# The Complexities of The Basics of PCA (Part II)

Dan Graham, Ph.D.
NESAC/BIO
University of Washington

*An NIH National Center for Research Resources*
*(NIH grant EB-002027)*
**http://www.nb.uw.edu**

NIH National Institute of Biomedical Imaging and Bioengineering

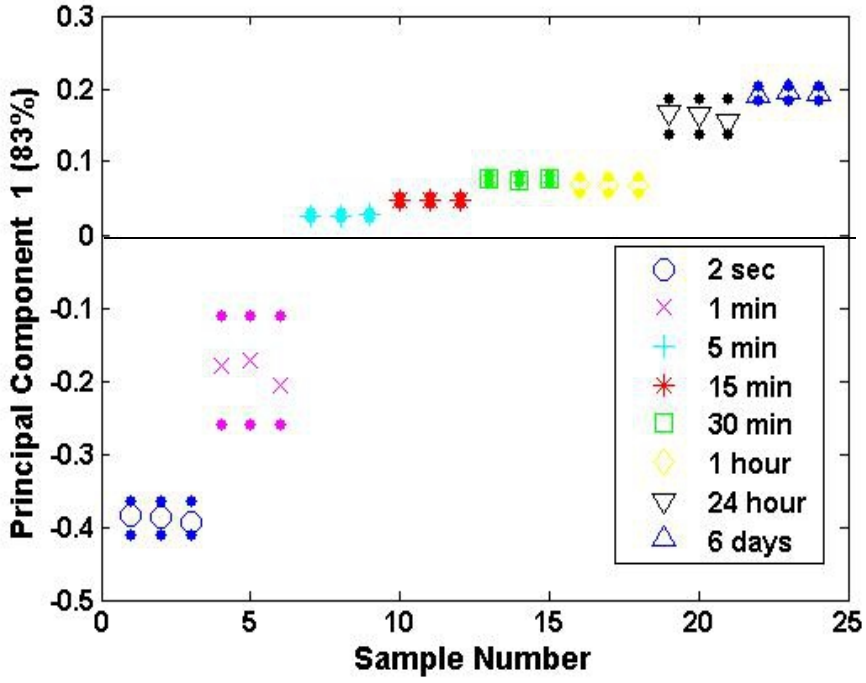
NESAC/BIO

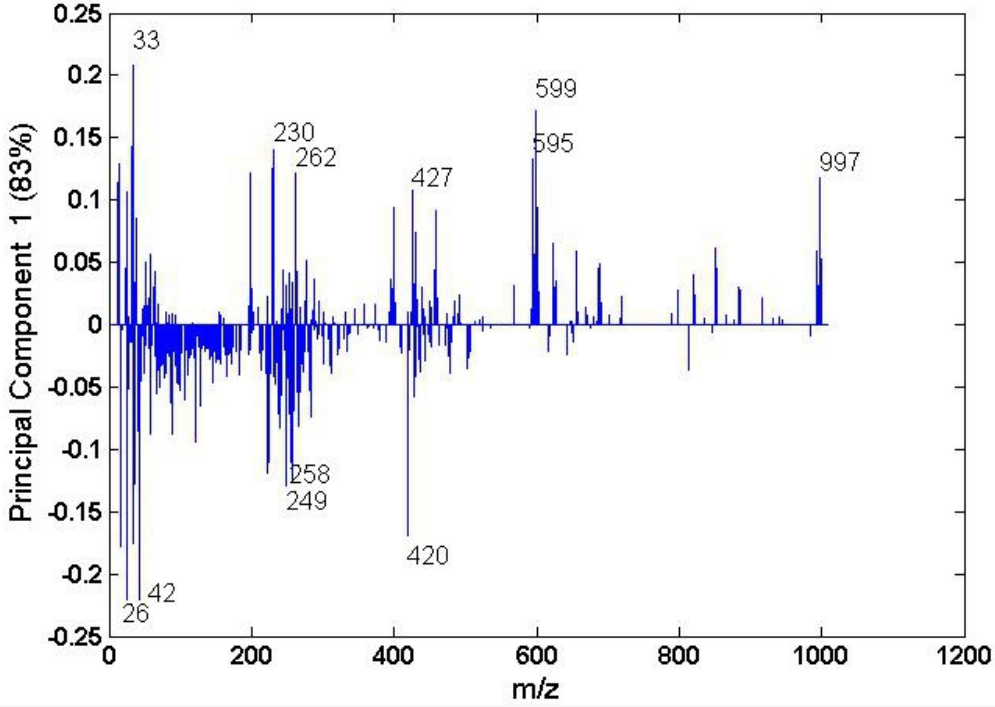
# Overview

- **Some basic complexities to think about for MVA of spectra**
- **Some basic complexities to think about for MVA of images**
- **Some examples of how MVA can be helpful**

# Raw Facts About Raw Data

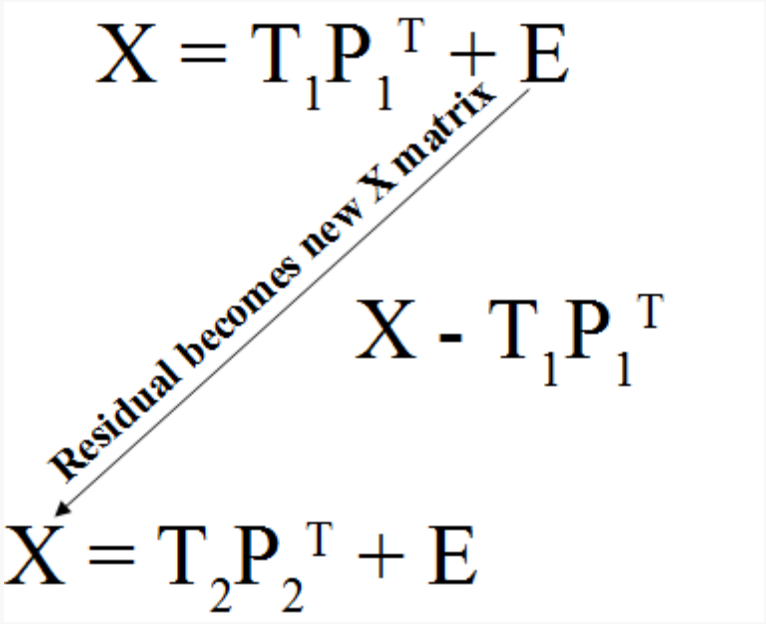
- **MVA methods are simply tools to help digest and analyze data**
- **It is always recommended to go back to the original data to verify any trends**
- **Most MVA methods are carried out on preprocessed data, not the original raw data**



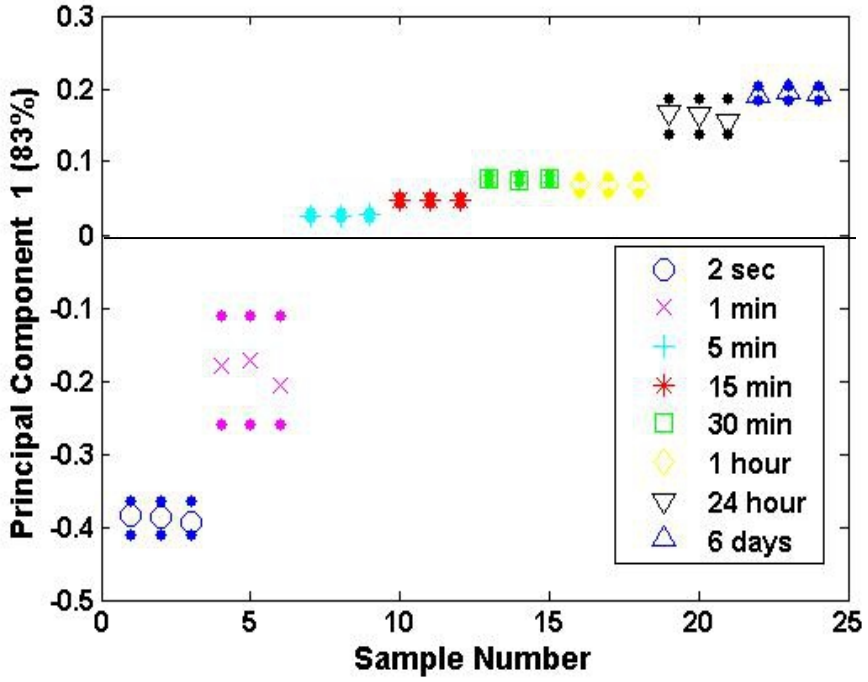



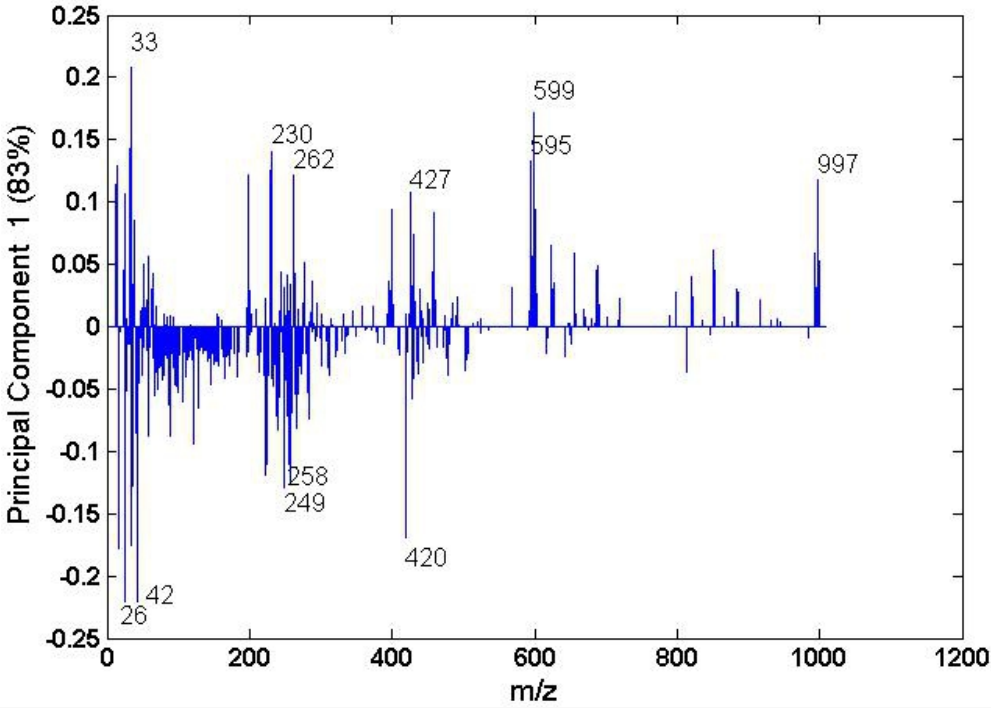
# Which "Raw" Data Should We Look At?

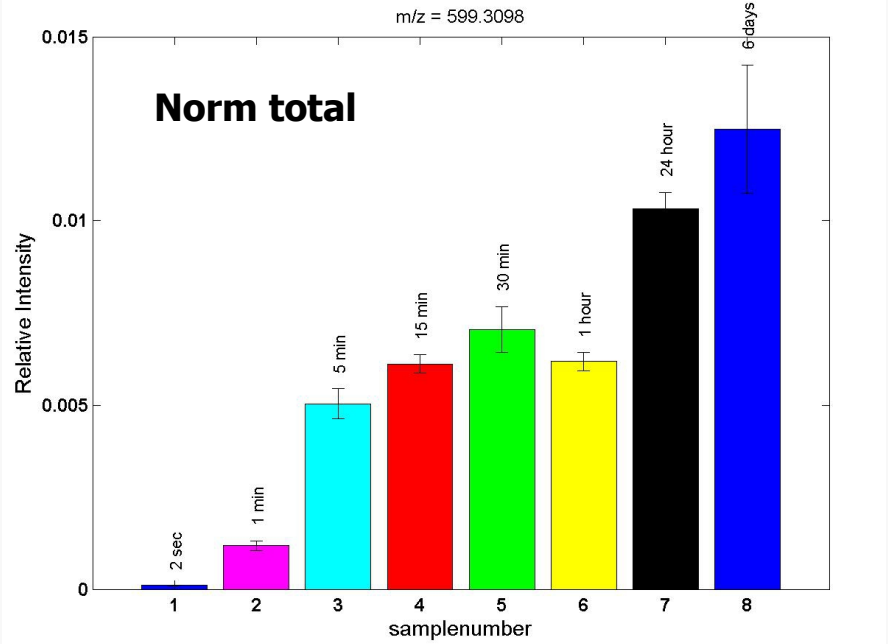
- **The original normalized counts?**
- **The preprocessed counts? (PCA is typically carried out on data that has been centered or scaled)**
- **What should be done when looking at PCs after PC1?**

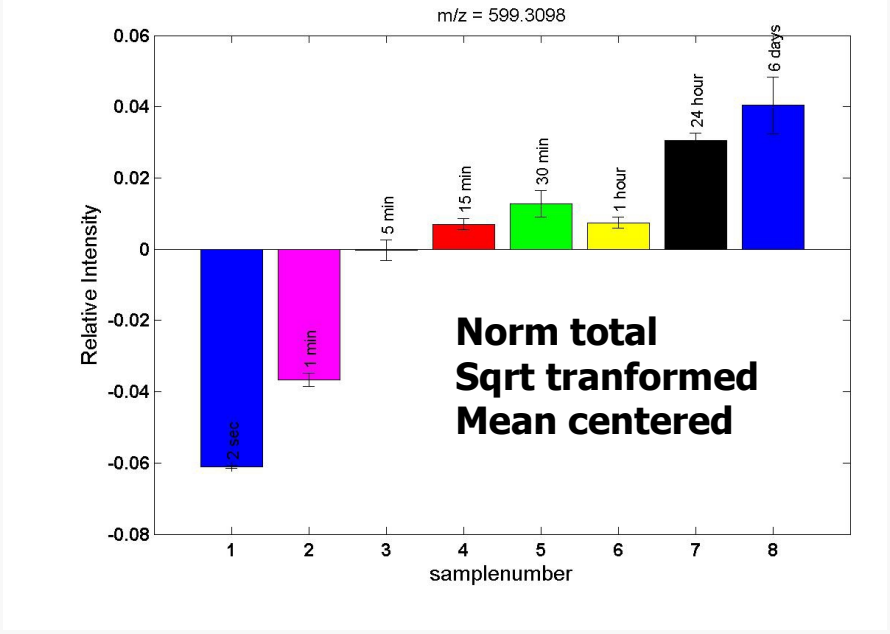


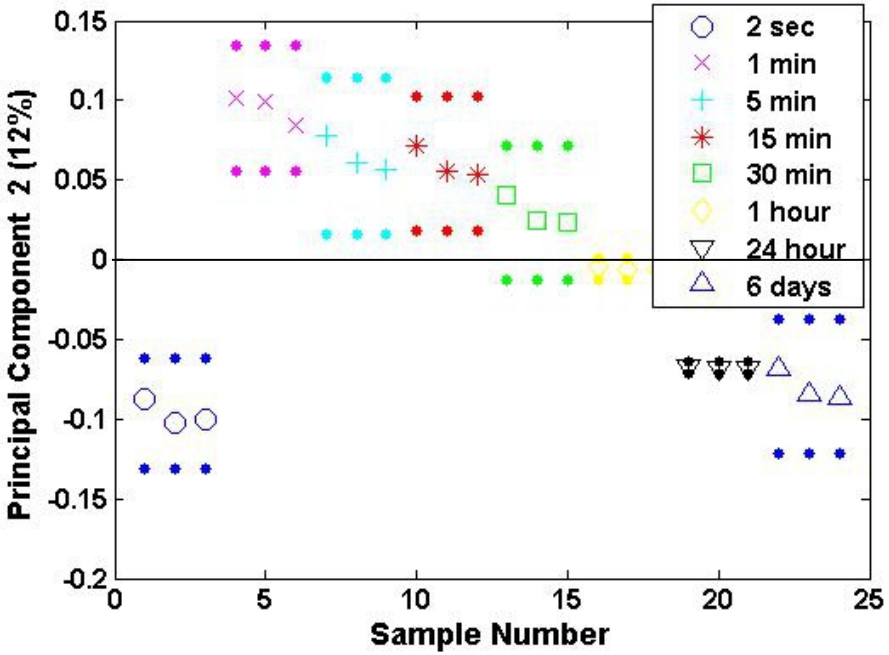

Each subsequent PC is calculated after subtracting the previous PCs from the data matrix. So what is the "raw" data for these PCs?

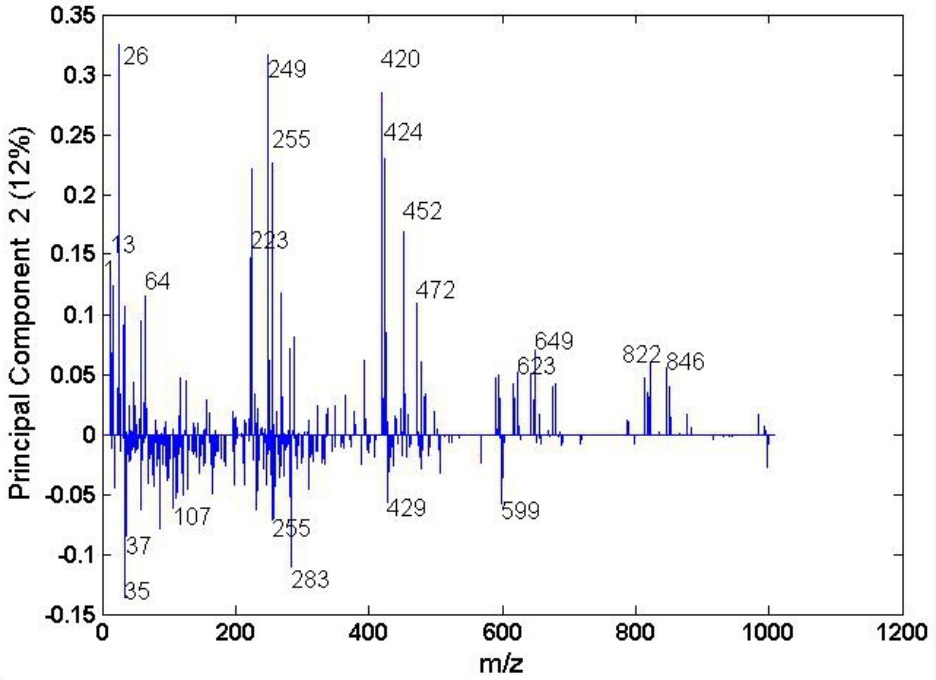

Principal Component 1 (83%)
Sample Number
2 sec
1 min
5 min
15 min
30 min
1 hour
24 hour
6 days

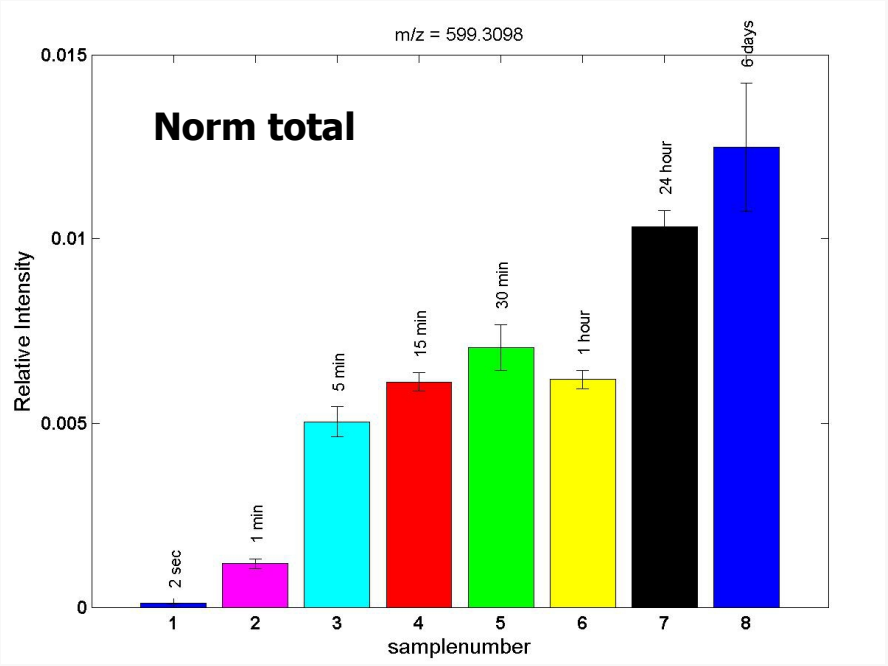

Principal Component 1 (83%)
m/z
33
599
230
262
595
427
997
258
249
420
42
26

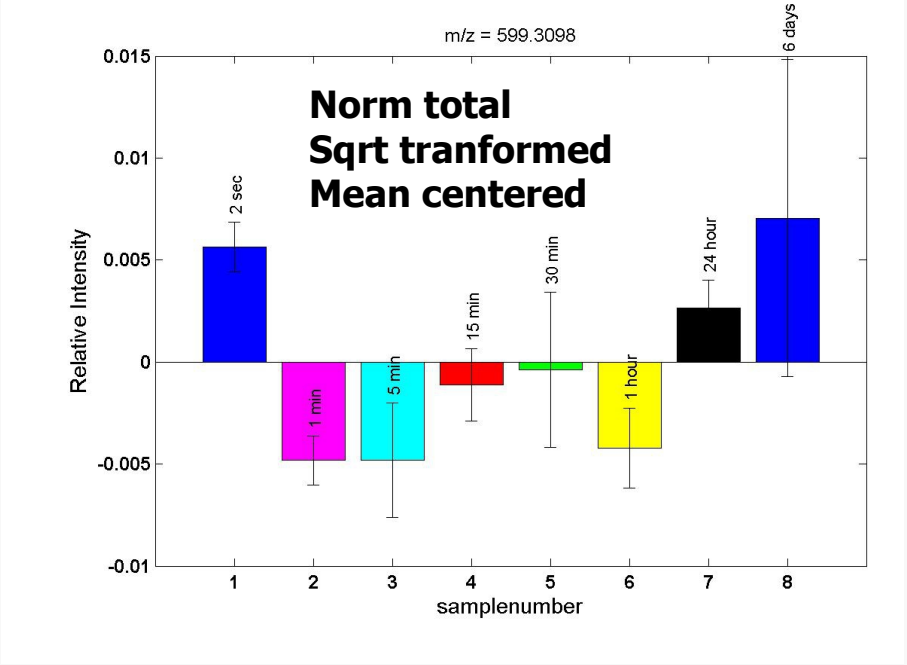

m/z = 599.3098
Norm total
Relative Intensity
samplenumber
2 sec
1 min
5 min
15 min
30 min
1 hour
24 hour
6 days


m/z = 599.3098
Norm total
Sqrt tranformed
Mean centered
Relative Intensity
samplenumber
2 sec
1 min
5 min
15 min
30 min
1 hour
24 hour
6 days

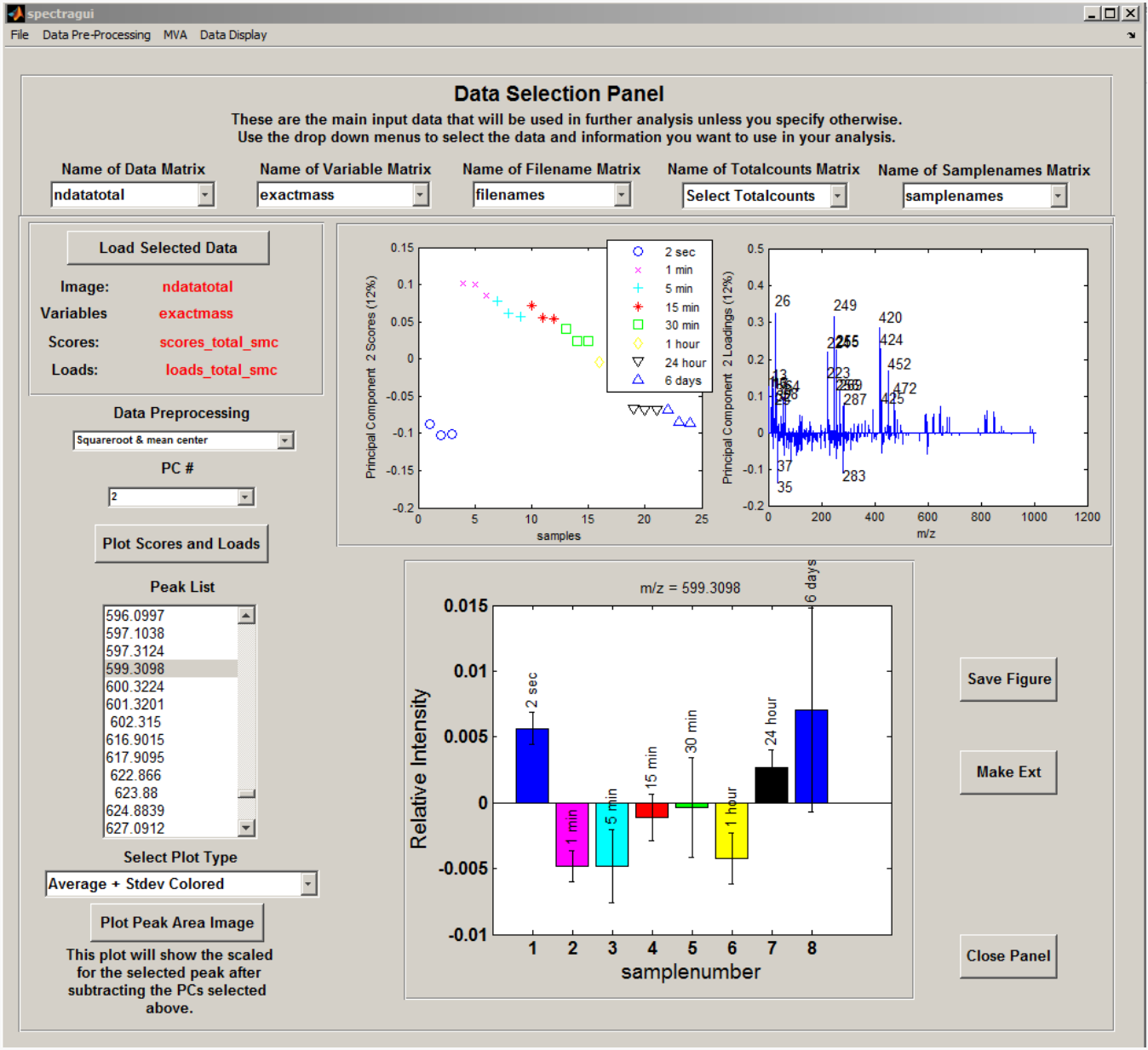

Principal Component 2 (12%)
Sample Number
2 sec
1 min
5 min
15 min
30 min
1 hour
24 hour
6 days


