Overview of single-cell ATAC computational tools

Ivan Berest

ivan.berest@biol.ethz.ch

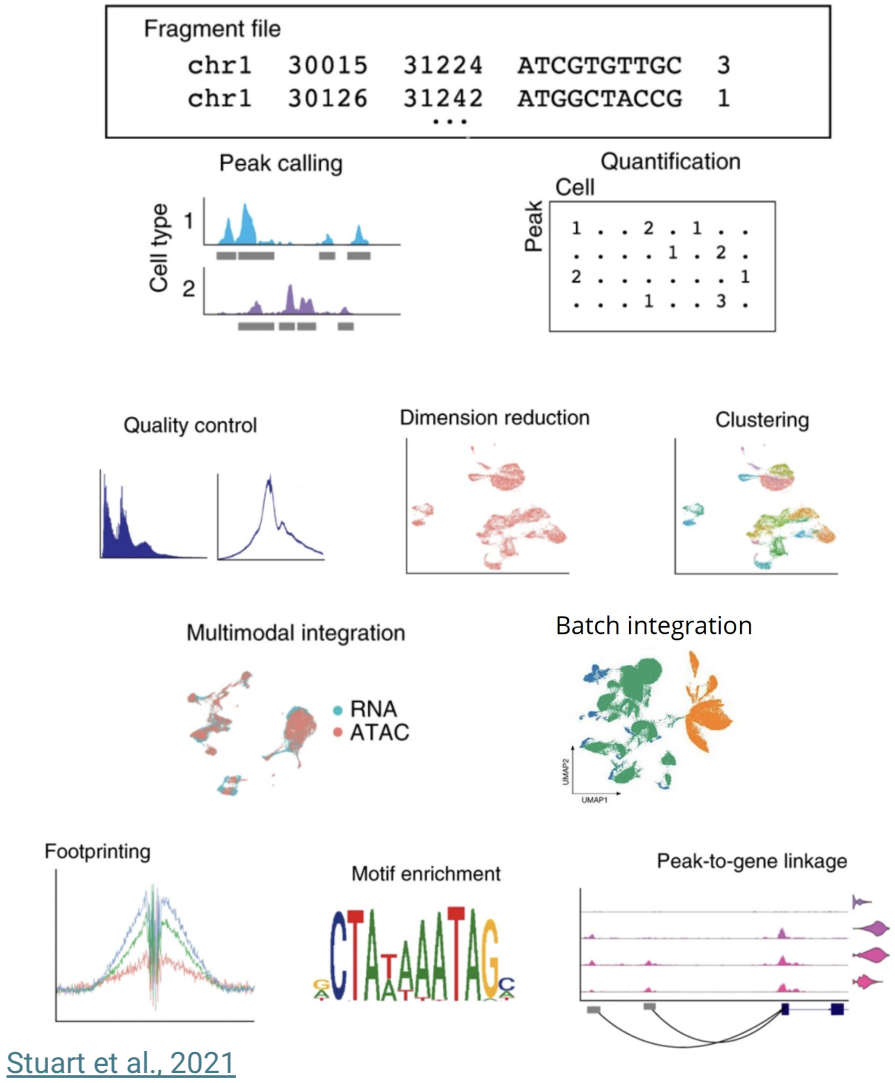
11.04.2024

- scATAC-seq processing workflows

ArchR, **Signac**, SnapATAC2, scATAC-pro, MAESTRO, EpiScanpy, ChrAccR,…

- TF related analyses
  - TF activity: chromVAR, BROCKMAN
  - Nucleosome/footprinting: DeNOPA, NucleoATAC, PRINT
  - TFBS prediction: maxATAC, scFAN
  - Embeddings: pycisTopic, CellSpace (TF/kmer based),SCALE
- Imputation & modelling: scOpen, AtacWorks, scBasset, BIRD (from RNA)
- CNV analysis: Alleloscope, epiAneufinder; Variant mapping: SCAVENGE
- Trajectory inference: Monocle3, STREAM, dynverse, Slingshot
- Annotation: Cellcanno, EpiAnno
- Gene regulation(GRNs): Cicero, DeepTFni, Dictys, SCENIC+, GRaNiE, Pando, SCARlink
- Data integration (modalities): GLUE, Seurat v3, WNN, MOJITOO, LIGER, scJoint, MOFA+, muon, FigR, scMVP, scDART, MultiVI, Cobolt, scMoMat,…

scATAC-seq processing workflow



Stuart et al., 2021

Cellranger-atac processing
(QC reads, mapping, filtering, peak calling)

Quality controls, filtering, normalization, dimensionality reduction, clustering

Integration between samples or modalities

Downstream analyses
(TF activity, TF footprinting, nucleosome, peak-to-gene, trajectory, GRN, ...)

ArchR

[Granja et al., 2021]

ArchR toolkit

Processing based on **arrow** files (similar to **hdf5**)

IterativeLSI better for rare cell types (TileMatrix != PeakMatrix)

All-in-one scATAC-seq analysis ecosystem

Data imputation with MAGIC

[ShinyArchR.UiO]

https://www.archrproject.com/

Signac analysis workflow

Package information	Package name	Signac	ArchR	SnapATAC	SCRAT	Dr.Seq2	Scasat
	Language	R/C++	R/C++	R	R	R/Python	R/Python
	Version	1.2.0	1.0.1	1.0.0	0.99.0	2.2.0	NA
Data input	Fragment file	Yes	Yes	Yes	No	No	No
	BAM file	No	Yes	Yes	Yes	Yes	Yes
	Count matrix	Yes	No	Yes	No	No	No
	Remote-hosted data support	Yes	No	No	No	No	No
Quality control	TSS enrichment	Yes	Yes	No	No	No	No
	Nucleosome signal	Yes	Yes	No	No	No	No
	Doublet prediction	No	Yes	No	No	No	No
	Genomic blacklist quantification	Yes	Yes	No	No	No	No
Quantification and peak calling	Genome-wide tile matrix	Yes	Yes	Yes	Yes	No	No
	Cell- or cluster-specific peak calling	Yes	Yes	Yes	No	No	No
	Peak matrix quantification	Yes	Yes	Yes	Yes	Yes	Yes
	Gene activity matrix quantification	Yes	Yes	Yes	Yes	No	No
Clustering and cell annotation	Dimensionality reduction	LSI	Iterative LSI	Jaccard + SVD	PCA	PCA	MDS
	Clustering	Yes	Yes	Yes	Yes	Yes	Yes
	Multimodal label transfer	Yes	Yes	Yes	No	No	No
	Differential accessibility	Yes	Yes	Yes	Yes	Yes	Yes
DNA motif analysis	DNA motif enrichment analysis	Yes	Yes	Yes	Yes	No	No
	Motif footprinting	Yes	Yes	No	No	No	No
Data visualization	Genome track plotting	Yes	Yes	No	No	No	No
	Interactive genome browser	Yes	Yes	No	No	No	No
	DNA motif visualization	Yes	No	No	No	No	No
	BigWig track support	Yes	No	No	No	No	No
Multimodal analysis	Multimodal data integration	Yes	Yes	Yes	No	No	No
	Paired scATAC + scRNA analysis	Yes	Yes	No	No	No	No
	Genome-wide peak-gene linking	Yes	Yes	Yes	No	No	No
	Paired scATAC + protein analysis	Yes	No	No	No	No	No
	Weighted nearest neighbor analysis	Yes	No	No	No	No	No
	Multi-assay support	Yes	No	No	No	No	No
	Spatial data support	Yes	No	No	No	No	No
	Cell hashing support	Yes	No	No	No	No	No
Mitochondrial genome analysis	Mitochondrial genotyping	Yes	No	No	No	No	No
	Mitochondrial clone identification	Yes	No	No	No	No	No
Software	Interface with Seurat	Yes	No	No	No	No	No
	Available on CRAN or Bioconductor	Yes	No	No	No	No	No
	Website with documentation	Yes	Yes	No	No	No	No
	User support forum	Yes	Yes	No	No	No	No
Supported operating systems	Linux	Yes	Yes	Yes	Yes	Yes	Yes
	macOS	Yes	Yes	Yes	Yes	Yes	Yes
	Windows	Yes	No	Yes	Yes	Yes	Yes

Designed to work with Seurat ecosystem

Modular design (easy to convert to/from)

