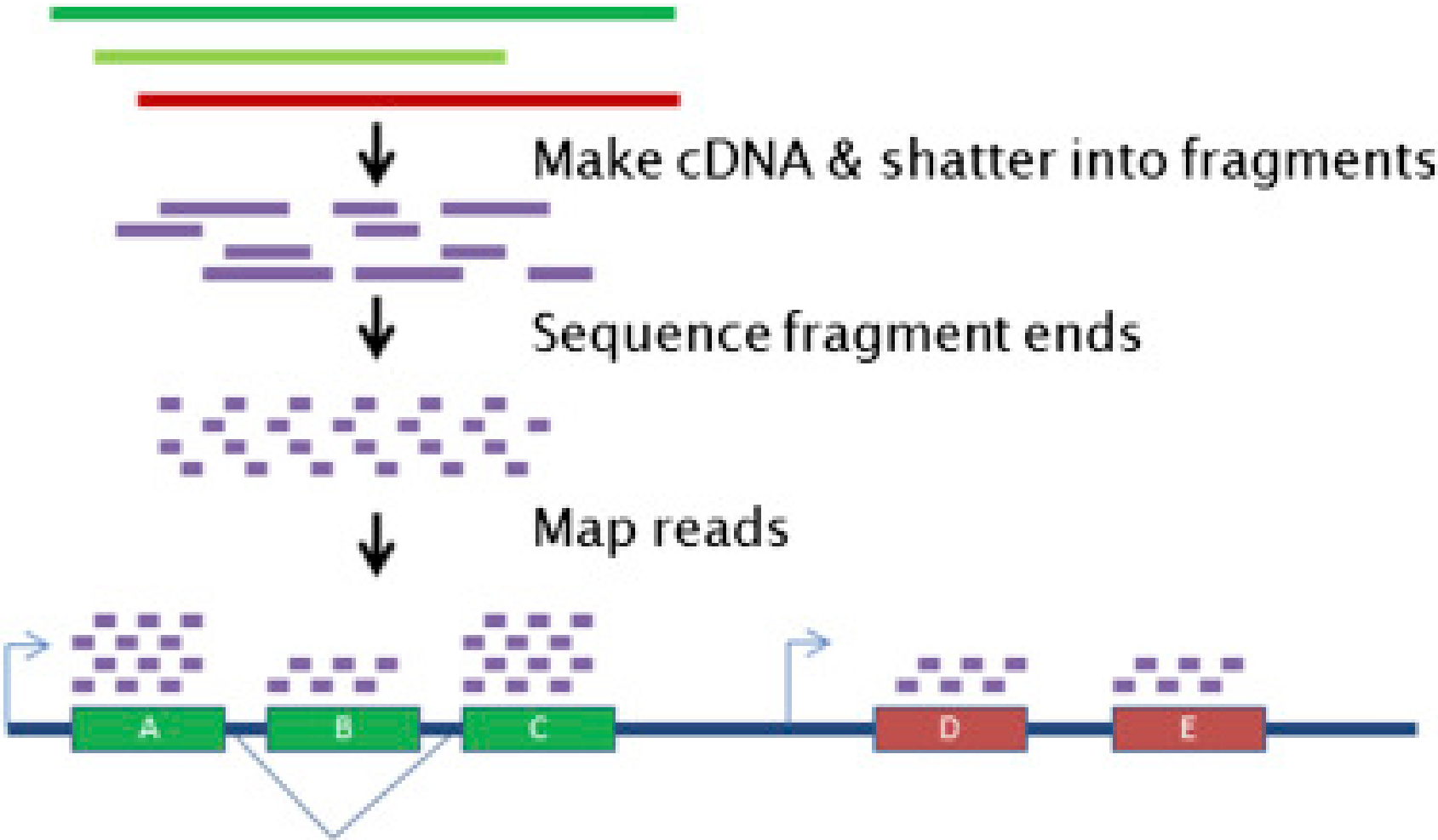
TCGA Report

Chen, Pin-Jui

Sep 26, 2025

Read and Count



RPKM, FPKM, and TPM

RPKM and FPKM formula :

$$\frac{total\ exon\ reads}{mapped\ reads\ (millions)\ *\ exon\ length(KB)}$$

TPM formula :

$$\frac{\frac{total\ exon\ reads}{exon\ length(KB)}}{\sum_i \frac{Gene_i\ mapped\ reads(millions)}{exon\ length(KB)}}$$

In the files GAV2.data and HiSeqV2.data, the data were converted using the $log_2(RPKM + 1)$ transformation.

Highly Variable Genes (HVG)

1. **vst**
   1. Fits a line to the relationship of log(variance) and log(mean) using **loess**.
   2. Standardizes the feature values using the observed mean and expected variance
   3. Clipping to a maximum and calculating feature variance.
2. **mean.var.plot**
   1. Calculates $log(\frac{variance}{mean})$
   2. Divides features into bins based on their average expression.
   3. Calculates z-scores for dispersion within each bin.
3. **dispersion**
   1. Selects the genes with the highest dispersion values.

