Cancer Immune Monitoring and Analysis Center
Precision Immunology Institute
Icahn School of Medicine at Mount Sinai (ISMMS)
Contact PI: Sacha Gnjatic, PhD (Sacha.Gnjatic@mssm.edu)

Performance Lab:
Human Immune Monitoring Center (HIMC)
Miriam Merad, MD, PhD, Director
Sacha Gnjatic, PhD, Professor, Co-Director
Seunghee Kim-Schulze, PhD, Facility Director
1470 Madison Avenue, New York, NY 10028

Assay type	Primary assay outputs	Pre-processing/ Normalization /QC	Initial analyses	Derived data outputs
scRNAseq	~20,000+ gene quantifications and ~5,000 cells per sample.	HIMC Sample QC Sequencing QC Analysis QC	Cell type identification, differential expression, molecular pathways and pseudotime.	Gene expression per cluster/cell type per sample.

1. Purpose of assay

High-dimensional single cell monitoring is a key strategy to elucidate complex phenotypic and functional characteristics of heterogeneous immune populations. This is a broadly applicable approach that can provide valuable insights into disease mechanisms and therapeutic responses and potentially identify correlative cellular biomarkers in the context of many different trials. Some of the key challenges in maximizing the impact of cellular immune monitoring are to increase the number of parameters that can be examined simultaneously in a single sample while minimizing experimental and technical variability. The Mount Sinai Cancer Immune Monitoring and Analysis Center (MS-CIMAC) has established a robust single-cell RNA sequencing (scRNAseq) platform addressing these challenges using the 10x Genomics Chromium scRNAseq workflow.

Droplet-based scRNAseq methods use microfluidics to partition a single-cell suspension into individual droplets, each of which contains a single cell and a single Gel Bead containing a set of barcoded reverse-transcription primers. The most widely adopted scRNAseq platforms all take advantage of unique molecular identifiers (UMI), a strategy for barcoding individual molecules of mRNA at the point of reverse transcription. Ultimately, this strategy allows for thousands of single-cell mRNA sequencing libraries per sample to be prepared in a single tube, where each library construct contains barcodes that indicate both the cell and individual mRNA molecule of origin.

The MS-CIMAC has invested heavily to establish a comprehensive scRNAseq pipeline led by Dr. Seunghee Kim-Schulze. It includes optimized sample processing SOPs and an efficient data processing pipeline. As one of the most active clinically-focused scRNAseq programs in the country, we routinely apply scRNAseq in a range of applications, most importantly in the characterization of rare clinical samples.

2. Manufacturer's Validation

The MS-CIMAC has engaged in extensive validation and use of the 10x Genomics platform for single-cell sequencing of RNA and TCR/BCR. However, because the protocol is well established by the manufacturer, a lot of information about specimen collection tubes and assay performance has already been queried and published as part of the 10x platform technical notes.

Document CG00052 tests the reproducibility of 10x scRNAseq across flow cells, using mice neuronal tissue replicates prepared and dissociated on separate days, sequenced on separate flow cells, or separate chips. The results showed biological variation exceeding all investigated sources of technical biases, with Pearson correlations consistently greater than 0.93. An example is provided below:

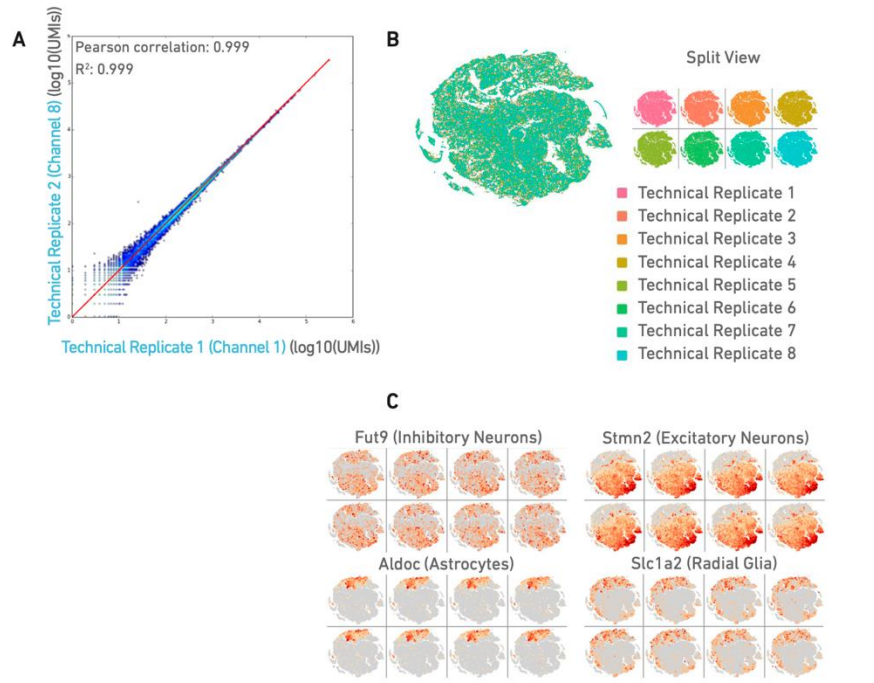

Figure 1. Gene expression correlation when the same sample is run on different input wells of the same chip. **A.** Scatter plot showing the correlation of average gene expression levels between wells 1 and 8 on a log10 scale. **B.** Merged (left panel) and split (right panel) tSNE clustering for 8 technical replicates. **C.** Expression profiles for known neuronal marker gene Fut9 (inhibitory neurons), Stmn2 (Excitatory neurons), Aldoc (Astrocytes), and Slc1a2 (radial glia).

In our own validation below, we performed same sample comparisons using PBMC, bone marrow mononuclear cells, and/or tumor tissue specimens, with similar results **(Fig. 1)**.

In addition, a variety of PMBC isolation methods were tested by 10x and were shown to perform with equal sensitivity for both 3' and 5' libraries, which represent the two main chemistries available. The figure below summarizes the findings **(Fig. 2).** To compare sensitivity, median genes and UMIs were plotted against sequenced raw read pairs. Median genes and UMIs captured across different anticoagulants and across replicates were comparable for both Single Cell 3' v3.1 and Single Cell 5' v2 libraries. At 20,000 read pairs per cell, both assays reached ~50% sequence saturation. UMAP plots for PBMCs collected with different anticoagulants showed significant overlap as evidenced by the consistent cluster structure across both Single Cell 3' v3.1 and Single Cell 5' v2 assays **(Fig. 3)**.

These results indicate that the 10x Chromium platform has a robust internal validation process for allowing new kits to be issued, and we therefore embarked on our own analytical validation with the expectation of reproducibility based on the vendor's characteristics.