Mitochondrial genotyping with mgatk

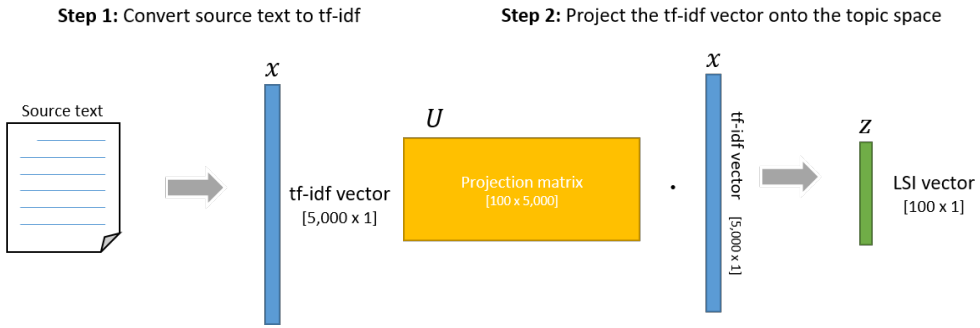
ChromatinAssay structure adds to the Seurat object to create multimodal objects

Stuart et al., 2021

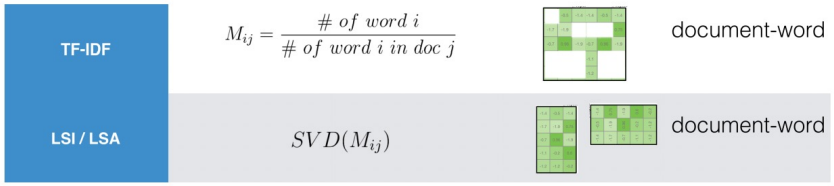
ArchRtoSignac

https://stuartlab.org/signac/

Dimensionality reduction

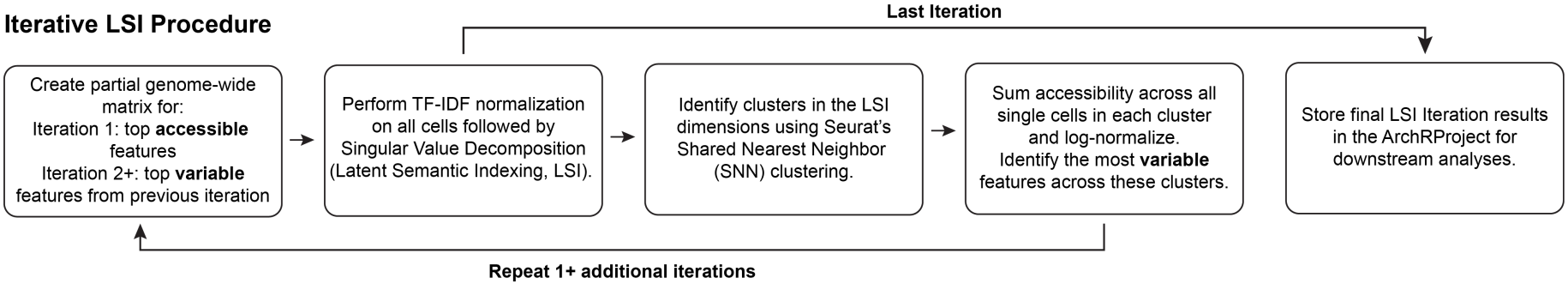


source



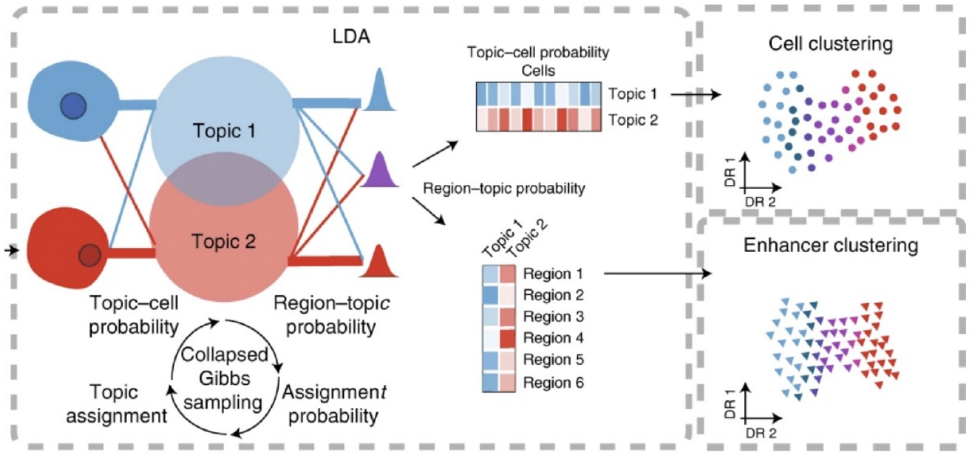
Signac: various TF-IDF implementations
log(TF x IDF);TF x log(IDF) – Cusanovich

ArchR: IterativeLSI – by iterations
identify most important features (peaks)



http://andrewjohnhill.com/blog/2019/05/06/dimensionality-reduction-for-scatac-data/

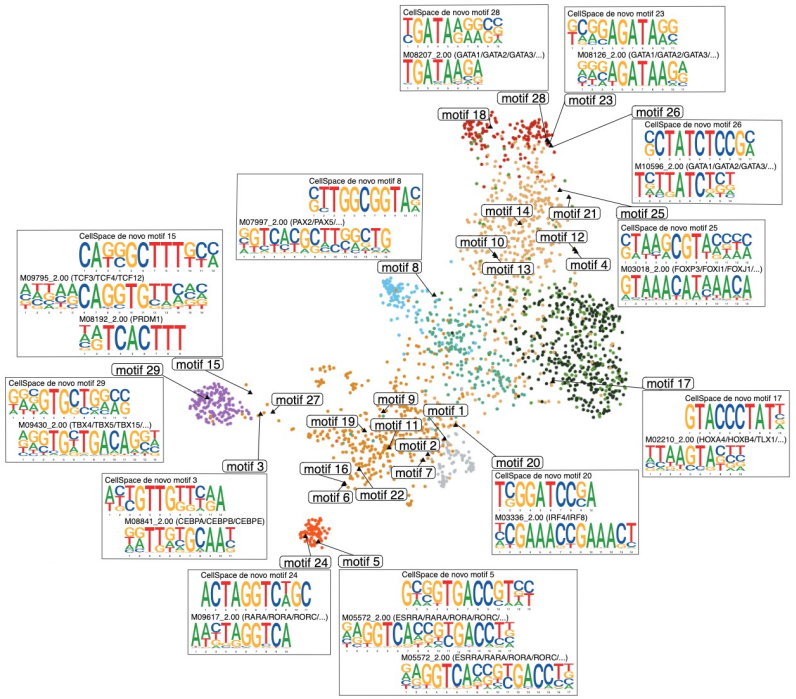
Interesting alternatives to standard LSI



Latent Dirichlet Allocation (LDA): unsupervised approach to simultaneously cluster cells and co-accessible regions into regulatory topics

What are regulatory topics?

