Program #: 218
Board #: B81

High-Dimensional Modeling of Peripheral Blood Mononuclear Cells from a Helios Instrument

Bagwell CB[1], Leipold M[2], Maecker H[2], and Stelzer G[3]; [1] Verity Software House, 45A Augusta Rd, Topsham, ME 04086;
[2]Stanford University, 299 Campus Dr., Stanford CA 94305; [3]Fluidigm, 7000 Shoreline Court, Suite 100, South San Francisco, CA 94080 USA

Verity Software House

Abstract

Analyzing data that comprises 48 correlated measurements on cells presents major issues for classical gating analysis systems. The number of two-dimensional measurement plots necessary to investigate all correlations is as high as 2,256 – a number that may require investigators to consider different analysis paradigms to extract useful information from the sample.

This study uses the technique of probability state modeling (PSM) to automatically model and extract information from PBMC samples. The major cell types defined in the model are 1) CD8 T cells, 2) CD4 T cells, 3) Gamma-Delta T cells, 4) B cells, 5) Monocytes, 6) NK cells, 7) Dendritic cells, 8) Basophils, 9) Granulocytes, and 10) Plasmablasts (PB). Within many of these cell types, populations are sub-classified into stages or categories and within each of these categories additional populations are defined with the TriCOM combinatory analysis system such as T_H1, T_H2, T_H17, NKT, T_{Reg}, and others. Other published approaches such as SPADE and t-SNE were evaluated for reproducibility and suitability for integration into GemStone.

The advantages of modeling systems like PSM are 1) scalability, 2) accuracy, 3) reproducibility, and 4) automation. A new force-directed graphics system summarizes much of the information in the model and thereby provides an informative and reproducible method for evaluating the results.

Introduction

To date, performing modeling on high-dimensional data sets with Probability State Modeling (PSM, GemStone) (1-5) has been mainly employed to identify specific populations such as PNH clones (6), stem cells (7), fetomaternal red cells (8), CD64 positive cells (9), T cells (10), and bone marrow B cells (11). With the advent of mass cytometry (12,13) and the greatly improved signal-to-noise characteristics of the Helios instrument (14), it is possible to model all the major populations present in peripheral blood mononuclear cell preparations (PBMC).

Objective 1. One of the main objectives of this study is to develop and present a GemStone model that automatically identifies all the major cellular populations in a set of PBMC samples from a Helios mass cytometer with 48 correlated measurements and when appropriate, stage them. A new force-directed graph will be presented that summarizes all the modeled populations and their respective percentiles.

Objective 2. Another objective of this study is to evaluate some other algorithms such as SPADE (15-17) and t-SNE (17-20). These methods can map n-dimensional data to two-dimensional display surfaces and can be evaluated by creating realistic simulated data where truth is known and then critically examining the analysis results. SPADE, spanning-tree progression analysis of density-normalized events, uses a density-dependent k-means clustering algorithm to down-sample data into nodes and then a spanning-tree graphical algorithm to connect related cluster nodes (15). The nodes can then be edited by the user via rotation, translocation, and normalization tools to improve their visual appearance and make them appear as clusters of grapes representing cellular ontogeny (21). Another common approach to visualizing n-dimensional data is to use the parametric t-SNE algorithm originally proposed by van der Maaten in 2009 (18), with improved performance using the Barnes-Hut algorithm (19), and ported by Amir to a MatLab-based system modified to accept cytometry FCS files (20).

Objective 3. The final objective for this study is to explore the possible integration of the t-SNE method within the context of GemStone modeling. High-dimensional modeling has the advantage of automatically identifying all normal n-dimensional cellular populations and t-SNE plots have the advantage of visualizing n-dimensional neighboring structures for both normal and abnormal populations in low-dimensional space (22). The integration of these two methods may allow investigators to better understand the true nature of the high-dimensional relationships embedded in the data.

