Methods 82 (2015) 55–63

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Methods for discovery and characterization of cell subsets in high dimensional mass cytometry data

Kirsten E. Diggins [a], P. Brent Ferrell Jr. [b], Jonathan M. Irish [a,c,*]

[a] Cancer Biology, Vanderbilt University School of Medicine, United States
[b] Medicine/Division of Hematology–Oncology, Vanderbilt University School of Medicine, United States
[c] Pathology, Microbiology and Immunology, Vanderbilt University School of Medicine, United States

ARTICLE INFO

Article history:
Received 16 January 2015
Received in revised form 24 April 2015
Accepted 6 May 2015
Available online 13 May 2015

Keywords:
Mass cytometry
Flow cytometry
Single cell biology
Unsupervised analysis
Machine learning

ABSTRACT

The flood of high-dimensional data resulting from mass cytometry experiments that measure more than 40 features of individual cells has stimulated creation of new single cell computational biology tools. These tools draw on advances in the field of machine learning to capture multi-parametric relationships and reveal cells that are easily overlooked in traditional analysis. Here, we introduce a workflow for high dimensional mass cytometry data that emphasizes unsupervised approaches and visualizes data in both single cell and population level views. This workflow includes three central components that are common across mass cytometry analysis approaches: (1) distinguishing initial populations, (2) revealing cell subsets, and (3) characterizing subset features. In the implementation described here, viSNE, SPADE, and heatmaps were used sequentially to comprehensively characterize and compare healthy and malignant human tissue samples. The use of multiple methods helps provide a comprehensive view of results, and the largely unsupervised workflow facilitates automation and helps researchers avoid missing cell populations with unusual or unexpected phenotypes. Together, these methods develop a framework for future machine learning of cell identity.

1. Introduction

1.1. High dimensional single cell biology

Single cell biology is transforming our understanding of the biological mechanisms driving human diseases and healthy tissue development [1]. Mass cytometry is a recently developed technology that enables simultaneous detection of more than 40 features on individual cells [2,3]. High dimensional mass cytometry measurements are single cell, quantitative, and well-suited to unsupervised computational analysis. New analysis tools have been created to take advantage of the massive amounts of data that result from high content single cell techniques like mass cytometry. Variations of many of these tools have been developed and applied for gene expression analysis, a field facing similar problems with data dimensionality. These tools draw on advances in machine learning and statistics that are not yet widely applied in biological studies. Many of these tools are complementary and address different aspects of data analysis, and it can be challenging for biologists to know when and how to use these tools to get the most out of their data. Advances have also been made in automating and standardizing the flow cytometry data analysis workflow [4–6]. Here, we present a modular workflow focused on high dimensional single cell analysis that combines multiple tools to provide a comprehensive view of both cells and populations. Rather than making the workflow fully automated, the goal here was to combine the complementary benefits of expert analysis and machine learning. This approach maintains single cell views, provides automatic population assignment for each cell, and facilitates statistical comparison of the key cellular features that characterized each population. This semi-supervised workflow facilitates comparison of populations discovered by different computational approaches, in different clinical samples, or using different biological features (e.g. RNA expression, cell surface protein expression, and cell signaling).

An advantage of traditional analysis in flow cytometry is the reliance on identification of known, prominent populations with strong supporting biology in the literature. Given the typical panel size for fluorescent experiments, this type of supervised analysis is fast and usually adequate. Unfortunately, expert manual gating has been shown to be particularly prone to inter-operator variability [7] and a tendency to overlook cell populations [8–10]. Recent

* Corresponding author at: Vanderbilt University School of Medicine, 740B Preston Building, 2220 Pierce Avenue, Nashville, TN 37232-6840, United States.
E-mail address: jonathan.irish@vanderbilt.edu (J.M. Irish).

http://dx.doi.org/10.1016/j.ymeth.2015.05.008
1046-2023/

efforts have developed new tools for high dimensional cytometry data that bring in elements of machine learning and statistical analysis, including clustering [11–14], dimensionality reduction [8], variance maximization [15], mixture modeling [6,16–18], spectral clustering [19], neural networks [20], and density-based automated gating [21]. Here, we highlight use of these tools in a sequential single cell bioinformatics workflow (Table 1). In particular, different tools address aspects of data visualization, dimensionality reduction, population discovery, and feature comparison. It can be valuable to apply multiple tools in order to view data in different ways and fully extract biological meaning at the single cell level (Fig. 1) and the population level (Figs. 2 and 3). After identifying cell subsets with the aid of computational tools, measured features, such as protein expression in the examples here, can be compared between and within the subsets. Traditional statistics used include medians, variance, and fold changes. Other statistical methods such as histogram statistics and probability binning have also been used to compare distributions in flow cytometry data [22–24].

1.2. Overview of the analysis workflow

The workflow presented here was applied to a CyTOF dataset from the analysis of healthy human bone marrow and a diagnostic sample of blood from a patient with acute myeloid leukemia (AML). The annotated FCS files and a step-by-step guide are available online from Cytobank (www.cytobank.org/irishlab) [25] and FlowRepository (http://flowrepository.org/experiments/640) [26]. This workflow was developed for use with high-dimensional mass cytometry data. However, it can also be applied to fluorescent flow cytometry data. The main steps presented consist of event restriction, population discovery, and population characterization. Each of these aspects of data analysis can be achieved with a variety of techniques (Table 1), and some tools address multiple steps. By sequentially combining three different techniques, this workflow draws on the strengths of specific tools, keeps biologists in touch with single cell views, and enables analysis of data from different studies and single cell platforms.

In the case of the example dataset here, the overall biological goal was to identify and compare three populations of cells: leukemia cells (AML blasts) and non-malignant cells (non-blasts) in the blood of a leukemia patient, and bone marrow cells from a healthy donor. In the analysis workflow, cell events were first manually gated based on event length and DNA content to include intact, single cells (Fig. 1) [11]. Next, visualization of stochastic neighbor embedding (viSNE) was used to identify and gate major subsets (Fig. 1). Gated cells from healthy bone marrow and AML were then analyzed by spanning-tree progression analysis of density-normalized events (SPADE) to discover and compare cell subsets (Fig. 2). Finally, the cell subsets identified by SPADE were further characterized using complete linkage hierarchical clustering and a heatmap in R (Fig. 3). The details of mass cytometry data collection and processing prior to initial cell selection (gating) are not covered in detail here. These early steps include experiment design, collection of data at the instrument (and instrument setup), any normalization, and transformation of the data to an appropriate scale (Table 1).

The initial event restriction step that begins the workflow focuses the analysis on populations of cells. The goal at this step is to remove events that do not contribute useful information while making minimal changes to the data and not over-focusing. Event restriction is traditionally performed using biaxial gating (Table 1), but given the high dimensionality of mass cytometry data, use of viSNE (Fig. 1) can simplify the process of distinguishing initial populations and avoid overlooking cells with unusual or unanticipated phenotypes. The second step, cell subset identification, is also traditionally performed by expert gating (Table 1). However, clustering tools such as SPADE [12] (Fig. 2), Misty Mountain [13], and Citrus [14], among others, can be used to automatically assign cells to groups or clusters in high dimensional data. In the workflow here, the goal is to find all the phenotypic clusters of cells in healthy bone marrow, AML blasts, and non-blast cells from AML blood (Fig. 2). As the final step, characterization of discovered cell

Table 1
A modular machine learning workflow for semi-supervised high-dimensional single cell data analysis.

