# Statistical Approaches for Proteomic Biomarker Discovery

Jeffrey S. Morris

**Department of Biostatistics**

**The University of Texas**

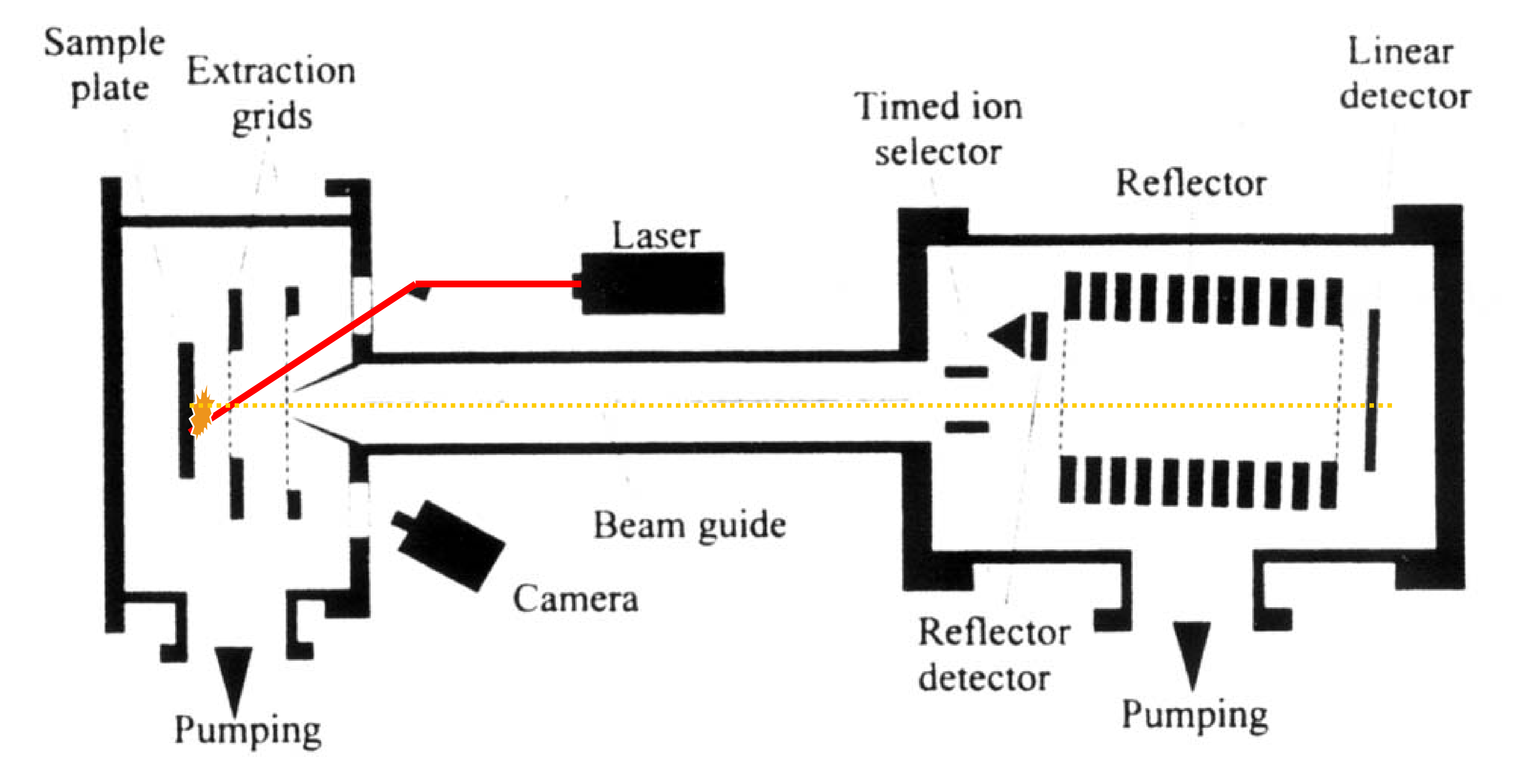
**MD Anderson Cancer Center**

**Houston, Texas**

# Outline

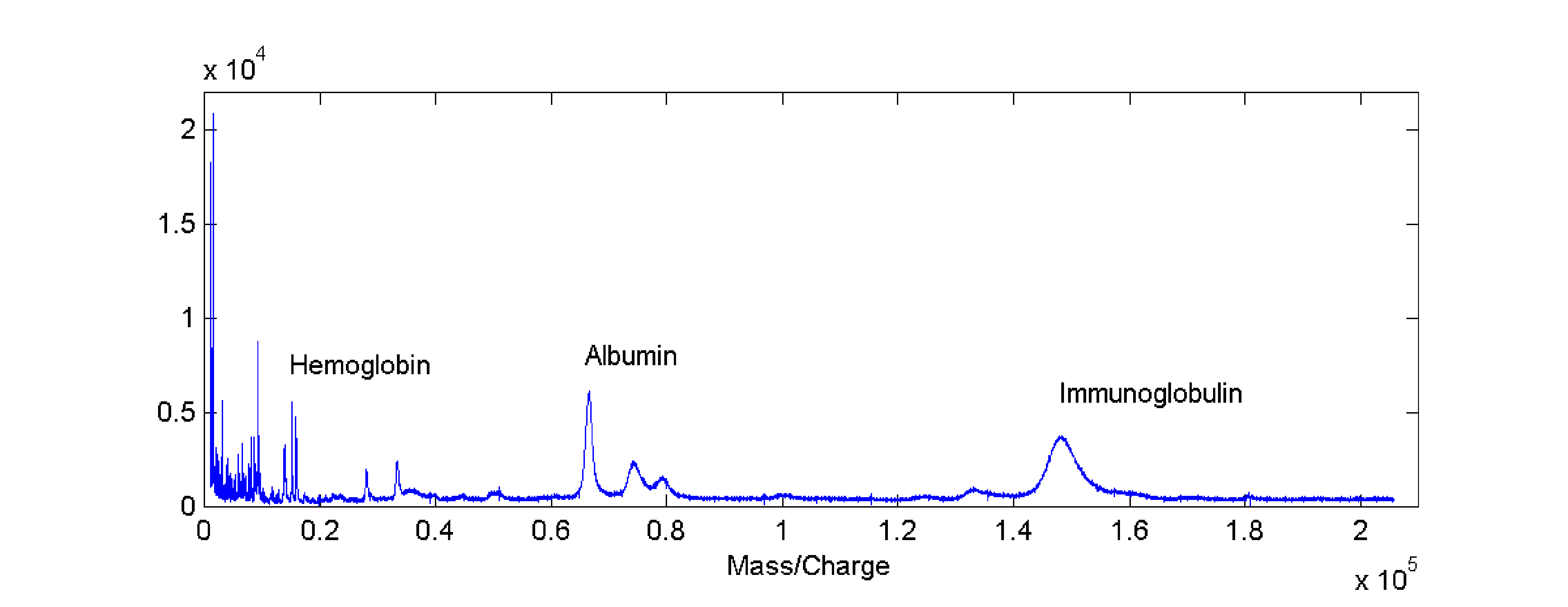
- Introduction to Proteomics
  - Proteomics vs. Genomics vs. Transcriptomics
  - Proteomic Methods: MALDI-MS/SELDI-MS/2DE
- Experimental Design Issues in Proteomics
- Feature Extraction Approach
  - Peak/Spot Detection via Avg Spectrum/Gel
  - Class comparison and Class prediction
- Functional Data Analysis Approach
  - Model/Inference
- Conclusions

# MALDI-TOF schematic



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

# Sample MALDI-TOF Spectrum



- g(*t*) = intensity of spectrum at m/z value *t*
- Intensity at peak (roughly) estimates the abundance of some protein with molecular weight of *t* Daltons

# Simulated spectra

- To study MALDI-TOF, and compare methods for analyzing them, we gave developed a simulation engine to produce realistic spectra (Coombes, et al. 2005)
  - – 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
  - – Also very instructive in how MALDI-TOF works, and why the data look the way they do.

# 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($\mu$,$\sigma$)

  $v_1$ = velocity after focusing

  $\delta$= delay time

- Equations

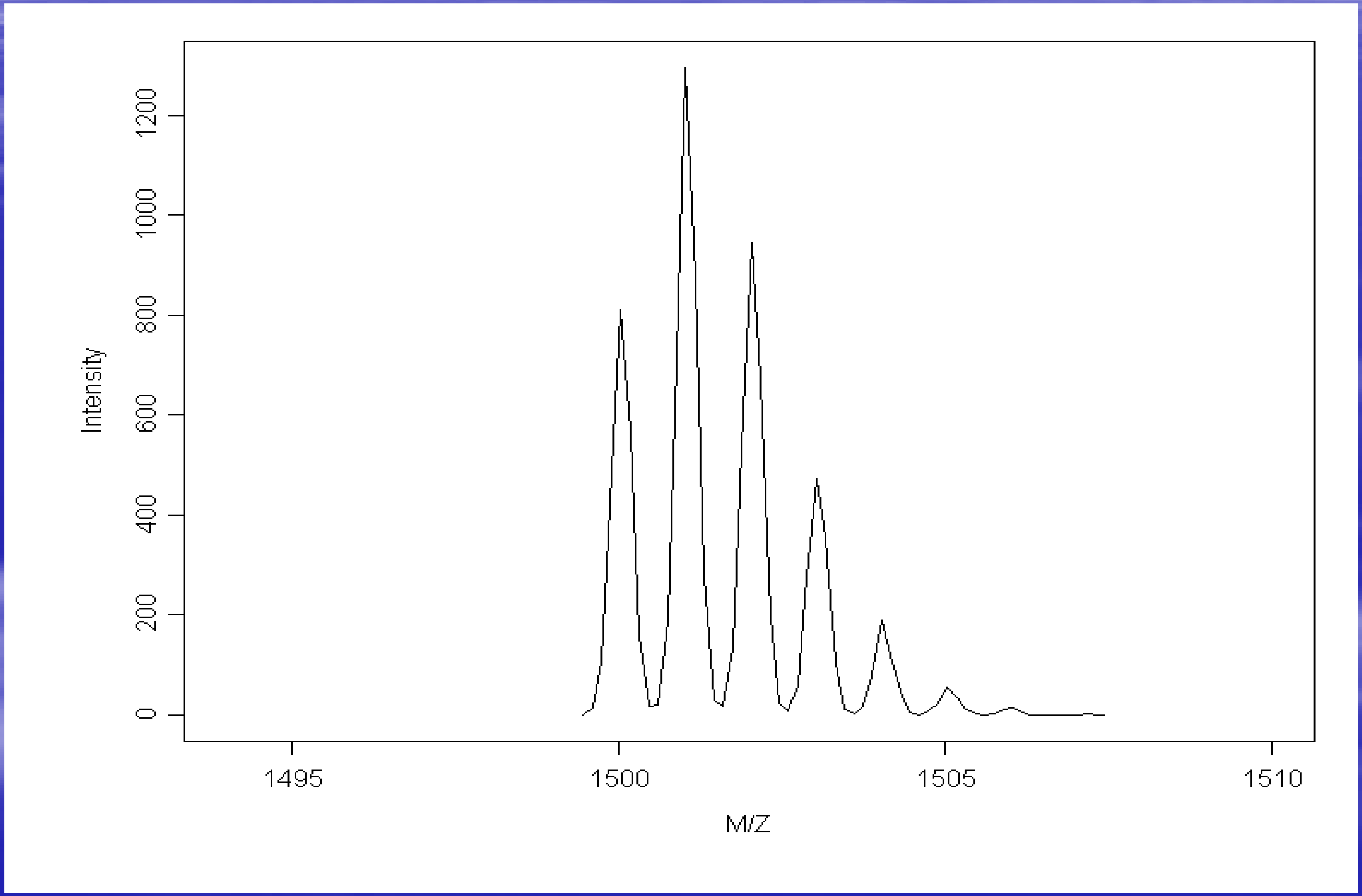
$$v_1^2 = v_0^2 + \frac{2qV_1}{mD_1}(D_1 - \delta v_0)$$

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

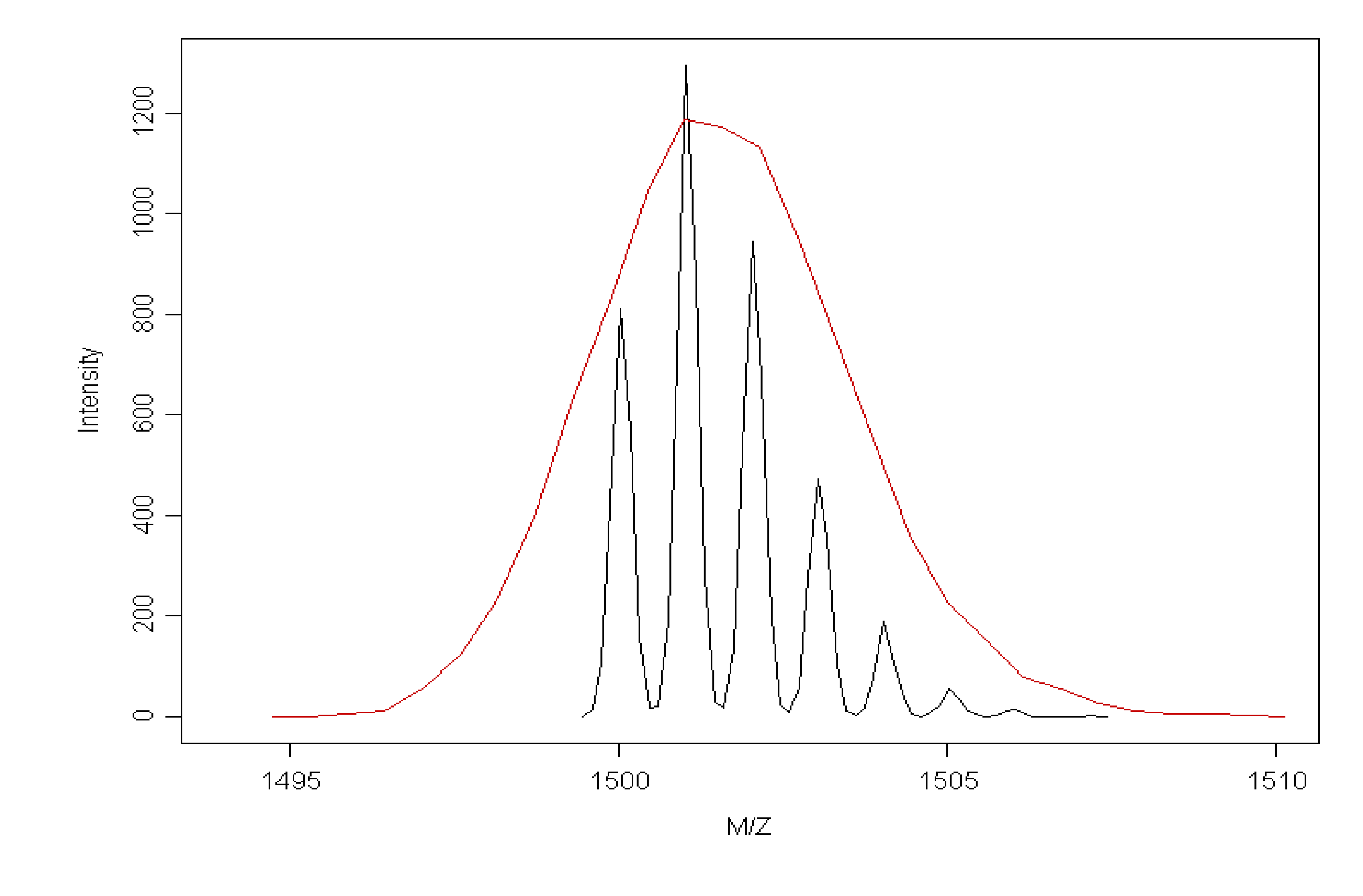
$$t_{ACCEL} = \frac{mD_2}{qV_2}\left( \frac{L}{t_{DRIFT}} - v_1 \right)$$

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

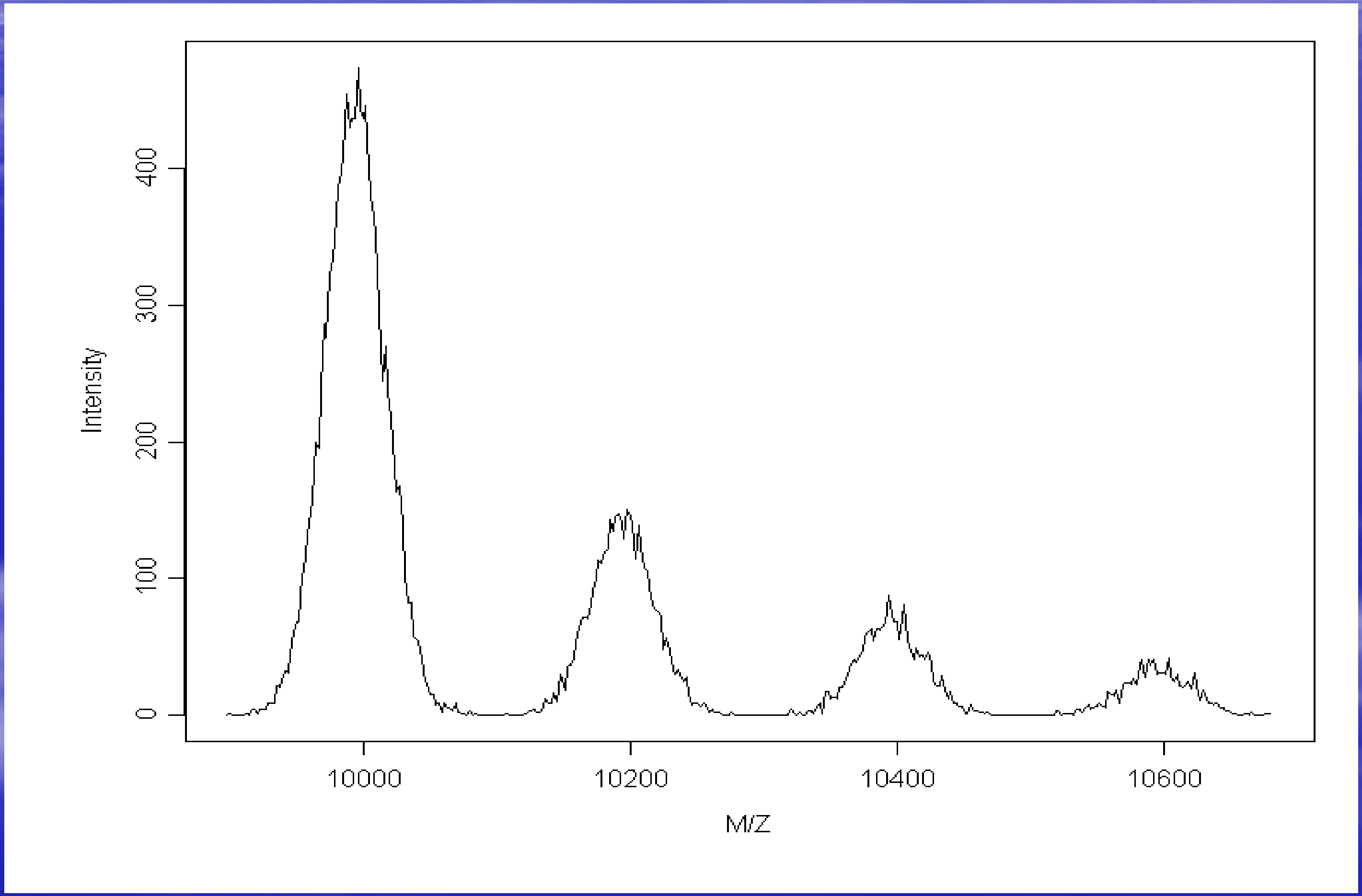
# Simulation of one protein, with isotope distribution



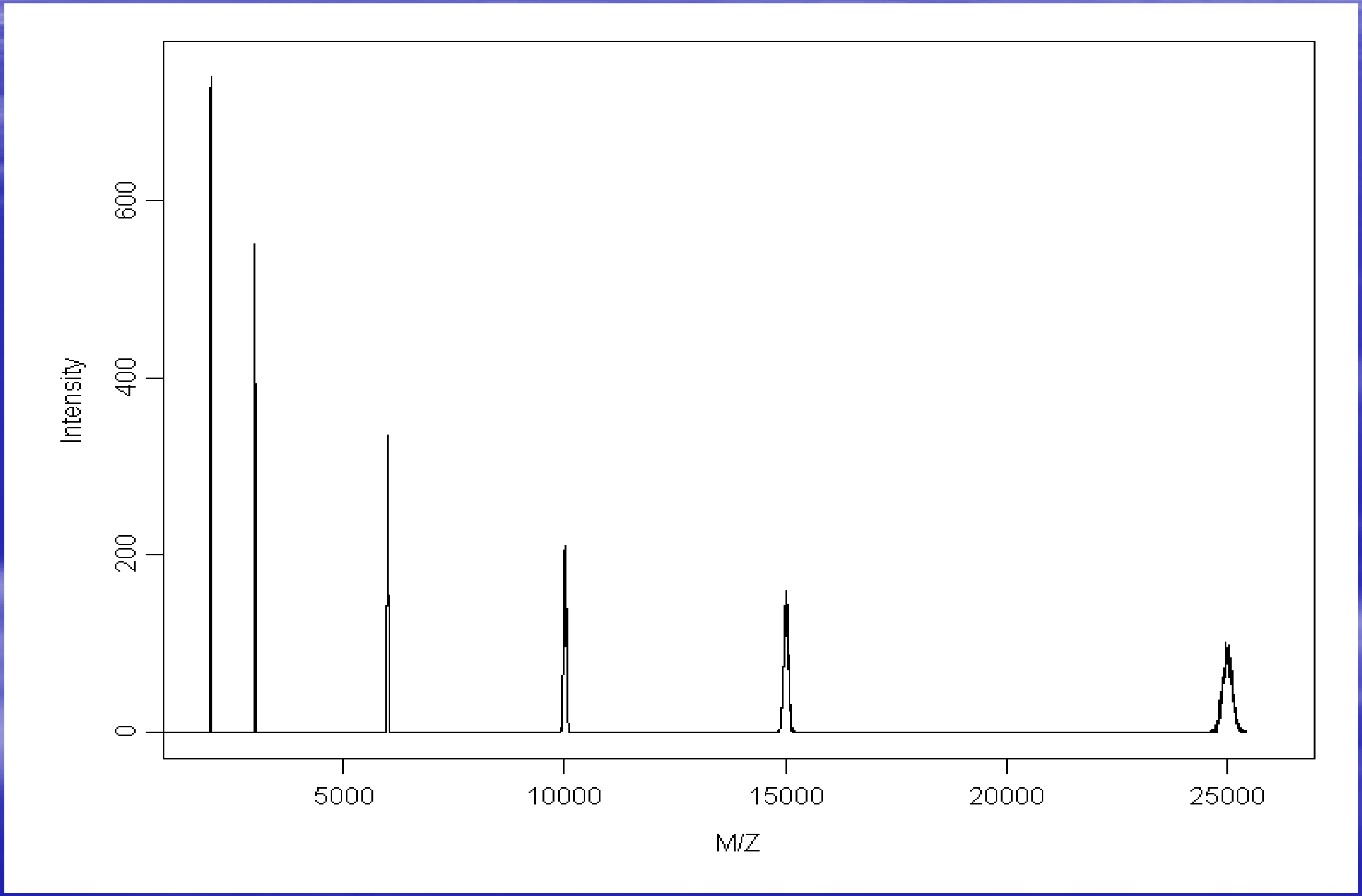
# Same protein simulated on a low resolution instrument



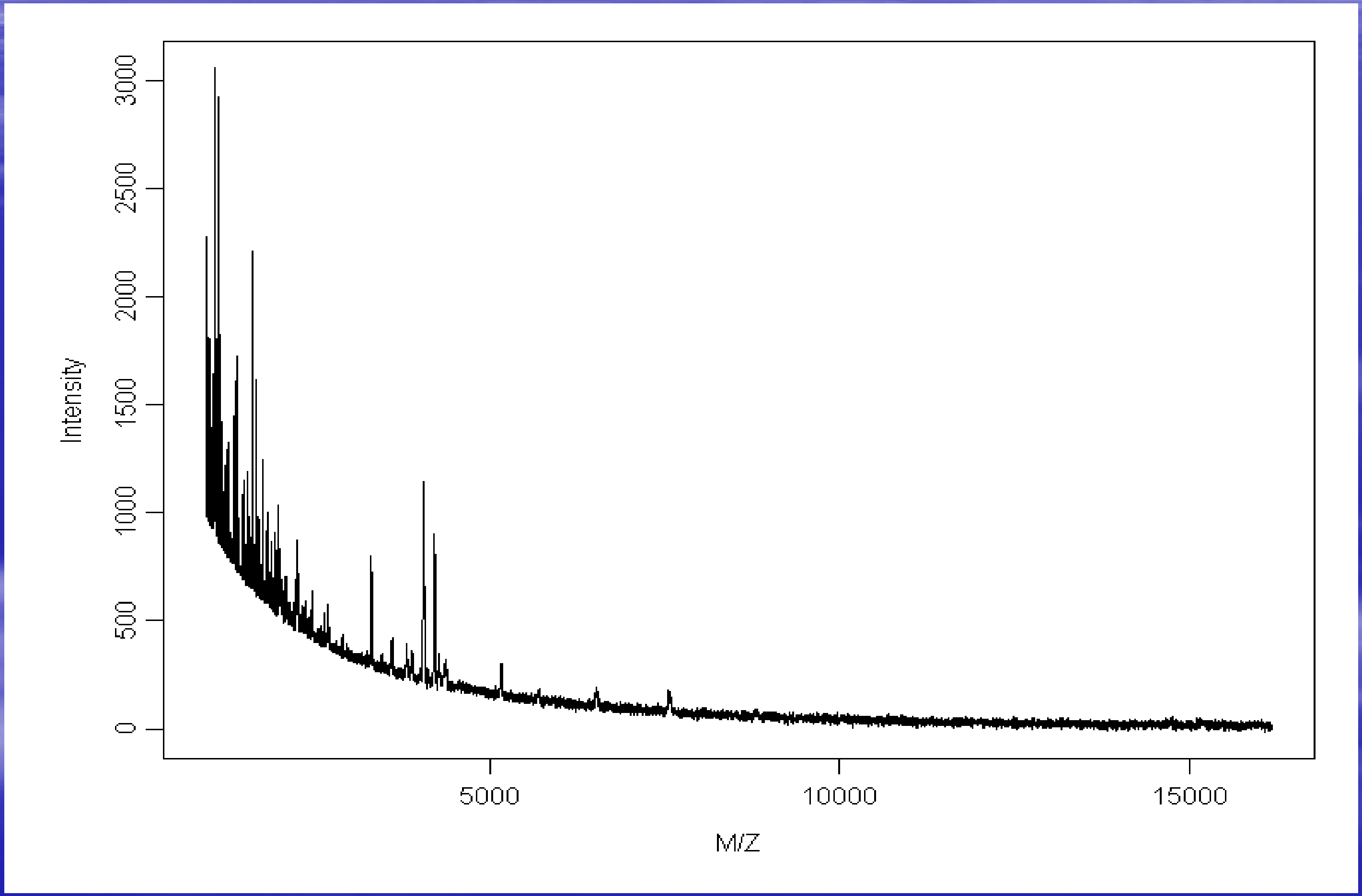
# Simulation of one protein with matrix adducts



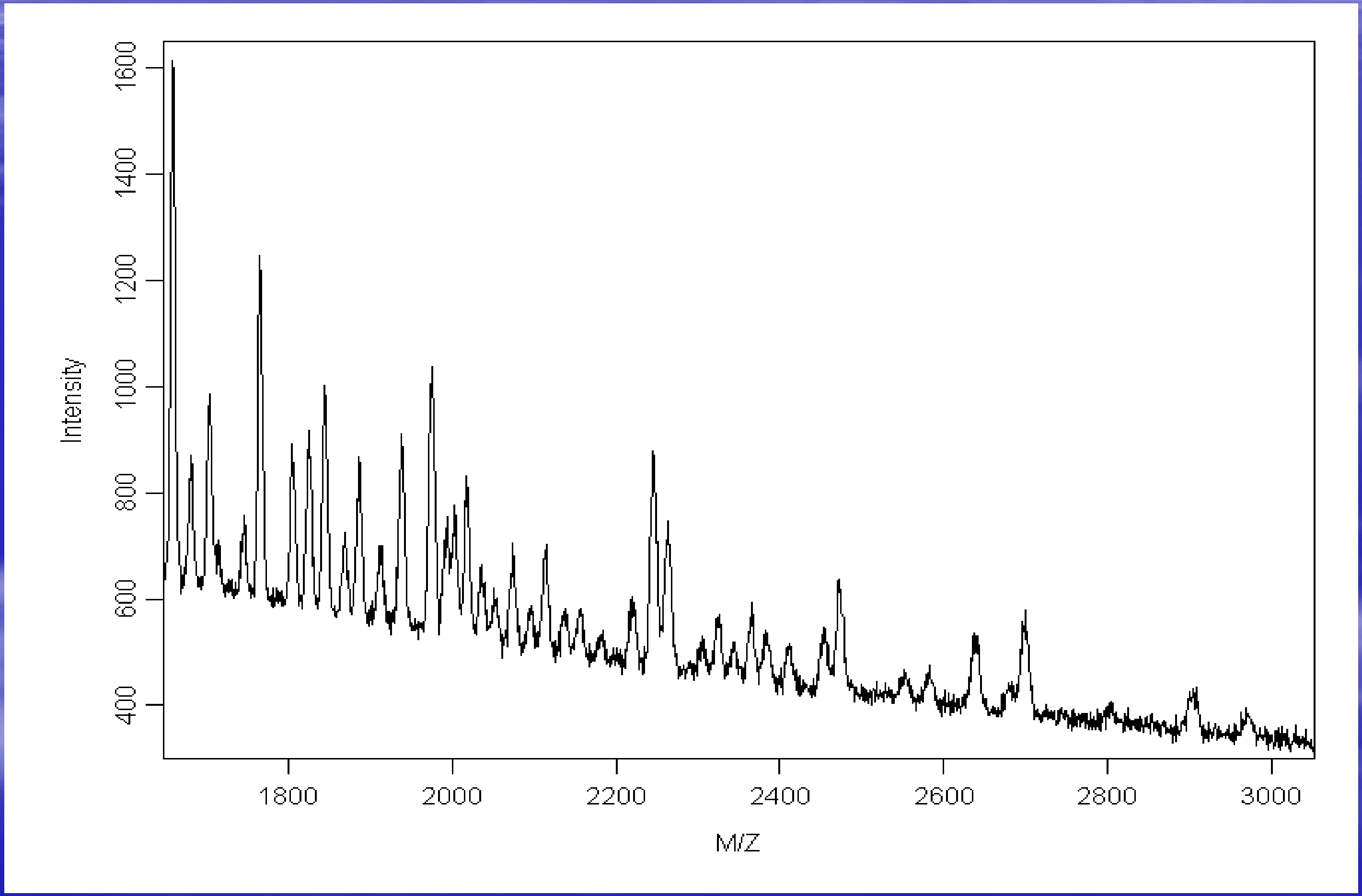
# Simulated calibration spectrum with equal amounts of six proteins



