**Introduction and Aims**

Bipolar Disorder causes a patient to fluctuate between manic and depressive states (<https://www.nimh.nih.gov/health/topics/bipolar-disorder/index.shtml#:~:text=Bipolar%20disorder%20(formerly%20called%20manic,three%20types%20of%20bipolar%20disorder)>.

Analysing gene expression data, such as “RNA sequencing data from peripheral whole blood”, might provide more insights into the causes and different development paths of Bipolar Disorder, which fills a gap in our current understanding (Whole blood transcriptome analysis in bipolar disorder reveals strong lithium effect).

This project aims to discover subtypes of patients and their biological significance.

**Work done so far, methods, and results**

F-test feature selection:

* Collaborated with fellow FYP student Zeng Yanxi
* To select genes that separate patients from controls well
* FDR-adjusted p-value <1%

PCA:

* Dimensionality reduction
* 83% of total variance

Classification:

* Logistic Regression, with 5-fold cross-validation:
* With PCA: Accuracy of about 83%
* Without PCA: Accuracy of about 51%
* Support Vector Machine with Linear Kernel, with 5-fold cross-validation:
* With PCA: Accuracy of about 85%
* Without PCA: Accuracy of about 85%
* Support Vector Machine with Radial Basis Function Kernel, with 5-fold cross-validation:
* With PCA: Accuracy of about 82%
* Without PCA: Accuracy of about 78%

Clustering:

* K-means clustering
* Optimal number of 3 clusters/subtypes determined using the within-cluster sum of square method

Gene Set Enrichment Analysis:

* What is GSEA?
* Results?

**Future work to complete the project**