K-mer Profiling: A Novel Approach for Cell Type Classification and Marker Discovery in scRNA-seq Data

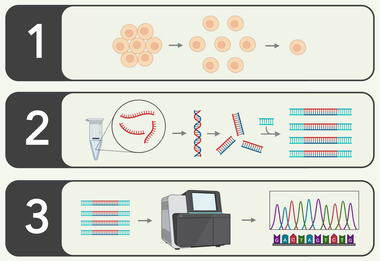
MEGAN HUANG[1], Yiqian Gu[2], and Matteo Pellegrini*[2]

[1]BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA;
[2]Department of Molecular, Cell and Developmental Biology, UCLA

Background: scRNA-seq Data Analysis

SINGLE-CELL RNA SEQUENCING

- *Provides detailed information on individual cells using gene expression*
- *High resolution method that allows for rare or transient cell type discovery*

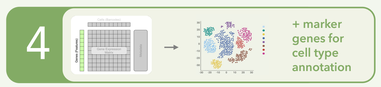


Getting scRNA-seq Data: A Short Guide

1. Harvest, dissociate, and isolate cells.
2. Extract RNA → cDNA, prepare library.
3. Sequence library to get raw reads.

Analyzing scRNA-seq Data With Pipelines

Pipelines are like assembly lines, but for processing or analyzing data – scRNA-seq pipelines process raw reads to annotate cell types for single cells, with these steps:

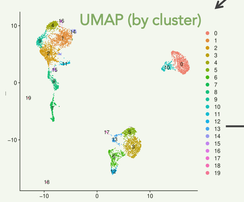


- Preprocessing: Gene Expression Matrix
- Downstream Analysis: Quality Control > Normalization > Feature Selection > Scaling + Regression > Dimensionality Reduction > Clustering > Marker Gene Identification > Cell Type Annotation

Standard Pipeline

scRNA-seq Raw Reads .fastq → **Cellranger** → *Gene Expression Matrix* *.mtx + 2 .txt*

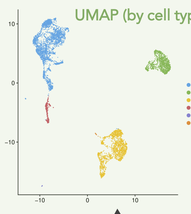
aligns reads to genes, counts genes, outputs gene expression matrix



used to annotate cell types on UMAP

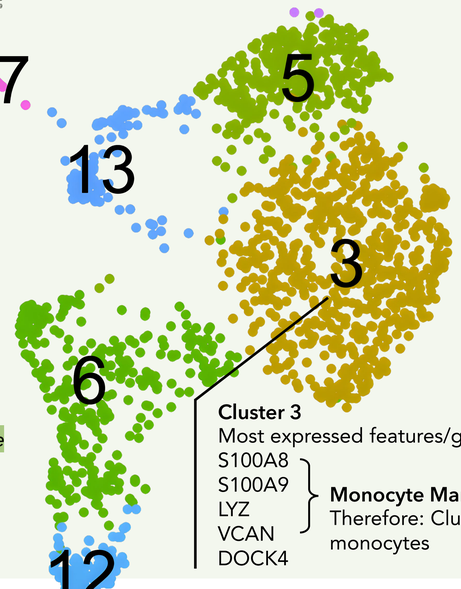
Gene Expression Matrix *.mtx + 2 .txt* → **Seurat** → outputs clusters (UMAP) and list of marker genes → *Marker Genes (Std.)* .csv

End Result
Cells are clustered and annotated by cell type.



How to Annotate a UMAP

UMAPs are visual representations of the clusters that individual cells fall into.
Cells are sorted into clusters based on feature expression similarity.
A cluster's cell type is identified using its marker features.



Cluster 3
Most expressed features/genes:
S100A8
S100A9
LYZ
VCAN
DOCK4

Monocyte Marker Genes
Therefore: Cluster 3 = monocytes

QUESTION

Can cell type markers be discovered using k-mer counts instead of the conventional gene counts? If so, can it uncover aspects of cellular identity that traditional gene expression analysis might miss?

