Getting started with GeoMx® Protein data analysis

Data Analysis Training

Applications Scientist

nanoString®

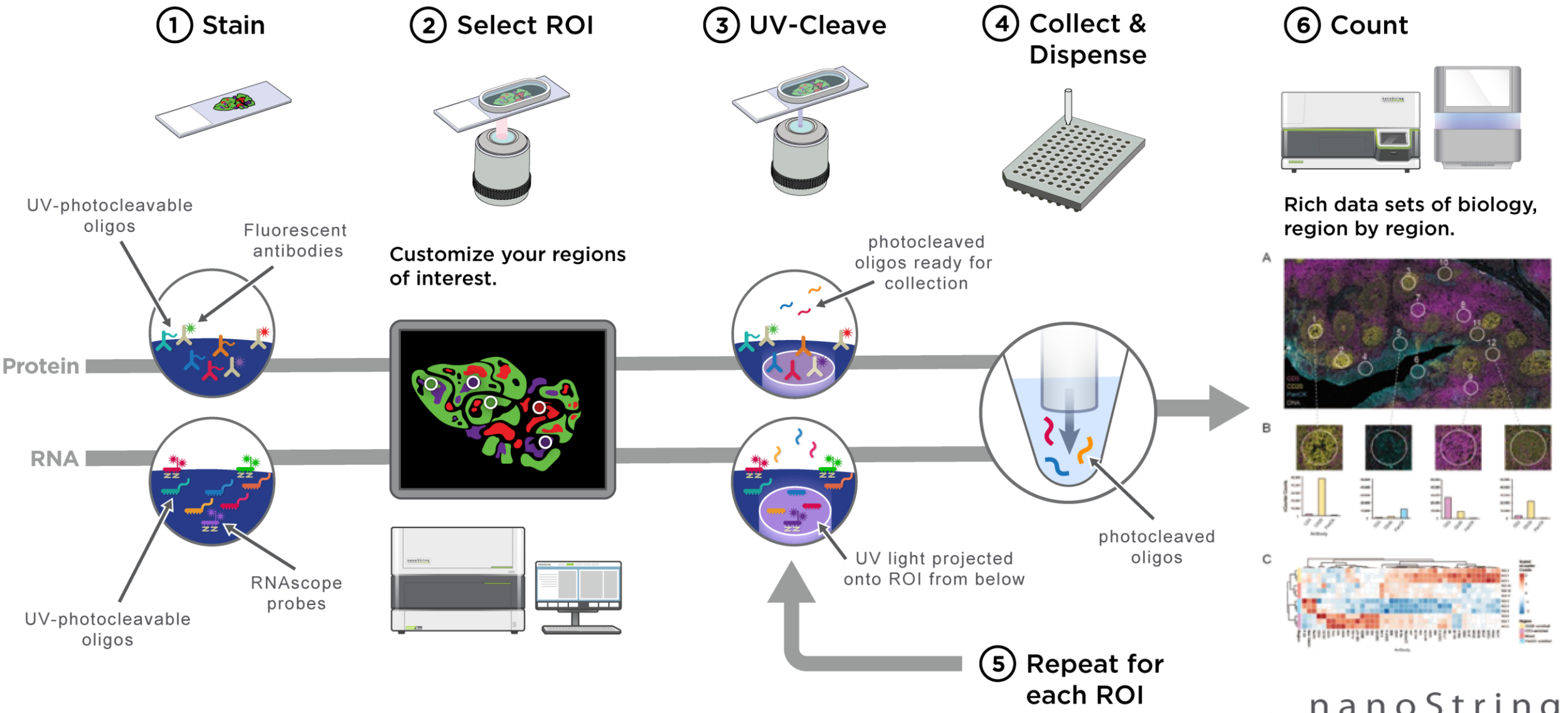
© 2022 NanoString Technologies, Inc., All rights reserved

Goals of this session

- Present a framework to analyze GeoMx® DSP Protein with nCounter® readout data
- Review the main steps of data processing
- Show Available resources
- Q&A

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

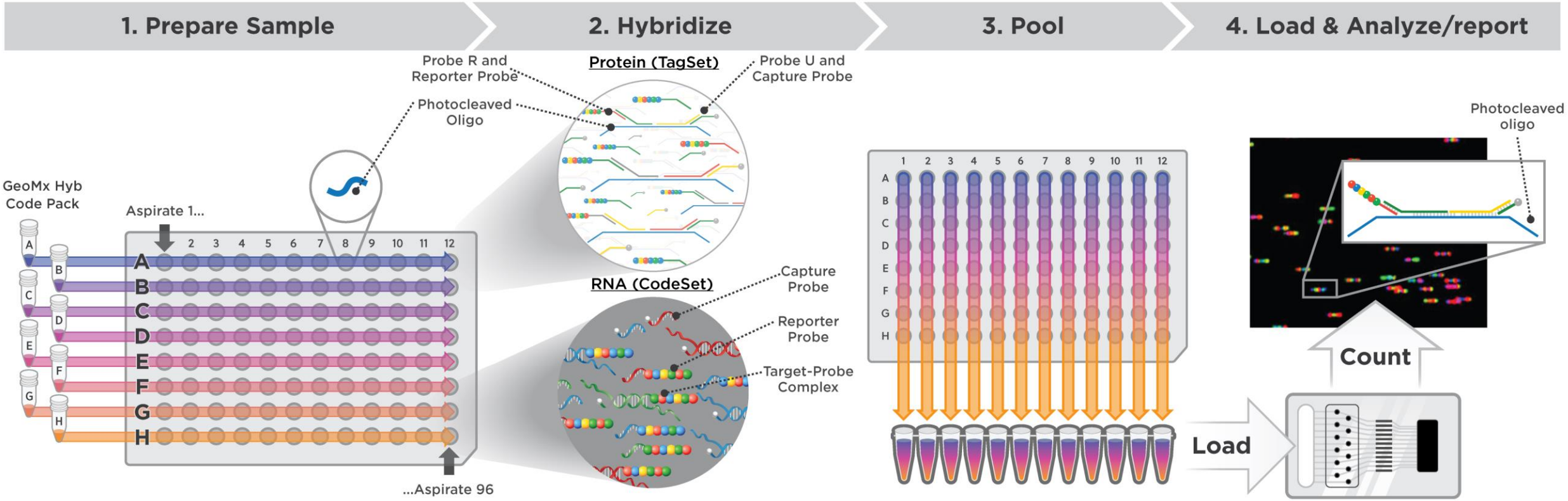
GeoMx Workflow



nanoString

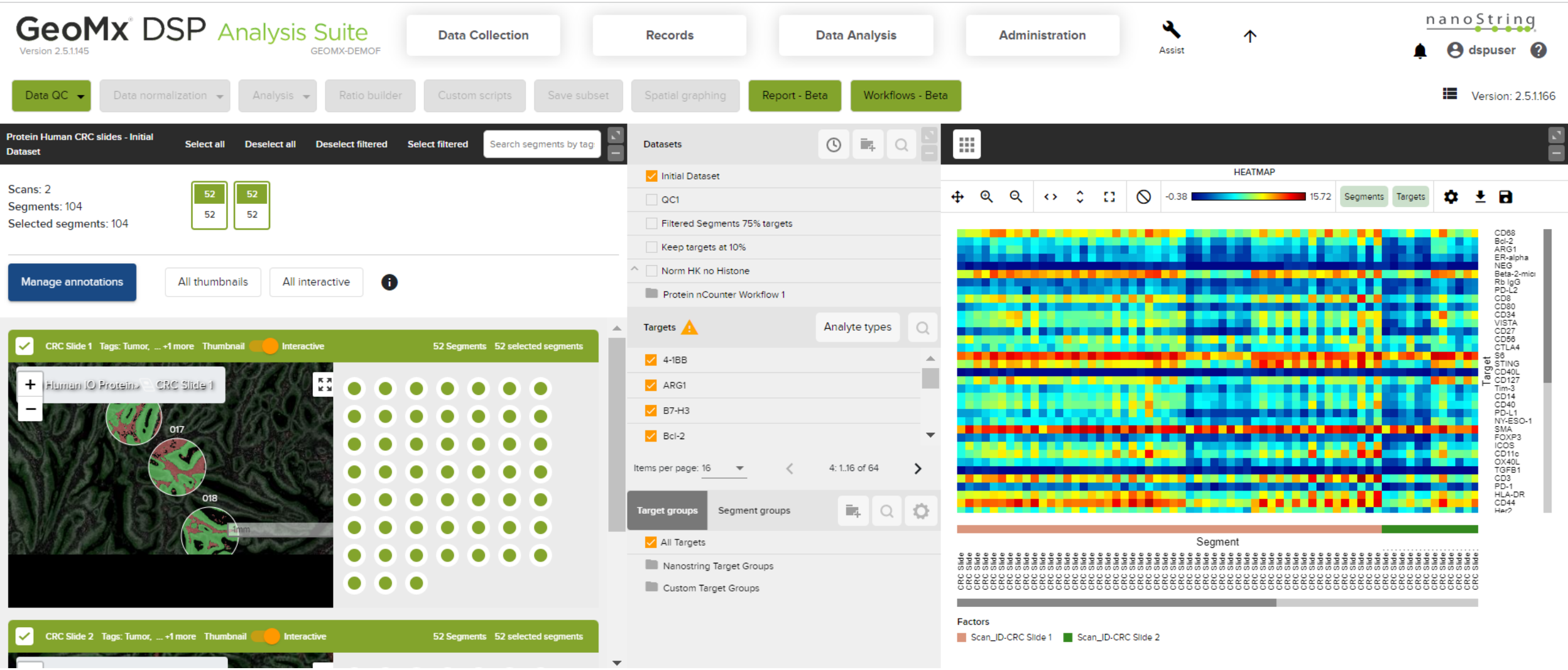
FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Now you have a plate with oligos...



FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

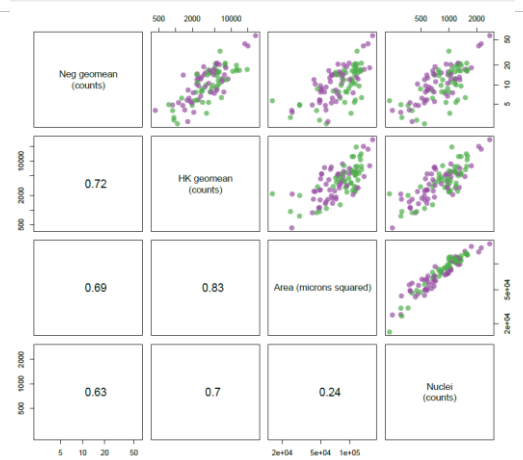
Using the DSPDA to analyse Protein nCounter data

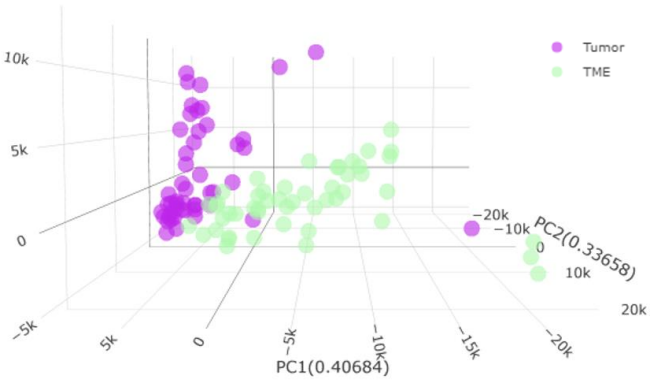


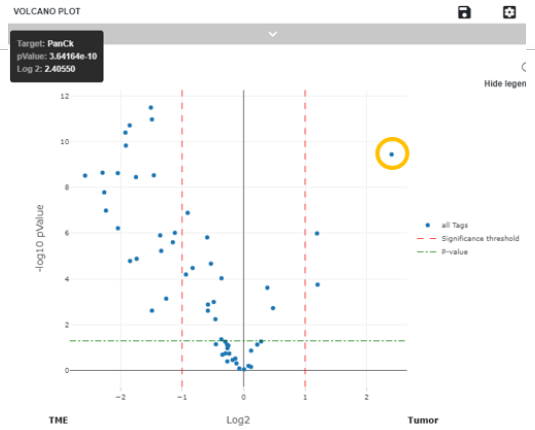
FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

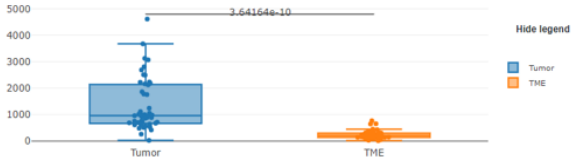
Iterative framework for Protein data analysis

QC and Normalize → Visualize and Cluster → Perform Statistics → Discover





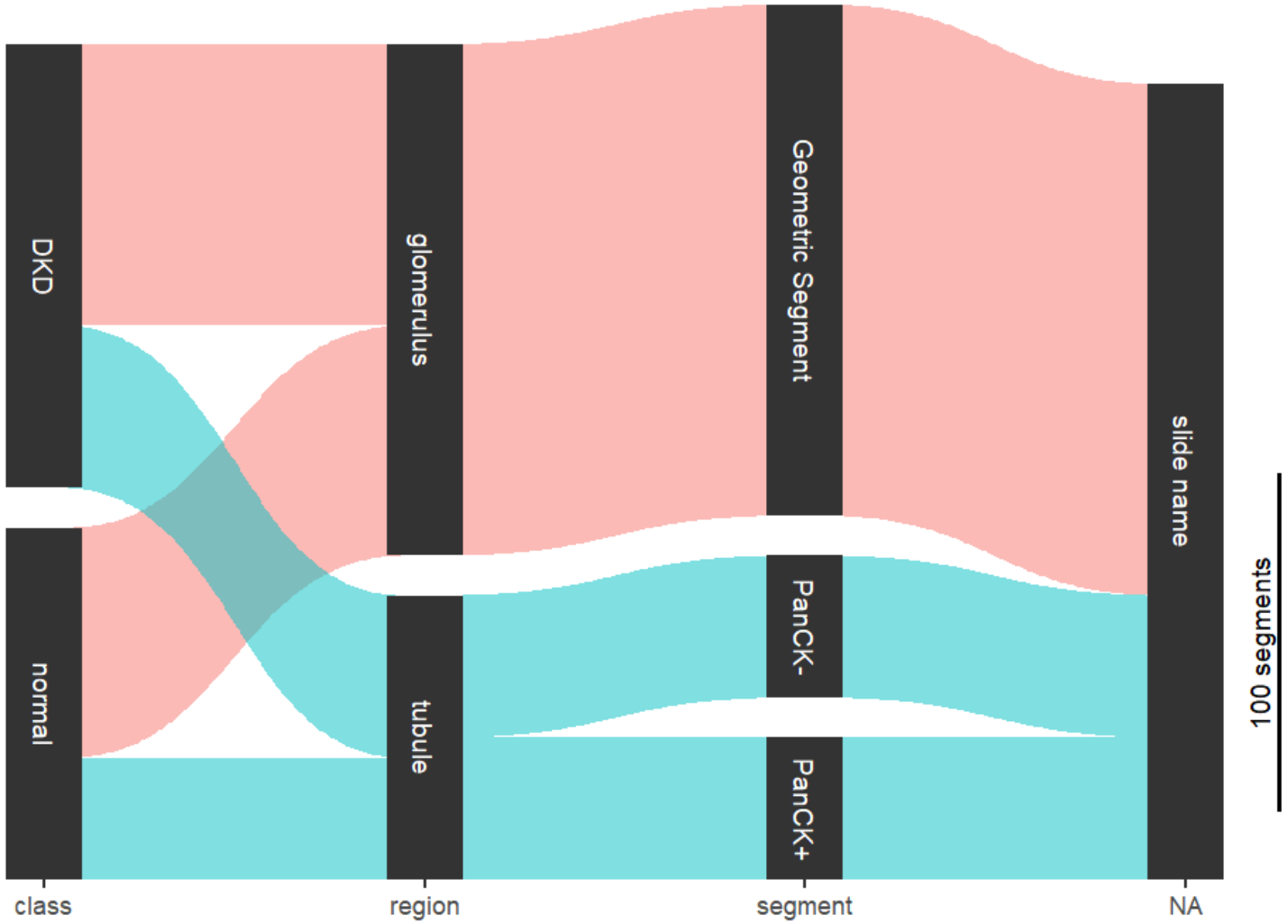




FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

First familiarize yourself with the study design

Sankey diagram*
(not in software)



Consider annotations

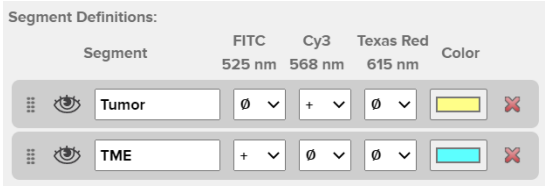
region glomerulus tubule

* Sankey diagram available in RStudio using GeoMX Workflows/GeoMX Tools

nanoString

Reviewing the Training Paradigm

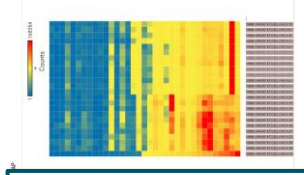
- A: CRC and Tonsil (not Cerebellum) tissue, three circular ROIs of 400 µm diameter each.



