Probabilistic modelling of single-cell methylation sequencing data

Keegan Korthauer

3 July 2023

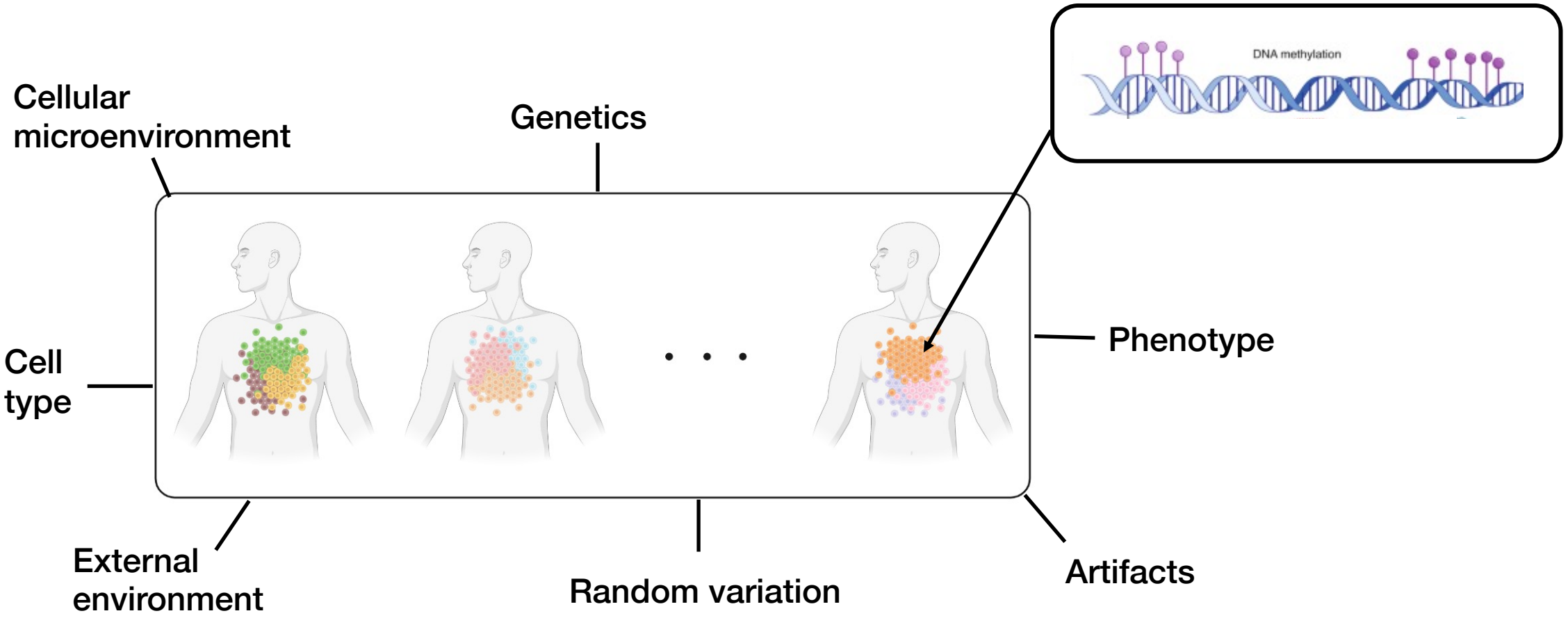
BIRS Workshop “Single-Cell Plus”

BC Children's Hospital Research Institute

THE UNIVERSITY OF BRITISH COLUMBIA

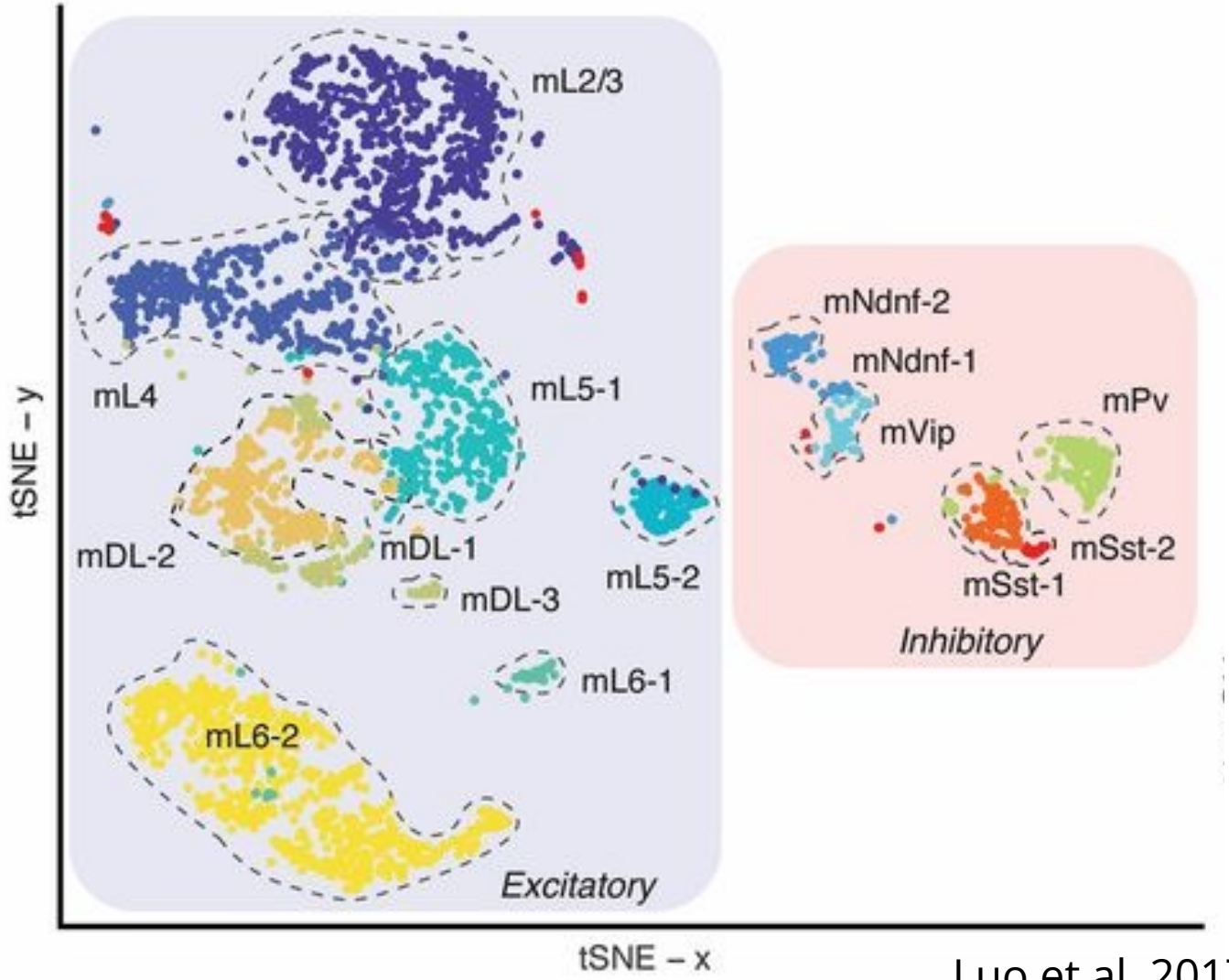
Department of Statistics
Faculty of Science

