Humanizing Phenomics 120
psomagen

Raw data

Report

2025-02-26

Dr. Jane Doe

10x Chromium 5’ GEM-X

1. Project Information

Customer	Dr. Jane Doe
Institute	Example University
Project ID	AN00012345
Platform	10x Chromium 5’ GEM-X
Organism	Mouse
Number of Samples	4
Sequencer	NovaSeq X Plus 10B (150 PE)

Table of Contents

2. Assay Description

2.1 Cell Capture and Library Preparation:
Single cells are partitioned and captured in droplets, where all generated cDNA share a common barcode. These droplets are nanoliter-scale Gel Beads-in-Emulsion (GEM), which are formed by combining gel beads, a master mix, and partitioning oil. Cells are delivered in a limiting dilution, where most GEMs are empty (90-99%), while the remainder primarily contain a single cell.

After capture, the cell is lysed and cDNA is synthesized. At this point, cDNA are pooled together by sample as each cell was uniquely barcoded. Library prep then proceeds using standard Illumina protocols with paired-end reads.

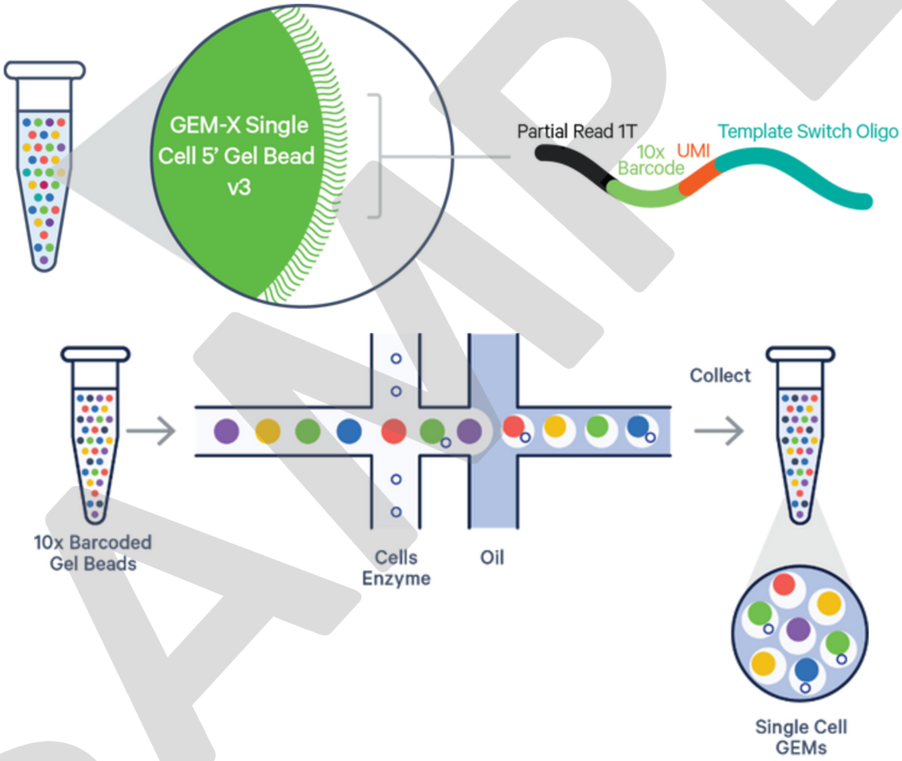


Figure 1: Schematic of 10x 5' Chromium cell capture.

2.2 Analysis:

After sequencing, samples are passed to cellranger mkfastq, a pipeline which demultiplexes raw Illumina base call (bcl) files into fastq files. The mkfastq module is a wrapper built around the standard Illumina tool bcl2fastq and optimized for single cell RNA-seq data sets.

After fastq files are generated, they are passed to the cellranger count pipeline, which performs alignment, filtering, barcode counting, and unique molecular identifier (UMI) counting. Cellranger count uses 10x barcodes to generate feature-barcode matrices, filter GEMs which do not contain cells, cluster cells, and perform basic gene expression analysis.

The feature-barcode matrices filtered for empty GEMs may then be loaded into standard bioinformatics tools such as scanpy (python) or Seurat (R) for advanced analysis.

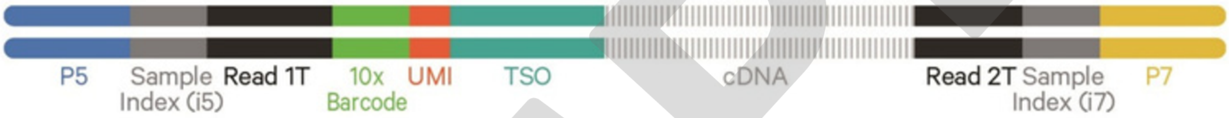


Figure 2: Example diagram of a 10x Chromium GEM-X 5' library read.

3. Data Access and Downloading

3.1 Download Links:

File name	File size	md5sum
AN00012345	xxx	xxxx

Report.zip - This is a zip file of analysis results.

md5sum: In order to verify the integrity of files, md5sum is used. If the values of md5sum are the same, there is no forgery, modification or omission.

Your data will be retained in our server for 3 months. Should you wish to extend the retention period, please contact us.

