Journal of Visualized Experiments

ExCYT: A Graphical User Interface for streamlining analysis of high-dimensional cytometry data

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE57473R3
Full Title:	ExCYT: A Graphical User Interface for streamlining analysis of high-dimensional cytometry data
Keywords:	flow cytometry; High-Dimensional Analysis; t-SNE; clustering; heat maps; Dimensionality Reduction
Corresponding Author:	Alexander Baras, MD, PhD Johns Hopkins University Baltimroe, Maryland UNITED STATES
Corresponding Author's Institution:	Johns Hopkins University
Corresponding Author E-Mail:	baras@jhmi.edu
First Author:	John-William Sidhom
Other Authors:	John-William Sidhom
	Debebe Theodros
	Benjamin Murter
	Jelani C Zarif
	Sudipto Ganguly
	Drew M Pardoll
Author Comments:	
Additional Information:	
Question	**Response**
If this article needs to be "in-press" by a certain date, please indicate the date below and explain in your cover letter.	

TITLE:

ExCYT: A Graphical User Interface for Streamlining Analysis of High-Dimensional Cytometry Data

AUTHORS & AFFILIATIONS:

John-William Sidhom[1,2,3], Debebe Theodros[1,2,4], Benjamin Murter[1,2], Jelani C. Zarif[1,2], Sudipto Ganguly[1,2], Drew M. Pardoll[1,2], Alexander Baras[1,2,5]

[1] The Bloomberg~Kimmel Institute for Cancer Immunotherapy, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[2] The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[3] Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[4] Department of Immunology, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[5] Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Jsidhom1@jhmi.edu
Dtheodros@jhmi.edu
Bmurter1@jhmi.edu
Jzarif1@jhmi.edu
Sgangul8@jhmi.edu
Dpardol1@jhmi.edu
Baras@jhmi.edu

KEYWORDS:

Flow Cytometry, High-Dimensional Analysis, t-SNE, clustering, heat maps, Dimensionality Reduction

SHORT ABSTRACT:

ExCYT is a MATLAB-based Graphical User Interface (GUI) that allows users to analyze their flow cytometry data via commonly employed analytical techniques for high-dimensional data including dimensionality reduction via t-SNE, a variety of automated and manual clustering methods, heatmaps, and novel high-dimensional flow plots.

LONG ABSTRACT:

With the advent of flow cytometers capable of measuring an increasing number of parameters, scientists continue to develop larger panels to phenotypically explore characteristics of their cellular samples. However, these technological advancements yield high-dimensional data sets that have become increasingly difficult to analyze objectively within traditional manual-based gating programs. In order to better analyze and present data, scientists partner with bioinformaticians with expertise in analyzing high-dimensional data to parse their flow cytometry data. While these methods have been shown to be highly valuable in studying flow cytometry, they have yet to be incorporated in a straightforward and easy-to-use package for scientists who lack computational or programming expertise. To address this need, we have

developed ExCYT, a MATLAB-based Graphical User Interface (GUI) that streamlines the analysis of high-dimensional flow cytometry data by implementing commonly employed analytical techniques for high-dimensional data including dimensionality reduction by t-SNE, a variety of automated and manual clustering methods, heatmaps, and novel high-dimensional flow plots. Additionally, ExCYT provides traditional gating options of select populations of interest for further t-SNE and clustering analysis as well as the ability to apply gates directly on t-SNE plots. The software provides the additional advantage of working with either compensated or uncompensated FCS files. In the event that post-acquisition compensation is required, the user can choose to provide the program a directory of single stains and an unstained sample. The program detects positive events in all channels and uses this select data to more objectively calculate the compensation matrix. In summary, ExCYT provides a comprehensive analysis pipeline to take flow cytometry data in the form of FCS files and allow any individual, regardless of computational training, to use the latest algorithmic approaches in understanding their data.

INTRODUCTION:

Advances in flow cytometry as well as the advent of mass cytometry has allowed clinicians and scientists to rapidly identify and phenotypically characterize biologically and clinically interesting samples with new levels of resolution, creating large high-dimensional data sets that are information rich[1-3]. While conventional methods for analyzing flow cytometry data such as manual gating have been more straightforward for experiments where there are few markers and those markers have visually discernable populations, this approach can fail to generate reproducible results when analyzing higher-dimensional data sets or those with markers staining on a spectrum. For example, in a multi-institutional study, where intra-cellular staining (ICS) assays were being performed to assess the reproducibility of quantitating antigen-specific T cell responses, despite good inter-laboratory precision, analysis, particularly gating, introduced a significant source of variability[4]. Furthermore, the process of manually gating population of interests, besides being highly subjective is highly time consuming and labor intensive. However, the problem of analyzing high-dimensional data sets in a robust, efficient, and timely manner is not one new to the research sciences. Gene expression studies often generate extremely high-dimensional data sets (often on the order of hundreds of genes) where manual forms of analysis would be simply infeasible. In order to tackle the analysis of these data sets, there has been much work in developing bioinformatic tools to parse gene expression data[5]. These algorithmic approaches have just been recently adopted in the analysis of cytometry data as the number of parameters has increased and have proven to be invaluable in the analysis of these high dimensional data sets[6-7].

Despite the generation and application of a variety of algorithms and software packages that allow scientists to apply these high-dimensional bioinformatic approaches to their flow cytometry data, these analytical techniques still remain largely unused. While there may be a variety of factors that have limited the widespread adoption of these approaches to cytometry data[8], the major hindrance we suspect in use of these approaches by scientists, is a lack of computational knowledge. In fact, many of these software packages (*i.e.,* flowCore, flowMeans, and OpenCyto) are written to be implemented in programming languages such as R that still require substantive programming knowledge. Software packages such as FlowJo have found

favor among scientists due to simplicity of use and 'plug-n-play' nature, as well as compatibility with the PC operating system. In order to provide the variety of accepted and valuable analytical techniques to the scientist unfamiliar programming, we have developed ExCYT, a graphical-user interface (GUI) that can be easily installed on a PC/Mac that pulls many of the latest techniques including dimensionality reduction for intuitive visualization, a variety of clustering methods cited in the literature, along with novel features to explore the output of these clustering algorithms with heatmaps and novel high-dimensional flow/box plots.

ExCYT is a graphical user interface built in MATLAB and therefore can either be run within MATLAB directly or an installer is provided that can be used to install the software on any PC/Mac. The software is available at https://github.com/sidhomj/ExCYT. We present a detailed protocol for how to import data, pre-process it, conduct t-SNE dimensionality reduction, cluster data, sort & filter clusters based on user preferences, and display information about the clusters of interest via heatmaps and novel high-dimensional flow/box plots **(Figure 1)**. Axes in t-SNE plots are arbitrary and in arbitrary units and as such as not always shown in the figures for simplicity of the user interface. The coloring of data points in the "t-SNE Heatmaps" is from blue to yellow based on the signal of the indicated marker. In clustering solutions, the color of the data point is based arbitrary on cluster number. All parts of the workflow can be carried out in the single panel GUI **(Figure 2 & Table 1)**. Finally, we will demonstrate the use of ExCYT on previously published data exploring the immune landscape of renal cell carcinoma in the literature, also analyzed with similar methods. The sample dataset we used to create the figures in this manuscript along with the protocol below can be found at https://premium.cytobank.org/cytobank/projects/875, upon registering an account.

PROTOCOL:

1. Collecting and Preparing Cytometry Data

1.1. Place all single stains in a folder by themselves and label by the channel name (by fluorophore, not marker).

2. Data Importation & Pre-Processing

2.1. To pause or save throughout this analysis pipeline, use the **Save Workspace** button at the bottom left of the program to save the workspace as a '.MAT' file that can later be loaded via the **Load Workspace** button. Do not run more than one instance of the program at a time. Therefore, when loading a new workspace, make sure to check there is no other instance of ExCYT running.

2.2. To begin analysis pipeline, first select type of cytometry (Flow Cytometry or Mass Cytometry – CYTOF), under the **File Selection Parameters** select number of events to sample from the file (for this example use 2000). Once data has been successfully imported, a dialogue box will pop up informing the user that the data has been successfully imported.

2.3. Press the **Auto-Compensation** button to conduct an optional auto-compensation step, as done by Bagwell & Adams[9]. Select the directory containing single stains. Select the unstained sample within the user interface dialogue.

2.3.1. Place a forward/side-scatter gate on any of the samples in this directory that will be used to select events to calculate the compensation matrix. It is recommended to use the unstained sample for this purpose. At this point, an algorithm has been implemented to set consistent thresholds at the 99th percentile of the unstained sample to define positive events in each of the single stains to calculate the compensation matrix. When this is finished, a dialogue box will inform the user that the compensation has been performed.

2.4. Next, press **Gate Population** and select the populations of cells of interest, as is the convention in flow cytometry analyses. When population of cells is selected, enter number of percentage of events downstream analysis (in this 10000 events).

2.5. Next, select the number channels to be used for analysis in the listbox in the far right of the Pre-Processing box (use the specific channels shown in the example).

3. **t-SNE Analysis**

3.1. Press the **t-SNE** button to have the program begin start to compute the reduced dimensionality data set for visualization in the window below the t-SNE button. To save image of t-SNE, press **Save TSNE Image**. On a machine with 8 CPU @ 3.4 GHz each and 8 GM RAM this step should take about 2 minutes for 10,000 events, 10 minutes for 50,000 events, and 20 minutes for 100,000 events.

3.2. To create a 't-SNE heatmap', as seen in several CYTOF publications[10-11], select an option from the **Marker-Specific t-SNE** pop-up menu (use the specific markers CD64 or CD3 as shown in the example). A figure will pop up showing a heatmap representation of the t-SNE plot that can be saved for figure generation.

3.3. Select areas of interest in the t-SNE plots by the user for further downstream analyses using the **Gate t-SNE** button.

4. **Cluster Analysis**

4.1. To begin clustering analysis, select an option in **Clustering Method** listbox (in this example us DBSCAN with a distance factor of 5 in dialogue box to the right of the listbox). Press the **Cluster** button.

4.2. Use one of the following options for automated clustering algorithms found in the 'Automated Clustering Parameters' panel:

4.2.1. *Hard KMEANS (on t-SNE)*: Apply k-means clustering to the reduced 2-dimensional t-SNE data and requires the number of clusters to be provided to the algorithm[12].

4.2.2. *Hard KMEANS (on HD Data)*: Apply k-means clustering to the original high-dimensional data that was given to the t-SNE algorithm. Once again, the number of clusters needs to be provided to the algorithm.

4.2.3. *DBSCAN:* Apply the clustering method of clustering, called Density-Based Spatial Clustering of Applications with Noise[13] that clusters the reduced 2-dimensional t-SNE data and requires a non-dimensional distance factor that determines the general size of the clusters. This type of clustering algorithm is well suited to cluster the t-SNE reduction as it is able to cluster non-spheroidal cluster that are often present in the reduced t-SNE representation. Additionally, due to the fact that it operates on the 2-dimensional data, it is one of the faster clustering algorithms.

4.2.4. *Hierarchical Clustering:* Apply the conventional hierarchical clustering method to the high-dimensional data where the entire Euclidean distance matrix is calculated between all events before providing the algorithm a distance factor that sets the size of the cluster.

4.2.5. *Network Graph*-Based: Apply a clustering method that has been most recently introduced into analyzing flow cytometry data when there are rare subpopulations that the user wants to detect[11,14]. This method relies on first creating a graph that determines the connections between all events in the data. This step consists of providing an initial parameter to create the graph, which is the number of k-nearest neighbors. This parameter generally governs the size of the clusters. At this point, another dialogue box pops up asking the user to employ one of 5 clustering algorithms that is applied to the graph. These include 3 options to maximize the modularity of the graph, the Danon Method, and a spectral clustering algorithm[14-18]. If one wants a generally faster clustering solution, we recommend Spectral Clustering or the Fast Greedy Modularity Maximization. While the Modularity Maximization methods along with the Danon method determine the optimal number of clusters, Spectral Clustering requires the number of clusters to be given to the program.

4.2.6. *Self-Organized Map:* Employ an artificial neural network to cluster the high-dimensional data.

4.2.7. *GMM – Expectation Maximization*: Create a Gaussian Mixture Model using Expectation Maximization (EM) technique to cluster the high-dimensional data.[19] This type of clustering method also requires the user to input the number of clusters.

