ANALYSIS OF GENES EXPRESSION USING UNSUPERVISED ML

Stanislav Liashkov

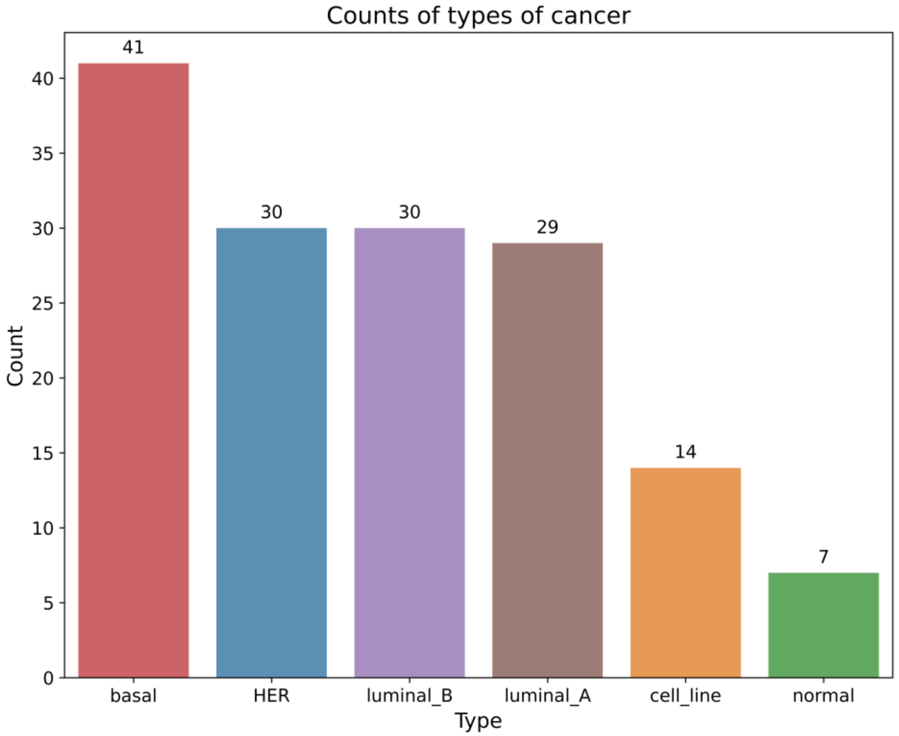
ABOUT DATA

We work with dataset from Curated Microarray Database (CuMiDa) that is called **GSE45827.** The dataset contains 151 samples of genes expression sequence with 5 (+1 healthy) different type of cancers as a label.

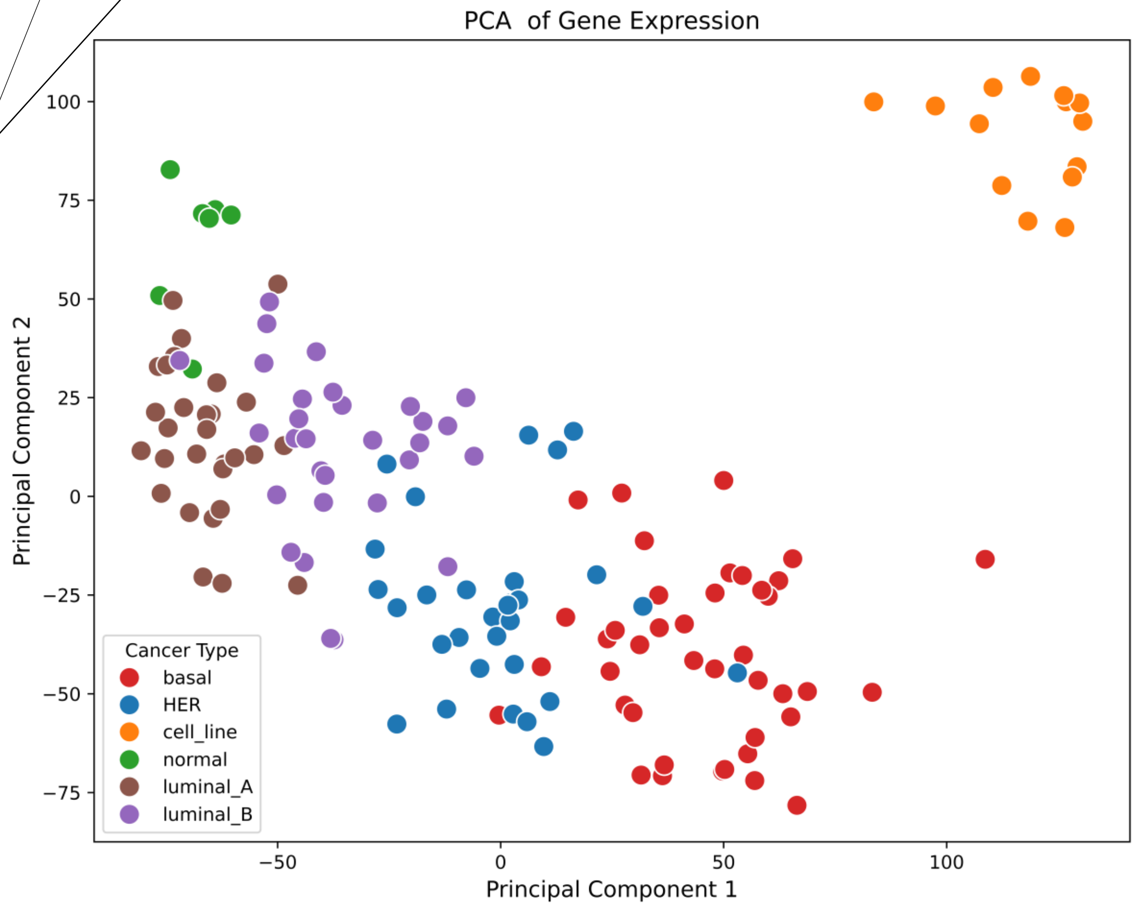
We aim to employ various unsupervised techniques and statistical tests to find most influential genes that affect the risk of cancers.

