**JCAP\_RNA\_SEQ\_APP readme.**

**Summary of the Main Functions of this App**

This app can take your counts matrix and phenotype data and perform several tasks when you click on the respective buttons.

The first task that the app can perform is differential expression and feature selection. Differential expression is performed using Limma-Voom. The top 200 hits are then subjected to feature selection, which is performed using a random forest classifier. 20 genes that are deemed to be most important in terms of how great of an impact randomizing/permutating their counts values have on the performance of the classifier are output on the page. The resulting table contains the Ensembl IDs, HGNC symbols, p-values, respective log fold changes , and if applicable, known full names of the related protein products of those top 20 genes. The table can be exported as a .csv file. This process also removes both genes that are lowly expressed among all samples and outlier samples.

The second task that this app can perform is to prepare a PCA plot of your data in which the points are colored with regard to the phenotype of interest (see the legend for details). Hovering your cursor over points on this plot will tell you what sample the points relate to.

The third task that this app can perform is to prepare a UMAP of your data colored in a similar fashion to the PCA plot. Hovering your cursor over points on this plot will tell you which sample the points relate to.

It is noteworthy that the first, second and third tasks make use of parallel processing for parts of their respective functions making the application quite fast.

The fourth task that this application performs is using EnrichR to map your top 20 genes to pathways. This can be performed for all of them, the upregulated genes, and the downregulated genes.

The fifth task that this app can perform is to prepare a volcano plot showing the relationship between the adjusted p-value and log fold change with regard to the top 20 genes. Hovering your cursor over points on this plot will inform you regarding what gene (HGNC symbol) a given point relates to.

The sixth task that the app can perform is to calculate the statistical power of your study, based on the number of samples in each group/class. If only two groups are present in your data, a t-test will be used. If more than two groups are present, an ANOVA test will be used.

The seventh task that the app can perform is to generate and evaluate a random forest classifier for your data using your top 20 genes as predictors. The Accuracy, AUC, Sensitivity, Specificity and Prevalence related to the model are shown in a downloadable table.

All plots have sliders and points on the plot that can be used for saving the plot (camera icon) or expanding it (a box icon).

**Prerequisites**

1.You need to have prepared a counts matrix from your raw sequencing reads. This counts matrix must be in a file that you can access on your device.

2. You need to have an accessible file containing your phenotype information.

Please be aware that the program can only look at one phenotype at a time with this app. However, after examining your data with regard to one phenotype, you can re-examine your data with regard to another phenotype.

The counts matrix and phenotype data can be in either .csv or .txt format.

**Mechanics/How to Use**

**Inputs and Main Functions**

1.On the top of the purple sidebar on the left of the interface, you will see a file browse button with the label “Input Counts” above it. Use this button to input your counts matrix. This can be in .txt or .csv format. When the file is uploaded, a blue ribbon will appear beneath the button that says ”Upload complete.”

2.Beneath that button, you will see another file browse button. This will have the label “Input Phenotype Data” above it. Use this button to input your phenotype data. This can also be either a .txt or .csv file. When the file is uploaded, a blue ribbon will appear beneath the button that says ”Upload complete.”

3.Below that you will see a button labeled “Perform Differential Expression Analysis.” Clicking this button after uploading your counts and phenotype data will cause your data to undergo quality control (removing outlier samples and lowly expressed genes), differential expression, followed by feature selection and annotation of the top 20 hits. The results of this process will be displayed in a table containing the Ensembl IDs, log fold changes, HGNC symbols, full name of the protein product and the adjusted p-value. The table can be viewed by clicking on the “ Differential Expression Results.” FYI, the tab is considered “clicked “ whenever it is bright blue. You will get a notification indicating the beginning and completion of this task.

4.Beneath that button, there is a button labeled “Display PCA Plot.” Clicking this button after uploading your counts and phenotype data will cause a PCA plot of your data to appear in the “PCA Plot” tab. Samples will be colored based on the phenotype. A legend is shown on the right of the plot. You will get a notification indicating the beginning and completion of this task.