**k-mer: a short nucleotide sequence of length k*

METHODS & WORKFLOW

K-mer Pipeline

A 10x Human PBMC dataset was used to test this pipeline.

scRNA-seq Raw Reads .fastq → **BKC** → counts k-mers for each cell, outputs k-mer expression matrix → .bkc → .txt → .mtx + 2 .txt

filter k-mers from matrix by count similarity

K-mer Expression Matrix .mtx + 2 .txt → **Seurat** → outputs clusters (UMAP) and list of marker k-mers → *Marker K-mers* .csv

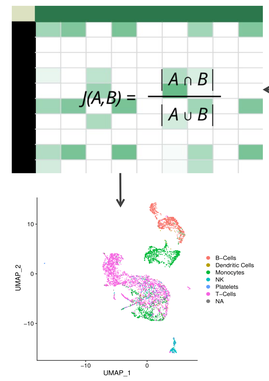
K-mer List .fasta → **Bowtie1 + Samtools** → outputs chromosomes + positions for each k-mer → *K-mer Alignments List* .bam

K-mer alignments can be inspected with this file using a genome viewer.

K-mer Alignments List .bam (no chr format) → **HTSeq + Ensembl** → for each chromosome + position, finds corresponding gene IDs, then gene names → *K-mer + Gene Name Key* .csv

Pipeline Comparison: Jaccard Similarity Matrix + UMAP Annotation

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$



Marker Genes (Std.) .csv

Marker Genes (K-mer) .csv

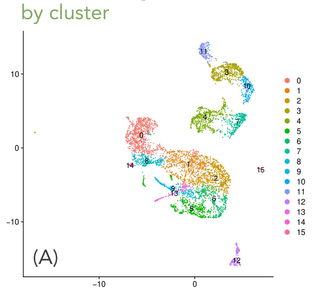
Why k-mers?

Short – may give more sensitive results compared to genes, which are longer.
Not exclusively found in genes - may reflect differences between cell types aside from gene expression.

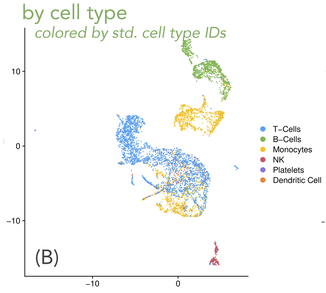
RESULTS

K-mer Pipeline UMAP

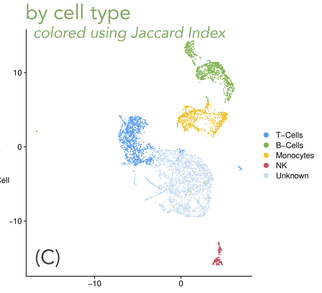
by cluster



by cell type
colored by std. cell type IDs



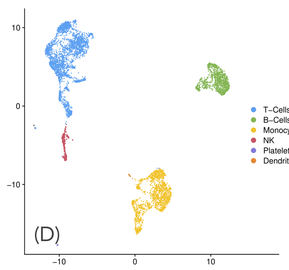
by cell type
colored using Jaccard Index



Integrative Genome Viewer (IGV)