4.2.8. *Variational Bayesian Inference for GMM*: Create a Gaussian Mixture Model but unlike EM, it can automatically determine the number of the mixture components k.[20] While the program does require a number of clusters to be given (larger than the expected number of clusters), the algorithm will determine the optimal number on its own.

4.3. To study a particular area of the t-SNE plot, press the **Select Cluster Manually** button to draw a set of user-defined clusters. Of note, clusters cannot share members (*i.e.,* each event can only belong to 1 cluster).

5. Cluster Filtration

5.1. Set(s) of clusters identified either manually or via one of the automatic methods described above can be filter via as follows.

5.1.1. To sort clusters (in the **Cluster Filter** panel) by any of the markers measured in the experiment, select an option from the **Sort** pop-up menu. To set whether the order is ascending or descending, press the **Ascending/Descending** button to the right of the **Sort** pop-up menu. This will update the list of Clusters in the 'Clusters (Filtration)' listbox and re-order them in descending order of median cluster expression of that marker. The percentage denoted in the 'Clusters (Filtration)' listbox denotes the percent of the population that this cluster represents.

5.1.2. To set a minimum threshold value for a given cluster across a certain channel, select an option from the **Threshold** pop-up menu (in this example us the marker CD65 and set a threshold at 0.75). Either type a value in the numerical box below the graph or use the slide-bar to set a threshold. Once threshold is set, press **Add Above Threshold** or **Add Below Threshold** to specify the direction of threshold. Once this threshold has been set, it will be listed in the Thresholds box next to the 'Cluster Filter' panel where the marker, the threshold value, and the direction will be listed so the user is aware of which thresholds are currently being applied. Finally, the t-SNE plot will update by blurring out clusters that do not meet the requirements of the filtration and the 'Clusters (Filtration)' listbox will update to show clusters that meet the filtration requirements.

5.1.3. To set a minimum threshold for frequency of a cluster, enter a numerical cut-off in the **Cluster Frequency Threshold (%)** box in the Cluster Filter panel (in this example use 1%).

6. Cluster Analysis & Visualization

6.1. To select clusters for further analysis and visualization, select clusters In **Clusters (Filtration)** listbox and press the **Select →** button to move them to the **Cluster Analyze** listbox.

6.2. To create heatmaps of clusters, select the clusters of interest in the **Cluster Analyze** listbox and press the **HeatMap of Clusters** button. When this button is pressed, a figure will pop up containing a heat map along with dendrograms on the cluster and parameter axes. The dendrogram on the vertical axis will group clusters by those that are closely related while the dendrogram on the horizontal axis will group markers that are co-associated. To save heatmap, press **File | Export Setup | Export.**

6.3. To create a 'High Dimensional Box Plot' or 'High Dimensional Flow Plot,' select the clusters of interest in the **Cluster Analyze** listbox and press either the **High Dimensional Box**

Plot button or the **High Dimensional Flow Plot** button. These plots can be used to visually assess the distribution of given channels of various clusters across all dimensions.

6.4. To show clusters in traditional 2D flow plots, select the transformation (linear, log10, arcsinh) and channel in the **Conventional Flow Plot** panel and press **Conventional Flow Plot.**

REPRESENTATIVE RESULTS:

In order to test the usability of ExCYT, we analyzed a curated data set published by Chevrier *et al.* titled 'An Immune Atlas of Clear Cell Renal Carcinoma' where the group conducted CyTOF analysis with an extensive immune panel on tumor samples taken from 73 patients[11]. Two separate panels, a myeloid and lymphoid panel, were used to phenotypically characterize the tumor microenvironment. The objective of our study was to recapitulate the results of their t-SNE and cluster analysis, showing that ExCYT could be used to come to the same conclusions as well as show additional methods of visualization and cluster analysis.

In the original manuscript, the group described 22 T cell clusters identified by the lymphoid panel and 17 cell clusters identified by the myeloid panel. In **Figures 3 & 4** of the publication, the group shows heatmaps of clusters, t-SNE plots with color-coded clustering solutions, and t-SNE heatmaps in subpanels A, B, & C. In order to perform the analysis, we obtained the manually gated data from Cytobank and sampled 2000 events from each file or took the entire file if it had less than 2000 events, following the analysis pipeline illustrated in the original manuscript. At this point, we sampled a total of 100,000 events via our post-gating subsampling parameter, conducted t-SNE analysis, and used a variety of clustering methods to explore the data in various ways.

First, we examined the myeloid panel by following the same analysis pipeline as the original manuscript by completing the t-SNE analysis and creating heatmaps of the various markers (**Figure 3A**). While the original manuscript normalized the t-SNE heatmaps to the 99th percentile of each marker, ExCYT does not do this type of normalization for its heatmaps. However, similar distributions of marker co-expression were observed as described in the original manuscript. We then applied a Network Graph-Based method of clustering the data by creating the graph with 100 k-nearest neighbors and clustering the graph via optimizing the modularity of the graph by using the Fast-Greedy implementation within ExCYT, where we found 19 sub-populations of cells (**Figure 3B**). When comparing the heatmap of these clusters created by ExCYT with the heatmap published in the original manuscript, we noted that we were able to identify similar clusters of myeloid cells (**Figure 3C**). Of note, the original manuscript identified and contrasted two sub-populations of myeloid cells that we identified in our analysis defined by HLA-DR^{int}CD68^{int}CD64^{int}CD36^{+}CD11b^{+} (Cluster 13) and HLA-DR^{+}CD4^{+}CD68^{+}CD64^{+}CD36^{-}CD11b^{-} (Cluster 18). Visualization by high-dimensional box plot of these two populations revealed statistically significant differences (Mann-Whitney) in the six markers mentioned (**Figure 1D**).

Next, we analyzed the lymphoid panel with a more conventional and faster hierarchical clustering approach. This approach yielded similar marker distributions via t-SNE heatmaps (**Figure 4A**). Furthermore, clustering of the data via hierarchical clustering (**Figure 4B**),

demonstrated similar clusters of lymphoid cells (**Figure 4C**). Of note, we also identified the unique regulatory T cell population from the original manuscript defined as $CD4^{+}CD25^{+}Foxp3^{+}CTLA\text{-}4^{+}CD127^{-}$ (Cluster 17) via our high-dimensional flow plot (**Figure 4D**).

Finally, we wanted to employ a method within ExCYT to quickly and quantitatively assess co-associations among markers. We began by using a hard k-means clustering algorithm to lay down 5000 clusters on the two-dimensional t-SNE data (**Figure 4E**). We then used the median expression of all the markers of all these clusters to create a heatmap from these clusters (**Figure 4F**). Since these heatmaps cluster rows as well as columns that are similar, this method of abstracting the data by applying a fine mesh of clusters and then creating a heatmap allows us to pick up co-associations easily, such as the co-association of Tim-3, PD-1, CD38, and 4-1BB.

FIGURE AND TABLE LEGENDS:

Figure 1: ExCYT Pipeline & Features. (A) ExCYT begins by importing raw FCS data, applying optional compensation, gating, and random subsampling prior to downstream analysis. This ensures all events being analyzed are relevant to the experiment being analyzed. t-SNE dimensionality reduction is then performed to visualize all events and t-SNE heatmaps can be generated to visualize phenotypic distributions. Finally, a variety of clustering algorithms can be applied on either t-SNE transformation or high-dimensional raw data. **(B)** Novel sorting and thresholding features allow users to quickly sort through possibly hundreds of clusters to find ones of interest. **(C)** Heatmaps of clusters can be created to examine how multiple clusters compare to each other as well as which markers co-associate. **(D)** Novel high-dimensional flow/box plots can be generated as a form of back-gating clusters on original data while appreciating the high-dimensional nature of the data.

Figure 2: ExCYT Graphical User Interface: The ExCYT graphical user interface allows for a streamline work flow working from the left to right of the panel as the user imports their data, conducts t-SNE dimensionality reduction, clustering, and final cluster analysis and visualization.

Figure 3: Recapitulation of Myeloid Sub-Populations from Chevrier *et al.* (A) Token t-SNE heatmaps of myeloid panel **(B)** t-SNE plot of myeloid panel color coded by Network-Graph clustering algorithm **(C)** Heatmap of clusters identified by clustering solution on myeloid panel **(D)** Comparative high dimensional box plot comparing contrasting myeloid subpopulations (Clusters 13 & 18) referenced in original manuscript

Figure 4: Recapitulation of Lymphoid Sub-Populations from Chevrier *et al.* (A) Token t-SNE heatmaps of lymphoid panel **(B)** t-SNE plot of lymphoid panel color coded by hierarchical clustering algorithm **(C)** Heatmap of clusters identified by clustering solution on lymphoid panel **(D)** High dimensional flow plot of identified regulatory T cell population (Cluster 17) in original manuscript **(E)** Clustering solution of 5000 cluster hard k-means analysis on t-SNE data **(F)** Heatmap of clusters identified by k-means clustering solution on lymphoid panel showing marker co-associations.

Table 1: Overview of All Functions Present in the ExCYT GUI

Table 2: Overview of Software-assisted Flow Cytometry Analysis Solutions

DISCUSSION:

Here we present ExCYT, a novel graphical user interface running MATLAB-based algorithms to streamline analysis of high-dimensional cytometry data, allowing individuals with no background in programming to implement the latest in high-dimensional data analysis algorithms. The availability of this software to the broader scientific community will allow scientists to explore their flow cytometry data in an intuitive and straightforward workflow. Through conducting t-SNE dimensionality reduction, applying a clustering method, being able to sort/filter through these clusters quickly, and make flexible, customizable heatmaps and high-dimensional flow/box plots, scientists will be able to not only understand the uniquely defined subpopulations in their samples but will be able to create visualizations that are intuitive and easily understood by their colleagues.

While the program is flexible in handling a variety of data types (conventional flow cytometry vs mass cytometry), there are a few considerations for optimal utility of the program. The first of these is regarding the data quality, specifically of flow cytometry data. Proper compensation and resolution of overlapping emission spectra is of paramount importance. Poorly compensated data can inadvertently lead to false co-associations of markers and formation of clusters that are not of true biological significance. Therefore, it is highly advisable that the input data is of sound quality before proceeding with the t-SNE analysis and further downstream analysis. Furthermore, use of the automatic compensation algorithm implemented in ExCYT requires clear single stains for all channels in order to accurately calculate the compensation parameters.

Another important consideration for use of ExCYT is when concatenating multiple FCS files into one analysis (as demonstrated in this manuscript), they must be comparable across all channels. First, this means that the same panel needs to be used across all samples *and* that there is no drift between samples across all channels. For example, if one were to read two samples on separate days and stained CD8 in FITC on both days but the voltage of the cytometer was set differently on one day resulting in a slightly shifted CD8 population, one could generate false clusters in the downstream analysis, as this shift was generated as a function of instrument variation and not due to biological significance. While future versions of ExCYT may be able to normalize samples to their single stains, at this point, careful consideration must be made that FCS files can be compared to each other before importing them into ExCYT.

Finally, the process of clustering is not one that is absolute/rigid. Different clustering algorithms and parameters can generate different clustering solutions. Whether the solution of the algorithm is appropriate is for the user to determine by synthesizing their understanding of the biology with the clustering solution. For example, when understanding the immune environment of tumors, one may be interested in macroscopic clusters (*i.e.,* T cells vs B cells vs Myeloid cells) while another may be interested in subpopulations of macroscopic clusters. The

resolution of the clusters is determined by the user and therefore, no single clustering solution is 'correct.' This is one of the main advantages of using the high dimensional flow plots available in ExCYT. The ability to visualize the distribution of a given cluster across all channels can help the user determine whether they have clustered in not only a biologically relevant way but in a way that is relevant to the scientific question being asked in the experiment. While our goal is to provide a plethora of methods used in the literature to cluster high-dimensional flow cytometry data while providing additional methods of clustering, we recommend using methods such as k-means and DBSCAN to explore the data via quickly iterating on cluster number and size and move towards network-graph and gaussian-mixed model approaches for more robust but more time-consuming approaches.

Given these considerations, ExCYT is still a highly flexible and valuable tool for exploring high dimensional cytometry data, and offers unique/differentiating features than other available packages available to conduct this type of analysis **(Table 2)**. First, ExCYT differentiates itself over most flow cytometry analysis approaches utilizing dimensionality reduction and clustering algorithms by its ability to be used without any scripting/programming knowledge. Additionally, by aggregating many clustering algorithms cited throughout the literature, we believe we provide the most options for clustering data. Finally, our unique feature of cluster filtration and sorting along with display via novel high dimensional flow plots, allows users to explore the characteristics of their clusters quickly and efficiently, making the process of 'discovering' rare subpopulations simple and efficient.