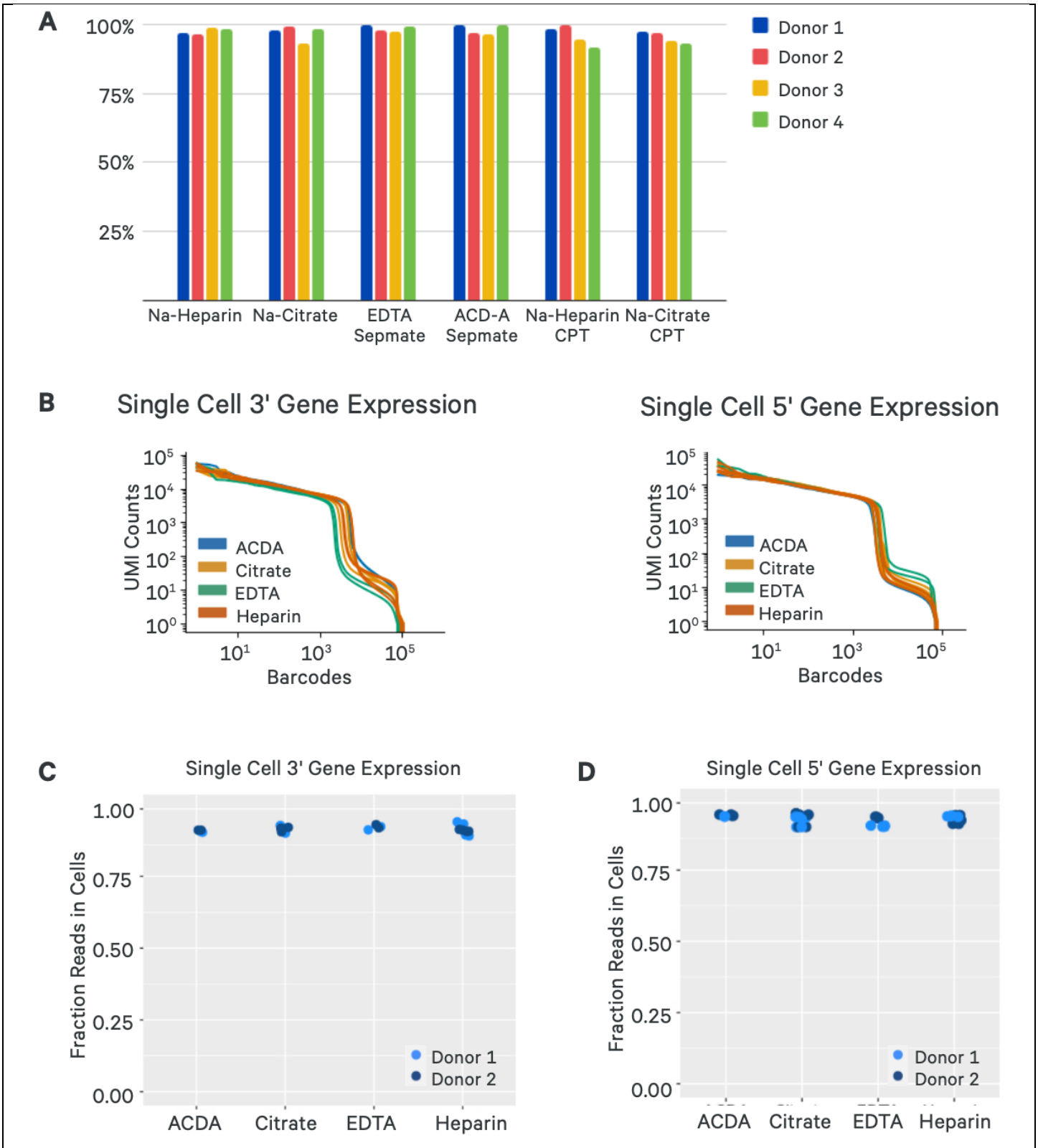

Figure 2. A. Cell viability and single cell behavior data for single cell 3' and 5' libraries. **B.** Rank plots contain data from both library types. **C&D.** Fraction reads in cells plots are for single cell 3' and 5' libraries, respectively.

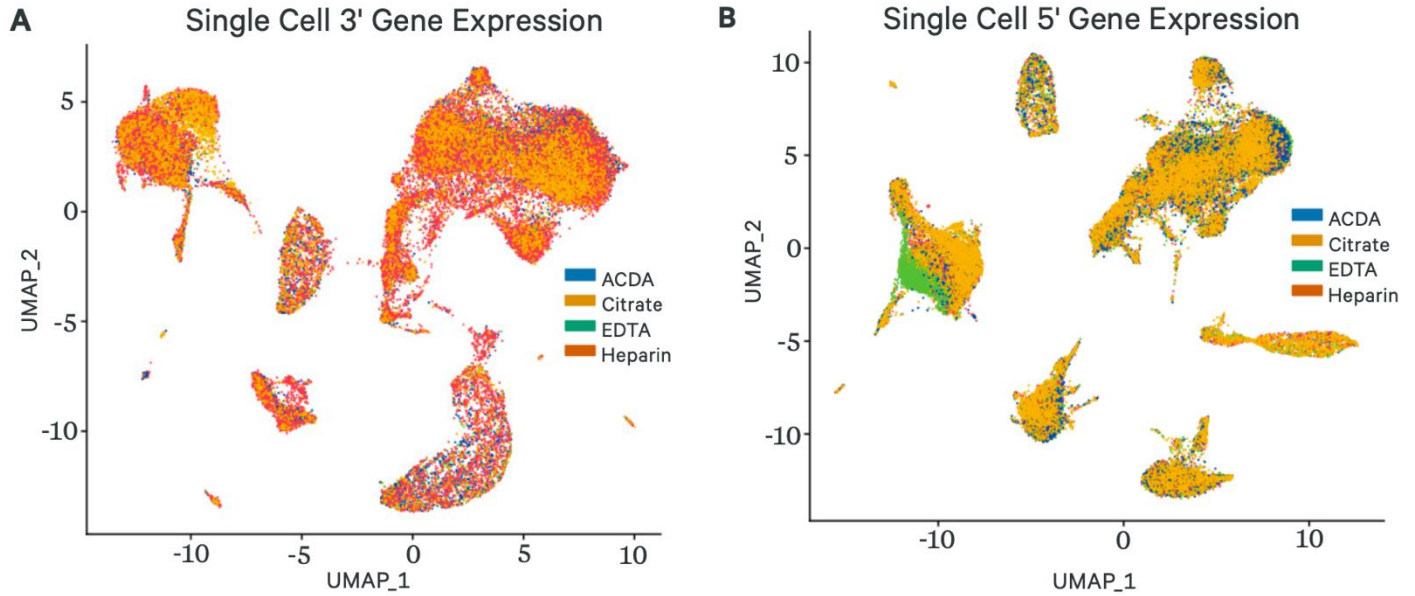

Figure 3. Dimensionality reduction UMAP showing the similarities between 3' and 5' assay in A and B, respectively.

3. Icahn School of Medicine at Mount Sinai Human Immune Monitoring Center Experience and Capacity

HIMC has been handling samples, quality controls, and library preparations since 2017. During this time, the number of samples has grown exponentially to satisfy the sequencing demand of the department of Oncological Sciences at Mount Sinai and both internal and external collaborators **(Fig 4A)**. These samples are represented mostly by human origin but also include other species such as mouse, zebrafish and ferret **(Fig 4B).** Further, these samples include a plethora of tissues including sorted cell types and cell lines, as well as different organs, compartments and whether the sample corresponds to tumor or others. **(Figs 5-7)**. This data indicates adequate capacity at Mount Sinai for this type of work, and applicability to a variety of clinical specimens.

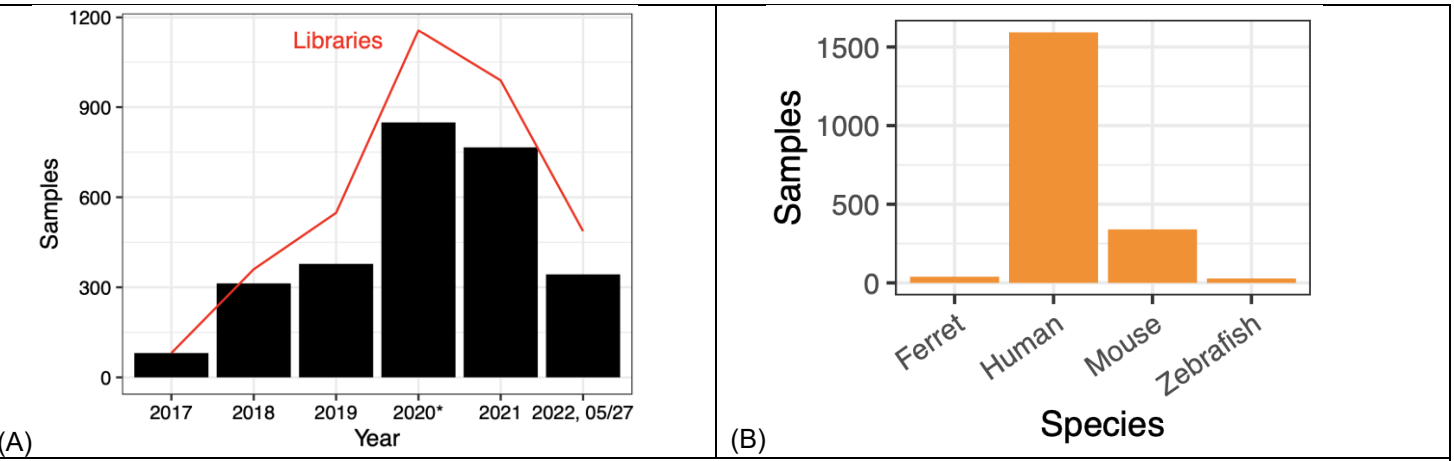

Figure 4. Samples and Libraries per year. A. Bar and line plot showing the number of samples sequenced since 2017. Processed at HIMC. In a red line, the number of libraries prepared. The x-axis shows the year and y-axis the numbers for both libraries and samples. B. Barplot showing the number of samples per species sequenced at HIMC.

