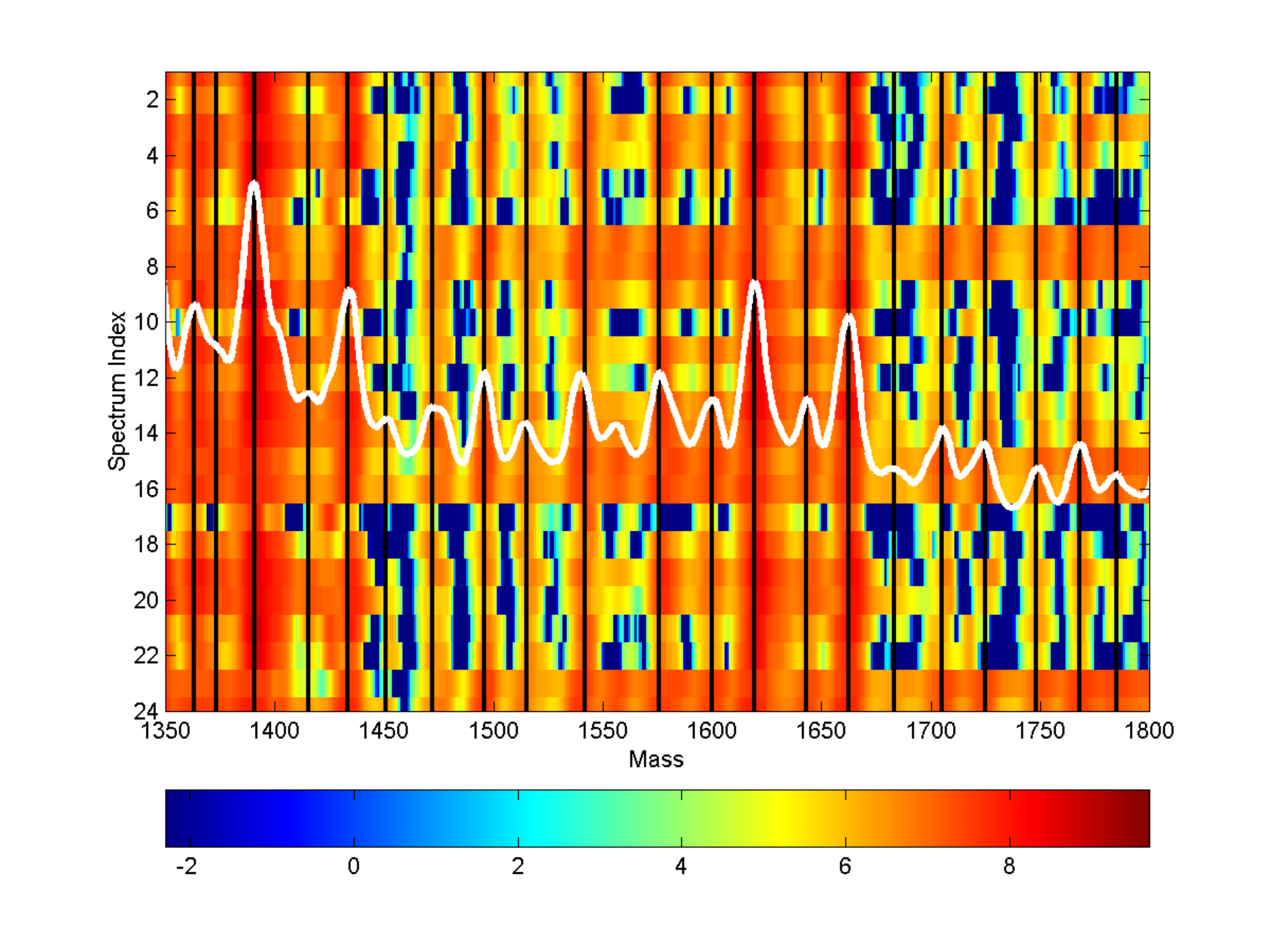
Our method finds peaks more reproducibly than Yasui et al.



Yasui et al. find many spurious peaks



Peaks found at least 10 times are visible in most spectra, and in the mean



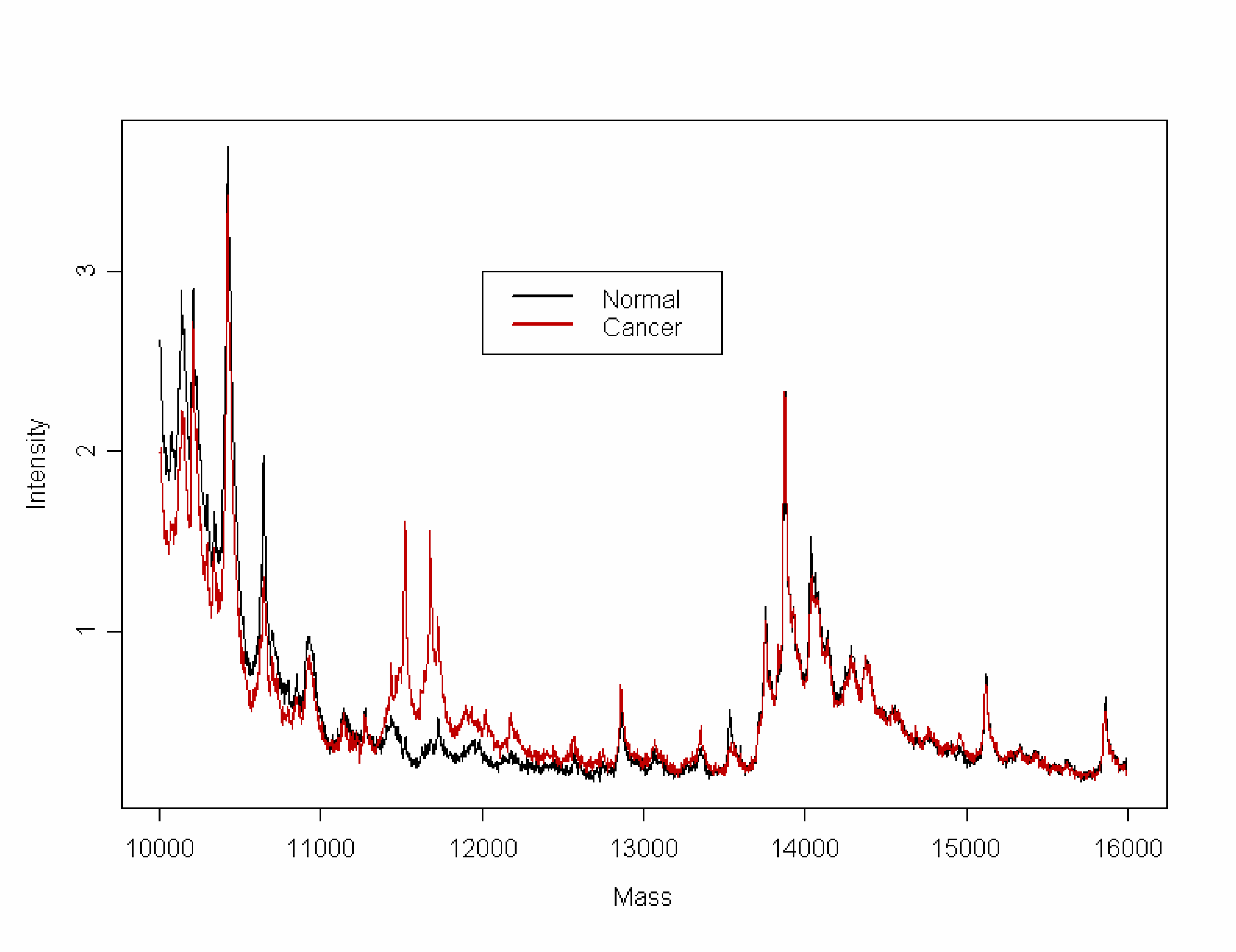
Using the mean spectrum

- We have started using the mean spectrum for peak finding
- Advantages:
  - Greater sensitivity, since noise should be reduced
  - Automatically accounts for minor calibration errors
  - Entirely avoids the problem of matching peaks across multiple spectra
  - Borrows strength across spectra, so it avoids ad hoc rules based on number of times a peak is seen with give signal-to-noise ratio.