Analysis step		Traditional	Additional methods[§]	Method here
Data collection	(1) Panel design	Human expert	–	–
	(2) Data collection	Human expert	–	–
Data processing	(3) Cell event parsing	Instrument software	Bead normalization and event parsing [39]	–
	(4) Scale transformation	Human expert	Logicle [47]	–
Distinguishing initial populations	(5) Live single cell gating (6) Focal population gating	Biaxial gating + human expert	No event restriction, AutoGate [61]	viSNE + human expert (Fig. 1)[†]
Revealing cell subsets	(7) Select features	Human expert	Statistical threshold [53]	Human expert[†]
	(8) Reduce dimensions or transform data	N/A	Heat plots [62], SPADE [12], *t*-SNE [63], viSNE [8], ISOMAP [27], LLE [29], PCA in R/flowCore [64]	SPADE[†], viSNE
	(9) Identify clusters of cells	Human expert	SPADE, *k*-medians, R/flowCore, flowSOM [65], Misty Mountain [13], JCM [30], ACCSENSE [66], DensVM [28], AutoGate, *Citrus* [14]	SPADE (Fig. 2)[†], viSNE + human expert (Fig. 1)
	(10) Cluster refinement	Human expert	Citrus, DensVM, R/flowCore	–
Characterizing cell subsets	(11) Feature comparison	Select biaxial single cell views	viSNE, SPADE, heatmaps [25,53], histogram overlays [25,53], violin or box and whiskers plots [64], wanderlust [31], gemstone	Heatmaps (Fig. 3A)[†], viSNE (Fig. 1C), SPADE (Fig. 2C)
	(12) Model populations	N/A	Median [53], JCM, PCA	–
	(13) Learn cell identity	Human expert	–	Human expert[†] (Figs. 1B, 2B, and 3B)
	(14) Statistical testing	Prism, excel	R/flowCore	–

[§] Methods with broad application (e.g. R/flowCore) are listed minimally at select steps based on particular strengths or published applications.
[†] Denotes the primary approach used at each step in the sequential analysis workflow shown here.

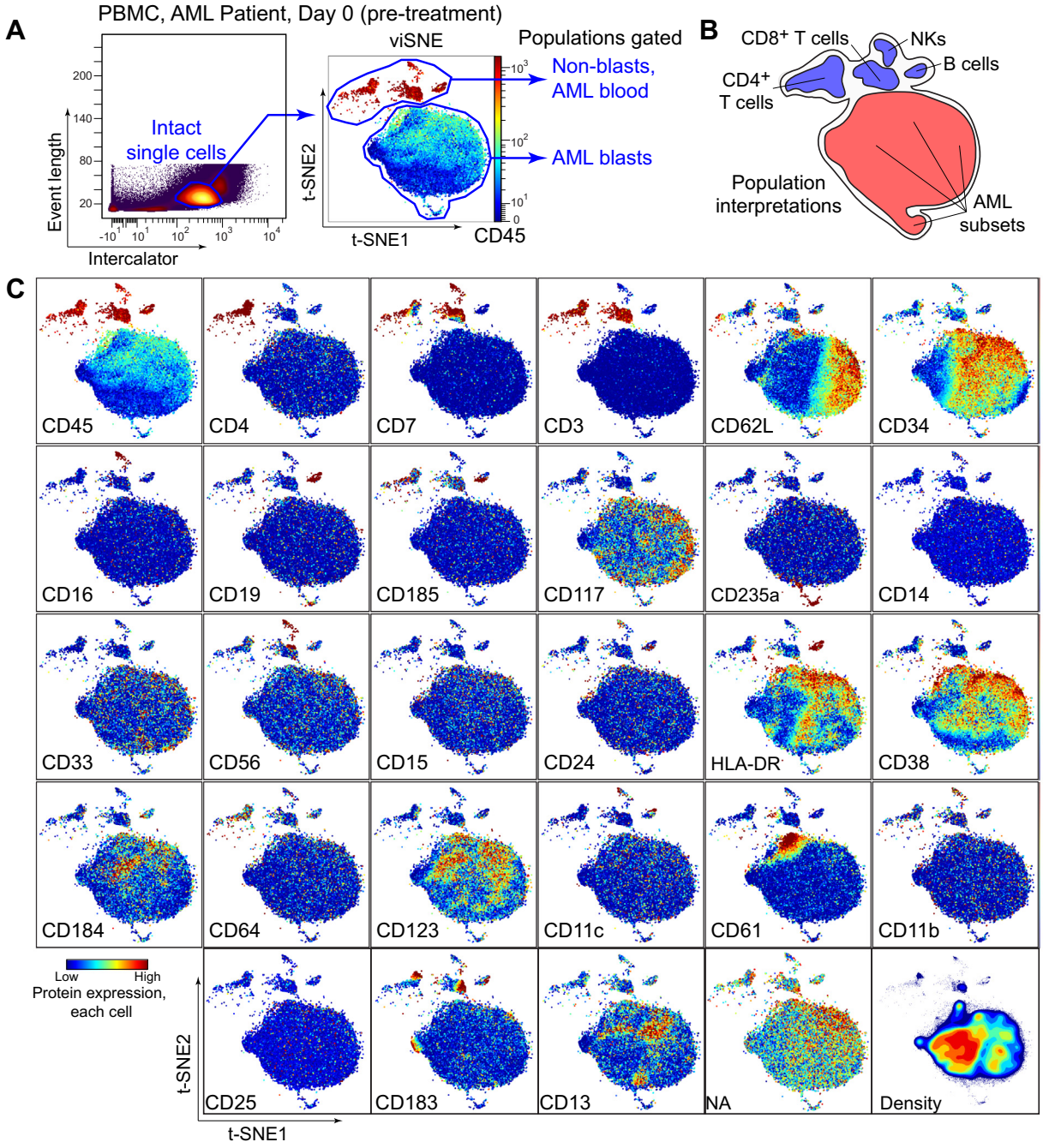

Fig. 1. Distinguishing initial populations with viSNE analysis of per-cell protein expression and expert gating. Plots show the use of viSNE to obtain a comprehensive single cell view and to initially distinguish cancerous and non-malignant cells in the blood of an AML patient. (A) Expert analysis of mass cytometry data identified intact single cells using event length and intercalator uptake. Subsequent viSNE analysis arranged cells along unitless *t*-SNE axes according to per-cell expression of 27 proteins. Expression of CD45 protein is shown for each cell on a heat scale. viSNE automatically arranged leukemia cells in one area of the map and facilitated selection of AML blast and non-blast cells by expert gating. Populations identified by viSNE and expert gating were subsequently analyzed by SPADE (Fig. 2). (B) Human interpretation of population identities based on viSNE analysis is shown. (C) Plots show expression of the 27 proteins, nucleic acid intercalator (NA), and density measured per cell.

subsets takes place downstream of manual gating or automated discovery tool implementation, and generally consists of feature expression comparison with heatmaps, violin plots, and histogram overlays for visualization, as well as data modeling and other statistical analysis. This workflow emphasizes integration of automated, unsupervised approaches with minimal human gating and processing. This type of semi-supervised cell population discovery and characterization can decrease human bias and variability and identify phenotypically unusual or rare cell subpopulations.