Principal Component 2 (12%)
m/z
26
249
420
255
424
452
13
223
64
472
649
623
822
846
107
429
599
37
255
283
35

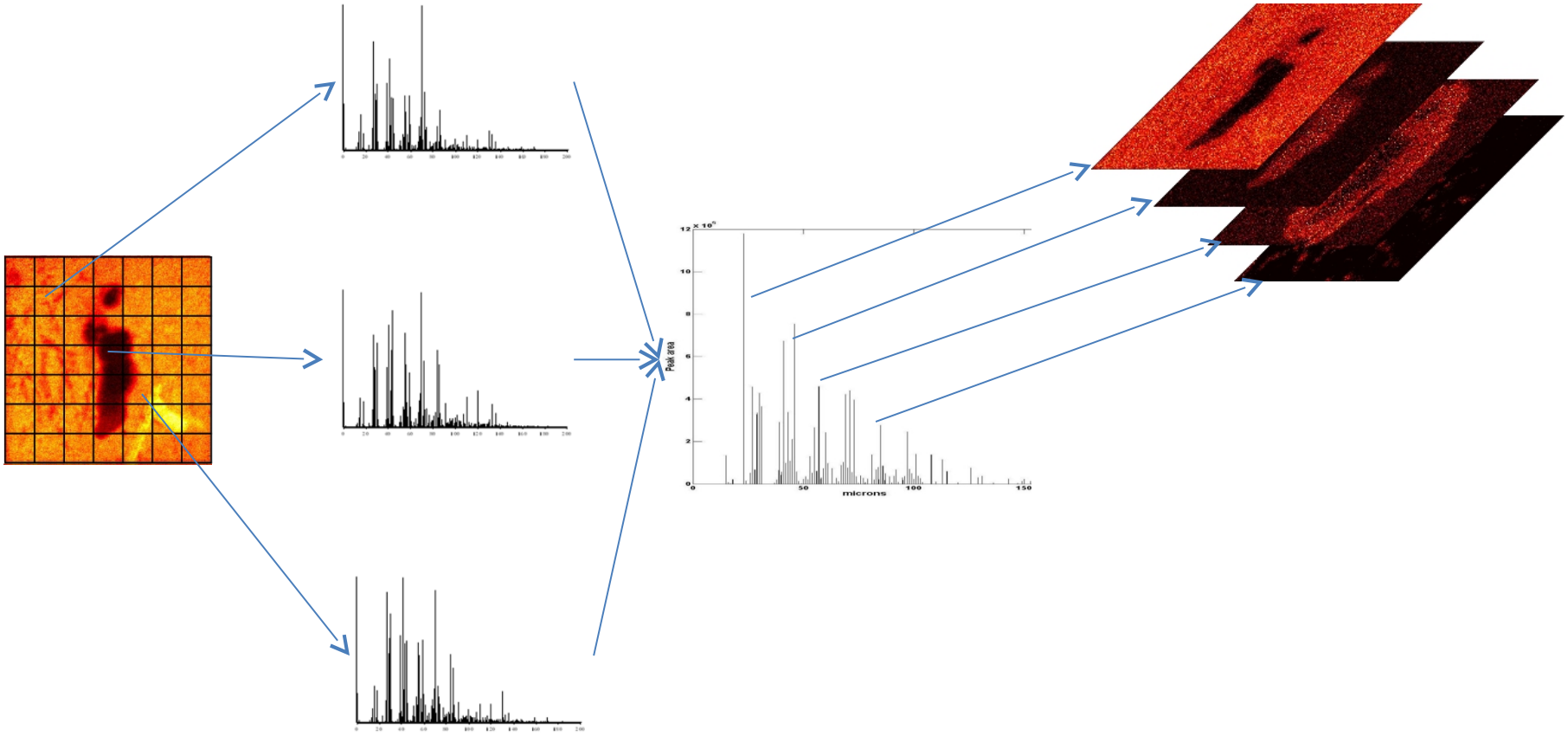

m/z = 599.3098
Norm total
Relative Intensity
samplenumber
2 sec
1 min
5 min
15 min
30 min
1 hour
24 hour
6 days


m/z = 599.3098
Norm total
Sqrt tranformed
Mean centered
Relative Intensity
samplenumber
2 sec
1 min
5 min
15 min
30 min
1 hour
24 hour
6 days

- **Recommended to look at preprocessed data for a given peak from a given PC**
- **NBToolbox has a function for this**

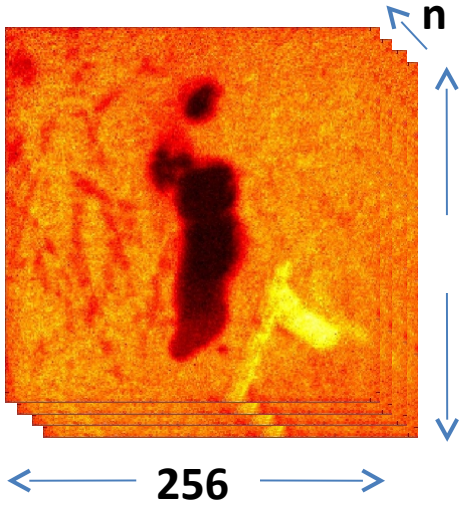


# ToF-SIMS Imaging Data



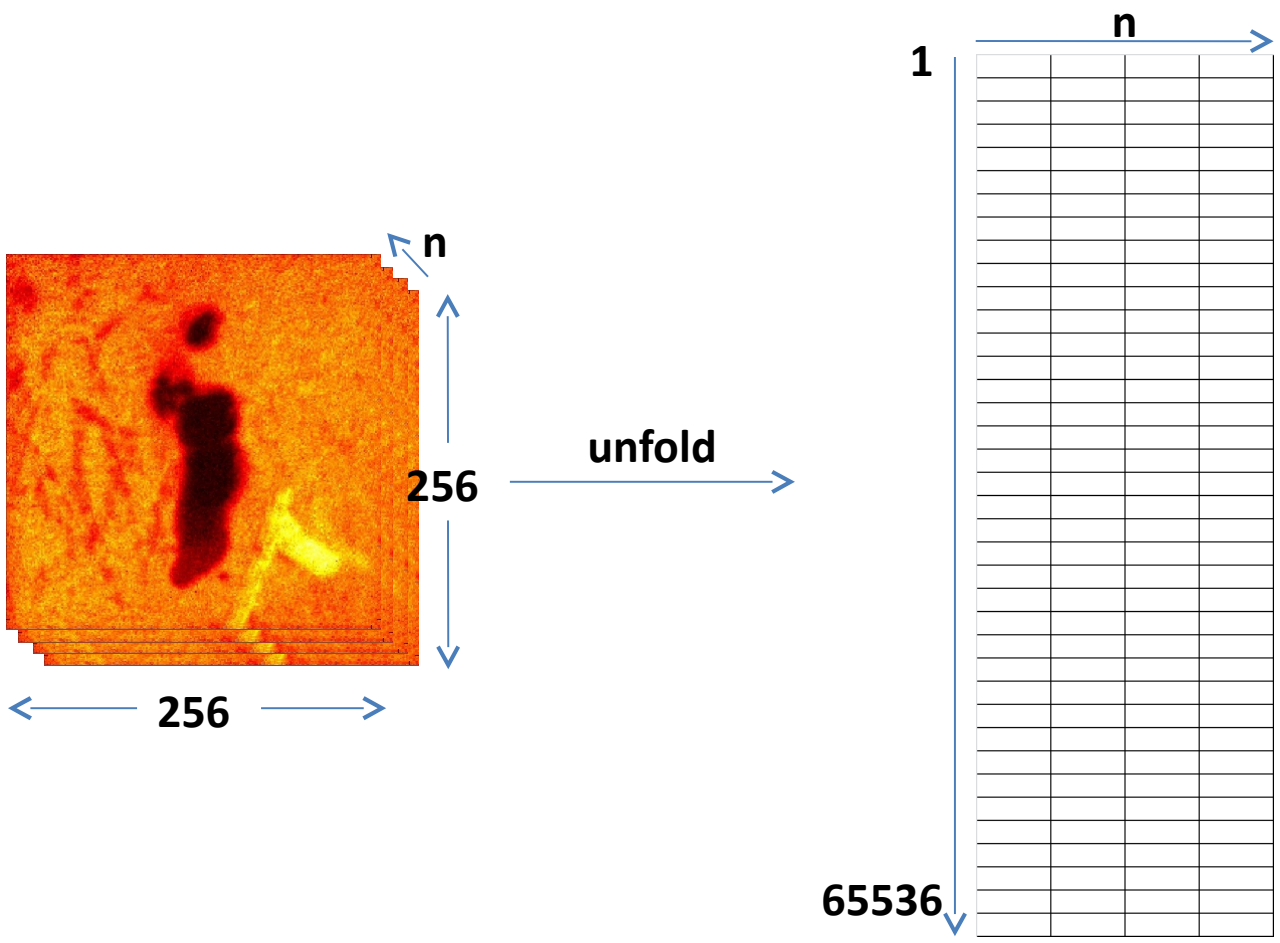
# ToF-SIMS Imaging Data

**A typical image is collected**
**At 256 x 256 pixels x n peaks**

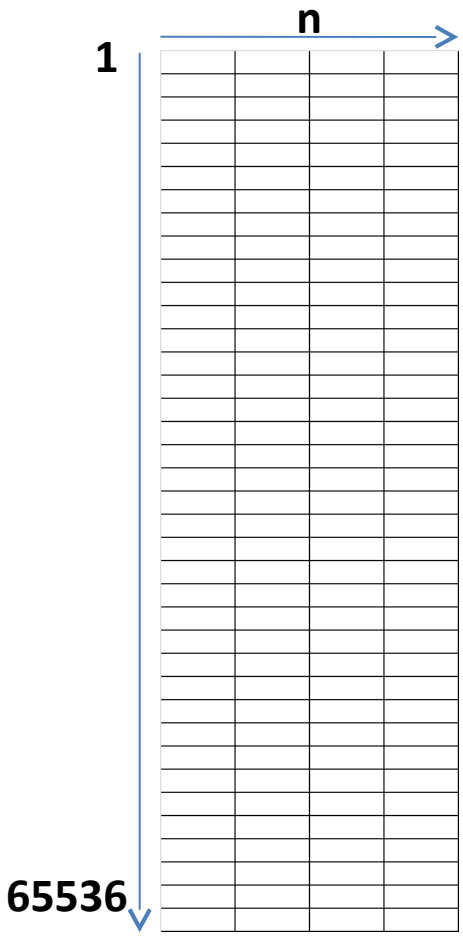



**= 65536 spectra x n peaks = Lots of data**

# ToF-SIMS Imaging Data: MVA



# ToF-SIMS Imaging Data: MVA

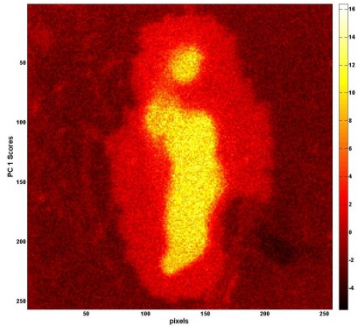



Pre-Processing

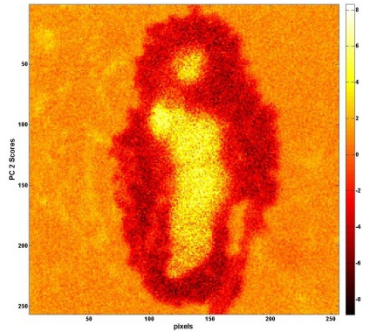
**MVA**

Re-fold

PC1 scores

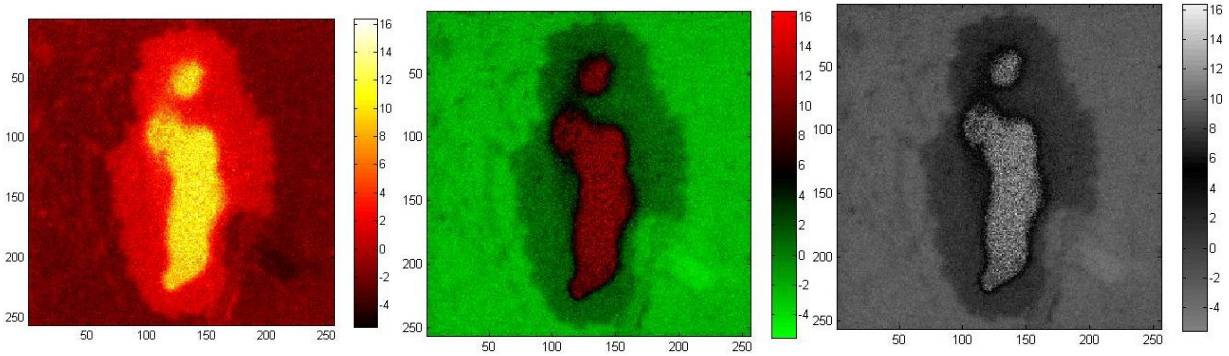


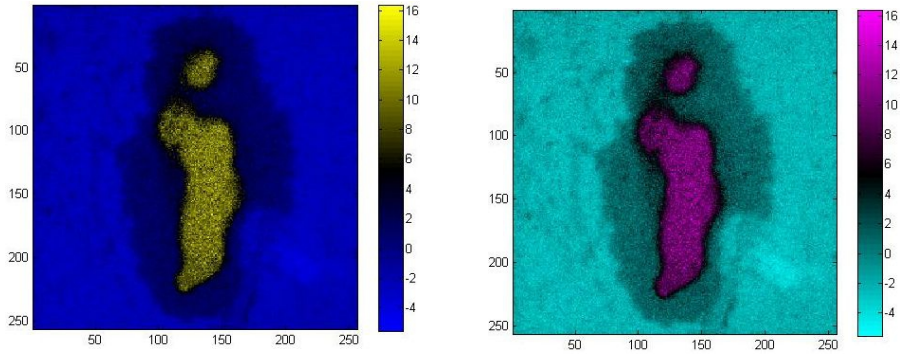
PC2 scores

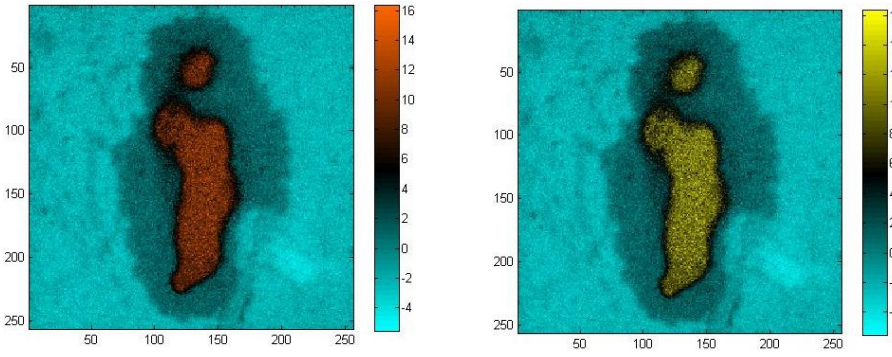


# ToF-SIMS Imaging Data: Displaying Scores

- **Use strong contrasting colors**
- **Best if 0 = black**
- **Greyscale can avoid color interpretation issues**
- **Provide color scale bar**



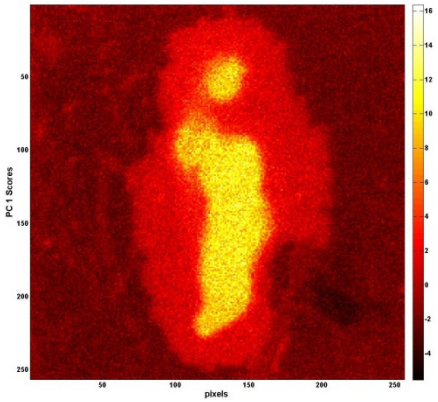



# ToF-SIMS Imaging Data: Displaying Scores

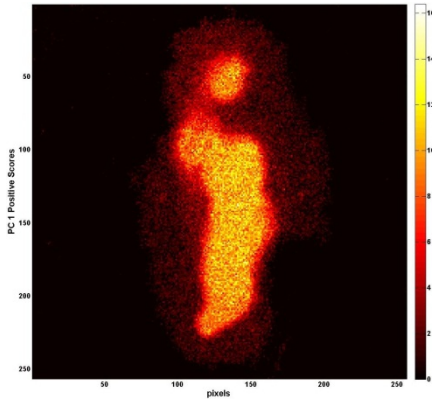
- Display positive and negative scores separately
- Must clearly label what is being displayed
- Valid mathematically as long as you keep track of the sign in both the scores and loadings plots

**Original PC1 Scores**

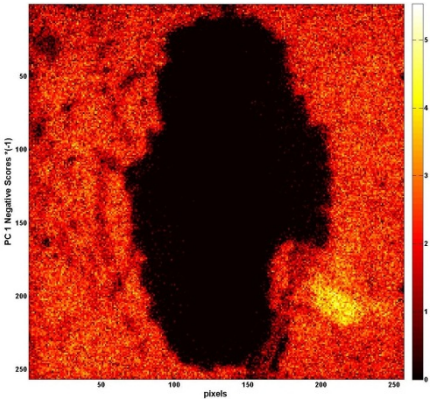



**Multiplied by -1 in order to display negative loadings from dark to light color**

**PC1 Positive Scores**



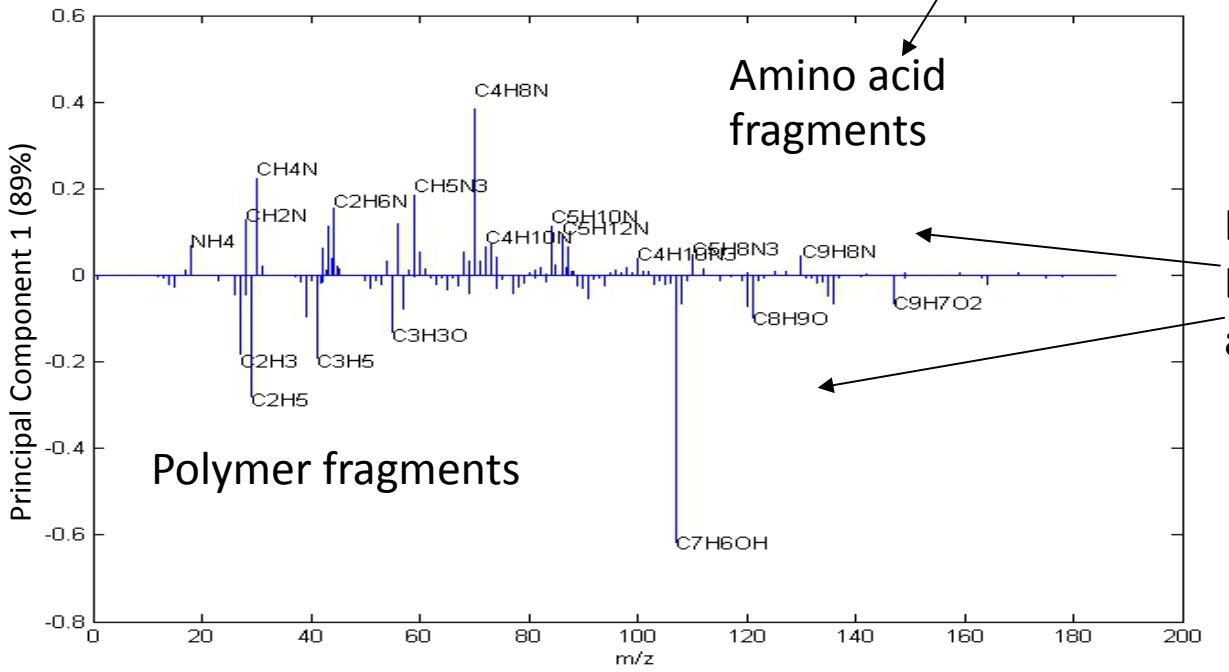

**PC1 Negative Scores (*-1)**



## ToF-SIMS Imaging Data:
## PCA Loadings

**% variance is shown**

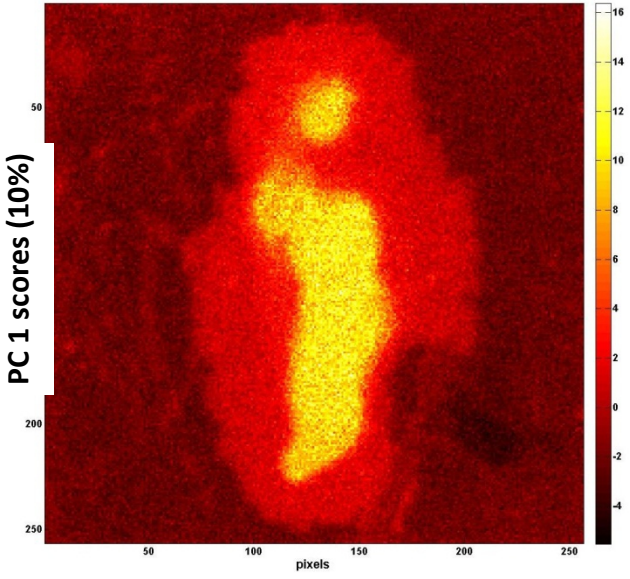
**Descriptor is added to highlight major differences**



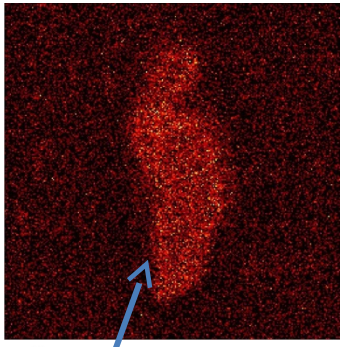

**Highest Loadings are labeled**

**Loadings are plotted versus m/z**

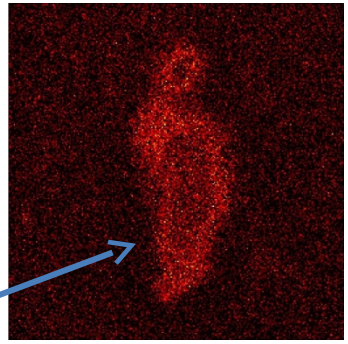
# ToF-SIMS Imaging Data:
# MVA – Interpreting Scores and Loadings

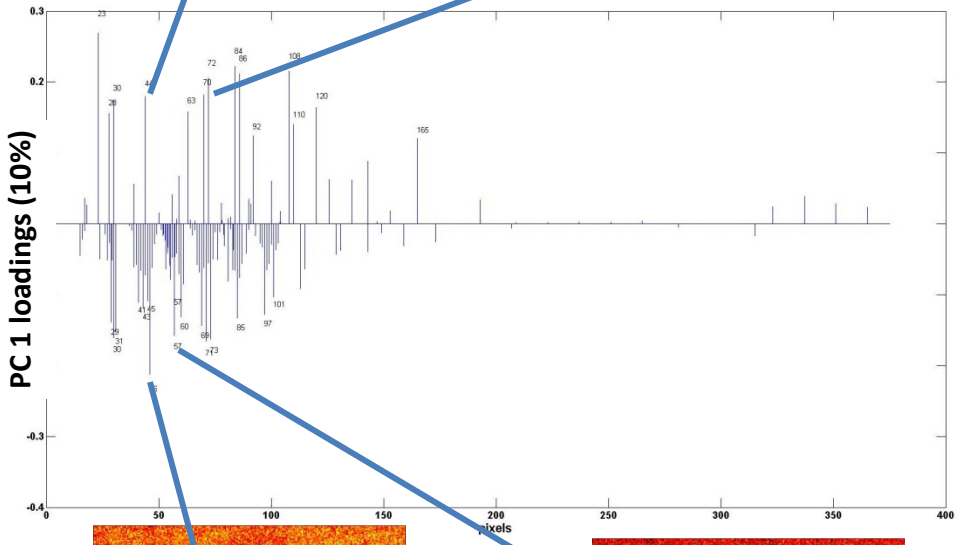



C2H6N

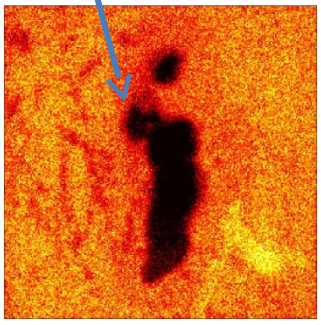


C4H8N

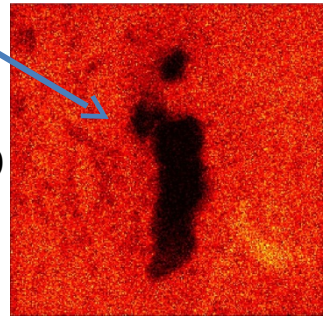





NO2



C3H5O

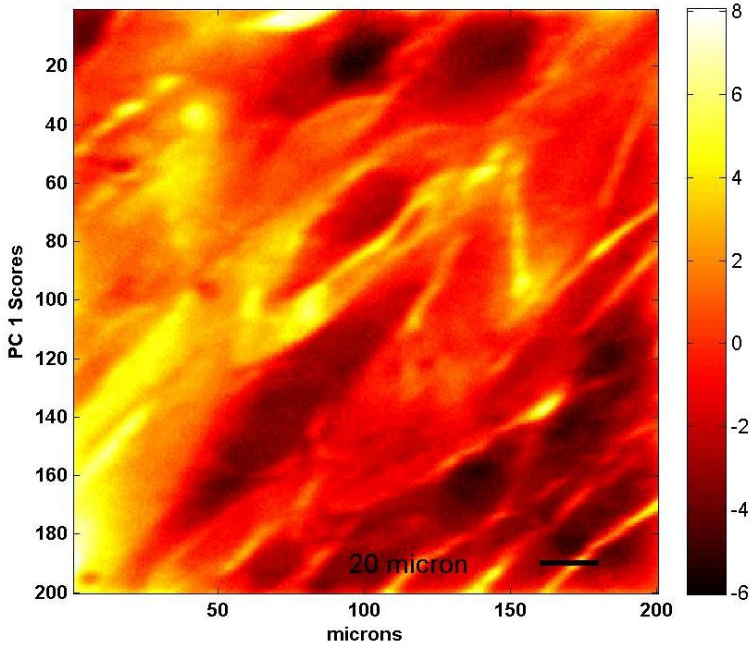


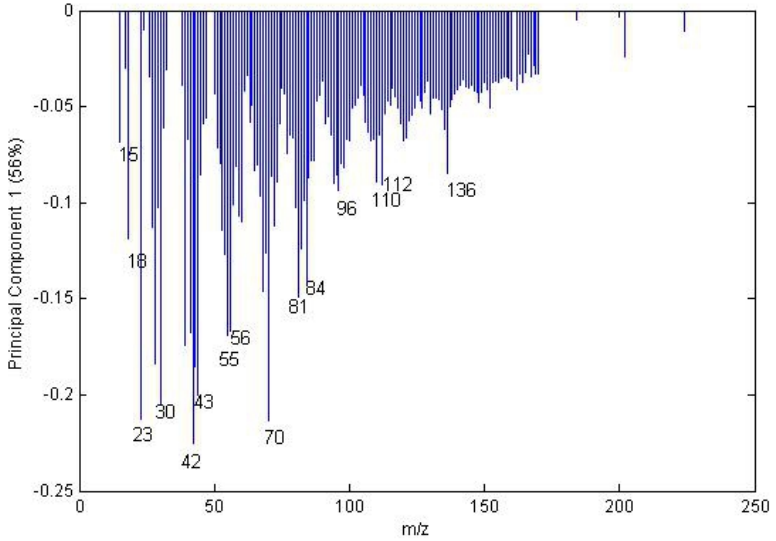
## Cell Imaging: Effects of data normalization

