High Dimensional Cytometry Data Visualization Using Parallel Coordinates

Jeffrey M. Hokanson[1,3], Chris Wakefield[2], John Weinstein[2], Michael Andreeff[1], Jared K. Burks[3]

[1]Department of Leukemia; [2]Department of Bioinformatics and Computational Biology; [3]Flow Cytometry & Cell Imaging Core Facilty; MD Anderson Cancer Center

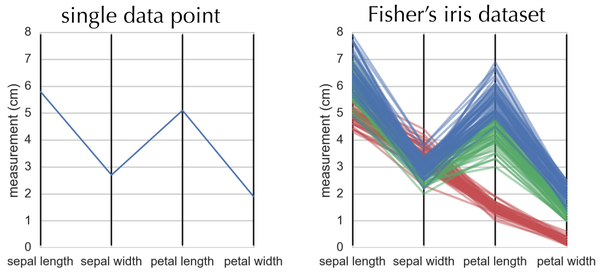
THE UNIVERSITY OF TEXAS MD Anderson Cancer Center

Introduction

Flow cytometry traditionally analyzed data using a hierarchy of manual gates drawn on two dimensional dot plots. With the increasing number of channels available with modern fluorescent and mass cytometers, this manual gating process has become time consuming. Identifying a population now requires viewing dozens of biaxial plots and correctly placing gates in each. Recent work has focused on reducing this burden by either automating the assignment of cells into clusters (i.e., phenograph, SWIFT, FLOCK) or providing new visualizations on which to draw manual gates (i.e., viSNE, SPADE). Both approaches have their drawbacks. Clustering may mislabel cells and data projections hide the original coordinates. Here we present a new a new alternative. Rather than replacing manual gating, we show how manual gating can be made easy for high dimensional data using an interactive parallel coordinates plot. This is particularly well suited for exploratory analysis of mass cytometry data sets.

Parallel Coordinates

Whereas dot plots show each data point as a dot in a two dimensional graph, a parallel coordinate plot shows each data point as a line. The parallel coordinate plot consists of multiple parallel axes, each representing one dimension of the data. Then, for each data point, a line is drawn between adjacent axes connecting the values of that data point.



Although parallel coordinate plots have existed for over a hundred years, this visualization became common with the advent of digital computers, popularized by Inselberg[2] in 1985. With computers, gates can be drawn on each axis to isolate specific clusters that would otherwise be hidden.

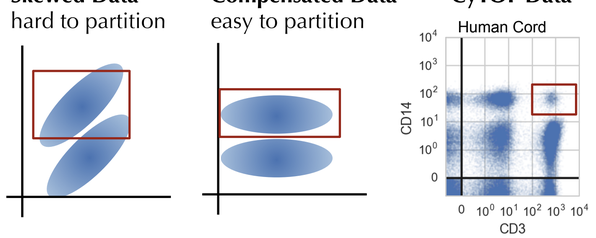
Parallel Coordinates and Flow Cytometry

Using parallel coordinates with conventional flow cytometry is challenging. Gates in a parallel coordinate plot correspond to high dimensional rectangles whose sides are parallel to the coordinate axes. Hence, if data is not properly compensated, it is impossible partition clusters using parallel coordinates. However, new instruments, such as the CyTOF, produce measurements that do not need to be compensated, making parallel coordinates beneficial.

