No Cell Left Behind

Recovering Discarded Cells via Self-Supervised Diffusion Denoising

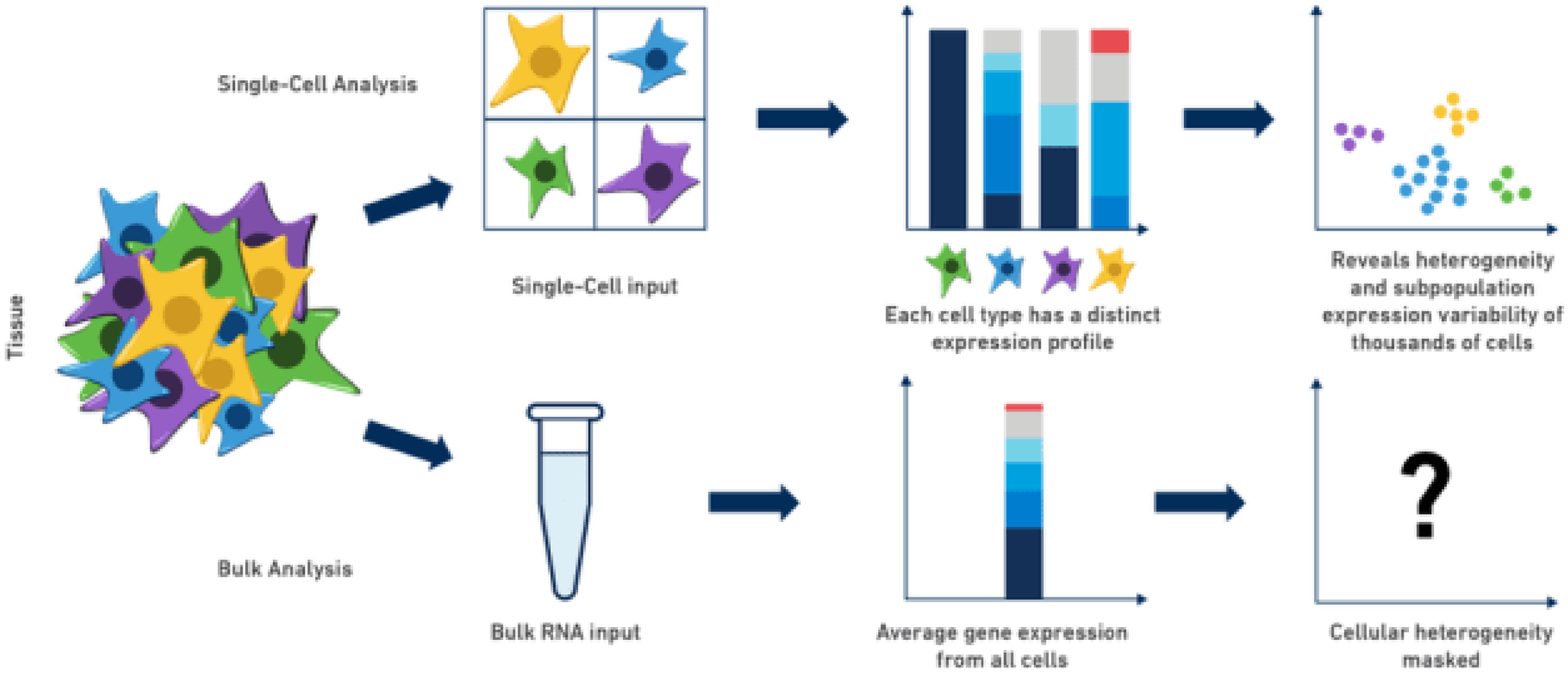
DEMUXLY | **JUYOUNG, NAYOUNG** | **PROF. JAEYOUNG CHUN**

2025.12.17

Agenda

BACKGROUND | OBJECTIVES | METHODS | RESULTS | DISCUSSION

BACKGROUND: scRNA-seq



https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

Stephanie Hicks. Welcome to the World of Single-Cell RNA-Sequencing. (2017). https://speakerdeck.com/stephaniehicks/welcome-to-the-world-of-single-cell-rna-sequencing?slide=3

BACKGROUND: library prep for scRNA-seq cost

10X Chromium Next GEM Single Cell 3’ Reagent Kits v3.1 (Dual Index)

Library Prep Cost

$2,500

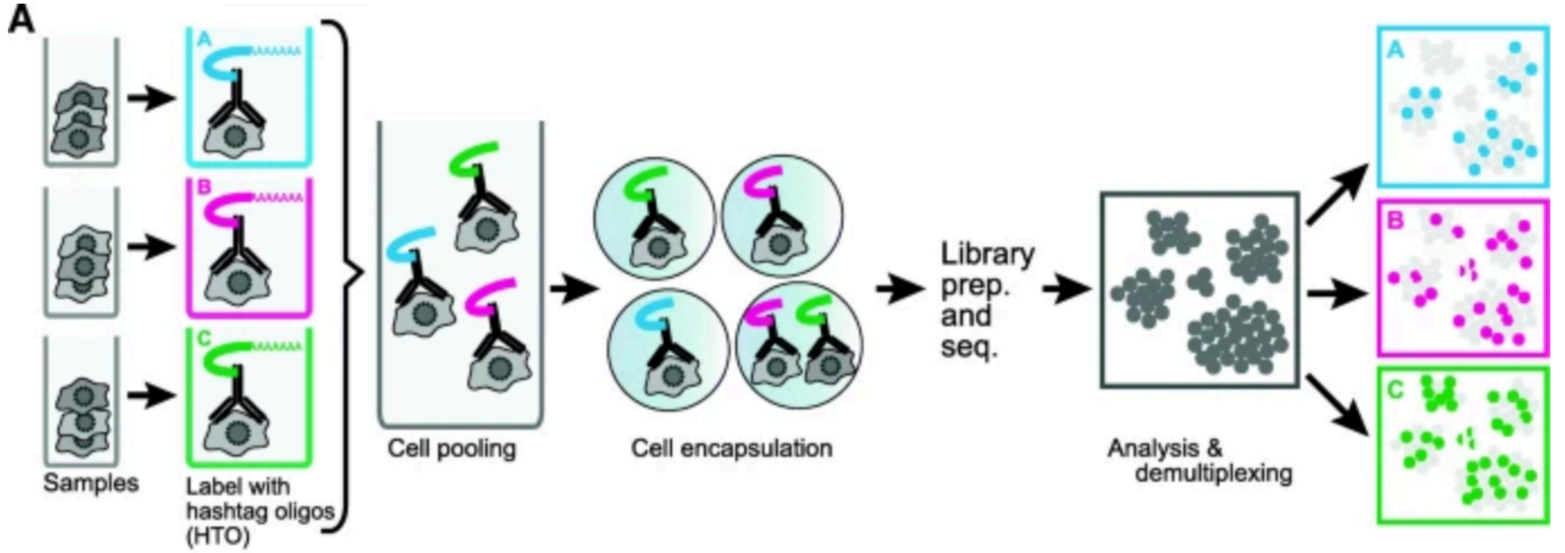
Johns Hopkins Medicine (2022).

