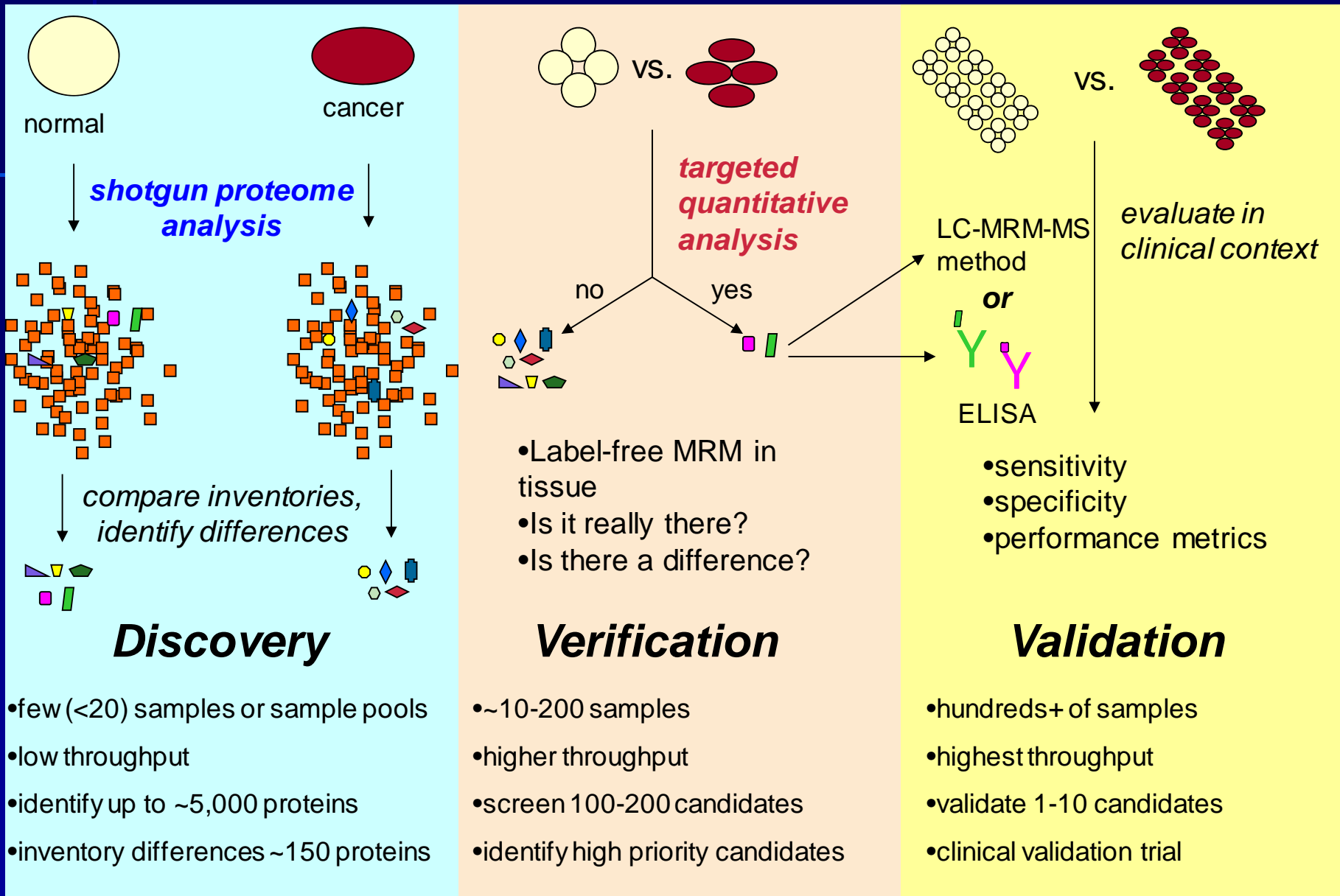


**Statistical Analysis Strategies
for
Shotgun Proteomics Data**

Ming Li, Ph.D.

Cancer Biostatistics Center
Vanderbilt University Medical Center

Ayers Institute Biomarker Pipeline



The Big Picture

Proteomics Tools

- Mass Spectrometry
- Database
- Bioinformatics
- Protein-separation Techniques



Research Goal

Biomarker Discovery

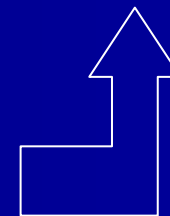


High Throughput Data

(MALDI MS, Shotgun, etc)



Challenge on
Data Analysis



Outline

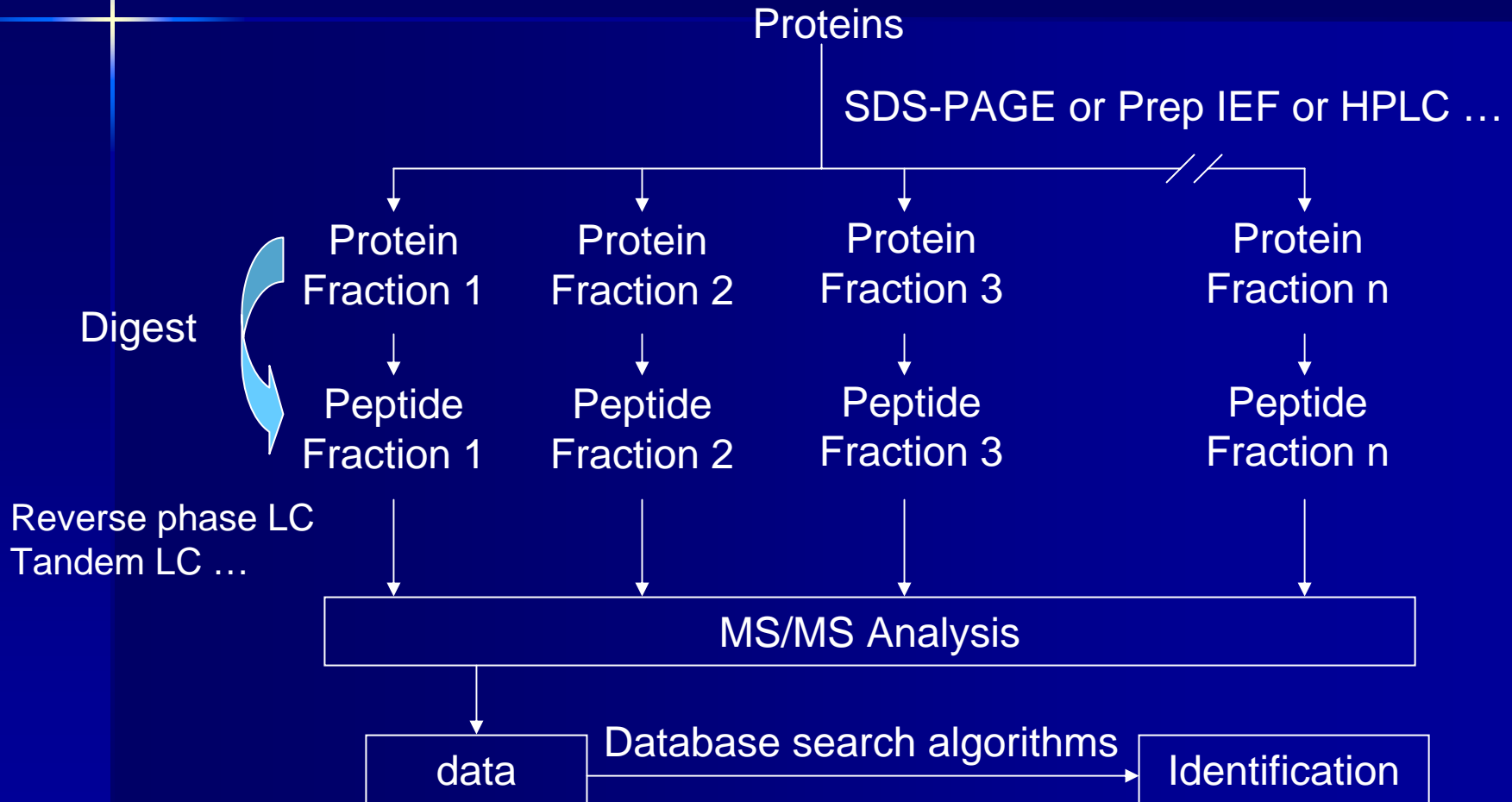
- **Background**
 - Shotgun Proteomics Techniques and Data
 - The Challenges for Statistical Analysis
- **Statistical Analysis Strategies**
 - Methods and Models
 - Case Study and Preliminary Results
- **Discussion and Future Work**

Background on Shotgun Proteomics

■ What is Shotgun Proteomics?

