Multiomics Analysis Software™

User Manual v2.0

BioLegend®

Table of Contents

It is highly recommended that this manual be read in its entirety before using this product.

Table of Figures

Introduction and General Information

Introduction to MAS

Multiomics Analysis Software™ (MAS) is a cloud-based program for the analysis of multiomics cytometry data generated through Next-Generation Sequencing (NGS) and related analysis pipelines. MAS offers the following:

- A simple and intuitive user interface for exploring multiomics single-cell data
- Accessibility to all users irrespective of bioinformatics background
- Accessibility through your web browser, reducing the need for extensive computational resources
- Sample demultiplexing based on Cell Hashing antibody staining
- Intuitive, flow cytometry-like gating on TotalSeq™ antibody staining
- Easy creation of UMAP (Uniform Manifold Approximation and Projection), tSNE (t-distributed stochastic neighbor embedding), and TriMap (dimensionality reduction technique based on triplet constraints) plots
- Cell clustering based on protein or RNA expression
- An interactive analysis module for the visualization of expression data and inter-cluster comparisons
- Easy sharing of any workspace with collaborators or BioLegend Technical Services

System Requirements

Multiomics Analysis Software is a cloud-based program and can be accessed by any computer connected to the Internet. MAS is optimized for performance with the following web browser(s):

- Google Chrome (version 74 or greater)

IMPORTANT

» Microsoft Internet Explorer is not supported and should not be used.

Account Activation, Forgotten Password, and Program Login

To request access to MAS, fill out the account request form available on the MAS home page at: https://www.biolegend.com/en-us/totalseq/MAS. Please allow 24-48 hours (~1 to 2 business days) for account activation.

To log in, navigate to https://mas.qognit.com/ and enter your credentials:

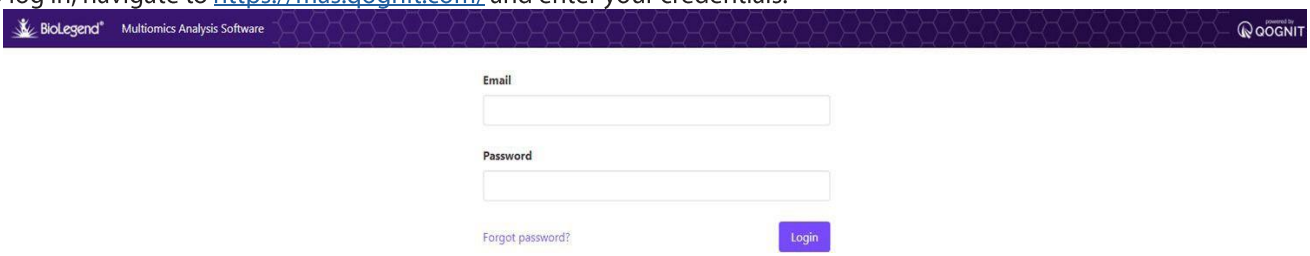

Figure 1. Login page.

Use the "Forgot password?" link to reset your password or if accessing MAS for the first time.

Data Security

Registered account users retain full control of any data they upload to the cloud servers. Account owners have sole permission and rights to add or delete data sets to the program. At no time will any other user be able to access your data, unless you explicitly grant them permission to do so via the share functionality in the program. This permission can also be revoked at any time. BioLegend hosts the Multiomics Analysis Software™ on virtual servers provisioned by world-class, leading cloud providers like Amazon EC2, Google Cloud, and Digital Ocean. BioLegend will only use cloud providers with data centers that are SOC 2 accredited, and who implement physical security safeguards that at minimum have:

- 24/7 physical security guard services and monitoring
- Physical entry restrictions to their facilities
- Biometric readers with two-factor authentication
- CCTV coverage of the facility

All incoming and outgoing data transfers through the Multiomics Analysis Software™ are encrypted during transit using modern, secure TLS (HTTPS) cryptographic protocols.

Supported Data and Input Files

MAS supports single-cell data that contains sequencing data for the following:

- TotalSeq™ Cell Hashing Antibodies (HTO) + 10x Genomics Gene Expression (5′ or 3′) + TotalSeq™ Antibodies (ADT) + V(D)J (optional with 5′ chemistry)
- TotalSeq™ Cell Hashing Antibodies (HTO) + 10x Genomics Gene Expression (5′ or 3′) + V(D)J (optional with 5′ chemistry)
- TotalSeq™ Cell Hashing Antibodies (HTO) + TotalSeq™ Antibodies (ADT)
- TotalSeq™ Cell Hashing Antibodies (HTO) only
- TotalSeq™ Antibodies (ADT) only

The inputs for MAS are count files generated by either Cell Ranger (**.tsv.gz*) or Seurat (**.rds*).

Required Cell Ranger Files:

- *Barcodes.tsv.gz*
- *Features.tsv.gz*
- *Matrix.mtx.gz*
- *VDJ contig (optional with 5′ chemistry)*

Cell Ranger Files Descriptions:

Barcodes.tsv.gz:

This file contains cell barcodes identified during sequencing. Each row represents a unique barcode while the row position number represents the index position for a given barcode.

	A
1	AAACCTGAGAACTGTA
2	AAACCTGAGAAGAAGC
3	AAACCTGAGAATAGGG
4	AAACCTGAGATTACCC
5	AAACCTGAGCACGCCT
6	AAACCTGAGCAGGCTA
7	AAACCTGAGCCACCTG
8	AAACCTGAGCCAGGAT
9	AAACCTGAGCCAGTTT
10	AAACCTGAGCCCAATT
11	AAACCTGAGCCGGTAA
12	AAACCTGAGCGGATCA
13	AAACCTGAGCTACCTA

Figure 2. Example field structure of barcodes.tsv.gz file (unzipped).

Features.tsv.gz:

This file contains information about TotalSeq™ ADTs (Antibody Derived Tag) used for staining and gene expression targets. In the example shown in Figure 3, we see HTO (Hash Tag Oligos), ADTs (Antibody Derived Tags), and Ensembl gene IDs listed in the first column. For experiments using Cell Hashing reagents, the second column must contain the sample ID, which is derived from the index primers used during sequencing library preparation. For ADTs, the second column must contain the epitope that the ADT identifies.

The third column contains the type of feature listed. As shown in Figure 3, there are Cell Hashing, Antibody Capture, and Gene Expression features.

IMPORTANT

» There are special naming conventions required for MAS to automatically delineate ADTs and HTOs in the Features.tsv.gz file. In the first column, HTOs are preferably named with "HTO_" prefixes. This will allow MAS to automatically identify the HTOs in a dataset. If the hashtags were used and their names do not start with "HTO_", the user is required to search and select the HTO features during the upload process (see section 1.3 Select HTO). In the third column, ADTs and HTOs need to be labeled "Antibody Capture", and Ensembl gene IDs need to be labeled "Gene Expression".

ENSG00000160285	LSS	Gene Expression
ENSG00000223901	AP001469.1	Gene Expression
ENSG00000215424	MCM3AP-AS1	Gene Expression
ENSG00000160294	MCM3AP	Gene Expression
ENSG00000228137	AP001469.2	Gene Expression
ENSG00000239415	AP001469.3	Gene Expression
ENSG00000182362	YBEY	Gene Expression
ENSG00000160298	C21orf58	Gene Expression
ENSG00000160299	PCNT	Gene Expression
A0403	CD207_Langerin	Antibody Capture
A0404	CD63	Antibody Capture
A0405	CD284_TLR4	Antibody Capture
A0406	CD304_Neuropilin-1	Antibody Capture
A0407	CD36	Antibody Capture
A0408	CD172a_SIRPa	Antibody Capture
A0409	CD85g_ILT7	Antibody Capture
A0410	CD38	Antibody Capture
A0418	CD243_ABCB1	Antibody Capture
A0419	CD72	Antibody Capture
A0420	CD158_KIR2DL1.S1.S3.S5	Antibody Capture
A0423	MERTK	Antibody Capture
A0427	Folate_Receptor_B_FR-B	Antibody Capture
HTO_7	HTO_7_d643Rest	Antibody Capture
HTO_10	HTO_10_d643Act	Antibody Capture
HTO_12	HTO_12_d670Act	Antibody Capture
HTO_8	HTO_8_d670Rest	Antibody Capture
HTO_9	HTO_9_d808Rest	Antibody Capture
HTO_13	HTO_13_d808Act	Antibody Capture

Figure 3. Example field structure of features.tsv.gz file (unzipped).

Matrix.mtx.gz:

	A
1	%%MatrixMarket matrix coordinate integer general
2	15 36000 539817
3	1 1 8
4	2 1 10
5	3 1 17
6	4 1 10
7	5 1 396
8	6 1 101
9	7 1 418
10	8 1 91
11	9 1 1306
12	10 1 29
13	11 1 104
14	12 1 53
15	13 1 81
16	14 1 21
17	15 1 36
18	1 2 629
19	2 2 11
20	3 2 13
21	4 2 8
22	5 2 8
23	6 2 18
24	7 2 48

Figure 4. Example field structure of matrix.mtx.gz file (unzipped).

This is a sparse matrix format file that stores a counts-per-cell-matrix. The first two rows in the matrix file represent header lines. Row two contains the number of lines in each file provided (features, barcodes, matrix), respectively. For matrix files, the count in row two does not include the first header row. All subsequent rows contain the following information:

1st number: Feature index (row position in features file)

2nd number: Barcode index (row position in barcodes file)

3rd number: Total Unique Molecular Identifier (UMI) count per cell and gene combination

The MTX format is a standard output of 10x Genomics Cell Ranger software. MAS supports MTX format files that are generated with Cell Ranger v3.1 or above.

Getting Started

Why MAS?

The goal of MAS is to support users with secondary analysis of single-cell proteogenomics data which includes enabling users to interactively explore and analyze protein, RNA, and/or clonotype/V(D)J data across cell types. In addition, users will be able to explore and discover protein and RNA expression in multiple cell types that are associated with sample level endpoint data. To accomplish this, MAS provides various data processing functions including the ability to remove unwanted events; demultiplex; identify highly expressed and variable genes and define cell types; use gating and/or clustering methods to identify cell populations of interest based on antibody and RNA expression; and create dimensionality reduction (projection) plots. MAS offers users an easy-to-use workflow that accomplishes these goals.

Recommended Workflow

BioLegend recommends the following workflow to accomplish the goals listed in the above section:

1. Upload input files, demultiplex, and QC (Quality Control) filtering
    a. Upload count data in sparse or RDS formats
    b. Select Hashtags if applicable
    c. Demultiplex data set
    d. Review QC filtering tab (UMI/Mito/Isotype/ADT filters)
    e. Review the Demux results tab if HTOs (Cell Hashtags) are included
2. Process RNA expression data (if Gene Expression data is included)
    a. Identify HEG/HVG and review results
    b. Normalize RNA Data
    c. Define Gene Modules (optional)
3. Add Sample Annotations and/or merge samples if desired.
4. Define cell types via gating and/or clustering
    a. Use gating to identify cells of interest (e.g., T cells as $CD3^+$ $CD19^-$) based on known lineage markers
    b. Apply clustering (in combination with gating if desired) to define, explore, and refine additional cell type definitions
5. Create projections (dimensionality reduction plots such as UMAP, tSNE, or Trimap)
6. Conduct analyses
    a. Cell type analysis: Compare cell types to each other by protein and/or RNA and explore individual markers
    b. Cross-sample: Perform paired or unpaired differential analysis across samples to compare protein/RNA expression in different cell types to user-provided endpoints

IMPORTANT

» Note that steps 4 and 5 may be iterated over. Projections may be created for specific cell types (e.g. only T cells). As such, having some cell types defined prior to projections is beneficial. However, having projections is also useful when applying clustering methods. Therefore, one may go back and forth between steps 4 and 5 depending on the dataset. This manual will guide users through this process.

1. Uploading Count Data, Demultiplexing, and QC Filtering

1.1 Loading Data

After login, click the "+" icon to create a new project:

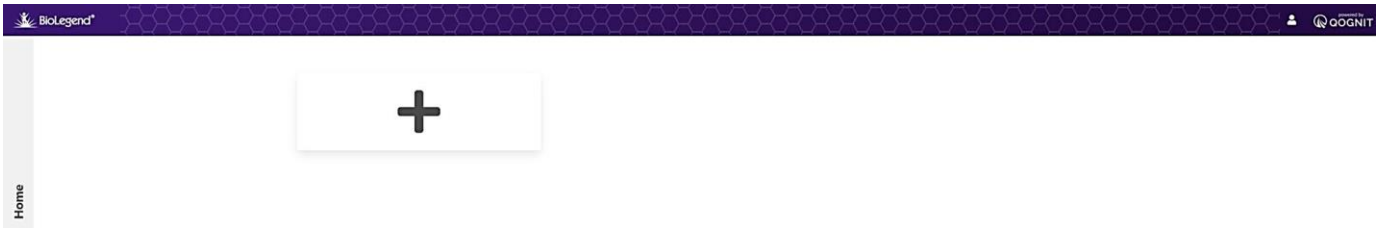

Figure 5. New project icon.

Enter a name for the dataset and click "Create":

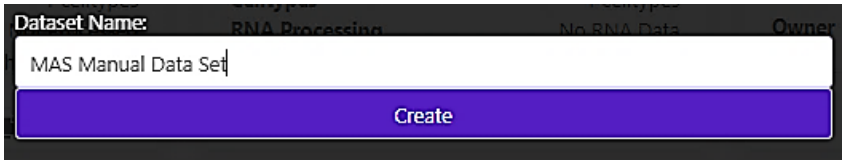

Figure 6. Naming a dataset.

Click the "Add New Fileset" button:

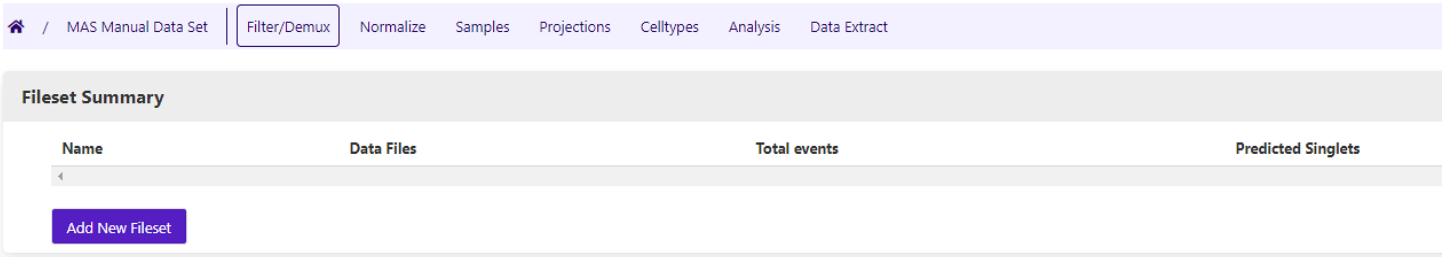

Figure 7. Adding new fileset.

If uploading 10x Genomics filtered count files, select the "10x" radio button and click "Select features, barcodes, and matrix files":

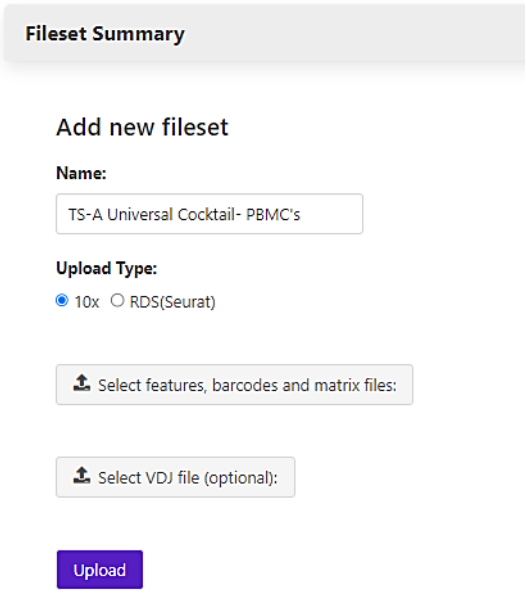

Figure 8. Upload options for file types.

Navigate to the folder containing the files, select all three files, and click "Open":

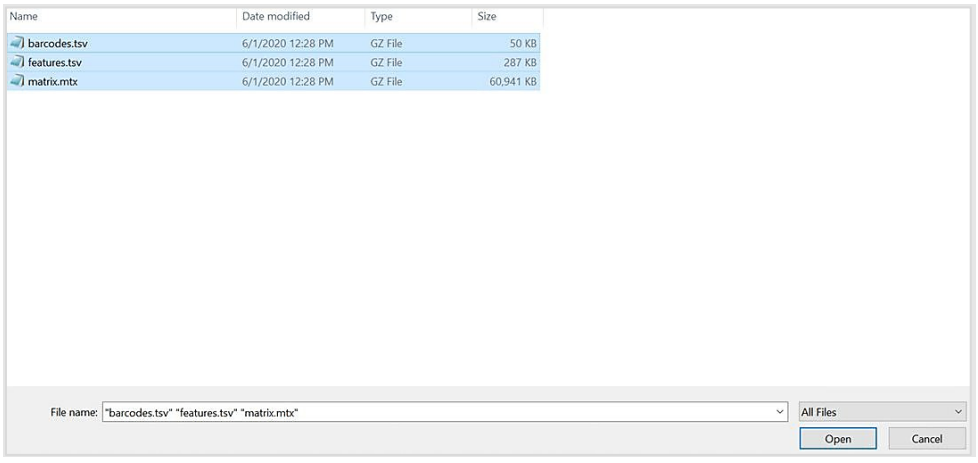

Figure 9. Selecting files to upload.

