`singleCellHaystack`

A clustering-independent method for predicting differentially expressed genes in single cell transcriptome data

Vandenbon and Diez, *Nature Communications*, 2020

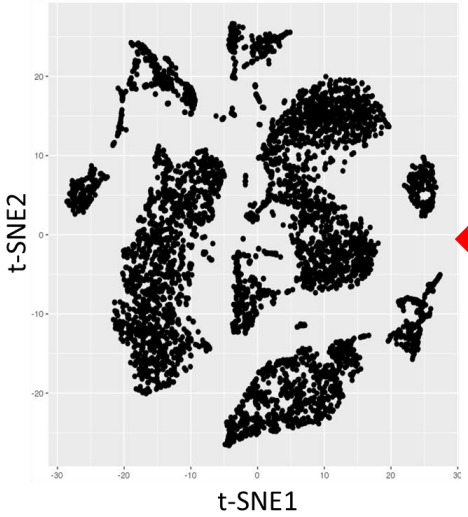
Also on GitHub and CRAN

Alexis Vandenbon alexisvdb@infront.kyoto-u.ac.jp
Institute for Frontier Life and Medical Sciences
Institute for Liberal Arts and Sciences
Kyoto University

Diego Diez
Immunology Frontier Research Center
Osaka University

Predicting DEGs in single-cell data

- Single-cell data is **high-dimensional**
- Often unclear:
  - **how many clusters** are there?
  - do **borders between clusters** make sense?
  - are there **sub-populations** within larger clusters?
- Comparing between many clusters is difficult and time consuming

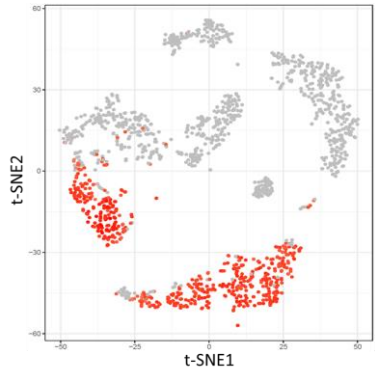


How many clusters??

