STAT428 Final Project

Breast Cancer Data Analysis

Xiaodan Zhang

Apr. 17th, 2013

http://www.zazzle.com/pink_ribbon_with_rose_breast_cancer_awareness_invitation-161809804078377799

What is Breast Cancer?

- "Breast Cancer is a type of cancer originating from breast issue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk."
- The **pink ribbon** is an international symbol of breast cancer awareness.

----By Wikipedia

- **Causes**: Inherited gene mutations;

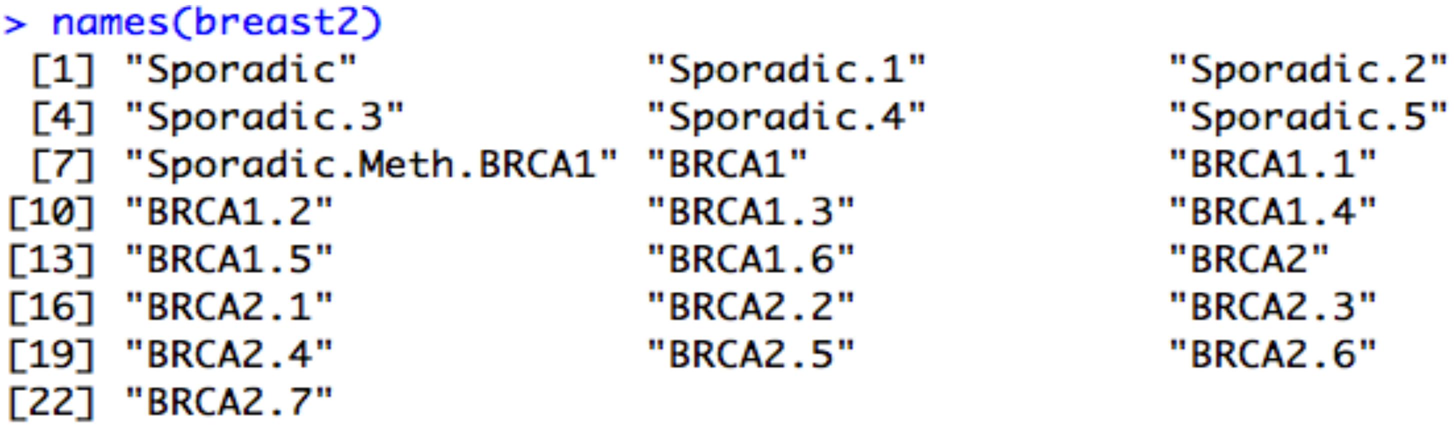
  Acquired gene mutations.

Project Goal

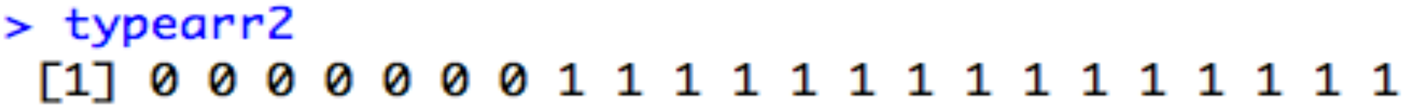
- Since **BRCA1** and **BRCA2** account for most cases of hereditary breast cancer in the United States and Europe, we aim to find which **genetic markers** play important roles in affecting these two inherited mutations.

First Step - Response array

- Rearrange the columns (3226x22 matrix)



```
> names(breast2)
 [1] "Sporadic"            "Sporadic.1"          "Sporadic.2"
 [4] "Sporadic.3"          "Sporadic.4"          "Sporadic.5"
 [7] "Sporadic.Meth.BRCA1" "BRCA1"               "BRCA1.1"
[10] "BRCA1.2"             "BRCA1.3"             "BRCA1.4"
[13] "BRCA1.5"             "BRCA1.6"             "BRCA2"
[16] "BRCA2.1"             "BRCA2.2"             "BRCA2.3"
[19] "BRCA2.4"             "BRCA2.5"             "BRCA2.6"
[22] "BRCA2.7"
```

