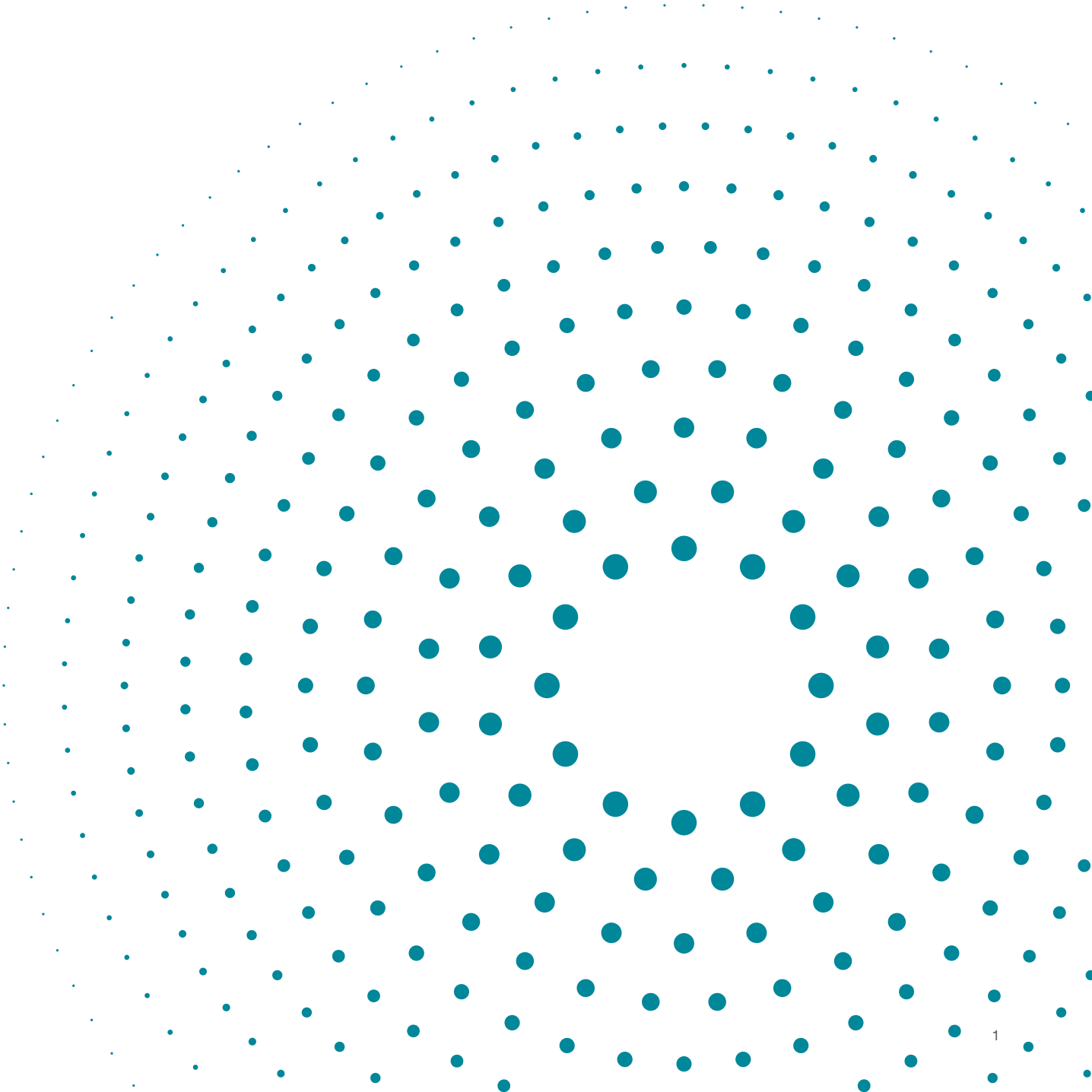
User Guide

Olink® Insights Stat Analysis



Introduction

Olink has extensive coverage of the plasma proteome and can deliver high quality data for over 1000 unique proteins. However, when it is time to perform data analysis, this amount of data can be overwhelming. To help with that, Olink provides a web-based application for basic data visualizations and statistical analyses based around the same tools and methods used by the Olink Data Science team — Olink Insights Stat Analysis.