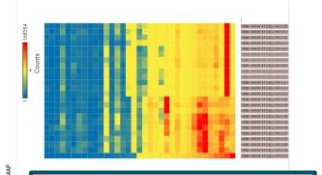
Tumor/Epithelial (PanCK+)

TME (DNA+, PanCK-)

- Data Analysis, Correlation & Clustering



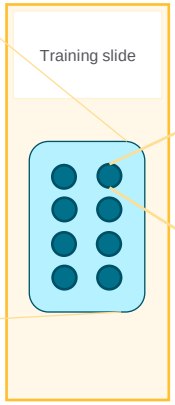
Data from Tumor segment

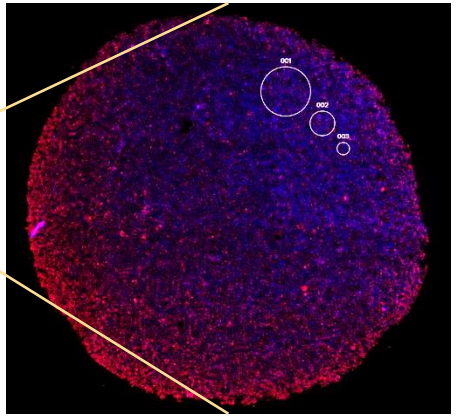


Data from TME segment

- B: Per cell line, three circular ROIs of 400, 200, and 100 µm diameter. Do not segment.

Colorectal Cancer	U251MG
Colorectal Cancer	HEK293
Cerebellum	RAJI
Tonsil	CCRF-CEM





- Known (+) and (-) markers for cell lines
- Reproducibility analysis within pellets

CRC and Tonsil tissue: 3 ROIs each with 2 segments for each ROI = 12 AOIs

Cell lines: 3 Geometric ROIs each = 12 AOIs

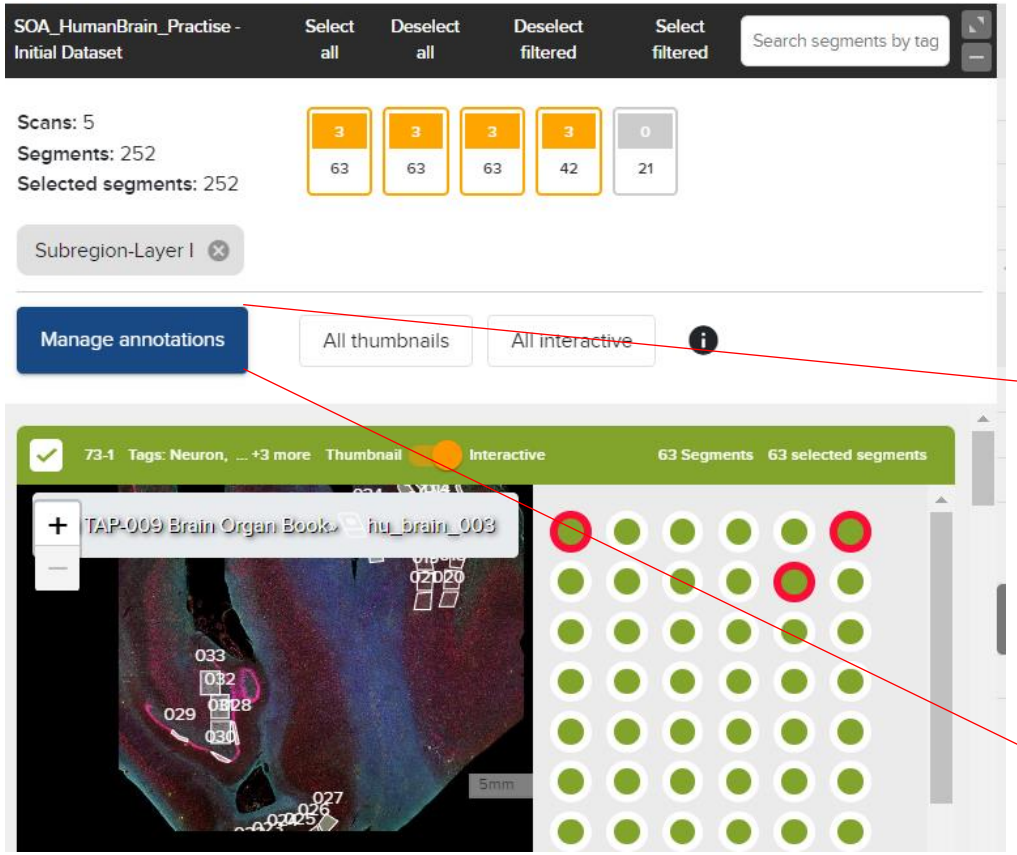
Total:

24 AOIs (2 rows in collection plate)

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Extend AOI annotation information

- Prepare sample groupings for later DE analysis

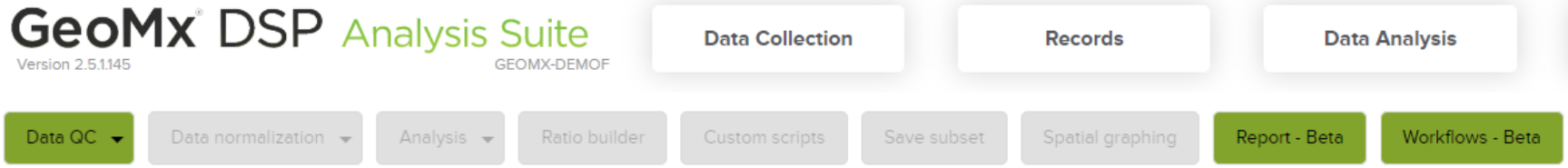


ROI (label)	Segment (Name/ Label)	Segment tags	LOT_Human_NGS_Whole_Transcriptome_Atlas_RNA	ROI_ID	Scan_ID	Subregion	Region	CellTypeInRegion
024	Full ROI		HWTA12001	024	Cortex_29	White Matter	White Matter	na
023	Full ROI		HWTA12001	023	Cortex_29	Layer V	Cerebral Cortex	na
022	Full ROI		HWTA12001	022	Cortex_29	Layer IV	Cerebral Cortex	na
021	Full ROI		HWTA12001	021	Cortex_29	Layer III	Cerebral Cortex	na
020	Full ROI		HWTA12001	020	Cortex_29	Layer II	Cerebral Cortex	na
019	Full ROI		HWTA12001	019	Cortex_29	Layer I	Cerebral Cortex	na
018	Neuropil	Neuropil	HWTA12001	018	Cortex_29	Layer V	Cerebral Cortex	Neuropil_Cerebral Cortex
018	GFAP	GFAP	HWTA12001	018	Cortex_29	Layer V	Cerebral Cortex	GFAP_Cerebral Cortex
018	Iba1	Iba1	HWTA12001	018	Cortex_29	Layer V	Cerebral Cortex	Iba1_Cerebral Cortex
018	Neuron	Neuron	HWTA12001	018	Cortex_29	Layer V	Cerebral Cortex	Neuron_Cerebral Cortex
017	Neuropil	Neuropil	HWTA12001	017	Cortex_29	Layer II/III	Cerebral Cortex	Neuropil_Cerebral Cortex
017	GFAP	GFAP	HWTA12001	017	Cortex_29	Layer II/III	Cerebral Cortex	GFAP_Cerebral Cortex
017	Iba1	Iba1	HWTA12001	017	Cortex_29	Layer II/III	Cerebral Cortex	Iba1_Cerebral Cortex
017	Neuron	Neuron	HWTA12001	017	Cortex_29	Layer II/III	Cerebral Cortex	Neuron_Cerebral Cortex
016	Full ROI		HWTA12001	016	Cortex_29	White Matter	White Matter	na
028	Neuropil	Neuropil	12002	028	77	Hilus	Hippocampus	Neuropil_Hippocampus
028	Neuron	Neuron	12002	028	77	Hilus	Hippocampus	Neuron_Hippocampus

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

Steps during data analysis



- Calibration
- QC & Filtering
  - QC of Initial Dataset
  - Filter Segments
  - QC of Filtered Initial Dataset (Notice the Correction Factors may change)
  - Background correction. Signal to noise Ratio and/or ERCC counts for all targets
  - Filter Targets
  - Filter Segments
- Normalization
  - Explore background consistency
  - Explore Signal Consistency
  - Compare IgG/HK to Area
  - Normalize your QC dataset
  - Post-normalization QC
- Statistical tests

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

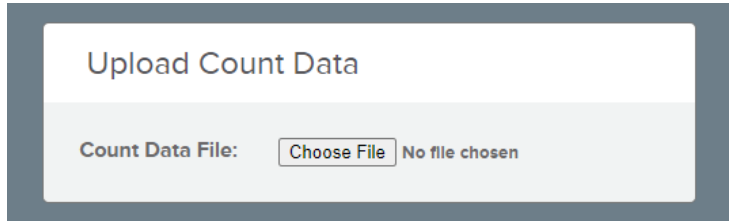
Uploading DCC files and Calibration

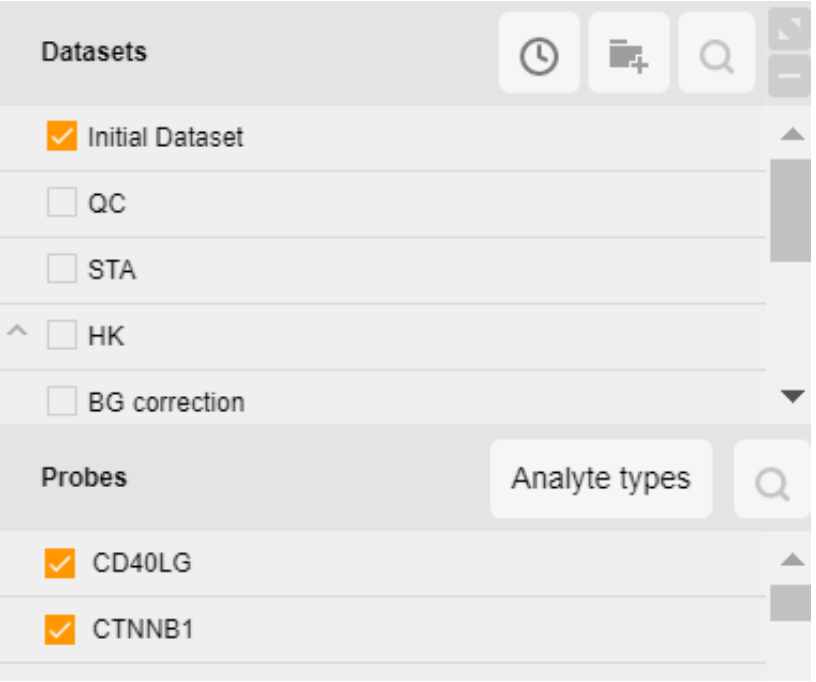
Data analysis

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Upload Counts and Calibration file to create an Initial Dataset

- Initial Data Set will be calibrated using the calibration file
- Calibration file and RCC files will be uploaded in a zipped format

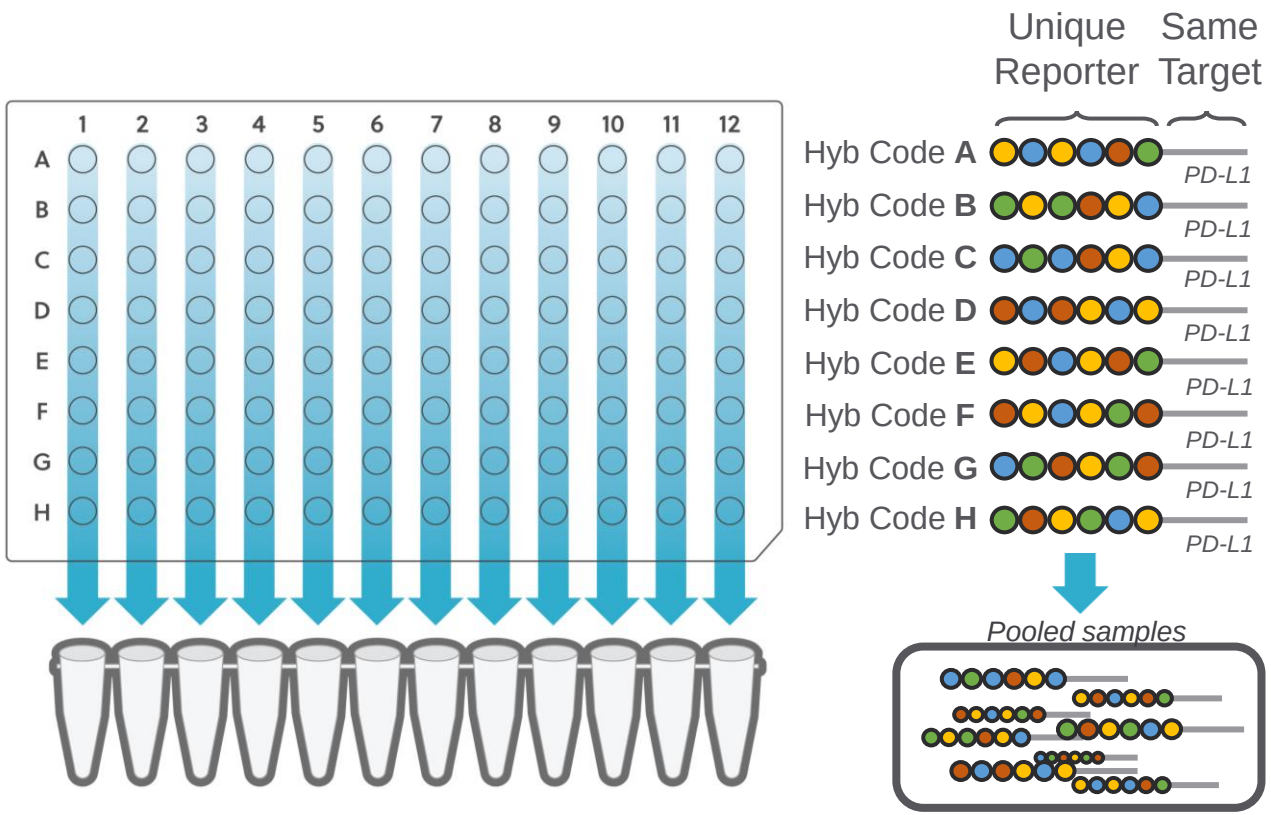




A calibration file must be uploaded to the GeoMx software
for each lot of Hyb Code
https://www.nanostring.com/products/geomx-digital-spatial-profiler/geomx-dsp/geomx-dsp-calibration-files

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Hyb Codes allow for multiplexing up to 96 AOIs per nCounter run



Pooled reporter counts are de-multiplexed by GeoMx Data Analysis software upon importing the data.

A calibration file must be uploaded to the GeoMx software for each lot of Hyb Code

NOTE: NanoString has already prepared calibration data for every available Hyb Code Lot

https://www.nanostring.com/products/geomx-digital-spatial-profiler/geomx-dsp/geomx-dsp-calibration-files

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Quality Control

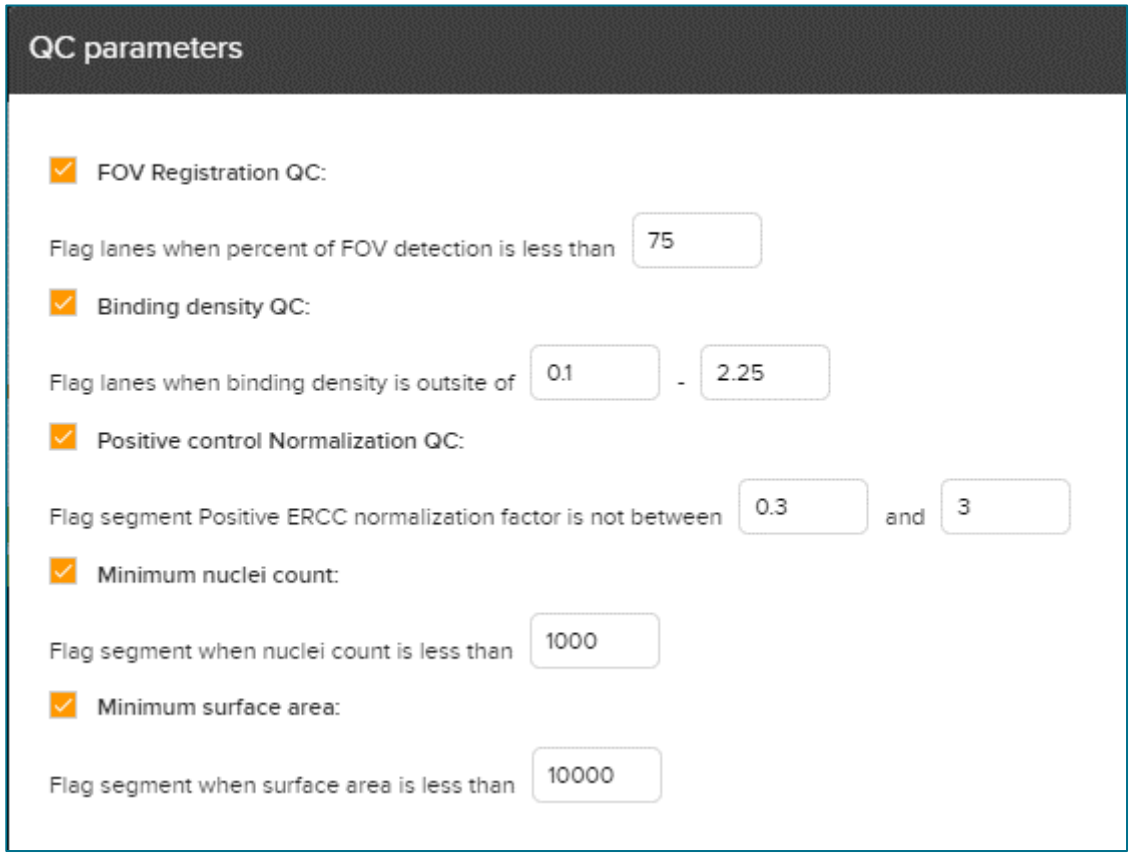
Data analysis

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Segment QC to start analysis

- Iterative QC process that can vary by study design
- Utilize visualizations to assist decision making

