Date: September 28, 2021

Cancer Immune Monitoring and Analysis Center
Dana-Farber Cancer Institute
Harvard Medical School
Contact PI: Catherine J. Wu, MD (e-mail: cwu@partners.org)

Performance lab:
Translational Immunogenomics Lab (TIGL)
Catherine J. Wu, MD (Faculty Advisor)
Kenneth J. Livak, PhD (Technology Lead Scientist)
Shuqiang Li, PhD (Technology Senior Scientist)

Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) Analytical Performance Report, version 1.1

While single cell RNA sequencing (scRNA-seq) has greatly advanced the resolution at which immune cell subsets can be defined, a major obstacle is the reconciliation of cell definitions obtained through scRNA-seq with classic flow cytometry-based cell definitions that have been developed by immunologists over the past decades. Differences in sample quality and other batch effects can introduce apparently distinct transcriptomic subclusters that do not have biological correlates. Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) enables the linking of cell clusters obtained through single cell transcriptomics to classical immunobiology by providing information on cell surface protein expression at the single cell level in addition to scRNA-seq profiles.[1] CITE-seq is thus a crucial technology to distinguish biologically relevant cell clustering from technical artifact.

This report describes CITE-seq of six cryopreserved, characterized peripheral blood mononuclear cell (PBMC) samples obtained from a patient with relapsed $JAK2^{V617F+}$ secondary AML after allogeneic stem cell transplantation with transient response to PD-1 blockade (Penter et al., *Blood Advances* 2021)[2]. We demonstrate that CITE-seq of six consecutive samples from the same patient yielded consistent, high quality sequencing results. Additionally, we show that CITE-seq revealed similar cell type kinetics of T cells and AML blasts before PD-1 blockade, at time of response to PD-1 blockade and at relapse as mass cytometry that was performed on the same samples. Using a second dataset (Oliveira et al., *Nature* 2021)[3], we demonstrate reproducibility of CITE-seq on multiple replicates.

1. Intended use

By providing information on cell surface marker expression, the CITE-seq assay overcomes several limitations of scRNA-seq and thus improves the annotation of scRNA-seq cell clusters. The limitations include the sparse nature of scRNA-seq data and subsequent drop out of expression information, the insufficient coverage of lineage-defining genes such as *CD4*, and the inability of RNA-based sequencing to capture protein isoforms used to define cell types like those of the *CD45* gene. CITE-seq can therefore help to inform about cell lineage and may reduce inconsistent cell type annotations across sequencing batches.

While CITE-seq is a powerful technology, it has several inherent technical limitations such as reduced signal to noise ratio, higher experimental costs, and longer turn-around time between sequencing and data analysis compared to flow or mass cytometry. Therefore, CITE-seq is not intended to substitute for other protein-based single-cell technologies and must be optimized for more specialized use cases.

Table 1. Summary of analytical performance findings for CITE-seq

Accuracy	Accurate assay performance determined by length distribution of libraries (**Fig. 1**) and QC metrics of sequenced data (**Table 3**) as compared to reference ranges provided by 10x Genomics. Similar distribution of cell types when compared to an independent analysis method (**Fig. 3, 4**).
Precision	As indicated by the Standard Deviation column in **Table 3**, the six samples analyzed here had comparable values across all the QC metrics.
Analytical sensitivity	Targeting 5,000 single cells, depending on collection and preparation method and recovery rate. Here, we start with 0.5 million cells for CITE-seq staining and load 17,000 cells for library preparation.
Analytical specificity including interfering substances	Preservation and storage condition (fresh vs frozen) of samples, amount of antibody per cell, and number of pooled antibodies could influence results. Here we processed cryopreserved samples for single cell RNA sequencing and cell surface marker staining with 67 antibodies according to an optimized protocol with 10x Genomics recommended reagents.
Reportable range	Relative surface marker expression based on the number of reads per antibody. We observed a median number of 1,734 antibody sequencing reads per cell (range 1,386 – 2,233). Results are sample and antibody dependent.
Reference interval (normal range)	Generated data represent the reference ranges of QC metrics of publicly available PBMC data reported by 10x Genomics, see **Table 3**.
Standardization, harmonization, reproducibility, and ruggedness	Standardized by pooling unique indexed samples and sequencing on single lane of NovaSeq6000 Sequencer. All analysis for a given sample set performed against the same reference database.
Quality control and improvement procedures	Processing of scRNA-seq and CITE-seq sequencing data was done with the cellranger (10x Genomics) pipeline. The QC metrics are summarized in **Table 3**.
Any other performance data	Not applicable.

2. Materials and methods

Commercially available CITE-seq antibodies labeled with distinct oligonucleotide barcodes were purchased from BioLegend (**Table 2**). Prior to staining they were pooled by mixing equal volumes of each antibody. The concentration of each antibody was 0.5 µg/µL prior to pooling.

