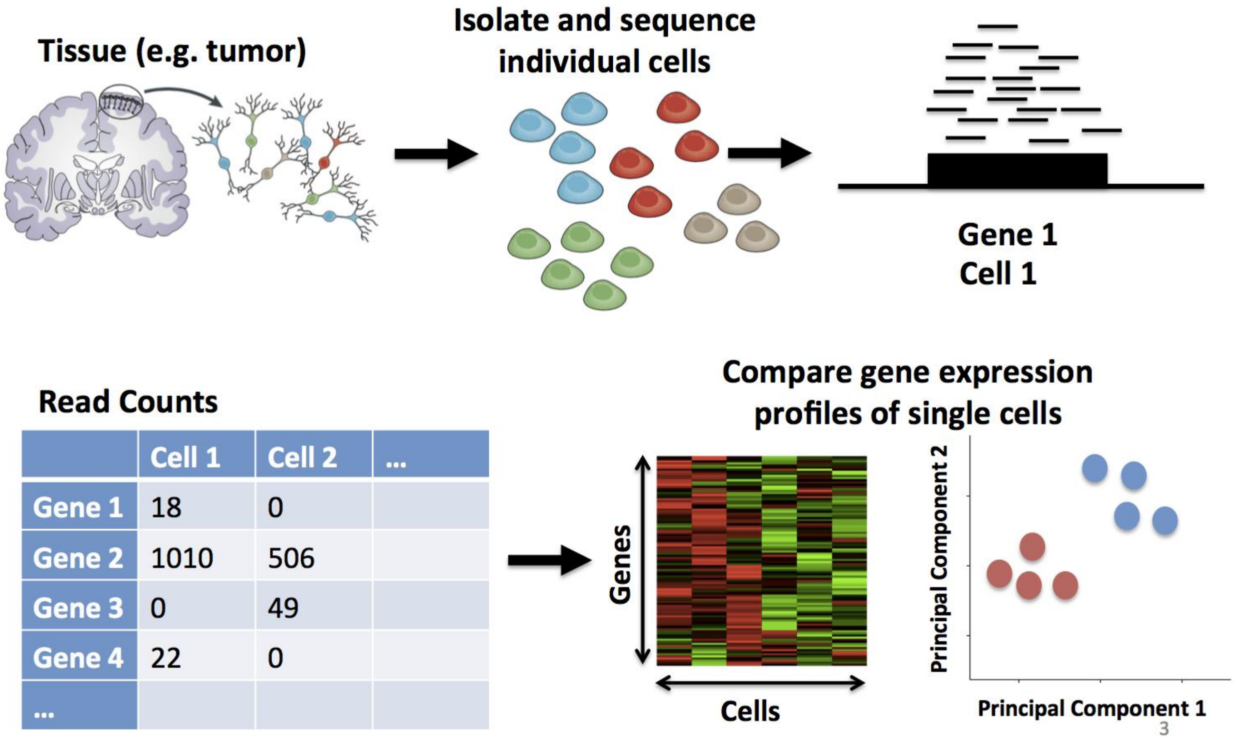
CIDER: an interpretable meta-clustering framework for single-cell RNA-seq data integration and evaluation

Hu, Ahmed & Yau, 2021
CSE590C - 2/7/21 (Ayse & Nicasia)

Single cell RNA sequencing (scRNASeq)

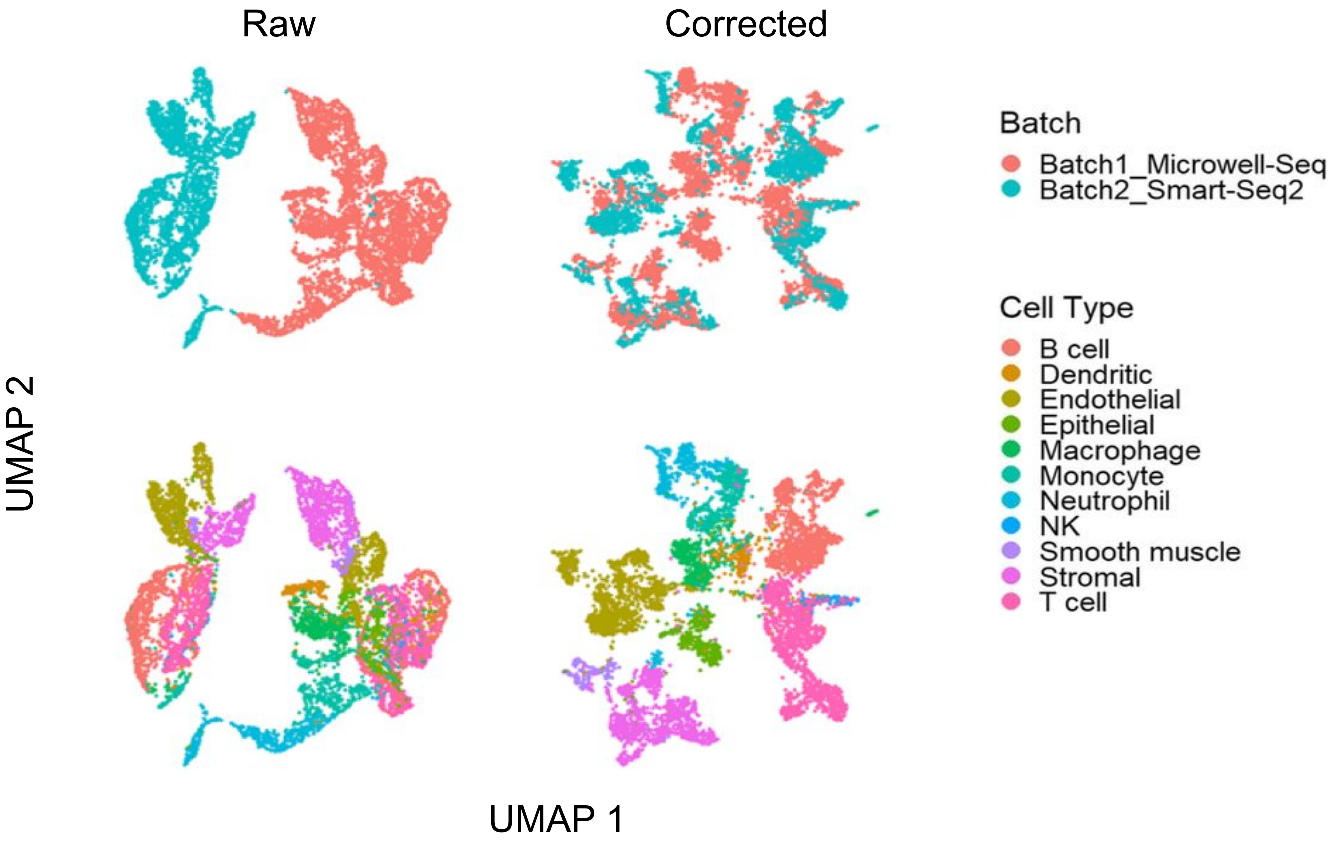

Tissue (e.g. tumor)
Isolate and sequence individual cells
Gene 1
Cell 1
Compare gene expression profiles of single cells
Genes
Cells
Principal Component 2
Principal Component 1