3.2 Download Instructions:

Data has been transferred via sFTP/Globus/hard drive. Log-in instructions are provided below:

3.3 Provided Files:

1. Raw data (per sample)
    a. fastq files
    b. QC statistics
    cOrderInfo.txt
    d. Program_info.txt
    e. Sample_list.txt
2. Counts output (per sample)
    a. Cloupe.cloupe
    b. Md5sum_check
    c. Report
    d. Web_summary.html
    e. Raw_feature_bc_matrix
        i. Matrix.mtx
        ii. Features.tsv
        iii. Barcodes.tsv
    f. Filtered_feature_bc_matrix
        i. Matrix.mtx
        ii. Features.tsv
        iii. Barcodes.tsv

3. Analysis (per project)
    - a. AN000012345_Seurat.rds (saved Seurat object)
    - b. Sample_QC
        - i. Violin plots of nCounts RNA, nFeatures RNA, and percent mitochondrial reads all cells (QC_project.png)
        - ii. Violin plots of nCounts RNA, nFeatures RNA, and percent mitochondrial reads per sample (QC_sample.png)
        - iii. Scaeerplots of nCounts RNA, nFeatures RNA by sample (QC_scaeerplot_sample.png)
        - iv. PCA plot by samples (QC_PCA_sample.png)
    - c. Clustering
        - i. Elbow plot for number of PCs/dimensions to select (QC_dims_elbow_plot.png)
        - ii. UMAPs by sample and cluster (UMAP_samples_clusters.png)
    - d. DGE (by cluster)
        - i. Significantly (FDR < 0.05) differentially expressed genes (DEGs_cluster.txt)
        - ii. Heatmap of top 5 DEGs/cluster (DEGs_heatmap.png)
    - e. Cell_ type_identification
        - i. UMAPs automatically annotated cell types (UMAP_SingleR_annota9on)
        - ii. Heatmap of predicted annotation probabilities (Cell_type_SingleR_main.png)
        - iii. Heatmap of predicted annotation probabilities (Cell_type_SingleR_fine.png)
        - iv. UMAPs final, manually annotated cell types (UMAP_final_annota9on.png)
        - v. Cell cluster marker genes dot plot (Dot_plot_markers.png)
        - vi. Cell counts per main cell type (Sample_cell_main_type_counts.png)
        - vii. Cell counts per main cell type (Sample_cell_subtype_counts.txt)
        - viii. Cell counts per cell subtype (Sample_cell_subtype_counts.png)

4. Quality Control

4.1 Captured Cells:

	Pre-live/dead selection		Post-live/dead selection		
	Cell number	Cell Viability	Cell number	Cell Viability	Live cell concentration
Control_1	2,670,716	91.9%	2,454,388	93.5%	2,282 cells/µL
Control_2	1,436,346	77.4%	1,111,732	89%	989 cells/µL
Treatment_1	1,947,115	90.9%	1,769,928	95.6%	1,681 cells/µL
Treatment_2	2,289,405	87.1%	1,994,072	89%	1,755 cells/µL

4.2 Sequencing and cellranger count:

	Control_1	Control_2	Treatment_1	Treatment_2
Estimated Number of Cells	29,575	27,466	20,376	29,359
Mean Reads per Cell	60,112	62,754	74,436	44,847
Median Genes per Cell	4,478	3,891	3,774	3,826
Number of Reads	1,777,823,593	1,723,588,291	1,516,706,297	1,316,665,225
Valid Barcodes	89.5%	90.4%	89.7%	89.6%
Sequencing Saturation	30.3%	31%	30.5%	29.5%
Q30 Bases in Barcode	94.6%	95%	93.9%	93.6%
Q30 Bases in RNA Read	94.8% 95.5%	93.2%	94.2%	93.9%
Q30 Bases in UMI	88.3%	96%	96.3%	96.2%
Reads Mapped to Genome	81.9%	89.7%	89.8%	88.4%
Reads Mapped Confidently to Genome		81.3%	80.1%	80.6%
Reads Mapped Confidently to Intergenic Regions	13.3%	11.4%	12.1%	12%
Reads Mapped Confidently to Intronic Regions	6.9%	7.3%	7.7%	7.3%
Reads Mapped Confidently to Exonic Regions	61.7%	61.8%	61.3%	63.1%
Reads Mapped Confidently to Transcriptome	64.3%	64.7%	64.1%	62.6%
Reads Mapped Antisense to Gene	5.8%	6.9%	6.7%	6.9%
Fraction Reads in Cells	96%	96.4%	94.4%	95.4%
Total Genes Detected	26,888	26,091	25,010	25,230
Median UMI Counts per Cell	15,183	16,196	15,052	15,177

4.3 Target cellranger QC Metric Thresholds:

- Estimated Number of Cells: 20,000/sample
- Mean Reads per Cell: 50,000/cell
- Fraction of Reads in Cells: >70%
- Valid Barcodes: >75%
- Valid UMIs: >75%
- Q30 Bases in RNA Read: >65%
- Reads Mapped Confidently to Intronic Regions: <10% (all cells); <25% (PBMCs only); <50% (nuclei only)
- Reads Mapped Confidently to Transcriptome: >30%
- Reads Mapped Antisense to Gene: <10%

5. Sequencing Analysis

5.1 Transcript Abundance and QC After filtering and removing empty GEM droplets, transcript frequency is assessed to determine the likelihood a droplet contains a single, healthy cell. In this case, nFeatures refers to the number of unique genes captured in each droplet, nCount refers to the number of RNA transcripts captured in each droplet, percent.mt refers to the percentage of transcripts which mapped to the mitochondrial genome, and percent.rb refers to the percentage of transcripts which mapped to ribosomal component genes.

The dataset was filtered for 300 < nFeature < 5000 and percent.mt < 5 to remove droplets which may contain more than one cell or dead/dying cells.

