Biomarker Discovery: Case Studies, Pitfalls, and Successes

Heidi M. Spratt, Ph. D.
Institute for Translational Science
Assoc. Prof., Depts. of Preventive Medicine and Community Health
Biochemistry and Molecular Biology
University of Texas Medical Branch

utmb Health
Bioinformatics Program

2 August, 2014

Biomarker Discovery

- Theory – combine data from multiple sources
  - Histology labs
  - Genomics
  - Proteomics
  - Metabolomics, etc
- To obtain useful biomarkers for
  - Disease discovery
  - Prognosis
  - Treatment
  - Drug discovery

Biomarker Discovery

- What is a Biomarker?
  - Urine test,
  - Blood test,
  - Tissue sample, or
  - Other bodily fluid test (Exhaled Breath Condensate, Bronchial Lavage)
- Can be used to assess disease status or treatment options

Biomarker Discovery

- Questions to be answered include:
  - Is there a panel of markers that will enable clinicians to accurately distinguish one disease from a more severe form of the disease?
  - Is there a panel of markers that will enable clinicians to predict which patients are more likely to survive a severe burn?
  - Is there a panel of markers that will enable clinicians to predict which subjects will progress to a more severe form of infection?

Biomarker Discovery

- Methods to create such panels – use Machine Learning/Classification algorithms
  - Take select data as input
  - Split the data into a training set and a test set (can use CV for this)
  - Run the algorithm
  - Create a classifier which accurately groups the data
  - Accuracy of the classifier is also computed

Biomarker Discovery

- Discovery
  - Performed on initial data to identify potential markers that are useful for prediction
- Qualification
  - Using the same samples as the discovery samples, can you verify the findings you have using a different technique
- Verification
  - Using completely new samples, does your biomarker panel work as well
- Problems frequently appear in the Qualification or Verification stages

Analytic Techniques

- Classification and Regression Trees
- Multivariate Adaptive Regression Splines
- Random Forests
- Support Vector Machines
- Generalized Path Seeker

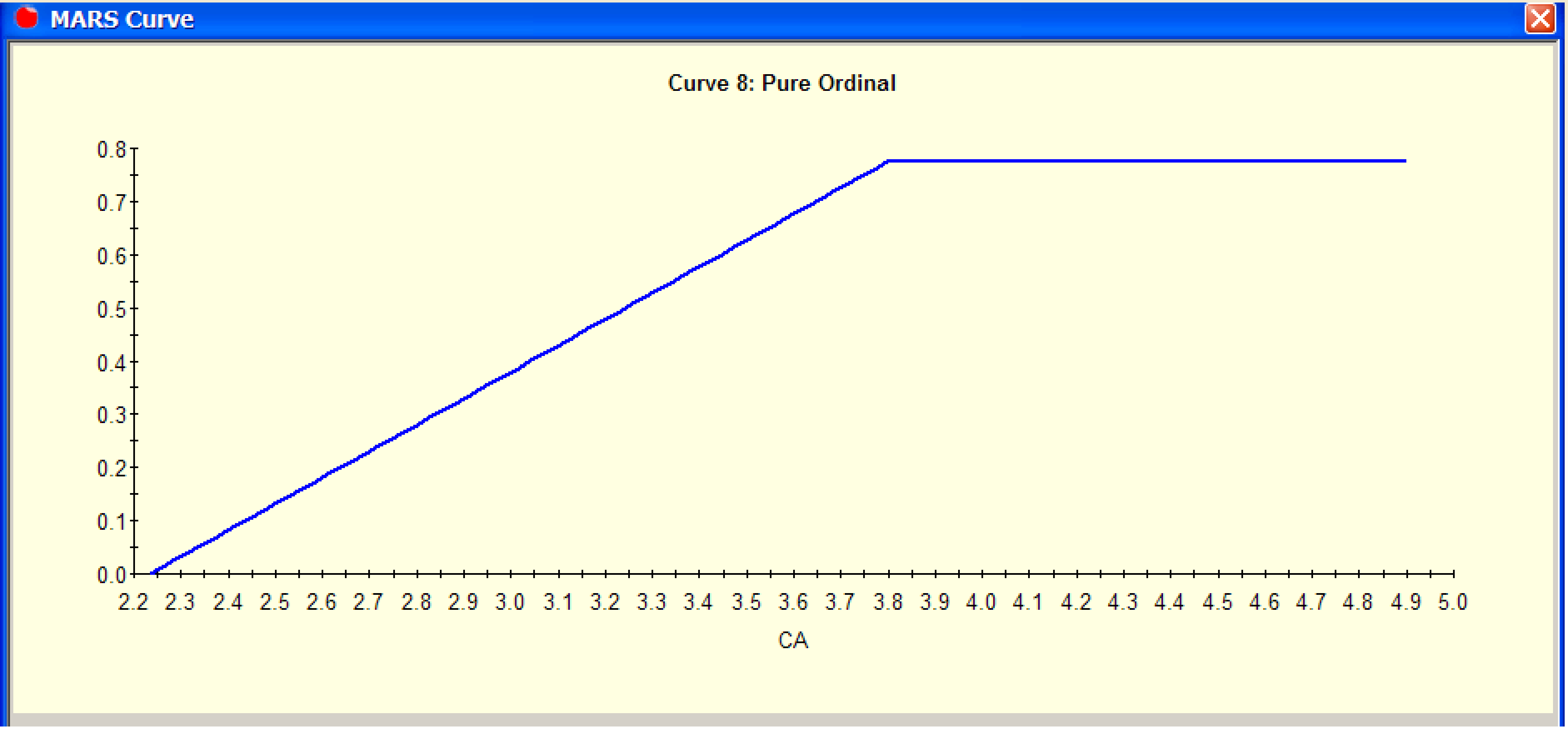
Analytic Techniques

- Classification and Regression Trees
- Multivariate Adaptive Regression Splines
- Random Forests
- Support Vector Machines
- Generalized Path Seeker