ACKNOWLEDGMENTS:

The authors have no acknowledgements.

DISCLOSURES:

The authors have nothing to disclose.

REFERENCES:

1. Benoist, C., & Hacohen, N. Flow cytometry, amped up. *Science*, **332** (6030), 677-678 (2011).
2. Ornatsky, O., *et al.* Highly multiparametric analysis by mass cytometry. *Journal of immunological methods*, **361**(1), 1-20 (2010).
3. Tanner, S. D., *et al.* Flow cytometer with mass spectrometer detection for massively multiplexed single-cell biomarker assay. *Pure and Applied Chemistry*, **80** (12), 2627-2641 (2008).
4. Maecker, H. T., *et al.* Standardization of cytokine flow cytometry assays. *BMC immunology*, **6** (1), 13 (2005).
5. Brazma, A., & Vilo, J. Gene expression data analysis. *FEBS letters*, **480** (1), 17-24 (2000).
6. Pyne, S., *et al.* Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, **106** (21), 8519-8524 (2009).
7. Ge, Y., & Sealfon, S. C. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics*, **28** (15), 2052-2058 (2012).

8. Venkatesh, V. Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information systems research*, **11** (4), 342-365 (2000).
9. Bagwell, C. B., & Adams, E. G. Fluorescence spectral overlap compensation for any number of flow cytometry parameters. *Annals of the New York Academy of Sciences*, **677** (1), 167-184 (1993).
10. Lavin, Y., *et al.* Innate immune landscape in early lung adenocarcinoma by paired single-cell analyses. *Cell*, **169** (4), 750-765 (2017).
11. Chevrier, S., *et al.* An immune atlas of clear cell renal cell carcinoma. *Cell*, **169** (4), 736-749 (2017).
12. Hartigan, J. A., & Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28** (1), 100-108 (1979).
13. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. Density-based spatial clustering of applications with noise. *International Conference Knowledge Discovery and Data Mining* (Vol. 240) (1996).
14. Levine, J. H., *et al.* Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, **162** (1), 184-197 (2015).
15. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, **2008** (10), P10008 (2008).
16. Le Martelot, E., & Hankin, C. Fast multi-scale detection of relevant communities in large-scale networks. *The Computer Journal*, **56** (9), 1136-1150 (2013).
17. Newman, M. E. Fast algorithm for detecting community structure in networks. *Physical review E*, **69** (6), 066133 (2004).
18. Hespanha, J. P. An efficient matlab algorithm for graph partitioning. *University of California*, 1-8 (2004).
19. Moon, T. K. The expectation-maximization algorithm. *IEEE Signal processing magazine*, **13** (6), 47-60 (1996).
20. Bishop, C. M. *Pattern recognition and machine learning*. Springer. (2006).

