

TappAS: A Computational Analysis Tool for Alternative Splicing Analysis and Differential Expression

Sean Driscoll

Video tutorial:

<https://youtu.be/YZBYJhxCVwk>

1. Introduction

Recent developments in sequencing technology and computational biology techniques have allowed us to research splicing variations at a much larger scale and with higher efficiency¹. However, the functionality of alternative splicing (AS) is still being explored. In order to potentially begin to understand the functionality of AS, I will be running a differential expression analysis on RNA-Seq data from mouse neural cells in order to answer the question of what percentage of differentially expressed genes code for more than one transcript and how does this compare to the overall percentage of genes that code for multiple transcripts. In other words, are genes that code for multiple transcripts more likely to be differentially expressed in the mouse neural tissue? Generating a list of differentially expressed genes that code for multiple transcripts can then be used to explore functionality through processes like gene ontology (GO).

To solve the research question proposed above, the application tappAS was used. TappAS is a free, Java GUI application that allows the user to analyze RNA-Seq data, specifically alternative splicing, and UTR processing all from a functional perspective. It is a project-based application that uses a comprehensive set of data analysis, filtering, visualization, and *ad hoc* query tools for these analyses. The application is free for commercial use and will run on most computers that have enough storage and computational resources. This manual will provide downloading and setup instructions for the application, along with a walkthrough on how to run a differential expression analysis on RNA-Seq data. For more documentation, refer to the published paper in Genome Biology².

The RNA-Seq dataset being used is built into the application that can be used as a demonstration dataset but is still able to be used to answer the research question proposed. It will be automatically downloaded with the tappAS application, but can also be downloaded from the “Availability of data and materials” section of the original paper². The dataset contains RNA-Seq data from two control samples of mouse neural stem cells (NSC) and two samples from mouse oligodendrocytes (OLD).

The minimum input files required are explained in more detail in section 4, but you will need both an experimental design and matrix data file (both in .tsv file format), plus an annotation features file for your specific species.

2. Download and install:

Download and install is the most difficult part of the application, so it will be explained in detail Below. To run the application there are a few requirements:

Operating System (One of the Following):

- Most recent version of Linux Ubuntu
- Most recent version of MAC OS X (used in this manual)
- Windows Version 7, 8, or 10

Computer Hardware:

- Minimum of 4 cores
- Minimum of 8 GBs of RAM
- Minimum of 20 GBs of available and unused disk space

Java:

- Java version 8 with update 40 or later. If a more recent version is used, it may fail to run the application

Software and Packages:

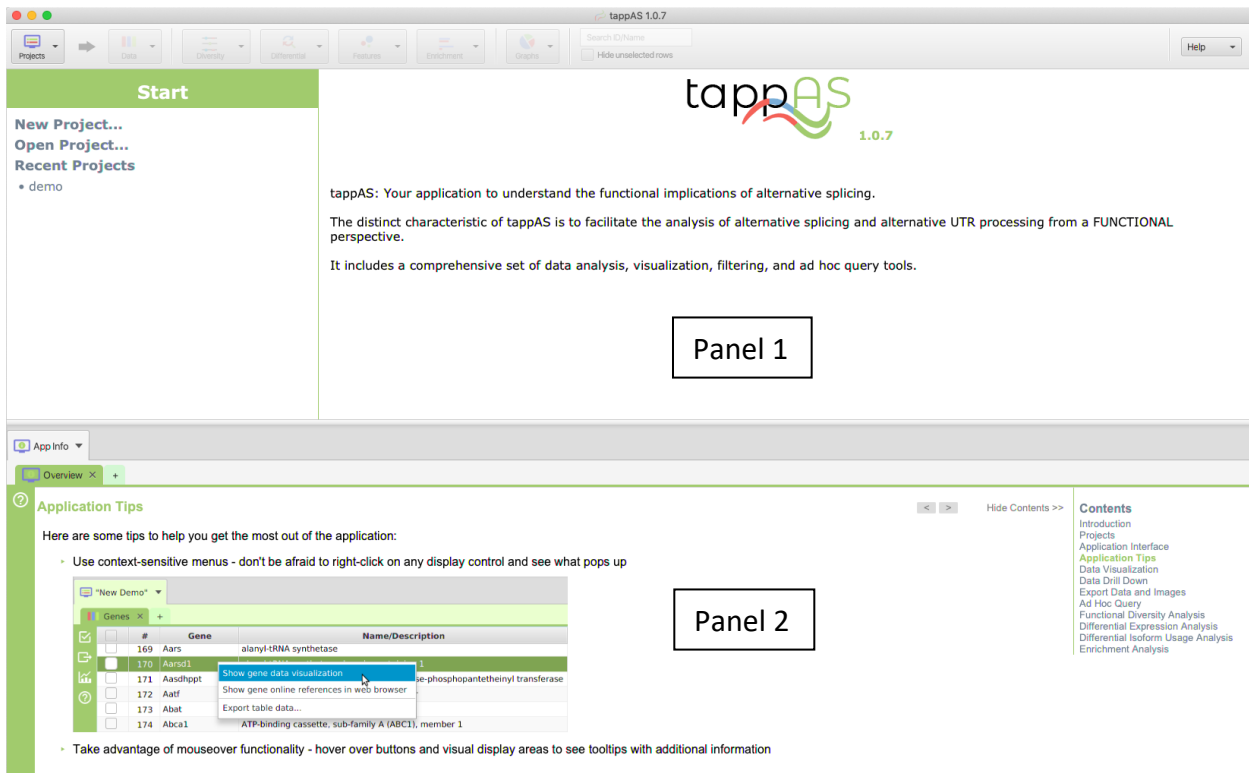
- R-script version 3.6.1 or later
- Many R packages are needed and if they are not previously installed, a pop-up message will appear and ask if you want to automatically install them. This may be very time consuming and depending on the computer hardware, it may fail. To avoid this, the creators recommend using the link below to automatically install the packages (enter into the R terminal). This may cause some issues for first time users.

```
o source("http://app.tappAS.org/resources/downloads/
  tappAS_packages.R")
```

- Note: The R version used is not RStudio and must be the actual R terminal

3. Launching the Program:

Double click on the **tappAS.jar** file to run the program. When it opens, tappAS will ask you where you want to store the applications data folder (automatically titled: “tappASWorkspace”). A common and recommended storage place would be the home directory or the desktop. When you select the file path and open the application, you will see the main home screen (a screenshot is shown below):



This is where all of your projects are stored and can be opened. The top panel is the main screen where you can select a project to open while the bottom panel shows application tips and content of the application (to the right of the panel).

4. Adding a Project

When you have opened the application and downloaded all appropriate packages, you are now ready to add a project and import your data. In the top left of the top panel, click add project. You will now be met with the “new project” window. A screenshot of the window is shown below.

1 Project Information

Name:

2 Annotation Data

Species:

☒ Use application's annotation features file - requires download first time used

☐ Specify annotation features file's location

3 Experiment Design and Data

Experiment Type:

Design File:

Matrix Data File:

Note: First group in design file must be the control and expression matrix MUST contain raw counts - see Help

☒ Normalize data by TMM method (recommended for raw counts)

☒ Expression Values Filter (recommended):

Low count values cutoff: CPM

Coefficient of variation cutoff: %

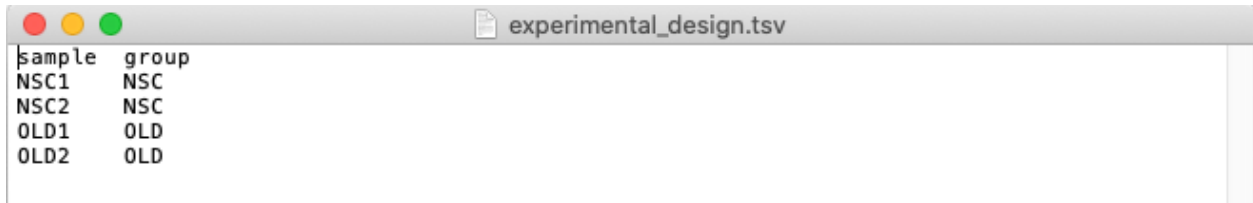
Transcripts Filter:

Options for Adding Project:

There are many different options you need to choose that depend on the type of data you have and the type of experiment that was done. Since the dataset used for this project is already built into the application, a detailed order of steps that correspond to the numbers in the screenshot above is outlined below for what inputting outside data would look like:

- 1) Chose the name for the project
- 2) Select the species you have run your RNA-Seq on and the chose the appropriate annotation features file. The tappAS application comes with annotation files for 5 different species: *Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Mus musculus*, and *Zea mize*. However, you are able to upload your own GFF3 annotation file if you have one for a species not mentioned above. You may also request one to be added on the tappAs website.
- 3) Chose the experiment type and upload required files. The format and explanation of what is required for this section is shown below:
 - a) **The experiment type.** You have 3 options to chose from depending on what the design of your experiment was: a case control, time-course single series, and time-course multiple series.
 - b) **The experimental design .tsv file.** This file will outline your experimental design with your control groups and replicate groups (example below).

- c) **The matrix data .tsv file.** This is your data file from the RNA-seq experiment. It must be in .tsv format (as with the other files) and must contain raw expression counts for one or more experimental groups (screenshot shown at end of section).
- d) The next two check boxes are for 1. normalizing the data by TMM method and 2. filtering the expression values by minimum count and coefficient of variation cutoffs. Both of these are recommended by the tappAS program.
- e) Once you are done selecting the project settings, click “OK”.



