Predicting Brain Cancer By Gene Expression

Anna Corcoran, Charlie Konen, Ethan Shang, Kethan Poduri, Daniel Cohen

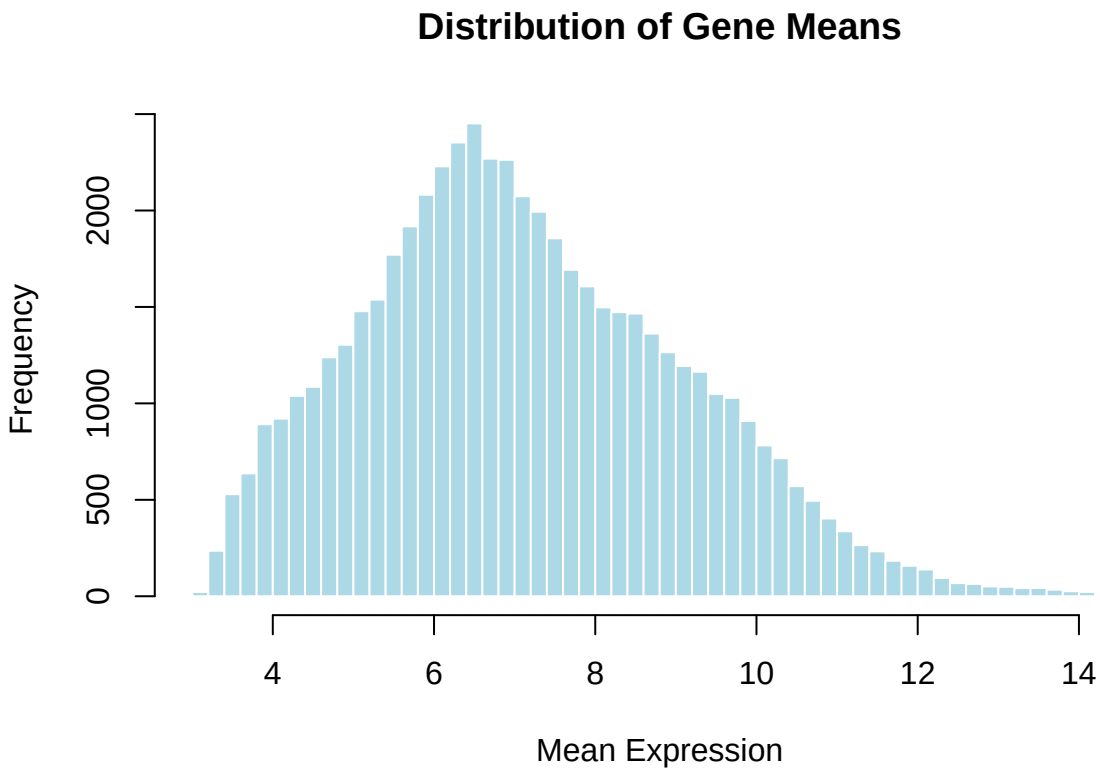
2025-12-15

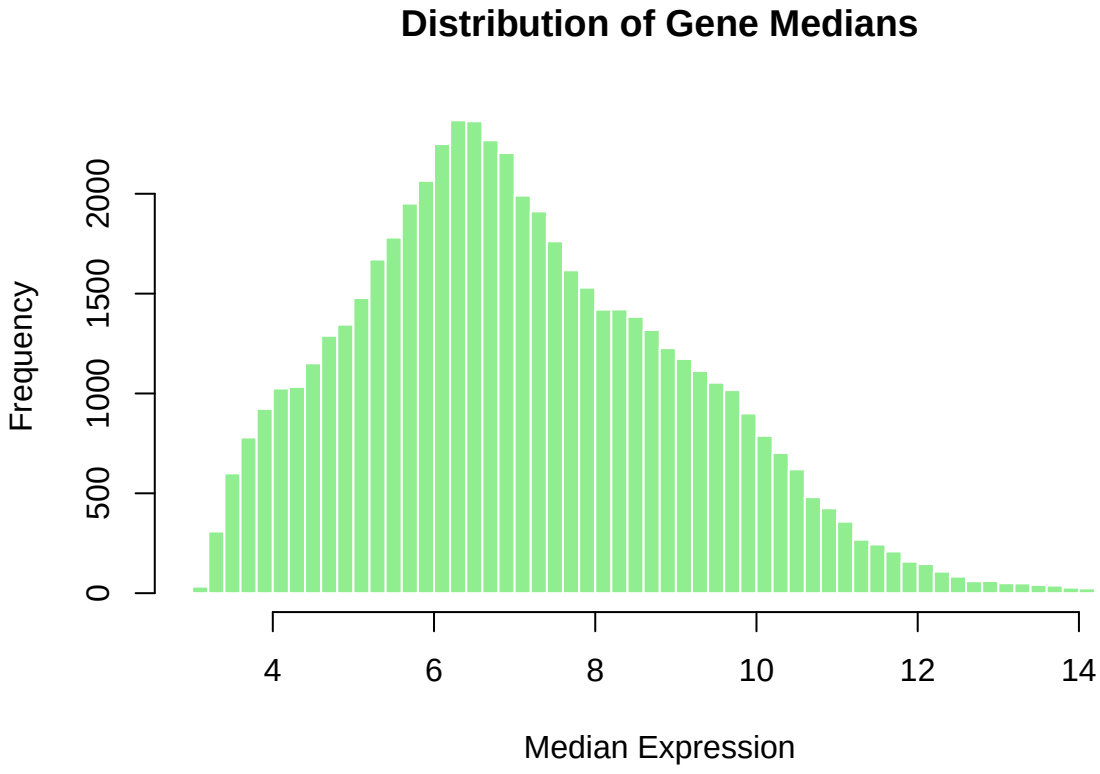
Introduction

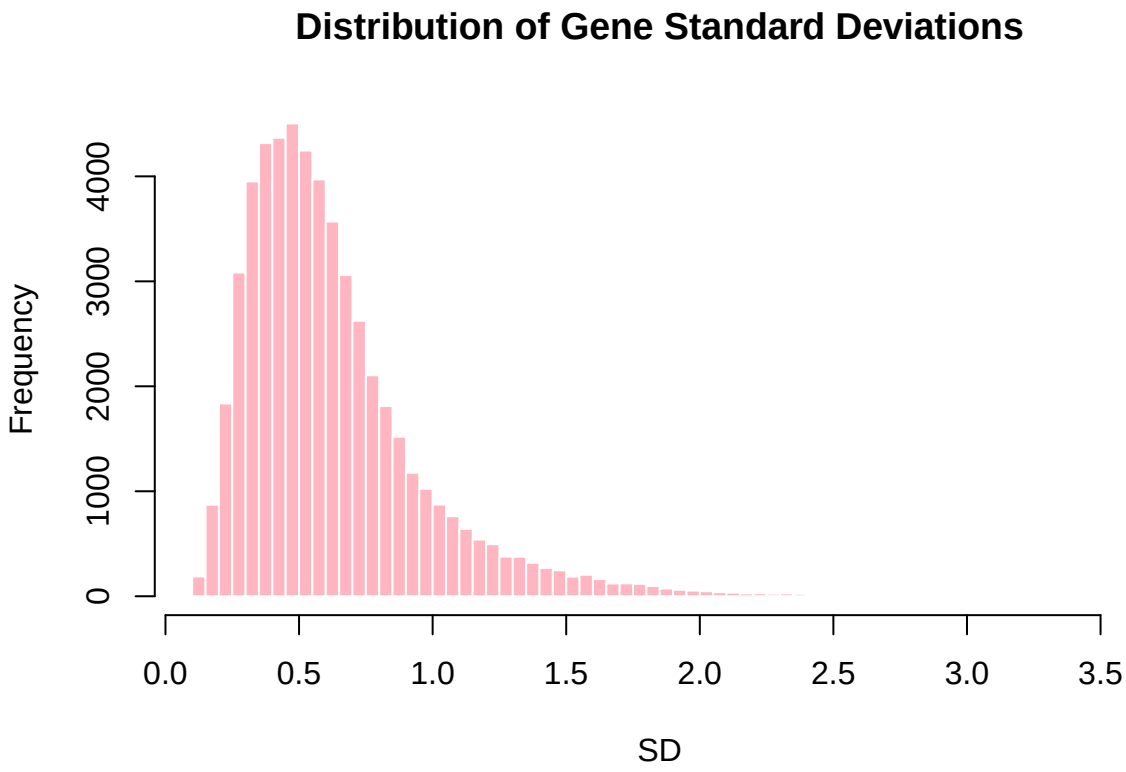
Brain cancer is extremely deadly, with a combined estimated 18,330 deaths in the United States in 2025 alone (SEER). Only 40% of adults diagnosed with a malignant brain tumor live for over a year, and less than 20% live for more than 5 years (Penfold). This project covers four types of brain cancer: ependymoma, glioblastoma, medulloblastoma, and pilocytic astrocytoma, with varying severity. Pilocytic astrocytoma is nearly always treatable, with a 96% 5-year survival rate (mayo). Glioblastoma, on the other hand, is incredibly deadly, with a median 5-year survival rate of only 5.6% in adults (ABTA), while medulloblastoma has a median 5-year survival rate of 80.6% (cancer.gov) and ependymoma of nearly 85% (cleveland clinic), making them relatively treatable. Early diagnosis is difficult due to nonspecific symptoms like headaches, weakness, confusion, and memory loss (Penfold) and requires a neurological exam; CT scan, brain MRI and/or PET scan to locate a mass; and then a brain biopsy and testing of the biopsy in lab to determine if the cells are cancerous (mayo 2). This is lengthy and expensive, delaying accurate prognoses. Brain biopsies are also invasive, requiring a hole to be drilled into the skull and a needle to be inserted into the hole (mayo 2). The goal of this project is to create a model that can accurately predict the state of a sample based on gene expression. This could then be used on circulating tumor DNA (ctDNA), which is cell-free DNA found in the bloodstream, to obtain accurate diagnoses earlier in progression without the invasion of a tumor biopsy (Kim). Early and accurate diagnoses can streamline treatment, getting patients into appropriate clinical trials and improving survival rates.

