Contributions of Statistics to Clinical Proteomic Research

Jeffrey S. Morris

The University of Texas
M.D. Anderson Cancer Center

Houston, TX, USA

http://biostatistics.mdanderson.org/Morris

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)

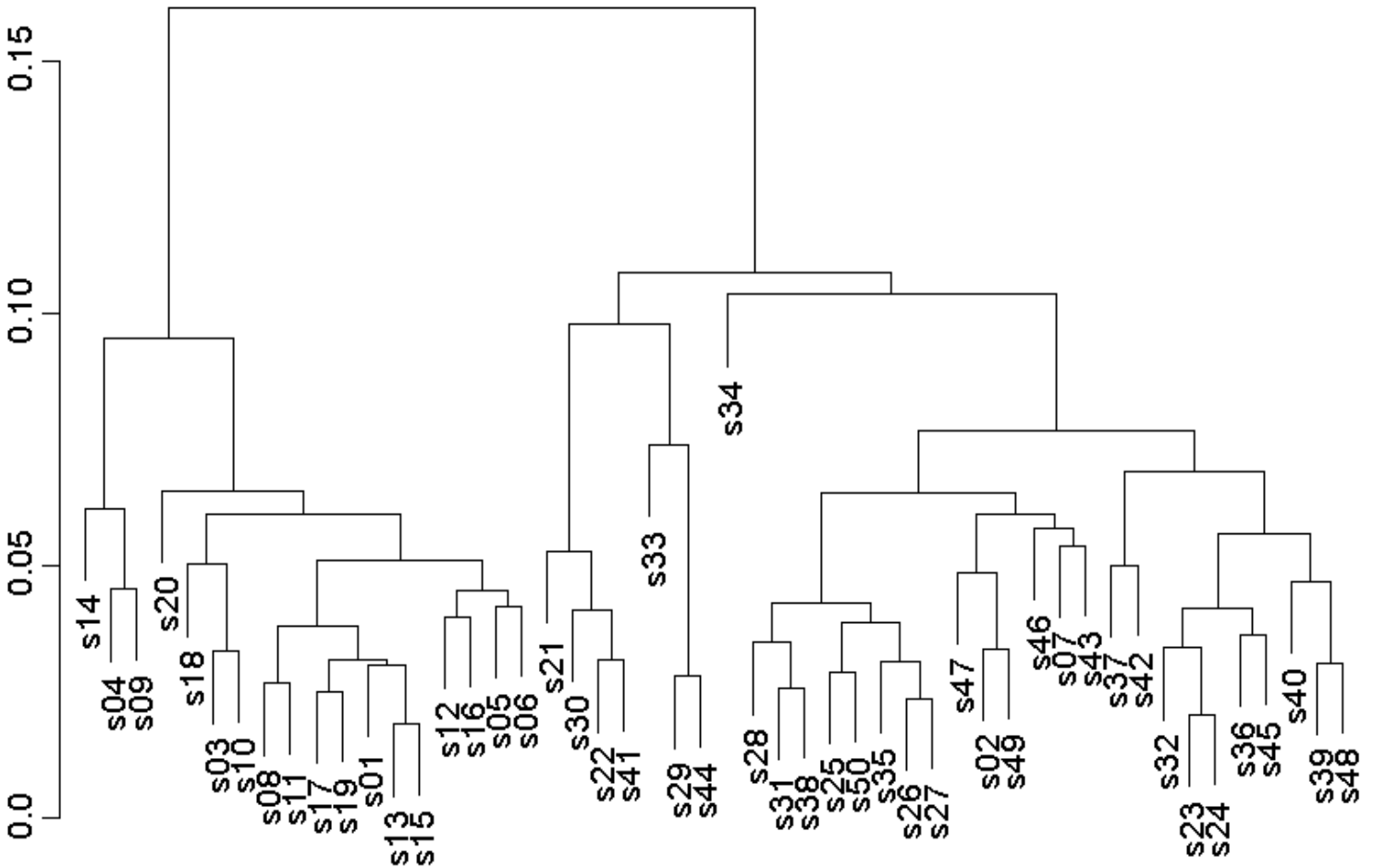
Design makes a difference

- 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

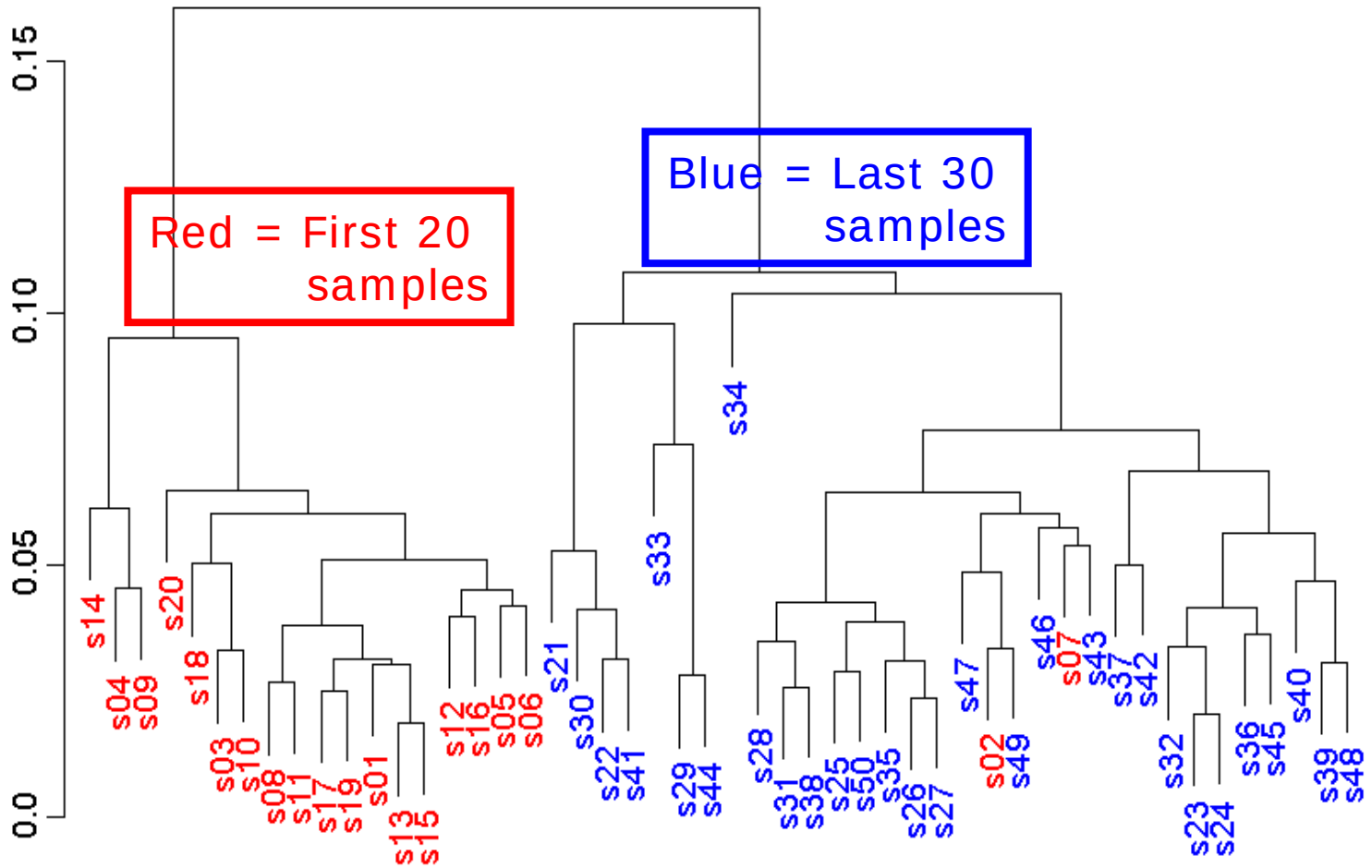
- All samples must be collected uniformly
  - Consistent protocol
  - Enforced at every collection site
- Failure to do this can (will) affect protein profiles
- The problem is particularly serious if sample handling is confounded with interesting variables (normal vs cancer)