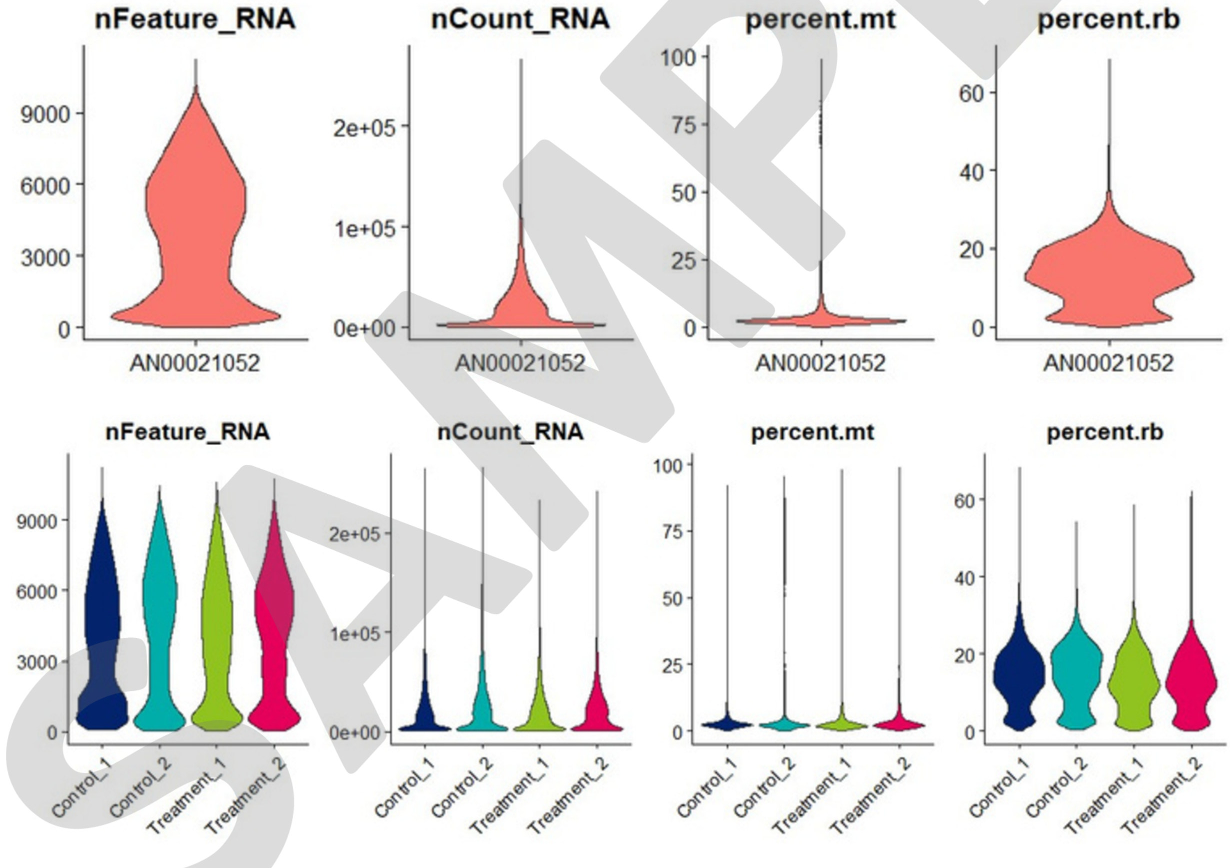


Figure 3: Pre-filtering transcript abundance in all cells by dataset (top) and by individual sample (bottom).

5.2 PCA:

After filtering, the dataset was normalized and scaled and then sample variance was visualized with principal component analysis (PCA).