1.3. Advantages of machine learning tools: dimensionality reduction, clustering, and modeling

Not all tools perform the same analysis functions. Three functions that are useful for high-content single cell analysis include dimensionality reduction, clustering of cells into populations, and modeling. SPADE and viSNE both include dimensionality reduction steps that project multi-dimensional data into a lower dimensional space for visualization and further interpretation. These algorithms

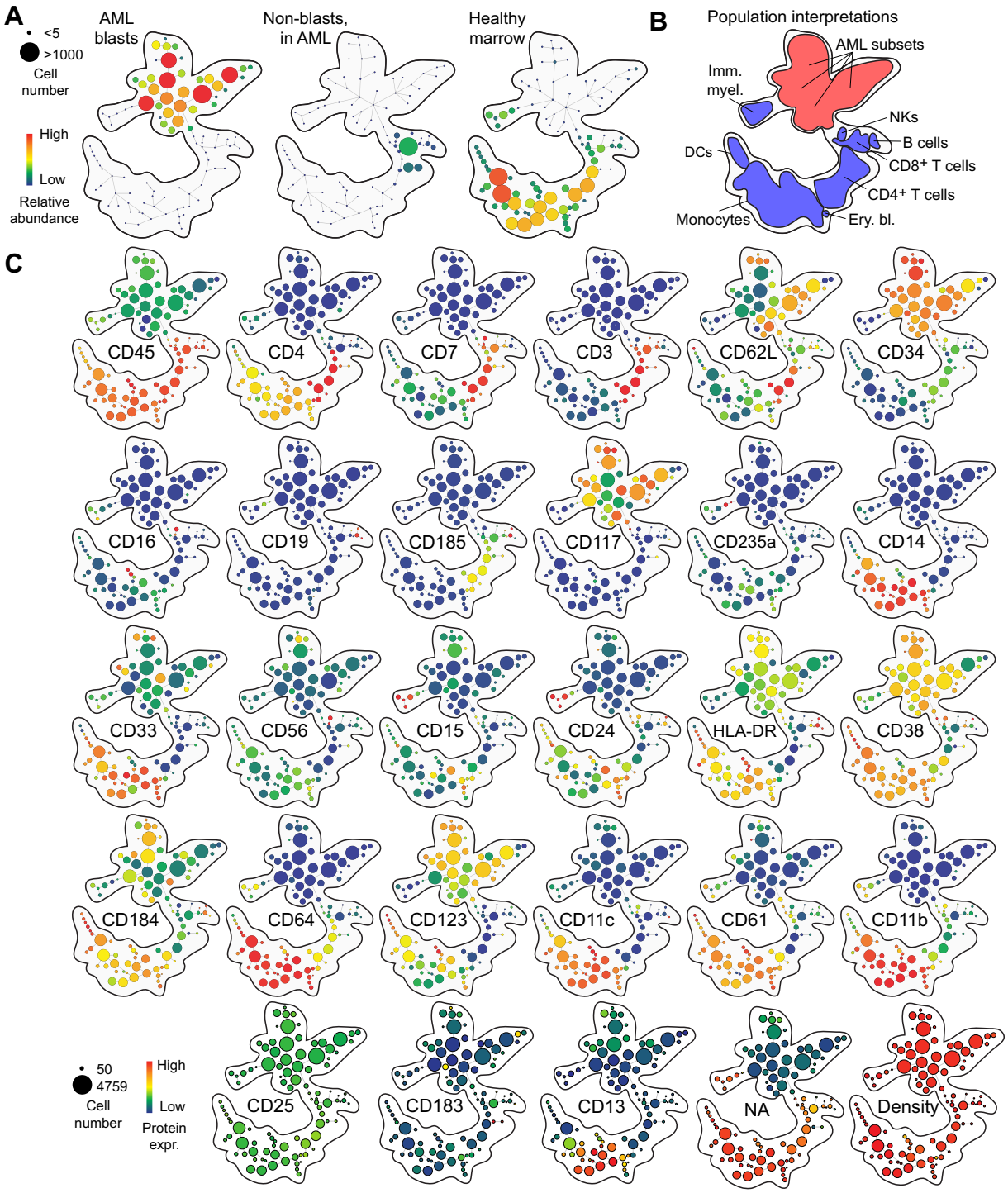

Fig. 2. Revealing cell subsets with SPADE analysis of population hierarchy, cell abundance, and median protein expression. Plots show the use of SPADE to reveal clusters of cell subsets in cell populations identified by expert analysis and viSNE (Fig. 1). (A) SPADE analysis identified distinct population clusters in each sample. Cell abundance is represented by size and color of each circle representing a population of cells. Phenotypically distinct cell subsets fell into different regions of the SPADE tree. (B) Human interpretation of population identities based on SPADE analysis is shown. (C) Plots show expression of the 27 proteins, nucleic acid intercalator (NA), and density measured per cell.

aim to preserve key high-dimensional phenotypic relationships between cells when visualizing and comparing them in 2D space. Depending on the structure of the data, other dimensionality reduction tools might be used (Table 1). Locally linear embedding (LLE) and isometric mapping (ISOMAP) are designed for the types of continuous phenotypic distributions seen in developmental progressions. ISOMAP accounts for geodesic distance in addition to local linear distances between high dimensional data points in order to reduce the dimensions of continuous and non-linear data [27,28]. A similar principle is applied with LLE, where locally linear embedding of similar data points in high dimensional space is preserved while allowing for a non-linear global embedding of the data during projection into low dimensional space [29]. In contrast, multidimensional scaling (MDS) and principal component analysis

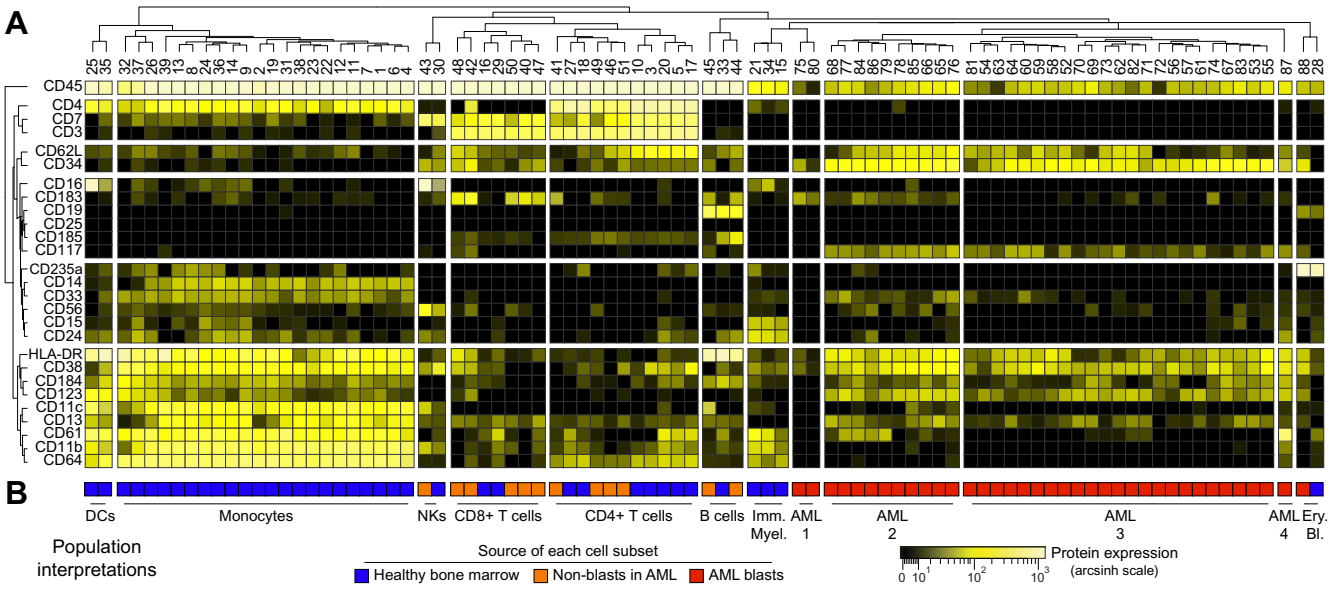

Fig. 3. Characterizing cell subsets with a heatmap analysis of median protein expression and hierarchical clustering of proteins and populations. A heatmap shows characterization of cell populations identified by SPADE (columns) according to median expression of 27 proteins (rows). For each sample analyzed in Fig. 2, cell populations identified by SPADE that contained at least 1% of total cells were included. Cell populations and proteins were arranged according to complete linkage hierarchical clustering. Heat intensity reflects the median expression of each protein for each cell population. (B) Each population contained cells from only the indicated source (healthy marrow, non-malignant cells in AML patient blood, and AML blasts). Human interpretation of population identities based on clustered heatmap analysis is shown.

