



# C-ZIPTF: stable tensor factorization for zero-inflated multi-dimensional genomics data

Neriman Tokcan



**University of Trento**

Masterclass Tensor Decompositions and Applications in Multi-Omics Data Analysis, November 2024



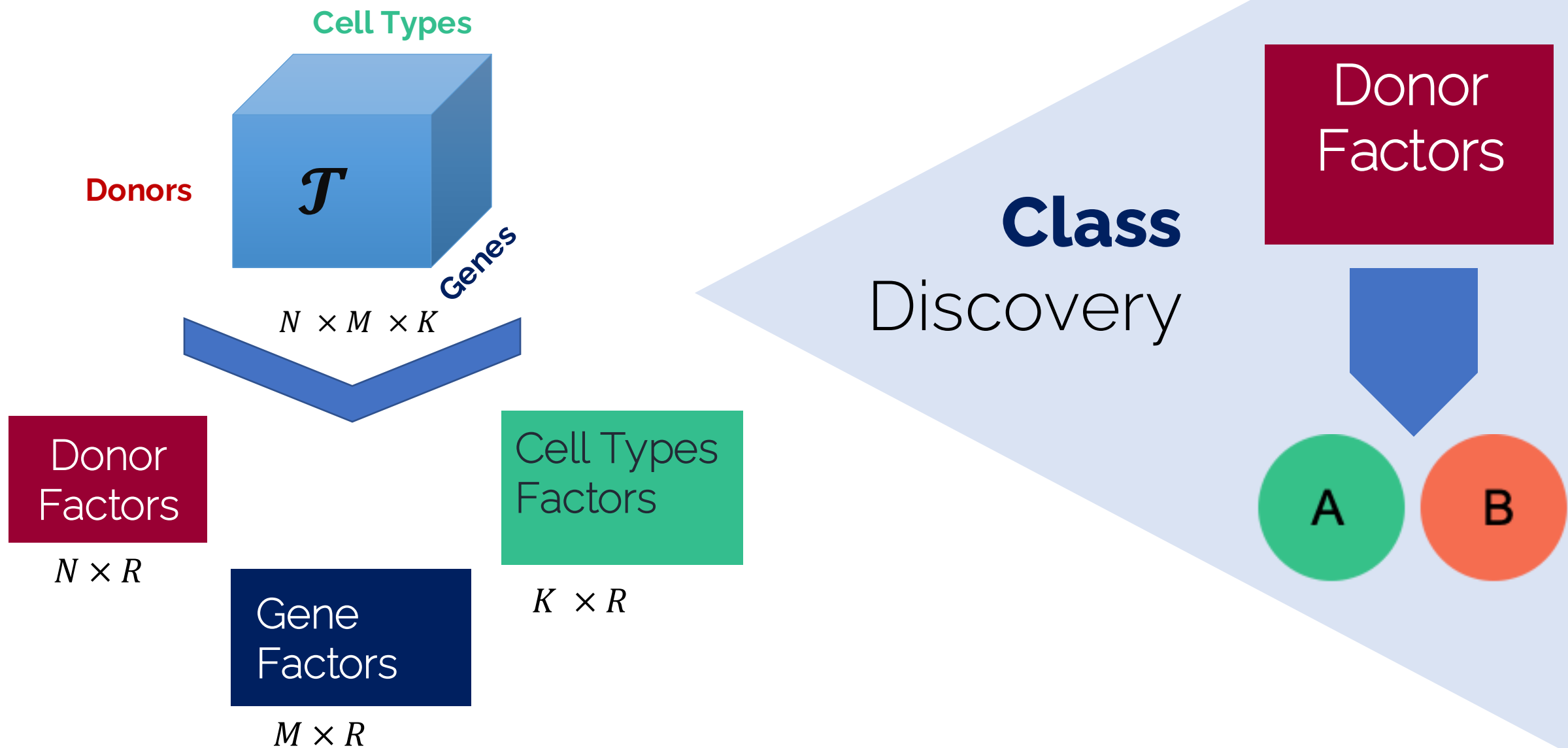


1

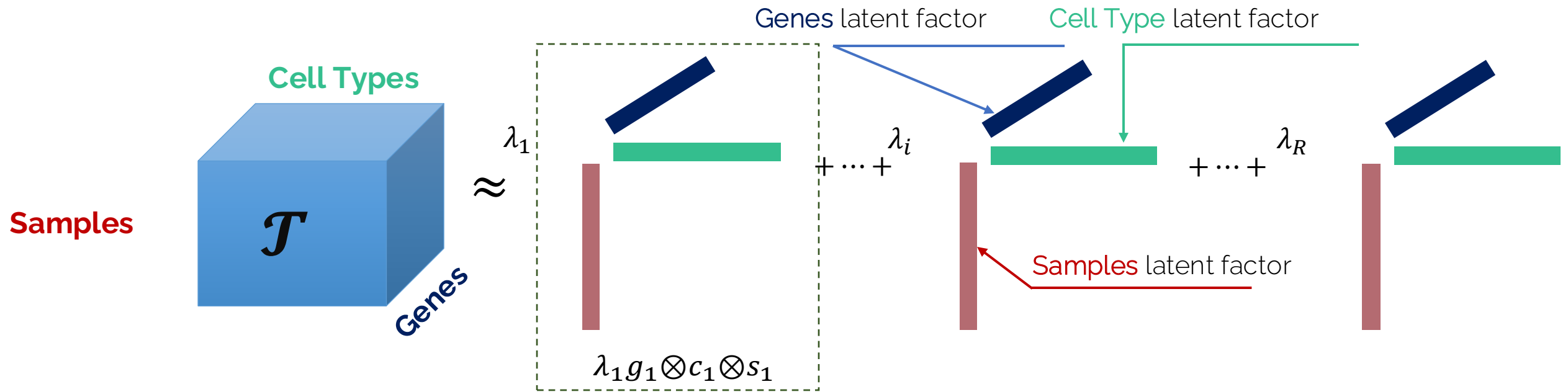
# Tensor Decomposition & Challenges



# CP Candecomp/Parafac Decomposition



# CP Decomposition



$\mathcal{T} = \mathcal{T}' + \varepsilon$  where

$$\mathcal{T}' = \sum_{r=1}^R \lambda_r g_r \otimes c_r \otimes s_r$$



noise assumption?  
rank selection?  
constraints?  
uniqueness?

$\mathcal{T}' = [G, C, S]$  where  $S = [s_1 \ s_2 \ \dots \ s_R]$ ,  $C = [c_1 \ c_2 \ \dots \ c_R]$ ,  $G = [g_1 \ g_2 \ \dots \ g_R]$

# CP-decomposition – traditional approaches

A common method for CP decomposition and other tensor-related optimization problems is **alternating least squares**. We want to solve the following problem:

$$\min_{\mathcal{J}'} \|\mathcal{J} - \mathcal{J}'\| = \sqrt{\sum_{i,j,k} \varepsilon_{ijk}^2} \quad \text{where } \mathcal{J}' = [G, C, S].$$

**Fit (explained variance)=**

$$1 - \frac{\|\mathcal{J} - \mathcal{J}'\|}{\|\mathcal{J}\|}$$

It is not a convex problem, but it can be given as 3 convex problems.

$$\begin{aligned} \min_G & \|\mathcal{J}^{(1)} - G(C \odot S)^T\| \\ \min_C & \|\mathcal{J}^{(2)} - C(S \odot G)^T\| \\ \min_S & \|\mathcal{J}^{(3)} - S(C \odot G)^T\| \end{aligned}$$

where  $\mathcal{J}^{(i)}$  is the mode-1 matricization of the tensor  $\mathcal{J}$ ,  $\odot$  denotes the "Khatri-Rao" product – matching column-wise Kronecker product  $C \odot G = [c_1 \otimes g_1 \quad c_2 \otimes g_2 \quad \dots \quad c_R \otimes g_R]$

Other loss functions?

Kullback-Leibner divergence

$$D_{KL}(\mathcal{J} \parallel \mathcal{J}') = \sum_{z \in Z} \mathcal{J}(z) \log \frac{\mathcal{J}(z)}{\mathcal{J}'(z)}$$



**Choice of the loss function or distance metric?**

**Stability of the factorization?**

# MLE & Bayesian approach

## Maximum Likelihood Approach

$$\operatorname{argmin}_{S,C,G} d(\mathcal{T}, \mathcal{T}')$$

s.t. constraints on the latent factor matrices  $S, C, G$

### loss functions:

$$d(\mathcal{T}, \mathcal{T}') = \|\mathcal{T} - \mathcal{T}'\|$$

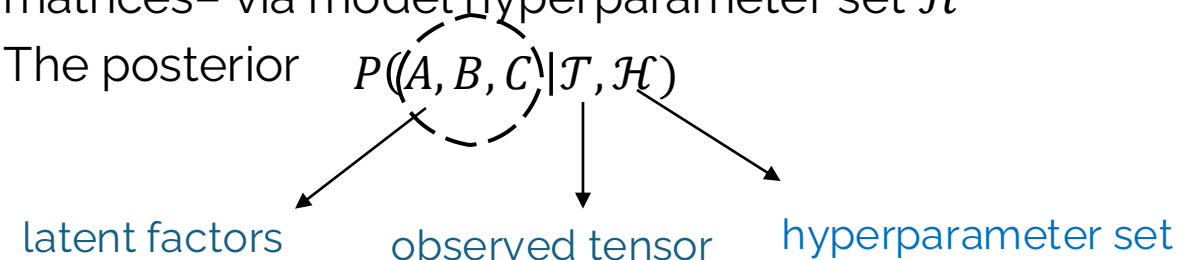
Kullback-Leibner divergence

$$D_{KL}(\mathcal{T} \parallel \mathcal{T}') = \sum_{z \in Z} \mathcal{T}(z) \log \frac{\mathcal{T}(z)}{\mathcal{T}'(z)}$$

## Bayesian Approach

Given the observed tensor  $\mathcal{T} \approx [A, B, C]$ . The goal is to estimate the posterior distribution of the factor matrices (A, B, C) given the observed tensor  $\mathcal{T}$  and any prior information you might have.

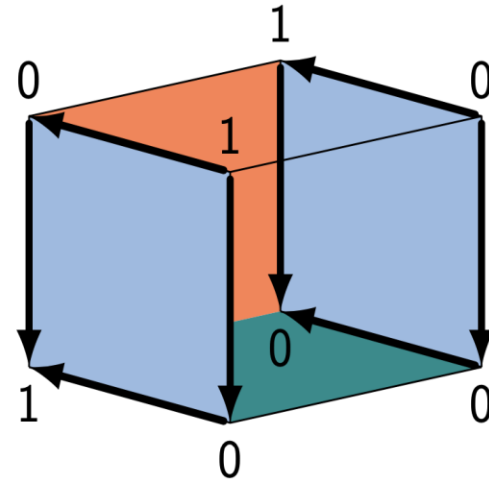
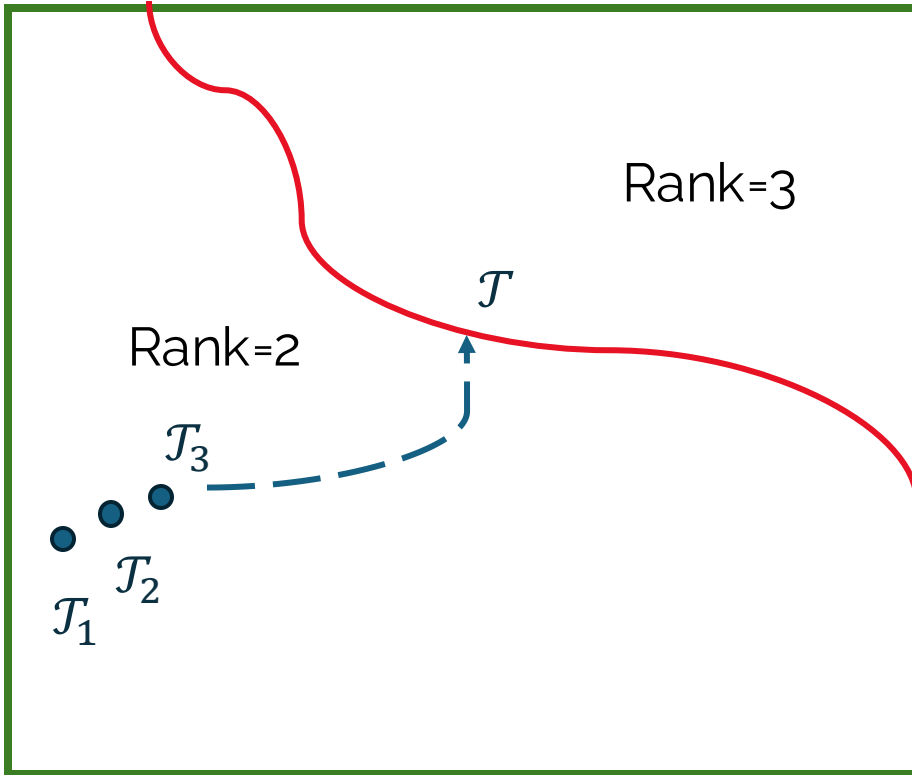
- Prior distributions are specified for the factor matrices – via model hyperparameter set  $\mathcal{H}$
- The posterior  $P((A, B, C) \mid \mathcal{T}, \mathcal{H})$



The posterior distribution is analytically intractable and must be approximated

Techniques like  
Markov Chain Monte Carlo (MCMC)  
Variational Inference (VI)

# Border rank



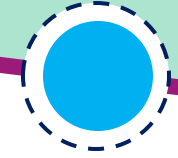
$$\mathcal{T} = \begin{bmatrix} 0 & 1 & | & 1 & 0 \\ 1 & 0 & | & 0 & 0 \end{bmatrix}$$

$$\mathcal{T} = e_2 \otimes e_1 \otimes e_1 + e_1 \otimes e_2 \otimes e_1 + e_1 \otimes e_1 \otimes e_2$$

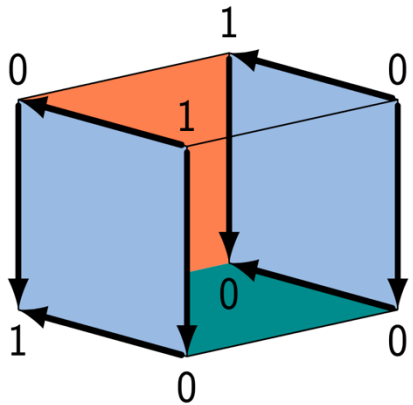
$\mathcal{T}$  has rank 3, but it can be approximated by rank 2 tensors

$$\lim_{n \rightarrow \infty} n \left( e_1 + \frac{1}{n} e_2 \right) \otimes \left( e_1 + \frac{1}{n} e_2 \right) \otimes \left( e_1 + \frac{1}{n} e_2 \right) - n e_1 \otimes e_1 \otimes e_1 = \mathcal{T}$$