Multivariate Adaptive Regression Splines (MARS)

- Nonparametric regression procedure that creates models based on piecewise linear regressions
- Searches through all predictors to find those most useful for predicting outcomes
- Optimal model is created by a series of regression splines called basis functions

MARS – Regression Spline



MARS Model Basics

- MARS builds models of the form

$$f(x) = \sum_{i=1}^{k} c_i B_i(x)$$

- Each basis function $B_i(x)$ takes one of the following three forms:
  - – a constant 1. There is just one such term, the intercept.
  - – a *hinge* function. A hinge function has the form max(0,*x* − *const*) or max(0,*const* − *x*). MARS automatically selects variables and values of those variables for knots of the hinge functions.
  - – a product of two or more hinge functions. These basis function can model interaction between two or more variables.

MARS

- Uses a two-stage process for constructing the optimal classification model.
- First half of the process involves creating an overly large model by adding basis functions that represent either single variable transformations or multivariate interaction terms.
- Model becomes more flexible and complex as additional basis functions are added.

MARS

- The process is complete when a user-specified number of basis functions have been added.
- Second stage, MARS deletes basis functions in order of least contribution to the model until the optimum one is reached.

MARS

- End result is a classification model based on single variables and interaction terms which will optimally determine class identity.
- Excels at finding thresholds and breaks in relationships between variables
- Is very well suited for identifying changes in the behavior of individuals or processes over time

MARS

- Advantages
  - Can take any form of predictor variable – continuous or categorical
  - Number of predictors not a problem
  - Fast algorithm
  - Handles missing values
- Disadvantages
  - Easy to overfit data
  - More than single interaction has no biological meaning

MARS Accuracy

- To access the model accuracy, we look at the prediction success rate and the ROC curves
- Ideally, we'd like a model that is 100% accurate at identifying our different groups of subjects
- We can check how well the model performs on our data, but it might be overfit
- What we need to do is somehow leave some of the data points out when we are building our model and test the accuracy of the model later on those left out patients
- This is done in one of 2 ways
  - Cross-validation
  - Leaving a proportion of the samples out from the beginning

Case Studies

- Hepatitis C Virus / Hepatocellular Carcinoma
- Dengue Fever / Dengue Hemorrhagic Fever
- Helicobacter Pylori / Peptic Ulcer Disease
- Invasive Aspergillosis

Hepatitis C Virus

- To be able to accurately classify patients with Hepatitis C Virus (HCV) versus patients with Hepatocellular Carcinoma (HCC)

Hepatitis C Virus

- One of the main causes of liver cancer in the US
- Causes an estimated 10,000 – 12,000 deaths per year in US
- Virus varies greatly in both course and disease outcome
- Many patients are asymptomatic
- Many have some degree of chronic hepatitis associated with a degree of fibrosis of the liver
- Prognosis for early stage HCV is good
- Prognosis for late stage poor as usually progresses to liver cirrhosis

Progression of HCV

- The time it takes to progress form one stage to the next is lengthy
- 20 years can pass before a person develops cirrhosis of the liver
- After 20-40 years, a very small percentage of the population will progress to HCC
- Little is known about which patients progress or why

Liver Cancer (HCC)

- Plagued with many of the same problems as HCV
- Cancer usually not detected until late stages
- Prognosis not good because treatment at late stages not very effective
- Early detection is wrought with problems due to many false positives and negatives
- Prognosis hopeful if detected early

