BECKMAN COULTER
Life Sciences

Use Machine Learning Algorithms to Explore the Potential of Your High Dimensional Flow Cytometry Data Example of a 20–color Panel on CytoFLEX LX

Objective

- Understand how to perform machine learning algorithms like viSNE and FlowSOM to identify phenotypes of populations/subsets present in the 20–color CytoFLEX LX flow cytometry data
- Build a computational flow cytometry data analysis pipeline with Cytobank
- Learn how to assess the quality of viSNE maps and FlowSOM clustering results
- Recognize how pre–processing steps can affect the result quality of machine learning algorithms

Due to the advancement of technology in the cytometry field, we are now able to measure many parameters simultaneously on a single cell level. The dimensionality of datasets has increased from traditional 4–5 color low parameter to 10–20 or more color with several commercially available instruments. The sample size per experiment is also getting increasingly larger in terms of the number of events and number of samples collected. This imposes a challenge in regards to data management, collaboration, and data visualization and analysis. Many machine learning algorithmic tools are developed for dimensionality reduction and clustering to handle this increase in data complexity (Figure 1).

Cytobank is a cloud–based analysis platform with integrated analysis algorithms, as well as a structured and secure content management system for flow cytometry and other single cell data. Cytobank's clustering, dimensionality reduction, and visualization tools (**SPADE**, **viSNE**, **CITRUS**, **FlowSOM**) leverage the scalable compute and collaborative power of the cloud, allowing large analyses to be done in parallel; the cloud–based storage provides the capability to automatically archive and easily share these data securely and safely[2].

In this tech note, we illustrate how machine learning algorithms can help analyze high dimensional data using example data from a CytoFLEX LX flow cytometer 20 color experiment.

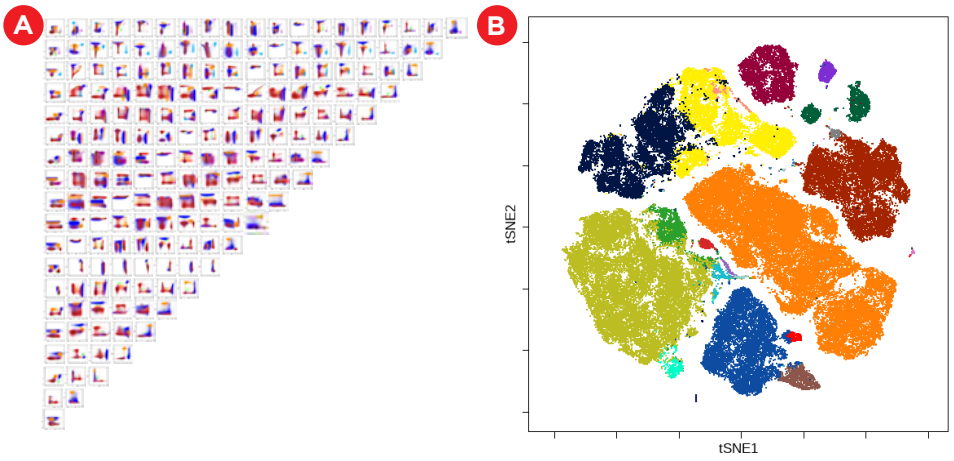


Figure 1: Comparison of biaxial plots with machine learning analysis of a 20–color panel cytometry data. A) N by N plot view of a 20–color panel cytometry data resulting in 190 plots. B) viSNE (or t–SNE) map view of the same data that visualize the 24–parameter information on one single map. NOTE: viSNE in Cytobank uses the Barnes–Hut implementation of the t–SNE algorithm [2,5].

Building a High Dimensional Flow Cytometry Data Analysis Pipeline with Cytobank

The unsupervised nature of algorithms such as viSNE, SPADE and FlowSOM increases the chance of discovering previously undescribed phenotypes. viSNE is a dimensionality reduction and visualization algorithm that creates a two-dimensional view of high-parameter data, allowing you to identify interesting and rare biological subsets. viSNE is often used as a first step for exploratory data analysis and in Cytobank can be easily linked to clustering algorithms[1]. FlowSOM is a Self Organizing Map algorithm that clusters phenotypically similar cells in a simple-to-interpret minimum spanning tree for fast, intuitive analysis and presentation of large high-dimensional data sets[5]. Combining clustering and viSNE data and displaying clustering data on a viSNE map alongside your native data can help with clustering algorithm optimization and with the assessment of cluster identity. In the next section we provide guidance on how to perfom the required steps as outlined in Figure 2.

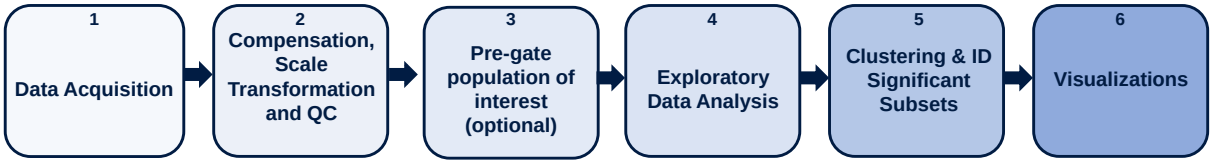


Figure 2: High Dimensional Flow Cytometry Data Analysis Pipeline with Cytobank. Key steps outlined for biomarker discovery analysis pipeline. 1) Data acquisition 2) Data compensation, transformation and quality control and 3) data cleanup and an optional pre-gating 4) Exploratory data analysis with viSNE 5) clustering and biomarker Discovery, 6)visualizations to communicate results.

1) Data acquisition

During the experiment setup and data acquisition process, it is important to consider the sample size required for statistical validity, as well as control samples depending on your research questions. If you run your samples in multiple batches, potential batch effects also need to be assessed and controlled.

2) Data compensation, transformation and quality control

Cytobank offers tools to prepare data for the subsequent analysis using machine learning algorithms. Alternatively, Kaluza Analysis Software and the Kaluza Cyotbank plugin may be used for fluorescence compensation and data transformation.