Revised algorithm

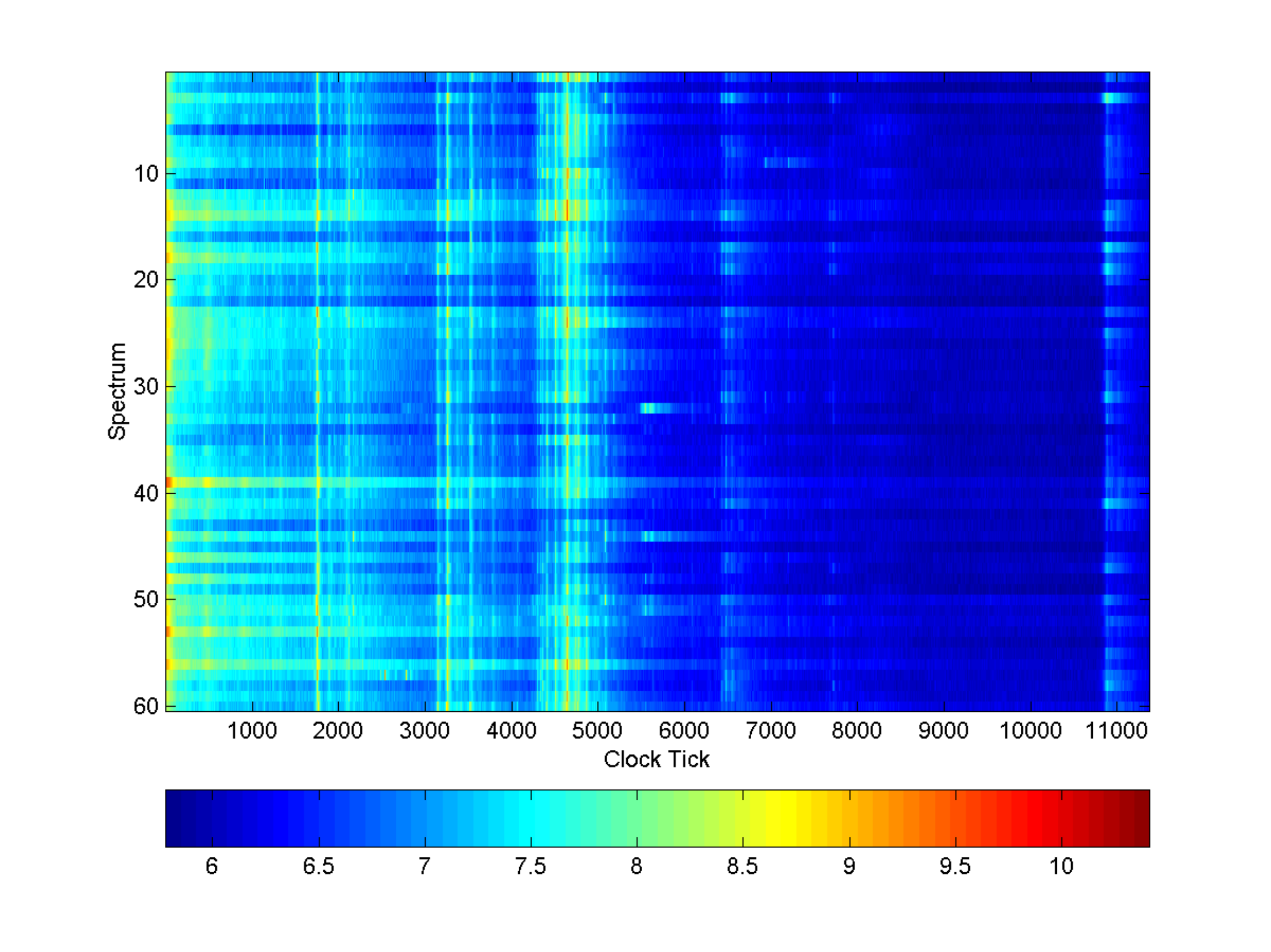
- Check that calibration is consistent
  - Interpolate to common time scale if needed
- Compute mean of raw spectra
- Apply wavelet method to denoise, baseline correct, and locate peaks in mean spectrum
- Quantify peaks in individual spectra
  - Apply wavelet method to denoise and baseline correct
  - Normalize by total ion current
  - Quantify by height (maximum) or area (sum)