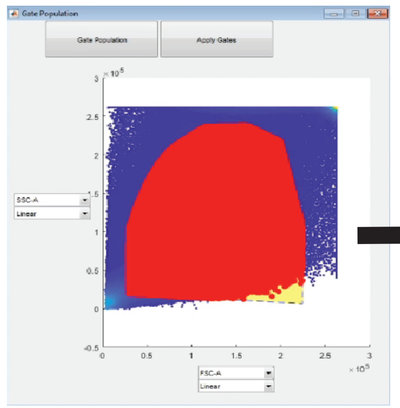
Figure 1

Click here to download Figure Figure 1.pdf

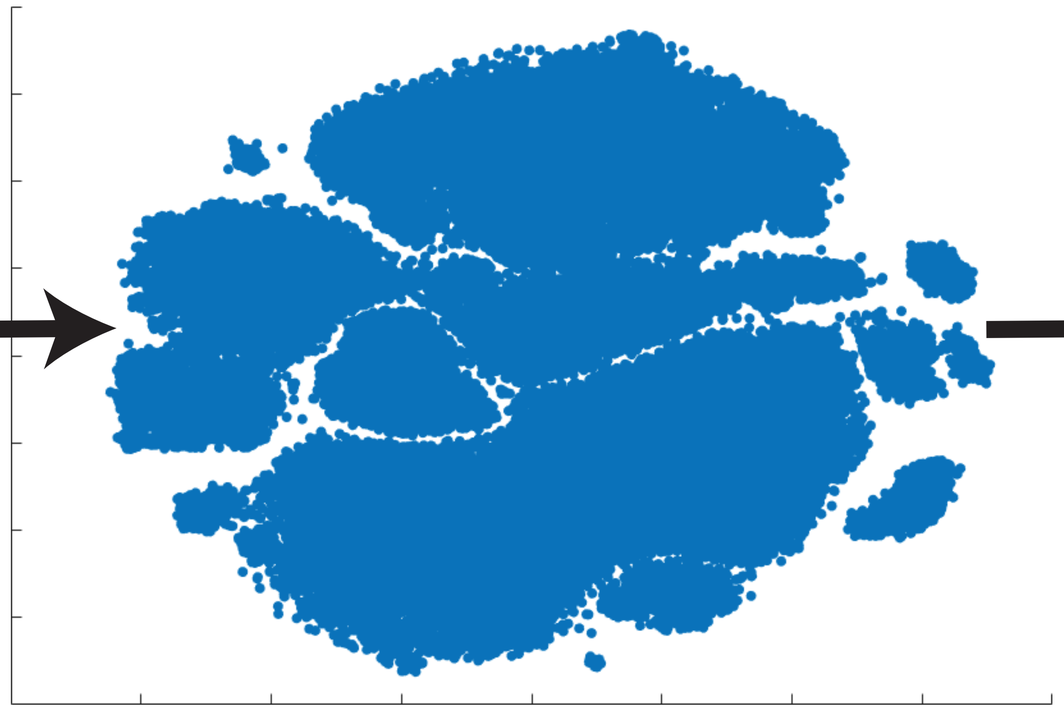
ExCYT Pipeline

A

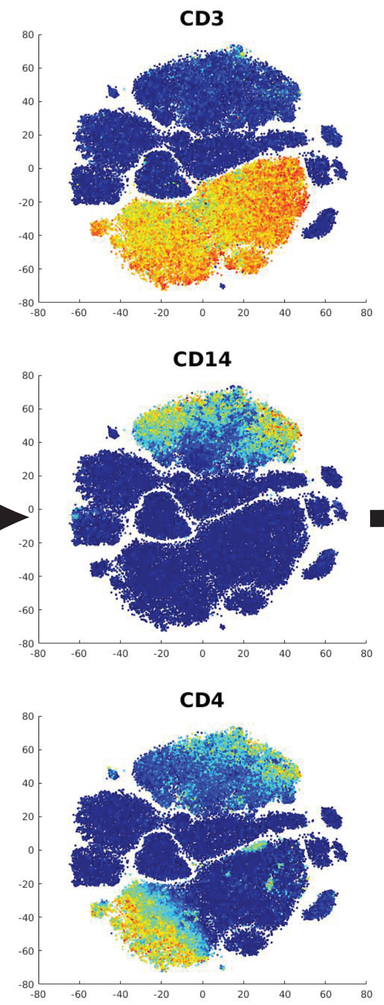
Loading FCS files & Pre-Processing



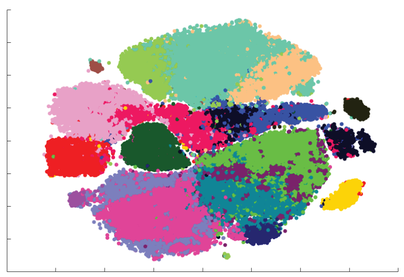
t-SNE Dimensionality Reduction

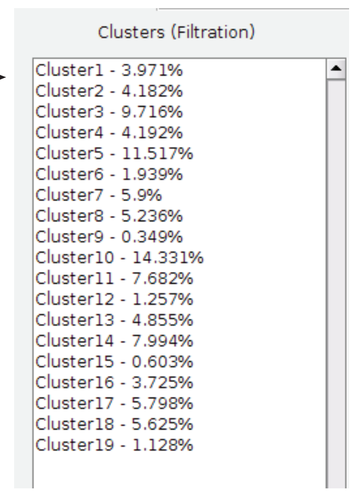


t-SNE Heat Maps



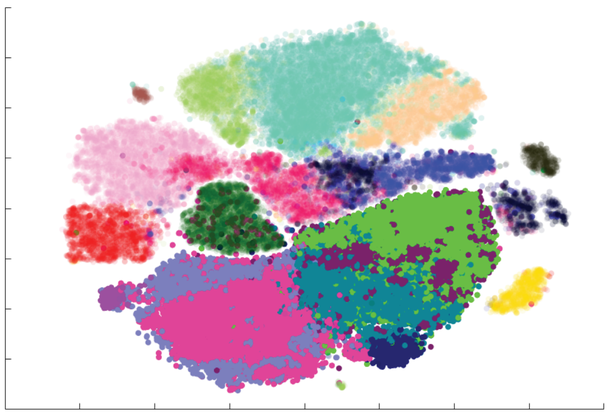
Clustering

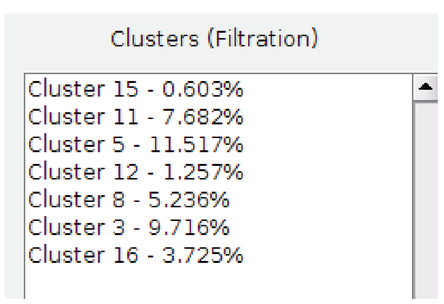




B

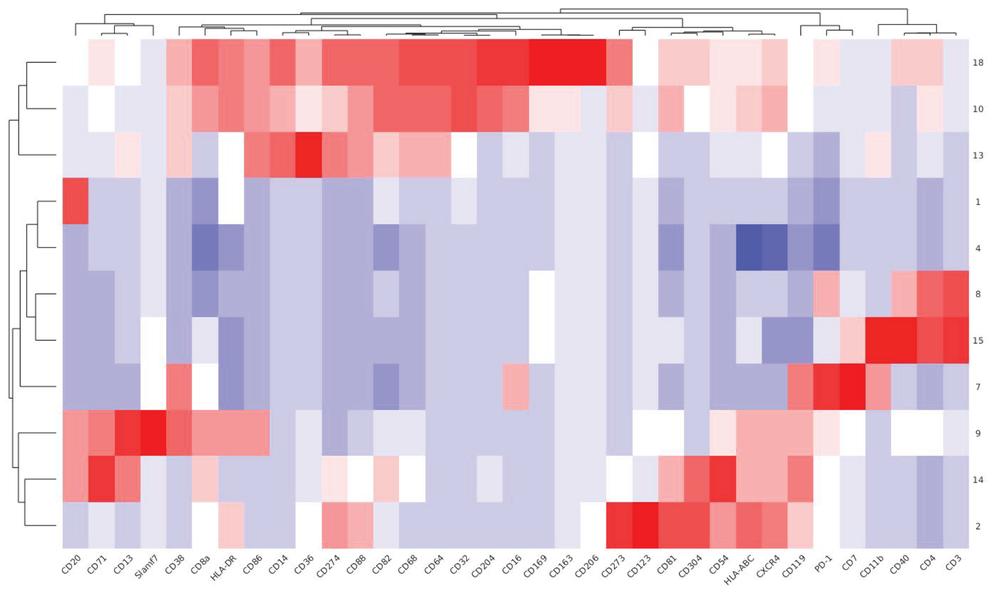
Cluster Sorting & Filtration





C

Cluster Heat Maps



D

High-Dimensional Flow/Box Plots

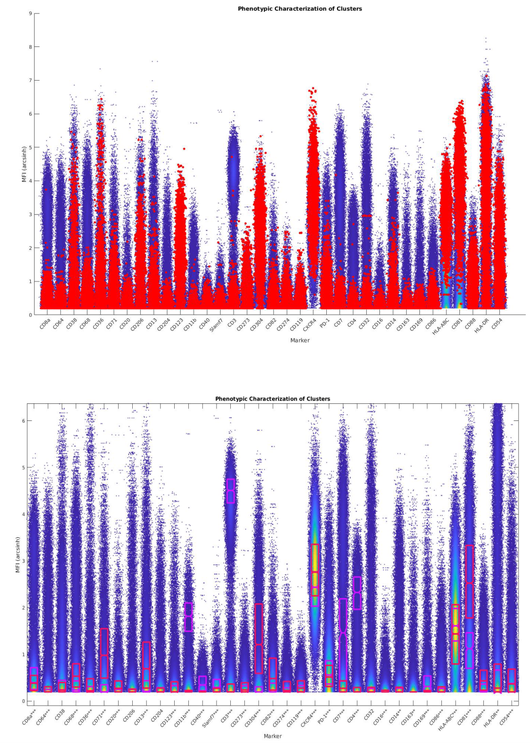


Figure 2

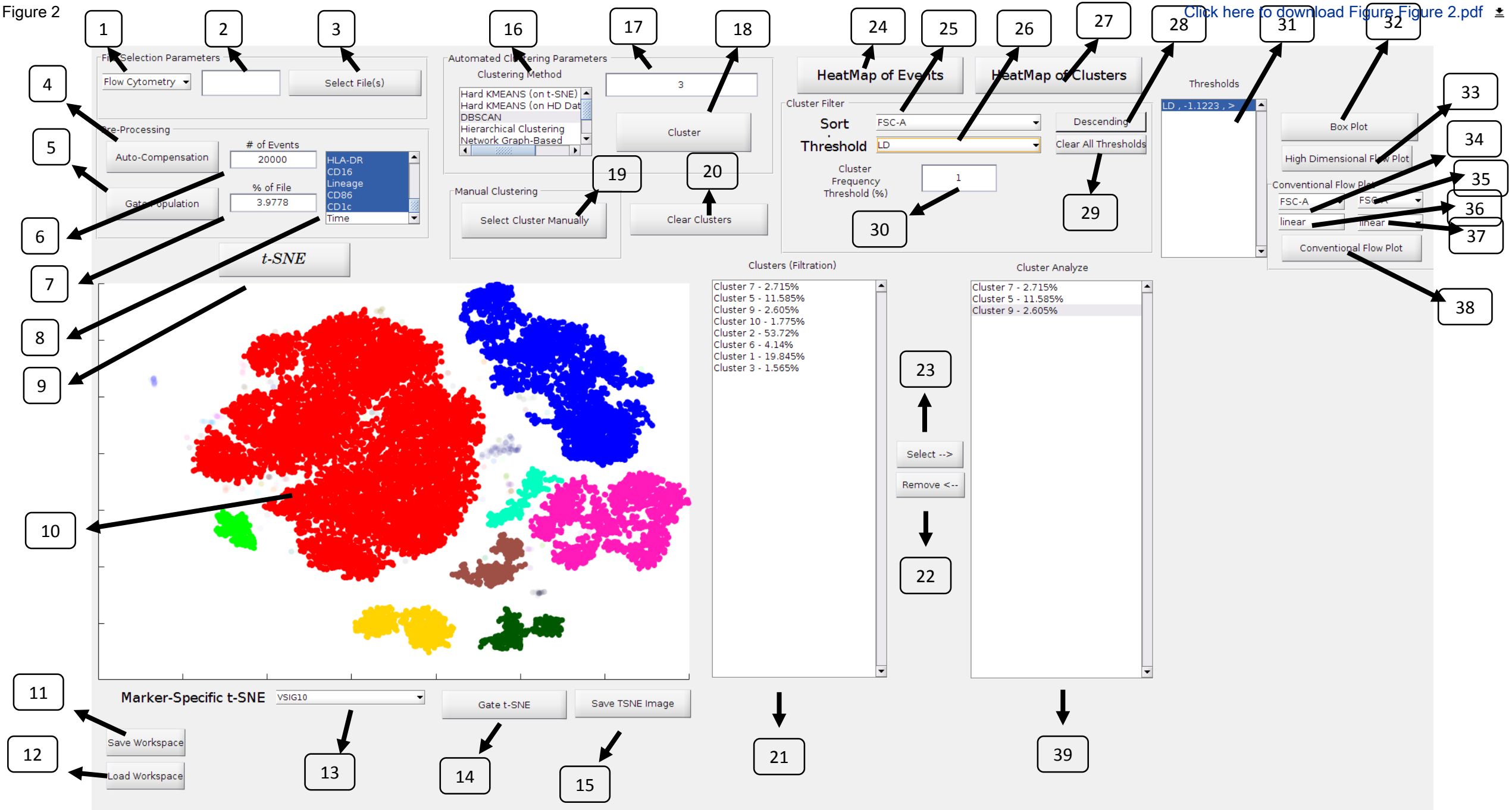
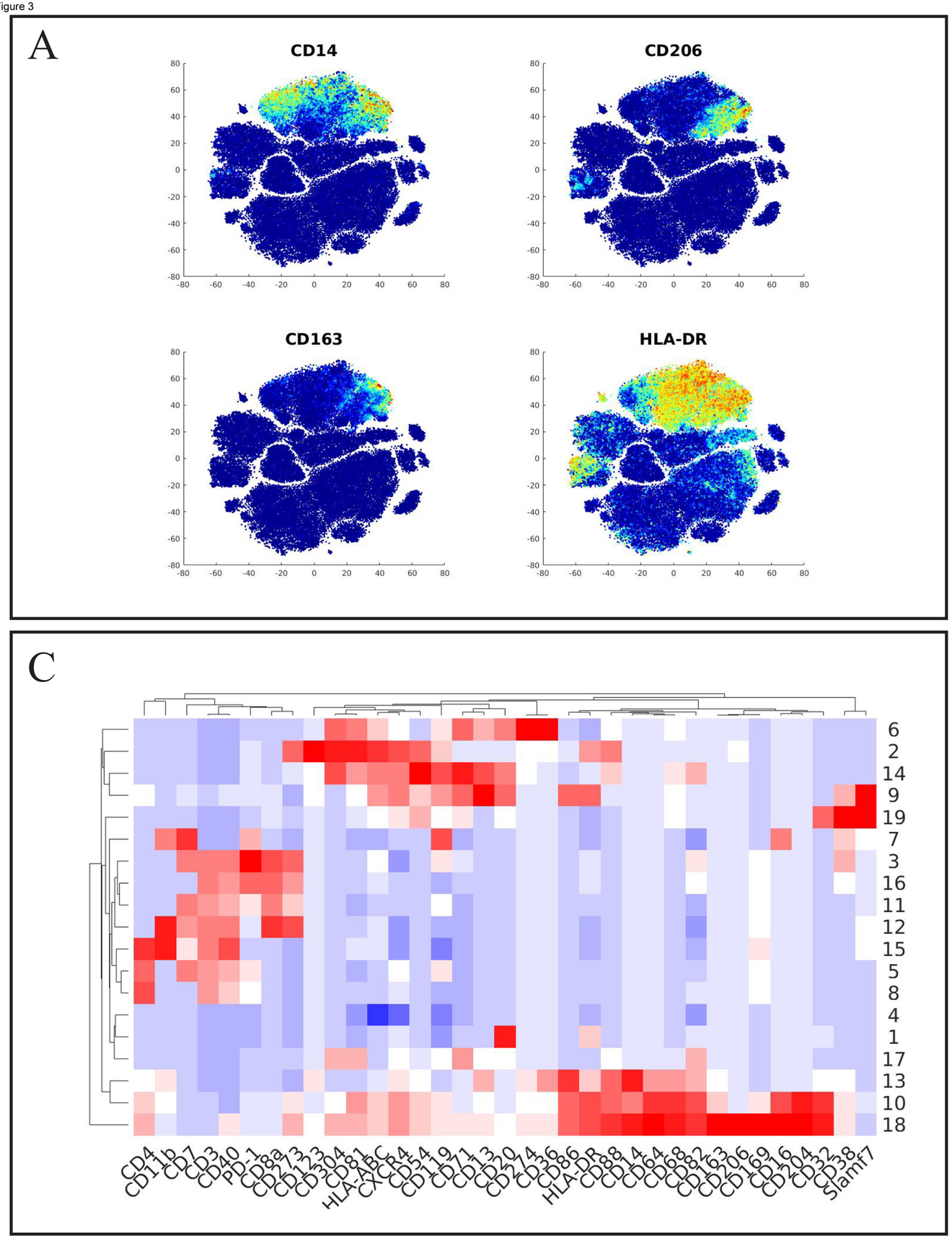
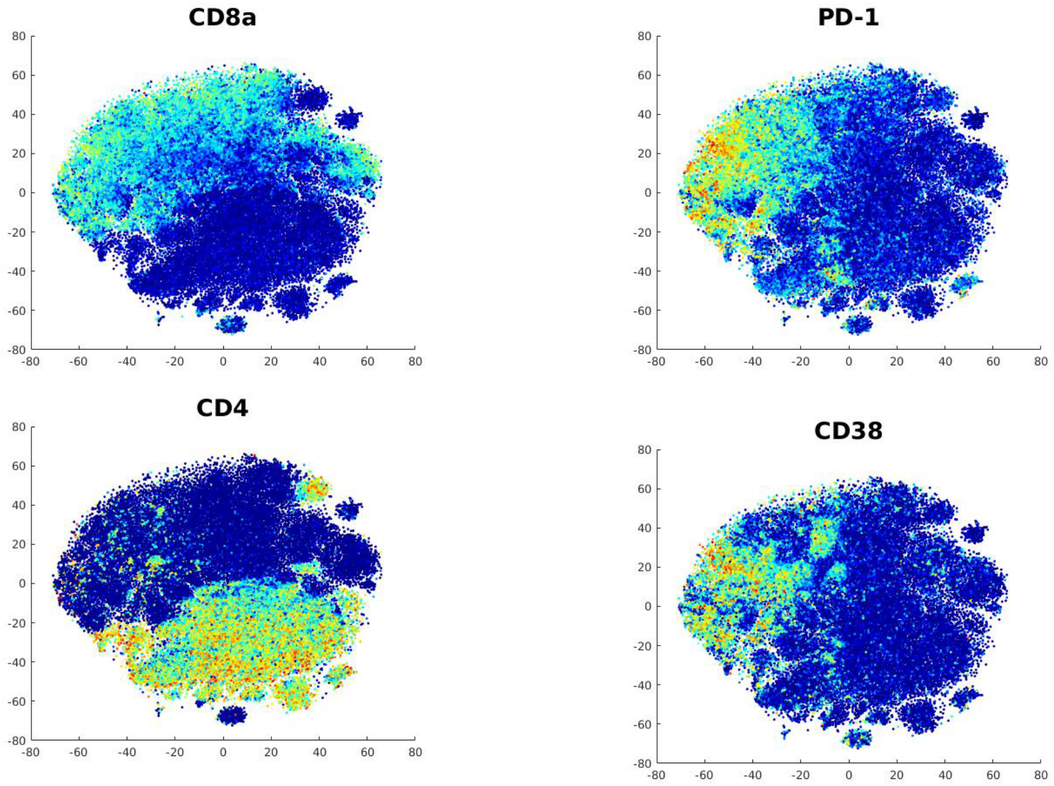


Figure 3

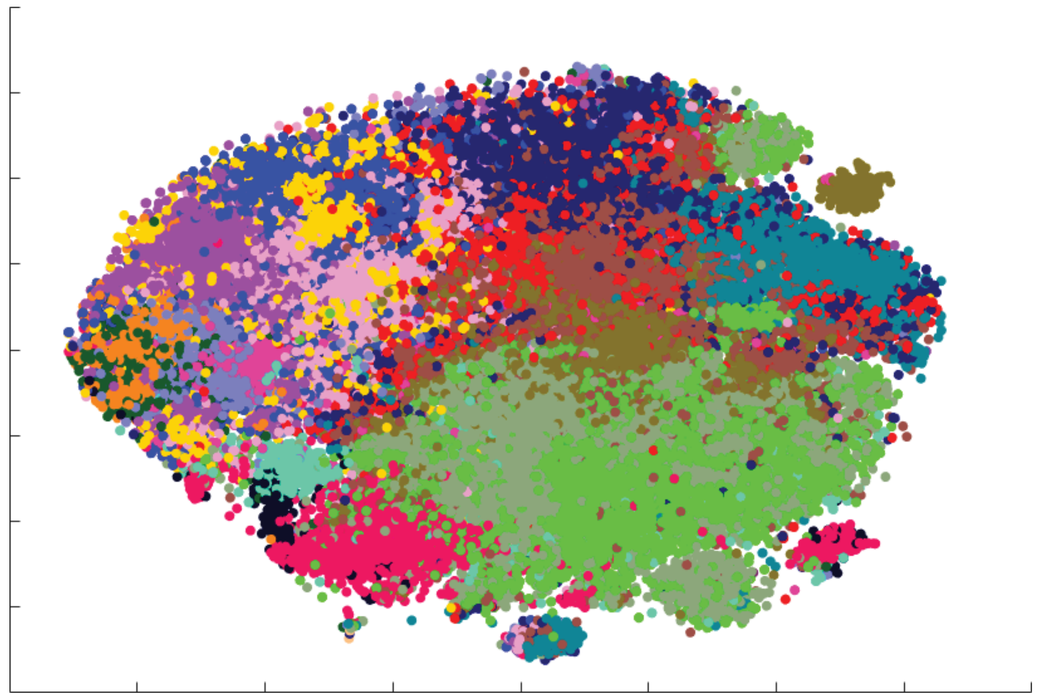
Click here to download Figure Figure 3.pdf