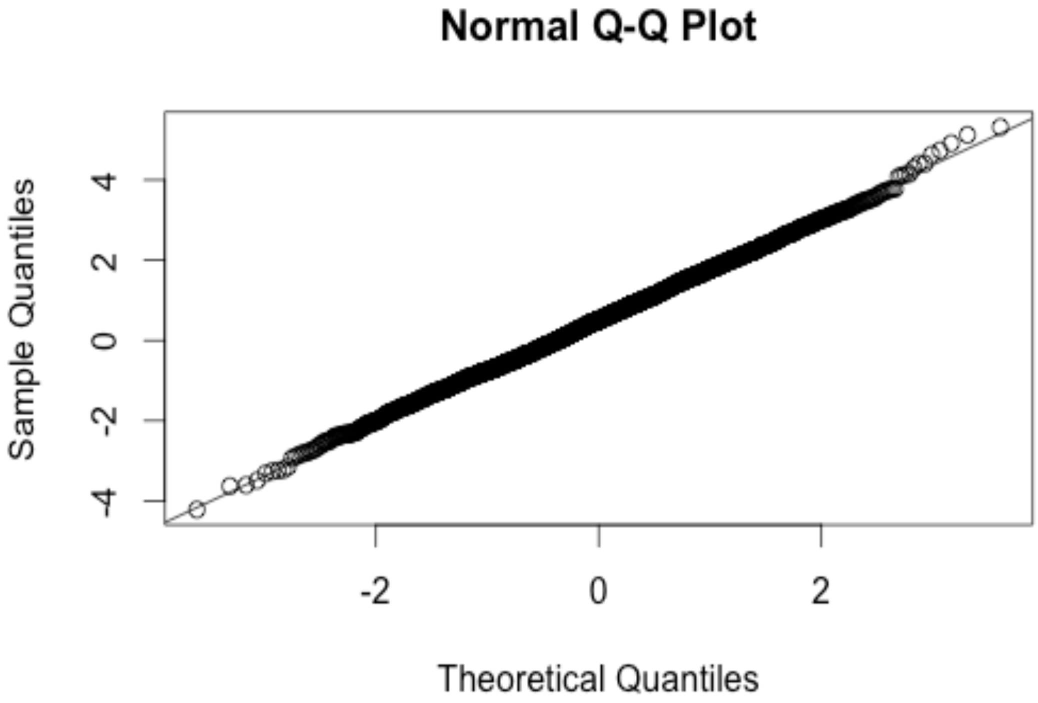


```
> typearr2
 [1] 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

- 0 means “healthy”; 1 means “BRCA”.

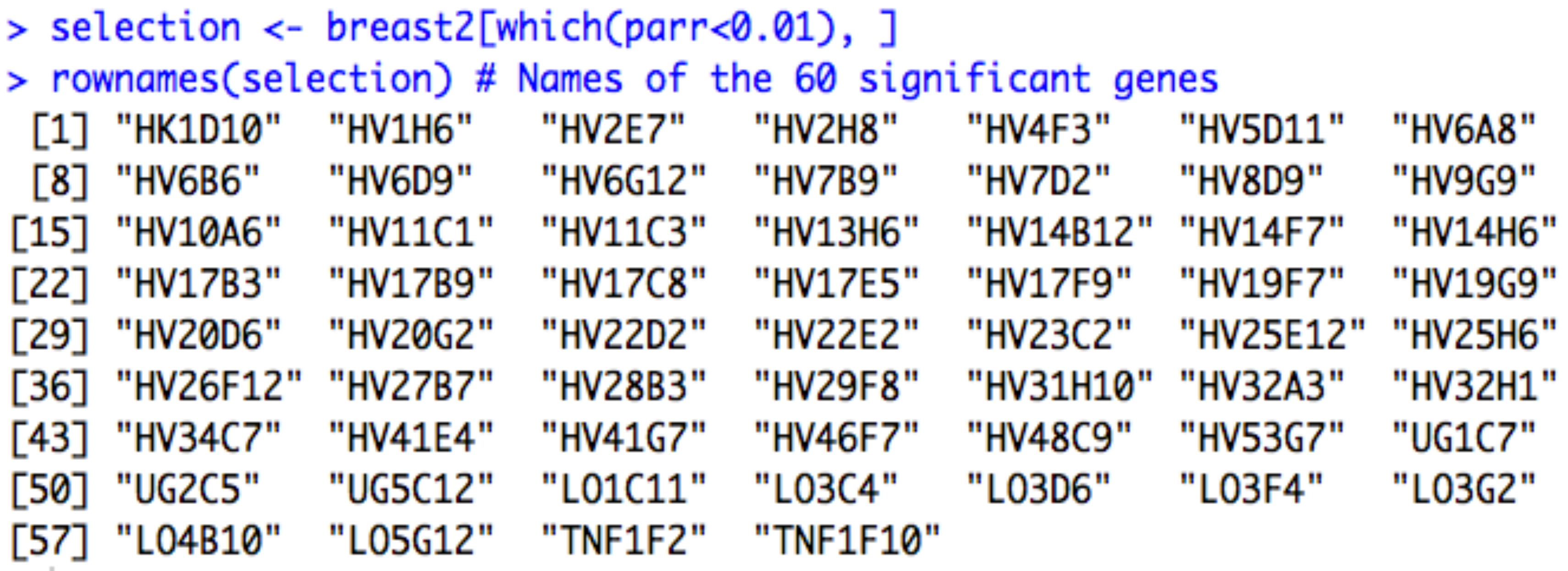
Second Step - Check Normality

- teststat2 <- mt.teststat (as.matrix(breast2), typearr2)
