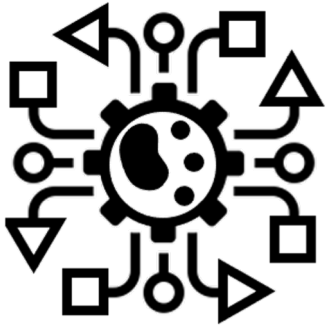


Introduction to single-cell multi-omics analysis

Advanced Topics in Single Cell Omics SciLifeLab-SIB Summer School 2021

Emma Dann
PhD @ Sanger Institute & EBI (UK)
ed6@sanger.ac.uk

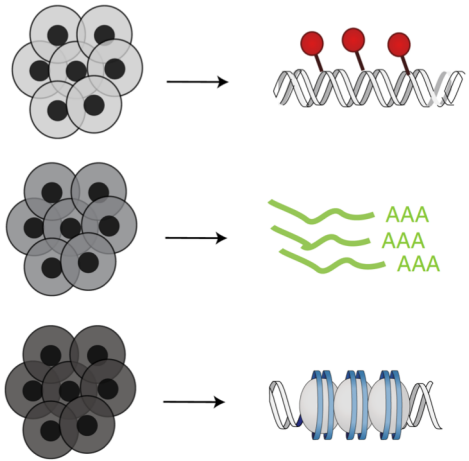
emdann @emmamarydann

What is single-cell multi-omics?

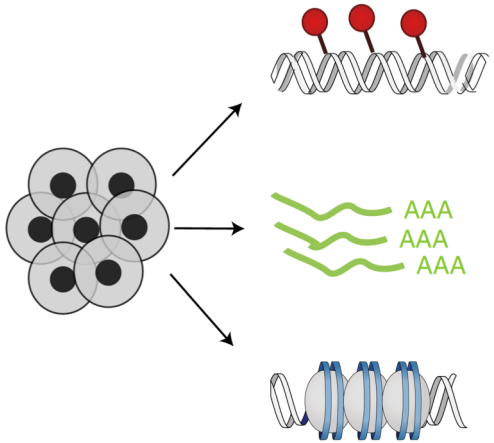
Joint analysis of two (or more!) datasets of measurements of different molecules from single-cells

What is single-cell multi-omics?

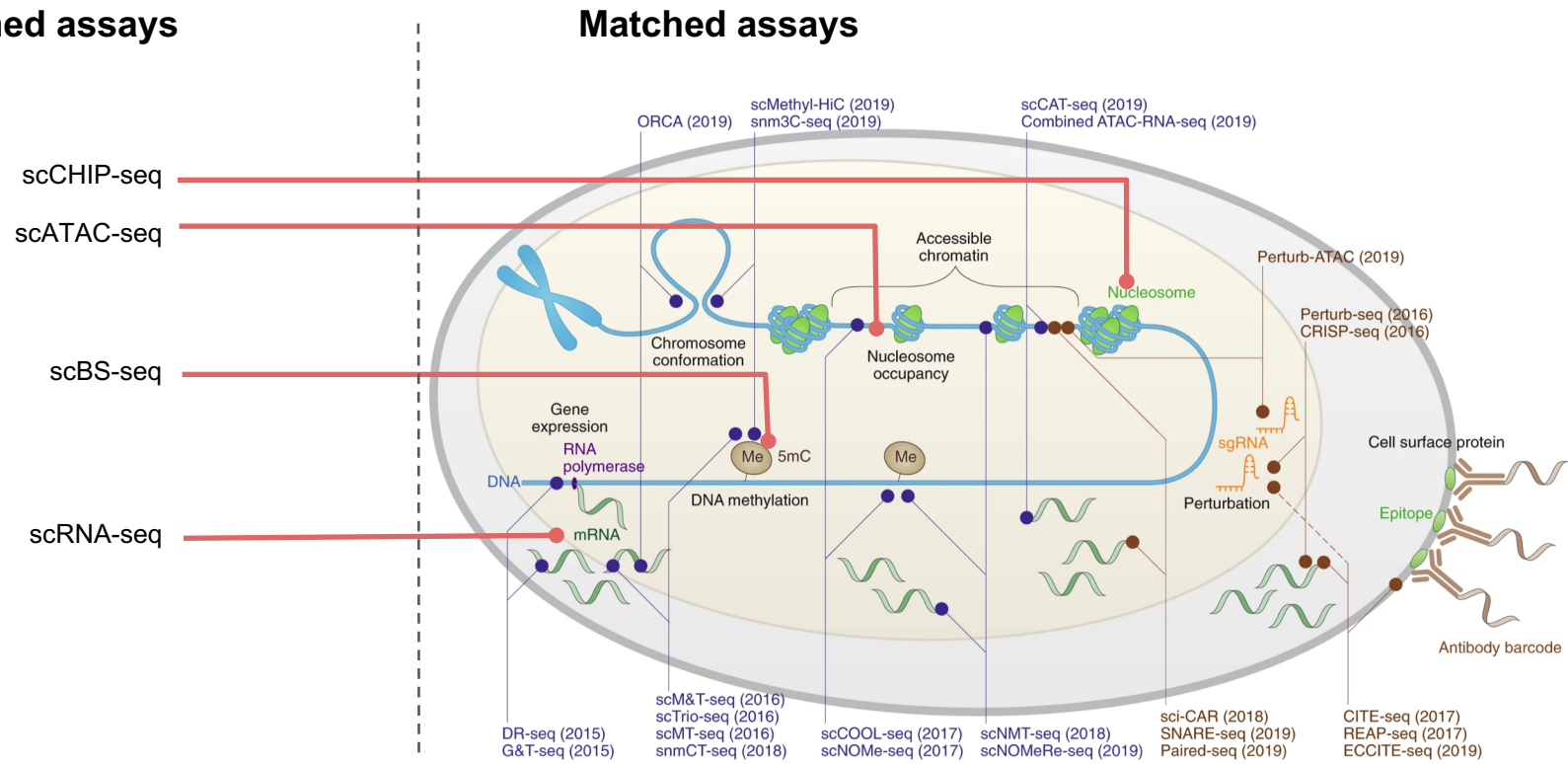
Unmatched assays



Matched assays

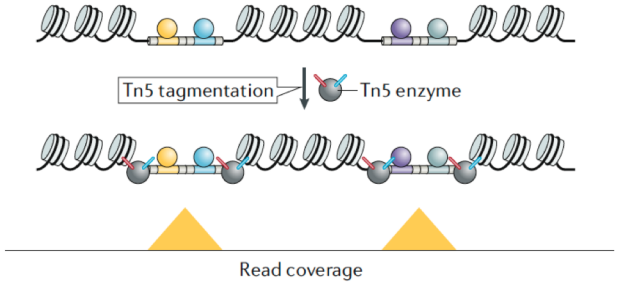


What is single-cell multi-omics?



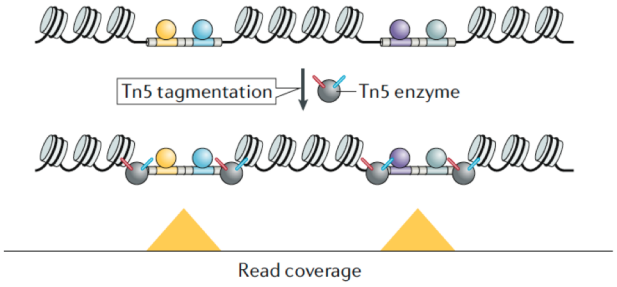
Zhu, Preissl & Ren Single-cell multimodal omics: the power of many, Nat Methods (2020)

scATAC-seq: chromatin accessibility



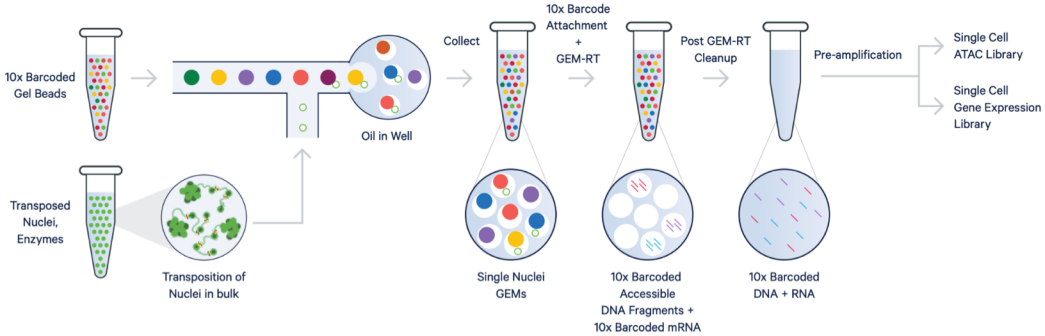
Minnoye et al. 2021 Chromatin accessibility profiling methods. Nat Rev Methods Primer

