



Research Techniques Made Simple: Mass Cytometry Analysis Tools for Decrypting the Complexity of Biological Systems

Tiago R. Matos^{1,2,3}, Hongye Liu^{1,2} and Jerome Ritz^{1,2}

Mass cytometry by time-of-flight experiments allow analysis of over 40 functional and phenotypic cellular markers simultaneously at the single-cell level. The data dimensionality escalation accentuates limitations, inherent to manual analysis, as being subjective, labor-intensive, slow, and often incapable of showing the detailed features of each unique cell within populations. The subsequent challenge of examining, visualizing, and presenting mass cytometry data has motivated continuous development of dimensionality reduction methods. As a result, an increasing recognition of the inherent diversity and complexity of cellular networks is emerging, with the discovery of unexpected cell subpopulations, hierarchies, and developmental pathways, such as those existing within the immune system. Here, we briefly review some frequently used and accessible mass cytometry data analysis tools, including principal component analysis (PCA); spanning-tree progression analysis of density-normalized events (SPADE); t-distributed stochastic neighbor embedding (t-SNE)—based visualization (viSNE); automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE); and cluster identification, characterization, and regression (CITRUS). Mass cytometry, used together with these innovative analytic tools, has the power to lead to key discoveries in investigative dermatology, including but not limited to identifying signaling phenotypes with predictive value for early diagnosis, prognosis, or relapse and a thorough characterization of intratumor heterogeneity and disease-resistant cell populations, that may ultimately unveil novel therapeutic approaches.

Journal of Investigative Dermatology (2017) 137, e43–e51; doi:10.1016/j.jid.2017.03.002

CME Activity Dates: 12 April 2017

Expiration Date: 11 April 2018

Estimated Time to Complete: 1 hour

Planning Committee/Speaker Disclosure: All authors, planning committee members, CME committee members and staff involved with this activity as content validation reviewers have no financial relationship(s) with commercial interests to disclose relative to the content of this CME activity.

Commercial Support Acknowledgment: This CME activity is supported by an educational grant from Lilly USA, LLC.

Description: This article, designed for dermatologists, residents, fellows, and related healthcare providers, seeks to reduce the growing divide between dermatology clinical practice and the basic science/current research methodologies on which many diagnostic and therapeutic advances are built.

Objectives: At the conclusion of this activity, learners should be better able to:

- Recognize the newest techniques in biomedical research.
- Describe how these techniques can be utilized and their limitations.
- Describe the potential impact of these techniques.

CME Accreditation and Credit Designation: This activity has been planned and implemented in accordance with the accreditation requirements and policies of the Accreditation Council for Continuing Medical Education through the joint providership of William Beaumont Hospital and the Society for Investigative Dermatology. William Beaumont Hospital is accredited by the ACCME to provide continuing medical education for physicians.

William Beaumont Hospital designates this enduring material for a maximum of 1.0 AMA PRA Category 1 Credit(s)[™]. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

Method of Physician Participation in Learning Process: The content can be read from the *Journal of Investigative Dermatology* website: <http://www.jidonline.org/current>. Tests for CME credits may only be submitted online at <https://beaumont.cloud-cme.com/RTMS-May17> – click ‘CME on Demand’ and locate the article to complete the test. Fax or other copies will not be accepted. To receive credits, learners must review the CME accreditation information; view the entire article, complete the post-test with a minimum performance level of 60%; and complete the online evaluation form in order to claim CME credit. The CME credit code for this activity is: 21310. For questions about CME credit email cme@beaumont.edu.

¹Division of Hematologic Malignancies, Dana-Farber Cancer Institute, Boston, Massachusetts, USA; ²Harvard Medical School, Boston, Massachusetts, USA; and ³Academic Medical Center, Department of Dermatology, University of Amsterdam, Amsterdam, The Netherlands

Correspondence: Tiago R. Matos, Department of Dermatology, Room L3-119, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands. E-mail: tiagoreismatos@gmail.com

Abbreviations: ACCENSE, automatic classification of cellular expression by nonlinear stochastic embedding; CITRUS, cluster identification, characterization, and regression; CyTOF, mass cytometry by time-of-flight mass spectrometry; FCS, Flow Cytometry Standard; PCA, principal component analysis; SPADE, spanning-tree progression analysis of density-normalized events; t-SNE, t-distributed stochastic neighbor embedding; viSNE, t-distributed stochastic neighbor embedding—based visualization

SUMMARY POINTS

- New methods are being continuously developed to analyze and best represent multidimensional, complex CyTOF data.
- Principal component analysis (PCA) provides a visualization of the data in three-dimensional space and identifies the parameters with the most variance among the dataset.
- Spanning-tree progression analysis of density-normalized events (SPADE) clusters cells into a minimum-spanning hierarchical tree for two-dimensional visualization.
- In t-distributed stochastic neighbor embedding (t-SNE)-based visualization (viSNE) and automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE), each single cell data point has a unique location in a two-dimensional representation, reflecting the cells' immunophenotypic similarity or differences in high-dimensional space.
- Cluster identification, characterization, and regression (CITRUS) identifies cellular features that correlate to an experimental endpoint of interest.