- qqnorm(teststat2)
- qqline(teststat2)



Third Step – T-tests

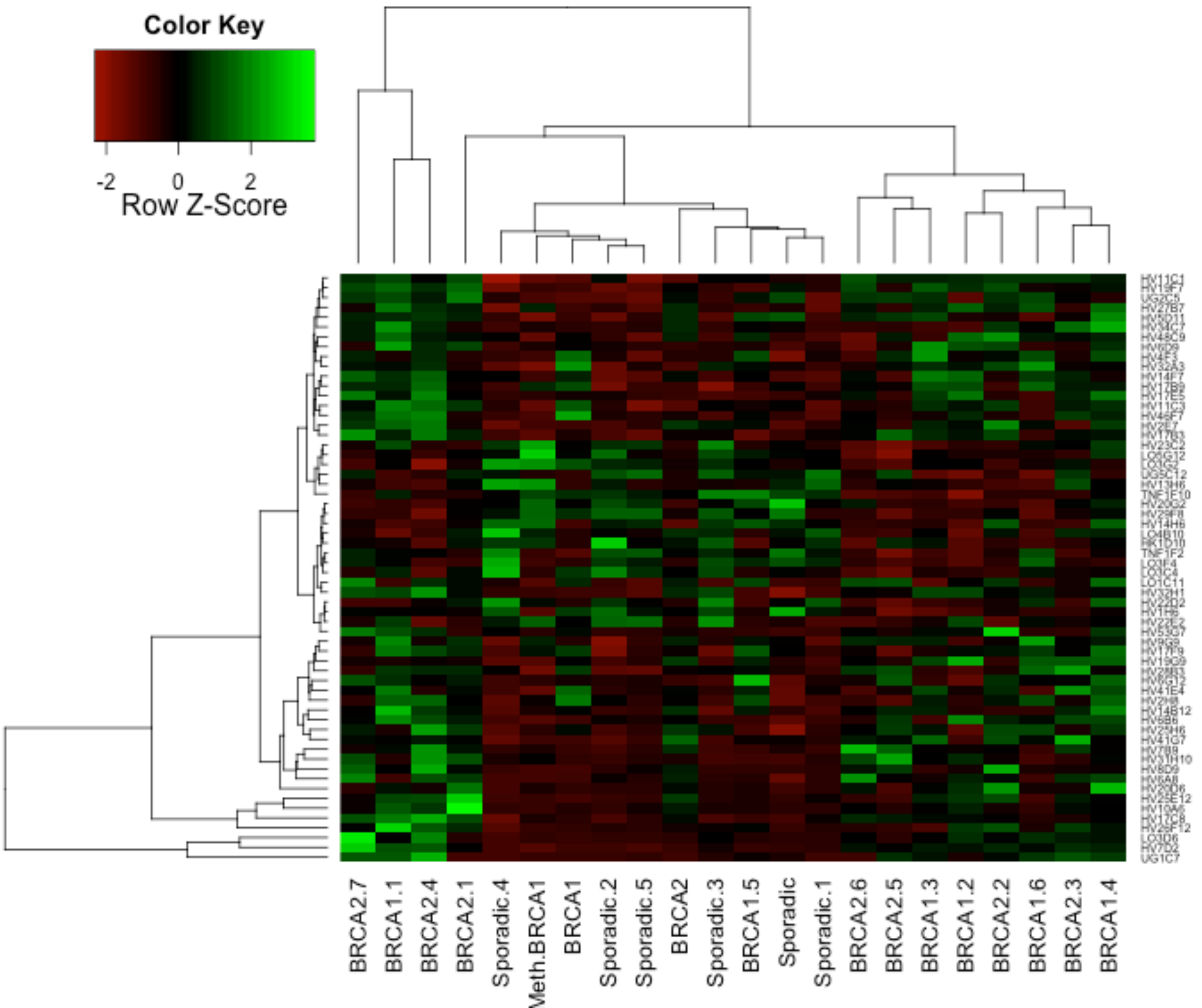
- Do **F-tests** first to check variances before performing T-tests.
- Two-sample T-test. ([1:7], [8:22])
- Use **0.01** as the significance level.