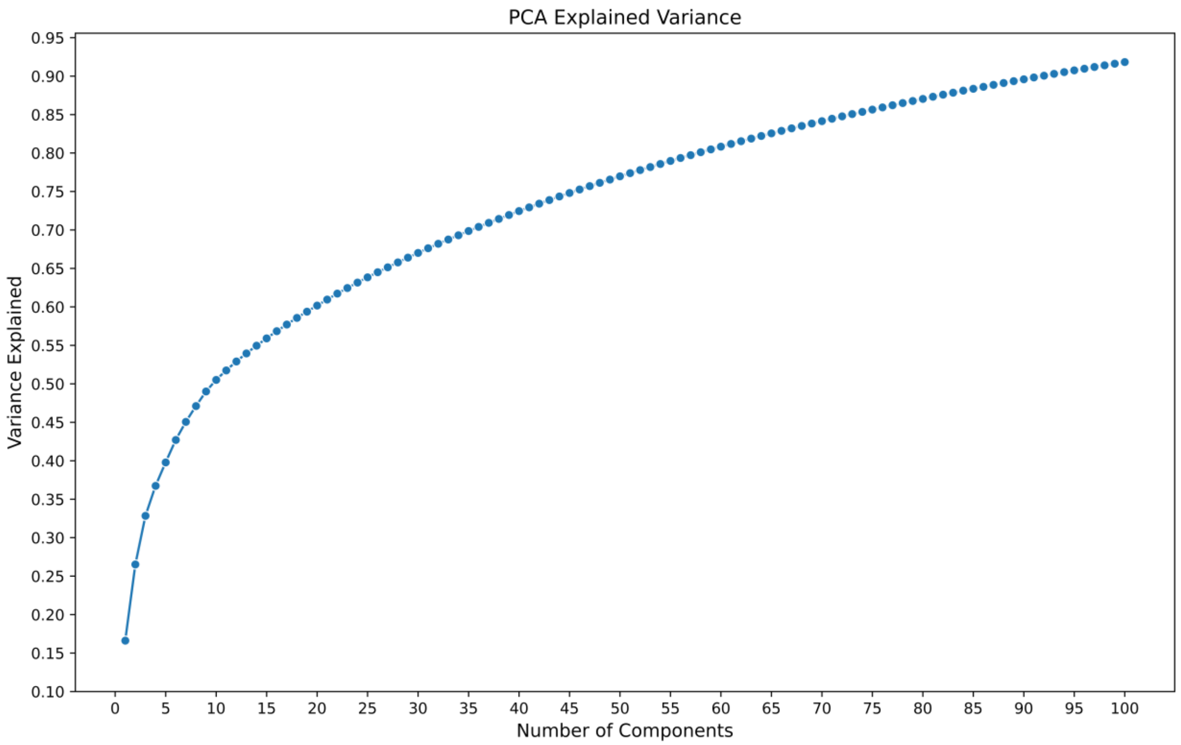
GSE45827 dataset info:

- 151 samples
- 54676 features (genes)
- 6 classes (unbalanced)
- Memory usage – 140 Mb