  - FOV registration QC
    - % of the fields of view (FOV) quantified by the instrument. Failures could be do to saturation, tilted cartridge or other issues.
  - Binding density QC
    - Density of barcodes on imaging cartridge. High BD could mean saturation, low BD could mean poor quality sample or assay failure.
  - Positive control normalization QC
    - ERCC spike in normalization factors should be similar for all segments, outliers could indicate insufficient segment size or assay failure
  - Minimum nuclei count
    - 20 nuclei is flagging minimum for protein; 200 for RNA
    - Requires nuclear stain
  - Minimum surface area
    - Both minimum nuclei count and minimum surface area can be user defined



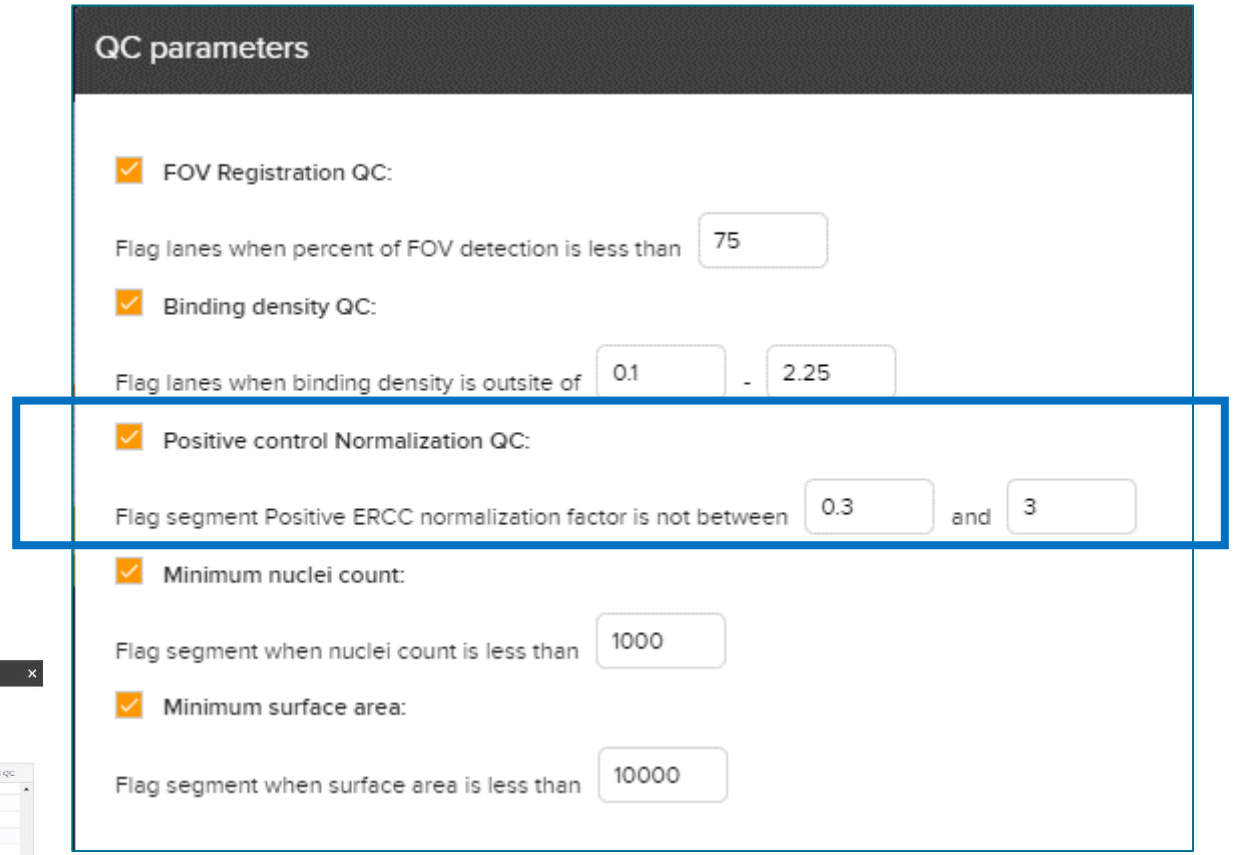
Note: Flags will show up as Tags associated with your AOIs, but will not automatically exclude them from your data set

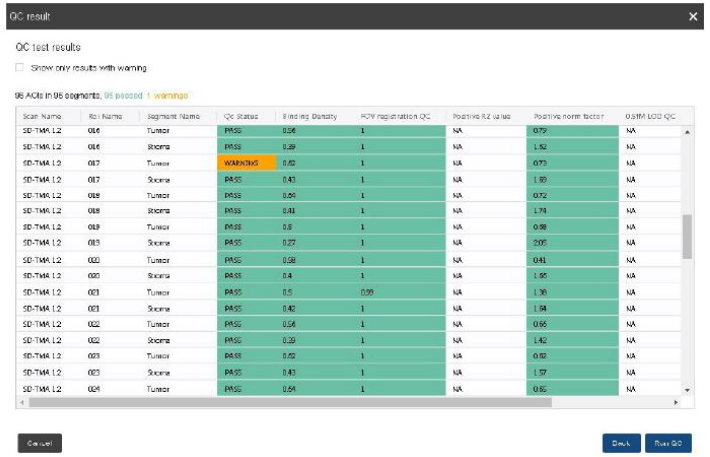
FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

Remove segments with flagged PosC QC

- Positive control normalization QC
  - ERCC spike in normalization factors should be similar for all segments, outliers could indicate insufficient segment size or assay failure
- Re do the QC to recalculate PosC Normalization Factors. Ensure they are all within acceptance range.





FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

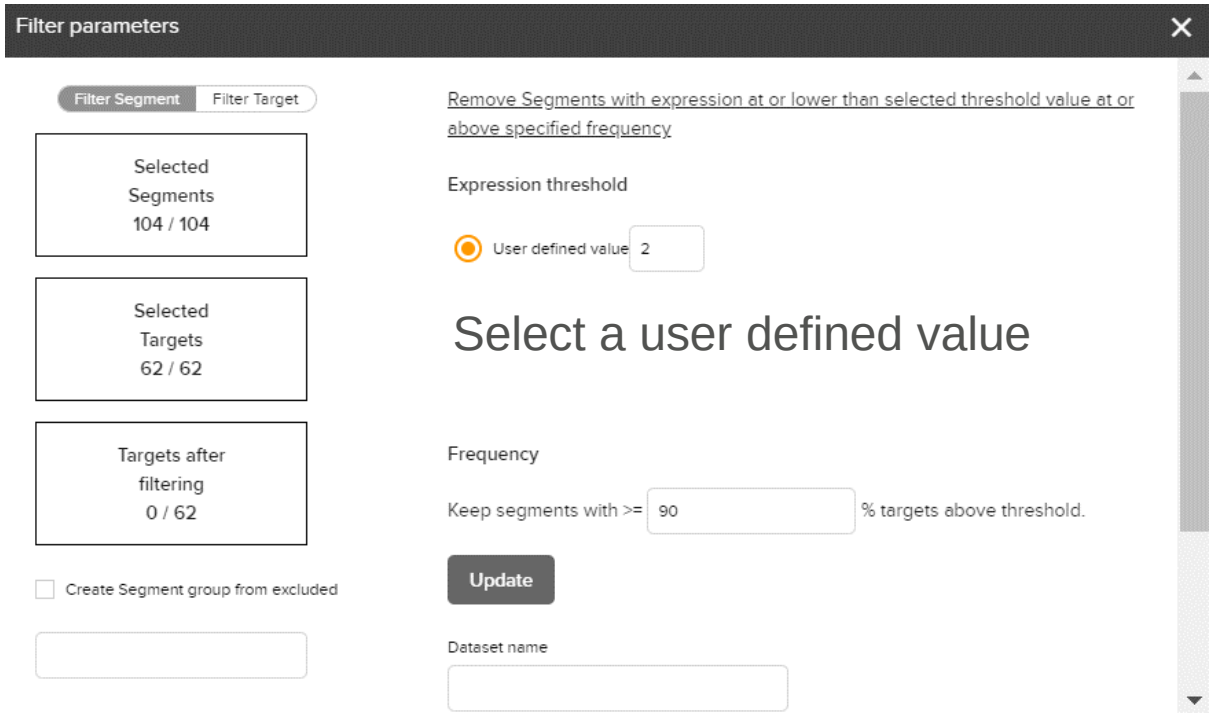
Filtering Segments and Targets

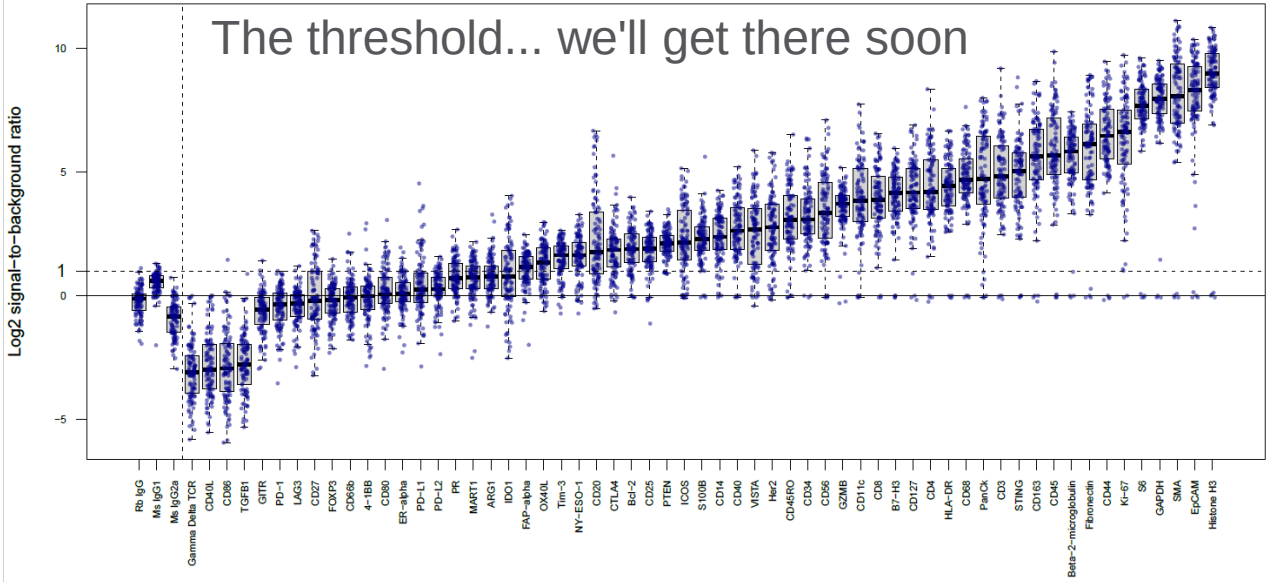
Data analysis

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Filter Segments and Targets

- Filtering is an important step to focus on the true biological signal of interest.
- Prior to downstream analysis, it is beneficial to remove segments that are low in signal relative to background.
- The Filter Parameters allow you to establish an expression threshold and a frequency at which the targets or segments are allowed to be below that threshold
  - Select an expression threshold.
  - Targets or segments with expression at or below the threshold will be deselected in the present dataset.





nanoString

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Filter Segments and Targets

- But first... you need to define the expression threshold

LOQ = Signal to Noise Ratio = Counts Target / Counts neg targets

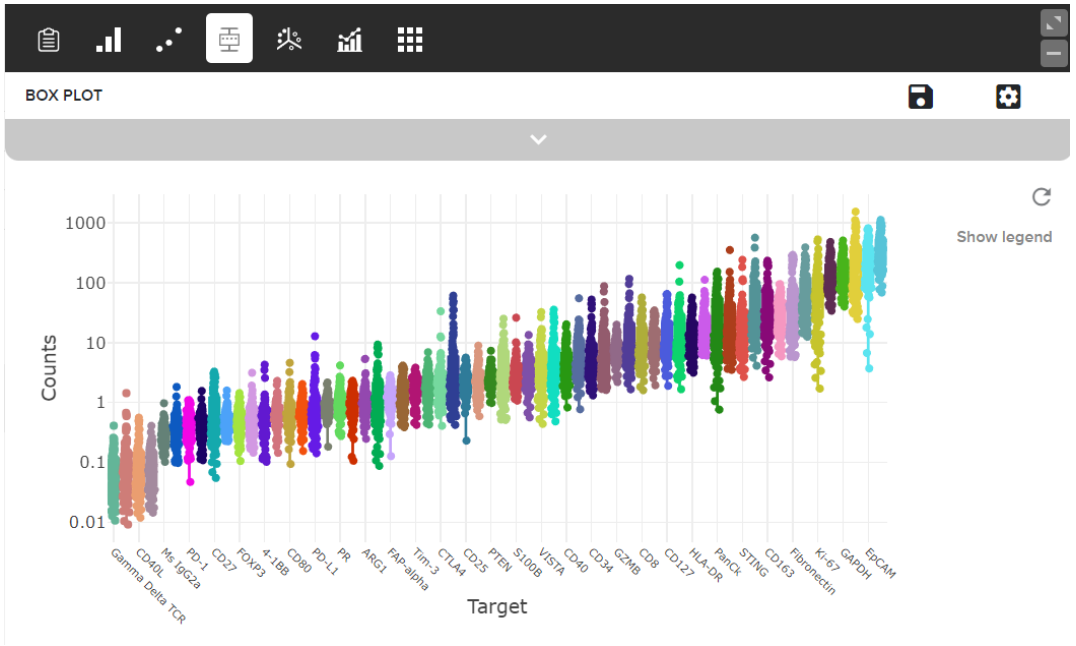
- Define LOQ **per segment**
- It is based on the negative target counts **per segment**
- We use this to **filter** low signal **segments** and **targets**
- Gene > LOQ in a segment = Gene detected in that segment

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

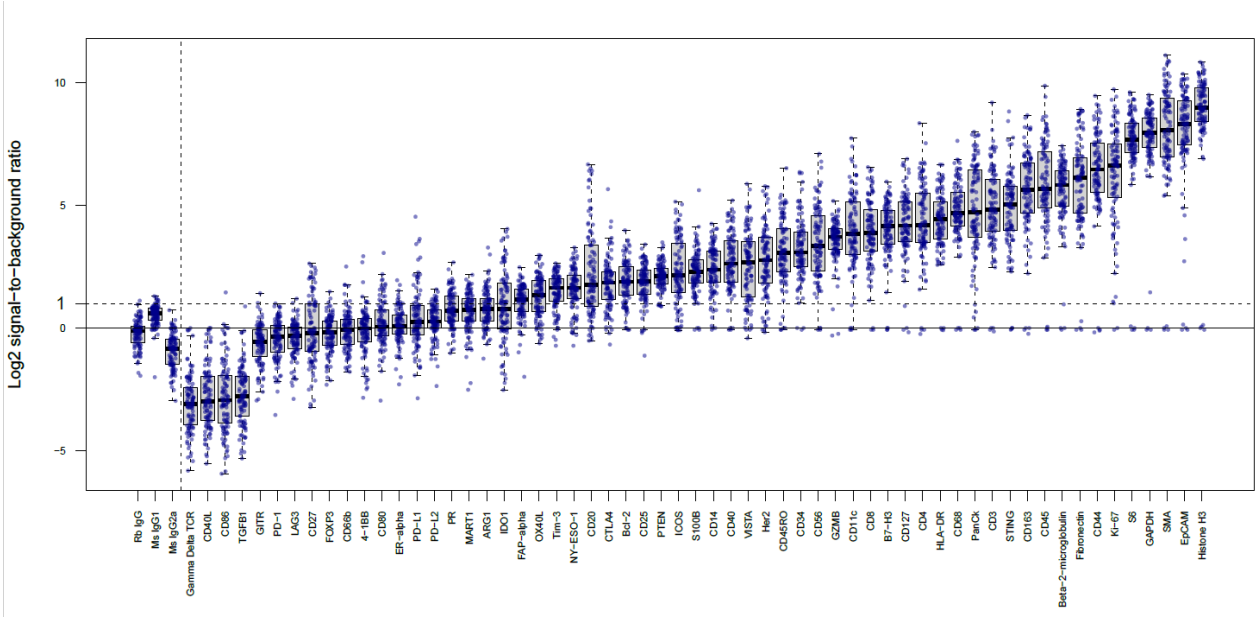
How to define the Expression Threshold

There are currently two available strategies to help you define the expression threshold

Use the built-in Background Correction Module



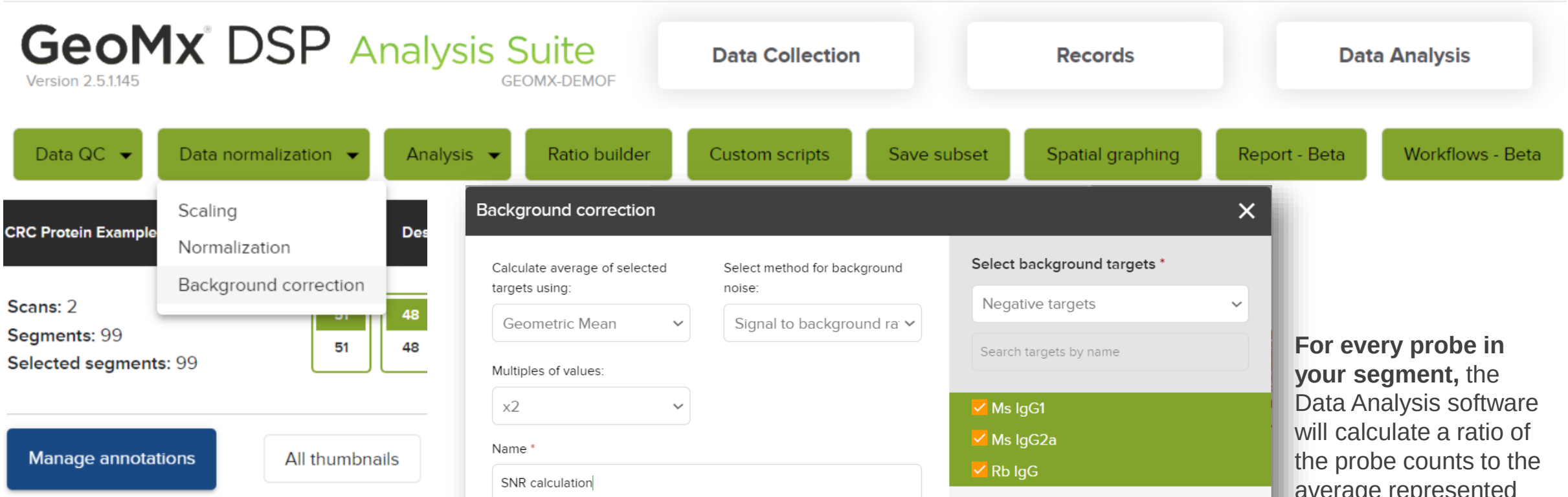
Use the Protein Normalization Custom Script