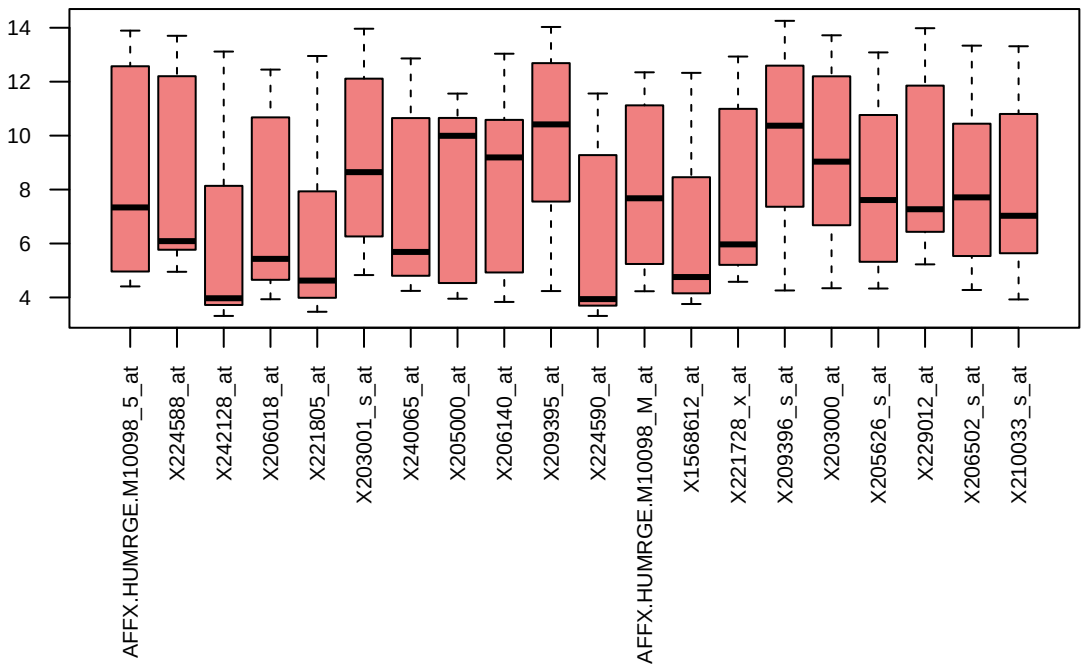
```
## function (x, ...)
## UseMethod("print")
## <bytecode: 0x11bd613b8>
## <environment: namespace:base>
```

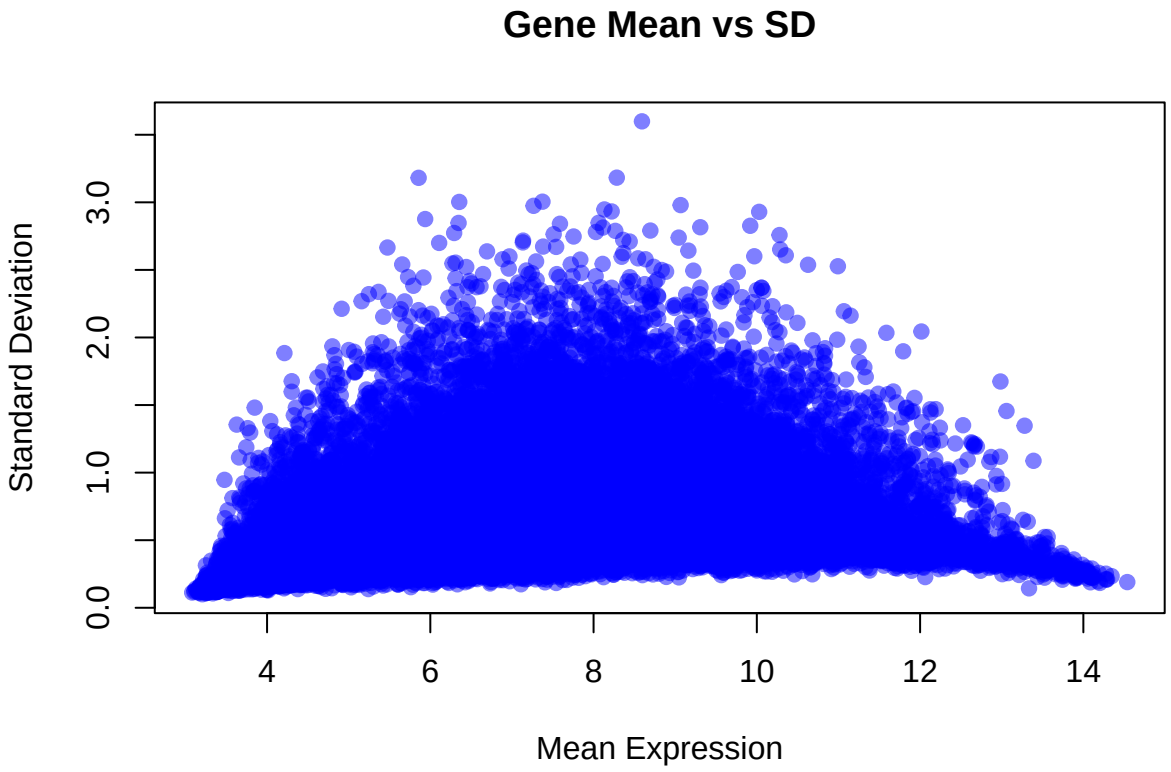
```
##                   Gene      Mean    Median        SD       Min       Max
## X1007_s_at X1007_s_at 12.276393 12.502518 0.7901601 10.156207 13.655639
## X1053_at     X1053_at  8.769583  8.786242 0.6733962  6.627878 10.716003
## X117_at       X117_at  7.722634  7.521674 1.0373393  6.222515 12.054143
## X121_at       X121_at  9.160209  9.194487 0.6153686  8.044421 10.407136
## X1255_g_at X1255_g_at  4.842069  4.462729 0.9220032  3.682762  7.404503
## X1294_at     X1294_at  7.968388  7.915062 0.6302601  6.560920 10.164655
```

