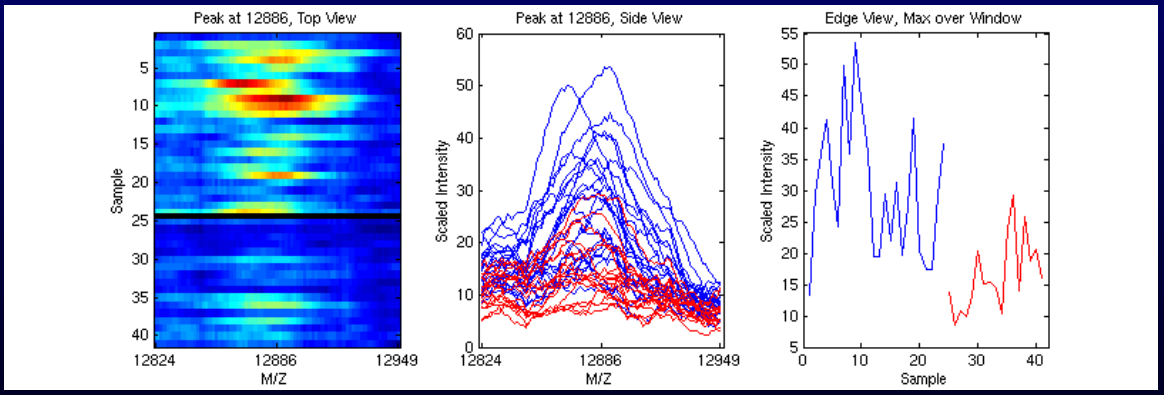
The Analysis of Proteomics Spectra from Serum Samples

Jeffrey S. Morris

Department of Biostatistics
MD Anderson Cancer Center



What Are Proteomics Spectra?

DNA makes RNA makes Protein

What Are Proteomics Spectra?

DNA makes RNA makes Protein

Microarrays allow us to measure the mRNA complement of a set of cells

What Are Proteomics Spectra?

DNA makes RNA makes Protein

Microarrays allow us to measure the mRNA complement of a set of cells

Mass spectrometry allows us to measure the protein complement (or subset thereof) of a set of cells

What Are Proteomics Spectra?

DNA makes RNA makes Protein

Microarrays allow us to measure the mRNA complement of a set of cells

Mass spectrometry allows us to measure the protein complement (or subset thereof) of a set of cells

Proteomics spectra are mass spectrometry traces of biological specimens

What Are Proteomics Spectra?

DNA makes RNA makes Protein

Microarrays allow us to measure the mRNA complement of a set of cells

Mass spectrometry allows us to measure the protein complement (or subset thereof) of a set of cells

Proteomics spectra are mass spectrometry traces of biological specimens

Why Are We Excited?

Profiles can be assessed using less invasive samples (serum, urine, nipple aspirate fluid) rather than tissue biopsies

Why Are We Excited?

Profiles can be assessed using less invasive samples (serum, urine, nipple aspirate fluid) rather than tissue biopsies

Spectra are cheaper to run on a per unit basis than microarrays

Why Are We Excited?

Profiles can be assessed using less invasive samples (serum, urine, nipple aspirate fluid) rather than tissue biopsies

Spectra are cheaper to run on a per unit basis than microarrays

Can run samples on large numbers of patients

Why Are We Excited?

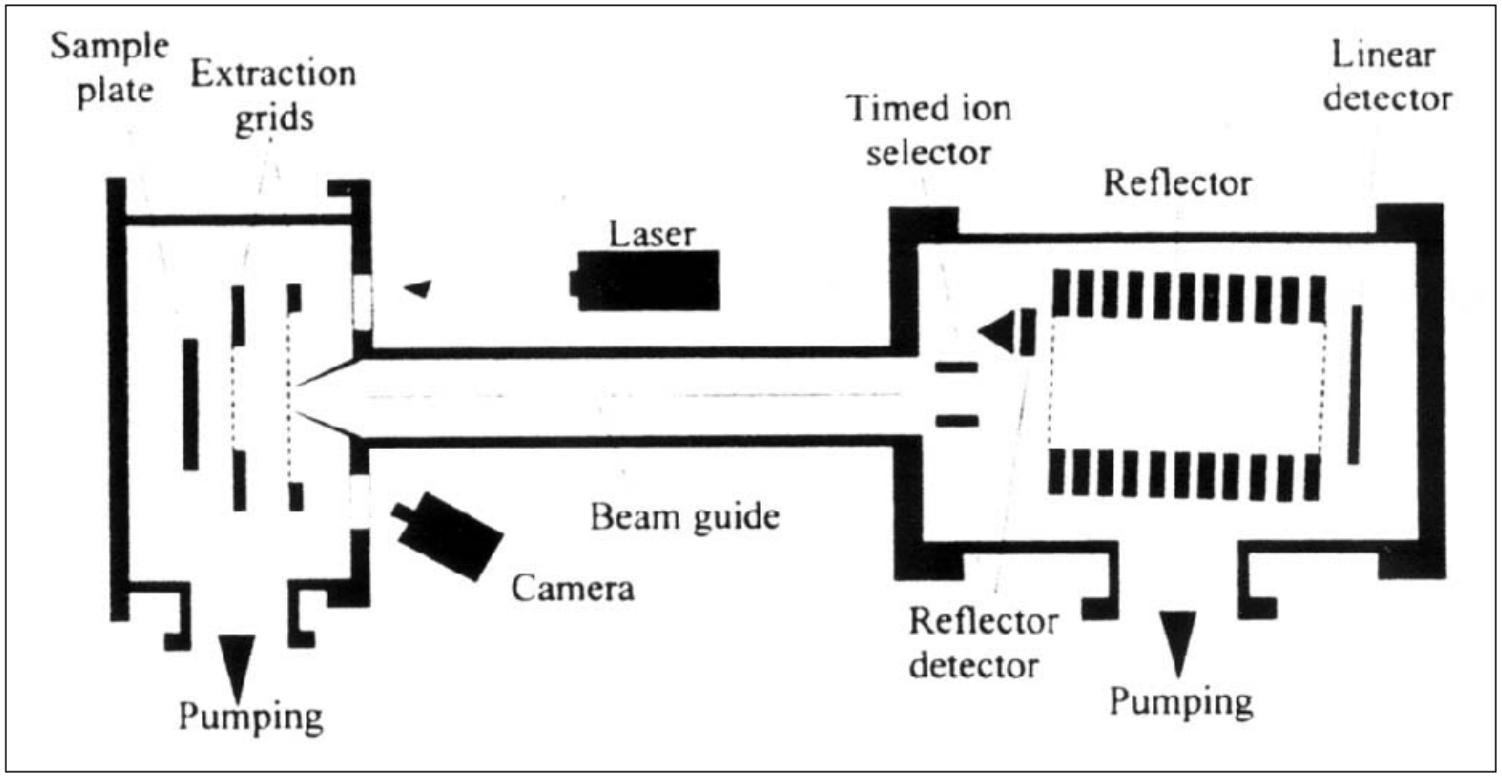
Profiles can be assessed using less invasive samples (serum, urine, nipple aspirate fluid) rather than tissue biopsies

Spectra are cheaper to run on a per unit basis than microarrays

Can run samples on large numbers of patients

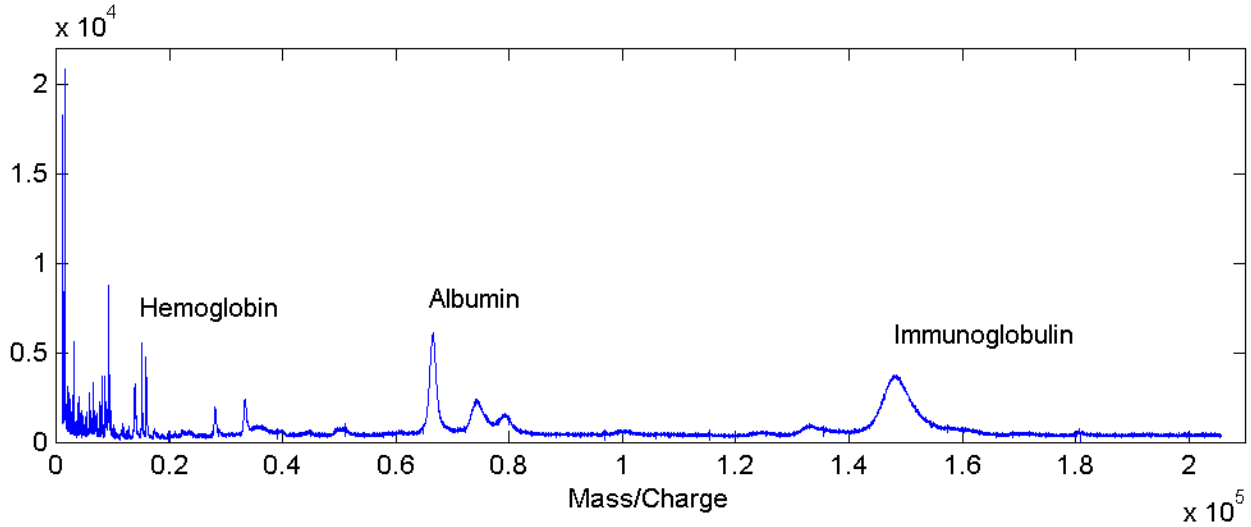
How Does Mass Spec Work?

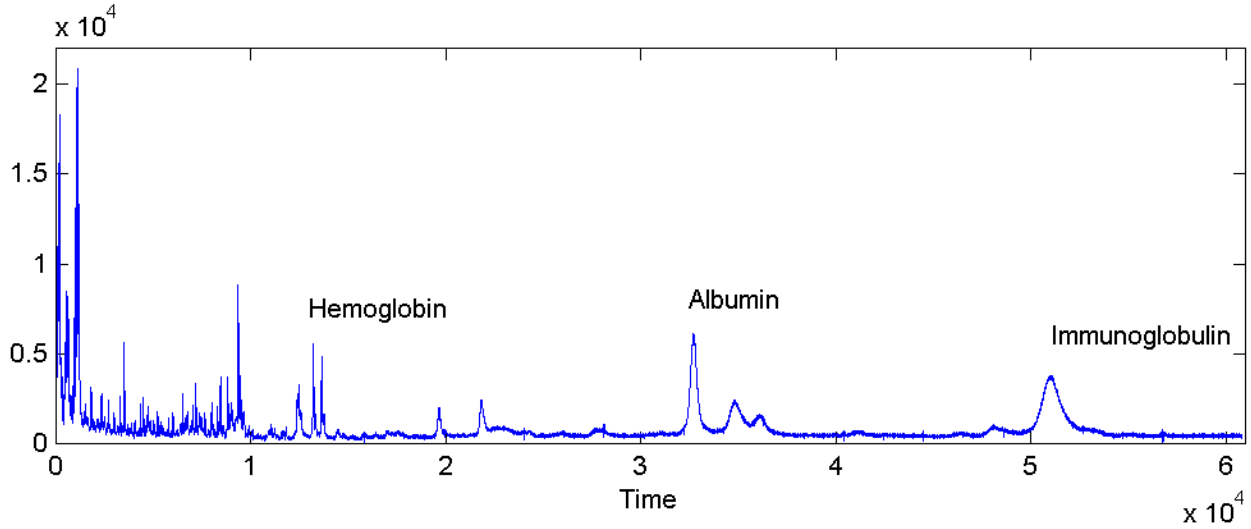
Block Diagram of a MALDI-TOF



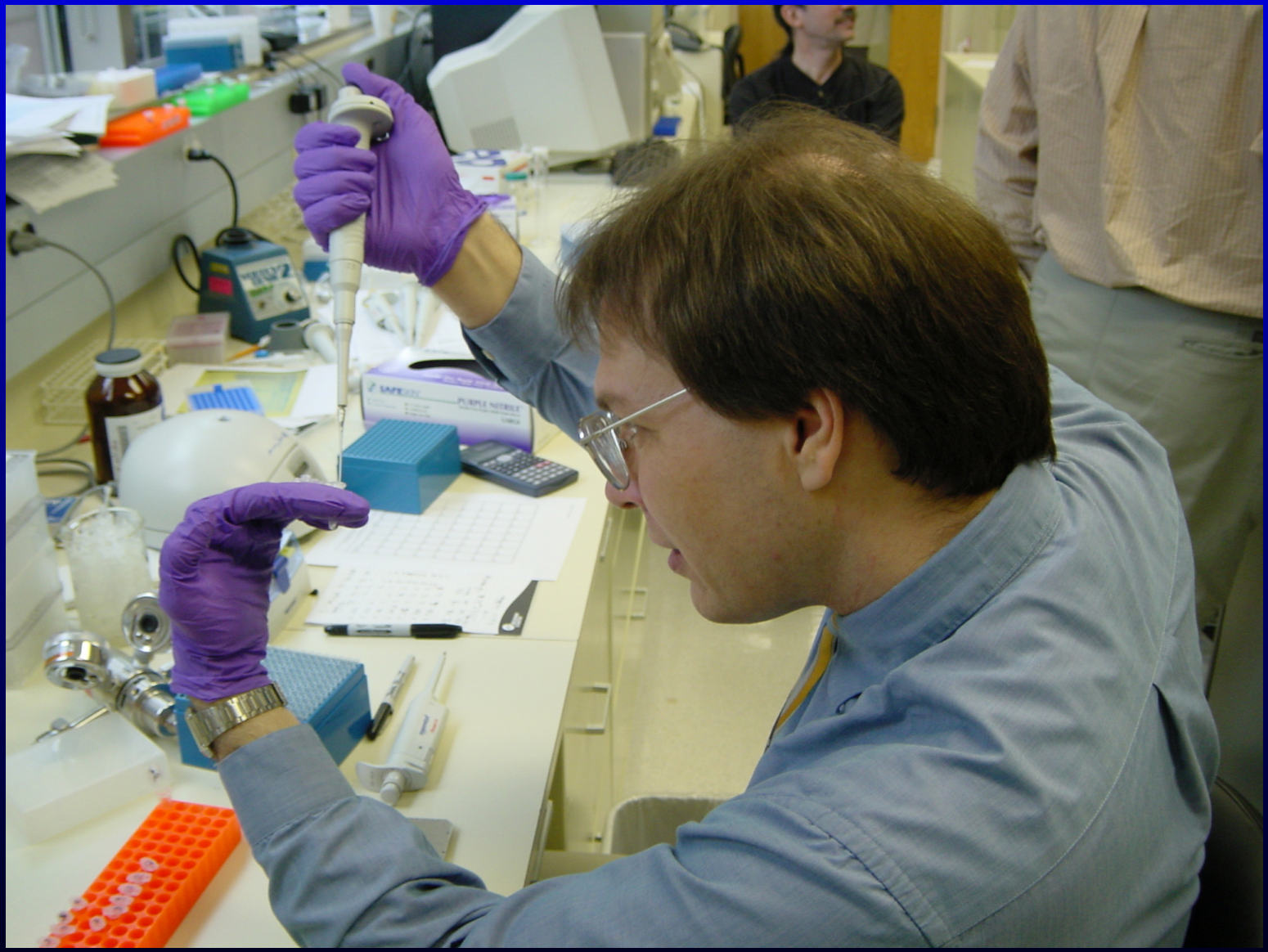
Vestal and Juhasz. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 892.