sample	group
NSC1	NSC
NSC2	NSC
OLD1	OLD
OLD2	OLD

Example of experimental design .tsv file



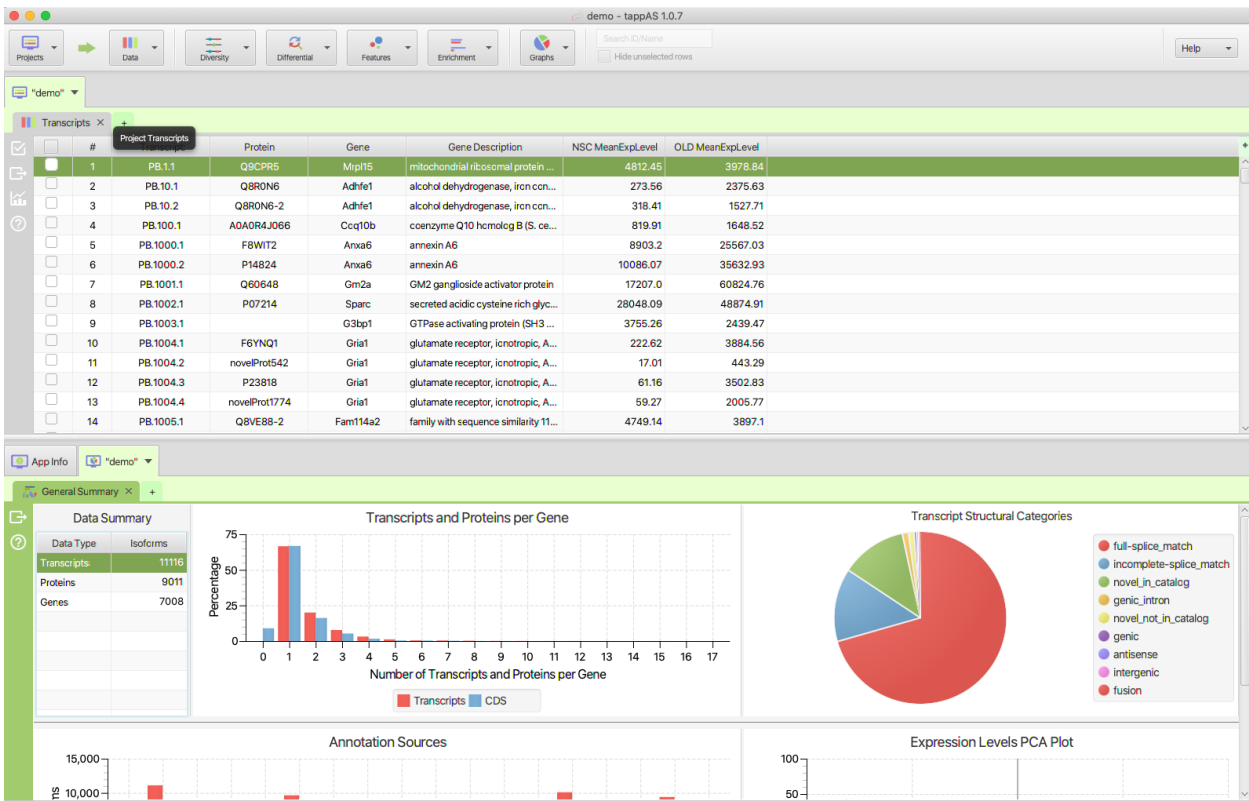
NSC1	NSC2	OLD1	OLD2
PB.1668.1	13090.25	8635.61	5911.69 6028.91
PB.972.1	2908.91	2220.34	2890.57 2836.8
PB.5017.1	3833.06	2861.24	2889.58 2544.89
PB.5017.2	691.37	353.12	81.23 148.82
PB.5017.3	3755.89	755.2	1005.15 795.3
PB.5017.4	416.69	250.44	323.03 436.99
PB.1463.1	1549	1138	459 464
PB.6522.1	1620.94	1134.45	1075.96 1427.3
PB.6522.2	20076.97	13833.76	11185.5 10275.35
PB.6522.3	2804.09	1569.78	1537.54 1282.35
PB.1565.1	1831	909	1191 1079
PB.7650.1	2317	1519	1987 2047
PB.7238.1	654	665	3970 4016
PB.4812.1	3138	2164	2348 2285
PB.3975.1	4374.77	2052.29	5177.9 4977.11
PB.3975.2	296.36	145.64	399.8 356.55

Example of an expression matrix .tsv file

Note: To access the dataset used in this project, choose the *mus musculus* for species and then for the annotation file, chose “demo 1”. This should automatically fill out the rest of the settings and no further inputs are required. Examples of the experimental design and expression matrix are shown above and in the video tutorial.

5. Viewing your data

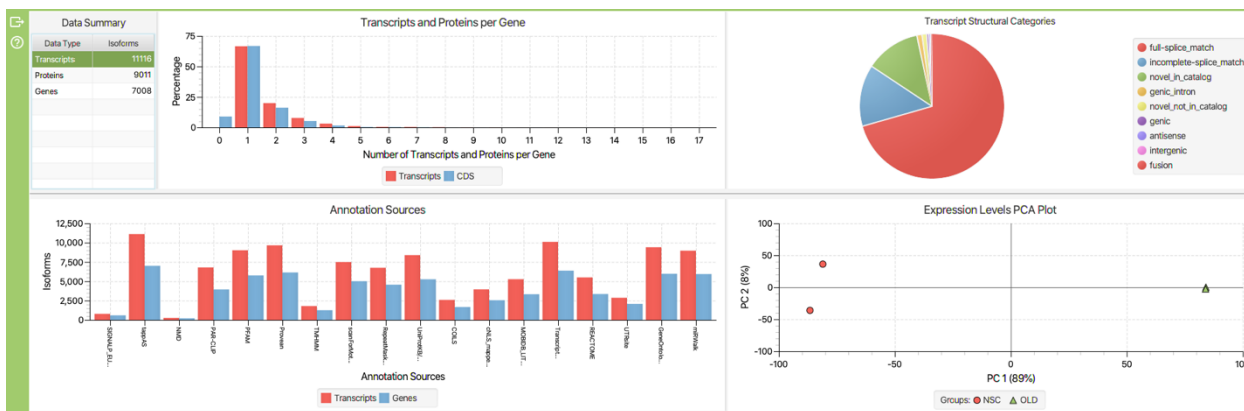
The process of uploading your data into the project may take a while but once it is uploaded (if it is successful) the application should look similar to the following screenshot:



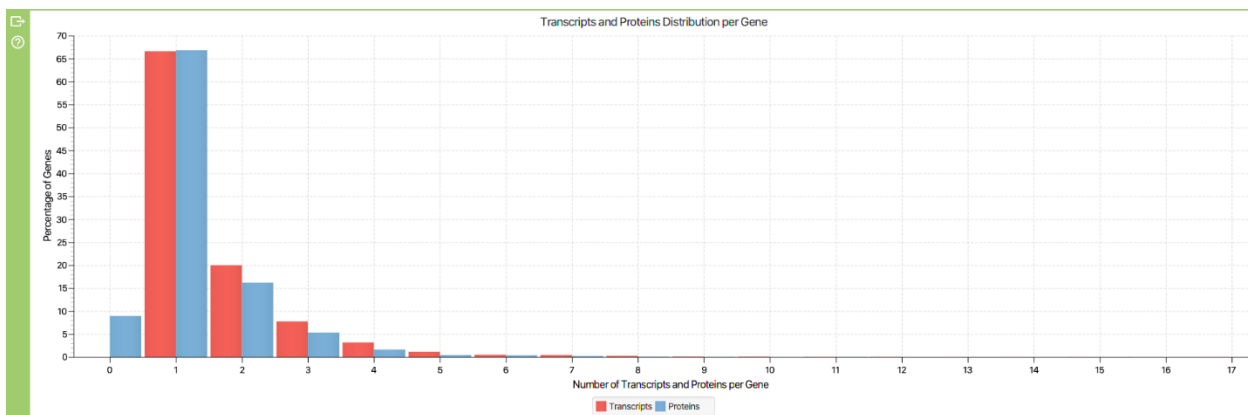
The top panel will show you a list with the following information (from left to right): the transcript, protein, gene, gene description, the mouse neural stem cells (NSC) mean expression level, and the mouse oligodendrocytes (OLD) mean expression level. The bottom panel will show a general summary of the data.

6. General Data Summary:

A screenshot of the lower panel with the general summary of the mouse neural tissue RNA-Seq data is shown below. The number of transcripts, proteins, and genes is shown on the upper left. A bar graph of the transcripts per gene (as a percentage) is shown to the right of that. A pie chart of the different categories of the transcript structures is shown in the upper right with a principal component analysis (PCA) plot shown below. Lastly, there is a bar graph showing the annotation sources of the genes (blue) and the genes (red) with the bars corresponding to the isoform number from that specific source.