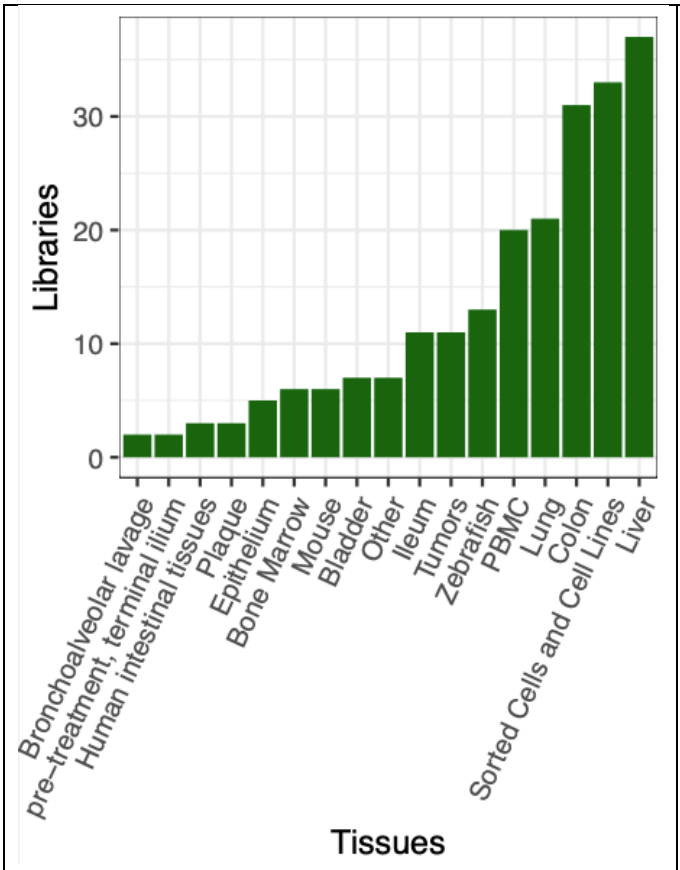

Figure. 5. Barplots showing the number of tissues sequenced at HIMC.

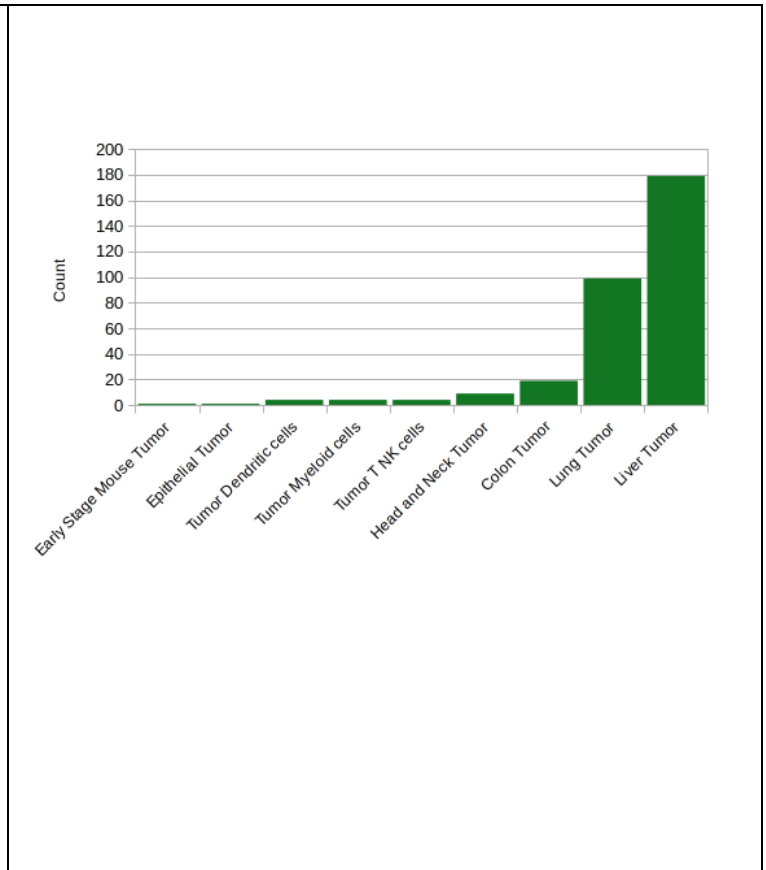

Figure. 6. Tumor samples sequenced at HIMC

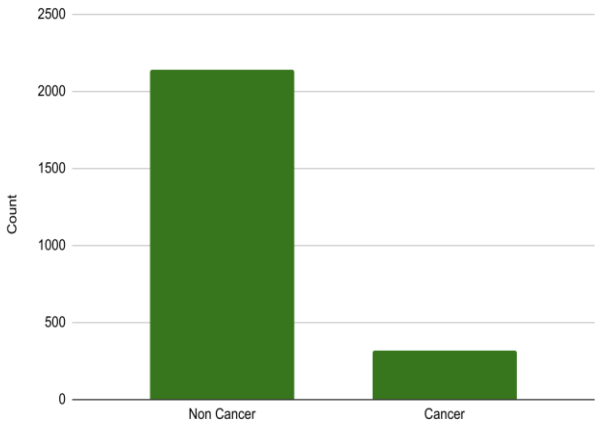

Figure 7. Cancer samples sequenced at HIMC

4. Assays performance characteristics based on Mt Sinai CIMAC's internal validation

Single cell RNAseq characterization	
(i) accuracy	Single cell RNAseq has a high accuracy of identifying cell types and sub-phenotypes. Studies that also perform CyTOF in parallel validate the cell type proportions identified by scRNAseq with 90% accuracy or more.
(ii) precision	The precision of intra-assay assays for sequencing is commonly >90%. Our cross-site validation comparing same sample sequenced at 3 different sites revealed a concordance and precision of >95%. (Fig. 14). This calculation can also be considered intra-batch with a standard variation <5%.
(iii) analytical sensitivity	The limit of single cell gene expression detection is estimated to be approximately 1000-3500 molecules per cell, though this also depends on the sequencing depth. Common quality control metrics during our experiments include cell count, gene count, gene number and mitochondrial content (Fig. 14). Assay sensitivity in terms of detection of rare cell populations is estimated to be populations of around 50 cells in every 5,000 with power of 95%. We also showed high analytical sensitivity across 3 different chemistries V1(5`), V2(3`), and V3(3`) with above a correlation of above 90% (Figure Chem).
(vii) standardization, harmonization, reproducibility and ruggedness	We have developed numerous QC and calibrator steps to improve standardization and to facilitate standardized outputs using 10x Genomics technologies. These can be applied to a variety of dissociated tissues and whole blood when QC is sufficient. All samples are barcoded and multiplexed when necessary to reduce costs. Validation results from cross-site comparison highlight technical and biological correlates of >95% with P<0.01 (Figure 12-14)
(ix) any other performance characteristics required for assay performance	A limitation of our scRNAseq is the sample quality may not be optimal for low cell input and low cell viability. Optimal cell concentration, volume, and viability per lane of the scRNAseq assay is 1M cells/mL in at least 50ul with a viability of >80%. The assay performs best with freshly prepared cells, but there is a cell fixation protocol from 10x Genomics (Single Cell Fixed RNA Sample Preparation Kit, CAT# 1000414) that can be used to preserve single cell isolates for long term storage before loading on the scRNAseq assay, currently under evaluation.

TECHNICAL VALIDATION

General materials and methods for assay characterization and validation. Assay validation was performed using NSCLC resection specimens. Following physical and enzymatic dissociation, samples were enriched for immune cells using magnetic beads, prior to microfluidic single-cell encapsulation, reverse transcription, and downstream processing.

Patient ID	Tissue type	# of 10x channels	Cell Viability	# of cells above UMI threshold	% of mito hi events (>20%) above UMI threshold	Median #UMI of cells above threshold	Median #genes of cells above threshold
R1027	Tumor	5	61.5%	21,344	34% (7,314)	3,490	1,205
R1027	Normal	5	69.4%	13,266	22% (2,928)	3,192	1,236
R1027	PBMC	5	96.8%	10,163	4% (482)	3,909	1,403
R1043	Tumor	5	75.9%	16,285	8% (1,329)	2,867	1,041
R1043	Normal	5	67.9%	12,683	13.5% (1,715)	3,757	1,362
R1043	PBMC	5	96.6%	7,315	2% (155)	4,033	1,313

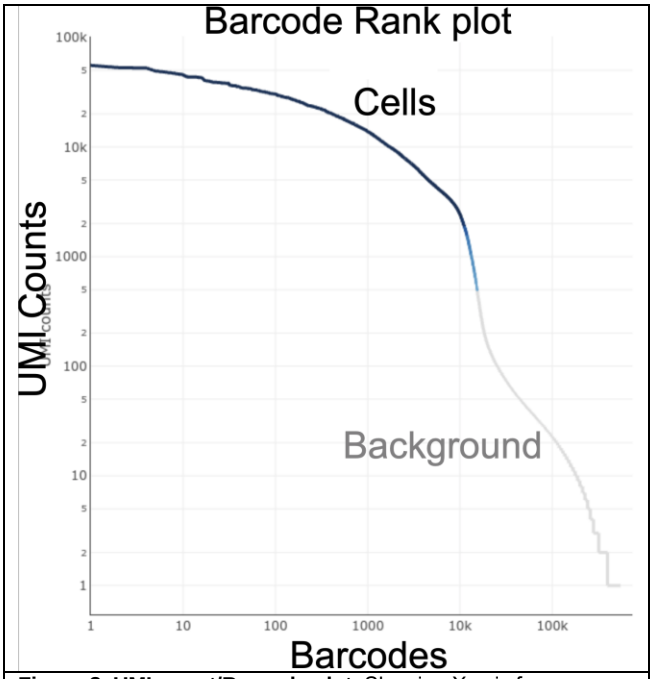

Figure 8. UMI count/Barcode plot. Showing X axis for barcodes and Y axis for UMI counts.

