Over Viewing SELDI Data to Determine Quality: what are the statistical and pre-analytical considerations?

- Over view a proteomic dataset using simple statistical methods.
- Discuss role of acute phase proteins and protein concentration in structuring and interpreting biomarker research.
- Discuss pre-analytic causes of assay variation in general and comment on relevance to Mass Spectrometry techniques.

Proteomic Dataset Overview

- Done in collaboration with Min Zhan Ph.D., Dept. of Epidemiology, Univ. Of Maryland Baltimore
- Sorace JM, Zhan M: A data review and re-assessment of ovarian cancer serum proteomic profiling.
  BMC Bioinformatics. 2003 Jun 09;4(1):24.
  PMID: 12795817

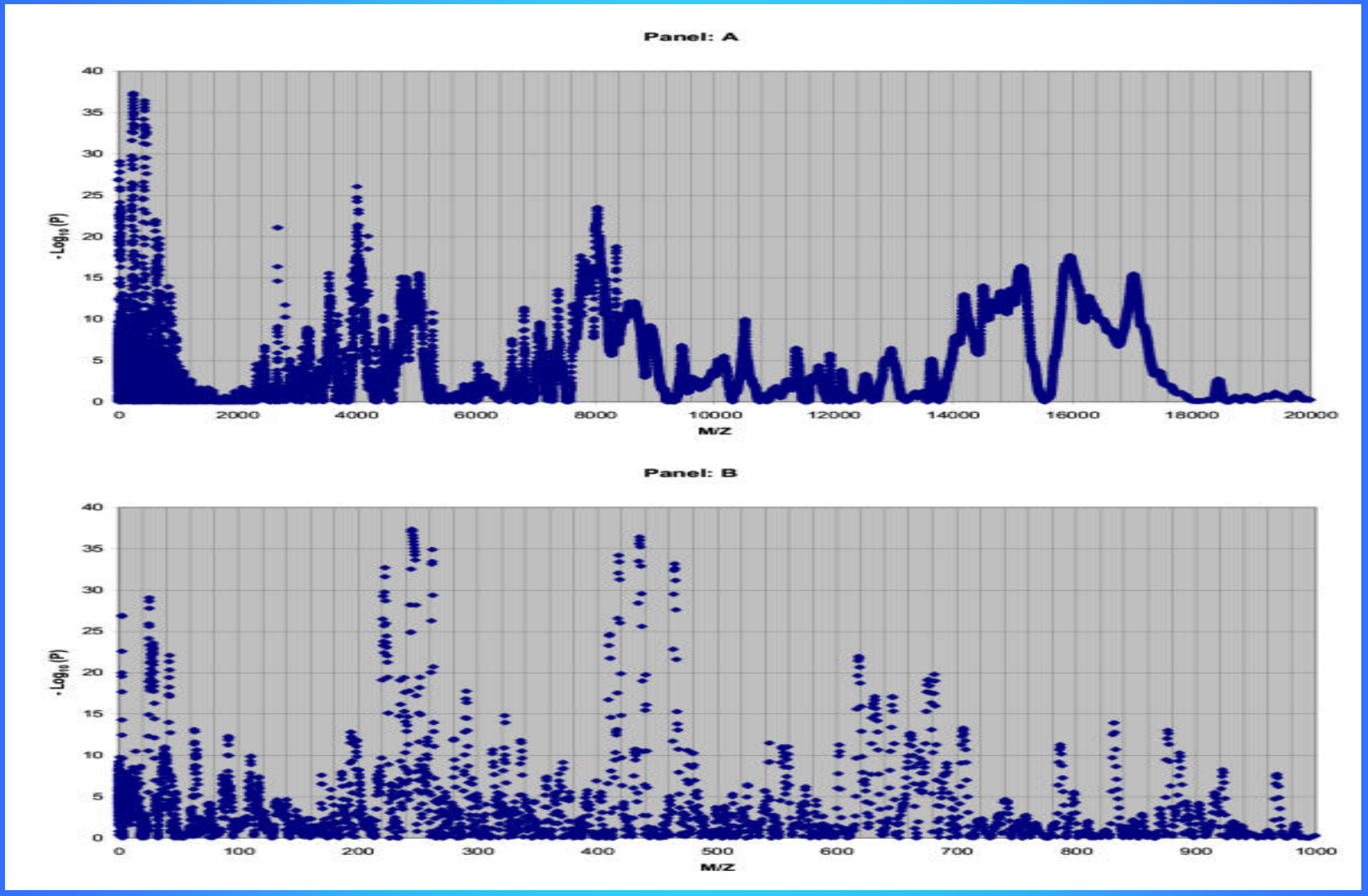
8-7-02 Clinicalproteomics Databank Dataset

- 161 Cancer and 91 Non-cancer serum samples
- 28 – stage 1, 20 – stage 2, 99 – stage 3, 12 – stage 4, 3- NSP
- Run on a Ciphergen WCX-2 Protein Chip Array
- 15,154 distinct M/Z values
- Posted without base line subtraction
- Rule presented that discriminates between the 2 groups with 100% sensitivity and specificity

SAS Analysis

- Will start with the second thing we did.
- In the first set of studies the data set was divided sample into a training set and a test set, will discuss monetarily.
- Performed Wilcoxon test on all M/Z values and plotted the results.

Wilcoxon test P-value M/Z Distribution



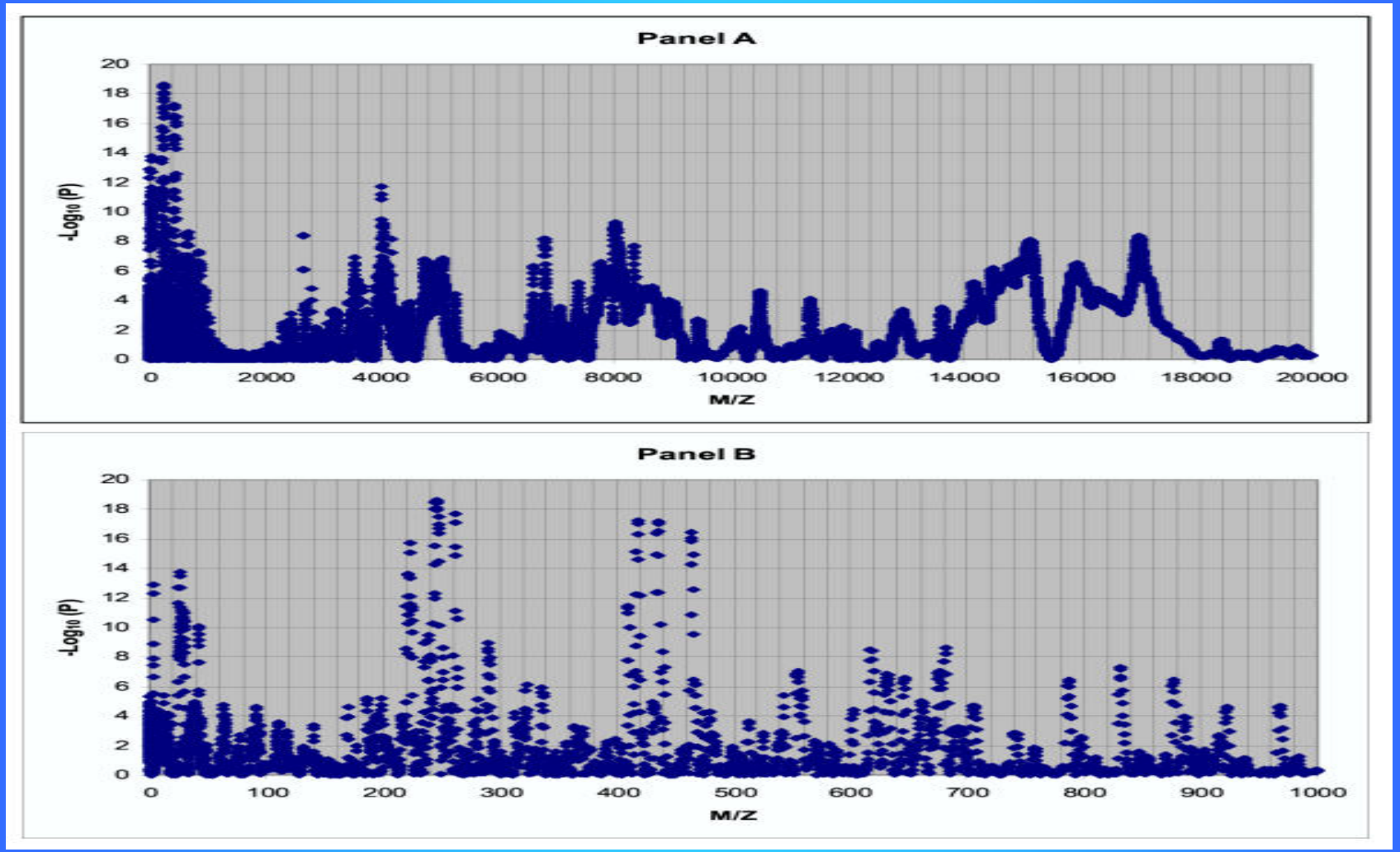
Wilcoxon Score Distribution

- Substantial statistical difference between these 2 groups.
- 3,591 M/Z p-values $<10^{-6}$
- Area of greatest statistical difference <1000 M/Z, particularly < 500 M/Z. This area typically considered noise. Also difficult to interpret M/Z values due to calibration considerations.
- However low molecular weight biomarkers of ovarian cancer have been described (LPA measured in plasma).

Traditional Statistics – Development of Classifiers

- First thing we did - randomly split samples into training and test sets (45/91 non-cancers and 80/162 cancers).
- Performed Wilcoxon Test at all M/Z values

Training Set P-Value M/Z Distribution



Traditional Statistics – Development of Classifiers 2

- Simply sorted the resulting p-values to determine most significant 100 M/Z values.
- These were further assigned to 12 bins by simply assigning consecutive M/Z values to the same bin.