INTRODUCTION

New methods are being developed to examine, visualize, and present the multidimensional complexity of cellular function and identity and the role of individual cells within biological systems. Mass cytometry by time-of-flight mass spectrometry (CyTOF)¹ currently has the capacity to allow investigation of 40 or more distinct parameters at the single-cell level (Figure 1). Although the technique has not yet been widely adopted within the field of investigative dermatology, it has potential to, for example, allow identification of cell signals for early diagnosis in cutaneous T-cell lymphoma, allow early detection or predict relapse in psoriasis and atopic dermatitis, and allow thorough characterization of drug-resistant cell populations in skin cancer, eventually unveiling new therapies. The large amount of data generated and potential of the technique to delineate rare cell subsets has driven the need to develop dimensionality reduction methods and analysis algorithms to best analyze and represent mass cytometry data. A significant limitation of traditional data clustering methods through biaxial plots and histograms, such as has been used to represent traditional flow cytometry data, is that pre-existing knowledge of the defining markers of each population is required. This limits the ability of researchers to discover unexpected cellular subsets and does not allow examination of system-level phenotypic diversity. Furthermore,

manual analysis of individual markers and combinations of markers is a subjective, slow, and labor-intensive process, which results in a significant scalability restriction and can introduce several inherent biases. Although CyTOF technology and experimental methodology have been described in detail in previous reviews (Doan et al., 2015; Matos et al., 2017), comprehensive understanding is also required with respect to the tools available for analysis of high-dimensional datasets to make meaningful use of the results. In this short review, we focus on some of the most commonly used and accessible novel CyTOF data analysis tools, including principal component analysis (PCA), spanning-tree progression analysis of density-normalized events (SPADE), t-distributed stochastic neighbor embedding (t-SNE)-based visualization (viSNE), automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE), and cluster identification, characterization, and regression (CITRUS).

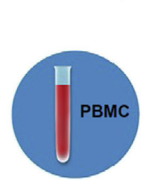
DIMENSION REDUCTION AND VISUALIZATION ALGORITHMS**PCA**

PCA is a well-established and widely used tool for visualizing multidimensional data that was adopted to analyze large mass cytometry datasets (Bendall et al., 2011; Jackson, 1991; Newell et al., 2012). PCA identifies those parameters among a certain dataset that present the most variance by generating linear combinations from a large list of parameters into new compound variables (principal components). As a result, the quotient of the relative variation of each principle component over the total variance gives an idea of the effectiveness of each component in separating out data points. In addition, PCA results in models that can be used to project new data points in linear time. For example, it allows graphical visualization of the expression intensity of several functional markers (y-axis) throughout the cell differentiation process (x-axis) (Figure 2). PCA also allows visualization of the data in three-dimensional space, often prominently displaying the first three data components of maximal variance. However, this feature can also be a limitation, because it may mask noteworthy biological differences that are more subtle variances in the data. Another constraint is the inherent assumption that the given data are parametric. PCA also represents the data through linear projections, which may not be representative of the inherent structure of the original data. To overcome this constraint, nonlinear methods such as t-SNE (described in following sections) were developed for high-dimensional data analysis. Newell et al. (2012) used PCA to represent simultaneously 25 markers from a single cell sample, hence quantifying the expression of functional markers among several CD8⁺ T-cell subsets. This representation method displayed a greater phenotypic and functional complexity among CD8⁺ T cells than previously appreciated (Figure 2). The holistic study of many functional and phenotypic markers and their expression levels through several differentiation subsets would not be possible by conventional manual analysis. This study also observed that subsets that develop in response to different viruses have distinct combinatorial patterns of cytokine expression, showing the remarkable

¹ The abbreviation "CyTOF", in addition to being the name of this technique, is also the name of a commercial product that enables researchers to use the method. The authors are in no way endorsing any specific commercial products.

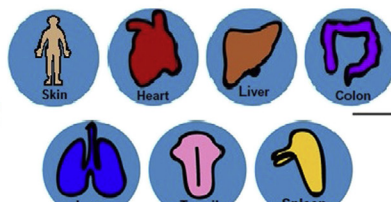
Sample Preparation

Blood sample

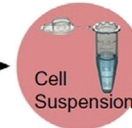


OR

Tissue sample



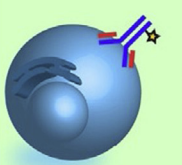
Cell isolation
(tissue digestion,
dead cell removal
& cell sorting, etc.)



Live/dead staining (Rh-103, Pt-195, Pt-194 or Pt-198)

Cell staining

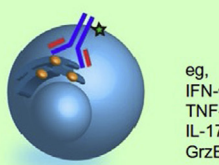
Surface staining
(differentiation & trafficking markers)



eg,
CD45RA
CCR7
CD127
CXCR5

Newell et al. *Immunity*, 2012.

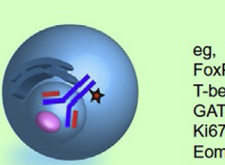
Intracellular staining
(cytokines & chemokines)



eg,
IFN- γ
TNF- α
IL-17
GrzB

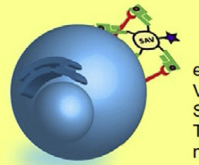
Newell et al. *Immunity*, 2012.

Intranucleus staining
(transcriptional factors)



eg,
FoxP3
T-bet
GATA3
Ki67
Eomes

Tetramer staining
(T cell specificity)



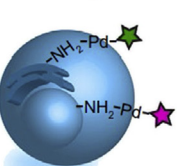
eg,
Viral-epitopes
Self-epitopes
Tumor-related
neo-antigen

Newell et al. *Nature Biotech*, 2014.

Binary/triply barcoding

Fixation before
barcoding

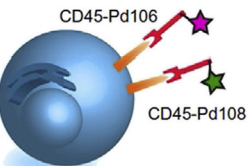
Mass-tag cell barcoding



Zunder et al. *Nature Protocols*, 2015.

Fixation after
barcoding

CD45-Pd cell barcoding



Mei et al. *J Immunol*, 2015.

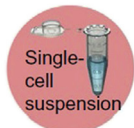
OR



DNA staining

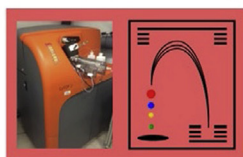
(Ir-191/193 or Rh-103 interchelator)

Cell acquisition



Adding
Four EQ
Beads

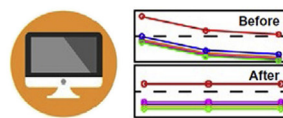
Cytometry by time-of-flight



Acquisition time: ~ hours to days

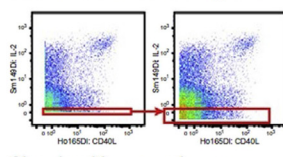
Data processing

Normalization



Finck et al. *Cytometry A*, 2013.

