Statistical Issues in Using Mass Spectra for Disease Classification

Hongyu Zhao

Department of Epidemiology and Public Health Yale University School of Medicine

Joint Work with Baolin Wu, Weichuan Yu, Kenneth Williams, David Ward, and Yale Keck Laboratory

Outline

- Introduction
- Some statistical problems
  - Data Pre-Processing
  - Classification
  - Visualization
- Examples
- Conclusions

Introduction

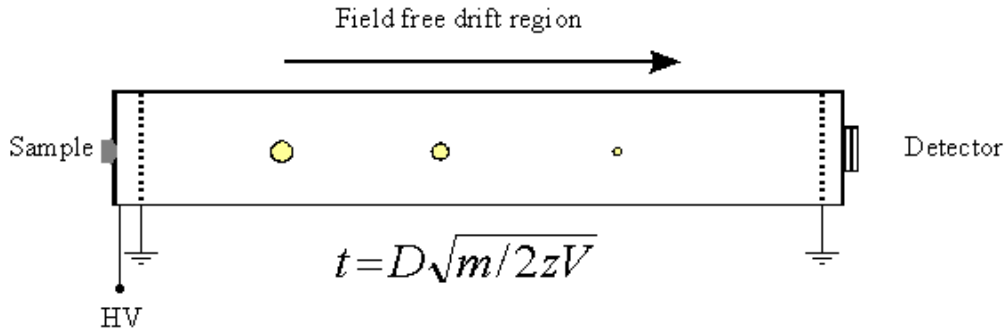
- Data Source:

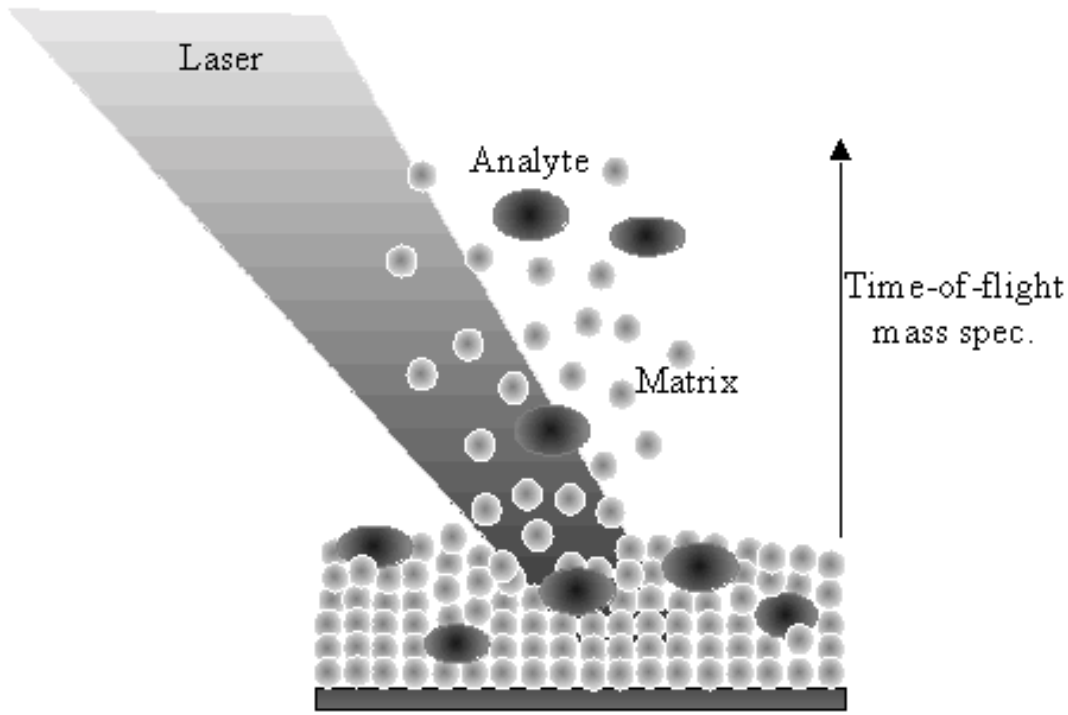
  Matrix-assisted laser desorption ionization (MALDI) Mass spectrometry (MS)

- Technology background

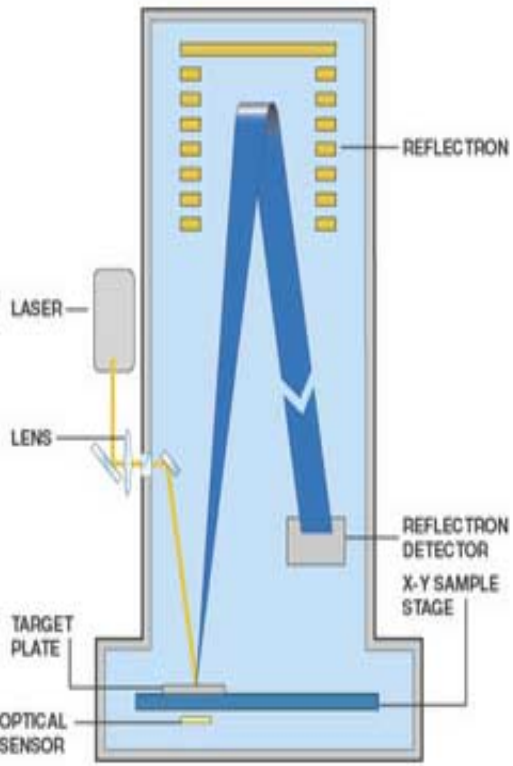
  MALDI-MS approach uses a nitrogen UV laser (337 nm) to generate ions from high mass, non-volatile samples such as peptides and proteins.

  Micromass' M@LDI™ and Q-Tof™ systems.

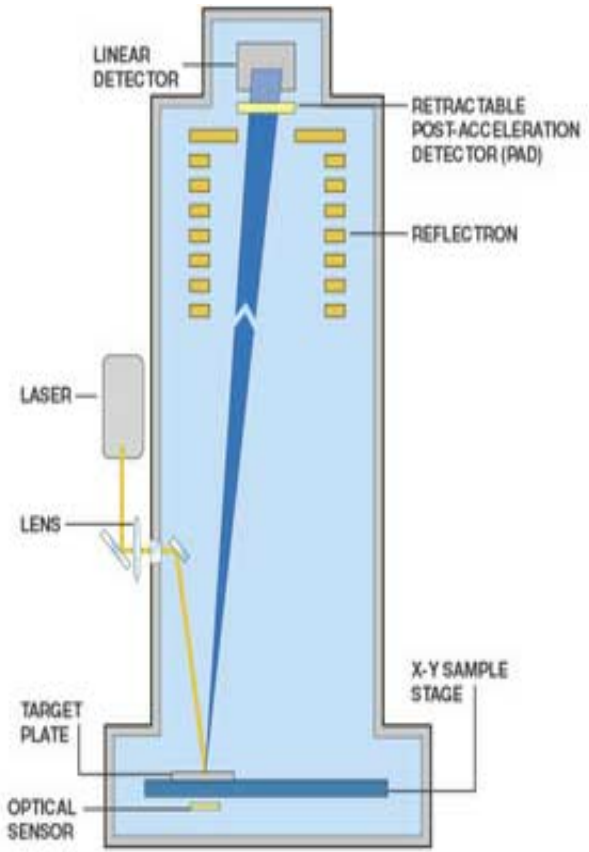

Field free drift region
Sample
Detector
HV
$t = D\sqrt{m/2zV}$


Laser
Analyte
Matrix
Time-of-flight
mass spec.

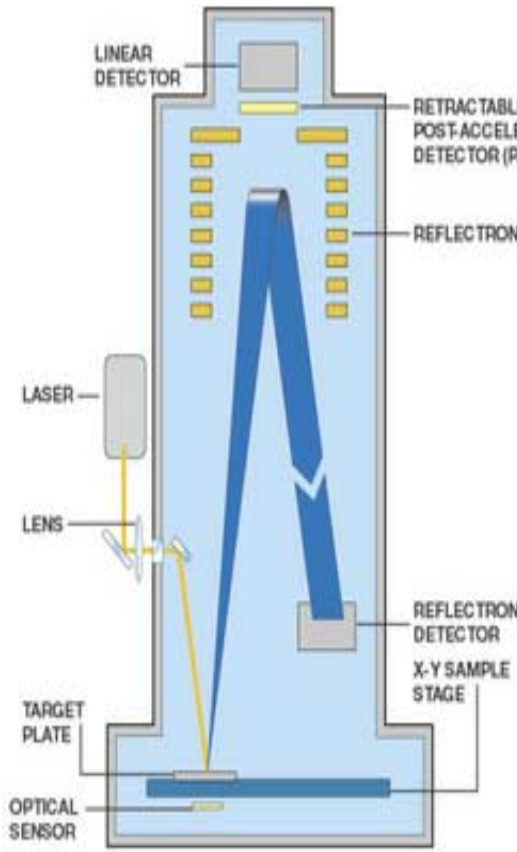
MS Instruments from Micromass



M@LDI *R*™



M@LDI *LR*™
Linear Mode



M@LDI *LR*™
Reflectron Mode

The M@LDI™ Family of 2D-Gel-MS Analyzers for The ProteomeWork™ System

Petricoin et al (2002) employed *Genetic Algorithms* and *Self-Organizing Maps* to analyze an MS dataset to distinguish ovarian cancer patients from normal individuals.

Baggerly et al. (2004): a careful reanalysis of Petricoin's data

Coombes et al. (2003):

Li et al (2002): LDA and bootstrap, breast cancer

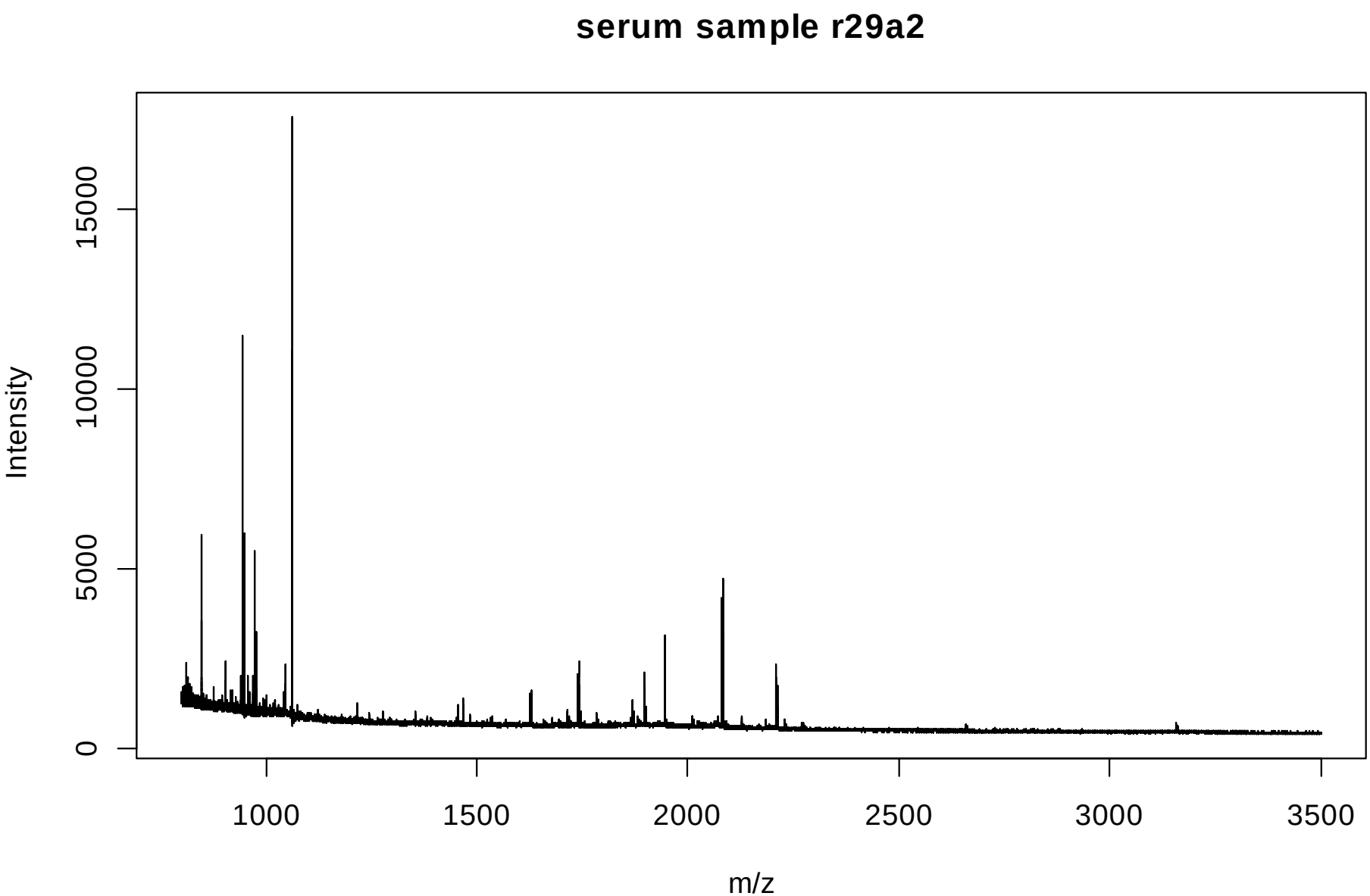
Adam et al. (2003), Qu et al. (2002), Yasui et al. (2003): Boosting, prostate cancer

Wu et al. (2003): Random Forest, ovarian cancer

Tibshirani et al. (2004): Peak probability contrast, nearest shrunken centroids, ovarian cancer

Many other studies

A typical plot for one serum sample



Data Characteristics

- **Simple Format**

  Paired intensity versus mass/charge datapoints, because MALDI-MS almost exclusively produces singly charged species, the mass/charge ratio is usually equal to the mass.

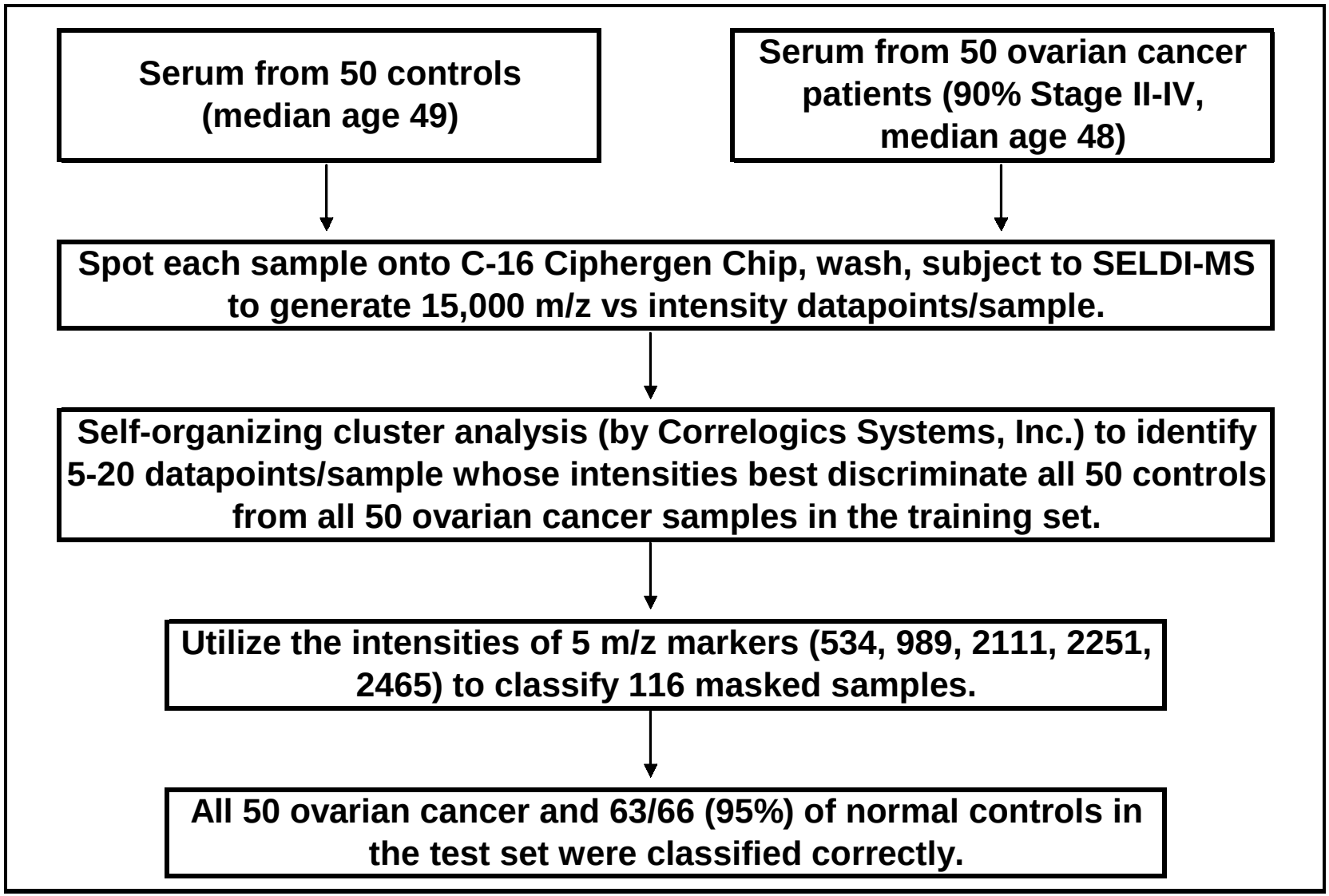
- **Goal**

  Find potential peptide/protein markers to distinguish cases from controls and to enable classification of future samples.