scATAC-seq: chromatin accessibility

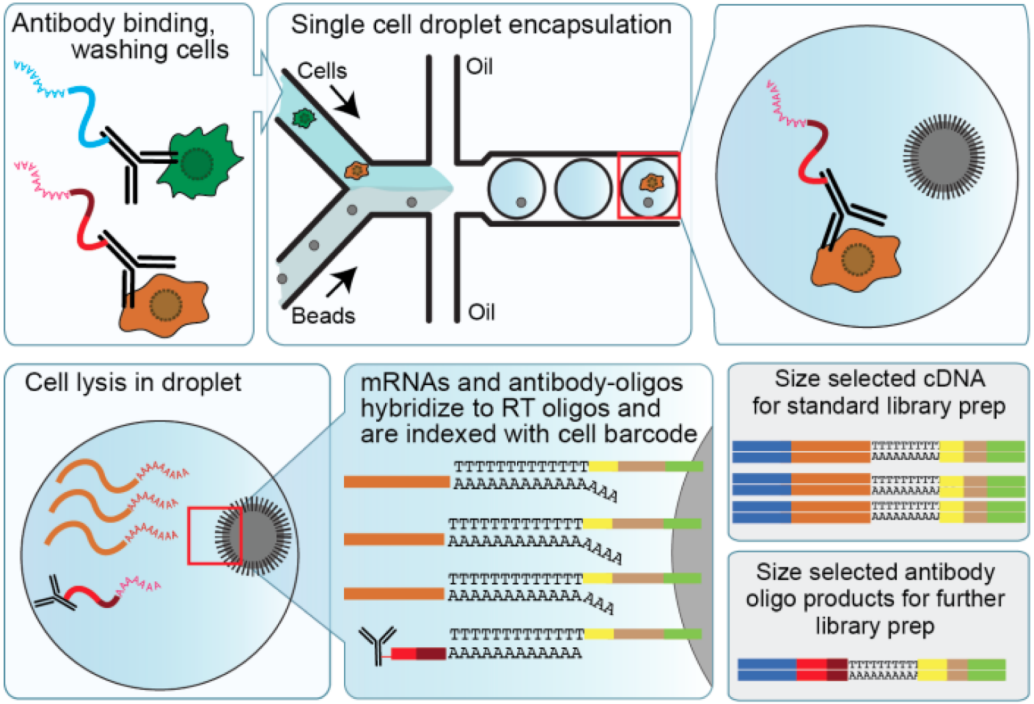


Minnoye et al. 2021 Chromatin accessibility profiling methods. Nat Rev Methods Primer

10X Genomics Multiome (scRNA+scATAC)

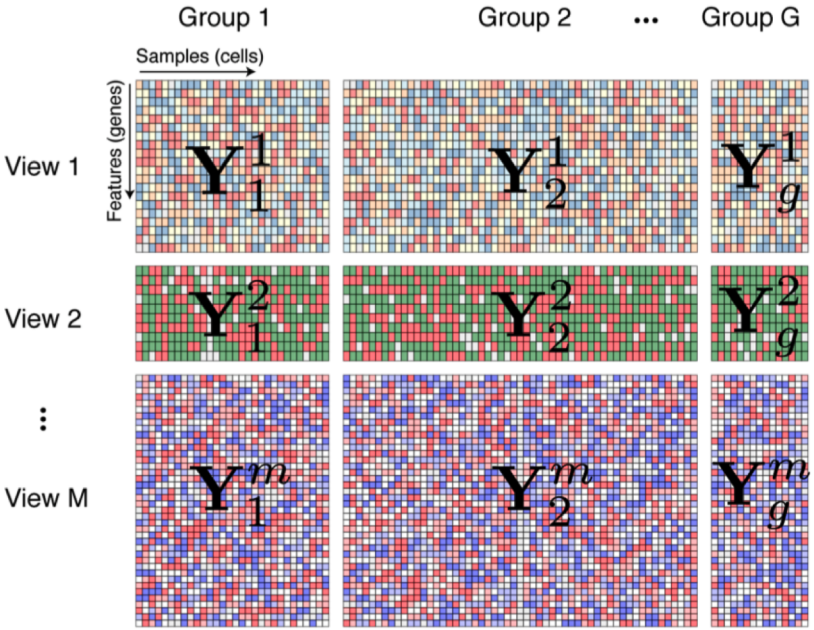


CITE-seq: mRNA expression and surface proteins



Stoeckius et al. (2017) Simultaneous epitope and transcriptome measurement in single cells

What does the data look like?



Common multi-omic analysis goals

A. **Verifying consensus across modalities**

A. **Co-embedding in meaningful latent space**

A. **Reconstructing missing/noisy data**

A. **Identifying statistical relationships between features**

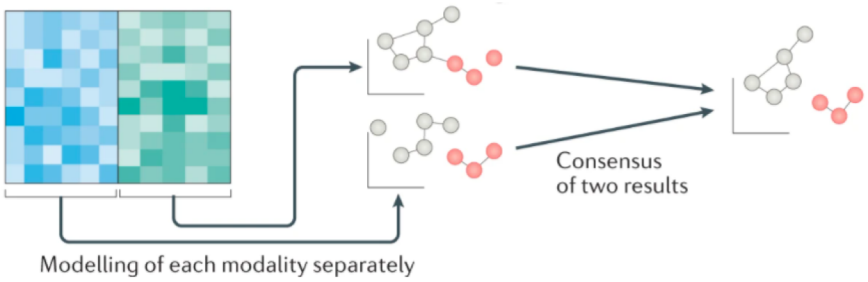
Common multi-omic analysis goals

A. **Verifying consensus across modalities**

A. Co-embedding in meaningful latent s

A. Reconstructing missing/noisy data

A. Identifying statistical relationships between features

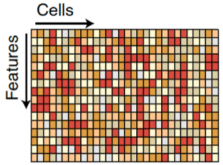


scATAC-seq preprocessing: from fragments to KNN graph

Raw data
(fragments.tsv.gz)

hg19_chr1	16205	16281	TTATGTCGTCTCAAAC-1	1
hg19_chr1	17124	17503	TGAGCCGGTATACGCT-1	1
hg19_chr1	235668	235711	CTTAATCCAAATAGTG-1	1
hg19_chr1	237712	237828	TCCGACTTCTTACGGA-1	1
hg19_chr1	237713	237792	TAGTCCCGTTAACTCG-1	1
hg19_chr1	237716	237782	GCCATAAGTGATCAGG-1	1
hg19_chr1	237716	237789	CCAATGATCCATCGAA-1	1
hg19_chr1	237721	237756	TGCGTAACAGGTGGTA-1	1
hg19_chr1	237722	237793	CCCAGAGCAAAGCTTC-1	1
hg19_chr1	237736	237782	GACCTTCTCACTGATG-1	3
hg19_chr1	521557	521596	AGATTCGGTTCTCGAA-1	1
hg19_chr1	521575	521611	TCACCACGTCCGTGCA-1	2
hg19_chr1	526022	526082	TGATGCAAGCCGCTGT-1	1
hg19_chr1	540966	541013	GTAGACTTCGTGGAAG-1	1
hg19_chr1	563390	563788	ACTGCAATCGTCCCAT-1	1
hg19_chr1	565288	565342	TCTCTGGTCCTGAAAC-1	2
hg19_chr1	565293	565322	TGAGCCGGTATACGCT-1	2

Tabular data

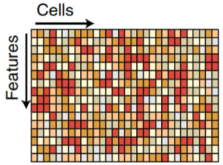


scATAC-seq preprocessing: from fragments to KNN graph

Raw data
(fragments.tsv.gz)

hg19_chr1	16205	16281	TTATGTCGTCTCAAAC-1	1
hg19_chr1	17124	17503	TGAGCCGGTATACGCT-1	1
hg19_chr1	235668	235711	CTTAATCCAAATAGTG-1	1
hg19_chr1	237712	237828	TCCGACTTCTTACGGA-1	1
hg19_chr1	237713	237792	TAGTCCCGTTAACTCG-1	1
hg19_chr1	237716	237782	GCCATAAGTGATCAGG-1	1
hg19_chr1	237716	237789	CCAATGATCCATCGAA-1	1
hg19_chr1	237721	237756	TGCGTAACAGGTGGTA-1	1
hg19_chr1	237722	237793	CCCAGAGCAAAGCTTC-1	1
hg19_chr1	237736	237782	GACCTTCTCACTGATG-1	3
hg19_chr1	521557	521596	AGATTCGGTTCTCGAA-1	1
hg19_chr1	521575	521611	TCACCACGTCCGTGCA-1	2
hg19_chr1	526022	526082	TGATGCAAGCCGCTGT-1	1
hg19_chr1	540966	541013	GTAGACTTCGTGGAAG-1	1
hg19_chr1	563390	563788	ACTGCAATCGTCCCAT-1	1
hg19_chr1	565288	565342	TCTCTGGTCCTGAAAC-1	2
hg19_chr1	565293	565322	TGAGCCGGTATACGCT-1	2