- A method of identifying proteins in complex mixtures using a certain separation/digestion method combined with mass spectrometry.
- More specifically, the proteins in the mixture are digested and the resulting peptides are separated (by HPLC, SCX, IEF, etc), tandem mass spectrometry (MS/MS) is then used to identify the peptides (LC-MS/MS) and matched back to proteins.

General Flow in Proteomic Analysis

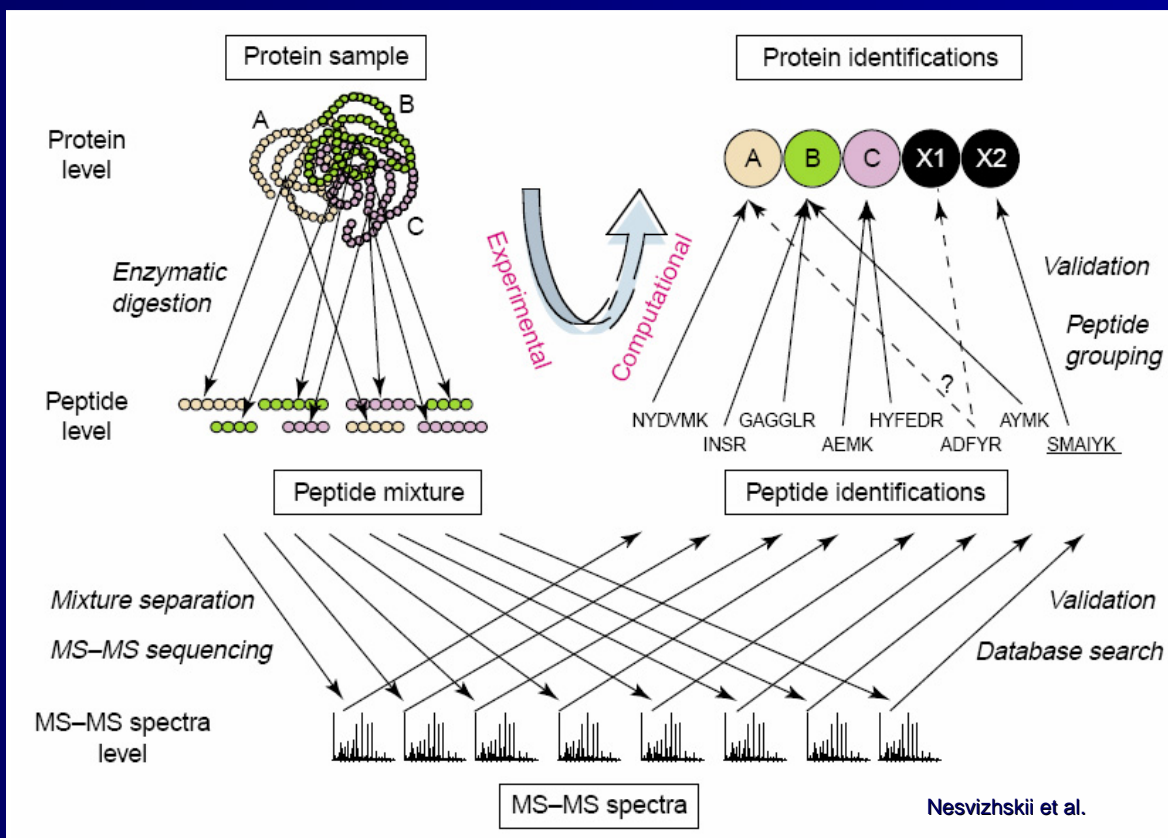


Shotgun 101

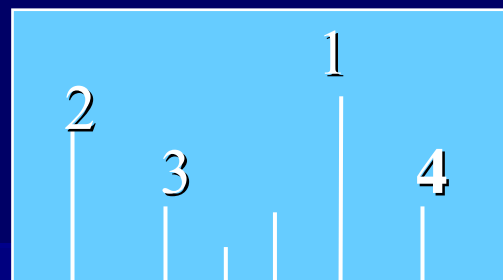
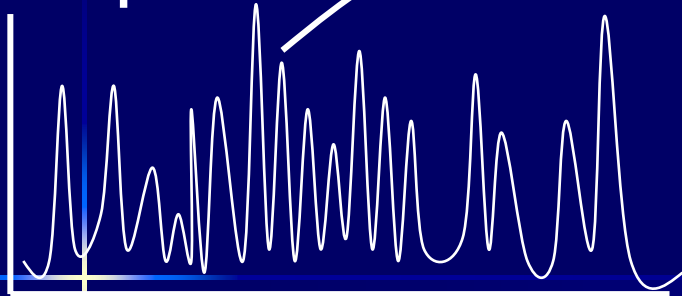
Design