(PCA) preserve linear, multi-dimensional variance. One of the advantages of PCA and other techniques, such as joint clustering and modeling [30], is the creation of a model that can be applied to newly analyzed samples. In addition to the unsupervised tools discussed here, population analysis techniques that include some supervision can be particularly useful for mapping features across known developmental progressions [31,32].

Notably, dimensionality reduction alone does not assign cells to groups. Here, dimensionality reduction with viSNE is used to aid expert interpretation of cluster identity. In this example, cells are projected onto a biaxial plot space by viSNE and then gated. Thus, viSNE is being used to see the phenotypic relationships of the cells according to all 27 protein features. This can help researchers visualize high dimensional data without losing rare populations that are best observed in single cell views. Following *t*-SNE or viSNE analysis, a human expert can look for cell clusters or major populations, as is the case here (Fig. 1), or a computational tool can identify cell clusters (Table 1), as with *t*-SNE + DensVM analysis [33]. As the workflow becomes increasingly unsupervised, it is especially important to include a single cell view early in the analysis so that expert can perform quality checks and get a sense of the overall biological results.

2. Data collection, processing, and initial population identification

2.1. Data collection

In mass cytometry, as with fluorescent flow cytometry, single cell suspensions are stained with metal-conjugated antibodies specific to molecules of interest. At the mass cytometer, cells are aerosolized and streamed single-file into argon plasma where they are atomized and ionized. The resulting ion cloud passes through a quadrupole to exclude low mass ions and enrich for reporter ions whose abundance is proportional to cellular features. These reporter ions are quantified by time of flight mass spectrometry [34] and recorded in an IMD format file. These data are typically parsed into single cell events and converted to a flow cytometry standard (FCS) file for analysis [35]. Many software programs can handle FCS files, including Cytobank (www.cytobank.org), FlowJo (www.FlowJo.com), R/Bioconductor, MATLAB, Cytoscape, and GenePattern (http://genepattern.broadinstitute.org/) [36]. Text files containing the expression matrix (where rows are cells and features are columns, and there is a median intensity value for each cell) can also be extracted directly from the IMD file from the cytometer or from the FCS file for manual analysis outside of flow cytometry analysis software. In Cytobank, export of text files with intensity values is available from the FCS file details page. An expression matrix can also be extracted from the FCS file in R and MATLAB using FCS file parsing functions. In R, the package "flowCore" can be used to extract the intensity values from the FCS file using the exprs() function [37]. In MATLAB, the tool "FCS data reader" includes the function fca_readfcs() to extract the intensity values of FCS files [38].

Here, the healthy human bone marrow sample analyzed was obtained as a de-identified sample left over from diagnostic analysis of non-cancerous tissue in the Vanderbilt Immunopathology core. Acute myeloid leukemia peripheral blood samples were collected from consented patients. In all cases, samples are collected in accordance with the Declaration of Helsinki following protocols approved by Vanderbilt University Medical Center (VUMC) Institutional Review Board. The patient blood sample evaluated here was collected at the time of diagnosis following initial evaluation and prior to any treatment.

2.2. Data processing and scale transformation

In order to prepare data for dimensionality reduction and analysis, initial processing steps aim to ensure the quality of cell events and perform appropriate scale transformations. Quality control varies by user and is especially important when conducting studies across time or using data from different instruments. Data normalization using internal bead controls can be applied as part of this data processing [39]. In this case, the two samples were collected sequentially on the same instrument and no signal normalization was required. Efforts are underway to facilitate comparison of data

among groups and centers and to report elements of panel design, instrument settings, data processing, and normalization. MIFlowCyt is a data standard set by International Society for Advancement of Cytometry (ISAC) that specifies the minimum amount of information that must be included in an FCS file to ensure reproducibility and transparency [40]. ISAC has also established a file format for classification results from flow cytometry data (CLR) [41] that handles cell classification from manual or automated identification and compliments the Gating-ML file format that was developed for sharing biaxial gate classifications [42]. Additionally, there have been efforts to standardize and compare computational flow cytometry data analysis tools. The FlowCAP project compares automated tools for cytometry data analysis using standardized datasets [5]. EuroFlow is a consortium of research groups that optimize flow cytometry protocols and analysis methods and set standards for the field of immunology and hematological studies [43]. Reporting of optimized antibody panels has also been standardized in the form of Optimized Multicolor Immunofluorescence Panels (OMIPs) [44]. Cytobank (www.cytobank.org) and FlowRepository (www.flowrepository.org) provide online access to annotated cytometry data files, including mass cytometry datasets [25,26].

Because cytometry data are log-normal, a log or log-like scale is typically used to visualize and interpret the data. Commonly used scales include inverse hyperbolic sine (arcsinh), logarithmic, and logicle (also referred to as "bi-exponential") scales [45]. Logicle or log-like scales more accurately represent the spread of data around 0 than logarithmic scales, given that modern cytometers can produce negative and zero values that cannot be transformed using logarithmic scales. The implementation of the arcsinh scale here was first used for fluorescent flow cytometry [46] and is now standard for mass cytometry. Typically, a cofactor is included as part of the arcsinh scale transformation as a way of setting a channel specific minimum significance threshold. The cofactor is set to an intensity value below which differences are not significant. For mass cytometry, cofactors typically range from 3 to 15 and depend on background and signal to noise with the detection channel and antibody-metal conjugate. In fluorescent flow cytometry, cofactors generally range from 25 to 2000 and are especially useful in correcting for channel specific differences in spreading of negative events that depend on fluorophore selection, compensation, and instrument setup. For fluorescent flow cytometry data, appropriate compensation must also be applied prior to analysis in order to correct for any spillover between channels. Algorithms have been developed for fluorescent cytometry to automatically determine scale transformations [47,48]. Applying an appropriate scale transformation prior to computational analysis is critical because it impacts quantification of distance between cells in the same way that it affects visualization of distance in biaxial plots.