- **Compensation:** Fluorescence spillover of fluorochromes into channels other than their designated detection channel is corrected by fluorescence compensation. Compensation artifacts can negatively impact downstream analysis if not corrected. Even if samples are acquired with the compensation matrix already applied, it is advisable to check and correct the compensation for all samples.

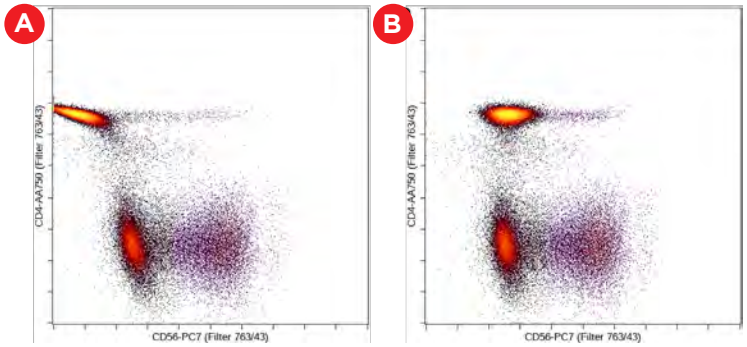


Figure 3: Cytometry data need to be properly compensated prior to the advanced machine learning analysis. Example of inappropriate (A) and appropriate compensation (B).

- **Scale transformation:** In flow cytometry, one of the following display scales are typically used, linear, log10, **arcsinh**, or biexponential transformation. Cytobank will automatically set the scale equation for fluorescent channels to arcsinh for data produced by most modern digital cytometers. It is very important to set the scales correctly for the advanced machine learning algorithms, especially the scale arguments. The algorithms will read the full range of your scale no matter if they are off scale or not, i.e. scale min and max do not matter, but the scale arguments value will affect the results. Please refer to our blog post on **how to scale cytometry data effectively.**

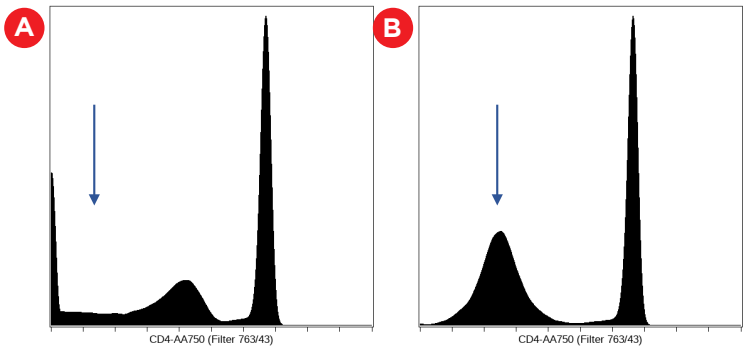


Figure 4: Scale transformation settings. An example of bad (A) and good scales (B). The CD4 channel was scaled improperly on the left, with a dip around the zero region, where all the negative values should form a single peak around the region, as shown on the right.

- **Quality control:** The stability of the sample acquisition process can be evaluated by displaying scatter and fluorescence versus time data. Gaps in data acquisition may indicate disruptions of the sample flow by blockages or bubbles. This display allows the user to restrict the analysis to parts of the acquisition when sample flow and signal detection were stable.

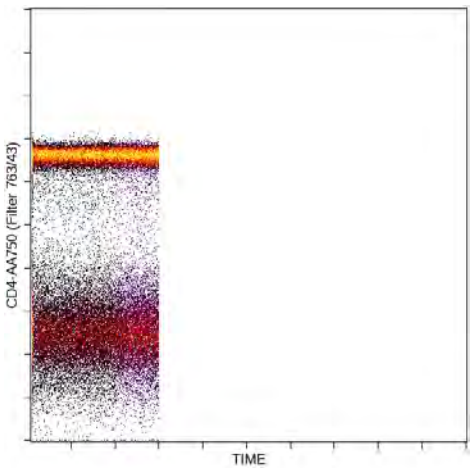


Figure 5: Quality control. An example of stable data acquisition shown by the constant fluorescence intensity over time and lack of gaps.

3) Data cleanup and an optional pre-gating

Remove unwanted events, such as from the dump channel, debris, doublets, and dead cells, by pre-gating. These events do not add information to the downstream analysis, could negatively impact data display and confound statistical results if not identified and excluded appropriately. Depending on the desired data visualization and research questions, it may be of value to pre-gate on the population of interest for further analysis.

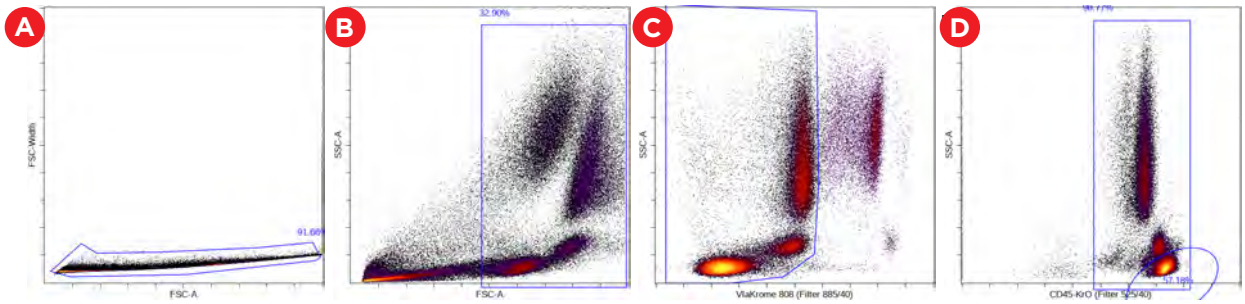


Figure 6: Data pre-processing by A) gate on single cells, B) gate out debris and C) gate on live cells D) gating on a major subset like CD45+ leukocytes or lymphocytes is optional.