Hierarchical clustering of serum protein profiles of brain cancer



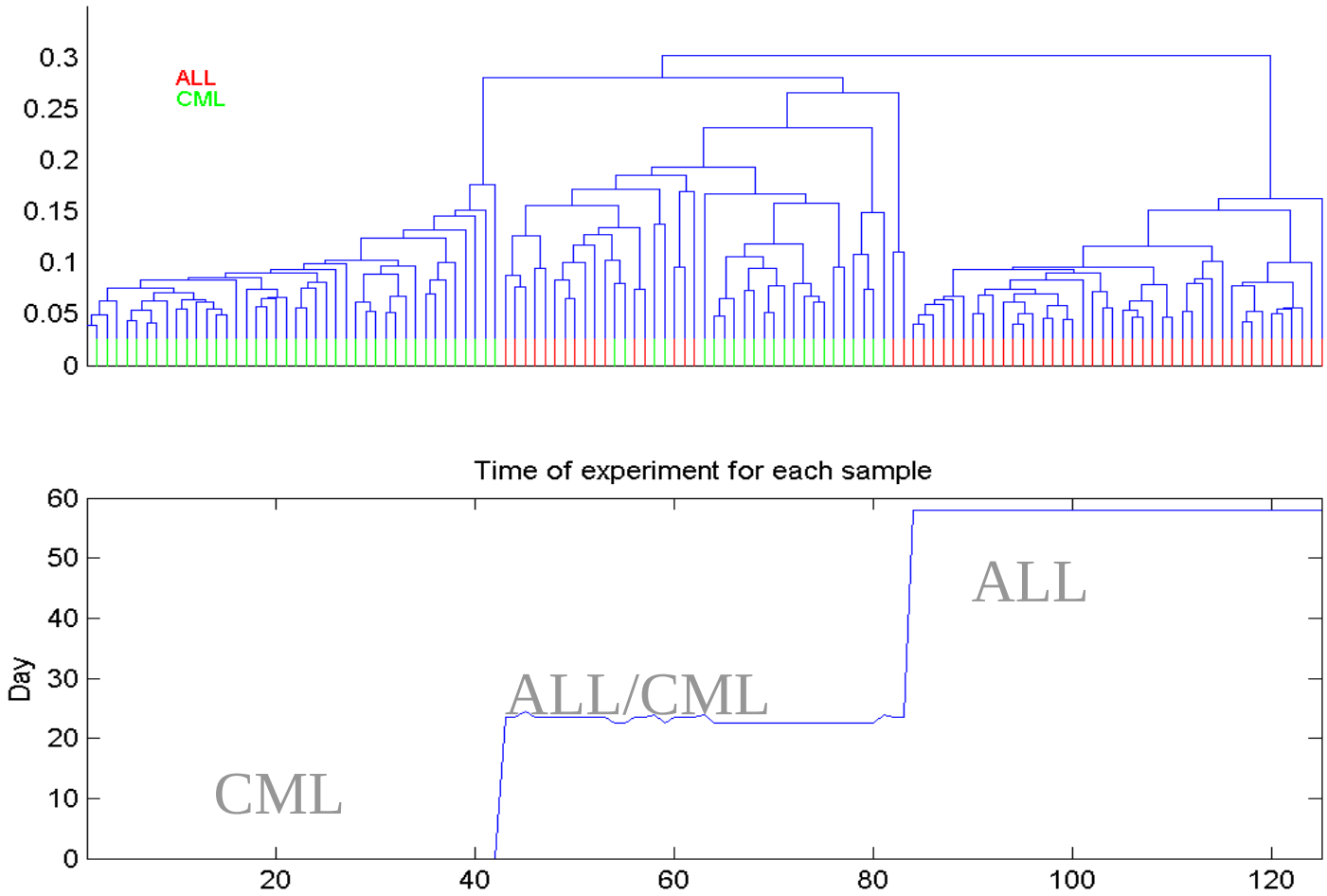
MALDI data from MDACC

Clustering reflects changes in the sample collection protocol



MALDI data from MDACC

Unsupervised methods often cluster samples by run date

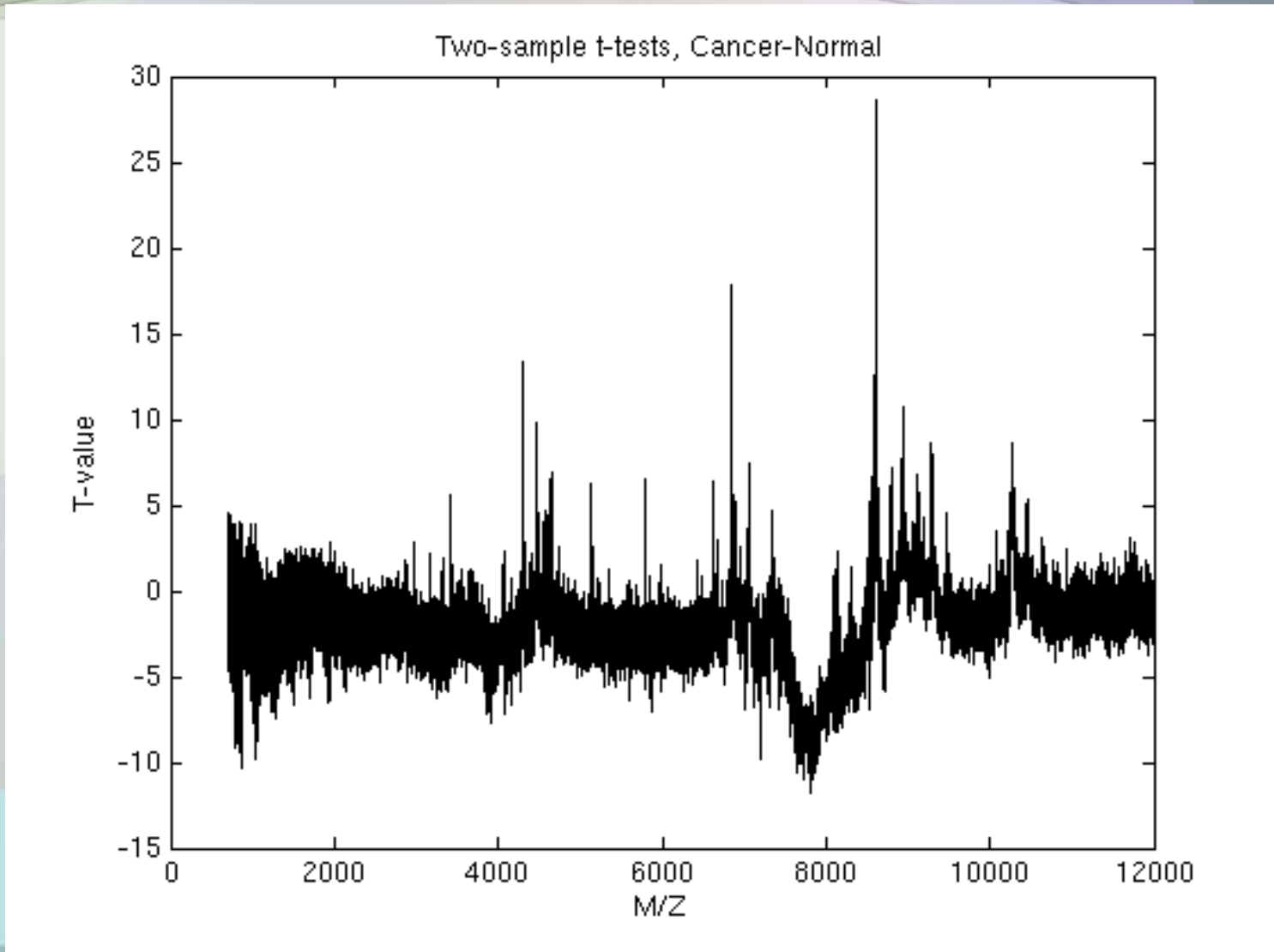


SELDI leukemia data from MDACC

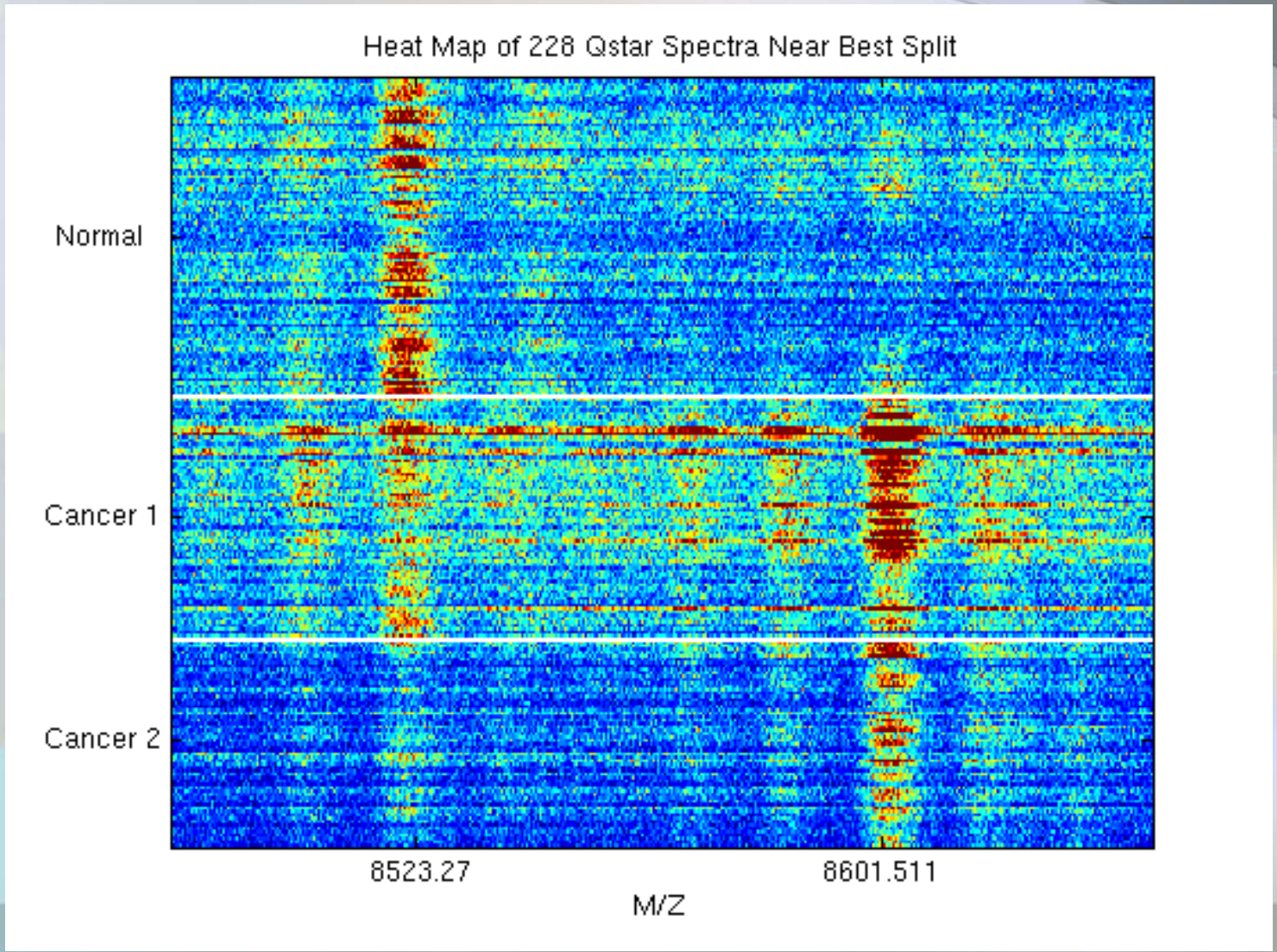
A cautionary tale

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

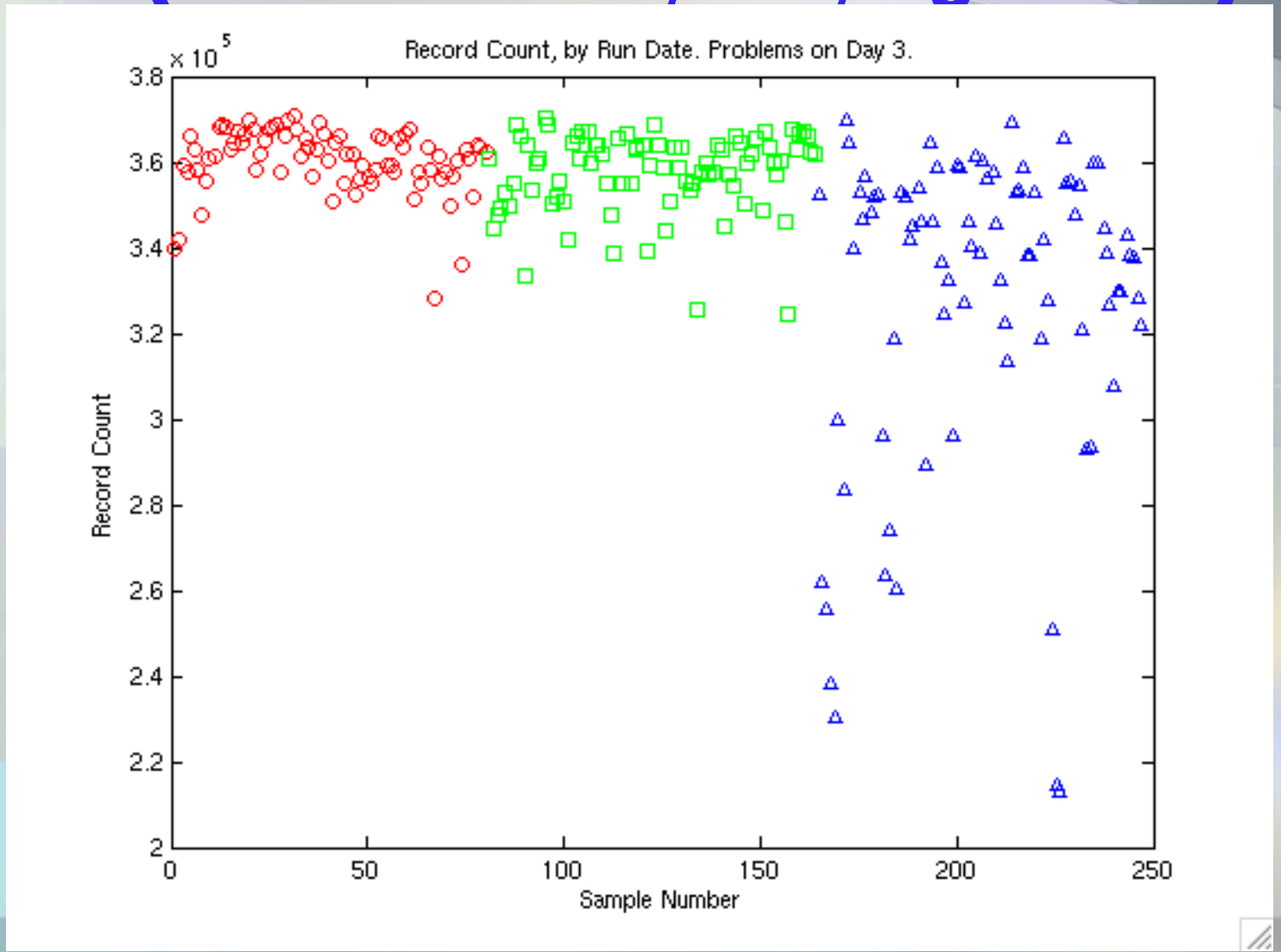
T-statistics identify separator at 8602D



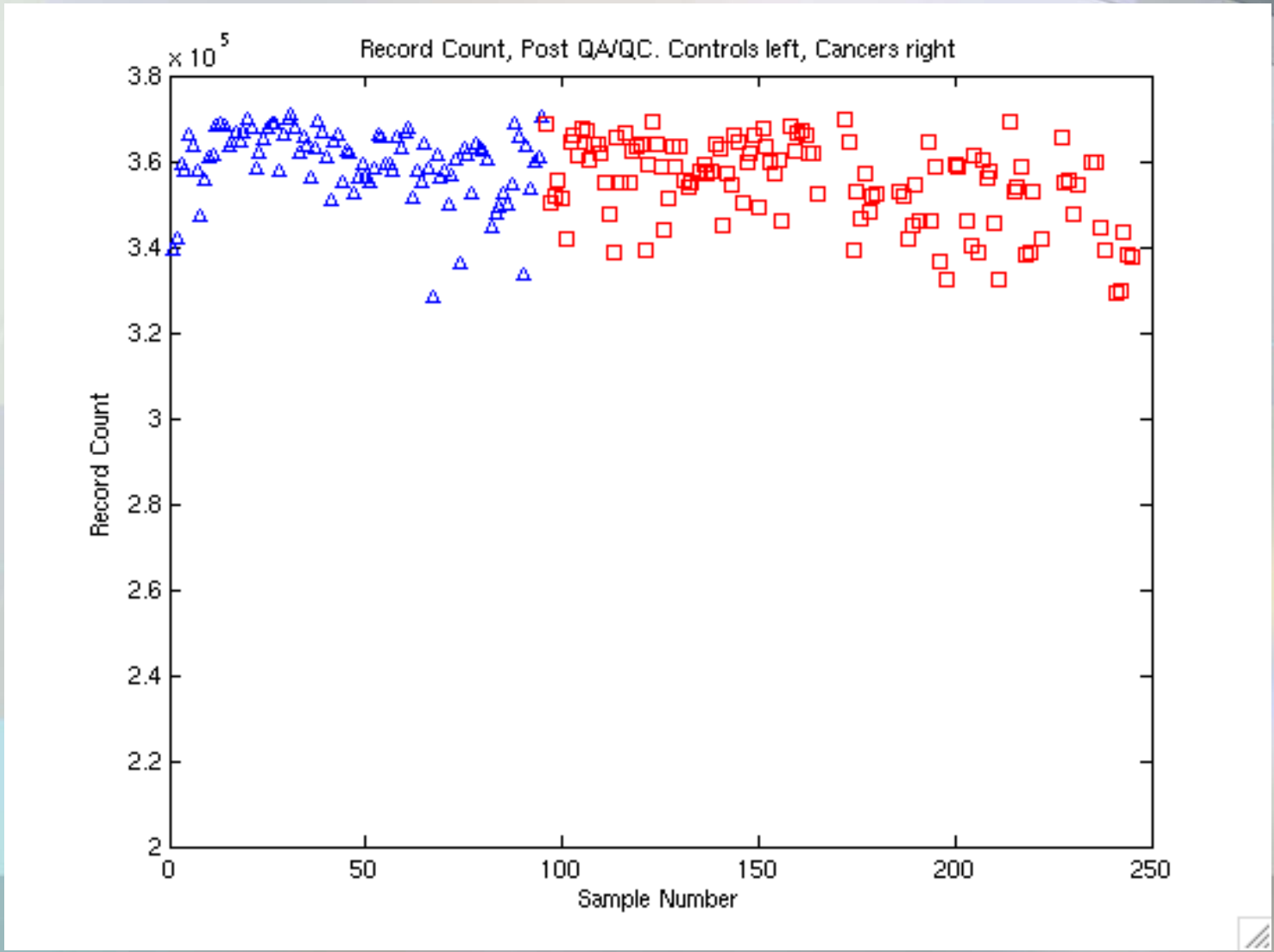
Heat map of raw data near 8602 Da: Why are there two cancer groups?



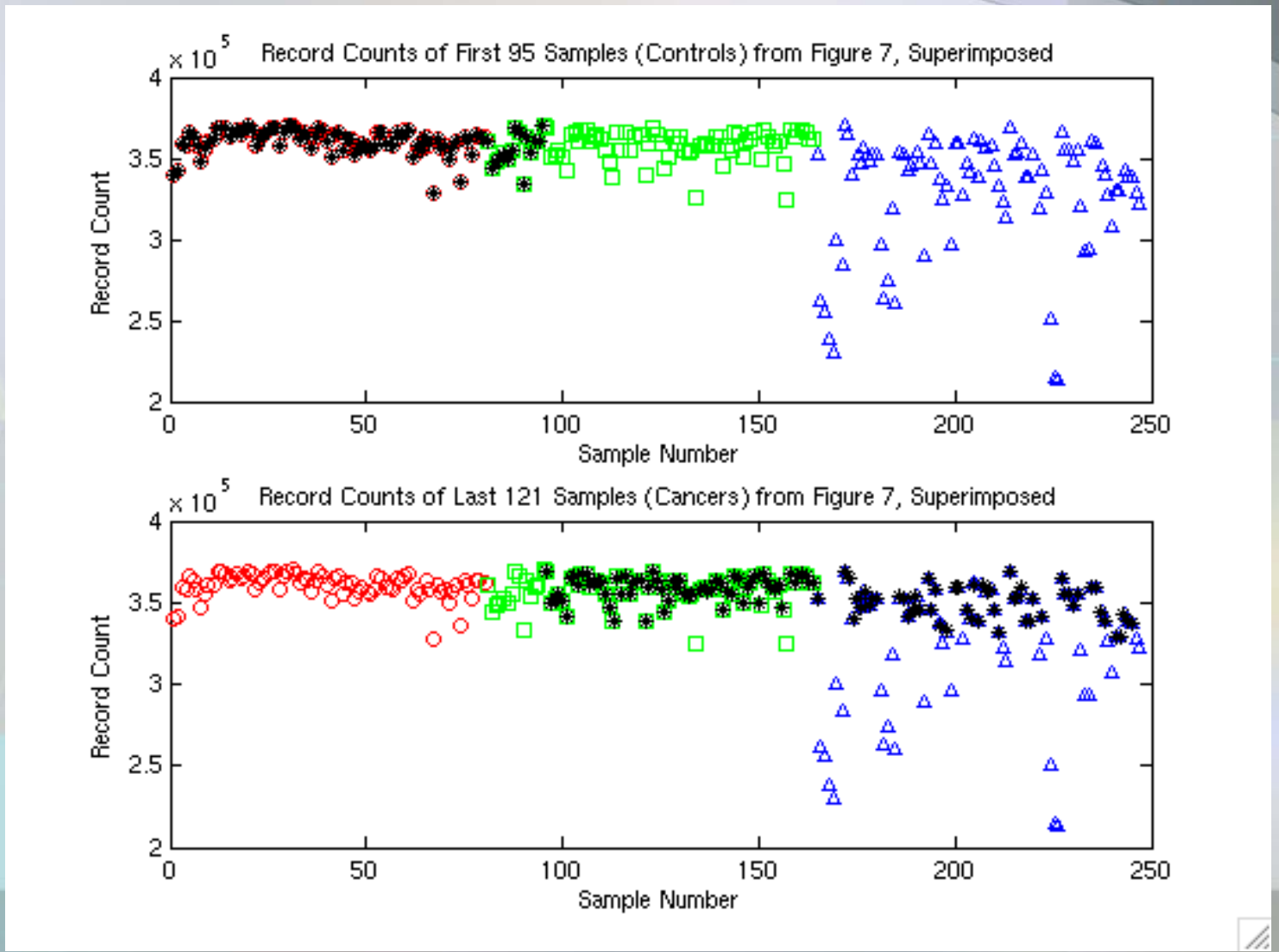
QC: Colors indicate run date (Conrads et al, ERC, Figure 6a)



QC: Colors indicate control/case (Conrads et al, ERC, Figure 7)



All controls were processed before all samples from cancer patients



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?

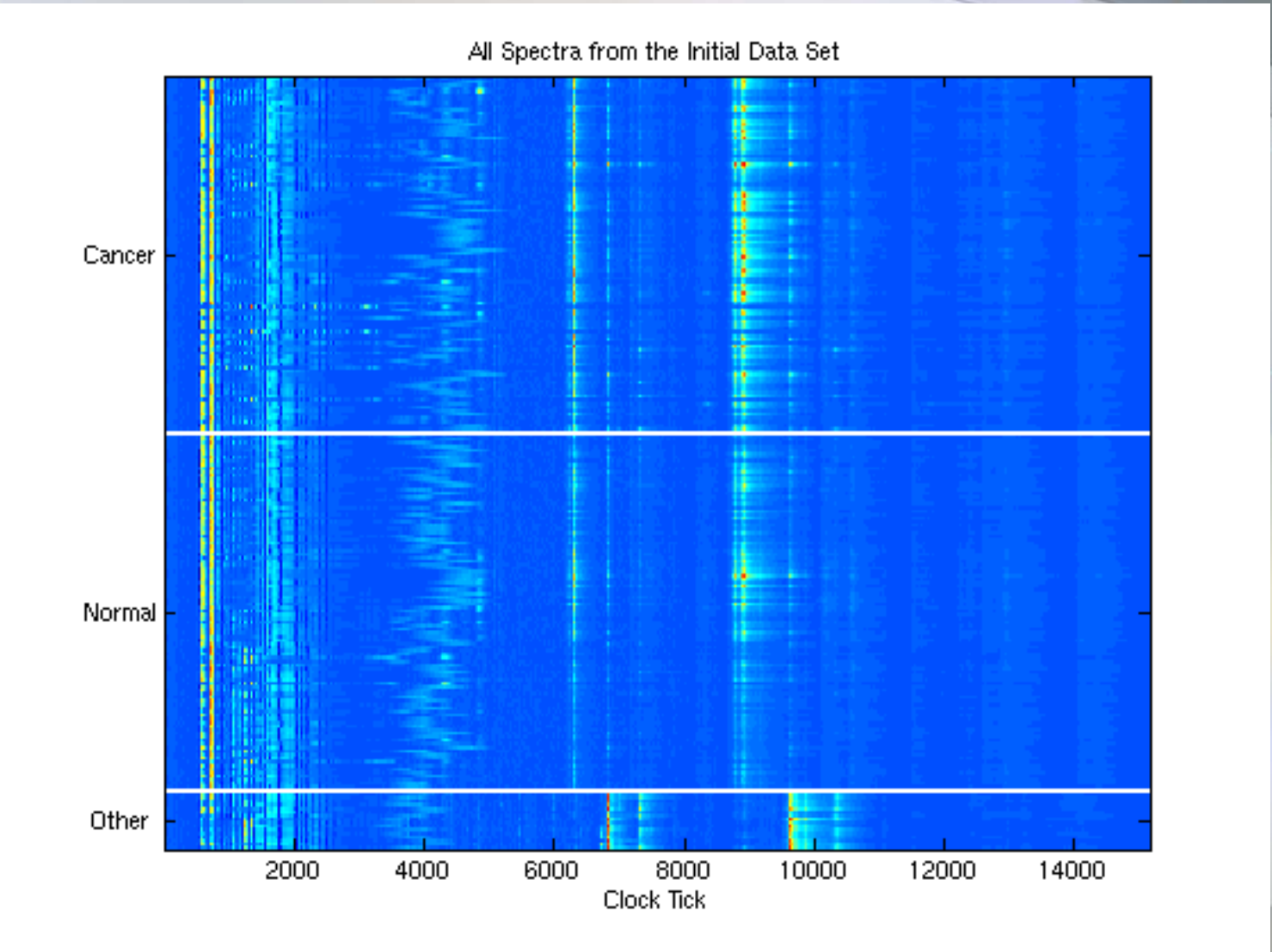
Look at the Data!!
Petricoin et al. (2002 Lancet)

- Collected SELDI 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

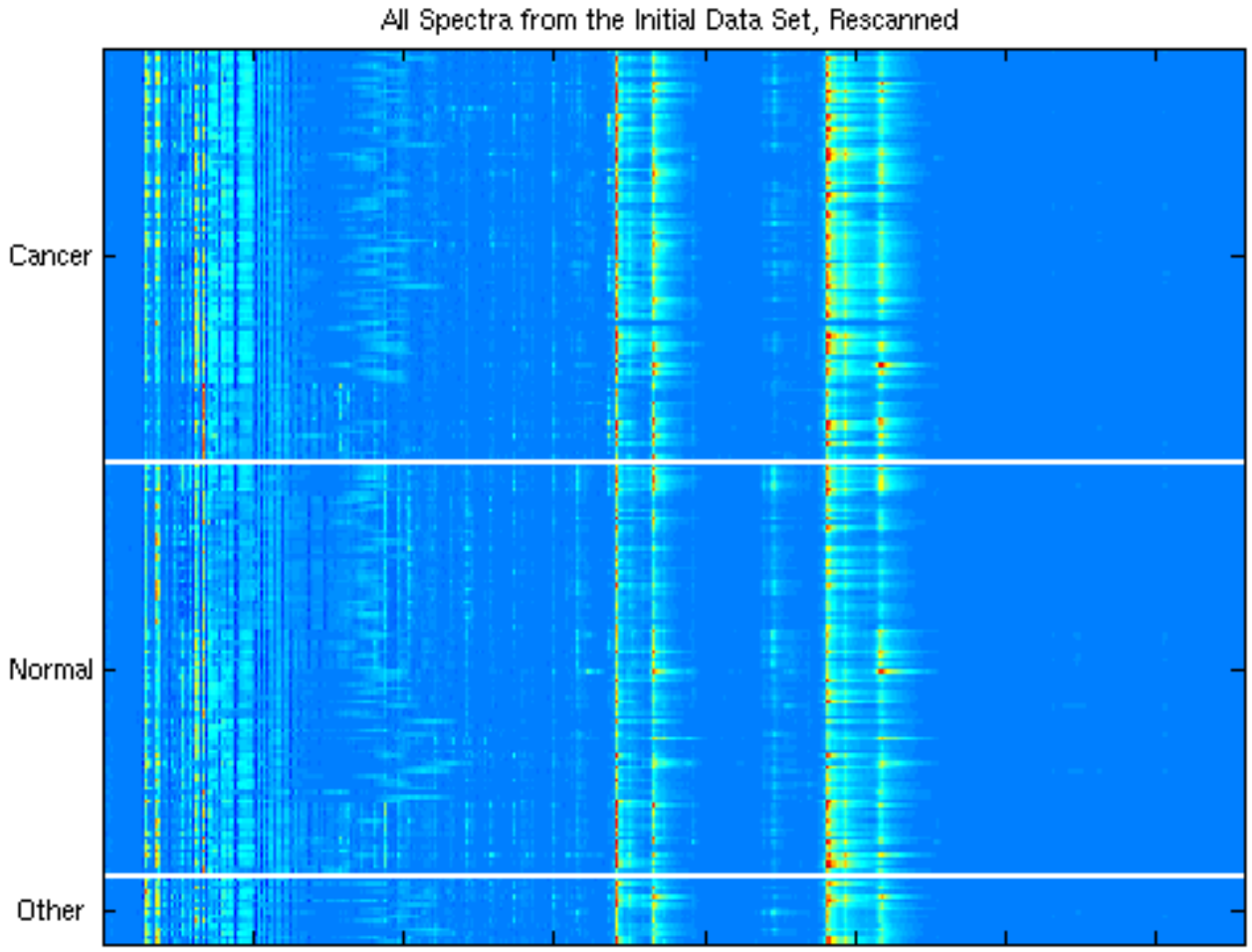
Petricoin Results

- 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”

