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

ProViz is distributed under the MIT License and use of ProViz constitutes acceptance of this 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 R and RStudio, clone the ProViz repository or download the code, 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.

- DT (0.17)
- dplyr (1.0.6)
- ggbeeswarm (0.6.0)
- ggplot2 (3.3.3)
- magrittr (2.0.1)
- plotly (4.9.3)
- readr (1.4.0)
- shiny (1.6.0)
- shinydashboard (0.7.1)

- shinyWidgets (0.6.0)
- tidyr (1.1.3)

Additionally, ProViz uses the SomaDataIO v5.1.0 package available from SomaLogic's SomaDataIO GitHub repository. Note that SomaDataIO requires the BioBase package from Bioconductor, which requires a slightly different installation procedure:

```
if (!requireNamespace("BiocManager", quietly = TRUE)) {
  install.packages("BiocManager")
}
BiocManager::install("Biobase")
```

Installing dependencies can be accomplished using the `remotes` package and the following series of R commands.

```
# install the remotes package
install.packages('remotes')

# install the individual dependencies
remotes::install_version('DT', version = '0.17',
                         repos = 'http://cran.us.r-project.org')
remotes::install_version('dplyr', version = '1.0.6',
                         repos = 'http://cran.us.r-project.org')
remotes::install_version('ggbeeswarm', version = '0.6.0',
                         repos = 'http://cran.us.r-project.org')
remotes::install_version('ggplot2', version = '3.3.3',
                         repos = 'http://cran.us.r-project.org')
remotes::install_version('magrittr', version = '2.0.1',
                         repos = 'http://cran.us.r-project.org')
remotes::install_version('plotly', version = '4.9.3',
                         repos = 'http://cran.us.r-project.org')
remotes::install_version('readr', version = '1.4.0',
                         repos = 'http://cran.us.r-project.org')
remotes::install_version('shiny', version = '1.6.0',
                         repos = 'http://cran.us.r-project.org')
remotes::install_version('shinydashboard', version = '0.7.1',
                         repos = 'http://cran.us.r-project.org')
remotes::install_version('shinyWidgets', version = '0.6.0',
                         repos = 'http://cran.us.r-project.org')
remotes::install_version('tidyr', version = '1.1.3',
                         repos = 'http://cran.us.r-project.org')
remotes::install_github('Somalogic/SomaDataIO@v5.1.0')
```

Once all dependecies have been installed successfully, open RStudio and navigate to the folder containing the ProViz R code. Open any of the R code files -

`global.R`, `ui.R`, or `server.R`. The code window in RStudio will have a button at the top-left that says ‘Run App’. Press this button to start ProViz. See RStudio’s Shiny page for details of using RStudio to run a Shiny application.

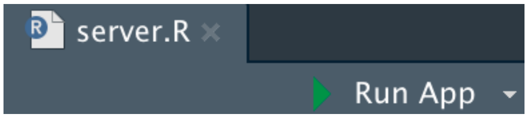


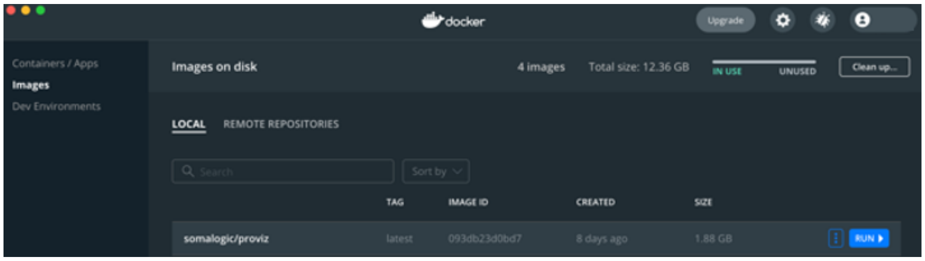
Figure 1.1: The Run App button on the server.R code window in RStudio.

Installation of R and RStudio on different operating systems (Windows, Mac OS, or Linux), may require administrator privileges as well as additional steps in order to complete the installation process. Contact your IT professional for assistance if you are not familiar with performing these types of operations.

1.1.2 DockerHub

ProViz can also be run using Docker or Docker Desktop using the image available through SomaLogic’s Dockerhub account. For more information on installing and running Docker Desktop, refer to the Docker Desktop home page.

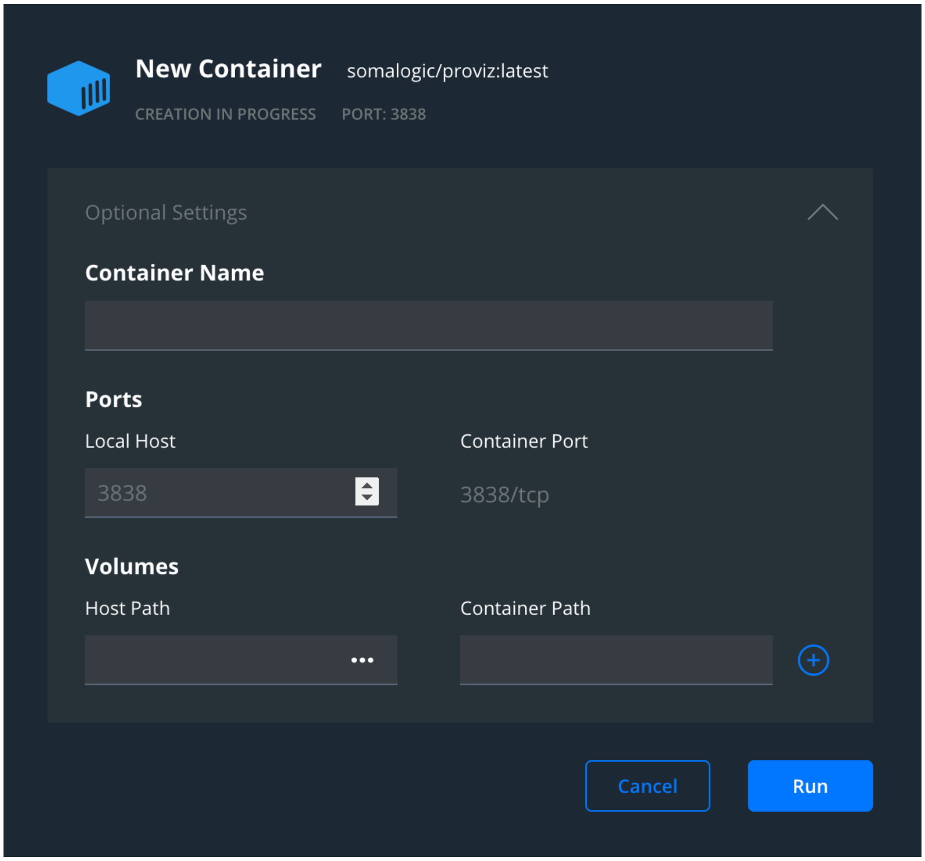
Once Docker Desktop is installed, the ProViz image can be pulled from dockerhub by executing `docker pull somalogic/proviz` from a terminal window on Mac or Linux, or from a command prompt on Windows. Once the image has been successfully pulled, the image name will be displayed in the Docker Desktop’s Dashboard application.



	TAG	IMAGE ID	CREATED	SIZE
somalogic/proviz	latest	093db23d0bd7	8 days ago	1.88 GB

Figure 1.2: Docker Dashboard showing the ProViz image available.

In Docker Desktop, the image can be started by clicking run. A dialog box will be presented in which a local port number needs to be set (Local Port) in order for a web browser to communicate with the running container. Any available port can be used, and the default 3838 is most likely a good option. (You may initially see a smaller dialog box with the port setting hidden. Clicking on the triangle at the right will expand the dialog to show all options.)



New Container somalogic/proviz:latest

CREATION IN PROGRESS PORT: 3838

Optional Settings

Container Name

Ports

Local Host: 3838

Container Port: 3838/tcp

Volumes

Host Path

Container Path

Figure 1.3: Setting the Local Port to 3838.

Once the Local Port is set, click Run. in the Dashboard, you will now see that the image is running. Clicking on Open in Browser will connect your web browser to the container through the Local Port set in the last step.

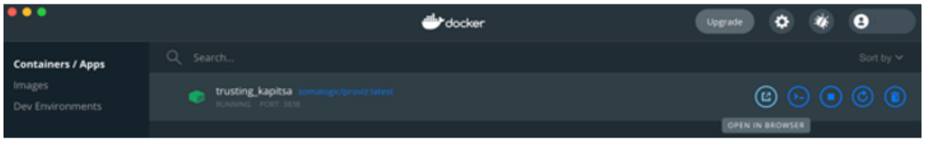


Figure 1.4: Docker Dashboard showing that the container is running.

When your browser opens, you should see the ProViz main panel.

