

# Statistical Analysis of Brain & Gene

## **Group-1**

Songfen Lu, Ziwei Li  
Dhanu Vardhan Singh Jhala

# Introduction

Aging, Dementia and Traumatic Brain Injury (TBI) Study

<b>Num of Samples</b>	377 (collected from 4 different brain regions with respect to 107 donors)	
<b>Age</b>	89.0+/-6.3	
<b>Gender</b>	F: 44	M: 63
<b>IHC + Protein Quantification</b>	32	
<b>Num of Gene expressions (Unnormalized Gene-level FPKM)</b>	50280	

# Pre-processing

## Dataset with all features

- 377 \* 50320

## NA check and replace

- only 1251 data are NAs
- replace NA with 0

## Recode factor variables

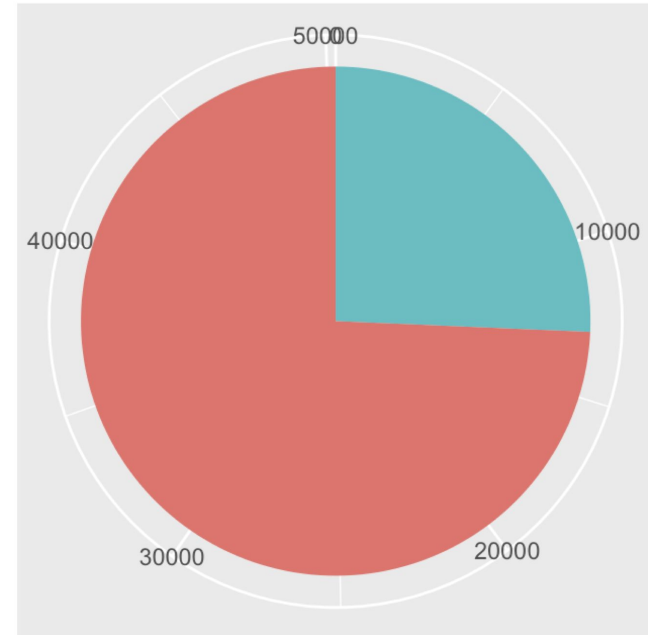
- convert into numeric

## Remove useless columns

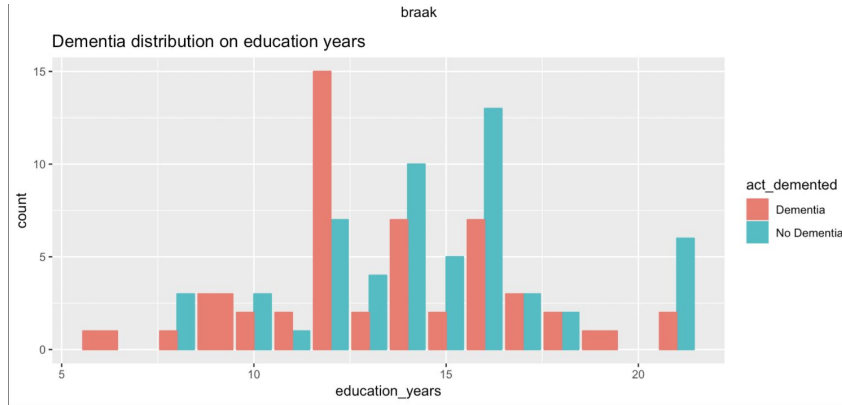
- define 80%-zero columns as useless

## Percentage of 80% zero columns

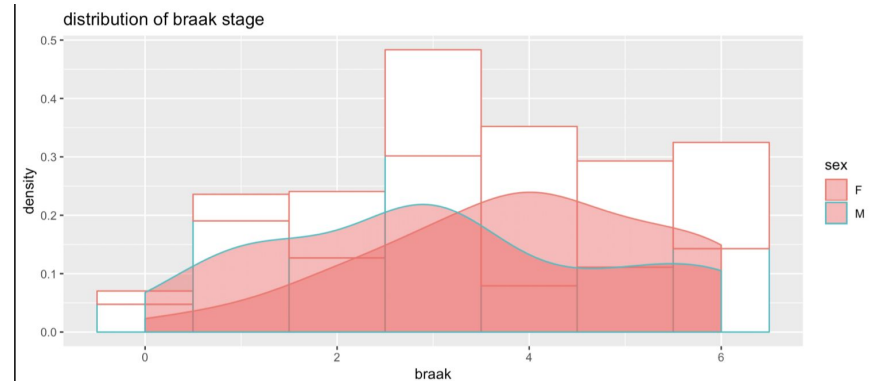
■ > 80%(25.67%) ■ <= 80%(74.33%)