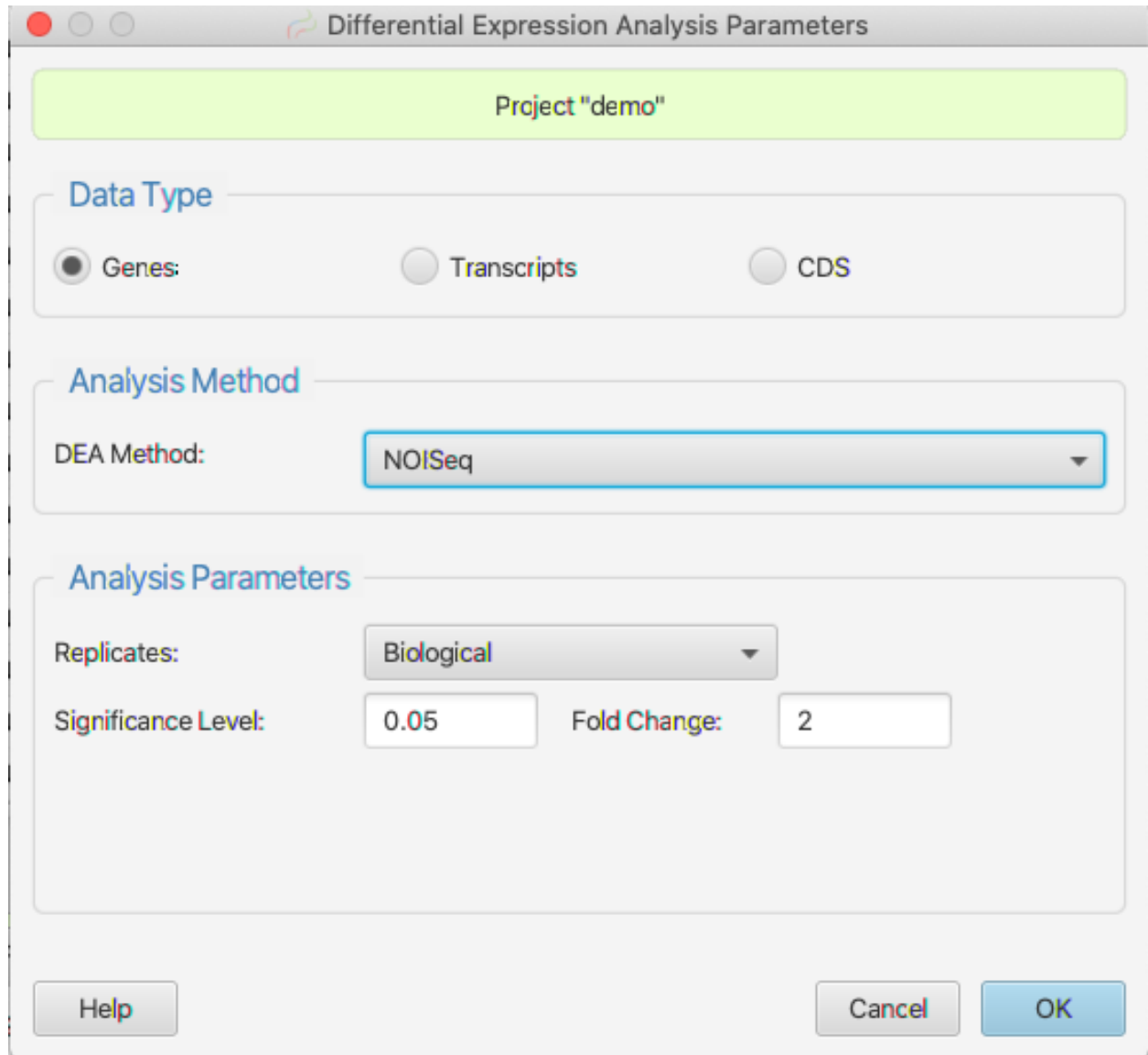
For the purpose of answering my research question, I will first focus on the “transcripts and proteins per gene” chart (zoomed in screenshot shown below). From this chart, we can see that about 65% of the genes code for one transcript. This leaves about 35% of the genes in this dataset that code for more than one transcript. This answers the first part of our research question: what is the percentage of genes in the dataset is that codes for more than one transcript? In the next section, I will be performing a differential expression analysis.



7. Performing a Differential Expression Analysis (DEA):

In the upper panel of the TappAS program, there are many options for different analyses you can perform. By selecting “differential”, you will see a drop-down menu that will allow you to choose two different kinds of differential analyses: a differential expression analysis, and a differential isoform usage analysis. For the scope of this project, we will be

choosing the former. You will then be given the parameters to be set. A screenshot of the parameters is shown below. For the “data type”, you can choose whether you want to perform the analysis on the genes, transcripts, or CDS’s. You can run all three or any combination, but each analysis will have to be independently. I chose to run the analysis on the genes and transcripts for this project. I kept the rest of the parameters to the default settings, as recommended, and ran the analysis. This may take a while. It took about 10 minutes in this manual.



