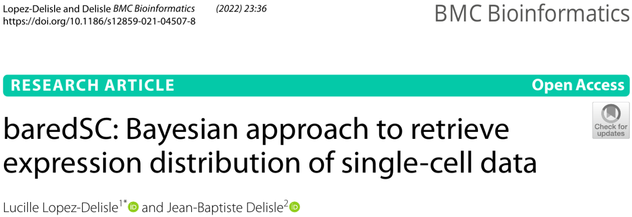
baredSC in Galaxy

Lucille Delisle

EDG 2023

2023-10-05



Lopez-Delisle and Delisle *BMC Bioinformatics* *(2022) 23:36*
https://doi.org/10.1186/s12859-021-04507-8

BMC Bioinformatics

RESEARCH ARTICLE **Open Access**

baredSC: Bayesian approach to retrieve expression distribution of single-cell data

Lucille Lopez-Delisle[1*] and Jean-Baptiste Delisle[2]

scRNA-seq

Cell Type A
Cell Type B
Bulk RNA Seq
SCRNA Seq

From perkinelmer website

scRNA-seq

Cell Type A
Cell Type B
Bulk RNA Seq
SCRNA Seq

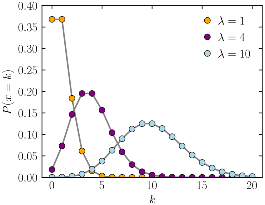
From perkinelmer website

- scRNA-seq:
  - Get a count for:
    - each cell
    - each gene
  - The matrix is very sparse:
    - About 360k mRNA per cell (source: qiagen), usually sequence 5-40k mRNA.
    - A 0 does not mean no expression.
    - The noise and sparsity can be explained by the Poisson distribution.
  - People usually display logNorm expression: $log(1 + 10^4 \frac{x}{N})$

scRNA-seq

From perkinelmer website

- scRNA-seq:
  - Get a count for:
    - each cell
    - each gene
  - The matrix is very sparse:
    - About 360k mRNA per cell (source: qiagen), usually sequence 5-40k mRNA.
    - A 0 does not mean no expression.
    - The noise and sparsity can be explained by the Poisson distribution.
  - People usually display logNorm expression: $log(1 + 10^4 \frac{x}{N})$

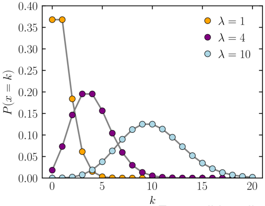


From wikimedia

A mRNA with a concentation of 10^{-4}
Sequence 10k mRNA ($\lambda = 1$)
Sequence 40k mRNA ($\lambda = 4$)

scRNA-seq

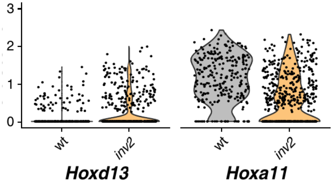
From perkinelmer website



From wikimedia

- scRNA-seq:
  - Get a count for:
    - each cell
    - each gene
  - The matrix is very sparse:
    - About 360k mRNA per cell (source: qiagen), usually sequence 5-40k mRNA.
    - A 0 does not mean no expression.
    - The noise and sparsity can be explained by the Poisson distribution.
  - People usually display logNorm expression: $log(1 + 10^4 \frac{x}{N})$

A mRNA with a concentation of 10^{-4}
Sequence 10k mRNA ($\lambda = 1$)
Sequence 40k mRNA ($\lambda = 4$)

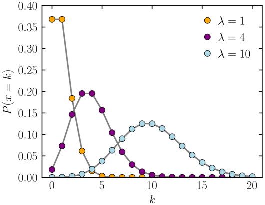


scRNA-seq

Cell Type A
Cell Type B
Bulk RNA Seq
SCRNA Seq

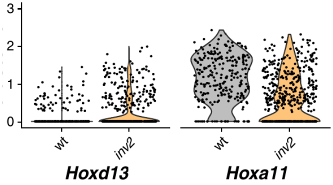
From perkinelmer website

- scRNA-seq:
  - Get a count for:
    - each cell
    - each gene
  - The matrix is very sparse:
    - About 360k mRNA per cell (source: qiagen), usually sequence 5-40k mRNA.
    - A 0 does not mean no expression.
    - The noise and sparsity can be explained by the Poisson distribution.
  - People usually display logNorm expression: $log(1 + 10^4 \frac{x}{N})$



From wikimedia

A mRNA with a concentation of 10^{-4}
Sequence 10k mRNA ($\lambda = 1$)
Sequence 40k mRNA ($\lambda = 4$)