- Image data taken in high spatial resolution mode
- Image data summed from depth profile using 150 slices (to increase counts)
- All peaks above background selected (negative ion data)
- PCA of:
  - Poisson scaled/mean centered data
  - Normalized/Poisson scaled/mean centered data

# Cell Data:Poisson scaled/mean centered

- All loadings are on one side
  - All peaks show a higher relative intensity for the areas with negative scores
  - Can be due to charging or topography
  - Not what we are interested in
  - Can be fixed by normalization

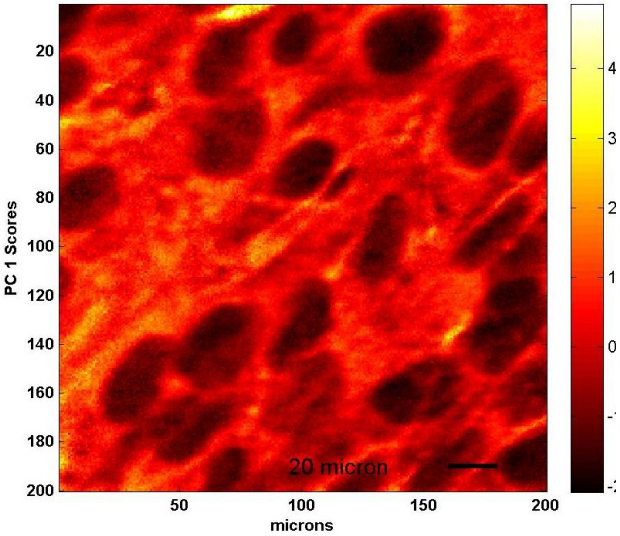


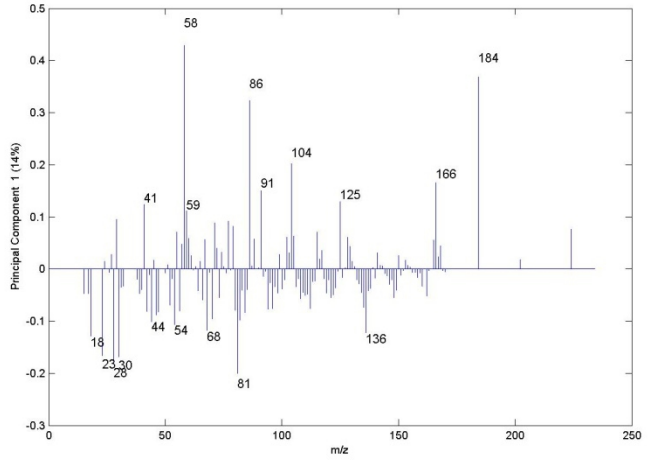



# Cell Data:Normalized/Poisson scaled/mean centered

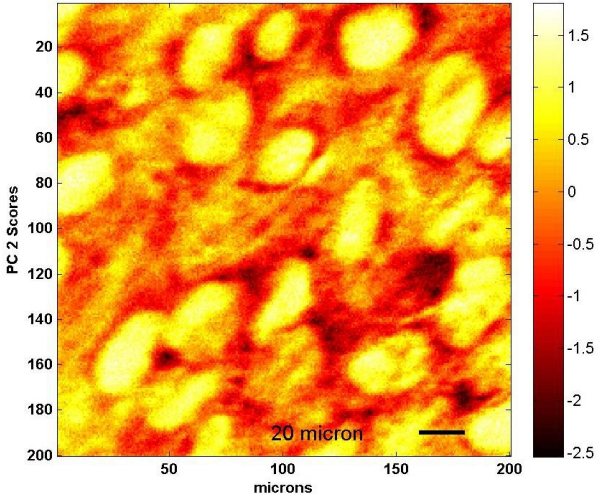
- After normalization to the total counts we see contrast due to cells
- PC1 now shows same contrast/information as PC2 from non-normalized data
  - Because we have normalized out the variance due to the total counts

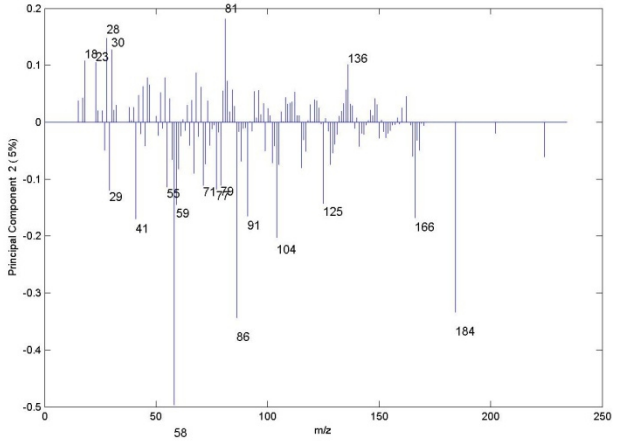
**PC1 normalized data**




**PC2 non-normalized data**



# A Note on Image Normalization

- Image normalization is sometimes helpful
- Should be used with care
  - Due to low count rates in SIMS images, normalization can
    - Accentuate noise
    - Possibly cause divide by zero errors if pixels have 0 counts

# Examples

- Spectra
  - ToF-SIMS analysis of proteins with various ion sources

Muramoto, et. al. J. Phys. Chem. C 2011, 115, 24247–24255

    - Determining protein orientation

Wang, Castner, B. Ratner, Jiang, Langmuir 20 (2004) 1877

    - Probing protein structure

Xia, May, McArthur and Castner, Langmuir, 18 (2002) 4090

    - PLS modeling to predict cell growth promotion

Analytica Chimica Acta, 1986, 185, 1-17

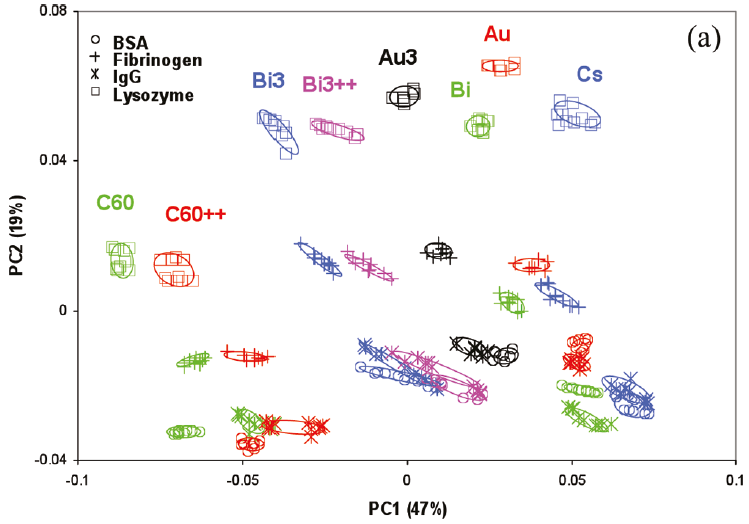
  - Imaging
    - Investigating cationization
    - PCA, MAF, MCR analysis of DNA microarray spots

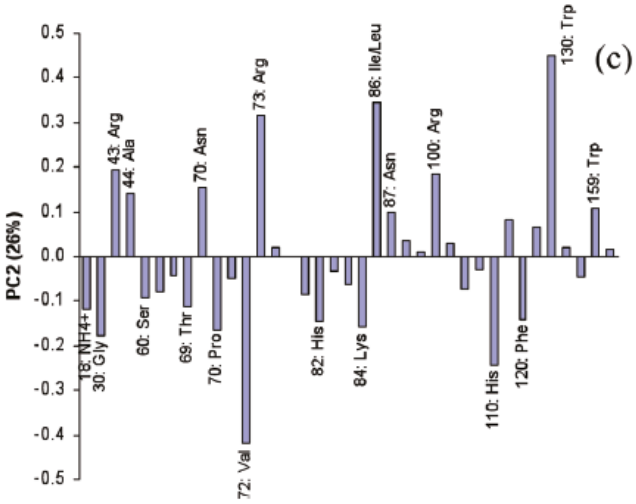
# Determining the Affects of Primary Ions

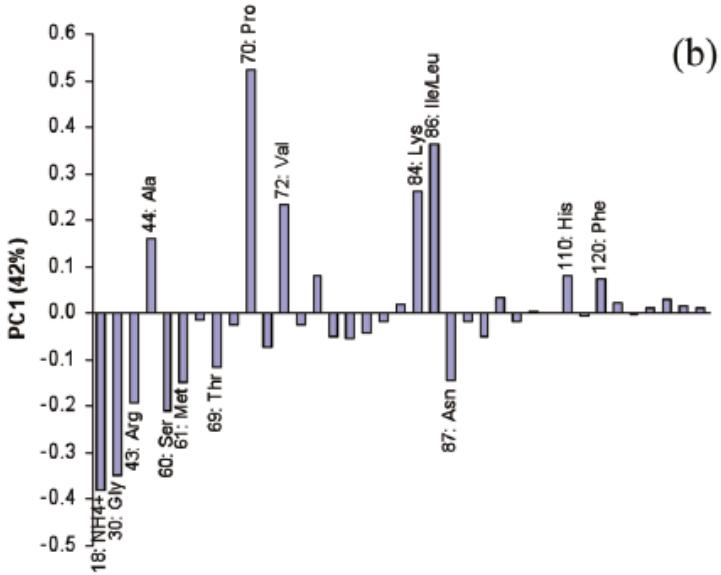
- Eight primary ion sources
- Four different proteins
- Selected AA related peaks
- Normalized to sum of selected peaks
- Mean centered

Muramoto, et. al. J. Phys. Chem. C 2011, 115, 24247–24255

# PCA Results



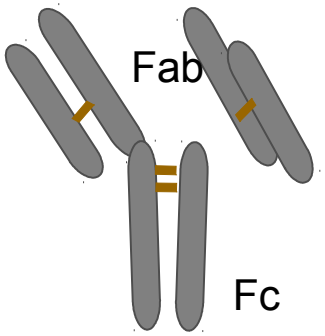



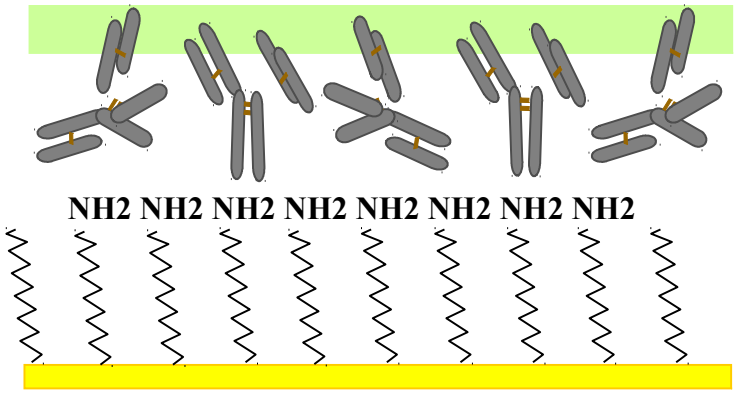

- Primary Ion caused largest variation
  - More fragmentation
    - Other trends were consistent between different primary ions
  - Important to keep primary ion consistent across data sets

Muramoto, et. al. J. Phys. Chem. C 2011, 115, 24247–24255

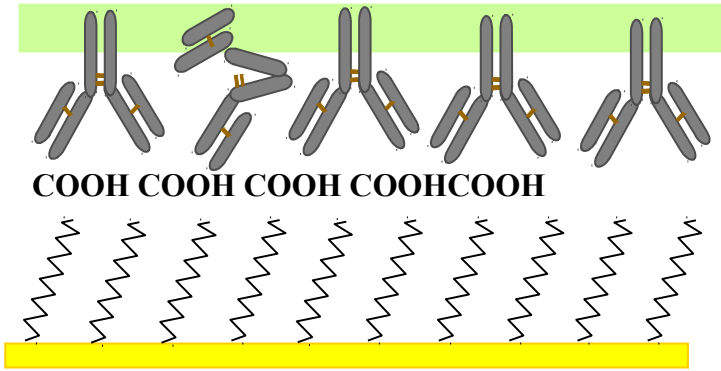
# ToF-SIMS Detection of IgG Orientation

- **Can SIMS be used to determine antibody orientation?**
- **Adsorb anti-hCG on self-assembled monolayers with different surface charges**
- **Anayze**
  - **Fab, Fc anti-hCG on Au**
  - **anti-hCG on SAMS**



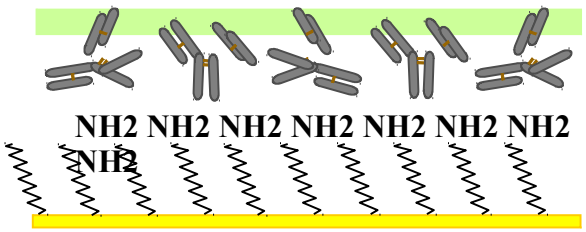

**anti-hCG on NH2- SAM**



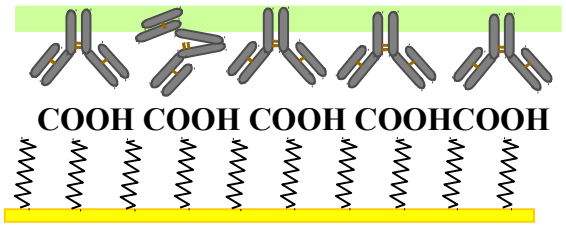

**anti-hCG on COOH- SAM**

Wang, Castner, B. Ratner, Jiang, *Langmuir* **20** (2004) 1877.

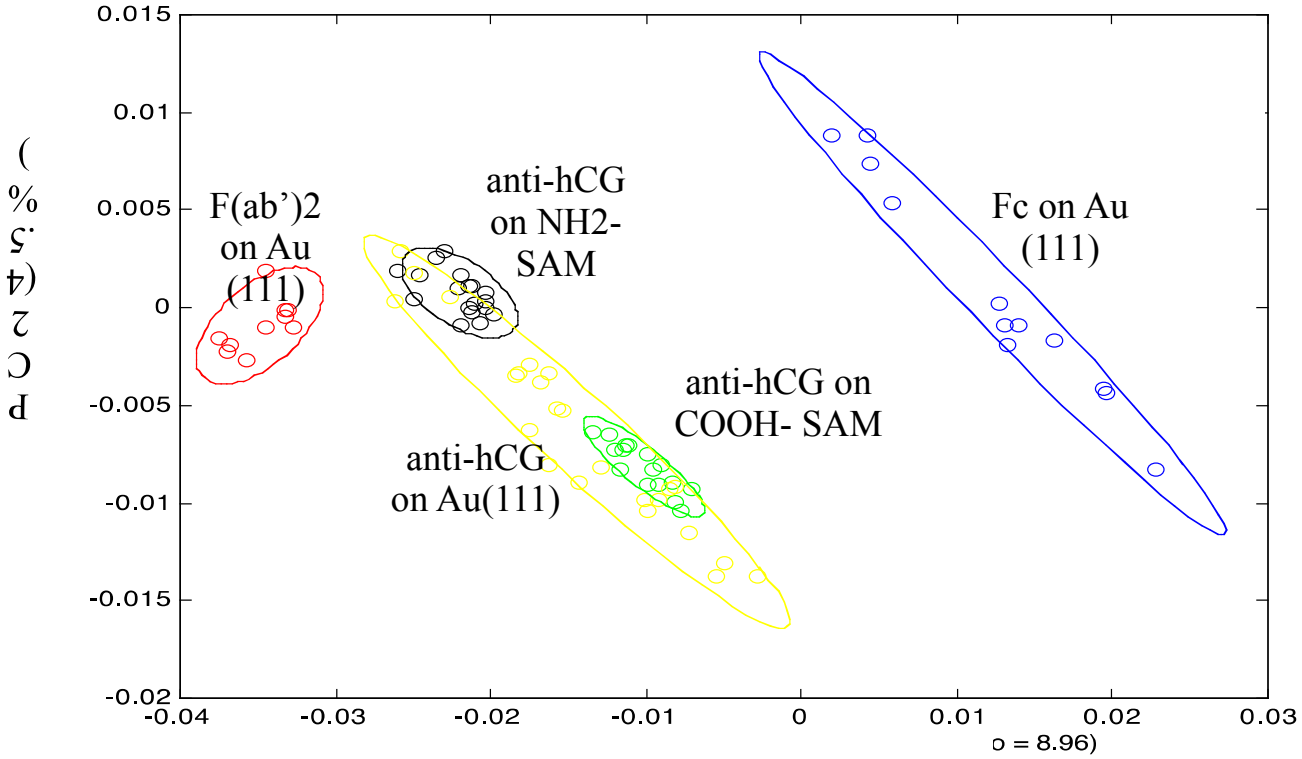
# ToF-SIMS Detection of IgG Orientation




anti-hCG on NH2- SAM

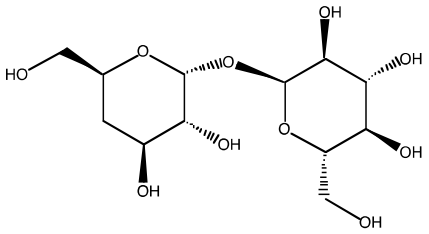



anti-hCG on COOH- SAM

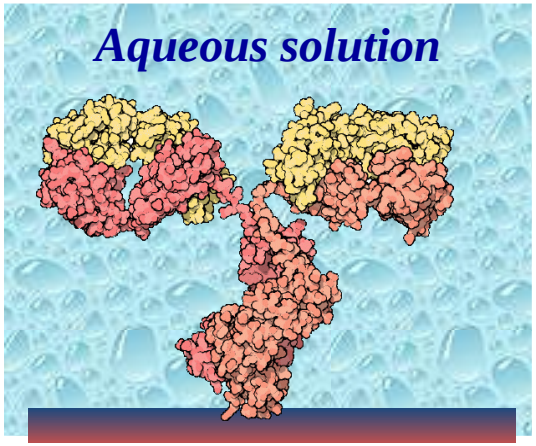


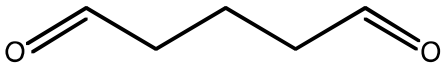

Wang, Castner, B. Ratner, Jiang, *Langmuir* **20** (2004) 1877.

# *Preserving Protein Structure*

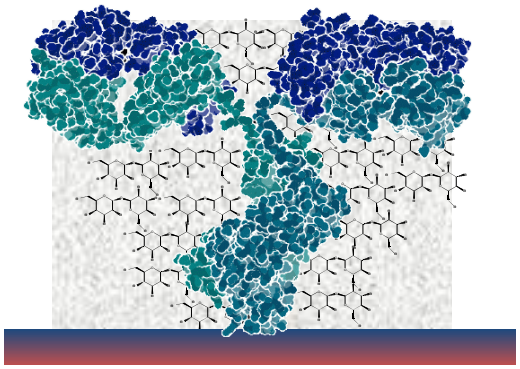


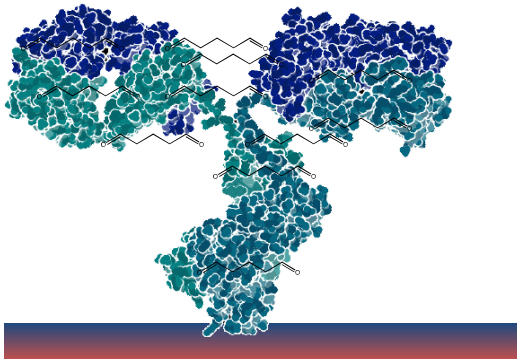

Trehalose




Glutaraldehyde

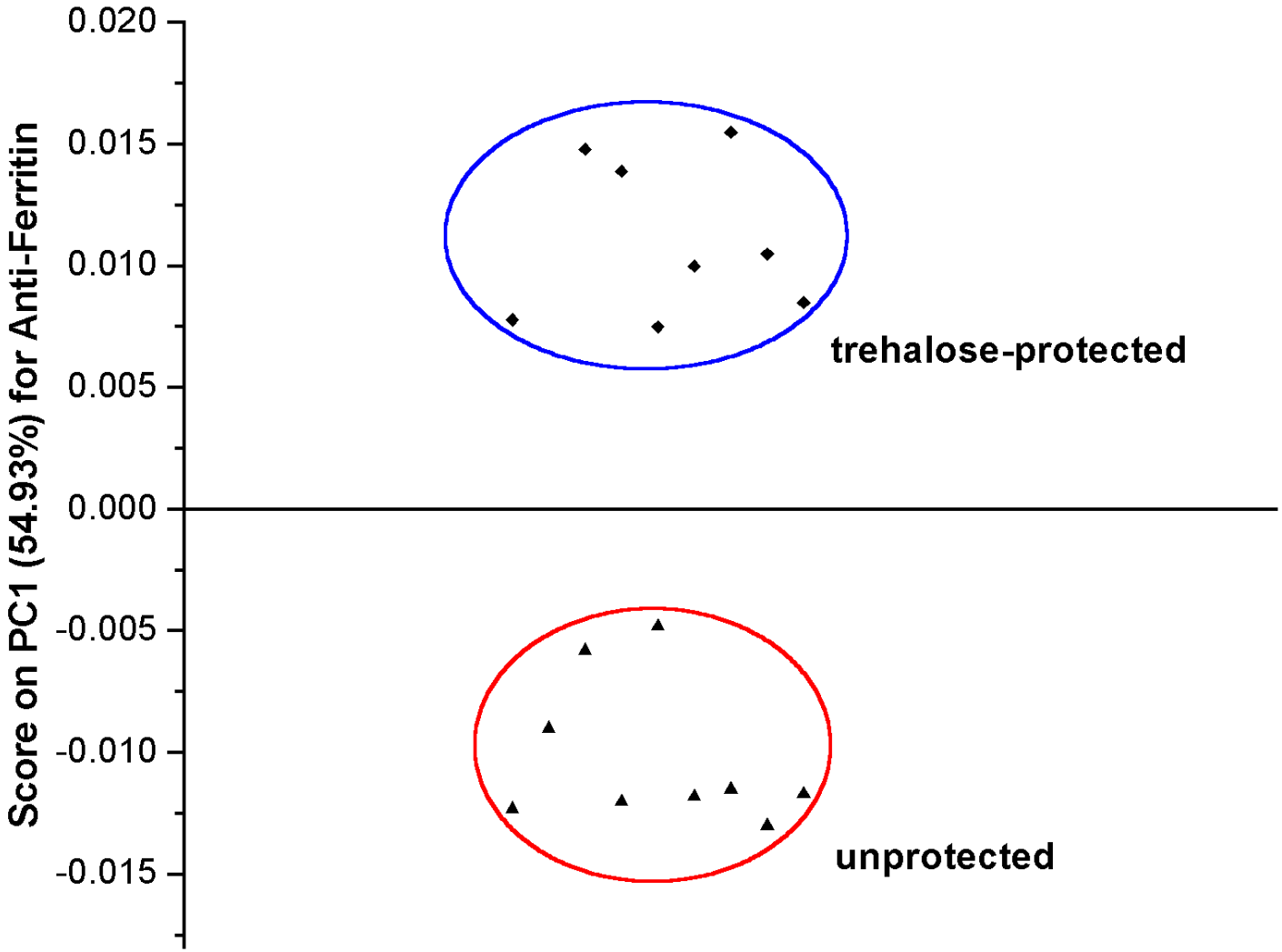




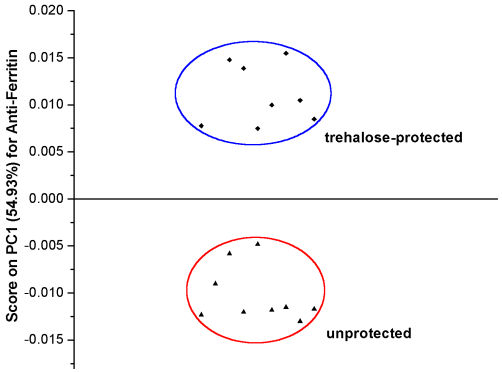
Xia, May, McArthur and Castner, Langmuir, **18** (2002) 4090.

