Analyzing multimodal single cell sequencing data like an Immunologist with hierarchical gating in CellEngine software

Matt Clutter[1], Sarah Asbury[2], David Riley[1], Pamela White[1], Jean-Philippe Goulet[2], Eustache Paramithiotis[2], Zach Bjornson[1]
[1] CellCarta, Fremont, CA, USA [2]CellCarta, Montreal, Canada

POSTER NUMBER: 38

CellCarta

ABSTRACT

Rapidly improving methods and decreasing sequencing costs are accelerating the use of single cell sequencing for deep immune monitoring. CITE-seq, a version of single cell sequencing, generates highly multiplexed protein and gene expression readouts[1]. Dimensional reduction and clustering are often used to define immune cell populations for CITE-seq analysis. **We examined the potential to identify cell types in CITE-seq data using hierarchical gating**, an approach favored by many Immunologists for the analysis of flow cytometry data. Compared to clustering, gated populations have clear definitions based on specific marker expression levels. **We compared cell type assignment by hierarchical gating vs. unbiased clustering; clustering did not separate every immune population we would typically gate in a cytometry experiment, however it did identify distinct subsets within our gated populations that our hierarchy failed to define.** Among others, the Vα7.2+ mucosal-associated invariant T (MAIT) cell population could be divided into two subpopulations based on numerous protein and gene features. **We propose a hybrid approach to cell type identification for CITE-seq** that is driven by hierarchical gating and uses clustering to refine the hierarchy. By leveraging analysis tools that are familiar to Immunologists, we highlight how these complex datasets can be analyzed more intuitively than many realize today.

METHODS

Two lab operators each thawed two PBMC samples from a single healthy donor. Samples were stained with the Biolegend TotalSeq-C universal human cocktail, then 14,000 cells per sample were captured on the 10X Genomics Chromium X using the 5' HT v2 workflow. RNA and Antibody libraries were generated and sequenced on an Illumina NovaSeq 6000 S2 flow cell to obtain approximately 28,000 RNA reads/cell and 12,000 Antibody reads/cell. 10X Genomics Cell Ranger was used to generate feature-barcode matrix H5 files. A CellCarta pipeline leveraging aspects of Seurat[2] was used to generate customizable FCS files from H5 files containing the following parameters for each cell:

- 137 protein features from TotalSeq antibody staining
- 295 sctype marker genes associated with various immune cell populations[3]
- 1,000 of the most highly variable genes selected across all cells of the experiment
- 164 gene set enrichment scores defined by the MSigDB Hallmark and GO Biological Processes collections, computed by scGSEA[4]
- 4 dimensionality reduction parameters (UMAP1, UMAP2, tSNE1, tSNE2)

The resulting 1,600-parameter FCS files were uploaded to CellCarta's cloud-based CellEngine software for analysis.

RESULTS

Cells from 4 replicate samples were gated using canonical protein markers from the CITE-seq data to identify 23 immune populations (**Fig 1A**). Population frequencies showed minor variation across the replicates (**Fig 1B**). Additional protein features were used to cluster gated populations and provide broader characterization (**Fig 1C**).

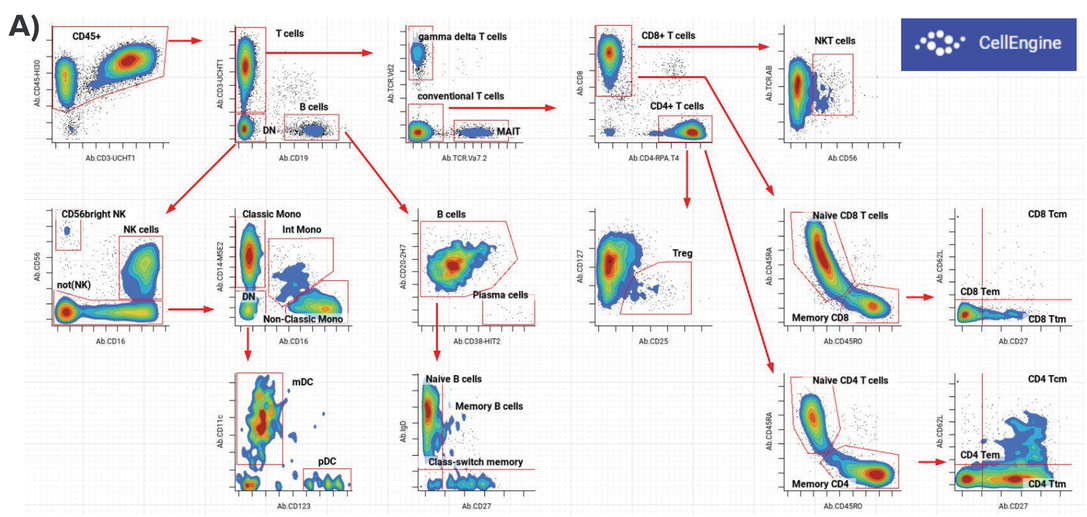

Figure 1. Identification of immune cell populations on CITE-seq data by hierarchical gating. (A) Gating hierarchy for 23 immune populations. **(B)** Population frequencies ± standard error of the mean, showing consistent cell type identification across the 2 replicate samples and 2 laboratory operators. **(C)** An expanded set of immune-relevant protein features.