# Numerical instability



## Border Rank



## Convergence Problems

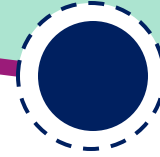
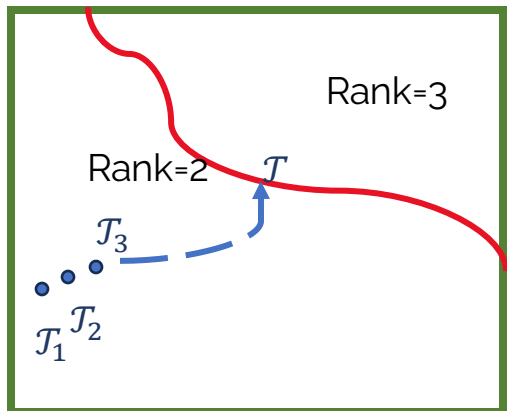
$$\min_{S,C,G} \left\| \mathcal{T} - \sum_{i=1}^r s_i \otimes c_i \otimes g_i \right\| \text{ where}$$

$$S=[s_1 \dots s_r], C = [c_1 \dots c_r], G = [g_1 \dots g_r].$$

$$\min_S \left\| \mathcal{T}^{(1)} - S(C \odot G)^T \right\|$$

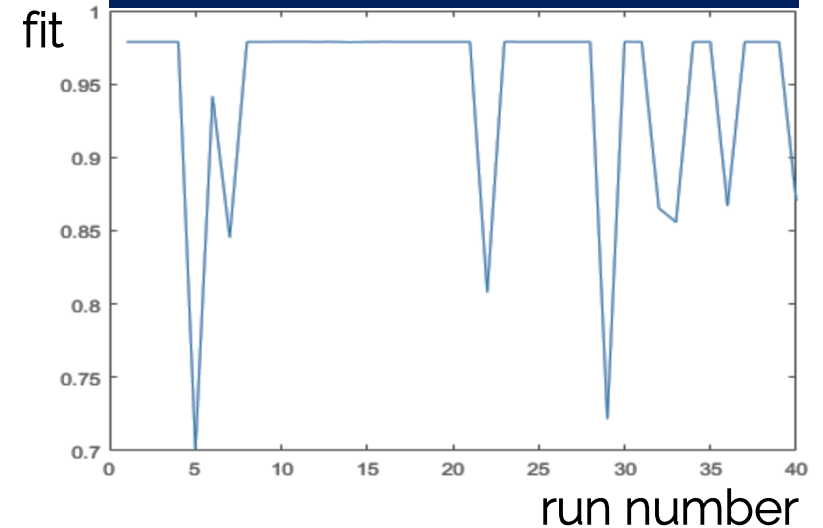
$$\min_C \left\| \mathcal{T}^{(2)} - C(G \odot S)^T \right\|$$

$$\min_G \left\| \mathcal{T}^{(3)} - G(C \odot S)^T \right\|$$



## Dependence on the Initial Guess

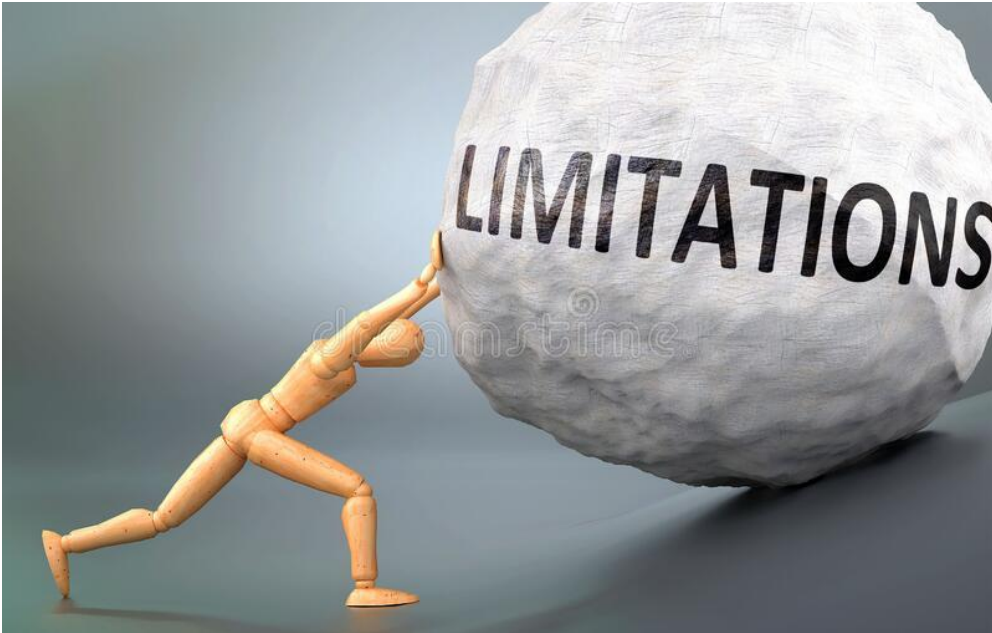
rank 3 approximation for tensor of size 3 x 4 x 5



\*Note:  $\|\mathcal{T}\| = \sqrt{\sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \sum_{k=1}^{I_3} \mathcal{T}_{i,j,k}^2}$



# Tensor decomposition: its limitations



- Decomposition is not **stable**
- **Convergence** is not guaranteed
- Rank selection is a **challenge**
- Uses assumptions that **do not hold** on real data sets
- **Needs a pipeline** for interpretation of latent factors

**Without custom pipeline:** *more capable than traditional methods*

**With custom pipeline:** *outperforms existing tensor methods*



Numerical instability and convergence problems



Rank selection is challenging



Interpretation of the factors can be difficult

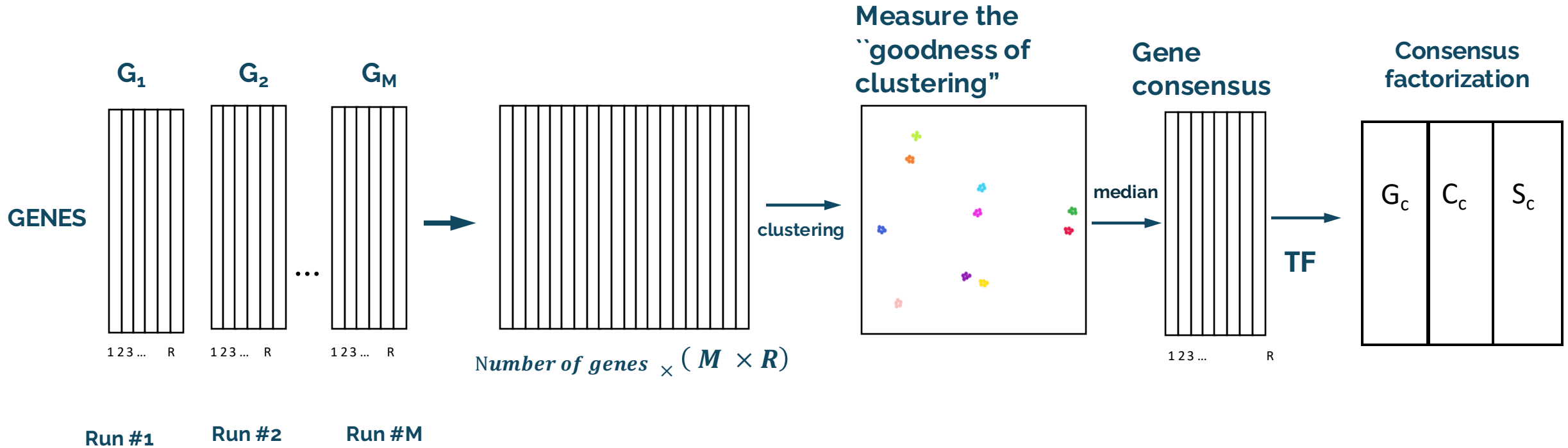


Incorporating true distribution of the data



# Consensus based tensor factorization

rank selection  
stable factorization

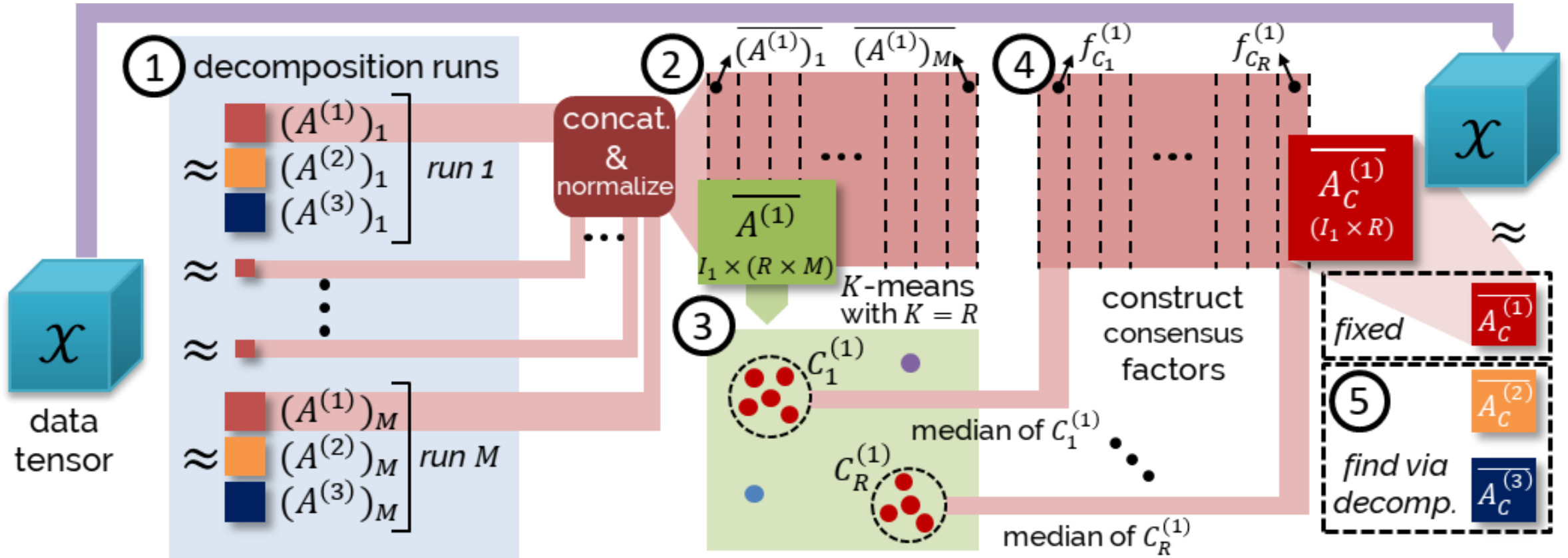


**connectivity matrix  $Connectivity_k$ ,  $1 \leq k \leq M$**   
 $Connectivity_k(i, j) = 1$  gene  $i$  and gene  $j$  belong to same cluster and 0 otherwise.

**Consensus matrix  $Consensus(i, j) =$**   
 probability of gene  $i$  and gene  $j$  cluster together  
 average of connectivity matrices

Evaluate dispersion between 0 and 1 and calculate **cophenetic correlation**

# Consensus based tensor factorization



# Bayesian Tensor Factorization

- The counts  $T$  are modeled as draws from a Poisson (or Zero Inflated Poisson) distribution. The mean for  $T_{ijk}$  is given by  $T'_{ijk}$  where

$$\mathcal{T} \approx \mathcal{T}' = [\mathbf{G}, \mathbf{C}, \mathbf{S}] = \sum_{r=1}^R \mathbf{g}_r \otimes \mathbf{c}_r \otimes \mathbf{s}_r$$

$$T_{ijk} \approx \text{Poisson}(\lambda = \sum_{r=1}^R g_{ri} c_{rj} s_{rk})$$

- We set a **Gamma Prior** on each entry of the factor matrices  $S$ ,  $C$  and  $G$ , and a gaussian prior on the **gate parameter** in the ZIP model that controls the zero inflation.

## Generative model:

$$S \sim \text{Gamma}(\alpha_s, \beta_s)$$

$$C \sim \text{Gamma}(\alpha_c, \beta_c)$$

$$G \sim \text{Gamma}(\alpha_g, \beta_g)$$

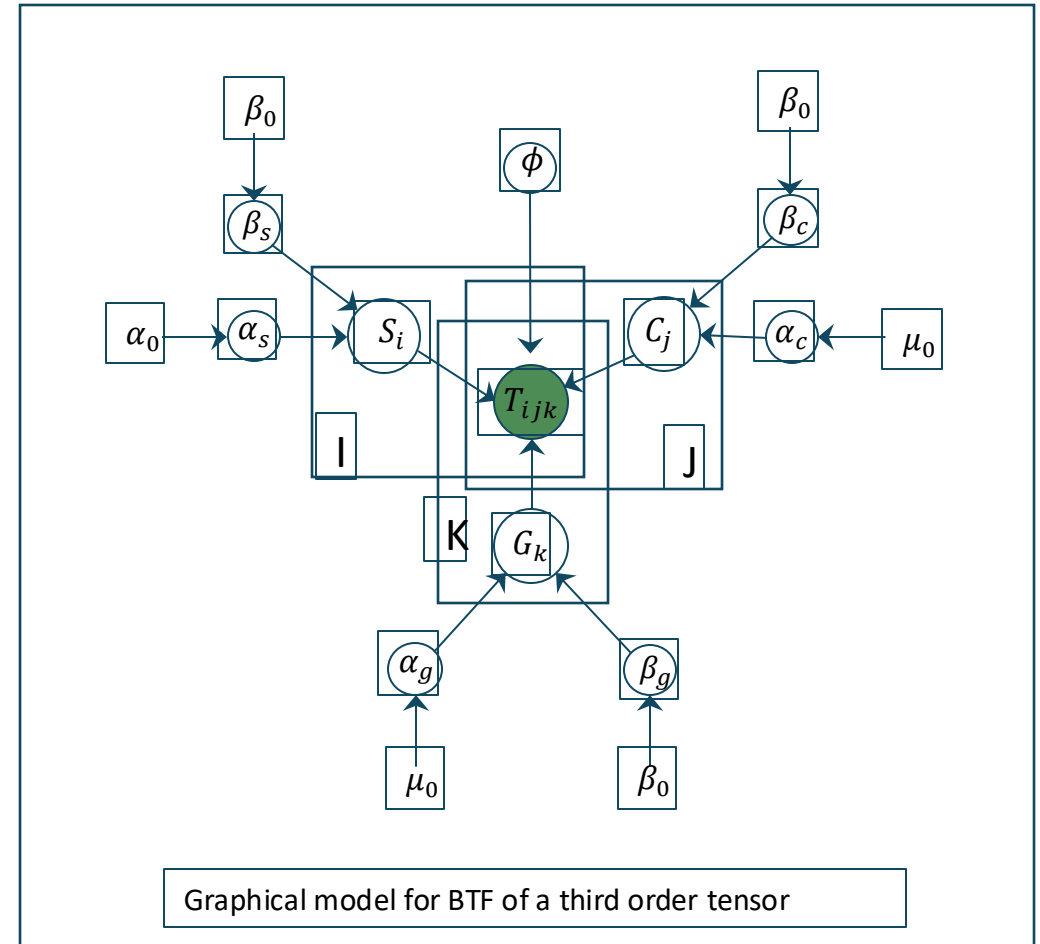
$$p \sim N(\mu, \sigma)$$

excess zeros

$$T_{ijk} \approx \text{ZIP}(\lambda = \sum_{r=1}^R g_{ri} c_{rj} s_{rk}, \phi = \text{sigmoid}(p))$$

- Potential benefits of Bayesian inference compared to the more prevalent maximum likelihood estimation approach include

- uncertainty quantification,
- incorporation of more realistic noise assumptions, and
- a principled way to include prior information



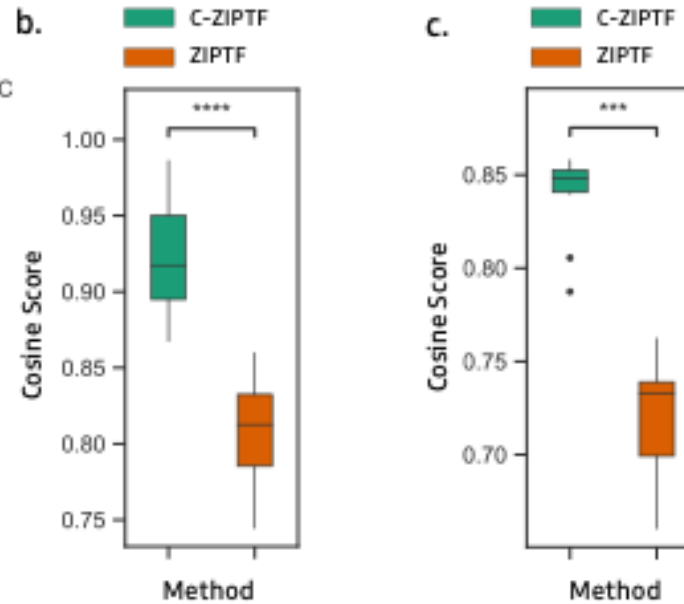
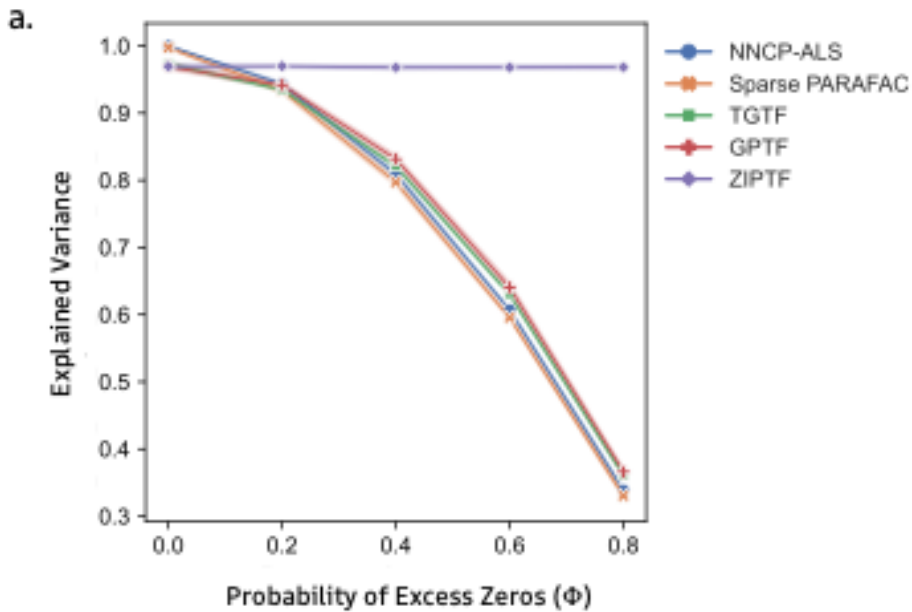
Note: To maximize the evidence lower bound (ELBO), we employ a stochastic optimization algorithm known as **the Black Box Variational Inference**. This algorithm operates by stochastically optimizing the variational objective using Monte Carlo samples from the variational distribution to compute the noisy gradients. It effectively alleviates the burden of analytic computations and provides a more efficient approach to ELBO maximization.



