# Site-of-Origin Prediction of Neuroendocrine Tumors

**Adriana Granados**
California State Polytechnic University, Pomona

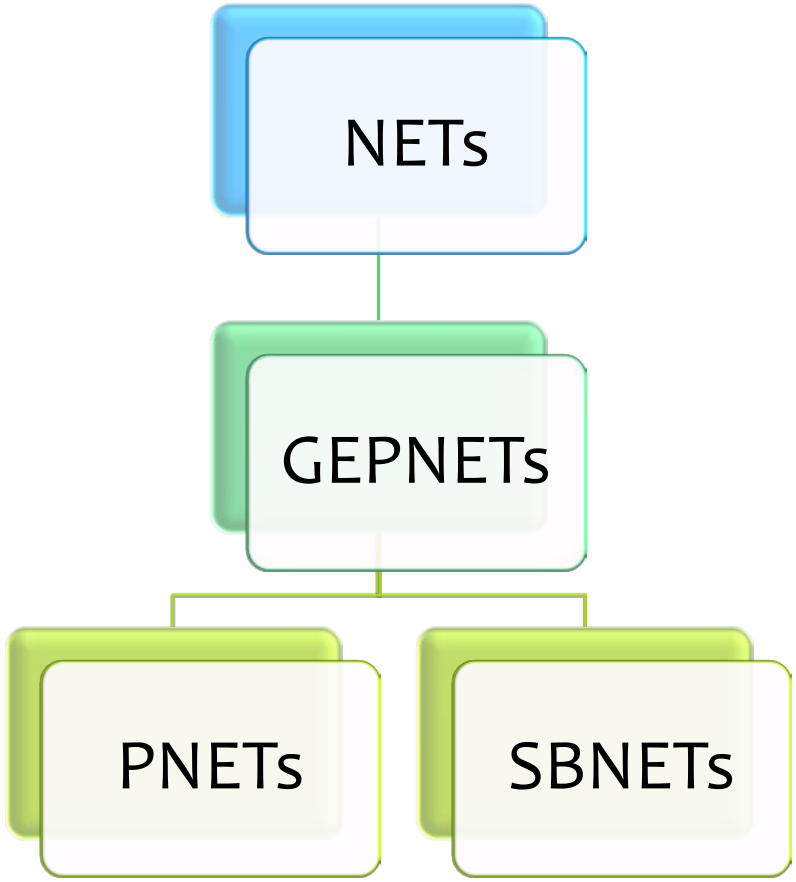
**Saniya Khullar**
Georgetown University

**Thomas Nemmers**
Creighton University

**Patrick Breheny, PhD**
Assistant Professor,
University of Iowa

# Biological Significance

* Neuroendocrine tumors (NETs) are rare tumors that develop in neuroendocrine cells.
* Gastroenteropancreatic neuroendocrine tumors (GEPNETs) originate in the gut or pancreas.
* Five-fold increase of incidence of GEPNETS in last 30 years

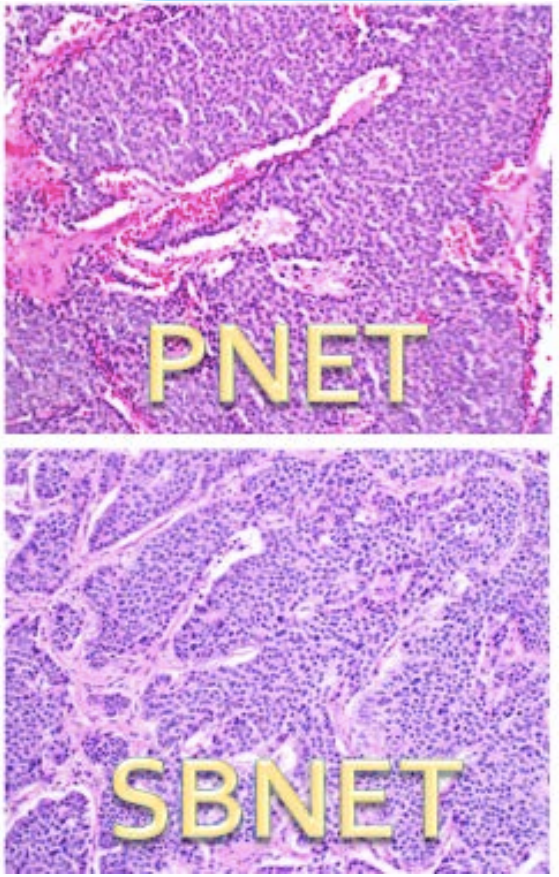


# Biological Significance Continued

* Critical need to systematically examine differences between PNETs and SBNETs to:
  * Improve diagnosis
  * Develop innovative treatment strategies
  * Prolong survival
* PNETs and SBNETs are two most common sites of malignant GEPNETs
* Once tumors have spread to other tissues, primary site becomes harder to distinguish

# How Do We Differentiate These?

* Current methods insufficient to identify primary site
* Methods to identify primary site:
  * Immunohistochemistry, or IHC
  * Gene Expression Classifier, or GEC



# Data Sets Collected

* Microarray Data: collection of gene expression for thousands of genes simultaneously

| Type of GEPNETs | Patients | Genes |
|---|---|---|
| Pancreatic | 5 | 22,011 |
| Small Bowel | 11 | |

* qPCR Data: collection of genes one at a time

| | Genes | Complete Cases | Incomplete Cases |
|---|---|---|---|
| Primary | 15 | 71 | 94 |
| Metastatic | 15 | 89 | 122 |

# Questions of Interest

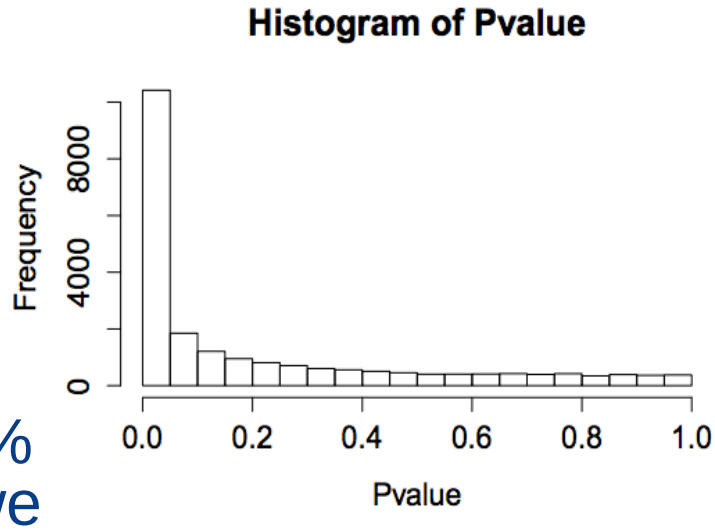
1. If we base a GEC from previously mentioned data sets and use these to predict site-of-origin for metastasized tumors, how accurate are the predictions?