# Interesting Statistical Overview



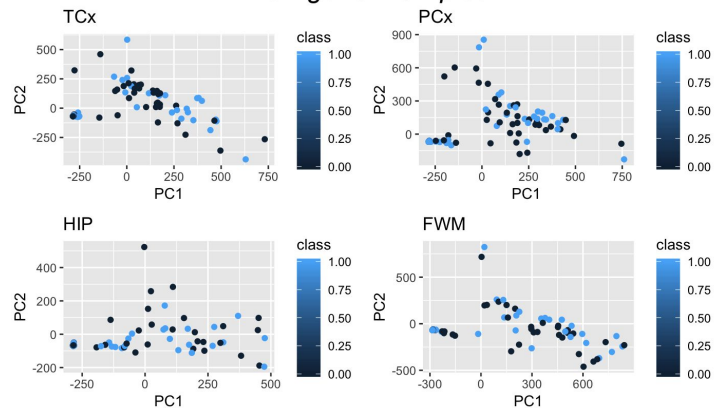
- Education years has a weak relationship with Dementia.
- Female are more likely to be diagnosed as higher stage.



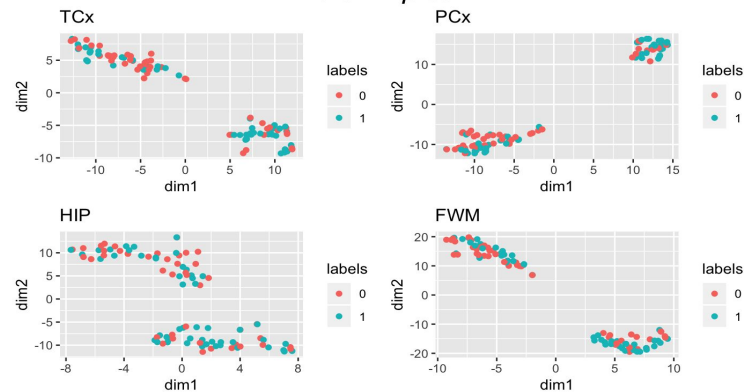
# PCA vs t-SNE on Dementia



Non-gene PCA plot



t-SNE plot





# Dementia Classification

# Dementia Classification

## ***Feature Engineering:***

*Feature Selection by T Test:* T test based pre processing was done. Thereafter we selected the top 10000 features. But is T test robust?

Lasso Regression & Random Forest: we applied lasso regression for feature selection. Random Forest based feature selection was also done to take top 10 features. This was to compare the two approaches.

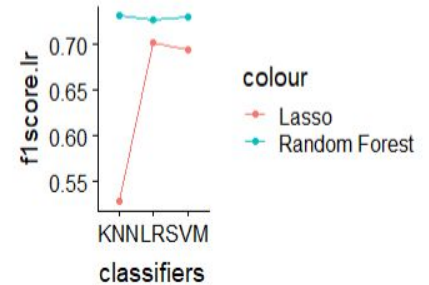
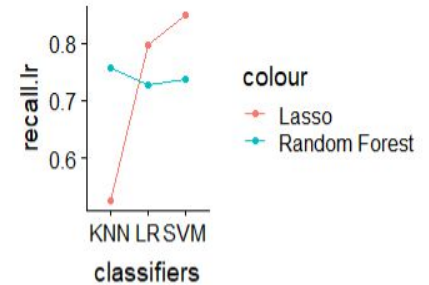
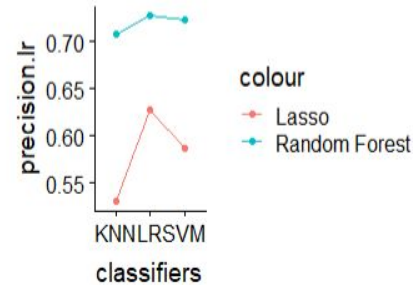
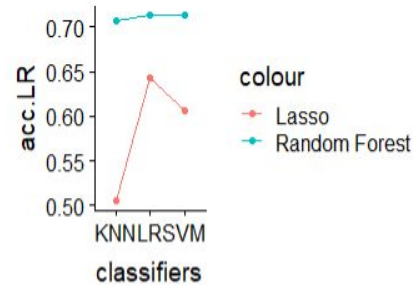
## ***Model Development:***