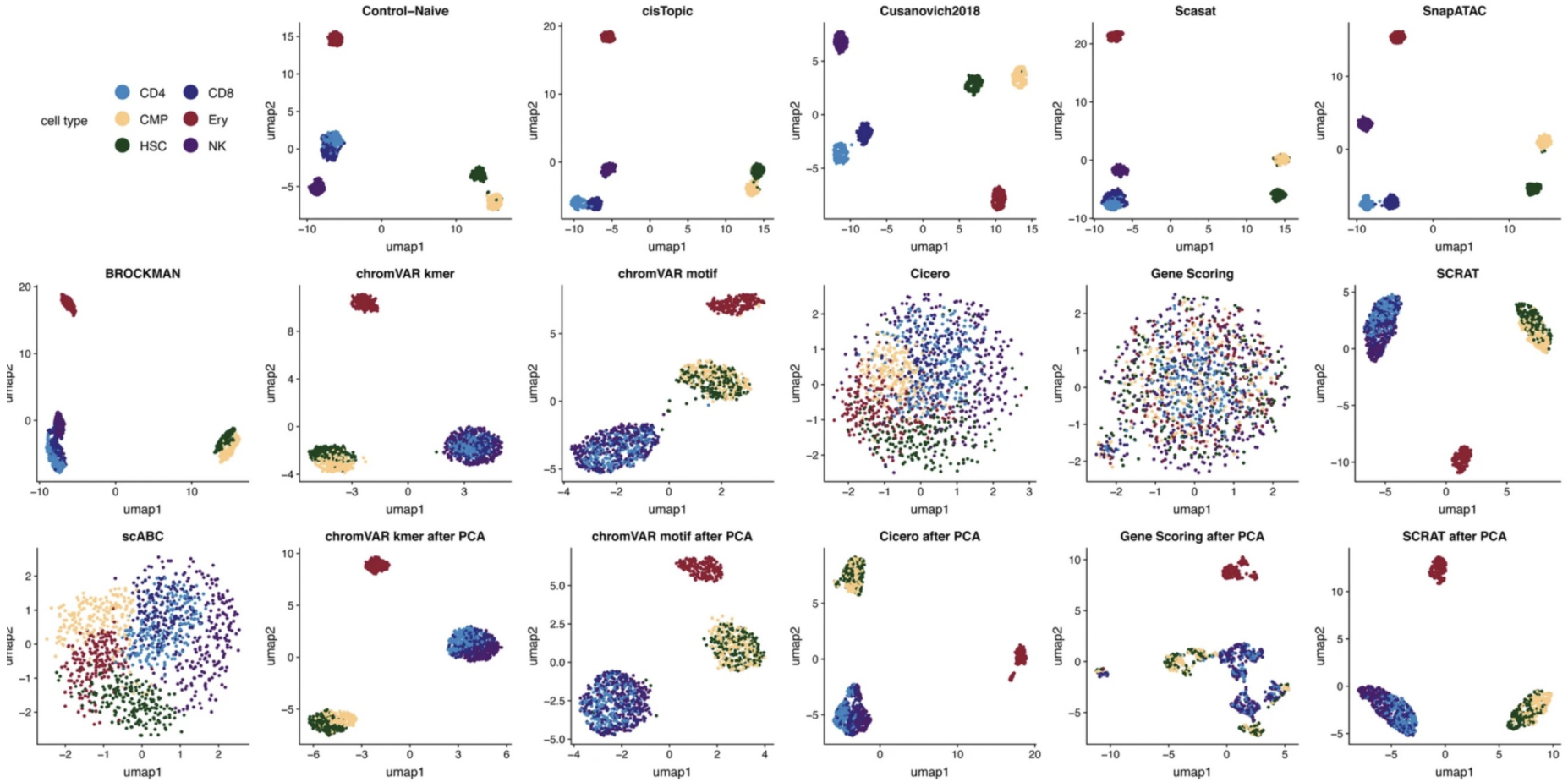
González-Blas et al., 2023



TFs as topics with StarSpace - CellSpace

Tayyebi et al., 2023

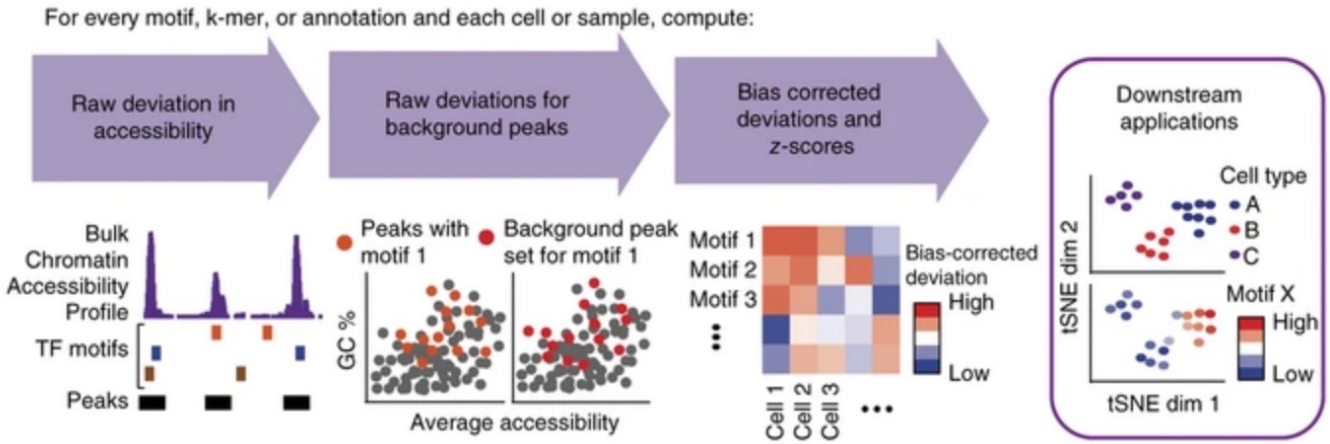
Final clustering is dependent on DR and feature space



Chen et al., 2019

Inferring TF activity from scATAC-seq

Schep et al., 2017



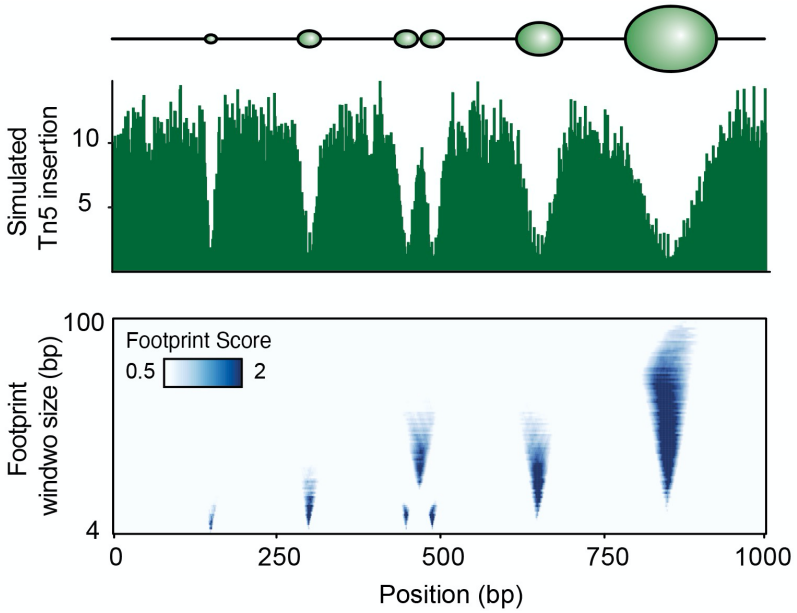
ArchR and Signac use chromVAR to define TF activity

BROCKMAN
TF activity uses k-mers frequencies

Important to define feature space (TF database):
JASPAR, HOCOMOCO, CisBP, UniProbe ...

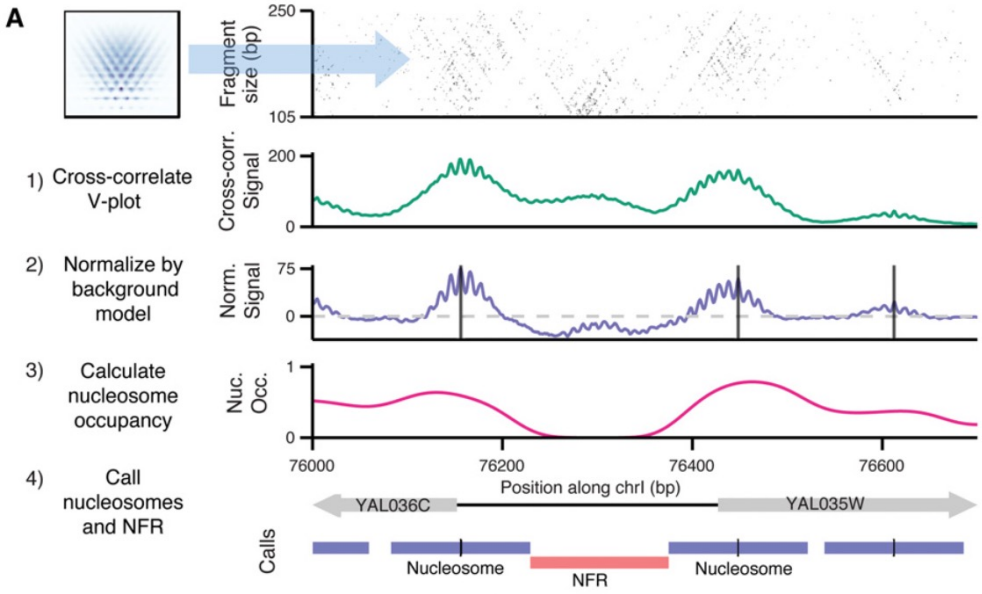
Jeff Vierstra effort to create non-redundant TF motif archetypes (v2.1 beta human).
Also provide scripts and procedure how to make archetypes for different organism

TF footprinting / Nucleosome occupancy



Multiscale footprinting framework (PRINT) calculate footprint score for subset of locations (enhancers or promoters)