Heterogeneity in DNA methylation (DNAm)



Motivating example: mouse neuronal scDNAm

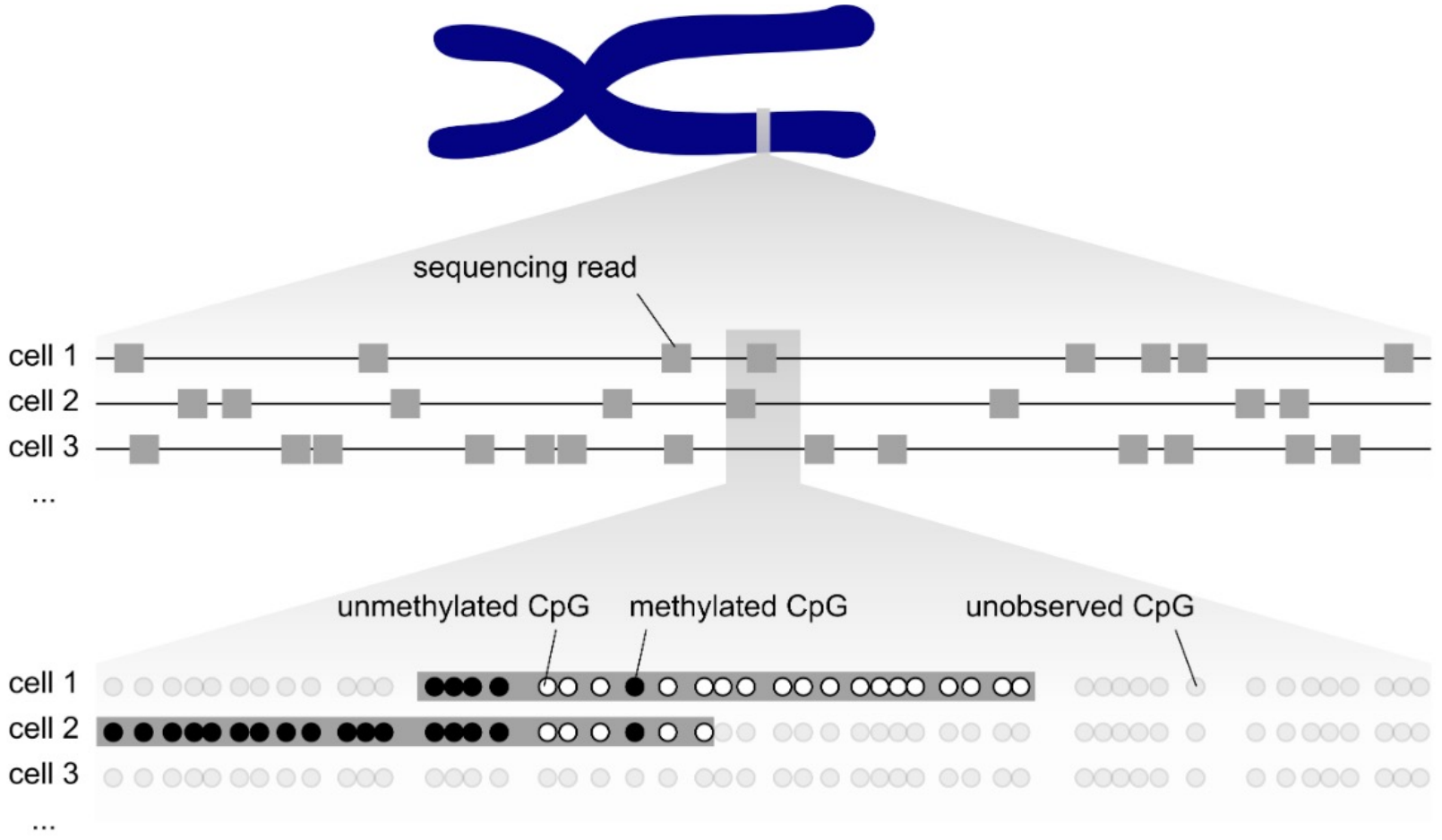
- tSNE of ~3K mouse neurons
- features: DNAm level in non-overlapping 100kb bins
- clusters annotated using gene body DNAm depletion in neuronal marker genes



Luo et al. 2017 (Science)