https://www.hopkinsmedicine.org/-/media/institute-basic-biomedical-sciences/transcriptomics-website-price-list-fy23.pdf

- Standard Throughput: ~10,000 cells per chip.
- Cost per Reaction is Constant:
  - Running 5,000 cells = $X (1 Kit)
  - Running 10,000 cells = $X (1 Kit)

BACKGROUND: Cell Hashing

Cell Hashing for “super-load”



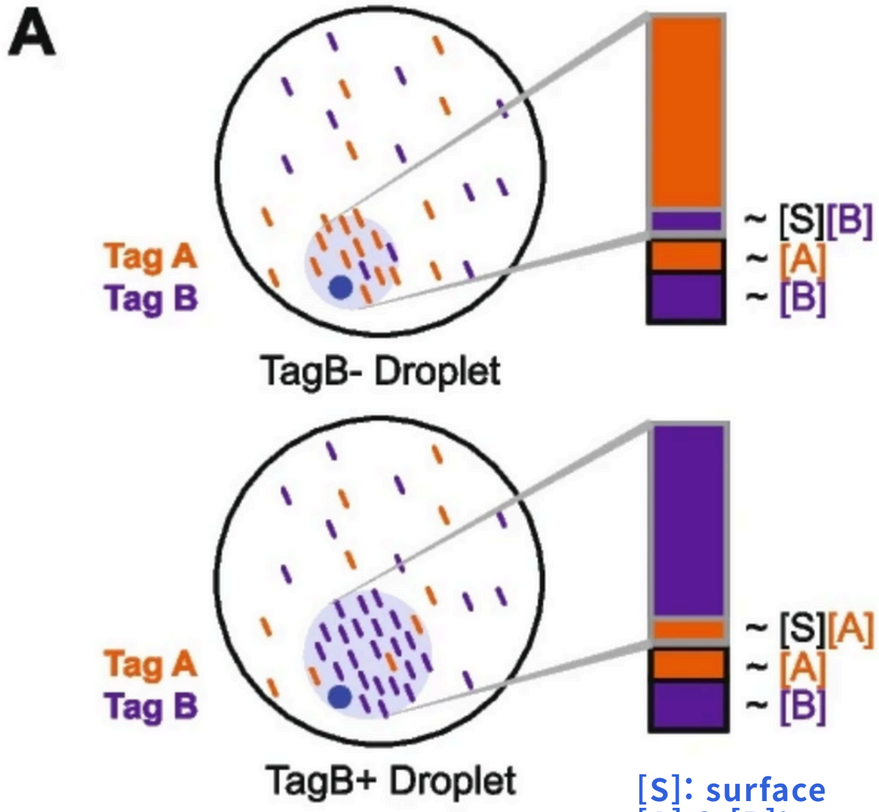
Demultiplexing the samples by unique antibody tag

Stoeckius, M., Zheng, S., Houck-Loomis, B. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol 19, 224 (2018). https://doi.org/10.1186/s13059-018-1603-1

Experiment Protocol: https://cdn.10xgenomics.com/image/upload/v1660261285/support-documents/CG000391_DemonstratedProtocol_CellLabelingwithCellMultiplexingOligo_RevB.pdf

BACKGROUND: Technical artifacts of Hashing

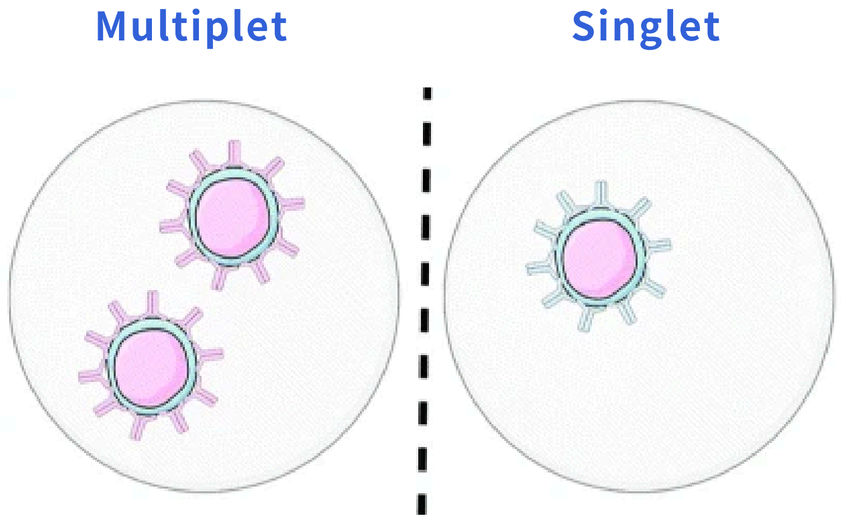
Ambient antibodies → Noise



[S]: surface
[A] & [B]: tag concentration

Zhu, Q., Conrad, D.N. & Gartner, Z.J. deMULTIplex2: robust sample demultiplexing for scRNA-seq. Genome Biol 25, 37 (2024). https://doi.org/10.1186/s13059-024-03177-y

Multiplets



Xin, H., Lian, Q., Jiang, Y. et al. GMM-Demux: sample demultiplexing, multiplet detection, experiment planning, and novel cell-type verification in single cell sequencing. Genome Biol 21, 188 (2020). https://doi.org/10.1186/s13059-020-02084-2

BACKGROUND: Previous Demultiplexing Algorithms

Mixture Model based Methods

- Fit data to fixed distributions
- Tools:
  - GMM-Demux
  - BFF
- **Rigid assumptions**
- **Inflexible to Data Heterogeneity**

Threshold-based Methods

- Assign cells based on hard confidence cutoffs
- Tools:
  - HTODemux
  - demuxmix
- **Low recall**

Contamination Aware Methods

- Contamination factor (e.g., cell surface) modeling
- Explicitly model ambient noise/background
- Tools:
  - deMULTIplex2
- **Low generalizability**