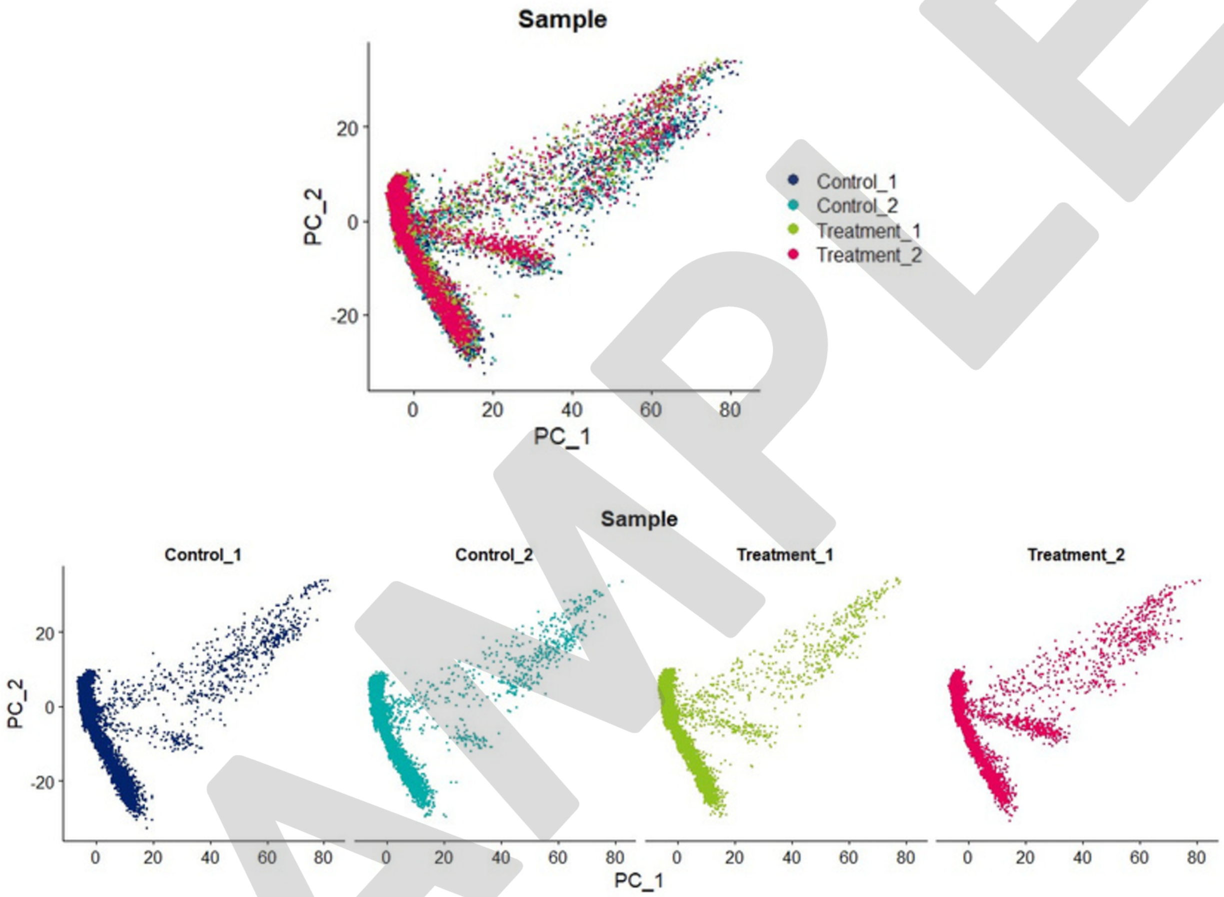


Figure 4: PCA plot of scaled RNA-seq data by dataset (top) and sample (bottom).

5.3 Cell Clustering:

The number of significant principal components, or dimensions, was determined by an elbow plot of the standard deviation for each dimension. Significant PCs are identified in the "upper arm" above the bend or "elbow", while non-significant PCs are found in the plateau region in the "lower arm". Too few dimensions may omit biological diversity while too many dimensions may create false clusters. Here, 30 dimensions were selected to maximize diversity with a resolution of 0.6.

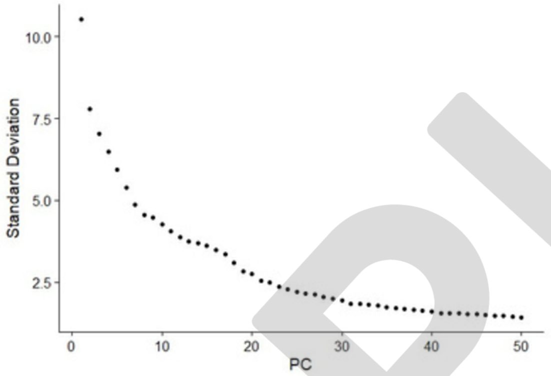


Figure 5: Example elbow plot.

After selecting the number of dimensions, cell neighbors and clusters are identified, which are visualized by uniform manifold approximation and projection (UMAP).

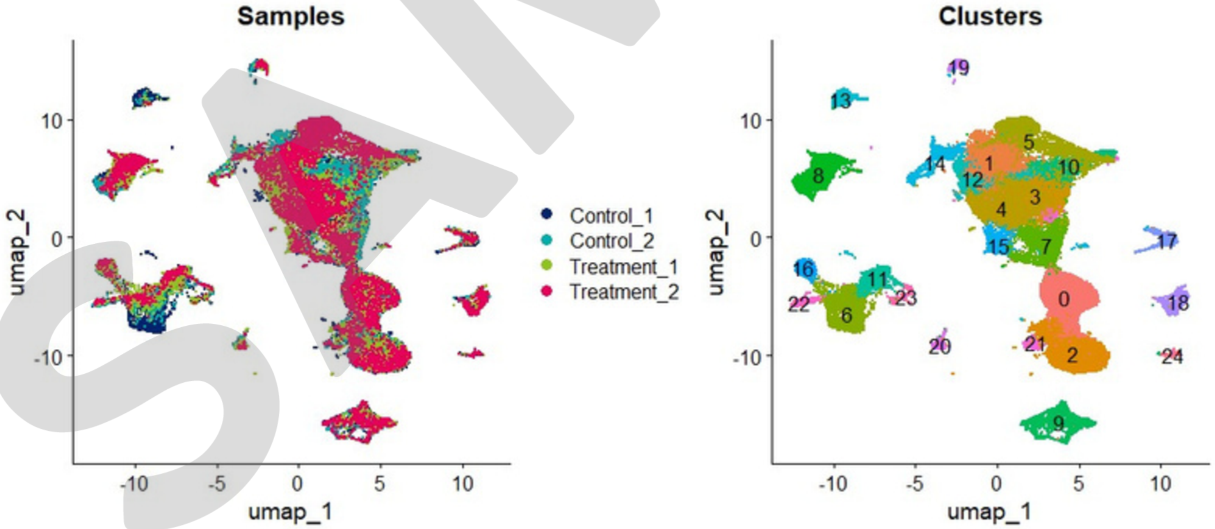


Figure 6: UMAPs grouped by sample (left) and by cluster (right).

5.4 Differentially Expressed Genes:

Differentially expressed genes (DEGs) were calculated by comparing expression of one cell cluster against all others, then repeating for each cluster. The top five most significantly enriched genes (FDR < 0.05) for each cluster were then visualized by heatmap.