What Do the Data Look Like?



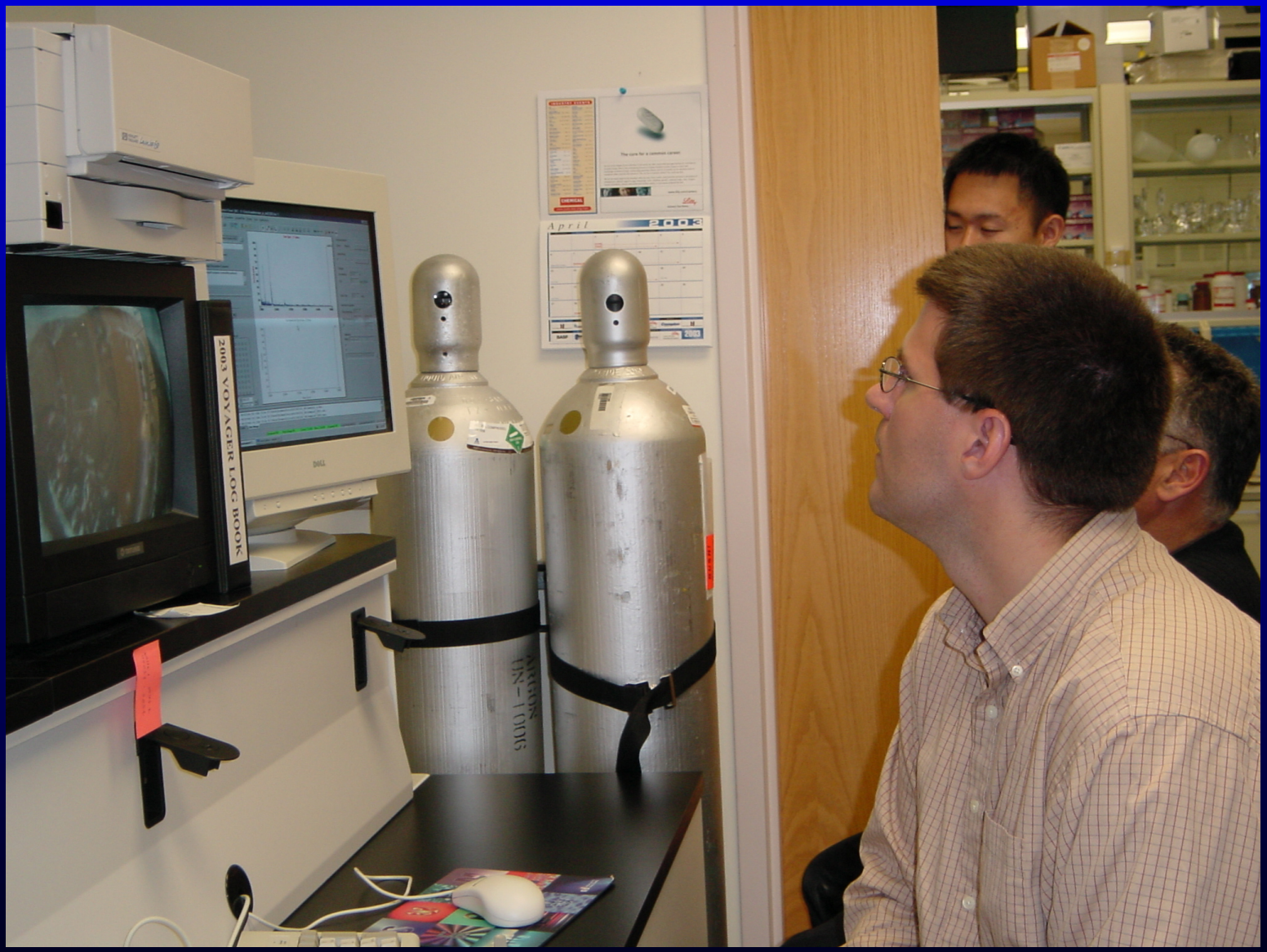


Learning: Spotting the Samples



What the Guts Look Like

Taking Data



Some Other Common Steps

Fractionating the Samples

Changing the Laser Intensity

Working with Different Matrix Substrates

SELDI: A Special Case

www.ciphergen.com

Precoated surface performs some preselection of the proteins for you.

Machines are nominally easier to use.

A Tale of Two Examples

Example 1 : Learning from the literature

Example 2 : Testing out our understanding

A story in pictures

Example 1: Feb 16 ’02 Lancet

MECHANISMS OF DISEASE

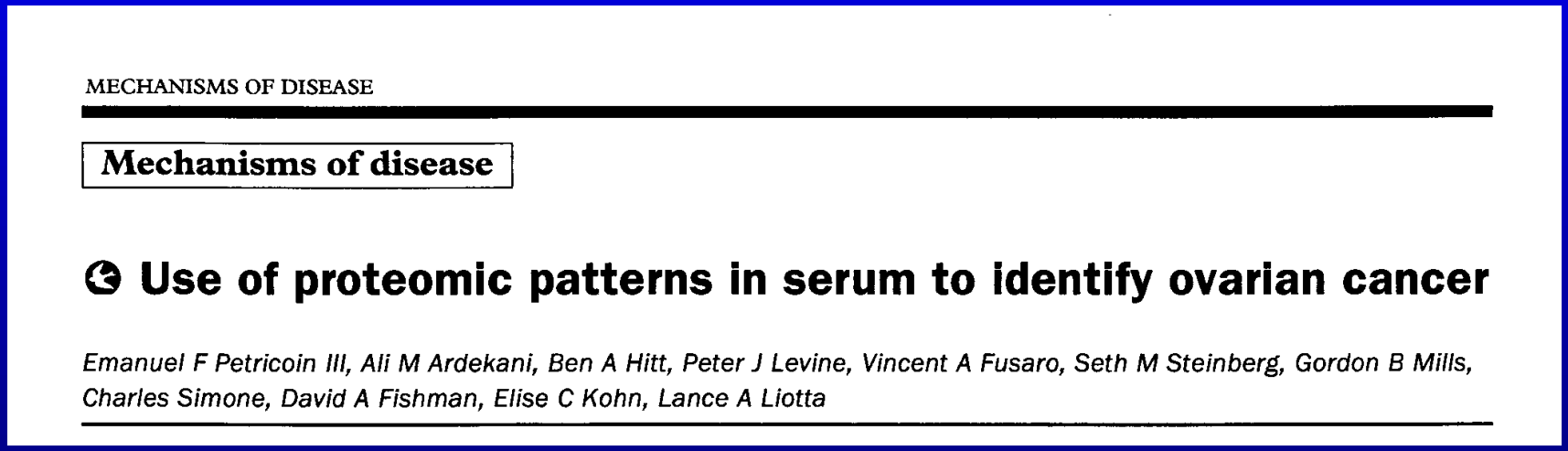
Mechanisms of disease

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

- 100 ovarian cancer patients
- 100 normal controls
- 16 patients with ’benign disease’

Example 1: Feb 16 ’02 Lancet



MECHANISMS OF DISEASE

Mechanisms of disease

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

- 100 ovarian cancer patients
- 100 normal controls
- 16 patients with ’benign disease’

Use 50 cancer and 50 normal spectra to train a classification method; test the algorithm on the remaining samples.

Their Results

- Correctly classified 50/50 of the ovarian cancer cases.
- Correctly classified 46/50 of the normal cases.
- Correctly classified 16/16 of the benign disease as ’other’.

Data at http://clinicalproteomics.steem.com

Large sample sizes, using serum

The Data Sets

3 data sets on ovarian cancer

Data Set 1 : The initial experiment. 216 samples, baseline subtracted, H4 chip

The Data Sets

3 data sets on ovarian cancer

Data Set 1 : The initial experiment. 216 samples, baseline subtracted, H4 chip

Data Set 2 : Followup: the same 216 samples, baseline subtracted, WCX2 chip

The Data Sets

3 data sets on ovarian cancer

Data Set 1 : The initial experiment. 216 samples, baseline subtracted, H4 chip

Data Set 2 : Followup: the same 216 samples, baseline subtracted, WCX2 chip

Data Set 3 : New experiment: 162 cancers, 91 normals, baseline NOT subtracted, WCX2 chip

The Data Sets

3 data sets on ovarian cancer

Data Set 1 : The initial experiment. 216 samples, baseline subtracted, H4 chip

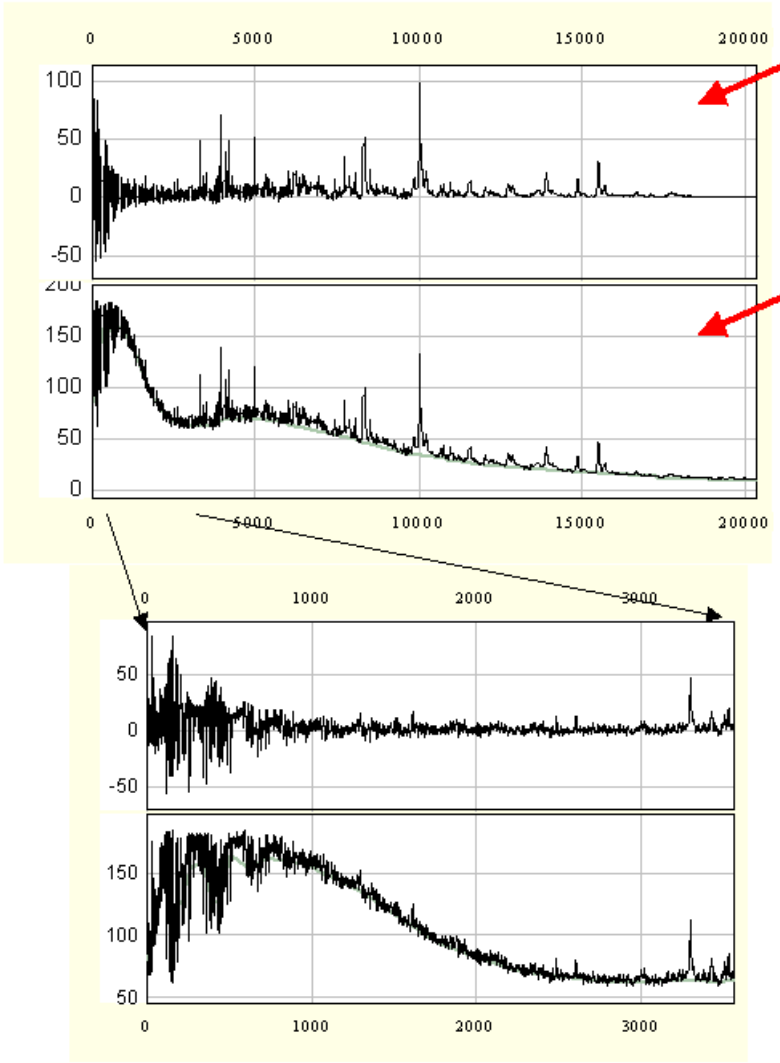
Data Set 2 : Followup: the same 216 samples, baseline subtracted, WCX2 chip

Data Set 3 : New experiment: 162 cancers, 91 normals, baseline NOT subtracted, WCX2 chip

A set of 5-7 separating peaks is supplied for each data set.