- **Difficulty**

  Huge number (~100,000) of features in a given dataset, comparatively small number (~100) of samples, and noisy background.

Petricoin et al Approach to Disease Biomarker Discovery in Serum



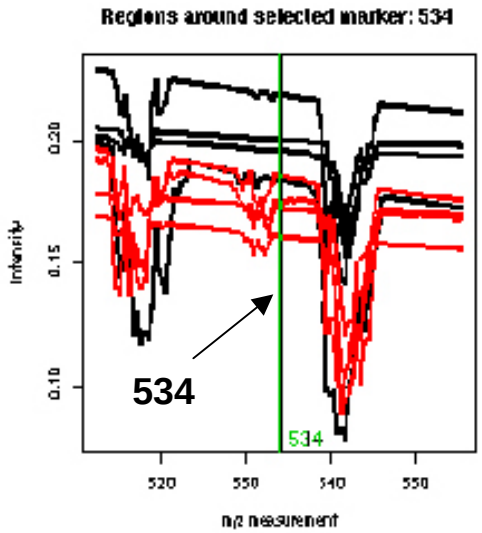
[1]From Petricoin et al, The Lancet 359, 572-577 (2002).

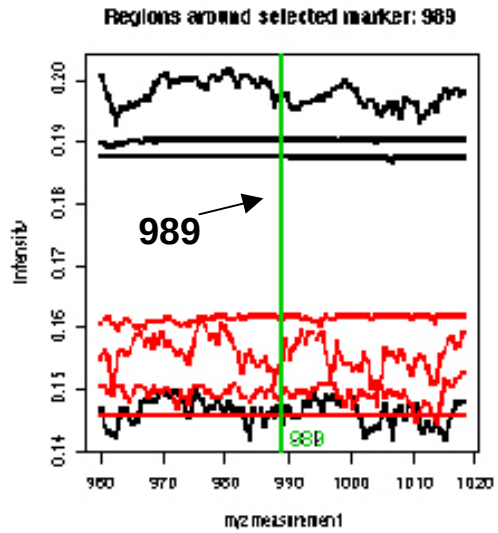
SELDI-MS Spectra for 4 Ovarian Cancer and Control Samples Around 5 Markers Identified by Petricoin (2002)

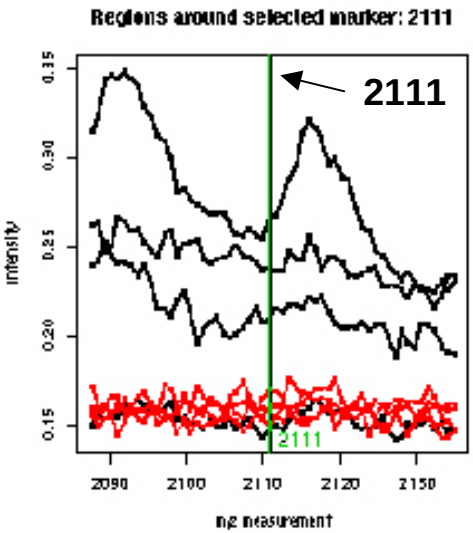
Ovarian Cancer # 1, 2, 3, 4

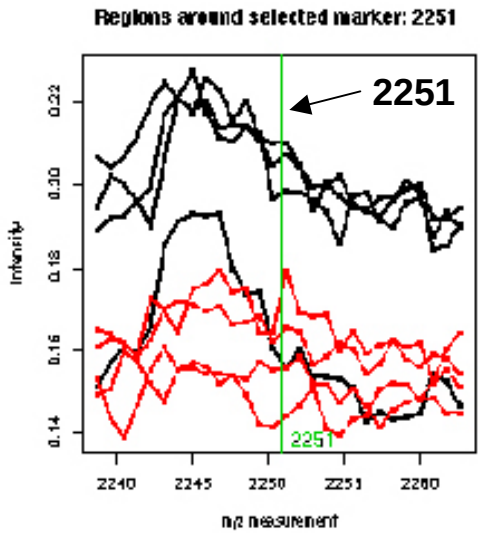
Controls #51, 52, 53, 54

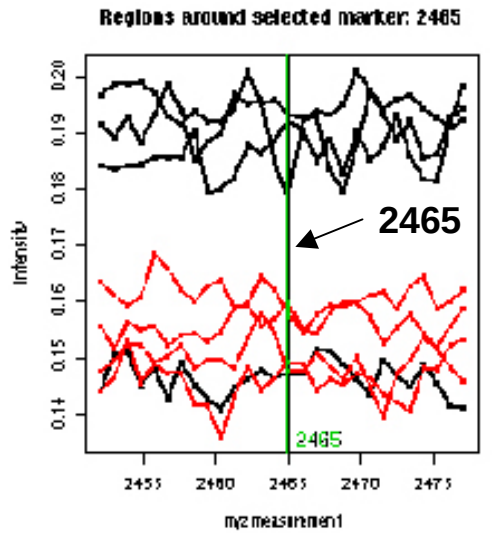
Spectra downloaded from NIH Clinical Proteomics Program Data Bank: http://clinicalproteomics.steem.com/



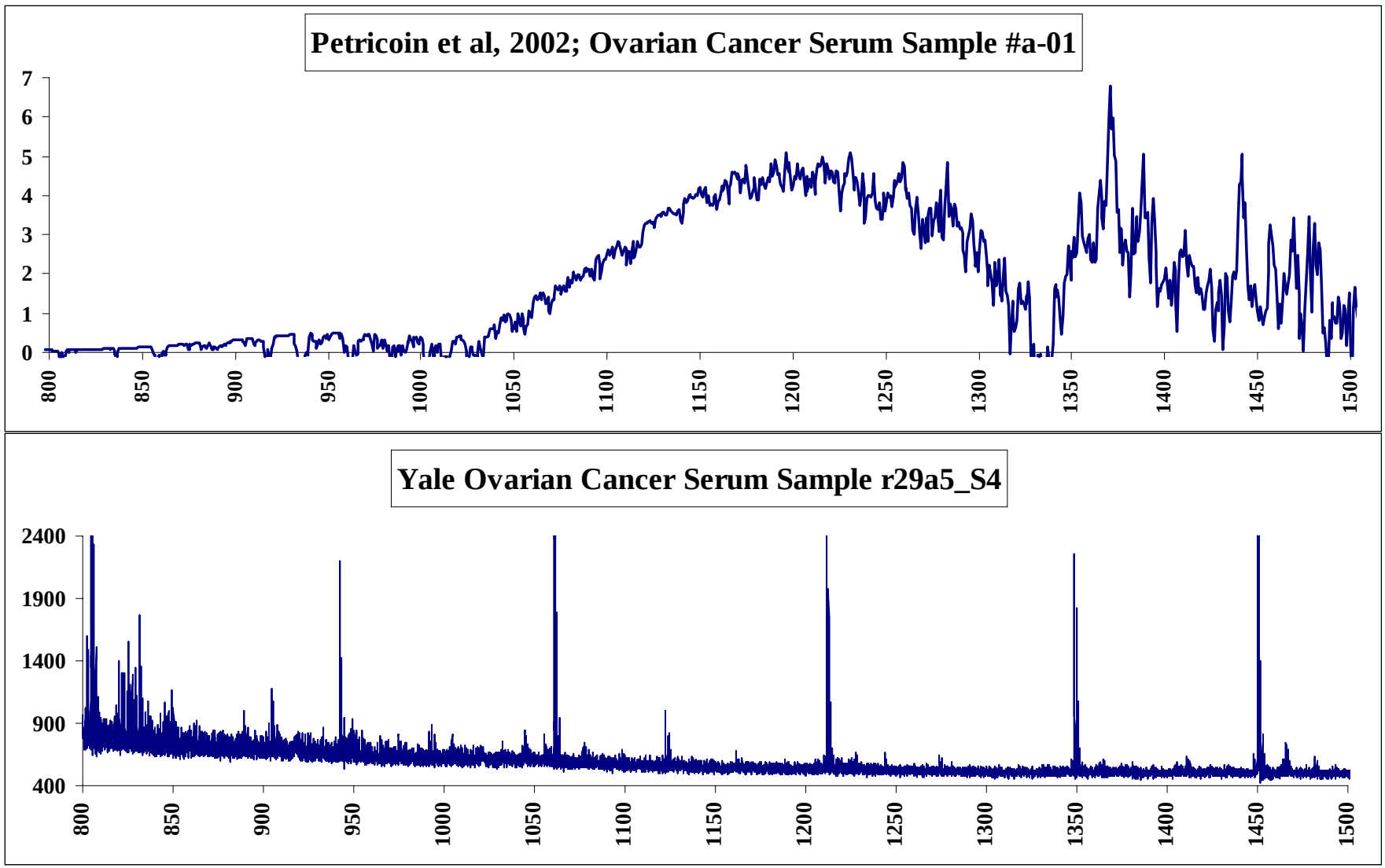




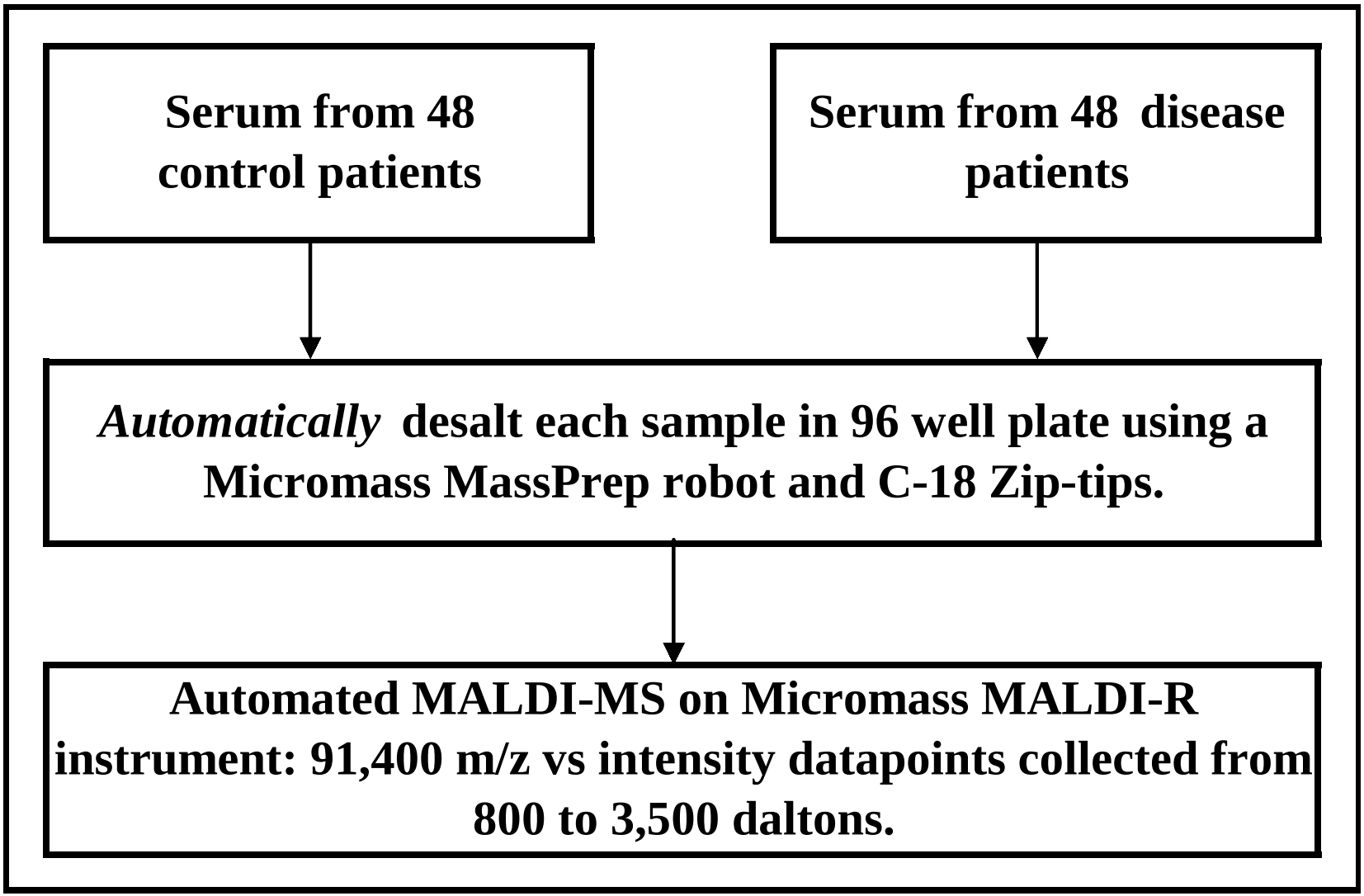




MALDI-MS From Ovarian Cancer Patient Serum Samples Obtained on a Ciphergen Protein Systems 2 (Top) and Micromass M@LDI-R Instrument (Bottom)

