RNAseqAuto User Guide

Version 3.0

(MATLAB R2020a or later)

Ion Channels and Human Disease Laboratory,

Florey Institute of Neuroscience and Mental Health

5th of January, 2021

Created by Lauren Eldershaw

Contents

This user guide can also be accessed by clicking the help symbol ⓘ in the app.

Overview of RNAseqAuto Workflow

The RNAseqAuto workflow was created to assist users in reviewing data output by the RNA-ssequencing analysis pipeline in the form of Galaxy tabular spreadsheets. The workflow allows the user to import data, choose between analysis for gene-level or exon-level data, and calculate the reads per kilobase per million mapped reads (RPKM).

Installing MATLAB for the first time

This code is written and tested to be compatible with MATLAB version R2020a or later. No additional toolboxes are required.

Running RNAseqAuto

Note: if you have downloaded the RNAseqAuto files in a .zip file be sure to extract the files first before attempting to run RNAseqAuto as to avoid an error message.

From File Explorer

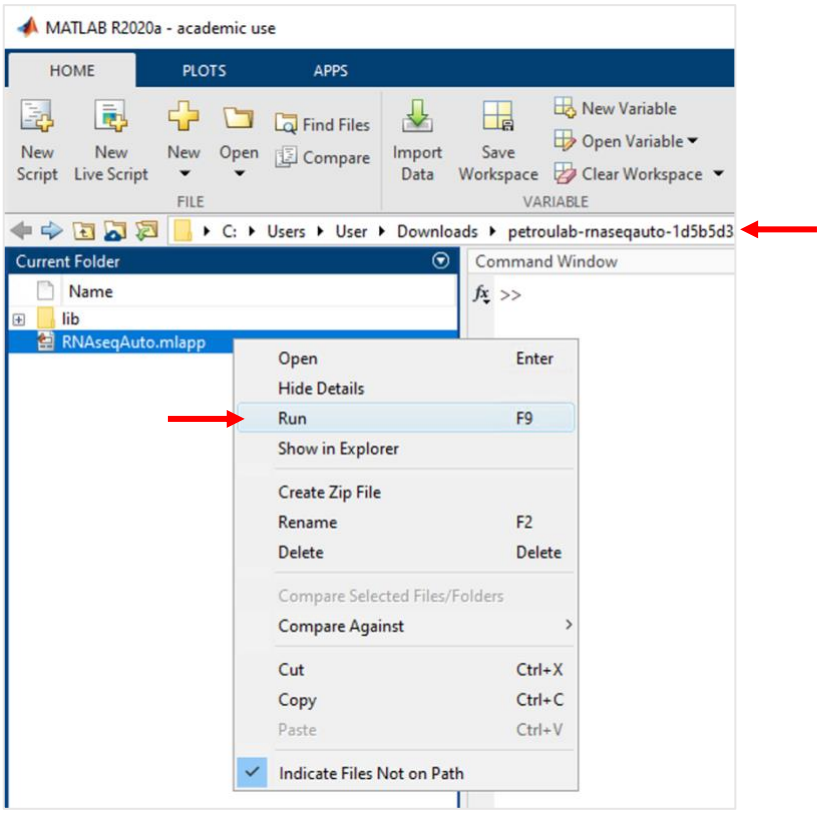
With MATLAB closed, navigate to the folder that the latest version of RNAseqAuto is stored in and double click on RNAseqAuto.mlapp to open the app.

From MATLAB

Open MATLAB – **version R2020a or later is required**.

Navigate to the folder that the latest version of RNAseqAuto is stored in by clicking on the file path section in the MATLAB window. Double click on the RNAseqAuto folder to open it. **There is no need to add the files to the path as RNAseqAuto will do this automatically.**

Right click on RNAseqAuto.mlapp and select ‘Run’:



The RNAseqAuto application will open up which requires the user to import their data file and change the necessary settings (as shown in the image below). See Input Settings for further details on how to adjust these settings to run the analysis.

Input Settings

Select Analysis Type

The user must select the appropriate analysis type as **'Gene'** or **'Exon'** data, depending on the data they are analysing to perform the correct calculations. See Run Analysis for details on the two analysis options.

