PeakVI: A Deep Generative Model for Single Cell Chromatin Accessibility Analysis

2021/11/03

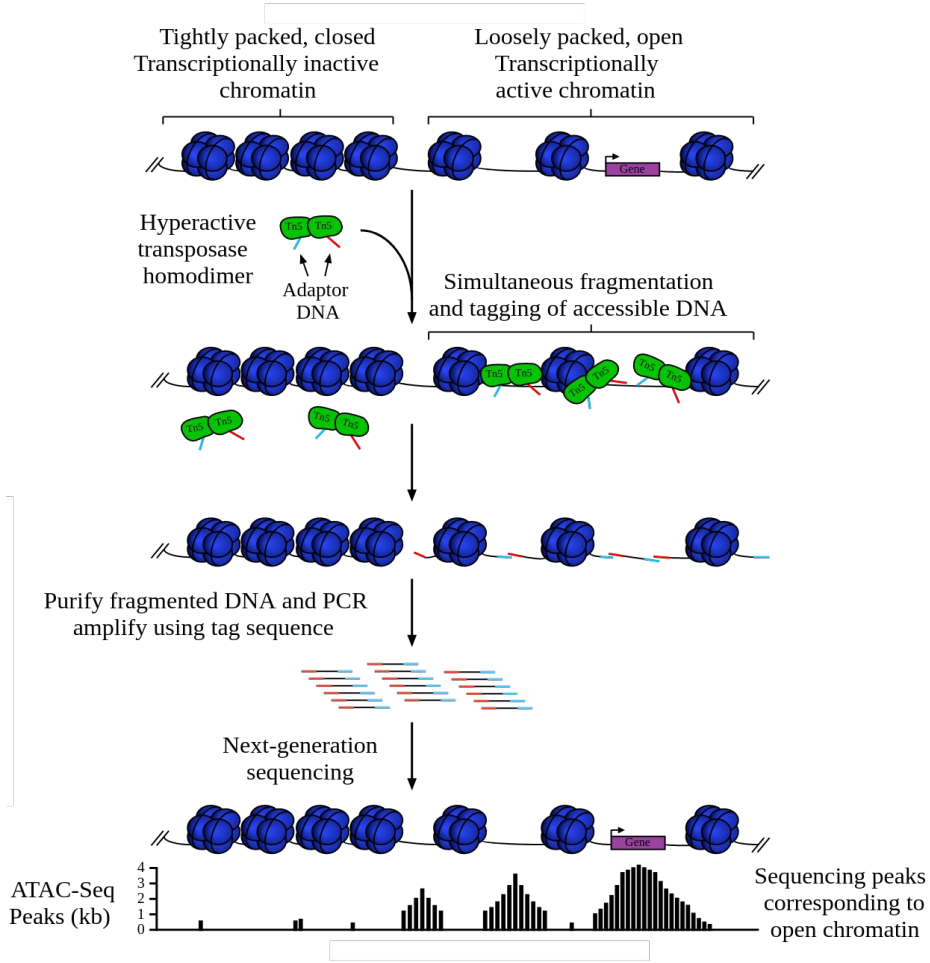
Jaime Abraham Castro Mondragon

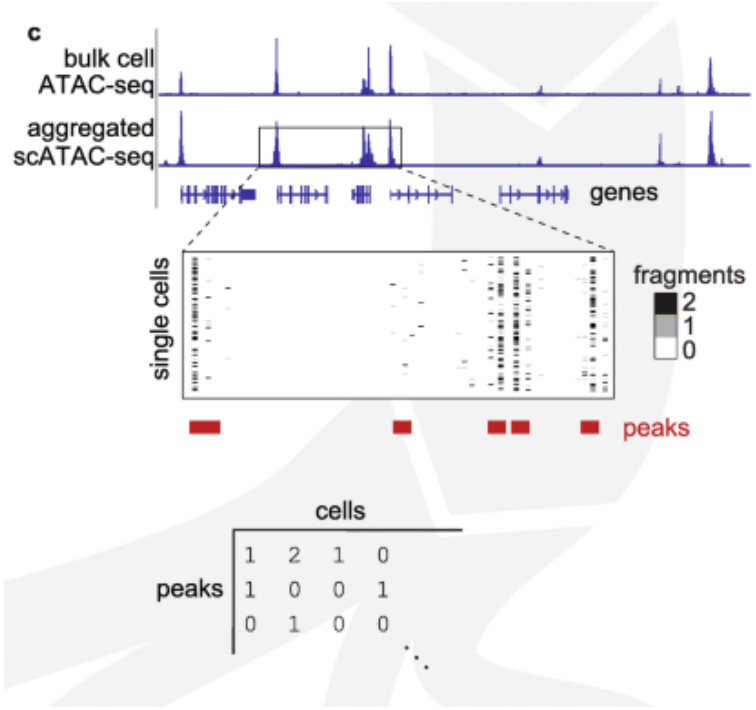
Ping-Han Hsieh

Introduction

- PeakVI proposes a probabilistic framework to analyze scATAC-Seq data.
- Difficulties in scATAC-Seq data analysis:
  - Limited sensitivity: 5-15% of accessible regions.
  - Limited coverage: 2 copies in a single cell.
  - High dimensionality: many genomic regions.
- Common task for scATAC-Seq data analysis:
  - Technical bias correction
    - Batch effect
    - Variation in sequence coverage
    - Width of DNA regions
  - Dimensionality reduction
  - Differential accessibility analysis

(sc)ATAC-seq (1)



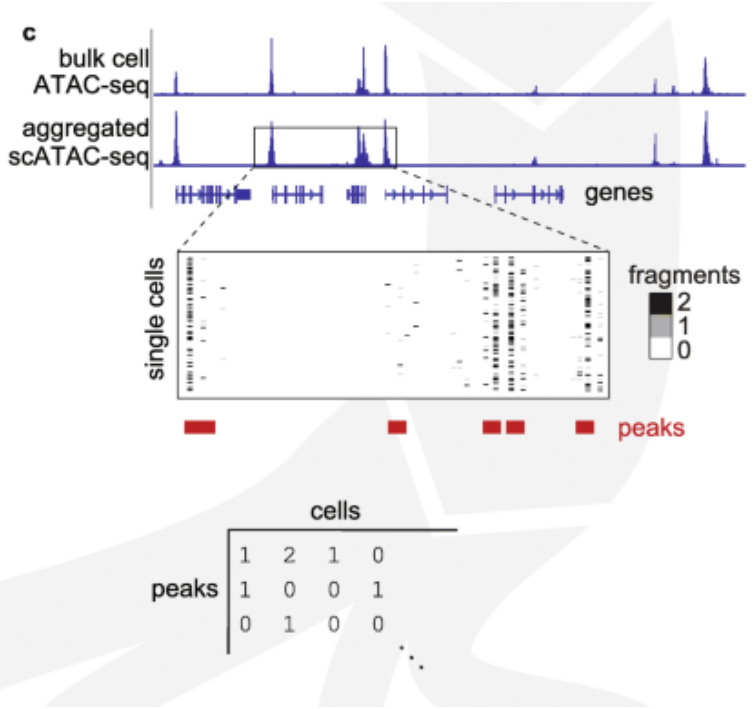


- ATAC-seq detects open-chromatin regions
- In scATAC-seq the peak signal is obtained by aggregation

1. https://commons.wikimedia.org/wiki/File:ATAC-Seq_Figure_.svg
2. Chen H et al. 2019 Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biology

(sc)ATAC-seq (2)

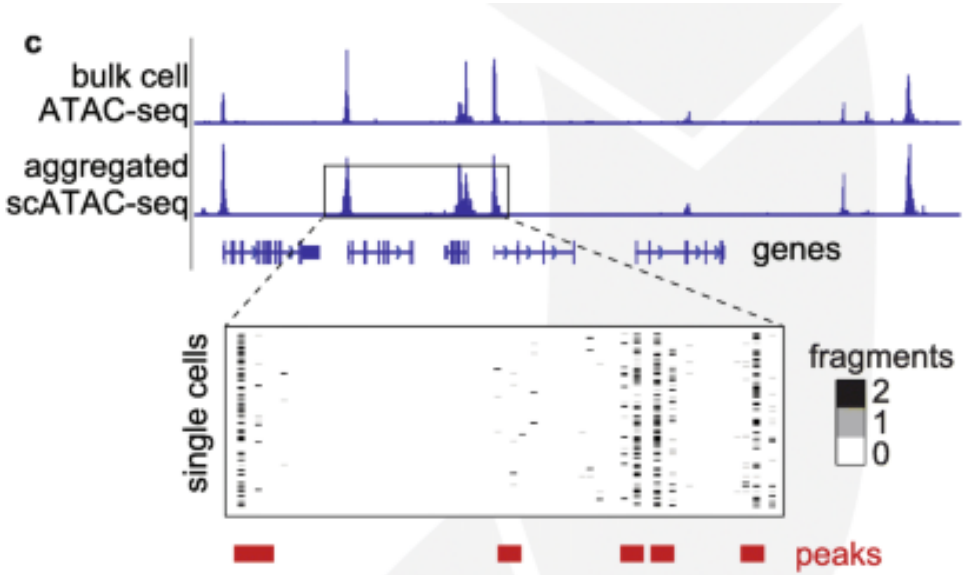
- Difficulties
  - 5-15% of open regions detected per cell
  - Max 2 fragments detected per cell (in humans, 2 alleles per region)
  - Multidimensional: millions of regions
- Computational Methods
  - Signal Aggregation
    - Interpretable
    - Lower resolution, more difficult to identify heterogeneity
  - Deep generative models:
    - Overfitting: more variables than observations
  - Text mining approaches:
    - Latent Dirichlet allocation (LDA) or latent semantic analysis (LSA).
    - Some biological/technical issues cannot be translated to a text mining framework (e.g., batch effect)