With this app it is easy to:

- View sample-wise distributions of NPX values
- Generate Principal Component Analysis (PCA) plots
- Generate heatmaps
- Perform basic statistical analyses: *t*-test (2 groups), Analysis of Variance (3+ groups)

The application can be accessed here: https://olinkproteomics.shinyapps.io/OlinkInsightsStatAnalysis/

To get started, you need an NPX data file generated from NPX manger and delivered to you by Olink Analysis Services or one of our certified core labs. If you do not have access to an NPX file but would still like to explore the features of the app, open the app, and on the **About** tab, select **Use example data**.

For questions, contact biostattools@olink.com.

Data and sample information input

Upload data into the application environment on the **NPX and sample information** tab. As long as the data file has been exported directly from NPX Manager or is the unmodified file that was delivered by Olink Analysis Service, the upload can be done in a few clicks. Uploading sample information is also easy.

NPX files

To upload an NPX file in the application environment, click **Browse** under the **Select NPX file** heading and select the NPX file you want to upload. If the file is in the proper format, basic details will be displayed in the NPX file box (number of samples detected, number of assays detected and number of panels detected) and additional tabs will get accessible.

If the file is not in a known NPX format, you will receive an error message and no additional tabs will be available. If this happens, ensure that the NPX file is in the exact format it was in when it was delivered from Olink Analysis Service or exported from NPX manager. If it is an .xlsx file, ensure that the first tab is the original spreadsheet.

Exclude QC Warnings

If the **Exclude QC warnings** check box in the Select NPX box on the NPX and sample information tab is checked, all visualizations will remove samples marked as having a QC warning.

If the box is checked before the start of analysis, these samples will be removed from the analysis.

Note: If the box is checked after analysis has already been performed, you need to click **Analyze** again.

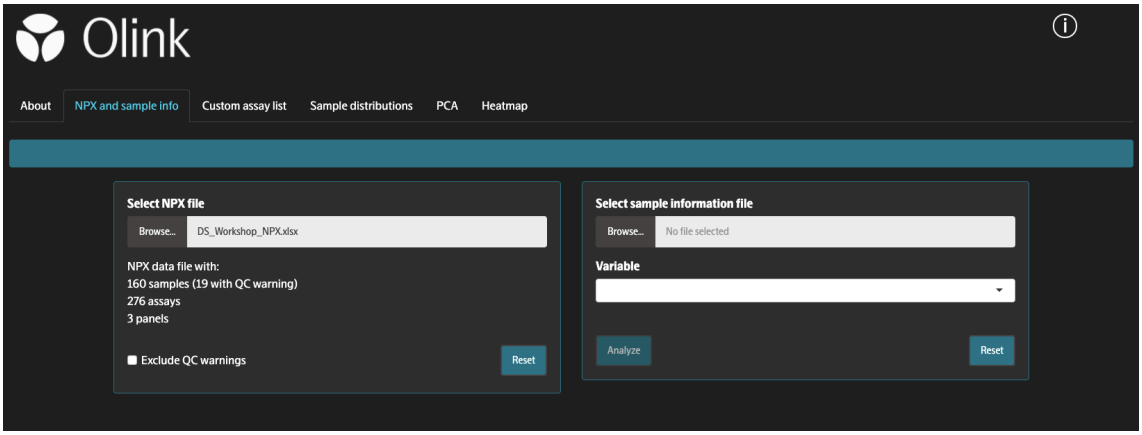


Figure 1 Successful upload of an NPX file. Counts of samples, assays and panels are shown below the file field.

Sample information

Sample information files are needed if you want to perform basic statistical analysis on your data, but not required if you only want to use the data visualization features. These files are just as easy to upload as NPX data but they must be in a required format. Acceptable file types for these files are .xls, .xlsx and .csv files.

Create a file

Create an Excel or comma-separated values (csv) file with at least two columns. The first column *must* be named **SampleID**. Note that this name is case-sensitive. ID's in the **SampleID** column of the sample information file must be *identical* to the Sample ID's listed in the NPX file. You can view these ID's by opening the NPX file and looking at the first column. All additional columns can have any names and will be used to select the variables in the application.

Note: Special characters will be removed from column names and column values and might therefore not appear identical to what is in the spreadsheet or .csv file. These column names and values will be used to annotate visualizations, so choose descriptive names formatted for your needs.