If uploading **.rds* files, select the "RDS" radio button and select a **.rds* file to upload:

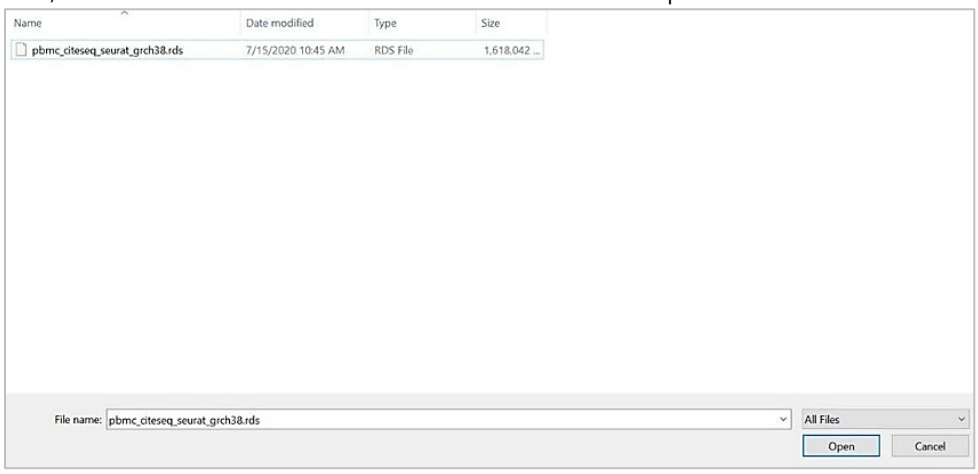

Figure 10. Selecting *.rds file(s).

Make sure the files selected are present in the correct order as shown in Figure 11. If changes are required, select the correct file from the drop-down menus:

Figure 11. Confirming file order.

If no HTOs are included in the data set, the "Select Hashing/Multiplexing Parameters" will display No Hashing/Multiplexing as seen in Figure 12. If this is correct, click "Upload" to proceed. If HTOs are included, proceed to 1.3 Select HTOs.

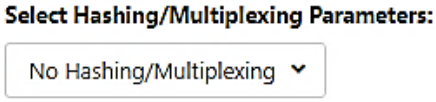

Figure 12. No HTOs included in dataset.

Click "Begin Filtering" to proceed to QC filtering results.

1.2 V(D)J Contig File (optional)

If the experiment contains V(D)J click on "Select VDJ File" to browse to the V(D)J contig file output from the Cell Ranger workflow. Open the respective file, if applicable. Then click "Upload" to proceed.

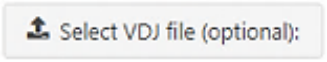

Figure 13. VDJ file selection.

1.3 Select HTOs

If HTOs are included in the data set, HTOs containing the "HTO_" prefix will be automatically selected from the features.tsv.gv file. Alternatively, if the HTOs were not denoted in the features.tsv.gz file, users can manually select the HTOs from the drop-down menu as seen in Figure 14, even if they do not follow the "HTO_" prefix naming convention. Once the HTOs are denoted in the dataset, proceed to upload the data.

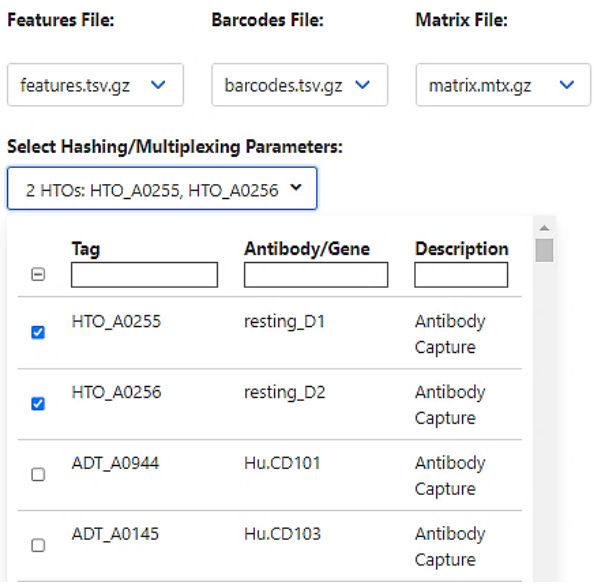

Figure 14. When selecting HTO features, those starting with 'HTO_' will be automatically selected.

Uploading will commence and may take up to several minutes to complete.

Figure 14. Upload progress bar

MAS will output an HTO statistics table (Figure 15). This table summarizes count statistics for each HTO. Any HTOs without a corresponding sequenced sample are expected to have very low counts. In these rare cases, the user should exclude the HTOs prior to executing the demultiplexing step by de-selecting the check boxes. The user should expect high values for the 95th percentile and above for each HTO, suggesting the cells are positive for staining for the respective HTO. Similarly, the lower percentiles should generally display low counts as this represents cells (belonging to other samples) that are negative for HTO staining.

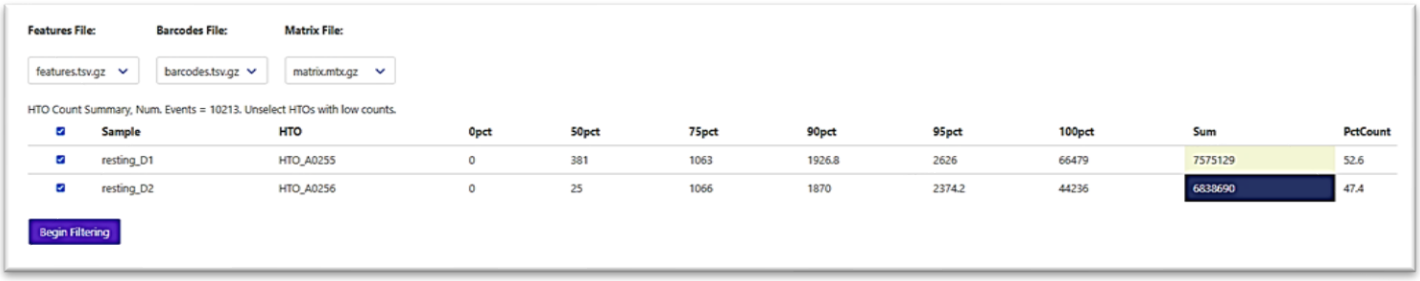

Figure 15. HTO statistics table.

Click "Begin Filtering" to proceed to QC filtering and HTO demultiplexing results.

2. Renaming, Sharing, and Deleting Projects

MAS offers users the ability to share projects with collaborators that have MAS accounts. Sharing and other basic project administrative functionalities can be accessed by first clicking the "Home" menu on the left and then hovering over the ☰ menu for a specific project. Alternatively, the "share" icon can also be accessed at the bottom of each of the project cards:

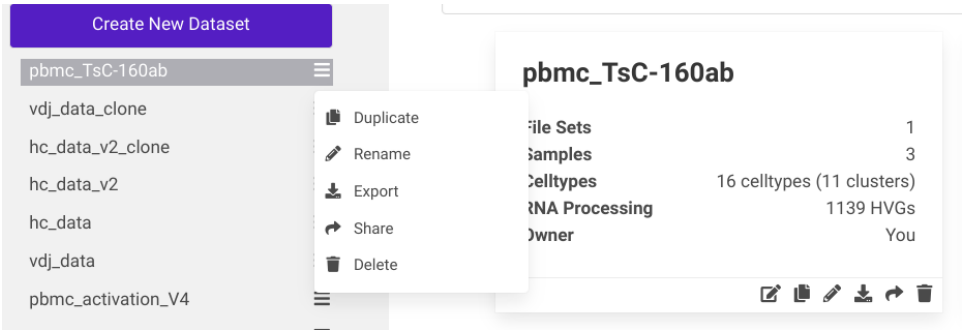

Figure 16. Experiment options.

To share a project, click on the "Share" icon and enter an email recipient:

Figure 17. Sharing experiments.

3. Filter/Demux Summary Page

After uploading the data, MAS will automatically present the User with the Filter/Demux tab. This tab will contain a file set summary that displays the file set(s) that have been uploaded (Figure 18). It will also display the number of Total Events and Predicted Singlets based on various QC filtering metrics, which will be covered in the sections below.

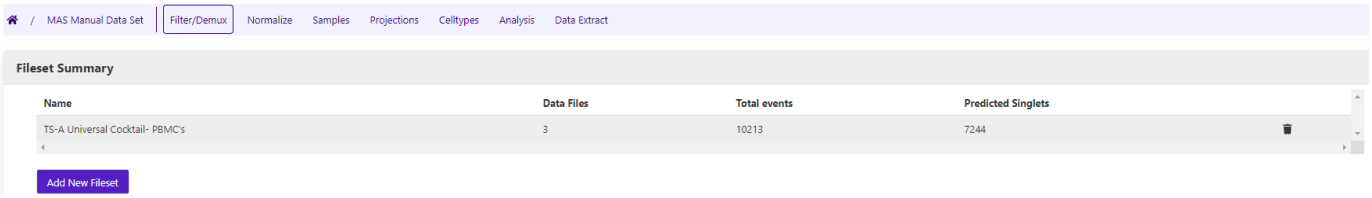

Figure 18. File set summary.

3.1 Filter Tab

Within the Filter/Demux Page, the Filter tab will display a table summarizing the various QC filters that have been applied to the data set. This section will cover these filters and elaborate on their functionality.

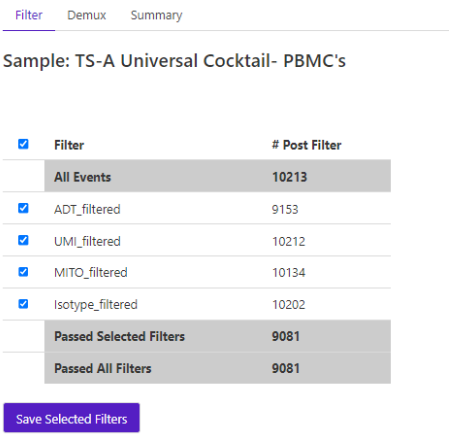

Figure 19. Filter tab summary table.

3.2 UMI Filter

MAS generates knee plots to visualize thresholds for filtering events (cells) based on UMI count complexity. Events are ranked on the x-axis according to the number of UMIs (Unique Molecular Identifier) detected. The y-axis displays the total number of UMIs for each cell. UMI-rich cells are found on the left-hand side of the x-axis, which are excluded, as these likely represent cell aggregates/doublets. On the right-hand side of the plot, past the "knee", are barcodes that have relatively low numbers of reads, likely debris, making them unsuitable for further analyses. MAS will generate knee plots ranking the UMIs for ALL the

events included in the analysis as well as individual plots for the UMIs specific for HTO, ADT, and RNA, if applicable. MAS will automatically create inclusion gates for each of the plots, but users have the option to edit the gating by selecting the individual plots and adjusting the gate accordingly.

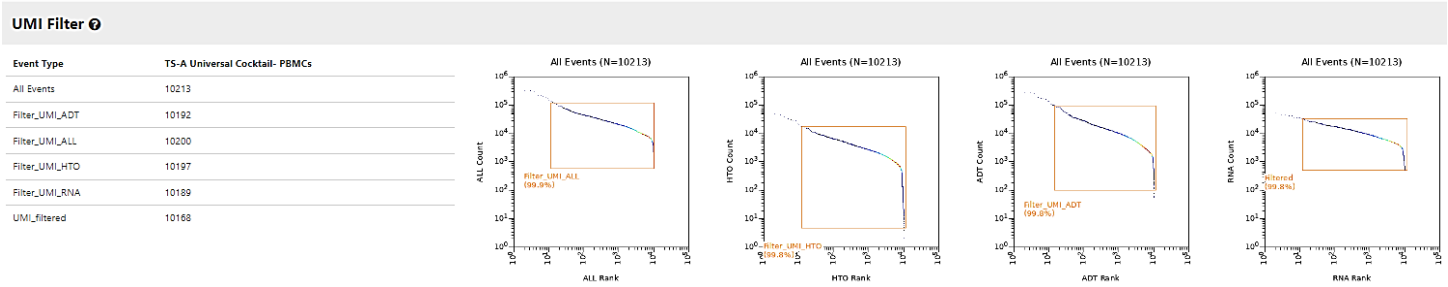

Figure 20. UMI filter knee plots.

3.3 MITO Filter

Typically, events with a high degree of mitochondrial reads are associated with dead or dying cells. MAS generates a plot that ranks RNA counts (x-axis) by the fraction of mitochondrial reads (y-axis). This allows the user to visualize the percentage of cells with high mitochondrial reads (top right) and can be used to exclude those cells from further analysis. MAS will automatically generate an inclusion gate setting the threshold at 0.8, but the user can change this gate by clicking on the plot and adjusting the gate location.

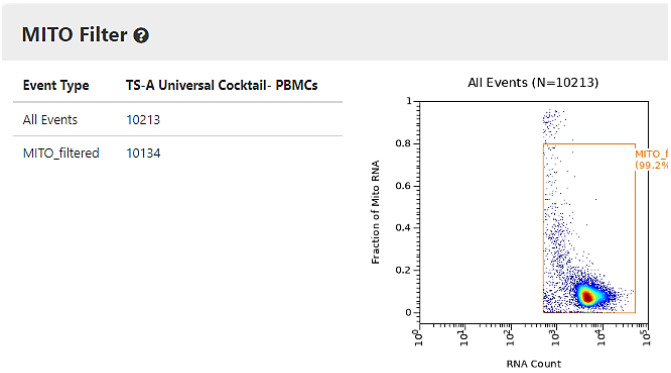

Figure 21. Mitochondrial gene UMI count plot.

3.4 Isotype Filter

If isotype control antibodies are included in the dataset, the isotype filter can be useful for removing particularly sticky cells from further analysis. In theory, isotypes should not be binding cells highly since they lack specificity. Therefore, these sticky cells may be binding to other antibodies non-specifically, and an inclusion gate can be set to filter out cells showing a high degree of isotype reads. It is recommended to set the positivity threshold to a minimum of two, meaning if a cell is positive for two or more isotypes (based on the gates set), it is excluded from further analysis. MAS will autogenerate the plots and inclusion gates, but the user can adjust the gates by clicking on the individual plots. The removal of cells positive for isotype control antibodies is not always necessary and may depend on the experimental sample(s). Furthermore, aggressive removal of cells showing staining for isotype control antibodies may result in removing certain cell types from the analysis (e.g., monocytes which are known to stain non-specifically).

IMPORTANT

» Note, isotype controls are not meant to be used for quantitative background subtraction. It is also not necessary to have an isotype control matching each antibody isotype included in the staining. Isotypes are meant to be used as an additional filter to identify and exclude potentially sticky cells. Typically, at least 2 isotypes of the most common isotype class in the staining is sufficient for filtering.

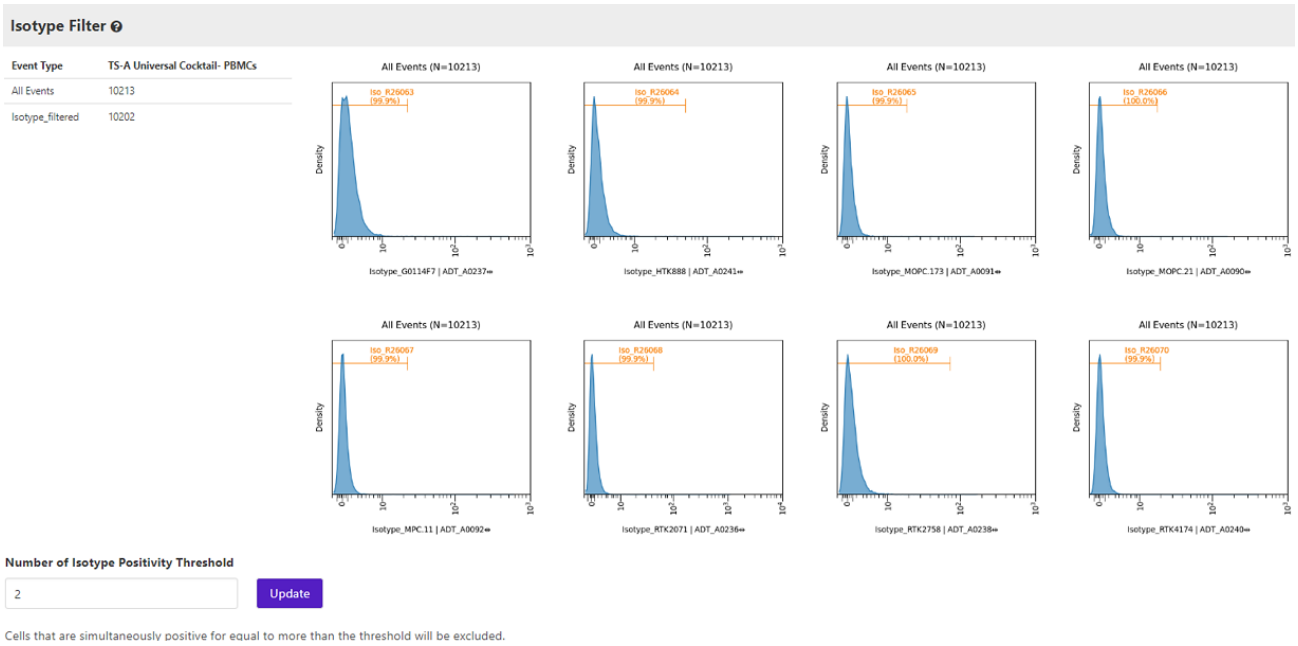

Figure 22. Isotype filter plots.

3.5 ADT Filter

It is recommended to remove cells that are double positive for mutually exclusive markers, such as CD3 and CD19. The double positive population(s) should not biologically exist, so in most cases these are likely sticky cells and should be removed from analysis. If mutually exclusive antibody markers are detected in the data, MAS will autogenerate plots and set gates excluding the double positive population (top right of plots). Users can adjust the gates by clicking on the individual plots.

