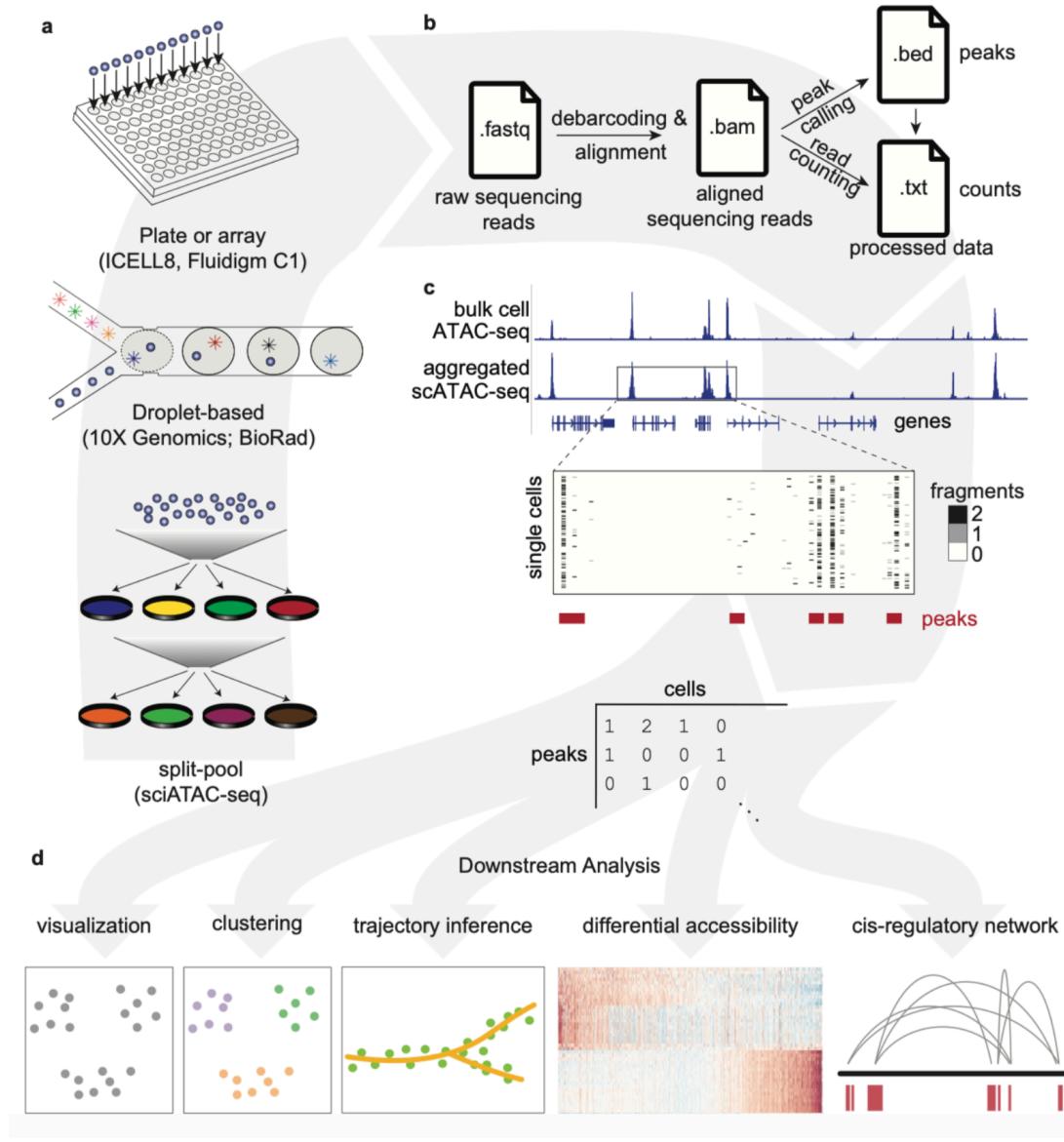


# **cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data**

E. Ravza Gur  
GMS course  
6/12/2019

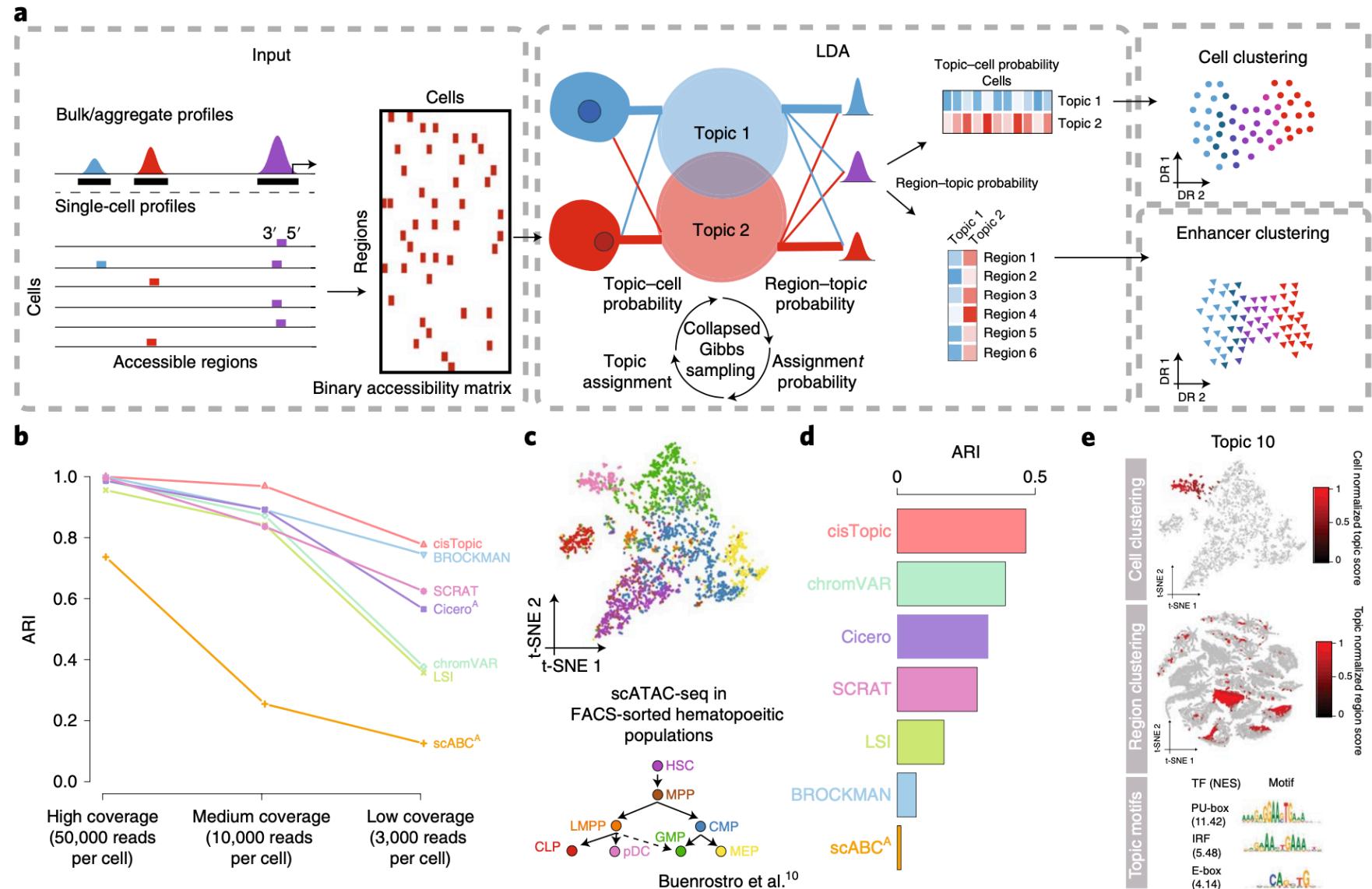
# Overview of single-cell ATAC-seq assays and analysis steps



# The cisTopic workflow

cisTopic consists of 4 main steps:

- (1) generation of a binary accessibility matrix as input for LDA;
- (2) LDA and model selection;
- (3) cell-state identification using the topic-cell distributions from LDA; and
- (4) exploration of the region-topic distributions.