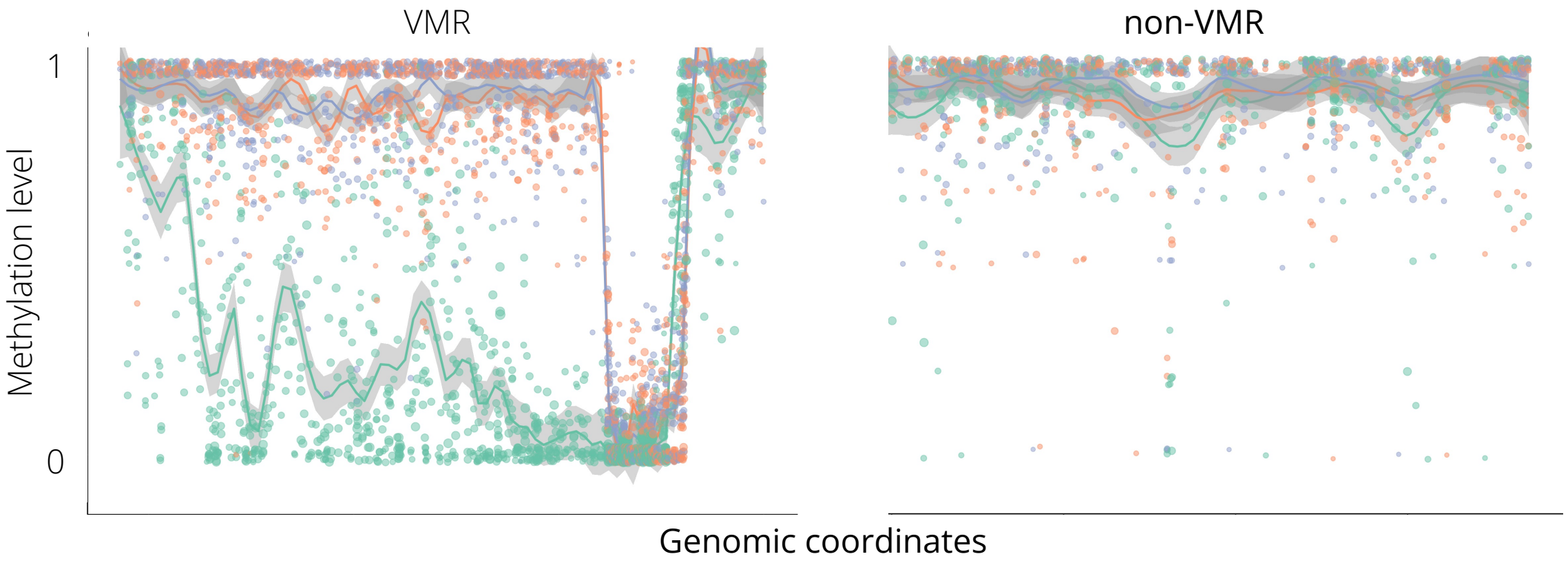
Key data science challenges

1. high dimensionality (~30M CpGs in human)
2. lack of independence of nearby CpGs
3. high sparsity (80-95+% missing)

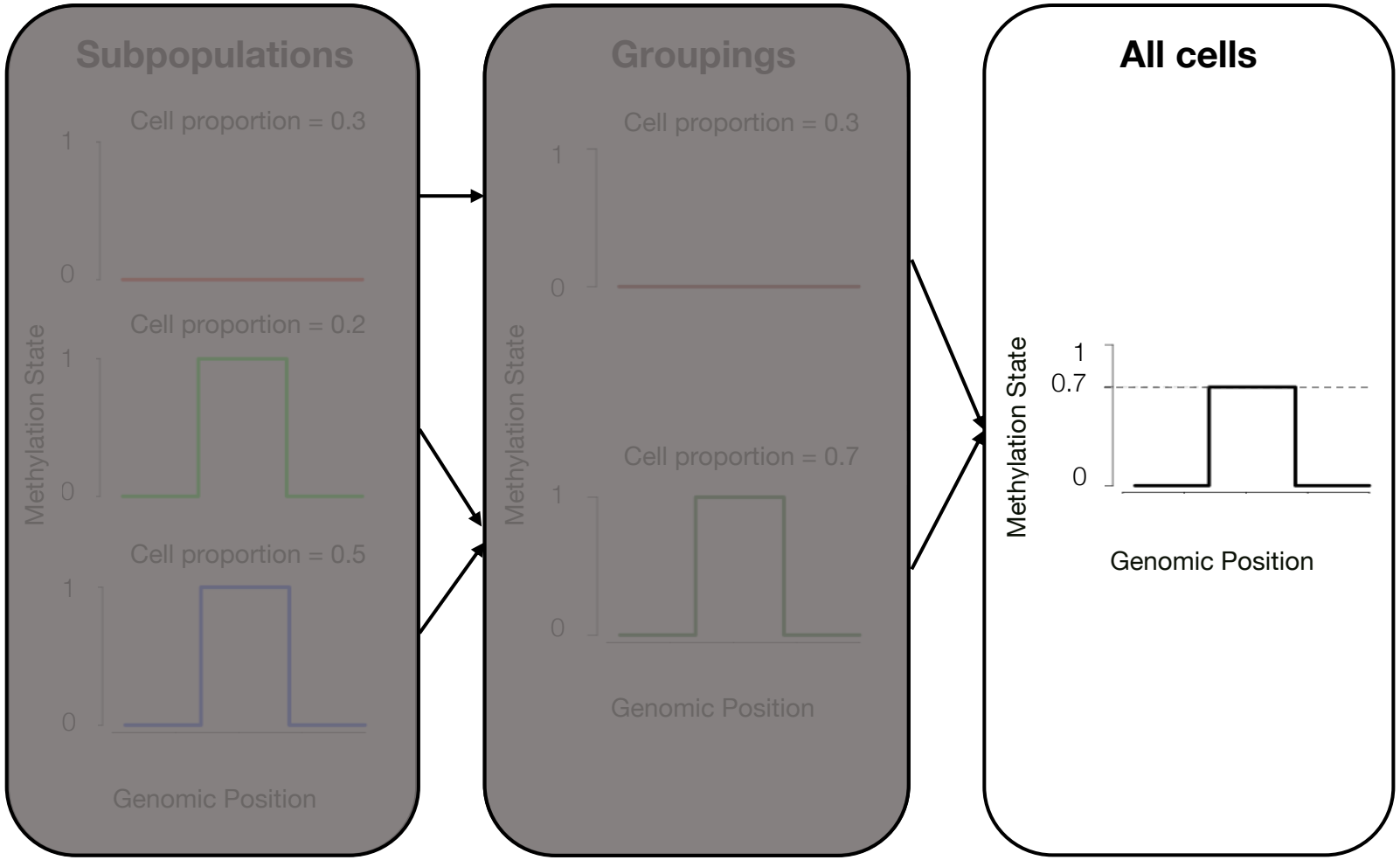


Adapted from Fig 1, Kremer et al. 2022 (bioRxiv)

