**Improving the accuracy of automated human cell classification in single cell mass cytometry**

Daniel Liu1,2, Nalin Leelatian1, Allison Rae Greenplate1,2, Kirsten Elizabeth Diggins1, Jonathan Michael Irish1,2

1Department of Cancer Biology, Vanderbilt University, Nashville, TN, USA.

2Department of Pathology, Microbiology, and Immunology, Vanderbilt University Medical Center, Nashville, TN.

Corresponding author: Dr. Jonathan M. Irish Ph.D., Assistant Professor of Cancer Biology, Vanderbilt University School of Medicine, 740B Preston Building, 2220 Pierce Avenue, Nashville, TN 37232-6840; Tel (615) 875-0965; E-mail: [jonathan.irish@vanderbilt.edu](mailto:jonathan.irish@vanderbilt.edu)

Funding sources: This study was supported by NIH/NCI R00 CA143231-03 (J.M.I. and A.R.G.), F31 CA199993 (A.R.G.), and the Vanderbilt-Ingram Cancer Center (VICC, P30 CA68485).

Running Title: Cytotyping accuracy improved by combined analysis of t-SNE axes and original features

**Abstract**

Identifying populations within a sample of cells reveals changes in cell signaling which has implications in the detection and treatment of diseases such as cancer. Mass cytometry has augmented the resolution by which we can study a sample of cells; however, learning cell identity through manual gating still remains a challenge. Gating is both time consuming and technically involved. Programs like SPADE were created to expedite this process as well as decrease the technical skills required to identify cells, but such programs are generally inaccurate. We found that the use of dimensionality-reducing cell similarity scores, tSNEs, improves the accuracy of the cell identification program SPADE in finding the same populations an expert would. tSNE was used in the automated identification of both peripheral blood mononuclear cells (PMBC) and glioblastoma multiforme (GBM) samples, and we found significantly increased accuracy for cell identification in both samples. While accuracy is improved, additional changes to automatic gating should be explored to further maximize the reliability of such programs.

**Introduction**

Advances in flow cytometry have enabled measurement of 35 or more proteins on millions of single cells from human tissues (Bendall et al., 2011) and raised the possibility of using machine learning algorithms to automatically identity populations of cells (Diggins et al., 2017) . However, current methods of cell identification rely on human experts manually identifying cell populations in a process called gating. Gating is an essential part of single cell data analysis. Once cells are identified as part of a population, analysis of protein expression or enrichment can be used to confirm cell identity (Diggins et al., 2015; Diggins et al., 2017) and to quantify functional properties of the cells, including signaling, cell cycle status, and more (Irish and Doxie, 2014). The abundance of cell types with different protein expression or signaling has been shown to be closely connected to the physiological state of patients, such as disease type (Myklebust et al., 2017), risk of poor outcome (Irish et al., 2004; Irish et al., 2010), and probability of response to treatment or recovery (Gaudilliere et al., 2014). However, the process of gating for cell subsets traditionally requires human experts to review cell expression profiles manually for individual features or pairs of features and recognize cell subsets based on patterns in the measured features and prior knowledge of biology (Diggins et al., 2015; Saeys et al., 2016). This type of manual gating on a one or two dimensional histogram has long been considered the gold standard, and immunologists have spent decades refining and discussing the order of feature analysis and the importance of individual features to cell identity (Roussel et al., 2016). More recently, automated clustering and machine learning tools have begun to challenge the manual gating approach to cell identification (Amir el et al., 2013; Irish, 2014). A well-established clustering tool in mass cytometry is spanning-tree progression analysis of density-normalized events (SPADE), which is typically used to assign cells into clusters based on the original measured features or a linear dimensionality reduction, such as a principal component analysis of the features (Qiu et al., 2011). t-SNE, viSNE, and related approaches were developed to create two-dimensional views of multidimensional data based on optimizing placement of cells in a low dimensional view according to a multidimensional similarity score (Amir el et al., 2013; Diggins et al., 2015). In a t-SNE map, cells form clusters based on their similarity to other cells, and the populations can be identified by an expert based on location and marker expression levels. Originally intended as a way to visualize populations and reveal hidden cell types, plots created by t-SNE have been observed to provide new ways to identify cell types by clustering on t-SNE axes (Becher et al., 2014; Diggins et al., 2017; Ferrell et al., 2016). The use of these new computational techniques in cytometry has raised the question of whether gating for cell identity on t-SNE axes is more or less accurate than gating on the “original” measured features. Expert manual gating is time-consuming and introduces the possibility of bias, particularly in tissue types that are less well-defined (Leelatian et al., 2017). But, it is not clear whether automated clustering tools should use the original measured features, t-SNE axes created based on similar analysis, or a combination of both. The goal of this study was to test whether the inclusion of the axes created in two dimensional t-SNE analysis improved or degraded the accuracy of clustering.