Next, the cells were clustered by the 22 antibody features used to gate immune populations in Figure 1. UMAP dimensionality reduction and Leiden Community Detection algorithms implemented in CellEngine resulted in consistent clustering across sample replicates (**Fig 2A**). By visualizing the frequency of each gated population assigned to each Leiden cluster **(Fig 2B**), we observed differences that required further investigation.

Case 1: Gated intermediate and non-classical monocytes were combined in a single cluster. Gating populations that are well established by literature may be justifiable even when they are not differentiated by clustering. However, clustering is useful as it prompts a closer look to ensure that the gating, which is performed on two parameters (i.e., CD14 x CD16), reflects broader differences expected of truly distinct populations. Looking across the 137 protein features, we observed numerous differences between these two populations (**Fig 3A**). Multiple genes associated with intermediate or non-classical monocytes corroborated the relevance of these gates (**Fig 3C**).

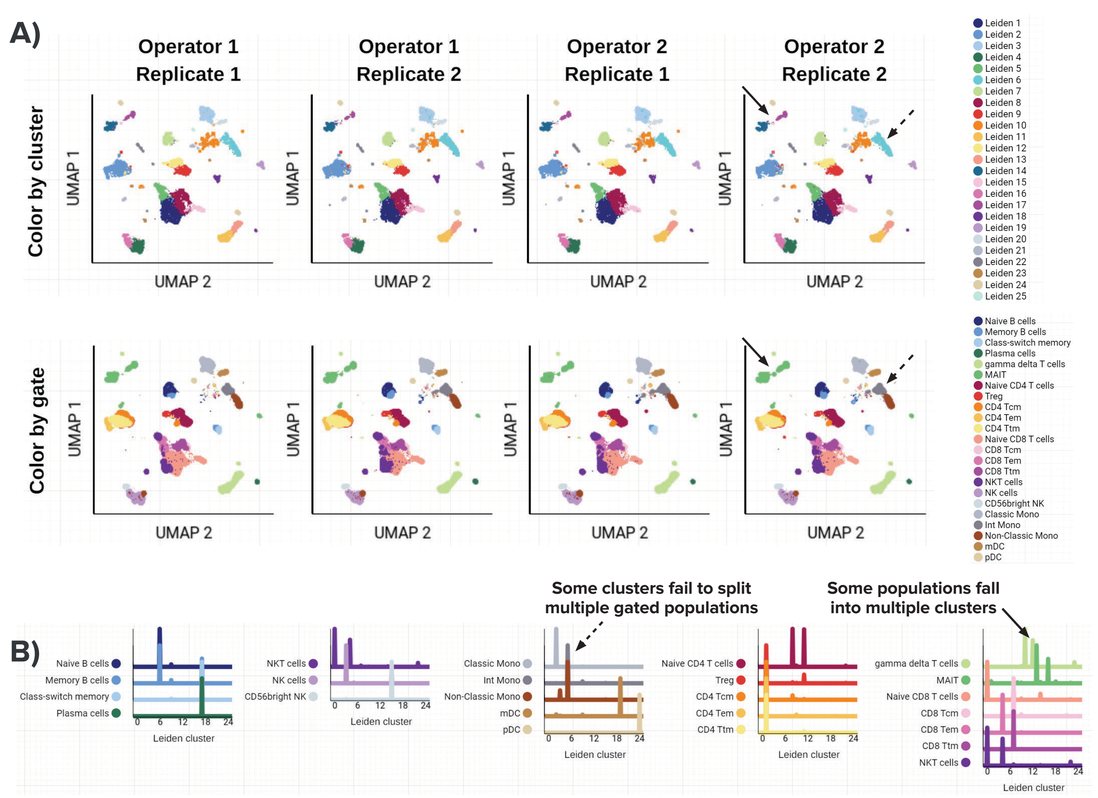

Figure 2. Comparing gated populations to unbiased cell clusters. Dimensionality reduction and Leiden clustering was performed. **(A)** Coloring cells by the 25 Leiden clusters (top) or the 23 gated populations (bottom) provided a global view of some differences. **(B)** Histogram plots more clearly showed cases where gating and clustering differed.

