CIDC 2.0 Bioinformatic Pipelines
CyTOF Validation

Essex Management

National Cancer Institute (NCI)

09/05/2024

Version 1.0

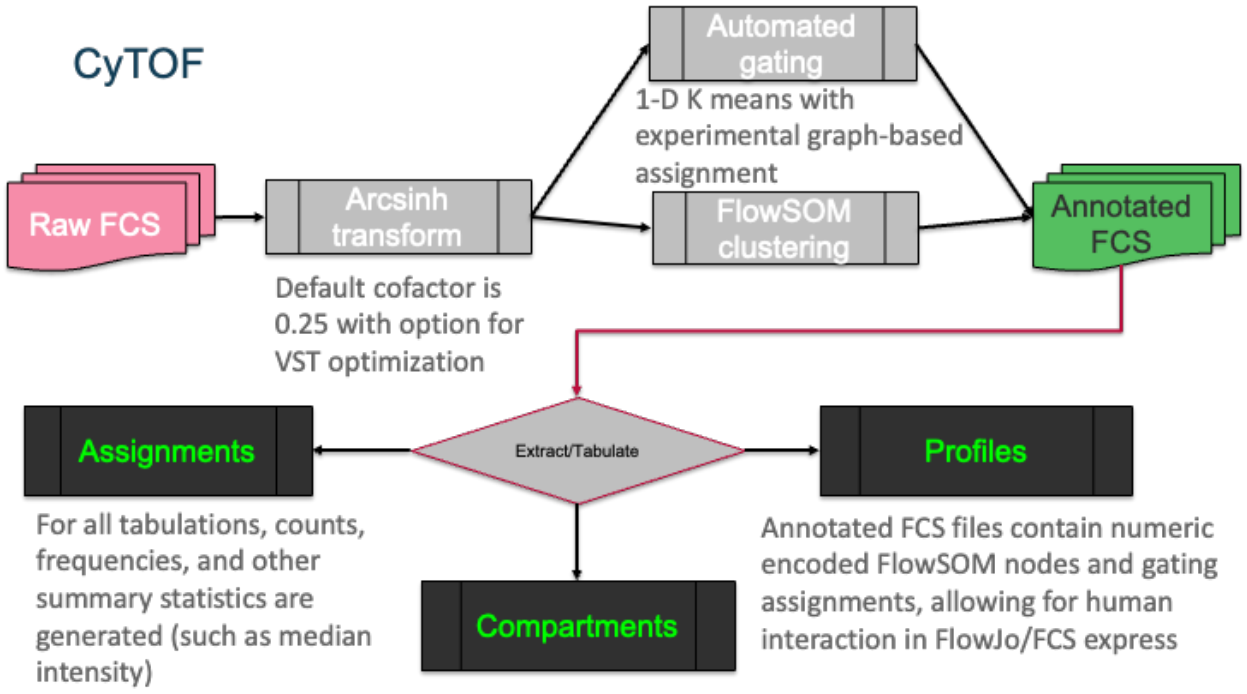
Draft

Submitted To:
Daoud Meerzaman
Computational Genomics and Bioinformatics Branch (CGBB)
National Cancer Institute
Center for Biomedical Informatics & Information Technology
Rockville, MD 20850

Submitted By:
Nick Renzette, Christina Coppola
Essex Management, LLC
11140 Rockville Pike | Suite 332
Rockville, MD 20852-3149
DUNS: 829872345
Cage Code: 5CYC9

1. INTRODUCTION

As part of the CIDC processes adopted from Dana Farber Cancer Institute, the National Cancer Institute (NCI) CIDC processed raw CyTOF data using a processing pipeline provided by a third party software vendor (Astrolabe). This process was evaluated and identified as an inefficiency that could be corrected to decrease data processing turnaround time and reduce long-term CIDC operational costs. A proposal was created by the CIDC Bioinformatics group, led by Dr. Daoud Meerzaman, in which the CyTOF processing pipeline (shown below) was brought within the CIDC cloud computing system, eliminating the need for a third party software vendor and the associated cost and processing complexity. This validation report describes the activities used to evaluate the in-house CyTOF bioinformatic pipeline. The goal of the validation is show to equivalency of the in-house and Astrolabe pipelines.



2. CYTOF PIPELINE – VALIDATION DATASET

Validation Dataset:

For validation, we will use a dataset previously analyzed with the Astrolabe processing pipeline. These data were collected for correlative analyses related to trial GU16-257. The full results have been reviewed by the bioinformatics team for completeness.

Sample	Data Location	Source File
CM5P0GCEY.01	gs://astrolabe-example--gu16-257/inputs/CM5P0GCEY.01.fcs	source.fcs

CM5P0GCH4.01	gs://astrolabe-example--gu16-257/inputs/CM5P0GCH4.01.fcs	source.fcs
CM5P0GCIV.01	gs://astrolabe-example--gu16-257/inputs/CM5P0GCIV.01.fcs	source.fcs
CM5P0GCKT.01	gs://astrolabe-example--gu16-257/inputs/CM5P0GCKT.01.fcs	source.fcs
CM5P0TR33.01	gs://astrolabe-example--gu16-257/inputs/CM5P0TR33.01.fcs	source.fcs
CM5P0TR8R.01	gs://astrolabe-example--gu16-257/inputs/CM5P0TR8R.01.fcs	source.fcs
CM5P0TRM9.01	gs://astrolabe-example--gu16-257/inputs/CM5P0TRM9.01.fcs	source.fcs
CM5P0TRTV.01	gs://astrolabe-example--gu16-257/inputs/CM5P0TRTV.01.fcs	source.fcs
CM5P2QBDT.01	gs://astrolabe-example--gu16-257/inputs/CM5P2QBDT.01.fcs	source.fcs
CM5P2QBO2.01	gs://astrolabe-example--gu16-257/inputs/CM5P2QBO2.01.fcs	source.fcs

3. CyTOF Pipeline – Validation Method

Validation Design: The validation design is intended to verify that the files generated by the in-house CyTOF pipeline matches the output of the Astrolabe pipeline. The CyTOF results uploaded to the Portal (and thus distributed to the CIMAC-CIDC network) are shown below:

- source.fcs – 'raw' fcs file supplied by the CIMAC data generator
- assignment.csv – matrix output – Channel x Cell Type
- compartment.csv – matrix output – Channel x Immune Compartment
- profiling.csv - – matrix output – Channel x Granular Cell Type
- cell_counts_assigment.csv – count of cells assigned to each cell type
- cell_counts_compartment.csv – count of cells assigned to each Immune Compartment
- cell_counts_profiling.csv - count of cells assigned to each granular cell type

For validation, the 'cell_counts' outputs (assignment & compartment) will be compared between the previous Astrolabe runs and the in-house CyTOF processing pipeline. These files are selected because they are summary files and thus are dependent on upstream results to be generated. Thus, concordance between these files from the two pipelines will increase the confidence that the overall output is concordant.