Thawing of cryopreserved PBMC samples
Slowly melt the contents of cryovials by placing the vials above the water of a water bath for about 10 minutes. Fill up each cryovial with 1 mL thawing medium (PBS + 10% grade II DNase I [cat. no. 10104159001, Sigma Aldrich] + 10% FBS) preheated to 37°C. Transfer contents of each cryovial to 15 mL Falcon tube. Slowly top up each Falcon tube with thawing medium over a period of about 3-5 minutes. Centrifuge Falcon tubes at 300×*g* for 5 minutes and discard supernatant. Resuspend cell pellet in preheated RPMI + 10% FCS + 10% DNase (cat. no. 07900, Stemcell Technologies). Cells were pelleted and resuspended in PBS + 0.04% bovine serum albumin before proceeding to antibody staining and processing on the 10x Genomics Chromium instrument.

CITE-seq procedure, library preparation, and sequencing
CITE-seq staining was performed at 4˚C as described by the CITE-seq protocol of the New York Genome Center Technology Innovation Lab (https://cite-seq.com/protocols/, version 2019-02-13). Briefly, $0.5x10^6$

cells were resuspended in 50 µL PBS + 0.04% ultrapure bovine serum albumin (BSA) (cat. no. AM2616, Invitrogen) and incubated with 10 µL Human TruStain FcX Fc Receptor Blocking Solution (cat. no. 422302, BioLegend) for 10 minutes. 33.5 µL of CITE-seq antibody pool (0.5 µL per antibody) were added to each sample and cells were incubated for 30 minutes. This was followed by 3 washing steps with PBS + 0.04% ultrapure BSA. Finally, cell concentrations were adjusted to 1,000/µL.

Subsequently, 17,000 cells per sample were loaded onto a Chromium Chip A (cat. no. 1000152, 10x Genomics). Single cell mRNA-derived and antibody-oligo-derived libraries were obtained using the Chromium Single Cell 5' Library & Gel Bead Kit (cat. no. 1000006, 10x Genomics). The Chromium i7 Multiplex Kit N Set A (cat. no. 1000084, 10x Genomics) was used to index different samples. Library preparations were performed as specified by the manufacturer's instructions. Quality control to assess library fragment length distribution and concentration was performed with a Bioanalyzer High Sensitivity DNA Kit (cat. no. 5067-4626, Agilent). Finally, libraries were sequenced on the NovaSeq6000 platform (Illumina) with paired end reads as follows: read 1, 28nt; read 2, 91nt; index 1, 8nt.

Data analysis
Cellranger count version 3.1.0 was used to process raw sequencing reads. Single cell RNA expression (scRNA-seq) and surface marker expression (CITE-seq) were processed together. Gene expression sequencing reads were aligned to the GRCh38-3.0.0 reference transcriptome. Further downstream analyses were performed with R studio and the Seurat package version 3.2.0 according to best practices described in the Seurat documentation (https://satijalab.org/seurat/articles/).[4,5]

Data availability
Single cell RNA sequencing data of Penter et al. can be accessed from GEO (GSE165496). CITE-seq data of Penter et al. can be accessed from GEO (GSE165822). Data of Oliveira et al. are available through the dbGaP portal (study ID: 26121, accession number: phs001451.v3.p1).

3. Results

We performed single cell RNA sequencing (scRNA-seq) and CITE-seq on six serially sampled peripheral blood mononuclear (PBMC) samples obtained from a patient with relapsed JAK2^{V617F+} secondary acute myeloid leukemia (AML). The samples were obtained before infusion of the PD-1 blocking antibody nivolumab (sample 1), at time of response to nivolumab (sample 2 and 3), at time of relapse (sample 4) and at time of unsuccessful re-exposure to nivolumab (sample 5 and 6). Separately, mass cytometry (CyTOF) was performed on the same six samples.

The antibody panel for CITE-seq consisted of 67 TotalSeq-C antibodies chosen to identify major cell lineages using canonic phenotypic markers and focuses on high resolution definition of T cell subpopulations and immune checkpoint molecules (**Table 2**).[6]

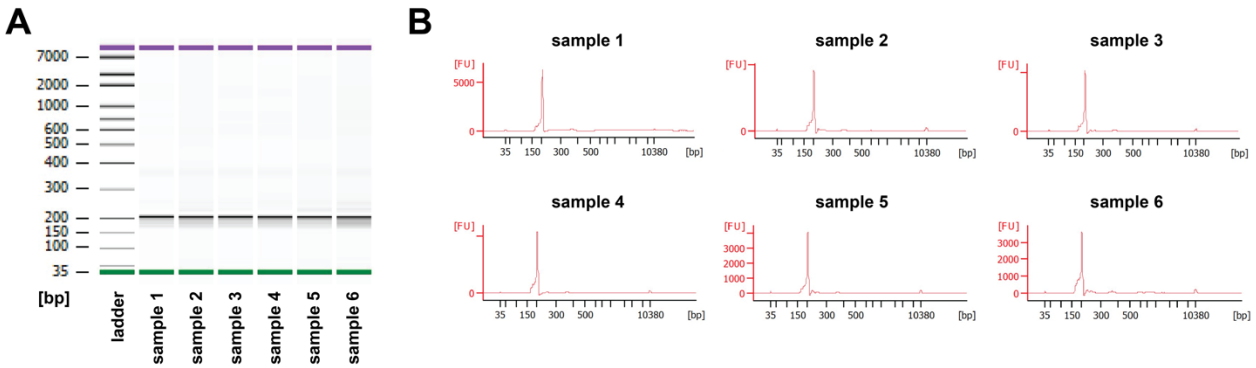

Figure 1. Quality control of CITE-seq libraries. A Fragment lengths **B** Peak distribution

Fragment length distribution of all six library preparations showed the typical Bioanalyzer traces expected for CITE-seq (**Fig. 1**).