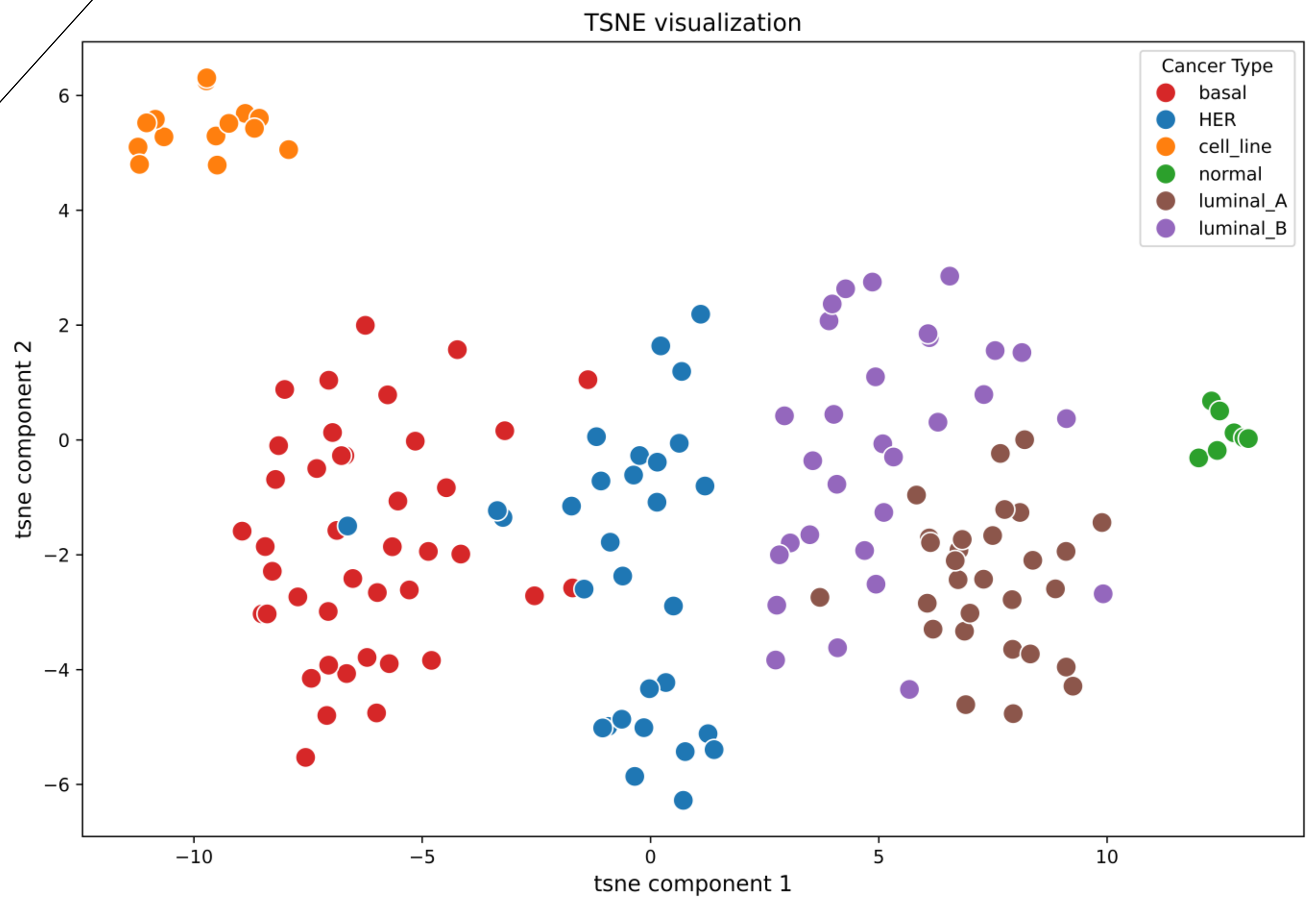
VISUALIZING HIGH DIMENSIONAL DATA: PCA





Using PCA on full gene matrix for visualization is not the best idea. First two components express only **27%** variance. Different types are not clearly separated on plot. That means that we need to try more sophisticated dimensionality reduction techniques.

VISUALIZING HIGH DIMENSIONAL DATA: T-SNE



T-SNE (t-distributed Stochastic Neighbour Embedding) - is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map.

This time, we have a better separation of types, though, some of clusters are stuck to each other. We see that **normal** and **cell line** classes are well separated from the rest which is a good starting point. However, we need to improve this visualization.

FEATURE SELECTION

Given the fact that we deal with very high dimensional data, we need a way to limit our analysis by considering only genes that matter in terms of risk of cancers. In order to select most influential genes, we will use combination of two tecnhiques from feature selection.

Mutual Information Criteria

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Mutual Information quantifies the amount of information obtained about one random variable through another random variable.

Chi-Squared statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Chi-Square Statistic tests for independence between two categorical variables. In feature selection, it identifies features whose distribution is related to the target variable's distribution. (more details in attached notebook)

We are going to score all genes we have using these two approaches. Finally, we will select those genes that satisfy:

- Chi-Square p-value < 0.01
- Mutual Infromation score >= 0.5

FEATURE SELECTION

Given the fact that we deal with very high dimensional data, we need a way to limit our analysis by considering only genes that matter in terms of risk of cancers. In order to select most influential genes, we will use combination of two tecnhiques from feature selection.

Mutual Information Criteria

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Mutual Information quantifies the amount of information obtained about one random variable through another random variable.

Chi-Squared statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Chi-Square Statistic tests for independence between two categorical variables. In feature selection, it identifies features whose distribution is related to the target variable's distribution. (more details in attached notebook)

We are going to score all genes we have using these two approaches. Finally, we will select those genes that satisfy:

- Chi-Square p-value < 0.01
- Mutual Infromation score >= 0.5

TUNING T-SNE PARAMETERS

After having important genes selected, I would like to introduce an additional step for achieving a decent 2D visualization of cancer clusters using t-SNE – optimization of t-sne hyperparameters.