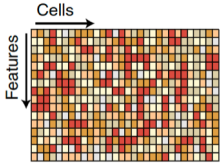
Tabular data



- Binning the genome into equally sized windows (10-50kb)
- Peak calling on pseudo-bulk profiles (MACS2)
  - Pseudo-bulk on first pass clustering on genomic bins
- Using known annotations for enhancers (e.g. in Drosophila genome)
- Other scATAC-specific feature extraction methods (BROCKMAN, scRegSeg)

scATAC-seq preprocessing: from fragments to KNN graph

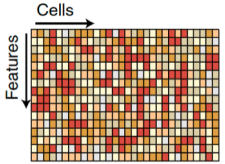
Tabular data



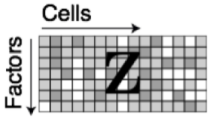
- *extreme* sparsity
- > 100k features
- Practically binary (most values are 1 or 0)

scATAC-seq preprocessing: from fragments to KNN graph

Tabular data



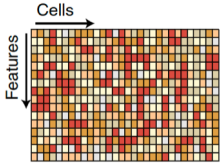
Reduced dimensions



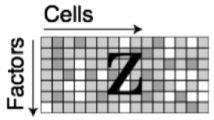
- *extreme* sparsity
- > 100k features
- Practically binary (most values are 1 or 0)

scATAC-seq preprocessing: from fragments to KNN graph

Tabular data

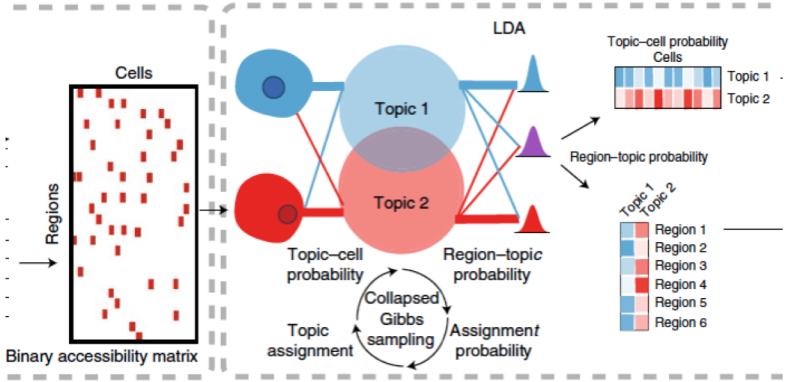


Reduced dimensions



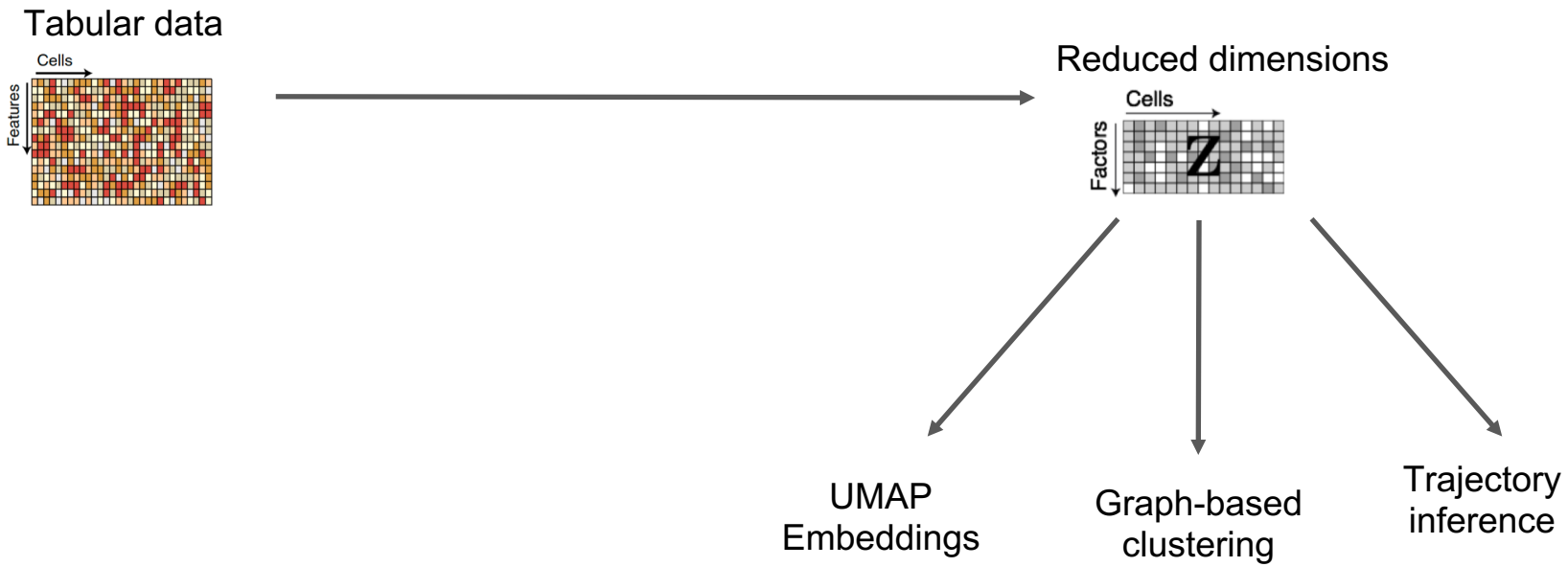
Adaptations of models used in text processing for topic extraction

- Latent Semantic Indexing
- Latent Dirichlet Allocation (cisTopic)



Bravo-Gonzales et al. (2019) cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data

scATAC-seq preprocessing: from fragments to KNN graph



Any questions?

Common multi-omic analysis goals

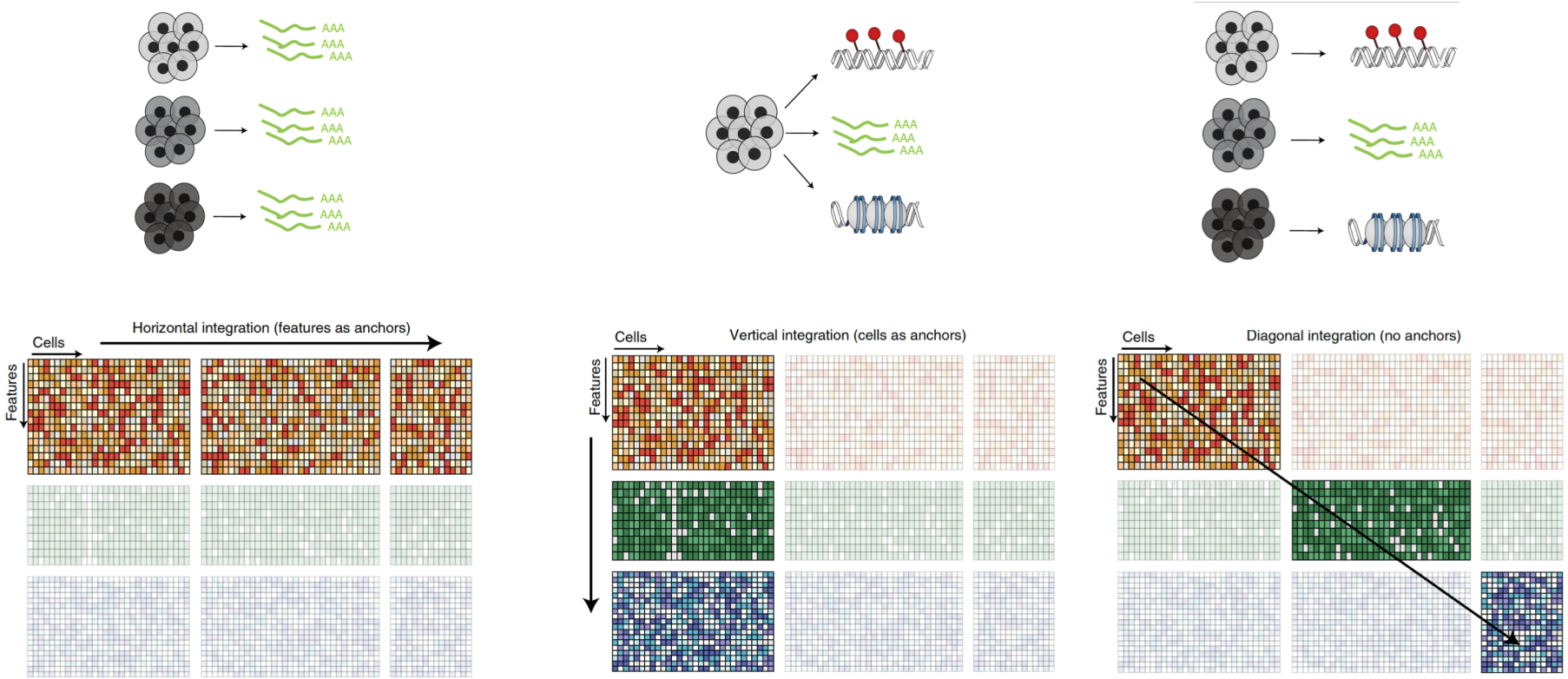
A. Verifying consensus across modalities

A. Co-embedding in meaningful latent space (*integration*)

A. Reconstructing missing/noisy data

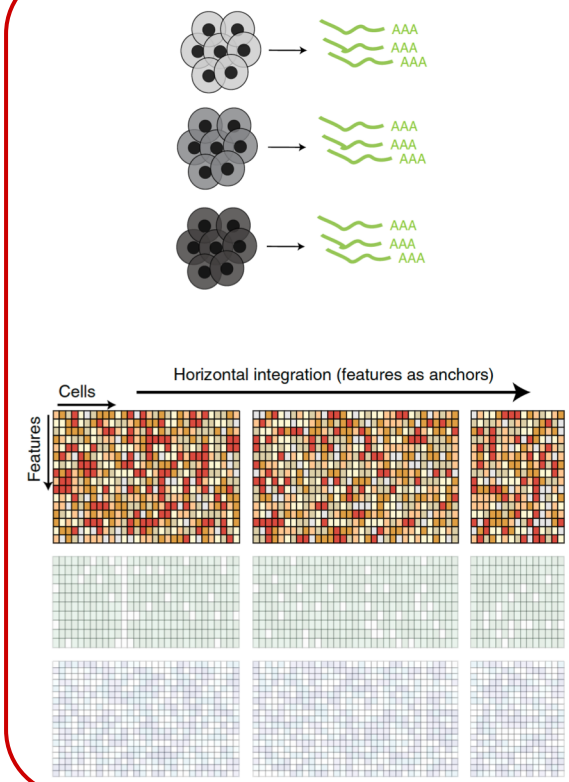
A. Identifying statistical relationships between features