The screenshot shows a macOS-style dialog box titled "Differential Expression Analysis Parameters". At the top, a light green bar displays "Project 'demo'". Below this, the "Data Type" section contains three radio buttons: "Genes" (selected), "Transcripts", and "CDS". The "Analysis Method" section features a dropdown menu labeled "DEA Method:" with "NOISeq" selected. The "Analysis Parameters" section includes a "Replicates:" dropdown menu set to "Biological", a "Significance Level:" text input field containing "0.05", and a "Fold Change:" text input field containing "2". At the bottom, there are three buttons: "Help", "Cancel", and "OK".

Differential Expression Analysis Parameters

Project "demo"

Data Type

☒ Genes ☐ Transcripts ☐ CDS

Analysis Method

DEA Method: NOISeq

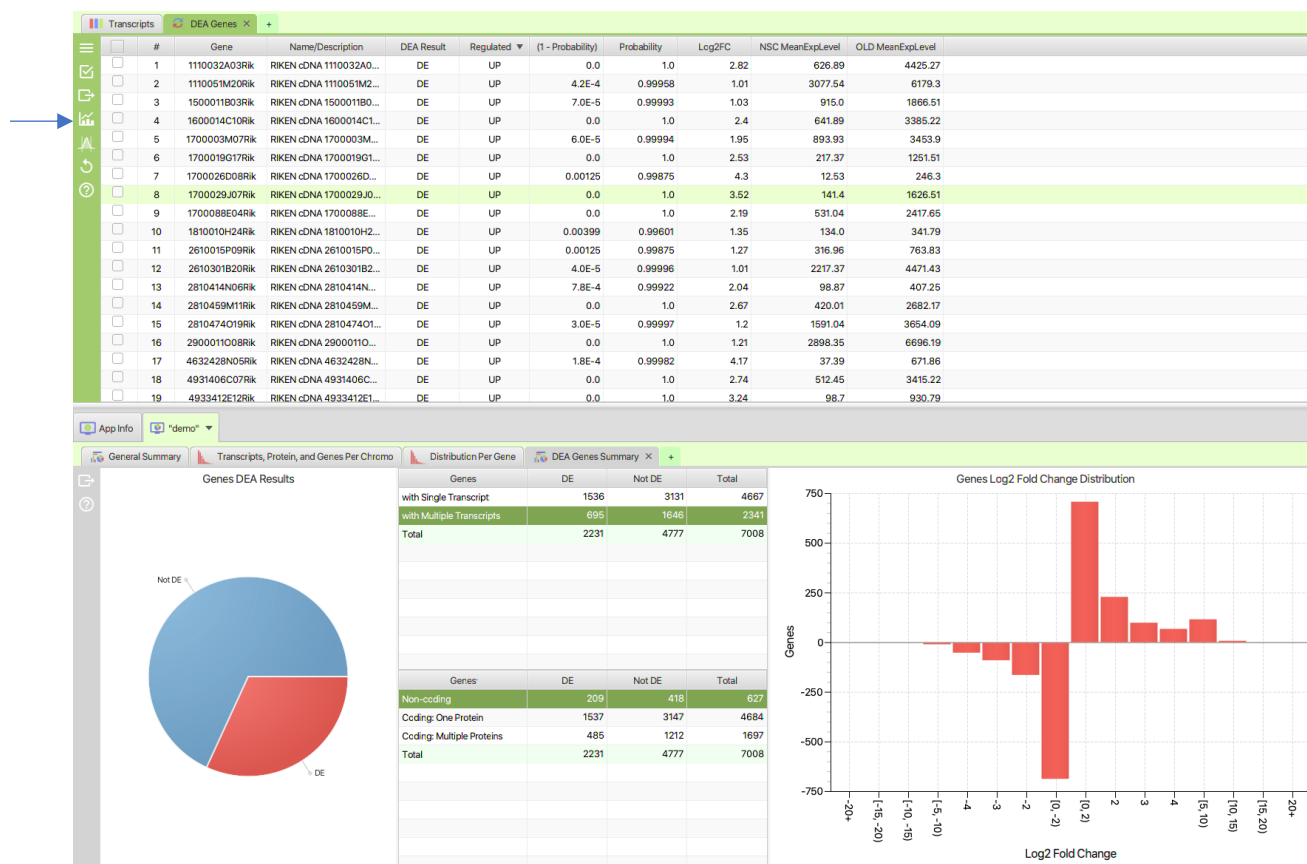
Analysis Parameters

Replicates: Biological

Significance Level: 0.05 Fold Change: 2

Help Cancel OK

When the analysis is done, you will be met with the following screen on the application. If the summary does not appear, you can click on the “data visualization options” tool on the left side of the screen (indicated by the blue arrow in the following screenshot).



What you have in the top panel (from left to right) is your list of genes, the name/description, the DEA result (either “DE” for differentially expressed or “not DE” for not differentially expressed), whether it was upregulated, downregulated, or had no change, the probability scores, the Log2 fold change, and the mean expression levels from the two groups. In the lower panel to the left, you have the percentage of genes overall that are DE vs not DE and, on the right, a bar graph of the number of genes from the Log2 fold change. In the middle of the two figures, we have a table with the number of genes DE, not DE, and the total. The upper portion of the table shows the genes with a single transcript, genes with multiple transcripts, and the total. The bottom portion of the table shows the number of genes that are DE or not DE that code for one protein, multiple proteins, or are non-coding. From the results of the DEA, we can see that about 31% of differentially expressed genes code for multiple transcripts, along with about 22% of those genes coding for multiple proteins. The percentage of genes that code for multiple transcripts that is DE (31%) is lower but close to the overall percentage of genes that code for multiple transcripts that I showed in the previous section (35%). To test whether there are fewer DE genes that code for multiple transcripts than expected by chance, the results were exported for further analysis in R. Exporting the results of the DEA can be done by clicking the

