Plot.PCA

The package is coded in R and takes as input two files:

1. Gene intensity file
2. Metadata file

Please note:

1. The input files should be csv format and I have incorporated several check measures. The first input file should have the 2nd column (i.e., gene symbol) with a header "symbol". If this is missing, the code will give an error.
2. The second input file should contain a column with header "sIdx". The IDs given in this column will be mapped to the headers of the first input file. If this does not match, the program will give an error.
3. The first input file may contain some junk data. Some missing values can also be present. Such genes are removed from the analysis. However, if these are present in more than 80% of the genes, the program will give an error.
4. If two or more genes show exactly similar values for all samples, such duplications will be removed. Suppose gene1 has intensities 0.1, 0.2, 0.8, 0.1 and gene2 also has intensities 0.1, 0.2, 0.8, 0.1. In this case, gene2 will be removed in the pre-processing. Such duplications increase the calculation time and interefere in identifying the top genes showing highest ranking.

The developed R code performs Principal Component Analysis (PCA) and returns a PCA plot. For PCA, the package utilizes prcomp function. Thus, the PCA calculations will be done by a singular value decomposition.

The aim, here, was to identify if the different time points differentiating when seen on the PCA plot, given the gene expression data for different time points (see example for input).

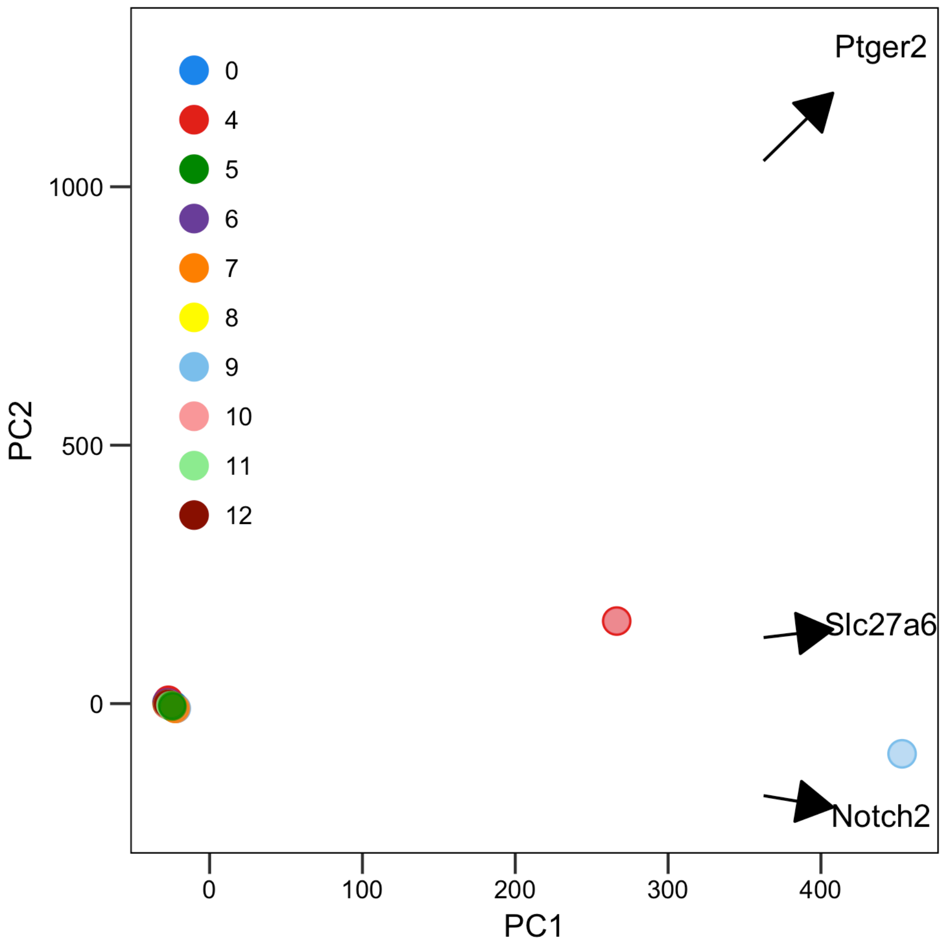

Figure 1. PCA plot showing all the 30 samples and 10 time points.

PC1 explained a variance of 71.8% and PC2 explained 8.6%. The samples S2 (timepoint = 9 hours) and S13 (timepoint = 4 hours) showed the maximum distance from all other samples. All other samples were clustered together. According to this plot, it is difficult to say that the time points are differentiating on the PCA.

