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

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

RESEARCH ARTICLE
Open Access

baredSC in Galaxy

Lucille Delisle

EDG 2023

2023-10-05
scRNA-seq

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

- **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

- 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

- 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: $\log(1 + 10^4 \frac{x}{N})$

If we know how to model the noise, can we denoise scRNA-seq?
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
Improve regular violin plots
baredSC_1d with real data

- Improve regular violin plots

Hoxd13

Hoxa11
bareSC_1d with real data

- Improve regular violin plots
Application of baredSC in study where both FACS and scRNAseq datasets are available

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

Raquel Rouco, Olimpia Bompadre, Antonella Rauseo, Olivier Fazio, Rodrigue Peraldi, Fabrizio Thorel & Guillaume Andrey

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

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

Figure 4: Influence of the Pen deletion on Pitx1 expression in hindlimb cell population. A. Pitx1 expression distribution across wildtype (red) and Pitx1Pen−/Pen− (cyan) hindlimb cells shows an increased proportion of non/low-expressing mutant cells and a decrease proportion of high-expressing cells. B. EGFP expression pattern in Pitx1GFP and Pitx1GFP;ΔPen in E12.5 embryos. C. FACS profile of Pitx1GFP (red) and Pitx1GFP;ΔPen (cyan) hindlimbs shows an increased number of EGFP non/low-expressing cells as well as a decrease of EGFP high-expressing cells. D. Pitx1 expression across all cluster s in Pitx1GFP and Pitx1GFP;ΔPen hindlimb. At the base of the distribution, the fold change in non/low-expressing cell number between wildtype and mutant is shown. Note the strong loss of expression and the accumulation of non/low-expressing cells in ICT and PPP clusters. E. F. UMAP (E) and quantification (F) of mesenchyme cell type proportions across conditions. (+) and (-) symbols indicate increase or decrease in cell proportions, stars indicate p<0.05.

The Pen enhancer contributes to Pitx1 regulatory landscape activation. The establishment of the active Pitx1 chromatin landscape includes changes in 3D conformation and the acetylation of specific cis-regulatory elements. Therefore, we asked whether the Pen enhancer itself is required to establish these features and specifically if its deletion would impact them. In sorted GFP+ and GFP− in Pitx1GFP;ΔPen, we first assessed using RNA-seq whether we could observe...
Application of baredSC in study where both FACS and scRNAseq datasets are available

Figure 4: Influence of the Pen deletion on Pitx1 expression in hindlimb cell population. A. Pitx1 expression distribution across wildtype (red) and Pitx1 Pen−/Pen− (cyan) hindlimb cells shows an increased proportion of non/low-expressing mutant cells and a decrease proportion of high-expressing cells. B. EGFP expression pattern in Pitx1 GFP and Pitx1 GFP;ΔPen in E12.5 embryos. C. FACS profile of Pitx1 GFP (red) and Pitx1 GFP;ΔPen (cyan) hindlimbs shows an increased number of EGFP non/low-expressing cells as well as a decrease of EGFP high-expressing cells. D. Pitx1 expression across all cluster s in Pitx1 GFP and Pitx1 GFP;ΔPen hindlimb. At the base of the distribution, the fold change in non/low-expressing cell number between wildtype and mutant is shown. Note the strong loss of expression and the accumulation of non/low-expressing cells in ICT and PPP clusters. E. UMAP (E) and quantification (F) of mesenchyme cell type proportions across conditions. (+) and (−) symbols indicate increase or decrease in cell proportions, stars indicate p<0.05.

The Pen enhancer contributes to Pitx1 regulatory landscape activation. The establishment of the active Pitx1 chromatin landscape includes changes in 3D conformation and the acetylation of specific cis-regulatory elements. Therefore, we asked whether the Pen enhancer itself is required to establish these features and specifically if its deletion would impact them. In sorted GFP+ and GFP− in Pitx1 GFP;ΔPen, we first assessed whether we could observe...
Application of baredSC in study where both FACS and scRNAseq datasets are available

Rouco et al. 2021 Fig 4

A. Pitx1: scRNA-seq

- Gain of low/non-expressing cells
- Loss of high-expressing cells

B. Normalized expression from scRNA-seq

- Density
- Log10(GFP)
- Fluorescence level

C. Normalized expression after baredSC

- Density
- Log10(Pitx1)

D. Normalized expression after baredSC

- Genotype
- FL Pitx1+/+
- HL Pitx1+/+
- HL Pitx1Pen−/Pen−

A. Fluorescence level

- Density
- Log10(GFP)

B. Fluorescence level with pseudo count

- Density
- Log1 + 0.01 GFP
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
Acknowledgements

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