1. Are there additional genes that may be of interest in a further study to help in diagnosing patients?

# Two Independent Sample t-Test

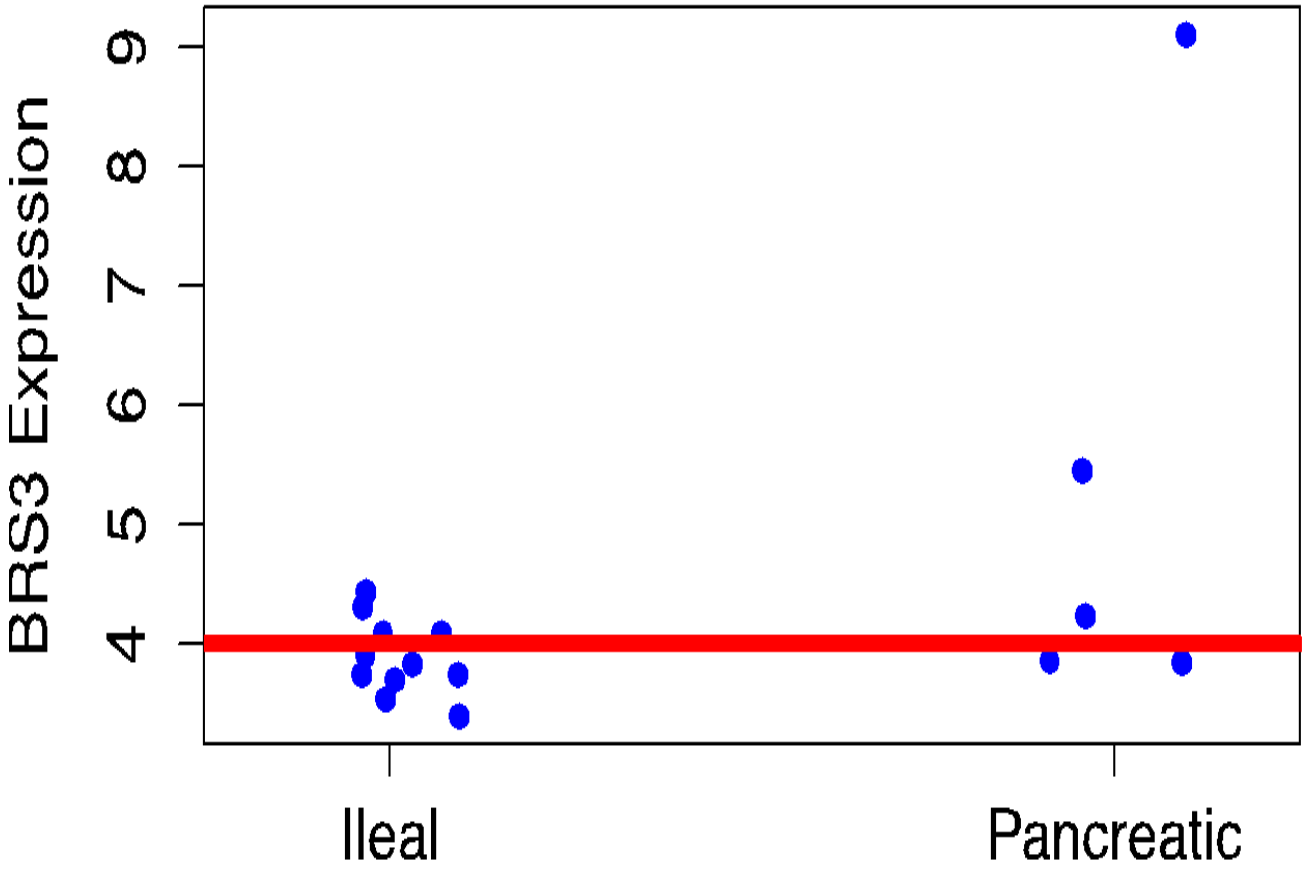
* The distribution of p-values from the tests is very right skewed.
  * A lot of genes are significant
* False Discovery Rate (FDR) (h = 22,011)

$$False\ Discovery\ Rate = \frac{\#\ of\ p-values < \alpha\ by\ random\ chance}{Total\ \#\ of\ p-values < \alpha} = \frac{h\alpha}{S}$$




* Looking at genes with a p-value < α:
  * False Discovery Rate (FDR) = % of these interesting genes that we expect to be misrepresented
  * Q-value is more specific and applies to each gene.