“export data/figures” to the left of the upper panel. This function allows us to use the exported .tsv file in R for analysis. I decided to run a chi-squared goodness of fit test for this analysis to see the contribution of genes that undergo AS to the set of DE genes. After running the chi-squared test, I got p-value of 2.2e-16. This shows that there is a significant relationship between transcript count and AS, and that there are significantly fewer than expected genes that undergo AS that are also DE.

8. Troubleshooting

These are some of the issues I ran into when opening the tappAS program and how I fixed them. For additional troubleshooting, visit the tappAS install pdf at:

<https://app.tappas.org/resources/downloads/install.pdf>

a. One of the main issues users run into is not having the proper installation hardware. Make sure you have Java version 8 (other version will not work) and windows 7, 8, and 10. Windows 11 will not work. To check the version of Java, open a terminal and type in the code:

```
java -version
```

This should output the current version of Java that you have.

b. Installation of the required R packages can take a very long time (30 minutes to multiple hours). Make sure you leave plenty of time for proper packages to download. A way to avoid this is by putting the following code in to the R terminal (reminder, it is the R terminal and not R studio):

```
source("http://app.tappAS.org/resources/downloads/tappAS_packages.R")
```

*Some packages from this source may be out of date and not available for download. If you find the name of the package that is not downloaded and search for it in Bioconductor, you should be able to find the code necessary to download the up-to-date package. This will make the opening of tappAS a lot quicker and smoother.

Citations:

1. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet* **102**, 11–26 (2018).
2. de la Fuente, L. *et al.* tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biology* **21**, 119 (2020).