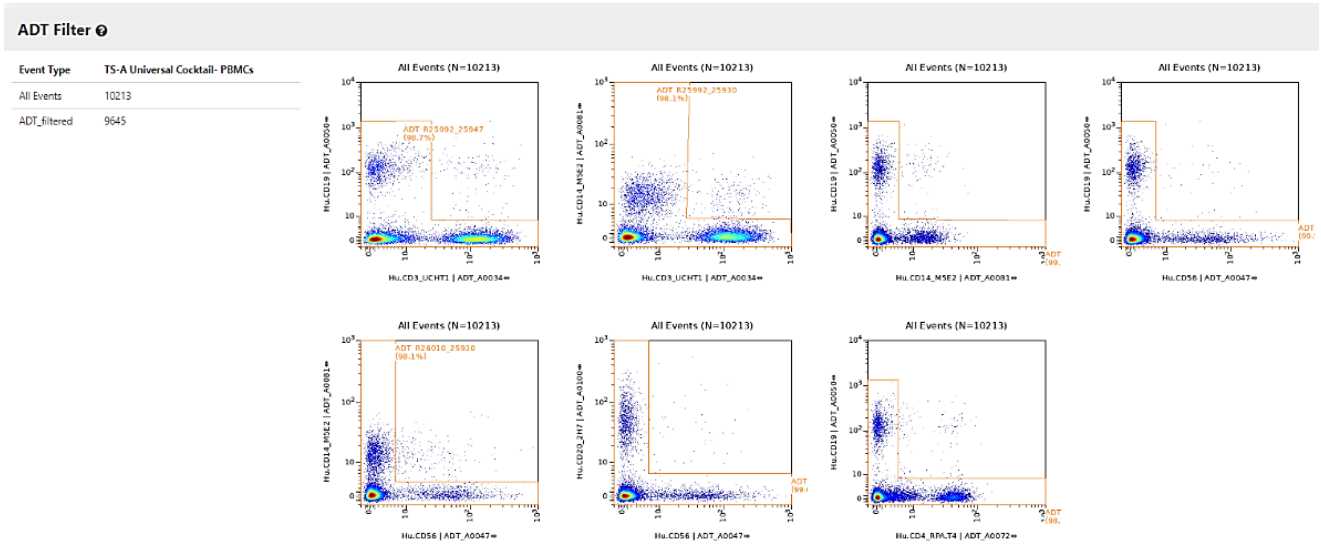

Figure 23. ADT filter: mutually exclusive markers.

3.6 Demux Tab

If HTOs were selected in the count upload steps, MAS by default will demultiplex by assessing the "Purity" of HTO expression for an event (i.e., cells that are positive for each of the HTOs and simultaneously negative for all the other HTOs). The Demux tab (Figure 24) will contain a table that denotes: the selected HTOs along with the respective number of "cells" positive (based on the threshold set in the histogram plots at the bottom of the page) for the corresponding HTO and negative for the remaining HTOs; the median value for this HTO; and other HTOs in these cells. The histogram plots are interactive, and the threshold can be further adjusted. If the gates are edited, click save at the top right of the plot to proceed. MAS will

automatically re-run the demultiplexing based on the updated threshold(s). The UMAP bivariate plots on the right will display the expression of the hashtags and can be useful for determining the correct threshold on the histograms. Clicking on a row will highlight cells in the UMAP positive for that respective HTO. Typically, each sample should be positive for a single HTO, and cells that are either negative or positive for multiple HTOs should be removed from the analysis.

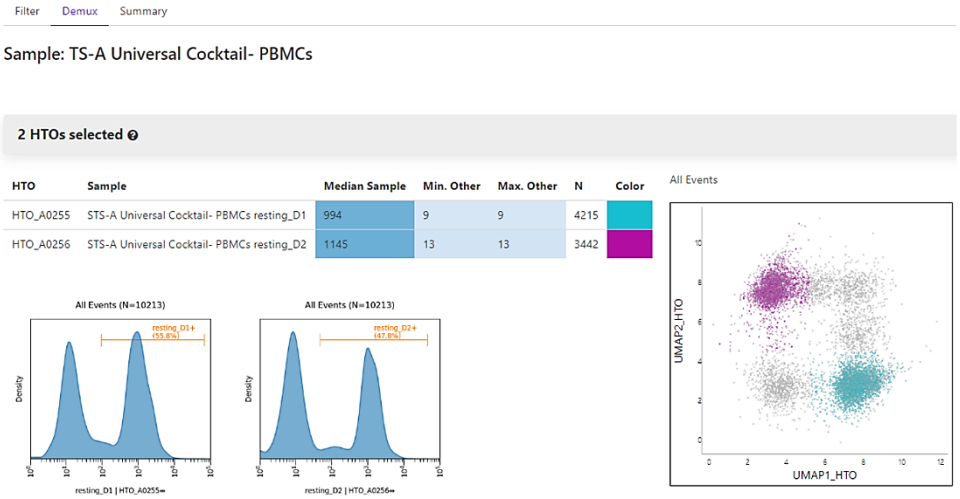

Figure 24. HTO demux results and adjustments.

3.7 Summary Tab

The Summary tab allows the user to examine the staining across the samples for individual ADT and HTOs (if applicable). The table on the left (as seen in Figure 25) provides statistics of the counts for each ADT/HTO for singlet cellular events. The percentiles provide a summary of the staining observed within the samples corresponding to each respective antibody. The columns highlighted in red and blue (one per sample) can be used to understand the uniformity of staining across the samples. The values shown here indicate if there is a higher or lower staining for a given sample compared to the rest. These values are 0.5 minus the Mann-Whitney U, which measures the overlap between two distributions. A value close to 0 indicates no bias, while a value close to 0.5 would indicate strong higher staining for a sample compared to the rest, and a value close to –0.5 would indicate exceptionally low staining for a sample compared to the rest. The user can search for a marker and click on the row to visualize the distribution based on the histogram and UMAP plots on the right.

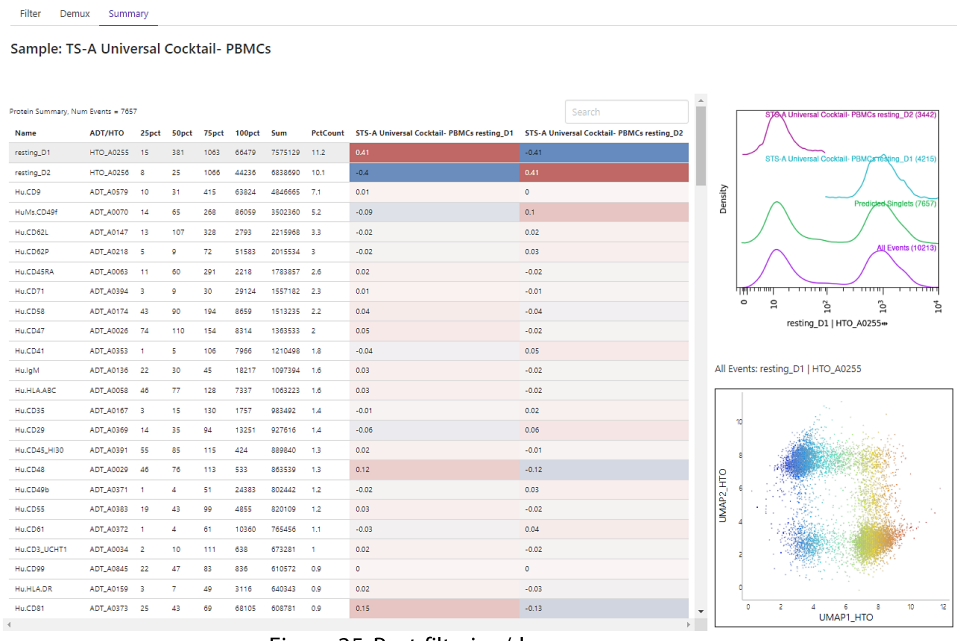

Figure 25. Post-filtering/demux summary.

Once the filtering and demultiplexing are complete, the user can proceed to the Normalization tab, as outlined below.

4. RNA Analysis

4.1 HVG/HEG & Normalization of Gene Expression Data

The Normalization page allows users to perform High Variability Genes (HVG) and High Expression Gene (HEG) selection. HVGs are the set of genes that display sufficient variance across the cells to be useful for downstream analyses (e.g. clustering, differential analysis). In addition, sometimes users may need to include genes that are highly expressed (HEG) in a small fraction of the cells but with low variability. Depending on the desired analysis, it may be beneficial to include these genes for downstream analysis. The data for the combination of HVG and HEG genes is then normalized to account for differences in cell size (e.g., T cells vs. Monocytes).

4.2 Normalize Tab

Click on the Normalize tab located at the top of the page. This will prompt MAS to automatically run the computation to identify HVGs included in the data set. This process may take a few minutes depending on the size of the data set. Once complete, the table below (as seen in Figure 26) will display a summary of the number of events included for analysis, discarded events, the number of HVGs identified, and whether normalization has been completed (Normalized Genes column). If normalization has been completed, the Normalized Genes column will display the normalization method used (scTransform or total_cell_count). Otherwise, it will ask to run HVG first until normalization is completed.

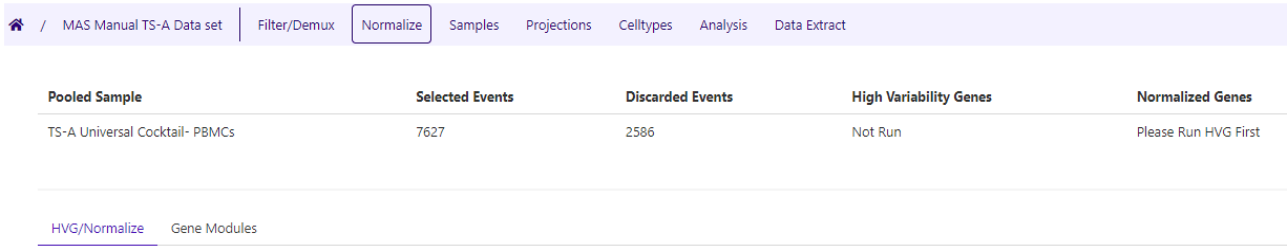

Figure 26. Normalize tab.

4.3 HEG/HVG Summary and Thresholding

Below the summary table, there is an HVG/Normalize subtab which shows bivariate plots and a hierarchy diagram displaying both the number and classification of genes selected based on the HVG computation. HVG calculates the frequency and variability of gene transcripts and selects a threshold for the inclusion of genes in downstream analyses. Only genes that are found to be Highly Expressed (HEG) and/or Highly Variable (HVG) are included. This is done to reduce computational burden and allow analysis to be completed in a reasonable amount of time. The default Median Gene UMI Threshold is set to 20 while the HVGs are set to 1000. The threshold set by MAS is adjustable via the Median UMI Threshold and # High Variability Genes options, however doing so may increase the amount of time required to complete analyses. If adjusting the thresholds, select the appropriate value(s) and click Refresh. This may take a few minutes to complete.

The hierarchy diagram located at the bottom left (Figure 27) contains counts for each type of gene identified (e.g., HVG+HEG+, etc.) and is further visualized by the bivariate plots above. Genes identified as being non-highly expressed and non-highly variable genes (HEV-HEG-) are excluded from subsequent analysis. Clicking on the rows of the hierarchy diagram will highlight the selected population(s) in the plots.

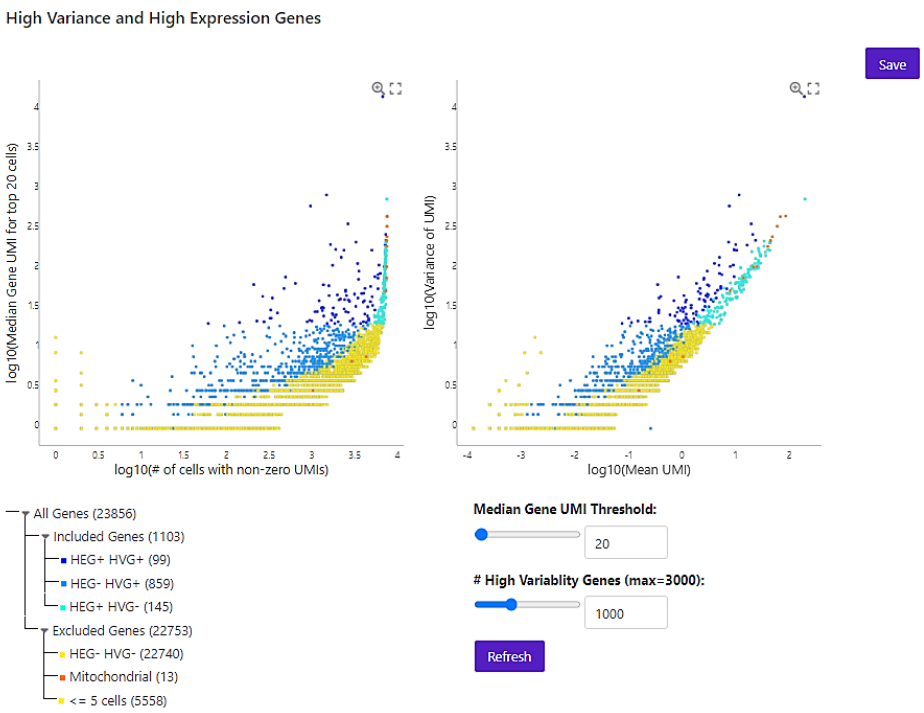

Figure 27. HVG analysis plot and thresholds.

4.4 Selection of HEG/HVG

The table presented in the HVG/Normalize subtab shows the inclusion status of each gene, as seen in Figure 28. Genes with a blue check box are identified as being included for subsequent visualization. Genes with an unchecked box are not included, although users can manually check a gene's box to force the gene to be included for downstream visualization. To search for genes, type in the gene symbol or Ensemble ID. Genes that are False for both "High_var" (HVG) and "High_exp_top" (HEG) in the table will not be accessible in subsequent visualizations. If and once additional genes are selected, click "Save". If no additional genes are selected, proceed to normalization.

Listing of Highly-variable genes (HVG). Click on **Refresh** above to edit.

ID	Gene	Nonzero cell count	Median_top	Mean	High_var	High_exp_top	HVG_rank
ENSG00000244734	HBB	6736	15104.0	192.0	True	True	1
ENSG00000206172	HBA1	959	634.5	7.7	True	True	2
ENSG00000188536	HBA2	1483	873.0	11.6	True	True	3
ENSG00000163736	PPBP	490	80.5	0.9	True	True	4
ENSG00000211675	IGLC1	344	44.0	0.4	True	True	5
ENSG00000158578	ALAS2	265	46.0	0.4	True	True	6
ENSG00000107317	PTGDS	208	65.0	0.4	True	True	7
ENSG00000105205	CLC	61	21.0	0.1	True	True	8
ENSG00000223609	HBD	142	21.5	0.1	True	True	9

Figure 28. Gene exclusion/inclusion for downstream analysis.

4.5 RNA Normalization and Summary

Once the gene selection steps are complete, make sure to save and then click "Normalize":

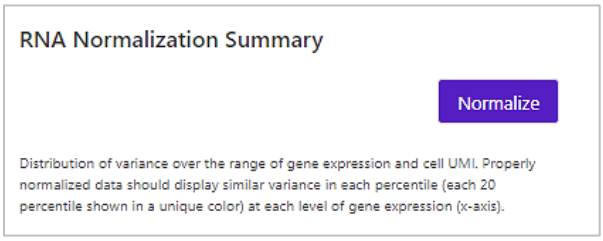

Figure 29. Normalization initiation.

This will display a drop-down menu, Figure 30, where users can select either scTransform or Total Cell Count normalization methods. scTransforms is the recommended method for normalization in MAS.

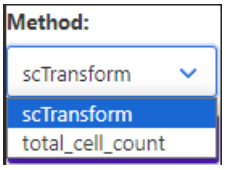

Figure 30. Selecting normalization method.

Click "Run Normalization" to initiate computation. A status page will appear displaying the estimated time to complete the normalization.

Once complete, MAS will display plots summarizing the results of the normalization (Figure 31). Since normalization is expected to adjust the RNA data to account for size differences, it is anticipated that following normalization, the variance in RNA data observed is similar irrespective of cell size. To assess this, variance is computed in the cells belonging to various quantiles of UMI (an indicator of cell size) at various levels (binned by the mean expression on a log scale) of gene expression. Within each gene bin, contribution to the variance from cells belonging to each percentile (in increments of 20) is shown in Figure 31 before (raw plot) and after normalization (normalized). Properly normalized data should show an approximately equal contribution from each percentile of cells within each gene bin.

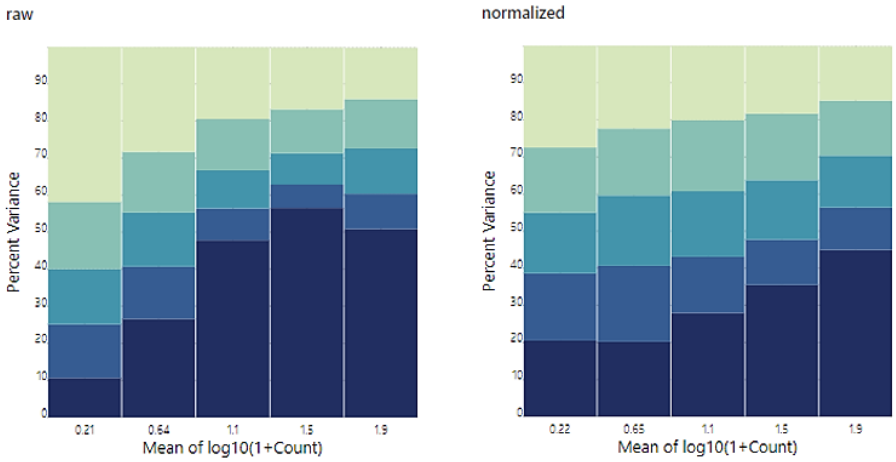

Figure 31. Normalization summary plots.