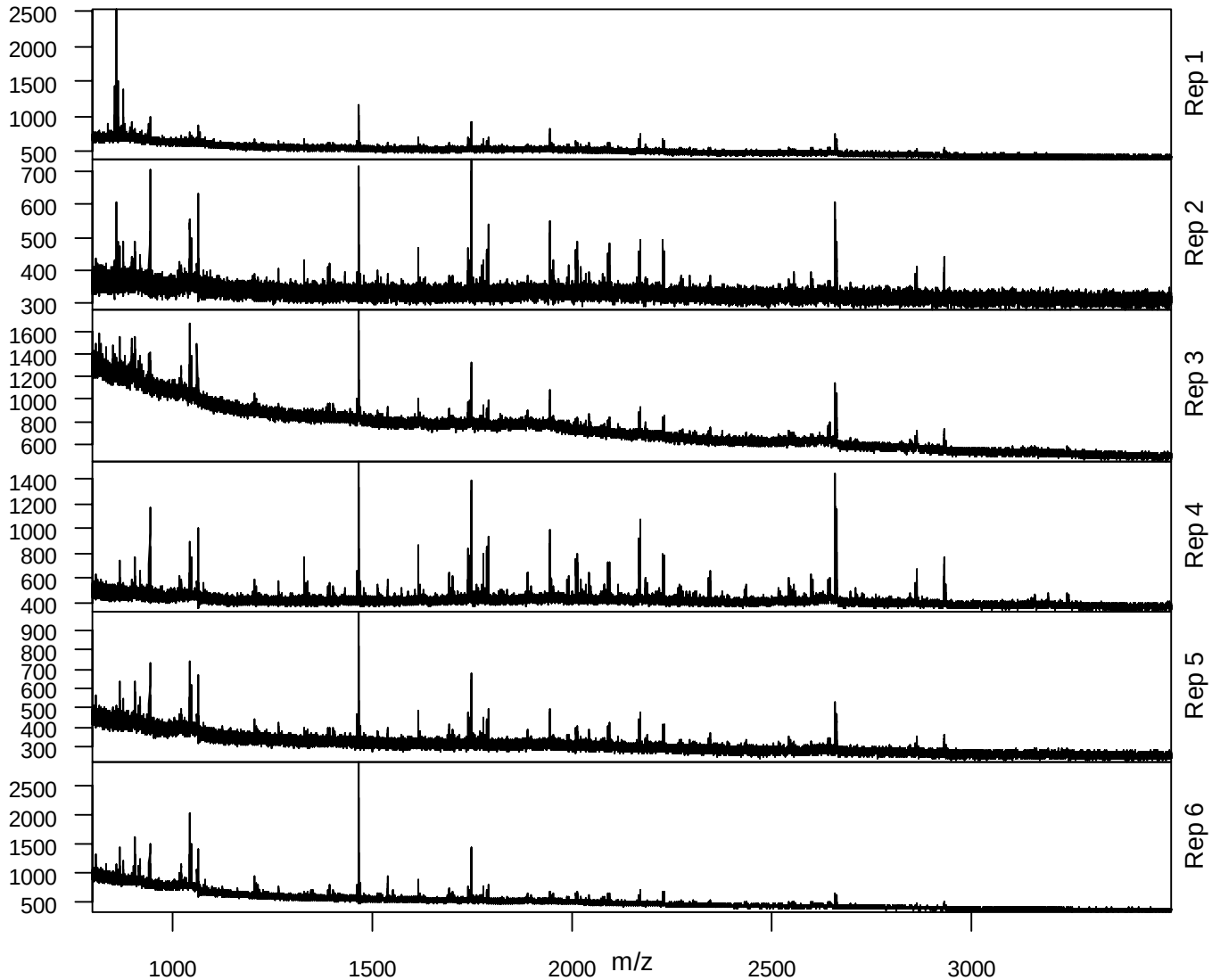


MALDI-MS Disease Biomarker Analysis of Serum



Data Reproducibility

Technical



Some Challenges in the Analyses

- Mass alignment
- Background subtraction
- Peak identification
- Normalization
- Classification

Challenges -- continued

- Visualization methods for this type of dataset

  We need an effective visualization method for MS dataset. Also this can serve as a reference to compare results from our algorithms.

- Software implementation

Data Pre-Processing

- **Mass alignment**

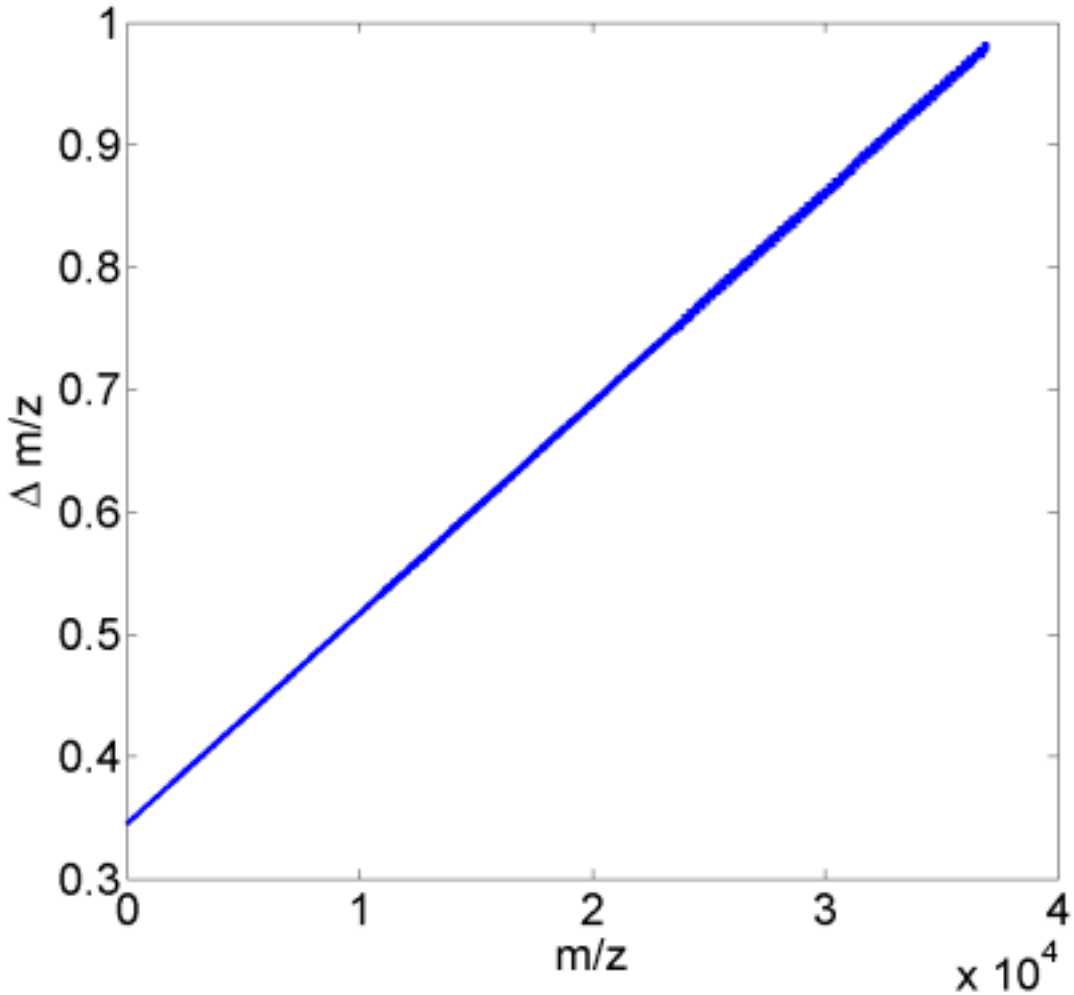
  Manually add Bradykinin to all samples so that its mass measurement serves as a common reference point for our mass alignment program.

- **Variable transformation to reduce variation**

  Log transformation to reduce the magnitude and variation of the intensities.

Why Take a log on the m/z Axis?

△m/z: a linear function of m/z index



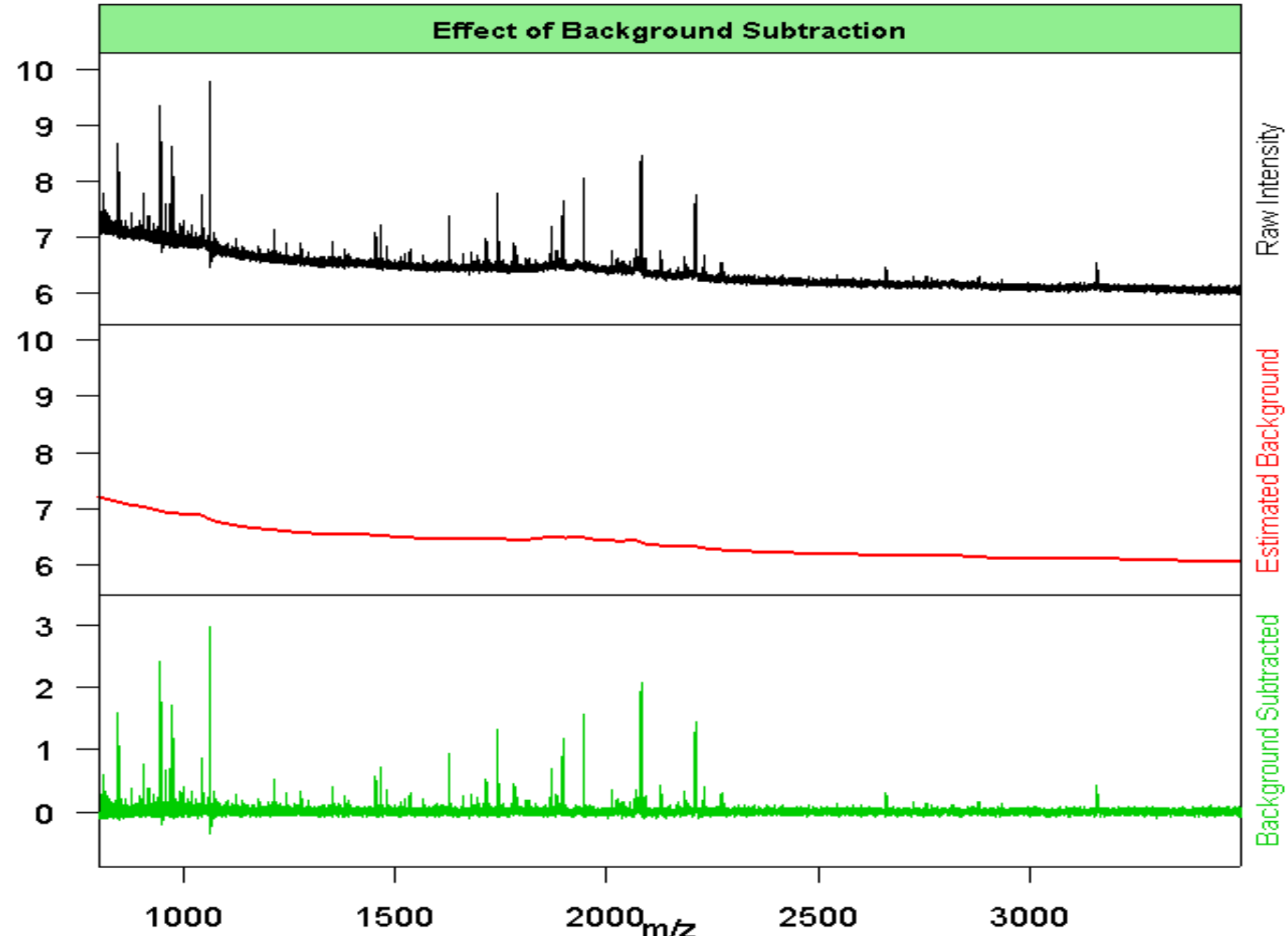
Data Pre-Processing

- **Background Subtraction**

  Chemical and electronic noise produces a background intensity which typically decreases with increasing mass.

  We can estimate a background intensity level to allow removal of this trend.

  The basic technique we use now is local robust polynomial regression.

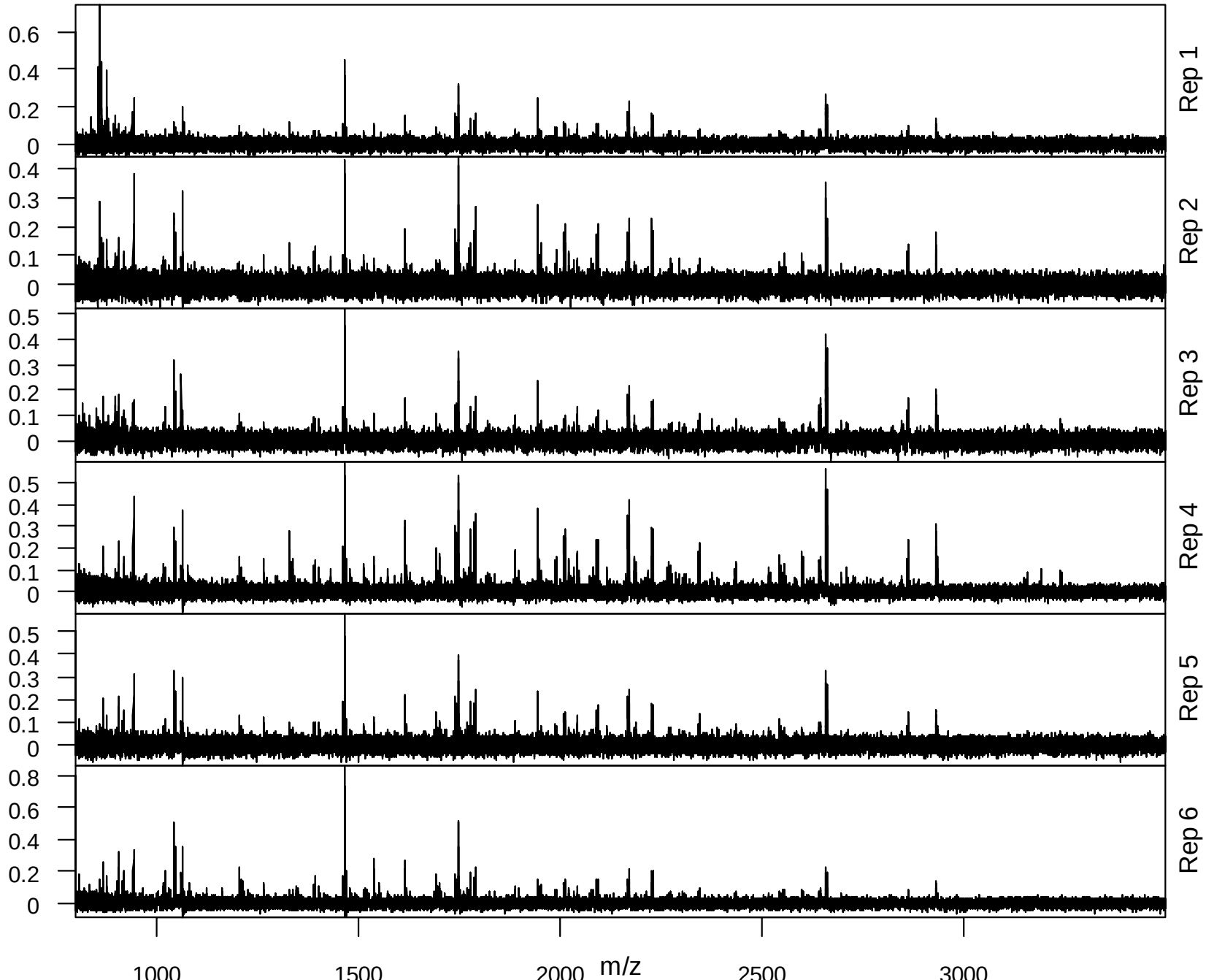

Effect of Background Subtraction
10
9
8
7
6
Raw Intensity
10
9
8
7
6
Estimated Background
3
2
1
0
Background Subtracted
1000
1500
2000
m/z
2500
3000