# BRS3 Gene Expression Threshold: 4

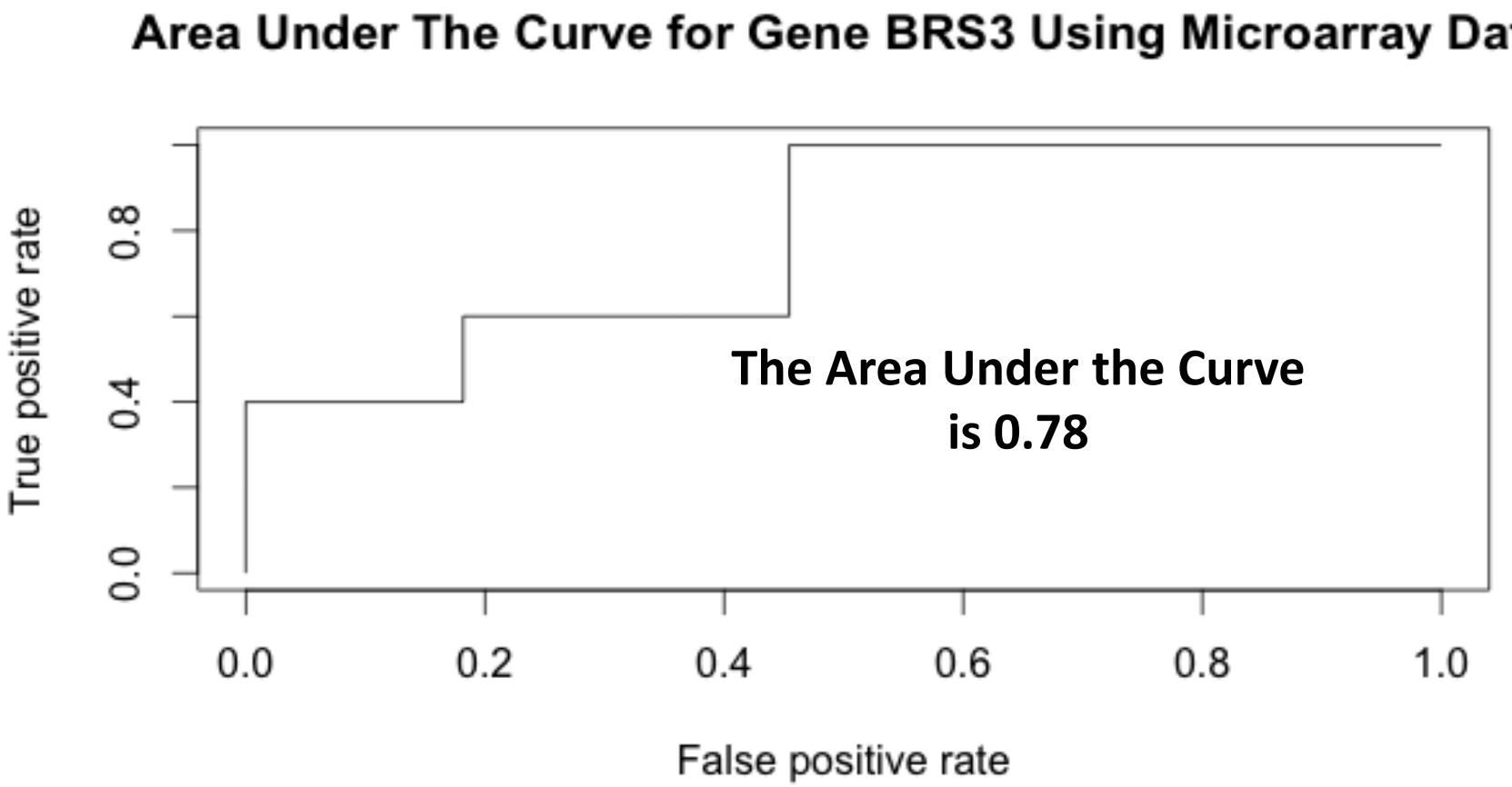



**BRS3 Gene Expression > 4 = Predicted as a PNET**

**BRS3 Gene Expression < 4 = Predicted as a SBNET**

# How Predictive are Certain Genes?

## Analyzing the Receiver Operator Characteristic Curve (ROC Curve)



# Logistic Regression

$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i + \ldots$, where

p = P(of one type of tumor),

$\beta_0$ is an "intercept",

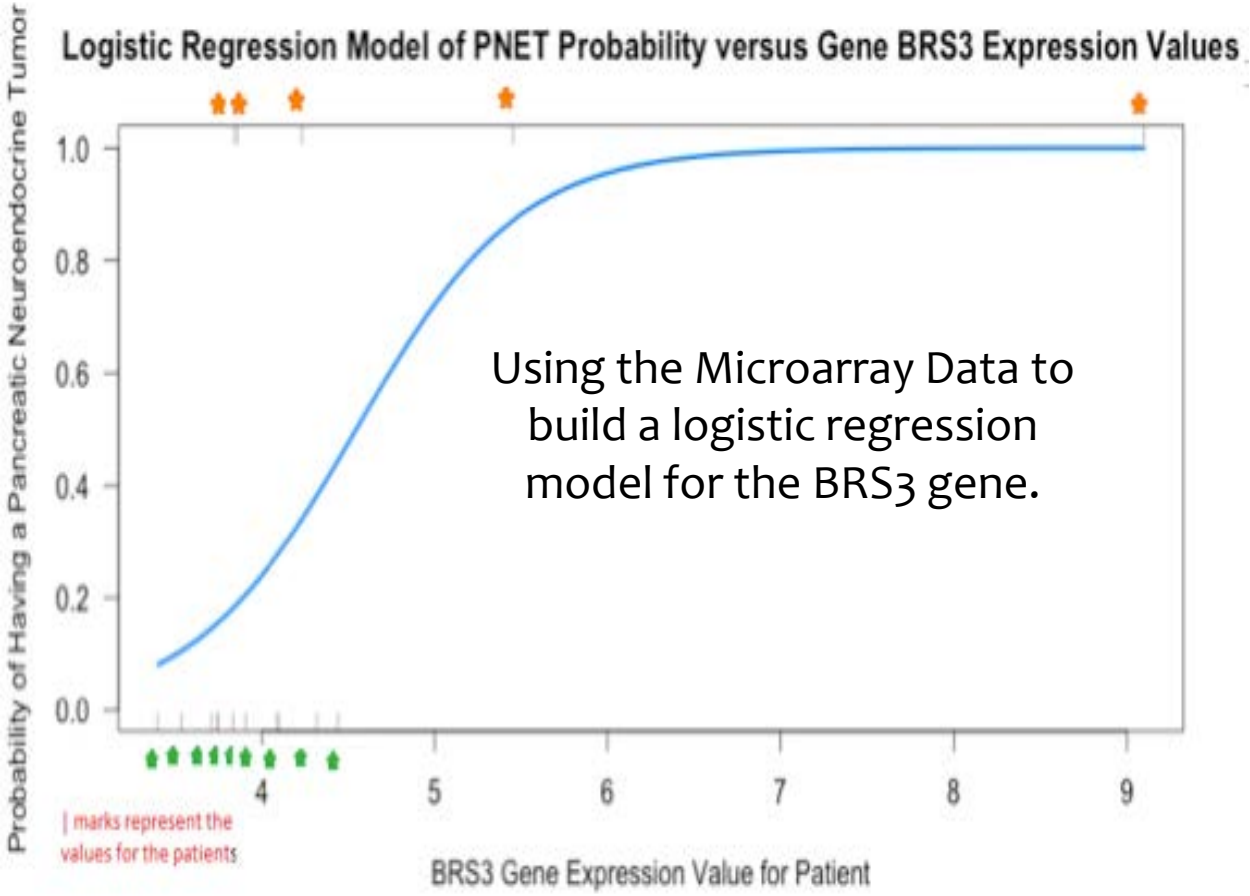
$\beta_i$ is the coefficient of $x_i$,

$x_i$ is the gene expression level for the ith gene, and

i$\geq$1.

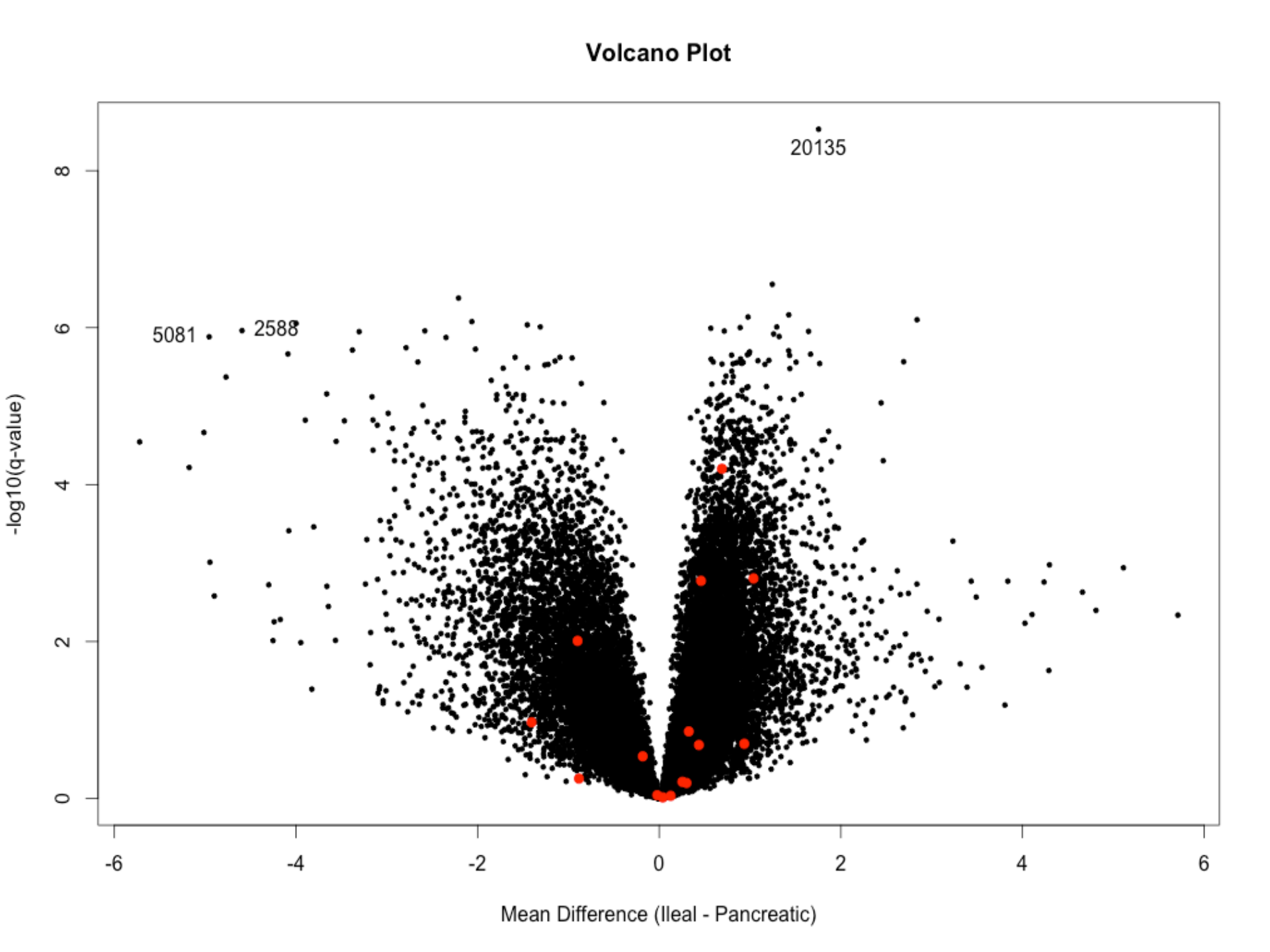
# BRS3 Example: Logistic Model

- Goal: add variables using forward selectio that reduce the AIC until cannot lower value
  - AIC is a maximum likelihood estimation
  - Add1 function is used to calculate the AIC.