4.6 Gene Module (optional)

The Gene Module is a feature that allows the user to denote a specific set of genes for which a combined score is computed for each cell. This signature score can be used during analysis to visualize and compare between cell types/samples. Gene sets, for instance, may be obtained from databases like GSEA (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp) corresponding to pathways, cell type, or immunologic states. MAS uses the previously published (see Reference, Gene Sets) method UCell to compute a combined score. Note that only genes that overlap between a gene set (e.g. from GSEA) and those selected by HVG/HEG are included in the computation.

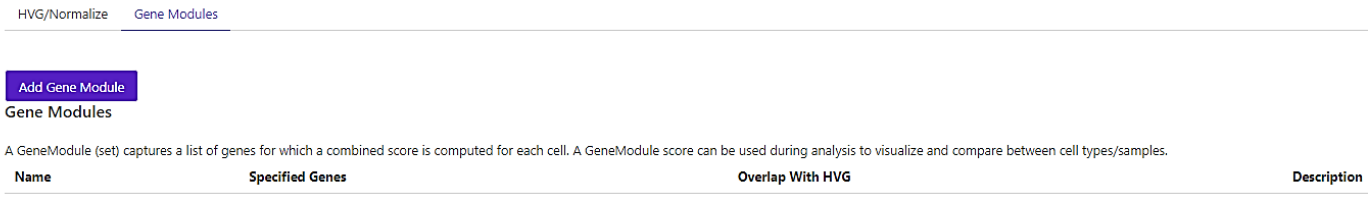

Figure 32. Adding gene modules.

To create a Gene Module, click "Add Gene Module". A pop-up window will appear (Figure 33) where the user can name the specific gene set, input gene names, and include a brief description. Note that only included genes in the HVG analysis can be used as an input. The gene names can be referenced in the gene column of the HVG table (Figure 28) and must be separated using a ",". Once complete, click "Create Gene Module". MAS will display an estimated timer while processing, and once complete, the Gene Module will appear in the column section of the Module page.

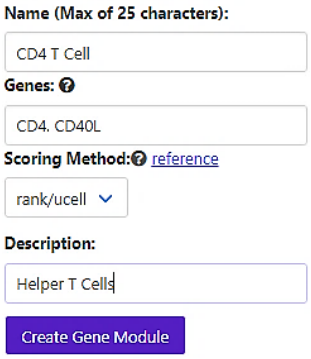

Figure 33. Defining gene module.

5. Sample Merging and Annotation

5.1 Sample Tab: Merging

MAS provides users the ability to merge demultiplexed sequences from the same or different sequencing lanes. This is useful if the same sample is run with multiple HTOs in the same lane or across different lanes (i.e., same count file or across count files). If RNA data is present, it is assumed that HVG and normalization processing have been completed and that the listing of HVGs and ADTs are the same across all samples.

To merge, select the samples to merge using the check boxes in the samples table (Figure 34) and then click merge. The sample can be renamed using the button on the right of the row. Repeat the process on multiple sample selections as needed. Samples that have been merged can be unmerged by clicking on the unmerge icon, , displayed at the end of row with a merged sample (Figure 35). Note: Sample annotation must be performed again after sample unmerging.

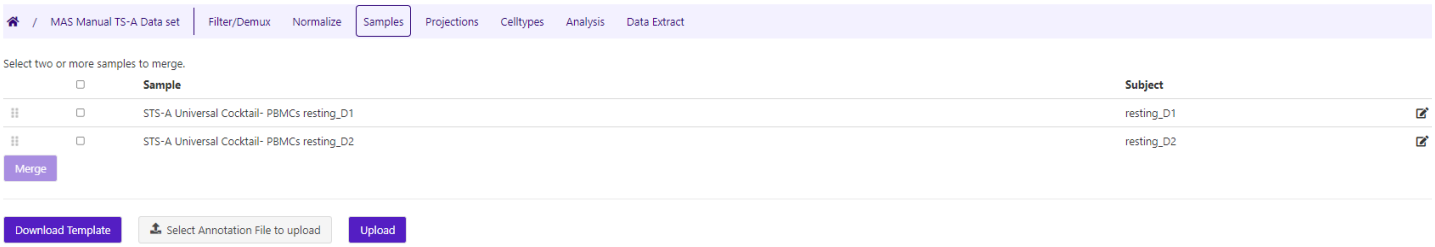

Figure 34. Sample merging.

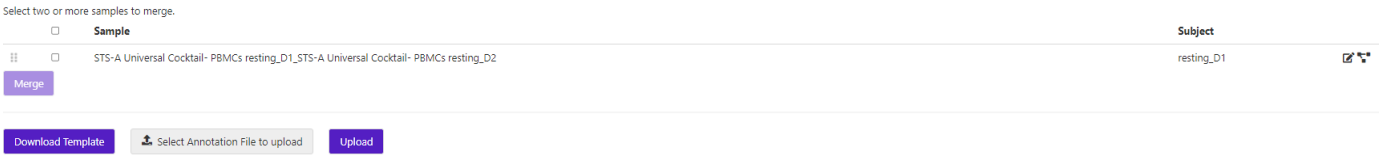

Figure 35. Sample unmerging.

5.2 Annotation Basics and Template

MAS provides the ability to annotate samples included in datasets. This can be helpful in downstream analysis such as Cross-Sample analysis, and/or when sharing projects with collaborators. To annotate samples, download the editable **.xlsx* file template using the "Download Template" button at the bottom of the Samples tab (Figure 36).

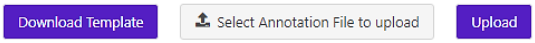

Figure 36. Annotation template options.

Open the **.xlsx* file which will display the metadata of the samples included in the analysis (Figure 37).

	A	B	C
1	**id**	**Sample**	**Subject**
2	381	STS-A Universal Cocktai	resting_D1
3	382	STS-A Universal Cocktai	resting_D1

Figure 37. Editable *.xlsx file.

The downloadable annotations file supports the addition of new columns:

	A	B	C	D	E
1	**id**	**Sample**	**Subject**	**New Column 1**	**New Column 2**
2	381	STS-A Universal Cocktail- PBMC's resting_D1	resting_D1	New Info	New Info
3	382	STS-A Universal Cocktail- PBMC's resting_D2	resting_D1	New Info	New Info

Figure 38. Newly added columns.

IMPORTANT

» Once new annotations have been added, save the file and re-upload it by first clicking the "Select annotation file upload" button to select the file, and then the "Upload" button to upload it (Figure 36). The webpage will now display the updated annotations.

6. Defining Cell Types by Gating and Clustering

Cell types can be defined and assigned to individual cells based on known cell lineages or functionality. Using the Celltypes tab, Users can assign cells for each sample by using a flow cytometry-like gating strategy through a graphical-user-interface (GUI) and/or by applying clustering methods.

6.1 Celltypes tab

To access Cell type definition, select "Celltypes" from the top-level menu:

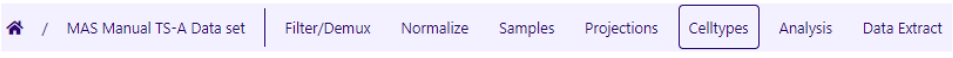

Figure 39. Celltypes tab.

Users will be directed to the following page:

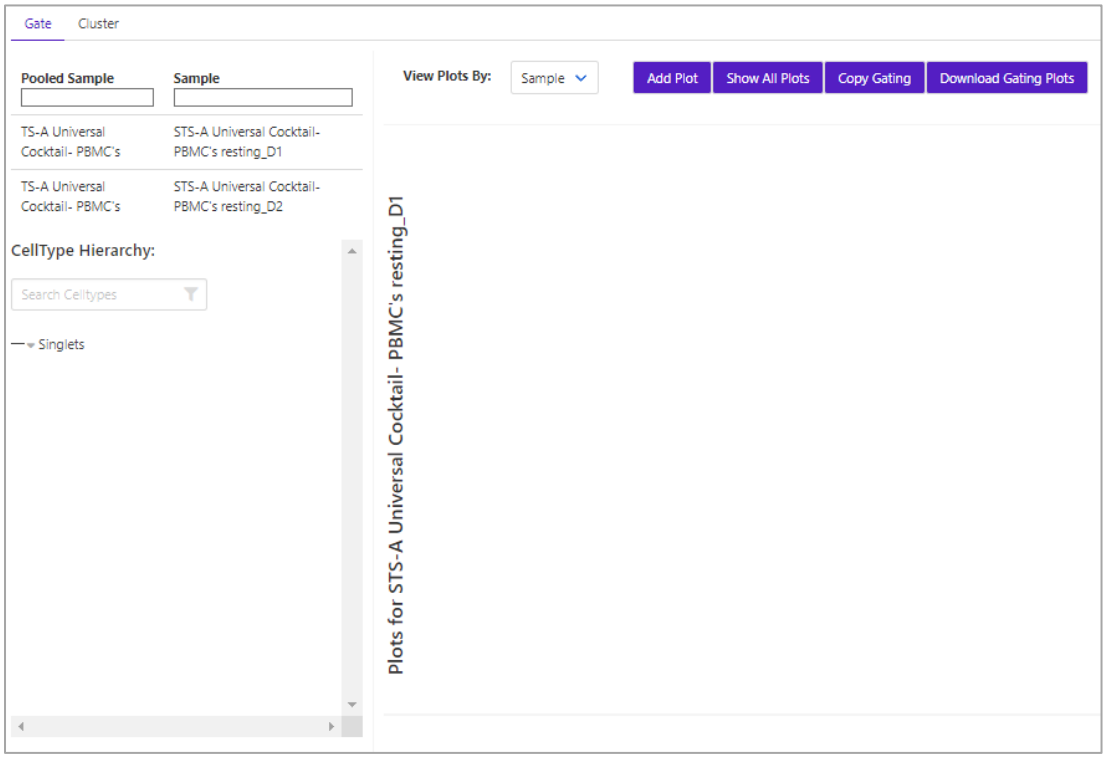

Figure 40. Gating page layout and hierarchy gating.

Cell types can be defined using gating and/or clustering. When gating, the user can draw geometric shapes (polygons, ellipses, quadrants, etc.) on one- or two-dimensional plots of protein/RNA expression to classify cells of distinct types. Gating is extensively used in the cytometry community for cell definition. With a diverse collection of markers, gating can be impractical to define cell types beyond a certain level of depth, in which case clustering methods can be utilized. Note that MAS allows for gating and clustering to be used interspersed within the same dataset. For example, one can define T cells using CD45 and CD3 expression, then employ clustering to define additional T cell subsets using additional markers.

By default, MAS will show the "Gate" subtab upon clicking "Celltypes" in the top-level menu. The user can click on the "Cluster" subtab to perform clustering. The functionality of both tabs is detailed below.

6.2 Gating Page Layout

The Gating page (Figure 40) contains sample names, the CellType Hierarchy tree, and plots for each sample that will appear with the subsequent gating that will be applied based on known cell lineage markers.

6.3 Defining Populations

To begin defining populations using gating, hover the cursor over "Singlets" from the CellType Hierarchy tree and select "Add plot" or click on the "Add Plot" button at the top of the page:

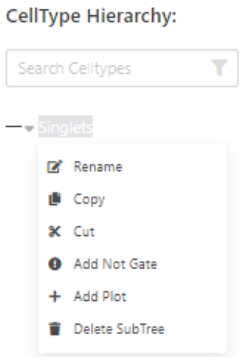

Figure 41. Singlets drop-down menu.

This will display the following window:

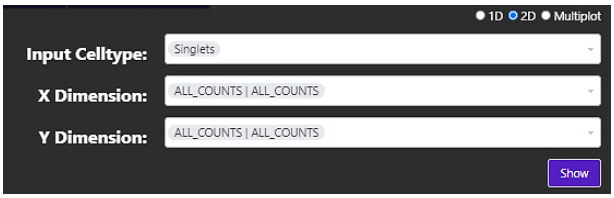

Figure 42. Creating a new plot.

Users can select one of three options for creating plots: 1D, 2D, or Multiplot. By default, the Input Celltype will be set to singlets. Once cell types are defined, those populations can be used as the "Input Celltype" when creating new gates.

6.4 1D Plots, 2D Plots, Multiplot, and Not Gates

1D plots

Select the 1D radio button:

Figure 43. Selecting a 1D plot.

Select an antibody or gene from the "X Dimension" drop-down menu. Users can search the menu by typing the name of the marker of interest:

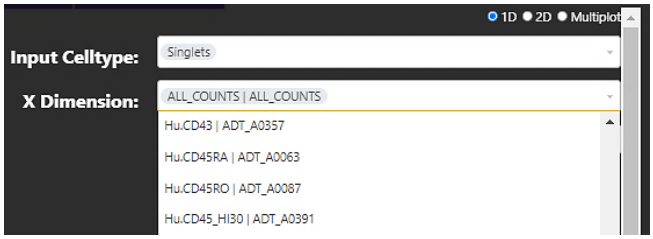

Figure 44. Defining the X dimension.

After selecting an antibody to display, click "Show" to display the plot:

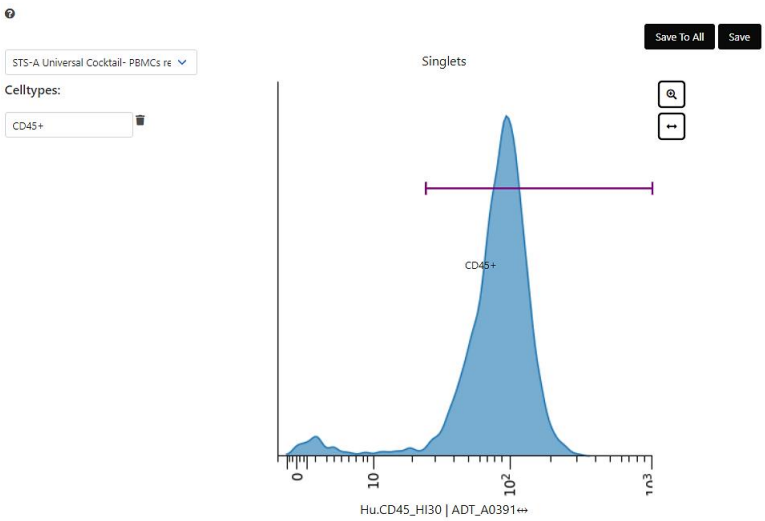

Figure 45. 1D plot example.

To draw a gate on the 1D plot, click  . Once a gate is drawn, a differential expression table will appear on the left-hand side of the window. This table shows the relative log fold-induction of each antibody for the cells captured inside a highlighted gate over cells not included in the gate. This table is useful for determining the expression levels of cells within the gate compared to all other cells. Gate names are editable, and it is recommended to provide each gate with a descriptive name.

When finished, click “Save To All” to save this gate to all samples or “Save” to save to only that sample.

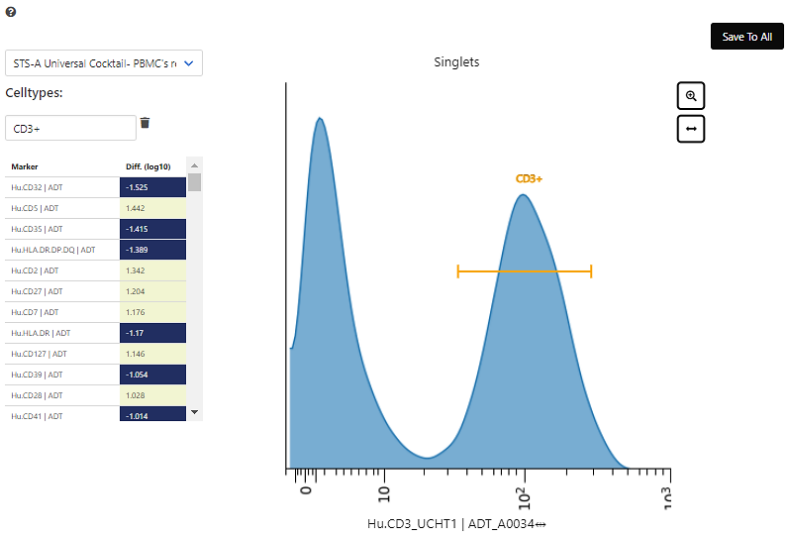

Figure 46. 1D plot: differential expression preview.

2D Plots

Select the 2D radio button and choose antibody markers for each dimension:

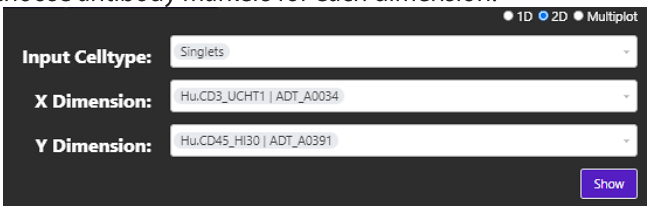

Figure 47. Selecting a 2D plot and defining X and Y dimensions.