Select Data Analysed Option

The user must select whether all or only select rows from the count data is to be analysed, from the options **'All data'** and **'Select genes/rows'**. Note: Both options are available in Gene-Level Analysis and only 'All data' is available in Exon-Level Analysis. See Gene-level Analysis for details on the option to select genes/rows.

Select Output Type

The user must select the output file types to be produced during the analysis. See Output File Types for details on the two output file options.

Browse for Counts + Matched Summary Files

Under **Input Files**, the user can select the gene/exon counts and matched summary files to be analysed. Clicking the '**Browse'** button brings up a dialog box for the user to select a single pair or multiple pairs of data files to be imported. The data files are to be **.tabular** files as exported from the Galaxy interface without renaming.

An example pair of counts + matched summary files:
Galaxy49-[featureCounts_on_data_7_and_data_17__Counts_(with_location)].**tabular**
Galaxy50-[featureCounts_on_data_7_and_data_17__Summary]**.tabular**

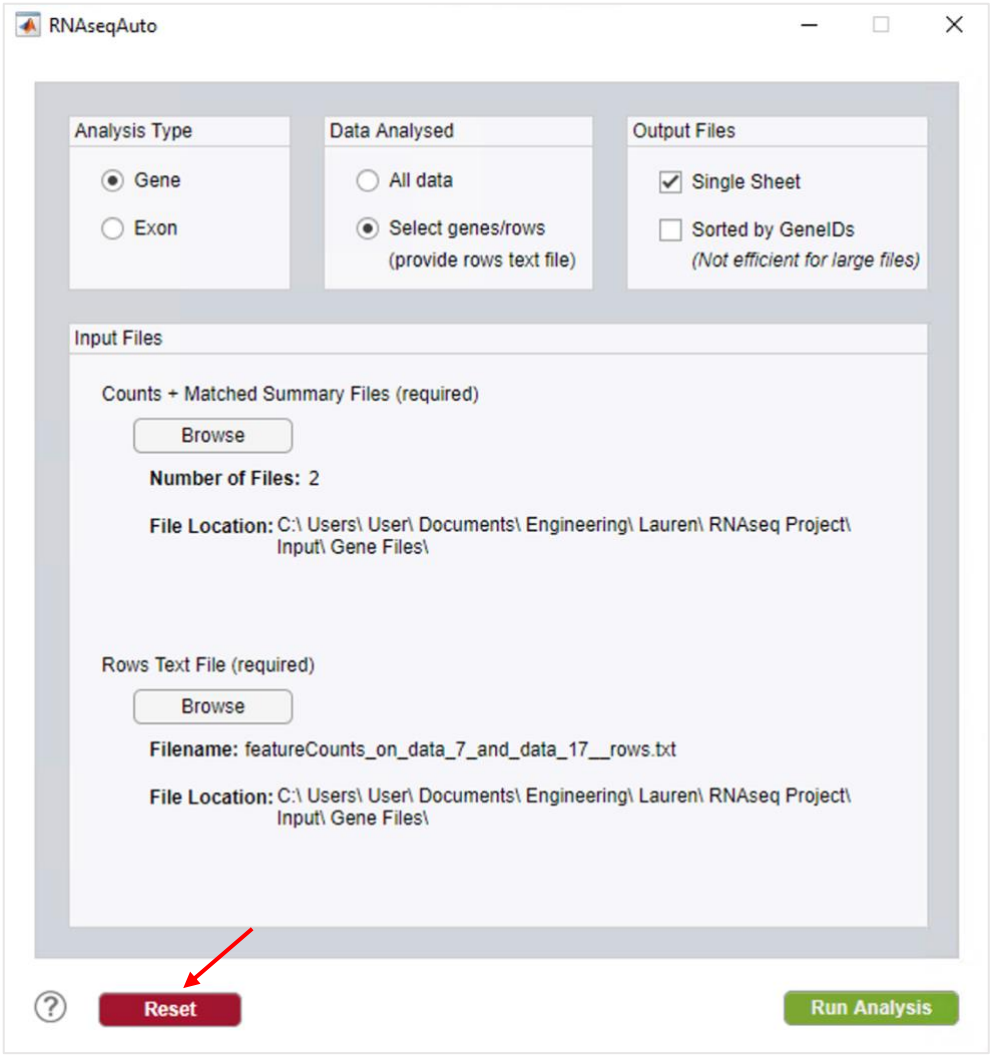
These file pairs are matched according to their dataset name shown in red in the example above. If the number of files selected is not even, and as such a full set of pairs has not been selected, an error will be thrown to notify the user (see Troubleshooting Errors for error messages).