Data Pre-Processing

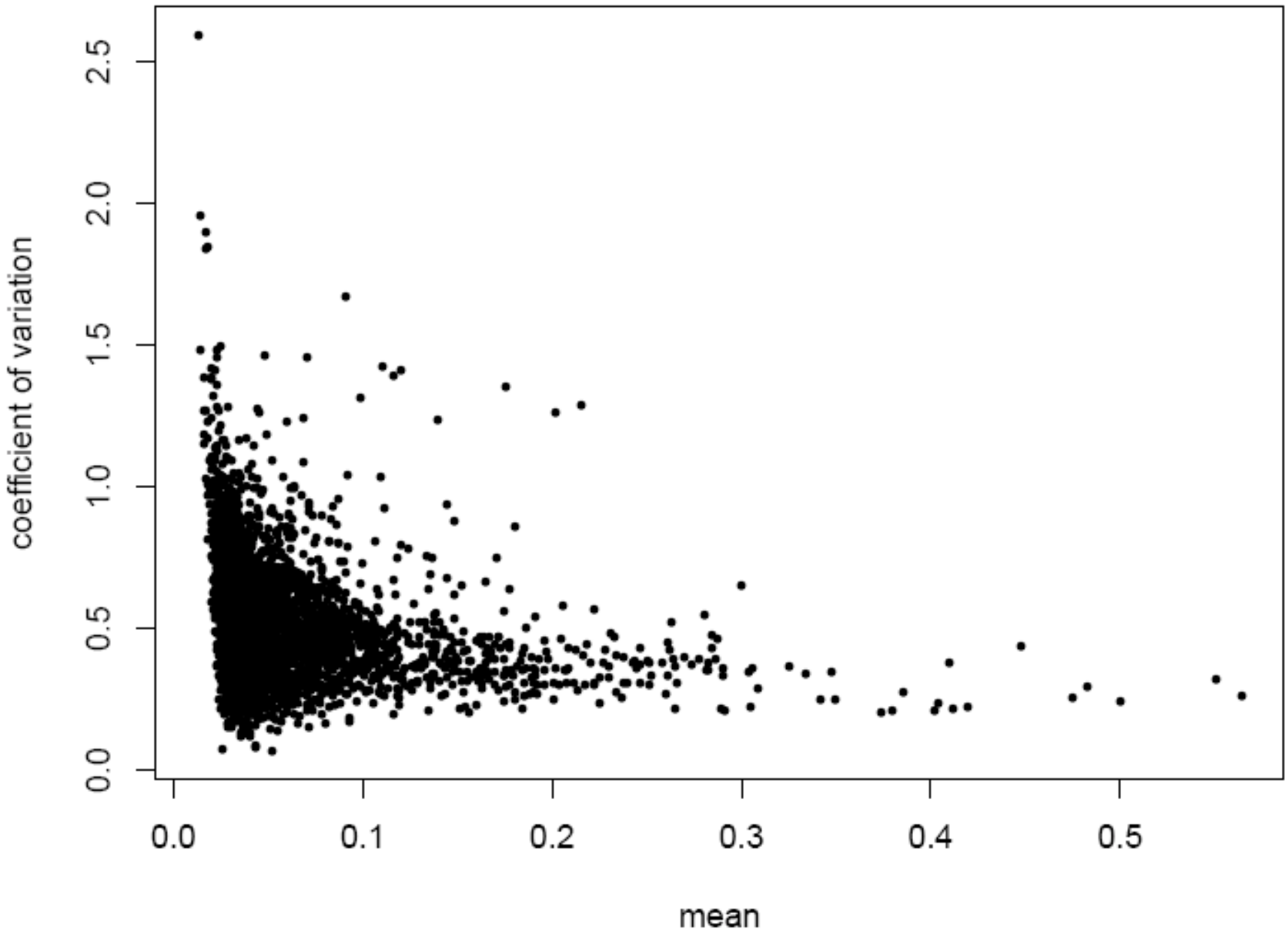
- **Normalization**

Micromass' M@LDI™ systems robotically take 40 shots in the sample with the final reported intensity being the sum of these individual spectra.

Due to variation in sample preparation and deposition on the target, matrix crystallization, and ion detection, samples are not directly comparable before normalization.


Rep 1
Rep 2
Rep 3
Rep 4
Rep 5
Rep 6
m/z
1000
1500
2000
2500
3000

	A2	A3	A4	A5	A6
A1	0.443	0.431	0.539	0.415	0.458
A2		0.445	0.615	0.481	0.541
A3			0.668	0.544	0.628
A4				0.668	0.690
A5					0.629


coefficient of variation
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.1
0.2
0.3
0.4
0.5
mean

Data Pre-Processing

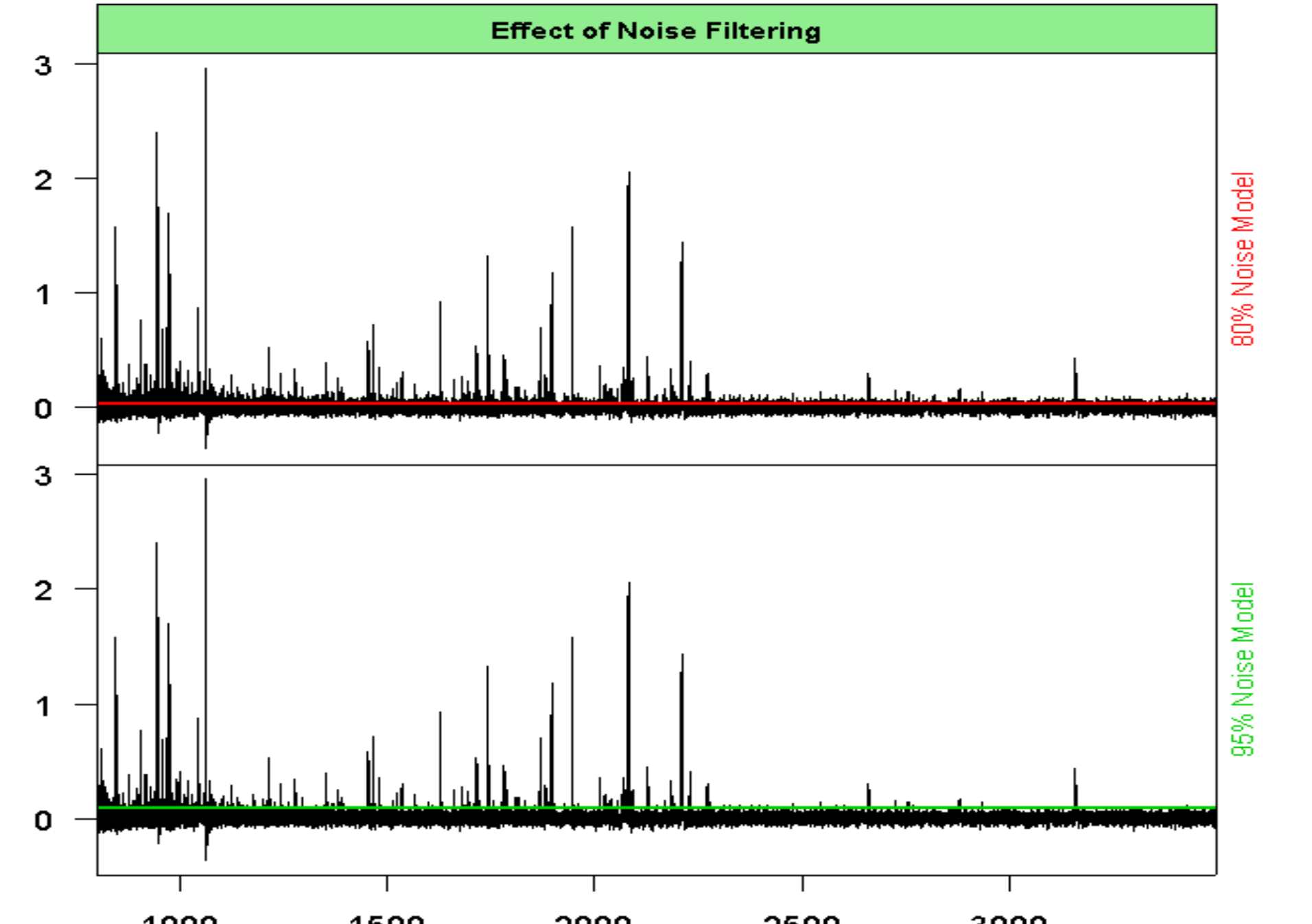
- **Peak Identification**

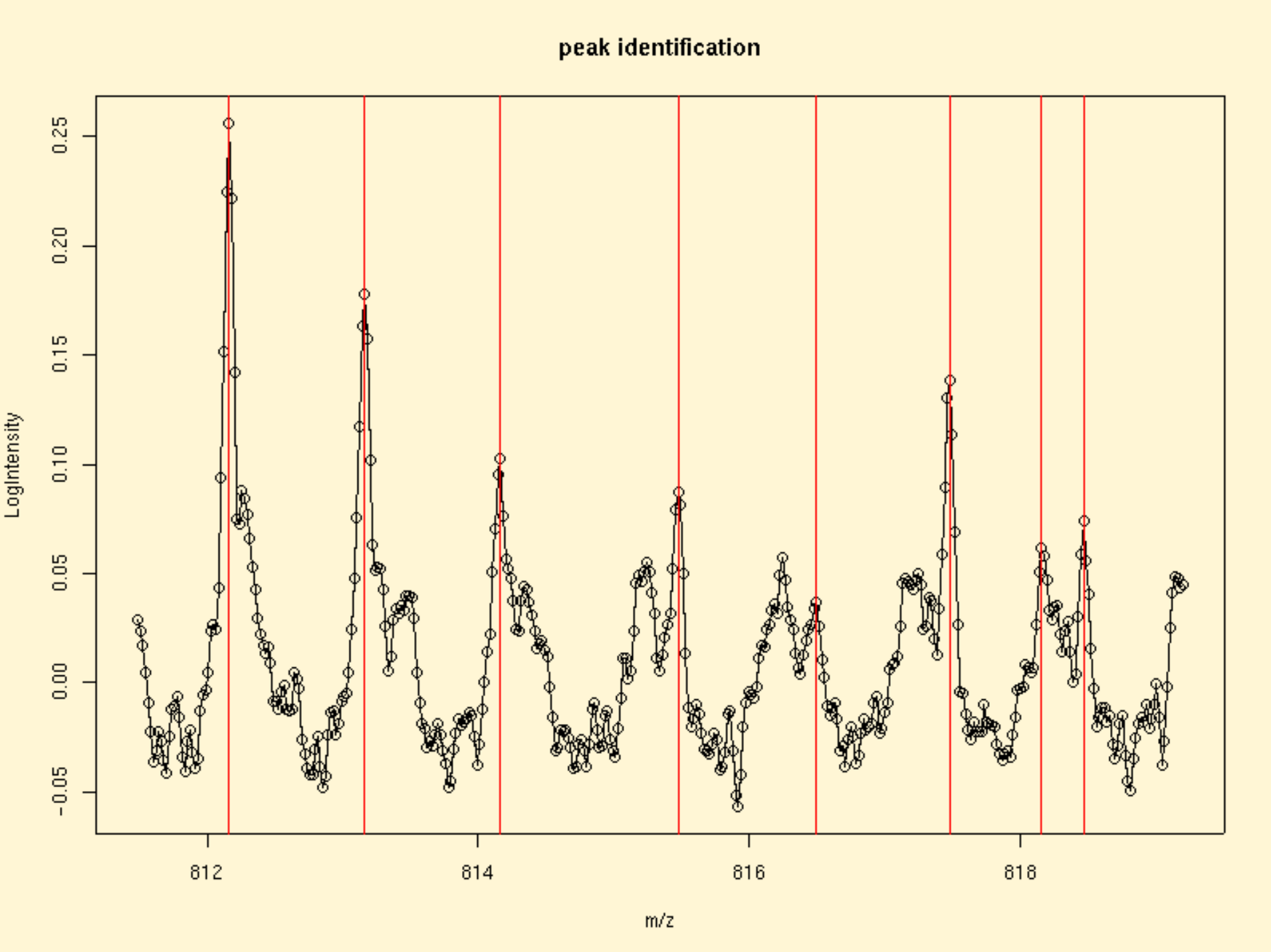
Yasui et al. Biostatistics 2003:

local maximum search, broad neighborhood control, plus thresholding, binary indicator for peak and non-peak

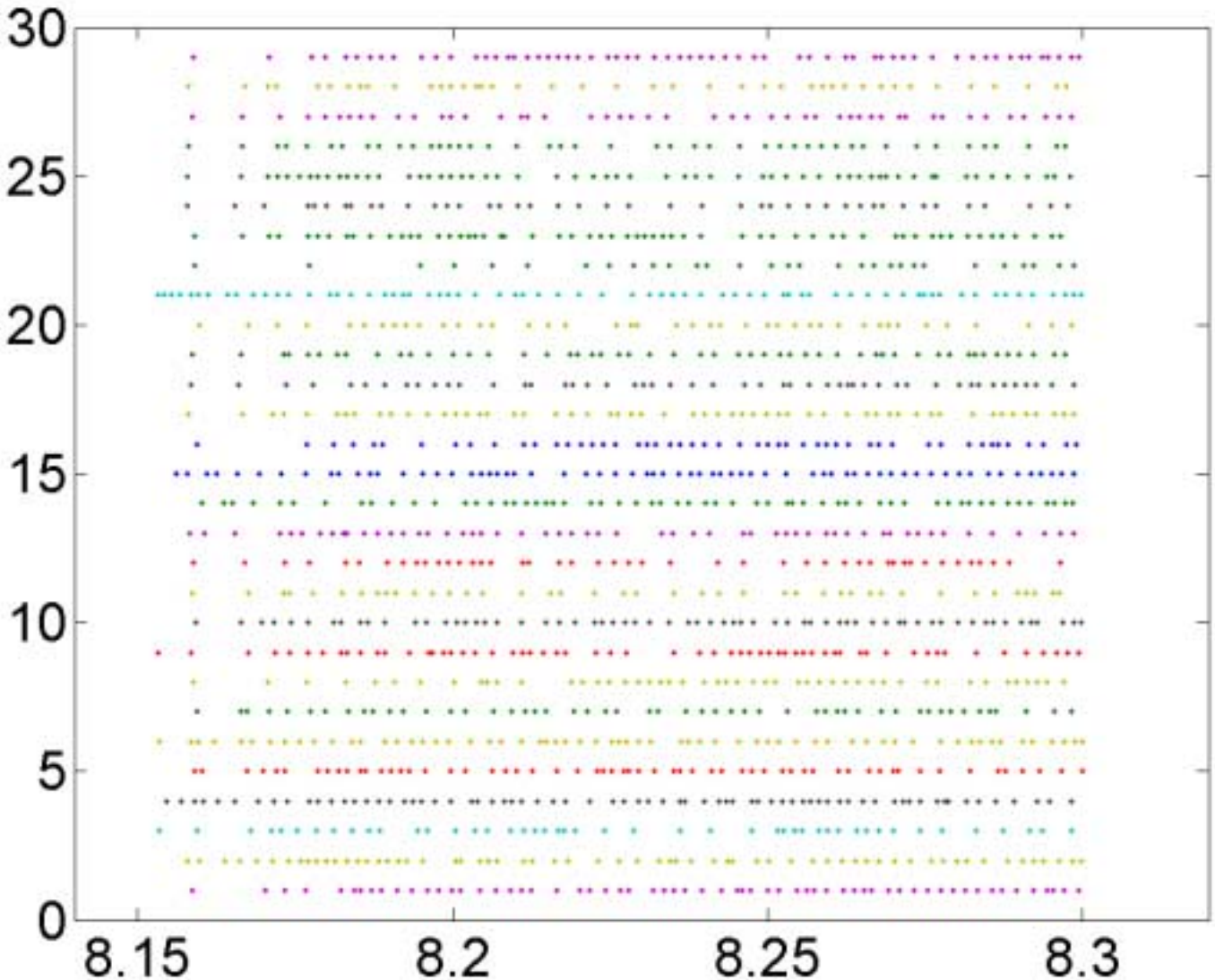
Coombes et al. Clinical Chemistry 2003:

noise estimation plus thresholding


Effect of Noise Filtering
3
2
1
0
80% Noise Model
3
2
1
0
95% Noise Model


peak identification
LogIntensity
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
812
814
816
818
m/z

Peaks before Alignment: Example



Yasui et al. Biostatistics 2003

Wagner et al. Proteomics 2003

Coombes et al. Clinical Chemistry 2003

Tibshirani et al. Unpublished manuscript 2004

Problem Formulation

Given two sets of peak points, find out a unique correspondence between them which minimizes a distance function.