When finished making the selections, click “Show” to display the plot:

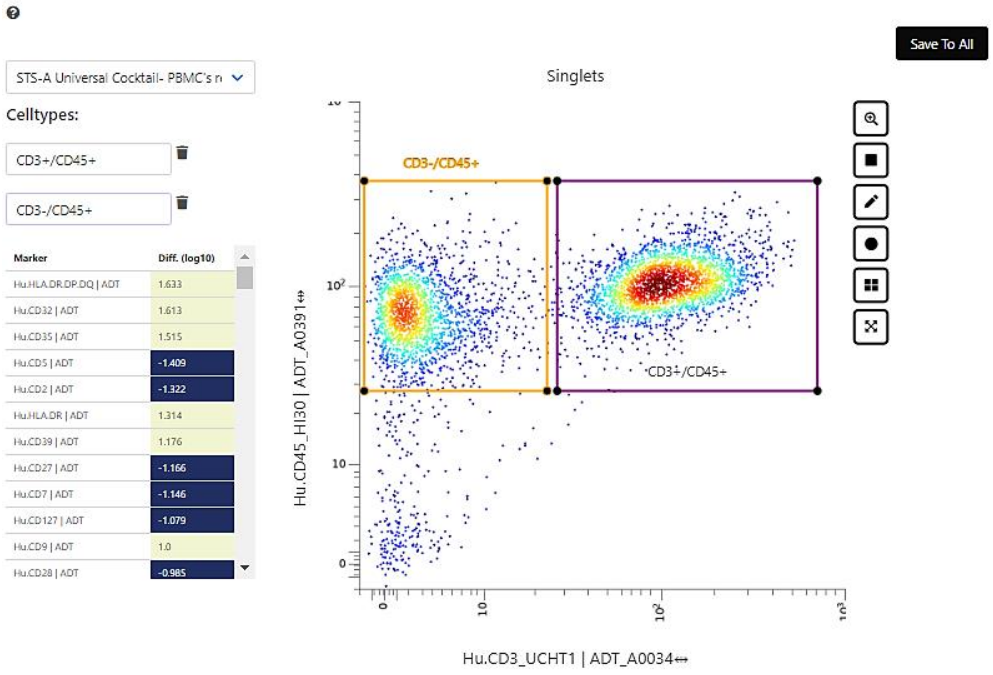

Figure 48. 2D plot: differential expression preview.

2D plots provide more options for drawing gates. To draw a gate, select one of the gate drawing options on the right-hand side of the plot. The available gate types are:

- Rectangle
- Polygon Free Draw
- Ellipse
- Quad
- Orthogonal Quad

Differential expression metrics are only displayed for Rectangle, Polygon Free Draw, and Ellipse gates. If multiple gates are included in a plot, select a gate to show the respective differential expression table. Click "Save To All" to finish.

Multiplot

Multiplot allows the user to view the expression of single antibodies among previously gating-defined cell populations. Select the Multiplot radio button, Input Celltypes, and an antibody from the X Dimension to display:

Figure 49. Selecting multiplot.

Click "Show" to display the plot:

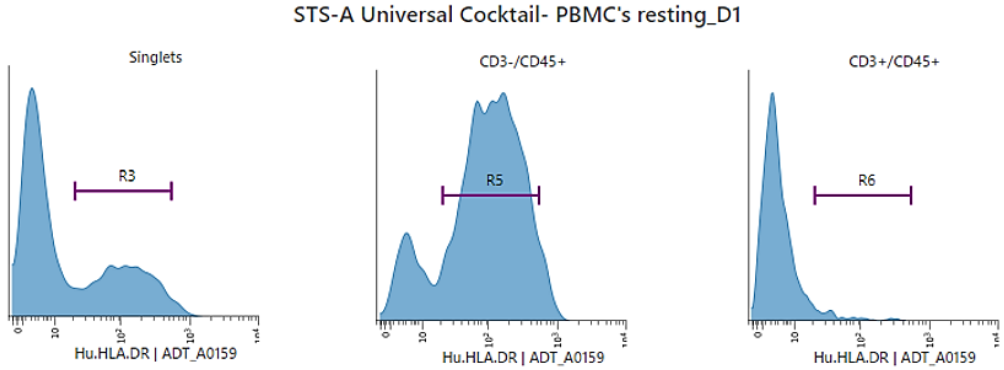

Figure 50. Multiplot example.

In the example shown in Figure 50, we used gating to identify CD3+CD45+ (T cells) and CD3-CD45+ cell populations. Using the Multiplot tool, we can visualize the expression of HLA-DR among both cell populations. HLA-DR is an MHC II cell surface receptor that acts as a ligand for TCR expressed on T cells.

Not Gates

Once gates are defined, Users can also select "Not Gates" (Figure 41) by hovering over the population in the hierarchy tree. This allows the users to create a new gate containing all of the cells not positive for the respective gate selected.

6.5 Additional Features Under Gating

Users are able to adjust plot parameters by clicking on the respective plot axis. This will display a pop-up menu with various options, as seen in figure 51. Changing the parameter values may be useful to adjust how the plot is visualized. The transformation of the data can also be adjusted to arcsinh, log or linear methods.

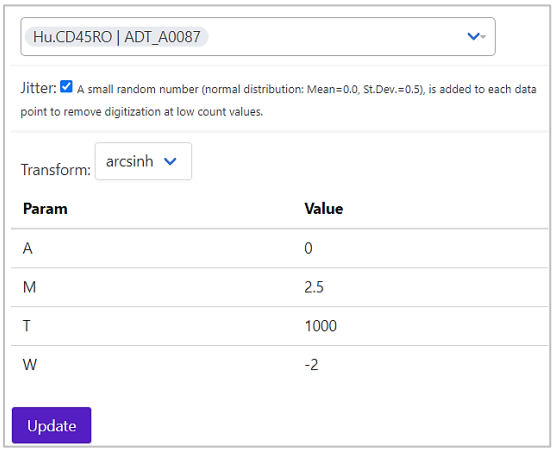

Figure 51. Adjusting plot parameters.

On the left-hand side in the Gate tab, cell type hierarchy is shown along with a box-and-whisker plot of the cell frequencies for each time across the samples. Depending on the dataset, this plot may provide a convenient mechanism to identify any outlier samples.

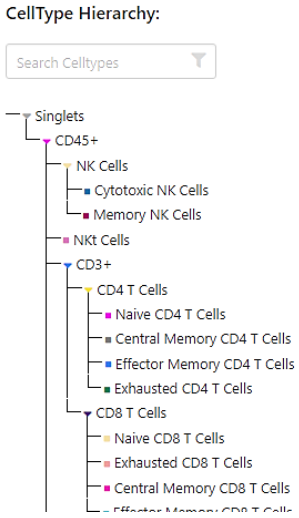

Figure 52. Cell type hierarchy.

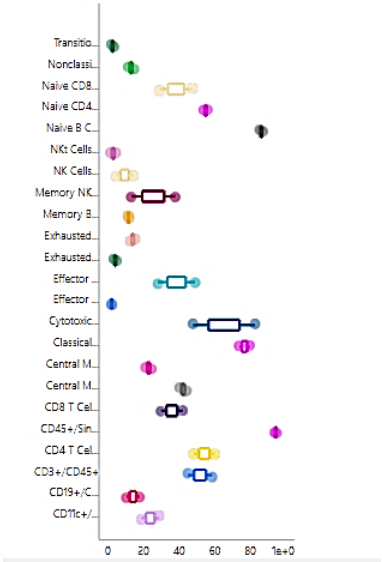

Figure 53. Box and whisker plots.

Gating plots are by default grouped by Sample. This grouping can be changed by setting "View Plots By" to "Plot" as seen in Figure 54. This will group all plots of the same type across the sample allowing for easier visual comparison across samples for the same gating markers and gates.

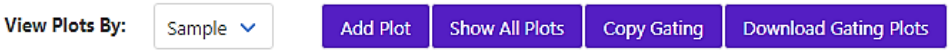

Figure 54. Plot view and options.

Users can also copy gating from a sample from a previous dataset by clicking the "Copy Gating" button and searching for a sample from a previous dataset as seen in Figure 55. Select the appropriate experiment under Source Sample. Click "Apply Gating" once the appropriate samples have been selected. Note that copying will only be successful if all the necessary gating markers (in the case of ADTs, also tagged to the same oligonucleotide tags) are present in the current dataset.

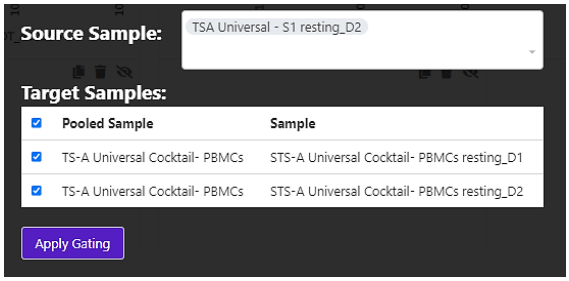

Figure 55. Copying gating to other experiments.

The gating plots generated can be downloaded by clicking the "Download Gating Plots" button. MAS will generate a PowerPoint containing images of the cell hierarchy and gating plots as seen in Figure 56.

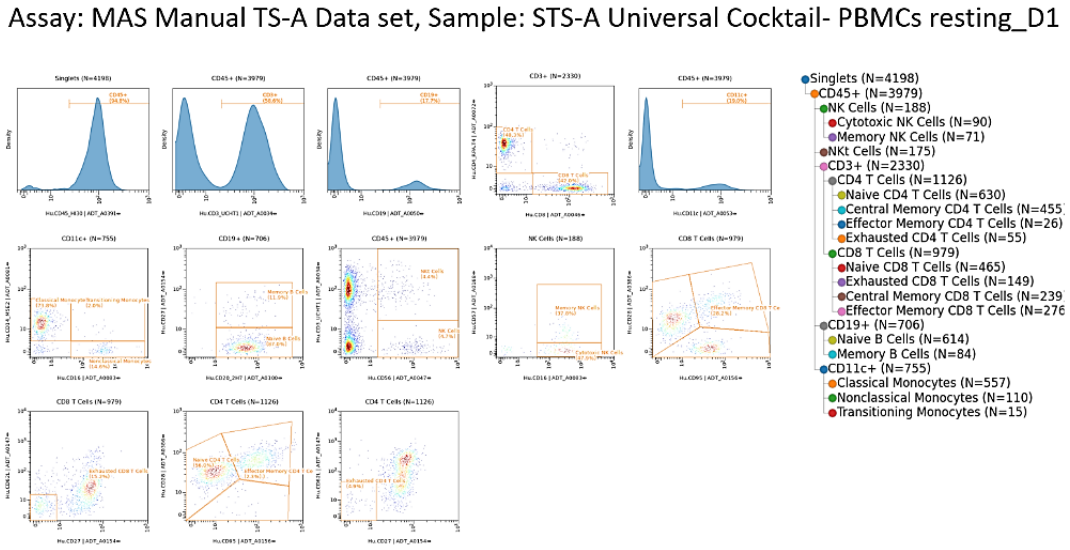

Figure 56. Downloaded gating plots example.

6.6 Clustering

In addition to gating, MAS also allows cell populations to be defined using clustering methods. Clustering of cells can aid in understanding the cellular heterogeneity within datasets and can be useful when cell lineage is not clear. Clustering can be applied to all or a subset of cells (defined by gating) using either PARC or Leiden community detection algorithms.

To explore the results from clustering, it is usually convenient to have a two-dimensional projection like t-SNE or UMAP ready to visualize top-ranking protein or RNA markers for each cluster. While it is possible to run clustering without creating projections, it is recommended that a projection for the cell type that will be clustered be created prior to the remaining steps in this section. Please see the next section for details relating to creating a projection.

To access the clustering page, click on “Cluster” within the Celltypes tab as seen in Figure 57.

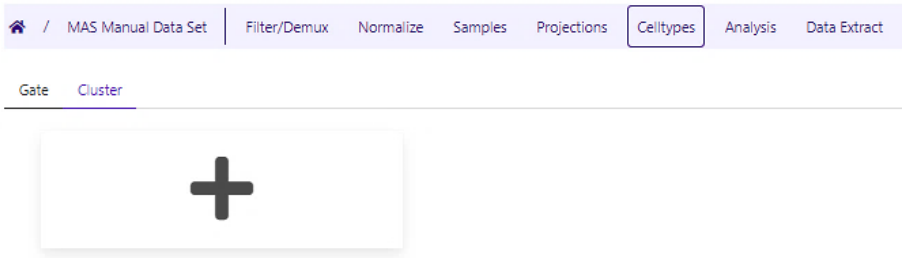

Figure 57. Accessing cluster subtab.

Users will be directed to the clustering page. Click the “+” icon to display the following page:

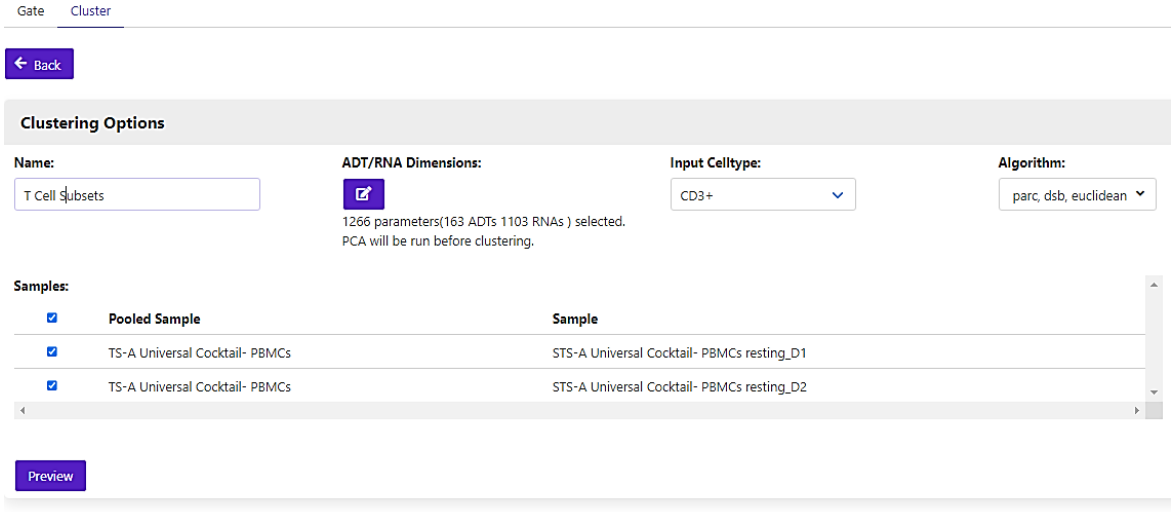

Figure 58. Defining a cluster.

Follow the steps below to set up the clustering process:

1. Provide a name for the cluster.
2. Select ADT/RNA dimensions to use as inputs for clustering by clicking on the button. The following pop-up screen will appear showing the variance for ADT markers sorted by total count (Figure 59).

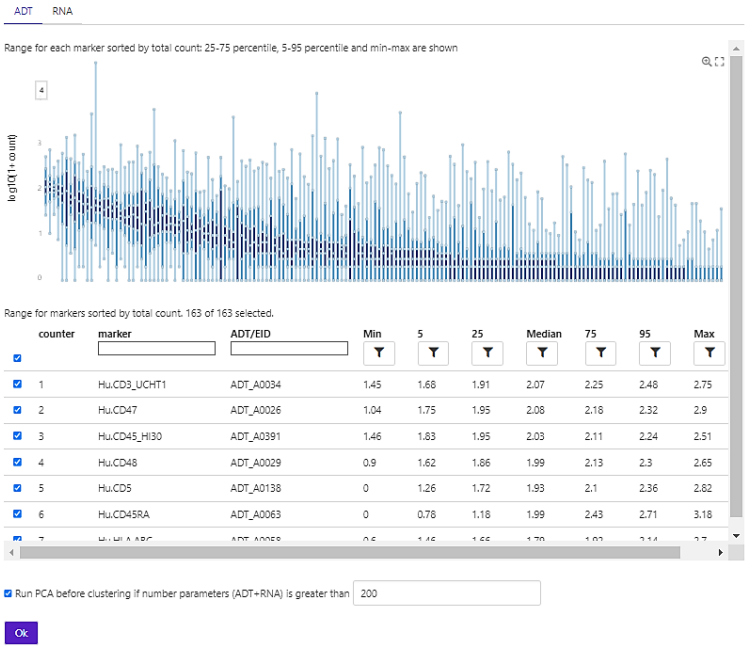

Figure 59. Selecting ADT dimensions.

Users can toggle to HVG RNA dimensions by clicking the RNA tab at the top of the pop-up window as seen in Figure 60.

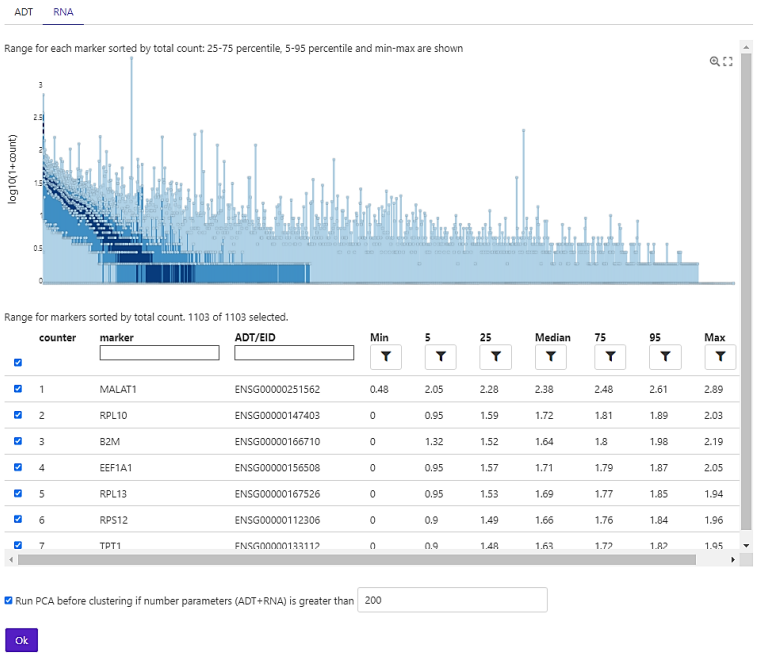

Figure 60. Selecting HVG dimensions.

- Any combination of ADT or HVGs can be used as inputs to clustering.
- The filters in the table(s) below the plot can be used to sub-select a subset of dimensions to include as inputs to the clustering. By default, all dimensions are selected.
- Make sure to deselect any isotype control antibodies included for ADT-based clustering.
- When the selected dimensions consist of only ADTs, the data is normalized by the background expression in the negative events (DSB normalization). If RNA dimensions are included in the selection, no additional normalization is performed other than log transforming the data.

3. Select the samples to cluster by clicking on the checkboxes next to the rows as seen in Figure 58.
4. Select the population of interest (input cell type).