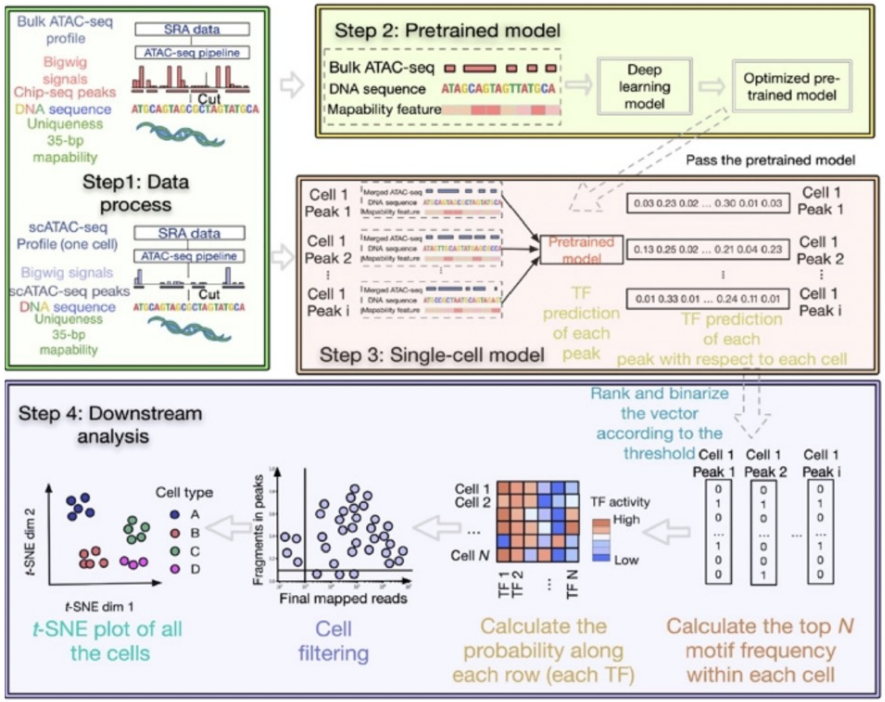
Hu et al., 2023

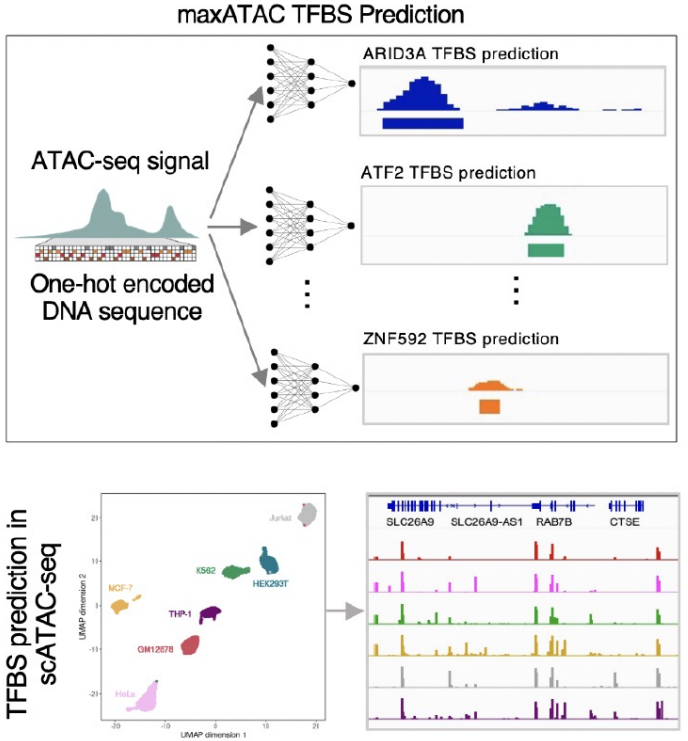


Use nucleosome-free (NFR) short reads and nucleosomal fragments to predict nucleosome occupancy

Schep et al., 2015

TFBS prediction using scATAC utilizing bulk ATAC





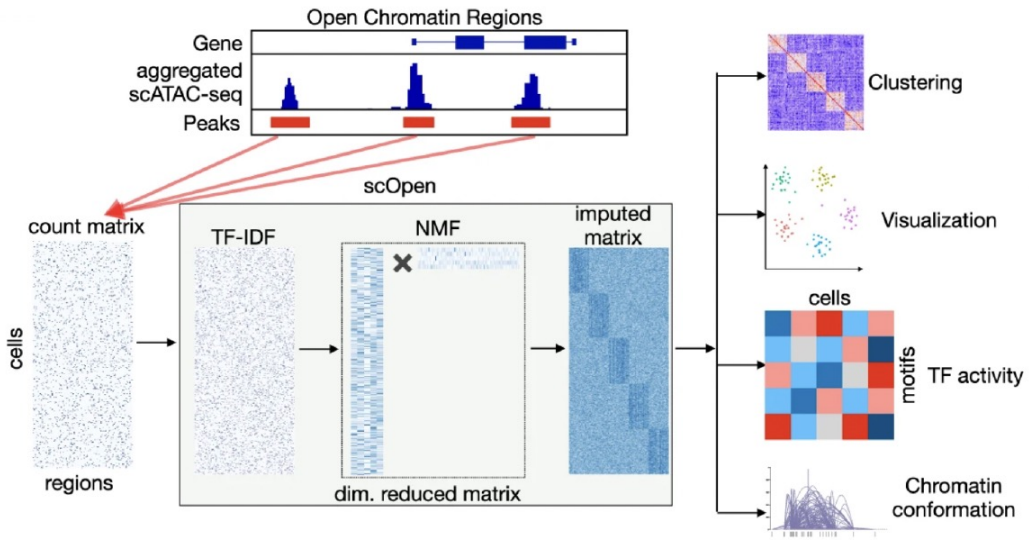
[scFAN](#) predict TF binding using DL on the scATAC data. Pretrained model use bulk ATAC, ChIP-seq and DNA sequencec

[maxATAC](#): Collection of DNNs for 127 TFs based on curated set of ATAC-seq and ChIP-seq data for TFBS prediction

[Fu et al., 2020](#)

[Cazares et al., 2023](#)

scATAC imputation and modelling



AtacWorks: deep-learning toolkit to denoise and improve peak calling from low quality (sc)ATAC data

Lai et al., 2021

scBasset: neural network to model scATAC data based on DNA sequence of the open chromatin

Yuan et al., 2022

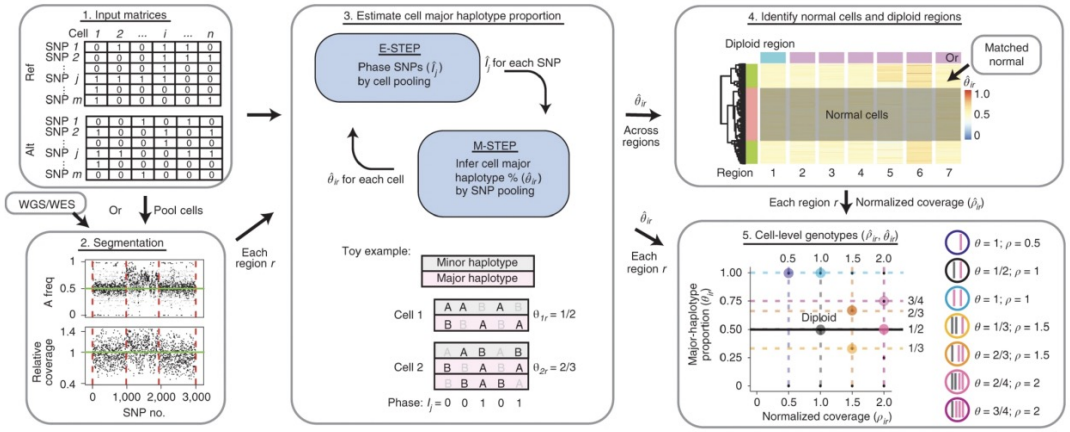
scOpen uses regularized NMF transformation to impute and denoise scATAC-seq data

Li et al., 2021

PoissonVAE: modelling fragment counts instead of Tn5 cuts for scATAC improve analysis