Acceptance Criteria:
Spearman Correlation: $\rho > 0.95$

Investigation of Discrepancies:
If the comparison metrics do not meet the acceptance criteria, an investigation will be carried out by members of NCI-CGBB and EM. The members to perform the investigation will be designated by Daoud Meerzaman, based on expertise and availability. The investigation should last no longer than 10 business days, at which time a report that outlines the problem and suggests solutions will be presented to Daoud Meerzaman.

4. CyTOF Pipeline – Validation Results

Reproducibility of In-house pipeline - Cell Subtype Assignments

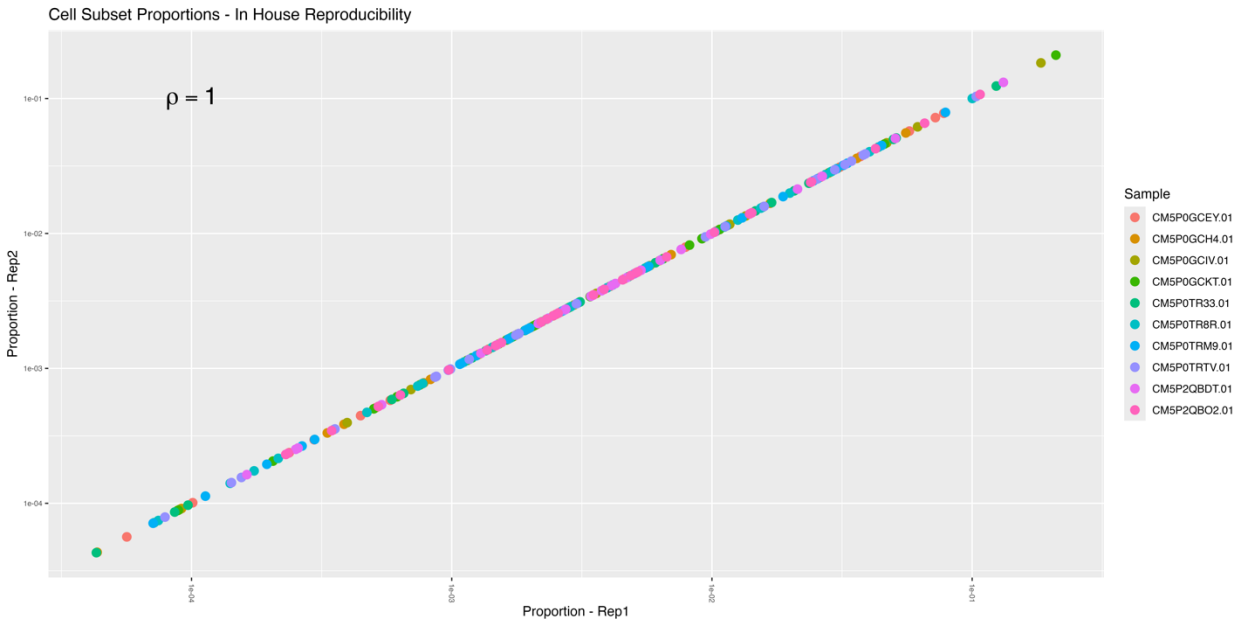


Figure 1: Scatter plot of cell assignment – Rep1 and Rep2 of In-house pipeline

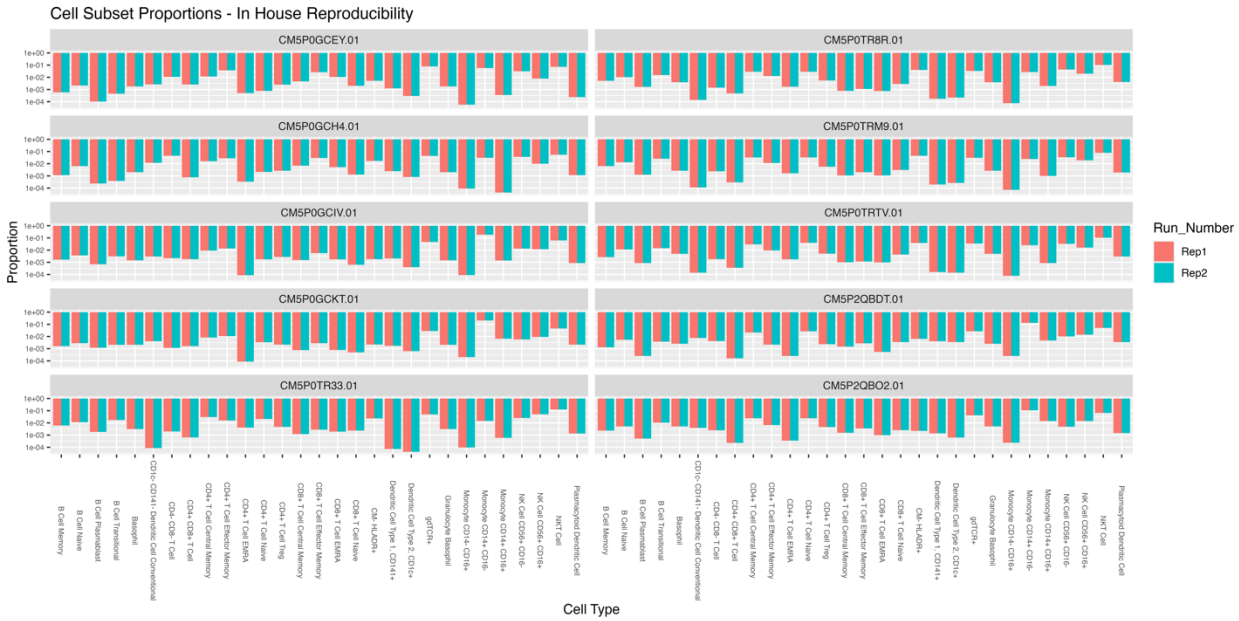


Figure 2: Cell assignment proportions for all validation samples – Rep1 and Rep2

Comparison of Astrolabe and In-house pipeline - Cell Compartments

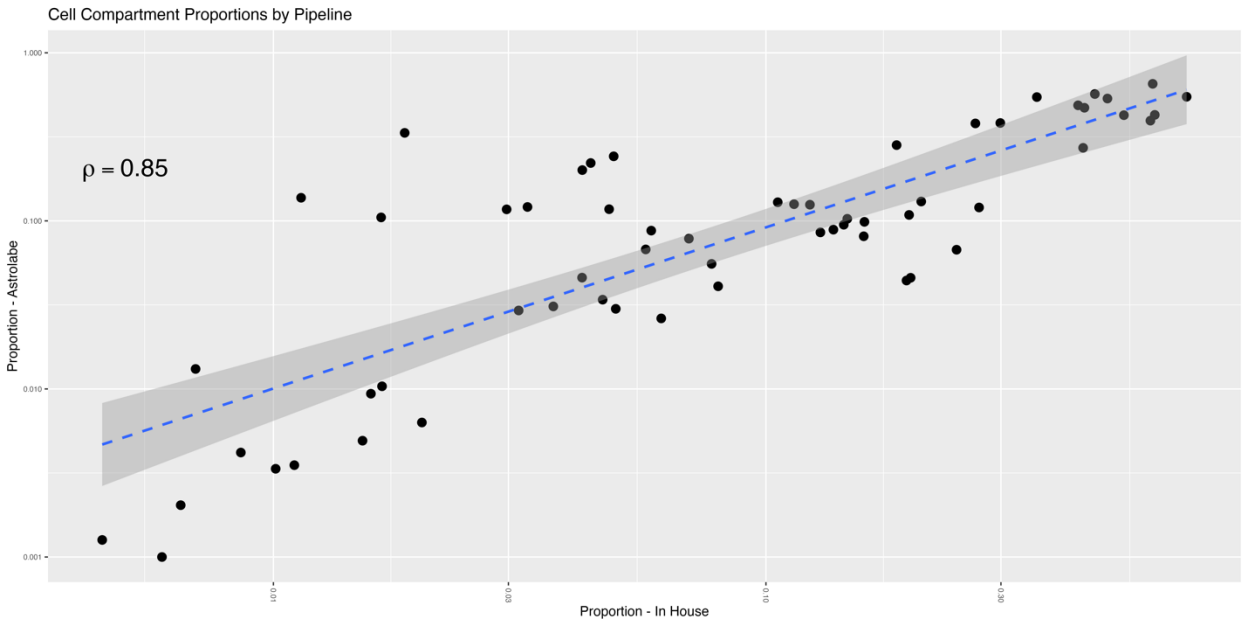


Figure 3: Scatter plot of cell compartment proportions

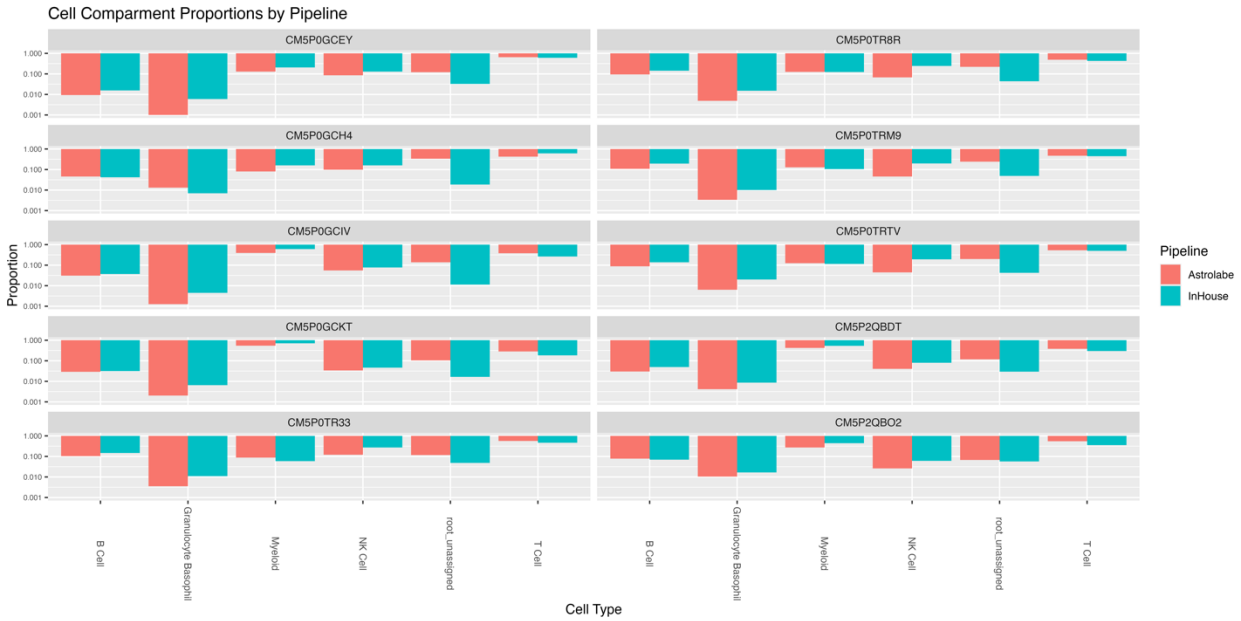