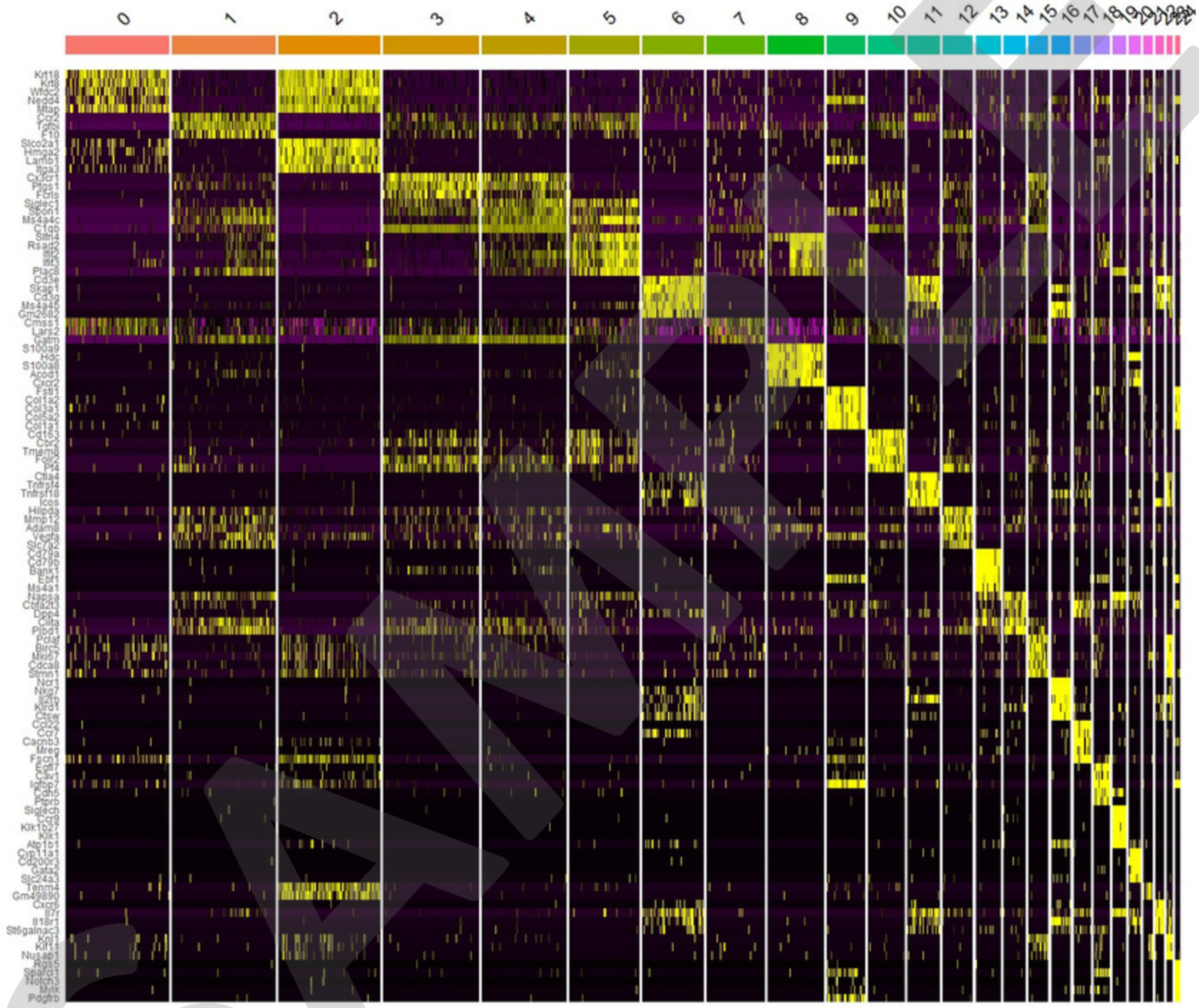


Figure 7: Top 5 DEG per cluster heatmap.

5.5 Cell Type Identification:

Cell cluster identification was achieved using comparison-based mapping of reference scRNA-seq datasets from multiple R packages (SingleR – mouse_rnaseq (MmRNAseq), ImmGen, Novershtern Hematopoietic; easybio – CellMarker2.0). Cell typing was confirmed by manual review and cluster-specific expression of known cell marker genes.

Final identification was given at two levels: (1) main and (2) subtype; e.g. main cell type for clusters 6 and 11 was T cells, which corresponded to CD4+ and Treg cell subtypes, respectively.

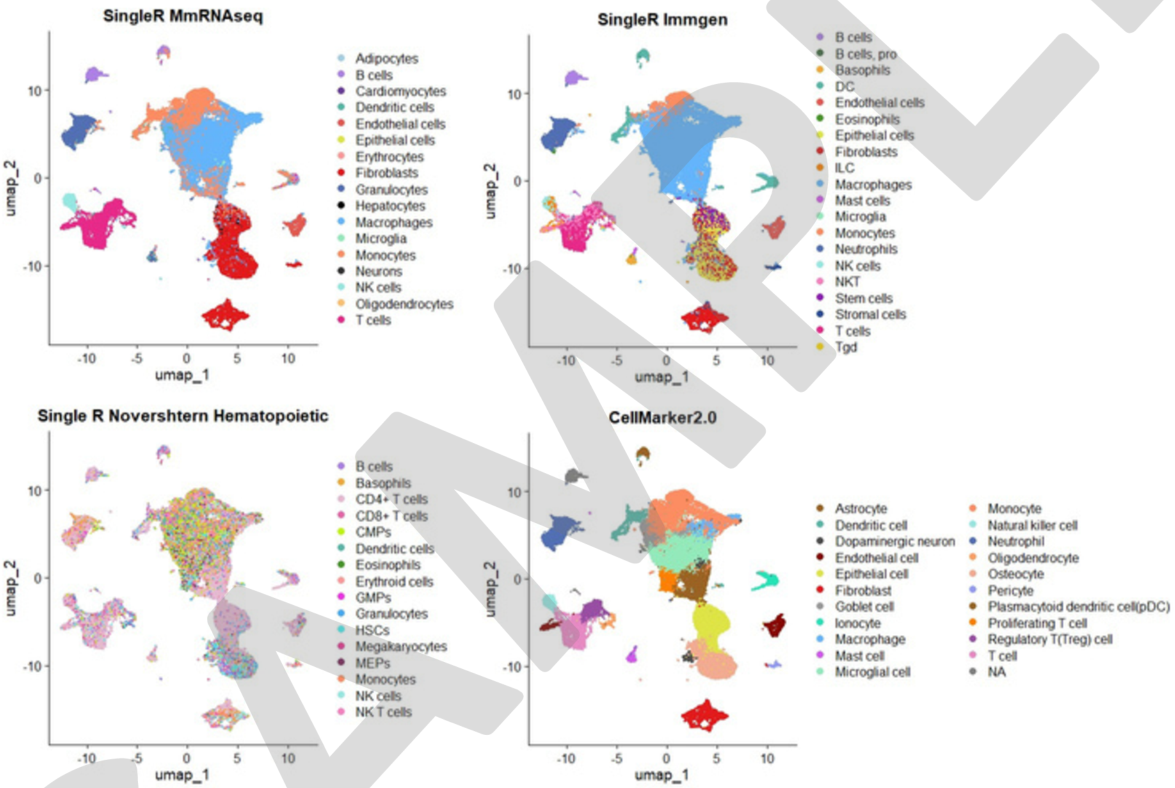


Figure 8: UMAPs of automated cell type identification.