Read Counts

	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

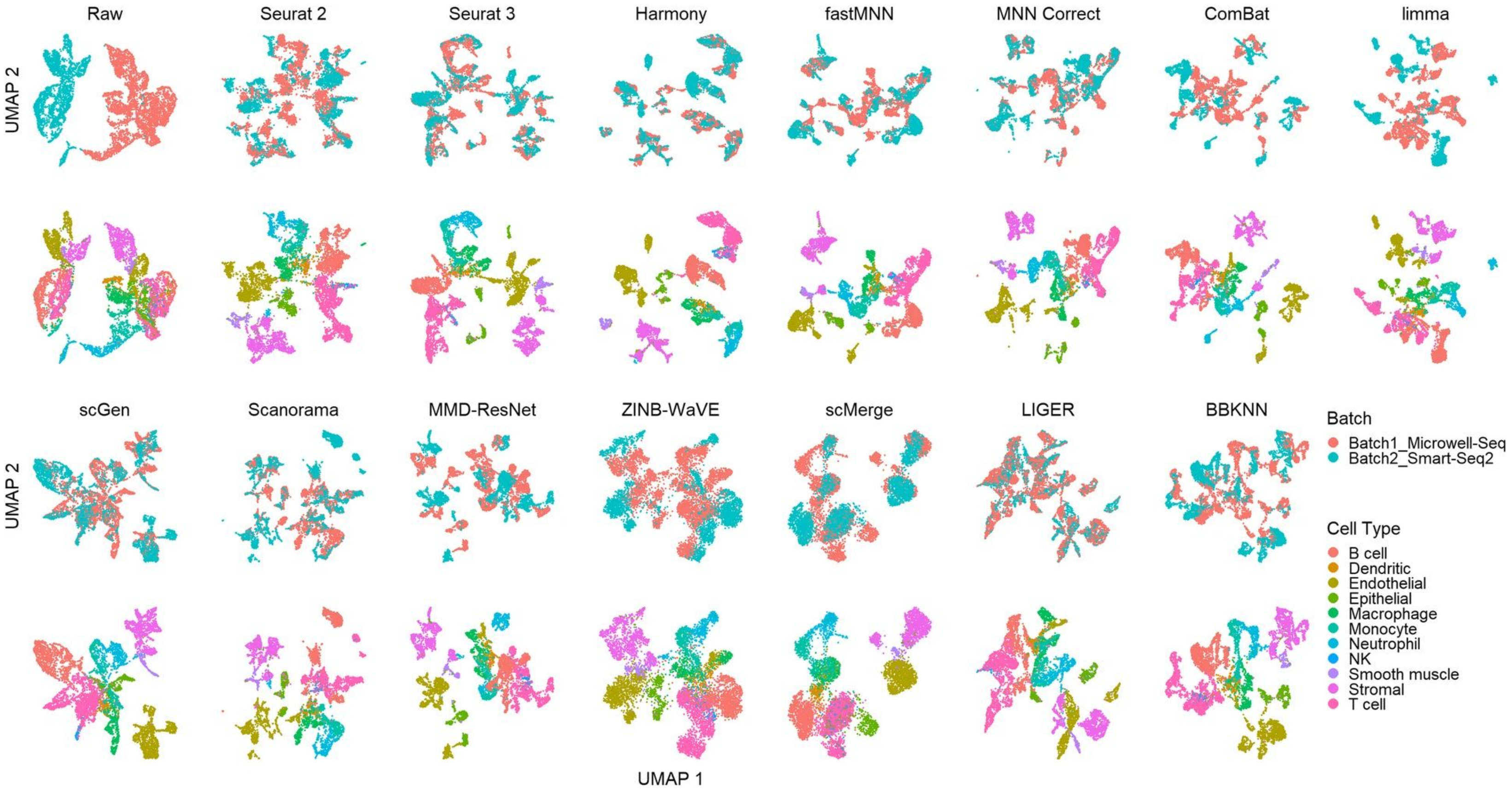
https://www.rna-seqblog.com/top-benefits-of-using-the-technique-of-single-cell-rna-seq/

scRNASeq - challenges with data integration



Tran, H.T.N., Ang, K.S., Chevrier, M. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 21, 12 (2020).

scRNASeq – current approaches



Tran, H.T.N., Ang, K.S., Chevrier, M. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 21, 12 (2020).

scRNASeq – current approaches

Clustering Workflows

Can identify cell populations in batch-effect- free datasets

Partition cells by inter-cell distance matrix using PCA or high variance genes (HGVs)

Examples: SC3, RaceID, Seurat v3

Performance degrades in datasets confounded by batch effects

Batch correction + clustering Workflows

Combines batch correction or integration methods and downstream clustering algorithms

Mutual nearest neighbors: Examples: Monocle3 pipeline, Scanorama, Seurat

Other approaches: Harmony, LIGER, ComBat, Conos

Performance can vary substantially across data types and scenarios

scRNASeq – current approaches

Limitations

Bias in initial selection:

- Integration algorithms work on the low-dimensional representation
- Can be affected by the bias in the initial selection of HVGs and PCs

Lack of interpretability:

- Difficult to determine why existing methods drive cells from different batches into the same cluster

To address these limitations, they introduced CIDER

CIDER contributions

1. New similarity metric: Inter-group Differential ExpRession (IDER) → clustering (CIDER)

2. Similar/superior performance compared with other clustering methods for scRNA-Seq data

3. CIDER as a ground-truth-free evaluation metric for other integration methods

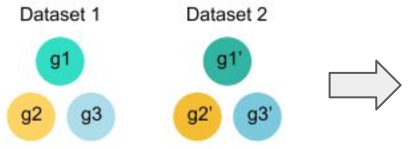
Inter-group Differential ExpRession (IDER) metric

Measures similarity between two groups of cells across datasets

IDER for g1 and g1’:

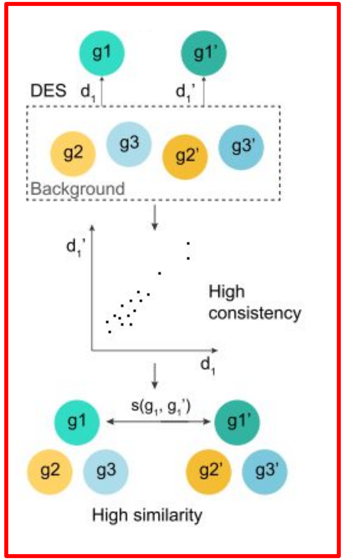
1. Separately, identify differentially expressed genes (DEGs) for g1 and g1’ each vs all other groups (limma-trend; can regress out confounders)
   → d1 and d1’ vectors (log2 fold change coeffs for each gene vs background)

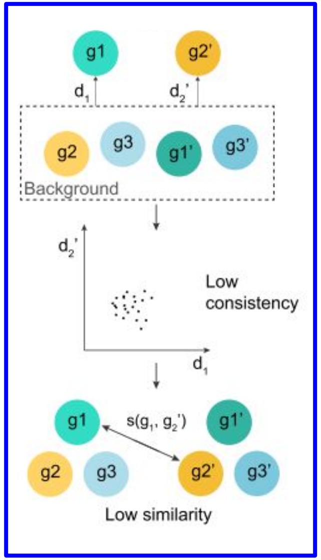
2. IDER(g1,g1’) = Pearson r(d1,d1’)
   similarity of the DEG vectors for g1 and g1’