The estimation for cell quality post sequencing is done using cell ranger with default settings. The UMI count/Barcode plots are inspected manually in order to validate the amplification process. The cutoff is identified by looking for the breaking point (elbow) between cells vs background. This cutoff is usually between 5000 and 100 UMI counts. This UMI threshold is determined by plotting barcodes in decreasing order of the number of UMIs associated with that particular barcode. A steep drop-off is indicative of good separation between the cell-associated barcodes and the barcodes associated with empty GEMs. **(Fig. 8).**

The quality controls for the quantification of cDNA and final gene expression library product are analyzed using the Agilent Bioanalyzer or Agilent Tapestation assays **(Fig. 9A&B).** These allow to quantify and identify the length of the pre and post amplification genetic material and visually inspect the cDNA/RNA profiles pre and post library prep, serving a critical role into ensuring correct genetic material is used. A negative control of a sample that fails to produce useful cDNA or library, is shown in **Fig. 9C**.

Table 2 below summarizes the QC and validation steps implemented at all steps of the scRNAseq assay and analysis pipeline, from sample processing, encapsulation, library preparation, sequencing, data processing and analysis.

Table 2

Sample processing	Library prep and sequencing	Data processing and analysis
Strict SOPs and detailed sample processing logs	qPCR validation of library contents and quality	Cellranger pipeline for read mapping and cell-barcode demultiplexing
Optimized tissue digestion protocols	Bioanalyzer/Tapestation validation of insert size (Figure 9)	Filtering of low UMI events
Optimized cell counting and loading procedure	Optimized sequencing protocol for scRNAseq libraries	Filtering of high mitochondrial RNA events

4.1. Initial Sample Quality Control (QC)

4.1.1. Types of biological samples accepted: The 10x Genomics single cell sequencing assay solution is compatible with polyadenylated mRNA molecules from eukaryotic cells.

4.1.2. Cell input requirements: The concentration, volume, and viability of the samples is very important for the success of the 10x Genomics single cell sequencing assay. The concentration of cells will ideally be between 0.5 million cells/mL and 2.0 million cells/mL, in a volume of at least 100uL for each lane of the 10x Genomics assay. For best results, the cell viability should be above 90%, but a viability as low as 75% is acceptable. Viability below 75% should undergo dead cell depletion in order to increase cell viability.

4.1.3. Other cell quality considerations: There are a number of factors besides cell concentration and viability that can impact the single cell assay. If there is a significant amount of debris in the sample, it is recommended that the sample be sequentially filtered through a 70um and 40um filter. Additionally, if high RBC content is observed, the sample should be treated with a RBC-lysis solution prior to loading on the 10x assay.

4.1.4. Recommended cell suspension buffer: The recommended cell resuspension solution is 1X PBS (calcium and magnesium free) containing 0.04% weight/volume BSA (400 µg/ml). BSA is added to minimize cell losses and aggregation.

4.2. Sample Multiplexing

4.2.1. Reasons to multiplex samples during single cell sequencing: Sample multiplexing is useful in cases where there are more samples than are able to fit within the 8 lanes of the 10x Genomics assay. Up to 12 samples can be stained with oligo-conjugated lipids or antibodies, washed, and pooled together in equal concentrations. This pool is then loaded on the 10x Genomics assay.

4.2.2. Sample considerations when multiplexing: The same sample preparation and quality recommendations are present as described in section 4.1, except that more cells are required for the staining and washing steps of the multiplexing assay. Due to the sample washing steps requiring the cells to be pelleted and re-suspended in wash buffer, it is recommended that each sample contain a minimum of 100,000 live cells at the start of the assay.

4.2.3. Recommended cell wash buffer: During the washing steps, it is recommended that 1X PBS (calcium and magnesium free) containing 1.0% weight/volume BSA (400 µg/ml) is used.

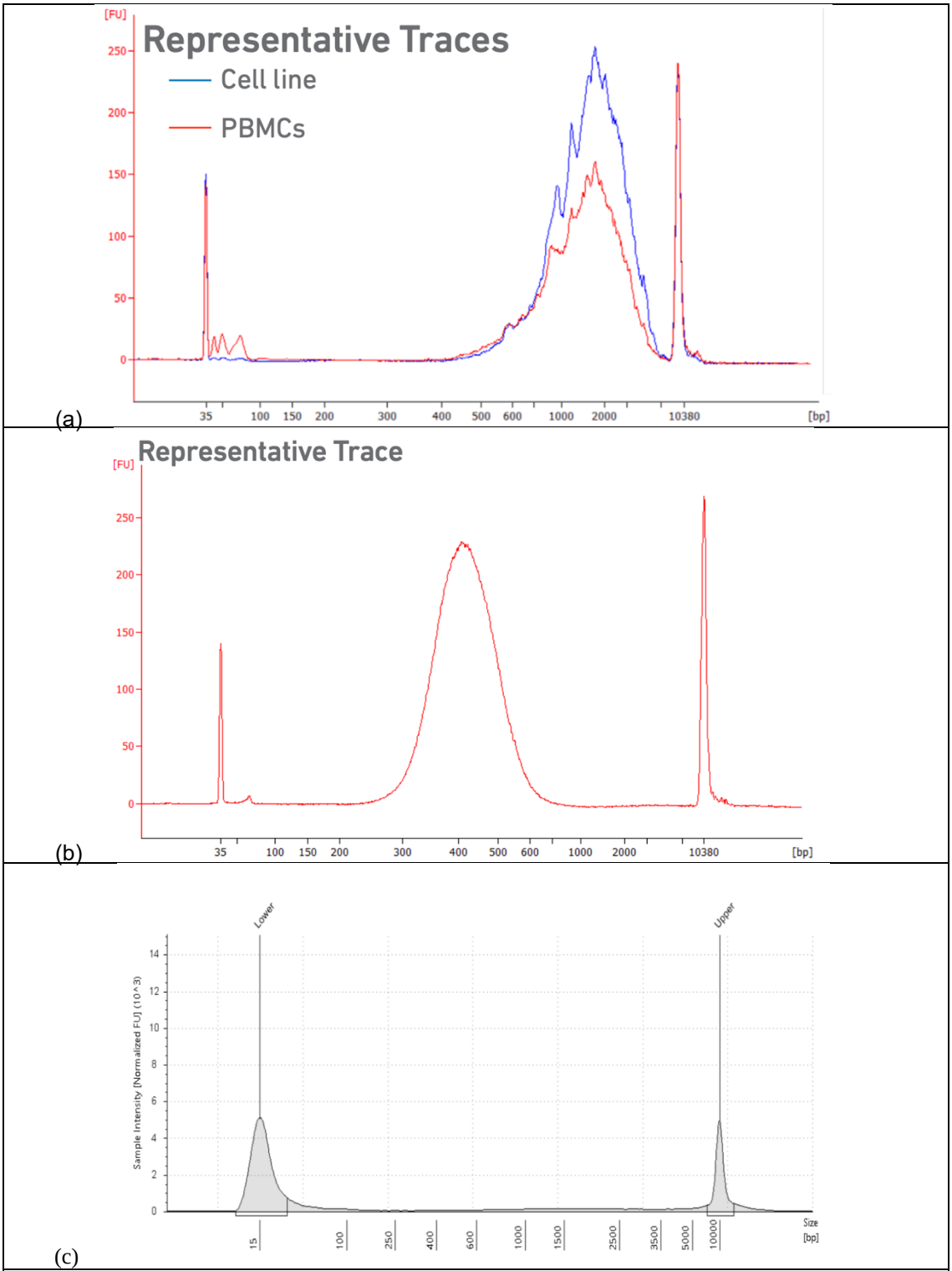

Figure 9: Quantification of cDNA (a) and final gene expression library product (b) using the Agilent Bioanalyzer or Agilent Tapestation assays. When a sample fails to produce useful cDNA or library, the tapestation result can be seen in (c).