Analysis

Integration

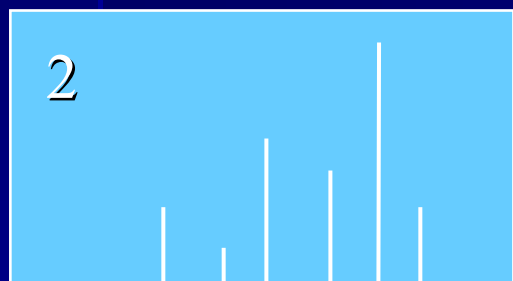


LC separation

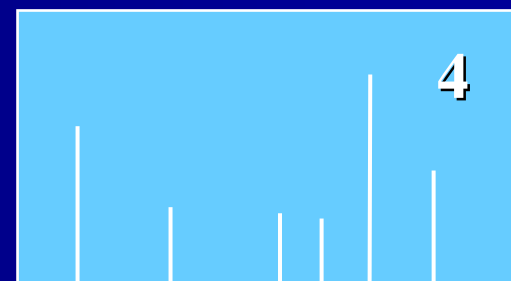


MS

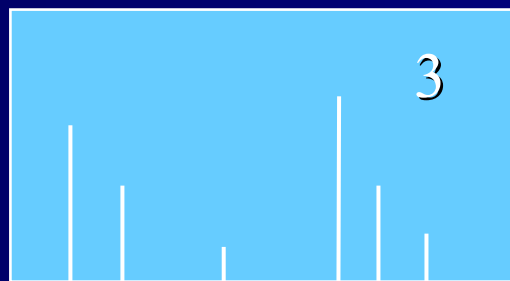
MS-MS
fragmentation



m/z



m/z



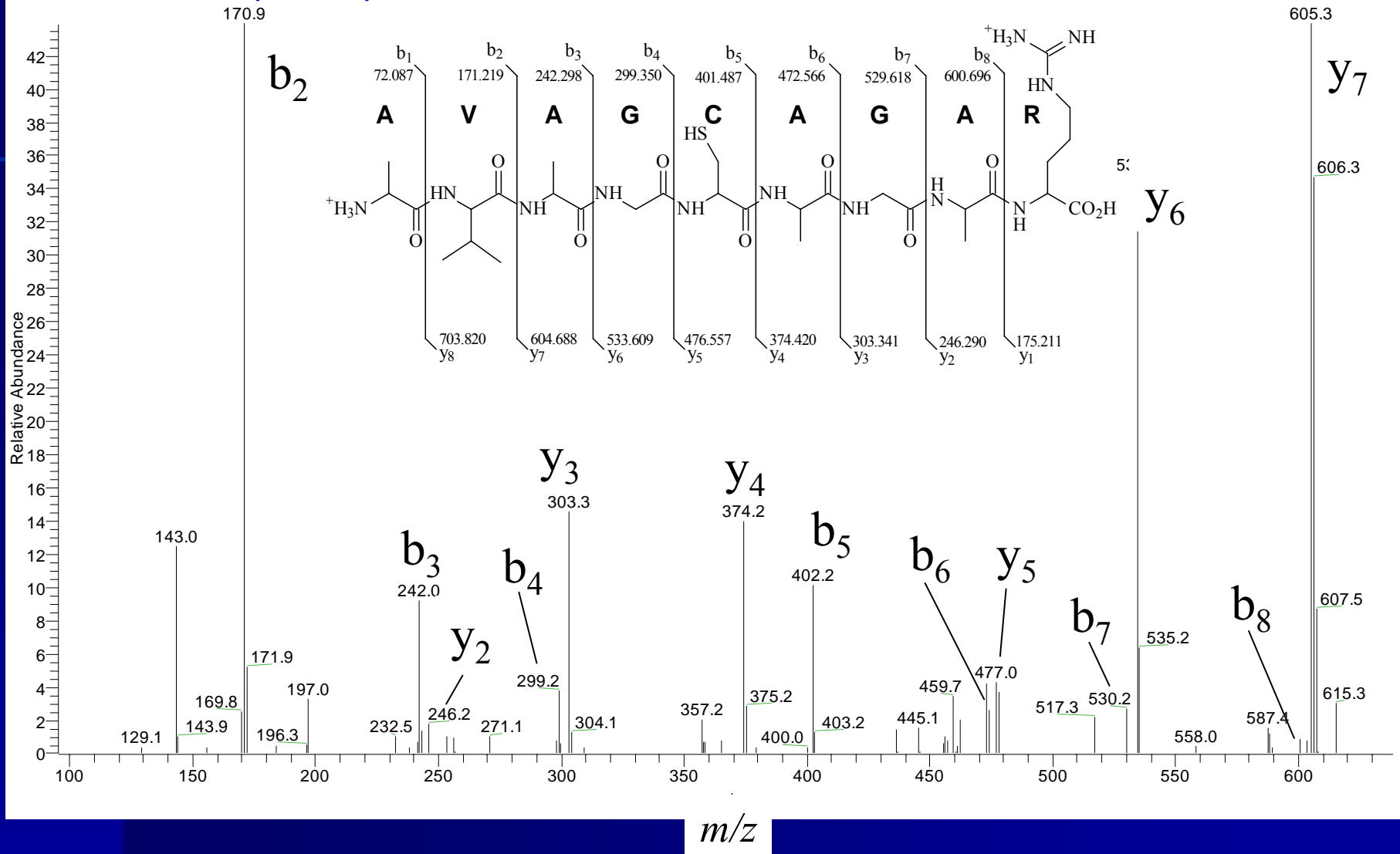
m/z

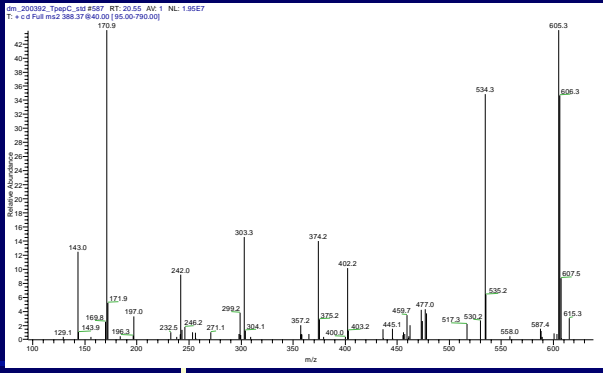


m/z

dm_200392_TpepC_std #587 RT: 20.55 AV: 1 NL: 1.95E7

T: + c d Full ms2 388.37@40.00 [95.00-790.00]





Actual MS-MS scan

Precursor peptide
[M+H]⁺ = 775.8

Get database sequences
that match precursor
peptide mass

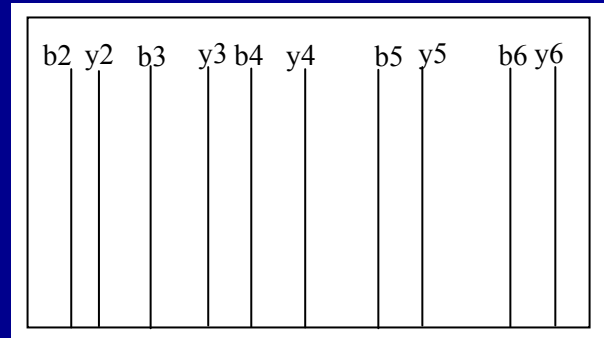
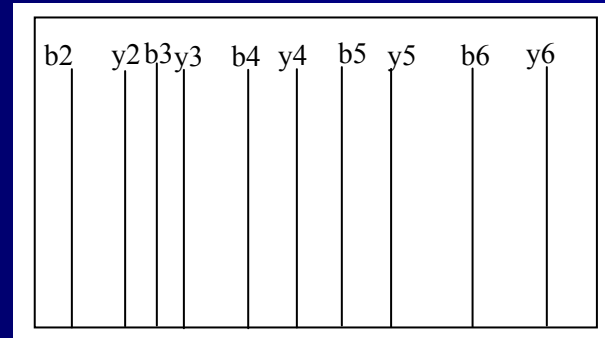
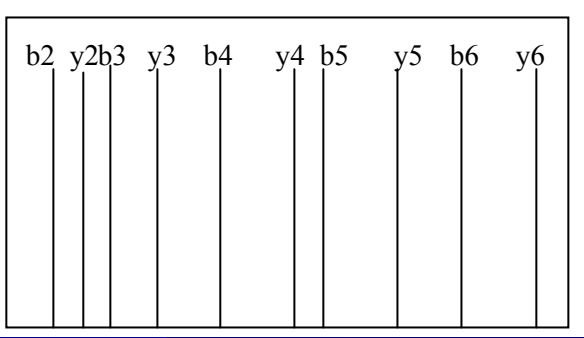
AVAGCAGAR
CVAAGAAGR
VGGACAAAR
etc...

Generate "virtual" MS-MS
spectra

AVAGCAGAR

CVAAGAAGR

VGGACAAAR



Compare virtual spectra
to real spectrum

Scoring

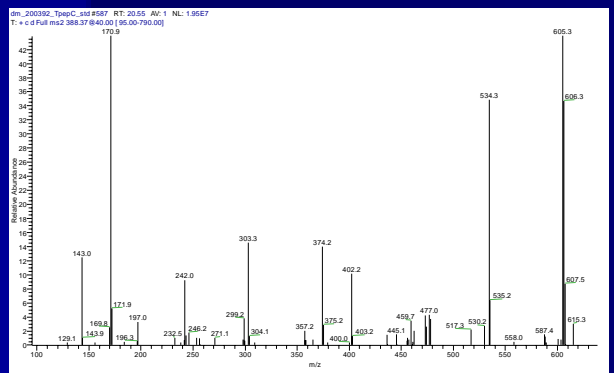
- Detect matches between theoretical b- and y-ions and actual spectrum ions
- Compute correlation scores
- Rank hits

Peptide score

AVAGCAGAR 2.56

CVAAGAAGR 0.57

VGGACAAAR 0.32



From Shotgun Proteomics to Biomarker Discovery

- **A Protein/Peptide Frequency-Based Analysis Approach**
 - Compare spectral identifications between groups;
 - The unit of measurement is the number of times a peptide observed in a single LC-MS/MS round of analysis that are matched to the peptide sequences;
 - The “counts” reflect the abundance of the protein from which these peptides are derived.

Shotgun Data and Statistical Challenge

Normal

Cancer

Protein	Sample 1	Sample m	Sample 1	Sample m
#1	57	65	108	160
#2	20	38	12	9
#3	85	67	8	22
#4	0	0	3	1
:	:	:	:	:	:	:
:	:	:	:	:	:	:
$\#N$	70	23	12	18

Statistical Analysis Goal

- Provide investigators a “winner” list of proteins for further study;
- The “winners” are the potential biomarkers of diagnostics, prognostics or therapeutic;
- The “winners” are selected by appropriate statistical models and procedures.