Variably Methylated Regions (VMRs)



Detection of subpopulations



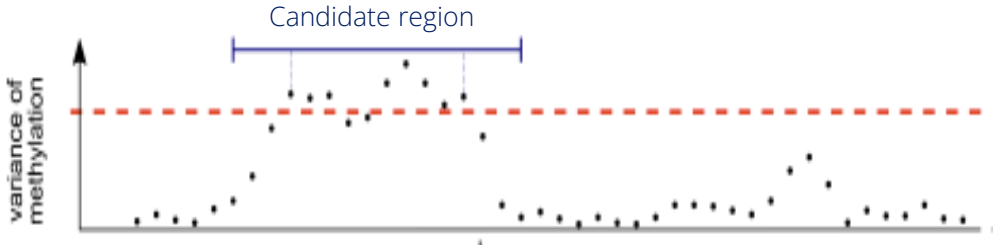
vmrseq step 1: scan genome for candidate regions

For each site j, compute variance across cells

$$\sigma_j^2 = \frac{\sum(\widehat{x_{ij} - \bar{x}_j})^2}{n_j - 1}$$

Relative methylation levels are used to adjust for uneven coverage biases
Kernel smoothing is used to borrow strength from nearby sites

Identify contiguous sites with high variance as candidate regions (CRs)



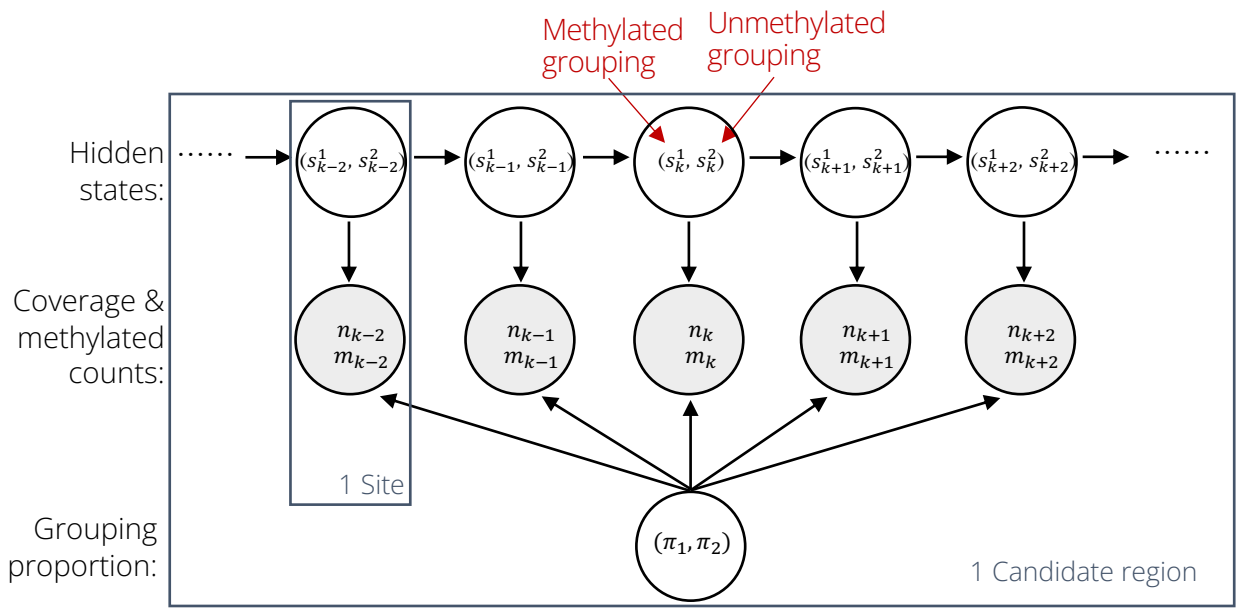
Adapted from Fig 2, Kremer et al. 2022 (bioRxiv)

Independent training data is used to simulate a **null distribution of variance** from cells of the same cell type

vmrseq step 2: compare likelihood of 2 vs 1-state

Decode **2-grouping** hidden Markov model

$$p(\vec{m}|\vec{n}, N_{\text{Grouping}} = 2)$$



(s_k^1, s_k^2) = (0, 0), (1, 0) or (1, 1)

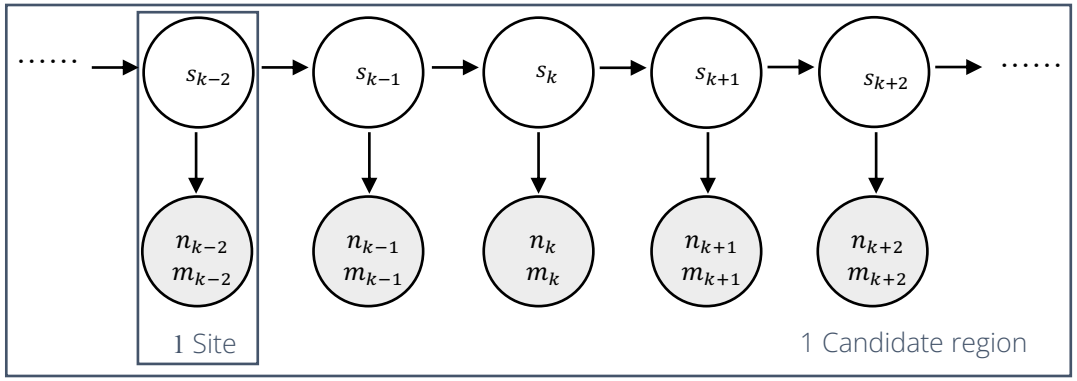
$>$?

OR

$<$?