K Nearest Neighbor, Logistic Regression and Support Vector Machine were implemented. Comparative Analysis done.

# Results

Results for Classifier				
	Accuracy	Recall	Precision	Fscore
<b>KNN All features (benchmark)</b>	0.553	0.545	0.58	0.562
<b>Lasso features</b>				
KNN	0.505	0.525	0.536	0.527
Logistic Regression	0.6436	0.797	0.626	0.702
SVM	0.606	0.848	0.587	0.694
<b>Random Forest features</b>				
KNN	0.707	0.757	0.707	0.731
Logistic Regression	0.712	0.727	0.727	0.727
SVM	0.712	0.737	0.722	0.73

- RF feature selection outperforms Lasso.
- LR and SVM performs similar with RF

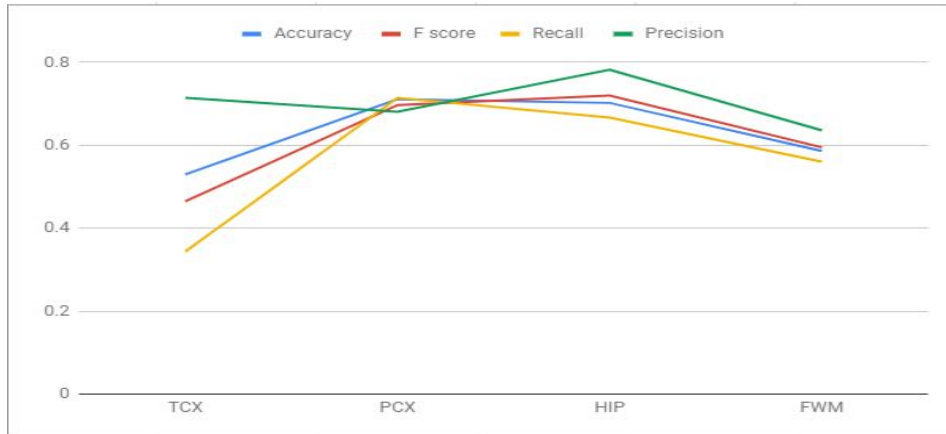




# Non Gene Features & Brain Region

Analysis done on non gene features: features “ihc\_at8\_ffpe” and “ihc\_tau2\_ffpe” are the most important non gene features. Furthermore, education years is the only donor information making significant contribution in predicting dementia.

Information collected from brain region HIP and PCX are most significant in predicting dementia.



	Accuracy	Recall	Precision	F score
Brain Region				
TCX	0.53	0.344	0.714	0.465
PCX	0.711	0.714	0.681	0.697
HIP	0.702	0.667	0.782	0.72
FWM	0.586	0.56	0.636	0.595


# Conclusion

- Gene information plays an important role then pathological or donor information.
- Logistic regression and support vector machine performs well. However KNN might not be a robust classifier.
- Information collected from Hippocampus and Parietal brain region plays important role compared to temporal neocortex and white matter of forebrain.
- Future work: Feature selection can be improved as features selected from lasso regression and Random Forest does not have much overlap. Furthermore, linear combination of features selected from two methods to experiment about the predictiveness of dementia.