Defining the integration axis

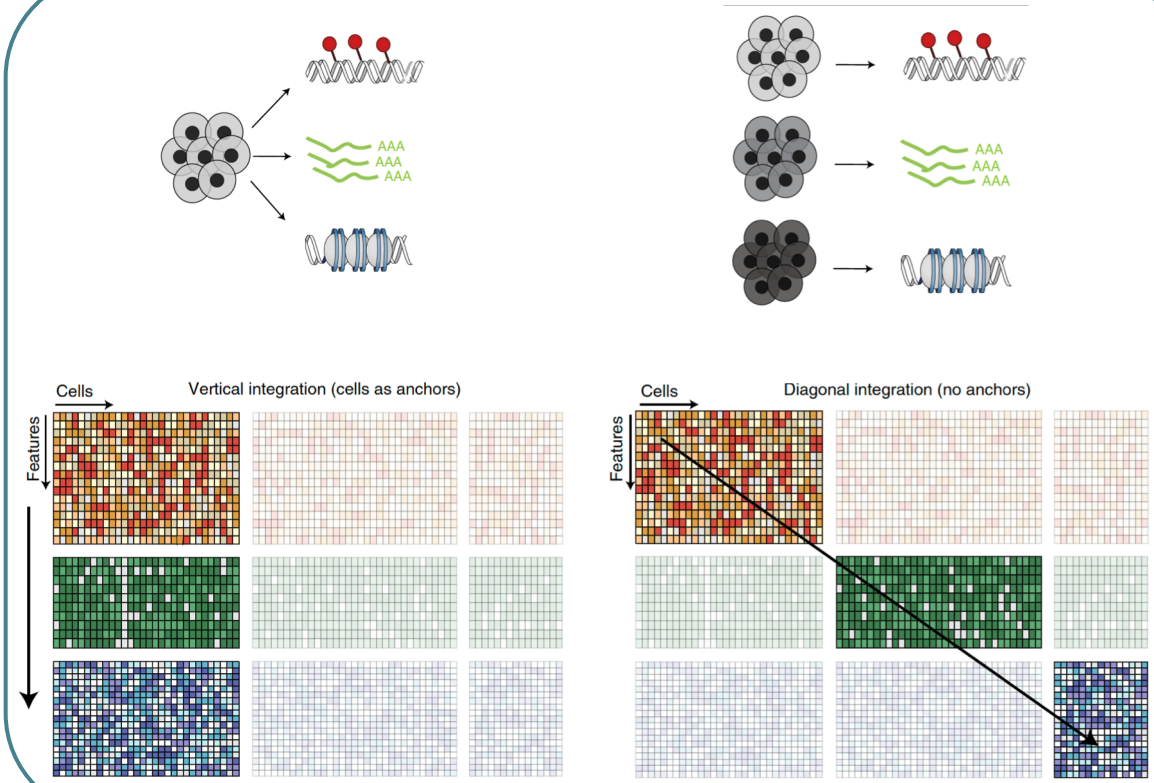


Argelaguet, Cuomo, Stegle and Marioni (2021) Computational principles and challenges in single-cell data integration. Nat Biotech

Defining the integration axis

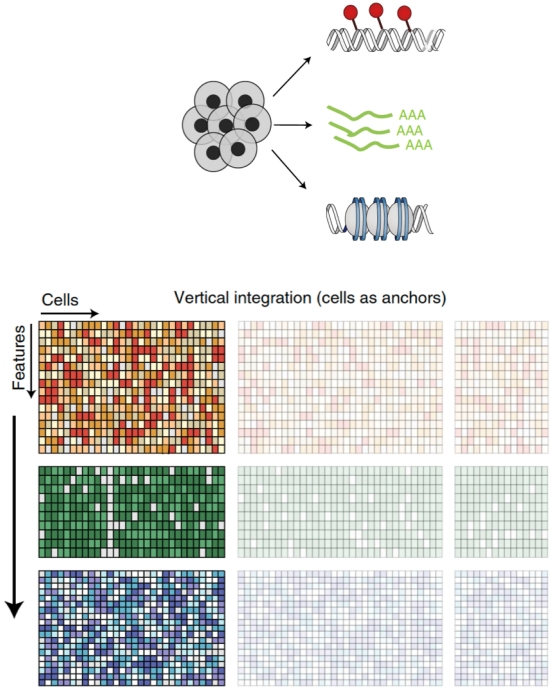


Batch correction, mapping to reference atlas

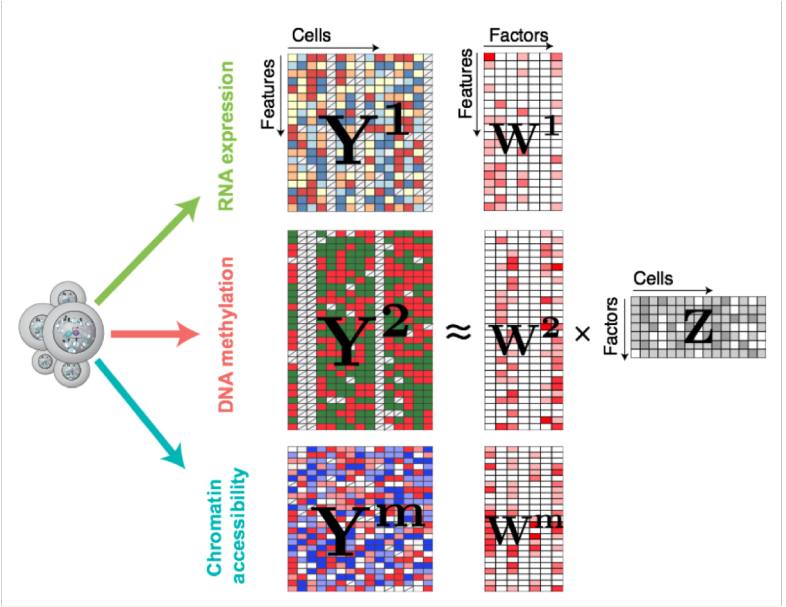


Multi-omics analysis

Vertical integration of matched multi-omics data



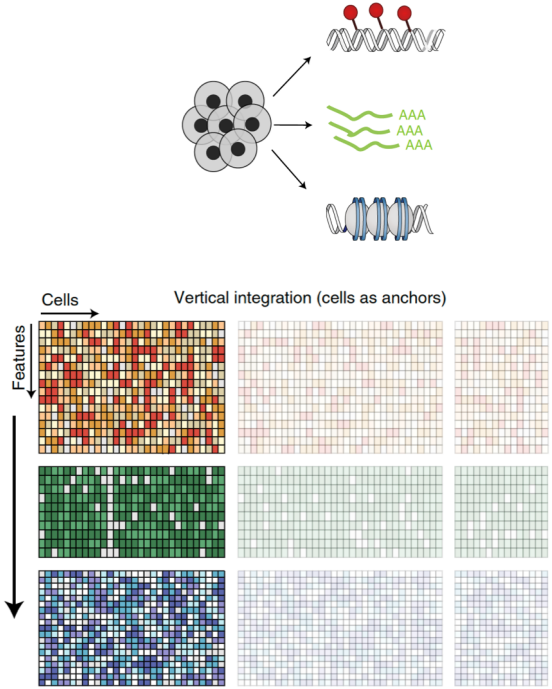
Multi Omics Factor Analysis (MOFA2)



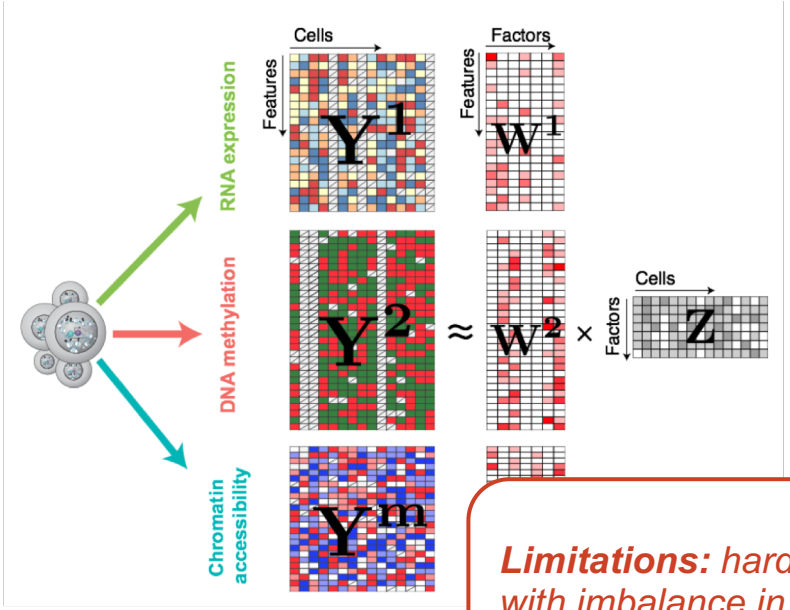
$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT}$$

Argelaguet, Velten et al. Mol Sys Biol 2018
Argelaguet, Arnol, Bredikhin et al. Genome Biology 2020

Vertical integration of matched multi-omics data



Multi Omics Factor Analysis (MOFA2)

