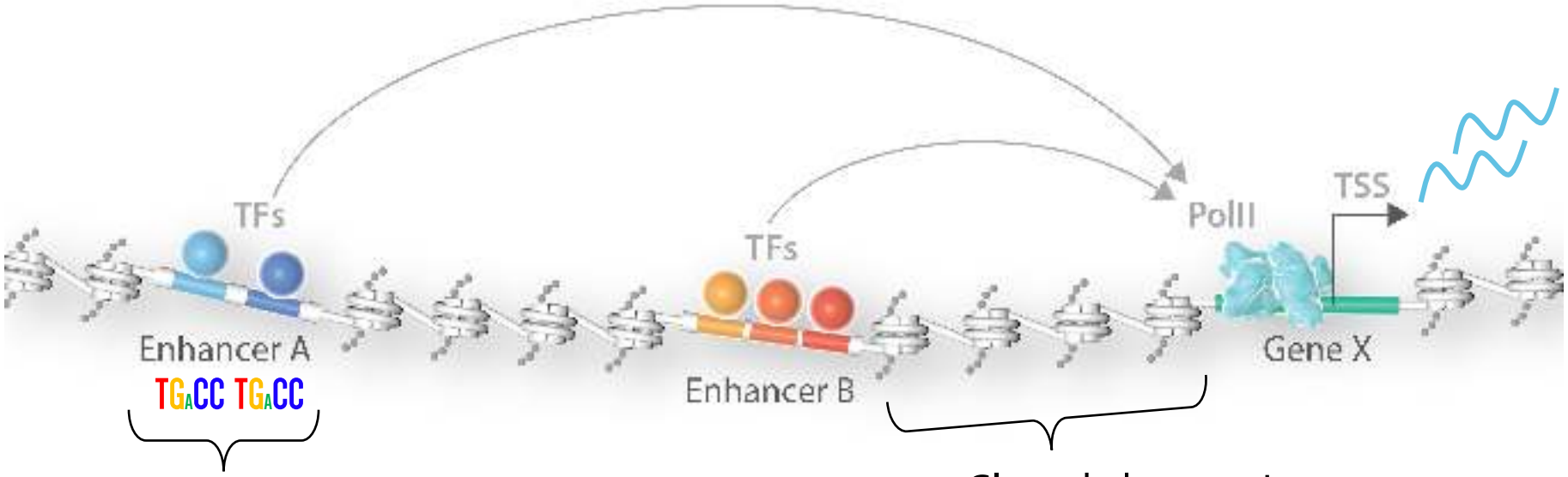
cisTopic

cis-Regulatory topic modelling of single-cell **epigenomes**

VIB KU LEUVEN

Deciphering gene regulatory programs



Nucleosome depleted chromatin bound by TFs

Closed chromatin

Studying accessible chromatin can reveal gene regulatory programs

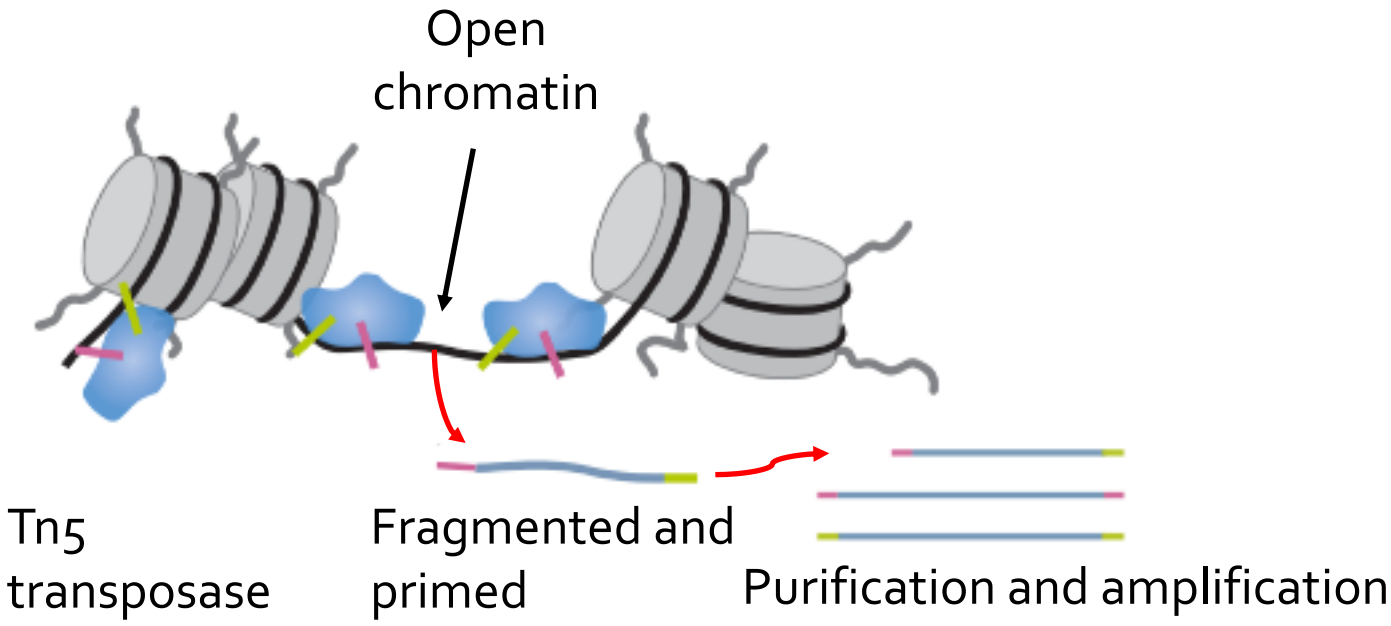
Measuring accessible chromatin

Table 1. Current approaches used to map distal enhancers genome-wide

Approach	Activity[1]	Specificity[2]
DNase I-seq	Open	Non
NA-seq	Open	Non
FAIRE	Open	Non
P300 (ChIP-seq)	Open	Enh. = Prom.
H3K4me1 (ChIP-seq)	Open	Enh. > Prom.
H3K4me2 (ChIP-seq)	Open	Enh. = Prom.
H3K27ac (ChIP-seq)	Active	Enh. > Prom.
H3K4me3 (ChIP-seq)	Active	Prom. > Enh.
Pol II (ChIP-seq)	Active	Prom. > Enh.
BRG1 (ChIP-Seq)	Active	Enh. = Prom.

Spicuglia & Vanhille. Chromatin signatures of active enhancers. (Nucleus, 2012)

Measuring accessible chromatin

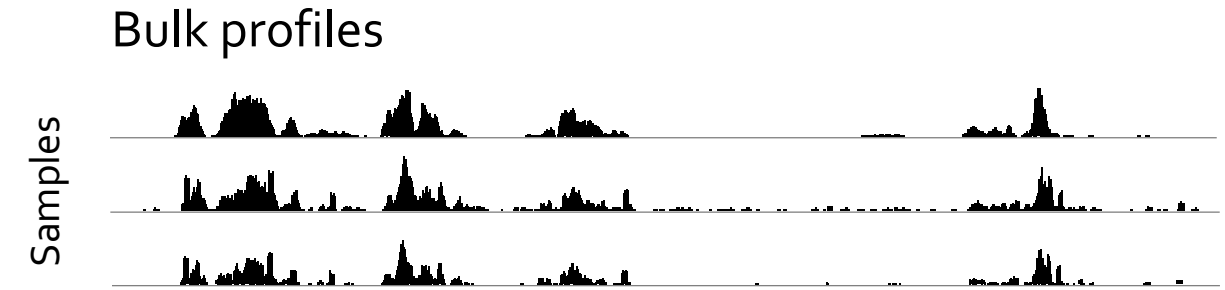


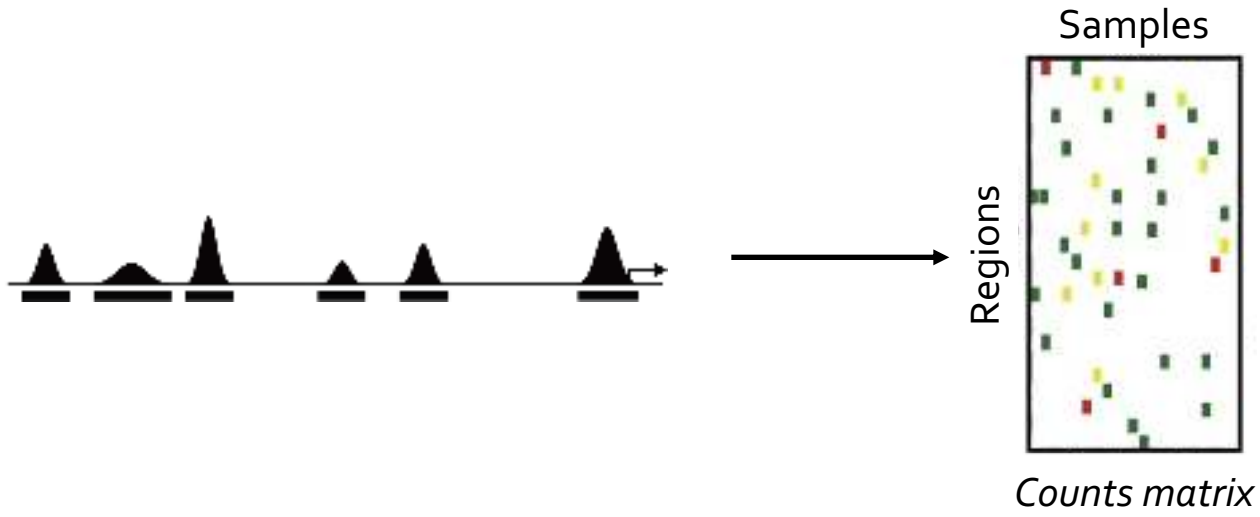
ATAC-seq

Buenrostro et al. ATAC-seq: A method for assaying chromatin accessibility Genome-wide. (Curr Protoc Mol Biol,, 2015)

Analysis pipeline

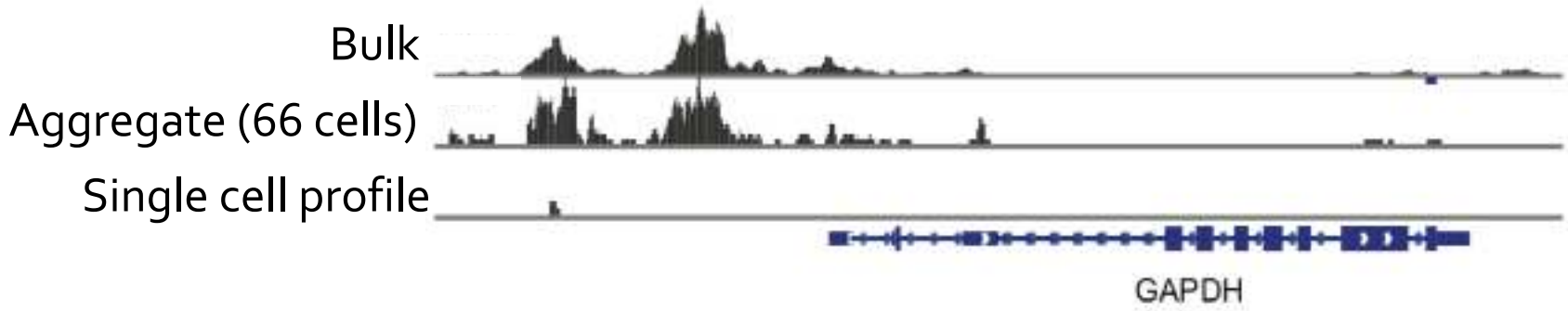
Software	Step
Fastq-mcf Fastqc	Cleaning + QC
Bowtie2	Mapping
MACS2	Peak calling
Subread	Feature counts





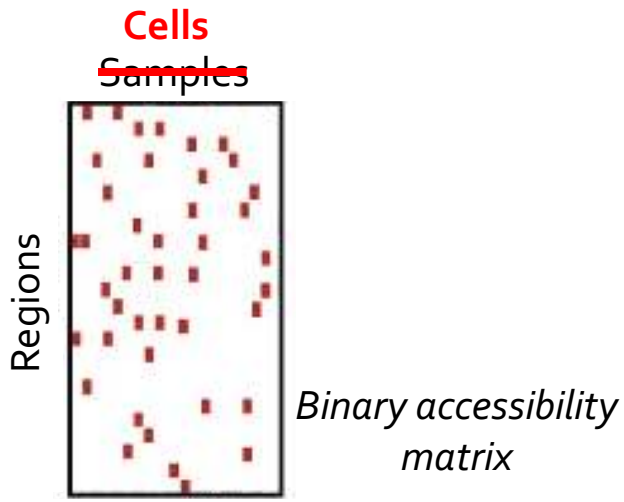
The challenges in single cell epigenomics

Sparsity

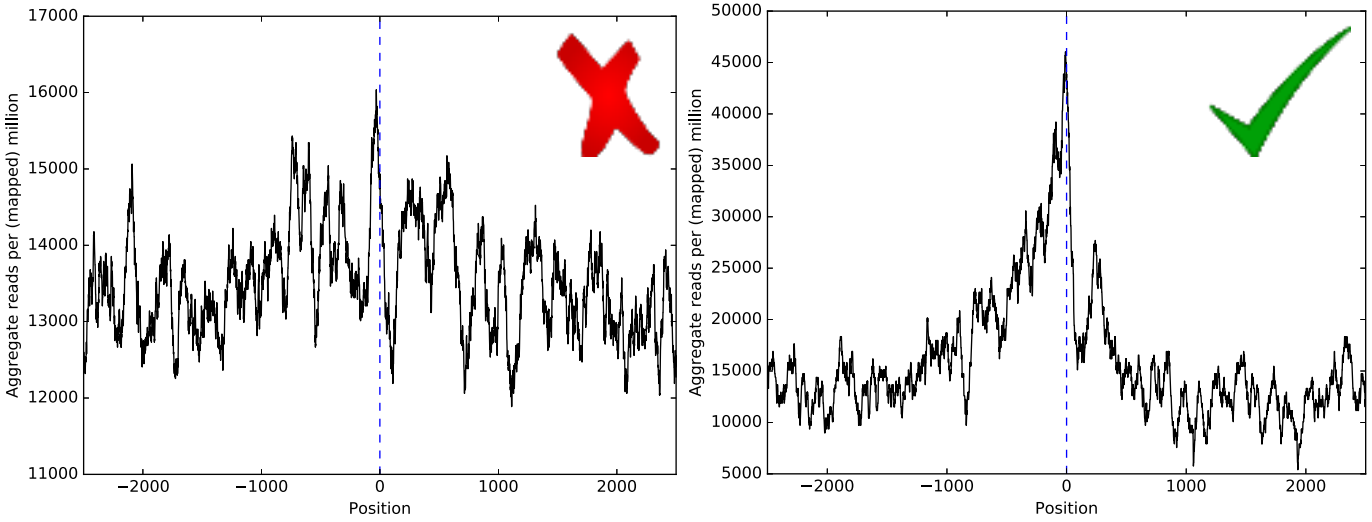


Scalability

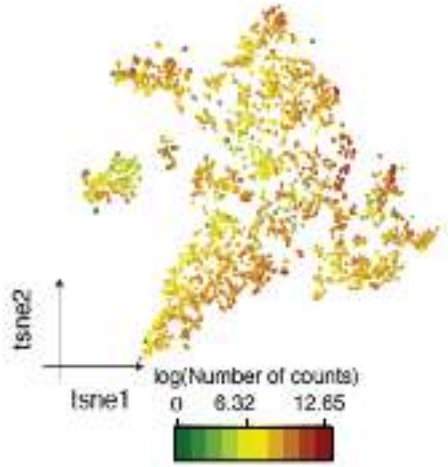
Up to 100,000 cells and more than 500,000 regions



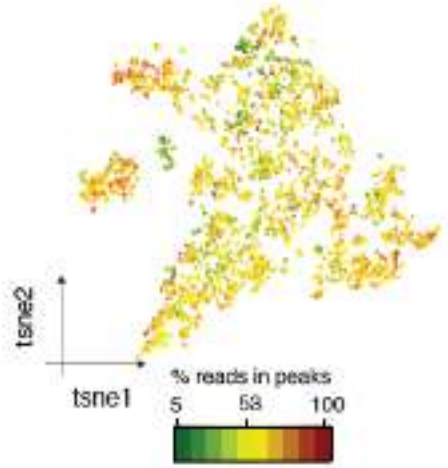
Analysis pipeline: Cell filtering



Aggregation of reads around TSS

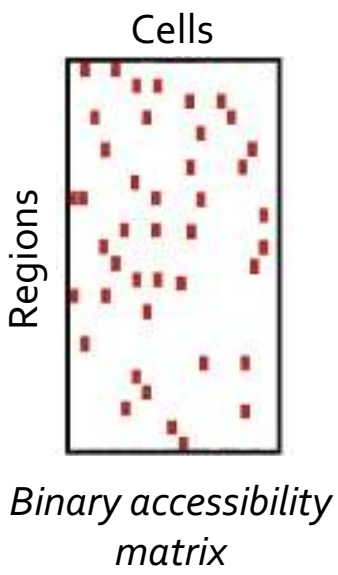


Number of reads

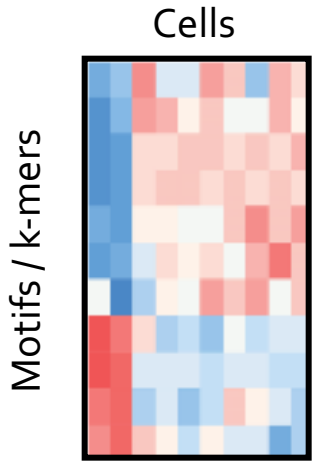


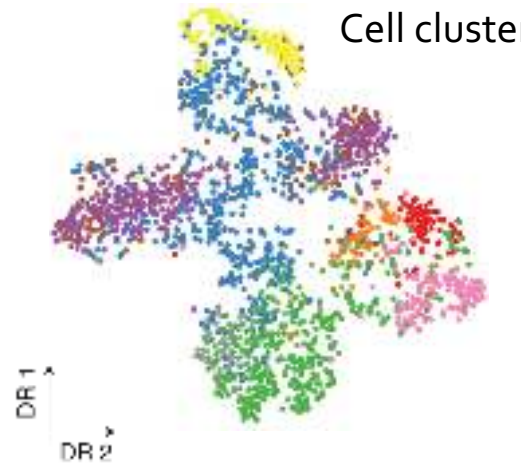
Percentage reads in peaks

Current approaches



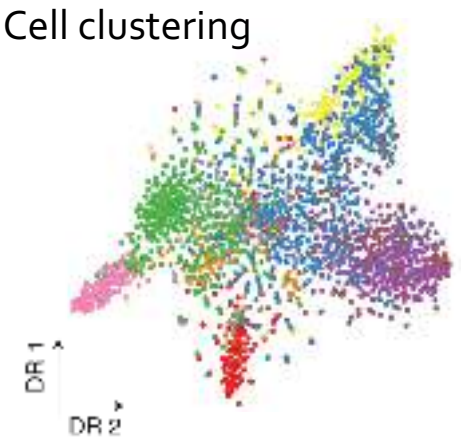
Region-based

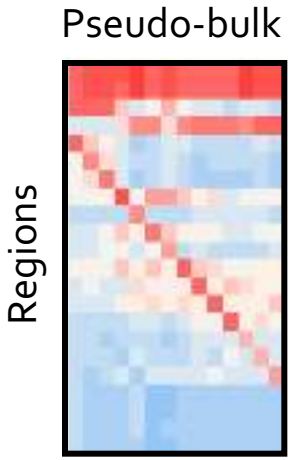




E.g. **SCRAT** (Ji *et al.*, 2017), **chromVAR** (Schep *et al.*, 2017) & **BROCKMAN** (de Boer & Regev, 2018)

Cell-based

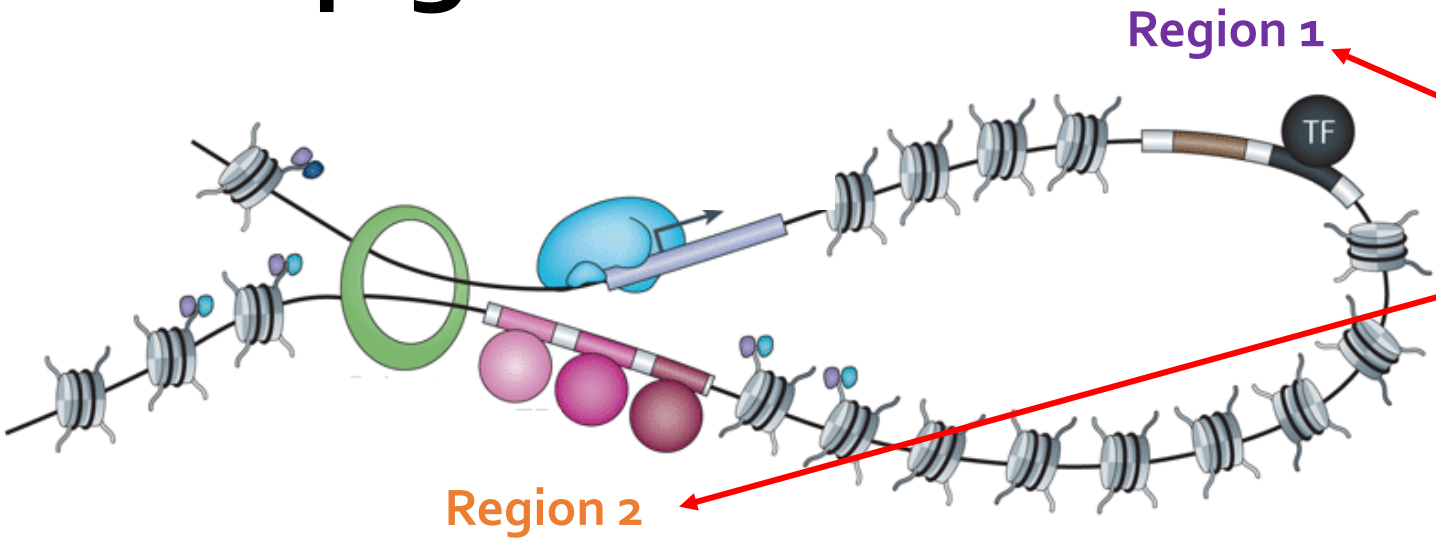


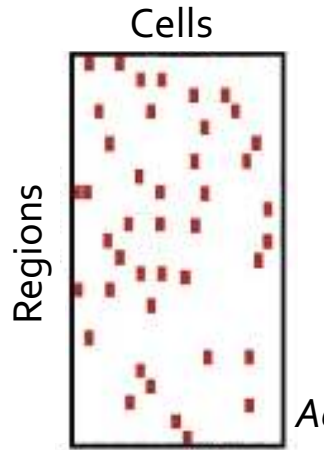


E.g. **LSI** (Cusanovich *et al.*, 2015), **scABC** (Zamanighomi *et al.*, 2018) & **Cicero** (Pliner *et al.*, 2018)

Can we co-optimize regions' and cells' clustering for better information retrieval from the data?