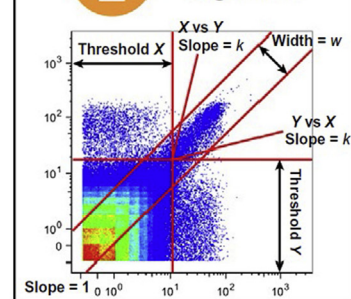
Randomization of signals



Signals with zero value are
rearranged to -1 ~ 0

After
cell acquisition &
data processing

Deconvolution
algorithm



Data analysis



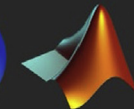
Debarcode samples

Softwares or web services

platforms for graphical plot of high-dimensional analysis



R



MATLAB

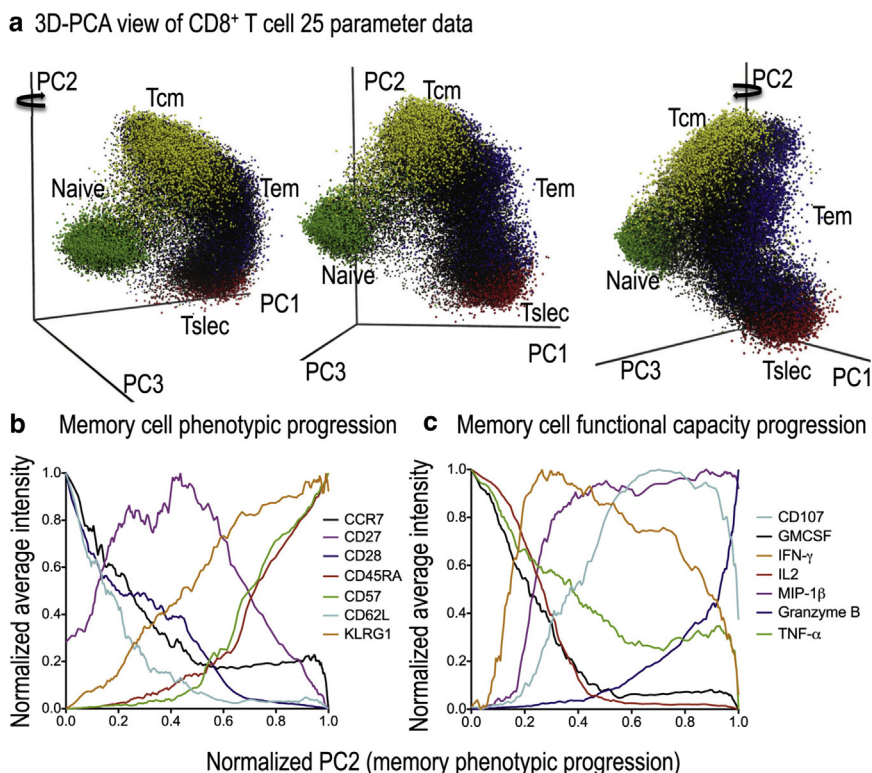


Cytobank

Figure 1. Mass cytometry experiment workflow. The experimental procedure can be separated into multiple steps, including sample preparation, cell staining, cell barcoding, acquisition of CyTOF data, data processing (randomization, normalization, and cell de-barcoding), and high-dimensional data analysis. Briefly,

Figure 2. 3D-principal component analysis (PCA) elucidates the phenotypic and functional complexity of CD8⁺ T-cell memory differentiation.

In this example, PCA was used to represent simultaneously 25 markers from a single cell sample, hence quantifying the expression of functional markers among several memory CD8⁺ T-cell subsets. (a) The cytometry dataset is plotted on the first three principal component axes and shown from three different perspectives (rotated around the PC2-axis). Cells are gated according to their surface memory markers, naive (green), central memory (Tcm, yellow), effector memory (Tem, blue), and short-lived effector (Tslec, red), which show main phenotypic clusters. (b and c) To analyze only memory cells, cells were gated to exclude the naive compartment (cells with low value for PC1). The average expressions for each (b) phenotypic and (c) functional parameter were normalized and plotted as a function of normalized PC2 values. In this way, the phenotypic progression of CD8⁺ memory T cells are represented by the x-axis (0 = early differentiated Tcm, 1 = mature Tslec), and the y-axis represents the average expression of each marker. The functional progression of these numerous markers during CD8⁺ T-cell memory differentiation would not be possible by conventional manual gating. Reprinted from Newell et al., 2012, with permission from Elsevier.



flexibility of CD8⁺ T cells in responding to pathogens. A similar approach could be used to study the diversity and complexity of skin-specific T cells or innate lymphoid cells, complementing the recent in situ topographic characterization of innate lymphoid cells in human skin (Brüggen et al., 2016).

SPADE

In contrast to PCA, which draws out the underlying variance within a dataset, the goal of clustering algorithms is to visualize common patterns within datasets. In the context of immune-phenotyping, automatic clustering algorithms aim to define the most prevalent cell populations by clustering cells based on markers expression similarity (Levine et al., 2015).