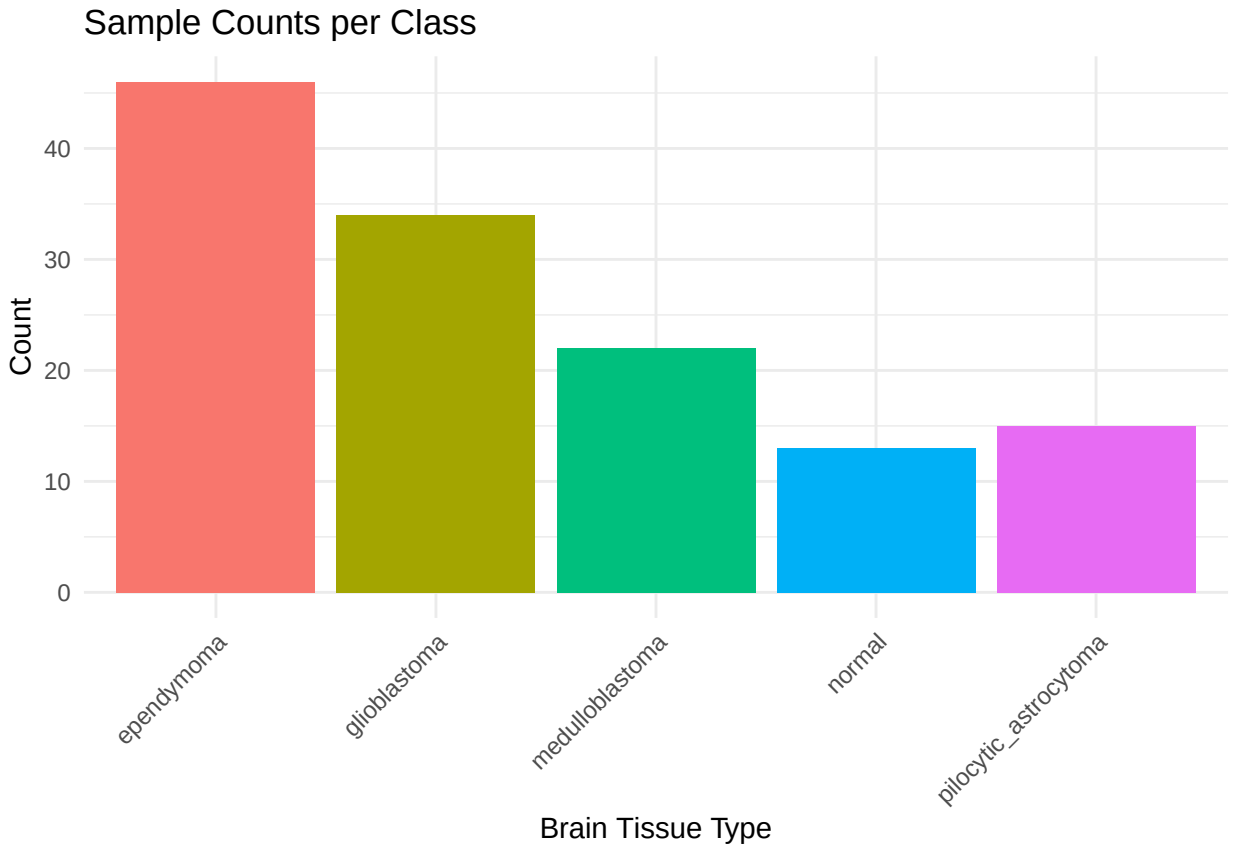

Distribution of Gene Means
Frequency
0
500
1000
2000
4
6
8
10
12
14
Mean Expression