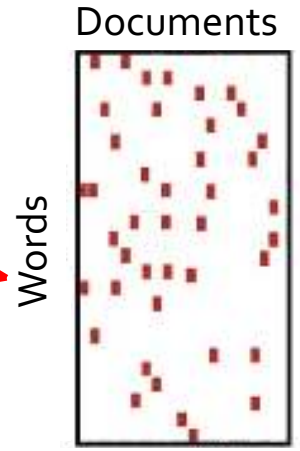
The link between topic models and single cell epigenomics





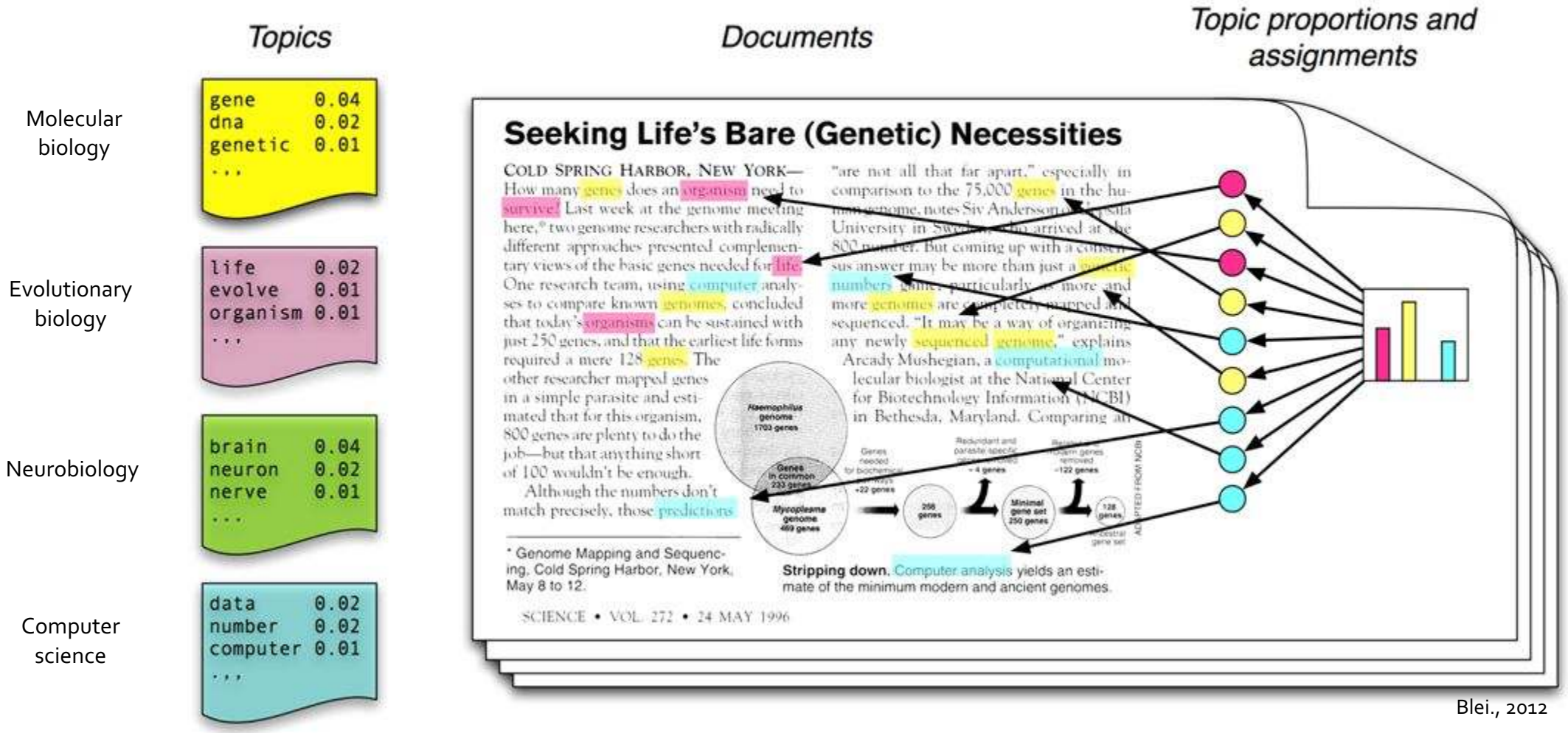
Accessibility matrix

We present SCENIC, a **computational method** for simultaneous **gene regulatory network** reconstruction and **cell-state** identification from **single-cell RNA-seq** data (**http://scenic.aertslab.org**). On a compendium of single-cell **data** from **tumors** and **brain**, we demonstrate that *cis*-regulatory **analysis** can be exploited to guide the identification of **transcription factors** and **cell states**. SCENIC provides critical biological insights into the mechanisms driving cellular heterogeneity.



Word-document coocurrence matrix

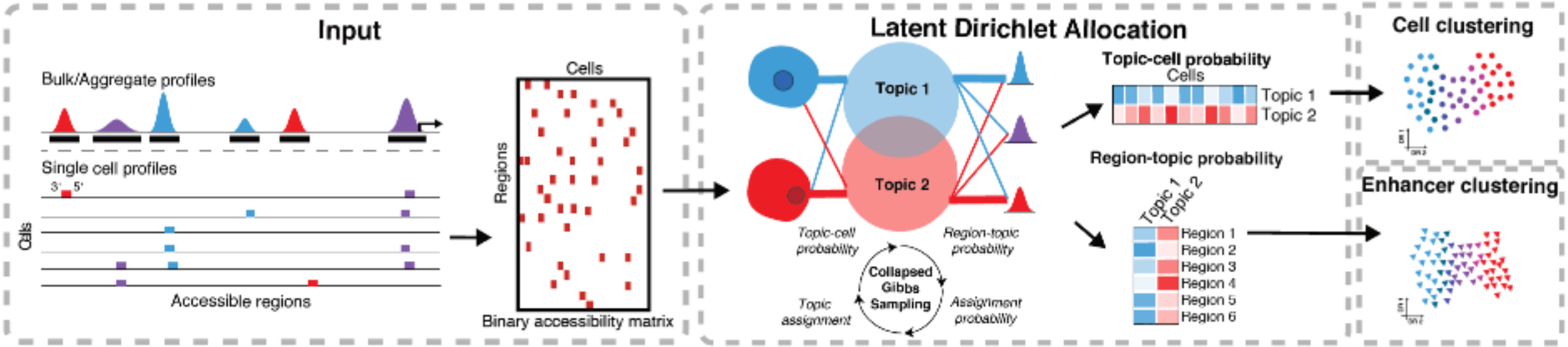
Latent Dirichlet Allocation (LDA)



Blei., 2012

Can we learn unsupervisedly cis-regulatory topics from single cell epigenomics data?

cisTopic: A new framework for single cell epigenomics data analysis



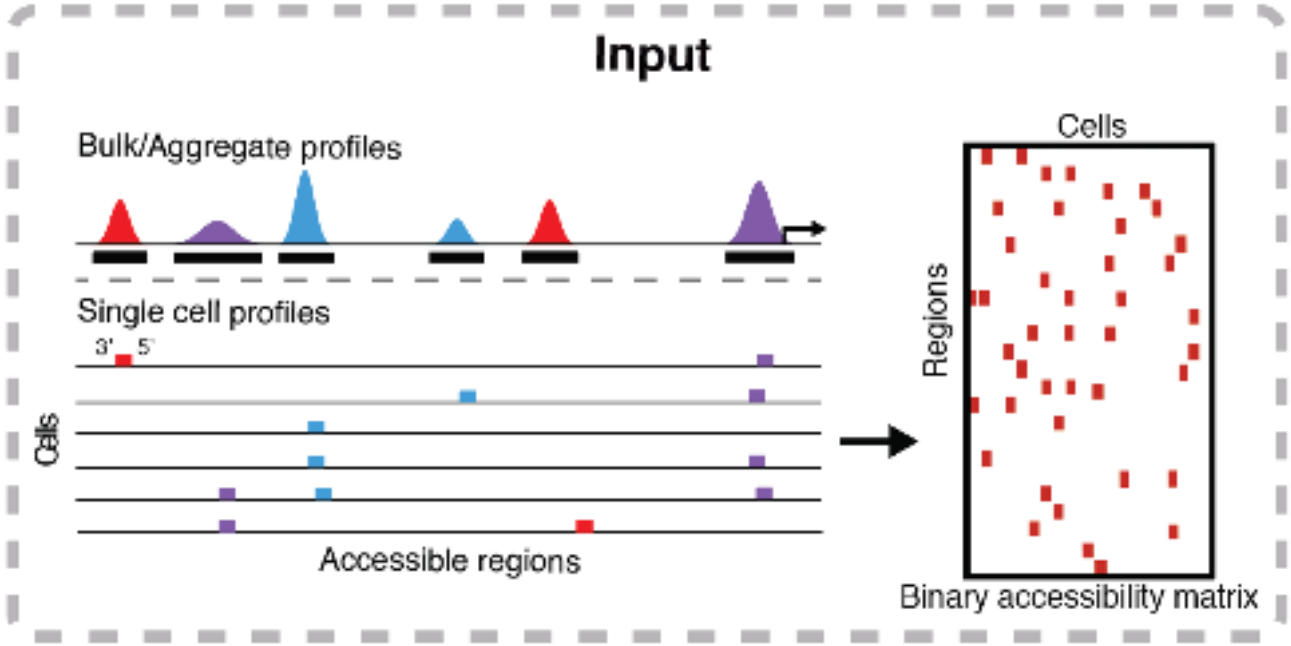
bioRχiv

Cis-topic modelling of single cell epigenomes

Carmen Bravo González-Blas, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, Stein Aerts

Posted July 16, 2018.

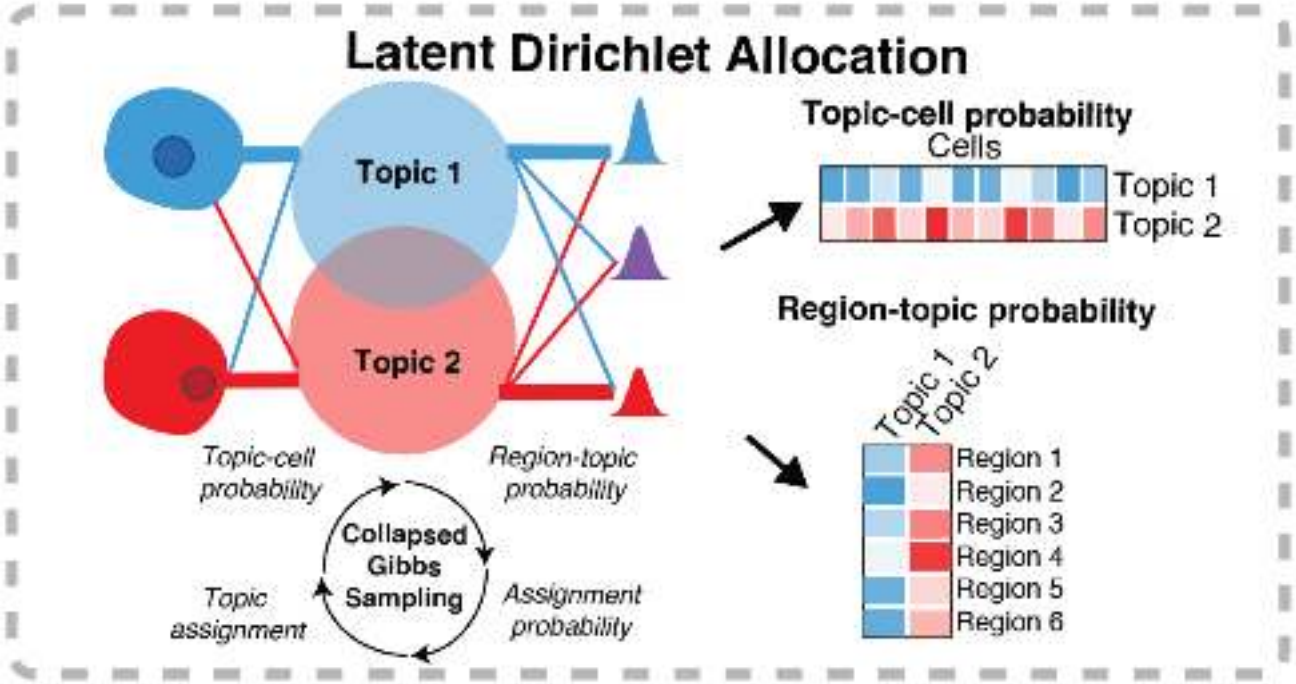
cisTopic: A new framework for single cell epigenomics data analysis



Input

- Single cell BAM files
- Regulatory regions (i.e. aggregate, bulk)

cisTopic: A new framework for single cell epigenomics data analysis



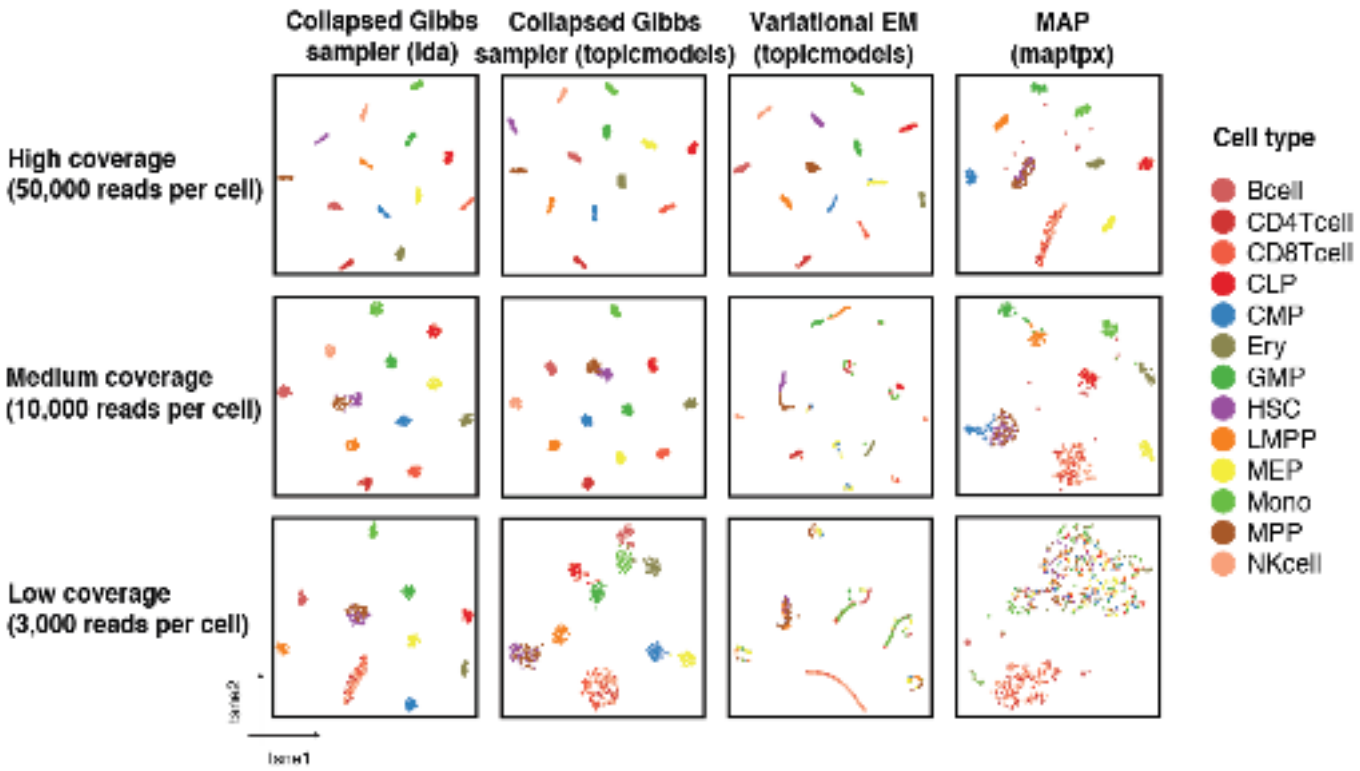
Parameter estimation with a Collapsed Gibbs sampler!

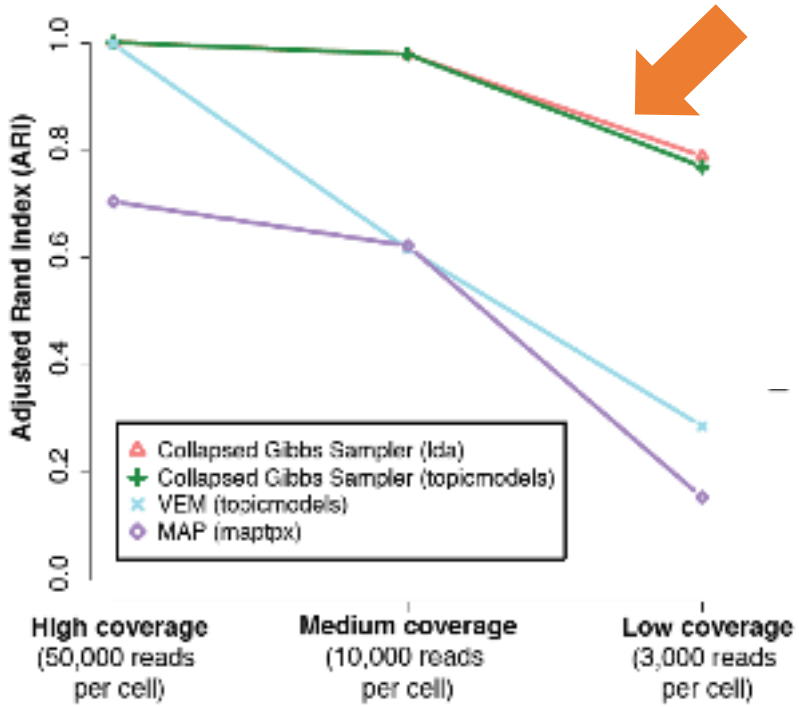
How much a region likes a topic

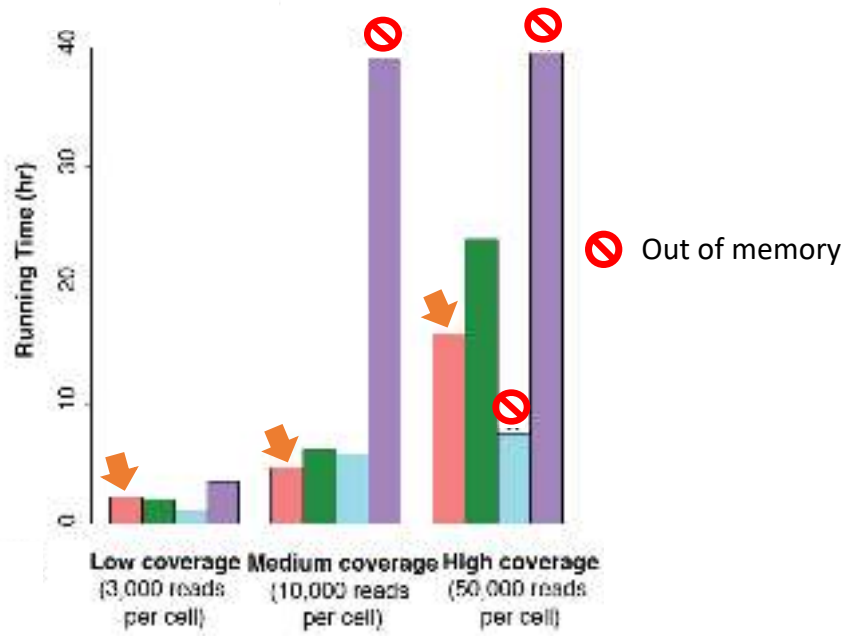
$$P(z_i|z_{-i},r) \propto \frac{n^{(r)}_{\bar{t},-i}+\beta}{n_{\bar{t},-i}+R\beta}\frac{n^{(\bar{c})}_{-i,t}+\alpha}{n^{(\bar{c})}_{-i}+T\alpha}$$

How much a cell likes a topic

A Collapsed Gibbs sampler

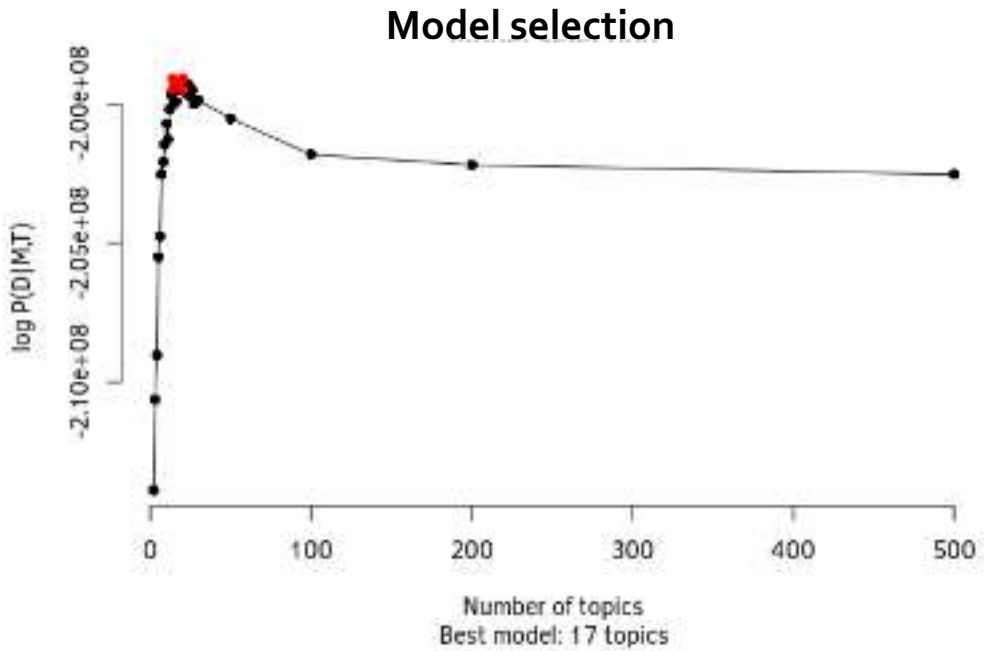




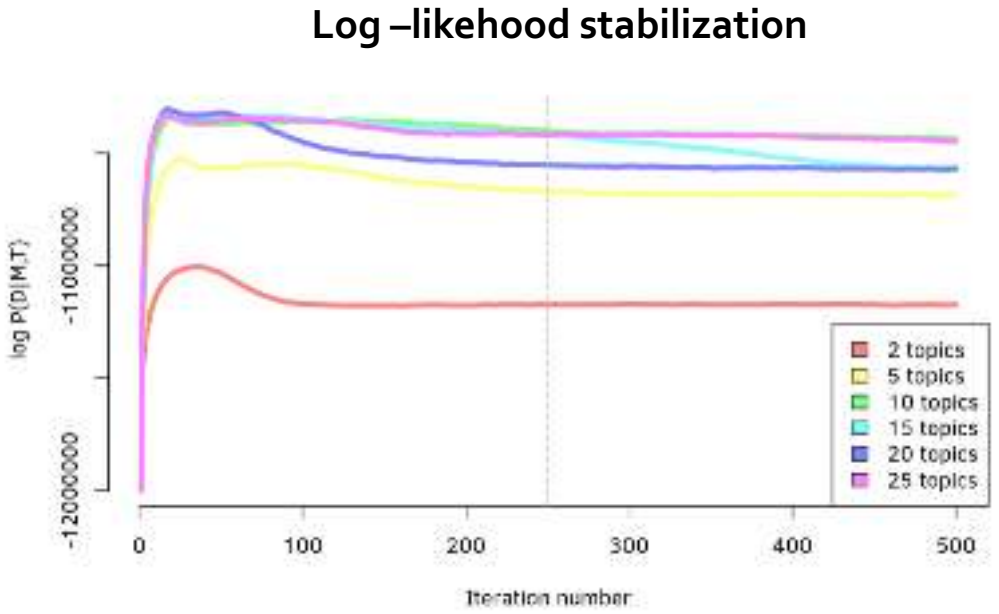


A Collapsed Gibbs sampler is the best option for parameter estimation in terms of accuracy, speed and memory efficiency.