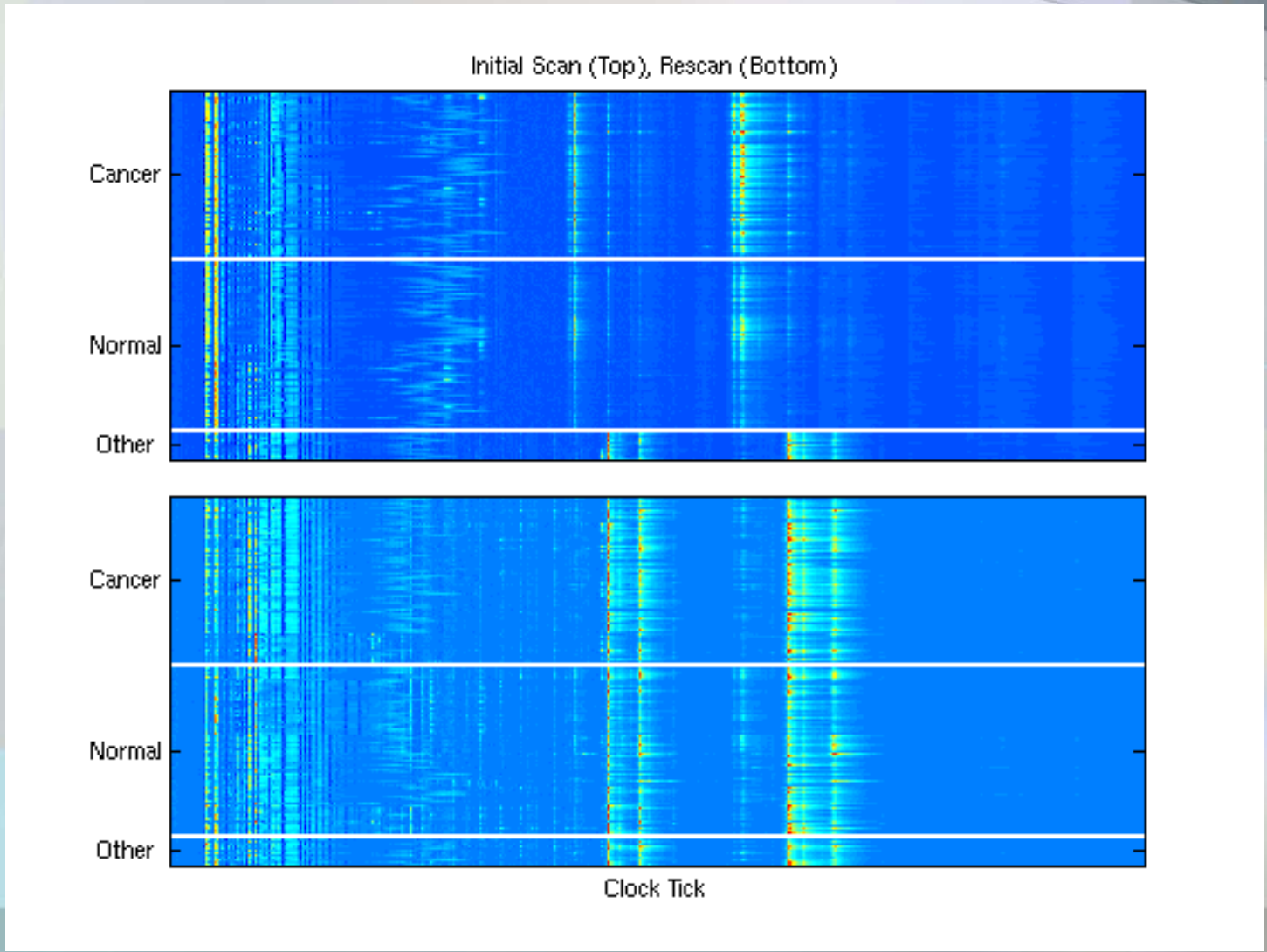
Some structure is visible in Heat Map



Structure disappears in Data Set 2 (same samples, different chip type)



Technology can overwhelm biology



Preprocessing

- We have found that preprocessing can be the most important step in the quantitative analysis process.
- It takes us from the raw data (spectra/gel images) to the meaningful scientific features we want to analyze (quantified peaks/spots)
- Important to get right, since subsequent analyses depend on results