BACKGROUND: Previous Demultiplexing Algorithms

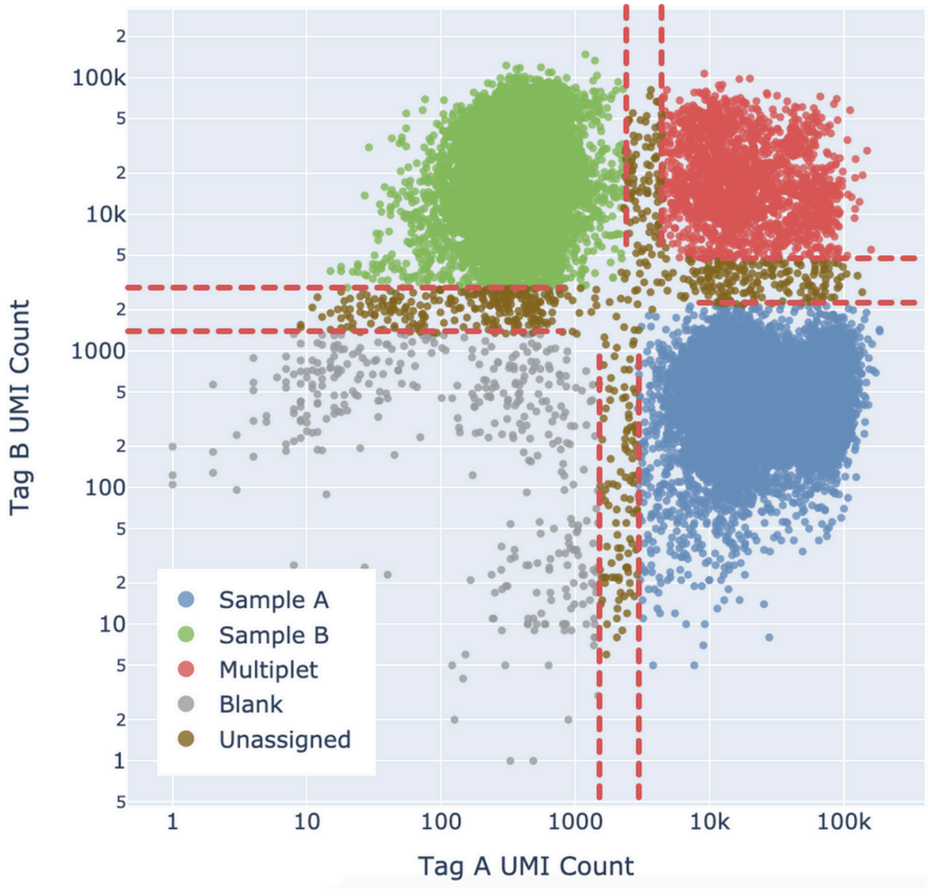
Cell Ranger (Multi)

#The Strict Judge
#High Precision

Confidence threshold-based
Equal variance assumption

Raw Noisy Input → conservative decision boundary

A strategy that selects only high **signal-to-noise ratio (SNR)** cells while discarding the rest



Cell Ranger's Cell Multiplexing Algorithm.
https://www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-3p-cellplex-algorithm

Rare but critical cases in biology

The case of Circulating Tumour Cells (CTCs; key indicators of metastasis and prognosis):

1–100 cells exists
along with **10^6–10^8 red blood cells**

Demultiplexing Objectives

Generalizability

across protocols and datasets

Sensitivity

to detect meaningful heterogeneity;
hard confidence assignment → Low rare cell detection

DeMUXly

Generalizability

across sample number, protocols and datasets

“confounder pattern inference through posterior distribution”

Sensitivity

no hard confidence assignment for rare cell detection

“self-supervised Learning”

Demultiplex by (1) Self-supervised Denoising → (2) Classification

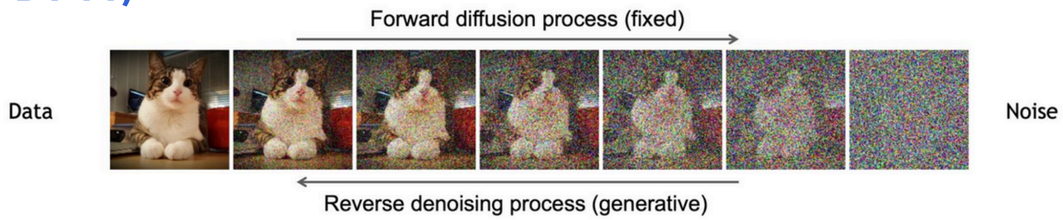
OBJECTIVES: Self-supervised Denoising model

Denoising Diffusion Probabilistic Model (DDPM)

Slow and heavy, but it precisely restores the data distribution

#Iterative Refinement

#Distribution Learning (Probabilistic)



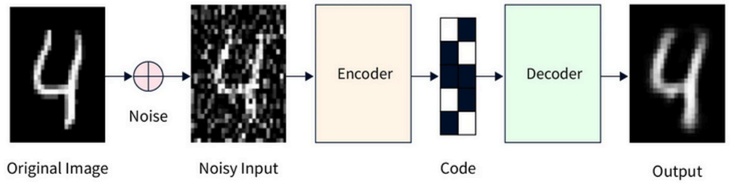
Hao Chen. Deep Learning: Diffusion. Lecture Slide (2025).
https://deeplearning.cs.cmu.edu/S25/document/slides/lec23.diffusion_s25.pdf

Denoising Autoencoder (DAE)

Fast and lightweight, but results can be blurry

#One-Shot Prediction

#Deterministic



Cathrine Jeeva. Exploring Denoising Autoencoders. Scaler Topics (2023).
https://www.scaler.com/topics/deep-learning/denoising-autoencoder/

OBJECTIVES: Self-supervised Denoising model

DDPM

#Iterative Refinement
#Distribution Learning
#Probabilistic

DAE

#One-Shot Prediction
#Deterministic

DeMUXly

“A self-supervised diffusion model can recover ambiguous cells discarded by hard thresholds by learning the underlying signal manifold to denoise confounders”

OBJECTIVES: Hypothesis

DeMUXly

Diffusion-Based Redemption Pipeline

1. Initial Screening:

- Generate ADT Count Matrix via Cell Ranger Multi.
- Apply initial labeling with a hard confidence threshold (> 0.9).
- Identify "Unassigned" or "Low-confidence" cells for rescue.