IDER matrix:

	g1’	g2’	g3’
g1	high	low	low
g2	low	high	low
g3	low	low	high

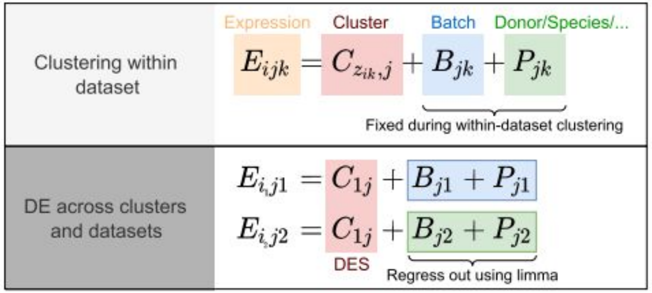




Clustering with IDER (CIDER)

Assumption: expression level is a linear combination of effects of:

- cluster (of interest)
- batch, donor, platform, etc. (confounders)



$$E_{ijk} = C_{z_{ik},j} + \underbrace{B_{jk} + P_{jk}}_{\text{Fixed during within-dataset clustering}}$$

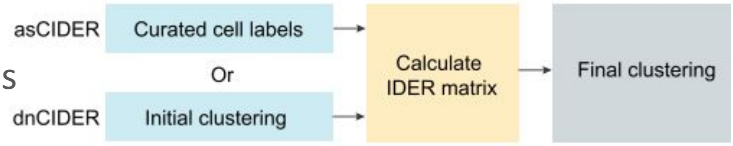
$$E_{i,j1} = C_{1j} + \boxed{B_{j1} + P_{j1}}$$

$$E_{i,j2} = C_{1j} + \underbrace{\boxed{B_{j2} + P_{j2}}}_{\text{Regress out using limma}}$$

CIDER algorithm:

1. Within dataset clustering → cluster effect only (confounding effects are constant)
   a. Unsupervised clustering algorithm (e.g., Louvain clustering) → (de novo) dnCIDER
   b. Curated annotations → (assisted) asCIDER

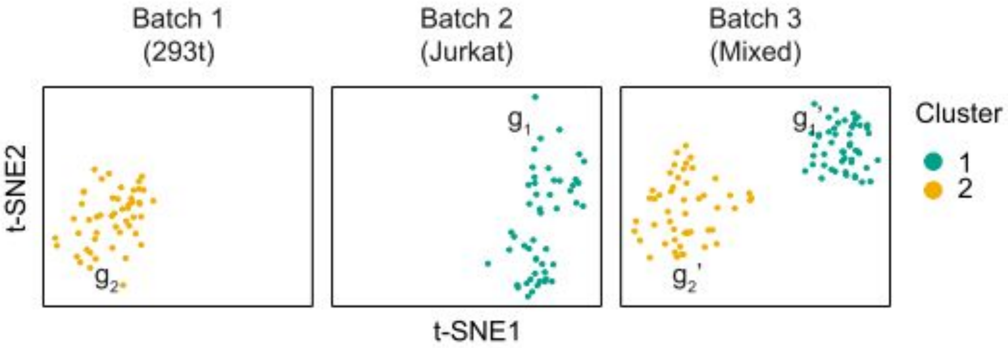
2. Compute IDER similarity matrix across all within-batch clusters to get cross-batch similarity → cluster similar groups across batches
   a. Similarity matrix S → distance matrix (1-S)
   b. Agglomerative clustering with complete linkage
3. (optional:) Use limma to regress out confounding effects

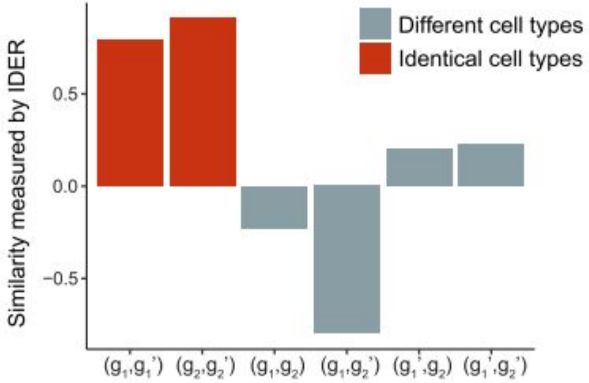


Simple example

Dataset1: Batch correction benchmarking dataset (Zheng et al 2017)

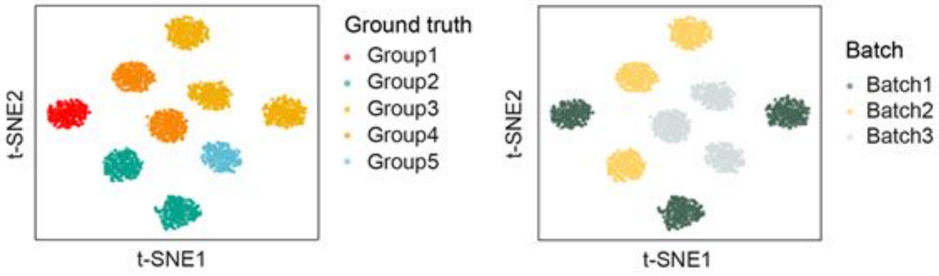
1. Only 293T cells
2. Only Jurkat cells
3. 1:1 mixture of 293T & Jurkat cells





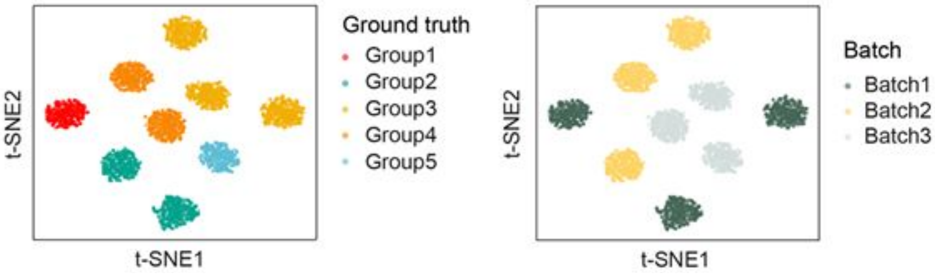
Benchmarking with simulated data

- 5 groups across 3 batches with non-identical populations

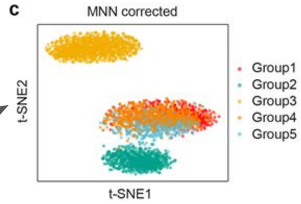


Benchmarking with simulated data

- 5 groups across 3 batches with non-identical populations

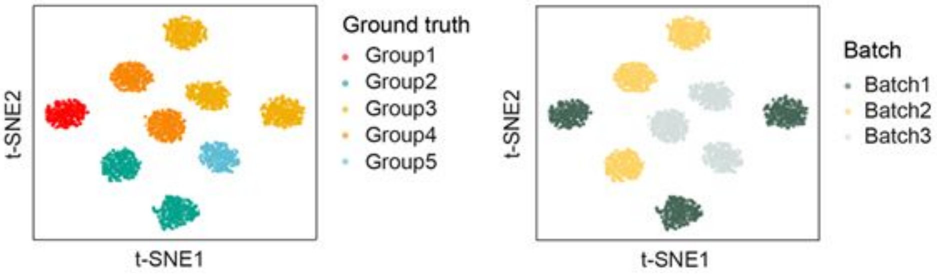


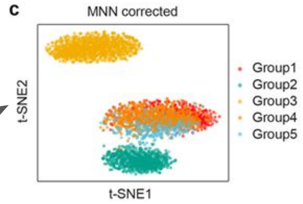
- Many alternative methods “overcorrect” for batch effects