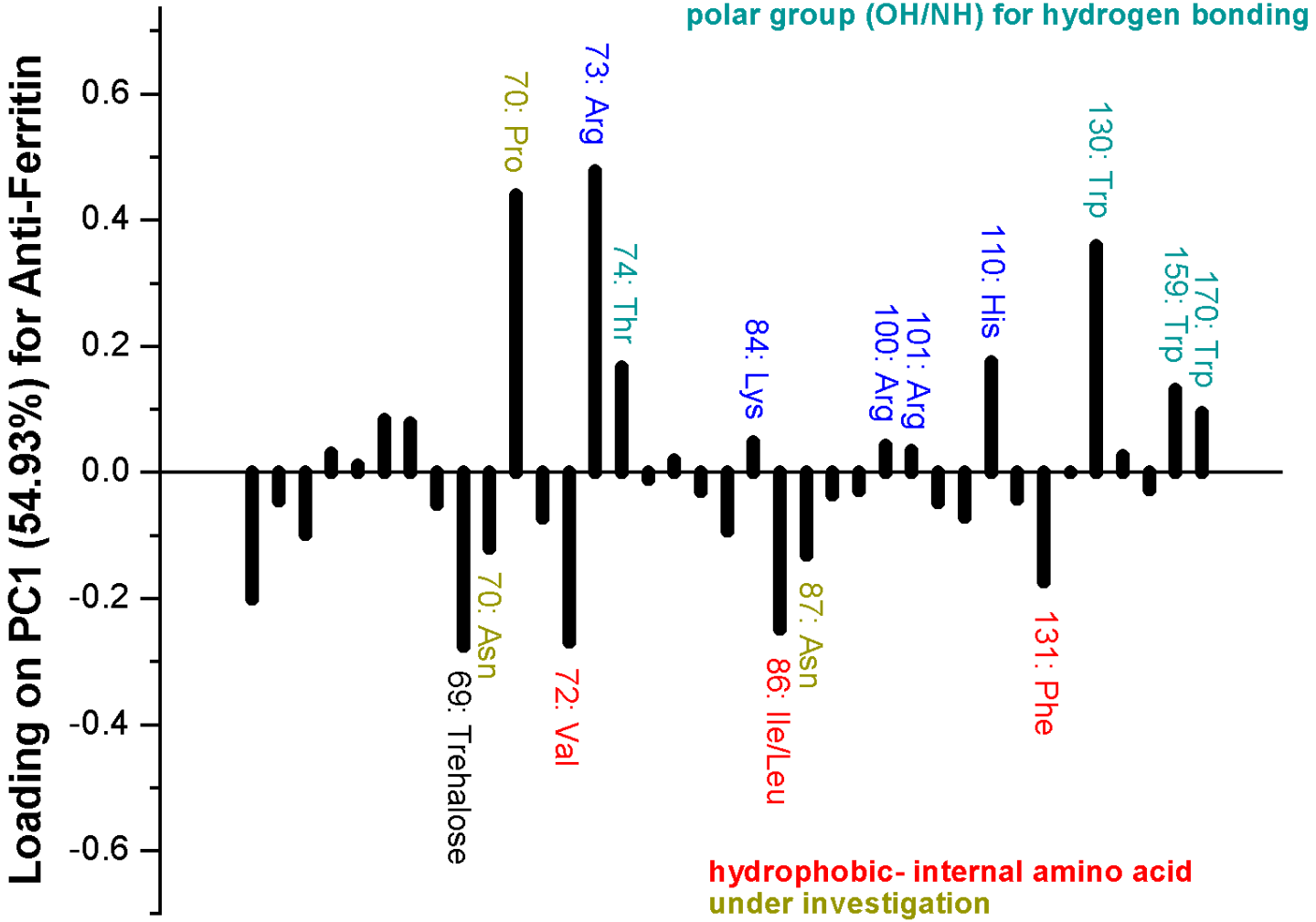
Xia & Castner, *JBMR* **67A** (2003) 179.

# *ToF-SIMS/PCA Results*



# PC1 Loadings



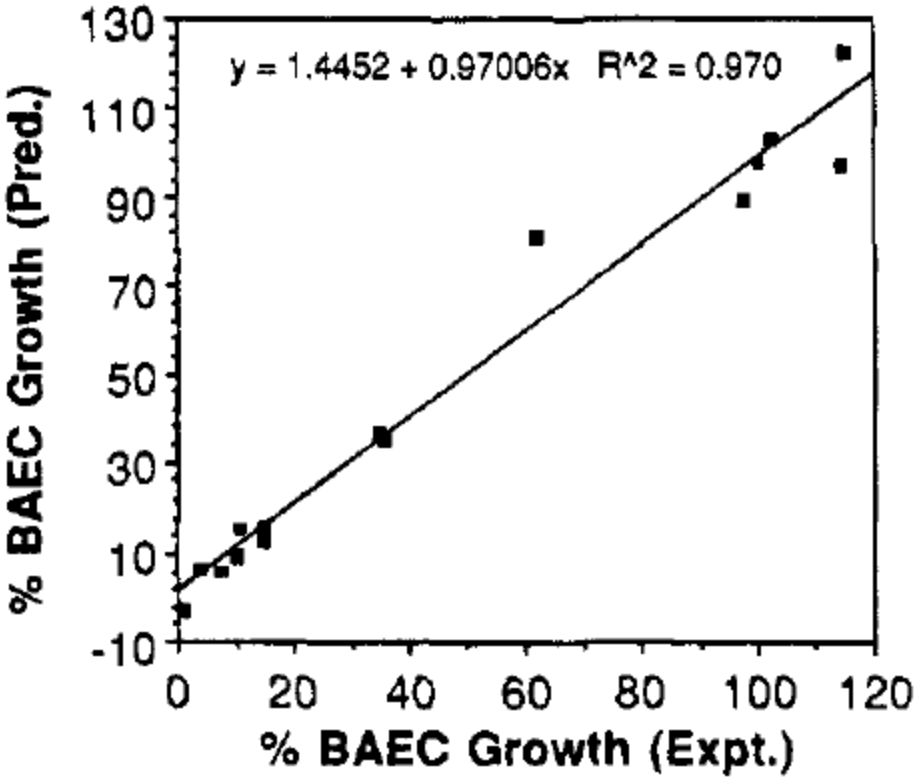
# *Modeling*

- Univariate:
  - y=mx+b
- Multivariate:
  - Y=XB+E
  - Partial least square (PLS)

Analytica Chimica Acta, 1986, 185, 1-17

# *Predicting Cell Growth From ToF-SIMS Spectra*

- Y=XB+E
  - Y=%BAEC growth
  - X=SIMS peak intensities

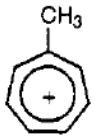



Analytica Chimica Acta, 1986, 185, 1-17

# *What Correlates With Cell Growth?*

- PLS allows you to determine what measured variables correlate with the measured response
- In this case, what peaks correlate with cell growth

**Table 4. Tentative Ion Structure Assignments for the SIMS Variables with the Largest Regression Coefficients[a]**

| *m/z* | ion structure | *m/z* | ion structure |
|---|---|---|---|
| | Positively Correlated with Cell Growth | | |
| 15+ (3.3) | $CH_3^+$ | 1− (8.7) | $H^-$ |
| 18+ (2.8) | $NH_4^+$ | 41− (3.4) | $CHCO^-$ |
| 27+ (7.8) | $C_2H_3^+$ | 42− (7.3) | $NCO^-$ |
| 29+ (5.6) | $C_2H_5^+$, $CHO^+$ | 43− (6.3) | $CH_3CO^-$ |
| 30+ (4.9) | $CH_2NH_2^+$, $NO^+$ | 55− (3.7) | $CH_3CHCO^-$ |
| 39+ (10.8) | $C_3H_3^+$ | 59− (4.2) | $CH_3COO^-$ |
| 41+ (13.4) | $C_3H_5^+$, $CHCO^+$ | 73− (7.2) | $HOCH_2CH{=}CHO^-$, $CH_3CH_2COO^-$ |
| 42+ (4.5) | $C_2H_2NH_2^+$, $C_2H_2O^+$ | | |
| 43+ (8.9) | $CH_3CO^+$, $C_3H_7^+$ | 80− (6.3) | $SO_3^-$ |
| 44+ (4.7) | $C_2H_4NH_2^+$ | 87− (7.0) | $C_2H_5COOCH_2^-$ |
| 45+ (3.0) | $C_2H_5O^+$ | 97− (13.6) | $HSO_4^-$ |
| 53+ (3.7) | $C_4H_5^+$, $C_3HO^+$ | | |
| 55+ (5.9) | $C_4H_7^+$, $C_3H_3O^+$ | | |
| 56+ (3.1) | $C_2H_4N_2^+$ | | |
| | Negatively Correlated with Cell Growth | | |
| 23+ (−4.2) | $Na^+$ | 13− (−9.0) | $CH^-$ |
| | | 24− (−5.9) | $C_2^-$ |
| 91+ (−3.6) |  | 25− (−19.0) | $C_2H^-$ |
| | | 57− (−5.8) | $C_3H_5O^-$ |
| 105+ (−4.5) |  | 71− (−6.6) | $C_3H_3O_2^-$ |

Analytica Chimica Acta, 1986, 185, 1-17

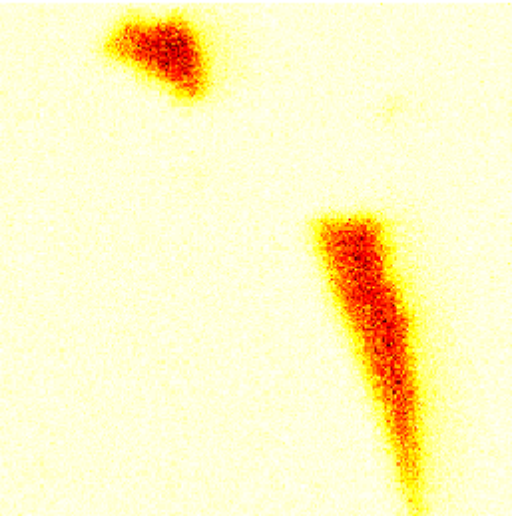
# PCA of ToF-SIMS Images: Solving Quality Control Issue

- PCA:Gold Delamination Problem
  - Noticed patches with missing gold on samples used for self assembly
  - Imaged patches with ToF-SIMS
  - Selected all peaks above background
  - Ran PCA on mean centered image data

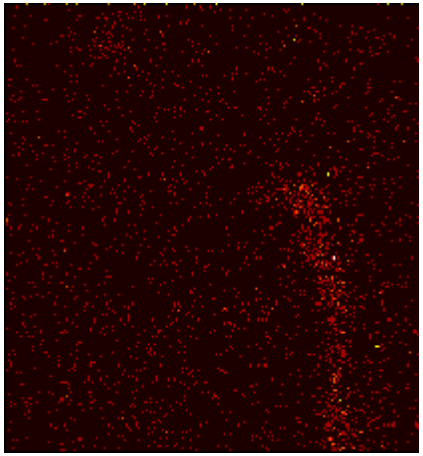
# Dodecanethiol on Au/Ti/Si

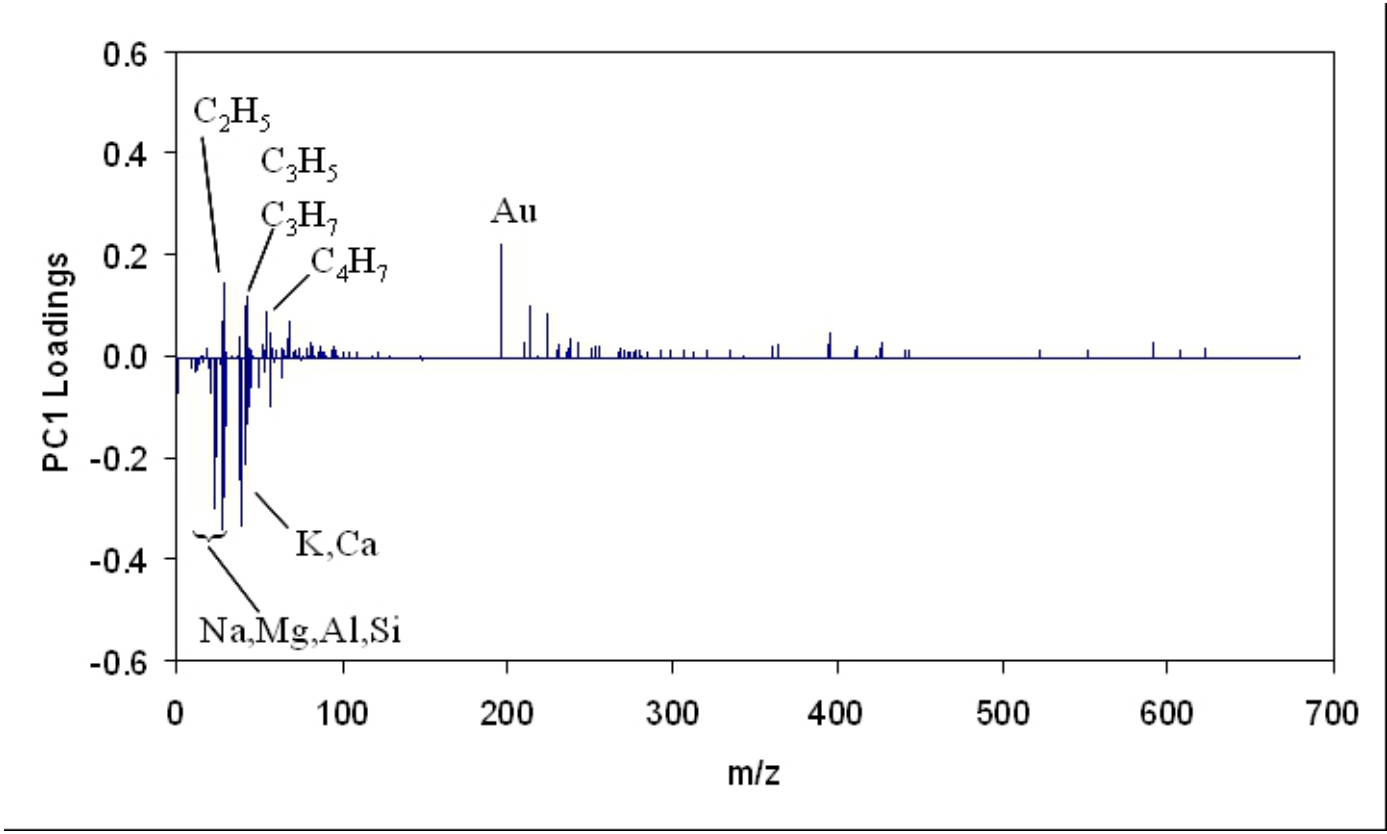
Image taken over delamination region

PC1 scores



Ti ion map

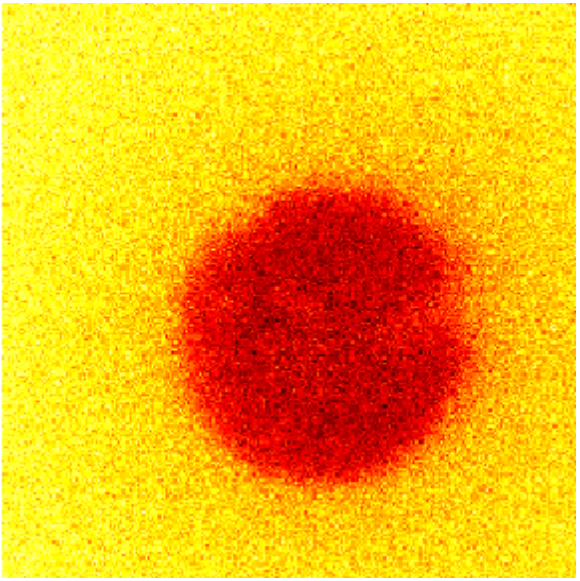


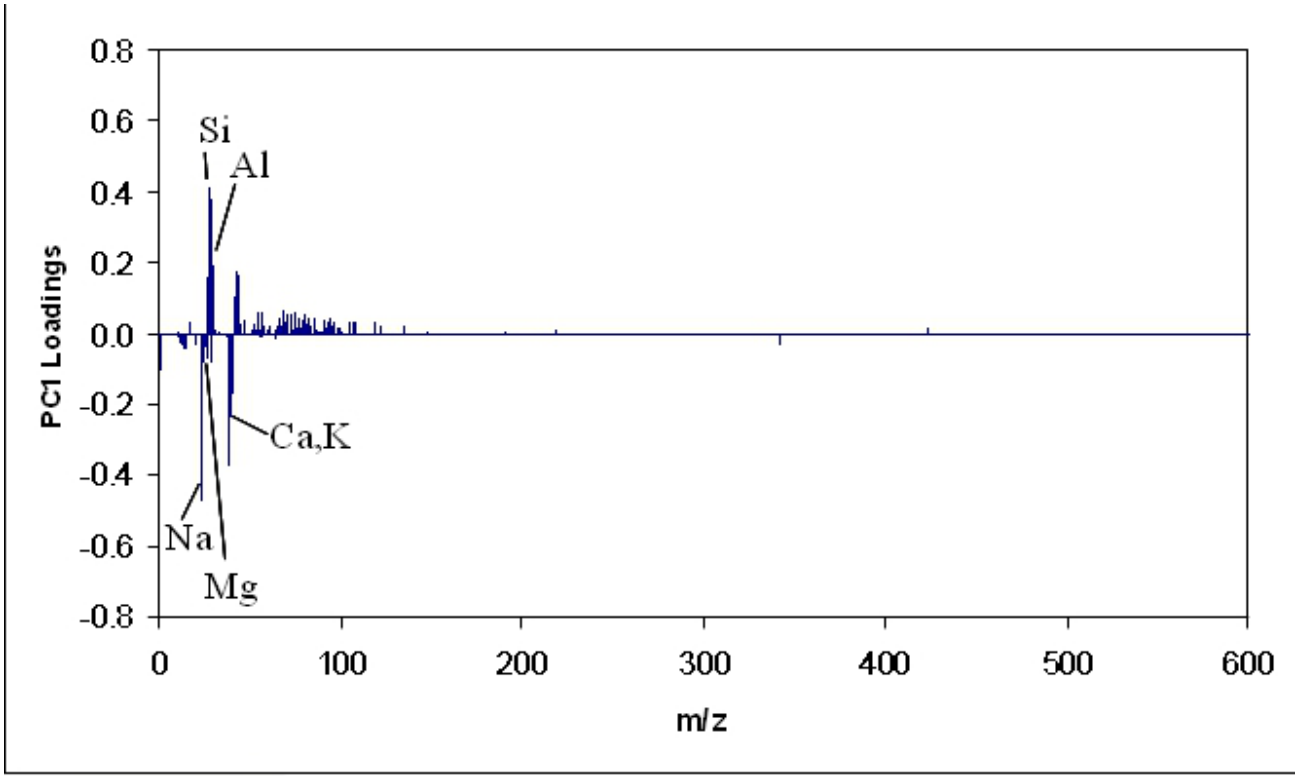


# PCA: Gold delamination problem

- Looked through gold sample preparation protocol
- Looked at samples after cleaning
- Noticed residue on the samples
- Imaged residue with ToF-SIMS
- Ran PCA on mean centered image data

# Bare Silicon wafer after cleaning
# Image of 'blob' on surface





# PCA: Gold delimitation problem

- PCA was able to quickly identify differences in the gold delamination areas
- Allowed generation of a hypothesis
- Isolation of the problem and adjustment to the sample preparation protocol

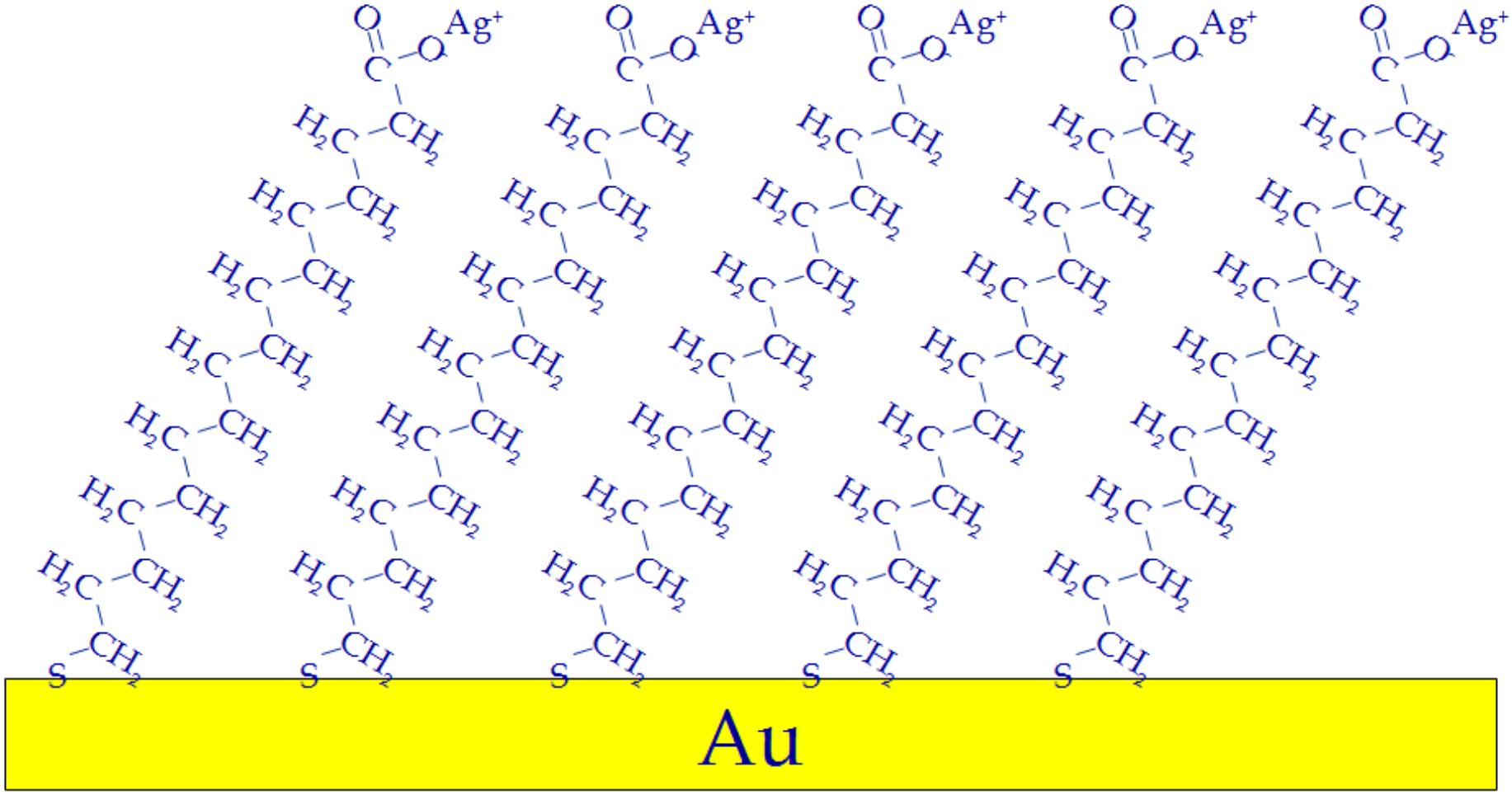
# PCA: Understanding Cationization

- It is known that silver ions (Ag+) can aid in the emission of large polymer fragments via cationization
  - Attachment of a cation to the molecule
- We wanted to:
  - Explore the cationization process
  - Determine if other cations also worked
  - Optimize new substrates for cationization

**Michel et al (2000) Langmuir, 16:6503-6509**

# Protein and polymer cationization

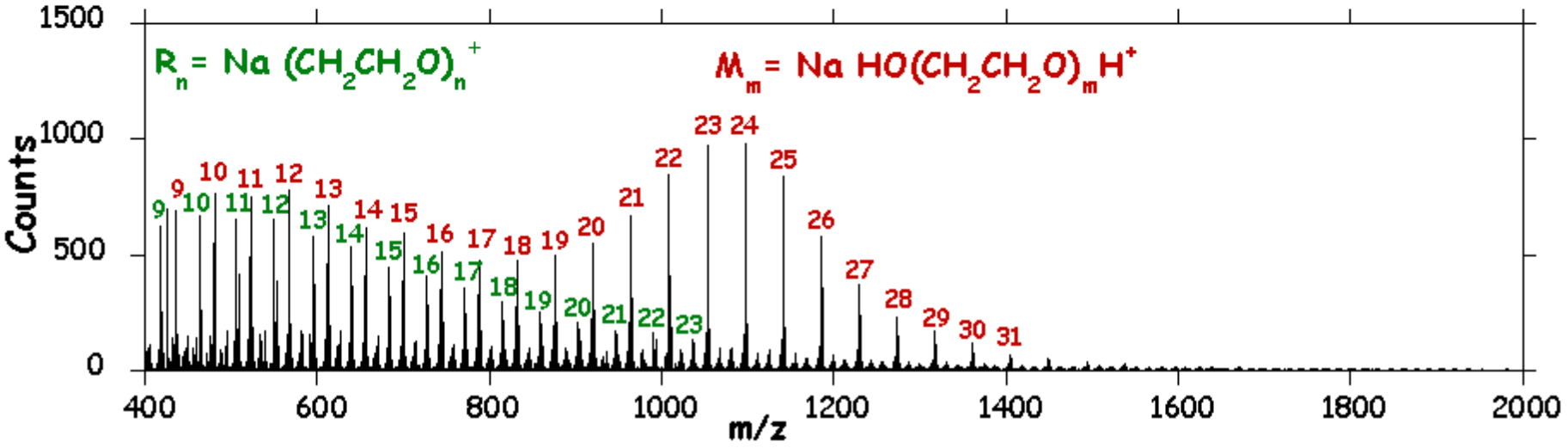
## Cations stabilize large fragments during SIMS analysis




Michel et al (2000) *Langmuir*, 16: 6503-6509

# Optimization of Cationization

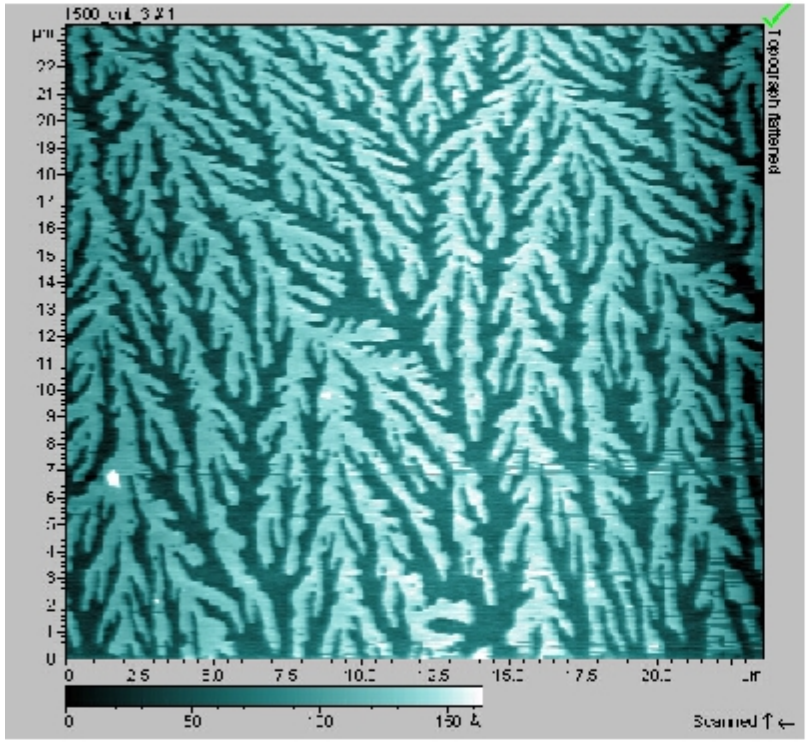
- How can we maximize the yield of high mass fragments with cationizing SAMs?
- What factors affect the high mass yield?

