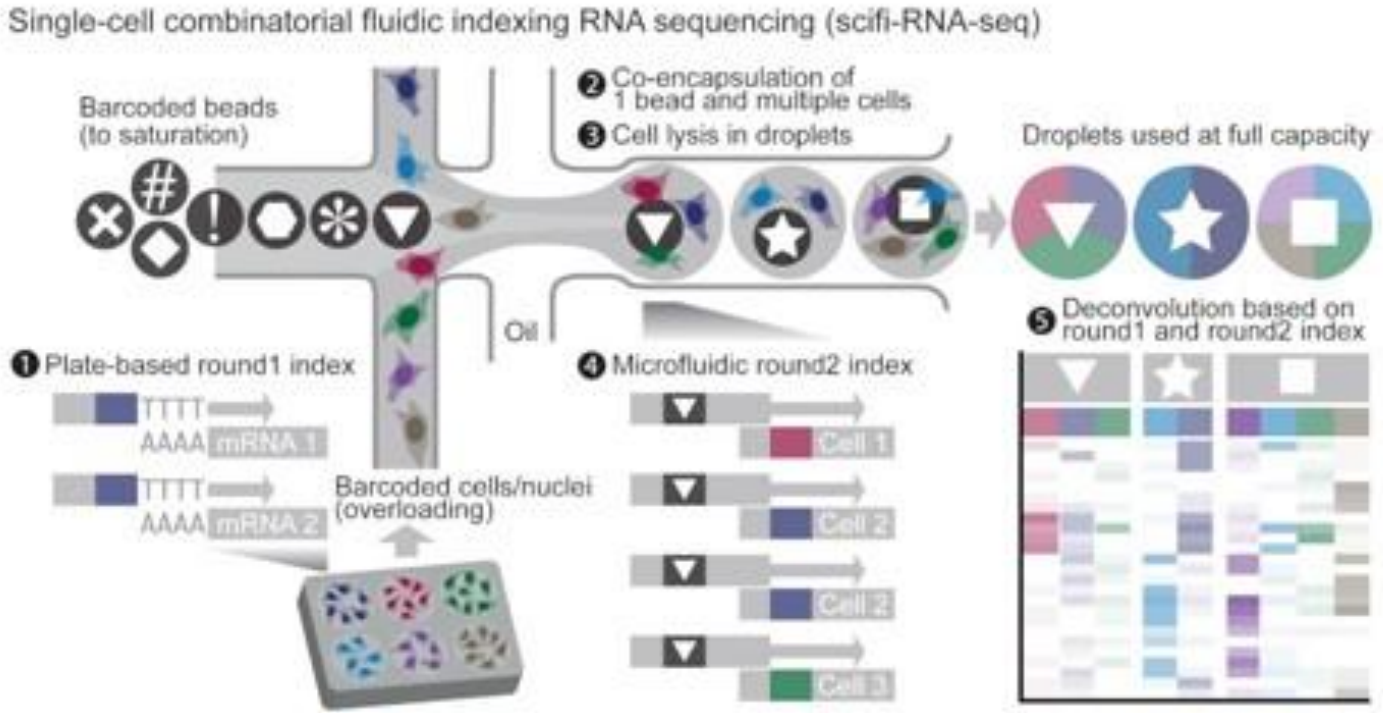
Identification of empty droplets in combinatorially indexed single cell RNA-seq data

Supervisor: Kholmatov Maksim
Students: Suleimanov Shakir, Lukina Maria, Grigoriants Vladimir

Introduction to single cell combinatorial indexing



Datlinger, P., Rendeiro, A.F., Boenke, T. et al. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. Nat Methods18, 635–642 (2021). https://doi.org/10.1038/s41592-021-01153-z