Differences may become obvious when using mean spectra



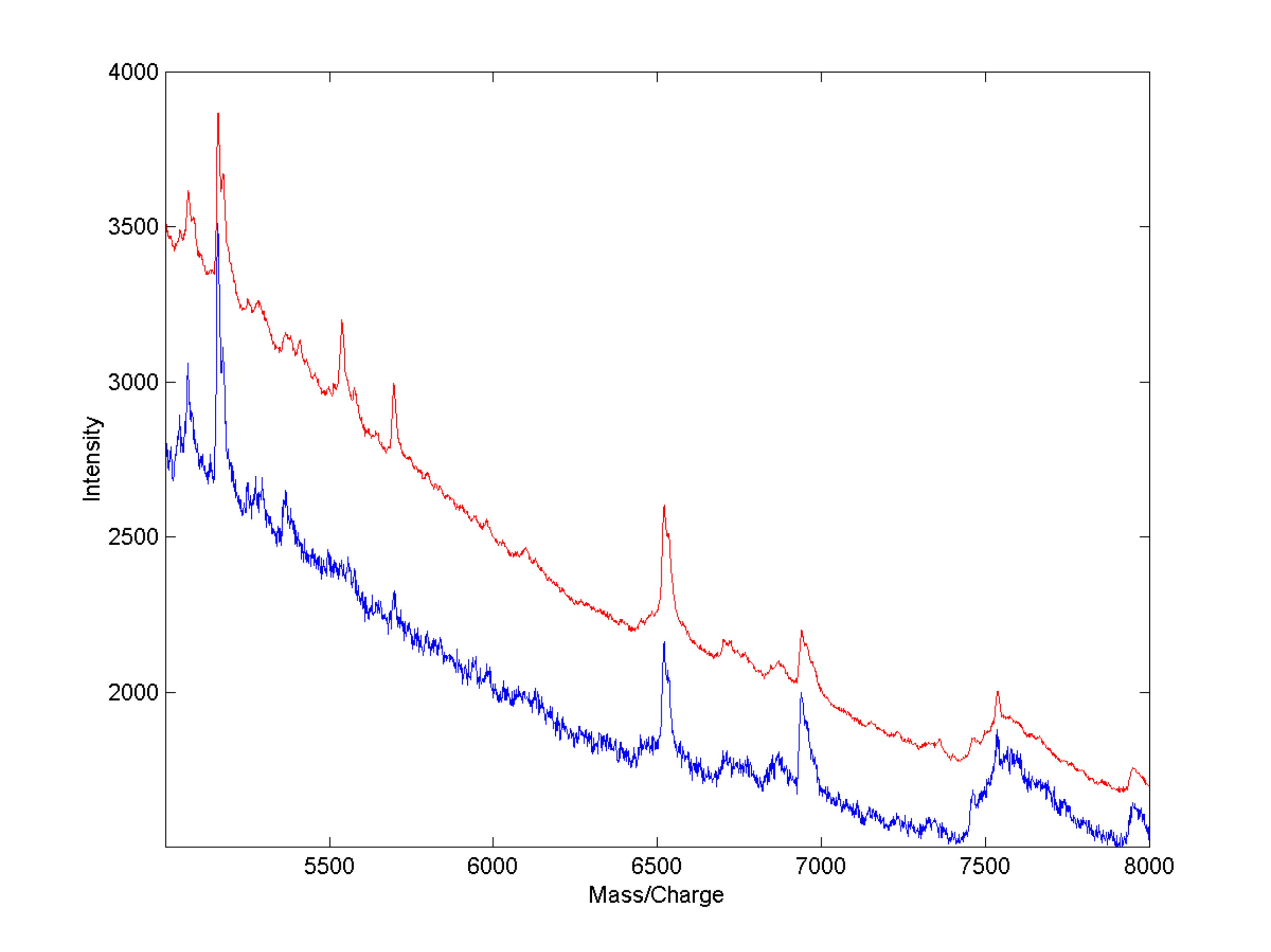
Pancreatic MALDI data from MDACC

Need to check approximate alignment across spectra before computing mean

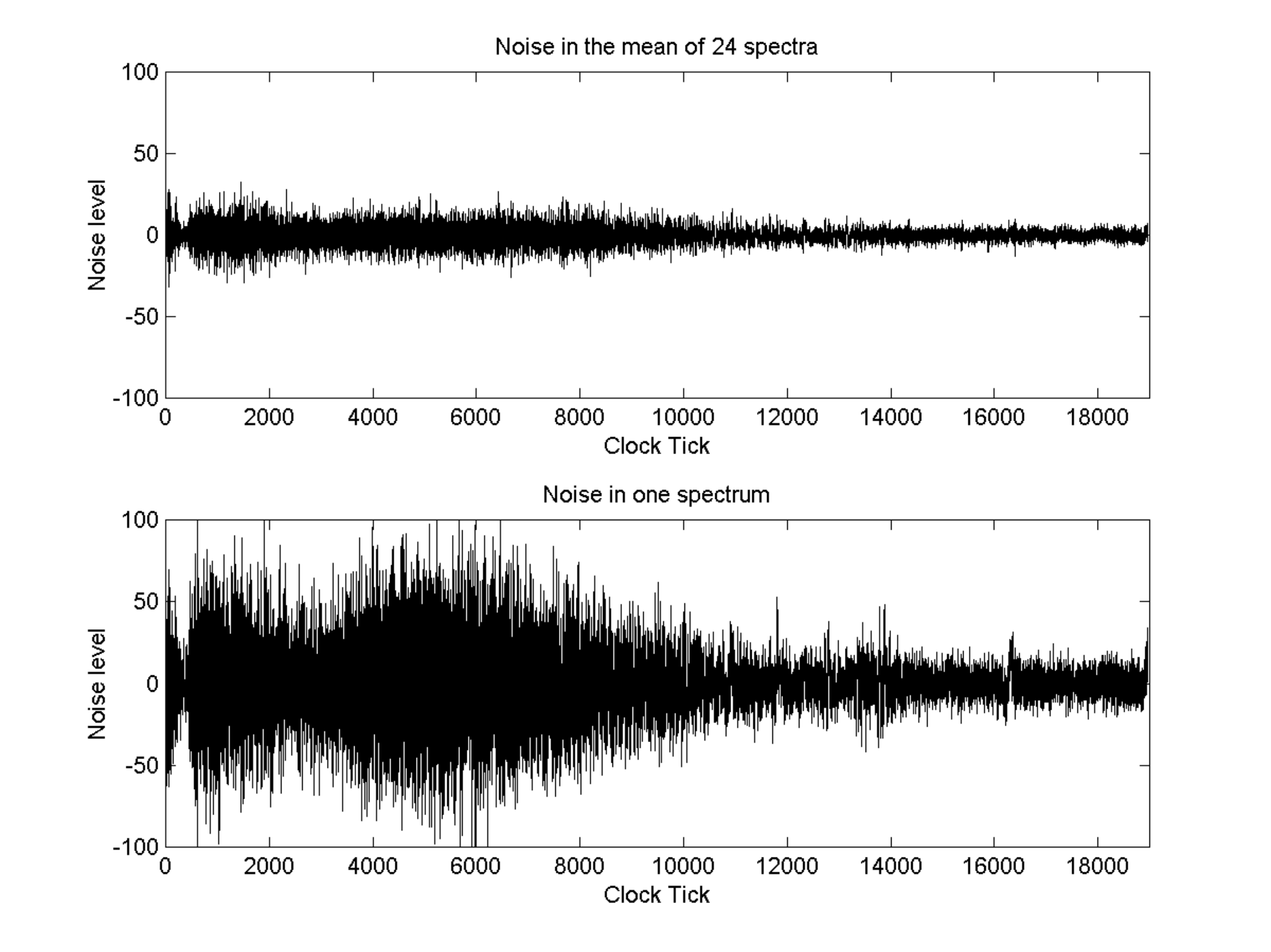


Pancreatic MALDI data from MDACC

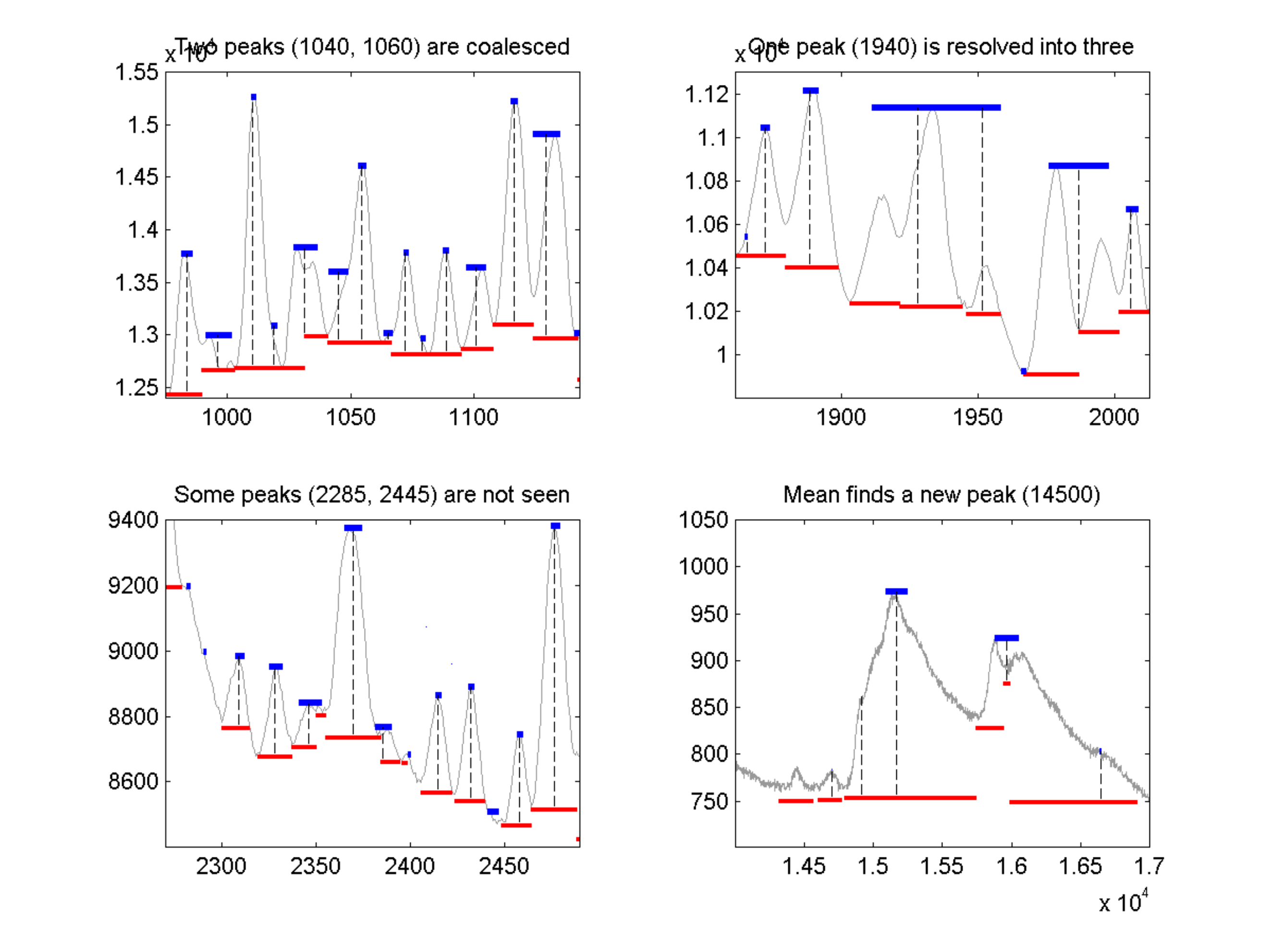
Noise goes down in the mean by the square root of the number of samples