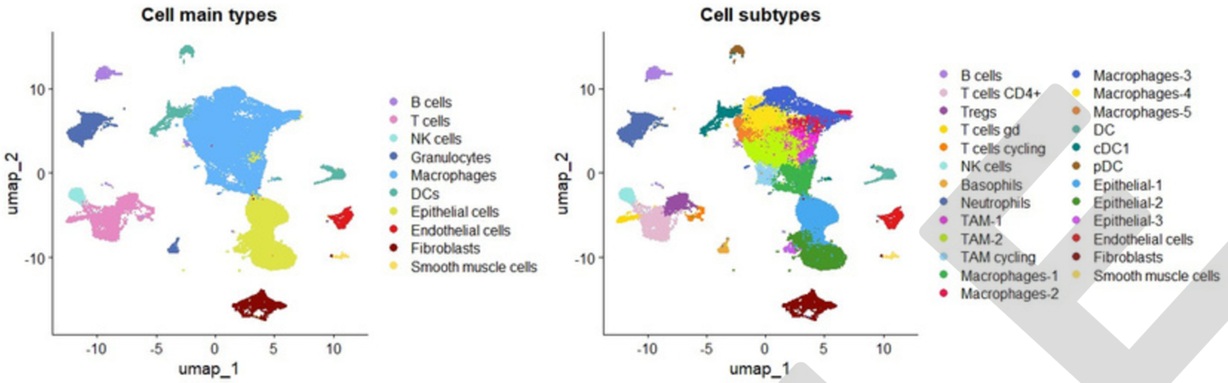


Figure 9: UMAPs of final cell type identification.

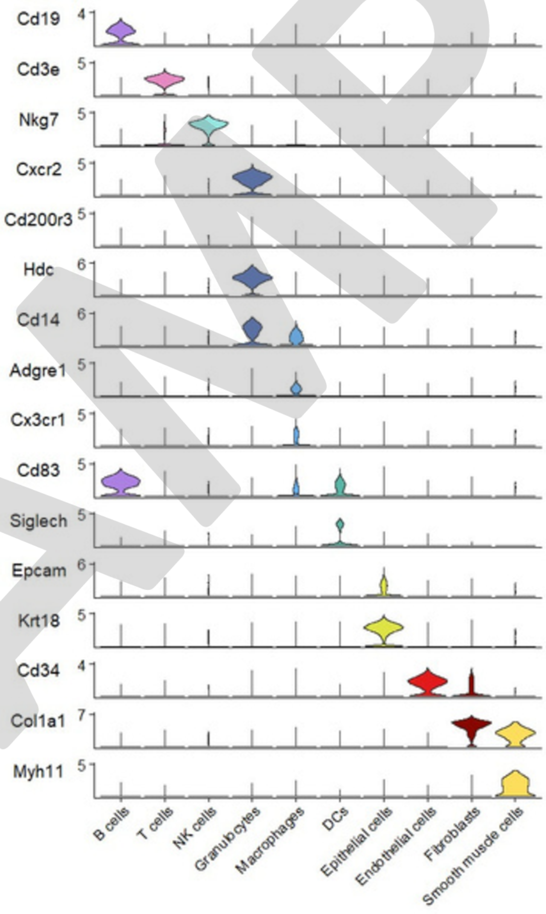


Figure 10: Marker gene expression for each main cell type.