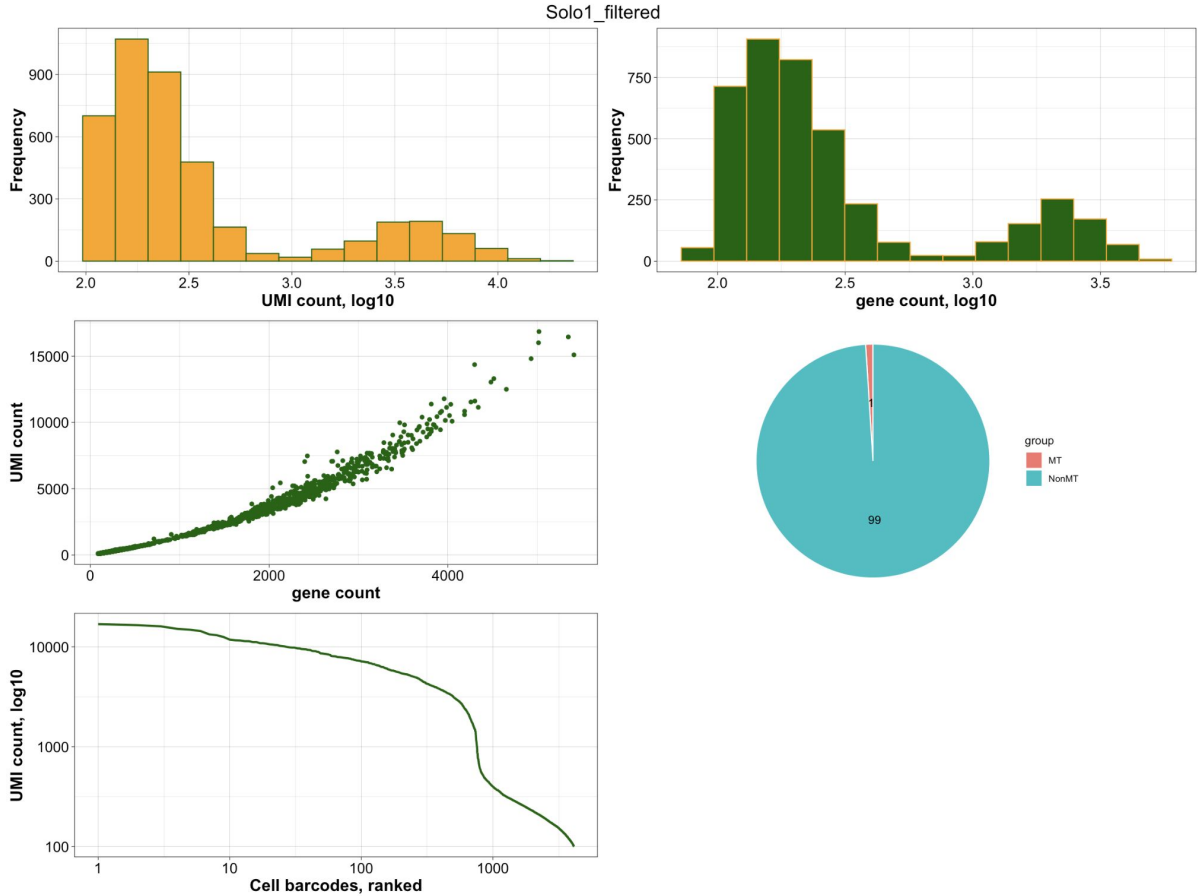
Aim and objectives

Specific aim: To assess the impact of uneven coverage distribution among samples on droplet classification

Objectives:

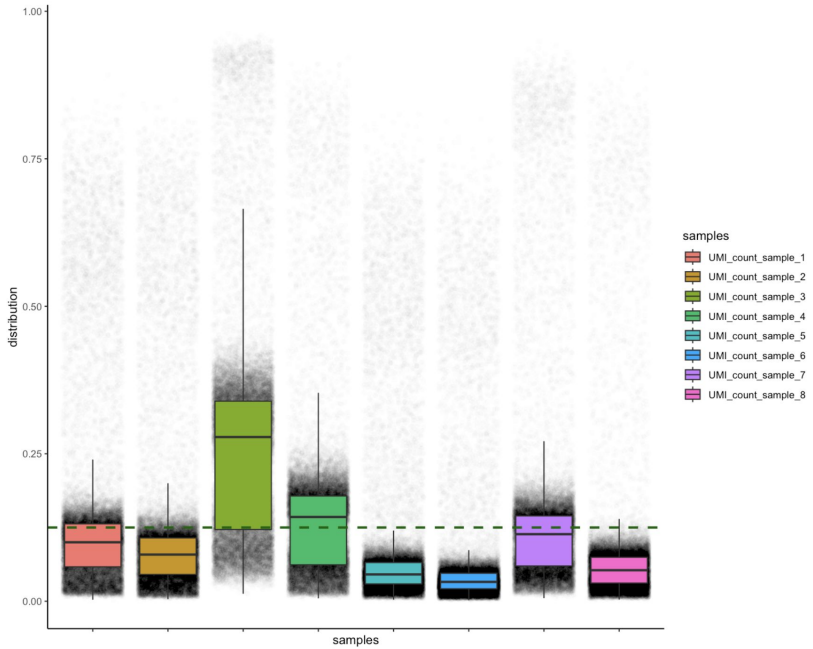
1. Get familiarized with the combinatorial indexing methodology in single-cell sequencing experiments and combinatorial indexing approach
2. Evaluate the distribution of UMIs across sample indices within droplets containing more than one nucleus.
3. Visualize the joint distribution of total coverage and the proportion of UMIs originating from each sample.
4. Identify samples exhibiting consistently lower coverage levels.
5. Investigate the impact of uneven UMI distribution between nuclei from different samples on EmptyDrops output.