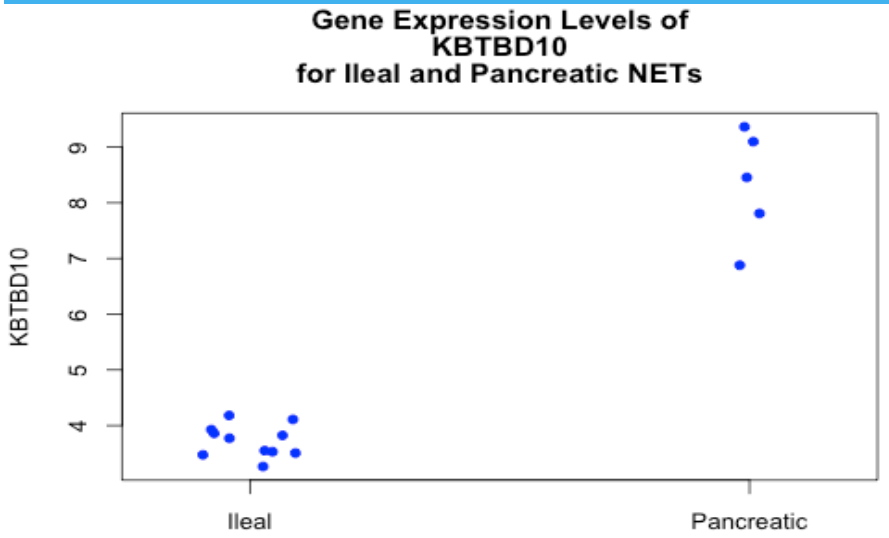


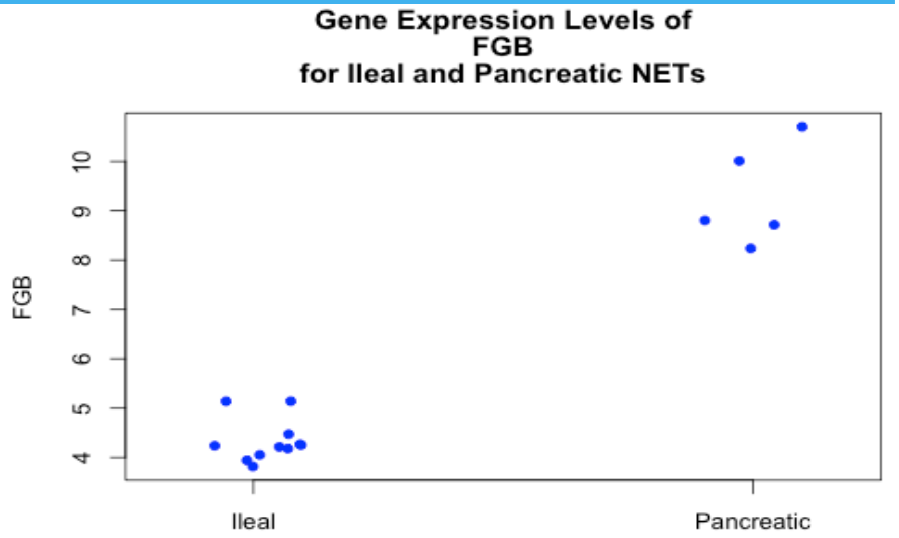
# Results

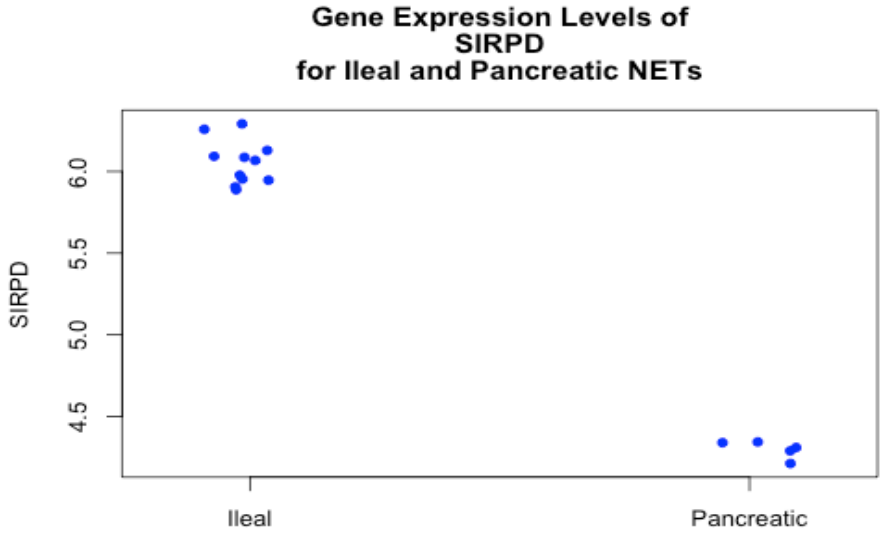
* Number of Q-values that are significant at an FDR = 0.01 is: 4,635
* One way of picking out from those 4,635 genes: volcano plot


Volcano Plot
20135
5081
2588
-log10(q-value)
0
2
4
6
8
-6
-4
-2
0
2
4
6
Mean Difference (Ileal - Pancreatic)

# Strip plots for 3 genes:



# GECs for qPCR Data

* Primary qPCR GEC:

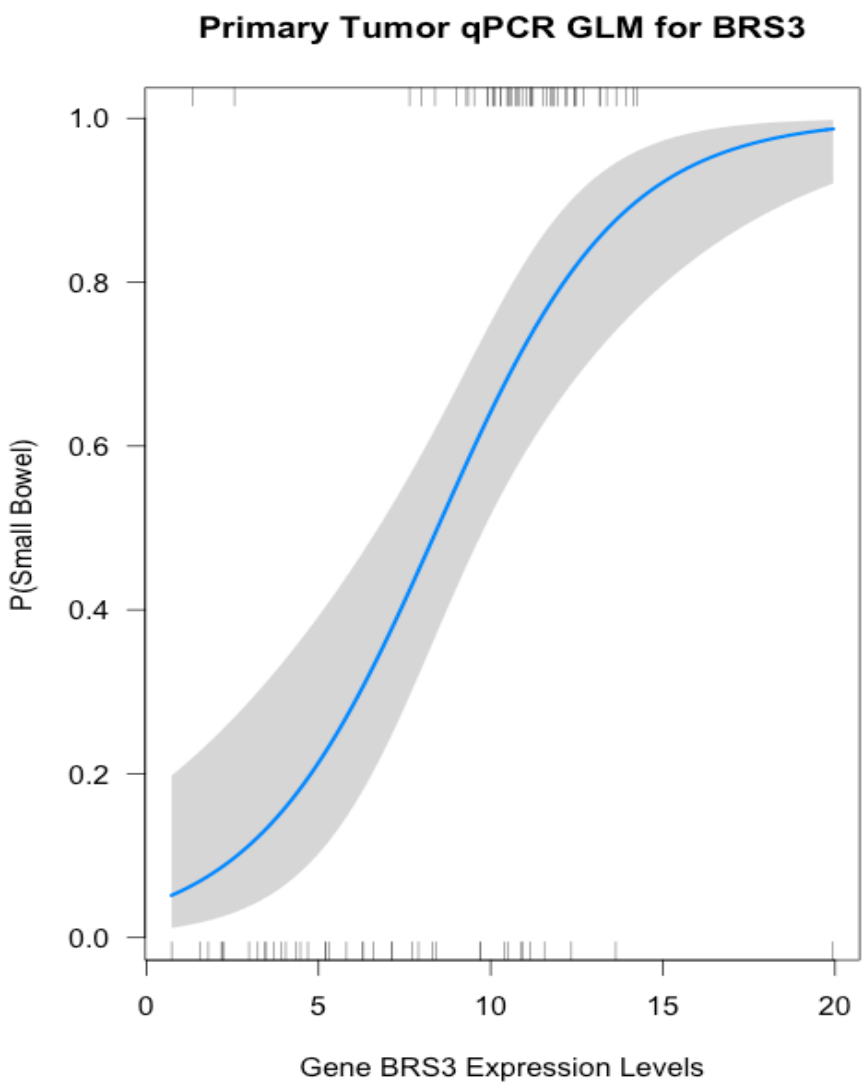
| | row.names | PNET accuracy | SBNET accuracy | Overall accuracy |
|---|---|---|---|---|
| 1 | BRS3 | 0.829 | 0.897 | 0.876 |
| 2 | OPRK1, GPR98, DRD1, SCTR, OXTR | 0.806 | 0.940 | 0.898 |

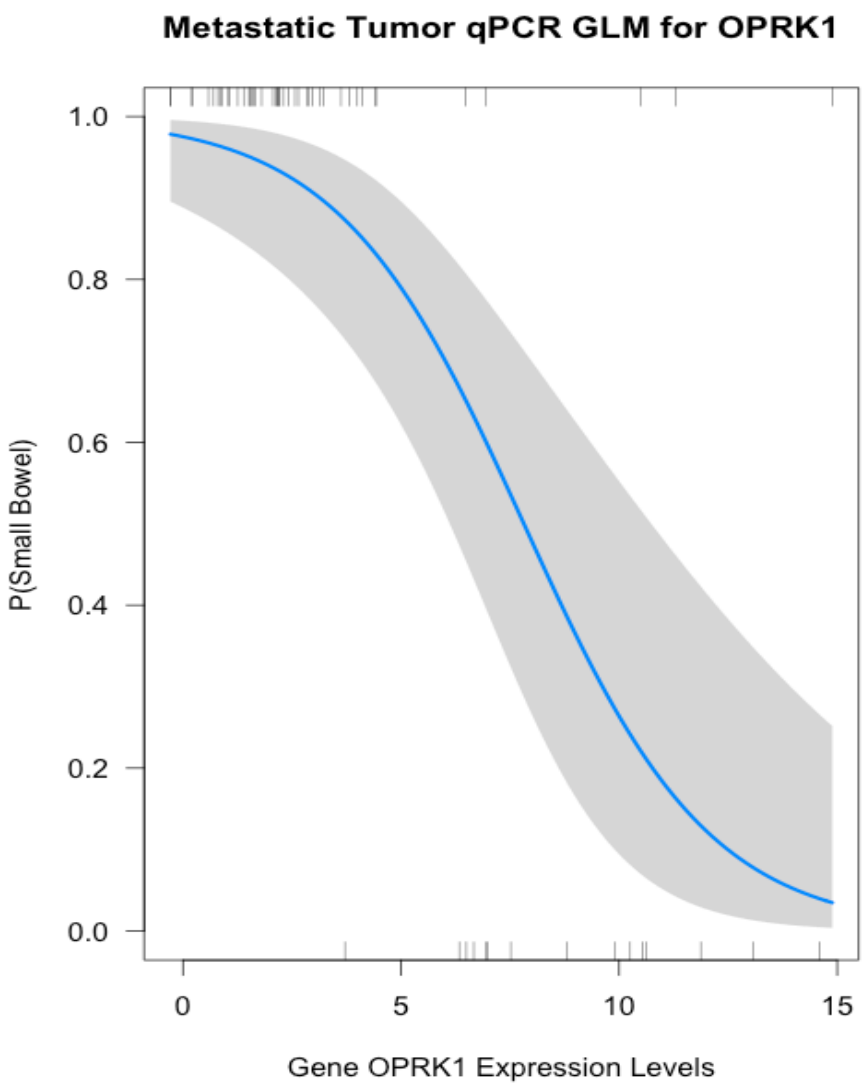
* Metastatic qPCR GEC:

| | row.names | PNET accuracy | SBNET accuracy | Overall accuracy |
|---|---|---|---|---|
| 1 | OPRK1 | 0.471 | 1.000 | 0.827 |
| 2 | OPRK1, GIPR | 0.588 | 0.914 | 0.808 |

* IHC: SBNET – 85.2%  PNET – 100%

# Logistic Regression Models



# GEC from Microarray Data

| | row.names | PNET Accuracy | SBNET Accuracy | Overall Accuracy | Self-Check Overall Accuracy | Difference |
|---|---|---|---|---|---|---|
| 1 | BRS3 | 80.0 | 91.0 | 87.6 | 81.3 | 6.3 |
| 2 | OPRK1 | 53.1 | 95.0 | 83.0 | 56.3 | 26.7 |
| 3 | DRD1 | 15.8 | 85.0 | 62.7 | 68.8 | 6.1 |
| 4 | GAPDH | 18.9 | 57.5 | 45.3 | 100.0 | 54.7 |
| 5 | GIPR | 10.5 | 91.5 | 65.8 | 68.8 | 3.0 |
| 6 | GPR98 | 0.0 | 100.0 | 68.1 | 68.8 | 0.7 |
| 7 | GRM1 | 54.1 | 58.8 | 57.3 | 62.5 | 5.2 |
| 8 | POLR2A | 59.5 | 87.5 | 78.6 | 56.3 | 22.3 |
| 9 | SSTR2 | 23.7 | 93.8 | 71.4 | 68.8 | 2.6 |
| 10 | SCTR | 64.9 | 64.3 | 64.5 | 100.0 | 35.5 |
| 11 | ADORA1 | 52.8 | 54.1 | 53.6 | 75.0 | 21.4 |
| 12 | OXTR | 56.8 | 82.9 | 74.3 | 100.0 | 25.7 |
| 13 | GPR113 | 54.1 | 62.8 | 60.0 | 93.8 | 33.8 |
| 14 | MUC13 | 0.0 | 100.0 | 68.1 | 68.8 | 0.7 |
| 15 | MEP1B | 0.0 | 100.0 | 67.5 | 75.0 | 7.5 |

- Overall prediction accuracy and the self-check values show that the Microarray Data agrees with Metastatic qPCR data.
  - Mean: 16.8
- However, there are a few genes that vary greatly, like GAPDH.