Quality metrics provided by cellranger count demonstrated reproducible results (**Table 3**). High quality cells were identified using standard cut-offs recommended by the Seurat documentation[4]:

Minimum cells per gene: 3
Minimum genes per cell: 200
Maximum genes per cell: 2,500
Maximum UMIs per cell: 10,000
Maximum percentage of mitochondrial genes: 20

Cell doublets were identified based on mutually exclusive expression of monocyte-, T, and B cell-associated genes (*CD3D, CD3E, CD3G, CD4, CD8A, CD8B* vs. *CD79A, CD79B, MS4A1, CD19*) and proteins (CD3, TCRab vs. CD19, CD20 vs. CD14).

In total, we obtained an estimated 46,678 single cell profiles with a median estimated number of 8,828 single cell profiles per sample (range 3,597 – 11,426). These had a median number of 1,734 antibody sequencing reads per cell (range 1,386 – 2,233).

Major cell lineages (T cells, B cells, monocytes, and AML blasts) could be identified with scRNA-seq. On CITE-seq, these cell types showed the expected immunophenotypes, for example, expression of CD3, CD4 and CD8a on T cells (**Fig. 2A-C**), CD19 and CD20 on B cells (**Fig. 2D, E**), CD4, CD14, CD16 and CD33 on monocytes (**Fig. 2F-H**) and lower expression of CD45 on AML blasts (**Fig. 2I**).

We used mass cytometry (CyTOF) to assess how well CITE-seq detection reproduces results compared to a validated phenotyping method and observed a very high correlation for $CD3^+$ T cells, AML blasts and $CD4^+$ and $CD8^+$ T cells (**Fig. 3, Fig. 4**) with a correlation coefficient ranging from 0.89 to 0.99. For a complete list of stained antibodies, see Fig. 5.

In order to assess reproducibility of CITE-seq profiles, we re-analyzed another published dataset of 64,791 single tumor-associated T cells from 5 melanoma biopsies[3]. T cells of each sample had been flow-sorted and subsequently sequenced in 2-4 replicates (**Fig. 6A**). Data were imported as a pre-processed Seurat object provided by the authors. Replicates within each of the 5 analyzed samples were highly similar (**Fig. 6B**) and had a mean Spearman correlation coefficient of 0.94 (range 0.65 – 1), thus demonstrating high reproducibility of CITE-seq across replicates.

In sum, we show that CITE-seq yielded reproducible and robust results that correlated well with results from an independent analytical assay.