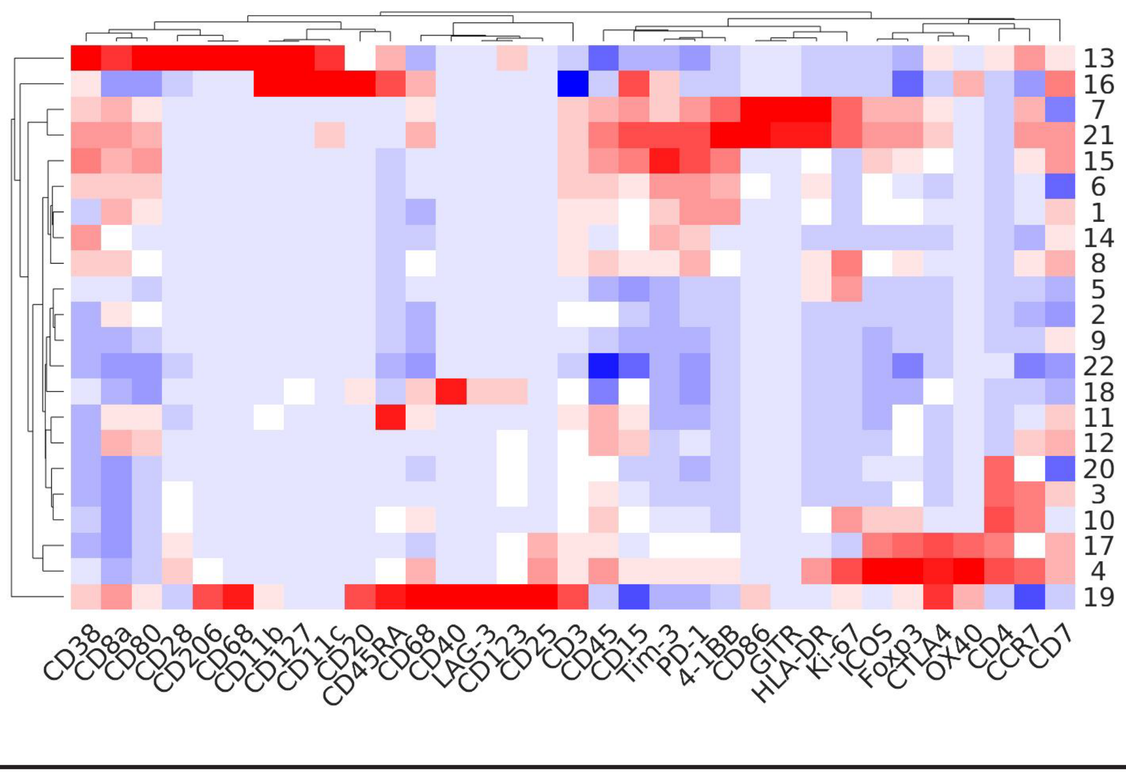
A



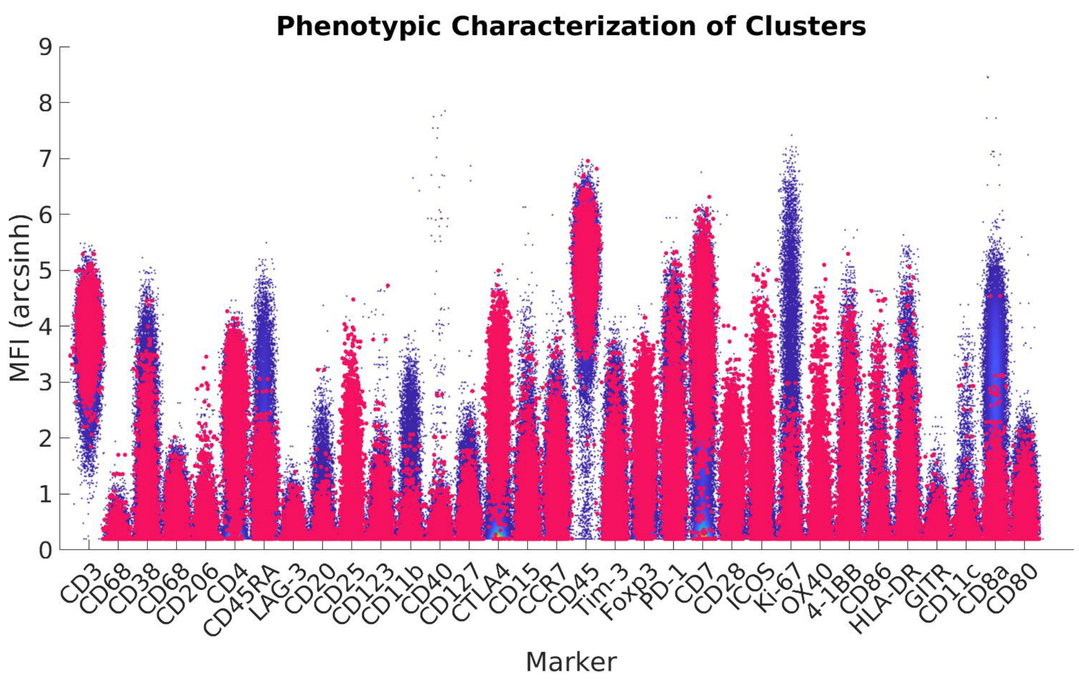
B



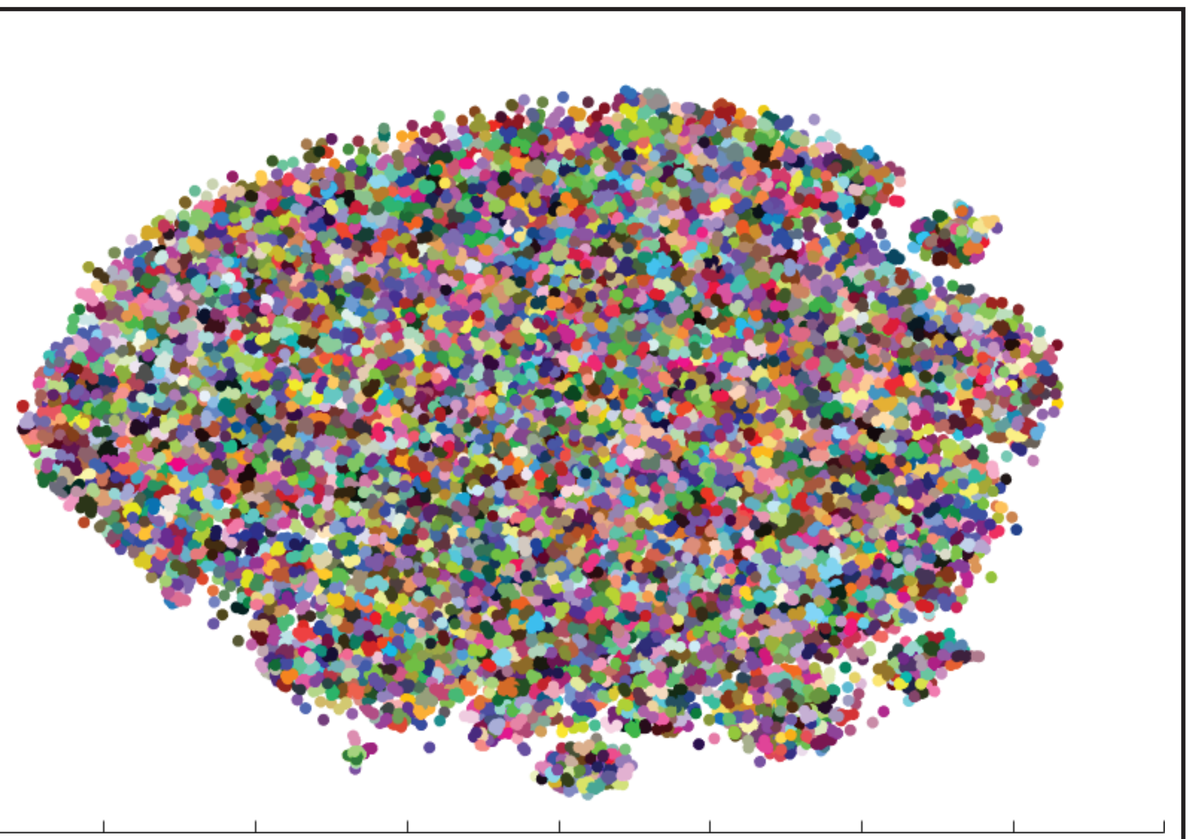
C



D



E



F

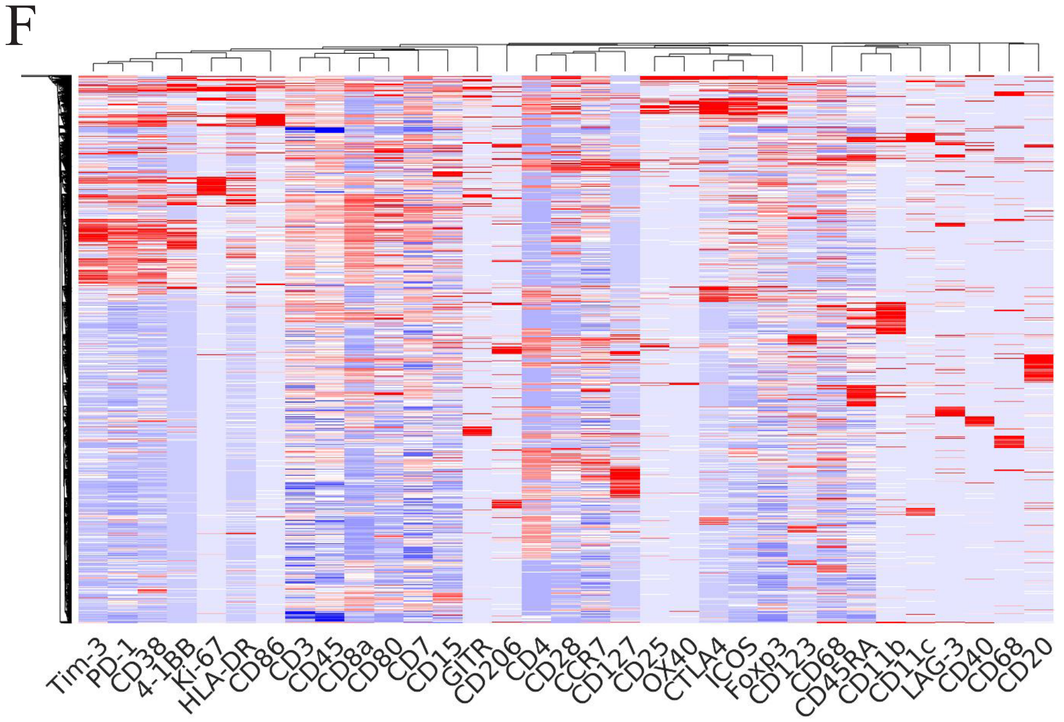


Table 1

Click here to download Table Table 1.xlsx

No.
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39

Description
Select type of Cytometry
Random subsampling of raw data
Select files for analysis
Auto-compensation of raw data based on directory of single stains provided to software
Gating to select events for t-SNE and clustering analysis
Random subsampling of gated data (absolute number)
Random subsampling of gated data (percent of gated population)
Select channels for analysis
Run t-SNE dimensionality reduction
t-SNE Window
Save workspace
Load Workspace
Create t-SNE heatmap on select marker
Gate t-SNE to re-do t-SNE analysis of select population
Save t-SNE window as image
Select Clustering Algorithm
Enter Clustering Parameter for given algorithm
Cluster Analysis
Draw Clusters Manually
Clear All Clusters to redo cluster analysis
Show Clusters under current filter conditions
Remove select clusters from Cluster Analyze listbox
Add cluster to Cluster Analyze listbox
Create conventional heatmap of all events in analysis
Sort clusters by select marker
Set threshold by select marker
Create conventional heatmap of select clusters from Cluster Analyze listbox
Flip order of sort
Clear all thresholds
Set frequency threshold for clusters
List of current thresholds active on 'Clusters (Filtration)' listbox
High Dimensional Box Plot
High Dimensional Flow Plot
Horizontal axis parameter for conventional flow plot
Vertical axis parameter for conventional flow plot
Data transformation for conventional flow plot on horizontal axis
Data transformation for conventional flow plot on vertical axis
Create conventional flow plot
Show Clusters for Analysis