Skewed Data
hard to partition

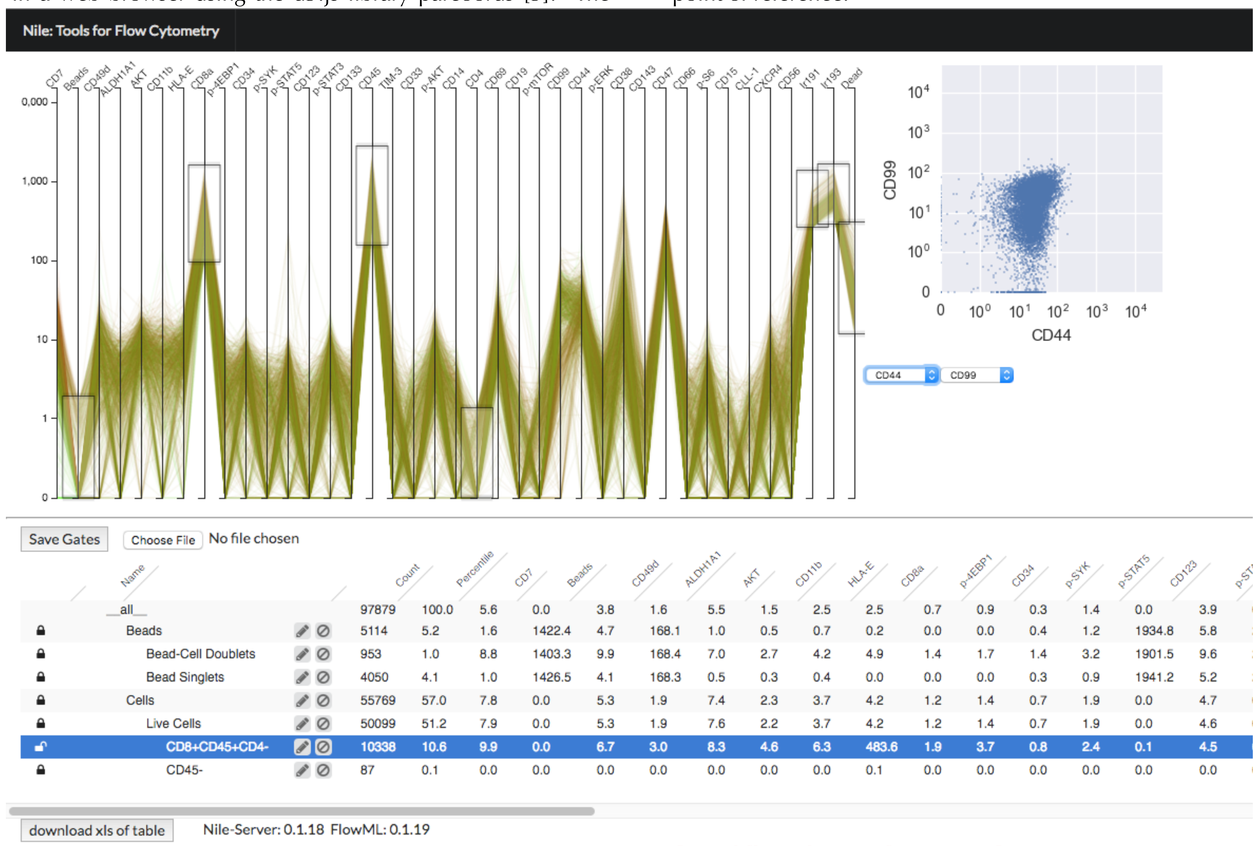
Compensated Data
easy to partition

CyTOF Data



An Interactive Parallel Coordinates Web App for Mass Cytometry

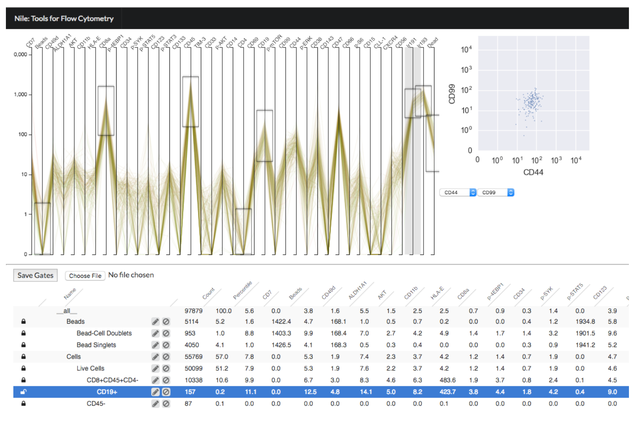
Although there are many implementations of parallel coordinates, none have been tailored to the specific needs of flow cytometry. We developed a web app that incorporates the traditional hierarchy of gates in addition to a parallel coordinates plot. This tool renders the parallel coordinate plot in a web browser using the d3.js library parcoords [5]. The user can draw gates along each axis (shown as rectangles below). The resulting gates are listed in the table below along with the counts, percentages, and the median expression in each marker. This tabular data can then be exported to Excel. A two dimensional dot plot is also show to provide a familiar point of reference.



Name	Count	Percentile	CD7	Beads	CD45rd	ALDH1A1	AKT	CD11b	HLA-E	CD8a	p-4EBP1	CD34	p-SYK	p-STAT5	CD123	p-ST
__all__	97879	100.0	5.6	0.0	3.8	1.6	5.5	1.5	2.5	2.5	0.7	0.9	0.3	1.4	0.0	3.9
Beads	5114	5.2	1.6	1422.4	4.7	168.1	1.0	0.5	0.7	0.2	0.0	0.0	0.4	1.2	1934.8	5.8
Bead-Cell Doublets	953	1.0	8.8	1403.3	9.9	168.4	7.0	2.7	4.2	4.9	1.4	1.7	1.4	3.2	1901.5	9.6
Bead Singlets	4050	4.1	1.0	1426.5	4.1	168.3	0.5	0.3	0.4	0.0	0.0	0.0	0.3	0.9	1941.2	5.2
Cells	55769	57.0	7.8	0.0	5.3	1.9	7.4	2.3	3.7	4.2	1.2	1.4	0.7	1.9	0.0	4.7
Live Cells	50099	51.2	7.9	0.0	5.3	1.9	7.6	2.2	3.7	4.2	1.2	1.4	0.7	1.9	0.0	4.6
CD8+CD45+CD4-	10338	10.6	9.9	0.0	6.7	3.0	8.3	4.6	6.3	483.6	1.9	3.7	0.8	2.4	0.1	4.5
CD45-	87	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0