	A	B	C	D	E
1	SampleID	SubjectID	Group	Treatment	Time
2	Sample1	ID1	Group 1	Treated	Day 1
3	Sample10	ID1	Group 1	Treated	Day 2
4	Sample100	ID1	Group 1	Treated	Day 3
5	Sample101	ID2	Group 2	Utreated	Day 1
6	Sample102	ID2	Group 2	Utreated	Day 2
7	Sample103	ID2	Group 2	Utreated	Day 3
8	Sample104	ID3	Group 2	Treated	Day 1
9	Sample105	ID3	Group 2	Treated	Day 2
10	Sample107	ID3	Group 2	Treated	Day 3
11	Sample108	ID4	Group 2	Untreated	Day 1
12	Sample109	ID4	Group 2	Untreated	Day 2
13	Sample110	ID4	Group 2	Untreated	Day 3

Figure 2 Example sample information file. Note that the first column is labeled SampleID and that these ID's match the sample ID's in your NPX data file.

Upload the file

Click **Browse** in the **Select sample information file** box and select the created sample information file. If you have an NPX file uploaded already, you will see a count of overlapping sample ID's found in both the sample information file and the NPX file in the message bar, as well as a count of each variable level from the variable selected in the **Variable** drop-down menu. If no NPX file has been previously uploaded, you will be asked to upload an NPX file to perform the analysis.

If analysis is performed and only a subset of sample ID's are present in the sample information file, only these samples will be shown in the visualizations and used for any statistical analysis.

Protein subsets

It is possible to restrict the data visualizations to a subset of proteins of interest. This can be done in two ways.

Create list on the Custom assay list tab

Another way to create a custom list of assays is to use the table on the **Custom assay list** tab. To do this, click and highlight assays you want to add to the list. All highlighted assays will be used when you select the **Selected assay list** in the data visualization **Panel** drop-down menu. This can be useful for quickly subsetting assays after data analysis has been performed. For example, after performing a *t*-test or ANOVA, you can create a custom list containing only statistically significant assays after completed analysis. To do this, navigate to the **Custom assay list** tab, click in the **Threshold** search box and select **Significant**. This will filter the table to contain only significant assays. Then click **Select all**. All significant assays are now highlighted and available in the **Panel** drop-down menu for each visualization tab.

Create list in spreadsheet

The first way is to create the list of assays that you want to use in the visualization in a spreadsheet. You only need a single column and can identify the assays using either the OlinkID, protein name or UniProt ID, all of which are listed in the original NPX file. The column containing the ID's *must* be appropriately named with either **OlinkID**, **Assay** or **UniProt**. These are case-sensitive. Once this file is created, click **Browse** on the **Custom assay list** tab and select the file. If successful, this list will be selectable in the **Panel** drop-down menus as **Uploaded assay list** on each visualization tab.

	A
1	OlinkID
2	OID00468
3	OID00556
4	OID00689
5	OID00677
6	OID00532
7	OID00553
8	OID00500
9	OID00729
10	OID00389
11	OID00744
12	OID00667
13	OID00722
14	OID00494
15	OID00478

Figure 3 Example custom assay list using OlinkID's. Note that the first column is labeled OlinkID and these ID's match those in the NPX data file.

Data visualizations

The application provides multiple global views of the data that can be useful on their own or combined with a sample information file.

Modebar buttons

Within each visualization there is a button bar in the top right corner of the plot that provides some useful features. The bar is visible when hovering over the plot. Each button and its function is described below.

- **Download as PNG**: Click to download the visualization as a .png file. The dimensions of the figure will be based on the scaling of the browser window.
- **Zoom**: Click to enter zoom mode. To zoom, click and drag on the visualization to draw a box. Drag

vertically to zoom on the y-axis and drag horizontally to zoom on the x-axis. Click **Reset axes** or double-click to zoom out.

- **Pan**: While zoomed, click to enter pan mode. It is now possible to click and drag to move the viewing window around the visualization.
- **Reset axes**: Click this button to zoom out.

Figure 4 Modebar buttons shown on visualizations. From left to right, buttons are Download as PNG, Zoom, Pan and Reset axes.

View panels

Every visualization tab also has options for selecting which panel to view. It is also possible to view all panels at once. However, for very large projects, the **Heatmap view** may be restricted to viewing only a single panel at a time. Data points can also be colored in each visualization using the **Color** drop-down menu. If analysis has not been performed, only **QC_Warning** will be available. If a sample information file has been uploaded and analysis performed, all variables from the sample information file will be available for coloring of data points.

Sample distributions tab

Three types of visualizations are available on the Sample distributions tab and all are useful for identifying samples that may have NPX distributions that do not look like the rest of the samples or could potentially be outliers.