4) Exploratory data analysis with viSNE

- **How to set up a viSNE run**
  - Select the pre-gated population and the files for all of the samples to be analyzed. Set the total number of events to the maximum number of events for viSNE, and select equal sampling.
  - Select the clustering channels to use for viSNE. You should exclude any channels used to define subsets during pre-gating, and include all channels that you want to use to define populations. Do not mix channels with linear scales and arcsinh scales in the channel selection, for example by selecting the time parameter or the scatter channels with the fluorescence channels together.
  - For fluorescent data, select the appropriate compensation matrix just as you would for any other analysis.
  - Adjust the number of iterations and perplexity based on the number of events. For fluorescent cytometry data, a good starting point for iterations is roughly 1000 iterations for every 100k events. NOTE: Mass cytometry data typically requires fewer iterations than fluorescent data. Perplexity can be thought of as a rough guess for the number of close neighbors (or similar points) any given event or observation will have, therefore it is specific to a given data set. Increasing perplexity can also have a big impact on algorithm run time. We recommend starting with a perplexity of 30 if the number of events in the viSNE is less than 1M, 50 for 1M to 1.5M events, and 70 for more than 1.5M events.
  - Run the viSNE. If you need to, set up multiple viSNEs runs in parallel on the cloud.
- **How to interpret the viSNE results**
  - **Assess the viSNE map quality**
    Before you can use viSNE for any exploratory data analysis, you'll want to check that you don't need to fine tune any of the advanced viSNE settings. Set up a working illustration that shows the viSNE map for every file on the rows colored by every channel on the columns. Assess the viSNE map quality by examining this working illustration to make sure that it has converged well.

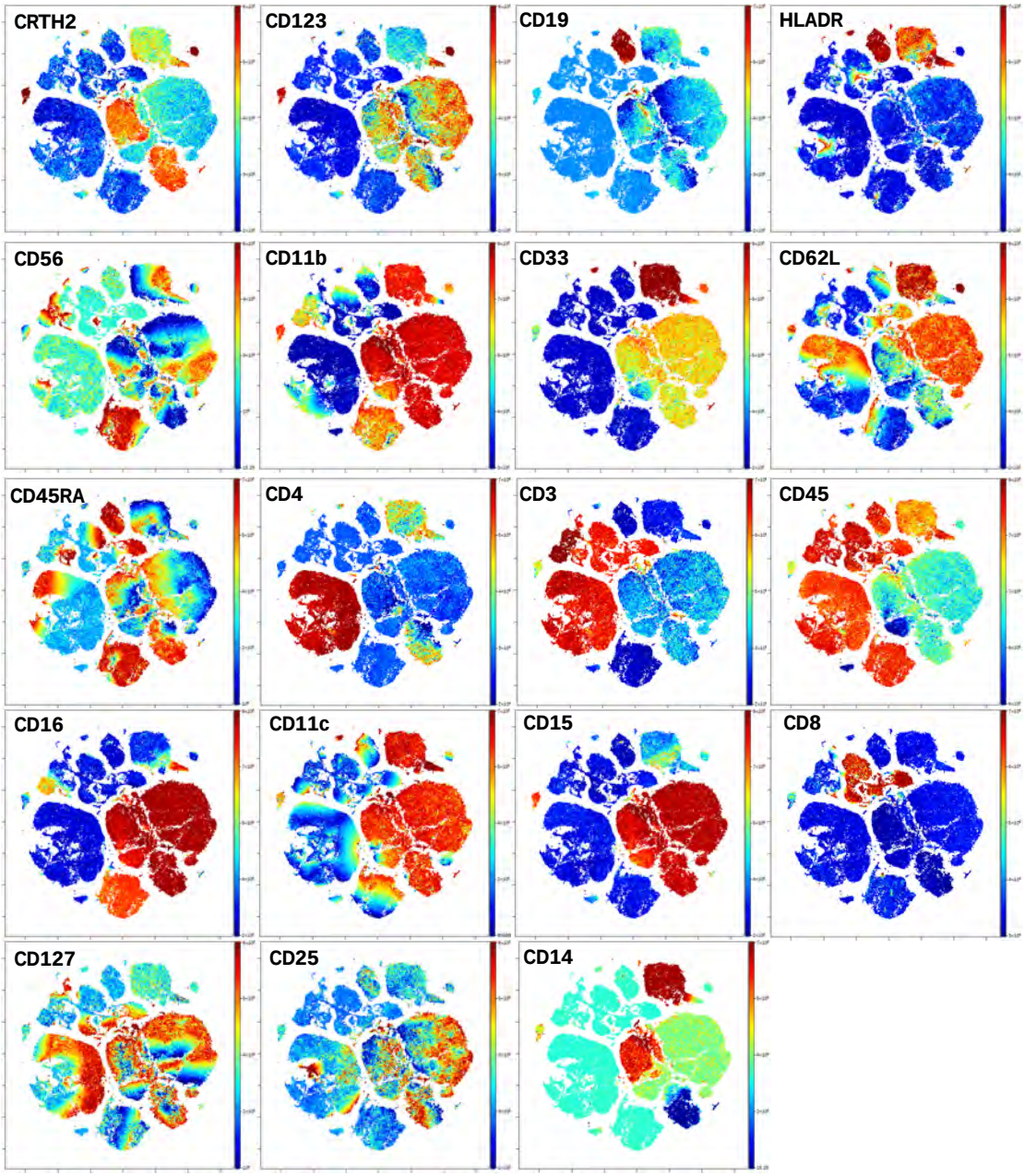


Figure 7: viSNE Visualization of 20–color flow cytometry data. viSNE ran on gated live single cells, 19 markers selected as viSNE channels. viSNE setting for 100,000 events: 1000 iteration, 30 perplexity and 0.5 theta (default settings). The intensity of each marker expression, which is labeled in the top left corner of each plot, is indicated by the plot's scale on the right.