Browse Row Text File (Select Genes/Rows option in Gene-level Analysis only)

In Gene-level Analysis there is an option to analyse only select genes/rows from the data file as listed in a separate rows text file. Under **Input Files**, the user can select the rows text file (**.txt**) which contains a list of numbers (one per line) which specify the rows in the counts file to be analysed. Clicking the '**Browse'** button brings up a dialog box for the user to select a rows file to be imported. See Gene-level Analysis for details on the data options.

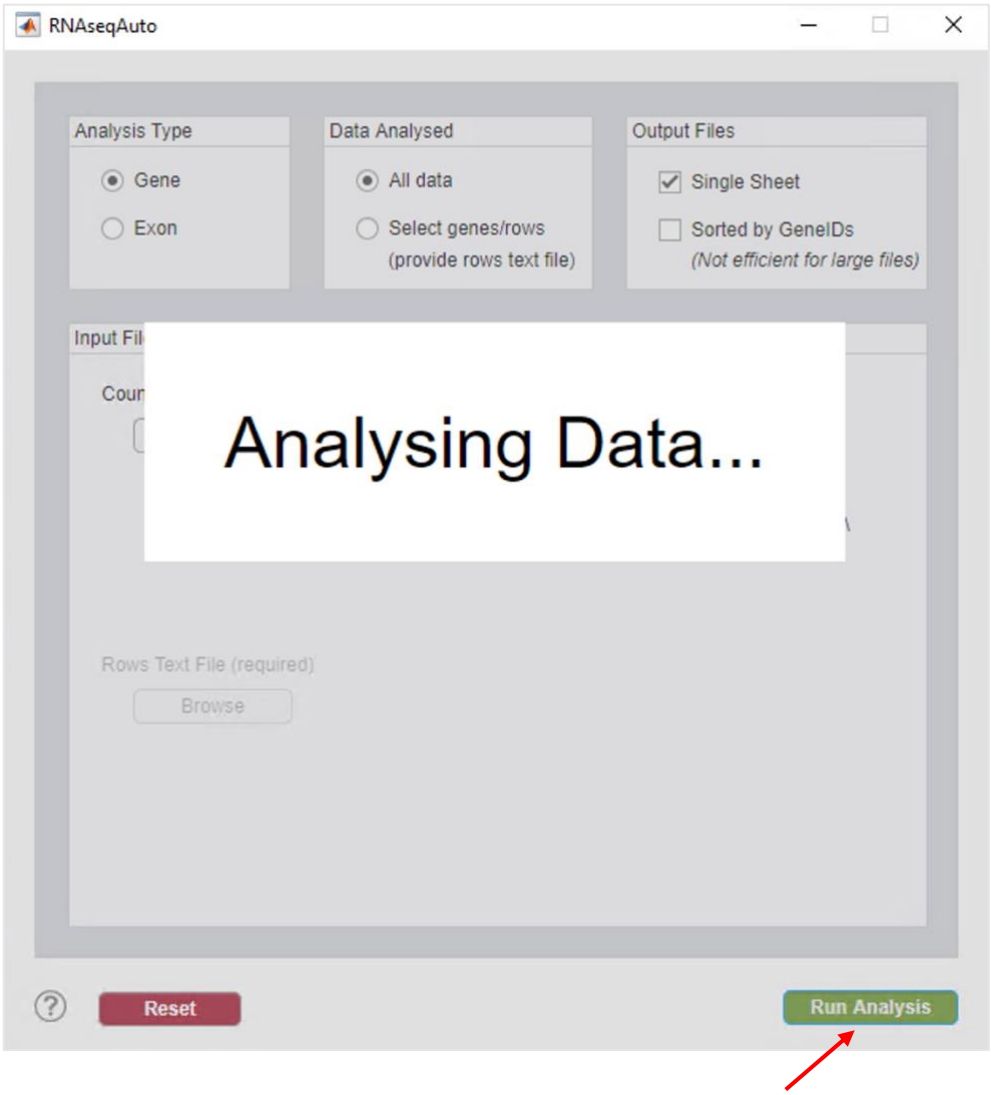
Reset Settings

Clicking the red '**Reset'** button in the bottom-left corner of the window resets the RNAseqAuto window and removes the files that have been selected.