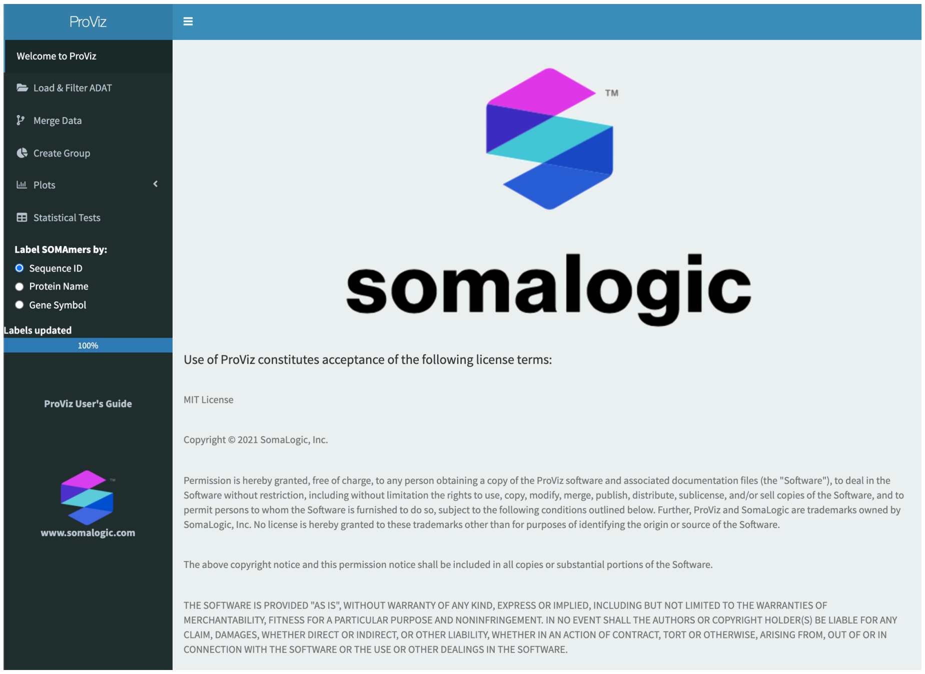
Installing Docker or Docker Desktop and setting configurations to enable the application to work may require administrator privileges as well as additional steps. Contact your IT professional for assistance if you are not familiar with performing these types of operations.

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

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.



Use of ProViz constitutes acceptance of the following license terms:

MIT License

Copyright © 2021 SomaLogic, Inc.

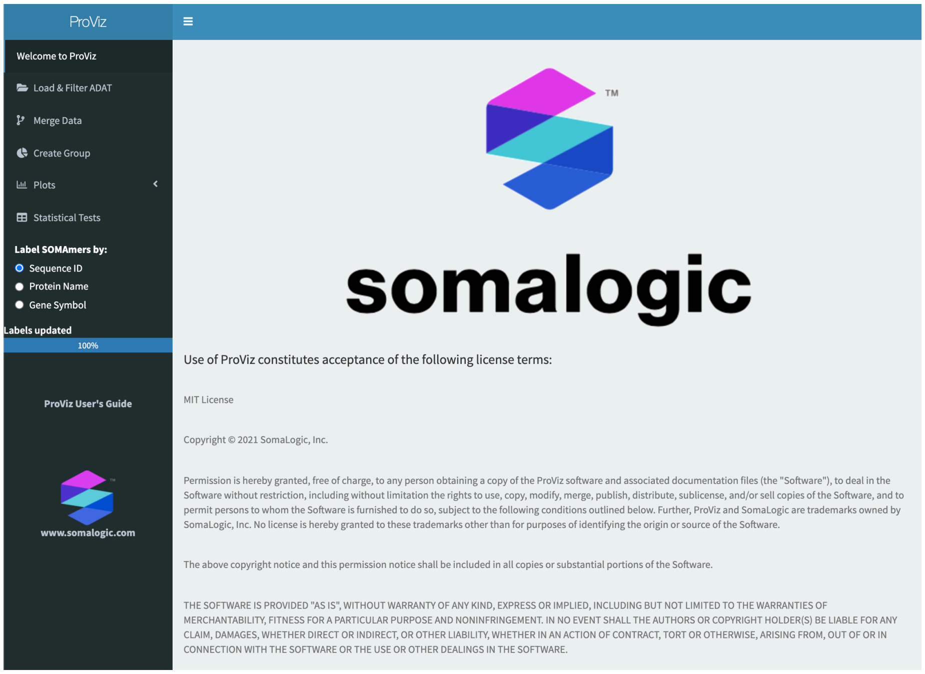
Permission is hereby granted, free of charge, to any person obtaining a copy of the ProViz software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions outlined below. Further, ProViz and SomaLogic are trademarks owned by SomaLogic, Inc. No license is hereby granted to these trademarks other than for purposes of identifying the origin or source of the Software.

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDER(S) BE LIABLE FOR ANY CLAIM, DAMAGES, WHETHER DIRECT OR INDIRECT, OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Figure 1.5: The main ProViz 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**: *Sequence ID*, *Protein Name*, or *Gene Symbol*. The latter two options are most likely more familiar to most users than Sequence IDs. Label types can be changed at any point when using ProViz. When labels are changed, there will be a short delay while the user interface is updated with the new labels, and progress bar indicates the status of the updates.



Label SOMAmers by:

● Sequence ID
○ Protein Name
○ Gene Symbol

Labels updated

100%

Use of ProViz constitutes acceptance of the following license terms:

MIT License

Copyright © 2021 SomaLogic, Inc.

Permission is hereby granted, free of charge, to any person obtaining a copy of the ProViz software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions outlined below. Further, ProViz and SomaLogic are trademarks owned by SomaLogic, Inc. No license is hereby granted to these trademarks other than for purposes of identifying the origin or source of the Software.

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

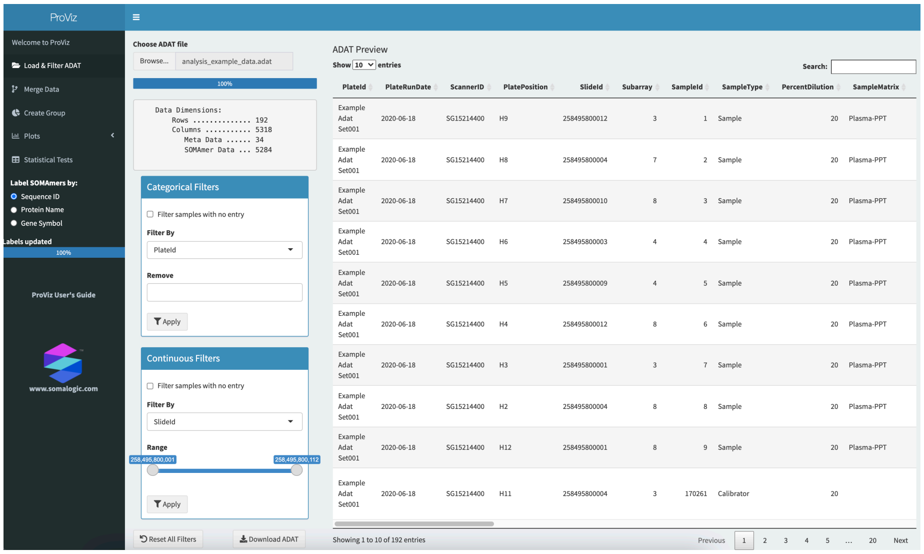
THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDER(S) BE LIABLE FOR ANY CLAIM, DAMAGES, WHETHER DIRECT OR INDIRECT, OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Figure 1.6: The main ProViz panel.

Chapter 2

Load and Filter ADAT

2.1 The Load and Filter ADAT Panel



Label SOMAmers by:

- ● Sequence ID
- ○ Protein Name
- ○ Gene Symbol

Labels updated 100%

Choose ADAT file: analysis_example_data.adat

100%

```
  Data Dimensions:
      Rows .............. 192
      Columns ........... 5318
         Meta Data ...... 34
         SOMAmer Data ... 5284
```

Categorical Filters

□ Filter samples with no entry

Filter By: PlateId

Remove:

Continuous Filters

□ Filter samples with no entry

Filter By: SlideId

Range: 258,495,800,001 – 258,495,800,112

ADAT Preview

Show 10 entries

Search:

PlateId	PlateRunDate	ScannerID	PlatePosition	SlideId	Subarray	SampleId	SampleType	PercentDilution	SampleMatrix
Example Adat Set001	2020-06-18	SG15214400	H9	258495800012	3	1	Sample	20	Plasma-PPT
Example Adat Set001	2020-06-18	SG15214400	H8	258495800004	7	2	Sample	20	Plasma-PPT
Example Adat Set001	2020-06-18	SG15214400	H7	258495800010	8	3	Sample	20	Plasma-PPT
Example Adat Set001	2020-06-18	SG15214400	H6	258495800003	4	4	Sample	20	Plasma-PPT
Example Adat Set001	2020-06-18	SG15214400	H5	258495800009	4	5	Sample	20	Plasma-PPT
Example Adat Set001	2020-06-18	SG15214400	H4	258495800012	8	6	Sample	20	Plasma-PPT
Example Adat Set001	2020-06-18	SG15214400	H3	258495800001	3	7	Sample	20	Plasma-PPT
Example Adat Set001	2020-06-18	SG15214400	H2	258495800004	8	8	Sample	20	Plasma-PPT
Example Adat Set001	2020-06-18	SG15214400	H12	258495800001	8	9	Sample	20	Plasma-PPT
Example Adat Set001	2020-06-18	SG15214400	H11	258495800004	3	170261	Calibrator	20	

Showing 1 to 10 of 192 entries

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 additional sample-related content.

The ADAT Preview table can be scrolled horizontally and vertically, as well as searched for specific content (**Search:** box in upper-right). The number of rows displayed per page can be adjusted with the **Show ... entries** box in the upper-left, and you can step through various pages by clicking on **Previous**, **Next**, or a specific page number at the bottom-right. SomaScan Assay data are not displayed in the preview table.

