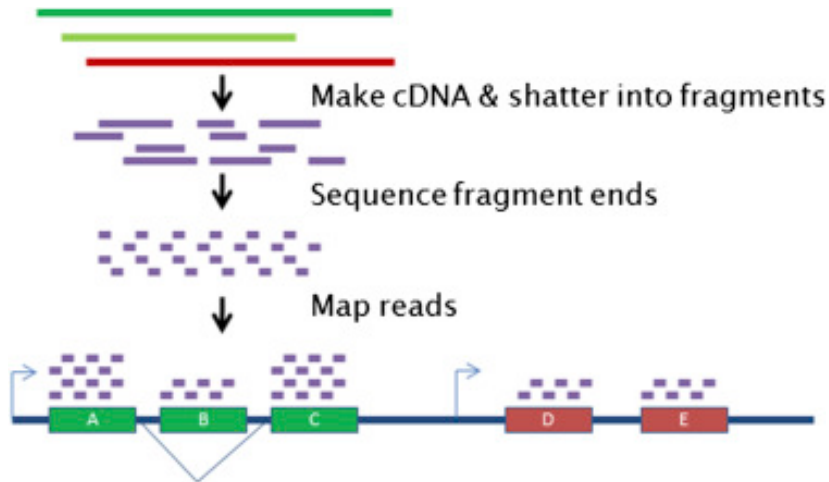


TCGA Report

Chen, Pin-Jui

Sep 26, 2025

Read and Count



RPKM, FPKM, and TPM

RPKM and FPKM formula :

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

TPM formula :

$$\sum_i \frac{\frac{\text{total exon reads}}{\text{exon length(KB)}}}{\text{Gene}_i \text{ mapped reads(millions)}}$$

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 $\log(\text{variance})$ and $\log(\text{mean})$ 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 $\log\left(\frac{\text{variance}}{\text{mean}}\right)$
- 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:

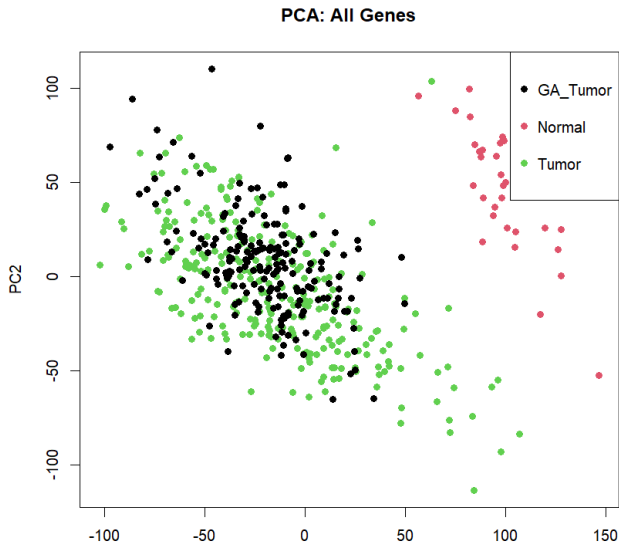
- Count overlapping DE genes between methods
- Optional: Venn diagram

Comparison of Gene Selection Methods

Expected Outcome

- PCA separation of groups with different gene sets
- Volcano plots showing significant genes
- Table / Venn diagram showing gene overlap

