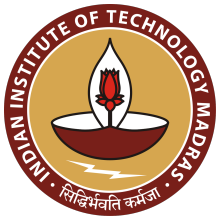
BT3041: Analysis and Interpretation of Biological Data

Assignment 3

Sumedh Sanjay Kangne (BE21B040)
Department of Biotechnology
Indian Institute of Technology, Madras

A BT3041: Analysis and Interpretation of Biological Data Assignment



May 18, 2024

Dataset

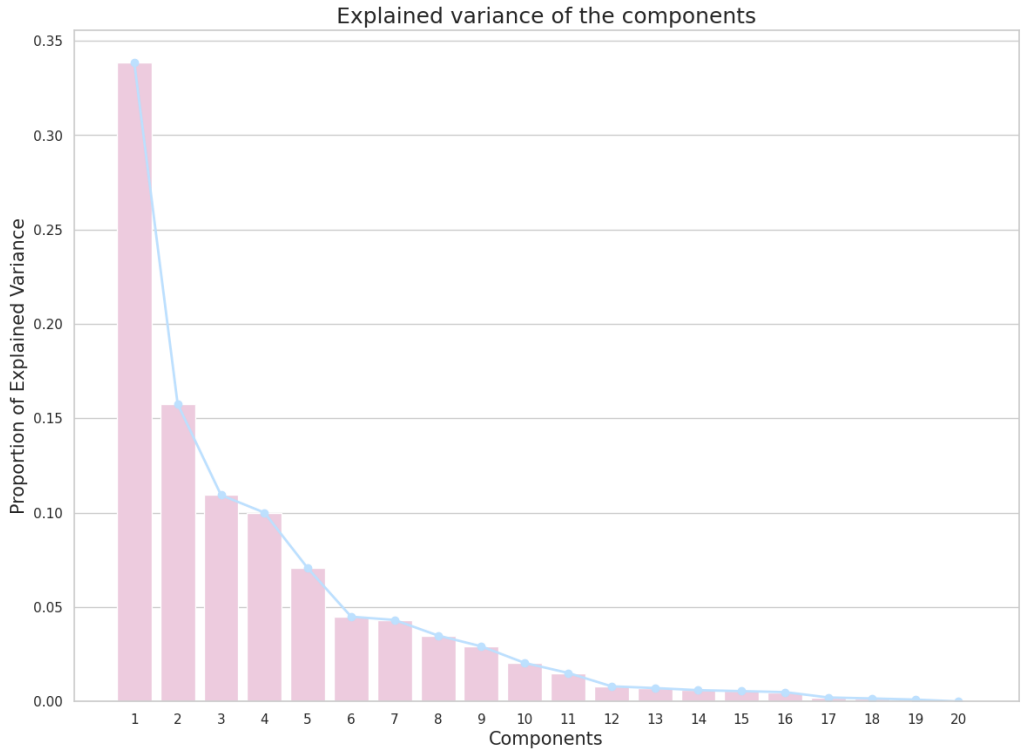
This dataset contains the expression values of 21487 genes measured across 20 different tissue/cell lines of Chinese Hamsters.

Gene	Bmp4_1	Cdkn3_1	Cnih1_1	Gmfb_1	Cgrrf1_1	Samd4a_1	Gch1_1	Wdhd1_1	Socs4_1	Mapk1ip1l_1
ERR2593198.bam	0	2228	2601	3811	657	1152	1420	2239	814	2196
ERR3374021.bam	0	3659	2801	3556	638	823	2223	1335	496	2879
SRR10572661.bam	0	5308	6317	13023	1551	2630	3245	1568	1849	4286
SRR16796949.bam	0	5770	9241	14418	1548	1677	3185	1943	2028	2226
SRR3401747.bam	8	3085	4160	2616	599	921	1223	1305	935	2786
SRR3401748.bam	2	1671	5624	3486	566	1053	1920	603	834	2342
SRR3401749.bam	7	1629	3322	2891	919	634	1709	1101	629	2788
SRR3401750.bam	12	844	2869	2701	830	634	1871	686	659	2760
ERR4184070.bam	3	5542	5019	9321	1007	2721	1119	4989	1815	7209
ERR4184071.bam	1	4244	3877	7367	803	2175	857	3671	1450	5685
ERR4184072.bam	1	3395	3503	7110	1057	1023	4395	1854	1333	4712

Each row consists of gene expression levels of 21487 genes whereas each column consists of the 20 different tissue/cell line samples.