We tried to (a) replicate their results, and (b) check consistency of the proteins found

We Can't Replicate their Results (DS1 & DS2)



They posted this

Baseline Subtraction ON

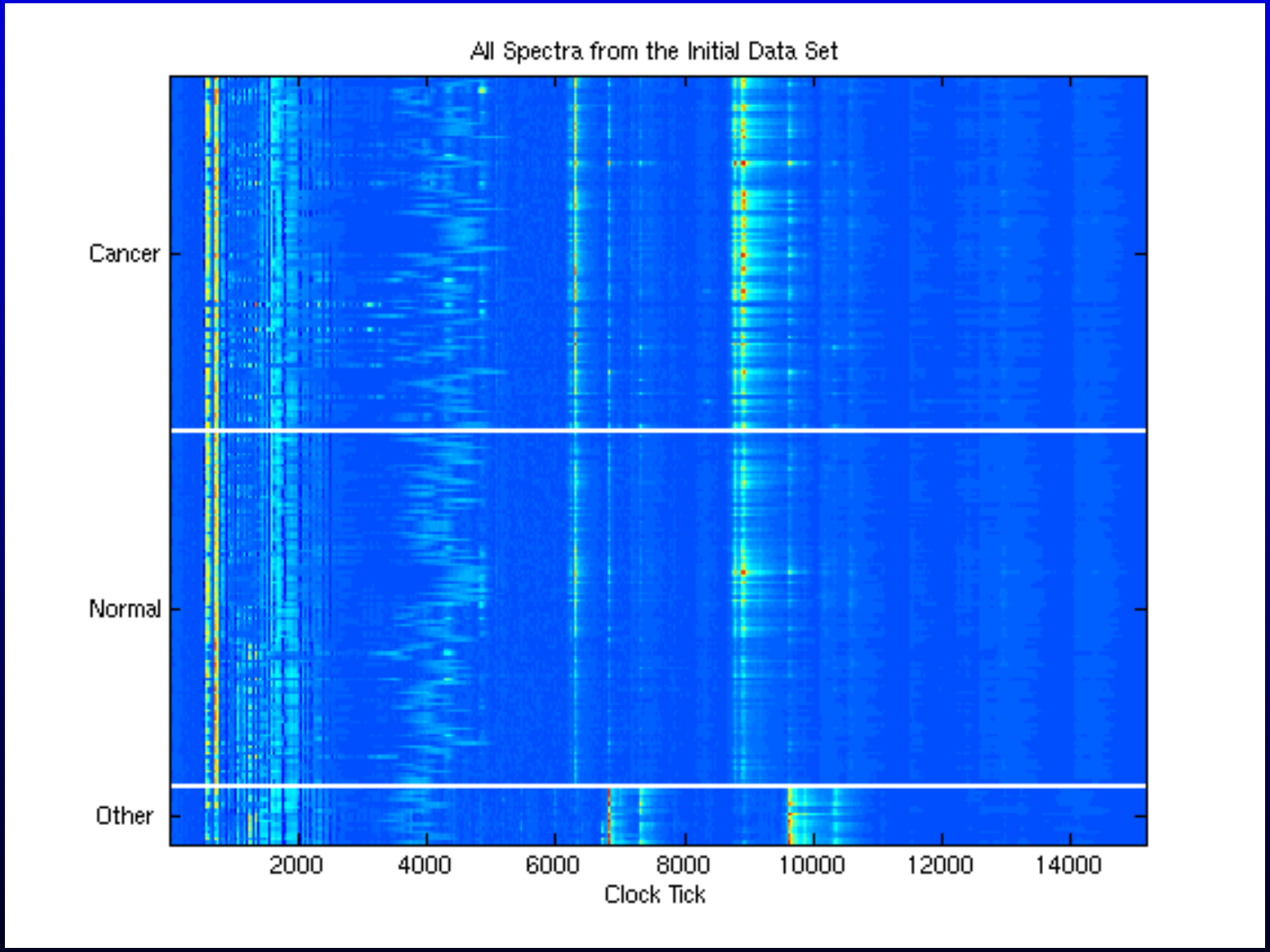
They analyzed this

Baseline Subtraction OFF

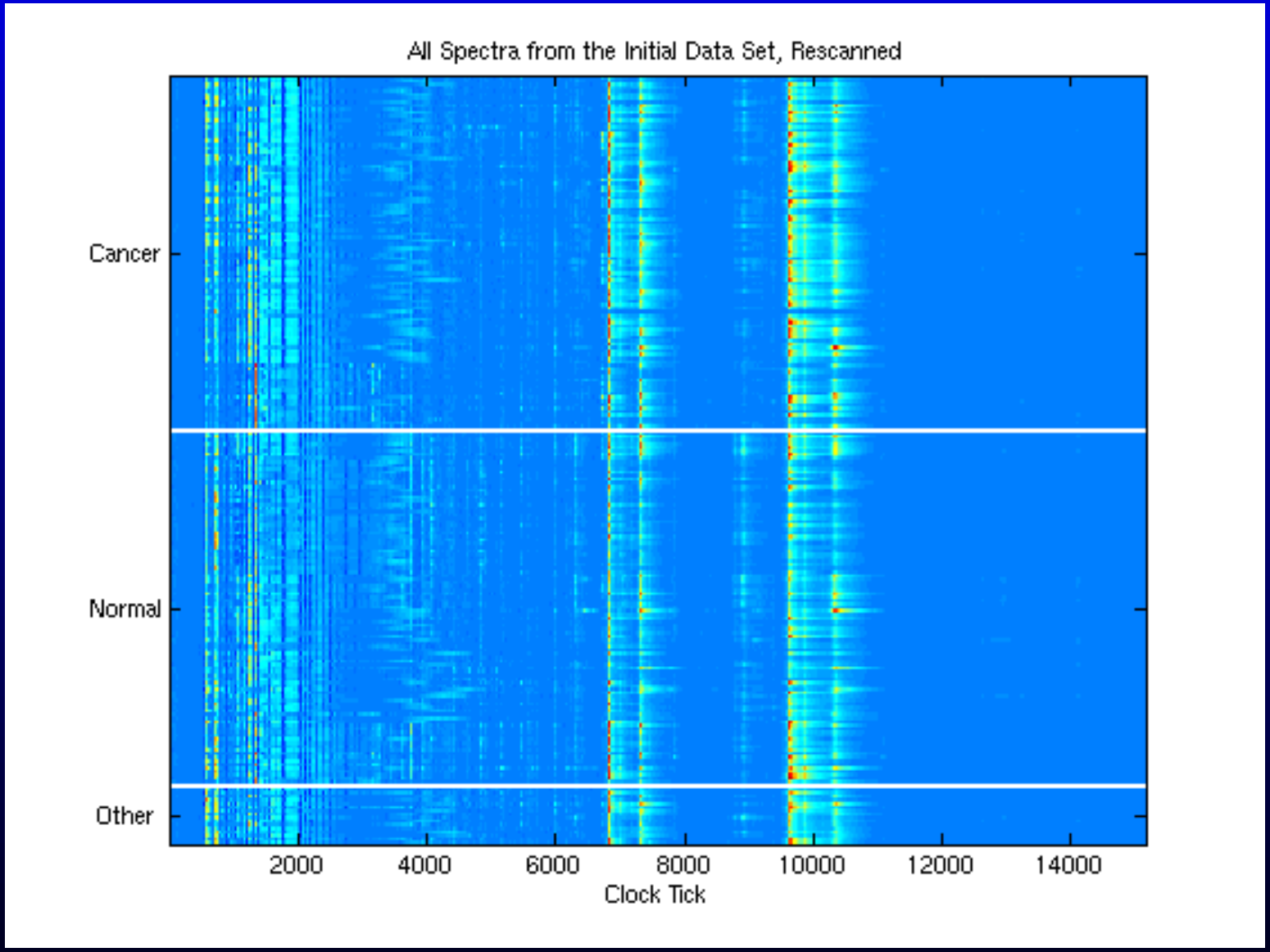
Zoomed in with Baseline Sub ON

Zoomed in with Baseline Sub OFF

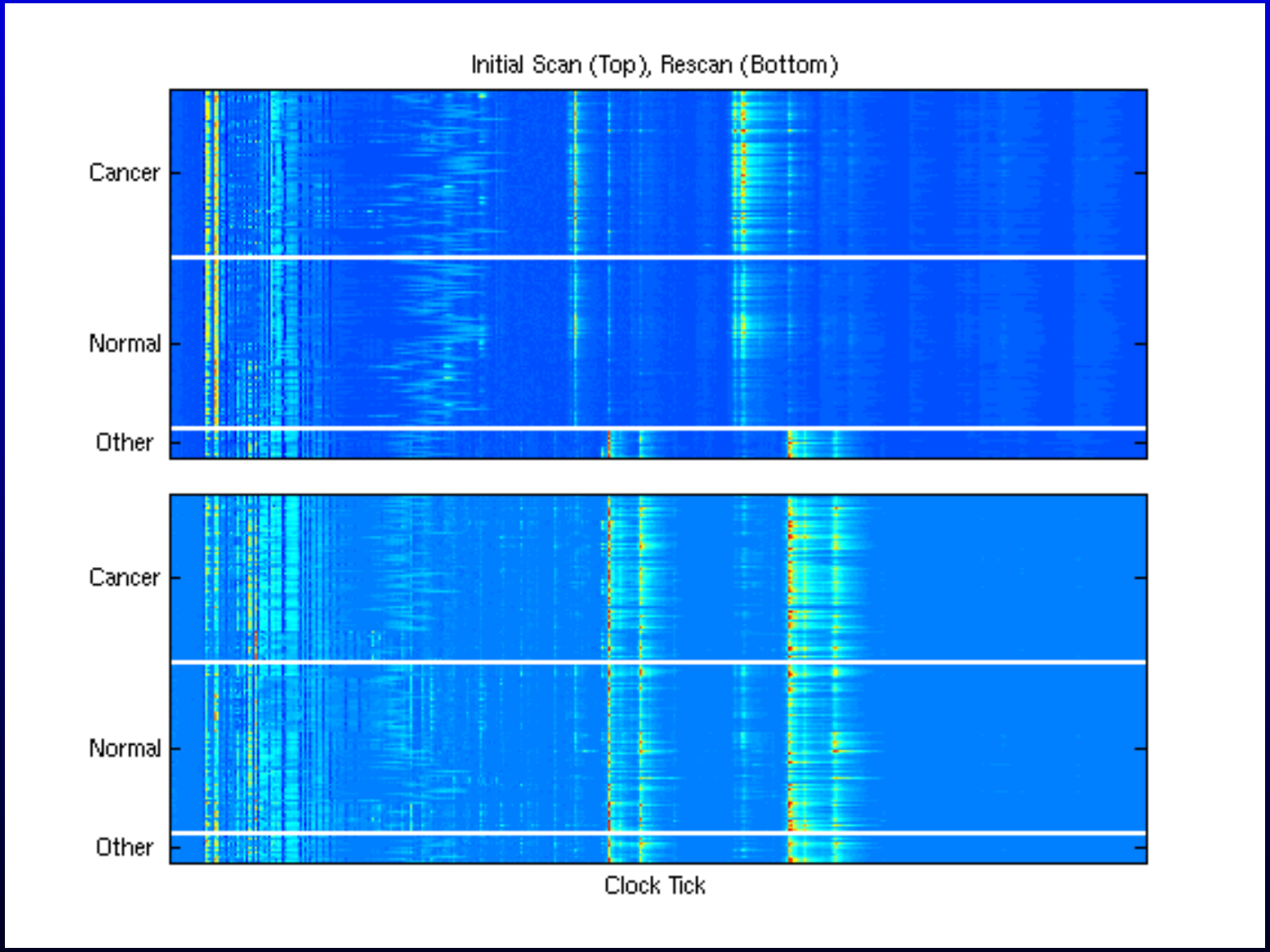
