ProViz User's Guide

Contents

ProViz

ProViz is a data visualization and analysis tool developed at SomaLogic. ProViz imports an ADAT file (SomaLogic's data file format) and allows users to perform various exploratory data analytic processes:

- Filter the ADAT to only the samples with desired characteristics
- Merge additional sample data with the data contained in the ADAT
- Create groups of data through aggregating samples or splitting at specific data points
- Create interactive boxplots, CDFs, and scatter plots
- Perform basic statistical tests including correlation, t-test, U-test, KS-tests, ANOVA, and Kruskal_Wallis.

License

Chapter 1

Installation and Prerequisites

1.1 Installation

1.1.1 GitHub

ProViz is a Shiny application written in R. The source code is available on GitHub in SomaLogic's ProViz repository.

To use ProViz, it is recommended to install RStudio, clone the ProViz repository, and run it using the RStudio IDE. See RStudio's Shiny page for details of using RStudio to run a Shiny application. RStudio's Shiny Server can also be used to run ProViz from a server, providing access through a managed web portal.

Some additional packages are required for ProViz. Following is a list of the required packages and the versions used during ProViz development and testing. All of the following packages are available through The Comprehensive R Network - CRAN.

- shiny (1.6.0)
- shinydashboard (0.7.1)
- shinyWidgets (0.6.0)
- DT (0.17)
- dplyr (1.0.5)
- ggbeeswarm (0.6.0)
- ggplot2 (3.3.3)
- magrittr (2.0.1)
- plotly (4.9.3)
- readr (1.4.0)
- tidyr (1.0.2)

Additionally, ProViz uses the SomaDataIO v5.0.0 available from SomaLogic's SomaDataIO GitHub repository.

1.1.2 DockerHub

A Docker image will be available soon.

1.2 Example Data

The example data used in this tutorial can be found at SomaLogic's SomaLogic-Data GitHub repository where you can find a description of the file and the ADAT file format. The `example_data.adat` file consists of 192 samples (including clinical samples and controls) and is representative of a typical ADAT file SomaLogic customer's receive.

1.3 ProViz Navigation

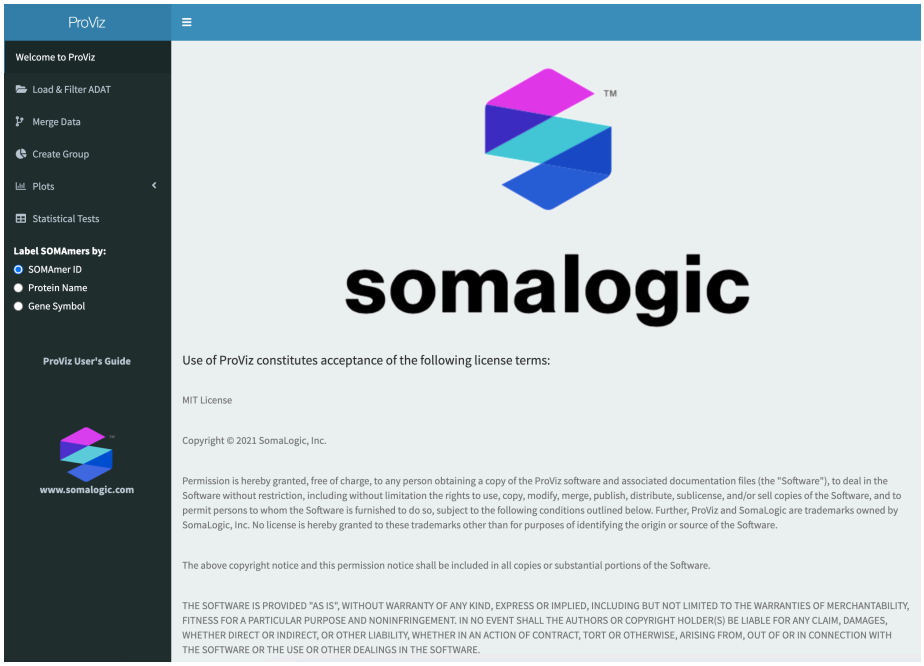


Figure 1.1: The main ProViz panel.

In ProViz, you can navigate to different panels using the navigation pane on the left by clicking on the name of the panel. Each chapter in this tutorial gives a brief tutorial on how to use the options in the corresponding panel.

Users can select how the SOMAmer-associated data will be listed in selection boxes across all ProViz panels by selecting an option under **Label SOMAmers by**: *SOMAmer ID*, *Protein Name*, or *Gene Symbol*. The latter two options are most likely more familiar to most users than SOMAmer IDs. Label types can be changed at any point when using ProViz.

Chapter 2

Load and Filter ADAT

2.1 The Load and Filter ADAT Panel

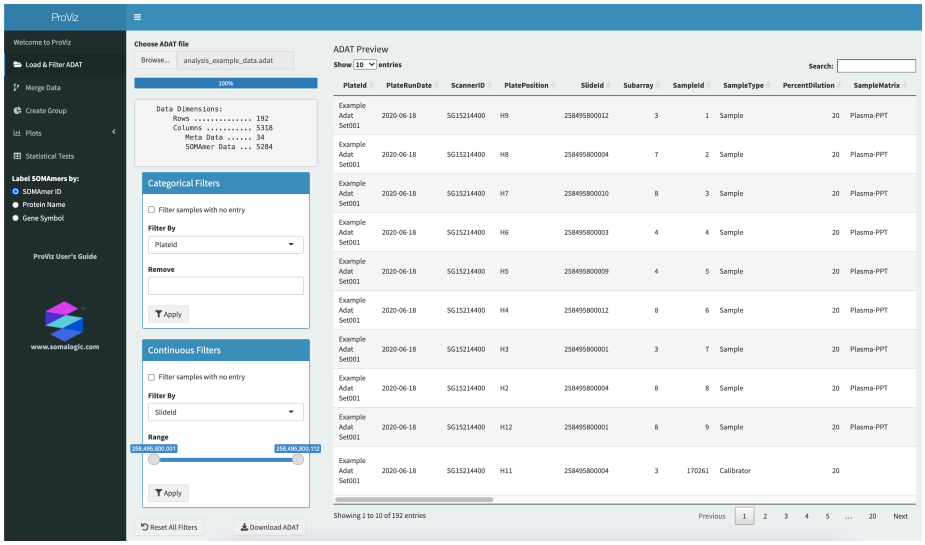


Figure 2.1: The ProViz Load and Filter ADAT panel after opening the `example_data.adat` file.

2.2 Load an ADAT file

To open an ADAT file, click on the **Browse...** button and locate an ADAT file. Once selected, the ADAT file will be opened and processed, and a preview of the content will be shown in the table at the right. ADAT files may have over

7,000 columns of SomaScan Assay data for every row in the file, so large files may take a few moments to load.

After the ADAT file is loaded, details of the file content are displayed. Here we see that there are 192 total rows (corresponding to individual samples), and 5,318 data columns. Of those data columns, 5,284 are SOMAmer data columns and 34 are Meta Data columns. These Meta Data columns contain assay-related data such as the Plate ID, Scanner ID, normalization data, and any sample-specific data that was submitted to SomaLogic with the samples. For the `example_data.adat` file, Sex and Age are included as well as various assay-related content.

The ADAT Preview table can be scroleld as well as searched for specific content (**Search:** box in upper-right). SomaScan Assay data are not displayed in the preview table.

2.3 Filter ADAT contents

Samples present in the ADAT file can be filtered based on Meta Data or SomaScan Assay data. When the ADAT file is processed, data columns are divided into Categorical (non-numeric) or Continuous (numeric) categories.

All filtering operations are cumulative.

2.3.1 Categorical Filters

Samples can be filtered from the ADAT based on categorical data using the Categorical Filters box. The column name can be selected from the **Filter By** selection box, which populates the **Remove** selection box with the unique categorical values in the ADAT file. Data values can be selected by clicking on one or more values in the **Remove** selection box. Once all selections are made, clicking the **Apply** button will remove those samples from the data, and the Data Dimensions box will be updated with new information.