# Synthetic tensor experiments for Zero-inflated Poisson Factorization

$$\chi = [A, B, C] = \sum_{r=1}^R a_r \otimes b_r \otimes c_r, \quad A \in \mathbb{R}^{I \times R}, B \in \mathbb{R}^{J \times R}, C \in \mathbb{R}^{K \times R}.$$

with elements drawn from a Gamma distribution  $\alpha = 3, \beta = 0.3$ , we generate  $\chi'$  by sampling from a ZIP distribution with mean  $\chi$  and varying probability extra zeros.  $(I, J, K, R) = (10, 20, 300, 9)$



b. Cosine similarity between factors obtained on repeat runs for ZIPTF and C-ZIPTF,  
 c. Cosine similarity between factors from 100 runs to original signal

$\Phi = 0.6$

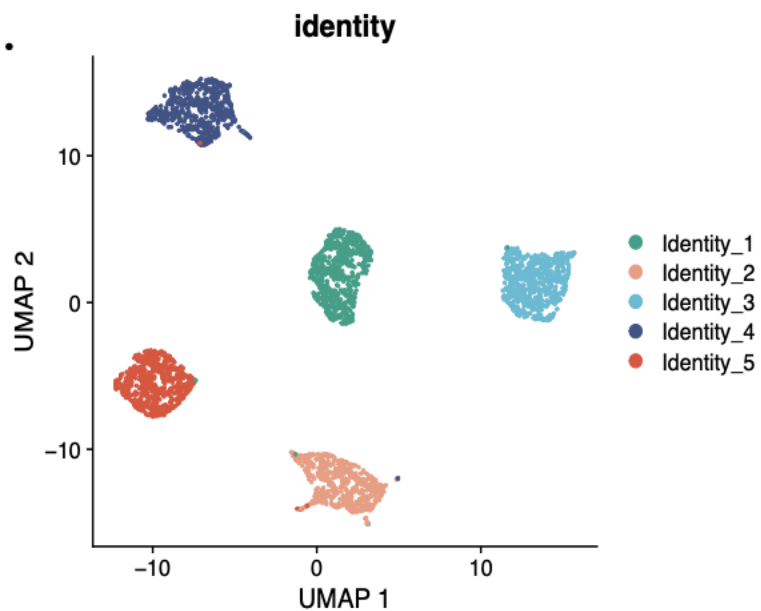
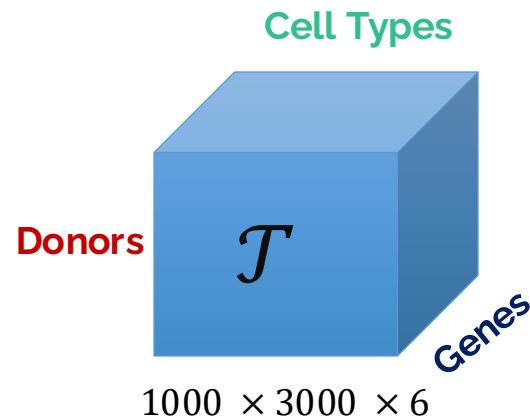
100 runs of ZIPTF and Consensus ZIPTF

ZIPTF: Zero-inflated Poisson factorization  
 NNCP-ALS: Non-negative CP decomposition via alternating least squares  
 GPTF: Gamma-Poisson tensor Factorization  
 TGTF: Truncated Gaussian tensor factorization

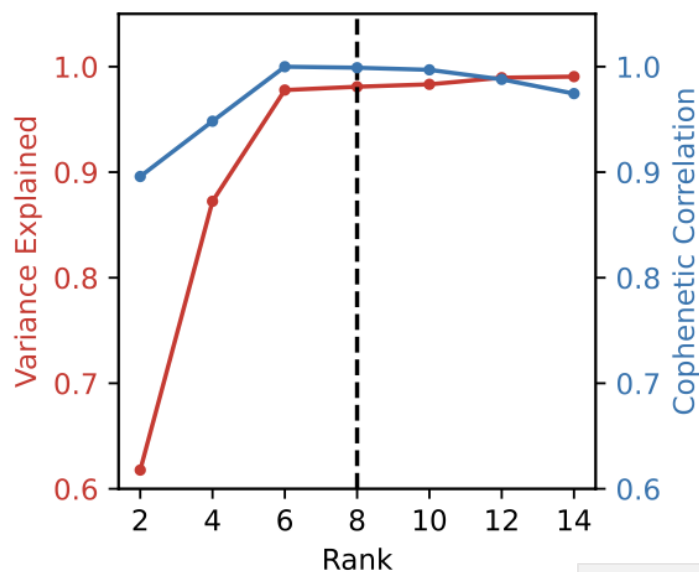
# Application 1: multi-donor multi-cell type expression data

Splatter simulation to generate the synthetic single-cell RNA sequencing dataset. The simulation framework utilizes a Gamma-Poisson hierarchical model with hyper-parameters estimated from real data.

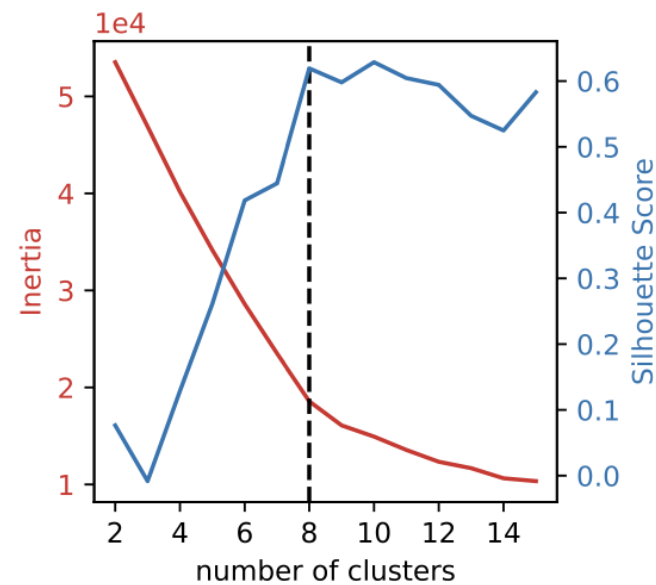
- 3,000 cells, 1,000 genes, six donors
- five gene expression programs defining cell type identities
- three gene expression programs defining donor-specific activity



simulated data

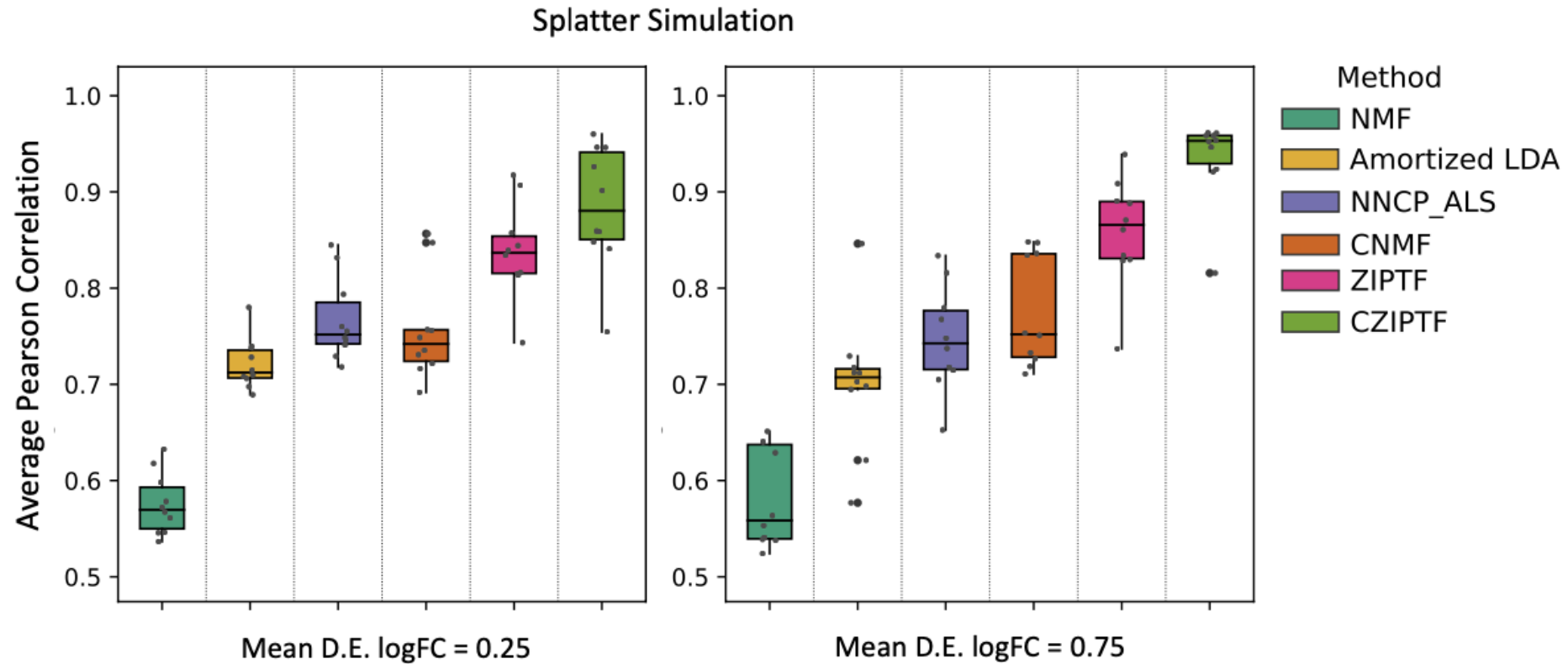


rank selection





# Recovery of gene expression programs



NMF: Non-negative matrix factorization  
LDA: Latent Dirichlet Allocation

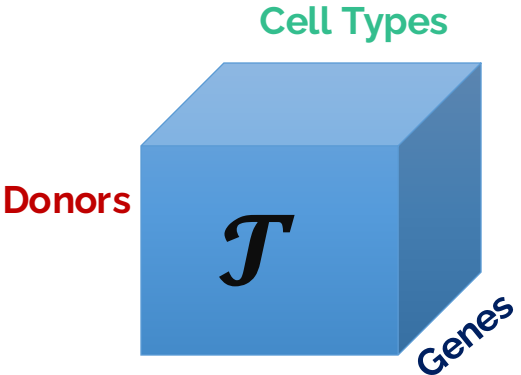
# Unsupervised discovery of disease subgroups and multicellular gene expression programs in the peripheral blood of patients with systemic lupus erythematosus (SLE)

dataset: C-ZIPTF to a multiplexed scRNA-seq (mux-seq) to profile over 1.2 million PBMCs from patients with systemic lupus erythematosus (SLE) and healthy controls

**Downsampled to 85,636 cells:** 8 SLE patients with flare, 8 SLE patients with managed disease, and 8 healthy controls.

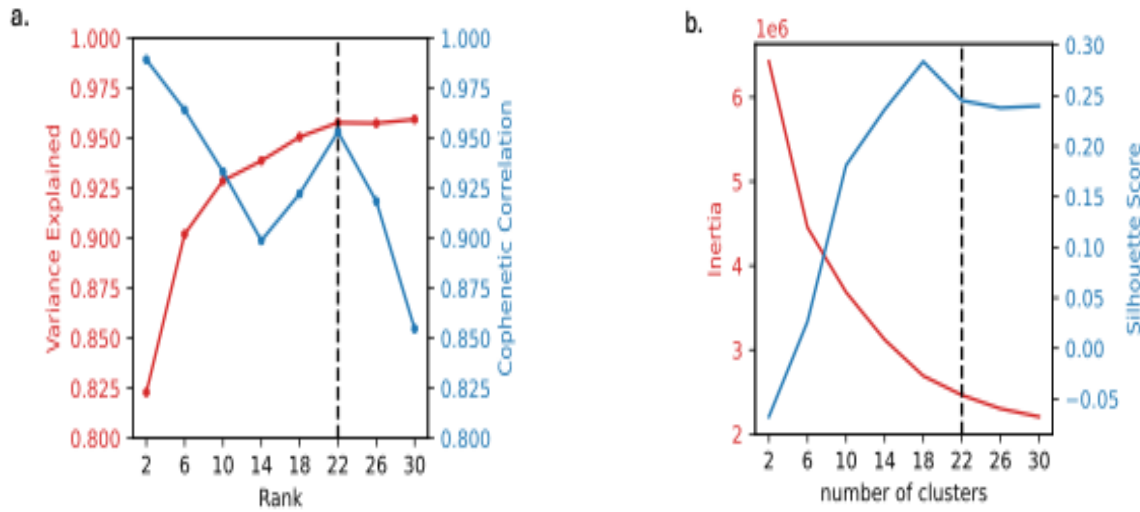
**cell types:** CD4-positive alpha-beta T cells, CD8-positive alpha-beta T cells, classical monocytes, conventional dendritic cells, and NK cells

## pseudobulk tensor



$donors \times cell\ types$   
 $\times genes$   
 $24 \times 5 \times 13,525$

## rank selection





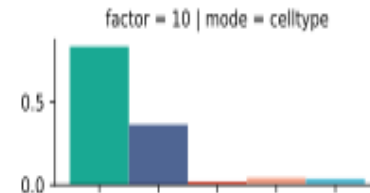
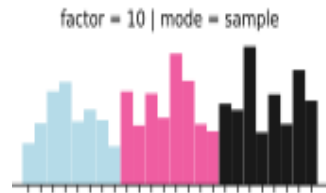
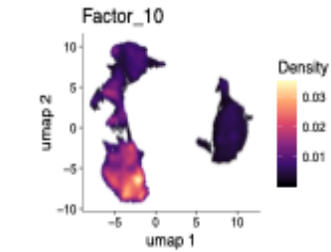
# Cell type identity Gene Expression Programs

sample factors

cell type factors

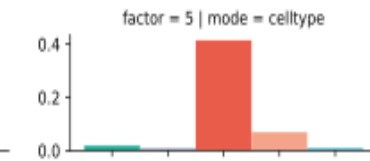
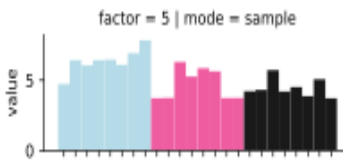
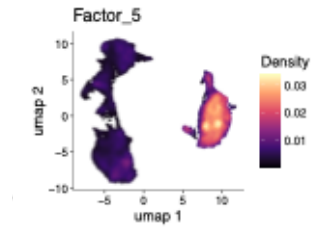
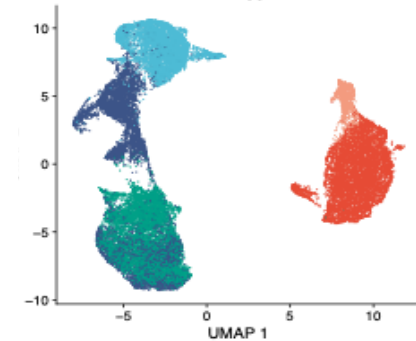
top 20 genes