Name (in GUI)
NA
NA
Select File(s)
Auto-Compensation
Gate Population
NA
NA
NA
t-SNE
NA
Save Workspace
Load Workspace
NA
Gate t-SNE
Save TSNE Image
Clustering Method
NA
Cluster
Select Cluster Manually
Clear Clusters
Clusters (Filtration)
Remove <--
Select -->
HeatMap of Events
Sort
Threshold
HeatMap of Clusters
Ascending/Descending
Clear All Thresholds
Cluster Frequency Threshold (%)
Thresholds
High Dimensional Box Plot
High Dimensional Flow Plot
NA
NA
NA
NA
Conventional Flow Plot
NA

Table 2

Name of Software/Package	**ExCYT**
Program Type	Matlab
Price to User	Free
Graphical User Interface	Yes
Dimensionality Reduction Techniques	t-SNE
Clustering Algorithms	K-Means DBSCAN Hierarchical Clustering Self-Organized Map Multiple Network-Graph Based Methods GMM - EM GMM - Variational Bayesian Inference
Ability to Sort/Filter Clusters	Yes
High Dimensional Flow Plots	Yes

CYT	**FCS Express**
Matlab	Stand-Alone Application
Free	$1,000
Yes	Yes
t-SNE,PCA	t-SNE, PCA, SPADE
K-Means GMM - EM Single Network-Graph Based Method (Phenograph)	K-Means
No	No
No	No

flowCore	**openCyto**
R	R
Free	Free
No	No
none	none
none	automation of manual gating workflow
No	No
No	No

FlowMeans

R

Free

No

none

K-Means

No

No

Materials Table

Click here to download Materials Table JoVE_Materials.xls

Name of Material/ Equipment	Company	Catalog Number	Comments/Description
Desktop	SuperMicro	Custom Build	Computer used to run analysis
MATLAB	Mathworks	N/A	Software used to develop ExCYT

1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article: ExCYT: A Graphical User Interface for streamlining analysis of high-dimensional cytometry data

Author(s): John-William Sidhom, Debebe Theodros, Benjamin Murter, Jelani C. Zarif, Sudipto Ganguly, Drew M. Pardoll, Alexander Baras

Item 1 (check one box): The Author elects to have the Materials be made available (as described at http://www.jove.com/author) via: [X] Standard Access [ ] Open Access

Item 2 (check one box):

[X] The Author is NOT a United States government employee.

[ ] The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.

[ ] The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

ARTICLE AND VIDEO LICENSE AGREEMENT

1. Defined Terms. As used in this Article and Video License Agreement, the following terms shall have the following meanings: "**Agreement**" means this Article and Video License Agreement; "**Article**" means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; "**Author**" means the author who is a signatory to this Agreement; "**Collective Work**" means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; "**CRC License**" means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode; "**Derivative Work**" means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; "**Institution**" means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; "**JoVE**" means MyJove Corporation, a Massachusetts corporation and the publisher of *The Journal of Visualized Experiments;* "**Materials**" means the Article and / or the Video; "**Parties**" means the Author and JoVE; "**Video**" means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. Background. The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. Grant of Rights in Article. In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4** and **7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the "Open Access" box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

4. Retention of Rights in Article. Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. Grant of Rights in Video – Standard Access. This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. Grant of Rights in Video – Open Access. This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this Section 6 is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. Government Employees. If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. Likeness, Privacy, Personality. The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

9. Author Warranties. The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

10. JoVE Discretion. If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have

full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

11. Indemnification. The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in *JoVE* or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

12. Fees. To cover the cost incurred for publication, JoVE must receive payment before production and publication the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

13. Transfer, Governing Law. This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to me one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement required per submission.

CORRESPONDING AUTHOR:

Name:	Alexander Baras
Department:	Pathology
Institution:	Johns Hopkins
Article Title:	ExCYT: A Graphical User Interface for streamlining analysis of high-dimensional cytometry data
Signature:	
Date:	10/16/2017

Please submit a signed and dated copy of this license by one of the following three methods:

1) Upload a scanned copy of the document as a pfd on the JoVE submission site;
2) Fax the document to +1.866.381.2236;
3) Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02139

For questions, please email submissions@jove.com or call +1.617.945.9051

Editorial comments #3:

1. Please provide the experimental values in the data processing. I have commented on what values are necessary in the attached manuscript. Please include these values as we need them in order to script and film. We cannot script and film a generalized protocol.

We have re-done the protocol section based on the above instruction and believe that they now better conform to what is described above with respect to the exact values/parameters used in the example.

2. Please use the attached link to upload the ExCYT files and the example data. This will ensure that users will be able to download and access both the software presented in the article as well the example data to fully replicate the protocol:
https://www.jove.com/account/file-uploader?src=17529313

We do not believe we can re-post the data from the site we got the data from directly in JOVE as this would likely constitute and breach of the data use agreement. We have provided the link to were the data for the example can be freely downloaded on line 111.

We generally prefer to use version control capable approaches to release software, such as is freely available in github. We have made ExCyt software freely available at the github link shown on line 99 of the manuscript.

Editorial comments #2:

1. As currently written, it is not clear how to replicate the protocol. Many essential step-by-step details are not included. We cannot film a generalized protocol and we need a specific example in order to film. Furthermore, we need explicit user input commands throughout in order to film, especially for computational protocols in a GUI. It is best to include all user input commands: File | Save | etc. There should be greater emphasis on Figure 2 and Table 1 in how the protocol is performed so the experiment is replicated.

3. The Protocol should contain only action items that direct the reader to do something. Please move the discussion about the protocol to the Discussion. Step 4.2 is only a discussion of the possible clustering methods. We cannot film this.

4. Please ensure that the entire protocol is written in the imperative tense as if showing someone how to perform the protocol. The Protocol should be made up almost entirely of discrete steps without large paragraphs of text between sections. Please simplify the Protocol so that individual steps contain only 2-3 actions per step and a maximum of 4 sentences per step.

The inclusion of more explicit user commands would also help allay reviewer concerns regarding the replication of the protocol.

We have re-done the protocol section based on the above instruction and believe that they now better conform to what is described above

Can the Chevrier data be provided as an example and used in the protocol as the example? This would greatly enhance the reproducibility of the protocol.

Yes, we have added a link to the data in the manuscript (lines 109-111)

2. Additional comments are in the manuscript.

We have addressed these in the manuscript.

5. Please place the superscripted numbered references before the punctuation.

We believe we have corrected these errors now.

6. Please ensure that all graphs have units and labels on the x and y axes: t-SNE heatmaps, Figure 3B, etc.

The axes of t-SNE scatter plots are arbitrary and in arbitrary units, which in many cases not noted in the figure. We noted this in the manuscript and have added more description for the coloring of the data points (lines 102-106) for clarity. We believe this is in line with how these type of data are shown in the literature.

7. Please use the following format for references. Please do not abbreviate journal titles.
[Lastname, F.I., LastName, F.I., LastName, F.I. Article Title. Source. Volume (Issue), FirstPage – LastPage (YEAR).]

We believe we have corrected this now.

Editorial comments #1:

We have attempted to address all editorial comments. We have added clarification to the editorial comments and the changes we made were needed below.

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues. The JoVE editor will not copy-edit your manuscript and any errors in the submitted revision may be present in the published version.

2. Please provide an email address for each author on the first page.

3. Please define all abbreviations before use.

4. Please remove all commercial language from your manuscript and use generic terms instead. All commercial products should be sufficiently referenced in the Table of Materials and Reagents.

5. Please ensure that all text in the protocol section is written in the imperative tense as if telling someone how to do the technique (e.g., "Do this," "Ensure that," etc.). The actions should be described in the imperative tense in complete sentences wherever possible. Any text that cannot be written in the imperative tense may be added as a "Note." However, notes should be concise and used sparingly. Please include all safety procedures and use of hoods, etc.

6. The Protocol should be made up almost entirely of discrete steps without large paragraphs of text between sections. The Protocol steps should contain only 2-3 actions per step and a maximum of 4 sentences per step.

7. Please ensure you answer the "how" question, i.e., how is the step performed? Alternatively, add references to published material specifying how to perform the protocol action.

8. For steps that involve software or analyzing tools, please make sure to provide all the details such as "click this", "select that", "observe this", etc. Please mention all the steps that are necessary to execute the action item. Please provide details so a reader may replicate your analysis including buttons clicked, inputs, screenshots, etc. This is the level of detail we're looking for. Please keep in mind that software steps without a graphical user interface cannot be filmed.

9. We cannot film the steps that are not fully described. Please do not highlight those steps which actions are only explained in the external references. Please keep those steps in the protocol but not highlighted. Please attention that your manuscript after acceptance will be the source for our script writers for the video production.

10. Please avoid usage of phrases such as "could be", "should be", and "would be" throughout the Protocol.

11. Please leave a blank line between all protocol steps as well as Notes.

12. Protocol: 1-6: Please use the imperative tense for all sentences in the Protocol steps. Please avoid usage of phrases such as "could be", "should be", and "would be", "may be", etc. throughout the Protocol. Please revise the protocol text to avoid the use of any personal pronouns (e.g., "we", "you", "our" etc.).

13. Protocol 1.2: Please avoid using personal pronouns in the protocol steps.

14. Protocol: 2.3: Please highlight complete sentences (not parts of sentences). Please ensure that the highlighted part of the step includes at least one action that is written in imperative tense.

15. Protocol: 2.4, 2.5: Please avoid the use of the personal pronouns in protocol steps.

16. Protocol: 3.1: Please specify the computer characteristics.

17. Protocol: 3.2: Please avoid the use of the personal pronouns in protocol steps.

18. Protocol: 4.2.1, 4.2.8: Please move the discussions to either the introduction or the Discussion section. Please ensure that all the protocol steps are written in the imperative tense. Please avoid using any personal pronouns in the protocol steps.

- In regards to details provided with clustering algorithms, this is not really a discussion. We believe that as a user is following the protocol, they should be aware of important details in the choice of the algorithm and should not have to go to the discussion of the manuscript to decide which algorithm to use.

19. Protocol: 4.3: Please move this step to the Discussion section.

20. Protocol: 5.3: Please highlight complete sentences (not parts of sentences).

21. Protocol: 6.1: please avoid using any personal pronouns in the protocol steps.

22. Protocol: 6.3: Please move the discussions to the Discussion section.

23. Protocol: 6.4: Please highlight the complete sentences.

24. After revising the protocol, please highlight 2.75 pages or less of the Protocol (including headings and spacing) that identifies the essential steps of the protocol for the video, i.e., the steps that should be visualized to tell the most cohesive story of the Protocol. Remember that non-highlighted Protocol steps will remain in the manuscript, and therefore will still be available to the reader.

25. Please include all relevant details that are required to perform the step in the highlighting. For example: If step 2.5 is highlighted for filming and the details of how to perform the step are given in steps 2.5.1 and 2.5.2, then the sub-steps where the details are provided must be highlighted.

Figure and Table legends:

26. Figures 1-4: Please add the corresponding colorbars to the graphs. What is each cluster?

- We tried multiple ways to denote cluster identification. Ultimately, since the number of clusters can get very high, and as we have seen in literature, color bars tend to be not very helpful in tracking clusters to t-SNE plots, we decided instead within the program for a user to identify a cluster by scrolling the mouse over the actual cluster. When one does this, the cluster number comes up. Also, when the user creates high-dimensional flow-plots, the color used in the plot is the same as the one used in the t-SNE plot. This has been how we have chosen to ‘color-code’ the outputs of the program. In a most recent submitted publication, our colleagues found the combination of t-SNE plots and high dimensional flow plots, they were able to accurately convey cluster identity.