Rule 1: Binning and Stepwise Discriminant Analysis

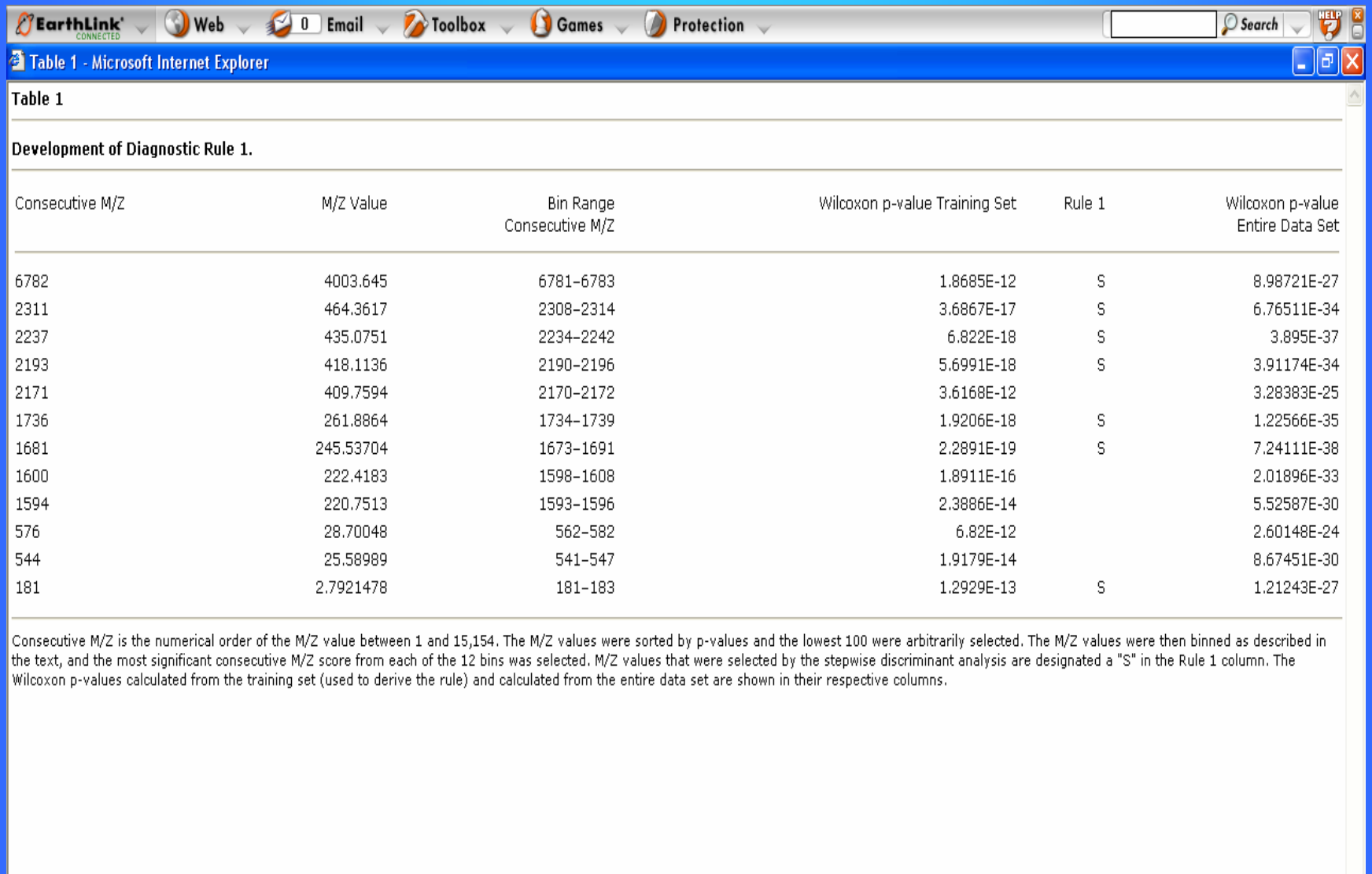


Table 1

Development of Diagnostic Rule 1.

Consecutive M/Z	M/Z Value	Bin Range Consecutive M/Z	Wilcoxon p-value Training Set	Rule 1	Wilcoxon p-value Entire Data Set
6782	4003.645	6781–6783	1.8685E-12	S	8.98721E-27
2311	464.3617	2308–2314	3.6867E-17	S	6.76511E-34
2237	435.0751	2234–2242	6.822E-18	S	3.895E-37
2193	418.1136	2190–2196	5.6991E-18	S	3.91174E-34
2171	409.7594	2170–2172	3.6168E-12		3.28383E-25
1736	261.8864	1734–1739	1.9206E-18	S	1.22566E-35
1681	245.53704	1673–1691	2.2891E-19	S	7.24111E-38
1600	222.4183	1598–1608	1.8911E-16		2.01896E-33
1594	220.7513	1593–1596	2.3886E-14		5.52587E-30
576	28.70048	562–582	6.82E-12		2.60148E-24
544	25.58989	541–547	1.9179E-14		8.67451E-30
181	2.7921478	181–183	1.2929E-13	S	1.21243E-27

Consecutive M/Z is the numerical order of the M/Z value between 1 and 15,154. The M/Z values were sorted by p-values and the lowest 100 were arbitrarily selected. The M/Z values were then binned as described in the text, and the most significant consecutive M/Z score from each of the 12 bins was selected. M/Z values that were selected by the stepwise discriminant analysis are designated a "S" in the Rule 1 column. The Wilcoxon p-values calculated from the training set (used to derive the rule) and calculated from the entire data set are shown in their respective columns.

Development of Rule 1

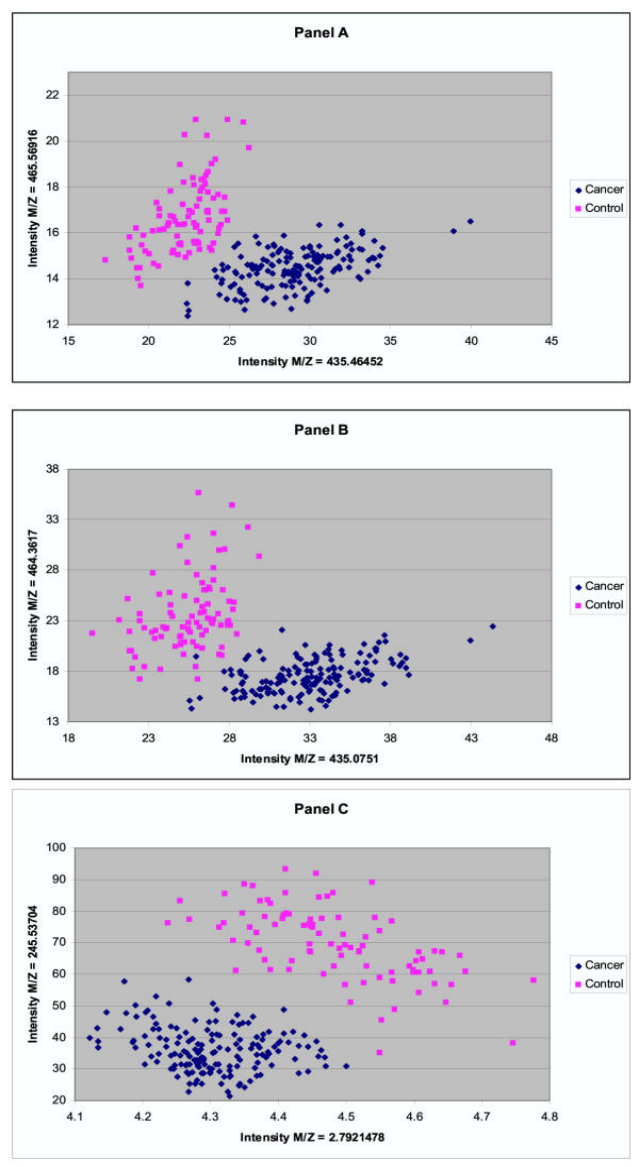
- Achieved 100% sensitivity and specificity without complex data mining approaches.
- 11 of 12 initial values all have M/Z values less than 500.
- 2 M/Z values 464.3617 and 435.0751 correspond to values in the clinincalproteomics database rule 465.56916 and 435.46452. Both pairs give excellent discrimination.

Development of Rule 1 Continued

- Also get excellent discrimination with 2.7921478 and 245.53704
- Of 7 M/Z values finally selected only one 4003.645 is > 500.

Low M/Z Discriminators

- Cancer in Blue
- Clinical Proteomics Rule: 465.56916 vs. 435.46452
- Rule 1: 464.3617 vs. 435.0751
- Rule 1: 245.53704 vs. 2.7921478
- Other groups using standard statistics have come to identical conclusions.



