GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

22 November 2005

Lecture 23: Design, Microarrays, and Proteomics

- Design Questions
- Sample Size
- Simulations
- Intro to MALDI
- A SELDI Case Study
- Applied Classification
- High-Res Qstar Data

Design Questions

- What is the question being asked?
- What types of arrays are being used?
- What size of effect is being looked for?
- How many arrays are needed?

Some common questions

Class Comparison – given classes with membership known a priori, find genes showing differences between classes.

Class Prediction – build a model characterizing known classes, and use the model to predict the class status of future samples.

Class Discovery – identify subsets of samples based on their clustering behavior.

Class Comparison: Two classes

n_{canc} Cancers, n_{cont} Controls

For a fixed number of samples, how should these be divided between cases and controls?

Each measurement is subject to variation σ^2, and we want to estimate the cancer/control contrast with maximal precision.

Contrast: $Avg(Cancer) - Avg(Control)$.

$$V(Contrast) = \frac{\sigma^2}{n_{canc}} + \frac{\sigma^2}{n_{cont}}$$

Information = Inverse Variance

Optimal variance: $2/n_{canc}$.

Say we have 15 cancer samples and 5 normal samples. How much information do we have about the contrast?

$$\frac{1}{5} + \frac{1}{15} = \frac{4}{15} = \frac{2}{7.5}$$

so we have slightly less information than would be present in an experiment with 8 samples from each group.

More General Inverse Variance

$$\frac{1}{k} + \frac{1}{3k} = \frac{4}{3k} = \frac{2}{1.5k}$$

Two key principles:

Replication and Balance

What if we have 3 groups?

AB contrast: $\frac{1}{n_A} + \frac{1}{n_B}$

AC contrast: $\frac{1}{n_A} + \frac{1}{n_C}$

BC contrast: $\frac{1}{n_B} + \frac{1}{n_C}$

Are the contrasts equally important? Can we assert how much more important some given contrasts are?

What if groupings overlap?

Treatment 1 (high/low) and Treatment 2 (high/low)?

Aim for roughly equal numbers of samples at each of the 4 possible combinations

These groups are “factors” and the combinations of all possible levels comprise factorial designs

What types of arrays do we have?

Affymetrix, or other single-channel arrays: nothing qualitatively new.

Two-color arrays: may have some new features associated with the natural pairing of samples.

Reference Designs for cDNA Arrays

Notation: Ratios are in Red/Green order

Comparing two groups, A and B

A_1/Ref, A_2/Ref, B_1/Ref, B_2/Ref

Focus on log ratios

$$Avg(\log(A/\mathrm{Ref})) - Avg(\log(B/\mathrm{Ref}))$$

Is this is a good design? (Can we do better?)

When can we do better?

If the only contrast of interest is A vs B (ie, the reference itself is not of secondary interest)

If we are unlikely to expand the contrast later (introducing, say, group C)

In this case, comparisons made using a reference are indirect, and direct comparisons may give more precision

How much better?

A_1/B_1, B_2/A_2 vs A_1/Ref, B_1/Ref (2 arrays each)

Say the variance associated with measuring a single log ratio is σ^2; we want to estimate $\log(A/B)$.

$$V(\tfrac{1}{2}\log(A_1/B_1) - \tfrac{1}{2}\log(B_2/A_2)) = \tfrac{1}{4}\sigma^2 + \tfrac{1}{4}\sigma^2 = \sigma^2/2.$$

$$V(\log(A_1/\mathrm{Ref}) - \log(B_1/\mathrm{Ref})) = \sigma^2 + \sigma^2 = 2\sigma^2$$

Direct comparison can be 4 times as precise.

This makes some assumptions about independence, but direct comparisons are never worse.

How can we extend this?

With two groups, work on balanced blocks

Same number of A and B, one each per array, equal numbers of arrays with A/B and B/A.

Note that the reference design didn't necessarily require dye swaps, but the direct comparisons do. (This assumes comparisons with the reference are not of interest!)

Very efficient in terms of numbers of arrays used for the amount of information obtained.

Doesn't work as nicely for clustering.

A Loop Extension

A_1/B_1, B_1/A_2, A_2/B_2, B_2/A_1

Every sample is used twice, once in red, once in green.

All pairs of samples can be compared through paths in the loop, cancelling out intervening terms:

$$\log(A_1/B_2) = \log(A_1/B_1) + \log(B_1/A_2) + \log(A_2/B_2)$$

pairs with terms “farther away” in the loop have their contrasts estimated less well.

Do we use Loops?

Rarely. Loops can be broken by bad arrays.

Loops are more complex in terms of analysis if we are interested in individual pairs than reference designs.

For contrasting two groups, randomized blocks work just as well.

Loops can require more uses of small amounts of RNA.

Aesthetically, however, they're quite nice.

Another Design Issue

Randomization.

This is underdiscussed in the array literature, but should be at least contemplated so as to avoid biases. This can help offset issues associated with run order, tech running the arrays, etc.

How many arrays do we need?

To do what?

Have a minimal level of power to detect an effect of a given size.

This is the classical problem of setting sample sizes, requiring decisions about sensitivity and specificity.

Numbers Needed

All told, we need to specify at least 4 parameters:

α, the significance level

$1 - \beta$, the statistical power

δ, the size of the effect we want to be able to see (e.g., 1 on a log scale)

σ, the standard deviation of the gene expression levels

An analytic approach

Given these, Simon et al (2002) show that

$$n = \frac{4(t_{\alpha/2}+t_{\beta})^2}{(\delta/\sigma)^2}$$

suffices, where the t distribution has $n-2$ degrees of freedom (this requires iteration in fitting).

In order to get started on the iteration, it pays to think big initially, so that

$$n = \frac{4(z_{\alpha/2}+z_{\beta})^2}{(\delta/\sigma)^2}$$

Using zs means that no iteration is required.

Some sample numbers

To account for multiple testing, they suggest starting with small values of α and β, such as $\alpha = 0.001$ and $\beta = 0.05$.

Of course, the value of σ will change from gene to gene, but some intermediate value from prior data (such as the median value) can be used to suggest the right order.

Using $\alpha = 0.001, \beta = 0.05, \sigma = 0.5, \delta = 1$, the target value of n is 26 using the z approximation, and goes up to 30 using the t distribution.

A limitation

The above approach assumes that the two groups to be contrasted will be present in roughly equal amounts ($n/2$). If the true ratio is to be $k:1$ instead of $1:1$, then the target size needs to be scaled by a factor of $(k+1)^2/4k$.

A simulation approach

Alexander Zien et al approach the problem of sample size determination through simulation, but they focus on a more involved model that explicitly incorporates multiple types of error (additive and multiplicative).

Their java computation applet is available at

http://www.scai.fhg.de/special/bio/howmanyarrays/

There is a similar applet at

http://bioinformatics.mdanderson.org

Their qualitative observations

Biological variation dominates technical variation

Measuring more samples is better than replicating measurements of the same samples

Sizes of classes should be as balanced as possible

Non-parametric tests are better

What their simulation requires

Estimates of variability:

multiplicative biological variability

multiplicative technical variability

additive technical variability

desired detectable fold change

desired detectable signal to noise ratio

What their simulation requires

numbers of samples in each class

numbers of genes on the array

numbers of genes expected to be “really different”

What it returns

Simulated false positive rates

Simulated false negative rates

sensitivity and specificity

For their default values, they found that they needed about 12-15 samples per class.

What Are Proteomic 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

Proteomic spectra are mass spectrometry traces of biological specimens

Why Are We Excited?

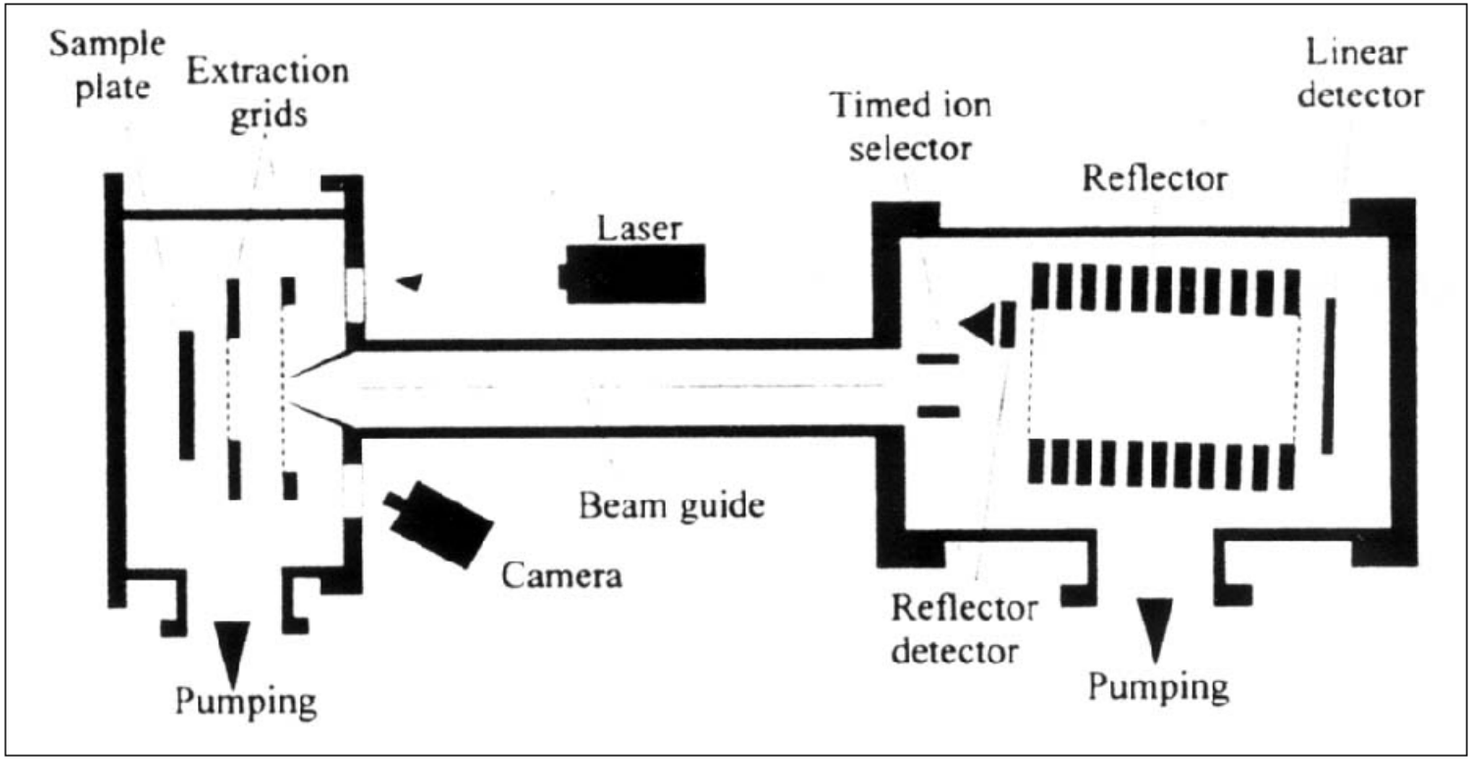
Profiles at this point are being assessed using serum and urine, not tissue biopsies

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

Can run samples on large numbers of patients

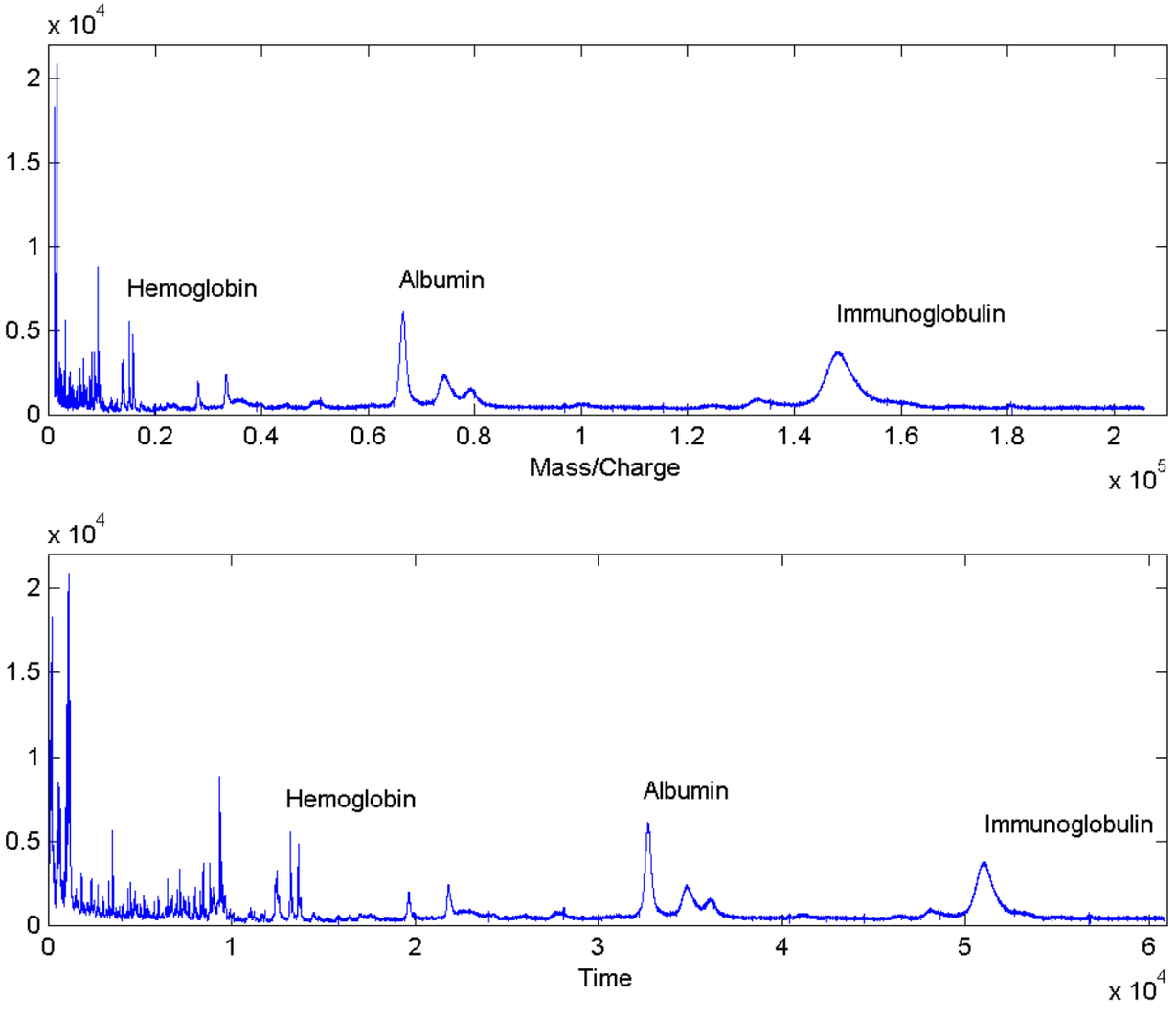
How Does Mass Spec Work?