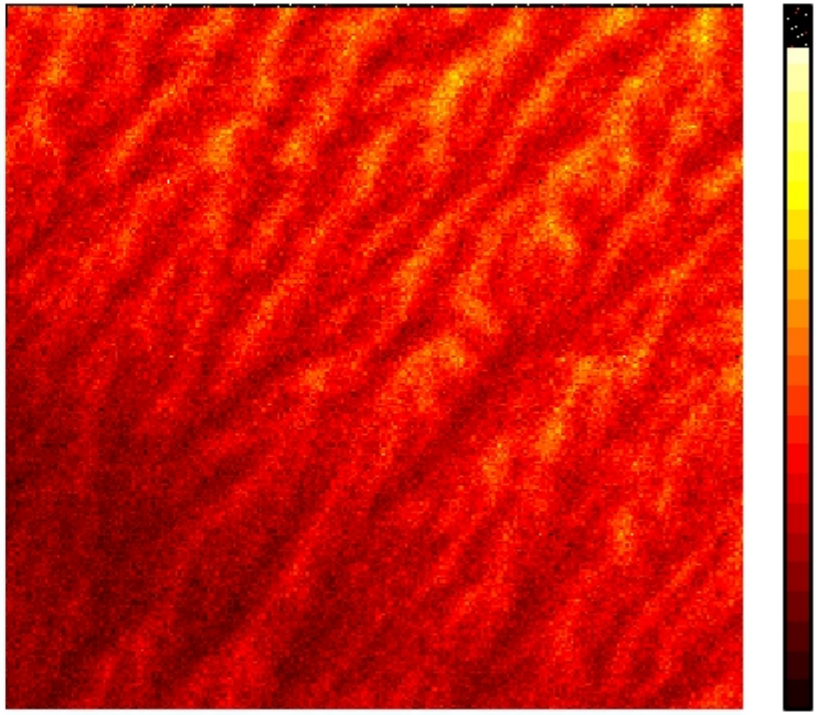



Michel et al (2000) Langmuir, 16:6503-6509

# PEG on COONa SAM

PEG.RAW: 0
Acq: 3 Apr 25

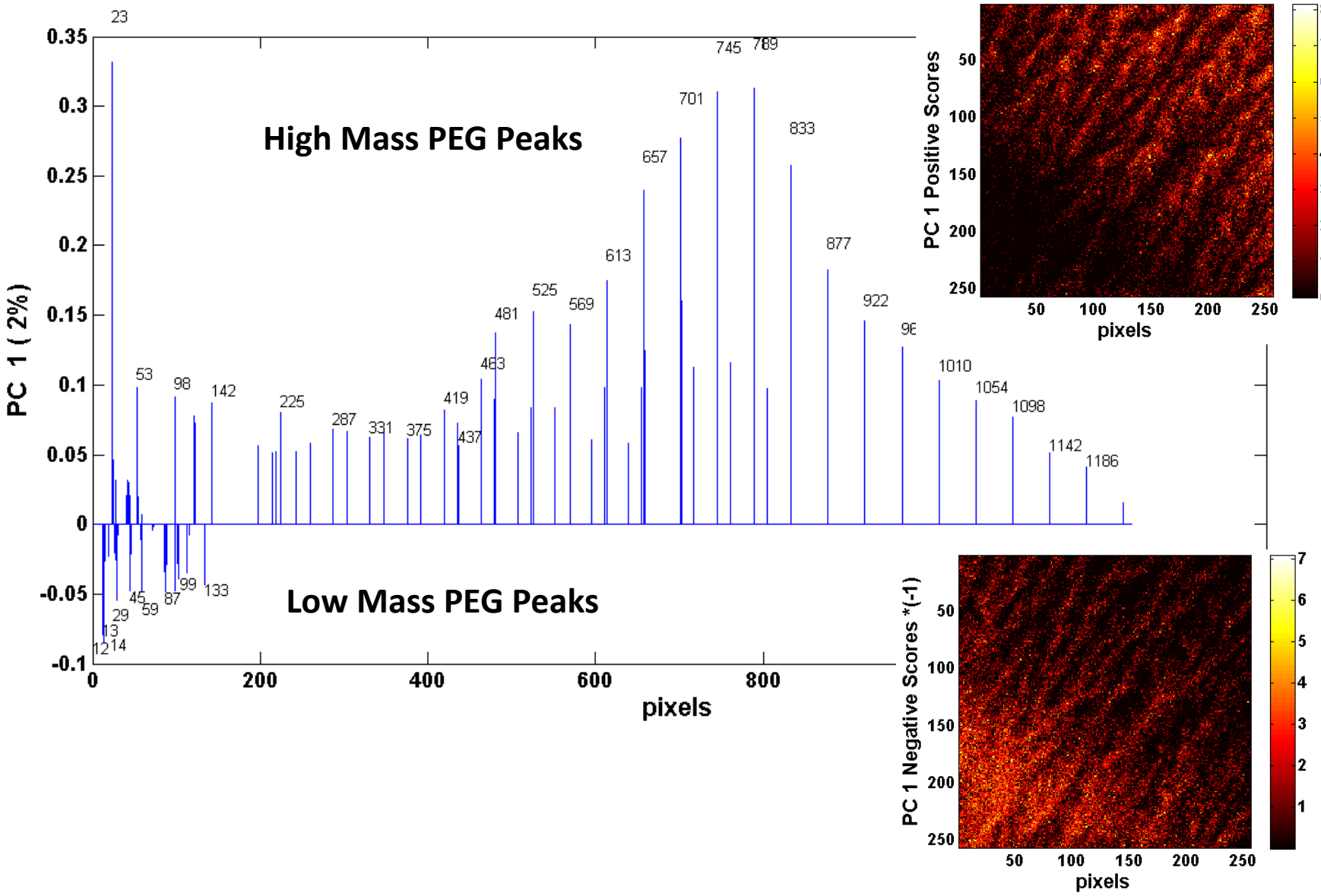


AFM Image

TOF-SIMS Image

# PCA of PEG on Cationizing SAM

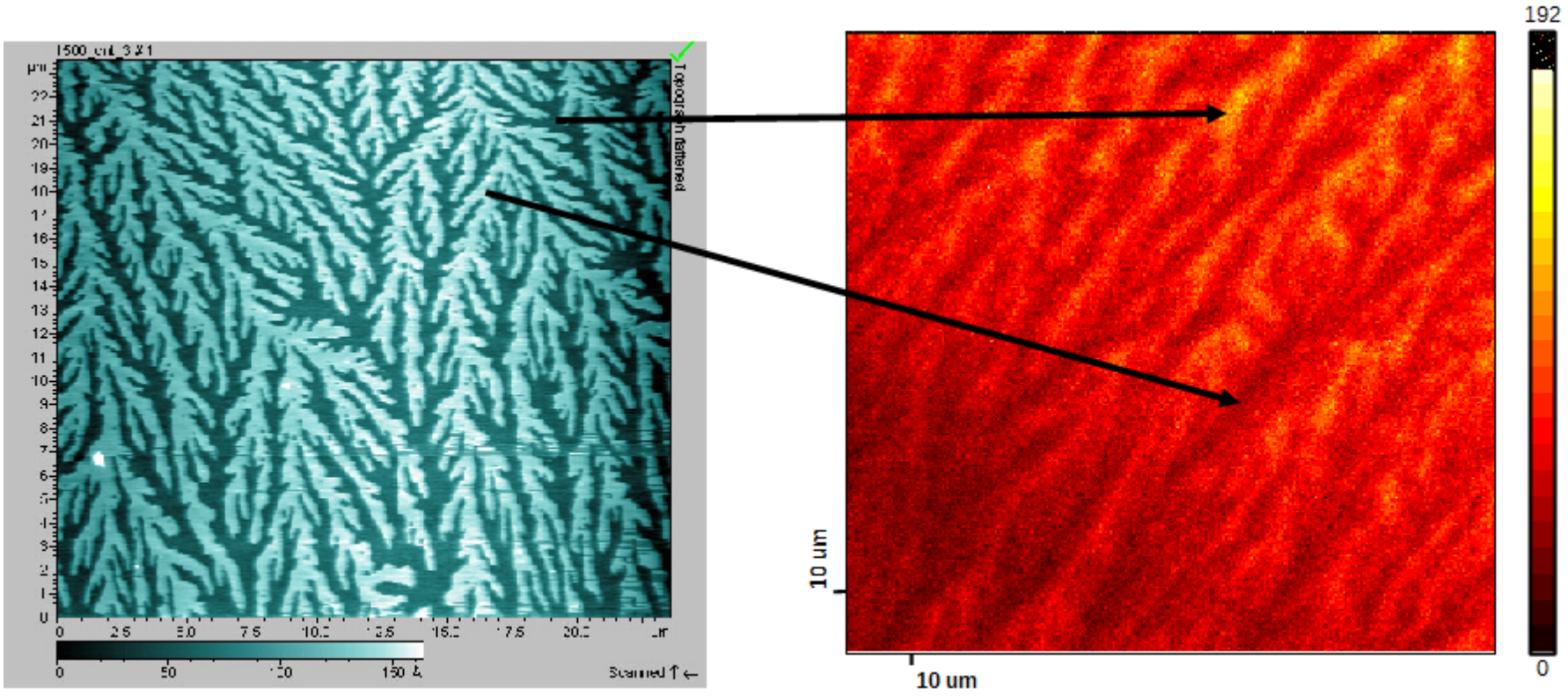


# What does it mean?

-It is believed that the high regions in the AFM image correspond with the dark regions in the TOF-SIMS image

-Thinner areas (brighter regions on SIMS image) produce higher yield of high mass fragments

-SIMS beam can actually sample through polymer to $Na^+$ ions

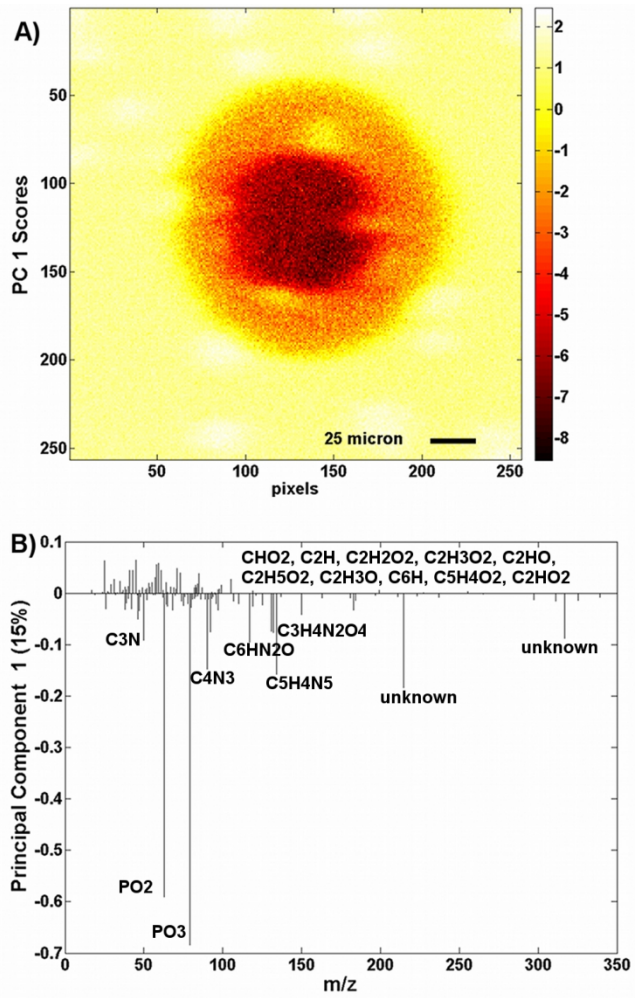


# MVA of DNA Microarray Spot

- Using ToF-SIMS to analyze DNA microarray spots
  - Check uniformity of spots
  - Look for chemical variation
  - Use results to feedback into spotting methodology to get better signal to noise
- Selected all peaks above background (negative ion data)
- Poisson scaled and mean centered
- Ran PCA, MAF and MCR

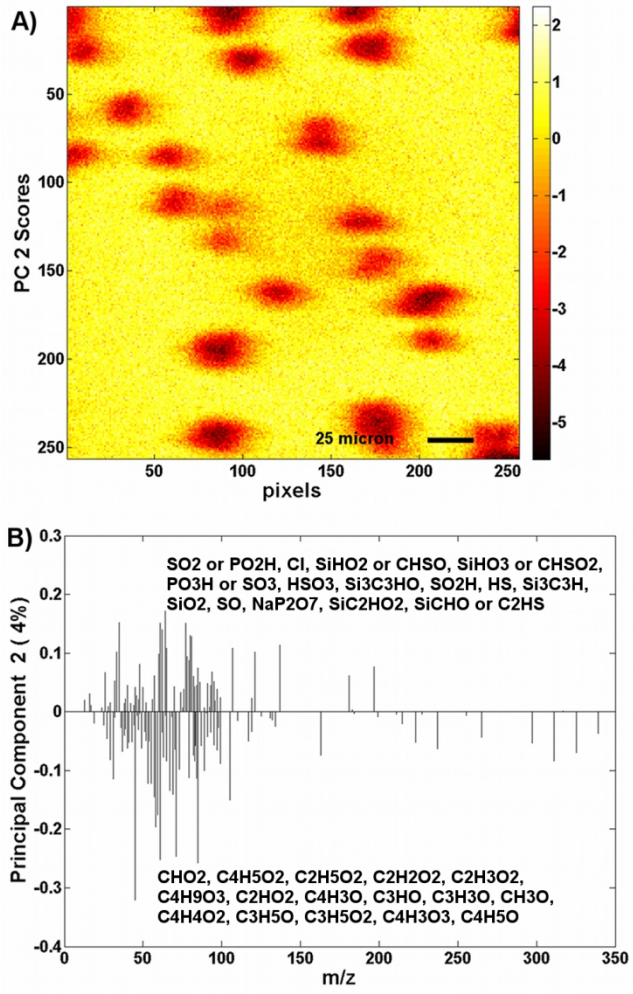
Data from Lara Gamble, Nicolas Vandecasteele

# PCA Results




- PC1
  - Separates DNA spot from background
  - DNA spot is not uniform
  - White dots appear across image

# PCA Results

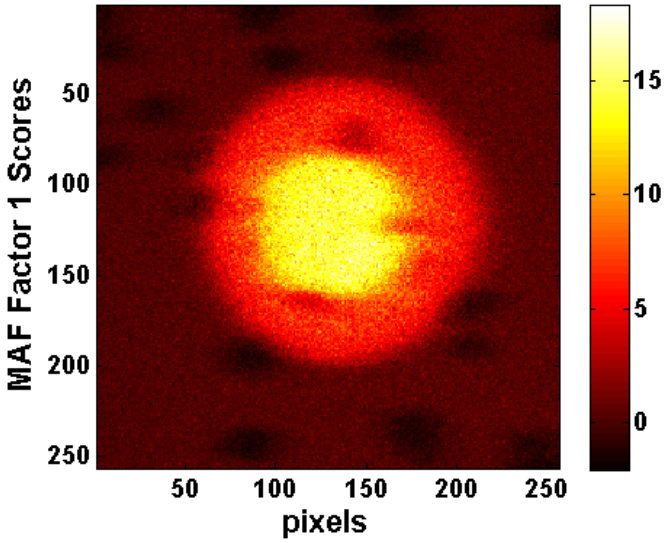


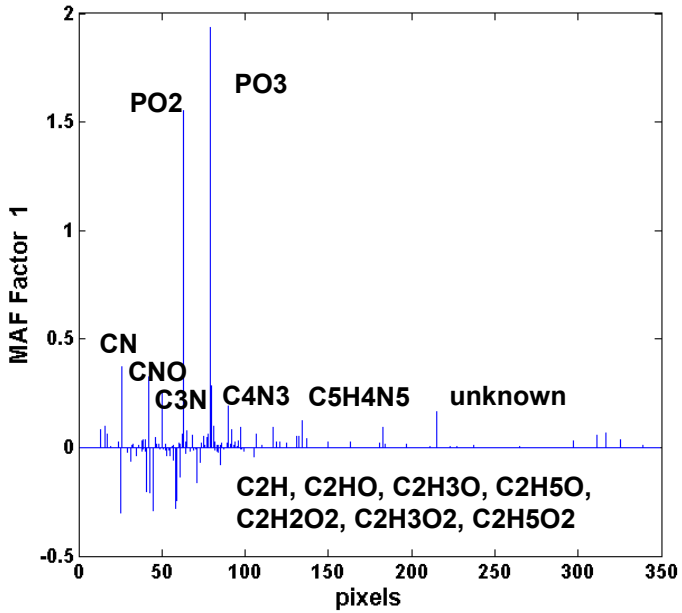

- PC2
  - Highlights dots
  - Peaks suggest presence of PEG
    - Known to crystallize and form agglomerate on some surfaces

# MAF of DNA Microarrays

- For PCA data matrix X is decomposes as:
  - S = UTX
    - S=scores
    - U = Loadings = eigenvector rotation of covariance matrix of X
- For MAF
  - U = eigenvector rotation of B where:
    - B=A-1V
      - V= covariance matrix of X
      - A=covariance matrix of shift matrix
        - Shift matrix = X-X shifted by 1 pixel in X or Y
    - Finds linear combination of peaks that maximize variation across the image while minimizing variation between neighboring pixels

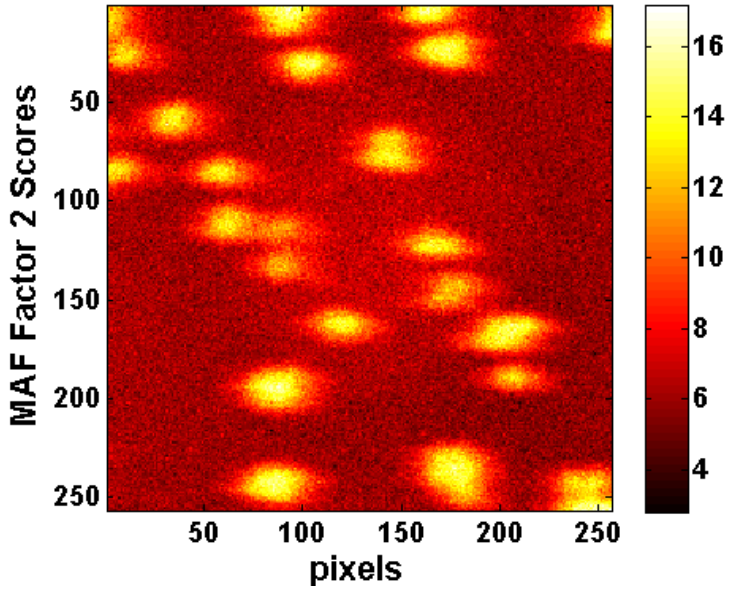
# MAF Results



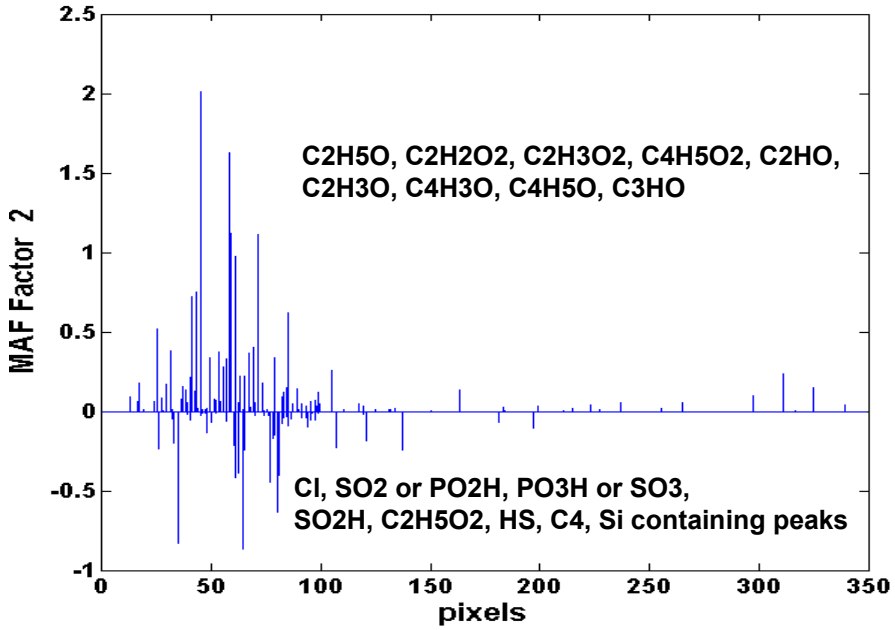




- Factor 1
  - Separates DNA spot from background
  - Contrast is inverted from PCA, but peaks are the same

# MAF Results



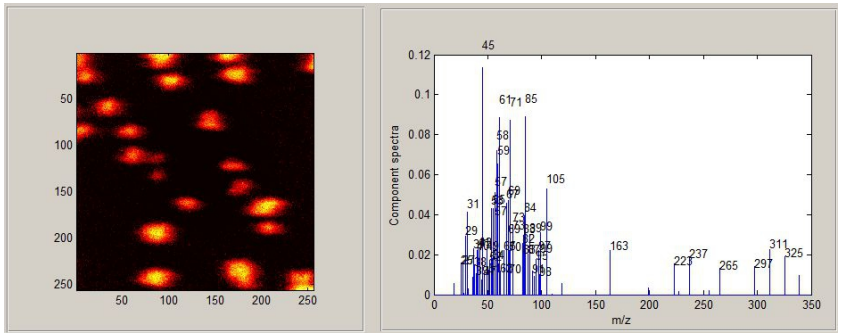

- Factor 2
  - Highlights dots
  - Peaks suggest presence of PEG
    - Consistent with PCA results

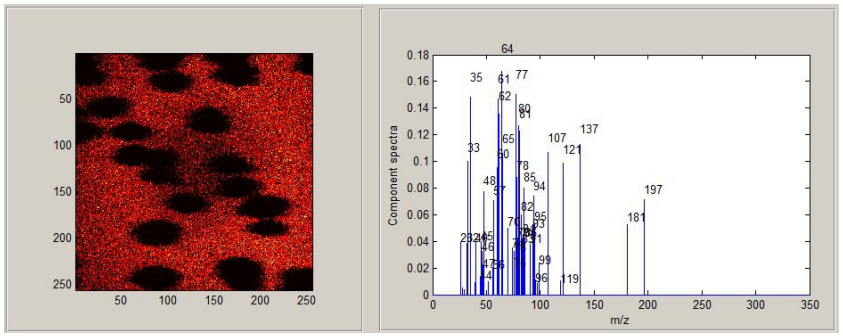


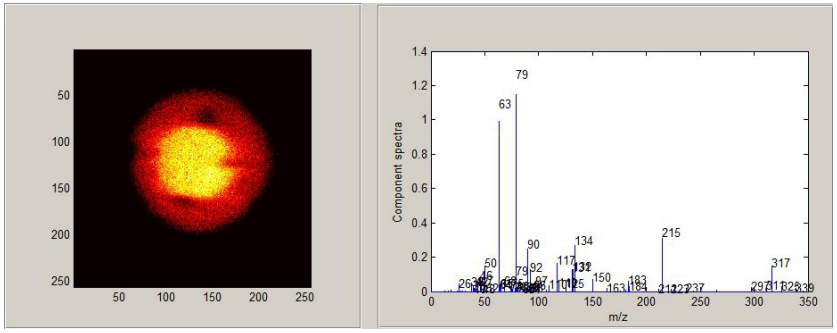
# MCR of DNA Microarrays

- MCR tries to describe the data as a linear model:
  - X=CFT+E
  - F=Spectra of pure components
  - C=concentration of components at each pixel
  - E=random error
- C and F are found using a least squares minimization
- Requires an initial guess of how many components there are
- For this data we used PCA results as initial guess for MCR of DNA microarray spot
  - MCR was run on Poisson scaled/mean centered data

