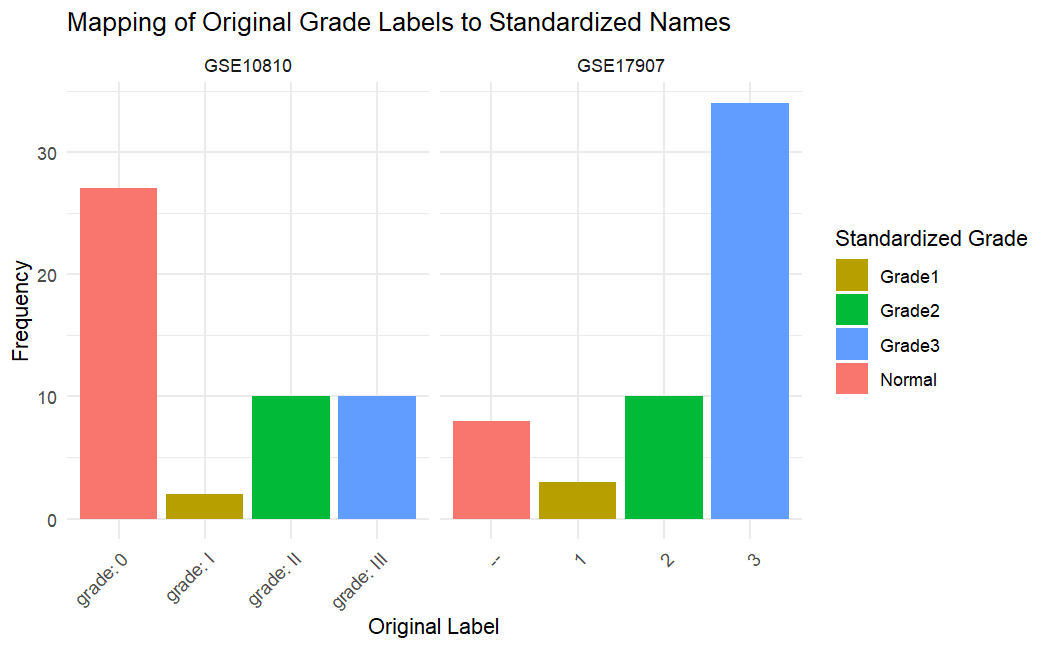
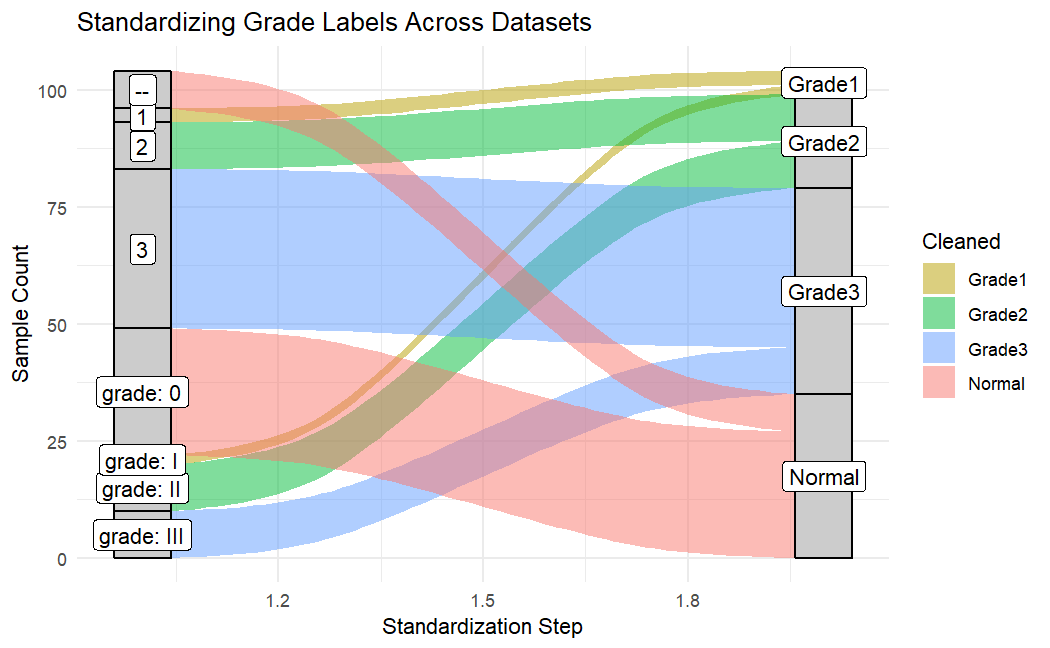
Visual Walkthrough of Data Pre-processing:

