Gene Expression in the Brain

2022-09-27

1. Abstract

2. Exploratory Data Analysis

For a good experimental design for gene differential analysis, measurements have to be accurate. Due to uncertainty in measurements, there is the need for replication. The most common types of replications are biological and technical replications. Biology replication involves taking measurements mostly in the same lab and with same technology for several samples of the same cell, while technical replication deals with repeating identical labs and protocols for sequencing on a single sample. While replication fosters confidence in our analysis as it enables us to quantify uncertainty, it posses yet a potential problem in gene differential analysis. For example, technical replication, if exists, could confound the true gene expression differentiation or can cause expression for reasons unrelated to the levels of the expression. Thus, it is highly imperative to investigate the presence of variations arising from replications before conducting further statistical analysis. A statistical tool for doing such investigation is called Exploratory Data Analysis (EDA).

The dataset has technical replicates. Measurements were taken from three labs viz Michigan, Irvine and Davis. We first make examine the distribution of genes by Lab for box sexes (Fig 1 & 2) and the look at same for the ACC & DLPFC brain regions by lab (Fig 3 & 4).

YUMENG'S CODES

We observe in Fig 1 & 2 that the distributions of the gene expressions vary by lab for both sexes. This is a clear indication of variability due to technical replication. Same story can be told about the distributions of the genes for the ACC & DLPFC brain regions. Further, we make a Multi-Dimensional Scale (MDS) plot to see how the various replicates cluster among each other for the 10 patients.