Basic quality metrics for samples

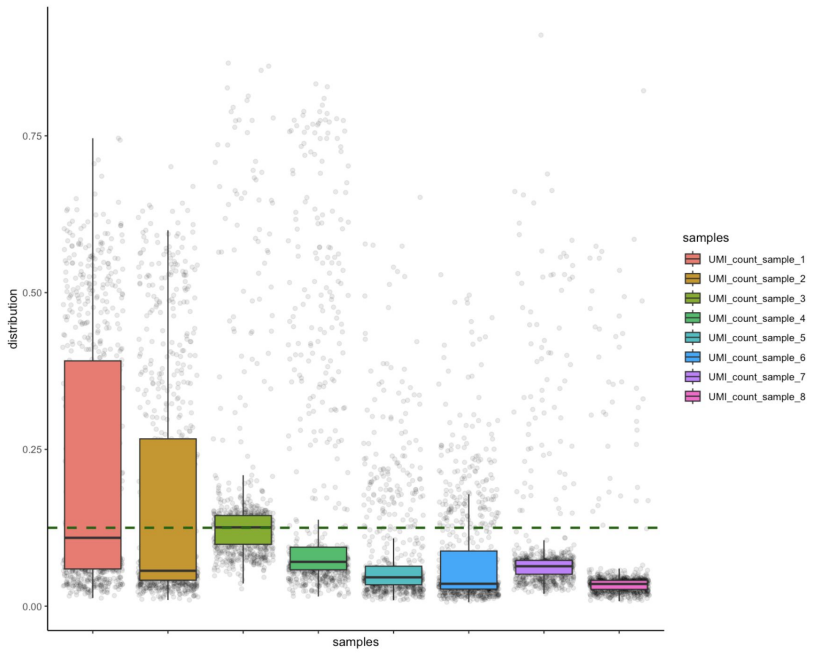


Results. UMI counts are uneven distributed in barcodes of all 8 samples

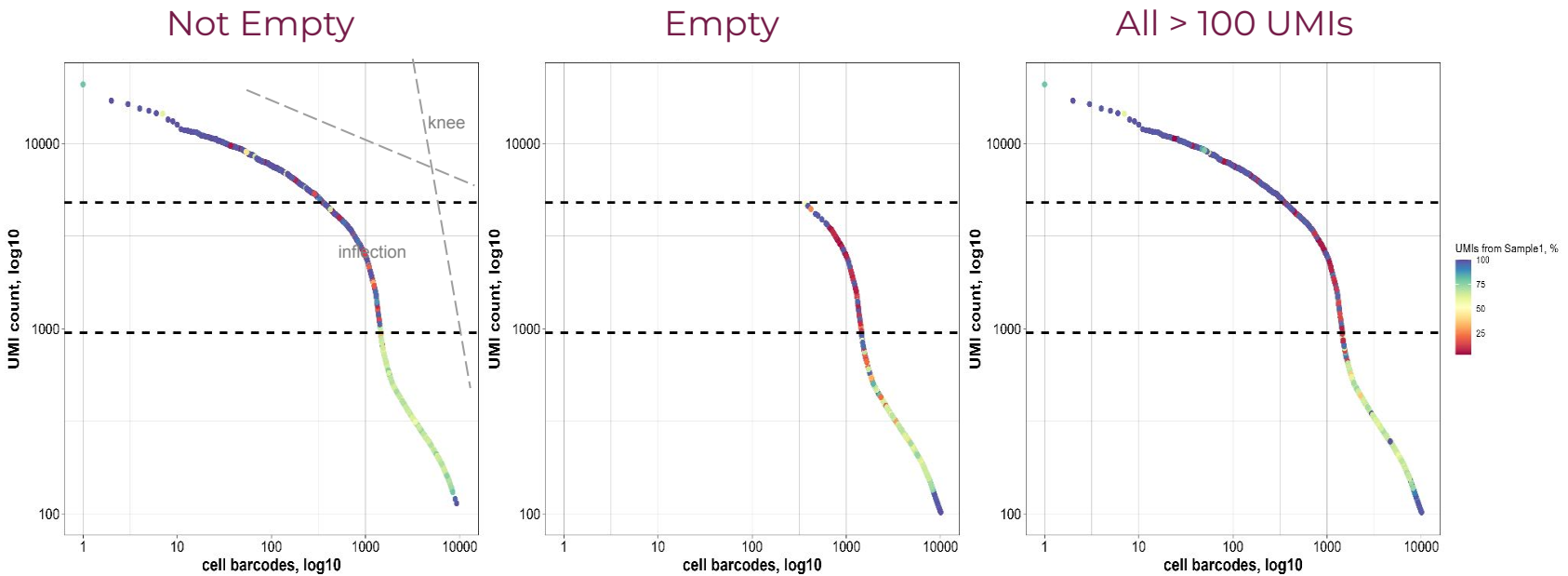
UMI distribution before filtering



UMI distribution after filtering