1. Chen H et al. 2019 Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biology

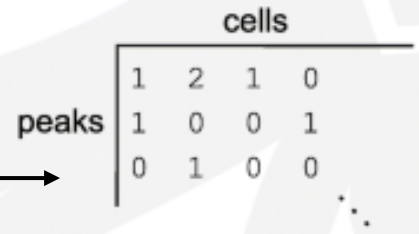
(sc)ATAC-seq (3)

- Differentially Accesible Regions → Cell heterogeneity
- Aggregation approaches
  - No individual cell specificity
- Linear models, RNA-seq model, classic statistical test:
  - Do not account for data sparsity



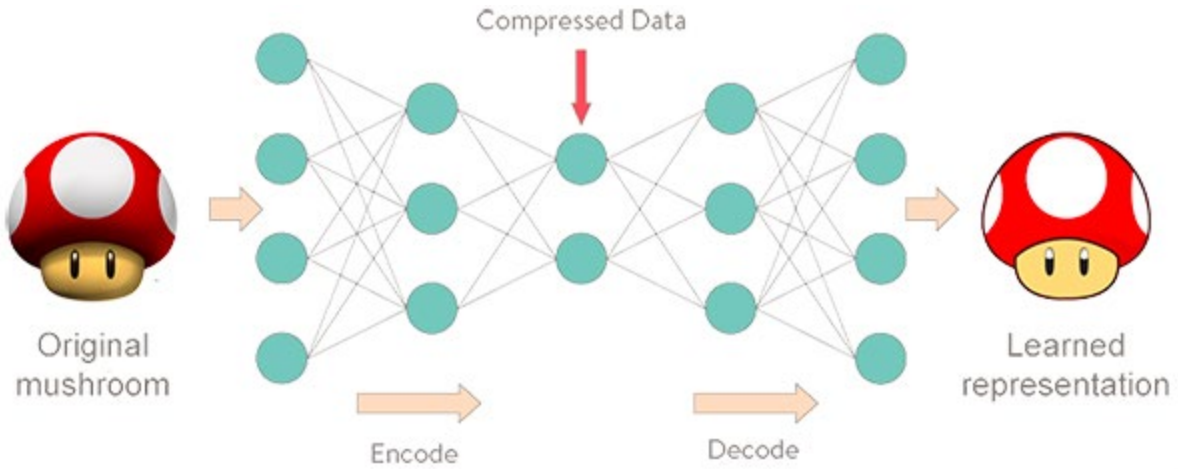
scATAC-seq matrix

- Entries may be the number of reads
- May be binarized: accessible or not



1. Chen H et al. 2019 Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biology

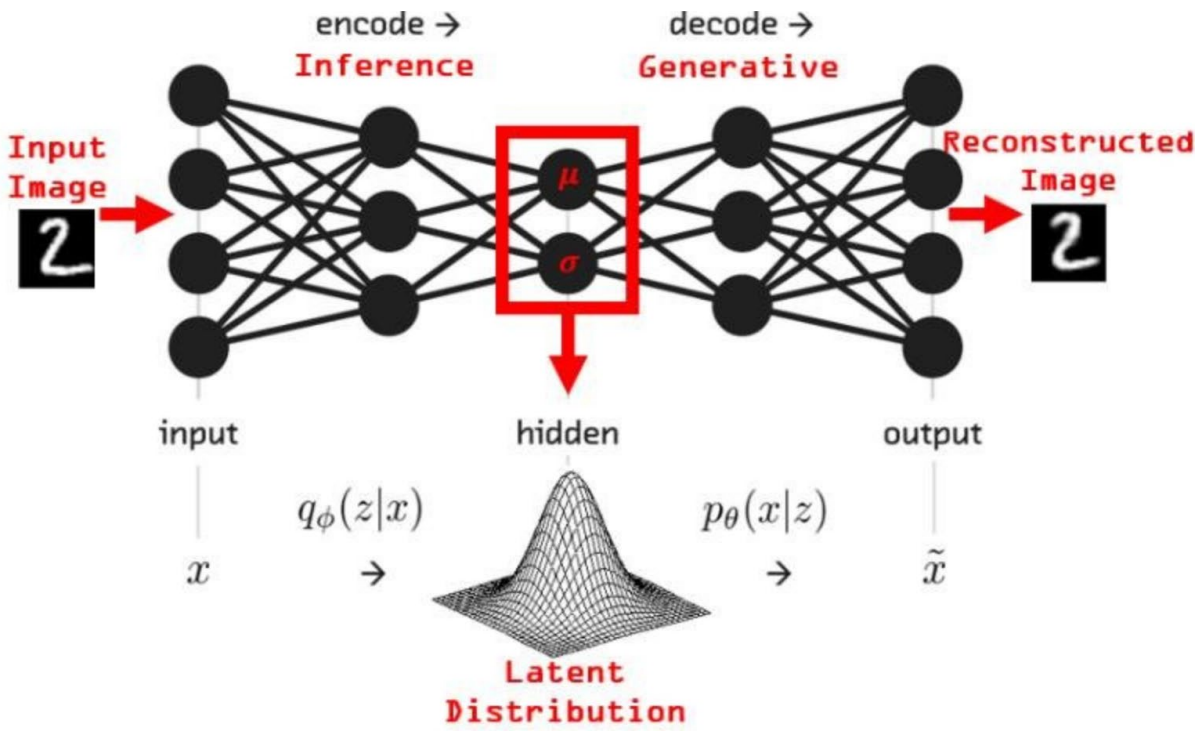
Autoencoders



- Original object is compressed (latent representation in a lower dimension) and reconstructed
- Output has the same dimension as input
- Useful for denoising

1. https://github.com/tchanda90/autoencoder

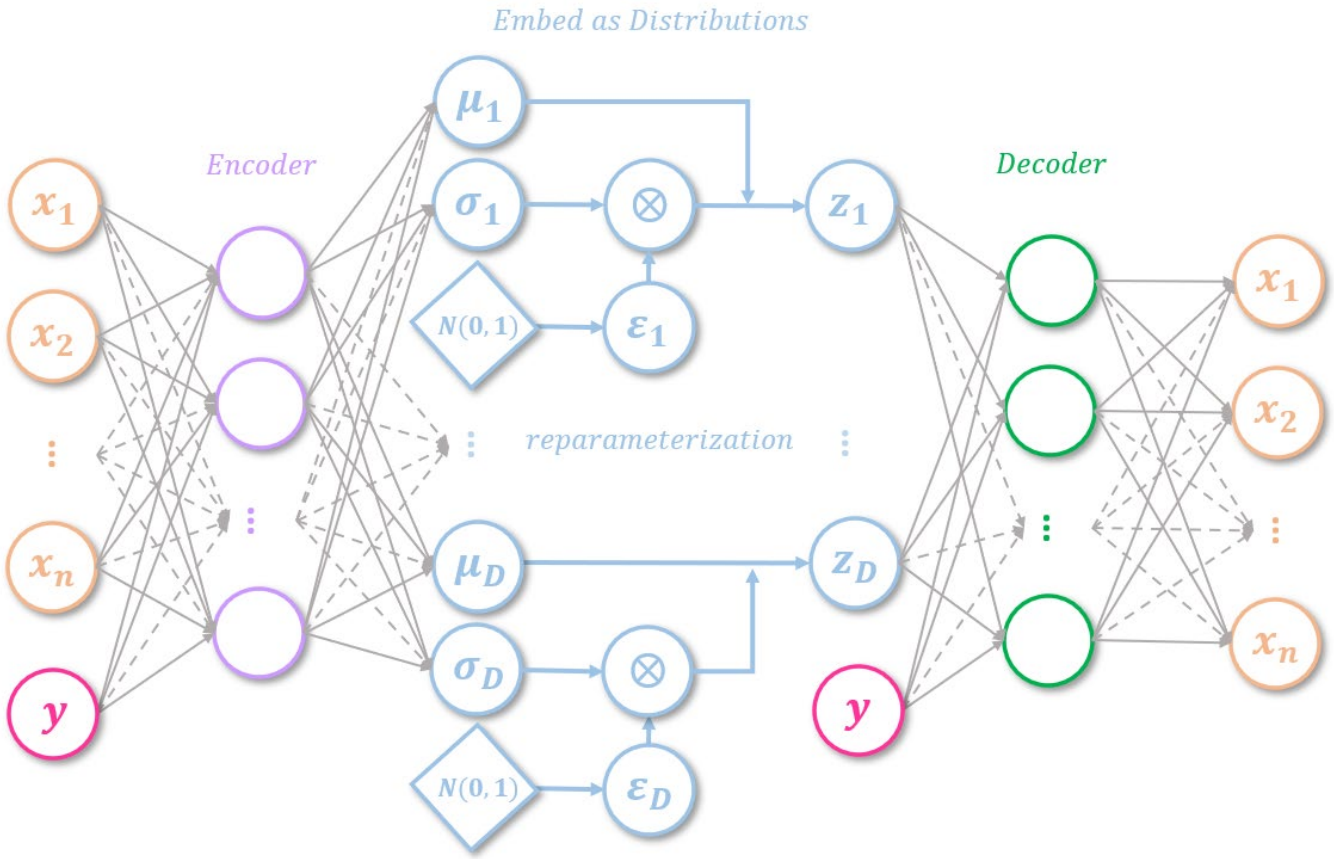
Variational Autoencoders



- The latent representation are parameters used to generate a distribution to sample points from it
- Probabilistic manner for describing an observation in latent space