SPADE was the first algorithm specifically developed for mass cytometry data analysis that includes both clustering and dimension reduction. In SPADE, cells are clustered into a hierarchical tree shape for two-dimensional visualization (Bendall et al., 2011; Qiu et al., 2011). Each cluster of a SPADE tree is graphically represented by a circular node in which the node size symbolizes the frequency of data points within that population (number of hits), and the node color shows the signal intensity of a selected marker (intensity of hit). Each node is connected to related nodes by a minimum-spanning tree algorithm. Minimum-spanning tree is a classic computer algorithm that searches for a tree-like graph through a set of spatial nodes by using the least possible number of connections. Such a tree might mimic a relational map of immune cell types in their cell development. This method

cells are isolated from blood or solid tissue samples. After the optional enrichment step, the cell suspension is stained with cisplatin or rhodium to distinguish dead cells. Principally, cells were probed with surface, intracellular, or intranuclear markers after tetramer staining (if applicable). Cells can be further barcoded by mass-tag cell barcoding (after cell fixation) or CD45-Pd cell barcoding (before cell fixation) systems. Fixed cells are stained with iridium or rhodium DNA-interchelator and resuspended in deionized water for subsequent acquisition on CyTOF. Collected data are converted into an FCS file, and metal signals are normalized. De-barcoded samples are loaded onto a bioinformatic platform (usually in a MATLAB or R environment or online server such as CytoBank.org) of choice for high-dimensional cytometric analysis. Adapted from Cheng and Newell, 2016, with permission from Elsevier. CyTOF, time-of-flight mass cytometry; FCS, Flow Cytometry Standard.

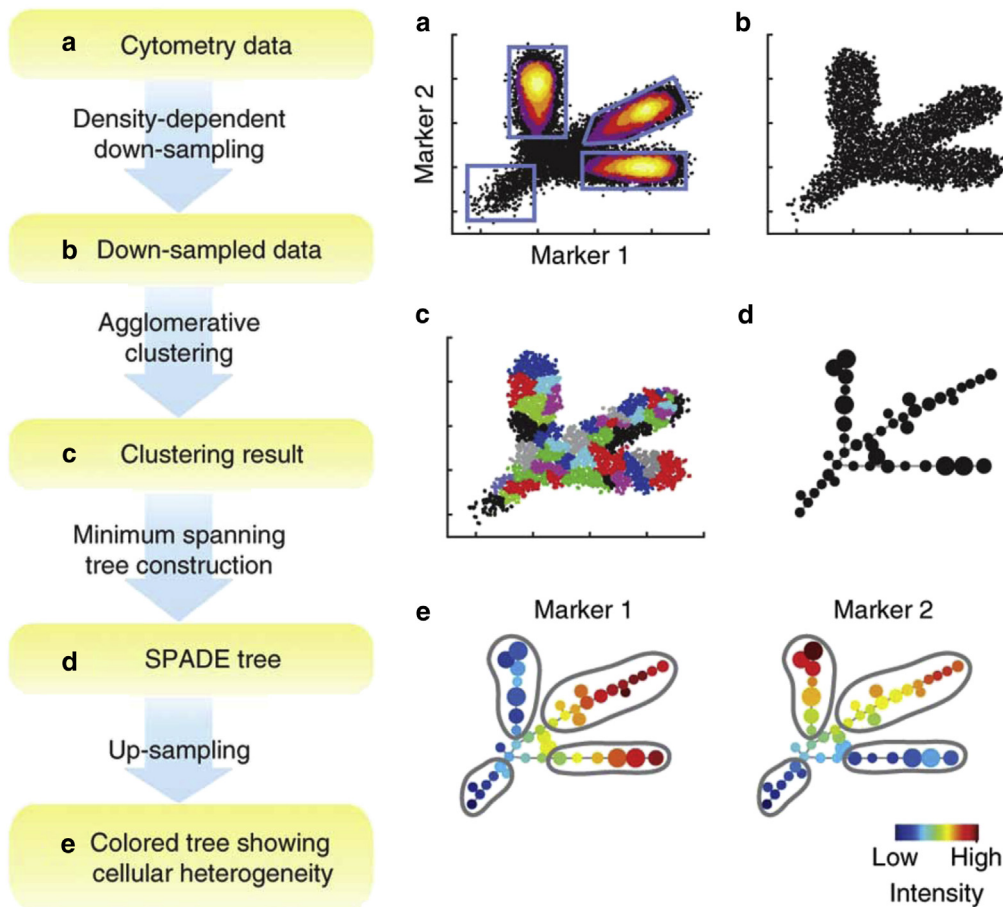


Figure 3. Flowchart of a SPADE tree construction and result visualization. This example illustrates how a SPADE tree is generated from raw cytometry data and how it displays the dataset results. (a) The cytometry dataset analysis by two parameters detects one rare population and three abundant populations. (b) The original data is subjected to density-density down-sampling. (c) Agglomerative clustering results of the down-sampled cells. (d) Minimum spanning-tree algorithm connects the cell clusters. (e) Colored SPADE trees. Nodes are colored by the median intensities of markers, allowing visualization of the markers' expression across the numerous heterogeneous cell populations. The final SPADE tree representation enables determination of how many different subsets are present within a dataset, the relative population density (size), the expression of various markers (color) within each subset, and the relationship among subsets (links). Reprinted by permission from Macmillan Publishers Ltd: *Nature Biotechnology* (Qiu et al., 2011), copyright 2011. SPADE, spanning-tree progression analysis of density-normalized events.

allows comparison between expression of markers among clusters and across distinct samples in a dimension-reduce, big-picture view of diverse cell populations (Figure 3). A limitation of SPADE is that it cannot reproduce the same representation of results when the same dataset is analyzed more than once, because the algorithm involves several stochastic steps. The rigid connections established within the data representation structure (graph) may mislead the positioning of some cell nodes, possibly obscuring the underlying biology. Lee et al. (2015) recently used CyTOF and SPADE to define in detail the phenotype and functional characteristics of distinct subsets of nasal dendritic cells in mice. SPADE clearly portrayed a map of the vast heterogeneity of nasal dendritic cells and identified new subsets that were not perceptible by manual gating on the basis of canonical marker expression. Moreover, it enabled simultaneous comparison of all subsets before and after stimulating factor (FMS)-related tyrosine 3 kinase ligand treatment, showing which subsets became functionally active and/or expanded in number of cells.