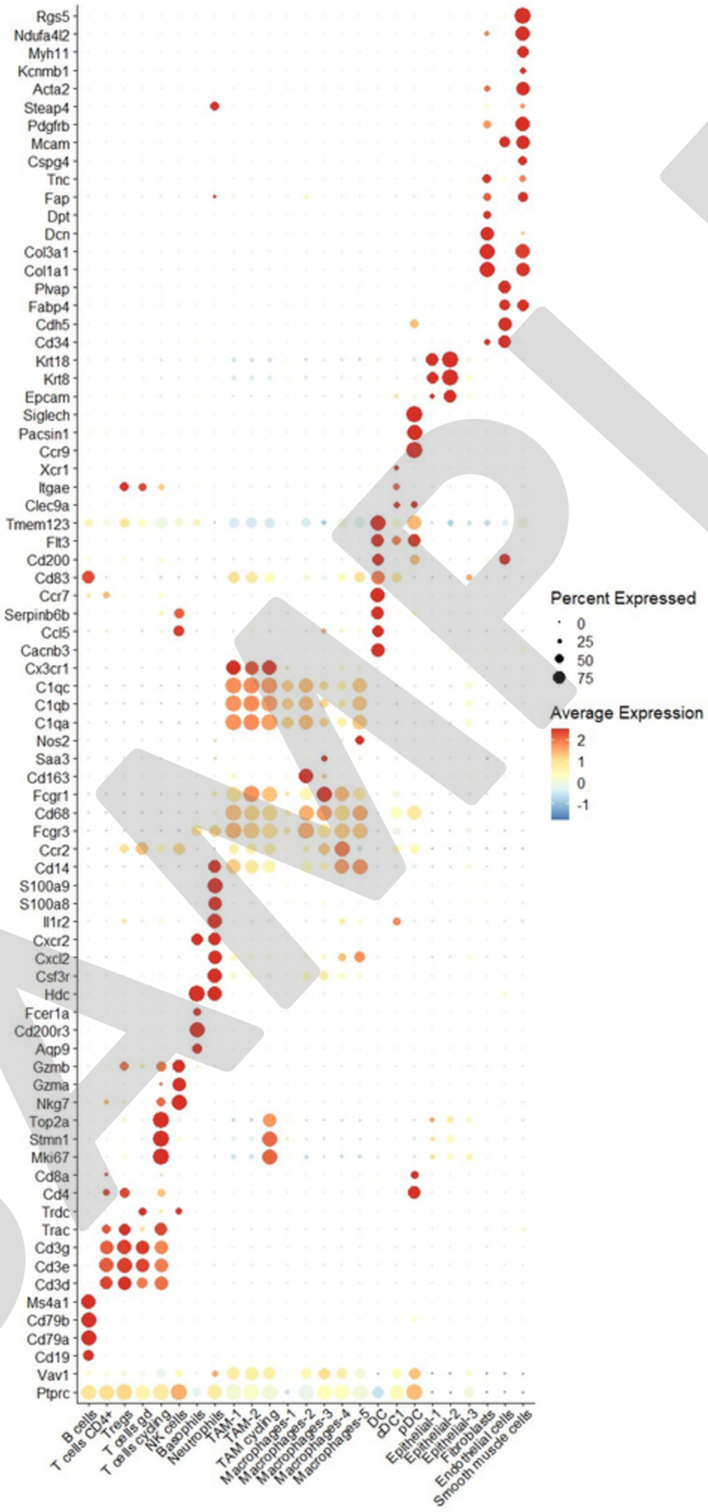


Figure 11: Dot plot of known marker genes and top DEGs are used to identify the cell subtype by cluster.

5.6 Cell Type Abundance:

Per-sample abundance of each main cell type is provided below, as a graph (absolute counts) and as a bar plot of relative abundance.

	Control_1	Control_2	Treatment_1	Treatment_2
B cells	1152	25	37	10
T cells	2867	1067	1002	381
NK cells	246	211	410	101
Granulocytes	757	802	1087	745
Macrophages	6408	4897	7487	5084
DCs	899	708	610	356
Epithelial cells	3304	3129	2427	1622
Endothelial cells	112	96	187	359
Fibroblasts	548	438	440	487
Mesenchymal stem cells	68	79	42	68

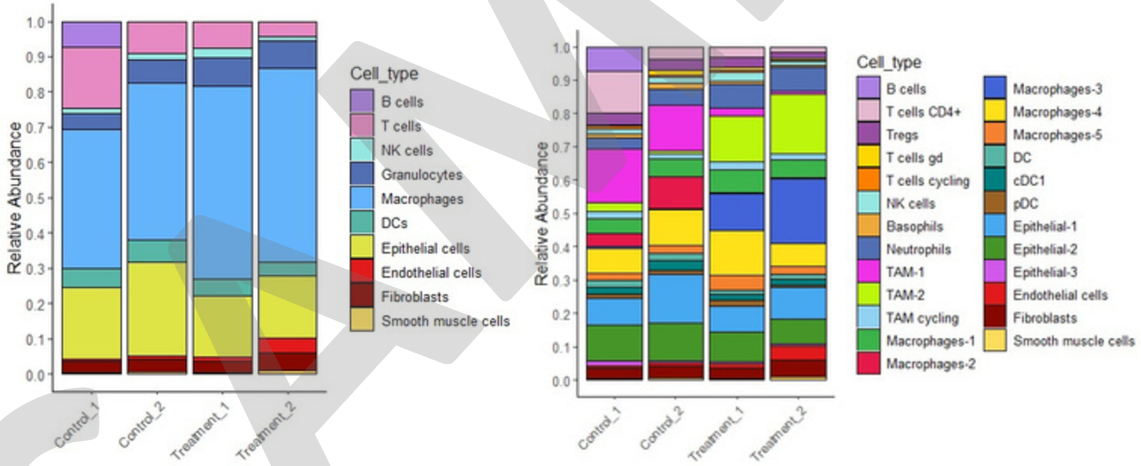


Figure 12: Cell abundance by sample for main cell type (left) and cell subtype (right).

6. Software and Package Description

6.1 FastQC:

FastQC performs quality checks on the raw sequences before analysis to ensure data integrity. The main function is importing BAM, SAM, or fastq files and providing quick overviews on which section(s) has problems. Output includes graphs on read quality and tables.

6.2 cellranger 8.0.1:

cellranger is a set of analysis pipelines that process 10x single cell RNA-seq outputs to generate fastq files, align reads, generate feature-barcode matrices and perform clustering and gene expression analysis. Individual pipeline components include:

- cellranger mkfastq: Demultiplexes raw base call (BCL) files generated by Illumina sequencers into FASTQ files. It is a wrapper around Illumina's bcl2fastq, with additional useful features that are specific to 10x libraries and a simplified sample sheet format.
- cellranger count: Takes FASTQ files from cellranger mkfastq and performs alignment, filtering, barcode counting, and UMI counting. It uses the Chromium cellular barcodes to generate feature-barcode matrices, determine clusters, and perform gene expression analysis. The count pipeline can take input from multiple sequencing runs on the same GEM well. cellranger count also processes Feature Barcoding data alongside Gene Expression reads.
- cellranger vdj: Takes FASTQ files from cellranger mkfastq or bcl2fastq for V(D)J libraries and performs sequence assembly and paired clonotype calling. It uses the Chromium cellular barcodes and UMIs to assemble V(D)J transcripts per cell. Clonotypes and CDR3 sequences are output as a .vloupe file which can be loaded into Loupe V(D)J Browser.
- cellranger aggr: Aggregates outputs from multiple runs of cellranger count, normalizing those runs to the same sequencing depth and then recomputing the feature-barcode matrices and analysis on the combined data. The aggr pipeline can be used to combine data from multiple samples into an experiment-wide feature-barcode matrix and analysis.
- cellranger reanalyze: Takes feature-barcode matrices produced by cellranger count or cellranger aggr and reruns the dimensionality reduction, clustering, and gene expression algorithms using tunable parameter settings.
- cellranger multi: Takes FASTQ files from cellranger mkfastq or bcl2fastq for any combination of 5' Gene Expression, Feature Barcode (cell surface protein or antigen) and V(D)J libraries from a single gem-well. It performs alignment, filtering, barcode counting, and UMI counting on the Gene Expression and/or Feature Barcode libraries. It also performs sequence assembly and paired clonotype calling on the V(D)J libraries. Additionally, the cell calls provided by the gene expression data are used to improve the cell calls inferred by the V(D)J library.