Statistical Analysis Strategy

- **Model the Count Data**
 - Poisson Regression Model
 - Quasi-likelihood Poisson Model (GEE method)
 - Rate Model (Poisson Model with Offset)
- **Handle Small Sample Size**
 - Provide Appropriate Test Statistics
 - Permutation Test
- **Deal with Multiple Comparison Issues**
 - Frequentist Approach (FDR)
 - Empirical Bayes Approach (LOCFDR)

Poisson Model for Count Data

- Poisson Regression Model (Poisson GLM)

- Y is Poisson random variable with mean u , then

$$P(Y = y) = \frac{e^{-u} u^y}{y!}$$

$$E(Y) = Var(Y) = u$$

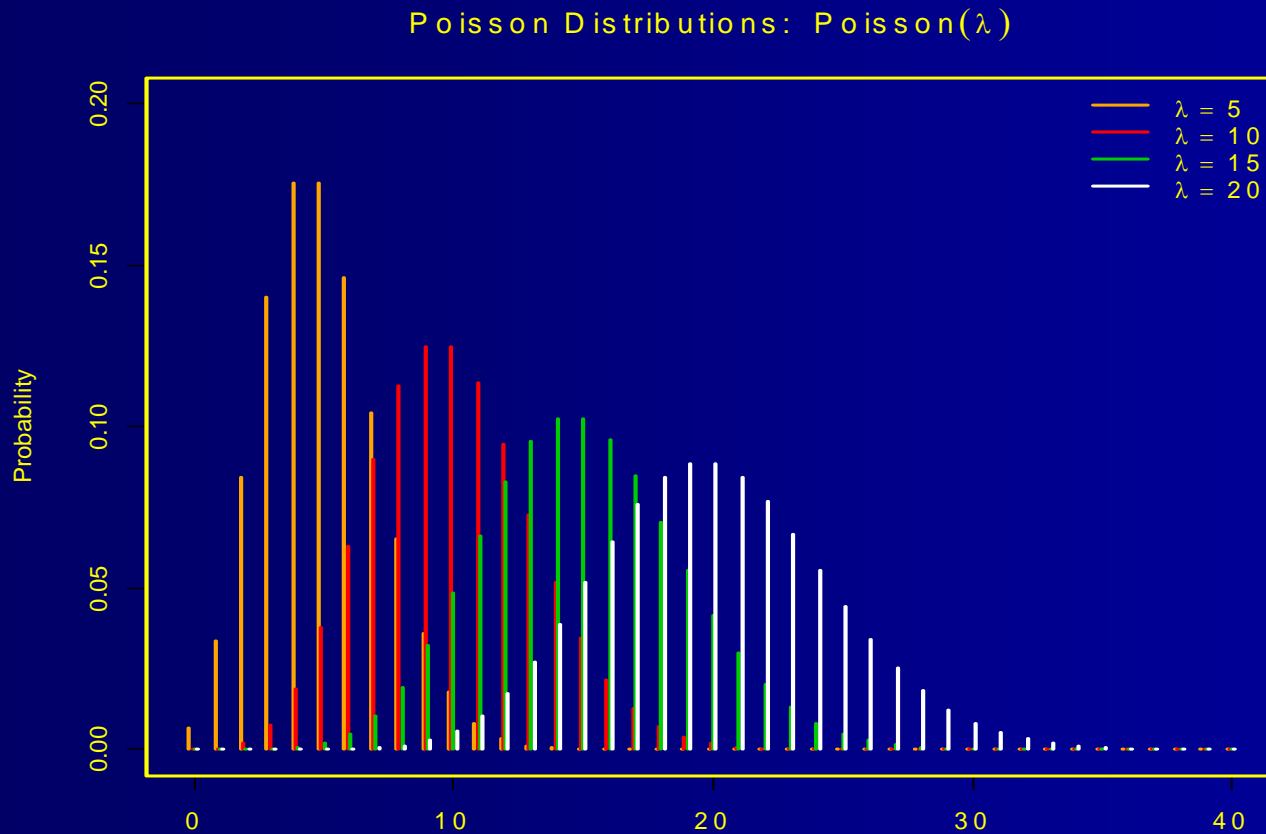
- Log link function: $\log(u) = \eta = \mathbf{x}^T \boldsymbol{\beta}$

(For our application: $\log(u) = \eta = \beta_0 + (\text{Group}) \times \beta_1$)

- By Newton-Raphson algorithm, get MLE of $\boldsymbol{\beta}$ and the 95%CI
- G-statistic and Pearson's χ^2 statistics

Graphical Presentation of Poisson Distribution

- May not be Flexible for Empirical Fitting Purpose.



Note: $\text{Poisson}(\lambda)$ has mean λ and SD $\sqrt{\lambda}$

Extend Poisson Model

■ Generalize Poisson Model by Allowing Dispersion

$$\text{Var}(Y) = \varphi E(Y) = \varphi u$$

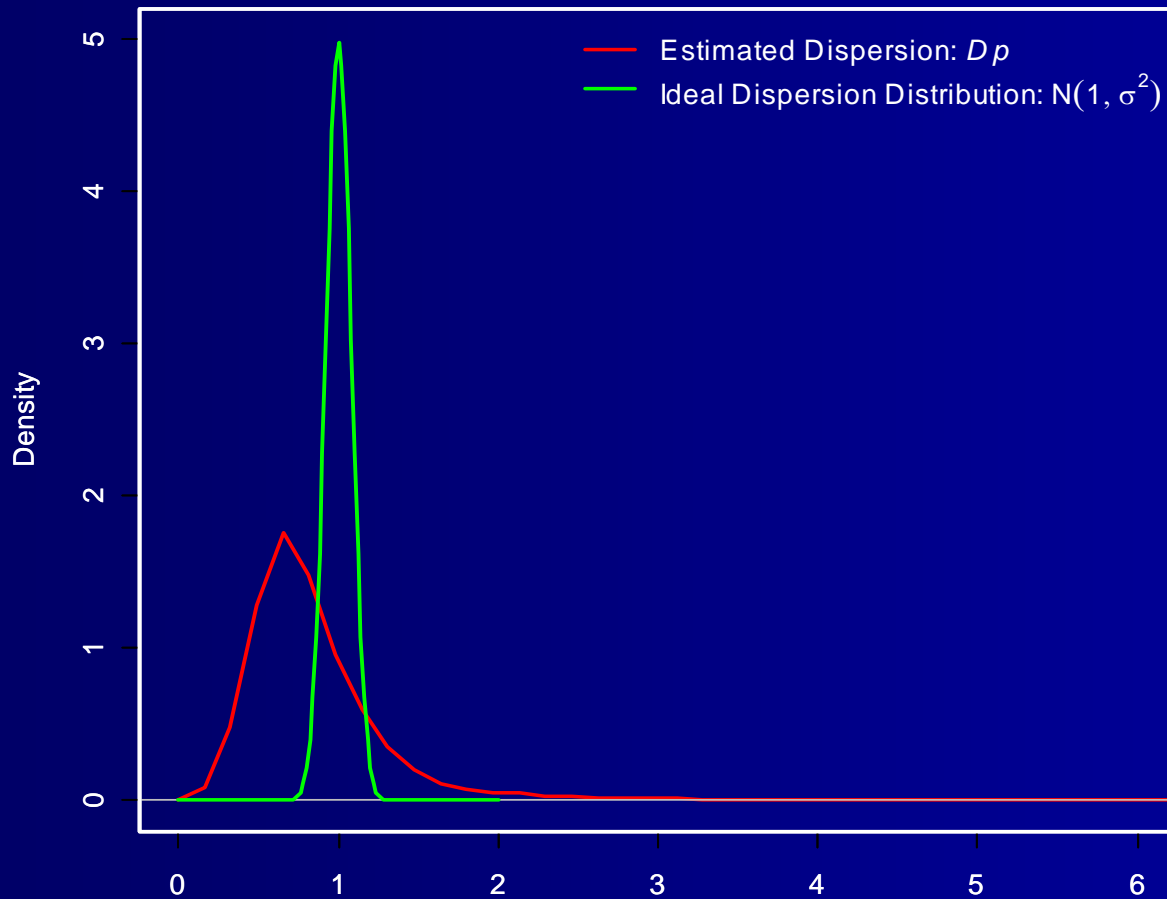
- $\varphi = 1$ (no dispersion), regular Poisson GLM is appropriate
- $\varphi > 1$ (over-dispersion) or $\varphi < 1$ (under-dispersion), the standard error estimation is not reliable, adjust standard errors by φ .