4.3. Validation of reproducibility across flow cells

To validate flowcell reproducibility within a sequencer, liver tissue from one human was used to prepare a single cell suspension. A single library was generated by loading the cell suspension on a microfluidic chip. The library was sequenced on two separate flowcells (Flowcell 1 and 2) on the same NovaSeq S1 sequencer. A perfect correlation was observed (Rho=1, R^2=1, P<0.05, **Fig. 10A and Fig10B**). To validate lane variability within a 10x chip **(**, liver tissue from one human was used to prepare a single cell suspension. In summary, two libraries were generated by loading the same cell suspension in two lanes (TENX LANE 1 and TENX_LANE_2) of the chip. Both libraries were sequenced on one flowcell and a near perfect correlation was observed (Rho=0.97, R^2=0.97, P<0.05, **Fig. 10C and Fig10D**).

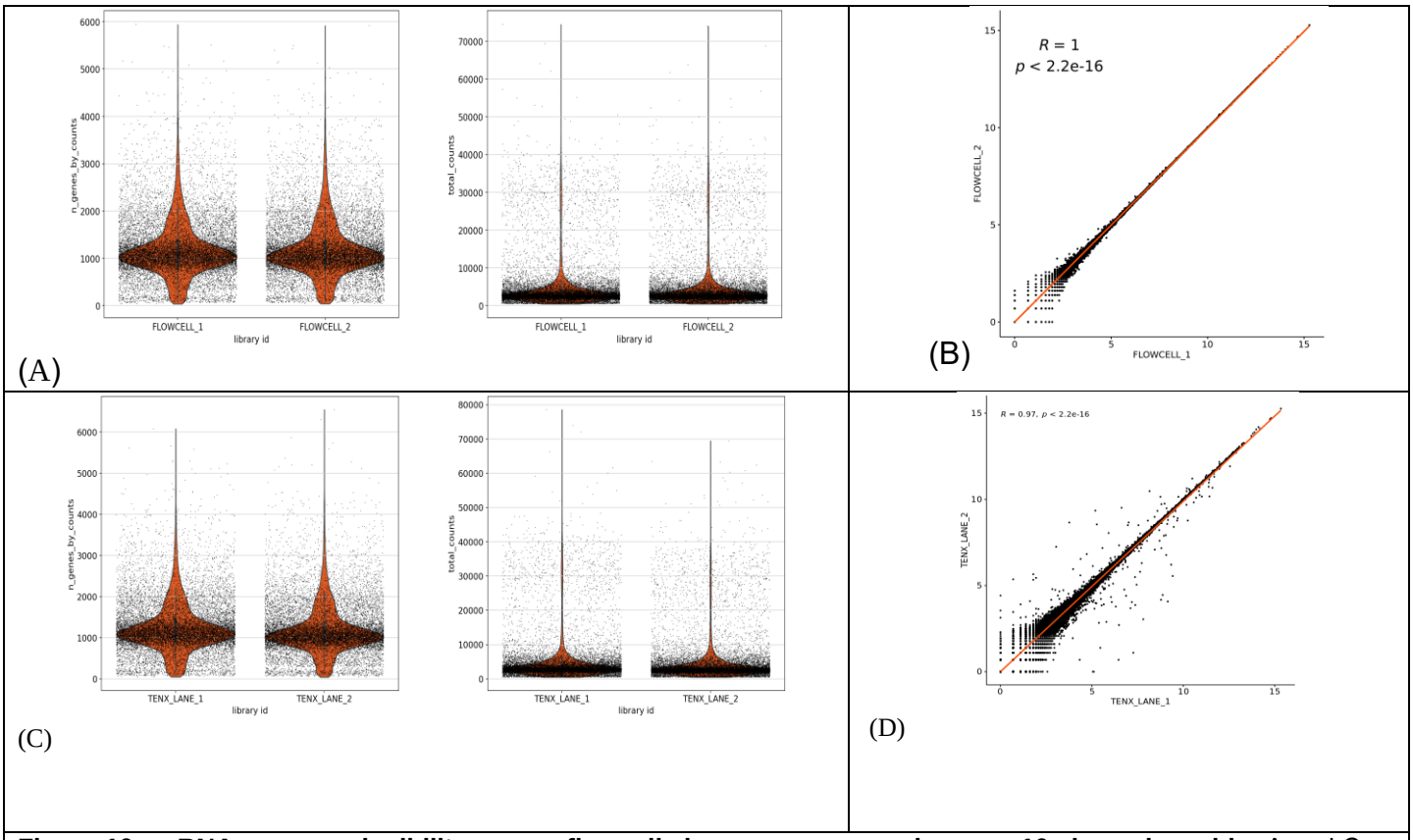

Figure 10. scRNAseq reproducibility across flowcells in a sequencer and across 10x lanes in a chip. A and C Distributions of the total number of genes (left) and total number of counts (right). B and D Scatter plot showing the correlation of average gene expression levels on a log10 scale.

4.5 Comparison across 10x Chromium scRNAseq chemistries

To validate that sequencing chemistries did not have a major impact in data quality and outcomes, we compared standard quality control metrics across 3 different chemistries (V1(5'), V2(3'), V3(3')) in an HCC cohort constituted by ~1000 samples. These metrics included gene counts, gene numbers and mitochondrial gene (Mt) quantification **(Fig. 11A-C)**. An in-depth review of the quality control metric distributions revealed that V1 and V3 generates a larger gene count and number per cell compared to V2 **(Fig. 11A&B).** However, V3 also showed a

larger percent of Mt genes compared to V2 and V1 **(Fig 11C).** In contrast, the data distribution of V1 was similar to V3 compared to V3 **(Fig. 11D)**. Further, we investigated the percent of gene variance attributed to chemistry and identified that chemistry have a quantifiable effect explaining approximately 25% of the total variation. As commonly observed by multiple published studies[1], most of the variance is explained by patient variability and smaller proportion is attributed to treatment and tissue types (<10%). Finally, we compared the correlation of the same samples ran on all 3 chemistries an found an average linear and non-linear correlation of ~0.9 **(Fig 11FGH).** These results show that despite variation due sequencing chemistry, outcomes or RNA quantifications per gene did not have a significant impact in data quality or gene quantifications. The current standard for sequencing is the 5' chemistry, which allows to perform scTCRseq on top of scRNAseq.