Standard Pipeline UMAP

by cell type



Standard Pipeline		Gene #	K-mer Pipeline: T-Cell	T-Cell	T-Cell	B-Cell	B-Cell	B-Cell	Mono	Mono	NK	Unknown						
			100	100	9	100	100	3	97	70	29	1	2	1	1	7	2	1
		Cluster	0	8	15	3	11	10	7	4	12	14	13	1	2	5	6	9
T-Cell	100	1	0.1049	0.1049	0	0.0695	0.0638	0	0.0102	0	0.0157	0	0	0	0	0	0	0
T-Cell	100	16	0.0989	0.0989	0	0.0752	0.0695	0	0	0	0	0	0	0	0	0	0	0
T-Cell	100	4	0.0989	0.0989	0	0.0752	0.0695	0	0	0	0	0	0	0	0	0	0	0
T-Cell	100	11	0	0	0.0092	0	0	0	0	0.0059	0.0238	0	0	0	0	0	0	0
T-Cell	100	19	0	0	0.0186	0	0	0	0.0207	0.0303	0.0573	0	0	0	0	0	0	0
B-Cell	100	0	0	0	0	0.0309	0.0309	0.0198	0.0051	0	0	0	0	0	0	0	0	0
B-Cell	100	10	0	0	0	0.0204	0.0256	0.0198	0.0051	0	0	0	0	0	0	0	0	0
Monocytes	100	13	0	0	0	0.0101	0.0101	0.0098	0.0591	0.0493	0.0157	0	0	0	0	0	0	0
Monocytes	100	12	0	0	0	0	0	0	0.1129	0.0759	0.0238	0	0	0	0	0	0	0
Monocytes	100	6	0	0	0	0	0	0	0.0478	0.0240	0	0	0	0	0	0	0	0
Monocytes	100	3	0	0	0	0	0	0	0.0314	0.0625	0	0	0	0	0	0	0	0
Monocytes	100	5	0	0	0	0	0	0	0.0102	0.0119	0	0	0	0	0	0	0	0
NK	100	7	0	0	0	0	0	0	0	0	0.0238	0	0	0	0	0	0	0
NK	100	8	0	0	0	0	0	0	0	0	0.0157	0	0	0	0	0	0	0
Platelet	100	18	0	0	0	0	0	0	0.0207	0.0240	0.0238	0	0	0	0	0	0	0
Dendritic Cell	100	17	0	0	0	0	0	0	0.0051	0.0059	0	0	0	0	0	0	0	0
T-Cell	100	14	0	0	0	0	0	0	0	0	0.0078	0	0	0	0	0	0	0
T-Cell	100	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T-Cell	100	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T-Cell	100	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **In (B) and (D), the major cluster groups (t-cells, b-cells, monocytes, NK) are maintained. The smaller groups (dendritic cells and platelets) were missed, though platelets were separated properly on the UMAP.**
- **Note that (C) highly resembles (B), indicating similar cell type identifications between the k-mer pipeline and the standard pipeline.**
- **IGV was used to inspect marker k-mers after they were aligned – though many k-mers aligned to genes, some aligned to intergenic regions preferentially, or even both!**

Conclusions & Future Directions

The k-mer pipeline successfully returns similar cluster patterns and cell-types to the standard pipeline. Many of the marker k-mers map to genes, but many preferentially map to intergenic regions as well. Gene markers produced by both pipelines vary greatly despite similar identifications for larger clusters, but some k-mer determined clusters possess too few genes for accurate clustering and identification.

- *Reduce data loss that occurs in the k-mer to gene processing step.*
- *Observe k-mers that map to intergenic regions, potentially reflecting alternative splicing or isoforms.*

Acknowledgements & References

Dr. Matteo Pellegrini (PI)

Yiqian Gu (Mentor)

I would like to thank Dr. Hoffman and the BIG Summer Program for their invaluable workshops, talks, seminars, and funding that made this poster possible. I also want to thank the DASL scholarship program for its support. Special thanks to Dr. Matteo Pellegrini and Yiqian Gu for their consistent guidance and encouragement, as well as the other members of the Pellegrini Lab for their feedback throughout this project.

5k Human PBMCs (Donor 4), Single Cell Immune Profiling Dataset by Cell Ranger v9.0.0, 10x Genomics, (2024, Nov 18). **Kokot M, Deorowicz S.** (2024). BKC. **Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, Satija R** (2023). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*. **Langmead B, Trapnell C, Pop M, Salzberg S.** (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R** (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **Anders S, Pyl P.T., Huber W.** (2014). HTSeq — A Python framework to work with high-throughput sequencing data. *bioRxiv*. **Dyer S, et. al (2025).** Ensembl 2025. *Nucleic Acids Research*. **Robinson J.T., Thorvaldsdóttir H, Winckler W, Guttman M, Lander E.S., Getz G, Mesirov J.P.** (2011). Integrative Genomics Viewer. *Nature Biotechnology*.