Noise goes down in the mean by the square root of the number of samples



Peak matching and mean peak finding give different results



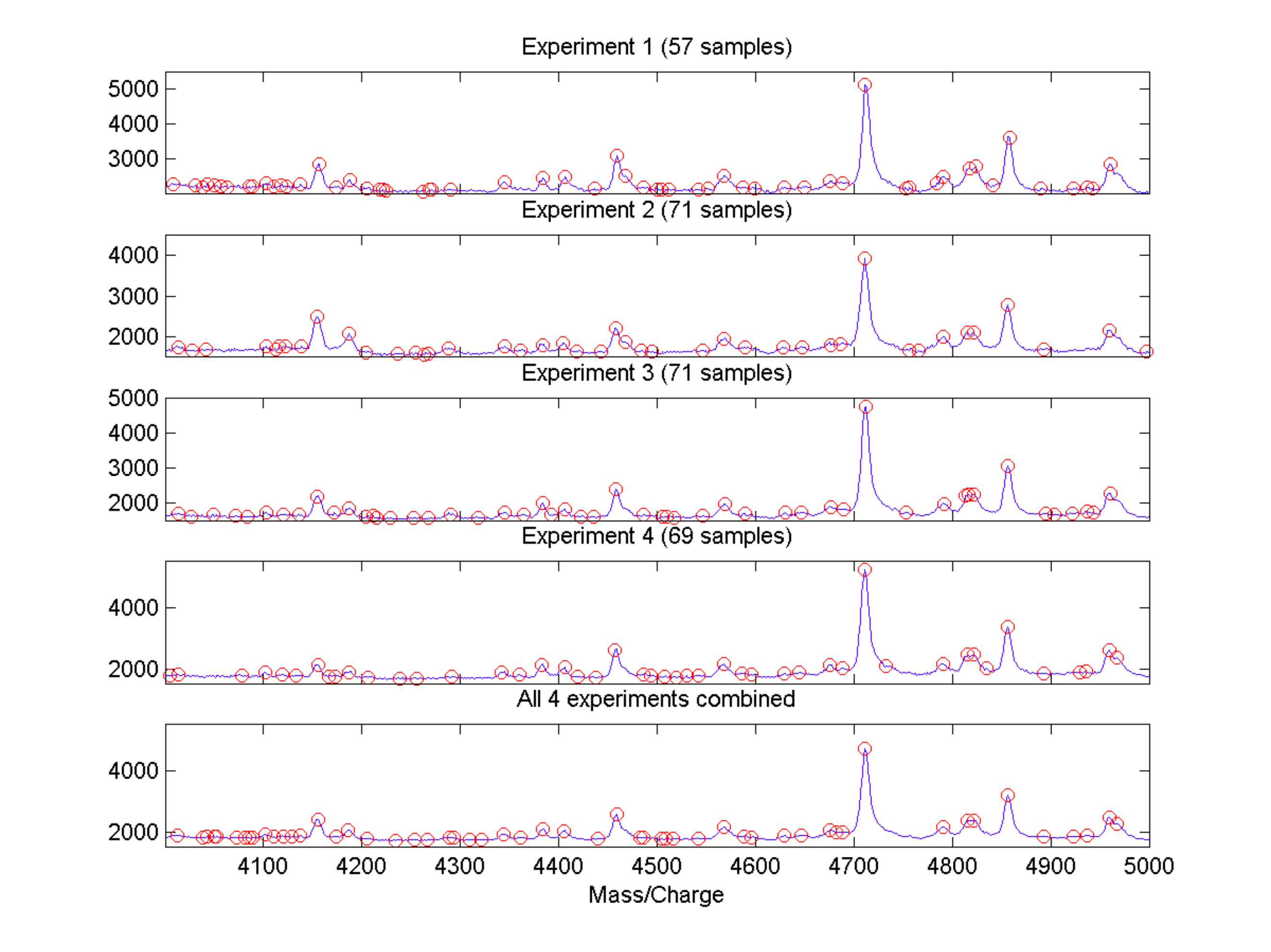
Two peaks (1040, 1060) are coalesced

One peak (1940) is resolved into three

Some peaks (2285, 2445) are not seen

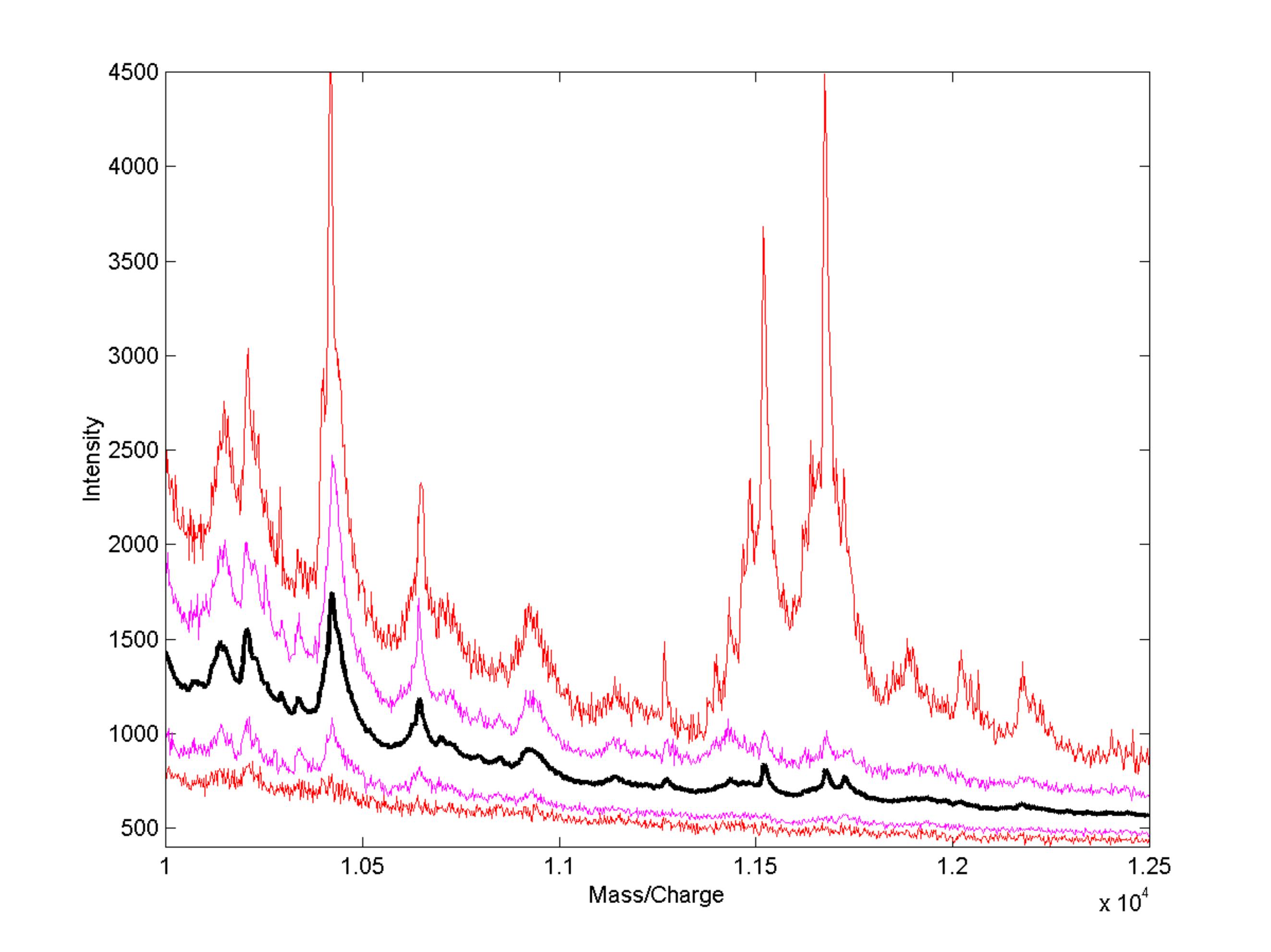
Mean finds a new peak (14500)