Run Analysis

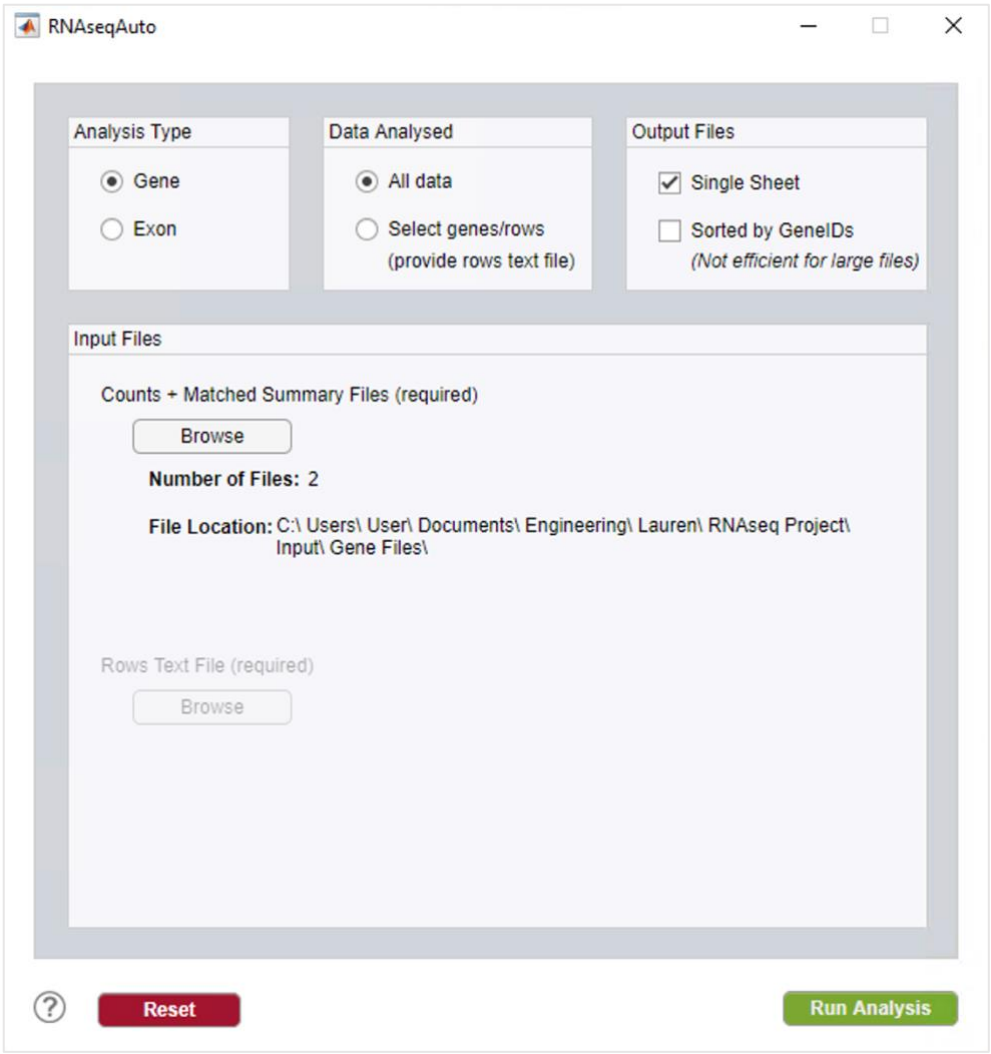
Click the green **'Run Analysis'** button to perform the analysis steps for the files selected. A loading screen will show until the analysis has concluded and the files have been created.