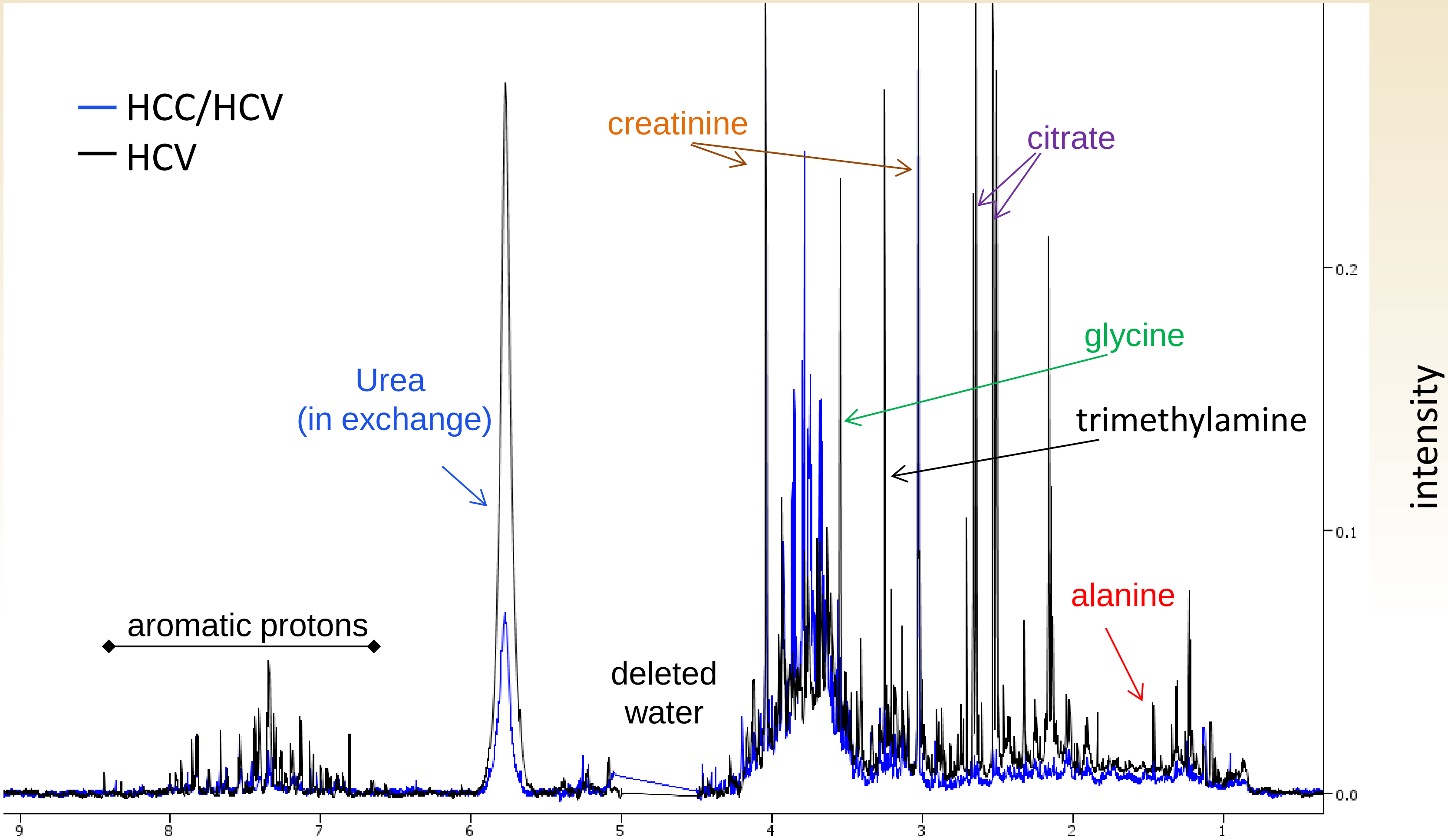
Experimental Set-up

- Patients
  - 27 late stage HCV
    - stage 2: 17
    - stage 3: 4
    - stage 4: 6
    - Patients co-infected with Hepatitis B, HIV, or HCC or those that have had an organ transplant are not considered for this study
    - Patients receiving antiviral therapy within the past year were excluded as well
    - Mix of races
  - 36 HCC
    - Caucasian, all stage 4 HCV
  - Patients actively consuming ethanol are excluded from both groups

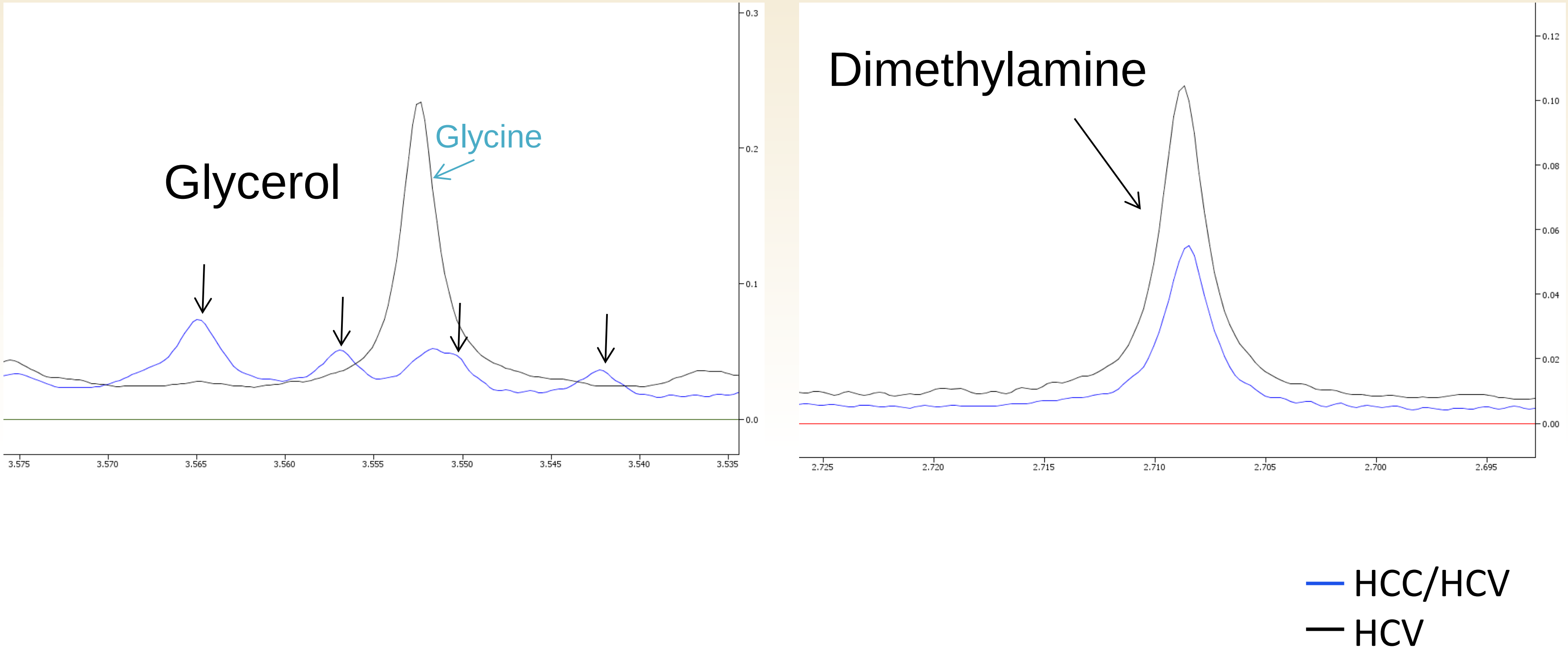
Experimental Methods

- Liver biopsies were performed on all patients to identify their stage of disease
- Typical blood chemistries were obtained for all patients as well
- Urine was collected from all patients
- NMR was performed on all samples