It would be nice to have a ***clustering-independent*** method for finding DEGs

singleCellHaystack methodology

Any ≥2D space

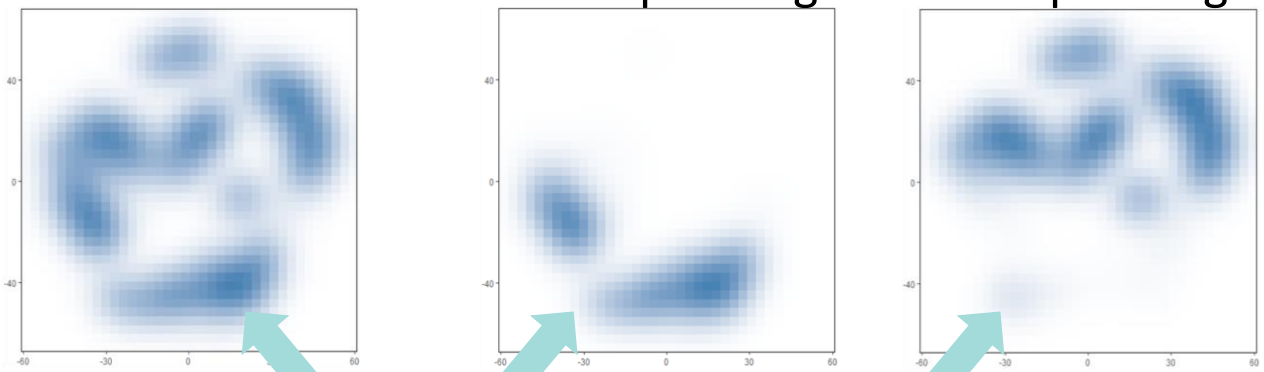


Expression of gene G

Distributions of…

All cells

Cells expressing G & **not** expressing G



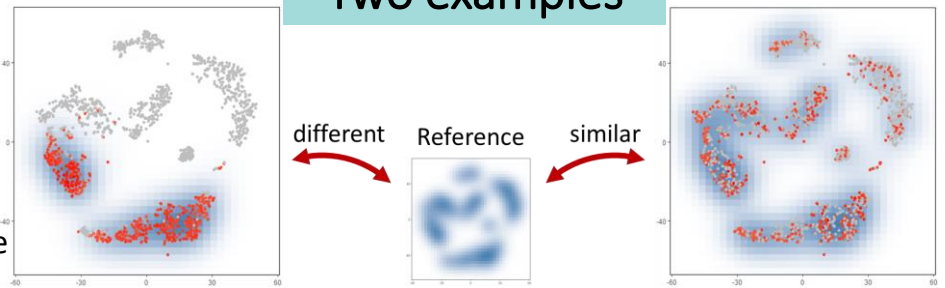
Comparison using **Kullback-Leibler divergence**