```
> selection <- breast2[which(parr<0.01), ]
> rownames(selection) # Names of the 60 significant genes
 [1] "HK1D10"  "HV1H6"   "HV2E7"   "HV2H8"   "HV4F3"   "HV5D11"  "HV6A8"
 [8] "HV6B6"   "HV6D9"   "HV6G12"  "HV7B9"   "HV7D2"   "HV8D9"   "HV9G9"
[15] "HV10A6"  "HV11C1"  "HV11C3"  "HV13H6"  "HV14B12" "HV14F7"  "HV14H6"
[22] "HV17B3"  "HV17B9"  "HV17C8"  "HV17E5"  "HV17F9"  "HV19F7"  "HV19G9"
[29] "HV20D6"  "HV20G2"  "HV22D2"  "HV22E2"  "HV23C2"  "HV25E12" "HV25H6"
[36] "HV26F12" "HV27B7"  "HV28B3"  "HV29F8"  "HV31H10" "HV32A3"  "HV32H1"
[43] "HV34C7"  "HV41E4"  "HV41G7"  "HV46F7"  "HV48C9"  "HV53G7"  "UG1C7"
[50] "UG2C5"   "UG5C12"  "LO1C11"  "LO3C4"   "LO3D6"   "LO3F4"   "LO3G2"
[57] "LO4B10"  "LO5G12"  "TNF1F2"  "TNF1F10"
```

Third Step – Heatmap

- Clear Clustering patterns.



Bio-functions

- HV41G7 - **Myotubularin Related Protein 4 (MTMR4)**

Gene type: protein coding

Loss of phosphatase activity in myotubularin-related protein 2 is associated with **Charcot-Marie-Tooth disease** type 4B1. CMT disease is a group of disorders passed down through families that affect the nerves outside the brain and spine.