Some Structure is Visible in DS1



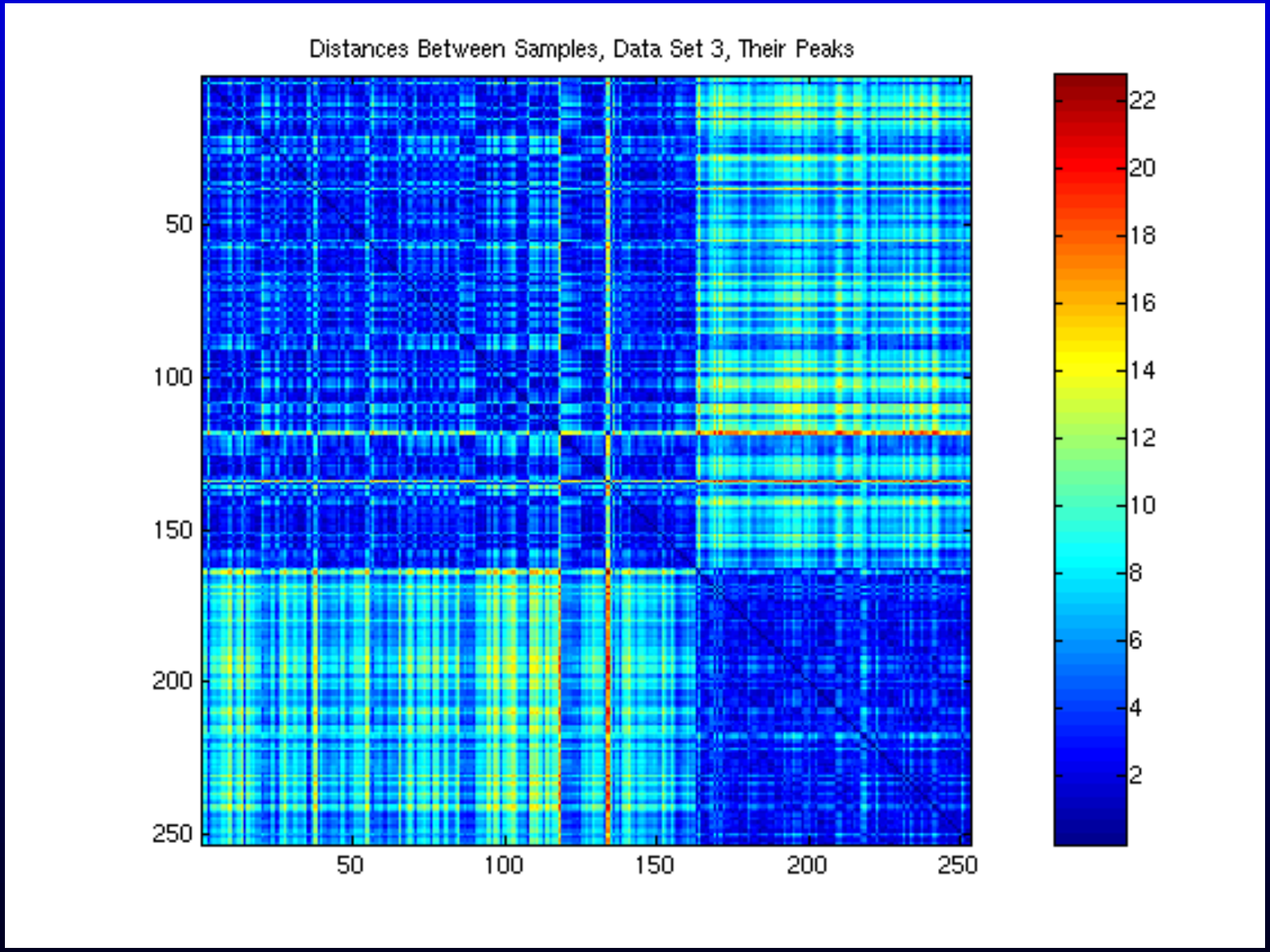
Or is it? Not in DS2



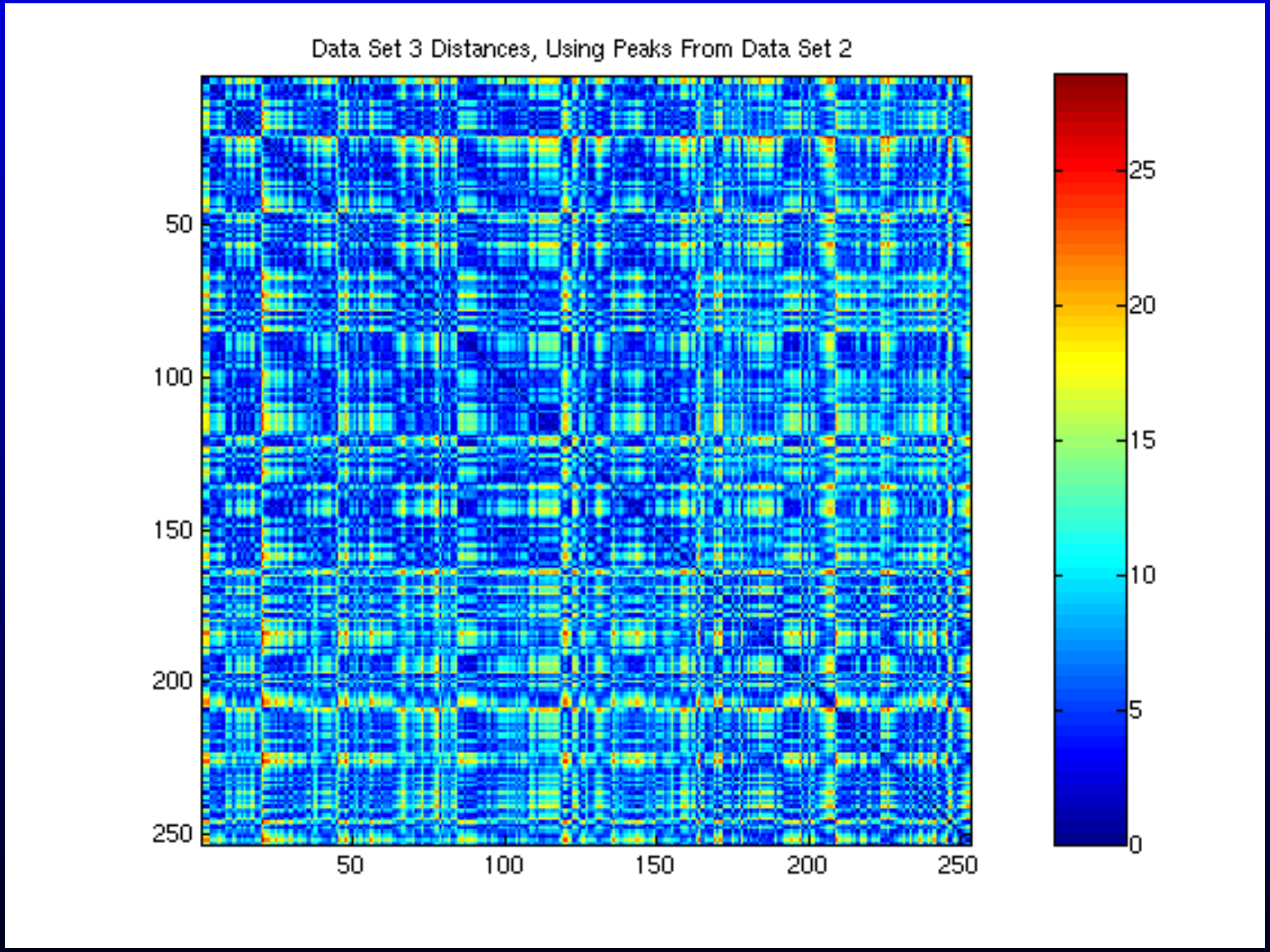
Processing Can Trump Biology (DS1 & DS2)



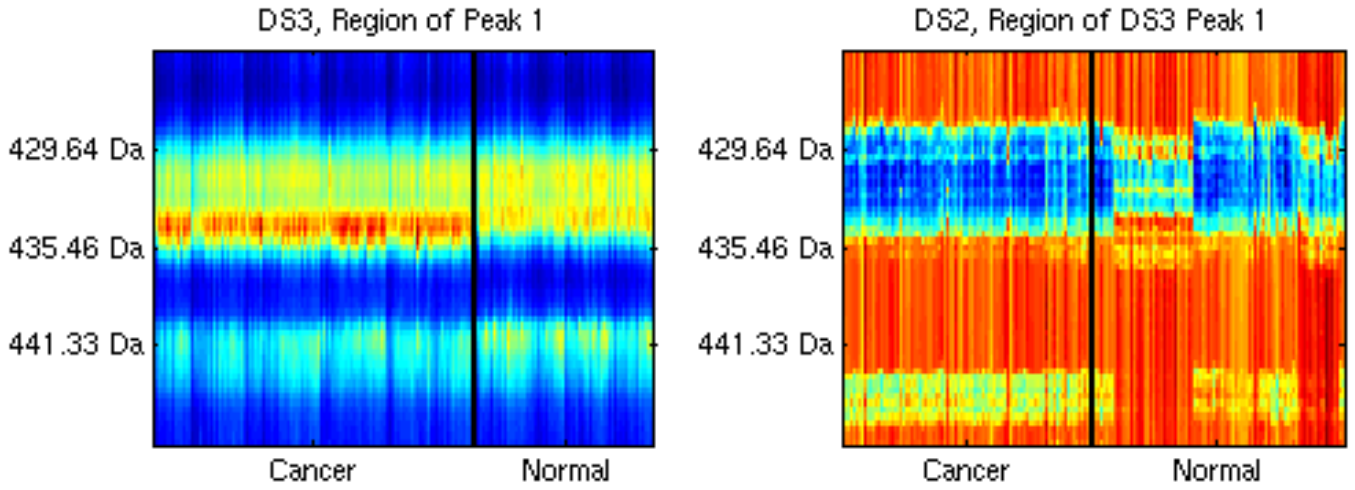
We Can Analyze Data Set 3!

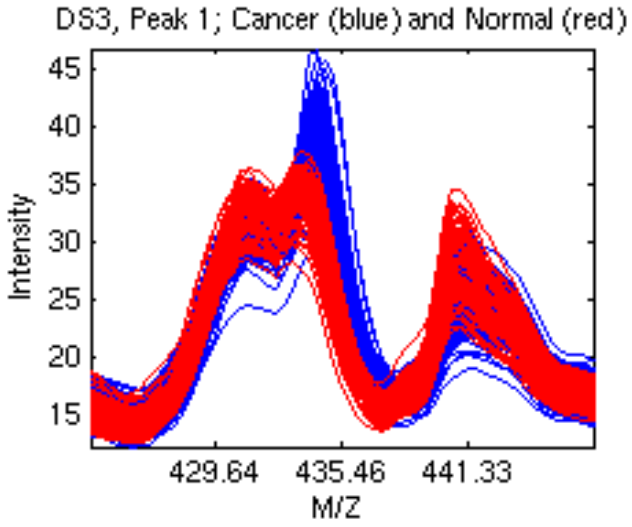


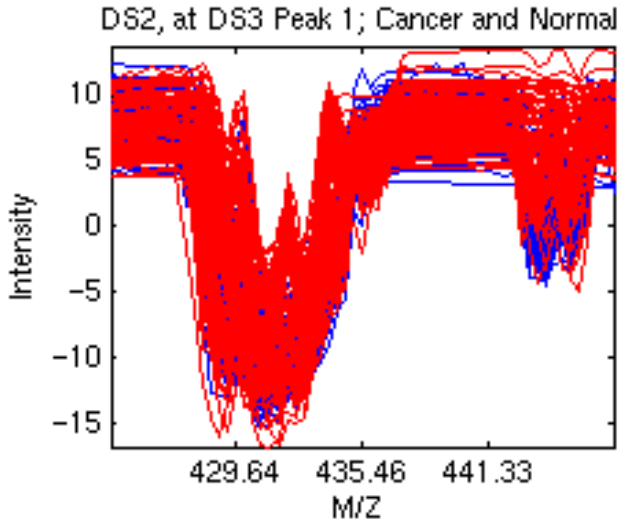
Do the DS2 Peaks Work for DS3?



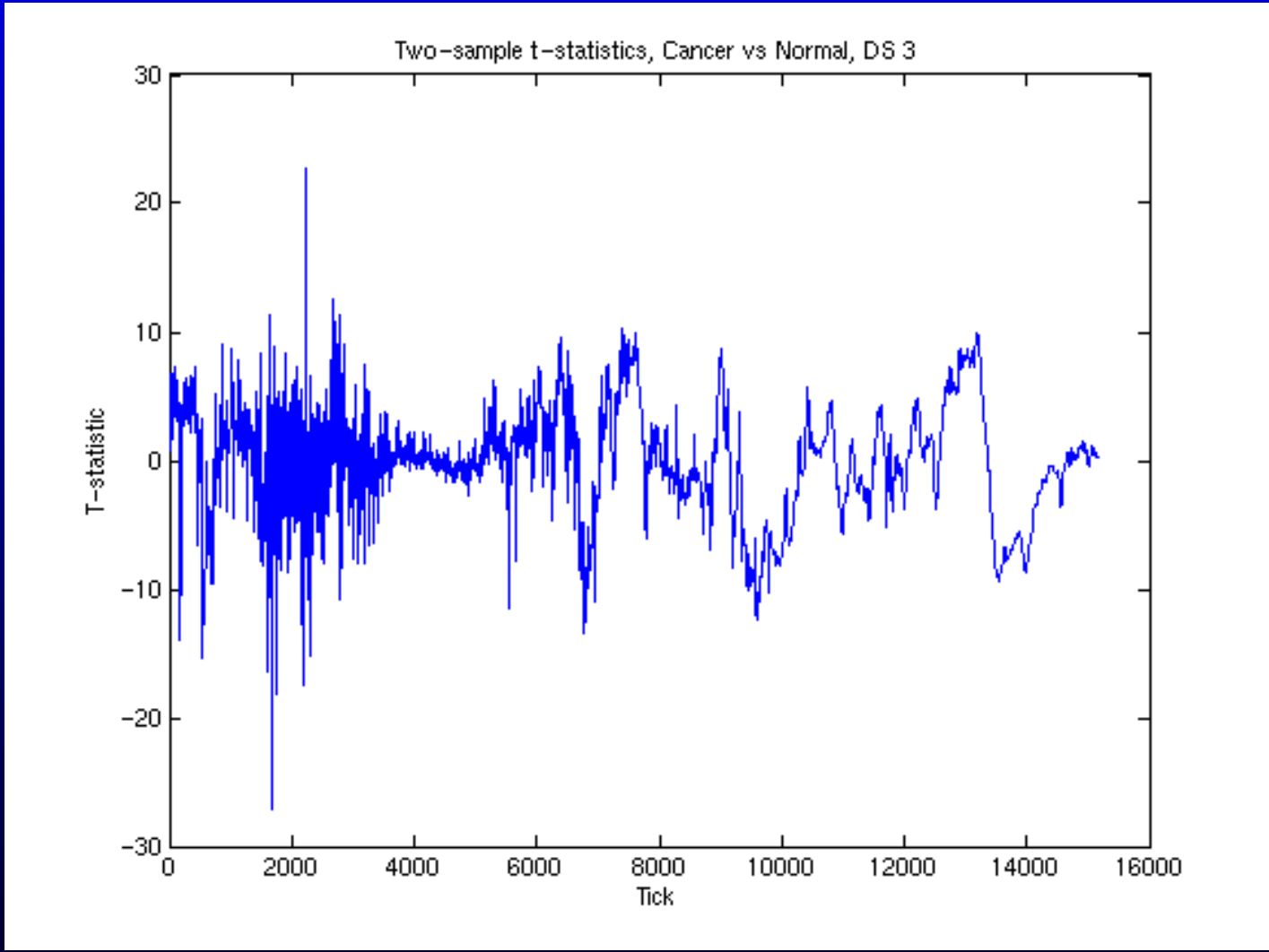
Do the DS3 Peaks Work for DS2?