Model selection

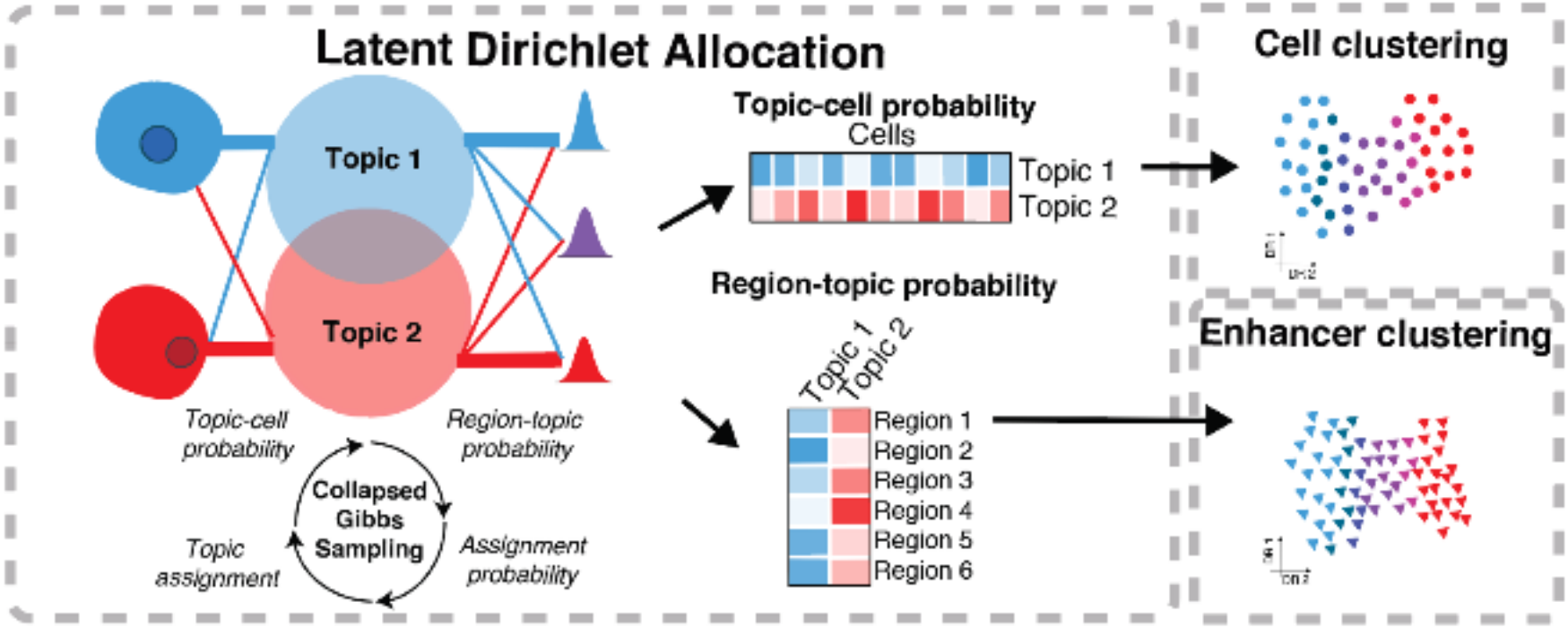


We select the model based on the **log–likehood**



Log–likehood must be stabilized before sampling (*burn-in*)

cisTopic: A new framework for single cell epigenomics data analysis



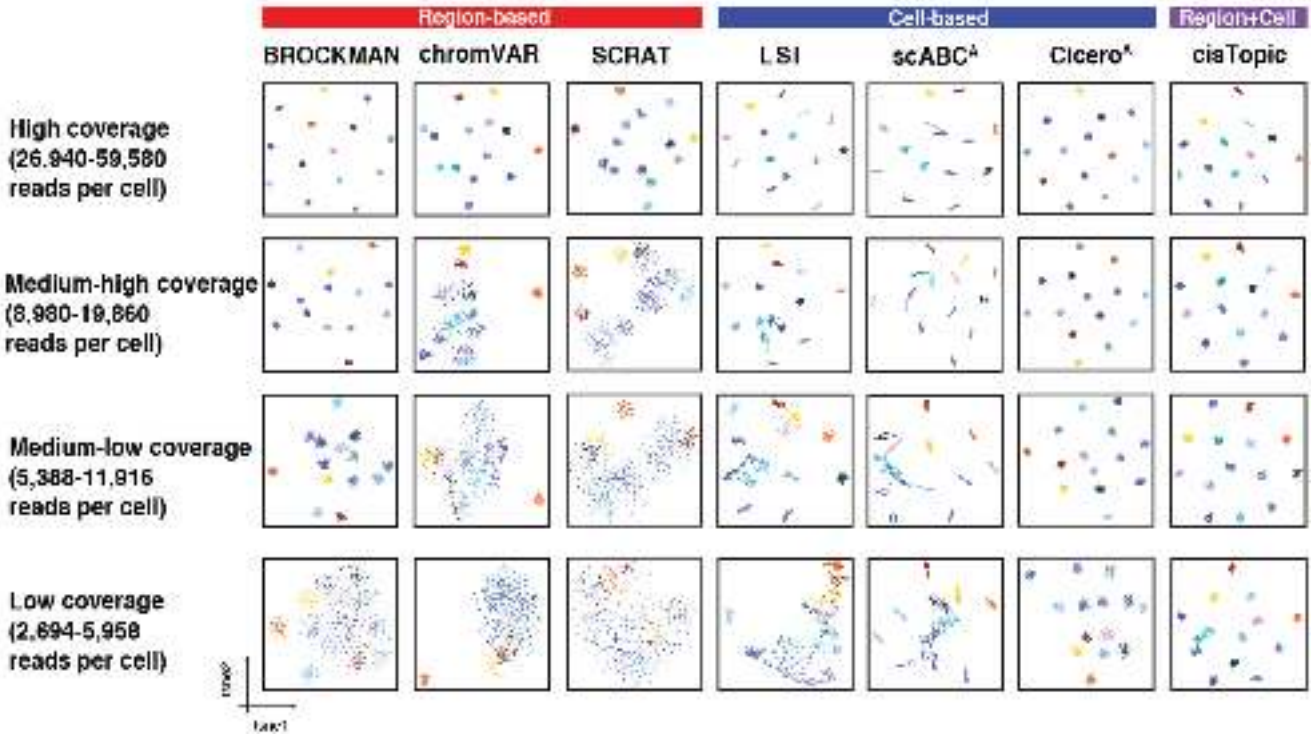
Topic-cell distributions can be used for identifying cell states

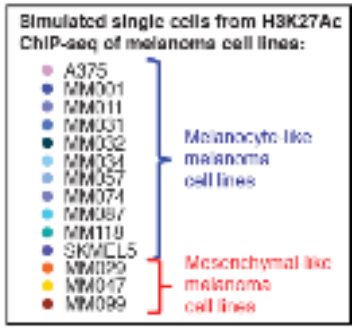
Region-topic distributions can be used to identify gene regulatory programs

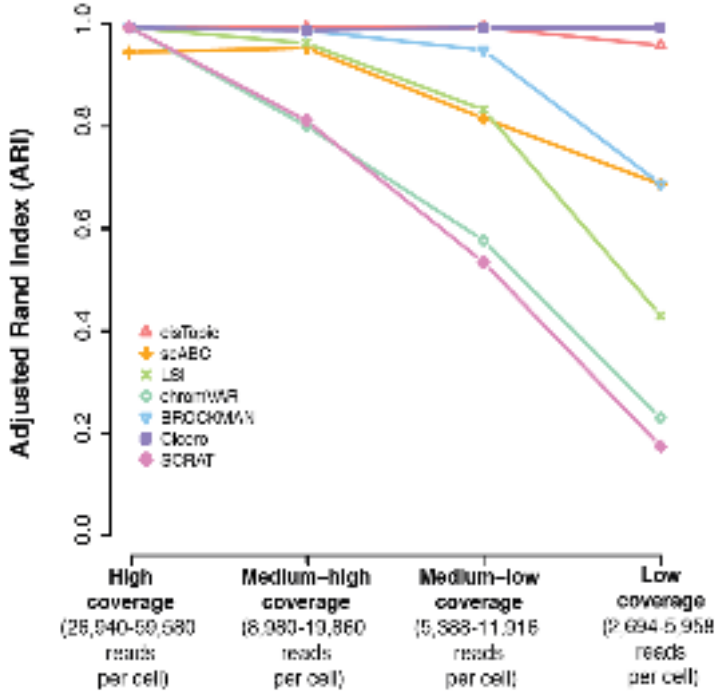
Application on data sets

Dataset		# cells	Feature
H3K27Ac melanoma (Verfaille et al.)	Human	700	✓ **Cell and region clustering** validation
ATAC-seq hematopoietic system (Buenrostro et al.)	Human	650	
scATAC-seq hematopoietic system (Buenrostro et al.)	Human	2,755	✓ Identification of **cell types** and master **GRNs** ✓ **Batch effect correction & technical bias robustness**
scnmC-seq frontal cortex	Human	2,784	✓ **Extension to other types of epigenomic data**
Human Brain (Lake et al.)	Human	34,520	✓ **De novo** discovery of **subpopulations**
Mouse Brain (Preissl et al.)	Mouse	3,034	✓ Topic-based **cross-species** comparison ✓ Validation of topics with external data
scATAC-seq SOX10 KD in melanoma	Human	598	✓ Analysis of **dynamic processes**
Mouse Cell Atlas (Cusanovich et al.)	Mouse	80,254	✓ **Scalability** to larger datasets

Validation on simulated data

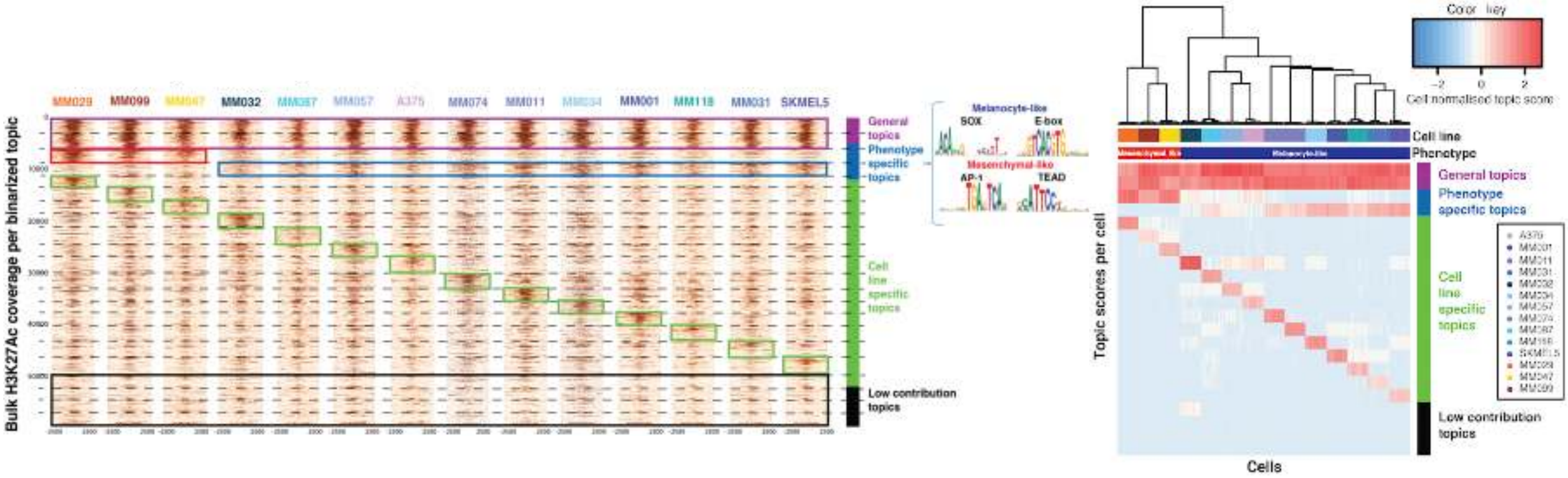






Melanoma cultures H3K27Ac data: Verfaillie, Kalender Atak, Imrichova et al., Nat Comms 2015

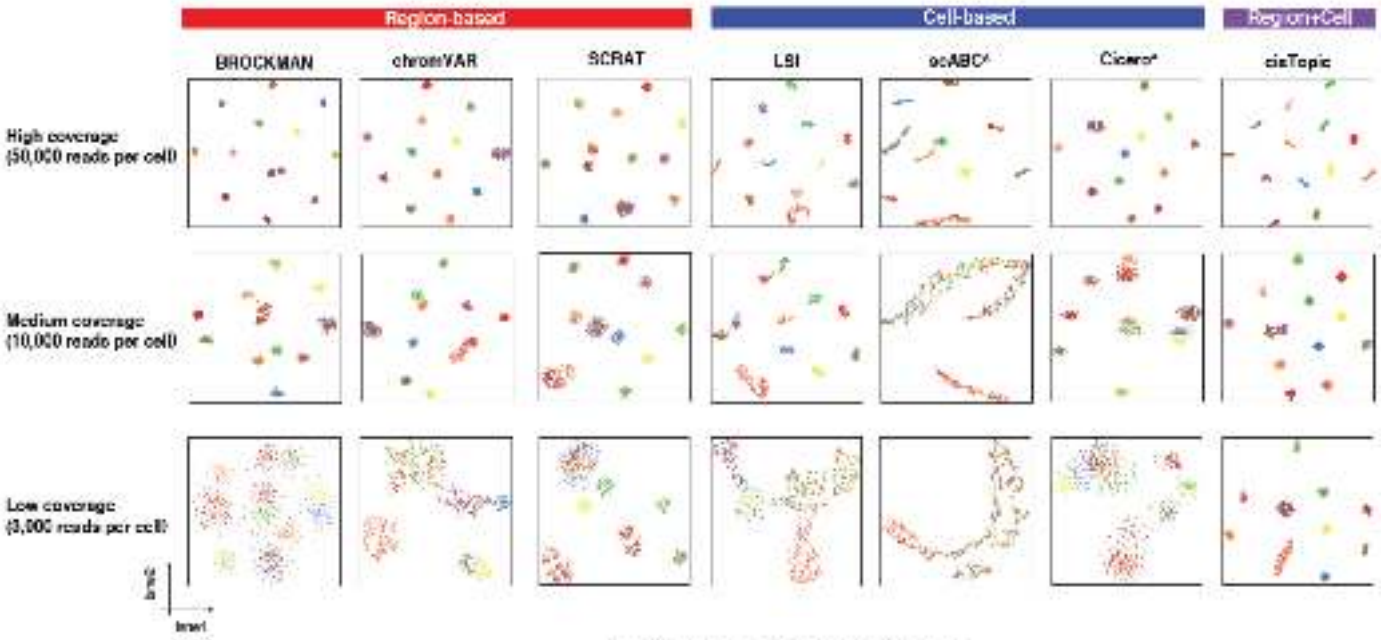
Validation simulated data

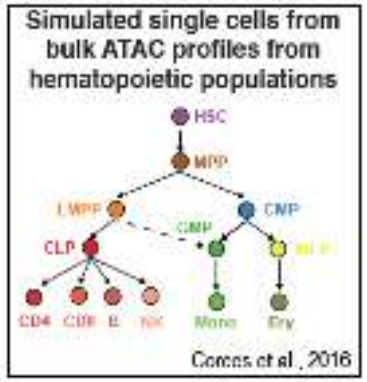


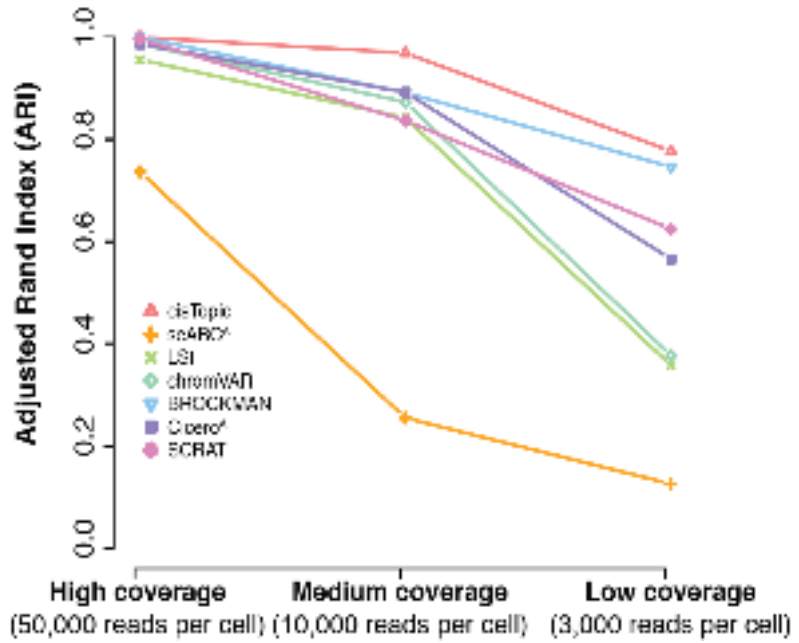
H3K27Ac data: Verfaillie, Kalender Atak, Imrichova et al., Nat Comms 2015

cisTopic also reveals biologically meaningful cell-type specific programs

Validation on simulated data

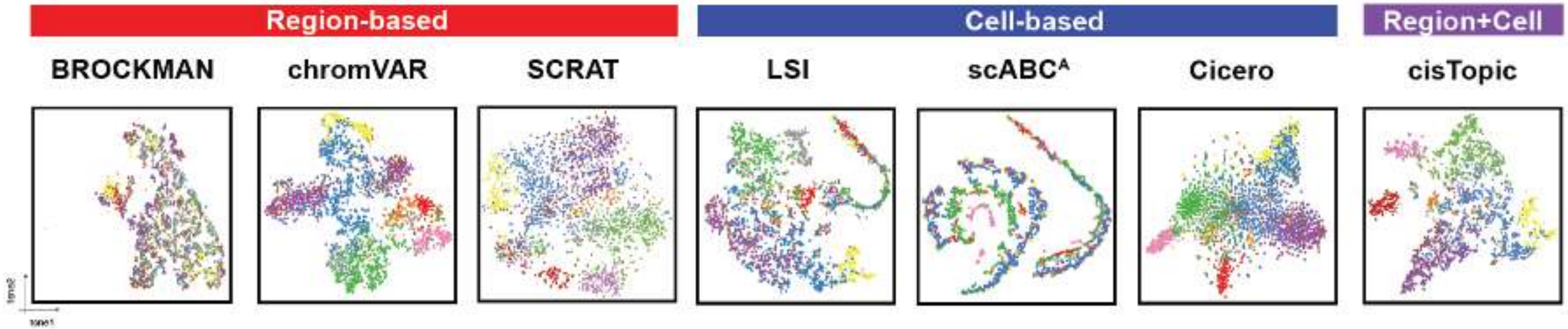




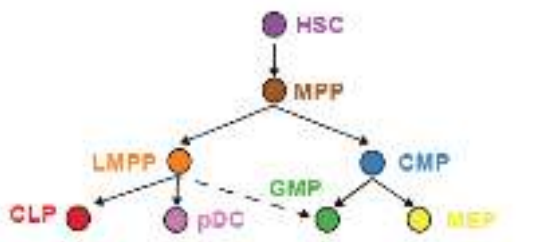


Hematopoietic system FACS bulk ATAC data: Corces et al., Nat Genetics 2016

cisTopic reconstructs dynamic trajectories in the hematopoietic system

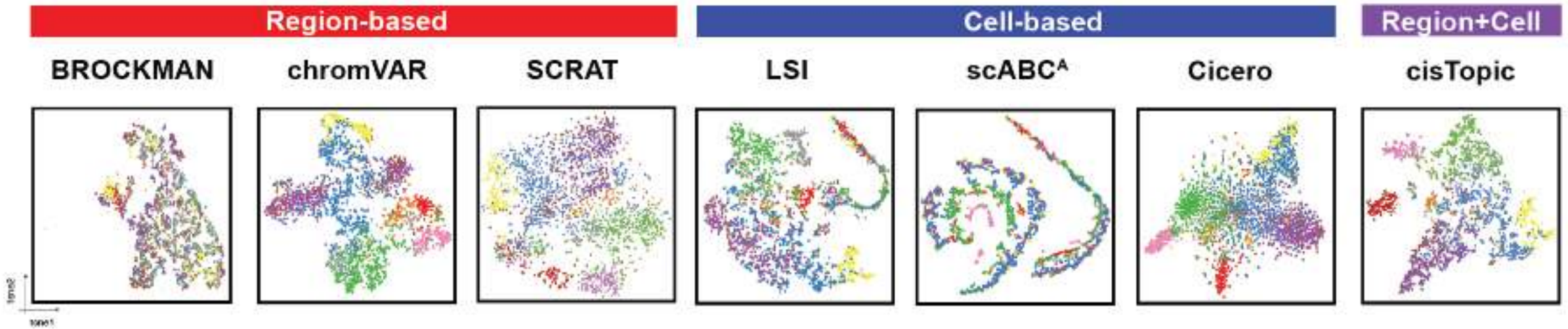


Single cell ATAC-seq in FAC-sorted hematopoietic populations

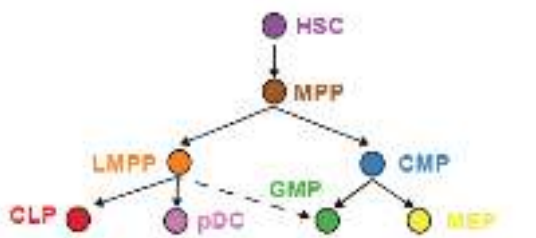


Buenrostro et al., 2018

cisTopic reconstructs dynamic trajectories in the hematopoietic system

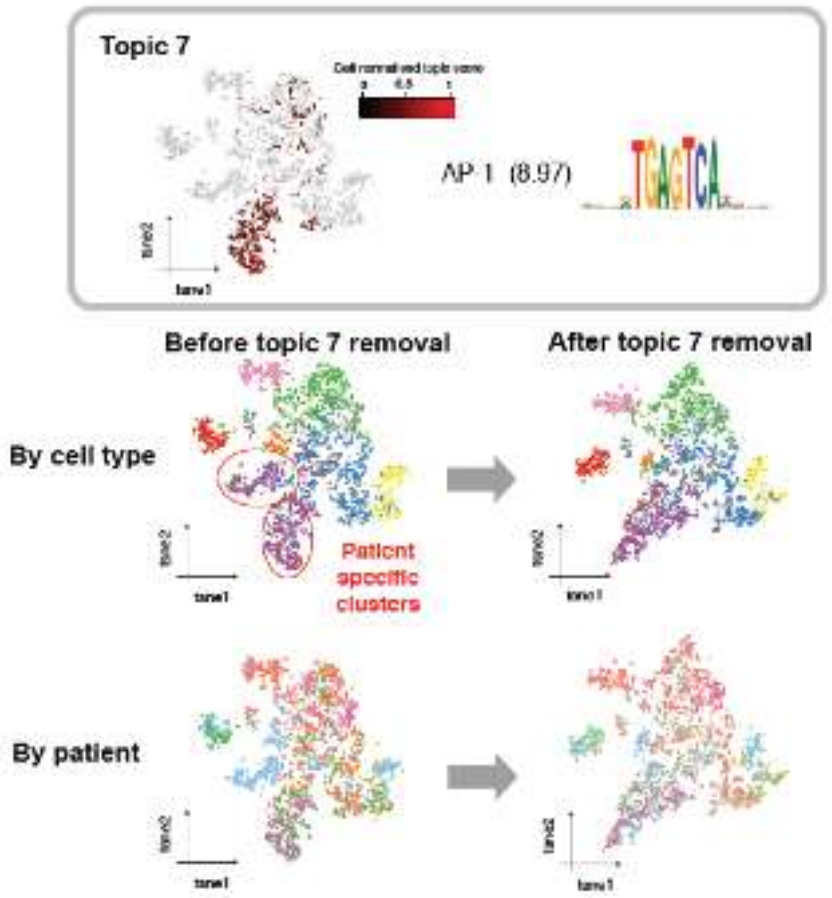


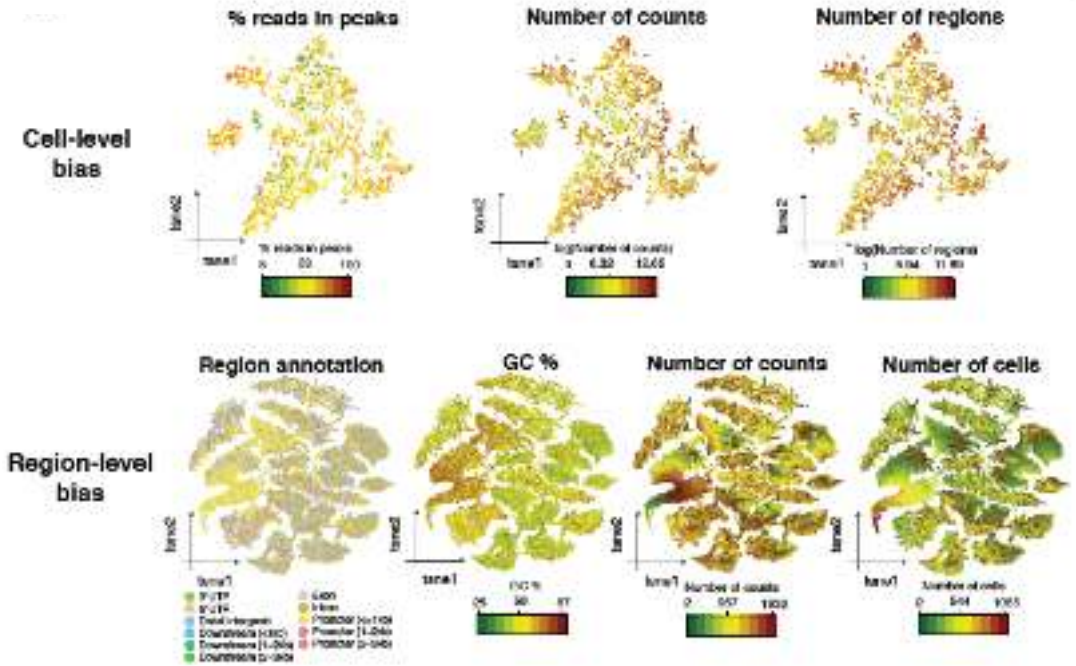
Single cell ATAC-seq in FAC-sorted hematopoietic populations