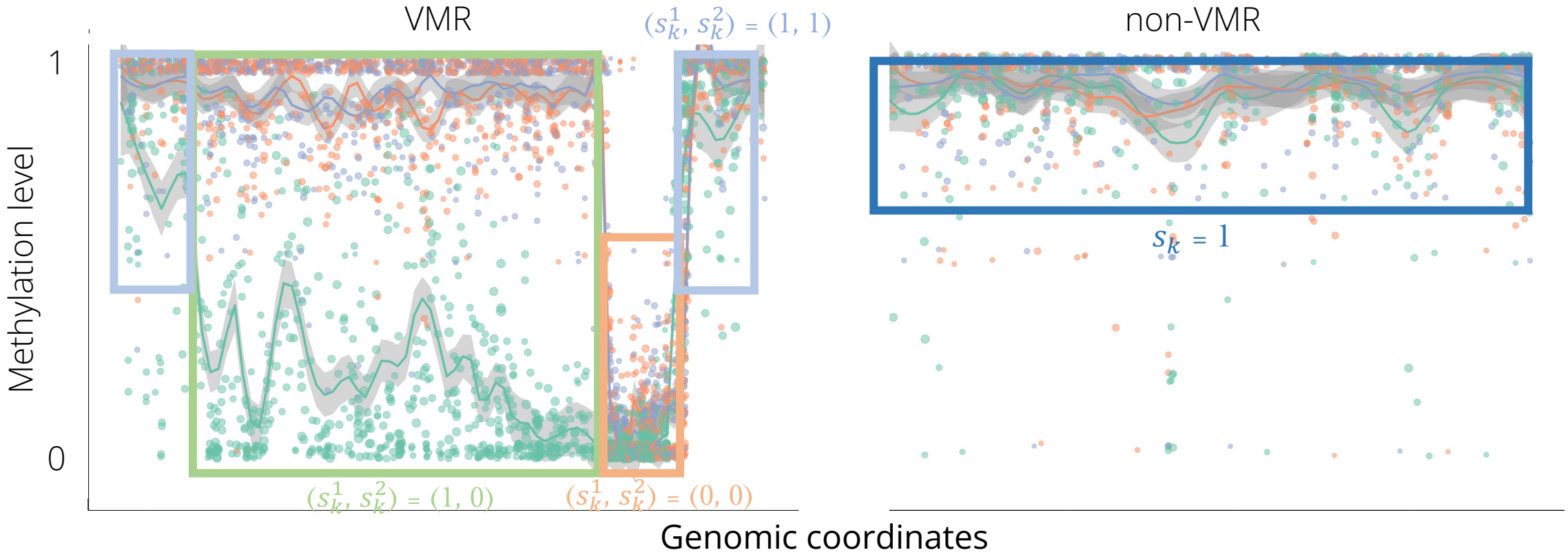
Decode **1-grouping** hidden Markov model

$$p(\vec{m}|\vec{n}, N_{\text{Grouping}} = 1)$$



s_k = 0 or 1

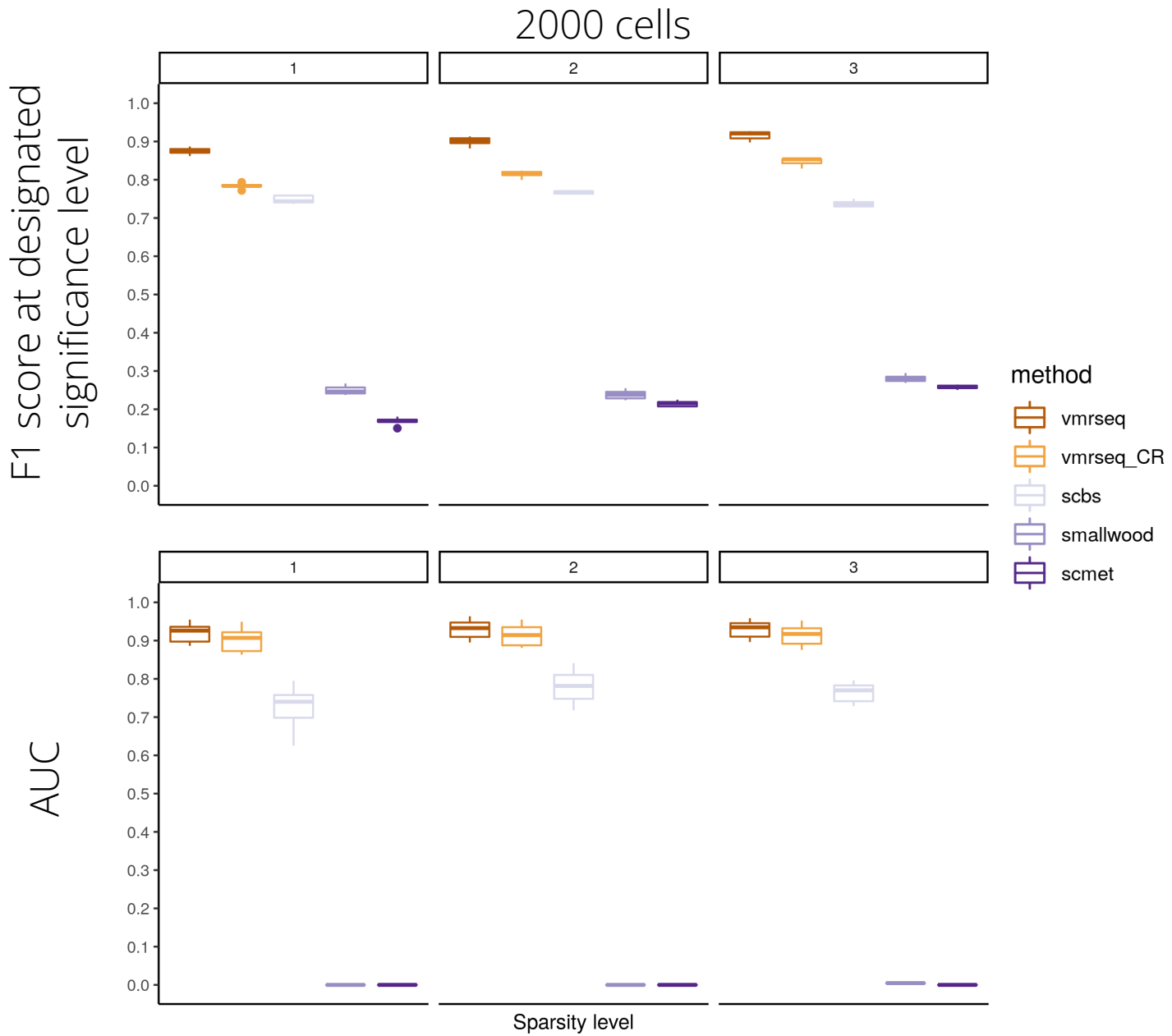
vmrseq output



$p(\vec{m}|\vec{n}, N_{\text{Grouping}} = 2) > p(\vec{m}|\vec{n}, N_{\text{Grouping}} = 1)$