- **Boxplot**: This figure has sample ID's along the x-axis and NPX values along the y-axis. Data is displayed as a box and whisker plot. Lower and upper ends of the boxes represent the 1st and 3rd quartiles of the data. The center line represents the median. Lower and upper fences show the minimum or maximum observed value that is within 1.5 times the inter-quartile range.
- **Density**: This figure shows NPX values along the x-axis and the relative proportion of assays along the y-axis. Each line indicates a separate sample. This plot is a smoothed version of a histogram. See Links on page 10 for more details.
- **Median vs IQR**: This figure reduces the distribution of NPX values for each sample down to two values; the median and the inter-quartile range (defined as the difference between the 75th and 25th percentiles). Horizontal and vertical dashed lines indicate +/- 3 standard deviations of all sample medians (x-axis) and IQR's (y-axis). These are meant only as a guide, not as a criteria for sample exclusion. Hover over a point to display the sample ID. Click on a point to add an arrow to the point and a label for the sample ID. Click on the point again to remove the label.

Figure 5 Dropdown menus available on the Sample distributions tab. The first menu on the left selects which panels to gather assays from. The second menu allows you to select which variable to use to color data points by. The last menu allows you to change the plot type.

PCA

Principal component analysis (PCA) is performed on the NPX data file and all samples are plotted as a scatterplot along the selected principal components. For plotting purposes, any samples with missing NPX values are imputed with the median NPX value from all samples for the given missing assay. All assay distributions are then centered at 0 and scaled to have a standard deviation of 1 before performing the PCA

analysis.

PCA plots can be useful for showing that samples separate or cluster based on variables of interest (e.g. by treatment, case/control, etc.). In addition, it can be used to identify potential outlier samples. Hover over a point to display the sample ID. Click on a point to add an arrow to the point and a label for the sample ID. Click on the point again to remove the label. See Links on page 10 for more information about PCA.

Heatmap

Heatmaps can sometimes provide useful overviews of large amounts of data for all samples. In this implementation, any samples with missing NPX values are imputed with the median NPX value from all samples for the given missing assay. All assay distributions are then centered at 0 and scaled to have a standard deviation of 1. Hierarchical clustering based on centered and scaled NPX values is performed on both samples and assays to determine row and column ordering. The centered and scaled NPX values are then used to color the heatmaps. Row side colors are added to the figure based on the available sample information. See Links on page 10 for more for more information about heatmaps.

Data analysis

This application provides the ability to easily perform basic statistical analysis on every assay. Results are then presented in a manageable table. Currently, two types of analyses are available and are automatically determined by the number of groups within a selected variable. If the selected analysis variable has two groups, a *t*-test will be performed on every assay. If the variable has more than two groups, Analysis of Variance (ANOVA) will be performed on every assay. All *p*-values are adjusted for multiple testing using the Benjamini–Hochberg method (see Links on page 10 for details). Performing the analysis is simple but a sample information file must be uploaded (see Data and sample information input on page 2).

After both the NPX file and associated sample information file have been uploaded, select the variable from the drop-down menu in the **Select sample information file** box that contains the variable you want to analyze. Then click **Analyze**, and a new tab will open with the results of either the *t*-test or ANOVA analysis.

Description of *t*-test

When two groups are detected in the selected variable from the sample information file, a *t*-test is performed for every assay in the NPX file. For a description of a *t*-test, see Links on page 10. This test is implemented using the t.test() function in base R. The *t*-test tab displays a table of the results from every assay (column descriptions are provided below). Click **Download full t-test results** at the top of the table to download the results. Tables are sorted by *p*-value by default. Click on a column name to sort the table by that column.

A volcano plot shows the NPX difference between the two groups on the x-axis and the -log10 of the nominal *p*-value on the y-axis is displayed below the table. Hover over a point to show the protein name. Click on a point in the volcano plot to label that point with the protein name. Click it again to remove the label.

Use the buttons at the top to toggle between the volcano plot and boxplots of raw NPX values for each group. Click on a row in the table to select the assay being plotted. A horizontal bar is displayed above the boxes if the difference is statistically significant.