2.3 Filter ADAT contents

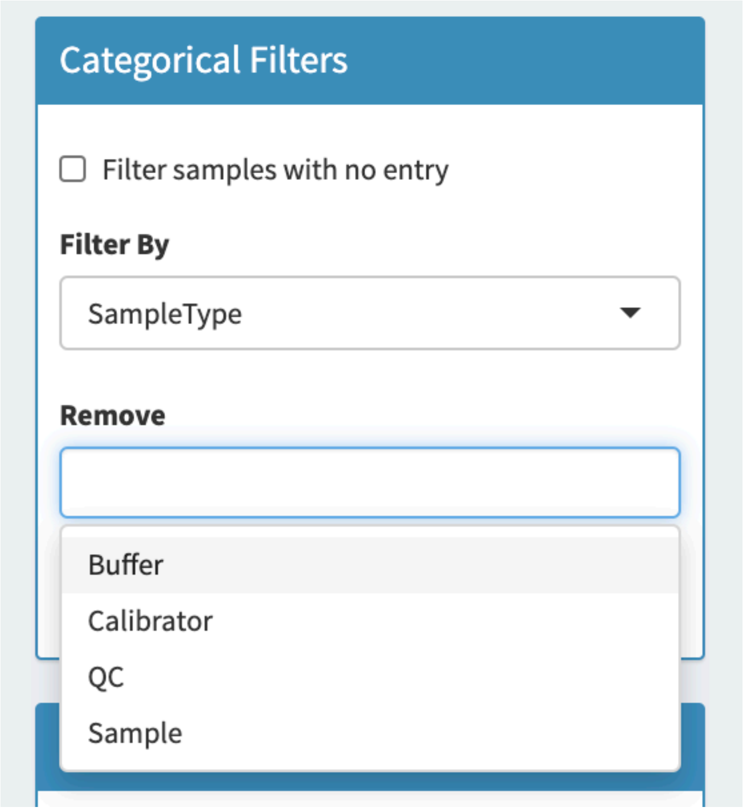
Samples present in the ADAT file can be filtered based on Meta Data or SomaScan Assay data. When the ADAT file is processed, data columns are divided into Categorical (non-numeric) or Continuous (numeric) categories.

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



Categorical Filters

□ Filter samples with no entry

Filter By: SampleType

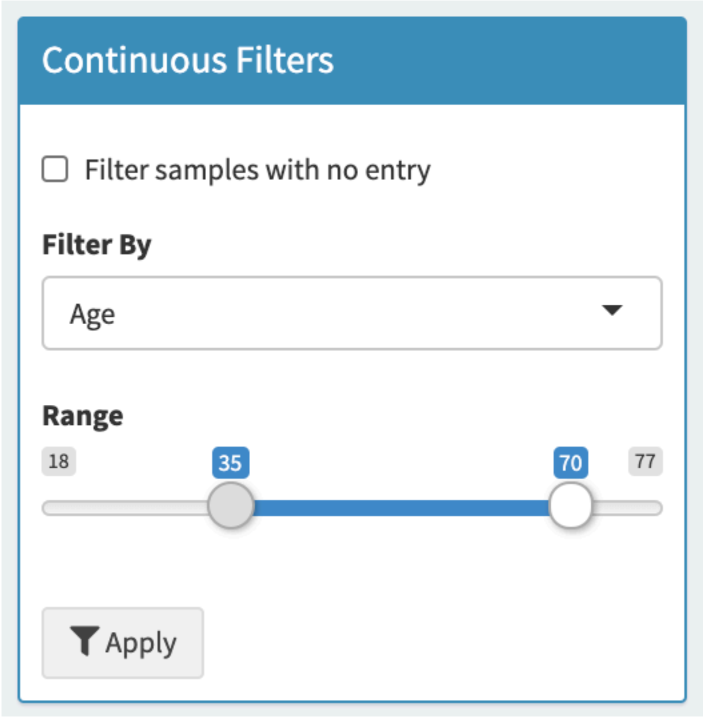
Remove

Figure 2.2: The categorical filters box.

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



Continuous Filters

□ Filter samples with no entry

Filter By

Age

Range

Figure 2.3: The continuous filters box.

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 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 selected 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.

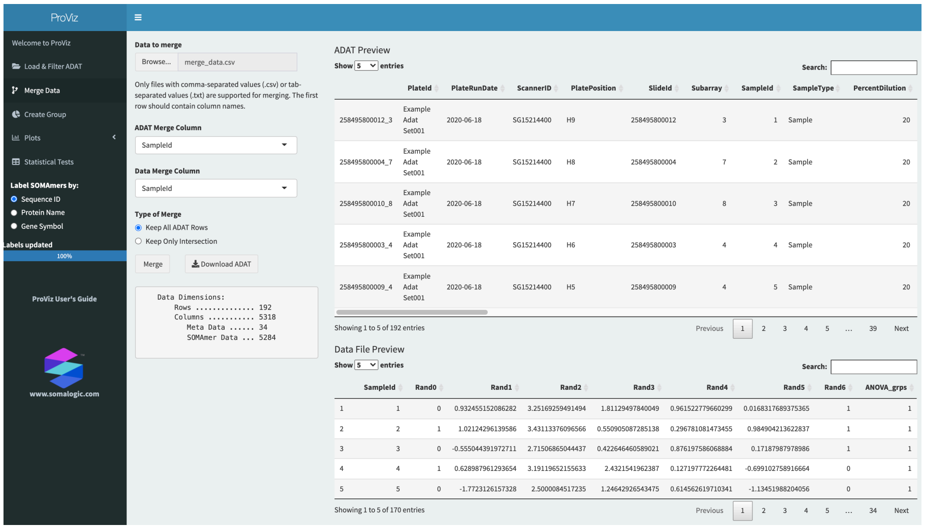
2.3.4 Download ADAT

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. To assist in this record keeping, an additional entry is made in the **^HEADER** section of the ADAT file listing the operations that were performed.

Chapter 3

Merge Data

Data pertaining to individual samples and stored in external files can be merged with the proteomic data in an ADAT file using the Merge Data panel.



Label SOMAmers by:
● Sequence ID
○ Protein Name
○ Gene Symbol

Data to merge
Browse... merge_data.csv

Only files with comma-separated values (.csv) or tab-separated values (.txt) are supported for merging. The first row should contain column names.

ADAT Merge Column
SampleId

Data Merge Column
SampleId

Type of Merge
● Keep All ADAT Rows
○ Keep Only Intersection

```
    Data Dimensions:
       Rows .............. 192
       Columns ........... 5318
          Meta Data ...... 34
          SOMAmer Data ... 5284
```

ADAT Preview

Show 5 entries   Search:

	PlateId	PlateRunDate	ScannerID	PlatePosition	SlideId	Subarray	SampleId	SampleType	PercentDilution
258495800012_3	Example Adat Set001	2020-06-18	SG15214400	H9	258495800012	3	1	Sample	20
258495800004_7	Example Adat Set001	2020-06-18	SG15214400	H8	258495800004	7	2	Sample	20
258495800010_8	Example Adat Set001	2020-06-18	SG15214400	H7	258495800010	8	3	Sample	20
258495800003_4	Example Adat Set001	2020-06-18	SG15214400	H6	258495800003	4	4	Sample	20
258495800009_4	Example Adat Set001	2020-06-18	SG15214400	H5	258495800009	4	5	Sample	20

Showing 1 to 5 of 192 entries   Previous 1 2 3 4 5 … 39 Next

Data File Preview

Show 5 entries   Search:

	SampleId	Rand0	Rand1	Rand2	Rand3	Rand4	Rand5	Rand6	ANOVA_grps
1	1	0	0.932455152086282	3.25169259491494	1.81129497840049	0.961522779660299	0.0168317689375365	1	1
2	2	1	1.02124296139586	3.43113376096566	0.550905087285138	0.296781081473455	0.984904213622837	1	1
3	3	0	-0.555044391972711	2.71506865044437	0.422646460589021	0.876197586068884	0.17187987978986	1	1
4	4	1	0.628987961293654	3.19119652155633	2.4321541962387	0.127197772264481	-0.699102758916664	0	1
5	5	0	-1.7723126157328	2.5000084517235	1.24642926543475	0.614562619710341	-1.13451988204056	0	1

Showing 1 to 5 of 170 entries   Previous 1 2 3 4 5 … 34 Next

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 in a one-to-one fashion between the ADAT file and the external data file. The columns do not need to have the same name. If there are duplicate IDs for the selected matching column in either the ADAT file or the external file, rows may be duplicated in order to match all-to-all, leading to a potentially confusing 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.

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 compatible. Careful construction of the external data file and selection of matching columns is critical.

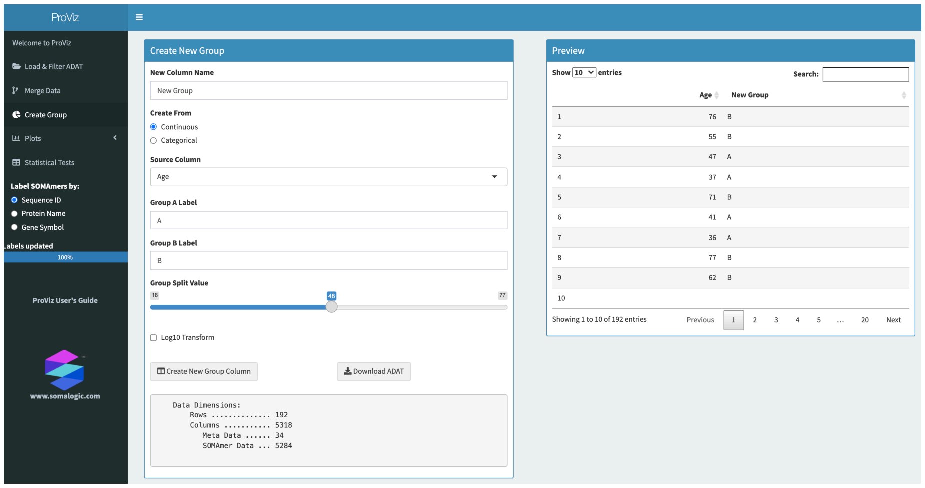
3.5 Download Adat

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.