2. Generative Denoising (Core)

- Train a Self-Supervised Diffusion Model on the data manifold.
- Learn confounder (noise) patterns to distinguish signal from background.
- Iterative Refinement: Denoise raw features into clean latent representations.

3. Final Classification

- Perform GMM (Gaussian Mixture Model) Clustering on the denoised features.
- Assign final cell identities without arbitrary hard cutoffs.

Model

1. Denoise
2. Classification

Data

1. Data Source
2. Preprocessing

Method Summary

Model

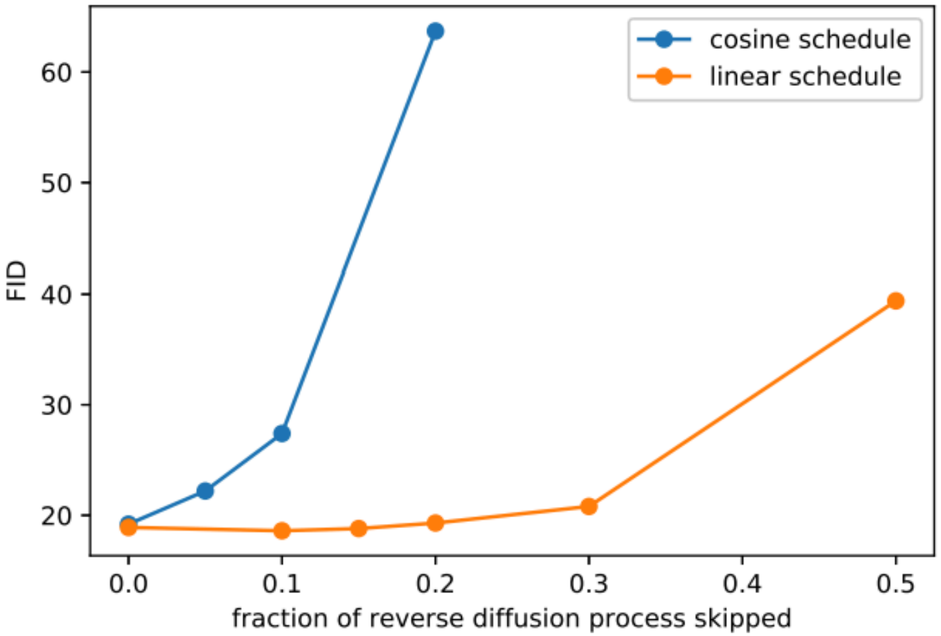
1. Denoise
2. Classification

Data

1. Data Source
2. Preprocessing

Basic DDPM

- Cosine Noise Schedule
- Foward Diffusion Proecess
- Reverse Process - x_0 prediction
- Conditional Scoring via Diffusion Consistency



Model

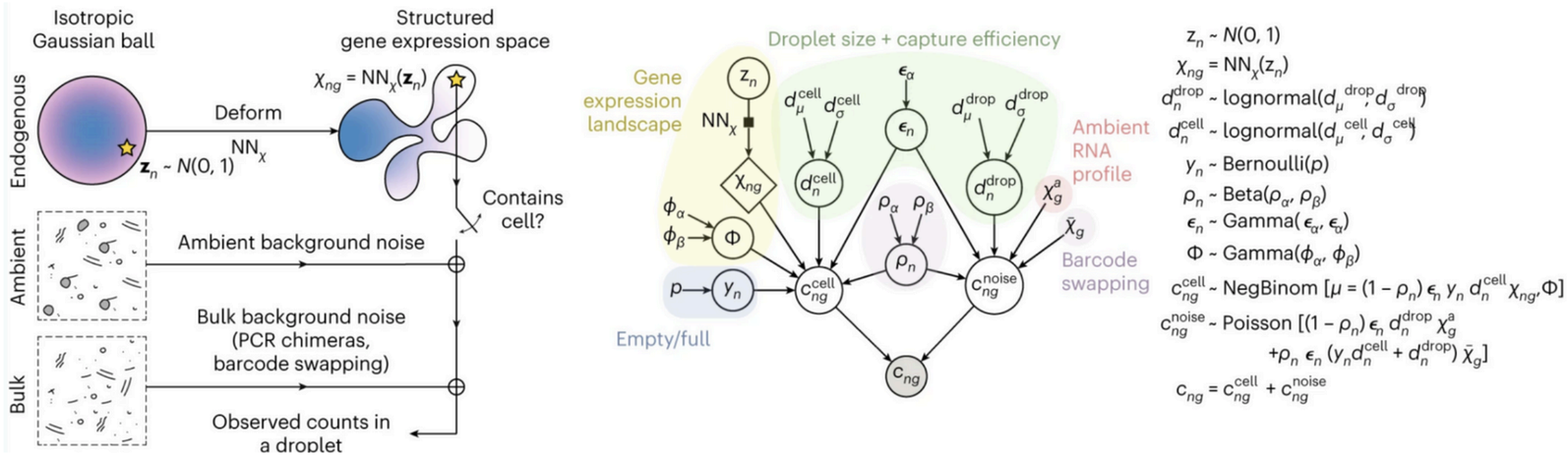
1. **Denoise**
2. Classification

Data

1. Data Source
2. Preprocessing

Binomial Likelihood

- Benchmarking CellBender
- **Likelihood-based scoring complements diffusion consistency**
- **Method focuses on label consistency under competing hypotheses.**



Cell-Bender - A generative model with likelihood-based decomposition

Model

1. Denoise
2. **Classification**

Data

1. Data Source
2. Preprocessing

Gaussian Mixture Model (GMM)

1D Gaussian Mixture Model (K=2)

- Posterior Probability
- Thresholding(0.99)

Binary Call per HTO

- Score > threshold → Positive
- Score ≤ threshold → Negative
- Result: binary HTO matrix (cell × HTO)
- Benchmarking Seurat's.