ClinicalProteomics Rule

Table 4 - Microsoft Internet Explorer

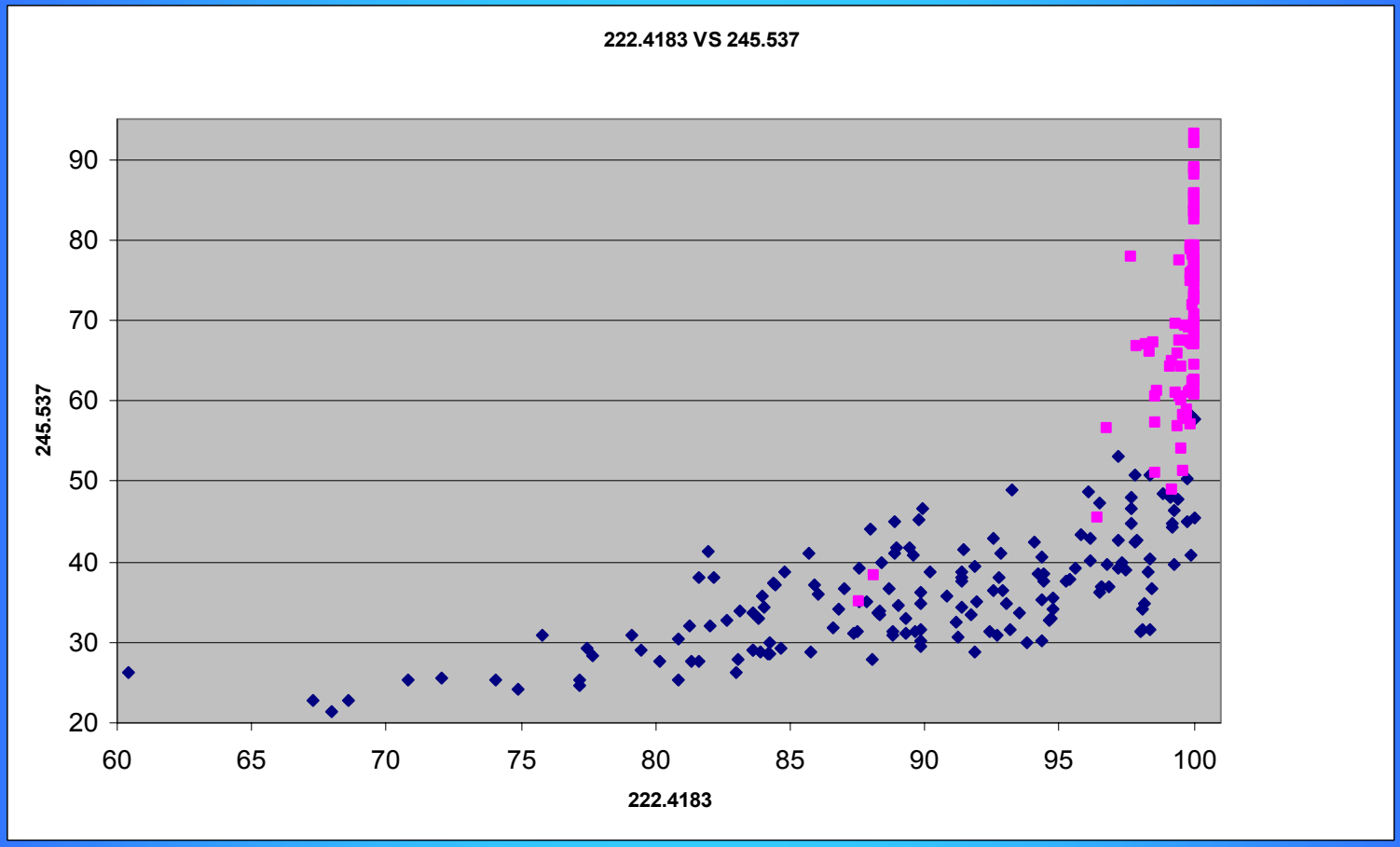
Table 4

Clinical Proteomics Program Databank Example Ovarian Rule.

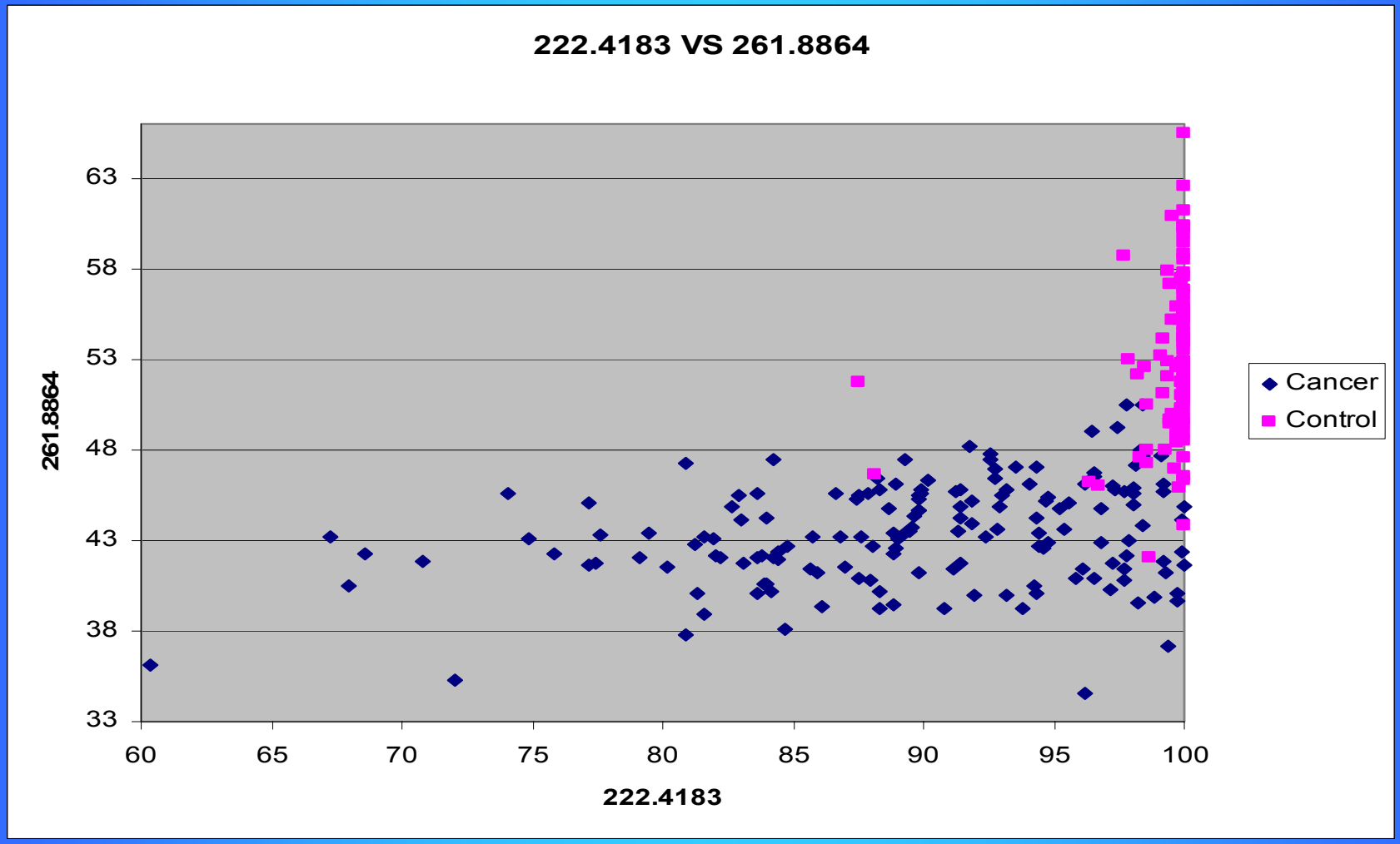
Consecutive M/Z Bin	M/Z-Value	P2_Wil
5632	2760.6685	0.239533474
15020	19643.409	0.521014657
2314	465.56916	2.49791E-28
8728	6631.7043	9.00537E-4
12704	14051.976	1.79156E-08
2238	435.46452	9.07922E-37
6339	3497.5508	1.40316E-06

Consecutive M/Z values and Wilcoxon p-values based on the entire dataset for the rule present on the Clinical Proteomics Program Databank website.

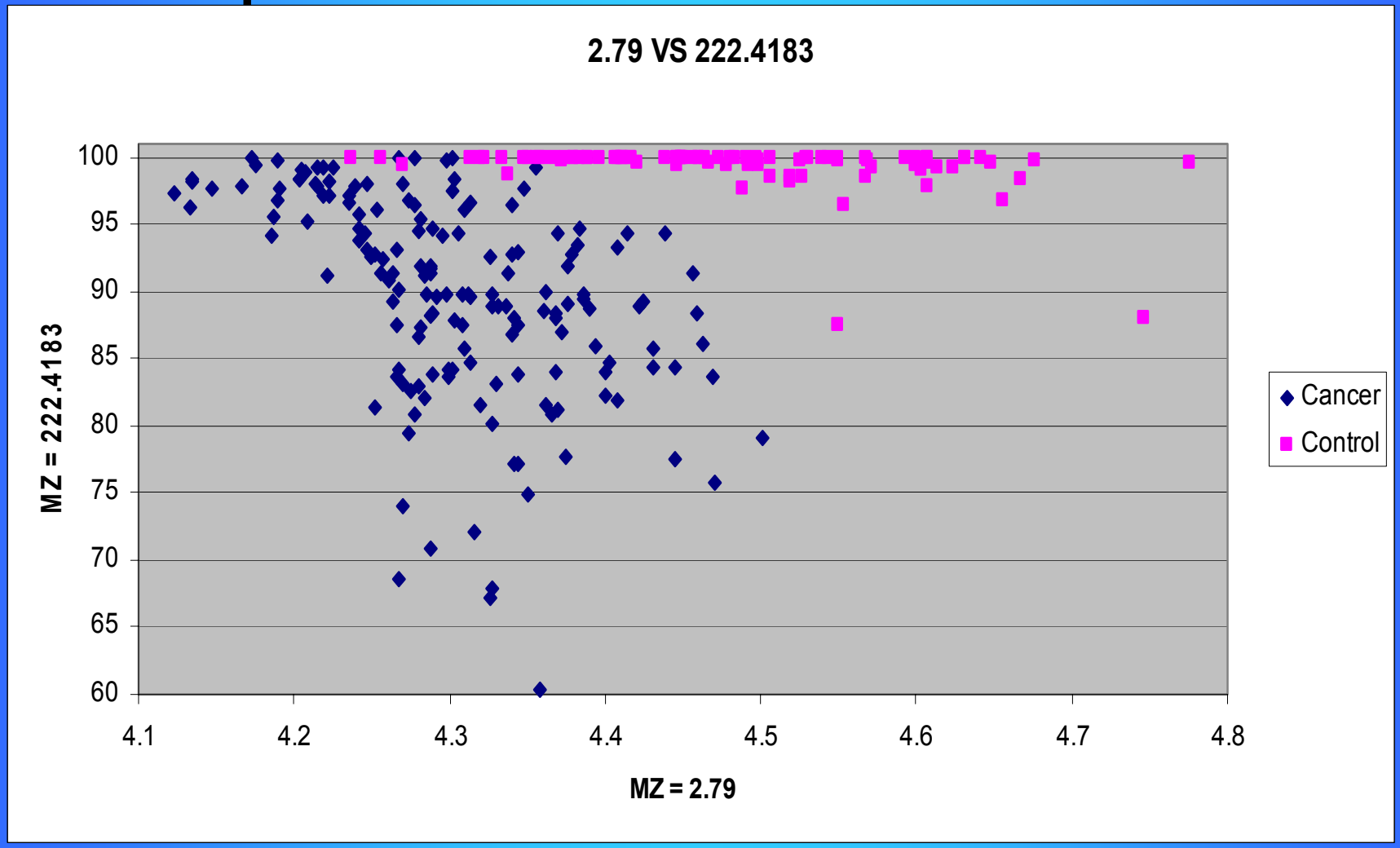
Additional Example of Low M/Z Separation 222.4183 VS 245.537