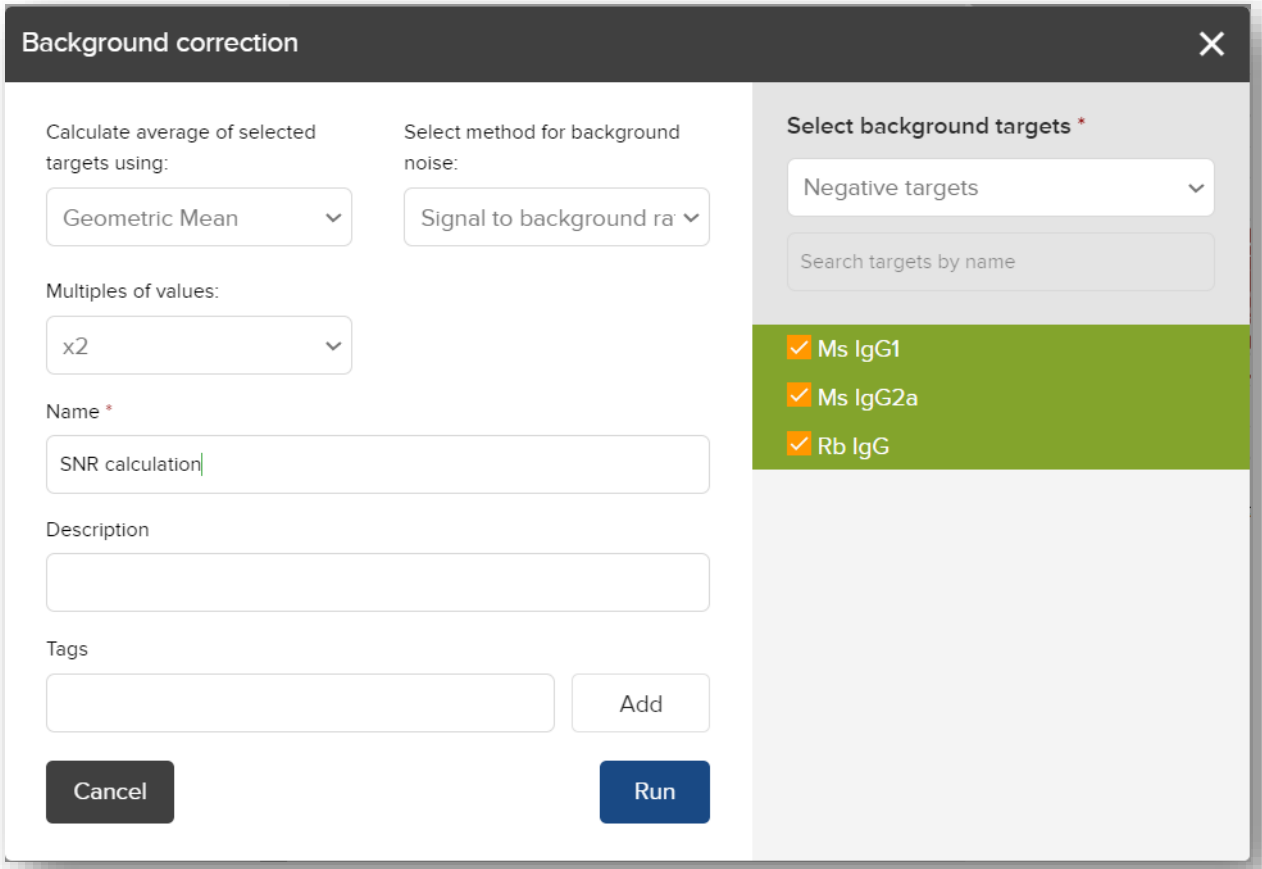


FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

Option 1: Background Correction





Signal to Noise Ratio
=
Target counts / Neg Targets counts

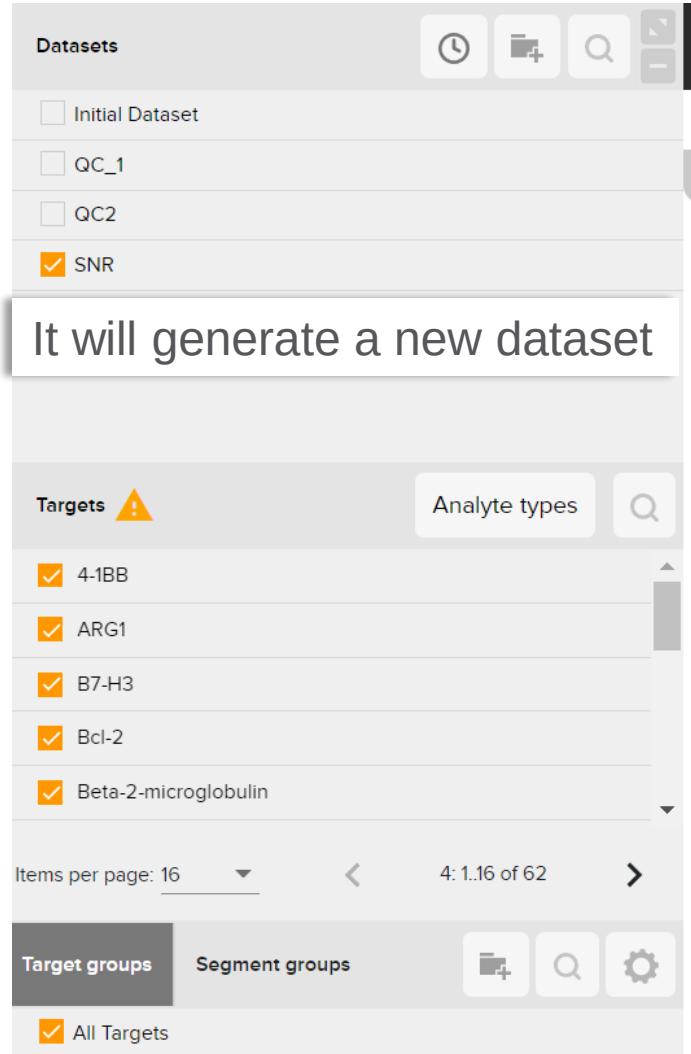
For every probe in your segment, the Data Analysis software will calculate a ratio of the probe counts to the average represented background.

- Select an averaging method
- Select the negative targets.

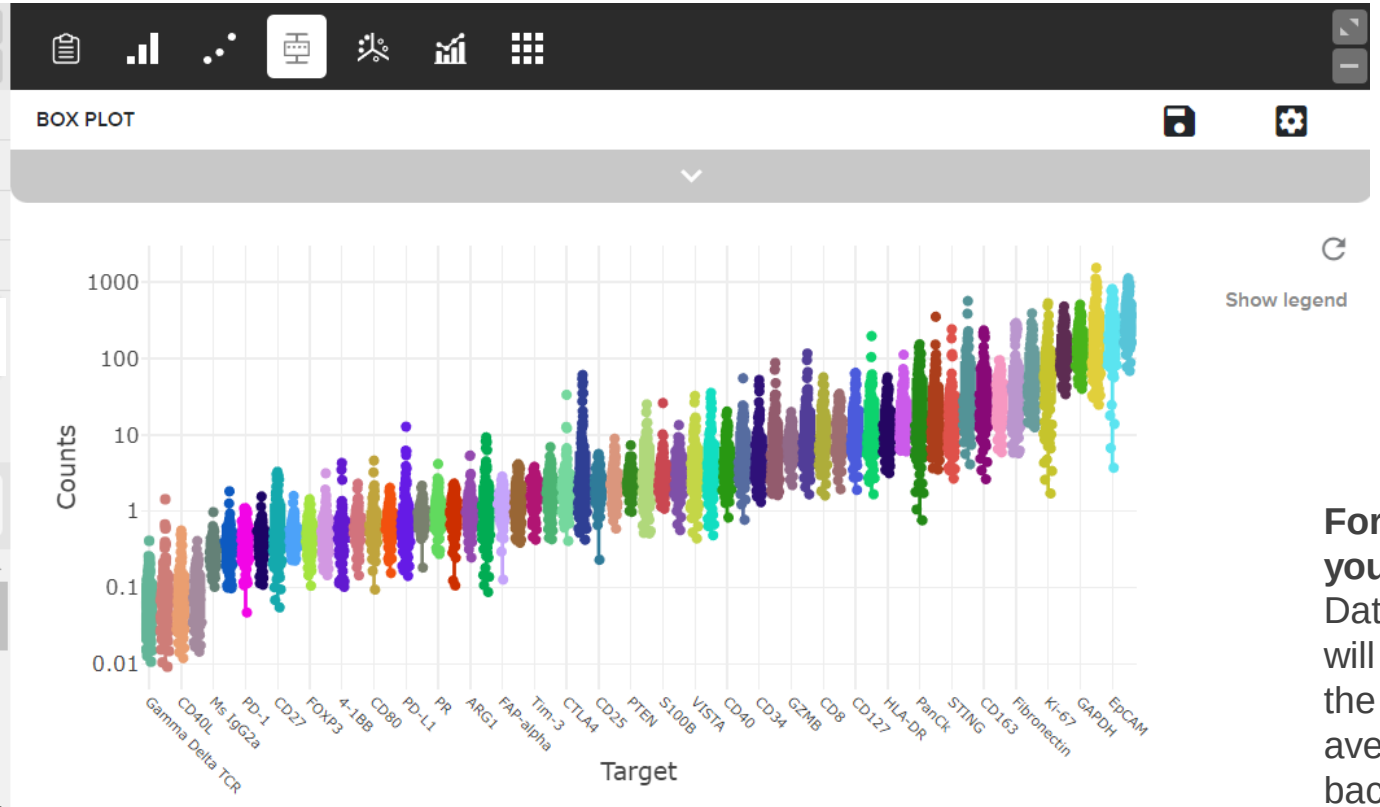
FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

Option 1: Background Correction



It will generate a new dataset



For every probe in your segment, the Data Analysis software will calculate a ratio of the probe counts to the average represented background.

Signal to Noise Ratio

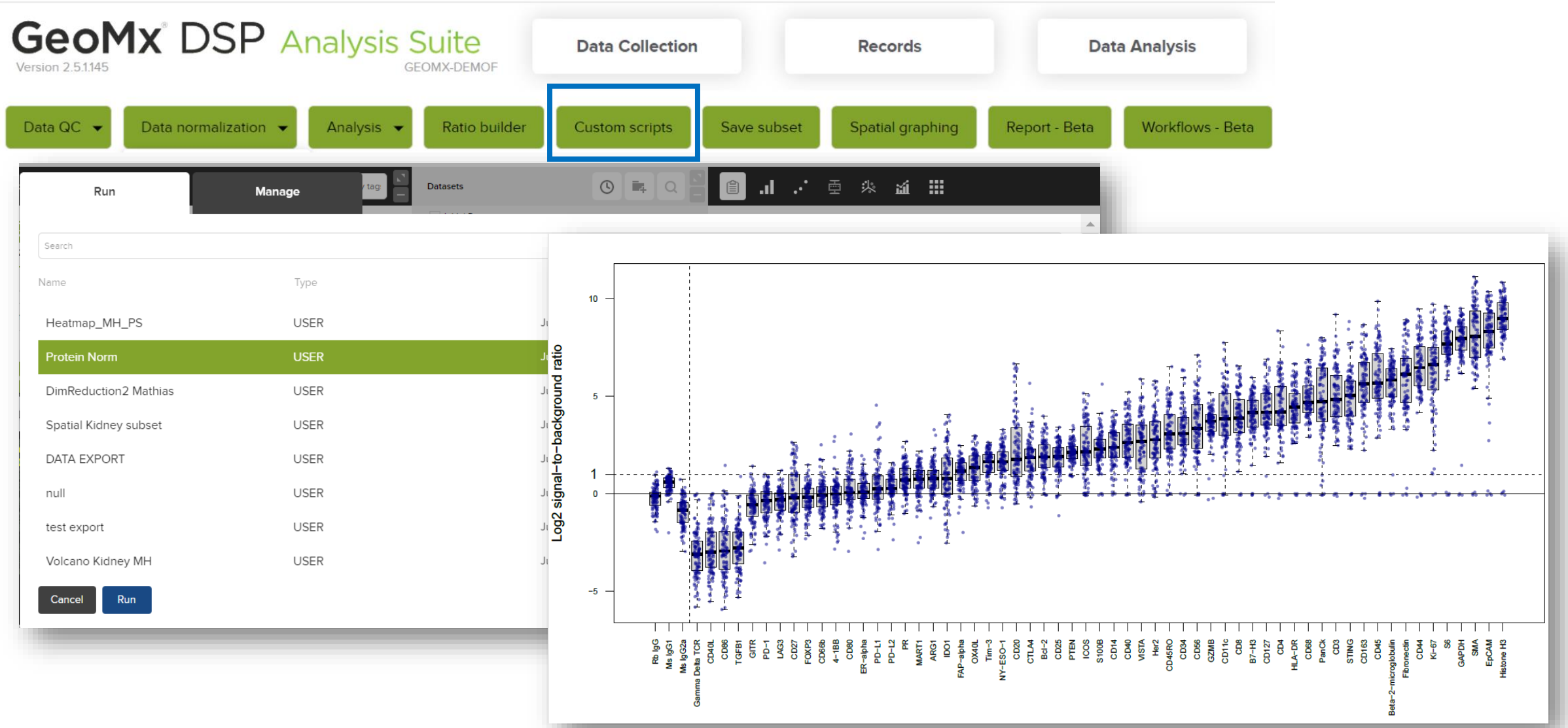
=

Target counts / Neg Targets counts

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

Option 2: Custom Script LOQ



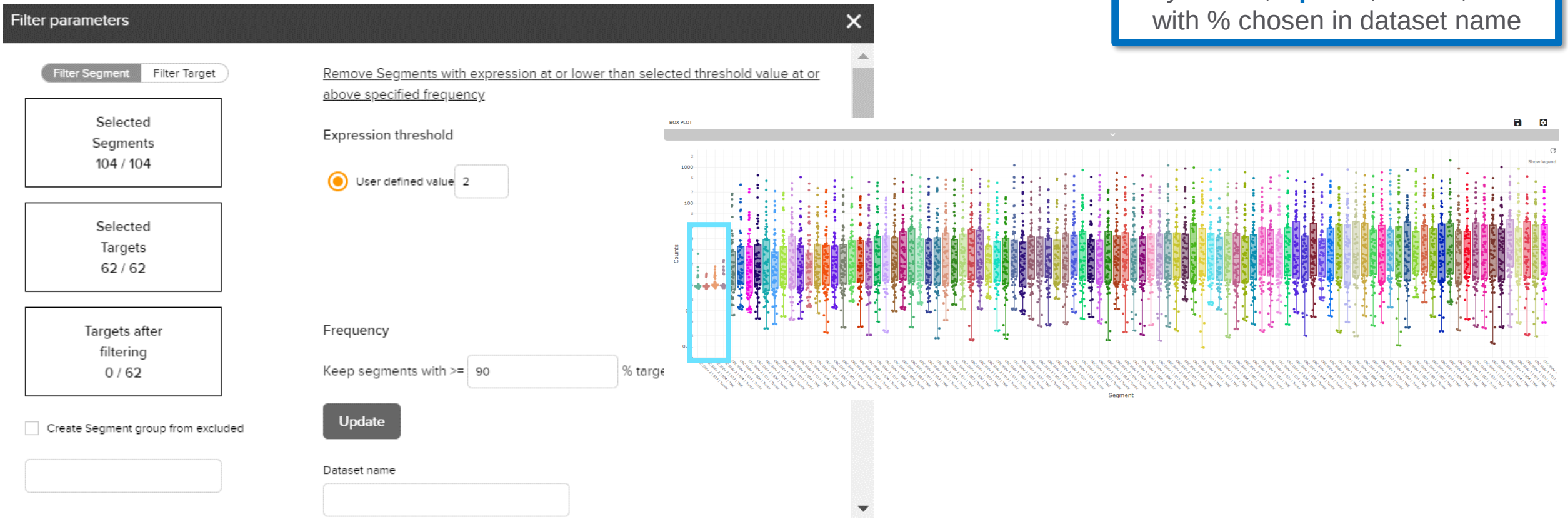
nanoString

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Filtering segments

- We look for segments for **relatively** lower target detection
- Choose a % cutoff

Filter segments
before filtering targets
Try 5-10%, **Update**, review, save
with % chosen in dataset name