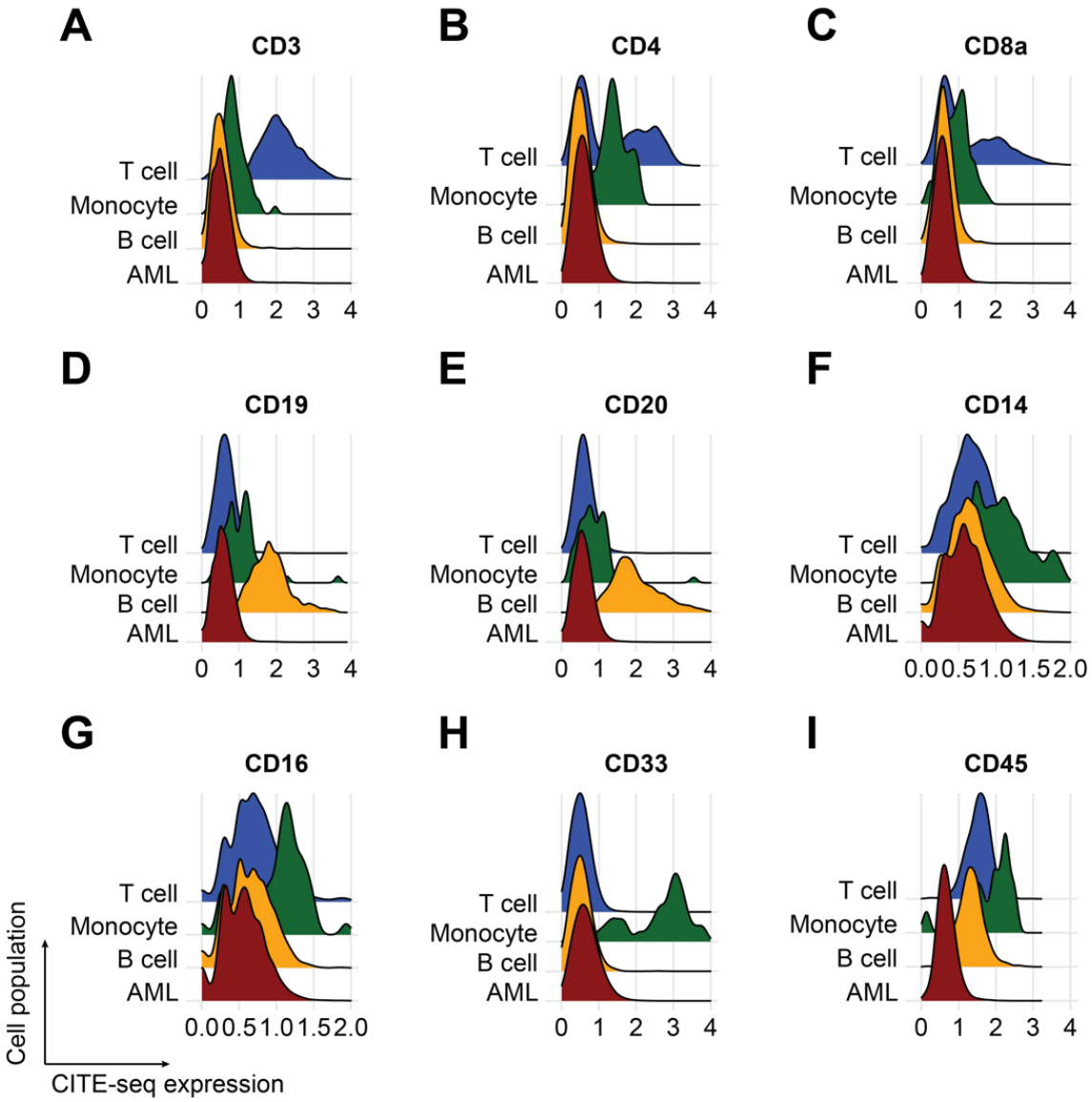

Figure 2 Immunophenotypes of major cell lineages quantified using CITE-seq. A-C T cell markers **D, E** B cell markers **F-H** Monocytic markers **I** Reduced expression of CD45 on AML blasts. CITE-seq expression is shown as normalized counts.

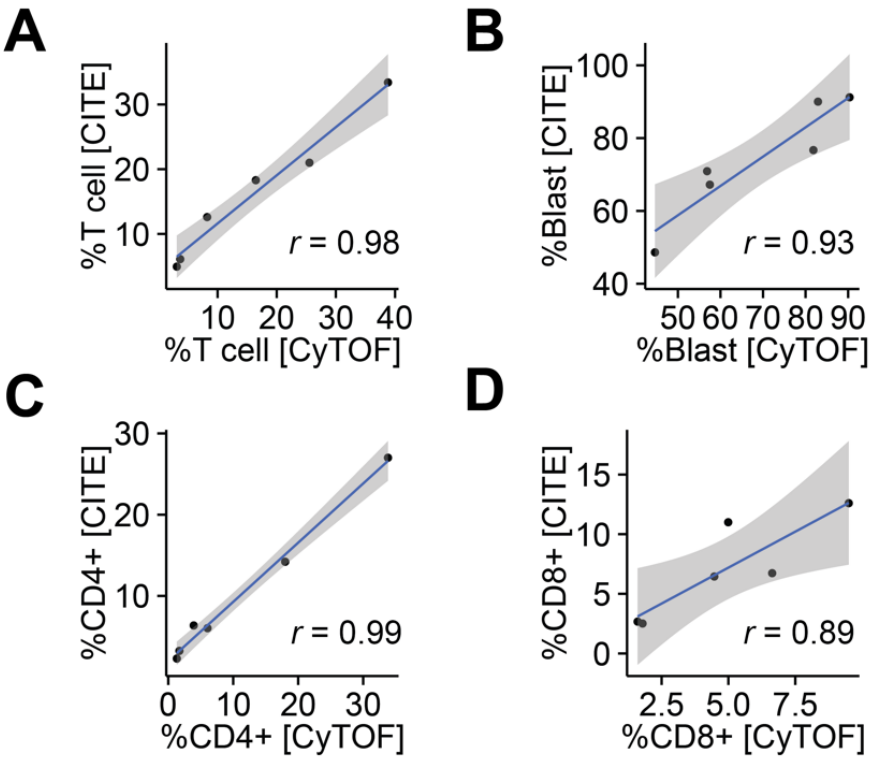

Figure 3 Correlation of cell type quantification with CITE-seq and mass cytometry (CyTOF). A T cells ($CD45^{hi}$ $CD3^{+}$) **B** AML blasts ($CD45^{lo}$ $HLA\text{-}DR^{+}$) **C** $CD4^{+}$ T cells and **D** $CD8^{+}$ T cells. Correlation coefficient calculated using Pearson's product-moment correlation.