In this example, the *SampleType* column is selected and the unique entries for this column are shown: Buffer, Calibrator, QC, and Sample. By clicking on Buffer, Calibrator, and QC, to select them, the assay control samples can be removed from the data file.

Checking the **Filter samples with no entry** box will remove all samples lacking a value for the selected column.

2.3.2 Continuous Filters

Samples can be filtered from the ADAT based on continuous data using the Continuous Filters box. The column name can be selected from the **Filter By** which will update the **Range** slider with the minimum and maximum values for that data column. By moving the minimum and maximum sliders, a desired

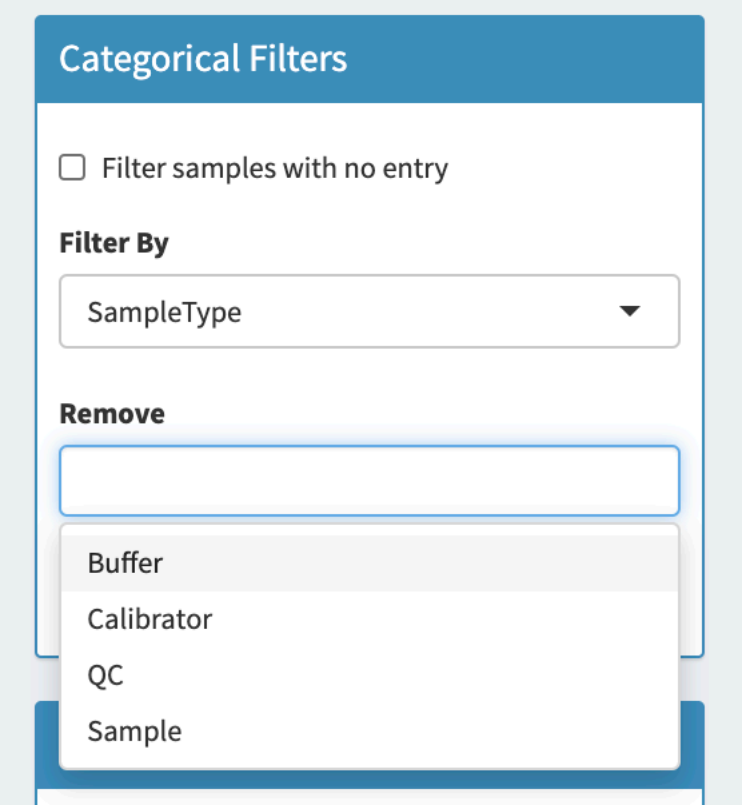


Figure 2.2: The categorical filters box.

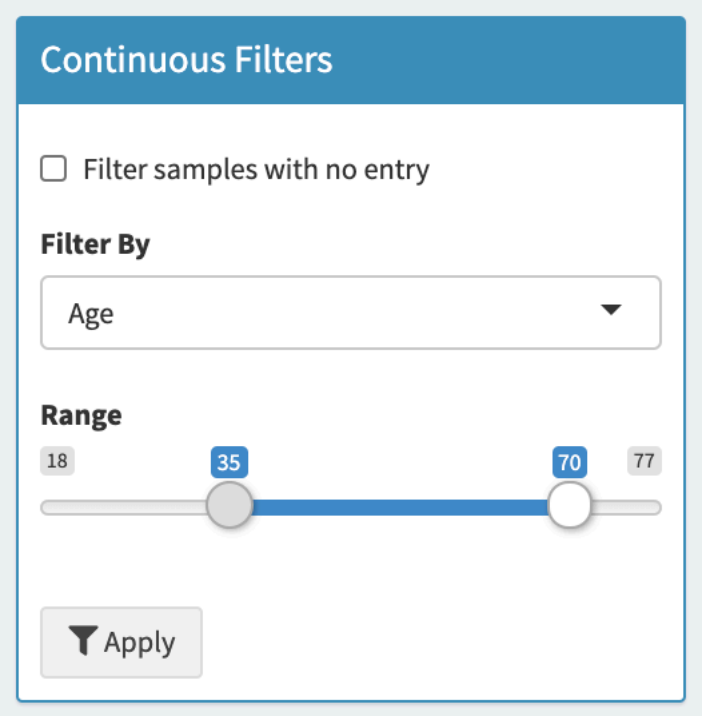


Figure 2.3: The continuous filters box.

range of the selected variable can be chosen and all samples with values outside that range will be removed.

In this example, the *Age* column is seleted and the sliders adjusted to keep only those samples with Age from 35 to 70.

Checking the **Filter samples with no entry** box will remove all samples lacking a value for the selected column.

2.3.3 Reset Filters

The original ADAT contents are always preserved and can be retrieved by clicking the **Reset All Filters** button.

2.3.4 Download ADAT

Once filtering has been performed, a version of the ADAT can be downloaded by clicking the **Download ADAT** button. As this new ADAT file may be significantly modified relative to the original, ensure that a new, unique, descriptive file name is specified so that the original ADAT is not overwritten. Keeping good notes regarding the filtering operations is essential to recall how the ADAT was modified when it is returned to at a later date.

Chapter 3

Merge Data

Data stored in external files can be merged with the proteomic data in an ADAT file using the Merge Data panel.

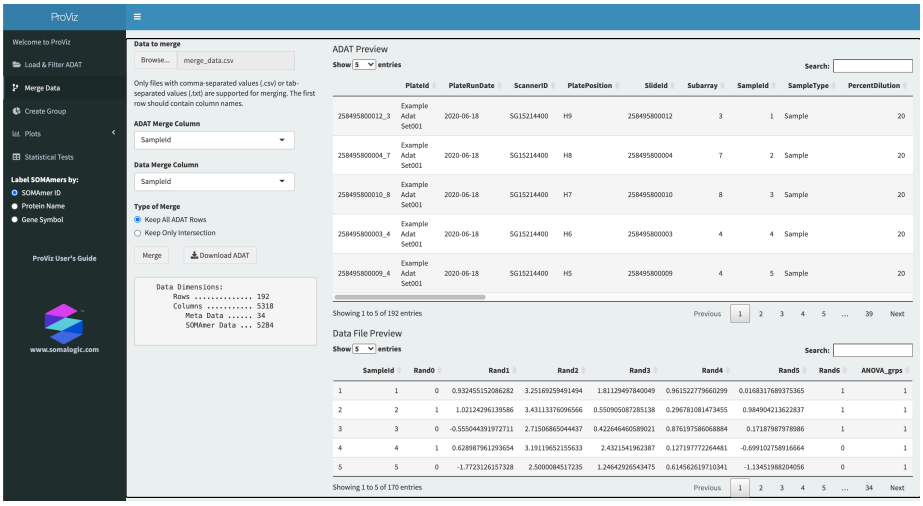


Figure 3.1: The Merge Data panel

3.1 Selecting a Data File to Merge

External data must be stored in a comma-delimited or a tab-delimited file, and the first row should contain column names. The file containing the data can be uploaded to ProViz by selecting the Browse button and navigating to the file. Once uploaded, a preview of the data file will be displayed below the ADAT Preview.

3.2 Selecting Columns in the ADAT and the Data File

In order to merge external data with the data in the ADAT file, each file should have one column that provides a way to match rows between the two files. Ideally, the columns should have all unique values and should match one-to-one between the ADAT file and the external data file. The target column in the ADAT can be selected with the **ADAT Merge Column** selection box, and the column in the external file can be selected with the **Data Merge Column** selection box. The columns do not need to have the same name.

In this example, each file has a column titled SampleId. All entries in this column in the external file are unique and correspond to the clinical samples' SampleId values in the ADAT.

3.3 Specifying the Type of Merge

The ADAT file typically contains control samples for which external data is not likely available, or the external data file may not have data for all clinical samples. Choosing *Keep All ADAT Rows* under **Type of Merge** will retain all rows in the ADAT file and include *NA* for those rows not found in the external file. Alternatively, only those rows found in both files can be retained by selecting *Keep Only Intersection*.

