Dimensionality Reduction Algorithms: How to Choose?

In this Application Note, you will learn:

- What to consider when choosing a dimensionality reduction tool
- viSNE, UMAP, opt-SNE and tSNE-CUDA: basic knowledge and Cytobank implementations
- Events limits for each algorithm in the Cytobank platform
- Tips on how to select the correct algorithm for your dataset

One of the biggest challenges when analyzing high-dimensional datasets in flow cytometry is capturing the expression of several markers simultaneously for each cell. When looking at just a few plots together, it might be quite easy to spot connections and identify relevant subsets of cells with a similar phenotype. Even when focusing on a relatively small number of colors compared to what modern technology enables (e.g., 15 colors), researchers need to look at more than 100 plots all together to have a general impression of how cells in that specific sample are behaving.

Dimensionality Reduction algorithms can transform data from a high-dimensional space into a low-dimensional space, so that the low-dimensional representation retains many meaningful properties of the original data. Translated into the flow cytometry world, dimensionality reduction allows for the creation of a 2D map where each cell is a point, and cells that are close to each other on this map have more similar phenotypes, meaning they express similar levels of several markers. Creating such maps allows for an easy visualization and rapid exploratory data analysis of multi-parametric datasets: researchers looking at those maps will be able to spot “islands” (i.e., sub populations) of cells with a defined phenotype and, by comparing maps from different samples, can follow the behavior of those cell subsets across samples.

Starting from the first publication of the viSNE algorithm1 in 2008, the flow cytometry field has progressively adopted several other dimensionality reduction algorithms, each with slightly different characteristics but serving the same purpose: to easily visualize results in a more intelligible way.

The new Dimensionality Reduction Suite, released in the Cytobank platform with Cytobank 9.0, provides users with the choice to select from among four different algorithms for dimensionality reduction. Let’s take a closer look at each of them.

viSNE

viSNE in Cytobank uses the Barnes-Hut implementation of the t-SNE algorithm.[2] viSNE works creating two new channels in your data, called tSNE1 and tSNE2, which contain information on the position of each cell in the high dimensional space. By plotting one against the other, you will obtain the viSNE map, where each point represents a single cell, and cells that are closer together are more similar to one another across the n-dimensions used to generate the viSNE analysis. Maximum number of events for a viSNE run in the Cytobank cloud is 1.3M for Premium and 2M for Enterprise.

Opt-SNE

Opt-SNE[3] is a much faster option compared to tSNE. The definition from the publication of Belkina's lab is "an automated toolkit for t-SNE parameter selection that utilizes Kullback-Leibler divergence evaluation in real time to tailor the early exaggeration and overall number of gradient descent iterations in a dataset-specific manner." What this means is that opt-SNE enables optimal[3] data resolution in t-SNE space by automating the selection of settings for each dataset, reducing time-to-results. This approach also eliminates the need for trial-and-error runs intended to empirically find the most favorable selection of t-SNE parameters, potentially saving many hours of computation time per research project. Each opt-SNE run in the Cytobank platform has a maximum of 3M total events.

tSNE-CUDA

tSNE-CUDA[4] is a GPU-accelerated implementation of t-SNE for visualizing datasets. tSNE-CUDA significantly outperforms other current implementations of the tSNE algorithm in terms of run time, dramatically reducing time-to-results. In the Cytobank cloud, you will have a maximum of 10M total events for each tSNE-CUDA run.

UMAP

UMAP[5] stands for Uniform Manifold Approximation and Projection algorithm. In the Cytobank platform we have implemented the GPU version of UMAP, which makes it even faster. Results of a UMAP run are similar to tSNEs results, as it provides a two-dimensional map where each point is a cell; nevertheless, it is widely believed to preserve global structure better than other solutions. Indeed, the more densely packed islands make it easy to observe global structure. In UMAP, distance between islands is meaningful: closer islands are more similar than islands farther apart. In the Cytobank cloud you can run up to 10M events per UMAP run.