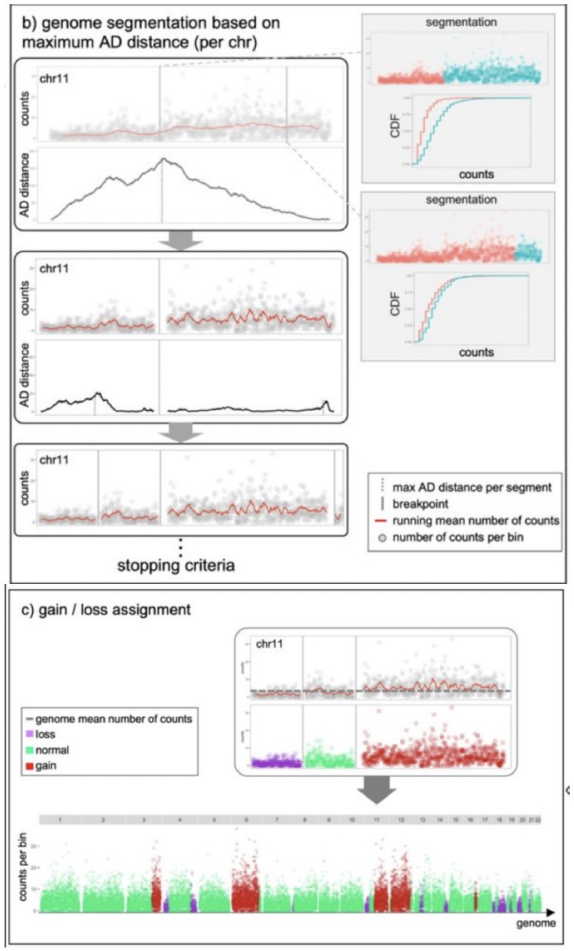
Martens et al., 2024

CNV analyses based on scATAC data



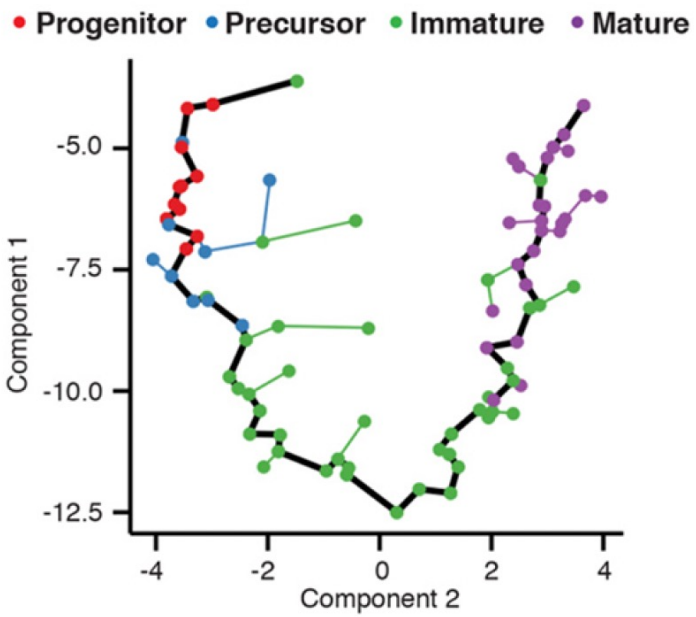
Alleloscope estimate cell major haplotype proportions, genotype each cell in each region

Wu et al., 2021



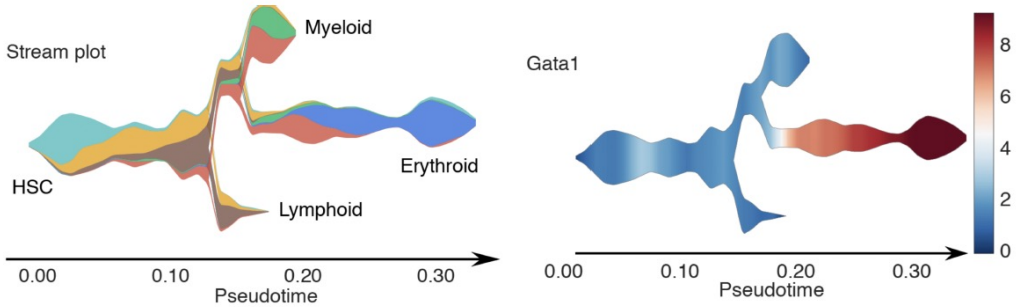
epiAneufinder: Similar goals, slightly different calculation algorithms (paper)

Trajectory analysis



Summarize chromatin accessibility changes and align on pseudotime
Monocle3 + Cicero toolkit

Pilner et al., 2018



STREAM: trajectory inference for mapping particular TFs, k-mers on the pseudotime

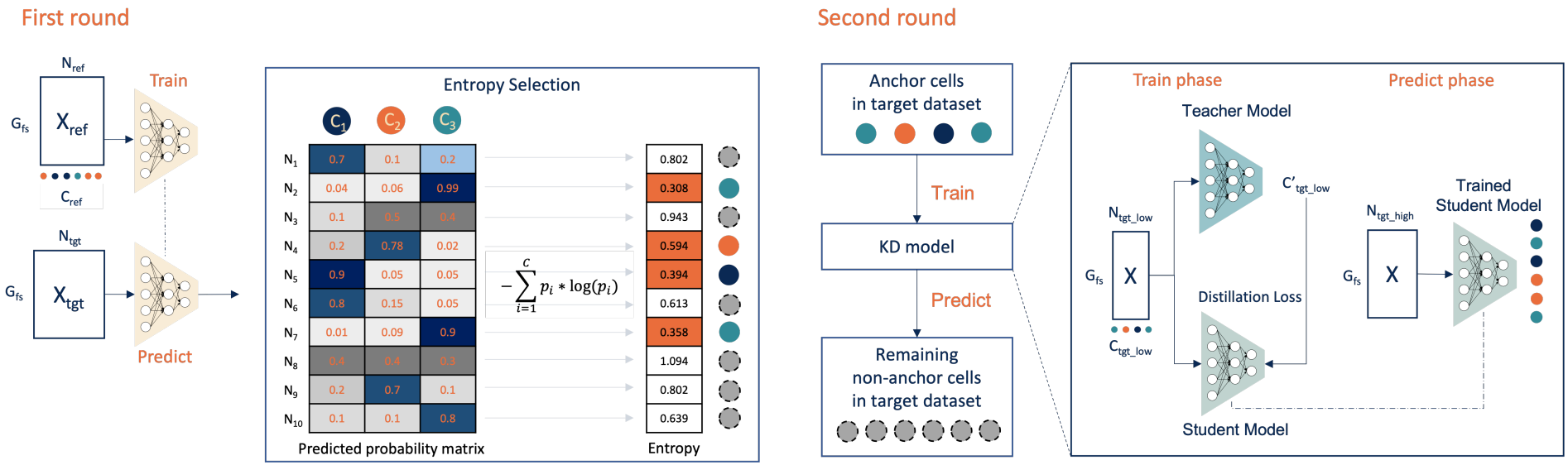
Chen et al., 2019

dynverse

We can use other trajectory methods for scATAC data (!not based on velocity)

Saelens et al., 2019

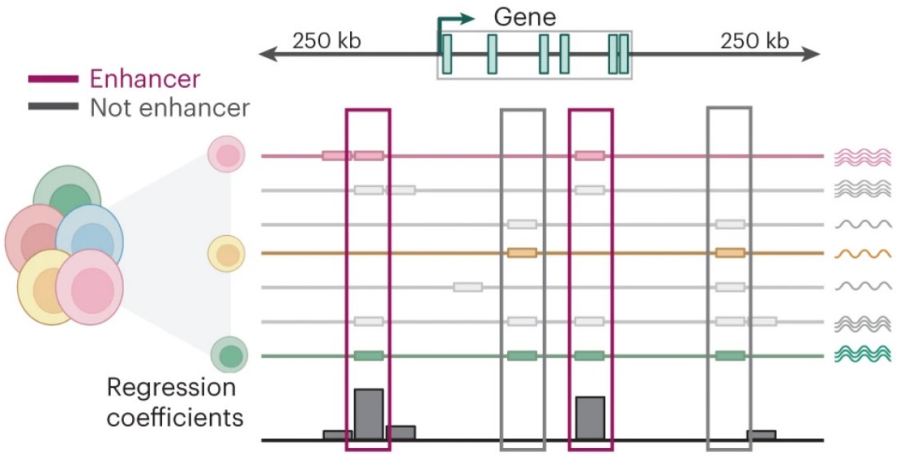
Celltype annotation



Cellcanno: supervised annotation of scATAC based on the reference annotated chromatin atlas (use of the published cell atlases)