In a nicely converged viSNE map, cells that have similar expression of clustering markers will form either separate viSNE islands (if they are very different from other cell types) or separate regions within viSNE islands. A poorly converged viSNE map will have overlapping and poorly formed islands that don't separate the expression of a single marker into distinct locations on the map (Figure 8A).
Cells expressing a given marker may appear in a string–like or spindly pattern. If the viSNE map is poorly converged, you should rerun viSNE with additional iterations (Figure 8B). viSNE maps can also be converged with the islands not very well separated. In this case, you should rerun viSNE with higher perplexity (Figure 8C). Rerun viSNE if needed, adjusting the settings based on your findings.

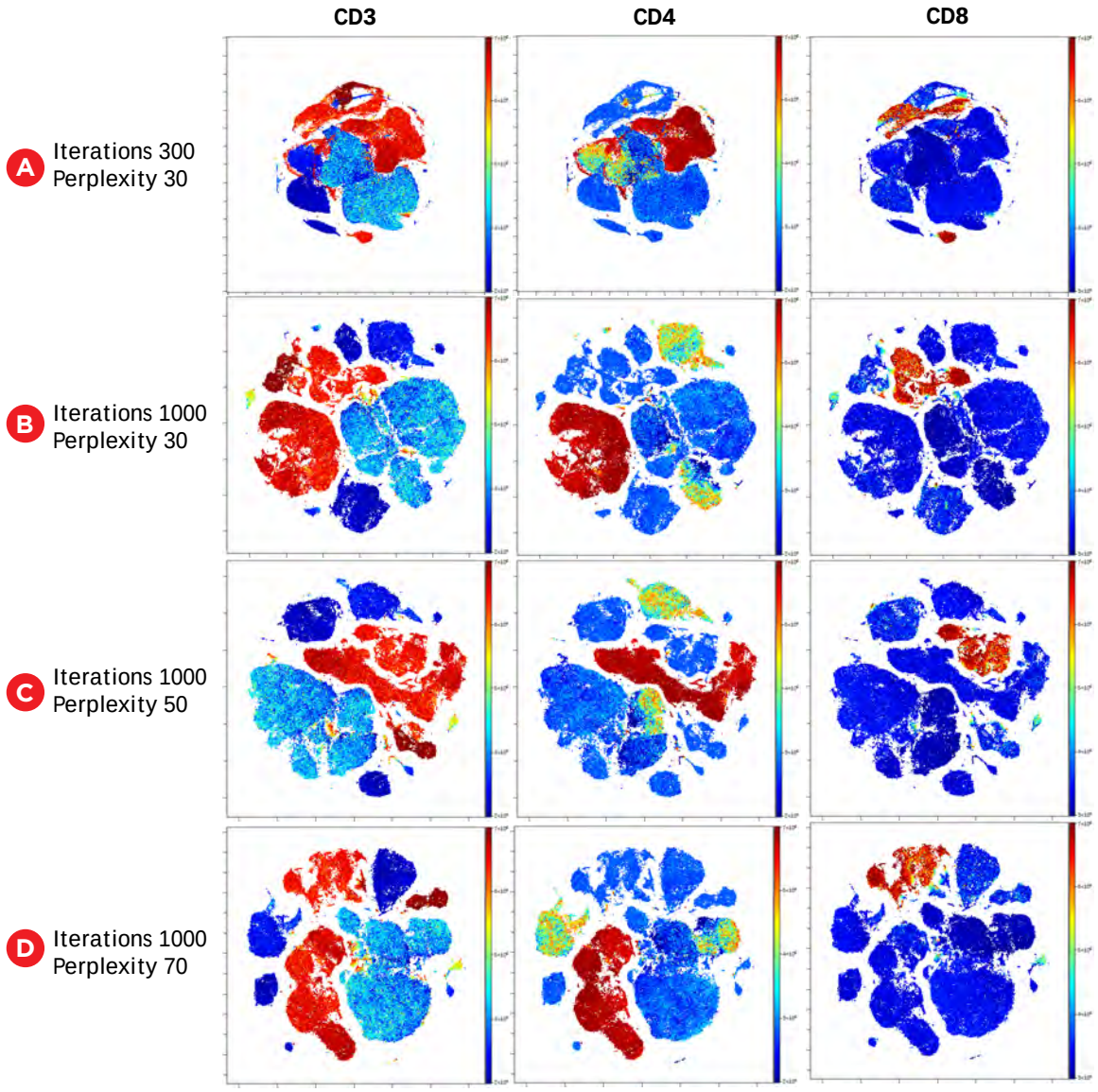


Figure 8: viSNE Visualization of 20-color flow cytometry data. viSNE ran on gated live single cells, 19 markers selected as viSNE channels. viSNE settings were chosen as indicated. The intensity of each marker expression, which is labeled on the top of each plot, is indicated by the plot's scale on the right.

TIPS FOR SUCCESS

Cancellations and failures for viSNE happen in only ~1% of all viSNE runs on the Cytobank platform. If you experience problems with cancellations and failures, it may be because your particular data set demands a lot of memory.

Some things you can try if you experience this are:

- rerun the original viSNE with the same settings
- rerun the viSNE with a lower perplexity and/or lower iterations
- select a subset of events before starting by gating on the time channel and then using the resulting population to **split files by population** if the number of selected events is much greater than the number of events sampled (visible in the setup page)

- **Exploring your data with viSNE**

  There are several visualizations you can create with viSNE to help explore and explain your results.

  - Color the viSNE map by expression of clustering markers. Create a working illustration of dot plots showing the tSNE axes colored by channel with all of the clustering channels on the columns. (see Fig. 7 above)
  - If you're looking at differences in functional marker expression, color the viSNE map by expression of functional markers not used for clustering. Create a working illustration of dot plots showing the tSNE axes colored by channel with the functional marker channels on the columns (Figure 9B). This can be done using a grid layout with files on the rows organized using sample tags for outcome groups.