Not all peaks appear in every data set:

One-to-one correspondence does not exist between every two data sets

- Basic fact: peak variation much smaller than distance of different peaks
- Construct a super set of all peaks and use it as the anchor of alignment --- every data set is aligned to this super set.
- Computational cost → down to M alignments

- Based on a distance threshold parameter
- Initialize the super set as the first data set
- For each point in the new candidate set, check its closest distance to the super set

  --- over the threshold:

  new peak, add to the super set

  --- below the threshold:

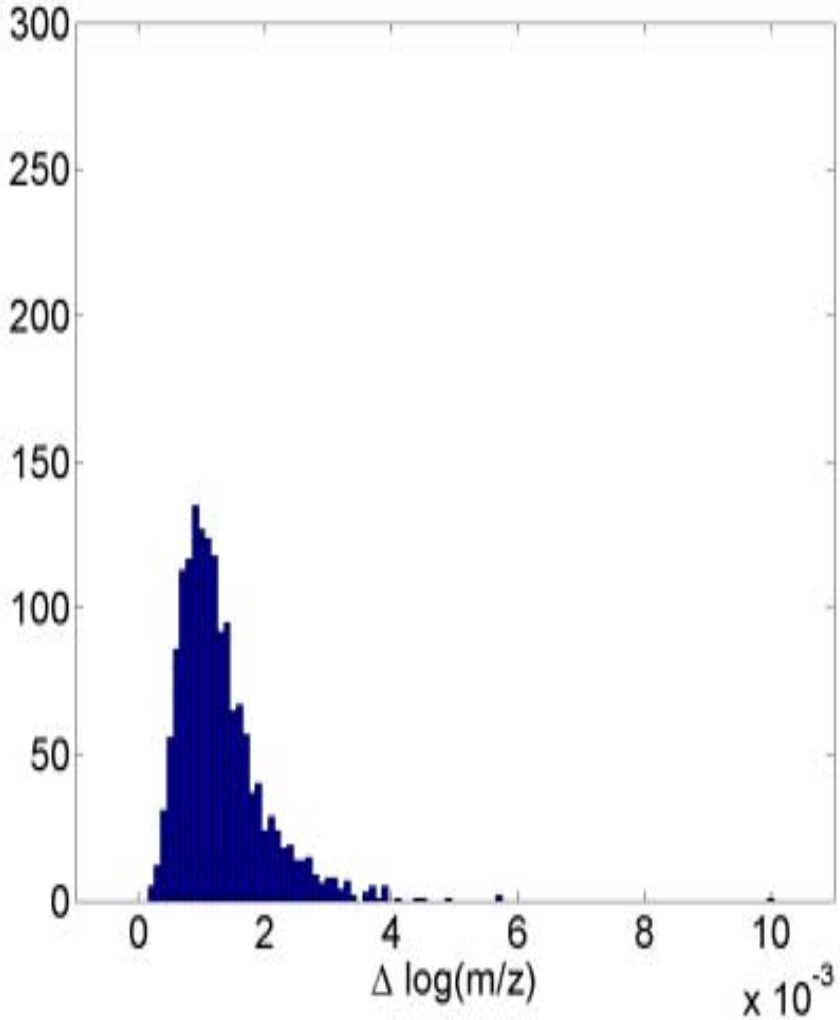
  already in the super set, ignore
- Continue till all data sets are processed

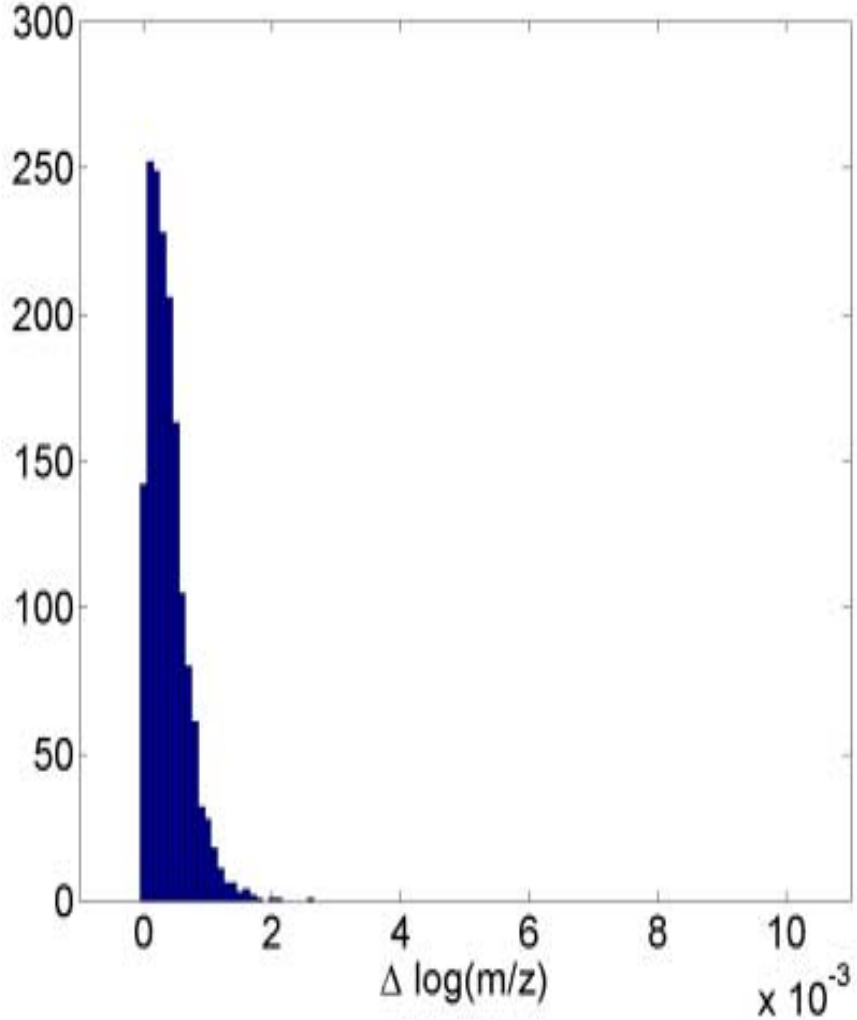
Does this distance threshold exist ?

Yes, e.g. 0.5 x mean of the neighboring peak distances

→ the statistics of the histogram

Peak Distance v.s. Peak Variation: mean of distance = 4 x mean of variation





Super Set Based Alignment Using Closed Point Matching

- For each point in the candidate data set, align it to its counterpart in the super set by using the closest distance criterion
- The existence of the counterparts in the super set is guaranteed.
- The framework is non-rigid point matching

Classification

1. Traditional approaches

   Dimension reduction, very hard to interpret the results.

2. Algorithm approaches

   CART (classification and regression trees), too much variation.

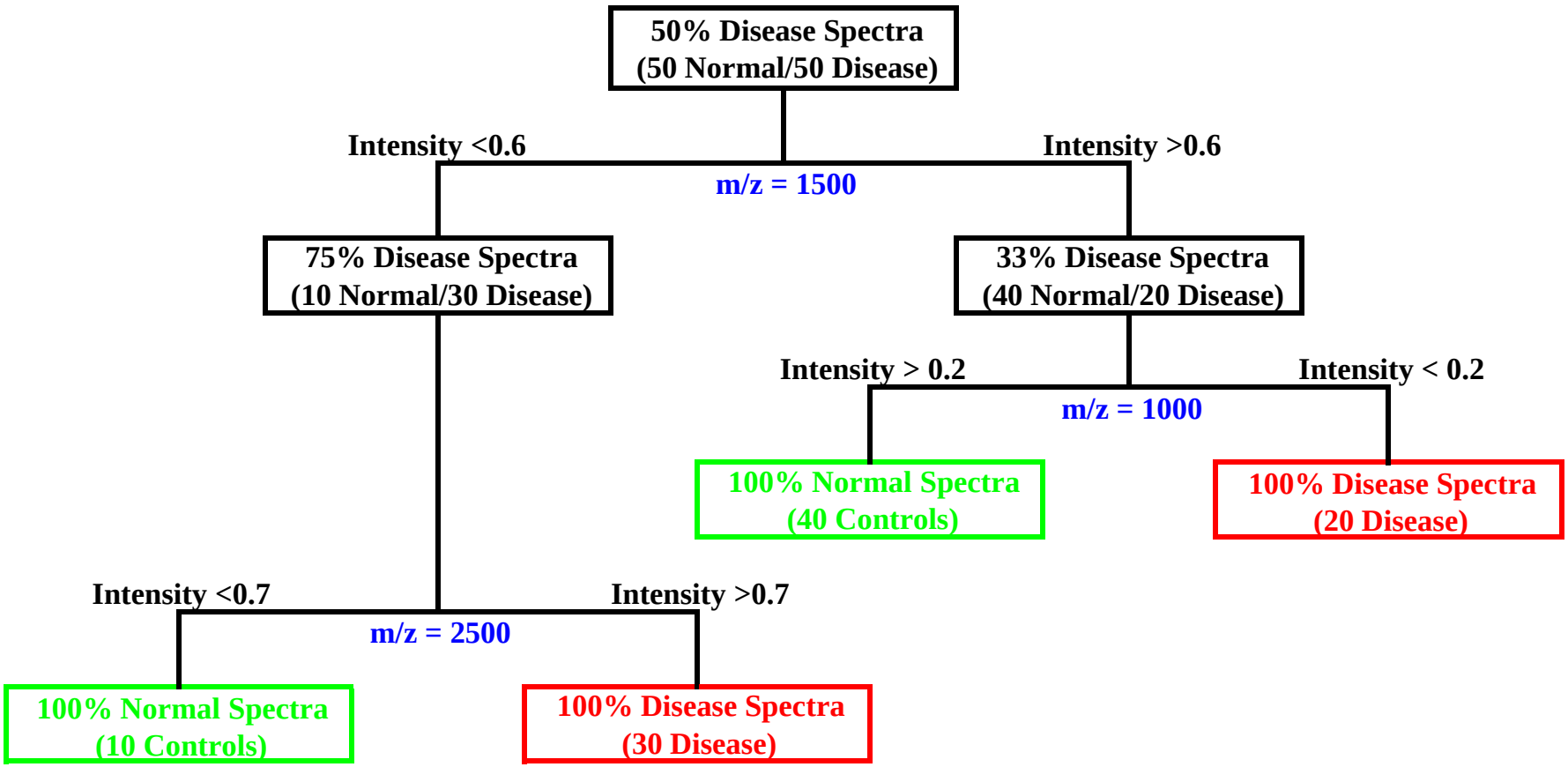
   Bagging, arcing (a.k.a. boosting), SVM, randomForest

LDA, QDA

NN-k

Support Vector Machine

Classification Trees



Bagging:

Sample with replacement to form N bootstrap samples .
Use to construct tree classifier, and predict using this classifier.
Final prediction is majority vote.

Boosting:

Arc-fs:

1) At first step, initialize $p_i^{(1)} = 1/n$
2) At the k-th step, using the current probabilities $p_i^{(k)}$, sample with replacement from sample S to get the training set S_k and construct tree classifier T_k using S_k .
3) Run S down T_k and let $d(i) = 1$ if i-th case is classified incorrectly, otherwise zero.
4) Define $\varepsilon_k = \sum_i p_i^{(k)} d(i), \beta_k = (1-\varepsilon_k)/\varepsilon_k$, and updated (k+1)st step probabilities by $p_i^{(k+1)} = p_i^{(k)} \beta_k^{d(i)} / \sum_j p_j^{(k)} \beta_k^{d(j)}$. If $\varepsilon_k = 0, \varepsilon_k \geq \frac{1}{2}$, we re-initialize all $p_i^{(k+1)} = 1/n$.
5) After K steps, the $\{T_1, ..., T_K\}$ are combined using weighted voting with T_k having weight $\log(\beta_k)$.

RandomForest

(Leo Breiman)

Combines the following two ideas:

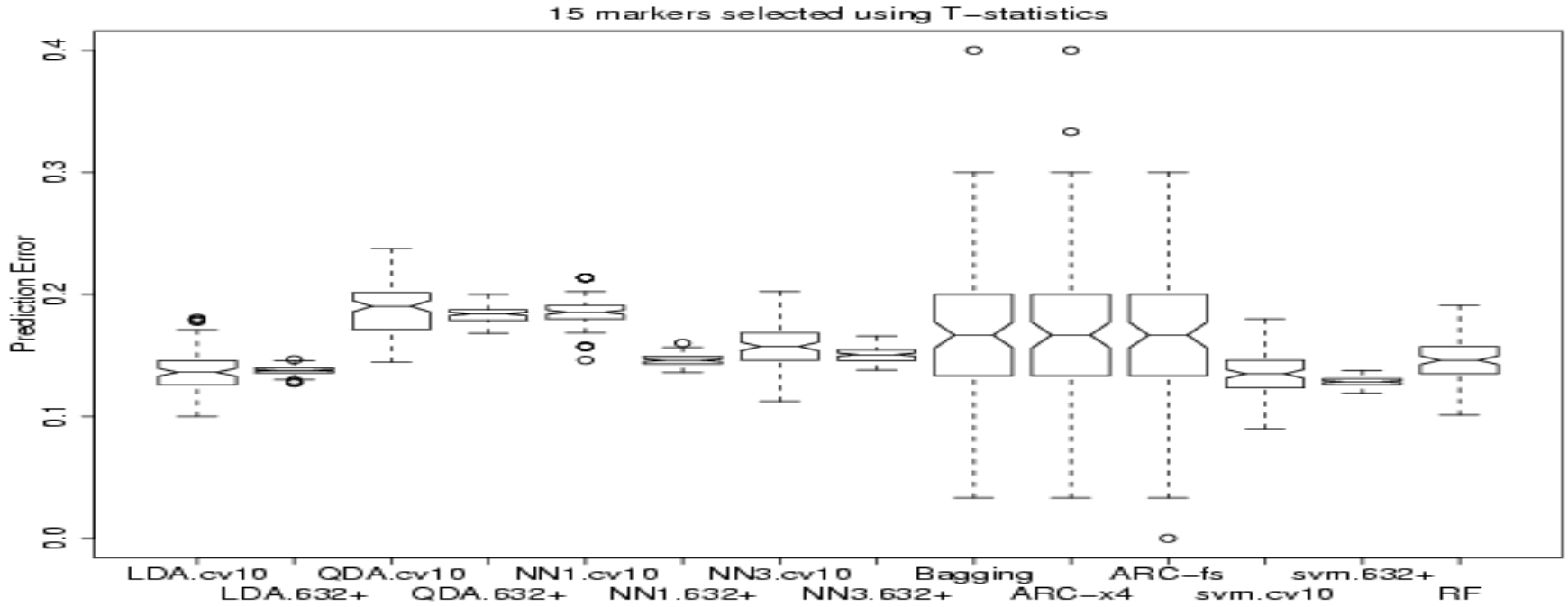
- Bagging pools of multiple classifiers from perturbed versions of the original dataset to increase predictive accuracy.
- Random feature selection from several best ones increases predictive accuracy.