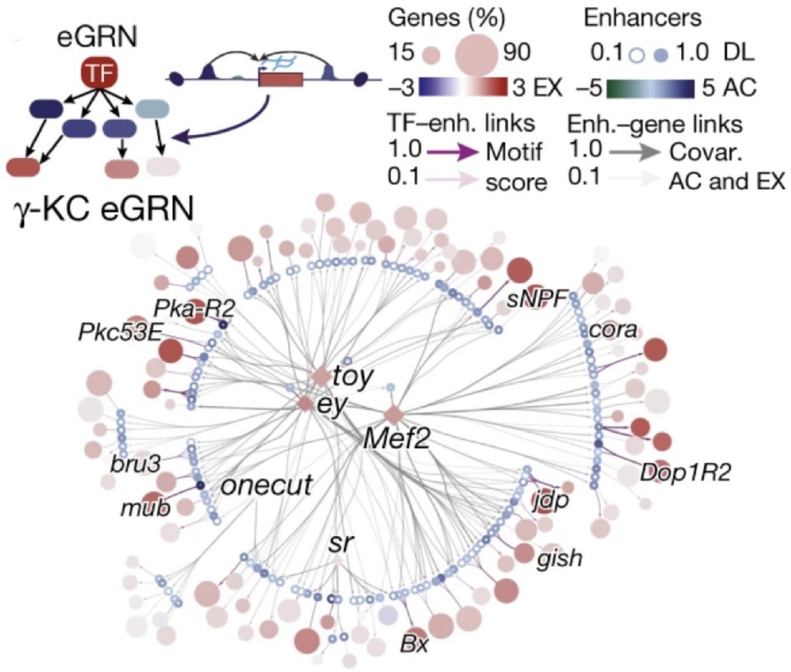
Ma et al., 2023

Gene regulation networks (GRN)



Link peaks and genes via proximity and correlation Cicero, SCARlink (multiome data, if unpaired used as reference)

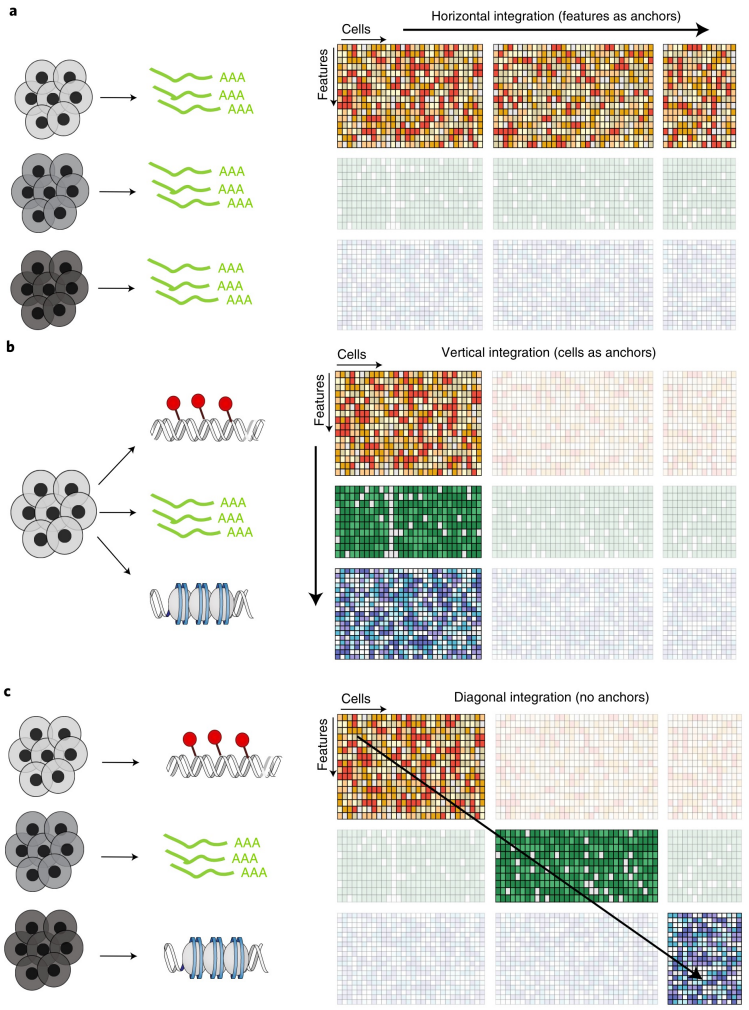
Mitra et al., 2024



Linking scATAC and scRNA data focusing around TF hubs (SCENIC+)

González-Blas et al., 2023

Argelaguet et al., 2021



Data integration methods

Horizontal integration: **different** cells/samples – **same** technology
Seurat v3, LIGER, Harmony

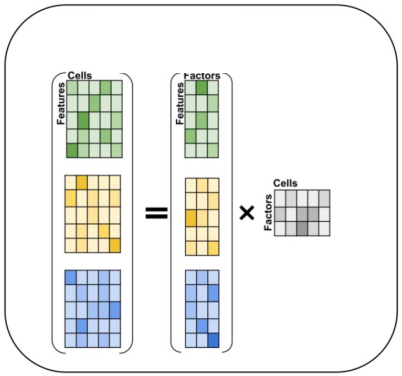
Luecken et al., 2022

Vertical integration: **same** cells/samples – **different** technology
WNN, MOFA+, MOJITOO, muon, scMVP, MultiVI, Cobolt

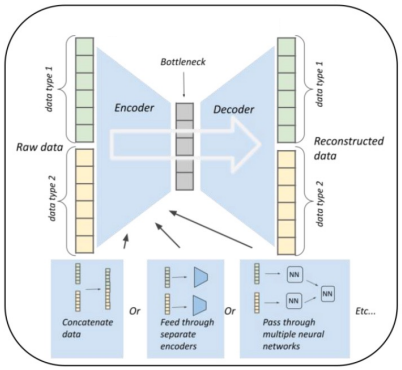
Diagonal integration: **different** cells/samples – **different** technology
GLUE, scJoint, scDART, FigR, MOFA+, scMoMat, MultiVI, SCOTCH

Lee et al., 2023

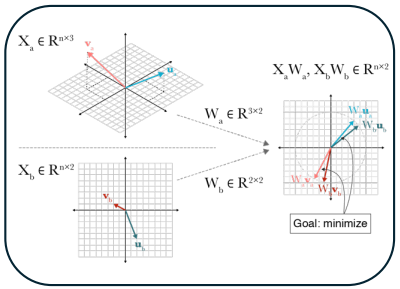
Data integration methods



Matrix factorization:
LIGER, **MOFA+,**
scMoMat

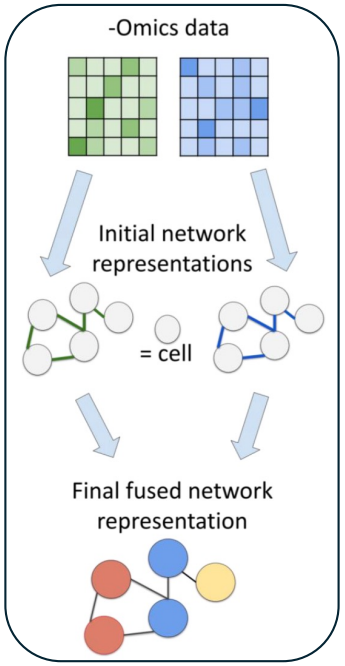


Neural networks:
GLUE, MultiVI
scJoint, scMVP,
scDART, MultiVI,
Cobolt



CCA based:
Seurat v3,
MOJITOO, **FigR**

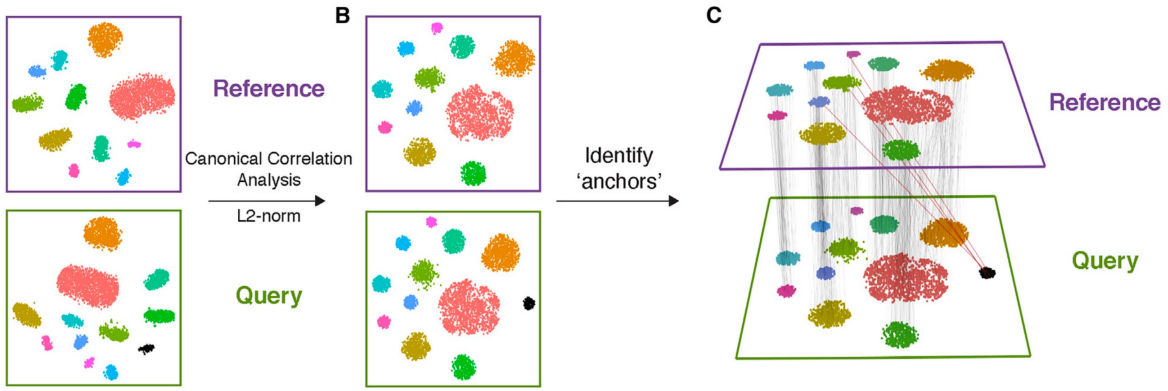
So many methods!!!
Choose what is fitting your current pipeline and comfortable to use!