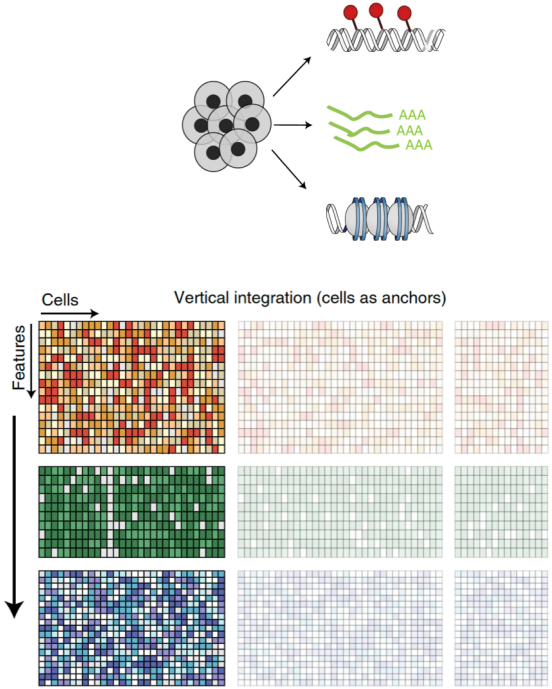


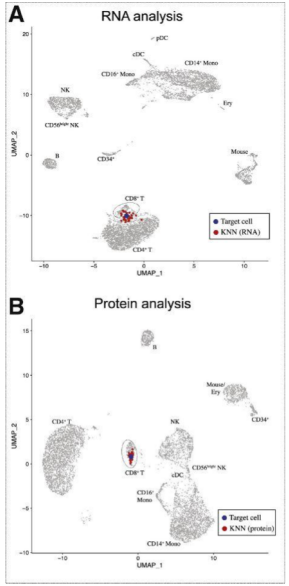
$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT}$$

Limitations: *hard to deal with imbalance in number of features between views*

Argelaguet, Velten et al. Mol Sys Biol 2018
Argelaguet, Arnol, Bredikhin et al. Genome Biology 2020

Vertical integration of matched multi-omics data

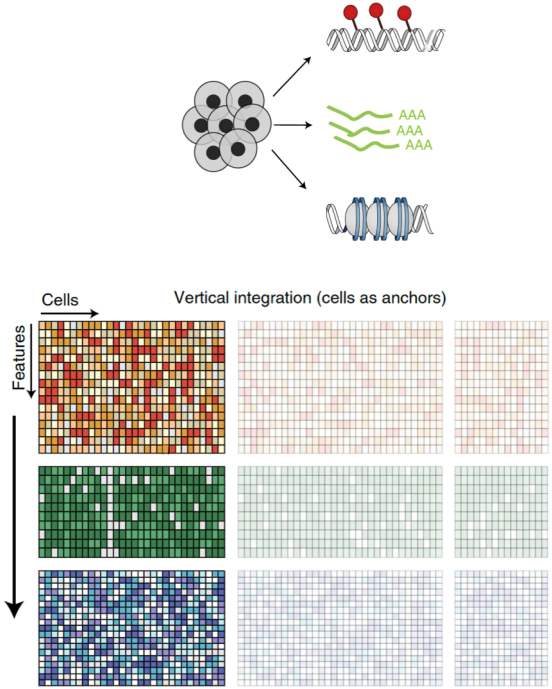




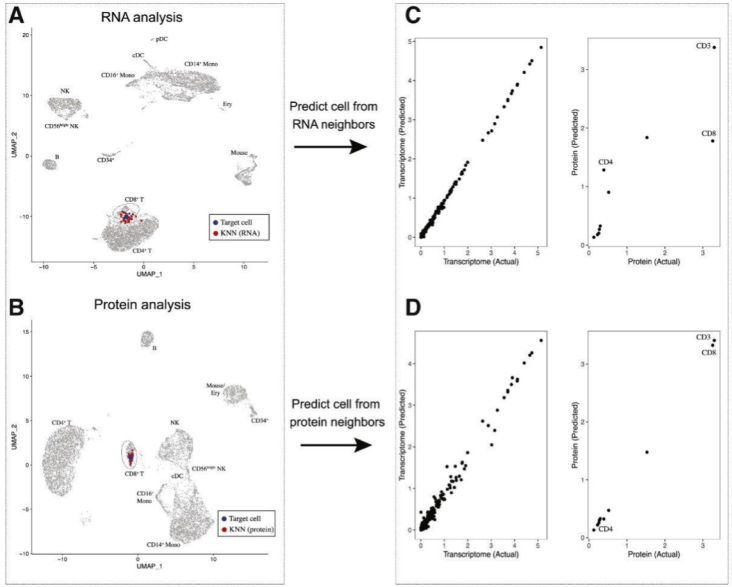
Weighted Nearest Neighbor (WNN) analysis