To test the efficacy of computational cell identification programs, two datasets were used in a series of computational experiments designed to compare manual gating as a “gold standard” with computational gating on original measured features, t-SNE axes, or both, simultaneously. Manual gating was performed both on the original measured features, as is traditional, and on t-SNE axes, as is becoming common in mass cytometry. SPADE was used as the clustering tool, and f-measure of grouping expert-assigned population identities used as the means of assessing accuracy of SPADE-identified cell populations. Two datasets were compared: 1) published data from glioblastoma multiforme (GBM) tumors (Diggins et al., 2017; Leelatian et al., 2017) and 2) published data from healthy human peripheral blood mononuclear cells (PBMCs) (Diggins et al., 2017). PBMCs consist of the cellular components of blood within the circulating pool of blood that have a round nucleus. PBMC populations functioned as a robust control group, since blood cell populations are well-studied and considered relatively easy for experts to identify. GBM is a deadly and relatively untreatable form of brain cancer that is thought to arise from neural origin cells that exhibit stem cell properties. It is the highest grade of glioma and is extremely heterogeneous. Identification of the population subsets in GBM is significantly more challenging than in peripheral blood and time consuming even for experts. The use of SPADE on GBM data thus provided a situation where cell identification would be significantly more challenging than in blood.

**Patients, Materials, & Methods**

Acquiring the samples

Single cell mass cytometry data set derived from one surgically resected human glioblastoma and one extraction of peripheral blood mononuclear cells were used in this study. This PBMC dataset was selected because the cell subsets within the sample are well-established, while the GBM dataset was selected because it contained both easily defined cell subsets as well as some poorly identifiable ones. Thus, the capability of the automatic programs in detecting less distinct populations could also be tested, with known cell types serving as internal controls.

Identifying the populations manually and automatically

To acquire the expert manual gating subsets, viSNE analysis was performed first on the single cell data from PBMCs to obtain a two-dimensional t-SNE plot (figure 1). Cell density and protein expression information were used to guide manual cell subset identification, resulting in 30 distinct subsets. These subsets were then compared to clustering-based computational algorithms (Qiu et al., 2011). It is hypothesized based on current knowledge in the field that the use of markers only would better define cell subsets than by t-SNE values alone or in conjunction with protein expression. For these reasons, 3 SPADE clustering settings (t-SNEs only, protein expression only, or t-SNEs with protein expression) were compared with expert gates. For SPADE analysis, the most optimal setting based on previous experiments was used: a target of 30 nodes (one node per true population) and a target of 1% down-sampling. To ensure that the effects of t-SNEs on the results SPADE was not due to truth being defined by viSNE, populations were identified based on biaxial gating as well (figure 1). To acquire expert biaxial manual gated subsets, the level of protein expression on single cells were plotted on a biaxial plot of select markers. The gating was done hierarchically, with the most general populations, such as T cells and B cells, defined first, and the more specific populations defined last resulting in a number of terminal populations. In the PBMC population, 19 populations were identified through biaxial gating. The tests for the 3 SPADE clustering algorithms were repeated for this expert gate as well, with the settings being 19 nodes and a target of 1% down-sampling

The same aforementioned techniques were applied to the GBM dataset as well. Manual gating based on viSNE found 21 populations for GBM, while gating based on biaxial plots found 15 distinct populations (Leelatian et al., 2017).. Thus, the SPADE settings used were 21 nodes and a target of 1% down-sampling for the viSNE based gates, whereas they were 15 nodes and a target of 1% down-sampling for the biaxial plot based gates.

Statistical Analysis

After the SPADE trees were obtained, the results were compared to the true populations with the f-measure test. 10 SPADE trees for each condition were compared, and the average and median f-measures were calculated. An ANOVA test and paired t-test was done to test whether the f-measures between conditions was statistically significant, and the standard deviation between conditions was calculated.