Block Diagram of a MALDI-TOF

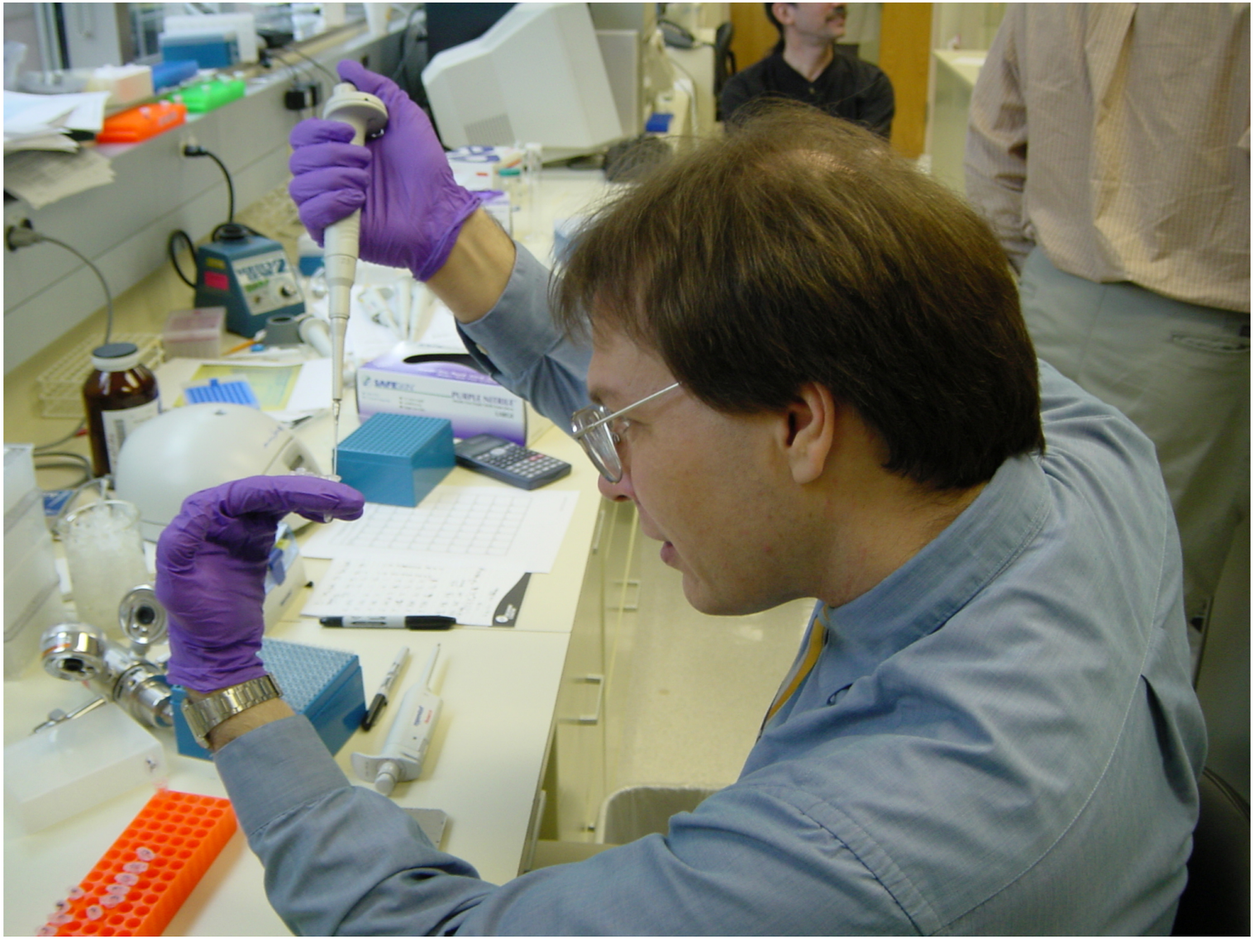


Vestal and Juhasz. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 892.

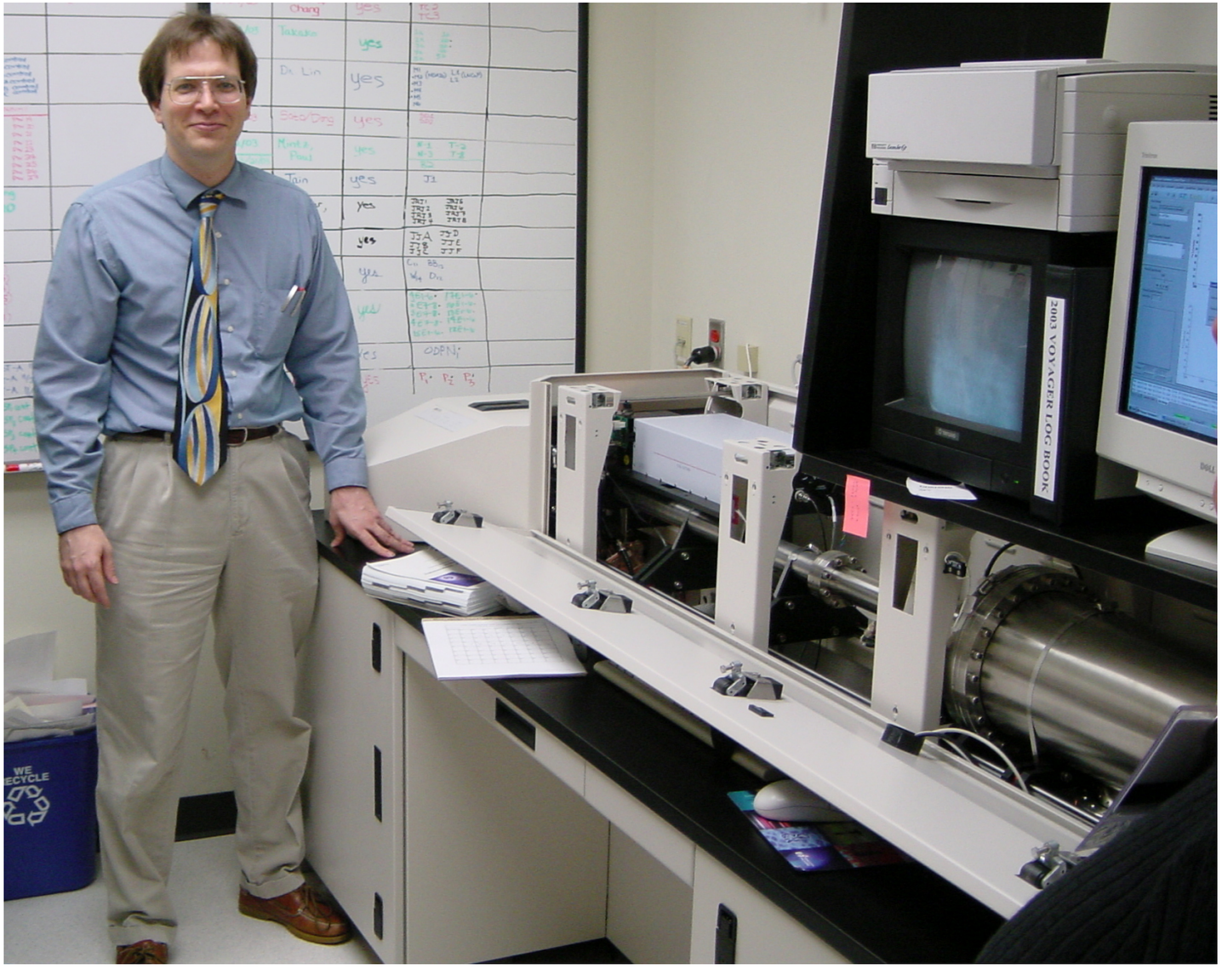
What Do the Data Look Like?



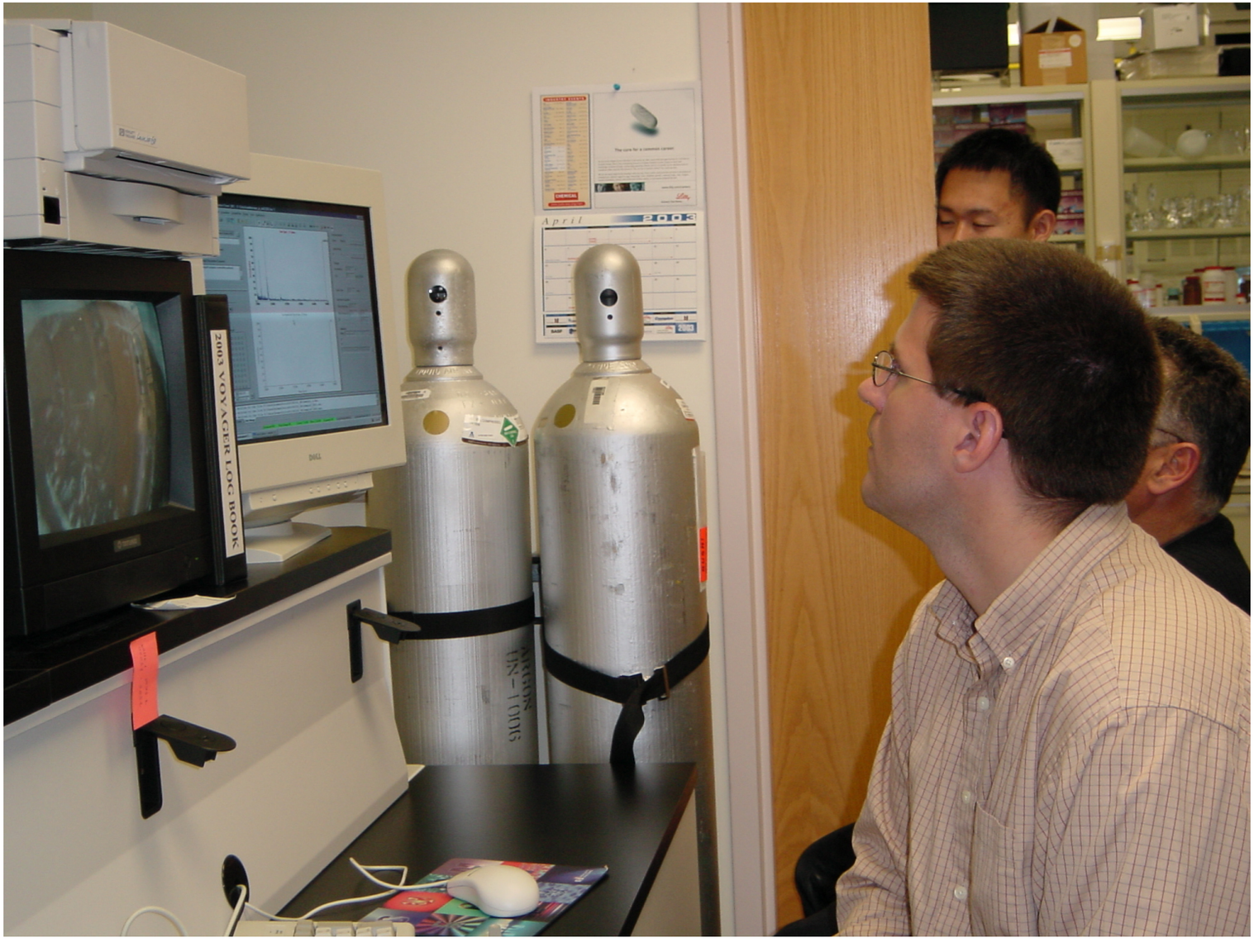
Learning: Spotting the Samples



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 SELDI Case Study

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://home.ccr.cancer.gov/ncifdaproteomics/ (used to be 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

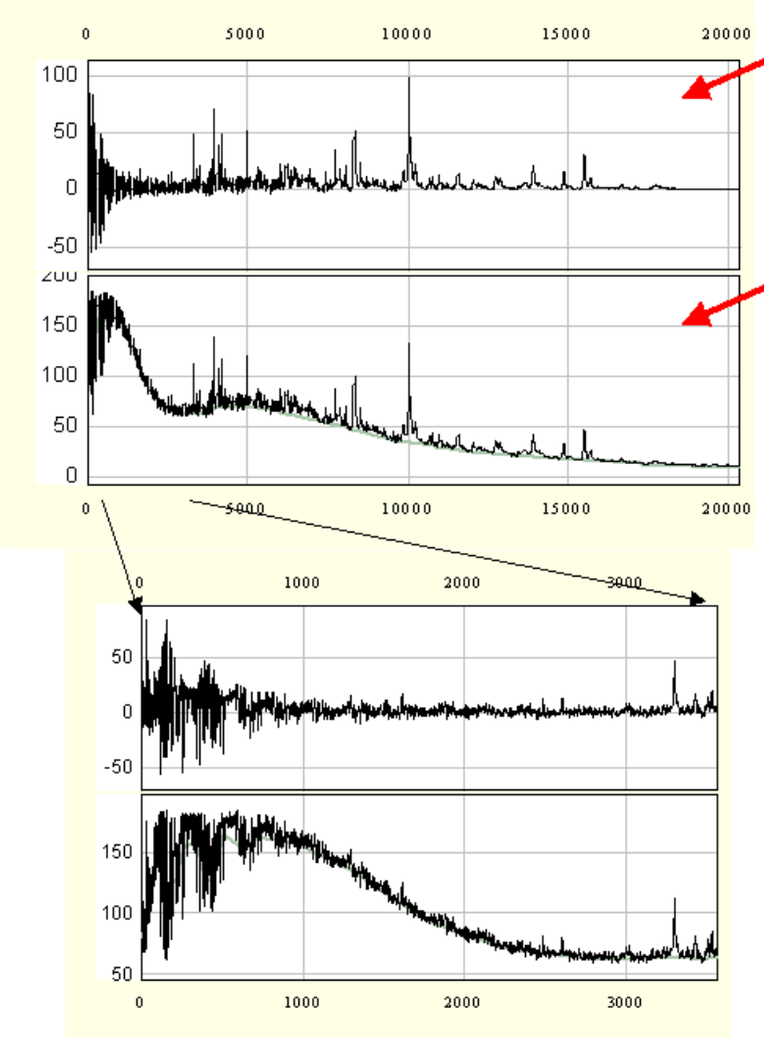
Data Set 2 – Followup: the same 216 samples, baseline subtracted, WCX2 chip