viSNE and ACCENSE

t-SNE allows visualization of high-dimensional data through a nonlinear reduction where each data point is given a location in a two- or three-dimensional map (Van

der Maaten et al., 2008). Newer t-SNE-based strategies have been developed to visualize complex mass cytometry data: viSNE and ACCENSE. In contrast to PCA, these algorithms effectively capture nonlinear relationships among the data, and unlike SPADE, they do not cluster cells into exclusive nodes. Each single-cell data point has a unique location in a two-dimensional representation, similar to a biaxial scatterplot, reflecting their proximity in high-dimensional space. viSNE is a graphical user-interfaced tool based on t-SNE, whereas ACCENSE has extended t-SNE with clustering of the resulting two-dimensional scatter data into density-based clusters. Thus, close proximity between any two cells is based on their immunophenotypic similarity, predefined by input markers. For example, to generate a map of the various memory T-cell subpopulation markers such as CD31, CD45RA, CD45RO, CCR7, or CD62-L can be selected. The algorithm then clusters cells according to their similarity within these markers.

A color scale can then be used to visualize each marker's relative expression in the population. However, this type of representation may obscure subtle density differences among cell populations. Another drawback of these algorithms is the inability to currently analyze large numbers of different samples simultaneously. For example, the current version of

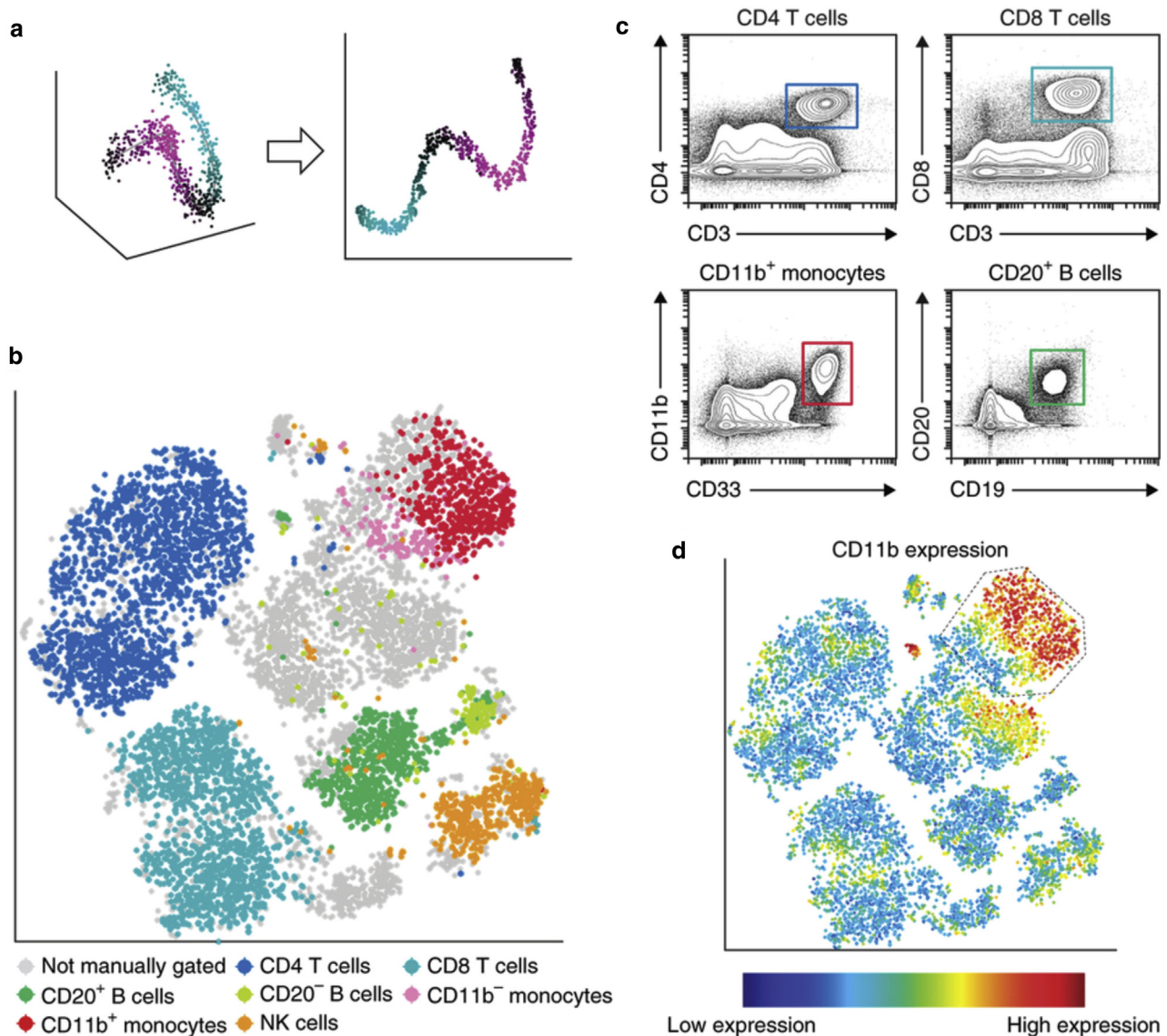


Figure 4. viSNE creates a map of the immune system. (a) In this example, viSNE projects a one-dimensional curve embedded in three dimensions (left) to two dimensions (right). The color gradient shows that nearby points in three dimensions remain close in two dimensions. (b) Application of viSNE to a healthy human bone marrow sample. viSNE automatically separates cells based on their subtype. Each point in the viSNE map represents an individual cell, and its color represents its immune subtype based on independent manual gating. The axes are in arbitrary units. (c) Biaxial plots representing the same data shown in b. Select subpopulations are shown with canonical markers where the square color matches the subtype in b. The actual gating used is more complex and uses a series of biaxial plots for each population. Note, unlike b, these plots do not separate between all subtypes in a single viewpoint. (d) The same viSNE map represented in b, but this time each cell is colored based on CD11b expression. Gated cells are all CD33 high and show a CD11b (maturity) gradient. Many of these cells were not classified as monocytes by manual gating (grey cells b). Reprinted by permission from Macmillan Publishers Ltd: *Nature Biotechnology* (Amir et al., 2013), copyright 2013. viSNE, t-distributed stochastic neighbor embedding—based visualization.

the commercial software Cytobank (Cytobank, Inc., Mountain View, CA) can analyze up to 2 million cells.