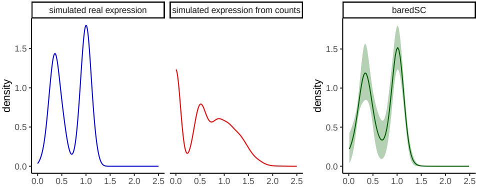
If we know how to model the noise, can we denoise scRNA-seq?

baredSC for a single gene (baredSC_1d)

- Goal: Find an estimation of the Probability Density Function (PDF) of the REAL expression for a given gene.
- Hypotheses:
  - Most of ’noise’ in scRNA-seq comes from sampling and can be explained by a Poisson law.
  - The PDF can be approximated by a Gaussian mixture model.
- Parameters
  - Number of Gaussians
  - Characteristics of Gaussians
- Strategy
  - Bayesian approach = evaluate the probability of the parameters given the data
  - We use Markov chain Monte Carlo for a fixed number of Gaussians and then combine different results using evidence.

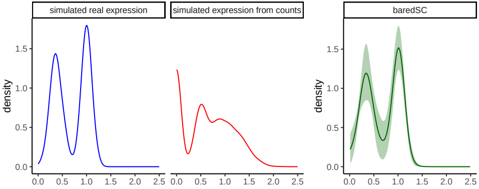
Test baredSC_1d using simulated data

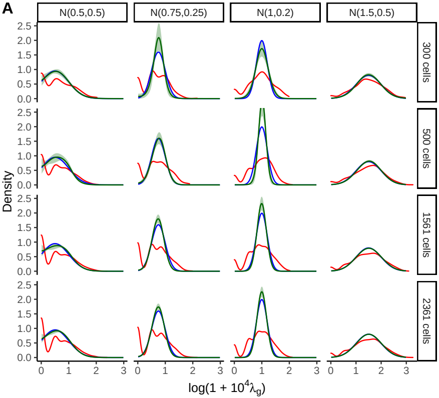
- Generate random expression following different distributions
- Use number of mRNA per cell quantified from a real dataset
- Simulate counts using Poisson
- Run baredSC_1d



Test baredSC_1d using simulated data

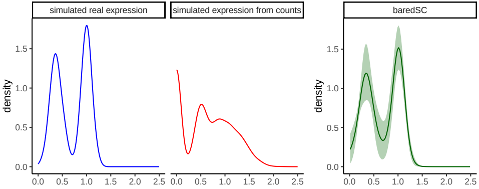
- Generate random expression following different distributions
- Use number of mRNA per cell quantified from a real dataset
- Simulate counts using Poisson
- Run baredSC_1d

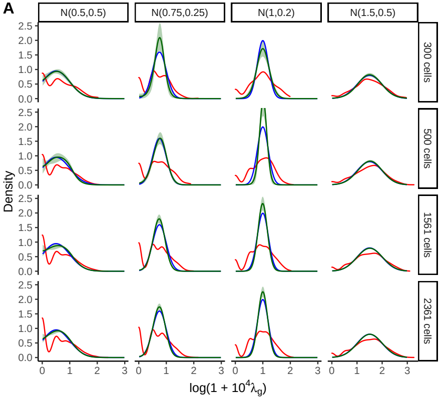


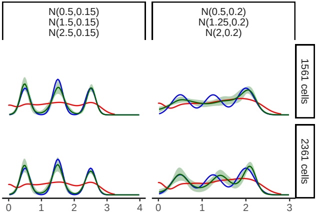


Test baredSC_1d using simulated data

- Generate random expression following different distributions
- Use number of mRNA per cell quantified from a real dataset
- Simulate counts using Poisson
- Run baredSC_1d