Column descriptions

- **Assay**: Protein name
- **OlinkID**: Unique Olink assay identifier
- **UniProt**: UniProt ID

- **Panel**: Olink panel name
- **Difference**: Estimate mean NPX difference between the two groups
- **Group 1**: Mean NPX value for group 1.
  Note: This column name will depend on the label in the group variable
- **Group 2**: Mean NPX value for group 2.
  Note: This column name will depend on the label in the group variable
- **p.value**: Nominal, unadjusted *p*-value from the *t*-test
- **Adjusted_pval**: Benjamini-Hochberg adjusted *p*-value
- **conf.low**: Lower 95% confidence interval for difference in means (unadjusted)
- **conf.high**: Upper 95% confidence interval for difference in means (unadjusted)
- **Threshold**: Indicator if the adjusted *p*-value is less than 0.05

Figure 6 The *t*-test tab shown after analysis and selection of a variable with two groups.

Description of ANOVA

If more than two groups are detected in the selected sample information variable, ANOVA will be performed on every assay. For a description of ANOVA, see Links on page 10 . The ANOVA analysis is performed in two steps. For a single assay, after the ANOVA model is fit, a global *F*-test is performed using the anova() function in R. The null hypothesis for this test is that all groups have equal mean NPX values. The alternative hypothesis is that at least one group has a mean different from the rest of the groups. *P*-values from this test are adjusted for multiple testing using the Benjamini-Hochberg method. Assays that have an adjusted global *F*-test *p*-value less than 0.05 are then moved on to a post-hoc analysis. This analysis is done to determine which specific groups differ from each other. These *p*-values are adjusted using the Tukey method and means are estimated using the emmeans package in R.

Note: In order for a post-hoc result to be significant, the assay must pass the global *F*-test *p*-value cut-off *and* have a Tukey adjusted *p*-value less than 0.05. Both tables can be downloaded by using the button at the top of each table.

Note: Post-hoc results with *p*-values less than 0.05 for an assay that did not pass the global *F*-test *p*-value threshold, should not be considered statistically significant. This is necessary for properly controlling the Type I error rate.

The ANOVA tab has tables for both the global *F*-test results and the post-hoc results. Use the selector in the top left corner of the tab to toggle between the tables. The columns are described below for these two tables. Tables are sorted by *p*-value by default. Click on a column name to sort the table by that column. The plot below the table display shows boxplots of raw NPX values for each group. Horizontal lines are displayed above the boxes with statistically significantly different means. To select the assay being plotted, click on a row in either of the tables.

Note: The post-hoc table will only show comparisons from assays that pass the global *F*-test *p*-value cutoff. If no assays pass this threshold, post-hoc results for the top 10 assays with the smallest *p*-values will be displayed. If the results are downloaded, this will include post-hoc test results for all assays.

Columns from the global *F*-test table

- **Assay**: Protein name
- **OlinkID**: Unique Olink assay identifier
- **UniProt**: UniProt ID
- **Panel**: Olink panel name
- **term**: Variable selected from the Sample Information file
- **df**: *F*-test degrees of freedom
- **sumsq**: *F*-test sum of squares
- **meansq**: *F*-test mean squares
- **statistic**: *F*-statistic
- **p.value**: Nominal, unadjusted *p*-value from the *t*-test
- **Adjusted_pval**: Benjamini-Hochberg adjusted *p*-value
- **Threshold**: Indicator if the adjusted *p*-value is less than 0.05

Columns from the post-hoc table

- **Assay**: Protein name
- **OlinkID**: Unique Olink assay identifier
- **UniProt**: UniProt ID
- **Panel**: Olink panel name
- **contrast**: Groups being compared
- **estimate**: Estimated difference in mean NPX for the groups being compared
- **conf.low**: Lower 95% confidence interval for difference in means (unadjusted)
- **conf.high**: Upper 95% confidence interval for difference in means (unadjusted)
- **Adjusted_pval**: Tukey adjusted *p*-value
- **Threshold**: Indicator if the adjusted *p*-value is less than 0.05 AND the assay is statistically significant based on the global *F*-test