http://charcotmarietoothdisease.org/charcot-marie-tooth-disease-information/charcot-marie-tooth-disease-basic-guide/

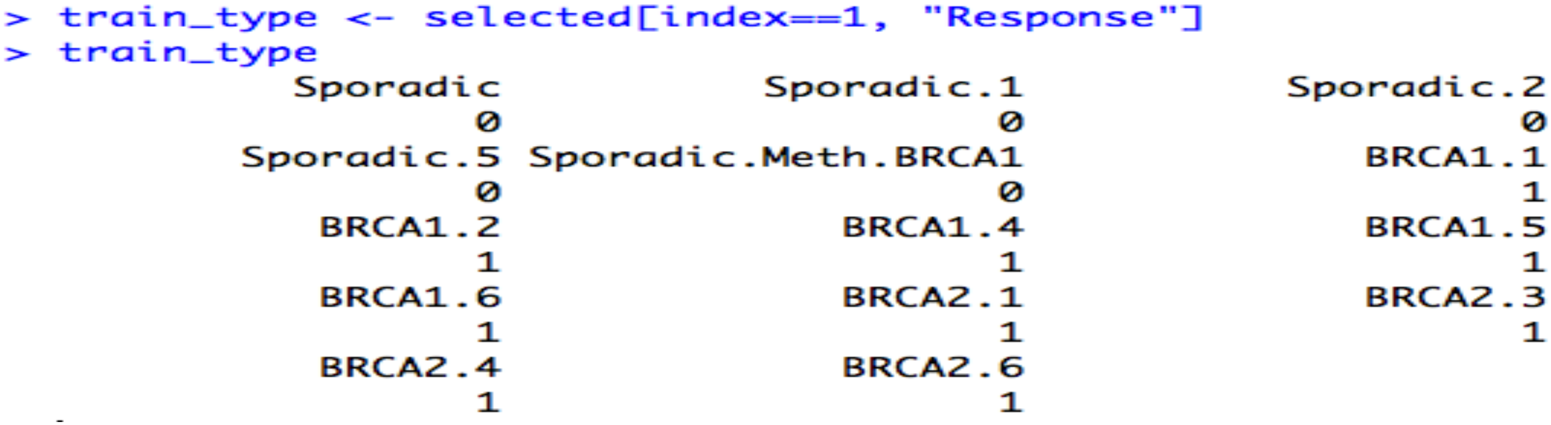
Bio-functions

- HV4F3 - **Transcription factor AP-2 gamma (activating enhancer-binding protein 2 gamma; TFAP2C)**
- **Gene type**: protein coding
- This encoded protein can act either a homodimer or heterodimer with other family members and is included during retinoic acid-mediated differentiation. It plays a role in the development of the eyes, face, body wall, limbs, and neural tube.

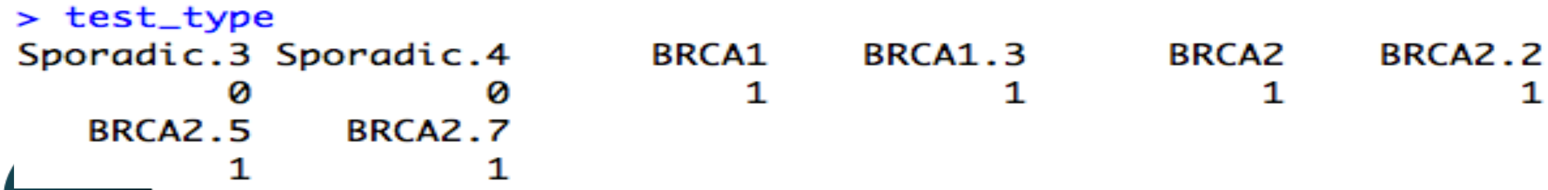
http://www.ncbi.nlm.nih.gov/gene/7022

Predict Breast Cancer with Selected Markers

- Use LDA to train the training data and test on the testing data.



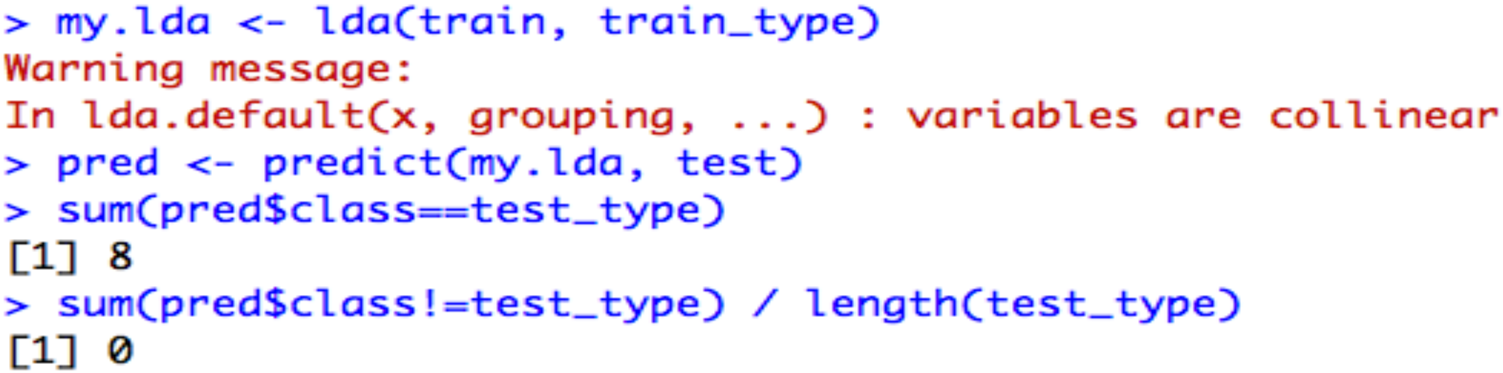
```
> train_type <- selected[index==1, "Response"]
> train_type
          Sporadic         Sporadic.1         Sporadic.2
                 0                  0                  0
        Sporadic.5 Sporadic.Meth.BRCA1            BRCA1.1
                 0                  0                  1
           BRCA1.2            BRCA1.4            BRCA1.5
                 1                  1                  1