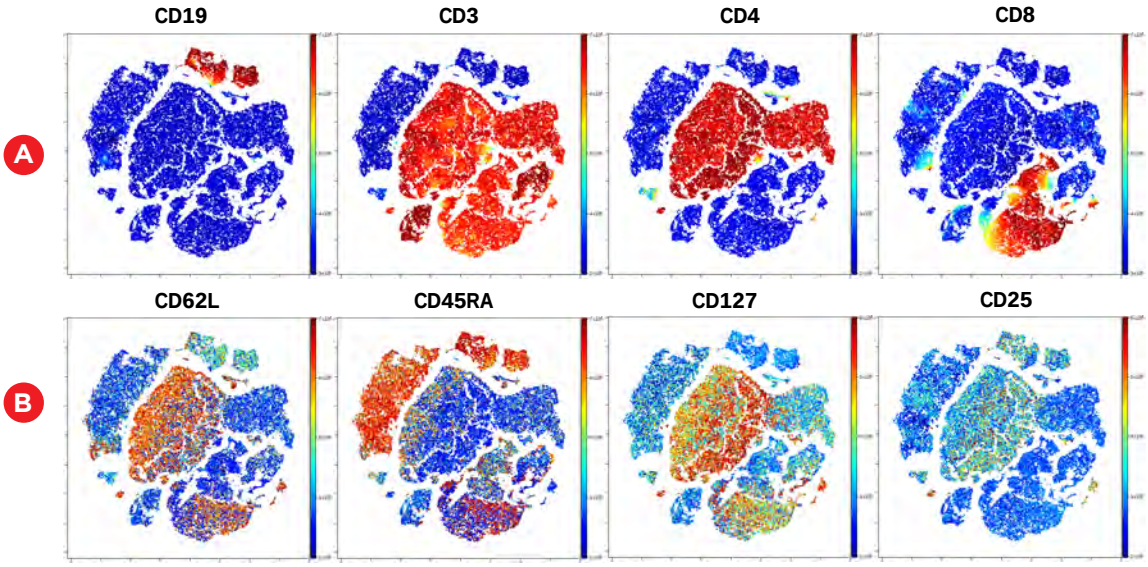


Figure 9: viSNE Visualization of 20-color flow cytometry data. viSNE ran on gated live lymphocytes, 6 markers were selected as viSNE channels viSNE setting for 100,000 events: 1000 iteration, 30 perplexity and 0.5 theta. The intensity of each marker expression, which is labeled on the top of each plot, is indicated by the plot's scale on the right. A shows population markers for orientation purposes B shows functional markers.

  - To keep track of the cell populations you are looking at, it may be helpful to manually gate the islands of the viSNE map and then display these on the working illustration. Do this by selecting 'show gates' and using the gate label 'gate name'.

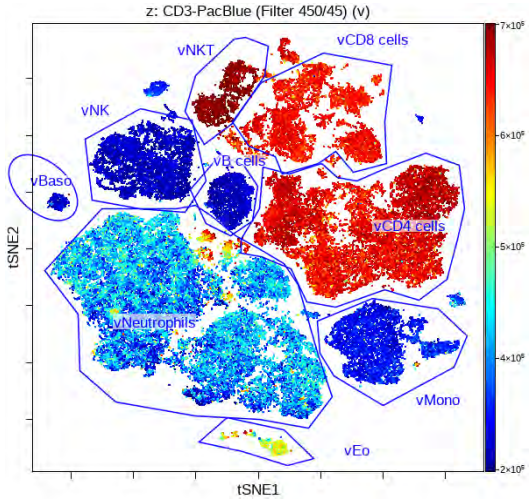


Figure 10: Gate on viSNE identified population. Use the same gating tools for dot plots on the gating window to gate on the viSNE identified population by showing the viSNE plots with third channels.

- **Examples of things gone wrong**
  - Failing to appropriately correct for fluorescence spillover and to optimize data transformation may impact the results of dimensionality reduction with viSNE.
  - Inadequate data transformation:

    Figure 11: A shows improper scaled data with negative populations for several of the markers displayed being split into 2 populations. Machine learning algorithms are unable to identify this artefact and as a result assume differences between the negative events. This can prevent the viSNE map from converging and may affect coloring by channel.

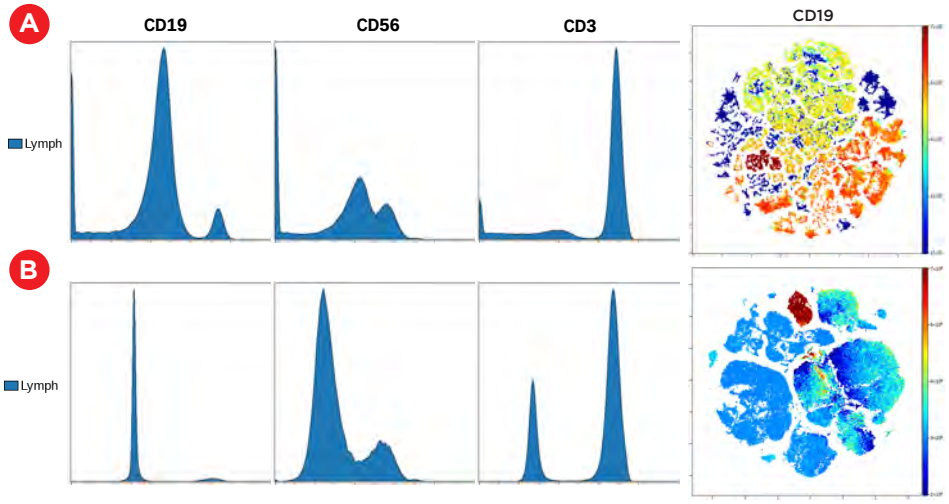


Figure 11: Effect of data transformation on viSNE visualization of 20-color flow cytometry data. viSNE performed as indicated in Figure 7. Representative histograms for lymphocyte markers show the data transformation. A shows suboptimal data transformation resulting in a poorly converged viSNE map and falsly high coloring for CD19. B shows the same data with optimized transformation settings.