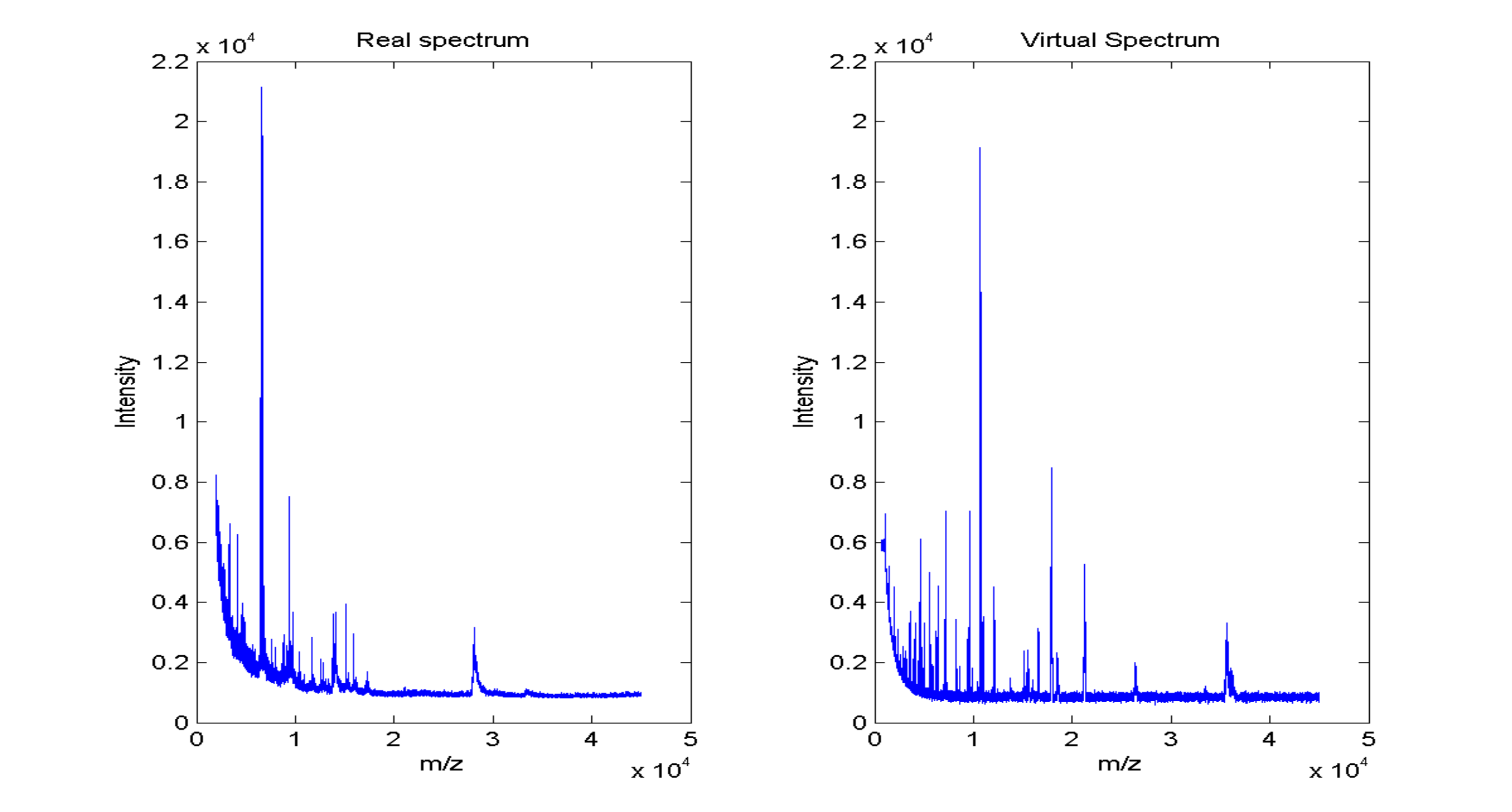
# Simulated spectrum with a complex mixture of proteins



# Closeup of simulated complex spectrum



# Real and Virtual Spectra



# Example: Pancreatic Cancer Study

- Koomen, et al. (2004)
- 256 blood serum samples – 141 pancreatic cancer, 115 normal controls
- 1 spectrum per sample.
- Samples (all fractions) run in 4 blocks on 4 different dates
- Goals:
  - Identify differentially expressed protein peaks.
- Must adjust for block effects on spectra

# Example:Organ-Cell Line Expt

- 16 nude mice had 1 of 2 cancer cell lines injected into 1 of 2 organs (lung or brain)
- Cell lines:
- A375P: human melanoma, low metastatic potential
  - PC3MM2: human prostate, highly metastatic
- Blood Serum extracted from each mouse – placed on 2 SELDI chips
- Samples run at 2 different laser intensities (low/ high)
- Total of 32 spectra (observed functions), 2 per mouse

# Example: Organ-Cell Line Expt

- Goal:

Find proteins differentially expressed by:

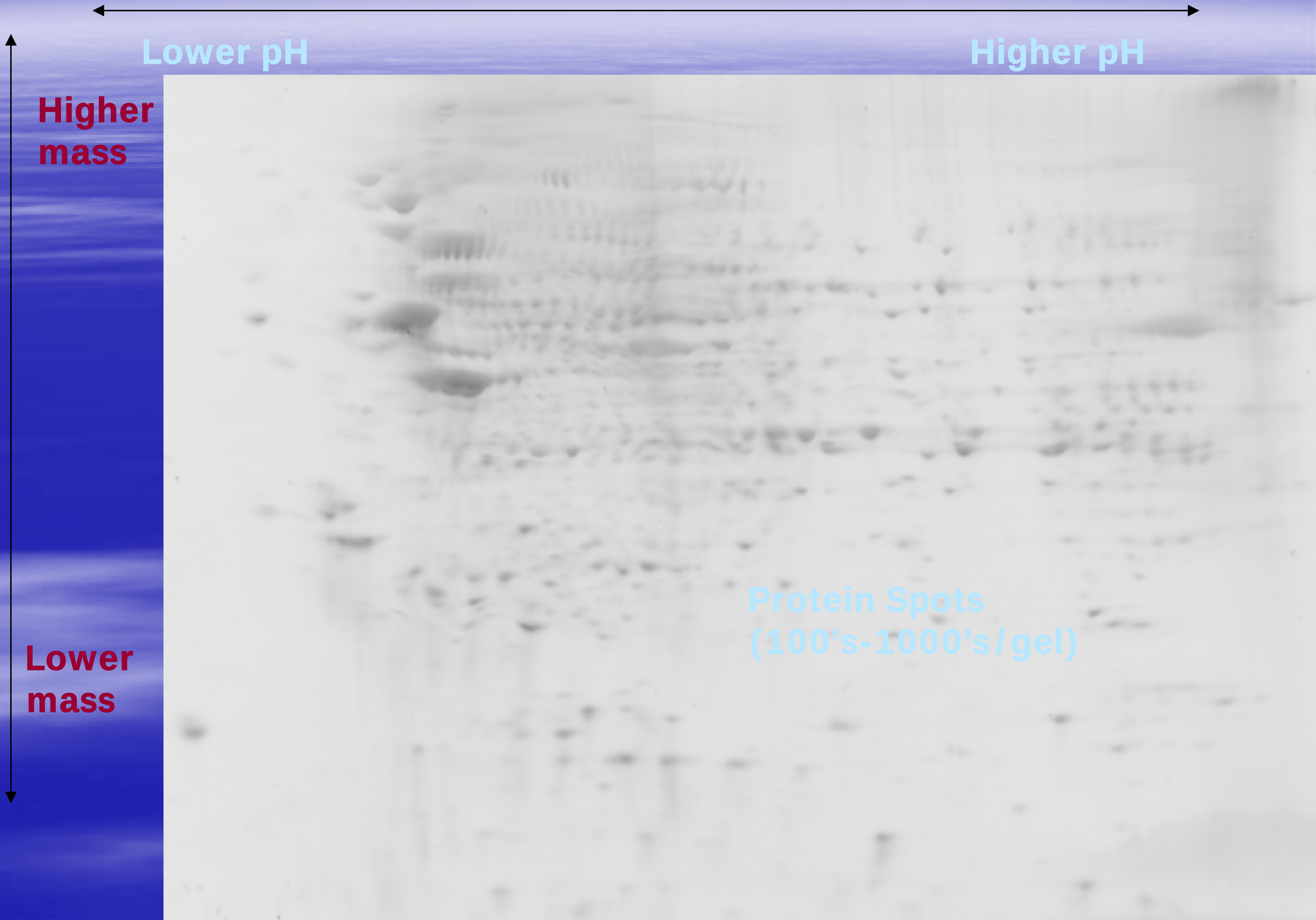
- – Host organ site (lung/brain)
- – Donor cell line (A375P/PC3MM2)
- – Organ-by-cell line interaction

- Combine information across laser intensities
  - – Must adjust for systematic laser intensity effect, and model correlation betwee spectra from same mouse.

# 2-D Gel Electrophoresis

- Method for separating proteins in biological sample based on isoelectric point (pH) and molecular mass.
- Used to identify proteins differentially expressed between treatment groups.
- Steps:
  1. Isoelectric focusing (IEF): pH gradient applied to gel, electric potential applied, causing proteins to migrate across polyacrylamide gel based on their pH
  2. Treated with SDS: denatures proteins and attaches negatively charged SDS molecules, with the amount proportional to protein's length (mass)
  3. Electric potential applied again, but in perpendicular direction, causing proteins to migrate. Friction of gel acts as sieve, so lighter proteins will travel further
  4. Stain applied to gel which binds to proteins.
  5. Gel image scanned into computer for quantitative analysis
  6. After analysis, cut out spots for identification by MS

# 2-D Gel Electrophoresis\



Lower pH

Higher pH

Higher mass

Lower mass

Protein Spots (100's-1000's/gel)

# 2-D Gel Electrophoresis

- First developed in 1975 by Patrick O’Farrell, PhD
- It is considered the “workhorse” of proteomics, yet its contribution to biomedical science has been limited by several factors
  - Major factor: The lack of efficient, effective, and automatic image processing algorithms.
  - There exist a number of commercial 2d gel image processing packages
- We will discuss the inadequacy of these commercial packages, until recently, and present some alternative approaches that work better.

# Major Areas of Statistical Input

1. Experimental Design
   - Prevent systematic bias and experimental variation from sabotaging a study
2. Quantitative Analysis
   - Data visualization (frequently a simple look at the data will reveal problems)
   - Preprocessing (extract and normalize protein signal from raw data)
   - Data Analysis (identify potential biomarkers and/or proteomic signatures for disease/response)

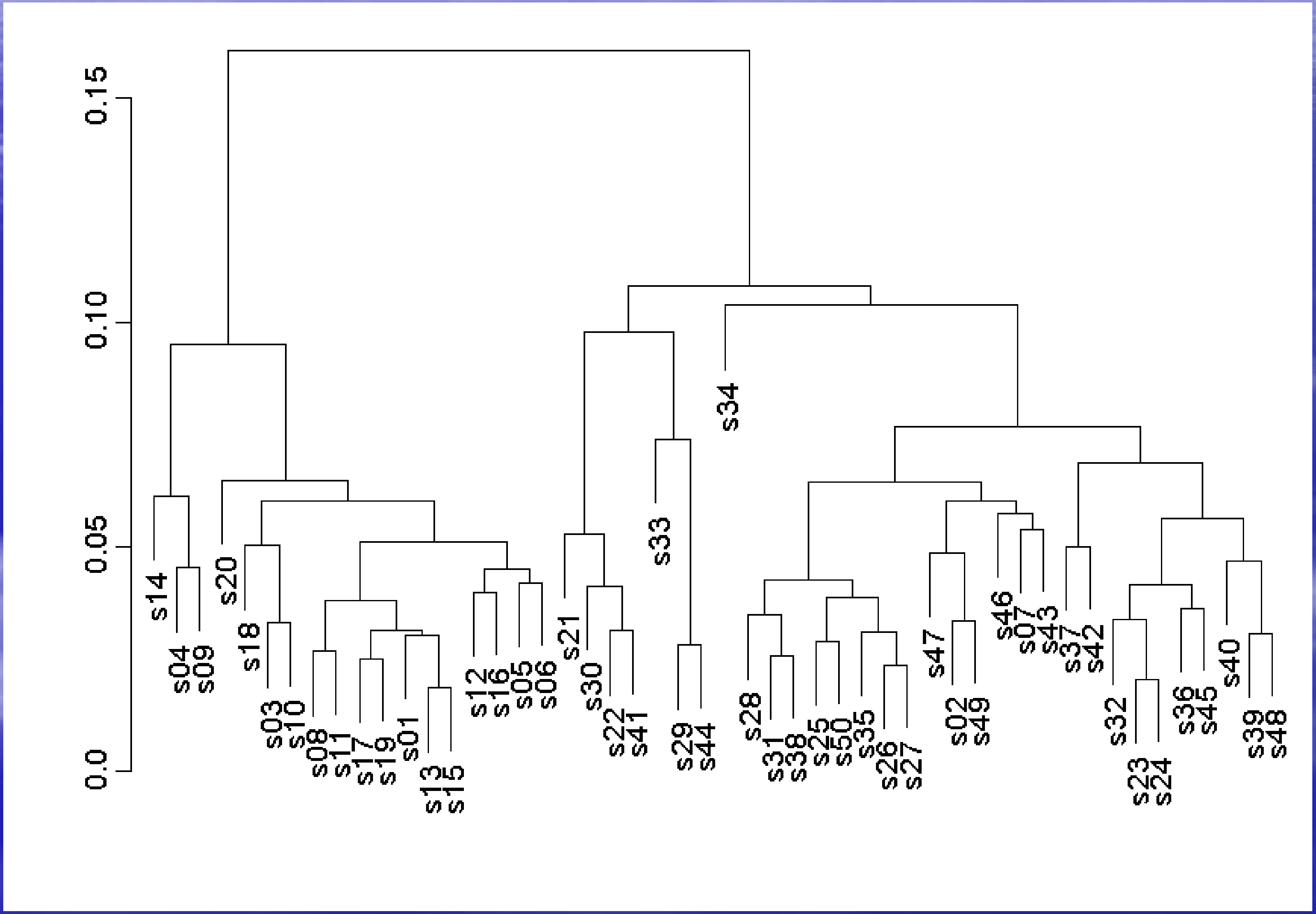
# Design makes a difference

- Selection of appropriate controls
  - see your local epidemiologist (specificity?)
- Sample size
  - make sure you have enough to find meaningful differences (or when constrained, at least find out how small of a difference you can detect)
- Sample collection and handling must be carefully controlled
- May want to Block on factors likely to impact data (e.g. run time)
- Randomization is needed at multiple points in the process

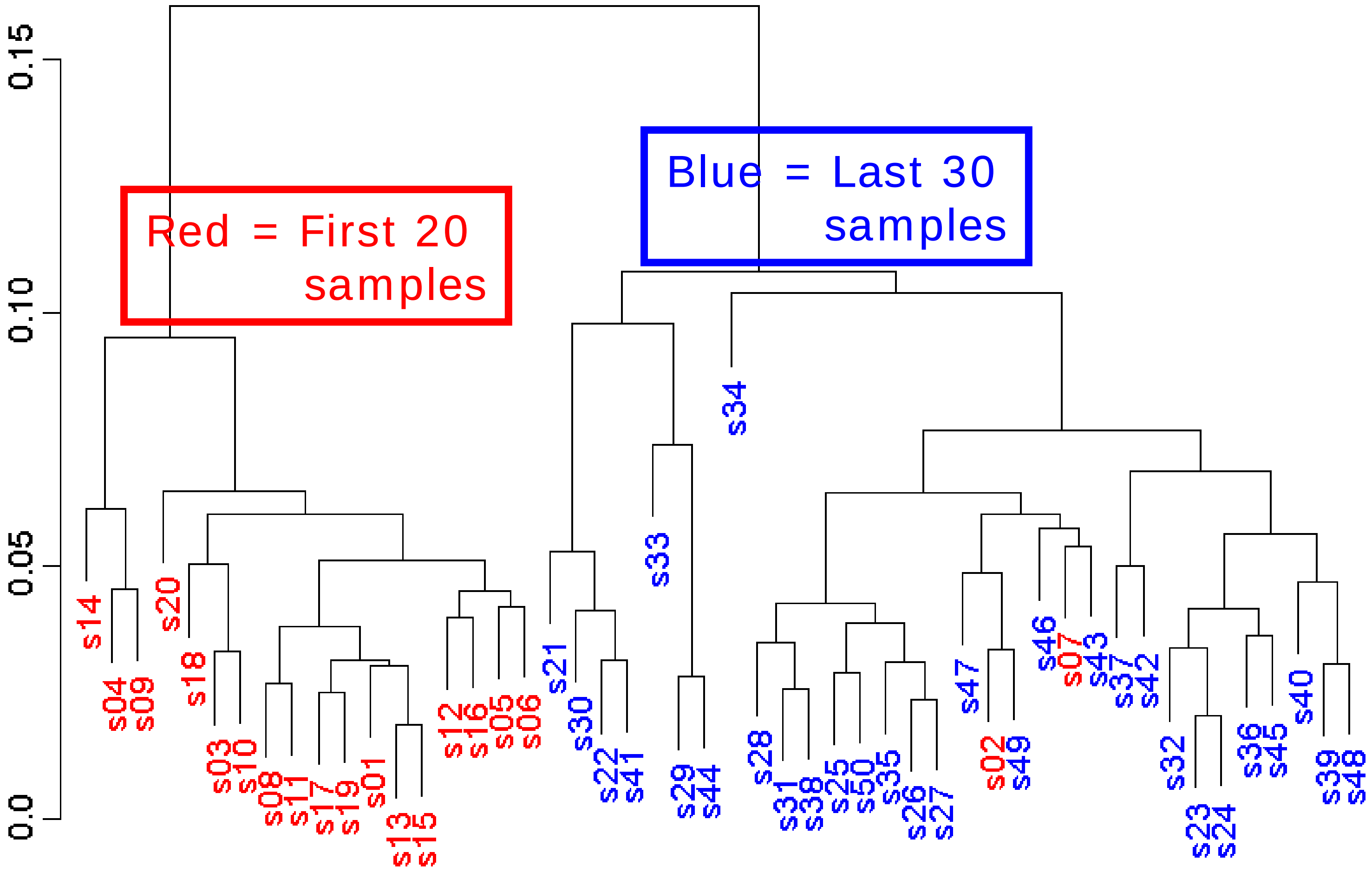
# Sample **handling** is critical

- All samples must be collected uniformly
  - Consistent protocol
  - Enforced at every collection site
- Failure to do this can (will) affect protein profiles
- The problem is particularly serious if sample handling is confounded with interesting variables (normal vs cancer)

# Hierarchical clustering of serum protein profiles of brain cancer



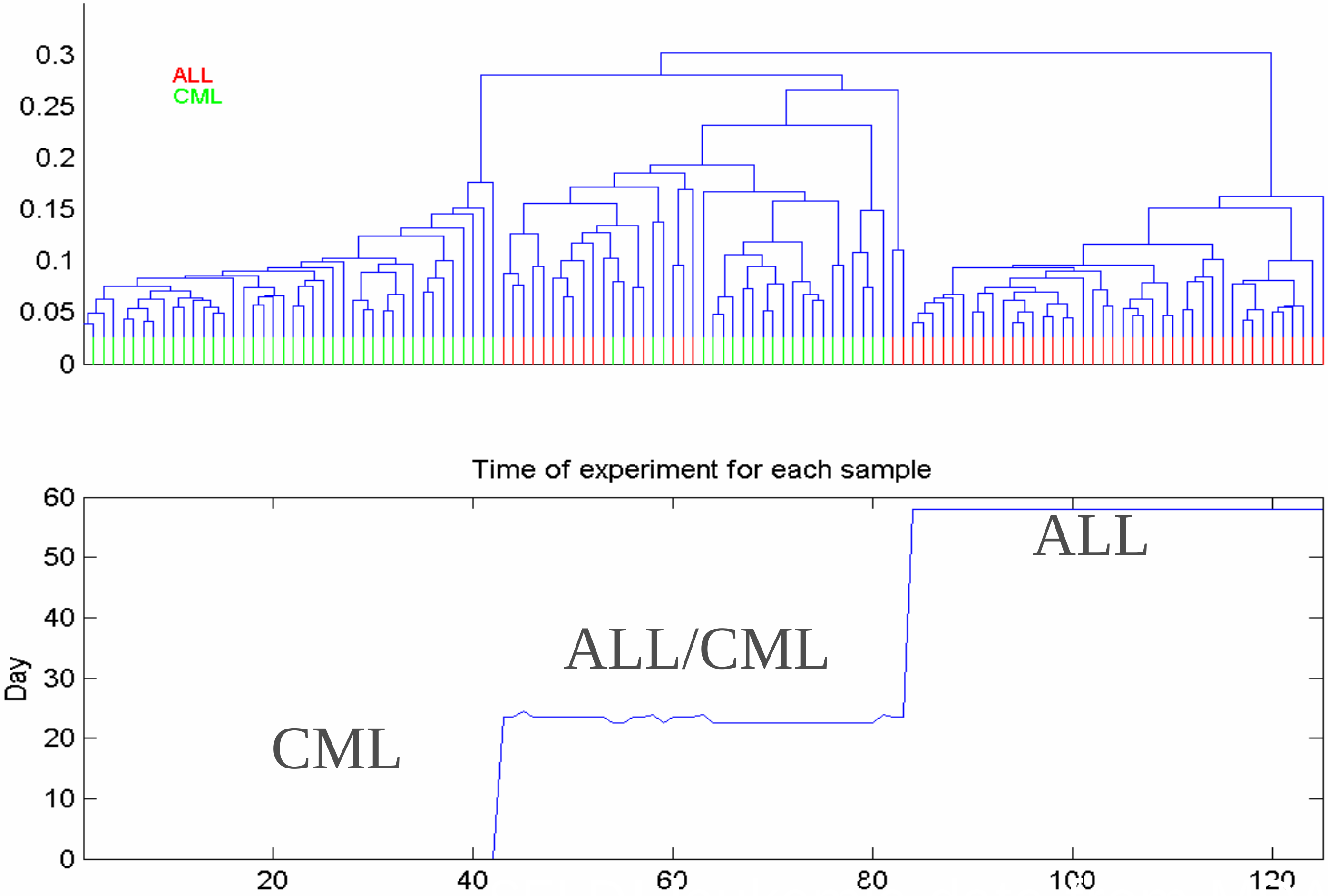
# Clustering reflects changes in the sample collection protocol



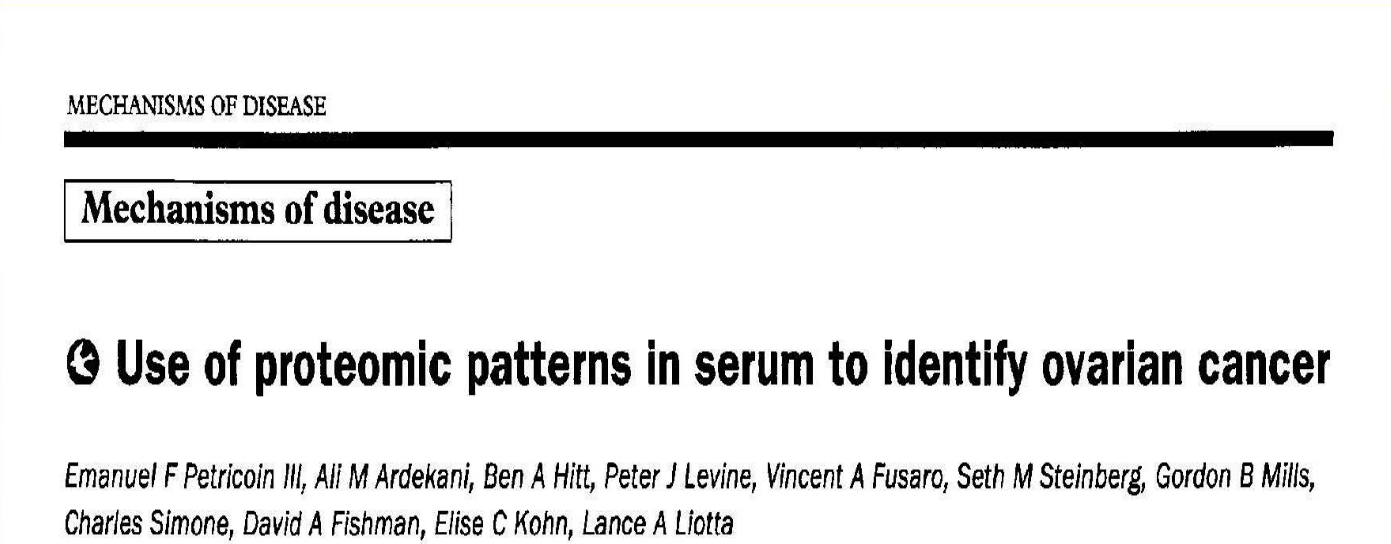
Red = First 20 samples

Blue = Last 30 samples

# Unsupervised methods often cluster samples by run date



# Case Study: Statistics Making a Difference



MECHANISMS OF DISEASE

Mechanisms of disease

**Use of proteomic patterns in serum to identify ovarian cancer**

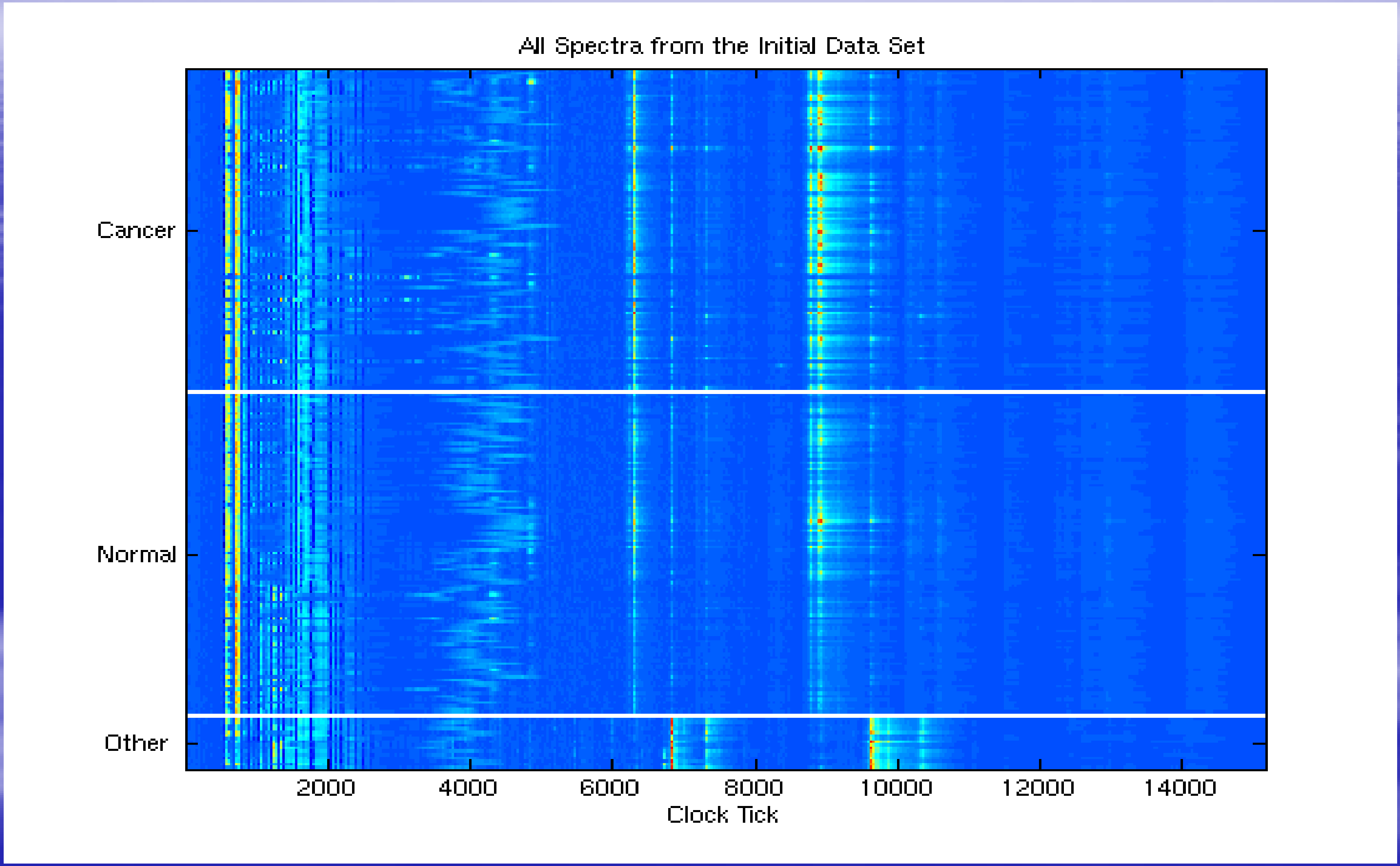
*Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta*

- Collected proteomics data on serum samples from
  - 100 women with ovarian cancer
  - 100 normal controls
  - 16 women with benign disease
- Selected 50 normal and 50 cancer
- Trained a statistical/computational algorithm to distinguish between the two types
- Tested the algorithm on the remaining samples

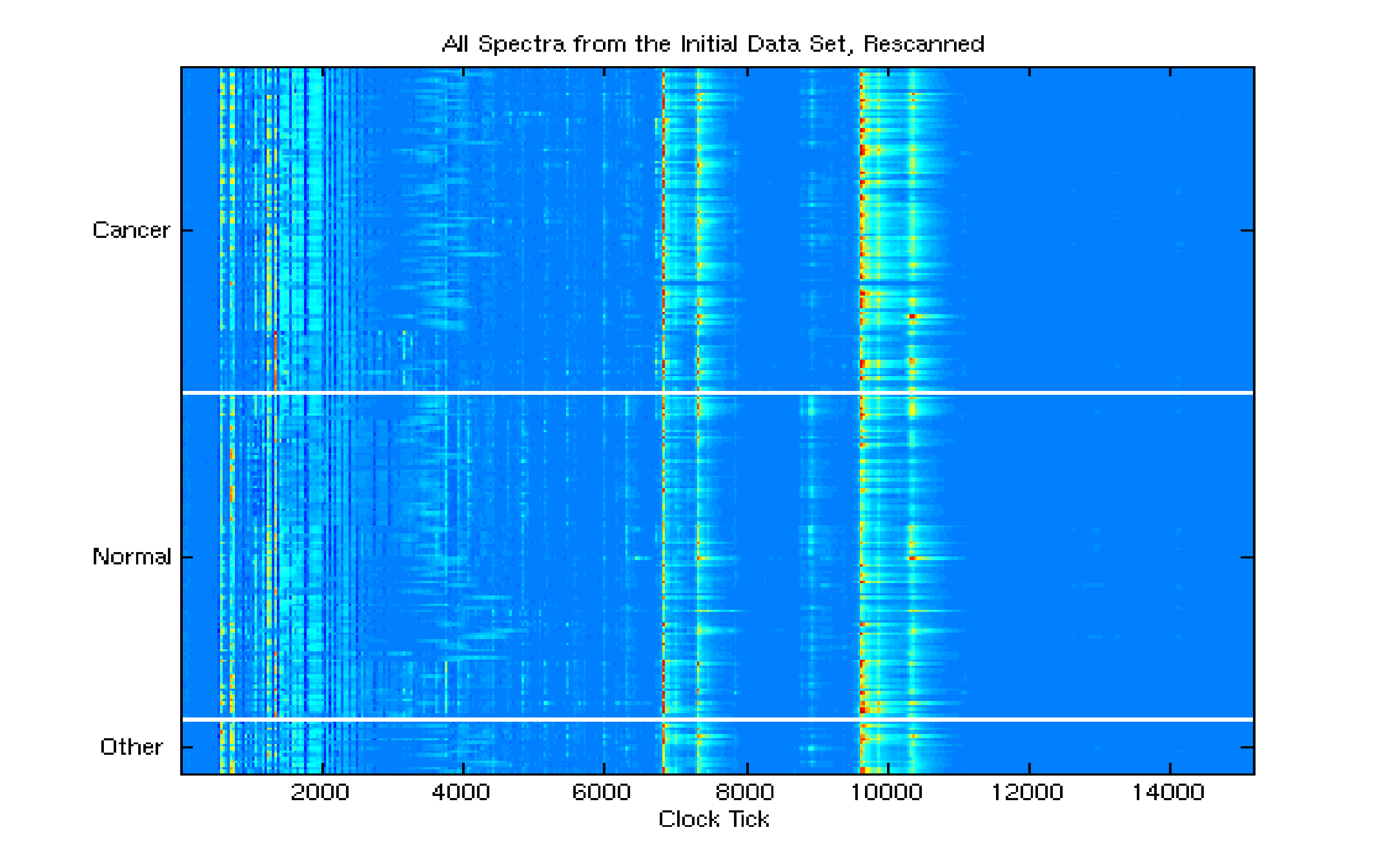
# Petricoin **Results**

- Results:
  - Correctly classified 50/50 of the ovarian cancer test cases as cancer
  - Correctly classified 47/50 normal samples as normal, with 3/50 classified as cancer
  - Correctly classified 16/16 benign disease as "neither normal nor cancer"
- Remarkable!! Can we identify ovarian cancer with a simple blood test? If so, then we can pretty much eliminate that disease, since it is easily treated when detected early.
- Ovacheck^TM (Correlogic, Quest Diagnostics, LabCorp): Company started to market ovarian cancer blood test based on these results