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

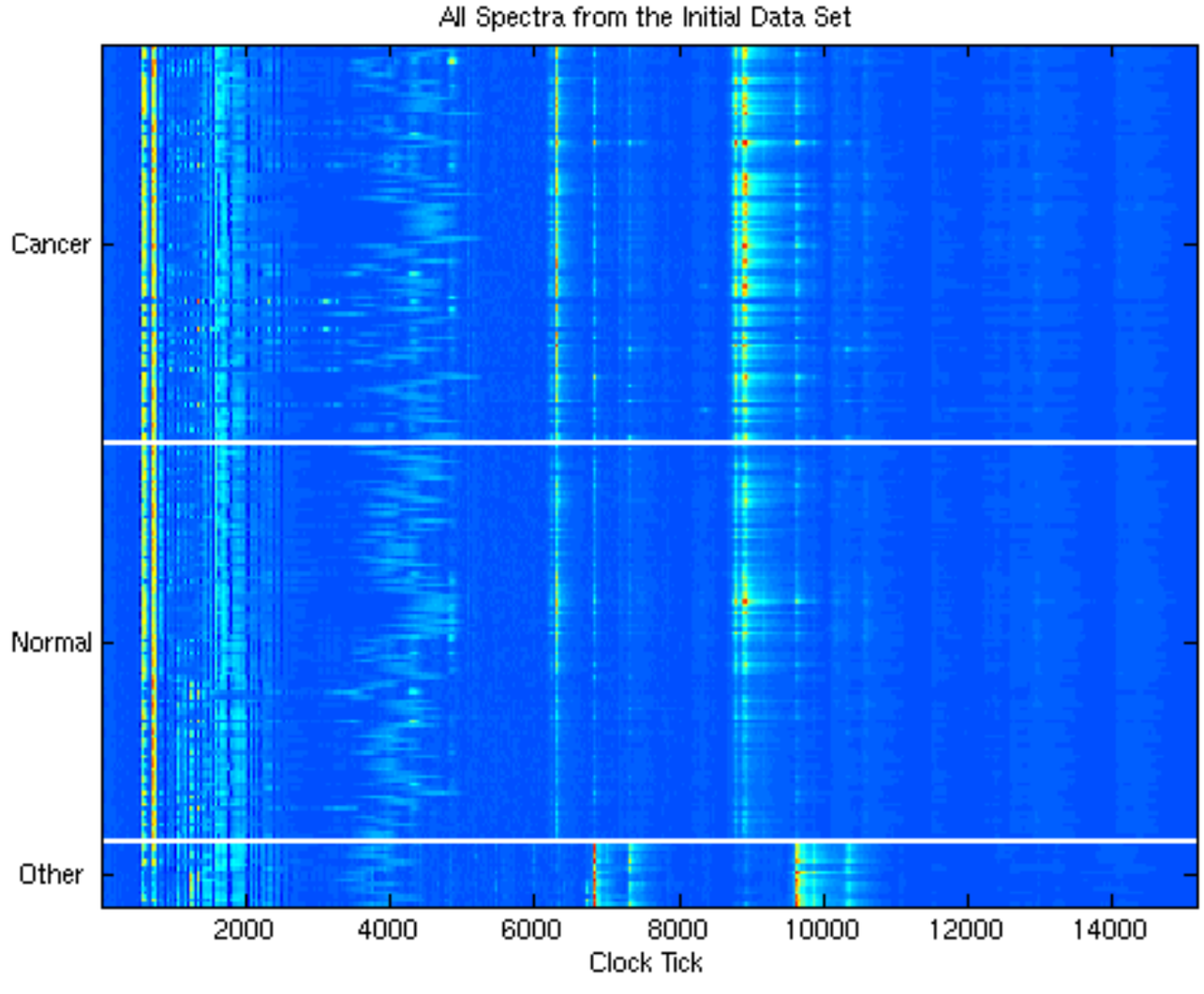
They analyzed this

Baseline Subtraction OFF

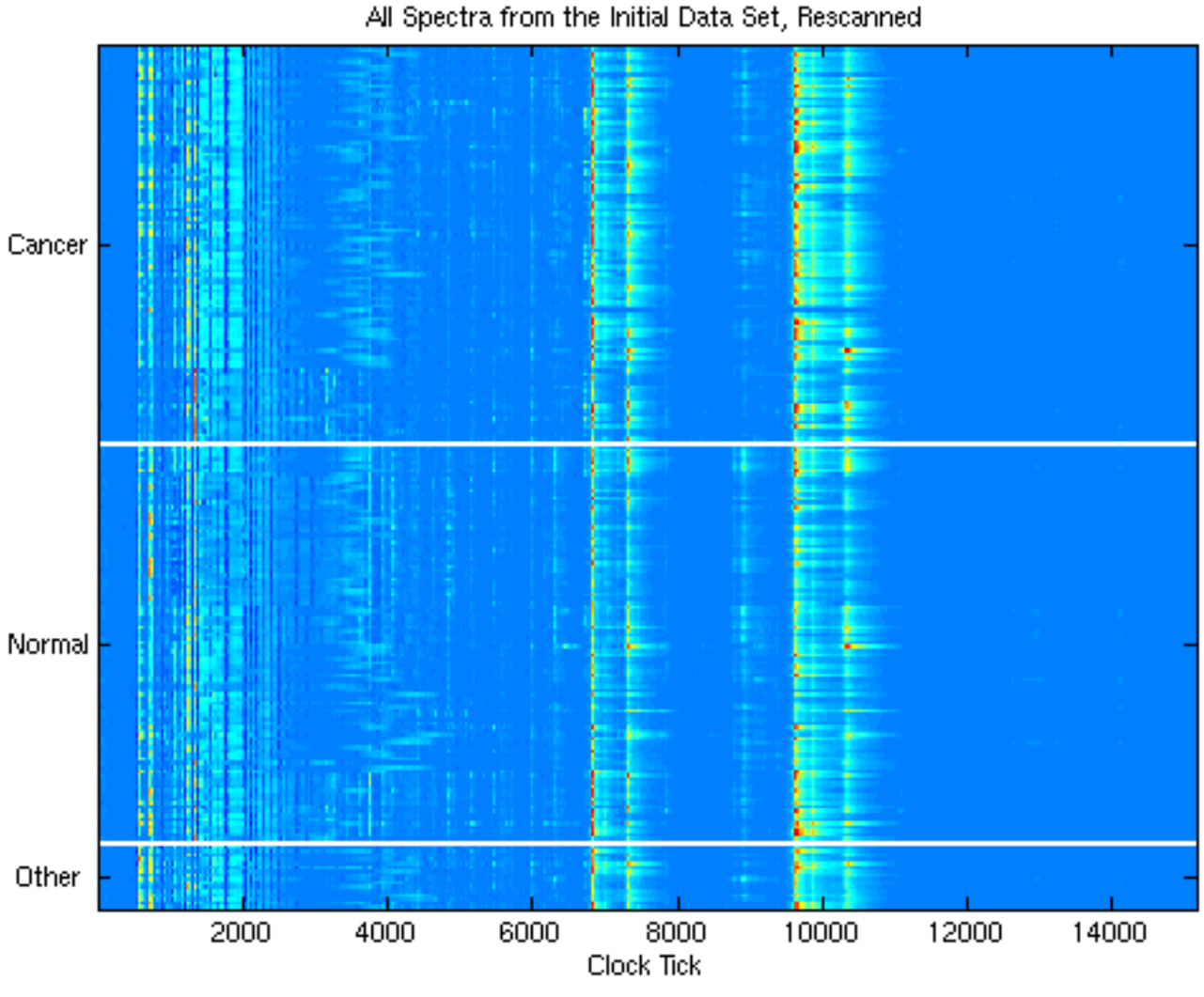
Zoomed in with Baseline Sub ON

Zoomed in with Baseline Sub OFF

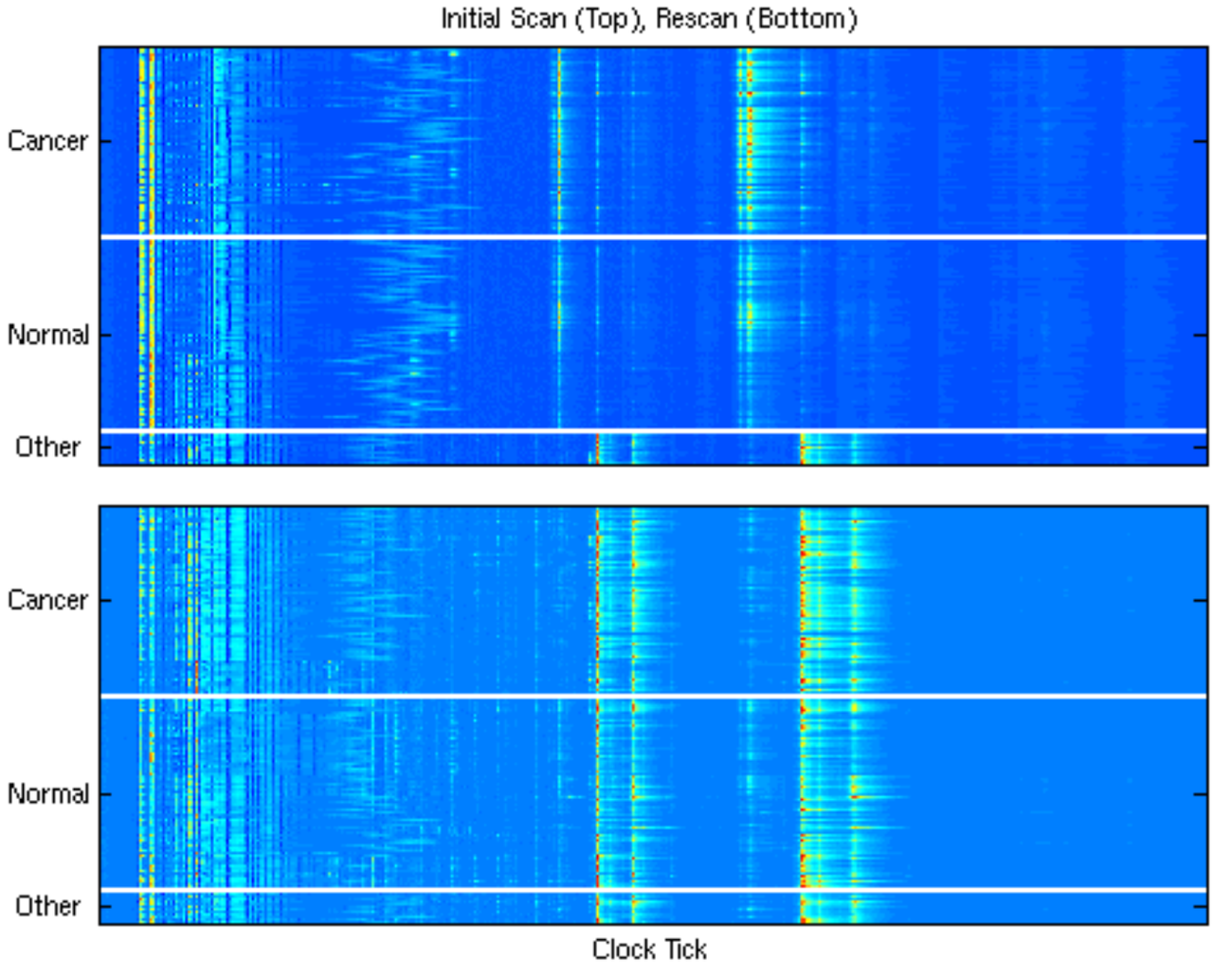
Some Structure is Visible in DS1



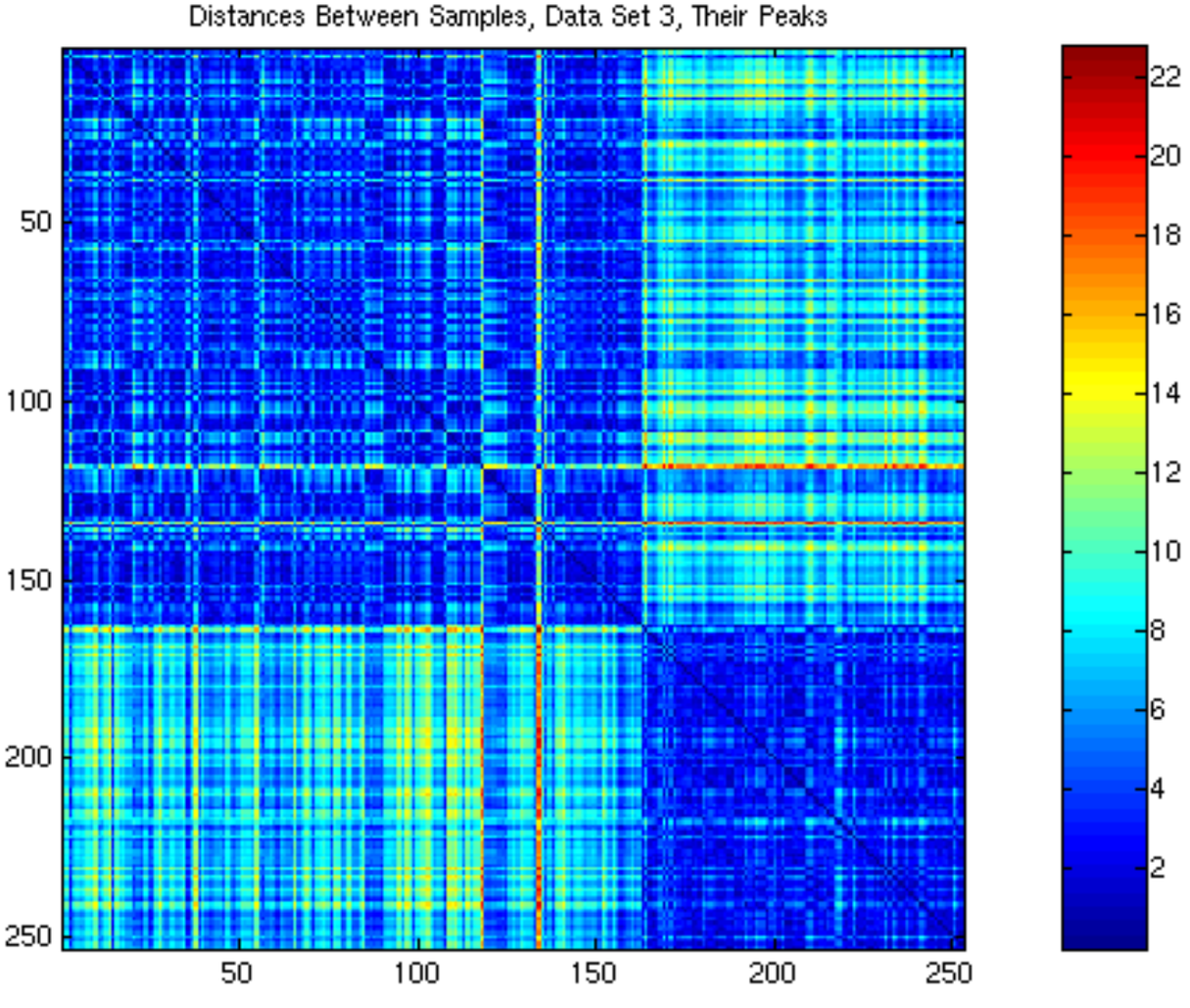