  - Inadequate fluorescence compensation may lead to the appearance of false positive or false negative events and will impact the coloring by channel. Figure 12A shows overcompensation of spillover from the CD4 signal into the CD56 channel, resulting in the appearance of extremely low fluorescence intensity of CD4+ cells in the CD56 channel. Undercompensation of the spillover of CD8 in the CD15 channel results in these cells showing in incorrect intermediate signal intensity for CD15. Figure 12B shows results after spillover correction.

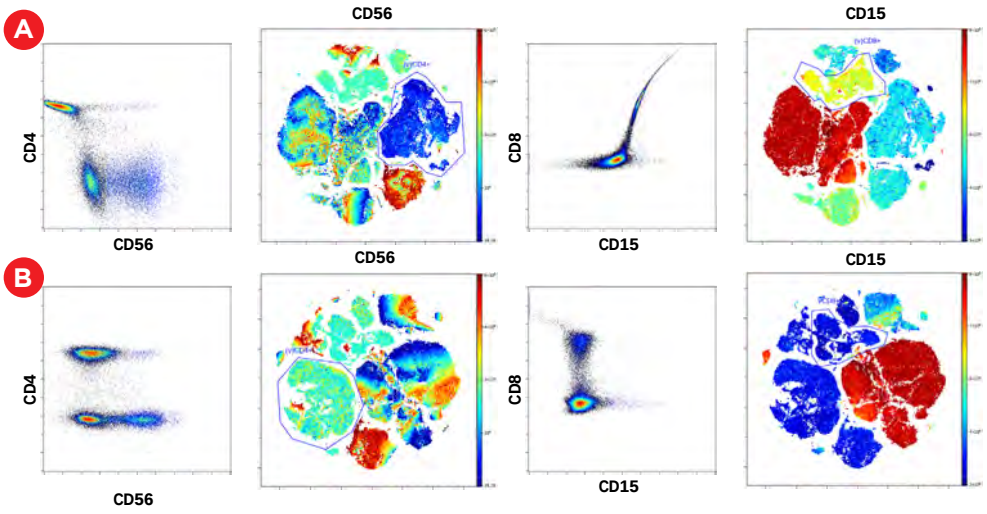
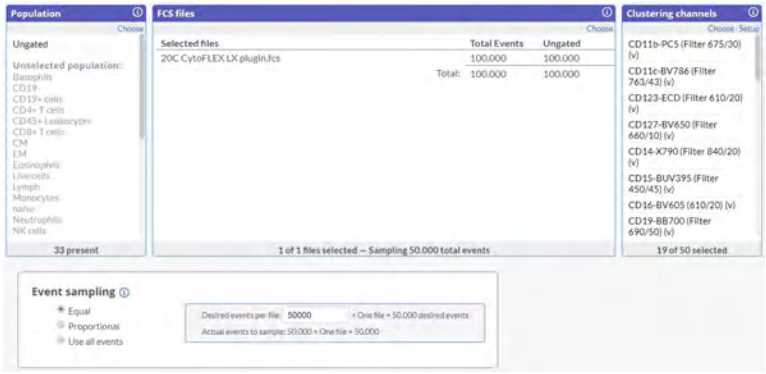


Figure 12: Effect of fluorescence spillover compensation on viSNE visualization of 20-color flow cytometry data. viSNE performed as indicated in Figure 7.

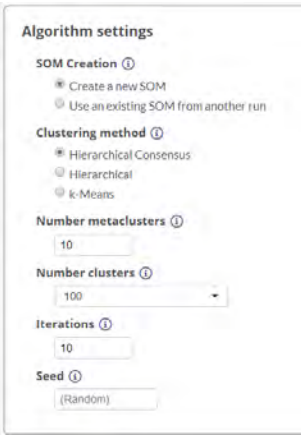
5) Clustering and biomarker discovery

Clustering algorithms such as FlowSOM and SPADE can be used to automatically identify cells of similar phenotype in an unbiased manner[3,4]. It can be performed on the original uploaded experiment or it can be performed on the generated viSNE experiment and viSNE identified islands.

- **Run FlowSOM on the original experiment**
  - **Set up the FlowSOM analysis on the original experiment**



- Select the pre-gated population and the files for all of the samples you determined you will include in this FlowSOM run. Set the total number of events to the proper number of events, and select equal sampling. FlowSOM is a fast algorithm that can process up to 4 million total events on the premium server.
- Select the clustering channels to use for FlowSOM. You should exclude any channels used to define subsets during pre-gating, and include all channels that you want to use to define populations. Be sure to not mix channels with linear scales and arcsinh scales with the channel selection, such as select the time parameter or the scatter channels together with fluorescence channels.
- For fluorescent data, select the appropriate compensation matrix just as you would for any other analysis.
- Choose the Algorithm settings.
  - SOM selection: You may choose to create a new SOM or use an existing SOM from a previous run.
  - Clustering method: choose a method that will be used for metaclustering. The default Hierarchical Consensus clustering works best in most cases.
  - Choose a target number of metaclusters: The number of metaclusters for a FlowSOM run determines how many metaclusters the clusters will be organized into. The correct number of metaclusters will depend on many factors and is data specific. A good starting point is to use a number that is slightly higher than the total number of major subsets to be discovered.