In optimization we need some metric that **quantifies the quality** of embedding.
I use metrics suggested in a research paper *"The art of using t-SNE for single-cell transcriptomics" by D. Kobak and P. Berens (2019).*

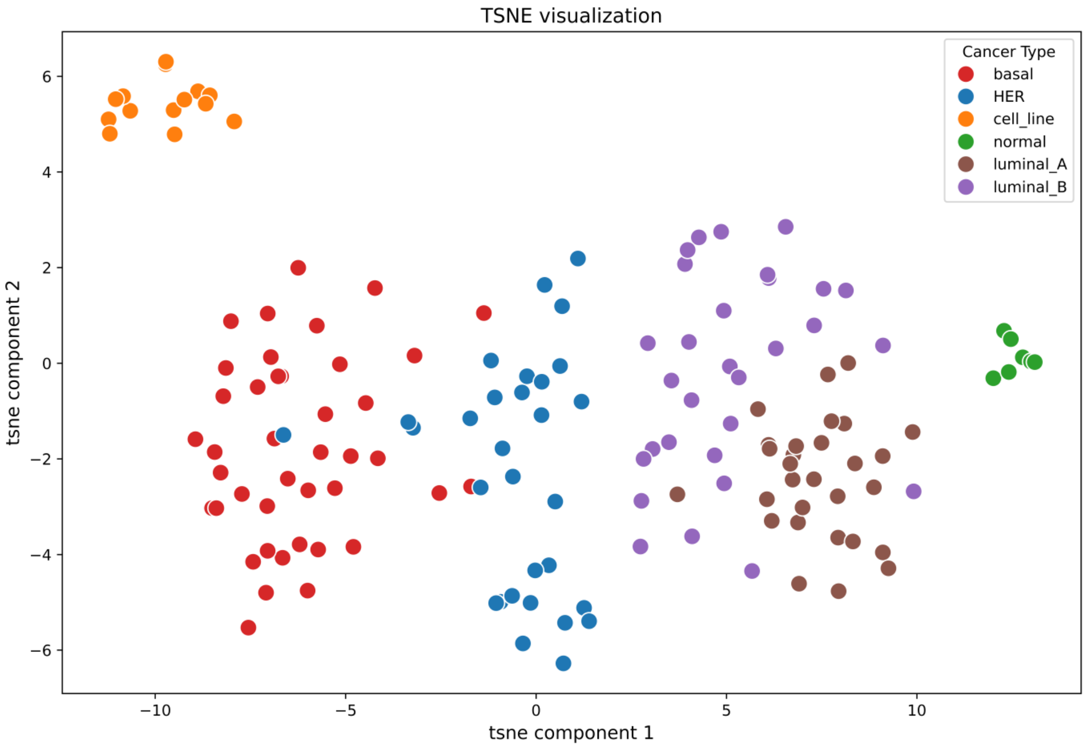
Metrics:

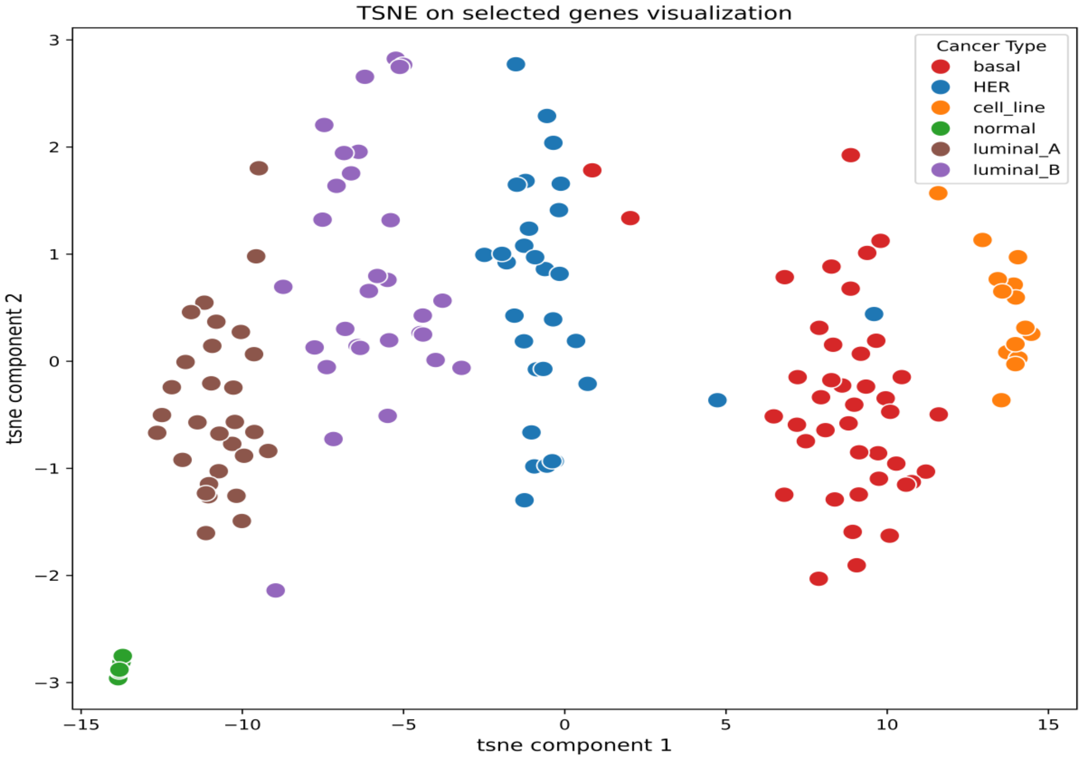
- ***KNN*** - The fraction of K-nearest neighbours in the original high-dimensional data that are preserved as K-nearest neighbours in the embedding
- **KNC -** The fraction of K-nearest class means in the original data that are preserved as K-nearest class means in the embedding.
- **CPD** - *Spearman correlation* between pairwise distances in the high-dimensional space and in the embedding.