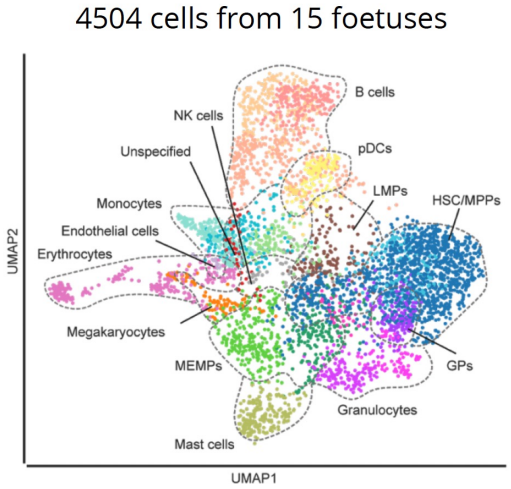


WNN, muon,
TREASMO(former scGREAT), SCOTCH
(with optimal transport)

Stanojevic et al., 2022

Integrate modalities/samples using “anchors”

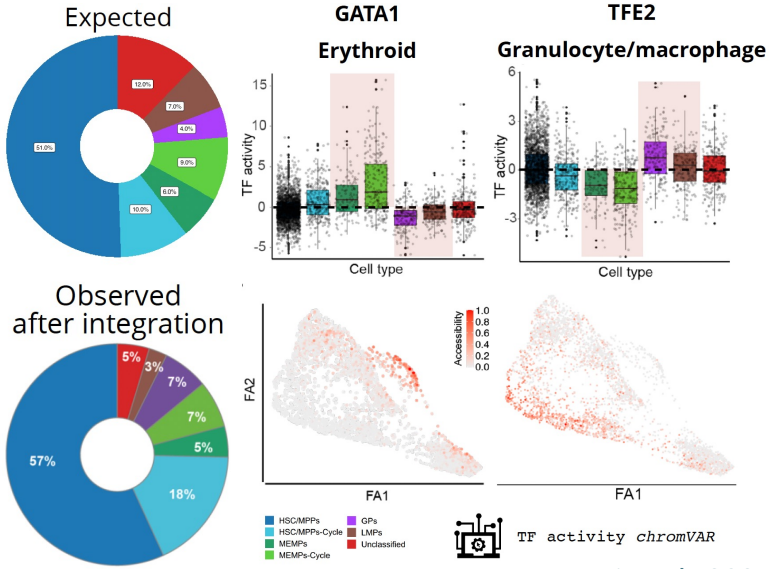




Transfer label between datasets of the same modality or different modalities using “anchors” (set of genes/peaks)

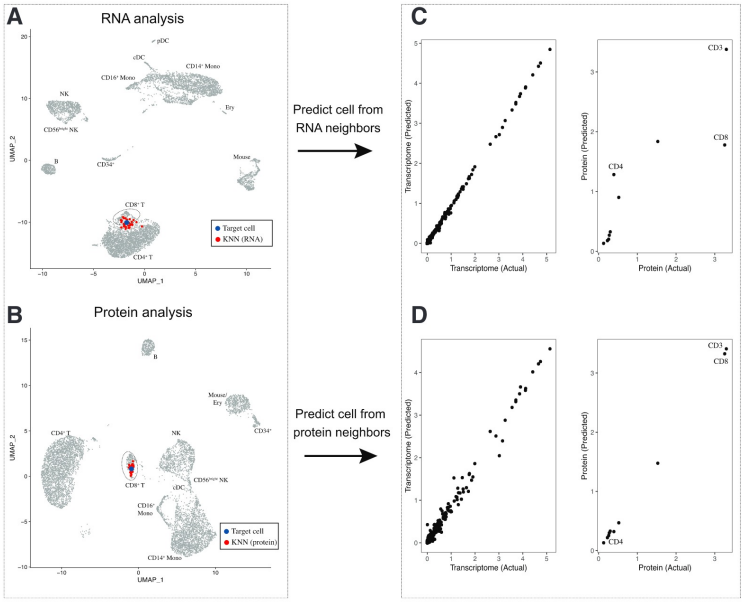
[Stuart et al., 2019]

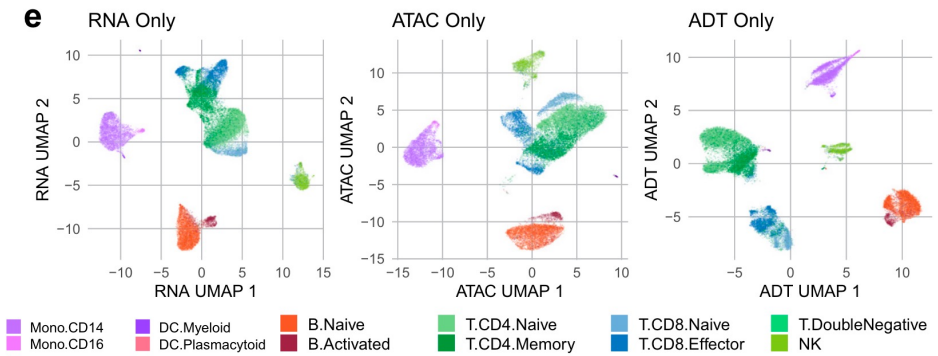
”Anchors” between ATAC and RNA?



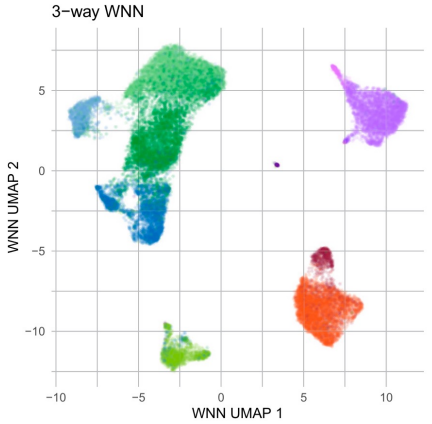
[Ranzoni et al., 2021]

Weighted nearest neighbors (WNN)





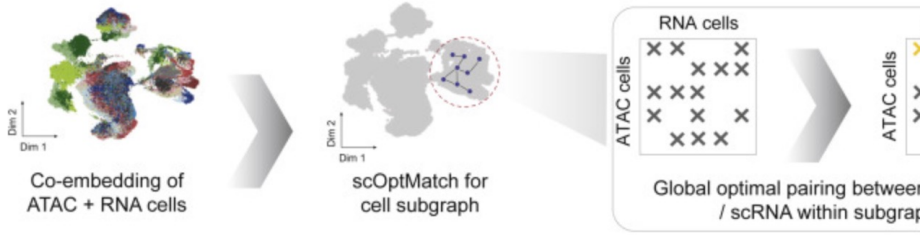
Swanson et al., 2021

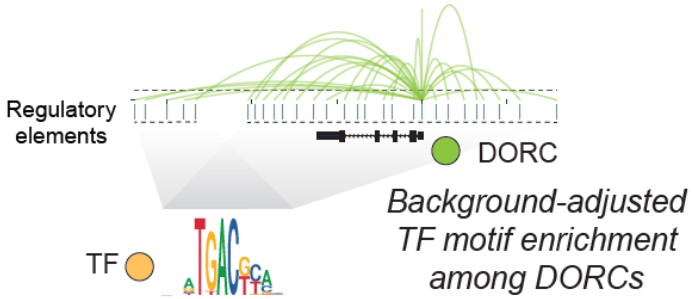


For each modality and each cell find neighbors on the knn graph and calculate weights for each modality -> create WNN graph using weighted information

Hao et al., 2021

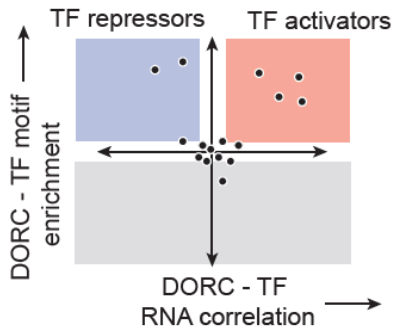
Functional inference of gene regulation (FigR)