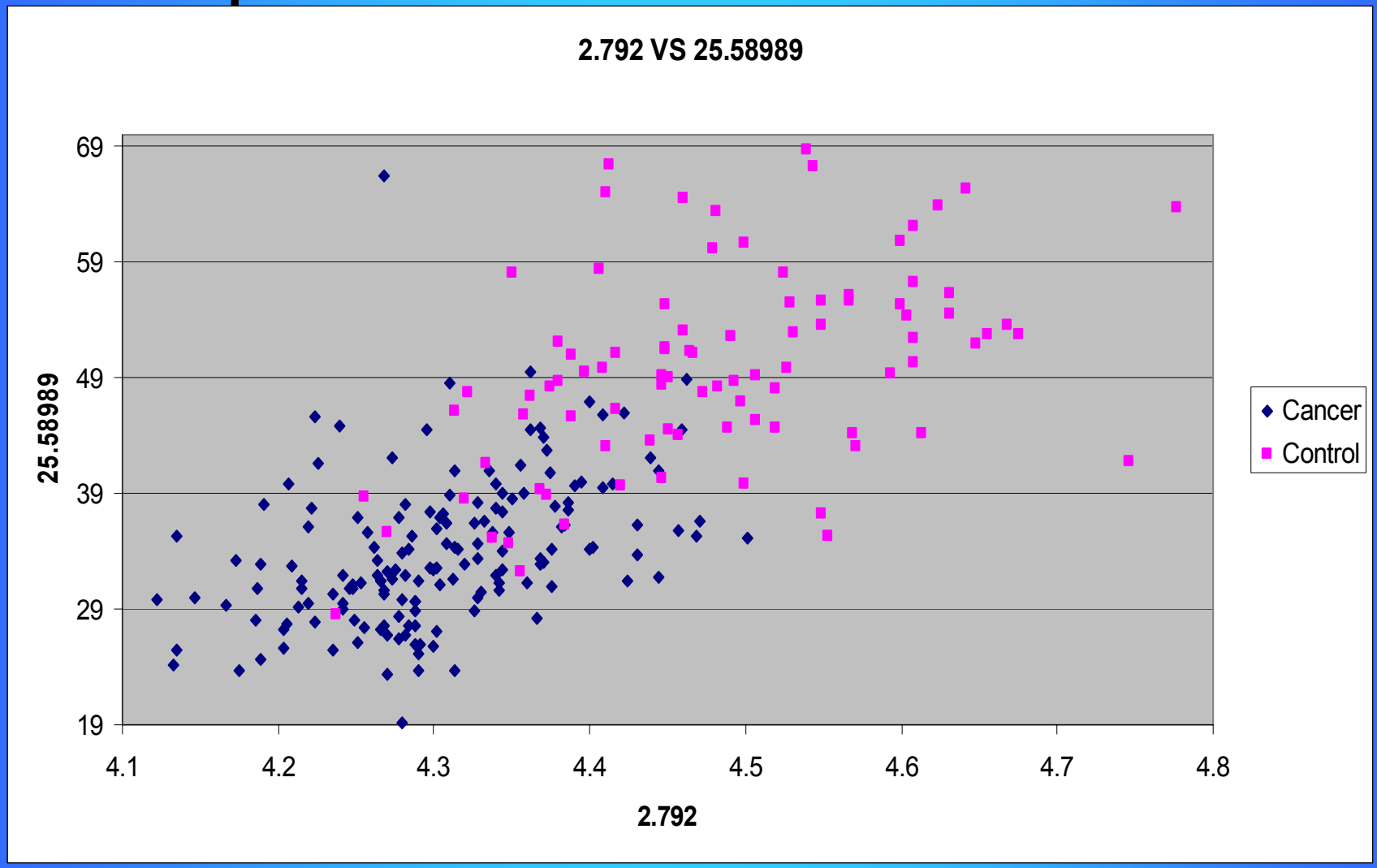
Additional Example of Low M/Z Separation 222.4183 VS 261.8864



Additional Example of Low M/Z Separation 2.79 VS 222.4184



Additional Example of Low M/Z Separation 2.79 VS 25.58989



Other Approaches 1

- Required M/Z greater than 2000 and $P < 10^{-6}$ used similar approach to get 13 M/Z values from 30 bins. For 96.25% sensitivity and 91.1% specificity on the test set.

Rule 2: Binning and Stepwise Discriminant Analysis

Table 2 - Microsoft Internet Explorer

Table 2

Development of Diagnostic Rule 2.

Consecutive M/Z	M/Z	Wilcoxon p-value	Rule 2
5534	2665.397	4.06E-09	S
6372	3534.072	1.26E-07	
6753	3969.469	4E-07	S
6772	3991.844	6.87E-09	S
6782	4003.645	1.87E-12	S
6802	4027.3	6.21E-10	S
6814	4041.526	1.86E-07	
6823	4052.213	8.33E-07	
6827	4056.967	9.38E-07	S
6836	4067.673	3.9E-07	
6852	4086.742	4.11E-07	
6934	4185.17	6.56E-09	
7383	4744.889	1.71E-07	S
7449	4830.124	2.89E-07	
7468	4854.802	8.22E-07	
7508	4906.962	5.45E-07	
7606	5035.93	1.41E-07	
8707	6599.823	4.96E-07	
8839	6801.495	6.46E-09	S
9439	7756.437	2.66E-07	
9457	7786.054	4.58E-07	S
9483	7828.934	6.23E-07	
9607	8035.058	4.94E-10	
9793	8349.266	2.04E-08	S
12910	14511.46	6.4E-07	
13036	14796.14	3.95E-07	S
13113	14971.48	1.76E-07	
13201	15173.13	7.88E-09	
13537	15955.47	3.13E-07	S
13987	17034.05	4.53E-09	S

The M/Z values were sorted by M/Z values greater than 2,000 and p-values less than 10^{-6}. Consecutive M/Z is the numerical order of the M/Z value between 1 and 15,154. The M/Z values were then binned as described in the text, and the most significant consecutive M/Z score from each of the 30 bins was selected. M/Z values that were selected by the stepwise discriminant analysis are designated "S" in the rightmost column.

Other Approaches 2

- Hybrid Rule combining rules 1 (M/Z values 400 to 500) and 2. Again achieved 100% sensitivity and specificity.
- Not surprising given that M/Z values in the ranges of 435 and 465 can give perfect discrimination.

How to Explain Signal in Noise ?

- Three general options, other than bias
- 1st statistically significant low M/Z values are legitimate tumor markers – optimistic interpretation but there is evidence suggesting these may exist in ovarian cancer.
- 2nd common fragment of a higher molecular weight multigene family.
- 3rd matrix is sensitive to initial conditions and may actually sum over differences.

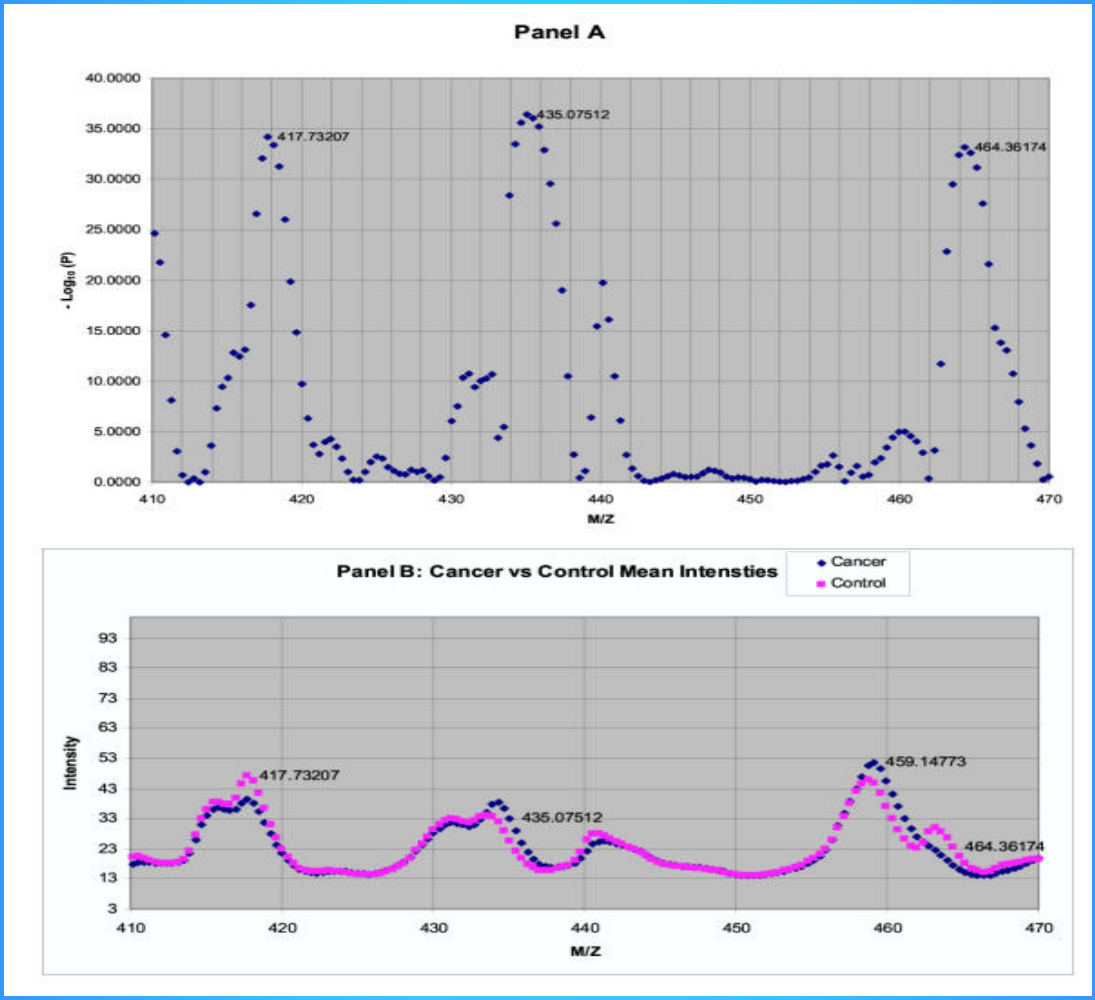
Hypothesis 1

- As noted in our Biomedcentral paper: “The disease process may influence the serum concentration of lipids, or other small molecules that either bind to the chip directly or through a complex formation with other macromolecules (e.g., binding to a receptor).”
- Lysophosphatidic Acid is a putative biomarker for ovarian cancer.
- Measured in plasma as it is a product of platelet activation.