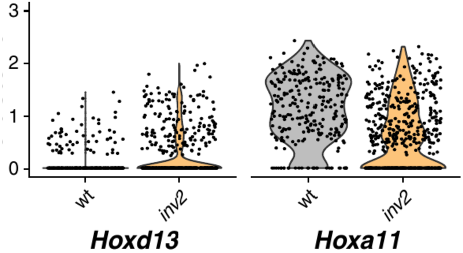






baredSC_1d with real data

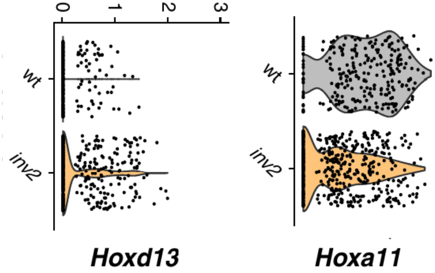
- Improve regular violin plots



Bolt et al. 2021

baredSC_1d with real data

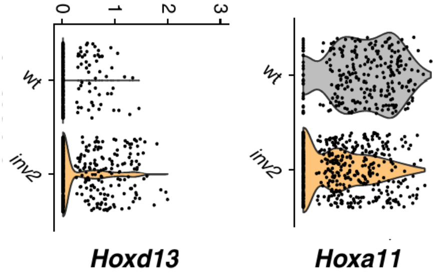
- Improve regular violin plots

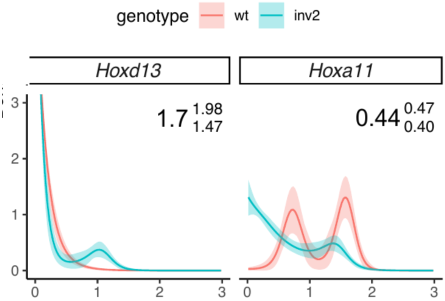


Bolt et al. 2021

baredSC_1d with real data

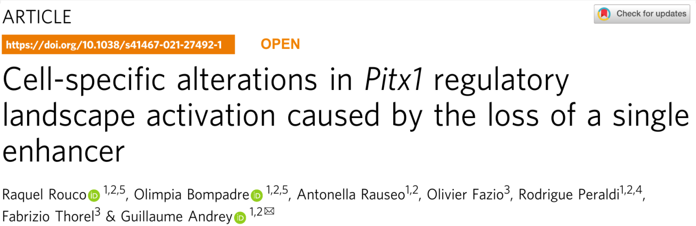
- Improve regular violin plots





Bolt et al. 2021

Application of baredSC in study where both FACS and scRNAseq datasets are available



ARTICLE

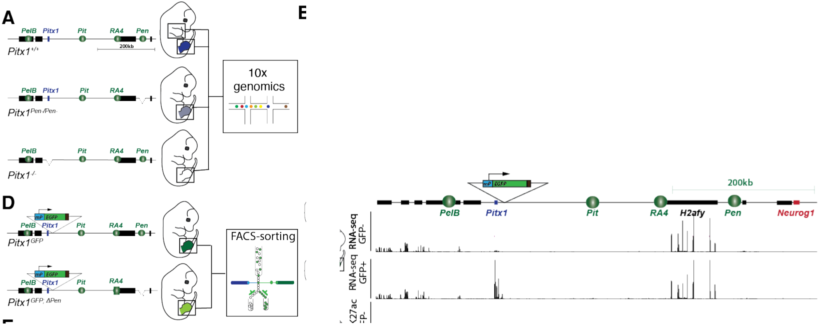
Check for updates

https://doi.org/10.1038/s41467-021-27492-1

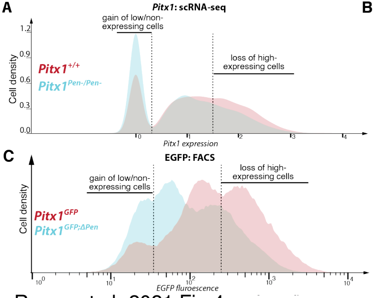
OPEN

Cell-specific alterations in *Pitx1* regulatory landscape activation caused by the loss of a single enhancer

Raquel Rouco[1,2,5], Olimpia Bompadre[1,2,5], Antonella Rauseo[1,2], Olivier Fazio[3], Rodrigue Peraldi[1,2,4], Fabrizio Thorel[3] & Guillaume Andrey[1,2]✉

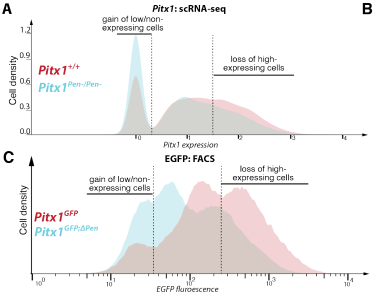


Application of baredSC in study where both FACS and scRNAseq datasets are available



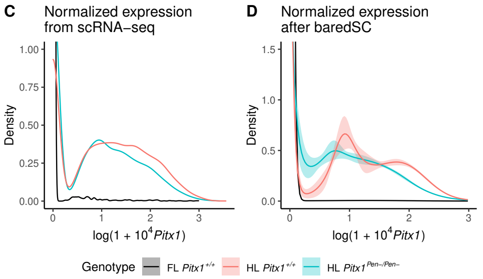
Rouco et al. 2021 Fig 4

Application of baredSC in study where both FACS and scRNAseq datasets are available