Model

1. Denoise
2. Classification

Data

1. Data Source
2. Preprocessing

Gaussian Mixture Model (GMM)

- 10X Genomics Dataset
- Peripheral blood mononuclear cells (PBMCs) from a healthy 19 year old female donor

	Gene 1	...	Gene m
Cell-1	12	...	2
...	...	....	...
Cell-n	0	....	5

GEX Count Matrix

	CMO301	CMO302
Cell-1	300	4
...	...	...
Cell-n	2	10

ADT Count Matrix

- GEX: how much each gene is expressed in each cell
- ADT: measures proteins on the cell surface instead of genes
- CMO: barcode to identify which sample each cell came from

Model

1. Denoise
2. Classification

Data

1. **Data Source**
2. Preprocessing

Cell Ranger Multi

- **Assignment:**
  - Singlets (CMO301, CMO302)
  - Background (Blank, Multiplet)
  - Unassigned
- **Pooled Multiplexed Sample using Cell Ranger**
  - Estimated Number of Cells: 13,446
  - Cells Assigned to a Sample: 12,577

	CMO301	CMO302
Cell-1	300	4
...	...	...
Cell-n	2	10

ADT Count Matrix

Multi

	Assignment
Cell-1	CMO301
...	...
Cell-n-1	Blank
Cell-n	Unassigned

Confidence Analysis

Model

1. Denoise

2. Classification

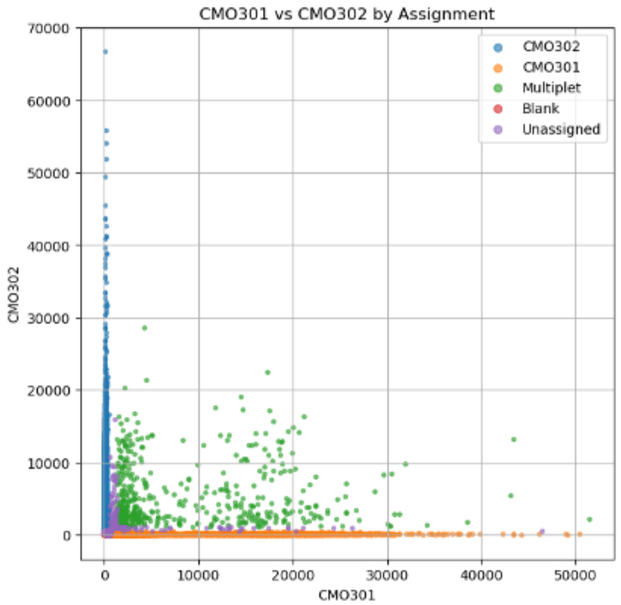
Data

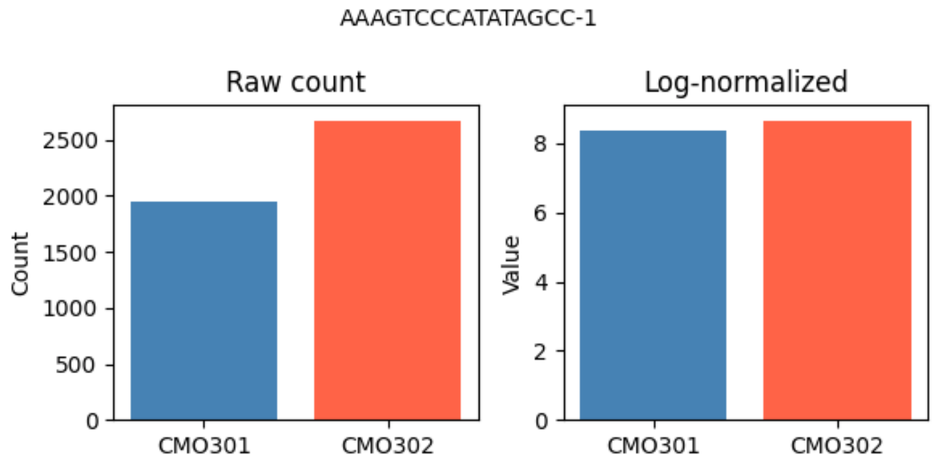
1. Data Source

2. **Preprocessing**

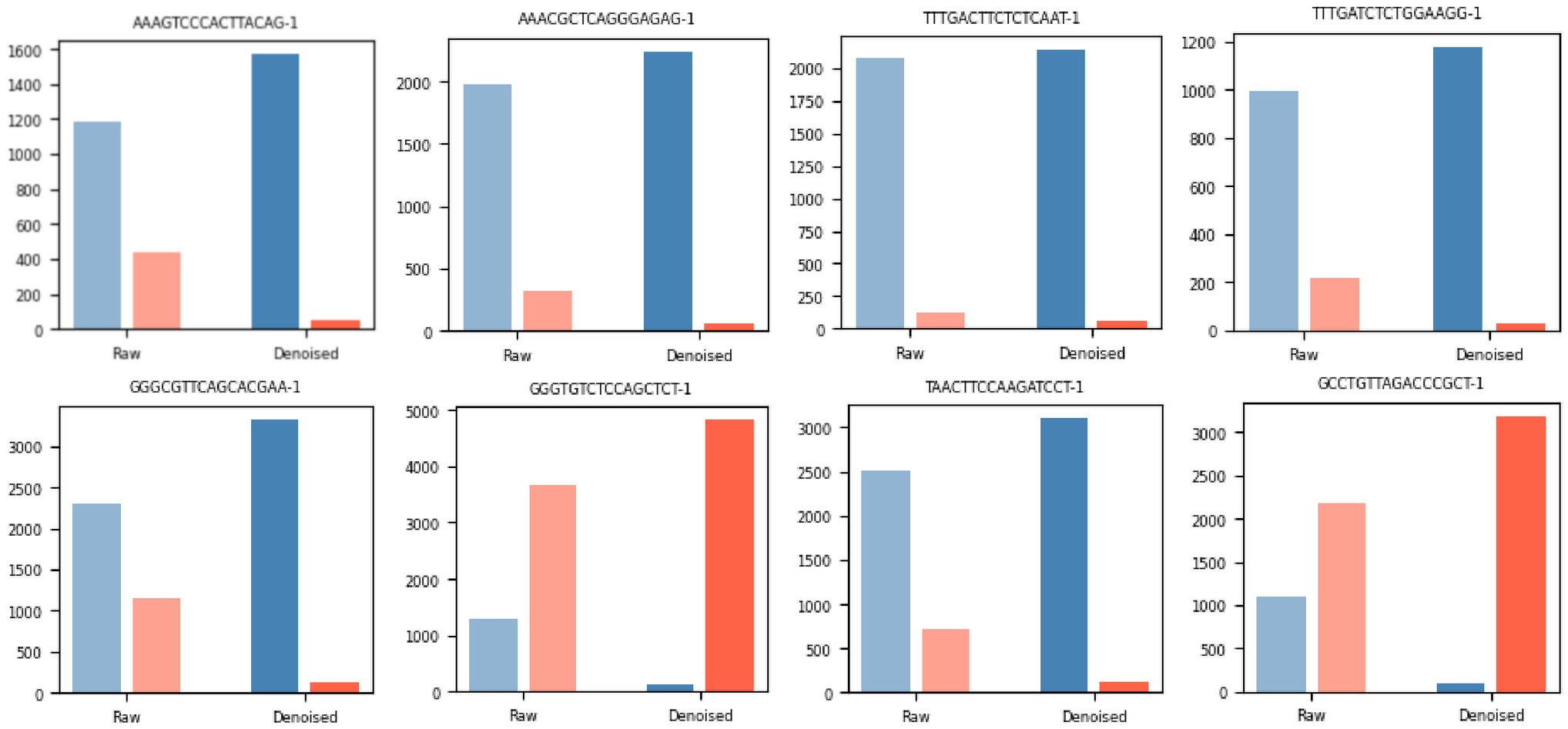
Normalization

- **Log-normalization**
  - CMO counts were adjusted by scaling each cell to a total of 10,000 and applying a log1p transformation.
  - This method stabilizes scale and variance while maintaining the relative proportions between CMO301 and CMO302.