Or is it? Not in DS2



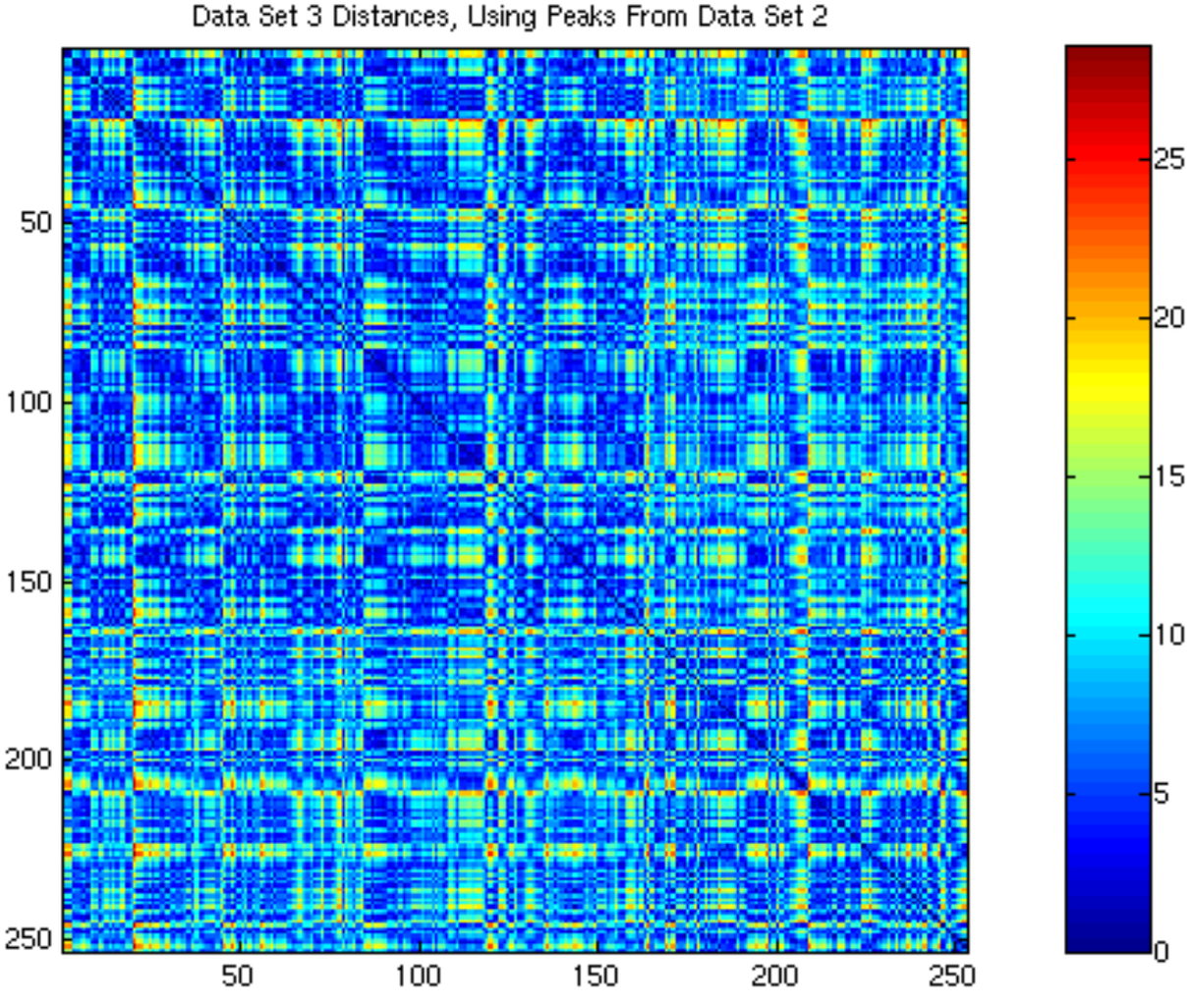
Processing Can Trump Biology (DS1 & DS2)



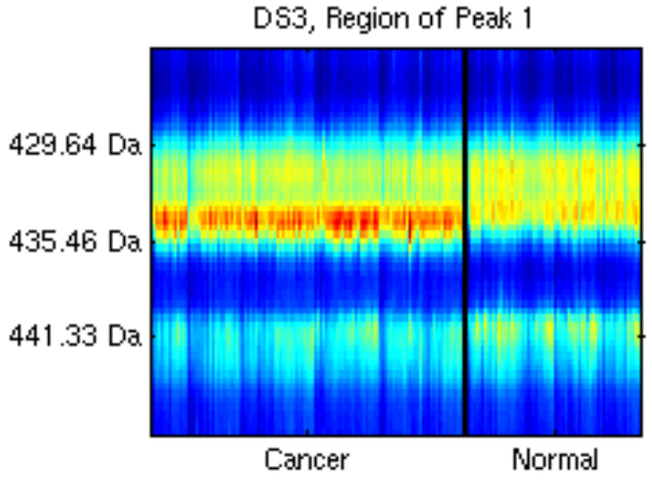
We Can Analyze Data Set 3!

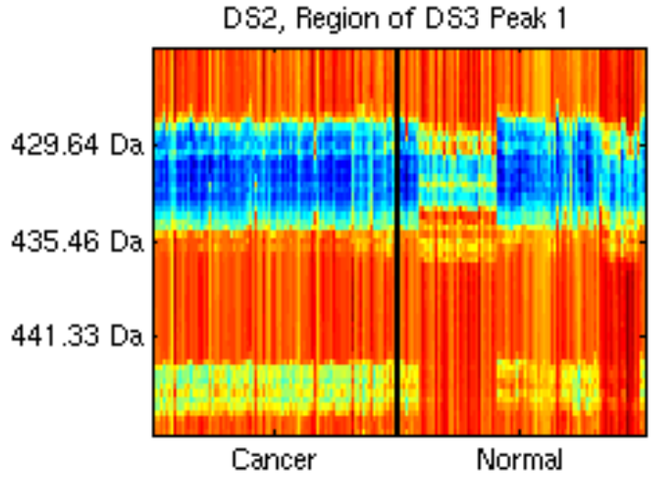


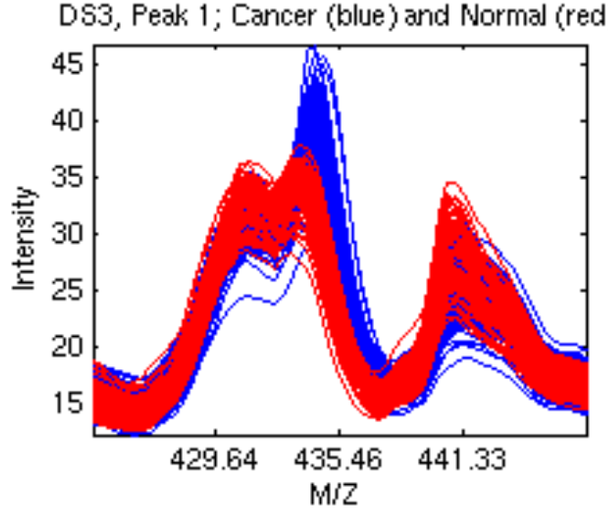
Do the DS2 Peaks Work for DS3?

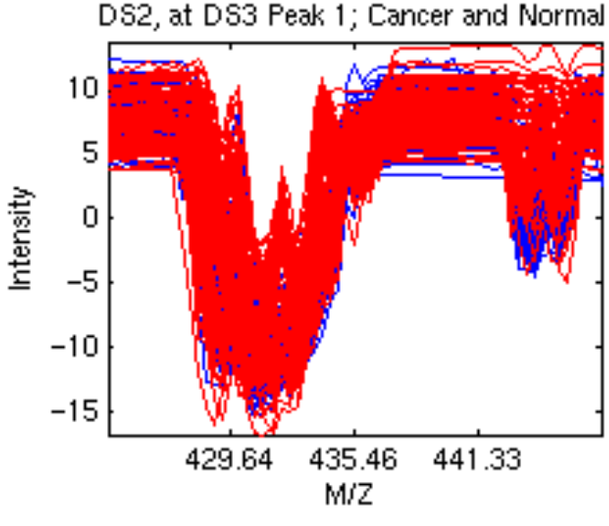


Do the DS3 Peaks Work for DS2?