Figure 4: Cell compartment proportions for all validation samples

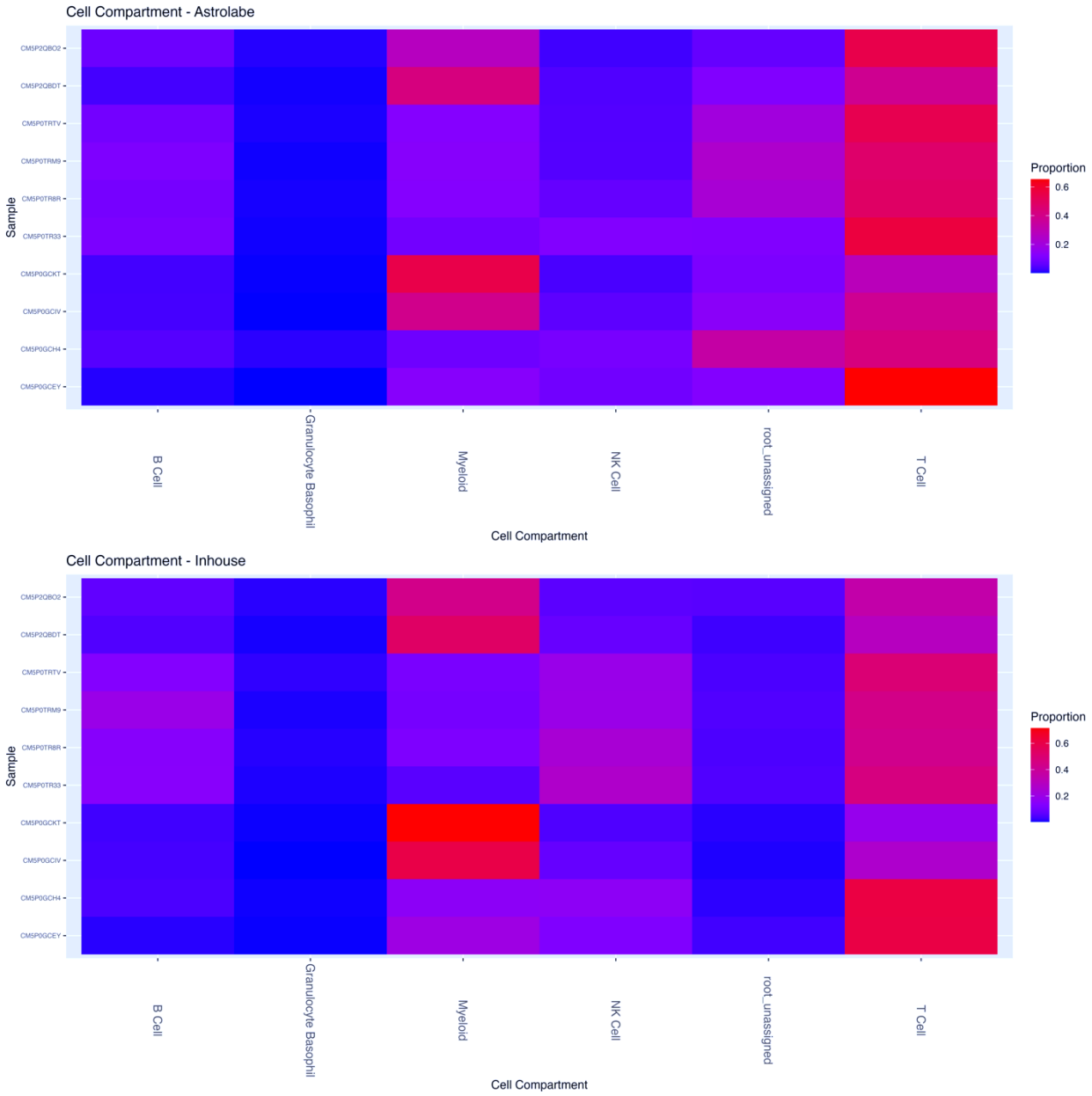


Figure 5: Heatmap of cell compartment proportions for Astrolabe and In-house pipelines
Notably difference observed in 'root_unassigned' (in-house label) vs. 'other' (Astrolabe) categories.
Data were re-analyzed with these categories removed (Figure 6).

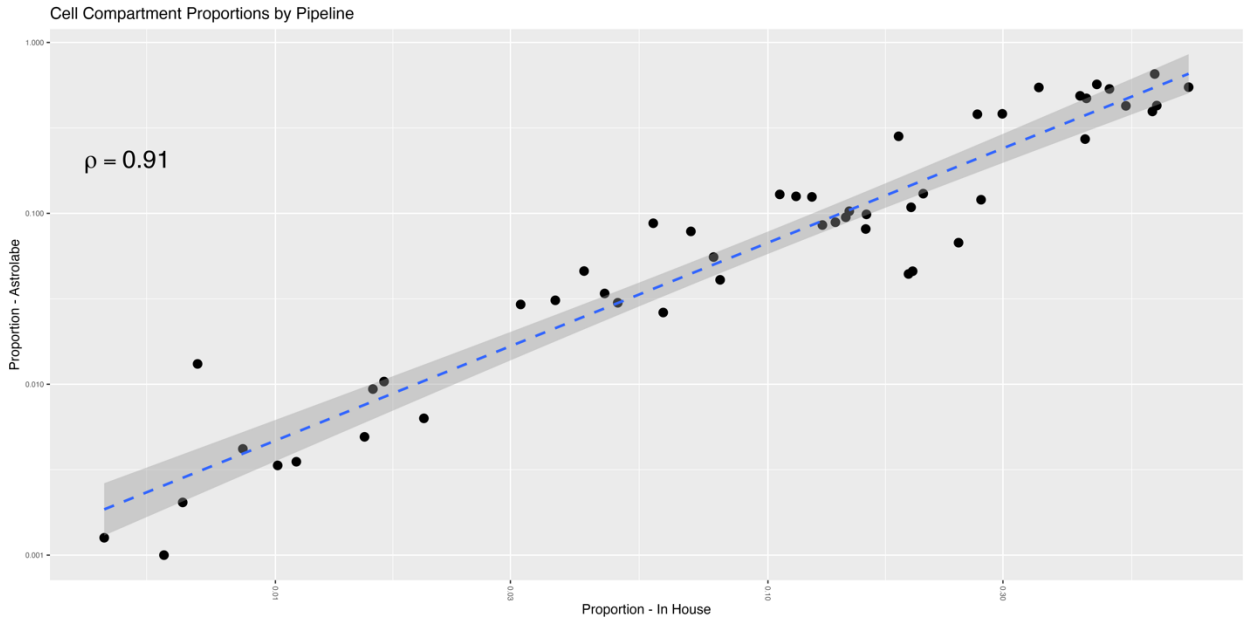


Figure 6: Scatter plot of cell compartment proportions – exclude 'root_unassigned'/'other' categories.