- (1) the probability of a region belonging to a topic (region-topic distribution) and
- (2) the contribution of a topic within a cell (topic-cell distribution)

Document 1: I had a peanut butter sandwich for breakfast.

Document 2: I like to eat almonds, peanuts and walnuts.

Document 3: My neighbor got a little dog yesterday.

Document 4: Cats and dogs are mortal enemies.

Document 5: You mustn't feed peanuts to your dog.

The LDA model discovers the different topics that the documents represent and how much of each topic is present in a document. For example, LDA may produce the following results:

Topic 1: 30% peanuts, 15% almonds, 10% breakfast... (you can interpret that this topic deals with food)

Topic 2: 20% dogs, 10% cats, 5% peanuts... ( you can interpret that this topic deals with pets or animals)

Documents 1 and 2: 100% Topic 1

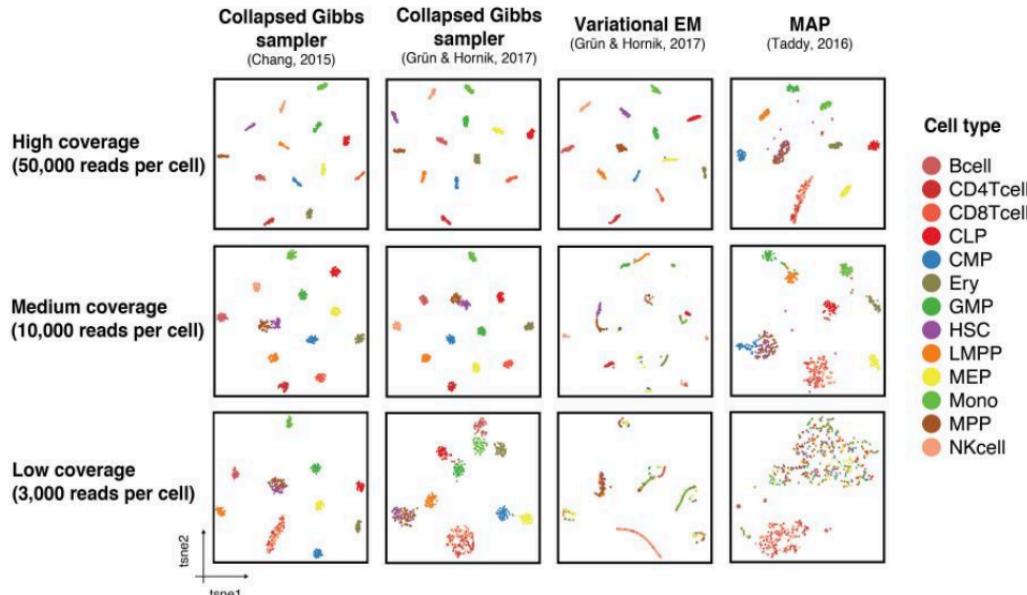
Documents 3 and 4: 100% Topic 2

Document 5: 70% Topic 1, 30% Topic 2

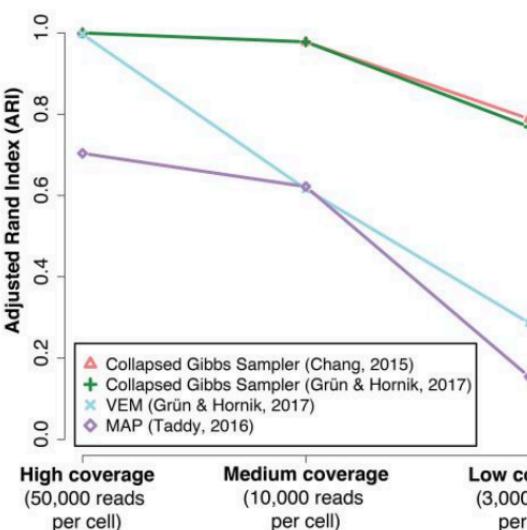
# Comparison of parameter estimation methods for latent Dirichlet allocation (LDA)

- Collapsed Gibbs Sampling
- Variational Expectation Maximization (VEM)
- Maximum a Posteriori (MAP)