Way of integrating **unpaired** scATAC and scRNA using scOptMatch (mix of CCA correlation and distance on KNN graph)

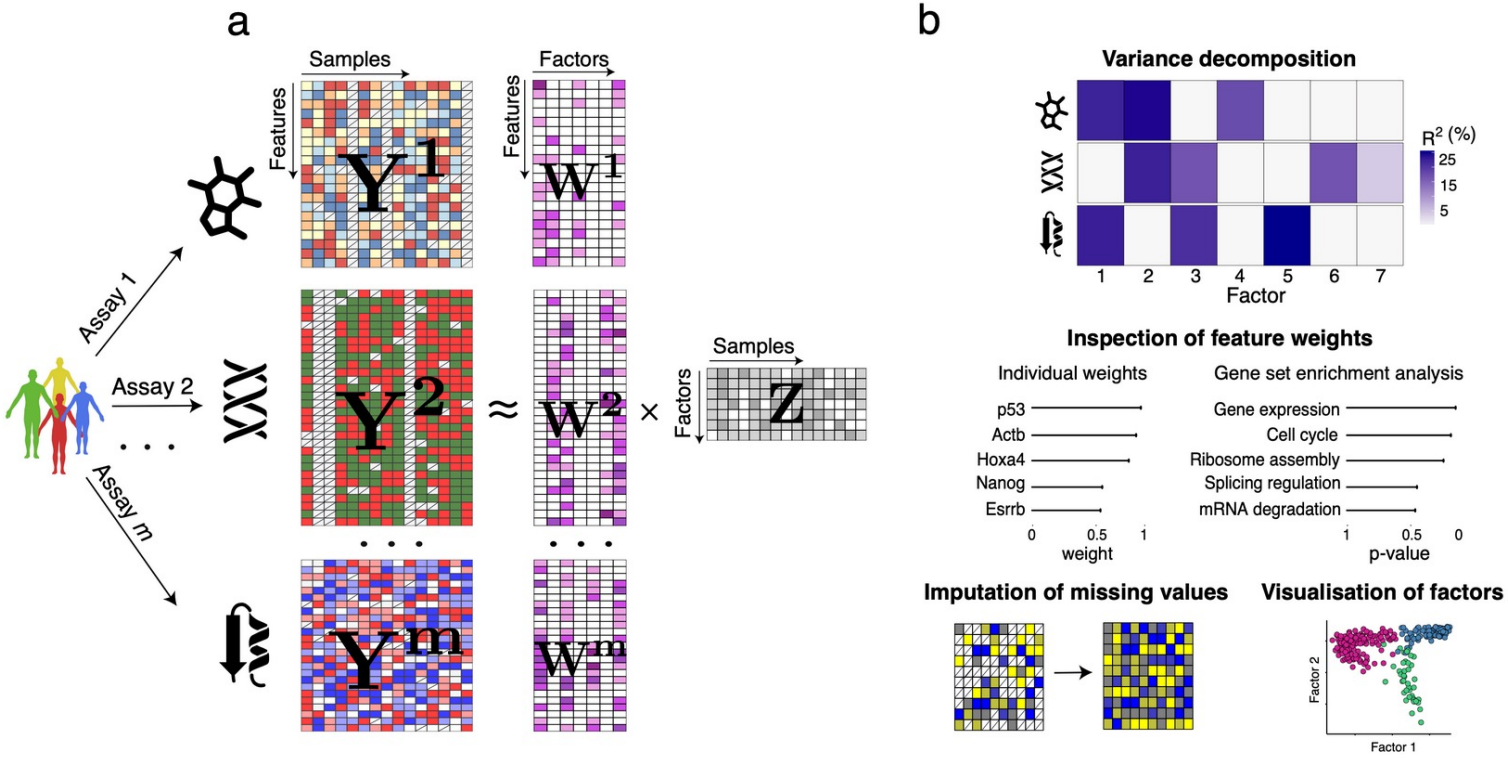
Introducing **DORC** (domains of regulatory chromatin): peak-gene links significantly correlated (ρ) and normalized by CG content and global accessibility



Kartha et al., 2022

MOFA+

Argelaguet et al., 2020



Probabilistic Bayesian framework summarizing feature space from modalities to a set of ”interpretable” factors (estimate weights per feature)

GLUE (old CLUE)

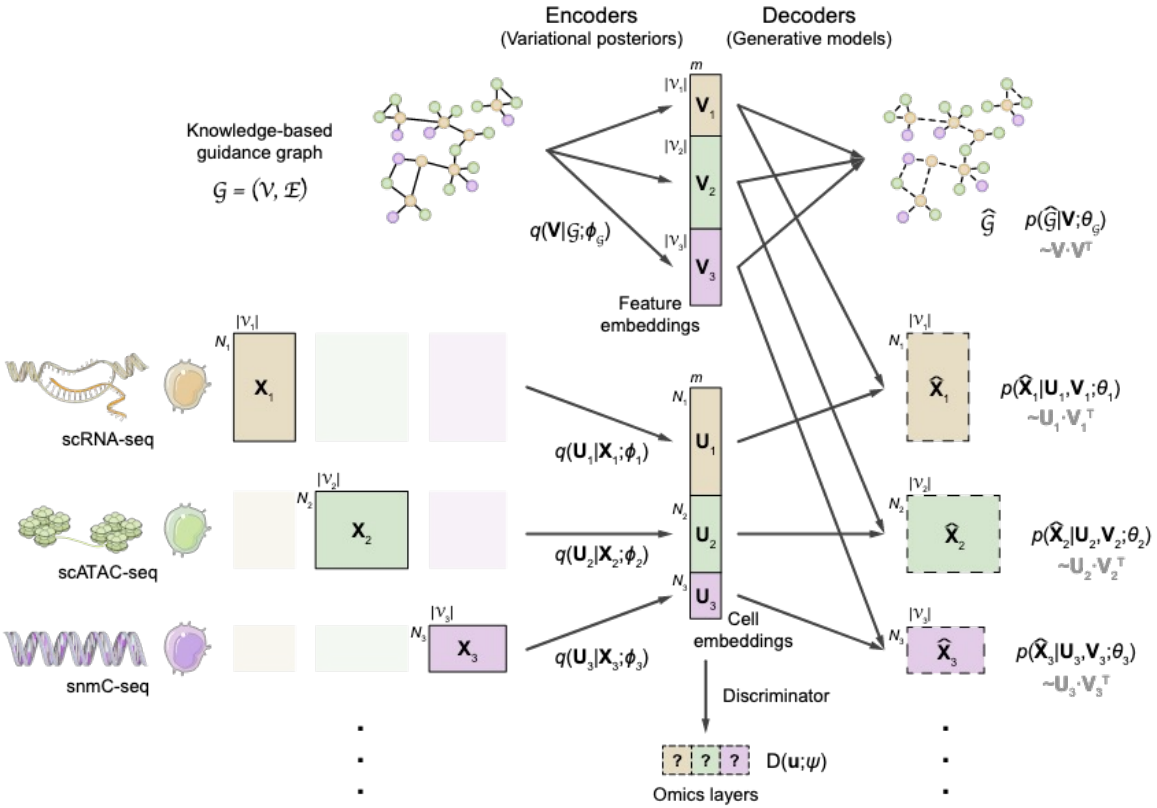
Task 2 - Match Modality

Given collections of cells in each modality, match profiles that originated from the same cell.

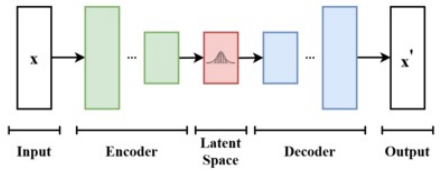
Winner in all categories: CLUE

Peking University, University of Washington, code

Zhi-Jie Cao, Xinming Tu, Chen-Rui Xia



Variational autoencoder (VAE) for each modality independently combined together with the guidance graph (prior knowledge of biology behind: openness in proximity of promoter – activation of the gene; DNA methylation – repress)



Cao & Gao, 2022

Interesting resources summarizing scATAC computational tools

- https://github.com/seandavi/awesome-single-cell (Epigenomics section)
- https://github.com/databio/awesome-atac-analysis (single-cell section)
- https://github.com/mdozmorov/scATAC-seq_notes
- https://www.singlecell.de/index.php/resources/software/ (scEpigenomics section)
- https://github.com/crazyhottommy/scATACseq-analysis-notes
- https://github.com/mikelove/awesome-multi-omics