Mean peak finding is consistent across batches of spectra



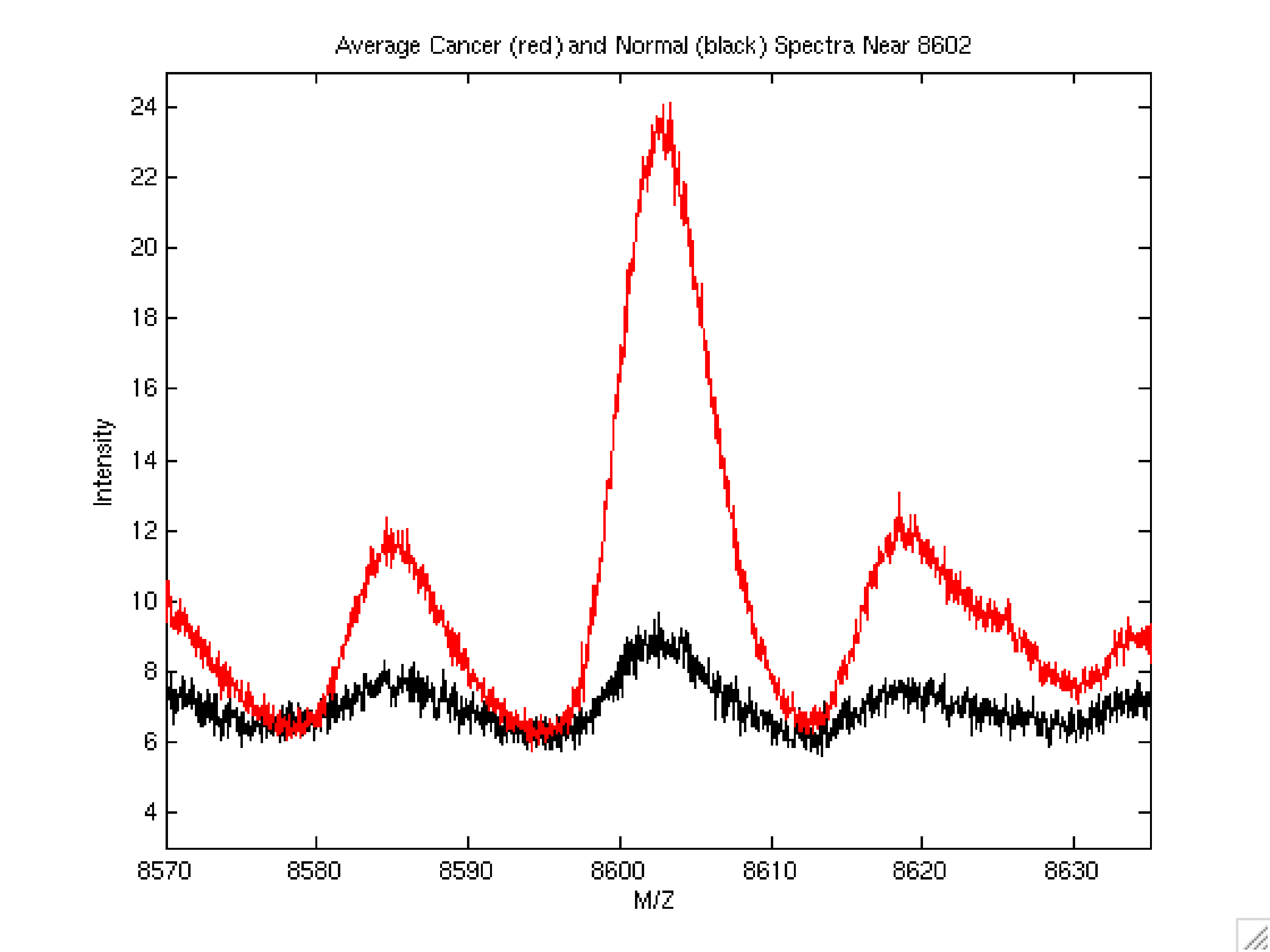
Pancreatic MALDI data from MDACC

The mean spectrum finds peaks that are only present in a few samples



Pancreatic MALDI data from MDACC

Consistent peaks with small S/N in individual spectra show up in the mean

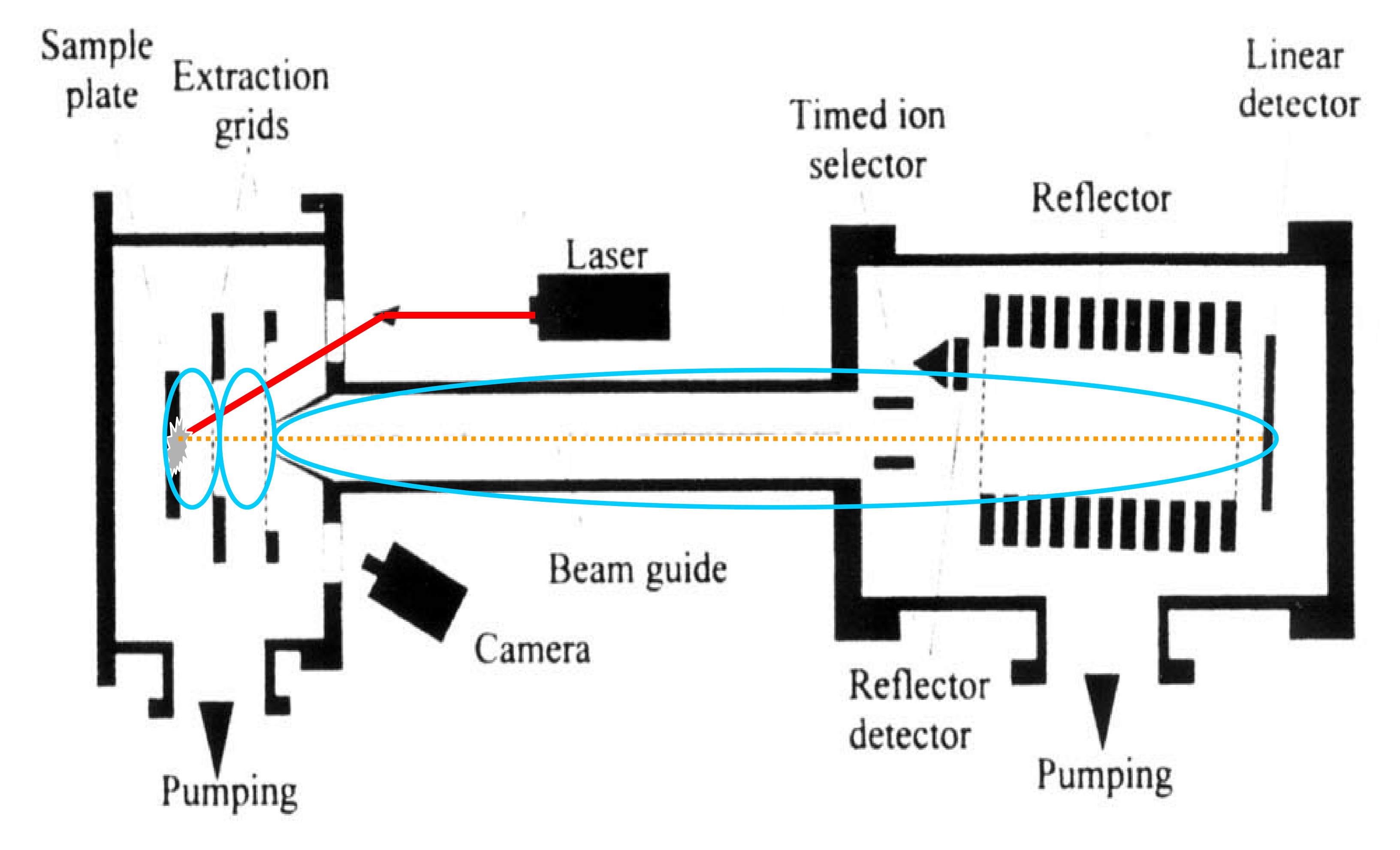


Ovarian Q-Star data from Conrad et al

Simulated spectra

- Difficult to evaluate processing methods on real data since we don’t know “truth”
- Have developed a simulation engine to produce realistic spectra
  - Based on the physics of a linear MALDI-TOF with ion focus delay
  - Flexible incorporation of different noise models and different baseline models
  - Includes isotope distributions
  - Can include matrix adducts, other modifications

MALDI-TOF schematic



Vestal and Juhasz. *J. Am. Soc. Mass Spectrom*. **1998**, *9*, 892.