■ Quasi-likelihood Poisson Model

- A more flexible modeling technique when over/under-dispersion occurs in count data
- No strong idea about the appropriate distributional form of the outcome variable while can specify the link and variance function for the model

Under/Over-Dispersion for Shotgun Data

Estimate Dispersion



Quasi-Likelihood Poisson Model

- More details about quasi-likelihood, define a score, U_i :

$$U_i = \frac{Y_i - u_i}{\phi V(u_i)}$$

Then

$$E(U_i) = 0; V(U_i) = \frac{1}{\phi V(u_i)}$$

$$-E \frac{\partial U_i}{\partial u_i} = -E \frac{-\phi V(u_i) - (Y_i - u_i)\phi V'(u_i)}{[\phi V(u_i)]^2} = \frac{1}{\phi V(u_i)}$$

Derive Quasi-likelihood

- These properties are shared by the derivatives of the log-likelihood, l' , which suggest we can use U in place of l' , define:

$$Q_i = \int \frac{y_i - t}{\phi V(t)} dt$$

Then define the log quasi-likelihood for all n observations as:

$$Q = \sum_{i=1}^n Q_i$$

Generalized Estimation Equation (GEE) Method

- Quasi-likelihood approach can be adapted for repeated measures and/or longitudinal experiment designs for shotgun proteomics research
- Generalized Estimation Equation (GEE) methods can be used for estimations
 - Mixed effect models for non-normal responses
 - A multivariate analogue of the equations for the quasi-likelihood models

$$\sum_i \left(\frac{\partial u_i}{\partial \beta} \right)^T \text{Var}(Y_i; \beta, \alpha)^{-1} (Y_i - u_i) = 0$$

Rate Model for Normalizing Shotgun Data

- **Why do We Need Rate Model?**
 - The number of events observed may depend on a size variable that determines the number of opportunities for the events to occur
 - For example: in some run of LC-MS/MS, the sample density might be different; we need “normalization” before comparison
 - Can be done within Poisson/Quasi-Poisson GLM by modeling the effect of the size variable

Rate Model for Normalizing Shotgun Data

- Rate Model

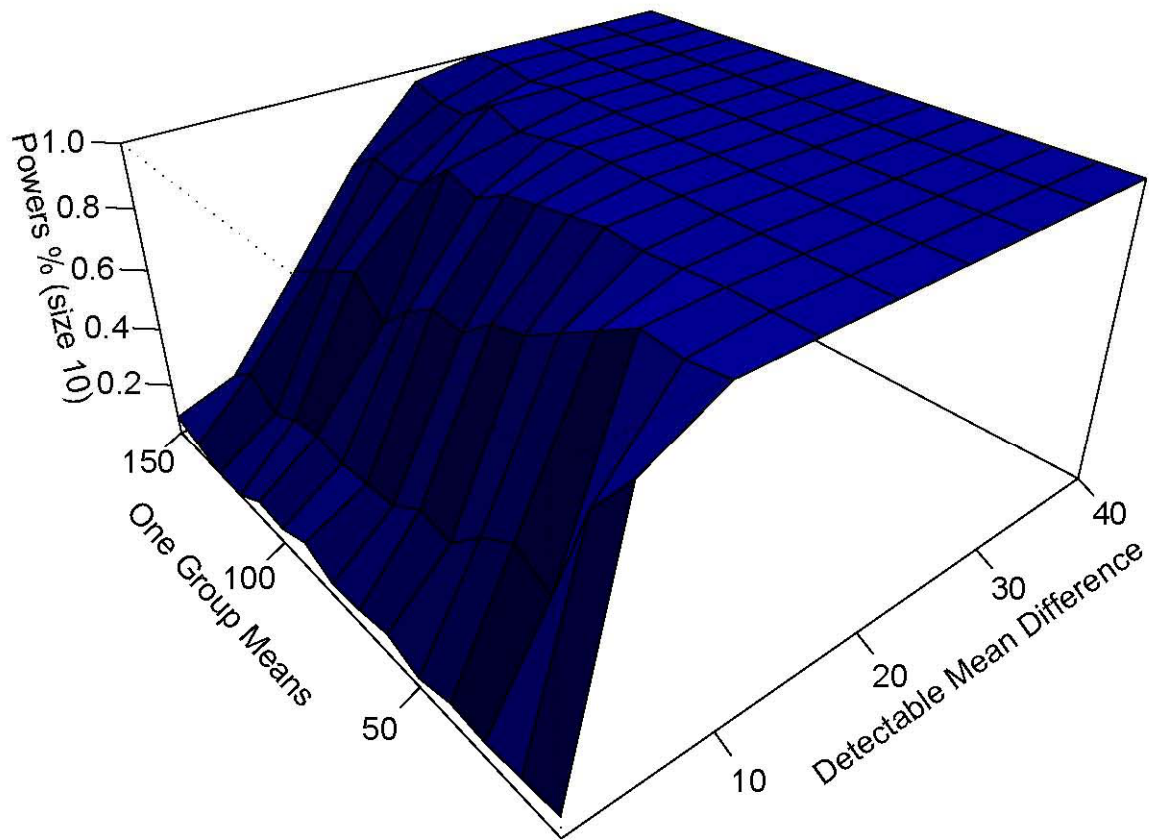
$$\text{Log } (u/w) = \eta = x \beta$$

$$\text{Log } (u) = \log(w) + x \beta$$

- In this manner, we are modeling the rate of spectral observing while still maintaining the count response for the Poisson/Quasi-Poisson model.
- We fix the coefficient as one by using an *offset*.

Handle Small Sample Size

- **Issues with Small Sample Size:**
 - Asymptotic Distribution May not be Hold
 - The χ^2 distribution is only an approximation that becomes more accurate as the sample size increases
 - Not possible to say exactly how large we need, but $N \geq 5$ is often suggested
 - Lack of Power if the Effect is not Strong
- **In the real world, due to financial/time constraints, the number of runs for shotgun data is usually small.**



Handle Small Sample Size

■ Some Solutions

– Appropriate Test Statistics: $D_S - D_L \sim \chi^2_{l-s}$

Model S: $\text{Log}(u) = \log(w) + \beta_0$

Model L: $\log(u) = \log(w) + \beta_0 + (\text{Group}) \times \beta_1$

– Permutation Test:

Reference distribution is obtained from the data

$$P = \{\# T_{perm} > T_{obs}\} / \{\# total\ permutation\}$$

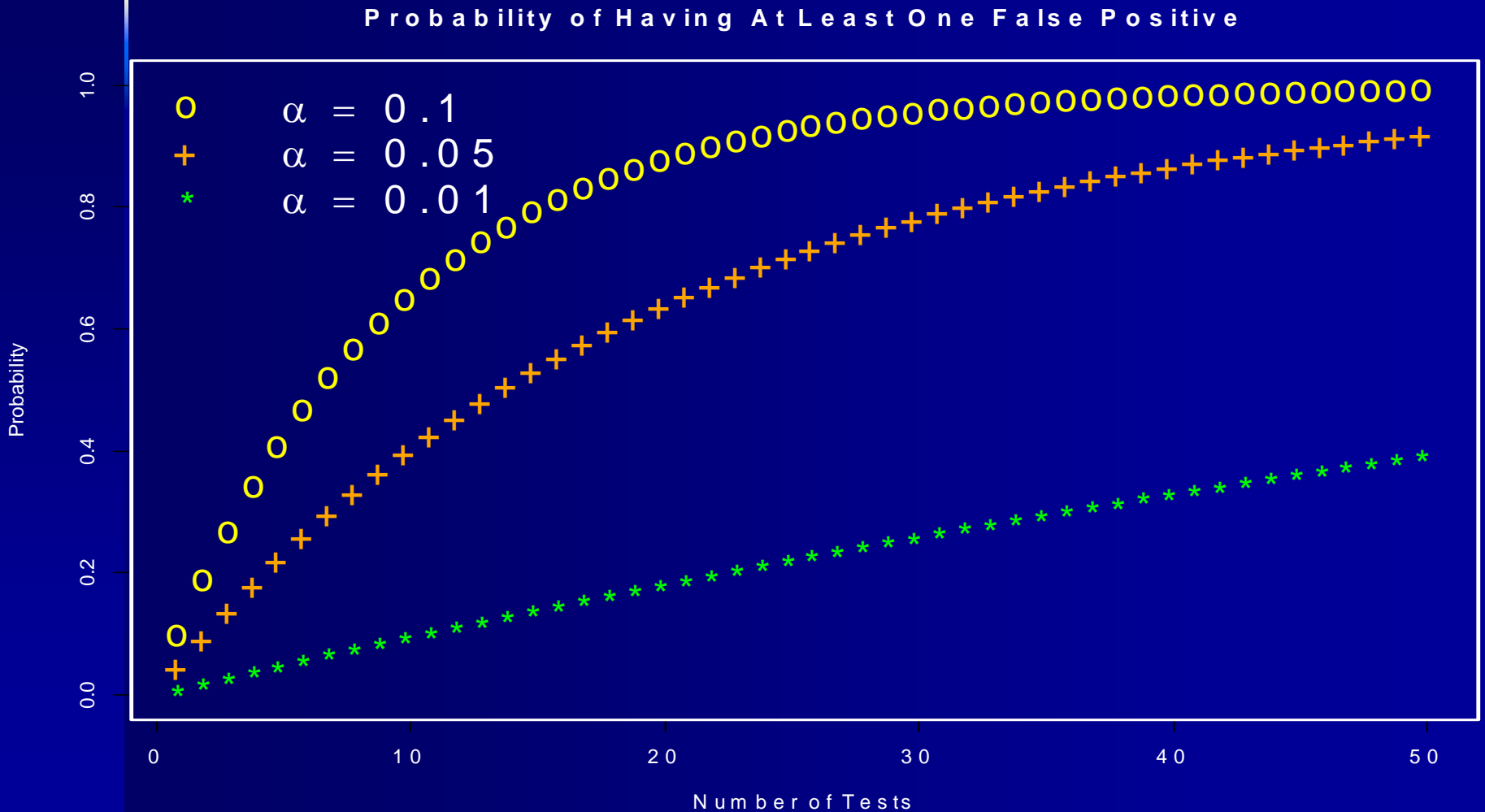
Statistical Analysis Strategy

- **Model the Count Data**
 - ✓ Poisson Regression Model
 - ✓ Quasi-likelihood Poisson Model (GEE method)
 - ✓ Rate Model (Poisson Model with Offset)
- **Handle Small Sample Size**
 - ✓ Provide Appropriate Test Statistics
 - ✓ Permutation Test
- **Deal with Multiple Comparison Issues**
 - Frequentist Approach (FDR)
 - Empirical Bayes Approach (LOCFDR)

Deal with Multiple Comparisons

- **Why Worry About It?**
 - **Selection Bias:** researchers tend to select significant ones to support their conclusions
 - **Inflate the False Positive Rate:** unadjusted P-values from the single-inference procedure result in increased type I error.

Multiple Tests Increase Chance of False Positive



Deal with Multiple Comparisons

- **Family wise error rate (FWER) methods**
 - Too conservative and less application values
 - Not suitable for large-scale simultaneous hypothesis testing problems arouse from “high through” technologies
- A desirable error rate to control might be the expected proportion of errors among the rejected hypotheses.

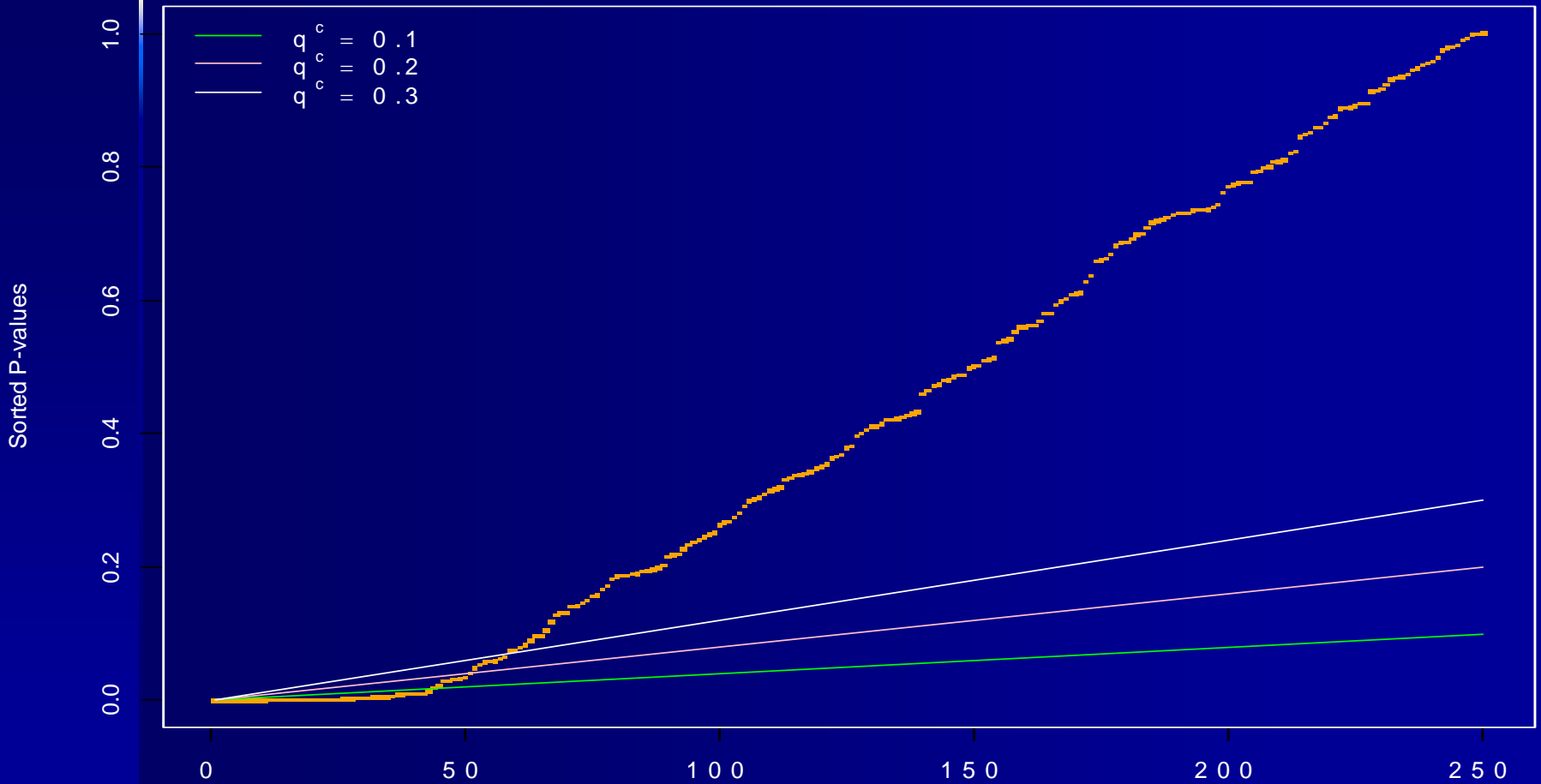
Approach 1: False Discovery Rate

- Define the FDR to be the expectation of Q , and control it under a value q^*

$$FDR = E(Q) = E\left(\frac{V}{V+S}\right) = E\left(\frac{V}{R}\right)$$

	Declared Non-significant	Declared Significant	Total
True Null hypothesis	U	V	m_0
Non-true Null hypothesis	T	S	$m - m_0$
	$m - R$	R	m