27. Figures 1-4: Please add a label and the corresponding unit to each axes.
    - The units of t-SNE analyses have no physical meaning. As t-SNE is dimensionality reduction technique, compressing 48 dimensions with units to 2 without dimensions means we cannot include units to axes.

28. If you are reusing figures from a previous publication, you must obtain explicit permission to re-use the figure from the previous publisher (this can be in the form of a letter from an editor or a link to the editorial policies that allows you to re-publish the figure). Please upload the text of the re-print permission (may be copied and pasted from an email/website) as a Word document to the Editorial Manager site in the "Supplemental files (as requested by JoVE)" section. Please also cite the figure appropriately in the figure legend, i.e. "This figure has been modified from [AUTHOR] et al.[REFERENCE]".

29. Please ensure that the references appear as the following: [Lastname, F.I., LastName, F.I., LastName, F.I. Article Title. Source. Volume (Issue), FirstPage – LastPage, doi: DOI (YEAR).] For more than 6 authors, list only the first author then et al.

30. Please abbreviate all journal titles.

31. Please revise the table of the essential supplies, reagents, and equipment. The table should include the name, company, and catalog number of all relevant materials in separate columns in an xls/xlsx file. Please list all the materials, equipment, instrument, and software used in your work.

Reviewers' comments:

We have address each reviewers comment below.

Reviewer #1:
Major Concerns:
The manuscript describes a pipeline for analysing multidimensional flow cytometry data, implemented as a "Software package" named Excyt.
While the manuscript seems to give a good description of how to use the pipeline, it is impossible to verify and use the software.
I see no link where it can be downloaded. It has not been published previously, as far as I can tell.
If this manuscript is the publication for this software, some details about implementation (in Materials and Methods) should be provided, as well as information
how to obtain it, e.g. a web page. At present, it is impossible to judge the software. At present only a Table 3 in Excel format exists, which can be dropped
anyway and incorporated into the main text, given how little info it has.
Also, I presume a video will be made how to use the software. It would be
nice if that were available so as to judge how well the pipeline is described (in words and video).
Overall, the paper looks good, but in the absence of the software, it is impossible to really review it.

Minor Concerns:
Name of the Package "Excyt". When searching for this, a lot of hits come up with "Excyte", a flow cytometry dedicated web site.
Something more unique might be better.
Table 1 and 2, especially table 1, should and can also be provided as PDF. I.e. what are the explanations of the Software User interface

made as Excel file? A 1-2 page word/pdf document would be better, text within cells can also be wrapped, doesn't need to be all single rows.

- While we understand that the name may be confused with other types of cytometry services, there is no software package that is termed ExCyt or a variation of it and since we have had a few people already use the software and know it by 'ExCyt,' we would prefer to keep it as this. In regards to how we provided information on how to use the software, we used a template from a previously accepted article in Jove that was describing a MATLAB GUI. (https://www.ncbi.nlm.nih.gov/pubmed/28745622)

COMMENTS AFTER REVIEWING SOFTWARE:
First about the "exe" file.
In the manuscript the authors state:
"80 we have
81 developed ExCYT, a graphical-user interface (GUI) that can be easily installed on a PC/Mac that
82 pulls many of the latest techniques including dimensionality reduction for intuitive visualization,
ExCYT is a graphical user interface built in MATLAB and therefore can either be run within
86 MATLAB directly or an installer is provided that can be used to install the software on any
87 PC/Mac. "

If the authors can only provide an .exe file for PC, this sentence is misleading (as it requires at least MatLab), and it cannot be installed without MATLAB on "any PC/Mac".

- We have modified the text to reflect that ExCYT is only available right now as a MATLAB script and PC executable. Many well known scientific packages are written to be only supported by one operating system in binary format (with source code provided) and we think this is acceptable, given we are providing this as a free tool, largely to academics who will either have a PC computer or a MATLAB license through their university.

Regarding the instructions in the manuscript.
Alas, with the information given there, I was not able to generate any plots.
Part of the problem is that text is overwritten in the interface, here is a screenshot.
Changing font preferences in MATLAB does not help.

Following the instructions (after importing facs files) I come to
"Finally, the user will be asked to select the channels they want to use for the analysis.
Generally, we recommend using stained channels as the ones to perform the downstream analyses on.
However, if the user believes the physical parameters of the cells are important to them, they can also select the forward and side-scatter channels."
But when clicking on the "t-SNE" button, nothing happens.
I cannot find any place where one is ask to select the channels one want to use.

At this point, the description is not adequate.
The authors should go through the whole manuscript with a user that has never used the software, and provide all steps necessary to really get this to work.

- Without screenshots and command line output from MATLAB, it is really difficult for us to address this bug as we cannot replicate it ourselves. We have given the program to several of our colleagues within the lab to use the software and they have not had these issues. They have been able to run through whole analyses easily without any major problems.

Reviewer #2:

Manuscript Summary:

Sidhom et al present a new cytometry data analysis software, ExCYT, that is a graphical user interface for MATLAB. They claim to offer a uniquely accessible tool for high dimensional flow cytometry data analysis.

Major Concerns:

The issues with the manuscript and the software fall broadly under three categories: 1) the rationale presented for development and use of ExCYT is severely dated or simply inaccurate; it likely reflects the authors limited experience with cytometry data analysis tools and their history. 2) the software is not really that unique. 3) its accessibility and usefulness will be compromised by how slow it is, how buggy it is, and the fact that it doesn't account for real world differences in how flow data will be presented to the software (it is not flexible enough to be used in different places, and doesn't have good error handling/user interaction).

- While our experience maybe different than the reviewer's, we do believe the area of cytometry is changing dramatically in the utilization of high-dimensional assays such as CyTOF and more capable flow cytometers, able to stain up to 48 colors simultaneously. Therefore, we believe the current state of the art in terms of analysis packages has much room for improvement and innovation in terms of how to analyze the larger dimensionality of these newly generated datasets. While there are implementation of t-SNE and clustering methods within other software packages, we believe our novelty does not necessarily lie within these capabilities but within how to explore the clustering solutions through ExCYT's ability to sort/filter clusters by user preferences and visualize their characteristics through heatmaps and novel high-dimensional flow plots. These features have already been appealing to colleagues who have already used ExCYT to create visualizations for presentations and have submitted manuscripts with ExCYT visualizations.

Specific issues:

1). Long abstract: Authors assert that it is difficult to analyze data objectively with traditional manual

gating. This is incorrect. Manual gating is actually the gold-standard against which automated approaches are compared, manual gating steps are key parts of high dimensional cytometry workflows that ensure compensation is correct, that antibodies are working, etc. So, the problem is not that manual gating is inaccurate, it is that it is inconvenient, and it is challenging to visualize cell populations in multidimensional space using 2D plots and manual gates.

- We appreciate the reviewers comment and their insights. However, we feel manual gating is by the nature of human annotation not an objective data driven process; however, we accept that despite this difficulty experienced flow cytometry analysts can produce reasonable results and interpretations from high dimension data given adequate training. Additionally, it would important to note that we have geared this tool to analyze data from technologies like CyTOF with 40+ channels (as demonstrated in the manuscript) for which manual gating is likely to be problematic.

2). Long abstract: Authors state that methods for high dimensional data analysis are not available in easy to use packages. This is fundamentally wrong. tSNE is implemented as an extremely to use platform in FlowJo v9. SPICE is a tool, available for many years, that also performs statistics for high dimensional data analysis. FlowJo v9 also has multiple easy to use clustering and binning tools. Gemstone and FCS express offer unique and easy implementations of other data visualizations. Further, it is trivial to export flow data to a text file and analyze it with GUI for graphing and statistical software (like JMP, and other MATLAB plugins). Cyotbank and Cytoflex also offer easy visualization. So the authors fundamental rationale for this tool - that it is an unmet need - ignores a myriad (really a myriad!) of tools that address high dimensional cytometry.

- We have included all these in the manuscript for completeness as recommended by the reviewer. However, we feel that in terms of opensource / freely available software we have yet to see a tool that allows the user to not only implement these routine analytics such as t-SNE and clustering algorithms but also provides a novel interface to sort/filter these clustering solutions while visualizing their complete characterization via heatmaps and novel high-dimensional flow cytometry plots. While nuanced as well (and reflected in Table 2), many of these packages that offer the ability to cluster usually just provide a simple K-means clustering algorithm which assumes ‘spheroidal’ clusters in high-dimensional space. This often is a simplistic clustering algorithm which fails to capture the non-spheroidal nature of the data in high-dimensional space. ExCYT provides alternatives to K-Means such as graph-based and gaussian mixed-models methods to better deal with the nature of high-dimensional single cell data. However, despite what we believe are improvements to the clustering options to the end user, we do believe our true novelty lies in ExCYT’s ability to explore the nature of these clusters in ways not provided by other software packages.

3). Long abstract: The authors claim that their compensation calculations will be more accurate than the instrument-calculated ones. What is the basis for this assertion?

- We have softened this comment in the manuscript by stating that we apply a method for more objective calculation of the compensation matrix as we determine the positive population in an algorithmic manner as oppose to asking the user to define the positive population manually.

4). Introduction: "manual gating (has) sufficed for experiments (with) a few markers ... that give visually discernible populations." This is incorrect. Manual gating works also for markers that are not clearly separated from negative, provided that proper controls are run. Manual gating is the basis of many data analyses from experiments using even 8-30 different markers (not just a few markers). Again, the problem is not necessarily gating, it is having a means to interpret and analyze the data.

- We have modified this section of the manuscript to clarify that manual gating is more straightforward when there are visually discernable populations as these populations do not need

isotype controls to determine what is positive/negative. However, we do maintain that in higher dimensional space, manual gating is problematic especially since many of these markers in large panels are often not canonical lineage markers but rather functional markers that will often have unimodal, gaussian distributions where isotype controls are needed to place appropriate gates in order to detect biologically meaningful shifts. As panels get larger and larger, as large as 48 colors, isotype controls for the majority of these channels would quickly become overwhelming and infeasible. As can be noted by many recent publications where high-dimensional single cell data is analyzed, such as the one used in this manuscript, manual gating is only used to identify live cells while the rest of the downstream analysis is done via automated clustering algorithms.

5). Introduction: "manual gating has failed to generated reproducible results..." This is an ABSURD statement. Manual gating has been used for years, on multiparametric datasets in fact. In fact, one might argue that newer high dimensional clustering algorithms are less reliable (or reliability has not been demonstrated). The GOLD STANDARD for gating is manual gating. Are authors aware of flowCAP effort? It compares automated algorithms to the consensus manual gating of humans, which has always been deemed the gold standard. How could automated or clustering-based methods be reliable any way if they are not comparable to each other, and often don't separate biologically distinct populations? The authors own data and the design of their software (to include many different clustering approaches) actually argue against their own point.

- In order to address the reviewer's concerns with the phrasing of the sentence, we have rephrased it to say, 'manual gating **can** fail to generate reproducible results.' This being said, the publication that was referenced that demonstrated this problem was on a very low dimensional flow cytometry measurement where the result should not be variable (measurement of quantity of antigen-specific T-cells based on low number of markers). We used this to highlight the problems with manual gating on a simple problem. The author is correct, we do cite that the clustering process in itself can yield different results when using various algorithms. However, this is stated in context of the type of data ExCYT is designed to analyze (high-dimensional flow cytometry geared at 20+ channels like the BD FACSymphony and CyTOF) where many functional markers can allow different clustering solutions depending on the algorithm employed. The intention of this statement in the manuscript is to highlight the problems with manual gating even in a simple setting and the need for more objective methods of segmenting flow cytometry data.

6). Introduction: the article referenced to support the claim that gating introduced variation is old. Closer analysis of this data, using a more modern interpretation, is that the instruments could not be sufficiently standardized to allow consistent manual gating. Modern automated gating or clustering approaches - which were not tested in that article - would actually fare worse.

- As previously mentioned in the prior question, we use this article to highlight the problem with manual gating approaches and the need to move towards more automated workflows (as other manuscripts on this topic have).

7). Introduction: The authors mention gene-expression studies as a model for a new generation of cytometry data analysis tools. They fail to recognize that most of the experience with those tools has been on bulk approaches rather than single cell technology. This is a fundamental difference... tools generated for bulk analysis ignore the added power and complications of single cell analysis. In fact, single cell gene expression studies rely on approaches developed for cytometry data originally. There is a fundamental lack of understanding of the landscape and history in the rationale they present for their tool.