Healthy and cancerous bone marrow samples were recently studied using viSNE. Healthy samples were graphically represented and contrasted with cancerous samples that exhibited an abnormal map. Marker expression patterns were analyzed from diagnosis to relapse, allowing identification of disease-specific subsets (Amir et al., 2013) (Figure 4). ACCENSE helped further characterize CD8⁺ T-cell subpopulations and identify new subsets that were not noticeable on biaxial plots (Shekhar et al., 2014) (Figure 5).

CLUSTERING AND CORRELATION WITH CLINICAL OUTCOME

CITRUS

CITRUS aims to not only showcase biological mechanisms and hierarchies and define population subsets, similar to previously described algorithms, but also to identify cellular features that correlate to an experimental endpoint of interest. Bruggner et al. (2014) showed that CITRUS accurately identified subsets of cells associated with AIDS-free survival in HIV-infected patients. CITRUS clusters cells based on similarity of cell markers. Such methods can delineate clusters of cells with specific behavioral

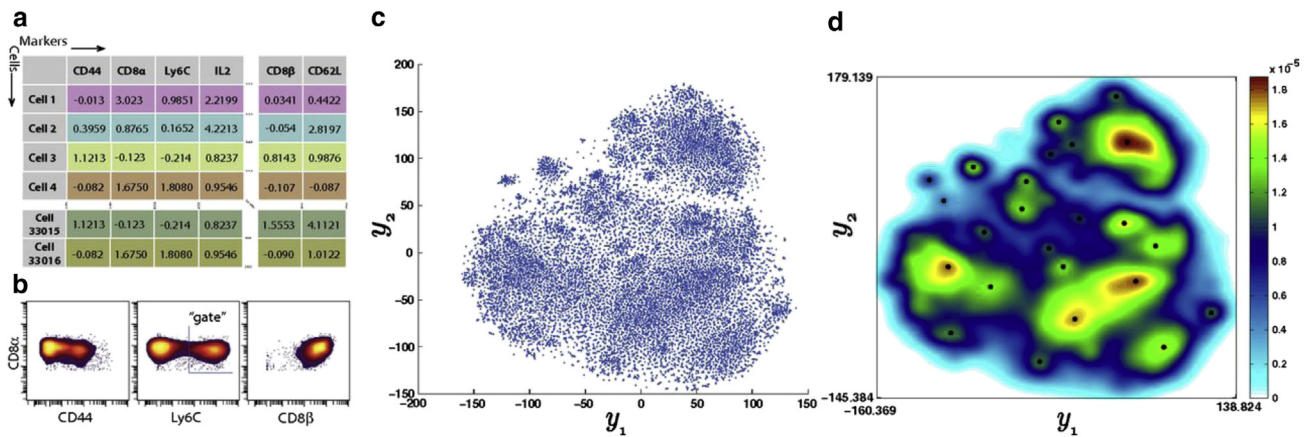


Figure 5. ACCENSE quantifies distinct subsets of CD8⁺ T cells. In this example, ACCENSE was able to stratify cells into phenotypic subsets by the local probability density of cells (ACCENSE map location of hits) based on the 35 markers studied. It identified a novel subset with a unique multivariate phenotype that is not distinguishable on a biaxial plot of markers. **(a)** Illustration of a sample mass cytometry dataset. Rows correspond to different cells, and columns correspond to the different markers. Entries correspond to transformed values of mass-charge ratios that indicate expression levels of each marker. **(b)** Biaxial plots exemplify the manual gating approach to identify cell subsets. **(c)** The two-dimensional t-SNE map of CD8⁺ T cells, where each point represents a cell ($n = 18,304$) derived by down-sampling the original dataset. **(d)** A composite map depicting the local probability density of cells as embedded in panel c, computed using a kernel-based transform. Local maxima in this two-dimensional density map represent centers of phenotypic subpopulations and were identified using a standard peak-detection algorithm. viSNE also generates a similar two-dimensional t-SNE map; however, each subpopulation has to be identified and demarcated manually, whereas ACCENSE automatically defines subsets based on local cell density in the map. Reprinted from [Shekhar et al., 2014](#), in accordance with permission rights to reprint material published in PNAS. ACCENSE, automatic classification of cellular expression by nonlinear stochastic embedding. t-SNE, t-distributed stochastic neighbor embedding; viSNE, t-distributed stochastic neighbor embedding-based visualization.

characteristics such as those that have activated specific signaling pathways, which can potentially be associated with clinical outcomes such as time to recovery or overall survival after a drug treatment or surgery. By specifying the outcome of interest for each sample from the vast cytometry data uploaded, regularized statistic algorithms can recognize cells exhibiting behavior predictive of outcome. The phenotype and behaviors of each cluster are delineated through distinct data representations including conventional biaxial plots that can result in a predictive model for the analyses or validation of future samples. To run correlation analysis with enough statistical power, CITRUS requires input from more than eight samples for each group. [Gaudillière et al. \(2014\)](#) applied this method to monitor patients undergoing hip replacement surgery to characterize the phenotypical and functional immune responses predictive of recovery from surgical trauma ([Figure 6](#)). Because the patients' clinical histories of post-surgical recovery were known, CITRUS identified intracellular signaling markers that strongly correlated with recovery from fatigue, functional hip impairment, and pain after surgery. Likewise, CITRUS could be used to analyze immune cell populations and molecular markers from patients with inflammatory or autoimmune diseases before and after receiving biologic treatments. It could show which cell subset or markers are predictive of therapeutic outcome, allowing early adjustments to treatment protocols in patients predicted to be nonresponders, reducing adverse effects and costs.