3.4 Merge

Once all selections are made, initiate the merge by pressing the **Merge** button. If an error occurs, a message will be displayed in the message box at the bottom of the panel, otherwise, the message box will contain updated data dimension information. The message box will provide the exact error message from the underlying R code, and may not be easily interpreted. Typically, if an error occurs, it is due to selecting columns in either the ADAT or the external file that are not campatible.

3.5 Download Adat

Once merging has been performed, a version of the ADAT can be downloaded by clicking the **Download ADAT** button. As this new ADAT file may be significantly modified relative to the original, ensure that a new, unique, descriptive file name is specified so that the original ADAT is not overwritten. Keeping good notes regarding the filtering operations is essential to recall how the ADAT was modified when it is returned to at a later date.

Chapter 4

Create Group

New groups can be created from existing data. Samples labeled with categorical data can be combined to create only 2 groups, or continuous values can be split to create 2 groups.

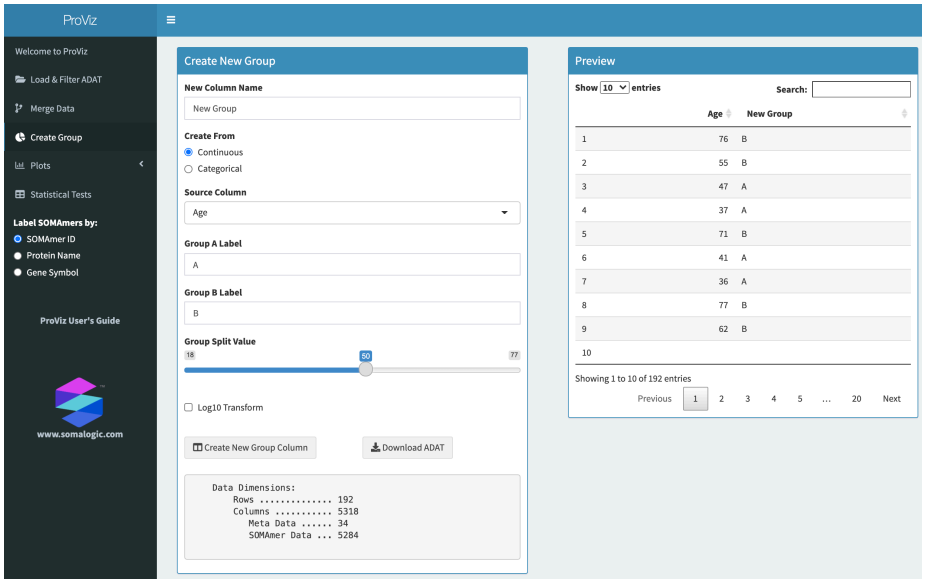


Figure 4.1: The Create Group panel.

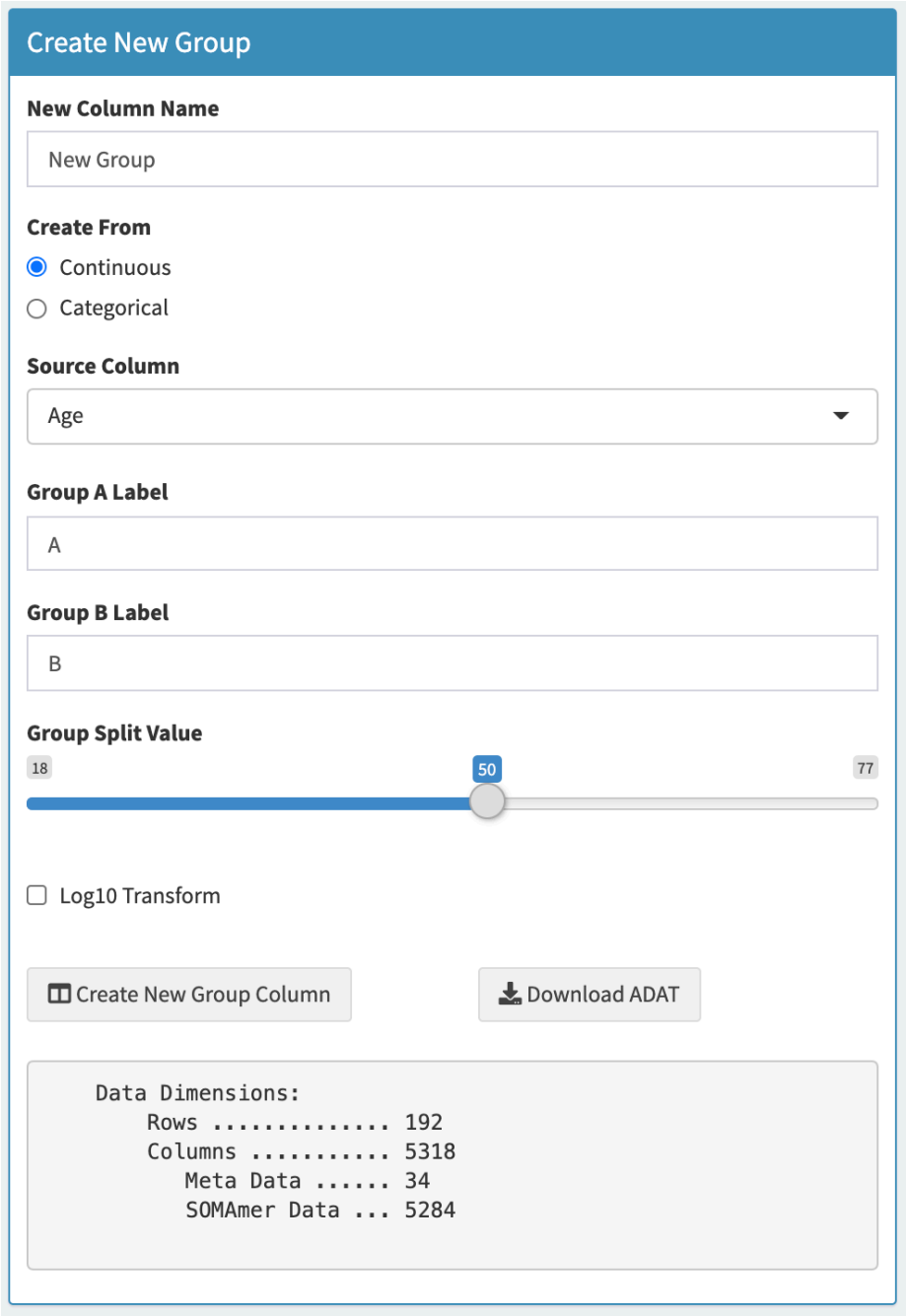


Figure 4.2: The Create New Group box with Continuous options selected.

4.1 Creating a New Group from Continuous Data

- The column containing the new group variable can be named in the **New Column Name** box.

- For splitting a continuous variable, select *Continuous* from the **Create From** options.

- The column containing the data to be split can be selected from the **Source Column** selection box.
- Labels for the two groups are specified under **Group A Label** and **Group B Label**.
- Adjust the **Group Split Value** slider to identify the split point for the data. Rows with values less than the chosen split will be given the label *Group A* and rows with values greater than or equal to the chosen split will be given the label *Group B*.
- Data can be log10 transformed by selecting **Log10 Transform**.
- A preview of the original data column and the newly created data column are shown at the right.
- When ready to create the new group, press the **Create New Group Column** button, and the new column will be added to the ADAT.

In this example, a column titled *New Group* will be created from the existing *Age* column. Rows with values less than 50 will be given the label *A* and values greater than or equal to 50 will be given the label *B*.

4.2 Creating a New Group from Categorical Data

- The column containing the new group variable can be named in the **New Column Name** box.

- For splitting a categorical variable, select *Categorical* from the **Create From** options.

- Clicking on the **Group A** or **Group B** select boxes will show the categorical values available from the chosen column. Selections can be made by clicking on the categorical values desired for that group.
- Labels for the two groups are specified under **Group A Label** and **Group B Label**.
- A preview of the original data column and the newly created data column are shown at the right.
- When ready to create the new group, press the **Create New Group Column** button, and the new column will be added to the ADAT.