Hypothesis 1

- LPA has been measured in plasma by first isolating a lipid band with TLC.
- Multiple family members.
- Plasma LPA band found to have increases using electrospray MS at M/Z values of 409, 433-437, 457, 481-482, 571, 599, and 619.
- The M/Z values discriminating between cancer and control in this dataset are associated with an increase at a M/Z value of about 435 and a decrease at about 464.

P-Values and Intensities for M/Z Values Between 410 and 470



Similar Hypothesis to Carrier Proteins Such as Albumin

- Mehta AI, Ross S, Lowenthal MS, Fusaro V, Fishman DA, Petricoin EF 3rd, Liotta LA. Biomarker amplification by serum carrier protein binding.
  Dis Markers. 2003-2004;19(1):1-10.
  PMID: 14757941
- Liotta LA, Ferrari M, Petricoin E Clinical proteomics: written in blood.
  Nature. 2003 Oct 30;425(6961):905. No abstract available.
  PMID: 14586448

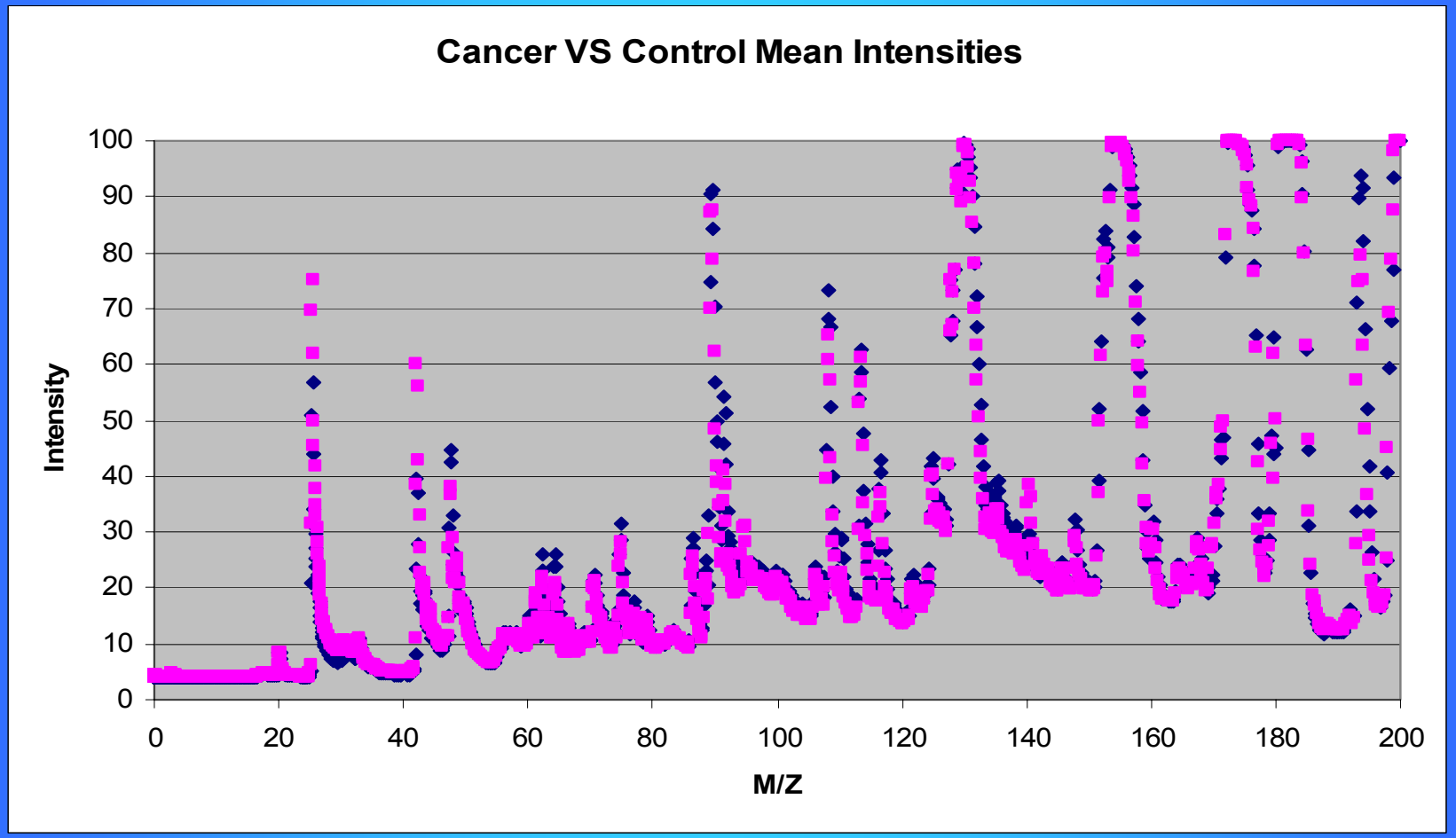
What Fraction Contains the Low Molecular Weight Biomarker?

- “Finally, the steps associated with sample collection, processing, and binding to the chip may represent a particularly fertile area for research. Any combination of such steps may significantly alter the molecular subset of the sample that can be successfully analyzed. “
- However fractionation may greatly complicate experimental design.
- Consider that serum involves the activation of the complement and coagulation pathways that generate low molecular weigh products and may complicate interpretation (e.g. reactive thrombocytosis).

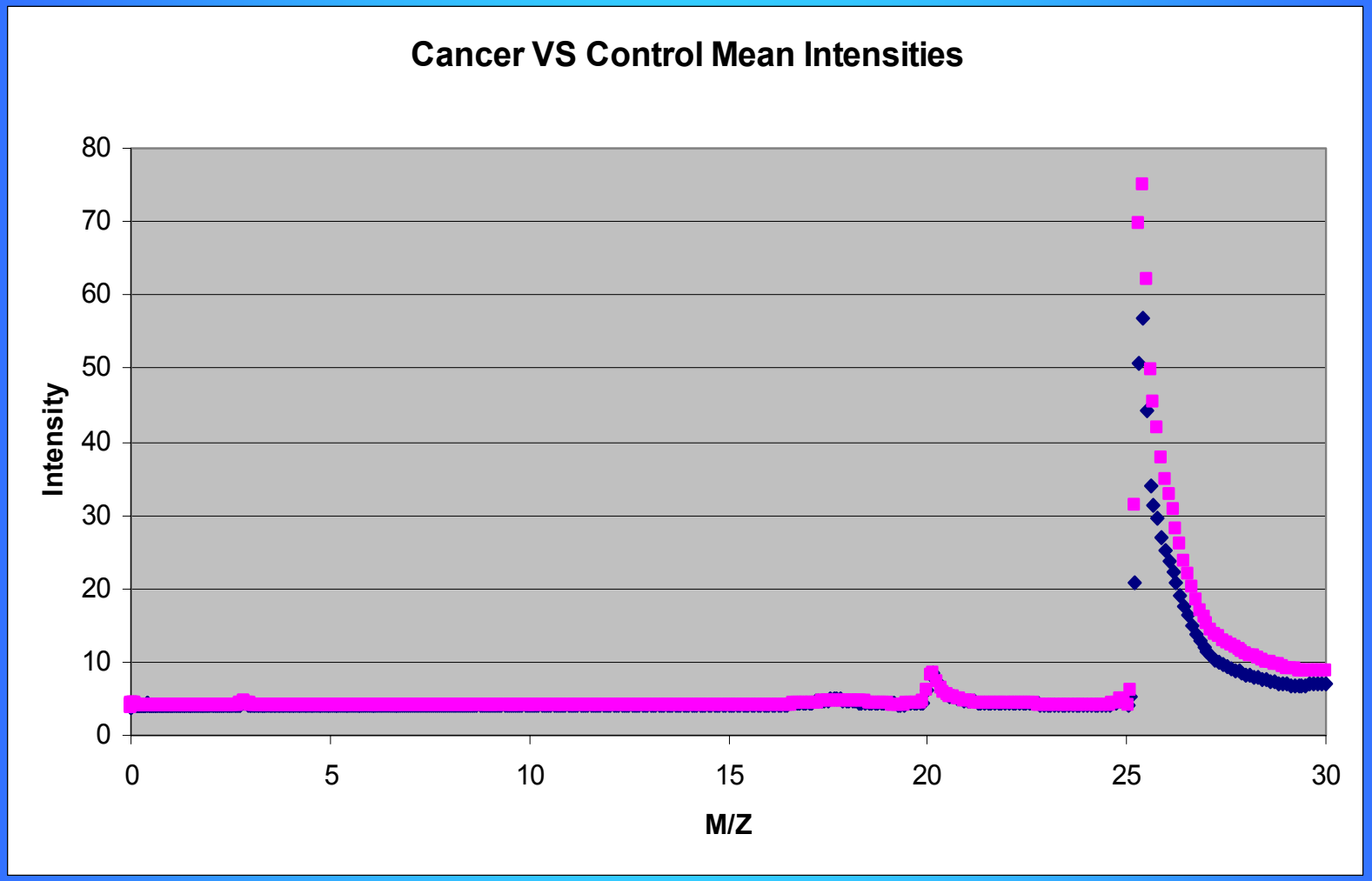
Hypothesis 2 and 3

- Speculative – all three proposals would benefit from identifying the peaks. Experiments spiking the sample with known proteins might be useful as would internal standards and confirmation with other measurement methods.
- Very difficult to apply to M/Z values of 2.79 and 25.58989.