# Acknowledgements

We'd like to say thank you to the following people for providing support throughout our project.

**Dr. Patrick Breheny**

**Dr. Gideon Zamba**

**Terry Kirk**

**Melissa Pugh, Joe Moen, John Van Buren**

**Jessica E. Maxwell, MD , MBA**

IOWA

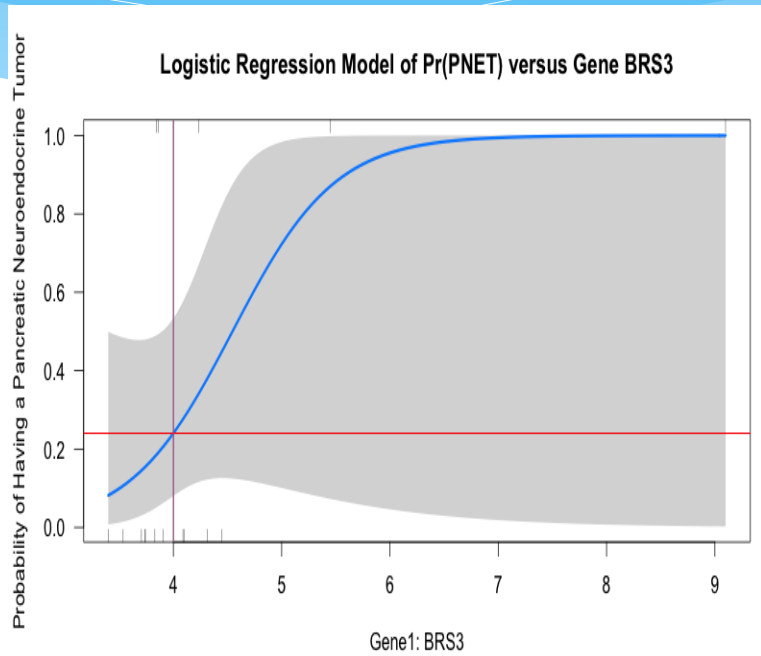
The University of Iowa College of Public Health

# Akaike Information Criterion

- AIC = $2k - 2\ln(L)$
  - k is the # of parameters
  - L is the maximized value of the Likelihood function
- "Add1" function in R
  - Which gene when added had a lower AIC.
- Every time an additional gene was added, we looked for smallest AIC.
- The AIC reflects on the amount of information lost if extra variables are added.
  - It is a measure used for model selection.
- Risks of Overfitting:
  - A good statistical model will be able to:
    - perform well on new data
    - will be able to generalize from the trend of the new data
  - Overfit models are built to closely resemble the training data with great accuracy and are like memorization
    - The overfit model will not be that effective at generalizing from the trend of the new data.

# Using a Logistic Regression to Enter Probabilities for the Data

- $log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) = -9.587 + 2.109(Mean\ Case\ Value\ for\ BRS3\ Gene)$
- For a **Mean Case Value of 4…**
- $log_e\left(\frac{PNET\ Probability}{1-PNET\ Probability}\right) = -9.587 + 2.109(4)$
- $log_e\left(\frac{PNET\ Probability}{1-PNET\ Probability}\right) = -9.587 + 8.436$
- $log_e\left(\frac{PNET\ Probability}{1-PNET\ Probability}\right) = -1.151$
- $\frac{PNET\ Probability}{1-PNET\ Probability} = e^{-1.151} = 0.3163203$
- $PNET\ Probability = 0.3163203 - (0.3163203)(PNET\ Probability)$
- $1.31632(PNET\ Probability) = 0.3163203$
- $PNET\ Probability = 0.2403 = 24.03\%$

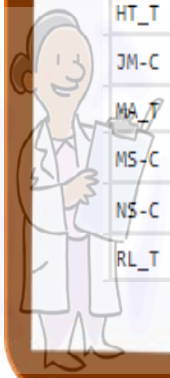


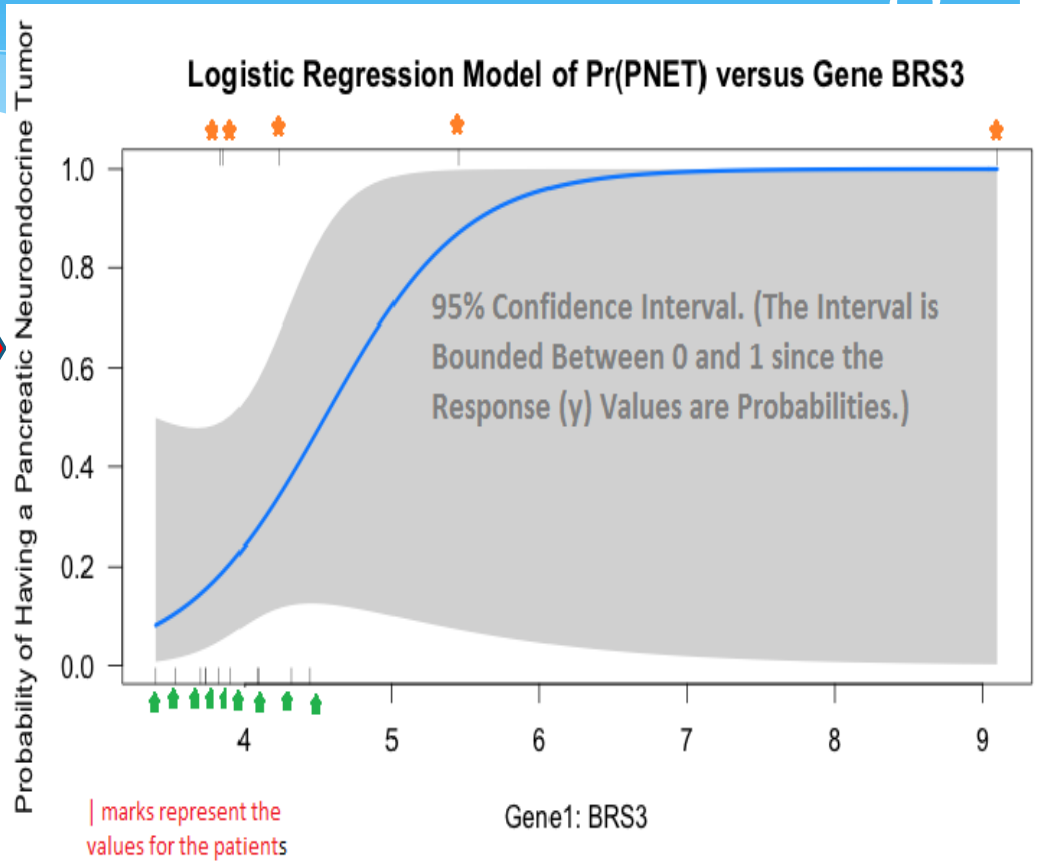

This process will be applied for the other data points in the Testing Data.