Differential Expression Analysis (DEA)

limma: Linear Models for Microarray Data

- **Purpose:** Identify differentially expressed genes in high-throughput data.
- **Original Use:** Designed for microarray data.
- **Now:** Also applied to RNA-seq data (with voom transformation).
- **Key Ideas:**
  - Uses **linear models** to estimate expression differences.
  - Applies **weighted least squares** to account for heteroscedasticity.
  - Uses **Empirical Bayes** to stabilize variance estimates and control false positives.
- **Output:** log2 fold change, p-value, and adjusted p-value (FDR).

Comparison of Gene Selection Methods

Objective

Compare three gene selection strategies on PCA and differential expression:

- All genes
- Highly Variable Genes (HVG)
- Sparse PCA-selected genes

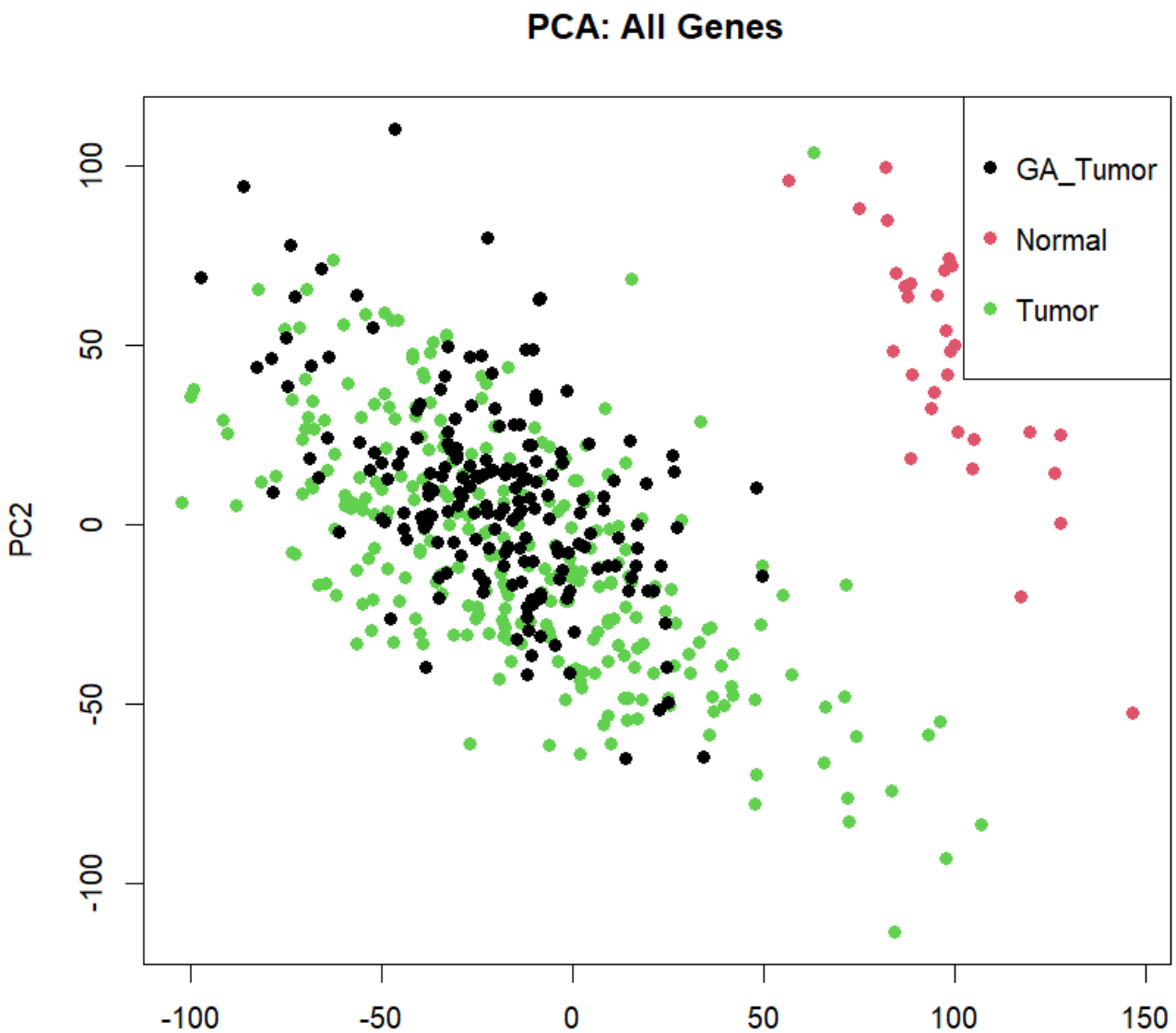
We perform PCA on HiSeq data, and then fit GA data onto the resulting PCA space.