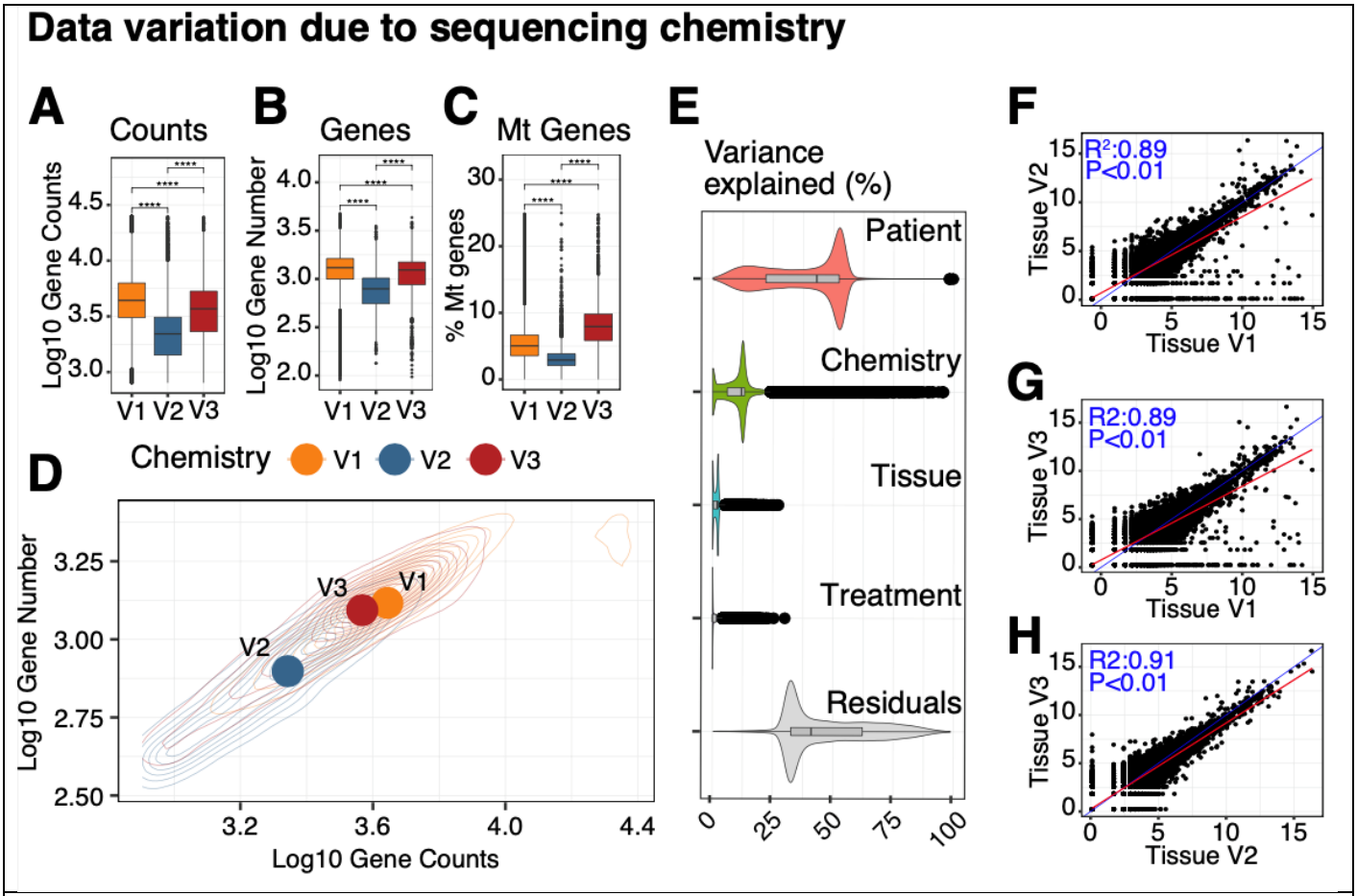

Figure 11. scRNAseq reproducibility across sequencing chemistry. Quality control metrics, indicative of mitochondrial genes, gene number and count are shown in boxplots colored by chemistry (A, B & C). D. Two-dimensional plot of the gene number and gene counts per chemistry. The x and y axis are in log10 scale. E. Variance explained by patient, chemistry, tissue, treatment or unexplained (Residuals) represented by boxplots. Each point represents a single gene. The cumulative variance across covariates for each gene is 1. Correlations of the expression across each chemistry for the same sample are shown in F, G and H, as scatterplots. A linear regression squared coefficient is shown in blue. The red line indicates linear regression and highlights the deviation from the abline shown in blue.

ANALYTICAL VALIDATION

4.6 Reproducibility of gene-gene correlation structure and biological correlates

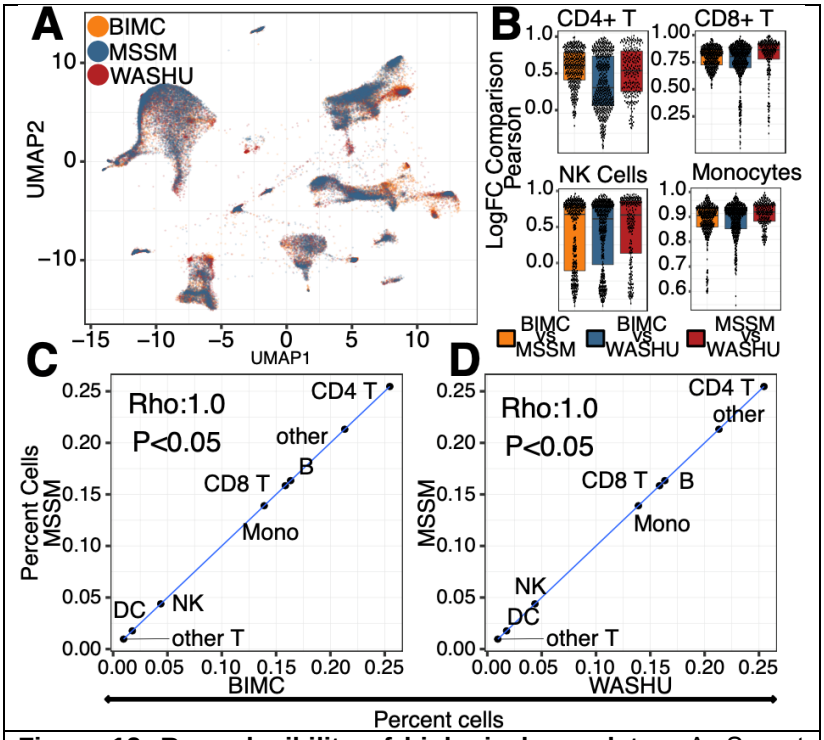

Figure 12. Reproducibility of biological correlates. A. Seurat Default UMAP based on ~ 30 Multiple Myeloma patient samples after reference mapping to PBMCs. B. Pearson correlation of the log fold change signature per cell type across three centers. C & D. Pearson correlation of percent of cells per cell type across three centers.

To evaluate the reproducibility across-site, we used a standardized algorithm for gene structure mapping implemented on Seurat V4, known as reference mapping/Azimuth[2], to map the single cell gene expression to a reference database. Then, we compared the mapping results of ~200 CD138-depleted bone marrow mononuclear cell (BMMC) samples across cell types and gene expression profiles, performed in three separate sites, as part of a network Immune Atlas study sponsored by the Multiple Myeloma Research Foundation involving Mount Sinai (ISMMS), Beth Israel Medical Center (BIMC), and Washington University (WashU). The dimensionality reduction PCA-UMAP and clustering based on the top 2500 most variable genes showed that were was not a significant difference between mapped samples per cluster **(Fig. 12A).** All 3 sites were identified in all clusters and each cluster was associated to a single cell type. However, to compare the gene signatures associated with each cluster we performed differential expression of each cell type specific cluster compared to the rest of cells in that sample to obtain a log fold change profile of genes per cluster. Then, the differential expression profiles per cluster per sample were correlated to each other and across centers, showing a range of variation between 0.2 to 1.0 with a mean of ~0.7 for CD4+ T and NK cells, ~0.85 for CD8+ T cells, ~0.9 for Monocytes **(Fig. 12B).** These results show that at minimum 70% of differentially expressed genes in a cell type are consistent between sites in average across ~200 samples. Next, we asked whether the cell type proportions were consistent between sites and correlated the percent of cell types for 18 samples (6 patients). The results indicate that the percent of cells identified between sites is extremely reproducible (Rho:1.0, P<0.05) despite differences in gene expression. Together, these results a clear correlation in gene-gene, differential expression and cell-type – cell type signatures between sites **(Figs. 12 and 13).** The major cell types were identified using refence mapping indicate high heterogeneity between patient samples **(Fig. 13)**, which can be attributed to sample processing (technical artifacts) or real variation dependent on biological factors (treatments, cancer degree, etc). However, cell type detection was sensitive enough to detect populations larger than 1% total cells in sample and others smaller <1% but with less confidence **(Fig. 13)**. Additionally, HIMC is currently developing new cell type identification methods and expanding single cell databases to refine the cell specific gene signatures in this assay.

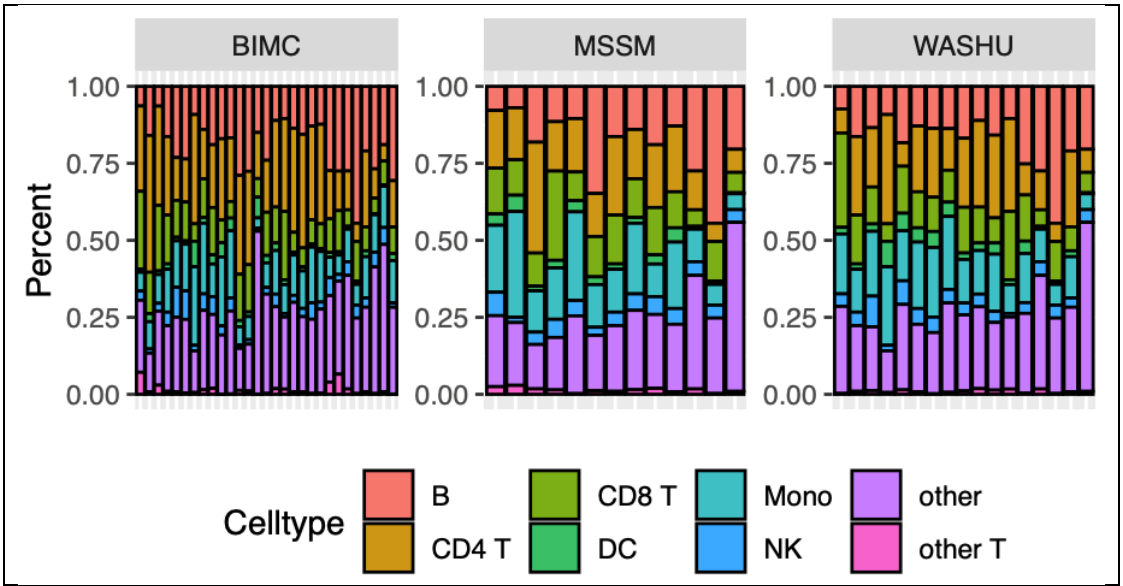

Figure. 13. Cell type proportions identified across sites. Barplot showing the stacked percent of cell types per sample. The colors indicate the cell type.