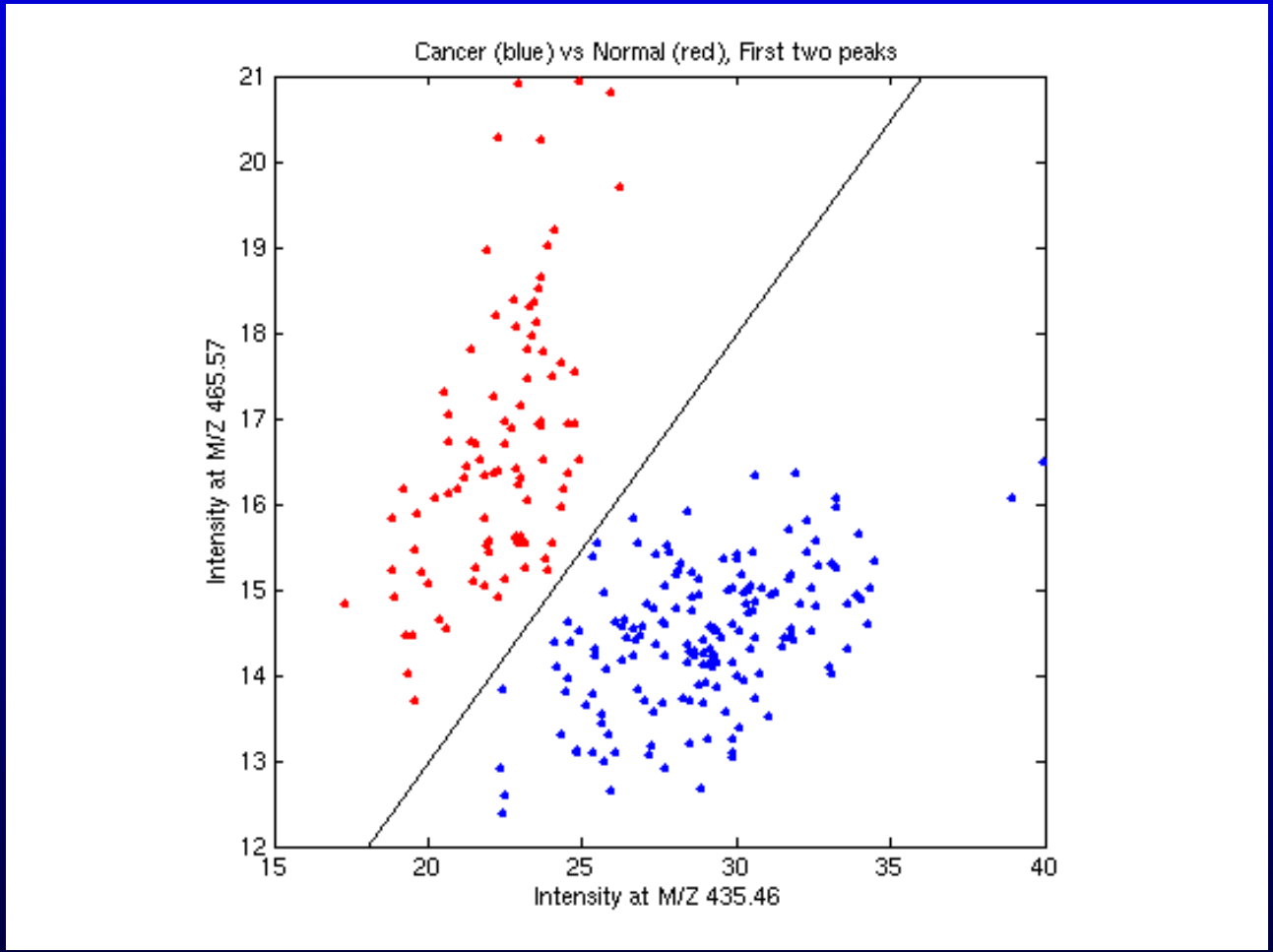


Which Peaks are Best? T-statistics



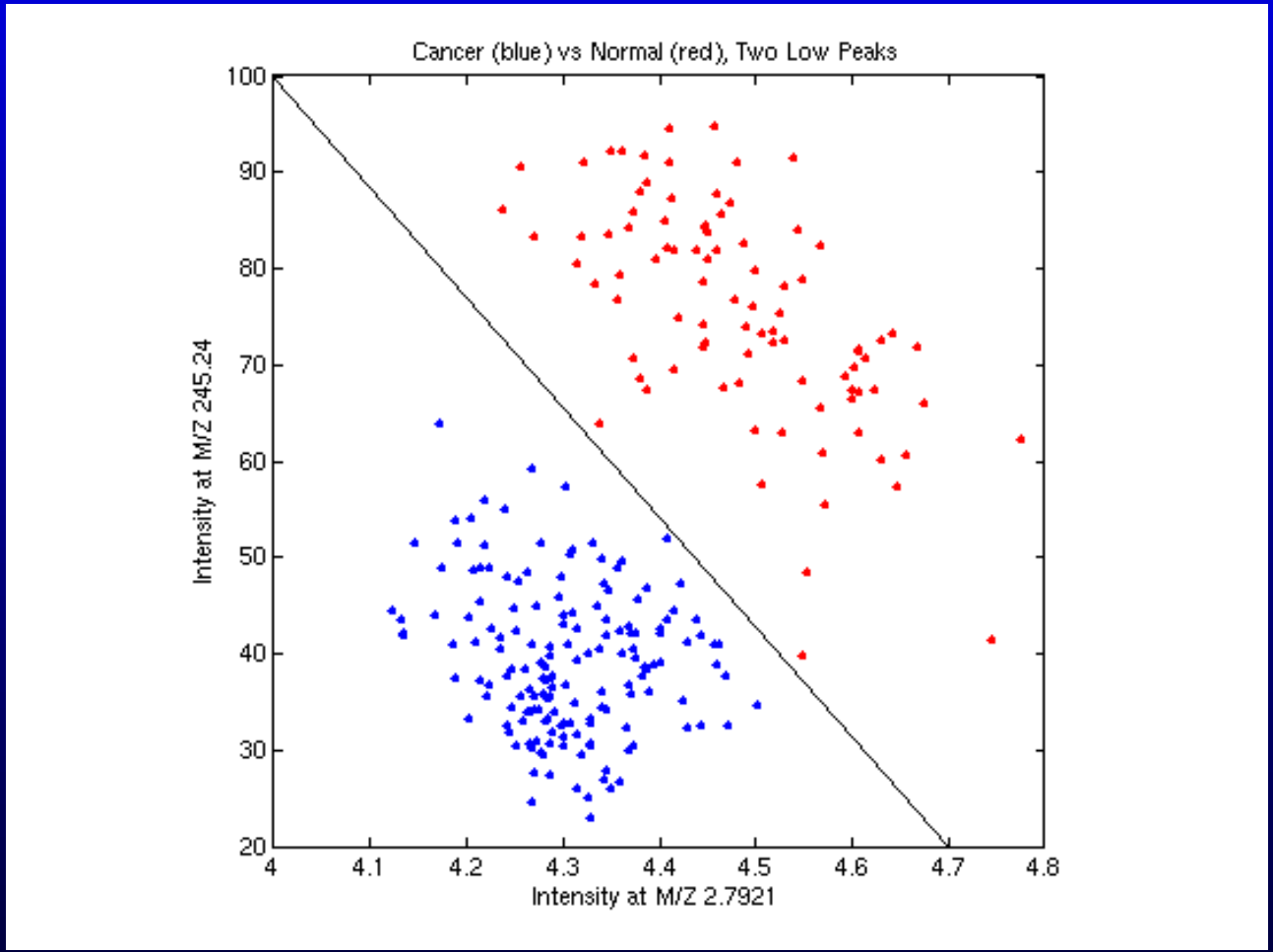
Note the magnitudes: t-values in excess of 20 (absolute value)!

One Bivariate Plot: M/Z = (435.46,465.57)



Perfect Separation. These are the first 2 peaks in their list, and ones we checked against DS2.

Another Bivariate Plot: M/Z = (2.79,245.2)



Perfect Separation, using a completely different pair. Further, look at the masses: this is the noise region.

Perfect Classification with Noise?

This is a problem, in that it suggests a qualitative difference in how the samples were processed, not just a difference in the biology.

This type of separation reminds us of what we saw with benign disease.

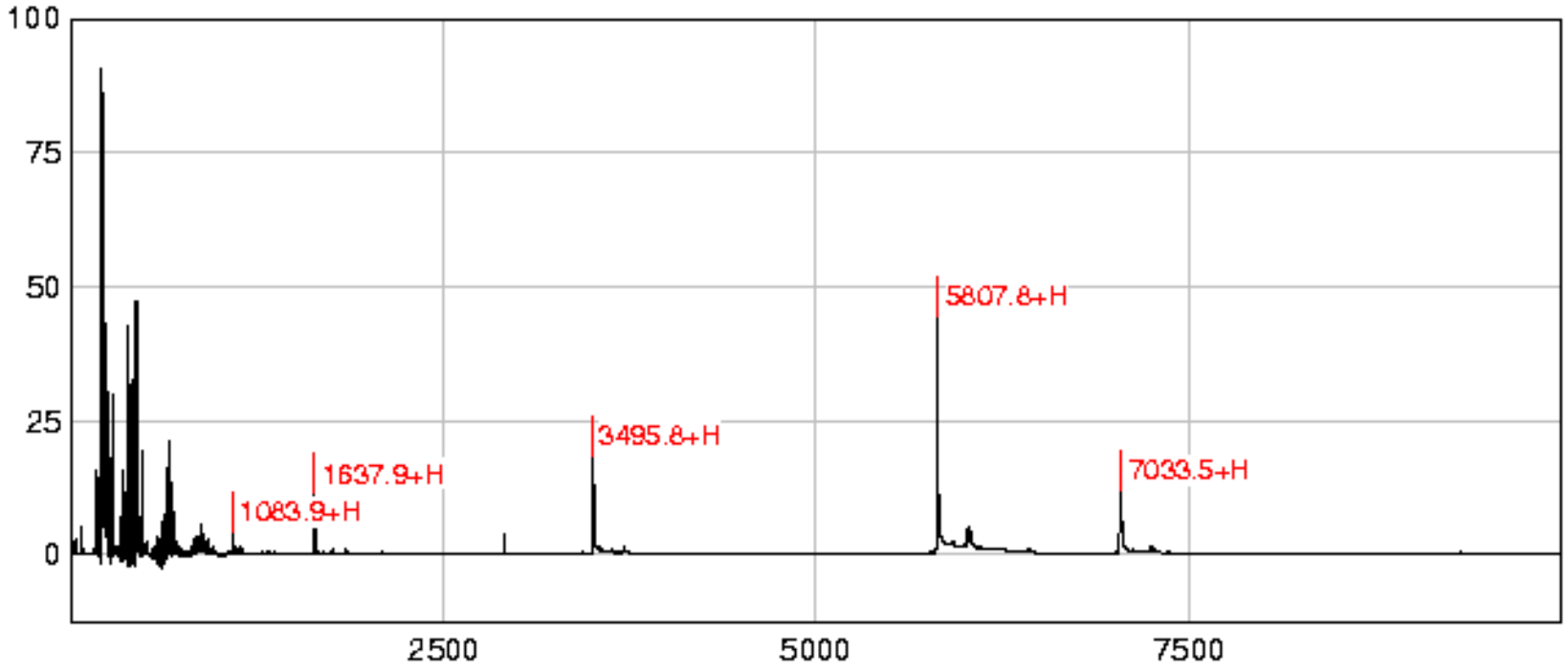
Mass Accuracy is Poor?

A tale of 5 masses...

Feb ’02 DS1	Apr ’02 DS2	Jun ’02 DS3
-7.86E-05	-7.86E-05	-7.86E-05
2.18E-07	2.18E-07	2.18E-07
9.60E-05	9.60E-05	9.60E-05
0.000366014	0.000366014	0.000366014
0.000810195	0.000810195	0.000810195

How are masses determined?

Calibrating known proteins



Calibration is the Same?

M/Z vectors the same for all three data sets.

Machine calibration the same for 4+ months?

What is the Calibration Equation?

The Ciphergen equation

$$\frac{m/z}{U} = a(t - t_0)^2 + b, \quad U = 20K, t = (0, 1, ...) * 0.004$$

What is the Calibration Equation?

The Ciphergen equation

$$\frac{m/z}{U} = a(t - t_0)^2 + b, \quad U = 20K, t = (0, 1, ...) * 0.004$$

Fitting it here

$$a = 0.2721697 * 10^{-3}, \quad b = 0, \quad t_0 = 0.0038$$

What is the Calibration Equation?

The Ciphergen equation

$$\frac{m/z}{U} = a(t - t_0)^2 + b, \quad U = 20K, t = (0, 1, ...) * 0.004$$

Fitting it here

$$a = 0.2721697 * 10^{-3}, \quad b = 0, \quad t_0 = 0.0038$$

These are the default settings that ship with the software!

Other issues

Q-star data different

clinical trials?

Example 2: Proteomics Data Mining

41 samples, 24 with lung cancer*, 17 controls.

20 fractions per sample.

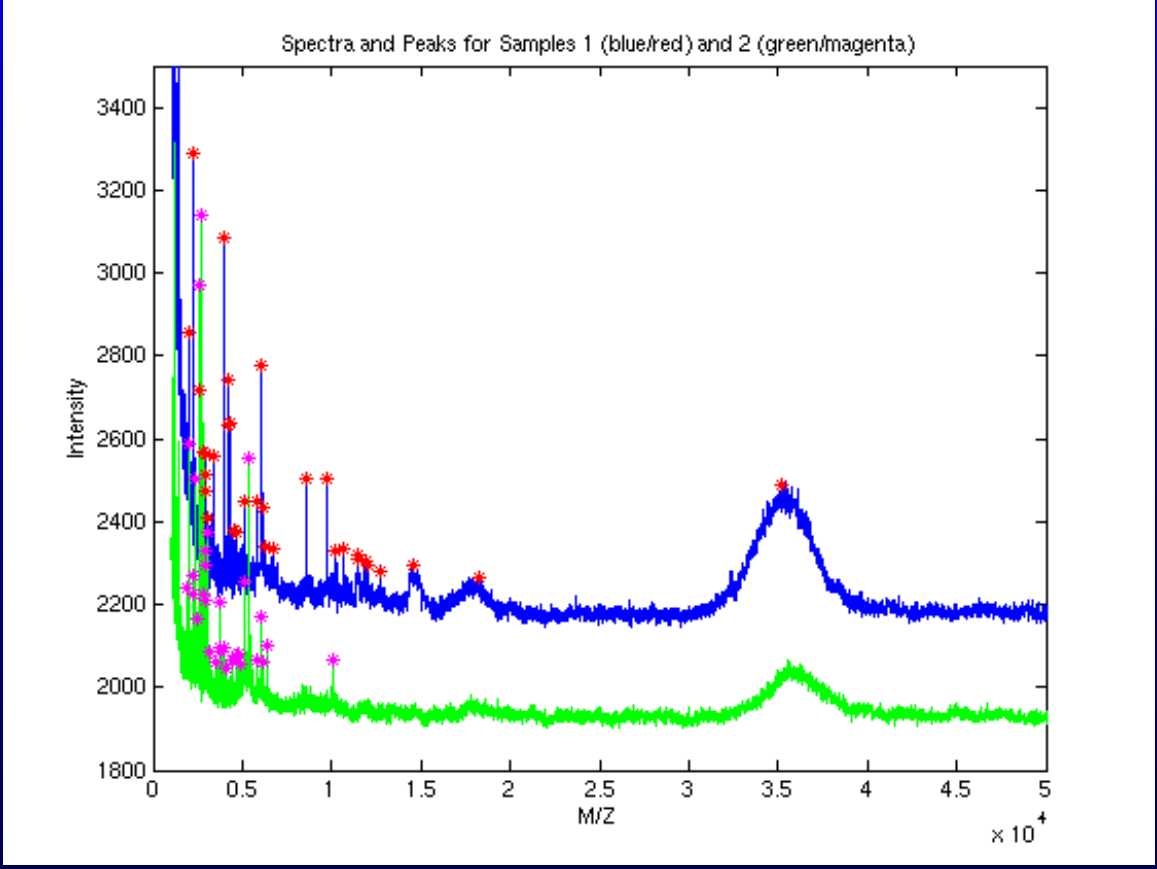
Goal: distinguish the two groups;