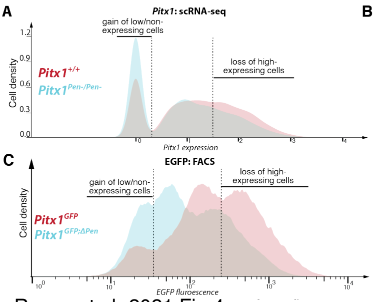


Rouco et al. 2021 Fig 4

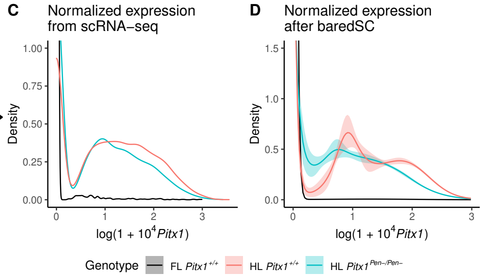
→

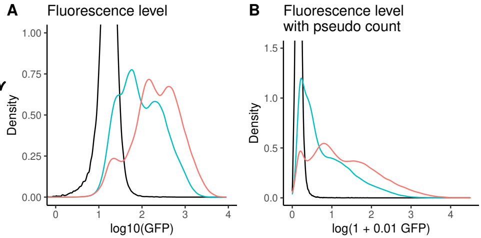


Application of baredSC in study where both FACS and scRNAseq datasets are available



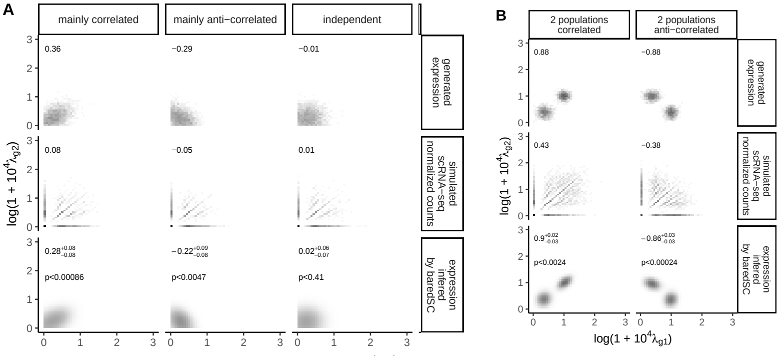
Rouco et al. 2021 Fig 4





baredSC_2d

- The same strategy used for a single gene can be extended to 2 dimensions for 2 genes using 2D gaussians.
- From the MCMC posteriors we can deduce a correlation coefficient.

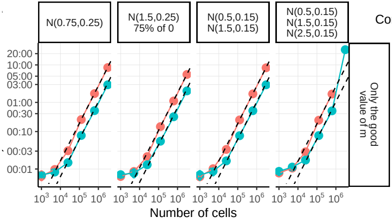


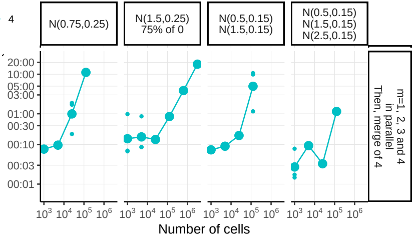
baredSC: Conclusions

- baredSC help to study the distribution of expression levels in a few genes of interest.
  - It could replace the widely used violin plots from normalized data.
  - It allows to retrieve the multi-modal expression distribution.

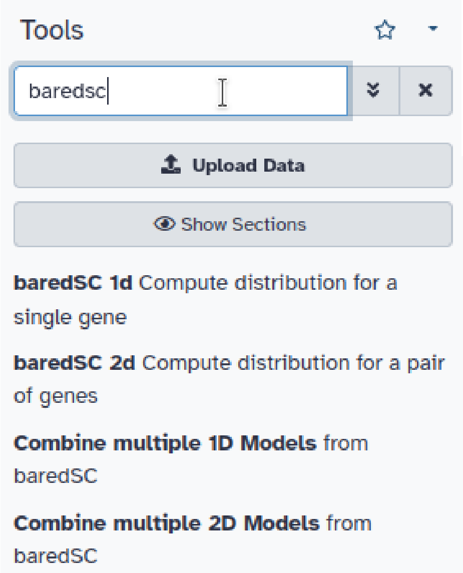
- baredSC in 2D allows better evaluation of the correlation between genes.

- Big disadvantage of baredSC is the computation time.





baredSC is already in Galaxy

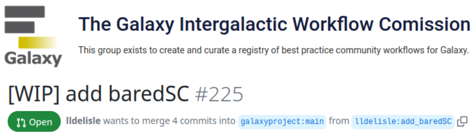


baredSC 1d Compute distribution for a single gene

baredSC 2d Compute distribution for a pair of genes

Combine multiple 1D Models from baredSC

Combine multiple 2D Models from baredSC



The Galaxy Intergalactic Workflow Comission

This group exists to create and curate a registry of best practice community workflows for Galaxy.

[WIP] add baredSC #225

Open **lldelisle** wants to merge 4 commits into `galaxyproject:main` from `lldelisle:add_baredSC`

Acknowledgements

- Jean-Baptiste Delisle
- Duboule's lab
- Andrey's lab
- tools-iuc

EPFL

Swiss National
Science Foundation