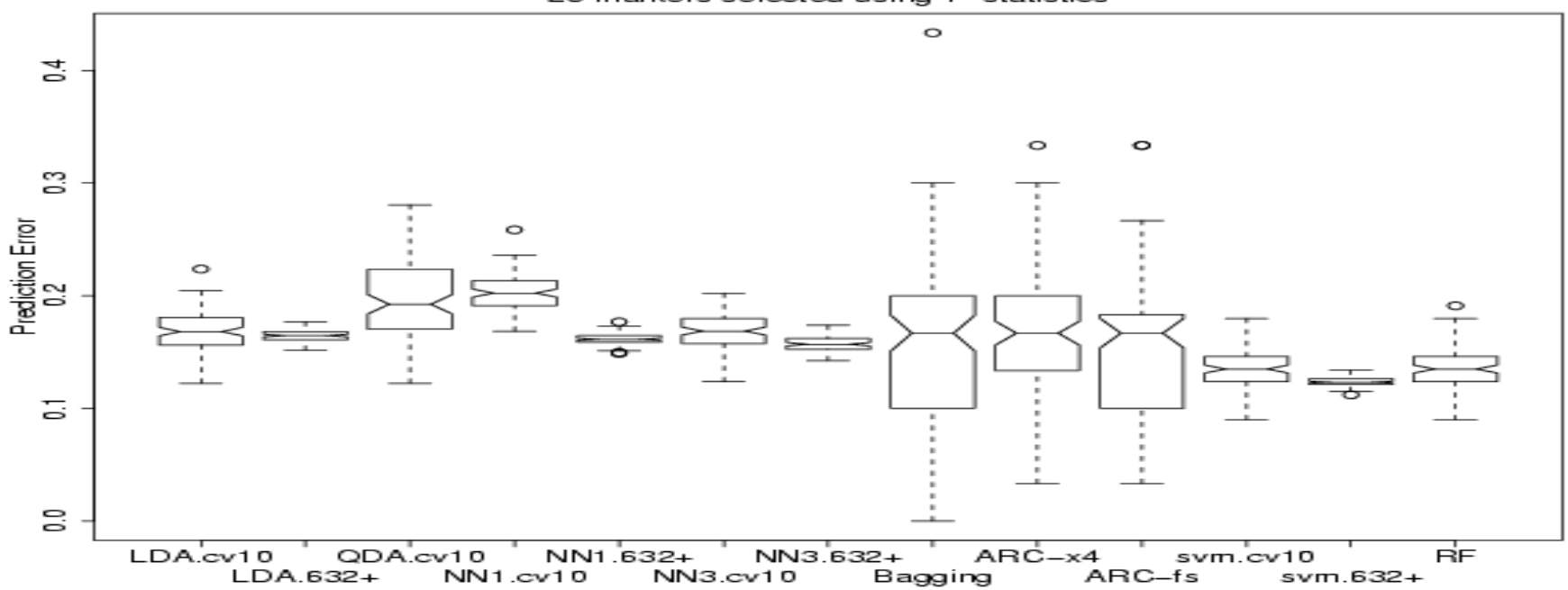
1) Sample with replacement to form N bootstrap samples.

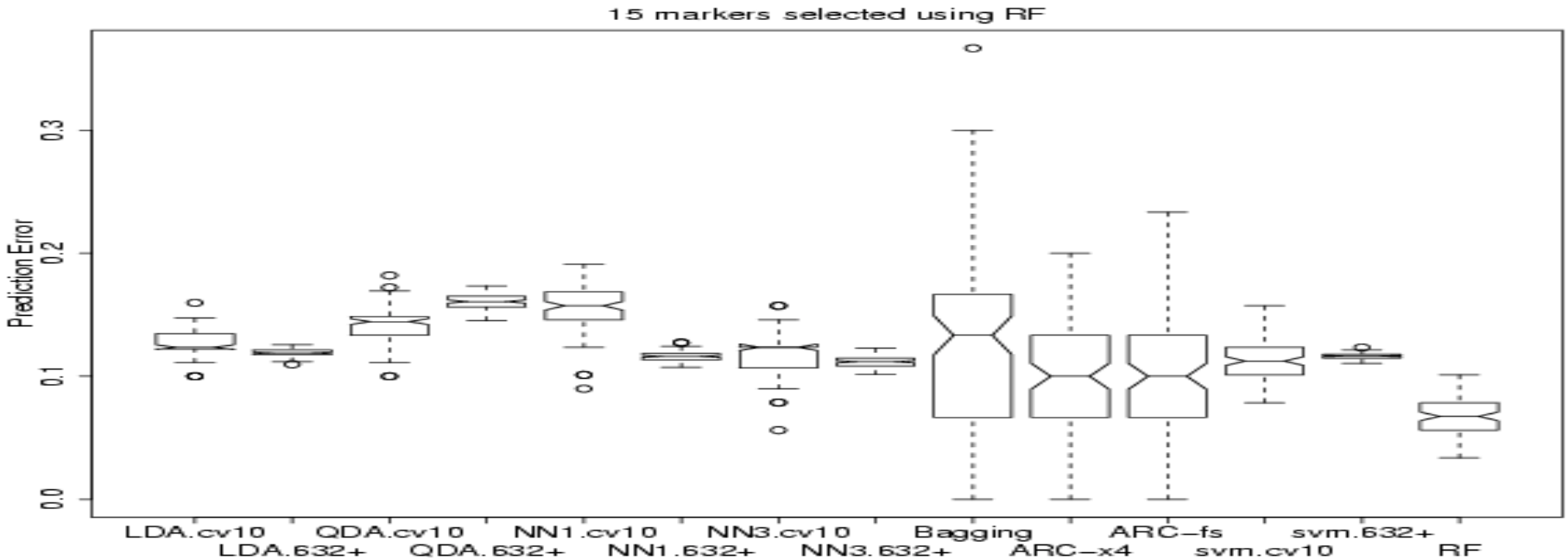
2) Use each sample to construct a Tree classifier to predict those samples that are not in the sample (called out-of-bag samples). When constructing each tree, at each node splitting, we first randomly select m variables and then we choose a best split from these m variables. These predictions are called out-of-bag estimators.

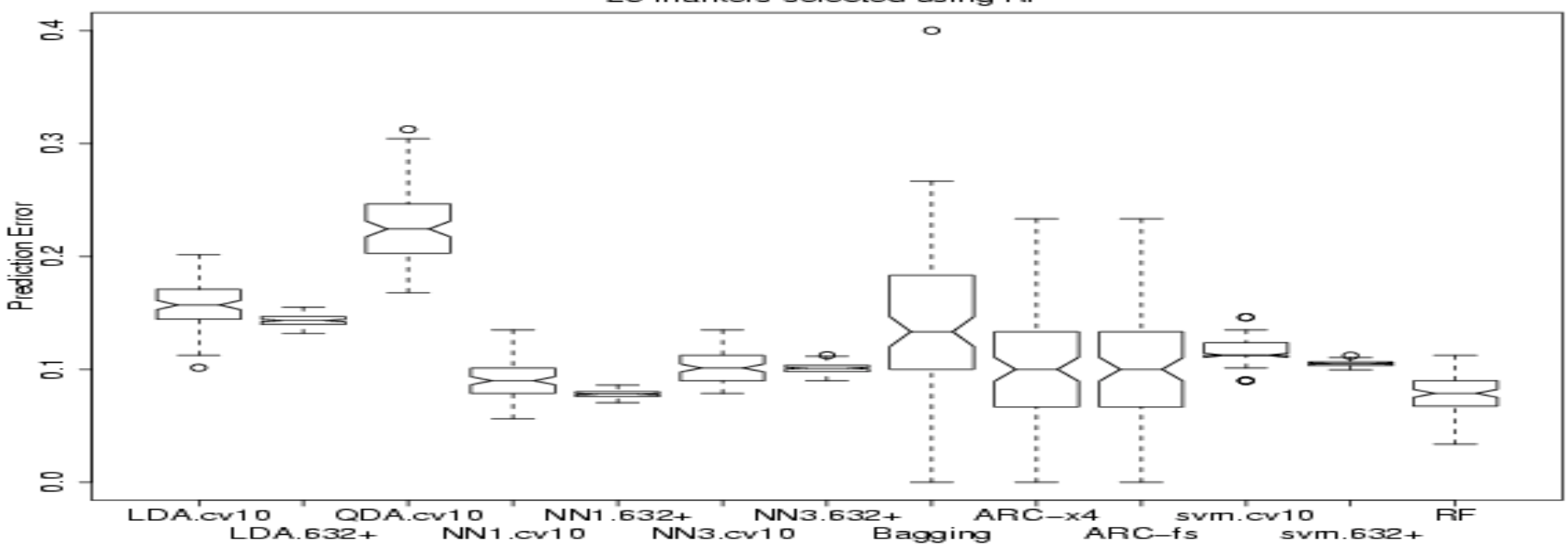
3) Final prediction is the average of out-of-bag estimators over all Bootstrap samples.

4) Before using each tree to predict out-of-bag samples, if we randomly permute the value for one variable for these out-of-bag samples, intuitively the prediction error is going to increase. And the amount of increase will reflect the importance of this variable.

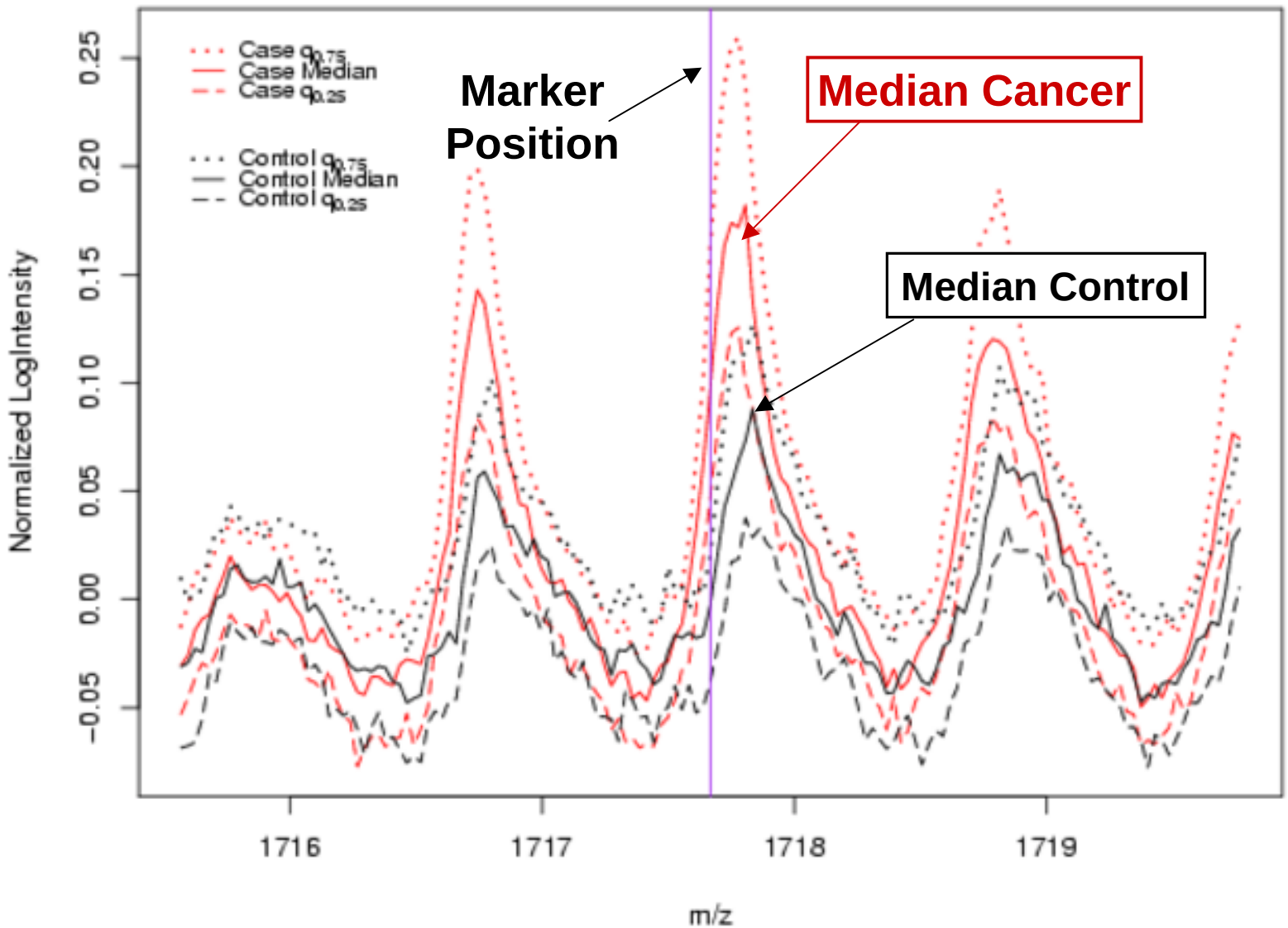

15 markers selected using T−statistics
Prediction Error
0.0
0.1
0.2
0.3
0.4
LDA.cv10
LDA.632+
QDA.cv10
QDA.632+
NN1.cv10
NN1.632+
NN3.cv10
NN3.632+
Bagging
ARC−x4
ARC−fs
svm.cv10
svm.632+
RF


25 markers selected using T−statistics
Prediction Error
0.0
0.1
0.2
0.3
0.4
LDA.cv10
LDA.632+
QDA.cv10
NN1.cv10
NN1.632+
NN3.cv10
NN3.632+
Bagging
ARC−x4
ARC−fs
svm.cv10
svm.632+
RF

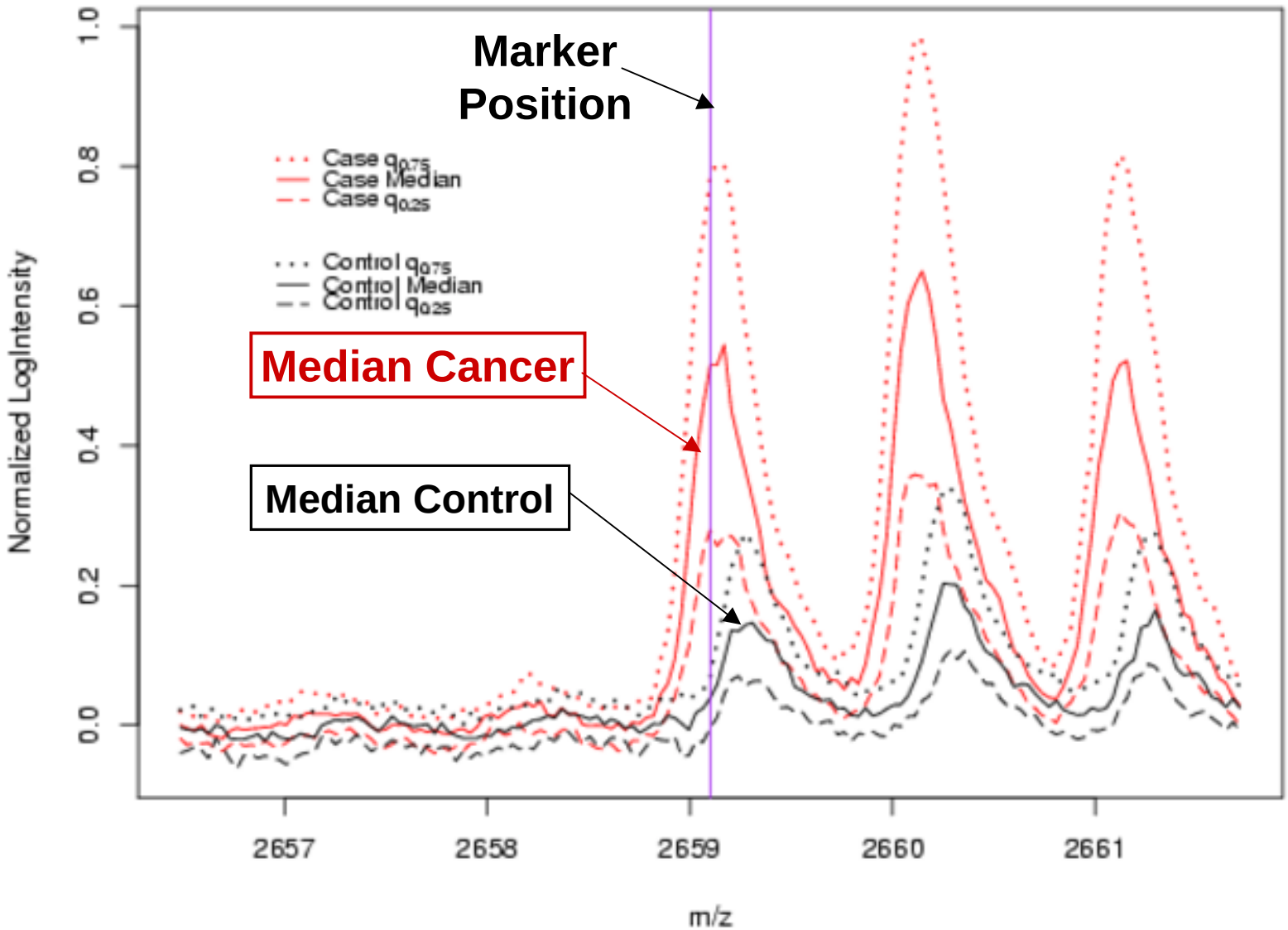

15 markers selected using RF
Prediction Error
0.0
0.1
0.2
0.3
LDA.cv10
LDA.632+
QDA.cv10
QDA.632+
NN1.cv10
NN1.632+
NN3.cv10
NN3.632+
Bagging
ARC−x4
ARC−fs
svm.cv10
svm.632+
RF