Cell Type



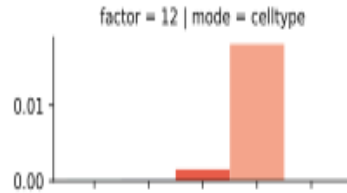
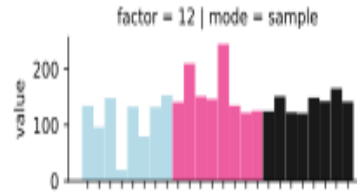
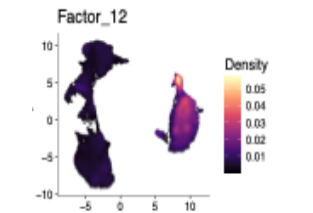
factor = 10 | mode = gene

CD3D, CD3G, CD27, CCR7, CD3E,  
NOSIP, LEPROTL1, OXNAD1, LEF1, ATF7IP2,  
TRABD2A, PASK, SELENOM, PIK3IP1, LCK,  
TOB1, SVIP, BCL11B, NELL2, SH3YL1



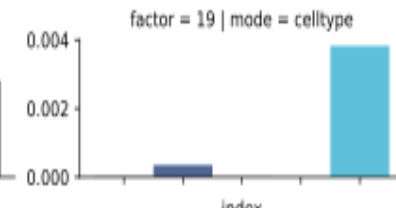
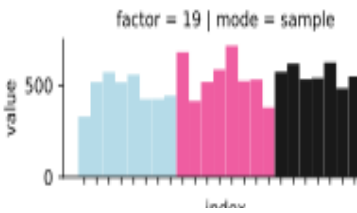
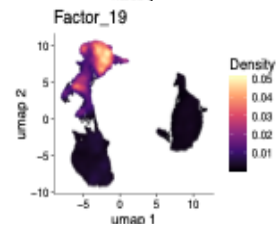
factor = 5 | mode = gene

S100A12, S100A8, RBP7, S100A9, CD14,  
CSTA, SERPINA1, CFD, APOBEC3A, TMEM176B,  
CLEC4E, ASGR1, CDA, FCN1, MNDA,  
VCAN, MS4A6A, NCF1, CYBB, CLEC12A



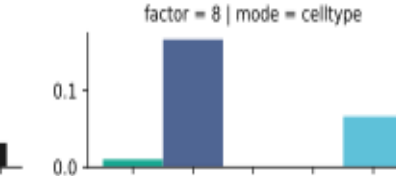
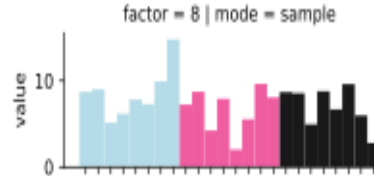
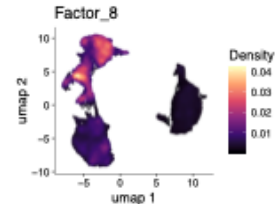
factor = 12 | mode = gene

CLEC10A, FCER1A, ENHO, HLA-DQB1, HLA-DQA1,  
HLA-DRB5, CD1C, HLA-DRA, HLA-DPB1, HLA-DPA1,  
PHACTR1, NDRG2, DNASE1L3, CD1E, CLIC2,  
CD74, CST3, HLA-DRB1, HLA-DMA, CPVL



factor = 19 | mode = gene

KLRF1, GNLY, CTSW, CLIC3, XCL2,  
GZMB, PRF1, CMC1, FGFBP2, SPON2,  
KLRD1, HOPX, KLRC1, PTGDS, S1PR5,  
AKR1C3, RAMP1, KLRB1, FCGR3A, IL2RB



factor = 8 | mode = gene

GZMK, CD8A, GZMH, CD8B, ZNF683,  
NKG7, GZMA, CST7, LYAR, LAG3,  
IL32, DUSP2, GZMM, KLRG1, PATL2,  
LAIR2, FCRL6, S100B, C1orf21, IFNG

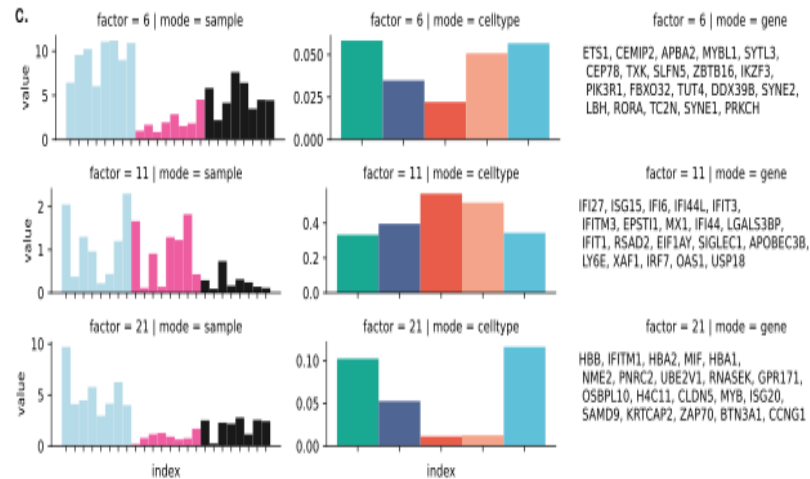
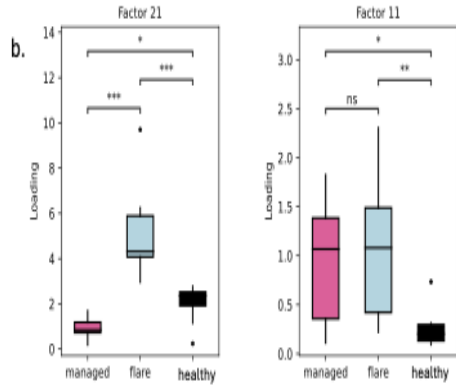
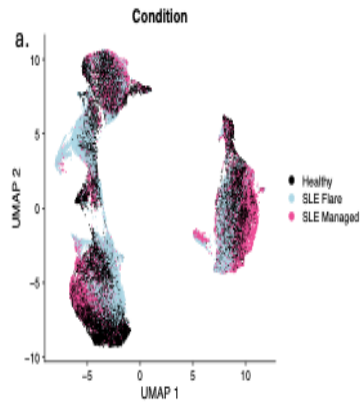
Disease

- Healthy
- SLE Flare
- SLE Managed

Celltype

- CD4-positive, alpha-beta T cell
- CD8-positive, alpha-beta T cell
- classical monocyte
- conventional dendritic cell
- natural killer cell

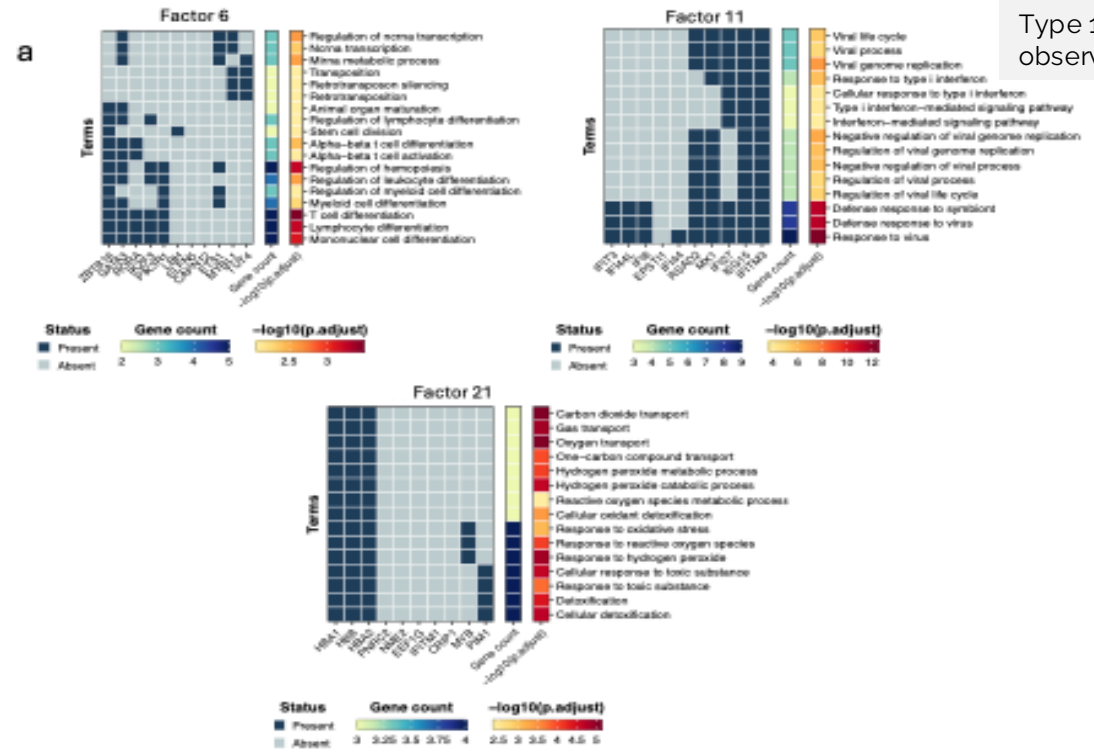
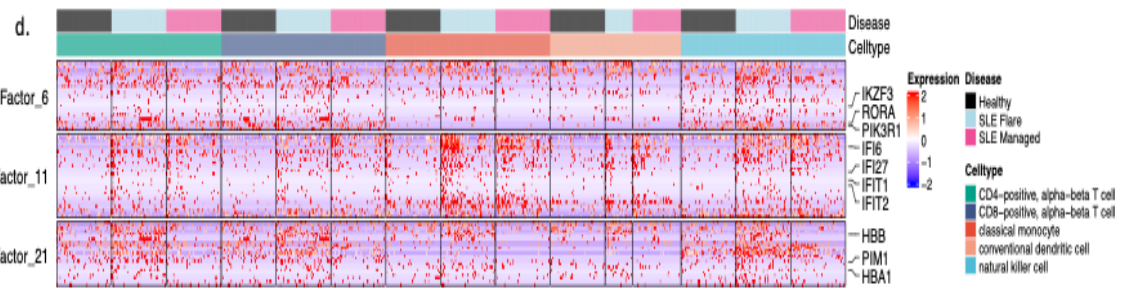
# Condition Specific Gene Expression Programs



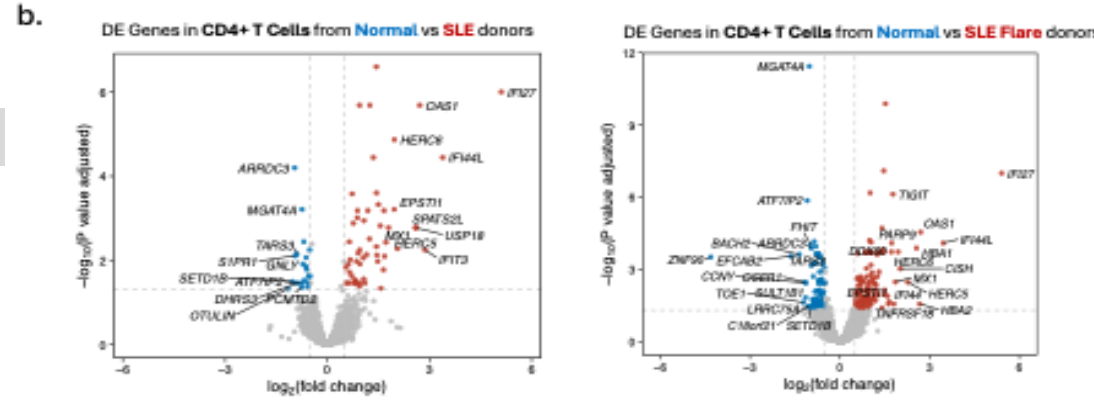
active flare

SLE patients

active flare



Type 1 –interferon pathway observed in viral infections



The number of differentially expressed genes (DEGs) was 55 when comparing all SLE patients against healthy donors and 122 when comparing SLE patients with an active flare against healthy donors

# Takeaway:

Scenarios where the source of intra-group **heterogeneity is unknown**, C-ZIPTF can highlight subgroups based on expression profiles and **identify the GEPs driving heterogeneity** that may be missed by supervised differential gene expression analysis.

Chafamo, Daniel, Vignesh Shanmugam, and Neriman Tokcan.  
"C-ziptf: stable tensor factorization for zero-inflated multi-dimensional genomics data."  
*BMC bioinformatics* 25.1 (2024): 323.



Vignesh Shanmugam


Broad Institute



Daniel Chafamo

Broad Institute  
Upenn Medical School



A microscopic image of a tumor microenvironment, showing a dense network of cells and blood vessels. A white circle with the number 2 is overlaid on the left side of the image.

2

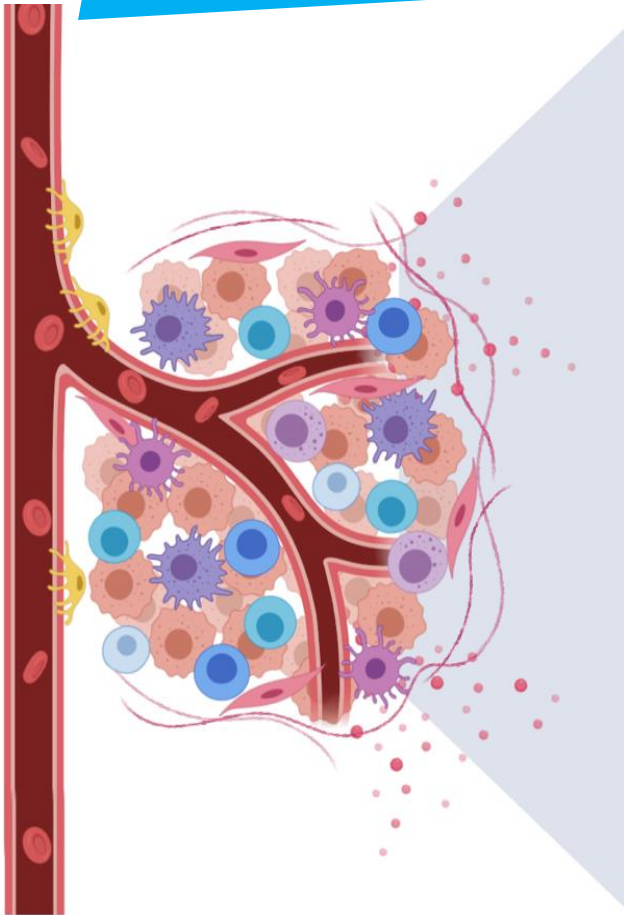
# Tumor

Microenvironments



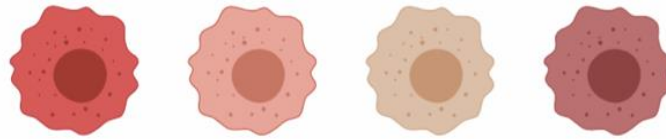
# Tumor microenvironment (TME)