Comparison of Gene Selection Methods

Workflow

1. **Differential Expression Analysis:**
   - Limma on each gene set
   - Obtain log2 fold-change and adjusted p-values
2. **Visualization:**
   - Volcano plots for each gene set
   - PCA plots using selected genes
3. **Overlap Analysis:**
   - Count overlapping DE genes between methods
   - Optional: Venn diagram

All gene PCA


PCA: All Genes
PC2
100
50
0
-50
-100
-100
-50
0
50
100
150
GA_Tumor
Normal
Tumor

All gene + DEA

Figure 1: DEA Volcano 4806

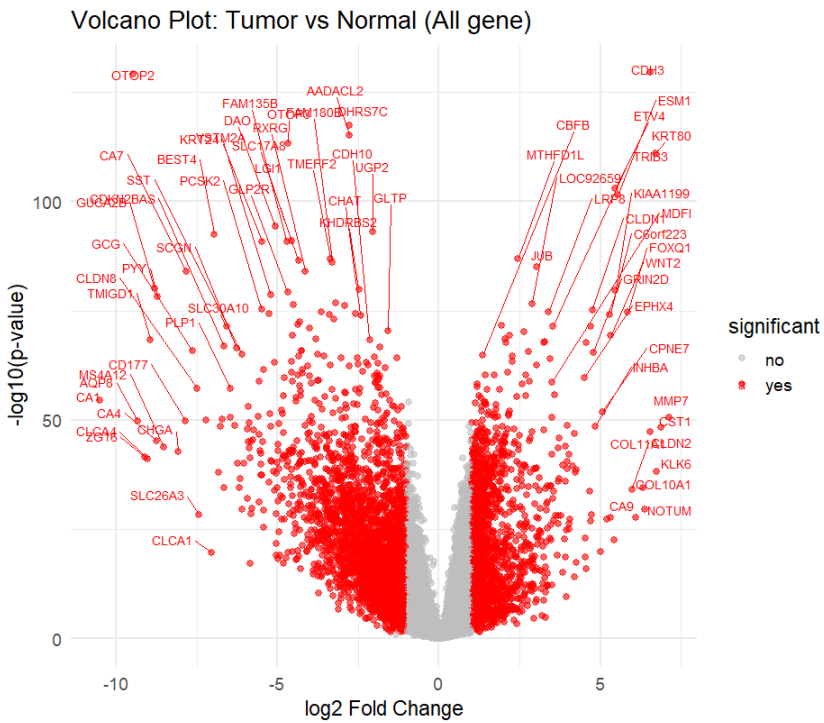
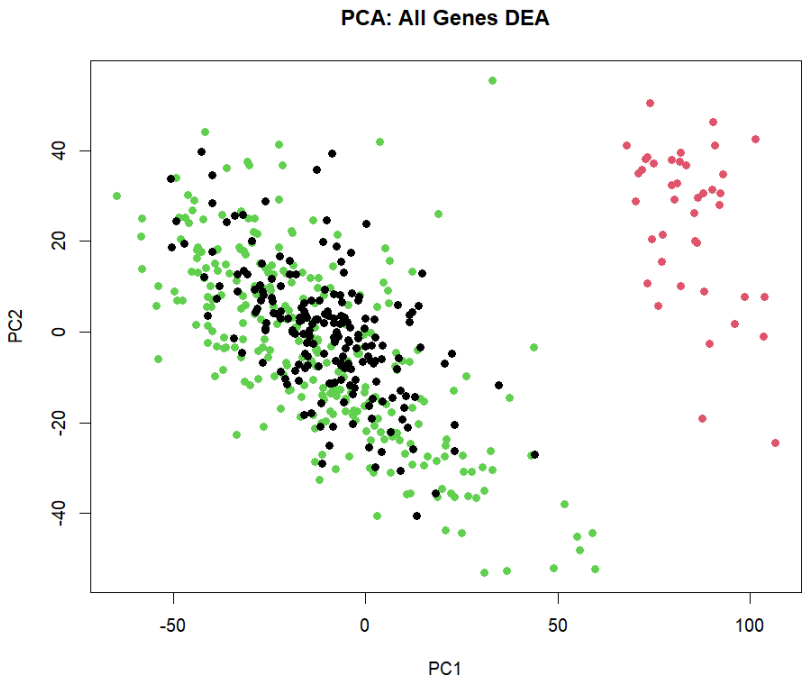
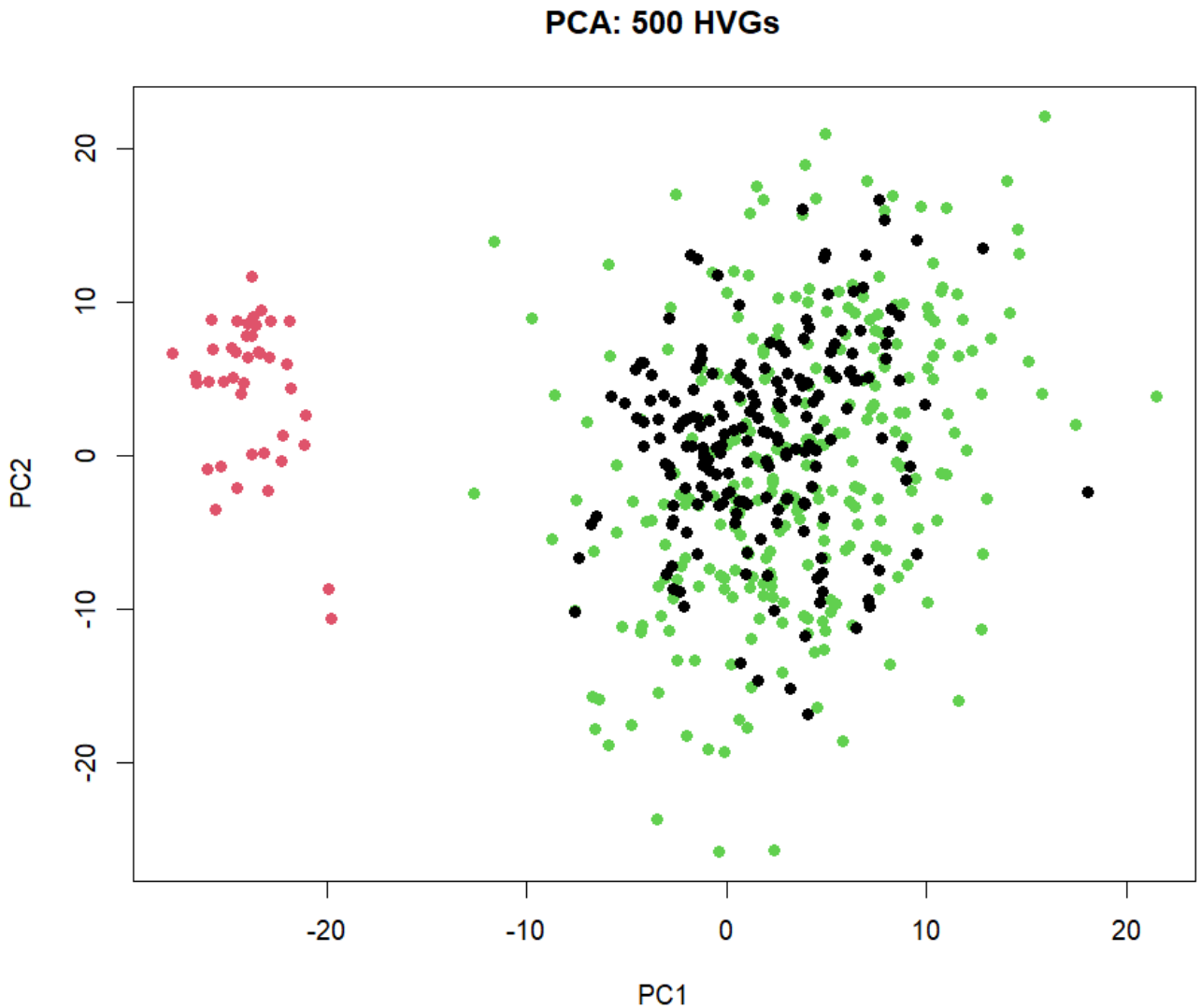