Semi-Statistical Model for Spectrum

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

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

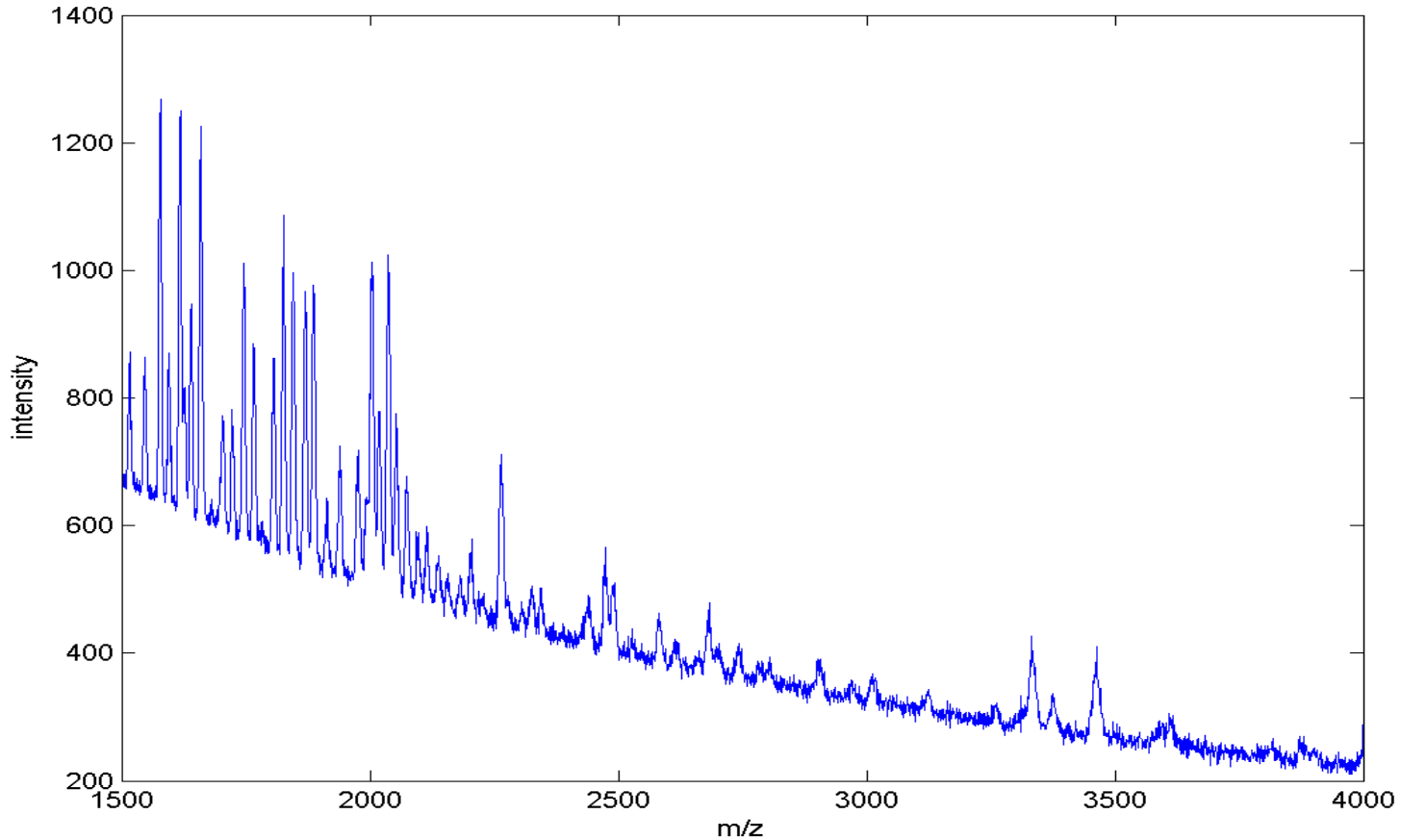
Preprocessing

- 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?**

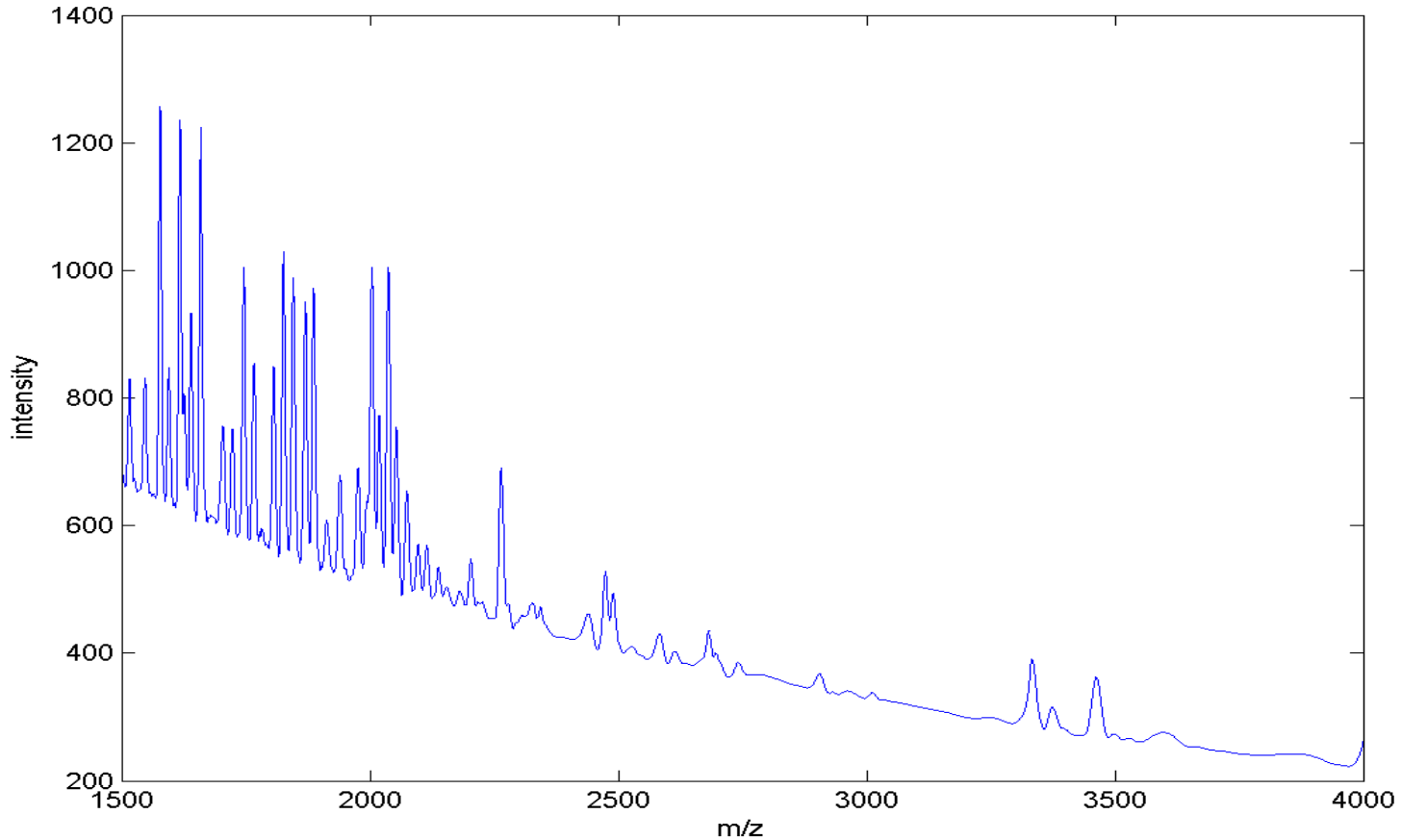
Denoising using Wavelets

- 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.**

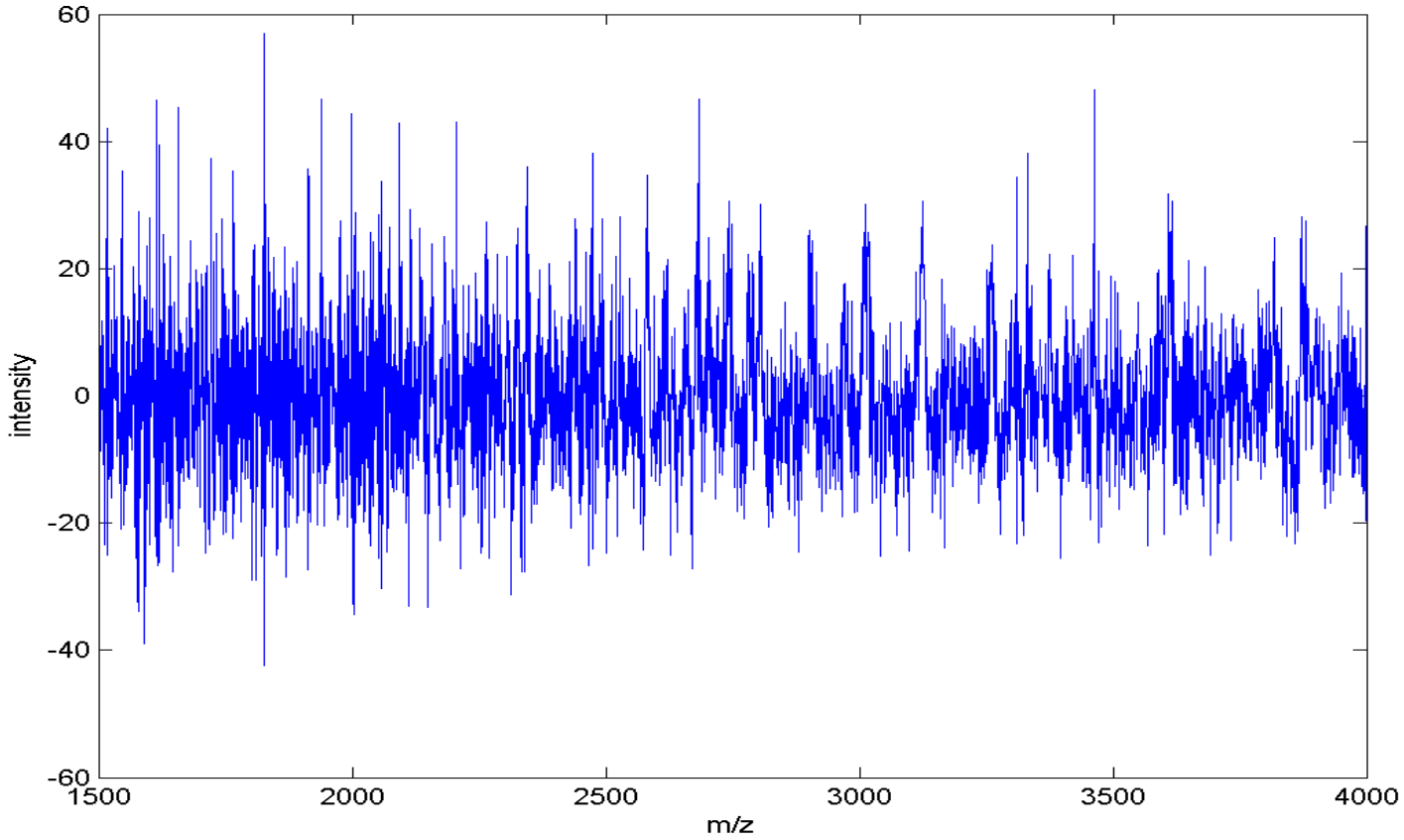
Raw Spectrum



Denoised Spectrum



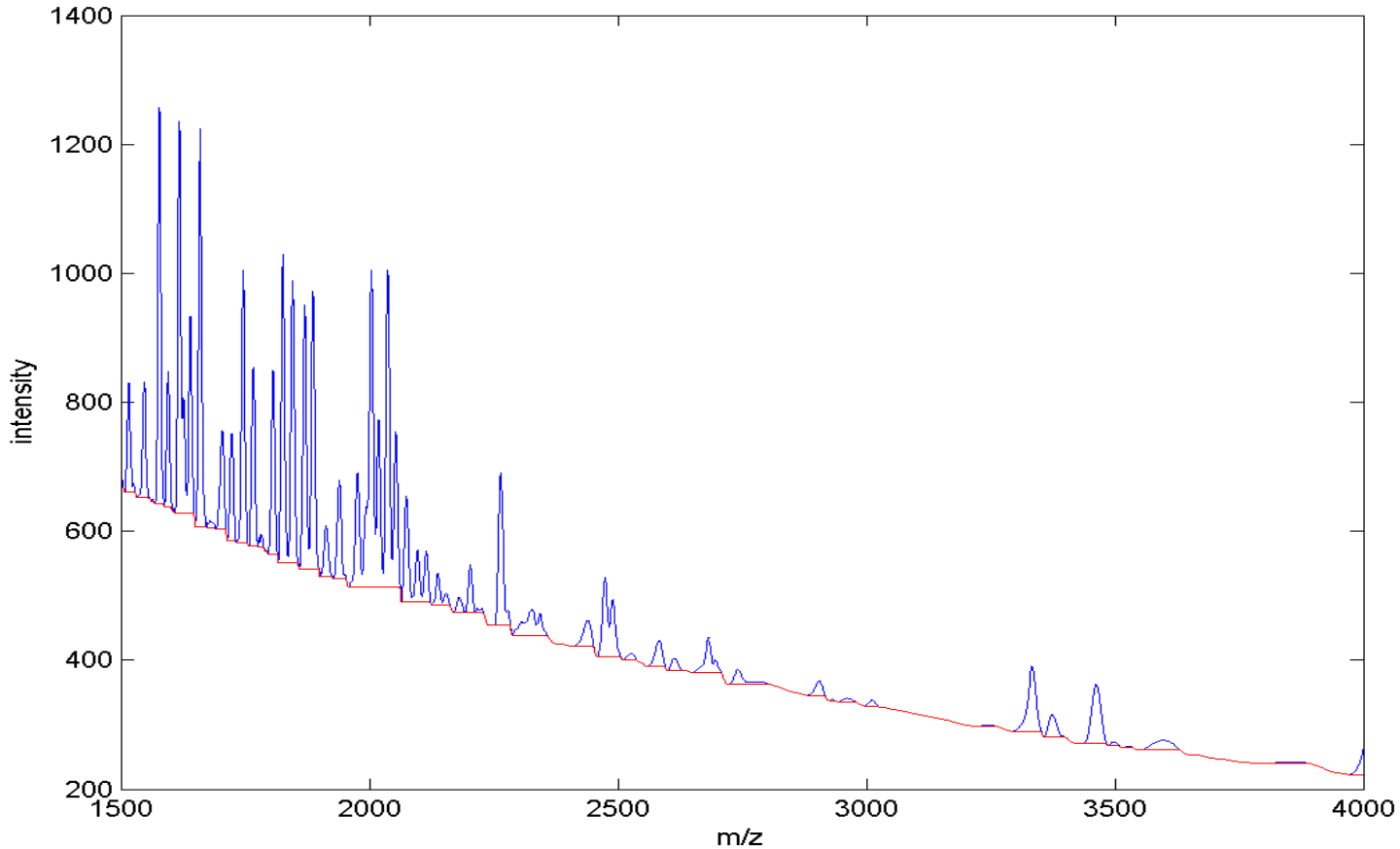
Noise



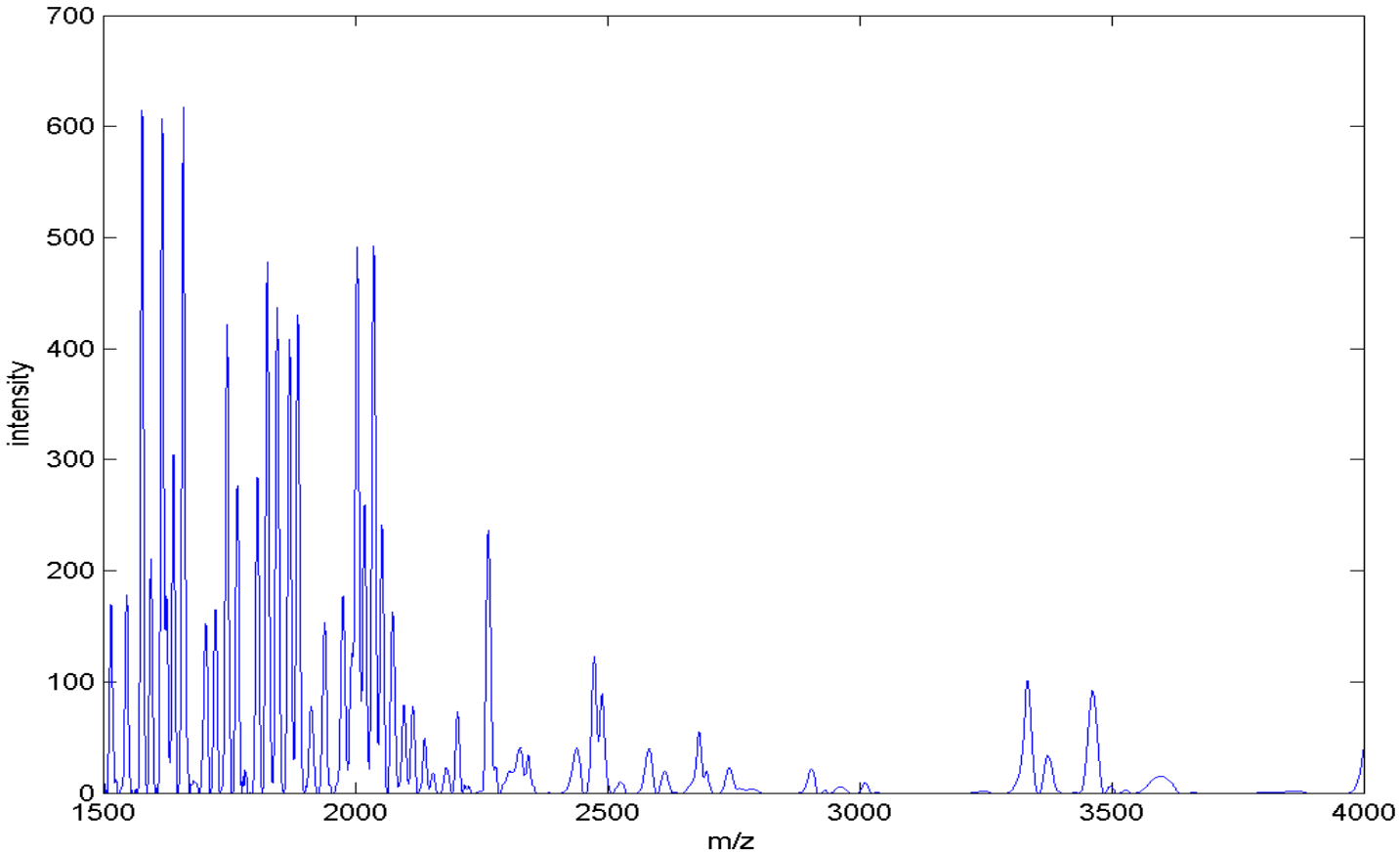
Baseline Correction & Normalization

- 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.**

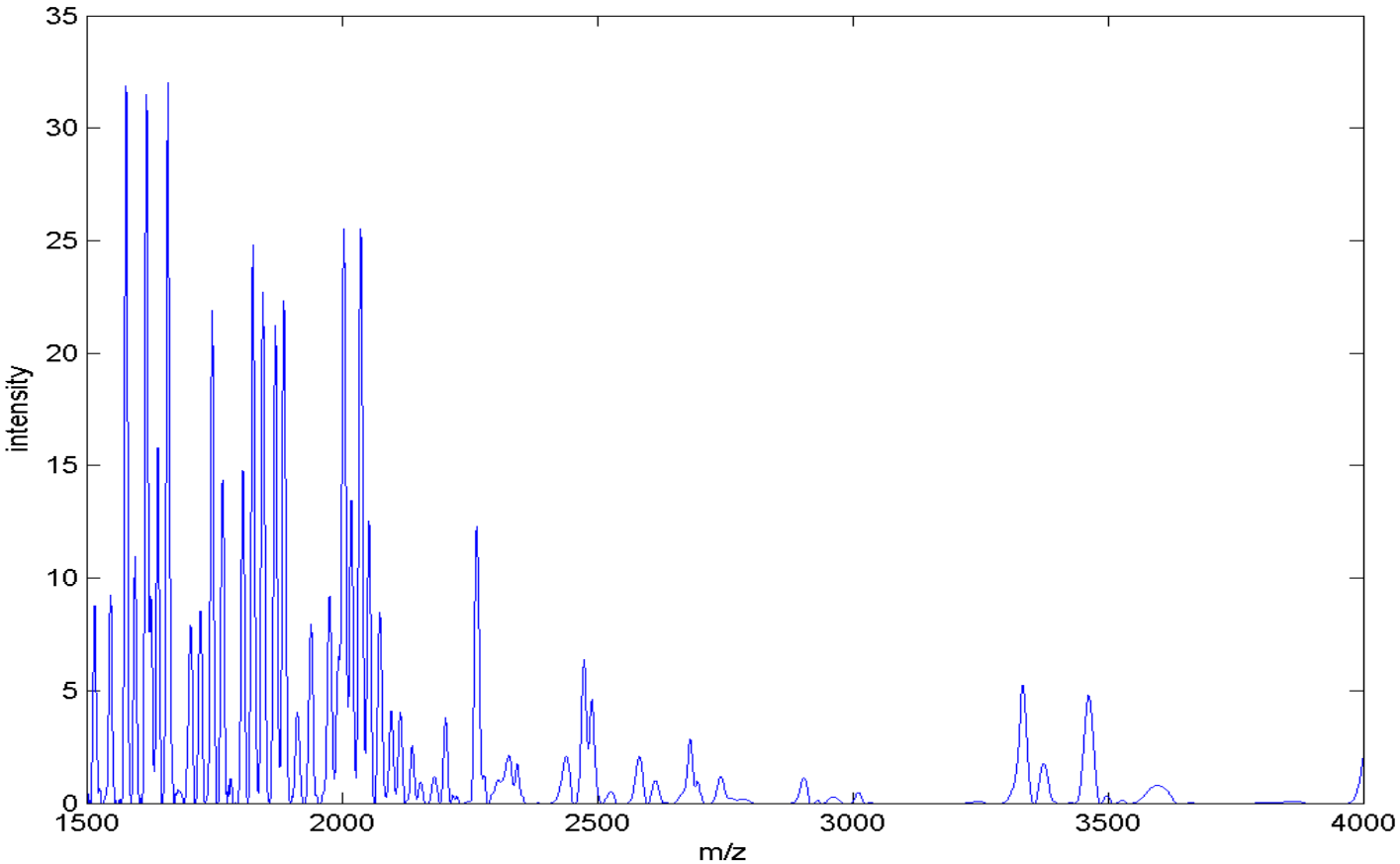
Baseline Estimate



Denoised, Baseline Corrected Spectrum



Denoised, Baseline Corrected, and Normalized



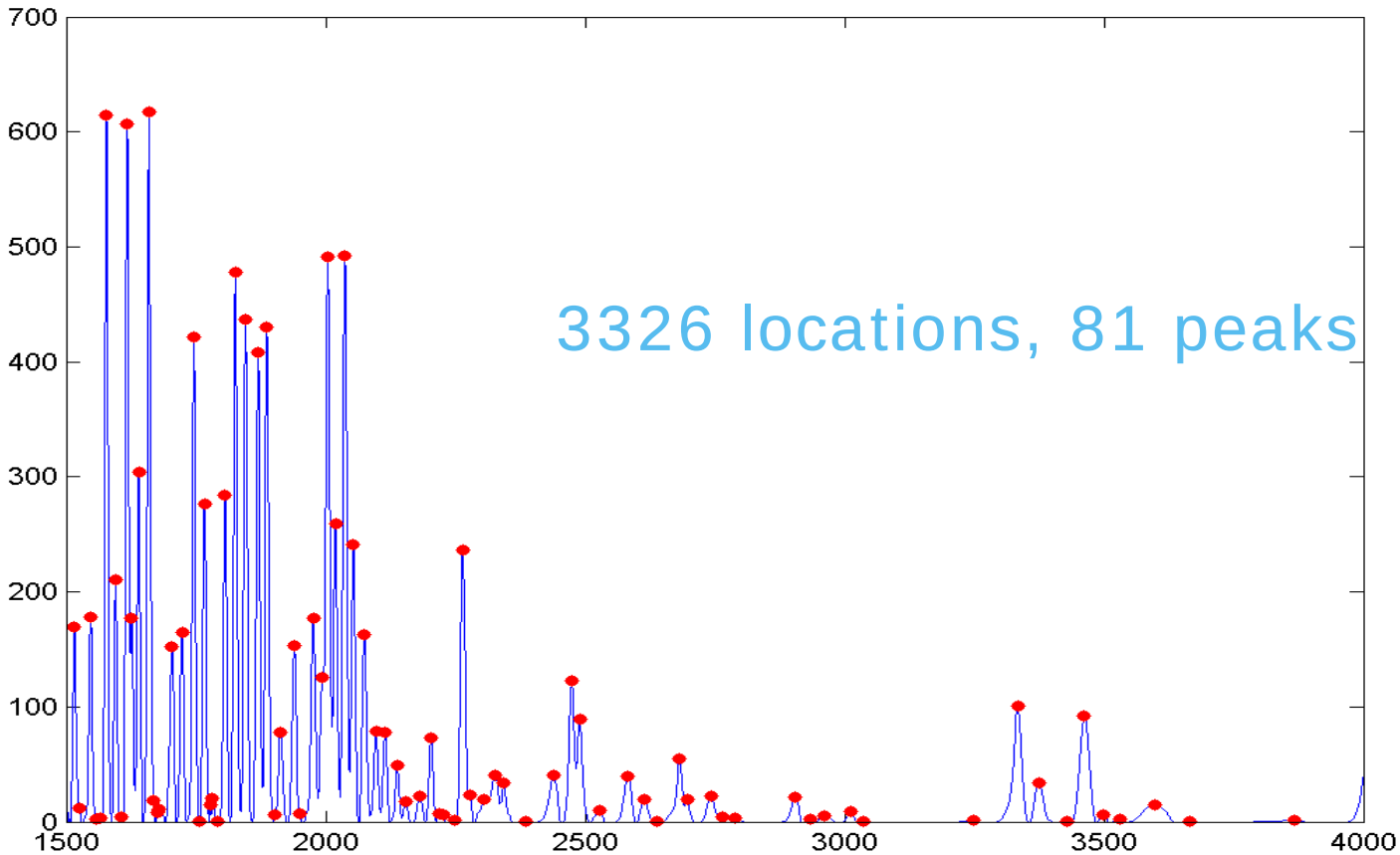
Protein Signal

- 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.**

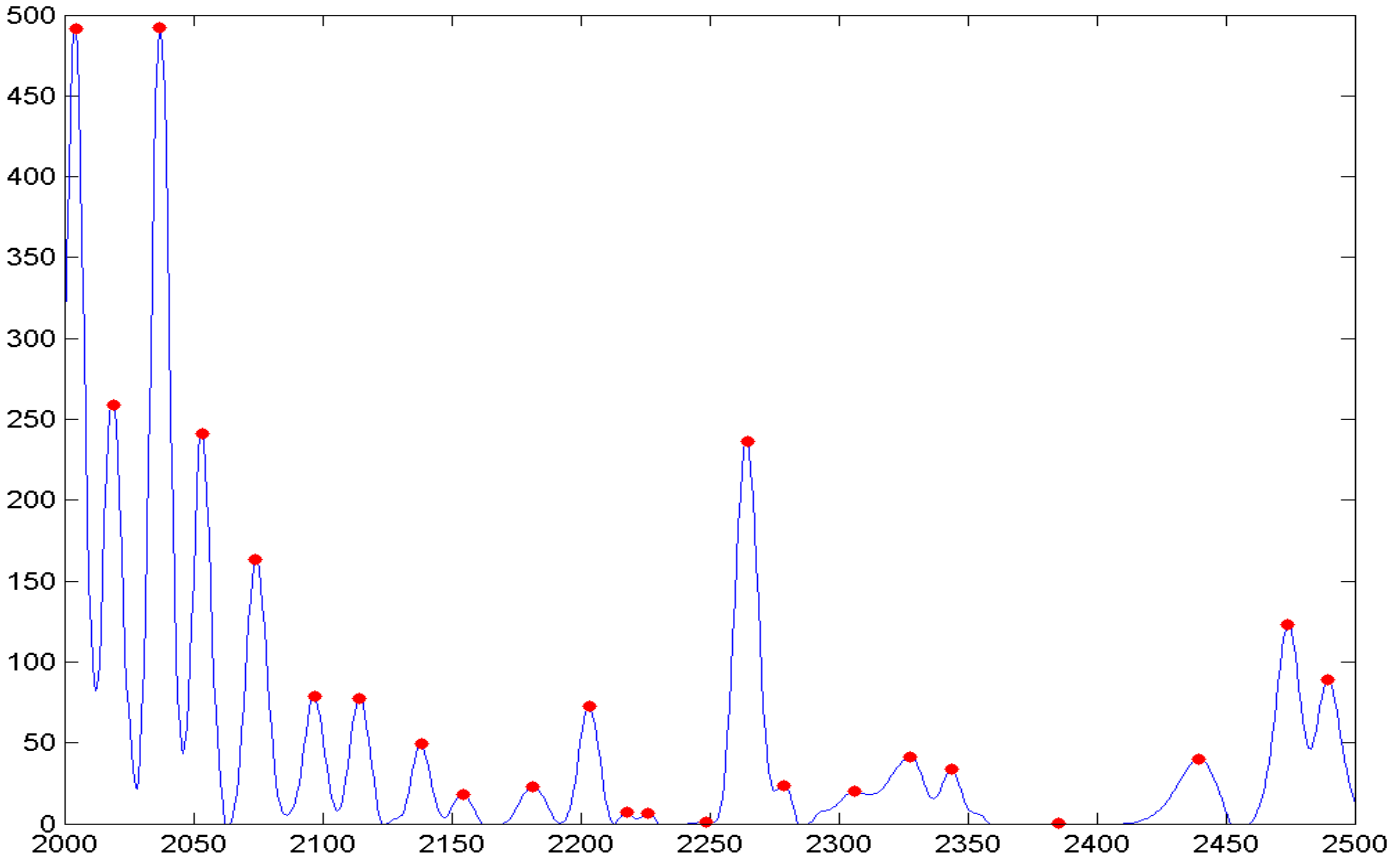
Peak Detection

- 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>δ 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.