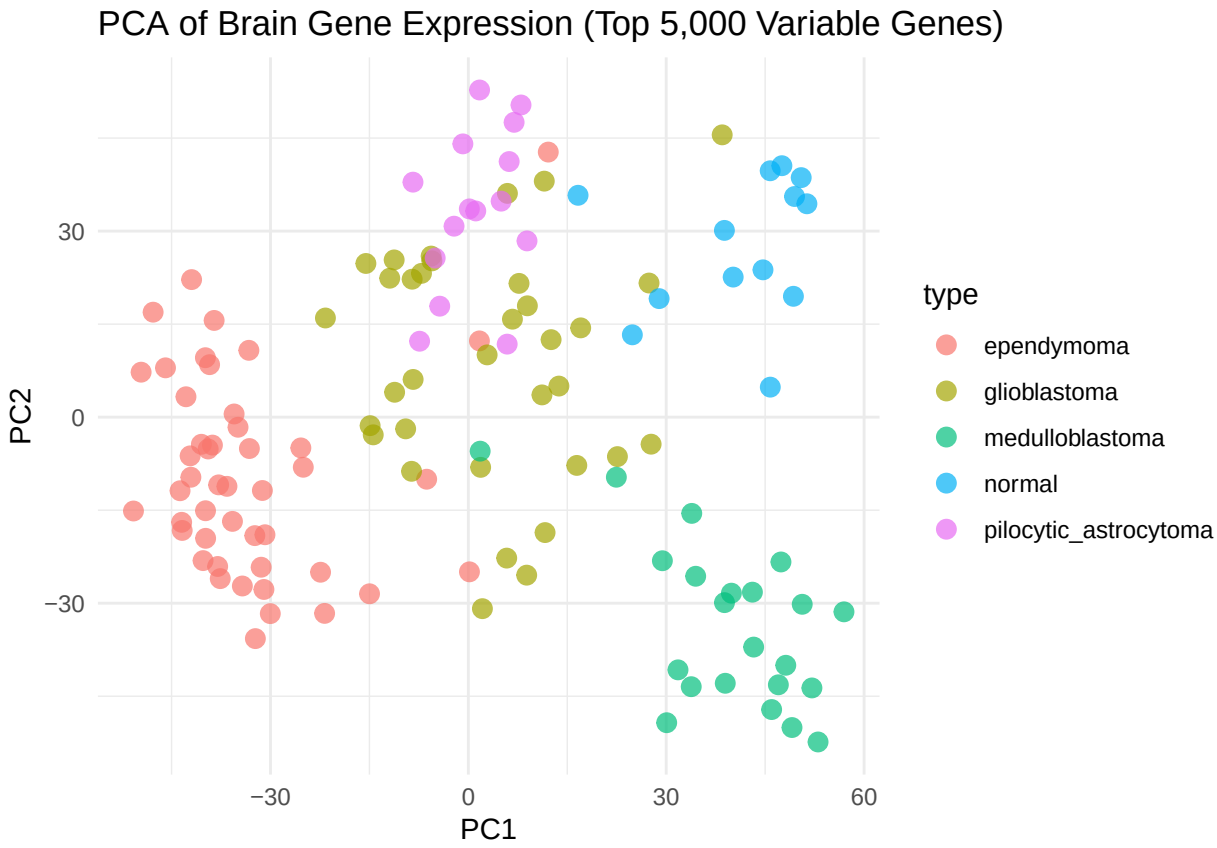

Distribution of Gene Medians
Frequency
0
500
1000
1500
2000
4
6
8
10
12
14
Median Expression


Distribution of Gene Standard Deviations
Frequency
0
1000
2000
3000
4000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
SD

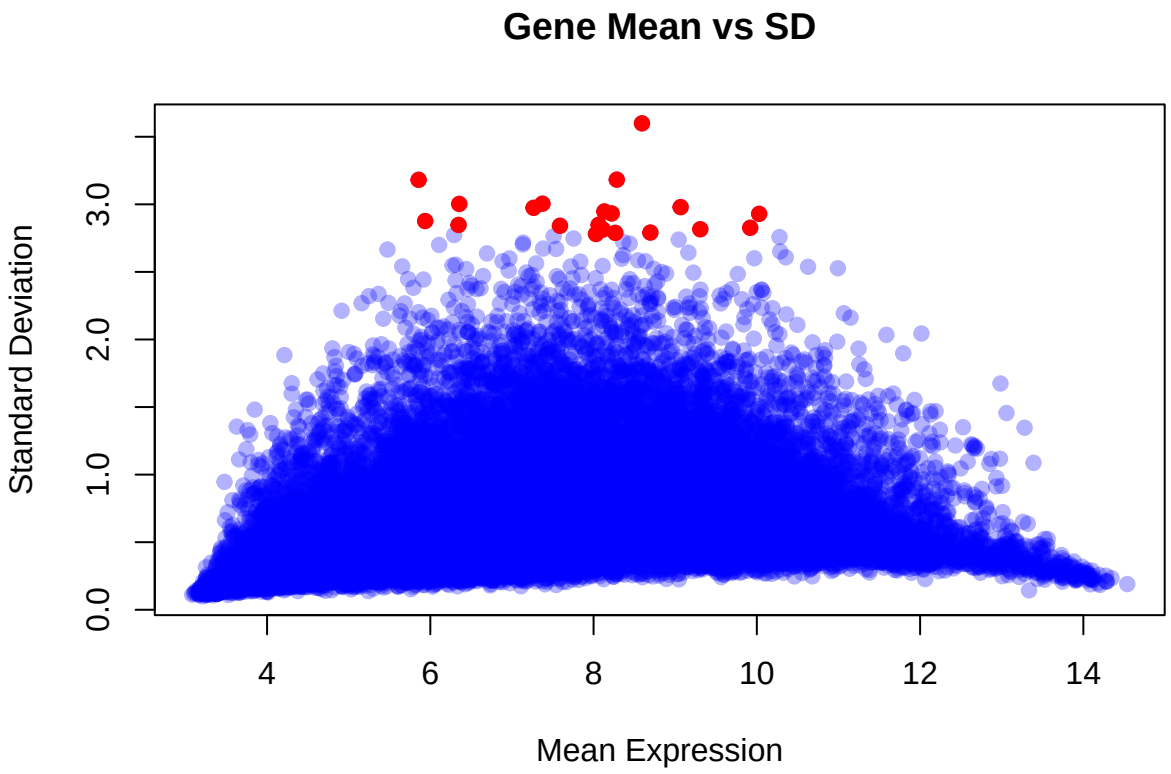

Top 20 Most Variable Genes
14
12
10
8
6
4
AFFX.HUMRGE.M10098_5_at
X224588_at
X242128_at
X206018_at
X221805_at
X203001_s_at
X240065_at
X205000_at
X206140_at
X209395_at
X224590_at
AFFX.HUMRGE.M10098_M_at
X1568612_at
X221728_x_at
X209396_s_at
X203000_at
X205626_s_at
X229012_at
X206502_s_at
X210033_s_at


Gene Mean vs SD
Standard Deviation
0.0
1.0
2.0
3.0
4
6
8
10
12
14
Mean Expression


Sample Counts per Class
Count
0
10
20
30
40
ependymoma
glioblastoma
medulloblastoma
normal
pilocytic_astrocytoma
Brain Tissue Type




```
##            Metric     Value
## 1   Mean of Means 7.1292377
## 2 Median of Means 6.9159918
## 3         Mean SD 0.6262881
## 4       Median SD 0.5495749
```