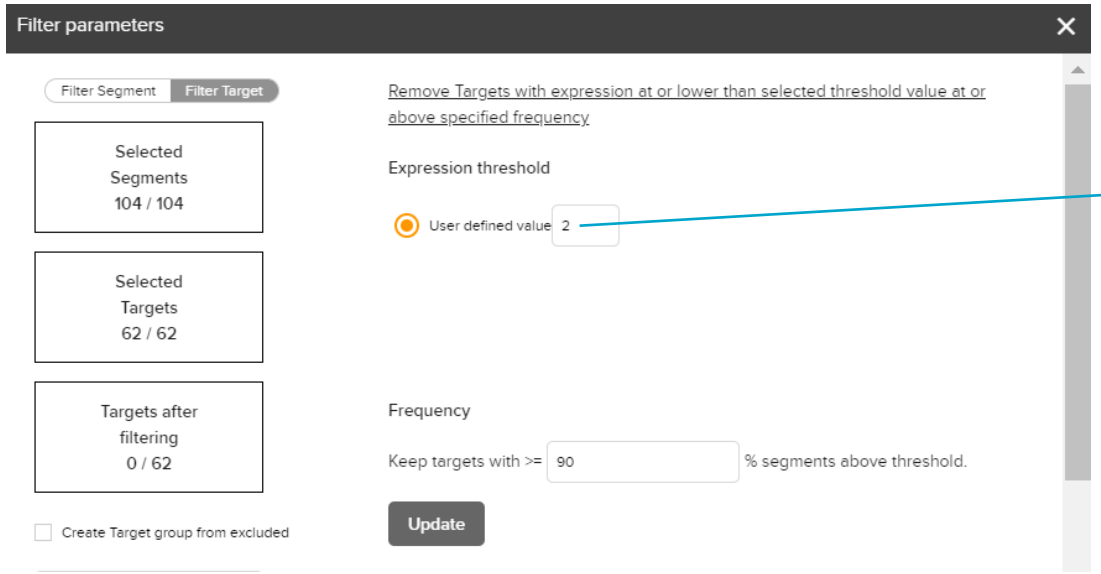
- Be careful!
  Better keep segments in first analysis rounds, to avoid losing relevant biology

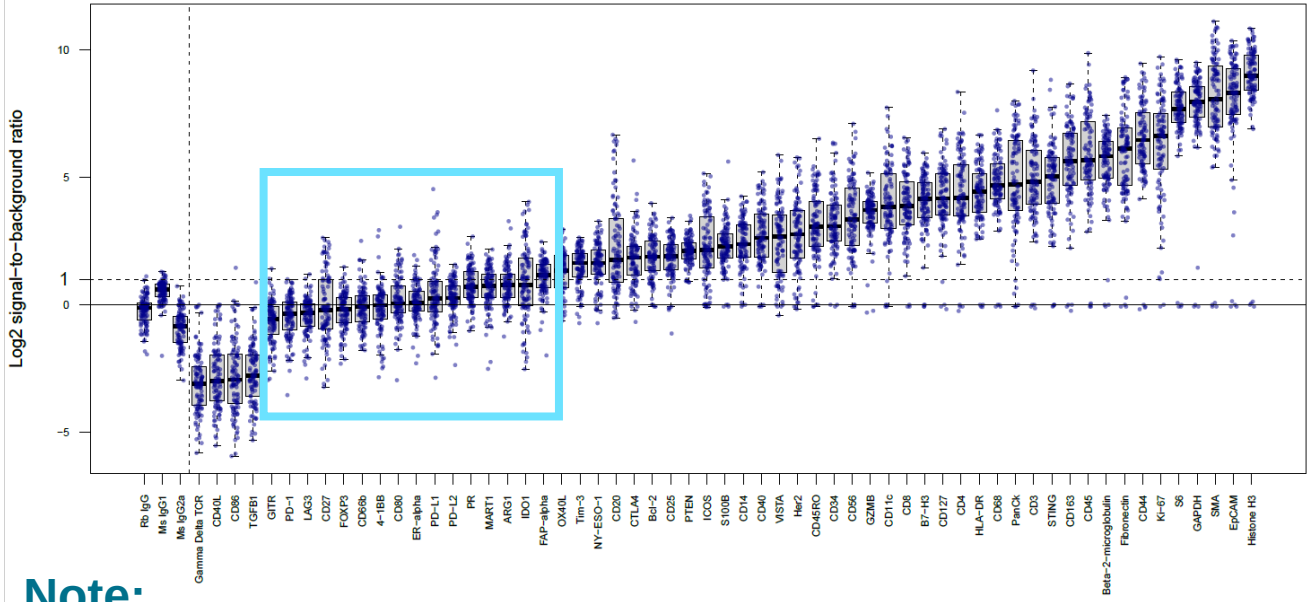
FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

Filtering targets

- We look for targets for **relatively** low detection in segments
- Choose a % cutoff





Filtering targets after filtering segments.
Use the Segment Filtered dataset.
Try 5-10%, Update, review, save with % chosen in dataset name

Note:

- Be careful!
- Take group sizes into consideration
- Make sure that you don't lose targets that are expressed only in a smaller number of AOIs
- Consider to start analysis with a subselection of AOIs

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Normalization

Data analysis

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Preamble to normalization

DSP counts respond to 3 things:

- Biology
  - Impact of AOI type, patient/tissue characteristics (e.g. HER2+ vs. HER2-)
  - Biological variability
- Technical variables
  - AOIs' signal levels, AOIs' background levels
- Noise

The point of normalization is to **strip the signal of technical variables from your data**

- Sometimes we can only normalize away a subset of the technical variables, and have to tolerate artifacts from other variables.
- Mitigation: use post-analysis QC and sanity checking to ensure your conclusions aren't driven by technical variables not flattened by normalization.

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Normalization: Identifying the Drivers of technical variability

AOI area,
AOI cell density,
nCounter efficiency

Specific count

Counts = True expression x Signal strength + Background

AOI binding (on target)

AOI binding (off target)

Area normalization:

- AOIs have uniform per-area "binding", both for on-target and off-target binding
- AOIs have consistent cell density

Nuclei normalization:

- AOIs have uniform per-cell "binding", both for on-target and off-target binding
- Nucleus counting algorithms don't always work as intended

IgG normalization:

- IgG's respond to nearly all factors going into signal strength
- AOIs' off-target binding doesn't diverge from their on-target binding

  (i.e. background levels and signal levels are redundant)

Housekeeper normalization:

- Housekeepers give a comprehensive readout of all factors contributing to AOIs' signal strength
- But in many experiments housekeepers are inconsistent across AOIs

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

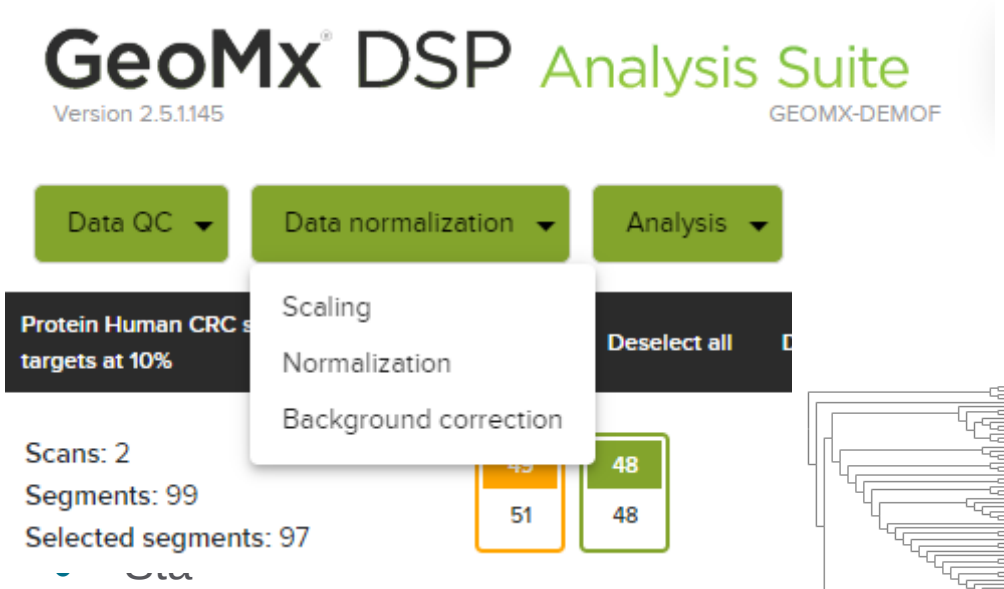
Normalize the data

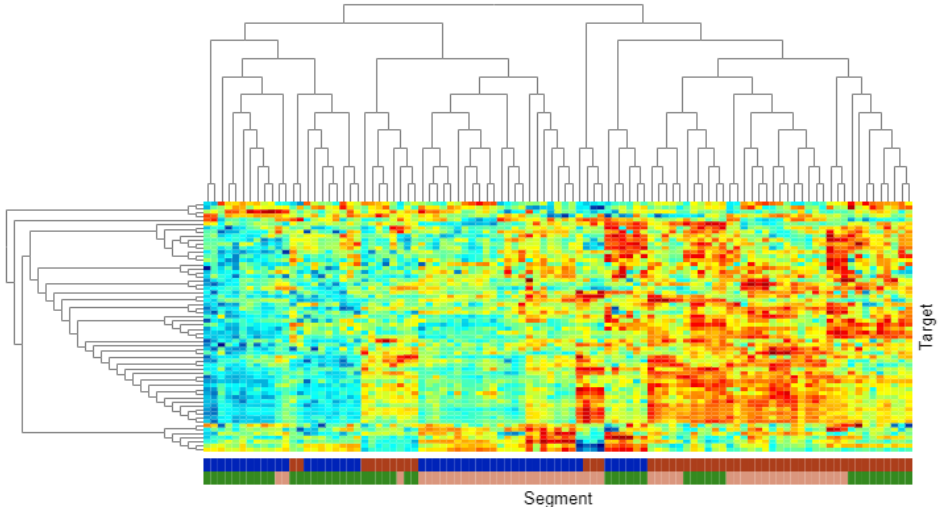
Evaluating Normalization Methods:

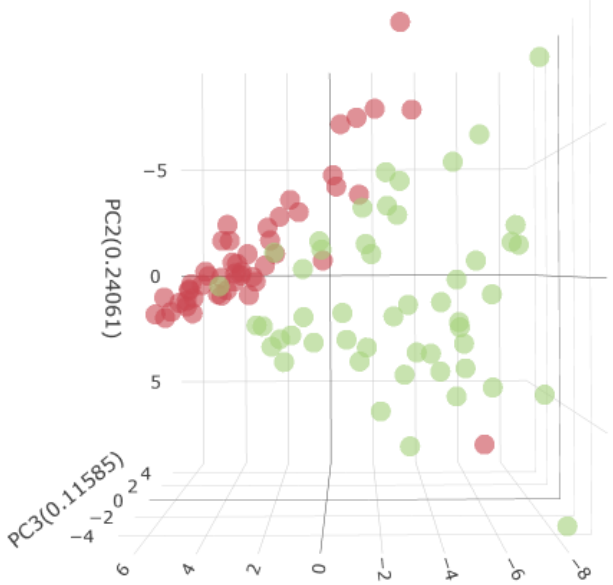
- Scaling to area OR nuclei count
- Content Normalization
- Background Normalization

- Two plots can help to compare these:
  - heatmaps
  - principal component plots

Check for technical bias / slide bias







Note: Do the normalization from the QC'ed counts, not from the SNR values !

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

How to chose the right Normalization strategy for your dataset

Before choosing a normalization method, we must QC the potential normalization factors.

The theory is simple: if multiple probes all accurately measure signal strength, they should be highly correlated with each other.

Evaluate Normalization for Protein

This script was designed for data from the GeoMx protein assay. It will also work with data from the GeoMx RNA assay.

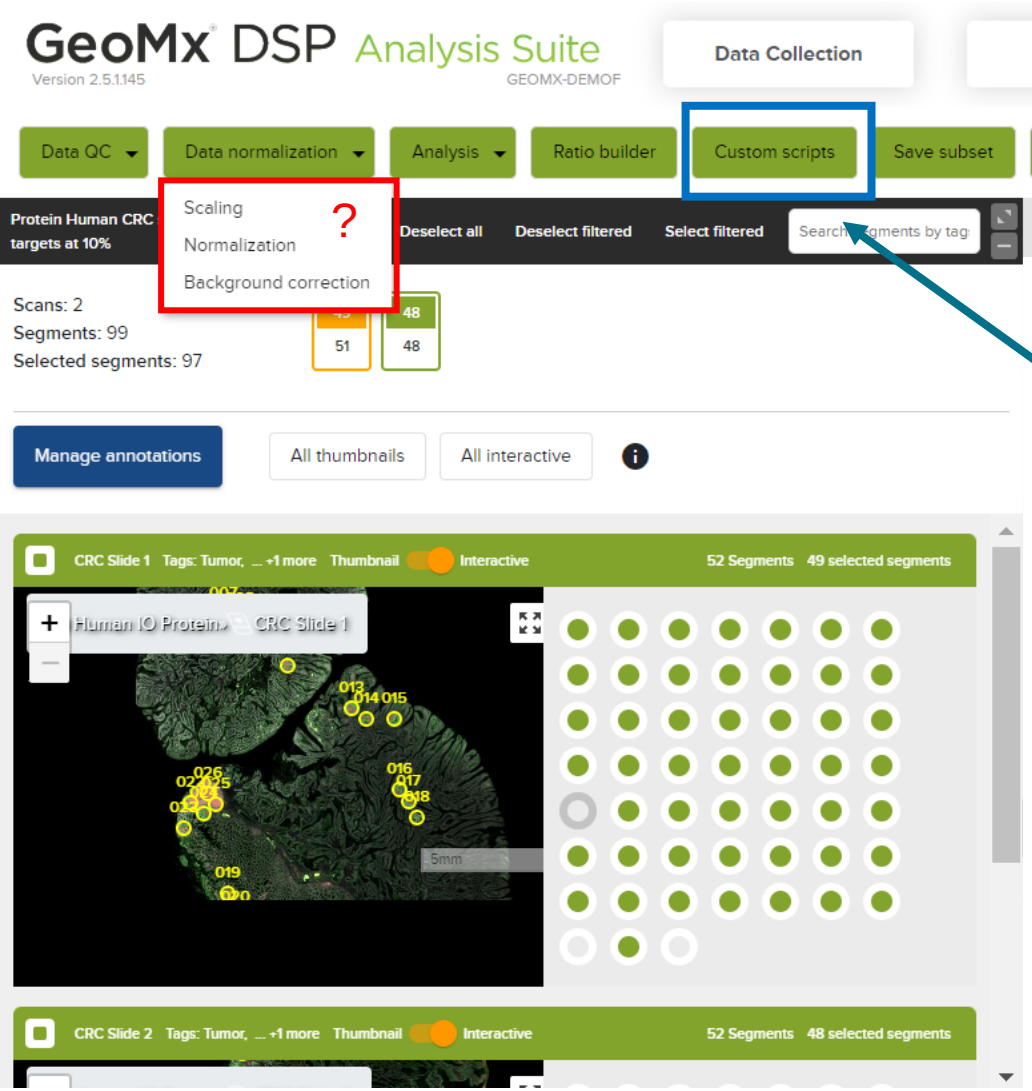
Using the Evaluate Normalization for Protein script

- Automatically identifies relevant variables from your segment annotations for plotting
- Arbitrarily assigns colors to the identified segment annotations
- Computes multiple normalization factors (negative control IgGs, housekeepers, area, nuclei) for comparison

- Produces a plot used to QC the negative control IgG's
- Produces a plot used to QC the housekeeping proteins
- Produces a plot used to QC all normalization factors computed in step 3

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

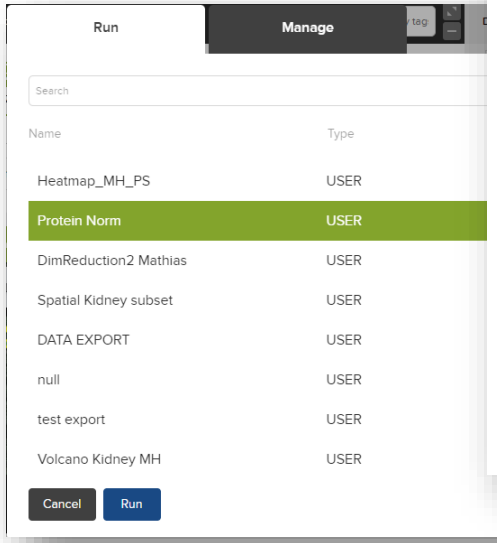
How to chose the right Normalization strategy for your dataset



Import

First, ensure that the **correct dataset** is selected; a **custom script** is run on the dataset, slides, and targets that are selected.