2.3. Initial population identification and quality assessment

Beginning data analysis with a single cell view reveals the quality of the data and allows experts to spot rare cell subsets or artifacts that can be obscured in aggregate analysis. It is valuable to review the single cell data to verify computational analysis results, and it is vital in publications to provide representative single cell views of findings. Here, intact single cells were gated by human analysis of event length and iridium intercalator uptake (Fig. 1). This initial gating might be accomplished various ways, such as use of cisplatin exclusion to identify live cells [49]. Event length is generally higher for the mass cytometry equivalent of 'cell doublets' that can occur when the signal from two cells is not well separated in time. Intercalator uptake helps mark all cells and is proportional to nucleic acid content [11,34]. A biaxial view of each channel was then used to evaluate data quality prior to computational analysis. If no intercalator positive events are seen in this view, it suggests that there were no cells in the sample or there was an error in DNA intercalator staining. Once intact, single cells have been identified (Fig. 1A), a quick check using traditional biaxial plots or histograms can be used to ensure there is no clear overstaining. Severe overstaining results in errors while collecting data on the cytometer because event length is too great and individual cell events cannot be distinguished. Additionally, checks could be included at this step for contaminant signals. Atomic mass contaminants, such as barium and lead, can be found in water, buffers or glassware. Collecting data for the corresponding channels (137 and 138 for Ba, 208 and 209 for Pb) can be used to track these contaminants. In summary, intact single cells are first gated by a human expert. This step may be automated, but it represents an opportunity for quality assessment and initial familiarization with the data prior to computational analysis.

3. Unsupervised machine learning tools

3.1. viSNE

viSNE is a cytometry analysis tool that employs *t*-stochastic neighbor embedding (*t*-SNE) in mapping individual cells in a two or three-dimensional map that is based on their high dimensional relationships [8,50]. viSNE can be used to provide a human readable two-dimensional (2D) view of cells that are arranged in a way that approximates high-dimensional phenotypic similarity. viSNE is implemented in MATLAB and Cytobank [25], and the Cytobank implementation of viSNE is shown here (Fig. 1). viSNE can be run using a single population of cells or multiple populations drawn from one or more files. However, cell features selected for analysis must have been measured on all cell populations in a comparable way and features must be measured on comparable scales. It is sometimes helpful to subsample cell events from populations to speed the analysis or test robustness. Sampling can be 'equal' with respect to the starting populations, in order to ensure that each cell population is represented on the viSNE map by the same number of cells, or 'proportional', so that each population is represented by a number of cells proportional to its abundance. When data are thought to contain rare cell subsets, subsampling should be avoided to preserve rare cells. Initial gating can be used to focus the analysis on a population of interest and increase its relative abundance. Here, equal numbers of cells were selected from the AML PBMC and healthy marrow files for the viSNE analysis.

The cell features selected for viSNE mapping affect the structure of the viSNE map. Markers that vary highly between cell subsets will polarize subsets, placing them farther apart in tightly grouped islands. Markers with low variance on subsets will cause those cells to be placed closer together on the map. Thus, including markers that are not expressed on any cells can result in compression of islands on the map and loss of subset polarization. Features that might contribute to clustering can be selected in an unsupervised manner based on variance. For example, features that vary more in disease than in healthy controls might be particularly useful in stratifying cells associated with distinct molecular subgroups [51]. Here, all 27 markers in the panel were included in viSNE mapping because all were expressed and variable on the cells in the samples. The displayed viSNE map shows cells from the AML patient file only (Fig. 1). The resulting viSNE map showed a broad distribution of heterogeneous $CD45^{lo}$ AML cells and several distinct islands of non-blast cells (Fig. 1B). Relative protein expression as heat intensity can be viewed for each marker in the panel and are shown here for the 27 markers on the panel (Fig. 1C). The two main populations of AML blast and non-blast cells were then manually gated from the viSNE map and exported as separate

FCS files for further comparison to healthy bone marrow cells using SPADE and heatmaps. All healthy marrow cells were exported from the viSNE analysis as no additional gating was required to identify major populations. Depending on the sample and biological question, it may be useful to gate initial major populations from several or all files in this step of the analysis.

The MATLAB implementation of viSNE can be accessed through the freely downloadable *cyt* tool (http://www.c2b2.columbia.edu/danapeerlab/html/cyt.html). *Cyt* employs a user interface that allows for selection of features for mapping, selection of files or gates to be mapped, an interface for visualizing parameter intensity on a heat scale, and a tool within the interface for manually gating populations resulting viSNE map.

3.2. SPADE

SPADE is an algorithm that includes dimensionality reduction, clustering of cells into populations (also referred to as 'nodes'), and visualization using a 2D minimum spanning tree. Data must be appropriately scaled and intact cell events gated prior to SPADE analysis as described above. Here, this is done prior to viSNE gating. SPADE has been implemented in Cytobank, R, Cytoscape (http://www.cytoscape.org/), and MATLAB. In R, the package "spade" includes functions to implement individual steps of SPADE and to execute a comprehensive SPADE analysis [52]. CytoSPADE is a plugin available for use in Cytoscape that provides a user interface with the R implementation (http://www.cytospade.org). The MATLAB implementation of SPADE requires the SPADE V2.0 MATLAB tool that is freely downloadable (http://pengqiu.gatech.edu/software/SPADE/index.html). Here, the Cytobank implementation of SPADE was used to compare populations identified in viSNE guided gating. User-defined parameters for SPADE analysis include downsampling, feature selection, and a target number of nodes. Target downsampling, which can be indicated as either a percentage of cells or an absolute number, specifies how much weight to give clusters of varying density. A lower downsampling percentage increases the likelihood that sparse regions of density will be given their own clusters rather than being grouped into clusters with regions of higher density. When a sample is thought to contain rare subsets of cells, entering a lower downsampling value can help distinguish these cells as a separate population [11,12]. Feature selection in SPADE can also be based on selecting highly variable or biologically relevant markers, as described above for viSNE. The number of nodes indicates the target number of clusters (i.e. cell subsets) that the algorithm should produce, and 200 nodes is a good default for standard mass cytometry datasets containing $\sim 10^5$–10^7 total intact single cells. Including more nodes in the analysis helps to assign rare subsets to their own clusters. These clusters can be easily combined in a process called "bubbling", in which a human expert manually refines the cluster identity. A table of basic statistics, such as median intensity of each feature, is generated for each population of cells identified by SPADE and can be downloaded as a text file. Cell subsets identified by SPADE can additionally be exported as individual FCS files for further analysis, as in the heatmap analysis shown here (Fig. 3).

In the example here, three populations were analyzed. The two populations of AML blast and non-blast cells identified by viSNE (Fig. 1) were compared with the population of healthy bone marrow cells stained with the same mass cytometry antibody panel. Here, a concatenated file containing all three populations was also included to allow visualization of all cells simultaneously on one tree (Fig. 2C). SPADE can initialize with a fixed or random seed and is random in the Cytobank implementation. The same random seed can be set from run to run in the MATLAB implementation. However, when new files are added to the analysis, a different tree can still stem from the same seed, which necessitates re-running the analysis to include any additional files. For this analysis, the downsampling percentage was set to 1%, the target number of nodes was 100, and the features used for clustering were all 27 measured markers in the panel. The resulting SPADE trees are shown in Fig. 2.

Including SPADE in this analysis workflow has several advantages. First, SPADE produces a visualization of population abundances by altering the sizes of each node depending on how many cells it encompasses. For example, it can be seen in the SPADE tree that the non-blast AML cells fall almost exclusively into one node, reflecting their relative homogeneity and the lack of normal immune cell populations in the AML patient's blood (Fig. 2). Clustering with SPADE also assigns each cell to a discrete group, which minimizes analysis variability and prevents loss of cells that are outside of gated regions in manual biaxial gating. In a standard SPADE analysis, the algorithm is asked to "over cluster", producing hundreds of relatively small clusters rather than grouping cells into fewer, larger groups. This over clustering gives high resolution to improve rare subset identification and allows for a thorough annotation and characterization of all potentially discrete biological populations in the heatmap analysis.