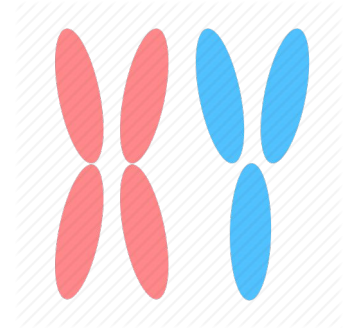
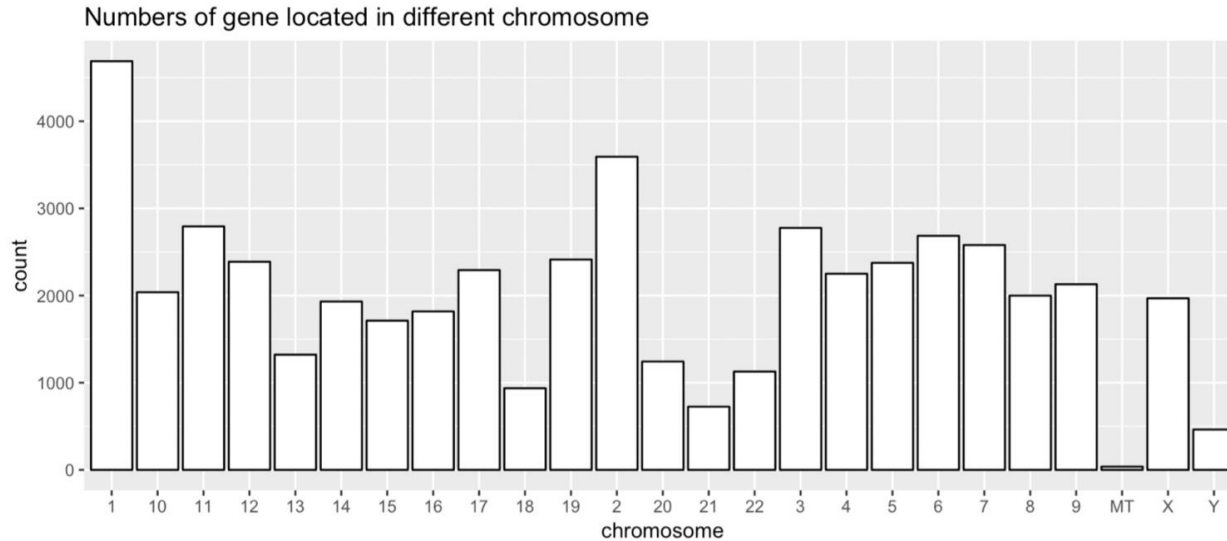


# Additional Analysis on gene and chromosomes

## Problems:

- Is sex dependent to gene expressions?
  - Is sexual related genes able to differentiate brain regions?
- 