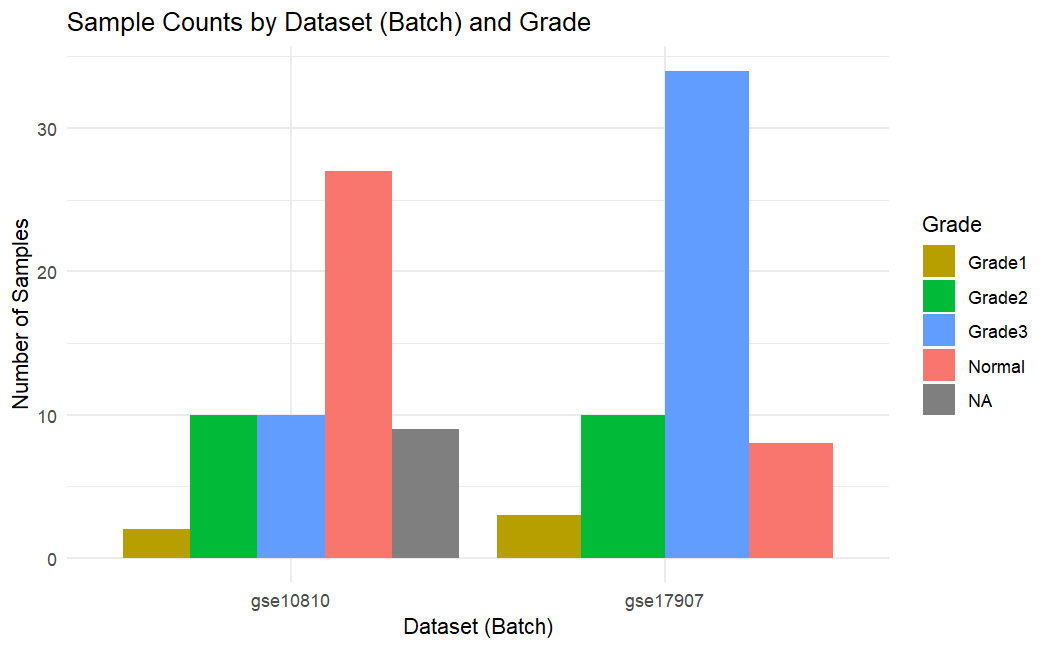
Data processing steps vary between datasets but here are steps we took for the combined(GSE17907, GSE10810).

The first step of data processing after loading up our dataset was mapping the original grade names to their standardized counterparts (see figure below).