**Results**

Gating for PBMC by viSNE resulted in 30 distinct populations found, hence a target of 30 nodes on the SPADE tree (Fig. 1A). Conversely, biaxial gating resulted in 19 distinct populations, hence a target of 19 nodes on the SPADE tree (Fig. 1B).

When SPADE trees (n=10) for each setting were compared to the true populations, the f-measures showed significant variation. When comparing SPADE trees obtained using marker and t-SNE information to true populations according to viSNE analysis, f-measure was on average 0.887 (Fig. 2B). When comparing SPADE trees obtained using t-SNE information only to true populations according to viSNE analysis, f-measure was on average 0.846 (Fig. 2B). When comparing SPADE trees obtained using marker information only to populations according to viSNE analysis, f-measure was on average 0.760 (Fig. 2B). Variance between trees was low as well. For SPADE by markers and t-SNE, the standard deviation was 0.011, for SPADE by t-SNE only that value was 0.0091, and for SPADE by markers only that value was 0.013. When comparing SPADE trees obtained using marker and t-SNE information to true populations according to biaxial gating analysis, f-measure was on average 0.724 (Fig. 2A). When comparing SPADE trees obtained using t-SNE information only to true populations according to biaxial gating analysis, f-measure was on average 0.724 (Fig. 2A). And, when comparing SPADE trees obtained using marker and t-SNE information to true populations according to biaxial gating analysis, f-measure was on average 0.715 (Fig. 2A). For SPADE by markers and t-SNE, the standard deviation was 0.024, for SPADE by t-SNE only that value was 0.015, and for SPADE by markers only that value was 0.022.

Gating for GBM by viSNE resulted in 21 distinct populations found, hence a target of 21 nodes on the SPADE tree (Fig. 3A). Conversely, biaxial gating resulted in 15 distinct populations, hence a target of 15 nodes on the SPADE tree (Fig. 3B).

When SPADE trees (n=10) for each setting were compared to the true populations, the f-measures showed statistically significant variation. When comparing SPADE trees obtained using marker and t-SNE information to true populations according to viSNE analysis, f-measure was on average 0.768 (Fig. 4B). When comparing SPADE trees obtained using t-SNE information only to true populations according to viSNE analysis, f-measure was on average 0.754 (Fig. 4B). When comparing SPADE trees obtained using marker information only to populations according to viSNE analysis, f-measure was on average 0.685 (Fig. 4B). The f-measures fell within a narrow range as well, as standard deviation was 0.015 for SPADE by markers and t-SNEs, 0.017 for t-SNEs only, and 0.019 for markers only. When comparing SPADE trees obtained using marker and t-SNE information to true populations according to biaxial gating analysis, f-measure was on average 0.840 (Fig. 4A). When comparing SPADE trees obtained using t-SNE information only to true populations according to biaxial gating analysis, f-measure was on average 0.853 (Fig. 4A). And, when comparing SPADE trees obtained using marker and t-SNE information to true populations according to biaxial gating analysis, f-measure was on average 0.751 (Fig. 4A). The f-measures fell within a narrow range as well, as standard deviation was 0.015 for SPADE by markers and t-SNEs, 0.020 for t-SNEs only, and 0.015 for markers only.

**Discussion**

Based on current understandings in the field of immunology and cancer biology, the results are inconsistent with what was expected. Marker information has been the primary information on which population identification has been based. Thus, t-SNE values were thought to “dilute” the quality of information available to SPADE. Since t-SNE reduces the number of dimensions that quantify the sample, the information is thought to be of lesser quality than information provided by protein markers. The dimensionality reduction would reduce the effectiveness of SPADE since the program would be given “bad” information. However, the opposite appears to be true based on the results. Regardless of what population was used as truth, the use of t-SNE information resulted in an automatically gated population that more closely resembled the true population. Using markers and t-SNEs generally provided the closest representation. However, sometimes the use of t-SNE values alone actually produced the best results. Nevertheless, the use of markers only always resulted in population identification that least resembled the true population as established by an expert. Most importantly, this signifies that the results are not circular. Essentially, the use of t-SNEs in SPADE does not enhance SPADE’s ability to identify the correct populations as a result of the aforementioned populations being based on viSNEs.