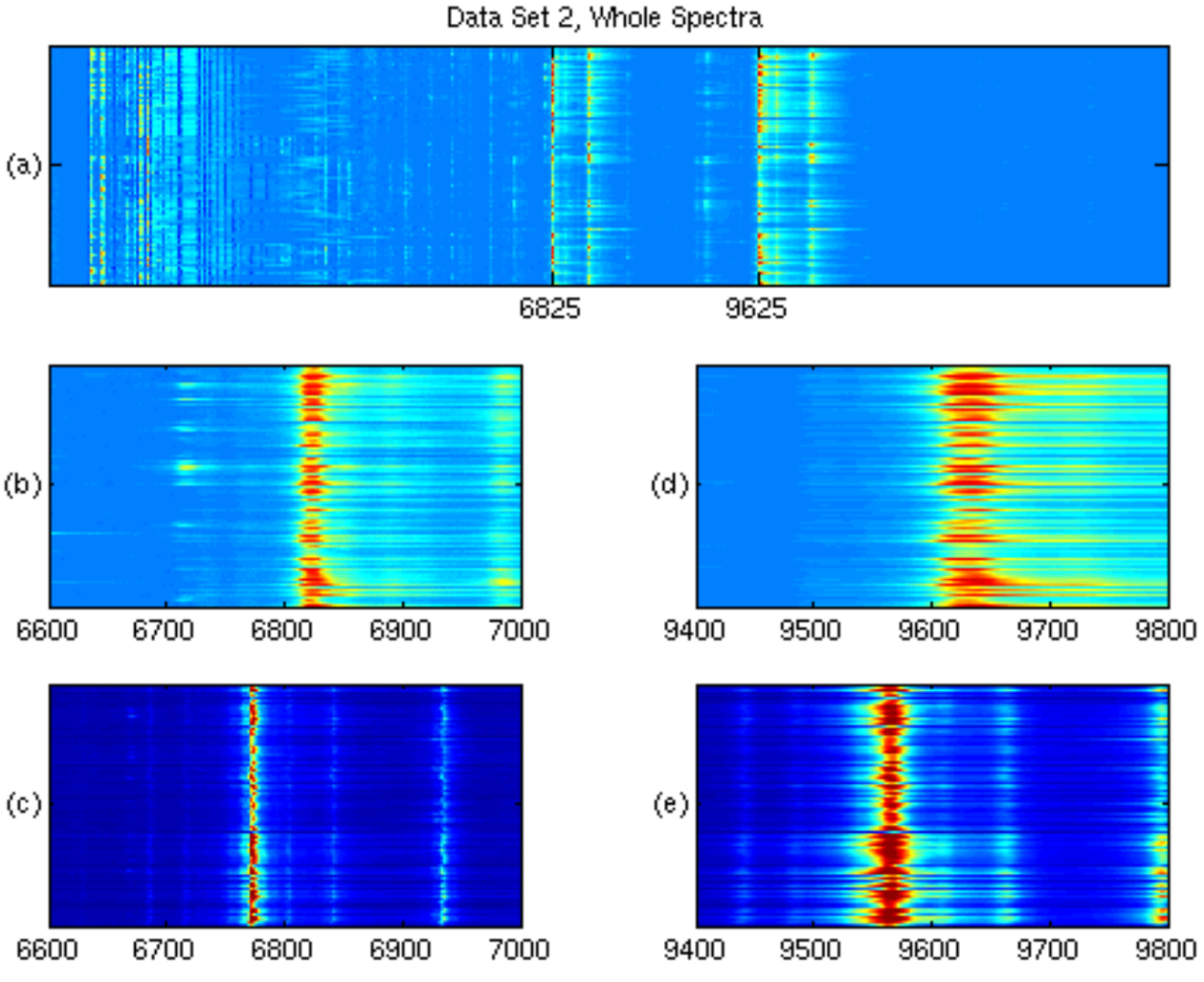




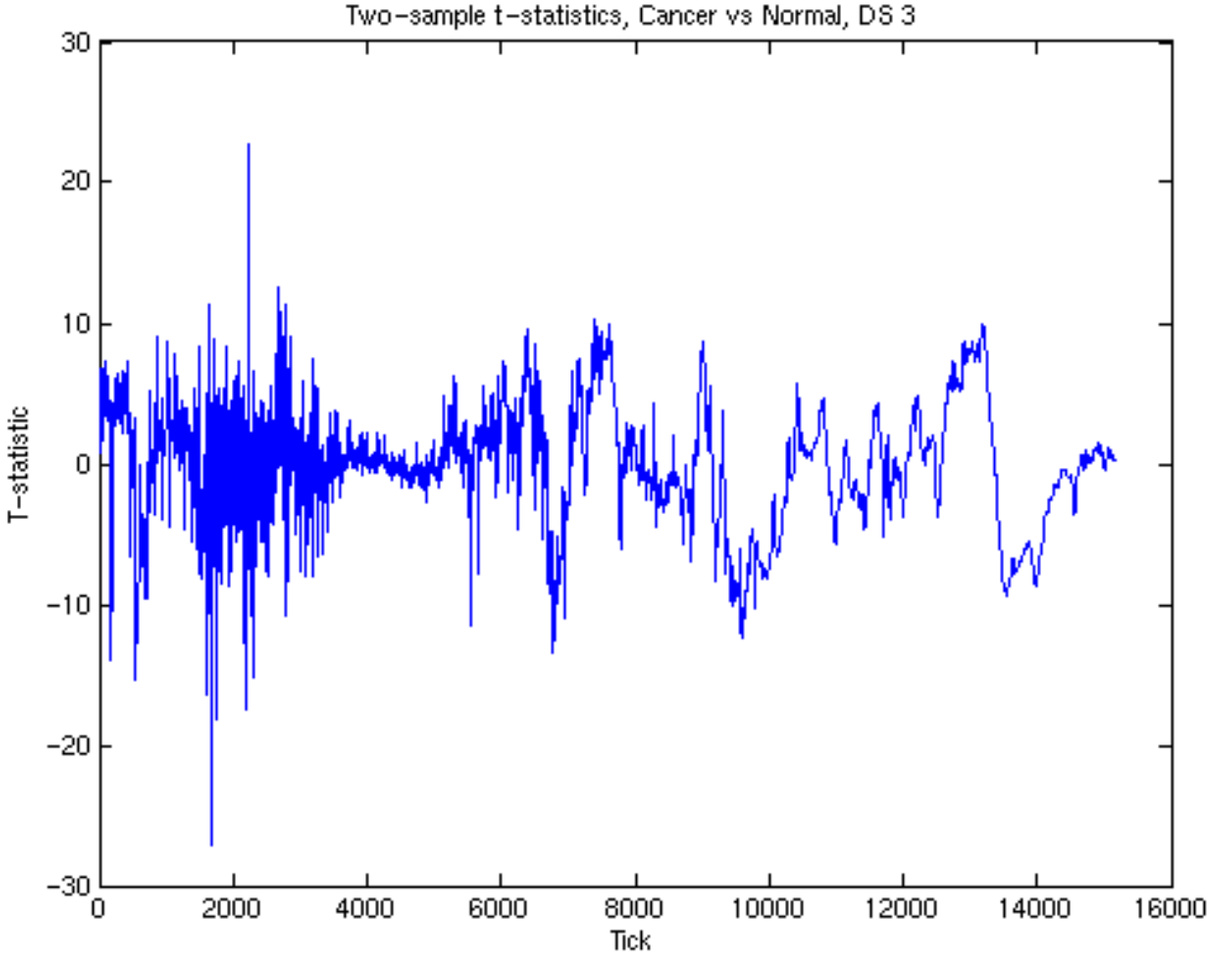




Peaks are Offset

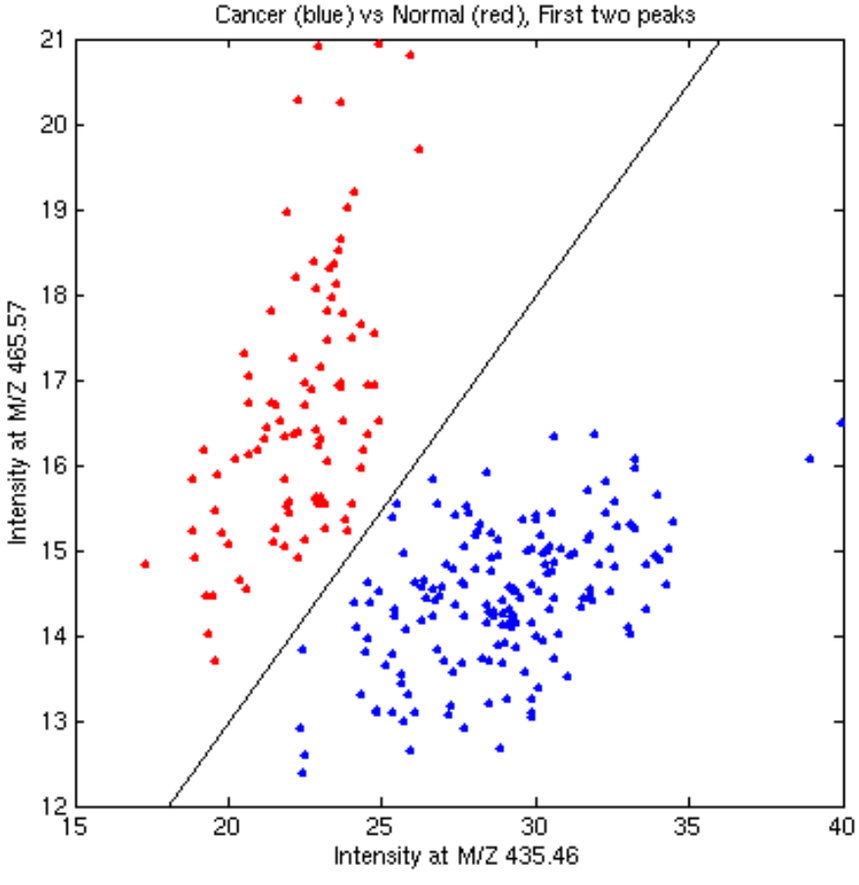


Which Peaks are Best? T-statistics



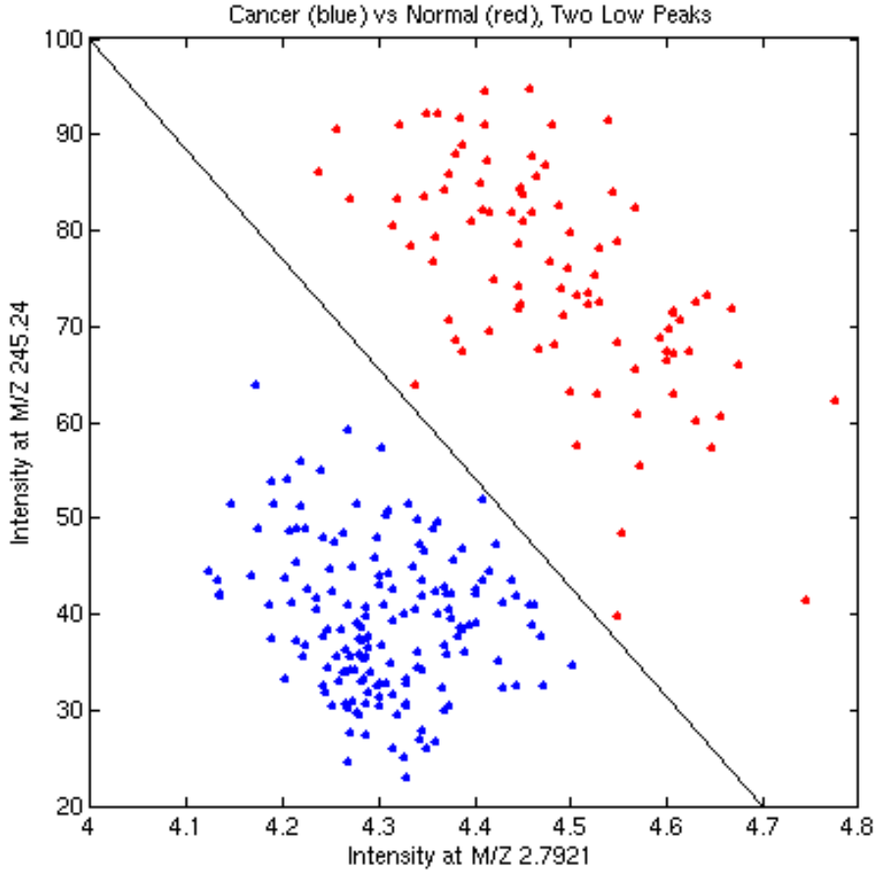
Note the magnitudes: t-values in excess of 20 (absolute value)!

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.

(Sorace and Zhan, BMC Bioinformatics, 2003)

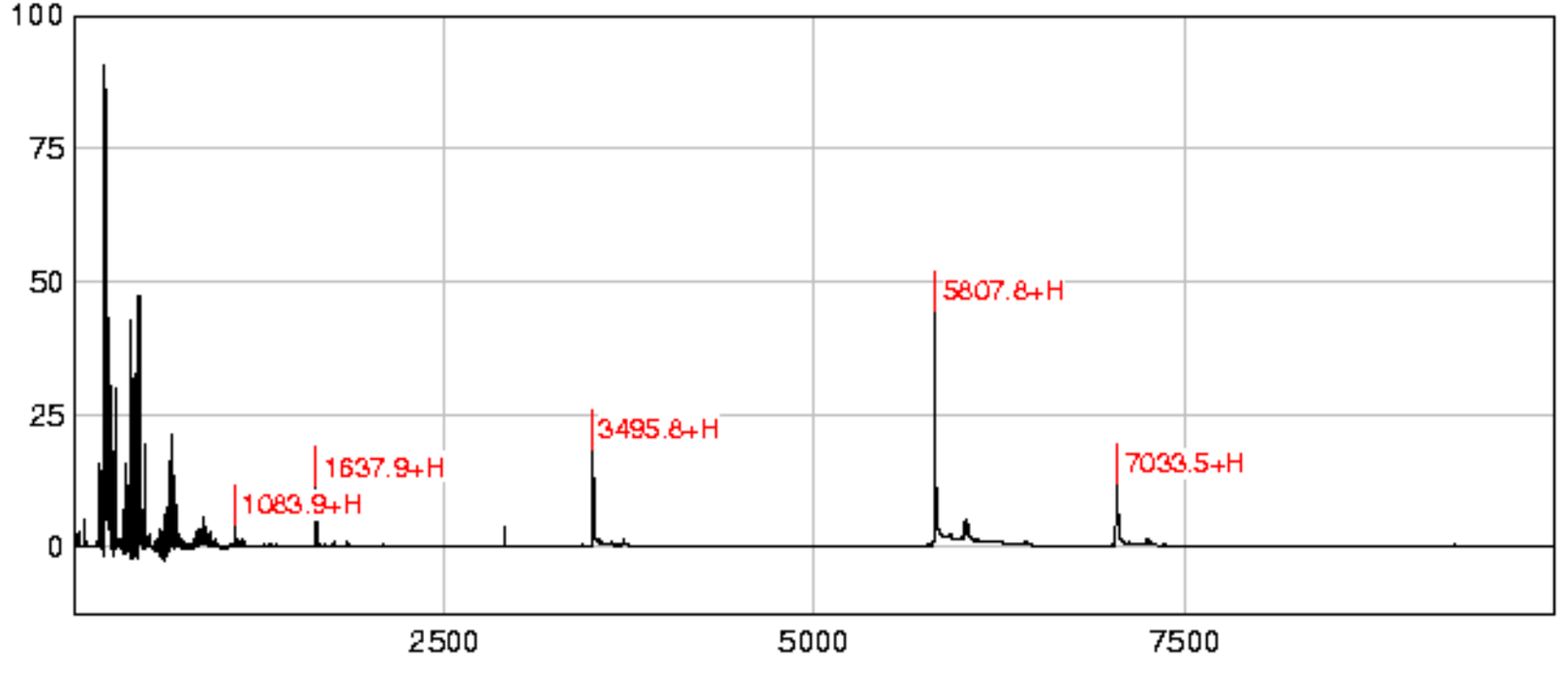
Mass Accuracy is Poor?

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

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

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!

Meanwhile...

In January 2004, Correlogic, Quest Diagnostics and Lab Corp announced plans to offer a “home brew” test called **OvaCheck**: samples would be sent in by clinicians for diagnosis.

Estimated market: 8 to 10 million women. Estimated cost: 100-200 dollars/test.

A Timeline

critiques available online, Jan 29

New York Times, Feb 3

Statement by the SGO, Feb 7

FDA letter to Correlogic, Feb 18

FDA letters to Quest and Lab Corp, Mar 2

editorials and features:
JNCI, Apr 7 & Jun 2, J Proteome Res, Apr 12, Nature Jun 3

The Response

Petricoin et al have written a rebuttal. This appeared as a comment on Sorace and Zhan (BMC Bioinformatics Mar ’04)

Several points. We focus on the two “most major”: (cited in Nature news feature Jun ’04).

1) Another group has found structure persisting across multiple data sets (DS2 & DS3), so our analysis is flawed.

2) Our focus has been on SELDI data, but it is their more recent high-resolution (Q-star) data that is the state of the art. We’re beating a dead horse.

Consistent Structure in DS2/DS3?

Recently, Zhu et al. (PNAS 2003,v100:14666-71) noted that they could find a set of m/z values that separated cancers from normals in both DS2 and DS3.

Use local smoothing (Gaussian kernel) and t-statistics to identify useful peaks in DS2. Keep only those with t-stats exceeding a certain magnitude threshold, 4.22, found using random field theory.

Choose a subset to get behavior on a training set (from DS2).

Final list of 18 m/z values.

We Were Somewhat Surprised...

10 of the 18 m/z values are less than 500 Da.

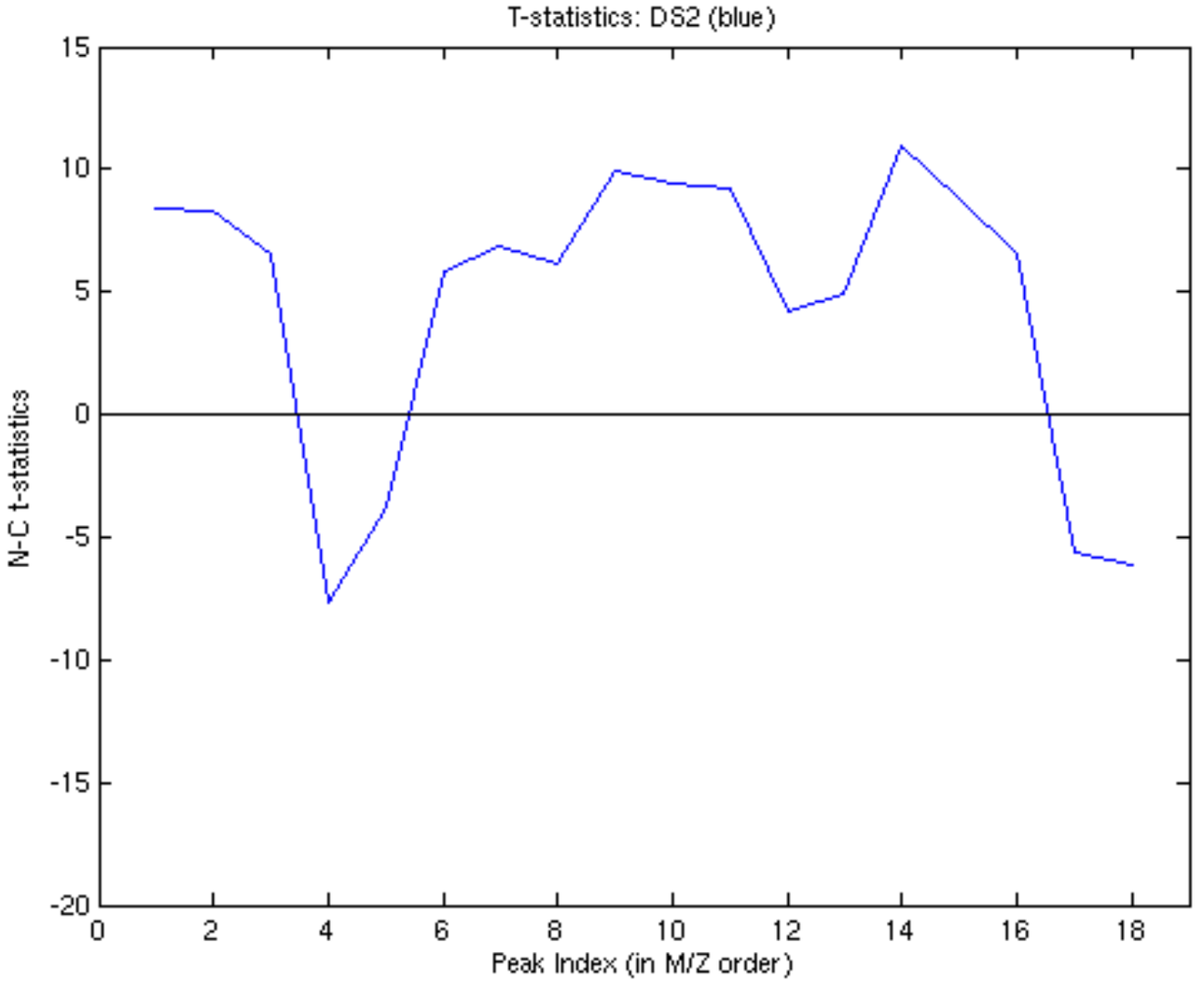
DS2 is baseline subtracted, and DS3 is not.

DS2 is offset relative to DS3.