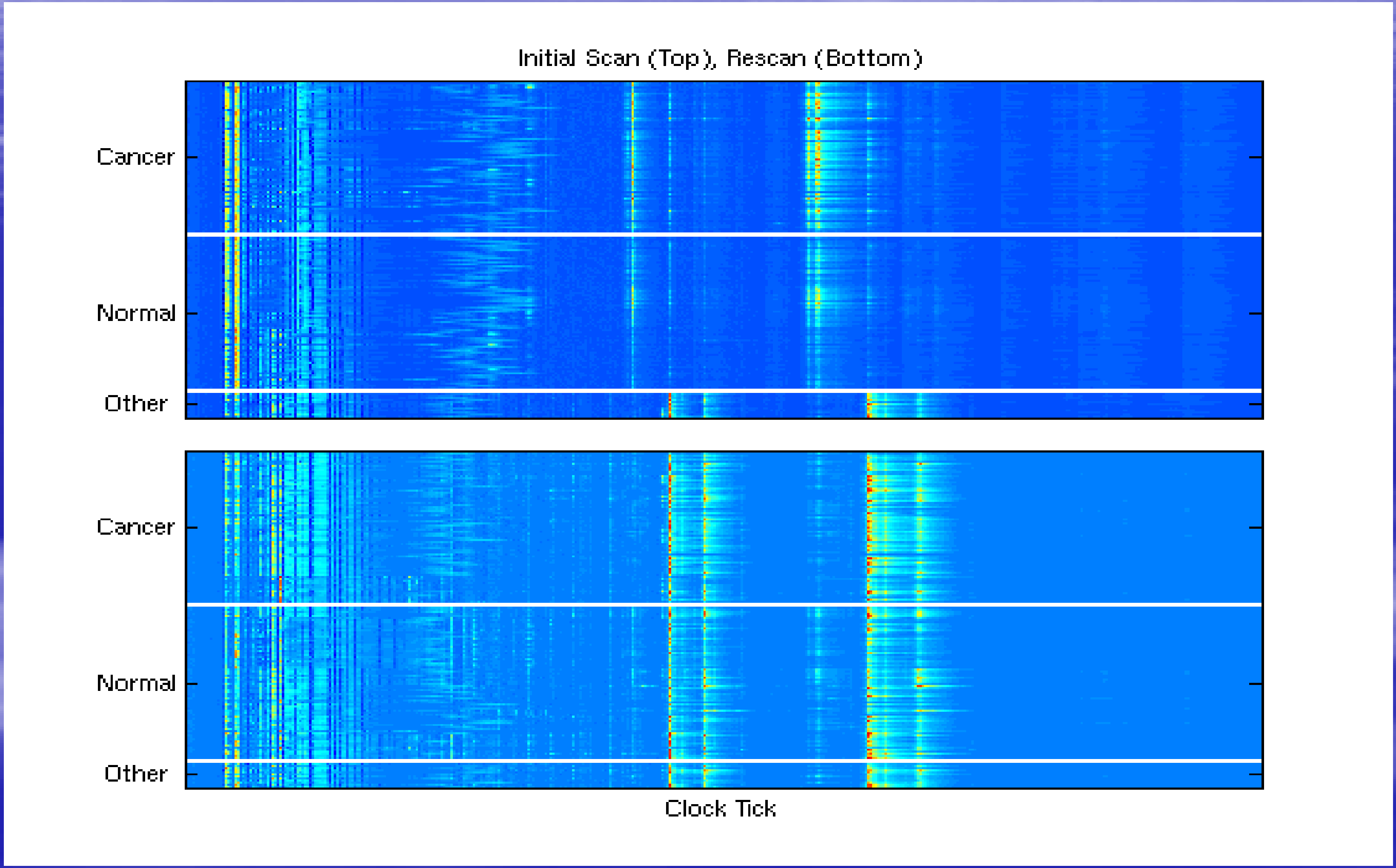
# Some structure is visible in Heat Map



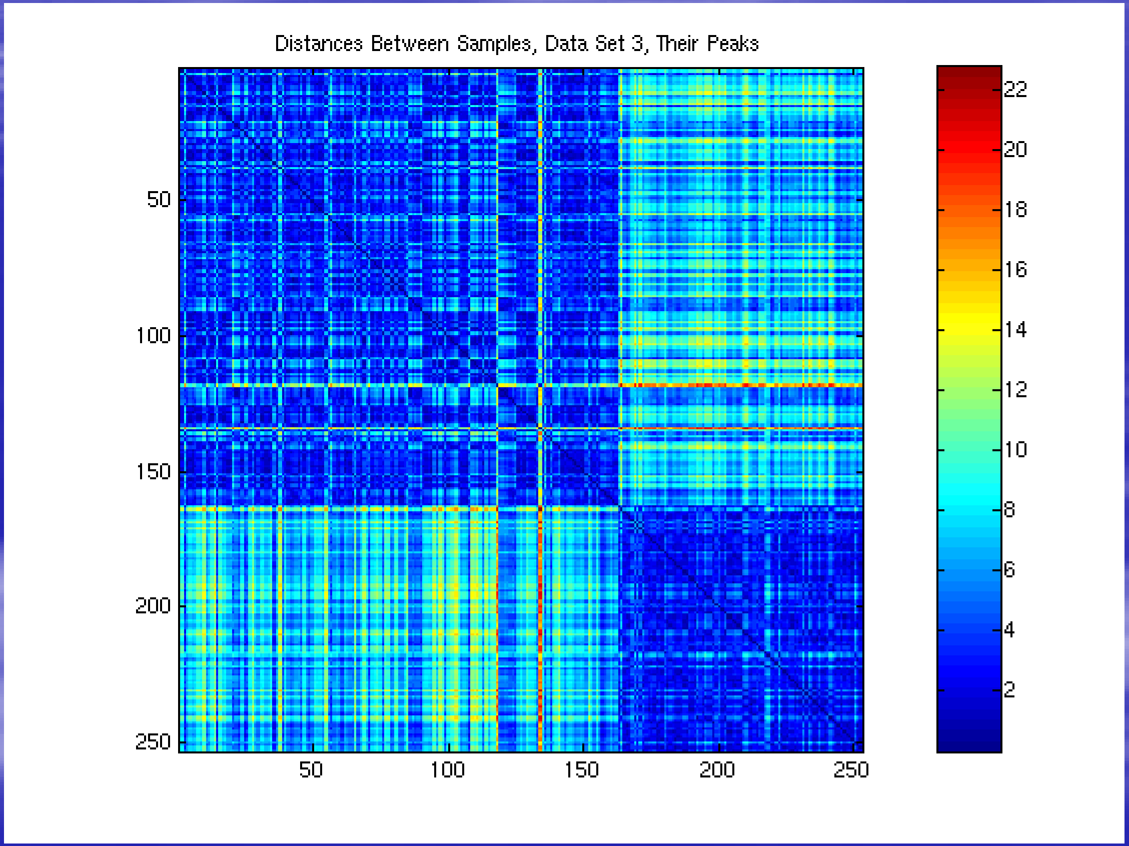
# Structure disappears for different chip type (same samples, different chip type)



# Any ideas what happened here?

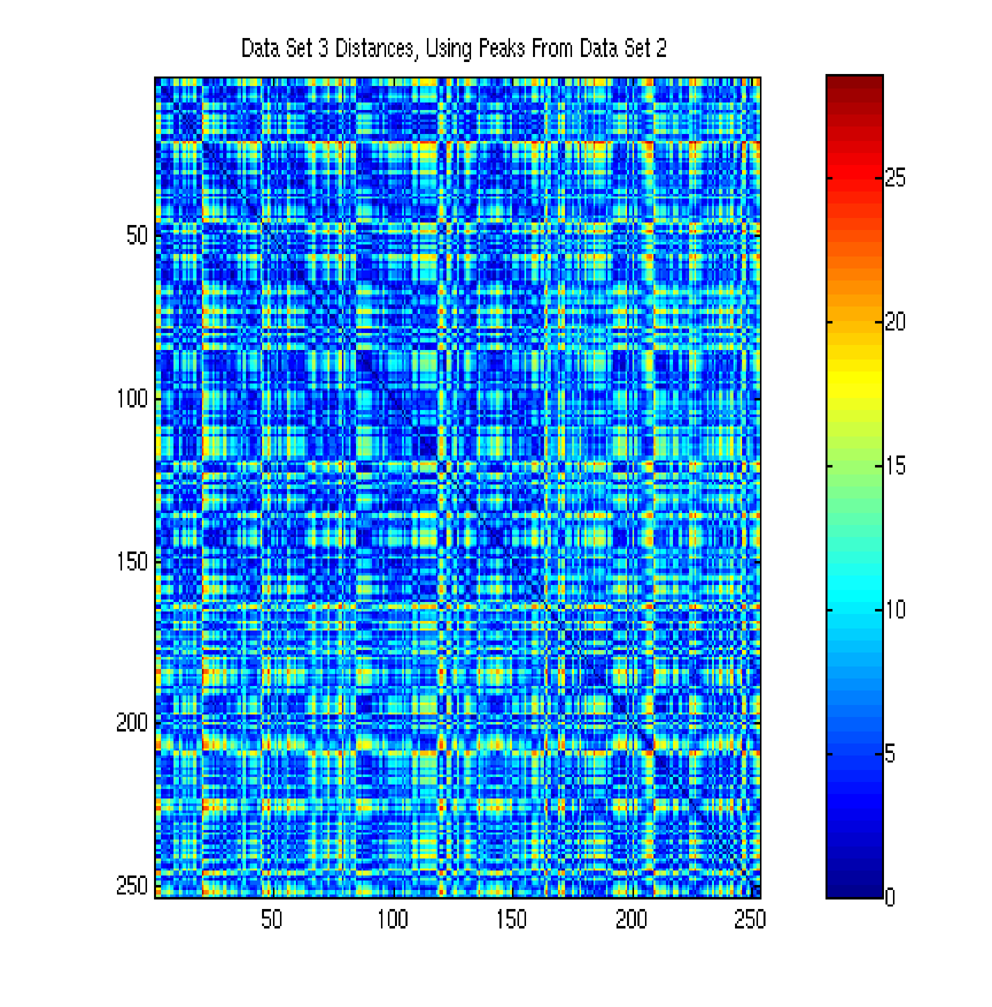


# A Follow-up Study



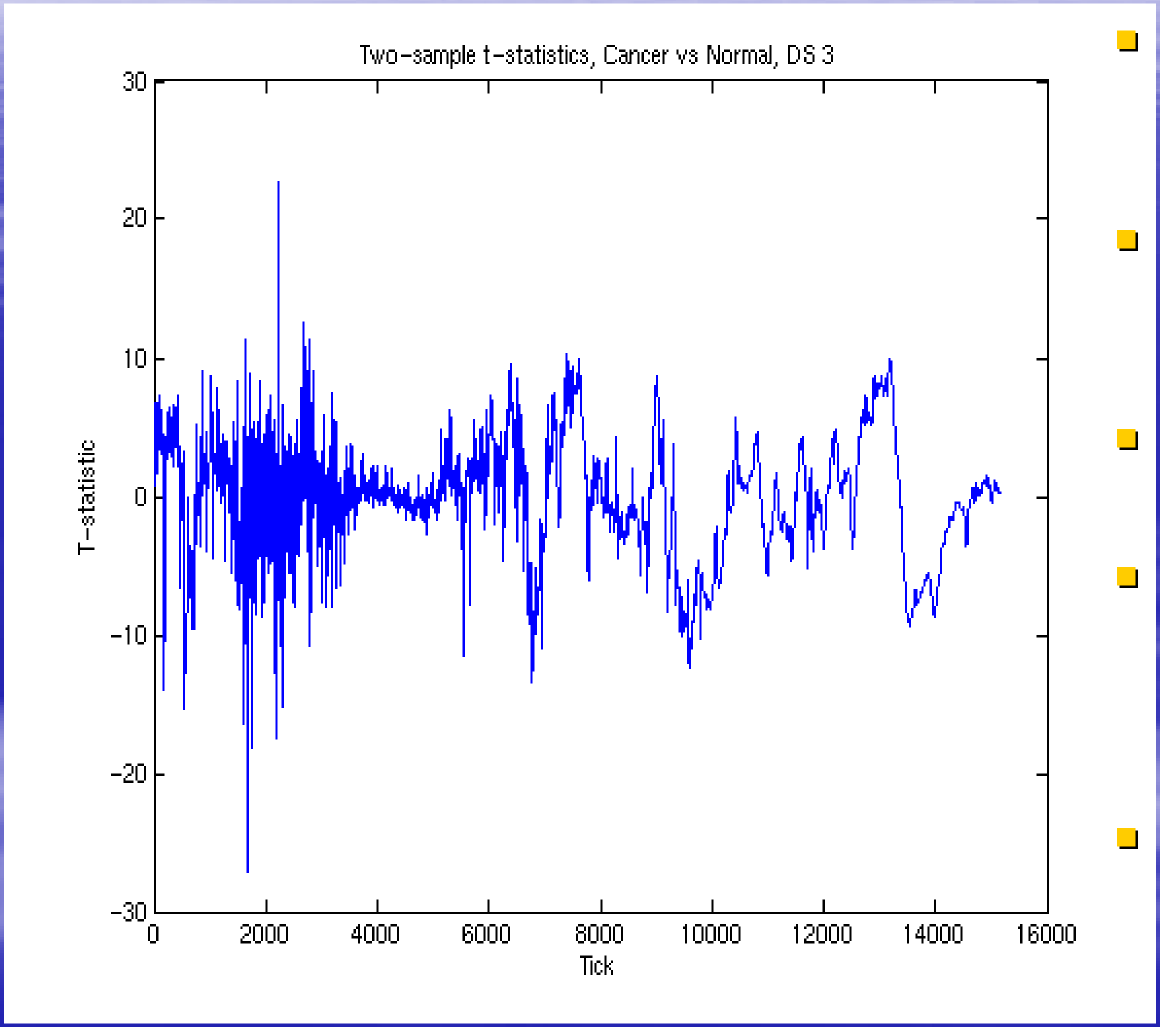
- They ran another study: Again, remarkable results
- Near perfect classification of cancers and normals

# A Follow-up Study



- Problem: algorithm from 2^nd data set does not work for 1^st data set
- Similarly, algorithm from 1^st data set does not work on 2^nd
- Pattern not reproducible!!
- Hmmm.  Not encouraging

# A Follow-up Study



- Plot of t-statistics separating cases and controls in data set 2
- MANY regions of spectrum separate cases/controls
- Including very low mass regions
- Can perfectly separate cases/controls with just two peaks, e.g. (2.79D, 245.2D)
- There is something funky with this data set!

# But PNAS paper found “reproducible signal”

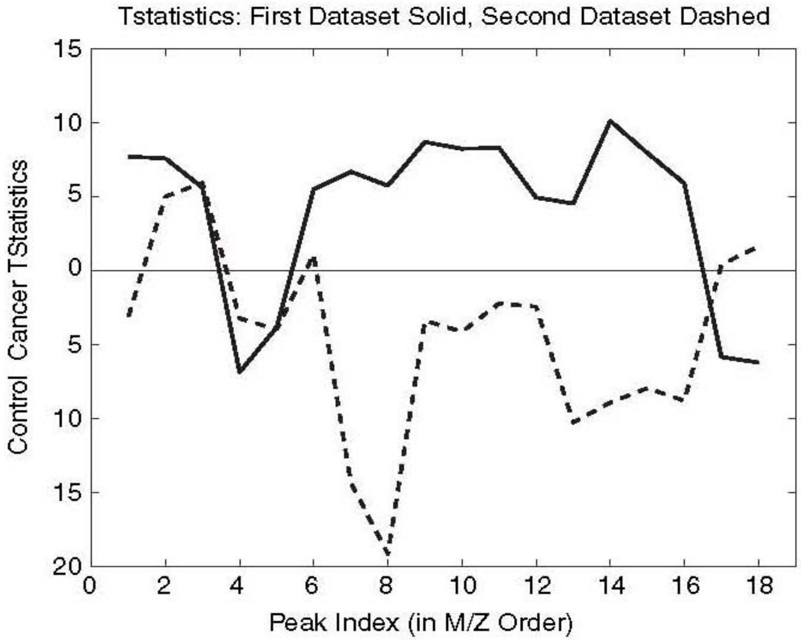


**Fig. 1.** Summary of *t* statistics at 18 published peaks. Peaks have *m*/*z* values as indicated in the text. The *t* statistics represent the difference in spectral intensity between cancer and unaffected spectra for the 18 reported *m*/*z* values. **Solid line** = *t* statistic values from the first dataset; **dashed line** = *t* statistic values from the second dataset. The magnitude and sign of the *t* statistics correspond to the relative protein expression of cancer and normal spectra for the two datasets; a change in sign indicates that the average spectral intensity at that *m*/*z* value was greater in cancer spectra for one dataset and for control spectra in the other.

- Zhu, et al. (2003 PNAS 100:14666-71)
- Reported that use of classification rule derived from 1st data set could accurately classify 2nd data set.
- Computed 2-sample t-statistics for 18 peaks contained in their sampling rule
- How then did they achieve such good classification on 2nd data set?

- From Baggerly, Morris, Edmondson, and Coombes (2005 ) JNCI 97(4): 307-309

# Simulation Study

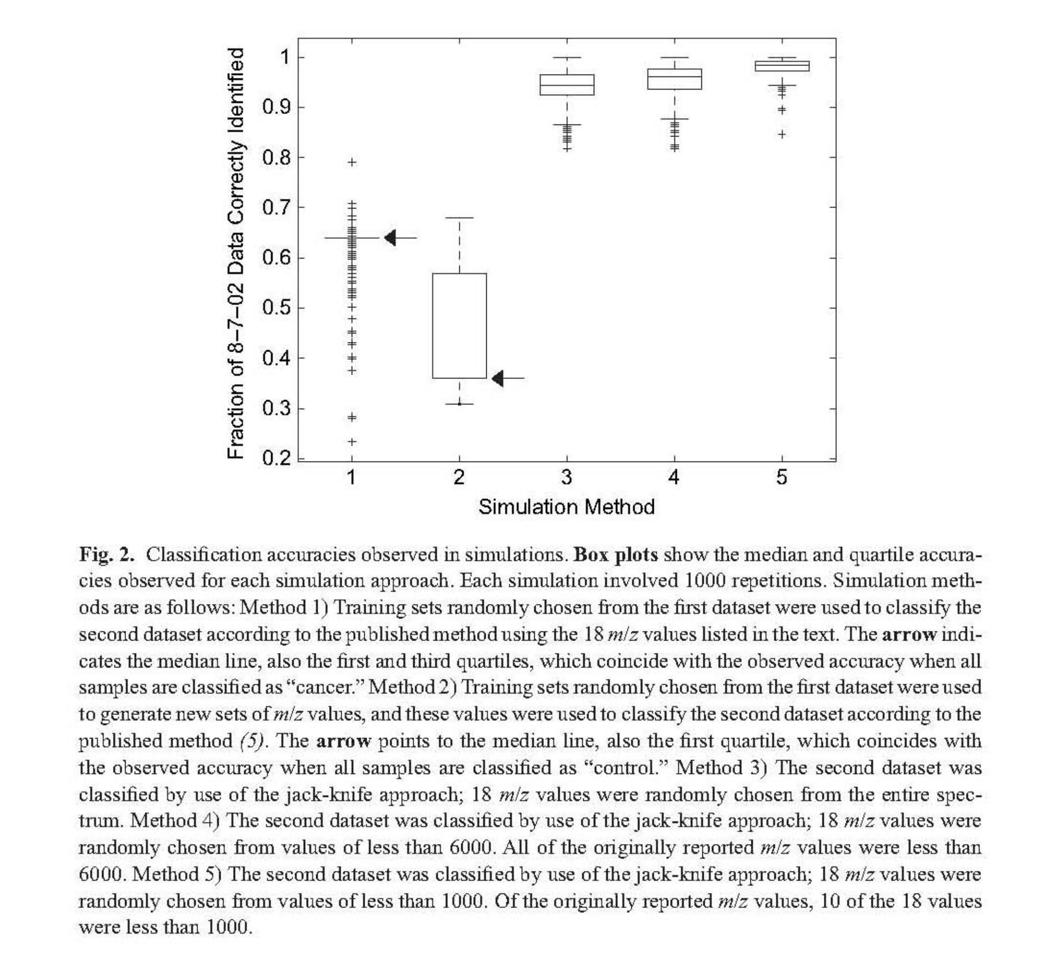


**Fig. 2.** Classification accuracies observed in simulations. **Box plots** show the median and quartile accuracies observed for each simulation approach. Each simulation involved 1000 repetitions. Simulation methods are as follows: Method 1) Training sets randomly chosen from the first dataset were used to classify the second dataset according to the published method using the 18 *m/z* values listed in the text. The **arrow** indicates the median line, also the first and third quartiles, which coincide with the observed accuracy when all samples are classified as "cancer." Method 2) Training sets randomly chosen from the first dataset were used to generate new sets of *m/z* values, and these values were used to classify the second dataset according to the published method *(5)*. The **arrow** points to the median line, also the first quartile, which coincides with the observed accuracy when all samples are classified as "control." Method 3) The second dataset was classified by use of the jack-knife approach; 18 *m/z* values were randomly chosen from the entire spectrum. Method 4) The second dataset was classified by use of the jack-knife approach; 18 *m/z* values were randomly chosen from values of less than 6000. All of the originally reported *m/z* values were less than 6000. Method 5) The second dataset was classified by use of the jack-knife approach; 18 *m/z* values were randomly chosen from values of less than 1000. Of the originally reported *m/z* values, 10 of the 18 values were less than 1000.

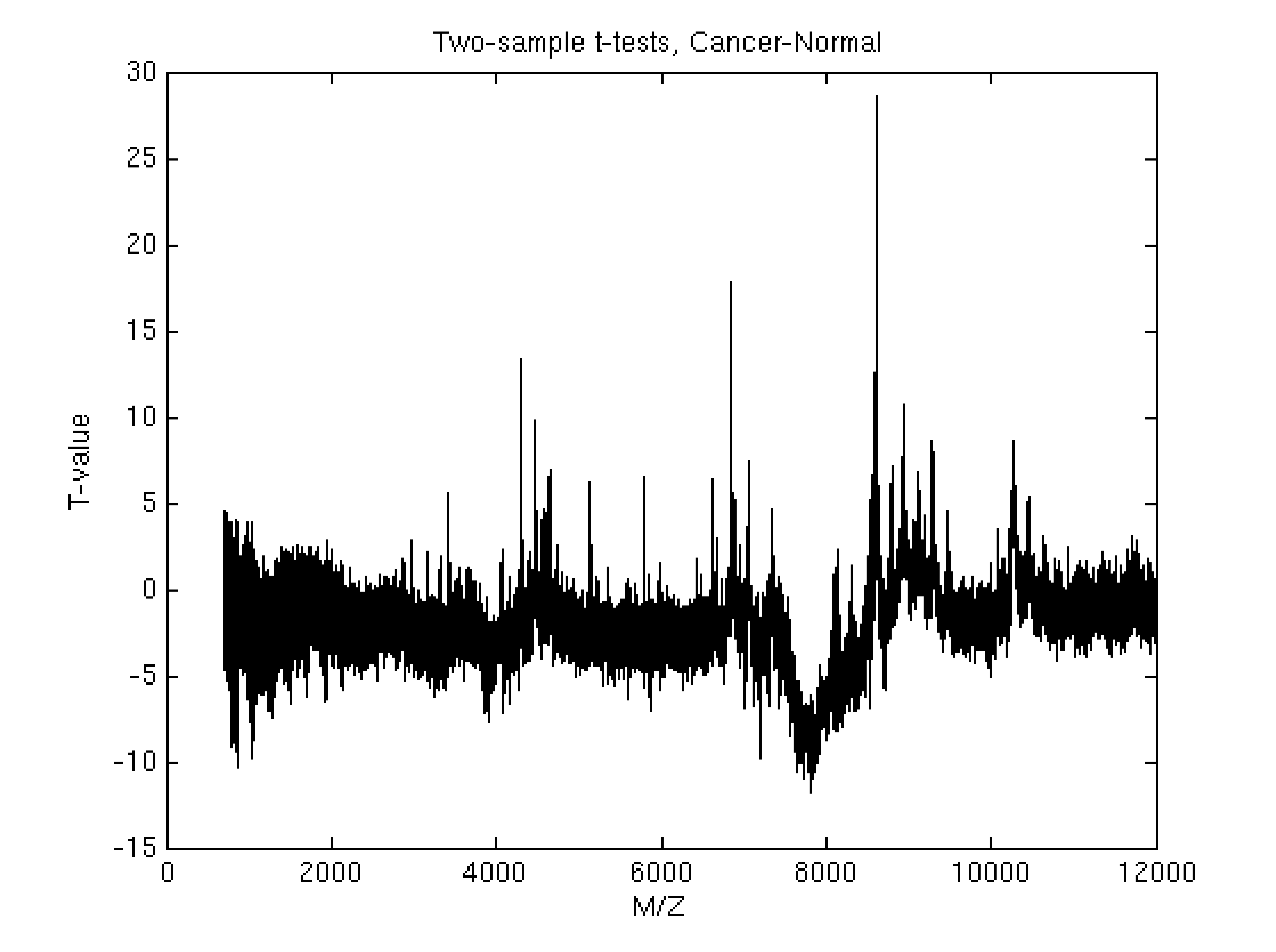
- We randomly selected 18 m/z values from spectra and built classification rule using data set 1, and then assessed its predictive accuracy on data set 2
- We obtained as good or better classification as Zhu, et al.'s model
  - 6% using whole spectrum
  - 14.8% using m/z <6000D
  - 56.2% using m/z <1000D
- Suggests systematic bias between cases/controls
- Cases and controls run in batches? Batch effect that looks like case/control effect?