Buenrostro et al., 2018

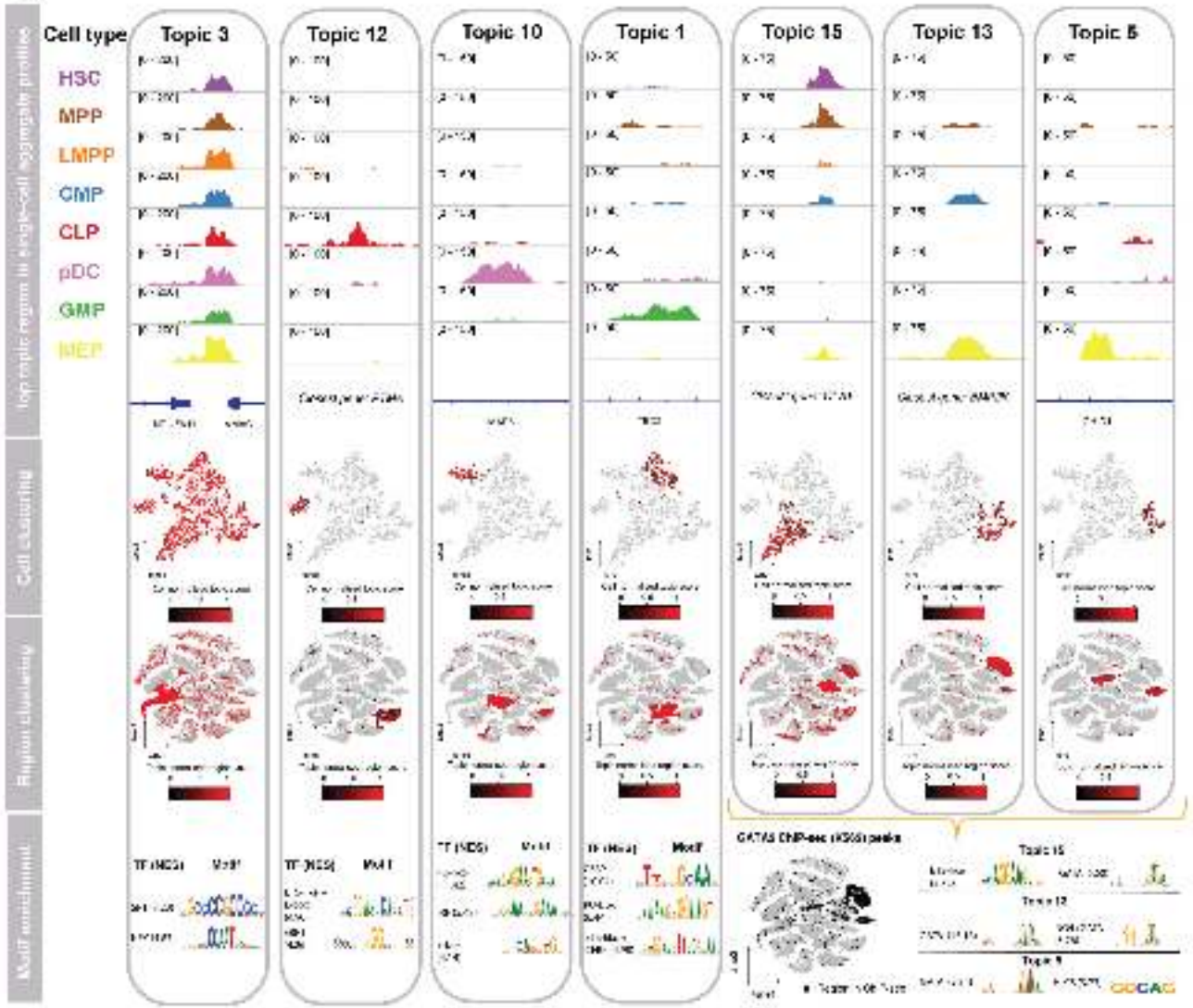
cisTopic is robust to batch effect and technical bias

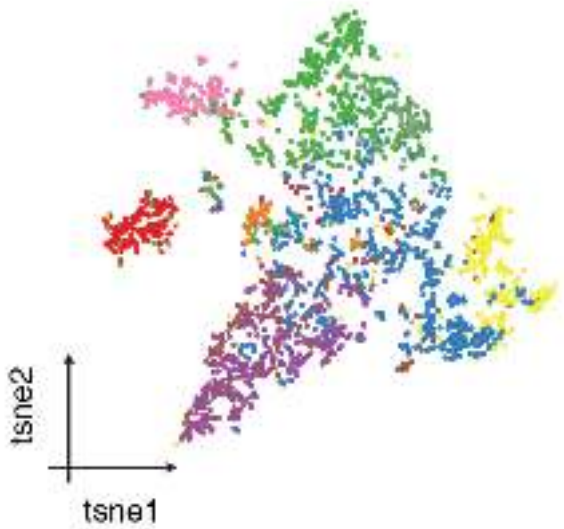




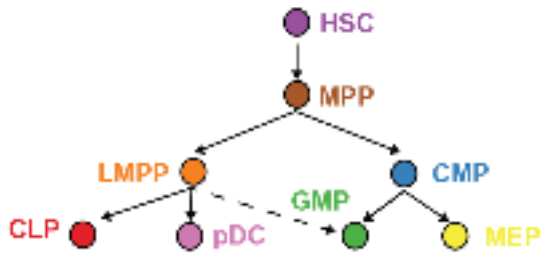
Hematopoietic system scATAC data: Buenrrostro et al., Cell 2018

cisTopic reveals dynamic GRNs in the hematopoietic system



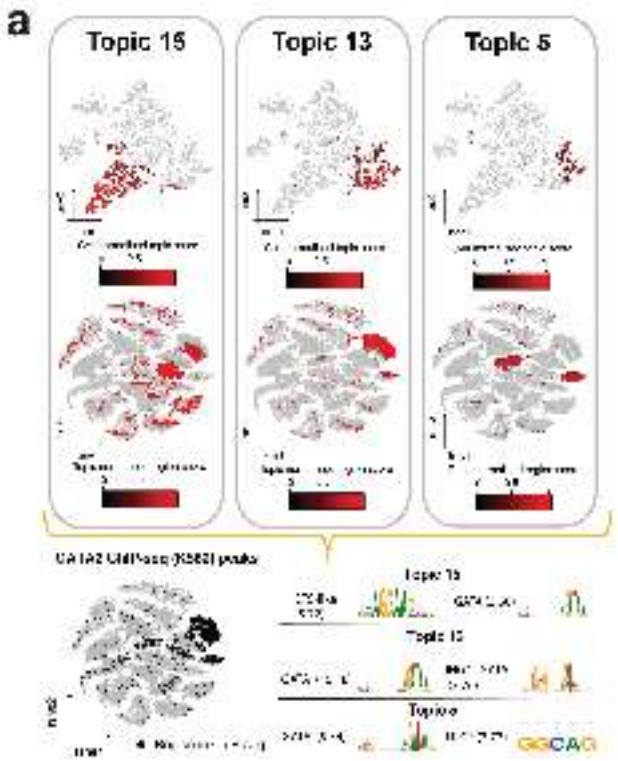


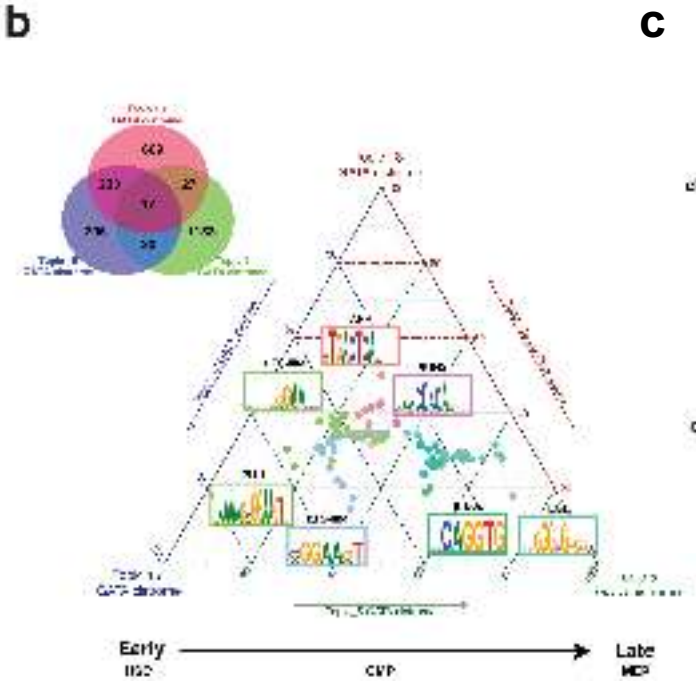
Single cell ATAC-seq in FAC-sorted hematopoeitic populations

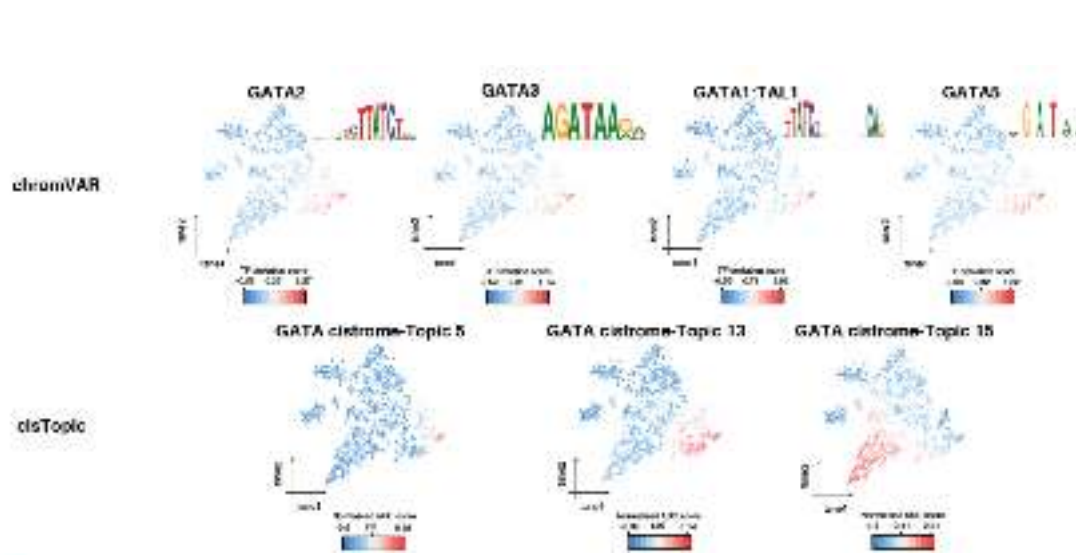


Hematopoietic system scATAC data: Buenrrostro et al., Cell 2018

cisTopic reveals dynamic GRNs in the hematopoietic system

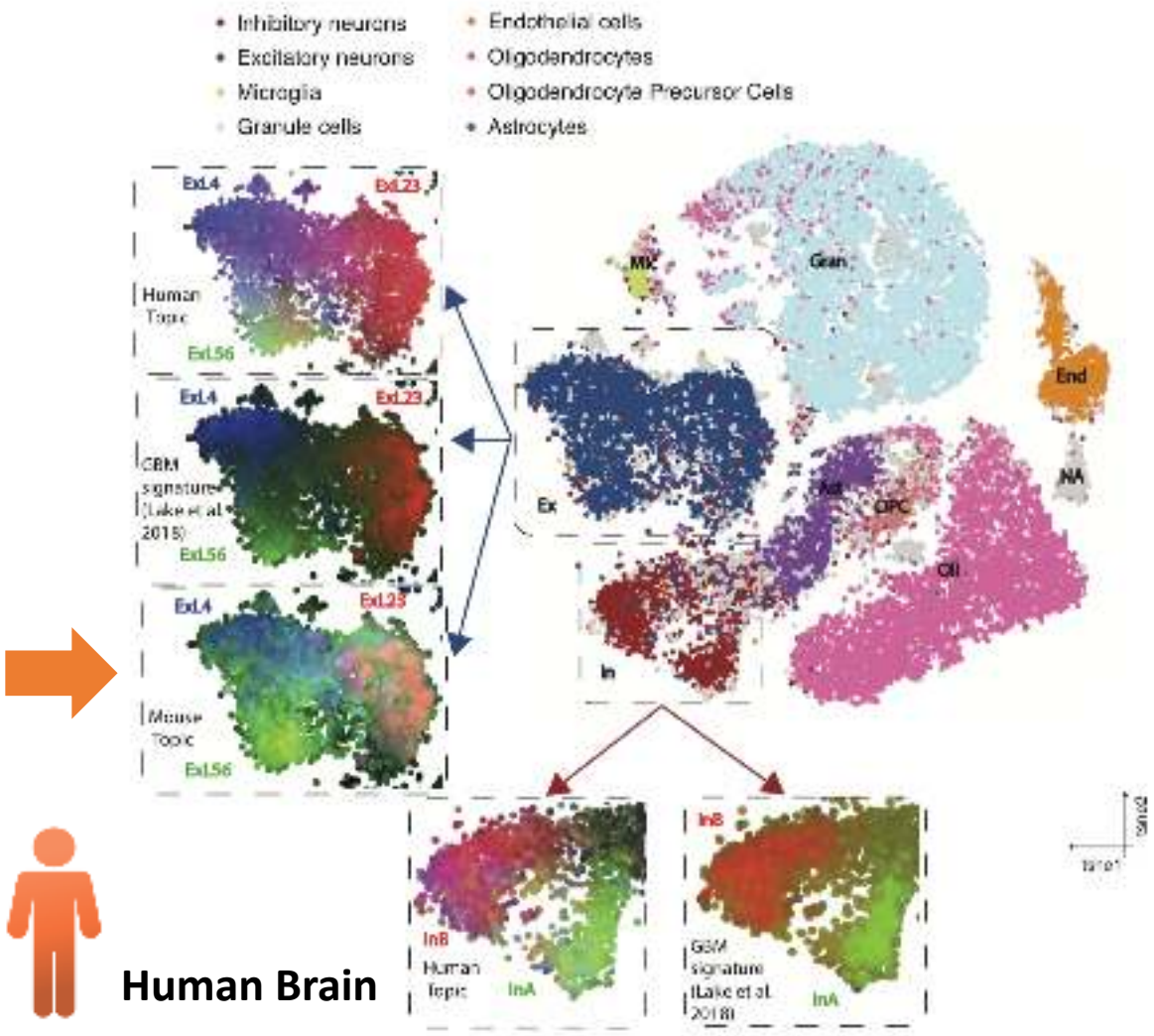






Hematopoietic system scATAC data: Buenrrostro et al., Cell 2018

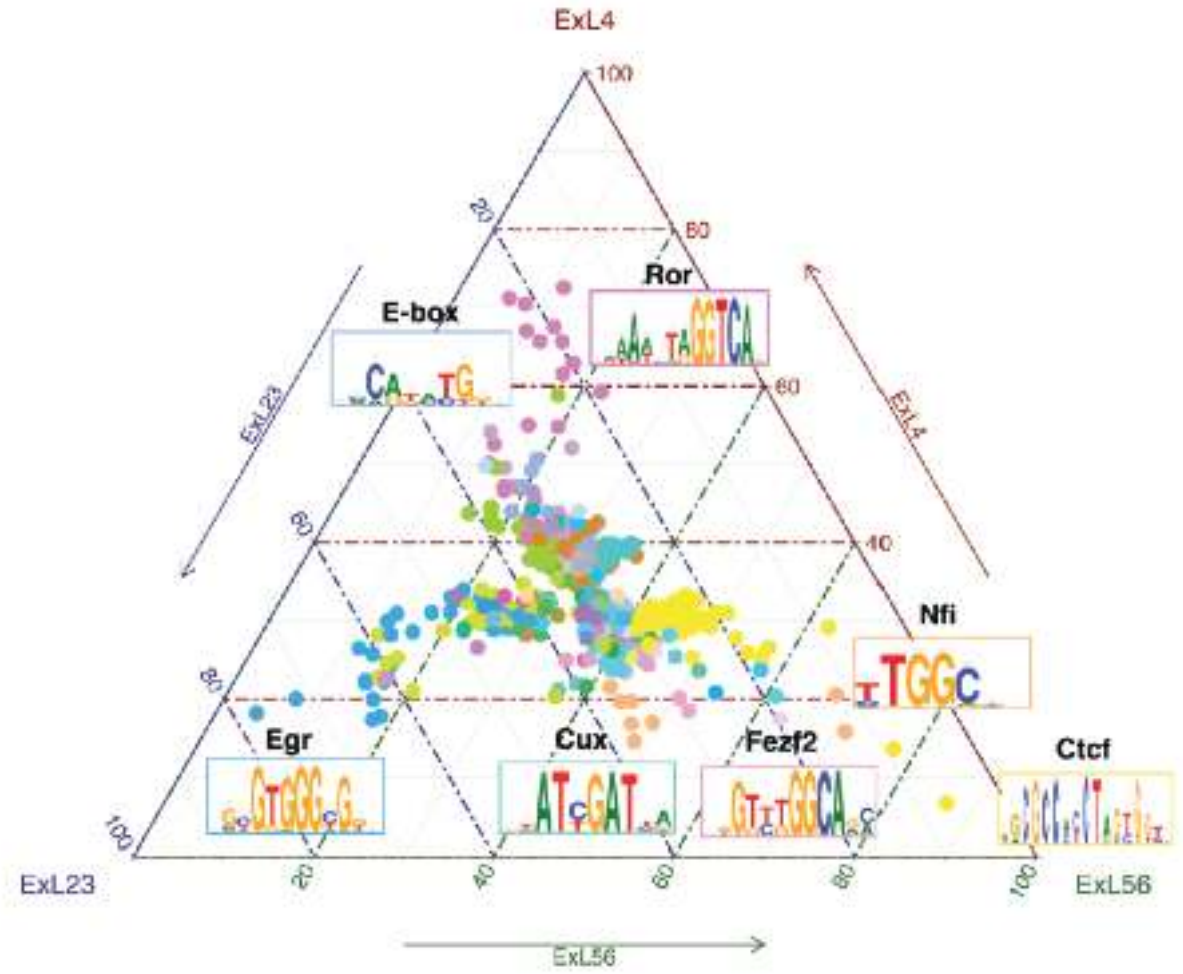
cisTopic reveals cell state conservation between mammalian brains

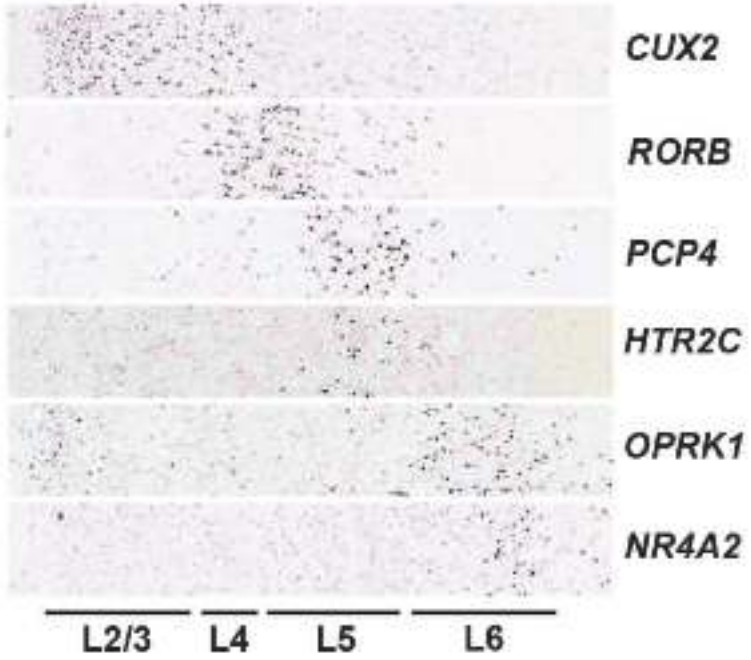


Human Brain

Human brain scATAC data: Lake et al., Nat Biotech 2017

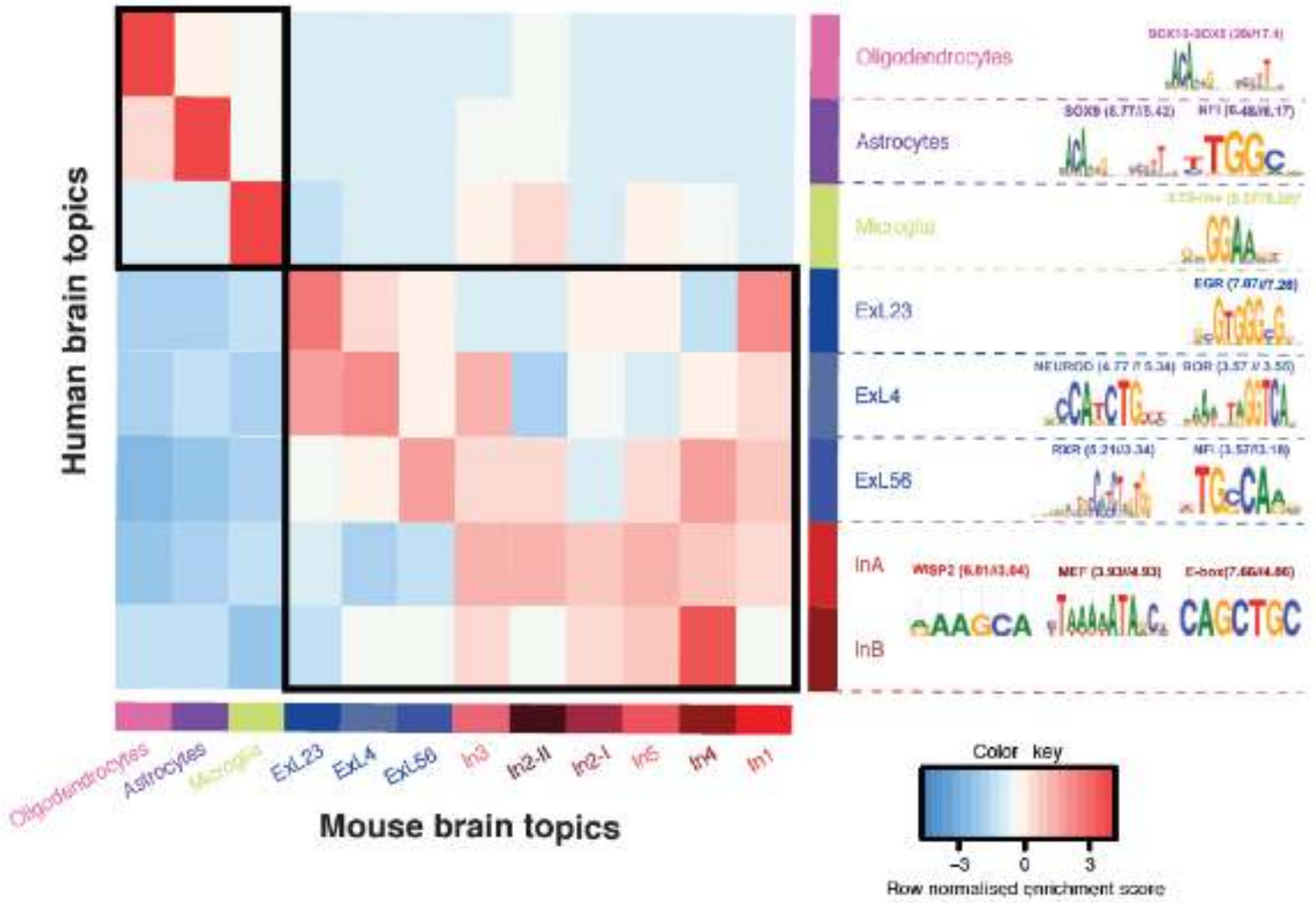
cisTopic reveals layer-specific GRNs





Lake et al., Science 2016

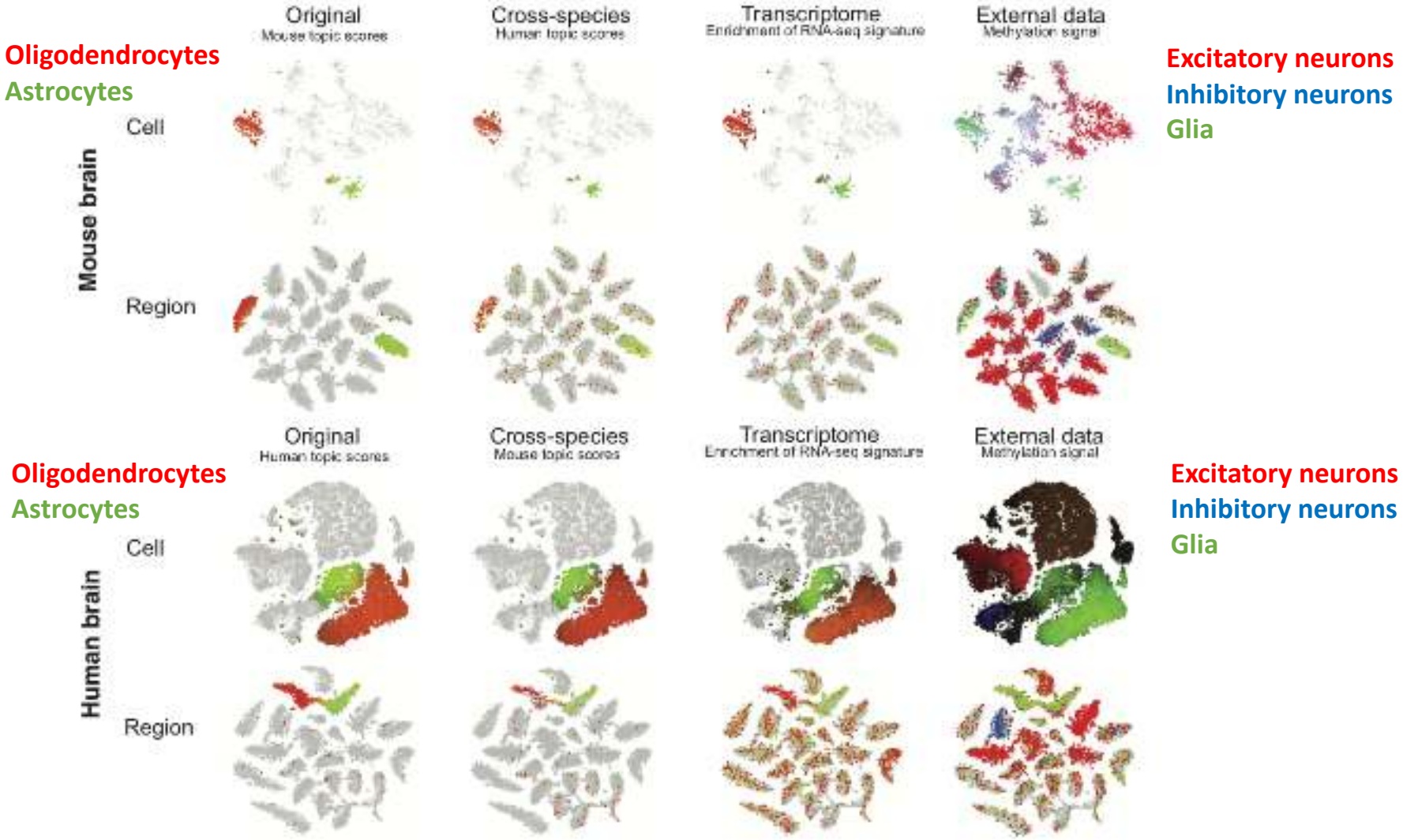
cisTopic reveals cell state conservation between mammalian brains