Given that, the findings suggest that the use of t-SNE enhances SPADE’s ability to identify populations. In PBMCs, the f-measures for SPADE demonstrate that the use of t-SNEs alone and t-SNEs with markers produced clustering that was better than clustering based on markers only (Fig. 2B). However, when biaxial gating was used as the truth in the PBMC data, that difference was less pronounced (Fig. 2A). This was possibly due to the fact that the populations identified based on biaxial plots, such as the CD16-, CD56 bright NK cells, are much rarer than what even would appear on a viSNE plot (Fig. 1). As a result, the SPADE program was probably unable to produce an accurate representation of the true population regardless of the information it was given since identification of such novel populations is beyond the capability of the program.

For the GBM data, the results were similar to the PBMC data; the use of t-SNEs in SPADE produced results that were better than using markers only. Interestingly, when the results for SPADE were compared to the true populations based on viSNE, t-SNEs only outperformed the use of markers and t-SNEs. It is possible that in populations less identifiable, such as GBM, the use of the t-SNE similarity scores may be the best information to base clustering on, and that the use of marker information in conjunction with t-SNE may actually impair the program’s ability to produce more accurate clustering.

Overall, the use of t-SNEs likely enhances SPADE’s population identification because more information is being provided to the program. By being given the cell-cell similarity score, SPADE can more accurately compare cells and put them in the correct population cluster. Additionally, the program would use the same information used by a human when they identify populations. When gating on viSNE, a human expert would utilize both marker and t-SNE information. By mimicking the human expert, SPADE’s ability to identify the true populations may be enhanced.

Improving the accuracy of the program had no adverse effect on the precision. The range for the f-measures was relatively small and did not vary between the three conditions (Fig. 2; Fig. 4). This means that the use of SPADE with t-SNE information will consistently generate relatively accurate gating. More importantly, this suggests that the use of automated population identification is a viable research tool, especially if more information can be provided to maximize the accuracy, since one will always know that the gating produced by the program will reliably find the sub-populations.

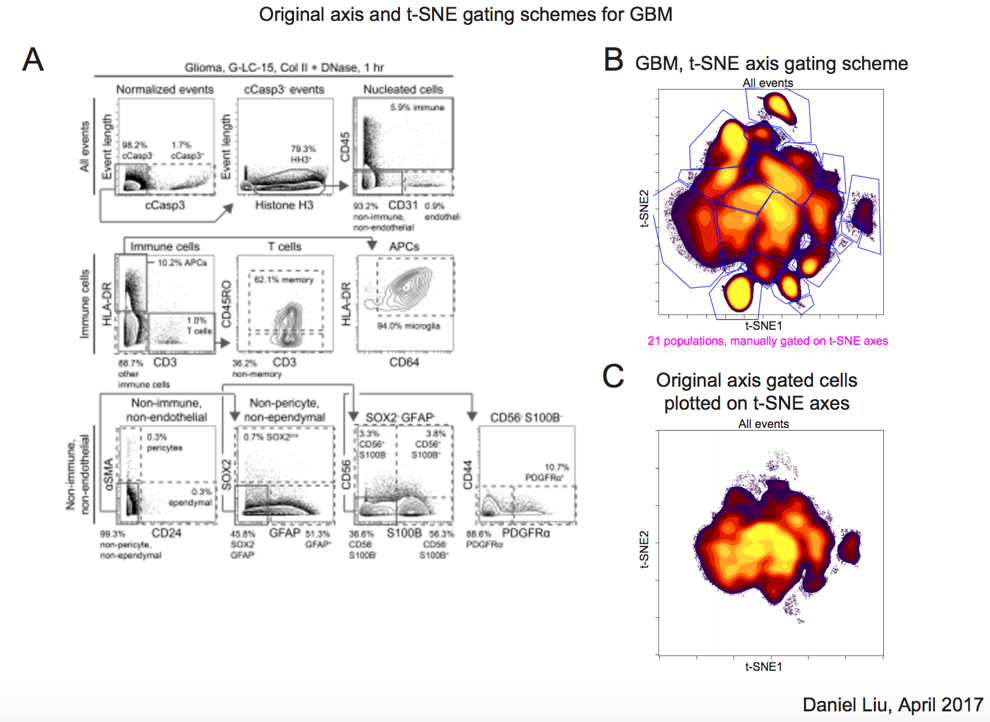
Nonetheless, SPADE is currently still unable to correctly identify all populations as a human would. As seen in the PBMC results compared to the biaxial gates, this gap may be due to the small “rare” populations found within the sample. Therefore, SPADE’s usefulness may be limited to easily identifiable populations within a sample, or for a beginner who is still learning how to identify population subsets. For people with more advanced gating skills or are researching populations with rarer phenotypes, manual gating might be the better or only possible option to identify certain cell populations.