Modeling the physics of MALDI-TOF

- Parameters
  - D_1 = distance from sample plate to first grid (8 mm)
  - V_1 = voltage for focusing (2000 V)
  - D_2 = distance between grids (17 mm)
  - V_2 = voltage for acceleration(20000 V)
  - L = length of tube (1 m)
  - v_0 = initial velocity ~ N(μ,σ)
  - v_1 = velocity after focusing
  - δ= delay time

- Equations

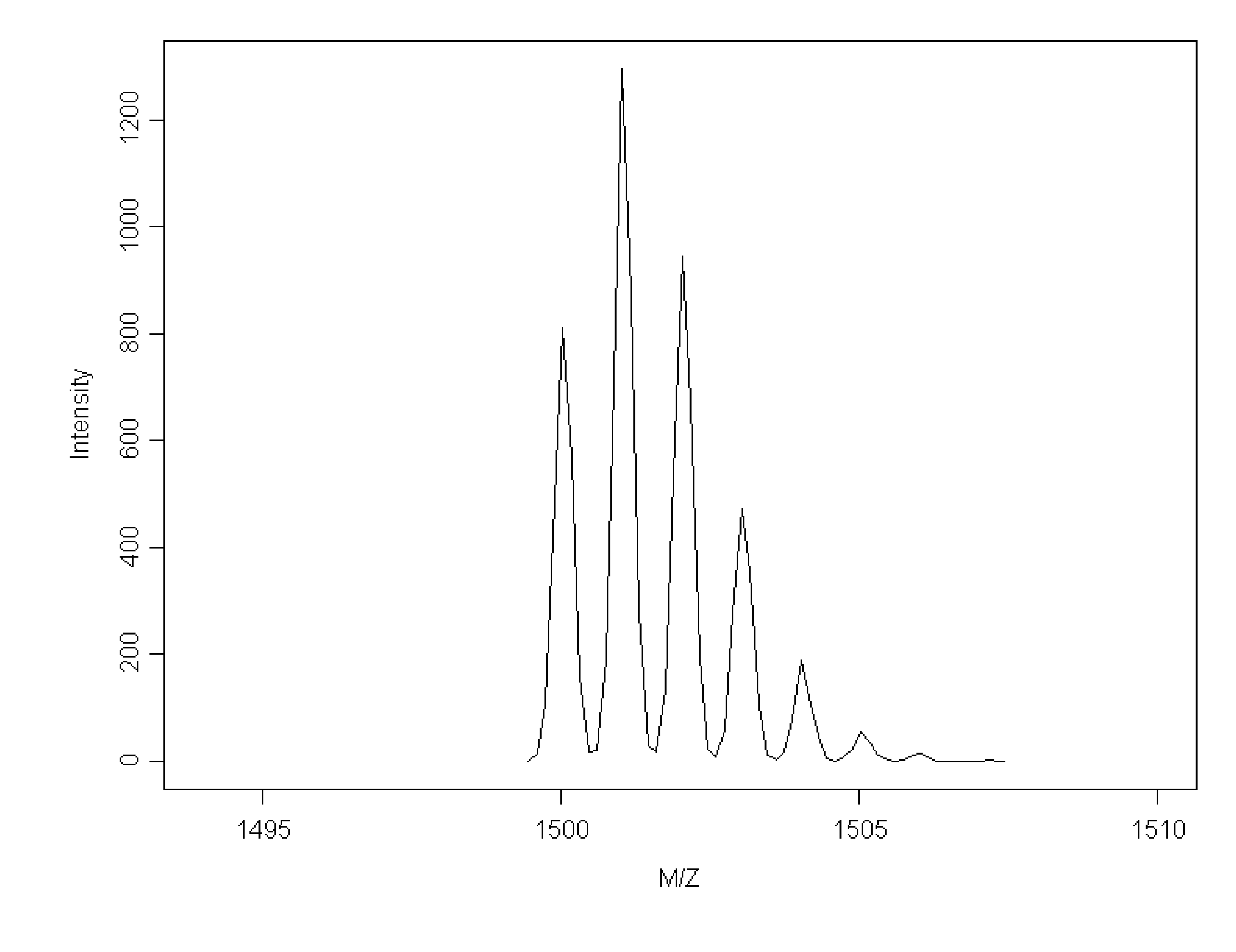
$$v_1^2 = v_0^2 + \frac{2qV_1}{mD_1}(D_1 - \delta v_0)$$

$$t_{DRIFT}^2 = L^2 / \left(\frac{2qV_2}{m} + v_1^2 \right)$$

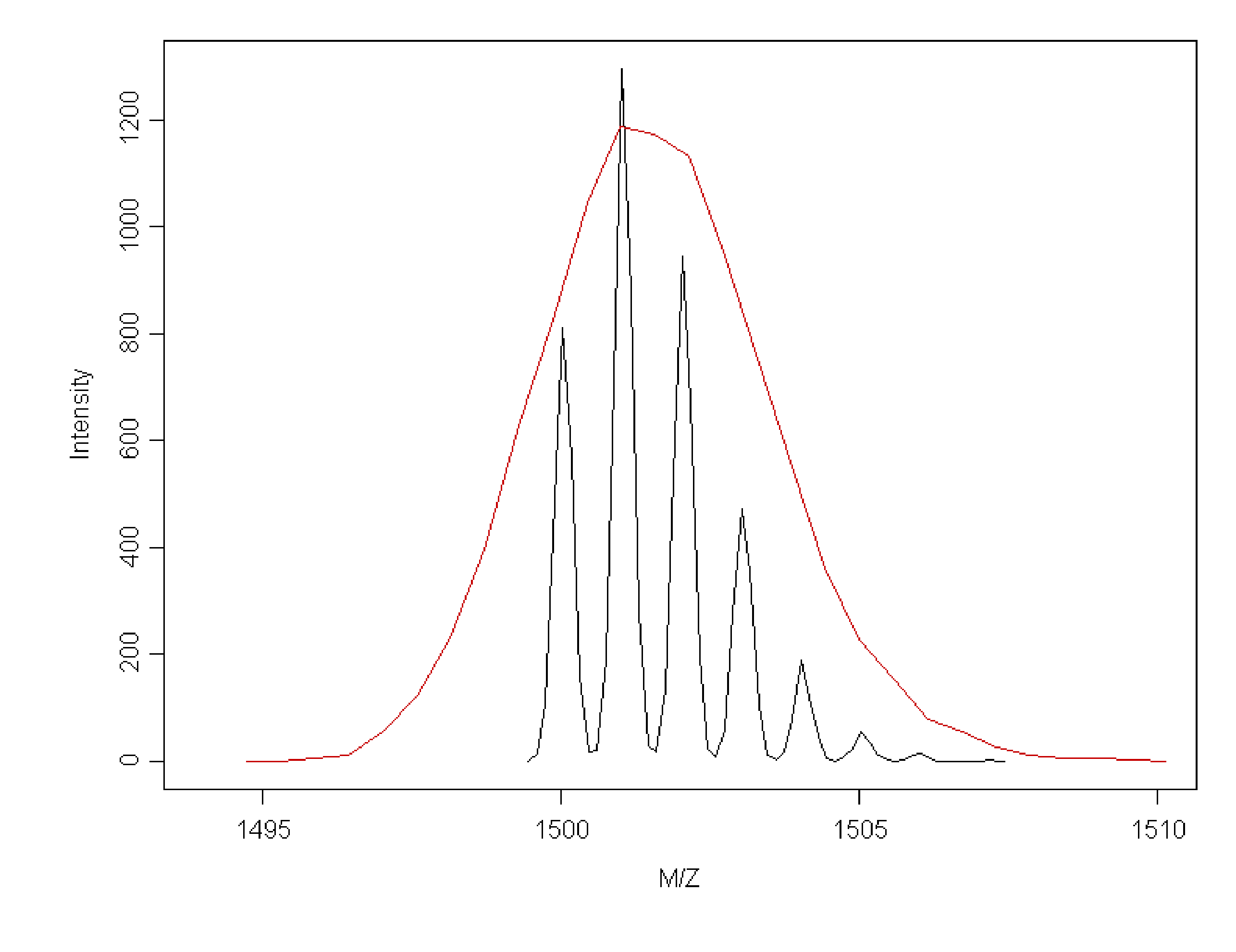
$$t_{ACCEL} = \frac{mD_2}{qV_2} \left(\frac{L}{t_{DRIFT}} - v_1 \right)$$