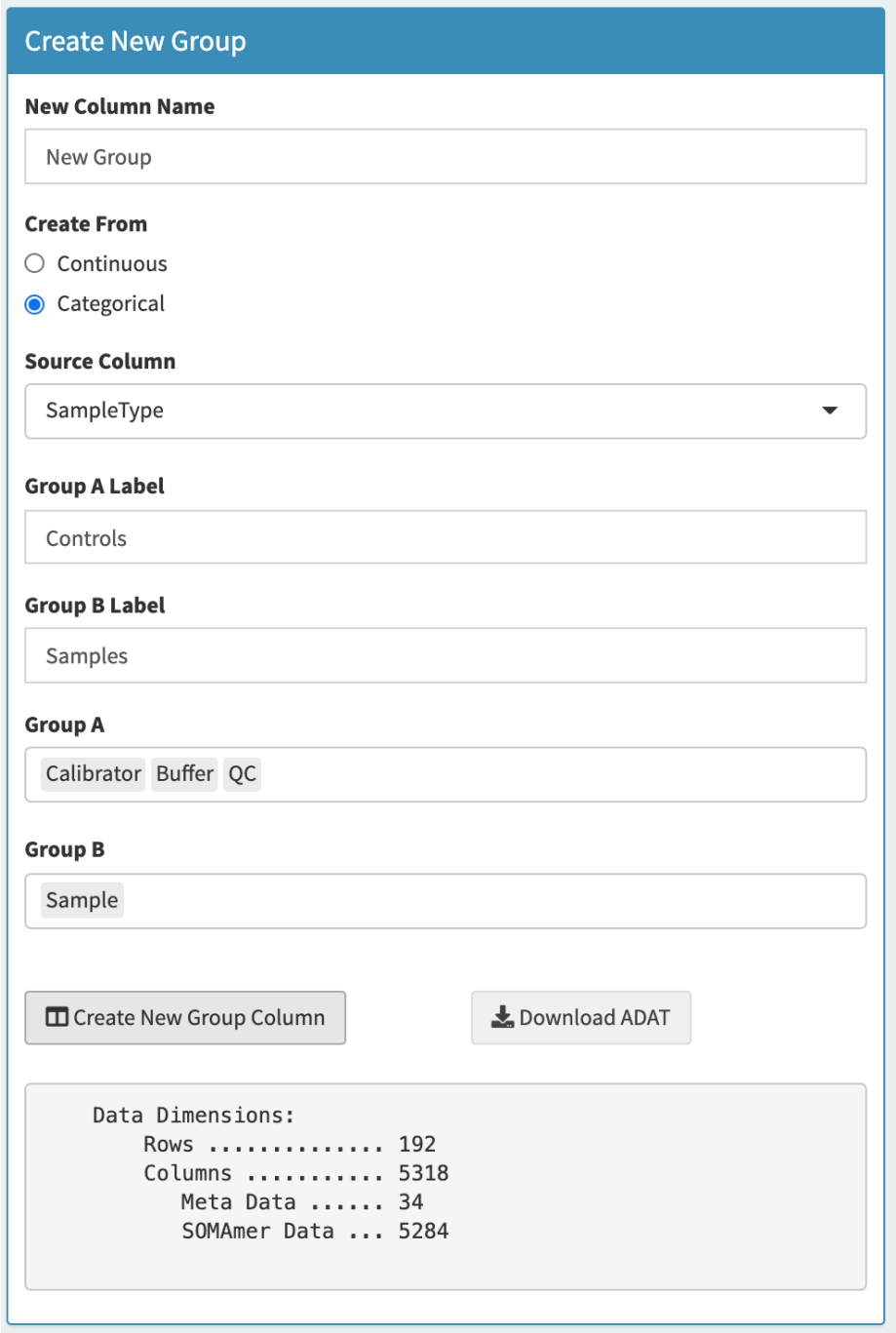


Figure 4.3: The Create New Group box with Categorical options selected.

In this example, a column titled *New Group* will be created from the existing *SampleType* column. Rows with values of *Calibrator*, *Buffer*, or *QC* will be given the label *Controls* and values of *Sample* will be given the label *Samples*. This will provide a useful label to distinguish between all controls and all clinical samples in the ADAT file.

4.3 Download ADAT

Once new groups have been created, a version of the ADAT can be downloaded by clicking the **Download ADAT** button. As this new ADAT file may be significantly modified relative to the original, ensure that a new, unique, descriptive file name is specified so that the original ADAT is not overwritten. Keeping good notes regarding the filtering operations is essential to recall how the ADAT was modified when it is returned to at a later date.

Chapter 5

Plots

ProViz provides tools to dynamically create a variety of plots with custom options. Plots can be created using SomaScan Assay data as well as meta data in the ADAT or data imported in the **Merge Data** panel. Plot features such as titles, colors, and lines can be added and customized using selectable options in the ProViz plotting panels.

5.1 Plot Features

At any point during plotting, moving the mouse over the plot will show a toolbar a the top of the plot.

Figure 5.1: Plot toolbar

The icons from left to right provide the following features.

- download as PNG - saves the image to a file
- zoom - when selected, the plot can be zoomed by clicking and dragging the mouse
- pan - when selected, the plot can be moved by clicking a dragging
- zoom in - zooms into the plot, keeping the current center
- zoom out - zooms into the plot, keping the current center
- autoscale - resets the coordinates of the plot if it has been zoomed or panned (similar to reset axes, below)
- reset axes - resets coordinates of the plot if it has been zoomed or panned (similar to autoscael, above)

- toggle spike lines - when selected, hovering over points on the graph will also include lines from teh point to the axes
- show closest data on hover - only informaiton about the closest data point will be shown when the point is hovered over
- compare data on hover - if multiple points are plotted at the same location, data for all points will be shown
- Plotly icon - a link to the Plotly website. Plotly is the tool used by ProViz to produce graphs.

5.2 Dynamic Plot Interactions

At any point during plotting, the plots provide dynamic interaction capabilities. Hovering on the plotted elements themselves, such as boxes in a box plot or points in any plot, provides additional information about the samples corresponding to that plotting element. Plots can be zoomed or panned by clicking and dragging - see the details above in **Plot Features**.

5.3 Boxplots

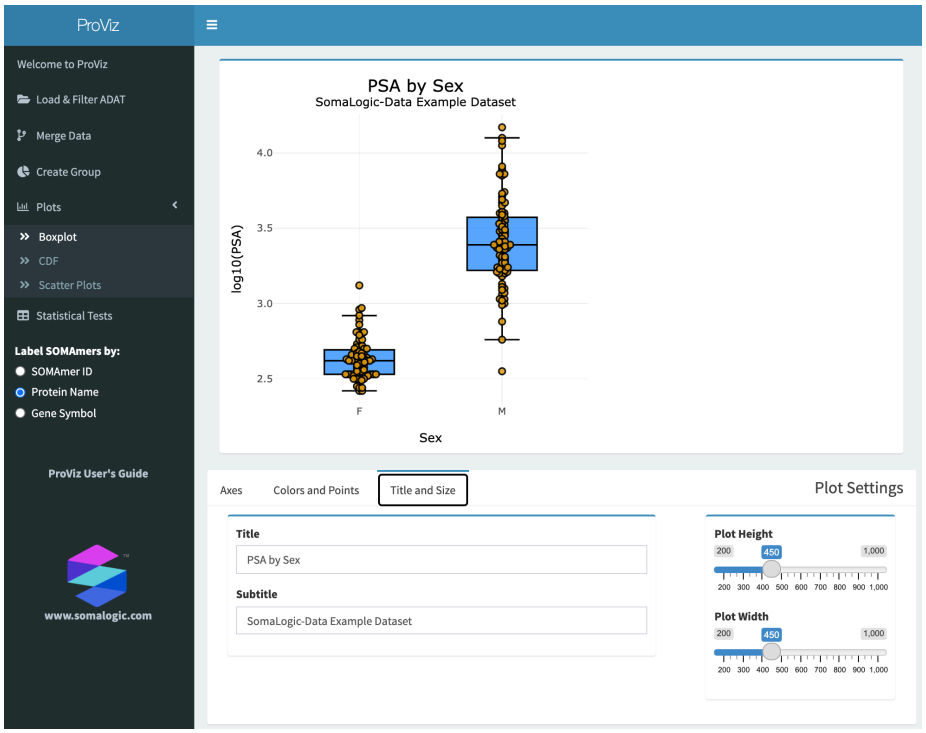


Figure 5.2: The Boxplot panel with a customized plot.