Tumors are complex cellular **ecosystems**



Complex cell-cell **interactions**

Diverse **malignant** cell states



*Genetic and non-genetic heterogeneity*

Diverse **non-malignant** cell types & states



**T cells**



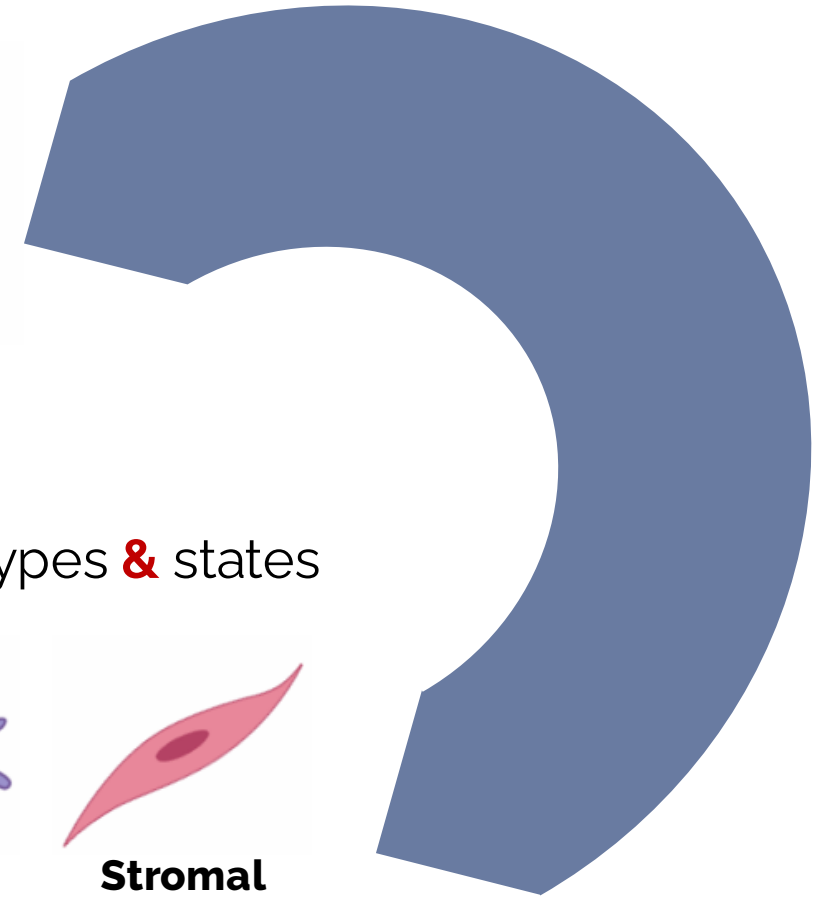
**NK cells**



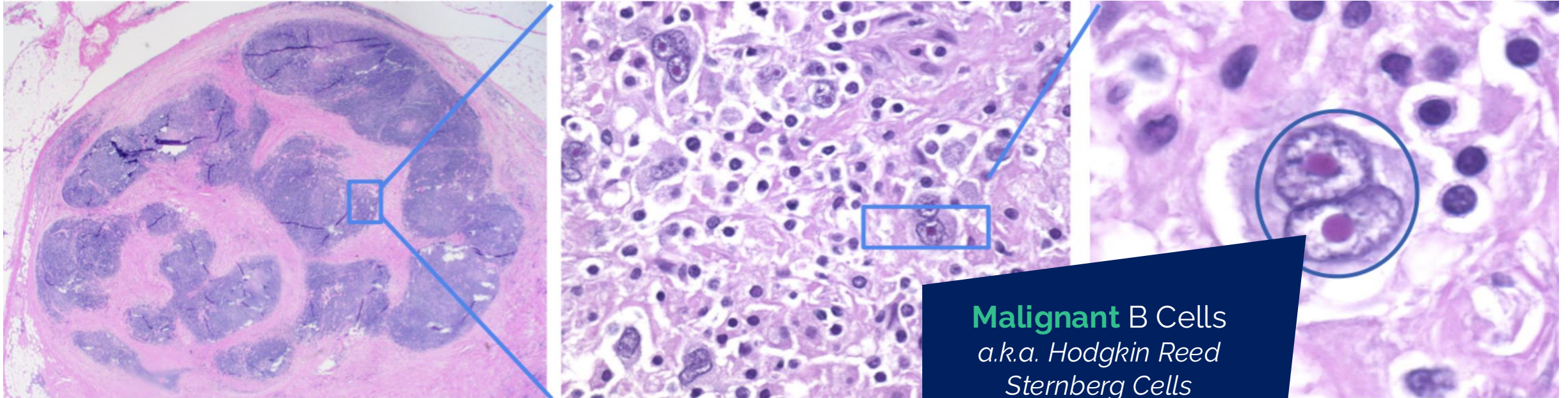
**Macs**



**Stromal**



# Classical Hodgkin Lymphoma (CHL)



**Malignant** B Cells  
*a.k.a. Hodgkin Reed  
Sternberg Cells*

**They comprise only ~1%  
of the tumor volume**

● Difficult to grow in culture

● Do not survive in immunodeficient mice

● Present in an extensive background of immune cells...  
*...yet continue to **grow** and **proliferate***

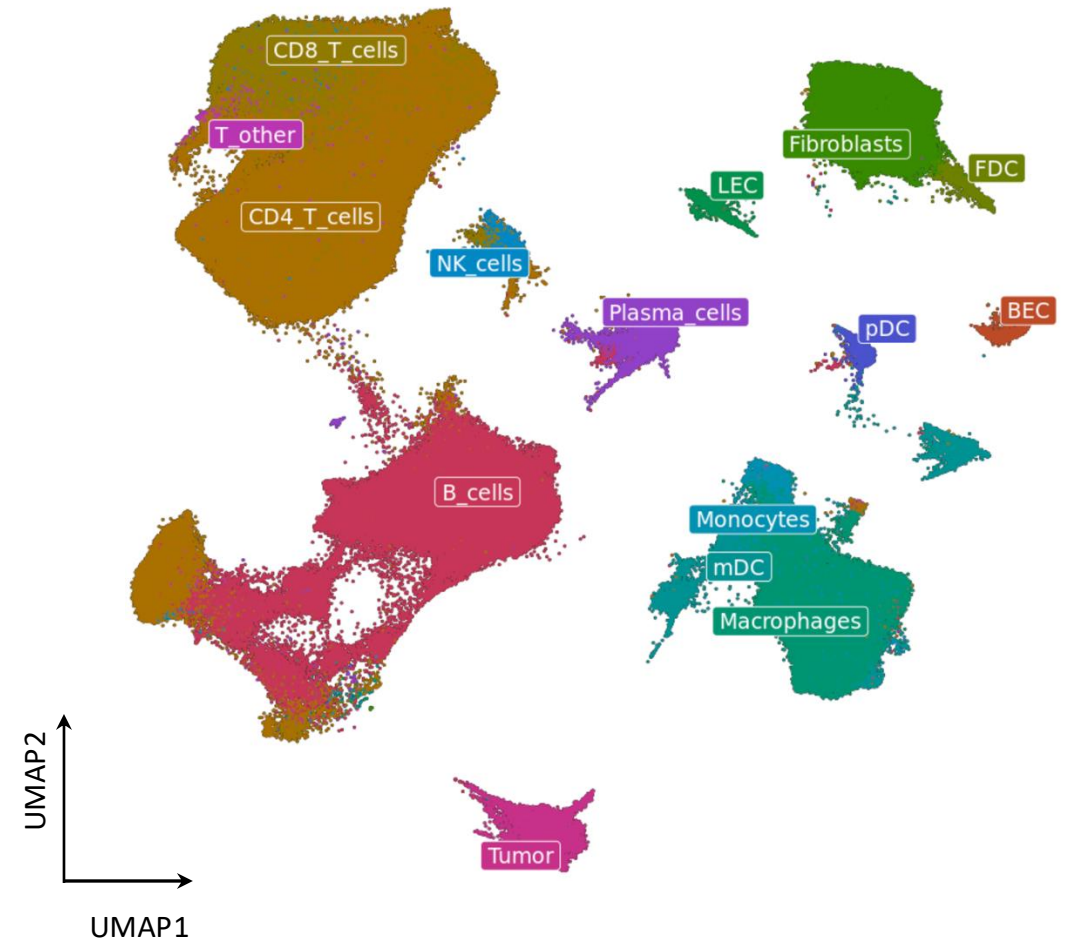
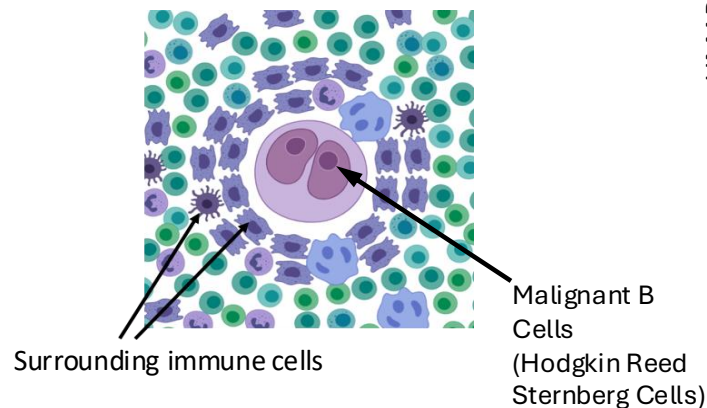
Strong  
**Dependence** on  
Microenvironment

## Application to multi-donor Hodgkin Lymphoma single nucleus dataset

- Malignant B cells of Hodgkin lymphoma are dependent on the native tissue microenvironment for survival and evade anti-tumor immunity
- We are interested in isolating the gene expression programs that characterize the altered cell states of immune and stromal cells in Hodgkin lymphoma patients.
- To address this question, we utilize a Single nucleus RNA-Seq (slide-tags protocol) dataset of 15 human samples from a clinically annotated patient cohort:

**10 Hodgkin lymphoma patients ( $\geq 2$  replicates)**  
**5 Epstein-Barr virus (EBV) positive**  
**5 Epstein-Barr virus (EBV) negative**  
**5 reactive lymph nodes ( $\geq 2$  replicates)**

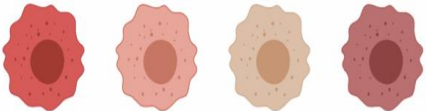
- We recover 320,000 nuclei after quality control (nUMI > 400, nGene > 200, %MT < 5, ambient RNA correction and doublet filtering)
- 15 cell types were annotated manually



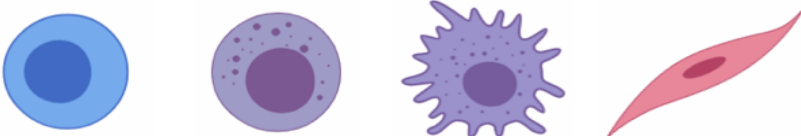


# Classic Hodgkin Lymphoma

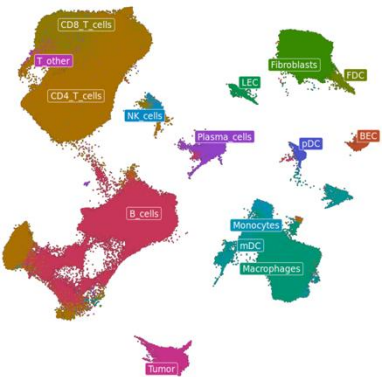
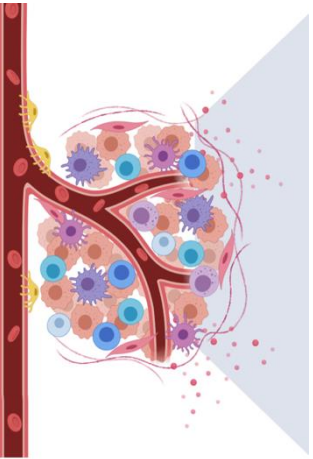
Diverse **malignant** cell states



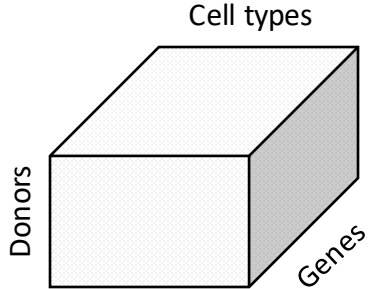
Diverse **non-malignant** cell types & states



- 15 cell types
- 19,875 genes
- 40 donors
- 2 conditions
  - Classic Hodgkin Lymphoma
  - Epstein-Barr Virus Positive
  - Epstein-Barr Virus Negative
  - Reactive lymph nodes

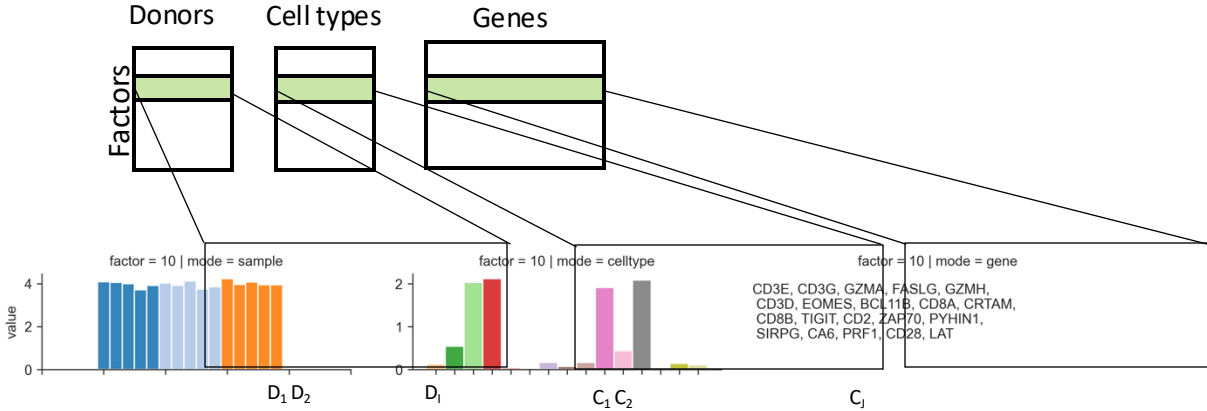


Single cell/nuclei



Pseudobulk tensor

19,875 x 15 x 40  
(Genes x Cell types x Donors)



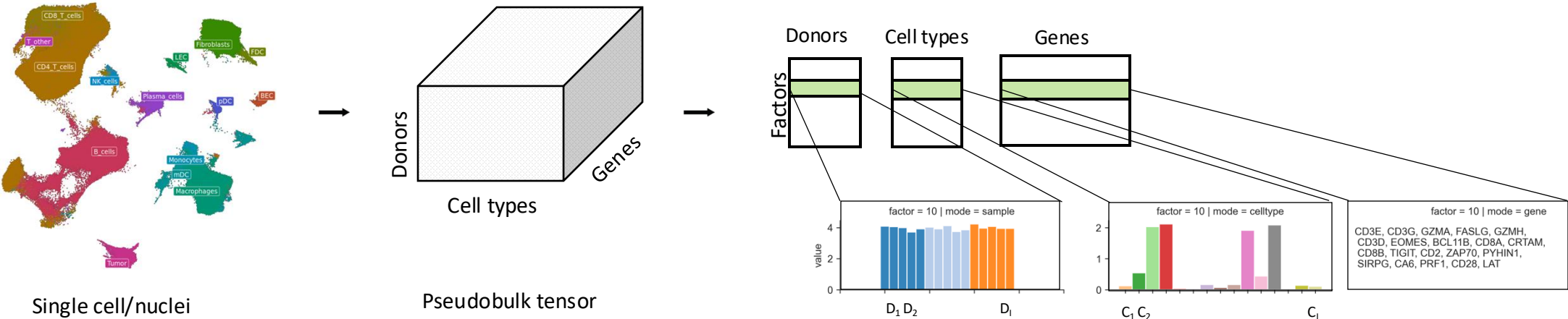
# Method Overview

## Benefits of the method:

- De novo discovery of gene expression programs that vary across cell types and donor conditions
- Unsupervised stratification of donors into subgroups and identification of GEPs that drive those stratifications

Our paper will be available on bioRxiv!