$$t_{FOCUS} = \frac{mD_1}{qV_1}(v_1 - v_0)$$

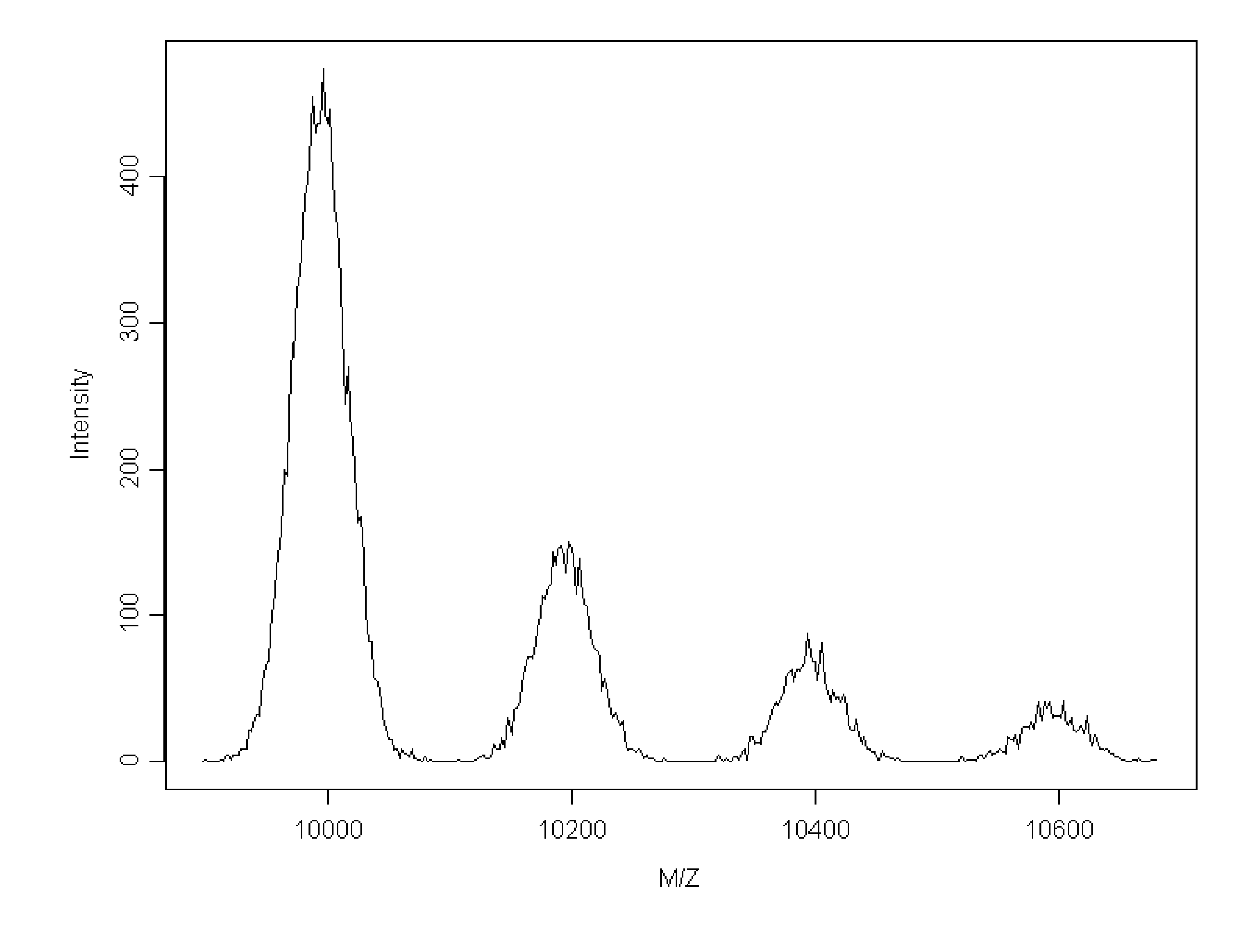
Simulation of one protein, with isotope distribution



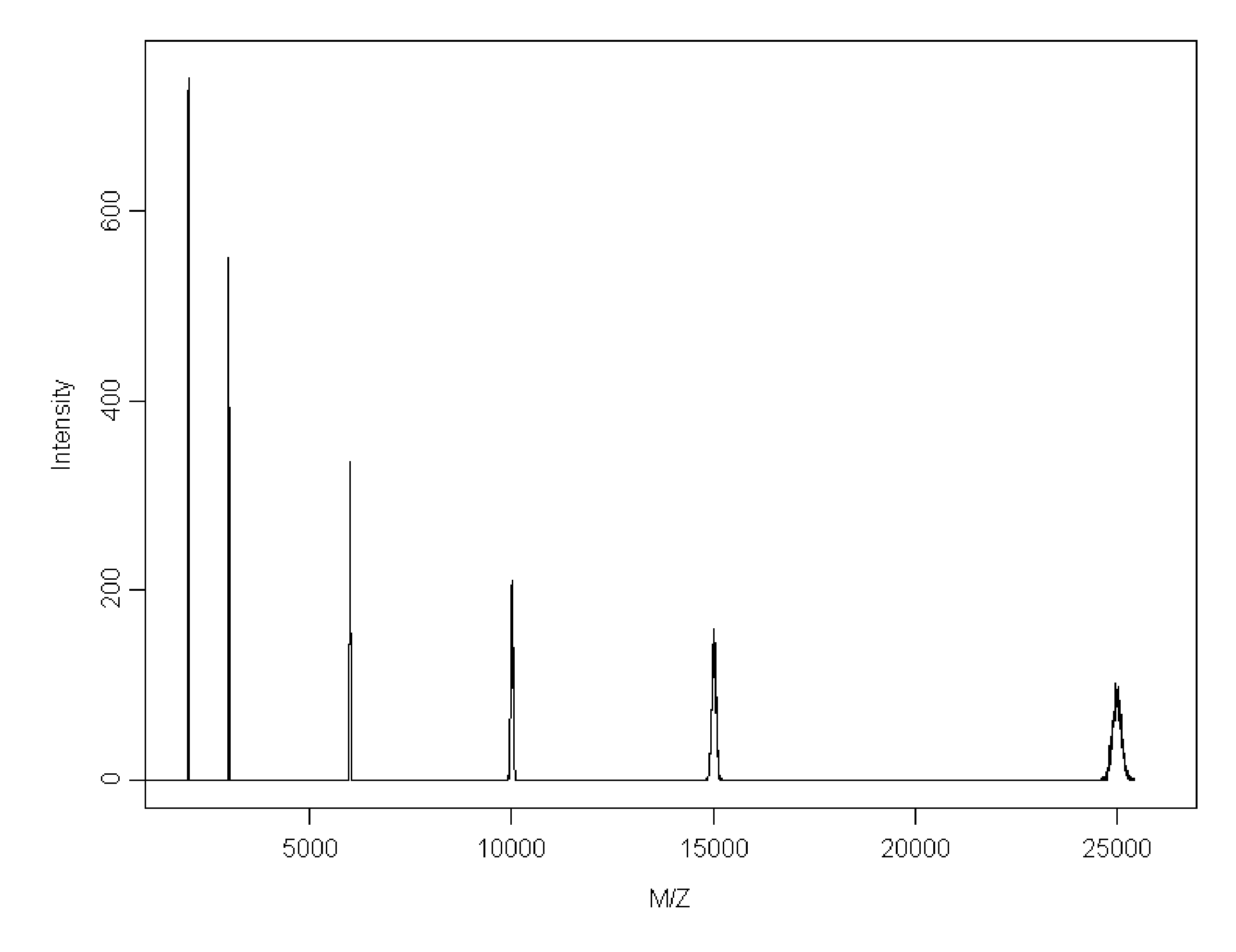
Overlay of the same protein simulated on a low resolution instrument