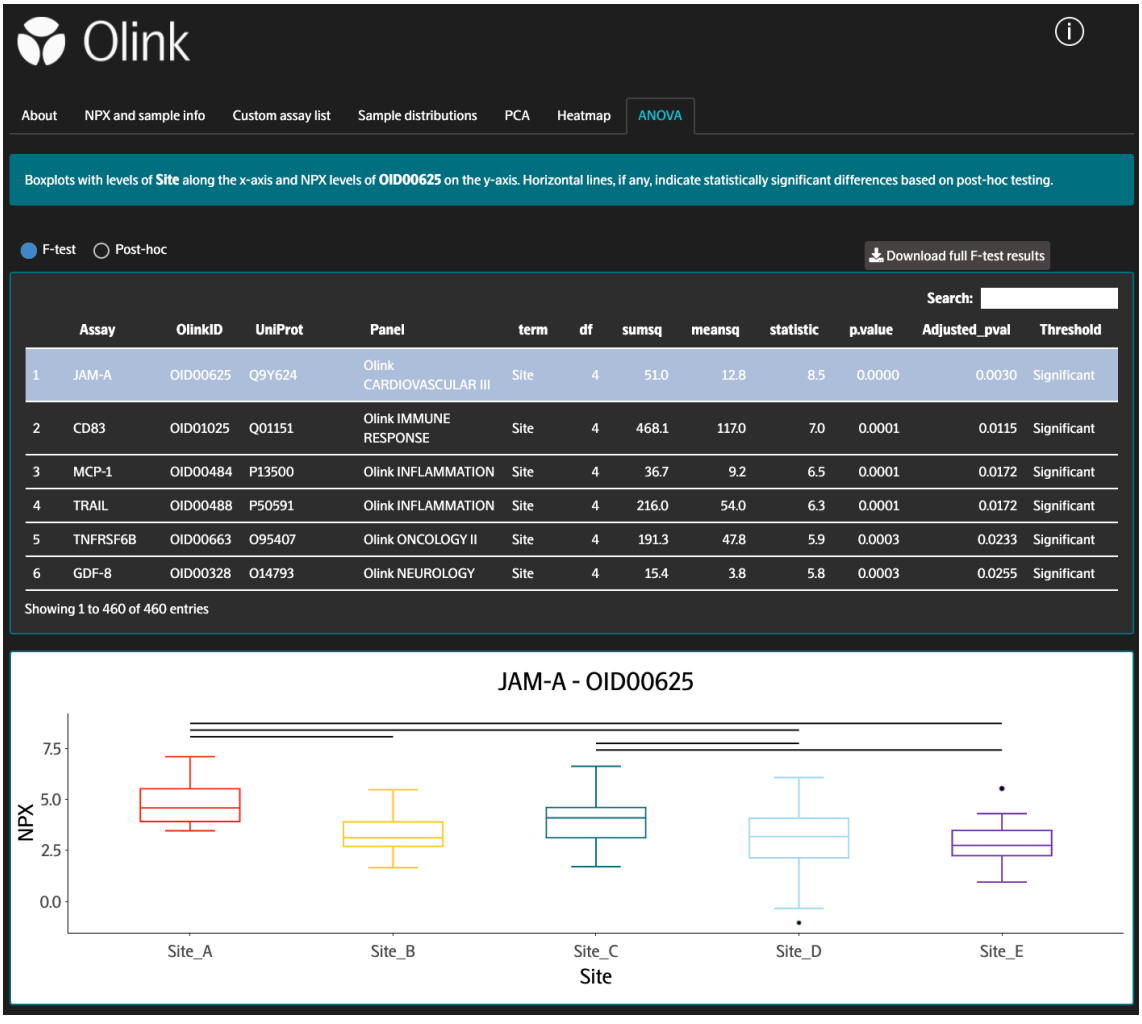


Figure 7 ANOVA tab shown after clicking the analyze button and selecting a variable with more than 2 groups.

Links

Some useful links are included below and provide additional details regarding statistical analyses and visualizations.

Density curves
https://en.wikipedia.org/wiki/Density_estimation

emmeans
https://cran.r-project.org/web/packages/emmeans/index.html

F-test
https://en.wikipedia.org/wiki/F-test

Heatmap
https://en.wikipedia.org/wiki/Heat_map

Interquartile range
https://en.wikipedia.org/wiki/Interquartile_range

p-value adjustment
https://en.wikipedia.org/wiki/False_discovery_rate
https://en.wikipedia.org/wiki/Multiple_comparisons_problem

Principal component analysis
https://en.wikipedia.org/wiki/Principal_component_analysis

t-test
https://en.wikipedia.org/wiki/Welch%27s_t-test

ANOVA
https://en.wikipedia.org/wiki/Analysis_of_variance

www.olink.com

Olink Proteomics, Dag Hammarskjölds väg 52B , SE-752 37 Uppsala, Sweden

1133, v1.0, 2020-06-02