1. https://www.kaggle.com/basu369victor/generate-music-with-variational-autoencoder

Conditional Variational Autoencoder



ELBO

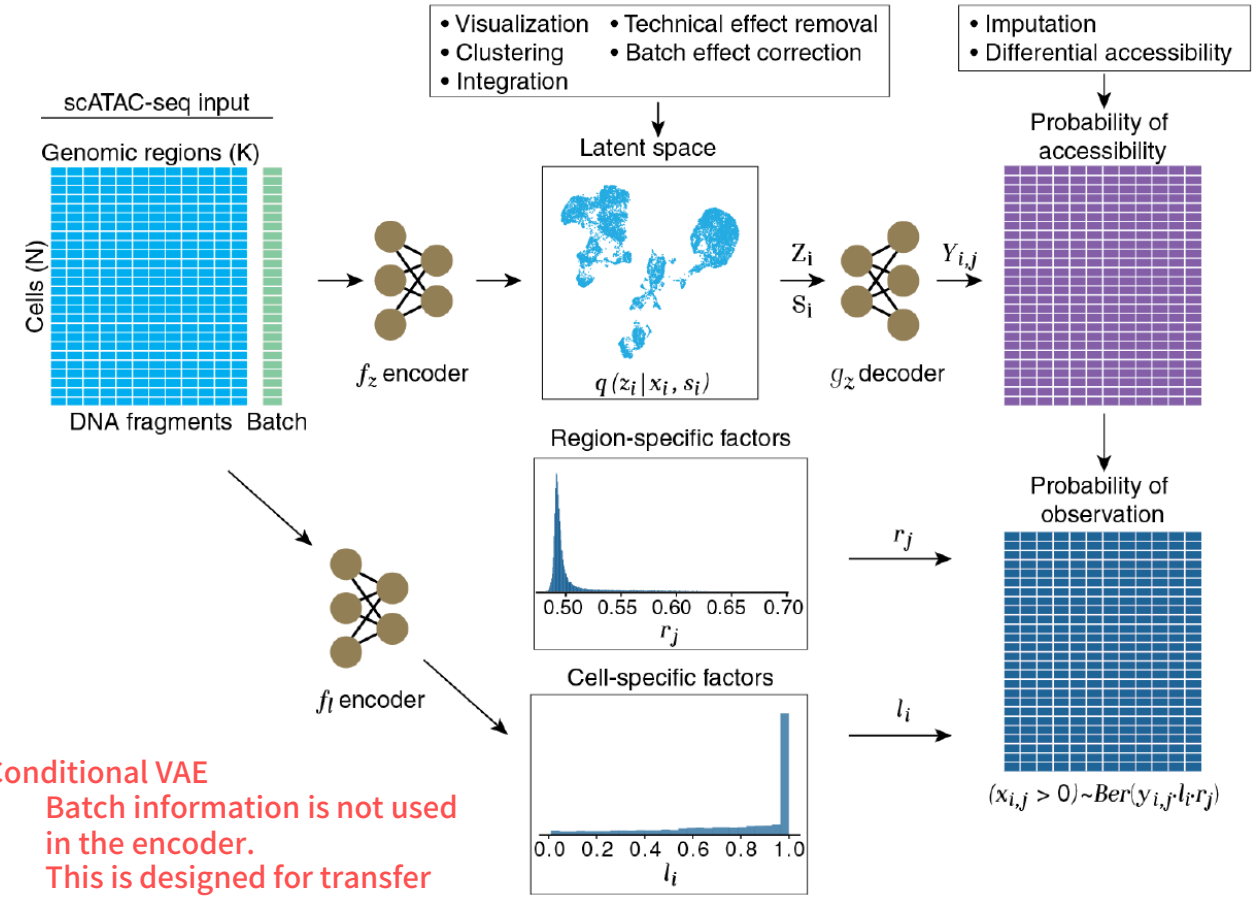
$$\log p_\theta(y|x) \geq \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(y|x,z)]$$
$$-KL(q_\phi(z|x,y)||p_\theta(z|x))$$

The reconstruction loss is a (log-) **likelihood function**:

- If the data distribution is Gaussian distribution, we could use **mean square error**.
- If the data distribution is Multinomial (Bernoulli) distribution, we could use **cross entropy**.
- If the data distribution is more complicated, e.g. Zero-inflated Negative Binomial (ZINB), the log-likelihood function is used instead.

Sohn *et al.,* NeurIPS, 2015

The PeakVI Model (1)



Conditional VAE
- Batch information is not used in the encoder.
- This is designed for transfer learning.

s_i:one hot-encoded batch information

The distribution of the observed scATAC-Seq is modeled as **Bernoulli distribution.**

The probability is parameterized with three parameters:

$$\pi_i = y_{ij} \cdot l_i \cdot r_j$$

: index for cell
index for genomic region

Region specific factor
(used to width of region/sequence content)

Cell specific factor
(used to correct batch effect)

Probability of accessibility

The PeakVI Model:

$$\begin{aligned}
(\mu_i, \sigma_i) &= f_z(x_i) \\
z_i &\sim \mathcal{N}(\mu_i, \sigma_i) \\
y_{ij} &= (g_z(z_i, s_i))_j \\
\ell_i &= f_\ell(x_i) \\
x_{ij} > 0 &\sim Ber(y_{ij} \cdot \ell_i \cdot r_j)
\end{aligned}$$

The PeakVI Model (2)

- Encoder (model latent representation)

$$FC\left(N, \sqrt{N}, 0.1, \text{leakyReLU}\right) \rightarrow$$
$$FC\left(\sqrt{N}, \sqrt{N}, 0.1, \text{leakyReLU}\right) \rightarrow$$
$$FC\left(\sqrt{N}, \sqrt{N}, 0.1, \text{leakyReLU}\right) \rightarrow$$
$$\left(FC\left(\sqrt{N}, \sqrt[4]{N}, 0.1, \text{Identity}\right), FC\left(\sqrt{N}, \sqrt[4]{N}, 0.1, \text{Identity}\right)\right)$$

- Decoder (reconstruction)

$$FC\left(\sqrt[4]{N} + S, \sqrt{N}, 0, \text{leakyReLU}\right) \rightarrow$$
$$FC\left(\sqrt{N}, \sqrt{N}, 0, \text{leakyReLU}\right) \rightarrow$$
$$FC\left(\sqrt{N}, \sqrt{N}, 0, \text{leakyReLU}\right) \rightarrow$$
$$FC\left(\sqrt{N}, N, 0, \text{sigmoid}\right)$$

- Encoder (model cell specific factor)

$$FC\left(N, \sqrt{N}, 0, \text{leakyReLU}\right) \rightarrow$$
$$FC\left(\sqrt{N}, \sqrt{N}, 0, \text{leakyReLU}\right) \rightarrow$$
$$FC\left(\sqrt{N}, \sqrt{N}, 0, \text{leakyReLU}\right) \rightarrow$$
$$FC\left(\sqrt{N}, 1, 0, \text{sigmoid}\right)$$

- Encoder (region specific factor)

$$softmax(\cdot)$$

not depending on the observed data

The PeakVI Model (3)

- Training procedure
  - Optimizer
    - AdamW (Adam with weight decay)
    - Learning rate: 0.0001
  - Minibatch
    - 128
  - Training/Validation Split:
    - 0.9/0.1
  - Warmup Epochs
    - 50 (slowly increase the coefficient of KL divergence in 50 epochs)
  - Early stopping

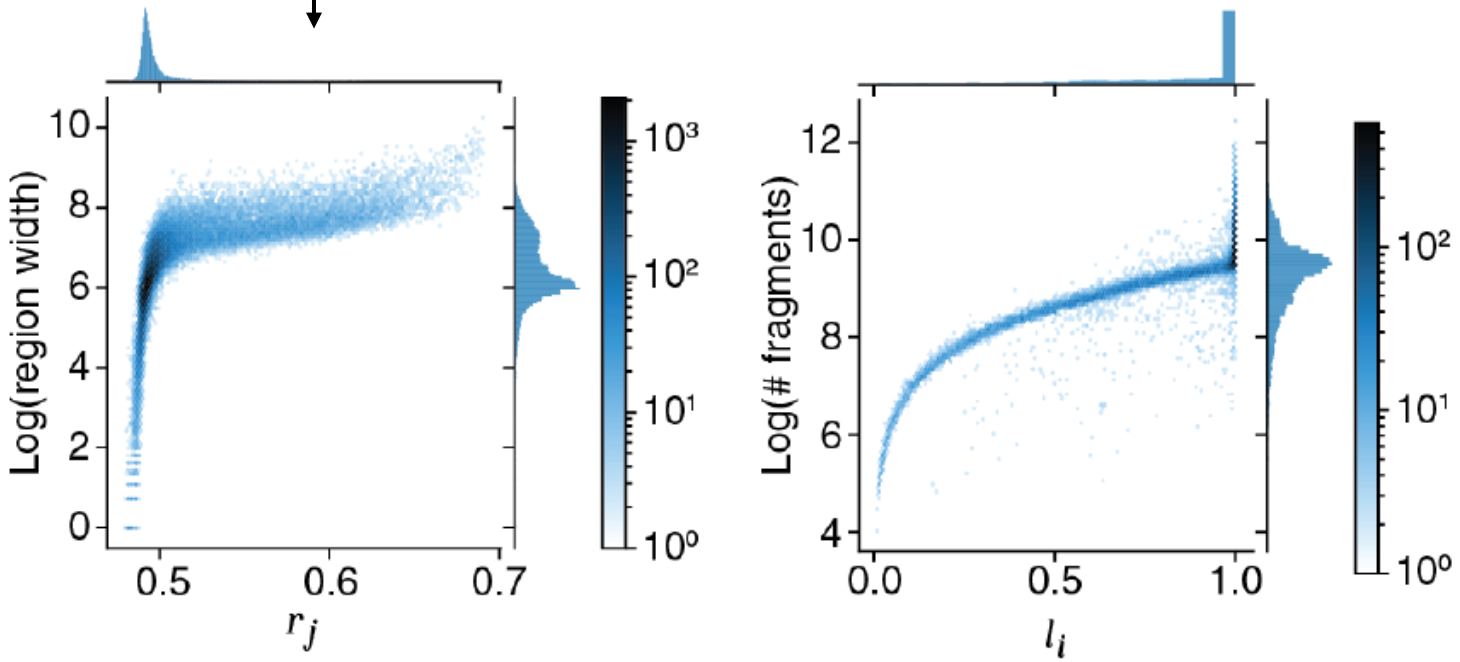
Benchmark Datasets

- Hematopoiesis data (Satpathey *et al.,*)
  - Consists of bone marrow and blood samples
  - Flow-sorted for cell subsets (contains cell labels).
  - Contains batch information for unsorted samples.
  - Used to evaluate **cell type clustering** and **batch correction**.