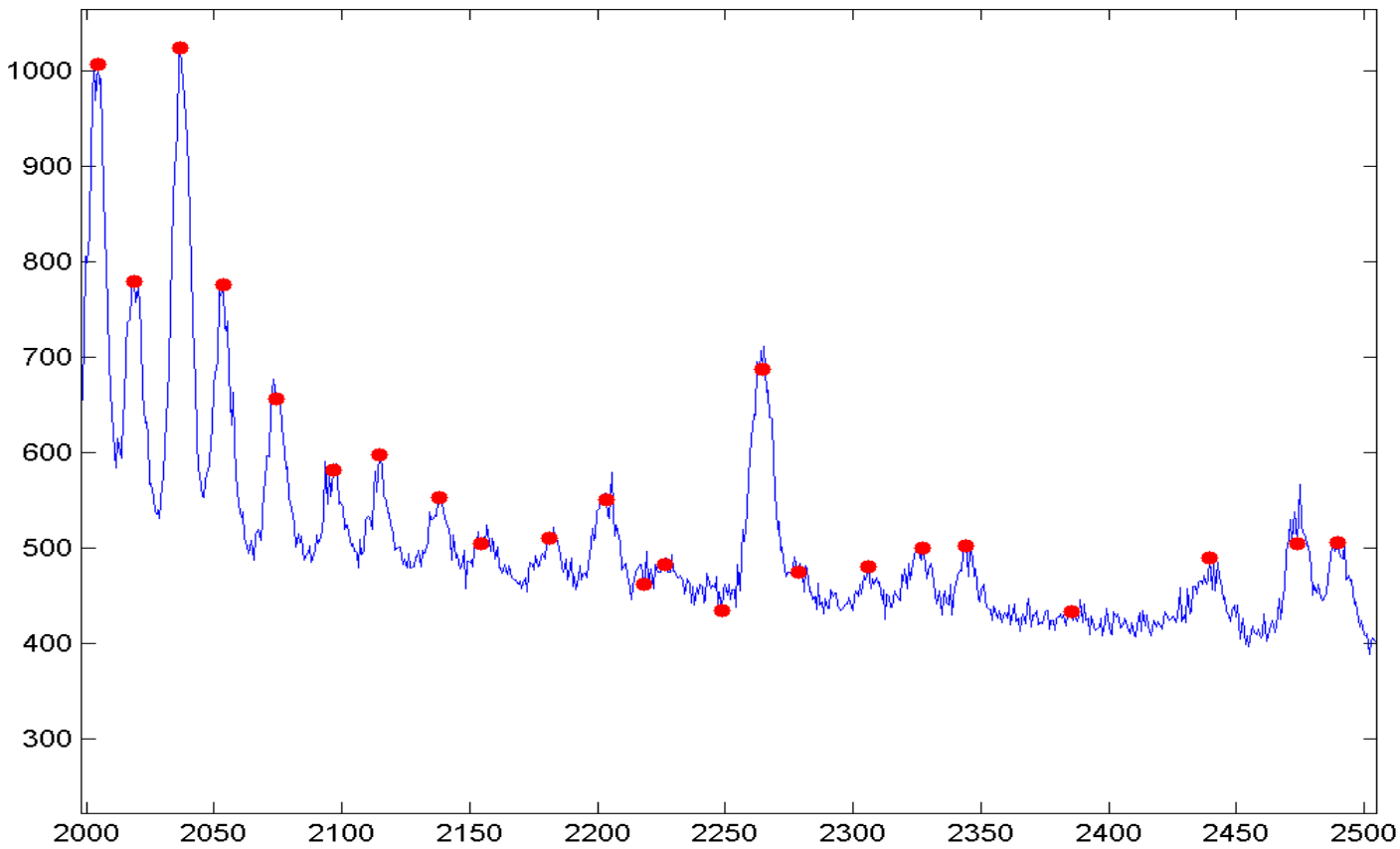
Peak Detection



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*.

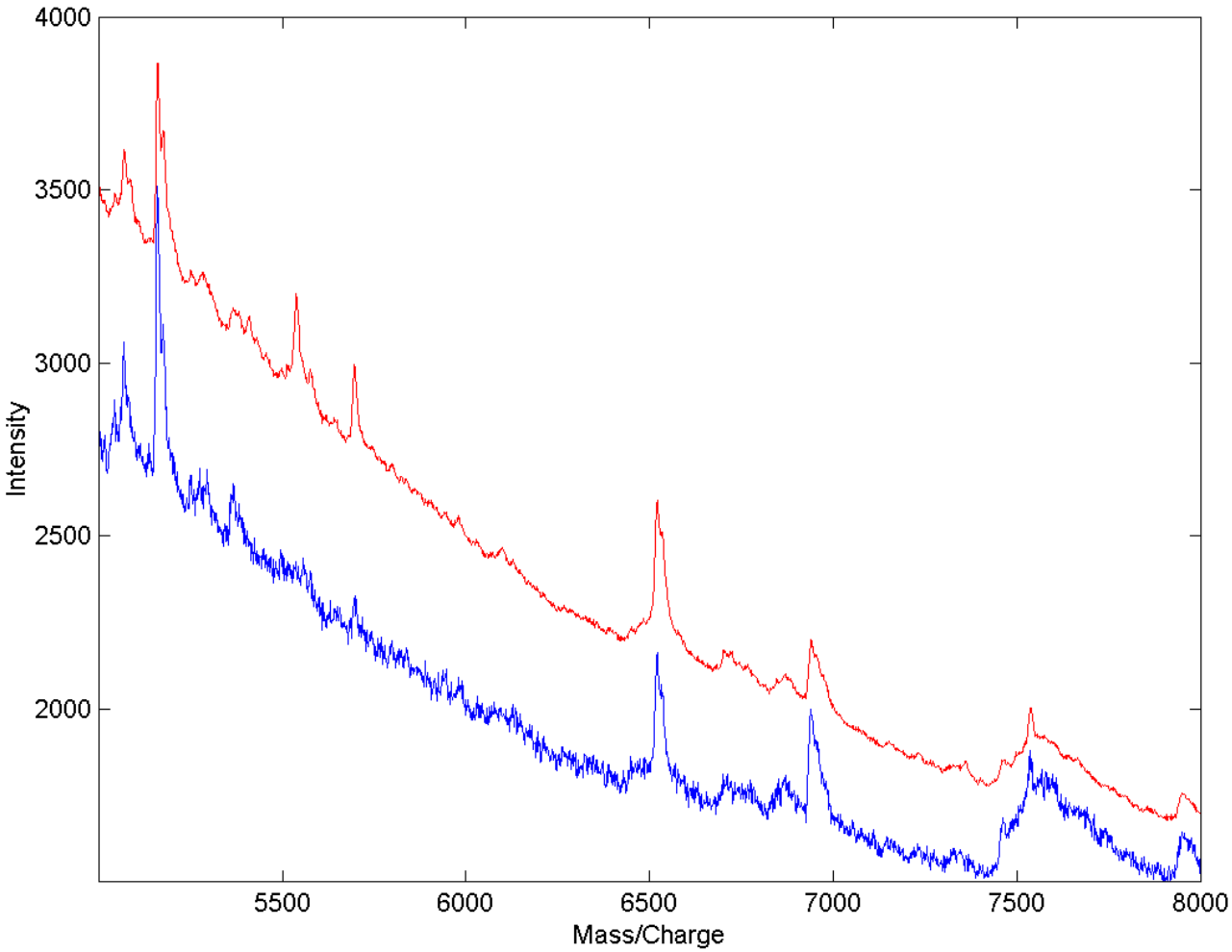
Peak Matching Problem

- 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.

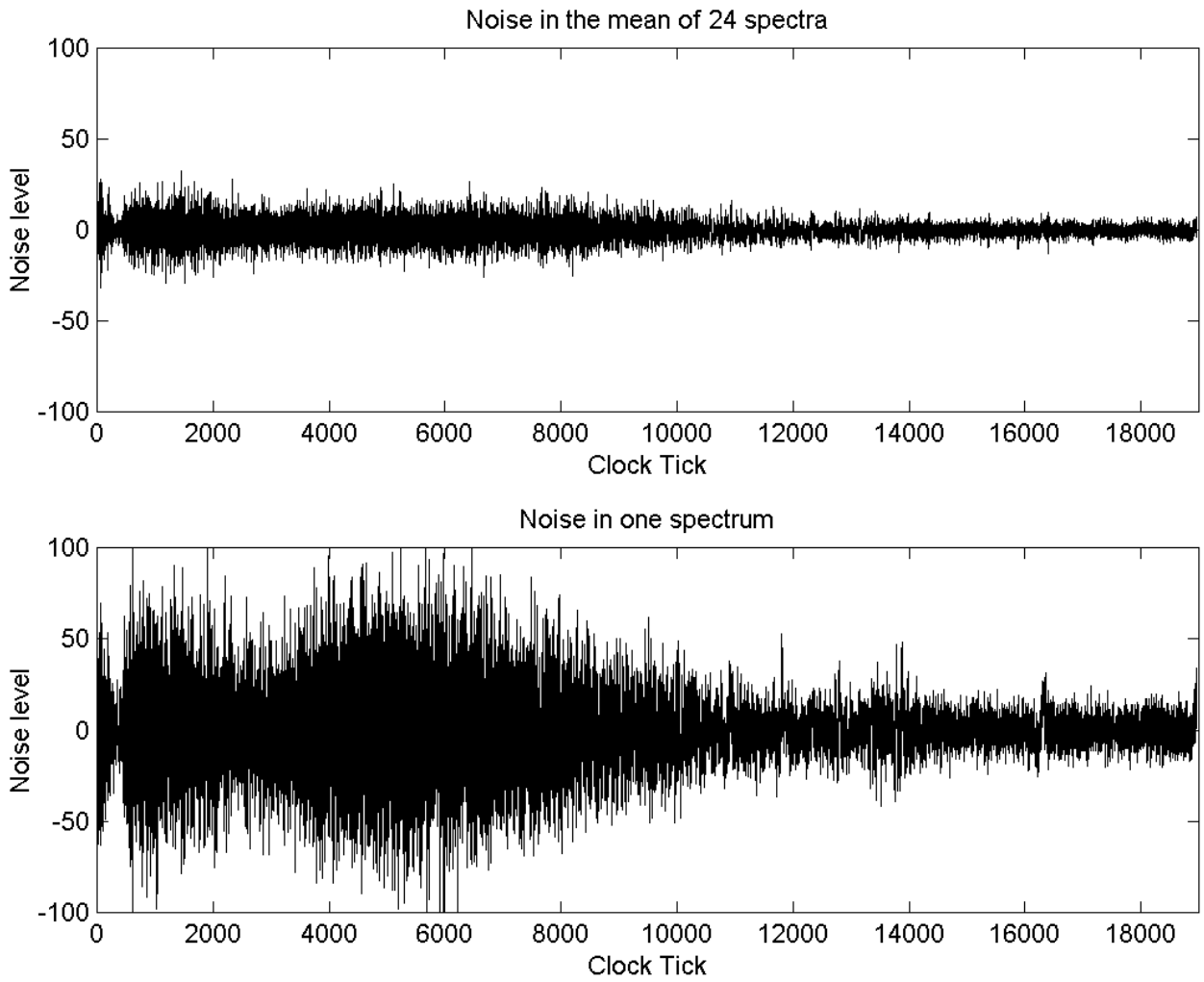
Advantages/Disadvantages

- **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**
  - May be less sensitive when prevalence of peak < 1/sqrt(n).

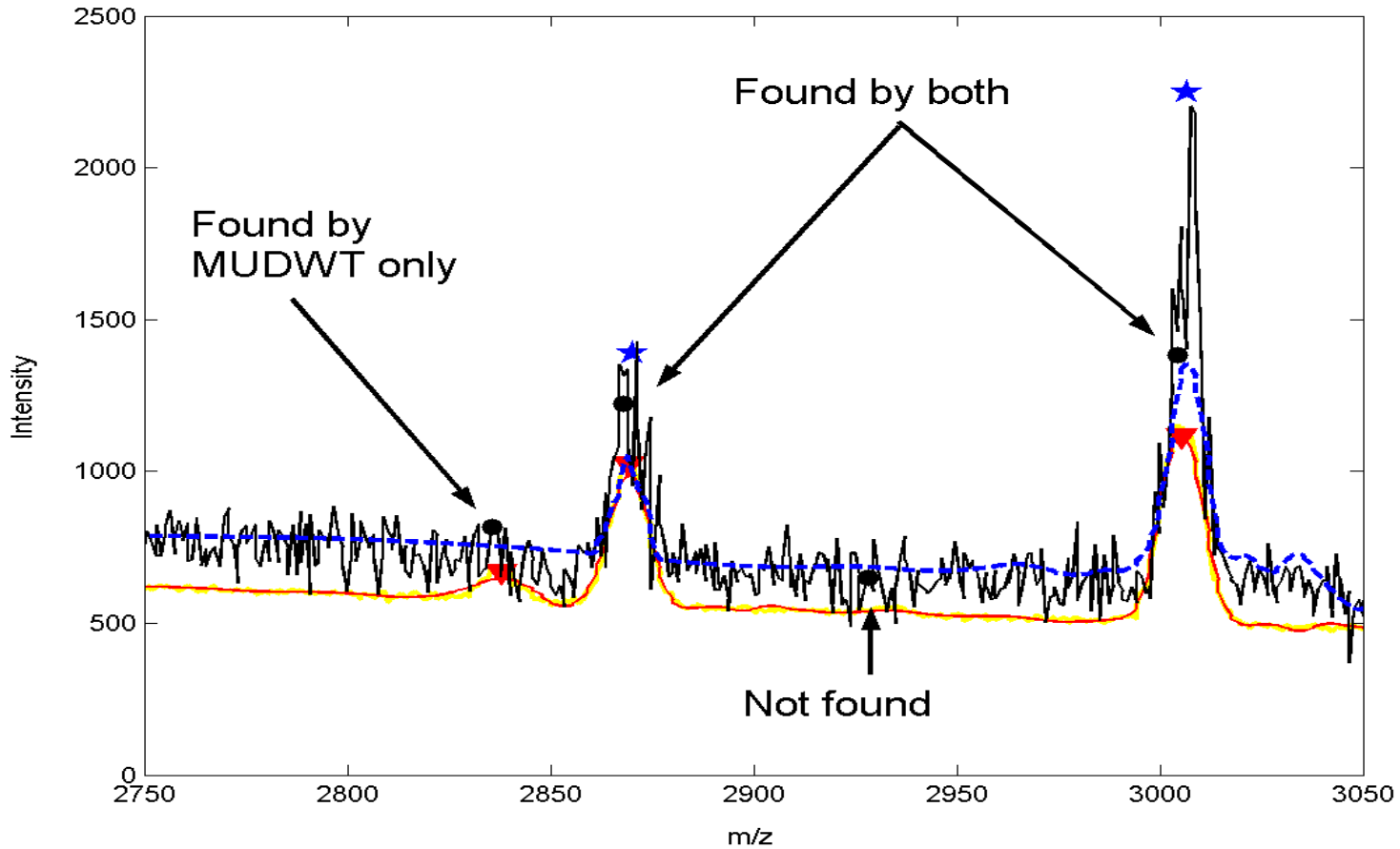
Noise reduced in mean spectrum



Noise reduced in mean spectrum



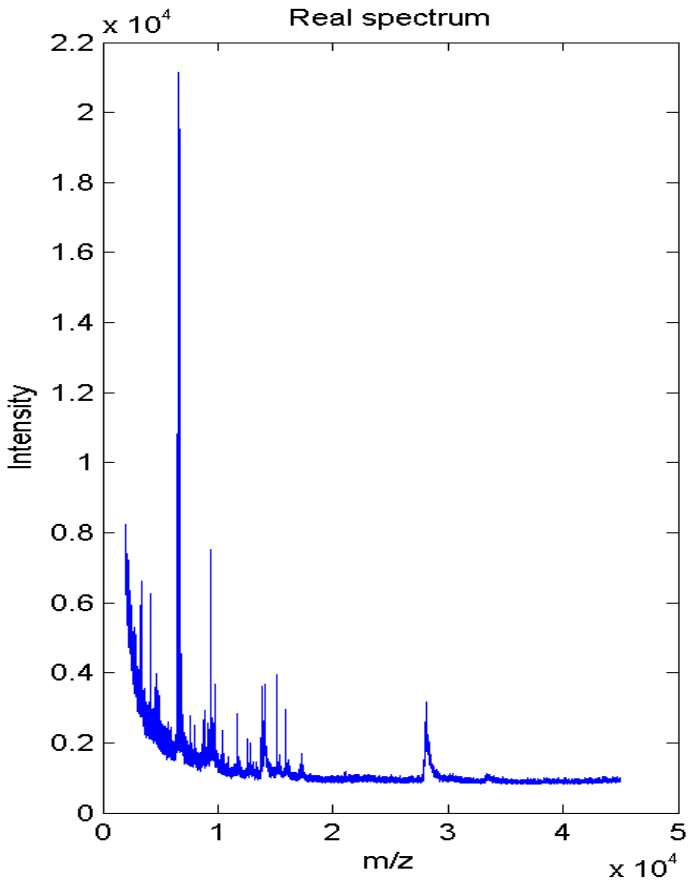
Sample Spectrum

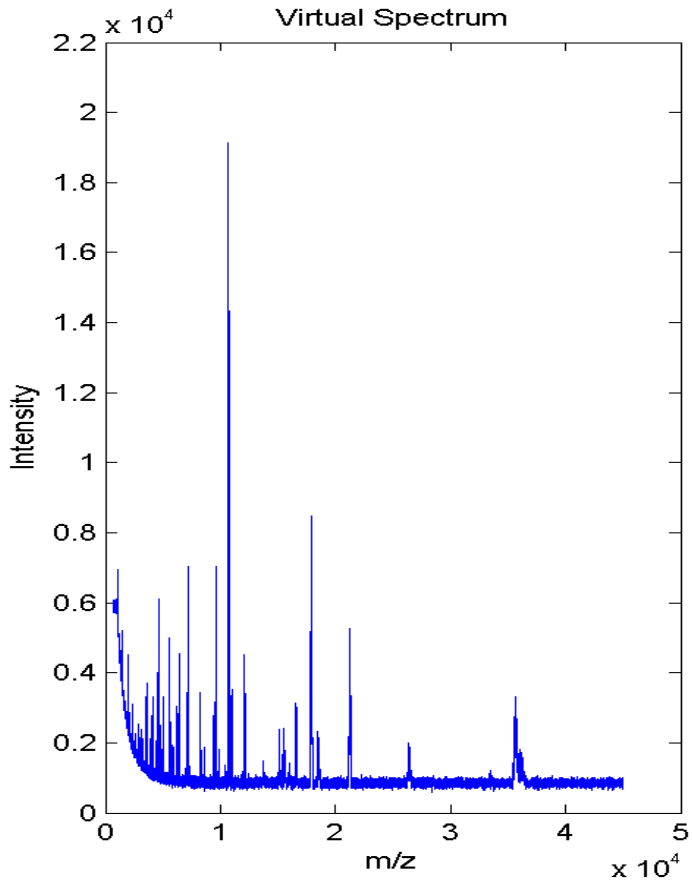


Simulated spectra

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

Real and Virtual Spectra





Simulation Results

	sensitivity	FDR	pv*
SUDWT (indiv. spectra)	**0.75**	**0.09**	**0.03**
MUDWT (mean spectrum)	**0.83**	**0.06**	**0.97**

*pv=the proportion of simulations with higher sensitivity

Simulation Results (by Prevalence)

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

Simulation Results (by abundance)

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

Preprocessing 2d gels

Usual Approach (e.g. PDQ, Progenesis)

- Background correct and normalize individual gels
- Detect spots on individual gels
- Match spots on each gel with spots on chosen reference gel
- Detect spot boundaries, quantify each spot on each gel by normalized spot volume.

Preprocessing 2d gels

Problems with Standard Approach:

1. Time consuming (run overnight?)
2. Complicated algorithms lead to many errors:
   - Detection errors (miss/split/merge)
   - Matching errors
   - Errors/variability in spot boundaries

   These errors tend to increase as more gels are run in a given experiment, encouraging researchers to run small studies that may be underpowered for detecting realistic differences
3. Requires hand editing (days/weeks?)
4. Results in many "missing spots": What to do about them?

