**Purpose**:

This document provides a write-up for class #5 homework the course “*Setting Bioinformatics Pipelines”.* The homework revolves around the following:

Instead of using the counts from Cellaranger-arc, use MACS2. Compare the obtained peaks. Focus on the QC control and the quality, number and distribution of the peaks. Is there any difference? Any clear advantage?

**Dataset(s):**

Multiome

**Results**:

Calling peaks from ATAC-seq data is the “first” step in the analysis pipeline. However, there are limited computional methods for this task in single-cell multiome. Here we are basically comparing two of the most used tools; MACS2 & Cellaranger-arc. In the pipeline, the two methods were applied and the generated peaks (counts) matrices were compared in terms of the number of peaks, and the distribution of the peaks.

We can see MACS3 producing more peaks than the other method, and about 60% of MACS3 peaks are shared (Figure 1). Also, the advantage of using Cellaranger-arc that it computes the “counts” fatser and in more details (Figure 2). For the distrubution, randomly subsets (~ 100) were selected to computed the distribution of the peaks due to computing resources. Based on the limited sample size that was tested, we can see somewhat comparable genome distrubution for the peaks using the two methods (Figure 3).

**Homework Limitation**:

The comparison was done using one dataset (multiome). Also, the peaks distribution comparison lacks quantification assessment. No comparison was done using the RNA dataset and how that would relate to or affect the study results.



Figure 1: Peaks numbers.

A diagram of a number of data

Description automatically generated with medium confidence

Figure 2: Counts by genome position.



Figure 3: Peaks distribution.