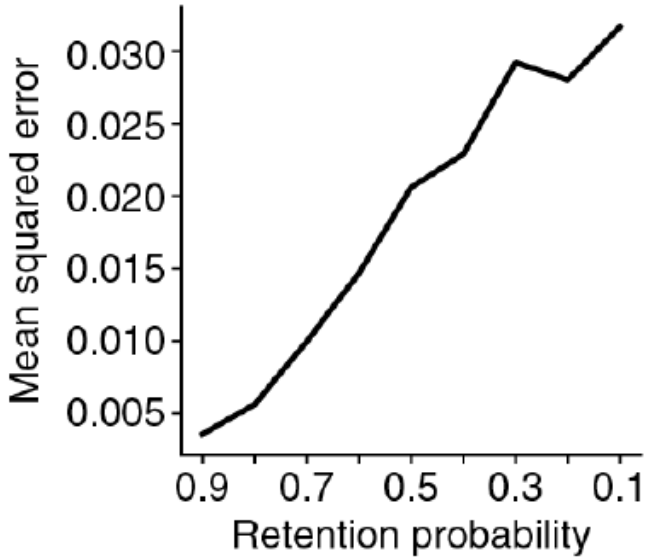
- Peripheral Blood Mono-nuclear cells (PBMCs) from 10x Genomics
  - Joint scRNA-Seq and scATAC-Seq
  - Use scRNA-Seq (**orthogonal modality**) to validate the result

PeakVI Captures Region and Cell Specific Confounders

Wider region has higher regional specific factor
(The technical bias is explained by , leaving clearer embedding for)



Corruption on the data leads to only small increase of error

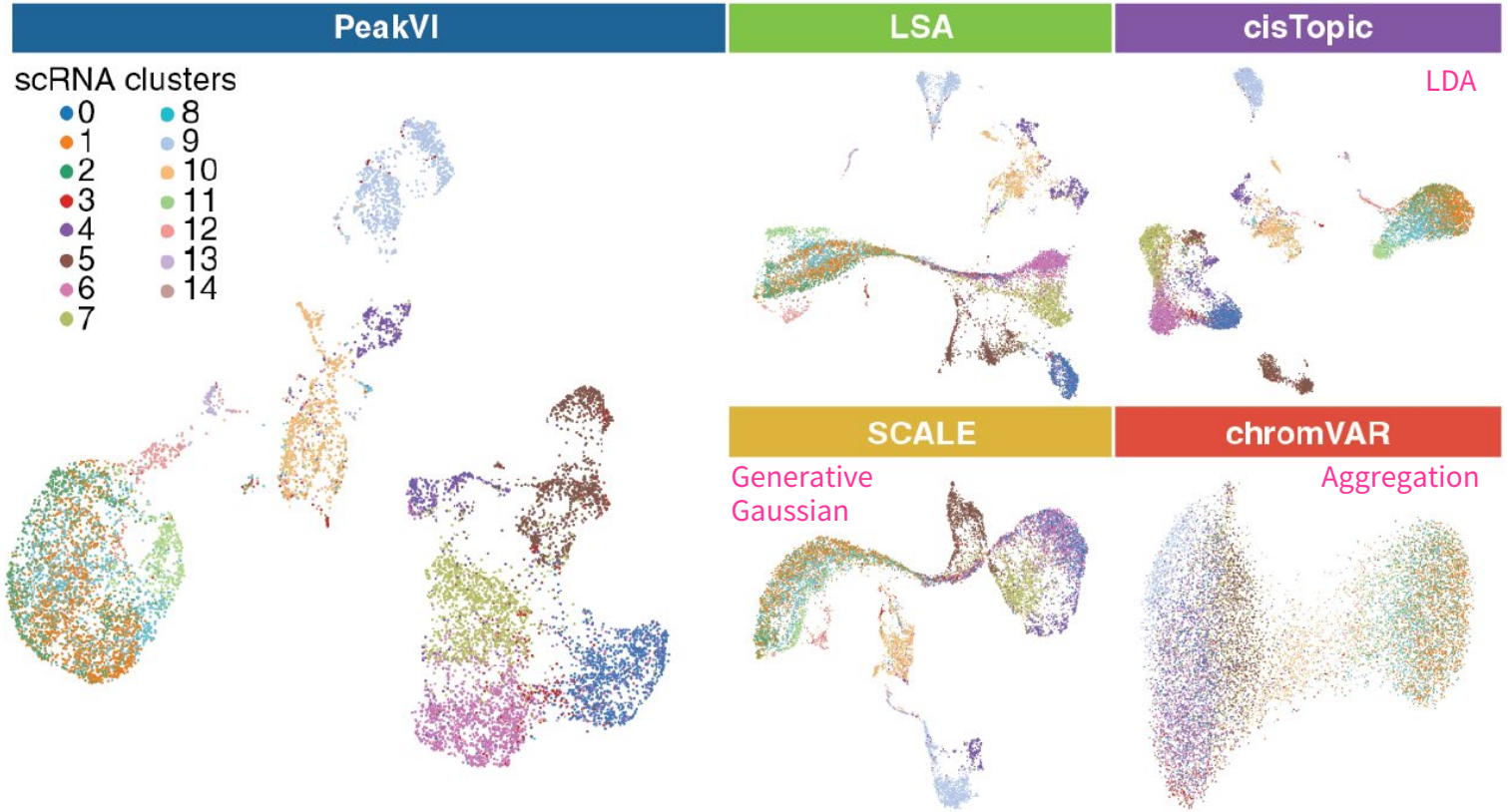


Method: Section 1.3

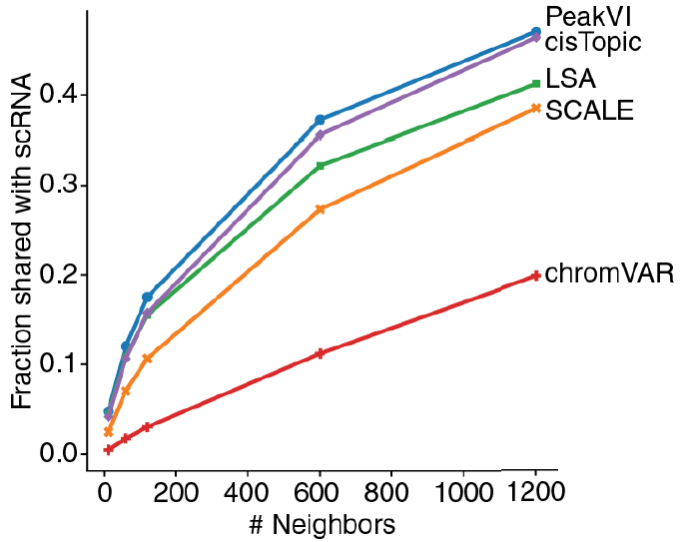
Larger library size has higher cell specific factor
(The technical bias is explained by , leaving clearer embedding for)

PeakVI Learns Batch Corrected Latent Representation (1)