Material and Methods

Synthetic Sample Files: Three replicate files and GemStone models from a single normal donor were selected from the CD8 T cell study (10). The models were used by GemStone to synthesize 10,000 CD3+ CD8+ events where the only measurement modulations were the stages: Naive (CCR7+ CD28+ CD45RA+), CM (CCR7- CD28+ CD45RA-), EM (CCR7- CD28- CD45RA-) and EF (CCR7- CD28- CD45RA+). There were no branches or any other populations in the data set.

Helios Sample Files: Four PBMC samples from Human Immune Monitoring Center at Stanford University were submitted to this study and labelled as Sample 1, 2, 3, and 4. All samples were from separate aliquots from a single draw of a normal donor. Samples 1 and 2 were thawed, stained, and run on the same day, with the same set of reagents; whereas, Samples 3 and 4 were run on later days. Any differences observed are due to variations from storage, thawing, staining, acquisition, and analysis. The 48 correlated measurements are shown in **Figure 1**.

CyTOF Immunophenotyping: This assay was performed in the Human Immune Monitoring Center at Stanford University. PBMCs were thawed in warm media, washed twice, resuspended in CyFACS buffer (PBS supplemented with 2% BSA, 2 mM EDTA, and 0.1% sodium azide), and viable cells were counted by Vicell. Cells were added to a V-bottom microtiter plate at 1.5 million viable cells/well and washed once by pelleting and resuspension in fresh CyFACS buffer. The cells were stained for 60 min on ice with 50 uL of the following antibody-polymer conjugate cocktail (CD45RA, CD20, CD33, CD28, CD24, CD161, CD38, CD11b, CCR6, CD94, CD86, CXCR5, CCR7, CD127, CD57, Live/Dead, HLA_DR, CD19, CD4, CD8, IgD, CD11c, CD3, CD85j, CD16, CD27, CD14, CXCR3, ICOS, TCRgd, PD-1, CD123, CD56, and CD25). All antibodies were from purified unconjugated, carrier-protein-free stocks from BD Biosciences, Biolegend, or R&D Systems. The polymer and metal isotopes were from Fluidigm. The cells were washed twice by pelleting and resuspension with 250 uL FACS buffer. The cells were resuspended in 100 uL PBS buffer containing 2 ug/mL Live-Dead (DOTA-maleimide) (Macrocyclics) containing natural-abundance Indium). The cells were washed twice by pelleting and resuspension with 250 uL PBS. The cells were resuspended in 100 uL 2% PFA in PBS and placed at 4C overnight. The next day, the cells were pelleted and washed by resuspension in fresh PBS. The cells were resuspended in 100 uL eBioscences permeabilization buffer (1x in PBS) and placed on ice for 45 min before washing twice with 250 uL PBS. If intracellular staining was performed, the cells were resuspended in 50 uL antibody cocktail in CyFACS for 1 hour on ice before washing twice in CyFACS. The cells were resuspended in 100 uL iridium-containing DNA intercalator (1:2000 dilution in PBS; Fluidigm) and incubated at room temperature for 20 min. The cells were washed twice in 250 uL MilliQ water. The cells were diluted in a total volume of 700 uL in MilliQ water before acquisition on the CyTOF (Helios, Fluidigm). The data presented in figures 2, 3, 5, and 6 were pre-gated on intact cells based on the iridium isotopes from the intercalator, then on singlets by Ir191 vs cell length, then on live cells (Indium-LiveDead minus population).

GemStone Analysis: GemStone (Verity Software House, Maine USA) V1.0.142 was used for all modeling analyses. The same model was loaded for all samples and all analyses were run automatically with no user input.

t-SNE Cyt Analysis: Version 2.0 of cyt was obtained from the website, http://www.c2b2.columbia.edu/danapeerlab/html/cyt-download.html. This version of "Cyt" required MatLab version B.6.0.267246. The analysis procedure was to 1) load the file of interest into the system, 2) set the cofactor to 1000 for fluorescence-based cytometry data and 5 for mass cytometry data, 3) select the measurements to analyze, 4) select and execute the bh-SNE algorithm, and 5) plot the two added bh-SNE parameters as scatter plots using all default settings. The specific measurement selections are indicated in the appropriate figure legends. Identification of subpopulations was done by either examining a series of measurement heat maps or using GemStone's encoded GS cumulative zone ids.