- From Baggerly, Morris, Edmondson, and Coombes (2005 ) JNCI 97(4): 307-309

# Maybe better technology would help?

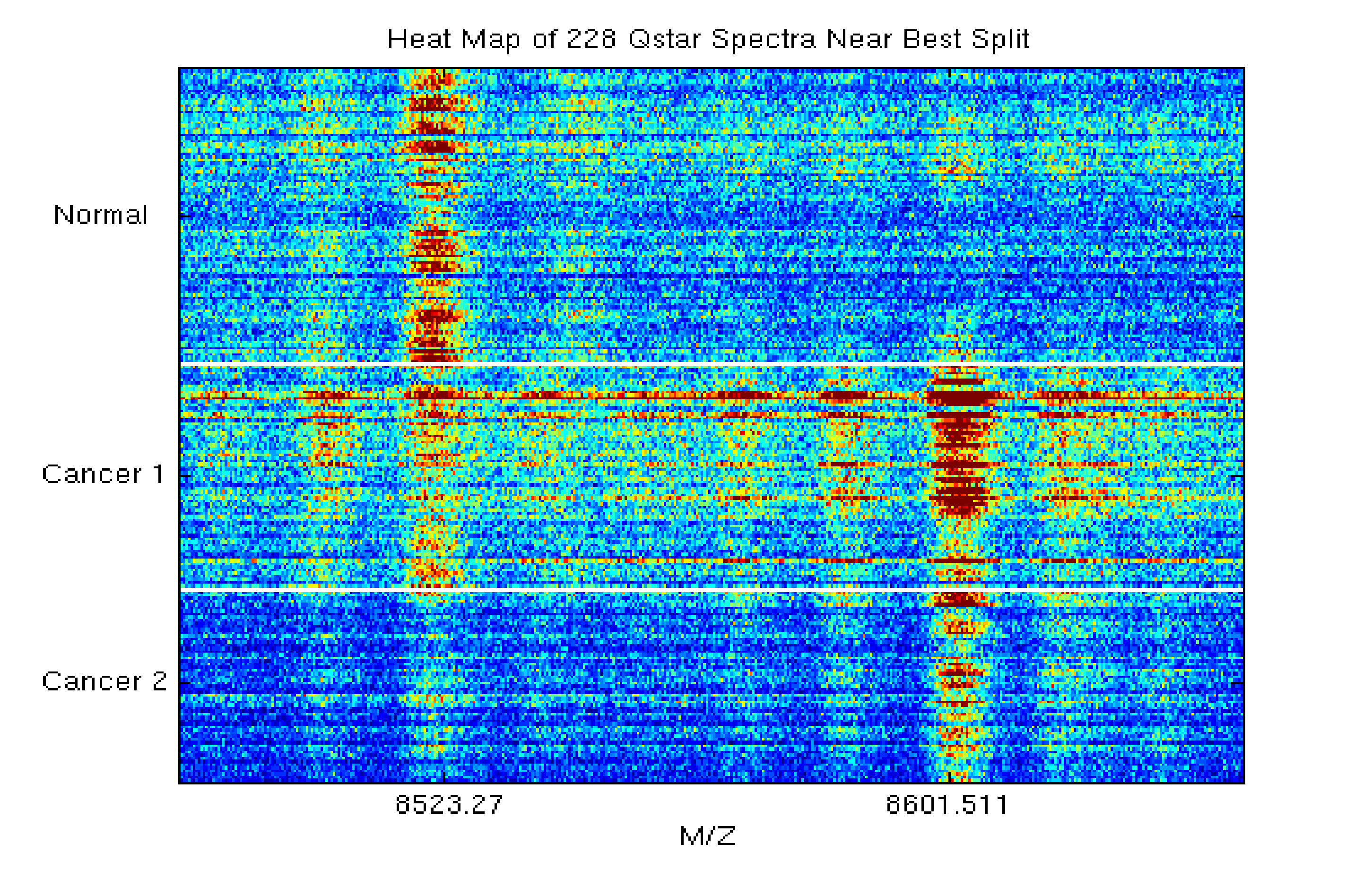
- Reference: Conrads et al., *Endocrine Related Cancer*, July 2004.
- Ovarian cancer
  - ~90 controls, ~160 cases
- Q-star instrument
  - high resolution
- Claim: can distinguish healthy women from cancer patients

# T-statistics identify separator at 8602D

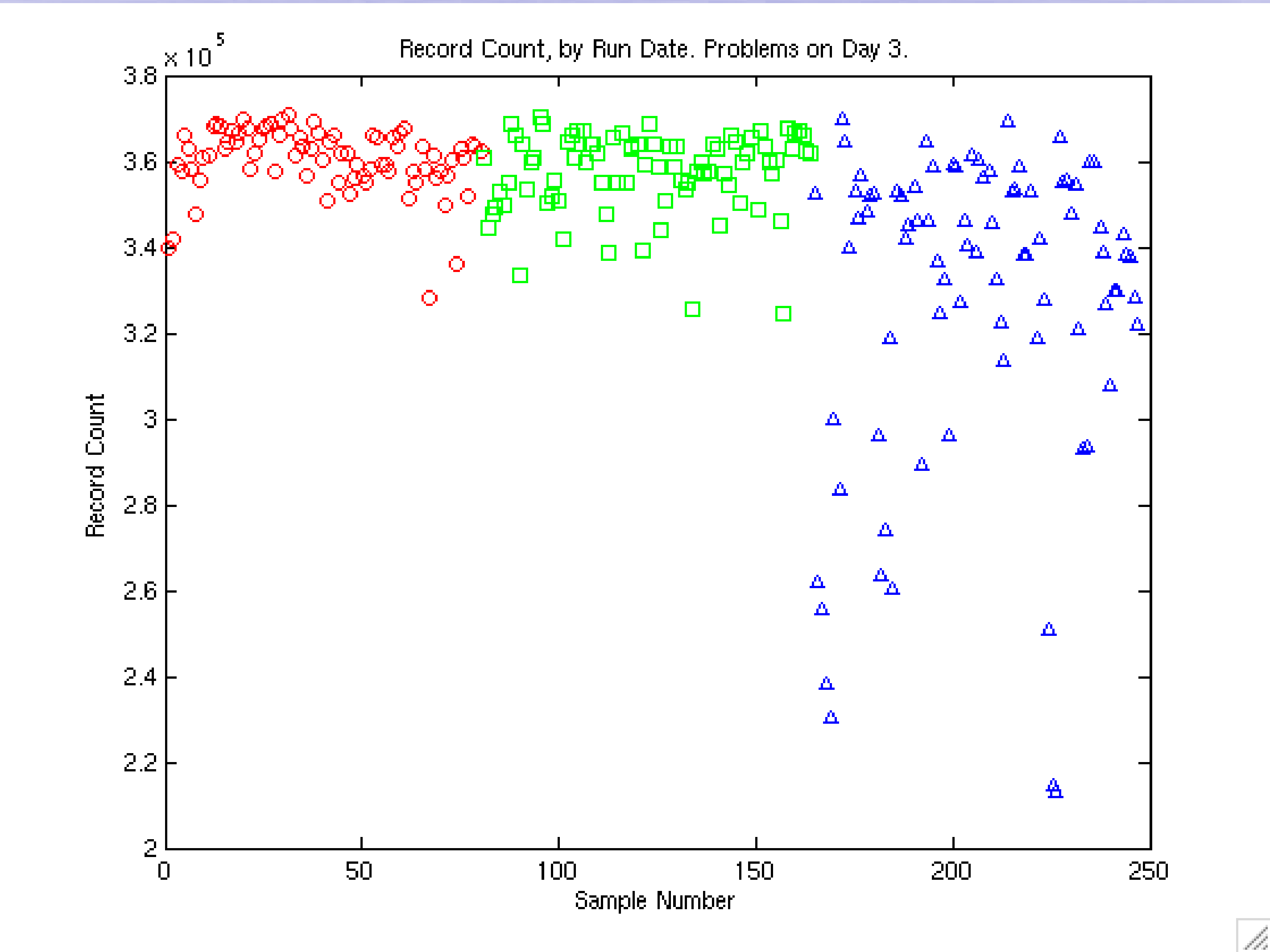


# Heat map of raw data near 8602

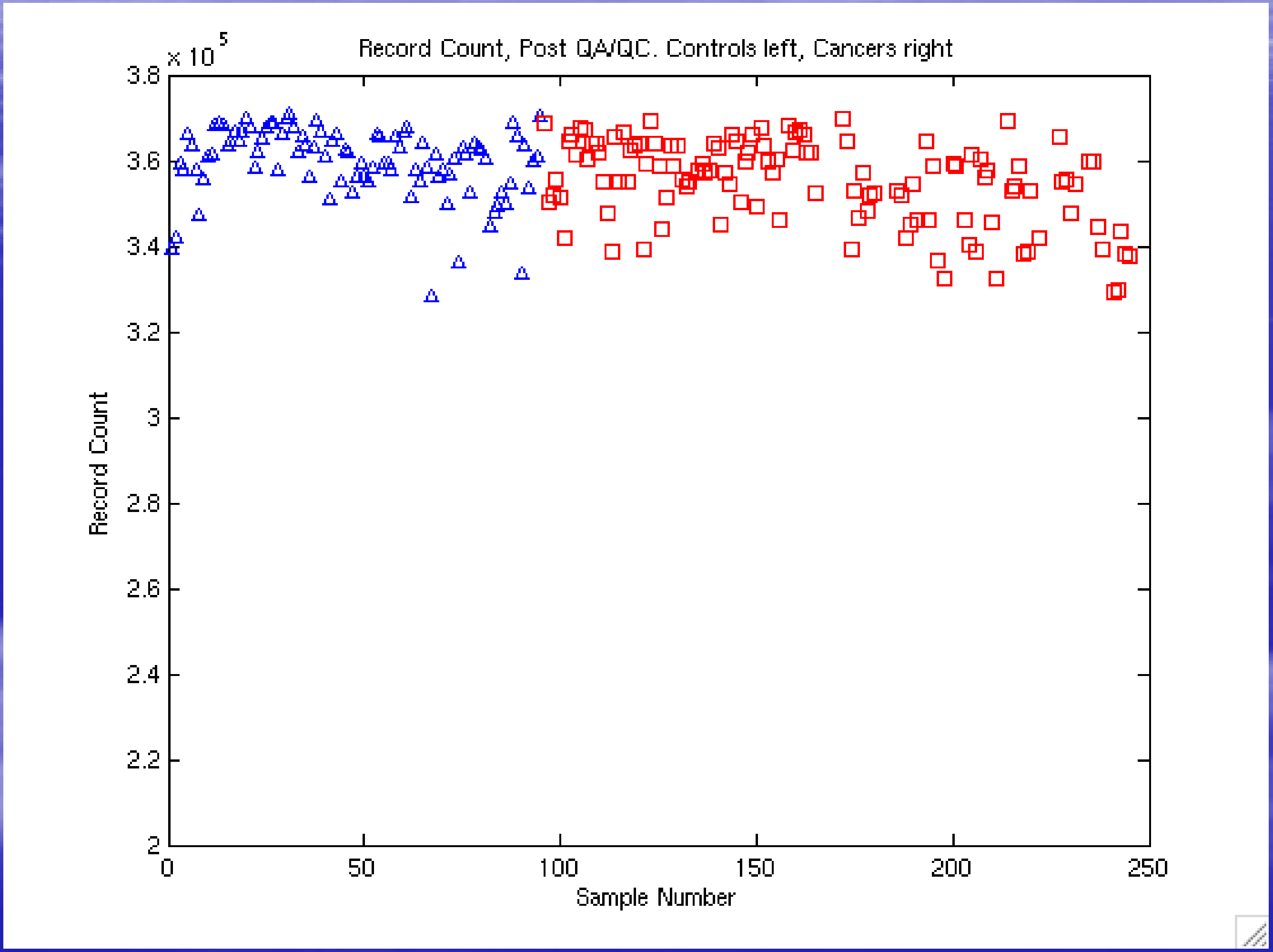
## Why are there two cancer groups?



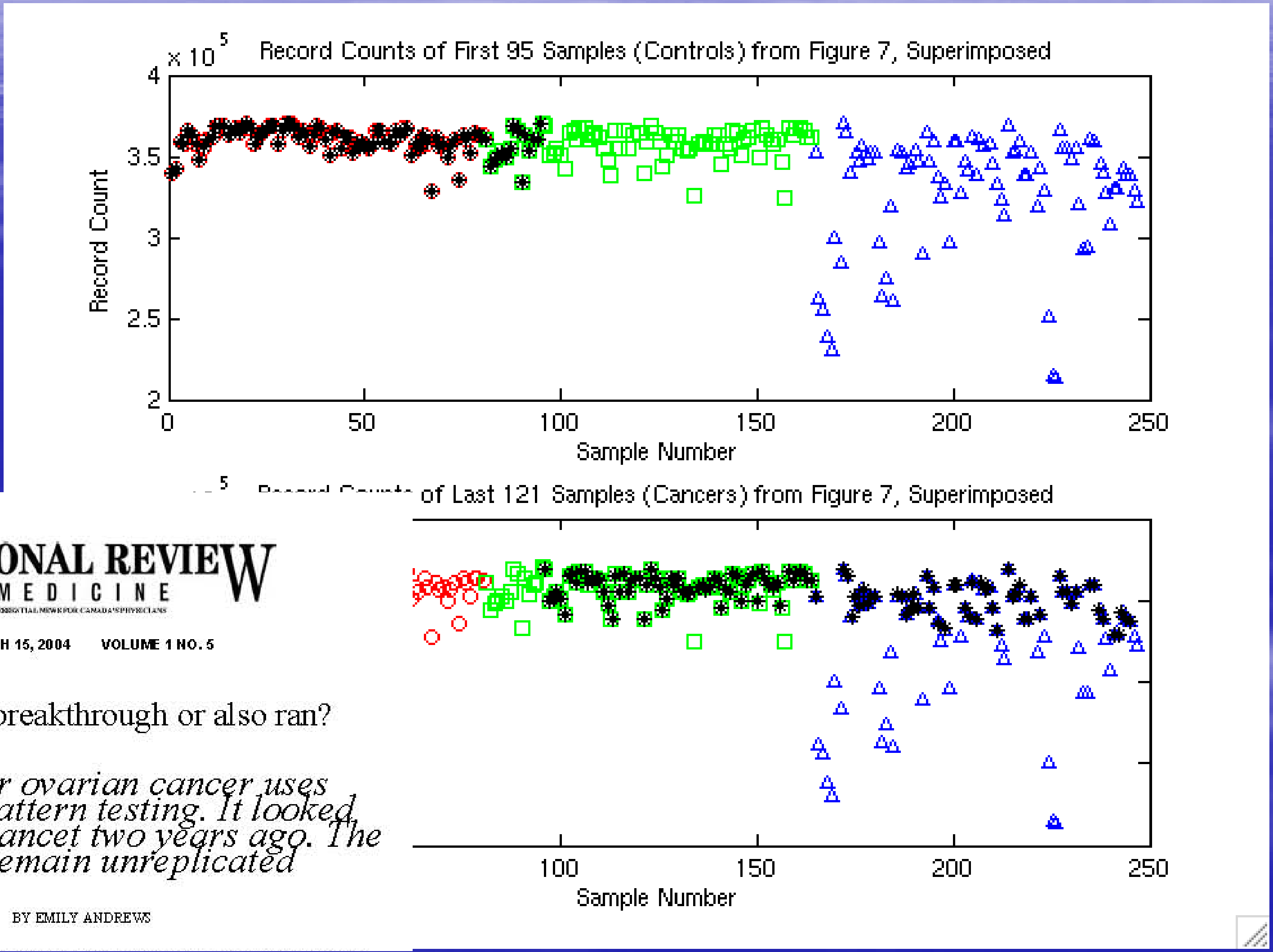
# QC: Colors indicate run date



# QC: Colors indicate control/case



# All controls were processed before all samples from cancer patients



**NATIONAL REVIEW of MEDICINE**

ESSENTIAL NEWS FOR CANADA'S PHYSICIANS

MARCH 15, 2004 VOLUME 1 NO. 5

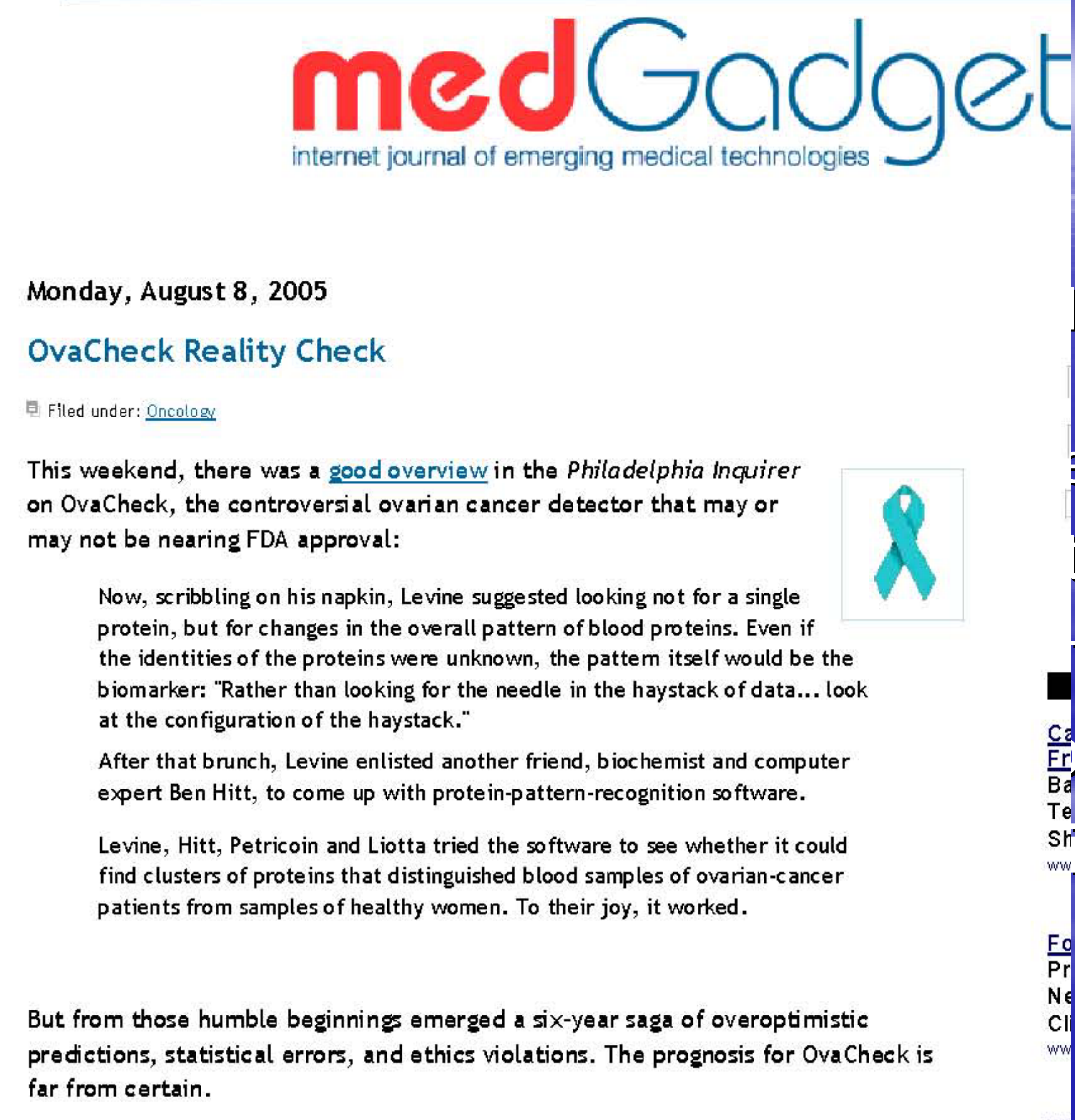
OvaCheck: breakthrough or also ran?

*The test for ovarian cancer uses proteomic pattern testing. It looked good in The Lancet two years ago. The results remain unreplicated*

BY EMILY ANDREWS

# Ovacheck

- Correlo be availabl
  - Consi ler FDA jurisdi
  - No fu ed
- 1/04: C t FDA
- 2/18/04 t
- 7/12/04 oval
- Current ing to see if it



medGadget
internet journal of emerging medical technologies

Monday, August 8, 2005

## OvaCheck Reality Check

Filed under: Oncology

This weekend, there was a good overview in the *Philadelphia Inquirer* on OvaCheck, the controversial ovarian cancer detector that may or may not be nearing FDA approval:

> Now, scribbling on his napkin, Levine suggested looking not for a single protein, but for changes in the overall pattern of blood proteins. Even if the identities of the proteins were unknown, the pattern itself would be the biomarker: "Rather than looking for the needle in the haystack of data... look at the configuration of the haystack."
>
> After that brunch, Levine enlisted another friend, biochemist and computer expert Ben Hitt, to come up with protein-pattern-recognition software.
>
> Levine, Hitt, Petricoin and Liotta tried the software to see whether it could find clusters of proteins that distinguished blood samples of ovarian-cancer patients from samples of healthy women. To their joy, it worked.

But from those humble beginnings emerged a six-year saga of overoptimistic predictions, statistical errors, and ethics violations. The prognosis for OvaCheck is far from certain.

Ca
Fr
Ba
Te
Sh
ww

Fo
Pr
Ne
Cl
ww

# Design lessons

- All samples must be processed using the same protocol
- Randomization should be performed
  - Before sample preparation steps
  - Before acquiring spectra/gels
- May also want to block on important factors – reduce variability – there are ways to filter out systematic block effects
- Same principles should be used for other sensitive laboratory instruments.

# Quantitative Analysis of Proteomics Data

- Look at raw data
- Pre-process
  - Calibration/Alignment
  - Background Corr.
  - Adjust Block Effects
  - Normalization
  - Peak/spot finding
  - Peak/spot quantification
  - Peak/spot matching across spectra/gels
- Look at processed data
- Clean things up
- Data Analysis
  - Clustering
  - T-test, ANOVA
  - Correlating with outcomes
  - Building predictive models
- Look at results
  - Identify proteins and validate them

**"Data is expensive, Analysis is cheap"**

# Data Analysis: Beware of Multiplicities!

- When performing biomarker detection, important to account for multiple tests when declaring biomarker “significant”
  - – If many peaks, $p<0.05$ gives lots of false +
  - – Methods available to control FDR
- When building discriminating model, important to properly validate model
  - – Independent validation samples/cross validation!!
  - – Internal vs. External CV: Cross-validate feature selection step!
  - – Are CV errors relevant for future data?

# Proteomics: Feature Extraction Approach

- Preprocess Data to align data, remove noise, and normalize spectra and gels.
- Extract relevant features from the data, i.e. detect all peaks and spots, and quantify each feature for each spectrum or gel.
  - Results in N x p matrix Y (p features, N spectra)
- Survey N x p matrix Y to find differentially expressed peaks (class comparison) or to build classifier (class prediction), while appropriately accounting for multiplicities.

# Statistical Model for Spectrum

$$Y_i(t_j) = B_i(t_j) + N_i S_i(t_j) + e_{ij}$$

# Statistical Model for Spectrum

$$Y_i(t_j) = \overbrace{B_i(t_j)}^{\text{Baseline Artifact}} + N_i S_i(t_j) + e_{ij}$$

# Statistical Model for Spectrum

$$Y_i(t_j) = \overbrace{B_i(t_j)}^{\text{Baseline Artifact}} + N_i \overbrace{S_i(t_j)}^{\text{Protein Signal}} + e_{ij}$$

# Statistical Model for Spectrum

$$Y_i(t_j) = \overbrace{B_i(t_j)}^{\substack{\text{Baseline}\\ \text{Artifact}}} + \underbrace{N_i}_{\substack{\text{Normal-}\\ \text{ization}\\ \text{Factor}}} \overbrace{S_i(t_j)}^{\substack{\text{Protein}\\ \text{Signal}}} + e_{ij}$$

# Statistical Model for Spectrum

$$Y_i(t_j) = \overbrace{B_i(t_j)}^{\text{Baseline Artifact}} + \underbrace{N_i}_{\text{Normalization Factor}} \overbrace{S_i(t_j)}^{\text{Protein Signal}} + \underbrace{e_{ij}}_{\text{additive noise (detector)}}$$

$$e_{ij} \sim N\{0, \sigma^2(t_j)\}$$

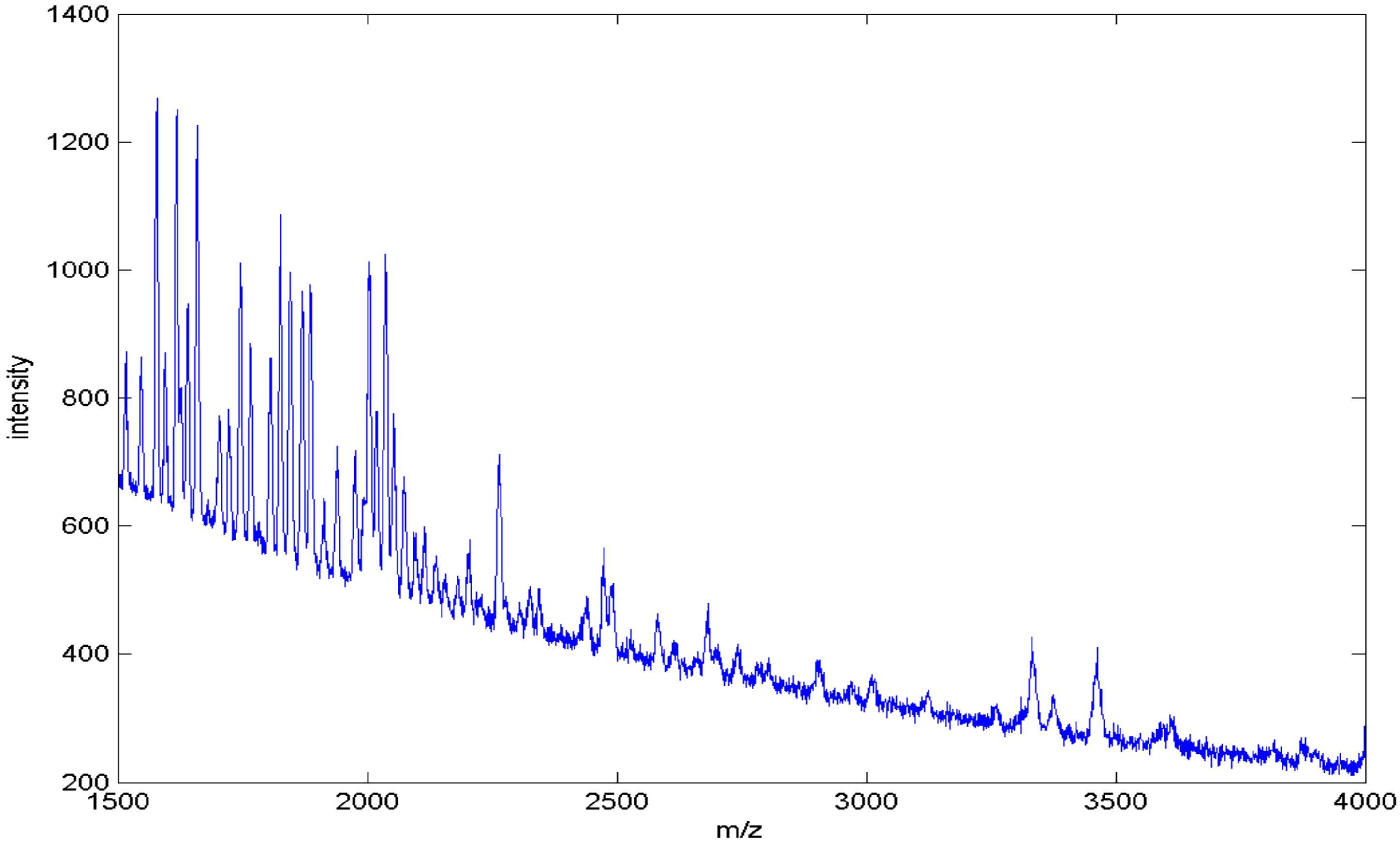
# Preprocessing

- **Goal:** Isolate protein signal $S_i(t_j)$
  - – Filter out baseline and noise, normalize
  - – Extract individual features from signal
- **Problem:**
  - – Baseline removal, denoising, normalization, and feature extraction are interrelated processes.
  - – Where do we start?

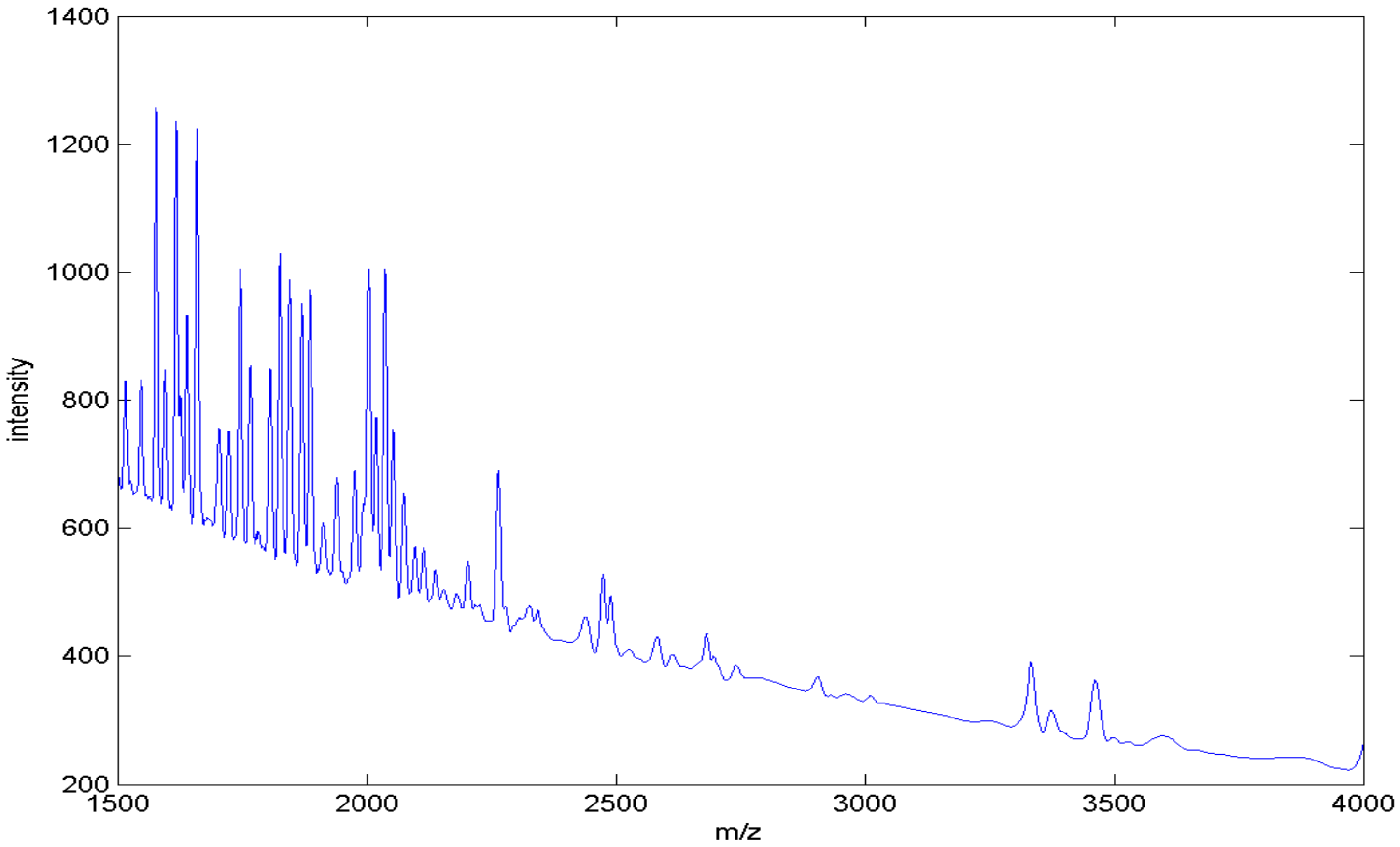
# Denoising using Wavelets

- First step: Isolate noise using wavelets
  - Wavelets: basis functions that can parsimoniously represent spiky functions
  - Standard denoising tool in signal processing
- Idea: Transform from time to wavelet domain, threshold small coefficients, transform back.
  - **Result:** Denoised function and noise estimate
  - **Why does it work?** Signal concentrated on few wavelet coefficients, white noise equally distributed. Thresholding removes noise without affecting signal.
- Does *much* better than denoising tools based on kernels or splines, which tend to attenuate peaks in the signal when removing the noise.