5.3.1 Axes

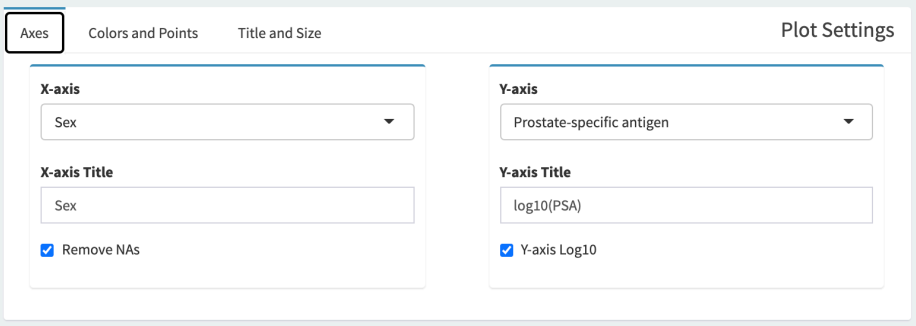


Figure 5.3: The Axes tab in the Plot Settings box.

A categorical variable that defines the boxes in the boxplot can be selected from the **X-axis** select box. If the categorical variable has missing values (NAs), those can be removed by selected **Remove NAs**.

A continuous variable can be selecged from the **Y-axis** select box. The continuous variable can be transformed by log10 by selecting **Y-axis Log10**.

5.3.2 Colors and Points

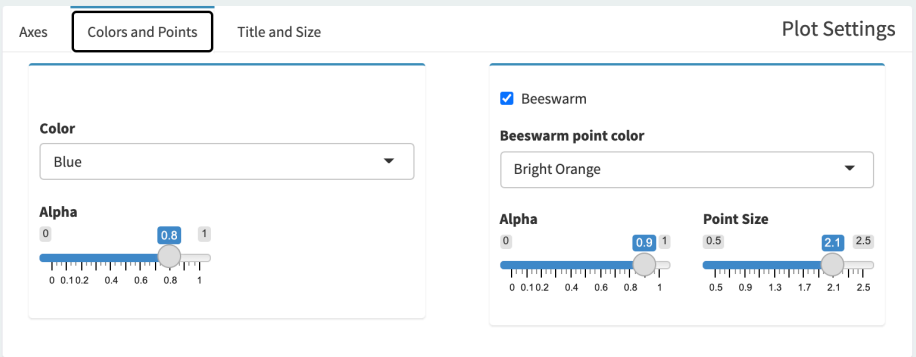


Figure 5.4: The Colors and Points tabs in the Plot Settings box.

Colors of boxes can be customized by selecting from the **Color** select box, and the degree of transparency can be adjusted with the **Alpha** slider on the left.

Individual points for each sample can be displayed by selected **Beeswarm**. Colors of the points can be customized by selecting a color from **Beeswarm point color** and the transparency can be adjusted with the **Alpha** slider on the right. The size of the points can be adjusted with the **Point Size** slider.

5.3.3 Title and Size

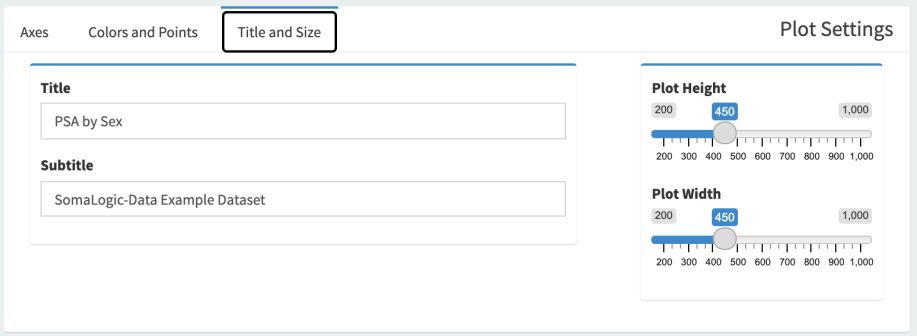


Figure 5.5: The Title and Size tab in the Plot Settings box.

A title and subtitle can be added to the plot by entering text into the **Title** and **Subtitle** boxes. The plot size can be adjusted using the **Plot Height** and **Plot Width** sliders.

5.4 CDF Plots

5.4.1 Axes

A continuous variable can be selected from the **X-axis** select box. The continuous variable can be log10 transformed by selecting **X-axis Log10**.

5.4.2 Lines and Points

A categorical variable can be selected from the **Color By** select box, which will produce one CDF for each category in that variable. If there are NAs for some samples, a CDF will also be produced for NA-containing set, or it can be removed by selecting **Remove NAs**. Line width can be adjusted using the **Line Width** slider, and point sizes can be adjusted with the **Point Size** slider. If the **Point Size** slider is set to *0*, points will be removed from the plot.

5.4.3 Title and Size

A title and subtitle can be added to the plot by entering text into the **Title** and **Subtitle** boxes. The plot size can be adjusted using the **Plot Height** and **Plot Width** sliders.

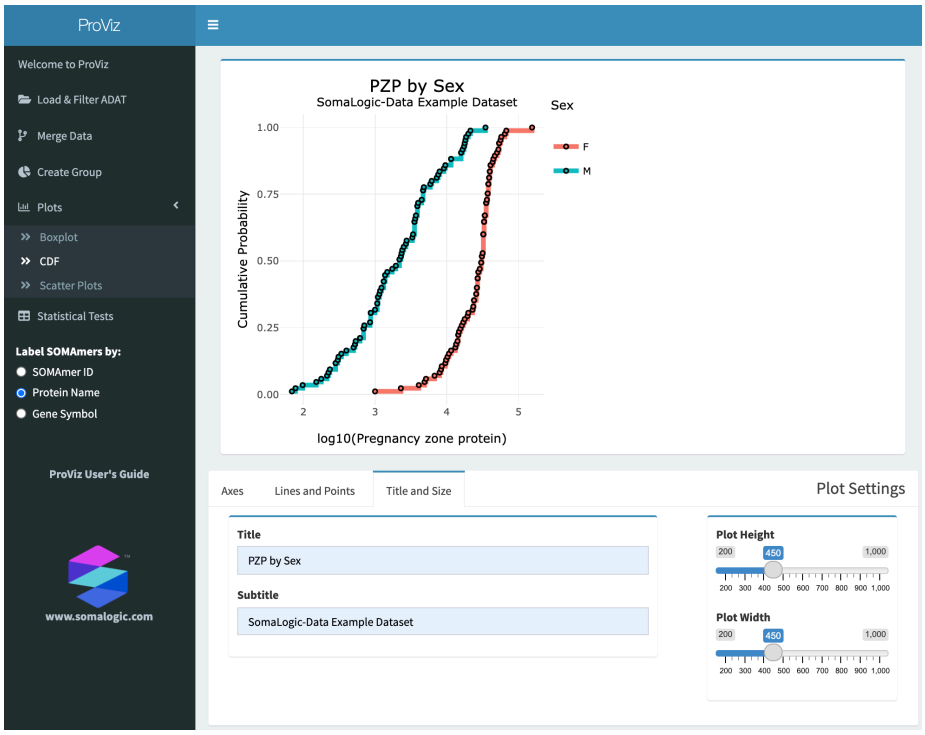


Figure 5.6: The CDF panel with a customized plot.