SPADE 3.0 Analysis: The Windows pre-compiled standalone version of SPADE 3.0 was obtained from the website, http://pengqiu.gatech.edu/software/SPADE/index.html. The analysis procedure was to 1) choose the folder containing the FCS file, 2) choose the "overlapping markers used for SPADE tree", 3) choose Arcsinh with cofactor of 1000, 4) leave all other options in their default state, including the 100 desired clusters, 5) choose "One click for all", and 6) choosing "overlay information by coloring nodes" with either a series of measurement heat maps or using GemStone's encoded GS cumulative zone ids.

Results

Figure 1. GemStone Model Design

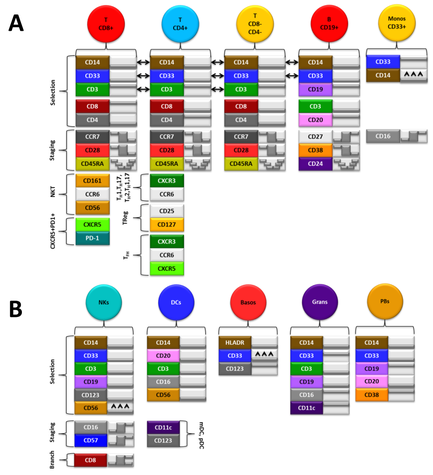

Figure 1. GemStone Model Design: The selection and staging of the T cells, B cells, and Monocytes are as shown in **Panel A**. To the right of each measurement label is a figure showing the general shape of the expression profile. Black double arrows show linked expression profiles. Staging for the T cells was done as published (10) with a step-down for CCR7 and for CD28 and three levels for CD45RA (high-low-high). Staging for the B cells was done with a step-up for CD27, step-down for CD38, and three-level for CD24 (high-low-high). Staging for the Monocytes was done with a three-level expression profile for CD16.

A number of subpopulations were identified for CD8+ and CD4+ T cells using multiple TriCOMs. The combinatory populations for the CD8+ cell type included NKT (CD161+CCR6+CD56+) and a currently unidentified subpopulation with the phenotype, CXCR5+PD1+. The populations for CD4+ cell type included T_H1 (CXCR3+CCR6-), T_H2 (CXCR3-CCR6-), T_H17 (CXCR3-CCR6+), $T_H1.17$ (CXCR3+CCR6+), T_{Reg} (CD25+ CD127-), and T_{FH} (CXCR3+CCR6-CXCR5+).

The selections and staging's for Natural Killer cells (NKs), Dendritic cells (DCs), Basophils (Basos), contaminating Granulocytes (Grans), and Plasma Blasts (PBs) are as shown in **Panel B**. NK selection was CD14- CD33- CD3- CD19- CD123- and CD56 dim/+. Staging for the NK cells was done with a step-up for CD16 and CD57. CD8 was considered a branched step-up.

Two subpopulations were identified for the Dendritic cells using TriCOM: pDC (CD123+CD11c-) and mDC (CD123-CD11c+).