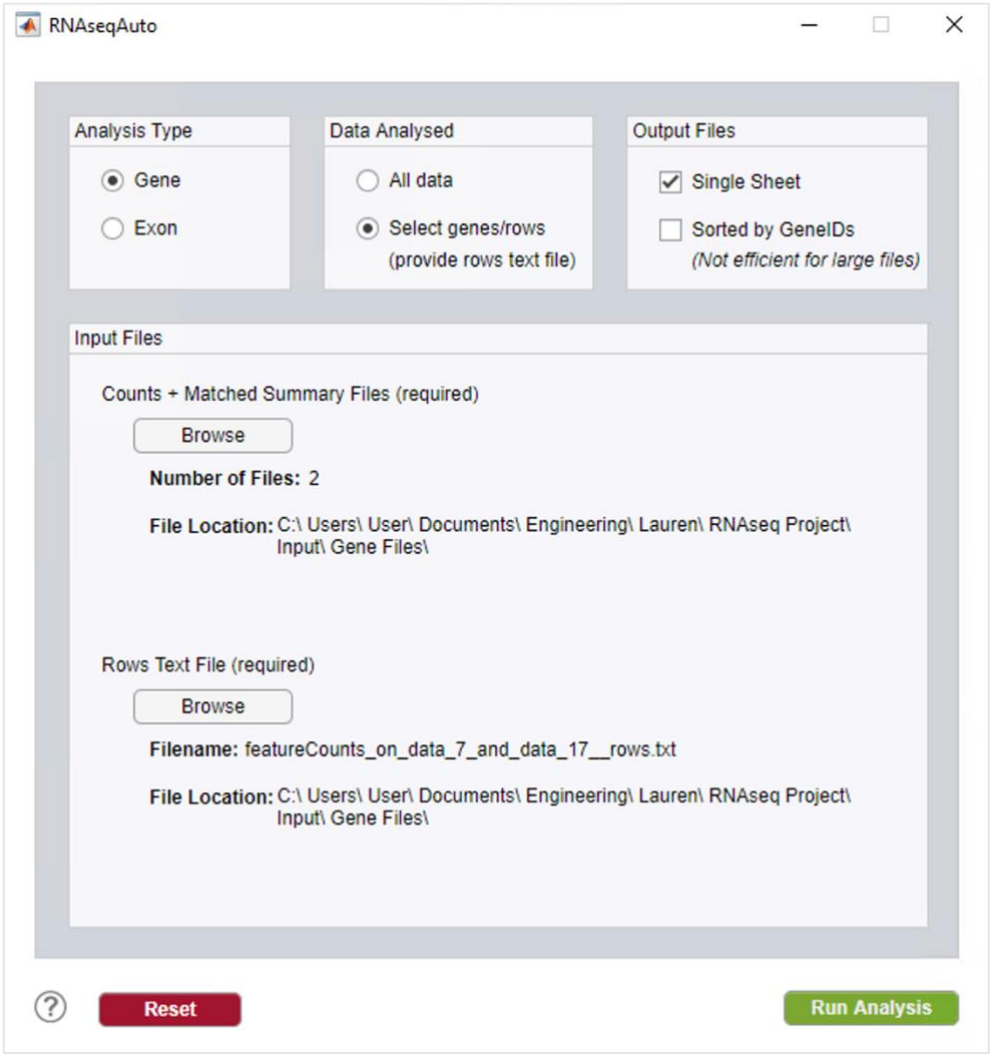


Gene-level Analysis

Analysing gene-level data requires at least one gene counts and summary file pair. This analysis type allows the user to select whether **'All data'** or only **'Select genes/rows'** in the file is to be analysed. If the 'Select genes/rows' option is chosen, a rows text file is also required which contains a list of numbers (one per line) which specify the rows in the counts file to be analysed.

During analysis, the total number of assigned reads is extracted from the summary file and used to calculate the reads per kilobase per million mapped reads (RPKM) for each of the rows (or selected rows) in the counts file. The calcuations are added to each row (including the intermediate steps) and saved into an Excel spreadsheet (.xlsx) according to the output type selected. See Output File Types for details on the two output file options.




RNAseqAuto
Analysis Type
Gene
Exon
Data Analysed
All data
Select genes/rows
(provide rows text file)
Output Files
Single Sheet
Sorted by GeneIDs
(Not efficient for large files)
Input Files
Counts + Matched Summary Files (required)
Browse
Number of Files: 2
File Location: C:\ Users\ User\ Documents\ Engineering\ Lauren\ RNAseq Project\ Input\ Gene Files\
Rows Text File (required)
Browse
Filename: featureCounts_on_data_7_and_data_17__rows.txt
File Location: C:\ Users\ User\ Documents\ Engineering\ Lauren\ RNAseq Project\ Input\ Gene Files\
Reset
Run Analysis