# MCR Results



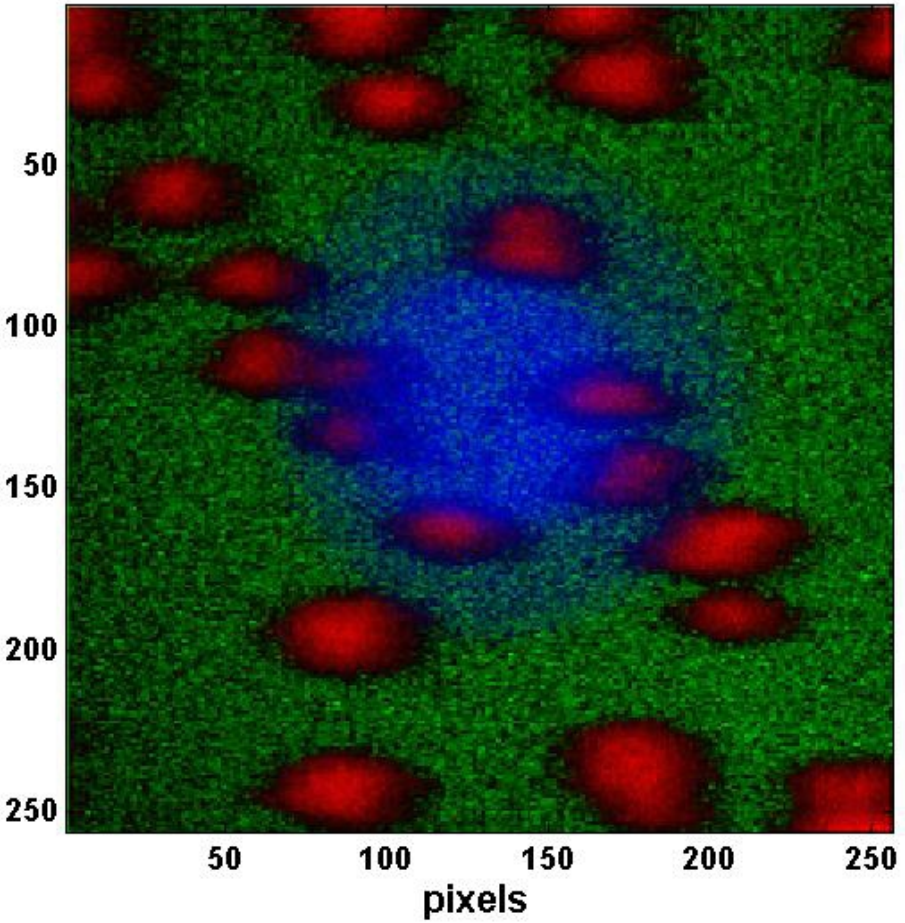




- Component 1 = PEG
- Component 2 = Silane linker/buffer salts
- Component 3 = DNA

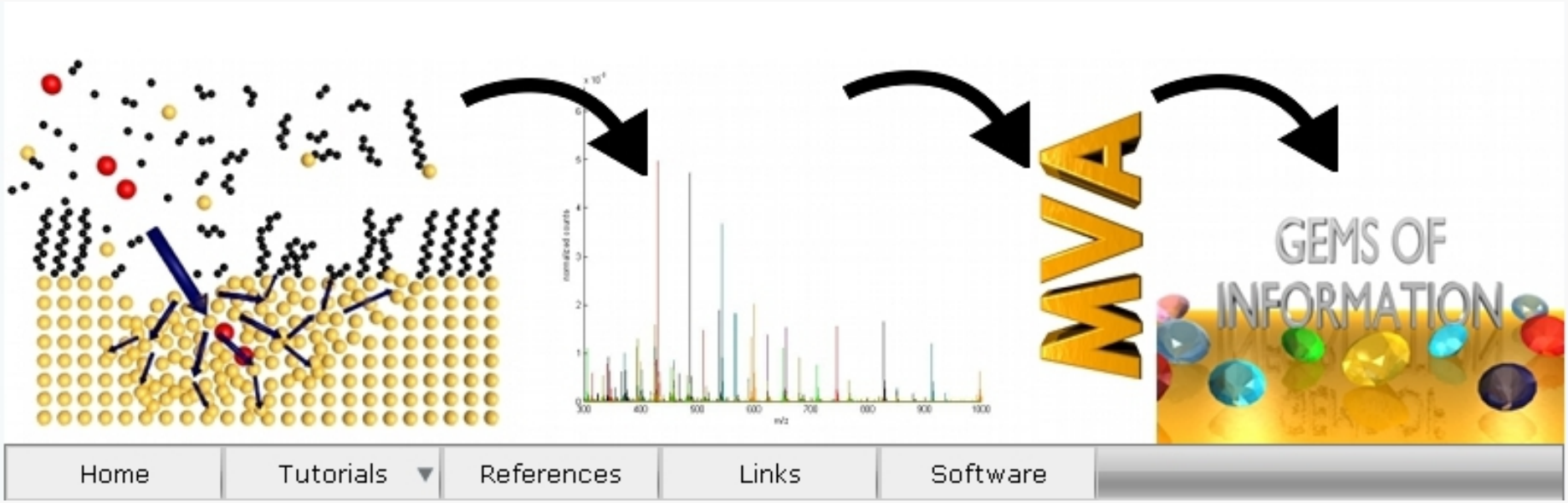
# RGB Overlay of MCR Components

- Red = PEG
- Green = Silane/Salts
- Blue = DNA



# Summary

- MVA methods are powerful tools to help with ToF-SIMS data analysis
- Successful ToF-SIMS experiments using MVA require good experimental plans
- The method you use should be chosen based on the needs of your analysis
- Data preprocessing should be done based off logical assumptions about the data




- **mvsa.nb.uw.edu**
- Tutorials
- References
- Links
- Software

djgraham@uw.edu

National Institute of Biomedical Imaging and Bioengineering

NESAC/BIO

# The Complexities of The Basics of PCA (Part II)

Dan Graham, Ph.D.
NESAC/BIO
University of Washington

*An NIH National Center for Research Resources*
*(NIH grant EB-002027)*
**http://www.nb.uw.edu**

NIH National Institute of Biomedical Imaging and Bioengineering

NESAC/BIO

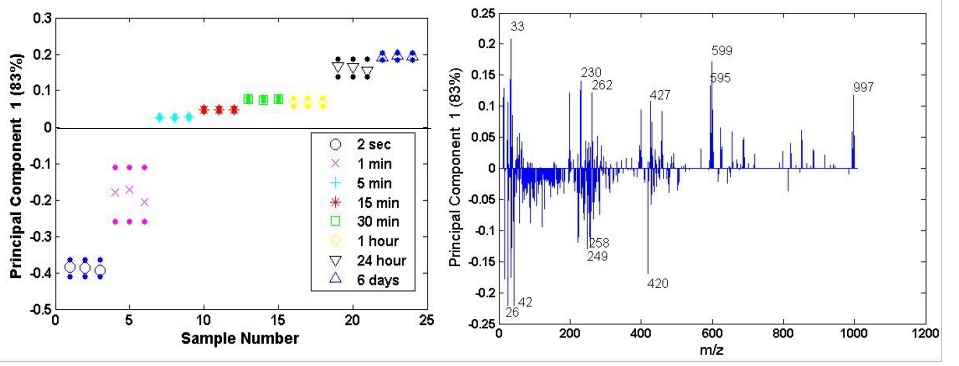
This is Part II of a 2 part tutorial about PCA. In this tutorial I will cover some ideas on how to verify trends in PCA results, some specific issues for dealing with image data sets, and then provide examples of how PCA can be applied to ToF-SIMS data.

# Overview

- **Some basic complexities to think about for MVA of spectra**
- **Some basic complexities to think about for MVA of images**
- **Some examples of how MVA can be helpful**

# Raw Facts About Raw Data

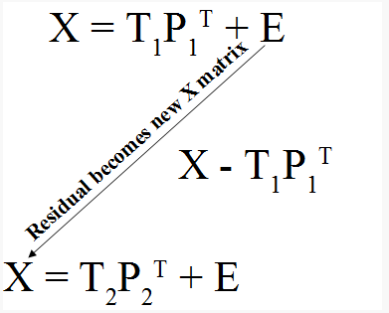
- **MVA methods are simply tools to help digest and analyze data**
- **It is always recommended to go back to the original data to verify any trends**
- **Most MVA methods are carried out on preprocessed data, not the original raw data**




It is often recommended that one goes back to the original data to verify the trends seen in PCA. However, we must first think about what data we should look at. PCA (and other MVA methods) are typically calculated from a preprocessed data matrix, not the original raw data.

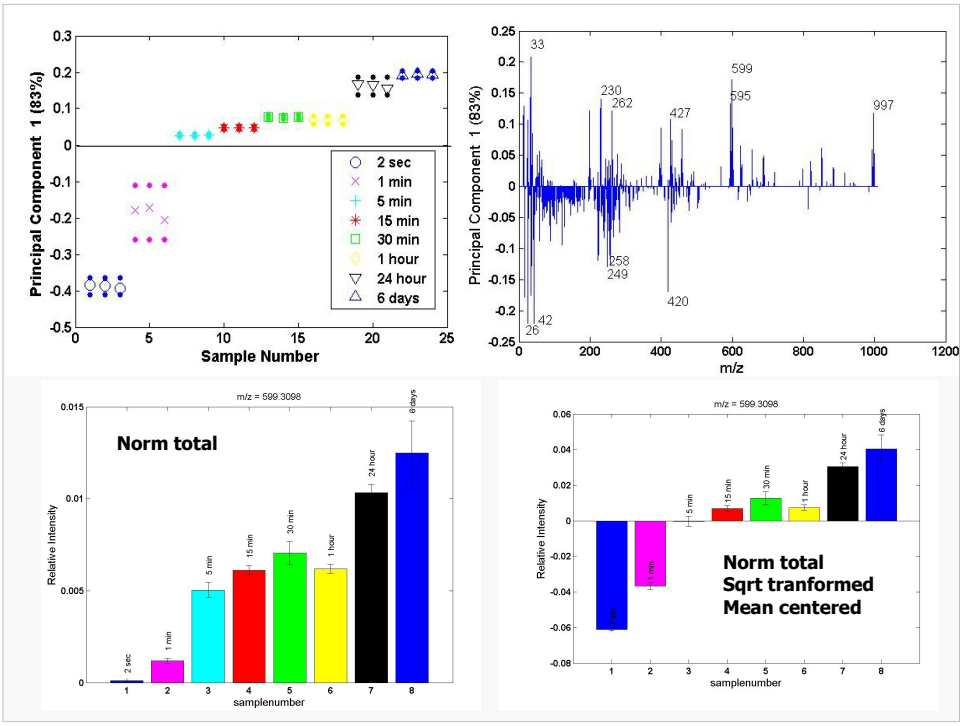
## Which "Raw" Data Should We Look At?

- **The original normalized counts?**
- **The preprocessed counts? (PCA is typically carried out on data that has been centered or scaled)**
- **What should be done when looking at PCs after PC1?**

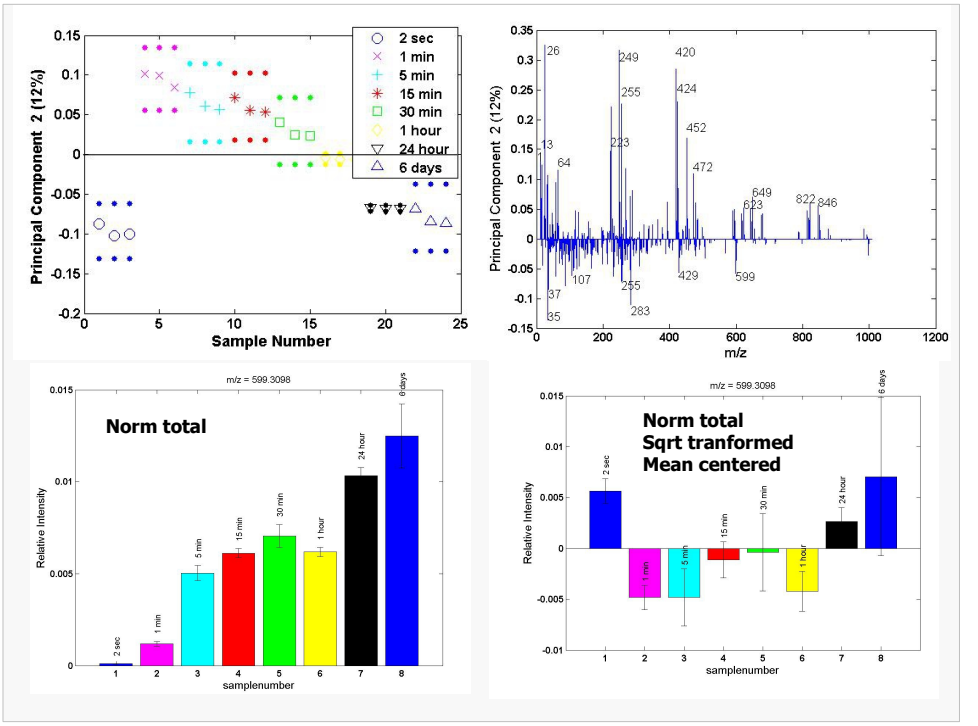



Each subsequent PC is calculated after subtracting the previous PCs from the data matrix. So what is the "raw" data for these PCs?

We also have to remember that most factor based methods are calculated sequentially, meaning that the first factor is found and then subtracted from the data set, and then the next factor is found and subtracted from the data set, and so forth. So the original normalized counts will only be reflective of PC1. However the most representative data of the trends seen in PC1 will be the preprocessed data matrix.



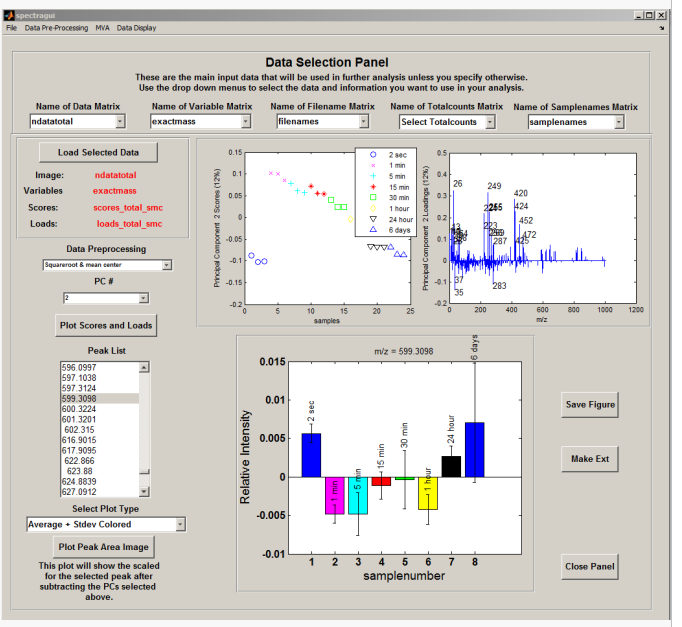

This slide shows an example from PCA of a series of samples created by exposing a gold surface to a dilute solution of dodecanethiol for various times ranging from a few seconds to several days. As can be seen in the top left figure, the PC1 scores increase with increasing assembly time. If we look at the peak at m/z 599 we see it has a high positive loading. If we plot the normalized counts for this peak for each sample, we see that it shows an increasing relative intensity with increasing assembly time as we would expect based on the PC1 scores. However, the scores show both positive and negative values. If we look instead at the preprocessed data for this peak, we see that the preprocessed relative intensity follows the trend seen in PC1 almost exactly. This makes sense since this is the data that PC1 was calculated from.




If we now look at PC2 we see a trend in the PC2 scores where the score value goes up after 2 seconds and then decreases over time. Once again if we look at the m/z 599 peak, the plot of the original normalized counts does not follow this trend at all. This is not unexpected because PC2 was calculated from the preprocessed data matrix after PC1 was removed from the data set. If instead we look at the preprocessed data after removing PC1 from the data set, we see that the preprocessed relative intensities for this peak for PC2 follows a trend similar to that seen in the PC2 scores. The trend is inverted from that seen in the scores plot because the m/z 599 peak has a negative loading on PC2.

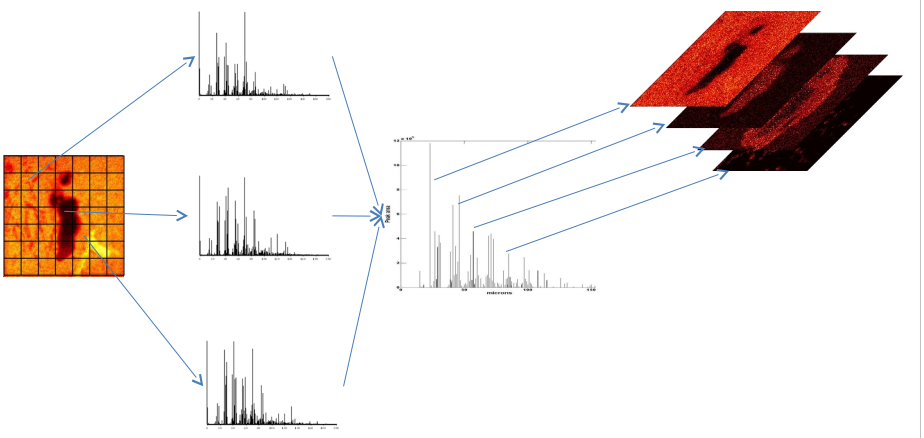
This would suggest that one way to look at the “original” data for a give PC is to look at the preprocessed data after subtracting any previous PCs from the data matrix. This process is mathematically sound. The challenge remains in how to interpret the differences one sees in a given PC.

- **Recommended to look at preprocessed data for a given peak from a given PC**
- **NBToolbox has a function for this**



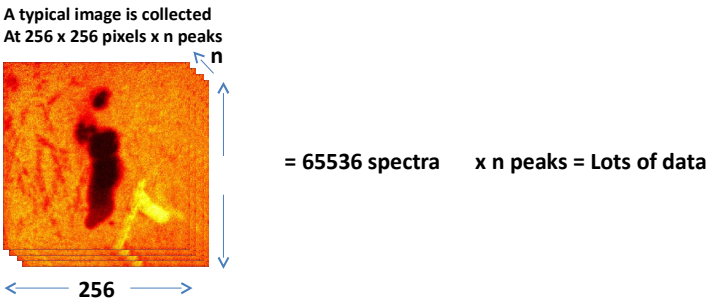

The NBtoolbox has a function called the PC Databrowser that enables looking at the preprocessed data after subtracting all previous PCs.

# ToF-SIMS Imaging Data



As you may know, modern ToF-SIMS instruments collect images and spectra simultaneously. So a given data set can be visualized as an image, or the spectra from a given pixel or set of pixels can be displayed. You can also sum the spectra from all pixels and get a total ion spectrum from the entire data set. One can also select a given peak and display the peak area image for any peak.

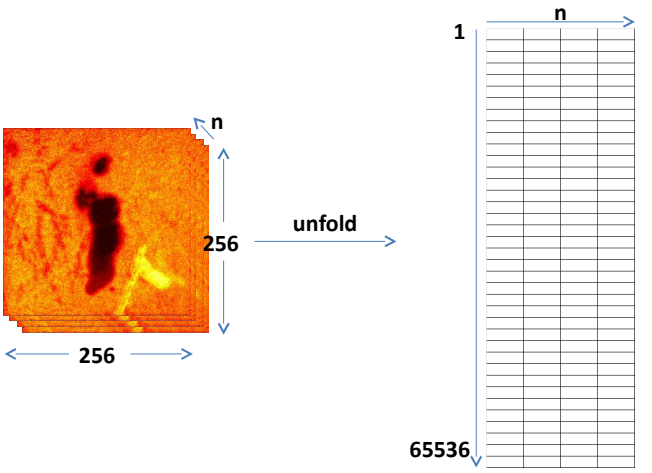
# ToF-SIMS Imaging Data



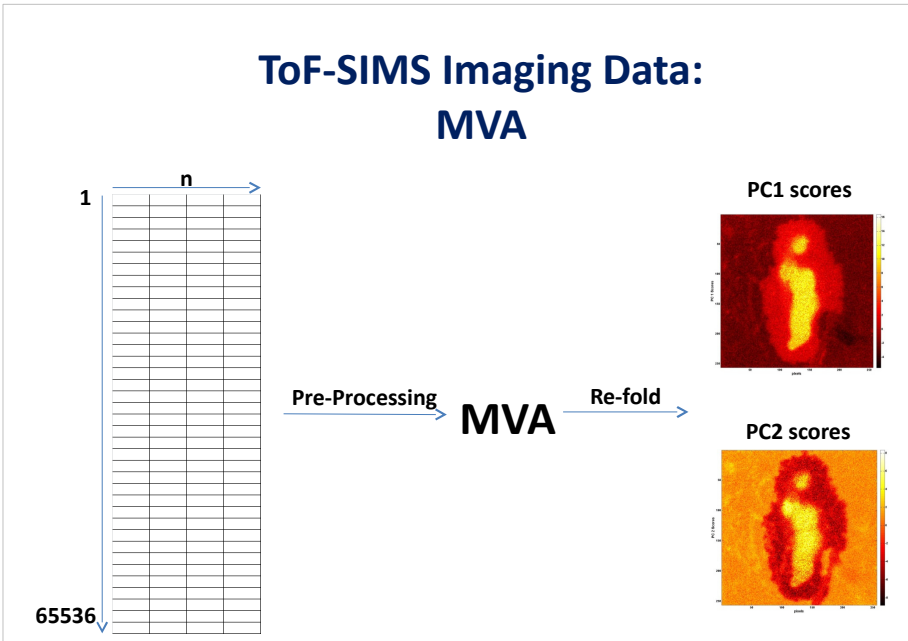

Imaging data sets can be quite large. Typical images are collected at 256x256 pixels. The full image stack would then be 65,536 x the number of peak area images (or data channels). This can easily result in hundreds of thousands of spectra. This is significant, because MVA methods treat all data sets as a series of spectra. More spectra means more computing time and more required computing power.

I recommend you install as much RAM as possible on your system and choose a computer with the best processor available within your budget.

# ToF-SIMS Imaging Data: MVA



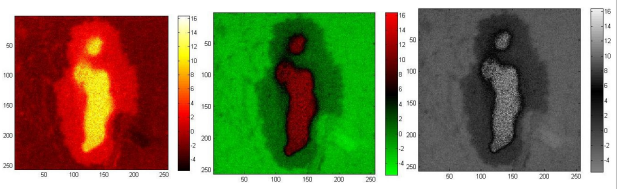

Before running MVA on a ToF-SIMS imaging data set, the data is unfolded into a 2d matrix where the rows contain spectra and the columns contain the peak areas for each spectra.

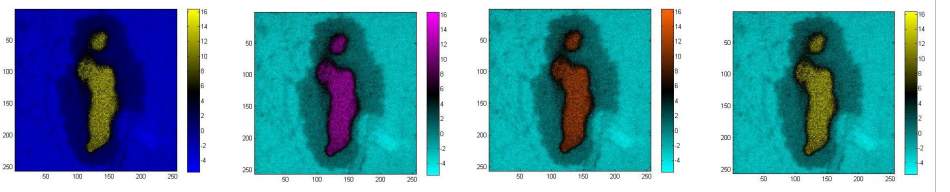



This unfolded data matrix is then preprocessed and used for MVA. The scores matrices can then be refolded into score images that can be displayed as desired.

# ToF-SIMS Imaging Data: Displaying Scores

- **Use strong contrasting colors**
- **Best if 0 = black**
- **Greyscale can avoid color interpretation issues**
- **Provide color scale bar**





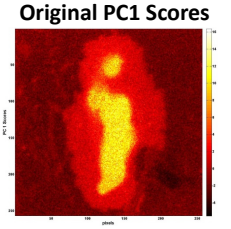
When displaying scores images it is useful to use strong contrasting colors. It is also helpful to use a color scheme where zeros are displayed as black (for some reason the figures above do not show this, but scores plots made in the imagegui do have black at zero). When choosing a color scheme, remember that some people are colorblind.

The imagegui contains several colormaps (some of which are displayed above).

# ToF-SIMS Imaging Data: Displaying Scores

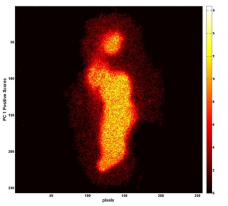
- **Display positive and negative scores separately**
- **Must clearly label what is being displayed**
- **Valid mathematically as long as you keep track of the sign in both the scores and loadings plots**

**Original PC1 Scores**

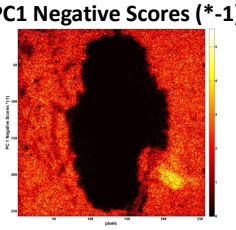


**Multiplied by -1 in order to display negative loadings from dark to light color**

**PC1 Positive Scores**

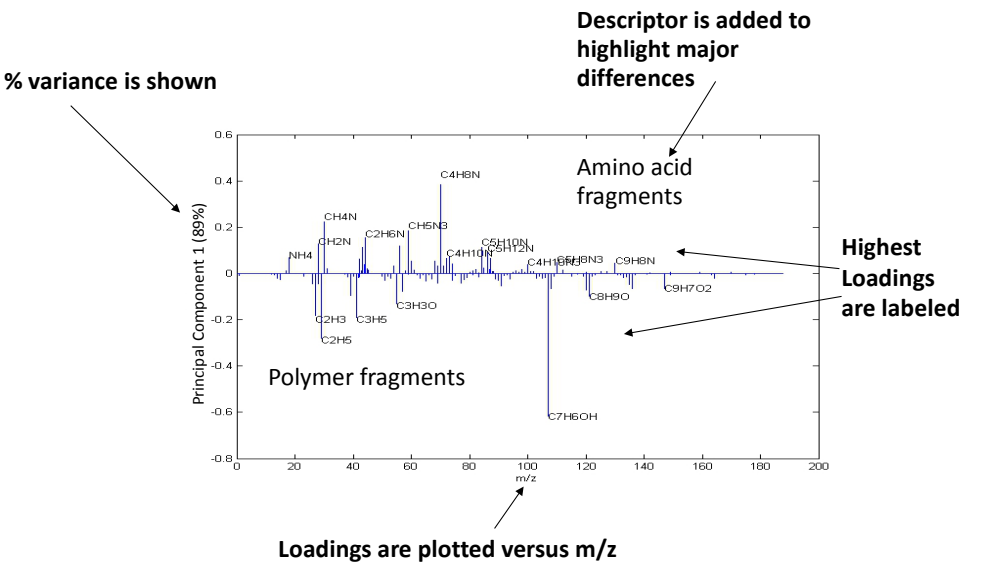


**PC1 Negative Scores (*-1)**



Another way of displaying score images that can be useful is to split the positive and negative ion scores into separate images. The NBToolbox allows plotting of positive and negative scores and loadings separately. In order to use the colorbars within matlab, the negative scores and loadings are multiplied by -1 in order to display them as positive values. This is valid as long as you remember to report what was done, and most importantly that you always remember to keep positive scores with positive loadings and negative scores with negative loadings and remember to multiply both the negative scores and loadings by -1.