Hao, Hao et al. Cell 2021

Vertical integration of matched multi-omics data

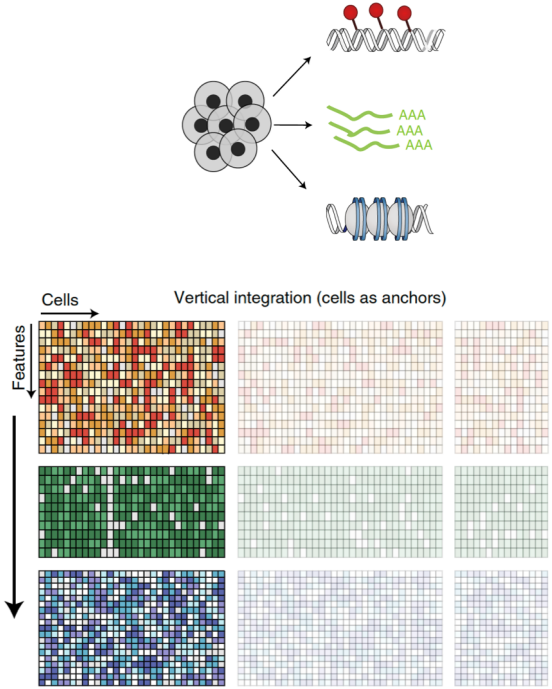


Weighted Nearest Neighbor (WNN) analysis

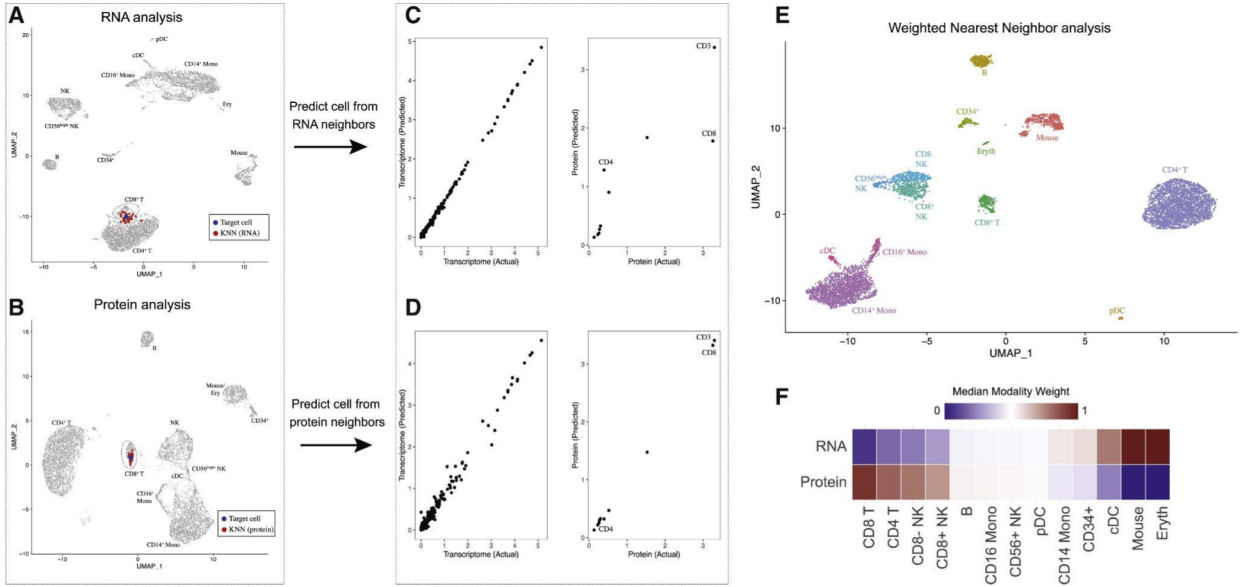


Hao, Hao et al. Cell 2021

Vertical integration of matched multi-omics data

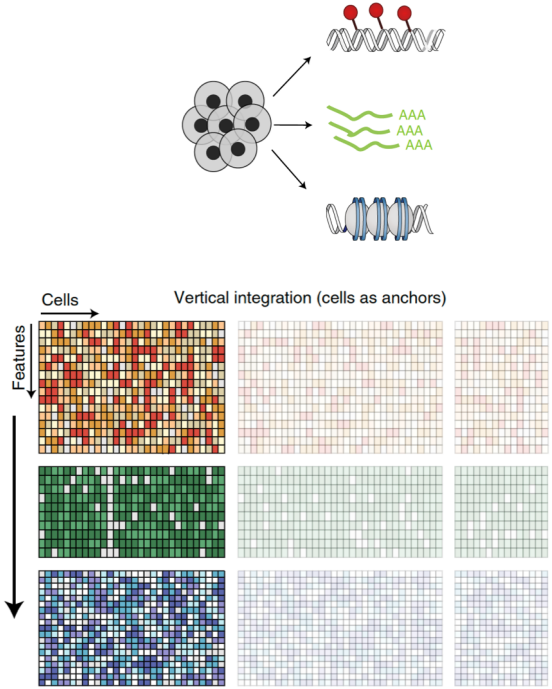


Weighted Nearest Neighbor (WNN) analysis

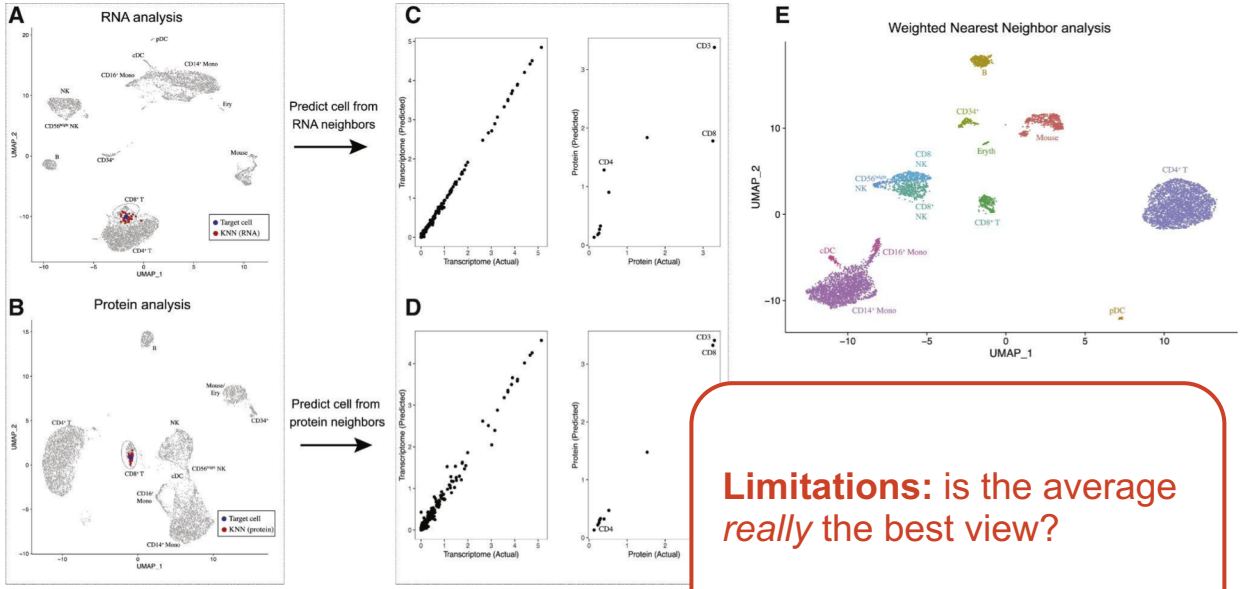


Hao, Hao et al. Cell 2021

Vertical integration of matched multi-omics data



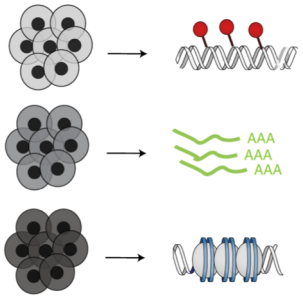
Weighted Nearest Neighbor (WNN) analysis



Limitations: is the average *really* the best view?

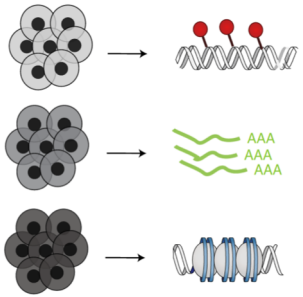
Hao, Hao et al. Cell 2021

Diagonal integration of unmatched multi-omics data

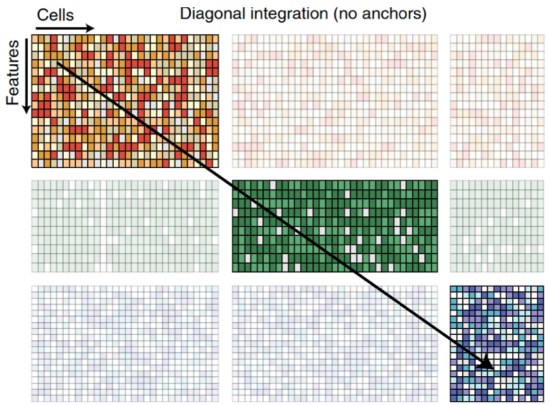


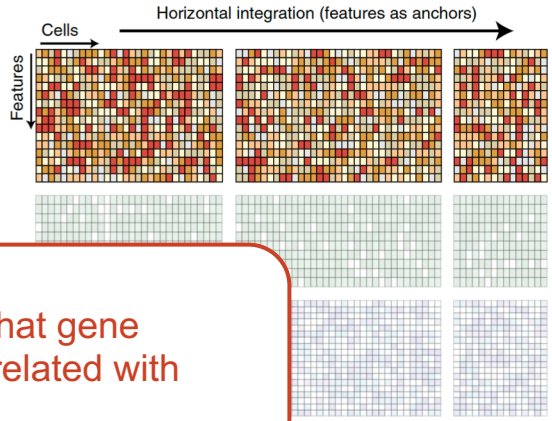
- Transform data to gene-level features (e.g. count ATAC fragments over gene bodies)
- Apply horizontal integration methods used for batch correction (Seurat CCA, LIGER)

Diagonal integration of unmatched multi-omics data



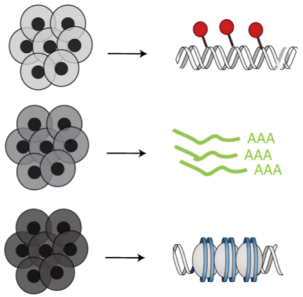
- Transform data to gene-level features (e.g. count ATAC fragments over gene bodies)
- Apply horizontal integration methods used for batch correction (Seurat CCA, LIGER)

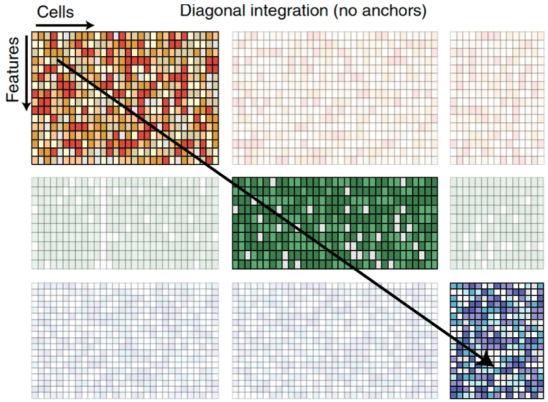




Limitations: assumption that gene accessibility is linearly correlated with gene expression

Diagonal integration of unmatched multi-omics data

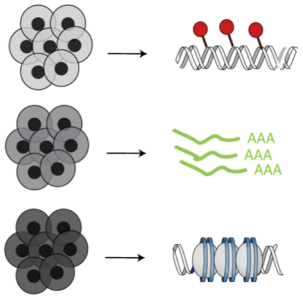


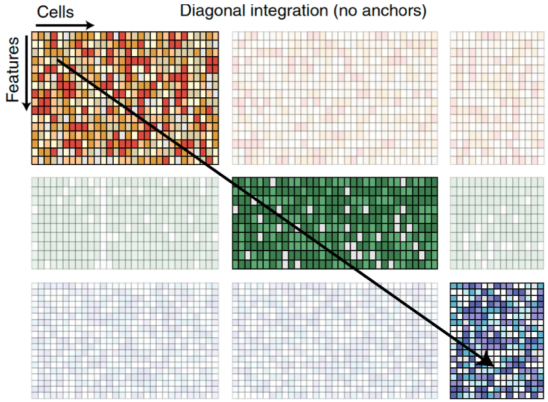


Integration with unpaired features
(in order of appearance on bioRxiv)

- MATCHER (Welch et al. 2017)
- MMD-MA (Liu et al. 2019)
- SCIM (Stark et al. 2020)
- UnionCom (Cao et al. 2020)
- Cross-modality autoencoders (Yang et al. 2021)
- SCOT (Demetci et al. 2020)
- BABEL (Wu et al. 2020)
- bindSC (Dou et al. 2020)
- MultiMAP (Jain et al. 2021)
- UINMF (Kriebel et al. 2021)
- MultiVI (Ashuach et al. 2021)
- ...

Diagonal integration of unmatched multi-omics data





Integration with unpaired features
(in order of appearance on bioRxiv)

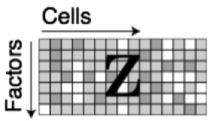
- MATCHER (Welch et al. 2017)
- MMD-MA (Liu et al. 2019)
- SCIM (Stark et al. 2020)
- UnionCom (Cao et al. 2020)
- Cross-modality autoencoders (Yang et al. 2021)
- SCOT (Demetci et al. 2020)
- BABEL (Wu et al. 2020)
- bindSC (Dou et al. 2020)
- MultiMAP (Jain et al. 2021)
- UINMF (Kriebel et al. 2021)
- MultiVI
- ...

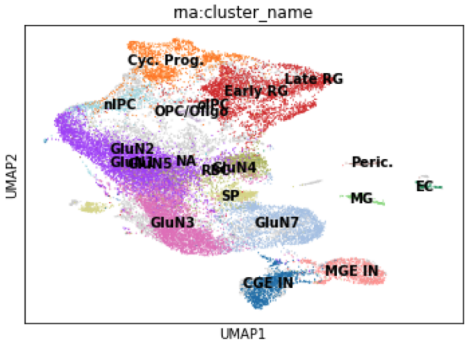
Limitations: assumption that cells lie on the same latent manifold

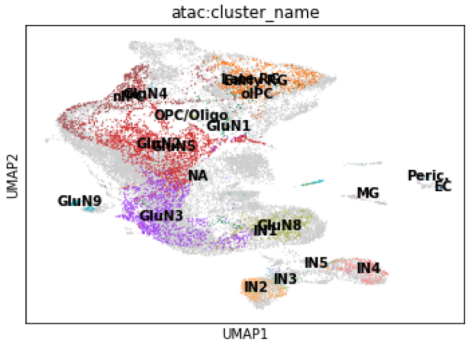
Any questions?