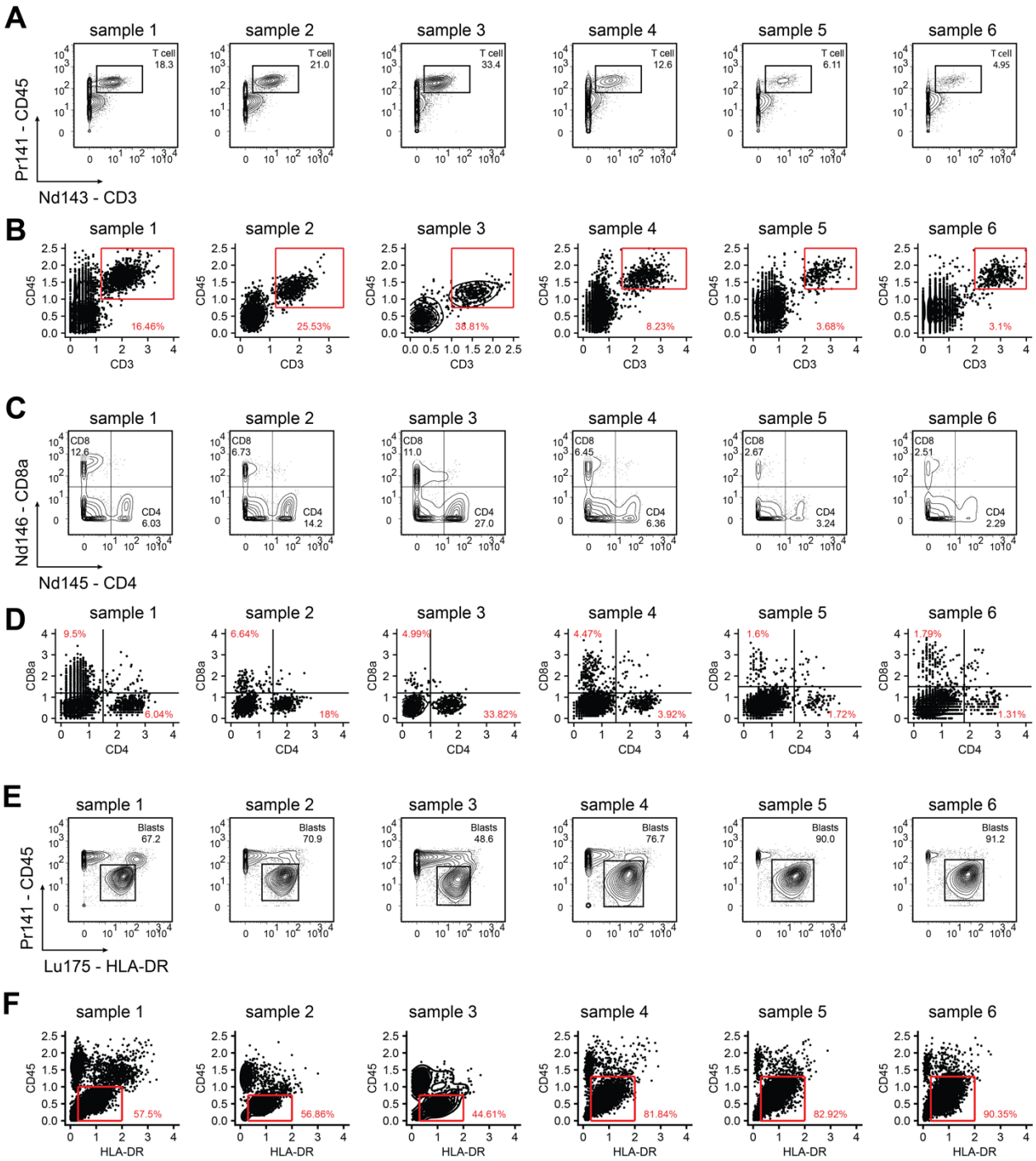

Figure 4 Correlation of cell type quantification with mass cytometry (CyTOF) and CITE-seq – raw data. **A** (CyTOF), **B** (CITE-seq) T cells ($CD45^{hi}$ $CD3^{+}$). **C** (CyTOF), **D** (CITE-seq) $CD4^{+}$ and $CD8^{+}$ T cells. **E** (CyTOF), **F** (CITE-seq) AML blasts ($CD45^{lo}$ $HLA\text{-}DR^{+}$). CyTOF data are represented as metal intensity. CITE-seq data are represented as normalized counts.