Preprocessing 2d gels

Our Approach

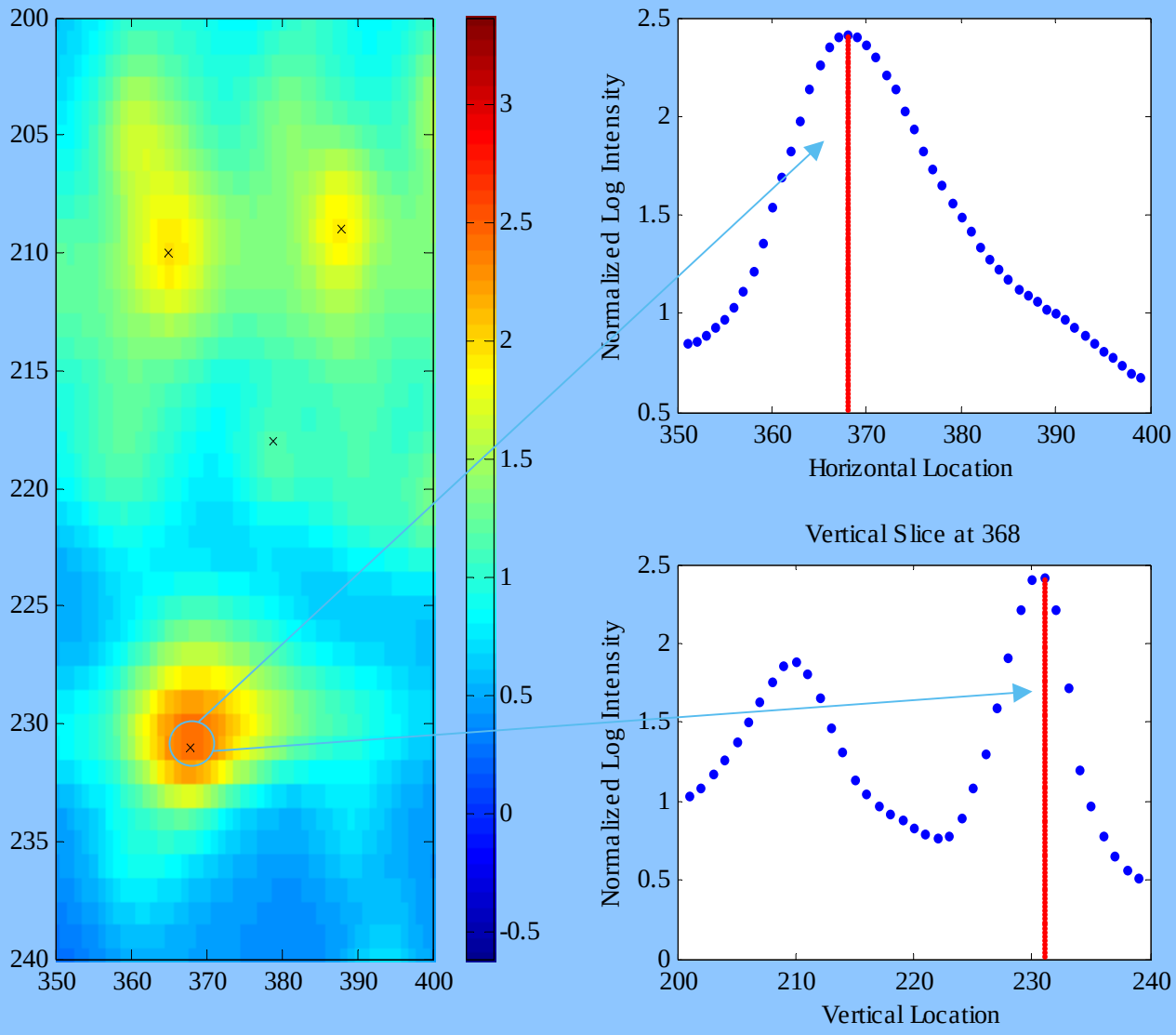
- Align gel images
- Compute average gel
- Denoise average gel using wavelets
- Detect spots on average gel using ***pinnacles***
- Background correct and normalize individual gels
- Quantify each spot on each gel by taking maximum pixel intensity in neighborhood of pinnacle

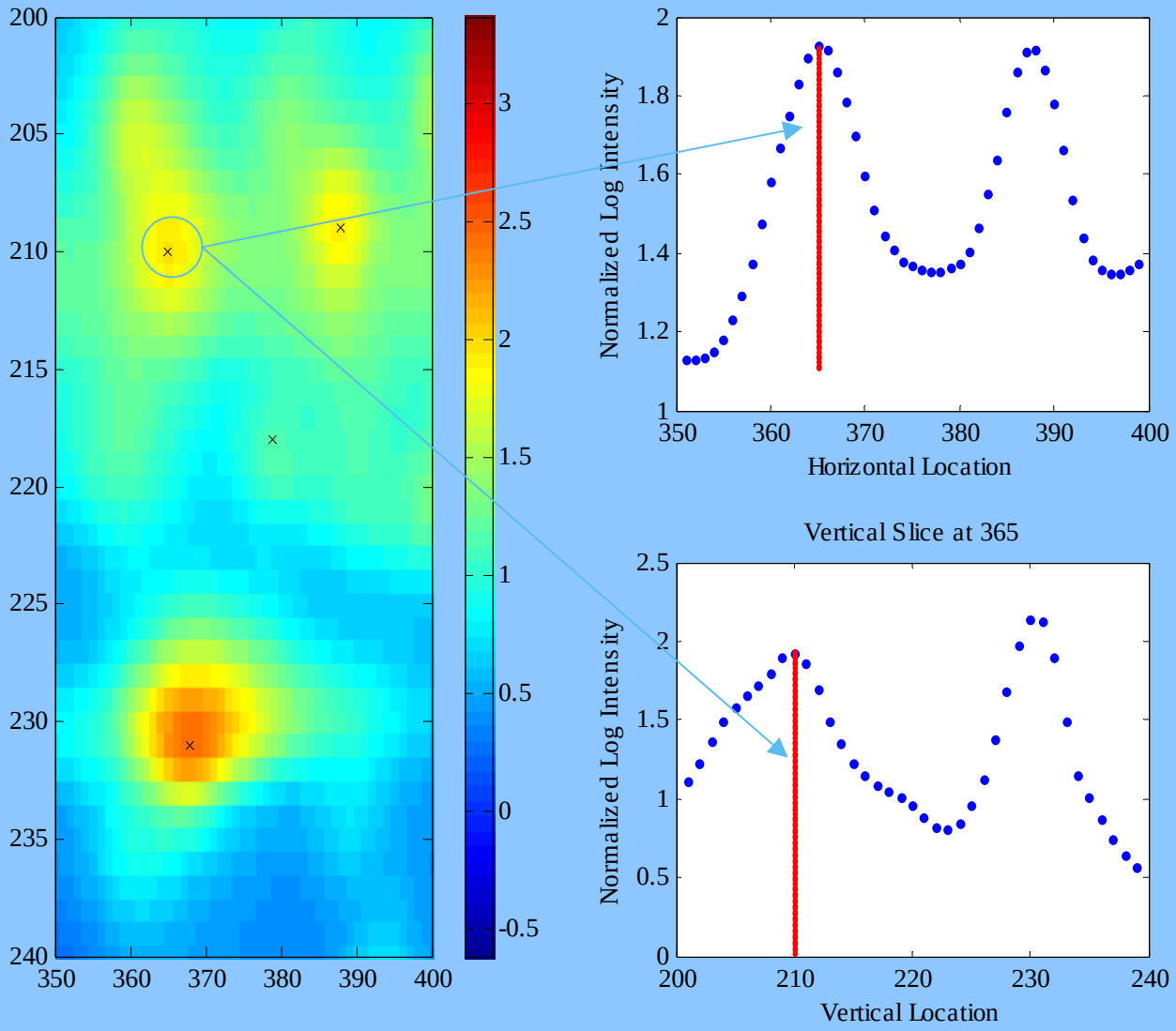
Gel Alignment

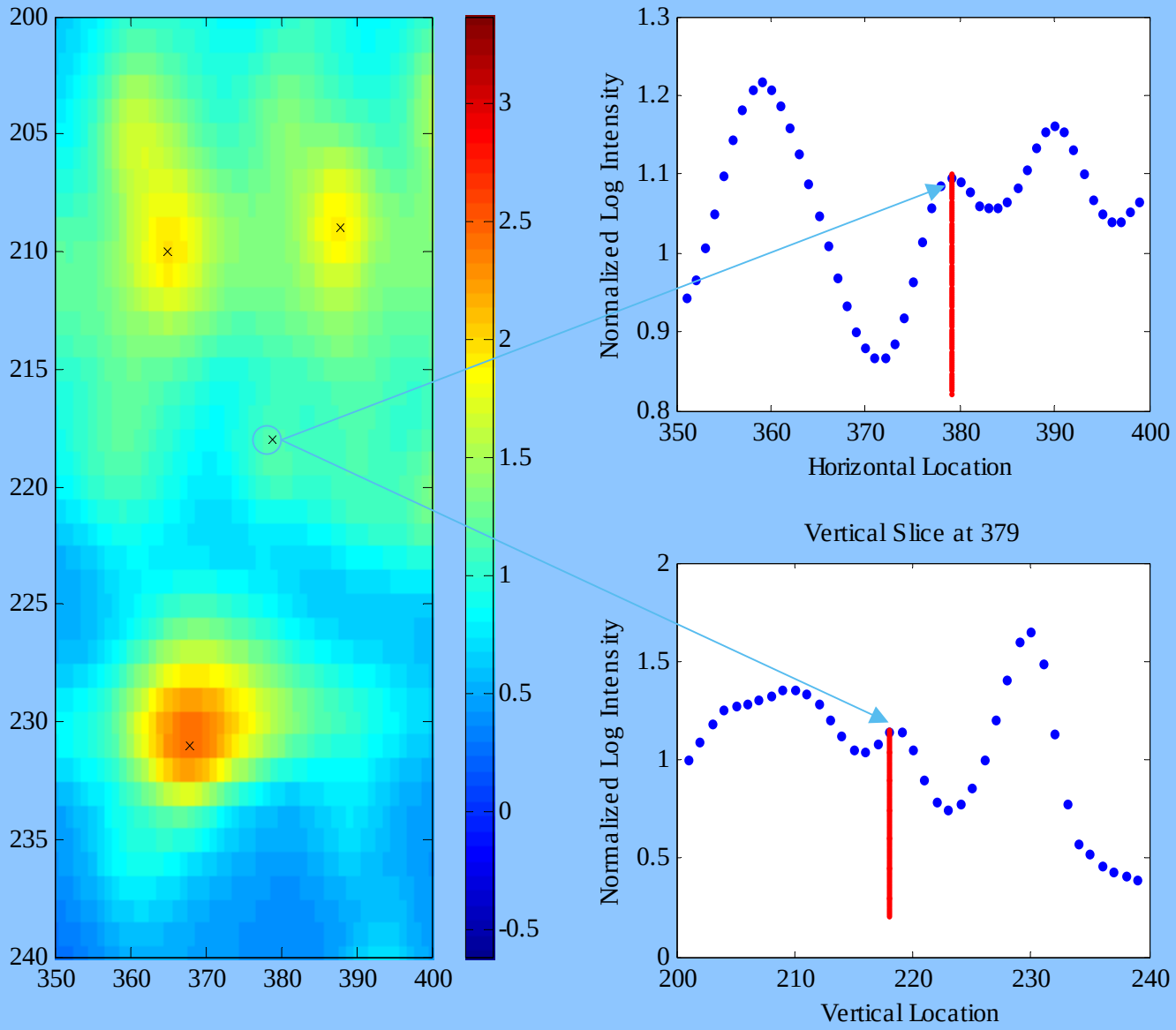
- Warp all gels to chosen reference gel so spots are aligned across gels
  - Easier and more accurate than matching detected spots, since warping algorithm can borrow strength from nearby regions of the gel when aligning spots
- We use TT900 (Nonlinear) to do the warping; other image registration programs are available and being developed.

Spot Detection

- Use of the average gel results in more **sensitivity** and **specificity** for spot detection
- Denoising the average gel using wavelets reduces the number of artifact spots found
- We identify spots based on their corresponding *pinnacles*
- A pixel location (x,y) on the gel is a *pinnacle* if:
  1. It is a *peak* (local maximum) in both the horizontal and vertical directions
  2. It has a pixel intensity above some minimum threshold (e.g. median intensity on gel)


Horizontal slice at 231
Normalized Log Intensity
Horizontal Location
Vertical Slice at 368
Normalized Log Intensity
Vertical Location


Horizontal slice at 210
Normalized Log Intensity
Horizontal Location
Vertical Slice at 365
Normalized Log Intensity
Vertical Location

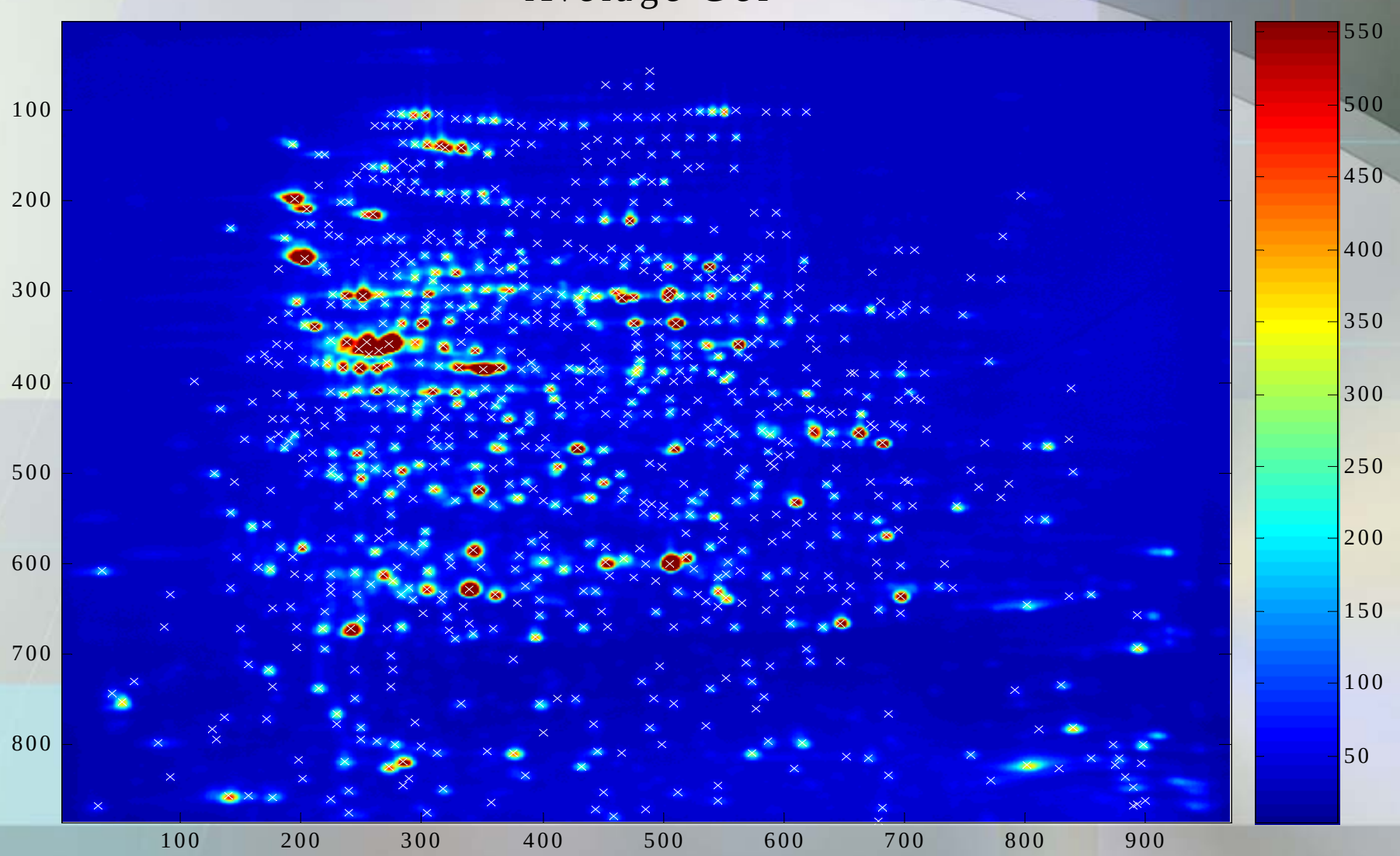

Horizontal slice at 218
Normalized Log Intensity
Horizontal Location
Vertical Slice at 379
Normalized Log Intensity
Vertical Location

Spot Detection

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

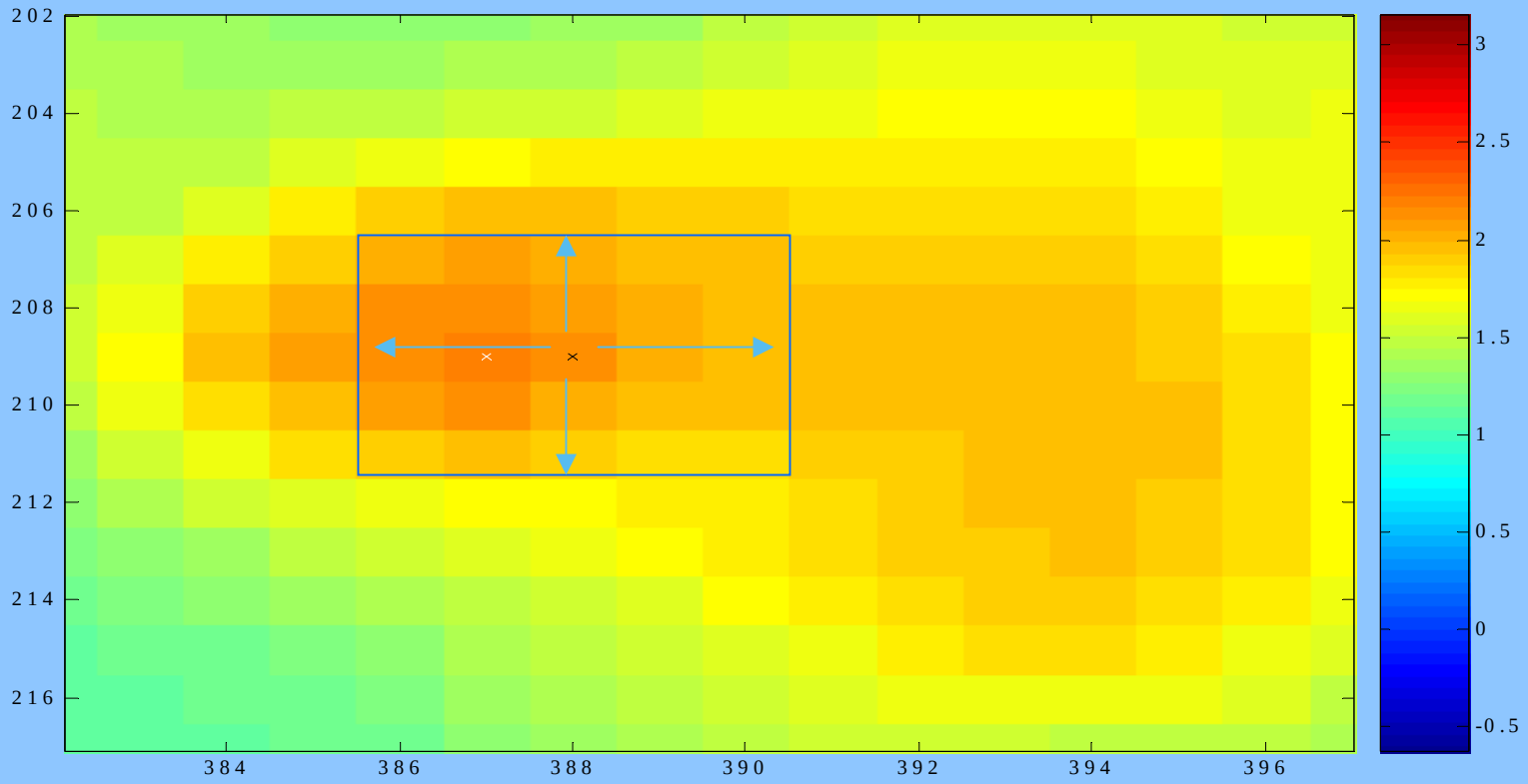
1. Unambiguous definition
2. Not affected by overlapping spots
3. No need to find spot boundaries
4. Seems to capture most real spots