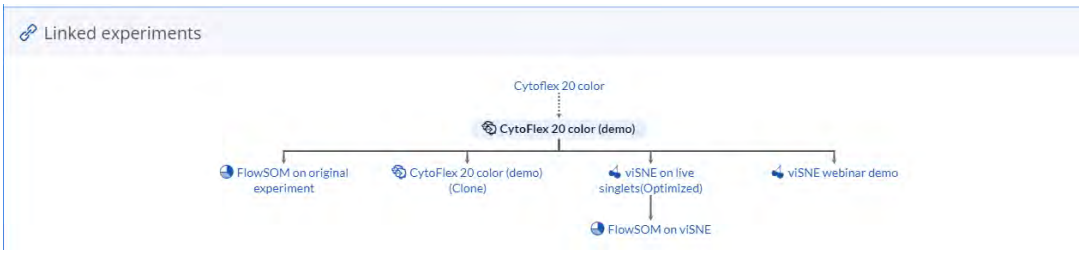


  - SOM selection: You may choose to create a new SOM or use an existing SOM from a previous run.
  - Clustering method: choose a method that will be used for metaclustering. The default Hierarchical Consensus clustering works best in most cases.
  - Choose a target number of metaclusters: The number of metaclusters for a FlowSOM run determines how many metaclusters the clusters will be organized into. The correct number of metaclusters will depend on many factors and is data specific. A good starting point is to use a number that is slightly higher than the total number of major subsets to be discovered.

- To configure the PDF output review our online support article How to Configure and Run a FlowSOM Analysis here
- Run FlowSOM by selecting the green button at the top. If you need to, set up multiple FlowSOM runs in parallel on the cloud by copying the settings and starting a new run.

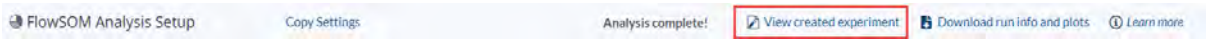
- **Run FlowSOM on the viSNE experiment**

In order to visualize the FlowSOM clusters on a viSNE plot, you need to initiate the FlowSOM run within the previously created viSNE experiment. You may access the viSNE experiment from the original experiment by going to Actions>View experiment summary. In the Linked experiments section select the viSNE experiment on which you wish to run FlowSOM and then follow the steps outlined in section **"Set up the FlowSOM analysis on the original experiment".**



Optimizing FlowSOM settings

After successfully **setting up and running a FlowSOM analysis**, you can perform exploratory or quantitative downstream analysis on your FlowSOM results. As part of the analysis process, you'll want to assess the quality of the FlowSOM analysis, which can be done by displaying FlowSOM metaclusters on a viSNE map and with heatmaps that show the marker expression of the FlowSOM metaclusters. When a FlowSOM run completes, a new experiment is created that contains the original files with new channels added for the FlowSOM cluster ID and metacluster ID. You can access this experiment either from the FlowSOM settings page for the run (click "View created experiment" on the page–level navbar), or from the Experiment Summary page of the originating experiment (click on the FlowSOM run name and you will be taken to the FlowSOM experiment).



Overlaying FlowSOM–identified metaclusters onto a viSNE map

The Working Illustration of the newly created experiment will be configured by default to set up dot plots with metaclusters overlaid on the viSNE map if you ran viSNE first, followed by FlowSOM. This visualization can help you assess the quality of the metaclustering and can inform how you may need to iterate various settings including target number of metaclusters and normalization.

You can compare overlaid clusters on a viSNE map to a contour plot of the viSNE map to gauge whether the number of clusters generated was sufficient to capture distinct, dense populations.

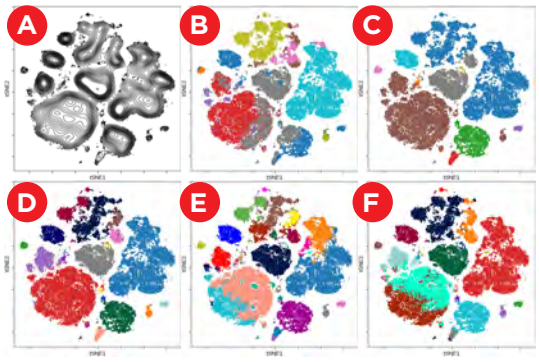


Figure 13: A: Contour plot of the viSNE map. B: manual gates overlaid on viSNE map C–E: FlowSOM metaclusters are overlaid on the viSNE map. Same seed was manually set for all FlowSOM runs. Clustering channels for FlowSOM are the same as the channels used for the viSNE experiment.

C: 10 metaclusters

D: 22 metaclusters

E: 35 metaclusters

F: 22 metaclusters, normalized.

Generally, the colors of the overlaid clusters should correspond to the viSNE continent definitions. If the same cluster spans multiple continents, that might indicate a need to increase the target number of clusters and/or metaclusters, or a need to enable or disable normalization. In the example above, you can see how a FlowSOM target number of metaclusters of 10 does not provide enough resolution, and a target of 35 provides too much resolution (though you could further explore this to see if, in fact, it is detecting relevant sub-populations). A target number of metaclusters of 22 appears to yield good correspondence between the FlowSOM metaclustering and the viSNE continents for the data set used here.

While normalization is often a helpful and necessary approach to transforming channel data prior to analysis, for some datasets it is not (see example above), and that can be rapidly detected via the display of FlowSOM metaclusters on a viSNE map, followed up by inspection of cluster expression. If you find that enabling normalization appears to diminish the resolution of the metaclustering, you can also try increasing the target number of metaclusters. There may be cases where enabling normalization appears to lower the resolution and result in poor quality, but by increasing the target number of metaclusters, you may be able to pull out sub-populations.

TIPS FOR SUCCESS

When a FlowSOM run completes, it writes FlowSOM_cluster_id and FlowSOM_metacluster_id channels into the newly generated files that comprise the FlowSOM analysis experiment. If you have run viSNE on the files first, the files in the FlowSOM analysis experiment will now contain all of the original channels and data, as well as the tSNE1 and tSNE2 channels from the viSNE run, and the new FlowSOM_cluster_id and FlowSOM_metacluster_id channels. The Working Illustration will be configured by default to set up dot plots with metaclusters overlaid on the viSNE map. (You can also run FlowSOM first, and then run viSNE – the same smart Working Illustration view will be set up following the viSNE run.)

6) Visualizations to communicate results

When running FlowSOM in an experiment containing viSNE results, you can take advantage of both the clustering and dimensionality reduction algorithms. You can use FlowSOM to get quantitative clusters to identify unique subsets and or marker expression and then use viSNE for visualization.