Figure 2: DEA PCA 4806



500 HVG PCA



HVG + DEA

Figure 3: DEA Volcano (369/500)

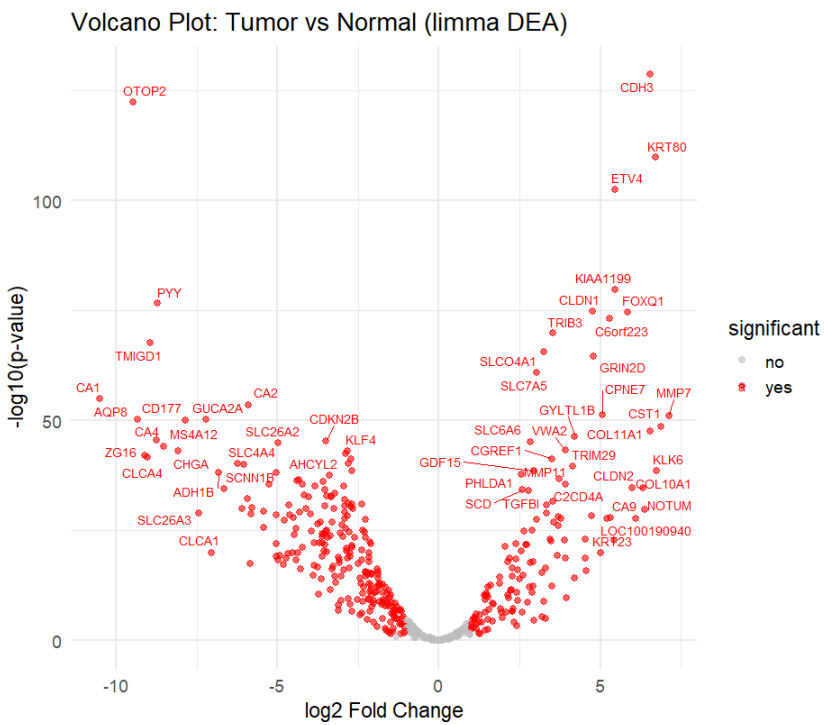
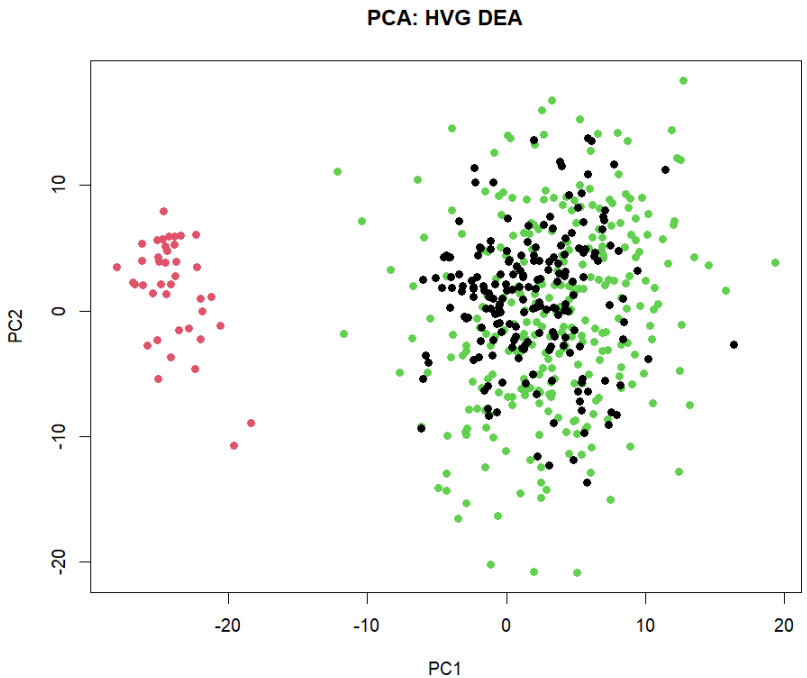
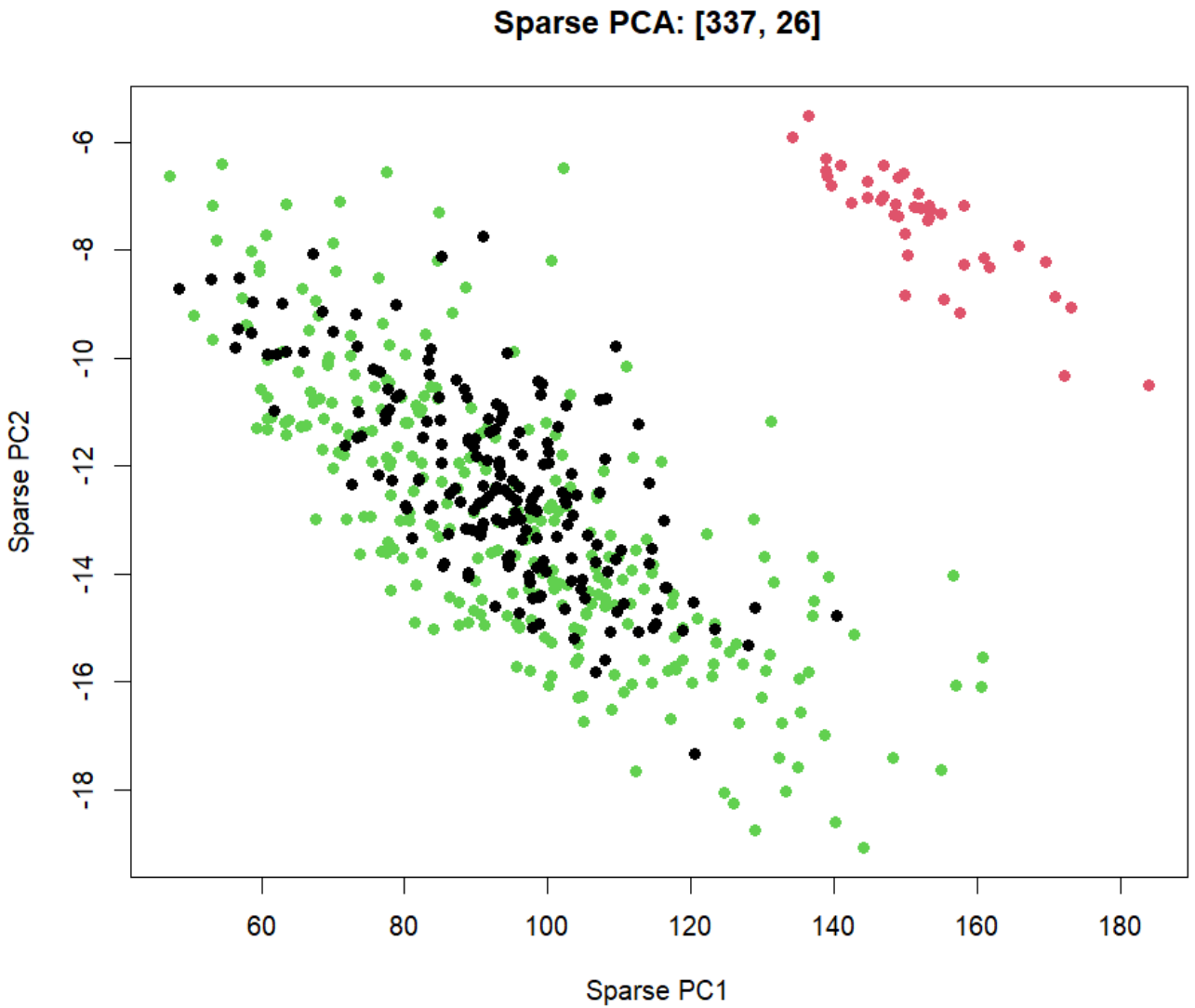