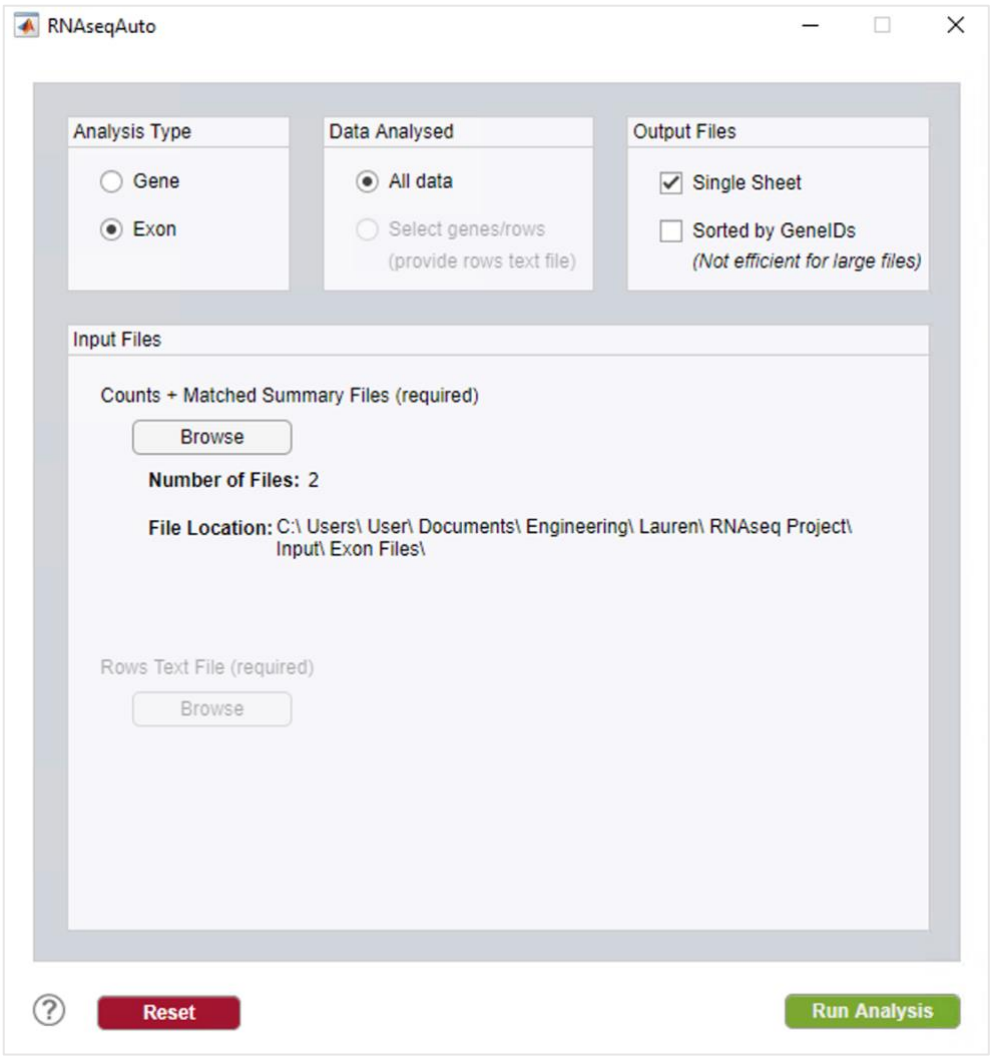
Exon-level Analysis

Analysing exon-level data only requires at least one exon counts and summary file pair. This analysis type only allows the user to analyse all the data in the file and does not allow the option to only analyse select genes/rows.

During analysis, the total number of assigned reads is extracted from the summary file and used to calculate the reads per kilobase per million mapped reads (RPKM) for each exon in the counts file. Since exon counts are separated into collections of reads with different start/stop sites, the total count for each exon is identified by grouping reads which overlap each other. The total size of the exon, or the **"super-exon length"**, is identified as the length of the exon from its earliest start site to its latest stop site.
The super-exon length and the combined number of counts (across the separate reads of the exon) are used to calculate the RPKM of each exon in each gene. The calcuations (including the intermediate steps) are added to each exon row and saved into an Excel spreadsheet (.xlsx) according to the output type selected. See Output File Types for details on the two output file options.