The goal of optimization is to try to keep all 3 metrics as high as possible. Therefore, we might use **the average** as a metric we optimize during hyperparameters tunning.

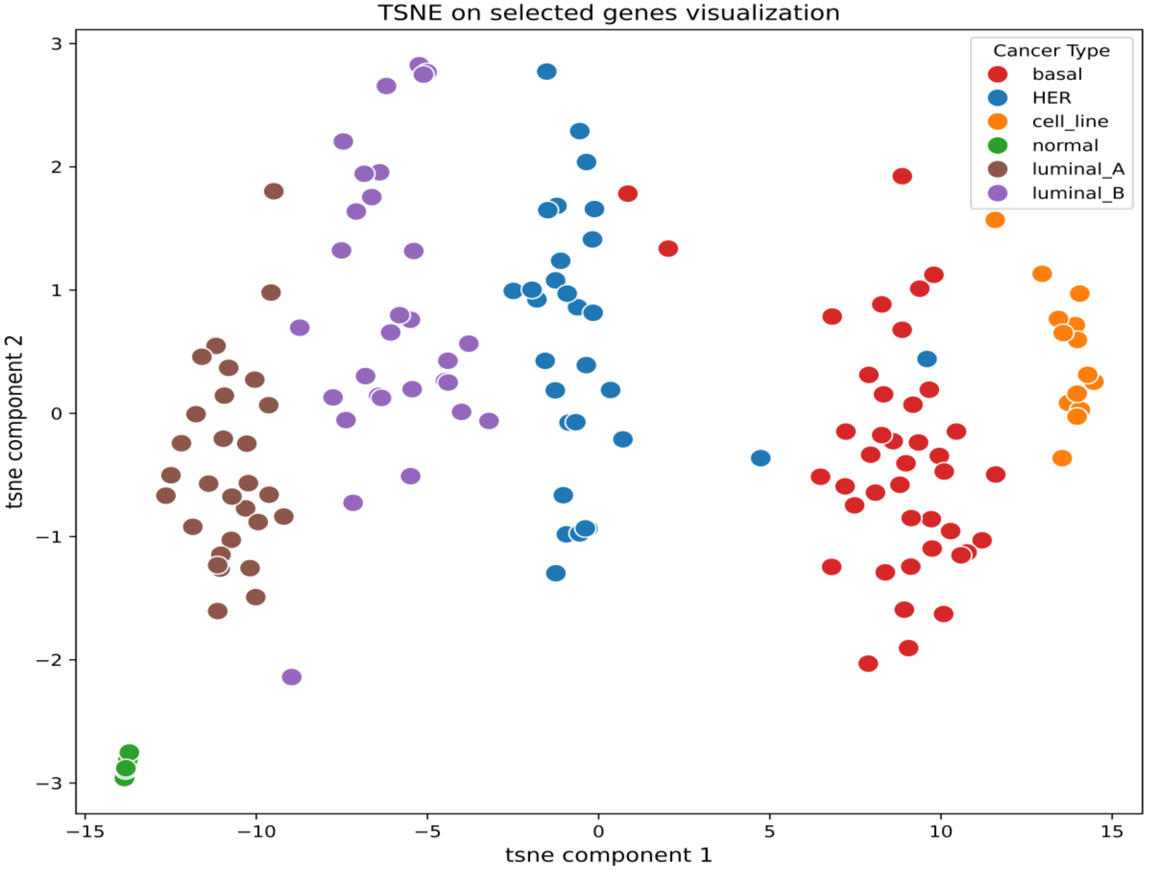
Hyperparameters tunning is perfomed by **Optuna** framework that uses Bayesian Optimization.

