CloudMapper: Accelerating Single-Cell RNA Sequence Alignment with a Scalable and User-Friendly Cloud-Based Platform

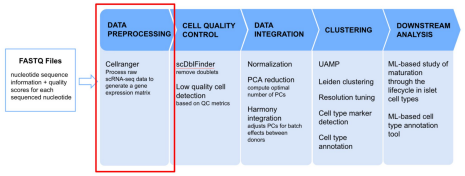
Kunwoo Park[1], Jiadong Bai[1], Yutong Lei[1], Seul Lee[2], Dongliang Leng[2], Jing Zhang[1], Wei Wang[3], Shuibing Chen[2], Chen Li[1]

[1]University of California Irvine, Irvine, CA; [2]Weill Cornell Medicine, New York, NY; [3]University of California Los Angeles, Los Angeles, CA

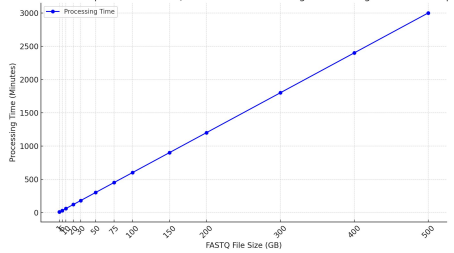
Introduction

Problem Statement

- **Single-cell RNA sequencing (scRNA-seq)** is a powerful method used to study the differences between individual cells, but the computational process of aligning RNA sequences to reference genomes remains a **major bottleneck**.
- Traditional alignment tools, like **CellRanger**, often require hours to days to process large datasets, even on high-performance computing (HPC) systems.
- Existing alignment workflows **demand significant technical expertise** to configure parallel computing environments like **SLURM** and manage multiple worker nodes.



Relationship Between FASTQ File Size and Processing Time in Single-Cell RNA-Seq



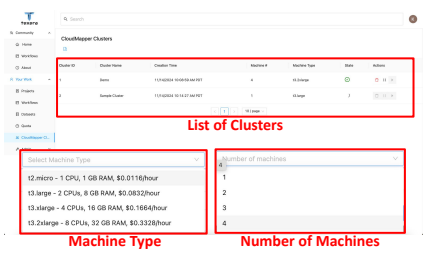
Objectives

User-friendly Web UI: Providing a cloud-based platform that enables bioinformaticians to accelerate scRNA-seq alignment using web services

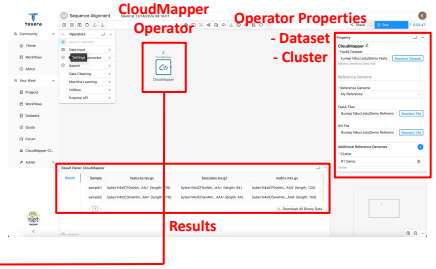
Elastic Usage of Public Cloud: Allowing users to dynamically construct a computing cluster based on the size and complexity of their datasets

How It Works

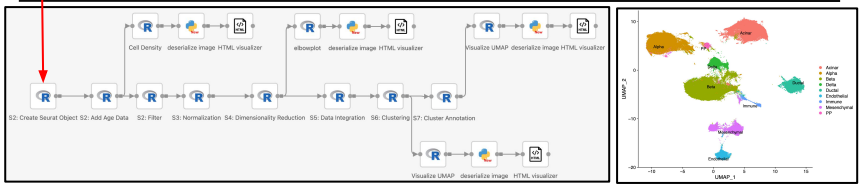
Launching Cloud Resources



Run Sequence Alignment

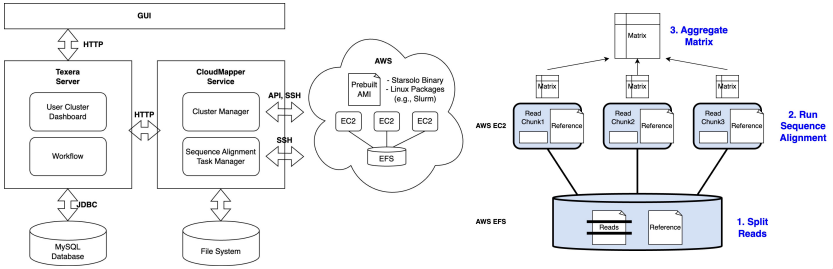


Downstream Data Analysis Pipeline



Architecture

- With just a few clicks, users can launch a cluster of multiple EC2 instances, without needing to worry about complicated settings, e.g., SLURM, network configurations, security policies, or storage setups.
- CloudMapper facilitates parallel processing of large scRNA-seq datasets by dividing a dataset into chunks, distributing the processing tasks across multiple instances, and merging the results into a unified count matrix.



Results

- A single server: 4.2 hrs
- CloudMapper: 0.3 hrs with $3.19 (using 16 VMs)
  - No local servers needed

# of VMs	Time to start cluster (hrs)	Data transfer time (hrs)	Running time (hrs)	Total time (hrs)	VM cost ($/hr)	Total cost ($)
1	0.07	0.21	4.2	4.5	0.33	1.49
2			2.2	2.5	0.67	1.67
4			1.1	1.4	1.33	1.89
8			0.6	0.9	2.66	2.53
16			**0.3**	0.6	5.32	**3.19**

30GB FASTQ file, 12 reference gene ids, t3.2xlarge

Future Work

- **Generalize CloudMapper Beyond scRNA-seq**: expanding its capabilities to support other data.
- **Intelligent Configuration Recommendations**: introducing an intelligent recommendation engine that suggests optimal cluster configurations based on user-defined constraints like time and budget.

Want to Join the Platform?

Texera Platform:
https://texera.dknet-ai.org

Texera Github:
https://github.com/Texera/texera

Dknet Webinar:
https://youtu.be/B81iMFS5fPc

Scan me!

National Institute of Diabetes and Digestive and Kidney Diseases