Variable Selection - Binary Case

Combine all different types of cancer into one cancer/not cancer binary

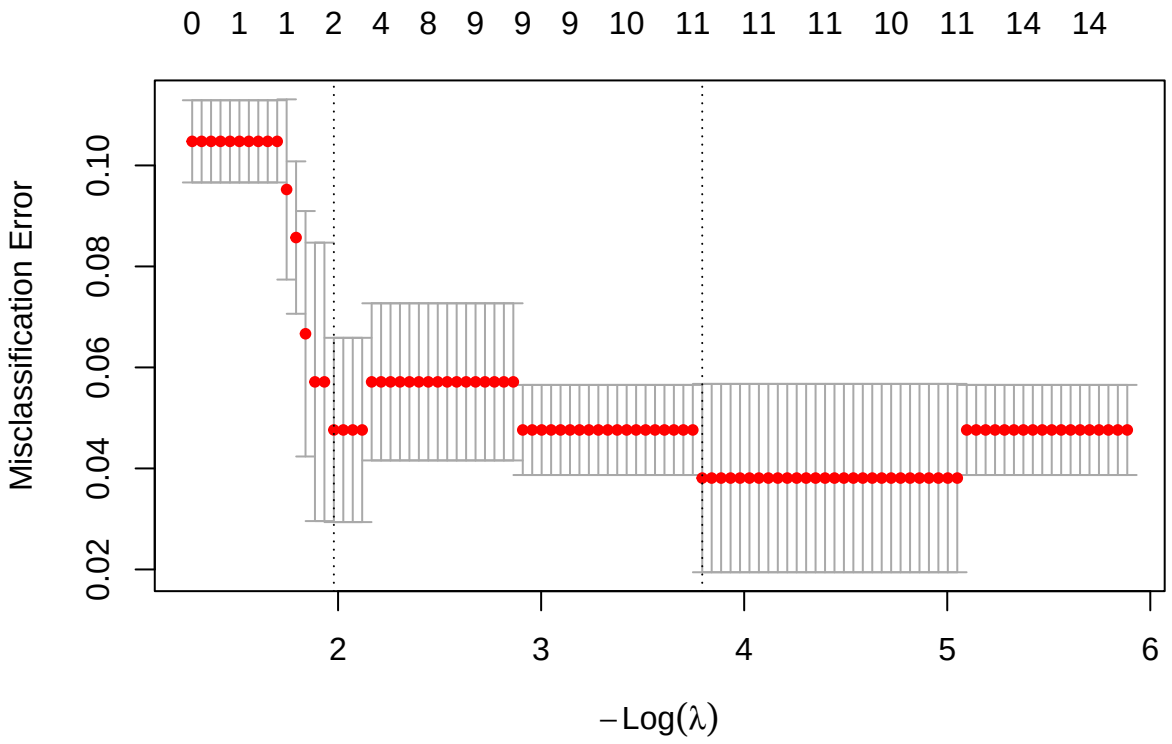
```
## y_bin
## non_cancer     cancer
##         13        117
```

Train/test data

3 - fold lasso

- 3 fold because there are only 13 non-sample cancer (what the kaggle dataset reccomended)

Choosing lambda:




```
## [1] 0.02251901
```

```
## [1] 0.1381745
```

Extract selected genes / accuracy

```
## [1] 12
```

```
## [1] 2
```

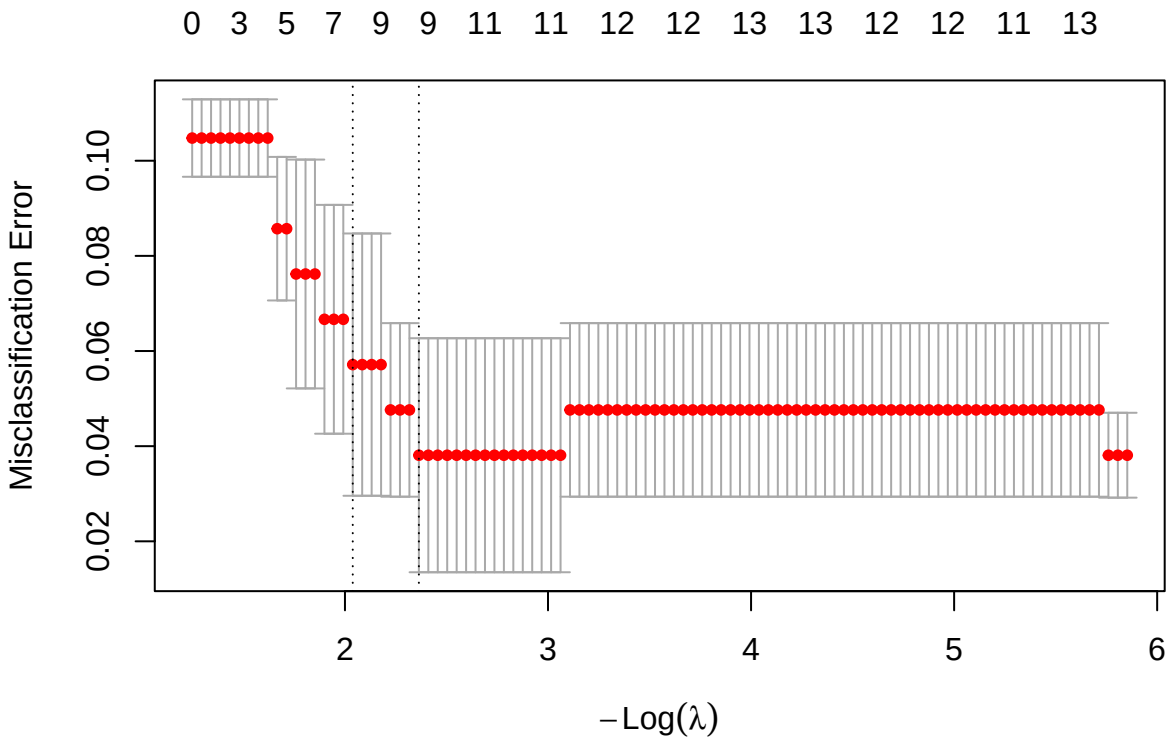
```
##                 Model     Lambda Num_Selected_Genes Test_Accuracy
## 1 Lasso (lambda.min) 0.02251901                 12          1.00
## 2 Lasso (lambda.1se) 0.13817454                  2          0.96
```

```
## [1] "X214145_s_at" "X216073_at"
```