- Click the **Custom scripts** tab
- Click **Manage**
- Under Manage, click **Add " + "**
- Load the **evaluate_normalization__options.R** script
- Click **Save** (in blue)
- Add a **name**
- Click **Add**



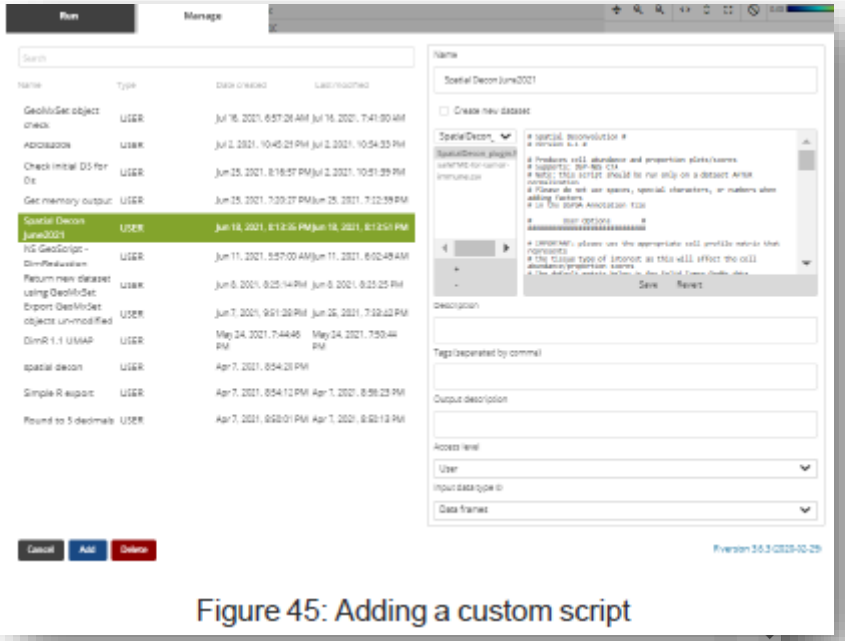


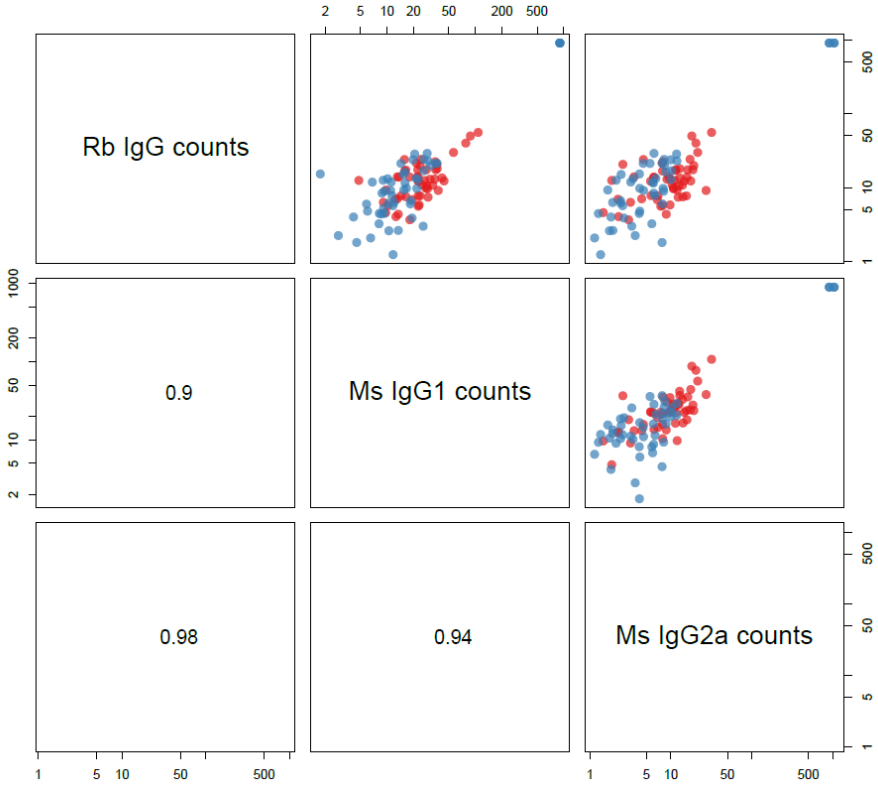
Figure 45: Adding a custom script

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Explore IgG consistency

- Do the negative control IgG's have high enough counts to be used in normalization?
  - These IgGs measure background, which can be a normalization method to compare slides and/or cell populations
  - ROIs with low (< 30) IgG geomeans will return unstable data if IgGs are used as the denominator in normalization because the counts are too low and less consistent.
- Are the IgG's internally consistent?
  - Assumption is that the IgG's independently measure ROIs' underlying background tendency.

how high are counts for neg probes?



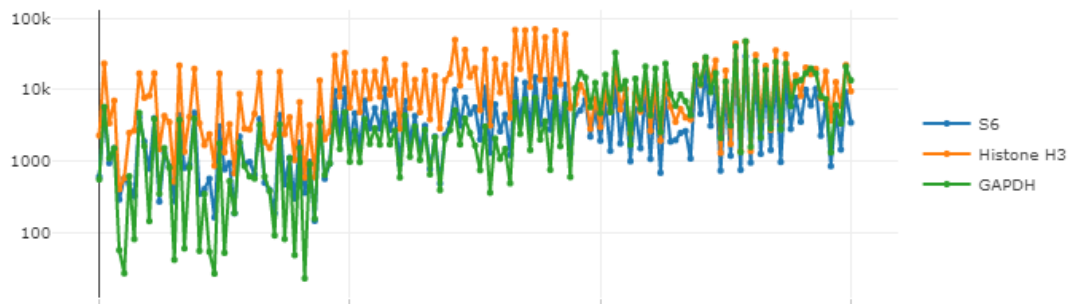
Numbers in the lower-left panels show the SD of the log2-ratios.

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

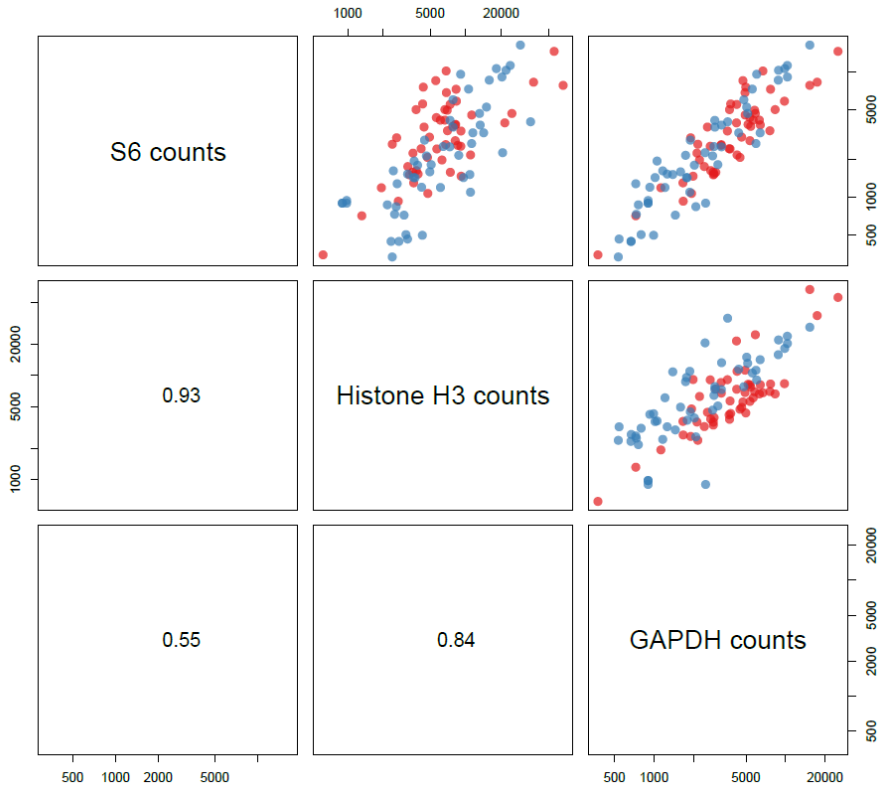
Explore housekeeper consistency

- Housekeeper normalization:
  - Assumes the housekeepers measure primarily signal strength
  - Under this assumption, housekeepers should be highly correlated (with consistent ratios between them).

Good correlations observed for the housekeepers



Can we exclude Histone H3? Tick away!

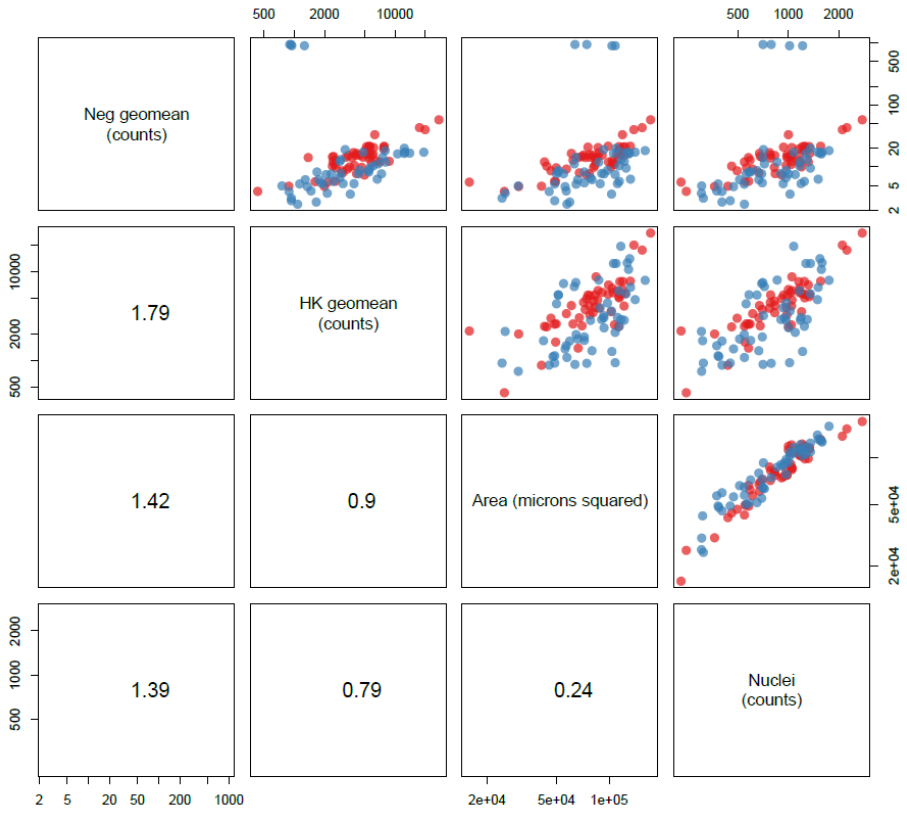


Numbers in the lower-left panels show the SD of the log2-ratios.

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Check for consistency between IgG's, Housekeepers, Area

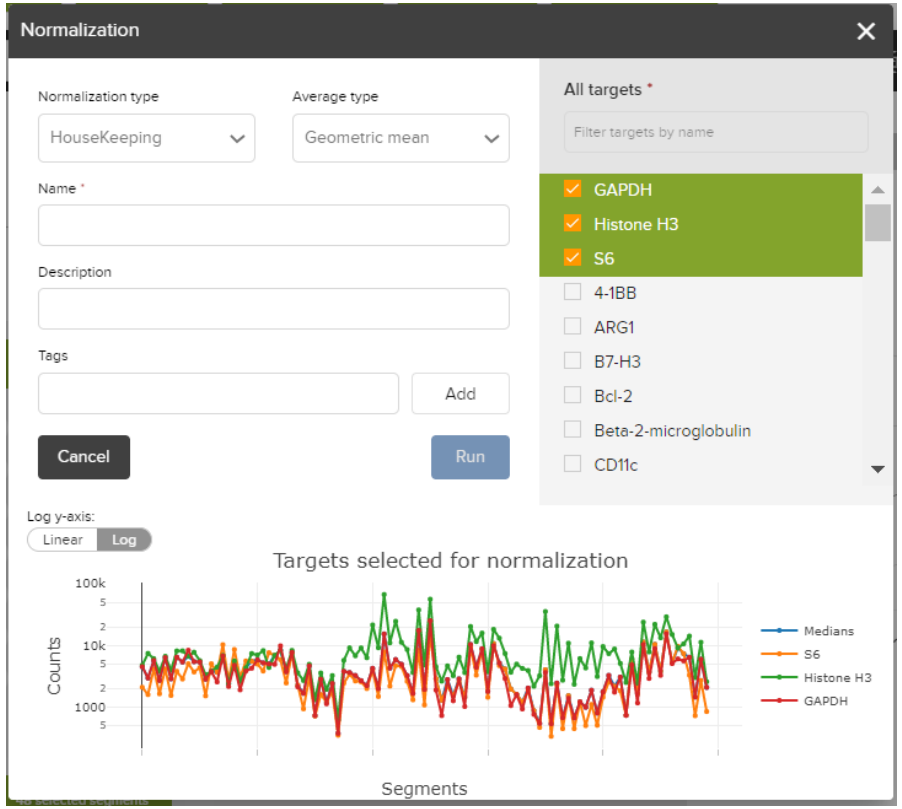
- How do we choose a normalization method?
  - Check how the different variables (IgG, HK, Area) behave with respect to each other before normalization.



Numbers in the lower-left panels show the SD of the log2-ratios.

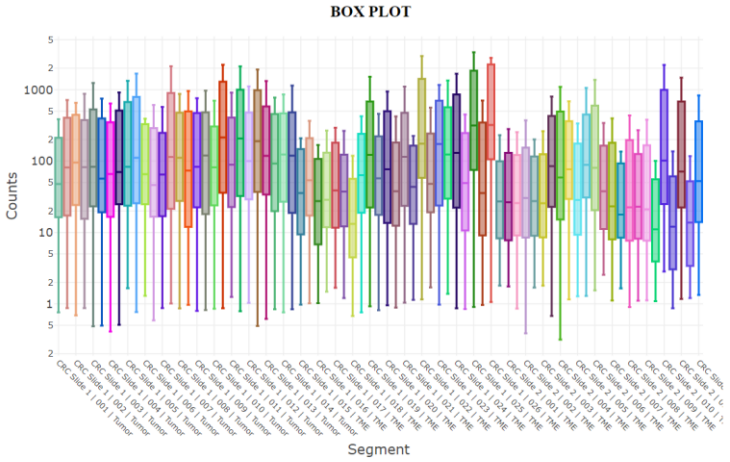
FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Normalization: post-analysis QC and sanity check

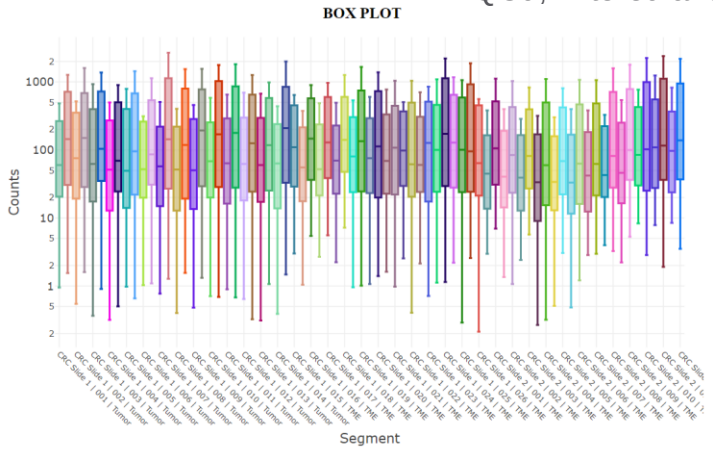


What it does

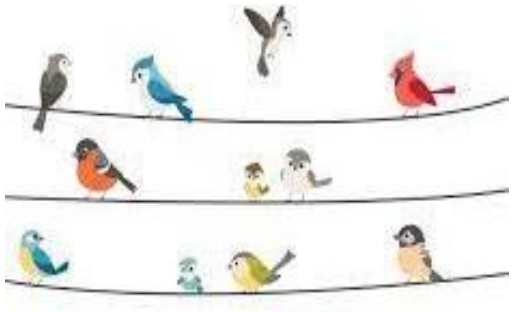
QCd and Filtered Dataset



QCd, Filtered and Normalized Dataset



How it feels like

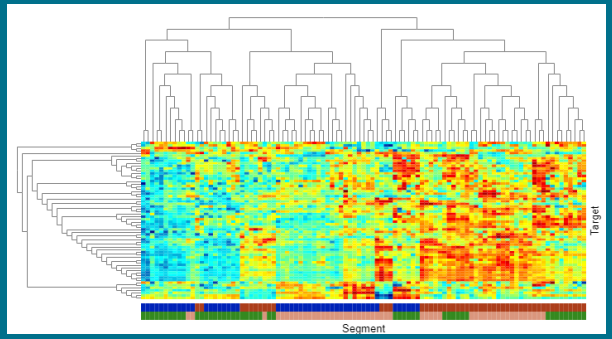


FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

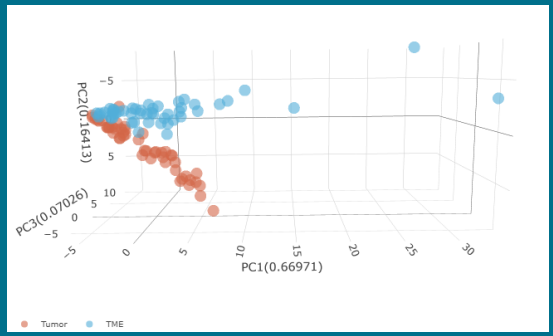
nanoString

Data Exploration

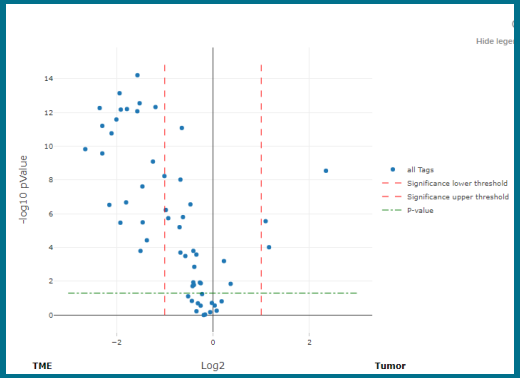
Cluster Analysis



PCA



Statistical Analysis



FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

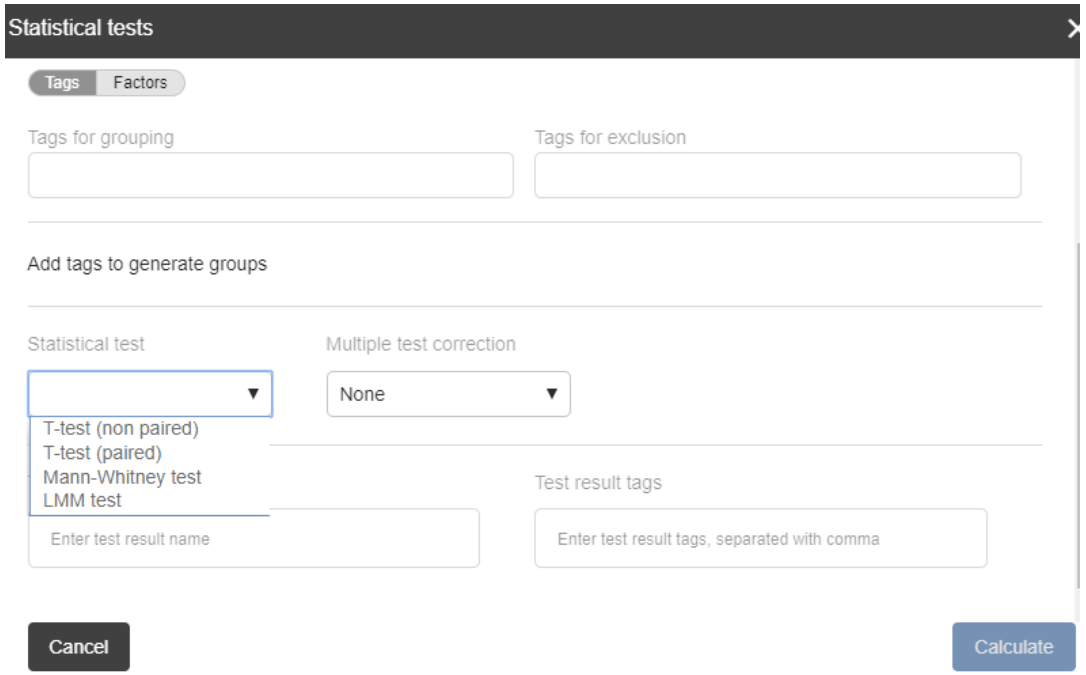
Statistical Analysis

Data analysis

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Statistical Tests for Differential Expression

