COMPUTATIONAL CYTOMETRY

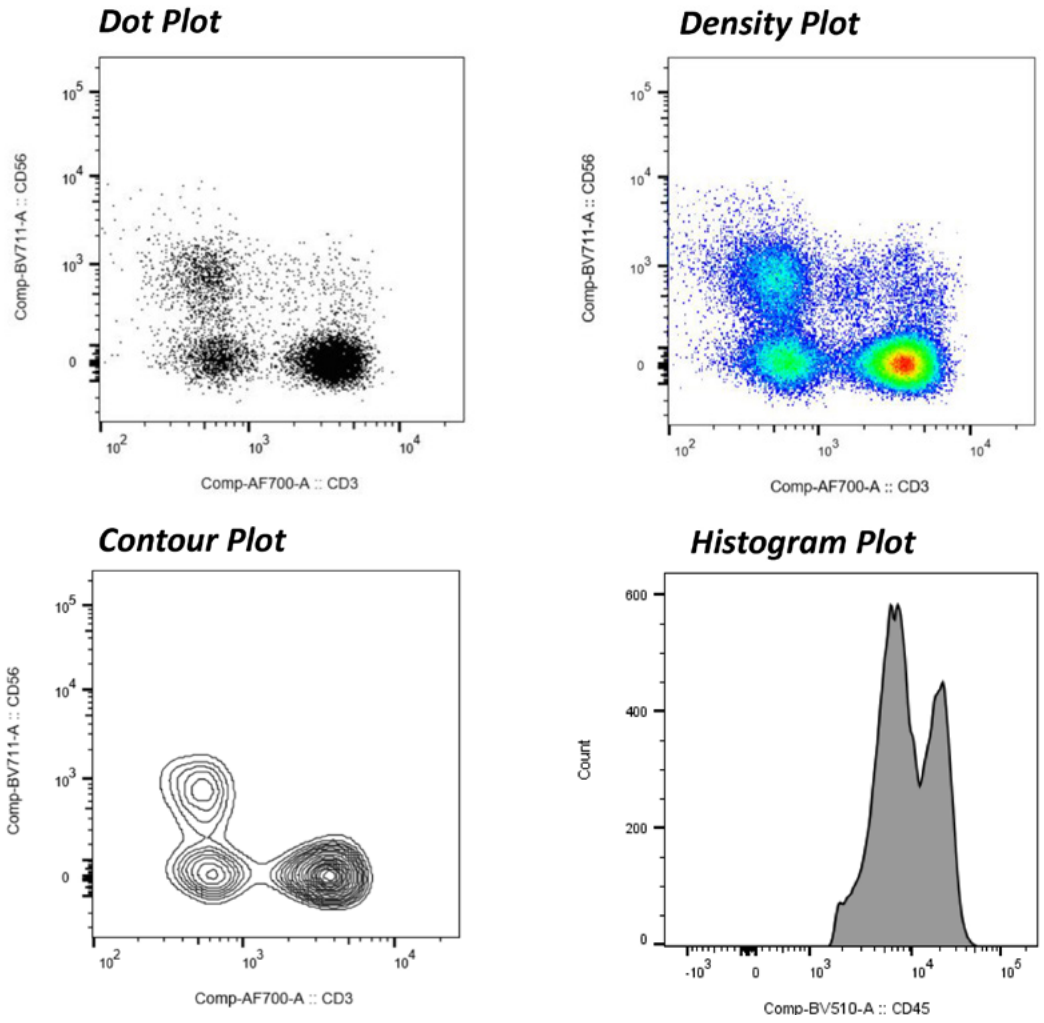
Flow Cytometry Data Analysis in the Era of Quantitative Data Science

Brought to you by: FlowMetric™

Computational Cytometry - Flow Cytometry Data Analysis in the Era of Quantitative Data Science

Flow cytometry is a widely used technique in biomedical research because it can characterize and measure many different cell types and is customizable, precise and quantitative. Flow cytometry has transformed the field of immunology and been instrumental to the progress of immunotherapeutics research and development.