Figure 2. Reproducible Progression Plots

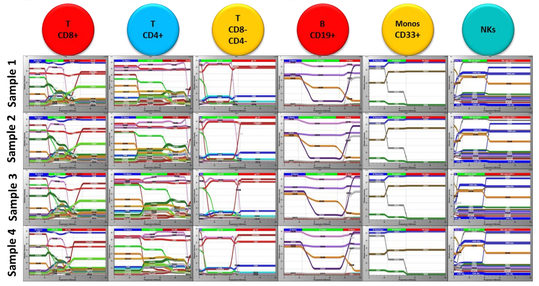

Figure 2. Progression Plots: Progression overlays from each sample for cell types T CD8+, T CD4+, T CD8-CD4-, B CD19+, Monos CD33+, and NKs are shown in the above figure. Details of the model design are shown in **Figure 1**. The x-axis represents the relative progression for each of the cell types. The modulations of a number of selected measurements are shown in each plot where the y-axis is the relative intensity of the measurements. Note the reproducibility of the high-dimensional modeling for the four samples from a single donor.

Figure 3. Model Summary Maps

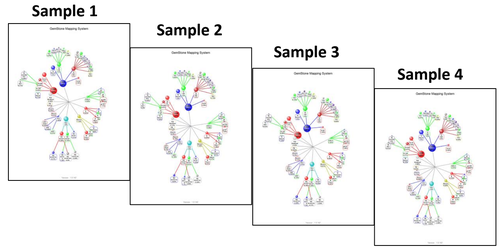

Figure 3. Model Force-Directed Summary Maps: These maps summarize the percentages for all the populations modeled from each of the four samples. The areas of the nodes are proportional to the number of events in the cell type, stage, or selected populations within the stages. Colors for the nodes are determined from the corresponding model cell type, stage, and triCOM combination colors. Parent-child nodes have a spring-like force where if the nodes are too close, they repel, and if they are too far, they attract. All other nodes have an inverse distance repelling force. These graphs have a self-assembly characteristic which results in a visually pleasing distribution of all the important populations defined in the model and allow a quick inspection of many of the modeling results.

Figure 4. GemStone, SPADE, t-SNE Reproducibility

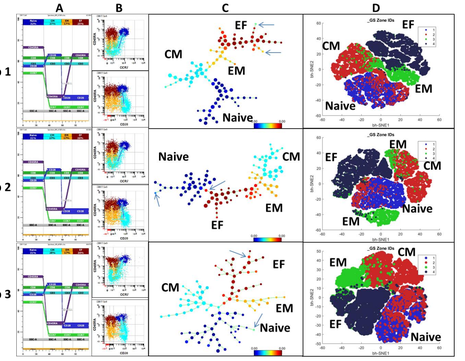

Figure 4. GemStone, SPADE, t-SNE Reproducibility: In order to assess and compare the reproducibility and specificity for GemStone, SPADE 3.0 and t-SNE; three replicate samples from a CD8 T cell study (10) were modeled by PSM as described in the manuscript. Each fitted model (see **Panel A Plots**) was then used to synthesize 30,000 CD8 T-cell events that only involved the measurements: CD3, SSC-A, CD4, CD8, CCR7 (CD197), CD28, and CD45RA. Since these data sets were synthesized, each event's stage was known and represented in the file as an added classification type of measurement with values ranging from 0 (Naive, dark blue), 1 (CM, aqua), 2 (EM, red/orange), to 3 (EF, brown). **Panel B** shows CCR7 vs. CD45RA and CD28 vs. CD45RA color-coded dot-plots with the appropriate progression vector arrows.

Each synthesized file was added to SPADE 3.0 and separately analyzed (see **Panel C Plots**). All setup parameters were in their default condition except fluorescence parameter transformations were given a hyperbolic sine cofactor of 1000. Only CCR7, CD28, and CD45RA were chosen for SPADE clustering in order to reduce the execution time. The orientation of the node positions were defined by the algorithm only. The general gestalt of the nodes was found to be dramatically different for each replicate. Most of the SPADE nodes were clustered properly except for a few (see blue arrows), which tended to be at the tips of the tree structures. In two of the three replicates, the Naïve-staged nodes were appropriately adjacent to the CM-staged nodes; however, in Rep 2, the Naïve-staged nodes were largely adjacent to the EF-staged nodes.

