Unsupervised clustering reveals cell types in 145 specimen melanoma PBMC cytometry dataset

Lee S. Leavitt MS, Gage Black PhD, Justin Jarrell PhD, Kristina Magee MD PhD, Hannah Selken, Li-Chun Cheng PhD, Ramji Srinivasan

Teiko.bio

Background

Gating has been the status quo to analyze cytometry datasets for decades. The challenge with this approach is the inability to scale to multiparameter (20-40+) panels. For instance, a 44-marker panel theoretically yields billions of potential cell types, requiring ~239 hours to fully gate.

Computational methods, including unsupervised analysis, are a credible alternative to make sense of these massive datasets. In the last decade, researchers have made several advances across normalization and visualization to analyze these large datasets. Phenograph (and its successor Fast-Phenograph) and Uniform Manifold Approximation and Projection (UMAP) are two of the critical advances that can generate automatic clusters, and help visualize high-dimensions, respectively.

These tools can find distinctive cell types that would be missed by manual gating. We've pioneered a new way of clustering, that accounts for biological reality, ensuring high-resolution identification of cell types. Our pipeline, applied to 85 melanoma patients uncovered a novel cell subset (CD4 T CD161+ CD39-) missed by conventional methods, which is linked to treatment response.

Methods

We analyzed an 85 patient, 145 PBMC specimen melanoma dataset from Massachusetts General Hospital using a 44-marker mass cytometry panel.

Gating

We performed manual gating of FCS files using CellEngine (CellCarta, Montreal, Canada) following the gating strategy found on `app.teiko.bio/projects/MGH333/overview`.

FastPG

We clustered cells using Fast-PhenoGraph (FastPG), which uses a K-nearest neighbors algorithm to identify the k cells with the most similar marker expression to each cell. After assigning neighbors, each pair of neighboring cells is assigned a Jaccard index weight, corresponding to the number of shared neighbors. Lastly, the Louvain algorithm is applied to form clusters that optimize the weight between connections.

Naming and Quality Control

We applied our naming algorithm to assign biologically relevant names to cell clusters by evaluating their marker expression. The algorithm assigned biologically relevant labels (e.g., "CD8+ T Memory" vs. "CD8+ T Naive" based on CD45RA), refined by marker-specific distinctions (e.g., CD161pos).

To ensure clusters were correct, we generated a cluster versus marker heatmap and series of UMAP plots, one for each cluster and marker (coloring only the selected cluster or marker).

Acknowledgements

We are grateful for the support from Genevieve Boland, MD and her colleagues at Massachusetts General Hospital for providing the PBMC samples analyzed in this study.

Conclusions

This dataset spans almost 29 million cells, and using our techniques, we uncovered distinct immune cell types that traditional gating would struggle to resolve. The specific immune cell type is a CD161-positive, CD39-negative subset of CD4+ T cells. This population showed statistically significant changes in marker expression across two functional markers: an increase in the expression of TBET, and a decrease in CCR7, in responder patients.

In the literature, we found these cells are associated with better prognosis of cancer therapy. Integrating unsupervised clustering with advanced quality control methods can overcome the inherent challenges of high-dimensional data analysis, allowing immunotherapy developers to find precise cell types associated with specific endpoints (i.e. dose or response).

Our work confirms findings from the literature

- Higher expression of CD161 and lack of CD39 and CCR7 in CD4 T cells are associated with Th17 phenotype and was previously reported with higher survival of cancer treatment. *Duurland, et al. J Immunother Cancer (2022) Jan;10(1):e003995.*
- T-bet expression is associated with Th1 phenotype and has been reported with better prognosis of NSCLC. *Laheurte, et al. Br J Cancer 121, 405–416 (2019)*

Cluster Naming

(1) Cluster cells with FastPG.

Each cell is assigned a cluster.

Clusters [48 X 48]
Each cluster contains ~500K-2M cells

	CD3	CD19	CD56	
[1]	4.2	0.1	0.3	# Cluster 1
[2]	0.2	3.8	0.2	# Cluster 2
[3]	0.3	0.2	4.1	# Cluster 3

(2) Manually gate cells.

Each cell is assigned a gate.

Median of MCV for Gated cell populations [36 X 48]

	CD3	CD19	CD56
T_CELL	4.0	0.2	0.3
B_CELL	0.3	3.9	0.2
NK_CELL	0.2	0.1	4.0

(3) Calculate cosine similarity * between the cluster expression table and the gated expression table.

Label cluster with cell type that has highest value.

Similarity matrix [36 X 48]

	Cluster1	Cluster2	Cluster3
T_CELL	0.98	0.15	0.20
B_CELL	0.12	0.99	0.14
NK_CELL	0.18	0.13	0.97

New Cluster Names
1_T_CELL
2_B_CELL
3_NK_CELL

Manually QC clusters

(4) Look at UMAP expression plots vs UMAP plots with labeled clusters
(see right panel for example)

We run FastPG to generate clusters. For every identified cluster, we calculate the median channel value (MCV) across all markers.

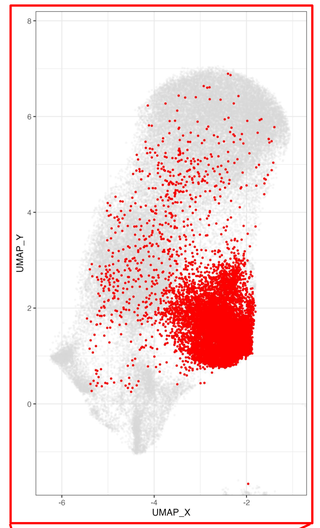
Separately, we manually gate populations of cells. We gather the MCV for all markers across all cells in each population.