The success of t-SNE in improving the accuracy by which SPADE is able to identify populations in GBM samples shows that there is potential for automatic gating to be further improved especially in samples with less easily identifiable populations. Additionally, it is possible that other similarity scores or identification-related information can be provided to produce more accurate population clustering. Perhaps, if SPADE and other automated gating tools can be improved by additional information, automatic gating can be used as a comparison point to double check for biases or even identify novel populations. Automatic gating presents a possible tool that would improve the study of diseases like GBM where the populations are not as well identified. This would teach us about the sub-populations and the cell-cell signaling that has clinical implications for treatment of disease.

** Figure 1 – Original measurement axis or t-SNE axis gating schemes defined PBMC cell subsets as “truth”.** Biaxial plots show identification of cell subsets in healthy human PBMC using (A) density of populations on original mass cytometry measurement axes and (B) density of populations on t-SNE axes. For (A), cell identity was assessed using 15 of the original measurement axes canonically accepted as defining major cell types, as in Figure 1 of (Nicholas et al., 2016), and 19 populations were identified. For (B), the t-SNE map was created using viSNE analysis of 28 of the original measurement axes available, and 30 populations were identified as dense clusters, as in Figure 1 of (Diggins et al., 2017). The populations identified by these two approaches were used subsequently as “truth” for f-measure calculations when assessing clustering strategies. (C) Cells identified in (A) are plotted on t-SNE axes from (B) and colored according to the identity based on the original measurement axes.

****

**Figure 2 – Clustering with t-SNE axes significantly improves accurate identification of human PBMC cell subsets.** F-measure was used to assess accuracy of SPADE clustering on original measurement axes (28 dimensions), t-SNE axes (2 dimensions), or both original measurement axes and t-SNE axes (30 dimensions) for PBMC populations defined by manual gating on (A) original measurement axes or (B) t-SNE axes (see Figure 1). A total of 10 SPADE clustering runs were used to automatically identify 30 SPADE populations for truth populations identified on the t-SNE map and 19 SPADE populations for truth populations identified on the original marker axes. Box and whisker plots show the median, 25th and 75th percentile, and the range for the f-measure across the 10 SPADE clustering runs (N = 10 for each of the six comparisons).

****

**Figure 3 – Original measurement axis or t-SNE axis gating schemes defined GBM cell subsets as “truth”.** Biaxial plots show identification of cell subsets in human glioblastoma multiforme tumor using (A) density of populations on original mass cytometry measurement axes and (B) density of populations on t-SNE axes. For (A), cell identity was assessed using 14 of the original measurement axes canonically accepted as defining major cell types, as in Figure 5 of (Leelatian et al., 2017), and 15 populations were identified. For (B), the t-SNE map was created using viSNE analysis of 13 of the original measurement axes available and 21 populations were identified as dense clusters, as in Figure 4 of (Leelatian et al., 2017). The populations identified by these two approaches were used subsequently as “truth” for f-measure calculations when assessing clustering strategies. (C) Cells identified in (A) are plotted on t-SNE axes from (B) and colored according to the identity based on the original measurement axes.

****

**Figure 4 – Clustering with t-SNE axes significantly improves accurate identification of human GBM tumor cell subsets.** F-measure was used to assess accuracy of SPADE clustering on original measurement axes (13 dimensions), t-SNE axes (2 dimensions), or both original measurement axes and t-SNE axes (15 dimensions) for GBM populations defined by manual gating on (A) original measurement axes or (B) t-SNE axes (see Figure 3). A total of 10 SPADE clustering runs were used to automatically identify 21 SPADE populations for truth populations identified by t-SNE axis gating and 15 SPADE populations for true populations identified by marker axis gating. Box and whisker plots show the median, 25th and 75th percentile, and the range for the f-measure across the 10 SPADE clustering runs (N = 10 for each of the six comparisons).