AVAILABILITY OF SOFTWARE

After mass cytometry data acquisition, each respective sample must be deconvoluted if samples were initially

barcoded and normalized based on calibration beads. CyTOF data can be exported in the form of Flow Cytometry Standard (FCS) files, which can be analyzed by standard flow cytometry software such as FCS Express (De Novo Software, Los Angeles, CA) and FlowJo (FlowJo, LLC, Ashland, OR), or by using cloud-based analysis tools, such as Cytobank. Cytobank allows transfer and storage of multiple CyTOF FCS files; attachment of related data including protocols, presentations, annotations, and images; and sharing of data and analysis with chosen collaborators. Additional analytical tools may be needed, including but not limited to dose response, heat maps, SPADE, viSNE, CITRUS, and dot and histogram overlays. viSNE can be publicly licensed to academic users as a MATLAB-based tool from the Dana Pe'er Lab of Computational Systems of Biology webpage (<http://www.c2b2.columbia.edu/danapeerlab/html/software.html>). The Nolan Laboratory also publically offers some of these algorithms, such as SPADE and CITRUS (<https://github.com/nolanlab>). ACCENSE is freely offered at <http://www.cellaccense.com>, and a PCA algorithm is included in the open-sourced R basic package (The R Foundation, Vienna, Austria). These analysis tools may provide different types of biological information about the same dataset, making it acceptable to concurrently use distinct tools in a single experiment ([Table 1](#)). Nevertheless, the accessibility to CyTOF analysis tools continues to improve, with numerous groups developing similar and new analysis methods for CyTOF data.

CONCLUSION

The continuous development and enhancement of analysis tools for mass cytometry expands our ability to study

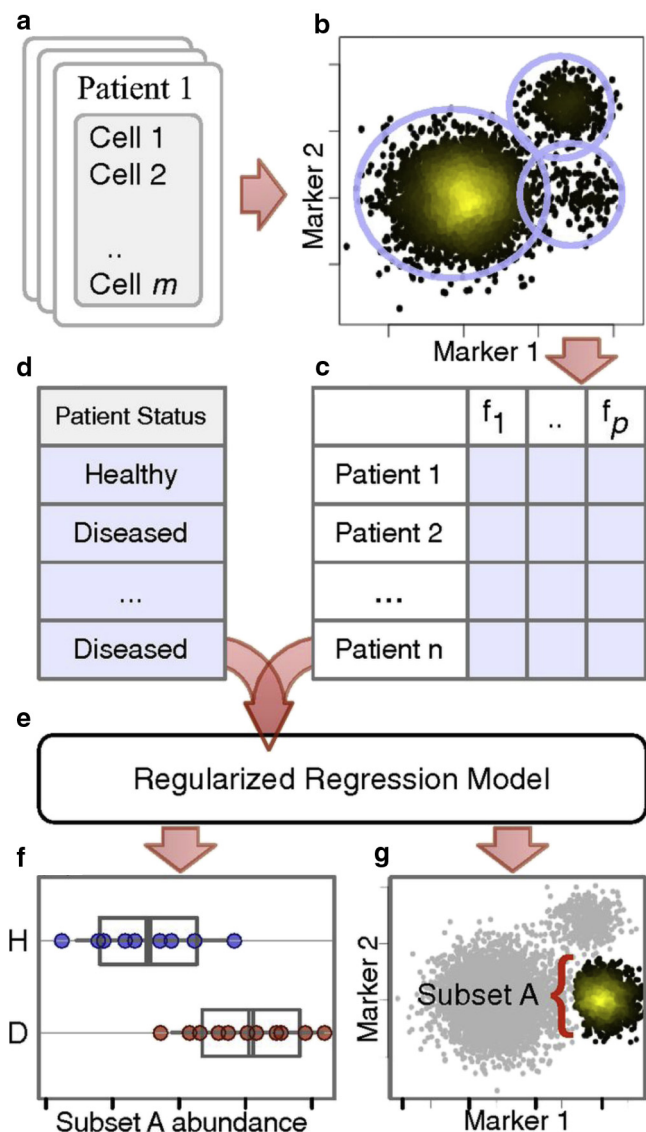


Figure 6. Overview of CITRUS. CITRUS enables correlation of the multitude of cellular parameters studied with clinical outcomes information. Cells from (a) all samples are combined (b) and clustered using hierarchical clustering. (c) Descriptive features of identified cell subsets are calculated on a per-sample basis and (d) used in conjunction with additional experimental metadata to train (e) a regularized regression model predictive of the experimental endpoint. (f) Predictive subset features are plotted as a function of experimental endpoint, (g) along with scatter or density plots of the corresponding informative subset. In this example, the abundance of cells in subset A was found to differ between healthy and diseased samples (f; subset A abundance in healthy patients and diseased patients). Scatter plots show that cells in subset A have high expression of marker 1 and low expression of marker 2 relative to all other cells (shown in gray). In this study, CITRUS identified T-cell subsets whose abundance is predictive of AIDS-free survival risk in patients with HIV. Reprinted from Bruggner et al., 2014, in accordance with permission rights to reprint material published in PNAS.

complex and heterogeneous biological systems at the level of individual cells. This enables our understanding of the progression and development of healthy and pathologic cells, such as in psoriasis, atopic dermatitis, and vitiligo. For example, why do some lesions reoccur only in the same anatomical site? Why are some areas of the body