^1H NMR of HCC/HCV and HCV Samples



Discriminatory Analytes



Significant Metabolites

- 1-Methylnicotinamide
- 2-Oxoglutarate
- 4-Hydroxyphenylacetate
- Acetate
- Alanine
- Carnitine
- Dimethylamine
- Ethanolamine
- Ferulate
- Formate
- Fumarate
- Glucose
- Glycerol
- Hypoxanthine
- O-Acetylcarnitine
- Quinolinate
- Taurine
- Tyrosine
- Xanthosine
- cis-Aconitate

MARS - Prediction Success

Training

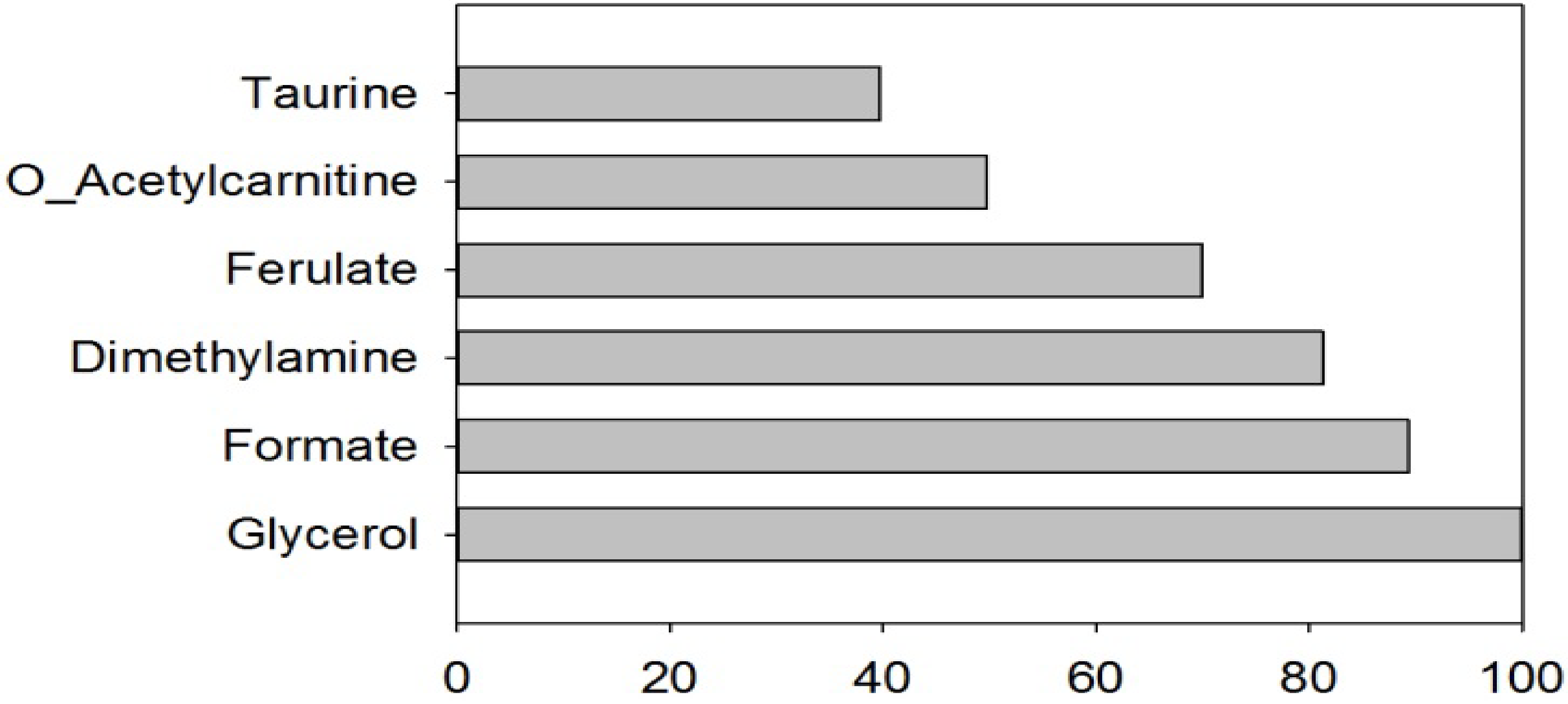
Class	Total	Prediction	
		HCV (n=32)	HCC (n=31)
HCV	27	1	26
HCC	36	30	6
Total	63	correct =96.30%	correct =83.33%

Testing

Class	Total	Prediction	
		HCV (n=32)	HCC (n=31)
HCV	27	3	24
HCC	36	28	8
Total	63	correct =88.89%	correct =77.78%

Variable Importance

Metabolites



% Variable Importance

Dengue Fever

- Mosquito-born flavivirus infection for which 2/5th’s of the world’s population is at risk
- Broad spectrum of diseases ranging from asymptomatic to a flu-like state
- 4 serotypes of disease
- 2 classes of disease
  - DF
  - DHF

Dengue Fever

- Infection is categorized by:
  - Fever
  - Headache
  - Muscle and joint pains
  - Characteristic skin rash that is similar to measles

Dengue Hemorrhagic Fever

- Infection is categorized by:
  - High fever
  - Coagulopathy
  - Vascular leakage
  - Hypovolemic shock
- No current drug therapy for DHF
- Fatality rates can exceed 20%
- Early and intensive therapy can bring it down to < 1%
- Must have had a prior DF infection with a different serotype

Study Data

- 52 patients
  - 30 DF
  - 22 DHF
- Created Model from Discovery data that has classification above 90% accuracy for training & testing data, with high AUC's
- Developed Selected Reaction Monitoring (SRM) assays for all markers in Discovery model
- Ran SRM assays on Qualification samples (same as Discovery)
- MARS results from SRM Qualification data presented