# Additional Analysis on Chromosomes



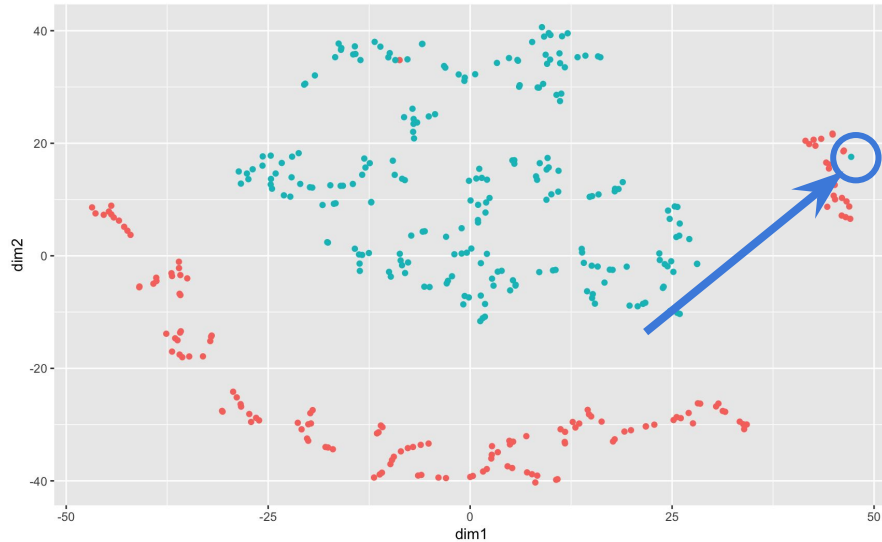
# Classification of gender

- using top features selected from p-value
- Female as Positive

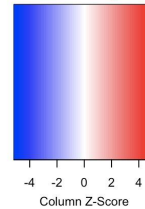
	<b>Classifier</b>	<b>Accuracy</b>	<b>Sensitivity</b>	<b>Specificity</b>	<b>F1-score</b>
Benchmark(all genes)	Knn	58.51%	41.56%	70.27%	52.62%
Top-30 genes on Sexual chromosome	Knn	100%	100%	100%	100%
	SVM(linear)	99.47%	100%	99.10%	99.54%
	SVM(sigmoid)	100%	100%	100%	100%
Top-30 genes on Non-Sexual chromosome	Knn	48.4%	28.57%	62.16%	45.37%
	SVM(linear)	84.57%	80.52%	87.39%	83.95%
	SVM(sigmoid)	85.64%	77.92%	90.99%	84.46%

# Visualization on gender

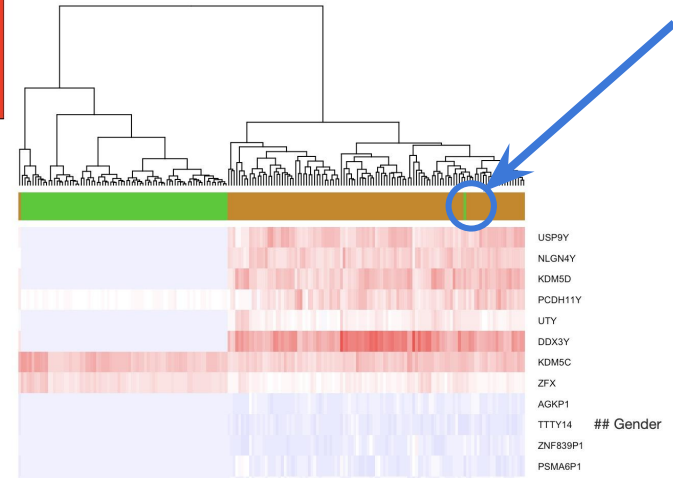
T-sne of 30 gene expressions  
on X,Y chromosomes



Color Key



Clustering by top 30 filtered features

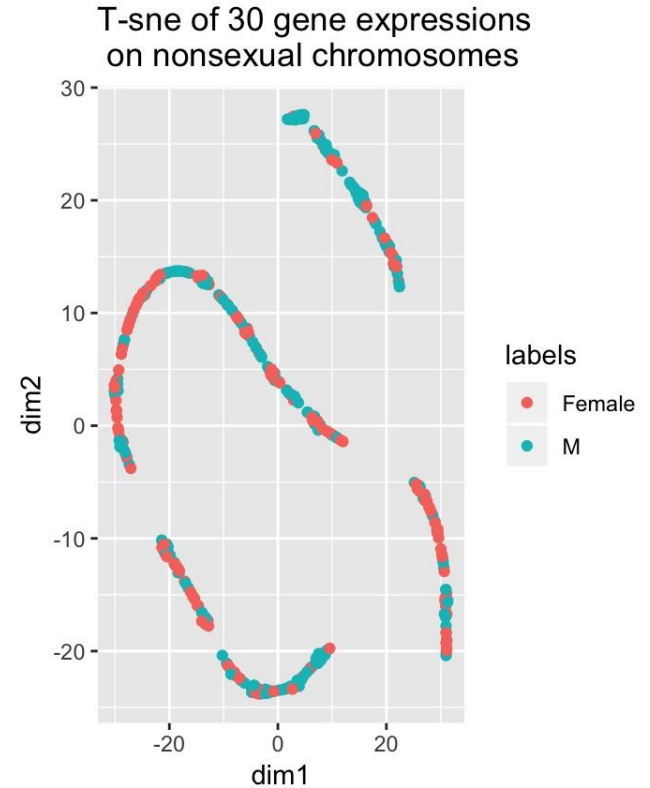
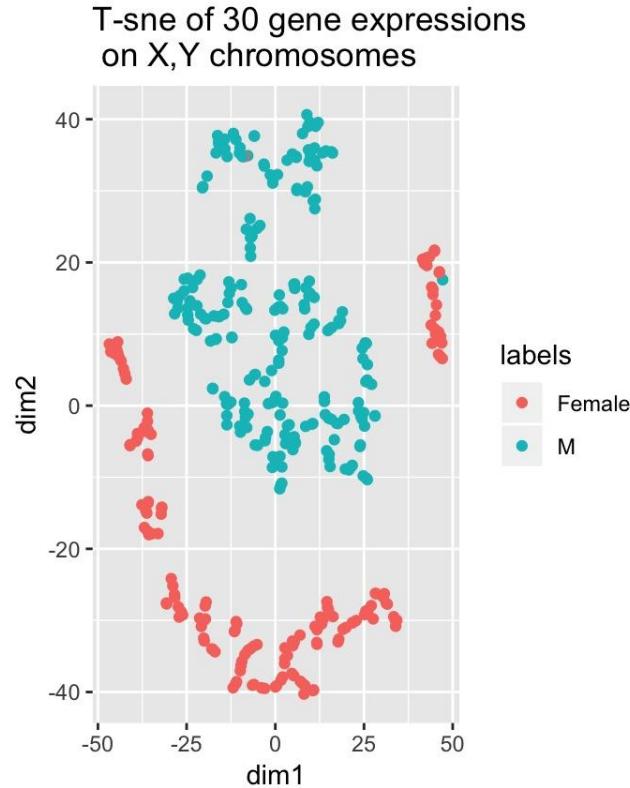