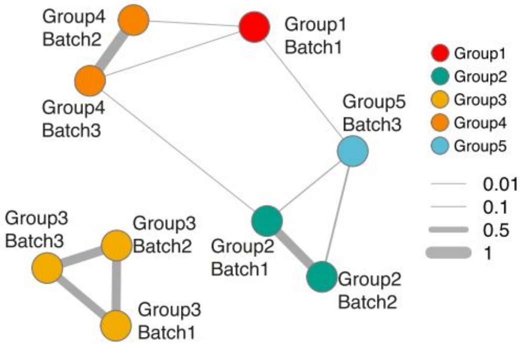
Benchmarking with simulated data

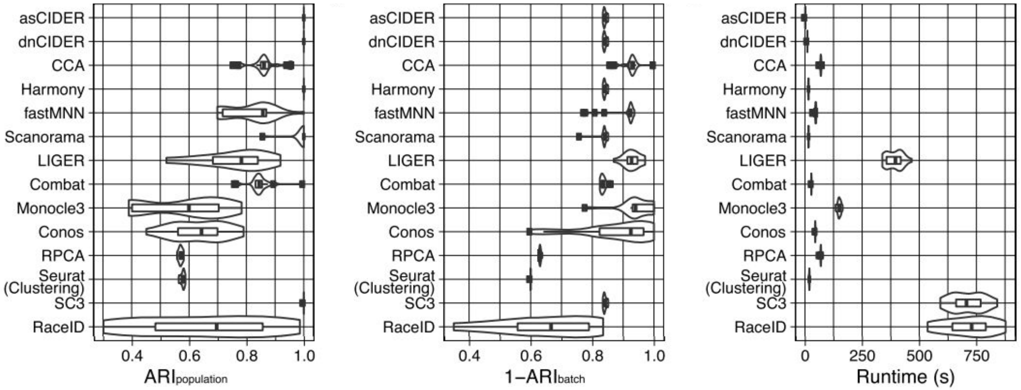
- 5 groups across 3 batches with non-identical populations





- Many alternative methods “overcorrect” for batch effects

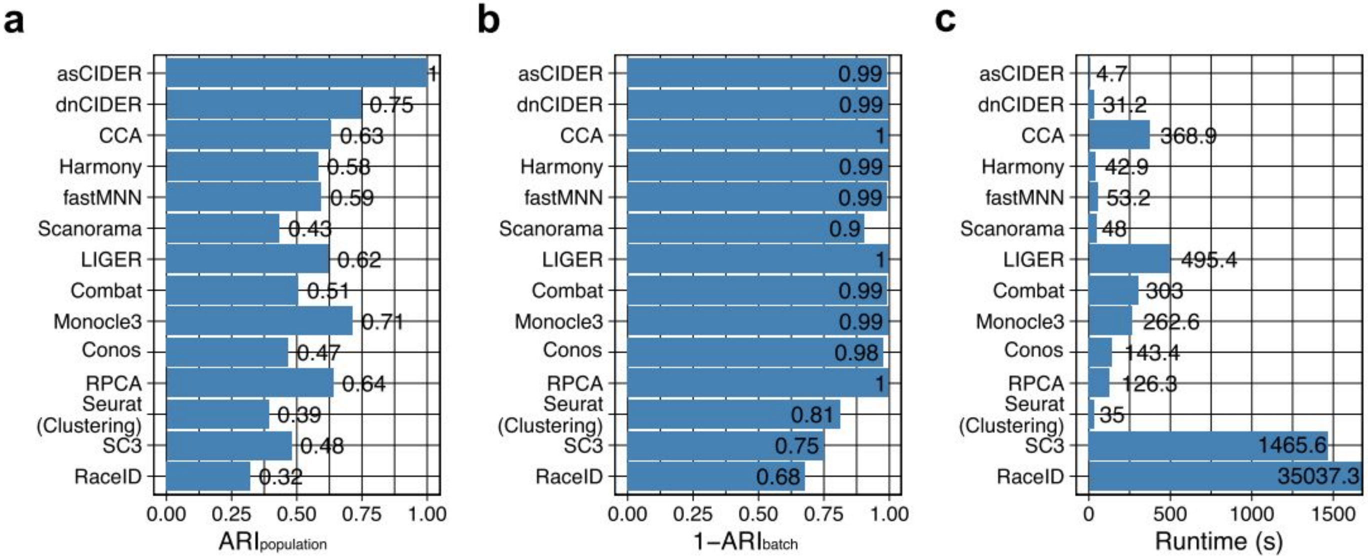




Benchmarking with real data: PBMCs

Dataset 3: human peripheral blood mononuclear cells (PBMCs)

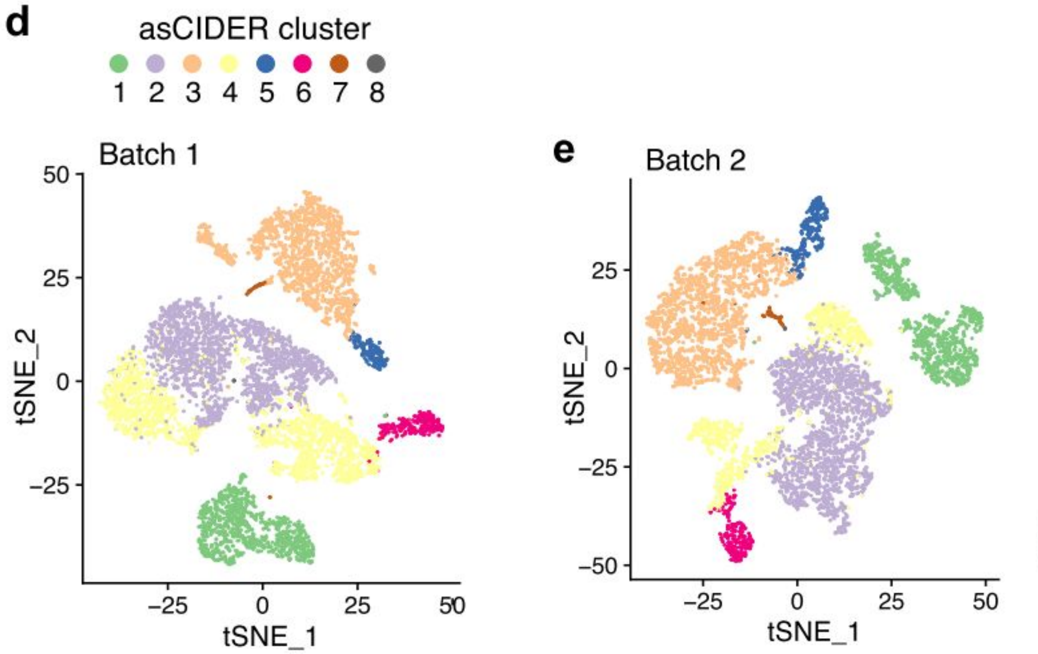
- 9 cell types/subtypes
- 2 techniques (10x 3’ and 5’ single-cell GE) as batches