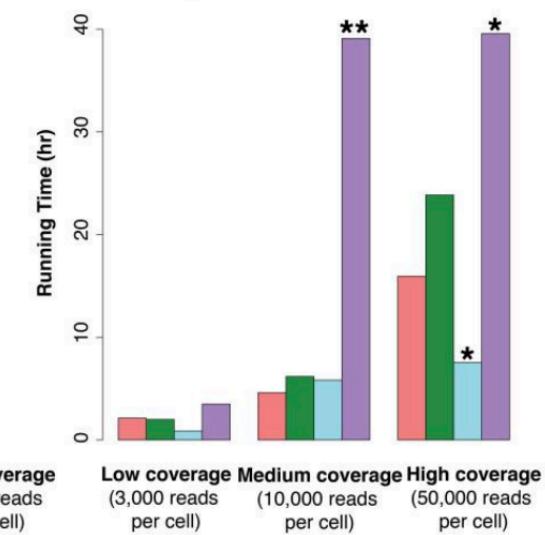
a Cell clustering with different parameter estimation algorithms for LDA



b Adjusted Rand Index



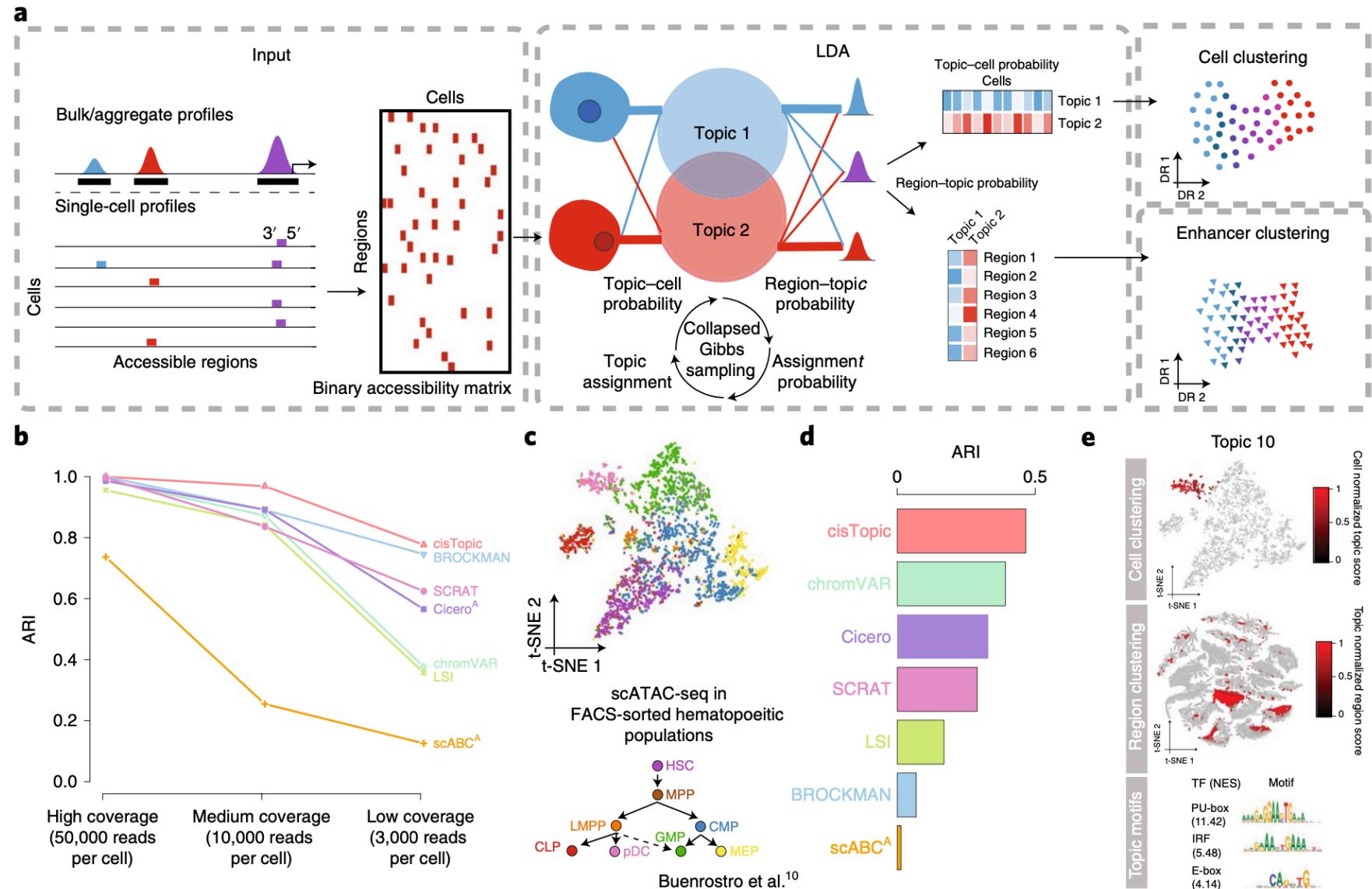
c Running time



# The cisTopic workflow

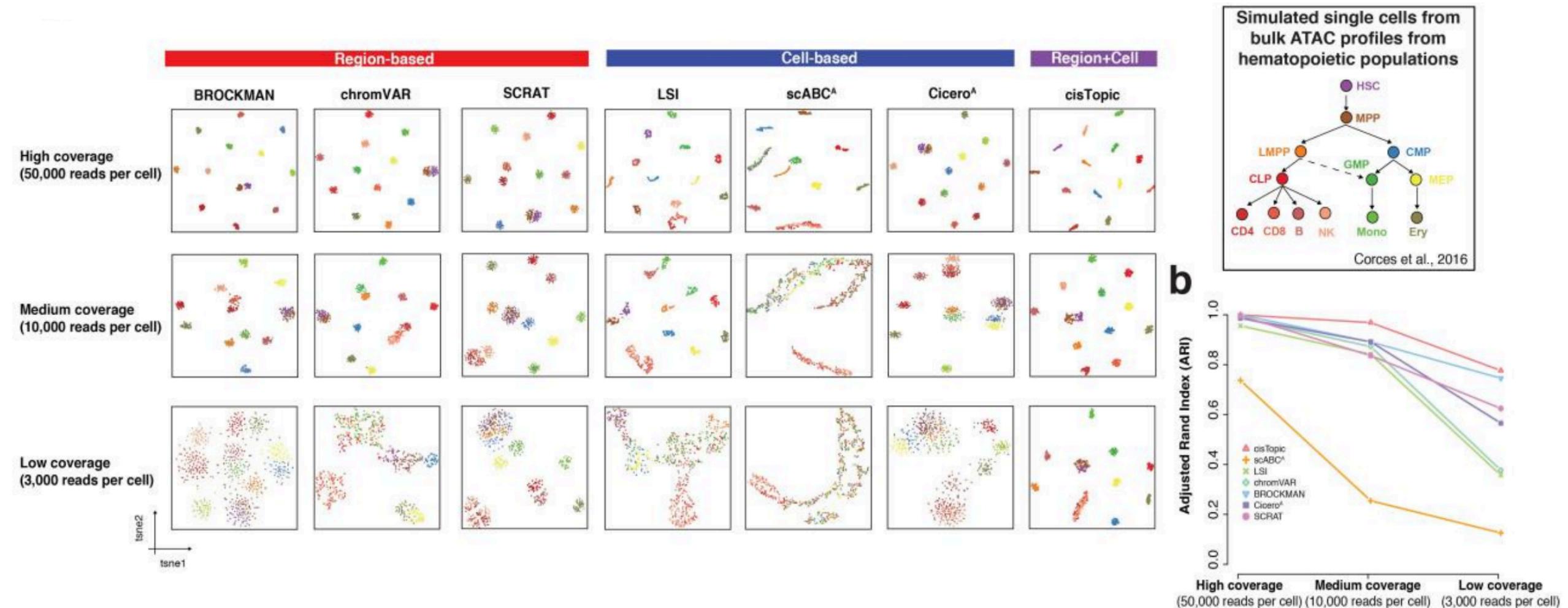
cisTopic consists of 4 main steps:

- (1) generation of a binary accessibility matrix as input for LDA;
- (2) LDA and model selection;
- (3) cell-state identification using the topic-cell distributions from LDA; and
- (4) exploration of the region-topic distributions.