IMPORTANT

» We recommend selecting populations that have had cells positive for mutually exclusive markers be removed.

5. Select an algorithm to use for clustering and make selections for clustering parameters. It is recommended to use PARC by default. The algorithm drop-down menu provides users with the ability to modify the following parameters under "Show Advanced Options":
   - ***Scale:*** DSB or Zscore can be selected as scaling options.
   - ***Distance Metric:*** Controls how distance is computed in the multi-dimensional space of the input data. By default, Euclidean is selected, but users can select Manhattan as an alternate distance metric if they expect any outlier cellular events.
   - ***Sample Size:*** The default sample size is set to 100,000.
   - ***Num Neighbors:*** The number of points in a neighborhood. Decreasing this number may reduce the total number of clusters created. Having said that, a larger number of clusters is not guaranteed. PARC in particular is designed to use this number to start the algorithm, but to a large extent, let the structure of the data determine the number of clusters. It is recommended to leave this value at 30.
   - ***Num Predicted Neighbors:*** This parameter comes into play only if the cells are sampled because there are more than 100,000 events across the samples. The number of neighbors is used when predicting the cluster membership when clustering algorithms are applied to individual samples. The cells are first pooled, sampled if the total count exceeds the sample size, and then the clustering algorithm is applied. The result is then used to reapply to the individual samples. The cluster ID for a cell is based on the closest neighbors.

- ***Resolution:*** Decreasing this number may result in fewer clusters. This parameter controls how tight a set of cells are to each other within a cluster.
- ***Small pop:*** The number of cells allowed in the smallest cluster. We recommend leaving this parameter at 30 unless some rare cell types are expected.

6. After making the selections, click on "Preview" to run the clustering algorithms and review the results. The preview summarizes the results from clustering allowing for close examination of the characteristics of each cluster, renaming, and merging before applying to the samples to create cell types. Figure 61 provides an overview of each of the elements of the clustering preview.

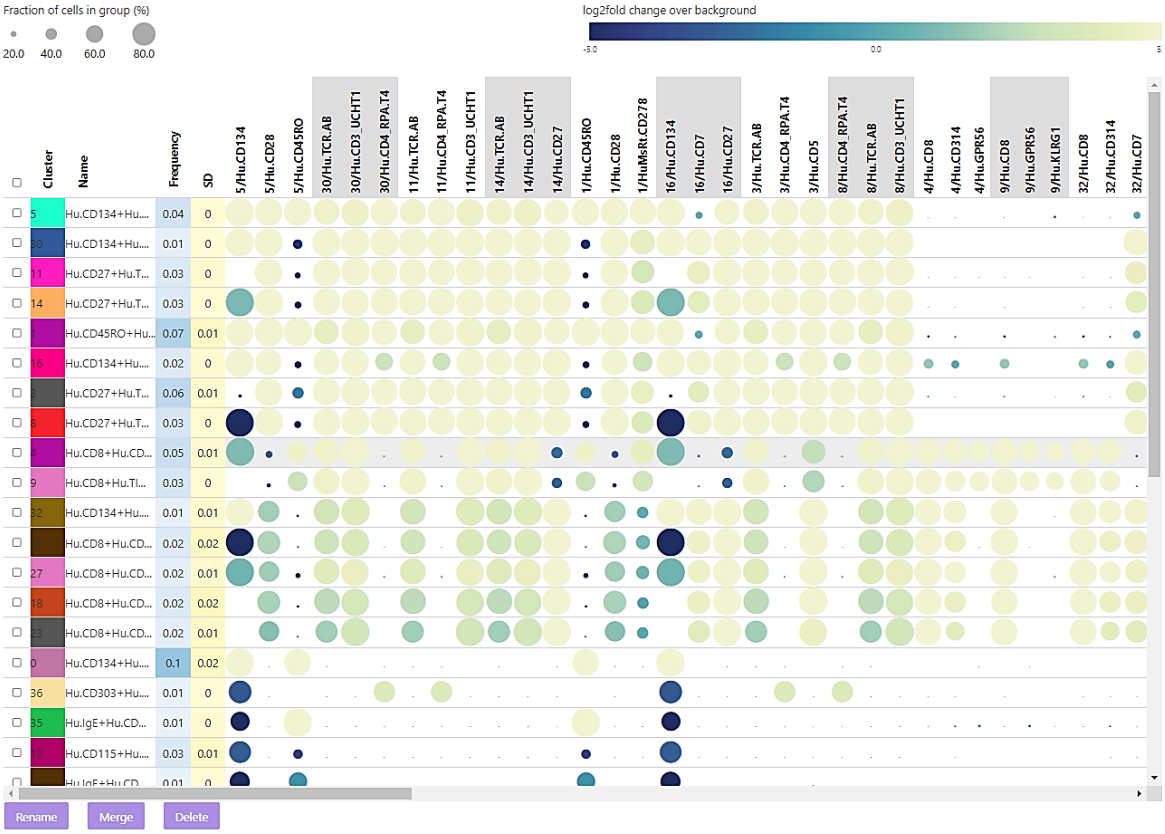

Figure 61. Cluster results preview.

For each cluster (i.e., each row), the cluster preview table provides an auto-generated name based on the top-ranking markers for that cluster along with the fraction of cells belonging to that cluster and the standard deviation of that frequency across the samples that have been pooled, followed by a summary of the top three genes per cluster via bubble plot. The color of a bubble is indicative of the differential expression (log fold change) of that marker in that cluster vs. all other clusters. The size of the bubble is indicative of the fraction of cells in that cluster that are positive for that marker. In the case of protein (ADT), positivity is measured above the background level, while in the case of RNA, positivity is simply the fraction of cells that have a positive UMI value in that cluster for that marker.

Clicking on a row will display the top 10 differentially expressed markers for that cluster. If a projection has been predefined, a feature plot with the projection colored by the marker expression is displayed for each of the top markers. Users can toggle through predefined projections below "Projection to Display".

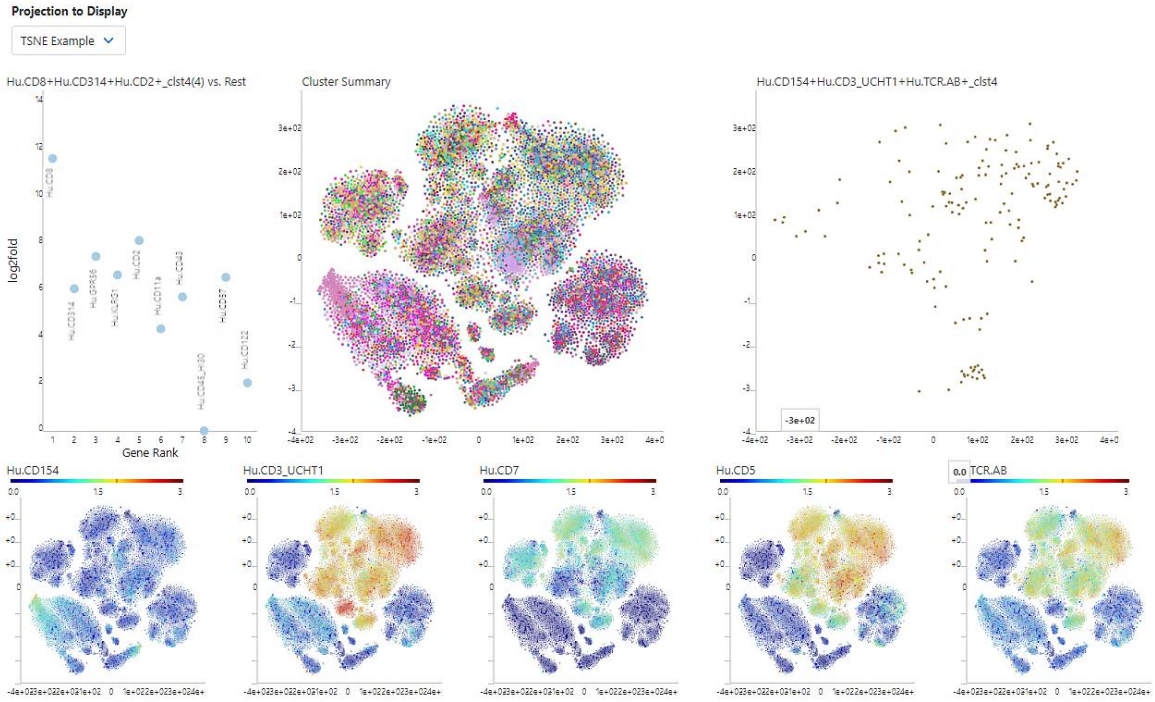

Figure 62. Cluster displayed on pre-defined projection.

If the automated naming for a cluster is not appropriate or a more conventional name is desired, click on the checkbox next to a cluster and click "Rename". Type a new name for the cluster and save.

If two clusters have similar patterns for the top-ranking markers and are therefore deemed to be similar to each other, they can be merged into a single cluster. Click on the checkbox next to the two (or more) clusters, click on merge, and provide a new cluster name to merge the clusters.

Once clustering is complete, click save. Depending on the workflow, Users can proceed to analysis or create projections if that step has not been completed.

7. Projections/Dimensionality Reduction Plots

Dimensionality reduction algorithms take in highly dimensional protein data and reduce it to two dimensions. This results in similar cells being grouped together and allows for increased ease of analysis. Dimensionality reduction plots act as a template with which to generate interactive analysis plots used in subsequent sections of this guide. After generating plots, users can cluster cells using community detection algorithms and project those communities (or clusters) onto dimensionality reduction plots created in this section.

7.1 Creating a Projection

From the top-level menu, click "Projections":

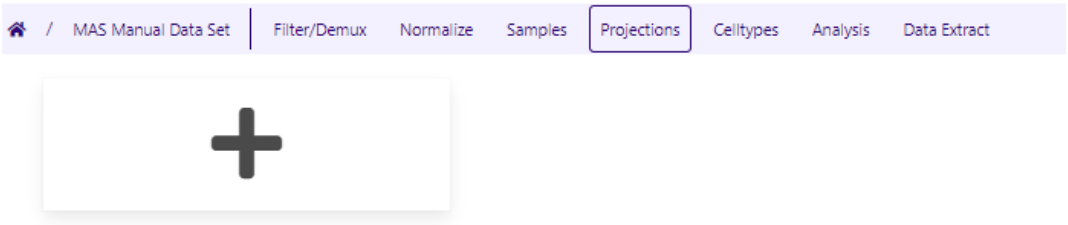

Figure 63. Accessing projections tab.

To create a UMAP, t-SNE or TriMap, click the "+" icon to display the following page:

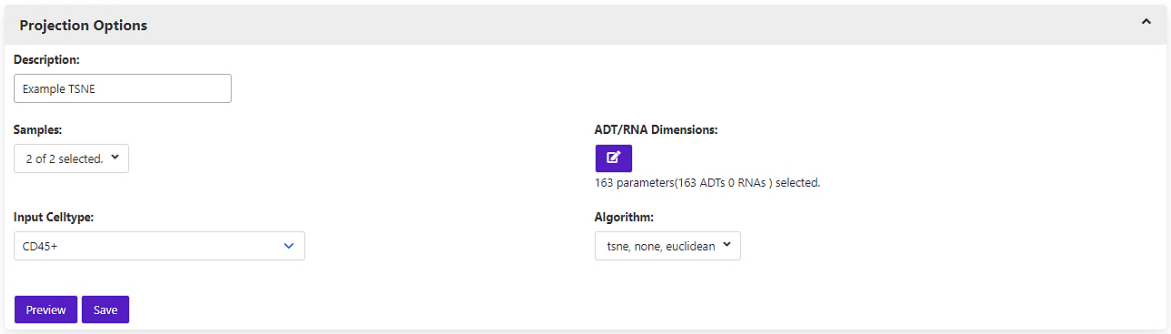

Figure 64. Setting up a new projection.

Follow the steps below to set up a dimensionality reduction plot:

1. Provide a description of the plot.
2. Select the samples to plot.
3. Select the population of interest (input cell type).

IMPORTANT

» We recommend selecting populations that have had cells positive for mutually exclusive markers removed.

4. Select ADT/RNA dimensions to use as inputs for clustering by clicking on the button. The following pop-up screen will appear showing the variance for ADT markers sorted by total count (Figure 65).

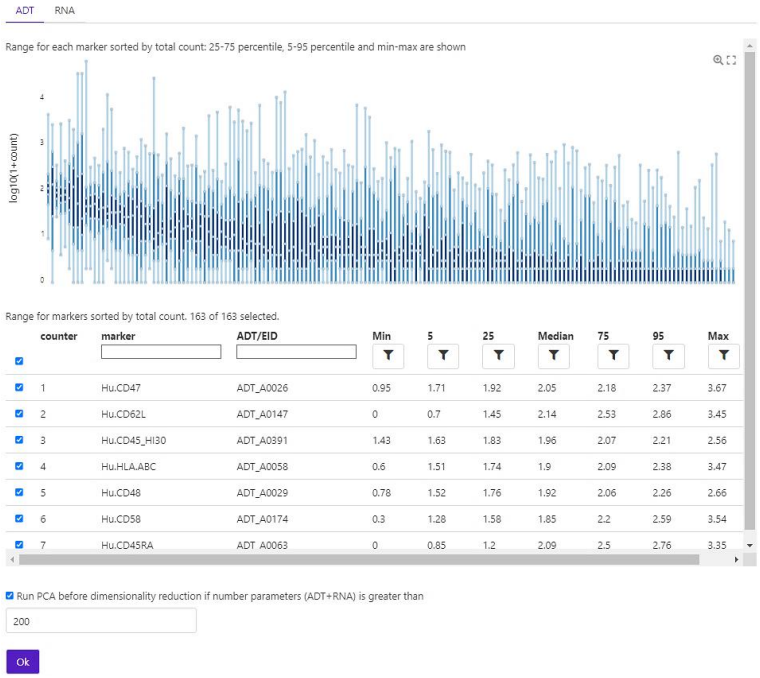

Figure 65. Defining ADT dimensions for projection.

Users can toggle to HVG RNA dimensions by clicking the RNA tab at the top of the pop-up window as seen in Figure 66. Users can select ADT or RNA markers, but it is recommended not to use both for the same projection. A common method is to perform ADT-based clustering by selecting relevant ADT markers and deselecting all RNA and any isotype control antibodies.

Figure 66. Defining HVG dimensions for projection.

IMPORTANT

» By default, all RNA and ADTs are selected. Since it is not recommended to use both RNA and ADT for the same projection, make sure to unselect all RNA or ADTs depending on the projection inputs (e.g., if using ADTs for the projection, unselect all RNA from the RNA dimensions tab). It is also recommended to deselect any isotype controls from the ADT parameter tab for ADT based projections.

5. Once the appropriate markers are selected, click "Ok".
6. Select a plot type under "Algorithm". Users can select between a UMAP, t-SNE, or TriMap. The algorithm drop-down menu provides users with the ability to modify the following parameters under "Show Advanced Options" - the most important tunable parameters (depending on the method) are marked by a "*" below.
   - ***Scale:*** None, DSB, and Zscore can be selected as scaling options. By default, UMAPs will be set to none.
   - ***Distance Metric:*** Controls how distance is computed in the ambient space of the input data. By default, Euclidean is selected, but users can select Manhattan as an alternate distance metric.
   - ***Num Neighbors*:*** This parameter controls how UMAP balances local versus global structure in the data. Higher values help resolve the global structure of the original data, but lose the local data structure.
   - ***Num Predicted Neighbors:*** This parameter comes into play when the cells were down-sampled because the pooled dataset is larger than the sample size of 500,000. If sub-sampling was applied, projection is first computed on the sub-sample, and then for the remaining cells, the locations in the projection space are interpolated based on the nearest neighbors. The number of neighbors used for the interpolation equals this parameter value.
   - ***Min. Distance*:*** The minimum distance between points in the final projection. This can be adjusted to control how loosely data points are packed.
   - ***Num Inliers (Trimap):*** Number of inliers included in projection.
   - ***Perplexity* (t-SNE):*** This is similar in concept to the number of neighbors to the defined local neighborhood and the most important parameter for t-SNE methods.
7. Once the projection options have been completed, click "Preview" to preview the projection. This may take a few minutes, but once complete, a preview of the plot will appear below.

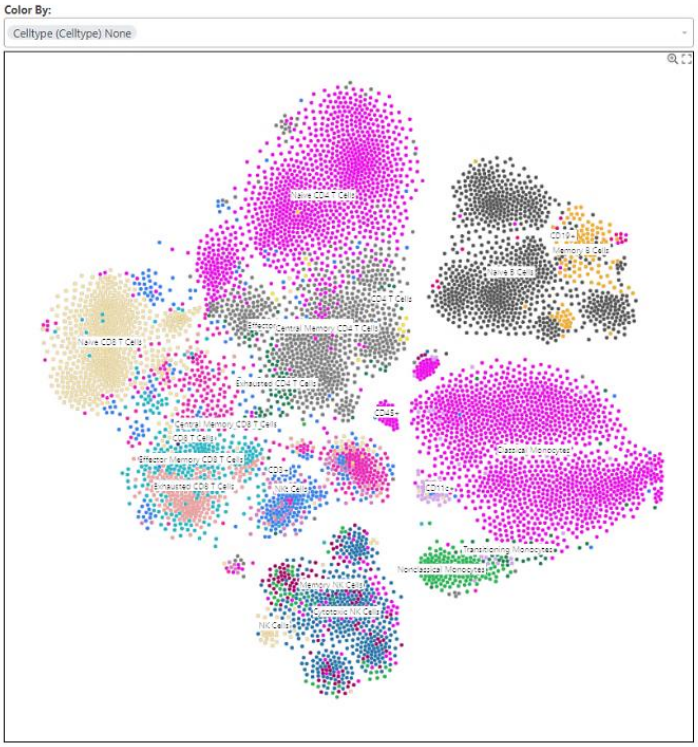

Figure 67. Preview of projection: example t-SNE.