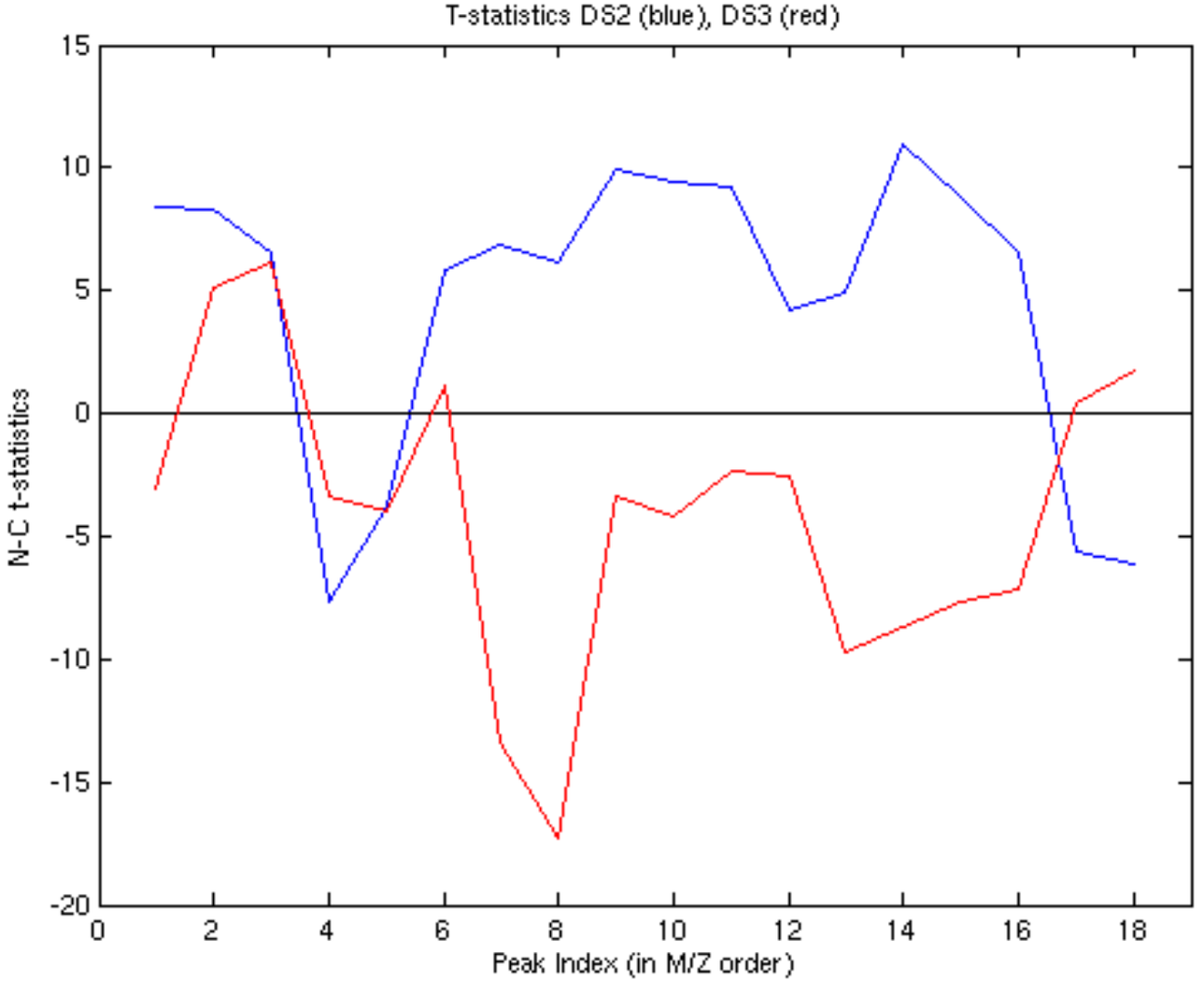
If the same nominal m/z values are used, we would think they are finding different proteins.

Try a simple test.

What Are the T-Statistics?



13/18 Flip Sign?!?



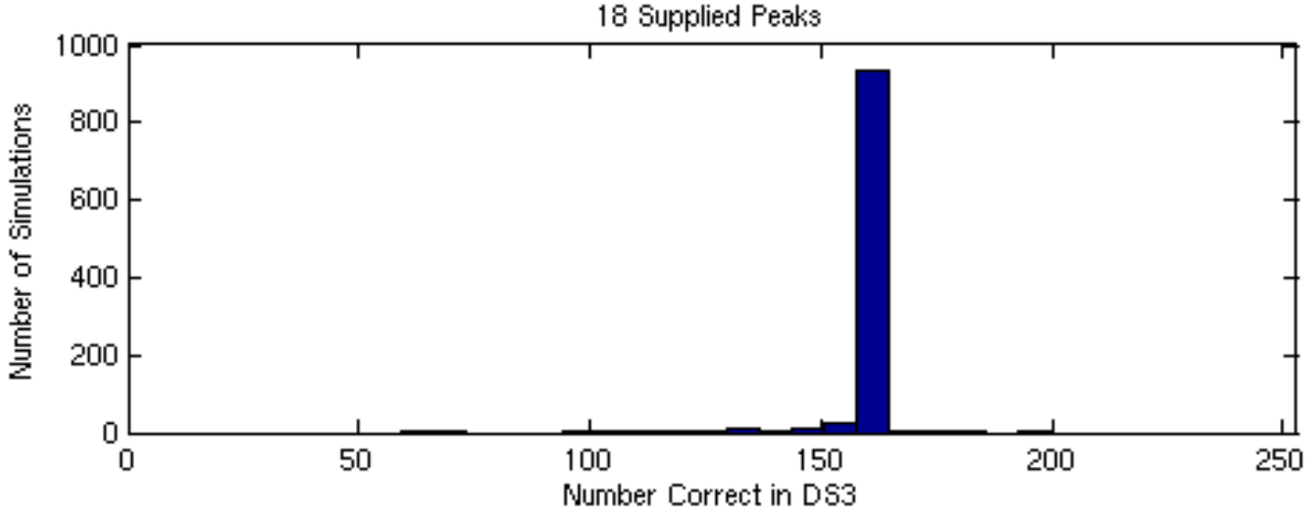
What Do Sign Flips Mean?

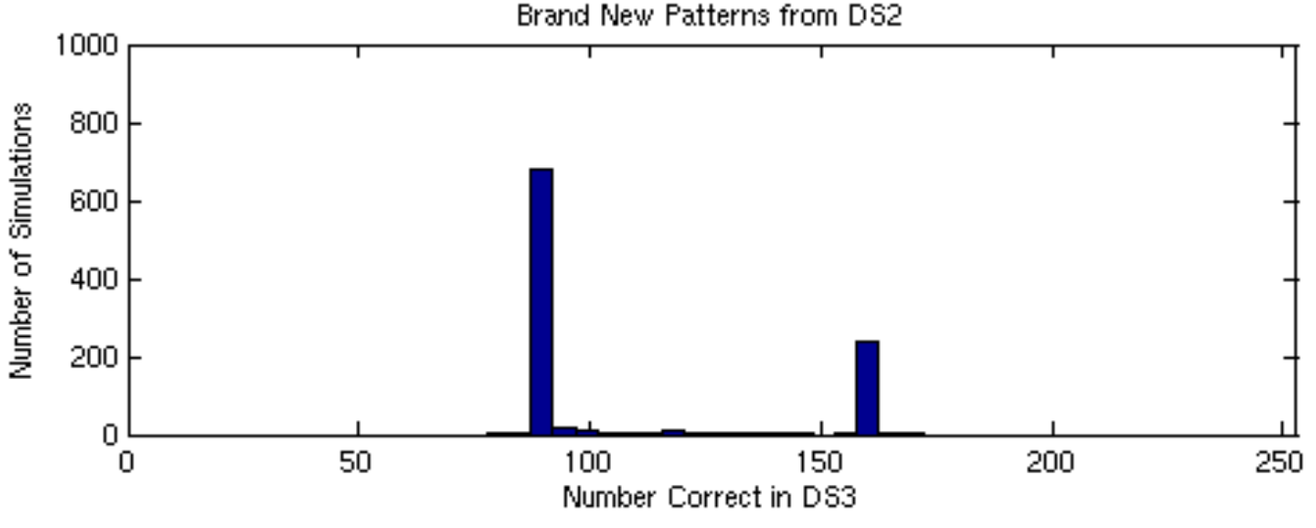
If intensities were higher in cancer in DS2, they're higher in controls in DS3.

This does not strike us as a consistent biological signature.

When simple and complex tests really disagree, question the complex test.

What Results Do We Get?





How was the Test Performed?

Given our results, we contacted Zhu et al.

There was a mixup.

> "the spectra in the second dataset were classified using a jack-knife approach where distances were computed between each spectrum and all of the other spectra in the second dataset, and the spectrum was classified according to the status of its 5 nearest neighbors in this set of spectra. Only the peak locations (m/z values) were retained across datasets, and these served to define the points at which the distances were computed. Further, the validation simulations used training sets drawn from the (August) dataset." – Wei Zhu, personal communication.

Why Did They Get Good Results?

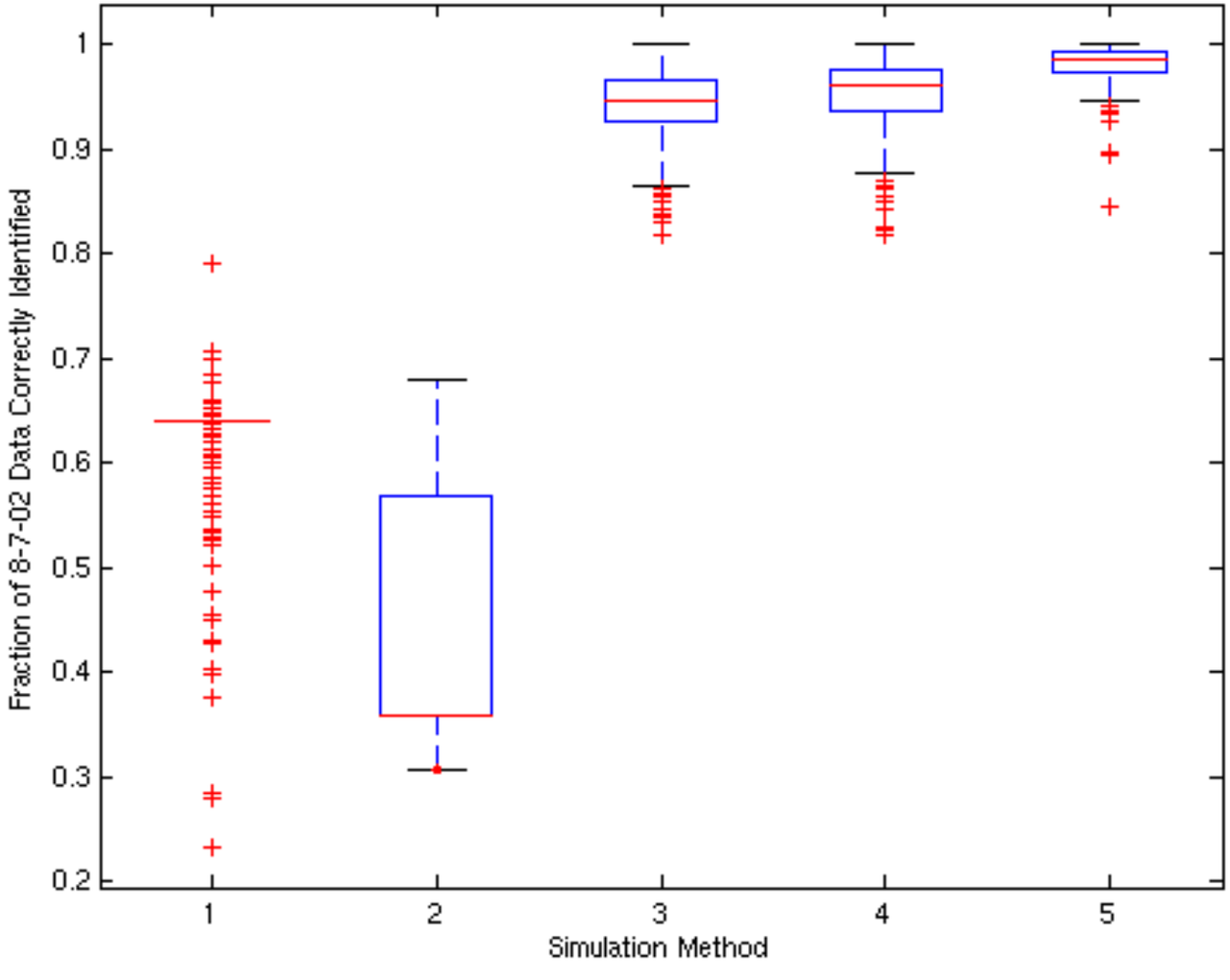
Even if they were chosen in an indirect way, these 18 m/z values work to separate DS2 and DS3. Does that mean that these values are “important”?

If the results look too good, try something that *shouldn’t* work.

DS3 is easy to classify.

Is it really easy?

Try *Random* Peak Sets



Classifying DS3 is Too Easy

We do just as good a job using randomly chosen m/z values.

This is disturbing, as it suggests that the differences between groups are *pervasive*, and if the differences were really that stark we should have seen them before.

Are We Beating a Dead Horse?

Qstar data is higher resolution.

They've added some QA/QC steps to remove bad spectra.

Still using patterns.

Reported results are even better.

Endocrine-Related Cancer (Jul '04) – 100% sensitivity and specificity.

What Does Qstar Data Look Like?