Further QC can be performed on the count data directly using publicly available algorithms to remove ambient gene expression or sequencing noise. These terms refer to counts that do not originate from a barcoded cell, but from other lysed cells or mRNA contaminants that may be included in the droplet-cell suspension prior to library construction. These added ambient counts can affect downstream analysis such as marker gene identification or other differential expression tests especially when levels vary between samples. These effects could explain the 70% accuracy of differential expression profiles observed previously for CD4+ T cells (**Fig. 12B**). However, it is possible to correct for these effects in our droplet-based scRNA-seq technologies such as Illumina 10x. This is achieved by the large numbers of empty droplets that are included in the sequencing, which can be used to model ambient RNA expression profiles. These methods include: (1) SoupX[3] uses this approach to directly correct the count data, (2) DecontX[4] or (3) ignoring genes detected in empty droplets all together in downstream analysis have also been used to tackle this problem[5]. All these cutting-edge tools are part of MSSM computational pipeline for scRNAseq analysis and also pre-installed and available as part of our High-Performance Computing (HPC) Resources, Icahn School of Medicine at Mount Sinai.

4.7 Validation of inter-site analytical reproducibility

To validate inter-site validation, a set of 6 patients (18 samples) were sequenced across all sites, additional to over 200 patients sequenced at each site. Post filtering of artifacts such a doublets or triples and removal of empty droplet followed by standard data cleaning resulted in non-significantly different distributions of mitochondrial genes **(Fig. 14A)**, gene features **(Fig 14B)** and gene counts **(Fig 14C)**. The correlation across biological samples (different patients) was consistently ranging between 0.7 and 1 (P<0.001), as shown by other studies[6]. However, the correlation between our three sites (BIMC, MSSM and WASHU) showed a correlation above 0.90, P<0.001 in average **(Fig. 14D)**. The highest correlation was value was Rho 0.99, P<0.001 **(Fig. 14 E, F and G)**. Together, these results show that scRNAseq at MSSM maintains high quality standards similar to BIMC or WASHU (Correlation above 95% cross-site). Further, the sequencing QC results from MSSM showed less inter-sample variability compared with BIMC or WASHU. As expected, the biological diversity was the main driver of variance, however, unsupervised clustering was sufficient to cluster samples based on their patient of origin **(Fig. 14D)**.

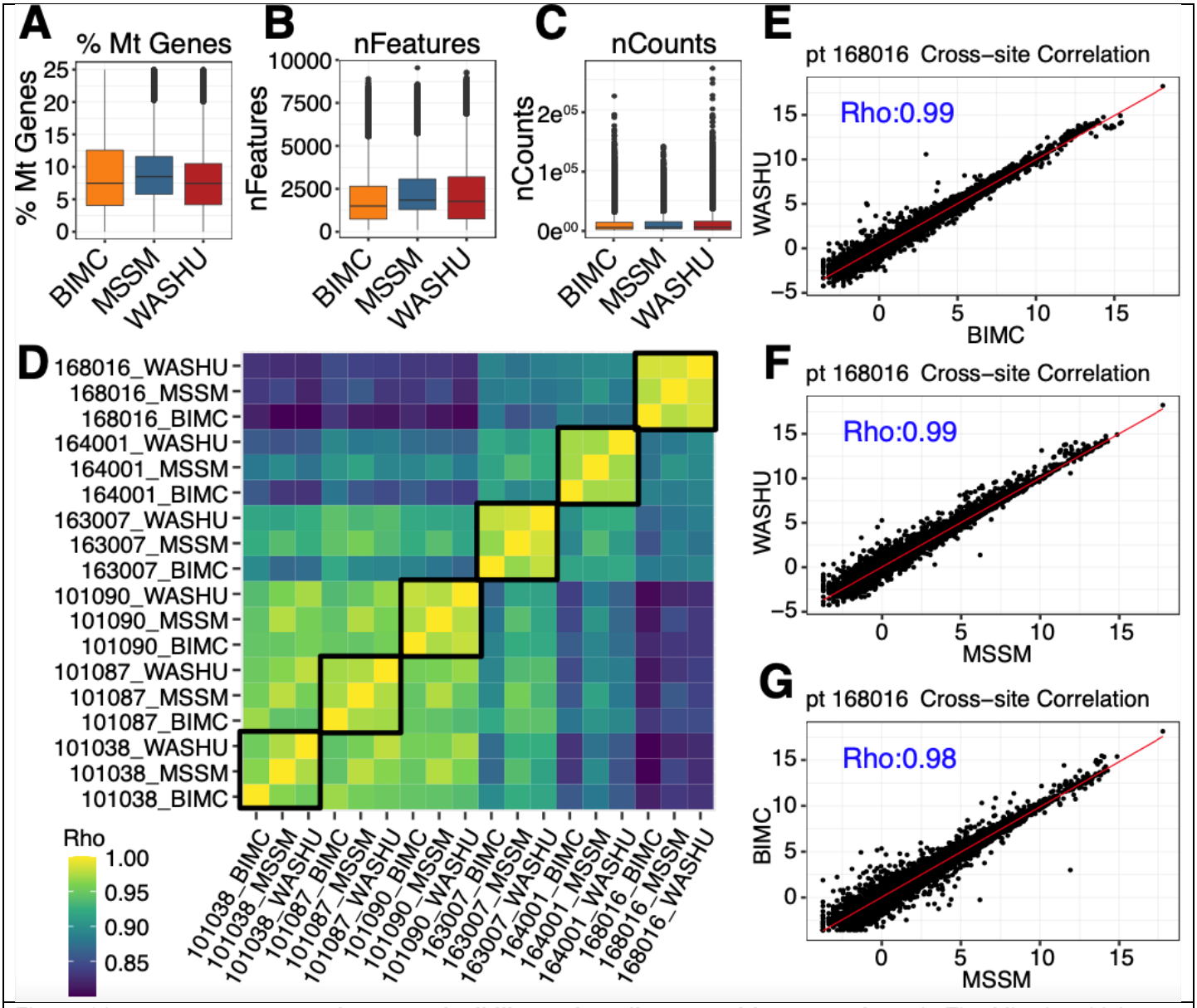

Figure 14. **ScRNAseq sequencing reproducibility and quality control between sites.** A. The Mitochondrial gene variation is shown in boxplots colored by site. A threshold of 25% or less was used. B. The number of genes or nFeatures distribution is shown in boxplots colored by site. The cutoff of at least 200 genes per cell or more was used. C. The number and variance in number of counts is shown by boxplots colored by site. The threshold used was at least 1000 counts per site. D. A Correlation heatmap based on the pseudo-bulk aggregate per patient is shown to represent the concordance between sequencing site and patient heterogeneity. The scale starts at Rho 0.8 and the black boxes area indicative of the same patient sample. Individual correlations for patient 168016 are shown in E, F and G. highlighting the correlation in aggregated log2 expression values across all 3 sites.

4.8 Validation of sensitivity in detection of unique cell clusters

To test the sensitivity of detecting cell clusters using unsupervised methods, we performed reduction of dimensionality (UMAP) using top 5000 variable genes, followed by clustering in 18 single samples corresponding to 6 patients across 3 institutions part of MMRF consortium. This analysis identified clusters that can be found across samples and rarer clusters common to only few samples or even unique samples **(Fig. 15)**. A quantification of the number of cells and percent of rare/unique clusters show that these correspond to 10% of the total number of cells **(Fig. 15C)**. Despite the number of cells for rare/unique clusters appear high in some cases **(Fig. 15A),** they represent a smaller percent of cells **(Fig. 15B),** when all cells are considered. These clusters can be further investigated and assigned to either: technical artifacts, dying cells, cell division and others.