Human brain scATAC data: Lake et al., Nat Biotech 2017

Mouse brain scATAC data: Preissl et al., Nat Neurosci 2018

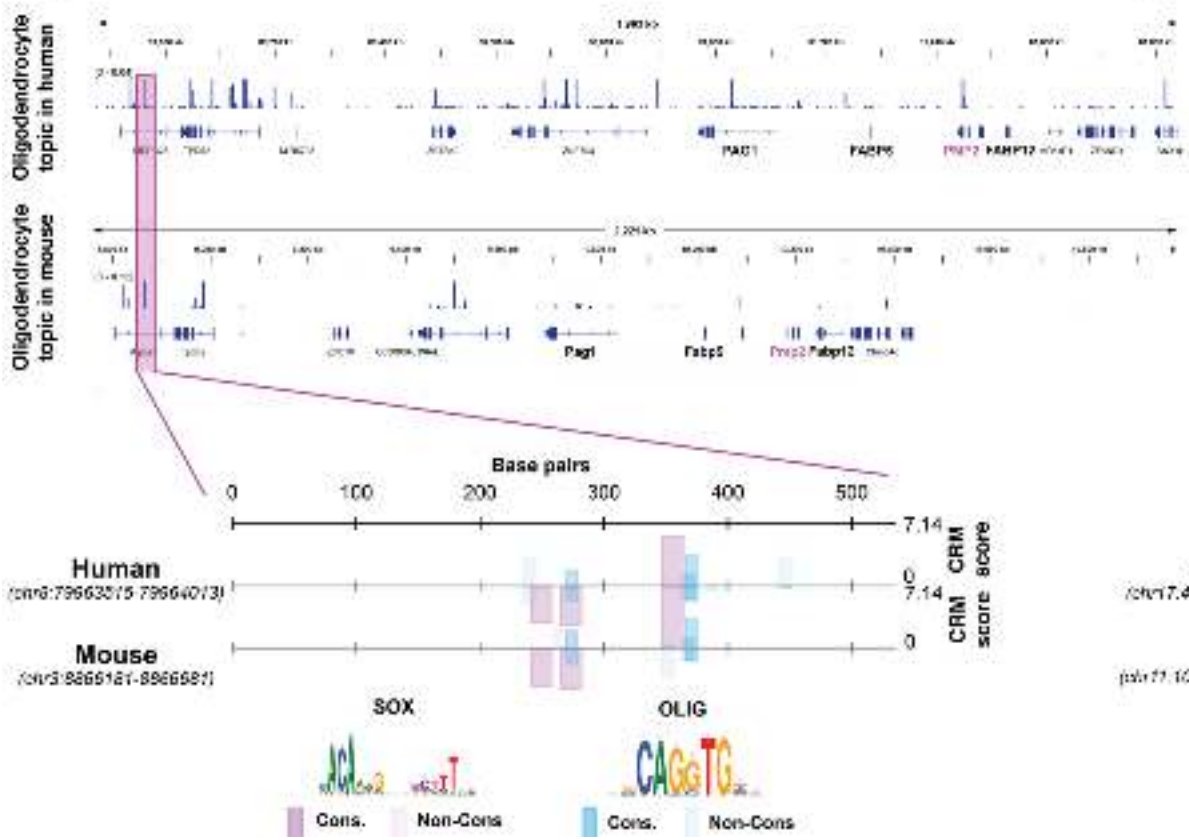
cisTopic reveals cell state conservation between mammalian brains

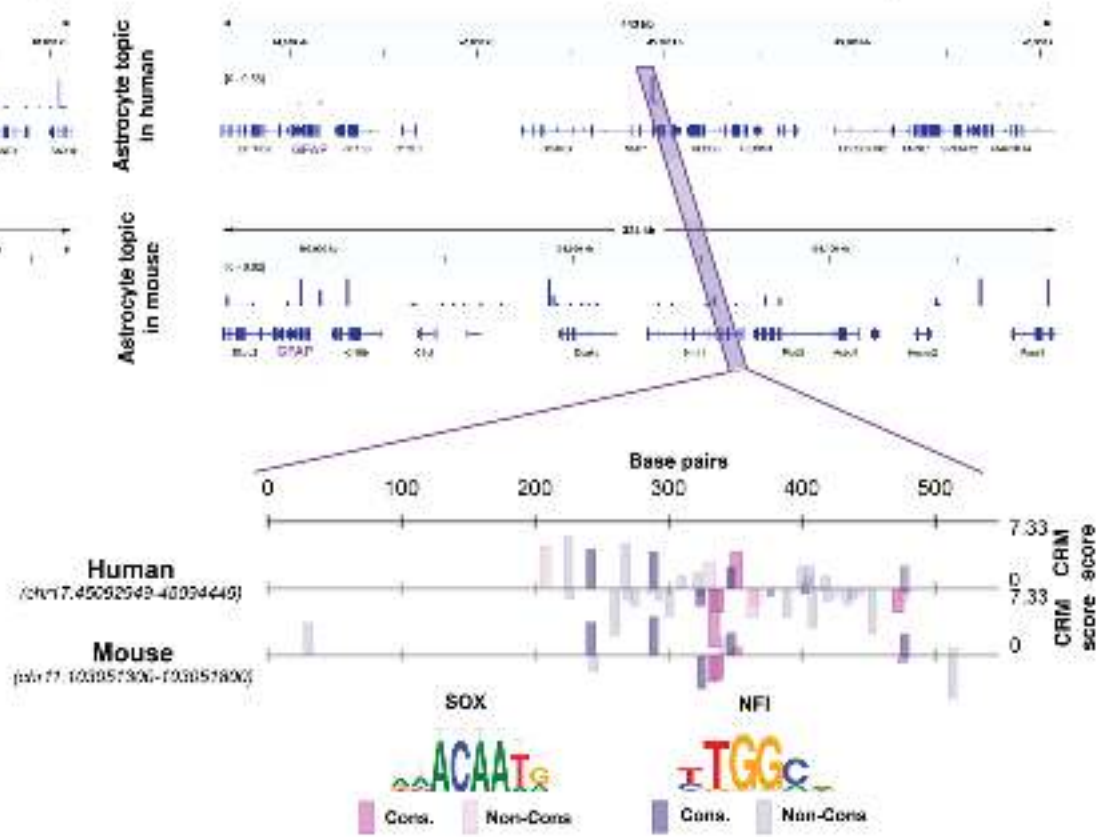


Human brain scATAC data: Lake et al., Nat Biotech 2017

Mouse brain scATAC data: Preissl et al., Nat Neurosci 2018

cisTopic reveals cell state conservation between mammalian brains

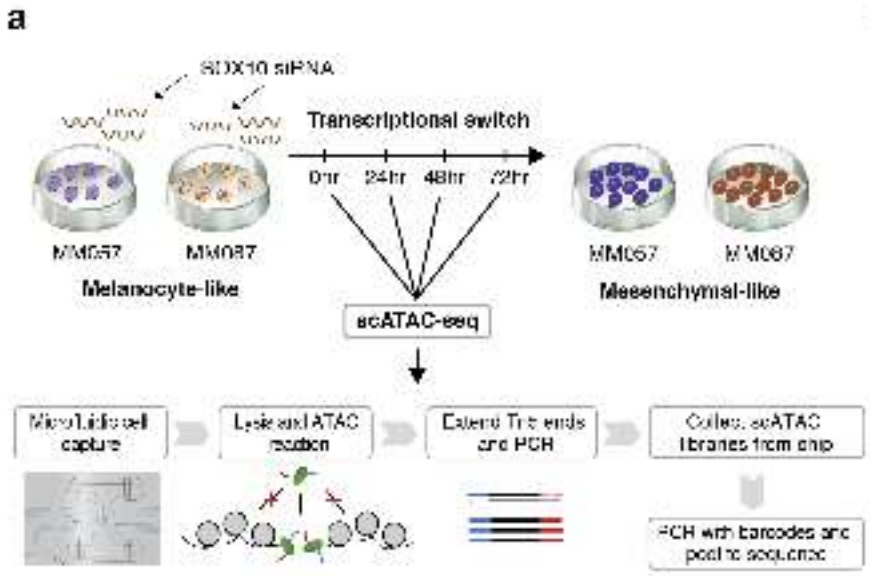




Human brain scATAC data: Lake et al., Nat Biotech 2017

Mouse brain scATAC data: Preissl et al., Nat Neurosci 2018

cisTopic maps a dynamic regulatory landscape during EMT-like transition in melanoma

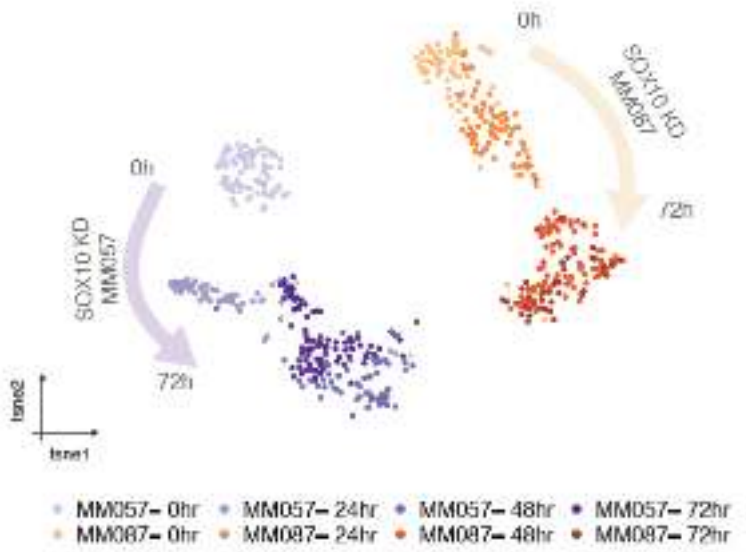


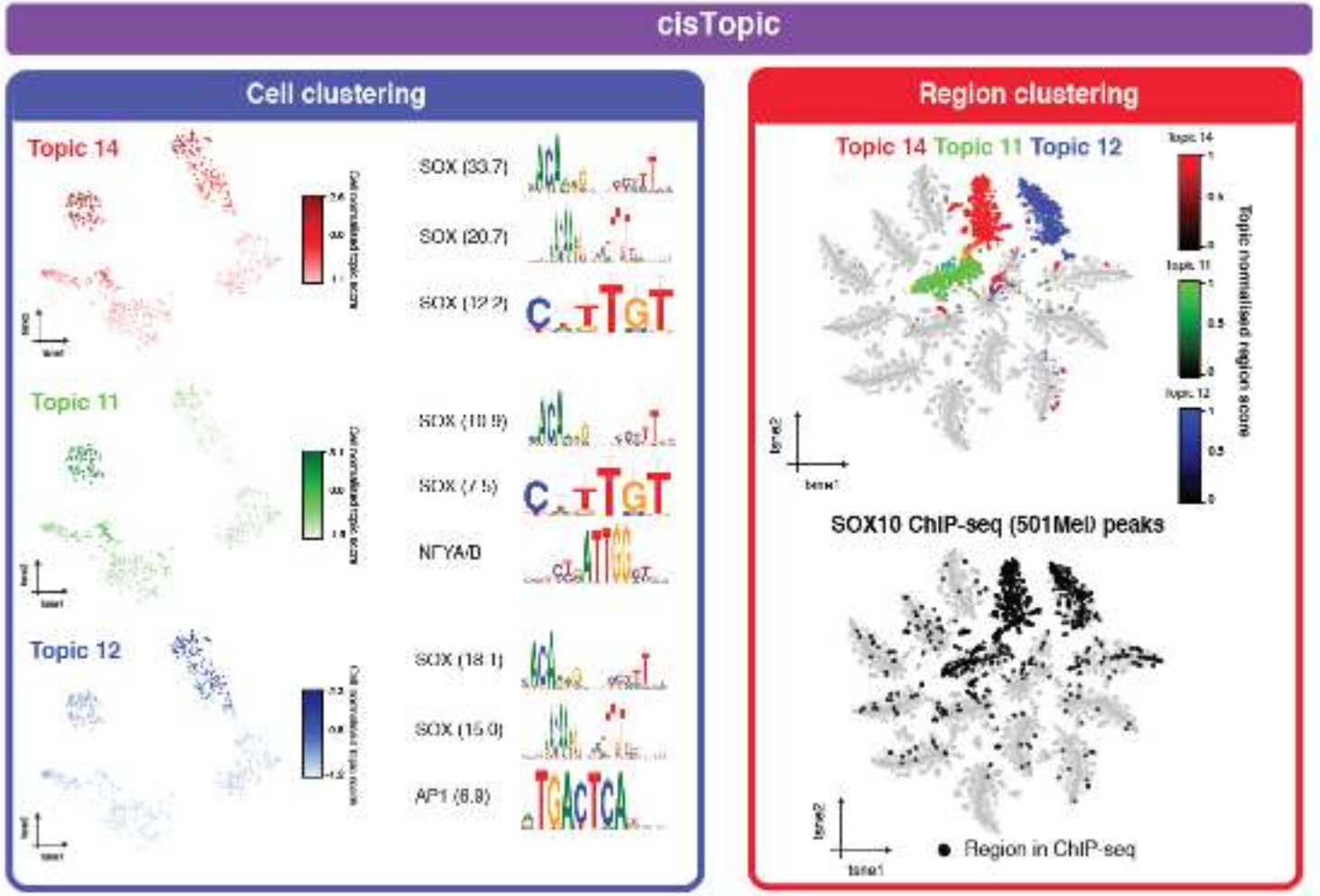
598 cells; 78262 regulatory regions

Melanoma SOX10KD scATAC data: Liesbeth Minnoye

Liesbeth

cisTopic maps a dynamic regulatory landscape during EMT-like transition in melanoma

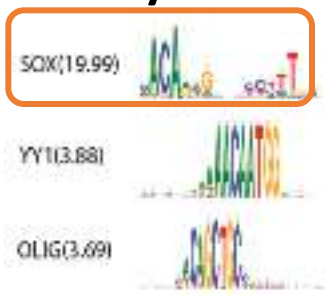




Melanoma SOX10KD scATAC data: Liesbeth Minnoye

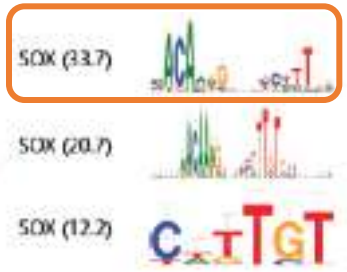
Finding SOXE cofactors

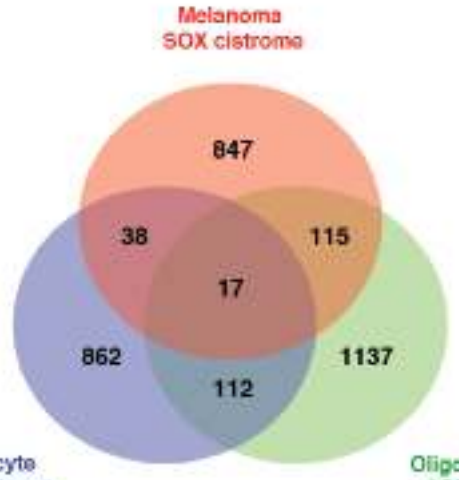
Oligodendrocytes

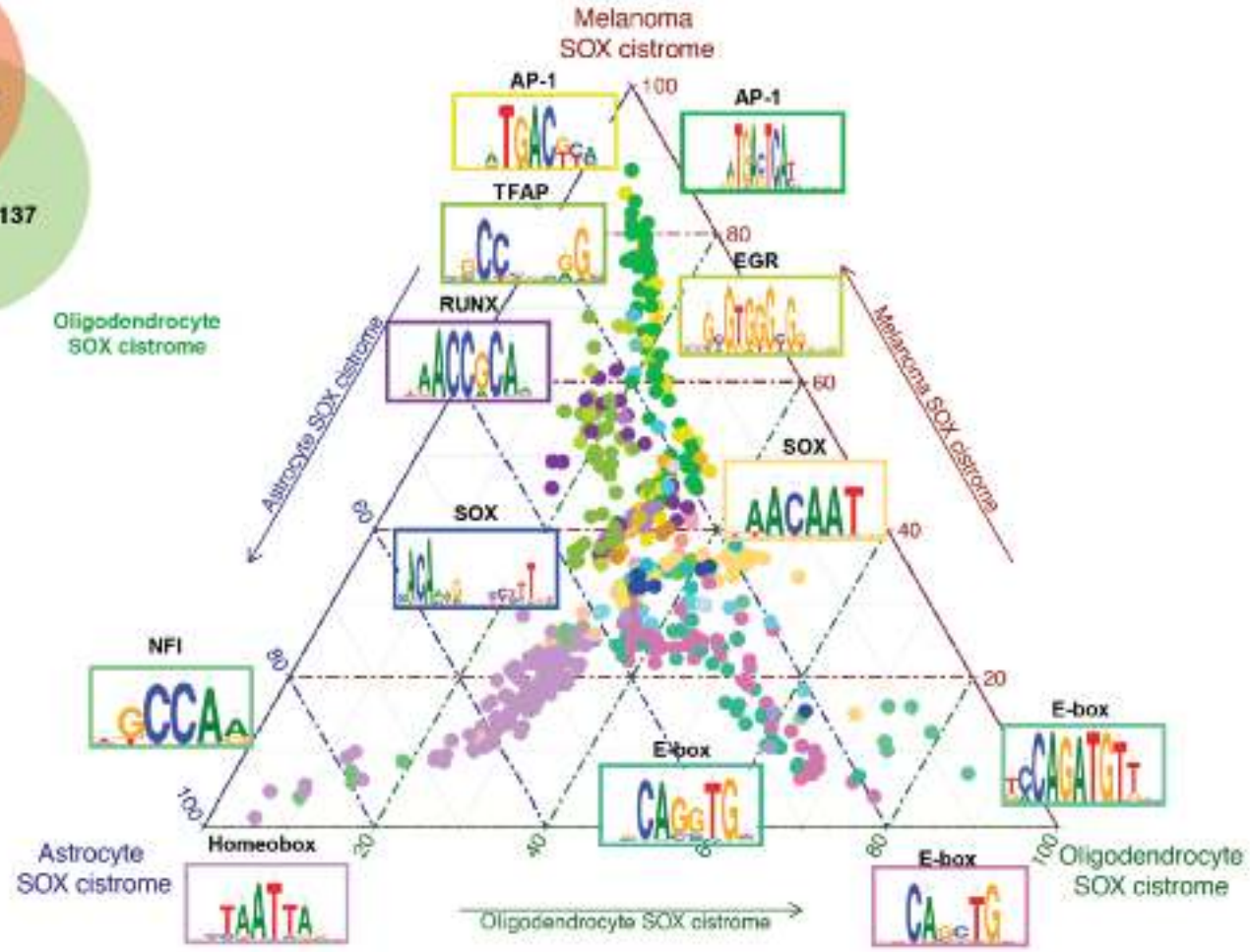


Astrocytes

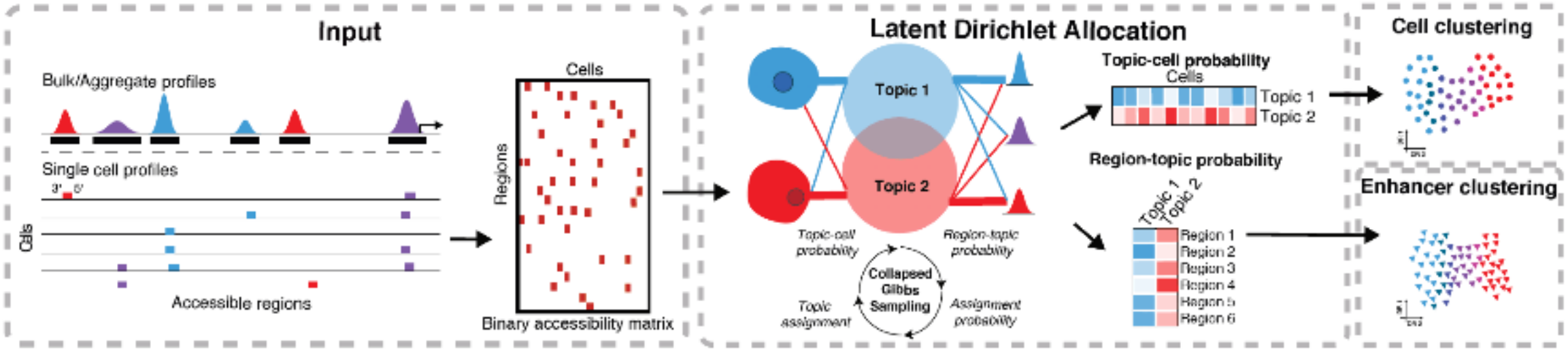
Melanoma







To take home…



- **scATAC-seq** can reveal regulatory programs in heterogeneous tissues and dynamic processes

- **cisTopic:** Probabilistic modelling of cis-regulatory programs from single cell epigenomics data

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Currently used in…

SCIENTIFIC REPORTS

Chromatin accessibility identifies diversity in mesenchymal stem cells from different tissue origins

Yen-Ting Ho[1], Takashi Shimbo[1], Edward Wijaya[1,2], Yuya Ouchi[1,2], Eiichi Takaki[1,2], Ryoma Yamamoto[1,2], Yasushi Kikuchi[1], Yasufumi Kaneda[3] & Katsuto Tamai[1]

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

Capturing cell type-specific chromatin structural patterns by applying topic modeling to single-cell Hi-C data

Hyeon-Jin Kim[1], Galip Gürkan Yardımcı[1], Giancarlo Bonora[1], Vijay Ramani[1,2], Jie Liu[1], Ruolan Qiu[1], Choli Lee[1], Jennifer Hesson[3,5], Carol B. Ware[3,5], Jay Shendure[1], Zhijun Duan*[4,5], and William Stafford Noble*[1,6]

EVENT

DSCB and C3BI Departments Seminar by Professor Stein Aerts, Department of Human Genetics, Laboratory of Computational Biology, KU Leuven

Location

Building: Metchnikoff (n 67)
Room: Salle Jules Bordet
Address: 25 Rue du Dr Roux, Paris, France

15 Feb 2019

11:00:00

Friday, February 15th

11:00 am – Jules Bordet meeting room, Metchnikoff building

"Deciphering gene expression programs at single-cell resolution"

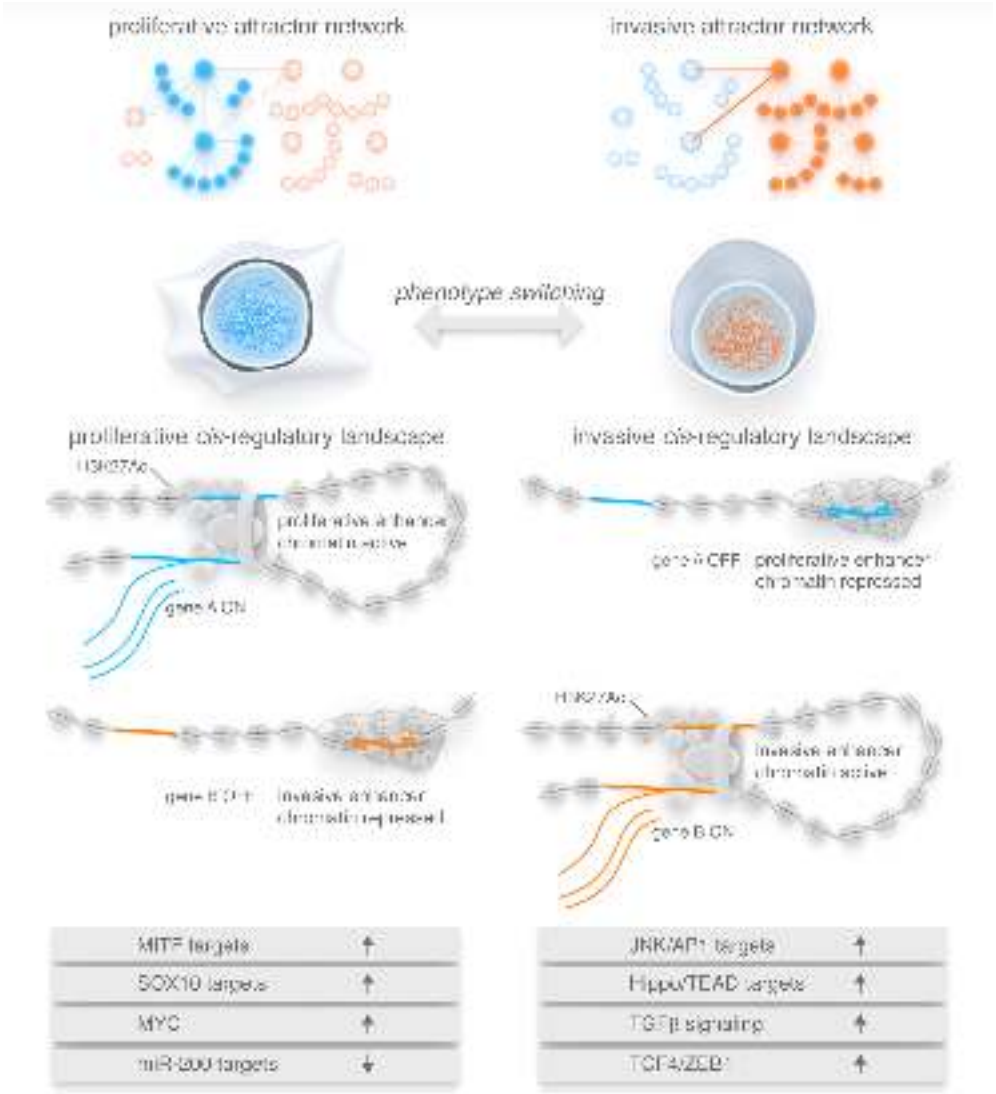
SUMMARY: Single-cell technologies are revolutionising biology and provide new opportunities to trace genomic regulatory programs underlying cell fate.