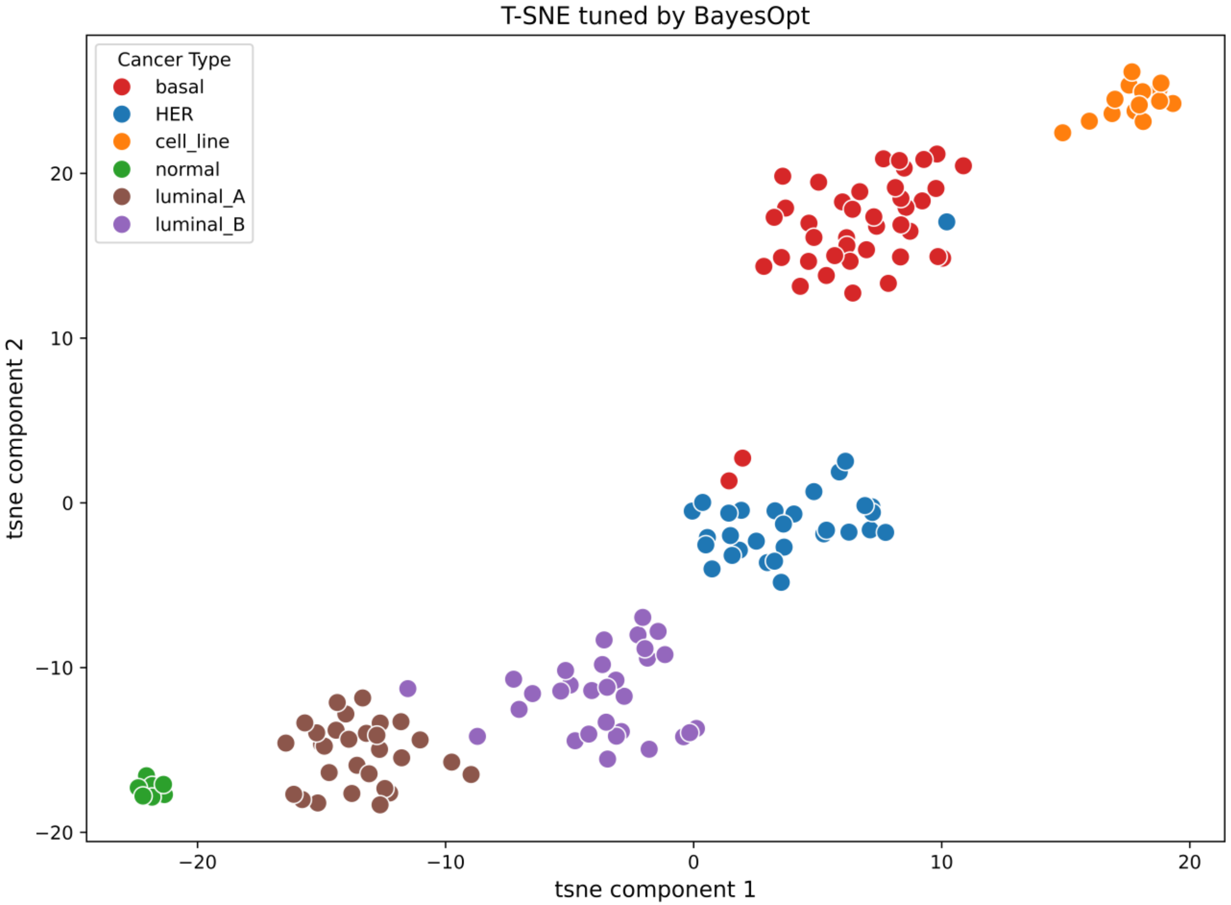
VISUALIZATION COMPARISONS



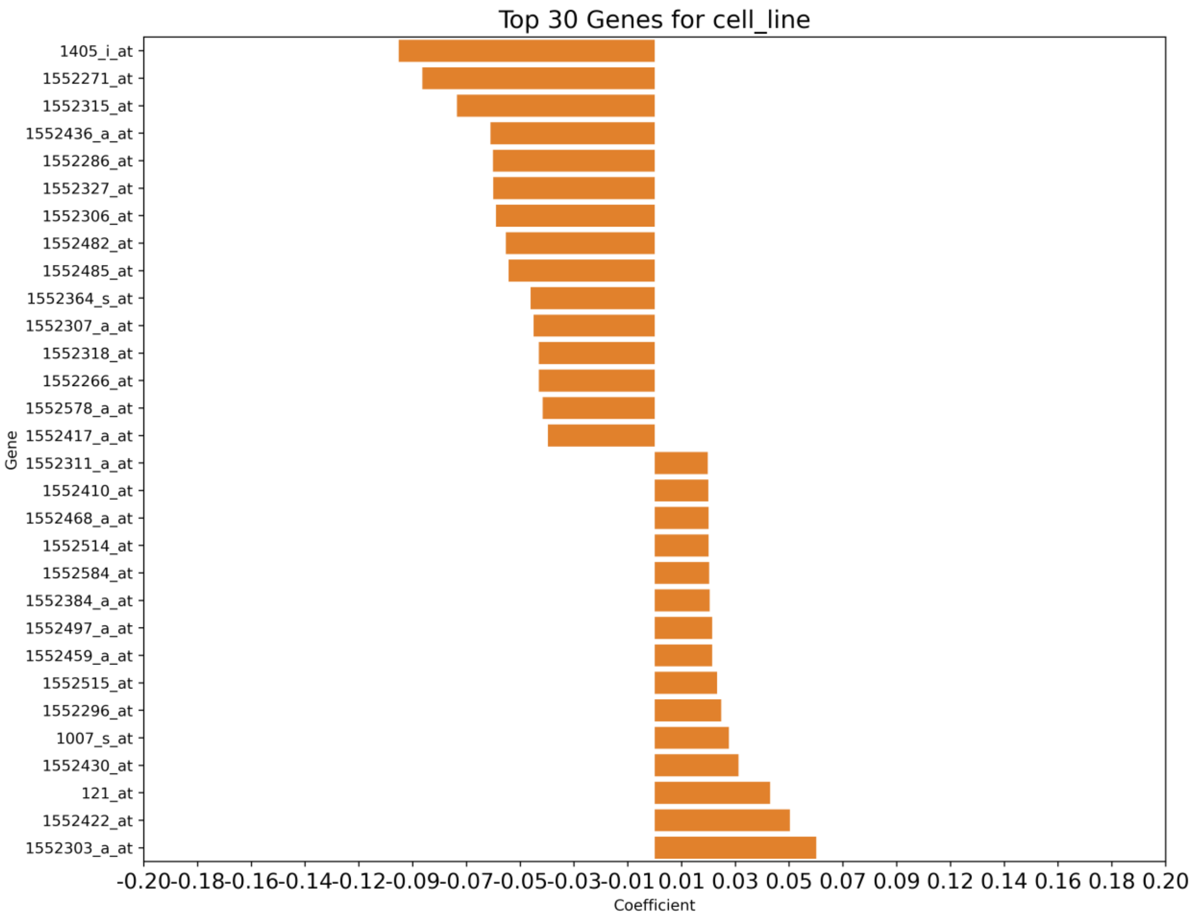


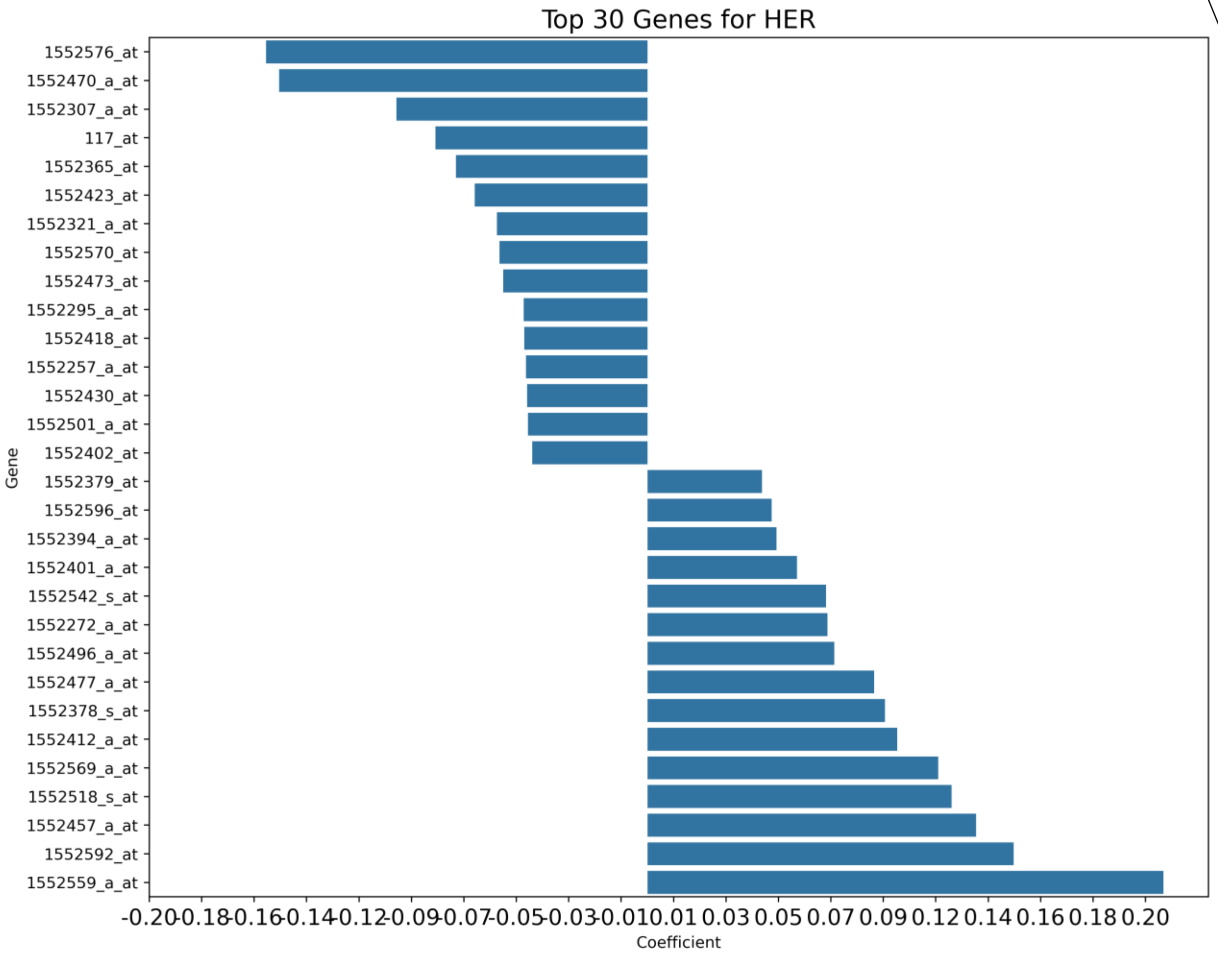