MARS - Prediction Success

Training

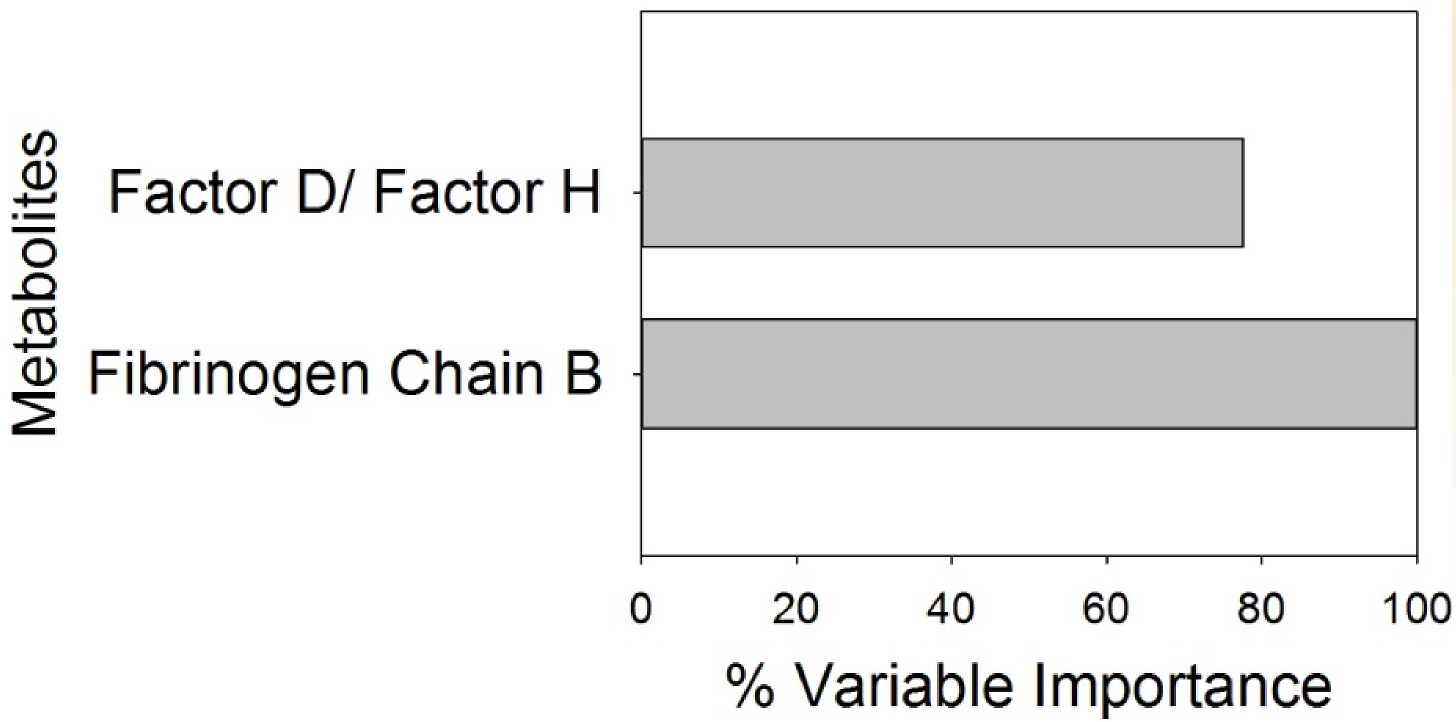
Class	Total	Prediction	
		DF (n=27)	DHF (n=25)
DF	30	25	5
DHF	22	2	20
Total	52	correct =83.33%	correct =90.91%

Testing

Class	Total	Prediction	
		DF (n=28)	DHF (n=24)
DF	27	23	7
DHF	36	5	17
Total	52	correct =76.67%	correct =77.27%

ROC AUC Training: .94; ROC AUC Testing: .78

Variable Importance



DF - Issues

- Verification samples do not work well with these predictors, or any others that we have worked with
- Verification samples were plasma and initial samples were serum
- Initial samples were all DF serotype 3, verification samples were all serotypes

H Pylori

- 30 H Pylori patients
- 30 H Pylori + Peptic Ulcer Disease patients
- Goal is to determine if there are any proteins/amino acids that discriminate between H Pylori patients that will develop PUD vs those H Pylori patients that will not (treatment & prognosis are different)