# Building the Alg Training D

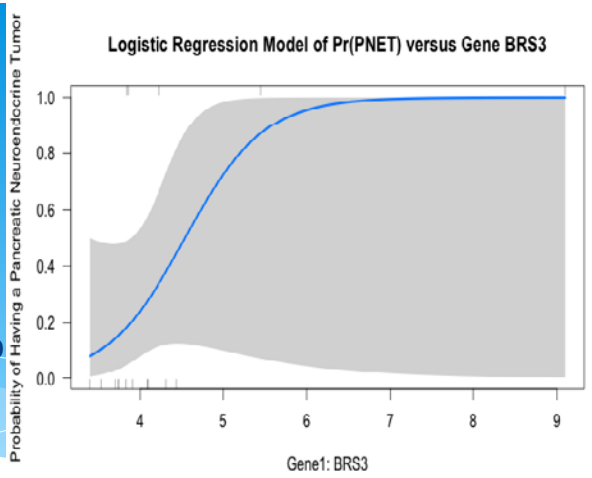
**Actual Data from MicroArray Data Set**

| Patient | Patient NET | BRS3 gene |
|---|---|---|
| 01_186_Tumor | Ileal | -0.79409762 |
| 03_209_Tumor | Ileal | -0.01393512 |
| 05_194_Tumor | Ileal | 0.11351113 |
| 07_160_Tumor | Ileal | -0.93106762 |
| 09_148_Tumor | Ileal | -0.24231137 |
| BH-C | Ileal | -0.42476074 |
| CD-C | Ileal | -0.58870262 |
| CE-C | Ileal | -0.50114949 |
| CM_T | Pancreatic | -0.09687387 |
| HM_T | Pancreatic | -0.47115262 |
| HT_T | Pancreatic | 1.11897988 |
| JM-C | Ileal | -0.62516512 |
| MA_T | Pancreatic | -0.48870387 |
| MS-C | Ileal | -0.59045137 |
| NS-C | Ileal | -0.23267199 |
| RL_T | Pancreatic | 4.76855238 |






Visreg was used.




```
Call:
glm(formula = response ~ explanatory, family = "bino

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0877  -0.7101  -0.5698   0.1404   1.8379

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.587      6.425  -1.492    0.136
explanatory    2.109      1.572   1.342    0.180

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19.875  on 15  degrees of freedom
Residual deviance: 14.812  on 14  degrees of freedom
AIC: 18.812

Number of Fisher Scoring iterations: 6
```

$$\ln\left[\frac{PNET\ Probability}{SBNET\ Probability}\right] =$$

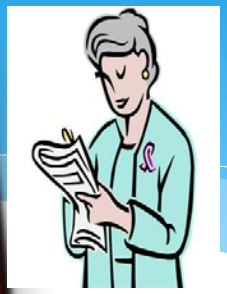
$$\hat{\alpha} + (\widehat{\beta})(Mean\ Case\ Value)$$

$$\ln\left[\frac{PNET\ Probability}{SBNET\ Probability}\right] =$$

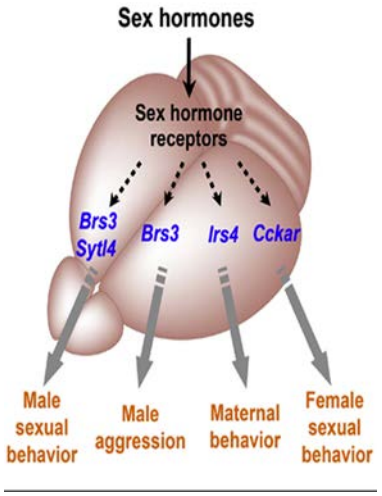
$$-9.587 + (2.109)(Mean\ Case\ Value\ for\ BRS3\ Gene)$$

## Testing Data (for BRS3)

| Patient # | Metastatic QPCR Data BRS3 Gene Expression Values |
|---|---|
| 1 | 6.272142 |
| 2 | NA |
| 3 | 1.421762 |
| 4 | 3.106775 |
| 5 | 11.584643 |
| 6 | NA |
| ... | ... |
| 59 | 10.877939 |
| 60 | 10.623849 |
| 61 | 11.849864 |
| 62 | 12.825 |
| 63 | NA |
| ... | ... |
| 145 | 10.09073 |
| 146 | 7.527785 |
| 147 | 11.627995 |
| 148 | 12.017324 |
| 149 | 11.810563 |
| 150 | 11.140289 |
| 151 | NA |
| 152 | NA |



# Using the logistic regression model to predict what each patient should be diagnosed with given that patient's

# The Logistic Regression Model Used to Get these Probabilities:

$$\ln\left[\frac{PNET\ Probability}{SBNET\ Probability}\right] =$$

$$-9.587 + (2.109)(Mean\ Case\ Value\ for\ BRS3\ Gene)$$

BRS3 Prediction Probabilities based upon BRS3 Gene Using the BRS3 Logistic Model

| Patient # | Probabilities for Metastatic QPCR data |
|---|---|
| 1 | 0.77799618 |
| 2 | NA |
| 3 | 0.98722256 |
| 4 | 0.96347842 |
| 5 | 0.10584873 |
| 6 | NA |
| ... | ... |
| 59 | 0.1566743 |
| 60 | 0.17929283 |
| 61 | 0.09087475 |
| 62 | 0.05093798 |
| 63 | NA |
| ... | ... |
| 145 | 0.23484277 |
| 146 | 0.61141829 |
| 147 | 0.10326051 |
| 148 | 0.08242893 |
| 149 | 0.09296672 |
| 150 | 0.13581532 |
| 151 | NA |
| 152 | NA |

| BRS3 Prediction Probabilities based upon BRS3 Gene Using the BRS3 Logistic Model | |
|---|---|
| Patient # | Probabilities for Metastatic QPCR data |
| 1 | 0.77799618 |
| 2 | NA |
| 3 | 0.98722256 |
| 4 | 0.96347842 |
| 5 | 0.10584873 |
| 6 | NA |
| ... | ... |
| 59 | 0.1566743 |
| 60 | 0.17929283 |
| 61 | 0.09087475 |
| 62 | 0.05093798 |
| 63 | NA |
| ... | ... |
| 145 | 0.23484277 |
| 146 | 0.61141829 |
| 147 | 0.10326051 |
| 148 | 0.08242893 |
| 149 | 0.09296672 |
| 150 | 0.13581532 |
| 151 | NA |
| 152 | NA |

FOR THE METASTATIC QPCR DATA:

* Probability > 50% → Pancreas Cancer is Predicted
* Probability < 50% → Small Bowel (Ileal) Cancer is Predicted

| Patient # | Prediction for Metastatic QPCR Data Patients |
|---|---|
| 1 | Pancreas |
| 2 | NA |
| 3 | Pancreas |
| 4 | Pancreas |
| 5 | Small Bowel |
| 6 | NA |
| ... | ... |
| 59 | Small Bowel |
| 60 | Small Bowel |
| 61 | Small Bowel |
| 62 | Small Bowel |
| 63 | NA |
| ... | ... |
| 145 | Small Bowel |
| 146 | Pancreas |
| 147 | Small Bowel |
| 148 | Small Bowel |
| 149 | Small Bowel |
| 150 | Small Bowel |
| 151 | NA |
| 152 | NA |