Data used to be at

http://www.radweb.mc.duke.edu/cme/proteomics/explain.htm

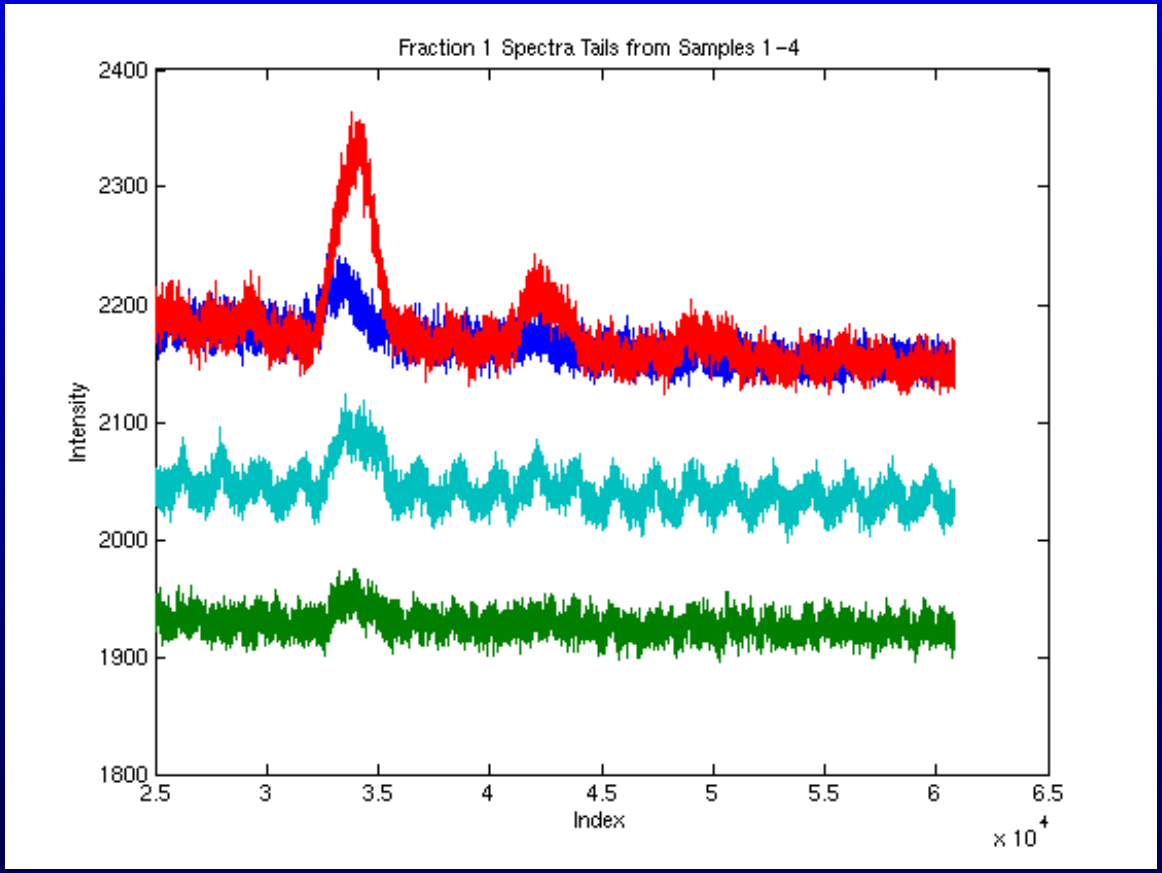
but the site has been retired. Send email to Ned Patz or Mike Campa at Duke if interested.

Raw Spectra Have Different Baselines



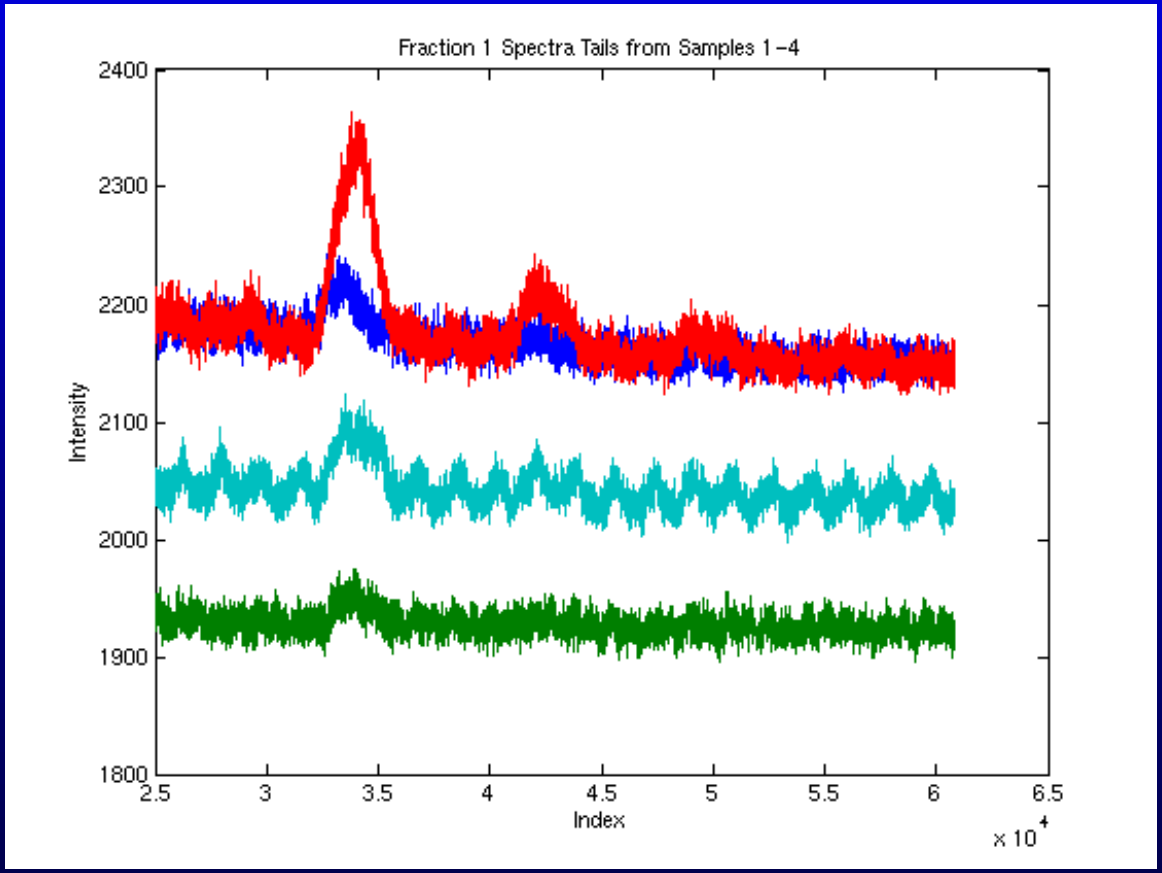
Note the need for baseline correction.

Oscillatory Behavior...



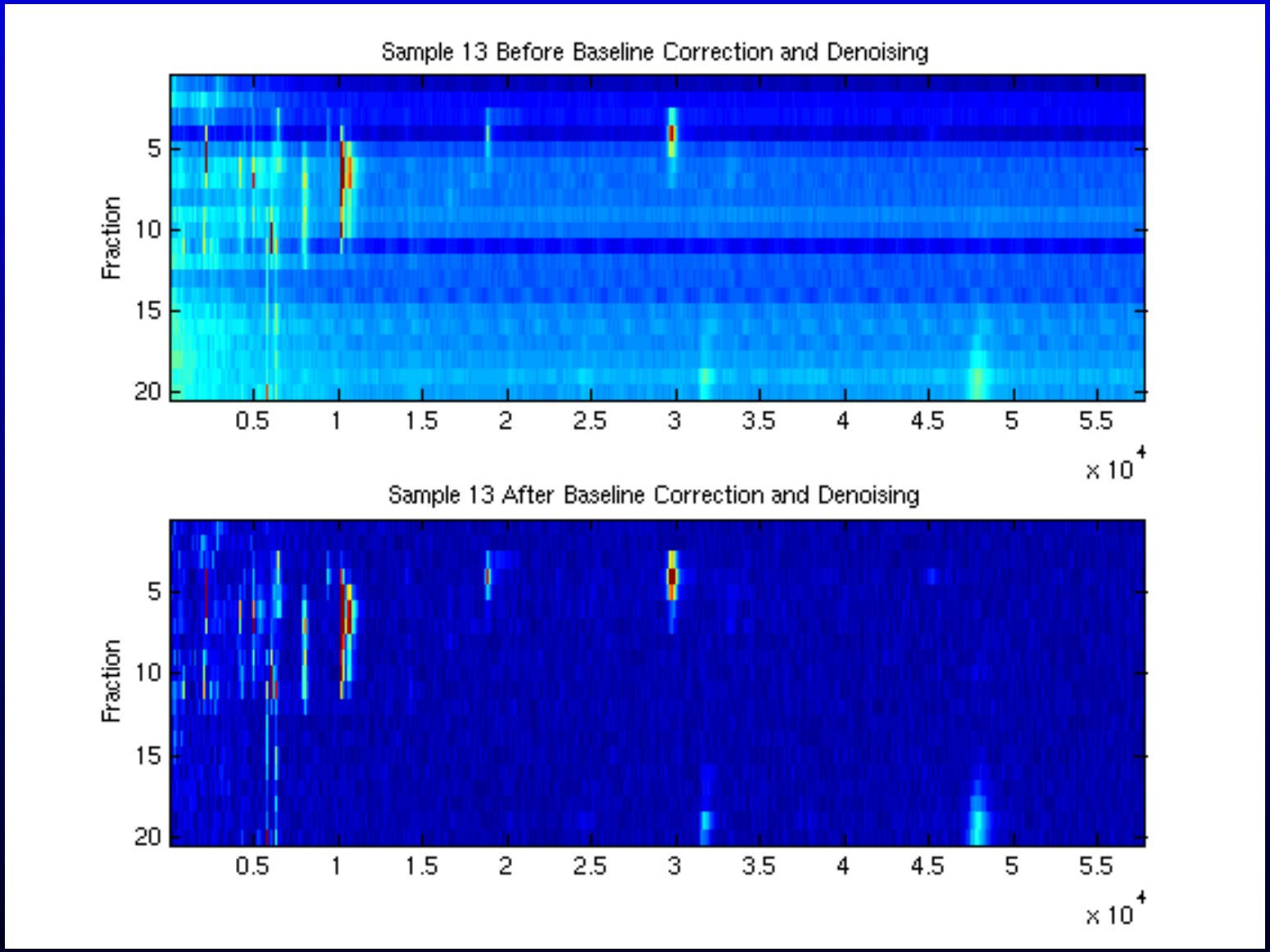
Roughly half the spectra have sinusoidal noise.

Oscillatory Behavior...

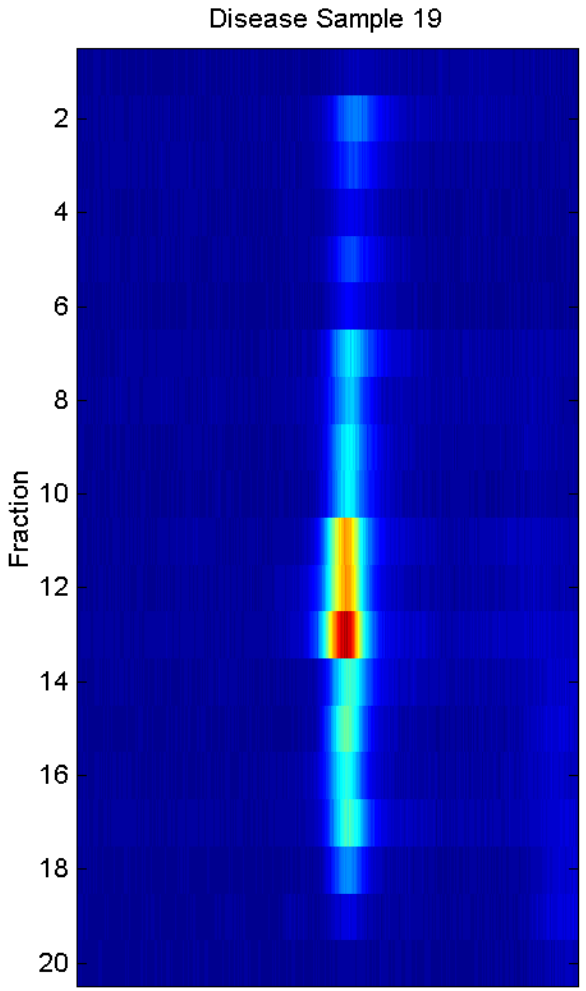


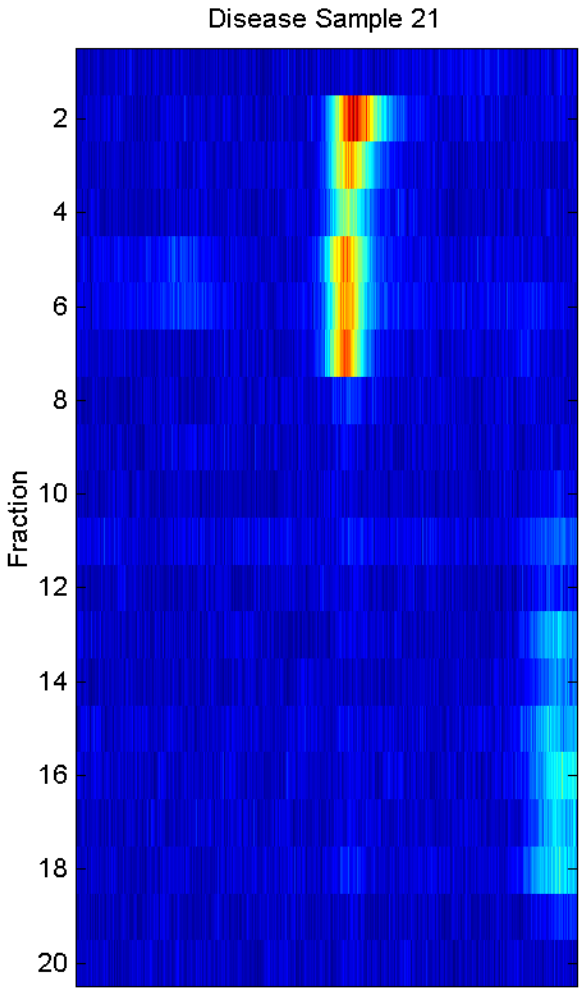
Roughly half the spectra have sinusoidal noise. We're seeing the A/C power cord.