H Pylori Data

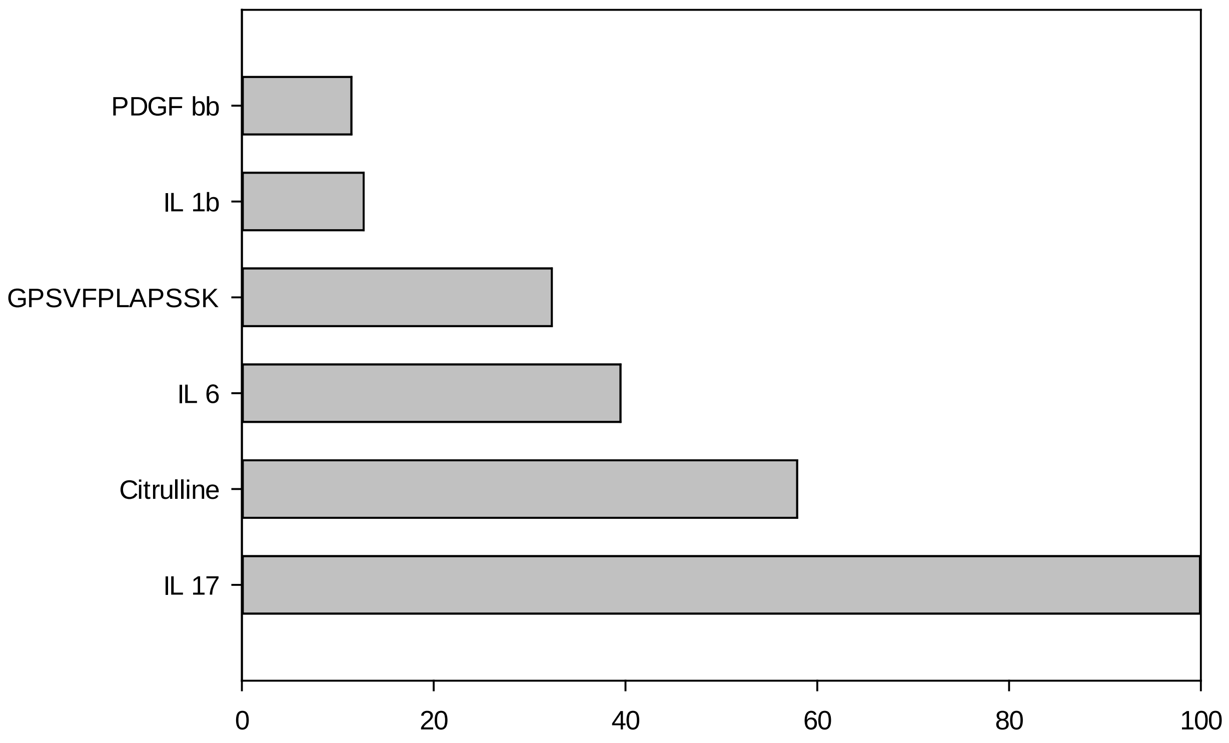
- Panel of Cytokines
- Panel of Amino Acids
- 2D gel Proteomics spots
- O16/O18 labeled peptides

Prediction Success

Actual Class	Total Cases	Percent Correct	H Pylori only N=30	PUD N=30
H. Pylori only	30	96.67	29	1
PUD	30	100	0	30

- Overall Accuracy: 98.33% Training Data only
- Results are similar for Testing data, with overall accuracy being 95%
- AUC is very high for both as well (> .95)