In this talk I will present several computational strategies for the analysis of single-cell RNA-seq and single-cell ATAC-seq data that exploit the genomic regulatory code, to guide the identification of transcription factors and cell states. I will illustrate these methods on several model systems, including the Drosophila brain. Finally I will discuss how single-cell analyses can contribute to cross-species comparisons of regulatory programs.

Hands-on

Exercise: cisTopic on simulated epigenomes

- Simulated single-cells from bulk H3K27Ac melanoma profiles
- 3 proliferative/melanocyte-like lines (MM001, MM011 and MM034) and 2 invasive lines (MM029 and MM047)
- Proliferative:: SOX10, MITF, TFAP2 // Invasive: AP-1
- 20 cells per simulated per profile
- 100 cells and ~112,000 accessible regions



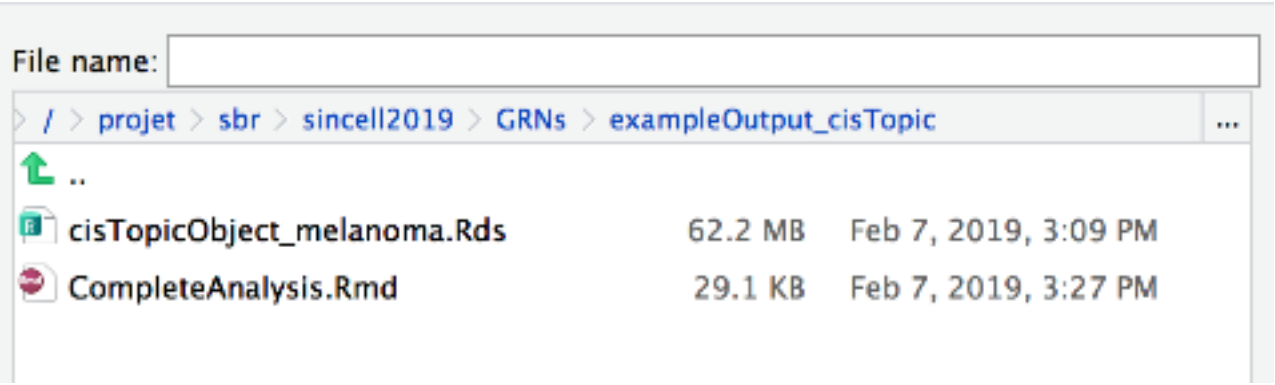
Our plan

1. Open rendered vignette and go all together over the code/outputs (Github)

Tutorial

You can find the rendered vignette here.

2. In the remaining time, explore tutorials (Course scripts and files in: /projet/sbr/sincell2019/GRNs)



Load data

```{r}
cisTopicObject <- readRDS('/projet/sbr/sincell2019/GRNs/exampleOutput_cisTopic/cisTopicObject_melanoma.Rds')
```

Initializing cisTopicObject (not run)

From BAM file

```{r, eval=FALSE}
pathToBams <- 'data/bamfiles/'
bamFiles <- paste(pathToBams, list.files(pathToBams), sep='')
regions <- 'data/regions.bed'
cisTopicObject <- createcisTopicObjectFromBAM(bamFiles, regions, project.name='scH3K27Ac_melanoma')

# If you want to rename cells
cell.names <- cisTopicObject@cell.names
new.cell.names <- sapply(strsplit(cell.names, split = ".", fixed=TRUE), "[", 3)
cisTopicObject <- renameCells(cisTopicObject, new.cell.names)
```

Single end: Count read if 5' falls within peak
Paired end: Count read if 5' or 3' fall within peak

From count matrix

```{r, eval=FALSE}
data(counts_mel)
cisTopicObject <- createcisTopicObject(counts_mel, project.name='scH3K27Ac_melanoma')
rm(counts_mel)
```

Add cell data

```{r, eval=FALSE}
data(cellData_mel)
cisTopicObject <- addCellMetadata(cisTopicObject, cell.data = cellData_mel)
rm(cellData_mel)
```

Build models (not run)

```r
```{r, eval=FALSE}
cisTopicObject <- runModels(cisTopicObject, topic=c(2, 5:15, 20, 25), seed=987, nCores=13, burnin = 120, iterations = 150, addModels=FALSE)
```
```

- **Number of topics** (`topic`): The number of topics are usually slightly bigger than the potential cell states in the data set. In the case of single cell epigenomics data the number of topics is low compared to other implementations (e.g. text classification). The running time will be affected by the number of topics.
- The Dirichlet hyperparameters **alpha** (`topic proportions`) and **beta** (`topic multinomials`): Alpha affects to the topic-cell contributions; a low alpha forces to pick for each cell a few topics with significant contribution, while a high alpha allows cells to have similar, smooth topic proportions. Beta affects to the region combinations; the lower the beta, the fewer regions a topic will have; the higher the beta, the less distinct these topics will be (i.e. there will be more overlap between the topics). By default, we select alpha as 50/number of topics and beta as 0.1 (as Griffiths & Steyvers, 2004).
- **Number of iterations and burnin**: For recording the assignments, it is necessary that the likelihood of the model is already stabilised. cisTopic counts with the function `logLikelihoodByIter` to check whether this parameters should be changed. The number of iterations affect the speed of the algorithm. Note that the burnin will be substracted from the number of iterations.

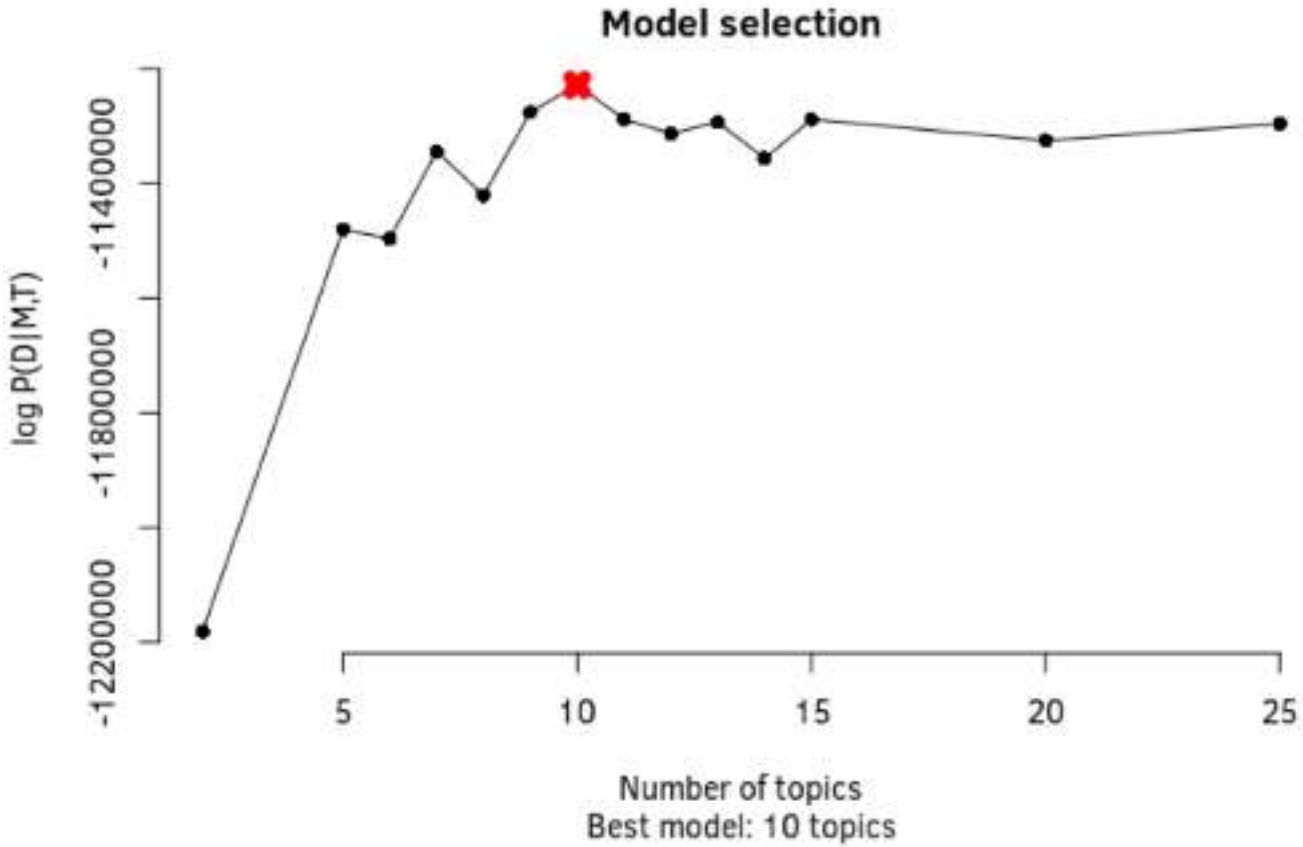
Build models (not run)

```r
```{r, eval=FALSE}
cisTopicObject <- runModels(cisTopicObject, topic=c(2, 5:15, 20, 25), seed=987, nCores=13, burnin = 120, iterations = 150, addModels=FALSE)
```
```

- The speed of this step is affected by:
    1. Size of the data set (number of regions and cells)
    2. Number of models
    3. Number of topics (>50 becomes slow)
    4. Number of CPUs
    5. Number of iterations
- With these settings it takes 3,5 minutes

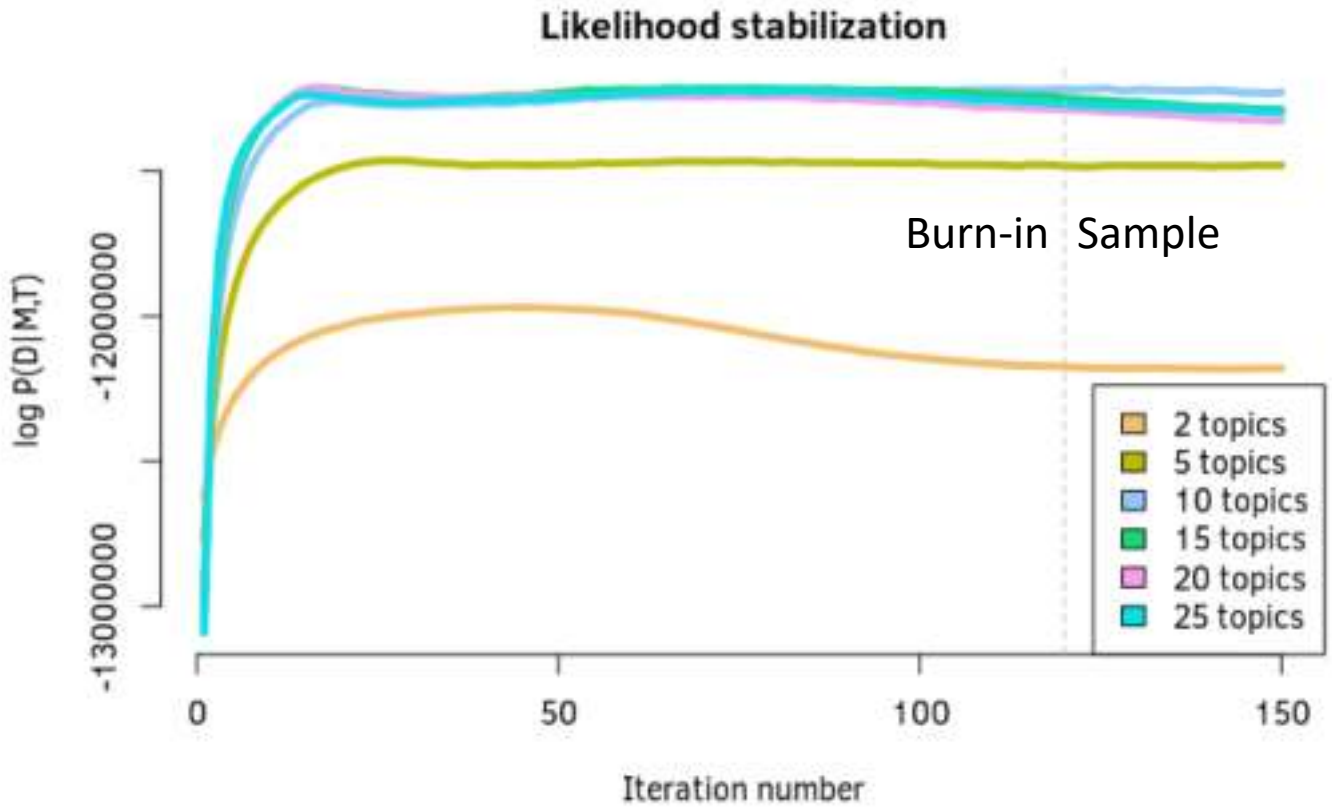
Model selection

```r
```{r, fig.show='hold', fig.align='center'}
cisTopicObject <- selectModel(cisTopicObject)
```
```



Check number of iterations and burn-in

```r
```{r, fig.show='hold', fig.align='center'}
logLikelihoodByIter(cisTopicObject, select=c(2,5,10,15,20,25))
```
```



Dimensionality reduction on cell-assignments

```r
```{r, eval=FALSE}
cisTopicObject <- runUmap(cisTopicObject, target='cell', method='Z-score')
```
```

Assignments normalized based on Z-score or Probability

See runtSNE, runDM and runPCA

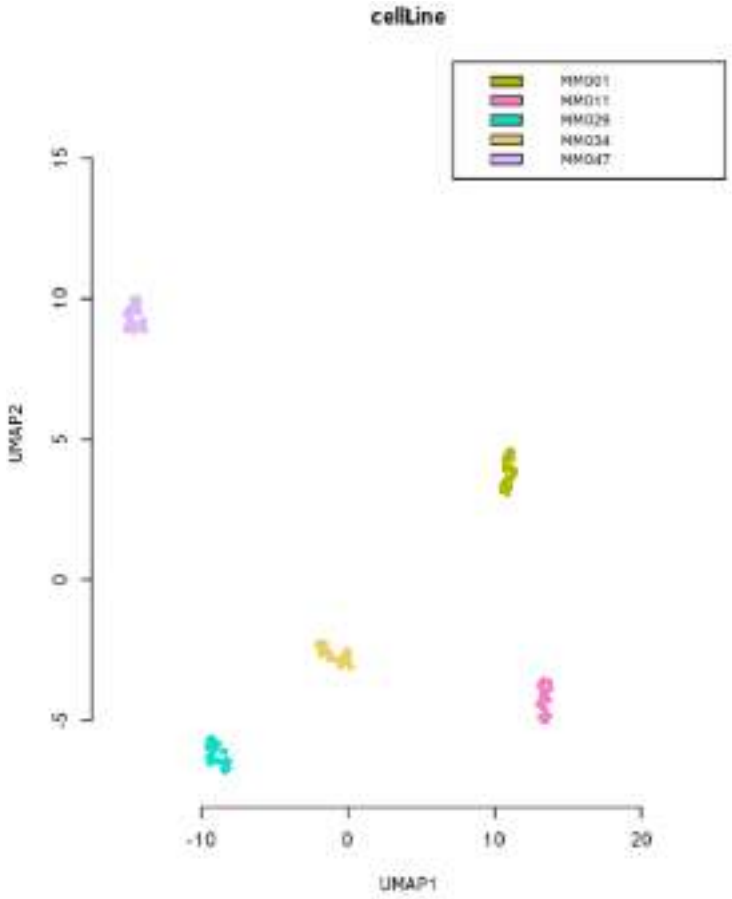
Plotting function

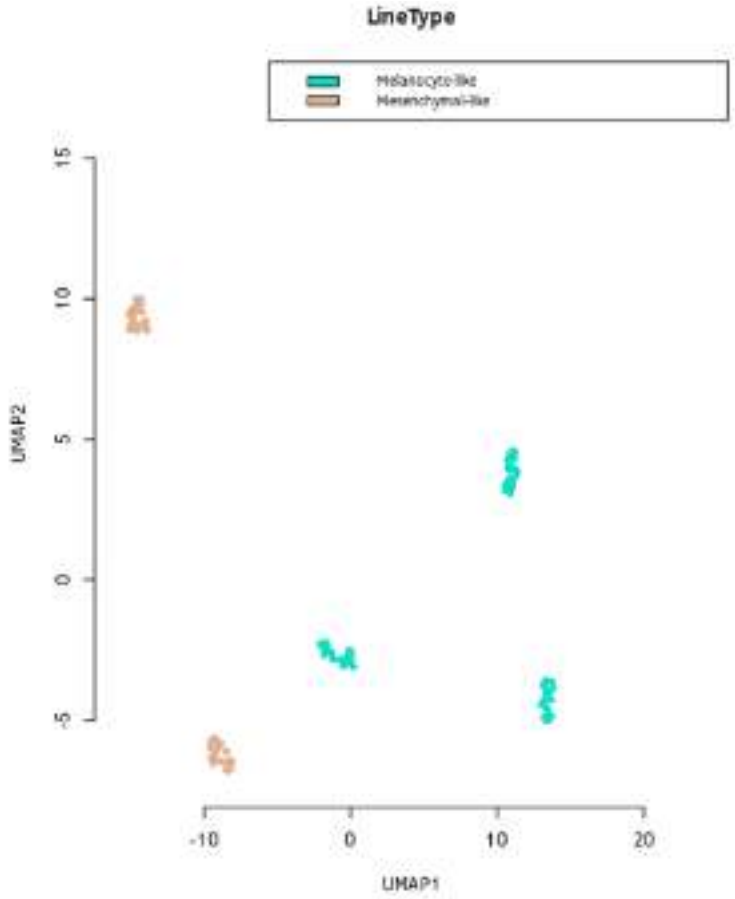
```r
```{r, fig.show='hold', fig.align='center'}
par(mfrow=c(1,3))
plotFeatures(cisTopicObject, method='Umap', target='cell', topic_contr=NULL, colorBy=c('cellline', 'LineType','pct_ReadsInPeaks'), cex.legend =
0.8, factor.max=.75, dim=2, legend=TRUE, col.low='darkgreen', col.mid='yellow', col.high='brown1', intervals=20)

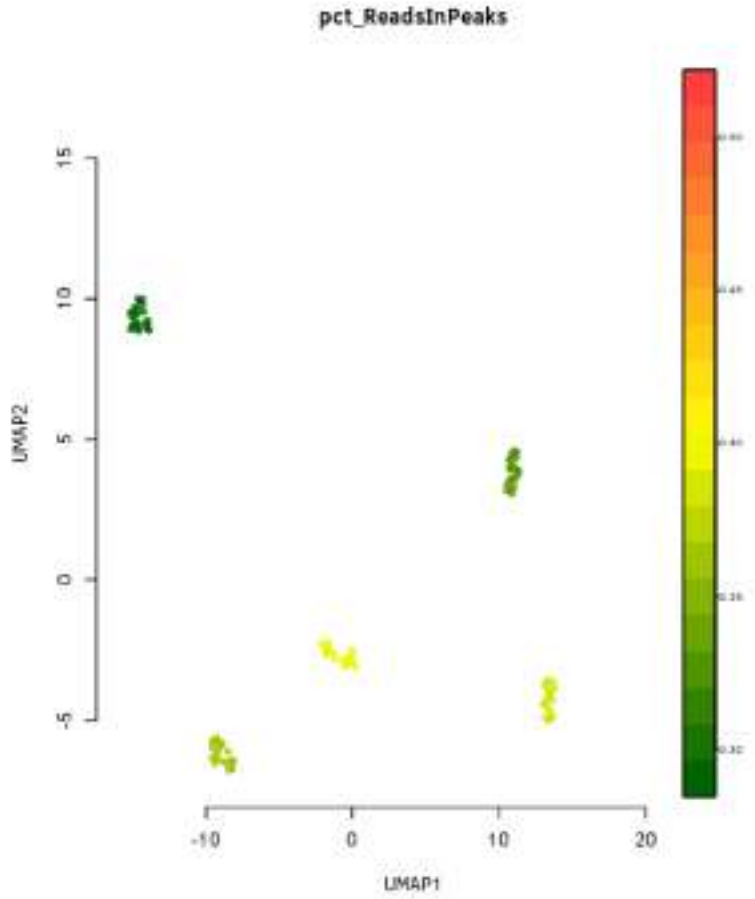
par(mfrow=c(2,5))
plotFeatures(cisTopicObject, method='Umap', target='cell', topic_contr='Probability', colorBy=NULL, cex.legend = 0.8, factor.max=.75, dim=2,
legend=TRUE)
```
```

Assignments normalized based on Z-score or Probability

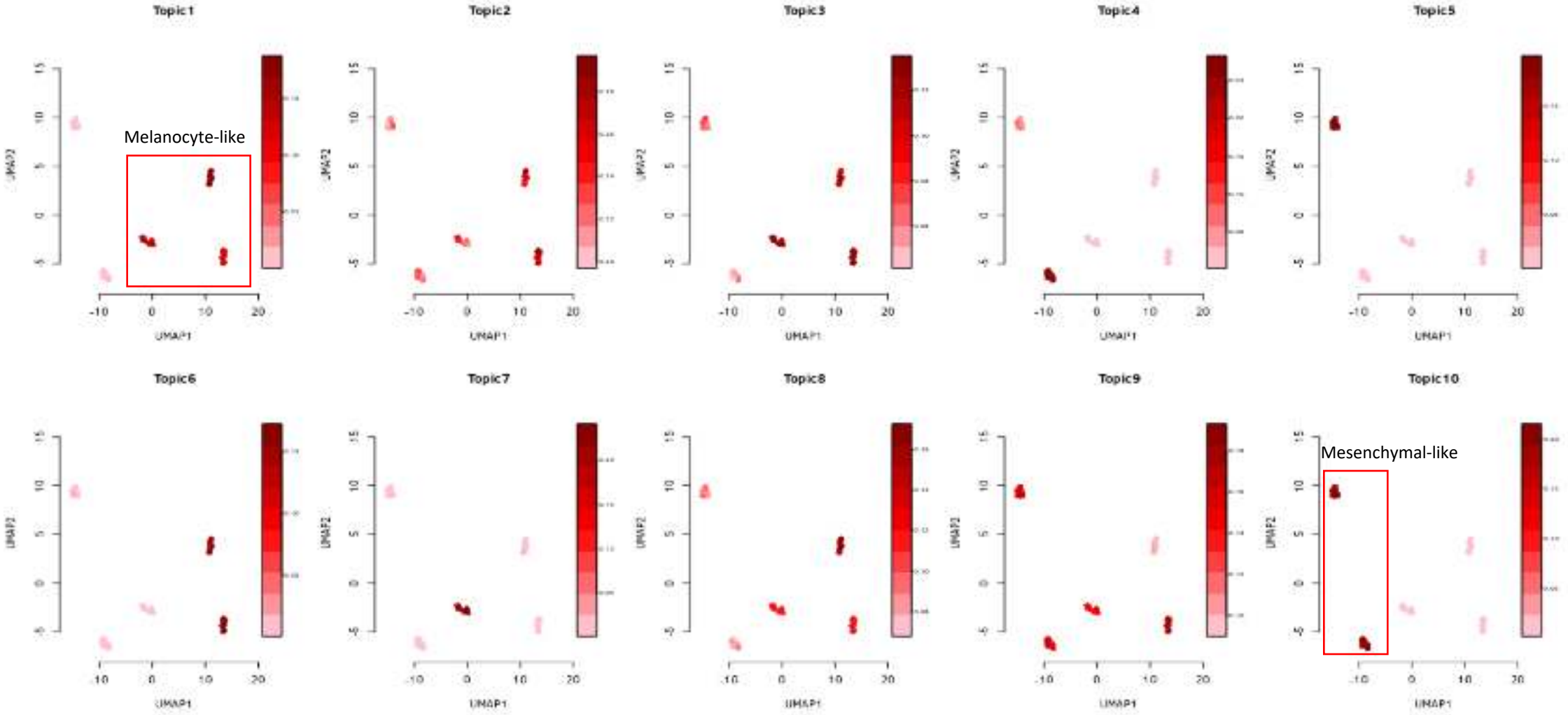
Dimensionality reduction on cell-assignments