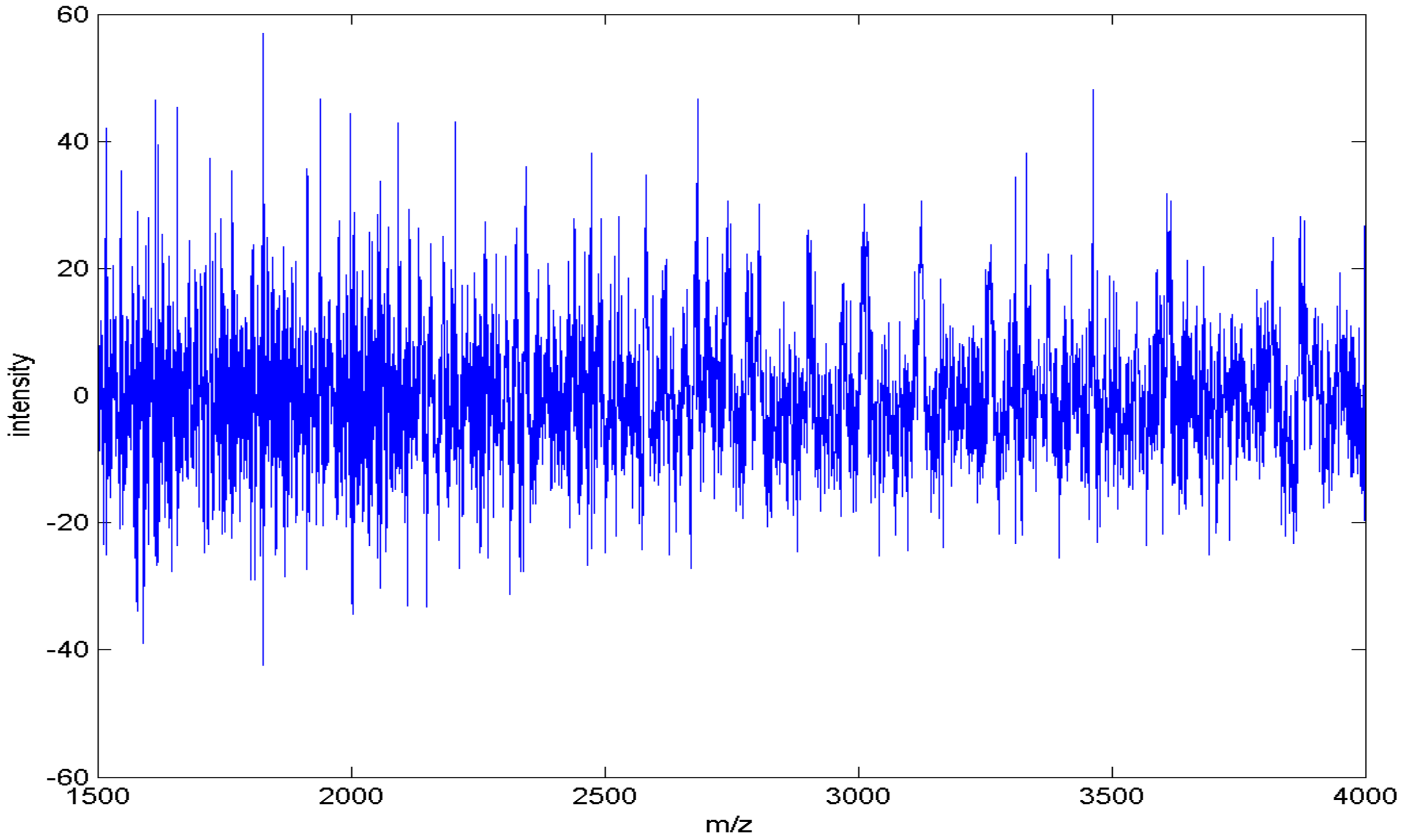
# Raw Spectrum



# Denoised Spectrum



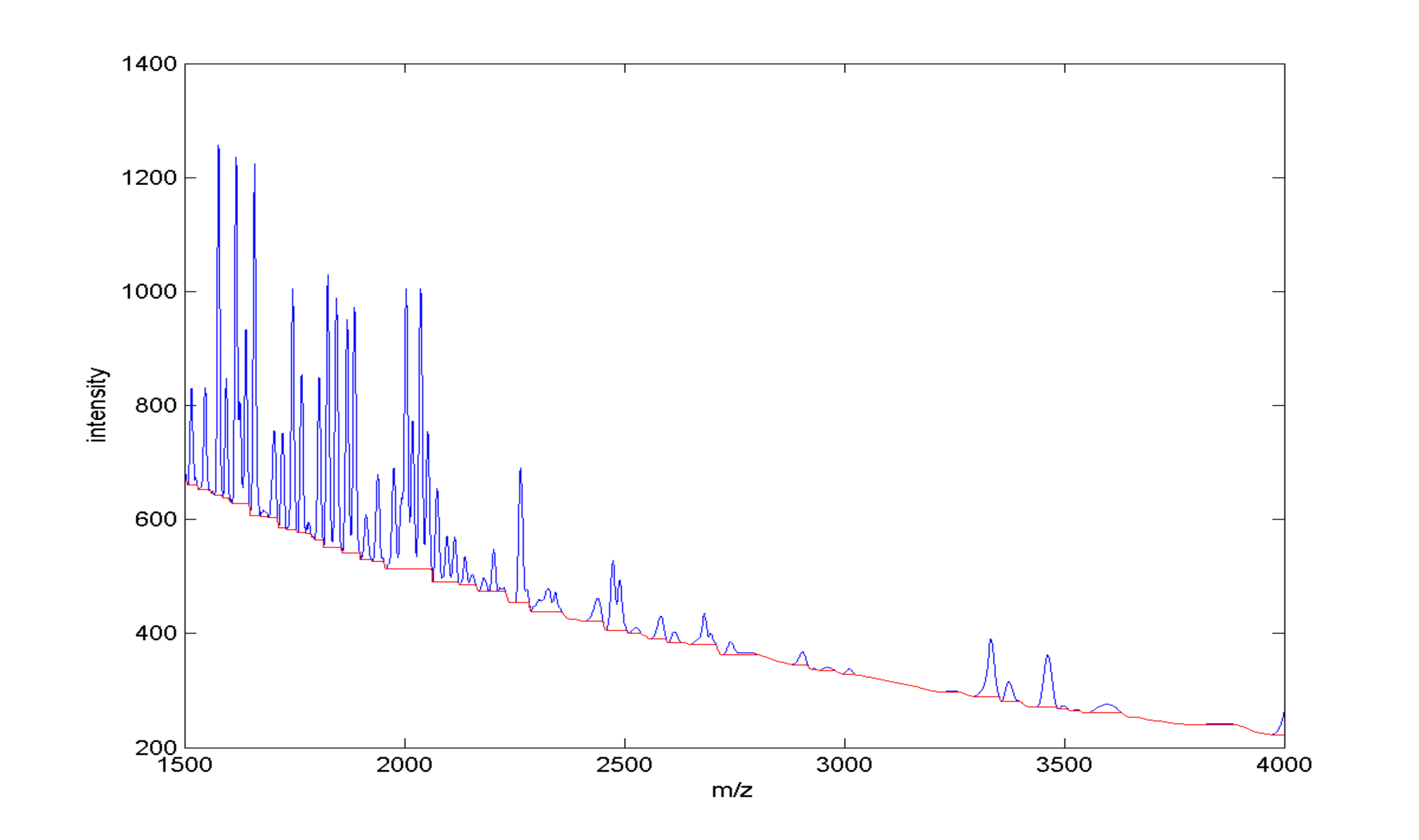
# Noise



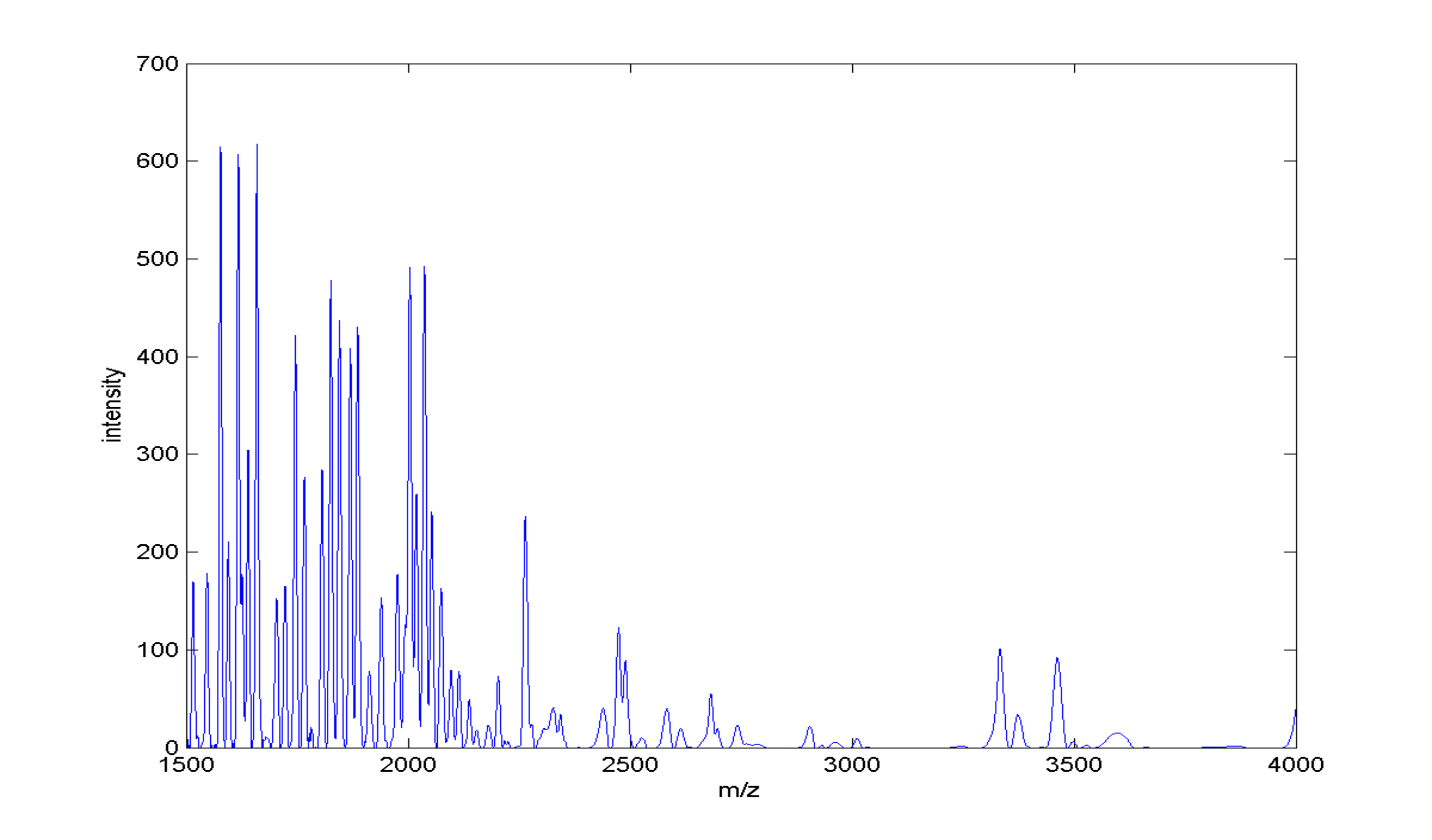
# Baseline Correction & Normalization

- **Baseline:** smooth artifact, largely attributable to detector overload.
  - Estimated by monotone local minimum
  - More stably estimated after denoising
- **Normalization:** adjust for possibly different amounts of material desorbing from plates
  - Divide by total area under the denoised and baseline corrected spectrum.

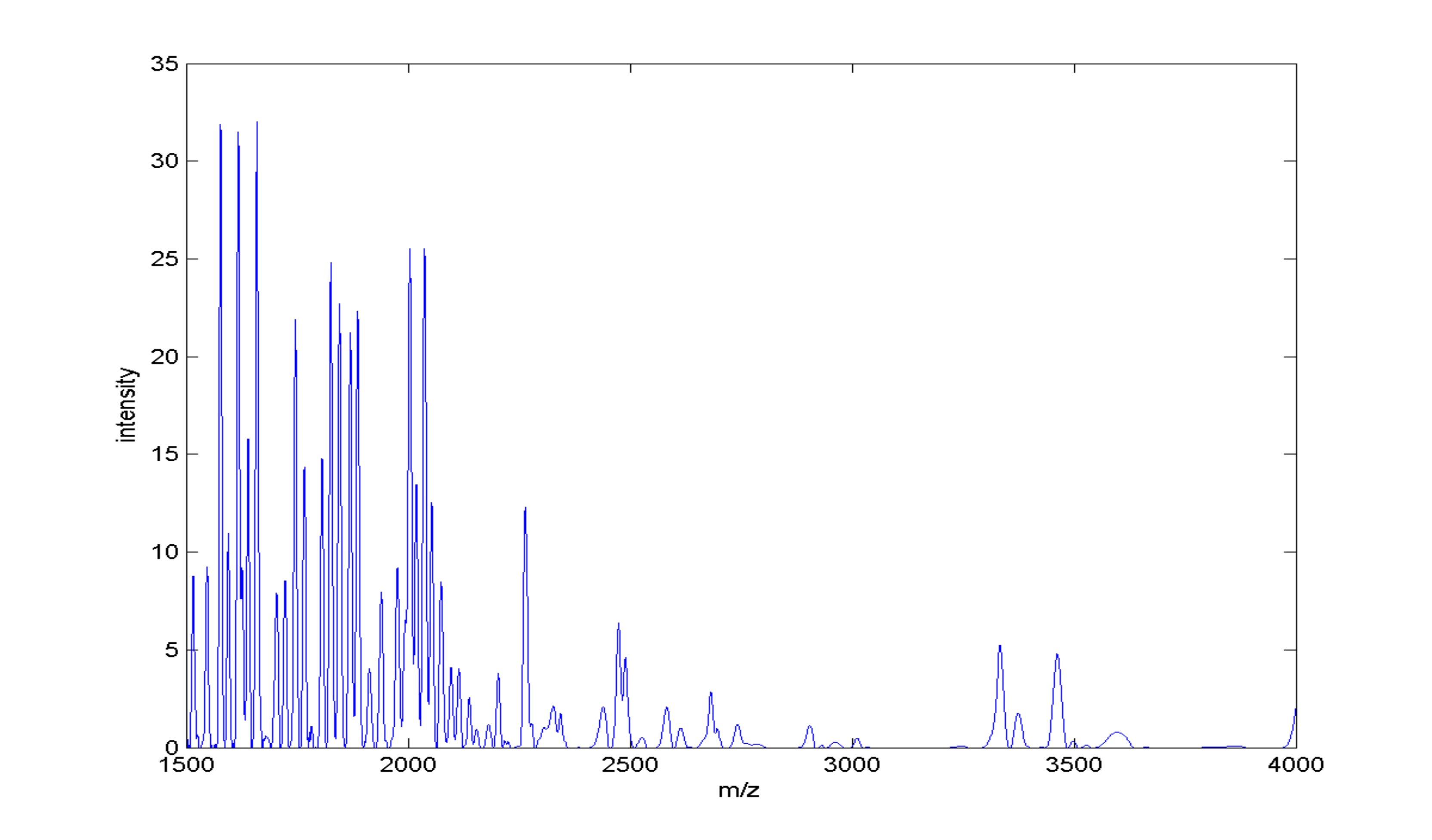
# Baseline Estimate



# Denoised, Baseline Corrected Spectrum



# Denoised, Baseline Corrected, and Normalized Spectrum



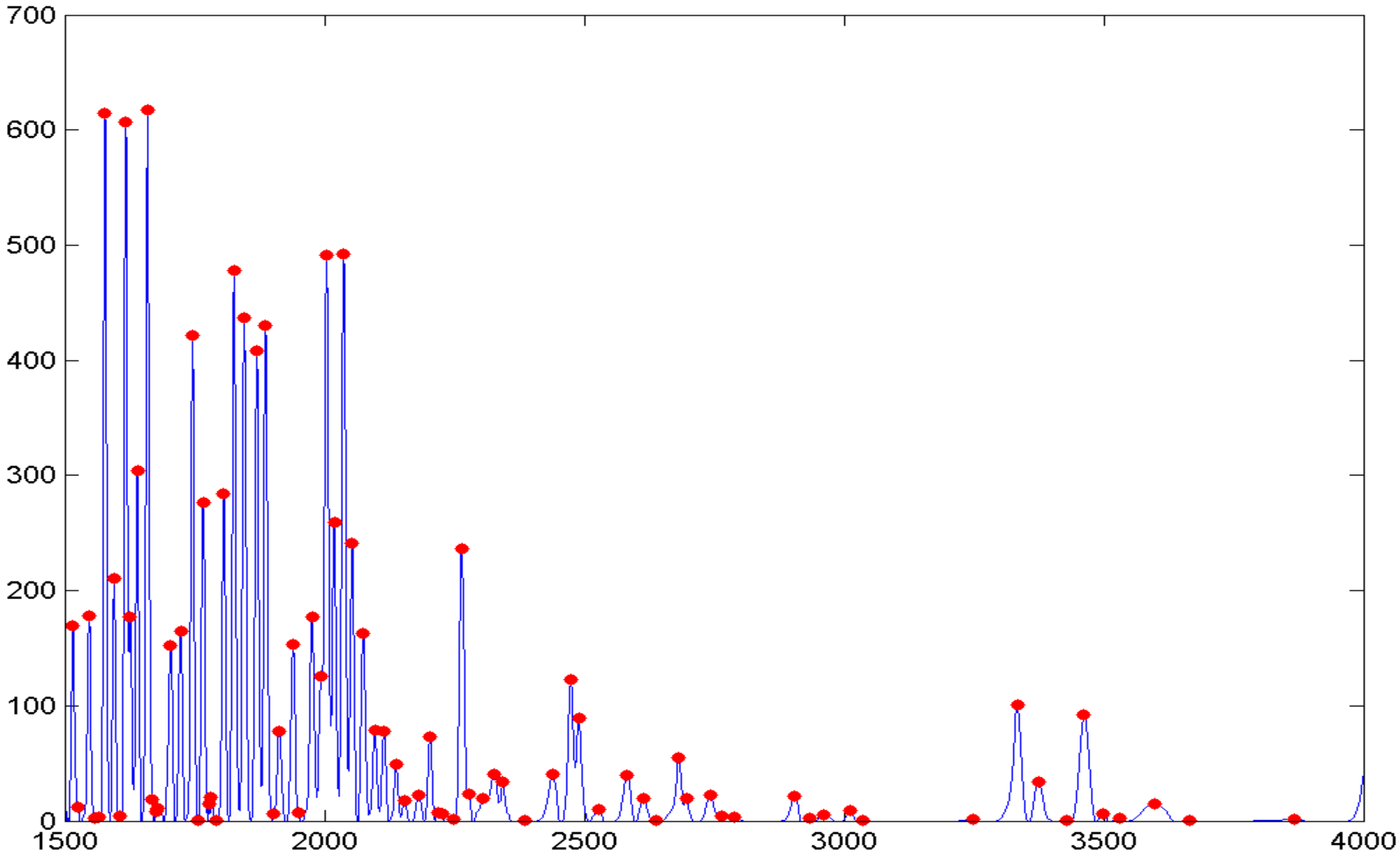
# Protein Signal

- **Ideal Form of Protein Signal**: Convolution of peaks
  - Proteins, peptides, and their alterations
  - **Alterations**: isotopes; matrix/sodium adducts; neutral losses of water, ammonia, or carbon
- Limitations of instrument used means we may not be able to resolve all peaks.
- Advantages of peak detection:
  - Reduces multiplicity problem
  - Focuses on units that are theoretically the scientifically interesting features of the data.

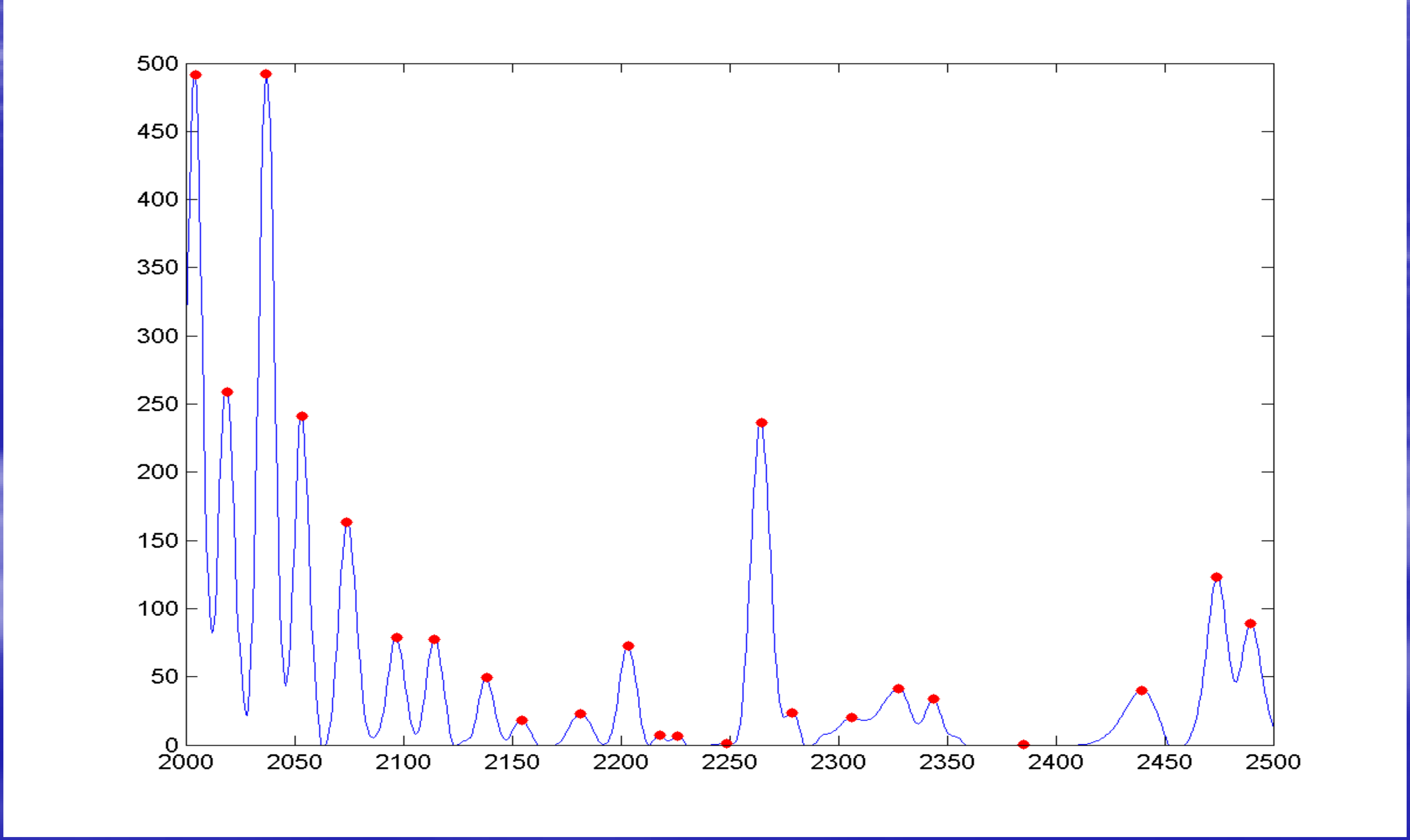
# Peak Detection

- Easy to do after other preprocessing
- Any local maximum after denoising, baseline correction, and normalization is assumed to correspond to a “peak”.
- May want to require S/N>$\delta$ to reduce number of spurious peaks.
  - We can estimate the noise process $\sigma(t)$ *by* applying a local median to the filtered noise from the wavelet transform.
  - Signal-to-noise estimate is ratio of preprocessed spectrum and noise.

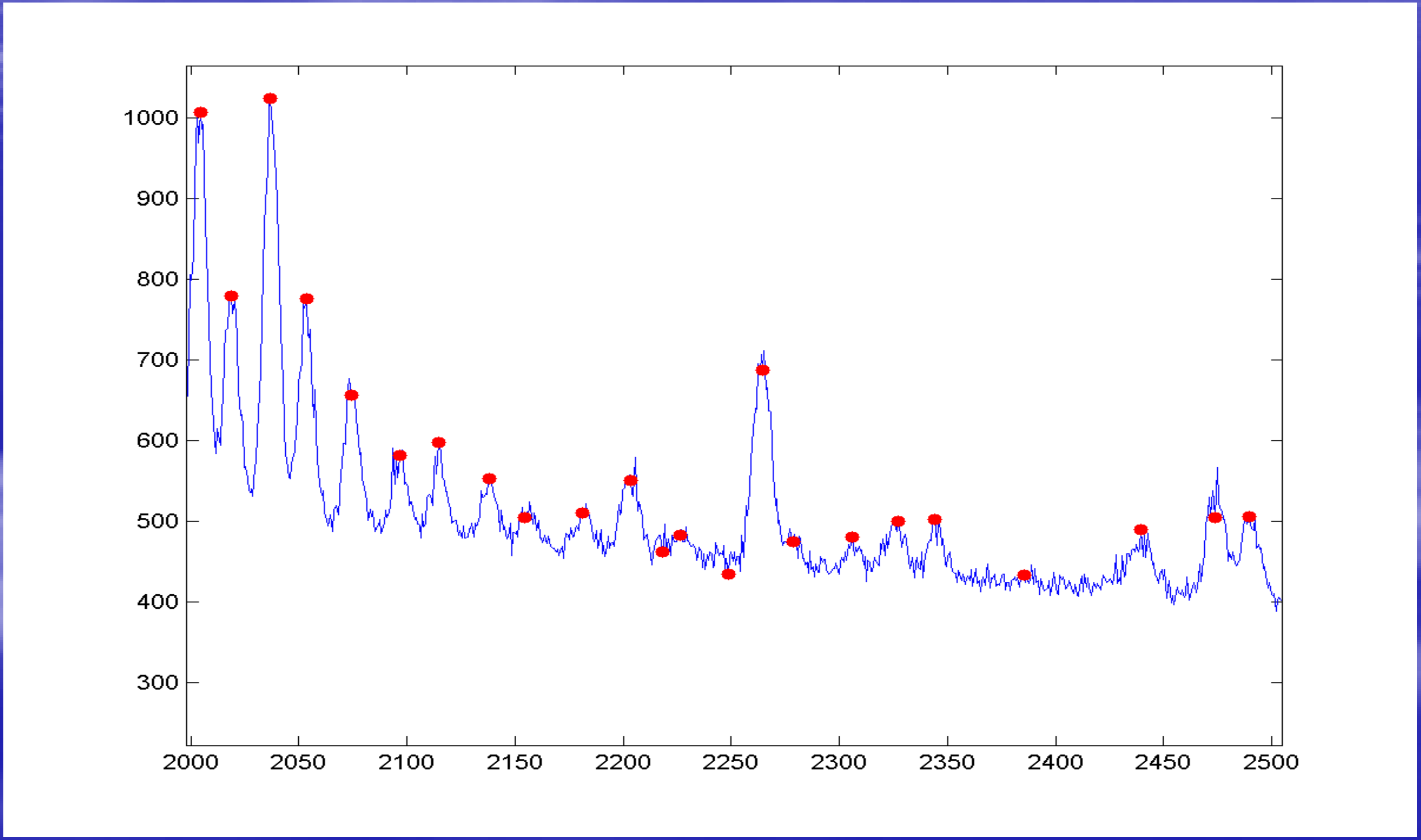
# Peak Detection



# Peak Detection (zoomed)



# Raw Spectrum with peaks



# Peak Quantification

- Two options:
  1. Area under the peak: Find the left and right endpoints of the peak, compute the AUC in this interval.
  2. Maximum intensity: Take intensity at the local maximum (may want to take log or cube root)
- Theoretically, AUP quantifies amount of given substance desorbed from the chip.
  - But it is very difficult to identify the endpoints of peaks

# Peak Quantification

- The maximum intensity is a practical alternative
  - No need for endpoints, should be correlated with AUP
  - Physics of mass spectrometry shows that, for a given ion with m/z value *x*, there is a **linear relationship** between the **number of ions** of that type desorbed from plate and the **expected maximum peak intensity** at *x.*
- Problem with both methods: Overlapping peaks that are not deconvolvable
  - Local maximum at *t* contains weighted average of information from multiple ions whose corresponding peaks have mass at location *t*.
  - Major problem – short of formal deconvolution, have not seen simple solution to this problem.

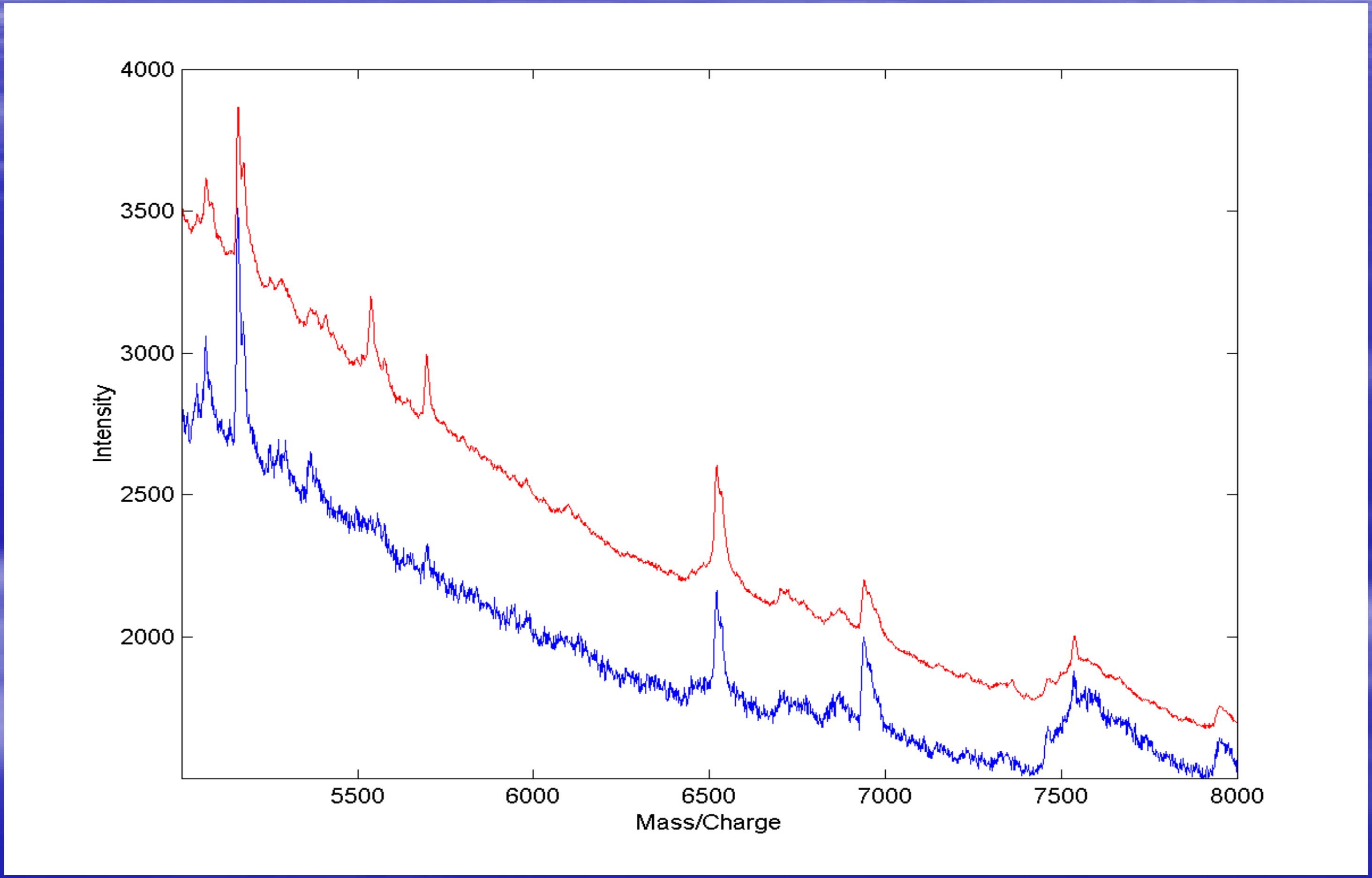
# Peak Matching Problem

- If peak detection performed on individual spectra, peaks must be matched across samples to get n *x* p matrix.
  - Difficult and arbitrary process
  - What to do about “missing peaks?”
- **Our Solution:** Identify peaks on **mean spectrum** (at locations $x_1, \ldots, x_p$), then quantify peaks on individual spectra by intensities at these locations.