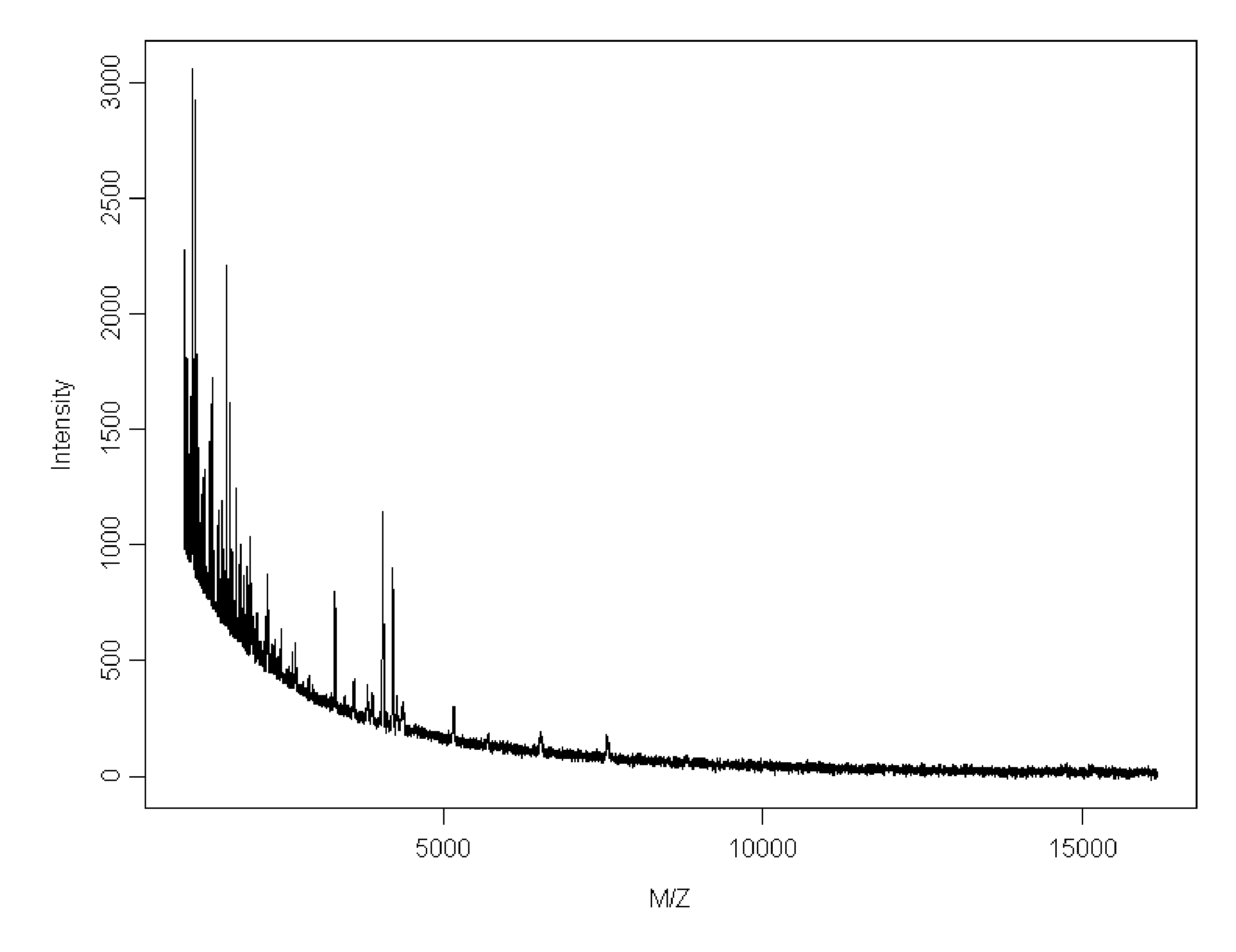
Simulation of one protein with decreasing numbers of matrix adducts



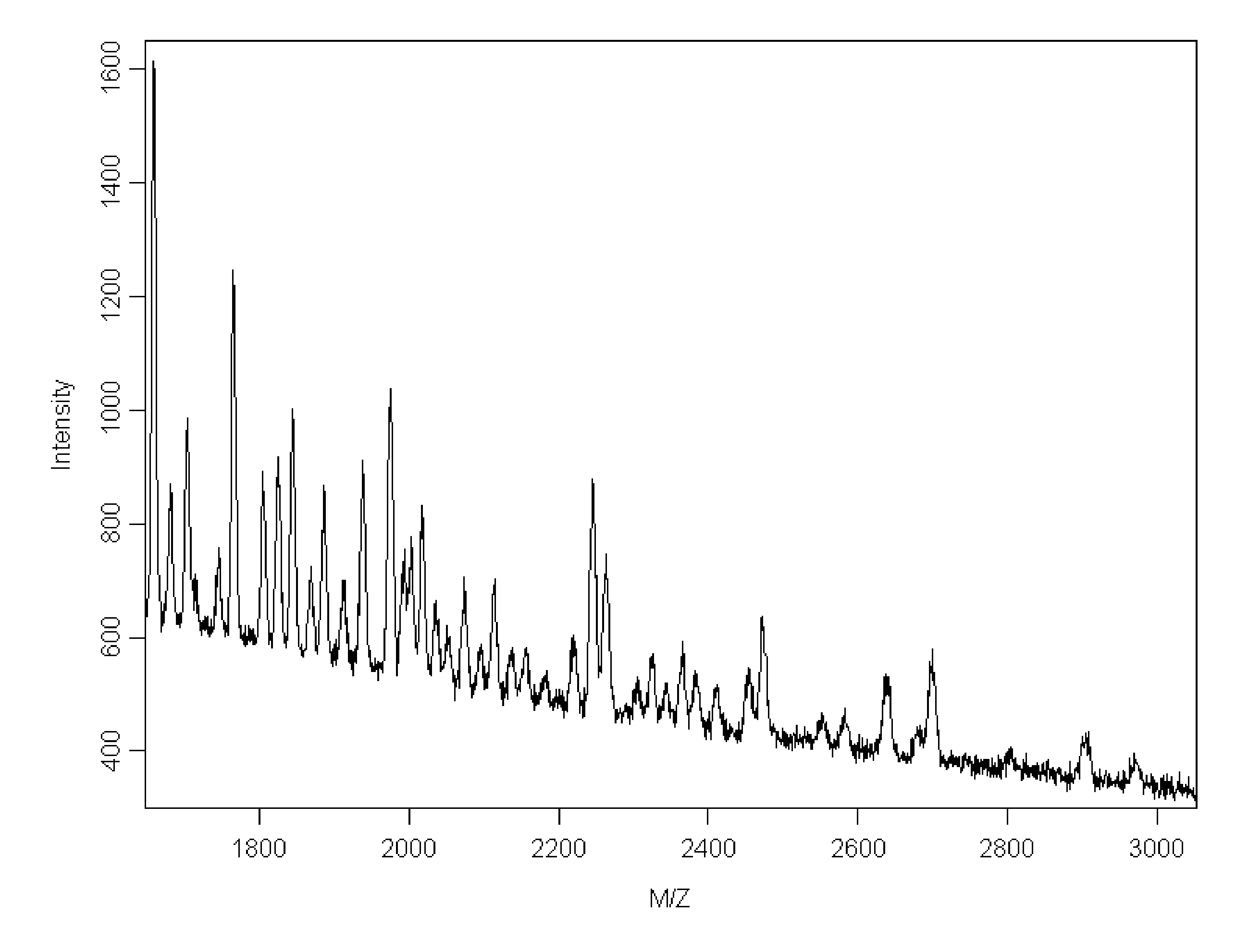
Simulated calibration spectrum with equal amounts of six proteins



Simulated spectrum with a complex mixture of proteins



Closeup of simulated complex spectrum



Open problems

- Better calibration?
  - Internal validation
- Better baseline correction?
- Alternative methods for normalization?
- Best method for quantification?
- Best statistical methods to use after done with preprocessing?
- Quality control/quality assurance?
- Ways to exploit simulations to test new methods?

Acknowledgements

- Bioinformatics
  - Keith Baggerly
  - Jeffrey Morris
  - Jing Wang
  - Lianchun Xiao
  - Spyros Tsavachidis
  - Thomas Liu
- Proteomics (MDACC)
  - Ryuji Kobayashi
  - David Hawke
  - John Koomen
- Ciphergen
  - Charlotte Clarke
- Biologists (MDACC)
  - Jim Abbruzzese
  - I.J. Fidler
  - Stan Hamilton
  - Nancy Shih
  - Ken Aldape
  - Henry Kuerer
  - Herb Fritsche
  - Gordon Mills
  - Lajos Pusztai
  - Jack Roth
  - Lin Ji