Figure 4: DEA PCA



Sparse PCA


Sparse PCA: [337, 26]
Sparse PC2
-6
-8
-10
-12
-14
-16
-18
60
80
100
120
140
160
180
Sparse PC1

SPCA + DEA

Figure 5: DEA Volcano (350/363)

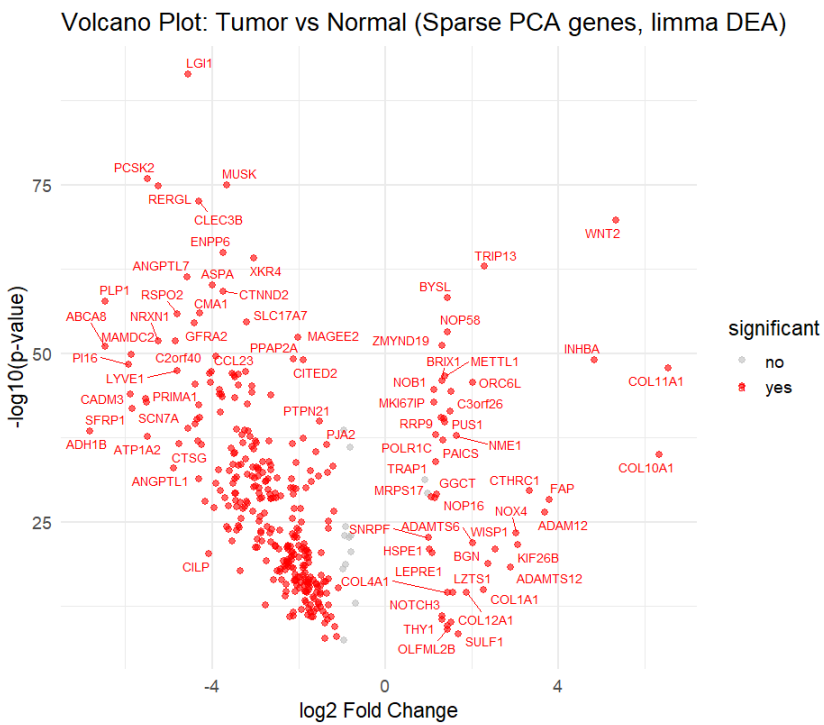
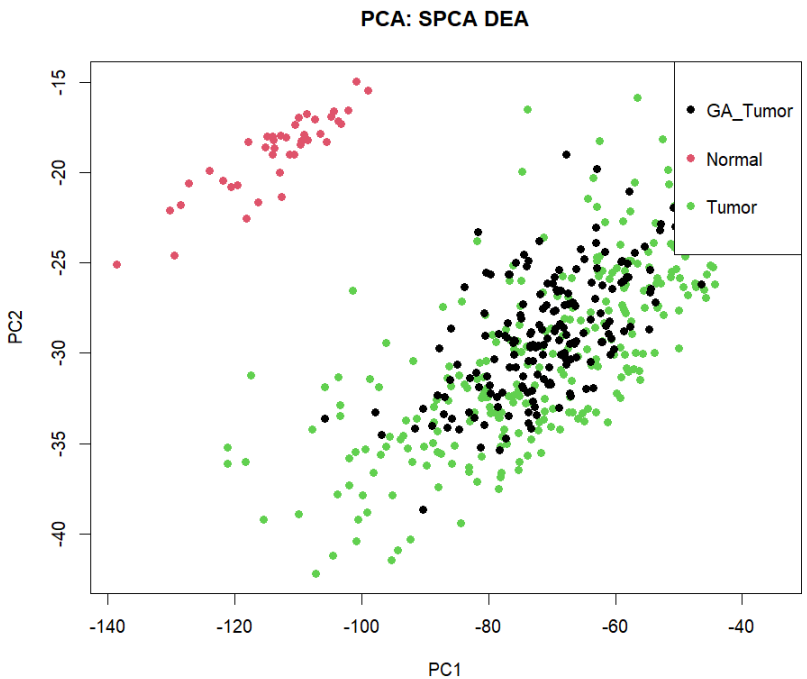
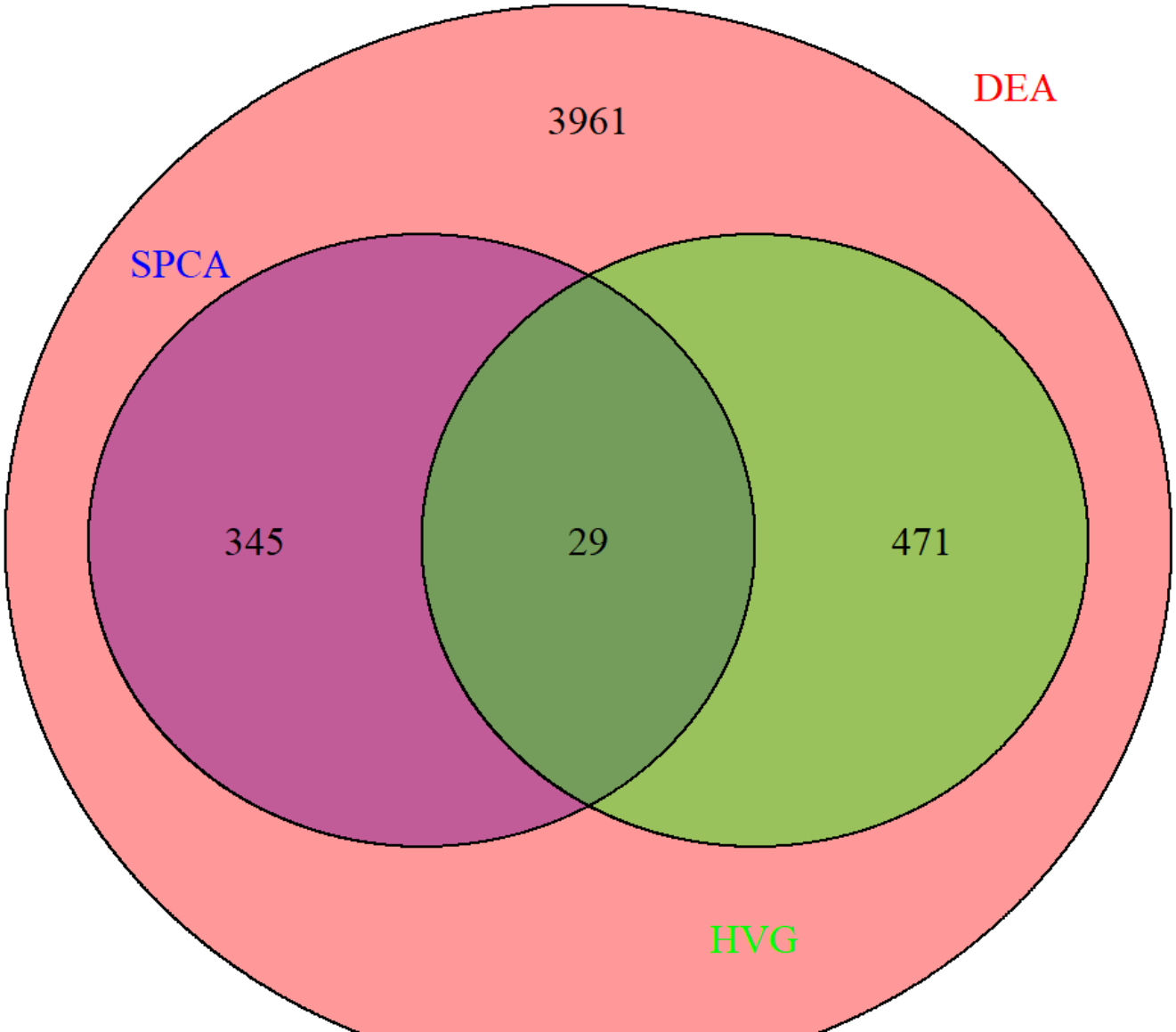