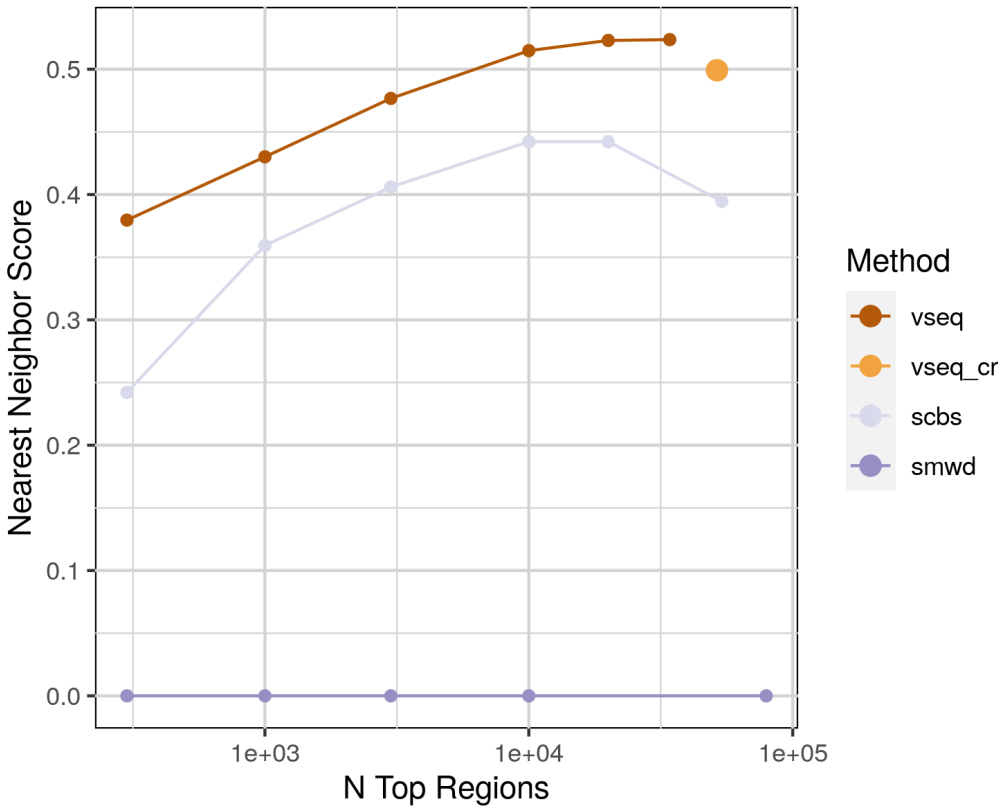
$p(\vec{m}|\vec{n}, N_{\text{Grouping}} = 2) < p(\vec{m}|\vec{n}, N_{\text{Grouping}} = 1)$

Clustering using vmrseq features increases simulation accuracy

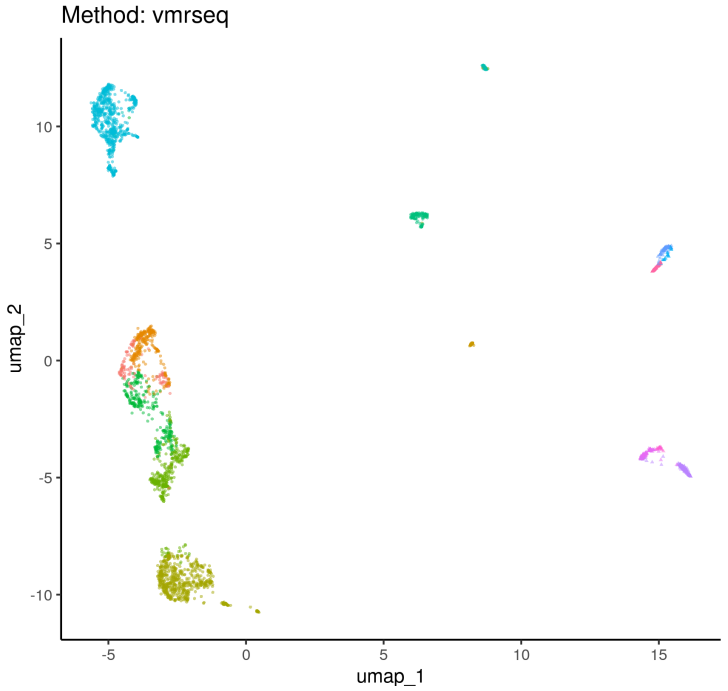


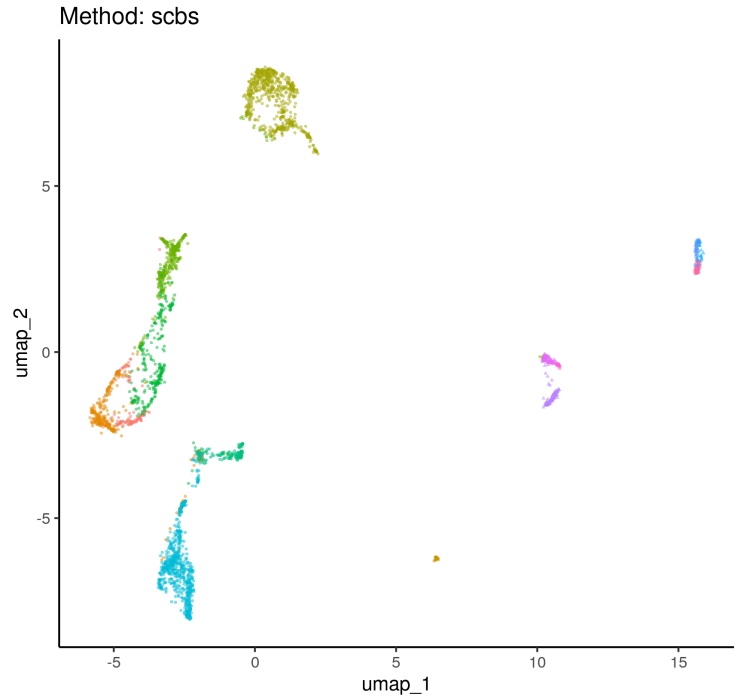
Shen & Korthauer (unpublished)