6.3 Seurat v5.1.0:

Seurat is an R package developed by the Rahul Satija lab for advanced analysis of single cell sequencing datasets. It can perform cell filtering, normalization, clustering, and gene expression analyses.

7. Glossary

Barcode – a unique identifying sequence for attached to all transcripts within a single cell **cellranger** – a modular, analytical pipeline created by 10x Genomics which demultiplexes raw reads, creates fastq files, alignment, filtering, barcode counting, and UMI counting

Chip – a plastic container which processes GEM cleanup and cDNA synthesis

Count – equivalent to transcripts in a single cell
Duplet/multiplet – occurs when 2+ cells are captured in the same GEM droplet

Feature – equivalent to genes in a single cell

GEM – Gel beads in Emulsion, droplets which capture single cells

Library – a group of samples prepared on the same chip for next generation sequencing

Loupe – an analytical tool created by 10x Genomics to check cell clustering and gene expression following cellranger count runs

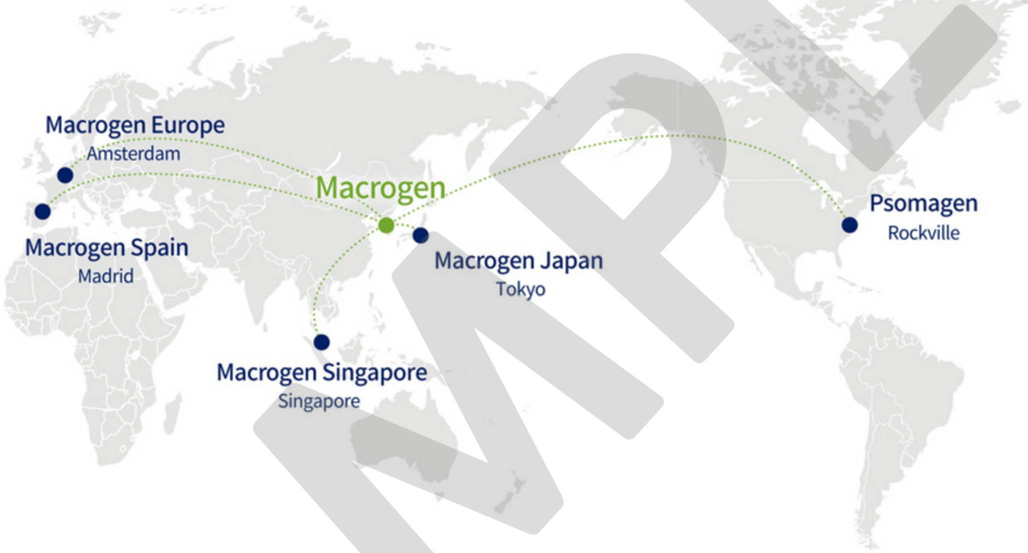
PCA – Principal Component Analysis, a dimensionality reduction technique which helps show relationship between cells

Sample – a group of single cells isolated from a single biological source (e.g. blood, tumor, mouse)

Seurat – an R package developed for advanced analysis of cellranger output files

UMAP – Uniform Manifold Approximation and Projection, a dimensionality reduction technique which helps show relationship between cells by distance between their transcription profiles

UMI – Unique Molecular Identifier, a unique 12 nt (3') or 10 nt (5') identifying sequence randomly created during first-strand synthesis next to the cell barcode. This sequence is unique for each transcript within a single cell.



HEADQUARTER

Macrogen, Inc.

Laboratory, IT and Business Headquarter & Support Center

[08511] 1001, 10F, 254, Beotkkot-ro,
Geumcheon-gu, Seoul, Republic of Korea
(Gasan-dong, World Meridian 1)
Tel: +82-2-2180-7000
Email1: ngs@macrogen.com(Overseas)
Email2: ngskr@macrogen.com
(Republic of Korea)
Web: www.macrogen.com
LIMS: dna.macrogen.com

SUBSIDIARY

Macrogen Europe

**Laboratory,
Business & Support Center**

Meibergdreef 57, 1105 BA, Amsterdam,
the Netherlands
Tel: +31-20-333-7563
Email: ngs@macrogen.eu

Psomagen (Macrogen USA)

**Laboratory,
Business & Support Center**

1330 Piccard Drive, Suite 103, Rockville,
MD 20850, United States
Tel: +1-301-251-1007
Email: inquiry@psomagen.com

Macrogen Singapore

**Laboratory,
Business & Support Center**

3 Biopolis Drive #05-18, Synapse,
Singapore 138623
Tel: +65-6339-0927
Email: info-sg@macrogen.com

Macrogen Japan

**Laboratory,
Business & Support Center**

16F Time24 Building, 2-4-32 Aomi,
Koto-ku, Tokyo 135-0064 JAPAN
Tel: +81-3-5962-1124
Email: ngs@macrogen-japan.co.jp

BRANCH

Macrogen Spain

**Laboratory,
Business & Support Center**

Av. Sur del Aeropuerto de Barajas,
28. Office B-2, 28042 Madrid, Spain
Tel: +34-911-138-378
Email: info-spain@macrogen.com