Each synthesized file was then added to the Cyt MatLab system. Only CCR7, CD28, and CD45RA were used to calculated the t-SNE (labelled as bh-SNE) parameters. Each of the three replicates appear in **Panel D Plots** where the four CD8 T cell stages are represented as Naïve (blue), CM (red), EM (green), and EF (dark indigo). The replicate patterns were also found to be stochastic in nature.

Figure 5. GemStone-Integrated t-SNE Plots

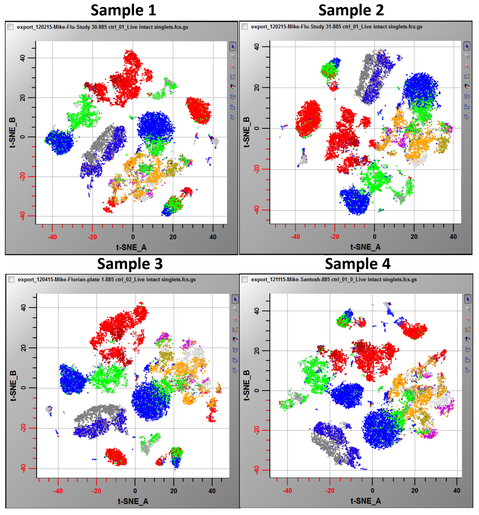

Figure 5. GemStone-Integrated t-SNE Plots: Four t-SNE plots were generated from the four same-donor samples. The markers involved in generating the two t-SNE calculated parameters, A and B, were CD45RA, CD20, CD33, CD28, CD24, CD161, CD38, CD11b, CCR6, CD94, CD86, CXCR5, CCR7, CD127, CD57, HLADR, CD19, CD4, CD8, CD11c, CD3, CD85j, CD16, CD27, CD14, CXCR3, PD-1, CD123, CD56, and CD25. The t-SNE related options were 20,000 events, perplexity=50.0, and theta=0.5. Coloring for each event came from the node colors shown in **Figures 3 and 6**. Although local neighboring structure is maintain by the algorithm, deciphering the specific populations can be challenging because of the stochastic nature of the algorithm.

Figure 6. Automated Labelling of the Integrated t-SNE Plot

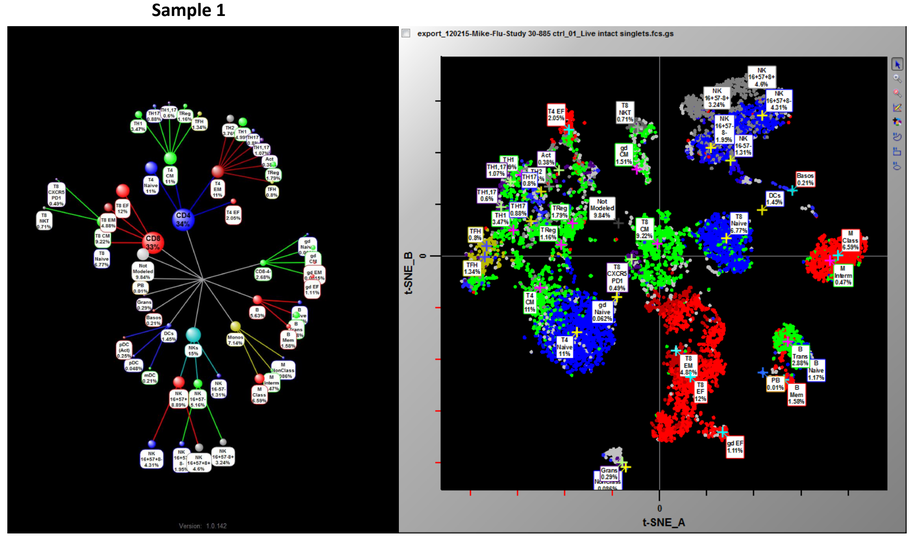

Figure 6. Automated Labelling of the Integrated t-SNE Plot: Sample 1's force-directed summary map is shown on the left-side of the figure. On the right is the associated labelled t-SNE plot with all the model-determined population landmarks identified.