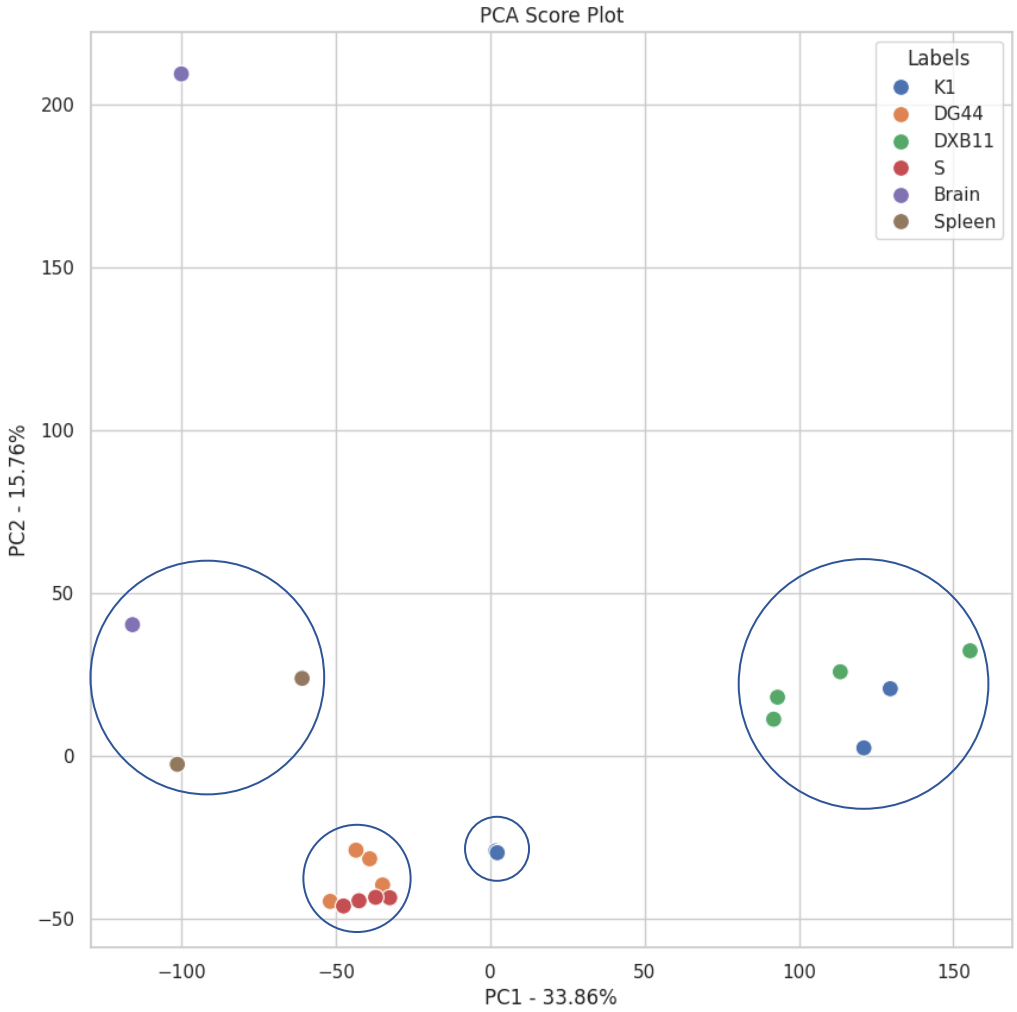
Performing PCA

Now performing PCA on the above-mentioned dataset reduces the total number of components to 20 (i.e. min(no. of rows, no. of cols)). The following scree plot shows the percentage variance of each Principal component.



The top 2 components have a variance of 33.86% and 15.76% respectively.

Further looking at the Score plot for PC1 vs PC2,



we can see that roughly 4 clusters can be formed. Cell line groups clustering together include **DG44 & S**, **Brain & Spleen,** and **K1 & DXB11.**

To see which genes contributed the most to PC1 and PC2, an eigenanalysis will provide a set of unit vectors (eigenvectors) and a corresponding vector of values (eigenvalues). The unit vectors define orthogonal directions in the same feature space as before, but optimized to capture the maximum amount of the variance of the input data in the fewest number of vectors. The eigenvalues tell you how much of the original variance is captured in the direction of the corresponding eigenvector. Therefore, a larger eigenvalue means a higher contribution.

Top 5 genes with the highest contribution to PC1:

1. LOC100770091_2
2. Spata21_1
3. Lrrfip2_1
4. Myo19_1
5. Bbs4_1

Top 5 genes with the highest contribution to PC2:

1. LOC100761590_1
2. LOC100766062_1
3. LOC100755925_1
4. Fbxo46_1
5. Ppp1r7_1

Final Analysis

Principal Component Analysis (PCA) is a powerful tool used in exploratory data analysis and for making predictive models. It is a method used to bring out strong patterns in a dataset by suppressing variations.

When applied to the gene expression data, PCA helped us understand the following:

1. **Variation in Gene Expression**: The principal components represent the largest sources of variation in the gene expression data. The first principal component represents the direction in the high-dimensional space along which the samples vary the most. We were thus able to see the trend of gene expression levels in a lower dimensional space.
2. **Sample Grouping**: The PCA plot can show whether samples within the same group have similar gene expression profiles. Samples from the same group would cluster together in the PCA score plot.
3. **Outlier Detection**: Outliers, or samples that have unique gene expression profiles, will appear as isolated points in the PCA score plot.

Thus, PCA allows us to visualize the trends of the higher dimensional data in a low dimensional space allowing us to deduce important insights from it.

Code

The link for the code and all the images used can be found here:

https://drive.google.com/drive/folders/1vuBeYWczmkf7Bh3ZE-sH7JkYWfYS6QwS?usp=drive_link

The End