Comparison of Astrolabe and In-house pipeline - Cell Subtype Assignments

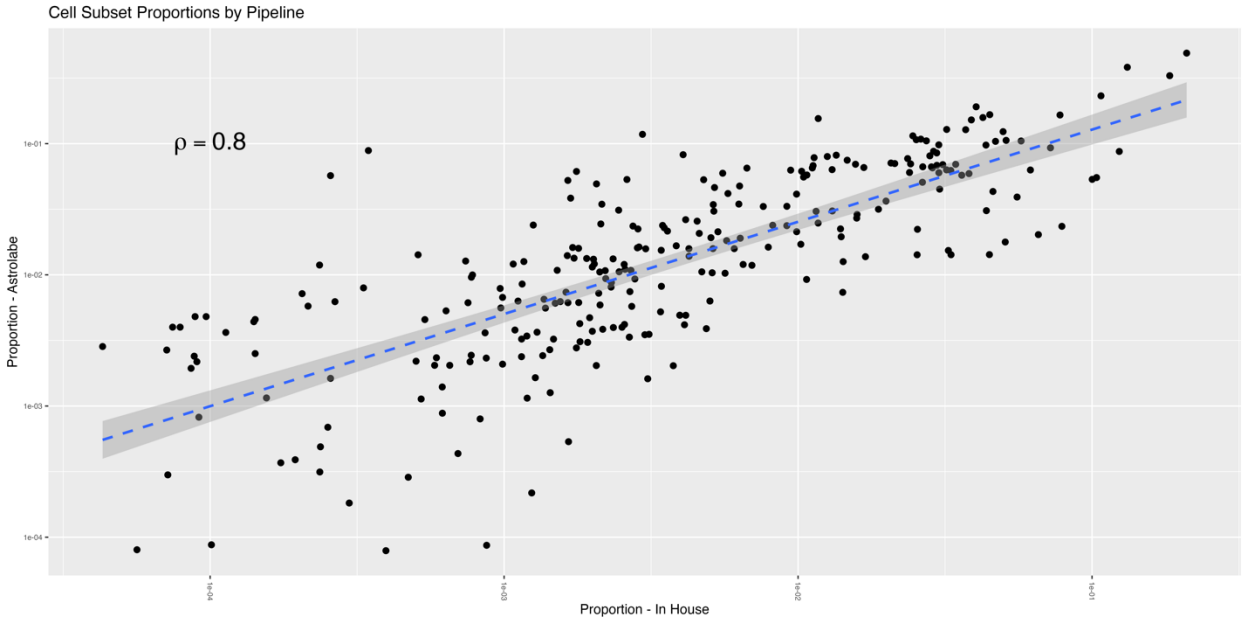


Figure 7: Scatter plot of cell assignment proportions

Figure 8: Cell assignment proportions for all validation samples

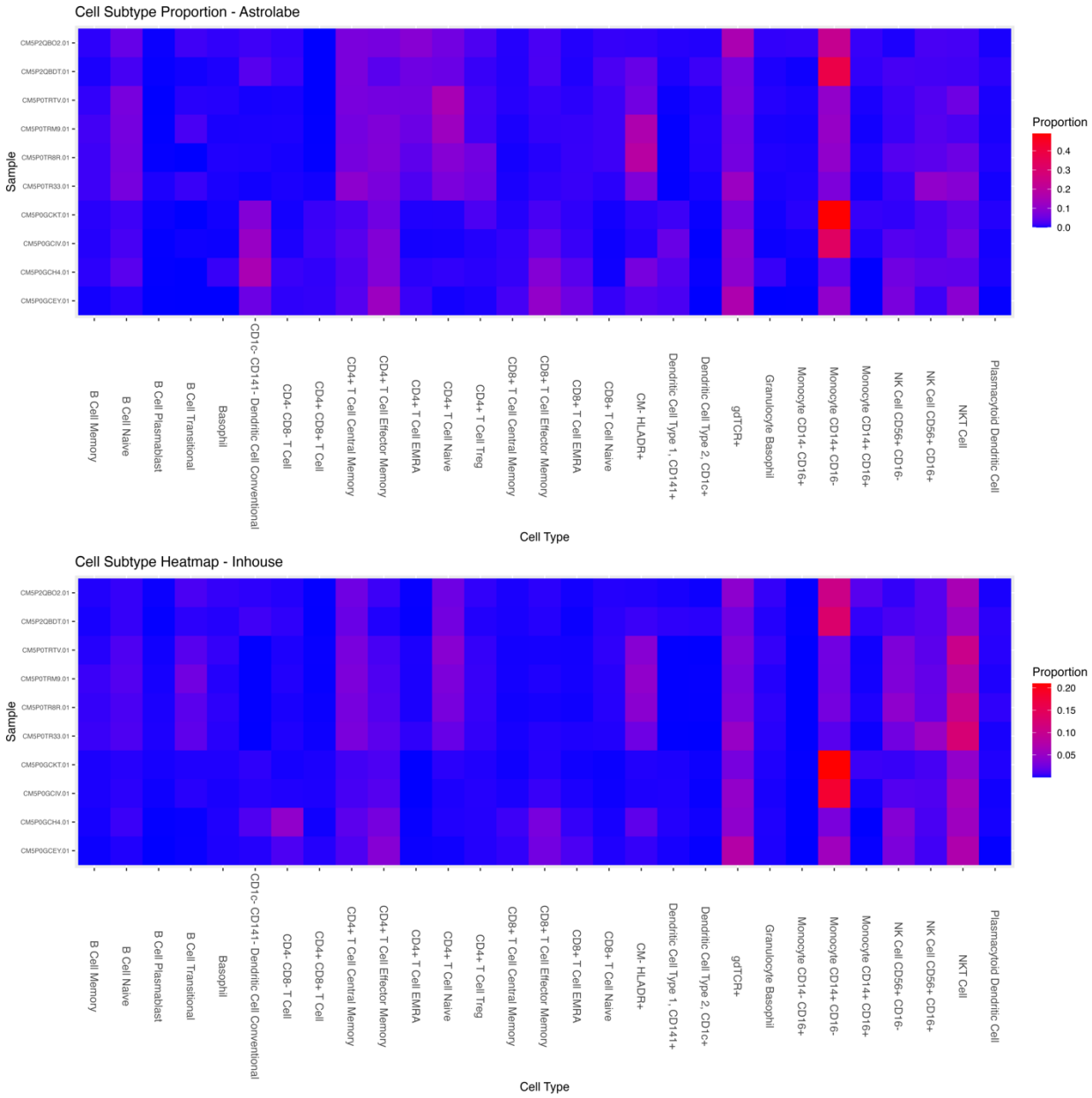


Figure 9: Heatmap of cell assignment proportions for Astrolabe and In-house pipelines