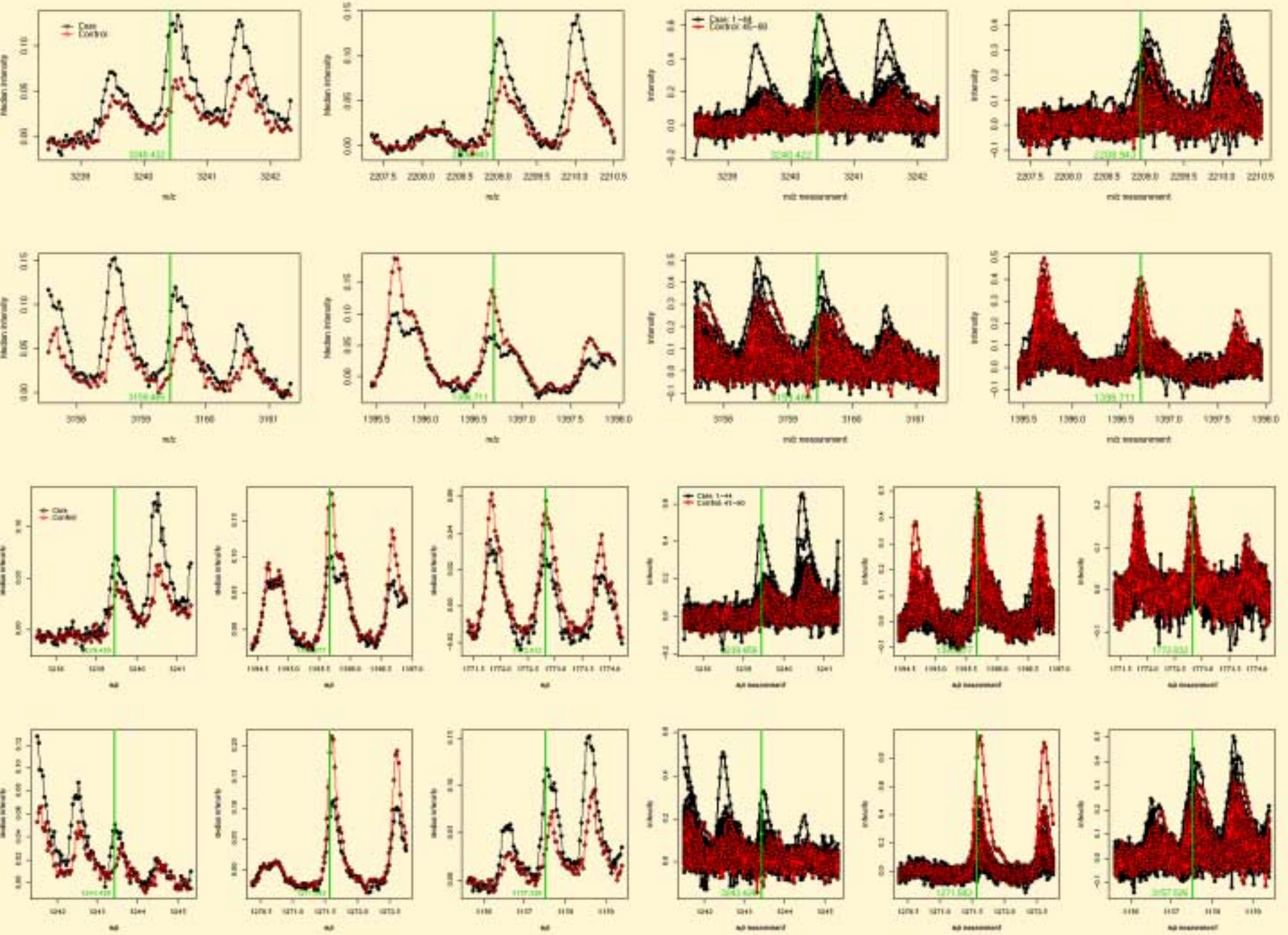

25 markers selected using RF
Prediction Error
0.0
0.1
0.2
0.3
0.4
LDA.cv10
LDA.632+
QDA.cv10
NN1.cv10
NN1.632+
NN3.cv10
NN3.632+
Bagging
ARC−x4
ARC−fs
svm.cv10
svm.632+
RF

m/z Versus Intensity Distribution for 129 Samples Around the 1718 Ovarian Cancer Marker (relative marker importance = 0.6)



m/z Versus Intensity Distribution for 129 Samples Around the 2659 Ovarian Cancer Marker (relative marker importance = 6.0)




Case
Control
Median intensity
m/z
Case: 1–44
Control: 45–60
Intensity
m/z measurement

How to appropriately assess prediction errors?

(1) Features need to be selected based on each training Sample, NOT pre-selected using all samples.

(2) Need to assess the impact of training set size.

(3) Need to examine the impact of feature selections.

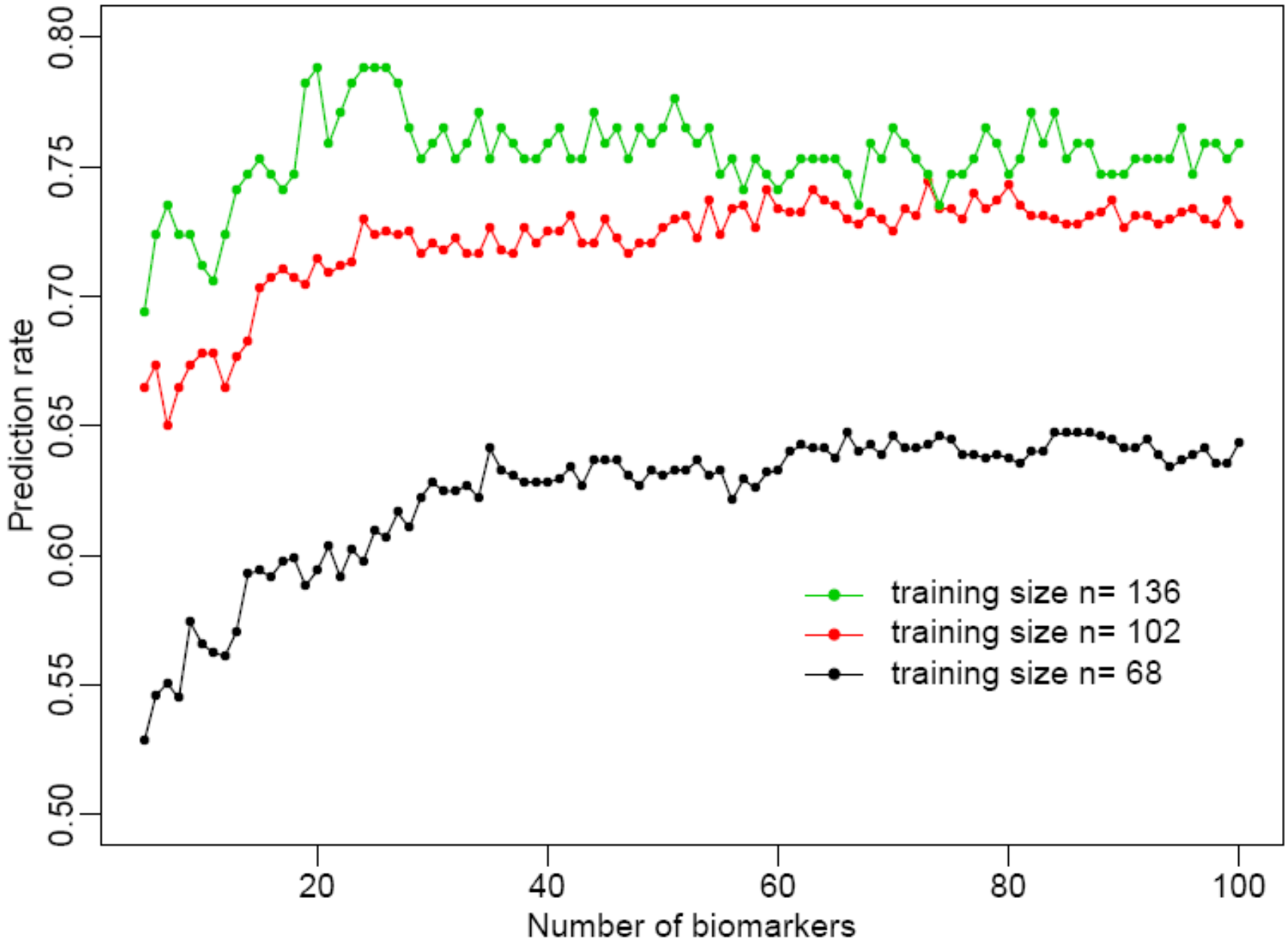
Approach

- Estimate error rate (Err) as a function of sample size n and number of features m: Err(n,m).

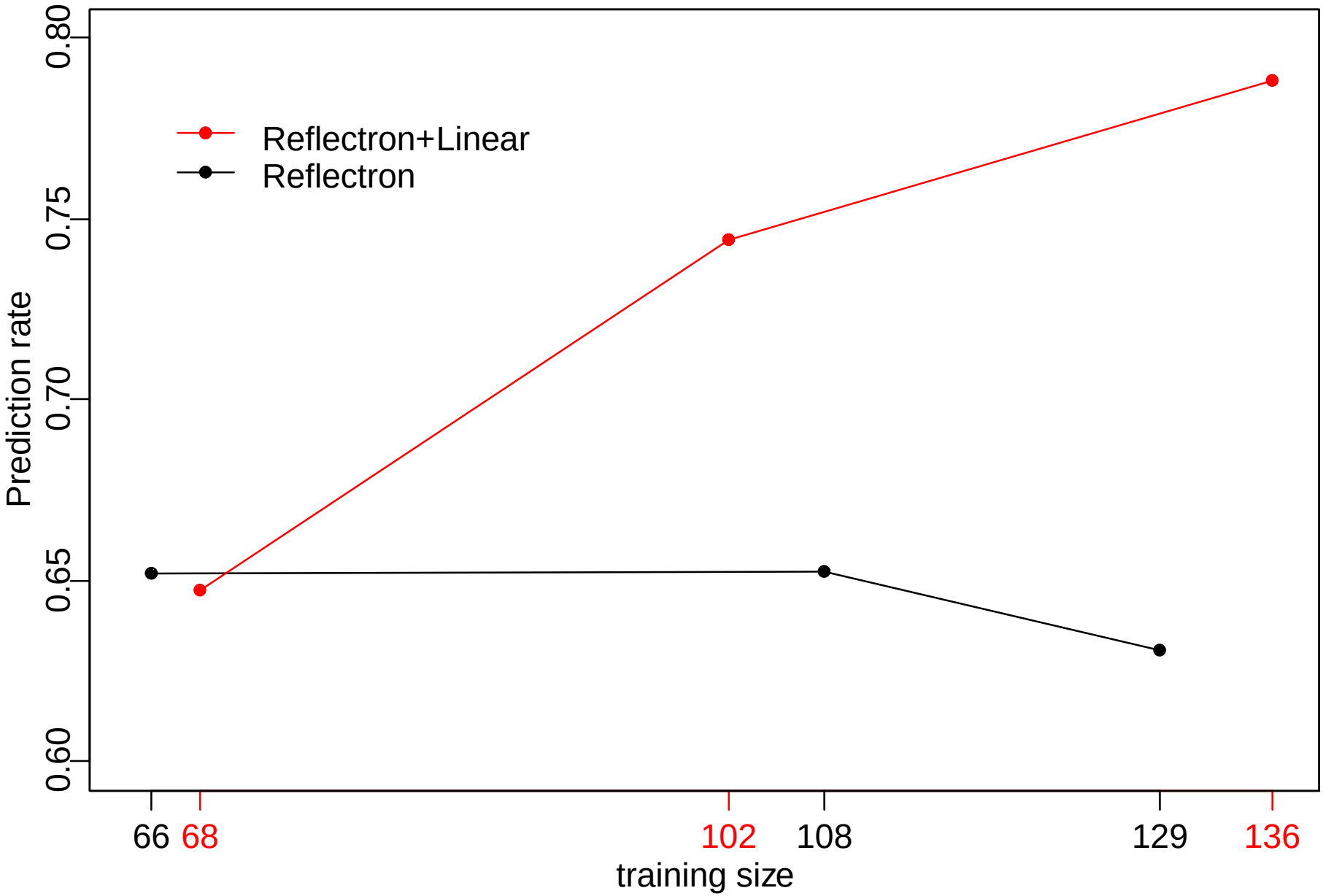
- We have a total of N=170 samples (77 normal + 93 ovarian cancer). Split 170 samples into 5 similar groups, each with 34 samples.

- Use *k* group as test set and other *5-k* groups as training set to estimate Err(136,m), there are a total of C(5,k) test sets.