# Analyzing differences between Actual NET and Predicted NET for the BRS3 Gene for the Metastatic QPCR patients:

| Prediction for Metastatic QPCR Data Patients Based on Probabilities using BRS3 gene | | Actual Site for Metastatic QPCR Data Patients | |
|---|---|---|---|
| Patient # | Prediction for Metastatic QPCR Data Patients | Patient # | Actual Site for Metastatic QPCR Patients |
| 1 | Pancreas | 1 | Small Bowel |
| 2 | NA | 2 | Small Bowel |
| 3 | Pancreas | 3 | Pancreas |
| 4 | Pancreas | 4 | Pancreas |
| 5 | Small Bowel | 5 | Small Bowel |
| 6 | NA | 6 | Small Bowel |
| ... | ... | ... | ... |
| 59 | Small Bowel | 59 | Small Bowel |
| 60 | Small Bowel | 60 | Small Bowel |
| 61 | Small Bowel | 61 | Small Bowel |
| 62 | Small Bowel | 62 | Small Bowel |
| 63 | NA | 63 | Small Bowel |
| ... | ... | ... | ... |
| 145 | Small Bowel | 145 | Small Bowel |
| 146 | Pancreas | 146 | Small Bowel |
| 147 | Small Bowel | 147 | Small Bowel |
| 148 | Small Bowel | 148 | Small Bowel |
| 149 | Small Bowel | 149 | Small Bowel |
| 150 | Small Bowel | 150 | Small Bowel |
| 151 | NA | 151 | Small Bowel |
| 152 | NA | 152 | Small Bowel |

| | ActualMetastatic | | |
|---|---|---|---|
| BRS3Predictions (Predictions based on gene) | Pancreas | Small Bowel | Row totals |
| Pancreas | **28** | 7 | 35 |
| Small Bowel | 7 | **71** | 78 |
| Column Total (Actual Totals for Metastatic) | 35 | 78 | 113 |

**Accuracy of Gene BRS3 for Predicting Metastatic QPCR results:**

| Gene | Probability(That the Prediction Says **Pancreatic** NET \| The Patient Actually Has Pancreatic N.E. Tumor). | Probability(That the Prediction Says **Ileal** \| The Patient Actually Has Ileal (Small Bowel) Cancer). | Overall Accuracy (# of Accurate Predictions) / (# of total predictions) |
|---|---|---|---|
| BRS3.dCT | 0% | 60% | 18.75% |
| | | | |

We omit the "NA" values because some data was missing for the BRS3 gene
The Results are Compared for the Site
This helps determine the accuracy of the BRS3 gene.

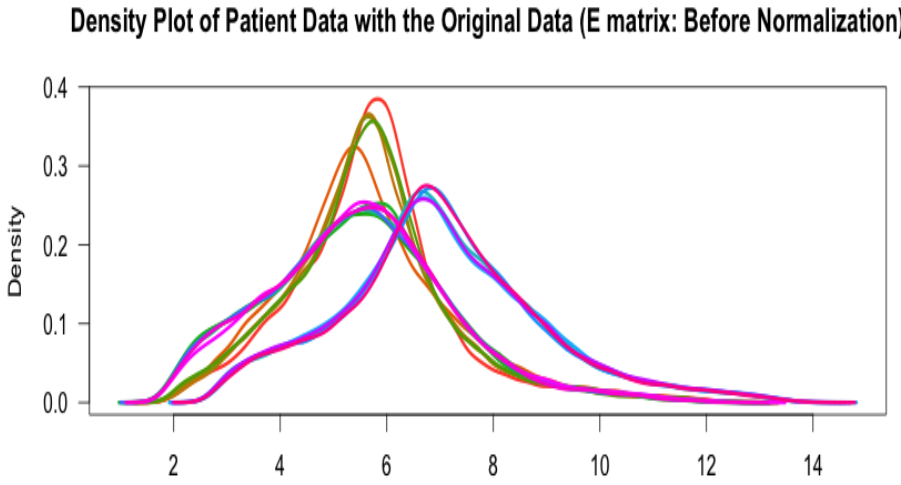
# Microarray → Microarray Model Results for the BRS3 gene:

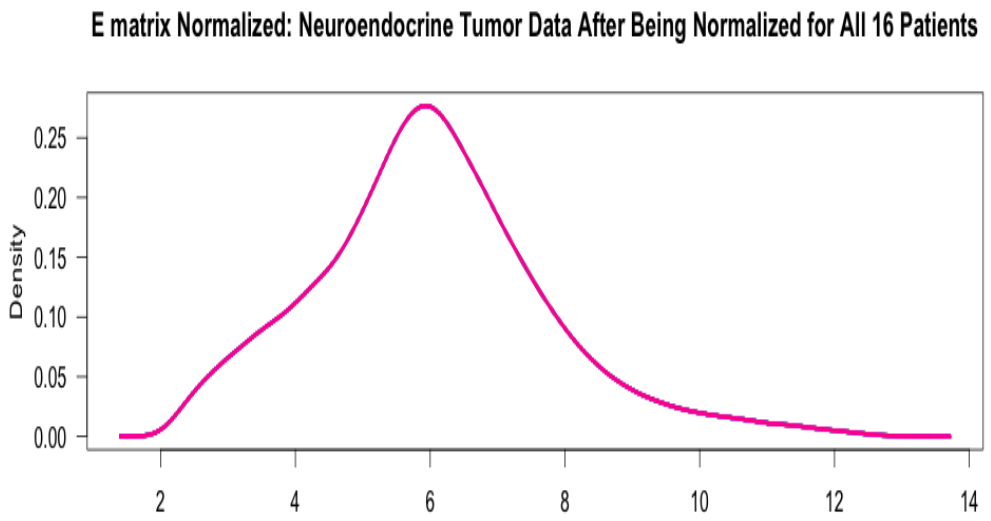
| | Actual Metastatic | | |
|---|---|---|---|
| BRS3 Predictions (Predictions based on gene) | Pancreatic | Ileal | Row totals |
| Pancreatic | **0** | 2 | 2 |
| Ileal | 11 | **3** | 14 |
| Column Total (Actual Totals for Metastatic) | 11 | 5 | 16 |

| Gene | Probability(That the Prediction Says **Pancreatic** NET \| The Patient Actually Has Pancreatic N.E. Tumor). | Probability(That the Prediction Says **Ileal** \| The Patient Actually Has Ileal (Small Bowel) Cancer). | Overall Accuracy (# of Accurate Predictions) / (# of total predictions) |
|---|---|---|---|
| BRS3.dCT | 0% | 60% | 18.75% |

# Normalization of Data

* Normalization was necessary for Microarray Data to account for possible differences between data collection for pancreatic and small bowel neuroendocrine tumors.
  * Centered the data at 0



# AUC

| | row.names | Primary qPCR | Metastatic qPCR | Microarray |
|---|---|---|---|---|
| 1 | BRS3 | 0.82239583 | 0.90402930 | 0.78181818 |
| 2 | OPRK1 | 0.05034446 | 0.07265625 | 0.36363636 |
| 3 | DRD1 | 0.61437908 | 0.60230263 | 0.49090909 |
| 4 | GAPDH | 0.53260329 | 0.75168919 | 0.00000000 |
| 5 | GIPR | 0.59663866 | 0.46148909 | 0.45454545 |
| 6 | GPR98 | 0.58216620 | 0.45614035 | 0.47272727 |
| 7 | GRM1 | 0.60504202 | 0.53817568 | 0.67272727 |
| 8 | POLR2A | 0.46640119 | 0.24831081 | 0.29090909 |
| 9 | SSTR2 | 0.63725490 | 0.50292398 | 0.38181818 |
| 10 | SCTR | 0.85116279 | 0.76447876 | 1.00000000 |
| 11 | ADORA1 | 0.44030612 | 0.41929429 | 0.09090909 |
| 12 | OXTR | 0.24786325 | 0.19914651 | 0.00000000 |
| 13 | GPR113 | 0.47697218 | 0.28898129 | 0.07272727 |
| 14 | MUC13 | 0.47572362 | 0.62230585 | 0.30909091 |
| 15 | MEP1B | 0.39285714 | 0.69120586 | 0.65454545 |