Two examples

A DEG

D_{KL}=1.08

low p value



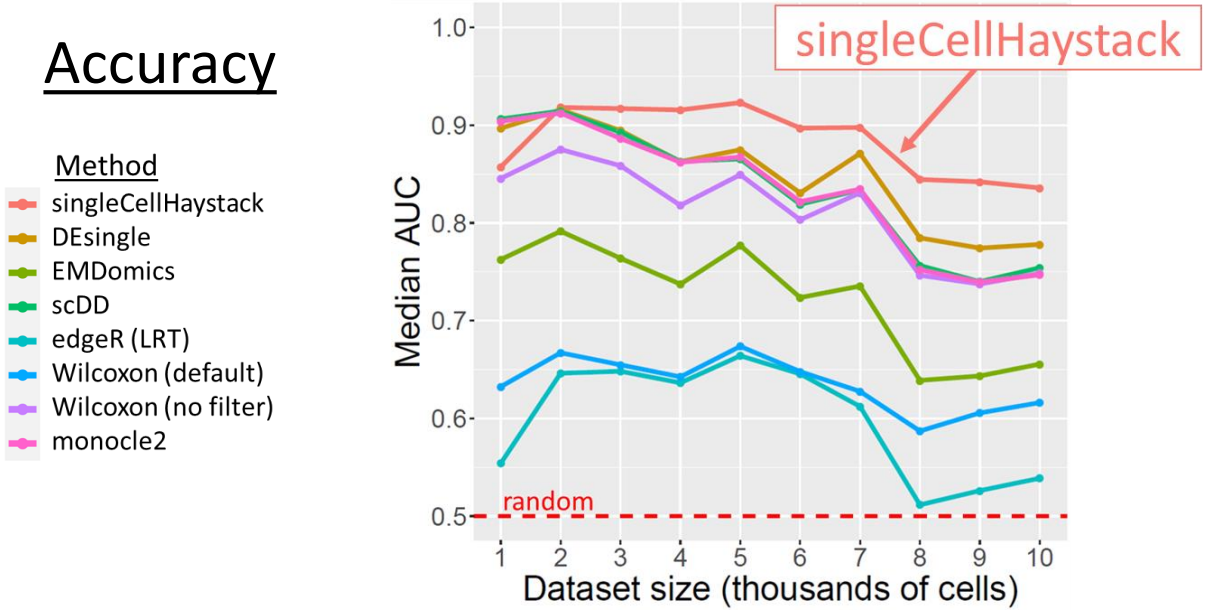
NOT a DEG

D_{KL}=0.07

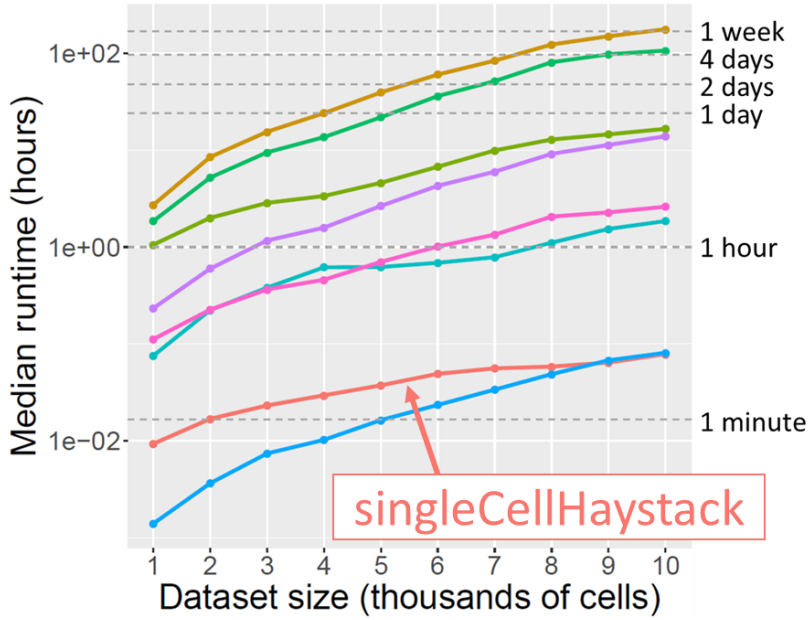
high p value

Comparison using artificial datasets

Accuracy

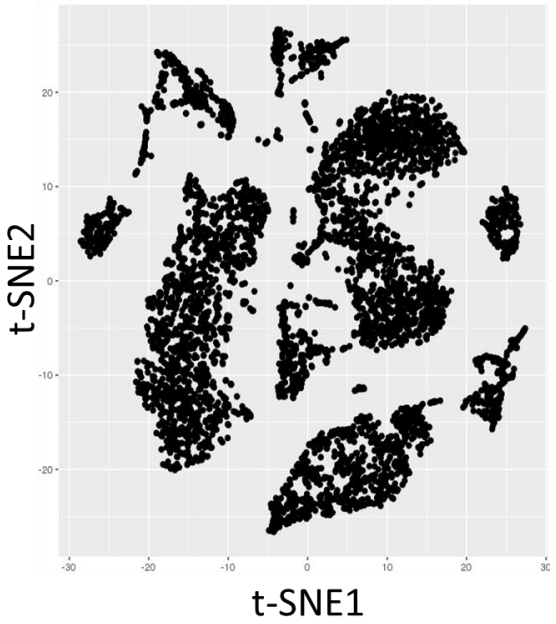


Runtimes



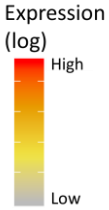
singleCellHaystack is accurate and fast

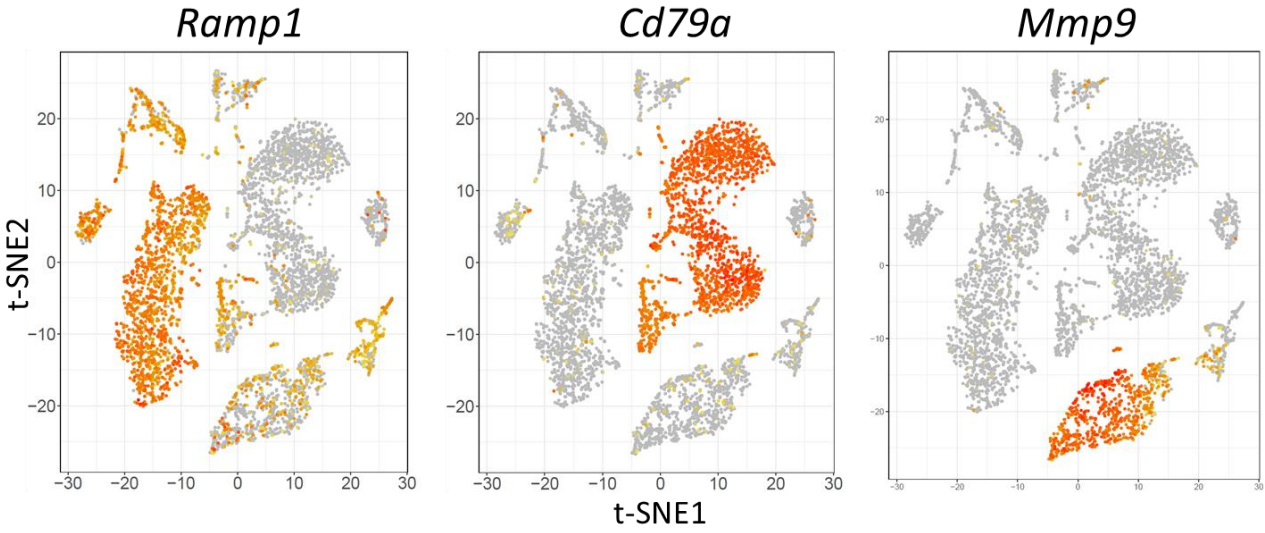
Application on real datasets



Tabula Muris bone marrow dataset

- 5,250 cells
- Input coordinates: 50 PC

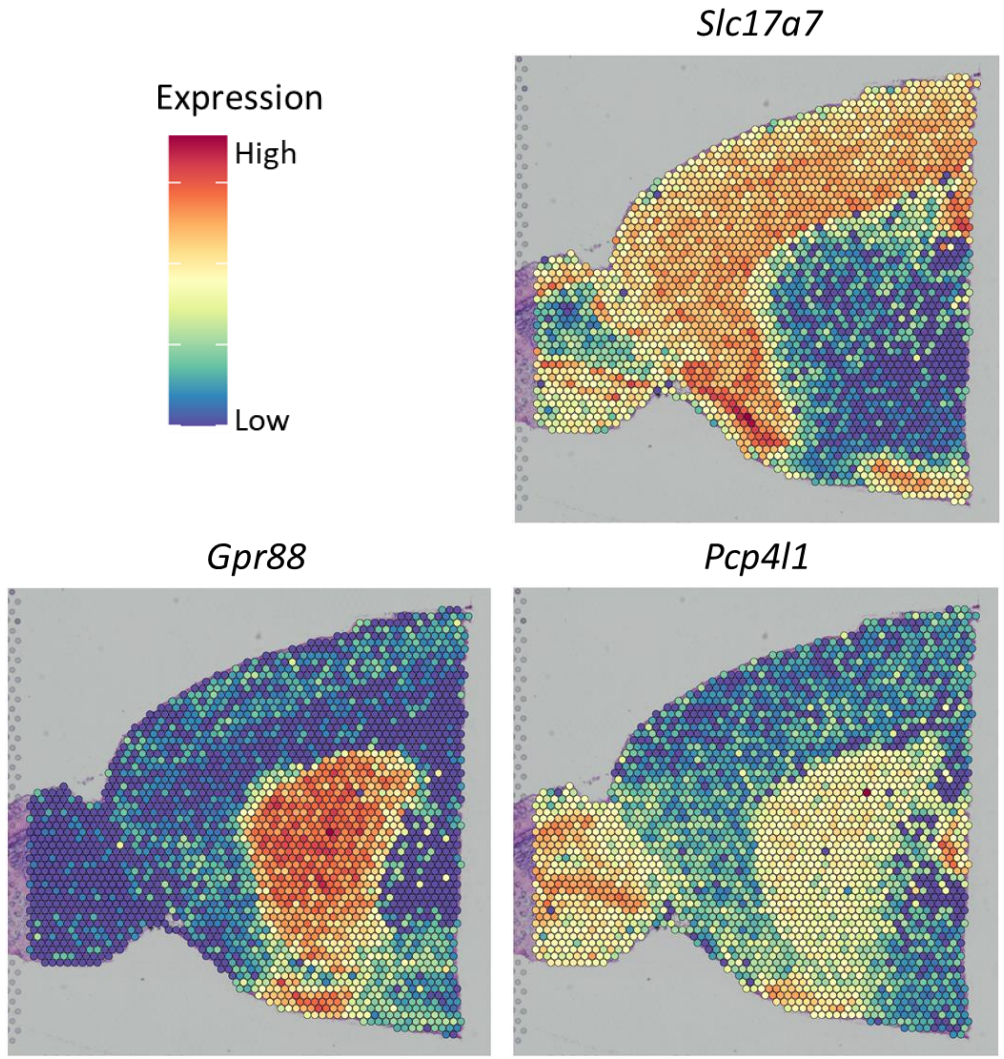




Application on spatial transcriptomics

Mouse anterior brain (10x Genomics Visium)

- 2,696 beads
- Input coordinates: 2D spatial coordinates



Summary singleCellHaystack

- Method for finding DEGs in single cell data (Vandenbon and Diez, *Nature Communications*, 2020)
- **Does not rely on clustering of cells**
- **Fast**
- Can find **any non-random expression pattern**
- High-scoring DEGs are often **known marker genes**
- Available as an **R package** on GitHub and CRAN

nature COMMUNICATIONS

GitHub