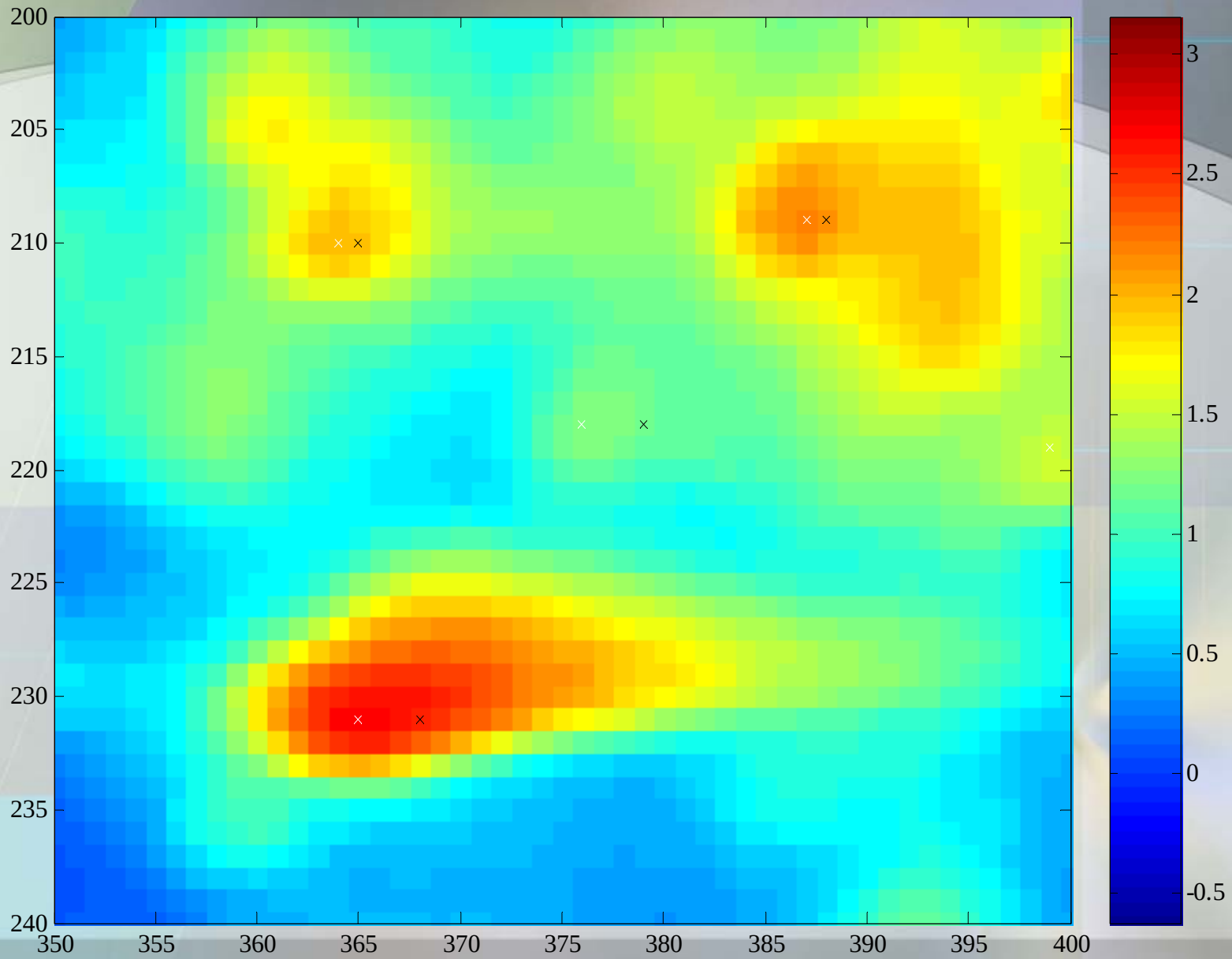
Results: Spot Detection

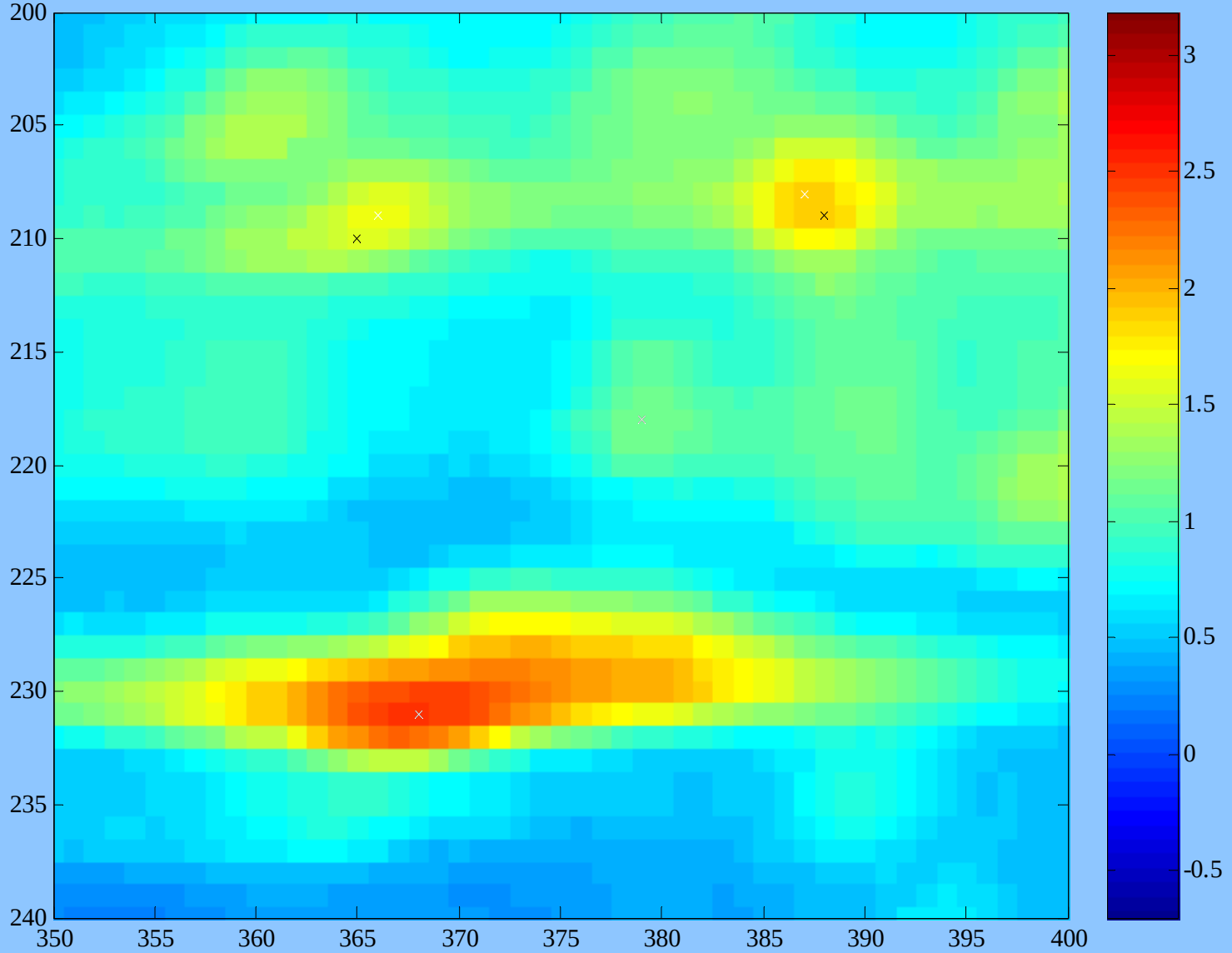


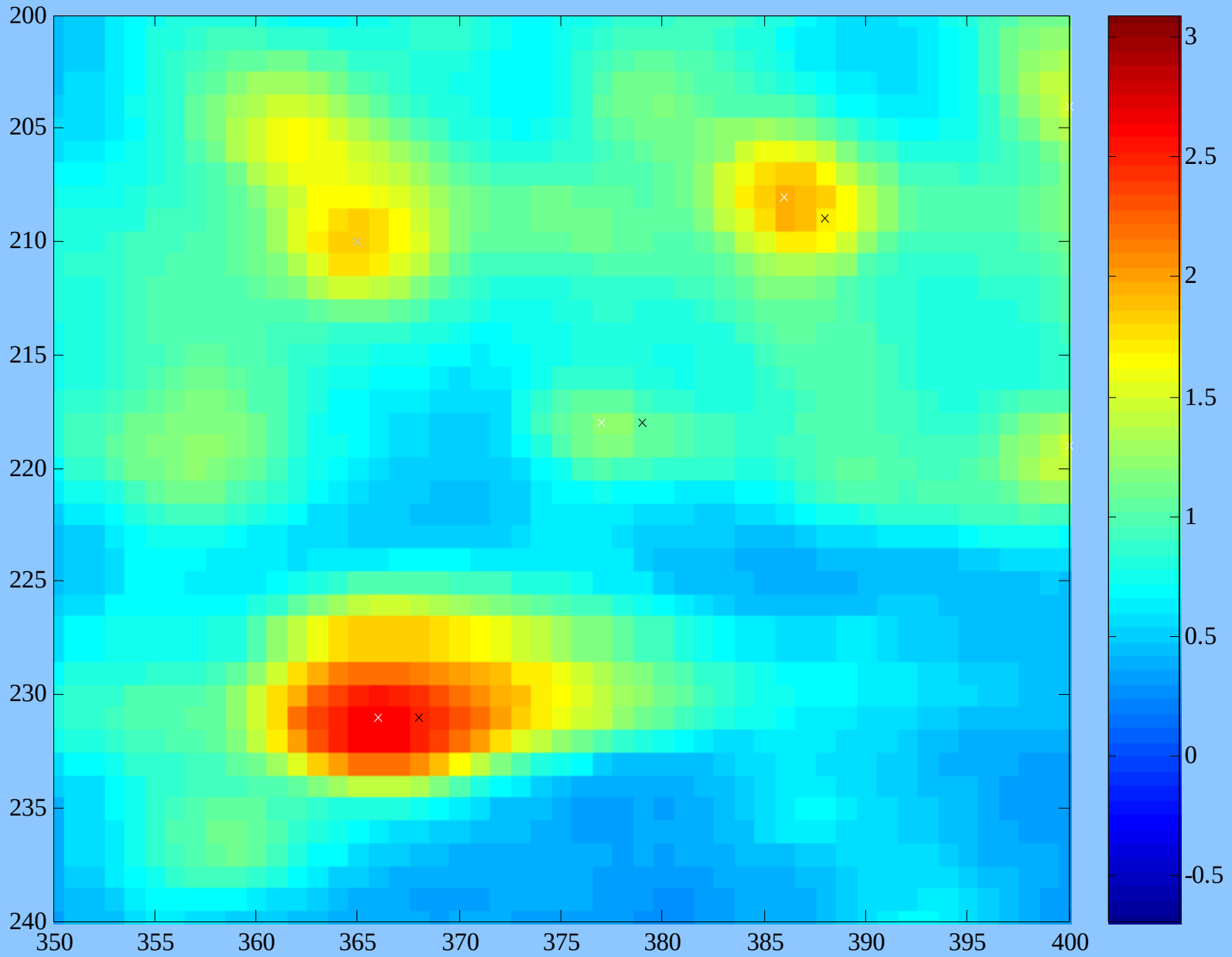
Spot Quantification

- 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

Spot Quantification

- **Why use pinnacle intensities? Why not use spot volumes?**
  1. Pixel intensity at a spot's pinnacle is highly correlated its volume.
  2. No need to detect spot boundaries, which reduces CV of quantification
  3. Much quicker and easier
  4. Results in spot intensities for each spot for every gel, i.e. no missing spots
- This approach leads to more reliable and precise spot quantifications

Validation: Dilution Series

- Nishihara and Champion (2002) conducted a dilution series experiment to validate PDQuest and Progenesis methods
- 4 replicate gels for each of 7 protein loads
  0.5μg, 7.5μg, 10μg, 15μg, 30μg, 40μg, 50 μg
- **Reliability** assessed by computing R^2 from regression of spot quantification on protein load
- **Precision** assessed by computing CV for 30μg load
- They only assessed set of 20 "representative" spots, and found good results (mean $R^2 = 0.98$, CV~15)

Results: 20 selected spots

Method	mean R^2	mean CV
Pinnacle	0.984	17.8
PDQuest	0.986	12.6
Progenesis	0.983	16.8

- Results of pinnacle method for 20 selected spots comparable
- What about the other ~1000 spots?

Results: Reliability (Linearity)

Method	spots detected	spots with $R^2 > 0.95$	median R^2
Pinnacle			
PDQuest	1448	636	0.936
Progenesis			

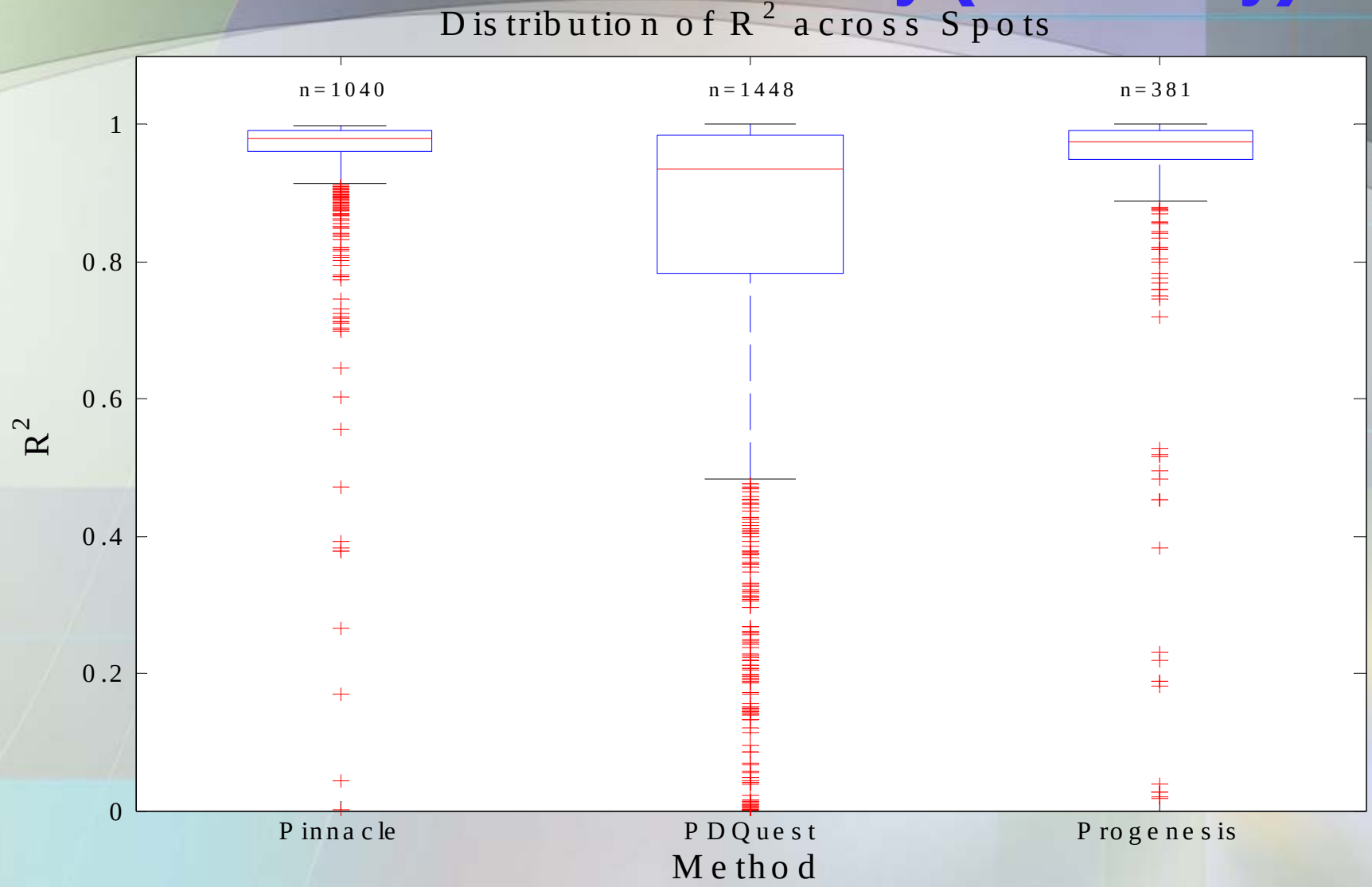
Results: Reliability (Linearity)

Method	spots detected	spots with $R^2 > 0.95$	median R^2
Pinnacle			
PDQuest	1448	636	0.936
Progenesis	381	286	0.975

Results: Reliability (Linearity)

Method	spots detected	spots with $R^2 > 0.95$	median R^2
Pinnacle	1040	853	0.980
PDQuest	1448	636	0.936
Progenesis	381	286	0.975