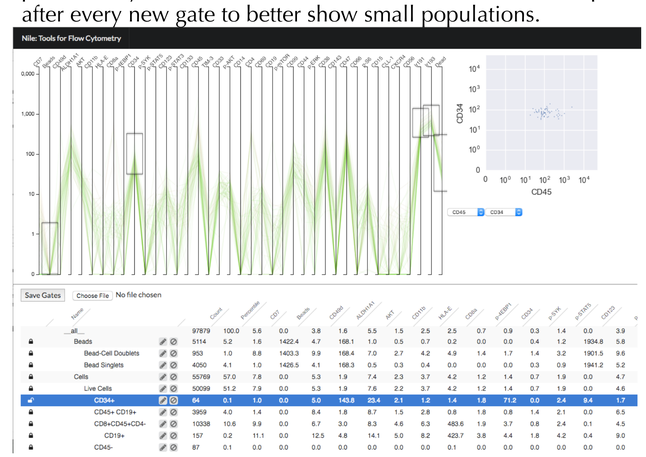
Defining Complex Gates

The example below shows how this parallel coordinate plot has allowed the user to identify a population with a complex phenotype. This plot shows eight gates that are simultaneously applied to the data, with the lines revealing the phenotype.



Identification of Rare Phenotypes

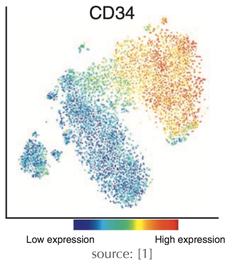
Parallel coordinates can also reveal rare populations. The plot below shows the CD34+ fraction of cells in healthy human peripheral blood as 0.1%. Although performance constrains the plot to only show a few thousand lines, these are resampled after every new gate to better show small populations.



Comparison to Other Exploration Tools

viSNE [1]

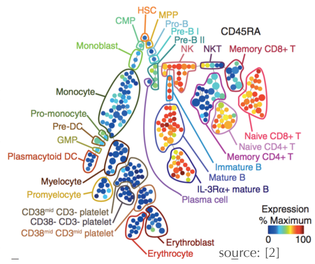
This visualization projects high dimensional data onto a low dimensional space using the nonlinear t-distributed Nearest Neighbor Embedding (t-SNE). Although this approach reveals high dimensional structure, inferring the phenotype of each cell is challenging. A typical approach assigns each point a color based on the expression of a single marker. By rotating through these plots for each marker, the phenotype of each section of the plot is assigned. Like two dimensional scatter plots, this approach still requires viewing multiple plots to reveal phenotype.



source: [1]

SPADE [4]

This approach first constructs a user specified number of cell clusters. From these clusters, it then builds a minimum spanning tree, grouping similar clusters together. Users then combine the clusters into distinct phenotypes. Like viSNE, identifying the phenotype requires rotating through markers.



source: [2]

Clustering

Clustering represents a fundamentally different question than exploratory data analysis. Clustering algorithms, such as SWIFT and phenograph seek to define labels for each data point. Assuming this is done accurately, it is helpful when comparing many samples. For data exploration, this may not be as useful since the clustering algorithm may not be aware of the considerations an expert user would provide when gating.

Conclusion

Interactive parallel coordinate plots are a powerful tool for data analysis as they provide information on every measurement for every data point in a single plot. This enables users to manually define gates in many coordinates simultaneously, speeding the exploration of flow cytometry data.

Try it Yourself !

This project is open source. A docker container image and instructions for installation can be found at:
https://hub.docker.com/r/cwakefield/nile/

References

[1] Amir ED, Davis KL, Tadmor MD, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nature biotechnology. 2013;31(6):545-552. doi:10.1038/nbt.2594.

[2] Bendall SC, Simonds EF, Qiu P, et al. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. Science (2011)332, 687-696. doi: 10.1126/science.1198704

[3] Inselberg, Alfred. The Plane with Parallel Coordinates. Visual Computer. 1985;1(4):69–91. doi:10.1007/BF01898350.

[4] Qiu P, Simonds EF, Bendall SC, et al. Extracting a Cellular Hierarchy from High-dimensional Cytometry Data with SPADE. Nature biotechnology. 2011;29(10):886-891. doi:10.1038/nbt.1991.

[5] https://syntagmatic.github.io/parallel-coordinates/