Average exon calculations are also performed by combining the length and counts of each exon within each gene to calculate the total gene length and read count, respectively. These calculations are recorded for each GeneID in a separate file (see Additional Output Files), including the average RPKM of the exons in the gene.

Note: The averageExon RPKM may be similar to the gene-level RPKM, however they are almost never the same (in eukaryotes). The gene-level RPKM and the averageExon RPKM are only the same when there is equal distribution of reads across all exons, however, we know that eukaryotic genes can have multiple transcript variants with alternative exons which affects the distribution of reads.


RNAseqAuto
Analysis Type
Gene
Exon
Data Analysed
All data
Select genes/rows
(provide rows text file)
Output Files
Single Sheet
Sorted by GeneIDs
(Not efficient for large files)
Input Files
Counts + Matched Summary Files (required)
Browse
Number of Files: 2
File Location: C:\ Users\ User\ Documents\ Engineering\ Lauren\ RNAseq Project\ Input\ Exon Files\
Rows Text File (required)
Browse
Reset
Run Analysis

Output Files

Each time the analysis is run a new folder is created, in the same folder as the imported data files, to hold the output files. These folders are named according to the date and time at which the analysis was run. For instance, the folder titled **'2020-04-15 12.30.00 PM RNAseqAuto'** contains the files from a RNAseqAuto analysis run on the 15th of April, 2020 at 12:30 PM.

Output File Types

There are two options for output files:

- **Single Sheet** – output_singleSheet.xlsx
  All data analysed and resulting calculations are saved in a single sheet of the Excel spreadsheet.
- **Sorted by GeneIDs** – output_sortedByGeneIDs.xlsx
  All data anlysed and resulting calculations are sorted into separate sheets of the Excel spreadsheet according to their respecitive GeneIDs.
  Note: creating the Sorted by GeneIDs file can be very slow for large data files, especially exon-level data files. The time to create this file depends on the size of the input files and the amount of memory your computer has. See Troubleshooting Errors for possible workarounds.

See Gene-level Analysis and Exon-level Analysis for more information on the output file contents.

Additional Output Files

- **Average Exon RPKM Calculations** – output_averageExon.xlsx *(Exon-level Analysis only)*
  Automatically output during Exon-level Analysis and contains the length, counts and average RPKM of each GeneID. See Exon-level Analysis for more information on the file contents.

Execution Log

The **execution_log.xlsx** file is generated and saved with the output files to provide details about the data analysed and the input settings that were used in the analysis.

It provides the:

- Analysis type (Gene or Exon)
- Data analysed option (All Data or Select genes/rows)
- Output file type (Single Sheet and/or Sorted by GeneIDs)
- Output folder – path to output folder
- Input folder – path to folder containing the input count/summary files
- List of count + matched summary files
- Row File (for Select genes/rows option of Gene-level Analysis only)

Troubleshooting Errors

Errors in the RNASeqAuto Window

"ERROR: Number of files selected isn't even (not paired)"

This error occurs when the number of files selected is not even, indicating that a full set of pairs has not been selected.

"ERROR: A row listed in the Rows Text File is not valid."

This error occurs when there is a row number listed in the Rows Text File that is not valid. Often this will occur when the row is the header (row 1).

"Invalid count files. See 'ERROR:...' in MATLAB window."

Refer to MATLAB error message "ERROR: '{dataset_name}' does not have exactly 1 count file."

Errors in the MATLAB Window

"ERROR: '{dataset_name}' does not have exactly 1 count file."

This error occurs when there is less than or more than one count file with the same dataset name as a summary file (eg. featureCounts_on_data_326_and_data_327).

"Error using writecell.
Unable to write to file '*{filename}*'. You may not have write permissions or the file may be open by another application."

This error may occur if your computer doesn't have enough memory to create Excel spreadsheets with a large number of Sheets, such the Sorted by GeneIDs output files. To overcome this, you can either use a computer with more memory or you can process your data in smaller parts instead of all at once.

Other Issues

RNASeqAuto runs for a long time and doesn't finish

This error may occur if the input files are very large and/or your computer doesn't have enough memory to create Excel spreadsheets with a large number of Sheets, such the Sorted by GeneIDs output files. To overcome this, you can either use a computer with more memory or process your data in smaller parts instead of all at once.