Now we find ourselves in an era in which flow cytometers can routinely measure 12 different cell parameters and beyond. Mass cytometry is a type of flow cytometry that uses a time-of-flight mass spectrometry readout, and this technique can measure upwards of 100 parameters. These advances in hardware and reagents have been coupled with a transformation in how flow cytometry data is analyzed, and now most flow cytometry users are learning how to use computational analysis to understand and visualize their large datasets .



1. Saeys Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology.* 2016 Jul;16(7):449.

Limitations of Manual Data Analysis

Experienced users may have begun using flow cytometry when most protocols looked at four to eight parameters. These protocols worked well for early immunophenotyping and intracellular cytokine staining applications, and these data were typically analyzed by making manual sequential gates on a series of two-dimensional plots to measure your cell populations of interest. Use of this sequential gating strategy can work well with robust staining and clearly defined cell populations. More often, manual gating is highly user dependent and subjective, and this subjectivity can imperil the quality and reproducibility of a study's findings. Manual analysis is also extremely cumbersome and inefficient for analysis of large datasets.

The flow cytometry community has moved toward using computational flow cytometry tools to analyze complex data and move the field toward improved standardization and reproducibility .

Machine Learning - Moving Flow Cytometry Data Visualization into the Modern Era

Computational biologists have come to realize that machine learning techniques are well suited to flow cytometry data visualization as the main objective of machine learning is to build models that make reasonable generalizations from input data. Machine learning algorithms can take high-dimensional data like those seen in large flow cytometry datasets and cluster data points together based on similarities in variables or use dimensionality reduction to create two-dimensional visualizations. These approaches can be combined such that data is first put through a dimensionality reduction algorithm before being used in a clustering algorithm, and numerous open source and commercial software packages are available to process and analyze raw data.

---

2. Finak G, Langweiler M, Jaimes M, et al. Standardizing Flow Cytometry Immunophenotyping Analysis from the Human Immuno-Phenotyping Consortium. *Science Reports*. 2016;6:20686.

Before Visualization - Experimental Design and Data Processing

A good research project begins with identifying an interesting question and designing a robust experiment that includes a reasonable sample size and a clear data analysis plan. A study that is too small will likely be underpowered for statistical analysis, and lack of planning for data analysis may result in measuring the wrong parameters, thus rendering your experiment meaningless. For flow cytometry, experimental design planning also means making sure the appropriate quality control measures, which include staining controls for compensation and normalization measurements, are in place.

During a flow cytometry run, the cytometer will need to be checked to make sure it is calibrated correctly and is working within its specifications. Cytometer users typically use some sort of general gating strategy during this acquisition phase to ensure they are collecting data on their cells of interest. Data acquisition is also a critical time for making sure that the data files are "evergreen" and will be compatible with data analysis in the near future or much further down the road. Evergreen data is a term that describes data that is collected for different cell markers that can be analyzed in different combinations by multiple methods. High dimensional flow cytometry is a powerful tool for collecting evergreen data as new cell subsets are always being defined, and this allows scientists to go back to their data and identify previously unknown cells.