5. To the right of that button, there is a button labeled “Display UMAP Plot.” Clicking this button after uploading your counts and phenotype data will cause a UMAP plot of your data to appear in the “UMAP” tab. Both of the plots can be saved via right clicking the plot or using the camera icon on the plot. You will get a notification indicating the beginning and completion of this task.

6. Beneath that button, there are 3 buttons that will allow you to use EnrichR to map your top genes to pathways if you click them after the table of the top 20 hits has been prepared.

The first of those buttons runs EnrichR on all 20 genes. It is labeled as "Enrich Pathways(All Genes)." The button beneath that button performs the analysis only on your upregulated genes. It is labeled as "Enrich Pathways(Upregulated Genes).” The last of those 3 buttons runs the analysis only on your downregulated genes. It is labeled as "Enrich Pathways(Downregulated Genes).” You will get a notification indicating the beginning and completion of this task.

7. The next button running down towards the bottom of the page is labeled "Display Volcano Plot." If this button is clicked after the table of the top 20 genes has been produced, it creates a volcano plot displaying the relationship between the adjusted p-value and log fold change of your top 20 genes. The plot will be shown in the "Volcano Plot” tab. You will get a notification indicating the beginning and completion of this task.

8. To the right of the “Display Volcano Plot” button there is a button labeled “Run Random Forest Analysis.” Clicking this button after running step 4 will cause the app to create a random forest classifier using your top 20 genes as predictors.” A notification will also appear on the screen when the process starts and is completed. The performance metrics of the model will be shown as a table in the “Random Forest Performance Metrics” tab. The table can be downloaded using the “Download Performance Metrics as CSV” button in the same tab.

9. Beneath the “Run Random Forest Analysis” button is a button labeled “Calculate Statistical Power.” This button will calculate the statistical power of your data if you click it after uploading your phenotype data and counts matrix. The results of the power calculation will be shown in the tab labeled “Power Calculation.” A notification will also appear on the screen when the process starts and is completed.

**Sliders**

Below button in #9, you will see sliders that will allow you to adjust parameters on the plots.

1.The first slider allows you to adjust how many PCA components are shown in the PCA plot.

2.The second slider allows you to adjust the number of UMAP neighbors in the UMAP plot.

3.The third slider moves the volcano plot along the adjusted p-value axis.

4.The fourth slider moves the volcano plot along the log fold change axis.

**Download Buttons**

1.The next set of buttons are download buttons. The first of the download buttons is labeled as “Export Results as CSV” and will allow you to download the” Differential Expression Results” table as a .csv file. **Note:** Do not click this button before the “Differential Expression Results” has been prepared.

2.The second (going towards the bottom of the page) download button is labeled as "Export EnrichR results for all 20 genes as a CSV." Clicking this button after performing pathway enrichment on all 20 of your top genes will allow you to download the results of enrichment as a .csv file.

3.The third (going towards the bottom of the page) download button is labeled as "Export EnrichR results for upregulated genes as a CSV." Clicking this button after performing pathway enrichment on your upregulated genes will allow you to download the results of enrichment as a .csv file.

4. The fourth and final download button is labeled as "Export EnrichR results for downregulated genes as a CSV." Clicking this button after performing pathway enrichment on your downregulated genes will allow you to download the results of enrichment as a .csv file.

**Notes:**

The “Perform Differential Expression Analysis,” “Display PCA Plot,” “Display UMAP Plot,” and “Calculate Statistical Power” buttons will not perform their respective tasks unless the counts matrix and phenotype file are uploaded. In the same way, the “Display Volcano Plot,” “Run Random Forest Analysis ” and the pathway enrichment buttons will not run if the Differential Expression Results table is not prepared.

You may encounter a slight lag in between the notification of plots being prepared and the plots appearing in their respective tabs. This depends on the capabilities of your computer.

This readme file can be found by clicking the “Readme” tab.