2 Strong genes, look at what they actually are:

```
##       PROBEID   SYMBOL ENTREZID                    GENENAME
## 1 214145_s_at     SPTB     6710 spectrin beta, erythrocytic
## 2   216073_at ANKRD34C   390616   ankyrin repeat domain 34C
```

Because pairwise gene–gene interactions (epistasis) can play an important role in tumor biology, we next asked whether the two genes selected by our initial Lasso model participate in meaningful interaction effects with other genes. Fitting all pairwise interactions among ~55,000 genes would require modeling billions of terms and is computationally infeasible, so we adopted a weak-hierarchy strategy centered on "anchor" genes. Specifically, we treated the two Lasso-selected probes (X214145_s_at and X216073_at) as anchors and constructed a high-dimensional design matrix that included all main effects for the ~55,000 genes, together with interaction terms between each anchor gene and every gene in the dataset (i.e., X214145_s_at×G_k and X216073_at×G_k for all k). We then fit a Lasso-penalized logistic regression model to this expanded feature set, allowing the penalty to select a sparse subset of main and interaction effects. This targeted interaction scan respects weak hierarchy, focuses on interactions involving genes already implicated by the main-effects model, and provides a computationally tractable way to screen for potential epistatic partners of the two anchor genes.




```
## [1] 0.09404014
```

```
## [1] 0.1302349
```

```
## [1] "X1554643_at:X216073_at"  "X1554771_at:X216073_at"
## [3] "X1556498_at:X216073_at"  "X1556881_at:X216073_at"
## [5] "X207284_s_at:X216073_at" "X233423_at:X216073_at"
## [7] "X236037_at:X216073_at"
```

```
## character(0)
```

Applying this interaction-augmented Lasso model revealed a focused set of epistatic candidates. Although the two anchor probes were selected based on their strong marginal association with cancer status, the interaction scan showed that only **X216073_at** exhibited evidence of meaningful combinatorial effects with other genes. Specifically, the model retained **seven interaction terms**, each representing an interaction between **X216073_at** and a distinct partner gene (X1554643_at, X1554771_at, X1556498_at, X1556881_at, X207284_s_at, X233423_at, and X236037_at). The absence of interaction terms involving X214145_s_at suggests its contribution to the phenotype is largely additive, whereas X216073_at may function as an interaction hub whose joint activity with multiple other genes enhances discrimination between cancer and normal samples. These retained interaction terms thus represent prioritized candidates for downstream biological interpretation and may point toward coordinated regulatory mechanisms underlying the cancer phenotype.

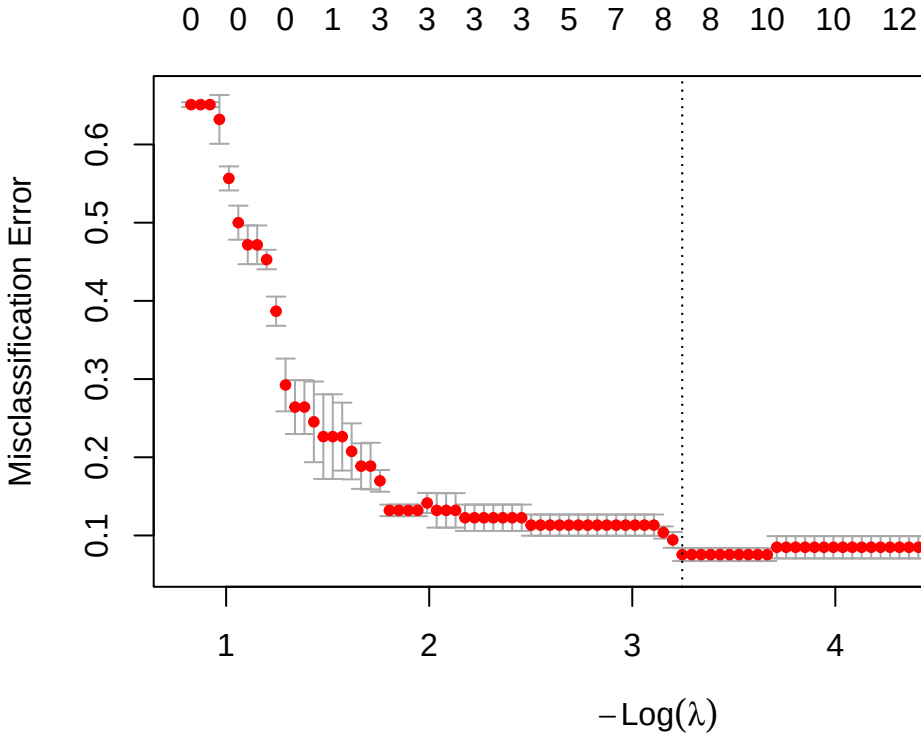
Binary Model Selection

Multinomial Logistic Regression Model

Plan: Use the original 5-class type as the multiclass outcome. Restrict predictors to highly variable genes (HVG candidates) using SD (top 5,000). Fit a multinomial LASSO (glmnet, family = "multinomial") on those HVGs to select top predictive genes. From these, choose the top 100 genes (by coefficient magnitude). Refit a multinomial logistic regression model using only those 100 genes (again via glmnet) and evaluate on a held-out test set.