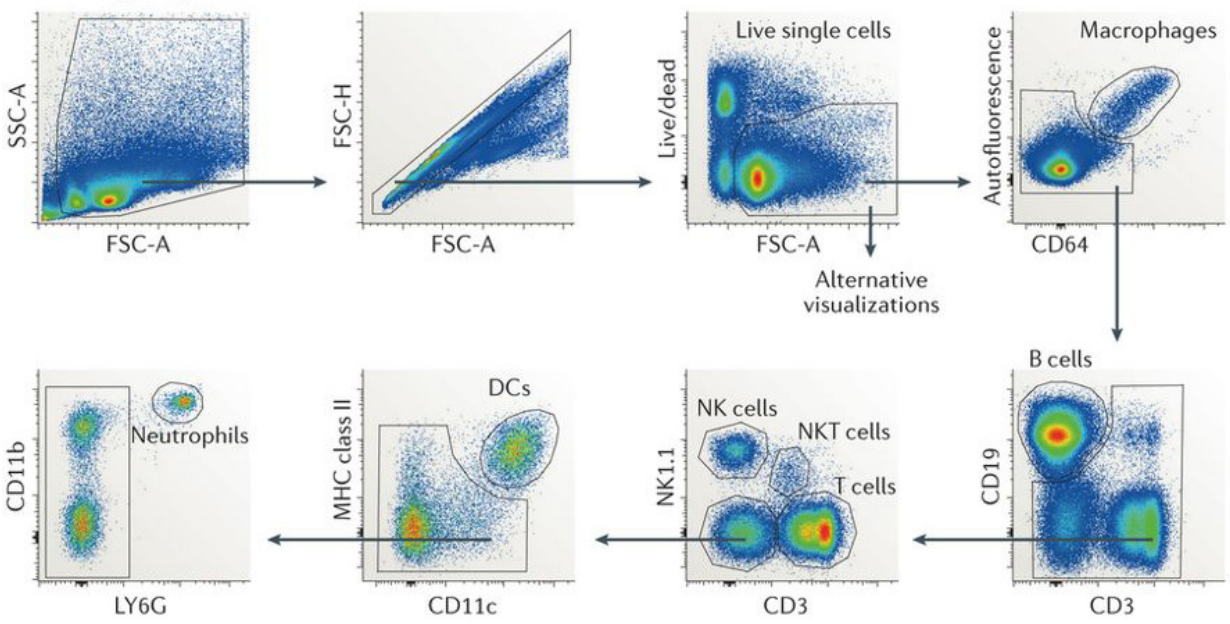
Data Analysis and Visualization

Dimensional Reduction at Work - Principle Component Analysis and t-SNE

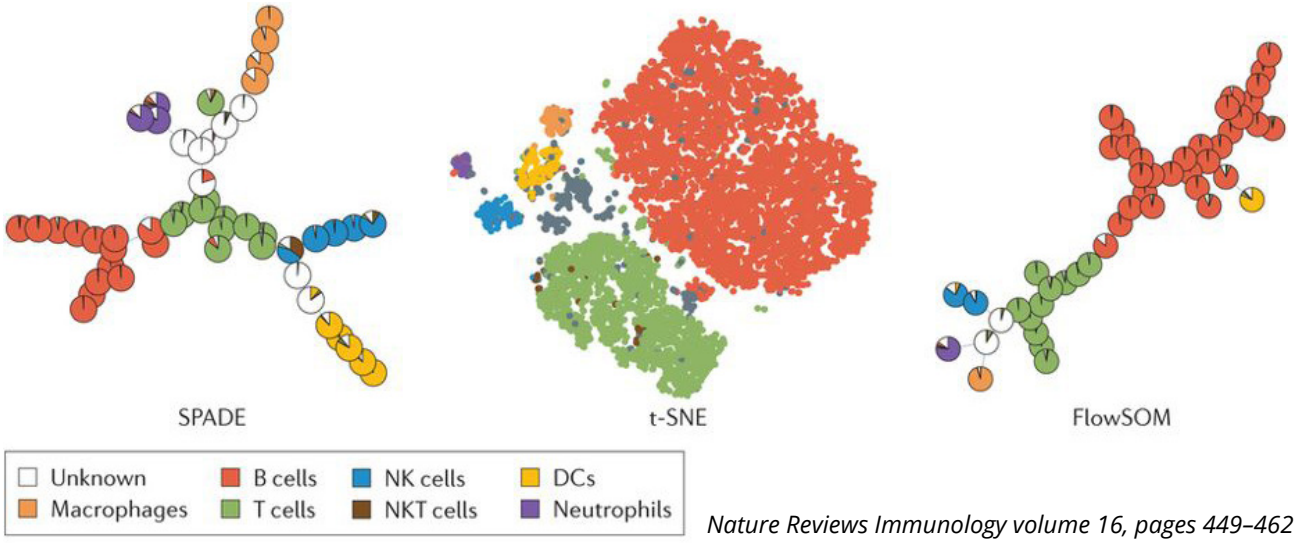
In flow cytometry, multiple parameters are measured for each cell, sample or patient, but representing this data in a multi-dimensional plot is not possible. Hence, dimensionality reduction has been used to simplify data visualization and observe trends. Principal component analysis (PCA) was the first dimensionality reduction method applied to flow cytometry data that is an unsupervised dimension-reduction method that generates new variables and plots them as combinations of original variables while maintaining variance from original data . T-stochastic neighbor embedding (t-SNE) is a more widely used dimensionality reduction method that makes a lower-dimensional plot, typically a single bivariate plot, while maintaining structure from high dimensional data . Both of these methods can plot an individual cell as a data point and provide valuable single cell data resolution. These methods can also be computationally intensive so may require that users decrease sample sizes, or only use a subset of data to run analysis within a reasonable time frame.