Benchmarking with real data: PBMCs

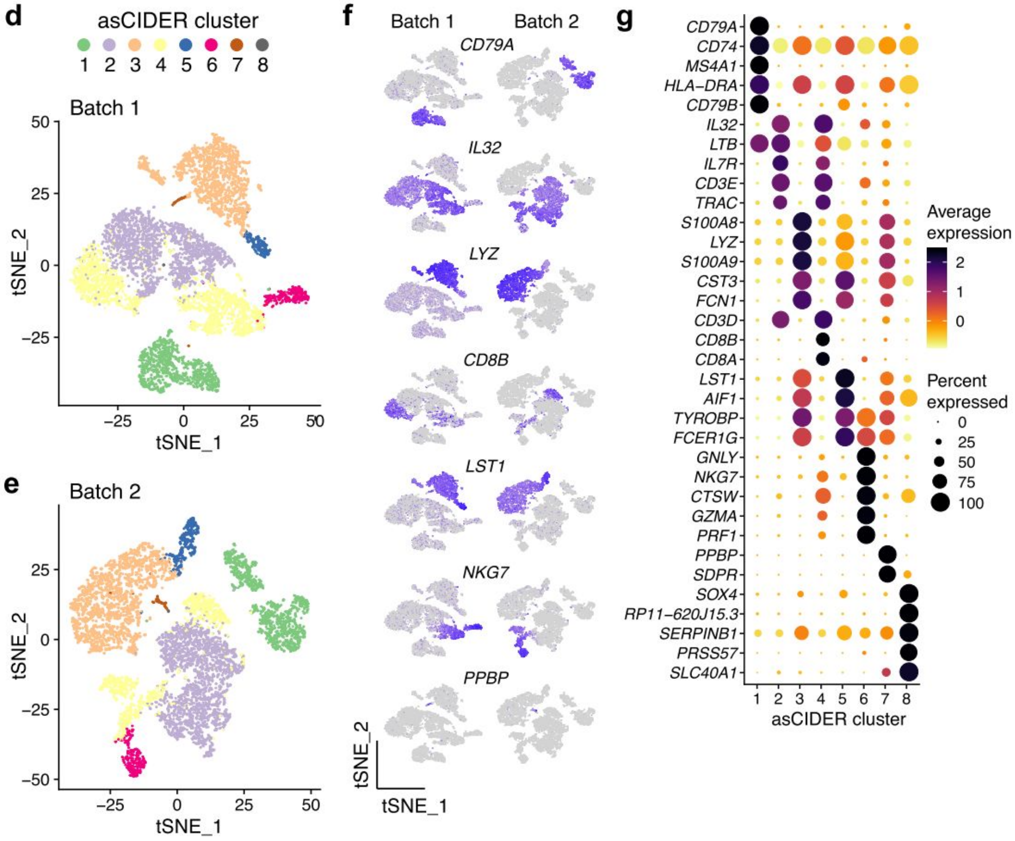
Dataset 3: human peripheral blood mononuclear cells (PBMCs)

- 9 cell types/subtypes
- 2 techniques (10x 3’ and 5’ single-cell GE) as batches



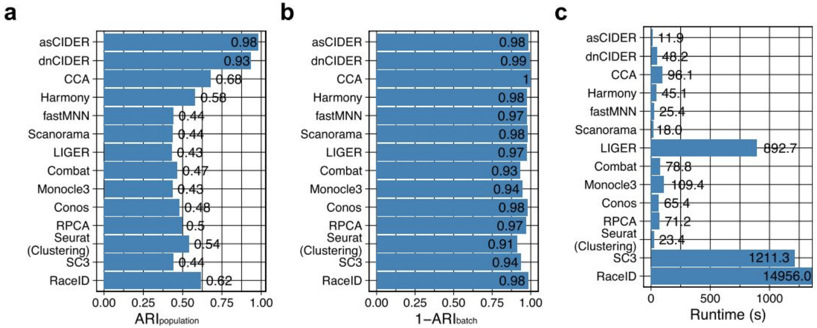
Benchmarking with real data: PBMCs

asCIDER could reveal the underlying relationships among initial clusters

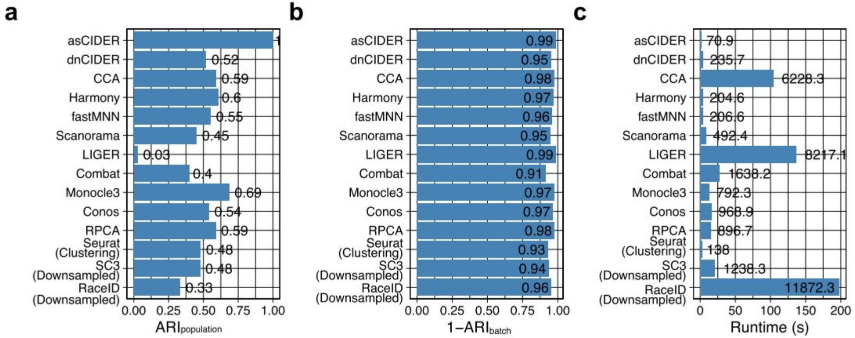


Benchmarking with real data

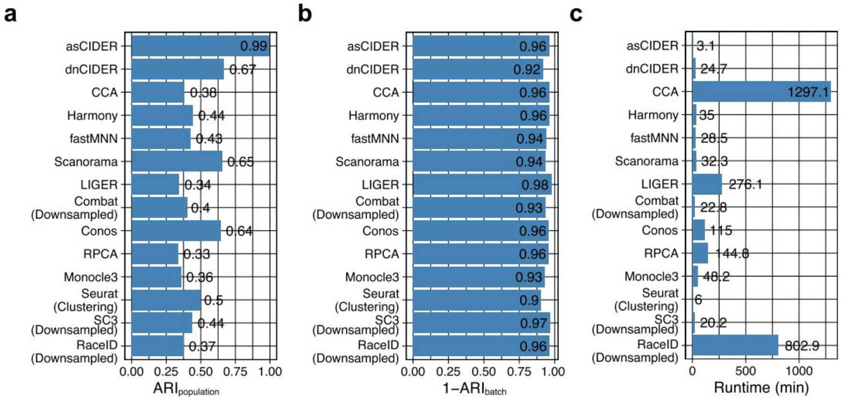
Dataset 4: human and mouse pancreatic data