Further, a log transformation was carried out on the data and performed PCA (Figure 2). After log transformation, PC1 explained a variance of 65.7% and PC2 explained 8.8%, which was slightly lower compared to the above case.

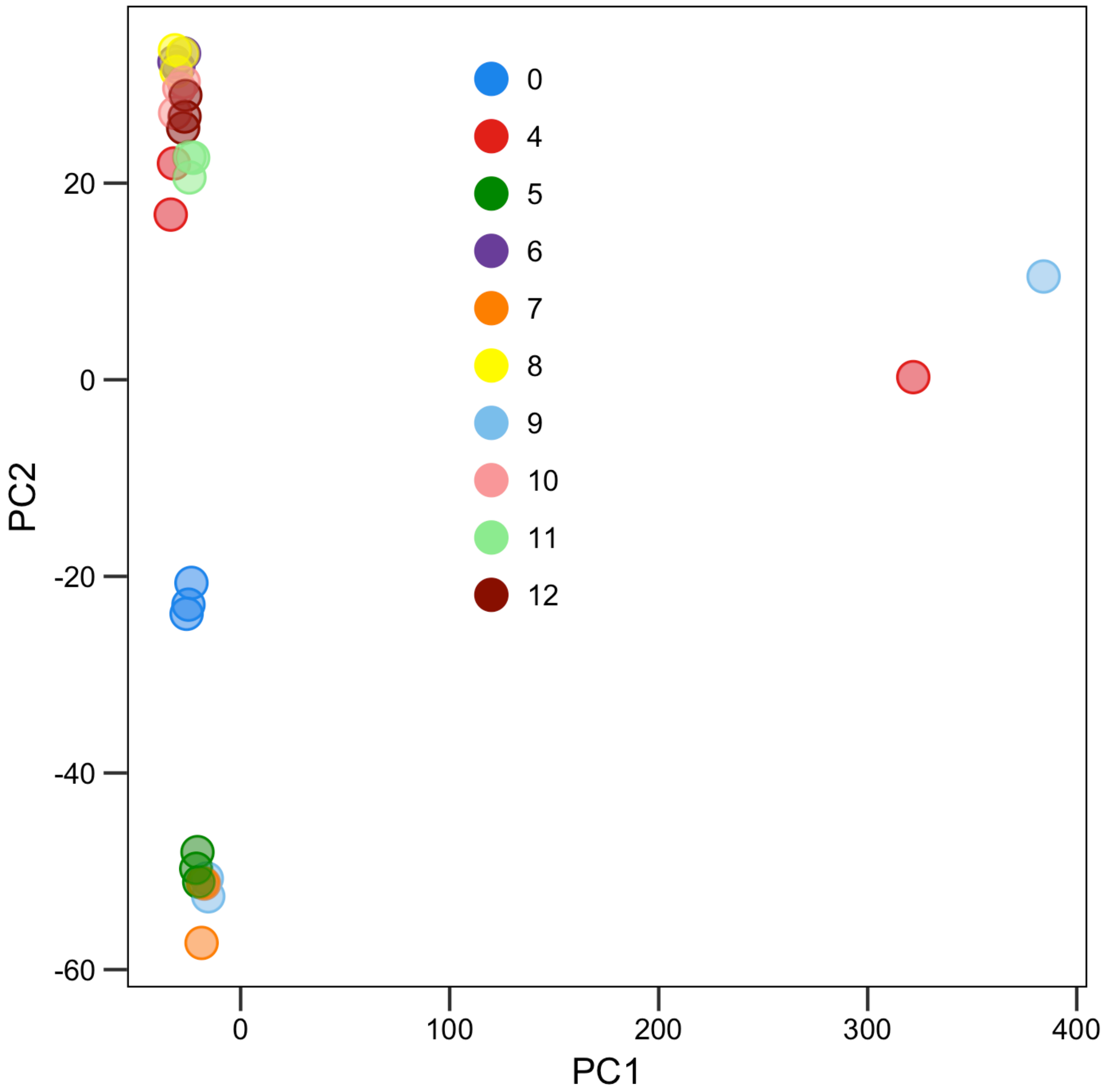

Figure 2. PCA plot obtained after log transformation of gene intensity values.

In this case, a total of four clusters could be seen in the plot. Let us label the clusters as A, B, C and D for easier discussion.