This figure demonstrates the utility of using computational analysis methods to understand and visualize the complexity of high dimensional flow data

Panel A: Shows the visualization of high dimensional flow using traditional flow data analysis methods. Once the live cell population has been identified it takes an additional 5 plots to visualize and decipher the cellular sub-populations of interest.



Panel B: Shows the power and utility of applying computational data analysis methods to the same high dimensional flow samples. Using the same gating strategy from Panel A (where we identify our live cell population) we can evaluate the same data using 3 different alternative algorithm-based techniques that reduces the data set from 5 plots down to one, thus showing that conventional approaches to data analysis does not scale well with an increasing number of correlated measurements.



Nature Reviews Immunology volume 16, pages 449–462 (2016)

3. Lugli E, Pinti M, Nasi M, Troiano L, Ferraresi R, Mussi C, Salvioli G, Patsekin V, Robinson JP, Durante C, et al. Subject classification obtained by cluster analysis and principal component analysis applied to flow cytometric data. *Cytometry A*. 2007;71A(5):334–44.

4. Amir el-AD, Davis KL, Tadmor MD, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol*. 2013;31(6):545-52.

Cluster Analysis - Automated Gating and Beyond

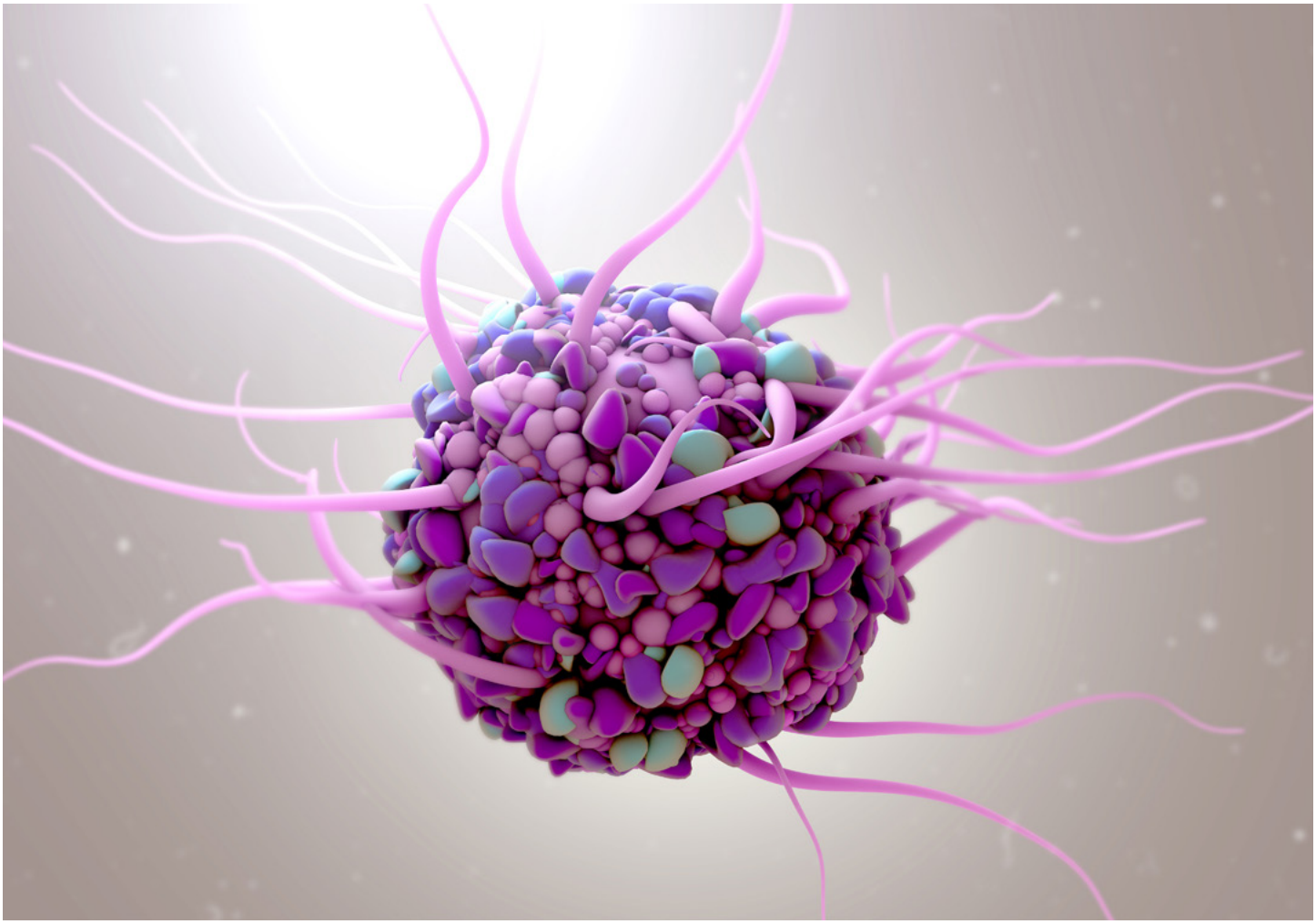
Cluster analysis is another unsupervised method for analyzing flow cytometry data that groups objects with similar profiles, such as cells with similar staining patterns . One important application of cluster analysis is as a method of automated gating . Several different automated gating techniques exist and can be applied to high-dimensional data or data that has undergone dimensionality reduction. The best cluster analysis option depends on the number of cell subsets being studied, how much is known about each subset, and whether or not you know the expected number of clusters. Cluster analysis is also a useful tool for identifying new cell types and following changes in cell subsets over time, and as such, can be valuable for defining novel biomarkers and monitoring the progress of clinical trials.