New Column Name: New Group

Create From: ● Continuous ○ Categorical

Source Column: Age

Group A Label: A

Group B Label: B

Group Split Value: 18 — 48 — 77

□ Log10 Transform

```
    Data Dimensions:
       Rows .............. 192
       Columns ........... 5318
          Meta Data ...... 34
          SOMAmer Data ... 5284
```

Show 10 entries   Search:

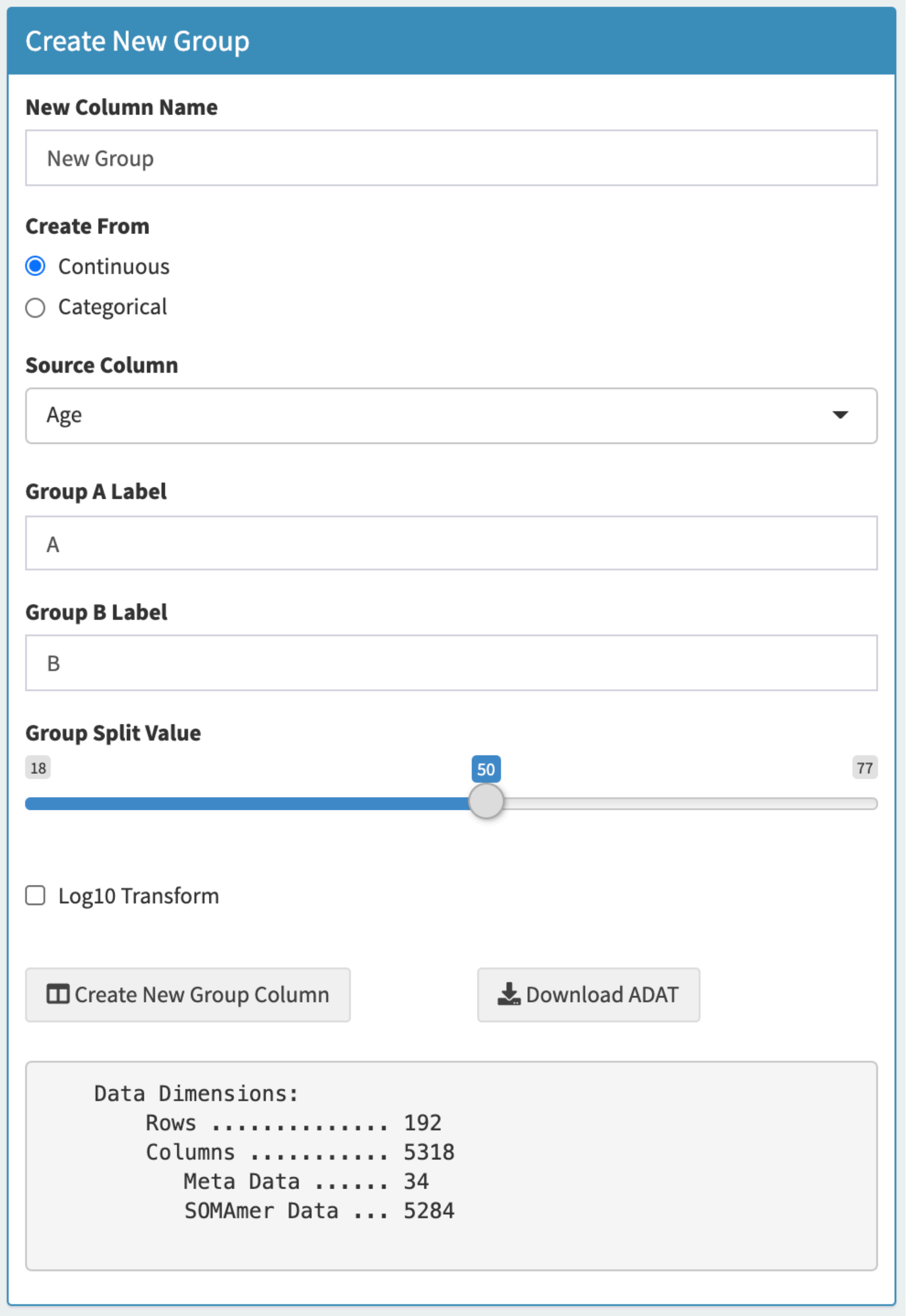
	Age	New Group
1	76	B
2	55	B
3	47	A
4	37	A
5	71	B
6	41	A
7	36	A
8	77	B
9	62	B
10		

Showing 1 to 10 of 192 entries   Previous 1 2 3 4 5 … 20 Next

Figure 4.1: The Create Group panel.

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.



Create New Group

New Column Name

New Group

Create From

● Continuous

○ Categorical

Source Column

Age

Group A Label

A

Group B Label

B

Group Split Value

□ Log10 Transform

```
Data Dimensions:
    Rows .............. 192
    Columns ........... 5318
       Meta Data ...... 34
       SOMAmer Data ... 5284
```

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

- 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. The content of this box is adjusted based on whether *Continuous* or *Categorical* is chosen from the **Create From** options.
- 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.

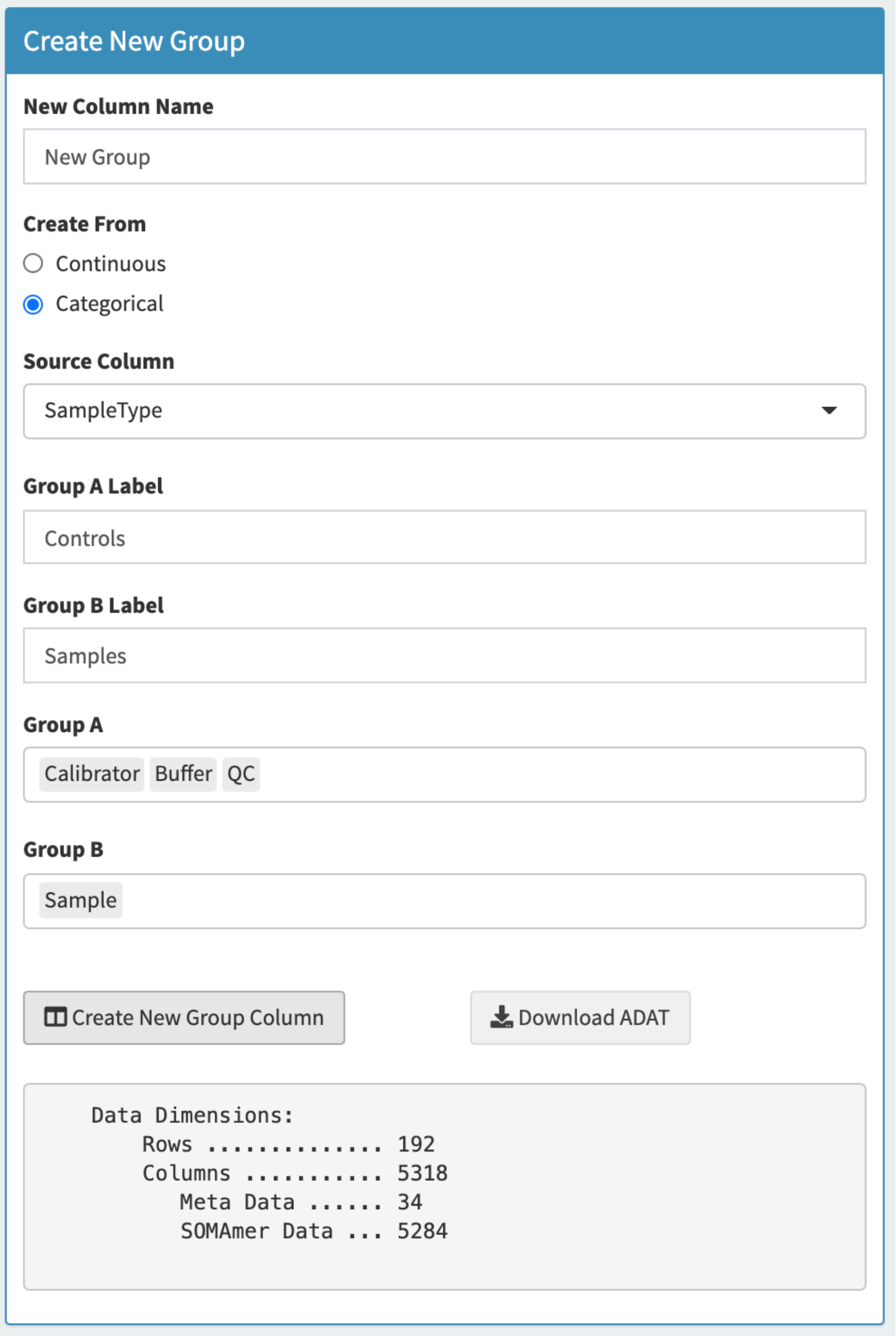
In this example, a column titled *New Group* will be created from the existing *Age* column. Rows with values less than 50 will be given the label *A* and values greater than or equal to 50 will be given the label *B*.

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.

