Benchmarking single cell RNA sequencing integration
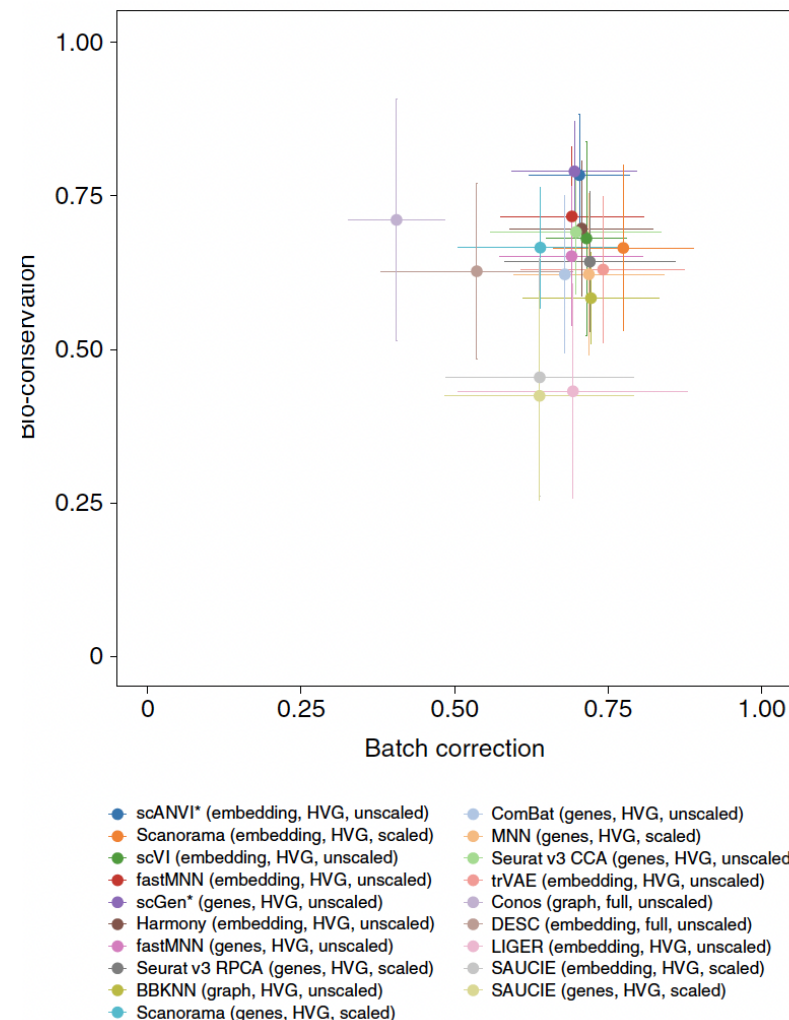
Xinran Miao
05/02/2022

Background

- **Dataset**: Single-cell RNA sequencing (scRNA-seq) datasets.
  - each dataset consists of a `cell` by `gene` expression matrix, where cells are annotated with batches and labels.

- **Goal**: Combining high-throughput sequencing datasets to produce a self-consistent version of the data for downstream analysis.

- **Reason**: Integrating scRNA-seq helps biological findings.

- **Challenges**:
  - Dealing with noise, sparsity, batch effects and rare cell types.
  - Evaluating integration methods.

Background

- (Luecken, Büttner, Chaichoompu, Danese, Interlandi, Müller, Strobl, Zappia, Dugas, Colomé-Tatché, and others, 2022) benchmarked 68 methods in 13 integration tasks.
- They used 14 metrics to evaluate integration methods on their ability to
  - remove batch effects, and
  - conserve biological variations.
- They provided guidelines to choose an integration method given a task.
- For all methods, they used default parameters.

Legend:
- scANVI* (embedding, HVG, unscaled)
- Scanorama (embedding, HVG, scaled)
- scVI (embedding, HVG, unscaled)
- fastMNN (embedding, HVG, unscaled)
- scGen* (genes, HVG, unscaled)
- Harmony (embedding, HVG, unscaled)
- fastMNN (genes, HVG, unscaled)
- Seurat v3 RPCA (genes, HVG, scaled)
- BBKNN (graph, HVG, unscaled)
- Scanorama (genes, HVG, scaled)
- ComBat (genes, HVG, unscaled)
- MNN (genes, HVG, scaled)
- Seurat v3 CCA (genes, HVG, unscaled)
- trVAE (embedding, HVG, unscaled)
- Conos (graph, full, unscaled)
- DESC (embedding, full, unscaled)
- LIGER (embedding, HVG, unscaled)
- SAUCIE (embedding, HVG, scaled)
- SAUCIE (genes, HVG, scaled)

Methods

- Integrating scRNA-seq datasets usually include two parts:

    - jointly **embedding** high-dimensional input onto a shared latent space, and
    - (soft) **clustering** cells that incorporates annotated information (e.g. cell type).

