**Use of the “Dimension Reduction” DSP DA plugin**

This vignette is a guide to running the Dimension Reduction DSP DA plugin and interpreting the resulting plots.

**Intended use**

This plug-in was designed for data from the GeoMx DSP Data, and may work best on high plex assays, such as the Cancer Transcriptome Atlas, but can be used with other assays.

This plugin does the following:

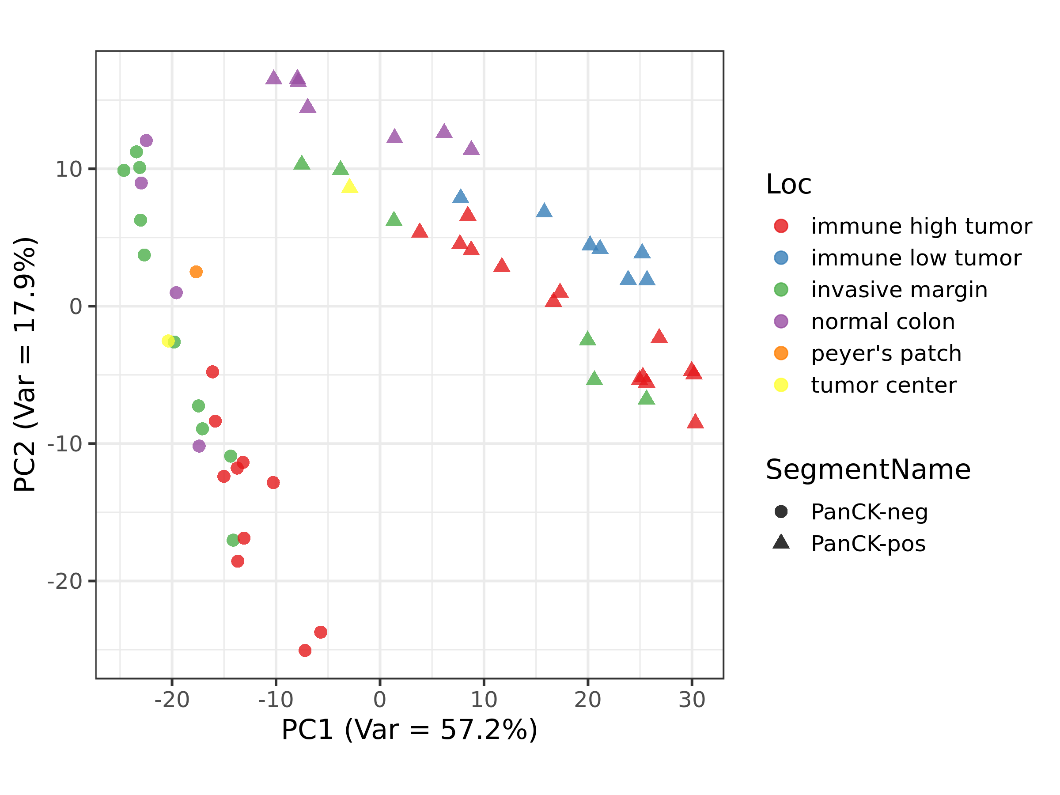
1. Performs dimension reduction analysis of the segments within the study, of the type specified by the user. Options include:
   1. PCA
   2. tSNE
   3. UMAP
2. Plots the resulting first two dimensions against each other as a scatter plot. There are options of users to control the color, shape, and color palette of the resulting plot. Plots are 6 in by 8 in, at 300 dpi PNG files.
3. If a PCA is being shown, it will also graph the cumulative proportion of variance explained by each principal component up to the first 15 components
4. Saves an updated annotation sheet with Dim1, Dim2, and in the case of PCA Dim 3 as well, for the purpose of re-graphing with external software as a CSV.

**Contents:**

Page 2: Loading into the DSP-DA

Page 3: Setting User Parameters

Page 4: Interpretation of resulting files and figures

****

For more example graphs see page 4

**Loading into the DSP-DA:**

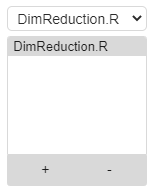
The DimReduction.R file may be loaded into the custom scripts section of the DSP-DA after you have a dataset processed and ready for analysis. To do so open the custom script section by clicking on the button shown below:



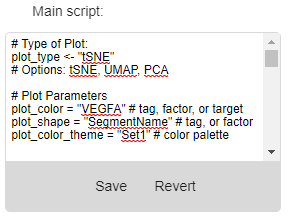
Select the “Manage” tab to open the area to load and edit scripts:



In the Management tab to add a new script and adjust parameters, fill out and then scroll to the bottom of the page. Use the “+” button to add the DimReduction.R file to the script:



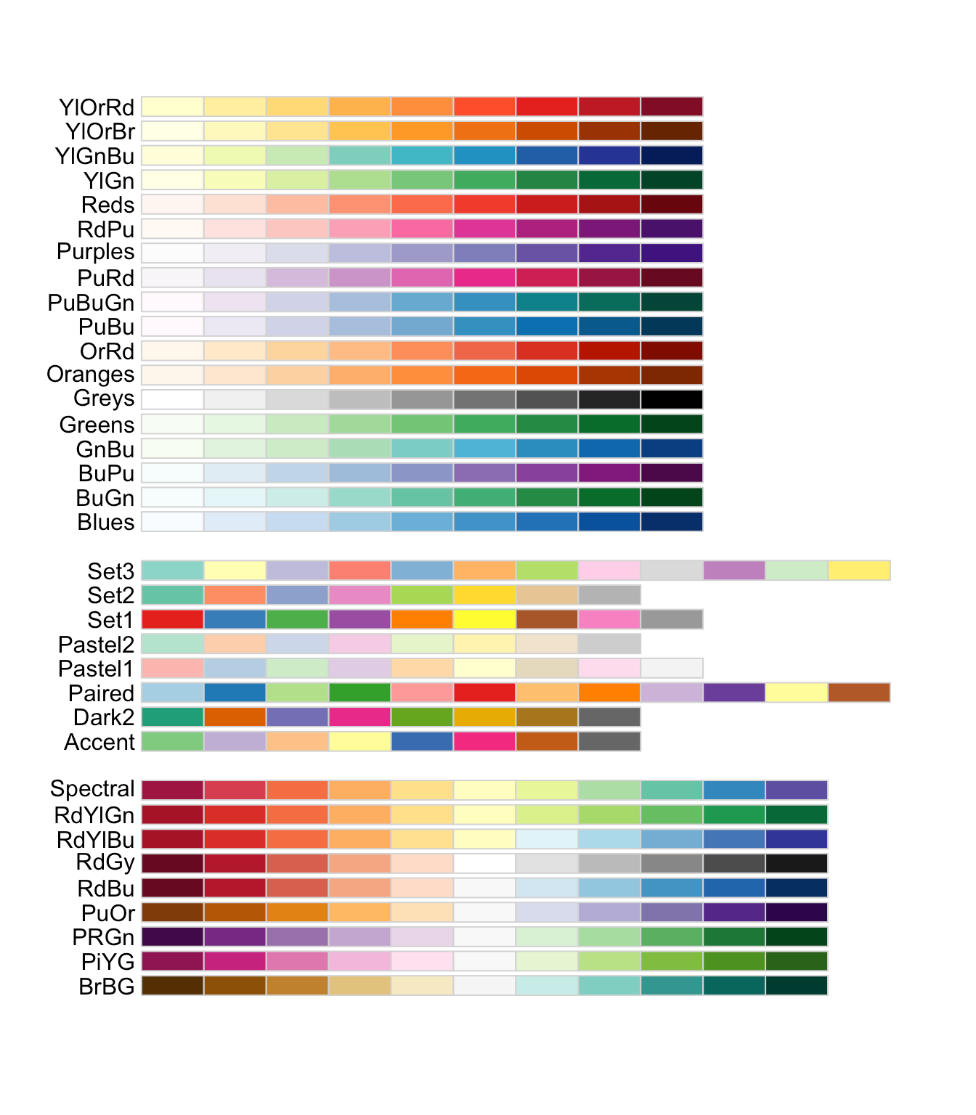
Once added the parameters of the script can be adjusted by editing the top lines in the script and hitting the “Save” button. You do not need to check the Create new dataset button.



Parameter options are described in full on the next page.

**Setting User Parameters:**

There are a few settings that can be adjusted easily by the user at the top of the plug-in script. These include:

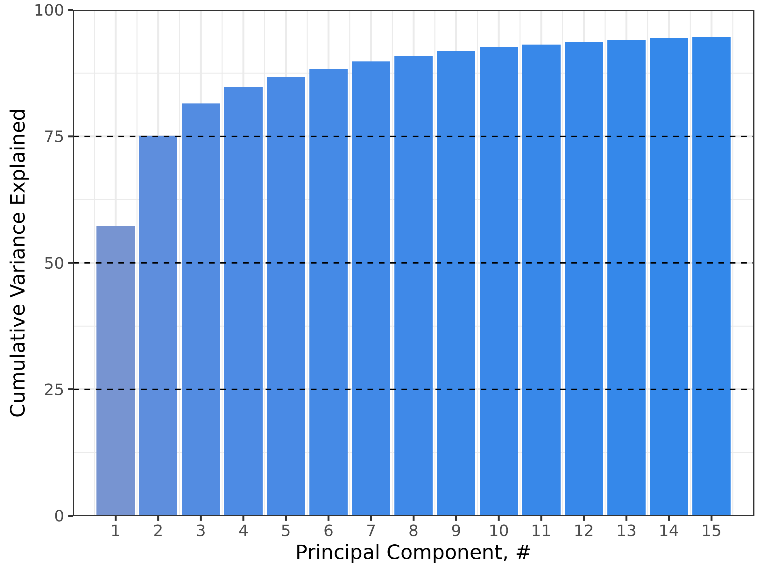
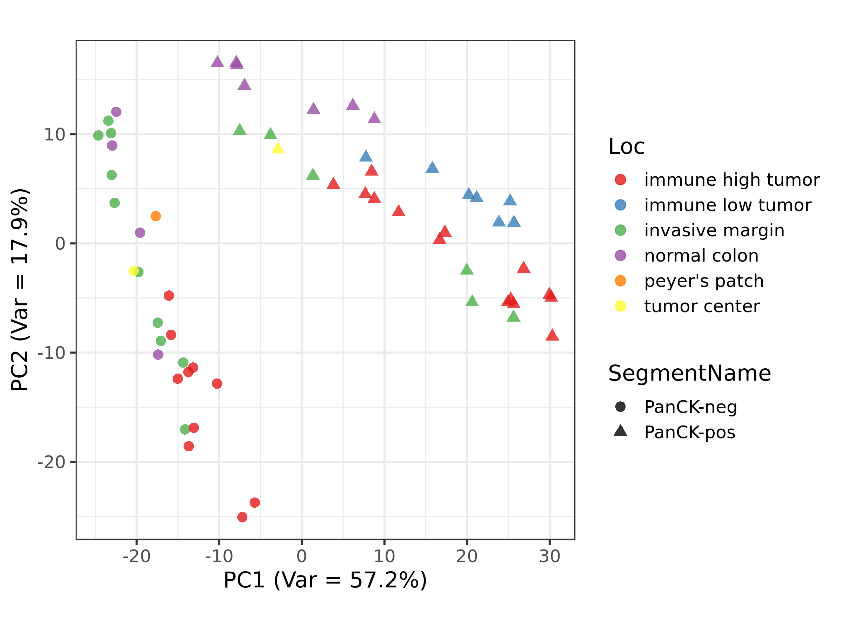
* plot\_type – set this to either “PCA”, “UMAP”, or “tSNE” based on user preference. No other values or methods are currently supported
* plot\_params (plotting parameter list):
  + color – set this to the name of an annotation tag, factor or the display name of a target within the data matrix. For example, you may have a factor named “Loc” that you want to visualize, this may be used with the color parameter. Alternatively, “VEGFA” can be used to color points by the continuous expression of VEGFA.
  + shape - set this to the name of an annotation tag or factor
  + color\_theme – the name of a color palette from the palettes listed below. For example “Dark2” or “Spectral”

**Palette Options:** Names to the left of each palette represent the text that can be used for the plot\_color\_theme variable. This is defaulted to “Set1” but any of the values shown here may be used. Note that palettes with light colors may be harder to distinguish on the graph.

**Interpretation of resulting files and figures**

**PCA**

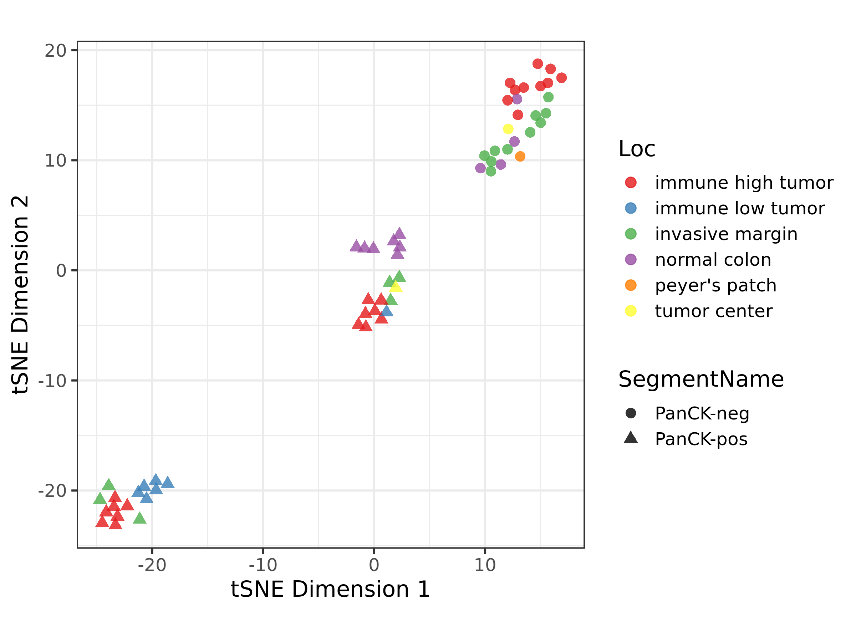
Principal component analysis (PCA), is a method for reducing high dimension data down to lower data spaces. It uses iteratively identifies the linear component that explains the most variation in a dataset, captures it and then looks for orthogonal vectors that would explain the next most amount of variation within the dataset. These principal components can then be used to visualize clusters of samples, as well as understanding the amount of variation that the analysis has captured for any given number of components. An example is shown below of the two graphs output by this method:

****

**tSNE**

tSNE (t-Distributed Stochastic Neighbor Embedding) is a method do cluster samples based on expression that is not linearly or orthogonally constrained like the PCA plot. However, it is a stochastic plot, and cannot be used to estimate where a new sample would fit within the defined clusters. As such, it is useful for data exploration, but less so for defining characteristics that may be shared in a new dataset.

Reference: L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

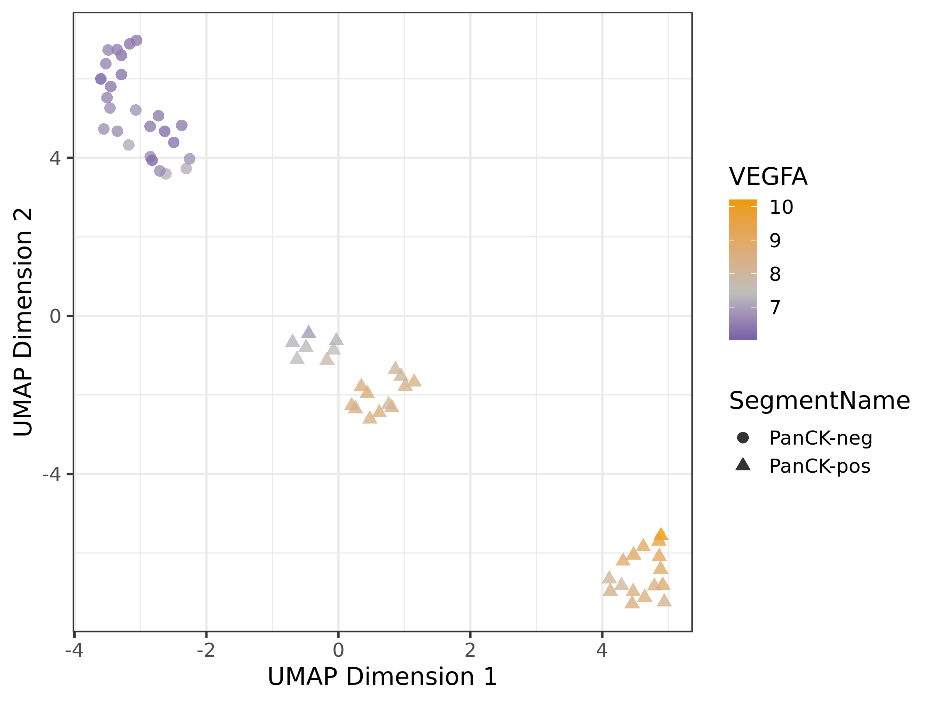


**UMAP**

UMAP (Uniform Manifold Approximation and Projection) is a method of dimension reduction developed for developing a reproducible method for graphing samples in a non-linearly constrained fashion. It has been heavily used by the single- For more information about this method see the reference below:

Reference: McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018

An Example of a UMAP for a set of samples produced is shown below:



In this particular example, color is being defined by the target “VEGFA”, shape is being set by the factor “SegmentName”, and the color palette used is “Set1”. Here we observe 3 clusters. As before, two clusters are PanCK-positive, which show higher expression of VEGFA than the PanCK-negative segments.

**CSV Outputs**

In addition to graphs, the plugin will also output a CSV file with new data columns depending on what variables were used.

New data columns:

* If color is set to a target name it shall be included in the table, with the log2 count values shown for the target based on the active data frame selected
* If UMAP or tSNE is selected 2 new columns (Dim1 and Dim2) will be added. If PCA is selected Dim3 will also be added. Dim1&2 represent the graphed values, Dim3 is added in case users are interested in graphing additional PCs.

These columns will appear at the end of the CSV file.