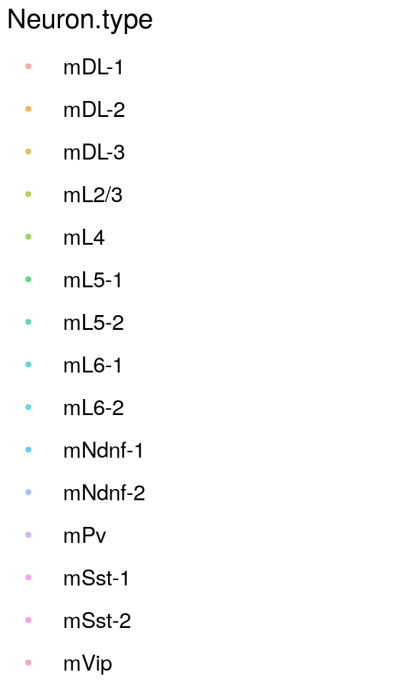
Clustering mouse neurons using vmrseq features yields increased cell type separation

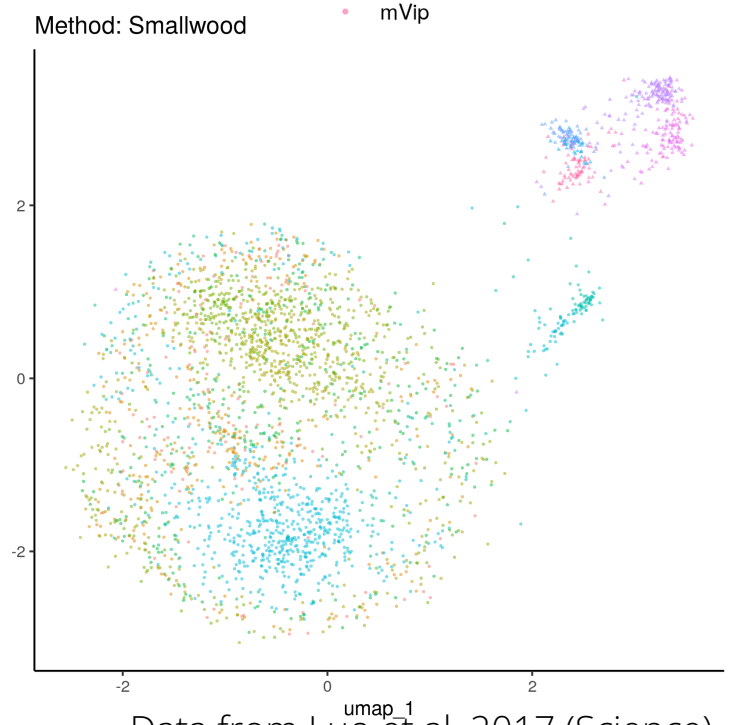


Shen & Korthauer (unpublished)