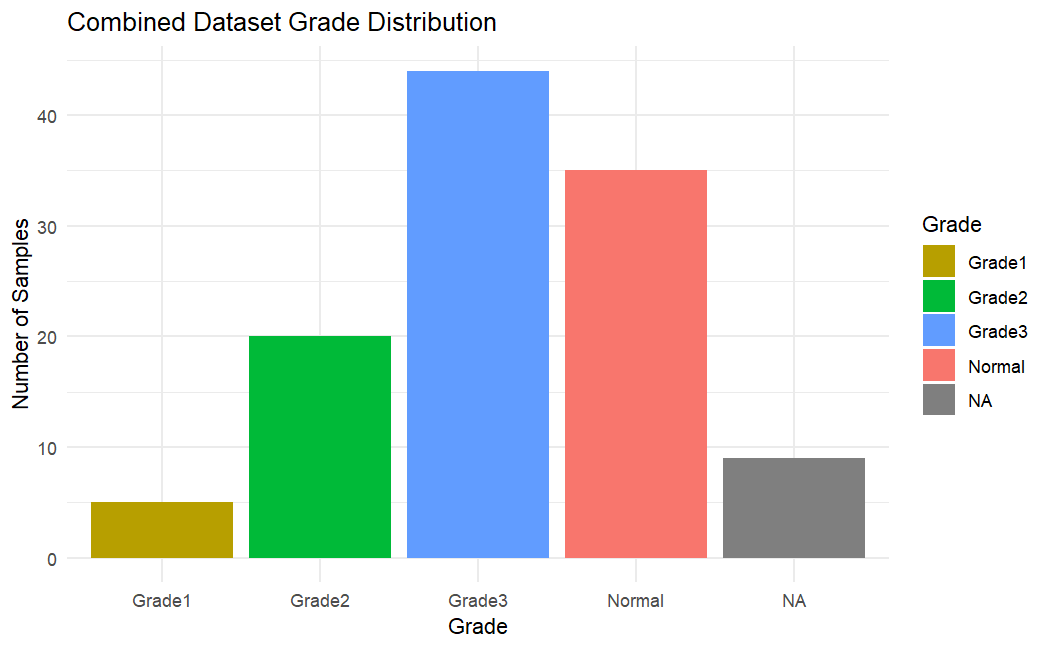


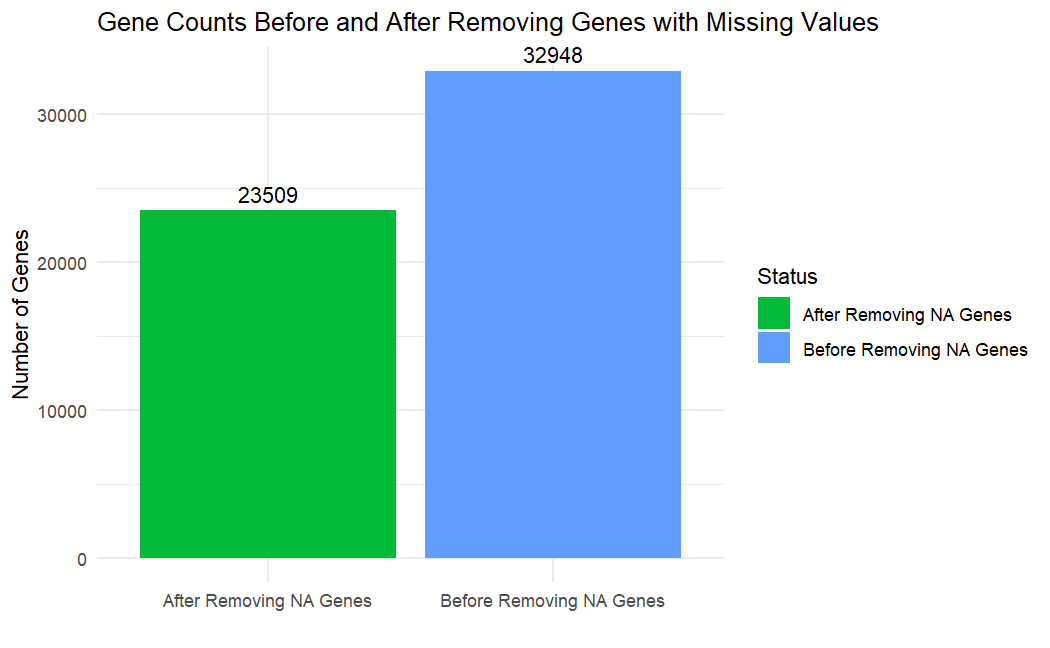
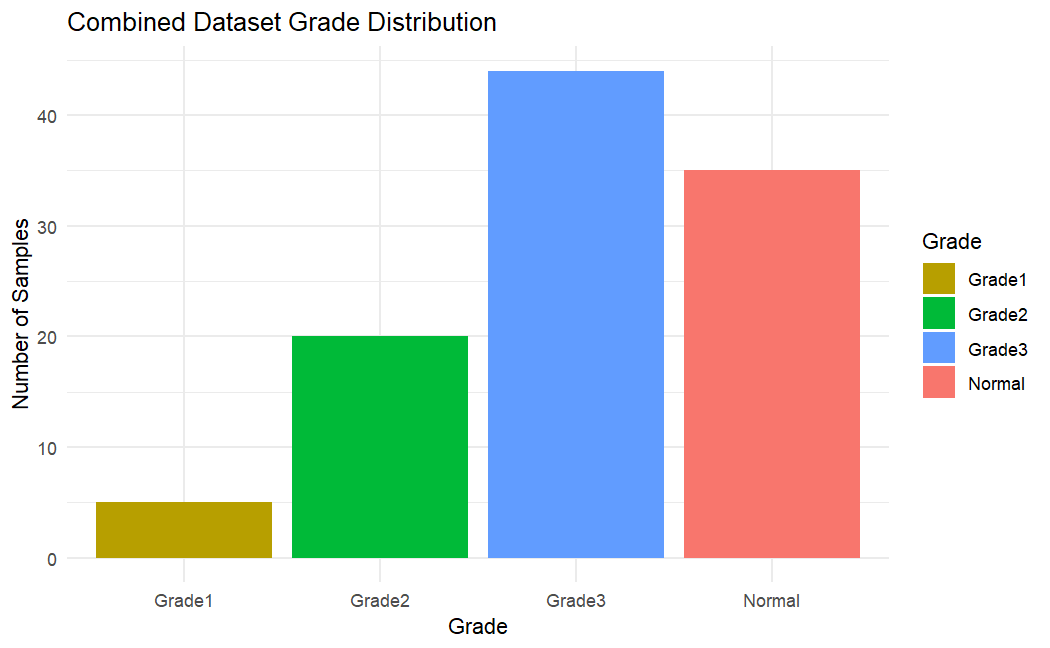
After the mapping process we standardized the grade labels to Grade1, Grade2, Grade3 and Normal visualized below.