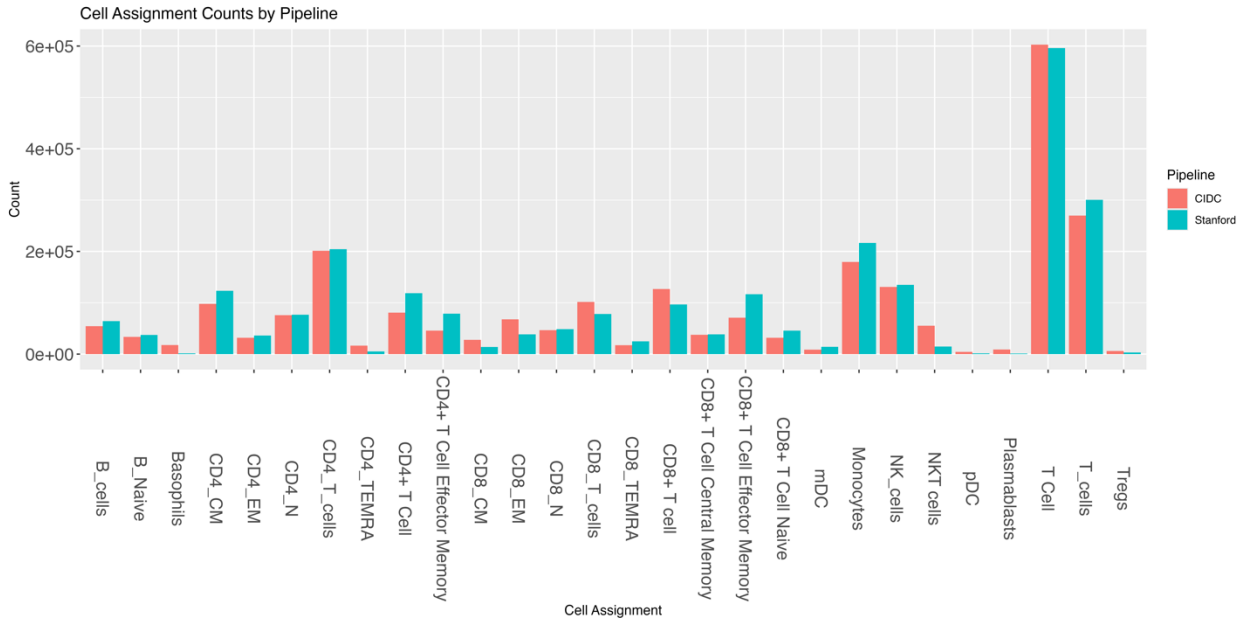
Conclusion: There is excellent qualitative and quantitative agreement between the previous Astrolabe CyTOF analysis pipeline and the in-house. Differences were noted in the calling of ‘unassigned’ or ‘other’ cells. These categories are reserved for cells with low-confidence calls (for example, due to weak or inconclusive signals). There was strong agreement in the pipelines for the other categories. Thus, it is recommended that the in-house pipeline replace the Astrolabe

pipeline. If requested, samples previously analyzed with the Astrolabe can be re-processed with the in-house pipeline.

Planned Future Addendum: The in-house CyTOF pipeline will be compared to results generated by the Stanford CIMAC which includes manual gating. **NOTE: Added on 09/05/2024 – See Section 5.**

5. CyTOF Pipeline – Validation Results – Addendum – Comparison to Stanford Results

As an addendum to the original validation report, results from the CIDC CyTOF pipeline were compared to those produced by Stanford's 'manual gating' processing pipeline. The Stanford approach is an excellent method for identifying and quantifying rare cells in the population, which may not be detectable using fully automated methods. The process does require highly trained personnel, though, to review and modify the processing parameters. Thus, the two pipelines offer different advantages for the analysis of CyTOF data. However, it is important to compare the results to understand the degree of agreement and identifying any areas where results may differ.