Except for: which integration method is the best

Outcome: co-embedding in joint latent space







Transferring cell type labels

Pseudotime ordering

Imputation of missing data

Common multi-omic analysis goals

A. Verifying consensus across modalities

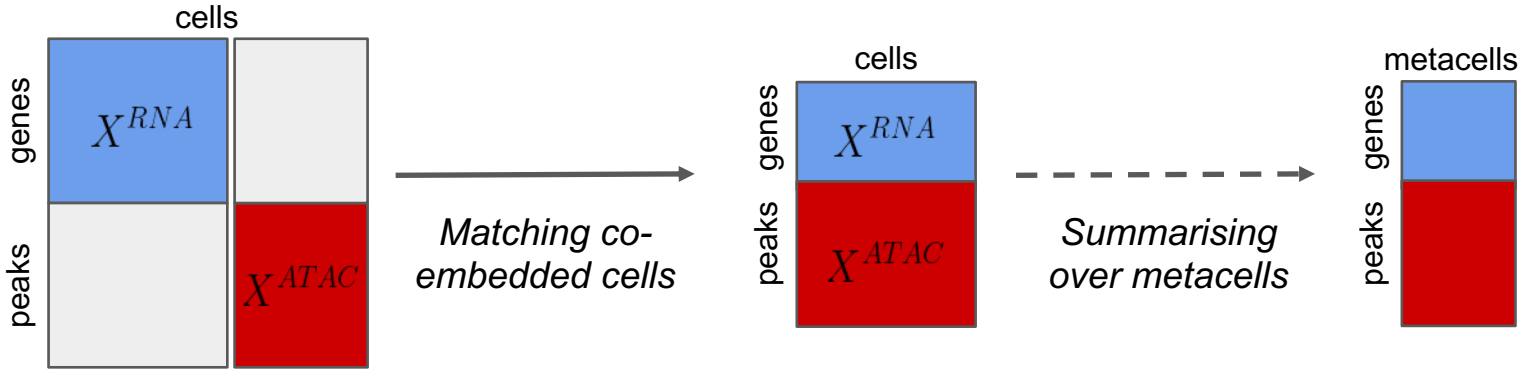
A. Co-embedding in meaningful latent space

A. Reconstructing missing/noisy data

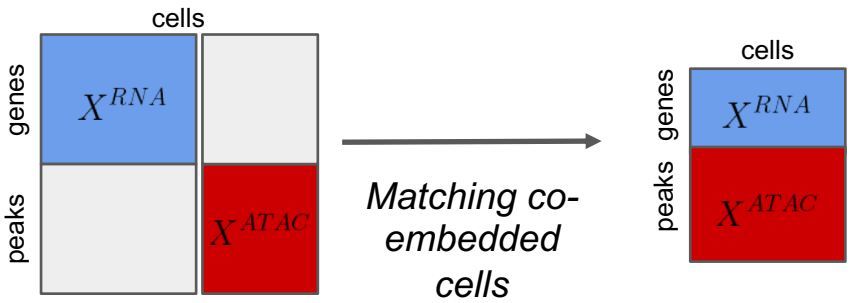
A. Identifying statistical relationships between features

Preprocessing for feature-wise analysis

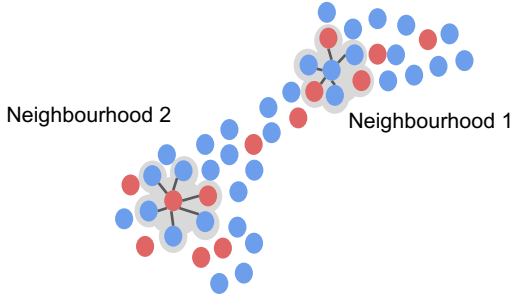
$$X_g^{RNA} = f(X_p^{ATAC})$$



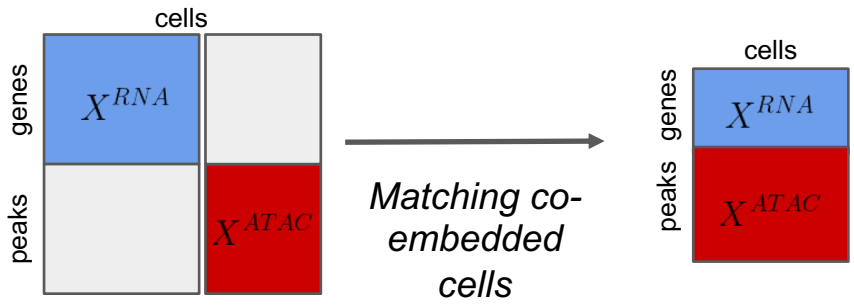
Preprocessing for feature-wise analysis



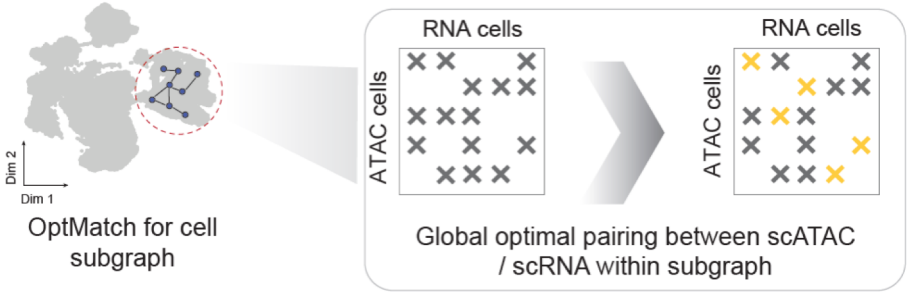
- Impute expression for scATAC cells as average of K-nearest neighbors



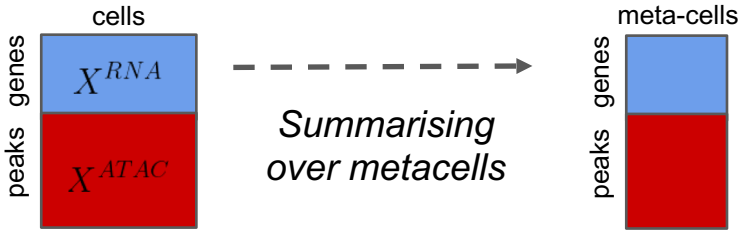
Preprocessing for feature-wise analysis



- Impute expression for scATAC cells as average of K-nearest neighbors
- Optimal matching of RNA and ATAC cells
  - Seurat anchors
  - Minimum-Cost Maximum-Flow bipartite graph matching (Stark et al. 2020 - https://github.com/ratschlab/scim)
  - OptMatch (Kartha et al. 2021 - https://github.com/buenrostrolab/stimATAC_analyses_code)



Preprocessing for feature-wise analysis



- Subsample (to representative or *optimally matched* cells)
- (Over)clustering
- Aggregate over KNN graph neighbourhoods
  - MetaCell (Baran et al. 2018 - https://github.com/tanaylab/metacell)
  - Milo (Dann et al. 2020 - https://github.com/MarioniLab/miloR)

Common multi-omic analysis goals

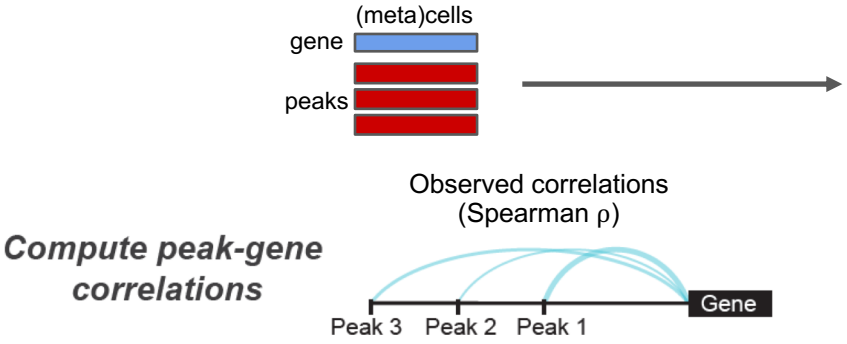
A. Verifying consensus across modalities