Discussion

Multi-cell type models such as the one presented here (see **Figure 1**) have a number of repetitive design characteristics. Each cell type begins with a set of constant expression profiles that select for the general cell type of interest. Since mass cytometers don't currently have correlated light-scatter measurements, lineage-negative markers are normally used to clean up lineage positive markers. For six of the ten cell types, a set of staging markers were used to further separate the events into stages. Once the events are selected into cell types and stages, it is very easy to inspect the other markers for other populations that may be present using a combinatory analysis routine called TriCOM. These marker combinations can then be selected and labelled and represented in subsequently graphs or database results.

After the modeling process has concluded, expression profile overlays such as shown in **Figure 2** can represent how selected markers modulate with stage. Since this study's four samples are from a single donor, the consistent nature of the overlays demonstrates the reproducibility of probability state high-dimensional modeling. Although these overlays represent marker correlations in a very compact form, there was need for a single graphical representation of all cell type, stage, and combinatory subpopulations' relative frequencies. The force-directed graphs shown in **Figure 3** are used for this summary data. The graphs self-organize due to attraction and repulsion forces of the nodes and present a very clear picture of the relative proportions of all the populations in one or more samples.

One of the objectives of this study was to evaluate the consistency and specificity of other algorithms such as SPADE (15) and t-SNE (18-20). Not shown in this study were the negative results from the Wanderlust algorithm (23) where it could not find the correct progression for any of the data shown in **Figure 4**. Although SPADE and t-SNE were both stochastic in nature (see **Figures 4 and 5**), t-SNE was found to be a better choice to integrate into GemStone because it did not have as many setup options and it also appeared to better separate the cellular populations defined in high-dimensional space.

The capability of observing the high-dimensional relationships of many of the subpopulations is helpful for developing models that are more consistent with the data. For example, the small isthmus of NK cells leading into the CD57- part of the NK distribution has the phenotype CD16- CD56 bright (see **NK inset in Figure 6**). As soon as they intersect the CD57- CD8- pole of the NK distribution, they appear to bifurcate into CD8- and CD8 dim populations. Both of these populations then appear to up-regulate CD57. The ability to easily integrate the modeling and t-SNE results allows the modeler to better represent observed patterns into concrete models. These models can then be shared with other scientists and their veracity subjected to closer scrutiny.

The relatively large gaps between the populations shown in **Figures 5 and 6** are mainly due to performing a t-SNE analysis on a single sample. If a study stochastically samples a series of files and then performs a t-SNE analysis, there will be an inevitable "blurring" effect on the cellular populations due to additional donor-to-donor intensity variabilities. The increased resolution of analyzing one sample should make it possible to better find aberrant or currently unidentified and unmodeled populations.

Summary

1. A single probability state model was constructed that automatically analyzed all the high-dimensional Helios files in this study.

2. A self-organizing force-directed graph was presented that provided clear summary graphs of the relative proportions of all the modeled cell types in the sample.

3. Both SPADE and t-SNE have stochastic characteristics when applied to replicate samples whereas probability state modeling results are very reproducible.

4. The integration of GemStone and t-SNE allows the automatic labelling of all the modelled populations in t-SNE plots.

5. The labelled subpopulations provide important landmarks in the t-SNE plots which aids in both subsequent model design and detection of aberrant or currently unknown populations.