False Discovery Rate Controlling Procedure



Approach 2: Local False Discovery Rate

■ Define Local FDR

– Null hypothesis: H_1, H_2, \dots, H_N

– Test statistic: z_1, z_2, \dots, z_N

$p_0 = \Pr\{\text{Null}\}$ $f_0(z)$ density if null

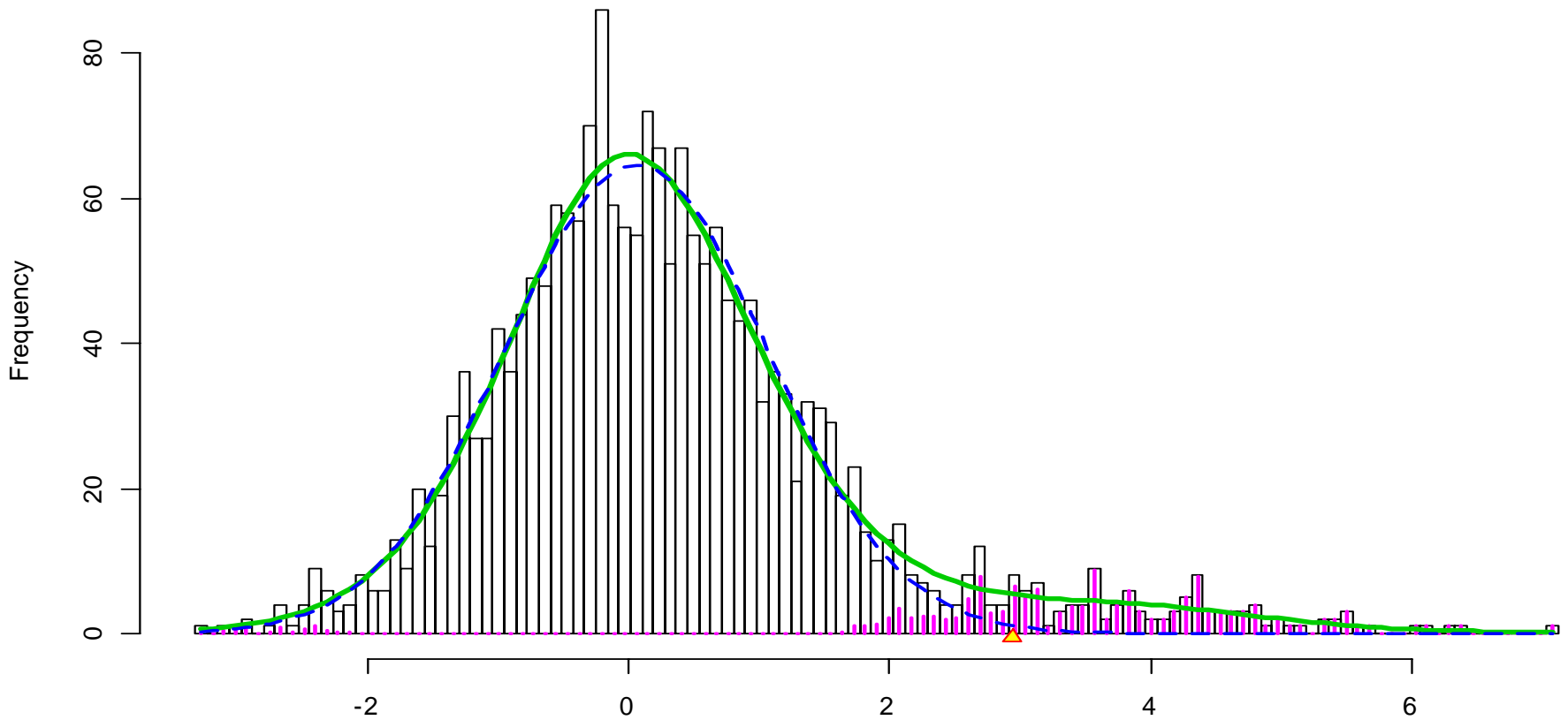
$p_1 = \Pr\{\text{non-null}\}$ $f_1(z)$ density if non-null

– Mixture density: $f(z) = p_0 f_0(z) + p_1 f_1(z)$

– Local FDR: $\text{fdr}(z) = \Pr\{\text{null}|z\} = p_0 f_0(z) / f(z)$

How to Apply Local False Discovery Rate

Histogram of Summary Statistics with Fitted Mixture Density



MLE: delta: 0.059 sigma: 1.006 p0: 0.928
CME: delta: 0.011 sigma: 0.966 p0: 0.908

FDR vs. LOCFDR

FDR	LOCFDR
Frequentist Approach	Empirical Bayes method
Works on P-values (null hypothesis tail area)	Works on the test statistics

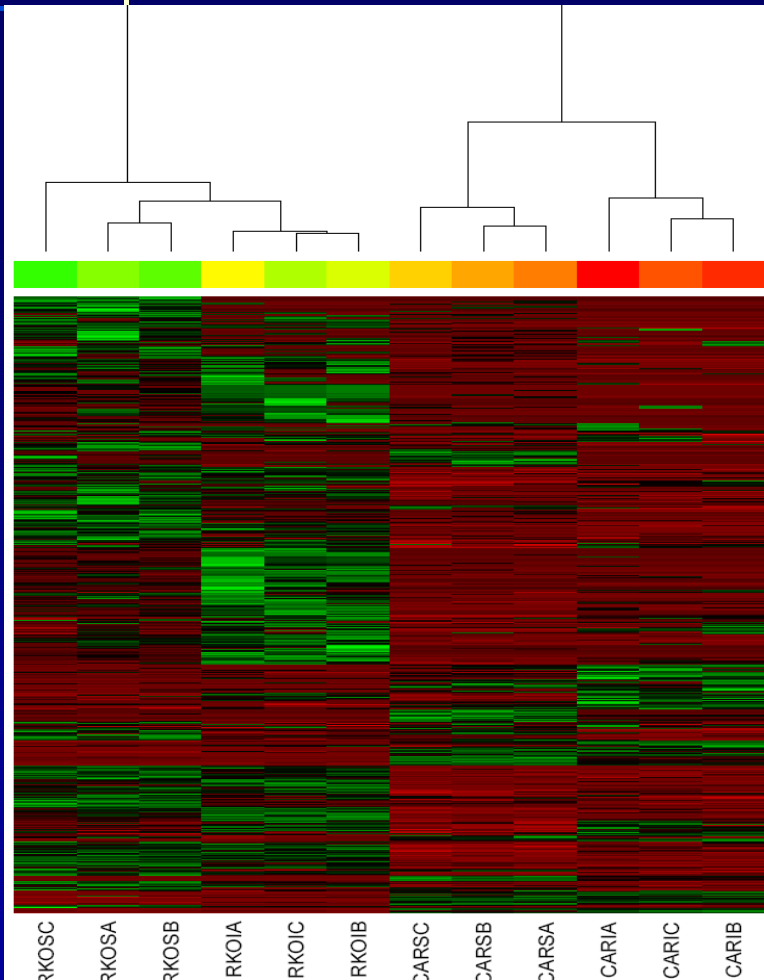
- Efron: “In practice, FDR and LOCFDR can be combined, using Benjamini-Hochberg algorithm to identify non-null cases, say with $q=0.1$, but also providing individual LOCFDR values for those cases.”