Baseline Adj: Fraction Agreement, Before & After

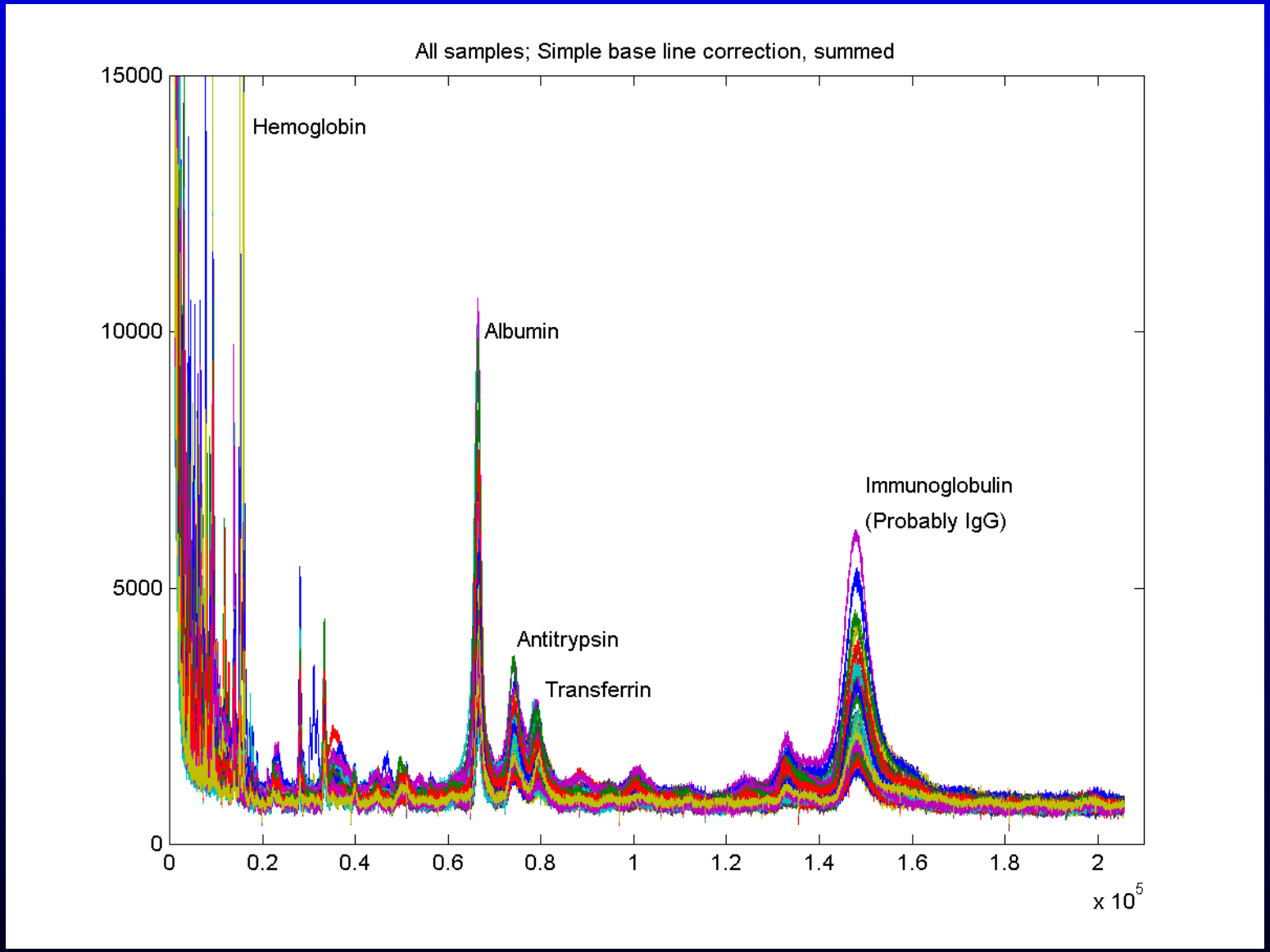


Fractionation is Unstable

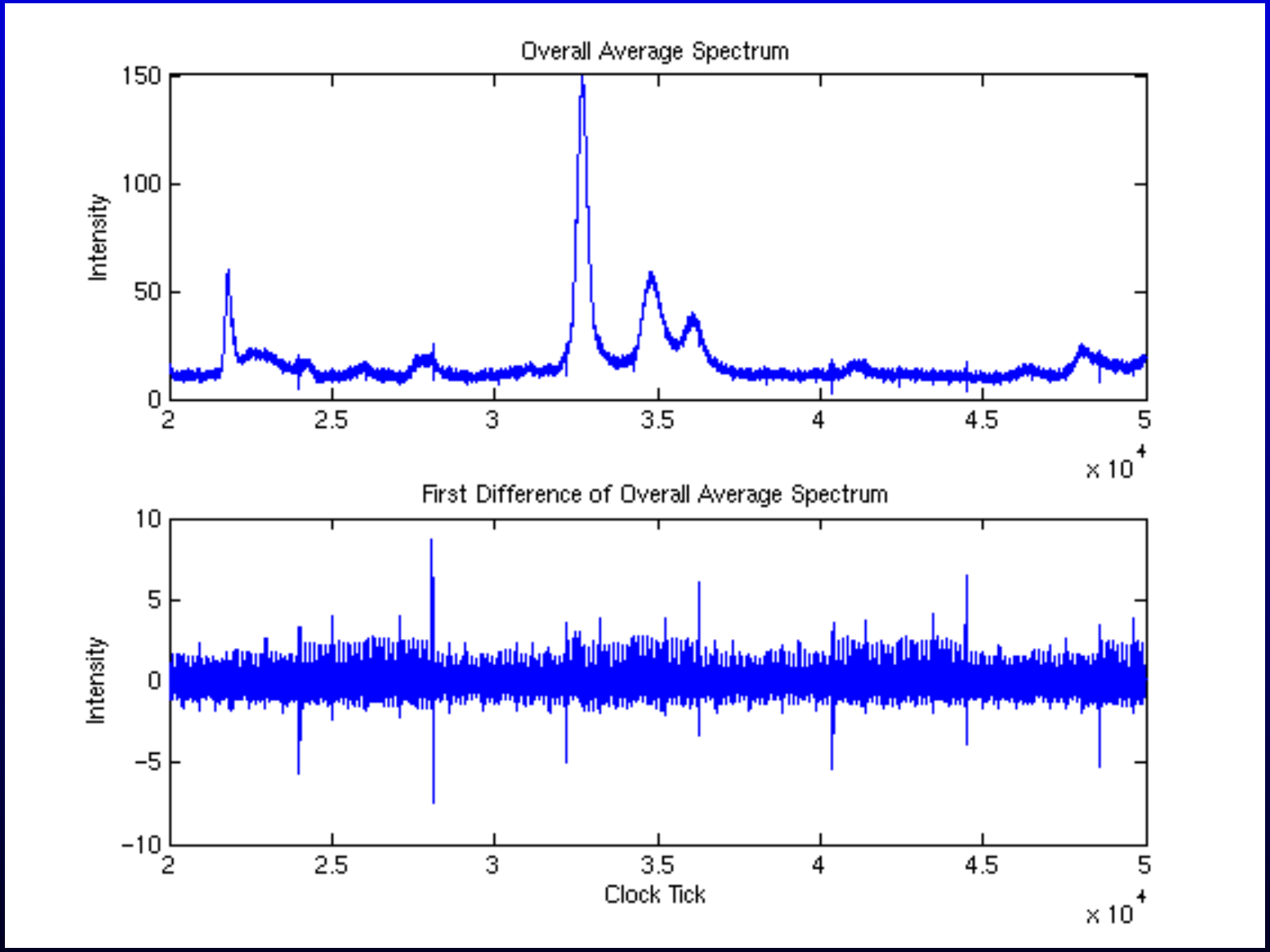




Unfractionating the Data

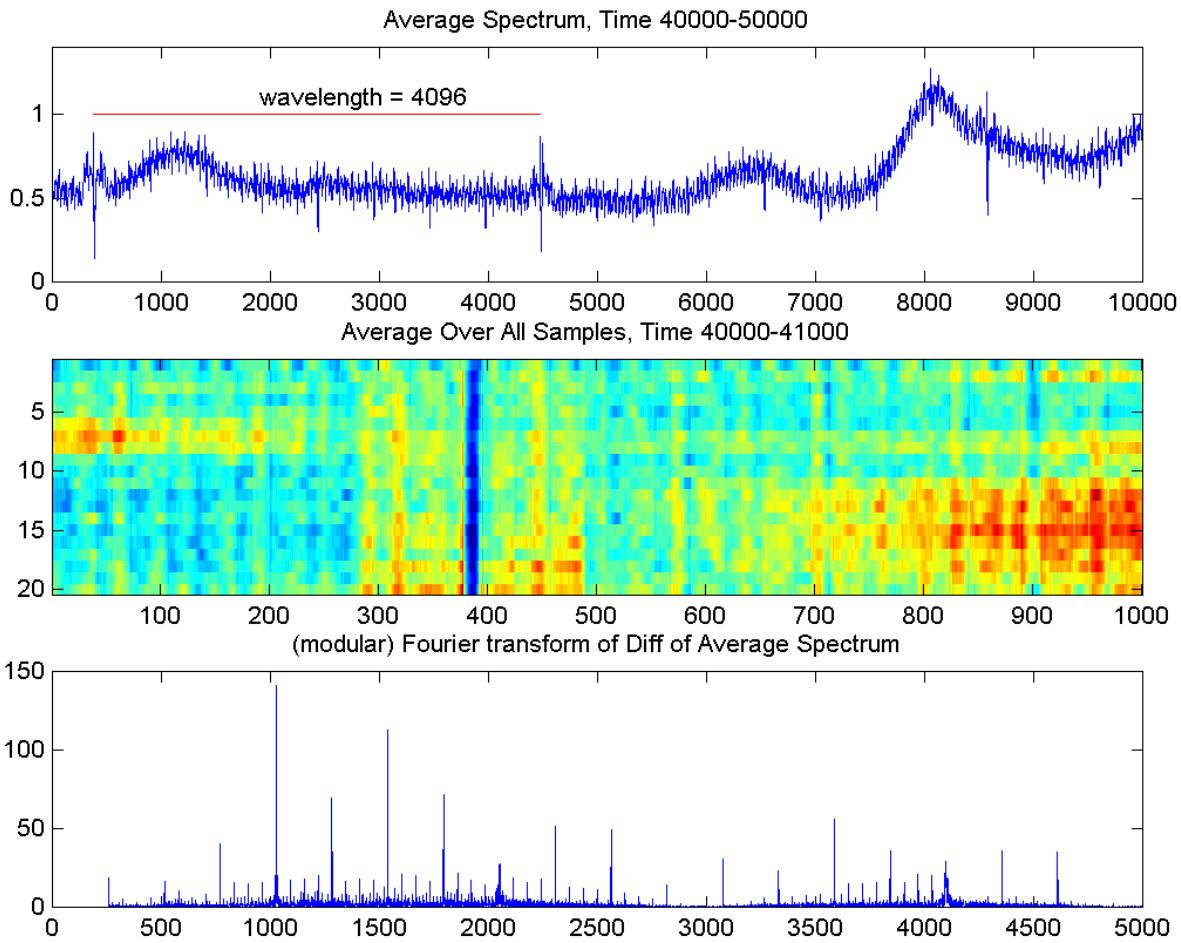


The Overall Average Shows Spikes. Difference It.



Computer Buffer?