- In this presentation, we aim to study the dependence of integration methods on the dimension of latent space.

- We focus on three methods that build deep generative models:
    - scVI (Lopez, Regier, Cole, Jordan, and Yosef, 2018),
    - scGen (Lotfollahi, Wolf, and Theis, 2019), and
    - scANVI (Xu, Lopez, Mehlman, Regier, Jordan, and Yosef, 2021).

Evaluation

- Average silhoutte width (ASW) measures the separation of clusters.
- We use modified batch ASW and cell type ASW to evaluate the ability of batch removal and biogical conservation, respectively.
- For both, the larger ASW we have, the better.

Real-data analysis: human immune cell integration

- Task: integrating 5 datasets of 10 batches (donors) with cells from peripheral blood and bone marrow, annotated by cell types
- According to (Luecken, Büttner, Chaichoompu, et al., 2022), scANVI is one of the best methods, evaluated by a weighted mean of 14 metrics.
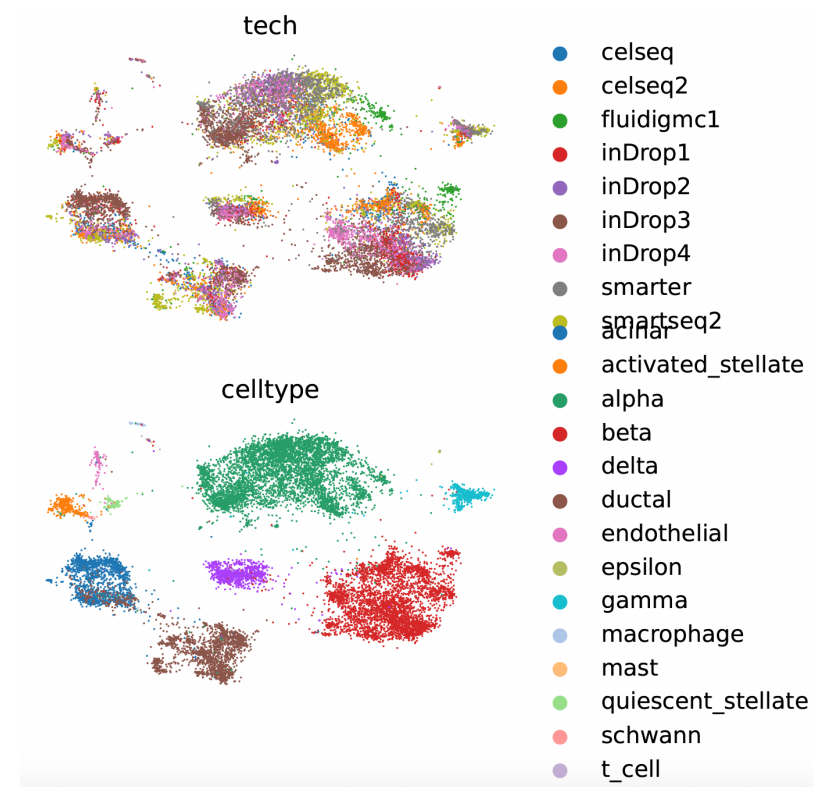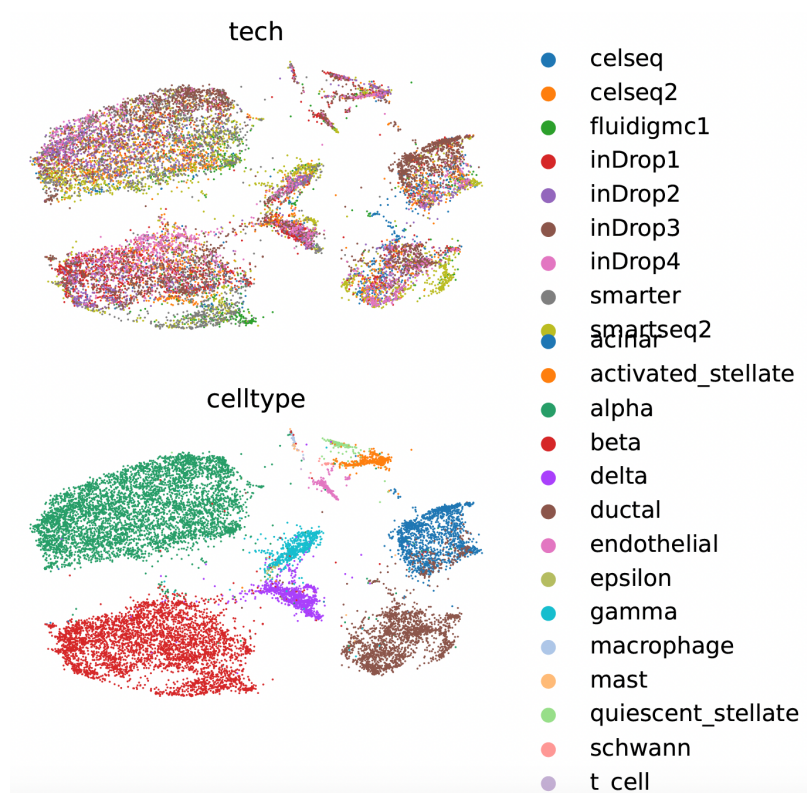
Results