Variable Importance

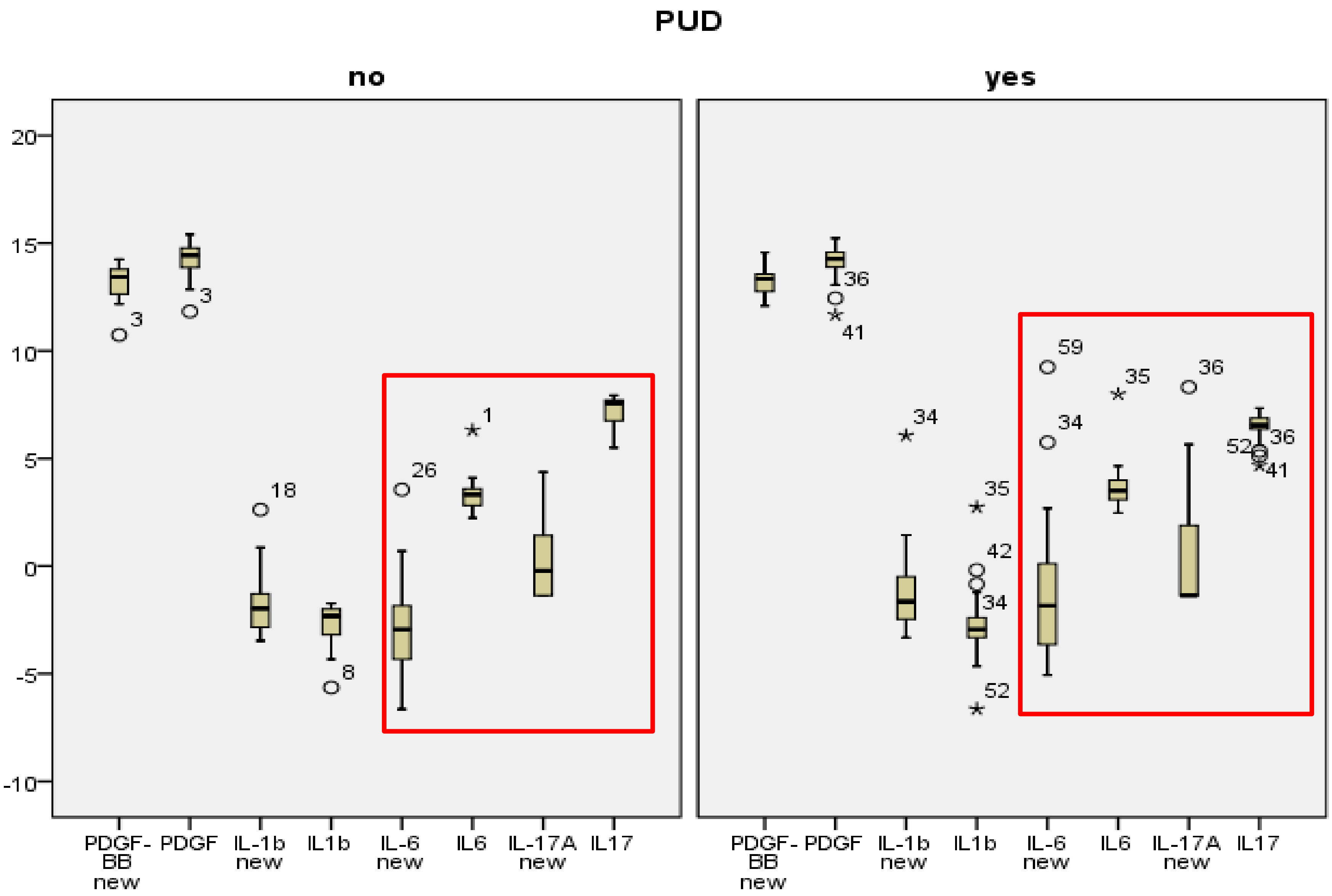


GPSVFPLAPSSK – Ig gamma-1 chain C region

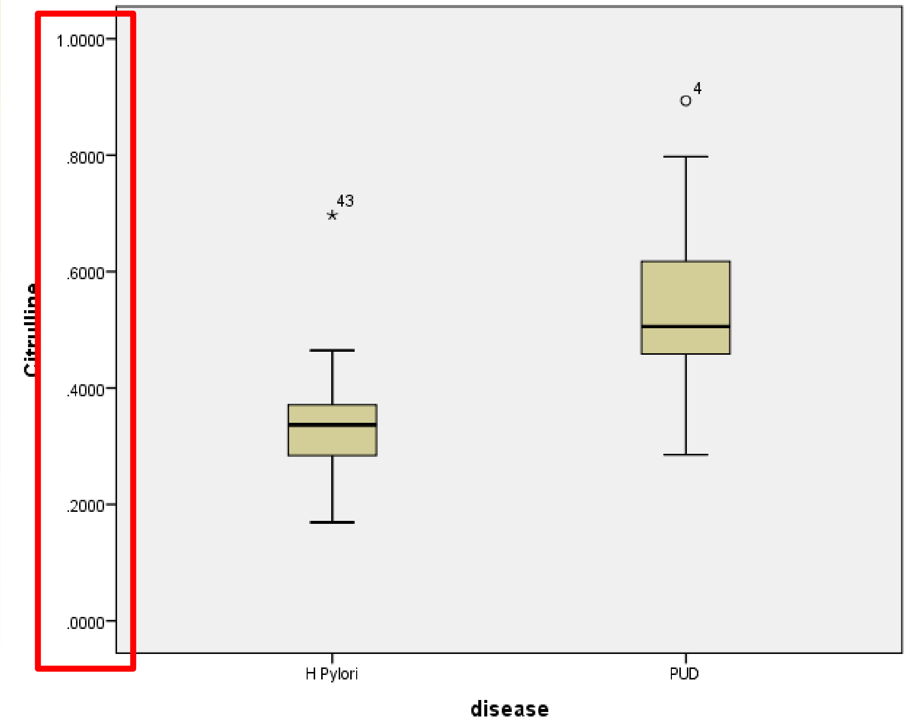
Japanese Sample Analysis

- Amino Acid Analysis
  - Same platform was used for AA as previous
  - Identified 38 amino acids in the samples
  - Since data is not normally distributed, even after log2 tranform, ran nonparametric tests to look for differences between H Pylori only and H Pylori + PUD
  - 1 AA significantly different : alpha-Amino-n-Butyric Acid
- Cytokine Analysis
  - Ran 4 of the original 8 cytokines
  - PDGD-bb, IL-1b, IL-17, and IL-6
  - None significantly different via Mann-Whitney

Cytokine Differences?

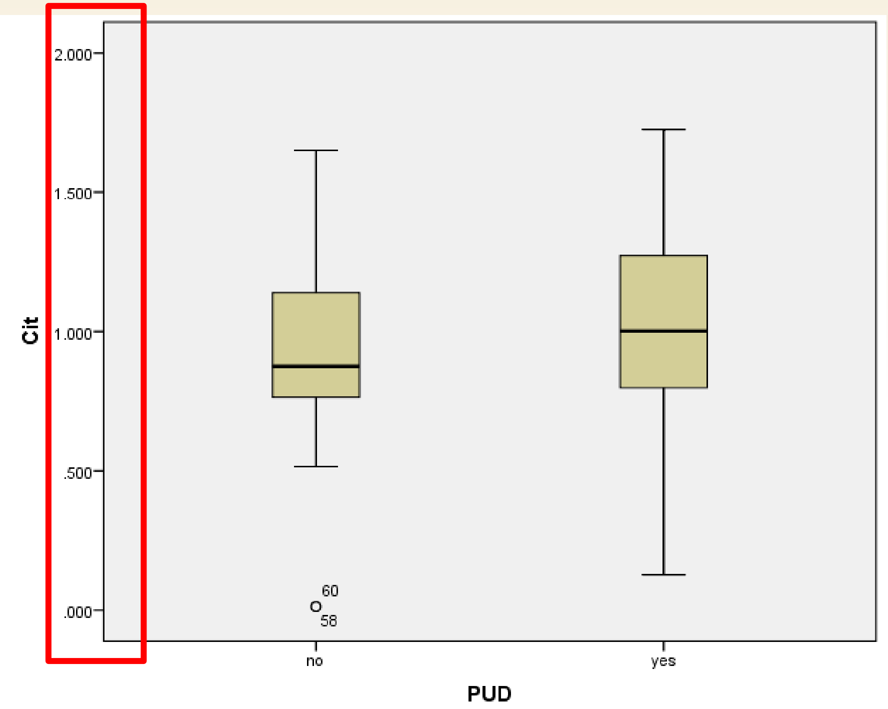


Amino Acid Differences?