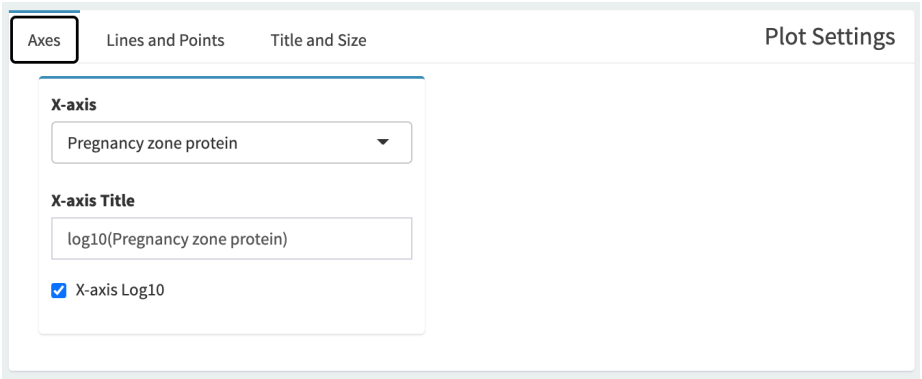


Figure 5.7: The Axes tab in the Plot Settings box.

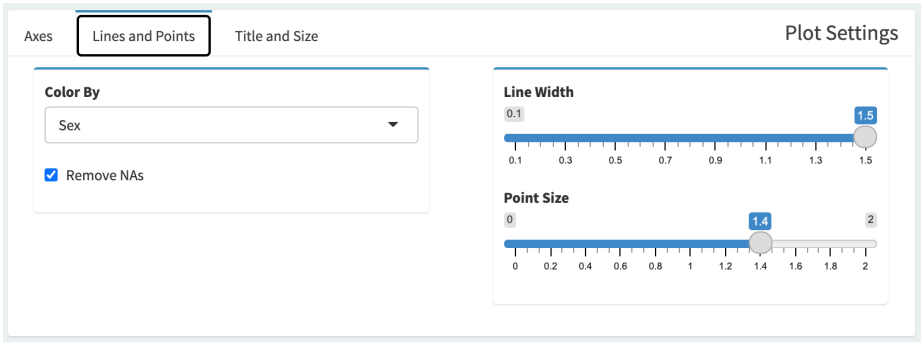


Figure 5.8: The Lines and Points tab in the Plot Settings box.

Figure 5.9: The Title and Size tab in the Plot Settings box.

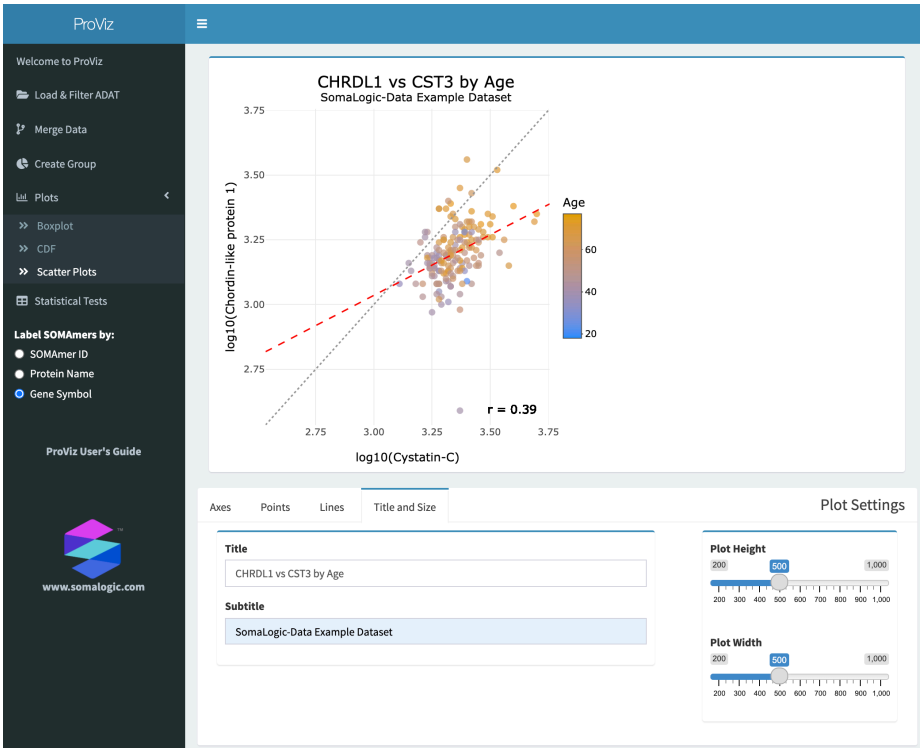


Figure 5.10: The Scatter Plot panel with a customized plot.

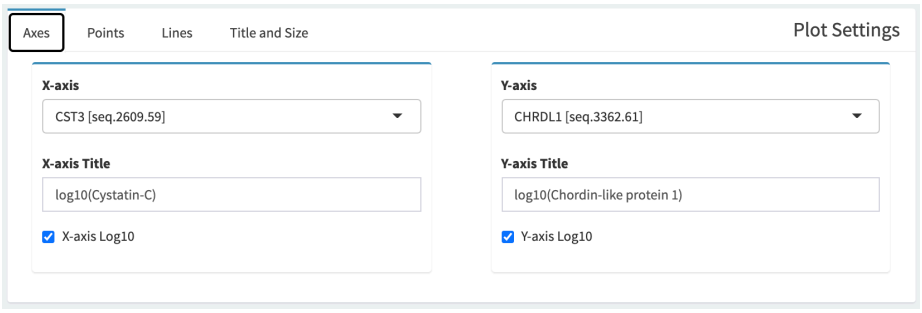


Figure 5.11: The Axes tab in the Plot Settings box.

5.5 Scatter Plots

5.5.1 Axes

To set the X- and Y-axes, select continuous variables from the **X-axis** and **Y-axis** select boxes. Selecting **X-axis Log10** or **Y-axis Log10** will log10 transform the data for that axis.

5.5.2 Points

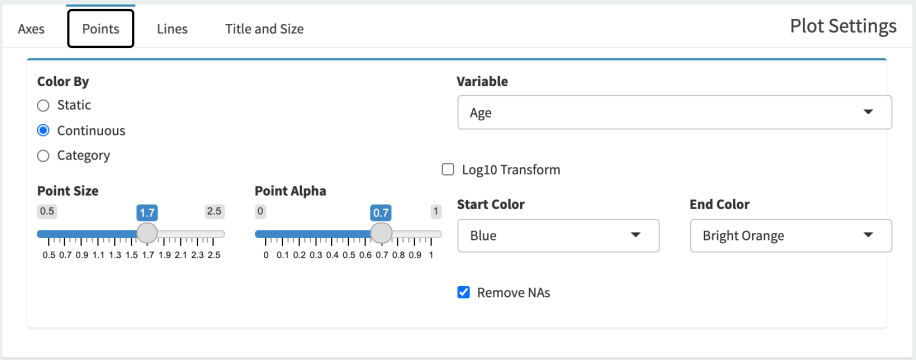


Figure 5.12: The Points tab in the Plot Settings box.

Individual points can be colored in multiple ways depending on the selection for **Color By**.

- If *Static* is selected, all points will be the same color, and that color can be chosen from the **Point Color** select box.
- If *Continuous* is selected, points can be colored by a gradient across two colors based upon the value of a chosen variable. Chose the variable using the **Variable** select box. The values for the variable can be log10 transformed by selecting **Log10 Transform**. The colors to use for the gradient can be chosen by selecting colors from **Start Color** and **End Color**.
- If *Category* is selected, points can be colored based upon a categorical variable. The variable can be chosen from the **Variable** select box.
- Points can also be sized by adjusting the **Point Size** slider, and their transparency adjusted using the **Point Alpha** slider. If there are NAs present, they can be removed by selecting **Remove NAs**.