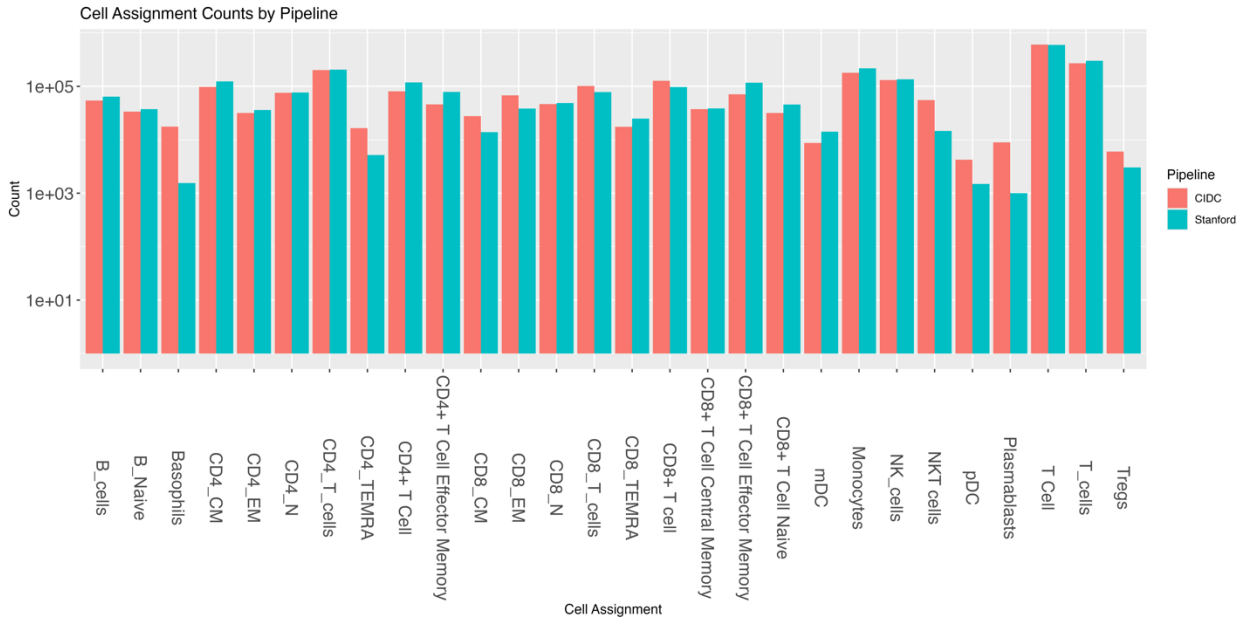
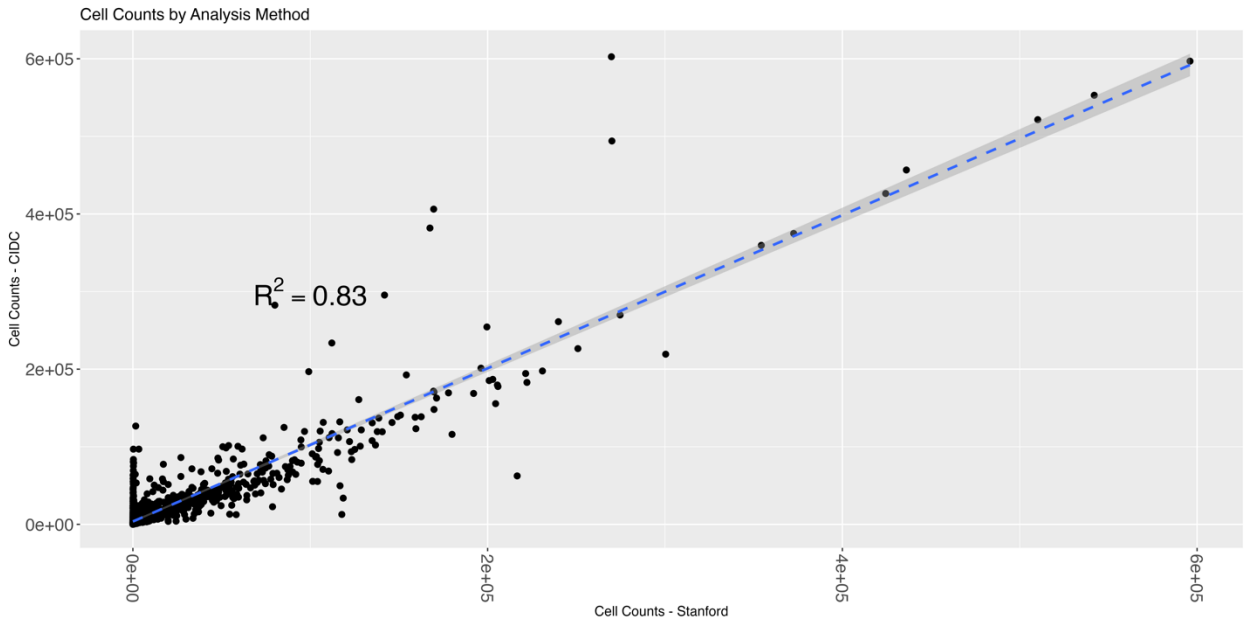


Figure 10: Cell assignment counts produced by CIDC and Stanford processing pipelines – Linear and Log Scales



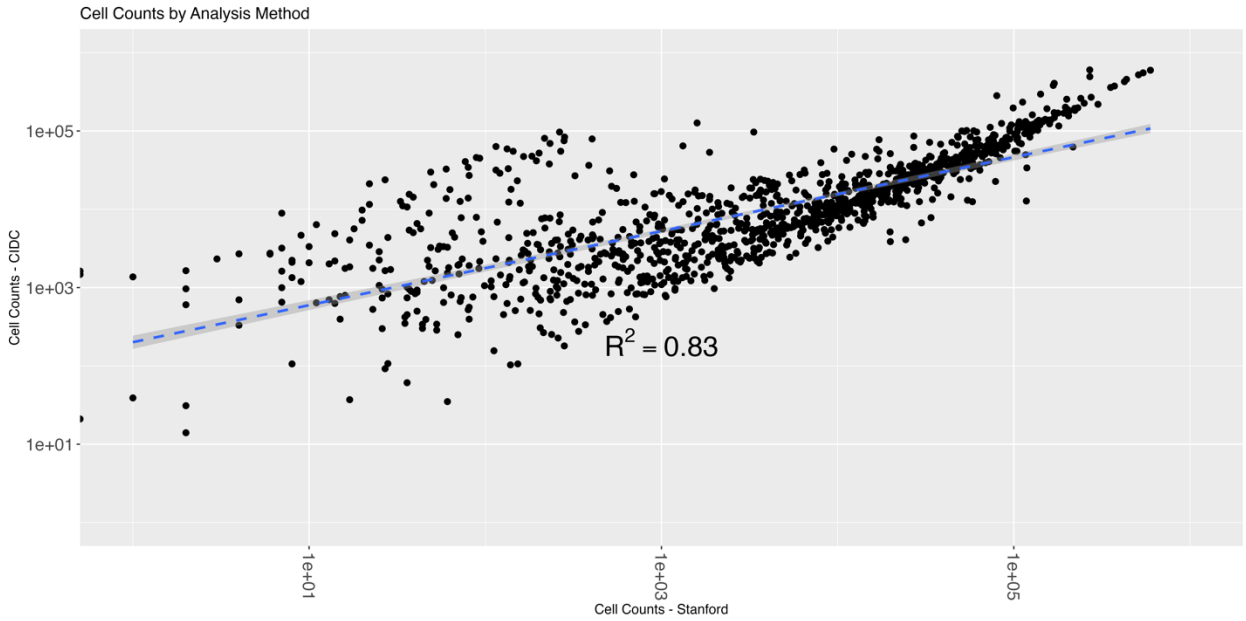


Figure 11: Scatter plot of cell assignment counts – Linear and Log Scales

The bar and scatter plots showed strong agreement between the two methods. However, it was noted that there was a notable divergent population of assignments that were called at low levels in the Stanford method and at higher levels in the CIDC method.