After standardization the datasets could now be merged here is a visualization of their “Grade” 

The phenotype data and gene expression data was taken from both datasets. From the phenotype data only the sample name, grade and which dataset was taken from was preserved create a new file “metadata”. All the corresponding expression data for each sample was taken and combined together known as “merged\_expr”. However, after merging there may still be many NA values visualized below in both the grade column of “metadata” and for some genes in “merged\_expr”. Thus, it’s essential to drop all these values before continuing (visualized below).





After removing NA values it’s crucial to remove lowly expressed genes, to:

**Reduce noise**

* Filtering removes low-expression genes that mostly reflect technical background rather than true biological signals.

**Improve statistical power**

* Eliminating uninformative genes reduces the number of statistical tests, increasing the chance of detecting true effects.

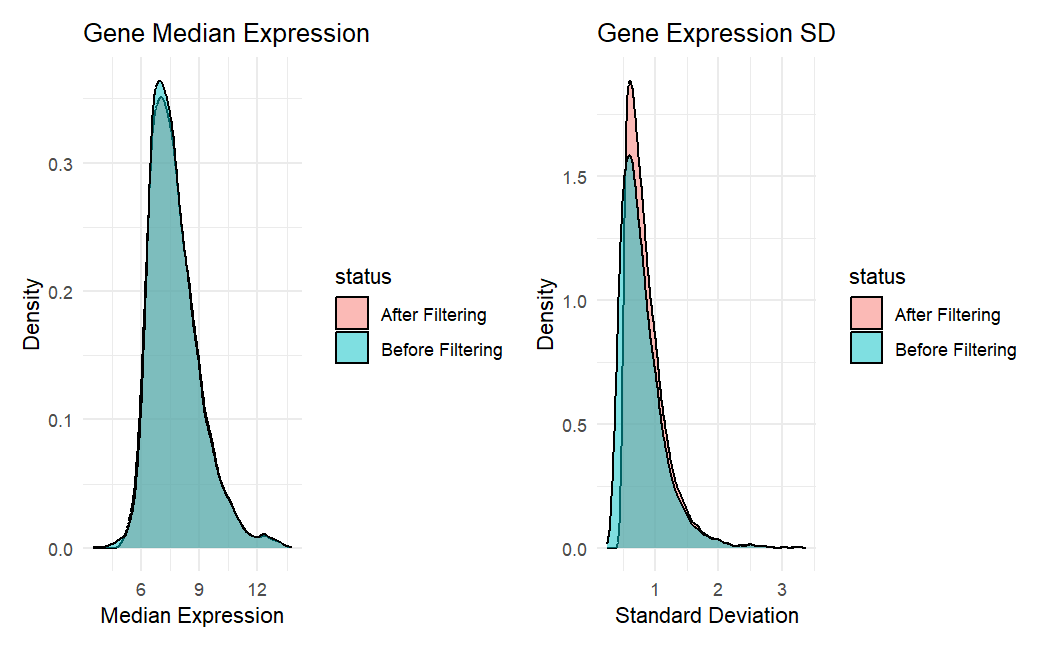
**Speed up analysis**

* Fewer genes mean faster computations and more efficient memory use in statistical and machine learning workflows.