Statistical Analysis Strategy

- **Model the Count Data**
 - √ Poisson Regression Model
 - √ Quasi-likelihood Poisson Model
 - √ Rate Model (Poisson Model with Offset)
- **Handle Small Sample Size**
 - √ Provide Appropriate Test Statistics
 - √ Permutation Test
- **Deal with Multiple Comparison Issues**
 - √ Frequentist Approach (FDR)
 - √ Empirical Bayes Approach (LOCFDR)

Case Study

A Global Shotgun Proteomic Analysis of Colorectal Carcinoma



■ Biological Materials

- RKO Cell Lines
- Rectal Adenocarcinoma Specimen

■ Mass Spectrometry Method

- Peptides Separated by Strong Cation Exchange (SCX) and Isoelectric Focusing (IEF)
- LTQ-Qorbitrap Mass Spectrometer (Thermo Electron, San Joase, CA)

■ Database Searching and Filtering

- MyriMatch Search Algorithm; IPI Human Database 3.31; IDPicker 2.0

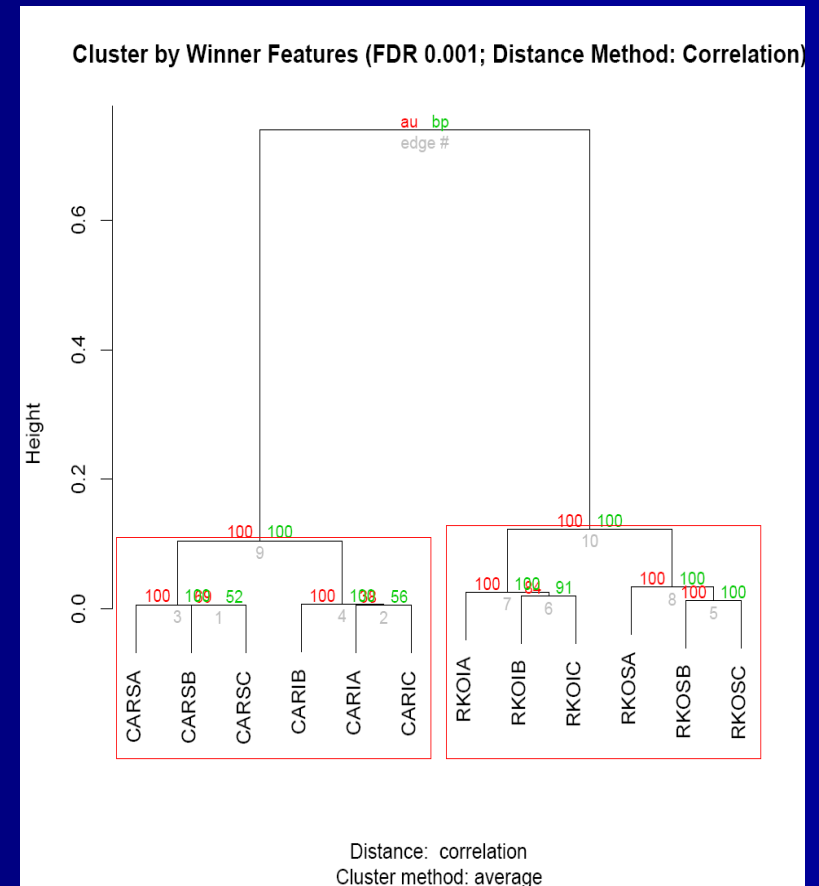
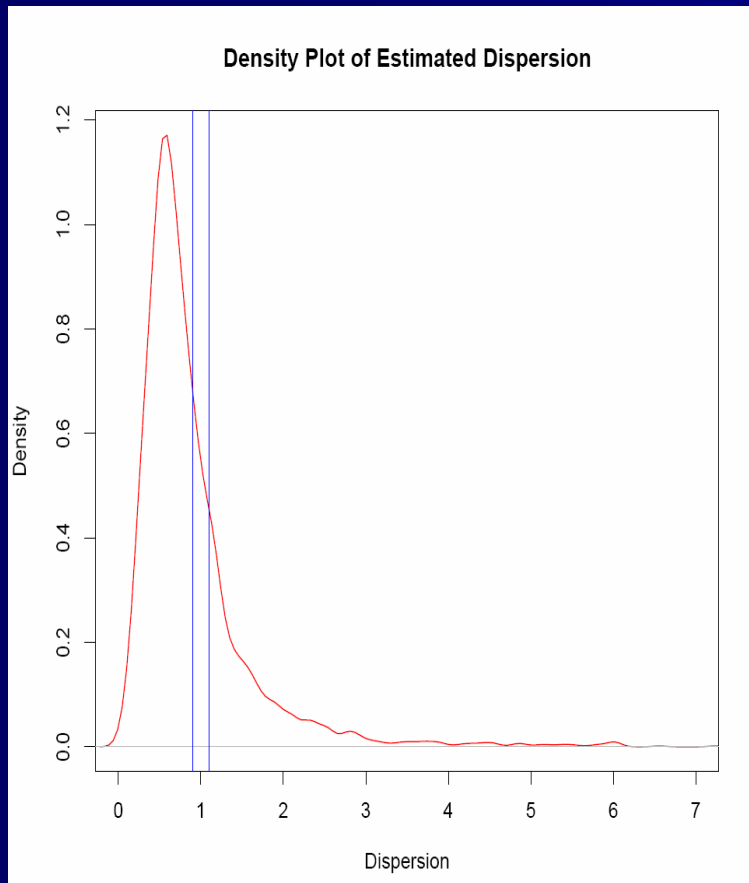
Case Study : Data Analysis

Preliminary Results

Protein.order	Protein	CAR_Counts	RKO_Counts	Poisson_Pvalue	Quasi_Pvalue	Pois_Pvalue_FD R	Quasi_Pvalue_FDR
1513	IPI00176193	342	2	5.72E-119	5.13E-12	8.73E-117	1.41E-08
768	IPI00020501	2267	37	0	1.84E-11	0	2.53E-08
2601	IPI00744256	2276	37	0	2.38E-11	0	2.18E-08
2648	IPI00784458	671	1	1.68E-237	3.50E-11	5.13E-235	2.40E-08
2398	IPI00465084	1326	1	0	9.68E-11	0	5.32E-08
2579	IPI00654755	759	2	9.84E-272	9.69E-11	3.38E-269	4.44E-08
280	IPI00007765	43	533	2.34E-84	1.21E-10	1.69E-82	4.76E-08
1325	IPI00072917	842	1	2.38E-297	2.61E-10	1.09E-294	8.96E-08
842	IPI00022200	847	1	3.23E-299	2.69E-10	1.78E-296	8.20E-08
2029	IPI00302592	1432	428	3.08E-183	2.73E-10	7.69E-181	7.49E-08
2438	IPI00473011	266	2	5.25E-91	3.62E-10	4.51E-89	9.05E-08
2319	IPI00418471	778	1	9.31E-279	5.95E-10	3.65E-276	1.36E-07
2509	IPI00554648	440	1	9.45E-158	7.26E-10	1.85E-155	1.53E-07
1597	IPI00216135	506	105	4.89E-87	9.86E-10	3.73E-85	1.93E-07
1983	IPI00299301	283	1	6.25E-101	1.26E-09	6.61E-99	2.31E-07
21	IPI00000230	574	101	3.69E-108	1.95E-09	4.41E-106	3.35E-07
2377	IPI00455050	547	101	1.93E-100	2.06E-09	1.97E-98	3.33E-07
871	IPI00022463	374	1	2.65E-132	2.51E-09	4.55E-130	3.84E-07
1757	IPI00220709	704	139	7.98E-124	2.56E-09	1.29E-121	3.69E-07

Case Study

Data Analysis Preliminary Results



Discussion and Future Study

- A “framework” Flexible to be Extend
 - Repeat measurement; trend test, and etc.
- Quality Control of the Data
 - Involve in the early step of data generation process
- Simulation Study Evaluating the Methods
- Add the Current Work to the “pipeline” of the Proteomic Research

Acknowledgement

- Rob Slebos
- Dan Liebler
- Pierre Massion
- Takefumi Kikuchi
- David Carbone
- Will Gray
- Yu Shyr
- Cancer Biostatistics Center
- Ayer Institute
- GI SPORE
- Lung SPORE

Thank You !