- When up to two groups are established by selecting tag combinations, the following tests may be run:
  - Paired T-test
    - useful for comparing two groups with a natural paired structure, e.g. pre vs. post-treatment, or Tumor vs. Immune ROIs from the same collection of samples.
  - Non paired T-test
    - useful for comparing two groups of independent samples
  - Mann-Whitney test
    - useful for data that is extremely skewed or heavy-tailed
  - LMM test
    - useful for data with repeated measurements from each sampling unit, e.g., multiple ROIs from each sample

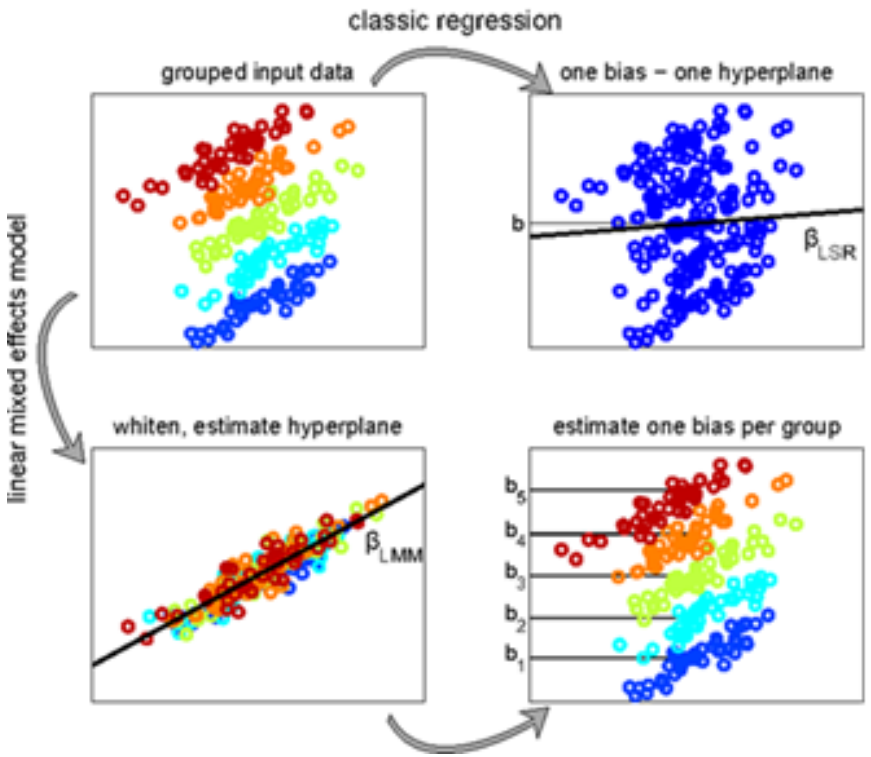


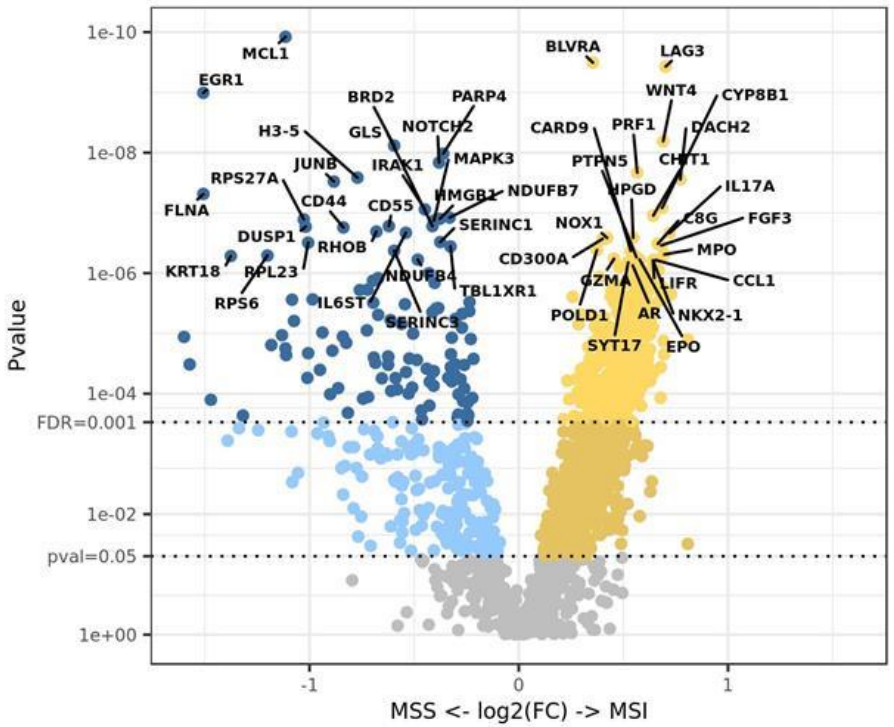
FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

Perform statistics : Differential Expression

- Two common experimental designs:
  - TMA: 1 ROI per Patient → T-test
  - Tissue section: multiple ROIs (e.g. 6-12) per Patient → **Linear Mixed Effect Model (LMM)**





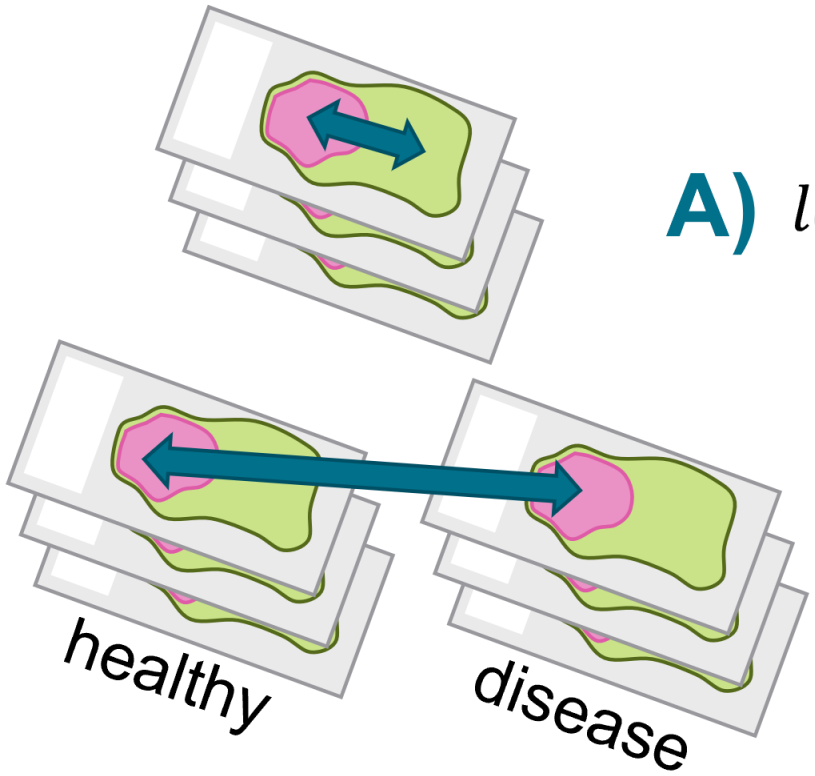
For a given comparison, we will subset the data down to the relevant segments and perform DE.

https://doi.org/10.1016/j.neuroimage.2011.03.061

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Pick one of two LMM flavors

- Tissue is your “batch” variable
- One question: test variable within the same tissue?



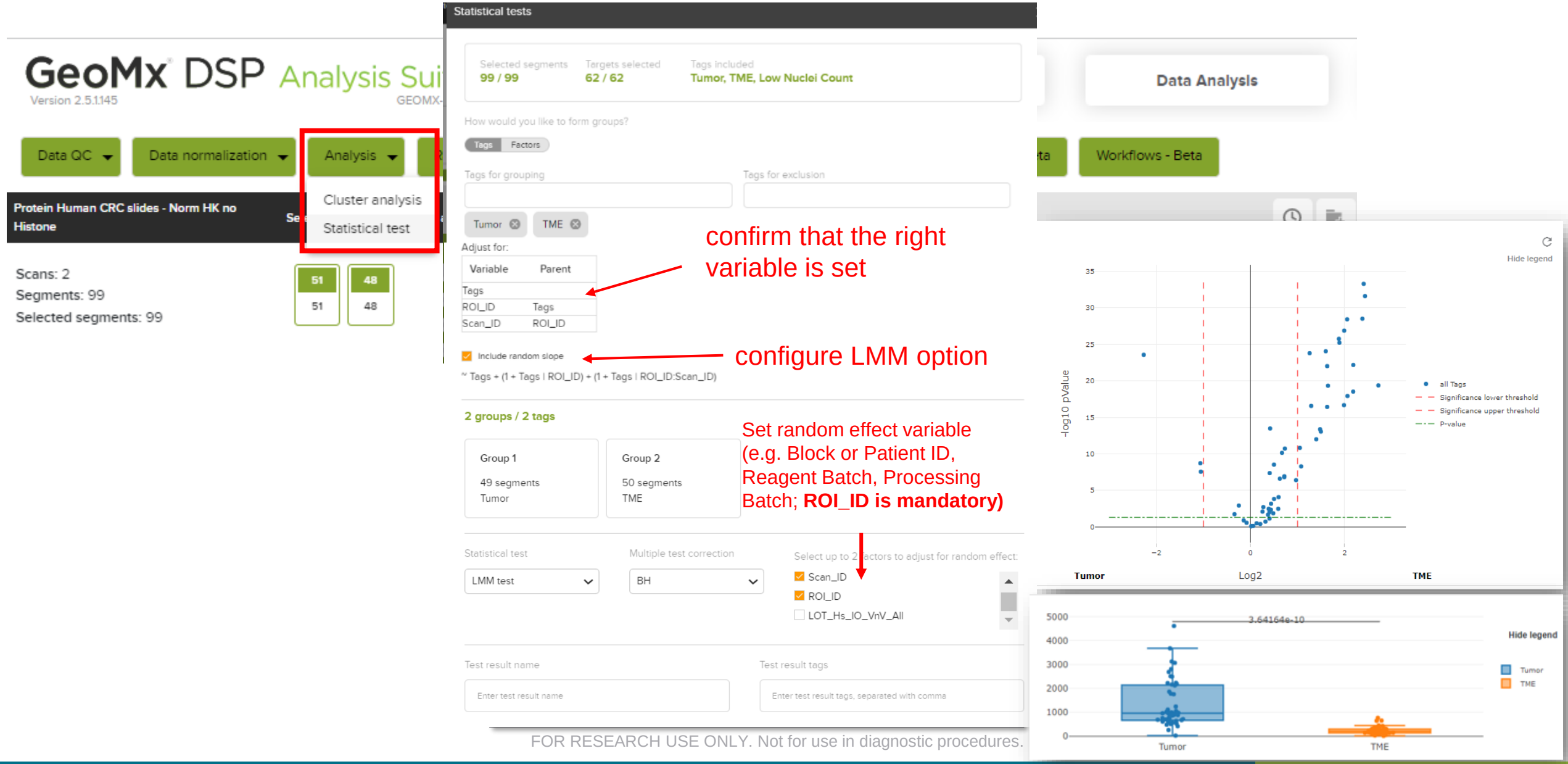
A) $log2(gene) \sim \overbrace{Region}^{\text{Test variable}} + (\overbrace{1 + Region}^{\text{Random Slope}} \mid \overbrace{Slide}^{\text{Random Intercept}})$

B) $log2(gene) \sim \overbrace{Disease}^{\text{Test variable}} + (1 \mid \overbrace{Slide}^{\text{Random Intercept}})$

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

Running the Statistical Analysis of choice



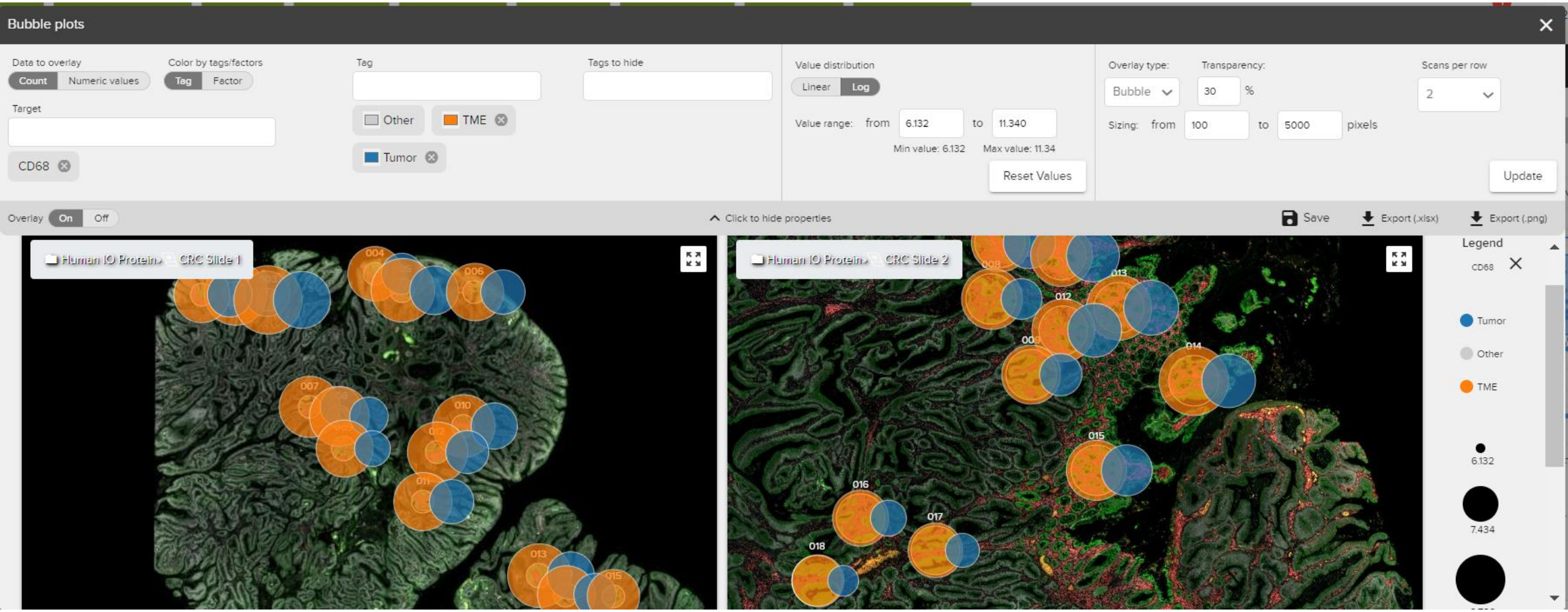
FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Other Features

Data analysis

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Spatial Graphing



FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

Workflows to streamline your analysis

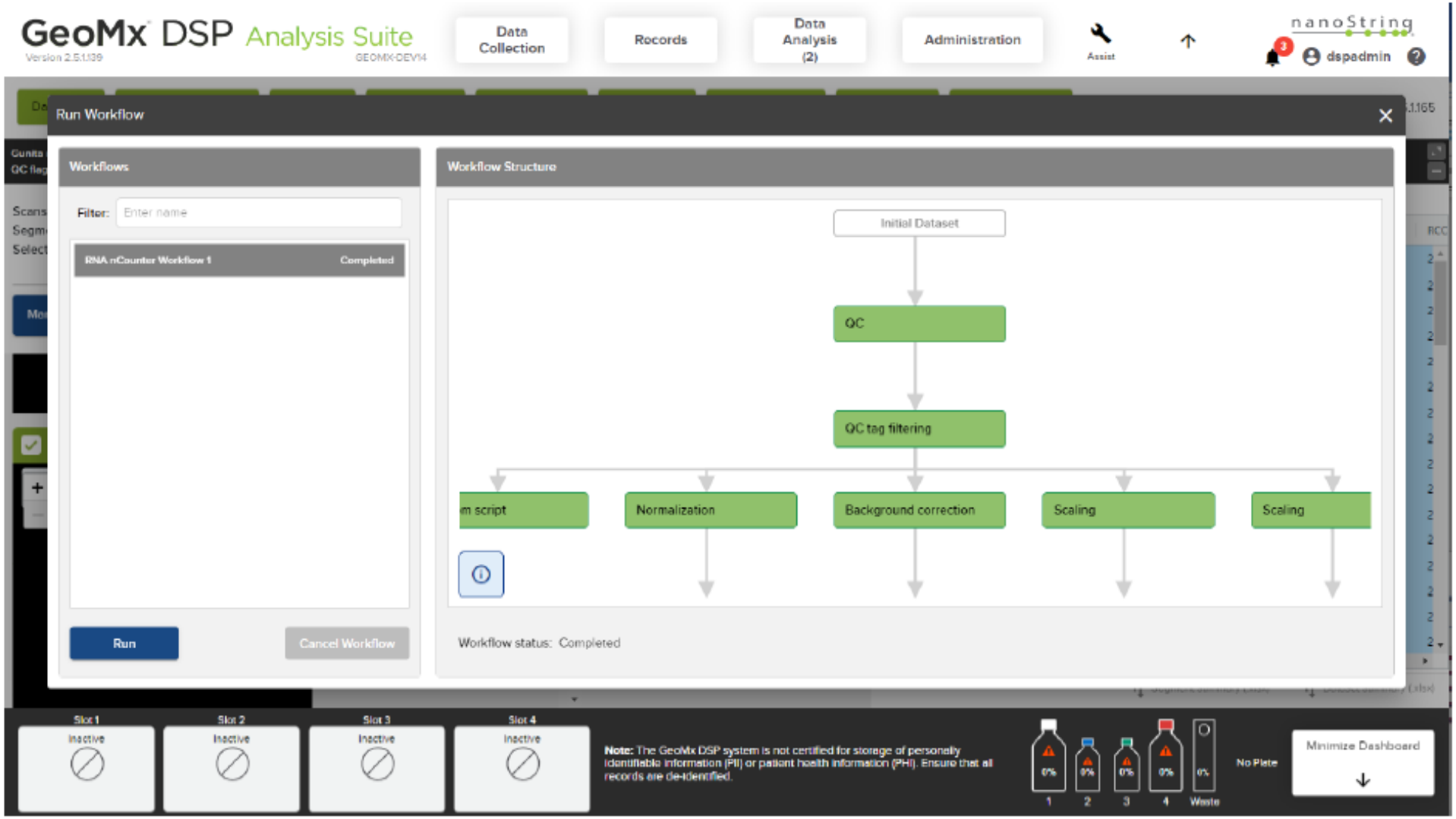


Figure 49: Run Workflow window

Workflows combine multiple data analysis steps into a single streamlined operation. Ten workflows are pre-loaded into the GeoMx DSP Data Analysis suite.
Their parameters can be customized to your specifications. At this time, it is not possible to create a new workflow from scratch.