- We vary the dimension of latent space, and evaluate `scvi`, `scanvi` and `scgen` using batch ASW and cell type ASW.
- Overall, `scgen` works worst.
- For `scvi` and `scanvi`, there is a trade-off between removing batch effects and conserving biological variations.
  - High-dimensional latent space leads to better batch removal and worse biological conservation.
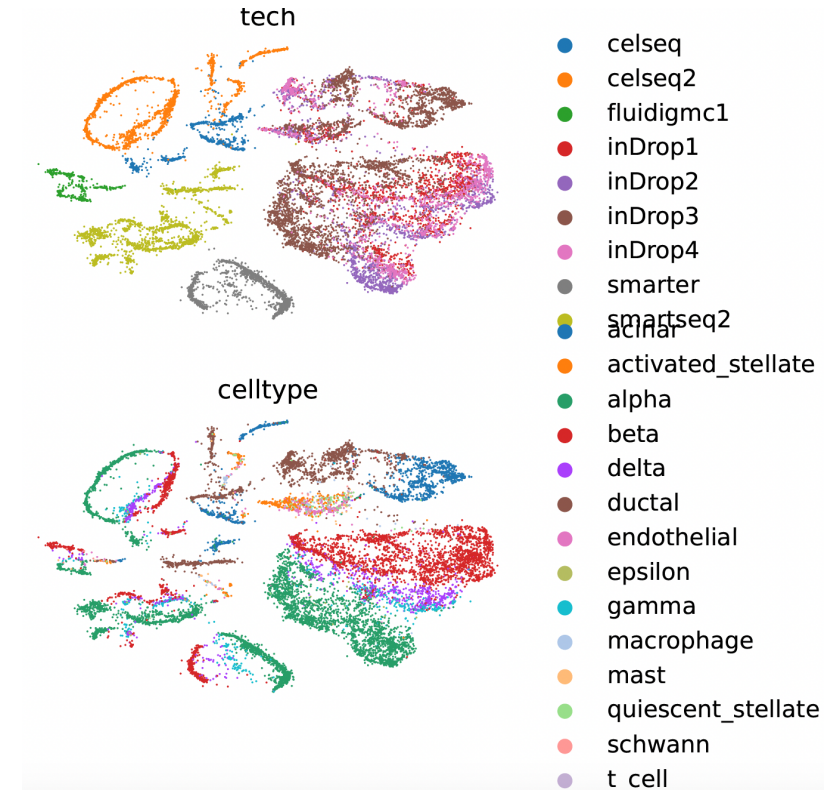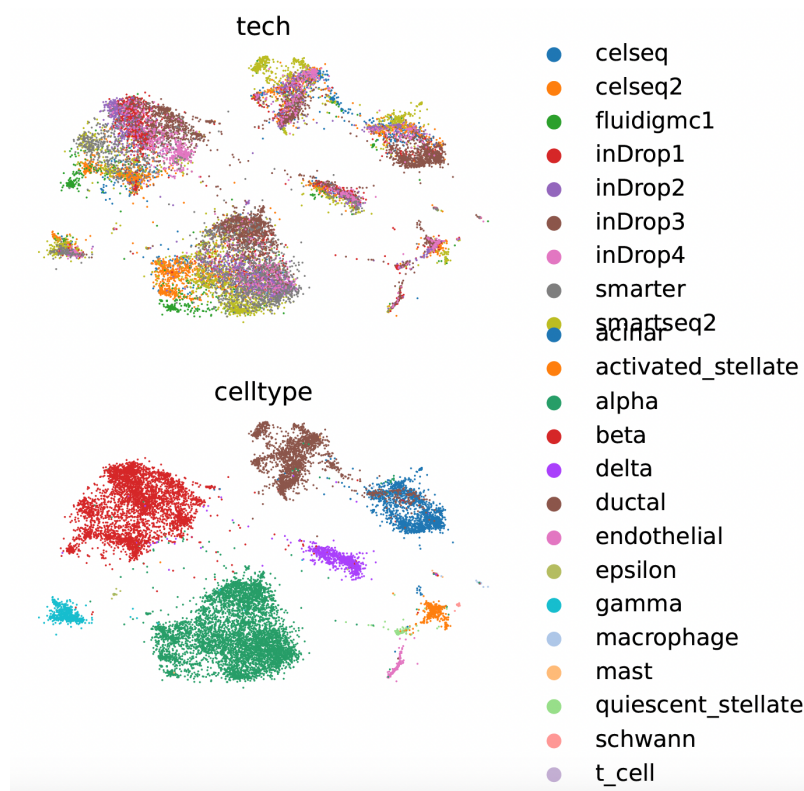 - The choice of latent space dimension doesn't affect the rank across methods.
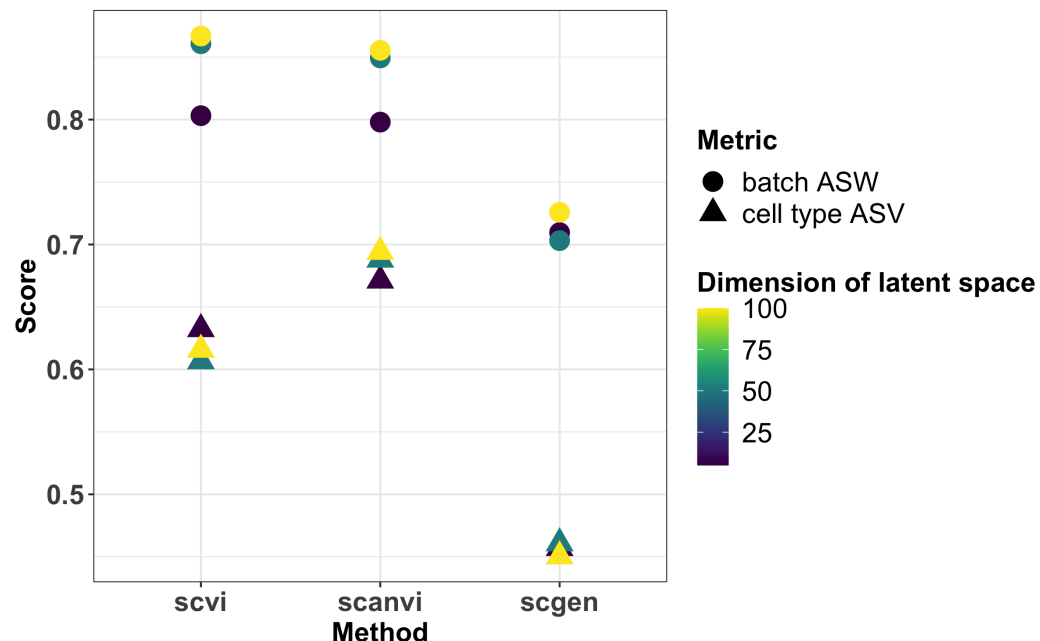
Results

Latent space fitted by `scvi` projected onto 2d space, with dimension of latent space (left) 5 and (right) 100, colored by (top) batches and (bottom) cell types.

Results

Latent space fitted by (left) scanvi and scgen projected onto 2d space, , colored by (top) batches and (bottom) cell types.

Takeaways

- For each method, the performance of integration methods is dependent on the dimension of latent space.
  - High-dimensional latent space leads to better batch removal and worse biological conservation.
  - That being said, it's hard to tell the difference in visualizations on 2d space.
- The choice of latent space dimension doesn't affect the rank across methods.

Reproducibility

- Results can be reproduced via https://github.com/XinranMiao/scRNA_int (not yet finished).

Thank You!
Questions?

References

Luecken, M. D., M. Büttner, K. Chaichoompu, et al. (2022). "Benchmarking atlas-level data integration in single-cell genomics". In: *Nature methods* 19.1, pp. 41-50.