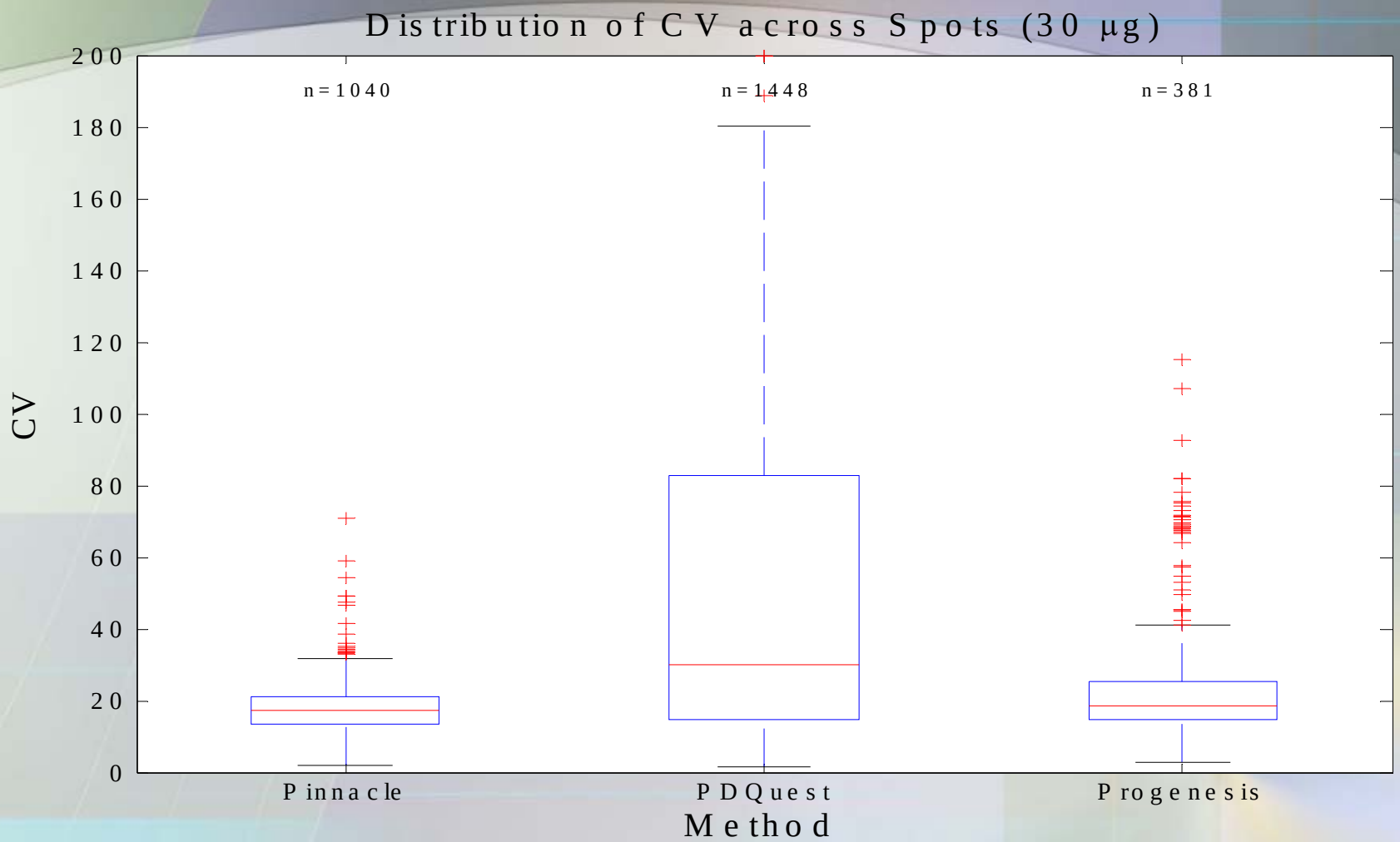
Results: Reliability (Linearity)



Results: Precision

Method	spots detected	spots with CV<20%	median CV (%)
Pinnacle	1040	723	17.2
PDQuest	1377*	519	29.9
Progenesis	367*	204	18.7

Results: Precision

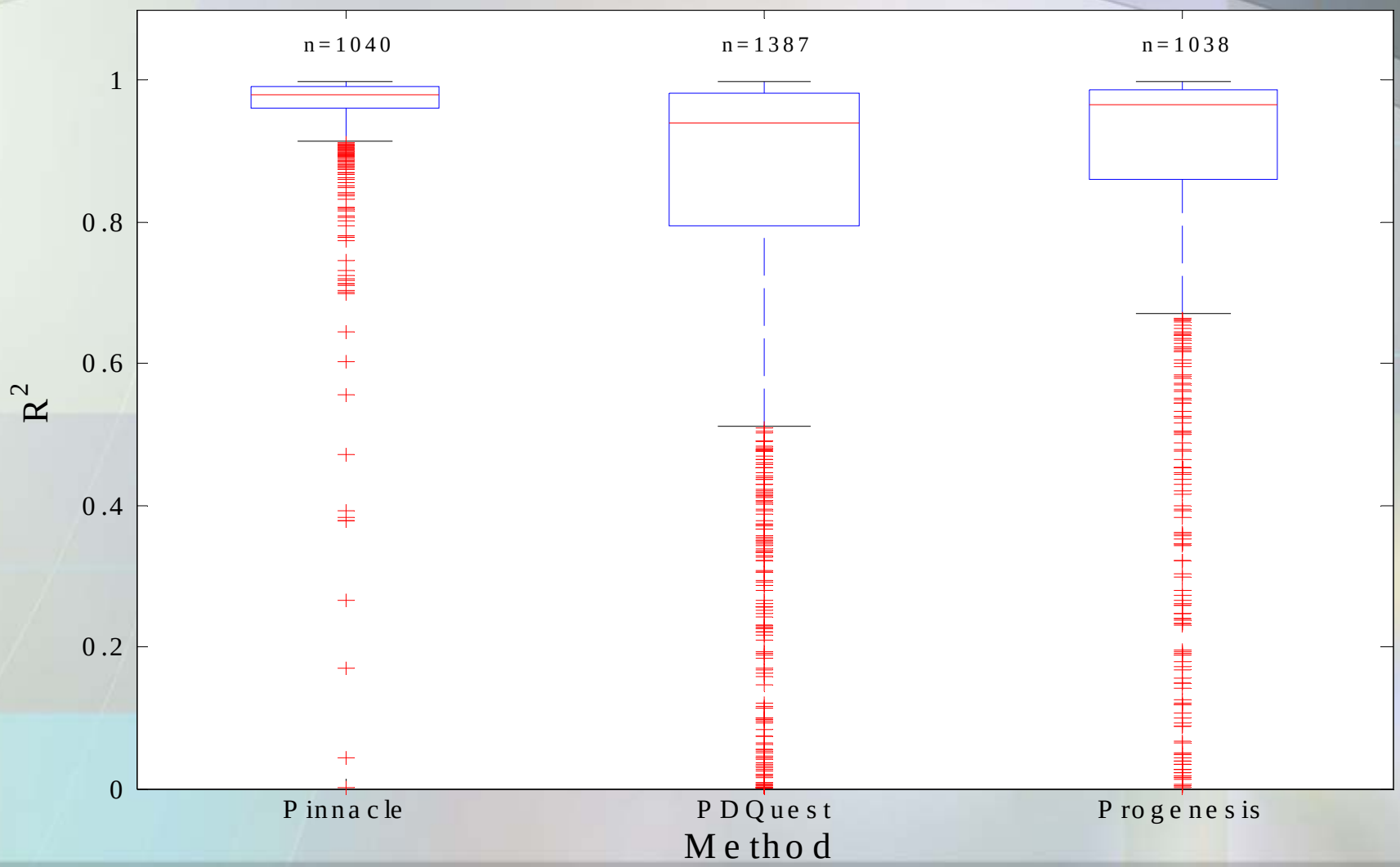


- **What if PDQuest and Progenesis are run on pre-aligned gels?**

Results: Reliability (Aligned Gels)

Method	spots detected	spots with $R^2 > 0.95$	median R^2
Pinnacle	1040	853	0.980
PDQuest (after alignment)	1387	639	0.940
Progenesis (after alignment)	1038	592	0.965

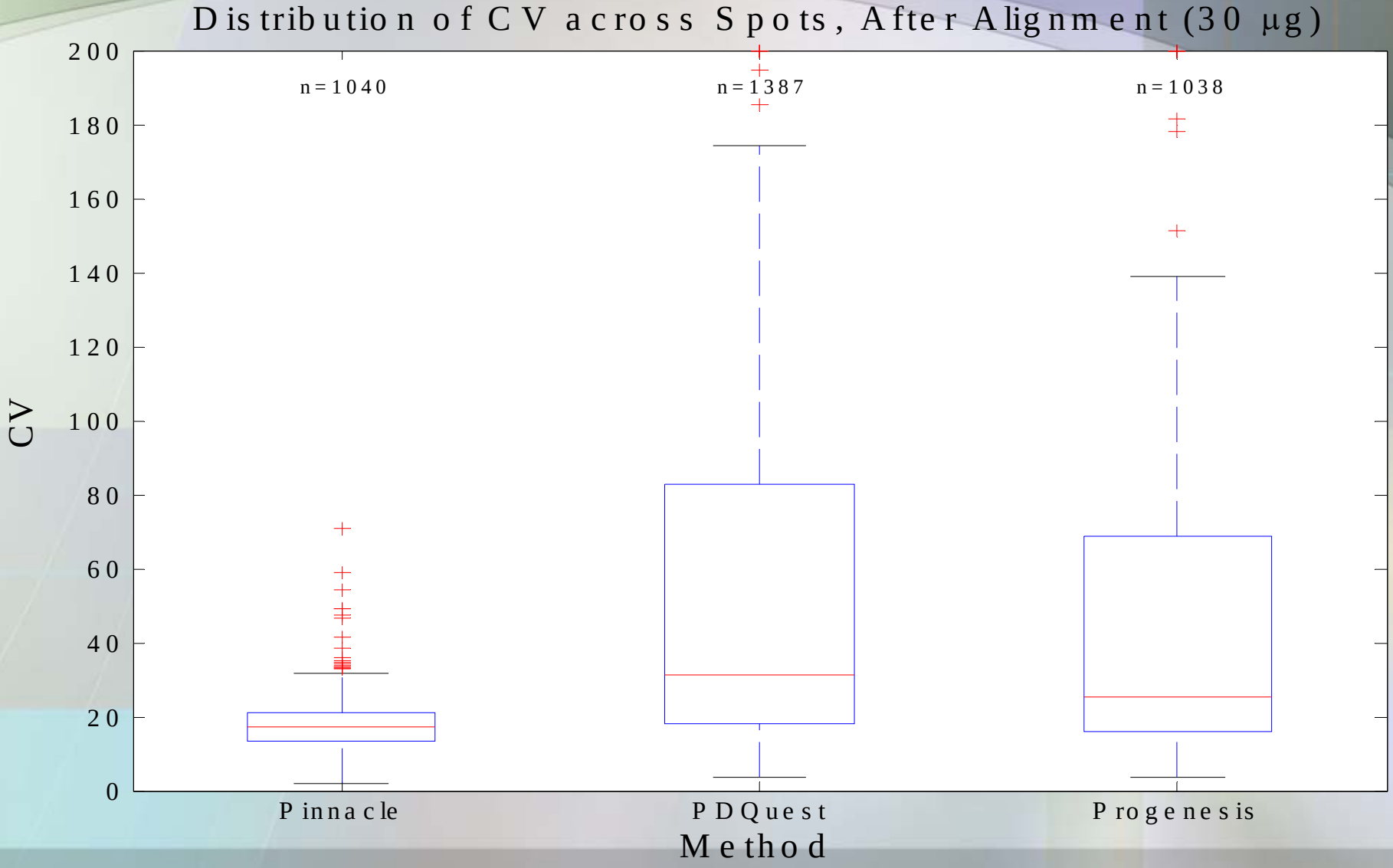
Results: Reliability (Aligned Gels)



Results: Precision (Aligned Gels)

Method	spots detected	spots with CV<20%	median CV (%)
Pinnacle	1040	723	17.2
PDQuest (after alignment)	1326*	392	31.5
Progenesis (after alignment)	942*	340	25.1

Results: Precisionc (Aligned Gels)



Advantages of our approach

- **Automatic** – After alignment, fully automated
- **Quick to implement** – <1 minute for 60 gels
- **Effective** – appears to work very well, finding most "real" spots
- **Sensitive** – 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 can eliminate artifacts limited to single gel
- **No missing spots** – we get quantifications for each spot on every gel
- **Reliable and Precise** – The use of the average gel and pinnacles results in more reliable and precise quantifications than standard approaches

Conclusions

- Statistical Input is valuable at all levels of Proteomics experiment
  - Experimental Design Phase
  - Preprocessing
  - Data Analysis and Discovery
- Principles:
  - Randomize! Randomize! Randomize!
  - Look at the Data!
  - Preprocessing is important!

Acknowledgements

- **Bioinformatics**
  - Kevin Coombes
  - Keith Baggerly
  - Phil Brown (U Kent)
  - Jianhua Hu
  - Jing Wang
  - Lianchun Xiao
  - Spyros Tsavachidis
  - Thomas Liu
  - Auston Wei
- **Proteomics (MDACC)**
  - Howard Gutstein (2D)
  - Ryuji Kobayashi
  - David Hawke
  - John Koomen
- **Ciphergen**
  - Charlotte Clarke
- **Biologists (MDACC)**
  - Howard Gutstein
  - Jim Abbruzzese
  - I.J. Fidler
  - Stan Hamilton
  - Nancy Shih
  - Ken Aldape
  - Henry Kuerer
  - Herb Fritsche
  - Gordon Mills
  - Lajos Pusztai
  - Jack Roth
  - Lin Ji

References:

Experimental Design:

1. Baggerly, KA, Morris JS, and Coombes KR: Reproducibility of SELDI Mass Spectrometry Patterns in Serum: Comparing Proteomic Data Sets from Different Experiments. *Bioinformatics*, 20(5): 777-785, 2004.
2. Baggerly KA, Edmonson S, Morris JS, and Coombes KR: High-Resolution Serum Proteomic Patterns for Ovarian Cancer Detection. *Endocrine-Related Cancers*, 11(4): 583-584, 2004.
3. Hu J, Coombes KR, Morris JS, and Baggerly KA: The Importance of Experimental Design in Proteomic Mass Spectrometry Experiments: Some Cautionary Tales. *Briefings in Genomics and Proteomics,* 3(4), 322-331, 2005.
4. Coombes KR, Morris JS, Hu J, Edmondson SR, and Baggerly KA: Serum Proteomics Profiling: A Young Technology Begins to Mature. *Nature Biotechnology,* 23(3): 291-292, 2005.
5. Baggerly KA, Coombes KR, and Morris JS. Are the NCI/FDA Ovarian Proteomic Data Biased? A Reply to Producers and Consumers. *Cancer Informatics*, 1(1): April 14, 2005.
6. Baggerly KA, Morris JS, Edmonson S, and Coombes KR: Signal in Noise: Evaluating Reported Reproducibility of Serum Proteomic Tests for Ovarian Cancer. *Journal of the National Cancer Institute*, 97: 307-309, 2005 (with commentary).

References:

Simulation:

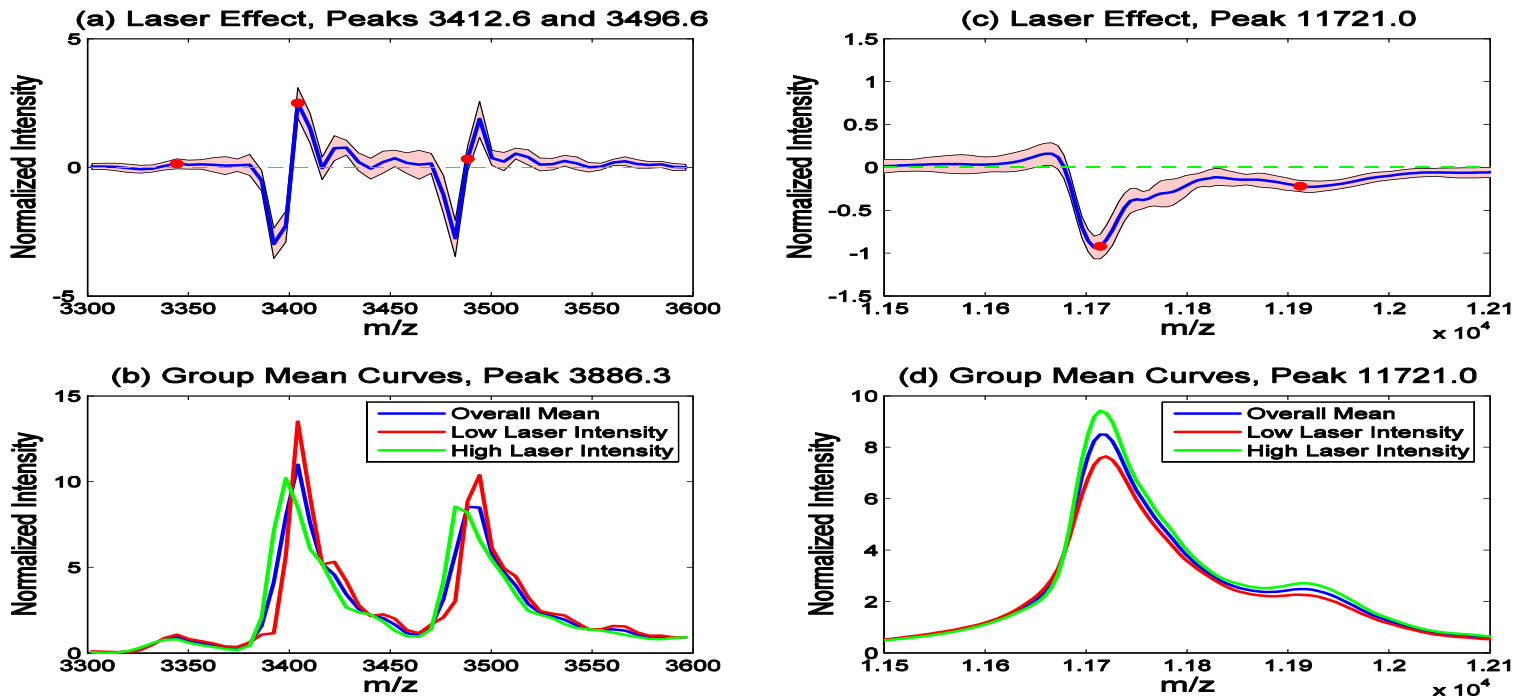
1. Coombes KR, Koomen, JM, Baggerly KA, Morris JS, and Kobayashi R: Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Informatics* 1, 2005.

Preprocessing:

1. Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, and Kuerer HM: Improved Peak Detection and Quantification of Mass Spectrometry Data Acquired from Surface-Enhanced Laser Desorption and Ionization by Denoising Spectra with the Undecimated Discrete Wavelet Transform. *Proteomics*, 5: 4107-4117, 2005.
2. Morris JS, Coombes KR, Kooman J, Baggerly KA, and Kobayashi R: Feature Extraction and Quantification for Mass Spectrometry Data in Biomedical Applications Using the Mean Spectrum. *Bioinformatics,* 21(9): 1764-1775, 2005.
3. Morris JS, Brown PJ, Baggerly KA, Herrick RA, and Coombes KR: Bayesian Methods for Analysing Mass Spectrometry Proteomic Data Using Functional Mixed Models. *Biometrics*, under revision.
4. Morris JS and Gutstein H: Fast, Automatic Pinnacle-Based Method for Detecting and Quantifying Spots in 2d Gel Data Using the Average Gel. In preparation.

- Links to papers and code can be found at **http://biostatistics.mdanderson.org/Morris**

Functional Mixed Models: SELDI Example



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

MALDI Example: Block Effect