A. Co-embedding in meaningful latent space

A. Reconstructing missing/noisy data

A. **Identifying statistical relationships between features**

Finding statistical relationships between features

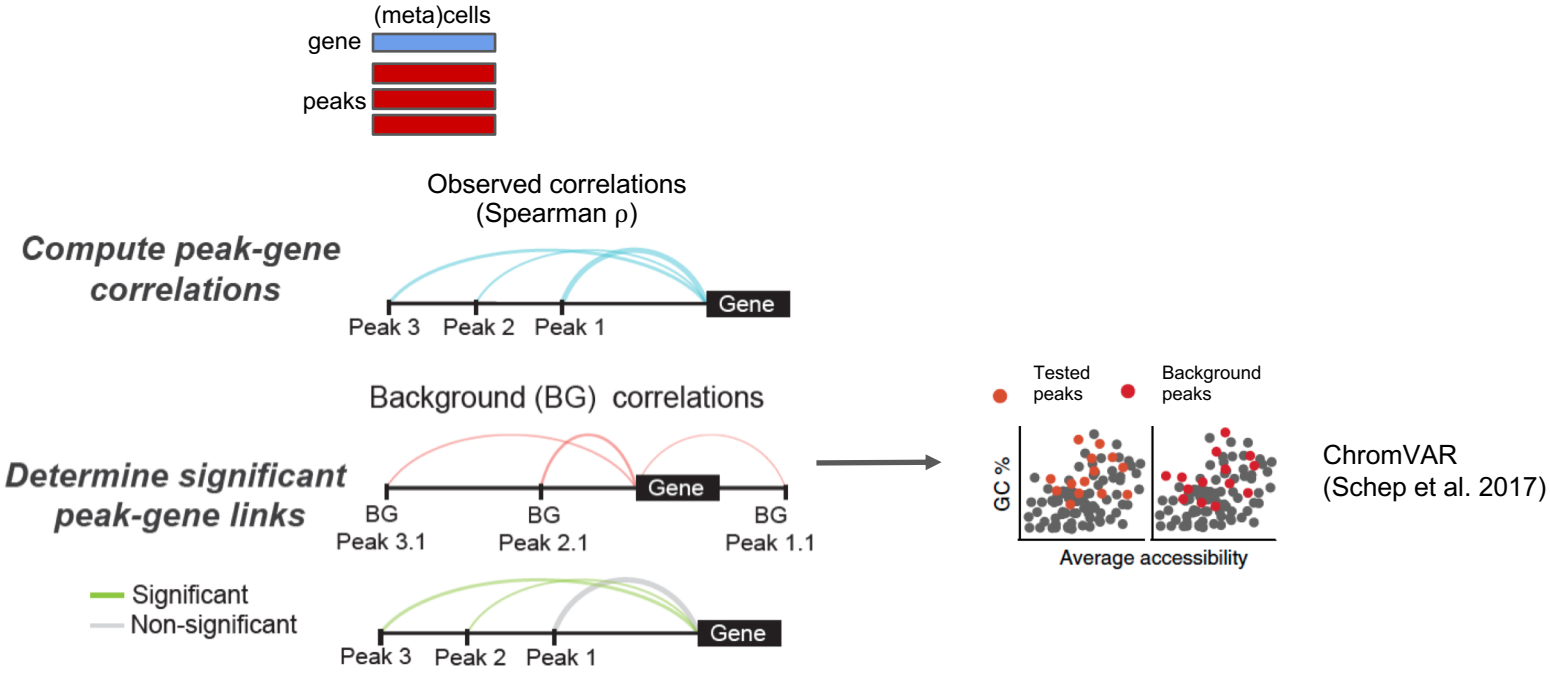


Feature selection

- Which genes? E.g. HVGs, marker genes, dynamic genes in pseudotime, …
- Which accessibility features? Should I aggregate peaks e.g. by TF motifs or genomic locus?
- Which feature pairs?

Ma et al. 2020, Kartha et al. 2021

Finding statistical relationships between features



Ma et al. 2020, Kartha et al. 2021

Downstream interpretation of peak-gene links

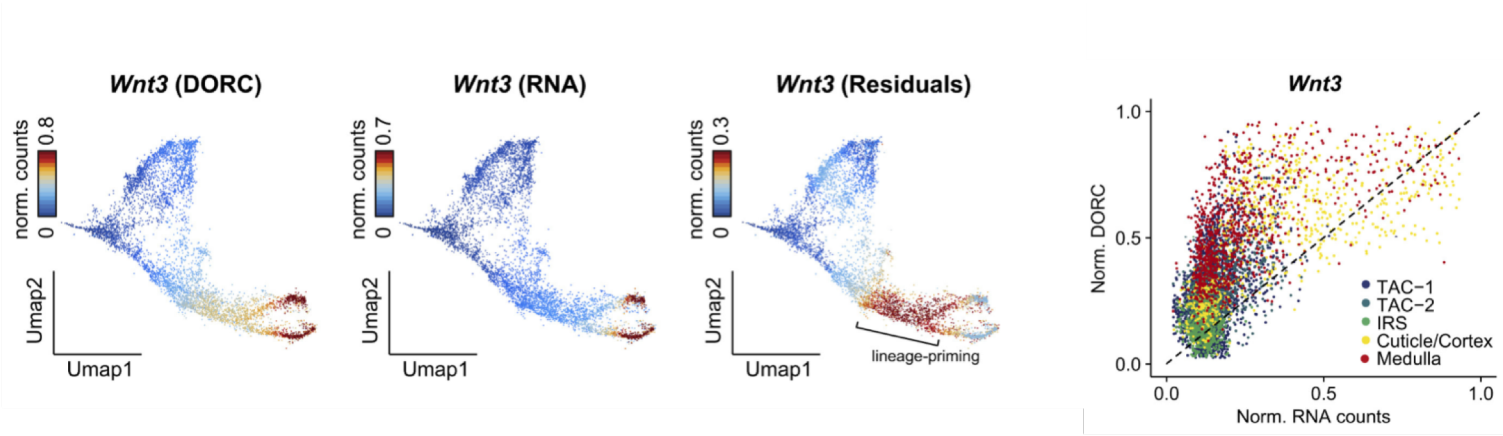
- **Validation:** Which peaks do we expect to be enriched in links? → Transcription Start Sites, enrichment in motifs for variable TFs
- Which genes show most regulatory elements linked?
- Pruning GRN inference links (e.g. SCENIC, CellOracle)
- Interpretation of GWAS hits

Working with multi-modal data

- Muon – python – extension of AnnData PMBio/muon
- MultiAssayExperiment – R/Bioconductor - extension of SummarizedExperiment waldronlab/MultiAssayExperiment
- Seurat v4/Signac – R - https://satijalab.org/seurat
- ArchR – R – specific to scATAC data https://www.archrproject.com/

Collection of resources as they come out: emdann/momicsTools

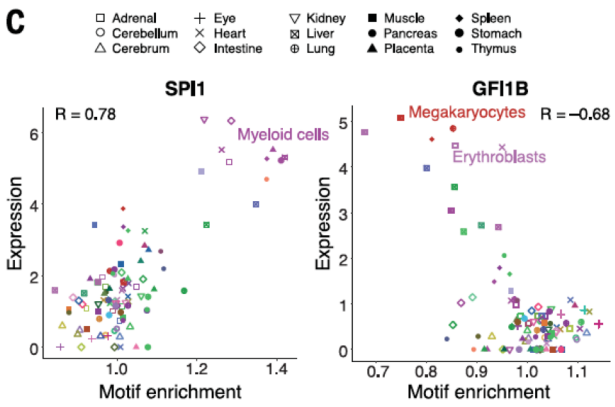
Limitations: assuming molecular changes are simultaneous



DORC = Domain of Open Regulatory Chromatin

Ma et al. (2020), Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin, Cell

Limitations: focus on positive regulation



Domke et al. (2020) A human cell atlas of fetal chromatin accessibility

Repressor factors: expression of a gene closes chromatin

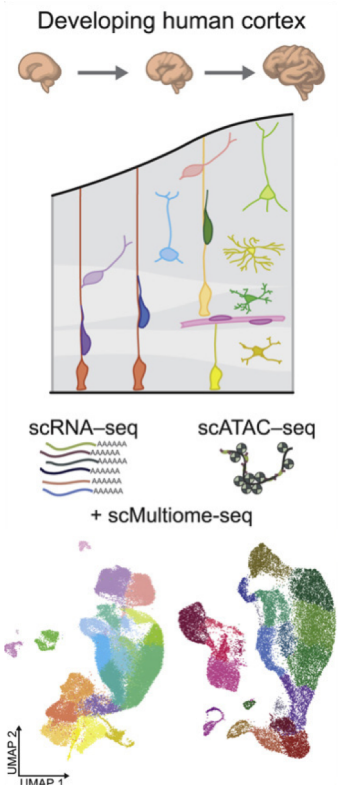
Silencer elements: accessibility of the locus silences a gene (allowing repressor TFs to bind?)

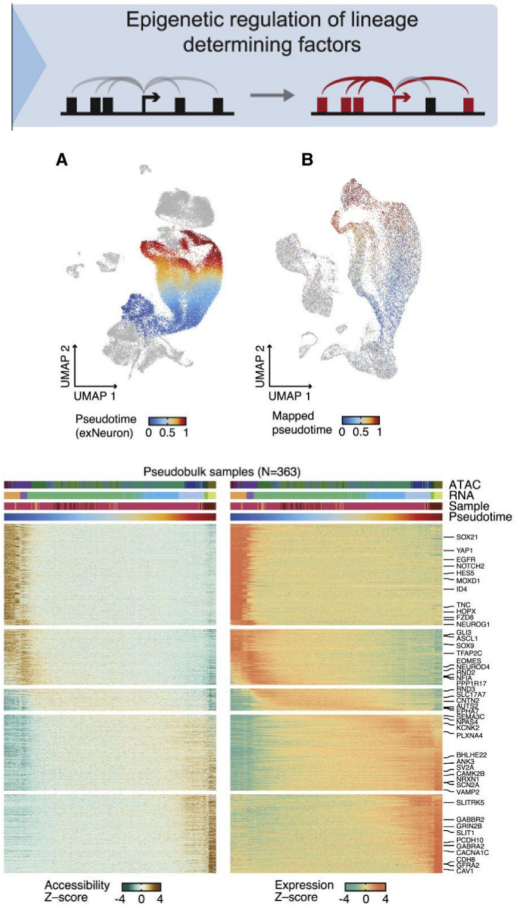
Take home messages

- **There is no state-of-the-art in multi-omics analysis:** new technology keeps coming and shifts the priority of data analysis

- **“Integration” is not the end, it’s the beginning:** cases that break the assumptions for co-embedding are possibly the most interesting

Group project





Group 1: diagonal integration of unmatched scRNA-seq and scATAC-seq dataset

Group 2: vertical integration of matched scMultiomie dataset

Trevino et al. (2021) Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution

Questions?