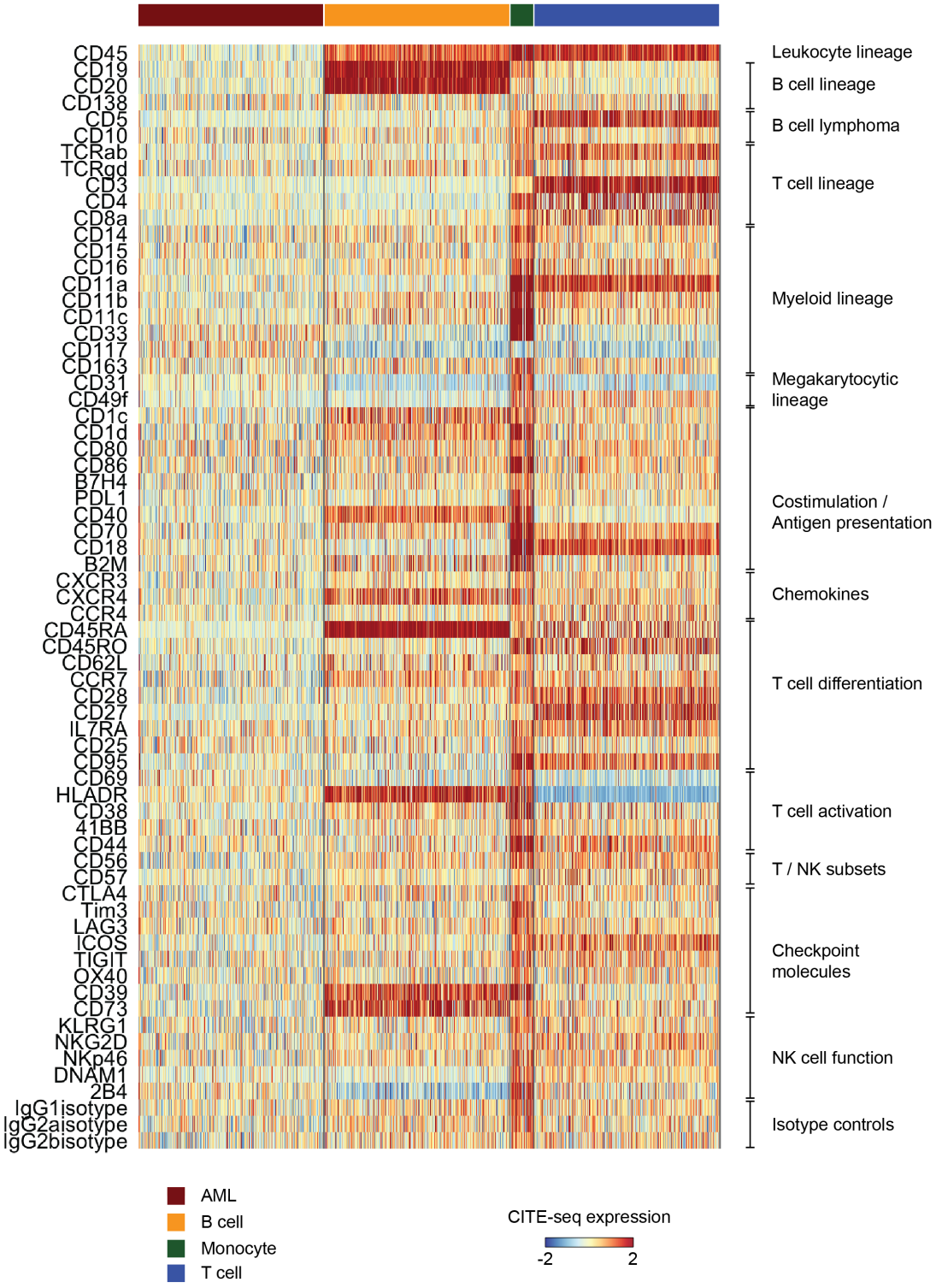

Figure 5 Overview of CITE-seq data by cell types. Shown are 500 single cells for AML, B cell, T cells and all available monocytes (n=61).

