



3000788 Intro to Comp Molec Biol

Lecture 17: Multi-omics integration

Fall 2025



Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today's agenda

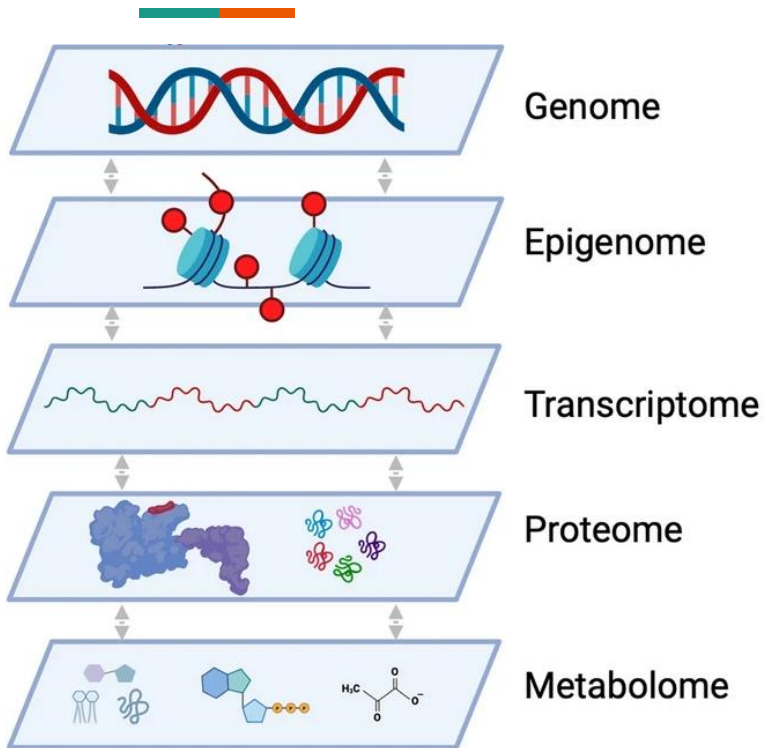


- Recap of omics data and biological information they convey
- Strategies for integrating multiple omics data
- Examples of multi-omics analysis and interpretation



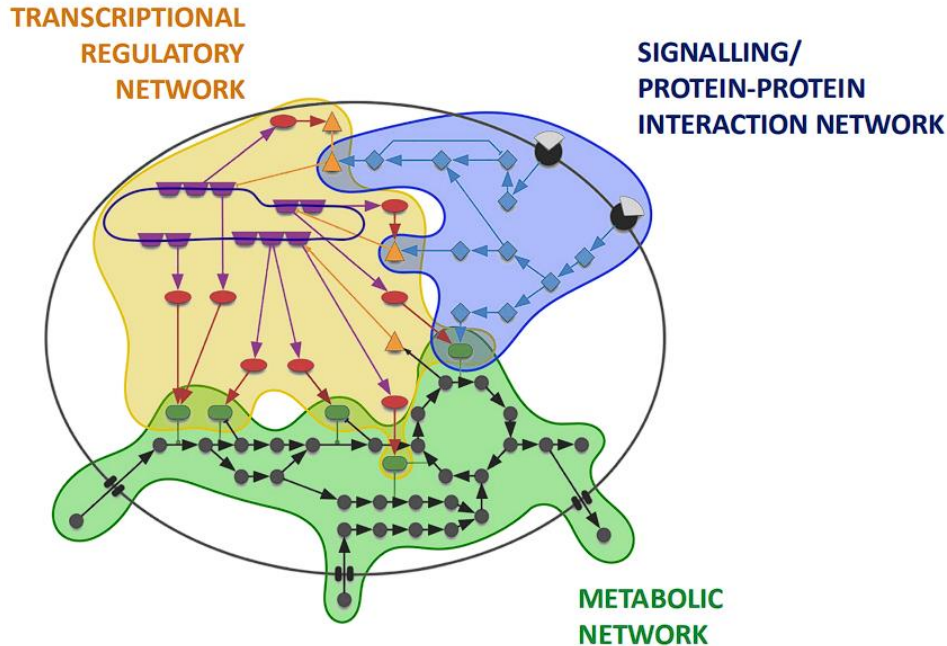
What do omics data tell us?

A simple causal relationship between omics



- Mutation in genome disrupt chromatin state / gene regulation / protein integrity
- Changes in epigenome, transcriptome, and proteome
- Induced changes in metabolome
- Changes can propagate back and forth

Systems view of omics data



- Chromatins, genes, proteins, and metabolites form a single regulatory network
- Change in one layer propagates to others
- Mechanistic explanation requires details from all layers

Genomics



- Mutation
 - Disrupt protein structure & function? (gene body)
 - Change in regulation / splicing? (intron, promoter, enhancer)
 - Change in epigenetics?
- CNV
 - Change in gene expression?
 - Change in regulation? (new gene copy in new promoter / enhancer context)
- SV
 - Combination of effects?
- Genomics tell us where changes occur, **but not the consequences**

Transcriptomics



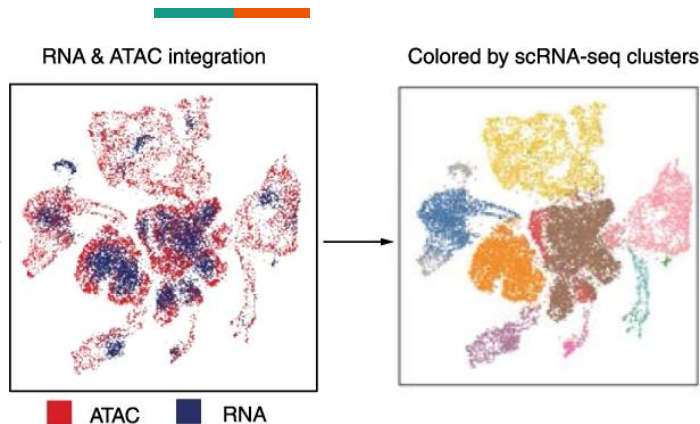
- Change in gene expression
 - Change in regulation?
- Change in pathway activity?
 - Systematic response?
 - Which component was affected first?
- Change in splicing (isoform detection)
- Transcriptomics tell us consequences of regulation, **but not how it occurs**