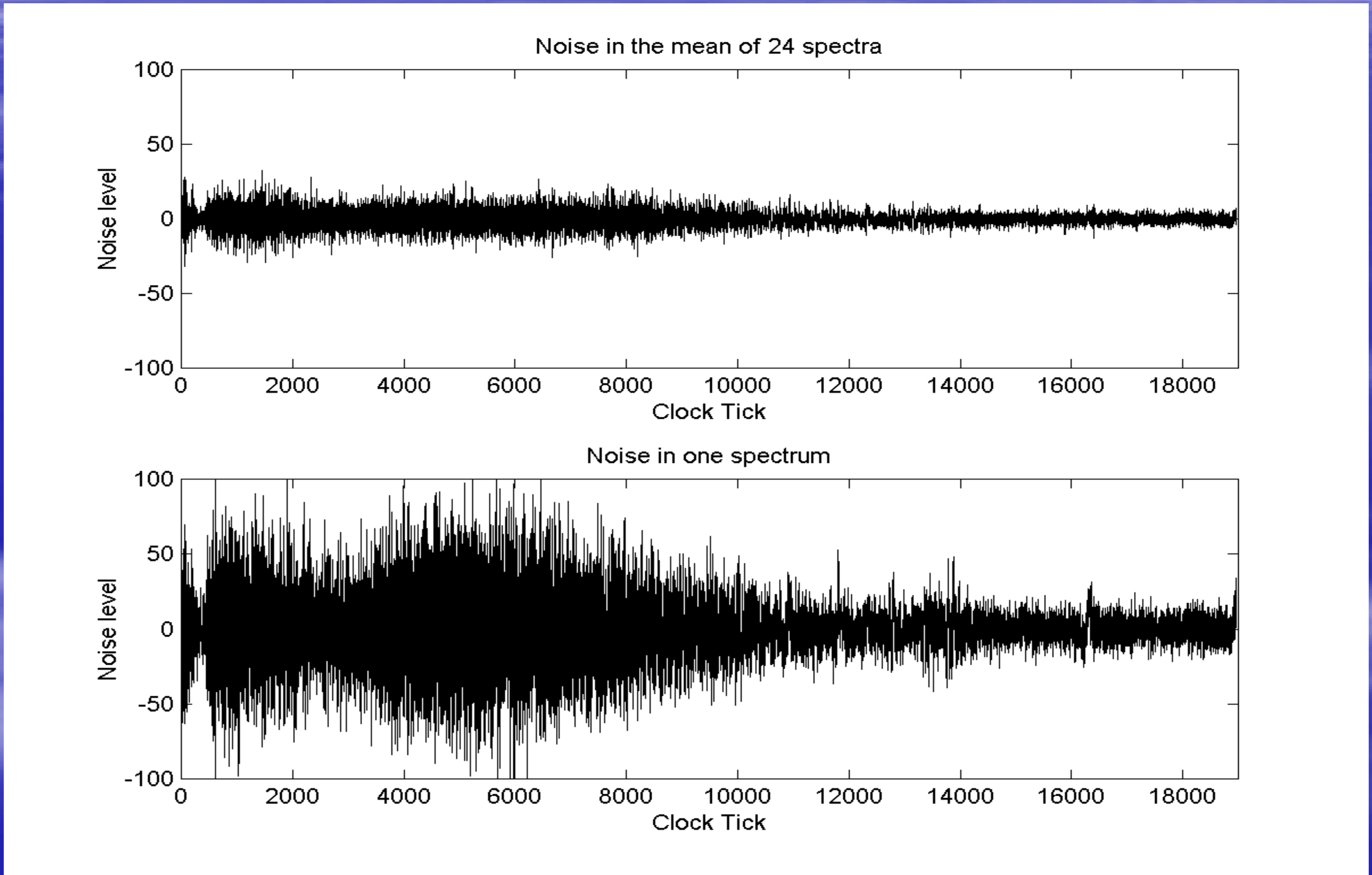
# Advantages/Disadvantages

- Advantages
  - Avoids peak-matching problem
  - Generally more sensitive and specific
    - Noise level reduced by sqrt(n)
    - Borrows strength across spectra in determining whether there is a peak or not (signals reinforced over spectra)
  - Robust to minor calibration problems
- Disadvantage
  - Tends to be less sensitive when prevalence of peak < 1/sqrt(n).

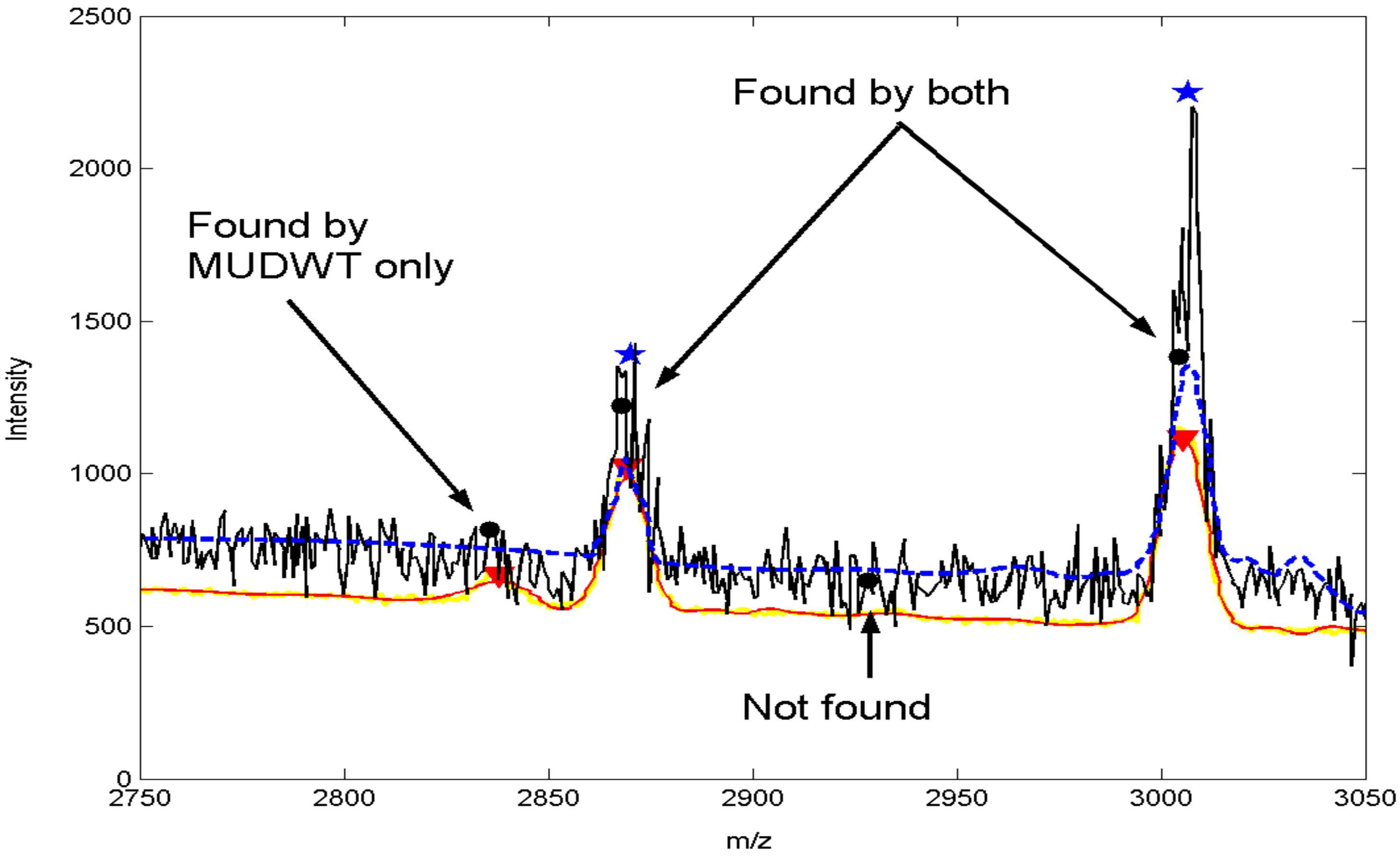
# Noise reduced in mean spectrum



# Noise reduced in mean spectrum



# Sample Spectrum



# Simulation Study

1. Generated 100 random virtual populations based on MDACC MALDI study on pancreatic cancer.
2. For each virtual population, generated 100 virtual samples, obtained 100 virtual spectra.
3. Applied preprocessing and peak detection method based on individual and average spectra
4. Summarized performance based on sensitivity (proportion of proteins detected) and FDR (proportion of peaks corresponding to real proteins).
   - Tricky to do – see paper for details.

# Simulation Results

## Overall Results

| | sensitivity | FDR | pv* |
|---|---|---|---|
| **SUDWT**<br>(indiv. spectra) | 0.75 | 0.09 | 0.03 |
| **MUDWT**<br>(mean spectrum) | 0.83 | 0.06 | 0.97 |

*pv=the proportion of simulations with higher sensitivity

# Simulation Results

## By Prevalence

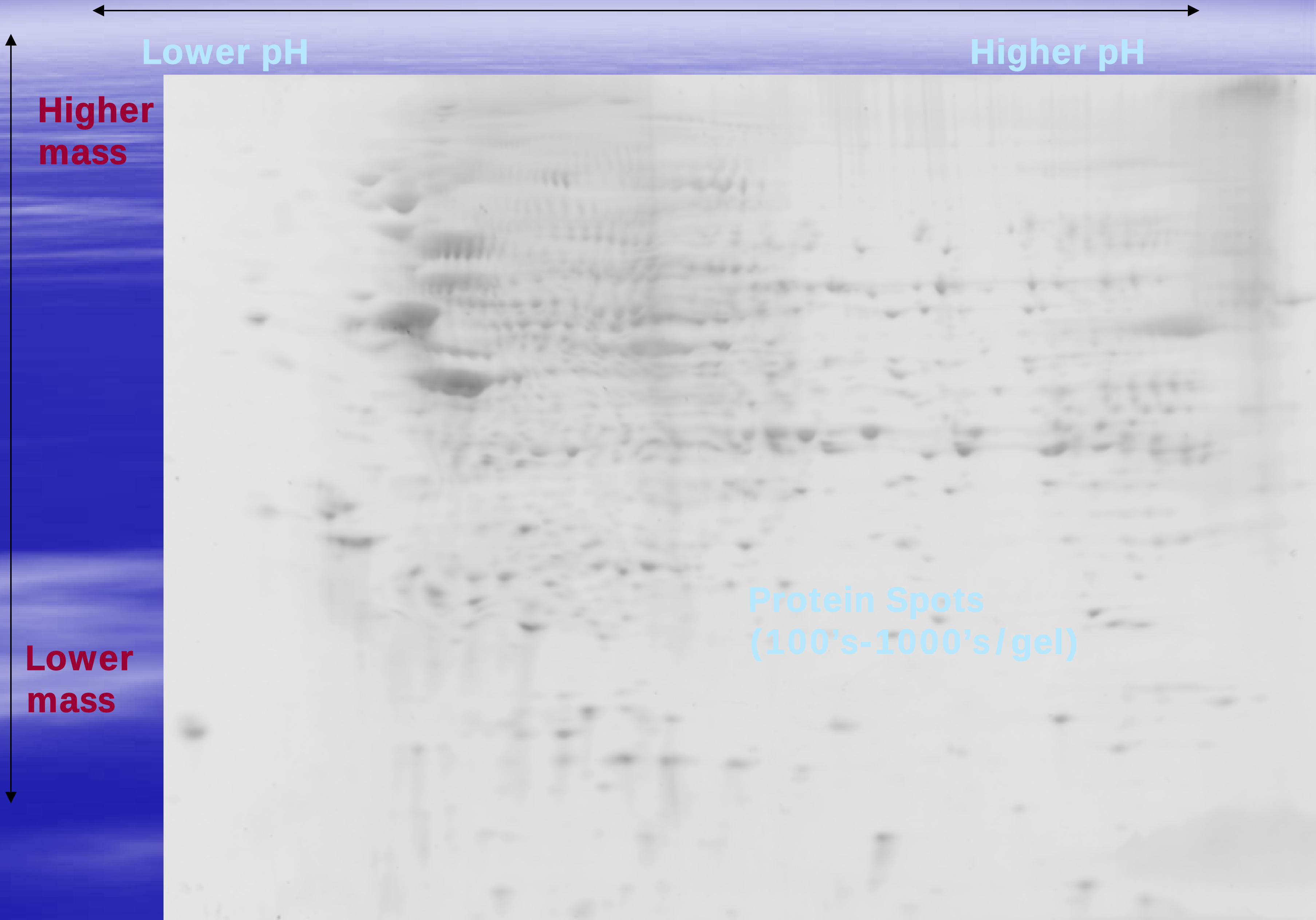
| $\pi$: | **<.05**<br>(14%) | **.05-.20**<br>(16%) | **.20-.80**<br>(40%) | **>.80**<br>(30%) |
|---|---|---|---|---|
| **sensitivity**<br>**(SUDWT)** | 0.43 | 0.74 | 0.81 | 0.82 |
| **sensitivity**<br>**(MUDWT)** | 0.38 | 0.74 | 0.93 | 0.97 |
| **pv**<br>**(MUDWT)** | 0.25 | 0.49 | 1.00 | 1.00 |

# Simulation Results

## By Abundance (mean log intensity)

| **log(μ):** | **<9.0**<br>(31%) | **9.0-9.5**<br>(27%) | **9.5-10**<br>(23%) | **>10**<br>(19%) |
|---|---|---|---|---|
| **sensitivity**<br>**(SUDWT)** | 0.68 | 0.75 | 0.78 | 0.82 |
| **sensitivity**<br>**(MUDWT)** | 0.78 | 0.84 | 0.85 | 0.88 |
| **pv**<br>**(MUDWT)** | 0.97 | 0.89 | 0.84 | 0.78 |

# 2-D Gel Electrophoresis\



Lower pH

Higher pH

Higher mass

Lower mass

Protein Spots
(100’s-1000’s / gel)

# Why Is Gel Analysis So Difficult?

- Usual Approach
  - Normalize individual gels
  - Detect spots and draw spot boundaries on individual gels
  - Match spots on each gel with spots on a chosen “reference” gel
  - Quantify spots by taking spot volumes

# Problems With The “Usual” Approach

- Complicated, error-prone algorithms
  - Spot detection errors (miss/split/merge)
  - Spot matching errors
  - Errors in spot boundary determination
- Errors tend to increase with number of gels
- Much hand editing required
  - Reduces objectivity and reproducibility of analysis
- Missing spots negatively impact statistical analysis

# What If We Could Eliminate the Complex Algorithms?

- Eliminate the need for spot matching
- Sum data across gels to objectively detect spots (create an “average” gel)
  - Detection power *increases* with more gels
- Eliminate need to draw spot boundaries
- Eliminate the problem of missing spots
- Eliminate the need for hand editing

# Preprocessing 2d gels

## Our Approach: *Pinnacle* Method

- Align gel images
- Compute average gel
- Denoise average gel using wavelets
- Detect spots on average gel using *pinnacles*
- Background correct and normalize individual gels
- Quantify each spot on each individual gel by taking maximum pixel intensity in neighborhood of pinnacle

# Image Registration

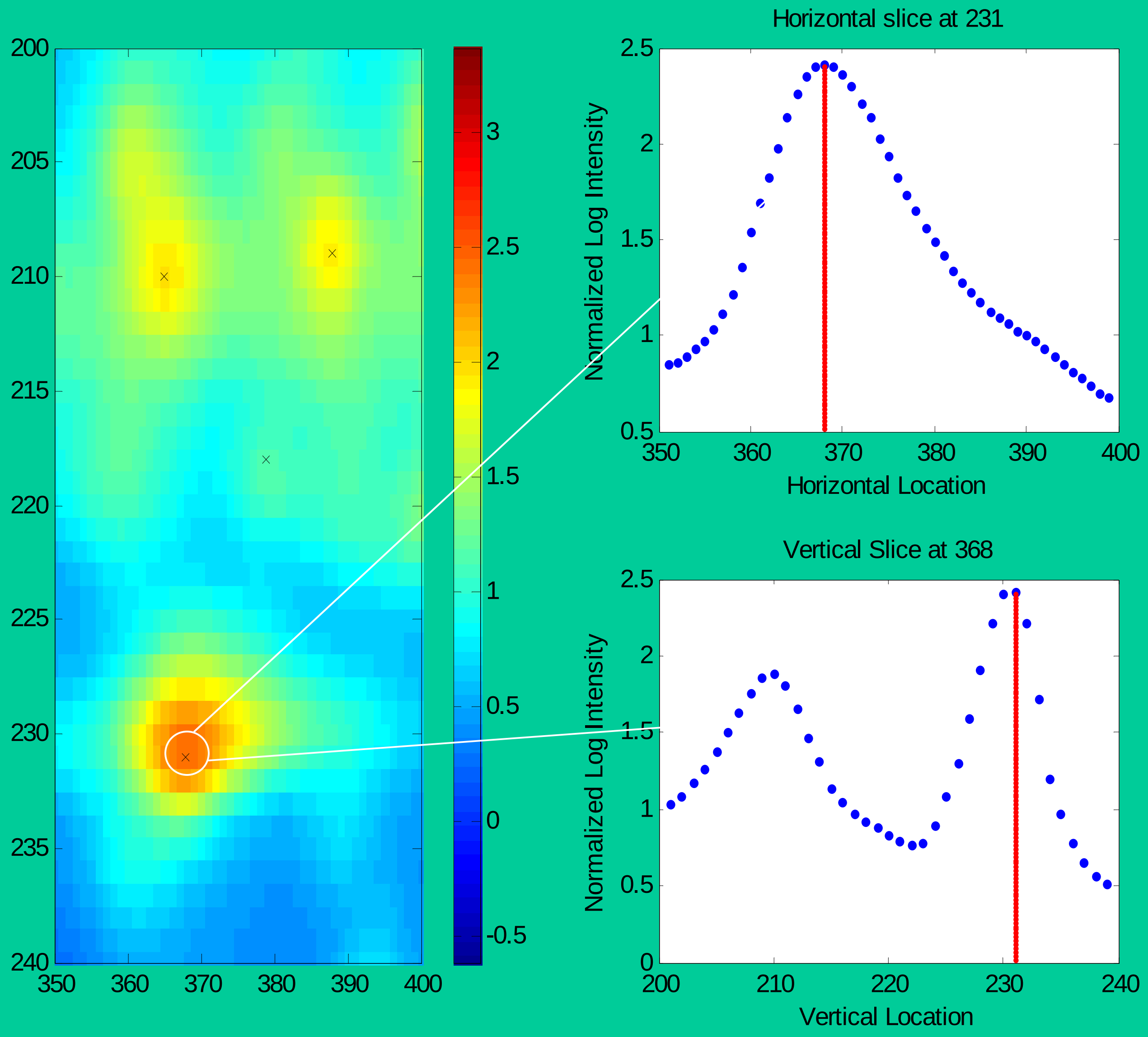
- Align all gels to chosen reference gel so spots are aligned across gels
  - Easier and more accurate than matching detected spots, since algorithm can borrow strength from nearby regions of the gel when aligning spots
- We use TT900 (Nonlinear Dynamics) to align gels; other image alignment programs are available

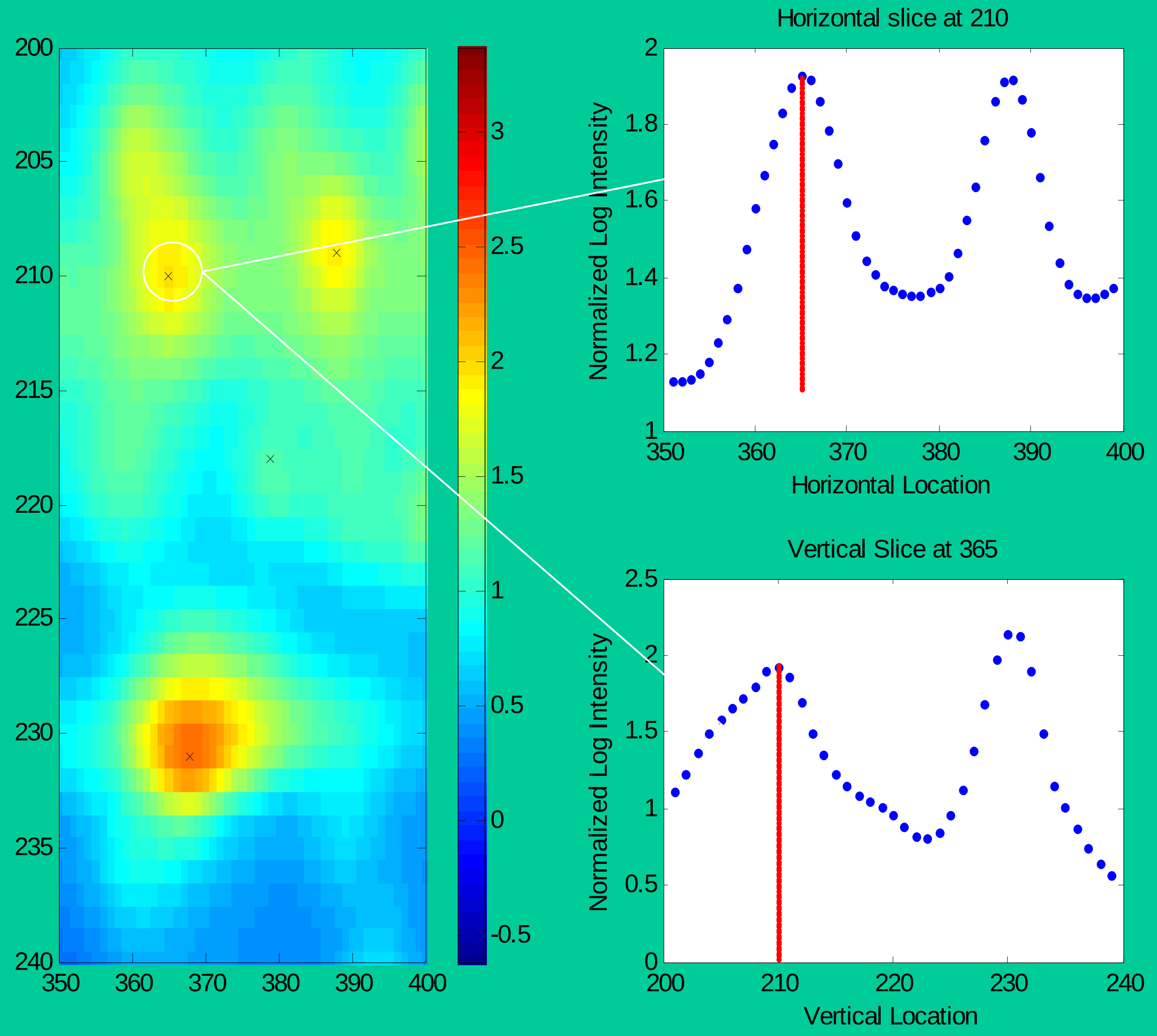
# Spot Detection

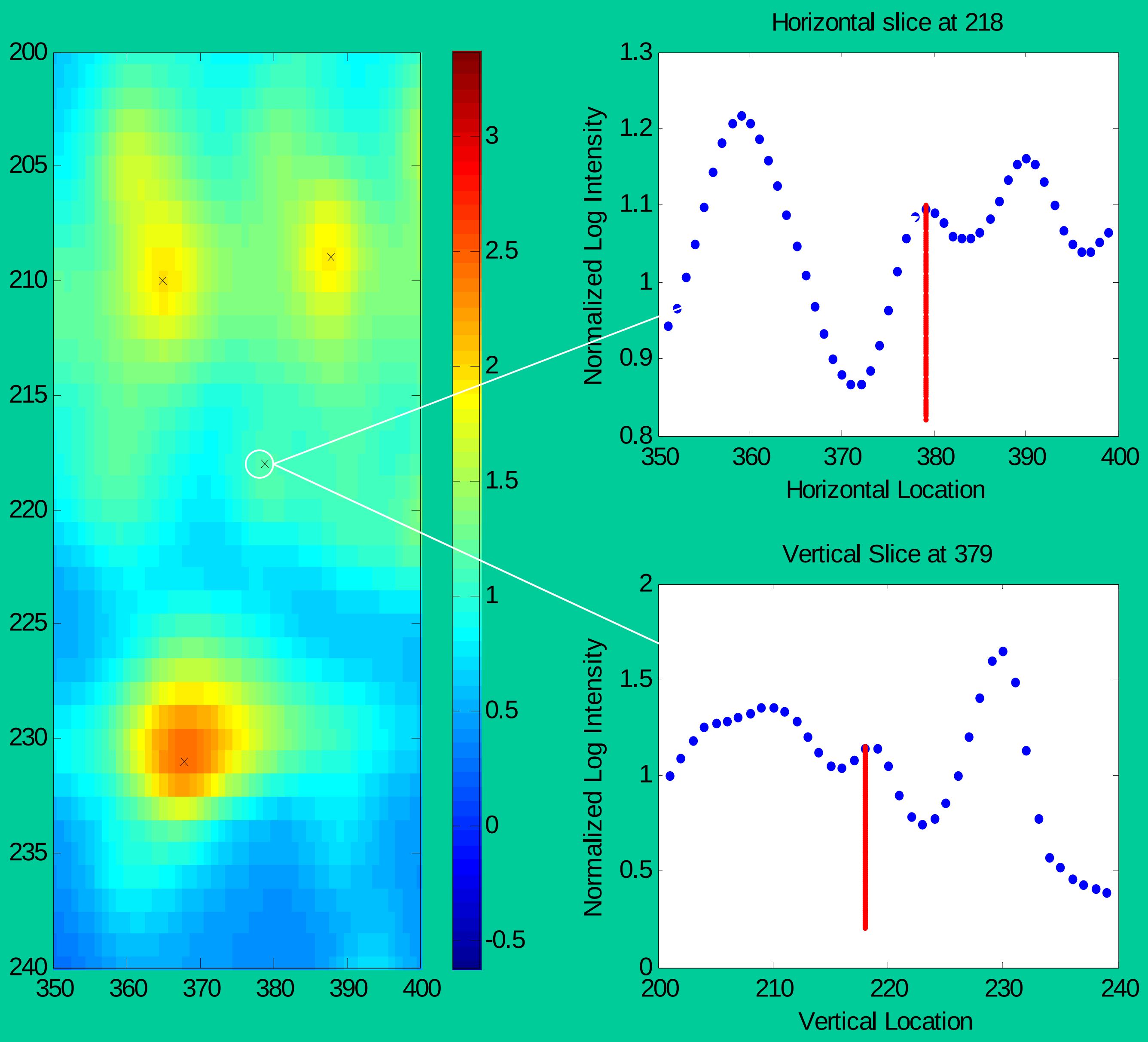
- First, compute the average gel ***Z***, by taking the mean over all $X_{ij}$ for each $i,j$
- Why use average gel for spot detection?
  - Avoids spot-matching problem (missing spots)
  - More sensitive and specific in identifying spots
    - "Real" spots will be present in multiple gels, so will be reinforced,
    - Artifacts will not, so tend to be averaged out.
  - The noise level reduced by $\sqrt{N}$
- Morris, et al. (Bioinformatics, 21:1764-1775, 2005) demonstrated this principle for peak detection for MALDI-MS (1-d case)
- Requires that gel images are aligned

# Pinnacle Method

- We identify spots based on their corresponding ***pinnacles***
- Location i,j on the gel is ***pinnacle*** if it is a ***peak*** (local maximum) in both the horizontal and vertical directions, AND
  - ***Intensity*** $Z_{ij}$>=***d*** : Must have certain minimum intensity (default is 75th percentile on gel)
- Also, combine together any pinnacles within +/- ***q*** pixel values apart (default q=2)


200
205
210
215
220
225
230
235
240
350
360
370
380
390
400
3
2.5
2
1.5
1
0.5
0
-0.5
Horizontal slice at 231
Normalized Log Intensity
2.5
2
1.5
1
0.5
350
360
370
380
390
400
Horizontal Location
Vertical Slice at 368
Normalized Log Intensity
2.5
2
1.5
1
0.5
0
200
210
220
230
240
Vertical Location


200
205
210
215
220
225
230
235
240
350
360
370
380
390
400
3
2.5
2
1.5
1
0.5
0
-0.5
Horizontal slice at 210
Normalized Log Intensity
2
1.8
1.6
1.4
1.2
1
350
360
370
380
390
400
Horizontal Location
Vertical Slice at 365
Normalized Log Intensity
2.5
2
1.5
1
0.5
0
200
210
220
230
240
Vertical Location



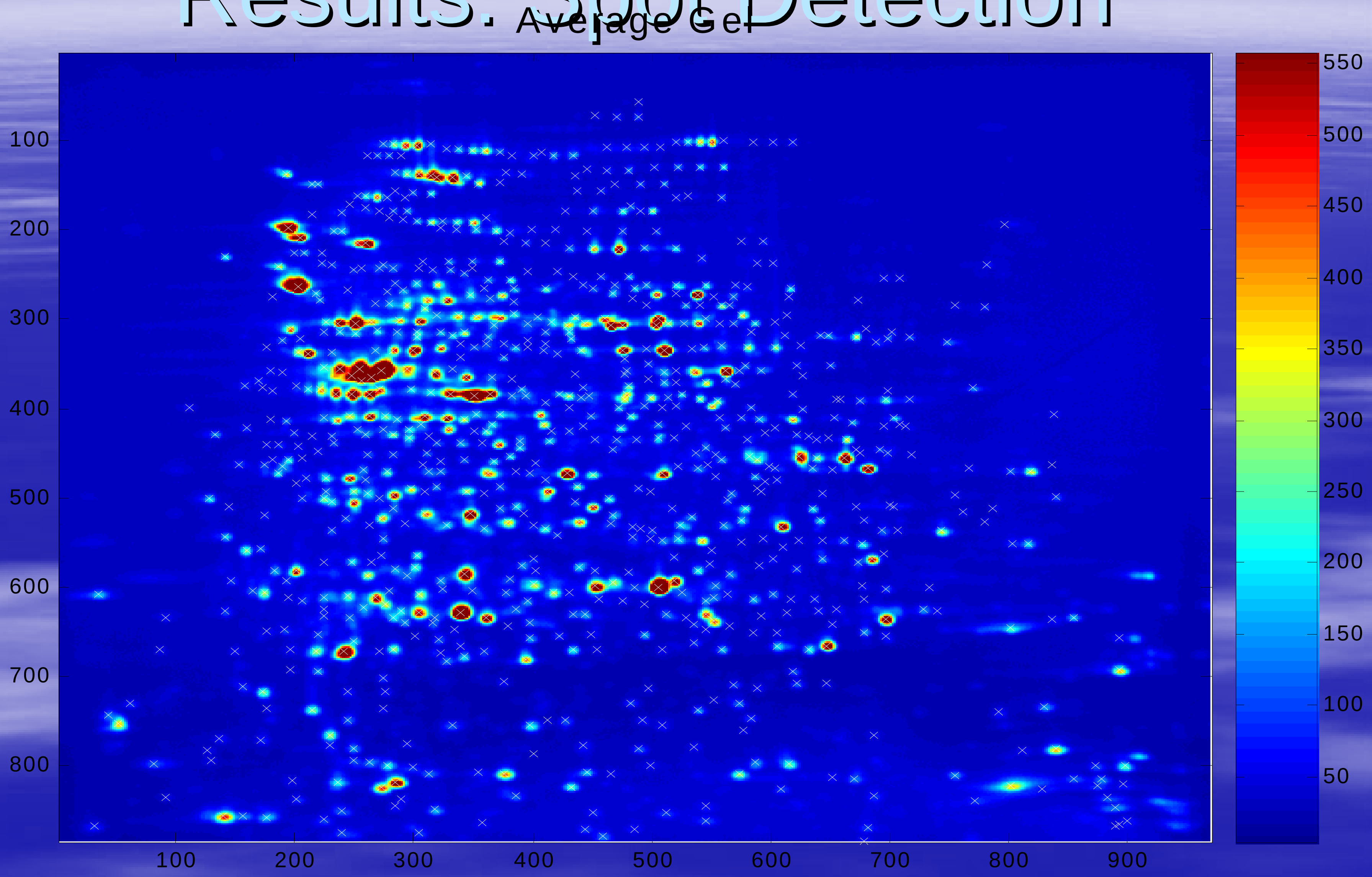
200
205
210
215
220
225
230
235
240
350
360
370
380
390
400
3
2.5
2
1.5
1
0.5
0
-0.5
Horizontal slice at 218
Normalized Log Intensity
1.3
1.2
1.1
1
0.9
0.8
350
360
370
380
390
400
Horizontal Location
Vertical Slice at 379
Normalized Log Intensity
2
1.5
1
0.5
0
200
210
220
230
240
Vertical Location