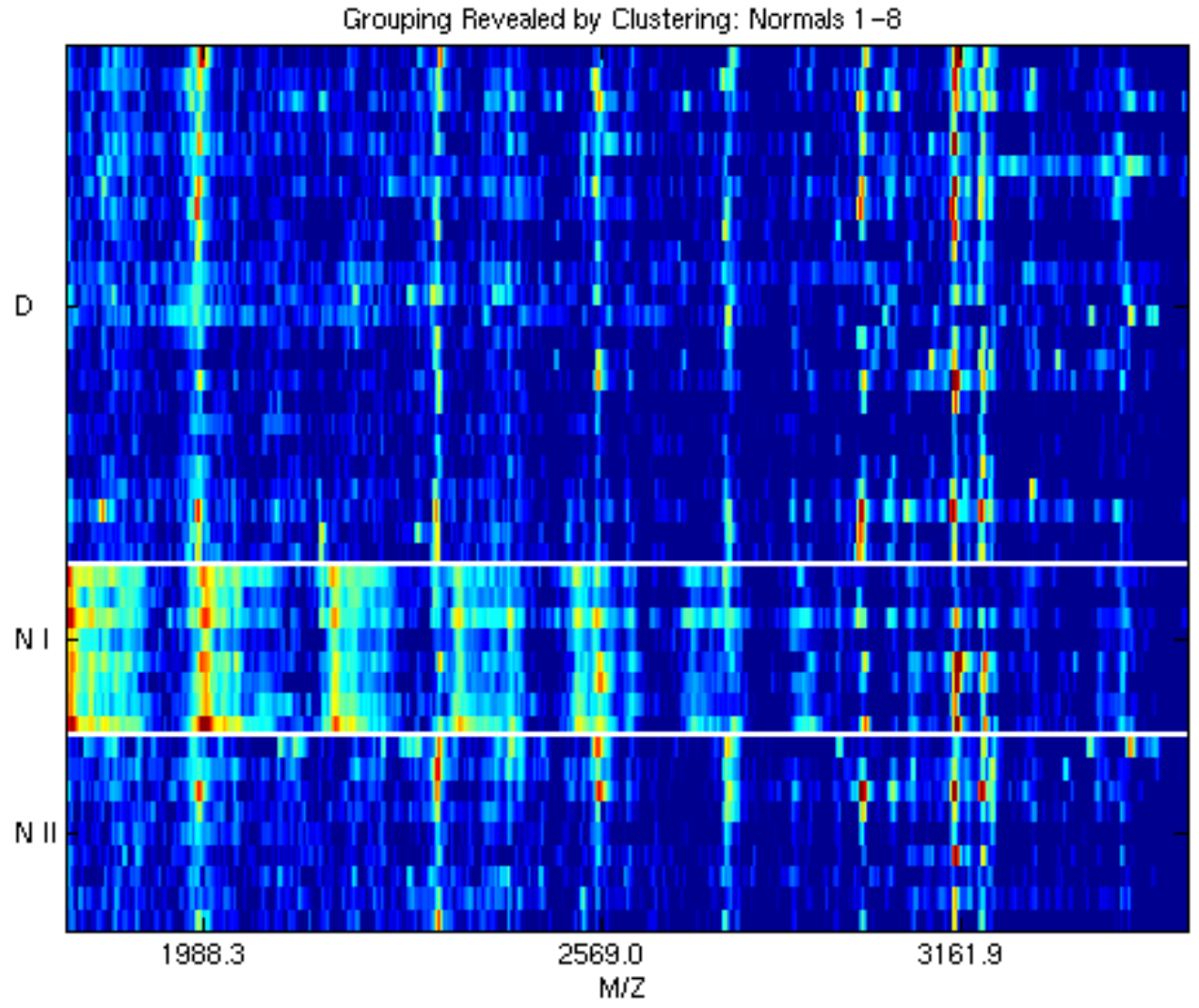
Spike spacing has a wavelength of $4096 = 2^{12}$.



Are We Done Cleaning Yet?

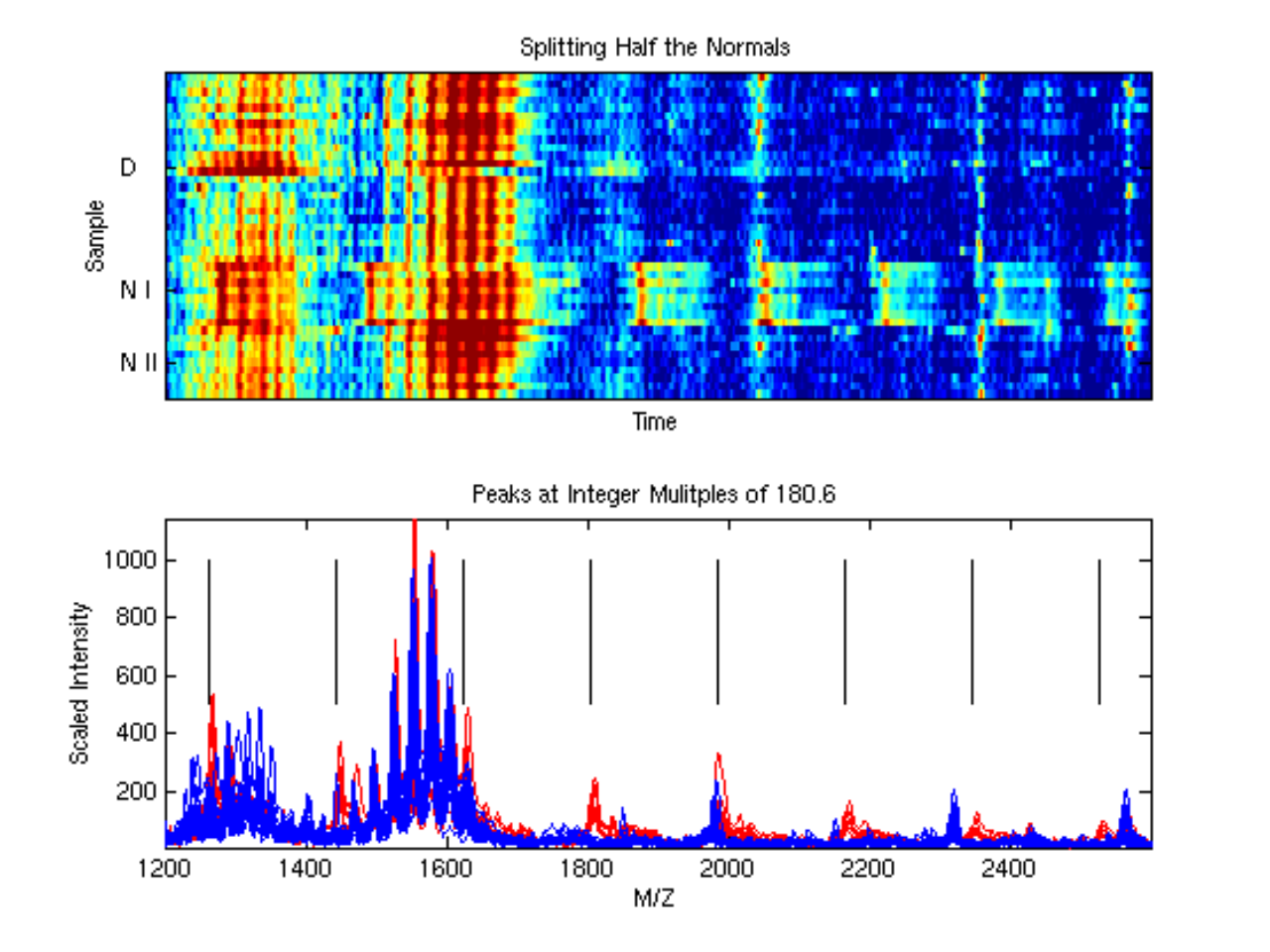
Give the problem a chance to be easy, try some simple clustering.

PCA Splits off Half the Normals



Peaks at Integer Multiples of M/Z 180.6!

This suggests a polymer. No Amino Acid dimers fit.



Cleaning Redux

- Baseline Correction and Normalization
- Inconsistent Fractionation
- Computer Buffers
- Polymers in some Normal Spectra
- Peak Finding (Use Theirs)

Data reduced to 1 spectrum/patient, with 506 peaks per spectrum.

Find the Best Separators

Peaks	MD	P-Value	Wrong	LOOCV
12886	2.547	≤ 0.005	11	11
8840, 12886	5.679	≤ 0.01	5	6
3077, 12886 74263	9.019	≤ 0.01	3	4
5863, 8143 8840, 12886	12.585	≤ 0.01	3	3
4125, 7000 9010, 12886 74263	23.108	≤ 0.01	1	1

There are 9 values that recur frequently, at masses of 3077, 4069, 5825, 6955, 8840, 12886, 17318, 61000, and 74263.

P-values are not from table lookups!

Testing Reality (Significance)

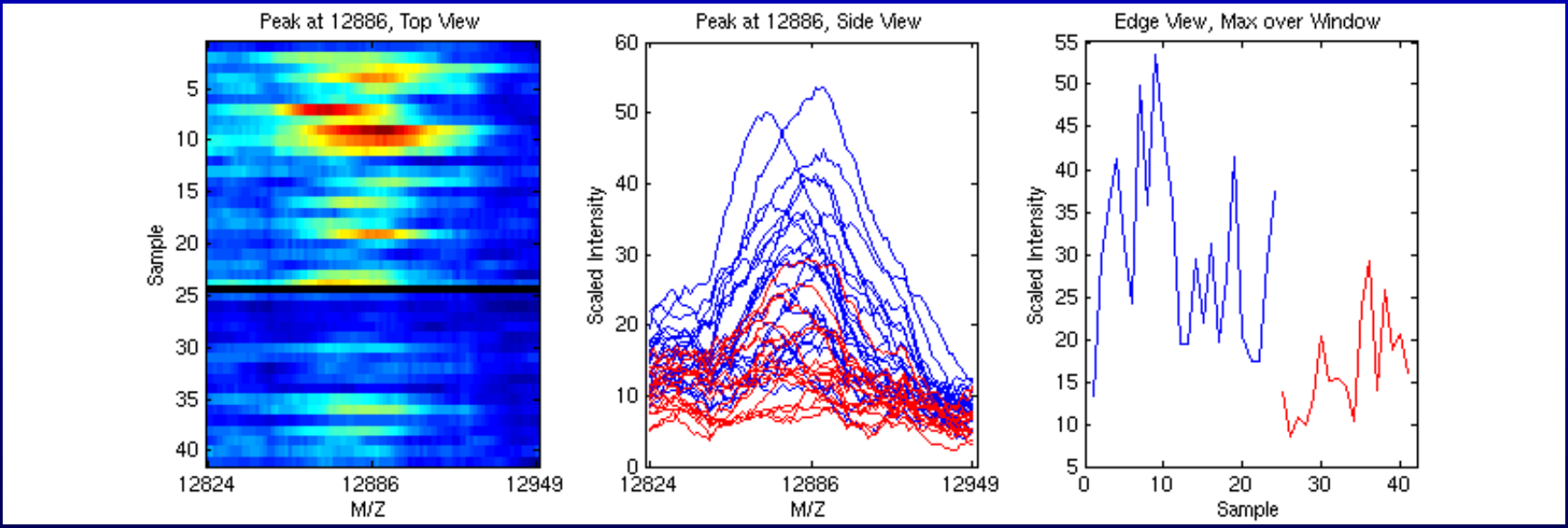
Generate a bunch of 'random noise' data matrices, each 41×506 in size.

For each matrix, split the 41 noise 'samples' into groups of 24 and 17.

Repeat our search procedure on the random data, and see how well we can separate things.

The Eyeball Test

We applied one last filtering step and actually *looked* at the regions identified. All 9 peaks listed above passed the eye test.



Blue lines = Cancers

Red lines = Controls

Punchlines

- There is no magic bullet here. (Too bad)

Punchlines

- There is no magic bullet here. (Too bad)
- Data preprocessing is extremely important with this type of data, and there is still much room for improvement.

Punchlines

- There is no magic bullet here. (Too bad)
- Data preprocessing is extremely important with this type of data, and there is still much room for improvement.
- Dimension reduction is critical; both to avoid spurious structure and to focus our attention on peaks.

Punchlines

- There is no magic bullet here. (Too bad)
- Data preprocessing is extremely important with this type of data, and there is still much room for improvement.
- Dimension reduction is critical; both to avoid spurious structure and to focus our attention on peaks.
- There is structure in this data (some peaks have been confirmed, and the writeup is in progress) and it can be found!

Punchlines

- There is no magic bullet here. (Too bad)
- Data preprocessing is extremely important with this type of data, and there is still much room for improvement.
- Dimension reduction is critical; both to avoid spurious structure and to focus our attention on peaks.
- There is structure in this data (some peaks have been confirmed, and the writeup is in progress) and it can be found!

Other Stuff

We were the only ones to notice the sinusoidal noise.

Other Stuff

We were the only ones to notice the sinusoidal noise.

and the clock tick.

Other Stuff

We were the only ones to notice the sinusoidal noise.

and the clock tick.

and we also won the competition...

Important Lessons

- Experimental Design Issues are Crucial

Important Lessons

- Experimental Design Issues are Crucial
  - Randomization
  - Uniform handling of samples
  - Blinding

Important Lessons

- Experimental Design Issues are Crucial
  - ⋆ Randomization
  - ⋆ Uniform handling of samples
  - ⋆ Blinding
- Careful Pre-Processing of Data is Essential

Important Lessons

- Experimental Design Issues are Crucial
  - Randomization
  - Uniform handling of samples
  - Blinding
- Careful Pre-Processing of Data is Essential
  - Calibration
  - Baseline Correction
  - Normalization

Important Lessons

- Experimental Design Issues are Crucial
  - ⋆ Randomization
  - ⋆ Uniform handling of samples
  - ⋆ Blinding
- Careful Pre-Processing of Data is Essential
  - ⋆ Calibration
  - ⋆ Baseline Correction
  - ⋆ Normalization
- Exploratory Data Analysis is Important: Look at the Data!

Important Lessons

- Experimental Design Issues are Crucial
  - ⋆ Randomization
  - ⋆ Uniform handling of samples
  - ⋆ Blinding
- Careful Pre-Processing of Data is Essential
  - ⋆ Calibration
  - ⋆ Baseline Correction
  - ⋆ Normalization
- Exploratory Data Analysis is Important: Look at the Data!
  - ⋆ Search for anomalies
  - ⋆ Confirm numerical results

References

On the *Lancet* data:

Baggerly, Morris and Coombes (2003), accepted by *Bioinformatics* pending revisions.

On the Proteomics Data Mining Conference data:

Baggerly, Morris, Wang, Gold, Xiao and Coombes (2003), *Proteomics*, **3(9)**:1677-1682.

pdf preprints are available.

The Deluge

Bladder Cancer

Pancreatic Cancer

Leukemia

Colorectal Cancer

Brain Cancer

Several show real structure, several show processing effects.

’If you’re not working on a proteomics project, you will be soon!’
Kevin Coombes to Bioinf section, 3/25/03

Collaborators

Keith Baggerly

Kevin Coombes

Jing Wang

David Gold

Lian-Chun Xiao

****

Ryuji Kobayashi

David Hawke

John Koomen