Original samples, raw, imputed

Japanese samples, raw, imputed



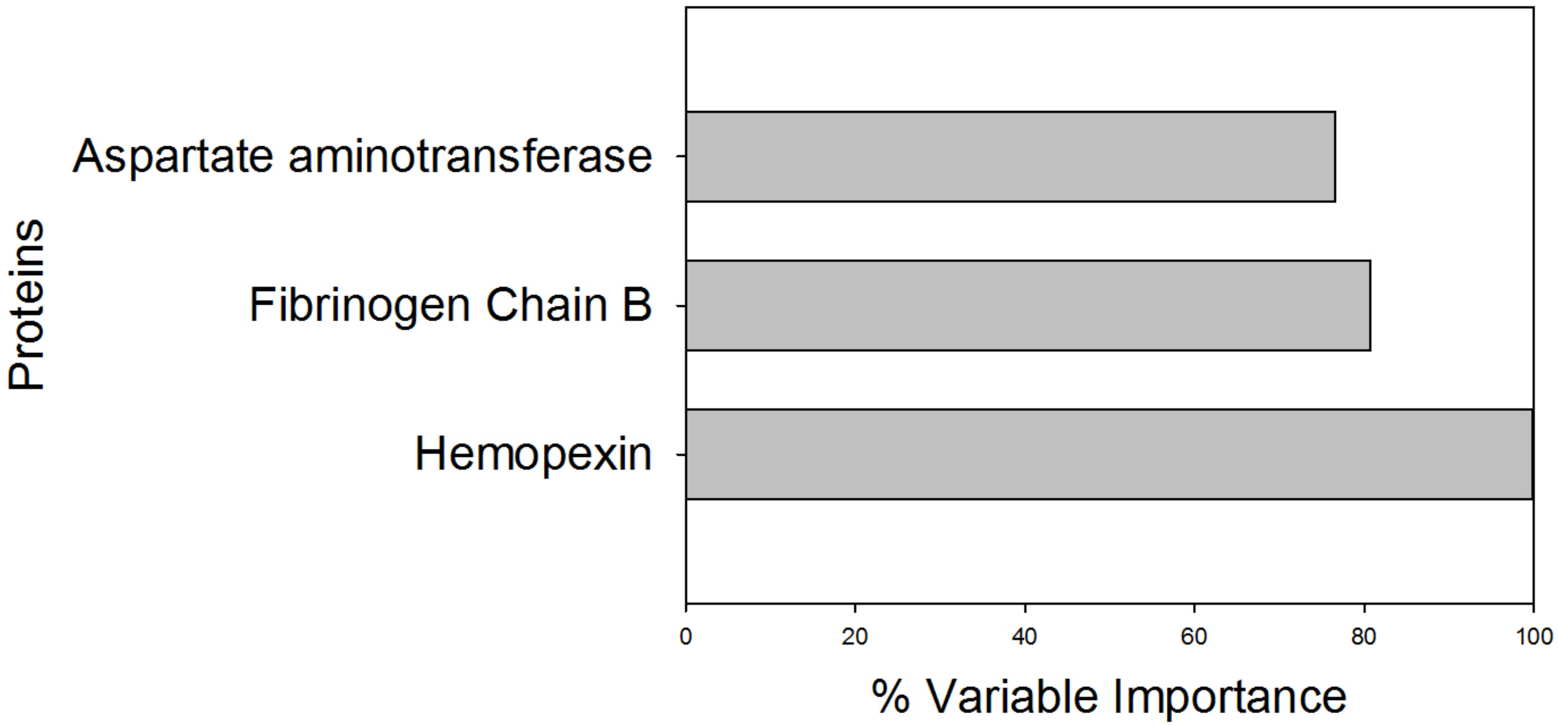
H Pylori Conclusions

- We attempted to verify our model on an independent dataset
- Our initial model included one amino acid and several cytokines
- The Japanese samples performed very differently with respect to citrulline, IL-17, and IL-6
- Maybe the diet of the Japanese is sufficiently different from Northern Americans that the Japanese samples can not be surrogates for the original samples
- An alternate validation set is being sought that will be similar to North Americans

Invasive Aspergillosis

- Invasive Aspergillosis is a secondary infection that appears in people with compromised immune systems
- Researchers are interested in learning if there are any predictors that will indicate if a person has invasive aspergillosis, since frequently the symptoms are masked by the primary infection
- 30 control individuals with a primary infection
- 30 individuals with a primary infection as well as invasive Aspergillosis

MARS Variable Importance



MARS - Prediction Success

Training

Class	Total	Prediction	
		Case (n=28)	Control (n=27)
Case	28	22	5
Control	27	6	20
Total	55	correct =78.57%	correct =77.78%

Testing

Class	Total	Prediction	
		Case (n=27)	Control (n=28)
Case	28	19	9
Control	27	8	19
Total	55	correct =67.86%	correct =70.37%

ROC AUC Training: .90; ROC AUC Testing: .73

Verification Results

- Model was not performing well at all
- Investigated if PI of study had matched properly – No!
- Looked at two groups of underlying disease – those with some form of Leukemia vs those with some other underlying disease
- Leukemia subjects behave differently than non-Leukemia subjects (and were present in differing amounts between the Qualification & Verification samples!)
- Non-Leukemia subjects do not show differential expression for any of the proposed markers (either in qualification or verification studies)
- For Qualification studies, Leukemia subjects show differential expression for the proposed markers of For Qualification samples, when the underlying disease is Leukemia, A1AG1, A2GL, AATC, ALBU, APOA1, APOC3, and HPX are significantly different between cases & controls
- For Verification samples, when the underlying disease is Leukemia, A2GL, AATC, FIBB, FIBA, FRIL, and HPT are significantly different

Take Home Message

- Sample size matters
- Make sure the samples you wish to use for Verification match as closely as possible to the samples for Discovery (both in type & location)
- Confounding factors make a difference – control for those when matching
- You will always get a model – make sure the most useful information is entered and not everything that is collected
- Overfitting is an issue when performing predictive modeling – make sure to deal with that appropriately

Acknowledgements

- Team
  - Allan Brasier
  - John Wiktorowicz
  - Adrian Recinos
  - Yingxin Zhao
  - Hyunsu Ju
  - Wendy Baker

- Supported in part by:
  - Clinical and Translational Science Award UL1TR000071 (Brasier)
  - NIAID Contract – Clinical Proteomics Center HHSN272200800048C (Brasier)
  - NCI K25 1K25CA127524-01A1 (Spratt)