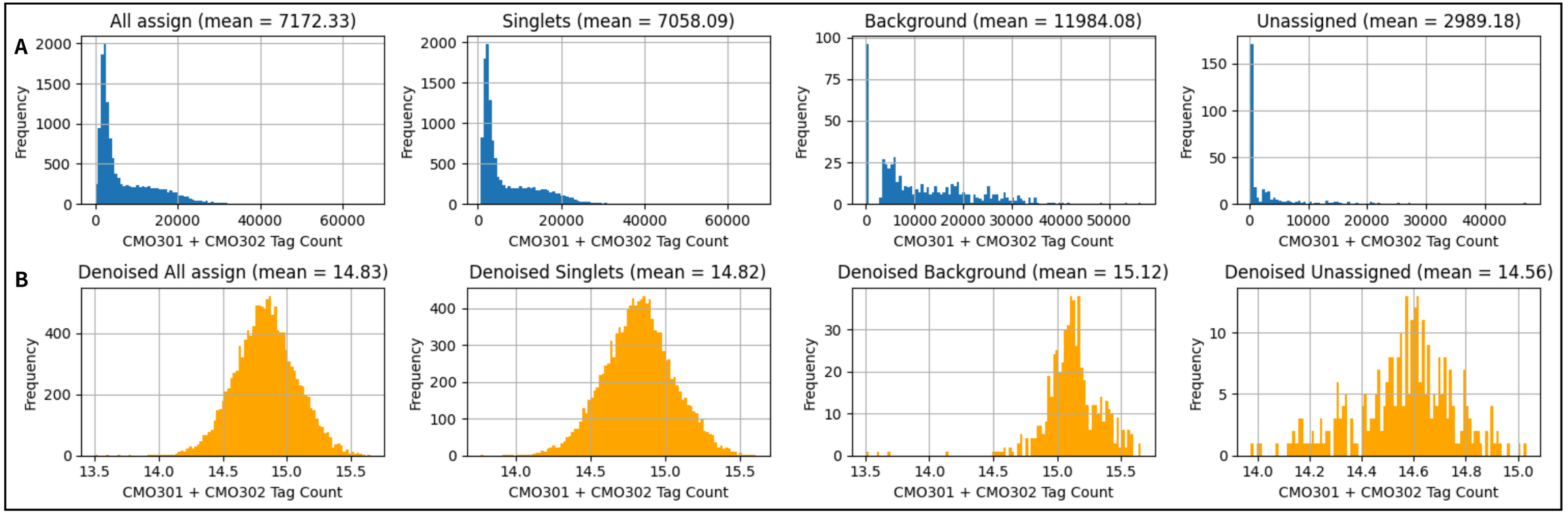




1. Denoising Result

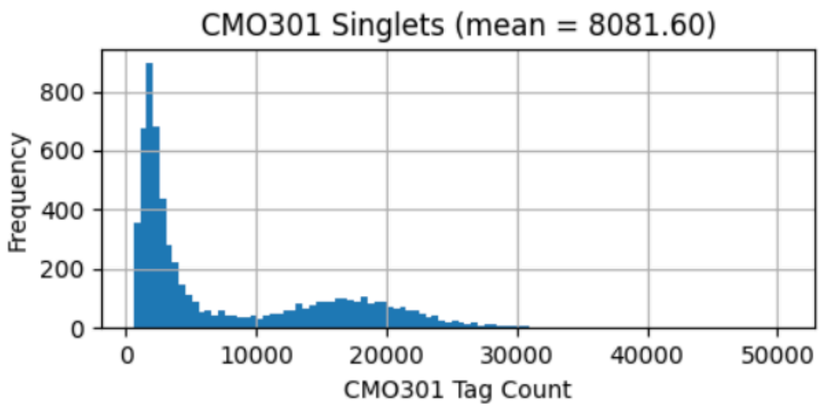


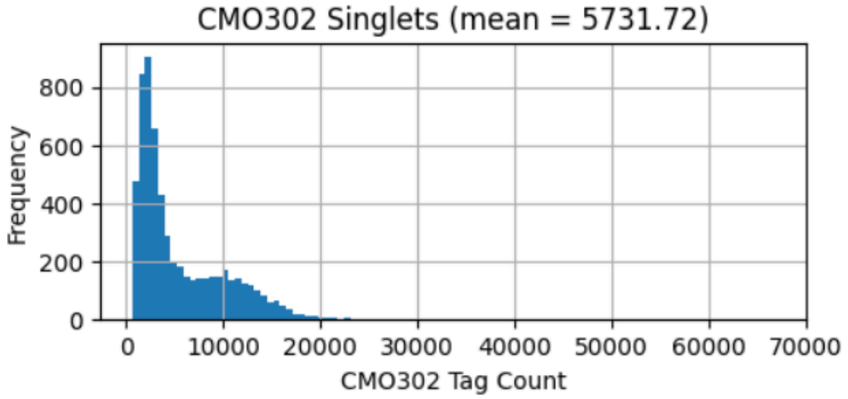
1. Denoising Result

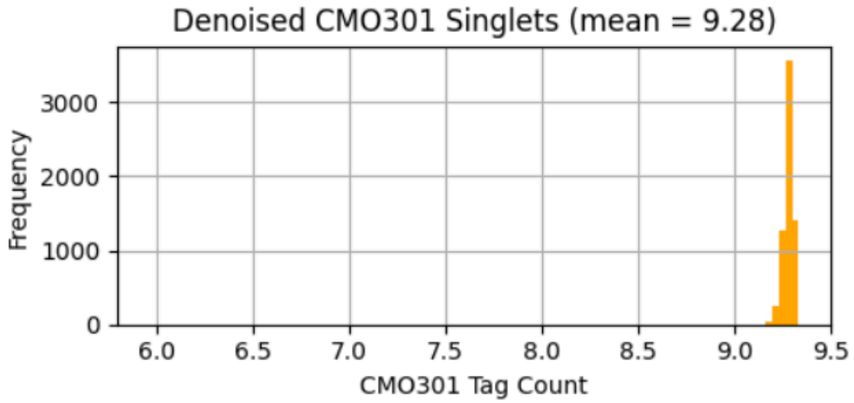


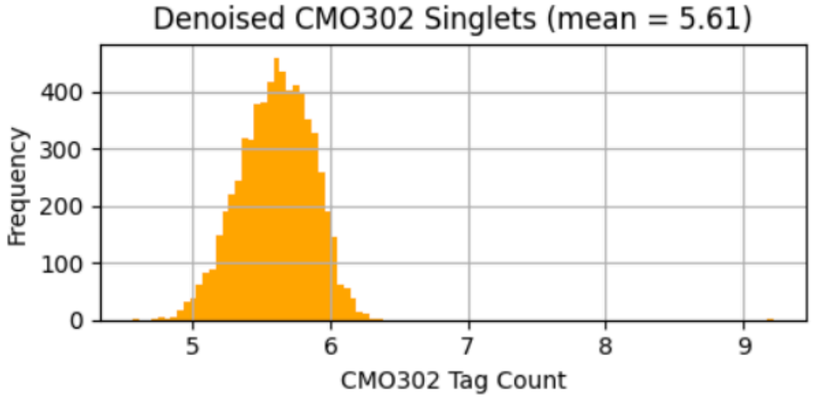
A - Raw CMO count(X axis) **B - Log CMO count(X axis)**