4. Characterizing and visualizing populations

4.1. Population heatmaps

With some algorithms it is not straightforward to compare the results of an analysis of one set of samples with the results from another set of samples. For example, with SPADE it is not straightforward to map a new sample onto an existing minimum spanning tree defined using different samples. Instead, a new SPADE analysis is generally run that includes both the new and old samples. In contrast, a heatmap can be used to compare populations identified in different analysis runs of SPADE or populations identified by different clustering techniques. Heatmaps also provide a compact view that facilitates comparing many populations according to a large variety of measured features. In heatmaps, different types of biological and clinical information can also be used to group populations or assessed for association with resulting groups [46,53]. While population heatmaps provide an intuitive, high-level view of the results, they can obscure variation within subsets [25]. To address this, statistics other than median expression can be shown in the heatmap, such as variance or the 95th percentile of expression [1,54].

For the last step in the workflow here, tables of statistics for the hundreds of cell subsets identified in the three starting populations (Fig. 2) were exported from SPADE as text files listing median expression of each feature for each cell subset. Cell subsets were excluded from further analysis if they contained less than 1% of the cells in the starting population. This arbitrary threshold was set in order to exclude sparse clusters where low cell number could potentially increase the error of reported medians. Here, the 1% threshold resulted in exclusion of approximately 5% of the total cell events from heatmap characterization. The table of statistics was then imported into R using the "read.table" function from the R.utils package [55] and visualized as a hierarchically clustered heatmap using the "heatmap.2" function in the gplots package (Fig. 3A) [56]. Output of a hierarchical dendrogram as part of the heatmap can be specified as one of the input parameters of the heatmap.2 function. The R package "stats" also offers a function called "heatmap" that performs the same function as heatmap.2 with slight differences in visualization options. After the clustered heatmap was generated, expert analysis was used to assign biological classifications to each group of populations in the hierarchical

clustering, and included the same populations seen in viSNE (Fig. 1B) and SPADE (Fig. 2B): dendritic cells (DCs), monocytes, natural killer cells (NKs), CD8+ T cells, CD4+ T cells, B cells, immature myeloid cells (Imm. myel.), four subsets of AML blast cells (AML1 through AML4), and erythroid blast cells (Ery. bl.) (Fig. 3B).

Use of a clustered heatmap in the workflow allows for simultaneous visualization of several markers for the same clusters (population of cells) from multiple files. Furthermore, nodes are hierarchically clustered, and this clustering can be pruned at various levels by the user to further group the nodes into biological populations. It is also important to note that the distance between nodes has quantitative meaning in the clustered heatmap dendrogram, as opposed to the distances on the SPADE tree that are for visualization purposes and not quantitative. Heatmap analysis therefore compliments the SPADE visualization by facilitating simultaneous visualization of nodes from multiple files and by quantifying phenotypic distances between the nodes.

4.2. Other packages and flowCore

There are many R packages designed for statistical and visual analysis of flow cytometry data, including flowCore [37], flowViz [57], flowStats [58], and flowClust [58], among others. These packages include functions for producing heat maps, histograms, bar plots, biaxial density plots, and are part of efforts to automate and standardize computational analysis of cytometry data [5,6]. Apart from the R packages designed for flow cytometry data analysis, other analysis and visualization packages can be applied to single cell data. For example, box and whisker plots or violin plots can be produced to show median, range, and the distribution of the feature in each subset.

5. Other considerations for automated flow cytometry data analysis

5.1. Algorithm selection

Three major considerations when choosing tools or algorithms for flow cytometry data analysis include (1) linear vs. non-linear measurement, (2) supervised or unsupervised approaches, and (3) need for modeling. The first consideration is whether a linear or non-linear method of dimensionality reduction is best for the data. Phenotypic relationships between cells may follow a 'creode', or necessary path, that is non-linear with respect to protein expression (i.e. co-expression or co-variance of molecules is not linearly correlated with important progressions in cellular identity or trajectories in data space) [10]. In this case, nonlinear dimensionality reduction tools may better preserve the high dimensional phenotypic relationships between cells compared to tools that assume a linear relationship between variables. The second consideration is whether an unsupervised or supervised method is needed. In an exploratory analysis where novel populations are anticipated, unsupervised approaches will minimizes the risk of overlooking the populations. Lastly, a consideration is whether or not the goal of analysis is to build a model. Mixture modeling tools can be implemented for analysis of flow cytometry data that will produce a model as output for downstream analysis. Additional issues to consider include (1) selection of features, which is generally initiated by hypotheses and pragmatic concerns and then narrowed to include those features with biologically meaningful variation [51], and (2) aspects of statistical power, including sample size, cluster density, and false discovery rate (FDR). It is vital to calculate FDR or a related statistic, such as the *f*-measure, in cases where a truth is known [5].

5.2. Scalability of workflow

Biomedical studies that employ flow and mass cytometry often accrue large numbers of samples over long periods of time. This and similar workflows can be adapted to accommodate data from these large studies. In order to account for experimental or instrument variability, normalization is necessary in these cases in order for compare samples run at different times or from different instruments. Bead normalization has been optimized for use with mass cytometry to control for machine variability between runs [39,59,60]. Polystyrene beads embedded with heavy metal isotopes are run with every sample as a standard that can be used to correct MI values for each event based on technical variability. When samples accrue over a long period of time, a key consideration is that new results may not be easily mapped back to the original viSNE map or SPADE tree without re-analysis. This is one advantage of heatmaps, which compare samples according to a simple 'model' of the data, such as median expression of selected features.

This workflow as presented includes manual intervention that could be prohibitive when analyzing many data files simultaneously. While all steps of this analysis could generally be batched and automated, human review of single cell data is advantageous at workflow breakpoints to verify computational results and spot artifacts. Cytobank and other flow cytometry data analysis software allow for rapid, simultaneous viewing and pre-processing of multiple files, including scale transformation and gating. viSNE analysis can currently be run on up to 800,000 cells in Cytobank, and this limit is pragmatic, not theoretical. Many files can be run simultaneously by subsampling cells equally or proportionally from each file prior to the viSNE run. SPADE can also be run on many files simultaneously, and data files with cluster information can be quickly downloaded in a compressed folder.

Import of text files into R and selection of nodes based on the number of cells they contain can be automated and batched for highly scalable and rapid heatmap analysis. However, a potential limitation of large-scale analyses is the visualization of all nodes simultaneously on the heatmap. It may be useful in these cases to segment the SPADE tree into major populations by "bubbling" and then building separate heatmaps from each bubble rather than for the whole tree. Depending on the expected prevalence of rare cells in the dataset, the user can request fewer nodes in the SPADE run in order to decrease the final number of clusters to be analyzed and visualized on the heatmap.

6. Conclusions

Data analysis in cytometry remains largely manual, supervised, and focused on large changes in magnitude of expression. As new tools are developed to assist in gating, reduce dimensionality, and automate analysis, it is important to show biologists the value of these tools and to integrate them into workflows that can become routine. The workflow presented here blends supervised and unsupervised analysis tools so that biologists can visualize results at the single cell level while still getting an accurate view of the big picture. Combining tools also allows the analyst to visualize data in multiple ways, which can be useful to extract the most meaning from a data set. Existing tools allow for identification of populations based on single cell expression profiles and characterization of these subsets using standard statistics, including expression magnitude, marker variance, and subset abundance. Going forward, tools that quantify cellular heterogeneity, identify critical population features, and assign biological identity to machine-identified subsets will be particularly useful in filling out the toolkit.