           BRCA1.6            BRCA2.1            BRCA2.3
                 1                  1                  1
           BRCA2.4            BRCA2.6
                 1                  1
```



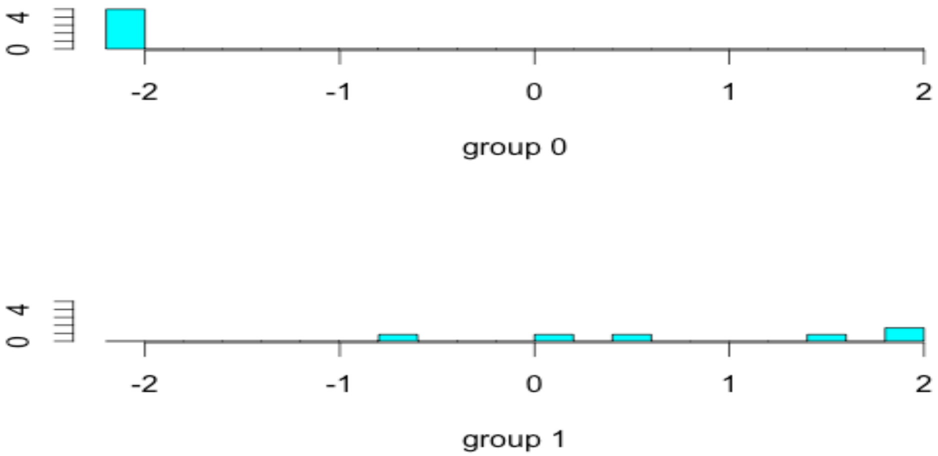
```
> test_type
Sporadic.3 Sporadic.4      BRCA1    BRCA1.3      BRCA2    BRCA2.2
         0          0          1          1          1          1
   BRCA2.5    BRCA2.7
         1          1
```

Predict Breast Cancer with Selected Markers

- Use LDA to train the training data and test on the testing data.



```
> my.lda <- lda(train, train_type)
Warning message:
In lda.default(x, grouping, ...) : variables are collinear
> pred <- predict(my.lda, test)
> sum(pred$class==test_type)
[1] 8
> sum(pred$class!=test_type) / length(test_type)
[1] 0
```



Limitations of the Model

- 1. For T-statistics, it may be affected by small or unstable variances.
- 2. For LDA, it implicitly assumes Gaussian distribution of data.
- 3. For LDA, it implicitly assumes the mean as the discriminating factor, not variance.
- 4. For LDA, it may overfit or underfit the data.

Thank you very much!