# Spot Detection

- **Benefits of using Pinnacles for Spot Detection:**

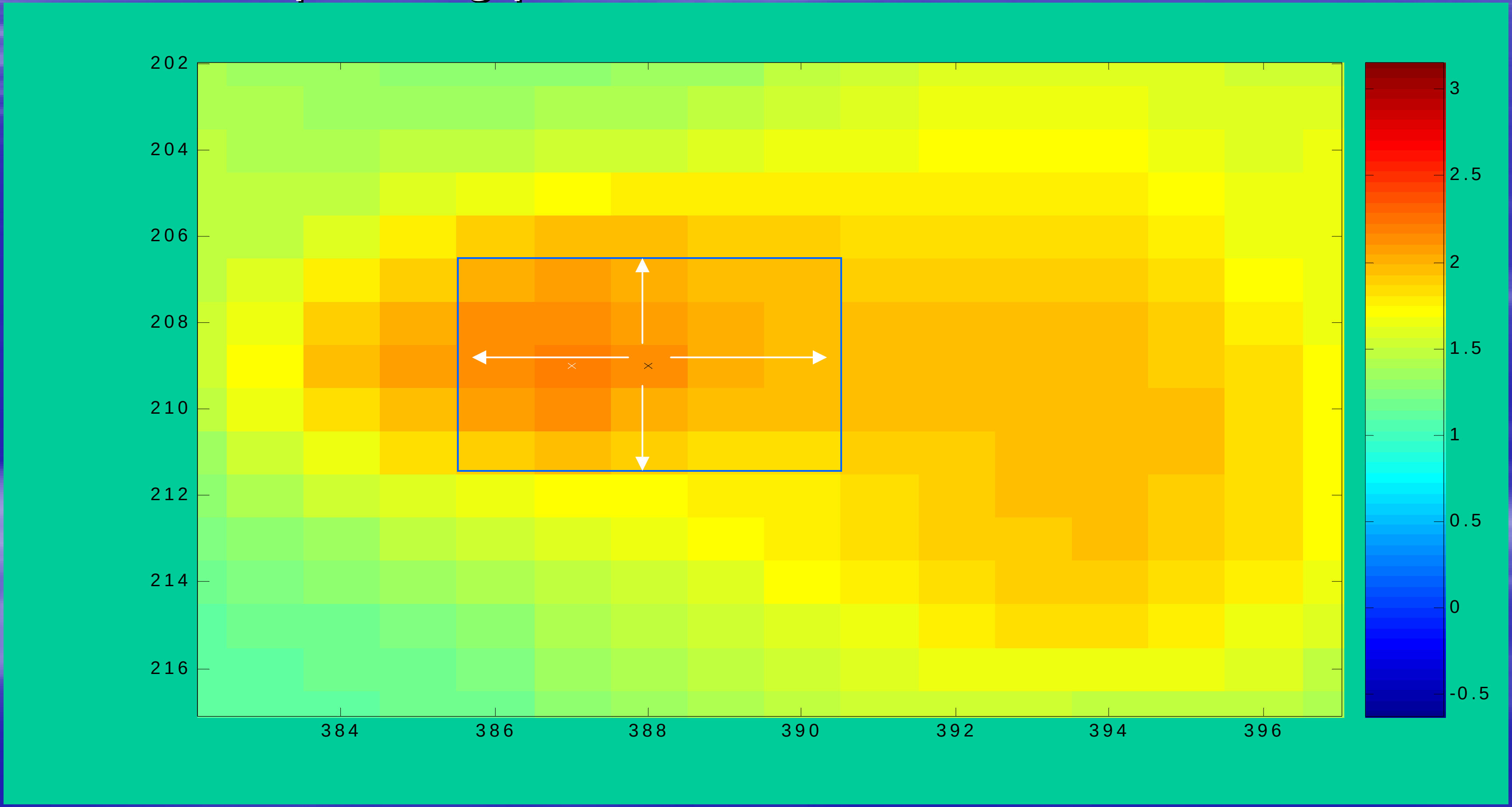
1. Unambiguous definition
2. Not affected by overlapping spots
3. No need to find spot boundaries
4. Excellent Sensitivity and Specificity

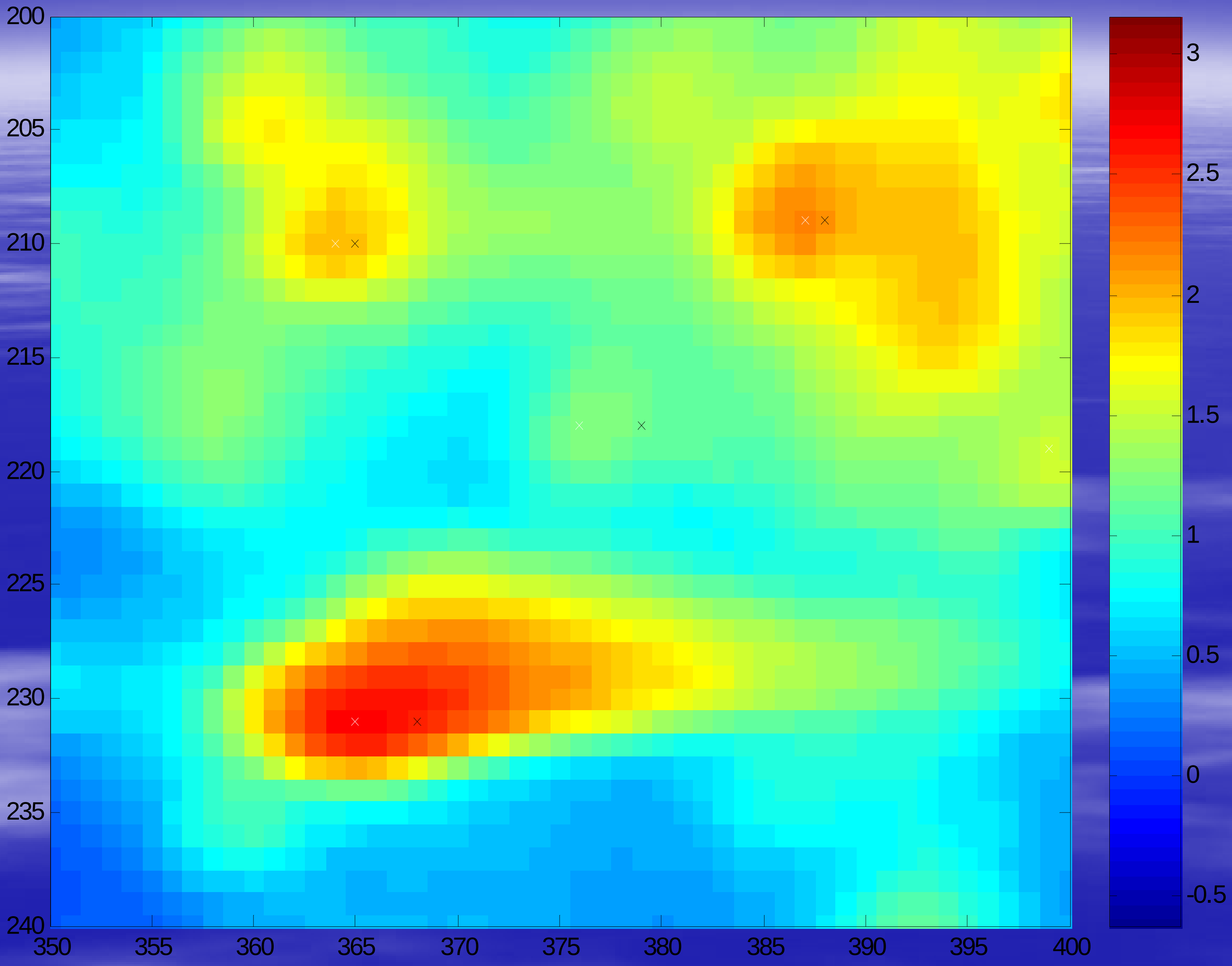
# Results: Spot Detection

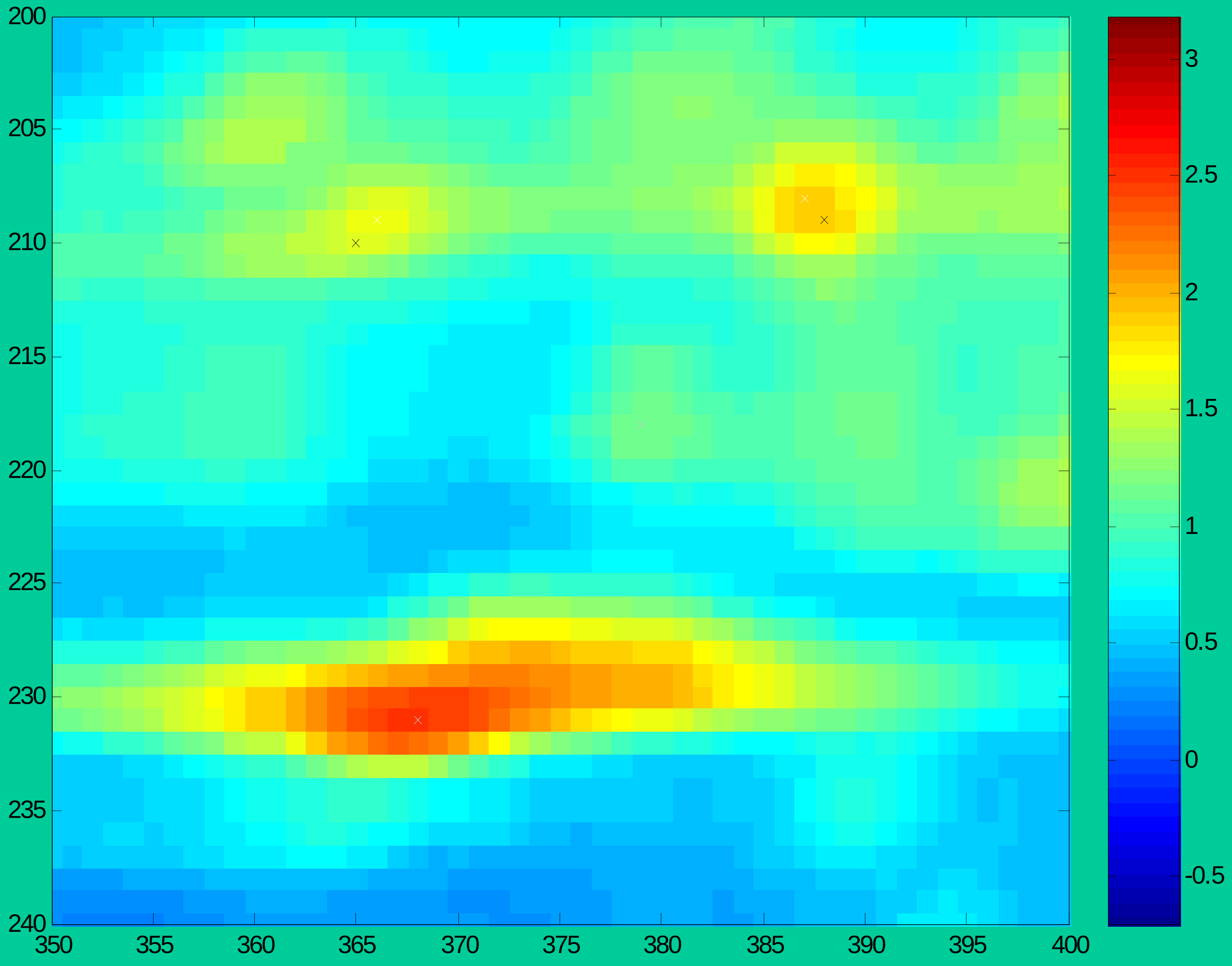


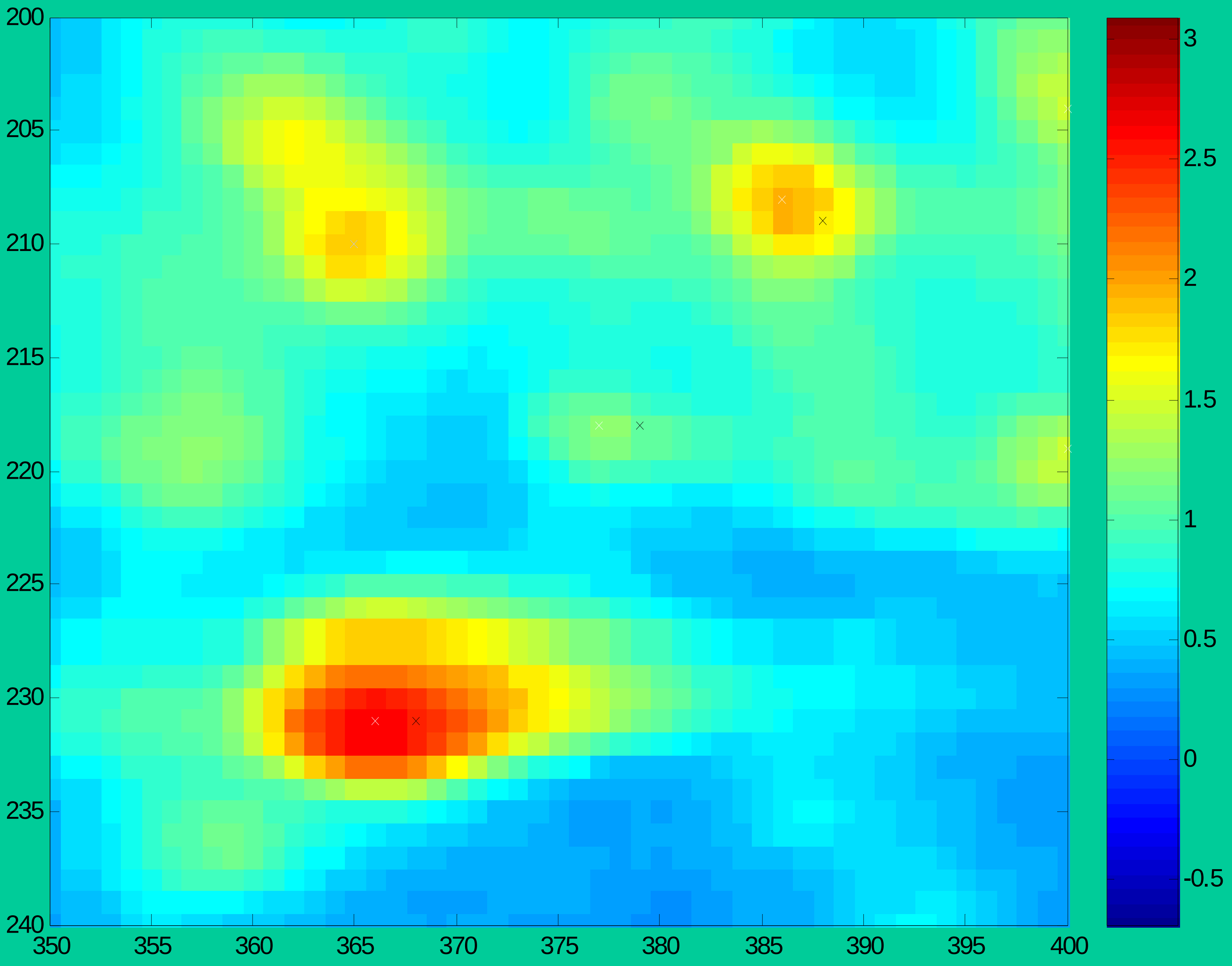
# Spot Quantification

- We quantify each spot for each gel by taking the maximum pixel intensity within a neighborhood around the corresponding pinnacle




200
205
210
215
220
225
230
235
240
350
355
360
365
370
375
380
385
390
395
400
3
2.5
2
1.5
1
0.5
0
-0.5


200
205
210
215
220
225
230
235
240
350
355
360
365
370
375
380
385
390
395
400
3
2.5
2
1.5
1
0.5
0
-0.5


200
205
210
215
220
225
230
235
240
350
355
360
365
370
375
380
385
390
395
400
3
2.5
2
1.5
1
0.5
0
-0.5

# Why Use Pinnacle Intensities for Spot Quantification?

- Pinnacle intensity highly correlated with volume
- No need to detect spot boundaries
  - Reduces complexity/error
  - Reduces CV of quantification
- No missing values
  - Pinnacle intensity for each spot in every gel
- This approach leads to more reliable and precise spot quantifications

# Normalization/Background Correction

- Background Correction:
  - Global: subtract minimum value on gel
  - Local: subtract minimum value within window around pinnacle (e.g. +/- 100 pixels)
- Normalization:
  - Total Volume Normalization: Divide each pinnacle by the sum of all pixels on the gel
  - Sum of Pinnacles Normalization: Divide each pinnacle by the sum of all pinnacles on the gel

# Validation: Dilution Series

- Nishihara and Champion (Electrophoresis, 2002) conducted a dilution series experiment to validate PDQuest and Progenesis PG240 software
- 4 replicate gels for each of 7 protein loads 0.5μg, 7.5μg, 10μg, 15μg, 30μg, 40μg, 50 μg
- We evaluated all spots in all gels
  - Initial study evaluated 20 selected spots
- Compared Pinnacle to PDQuest, PG240, and recently SameSpots

# Parameters Evaluated - 1

- Number of Spots Detected
  - Pinnacle method – all identifed pinnacles
  - PDQ and Progenesis - unmatched spots and spots not present in at least 3 out of 4 replicates in one treatment group excluded
  - Aligned group – to determine the effect of alignment alone
  - Determined # of spots present in all gels
- Match Percentage
  - Random sampling of 10% of all spots

# Parameters Evaluated – 2

- Reliability assessed by computing $R^2$ from regression of spot quantification on protein load
  - Linearity of quantification over different protein loads
- Precision assessed by computing CV for 30 µg load

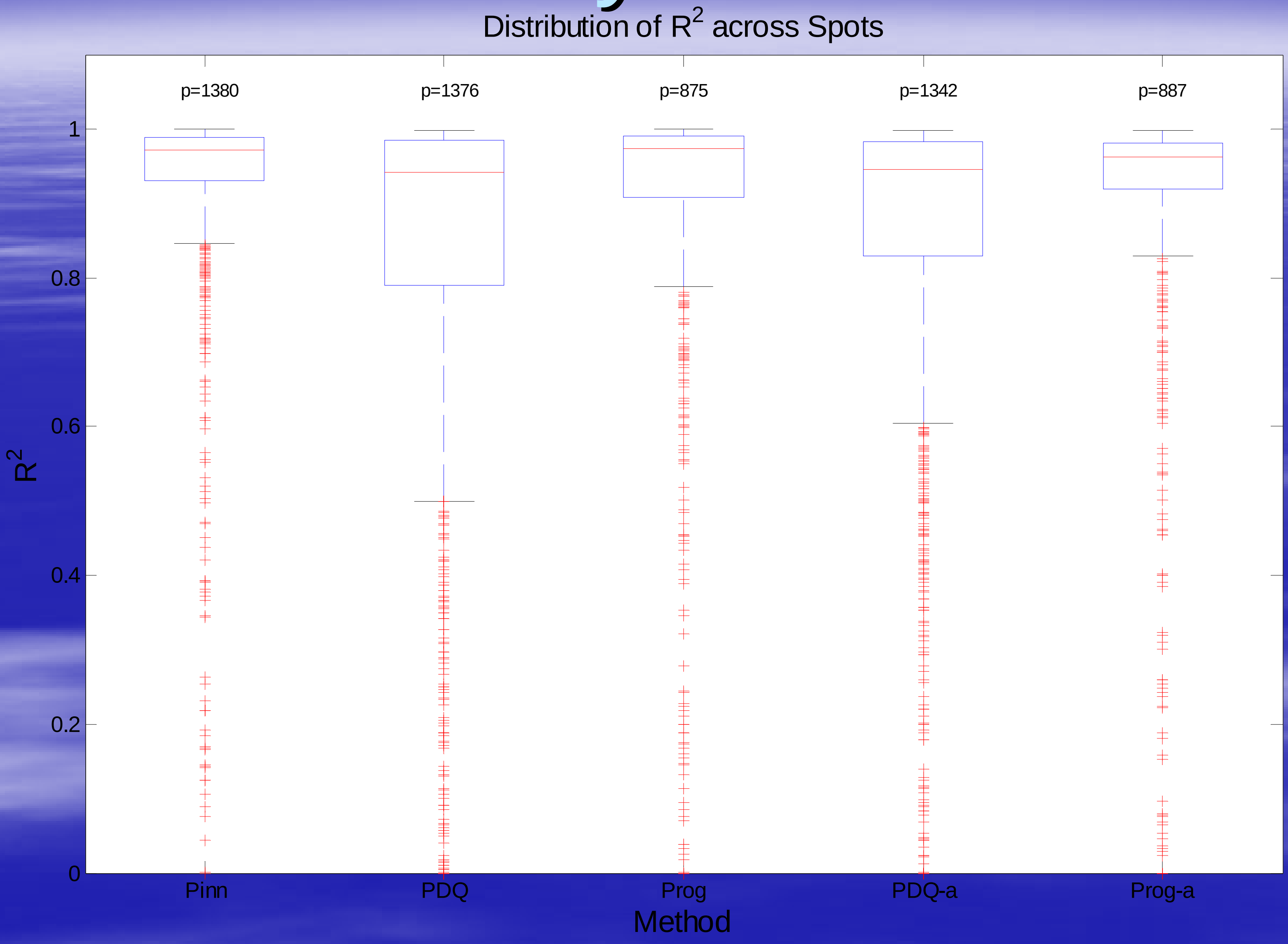
# Spot Detection and Matching Nishihara and Champion

| Analysis Method | # Spots Detected | # Spots Selected | # Spots All Gels | Match % |
|---|---|---|---|---|
| Pinnacle | 1380 | 1380 | 1380 | 100 |
| PDQuest | 2692 | 1376 | 377 | 60 |
| PG240 | 1986 | 875 | 271 | 84 |
| PDQ-a | 2636 | 1342 | 385 | 71 |
| PG240-a | 2006 | 887 | 312 | 80 |
| SameSpots | 688 | 688 | 688 | 100 |

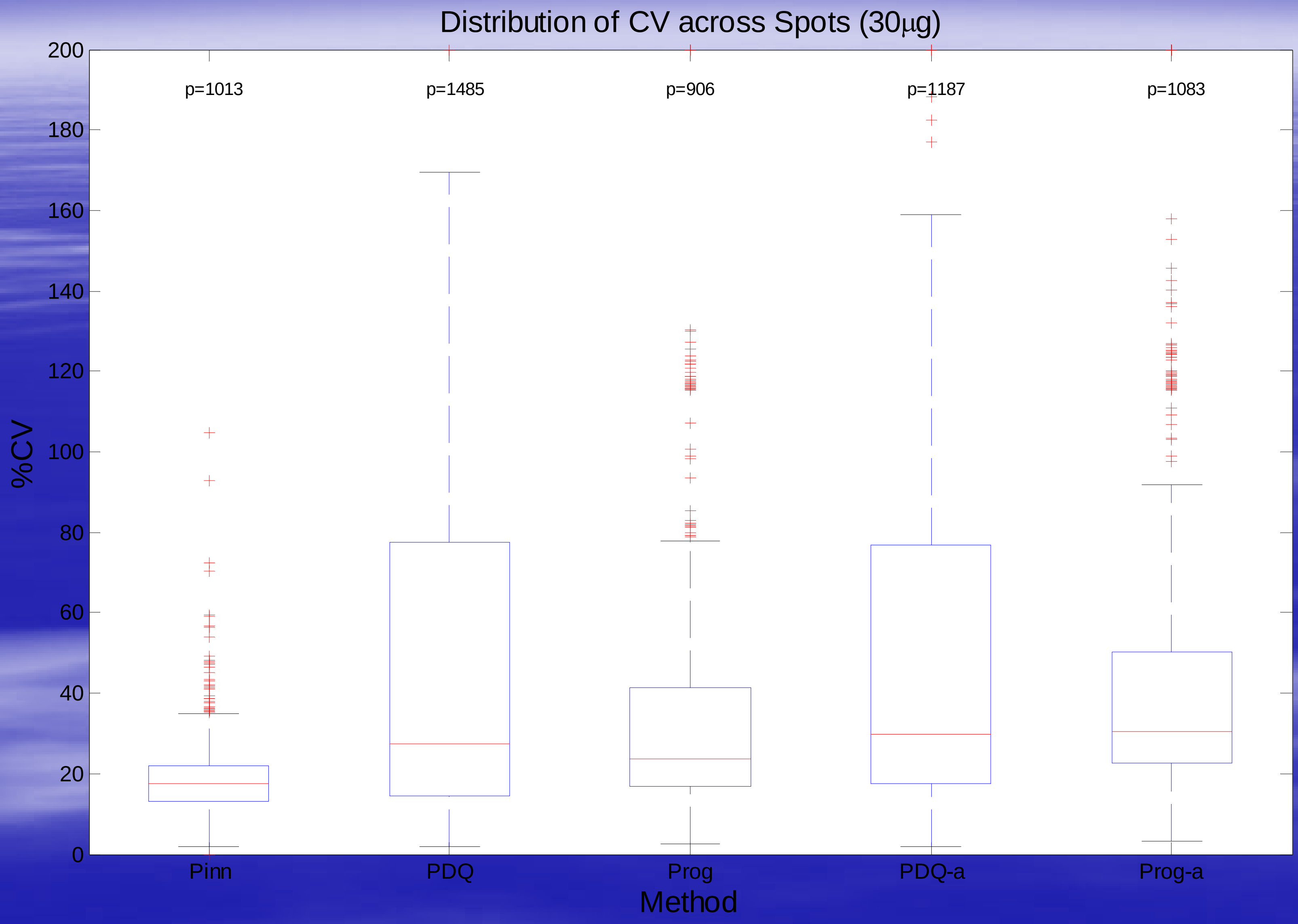
# Reliability and Precision
# Nishihara and Champion

| Analysis Method | # Spots Selected | # Spots $R^2 >$ 0.90 | Mean $R^2$ | # Spots %CV < 20 | Mean %CV |
|---|---|---|---|---|---|
| Pinnacle | 1403 | 1203 | 0.924 | 983 | 20.0 |
| PDQuest | 1376 | 847 | 0.835 | 498 | 54.7 |
| PG240 | 875 | 666 | 0.883 | 304 | 40.3 |
| PDQ-a | 1342 | 869 | 0.850 | 415 | 55.7 |
| PG240-a | 887 | 713 | 0.894 | 144 | 47.4 |
| SameSpots | 688 | 646 | 0.956 | 464 | 20.2 |

# Reliability – N and C



# Precision - N and C



# Validation: Homegrown Dilution Series

- Extract of SH-SY5Y neuroblastoma cell line
- 3 replicate gels for each of 6 protein loads
  - 5 µg, 10 µg, 25 µg, 50 µg, 100 µg, 200 µg
- Evaluated all spots in all gels

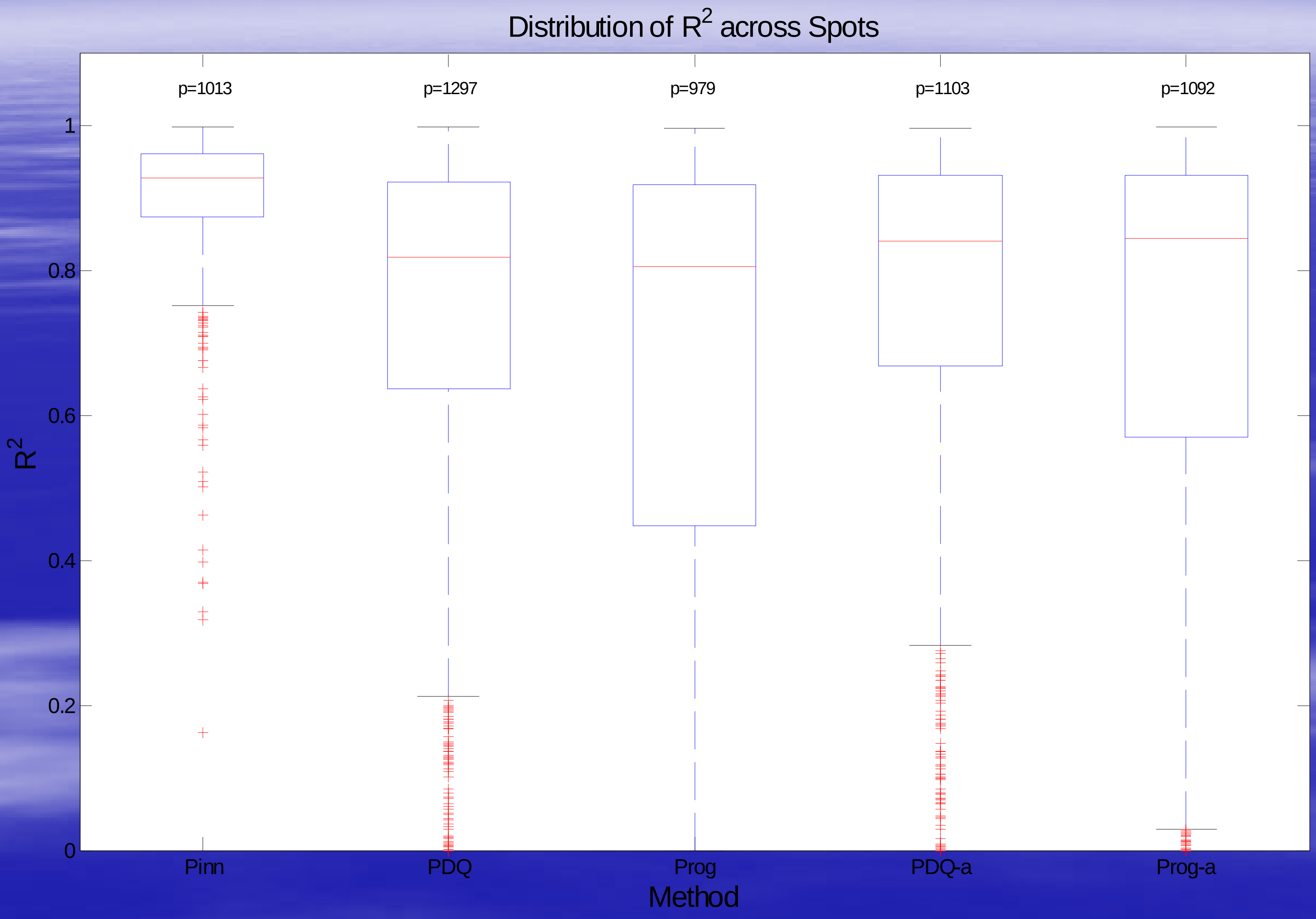
# Spot Detection and Matching Homegrown Dilution Series

| Analysis Method | # Spots Detected | # Spots Selected | # Spots All Gels | Match % |
|---|---|---|---|---|
| Pinnacle | 1013 | 1013 | 1013 | 100 |
| PDQuest | 2666 | 1297 | 40 | 45 |
| PG240 | 1891 | 979 | 51 | 30 |
| PDQ-a | 2243 | 1103 | 80 | 64 |
| PG240-a | 1730 | 1092 | 143 | 43 |
| SameSpots | 1037 | 1037 | 1037 | 100 |