5.5.3 Lines

A linear regression line can be added to the plot by selecting **Regression Line**. The width of the line can be adjusted with the **Regression Line Width** slider,

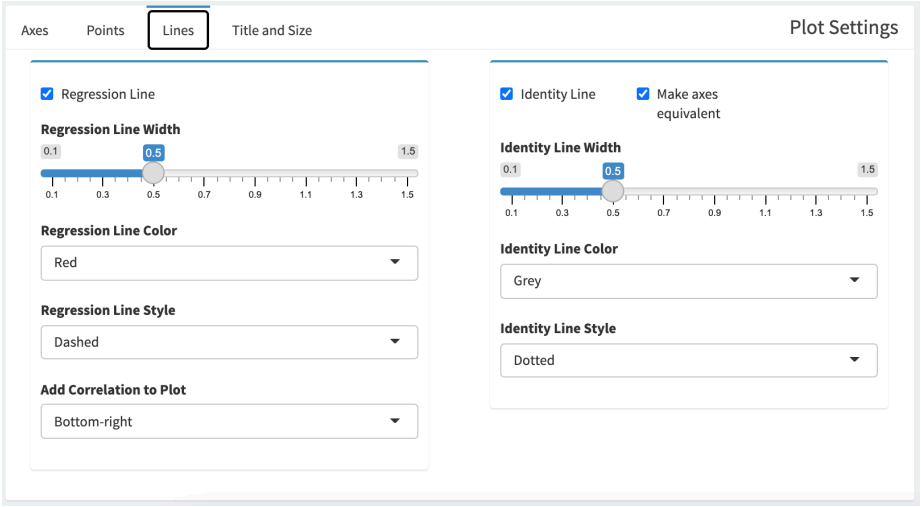


Figure 5.13: The Lines tab in the Plot Settings box.

it can be colored by selecting a color from **Regression Line Color**, and the plotting style (solid, dashed, dotted, etc.) can be selected from the **Regression Liney Style** select box. An R^2 of the correlation can be added to the plot by selecting an position option from the **Add R^2 to Plot** select box.

An identity line can be added to the plot by selecting **Identity Line**. Identity lines are useful to illustrate the expected, perfect one-to-one relationship between two variables. Selecting **Make axes equivalent** will create a square plot in which the extents of the X- and Y-axes are the same. The identity line can be adjusted for width with the **Identity Line Width** slider, the color can be selected from the **Identity Line Color** select box, and the plotting style selected from the **Identity Line Style** select box.

5.5.4 Title and Size

A title and subtitle can be added to the plot by entering text into the **Title** and **Subtitle** boxes. The plot size can be adjusted using the **Plot Height** and **Plot Width** sliders.

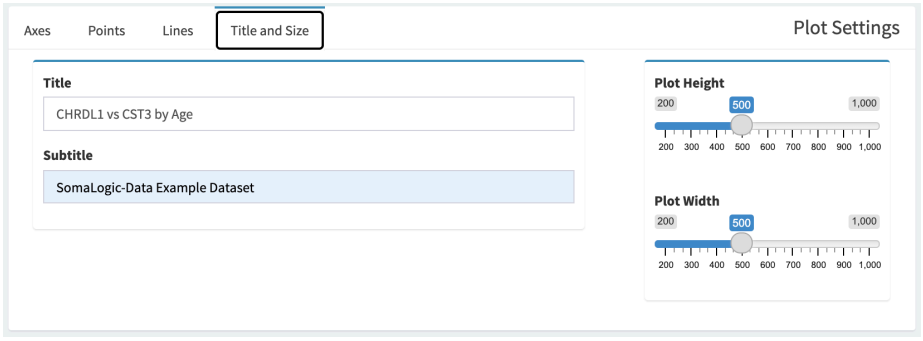


Figure 5.14: The Title and Size tab int he Plot Settings box.

Chapter 6

Statistical Tests

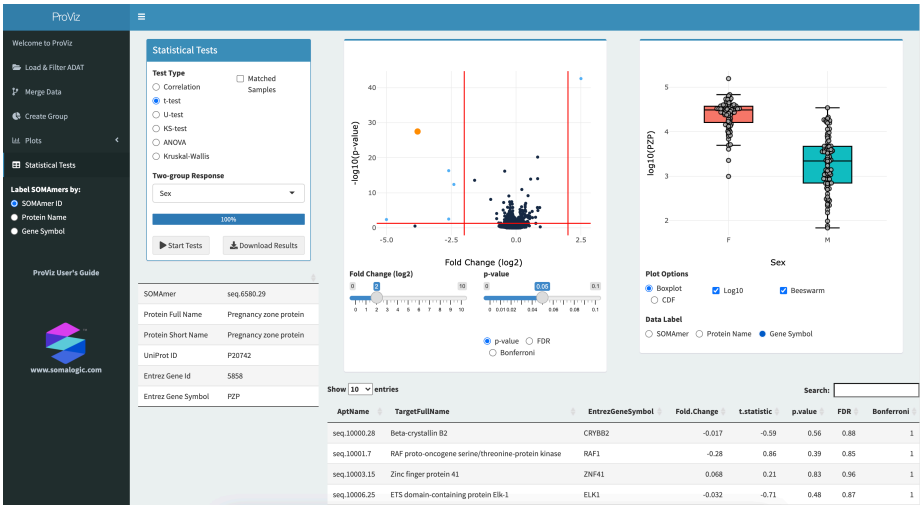


Figure 6.1: The Statistical Tests panel.

6.1 Statistical Tests

Different statistical tests for comparing continuous, two-group data, or multi-group data can be performed on the Statistical Tests panel. The tests available are:

- correlation - compares two continuous variables using either Pearson's or Spearman's rank-based methods
- t-test - compares the means of two groups; a paired t-test is also available

- U-test - compares the medians of two groups; this is a non-parametric alternative to the t-test; a paired U-test is also available
- KS-test - compares the distributions of data between two groups
- ANOVA - Analysis of Variance test of the means of multiple groups; repeated measures ANOVA is also available
- Kruskal-Wallis - compares the medians of multiple groups; this is a non-parametric alternative to ANOVA
- Friedman's Test - a non-parametric alternative to repeated measures ANOVA

6.1.1 Selecting a Response

The type of test can be selected from **Test Type**.

The variable containing the response or the grouping labels can be selected from the select box titled **Continuous Response**, **Two-group Response**, or **Multi-group Response**. The label of the select box will change as different test types are selected. Also, the contents of the select box will be updated with varaibles consistent with the selected test. After selecting a new response variable, tests will be performed for all SOMAmers in the ADAT, which could take a minute or two.

If a paired t-test, paired U-test, repeated measures ANOVA, or Friedman's test are required, check the Matched Samples box and select the variable which defines how samples are matched across groups. **All matched tests require a complete dataset in which each treatment group consists of the same number of matched subjects, and all subjects have measurements in each group. If your dataset is incomplete and does not have all measurements for all subjects across all groups, ProViz will attempt to adjust the data as necessary. This may, however result in errors.**

Once a test has been conducted, the *Statistical Results Table* will contain all test results and can be downloaded as a comma-separated file (.csv) by clicking the **Download Results** button.

6.1.2 Plots

After the statistical test is complete and the results table has been loaded, a volcano plot is displayed. For the t-test and U-test, this plot illustrates each SOMAmer as a point with the X-axis being the log2(fold change) and the Y-axis being -log10(p-value). For a correlation test, the X-axis will illustrate the Pearson's Correlation Coefficient. For ANOVA and Kruskal-Wallis, the X-axis will illustrate the maximum fold-change between groups. The left slider (**Fold Change (log2)** for all tests except correlation, and **Correlation** for the correlation test) and the **p-value** slider can be used to adjust the vertical and