**Bibliography**

Amir el, A. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nature biotechnology *31*, 545-552.

Becher, B., Schlitzer, A., Chen, J., Mair, F., Sumatoh, H. R., Teng, K. W., Low, D., Ruedl, C., Riccardi-Castagnoli, P., Poidinger, M.*, et al.* (2014). High-dimensional analysis of the murine myeloid cell system. Nature immunology *15*, 1181-1189.

Bendall, S. C., Simonds, E. F., Qiu, P., Amir el, A. D., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I.*, et al.* (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science *332*, 687-696.

Diggins, K. E., Ferrell, P. B., Jr., and Irish, J. M. (2015). Methods for discovery and characterization of cell subsets in high dimensional mass cytometry data. Methods *82*, 55-63.

Diggins, K. E., Greenplate, A. R., Leelatian, N., Wogsland, C. E., and Irish, J. M. (2017). Characterizing cell subsets using marker enrichment modeling. Nature methods *14*, 275-278.

Ferrell, P. B., Jr., Diggins, K. E., Polikowsky, H. G., Mohan, S. R., Seegmiller, A. C., and Irish, J. M. (2016). High-Dimensional Analysis of Acute Myeloid Leukemia Reveals Phenotypic Changes in Persistent Cells during Induction Therapy. PLoS One *11*, e0153207.

Gaudilliere, B., Fragiadakis, G. K., Bruggner, R. V., Nicolau, M., Finck, R., Tingle, M., Silva, J., Ganio, E. A., Yeh, C. G., Maloney, W. J.*, et al.* (2014). Clinical recovery from surgery correlates with single-cell immune signatures. Science translational medicine *6*, 255ra131.

Irish, J. M. (2014). Beyond the age of cellular discovery. Nature immunology *15*, 1095-1097.

Irish, J. M., and Doxie, D. B. (2014). High-dimensional single-cell cancer biology. Current topics in microbiology and immunology *377*, 1-21.

Irish, J. M., Hovland, R., Krutzik, P. O., Perez, O. D., Bruserud, O., Gjertsen, B. T., and Nolan, G. P. (2004). Single cell profiling of potentiated phospho-protein networks in cancer cells. Cell *118*, 217-228.

Irish, J. M., Myklebust, J. H., Alizadeh, A. A., Houot, R., Sharman, J. P., Czerwinski, D. K., Nolan, G. P., and Levy, R. (2010). B-cell signaling networks reveal a negative prognostic human lymphoma cell subset that emerges during tumor progression. Proc Natl Acad Sci U S A *107*, 12747-12754.

Leelatian, N., Doxie, D. B., Greenplate, A. R., Mobley, B. C., Lehman, J. M., Sinnaeve, J., Kauffmann, R. M., Werkhaven, J. A., Mistry, A. M., Weaver, K. D.*, et al.* (2017). Single cell analysis of human tissues and solid tumors with mass cytometry. Cytometry B Clin Cytom *92*, 68-78.

Myklebust, J. H., Brody, J., Kohrt, H. E., Kolstad, A., Czerwinski, D. K., Walchli, S., Green, M. R., Troen, G., Liestol, K., Beiske, K.*, et al.* (2017). Distinct patterns of B-cell receptor signaling in non-Hodgkin lymphomas identified by single-cell profiling. Blood *129*, 759-770.

Nicholas, K. J., Greenplate, A. R., Flaherty, D. K., Matlock, B. K., Juan, J. S., Smith, R. M., Irish, J. M., and Kalams, S. A. (2016). Multiparameter analysis of stimulated human peripheral blood mononuclear cells: A comparison of mass and fluorescence cytometry. Cytometry Part A : the journal of the International Society for Analytical Cytology *89*, 271-280.

Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., and Plevritis, S. K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. In Nat Biotechnol, pp. 886-891.

Roussel, M., Greenplate, A. R., and Irish, J. M. (2016). Dissecting Complex Cellular Systems with High Dimensional Single Cell Mass Cytometry. In Experimental Approaches for the Investigation of Innate Immunity, (WORLD SCIENTIFIC), pp. 15-26.

Saeys, Y., Gassen, S. V., and Lambrecht, B. N. (2016). Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nature reviews Immunology *16*, 449-462.