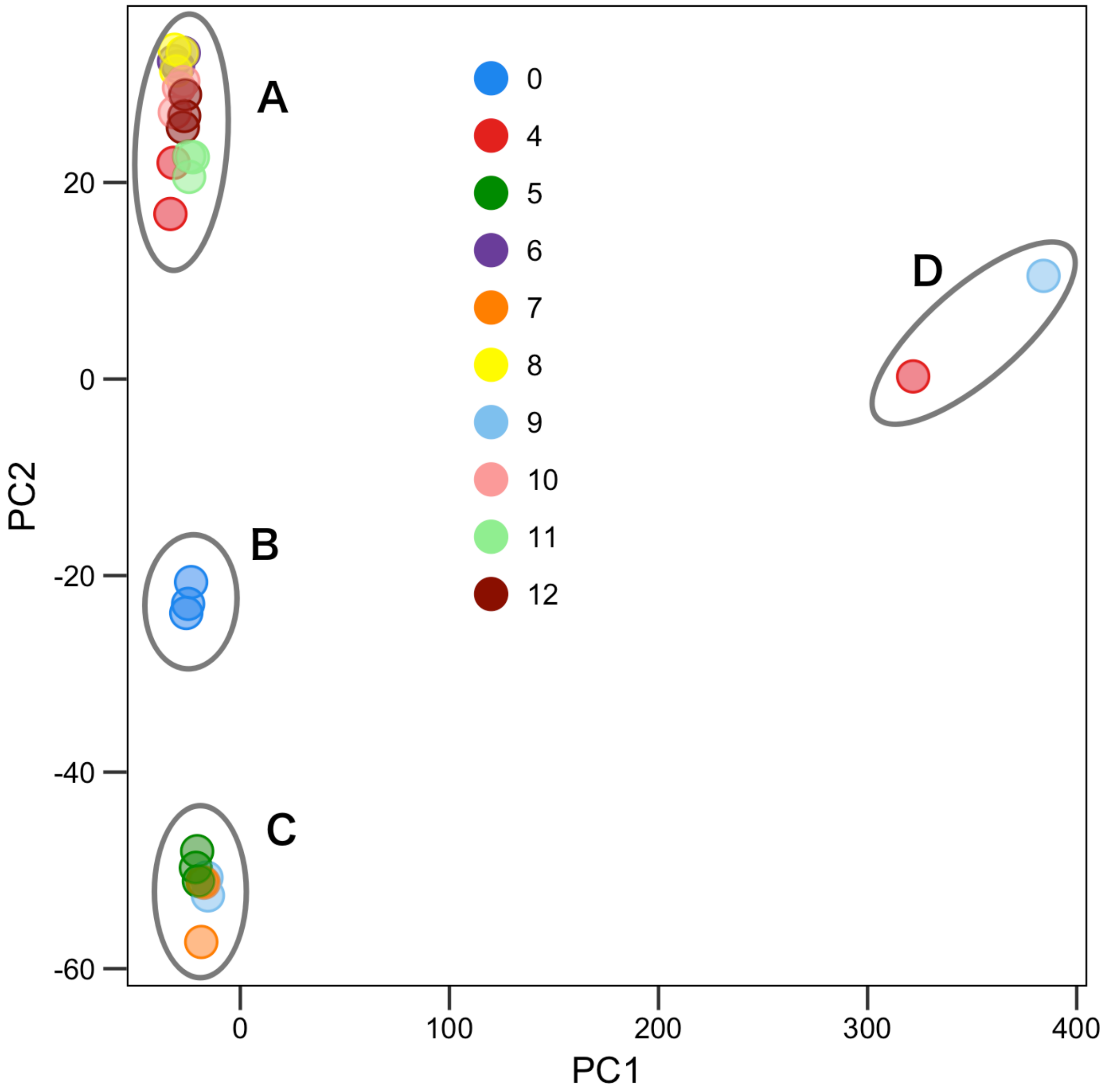

Figure 3. PCA plot obtained after log transformation of gene intensity values showing clustering.

The clusters A, B and C are separated on the basis of PC2, which explained only 8% variance. It is to be noted here that these clusters are separated according to time points. The cluster A consist of samples from timepoints- 4, 6, 8, 10, 11, 12. Cluster B consists of samples from 0 hours only. Cluster C consist of samples from timepoints 5, 7 and 9.

Since only two samples were having large distances from the rest of the samples, other samples seemed to cluster together in Figure 1 and 2. To overcome this problem, I removed these two samples (S2 and S13) from the data and re-analyzed.

To remove these samples, the following commands were used:

```
row.names.remove <- c("S2", "S13")
data = data[!(row.names(data) %in% row.names.remove), ]
```

(Refer to script.R in example folder)

The PC1 and PC2 could explain 40.8% and 19.2%, respectively. A clear difference between different samples based on time points is now visible on the plot (Figure 4).