Applied to PBMC data, cluster defined by scRNA-Seq



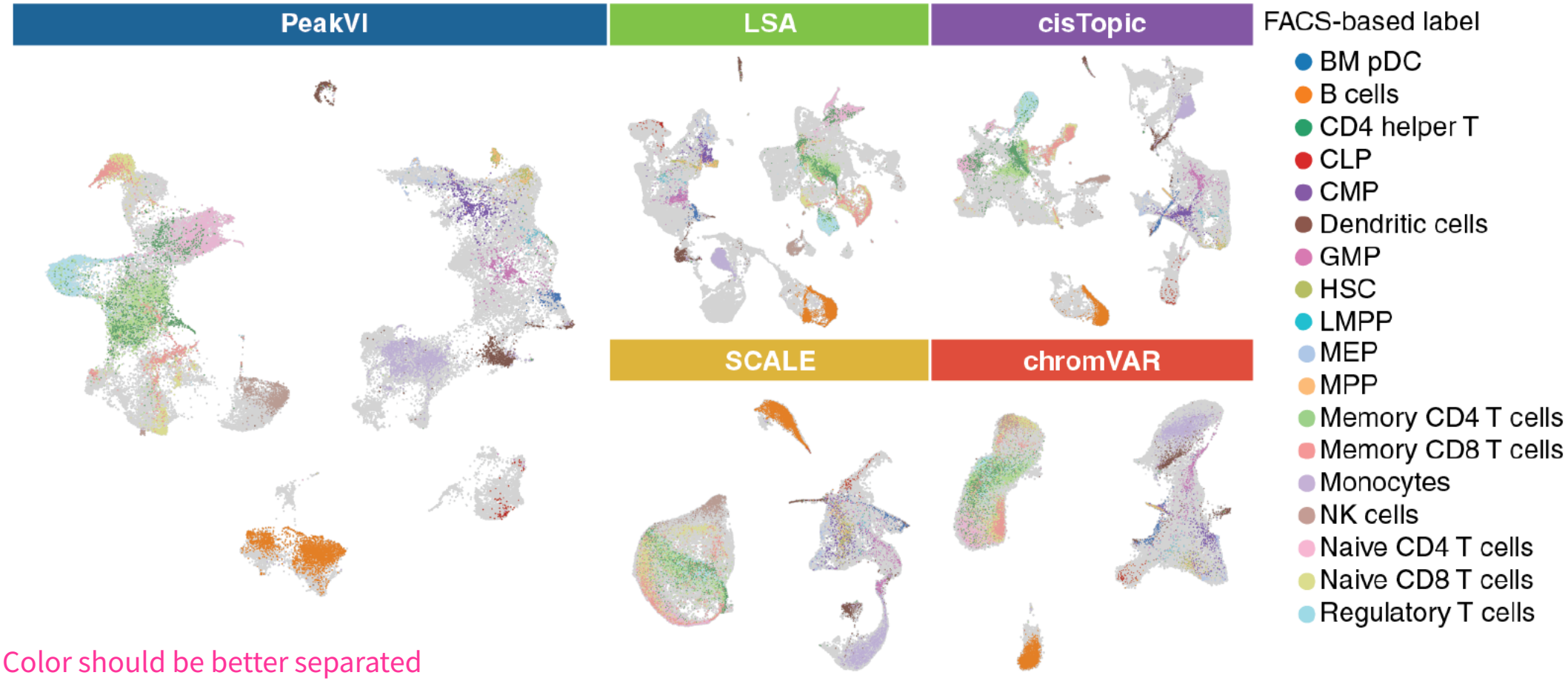
Color should be better separated



Check if the K nearest neighbors belong to the same cluster defined by scRNA-Seq

PeakVI Learns Batch Corrected Latent Representation (2)

Applied to Hematopoiesis data, cluster defined by FACs-based Label

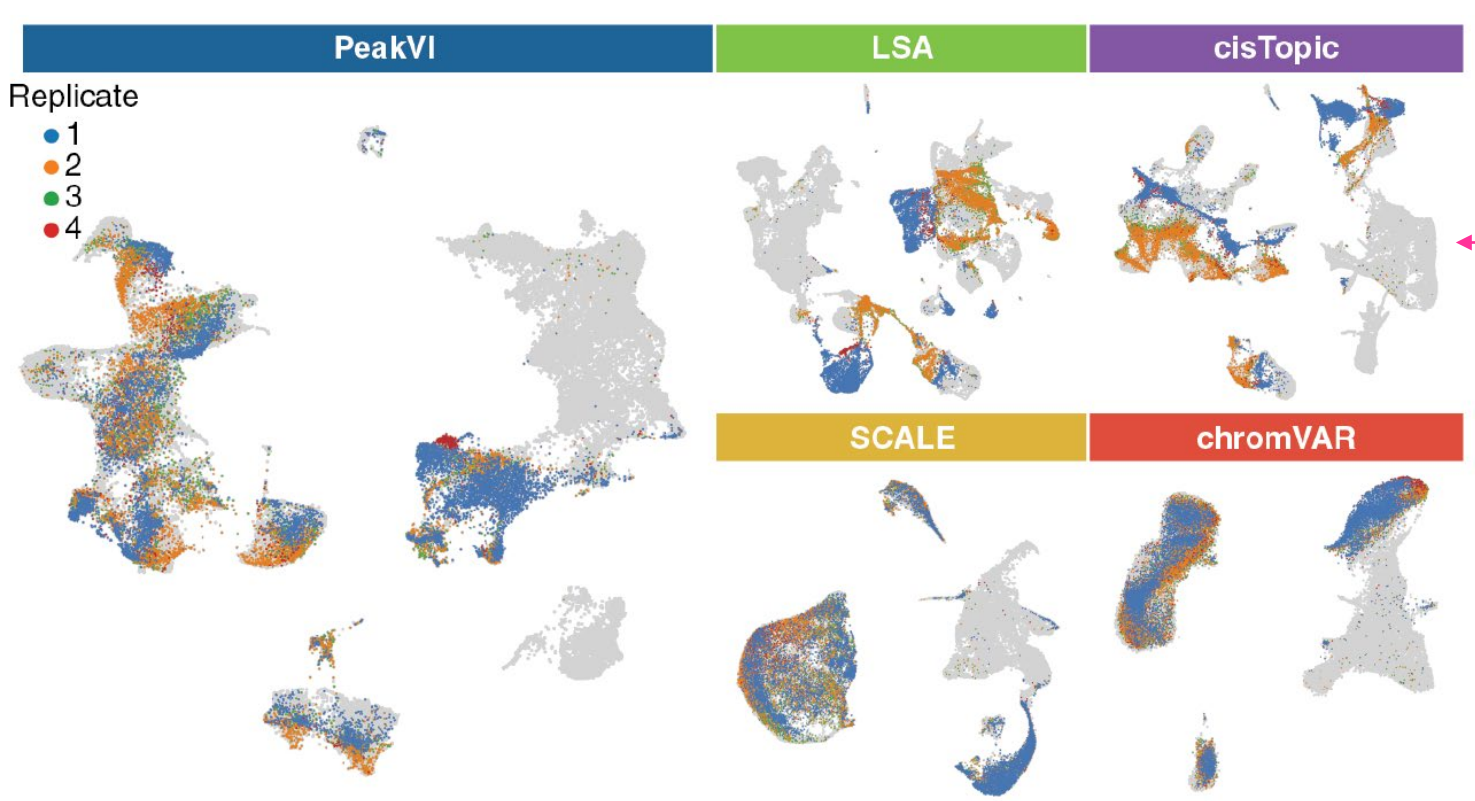


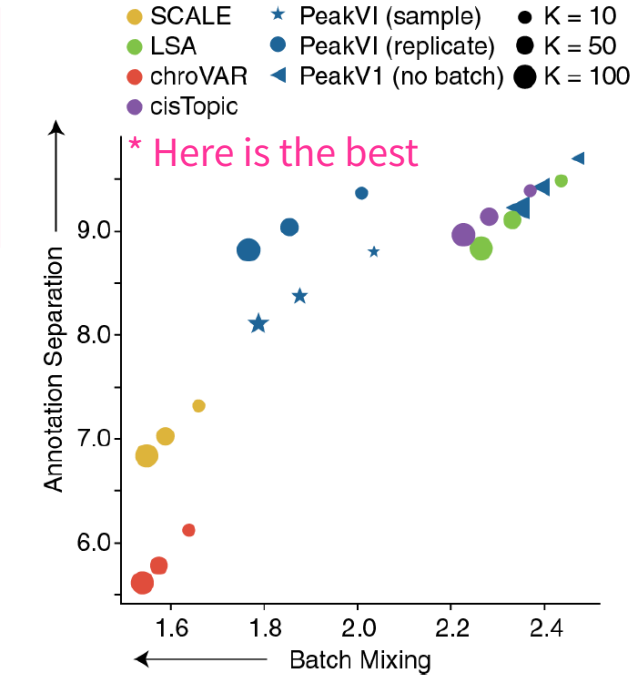
Color should be better separated

PeakVI Learns Batch Corrected Latent Representation (3)

Applied to Hematopoiesis data, cluster defined by Batches

Text-mining based approached does not integrate batch effects in their model





Color should be better mixed

PeakVI provides a balance representation to preserve biological variability and reduce technical bias.

* Cells separated by cell-label and by batch

Differential Accessibility Analysis (1)

- Computed based on Bayesian Factor (odds ratio instead of probability for Pvalue)
- Given two population A and B:
  1. Sample cells from each population with replacement.
  2. Apply forward pass of PeakVI model.
     - Recall that the latent representation is a distribution for each cell, therefore we could sample multiple times for one cell (**increasing statistical power** beyond original sample size)
  3. Compute absolute difference between the average accessibility across two population:

$$\Delta_j = (\bar{y_A})_j - (\bar{y_B})_j$$

For more detail comparison between Bayesian Factor and Pvalue
Assaf and Tsionas Tourism Management, 2018.