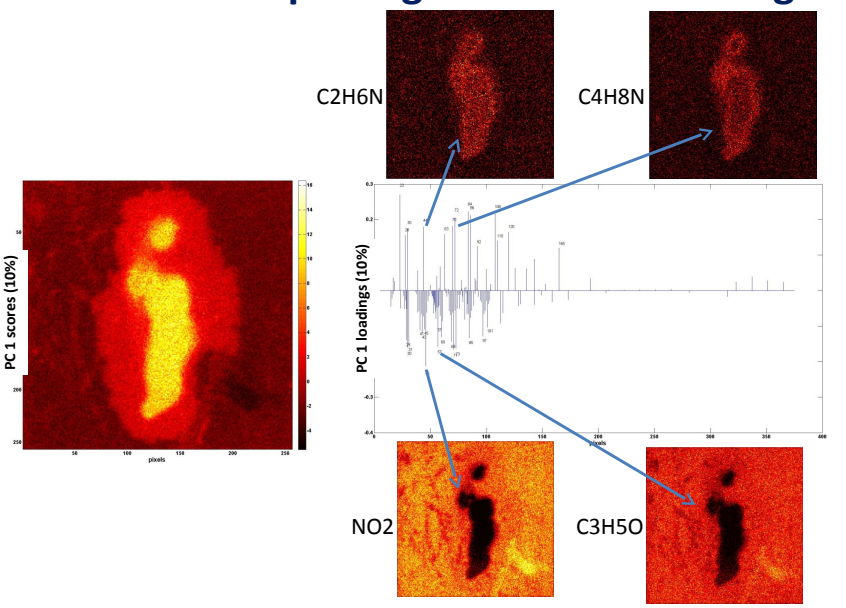
**ToF-SIMS Imaging Data:**
**PCA Loadings**




Loadings plots from image data sets can be displayed the same way as they are for spectral data sets. As usual it is important to label the axes to show what is being plotted. I find it easiest to interpret scores and loadings 1 PC at a time.

By plotting the loadings versus the peak masses, the loadings plot will look similar to a mass spectrum. This also can help space out the peaks for easier labeling. Only label the highest loading peaks to keep the plot from being too busy. You can summarize the trends in the loadings by adding text above and below the peaks. If there are other peaks of interest in the loadings, you can show them in a separate table.

# ToF-SIMS Imaging Data:
# MVA – Interpreting Scores and Loadings




As usual, positive scores correspond with positive loadings. For imaging data sets that means that peaks with positive loadings will show peak area images with higher relative intensities in areas with positive scores. Peaks with negative loadings with show peak area images with higher relative intensities in areas with negative scores.

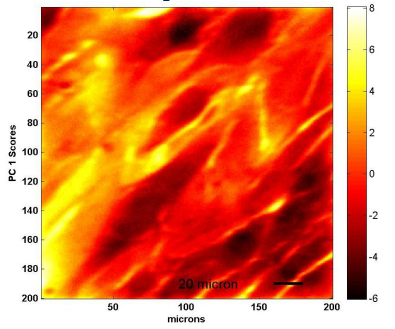
### Cell Imaging: Effects of data normalization

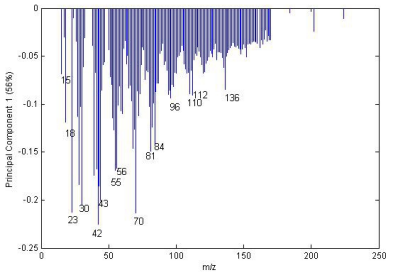
- Image data taken in high spatial resolution mode
- Image data summed from depth profile using 150 slices (to increase counts)
- All peaks above background selected (negative ion data)
- PCA of:
  - Poisson scaled/mean centered data
  - Normalized/Poisson scaled/mean centered data

Here I provide an example of an ToF-SIMS image data set to show the affects of normalization. This data set consists of a ToF-SIMS image from cells on silicon. The image was constructed by summing 150 slices of a depth profile to improve image contrast. All peaks above 3x the background were selected. PCA was run separately the data set preprocessed in 2 different ways. For one analysis the data was Poisson scaled and mean centered. For the other analysis, the data was Normalized to the total counts, then Poisson scaled and mean centered.

# Cell Data:Poisson scaled/mean centered

- All loadings are on one side
  - All peaks show a higher relative intensity for the areas with negative scores
  - Can be due to charging or topography
  - Not what we are interested in
  - Can be fixed by normalization

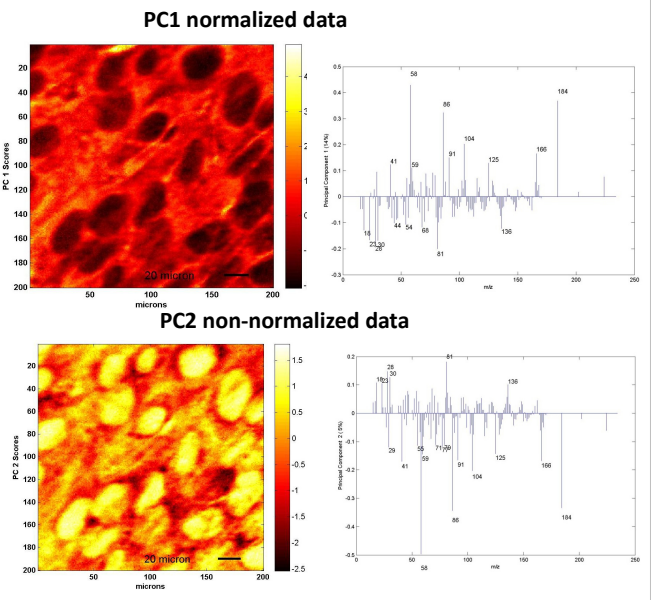



For the data set that was Poisson scaled and mean centered it is seen that the loadings for all peaks are negative. For this data set this means that all peaks have a higher relative intensity in pixels with negative scores. This is usually caused by issues with charging or topography in the image. Normalizing the data will remove the differences due to the overall intensity and therefore will eliminate this PC.

## Cell Data:Normalized/Poisson scaled/mean centered

- After normalization to the total counts we see contrast due to cells
- PC1 now shows same contrast/information as PC2 from non-normalized data
  - Because we have normalized out the variance due to the total counts




The top two figures are the PC1 PCA scores and loadings from the same data set after normalization, Poisson scaling and mean centering. As seen in these figures, we can now see the cell nuclei (dark areas) and cell membranes (brighter areas).

The bottom two figures are the PC2 PCA scores and loadings from the data set without normalization (same data as on the previous slide). Though the contrast in the scores plot is inverted, one can see that PC2 from the non-normalized data is the same as PC1 from the normalized data set. This is what we should expect because by normalizing the data, we removed the variance that was originally being captured by PC1. The next greatest variation in the data is what is being captured in PC2. So when you normalize the data what used to be in PC2 becomes PC1. This can be confusing, but if you think through it a couple of times it will make sense.

# A Note on Image Normalization

- Image normalization is sometimes helpful
- Should be used with care
  - Due to low count rates in SIMS images, normalization can
    - Accentuate noise
    - Possibly cause divide by zero errors if pixels have 0 counts

Though image normalization can sometimes be useful, it should be used with care. Due to the low count rates of ToF-SIMS images normalization can sometimes introduce artifacts and accentuate noise in the data. There is also the possibility of divide by zero errors if pixels have 0 counts. In the Imagegui divide by zero errors are avoided by simply keeping pixels with 0 counts as 0.

# Examples

- Spectra
  - ToF-SIMS analysis of proteins with various ion sources

Muramoto, et. al. J. Phys. Chem. C 2011, 115, 24247–24255

    - Determining protein orientation

Wang, Castner, B. Ratner, Jiang, Langmuir 20 (2004) 1877

    - Probing protein structure

Xia, May, McArthur and Castner, Langmuir, 18 (2002) 4090

    - PLS modeling to predict cell growth promotion

Analytica Chimica Acta, 1986, 185, 1-17

- Imaging
  - Investigating cationization
  - PCA, MAF, MCR analysis of DNA microarray spots

The rest of this tutorial will provide examples of how MVA has been applied to various systems

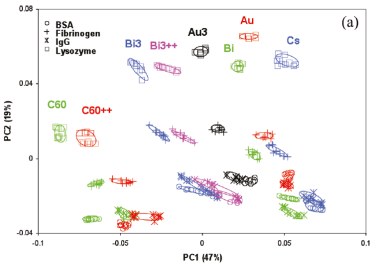
## Determining the Affects of Primary Ions

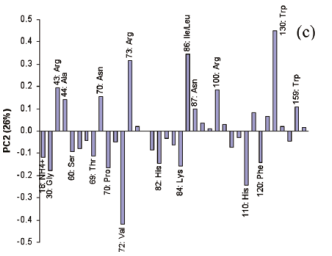
- Eight primary ion sources
- Four different proteins
- Selected AA related peaks
- Normalized to sum of selected peaks
- Mean centered

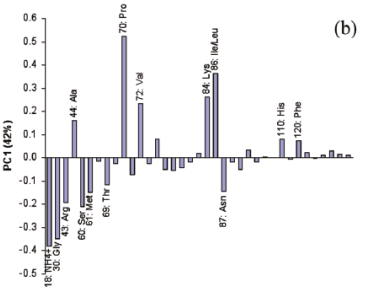
Muramoto, et. al. J. Phys. Chem. C 2011, 115, 24247–24255

In this example we were interested in determining whether proteins could be distinguished using different ToF-SIMS primary ion sources. We knew from previous data that single component proteins could be identified based off the ToF-SIMS fragmentation pattern of the amino acid fragments. We now wanted to test to see if this held true with different primary ion sources. For this we tested 5 different primary ions (C60+, C60++, Bi3+, Bi3++, Au3++, Au+, and Cs+) and 5 different proteins (BSA, Fibrinogen, IgG, Lysozyme).

# PCA Results




- Primary Ion caused largest variation
  - More fragmentation
    - Other trends were consistent between different primary ions
  - Important to keep primary ion consistent across data sets

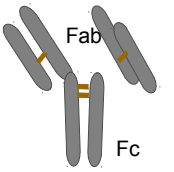
Muramoto, et. al. J. Phys. Chem. C 2011, 115, 24247–24255

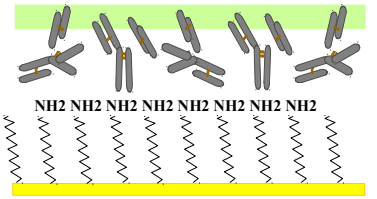
The top left plot shows the PC1 vs PC2 scores. The bottom left plot shows the PC1 loadings and the upper right plot shows the PC2 loadings. As can be seen in the data, PC1 separates the samples mainly based on the primary ion source used. This means that the greatest difference within the data set are caused by the type of primary ion used. It is noted from the loadings on PC1 that the C60 primary ions seem to cause more fragmentation resulting in an increase in lower mass fragments.

PC2 mainly captures the differences between the proteins. It is noted that the trend in the PC2 scores for the different proteins is consistent across all ion sources. This means that the trends are consistent and that one should make sure they use the same ion source for a given set of experiments.

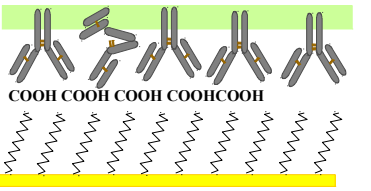
# ToF-SIMS Detection of IgG Orientation

- **Can SIMS be used to determine antibody orientation?**
- **Adsorb anti-hCG on self-assembled monolayers with different surface charges**
- **Anayze**
  - **Fab, Fc anti-hCG on Au**
  - **anti-hCG on SAMS**




anti-hCG on NH2- SAM




anti-hCG on COOH- SAM

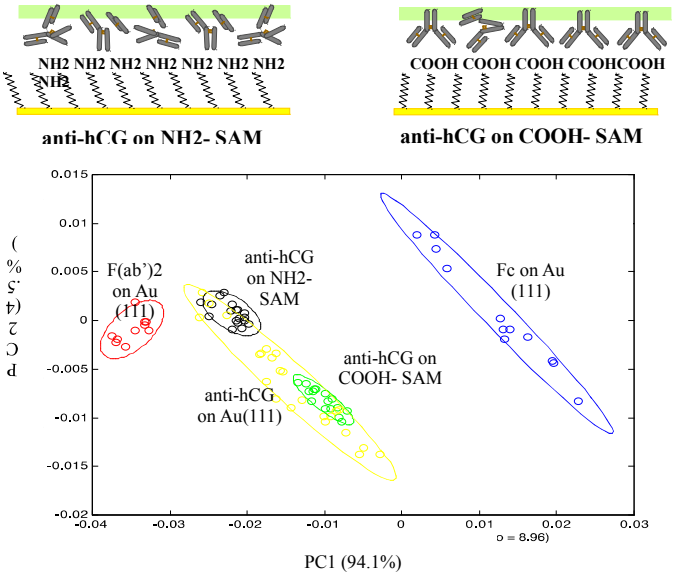
Wang, Castner, B. Ratner, Jiang, *Langmuir* **20** (2004) 1877.

In this example, PCA was used to help determine the orientation of an antibody on two different charged surfaces. Self-assembled monolayers of amine terminated (positive) and carboxylic acid terminated (negative) thiols on gold were used to create charged surfaces. Since antibodies have a dipole, it is expected that the FC section of the antibody would be attracted more to the positively charged SAM and the FAB section would be attracted to the negatively charged SAM.

SIMS data was collected on:
FC fragment on gold (control for FC fragment)
FAB fragment on gold (control for FAB fragment)
Anti-hCG on gold (random orientation)
Anti-hCG on amine terminated SAM
Anti-hCG on COOH terminated SAM

All data was analyzed using PCA.

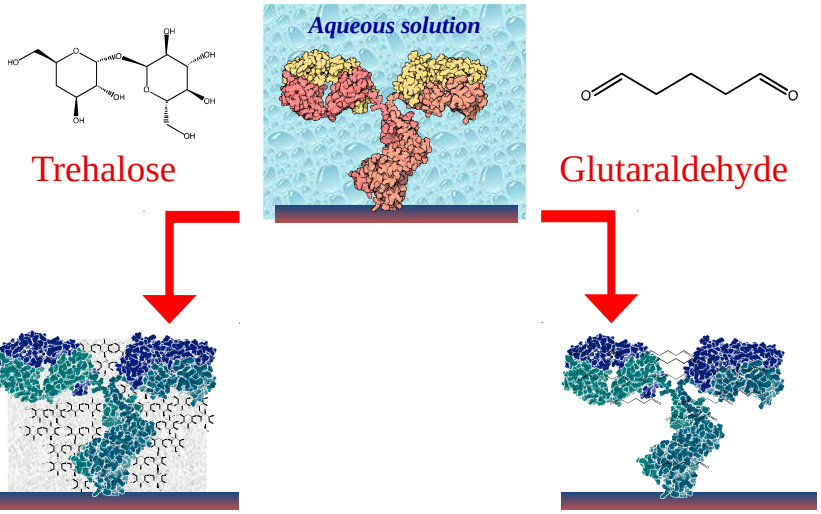
# ToF-SIMS Detection of IgG Orientation




Wang, Castner, B. Ratner, Jiang, *Langmuir* **20** (2004) 1877.

This slide shows the PC1 vs PC2 scores plot for all samples. As seen in the figure, PC1 clearly separates the FC (positive scores on PC1) and FAB (negative scores on PC1) controls. The anti-hCG on the COOH SAM samples are seen to be closer to the FC control than the FAB control, suggesting those samples are spectrally more similar to the FC control. The anti-hCG on the amine terminated SAM samples are located closer to the FAB control, suggesting those samples are more similar spectrally to the FAB control. This suggests that the antibody shows some orientation on the charged surfaces. Since the samples on the COOH SAM do not overlap completely with the FC control and the samples on the amine SAM do not overlap completely with the FAB control it is clear that the antibodies are not completely oriented. However, there is clearly some orientation. Also, the anti-hCG on gold shows scatter between both controls suggesting a random orientation. The orientation of the antibody was further verified using other analytical methods.

# *Preserving Protein Structure*

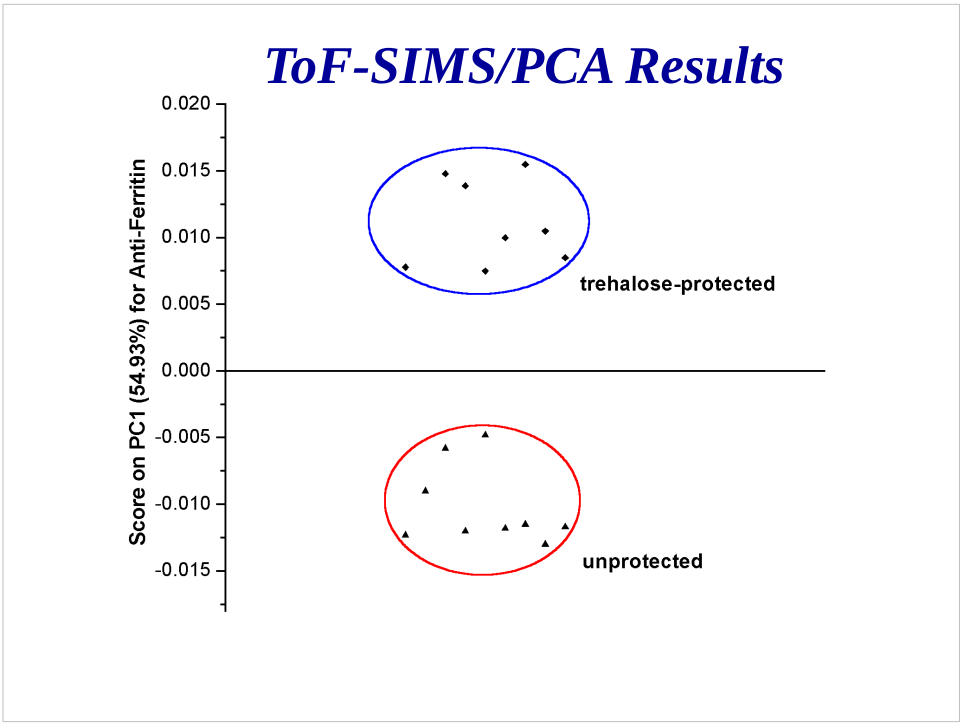



Xia, May, McArthur and Castner, Langmuir, **18** (2002) 4090.

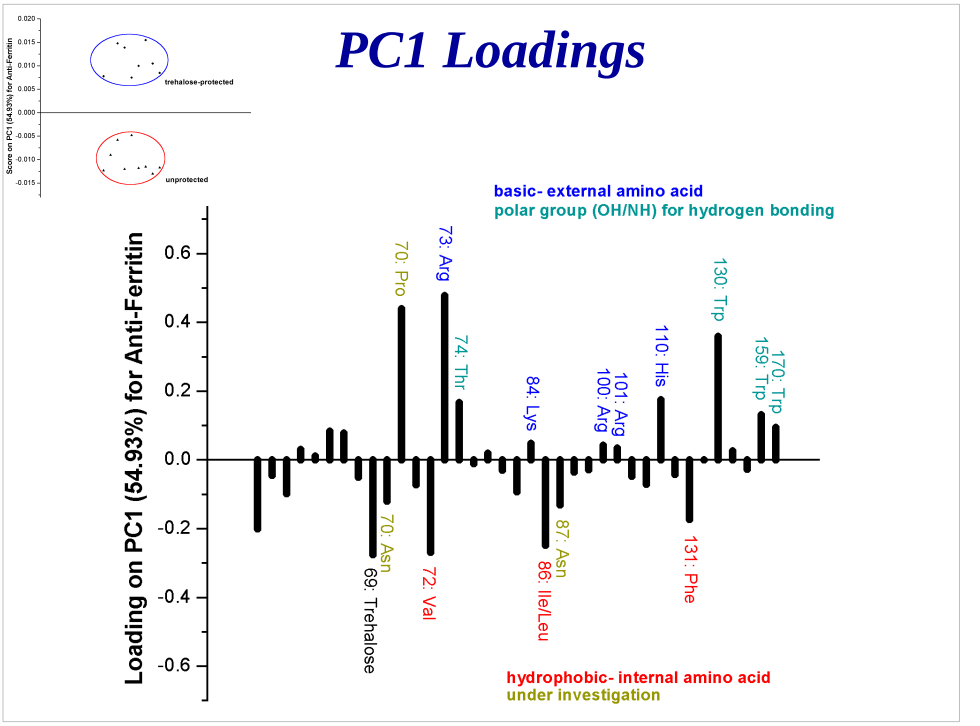
Xia & Castner, *JBMR* **67A** (2003) 179.

In this example PCA was used to investigate if Trehalose could be used to protect a protein in an ultrahigh vacuum environment. One concern of analyzing biological things in an ultrahigh vacuum environment is that they will change their structure upon exposure to an ultrahigh vacuum environment, so you would not be analyzing its true structure.

For this experiment ToF-SIMS spectra were acquired on an unprotected protein and on a sample with protein protected using trehalose.




This plot shows the PC1 scores obtained from this data. The protected and unprotected samples are clearly different from themselves. The circles on this plot are just showing the groups and do not represent 95% confidence limits.




This is the PC1 loadings plot for this data. The scores plot has been inset in the top left corner. As seen in the loadings plot the trehalose protected samples (positive scores) correspond with fragments from basic and polar amino acids. These amino acids tend to be on the outside of a protein in normal aqueous environments. The unprotected samples (negative scores) correspond with fragments from hydrophobic amino acids. These hydrophobic amino acids tend to be on the inside of proteins in normal aqueous environments. This data suggest that the unprotected samples are somewhat denatured, exposing the inner hydrophobic amino acids, and that the trehalose is able to protect the protein from denaturation keeping the protein a more native state.

## *Modeling*

- Univariate:
  - y=mx+b
- Multivariate:
  - Y=XB+E
  - Partial least square (PLS)

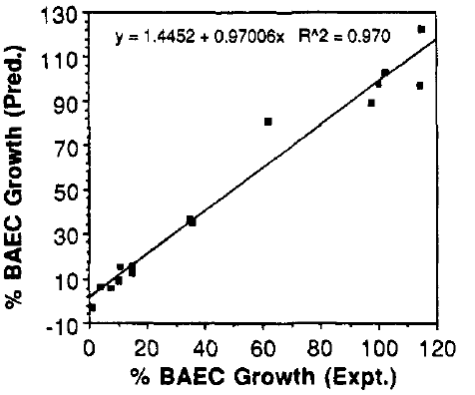
Analytica Chimica Acta, 1986, 185, 1-17

In this example I show how PLS modeling can be used to gain insight into what chemical moieties affect cell growth. PLS is a way of creating a predictive modeling in a multivariate way. It is the same idea as fitting a line to a set of data with 2 variables. For this you can use the equation 'y=mx+b' to create a linear model of how x and y are related. One can then predict a value of y based on a given x.

In PLS the same thing is done using multiple variables. In this case the equation is Y = XB + E, where Y, X, B and E are all matrices. X is a matrix of measured values (in this case ToF-SIMS peak intensities). Y is a matrix of outcomes (in this case cell growth data).

# *Predicting Cell Growth From ToF-SIMS Spectra*

- Y=XB+E
  - Y=%BAEC growth
  - X=SIMS peak intensities




Analytica Chimica Acta, 1986, 185, 1-17

In this study the authors fit their data using PLS to create a predictive model for cell growth. They used a process called cross validation to make sure the model was robust. As seen in this figure, they obtained a very good correlation between predicted and measured cell growth.

# *What Correlates With Cell Growth?*

- PLS allows you to determine what measured variables correlate with the measured response
- In this case, what peaks correlate with cell growth

**Table 4. Tentative Ion Structure Assignments for the SIMS Variables with the Largest Regression Coefficients[a]**

| *m/z* | ion structure | *m/z* | ion structure |
|---|---|---|---|
| | Positively Correlated with Cell Growth | | |
| 15+ (3.3) | $CH_3^+$ | 1− (8.7) | $H^-$ |
| 18+ (2.8) | $NH_4^+$ | 41− (3.4) | $CHCO^-$ |
| 27+ (7.8) | $C_2H_3^+$ | 42− (7.3) | $NCO^-$ |
| 29+ (5.6) | $C_2H_5^+$, $CHO^+$ | 43− (6.3) | $CH_3CO^-$ |
| 30+ (4.9) | $CH_2NH_2^+$, $NO^+$ | 55− (3.7) | $CH_3CHCO^-$ |
| 39+ (10.8) | $C_3H_3^+$ | 59− (4.2) | $CH_3COO^-$ |
| 41+ (13.4) | $C_3H_5^+$, $CHCO^+$ | 73− (7.2) | $HOCH_2CH{=}CHO^-$, $CH_3CH_2COO^-$ |
| 42+ (4.5) | $C_2H_2NH_2^+$, $C_2H_2O^+$ | | |
| 43+ (8.9) | $CH_3CO^+$, $C_3H_7^+$ | 80− (6.3) | $SO_3^-$ |
| 44+ (4.7) | $C_2H_4NH_2^+$ | 87− (7.0) | $C_2H_5COOCH_2^-$ |
| 45+ (3.0) | $C_2H_5O^+$ | 97− (13.6) | $HSO_4^-$ |
| 53+ (3.7) | $C_4H_5^+$, $C_3HO^+$ | | |
| 55+ (5.9) | $C_4H_7^+$, $C_3H_3O^+$ | | |
| 56+ (3.1) | $C_2H_4N_2^+$ | | |
| | Negatively Correlated with Cell Growth | | |
| 23+ (−4.2) | $Na^+$ | 13− (−9.0) | $CH^-$ |
| | | 24− (−5.9) | $C_2^-$ |
| 91+ (−3.6) |  | 25− (−19.0) | $C_2H^-$ |
| | | 57− (−5.8) | $C_3H_5O^-$ |
| 105+ (−4.5) |  | 71− (−6.6) | $C_3H_3O_2^-$ |

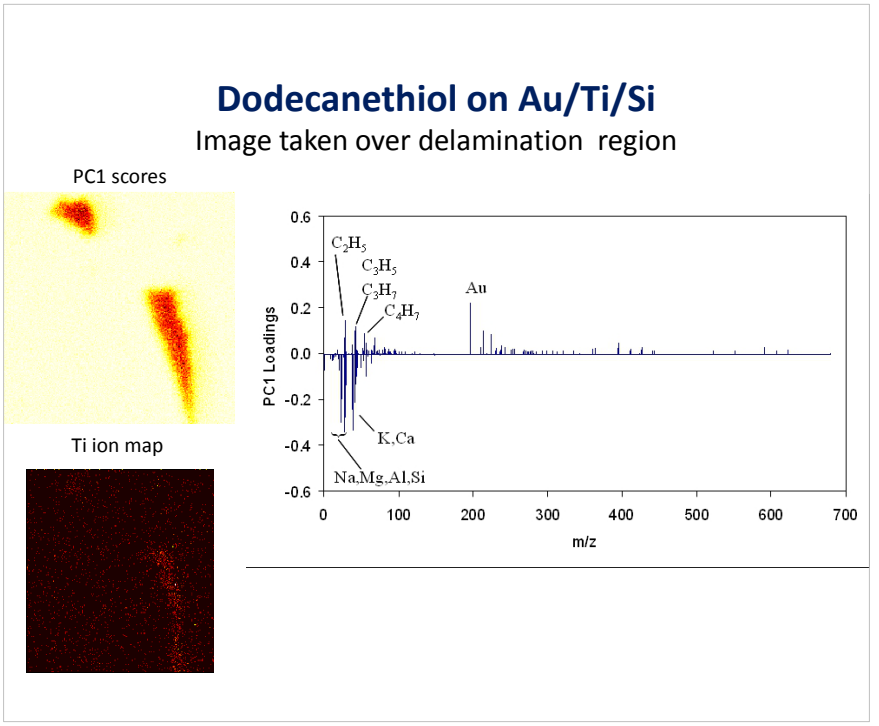
Analytica Chimica Acta, 1986, 185, 1-17

The PLS regression coefficients provide information about which peaks from the SIMS spectra correlated positively with cell growth and which peaks correlated negatively with cell growth. In this case it was found that peaks that contained nitrogen and oxygen correlated positively with cell growth, while some small hydrocarbons, cyclic structures and sodium correlated negatively with cell growth.