Title: ***"Genome-scale spatial mapping of the Hodgkin lymphoma microenvironment identifies factors required for tumor cell survival"***

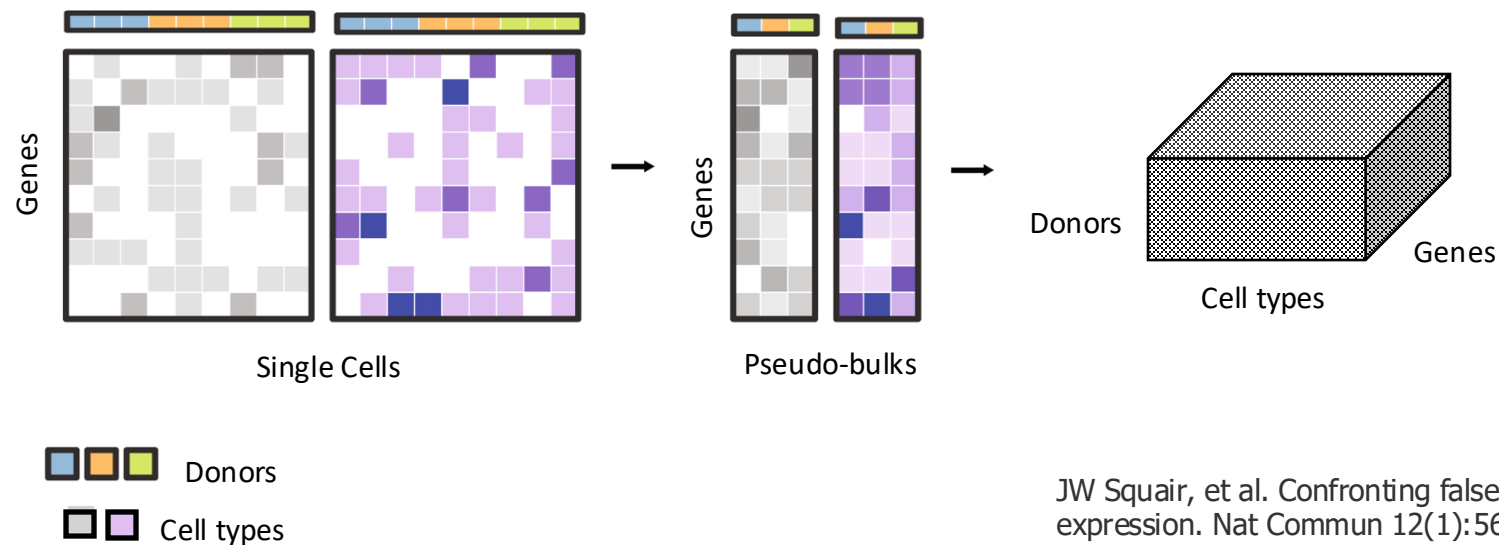


# Pseudobulk tensor formation

- Given a cell by gene matrix, wherein cells are annotated by cell type and donor, we create a pseudobulk tensor by aggregating the **raw** counts for each cell type, donor, and gene.
- The resulting pseudobulk data tensor has dimensions  $S \times C \times G$ , where  $S$  denotes the number of samples(donors),  $C$  the number of cell types and  $G$  the number of genes.
- We normalize the tensor such that each sample-cell type pair has a total of 10,000 counts.
- For the Poisson and Zero Inflated Poisson models we round the counts to the nearest integer to align with the support for the models.

## Gene filtering

- In order to facilitate biological interpretability of factors and reduce noise in the tensor formed we removed genes using the following to criteria:
  1. Filter out genes that we not provided with HGNC (HUGO Gene Nomenclature Committee) symbols
  2. Filter out genes with less than 10 total count across all cells



JW Squair, et al. Confronting false discoveries in single-cell differential expression. Nat Commun 12(1):5692.



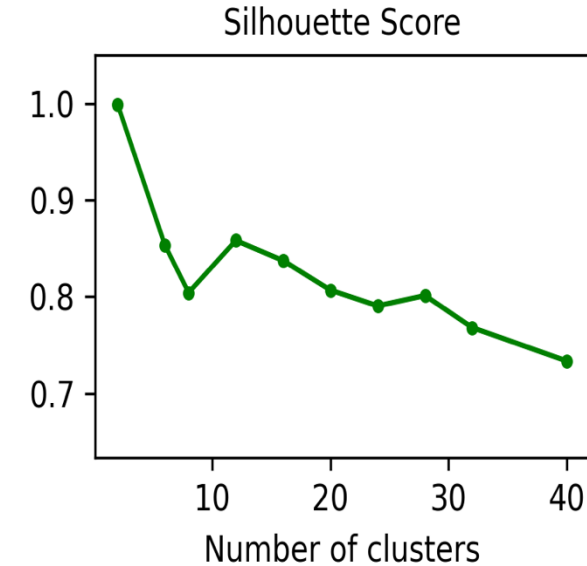
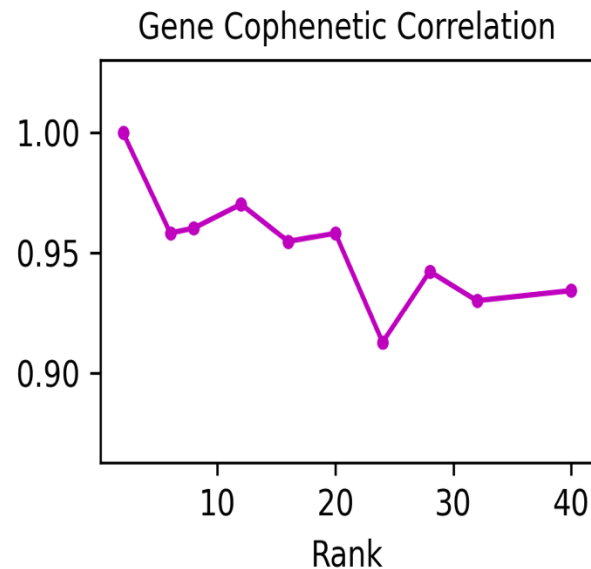
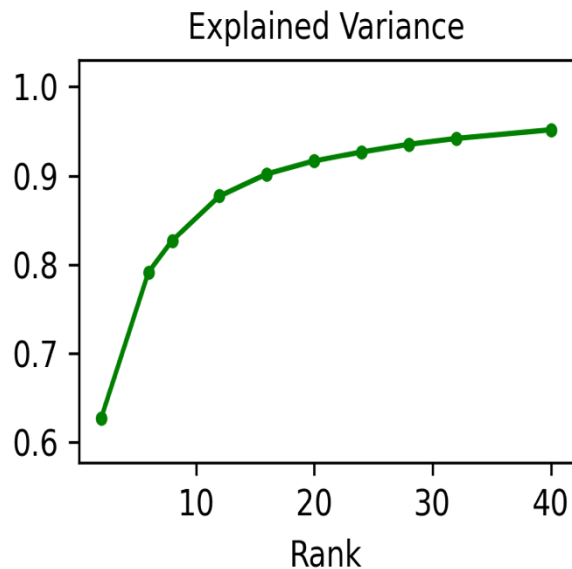
# Application to multi-donor Hodgkin Lymphoma single nucleus dataset

By pseudo-bulking we created a tensor with dimensions 40 x 15 x 19,875 (Donors x Cell types x Genes)

Explained variance of factorization went from a low of 0.627 at rank 2 to a high of 0.952 at rank 40

We run the algorithm 100 times to check the stability

At which rank we have stability?



# Assigning genes to factors

To select factor specific genes, we use two metrics:

## 1. Entropy

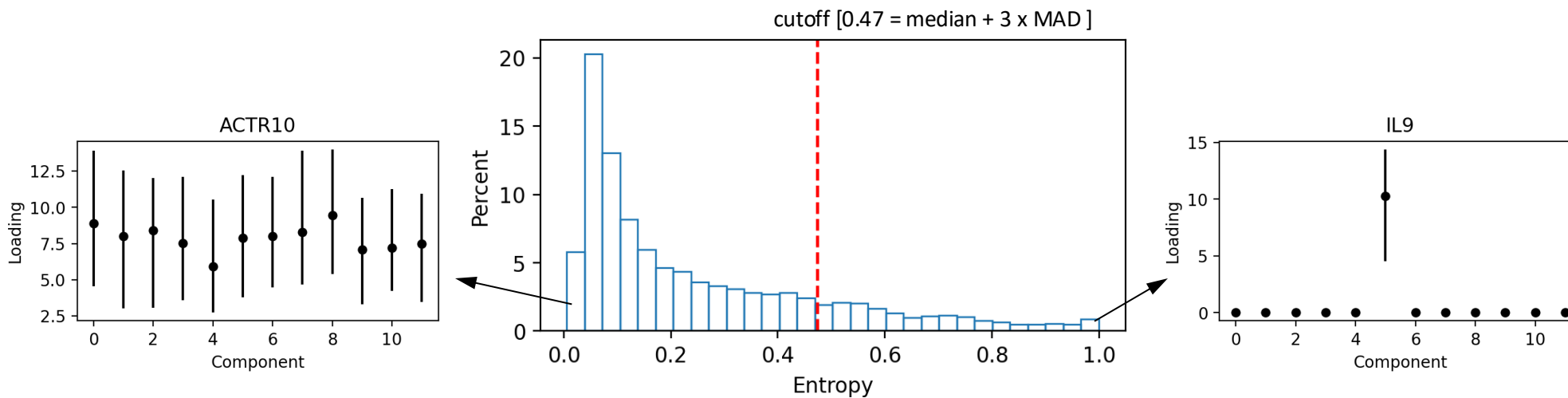
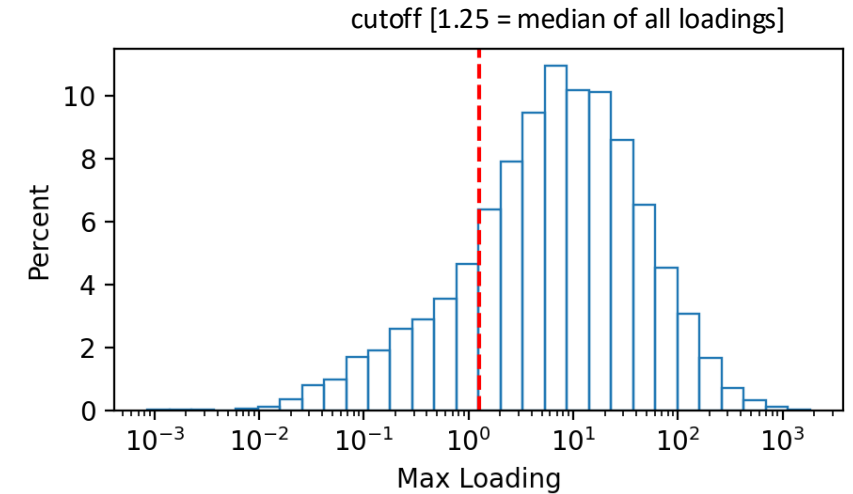
$$Entropy(g_i) = 1 + \frac{1}{\log_2 R} \sum_{j=1}^R P(g_i, f_j) \log_2 P(g_i, f_j)$$

where  $R$  = rank and  $P(g_i, f_j)$  is the probability that the gene  $i$  contributes to factor  $j$ .

- The higher the entropy the more factor-specific the corresponding gene.
- We set a threshold of median + 3\*Median Absolute Deviation to filter for high entropy genes

## 2. Max Loading

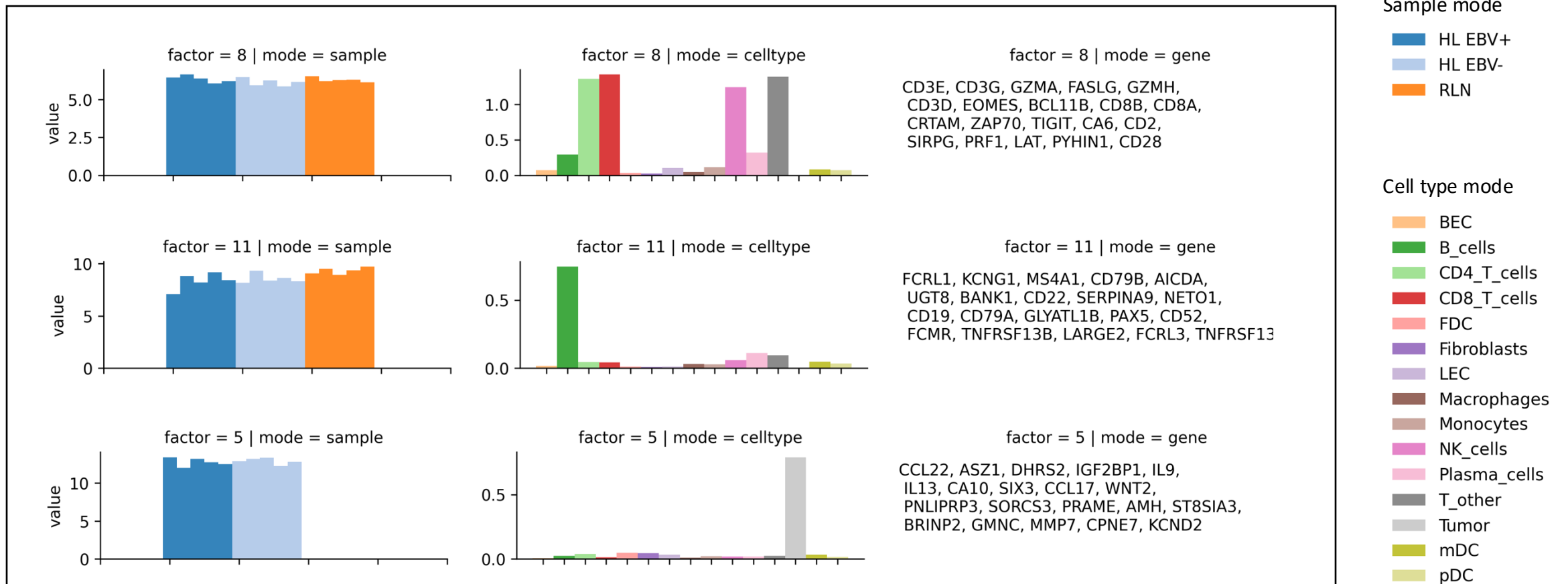
- To exclude genes that are overall too lowly expressed we filter out genes whose maximum loading across all factors is less than the median of all loadings.



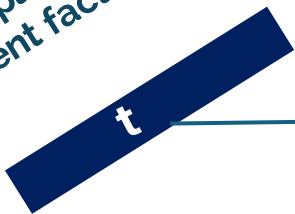
# Interpretation of recovered factors - Rank 12

40x 15 x 19,875  
Donors x Cell types x Genes

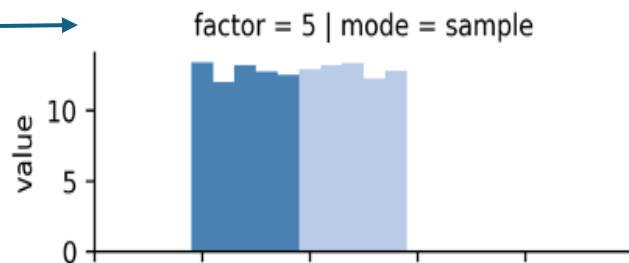
- At lower ranks, the factors mostly pick up cell type identity gene expression programs.
- The genes associated with the factors are canonical cell type marker genes.
- Factor 5 which loads only on tumor cells from Hodgkin Lymphoma samples, identifies the tumor identity gene expression program.