Dataset 5: COVID-19 study

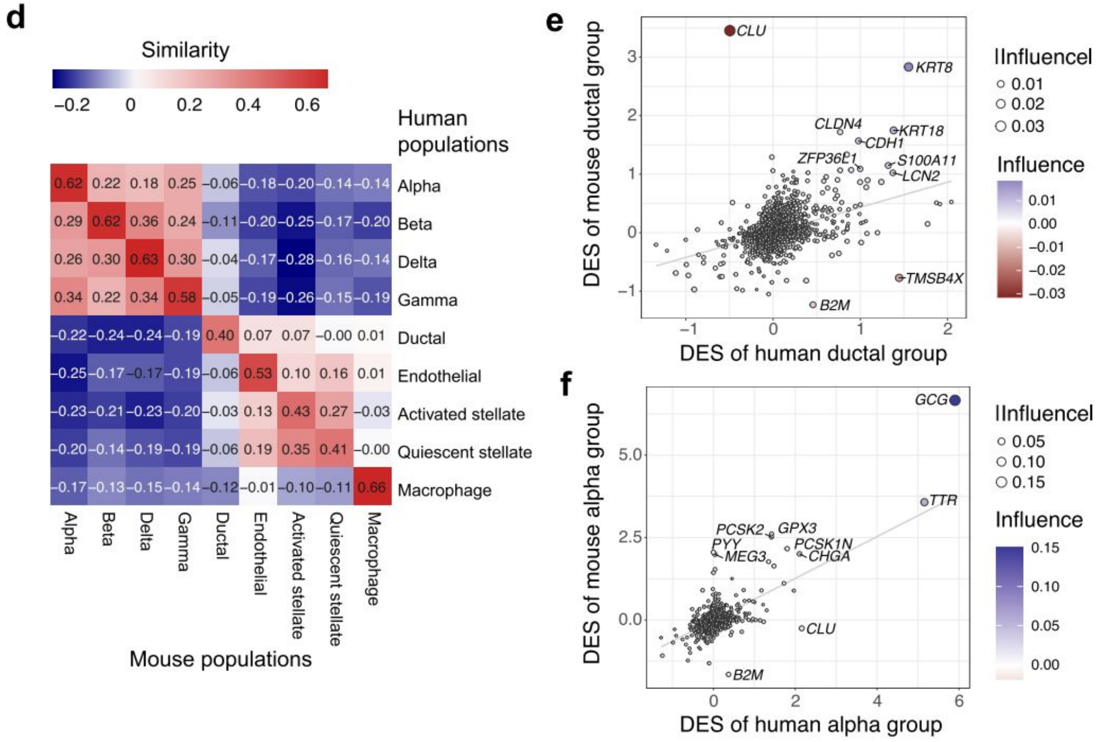


Dataset 6: breast cancer data



Benchmarking with real data: human vs mouse pancreatic cells

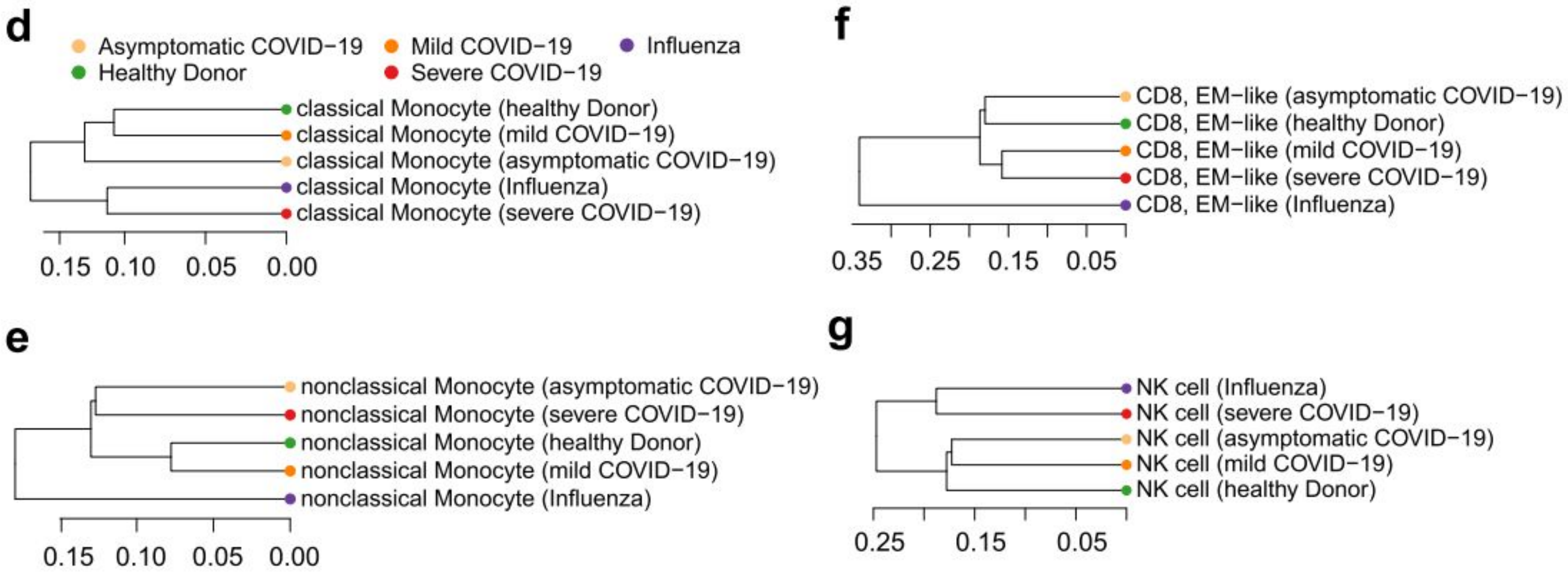
2 mouse samples, 4 human samples → both species and donor effect



	Alpha	Beta	Delta	Gamma	Ductal	Endothelial	Activated stellate	Quiescent stellate	Macrophage
Alpha	0.62	0.22	0.18	0.25	−0.06	−0.18	−0.20	−0.14	−0.14
Beta	0.29	0.62	0.36	0.24	−0.11	−0.20	−0.25	−0.17	−0.20
Delta	0.26	0.30	0.63	0.30	−0.04	−0.17	−0.28	−0.16	−0.14
Gamma	0.34	0.22	0.34	0.58	−0.05	−0.19	−0.26	−0.15	−0.19
Ductal	−0.22	−0.24	−0.24	−0.19	0.40	0.07	0.07	−0.00	0.01
Endothelial	−0.25	−0.17	−0.17	−0.19	−0.06	0.53	0.10	0.16	0.01
Activated stellate	−0.23	−0.21	−0.23	−0.20	−0.03	0.13	0.43	0.27	−0.03
Quiescent stellate	−0.20	−0.14	−0.19	−0.19	−0.06	0.19	0.35	0.41	−0.00
Macrophage	−0.17	−0.13	−0.15	−0.14	−0.12	−0.01	−0.10	−0.11	0.66