- The column containing the data to be split can be selected from the **Source Column** selection box. The content of this box is adjusted based on whether *Continuous* or *Categorical* is chosen from the **Create From** options.
- Labels for the two groups are specified under **Group A Label** and **Group B Label**.
- 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.
- A preview of the original data column and the newly created data column are shown at the right.



Create New Group

New Column Name: New Group

Create From

○ Continuous

● Categorical

Source Column: SampleType

Group A Label: Controls

Group B Label: Samples

Group A: Calibrator Buffer QC

Group B: Sample

```
Data Dimensions:
    Rows .............. 192
    Columns ........... 5318
       Meta Data ...... 34
       SOMAmer Data ... 5284
```

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

- 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 *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 at the top of the plot.

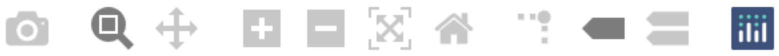


Figure 5.1: Plot toolbar

The icons from left to right provide the following features.

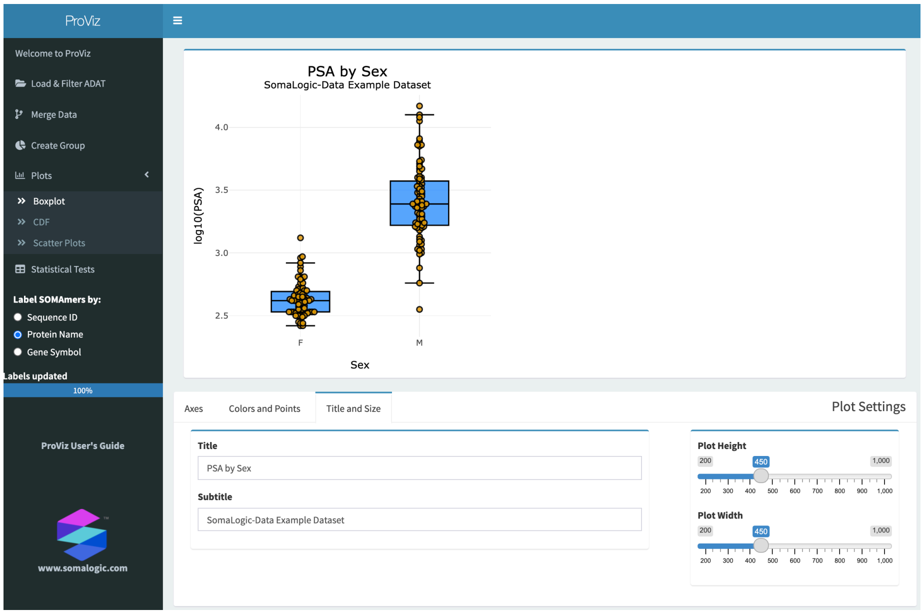
- download as PNG - saves the image to a file
- zoom - when selected, the plot can be zoomed by clicking and dragging the mouse
- pan - when selected, the plot can be moved by clicking a dragging
- zoom in - zooms into the plot, keeping the current center
- zoom out - zooms into the plot, keping the 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 autoscale, 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 information 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



Label SOMAmers by:
○ Sequence ID
● Protein Name
○ Gene Symbol

Labels updated 100%

Title: PSA by Sex

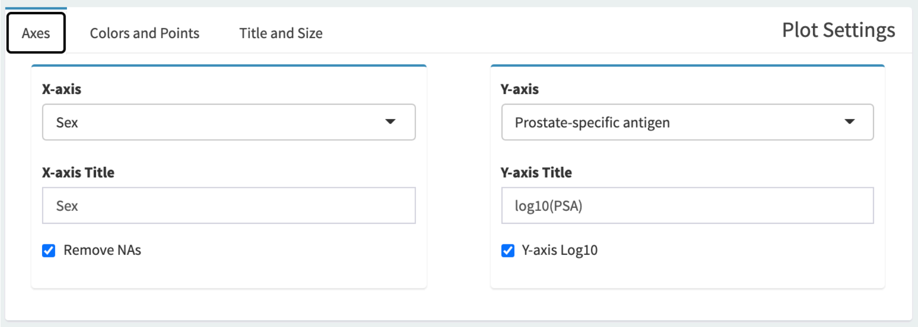
Subtitle: SomaLogic-Data Example Dataset

Plot Height: 450

Plot Width: 450

Figure 5.2: The Boxplot panel with a customized plot.

5.3.1 Axes



X-axis: Sex

X-axis Title: Sex

☑ Remove NAs

Y-axis: Prostate-specific antigen

Y-axis Title: log10(PSA)

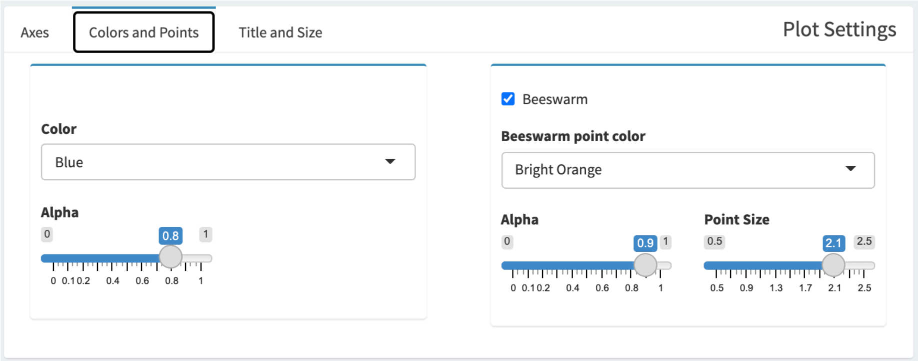
☑ Y-axis Log10

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 selected 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



Color: Blue

Alpha: 0.8

☑ Beeswarm

Beeswarm point color: Bright Orange

Alpha: 0.9

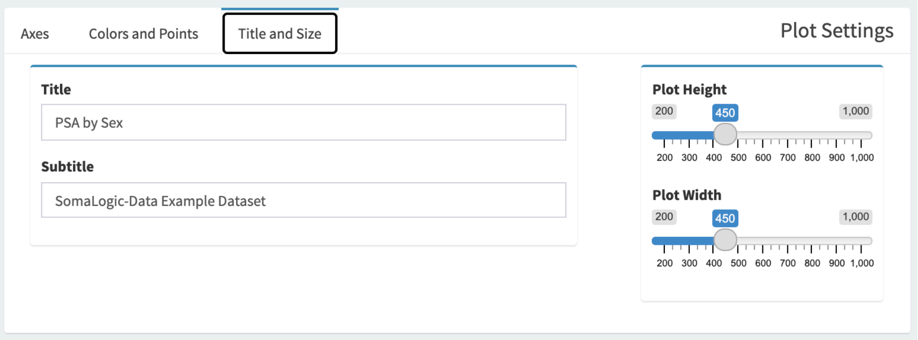
Point Size: 2.1

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



Title: PSA by Sex

Subtitle: SomaLogic-Data Example Dataset

Plot Height: 450

Plot Width: 450

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

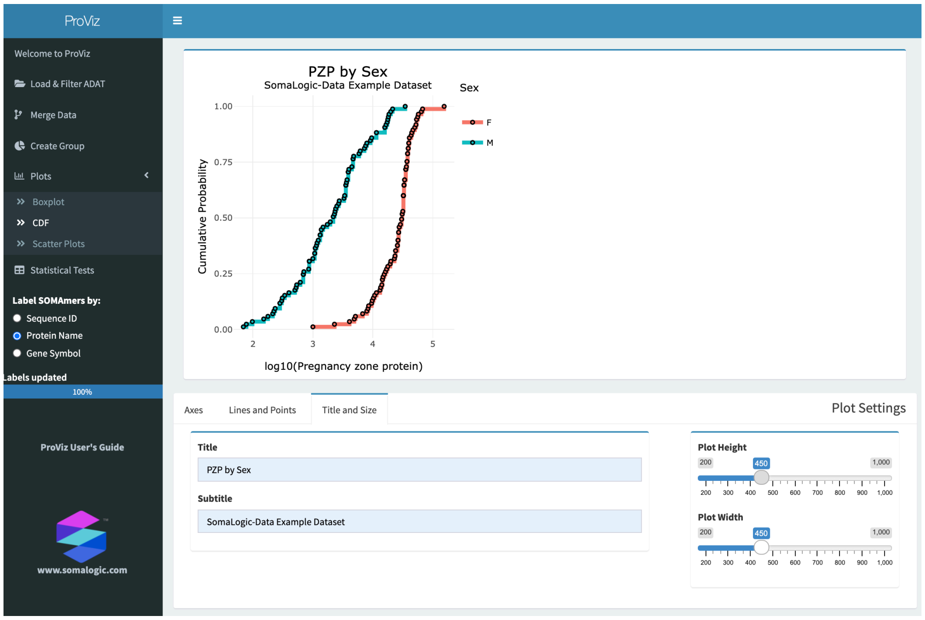
A continuous variable can be selected from the **X-axis** select box. The continuous variable can be log10 transformed by selecting **X-axis Log10**.

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.