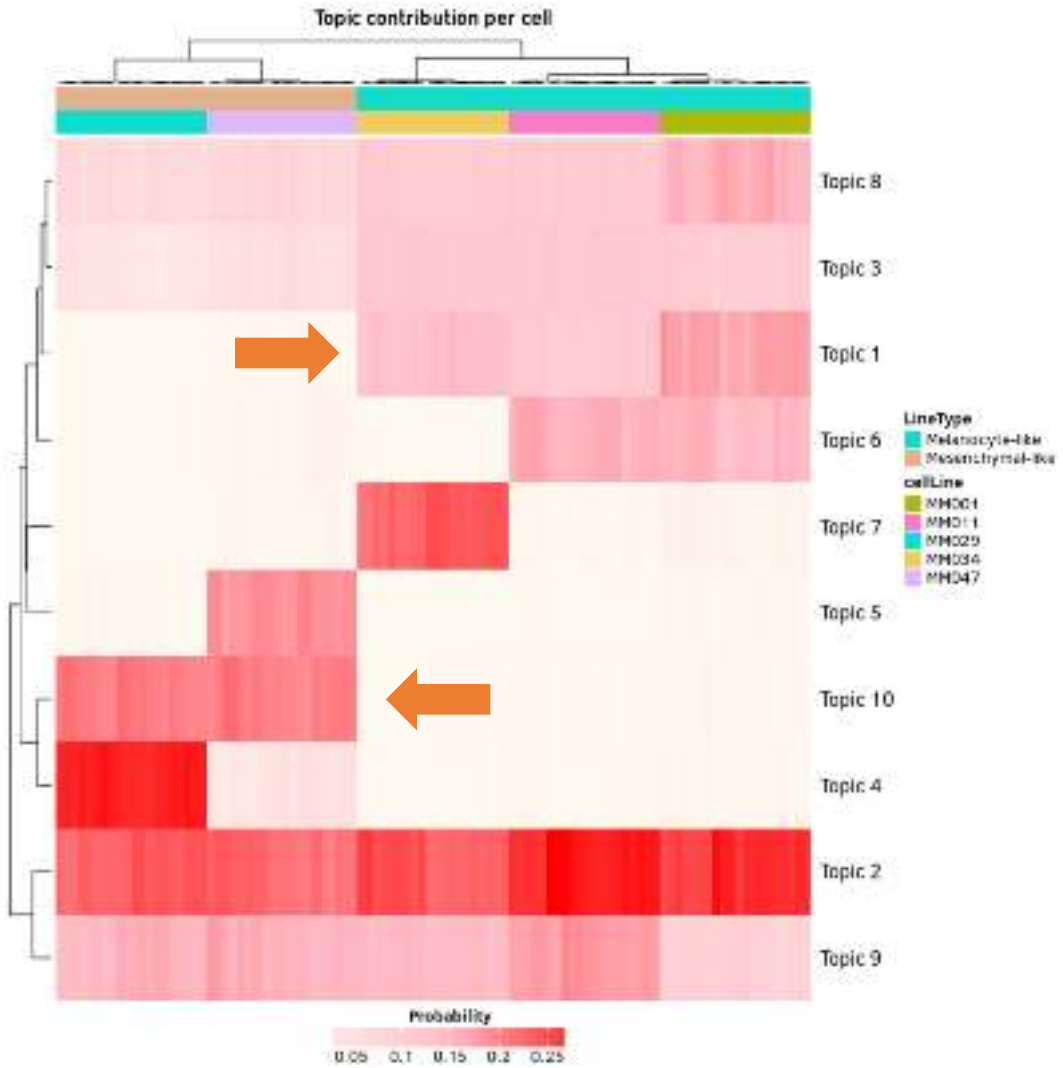


Dimensionality reduction on cell-assignments



Heatmap based on cell-assignments

```r
```{r, fig.show='hold', fig.align='center'}
cellTopicHeatmap(cisTopicObject, method='Probability', colorBy=c('LineType', 'cellLine'))
```
```



Enrichment of epigenomic signatures on cells

```{r, eval=FALSE}
pred.matrix <- predictiveDistribution(cisTopicObject)
```

Probability of a region in a cell $$P(r_i|c_j) = \sum_{k=1}^{K} P(r_i|T_k)P(T_k|c_j)$$

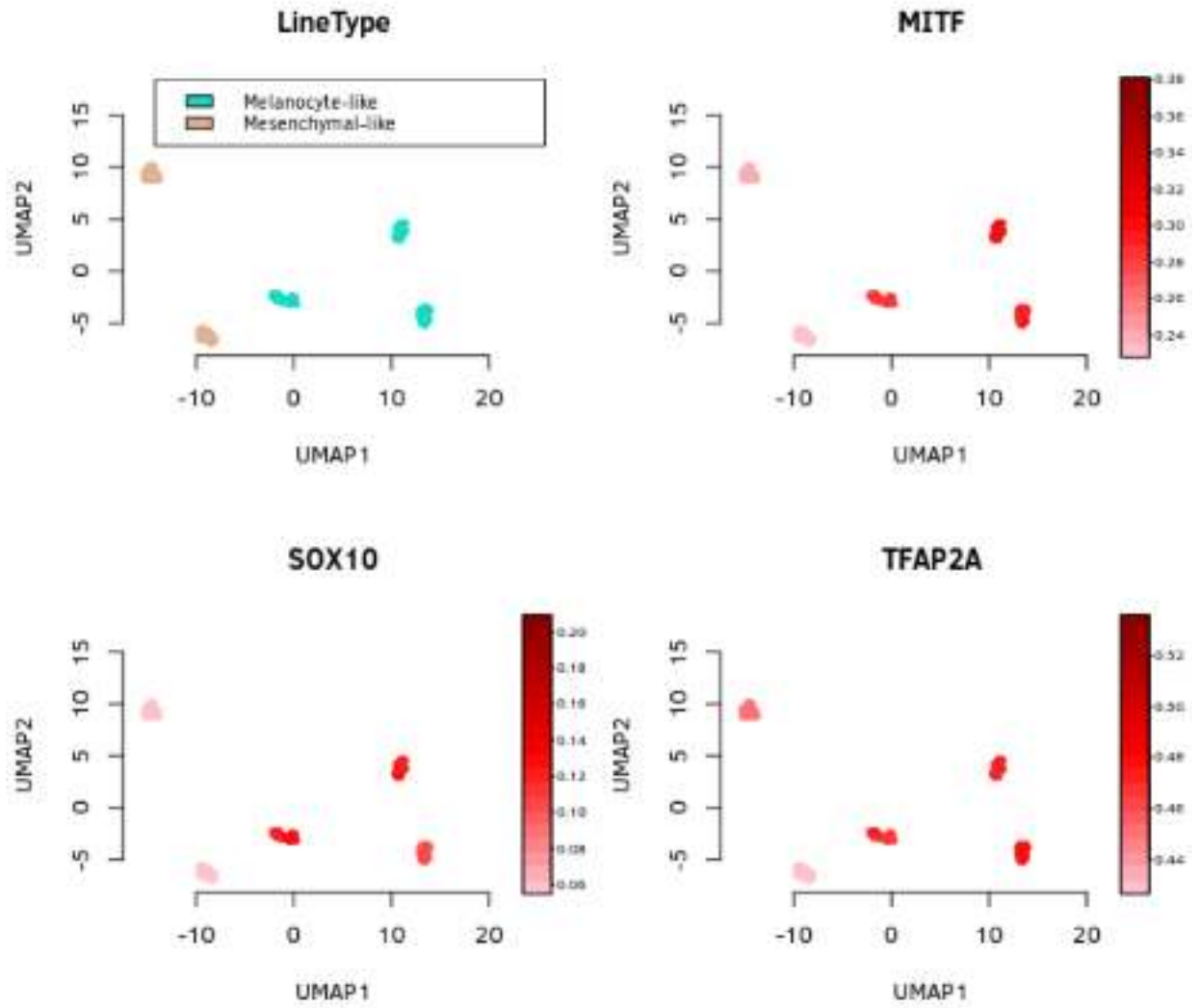
```{r, eval=FALSE}
# Obtain signatures
path_to_signatures <- 'data/ChIP-seq_signatures/'
ChIP_Seq_signatures <- paste(path_to_signatures, list.files(path_to_signatures), sep='')
labels  <- c('MITF', 'SOX10', 'TFAP2A')
cisTopicObject <- getSignaturesRegions(cisTopicObject, ChIP_Seq_signatures, labels=labels, minOverlap = 0.4)

# Compute cell rankings
library(AUCell)
aucellRankings <- AUCell_buildRankings(pred.matrix, plot=FALSE, verbose=FALSE)

# Check signature enrichment in cells
cisTopicObject <- signatureCellEnrichment(cisTopicObject, aucellRankings, selected.signatures='all', aucMaxRank = 0.1*nrow(aucellRankings), plot=FALSE)
```

Signatures (e,g, ChIP-seq regions) are mapped to the regions in the data set based on their overlap (i.e. a region is mapped to another region if the overlap is at least of 40% by default). Enrichment of these region signatures in each cell is performed with AUCell, using the mapped region sets as ‘gene sets’ and the cell-rankings based on the region probability per cell.

Enrichment of epigenomic signatures on cells



MITF, SOX10 and TFAP2A ChIP-seq signatures are enriched in melanocyte-like melanoma cell lines.

Calculation of topic scores for regions

```r
```{r, eval = FALSE}
cisTopicObject <- getRegionsScores(cisTopicObject, method='NormTop', scale=TRUE)
```
```

Assignments normalized based on Z-score, Probability or NormTop

NormTop (recommended) $\beta_{w,k}(\log \beta_{w,k} - 1/K \sum_{k'} \log \beta_{w,k'})$

Export to bigwig

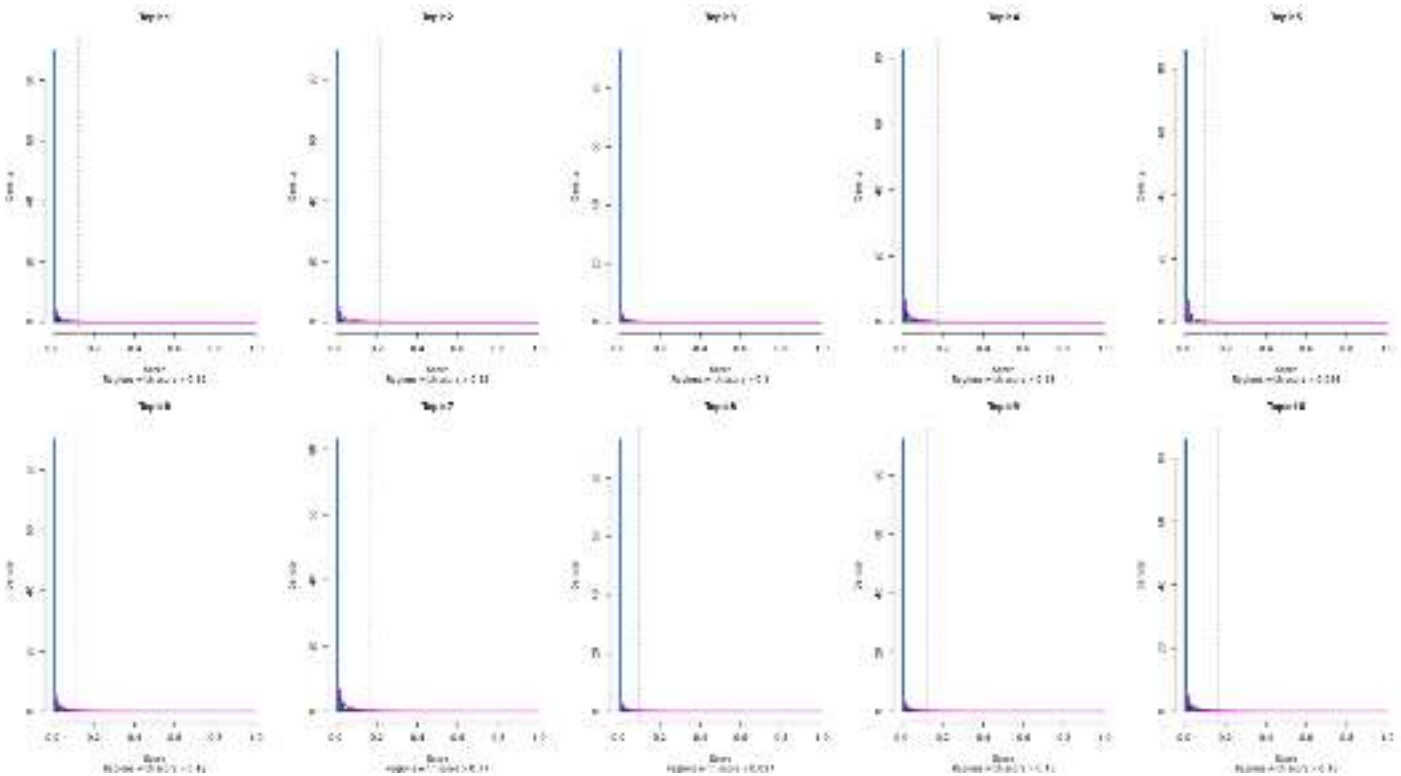
```r
```{r, eval = FALSE}
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene

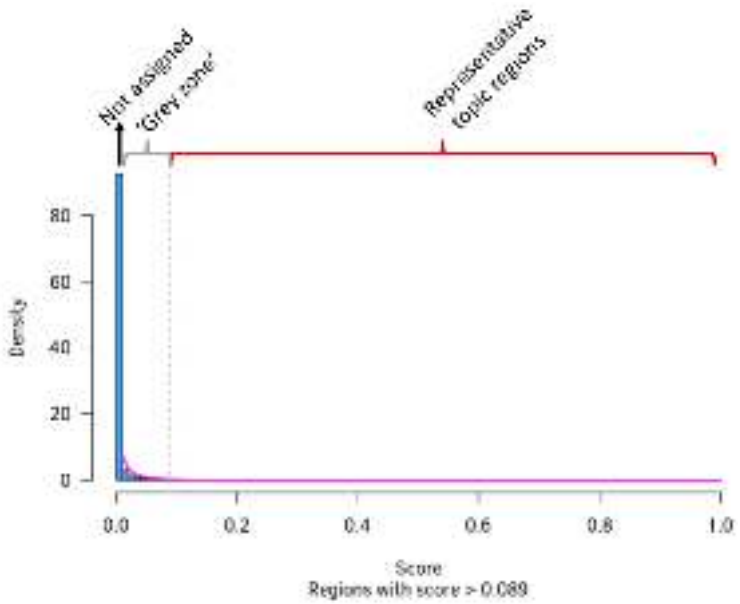
getBigwigFiles(cisTopicObject, path='output/cisTopics_asBW', seqlengths=seqlengths(txdb))
```
```

Binarize topics

```r
```{r, fig.show='hold', fig.align='center'}
par(mfrow=c(2,5))
cisTopicObject <- binarizecisTopics(cisTopicObject, thrP=0.975, plot=TRUE)
```
```

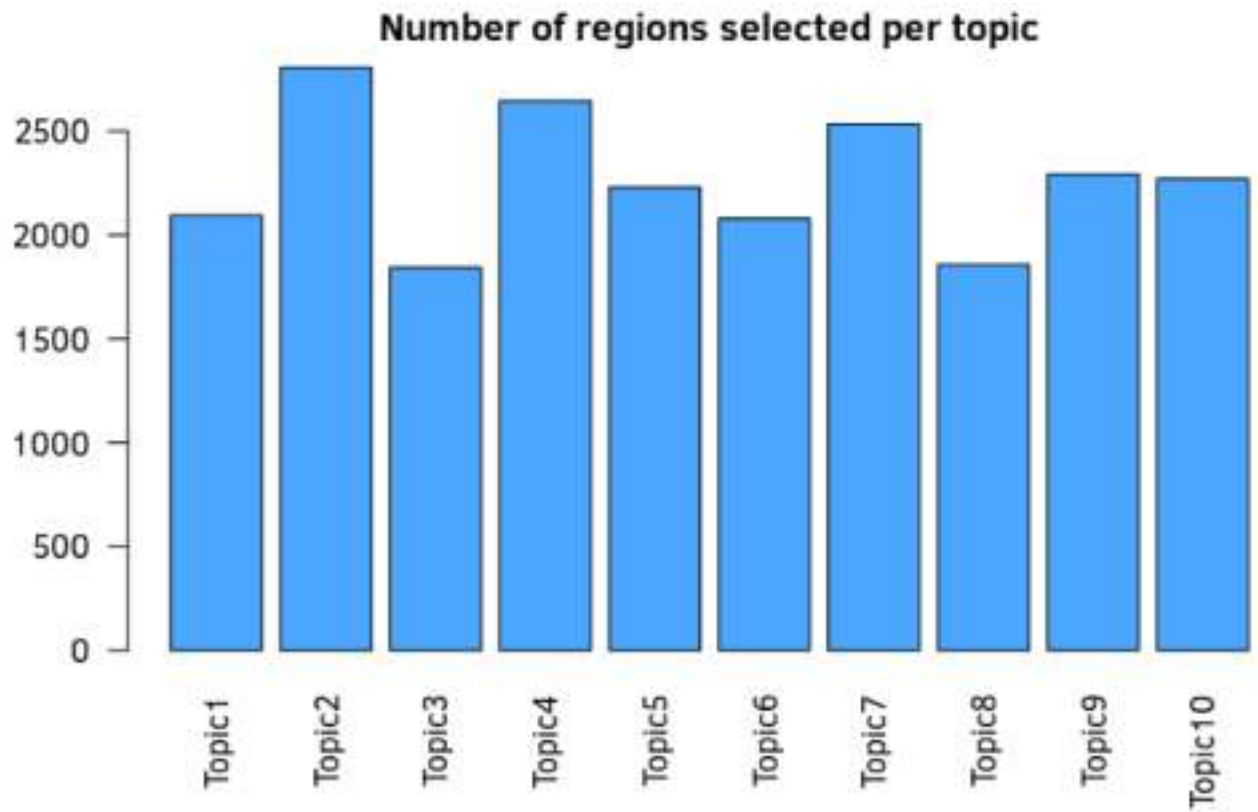
Take top regions or fit to a gamma distribution





Binarize topics

```r
```{r, fig.show='hold', fig.align='center'}
par(mfrow=c(2,5))
cisTopicObject <- binarizecisTopics(cisTopicObject, thrP=0.975, plot=TRUE)
```
```



Get bed files

```r, eval=FALSE}
getBedFiles(cisTopicObject, path='output/cisTopics_asBed')
```

Region clustering (based on region-topic scores) and visualization (not run)

```
```{r, eval=FALSE}
cisTopicObject <- runtSNE(cisTopicObject, target='region', perplexity=200, check_duplicates=FALSE)
```
```

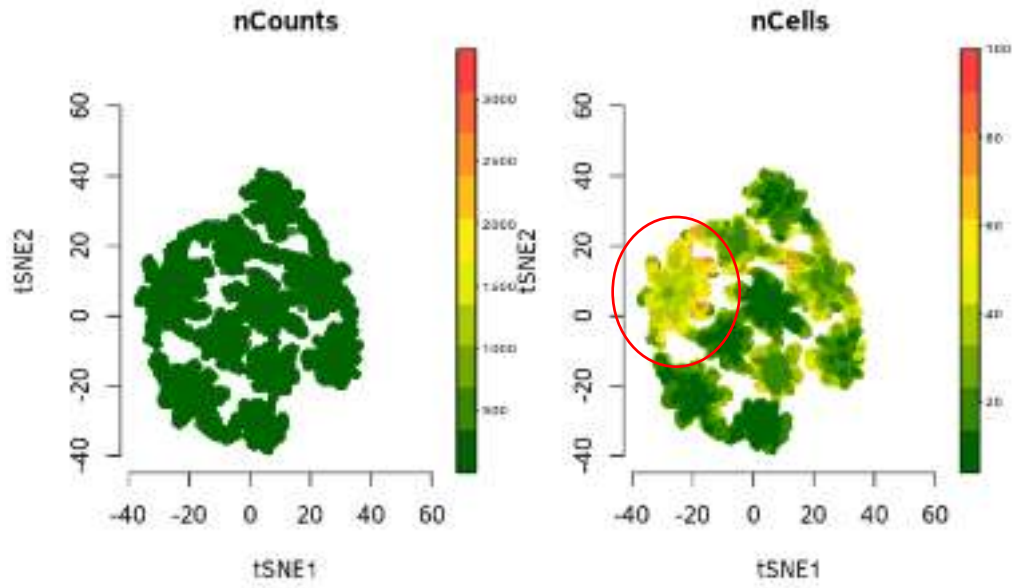
Assignments normalized based on NormTop, Z-score or Probability

NOTE: We only use high confidence regions (at least present in one binarized topic)

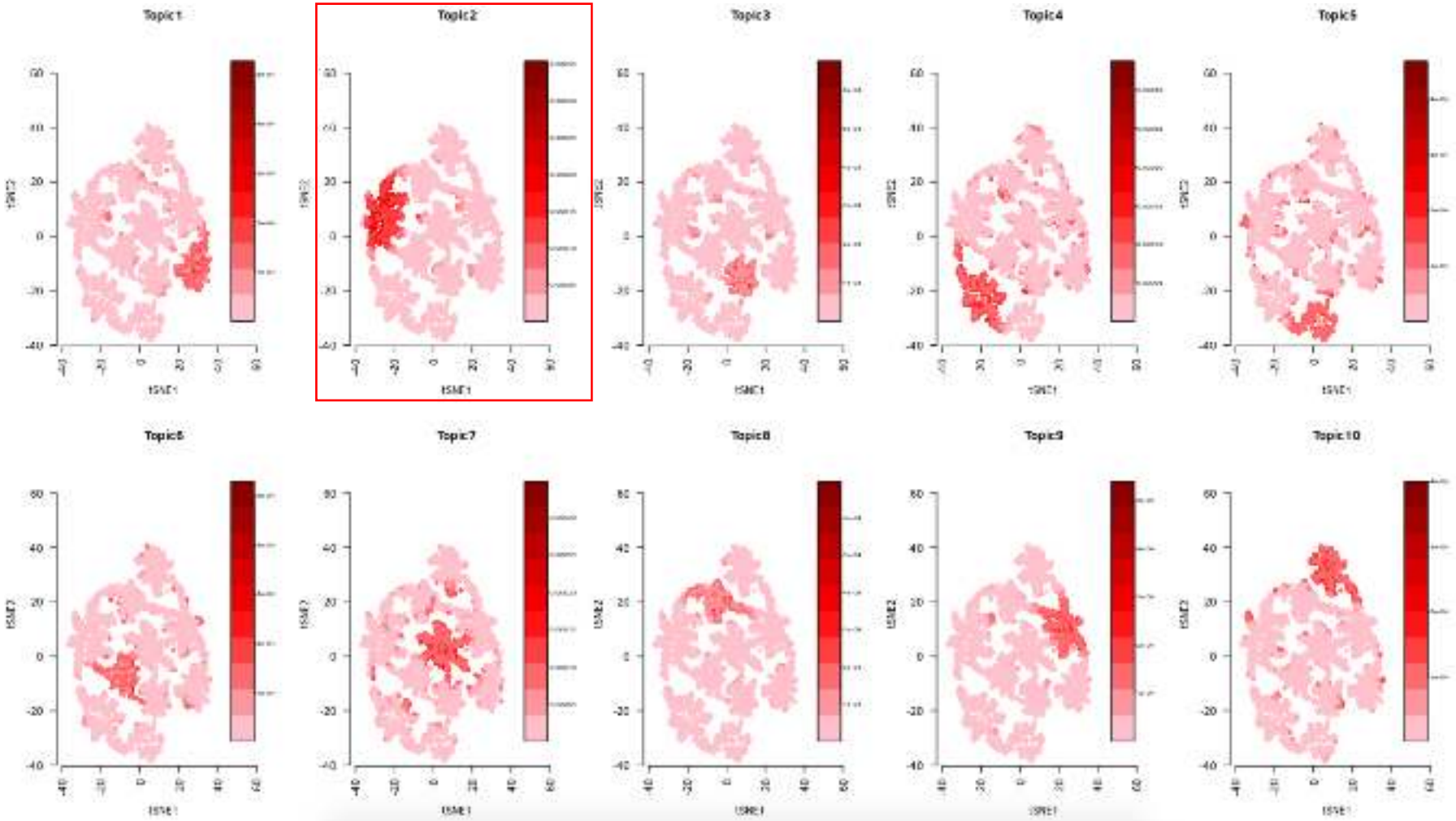
Region clustering (based on region-topic scores) and visualization

```r
```{r, fig.show='hold', fig.align='center'}
par(mfrow=c(1,2))
plotFeatures(cisTopicObject, method='tSNE', target='region', topic_contr=NULL, colorBy=c('nCounts', 'nCells'), cex.legend = 0.8, factor.max=.75, dim=2,
legend=TRUE, col.low='darkgreen', col.mid='yellow', col.high='brown1', intervals=10)

par(mfrow=c(2,5))
plotFeatures(cisTopicObject, method='tSNE', target='region', topic_contr='Probability', colorBy=NULL, cex.legend = 0.8, factor.max=.75, dim=2,
legend=TRUE)
```
```