Table 1. Analysis algorithms for mass cytometry data

Algorithm Name	Type of Information	Advantages	Limitations	Reference
PCA	Parameters with most variance within dataset	Visualization in 3D space	<ul style="list-style-type: none">May miss subtle variances within dataAssumes that data are parametricData representation is restricted to linear projections	Jackson, 1991; Newell et al., 2012
SPADE	Cell population hierarchies	<ul style="list-style-type: none">Delineates the presence of rare cell typesCan compare clusters and expression markers among cell subsets and across samples	<ul style="list-style-type: none">Lacks reproducibilityRigid structure connectivity	Qiu et al., 2011
t-SNE—based viSNE and ACCENSE	Cell subset heterogeneity	Single-cell representation (without clustering)	Unfeasible to analyze vast numbers of cells ¹	Amir et al., 2013; Shekhar et al., 2014
CITRUS	Allows correlation between sample and clinical outcome	Correlation to experimental endpoint of interest	More than eight samples per group are required	Bruggner et al., 2014

Abbreviations: 3D, three-dimensional; ACCENSE, automatic classification of cellular expression by nonlinear stochastic embedding; CITRUS, cluster identification, characterization, and regression; PCA, principal component analysis; SPADE, spanning-tree progression analysis of density-normalized events; t-SNE, t-distributed stochastic neighbor embedding; viSNE, t-distributed stochastic neighbor embedding—based visualization.

¹The cloud-based software Cytobank viSNE can currently handle up to 2 million cells.

more commonly affected? What distinguishes pathological cells in a stable versus a progressive disease? Disease-specific cell subsets can be identified, characterized, monitored during treatment, and perhaps screened for early biomarkers predictive of relapse risk. Differences may be revealed among cells responsible for the clinical heterogeneity of cutaneous T-cell lymphoma, ultimately unveiling disease-specific biomarkers and personalized novel therapeutic approaches. Innovative therapies can be studied to specifically target malignant cell populations resistant to conventional treatments, such as in melanoma. Ultimately, the goal of this article is to demystify the developing tools for mass cytometry and its data analysis so these technologies can be adopted and the results understood to address these and other important research questions in the coming years.

CONFLICT OF INTEREST

The authors state no conflict of interest.

MULTIPLE CHOICE QUESTIONS

- What is the best CyTOF data analysis tool?
 - Same methods as for flow cytometry
 - The analysis method depends on specific experimental goals.
 - Manual clustering methods through biaxial plots
 - Comparisons of marker expression using histograms
- Identify one advantage of principal component analysis (PCA).
 - Displays data in two-dimensional representation
 - Results are represented through linear projections
 - Identifies parameters with the most variance
 - Capable of analyzing only a few parameters
- Which of the following is a limitation of spanning-tree progression analysis of density-normalized events (SPADE)?
 - Incapable of reproducing the same representation of results when analyzed more than once
 - Represents cell subset hierarchies
 - Assumes that data is parametric
 - Does not allow comparing marker expression among subsets and samples
- Select one advantage of t-distributed stochastic neighbor embedding (t-SNE)-based visualization (viSNE) and automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE).
 - It captures only nonlinear relationships among the dataset.
 - It clusters cells into exclusive nodes.
 - It allows representation of single cells without clustering
 - It allows the researcher to identify clusters capable of predicting the sample's outcome.
- What is the novel application of analyzing data using cluster identification, characterization, and regression (CITRUS)?
 - It is capable of displaying data in three-dimensional representations.
 - It identifies rare cell subsets with the highest expression of studied markers.
 - It allows the researcher to define cell population hierarchies.
 - It identifies cellular features that correlate to an experimental endpoint of interest.

ACKNOWLEDGMENTS

We would like to thank Dr. Jodi L. Johnson for helpful comments, critical reading of the manuscript, and editorial assistance.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to this paper. Teaching slides are available as supplementary material.

REFERENCES

- Amir E-AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* 2013;31:545–52.
- Bendall SC, Simonds EF, Qiu P, Amir ED, Krutzik PO, Finck R, et al. Single cell mass cytometry of differential immune and drug responses across the human hematopoietic continuum. *Science* 2011;332(6030):687–96.
- Brüggen MC, Bauer WM, Reininger B, Clim E, Captarencu C, Steiner GE, et al. In situ mapping of innate lymphoid cells in human skin: evidence for remarkable differences between normal and inflamed skin. *J Invest Dermatol* 2016;136:2396–405.
- Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. *Proc Natl Acad Sci USA* 2014;111:E2770–7.
- Cheng Y, Newell EW. Deep profiling human T cell heterogeneity by mass cytometry. *Adv Immunol* 2016;131:101–34.
- Doan H, Chinn GM, Jahan-Tigh RR. Flow cytometry II: mass and imaging cytometry. *J Invest Dermatol* 2015;135:e36.
- Gaudillière B, Fragiadakis GK, Bruggner RV, Nicolau M, Finck R, Tingle M, et al. Clinical recovery from surgery correlates with single-cell immune signatures. *Sci Transl Med* 2014;6:255ra131.
- Jackson JE. PCA with more than two variables. In: Jackson JE, editor. *User's guide to principal component analysis*. New York: John Wiley and Sons; 1991. p. 26–62.
- Lee H, Ruane D, Law K, Ho Y, Garg A, Rahman A, et al. Phenotype and function of nasal dendritic cells. *Mucosal Immunol* 2015;8:1083–98.
- Levine JH, Simonds EF, Bendall SC, Davis KL, Amir el-AD, Tadmor MD, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 2015;162:184–97.
- Matos TR, Liu H, Ritz J. Experimental methodology for single-cell mass cytometry. *J Invest Dermatol* 2017;137:e31–8.
- Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM. Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes. *Immunity* 2012;36:142–52.
- Qiu P, Simonds EF, Bendall SC, Gibbs KD Jr, Bruggner RV, Linderman MD, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol* 2011;29:886–91.
- Shekhar K, Brodin P, Davis MM, Chakraborty AK. Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). *Proc Natl Acad Sci USA* 2014;111:202–7.
- Van der Maaten L, Hinton J. Visualizing data using t-SNE. *J Mach Learn Res* 2008;9(85):2579–605.