Title

PZP by Sex

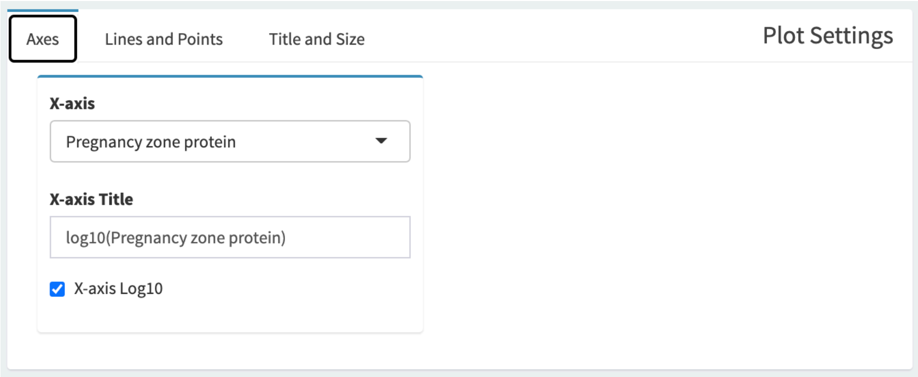
Subtitle

SomaLogic-Data Example Dataset

Plot Height: 450

Plot Width: 450

Figure 5.6: The CDF panel with a customized plot.



X-axis

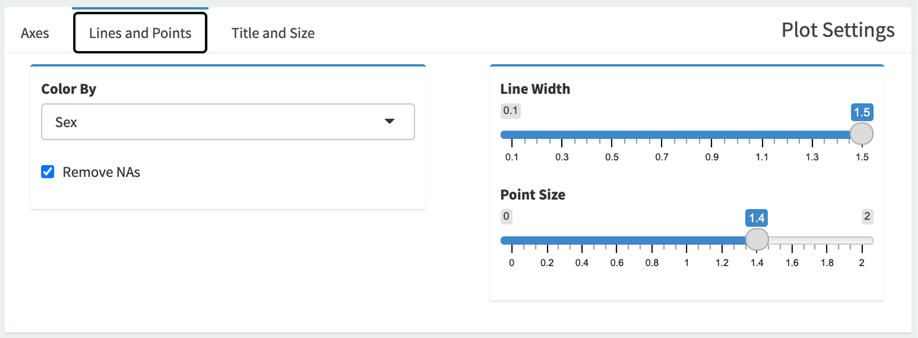
Pregnancy zone protein

X-axis Title

log10(Pregnancy zone protein)

☑ X-axis Log10

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



Color By

Sex

☑ Remove NAs

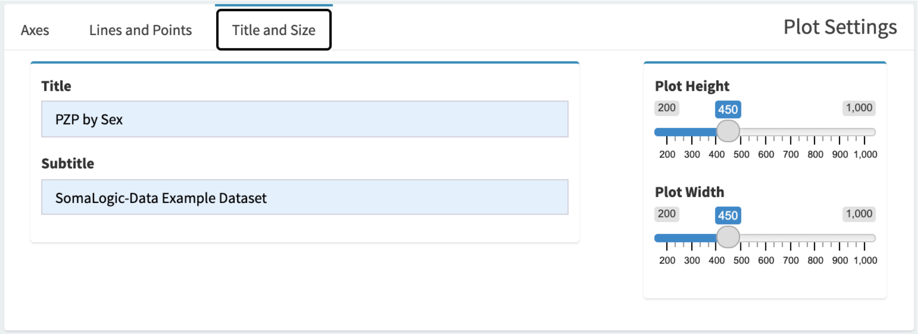
Line Width

1.5

Point Size

1.4

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



Title

PZP by Sex

Subtitle

SomaLogic-Data Example Dataset

Plot Height

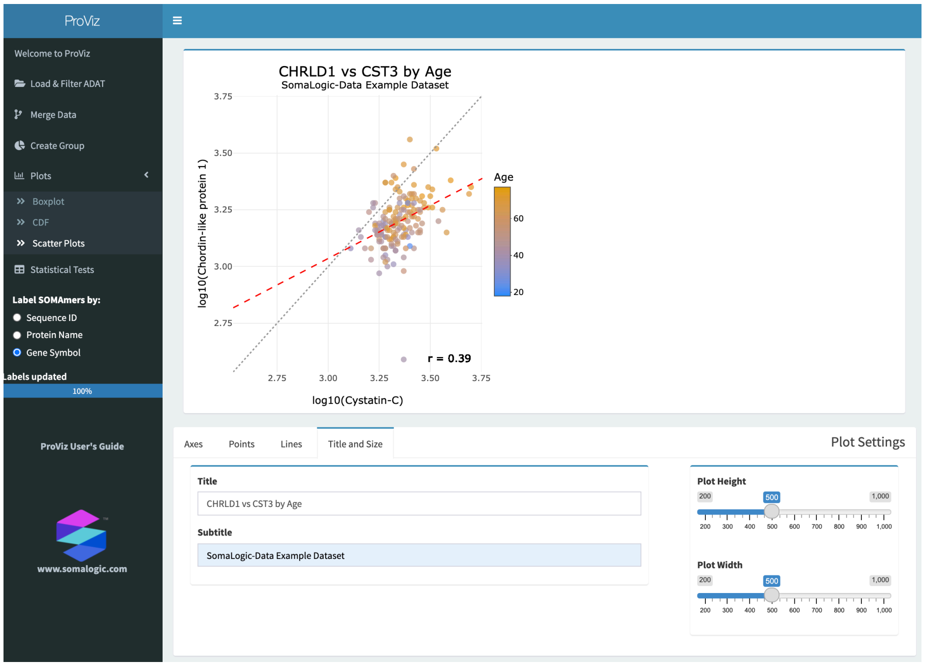
450

Plot Width

450

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

5.5 Scatter Plots



Title: CHRLD1 vs CST3 by Age

Subtitle: SomaLogic-Data Example Dataset

Plot Height: 500 (range 200 – 1,000)

Plot Width: 500 (range 200 – 1,000)

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

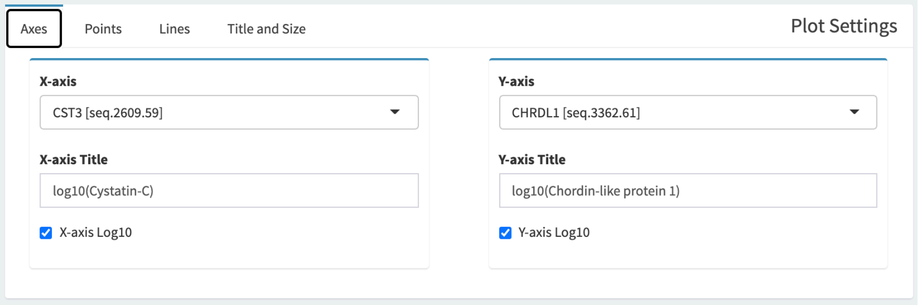
5.5.1 Axes

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

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



X-axis

CST3 [seq.2609.59]

X-axis Title

log10(Cystatin-C)

☑ X-axis Log10

Y-axis

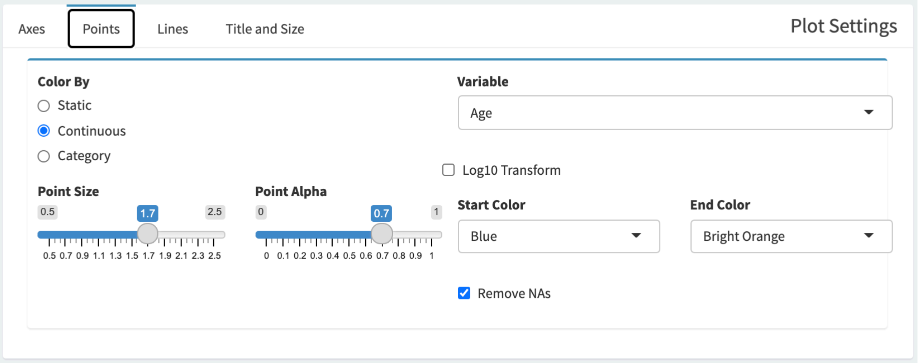
CHRDL1 [seq.3362.61]

Y-axis Title

log10(Chordin-like protein 1)

☑ Y-axis Log10

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



Color By

○ Static
● Continuous
○ Category

Variable

Age

□ Log10 Transform

Point Size 1.7

Point Alpha 0.7

Start Color

Blue

End Color

Bright Orange

☑ Remove NAs

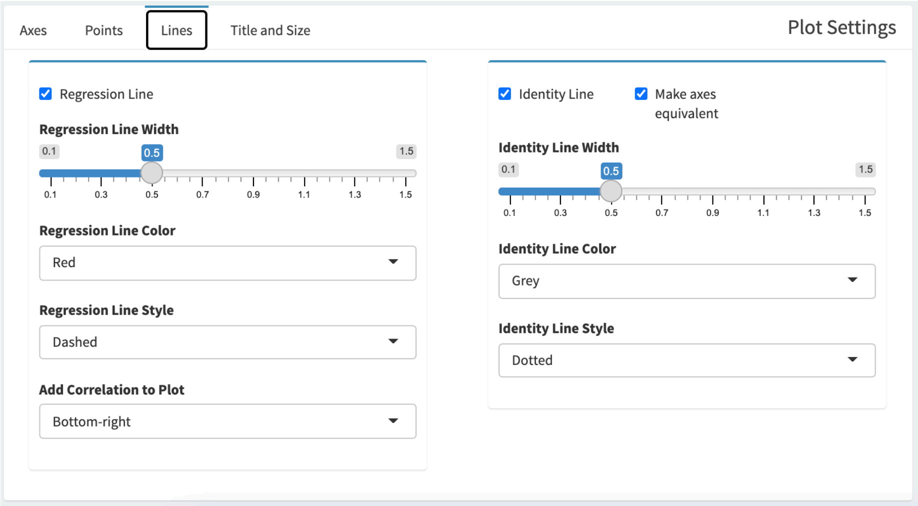
Figure 5.12: The Points tab in the Plot Settings box.