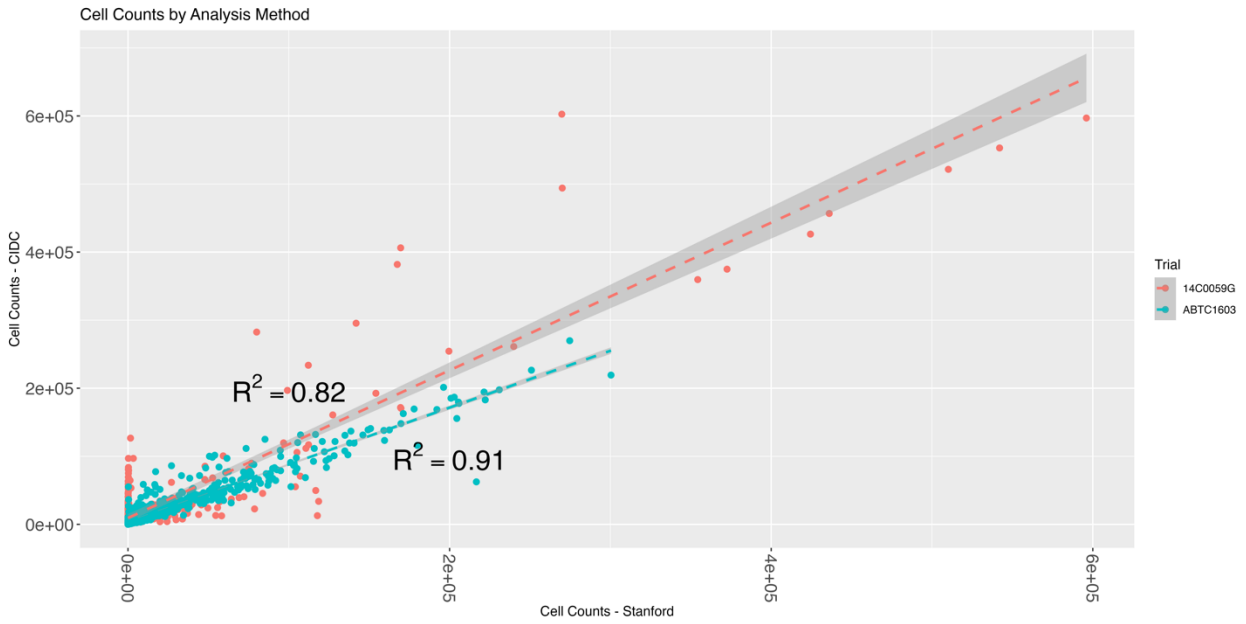


Figure 12: Scatter plot of cell assignment counts – Linear and Log Scales – Separate by Trial

By separating the data by trial, it became apparent that there was a stronger correlation for the ABTC-1603 dataset as compared to the 14-C-0059G dataset. Further, the 14-C-0059G dataset showed a marked population of the CIDC_High / Stanford_Low counts. This was investigated further and shown to be concentrated in cells in the T cell lineage. It is noted that the CIDC pipeline allows for 19 distinct T cell assignments, whereas the Stanford analysis allowed for 481 distinct T cell assignments. Thus, these results likely reflect the known differences in methodologies, highlighting the advantage of the Stanford method for detecting and quantifying rare cell population subsets within broader categories. The CIDC method, though, is shown to provide results in good agreement with this method in a fully automated and highly reproducible fashion.

6. CyTOF Pipeline – Appendix I – Tabular Results

Sample	Cell_Compartment	Proportion - In House	Proportion - Astrolabe
	B Cell	0.016	0.009
	Granulocyte		
	Basophil	0.006	0.001
	Myeloid	0.207	0.130
	NK Cell	0.129	0.085
	root_unassigned	0.033	0.121
CM5P0GCEY	T Cell	0.610	0.654
	root_unassigned	0.018	0.334
	B Cell	0.042	0.046
	T Cell	0.616	0.427
	Myeloid	0.158	0.081
	NK Cell	0.158	0.099
	Granulocyte		
CM5P0GCH4	Basophil	0.007	0.013
	root_unassigned	0.011	0.137
	B Cell	0.037	0.031
	T Cell	0.266	0.380
	Myeloid	0.603	0.395
	NK Cell	0.078	0.055
	Granulocyte		
CM5P0GCIV	Basophil	0.004	0.001
	root_unassigned	0.017	0.105
	B Cell	0.031	0.029
	T Cell	0.184	0.282
	Myeloid	0.715	0.547
	NK Cell	0.047	0.034
	Granulocyte		
CM5P0GCKT	Basophil	0.006	0.002
	root_unassigned	0.044	0.221
	B Cell	0.144	0.095
	T Cell	0.430	0.487
	Myeloid	0.123	0.125
CM5P0TR8R	NK Cell	0.244	0.067

	Granulocyte		
	Basophil	0.015	0.005
	root_unassigned	0.048	0.117
	B Cell	0.146	0.103
	T Cell	0.465	0.569
	Myeloid	0.058	0.088
	NK Cell	0.271	0.120
	Granulocyte		
CM5P0TR33	Basophil	0.011	0.004
	root_unassigned	0.049	0.242
	B Cell	0.195	0.108
	T Cell	0.443	0.471
	Myeloid	0.106	0.129
	NK Cell	0.197	0.046
	Granulocyte		
CM5P0TRM9	Basophil	0.010	0.003
	root_unassigned	0.042	0.201
	B Cell	0.137	0.089
	T Cell	0.494	0.535
	Myeloid	0.114	0.126
	NK Cell	0.193	0.044
	Granulocyte		
CM5P0TRTV	Basophil	0.020	0.006
	root_unassigned	0.030	0.117
	B Cell	0.050	0.030
	T Cell	0.299	0.382
	Myeloid	0.533	0.426
	NK Cell	0.080	0.041
	Granulocyte		
CM5P2QBDT	Basophil	0.009	0.004
	root_unassigned	0.057	0.068
	B Cell	0.070	0.078
	T Cell	0.355	0.545
	Myeloid	0.441	0.272
	NK Cell	0.061	0.026
	Granulocyte		
CM5P2QBO2	Basophil	0.017	0.010