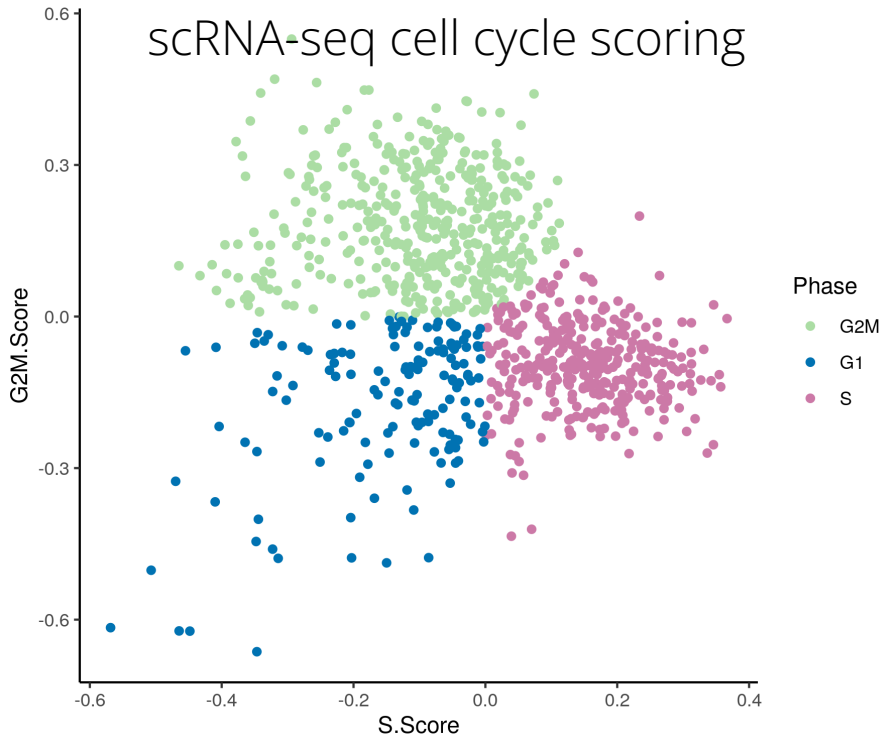




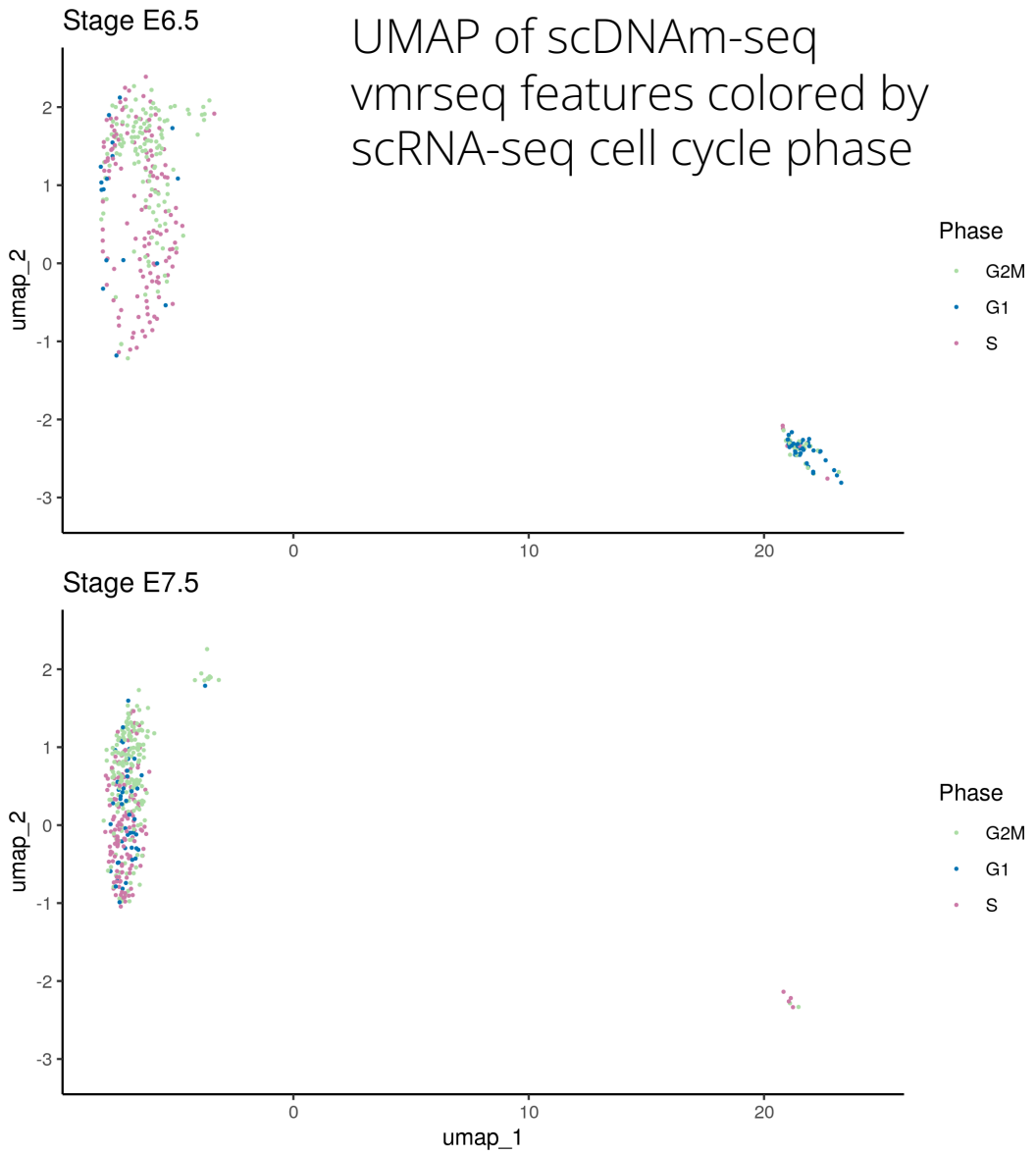


Data from Luo et al. 2017 (Science)

vmrseq features reveal heterogeneity associated with cell cycle

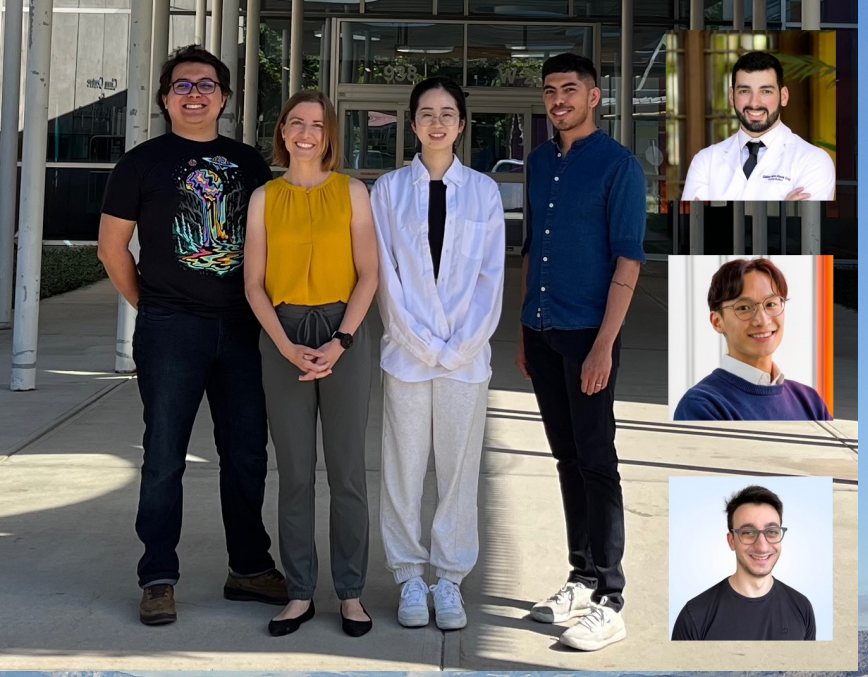


UMAP of scDNAm-seq vmrseq features colored by scRNA-seq cell cycle phase



Mouse gastrulation multi-omic profiling with single-cell DNAm and RNA-seq by Argelaguet et al. 2019 (Nature)

Shen & Korthauer (unpublished)



BC Children's Hospital Research Institute

THE UNIVERSITY OF BRITISH COLUMBIA

Department of Statistics
Faculty of Science

NSERC CRSNG

Korthauer Lab

Giuliano Netto Flores Cruz

Erick Isaac Navarro Delgado

Irvin Ng

Marco Tello Palencia

Ning Shen

Giuseppe Tinti Tomio

Funding

NSERC Discovery Grant

BCCHR IGAP

BCCHR Establishment Award