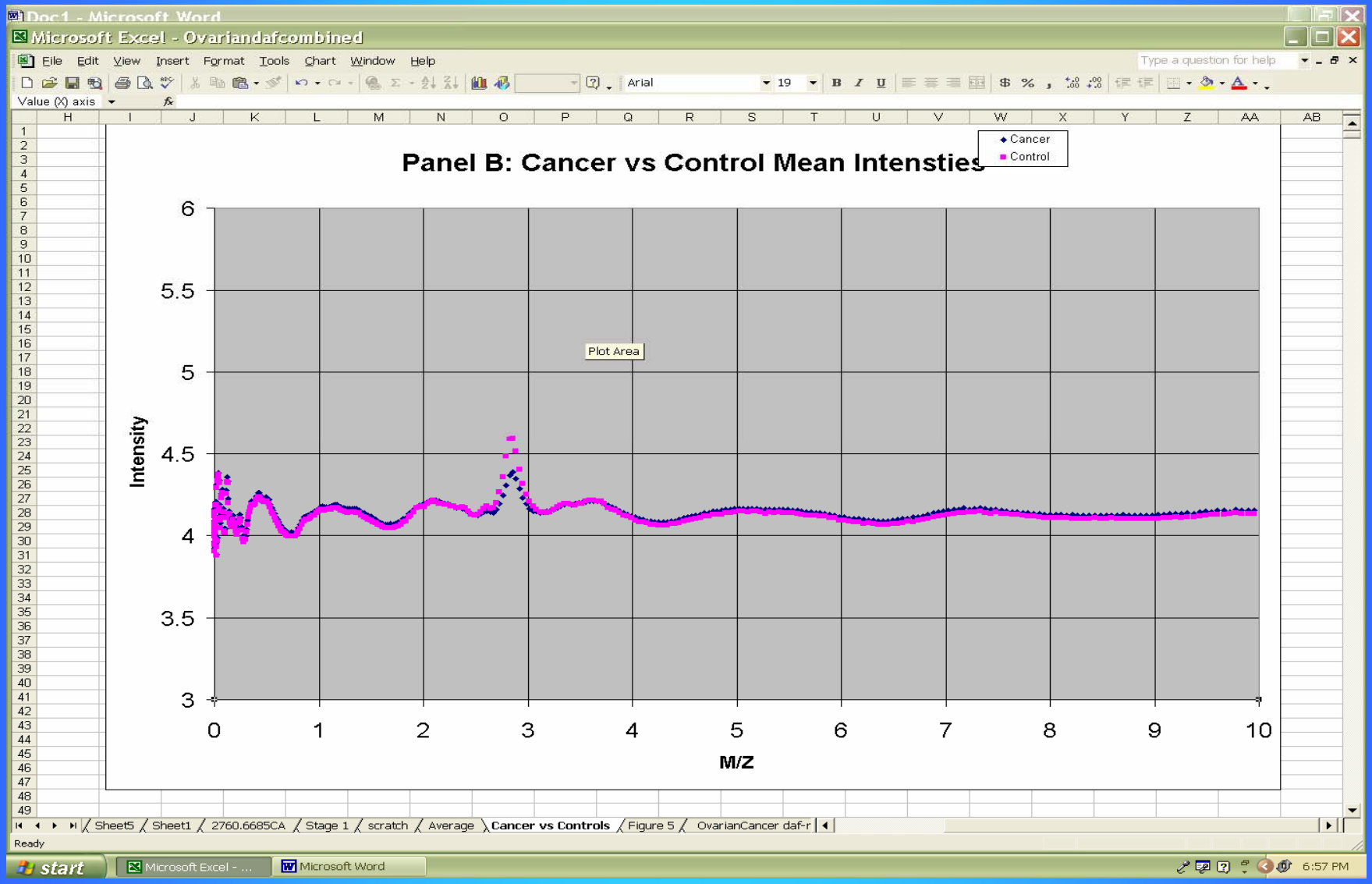
Early Peaks 1



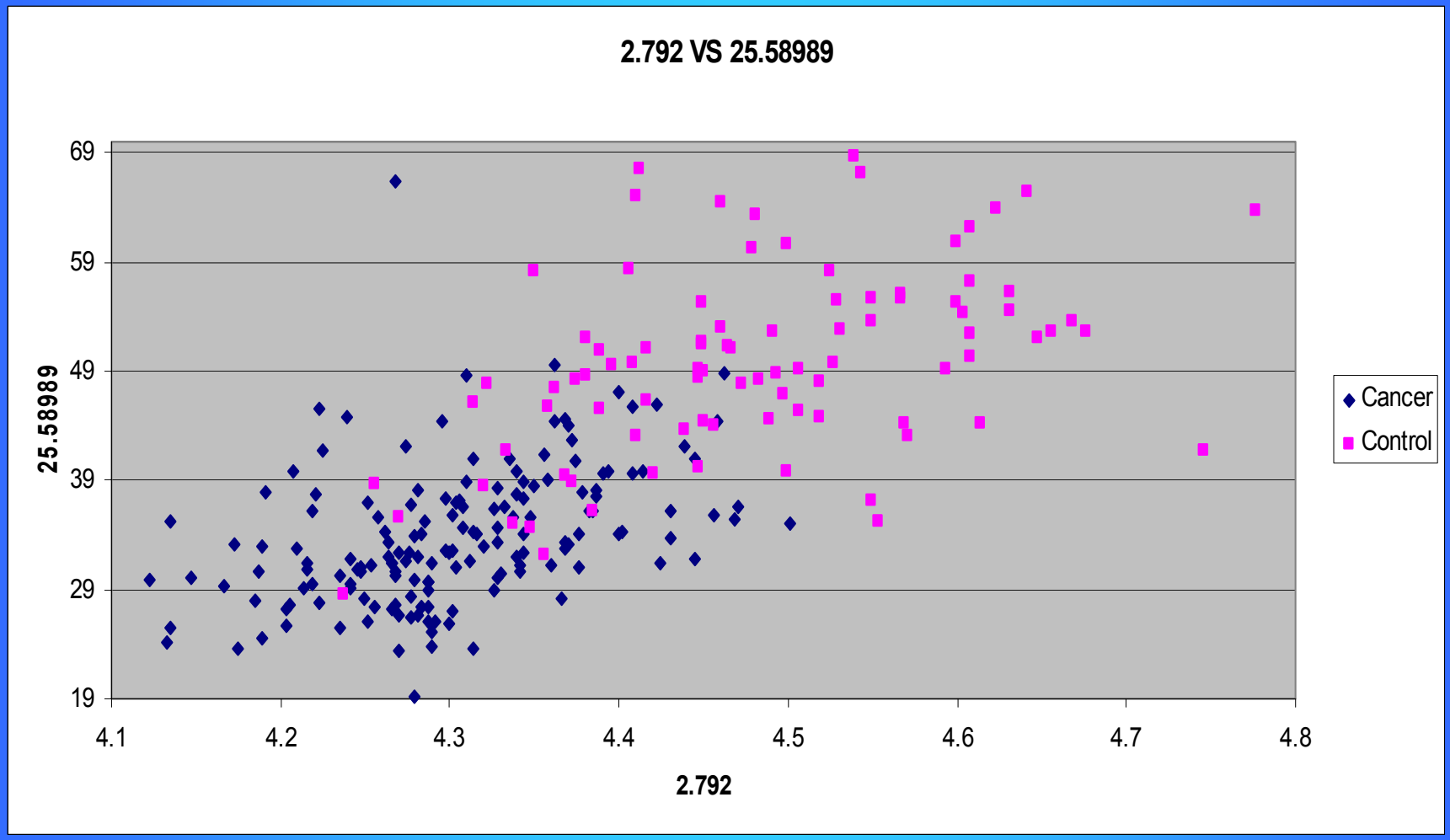
Early Peaks 2