8. Users can select the ADT expression to be highlighted by selecting the "Color by" option at the top of the plot. This will also generate an ADT expression plot on the top right for the selected marker:

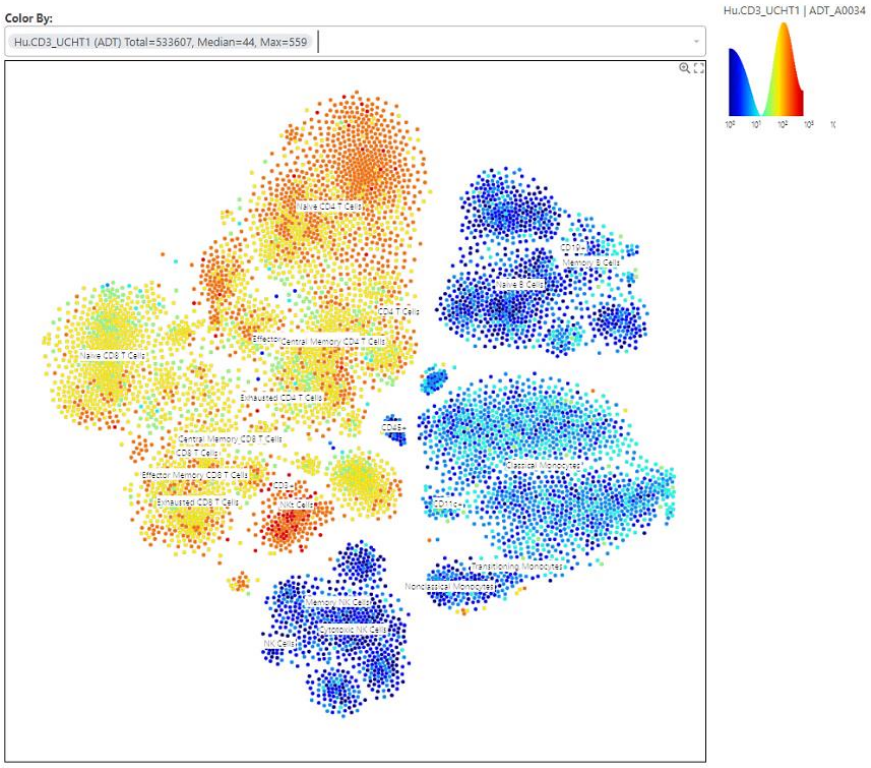

Figure 68. t-SNE colored by CD3 ADT expression example.

9. Once the projection is complete, click "Save".

7.2 Tuning a Projection

When creating a projection, there are a few important tunable parameters that will dictate how the projection looks and how accurately the data is represented. For a UMAP or t-SNE, tuning the Num Neighbors (1-200) or perplexity (1-100) changes the balance between local and global aspects of the data. Another tunable feature of a UMAP is Min. Distance (ranging from 0.0 - .99) which for simplicity's sake, resembles how close or far the dots/cells are in relation to each other.

Users should adjust these parameters until they find a projection that aligns with the biology of the sample(s). Usually, different iterations of the projection are needed to settle on one that more accurately represents the data. For example, we would expect similar cells to cluster near each other and to have enough resolution in the visualization to see the cells forming discernable clusters.

8. Cell Type and Cross Sample Analysis

Once cell types have been defined and the desired projections have been computed, the Analysis tab can be used to conduct various modes of interactive analysis. This includes interacting with the data to explore expression patterns, computing and visualizing differential expression patterns (either between cell types or samples), and/or creating single-cell overlay plots to visually highlight expression patterns.

Currently, MAS provides two types of analyses. The first, cell type analysis, is for visualizing single-cell data using one of the projections for each sample and for comparing and contrasting protein or RNA expression across cell types within each sample. The second, cross-sample, analysis is for exploring protein/RNA expression in various cell types that are associated with sample endpoints; for example, which protein or RNA markers are differentially expressed in which cell type between two subject groups like responders and non-responders.

8.1 Celltype Analysis

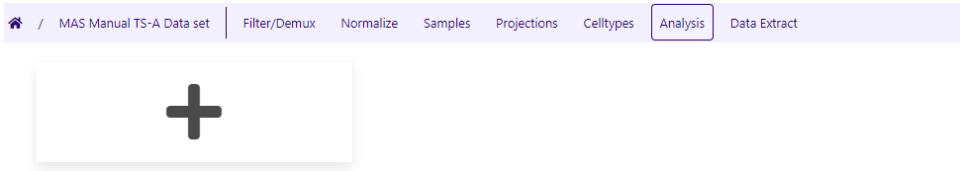

Figure 69. Analysis tab.

1. Within the Analysis tab, click on the "+" icon to display the following page:

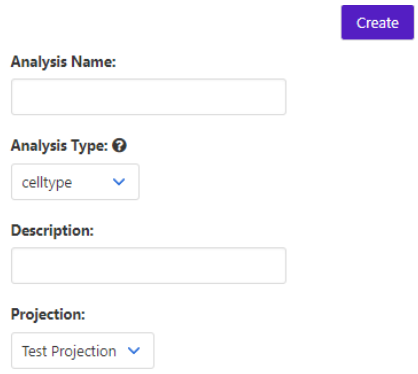

Figure 70. Defining the analysis and selecting the projection.

2. Name the analysis and select either celltype or crosssample under the Analysis Type option. The celltype analysis can be used for interactive exploration of the data using UMAP or t-SNE projections.
3. Provide a brief description of the analysis and select the appropriate projection to use. Note that only previously defined projections can be used as the input.
4. Once complete click "Create".
5. MAS will generate a dashboard (Figure 71) that provides a comprehensive view of all aspects of the dataset.

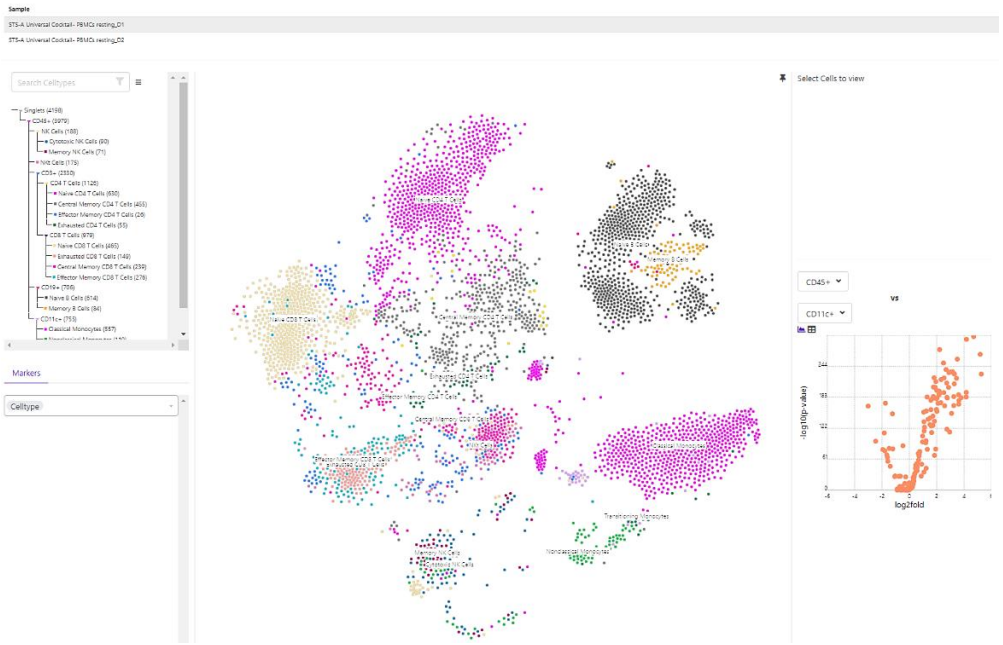

Figure 71. Analysis dashboard.

6. At the top of the dashboard, users can select which sample to view by clicking on the row with the respective sample name.
7. Users can search and select for previously defined CellTypes using the cell type tree found on the top left of the dashboard. Clicking on the populations in the tree will highlight that population in the projection plot. Note that multiple populations can be highlighted simultaneously. To unhighlight, unselect the population. Users can also add/remove annotations or change population colors by clicking on the ≡ button in Celltypes Hierarchy.
8. The bottom left Markers table (Figure 72) allows the user to select ADT or RNA markers, and color the cells in the projection by their respective expression of that marker. Click the thumbnail on the top left to save the projection on the bottom Analysis dashboard.

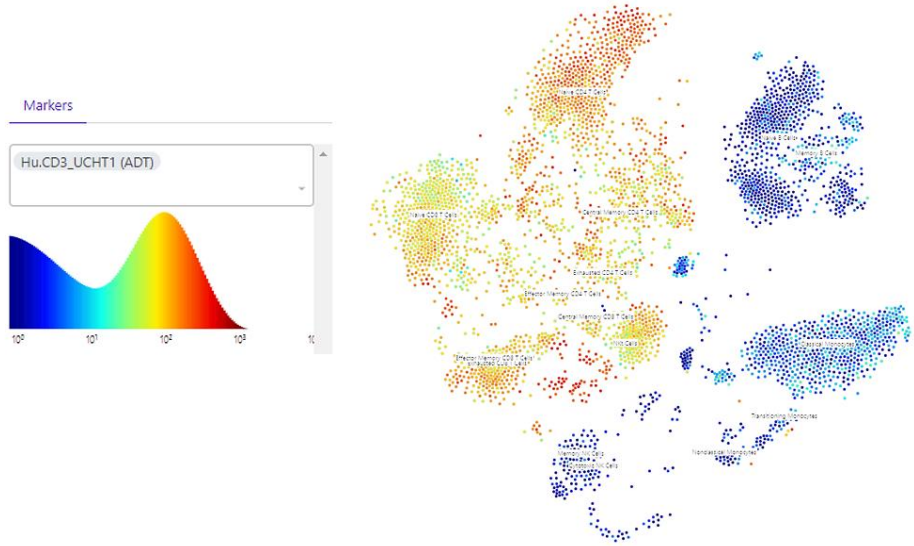

Figure 72. Highlighting markers on a projection.

9. Within the projection, users can click on the plot and create a gate around a population of interest. The table to the top right will display a differential analysis of the selected population as seen in Figure 73.

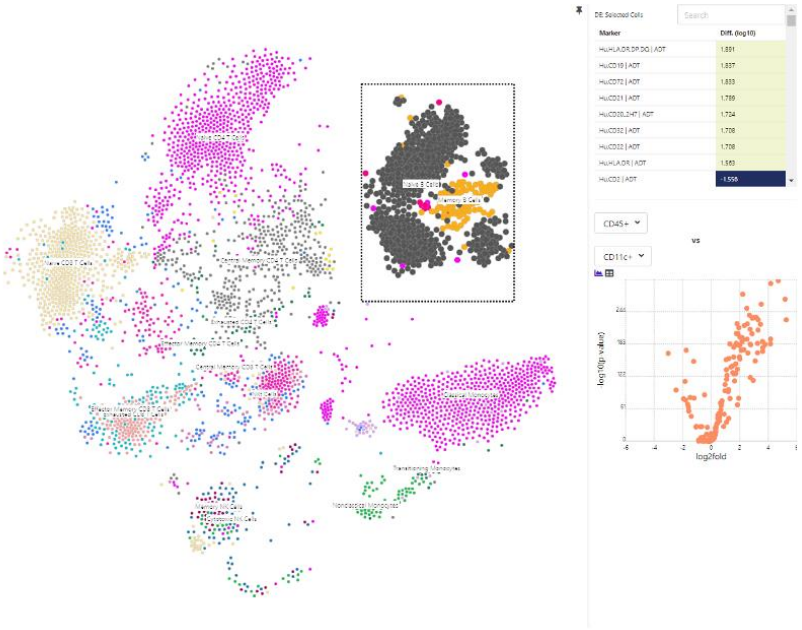

Figure 73. Selecting a subpopulation and summary statistics.

10. At the bottom right corner of the dashboard (Figure 73), users can select previously defined populations and conduct differential analysis between those cell types. MAS will display a volcano plot which plots statistical significance (p-value) against the magnitude of change (fold change). Typically, upregulated genes are towards the right while downregulated genes are towards the left. The most statistically significant genes are found at the top of the plot. Clicking on the button (Figure 74) will change the view from a volcano plot to a table.

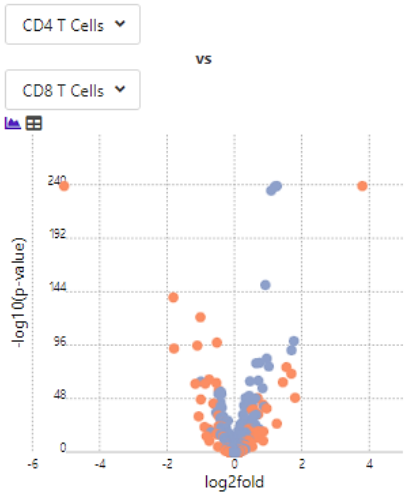

Figure 74. Volcano plot.

8.2 Editing Color Schemes

In the Analysis tab, users can edit the coloring scheme for the projections by clicking on the setting button at the top menu bar as seen in Figure 75. Users have the option to edit the color schemes used for Celltypes, Markers, or Samples. To change colors, select the population of interest and adjust the color. Once complete, click “Save”.

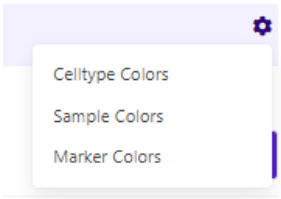

Figure 75. Color adjustment options.

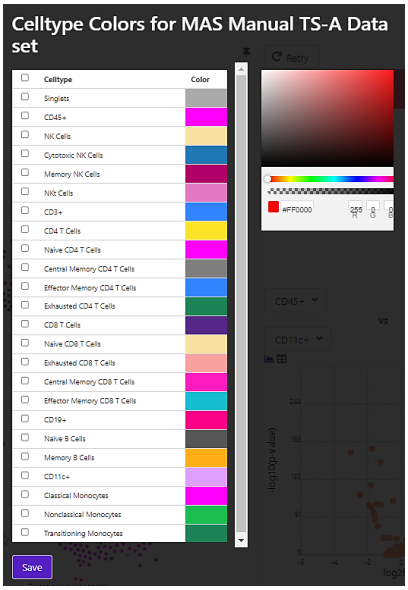

Figure 76. Changing Celltype colors.

8.3 Cross Sample Analysis

As mentioned above, cross sample analysis is used to explore protein or RNA dimensions that may be associated with sample endpoints or characteristics in various cell types. Unlike cell type analysis, the focus here is not to explore single-cell data via projections, but to analyze aggregate measurements like median values for each protein or RNA in each cell type and compare these values to sample characteristics. For example, the samples may be split into multiple treatment or response groups; or may have been obtained before and after treatment from each subject. On the cross-sample analysis page, one can specify the endpoint and related parameters and prompt MAS to compare all protein and RNA dimensions in each cell type to the specified endpoint.

1. To create a cross sample analysis, click on the "+" icon in the Analysis tab.
2. Provide an analysis name, select "crosssample" under analysis type, provide a description, and click on create.

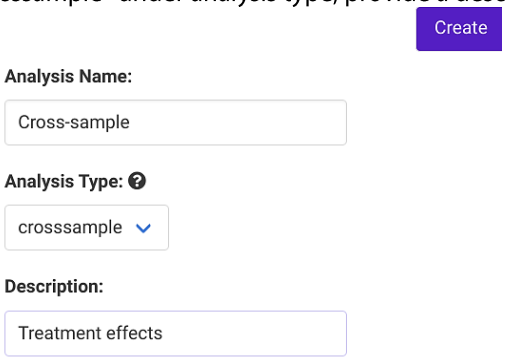

Figure 77. Cross sample analysis.

3. Select the dependent variable, i.e. the endpoint to which protein and RNA data will be compared against.
    a. In this example, the six samples fall into three treatment buckets. "Treatment" is selected as the dependent variable. In a different case, the samples may be paired; for example, before and after treatment samples from the same subject. In this case, protein/RNA levels can be compared between pre- and post- sample

paired on the subject. Pairing variable, if it exists can be selected under “Pairing”. If the comparisons should be done independently in subsets of samples, the variable that should be used to sub-group the sample can be selected under “Grouping”. Currently, only median values for each marker in each cell type are assessed. Finally, the DE Threshold value will be used to keep only dimensions in each cell type for which the difference between groups being compared is greater than this value.

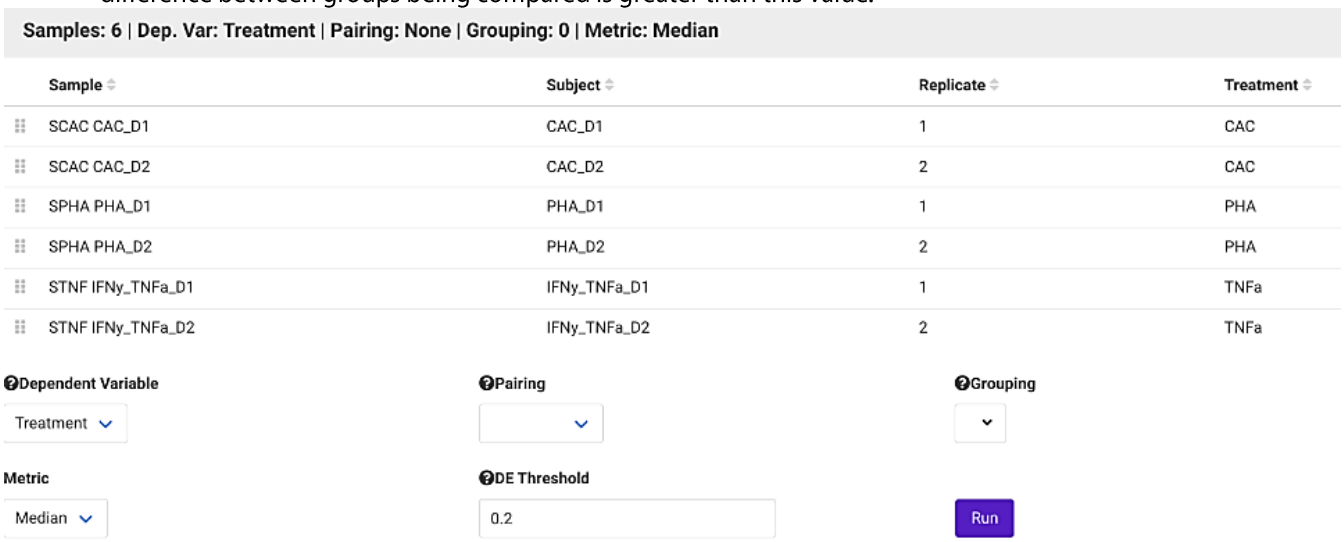

Figure 78. Defining cross sample parameters.