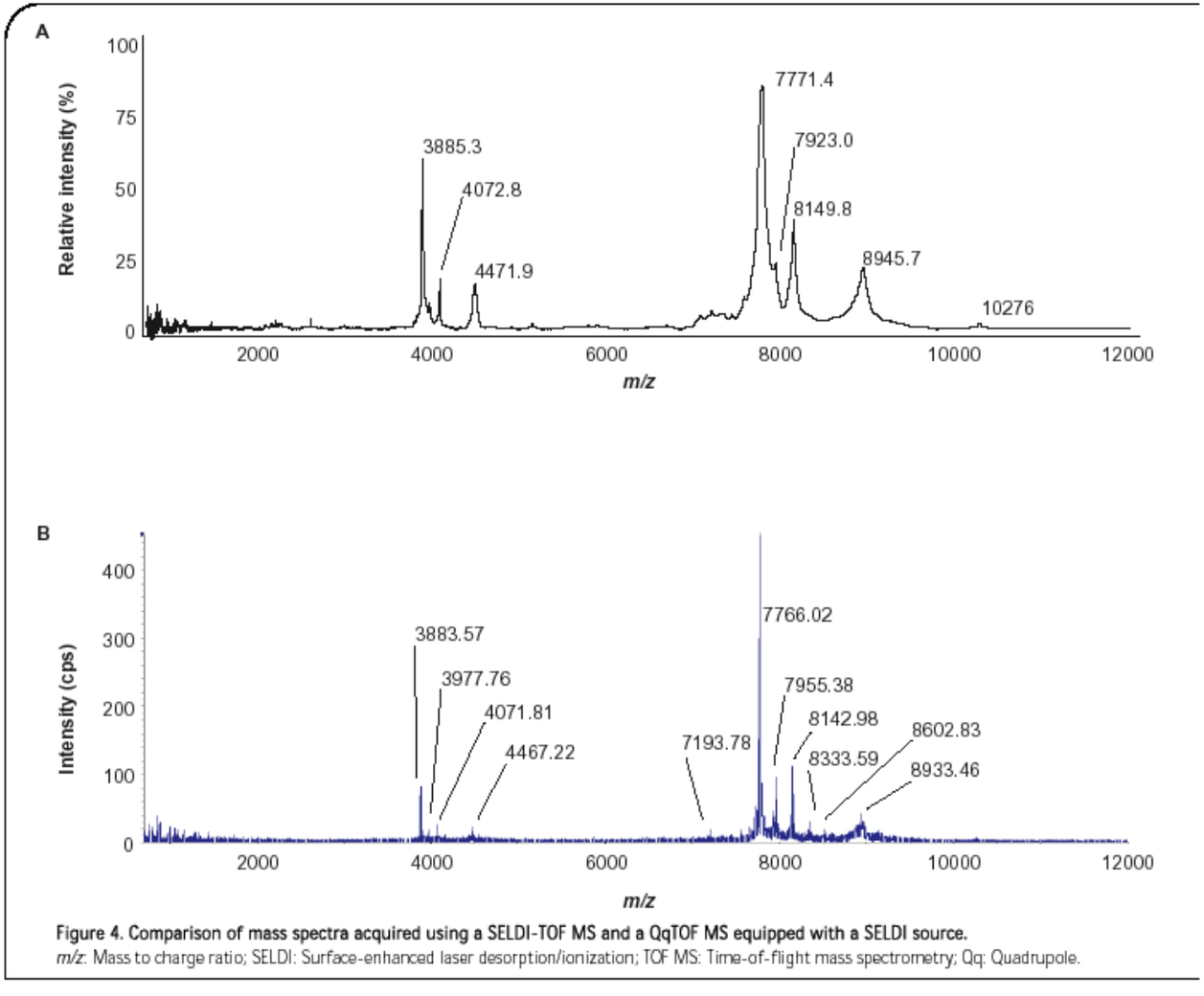
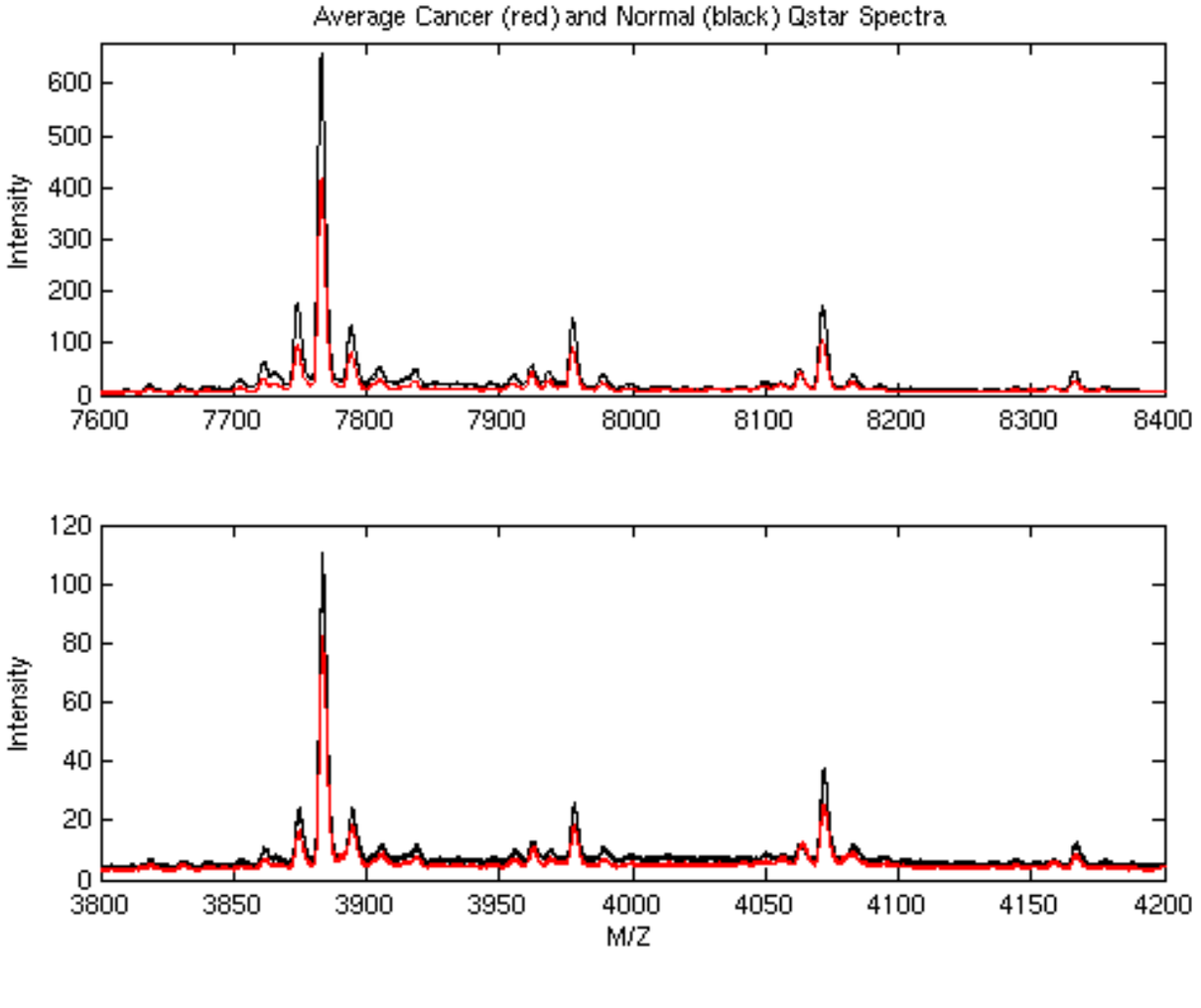


Figure 4. Comparison of mass spectra acquired using a SELDI-TOF MS and a QqTOF MS equipped with a SELDI source.
m/z: Mass to charge ratio; SELDI: Surface-enhanced laser desorption/ionization; TOF MS: Time-of-flight mass spectrometry; Qq: Quadrupole.

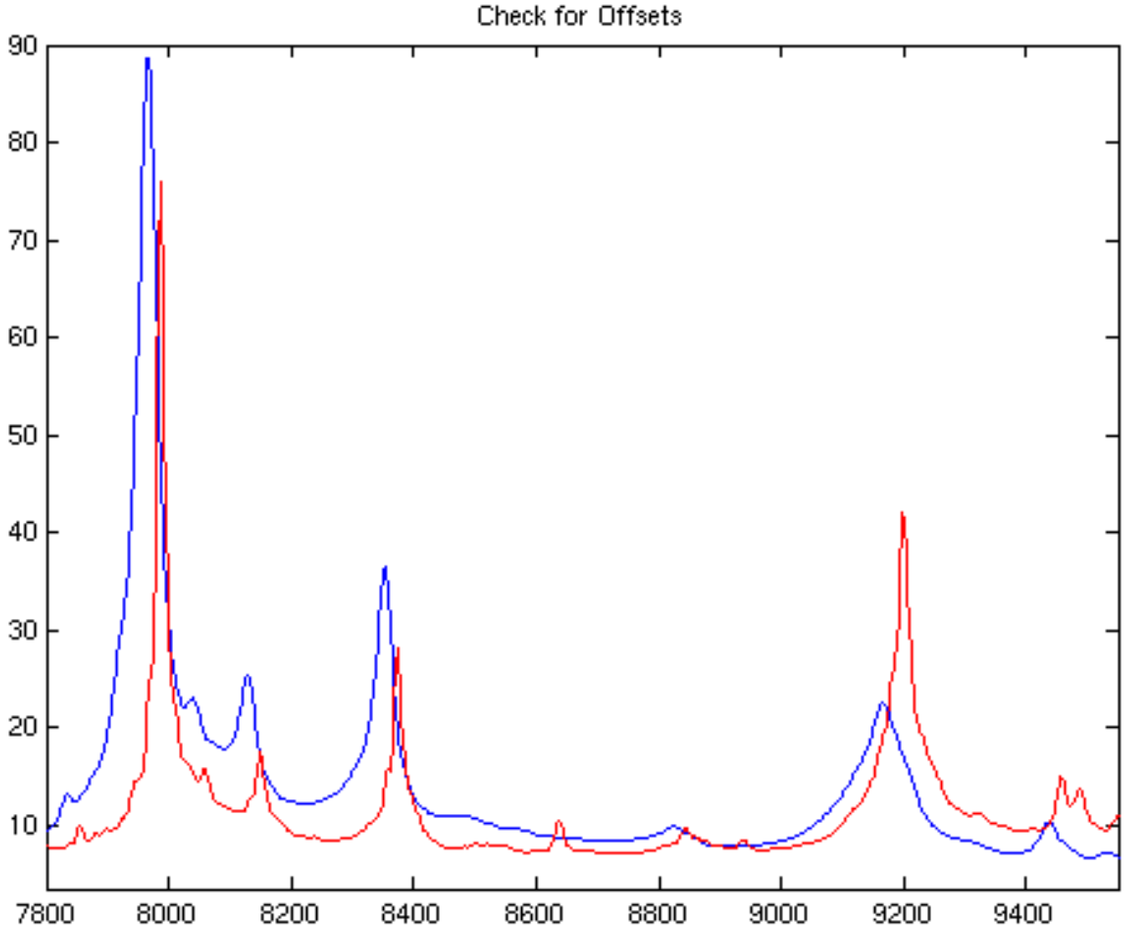
Conrads et al, Expert Rev Mol Diag (2003), 3, 411

Some Neat Structure: Doubling!



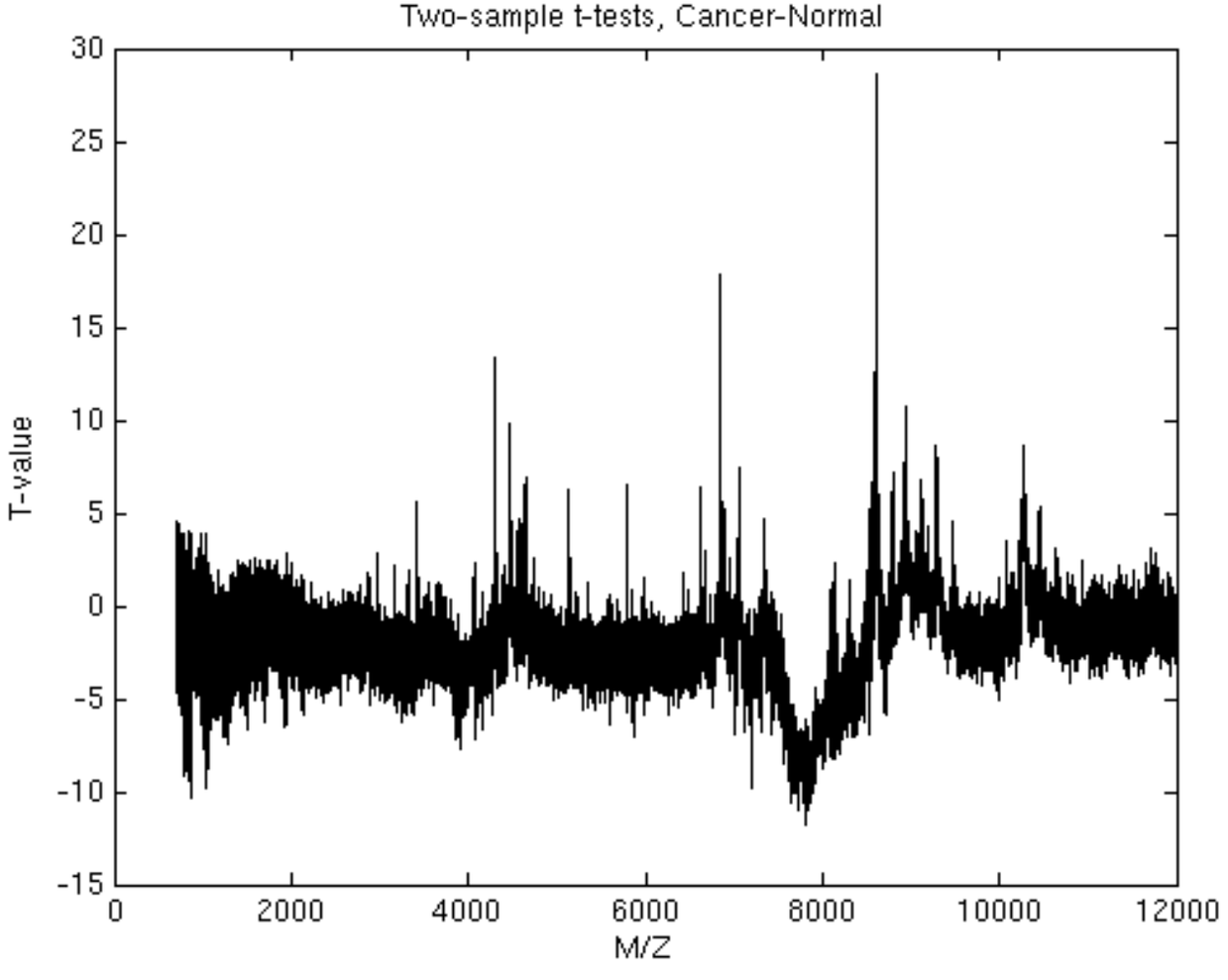
The calibration is correct!

SELDI Calibration En Passant



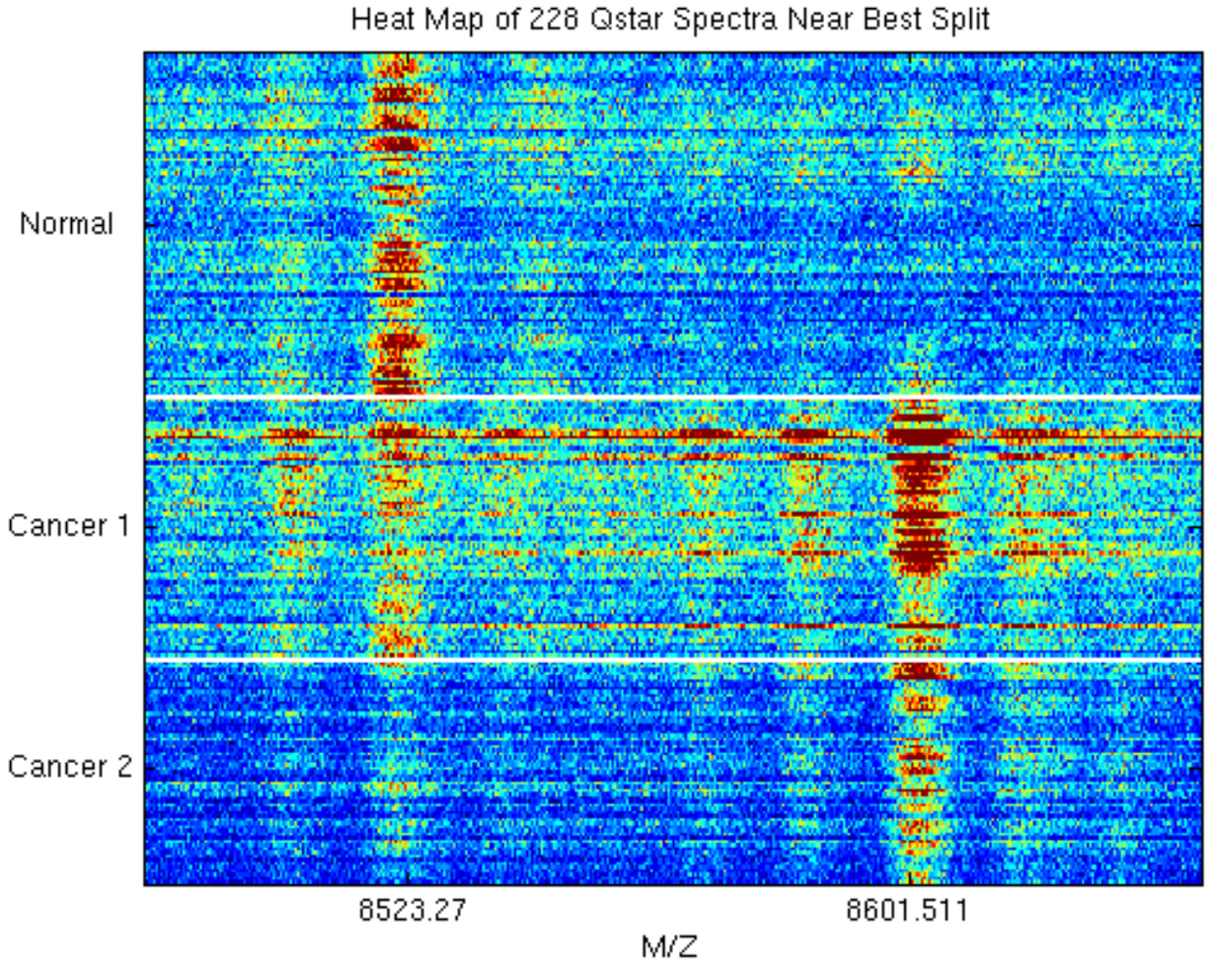
The calibration is incorrect! (between 2.5-4%).