Differential Accessibility Analysis (2)

- Given two population A and B:
  4. Randomly pair samples into pairs:

$$\left\{(y_A, y_B)^i \,|\, i \in [N]\right\}$$

  4. Count how many pairs yield the difference between estimates greater than a threshold:

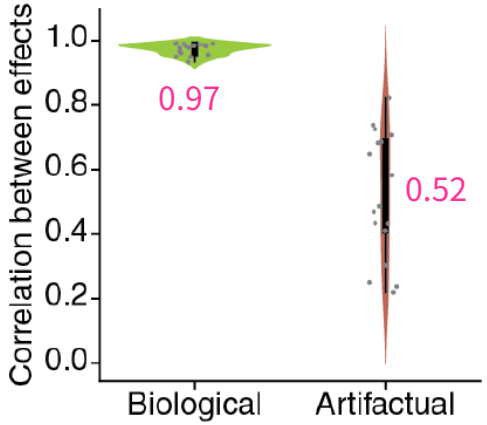
$$\frac{1}{N}\sum_i^N \mathbb{1}\left((y_A)_j^i - (y_B)_j^i > \delta\right)$$

  4. Compute Bayesian Factor

$$BF_j = \log \frac{p_{DA}}{1-p_{DA}}$$

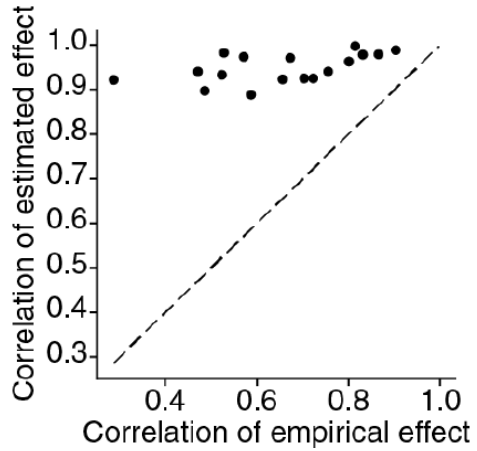
Differential Accessibility Analysis (3)

Experiment



Correlation between the **estimated effect** from PeakVI and **empirical effect** from observations

- Similar result for biological variability
- Less similar for technical variability

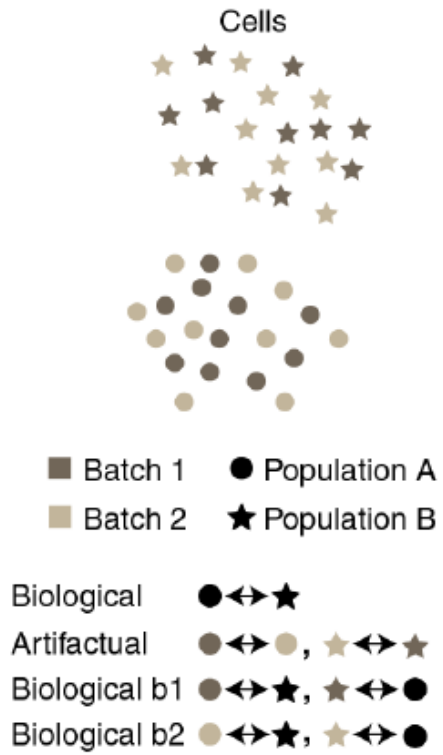


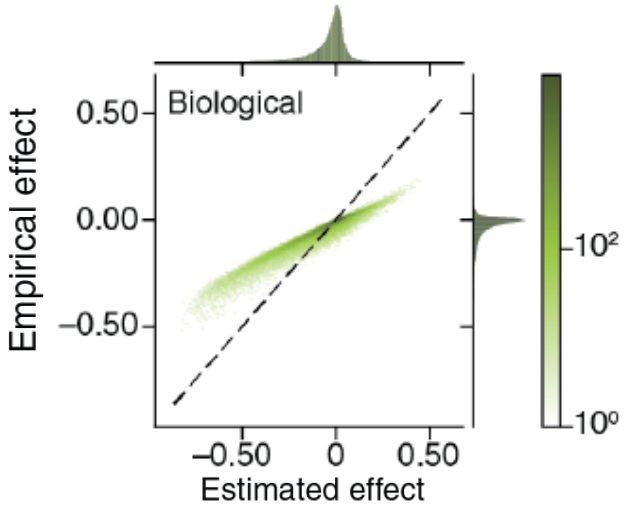
Correlation between **biological b1** and **biological b2:**

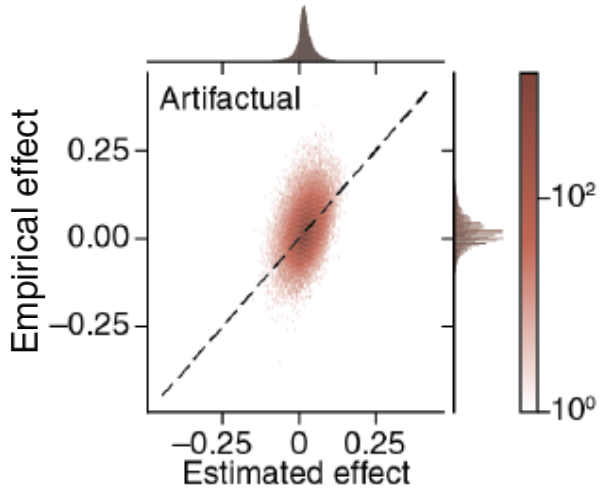
- We expect to have high correlation as the batch effect should not overshadow the biological variability
- But the empirical effect failed to deliver this.

Differential Accessibility Analysis (4)

Experiment



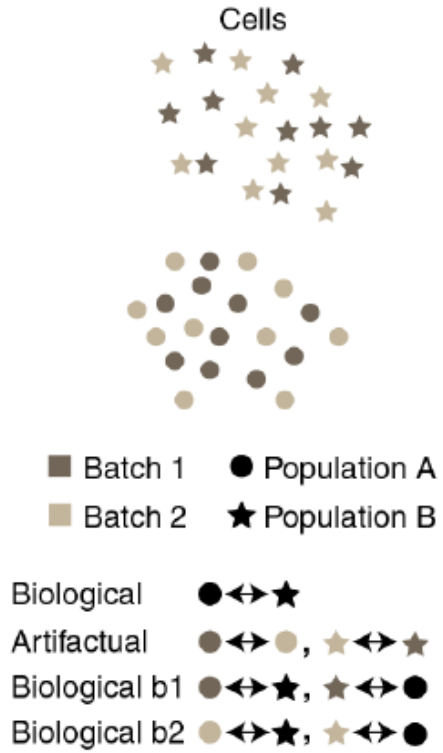




The distributions for the empirical effect on the artifactual effect is wider, but we expect the artifactual variability to be small

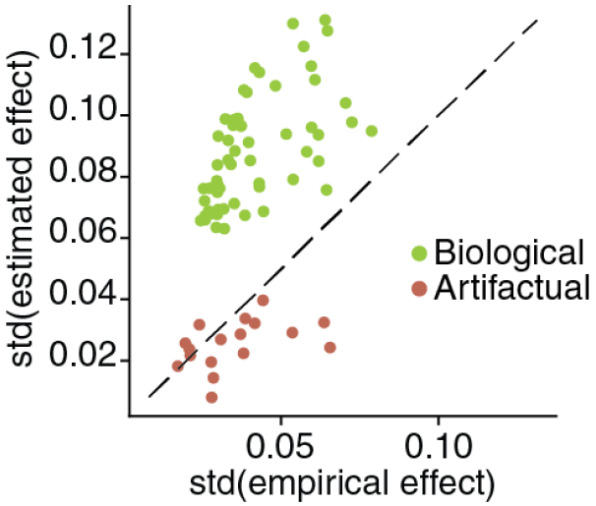
Differential Accessibility Analysis (5)

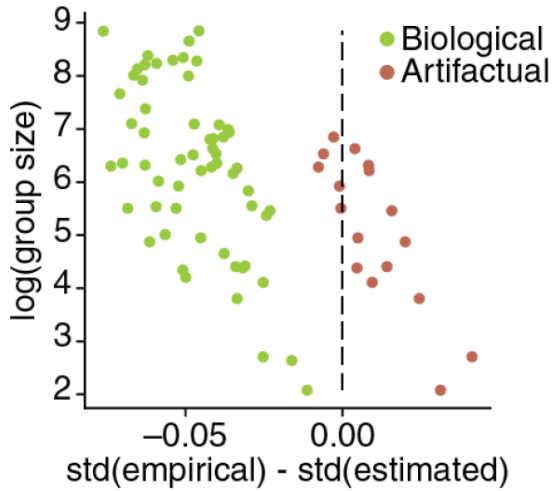
Experiment



Emperical effect

$$X_{C_j} = \sum_{i \in C} \mathbb{1}\left(x_{ij} > 0\right)$$

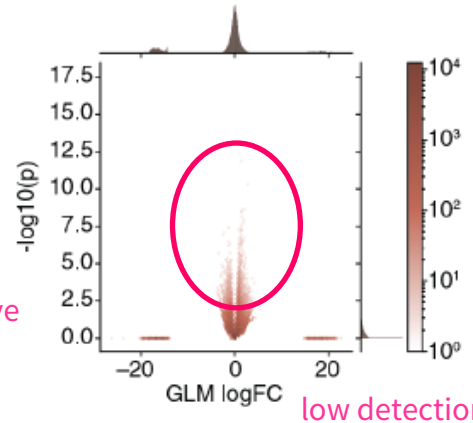




PeakVI amplifies the biological variability, and silence the variability caused from the product of noise.