VISUALIZATION COMPARISONS



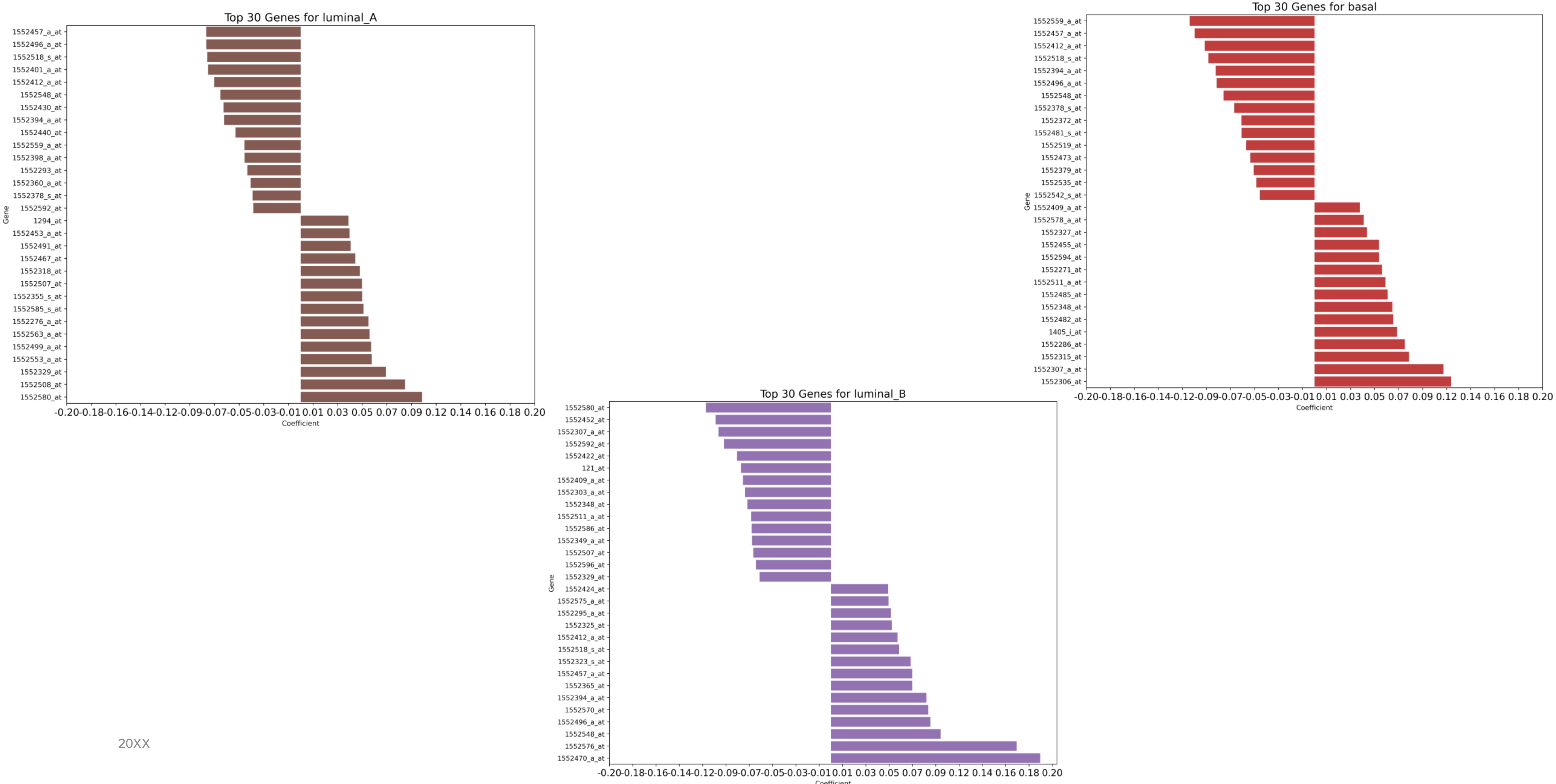


DISCOVER GENES AFFECT BY LOGISTIC REGRESSION





DISCOVER GENES AFFECT BY LOGISTIC REGRESSION



Top 30 Genes for luminal_A

Top 30 Genes for basal

Top 30 Genes for luminal_B