## PCA of ToF-SIMS Images: Solving Quality Control Issue

- PCA:Gold Delamination Problem
  - Noticed patches with missing gold on samples used for self assembly
  - Imaged patches with ToF-SIMS
  - Selected all peaks above background
  - Ran PCA on mean centered image data

This examples shows how PCA was used to help solve a problem with gold delaminating from silicon wafers.
While working with gold coated silicon wafer pieces for self-assembly of alkanethiols, we noticed that there were patches on the surface where the gold had delaminated. This started happening fairly regularly, so we wanted to figure out what was causing the problem. So we took ToF-SIMS images of areas where the gold had delaminated and then ran the data through PCA.



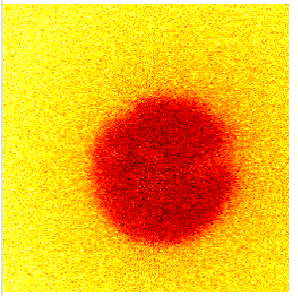

The figure on the top left shows the PC1 scores. The loadings plot on the right shows that the dark areas, where the gold has delaminated, correspond with a series of salt ions. We also checked to see if there was still any titanium on the surface since we used a titanium underlayer, and we only saw traces of titanium in the delamination areas.

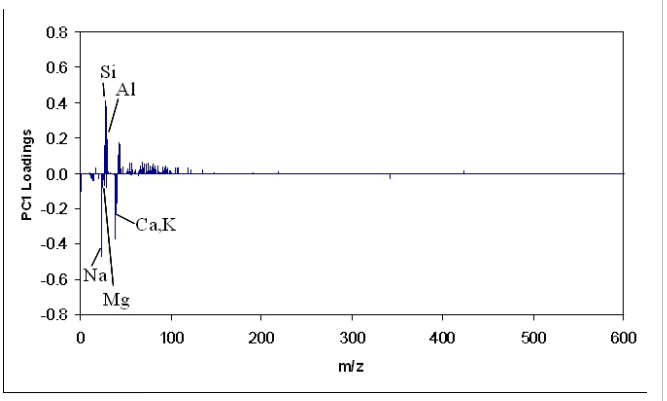
## PCA: Gold delamination problem

- Looked through gold sample preparation protocol
- Looked at samples after cleaning
- Noticed residue on the samples
- Imaged residue with ToF-SIMS
- Ran PCA on mean centered image data

We went back through our sample preparation protocol and analyzed samples from various stages within the protocol. We noticed some residue on our silicon wafers after cleaning, so we analyzed some of these spots with SIMS and processed the data with PCA.

## Bare Silicon wafer after cleaning
## Image of 'blob' on surface






This slide shows the PC1 scores and loadings from a residue spot on a silicon wafer. As seen in the figure the dark area where the residue (negative scores) corresponds with the same set of salt ions as was seen on the gold sample.

In looking back through the sample preparation protocol we found that the dicing saw we used to cut our wafers had been switched from using DI water to using house water. Since the house water was not filtered it contained salts that ended up being deposited onto the silicon.

Based on this information we added a water soak to our cleaning protocol to remove the salts and were able to eliminate the problem of the gold delamination.

## PCA: Gold delimitation problem

- PCA was able to quickly identify differences in the gold delamination areas
- Allowed generation of a hypothesis
- Isolation of the problem and adjustment to the sample preparation protocol

I think this is a good example of how PCA was able to help us quickly summarize large data sets and aid us in coming up with hypotheses that could be tested in order to solve a real problem.

# PCA: Understanding Cationization

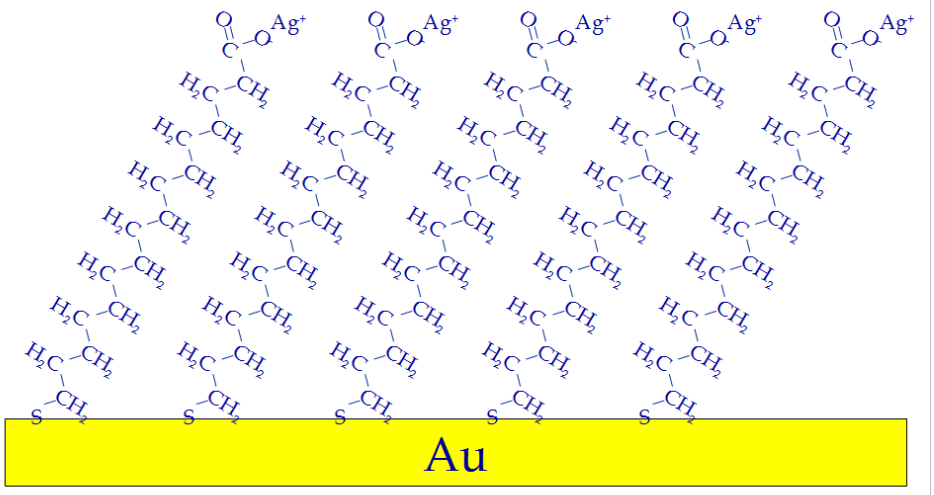
- It is known that silver ions (Ag+) can aid in the emission of large polymer fragments via cationization
  - Attachment of a cation to the molecule
- We wanted to:
  - Explore the cationization process
  - Determine if other cations also worked
  - Optimize new substrates for cationization

**Michel et al (2000) Langmuir, 16:6503-6509**

In this study we used PCA to help us understand more about the process of cationization and to determine how to optimize substrates for cationization. Years ago it was found that if you put thin layer of a polymer onto etched silver, you would get large polymer chains (~3000 amu) and fragments that contained the addition of a silver cation (thus cationization). We wanted to see if we could optimize this process using self-assembled monolayers.

# Protein and polymer cationization

## Cations stabilize large fragments during SIMS analysis

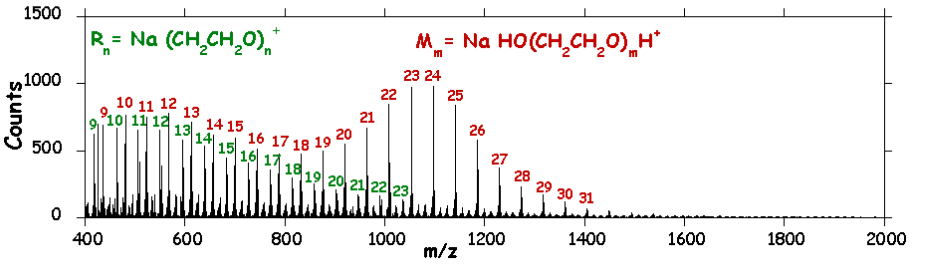



Michel et al (2000) *Langmuir*, 16: 6503-6509

For this we used carboxylic terminated self-assembled monolayers and exchanged the COOH hydrogen with various cations (Na+, K+, Ca+, Ag+, etc). It was found that these SAM substrates also worked for cationization.

# Optimization of Cationization

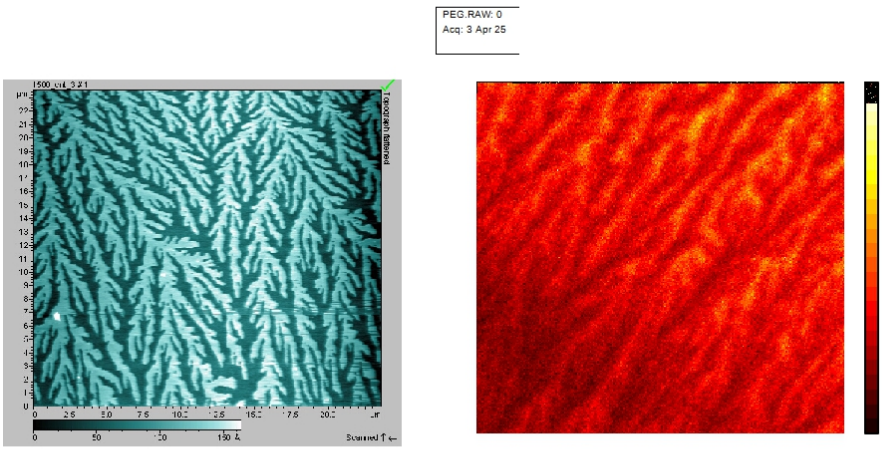
- **How can we maximize the yield of high mass fragments with cationizing SAMs?**
- **What factors affect the high mass yield?**




**Michel et al (2000) Langmuir, 16:6503-6509**

This plot shows the molecular weight distribution of a PEG 1000 polymer obtained using a Na+ cationizing SAM. With a working model system, we then wanted to see how we could optimize the yield of this higher mass molecules and fragments. During this process we noticed some oddities in the way the PEG polymers coated the surface.

# PEG on COONa SAM

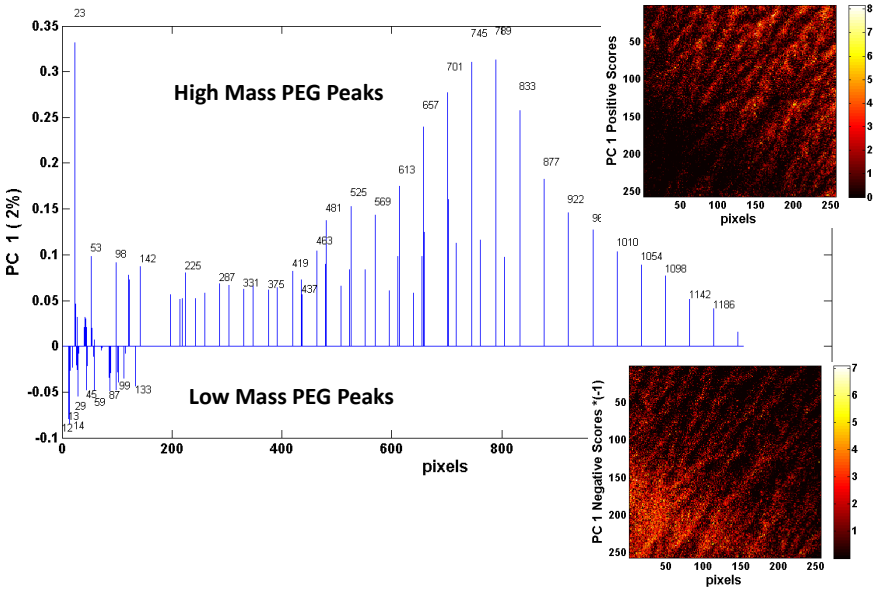



AFM Image

TOF-SIMS Image

These figures show AFM and ToF-SIMS images from a surface coated with PEG. This type of crystal formation is typical with PEG chains under certain conditions. We wanted to determine where the highest yield of the high mass peaks originated (this time from a PEG 800 polymer). So we ran PCA on the ToF-SIMS image.

# PCA of PEG on Cationizing SAM



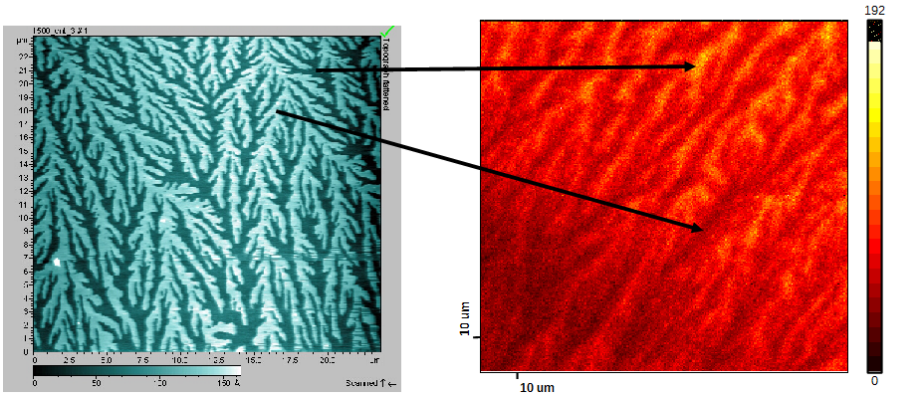

This slide shows the scores and loadings from the ToF-SIMS image. The upper right image shows the PC1 positive scores, the lower right image shows the PC1 negative scores.

# What does it mean?

**-It is believed that the high regions in the AFM image correspond with the dark regions in the TOF-SIMS image**

**-Thinner areas (brighter regions on SIMS image) produce higher yield of high mass fragments**

**-SIMS beam can actually sample through polymer to $Na^+$ ions**




From comparing the AFM and ToF-SIMS PCA results, we determined that the higher (thicker) areas in the AFM images corresponded with the darker regions in the ToF-SIMS images. The thinner areas in the AFM image corresponded with the brighter areas in the ToF-SIMS image. The brighter areas in the ToF-SIMS image corresponded with areas with more high mass PEG peaks. This means that one gets better yield of high mass peaks with thin polymer layers where the primary ion beam can penetrate the layer and there can be direct interaction with the underlying cations.

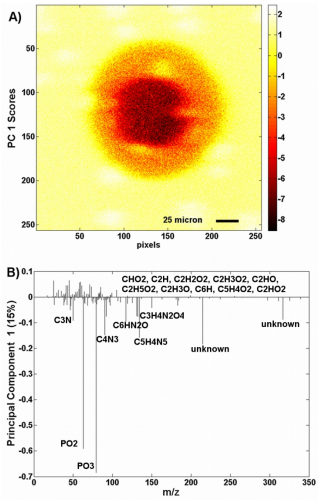
# MVA of DNA Microarray Spot

- Using ToF-SIMS to analyze DNA microarray spots
  - Check uniformity of spots
  - Look for chemical variation
  - Use results to feedback into spotting methodology to get better signal to noise
- Selected all peaks above background (negative ion data)
- Poisson scaled and mean centered
- Ran PCA, MAF and MCR

Data from Lara Gamble, Nicolas Vandecasteele

This last example is used to compare the results one can obtain from PCA, MAF and MCR. This data comes from the analysis of DNA spotted onto a standard microarray slide. The slide was coated with a silane linker to which a PEG polymer was attached to create a “non-fouling” background. The purpose this study was to determine if the DNA was being spotted uniformly and help optimize the performance of the slides by understanding the chemistry of the surface.

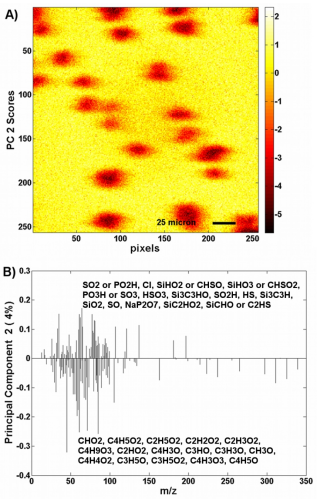
# PCA Results




- PC1
  - Separates DNA spot from background
  - DNA spot is not uniform
  - White dots appear across image

PCA PC1 shows the DNA spot and some white spots across a uniform background. Several things are noted, include the fact that the DNA spot is not uniform. The loadings plot shows that the DNA spot (negative scores) corresponds with peaks expected from DNA bases. The background (positive scores) corresponds with peaks related to PEG.

# PCA Results




- PC2
  - Highlights dots
  - Peaks suggest presence of PEG
    - Known to crystallize and form agglomerate on some surfaces

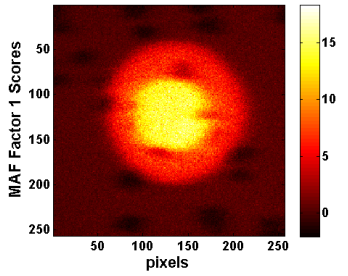
PCA PC2 highlights the spots across the surface (negative scores). These areas correspond with peaks related to PEG. The positive scores correspond with peaks related to the silane linker.

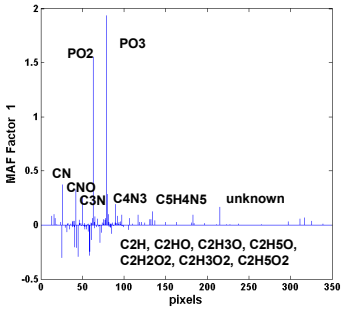
# MAF of DNA Microarrays

- For PCA data matrix X is decomposes as:
  - S = UTX
    - S=scores
    - U = Loadings = eigenvector rotation of covariance matrix of X
- For MAF
  - U = eigenvector rotation of B where:
    - B=A-1V
      - V= covariance matrix of X
      - A=covariance matrix of shift matrix
        - Shift matrix = X-X shifted by 1 pixel in X or Y
    - Finds linear combination of peaks that maximize variation across the image while minimizing variation between neighboring pixels

This slide describes the basic background for MAF. MAF is calculated using what is called a shift matrix. By using this matrix MAF is able to find a linear combination of peaks that maximize variation across the image while minimizing variation between neighboring pixels. MAF is scale independent, meaning that you will get the same answer regardless of how you scale the original matrix.

# MAF Results

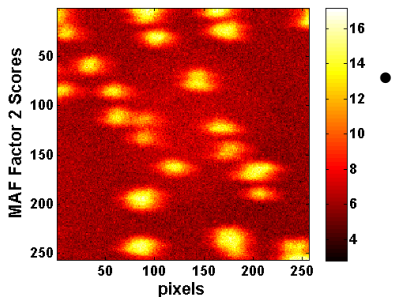


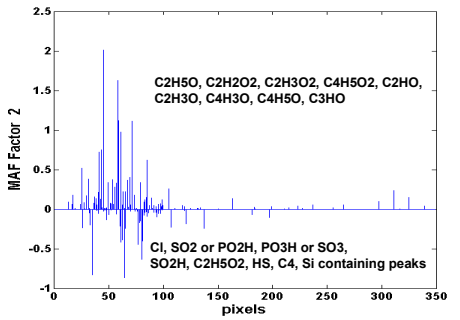

- Factor 1
  - Separates DNA spot from background
  - Contrast is inverted from PCA, but peaks are the same

MAF factor 1 looks very similar to PCA PC1, except that the contrast is inverted. The peaks corresponding to the various regions is the same.

# MAF Results




- Factor 2
  - Highlights dots
  - Peaks suggest presence of PEG
    - Consistent with PCA results



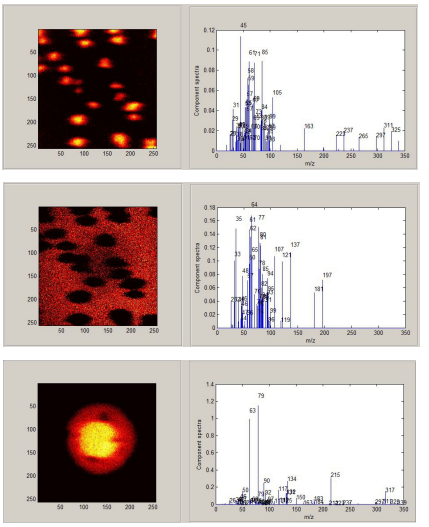

MAF Factor 2 is also very similar to PCA PC2. The results show the same thing.

# MCR of DNA Microarrays

- MCR tries to describe the data as a linear model:
  - X=CFT+E
  - F=Spectra of pure components
  - C=concentration of components at each pixel
  - E=random error
- C and F are found using a least squares minimization
- Requires an initial guess of how many components there are
- For this data we used PCA results as initial guess for MCR of DNA microarray spot
  - MCR was run on Poisson scaled/mean centered data

This slide provides a basic summary of MCR. MCR is often stated as being "better" than other methods because it can find "pure" components. I do not agree that MCR is "better" than other methods. It is just another method that can be helpful for ToF-SIMS. HOWEVER, MCR must be used with caution. There are in infinite number of solutions possible with MCR. The answer you get is completely dependent on the initial guess you use. Furthermore, I have never seen a data set where MCR has provided any new information that cannot be learned from PCA. In my opinion MCR is only useful to display the components you find after doing a thorough analysis using PCA.

# MCR Results



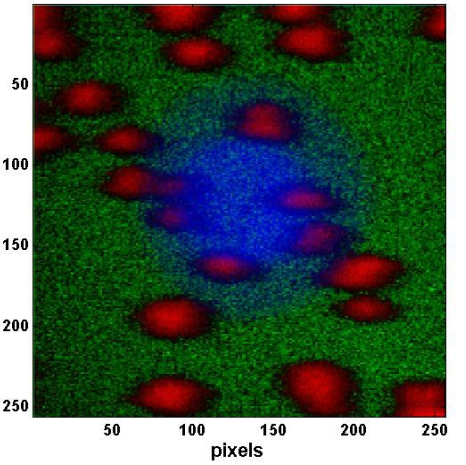
- Component 1 = PEG
- Component 2 = Silane linker/buffer salts
- Component 3 = DNA

The components shown on this slide are the result of MCR using the first 3 PCs as the initial guess. MCR picks out the same 3 components seen with PCA or MAF, namely the DNA spot, the PEG spots and the silane linker background.

MCR does provide nice looking images and WHEN it finds a logical solution it provides spectra that typically contain only peaks from the component it finds. However, in my opinion there is no new information provided by running MCR. As I mentioned before, you should use MCR with caution since it will always give you an solution, but you have to determine if the solution makes sense.

# RGB Overlay of MCR Components

- Red = PEG
- Green = Silane/Salts
- Blue = DNA



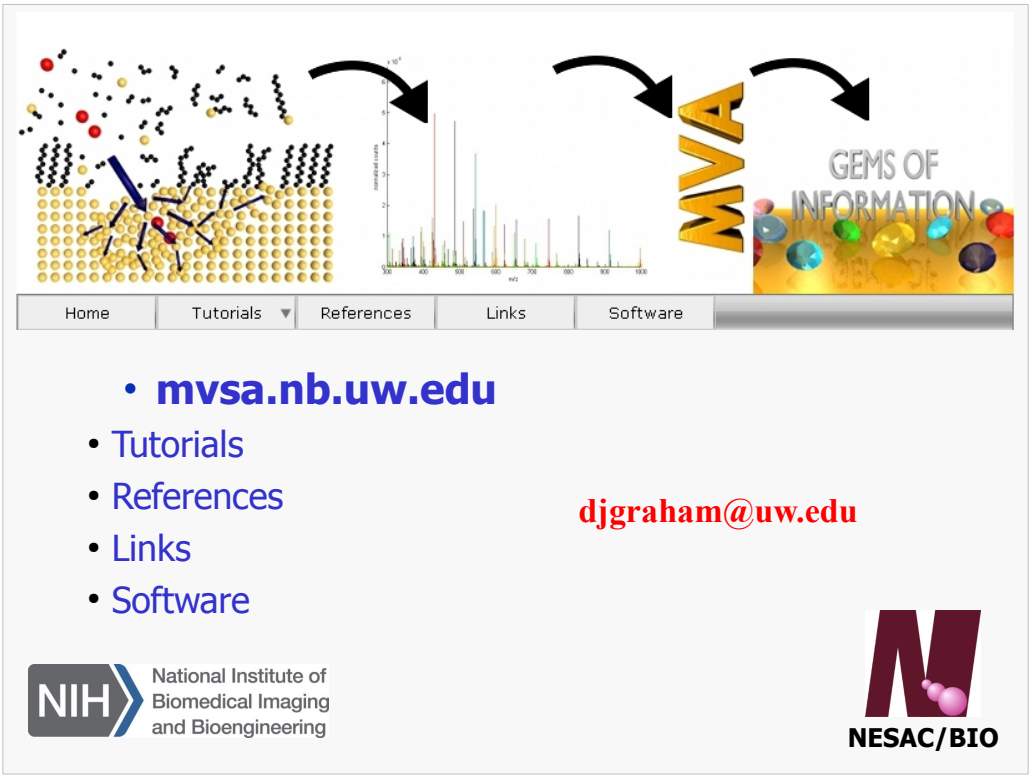

MCR does allow you to make very nice overlay images if you overlay the components it finds.

# Summary

- MVA methods are powerful tools to help with ToF-SIMS data analysis
- Successful ToF-SIMS experiments using MVA require good experimental plans
- The method you use should be chosen based on the needs of your analysis
- Data preprocessing should be done based off logical assumptions about the data

In summary, MVA can be very useful. You should choose the method you use based on what you want to learn about your data. I find that PCA is always a good starting point, and often is all you will need. Other methods can be useful depending on the goal of your analysis.

Make sure you understand what you are doing and that you understand the assumptions you are making when you use any MVA method.




Visit the website to learn more.