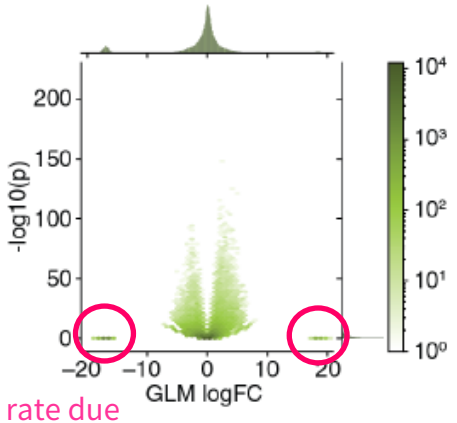
Differential Accessibility Analysis (6)

comparison between two batch of NK-cell



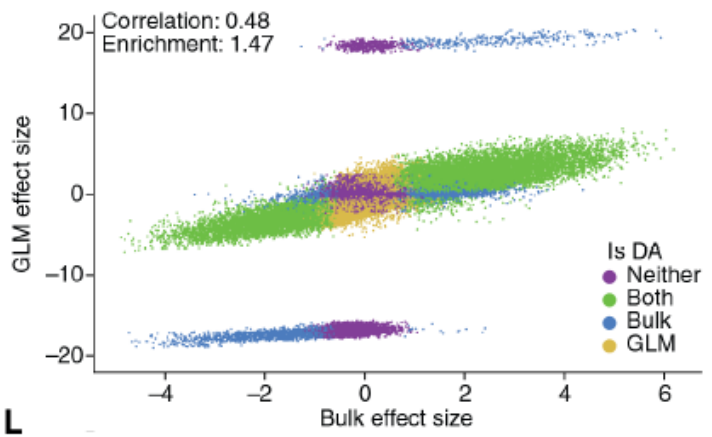
910 regions false positive

comparison between NK-cell and B-Cell



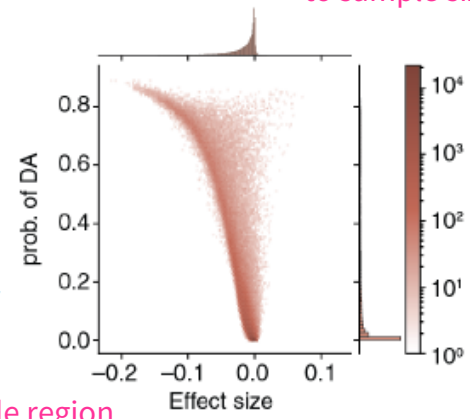
low detection rate due to sample size

comparison with bulk ATAC-Seq (Calderon et al.,)

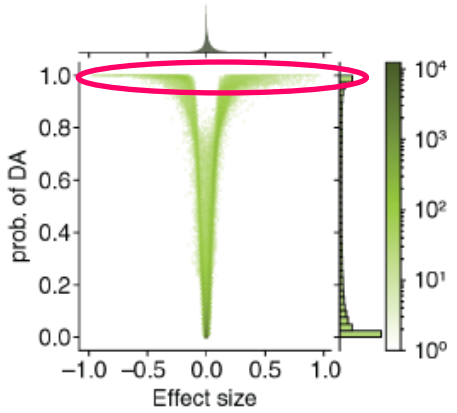


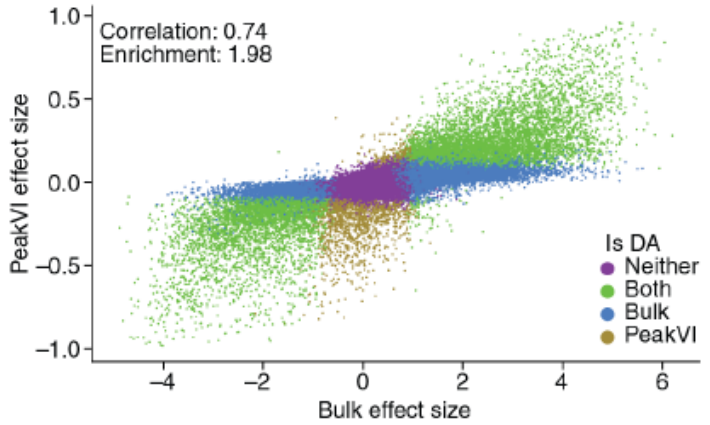
Estimated effect size

$$\Delta_j = (\bar{y_A})_j - (\bar{y_B})_j$$



no differentially accessible region (prob > 0.9995, BF > 3.2)

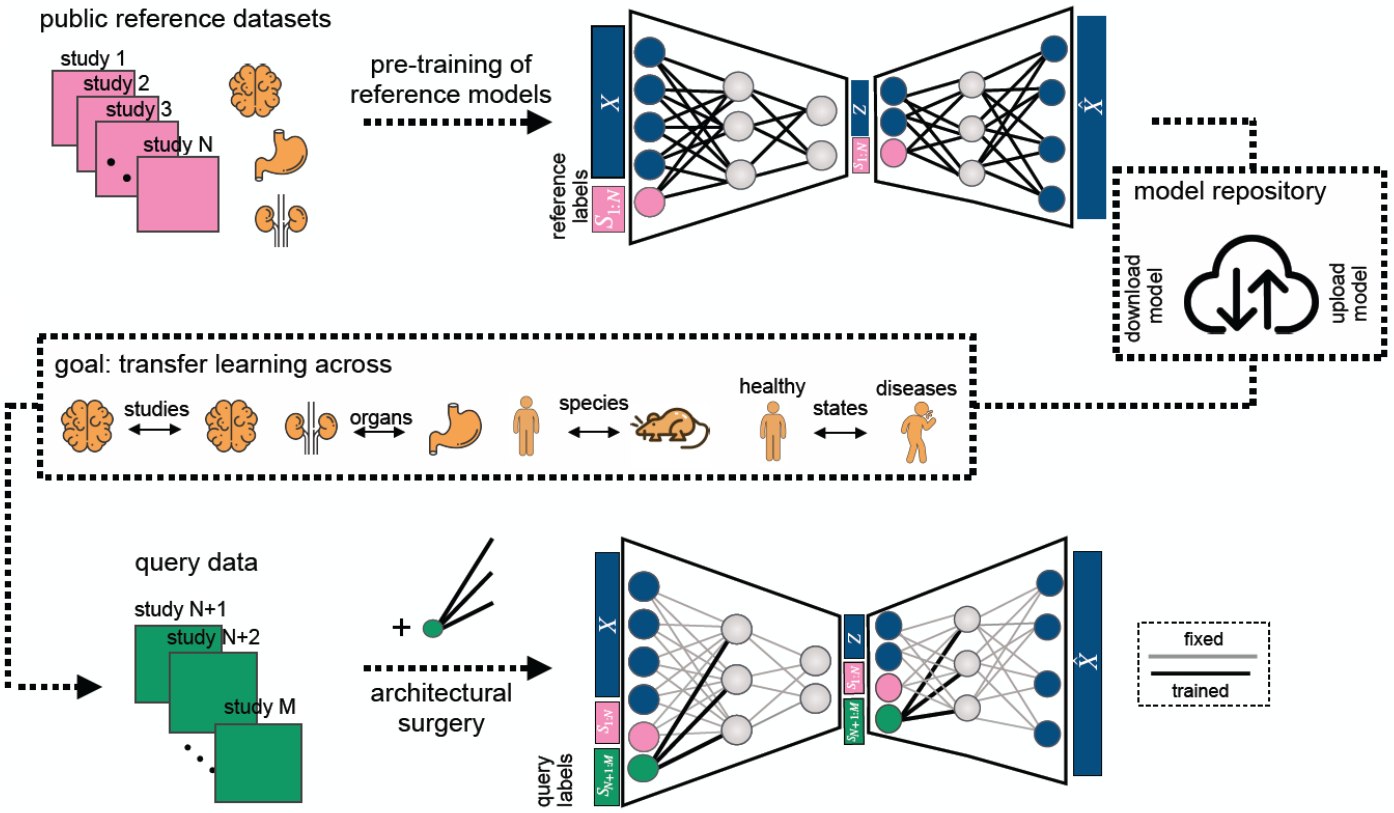




the differential accessibility identified by PeakVI is more in line with bulk ATAC-Seq

Transfer Learning (Query to Reference Integration) (1)

scArches



Fix the weight for the network for the reference, add additional dimensionality (batch information) to the network, and retrain it.

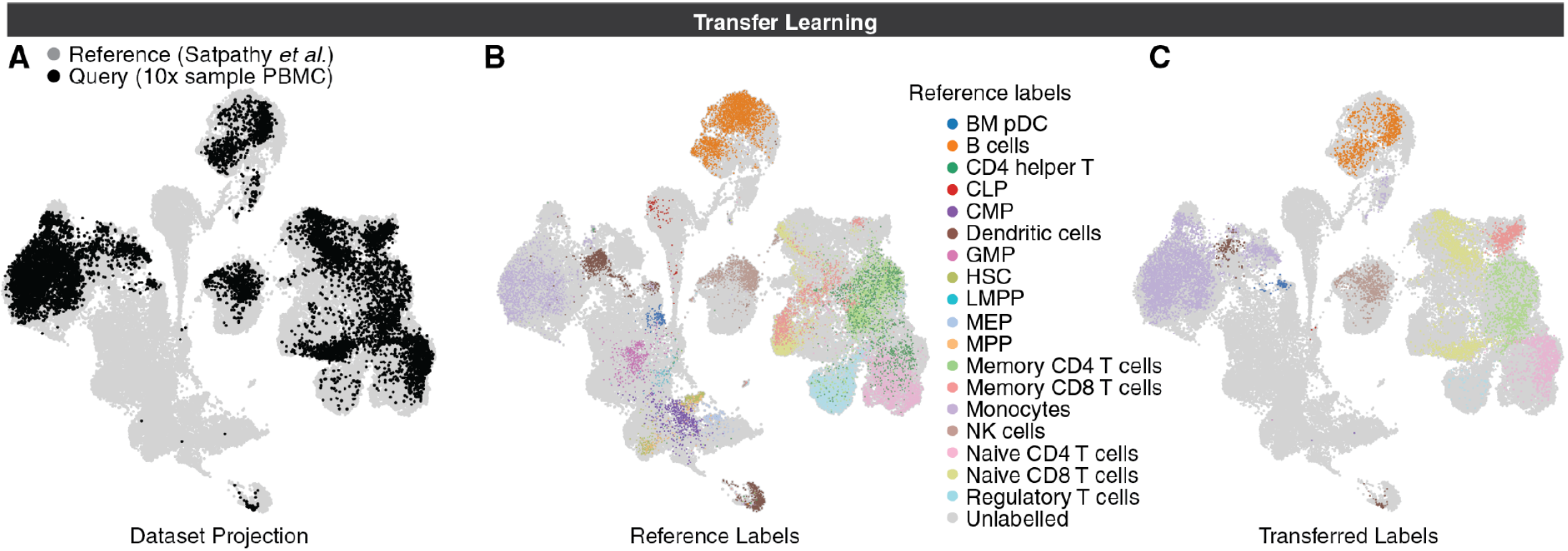
For PeakVI, the default is to only include batch information in the decoder, so the projection of the latent coordinates remains the same for the reference.