Benchmarking with real data: COVID-19

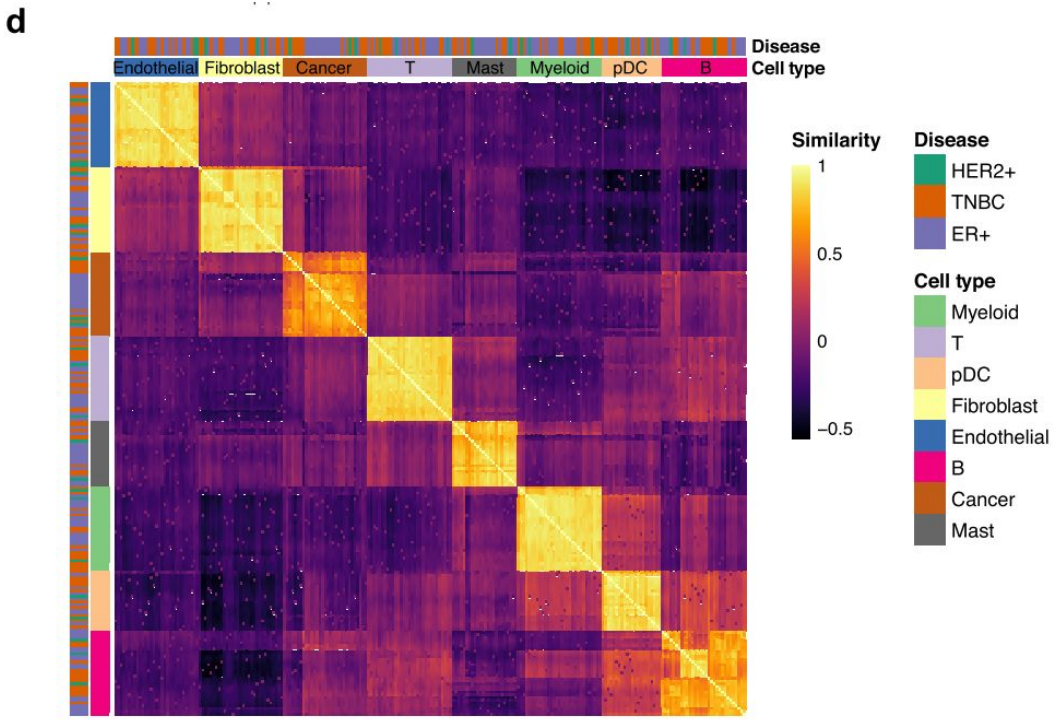
PBMCs collected from healthy donors, patients with severe influenza, and patients with various severity of COVID-19 (asymptomatic, mild, and severe)



Benchmarking with real data: breast cancer

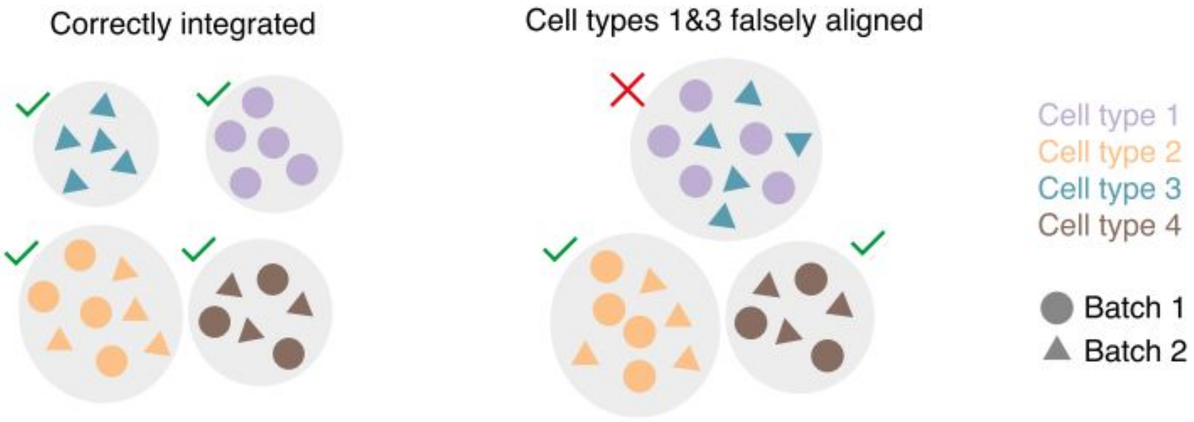
170K cells from 31 breast cancer patients

Two samples per patient: before and after treatment



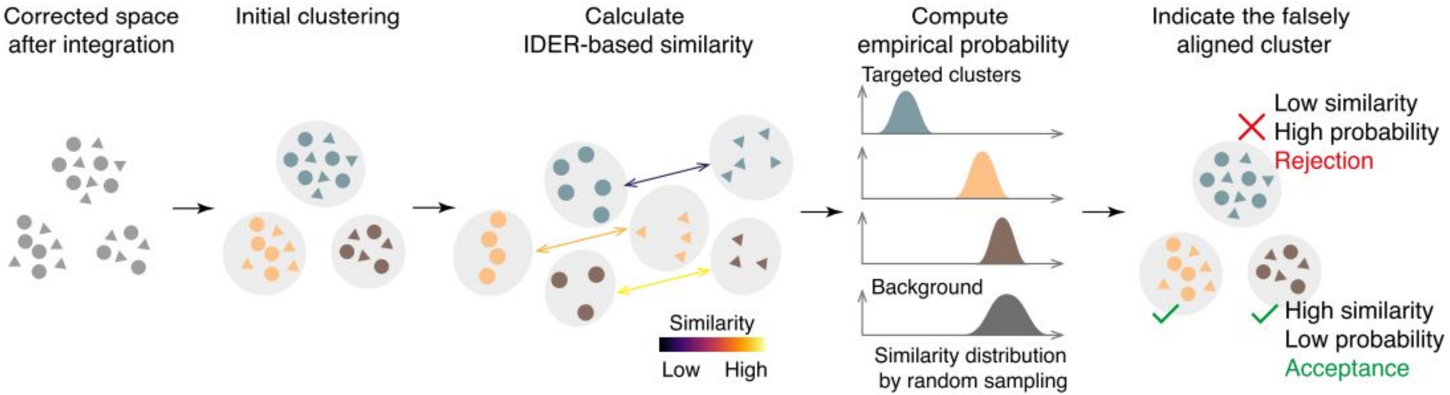
CIDER as a ground-truth-free test metric of integration

- Common issue for integration methods: incorrect alignment - sometimes groups are merged that shouldn't have been
- Other existing metrics require predefined cell populations (e.g., cLISI: Cell-type local inverse Simpson Index)