1. Denoising Result



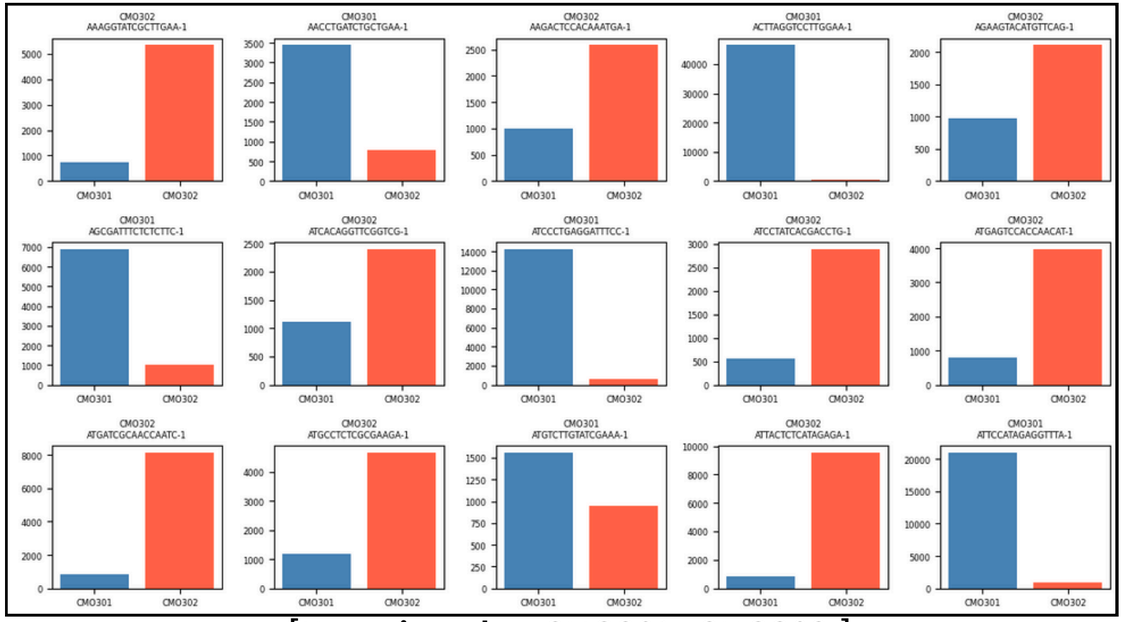




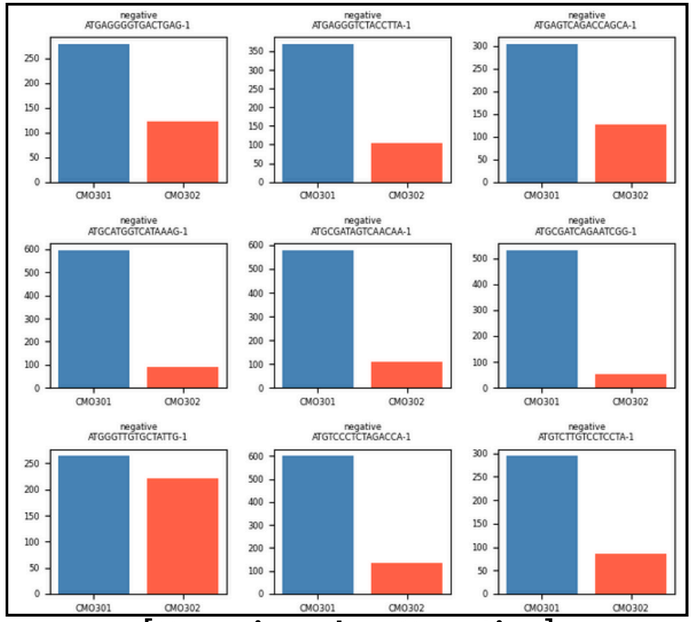


2. Resurrection of Unassigned (n = 308)

Assignment	Unassigned
HTO_call	
CMO301	41
CMO302	42
negative	225



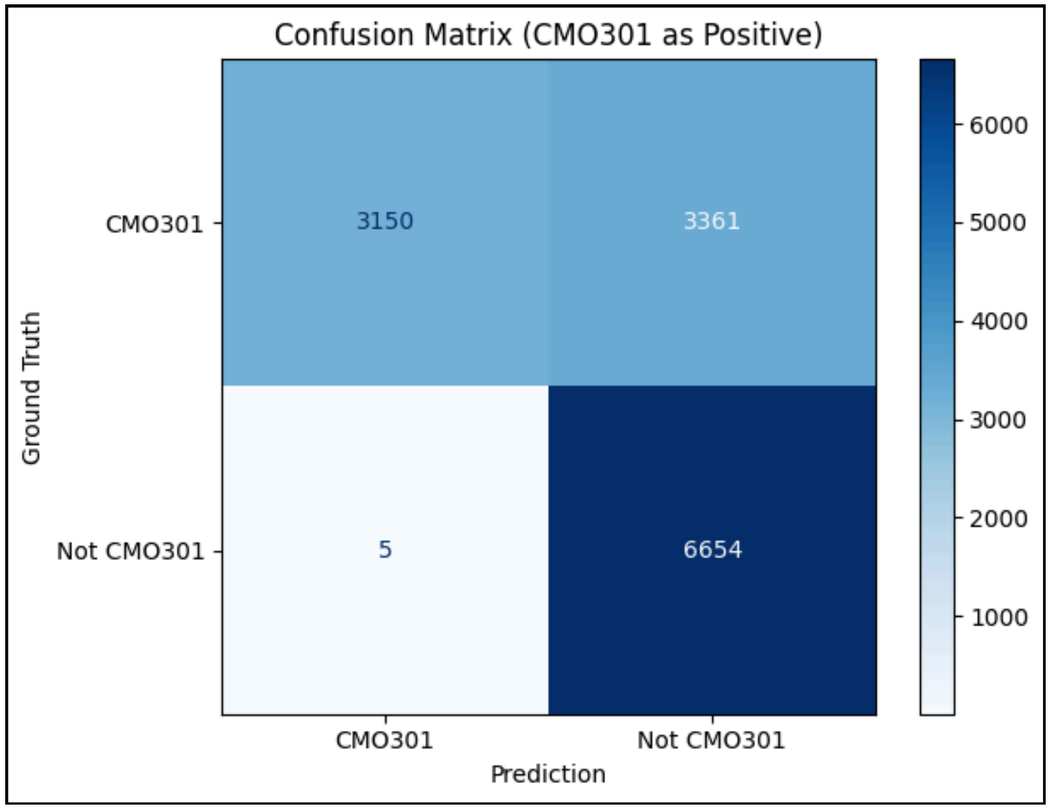
[Unassigned → CMO301, CMO302]

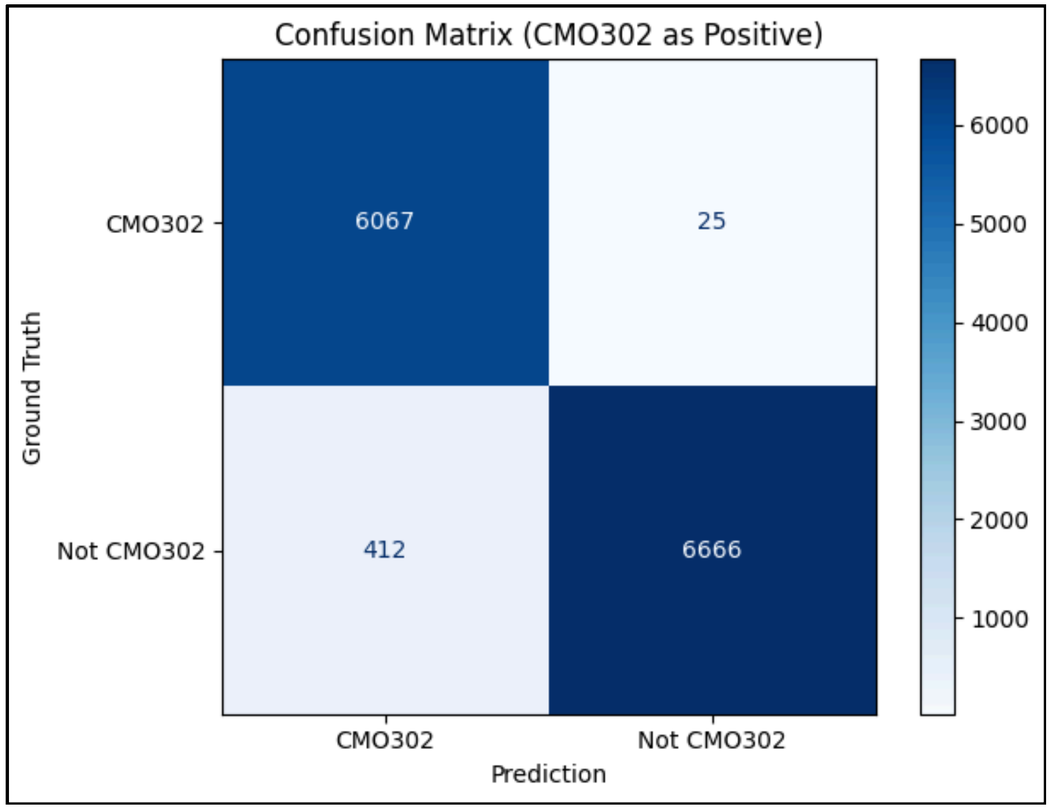


[Unassigned → Negative]

Diffusion reconstruction reveals that a subset of previously unassigned cells contains recoverable CMO301/CMO302 signals and most were classified as negative cells.

3. Classification Per Assignment





The diffusion model preserves HTO signal separability, but the conservative GMM 0.99 cutoff causes many true CMO301 cells to be classified as negative.

Cell Ranger

#The Strict Judge
#High Precision

Raw Noisy Input → conservative decision boundary

A strategy that selects only high signal-to-noise ratio (SNR) cells while discarding the rest

DeMUXly

#The Restorer
#High Yield & Recovery

Refined Input → clearly clustering and selective resurrection

Diffusion training learns global confounders → Recovers low-SNR cells

Limitations

- Validated primarily on 2-sample cases (Lack of extensive testing on multi-sample (>3) scenarios)
- GMM assumptions (Gaussian) may not fully align with Diffusion latents, potentially failing to capture overdispersion.
- Pipeline is not yet fully end-to-end self-supervised.
- Potential risk of doublet misclassification

Future Directions

- Perform benchmarking on high-complexity datasets to ensure scalability and robustness.
- Develop an End-to-End Learning framework
- Explore non-parametric clustering
- Assess rescued populations' impact on Downstream analysis clustering and differential expression.

Implication

Recovering discarded cells via generative denoising

Future Work

Achieving end-to-end optimization and faster inference

Use Case

Noisy datasets requiring maximum cell recovery

DeMUXly

Thanks for your attention

If you have any question,
please do not hesitate to ask.

Email
nayoungku1@gmail.com
juyoung4805@naver.com

GitHub
TBA