Samples latent factor



Samples latent factor



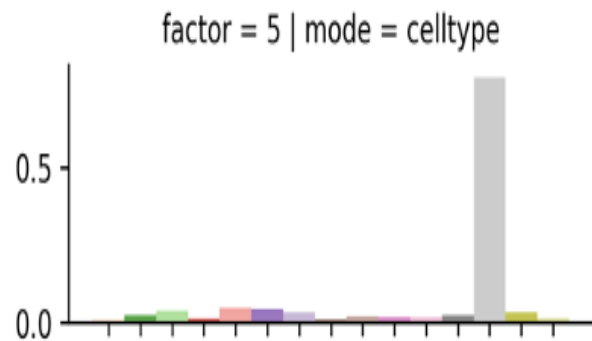
- HL EBV+
- HL EBV-
- RLN

HL: Hodgkin lymphoma  
 EBV+: Epstein-Barr virus positive  
 EBV-: Epstein-Barr virus negative  
 RLN: Reactive Lymph node

Cell types latent factor

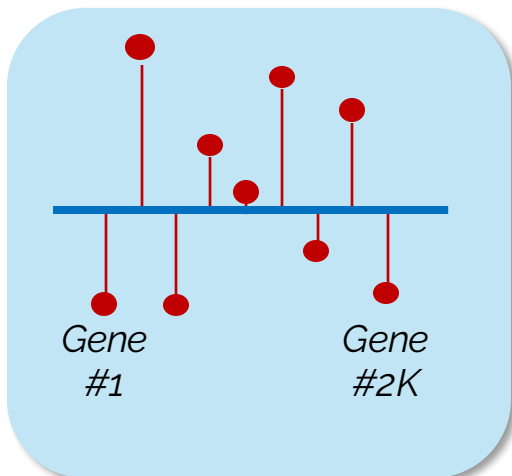


Cell types latent



- BEC
- B\_cells
- CD4\_T\_cells
- CD8\_T\_cells
- FDC
- Fibroblasts
- LEC
- Macrophages
- Monocytes
- NK\_cells
- Plasma\_cells
- T\_other
- Tumor
- mDC
- pDC

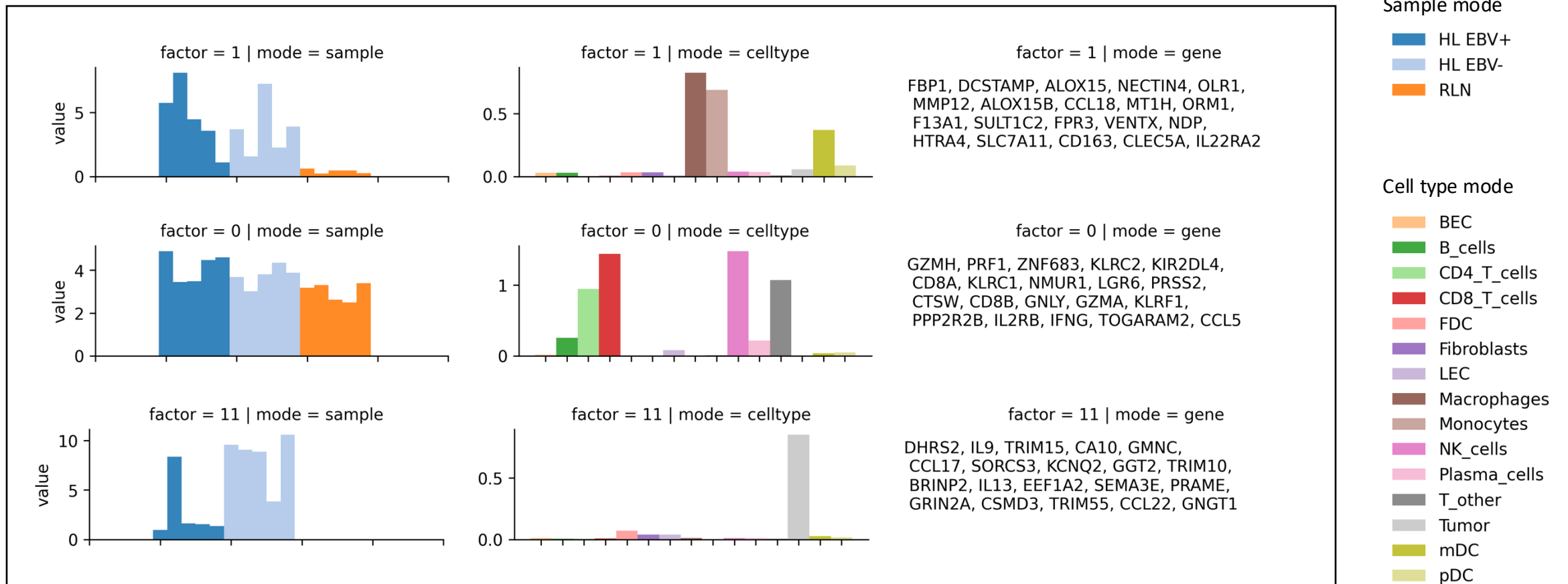
genes latent factor



JAK/STAT pathway promotes tumor cell proliferation and survival of tumor cells

# Interpretation of recovered factors - Rank 20

- At intermediate ranks, factors begin to pick up GEPs that characterize cell type specific sample heterogeneity
- Some factors continue to pick up cell type identity programs that are conserved across all samples
- Factors corresponding to Tumor cell type split by mostly EBV+ and mostly EBV-

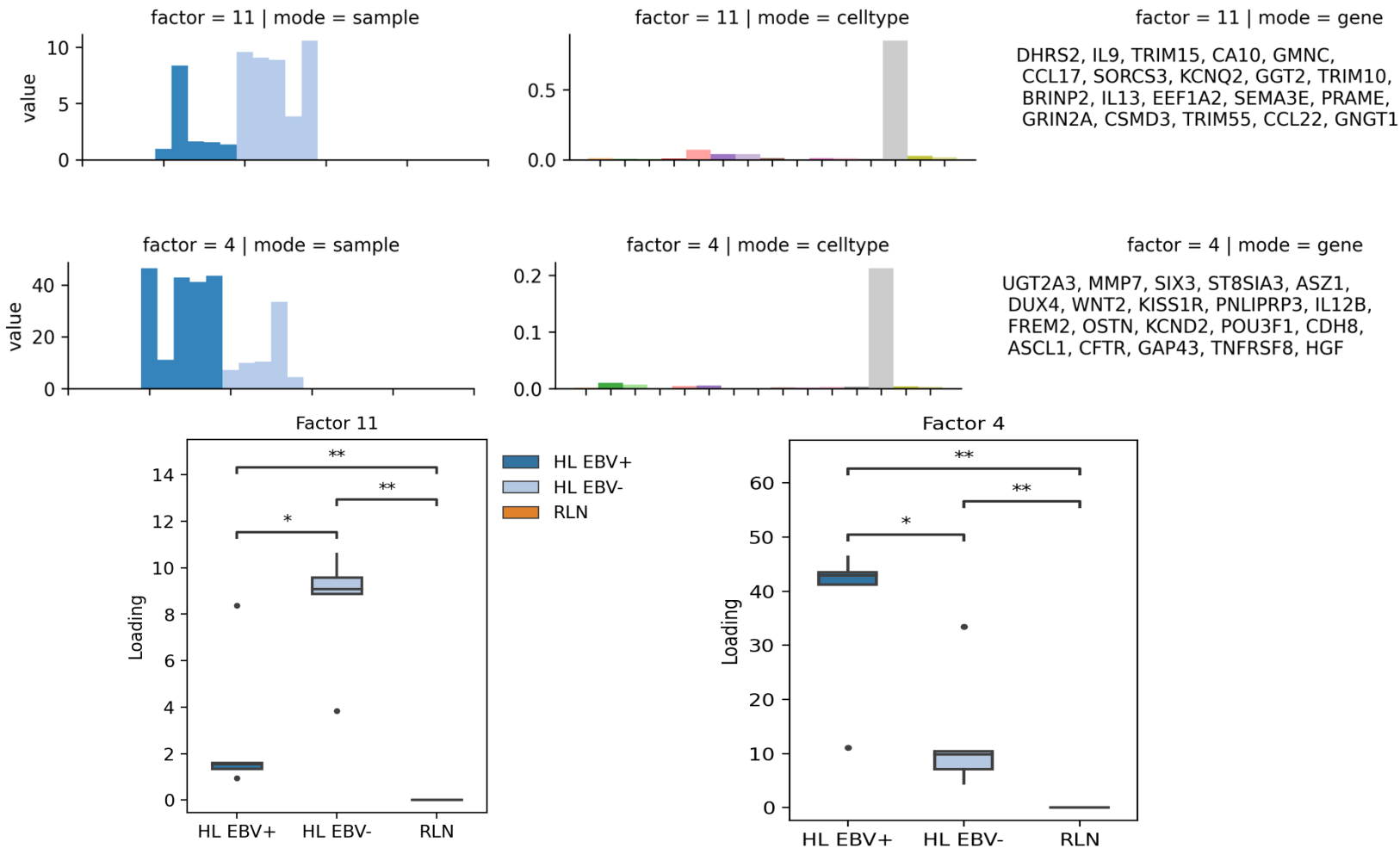




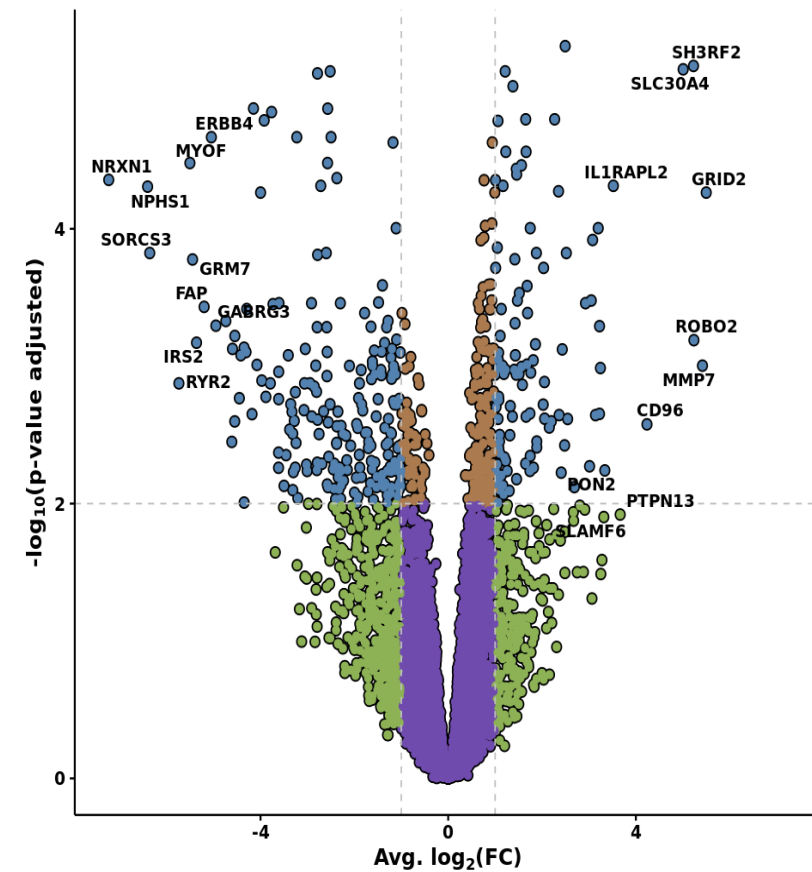
# C-ZIPTF identifies heterogeneity in tumor cell type identity gene expression program

- We identify factors that are capture heterogeneity in tumor cell type identity across different patients
- These subgroupings are mostly in line with clinically annotated EBV status
- The gene programs captured concur with the genes that are recovered by DE testing

## Factors from Rank 20

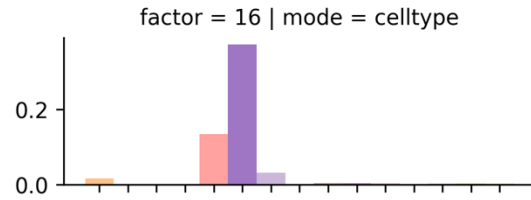
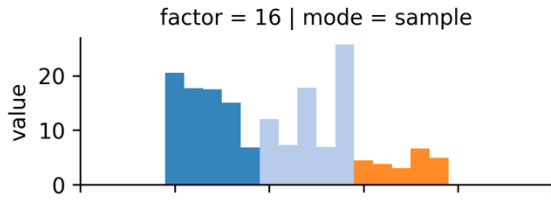


## DE Genes in Tumor: EBV neg(<-->) vs EBV pos(<-->)



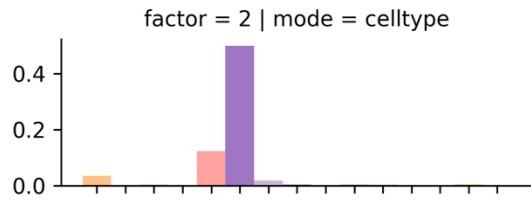
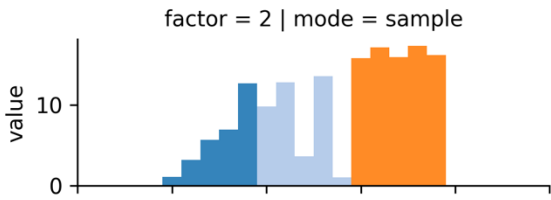
# C-ZIPTF identifies gene expression programs for cancer associated fibroblasts

## Factors from Rank 20



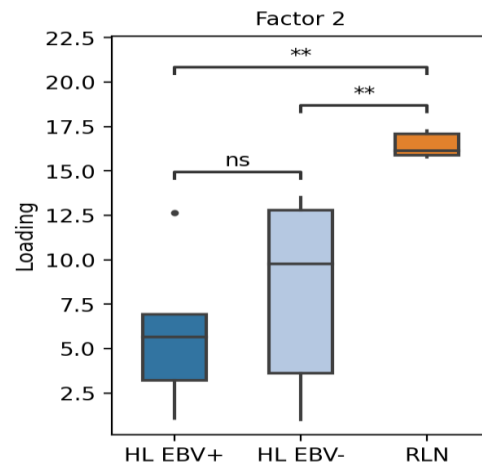
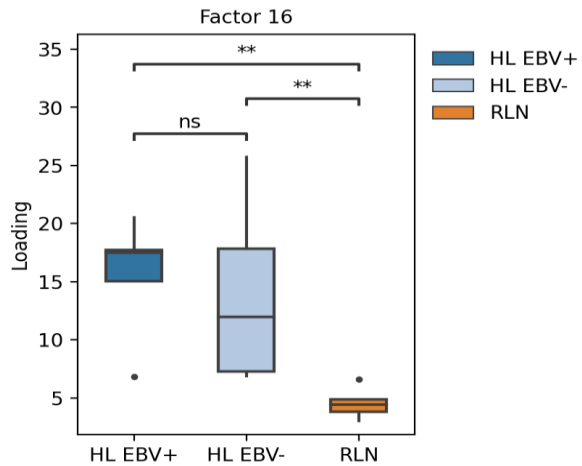
factor = 16 | mode = gene

SFRP2, THBS2, PLA2G2A, COL3A1, POSTN, MMP1, GPC6, LRRC15, COL1A1, SULF1, CA12, COMP, STEAP2, COL1A2, COL8A1, COL16A1, FBLN1, ITGBL1, DCN, WT1

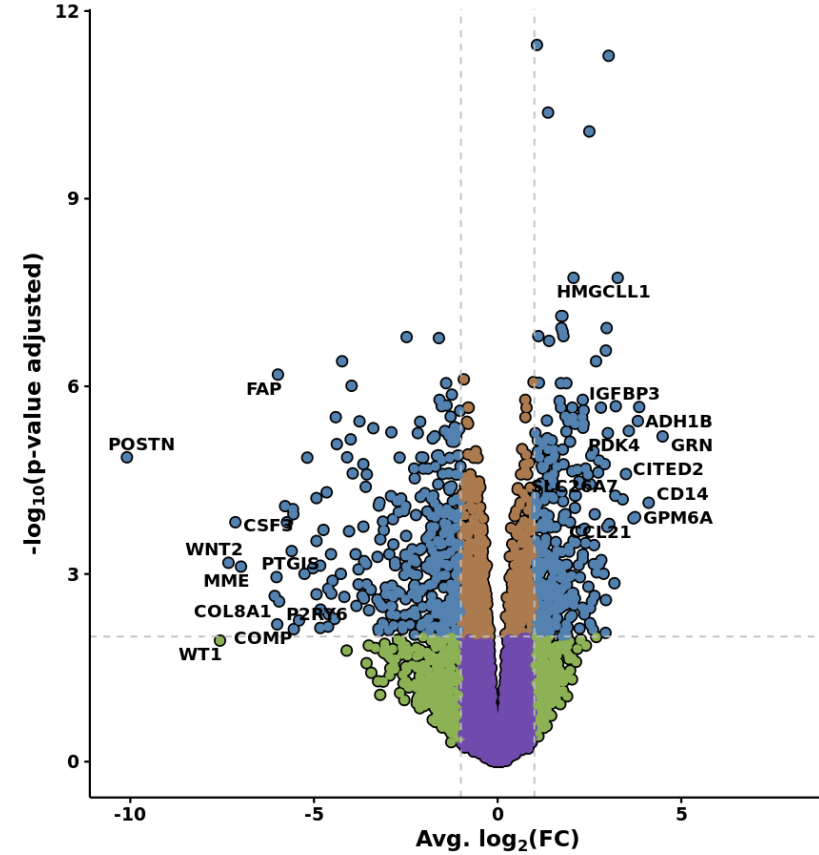


factor = 2 | mode = gene

ADH1B, ADRA1A, GREM1, CCL21, SLC26A7, ID4, TNFSF11, CLSTN2, SLITRK2, BMP4, CXCL12, C7, CXCL14, ADAM33, NGF, PTPRQ, APLNR, COL14A1, DES, RASL11A