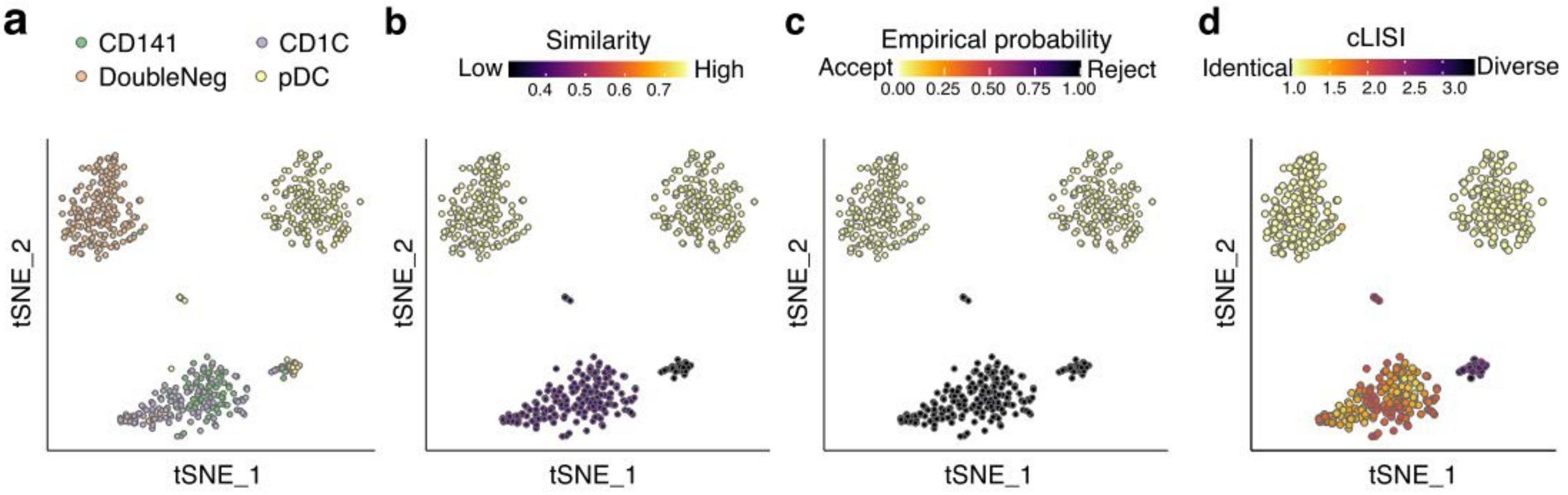
Embedding CIDER into a workflow to evaluate integration:

- Other method: Perform batch correction and learn cross-batch clusters
- Apply IDER metric to cross-batch clusters:
  - For each learned cluster, split by batch
  - Compute IDER similarity for each pair ⟶ higher similarity=better integration
  - Compare pairs’ similarity to distribution of similarities for random partitions within the cluster


Corrected space after integration
Initial clustering
Calculate IDER-based similarity
Similarity
Low
High
Compute empirical probability
Targeted clusters
Background
Similarity distribution by random sampling
Indicate the falsely aligned cluster
Low similarity
High probability
Rejection
High similarity
Low probability
Acceptance

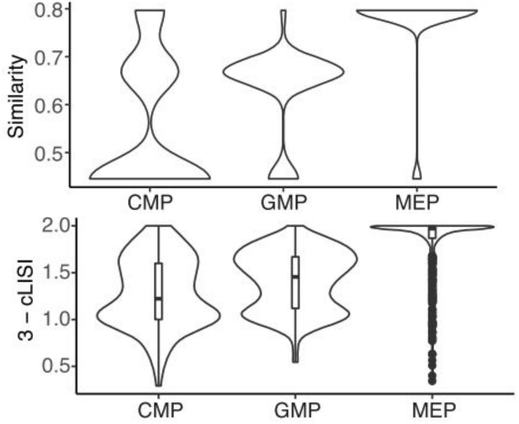
Using CIDER to evaluate CCA integration on a dendritic cell dataset

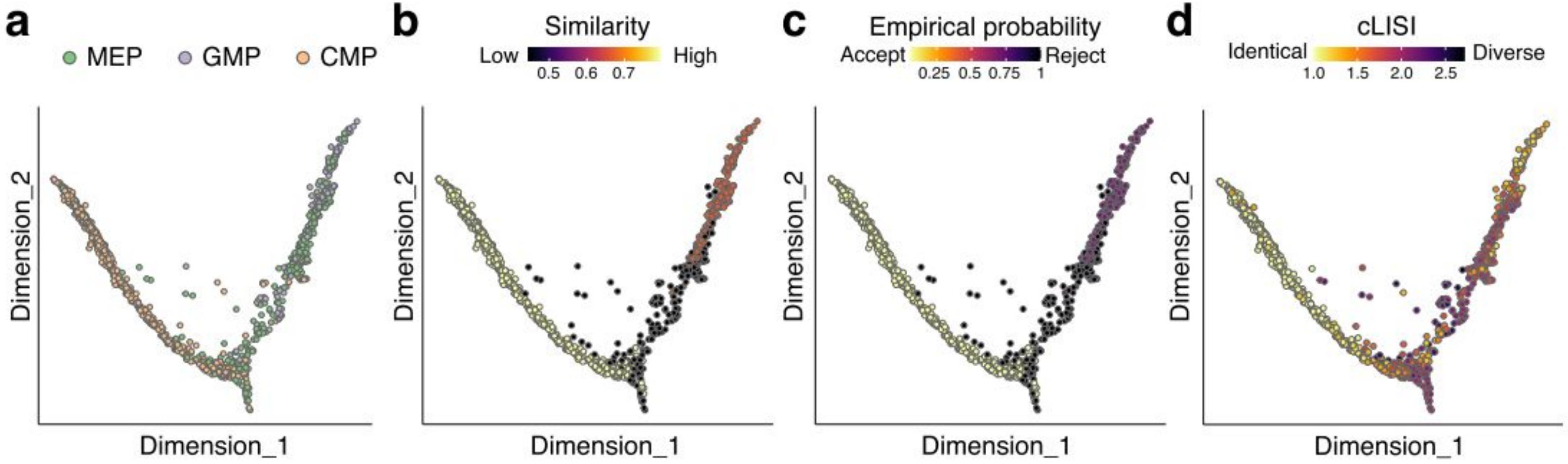
- CD141 & CD1C are prone to being merged by batch correction methods
- CIDER has similar results with cLISI (but doesn’t require labels to calculate)



Using CIDER to evaluate mouse hematopoietic progenitor data (continuous data structure)

- Goal: Use CIDER to evaluate local biological heterogeneity without predefined annotations
- Mouse hematopoietic progenitor data (common myeloid, megakaryocyte/erythrocyte, and granulocyte/macrophage progenitor cells) from 2 platforms (MARS-seq and Smart-Seq2)





Discussion

Summary:

- Introduced IDER, a differentially expressed gene-based similarity metric, which can be used to identify cross-batch clusters
- Both dnCIDER and asCIDER were evaluated on a wide array of benchmarks (dnCIDER was often much better)
- IDER metric can be used to evaluate other batch-correction methods in the absence of ground truth labels

Limitations:

- Developed for scRNA-Seq - currently not designed for multi-modal data
- Linear approach
- Group-level analysis assumes coarse-grained clusters (not continuous data)

Discussion topics

- Worse performance for dnCIDER vs asCIDER – how do we feel about that, given that one of their presented advantages is not needing labels?

- This space is quite saturated (e.g., all the methods they benchmarked against)
  - What does a new method need to achieve to really be worth using? Did this paper meet that standard?
  - Where should the field go next?

- Circular benchmarks: Most “ground truth” labels are actually the output of clustering methods/previously found gene signatures which are used to identify cell types, so new methods benchmark against these