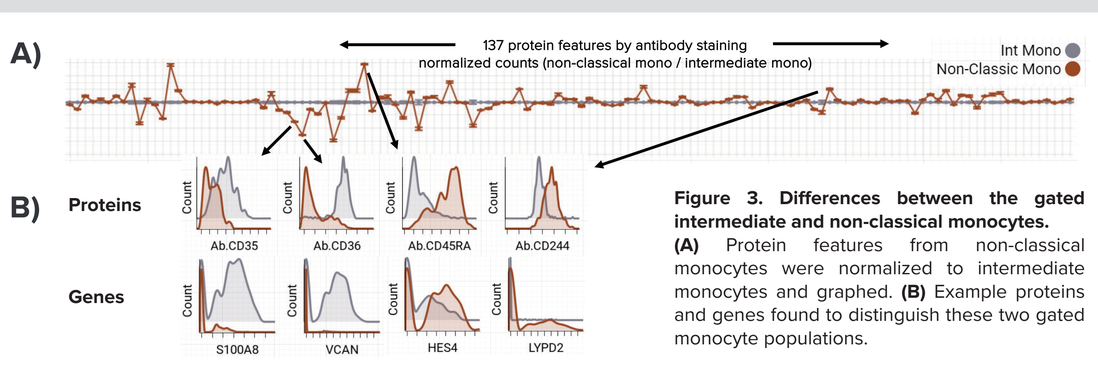

Figure 3. Differences between the gated intermediate and non-classical monocytes. **(A)** Protein features from non-classical monocytes were normalized to intermediate monocytes and graphed. **(B)** Example proteins and genes found to distinguish these two gated monocyte populations.

Case 2: Gated MAIT cells were split into two clusters. In cases where multiple clusters are identified within a single gated population, protein and gene features can be inspected to evaluate how distinct the clusters are. Leiden 17 showed similarity to Leiden 14 (Spearman ρ= 0.67), but was much more similar to effector memory CD8 T cells (ρ=0.95) and NKT cells (ρ=0.93) (**Fig 4A**). Multiple proteins and genes distinguished Leiden 14 from 17 (**Fig 4B**), suggesting that our gating strategy should be updated to subset the Vα7.2+ MAIT population. To survey the biological differences between the Leiden 14 and 17 MAIT cells, we compared 164 single cell gene set enrichment scores; Leiden 17 MAITs showed hallmarks of cell migration and extravasation with differentiated levels of adhesion proteins (**Fig 4C**).

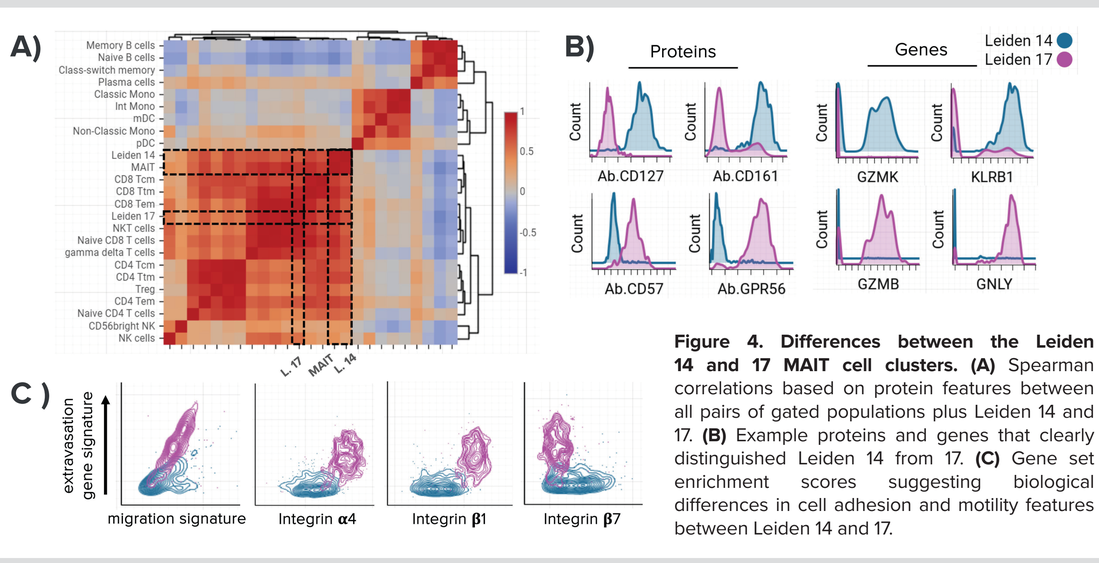

Figure 4. Differences between the Leiden 14 and 17 MAIT cell clusters. (A) Spearman correlations based on protein features between all pairs of gated populations plus Leiden 14 and 17. **(B)** Example proteins and genes that clearly distinguished Leiden 14 from 17. **(C)** Gene set enrichment scores suggesting biological differences in cell adhesion and motility features between Leiden 14 and 17.

KEY TAKEAWAYS

- Gating is familiar to Immunologists and defines populations based on established markers.
- When clustering fails to split two or more gated populations, the underlying markers can be investigated to understand whether the gating strategy is appropriate.
- Clustering can uncover distinct cell subsets within gated populations that the gating strategy missed.
- Gating and clustering together form a superior approach for population identification that makes the analysis of complex CITE-seq datasets more intuitive and comprehensive.
- CellEngine enables hybrid gating plus clustering in a single environment.

REFERENCES

1. Stoeckius et al. Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods 14:865-8 (2017).
2. Hao and Hao et al. Integrated analysis of multimodal single-cell data. Cell 184:3573-87 (2021).
3. Ianevski et al. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat. Commun. 13, 1246 (2022).
4. Franchini et al. Single-cell gene set enrichment analysis and transfer learning for functional annotation of scRNA-seq data. NAR Genom. Bioinform. 5(1):lqad024 (2023).

Please send inquiries to
Matt Clutter, PhD
Director, Global R&D
matt.clutter@cellcarta.com