Novel Graph-based Model Algorithms for scRNASeq Analysis

Project Name: Novel Graph-based Model Algorithms for scRNASeq Analysis
Project PI: Dr. Chao Zhang
Lab Contact: Yuchen Liu
Name: Saumya Pothukuchi
Project Start Date: 06/04/2024

Background

The lab is currently developing a novel graphical-based model for single-cell RNA-seq analysis (scGND).
Data integration in single-cell RNA sequencing (scRNA-seq) addresses the challenge of combining datasets from different sources, conditions, or technologies to uncover insights that are not discernible from isolated datasets. It is crucial due to the potential variability in scRNA-seq data, stemming from technical noise, batch effects, and biological diversity. By integrating data, researchers aim to enhance the robustness of biological conclusions, increase the statistical power of analyses, and reveal conserved or divergent patterns across conditions or species. Data integration enables the comparison and combination of single-cell datasets to explore cellular heterogeneity, developmental trajectories, and disease mechanisms on a scale and depth previously unattainable. The main challenge in data integration is to remove batch effects and technical noise from the real biological diversity. Many established packages, such as Seurat, scVI, Harmony and scGPT, approach the data integration problem from different aspects.
The performance of scGND needs to be evaluated and benchmarked.

Hypothesis

To evaluate the performance of this new method, the plan is to compare it with various scRNA-seq data integration packages using diverse benchmarking datasets. Prior to that, the novel deep learning-based method needs to be further optimized. Multiple packages will be run using both Python and R for comparative study with the novel method. Additionally, databases need to be benchmarked and key benchmarking scRNA-seq datasets for data integration tasks need to be selected.

Study Design

Collaborative project where 6 scRNASeq mouse samples were sequenced but only 5 are in use since the first sample was processed differently and skews the results. Data was collected from stem cells to study leukemia and hence the hotspot chosen is JAK2 and C57BL6J wild type.

Samples need to be analysed using Seurat package and data needs to be checked to be of appropriate quality using standard QC metrics. Additionally other packages such as MNN, fastMNN, scmap, limma, scMerge, LIGER, scGEN, MMD-ResNet, ZINB-WaVE, etc need to be learnt to benchmark datasets with scGND.

Deliverables

https://github.com/BU-BMSIP/Zhang-scGND

Projected Timeline

Week 1-2:

- Familiarisation with Seurat and running complete analysis on 5 scRNASeq samples

Week 3-5:

- Reanalysis with different parameters to ensure optimal results of samples. Familiarising with half of the tools to be compared. Selection of optimal datasets for benchmarking.

Week 6-7:

- Running analysis using previous tools with selected datasets. Familiarisation with rest of the tools.

Week 8-10:

- Running analysis with rest of the tools with same datasets. Compiling comparative study results.

References

Liu, Yu-Chen, et al. "scGND: Graph neural diffusion model enhances single-cell RNA-seq analysis." bioRxiv (2024): 2024-01.