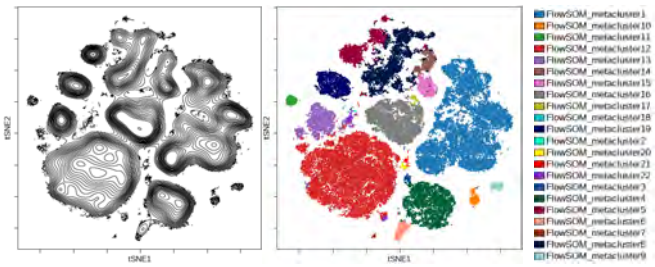


Figure 14: Contour map of viSNE map compared toFlowSOM metaclusters overlaid on the viSNE map.

FlowSOM was performed on the viSNE experiment with 22 metaclusters. Clustering channels for FlowSOM are the same as the channels used for the viSNE experiment.

- **Identify clusters using heatmaps to visualize marker expression**

Now that you have optimized your clustering algorithm run, you can proceed to view the expression of markers of interest in heatmap format. On the Working Illustration page, under "Plot Controls" under the illustration tab, select "Plot Type" –– "Heatmap". Make sure to select the appropriate channels for your heatmap in the "Channels" dimension. Adjust the other dimensions as needed. In order to create a heatmap with multiple channels, use the **channels figure dimension box** within the Working Illustration. Make sure to set the **X axis illustration setting to Panel/Channel Values**. This setting defers control of the x axis to the channels figure dimension box and allows for figures with many channels. A common mistake is to not set this axis setting correctly, resulting in every square of the heatmap having the same data value.

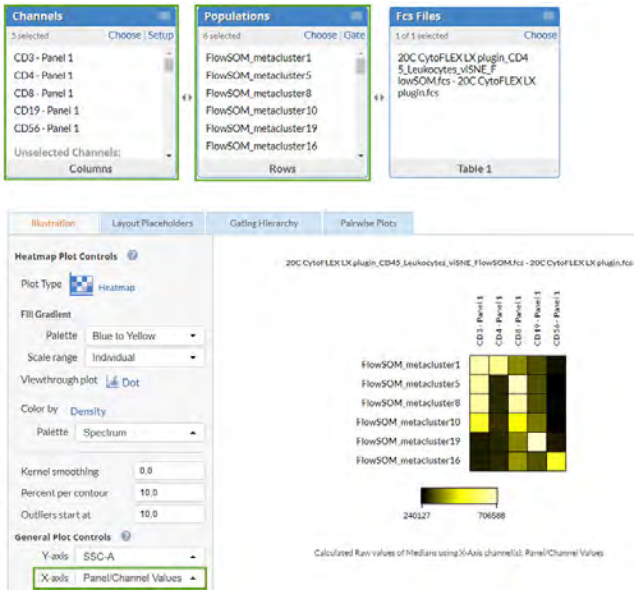


Figure 15: Example of how to create a heatmap to compare marker expression between metaclusters. Important settings are indicated by green boxes. FlowSOM metaclusters are listed in rows and marker expression on the column. Displayed are raw median intensity values.

The scale coloring range is the range of the color bar below the heatmap.

Global (Automatic): The coloring range is the same across heatmaps.

Individual: The coloring range is calculated per heatmap.

Global (User–defined): The coloring range is user–defined and global with an inflection point. Currently it is forced to use both sides of the color scale for this configuration.

The desired statistic and equation (e.g. fold) can be chosen using the illustration controls panel on the left side of the working illustration. Heatmaps can be made with any statistic, such as event counts, percent in gate, medians, as well as ratios of these statistics compared between samples (e.g. fold change of event counts).

By default, heatmaps display statistical values and color themselves according to raw, unscaled data values. In order to view the data in the biologically more useful transformed space, the transformed ratio should be used.

When using a comparison equation, make sure to set the control appropriately. A control of table's minimum is usually appropriate to visualize an entire heatmap in scaled mode. When using a single heatmap with many channels, however, a control that normalizes each channel to itself is likely desired. Controls using column or row would be indicated for this, depending on how channels are arranged.

TIPS FOR SUCCESS

While it is possible to perform viSNE, SPADE and FlowSOM calculations using desktop software, the computer cannot be used for other tasks during this time and the available compute may limit the number of events that can be analyzed per sample, requiring down-sampling. Also, only one run at a time can be performed, slowing down optimization of algorithm settings. Cloud-based data analysis does not suffer from these limitations, multiple runs can be run in parallel, ie varying algorithm settings, and do not utilize desktop compute power.

Summary

Combining dimensionality reduction using viSNE, clustering with FlowSOM and visualization of expression patterns using heatmaps can be combined into a data analysis pipeline that will allow unbiased identification of cellular subsets.

Performing this integrated workflow on the Cytobank platform allows users to easily combine algorithms and take advantage of smart visualization options. Running analyses in the cloud allows the user to perform different runs in parallel to optimize settings faster.

REFERENCES:

1. Amir, E. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., & Pe'er, D. (2013). ViSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nature Biotechnology, 31(6), 545–552. https://doi.org/10.1038/nbt.2594
2. Kotecha, N., Krutzik, P. O., & Irish, J. M. (2010). Web-Based Analysis and Publication of Flow Cytometry Experiments. Current Protocols in Cytometry, 53(1), 10.17.1–10.17.24. https://doi.org/10.1002/0471142956.cy1017s53
3. Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., & Plevritis, S. K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nature Biotechnology, 29(10), 886–891. https://doi.org/10.1038/nbt.1991
4. van der Maaten, L. (2014). Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research.
5. Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., & Saeys, Y. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data: FlowSOM. Cytometry Part A, 87(7), 636–645. https://doi.org/10.1002/cyto.a.22625

For Beckman Coulter's worldwide office locations and phone numbers, please visit Contact Us at beckman.com

FLOW-6309APP02.20