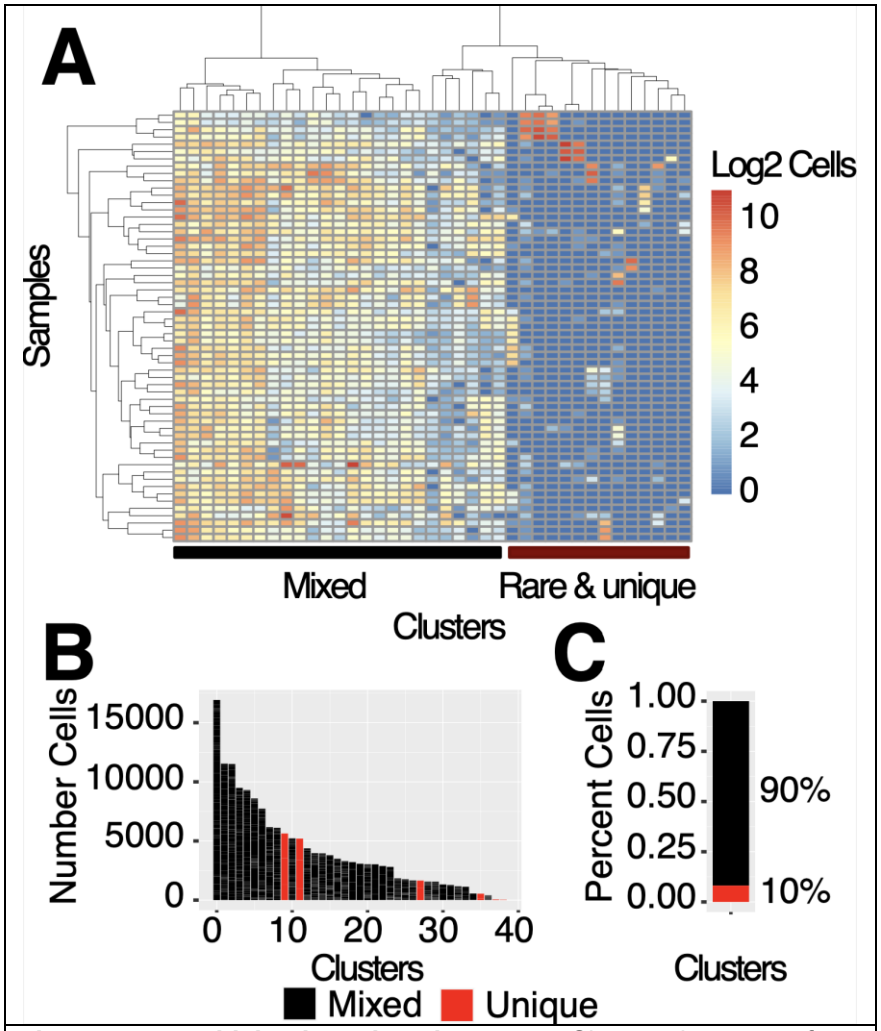

Figure 15. Sensitivity detecting clusters. A. Shows a heatmap of clusters (x-axis) and samples (y-axis) with values in log2 Cells number scale. **B.** Shows the number of cells in each cluster, x-axis: number of cells, y-axis:number of clusters. **C.** Barplot showing the percent of unique to sample clusters.

4.9 Validation of TCRseq sensitivity

To test the capabilities of HIMC into identifying TCRseq-like clones in either bulk or scRNAseq and establish the reproducibility between both assays regarding T-cell clones, we used the TRUST4 and MIXCR algorithms, which reconstruct the CDR3 regions per clonotype. The samples for these sections were obtained by TCR sequencing

of the same samples using single cell TCRseq and bulk TCRseq (n=10). There was no significant difference between both algorithms. Here, we identified a greater number of clones in bulk than single cell **(Fig. 16A)**, as expected due to library size. The quantification of bulk clones compared to single cell showed that bulk RNAseq can identify 3 times more unique TCRs compared to single cell (71% vs 29% (shared), respectively), also due library size and total number of cells per assay. These comparisons were only possible by sequencing the same samples with both bulk and single cell TCR-seq. More importantly, when we use the identified clones to investigate the difference between biological conditions, we observe similar to identical results from either bulk or single cell RNAseq regarding the directionality of the changes per studied condition **(Fig. 16B&C)**. Further, a correlation of clones identified by both assays show a high reproducibility of 0.94 Spearman Rho **(Fig 16D)**. In summary, these results show that both bulk and single cell TCR reconstruction can be used to extract biologically relevant information about T-cell clonality in either assay. Although single cell assays capture less cells than bulk TCR sequencing due library size/coverage, they represent a major opportunity to extract matching transcriptional profiles for the single cells and higher clonal confidence due the availability of both alpha and beta TCR chains.

Bulk-TCRseq – scRNAseq Comparison

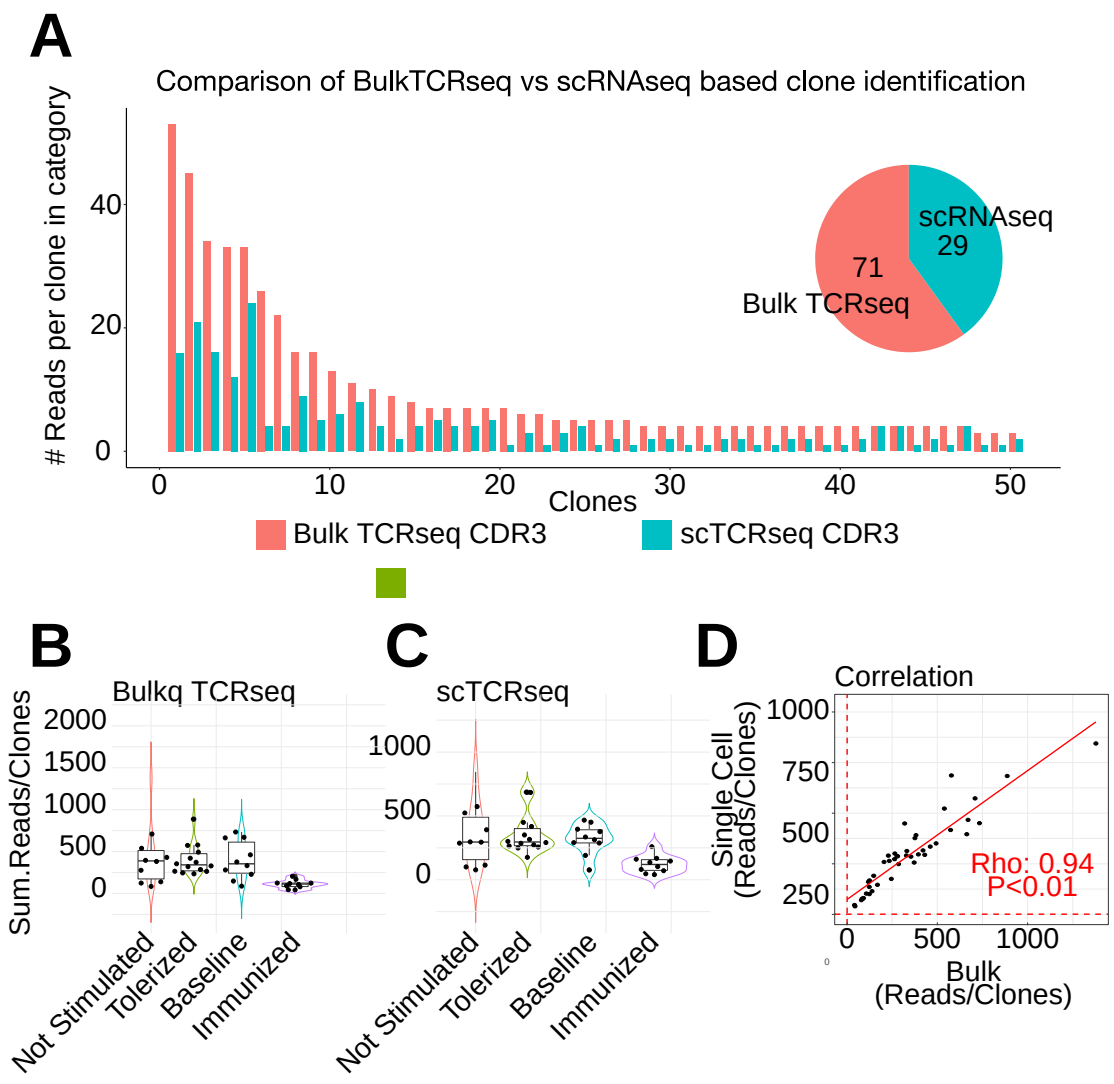

Figure 16. Comparing bulk to single cell TCRseq. A. Shows the number of reads per clone identified by either single cell or bulk TCRseq. **B.** Boxplots showing the proportion of reads per clone in 4 different conditions using bulk TCRseq. **C.** Boxplots showing the proportion of reads per clone in single cell TCRseq. **D.** Correlation of identified clones between bulk and single cell TCRseq.

5. References

1 Lopes, M. B. & Vinga, S. Tracking intratumoral heterogeneity in glioblastoma via regularized classification of single-cell RNA-Seq data. *BMC Bioinformatics* **21**, 59, doi:10.1186/s12859-020-3390-4 (2020).
2 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587 e3529, doi:10.1016/j.cell.2021.04.048 (2021).
3 Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, doi:10.1093/gigascience/giaa151 (2020).
4 Angelidis, I. *et al.* An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nat Commun* **10**, 963, doi:10.1038/s41467-019-08831-9 (2019).
5 Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol* **21**, 57, doi:10.1186/s13059-020-1950-6 (2020).
6 Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, e8746, doi:10.15252/msb.20188746 (2019).