Region clustering (based on region-topic scores) and visualization



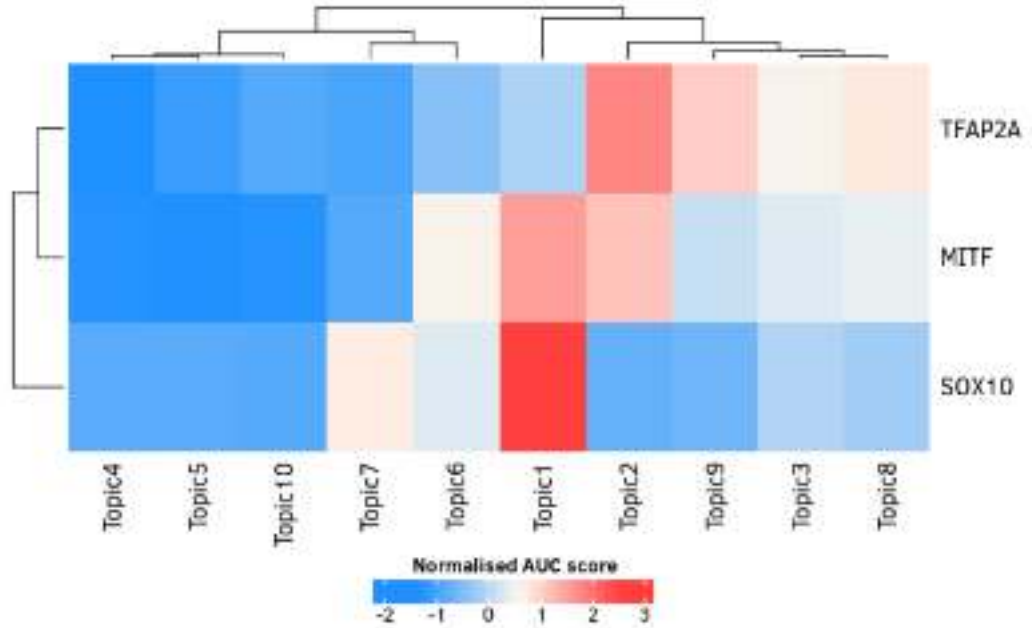
Enrichment of signatures within topics

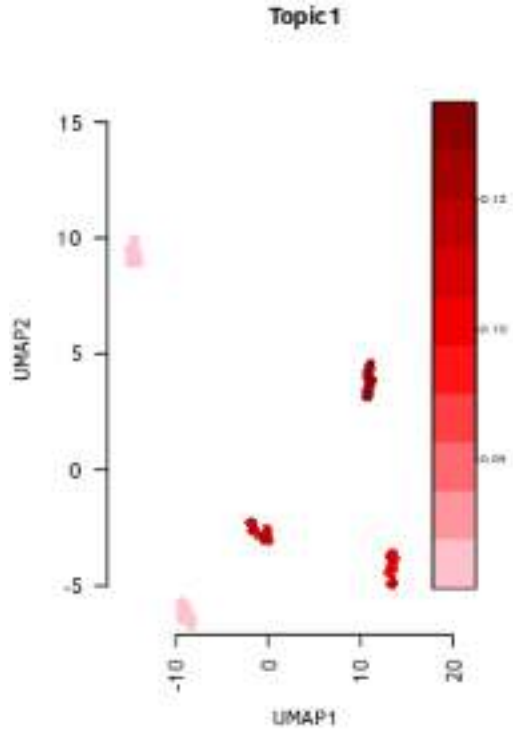
```r
# Obtain signatures (if it has not been run before)
path_to_signatures <- 'data/ChIP-seq_signatures/'
ChIP_Seq_signatures <- paste(path_to_signatures, list.files(path_to_signatures), sep='')
labels  <- c('MITF', 'SOX10', 'TFAP2A')
cisTopicObject <- getSignaturesRegions(cisTopicObject, ChIP_Seq_signatures, labels=labels, minOverlap = 0.4)
```

```r
signaturesHeatmap(cisTopicObject)
```

AUCell is also used for estimating the enrichment of signatures within topics, using as ‘gene sets’ the mapped signatures regions and the regions-rankings per topic as rankings.

Enrichment of signatures within topics



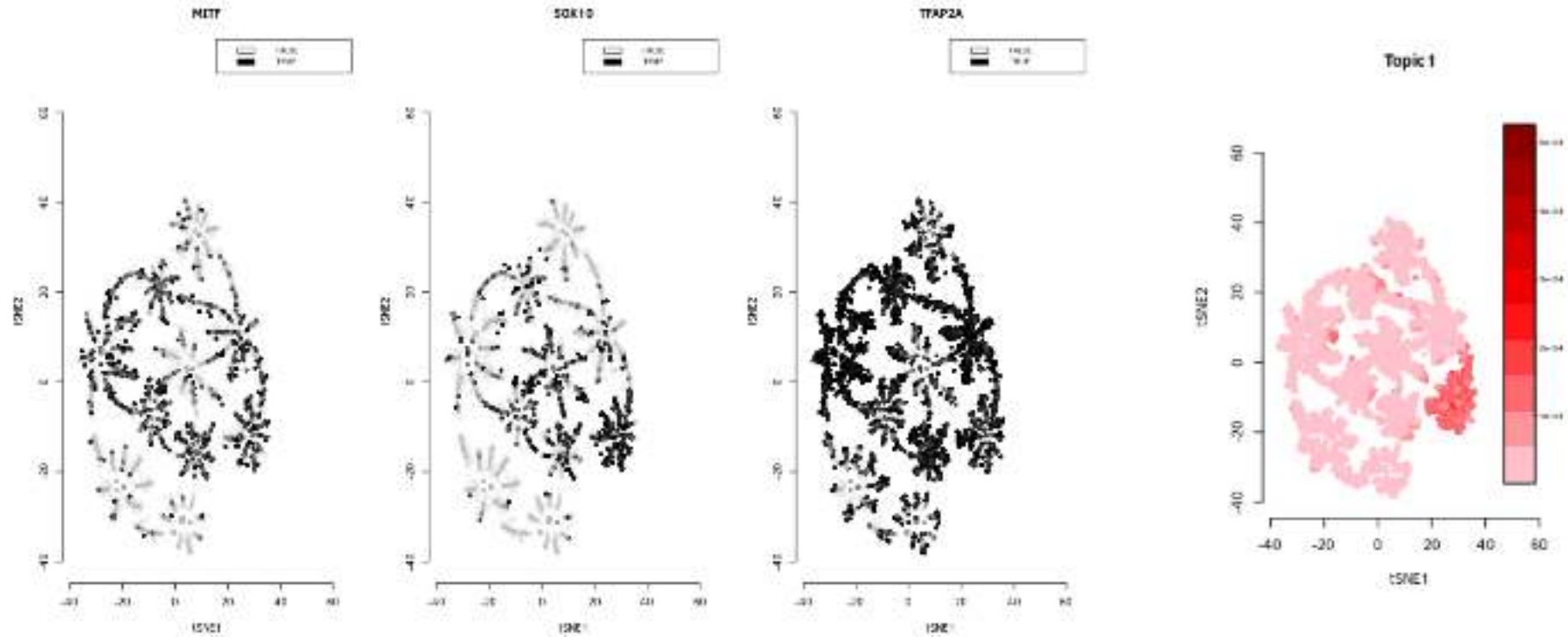


Topic 1 is a melanocyte-like topic and is enriched in SOX10 binding sites.

Showing overlapping regions in tSNE

```r
```{r, fig.show='hold', fig.align='center'}
colVars <- list()
colors <- c(adjustcolor('grey', alpha.f=0.05), 'black')
names(colors) <- c('FALSE', 'TRUE')
colVars[['MITF']] <- colors
colVars[['SOX10']] <- colors
colVars[['TFAP2A']] <- colors

par(mfrow=c(1,3))
plotFeatures(cisTopicObject, method='tSNE', target='region', topic_contr=NULL, colorBy=c('MITF', 'SOX10', 'TFAP2A'), cex.legend = 0.8, factor.max=.75,
dim=2, legend=TRUE, col.low='darkgreen', col.mid='yellow', col.high='brown1', intervals=20, colVars = colVars)
```
```

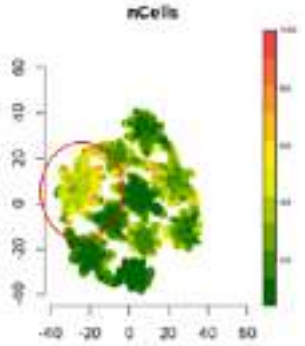


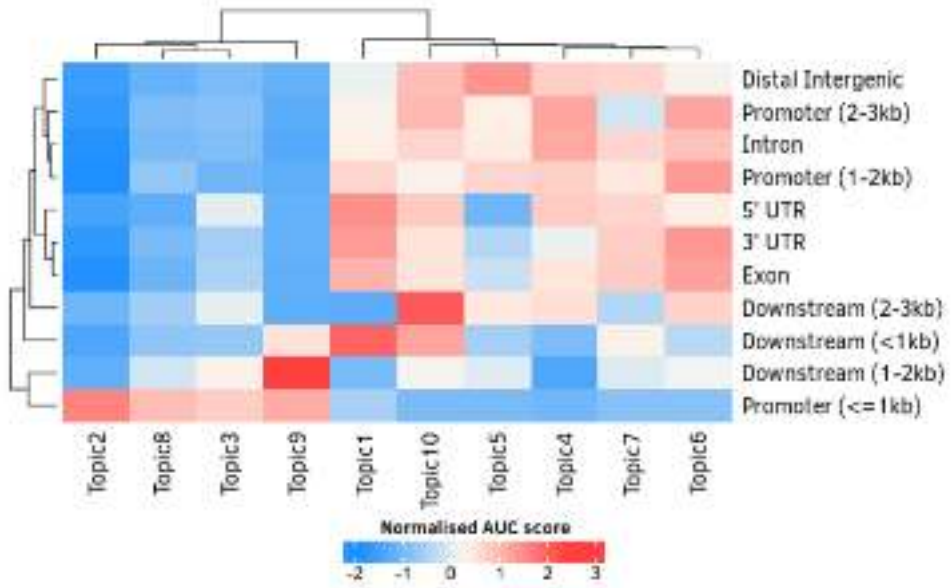
Annotation

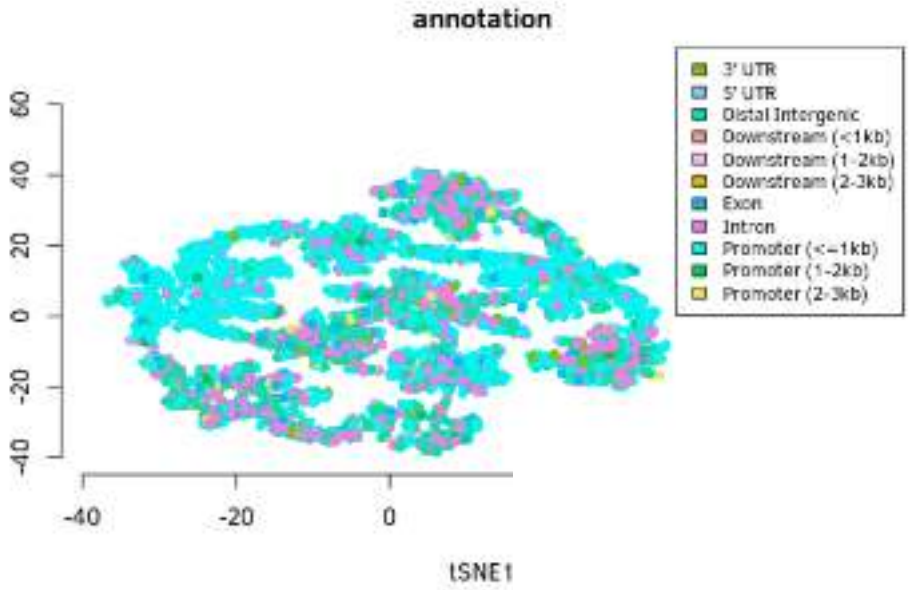
```r
```{r, eval=FALSE}
library(org.Hs.eg.db)
cisTopicObject <- annotateRegions(cisTopicObject, txdb=TxDb.Hsapiens.UCSC.hg19.knownGene, annoDb='org.Hs.eg.db')
```
```

Annotates region to closest gene and classifies the type of region (i.e. promoter, distal, …), These labels can be used as a signature (whose enrichment can be estimated in the topics).

```r
```{r, fig.show='hold', fig.align='center'}
par(mfrow=c(1,1))
signaturesHeatmap(cisTopicObject, selected.signatures = 'annotation')
plotFeatures(cisTopicObject, method='tSNE', target='region', topic_contr=NULL, colorBy=c('annotation'), cex.legend = 0.8, factor.max=.75, dim=2,
legend=TRUE, intervals=20, colVars = colVars)
```
```







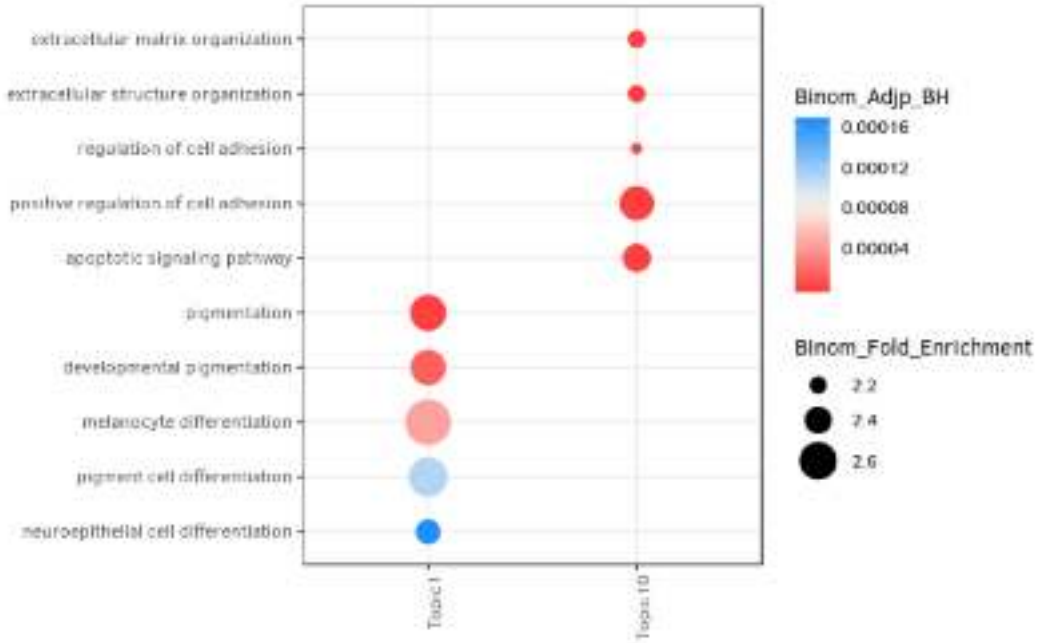
Topic 2 (and other general topics) are enriched in promoters

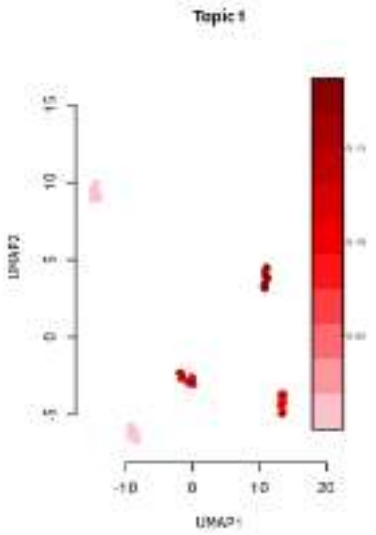
Using rGREAT for GO annotation

```r
cisTopicObject <- GREAT(cisTopicObject, genome='hg19', fold_enrichment=2, geneHits=1, sign=0.05, request_interval=10)
```

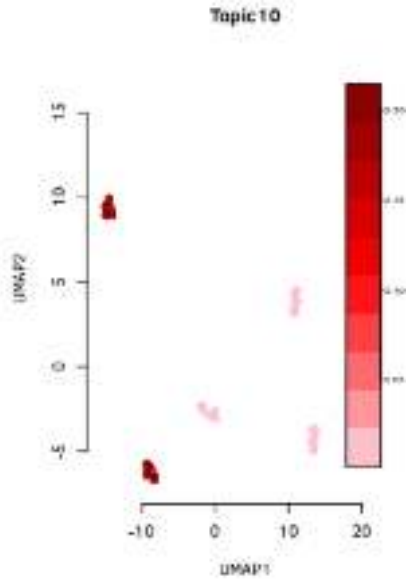
GO topic annotation (based on binarized topics, not run)

```r
ontologyDotPlot(cisTopicObject, top=5, topics=c(1,10), var.y='name', order.by='Binom_Adjp_BH')
```





Melanocyte-like



Mesenchymal-like

Motif enrichment with RcisTarget (not run)

Convert region to ctx regions (overlap 40%)

```{r, eval=FALSE, message=FALSE}
cisTopicObject <- binarizedcisTopicsToCtx(cisTopicObject, genome='hg19')
```

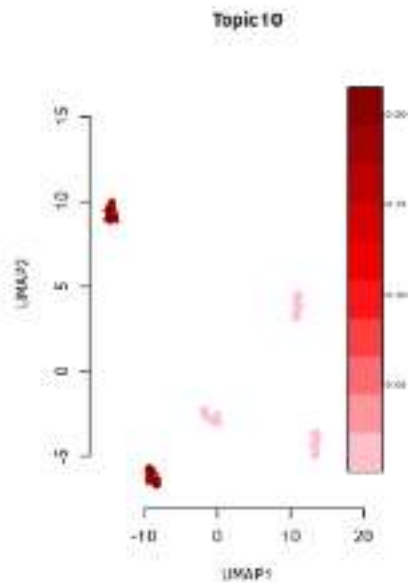
```{r, eval=FALSE}
cisTopicObject <- scoredRegionsToCtx(cisTopicObject, genome='hg19')
```

Run RcisTarget per topic

```{r, eval = FALSE}
date()
pathToFeather <- "feather/hg19-regions-9species.all_regions.mc8nr.feather"
cisTopicObject <- topicsRcisTarget(cisTopicObject, genome='hg19', pathToFeather, reduced_database=FALSE, nesThreshold=3, rocthr=0.005, maxRank=20000,
nCores=5)
date()
```

Motif enrichment with RcisTarget

```r
```{r, fig.show='hold', fig.align='center'}
Topic10_motif_enr <- cisTopicObject@binarized.RcisTarget[[10]]
DT::datatable(Topic10_motif_enr[,-c("enrichedRegions", "TF_lowConf"), with=FALSE], escape = FALSE, filter="top", options=list(pageLength=5))
```
```



Mesenchymal-like

Show 5 entries Search:

	logo	cisTopic	motif	NES	AUC	TF_highConf	nEnrRegions	rankAtMax
1		Topic_10	bocomnen__JUN_MOUSE.H11MO.0.A	6.75	0.0243	JUN (inferredBy_Orthology).	647	18877
2		Topic_10	cisbp__M6318	6.68	0.024	JUND (directAnnotation).	639	19792
3		Topic_10	bocomnen__FOS_HUMAN.H11MO.0.A	6.67	0.024	FOS (directAnnotation).	649	18877
4		Topic_10	bocomnen__JUNB_MOUSE.H11MO.0.A	6.66	0.024	JUNB (inferredBy_Orthology).	657	19792
5		Topic_10	cisbp__M4526	6.65	0.0239	SMARCC1 (directAnnotation).	662	18543

Formation of topic-specific cistromes

Get cistromes (i.e. regions enriched for motifs linked to a specific TF)

```{r, eval = FALSE}
cisTopicObject <- getCistromes(cisTopicObject, annotation = 'Both', nCores=5)
```

Enrichment of cistromes within cells (as previously explained for signatures in cells)

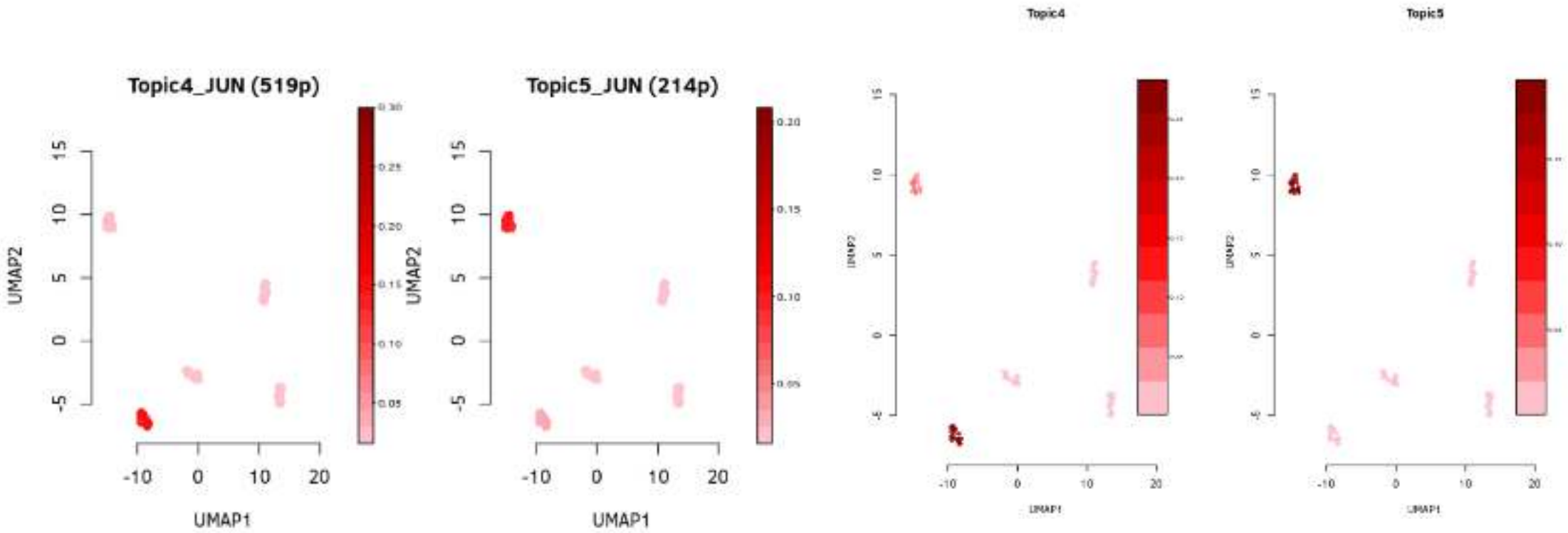
```{r, eval=FALSE}
# Compute AUC rankings based on the predictive distribution
pred.matrix <- predictiveDistribution(cisTopicObject)

library(AUCell)
aucellRankings <- AUCell_buildRankings(pred.matrix, plot=FALSE, verbose=FALSE)
```

```{r, eval=FALSE}
cisTopicObject <- getCistromeEnrichment(cisTopicObject, topic=4, TFname='JUN', aucellRankings = aucellRankings, aucMaxRank = 0.05*nrow(aucellRankings),
plot=FALSE)
cisTopicObject <- getCistromeEnrichment(cisTopicObject, topic=5, TFname='JUN', aucellRankings = aucellRankings, aucMaxRank = 0.05*nrow(aucellRankings),
plot=FALSE)
```

Visualization of topic-specific cistromes

```r
```{r, fig.show='hold', fig.align='center'}
par(mfrow=c(1,2))
plotFeatures(cisTopicObject, method='Umap', target='cell', topic_contr=NULL, colorBy=c('Topic4_JUN (519p)','Topic5_JUN (214p)'), cex.legend = 0.8,
factor.max=.75, dim=2, legend=TRUE, intervals=20)
```
```



There are cell-line specific (mesenchymal-like) AP-1 regions

Conclusions

- Melanoma cell-lines and classes can be identified based on their epigenomic profile
- cisTopic finds general, line type-specific and cell line-specific topics
- General topics are enriched for promoters, showing ubiquitous accessibility across the lines and a higher GC content
- Line type-specific topics reveal SOX10, MITF and TFAP as master regulators of the melanocyte-like state and AP-1 as master regulator of the mesenchymal-like state
- TF can have different targets depending on the cell type (likely dependent on cofactors)


VIB
KATHOLIEKE UNIVERSITEIT LEUVEN
GENOMICS CORE LEUVEN
KATHOLIEKE UNIVERSITEIT LEUVEN NANOCENTER
VSC
Vlaams Supercomputer Centrum
fwo
erc
European Research Council