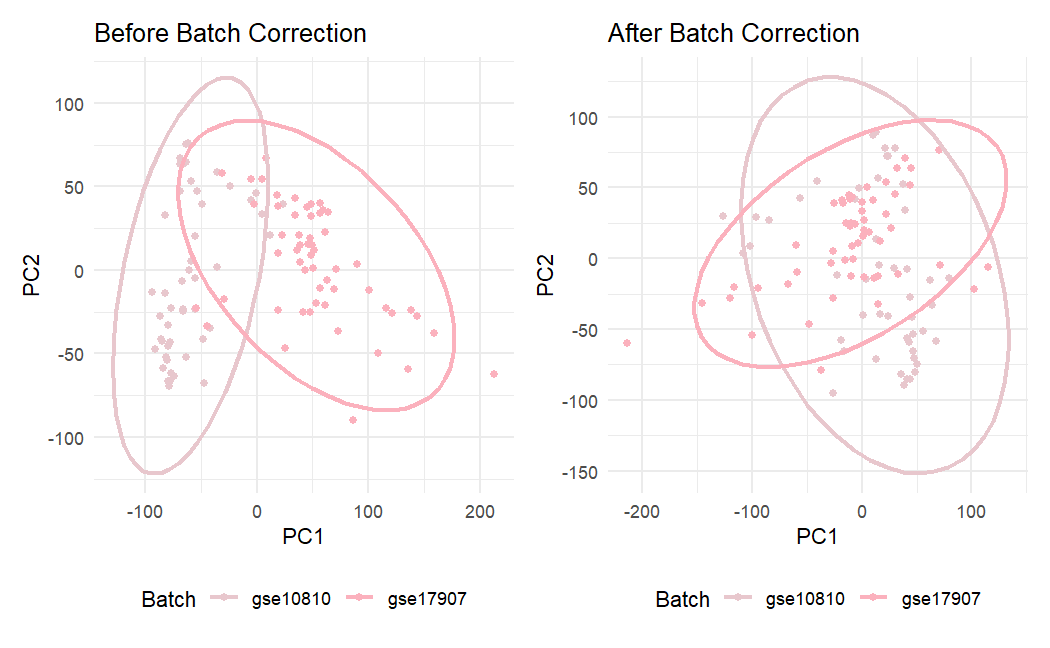
**Focus on biologically relevant genes**

* Highly expressed and variable genes are more likely to be functionally important and biologically interpretable.

Below is a plot showing the before and after of filtering which removes low median expression (genes not expressed in most samples) and then low standard deviation (genes with little variation across samples).



After this batch correction is the next necessary step. Batch correction is done to remove technical differences (like different labs or equipment) to bring the focus back to the biological differences. The left plot below shows the two dataset clearly separated along PC1 suggesting that batch effects are prevalent instead of true differences. Thus the right plot shows the two datasets more closely overlapping in the PCA space, suggesting the “batch” has been corrected for.



After batch correction the data is split into two dataset in a 20/80 split. Where 20% of the data is randomly selected and split into whats known as the testing set. The testing set is used for the final validation to determine how well it works on unseen data, once split it shouldn’t be tampered with until the its used to test the model. The remaining 80% is known as the training data and is what the models are trained on and evaluated (below are the splits for the combined dataset).

|  |  |  |  |
| --- | --- | --- | --- |
| **Training set grade counts:** | | | |
| Normal | Grade1 | Grade2 | Grade 3 |
| 28 | 4 | 16 | 35 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Testing set grade counts:** | | | |
| Normal | Grade1 | Grade2 | Grade 3 |
| 7 | 1 | 4 | 9 |

After splitting the data into test/train split is Resampling. Resampling is a method used to adjust the distribution of a data to balance the different classes. Classes referring to Normal, Grade1, Grade2 and Grade3. From the image below we can observe there is a distinct lack of Grade 1 within the original compared to all other classes thus resampling was used to rebalance the classes. An imbalanced dataset can be a major issue in modeling, for instance what if a dataset contained 95% cancer sample and 5% non-cancer, a model could learn to predict cancer 100% of the time and still have high accuracy (other metrics can help decern this but we’ll cover that in later modules). Therefore, as a precautionary step resampling is necessary.

In the bar chart below we can see how the data is rebalanced after SMOTE a a popular resampling method is used. Creating more samples in the Grade1 column, note that resampling is done after the train/test split to ensure there isn’t data leakage and the final test is blind.