##HVGs Setup multiclass response and HVG candidates

Train/test split by 5 class type



Multinomial LASSO to select predictive HVGs

```
## [1] 0.03892134
```

```
## [1] 0.03892134
```

Pull out nonzero features, keep the top 100 by coefficient magnitude (HVG)

```
## [1] 47
```

```
## [1] 47
```

```
## [1] "X204933_s_at" "X204932_at"   "X220156_at"   "X230695_s_at" "X244364_at"
## [6] "X235334_at"
```

Fit multinomial logistic regression model with top 100 HVGs

```
## [1] 0.002176018
```

```
## [1] 0.008784623
```

Predictions and performance

```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction                ependymoma glioblastoma medulloblastoma normal
##   ependymoma                       9            0               0      0
##   glioblastoma                     0            5               0      0
##   medulloblastoma                  0            0               4      0
##   normal                           0            1               0      2
##   pilocytic_astrocytoma            0            0               0      0
##                          Reference
## Prediction                pilocytic_astrocytoma
```

```
##   ependymoma                                0
##   glioblastoma                              0
##   medulloblastoma                           0
##   normal                                    0
##   pilocytic_astrocytoma                     3
##
## Overall Statistics
##
##                Accuracy : 0.9583
##                  95% CI : (0.7888, 0.9989)
##     No Information Rate : 0.375
##     P-Value [Acc > NIR] : 2.452e-09
##
##                   Kappa : 0.9447
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: ependymoma Class: glioblastoma
## Sensitivity                      1.000              0.8333
## Specificity                      1.000              1.0000
## Pos Pred Value                   1.000              1.0000
## Neg Pred Value                   1.000              0.9474
## Prevalence                       0.375              0.2500
## Detection Rate                   0.375              0.2083
## Detection Prevalence             0.375              0.2083
## Balanced Accuracy                1.000              0.9167
##                      Class: medulloblastoma Class: normal
## Sensitivity                          1.0000       1.00000
## Specificity                          1.0000       0.95455
## Pos Pred Value                       1.0000       0.66667
## Neg Pred Value                       1.0000       1.00000
## Prevalence                           0.1667       0.08333
## Detection Rate                       0.1667       0.08333
## Detection Prevalence                 0.1667       0.12500
## Balanced Accuracy                    1.0000       0.97727
##                      Class: pilocytic_astrocytoma
## Sensitivity                                 1.000
## Specificity                                 1.000
## Pos Pred Value                              1.000
## Neg Pred Value                              1.000
## Prevalence                                  0.125
## Detection Rate                              0.125
## Detection Prevalence                        0.125
## Balanced Accuracy                           1.000

##  Accuracy
## 0.9583333

##               Lambda Num_Selected Test_Accuracy
## Accuracy 0.008784623           47     0.9583333

##                        Reference
## Prediction              ependymoma glioblastoma medulloblastoma normal
```

```
##   ependymoma                    9            0              0      0
##   glioblastoma                  0            5              0      0
##   medulloblastoma               0            0              4      0
##   normal                        0            1              0      2
##   pilocytic_astrocytoma         0            0              0      0
##                        Reference
## Prediction              pilocytic_astrocytoma
##   ependymoma                                0
##   glioblastoma                              0
##   medulloblastoma                           0
##   normal                                    0
##   pilocytic_astrocytoma                     3
```

Methodology: For subtype prediction, we first restricted attention to highly variable genes by computing the standard deviation of each probe across samples and retaining the top 5,000 most variable probes. Using these 5,000 HVGs as predictors and the five-class tissue label as the response (normal plus four tumor subtypes), we fit a LASSO-penalized multinomial logistic regression model via glmnet with 3-fold cross-validation. From the resulting model at the $\lambda.1se$ penalty level, we ranked all nonzero features by the maximum absolute coefficient across classes and selected the top 100 probes as our final HVG feature set. We then refit a multinomial logistic regression model using only these 100 probes and evaluated predictive performance on a held-out test set using overall accuracy and the confusion matrix.

Results: A sparse multinomial LASSO classifier using only 47 gene expression features achieved 95.8% test accuracy across five brain tissue types. Three tumor subtypes (ependymoma, medulloblastoma, and pilocytic astrocytoma) were classified with perfect sensitivity and specificity. The only misclassifications involved a single glioblastoma sample predicted as normal. This result demonstrates that a low-dimensional gene panel is sufficient for near-perfect discrimination among major pediatric and adult brain tumor classes.

References: https://seer.cancer.gov/statfacts/html/brain.html Penfold C, Joannides AJ, Bell J, Walter FM. Diagnosing adult primary brain tumours: can we do better? Br J Gen Pract. 2017 Jun;67(659):278-279. doi: 10.3399/bjgp17X691277. PMID: 28546414; PMCID: PMC5442949. https://www.mayoclinic.org/diseases-conditions/astrocytoma/survival-rates/gnc-20591685 https://www.abta.org/tumor_types/glioblastoma-gbm/#:~:text=Prognosis,ages%2040+):%205.6%25* https://www.cancer.gov/rare-brain-spine-tumor/tumors/medulloblastoma https://my.clevelandclinic.org/health/diseases/23147-ependymoma https://www.mayoclinic.org/diseases-conditions/brain-tumor/diagnosis-treatment/drc-20350088 Kim H, Park KU. Clinical Circulating Tumor DNA Testing for Precision Oncology. Cancer Res Treat. 2023 Apr;55(2):351-366. doi: 10.4143/crt.2022.1026. PMID: 36915242; PMCID: PMC10101787.