- can be used to differentiate the sex of observations
- do not follow a linear pattern strictly

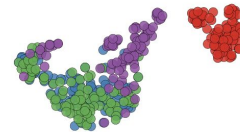
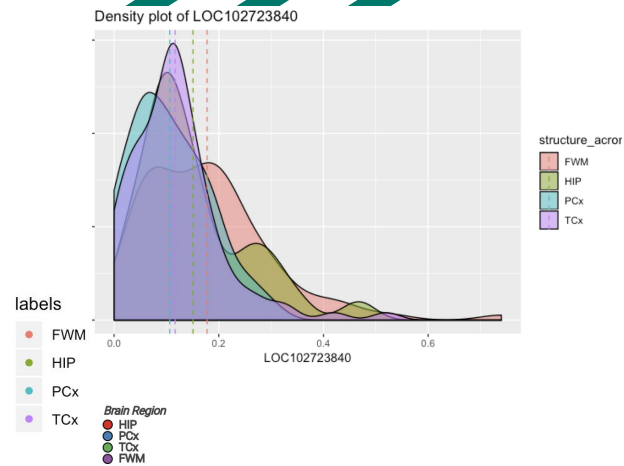
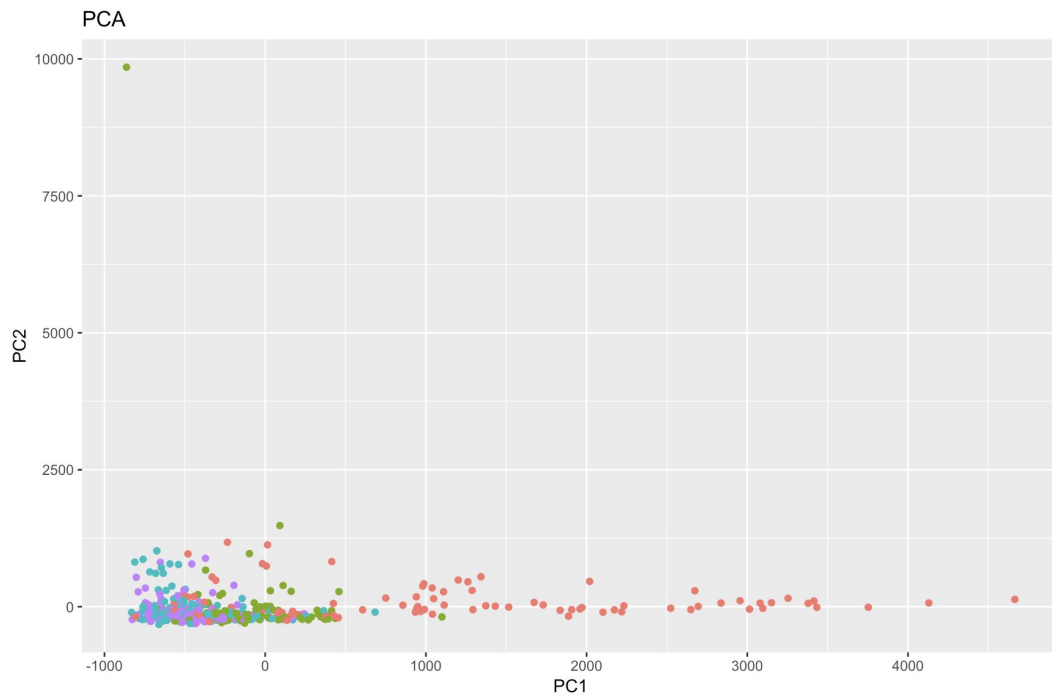
# Sexual vs Non-sexual