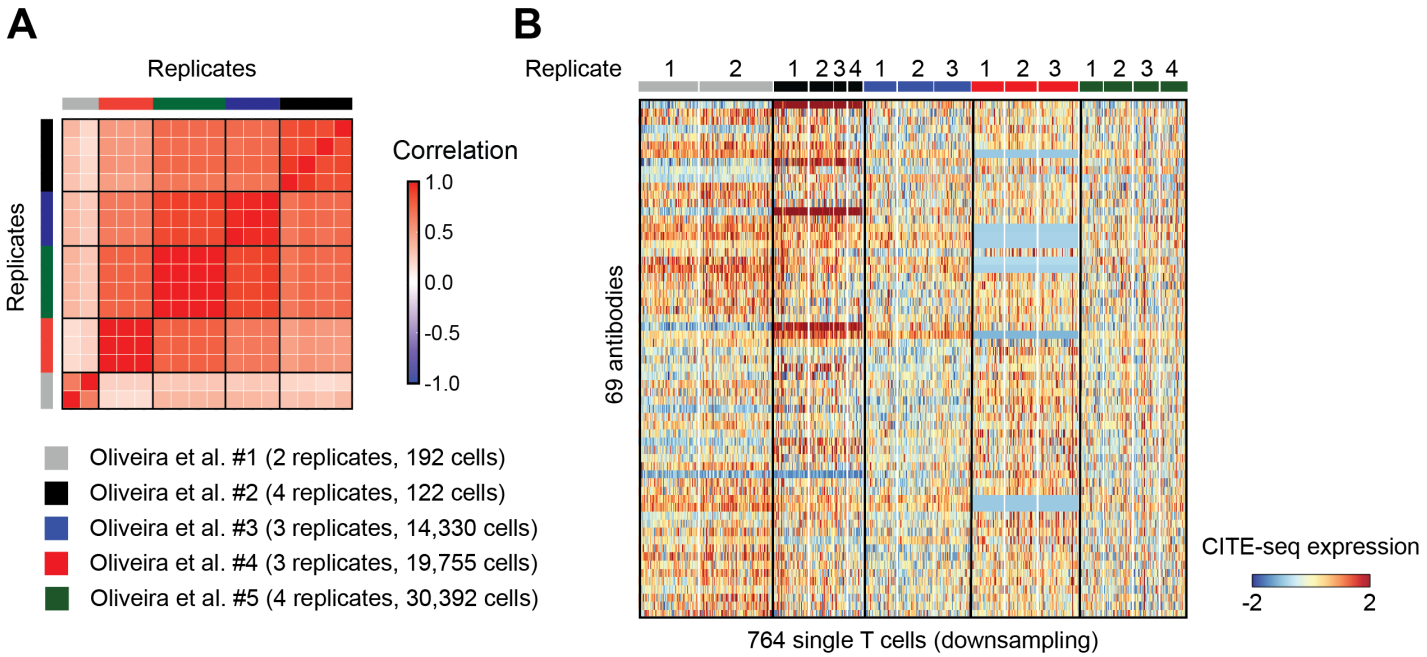

Figure 6 Reproducibility of CITE-seq on multiple replicates. A. Spearman correlation of mean expression of 69 markers between replicates of 5 samples (Oliveira et al., *Nature* 2021) with a total of 64,791 single cell profiles. **B.** Raw data used for calculation of correlation (shown using downsampling to improve visibility of samples with fewer cells).

Table 2. TotalSeq™-C human antibody panel used by TIGL.

Function	Antigen	Clone
Leukocyte lineage	CD45	HI30
B cell lineage	CD19	HIB19
	CD20	2H7
	CD138	DL-101
B cell lymphoma	CD5	UCHT2
	CD10	HI10a
T cell lineage	TCRab	IP26
	TCRgd	B1
	CD3	UCHT1
	CD4	RPA-T4
	CD8a	RPA-T8
Myeloid lineage	CD14	M5E2
	CD15	W6D3
	CD16	3G8
	CD11a	TS2/4
	CD11b	ICRF44
	CD11c	S-HCL-3
	CD33	P67.6
	CD117	104D2
	CD163	GHI/61
Megakaryocytic lineage	CD31	WM59
	CD49f	GoH3
Costimulation / Antigen presentation	CD1c	L161
	CD1d	51.1
	CD80	2D10
	CD86	IT2.2
	B7H4	MIH43
	PDL1	29E.2A3
	CD40	5C3
	CD70	113-16
	CD18	TS1/18
	B2M	2M2
Chemokines	CD183	G025H7
	CD184	12G5
	CD194	L291H4

Function	Antigen	Clone
T cell differentiation	CD45RA	HI100
	CD45RO	UCHL1
	CD62L	DREG-56
	CCR7	G043H7
	CD28	CD28.2
	CD27	O323
	IL7RA	A019D5
	CD25	BC96
	CD95	DX2
T cell activation	CD69	FN50
	HLADR	L243
	CD38	HIT2
	41BB	4B4-1
	CD44	BJ18
T / NK subsets	CD56	QA17A16
	CD57	QA17A04
Checkpoint molecules	CD279	EH12.2H7
	CTLA4	BNI3
	Tim3	F38-2E2
	LAG3	11C3C65
	ICOS	C398.4A
	TIGIT	A15153G
	CD134	Ber-ACT35
	CD39	A1
	CD73	AD2
NK cell function	KLRG1	2F1/KLRG1
	NKG2D	1D11
	NKp46	9E2
	CD226	11A8
	CD244	C1.7
Isotype controls	IgG1isotype	MOPC-21
	IgG2aisotype	MOPC-173
	IgG2bisotype	MPC-11

Overview of 67 TotalSeq™-C antibodies purchased from BioLegend. Antibodies were pooled at standard concentration of 0.5 µg/µl.

Table 3. Quality metrics provided by cellranger.