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



☑ Regression Line

Regression Line Width: 0.5 (0.1–1.5)

Regression Line Color: Red

Regression Line Style: Dashed

Add Correlation to Plot: Bottom-right

☑ Identity Line ☑ Make axes equivalent

Identity Line Width: 0.5 (0.1–1.5)

Identity Line Color: Grey

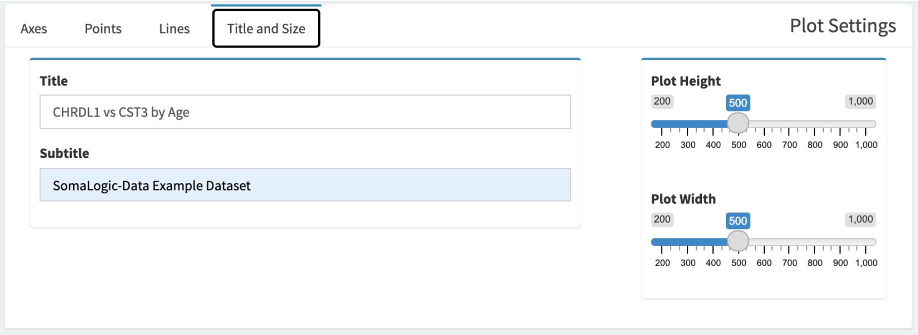
Identity Line Style: Dotted

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

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, 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 Line Style** select box. The correlation can be added to the plot by selecting an position option from the **Add Correlation 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, known as the the concordance. Selecting **Make axes equivalent** will create a square plot in which the extents of the X- and Y-axes are the same. The width of the identity line can be adjusted 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



Title
CHRDL1 vs CST3 by Age

Subtitle
SomaLogic-Data Example Dataset

Plot Height
500

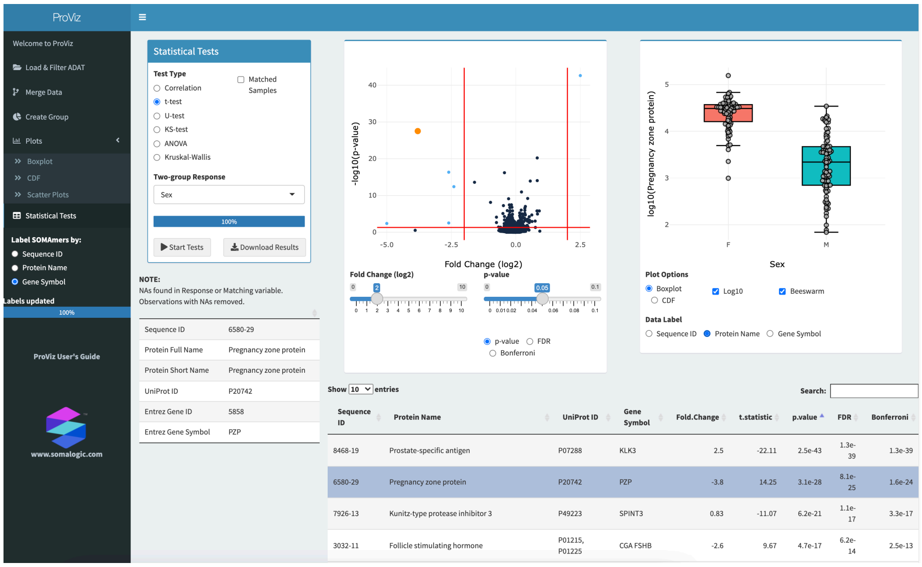
Plot Width
500

Figure 5.14: The Title and Size tab int he 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.

Chapter 6

Statistical Tests



Test Type
○ Correlation
● t-test
○ U-test
○ KS-test
○ ANOVA
○ Kruskal-Wallis

□ Matched Samples

Two-group Response: Sex

Label SOMAmers by:
○ Sequence ID
○ Protein Name
● Gene Symbol

Labels updated 100%

NOTE:
NAs found in Response or Matching variable.
Observations with NAs removed.

Sequence ID	6580-29
Protein Full Name	Pregnancy zone protein
Protein Short Name	Pregnancy zone protein
UniProt ID	P20742
Entrez Gene ID	5858
Entrez Gene Symbol	PZP

Fold Change (log2): 0 – 2 – 10
p-value: 0 – 0.05 – 0.1
● p-value ○ FDR ○ Bonferroni

Plot Options
● Boxplot ○ CDF
☑ Log10
☑ Beeswarm

Data Label
○ Sequence ID ● Protein Name ○ Gene Symbol

Show 10 entries

Search:

Sequence ID	Protein Name	UniProt ID	Gene Symbol	Fold.Change	t.statistic	p.value	FDR	Bonferroni
8468-19	Prostate-specific antigen	P07288	KLK3	2.5	-22.11	2.5e-43	1.3e-39	1.3e-39
6580-29	Pregnancy zone protein	P20742	PZP	-3.8	14.25	3.1e-28	8.1e-25	1.6e-24
7926-13	Kunitz-type protease inhibitor 3	P49223	SPINT3	0.83	-11.07	6.2e-21	1.1e-17	3.3e-17
3032-11	Follicle stimulating hormone	P01215, P01225	CGA FSHB	-2.6	9.67	4.7e-17	6.2e-14	2.5e-13

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 is also available by checking **Matched Samples** and providing a properly constructed **Matching Variable**
- 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 by checking **Matched Samples** and providing a properly constructed **Matching Variable**
- 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; this test is available instead of Kruskal-Walllis when **Matched Samples** is checked

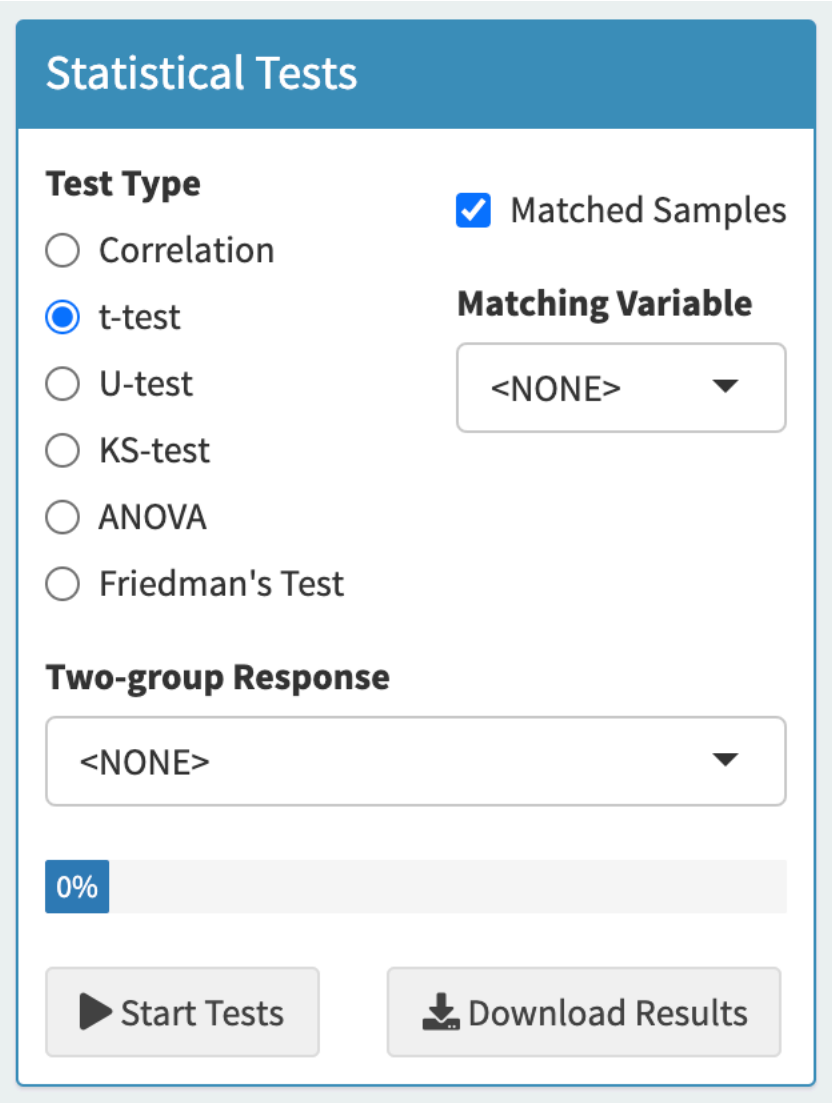
6.2 Selecting a Response

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** (for **Correlation tests**), **Two-group Response** (for **t-tests**, **U-tests**, or **KS-tests**), or **Multi-group Response** (for **ANOVA**, **Kruskal-Wallis** or **Friedman's Test**). 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 variables 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.

6.3 Paired, matched or repeated measures tests

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**



Statistical Tests

Test Type

○ Correlation

● t-test

○ U-test

○ KS-test

○ ANOVA

○ Friedman's Test

☑ Matched Samples

Matching Variable

<NONE>

Two-group Response

<NONE>

0%

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

to adjust the data as necessary. This may, however result in errors. If missing values (NAs) are found in any group labels (Two-group Response or Multi-group Response), or any Matching Variable, the observations with those missing values will be removed and a notice will be displayed.

The following tables illustrate how Grouping and Matching variables should be constructed for the 2-group and multi-group tests, as well as their paired or repeated measures counterparts.