We calculate a cosine similarity between the cluster cells and the gated cells matrices. This results in a matrix where the columns are clusters and the rows are gated cell populations. Each cluster population has a similarity score for every gated cell population. We identify the gated cell population with the highest score for each cluster. We use this to rename the cluster.

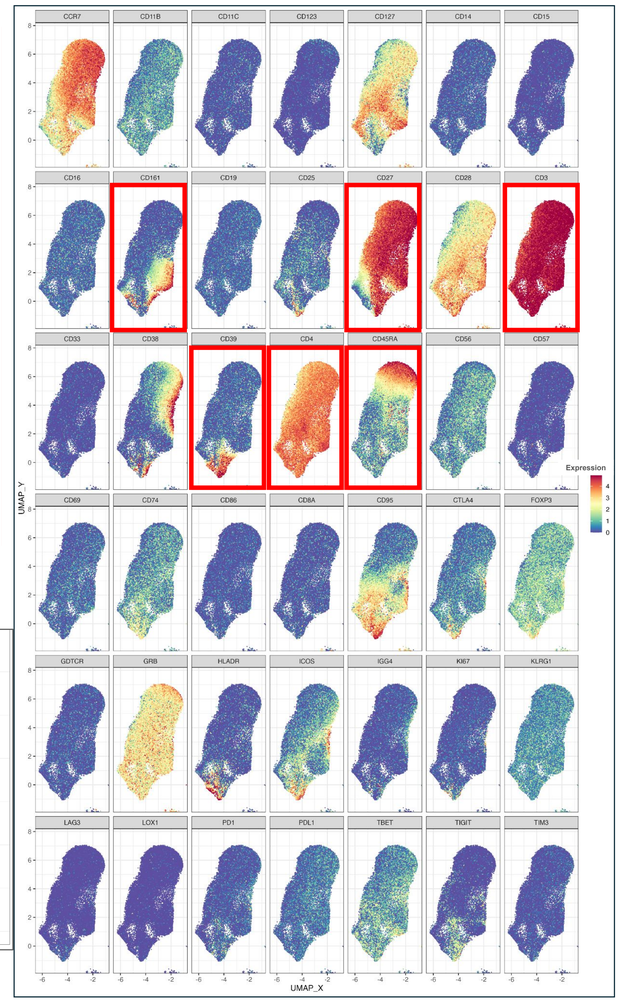
* Cosine similarity is the dot product of two vectors divided by the product of their lengths, i.e. (A * B) / (|| A || * || B ||)

CD161+ CD39- CD4+ T Cell Subset Discovered with High-Dimensional Cytometry, FastPG, & UMAP

Red CD4+ T Island: 1.8M cells



Start with ~29M cells



After the clustering of ~29M cells and automatic naming has finished, we zoom in on each cluster to inspect whether the name aligns with marker expression.

Here, we show Cluster-16 where the expression does not follow a cell type identified by gating.

In the example, we zoom in on the T cell island, and color Cluster-16 cells red and all other cells grey. Next, we create a UMAP plot for each markers' expression. Blue is low expression, and red is high expression.

Our algorithm named Cluster-16 as a CD4 T central memory cell due to its expression of CD27 and lack of expression of CD45RA.

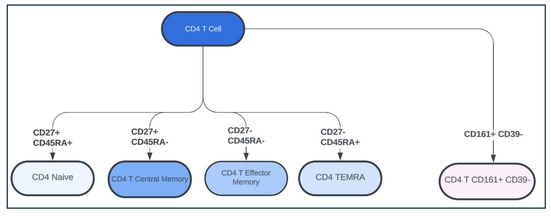
To verify this cluster name, we look at the labelling of CD3 and CD4 to identify whether it is a CD4 T cell. Next we look at the markers for T cell subsets, CD27 and CD45RA.

This cluster has low expression of CD45RA which matches central memory cell labelling. When looking at the CD27 expression the expression was not uniform and showed variability.

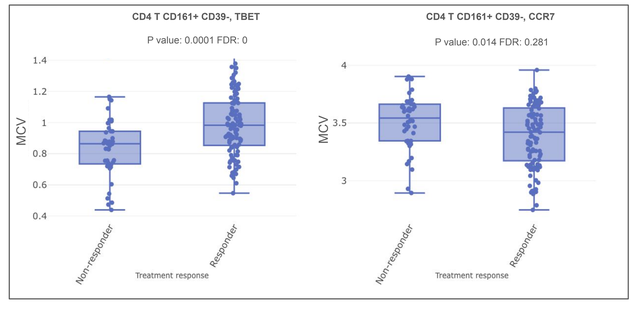
We next look at other markers that explain the shape of the cluster. In this cluster, we found high expression of CD161 and low expression of CD39.

This leads us to believe we have identified a cell type outside of traditional immunology, which would have been missed by traditional gating.

Responder and Non-Responder Show Statistically Significant Differences in the Expression of TBET and CCR7



A gating tree showing the traditional method to classify cells within CD4 T cells in blue. In pink we show our newly discovered non traditional T cell.



This dataset is available on our dashboard at: app.teiko.bio/projects/MGH333 | teiko.bio | SMS at (385) 355-9108