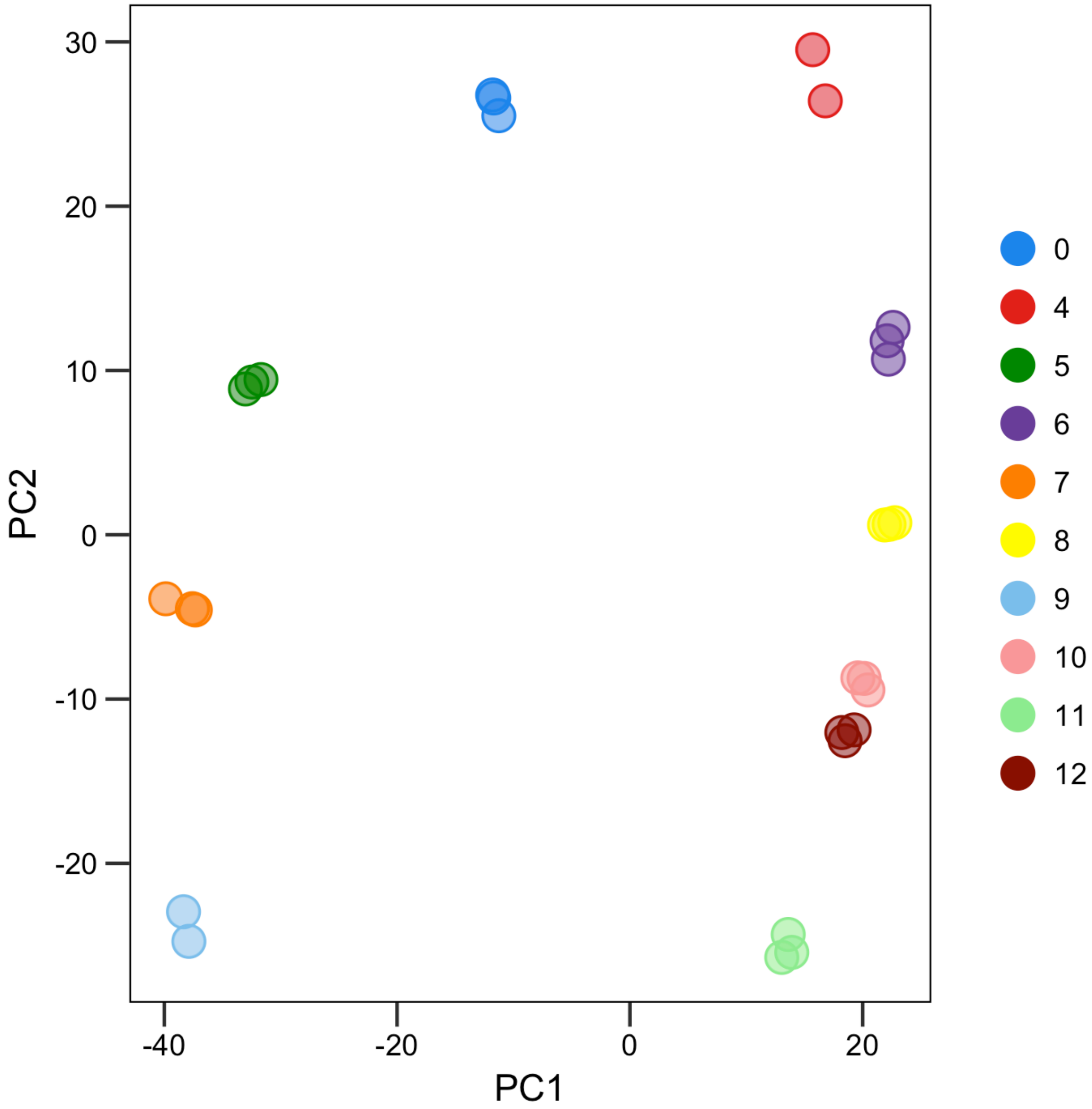

Figure 4. PCA plot after log transformation and removing two samples (S2 and S13).

This shows that the time points at which samples were collected are differentiating when seen on a PCA plot. As also previously observed (Figure 3), the time points 5, 7 and 9 were closer to each other (separated mainly on PC2). Similarly, the time points 4, 6, 8, 10 and 12 were also closer to each other. Time point zero was distant from all other time points. The genes "Lonrf3", "C1rb" and "Cdh15" contributed maximum to the separation. The plot suggests that in the given dataset, the genes show different intensity (or expression) at different time points.

In various scenarios, such as in temporal gene expression analysis, it is important to see if different timepoints are clustered separately on a PCA. For example, when given antibiotic treatment to bacteria, it would be interesting to note that the importance of time for which the treatment is given.