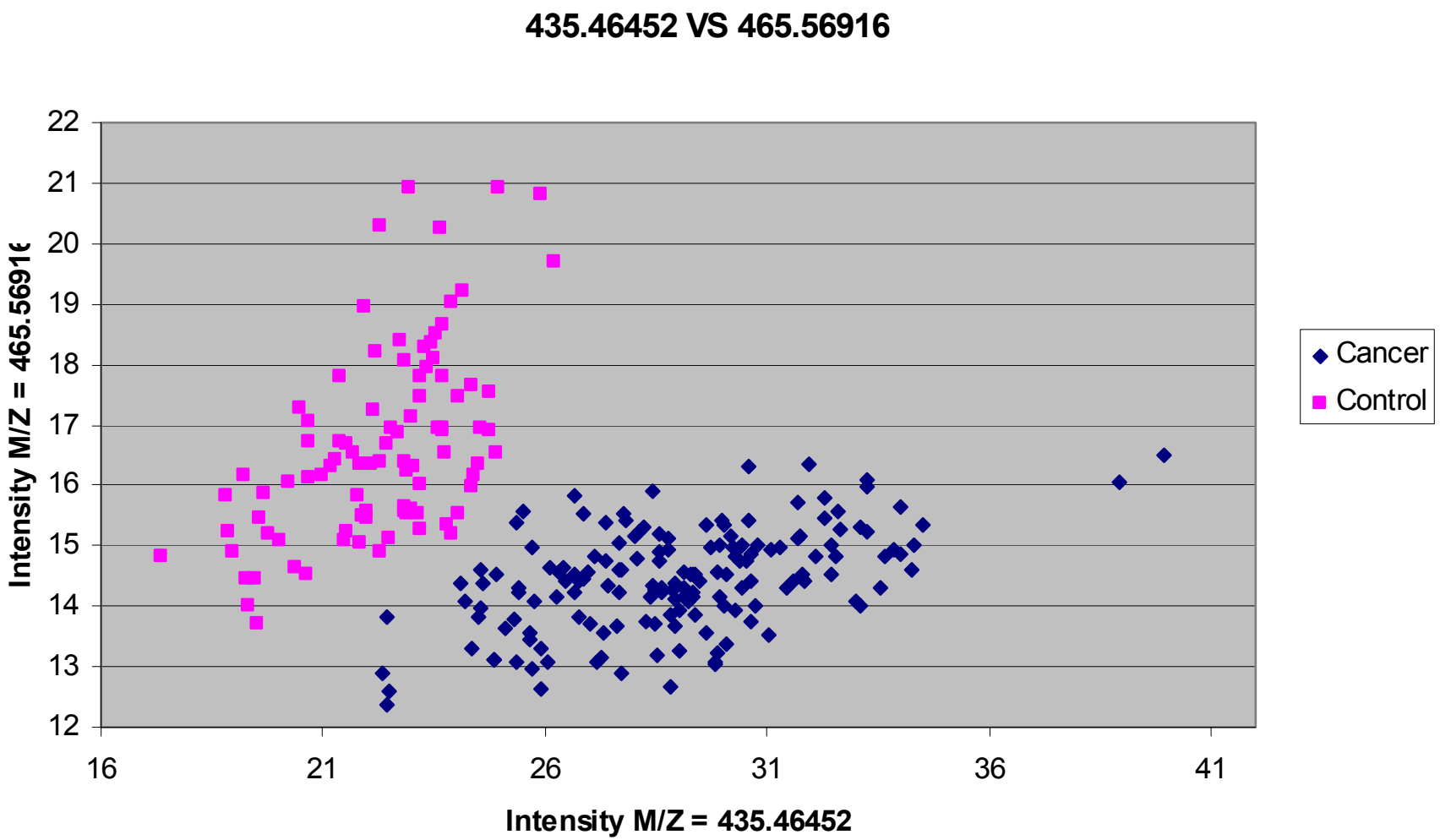
Early Peaks 3



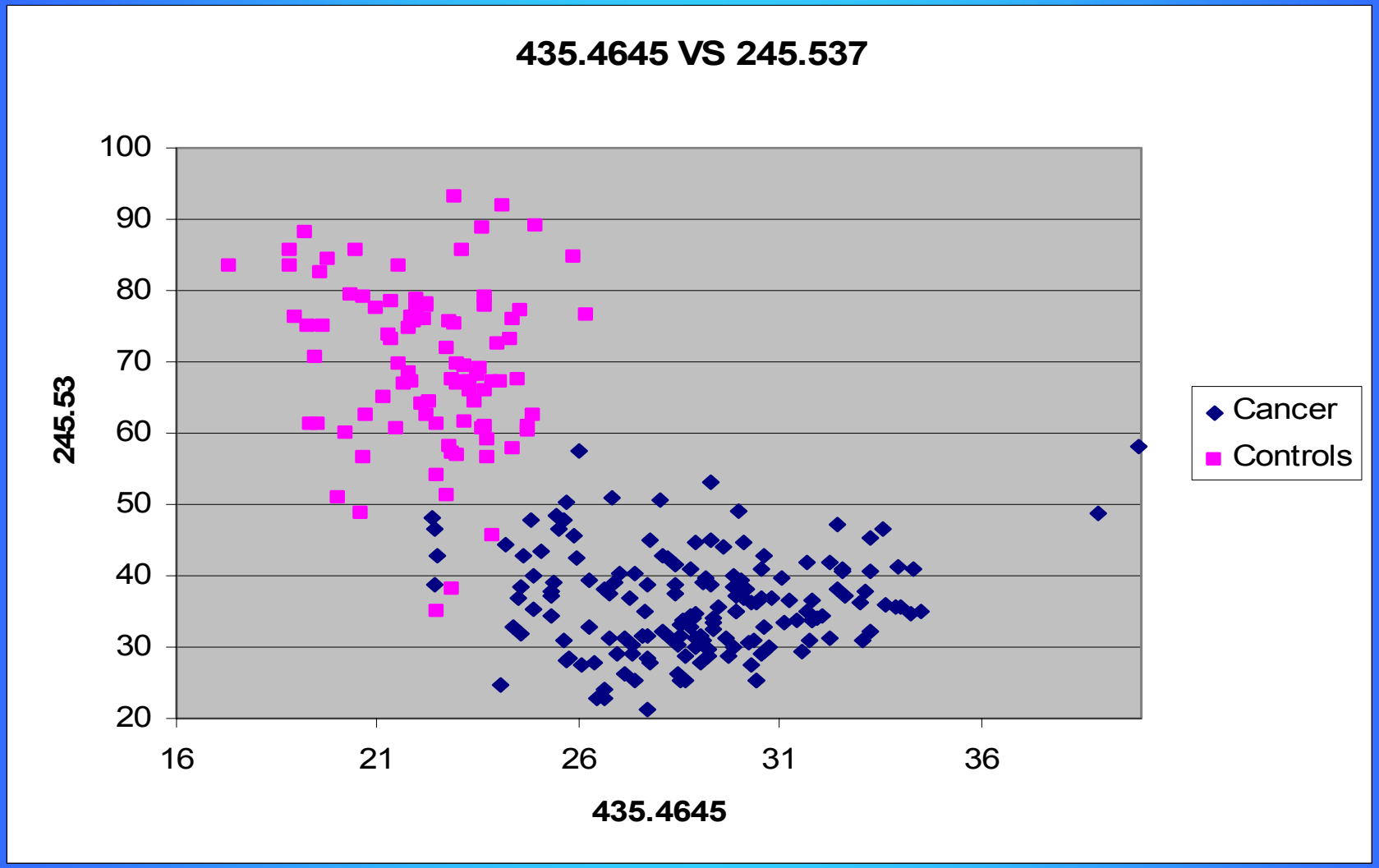
Additional Example of Low M/Z Signal 2.79 VS 25.58989