For a two-group test, such as a t-test, U-test, or KS-test, samples must be divided into 2 groups. In Table 6.1, each sample is assigned to either *Group_A* or *Group_B*. If each subject is unique - no repeated samples are taken from the same subject - then only the *SampleId* and *Group_Id* columns are needed. In the example in Table 6.1, there would be 10 subjects, 5 in *Group_A* and 5 in *Group_B*, and the *Matching_Id* column is not needed.

If, however, samples were taken from each subject twice - for example, a pre-treated and post-treated sample taken from every subject - an additional column is required to indicate which samples came from the same subject. In the example in Table 6.1, there would be only 5 subjects, and a sample was taken from each subject twice. The *Matching_Id* column indicates which samples came from the same subjects across the 2 groups.

For a multi-group test, such as ANOVA, Kruskal-Wallis, repeated measures ANOVA and Friedman's Test, the Grouping and Matching works similarly but with more than two groups.

If each subject is unique - no repeated samples are taken from the same subject - then only the *SampleId* and *Group_Id* columns are needed. In the example shown in Table 6.2, there would be 20 subjects, 5 in each group and the *Matching_Id* column is not needed.

If, however, samples were taken from each subject twice - for example, a time course study with samples taken from each subject at each of 4 timepoints - an additional column is required to indicate which samples came from the same subject. In the example in Table 6.2, there would be only 5 subjects, and a sample was taken from each subject four times. The *Matching_Id* column indicates which samples came from the same subjects across the 4 groups.

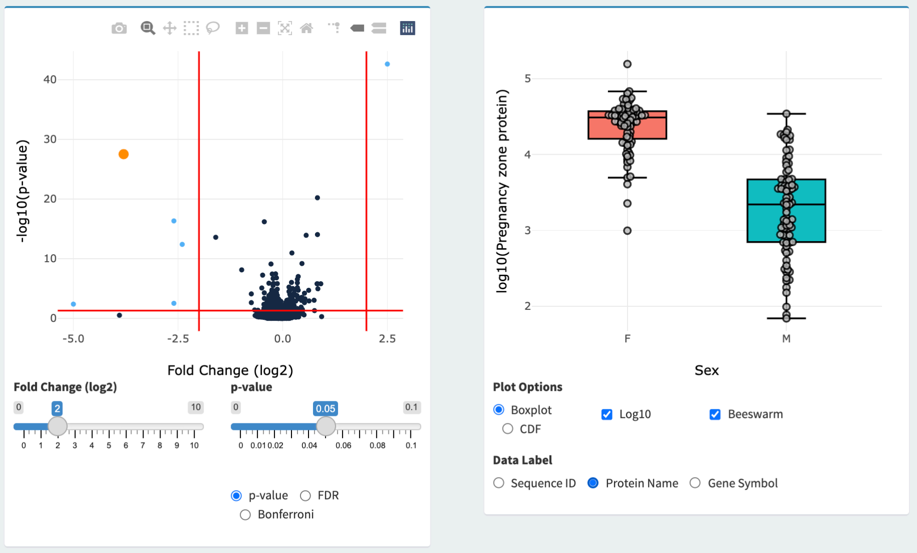
Once a test has been conducted, the *Statistical Results Table* will contain all test results and can be downloaded as a comma-separated file (.csv) by clicking the **Download Results** button.

6.4 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

Table 6.1: Two-Group Test Grouping and Matching Variable Example

SampleId	Group_Id	Matching_Id
Sample_1	Group_A	Subject_1
Sample_2	Group_A	Subject_2
Sample_3	Group_A	Subject_3
Sample_4	Group_A	Subject_4
Sample_5	Group_A	Subject_5
Sample_6	Group_B	Subject_1
Sample_7	Group_B	Subject_2
Sample_8	Group_B	Subject_3
Sample_9	Group_B	Subject_4
Sample_10	Group_B	Subject_5
Sample_1	Group_A	Subject_1
Sample_2	Group_A	Subject_2
Sample_3	Group_A	Subject_3
Sample_4	Group_A	Subject_4
Sample_5	Group_A	Subject_5
Sample_6	Group_B	Subject_1
Sample_7	Group_B	Subject_2
Sample_8	Group_B	Subject_3
Sample_9	Group_B	Subject_4
Sample_10	Group_B	Subject_5



Fold Change (log2): 0 [2] 10

p-value: 0 [0.05] 0.1

● p-value ○ FDR
○ Bonferroni

Plot Options

● Boxplot
○ CDF
☑ Log10
☑ Beeswarm

Data Label

○ Sequence ID ● Protein Name ○ Gene Symbol

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

Table 6.2: Multi-Group Test Grouping and Matching Variable Example

SampleId	Group_Id	Matching_Id
Sample_1	Group_A	Subject_1
Sample_2	Group_A	Subject_2
Sample_3	Group_A	Subject_3
Sample_4	Group_A	Subject_4
Sample_5	Group_A	Subject_5
Sample_6	Group_B	Subject_1
Sample_7	Group_B	Subject_2
Sample_8	Group_B	Subject_3
Sample_9	Group_B	Subject_4
Sample_10	Group_B	Subject_5
Sample_11	Group_C	Subject_1
Sample_12	Group_C	Subject_2
Sample_13	Group_C	Subject_3
Sample_14	Group_C	Subject_4
Sample_15	Group_C	Subject_5
Sample_16	Group_D	Subject_1
Sample_17	Group_D	Subject_2
Sample_18	Group_D	Subject_3
Sample_19	Group_D	Subject_4
Sample_20	Group_D	Subject_5

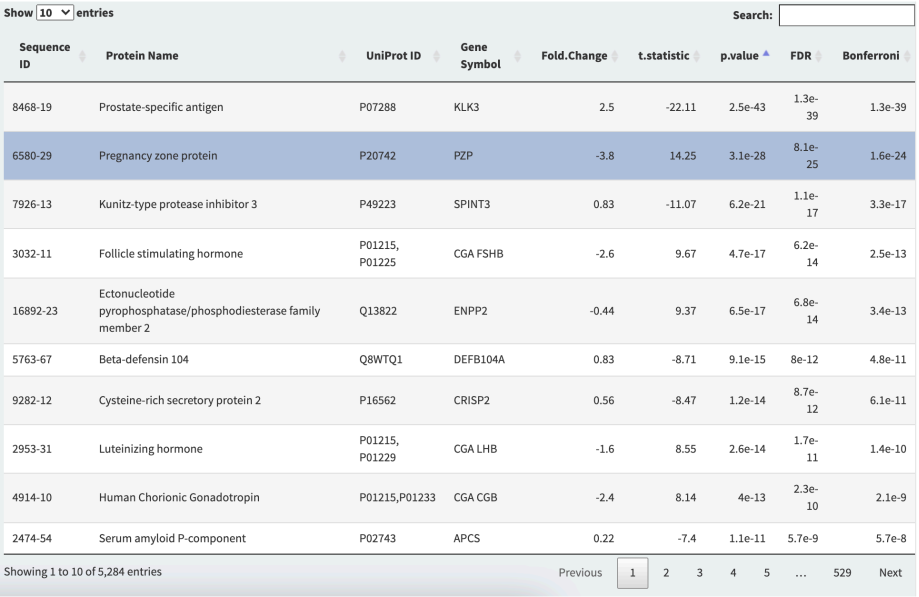
each SOMAmer-detected protein as a point with the X-axis being the log2(fold change) and the Y-axis being -log10(p-value). For a correlation test, the X-axis will illustrate the Pearson's Correlation Coefficient. For ANOVA, Kruskal-Wallis, and Friedman's Test, the X-axis will illustrate the maximum fold-change between medians of all 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 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 reagent associated with that point. Additionally, a table will be displayed below the **Statistical Tests** box with summary information as well as a plot to the right illustrating the data behind the selected point. For correlation tests, a scatter plot of the SOMAmer-detected protein data versus the response will be displayed. For all other tests, boxplots or CDF plots for that SOMAmer-detected protein 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 Sequence 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.

6.5 Statistical Results Table

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



Sequence ID	Protein Name	UniProt ID	Gene Symbol	Fold.Change	t.statistic	p.value	FDR	Bonferroni
8468-19	Prostate-specific antigen	P07288	KLK3	2.5	-22.11	2.5e-43	1.3e-39	1.3e-39
6580-29	Pregnancy zone protein	P20742	PZP	-3.8	14.25	3.1e-28	8.1e-25	1.6e-24
7926-13	Kunitz-type protease inhibitor 3	P49223	SPINT3	0.83	-11.07	6.2e-21	1.1e-17	3.3e-17
3032-11	Follicle stimulating hormone	P01215, P01225	CGA FSHB	-2.6	9.67	4.7e-17	6.2e-14	2.5e-13
16892-23	Ectonucleotide pyrophosphatase/phosphodiesterase family member 2	Q13822	ENPP2	-0.44	9.37	6.5e-17	6.8e-14	3.4e-13
5763-67	Beta-defensin 104	Q8WTQ1	DEFB104A	0.83	-8.71	9.1e-15	8e-12	4.8e-11
9282-12	Cysteine-rich secretory protein 2	P16562	CRISP2	0.56	-8.47	1.2e-14	8.7e-12	6.1e-11
2953-31	Luteinizing hormone	P01215, P01229	CGA LHB	-1.6	8.55	2.6e-14	1.7e-11	1.4e-10
4914-10	Human Chorionic Gonadotropin	P01215,P01233	CGA CGB	-2.4	8.14	4e-13	2.3e-10	2.1e-9
2474-54	Serum amyloid P-component	P02743	APCS	0.22	-7.4	1.1e-11	5.7e-9	5.7e-8

Showing 1 to 10 of 5,284 entries

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