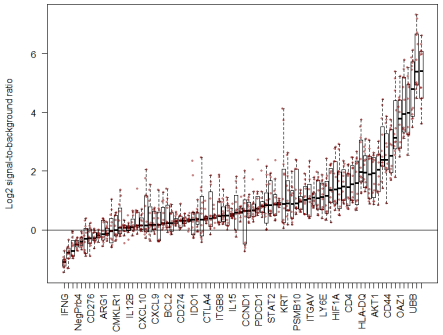
FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

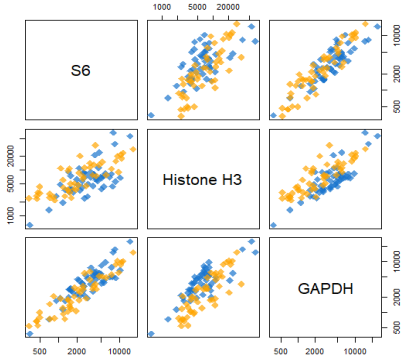
Custom Plugins

Data analysis

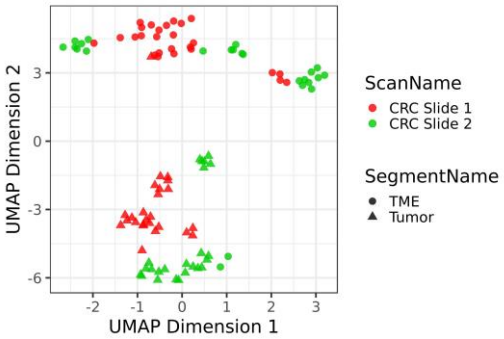
Current custom plugins for the DSPDA

Evaluate Normalization for Protein

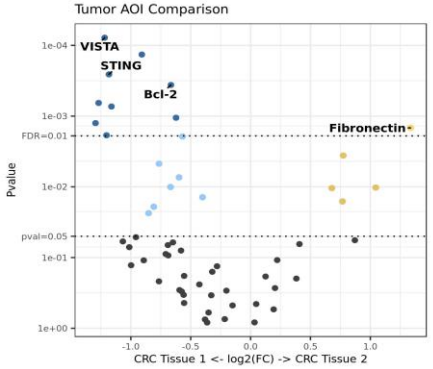




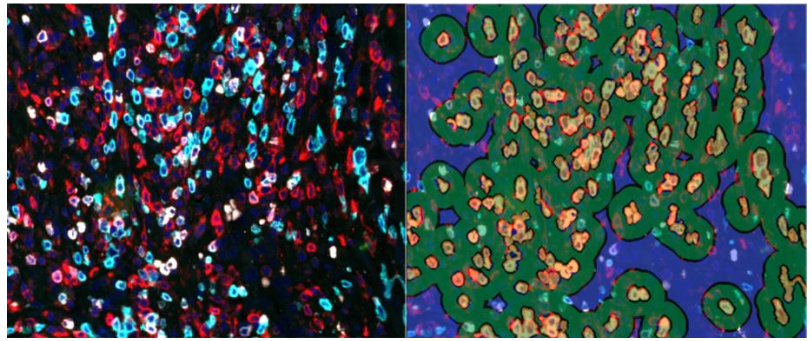
Dimension Reduction



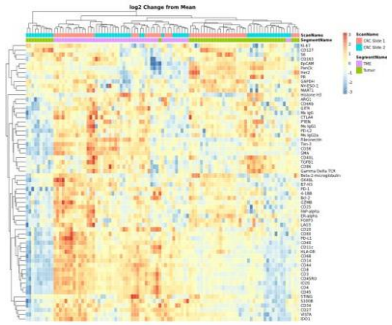
Volcano Plot



Rare Cell Contouring



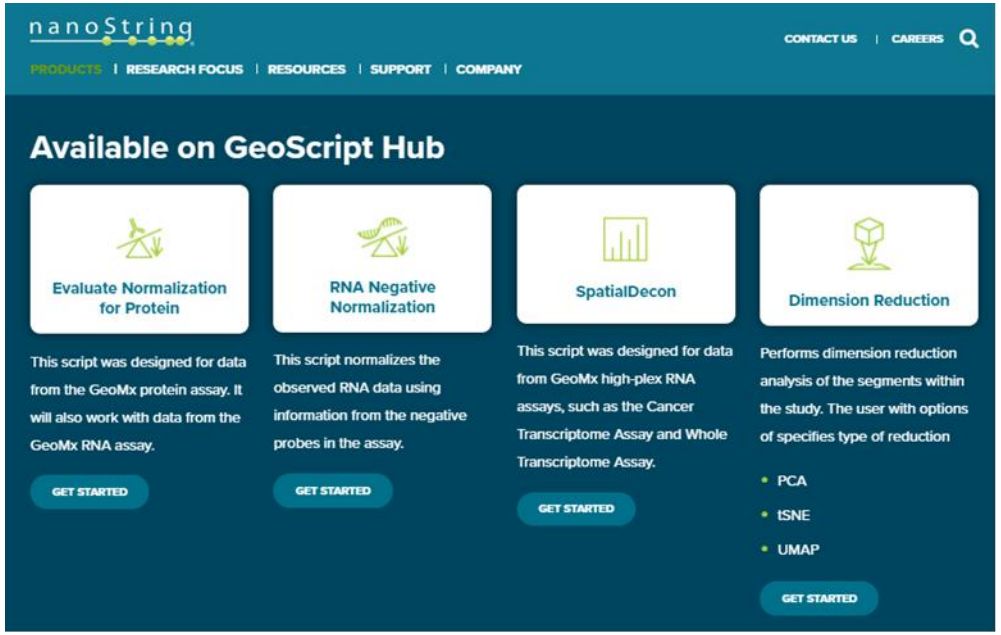
Heatmap

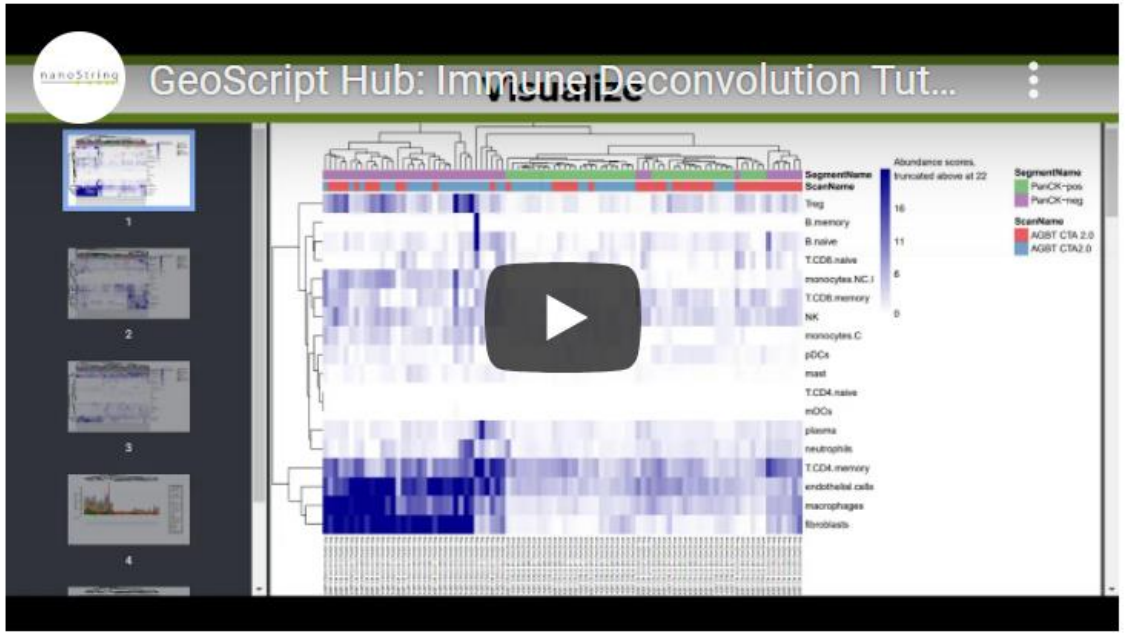


FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Purpose of the GeoScript Hub

- Create customizable figures and/or perform advanced applications within our software
  - https://www.nanostring.com/products/geomx-digital-spatial-profiler/geoscript-hub/
  - Whitepapers
  - Bioconductor community





Three analysis avenues to consider

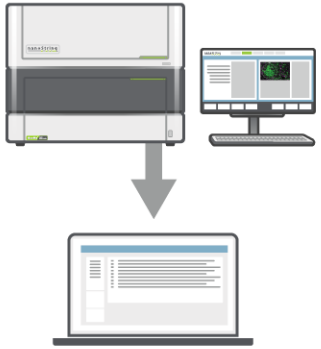
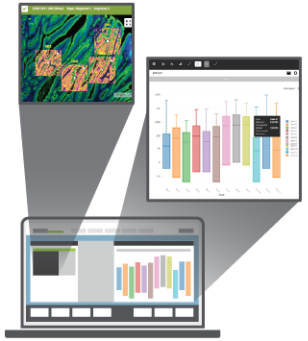
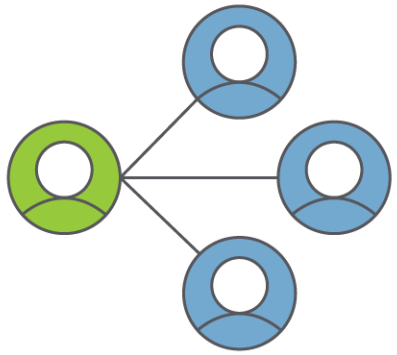
I want to analyze my data and do not code	I have biostats/bioinformatics experience/support	I want to outsource my data analysis
GeoMx DSP Analysis Suite	For NGS RNA Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS	sDAS SPATIAL DATA ANALYSIS SERVICE
nanoString GeoScript Hub	R	R
NANO U	NANO U	NANO U

nanoString

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Process for the Spatial Data Analysis Service (sDAS)

- We have launched!
- Please interface with your field team

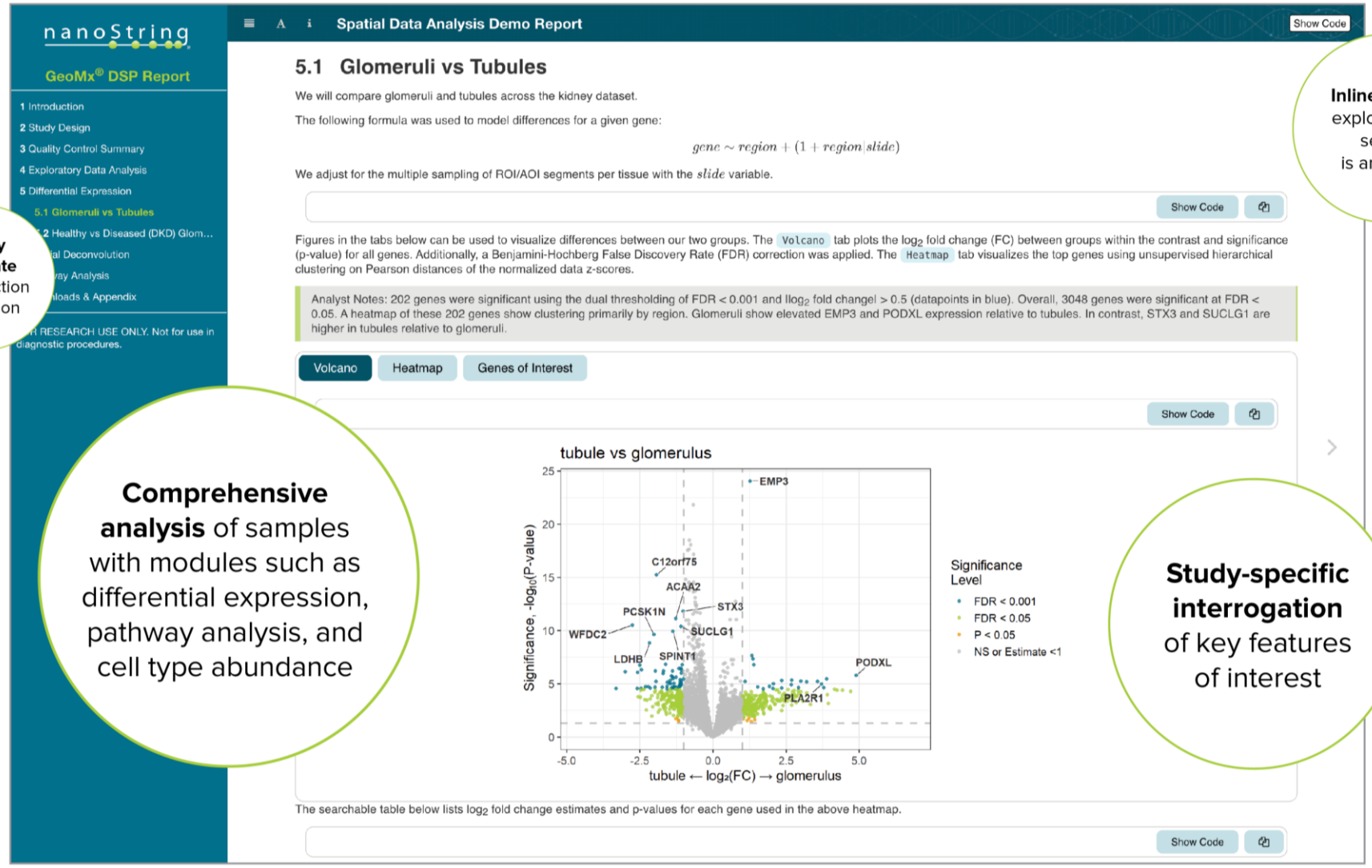
Obtain Customer Files	Study QC	Report Analysis	Customer Report Walkthrough
Day 0	**1-2 Days**	**3 Weeks**	
			
Customer exports and sends GeoMx data via Box* to NanoString	A NanoString computational biologist reviews input data for accuracy and likelihood of analysis success. Analysis continues unless concerns are identified.	The computational biologist performs statistical analysis addressing pre-specified hypothesis provided by the customer. Results of full report are peer reviewed internally for accuracy.	Report is returned to customer and meeting is scheduled to review report output and discuss the findings.

* Box is the default data-sharing method, but other methods shall be considered based on customer need.
Additional fees may apply if infrastructure needs to be stood up or maintained. Reports available on Box for up to 1 year after analysis.

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

Presentation-ready data visualizations and analysis report



Inline Code to explore how a section is analyzed*

Easily navigate from section to section

Comprehensive analysis of samples with modules such as differential expression, pathway analysis, and cell type abundance

Study-specific interrogation of key features of interest

https://nanostring.com/resources/spatial-data-analysis-demo-report

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nanoString

Takeaways and next steps

- Translatable framework: QC/norm, visualize/cluster, statistics, discovery
- Three avenues to explore data
  - DSPDA
  - Opensource code
  - sDAS
- https://university.nanostring.com/
  - See in particular (as refresher)
  - https://university.nanostring.com/geomx-dsp-data-analysis-user-manual
  - https://university.nanostring.com/normalization-of-geomx-dsp-protein-data-using-a-custom-script
  - https://university.nanostring.com/new-features-of-geomx-software-v25-including-data-analysis-workflows

FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

Q&A

nanoString®