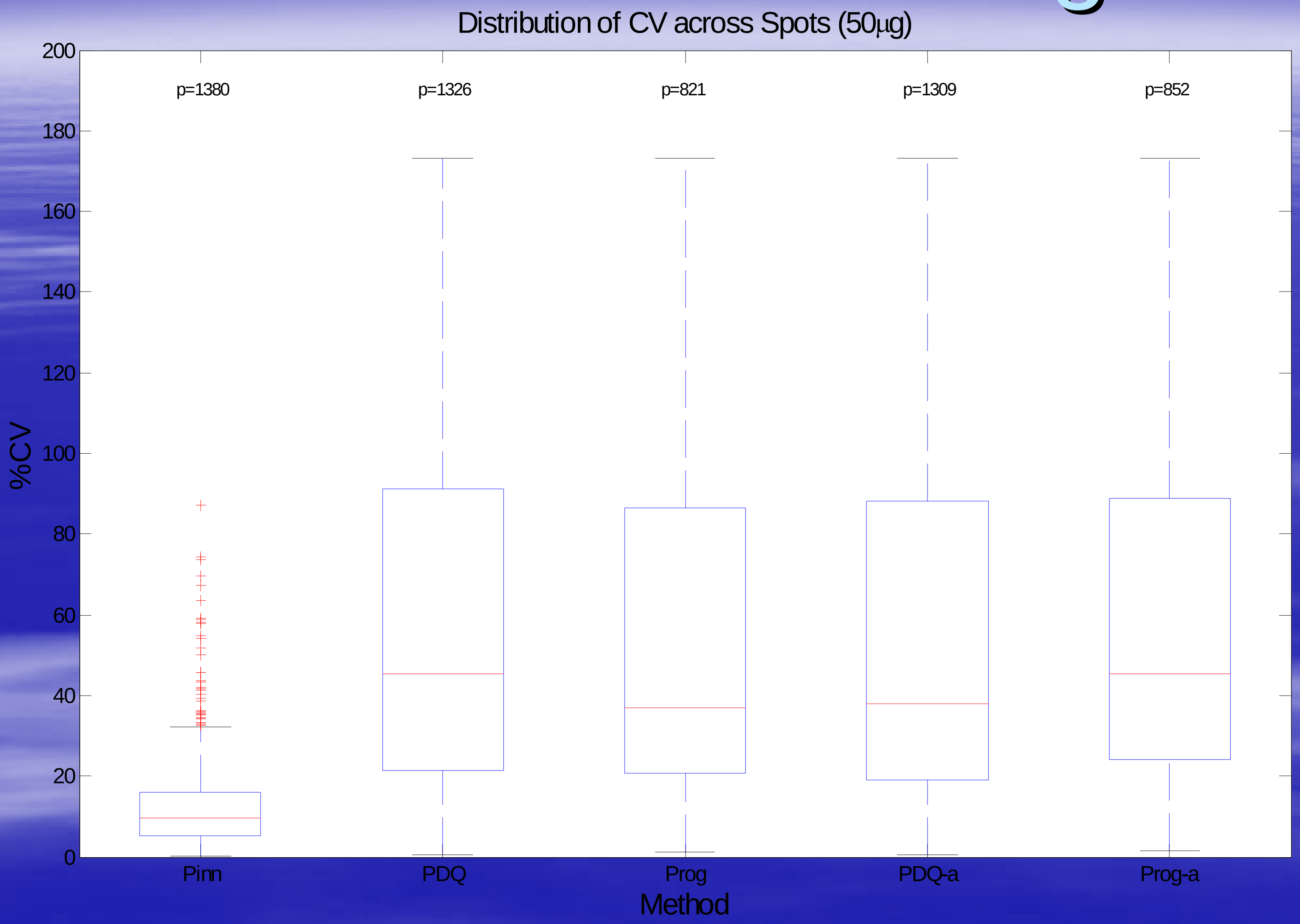
# Reliability and Precision Homegrown Dilution Series

| Analysis Method | # Spots Selected | # Spots $R^2 >$ 0.90 | Mean $R^2$ | # Spots %CV < 20 | Mean %CV |
|---|---|---|---|---|---|
| Pinnacle | 1013 | 663 | 0.889 | 859 | 26.6 |
| PDQuest | 1297 | 406 | 0.735 | 267 | 64.4 |
| PG240 | 979 | 295 | 0.662 | 188 | 53.2 |
| PDQ-a | 1103 | 391 | 0.753 | 272 | 58.8 |
| PG240-a | 1092 | 384 | 0.698 | 182 | 59.9 |
| SameSpots | 1037 | 501 | 0.804 | 400 | 29.9 |

# Reliability – Homegrown



# Precision - Homegrown



# Effect of Increasing Gel Numbers on Match Percentage

| Experiment | PDQuest | Progenesis |
|---|---|---|
| 3 gels | 88% | 96% |
| 9 gels | 70% | 66% |
| 27 gels | 42% | 21% |
| | | |
| 3 aligned gels | 82% | 89% |
| 9 aligned gels | 75% | 71% |
| 27 aligned gels | 52% | 35% |

# Advantages of Our Approach

- Automatic – After alignment, fully automated
- Rapid – <1 minute for 60 gels
- Sensitive and Specific – use of average gel borrows strength across gels, allowing one to find fainter spots, thus increasing realized dynamic range of gel
- Robust – use of average gel minimizes artifacts
- No Missing Spots – quantifications for each spot on every gel
- Reliable and Precise – use of the average gel and pinnacles results in more reliable and precise quantifications than standard approaches
- No Spot Mismatching – significant issue with other automatic methods

# Proteomics: Feature Extraction Approach Identifying Significant Features

- Class comparison
  - Perform any statistical test on columns of Y – obtain test statistics or p-values(like microarrays)
  - To control for multiple testing, use FDR (false discovery rate) based method to find appropriate threshold for determining significance.
    - Global FDR control: control expected proportion of false discoveries
    - Local FDR estimation: For each feature, estimate probability of being false discovery if called significant
- Class prediction can also be done, but IMHO proteomics assays not yet ready to be used in clinical applications.
  - Be sure to properly validate your classifier (with external, not internal CV) for accurate estimates of prediction error

# Proteomics: Feature Extraction Approach

- Advantages of feature extraction approach:
  - Meaningful dimension reduction: reduces high dimensional functions/images to simple matrix.
  - Computationally efficient: computing time and memory.
  - Flexibility: can apply any statistical method to N x p matrix
  - If effective, should capture biologically meaningful information in the data.
- Disadvantages:
  - Potential discoveries missed from features not detected.
  - Difficult to model systematic functional effects of nuisance factors.
- Alternative approach: model entire spectrum/image as function

# Proteomics: Functional Modeling Approach

- Preprocess spectra/images
- Apply functional model to spectra/images
  - Model must be flexible enough to capture complex features in data
  - Must be computationally efficient enough to handle very large functions/images
  - Wavelet-based functional mixed model (yesterday's talk) seems to work well.
- Perform model-based inference to identify significant features or perform classification.

# Functional Mixed Model

$\boldsymbol{Y(t)}$ = set of $N$ spectra, stacked as rows.

$$U_i(t) \sim GP(0,Q)$$

$$E_i(t) \sim GP(0,S)$$

$$\underbrace{Y(t)}_{\substack{\text{N}\\\text{functions}}} = \overbrace{X}^{N \times p}\ \underbrace{B(t)}_{\substack{\text{p}\\\text{functions}}} + \overbrace{Z}^{N \times m}\ \underbrace{U(t)}_{\substack{\text{m}\\\text{functions}}} + \underbrace{E(t)}_{\substack{\text{N}\\\text{functions}}}$$

- *Q* and *S* are covariance surfaces describing the how the random effect curves/residual error processes vary across replicates.
- For image data, Y, X, U, and E are functions of both pH ($t_1$) and molecular mass ($t_2$)
- Model fit using Bayesian, wavelet-based method
- Yields posterior samples for all functional parameters

# Model: MALDI Example

Let $Y_i(t)$ be the MALDI spectrum *i*

$$\log_2\{Y_i(t)\} = B_0(t) + \sum_{j=1}^{4} X_{ij} B_j(t) + \sum_{k=1}^{16} Z_{ik} U_k(t) + E_i(t)$$

- **$X_{i1}$=*1* for lung, *-1* brain. $X_{i2}$=*1* for A375P, *-1* for PC3MM2**

  **$X_{i3} = X_1 * X_2$ $X_{i4}$=*1* for low laser intensity, *-1* high.**

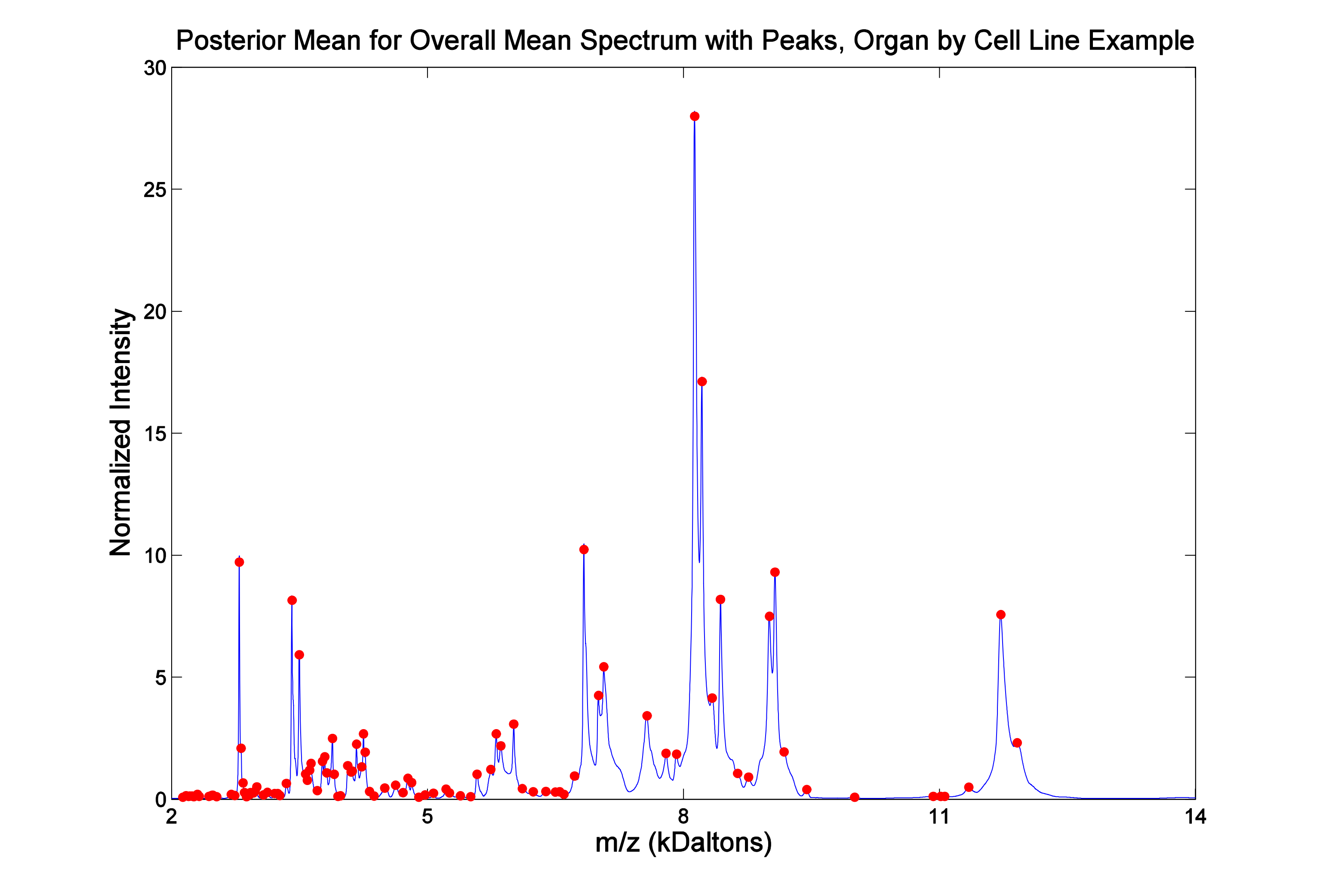
- $B_0(t)$ = **overall mean** spectrum $B_1(t)$ = **organ main effect** function

  $B_2(t)$ = **cell-line main effect** $B_3(t)$ = **org x cell-line int** function

  $B_4(t)$ = **laser intensity effect** function

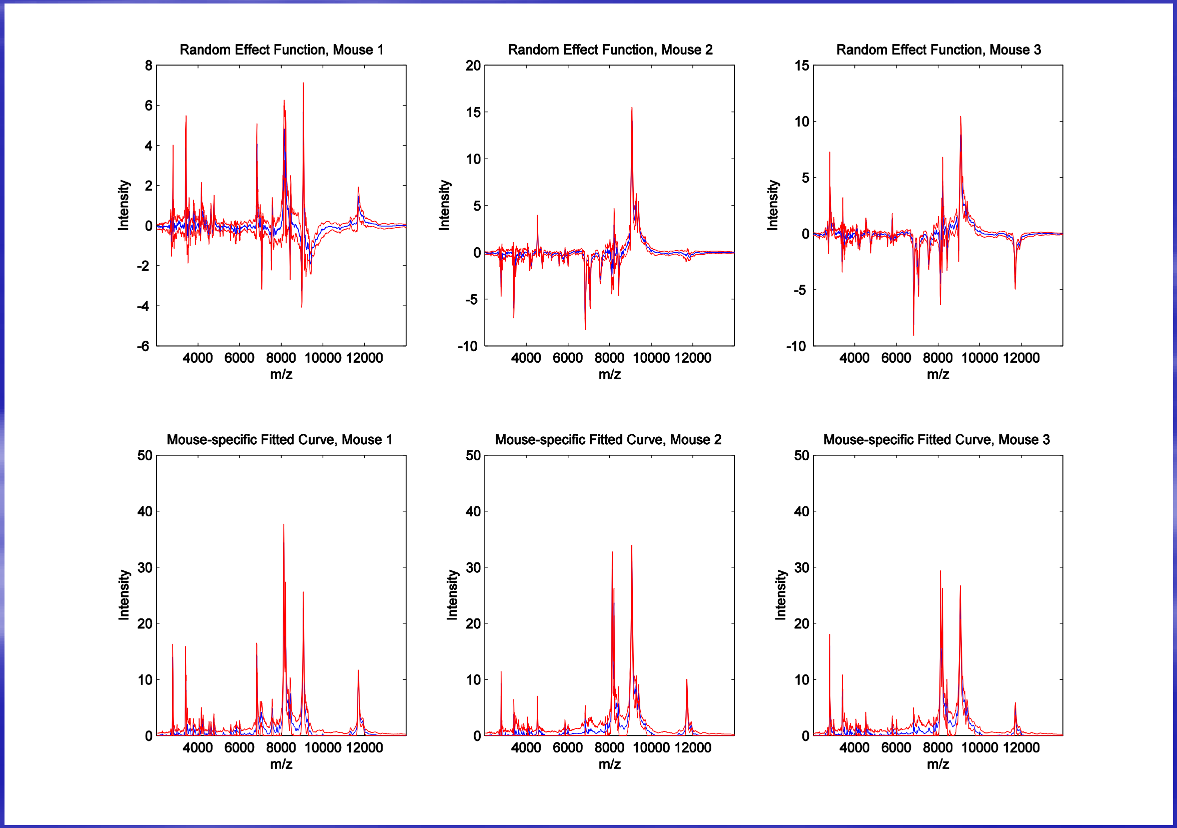
- $Z_{ik}$=*1* if spectrum *i* is from mouse *k* (*k=1, …, 16*)
- $U_k(t)$ is random effect function for mouse *k*.

# Adaptive Regularization



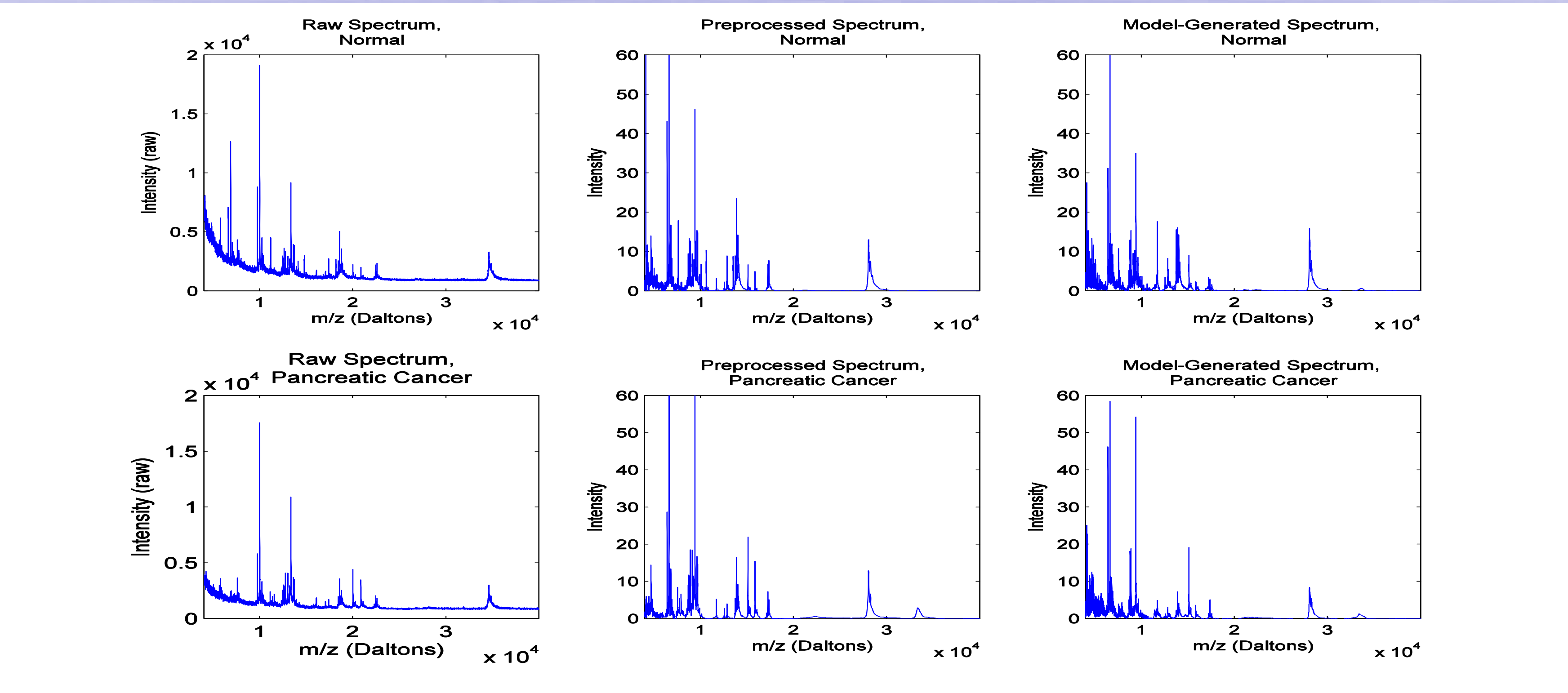
# Adaptive Regularization

- **Posterior samples/estimates of random effect functions *$U_i(t)$* are also *adaptively regularized* from Gaussian prior, since each wavelet coefficient has its own random effect & residual variance**



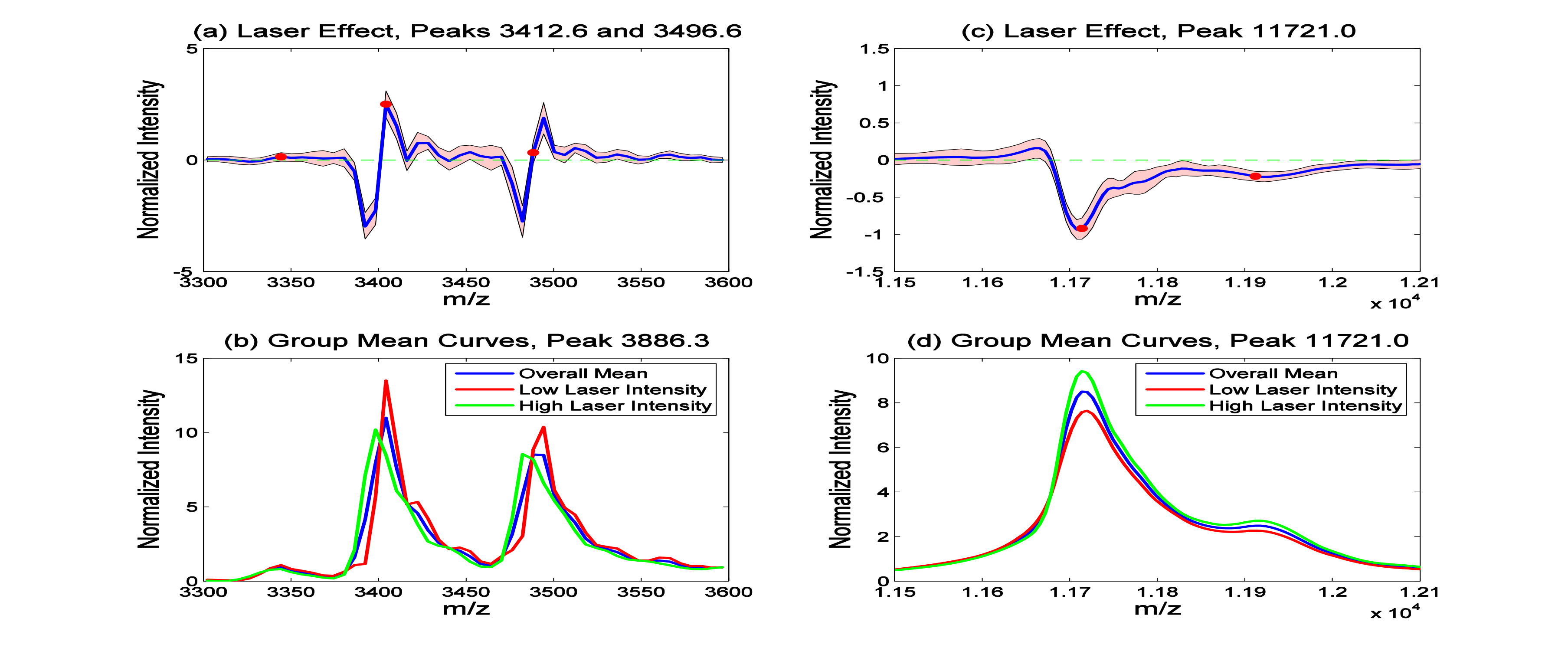
- Able to preserve spikes in random effect functions, as well
- Important for estimation of random effect functions AND for valid inference on fixed effect functions.

# Results: MALDI Example



- Draws of spectra from posterior predictive distribution yield data that looks like real MALDI data (3rd column), indicating reasonable model fit.

# Modeling Block Effects



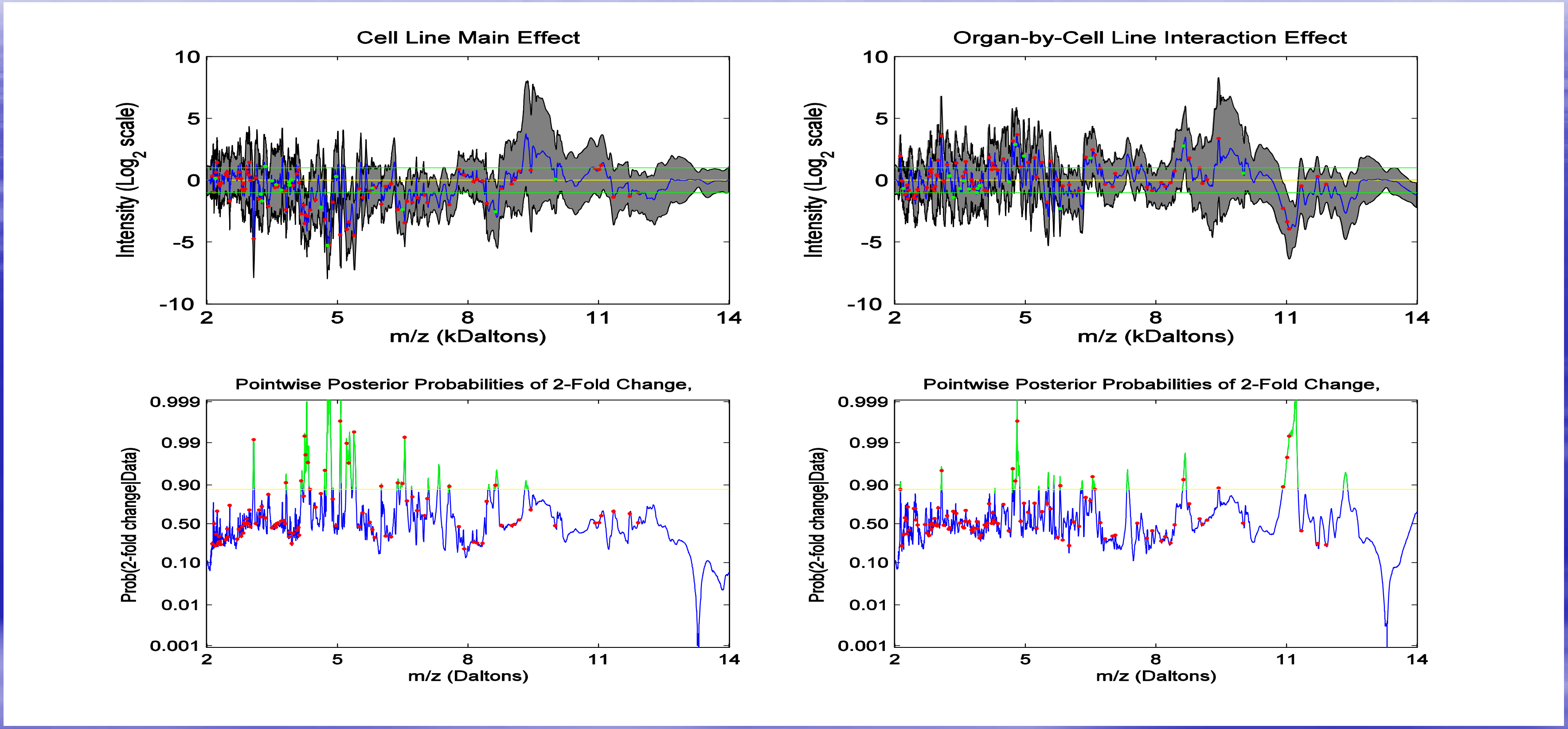
- **Inclusion of nonparametric functional laser intensity effect is able to adjust for systematic differences in the *x* and *y* axes between laser intensity scans**

# Bayesian Inference:

## Class Comparison

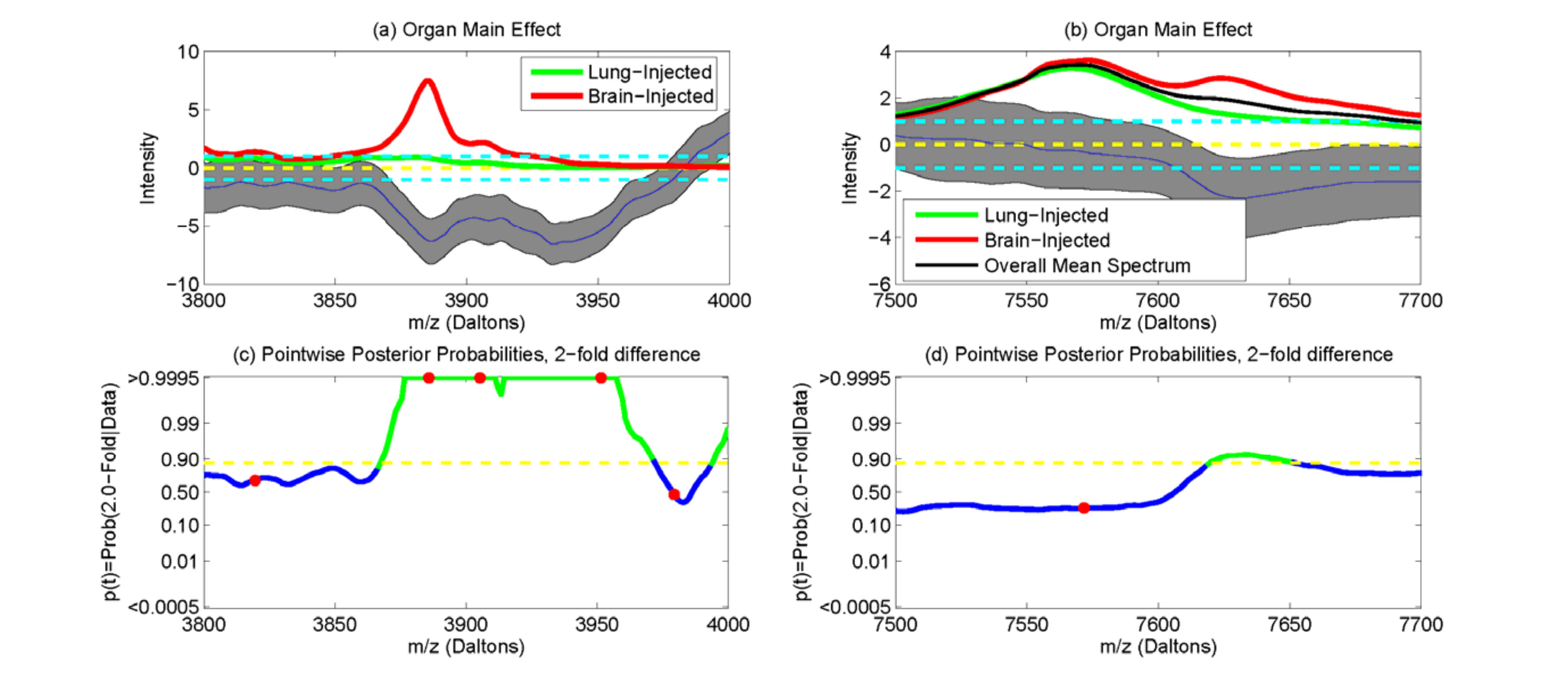
- **WFMM outputs posterior samples of fixed effect functions $B_i(t)$, which measures the effect of factor *i* on each location t of the spectra.**
- **Flag regions of t with $|B_i(t)|$ large as potential biomarkers**
- **Given desired effect size $\geq \delta$, compute pointwise posterior probabilities of effect size for factor *i* being at least $\delta$ : $p_i(t)=Pr\{|B_i(t)|>\delta|Y\}$**
  - **These quantities are Bayesian local FDR estimates at different regions of curve (false discovery rate computed across regions of curves, not genes).**
  - **Can find cut point on local FDR to control Bayesian estimate of global FDR at level α.**

# Results: MALDI Example



- Using $\alpha$=0.05, $\delta$=1 (2-fold expression on $\log_2$ scale), we flag a number of spectral regions.

# Results: MALDI Example



- **3900 D (~100-fold) (CGRP-II): dilates blood vessels in brain**
- **7620 D (~5-fold) (neurogranin): active in synaptic modeling in brain (Not detected as peak)**

# Conclusions

- Proteomic data are complex, requiring multi-step analysis procedure
- Preprocessing important to remove artifacts from data and get data on common scale
- Feature extraction approach quick and easy, but could miss stuff
- Functional modeling does not require feature extraction, but involves complex modeling and is computationally intensive.
- Each method has its merits: simulation studies and thorough comparisons are required to assess the cost-benefit tradeoff between the two methods.

# Acknowledgements

## Statistical Collaborators

*Raymond J. Carroll*

*Jianhua Hu*

*Keith A. Baggerly*

*Kevin R. Coombes*

*Richard C. Herrick* (computing)

## Biomedical Collaborators

*Stanley Hamilton*

*James Abbruzzesse*

*Ryuji Kobayashi*

*John Koomen*

*Nancy Shih*

*Howard Gutstein*

*Brittan Clark*

*Josh Fidler*

*Donghui Li*

A number of papers describing both feature extraction and functional mixed model methods, plus papers giving overviews of proteomics and proteomic data analysis are available on my website (http://biostatistics.mdanderson.org/Morris)

The code for performing MALDI peak detection (PrepMS and Cromwell) and for performing MALDI-MS simulations are also available. Spot detection and quantification software for 2d gels (Pinnacle) will be available soon. Software for fitting the WFMM is also available on the web, and will be updated to make it more user friendly in the future.