4. Once the parameters have been set up, click “Run” - this may take several minutes.
5. The results are displayed on a plot of effect size vs. -$\log_{10}$(p-value); when comparing two groups this is commonly known as the volcano plot. Each point on this plot corresponds to a protein or RNA marker in a cell type. The points are colored by the cell type. The x-axis is the differential expression between the groups when there are two groups. In the case of multiple groups (or a numerical dependent variable), it is the standard deviation across all samples. In the case of two groups, p-value is obtained from the Wilcoxon test. For multi-groups, it is obtained from one-way ANOVA, and in the case of a numerical dependent variable, it is obtained from linear regression. The results are also shown in the table on the right-hand side.

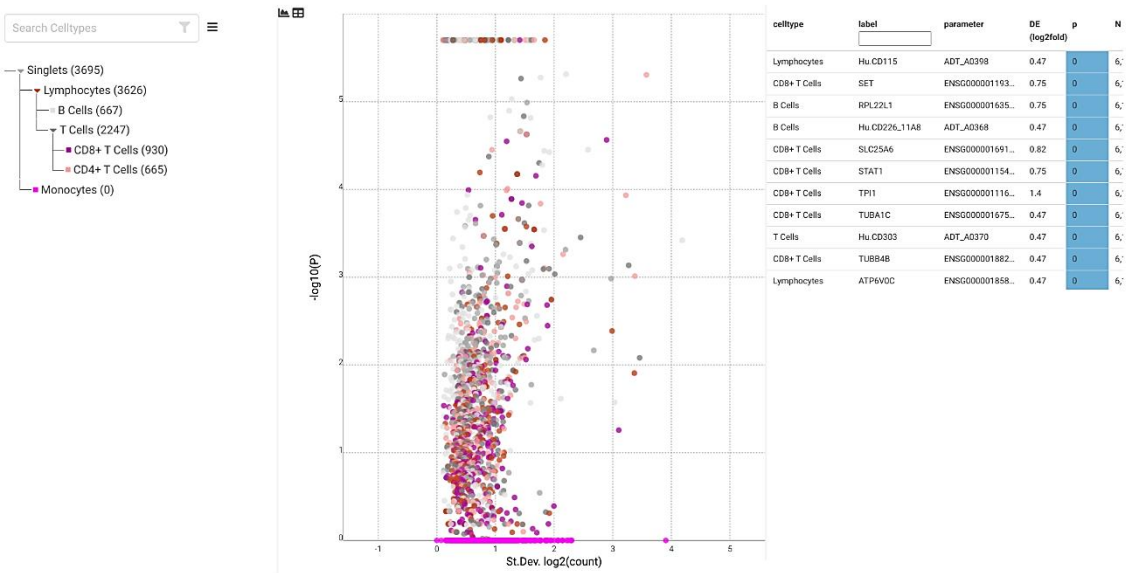

Figure 79. Cross sample analysis results.

6. By clicking on any defined population in the cell type tree on the left-hand side, all points corresponding to that cell type can be highlighted in the plot and filtered in the table to the right. In addition, an individual dimension can be

visualized by clicking on a point in the plot on a row in the table.

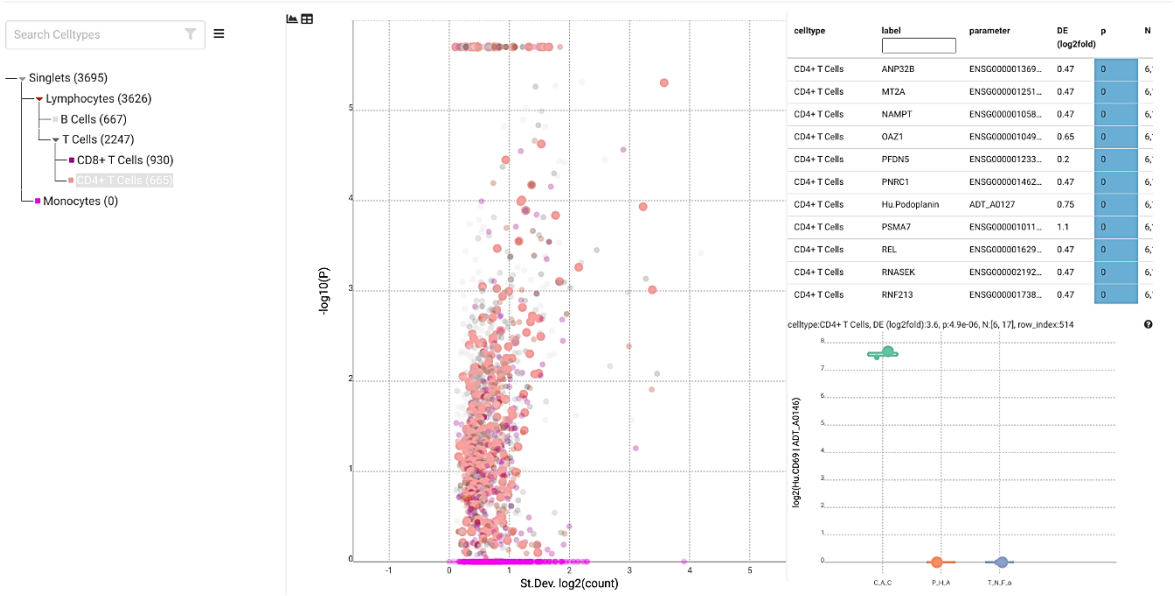

Figure 80. Highlighting populations on volcano plot.

7. Clicking on multiple points in the boxplot will display the underlying single-cell data as histograms.

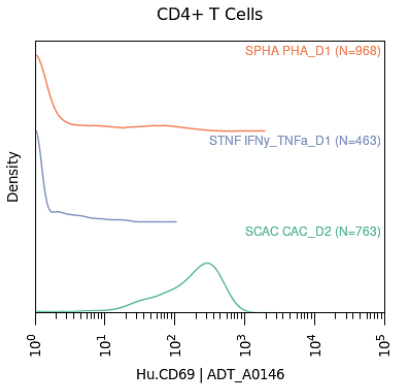

Figure 81. Histogram data in cross sample analysis.

8.4 Exporting Analysis Plots

Plots that were saved in the Analysis dashboard can be exported by clicking the download button at the top of the analysis page. Users can select which plots to included and download in a compressed folder. The folder will contain PNG file images of the exported plots.

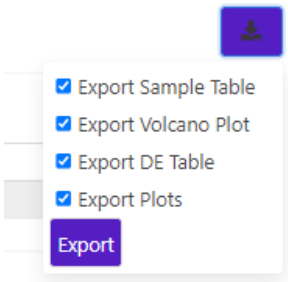

Figure 82. Exporting analysis plots.

9. Generating Reports and Data Extraction

MAS provides multiple mechanisms to export the generated plots as images (PNG files) and/or the data in a variety of formats (e.g. h5ad, csv, Loope, etc). Within the Data Extract tab, users can access the Reporting subtab to select plots to download as images, or use the Data Extract subtab to export the data.

Figure 83. Data extract and reporting.

9.1 Reporting

In the Report subtab, users can create overlay plots of protein and RNA markers across samples or cell types.

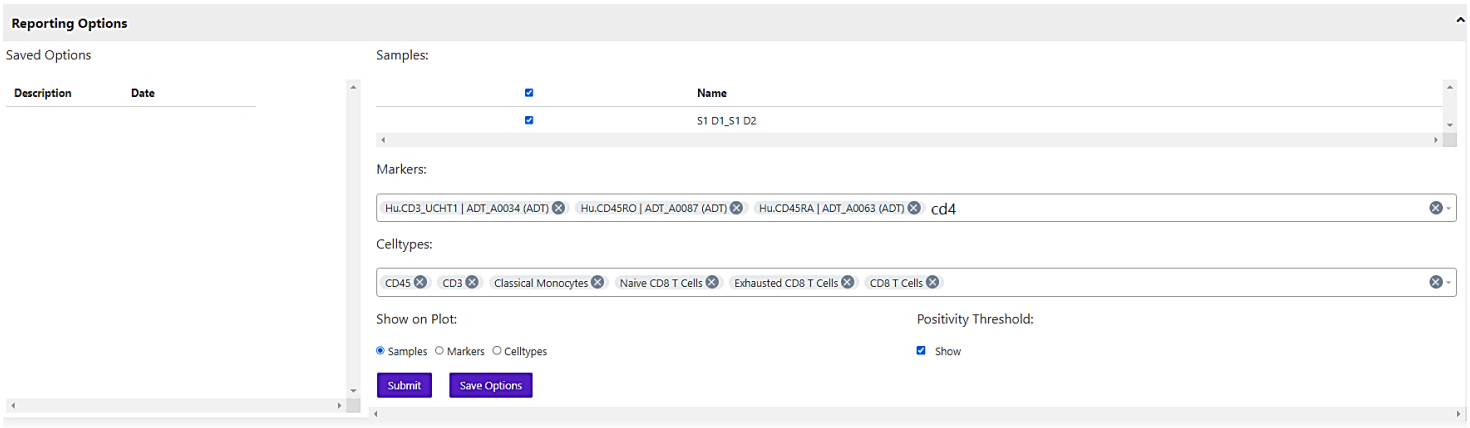

Figure 84. Reporting options.

1. To begin, first select the Samples by checking the box to the left as seen in Figure 84.
2. Then, select the appropriate Markers and Celltypes to plot by selecting them in the drop-down menu.
3. Select whether to show the Samples, Markers, or Celltypes on the plots.
4. By default, the "show positivity threshold" option will be checked. This will generate a table with the positivity values used and can be downloaded as an excel file by clicking the download button at the bottom left of the Report tab.
5. Click "Submit" to generate the report which will be displayed below.
    a. If positivity threshold was selected, the threshold will be displayed and can be adjusted by clicking on the plot(s).
6. Click "Save" to name and save the report configuration.
    a. Saved reports can be accessed via the "Saved Options" table by clicking on the name of the report.
7. Download the generated plots by clicking the download button located to the right of the plots.

9.2 Data Extract

The second sub-tab, Data Extract, allows the user to download the data in various formats.

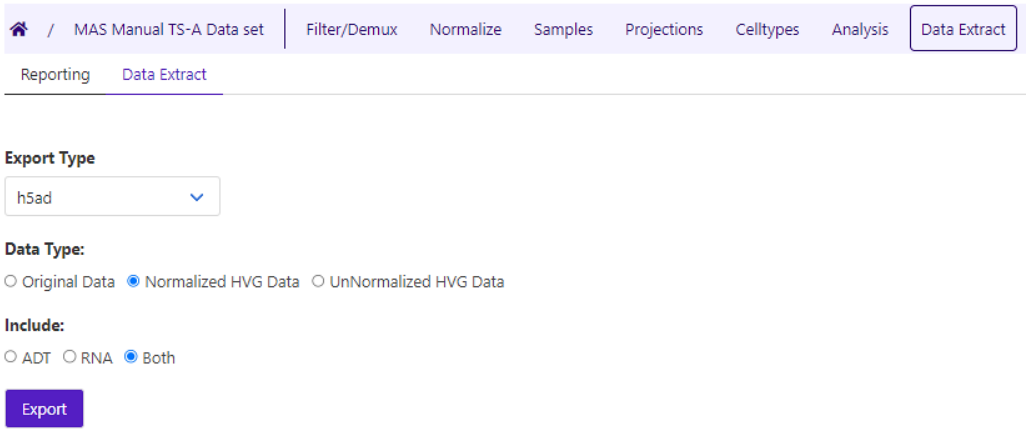

Figure 85. Data extraction.

1. To begin, select the "Export Type" to select the format of the data. Users have the option of selecting h5ad, csv, Metric data, Loupe annotation, and FCS file types as seen in Figure 86.

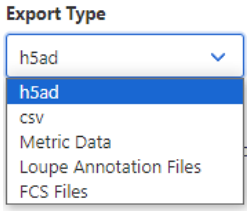

Figure 86. Data format options.

2. Select the “Data Type”. Users can pick between the original, normalized, or unnormalized HVG data.
3. Finally, select whether to include ADT, RNA or Both types of data in the export.
4. Once the selections are complete, select “Export” and an automatic download should begin.

The most common format used is usually the h5ad format. This will include the single-cell data (along with the cell type annotations) and any projections that have been created. Similarly, the csv format option will also provide the same data but in a comma-separated format. The Metric data format provides median values for protein and RNA for each cell type in each sample. This data is provided as an OData stream which can be pasted into Excel, Spotfire, tableau, R, or Python programs to connect to the data and conduct any other downstream analysis. A key advantage of using an OData link instead of a static file is that any changes to gating/cell types, in MAS for instance, will automatically be reflected in the OData stream. The downstream analysis tool (e.g. Excel or Python) only needs to refresh to obtain the latest version of the data.

Algorithm References

Debris

MAS uses a knee plot to filter cells based on UMI complexity. A knee plot is a standard quality control tool created in single-cell RNA experiments and is used to determine thresholds for determining which cells will be retained for analyses. Cells are ordered on the x-axis according to the number of UMIs detected. The y-axis displays the total number of UMIs for each cell. The threshold is an adjustable parameter. Thresholding is performed by identifying the inflection point of the plot or "knee". UMI-rich cells are found on the left-hand side of the plot and are excluded as these represent likely cell aggregates. On the right-hand side of the plot beyond the "knee" are barcodes that have relatively low numbers of reads and are therefore considered either to have failed in the capture or to be too noisy for further analysis. See https://liorpachter.wordpress.com/tag/knee-plot/ for more information. MAS sets a threshold based on the data provided.

Please see the following reference for more information:
Páll Melsted, A. Sina Booeshaghi, Fan Gao, Eduardo Beltrame, Lambda Lu, Kristján EIdjárn Hjorleifsson, Jase Gehring, Lior Pachter, Modular and efficient pre-processing of single-cell RNA-seq bioRxiv 673285; doi: https://doi.org/10.1101/673285

Purity Doublets

Cell Hashing demultiplexing algorithm. Please see the following reference for more information:

Stoeckius, M., Zheng, S., Houck-Loomis, B. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single-cell genomics. Genome Biol 19, 224 (2018). https://doi.org/10.1186/s13059-018-1603-1

DemuxEM

Cell Hashing demultiplexing algorithm. Please see the following reference for more information:

Gaublomme, J.T., Li, B., McCabe, C. et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. Nat Commun 10, 2907 (2019). https://doi.org/10.1038/s41467-019-10756-2

Gene Sets

Gene Module algorithm. Please see the following reference for more information:
Massimo Andreatta, Santiago J. Carmona UCell: Robust and scalable single-cell gene signature scoring, Computational and Structural Biotechnology Journal, Volume 19, 2021, Pages 3796-3798 (https://www.sciencedirect.com/science/artide/pii/S2001037021002816)

TriMap

Dimensionality reduction algorithm. Please see the following reference for more information:

Ehsan Amid, Manfred K. Warmuth *TriMap: Large-scale Dimensionality Reduction Using Triplets*, arXiv:1910.00204 https://arxiv.org/abs/1910.00204

UMAP

Dimensionality reduction algorithm. Please see the following reference for more information:

McInnes, L, Healy, J, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426, 2018 https://arxiv.org/abs/1802.03426.

t-SNE

Dimensionality reduction algorithm. Please see the following reference for more information:

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 https://www.jmlr.org/papers/v9/vandermaaten08a.html

Leiden Community Detection Algorithm

Community detection algorithm. Please see the following reference for more information:

Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z

PARC Community-partitioning Algorithm

Phenotyping by Accelerated Refined Community-partitioning (PARC) Algorithm. Please see the following reference for more information:

Shobana V Stassen, Dickson M D Siu, Kelvin C M Lee, Joshua W K Ho, Hayden K H So, Kevin K Tsia, PARC: ultrafast and accurate clustering of phenotypic data of millions of single cells, Bioinformatics, Volume 36, Issue 9, 1 May 2020, Pages 2778–2786, https://doi.org/10.1093/bioinformatics/btaa042

Troubleshooting

For any issues, questions, or feedback please reach out to BioLegend Technical Services:

Phone Toll-Free (US & Canada): 1.877.273.3103

Phone (International): 1.858.768.5801

For email inquiries, please use the following form: https://www.biolegend.com/en-us/contact-technical-service

BioLegend products are manufactured in an ISO 13485:2016 certified facility following GMP compliant procedures, ensuring the highest quality and stability.

Multiomics Analysis Software™ is provided by BioLegend Inc.
8999 BioLegend Way
San Diego, CA 92121
Tel: 1.858.768.5800
Tel US & Canada Toll-Free: 1.877.Bio-Legend
(1.877.246.5343) Fax: 1.877.455.9587
Email: info@biolegend.com
biolegend.com

For a complete list of worldwide BioLegend offices and distributors, please visit our website at: biolegend.com