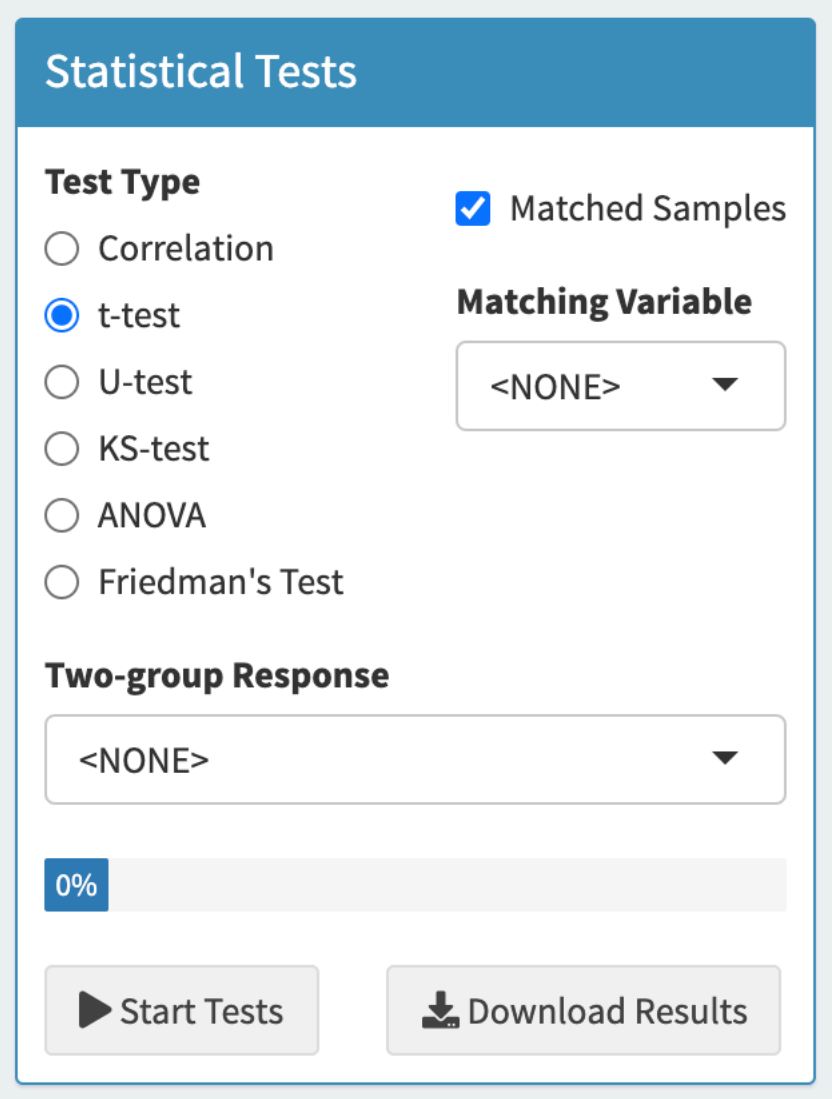


Figure 6.2: The Settings box in the Statistical Tests panel.

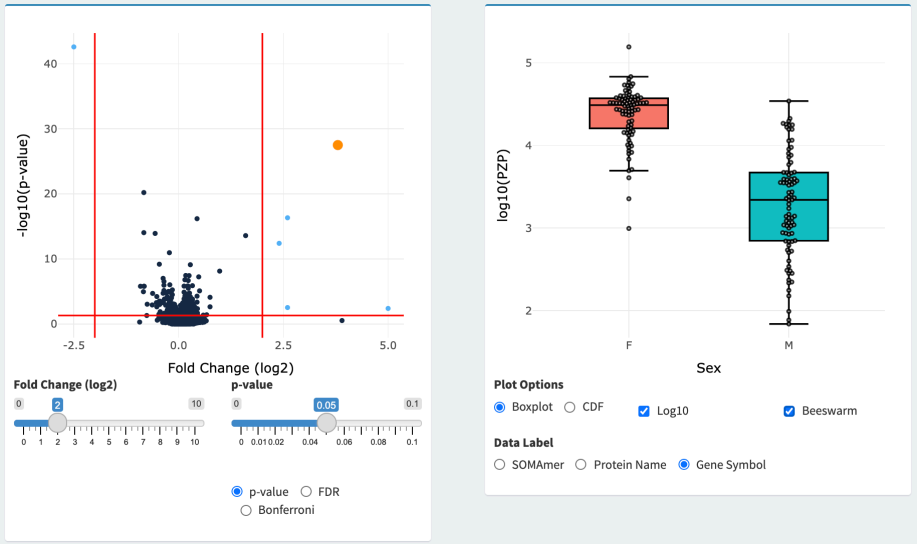


Figure 6.3: The Volcano Plot and Distribution Plot in the Statistical Tests panel.

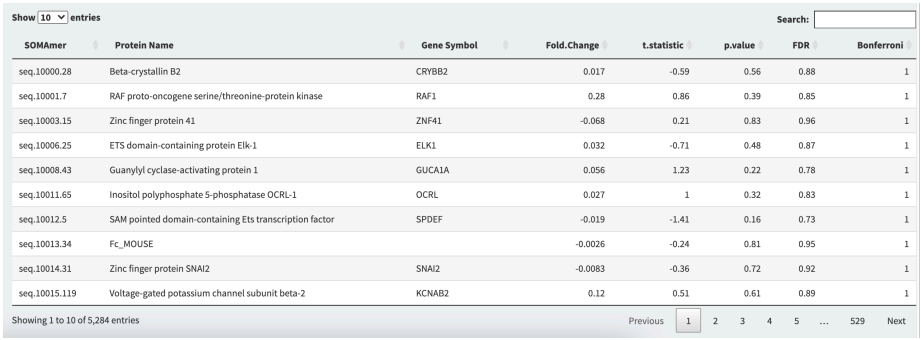
horizontal red lines in the plot, which represent cutoff values for identifying biomarkers of interest. Selecting *p-value*, *FDR*, or *Bonferroni* will determine which type of p-value correction is used for plotting.

Hovering over points on the volcano plot will provide a pop-out detailing information of the SOMAmer associated with that point. Additionally, selecting a point by clicking on it will result in a plot illustrating the selected point's details. For correlation tests, a scatter plot of the SOMAmer data versus the response will be displayed. For all other tests, boxplots or CDF plots for that SOMAmer are shown - the plot can be changed between boxplot and CDF by selecting under **Plot Options**. Data can be log10 transformed by selecting **Log10** and individual points can be displayed by selecting **Beeswarm**. Hovering over points on this plot will provide additional information about the specific sample associated with that point, and hovering over the boxplot will provide summary statistics. The Y-axis of the plot can be labeled with the SOMAmer Id, Protein Name, or Gene Symbol by selecting the appropriate item under **Data Label**.

For matched tests, matched observations across groups can be connected with lines by selecting the **Plot Matched** checkbox on the Distribution Plot.

6.1.3 Statistical Results Table

Results of the statistical test performed for all SOMAmers will be displayed in the **Statistical Results Table**. All protein identifiers (*SOMAmer Id*, *Protein Name*, *Gene Symbol*) will be displayed along with *Fold Change*, *Maximum Fold*



Show 10 entries Search:

SOMAmer	Protein Name	Gene Symbol	Fold.Change	t.statistic	p.value	FDR	Bonferroni
seq.10000.28	Beta-crystallin B2	CRYBB2	0.017	-0.59	0.56	0.88	1
seq.10001.7	RAF proto-oncogene serine/threonine-protein kinase	RAF1	0.28	0.86	0.39	0.85	1
seq.10003.15	Zinc finger protein 41	ZNF41	-0.068	0.21	0.83	0.96	1
seq.10006.25	ETS domain-containing protein Elk-1	ELK1	0.032	-0.71	0.48	0.87	1
seq.10008.43	Guanylyl cyclase-activating protein 1	GUCA1A	0.056	1.23	0.22	0.78	1
seq.10011.65	Inositol polyphosphate 5-phosphatase OCRL-1	OCRL	0.027	1	0.32	0.83	1
seq.10012.5	SAM pointed domain-containing Ets transcription factor	SPDEF	-0.019	-1.41	0.16	0.73	1
seq.10013.34	Fc_MOUSE		-0.0026	-0.24	0.81	0.95	1
seq.10014.31	Zinc finger protein SNAI2	SNAI2	-0.0083	-0.36	0.72	0.92	1
seq.10015.119	Voltage-gated potassium channel subunit beta-2	KCNAB2	0.12	0.51	0.61	0.89	1

Showing 1 to 10 of 5,284 entries Previous 1 2 3 4 5 … 529 Next

Figure 6.4: The Statistical Results Table in the Statistical Tests panel.

Change, or r (Pearson's correlation coefficient), depending on the chosen test. Additional columns contain, the test statistic, p-value, and p-values adjusted for multiple testing by False Discovery Rate (FDR) or Bonferroni correction. The table can be sorted by clicking on the double arrow next to any column name, or searched for protein name or gene symbol. Selecting a row in the table will result in that point being illustrated in the volcano plot as well as a distribution or scatter plot.