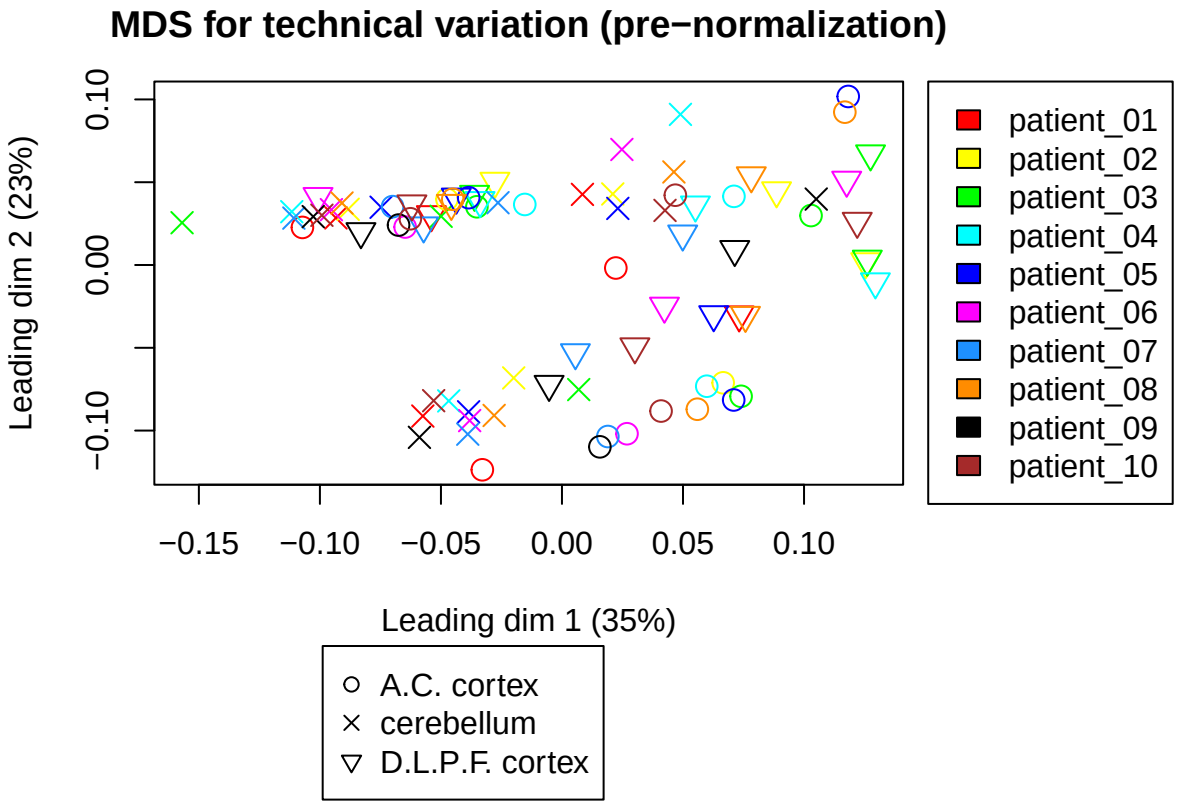
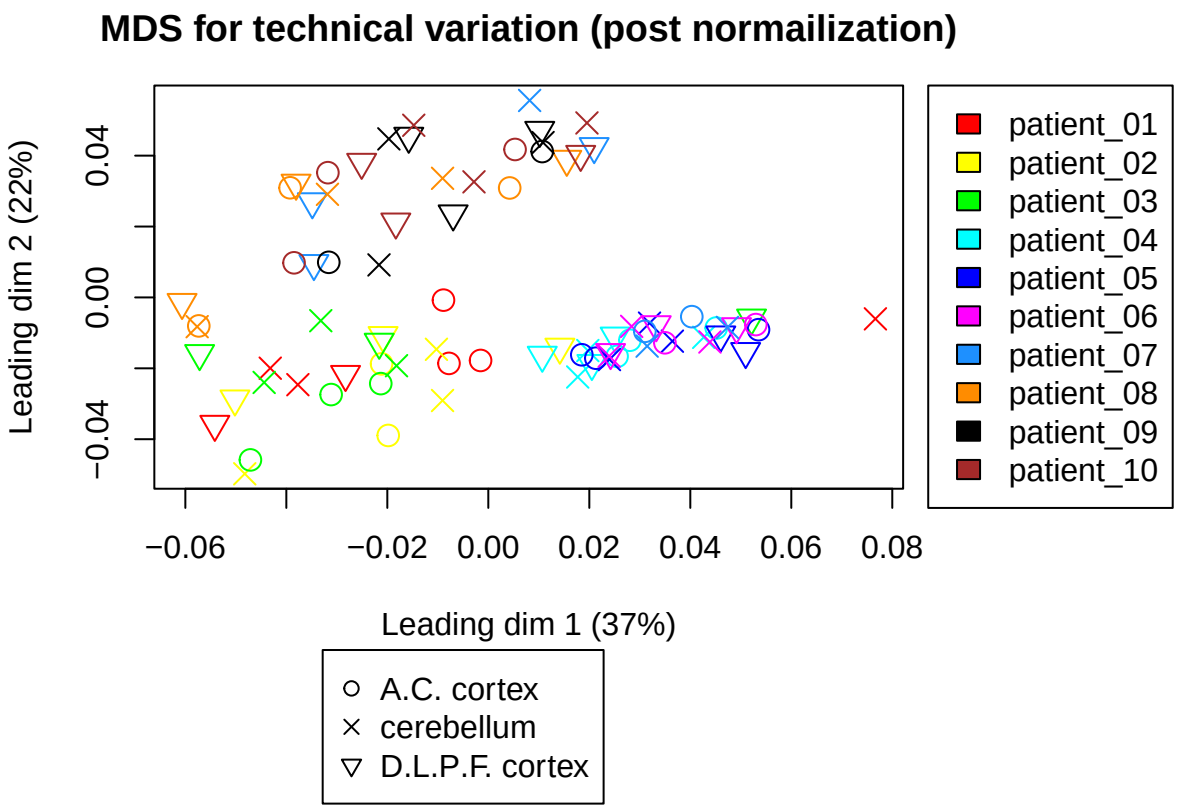


Fig 5 show the first two dimensions of the MDS. The first dimension explains 35% while the second explains 23% of the total variation explained by the MDS. Clearly, there is significant sparsity for the replicates of each patient as the various measurements across the three labs are farther apart.

The issue of variation due to technical replications is one that had left researches with serious decisions to make in the past years. Should the scientist throw the data away and start all over? Recent developments have provided a good solution to handle this problem viz Normalization. There have been several normalization techniques that scientists have employed in reducing technical variations including the use of housekeeping (control) genes as baseline or the expression level from a particular quantile of the distribution of gene expression values of each sample as well as using variance stabilizing transform from statistical modeling. We employ one of the Normalization techniques which would be discussed in the subsequent section and visualize the MDS plot post normalization in Fig 6.



We observe that, normalizating the data has resulted in significant reduction in the variation due to technical replication (comparing Fig 5 and 6). This gives us some confidence in assessing the true gene differentiation as we perform the appropriate statistical analyses in the rest of the sections of this report.