All gene PCA



All gene + DEA

Figure 1: DEA Volcano 4806

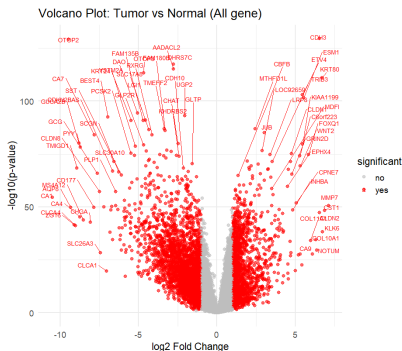
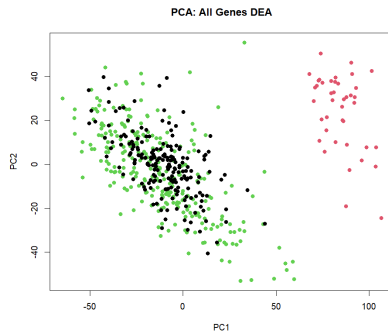
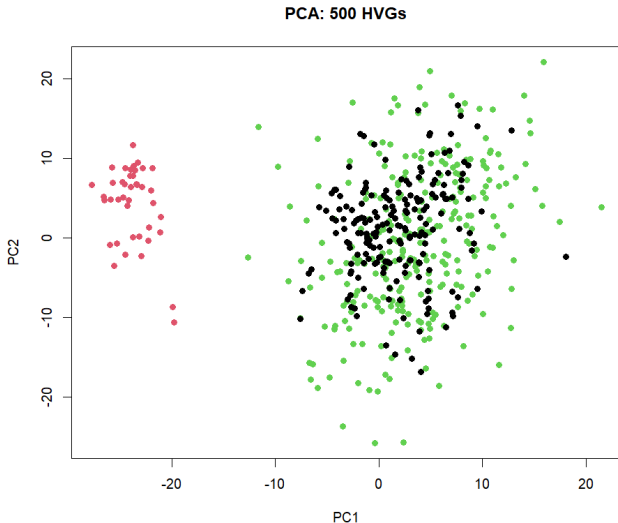


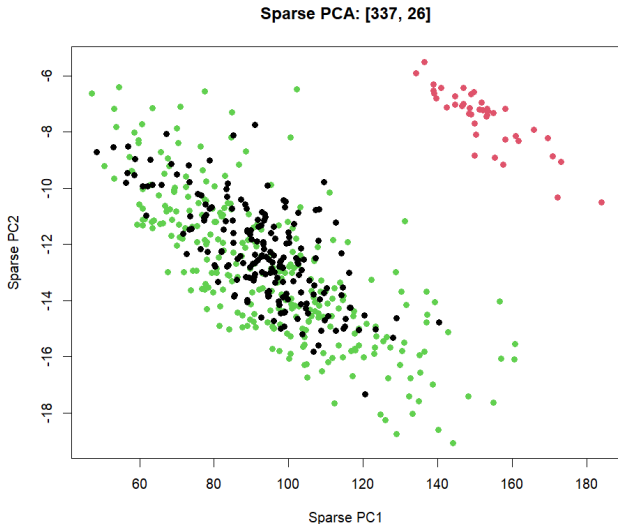
Figure 2: DEA PCA 4806



500 HVG PCA

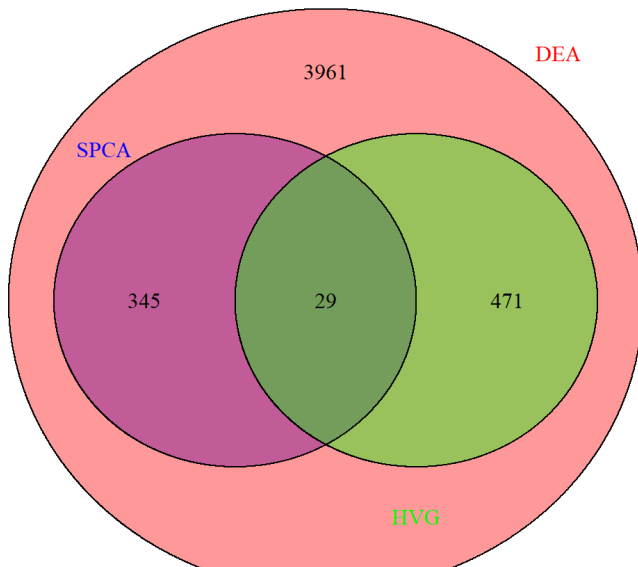


Sparse PCA



- 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