Stochastic Nature of Visualization Methods

Most of these machine learning-based methods used in flow cytometry are stochastic in nature, which means they predict outcomes and include a certain degree of randomness. As such, different results may emerge from each run, so it is critical for data to be run through an algorithm multiple times in order to be confident in the trends that are being observed.

5. Qiu,P. etal. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat.Biotechnol*. 2011;29, 886–891.

6. Lo,K., Brinkman,R.R. & Gottardo,R. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A* 2008;73, 321–332.



Let Data Reveal the Biology

These new computational flow cytometry analysis methods seem daunting to many flow cytometry users who are now swimming in terabytes of flow cytometry data, and data analysis has become the primary bottleneck in completing these projects. Some understanding of computer coding and statistics is needed to be able to design appropriate experiments and select the correct analysis methods. As a result, computational biologists and immunologists find themselves collaborating more and more. Also, laboratories and commercial enterprises with expertise in computational flow cytometry are finding themselves in high demand.

Many scientists have also come to appreciate that these new methods have given them a completely different perspective in looking at their data, that is, they let the data reveal the biology. Now, new cell subsets can be defined and understood in terms of cell development and differentiation. Insights into immunotherapeutic mechanisms can also be gained through longitudinal clinical studies, and decades-old samples can be thawed and re-analyzed with these contemporary methods to understand both basic biology and pathological processes. Most importantly, these methods are shifting the mindset of the flow cytometry user community toward better practices in standardization, harmonization, and reproducibility.

FlowMetric™

Quantifying Biological Response Through Cytometry