The left graph

- displays a wide range of overlapping
- dots follow exactly the same pattern
- can be clustered into 4 groups

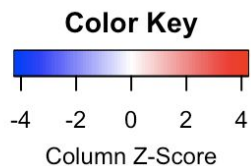


# PCA on sexual chromosome genes

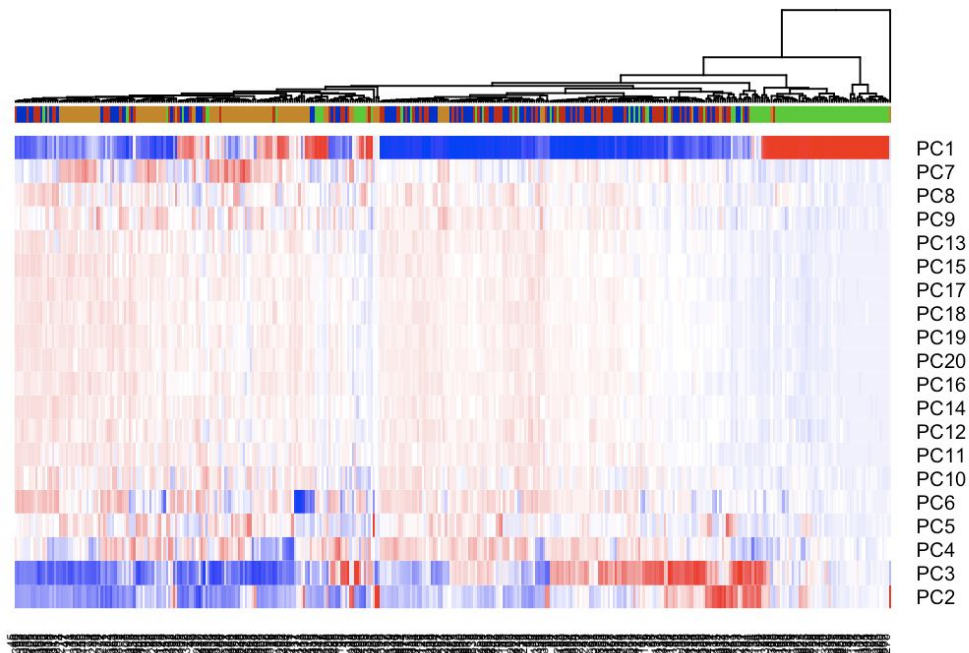




# Brain region clustering



**Clustering Brain Region according to PCA result**



# Conclusion

- ❑ Gene expressions are significant predictors for gender.
- ❑ In particular, the genes on X and Y chromosomes can exactly tell the sex information for a human.
- ❑ Genes located on other chromosomes also contains sex information but it may be not accurate, probably affected by other body mystery.
- ❑ We can separate samples from HIP and FWM according to X and Y chromosomes located genes whereas the other two cortical regions overlap(PCx and TCx).



# Thanks :-)

Any questions?