Figure 6: DEA PCA [208,22]

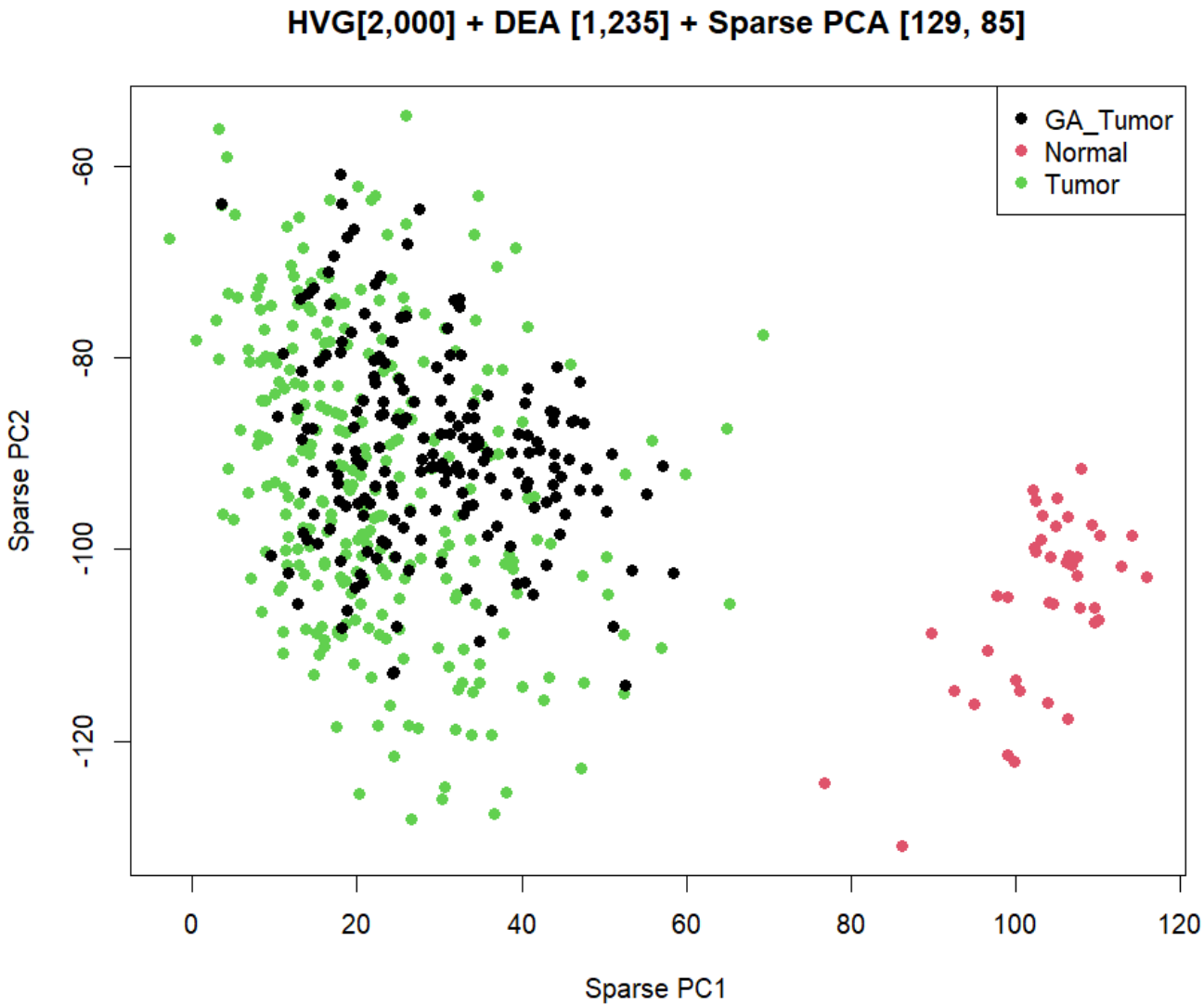


1. All gene PCA vs DEA cosine similarity :
   PC1 = 0.997 ; PC2 = 0.953
2. HVG vs HVG + DEA cosine similarity :
   PC1 = -0.977 ; PC2 = 0.919
3. SPCA vs SPCA + DEA cosine similarity :
   PC1 = -0.944 ; PC2 = 0.924

Intersection DEA, HVG, SPCA



HVG + DEA + SPCA


HVG[2,000] + DEA [1,235] + Sparse PCA [129, 85]
GA_Tumor
Normal
Tumor
Sparse PC2
-60
-80
-100
-120
0
20
40
60
80
100
120
Sparse PC1