- Respectfully, we are acutely aware of this distinction between single cell data and data derived from homogenates of cells. I think it is important to understand that these algorithms, while initially applied to bulk gene expression data, are equally applicable in the single-cell scenario with a proper understanding of the algorithms. In the case of the manuscript, t-SNE and clustering methods are initially used at the single-cell level to abstract the data into higher-concept clusters,

on which traditional heatmaps are then applied to understand the cluster to cluster relationships. This is in fact a workflow adopted by many manuscripts, including the one used as the demonstration for ExCYT.

8). Introduction: Authors claim that high dimensional bioinformatic approaches for flow cytometry data remain largely unused. This is an overstatement... I could probably come up with a list of 20 papers doing this.

- While there are many high-impact manuscripts that utilize these advanced techniques, there are thousands of more papers that use simple gating strategies and delineate these in order to quantify and track populations of cells between samples. Furthermore, day to day analysis within the lab environment typically does not consist of using these more advanced algorithms to analyze data. In fact, these types of analyses are usually conducted by a bioinformatics team when a publication is being assembled and not part of a routine data analysis pipeline done on a daily basis. Our hope is that ExCYT is easily enough used that it could become a tool for daily and easy to use to explore data right off the cytometer.

9). Introduction: The authors reference a paper from 18 years ago (!) to explain the reasons why high dimensional approaches are supposedly under-used in the literature. 18-years ago was ages ago... the most advanced cytometry labs were still working in no more than 10-12 color space, not 30-color space as they are now. Flow cytometry was the only single cell technology back then, not as now where it is one of dozens.

- This particular reference, while indeed is an old one is not just commenting on the use of flow cytometry but rather attempting to understand the acceptance of new technologies. The concepts this particular manuscript is describing are generalizable to understanding the hurdles in disseminating new technologies.

10). Introduction: Interestingly, the authors cite flowjo but do not include in the table comparing exCYT to other tools. I suspect this is because many of the features of exCYT are available in FlowJo and better implemented.

- In fact, we did include Flowmeans (a flowjo add-on) which is a limited cluster package for FlowJo relative to what we present in the manuscript. We would interpret the above comments as referring to this particular FlowJo add-on as the base FlowJo is unlikely to be adequately geared to handle very high dimensional flow data. Furthermore, our tool was developed because while we have licenses for FlowJo in our lab, it proved inadequate to do the depth of the analysis we needed, and therefore, developed ExCYT first for ourselves and then considered that it might be of interest to others. We noted and compared to FCSexpress that had a similar profile of add-ons with respect to SPADE, t-SNE, PCA, etc. Again, the salient distinction in our tool is the ability to explore, characterize, and visualize the clusters, not just perform the clustering step.

11). Protocol: How does the software handle errors generally, what feedback is provided to users? I found this lacking in my use of the software.

- We accept this was the first version of the software and have since implemented and will continue to make more robust implementations while maintaining a current Github repository (as can be seen in the latest version of the software).

12). Protocol: The authors do not use the standardized method for naming detectors on the cytometer, so they cannot account for when people use different synonymous dyes across different experiments. This will complicate how compensation tubes are imported into the tool. I expect the software will have trouble across different types of cytometers or different experimental designs.

- We don’t explicitly name the detectors within the ExCYT package. They are imported from the name metadata within the FCS files.

13). Protocol: What is the method for down-sampling? If it is random, then wouldn't evaluation of low frequency populations by exCYT be hampered?

- While true, this is a universal issue with physical and computational down-sampling. We will consider more robust technique to re-introduce rare population in later versions of the software; however, we focused on clustering and cluster exploration, characterization, and visualization for this manuscript and the version of the tool associated with it.

14). Protocol: Only a single gate can be set for compensation tubes... if the population shifts or if different beads or cells are needed for compensation, compensation will be wrong.

- The single gate is set on the fsc/ssc gate, where we would not expect shifts if the samples were run at the same time on the cytometer.

15). Protocol: The authors assert that a threshold at 99th percentile is objective. How is this objective? It is actually consistent more than objective, and it is not really necessary because of the mathematics of compensation.

- We changed the wording to reflect that this threshold is consistent, rather than objective across samples. We do disagree however that it is necessary to set a consistent gate when selecting the positive population for compensation. It directly affects the computation of the compensation matrix as it can change the average value for the population which is used to measure the ‘bleed’ across channels.

16). Protocol: The times claimed for tsne analysis are really slow. FlowJo is much faster (on the same computer). I tested tsne time on exCYT and found it to be very slow, with poor feedback about progress.

- It is unclear what comparison in compute time is being made because of difference in gating and down-sampling using in FlowJo vs what ExCYT did for downstream t-SNE. We accept that the algorithm speed could be improved; we will consider for later version of the program; however, this is the standard t-SNE implementation within MATLAB which has not been reported as being “slow” by others colleagues that have been using it fairly routinely.

17). Protocol: There are a lot of parameters that customize tSNE that are not available in this implementation.

- We have added distance metric and perplexity to the user interface, with suggested default values.

18). Protocol: I have never seen anyone cluster tsne data, and I have reviewed many applications of tsne and algorithm development that includes tsne. Essentially the tsne process is already clustering the data. Populations should be easy to identify at this point - if tsne worked well and was the best tool for the dataset - simply by gating or overlaying markers. There is no need for additional clustering.

- Respectfully, t-SNE is a dimensionality reduction technique and not a clustering solution, typically used for visualization purposes. The notion above regarding "no need for additional clustering" is problematic. Again, we focus this tool on the exploration, characterization, and visualization of clustering solutions as derived from high dimensional representation of the data and lower dimension representations such as t-SNE.

19). Protocol: The number of choices for clustering is good, but they don't provide data showing performance of each on a multi parameter dataset. How do I know other clustering methods work and

what their output looks like? Also, when I used software tsne graphic never showed up, nor did the other clustering tools give any output.

- Given what is known about clustering algorithms, especially as applied to single-cell high-dimensional data, there is little yet known about general optimal parameters that can be applied across all experiments. If so, we would not give the user options to use various clustering algorithms and thresholds for those algorithms. We believe the evaluation of a clustering solution needs to be informed by an understanding of the biology, and this is why the exploration of these clusters is vitally important to implement a clustering solution. This is why we have focused the attention of the algorithm to exploration, characterization, and visualization of the clustering solutions because we believe ultimately, the scientist doing the experiments needs to guide the analysis with their understanding of the experiment and the biology. As for the bug the reviewer is citing, it is unclear to us where this is happening. We have not experienced this issue. If the user is running ExCYT in MATLAB, we would like to see the command line output for when they attempt to cluster the data.

20). Figure with ExCYT pipeline: The initial pre-processing gate (A) makes no sense at all. It appears to be "selecting" almost all of the events, without regard to any underlying distribution, which can't even be seen in the histogram.

- This figure is largely just meant to be a schematic; interestingly it does highlight our perceived difficulties with gating in high dimensional data. In our hands, we use a liberal gating strategy at first in the FSC/SSC domain, and let the dimensionality reduction and clustering algorithms remove noise, if it is present.

21). Figure with ExCYT pipeline: I don't see good separation in Tsne between CD4 and CD8 cells which all appear on the same "island" in the visualization. Without information on expression of each cluster (which is not visible at this resolution) I can't judge accuracy of clusters. Notably though, some clusters are dispersed widely in different parts of Tsne plot (like the purple one). So, they are clusters by clustering algorithms but not by tsne. which is right? I think you need a dataset with more clearly defined cell types to first validate this software, then test it on more complex datasets. For informatics experts like the authors, I am very disappointed to see a lack of attention to validation.

- The example the reviewer is citing is in regards to a CD3+ population of T-cells. Given the fact that this is whole tumor data, in regards to the heterogenous nature of the tumor (myeloid, tumor, stroma, and lymphocyte cells), the algorithm will naturally create these large islands first by major cell type. This is seen commonly in many publications, including the one we pulled this data from. At the risk of detracting from the novelty in this manuscript, we are not really developing a new algorithm, and therefore do not see the need to approach the manuscript with that rigor. Instead we are developing a tool to aid in the exploration, characterization, and visualization of high-dimensional cytometry with currently established techniques in a reasonable UI. As for the validation data set, we felt this was one of the larger datasets we could find that had a high-impact manuscript to complement it with vetted conclusions. As shown in the manuscript, we validated the use of ExCYT in coming to these same conclusions.

22). High dimensional flow/box plots provide little information, nor do the authors provide any guidance on why they are valuable or how they should be interpreted.

- Respectfully, various non-informatics users of the same tools noted above by the reviewer at our institution have them informative and have used them to create figures for presentations and manuscripts currently in review.

Reviewer #3:

Manuscript Summary:
This articles describes ExCYT, a toolbox written in Matlab with a GUI to analyze flow cytometry data. Many cluster algorithms are provided.
The manuscript is well written and will likely provide some useful tools for biologists.

Major Concerns:
None

Minor Concerns:
- Not all discussed cluster methods have references. Perhaps provide a reference to the methods and an application in FC for each?
- Only a limited number of data transformation methods are implemented (linear, log10, arcsinh). Some other commonly-used transformations are discussed in "Optimizing transformations for automated, high throughput analysis of flow cytometry data", i.e. linlog, gen arcsinh, biexponential, gen Box-Cox. I would be useful to add these as well, with hyperparameters to be selected by the user.
- Likewise, only the Euclidean distance is implemented for clustering. Some others might work better, such as the squared Euclidean, Mahalanobis, Manhattan etc. These should ideally also be provided.

We appreciate the suggestions from this reviewer. While our goal is to maintain simplicity of the program for non-technical users, we have added some more functionality by allowing the user to specify the distance metric and perplexity of the t-SNE analysis. In terms of the provided transformations, we have selected those that are commonly found and used by biologists in programs such as FlowJo. Our goal was to create an environment with some of the familiarities from FlowJo with the capabilities of more sophisticated tools.

Reviewer #4:
The authors provide a graphic user interface for exploring multi-parameter cytometry datasets and apply a large variety of clustering algorithms. Using the interface does not require any knowledge in programming, making the powerful dimensionality reduction and clustering techniques available to a broad spectrum of researchers. With the continuing increase of high-dimensional cytometry the open access "ExCyt"-Software will prove useful in the visualization and discovery of underlying cell population structures.
Overall the manuscript is well-written, clearly detailing the use of the program in a step-wise protocol, thereby, enabling new users to quickly understand the data analysis process. A special focus is put on introducing short introduction of the implemented clustering algorithms that are likely unfamiliar to the intended users of "ExCyt".
In the second part of the manuscript the authors employ their software to re-analyze a CyTOF dataset and convincingly recreate the published results, highlighting the usability of "ExCyt" in the analysis of real-world high-dimensional cytometry data.
The ExCyt GUI presents itself clearly organized and user-friendly. It allows the straightforward creation of complex figures, especially the novel high-dimensional flow and box plots provide a powerful visualization of differences between clusters. The manual refinement of automatically generated clusters by manual gating is a valuable feature for users accustomed to the traditional analysis of cytometric data.

Major:
Bugs:
-Canceling manual tSNE-gating erases graph

- Unfortunately, we cannot replicate this bug.

-Inappropriately named single stain samples can be selected without warning

- We included further error-checking functionality in the importation of single-stain files. The program now checks that the user selected fcs files, then checks for an 'Unstained.fcs' file, and then finally checks to make sure the single-stain file names are in concordance with the panel in the main file for analysis.

-Non-integer/non-numeric arguments are allowed as input for clustering arguments (e.g. # of clusters = 1.2)

- We have fixed this and now the program throws back an error to the user if a non-integer value is given for a clustering algorithm that requires an integer input.

Suggestions:
-Add color legend for clusters
-Improve documentation of auto-compensation
oNaming Single Stain samples by Fluorophore does not work
oName has to match $PnN Keyword in .FCS header
-Editing of parameter names after data import
-Allow users to set tSNE parameters (Perplexity, Iterations, …)
-Display automatically calculated compensation matrix for user evalution
-Indicator for status of auto-compensation
-Add warning for unsaved workspace when exiting
-Add Import of .lmd files
Minor:
-The sentence in line 261-263 needs clarification. Furthermore in line 151 "speculations" should be changed to "specifications".

- We thank the reviewer for suggestions. We have given the user the ability to set tSNE parameters as well corrected the typo found by the reviewer. We will work on including more of these features in future releases.