In the Cytobank platform, there are several features you can use to collect meaningful insights from a dimensionality reduction analysis. You can Color-by-Channel the map, to view single-cell expression of measured parameters (Figure 1A), gate and isolate cells populations of interest (Figure 1B), Color-by-Sample, to compare maps from different samples (e.g., different treatment conditions) (Figure 1C) or Color-by-Density, to highlight abundance of cell populations (Figure 1D). All these options can be used for any of the dimensionality reduction algorithms, as the basic principles behind them are the same.

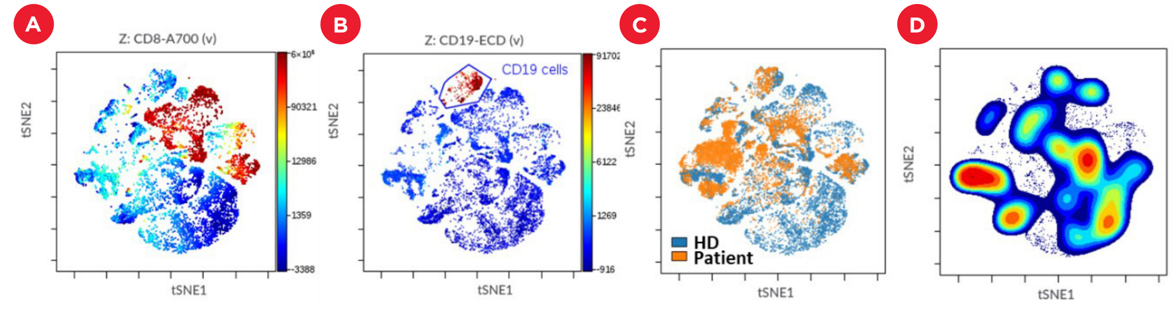

Figure 1. viSNE (in the Cytobank platform) was run with default settings on CD45+ cells manually gated in two samples from a healthy donor and a patient. Results have been visualized colored by channel **(A)**, gated **(B)**, overlaid and colored by sample **(C)** or colored by density **(D)**.

One last general point to mention: as we have seen, in the Cytobank platform, the tSNE-CUDA and UMAP implementations are GPU - based (graphics processing unit), which means they are much faster than the original algorithms because they use the massively parallel processing power of the GPU to boost computational performance. Nevertheless, it also means they might be particularly hard to install locally, even for experienced programmers, and they might have specific hardware requirements to run effectively. Therefore, one of the greatest advantages of the Cytobank platform is that you can access the computational power of the cloud, allowing the use of faster implementations of dimensionality reduction algorithms that you might otherwise not have the hardware, resources or programming knowledge to use.

Which algorithm should I use?

After briefly exploring the basic characteristics of each algorithm, you might want to determine which algorithm you should choose for the exploration of your data. Let's review a few important points that might help you make your selection.

- tSNE-CUDA might be a fast option to explore large datasets, as it works very well with default settings on most tested datasets; nevertheless, if you want to be able to include the seed for your run in your paper, be aware that the only two algorithms that will allow you to set the seed are opt-SNE and the standard viSNE.
- In terms of visualization, tSNE-CUDA islands are bigger and less compact, and the points are more spread out and don't overlap each other as much as in UMAP. Therefore, it will be easier to discriminate subpopulations or subtle differences sample-to-sample within the tSNE-CUDA islands than the dense UMAP islands. (Figure 2)

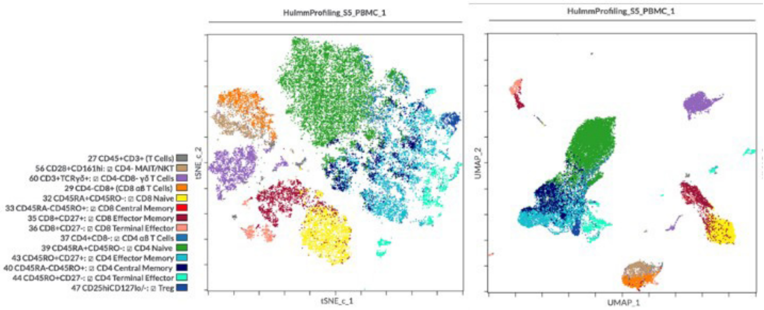

Figure 2. Both illustrations show manually gated T cell populations overlaid on dimension reduction. The CD4 populations (blue and green) are more spread out in the tSNE-CUDA than the UMAP.