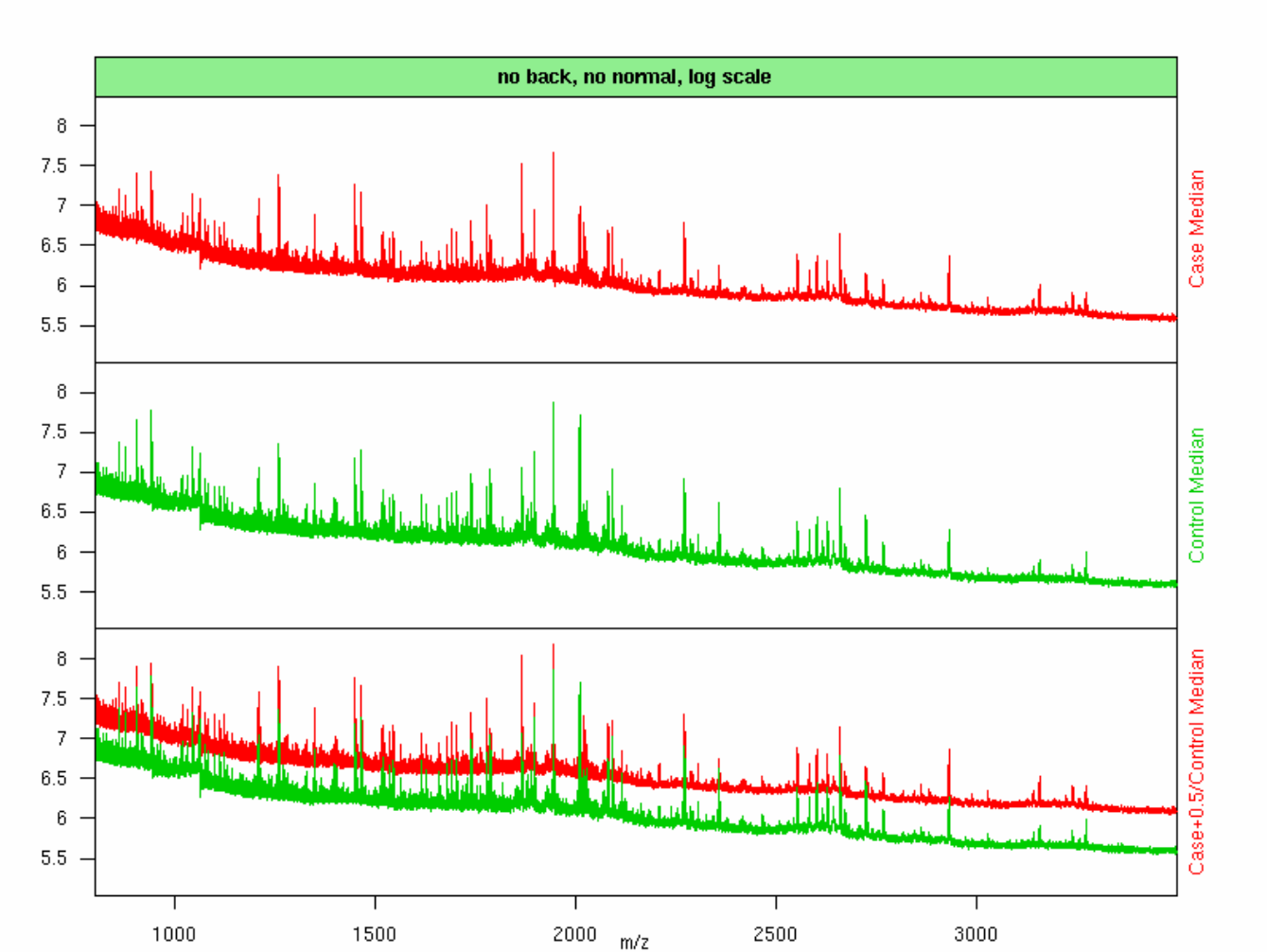
Reflectron+Linear Data

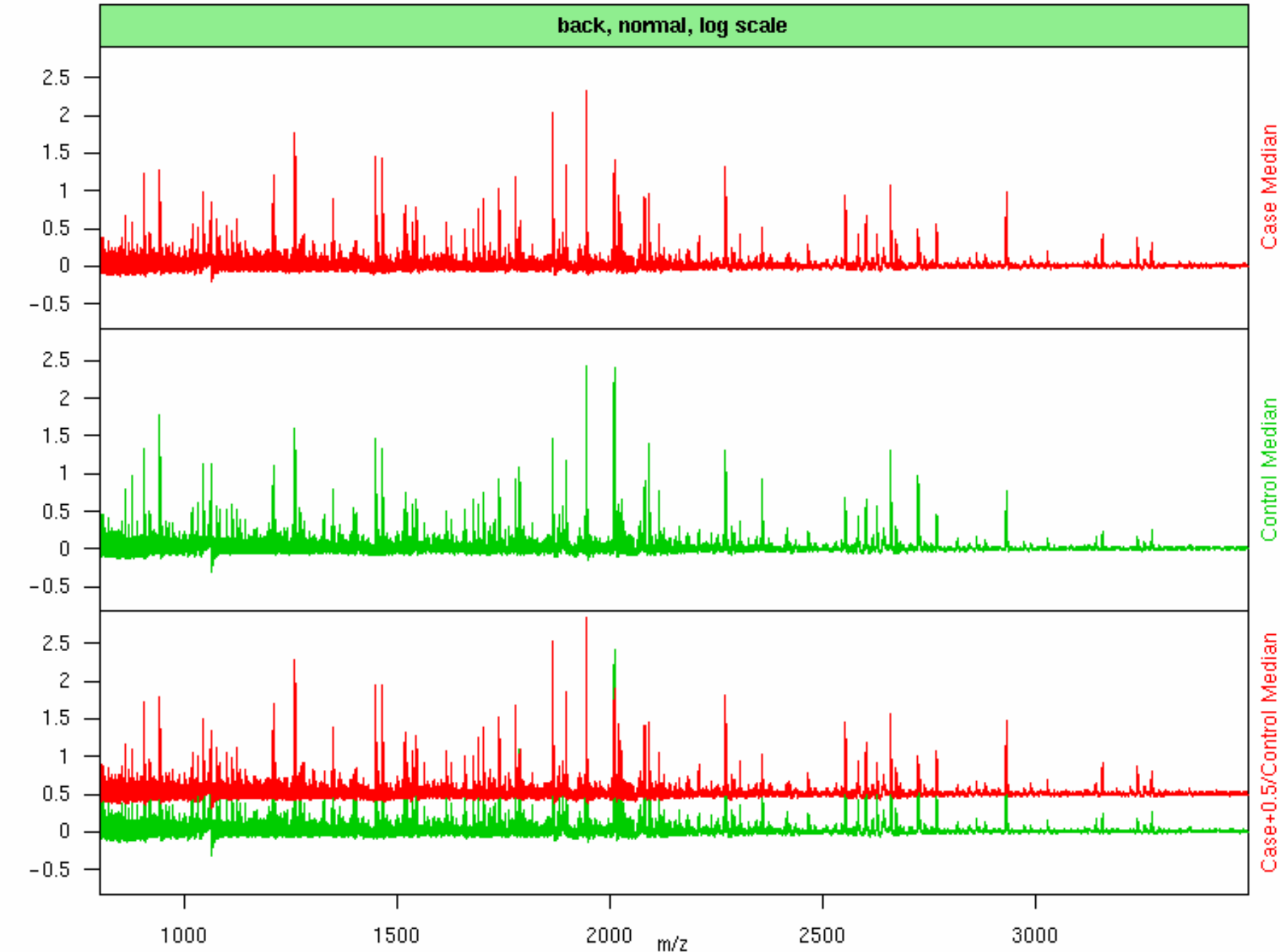


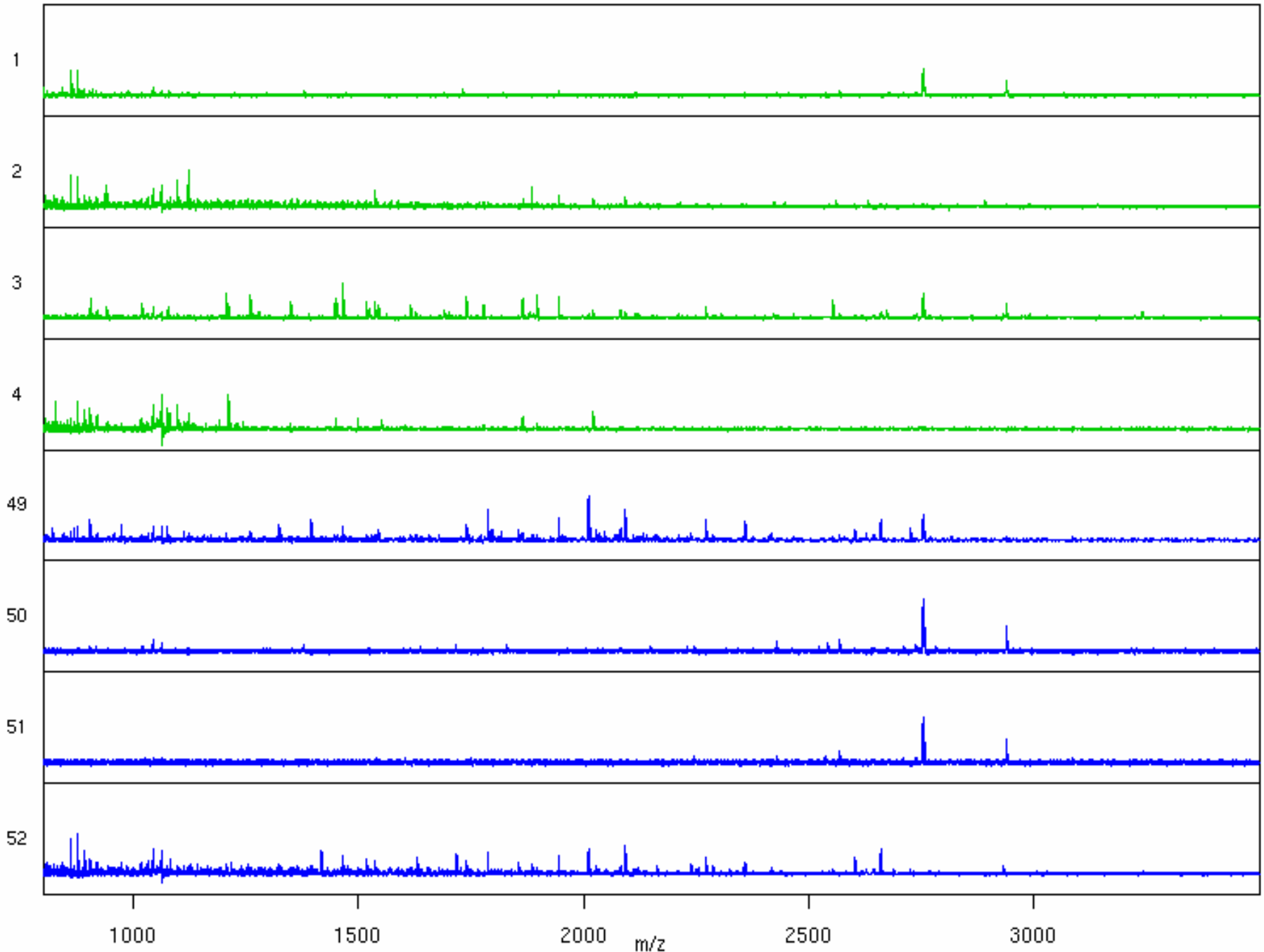
From Reflectron to Linear

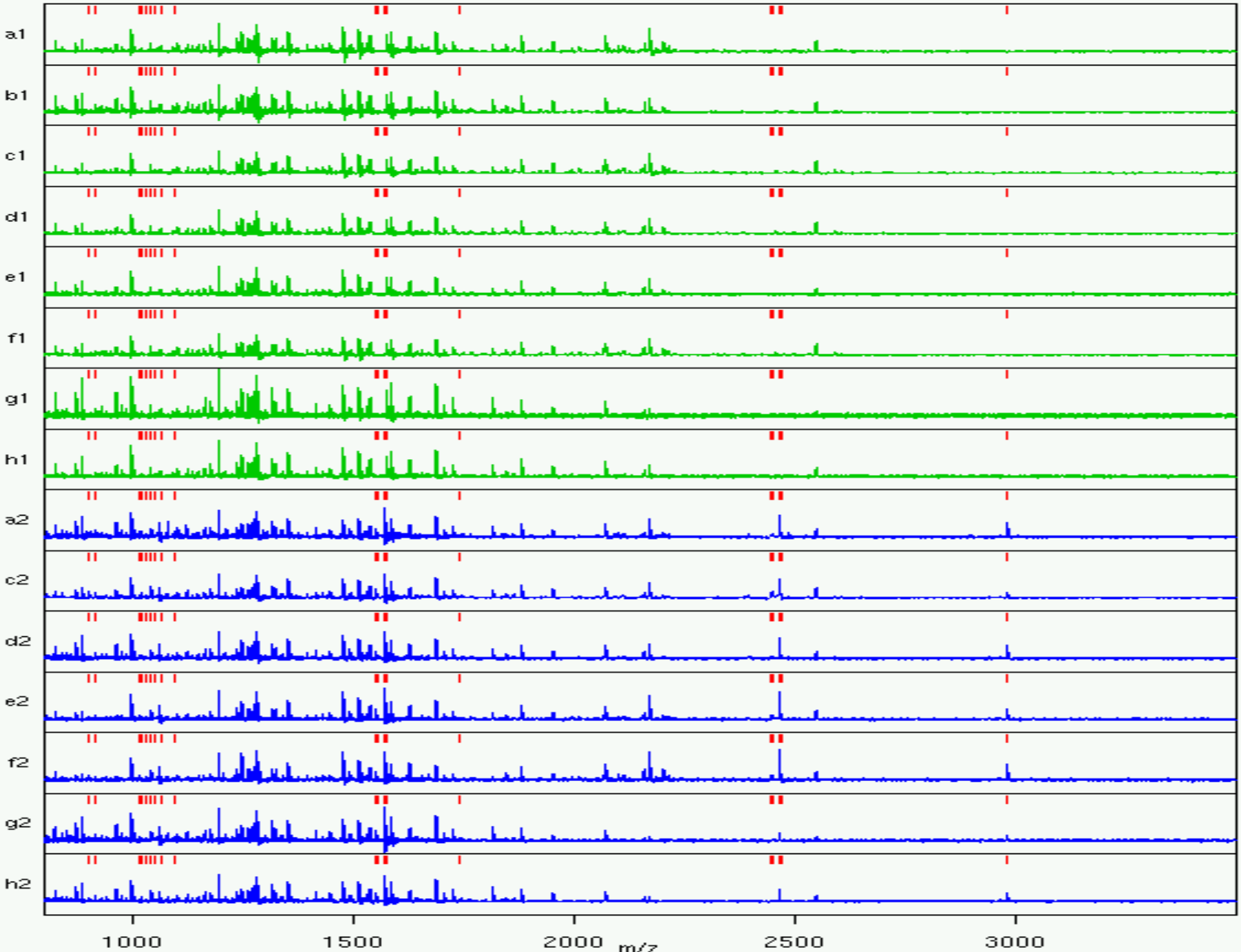


- Visualization methods -- some plots




back, normal, log scale
2.5
2
1.5
1
0.5
0
-0.5
Case Median
Control Median
Case+0.5/Control Median
1000
1500
2000
2500
3000
m/z


1
2
3
4
49
50
51
52
1000
1500
2000
m/z
2500
3000


a1
b1
c1
d1
e1
f1
g1
h1
a2
c2
d2
e2
f2
g2
h2
1000
1500
2000
m/z
2500
3000

Software Implementation

- ***R*** software

  All our current analyses are based on ***R*** software (http://cran.r-project.org).

Conclusions

- MS is a promising yet **challenging** technology for disease diagnosis and prognosis
- Many statistical problems need to be resolved to make the most and appropriate use of MS data
  - Initial alignment
  - Normalization
  - Peak identification and alignment
  - Classification
  - Visualization
- Data being collected can be of much higher dimension and complexity
- MS data need to be related to biology, and considered in combination with gene expression data as well as sequence and protein data to understand the results from such analyses

Yale/NHLBI
Proteomics Center

301 BCMM
P.O. Box 9812
New Haven, CT
06536-0812

Contact Us

Yale/NHLBI Proteomics Center

Contact the Director

Important Notices:

- Yale University **press release** (10/28/2002) and **Bulletin article** (11/1/2002) announcing funding of the Yale/NHLBI Proteomics Center
- NHLBI press release announcing formation of 10 **Proteomics Centers** (10/9/2002)
- **Yale Cancer Center Press Release** (9/20/02) announcing funding of an NIH High End Instrumentation Grant to purchase an FT-ICR mass spectrometer for the Keck Laboratory

General Information About the Center:
Center Members

Description and History

The Yale/NHLBI Proteomics Center Contract

- Fact Sheet
- Abstract
- Listing of Research Projects
- Research Project Summaries

Background Information on the Human Proteome and the Yale/NHLBI Proteomics Center

Positive Impact of the Center

Description of the HHMI Biopolymer/Keck Foundation Biotechnology Resource Laboratory at Yale University

Information About Mass Spectrometry:

- FAQ about MS
- FAQ about FTICR

Information and Technical Reports on Peptide/Protein Profiling Methodologies

Information on Research in the Center

- Hematopoiesis Overview
- Gene Regulation in Hematopoiesis
- Myeloid Database

Yale Proteomic Web Sites:

- Mark Gerstein Laboratory
- Analysis of mRNA and protein expression data

Proteomics Tools:
Analysis of Protein:Protein Interactions

Peptide/Protein Profiling Services:

- MALDI-MS Based Serum Peptide Disease Marker Discovery
- Instructions for Submitting a 96-Well Plate for MALDI-MS Peptide Disease Biomarker Screening

Acknowledgements

- **Yale/NHLBI Proteomics Center**
- **Keck Foundation Biotechnology Resource Laboratory at Yale U.**
- **David Fishman (Northwestern University)**
- **Gil Mor, Yale U.**
- **David Ward, Yale U.**