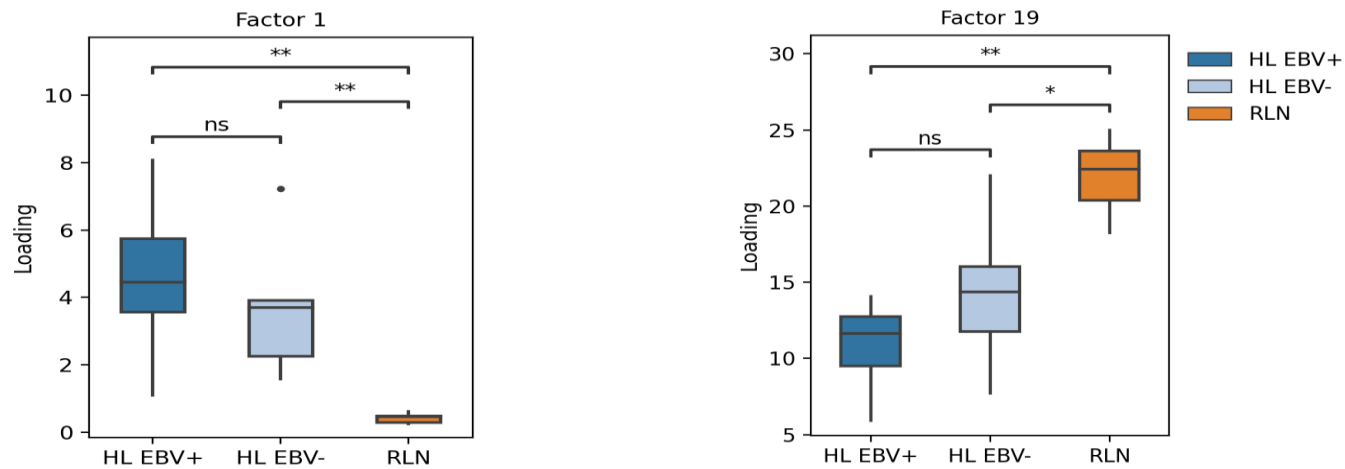
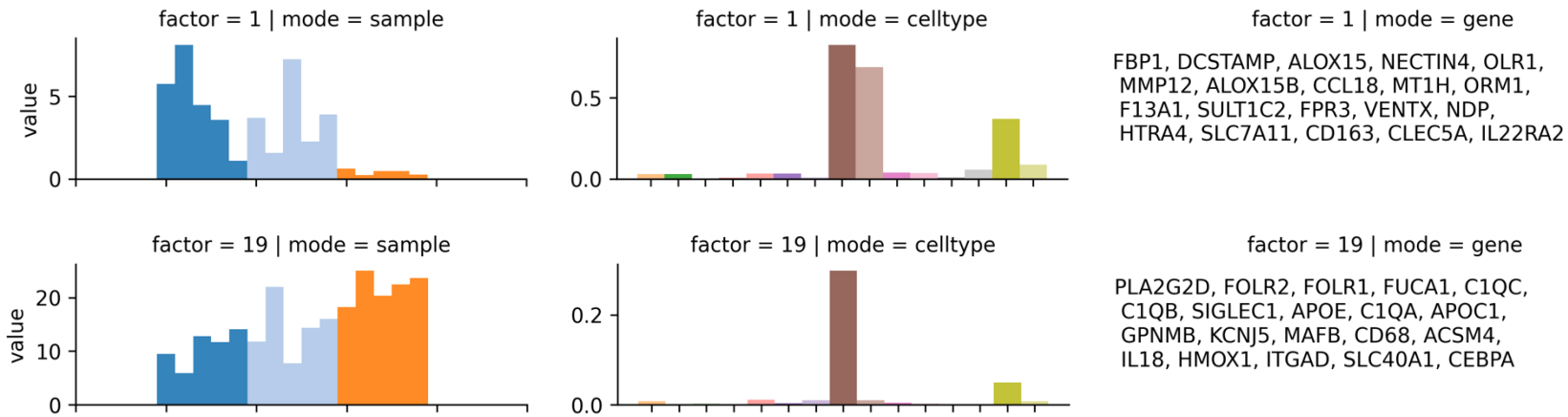


## DE Genes in Fibroblasts: HL(<---) vs RLN(--->)



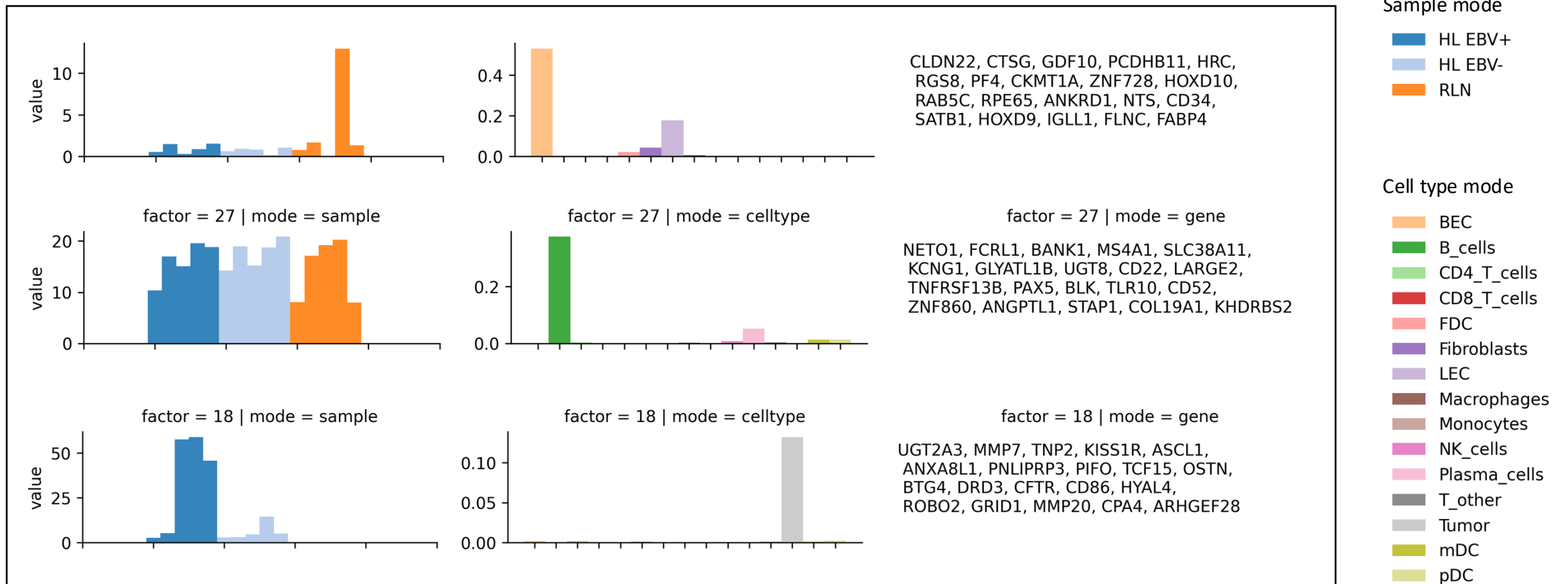
# C-ZIPTF identifies gene expression program upregulated in monocytes from tumor samples

Factors from Rank 20



# Interpretation of recovered factors - Rank 40

- At much higher ranks, the factors begin to break down to individual donor specific gene expression programs.
- Again, some factors continue to pick up cell type identity programs that are conserved across all samples
- Tumor cell type GEPs subdivide the HL samples at higher resolution



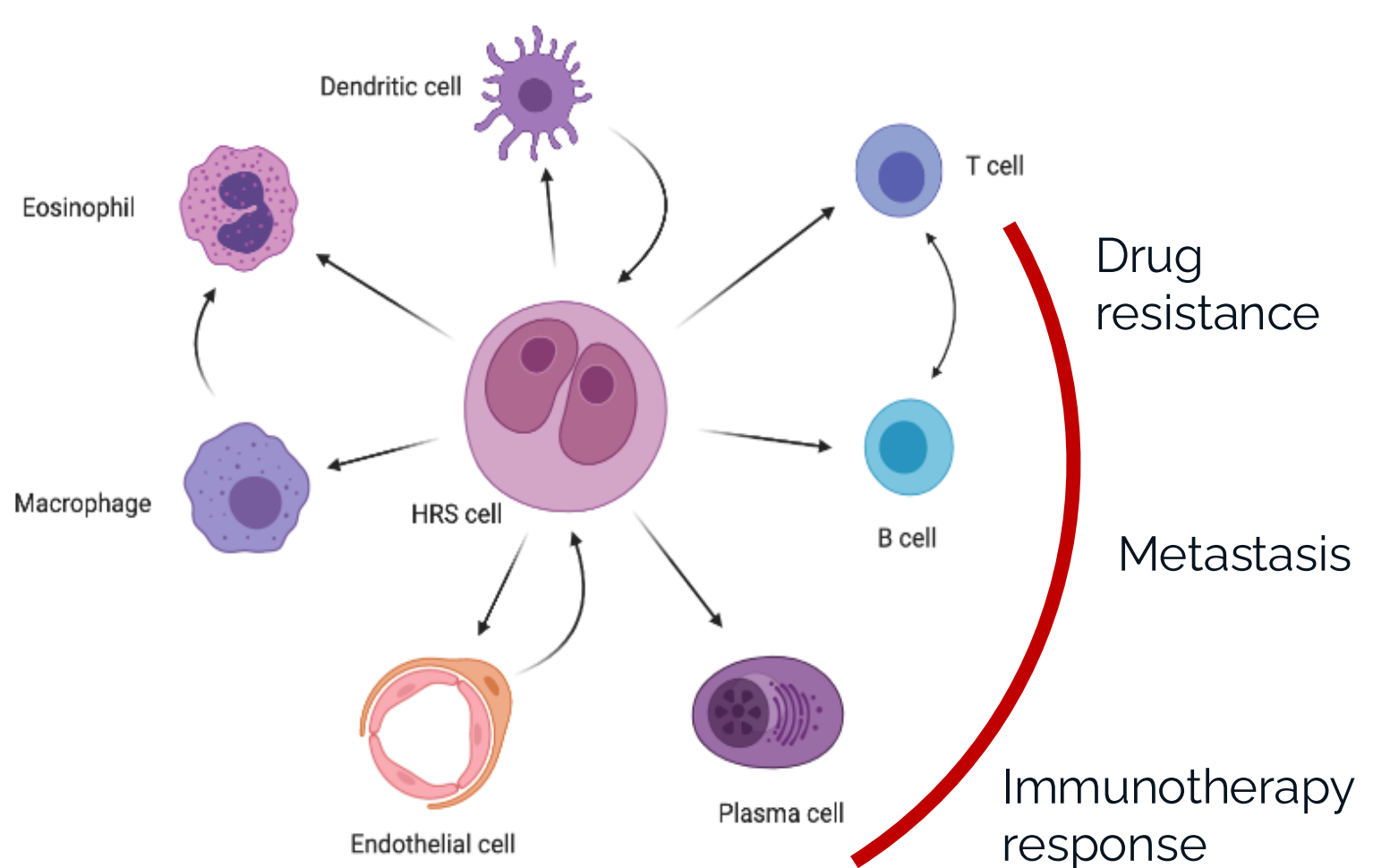
# Malignant B Cells of CHL **depend** on the microenvironment

What are the cell **types/states** within the TME of CHL?

What cell types/states are **enriched** around HRS cells?

TME-specific **survival & growth** signaling?

Mechanisms of immune **evasion**?

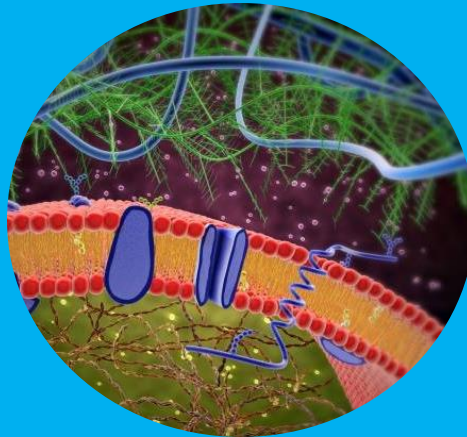
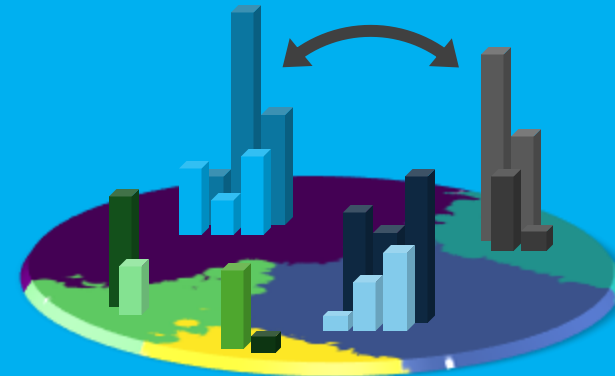




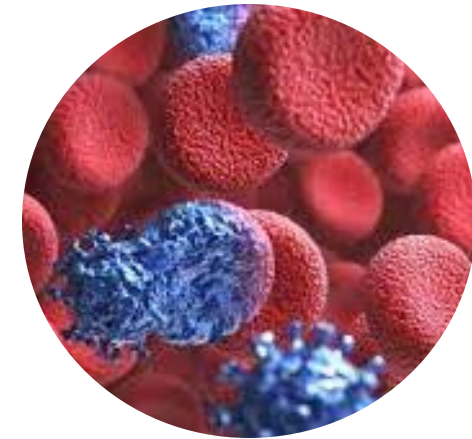
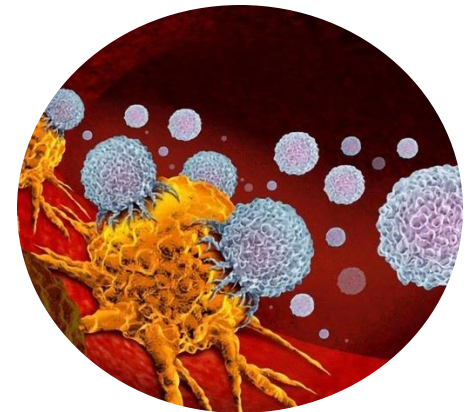
## Inter- & Intra-patient Heterogeneity



## Cross-talks & Pathways within the TME



## Immunotherapy Targets

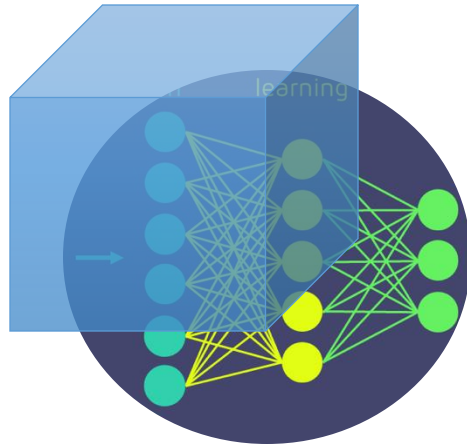




3

# Future Projects & Goals

**Future** long-term adventures



Integrations of **Tensor Methods** and **Deep Learning** approaches



Theoretical improvements of **Tensor Algorithms**



Applications in **additional domains**





# Thanks!

Neriman Tokcan <[neriman.tokcan@umb.edu](mailto:neriman.tokcan@umb.edu)>