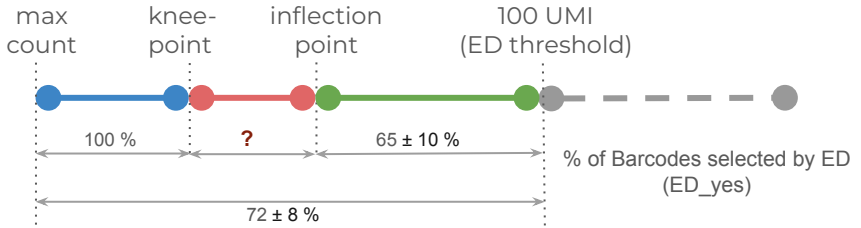


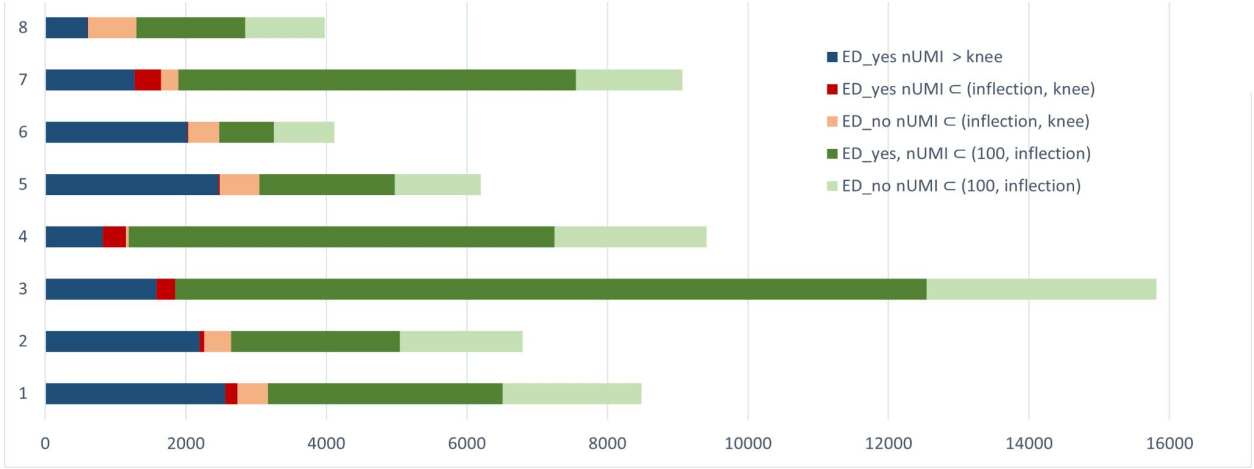
UMI uneven distribution impacts two-sample based EmptyDrops classification of non-empty/empty droplets



Inconsistency in EmptyDrops filtration in different samples occur mostly between inflection and knee points