- On the other hand, the shape of a UMAP embedding may look fuller/more clear for smaller samples than for tSNEs (Figure 3).

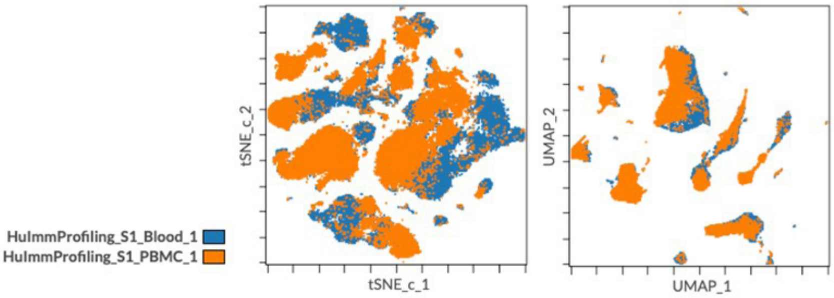

Figure 3. Blood and PBMC from the same subject overlaid by tSNE-CUDA (left) and UMAP (right). This is gate01 from the Fluidigm MaxPar dataset. Differences are clearer in tSNE-CUDA.

- Importantly, as you can see from Figure 4, for samples with low event counts, it may be easier to see underlying data patterns in UMAP, if events on tSNE are too sparse.

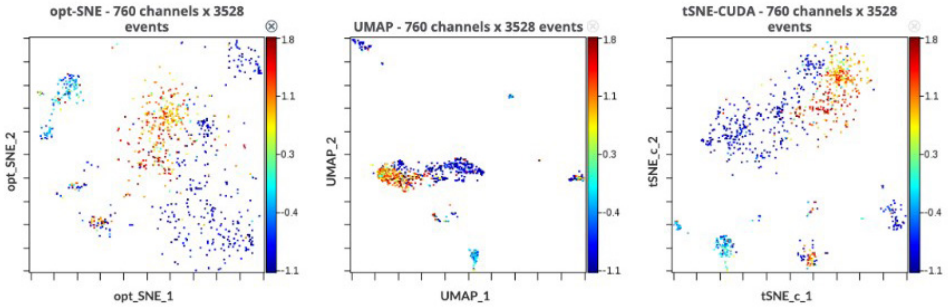

Figure 4. CITE-Seq data, 760 channels and 3528 events included in dimensionality reduction from 4 files.

- It is also important to keep in mind that UMAP can preserve the global structure better than tSNE because the default initialization method of UMAP can keep the global structure. Distance among cell islands in UMAP is meaningful, while it is not for the tSNEs algorithms; therefore, UMAP is better for investigating relationships between cell clusters.

In conclusion, we have highlighted the major differences among the different dimensionality reduction methods and shown specific characteristics that can guide you in selecting the best algorithm to fit your datasets and your research needs.

To learn more about how to configure and run a dimensionality reduction analysis, the effect of settings on run time, or how to compare results of algorithms for dimensionality reduction, please visit our support site at **support.cytobank.org**.

References

1. Van der Maaten and Hinton, Journal of Machine Learning Research, 2008
2. Van der Maaten, Journal of Machine Learning Research, 2014
3. Belkina A. et al., Nat Communications, 2019
4. Chan DM et al., Journal of Parallel and Distributed Computing, 2019
5. Corey J. Nolet. et al., arXiv, 2020, arXiv:2008.00325 **http://arxiv.org/abs/2008.00325**

For research use only. Not for use in diagnostics procedure.

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

21.07.2150.FLOW