Acknowledgments

This study was supported by R25 CA136440-04 (K.E.D.), NIH/NCI K12 CA090625 (P.B.F.), R00 CA143231-03 (J.M.I.), the Vanderbilt-Ingram Cancer Center (VICC, P30 CA68485), and VICC Young Ambassadors and VICC Hematology Helping Hands awards. Thanks to Mikael Roussel for helpful discussions of myeloid cell identity markers.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ymeth.2015.05.008.

References

[1] J.M. Irish, D.B. Doxie, Curr. Top. Microbiol. Immunol. 377 (2014) 1–21.
[2] D.R. Bandura, V.I. Baranov, O.I. Ornatsky, A. Antonov, R. Kinach, X. Lou, S. Pavlov, S. Vorobiev, J.E. Dick, S.D. Tanner, Anal. Chem. 81 (2009) 6813–6822.
[3] O. Ornatsky, D. Bandura, V. Baranov, M. Nitz, M.A. Winnik, S. Tanner, J. Immunol. Methods 361 (2010) 1–20.
[4] G. Finak, J. Frelinger, W. Jiang, E.W. Newell, J. Ramey, M.M. Davis, S.A. Kalams, S.C. De Rosa, R. Gottardo, PLoS Comput. Biol. 10 (2014) e1003806.
[5] N. Aghaeepour, G. Finak, C.A.P.C. Flow, D. Consortium, H. Hoos, T.R. Mosmann, R. Brinkman, R. Gottardo, R.H. Scheuermann, Nat. Methods 10 (2013) 228–238.
[6] S. Pyne, X. Hu, K. Wang, E. Rossin, T.I. Lin, L.M. Maier, C. Baecher-Allan, G.J. McLachlan, P. Tamayo, D.A. Hafler, P.L. De Jager, J.P. Mesirov, Proc. Natl. Acad. Sci. U.S.A. 106 (2009) 8519–8524.
[7] H.T. Maecker, A. Rinfret, P. D'Souza, J. Darden, E. Roig, C. Landry, P. Hayes, J. Birungi, O. Anzala, M. Garcia, A. Harari, I. Frank, R. Baydo, M. Baker, J. Holbrook, J. Ottinger, L. Lamoreaux, C.L. Epling, E. Sinclair, M.A. Suni, K. Punt, S. Calarota, S. El-Bahi, G. Alter, H. Maila, E. Kuta, J. Cox, C. Gray, M. Altfeld, N. Nougarede, J. Boyer, L. Tussey, T. Tobery, B. Bredt, M. Roederer, R. Koup, V.C. Maino, K. Weinhold, G. Pantaleo, J. Gilmour, H. Horton, R.P. Sekaly, BMC Immunol. 6 (2005) 13.
[8] A.D. Amir el, K.L. Davis, M.D. Tadmor, E.F. Simonds, J.H. Levine, S.C. Bendall, D.K. Shenfeld, S. Krishnaswamy, G.P. Nolan, D. Pe'er, Nat. Biotechnol. 31 (2013) 545–552.
[9] P.O. Krutzik, M.R. Clutter, G.P. Nolan, J. Immunol. 175 (2005) 2357–2365.
[10] J.M. Irish, Nat. Immunol. 15 (2014) 1095–1097.
[11] S.C. Bendall, E.F. Simonds, P. Qiu, A.D. Amir el, P.O. Krutzik, R. Finck, R.V. Bruggner, R. Melamed, A. Trejo, O.I. Ornatsky, R.S. Balderas, S.K. Plevritis, K. Sachs, D. Pe'er, S.D. Tanner, G.P. Nolan, Science 332 (2011) 687–696.
[12] P. Qiu, E.F. Simonds, S.C. Bendall, K.D. Gibbs Jr., R.V. Bruggner, M.D. Linderman, K. Sachs, G.P. Nolan, S.K. Plevritis, Nat. Biotechnol. 29 (2011) 886–891.
[13] I.P. Sugar, S.C. Sealfon, BMC Bioinform. 11 (2010) 502.
[14] R.V. Bruggner, B. Bodenmiller, D.L. Dill, R.J. Tibshirani, G.P. Nolan, Proc. Natl. Acad. Sci. U.S.A. 111 (2014) E2770–E2777.
[15] E.W. Newell, N. Sigal, S.C. Bendall, G.P. Nolan, M.M. Davis, Immunity 36 (2012) 142–152.
[16] T.R. Mosmann, I. Naim, J. Rebhahn, S. Datta, J.S. Cavenaugh, J.M. Weaver, G. Sharma, Cytometry A (2014).
[17] I. Naim, S. Datta, J. Rebhahn, J.S. Cavenaugh, T.R. Mosmann, G. Sharma, Cytometry A (2014).
[18] X. Chen, M. Hasan, V. Libri, A. Urrutia, B. Beitz, V. Rouilly, D. Duffy, E. Patin, B. Chalmond, L. Rogge, L. Quintana-Murci, M.L. Albert, B. Schwikowski, C. Milieu Interieur, Clin. Immunol. (2015).
[19] H. Zare, P. Shooshtari, A. Gupta, R.R. Brinkman, BMC Bioinform. 11 (2010) 403.
[20] D.L. Tong, G.R. Ball, A.G. Pockley, Cytometry A (2015).
[21] Y. Qian, C. Wei, F. Eun-Hyung Lee, J. Campbell, J. Halliley, J.A. Lee, J. Cai, Y.M. Kong, E. Sadat, E. Thomson, P. Dunn, A.C. Seegmiller, N.J. Karandikar, C.M. Tipton, T. Mosmann, I. Sanz, R.H. Scheuermann, Cytometry B 78 (Suppl. 1) (2010) S69–S82.
[22] M. Roederer, W. Moore, A. Treister, R.R. Hardy, L.A. Herzenberg, Cytometry 45 (2001) 47–55.
[23] C.B. Bagwell, J.L. Hudson, G.L. Irvin 3rd, J. Histochem. Cytochem. 27 (1979) 293–296.
[24] W.R. Overton, Cytometry 9 (1988) 619–626.
[25] N. Kotecha, P.O. Krutzik, J.M. Irish, in: J. Paul Robinson et al. (Eds.), Current Protocols in Cytometry/Editorial Board. Chapter 10, 2010. Unit10 17.
[26] J. Spidlen, K. Breuer, R. Brinkman, in: J. Paul Robinson et al. (Eds.), Current Protocols in Cytometry/Editorial Board. Chapter 10, 2012. Unit 10 18.
[27] J.B. Tenenbaum, V. de Silva, J.C. Langford, Science 290 (2000) 2319–2323.
[28] B. Becher, A. Schlitzer, J. Chen, F. Mair, H.R. Sumatoh, K.W.W. Teng, D. Low, C. Ruedl, P. Riccardi-Castagnoli, M. Poidinger, M. Greter, F. Ginhoux, E.W. Newell, Nat. Immunol. (2014) 1181–1189.
[29] S.T. Roweis, L.K. Saul, Science 290 (2000) 2323–2326.
[30] S. Pyne, S.X. Lee, K. Wang, J. Irish, P. Tamayo, M.D. Nazaire, T. Duong, S.K. Ng, D. Hafler, R. Levy, G.P. Nolan, J. Mesirov, G.J. McLachlan, PLoS One 9 (2014) e100334.
[31] S.C. Bendall, K.L. Davis, A.D. Amir el, M.D. Tadmor, E.F. Simonds, T.J. Chen, D.K. Shenfeld, G.P. Nolan, D. Pe'er, Cell 157 (2014) 714–725.
[32] M.S. Inokuma, V.C. Maino, C.B. Bagwell, J. Immunol. Methods 397 (2013) 8–17.