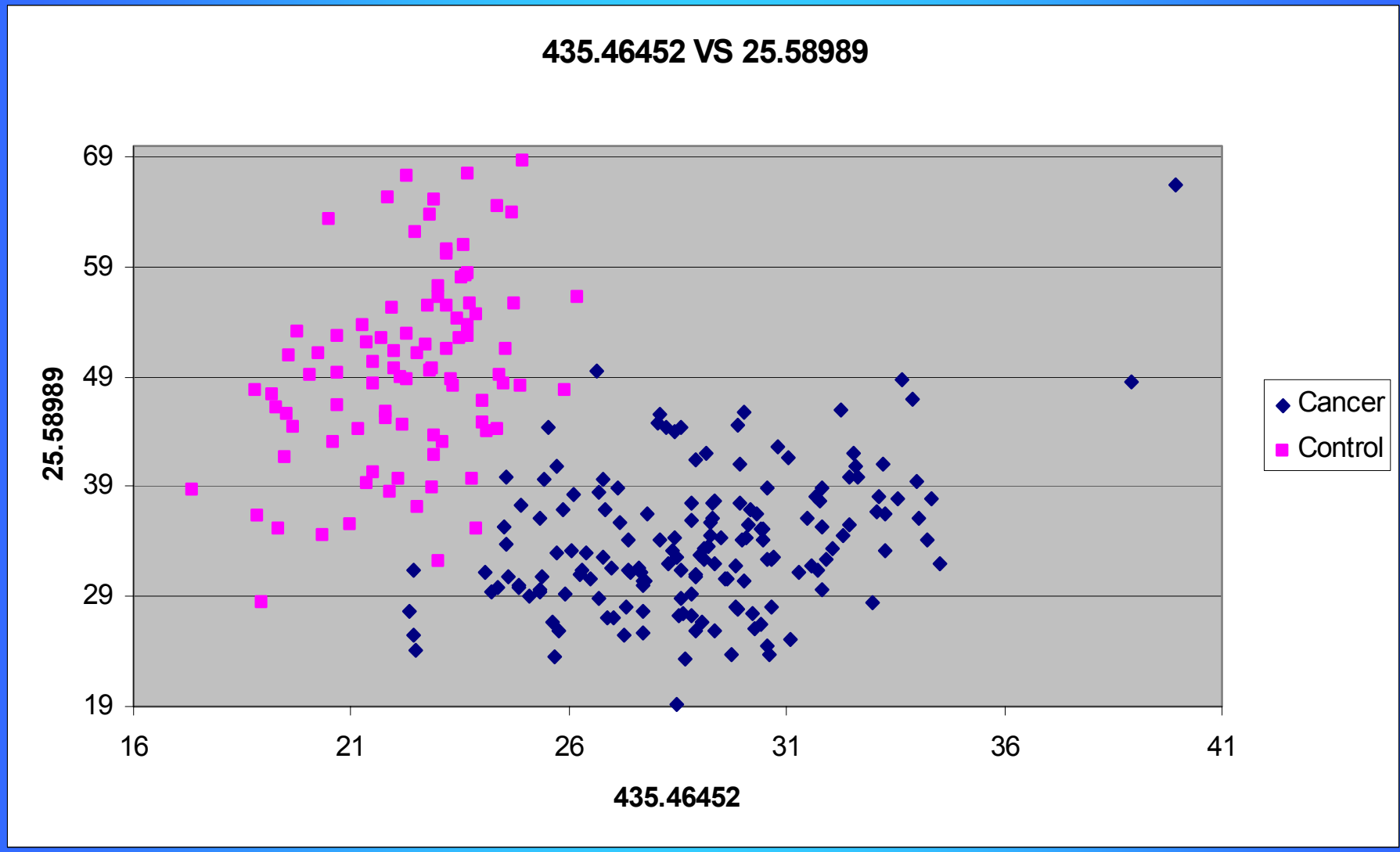
Biology or Bias? 1



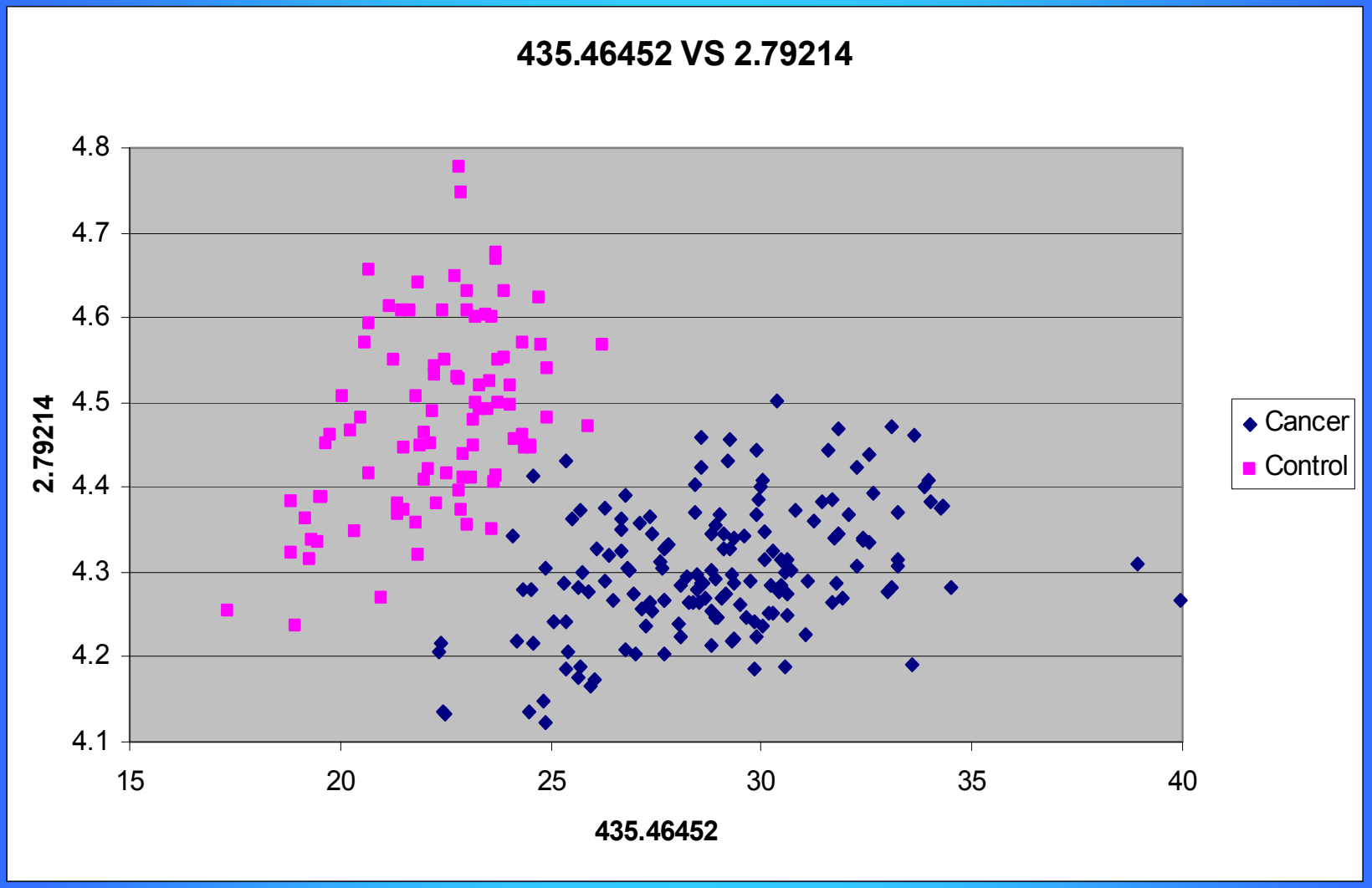
Biology or Bias? 2



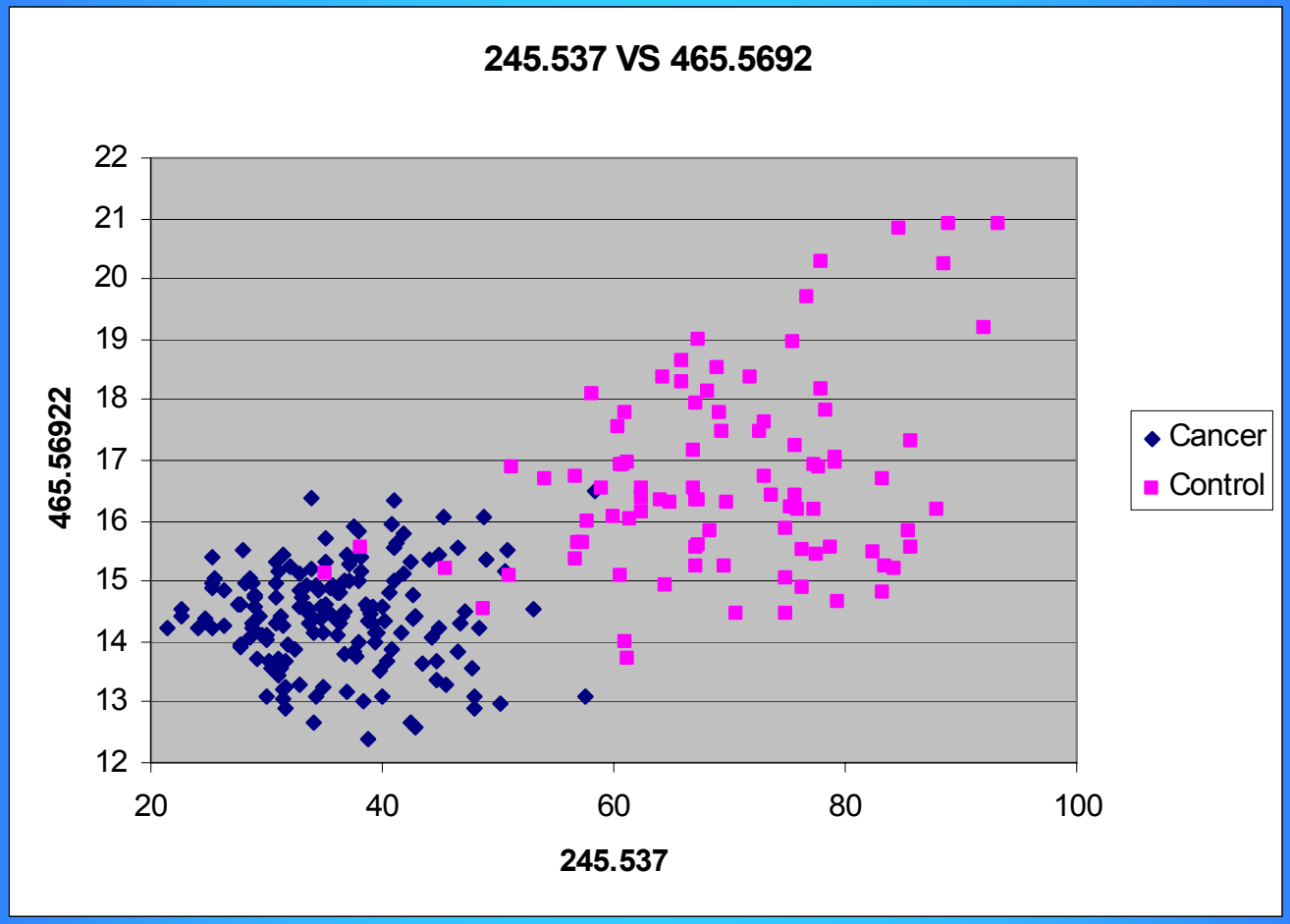
Biology or Bias? 3



Biology or Bias? 4



Biology or Bias? 5



Conclusions

- Data set has an experimental bias, or at a minimum the identity of the molecules associated with M/Z values of 2.79 and 25.59 must be established to rule out a bias.
- Additional data for the control samples similar to that posted for the cancer samples (e.g. subject age, SELDI Chip ID) would be useful.
- More information on these datasets in subsequent talks.
- Cannot conclusively establish sensitivity and specificity.
- Low M/Z biomarkers may be hypothesized (partially consistent with some literature).
- If there are low M/Z tumor markers may need to rethink approach (sample fractionation, use Mass Spec methods that are more accurate in the low M/Z region).

Consider Protein Concentration 1

- First as published by Ele_f_therios P. Diamandis consider the role of protein concentration, and known “internal controls” (should the experiment detect a known biomarker or acute phase protein). See Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: Opportunities and potential limitations.
  Mol Cell Proteomics. 2004 Feb 28 [Epub ahead of print]
  PMID: 14990683

Consider Protein Concentration 2

- Using current sample preparations most binding sites are occupied by a few prevalent proteins such as albumin.
- Consider spiking experiments (to define sensitivity) as well as sample fractionation.

Consider Acute Phase Proteins 1

- Acute Phase proteins are gene families that are known to be generally altered by any illness.
- Partial list includes: Amyloid P Component, Transferrin, Serum Albumin, Serum Amyloid A
- http://fred.hmc.psu.edu/ds/retrieve/fred/meshdescriptor/D000209

Consider Acute Phase Proteins 2

- Many known to be associated with malignancy.
- Haptoglobin is a classic example with changes noted in ovarian cancer (e.g. increased Haptoglobin-alpha subunit).
- May also be viewed as an internal control (i.e. with a given design should they have been detected).
- Alterations in post-translational modifications have been reported.

Consider Acute Phase Proteins 3

- Very difficult to know how other prevalent disease (e.g. hepatitis C, HIV) might influence any of these results.
- Haptoglobin itself is altered by Hepatitis B virus among others.
- Studies with small sample sizes in the non-cancer population will tend to overestimate specificity as the confounding disease(s) are not adequately represented.

Consider Acute Phase Proteins 4

- Collect a panel of samples from patients with a variety of prevalent non-malignant disease (also visit the ICU).
- Determine the differential diagnosis of a given pattern.
- Develop diagnostic algorithms that include non-malignant disease.

Consider Biological Response Modification

- Improve diagnostic accuracy by using a therapeutic challenge of an anti-neoplastic to increase biomarker signal.
- May enhance any biomarker protein or nucleic acid based, thus allowing a coordinated approach.
- May be linked to any type of therapy but recently developed anti-growth factor receptor and anti-angiogenesics may have greatest safety margin.

The Pre-analytic Challenge

- There are a great many causes of variation that may influence a clinical assay including:
- Exercise – increase free fatty acids, and muscle enzymes
- Diet – glucose etc.
- Drugs and alcohol
- Posture can result in a difference in protein concentration of 8% to 10 %, and a 2-fold change nor-epinephrine concentration
- Tourniquet time

Age

- Data reported in our paper found on http://ncifdaproteomics.com/methods.php
- Given that experimental bias prevents the assignment of specificity or sensitivity, one should not assume that age isn't a confounding variable.

The Pre-analytic Challenge

- Very little data regarding how these influence Mass Spec approaches.
- Studies in this area are not glamorous but are needed.
- Hard to control for in small scale studies, usually require carefully controlled multi-center trials, if not actual implementation, before these factors are fully understood.
- A positive test does not equal a known disease!

Summary

- In interpreting these results the use of readily available statistical methods are mandatory.
- Always include the noise region.
- Considerable need to better understand pre-analytic causes of variation.