Where are the Best Separators Here?



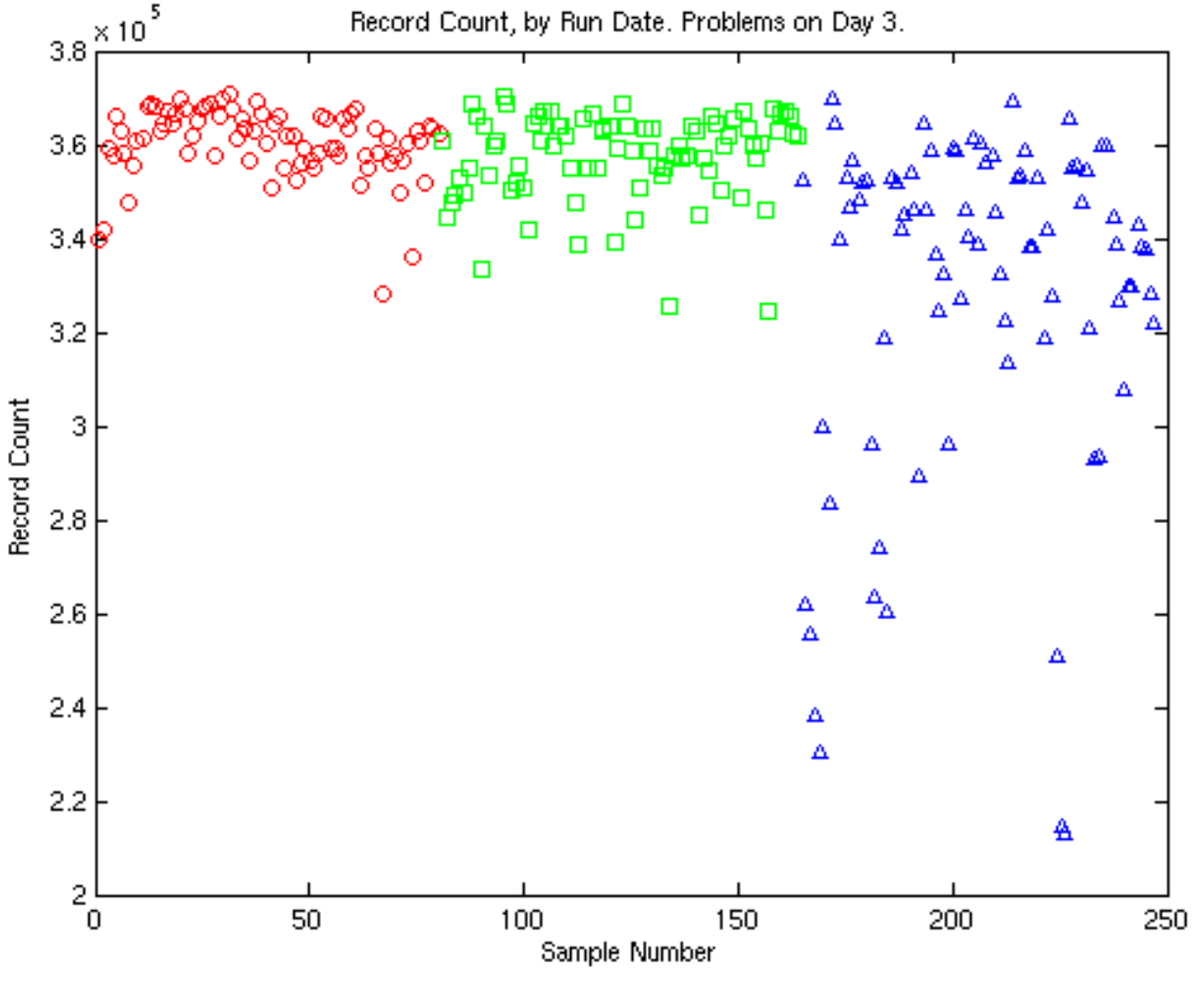
Best near 8602 Da.

Is it real? Looking at all the data



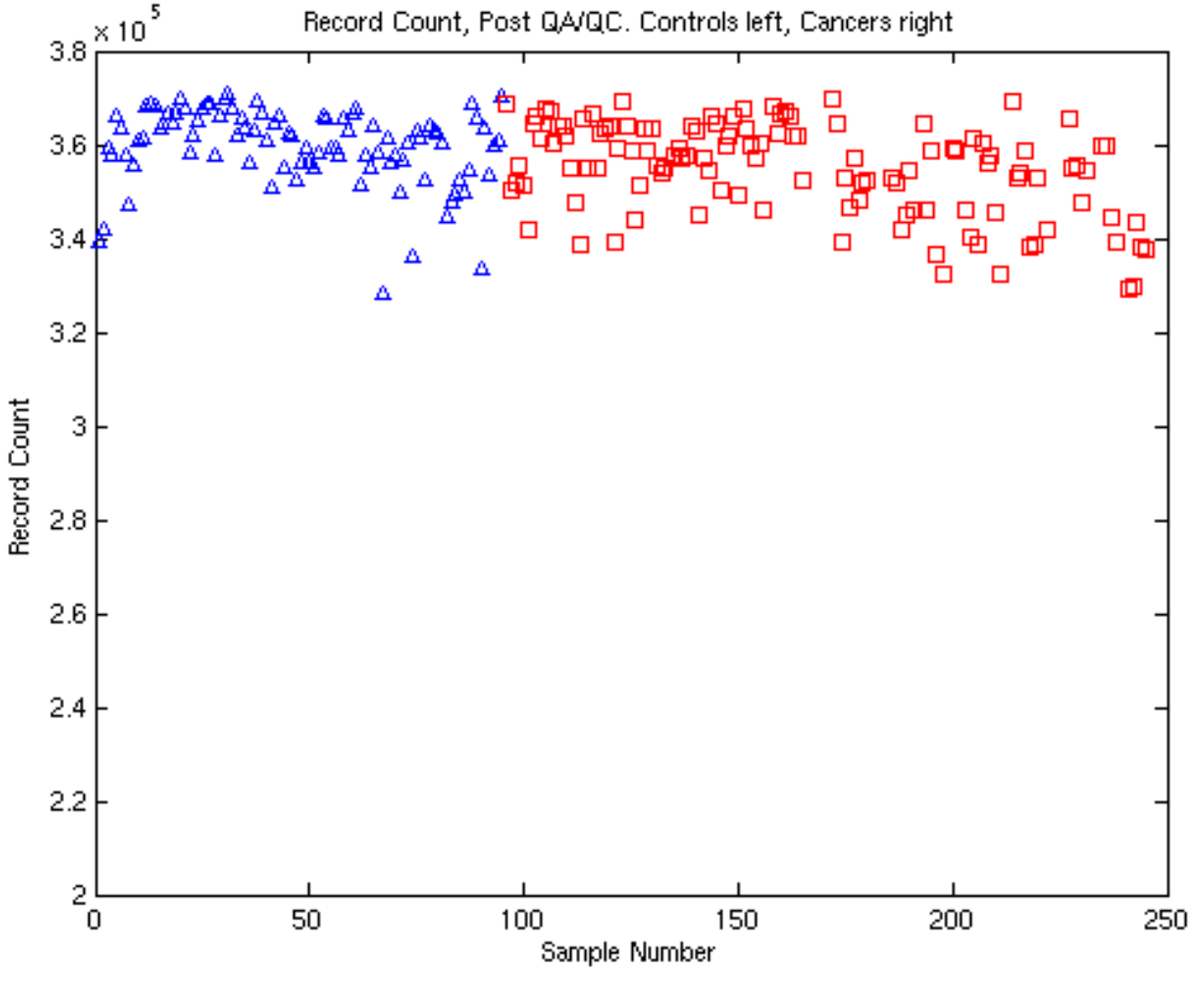
Heat map of the 8602 Da peak. Second peak 80 Da lower?

What's Going On? Part I



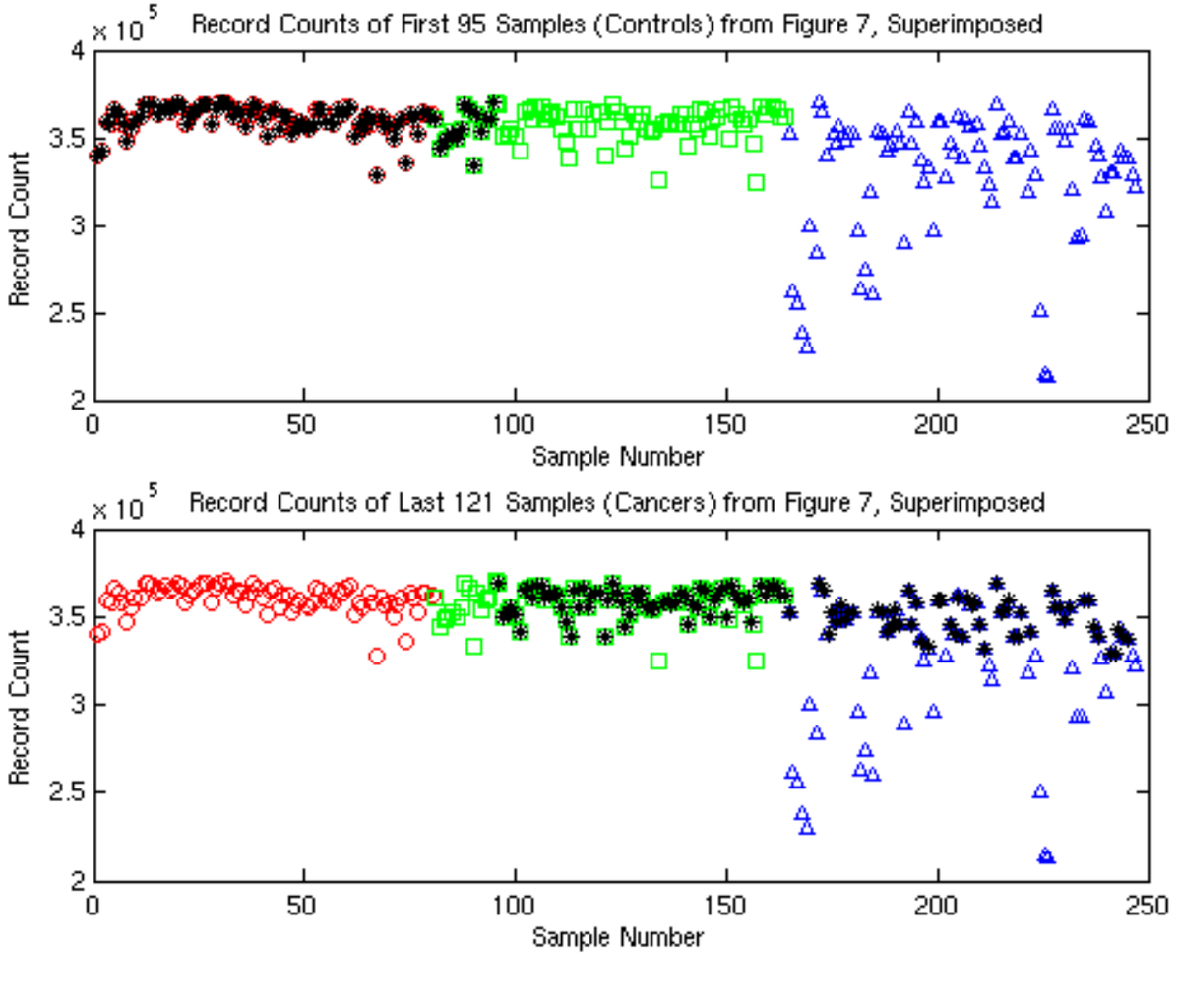
Conrads et al, ERC (Jul '04), Fig 6a

What's Going On? Part II



Conrads et al, ERC (Jul '04), Fig 7

What's Going On? Part III



Conrads et al, ERC (Jul '04), Fig 6a & 7

That Horse Looks Alive...

All of the controls were run before all of the cancers.

Given the time trend in the data, this biases the results – the cancer samples were more affected by the worsened problem on Day 3.

A better machine will not save you if the experimental design is poor.

What's up with the data?

Petricoin et al, March 04: "If the authors had contacted us, we could have elaborated, as previously stated on our website, that the SELDI-TOF MS data was produced by randomly commingling cases and controls. On any given 8 spot ProteinChip array, both cases and control samples were applied in random spot locations, to minimize systematic chip-to-chip variation."

Liotta et al, Feb 05: "In contrast, the goal of data set 8-7-02 was to determine whether the between-day process variance was more or less than the variance between the case and control groups. For data set 8-7-02, case and control samples were run in separate batches on separate days and not commingled."

So, Are We Ready for Prime Time?

If this means telling a woman that she needs an oophorectomy based on these tests, then no way.

In July of 2004, the FDA ruled that OvaCheck could not be made available under the homebrew exemption as the program was a “device” that needed to be more tightly regulated.

What About Proteomic Profiling in General?

Our results DO NOT say that the approach can't work. (There are some other arguments on that front.)

Our results DO say that experimental design and randomization are called for at the processing as well as at the collection stages.

Some well-designed studies have appeared (eg, Zhang et al, Cancer Res, 2004) and do show some gains in predictive accuracy.

and There's Fun Stuff Beyond Design!

Processing spectra to find peaks

Calibration of spectra

Baseline subtraction and Normalization of spectra

Deconvolution of spectra to account for specifics of MS structure (multiple charge states, common modifications)

Separating cancers from controls

and this goes for other types of MS as well: LC-MS, FT-ICR, 2d-gels