[33] B. Becher, A. Schlitzer, J. Chen, F. Mair, H.R. Sumatoh, K.W. Teng, D. Low, C. Ruedl, P. Riccardi-Castagnoli, M. Poidinger, M. Greter, F. Ginhoux, E.W. Newell, Nat. Immunol. 15 (2014) 1181–1189.
[34] O. Ornatsky, V.I. Baranov, D.R. Bandura, S.D. Tanner, J. Dick, J. Immunol. Methods 308 (2006) 68–76.
[35] J. Spidlen, W. Moore, D. Parks, M. Goldberg, C. Bray, P. Bierre, P. Gorombey, B. Hyun, M. Hubbard, S. Lange, R. Lefebvre, R. Leif, D. Novo, L. Ostruszka, A. Treister, J. Wood, R.F. Murphy, M. Roederer, D. Sudar, R. Zigon, R.R. Brinkman, Cytometry A 77 (2010) 97–100.
[36] J. Spidlen, A. Barsky, K. Breuer, P. Carr, M.D. Nazaire, B.A. Hill, Y. Qian, T. Liefeld, M. Reich, J.P. Mesirov, P. Wilkinson, R.H. Scheuermann, R.P. Sekaly, R.R. Brinkman, Source Code Biol. Med. 8 (2013) 14.
[37] H.P. Ellis B, F. Hahne, N.L. Meur, N. Gopalakrishnan, J. Spidlen, R Package Version 1.34.3, 2015.
[38] L. Balkay, MATLAB Central File Exchange, 2014 (retrieved).
[39] R. Finck, E.F. Simonds, A. Jager, S. Krishnaswamy, K. Sachs, W. Fantl, D. Pe'er, G.P. Nolan, S.C. Bendall, Cytometry A 83 (2013) 483–494.
[40] J.A. Lee, J. Spidlen, K. Boyce, J. Cai, N. Crosbie, M. Dalphin, J. Furlong, M. Gasparetto, M. Goldberg, E.M. Goralczyk, B. Hyun, K. Jansen, T. Kollmann, M. Kong, R. Leif, S. McWeeney, T.D. Moloshok, W. Moore, G. Nolan, J. Nolan, J. Nikolich-Zugich, D. Parrish, B. Purcell, Y. Qian, B. Selvaraj, C. Smith, O. Tchuvatkina, A. Wertheimer, P. Wilkinson, C. Wilson, J. Wood, R. Zigon, F. International Society for Advancement of Cytometry Data Standards Task, R.H. Scheuermann, R.R. Brinkman, Cytometry A 73 (2008) 926–930.
[41] J. Spidlen, C. Bray, I.D.S.T. Force, R.R. Brinkman, Cytometry A 87 (2015) 86–88.
[42] J. Spidlen, R.C. Leif, W. Moore, M. Roederer, F. International Society for the Advancement of Cytometry Data Standards Task, R.R. Brinkman, Cytometry A 73A (2008) 1151–1157.
[43] T. Kalina, J. Flores-Montero, V.H. van der Velden, M. Martin-Ayuso, S. Bottcher, M. Ritgen, J. Almeida, L. Lhermitte, V. Asnafi, A. Mendonca, R. de Tute, M. Cullen, L. Sedek, M.B. Vidriales, J.J. Perez, J.G. te Marvelde, E. Mejstrikova, O. Hrusak, T. Szczepanski, J.J. van Dongen, A. Orfao, C. EuroFlow, Leukemia 26 (2012) 1986–2010.
[44] Y. Mahnke, P. Chattopadhyay, M. Roederer, Cytometry A 77 (2010) 814–818.
[45] L.A. Herzenberg, J. Tung, W.A. Moore, L.A. Herzenberg, D.R. Parks, Nat. Immunol. 7 (2006) 681–685.
[46] J.M. Irish, J.H. Myklebust, A.A. Alizadeh, R. Houot, J.P. Sharman, D.K. Czerwinski, G.P. Nolan, R. Levy, Proc. Natl. Acad. Sci. U.S.A. 107 (2010) 12747–12754.
[47] W.A. Moore, D.R. Parks, Cytometry A 81A (2012) 273–277.
[48] D. Parks, M. Roederer, W.A. Moore, Cytometry A 59A (2004) 87.
[49] H.G. Fienberg, E.F. Simonds, W.J. Fantl, G.P. Nolan, B. Bodenmiller, Cytometry A 81 (2012) 467–475.
[50] L. van der Maaten, G. Hinton, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[51] J. Irish, R. Hovland, P. Krutzik, O. Perez, O. Bruserud, B. Gjertsen, G. Nolan, Cell 118 (2004) 217–228.
[52] Q.P. Linderman M, E. Simonds, Z. Bjornson, R Package Version 1.14.0, 2011 <http://cytospade.org>.
[53] J.M. Irish, R. Hovland, P.O. Krutzik, O.D. Perez, O. Bruserud, B.T. Gjertsen, G.P. Nolan, Cell 118 (2004) 217–228.
[54] N. Kotecha, N.J. Flores, J.M. Irish, E.F. Simonds, D.S. Sakai, S. Archambeault, E. Diaz-Flores, M. Coram, K.M. Shannon, G.P. Nolan, M.L. Loh, Cancer Cell 14 (2008) 335–343.
[55] H. Bengtsson, 2015 <http://cran.r-project.org/web/packages/R.utils/R.utils.pdf>.
[56] B.B. Gregory, R. Warnes, L. Bonebakker, R. Gentleman, W.H.A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz, B. Venables, R Package Version 2.17.0, 2015 <http://cran.r-project.org/web/packages/gplots/index.html>
[57] G.R. Ellis B, F. Hahne, N.L. Meur, D. Sarkar, R Package Version 1.31.1, 2015.
[58] H.F. Lo K, R. Brinkman, R. Gottardo, BMC Bioinform. 10 (2009).
[59] A.I. Abdelrahman, S. Dai, S.C. Thickett, O. Ornatsky, D. Bandura, V. Baranov, M.A. Winnik, J. Am. Chem. Soc. 131 (2009) 15276–15283.
[60] A.I. Abdelrahman, O. Ornatsky, D. Bandura, V. Baranov, R. Kinach, S. Dai, S.C. Thickett, S. Tanner, M.A. Winnik, J. Anal. At. Spectrom. 25 (2010) 260–268.
[61] S. Meehan, G. Walther, W. Moore, D. Orlova, C. Meehan, D. Parks, E. Ghosn, M. Philips, E. Mitsunaga, J. Waters, A. Kantor, R. Okamura, S. Owumi, Y. Yang, L.A. Herzenberg, L.A. Herzenberg, Immunol. Res. 58 (2014) 218–223.
[62] J.M. Irish, J.H. Myklebust, A.A. Alizadeh, R. Houot, J.P. Sharman, D.K. Czerwinski, G.P. Nolan, R. Levy, Proc. Natl. Acad. Sci. U.S.A. 107 (2010) 12747–12754.
[63] S.R. Geoffrey Hinton, Advances in Neural Information Processing Systems, 2002.
[64] F. Hahne, N. LeMeur, R.R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain, R. Gentleman, BMC Bioinform. 10 (2009) 106.
[65] S. Van Gassen, B. Callebaut, M.J. Van Helden, B.N. Lambrecht, P. Demeester, T. Dhaene, Y. Saeys, Cytometry A (2015).
[66] K. Shekhar, P. Brodin, M.M. Davis, A.K. Chakraborty, Proc. Natl. Acad. Sci. U.S.A. 111 (2014) 202–207.