	PBMC 10x	sample 1	sample 2	sample 3	sample 4	sample 5	sample 6	Median	Standard deviation
Sequencing									
Estimated number of cells	8,258	8,578	4,831	3,597	9,078	9,168	11,426	8,828	2,958
Mean reads per cell	54,262	28,131	40,992	54,029	25,203	29,091	21,561	28,611	12,137
Median genes per cell	1,508	1,257	531	430	1,122	1,053	1,408	1,088	397
Number of Reads	448,096,586	241,311,099	198,032,810	194,343,205	228,794,037	266,705,239	246,360,747	235,052,568	28,405,449
Valid Barcodes	88.1%	91.30%	88.10%	88.30%	89.90%	89.10%	92.30%	89.50%	1.68%
Valid UMIs	not reported	99.90%	99.90%	99.90%	99.90%	99.90%	99.90%	99.90%	0.00%
Sequencing Saturation	82.0%	59.70%	77.70%	76.60%	60.20%	62.50%	53.30%	61.35%	9.90%
Q30 Bases in Barcode	96.4%	96.80%	96.60%	96.70%	96.70%	96.70%	96.70%	96.70%	0.06%
Q30 Bases in RNA Read	92.0%	87.80%	85.50%	83.90%	88.20%	88.00%	89.30%	87.90%	2.01%
Q30 Bases in Sample Index	93.1%	88.90%	96.00%	94.90%	93.60%	93.70%	95.00%	94.30%	2.51%
Q30 Bases in UMI	96.3%	96.40%	96.40%	96.60%	96.40%	96.40%	96.40%	96.40%	0.08%
Total Genes Detected	20,875	20,077	19,327	17,924	20,257	20,320	20,353	20,167	955
Median UMI Counts per Cell	4,699	5,862	1,330	1,332	3,884	3,483	5,880	3,684	2,035
Reads Mapped to Genome	91.8%	85.00%	77.00%	70.50%	84.80%	84.60%	87.70%	84.70%	6.52%
Reads Mapped Confidently to Genome	79.5%	73.90%	64.60%	60.30%	74.10%	73.20%	80.70%	73.55%	7.38%
Antibody Sequencing									
Number of Reads	133,509,587	67,740,895	58,433,701	66,434,813	65,068,538	58,140,419	43,364,409	61,751,120	9,050,539
Valid Barcodes	98.2%	98.50%	98.30%	98.30%	98.30%	98.30%	98.30%	98.30%	0.08%
Valid UMIs	not reported	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0.00%
Sequencing Saturation	49.5%	50.80%	26.30%	34.40%	27.20%	29.20%	22.60%	28.20%	10.10%
Q30 Bases in Barcode	96.5%	96.60%	96.70%	96.60%	96.60%	96.60%	96.60%	96.60%	0.04%
Q30 Bases in Antibody Read	94.2%	96.70%	96.50%	96.60%	96.50%	96.60%	96.60%	96.60%	0.08%
Q30 Bases in Sample Index	93.9%	94.50%	94.80%	94.50%	95.70%	94.30%	90.20%	94.50%	1.93%
Q30 Bases in UMI	95.9%	96.10%	96.00%	96.00%	96.00%	96.10%	95.90%	96.00%	0.08%
Antibody Application									
Fraction Antibody Reads	97.0%	97.00%	97.00%	97.10%	97.00%	97.00%	97.00%	97.00%	0.04%
Fraction Antibody Reads Usable*	59.9%	17.60%	18.50%	11.00%	24.90%	26.50%	36.90%	21.70%	8.96%
Antibody Reads Usable per Cell*	9,679	1,386	2,233	2,039	1,787	1,681	1,399	1,734	340
Fraction Reads in Barcodes with High UMI Counts	1.5%	41.50%	7.10%	17.80%	8.80%	9.20%	3.30%	9.00%	14.00%
Fraction Unrecognized Antibody	3.0%	3.00%	3.00%	2.90%	3.00%	3.00%	3.00%	3.00%	0.04%

Antibody Reads in Cells*	62.9%	18.50%	19.40%	11.70%	26.20%	27.90%	38.90%	22.80%	9.43%
Median UMIs per Cell	3,996	840	1,404	1,082	1,099	1,090	945	1,086	190

*Depends on total antibody, number of antibodies and sequencing depth.

As comparator the PBMC dataset from 10x Genomics (vdj_v1_hs_pbmc2_5gex_protein) is shown. Of note, the PBMC dataset was sequenced with greater sequencing depth.

3. References

1. Stoeckius M, Hafemeister C, Stephenson W, et al. Large-scale simultaneous measurement of epitopes and transcriptomes in single cells. *Nat Methods*. 2017;14(9):865–868.
2. Penter L, Gohil SH, Huang T, et al. Coevolving JAK2V617F+ relapsed AML and donor T cells with PD-1 blockade after stem cell transplantation: an index case. *Blood Adv*. 2021; bloodadvances.2021004335.
3. Oliveira G, Stromhaug K, Klaeger S, et al. Phenotype, specificity and avidity of antitumour CD8+ T cells in melanoma. *Nature*. 2021;596(7870):119–125.
4. RStudio Team. RStudio: Integrated Development Environment for R. Boston, MA: RStudio, PBC.; 2020.
5. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21.
6. Cossarizza A, Chang H-D, Radbruch A, et al. Guidelines for the use of flow cytometry and cell sorting in immunological studies (second edition). *Eur J Immunol*. 2019;49(10):1457–1973.