References

1. Bagwell C; Bagwell, CB., assignee. Probability state models. 2007.
2. Bagwell C. Probability state modeling: a new paradigm for cytometric analysis. In: Litwin V, Marder, P., editor. Flow cytometry in Drug Discovery and Development. Hoboken NJ: John Wiley and Sons Inc; 2010. p 281.
3. Bagwell CB. Breaking the dimensionality barrier. Methods Mol Biol 2011;699:31-51.
4. Bagwell C. A new paradigm for cytometric analysis. In: Kottke-Marchant K, Davis, B H., editor. Laboratory Hematology Practice: Wiley-Blackwell Publishing Ltd; 2012.
5. Bagwell CB, Hunsberger BC, Herbert DJ, Munson ME, Hill BL, Bray CM, Preffer FI. Probability state modeling theory. Cytometry Part A 2015:n/a-n/a.
6. Miller DT, Hunsberger BC, Bagwell CB. Automated analysis of GPI-deficient leukocyte flow cytometric data using GemStone. Cytometry B Clin Cytom 2012;82:319-24.
7. Herbert D, Miller D, Bagwell C. Automated analysis of flow cytometric data for CD34+ stem cell enumeration using a probability state model. Cytometry B Clin Cytom 2012;828:313-318.
8. Wong L, Hunsberger BC, Bruce Bagwell C, Davis BH. Automated quantitation of fetomaternal hemorrhage by flow cytometry for HbF-containing fetal red blood cells using probability state modeling. Int J Lab Hematol 2013;35:548-54.
9. Wong L, Hill BL, Hunsberger BC, Bagwell CB, Curtis AD, Davis BH. Automated analysis of flow cytometric data for measuring neutrophil CD64 expression using a multi-instrument compatible probability state model. Cytometry B Clin Cytom 2014.
10. Inokuma MS, Maino VC, Bagwell CB. Probability state modeling of memory CD8(+) T-cell differentiation. J Immunol Methods 2013;397:8-17.
11. Bagwell C, Hill B, Wood B, Wallace P, Alrazzak M, Kelliher A, Preffer F. Human B-Cell and Progenitor Stages As Determined by Probability State Modeling of Multidimensional Cytometry Data. Cytometry B Clin Cytom 2015.
12. Tanner SD, Bandura DR, Ornatsky O, Baranov VI, Nitz M, Winnik MA. Flow cytometer with mass spectrometer detection for massively multiplexed single-cell biomarker assay. Pure and Applied Chemistry 2008;80:2627-2641.
13. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, Pavlov S, Vorobiev S, Dick JE, Tanner SD. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. Anal Chem 2009;81:6813-22.
14. Paul N. Case studies and patient analysis utilizing mass cytometry. 2016 05May2016; Boston, Merck Research Laboratories.
15. Qiu P. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nature Biotechnology 2011;29:6.
16. Linderman M. CytoSPADE: high-performance analysis and visualization of high-dimensional cytometry data. Bioinformatics 2012;28:2.
17. Mair F, Hartmann FJ, Mrdjen D, Tosevski V, Krieg C, Becher B. The end of gating? An introduction to automated analysis of high dimensional cytometry data. Eur J Immunol 2016;46:34-43.
18. van der Maaten L. Learning a Parametric Embedding by Preserving Local Structure. AISTATS. TICC, Tilburg University P.O. Box 90153, 5000 LE Tilburg, The Netherlands; 2009.
19. van der Maaten L. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research 2014;15:1-21.
20. Amir el AD. viSNE and Wanderlust, two algorithms for the visualization and analysis of high-dimensional single-cell data: Columbia University; 2014.
21. Bendall SC, Simonds EF, Qiu P, Amir el AD, Krutzik PO, Finck R, Bruggner RV, Melamed R, Trejo A, Ornatsky OI and others. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 2011;332:687-96.
22. Ferrell PB, Jr., Diggins KE, Polikowsky HG, Mohan SR, Seegmiller AC, Irish JM. High-Dimensional Analysis of Acute Myeloid Leukemia Reveals Phenotypic Changes in Persistent Cells during Induction Therapy. Plos One 2016;11:e0153207.
23. Bendall S, Davis K, Amir el A, Tadmor M, Simonds E, Chen T, Shenfeld D, Nolan G, Pe'er D. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell 2014;157:714-25.