But for the evaluation, the authors add batch information to the encoder and use **scArches** to do transfer learning.

Lotfollahi *et al.*, bioRxiv 2020

Transfer Learning (Query to Reference Integration) (2)

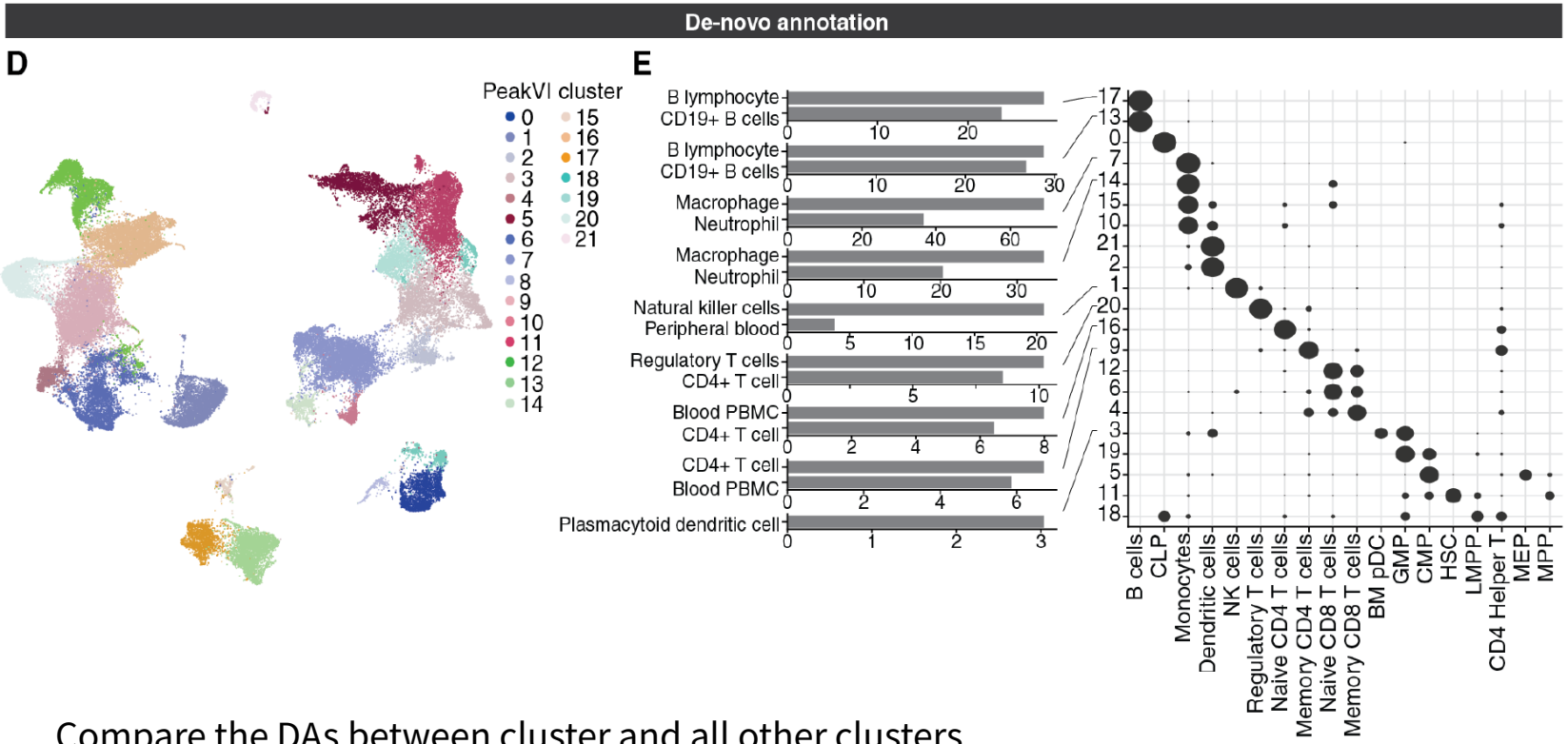
When cell type information is available



PeakVI can be used to infer the cell types of query dataset while considering the additional batch information that is not provided in the reference, without the need to retrain the entire model.

Transfer Learning (Query to Reference Integration) (4)

When cell type information is not available, but we believe each cell type correspond to known gene signature



Clusters defined by **cluster majority vote** (or K-nearest neighbors)

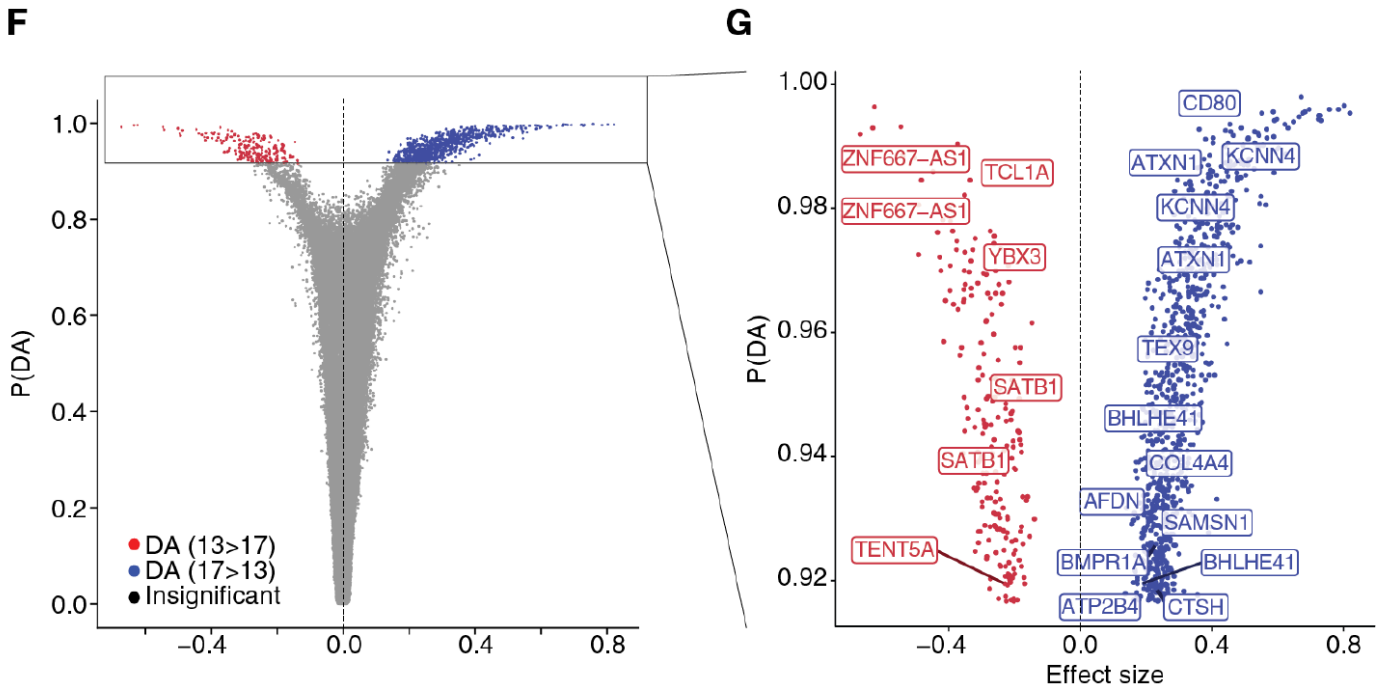
Generate many cluster labels
Select a random generation A
  Repeat (for 1 to C):
    Locate nearest neighbor cluster in A to all generations
    Label the nearest neighbor
    Remove the nearest neighbor from the pool
Repeat

Charkhabi *et al.*, IJMLC, 2015

1. Compare the DAs between cluster and all other clusters.
2. Use **enrichr** to associate the DA region to the genes.
3. Use **ARCHS4** collections to associate with cell type specific gene signatures.

Transfer Learning (Query to Reference Integration) (5)

Refine the clusters, for instance, naïve B-Cell and memory B-Cell are both annotated as B-Cell in the previous step:



1. Extract the clusters corresponds to the same cell type (e.g. cluster 13 and 17)
2. Perform DA region analysis.
3. Annotate the DA region with genes, annotate the clusters with our prior knowledge of those genes.

Discussion and Conclusion

- PeakVI proposes a probabilistic framework to analyze scATAC-Seq data.
  - Prior preprocessing is not necessary
  - Consider batch effect, number of reads (library size), technical confounder (width of genomic region)
  - Provide better statistical method for differentially accessibility analysis especially for cell with low sample size.
  - Provide transfer learning to annotate cells in an online fashion (retraining the entire model is not required).
- The manuscript is in general very well written.
  - The effect size comparison is a bit unfair, but GLM does suffer from sample size.
- We can try using hierarchical VAE to model scATAC-Seq and scRNA-Seq jointly with dependent relation in latent representation.
  - For joint analysis without the dependency, one can check MultiVI (Ashuach et al., 2021)