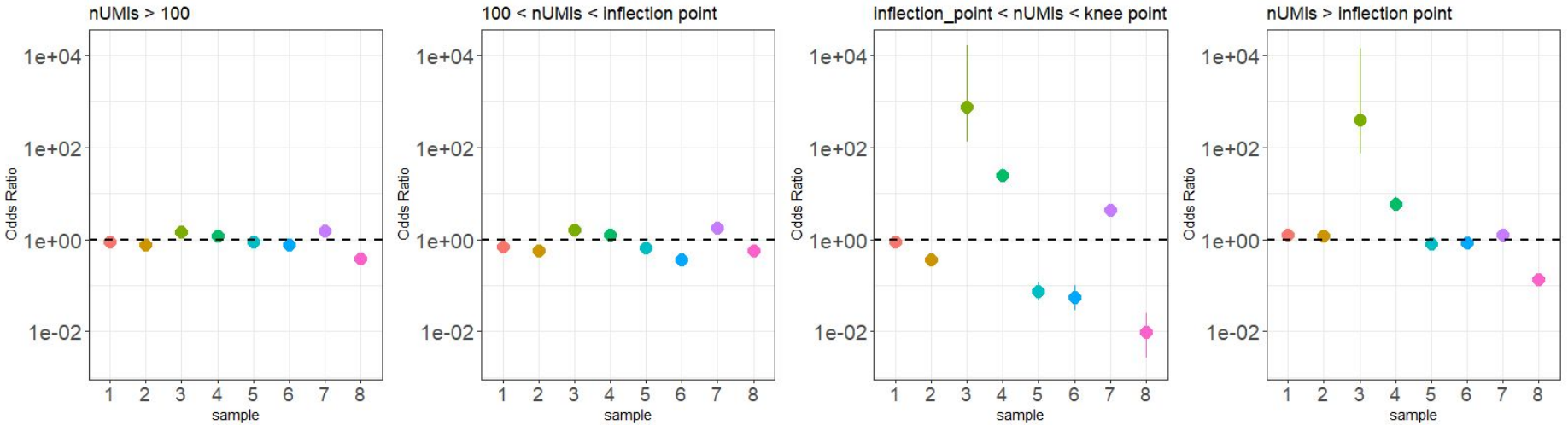
Empty and not empty barcodes distribution in different ranges





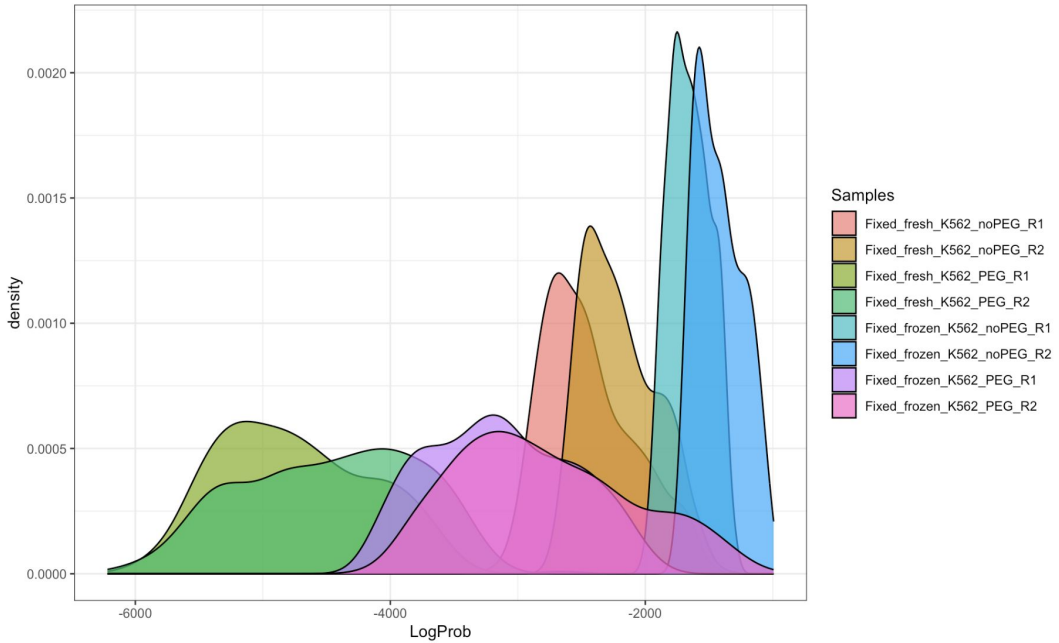
Number of selected barcodes within inflection-knee range varies dramatically between samples

Odds ratio for each sample vs “summarized” sample



Confidence of EmptyDrops in classifying droplets within the intermediate range varies across samples, correlating strongly with sample coverage

Distribution of EmptyDroplets Log Probabilities



Conclusions

https://github.com/maxim-h/bi-kho2-2024

- **Development of Analysis Functions:** We created functions to analyze basic quality metrics of individual samples in SUMseq.
- **UMI Distribution Hypothesis:** Tested if UMI distribution across barcodes from different samples aligns with the overall distribution or that of empty droplets.
- **Knee Plot Efficacy:** Knee plots effectively highlighted differences in UMI distributions and identified potentially full droplets.
- **Classification Variability:** Significant variability in droplet classification in intermediate UMI regions indicates the need for more sensitive and computationally efficient methods.
- **Sample-Specific Variations:** Discrepancies in droplet detection suggest that factors beyond coverage influence classification accuracy, necessitating further investigation.

Future Directions:

- Analysis of the differences between samples excluding the impact of the sample coverage
- Analysis of the distribution of the barcodes from several samples regarding inflection and knee point of the samples
- Tuning the parameters of the EmptyDrops analysis regarding the individual sample metrics