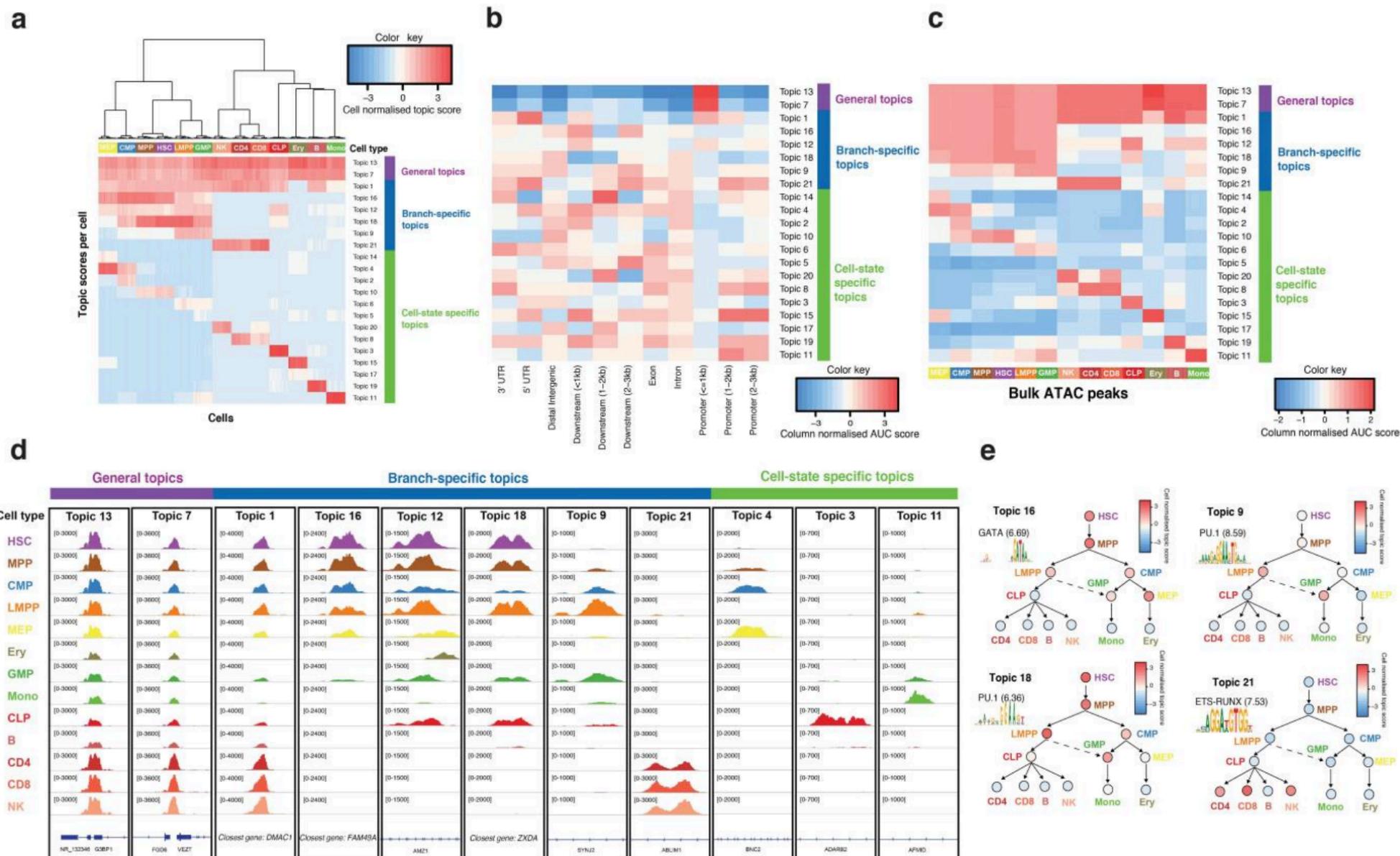


- (1) the probability of a region belonging to a topic (region-topic distribution) and
- (2) the contribution of a topic within a cell (topic-cell distribution)

# Benchmark of single-cell ATAC-seq analysis methods using simulated single cells from bulk ATAC-seq profiles from the hematopoietic system.



# cisTopic on single-cell ATAC-seq data simulated from bulk ATAC-seq profiles from the hematopoietic system.



Thank you for your attention.