Epigenomics



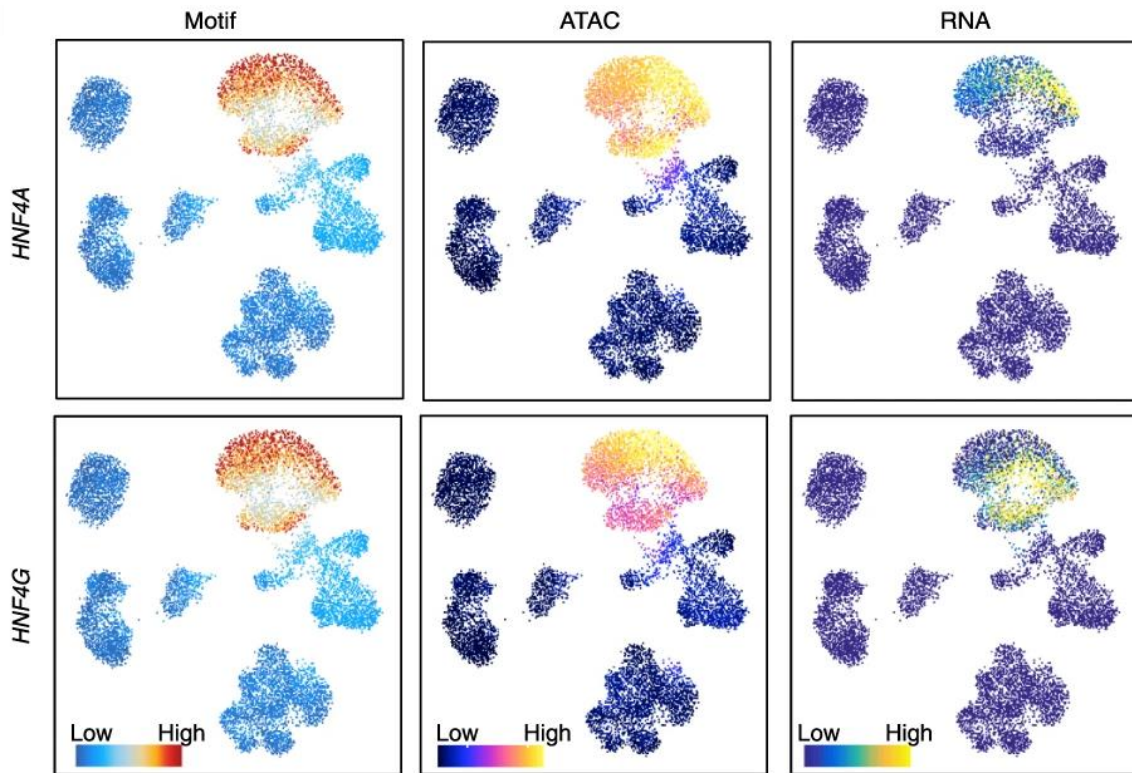
- Change in DNA / histone modification
 - Change in chromatin state?
- Change in transcription factor occupancy
 - Due to increased TF expression?
 - Due to increased chromatin accessibility?
 - Due to another cofactor of TF?
- Change in chromatin accessibility
 - Change in DNA binding protein occupancy?
 - Change in gene expression?
- Epigenomics **link genome and transcriptome by providing mechanisms**

Mechanistic understanding from multi-omics



Qu, J. et al. Nature Communications 13:4069 (2022)

- Correlated TF motif occupancy, accessibility, and RNA expression

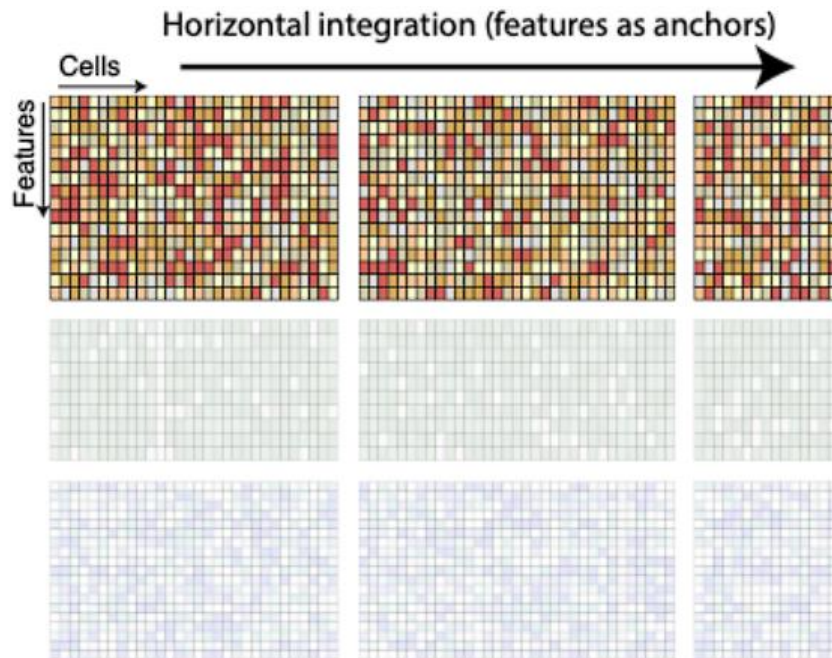
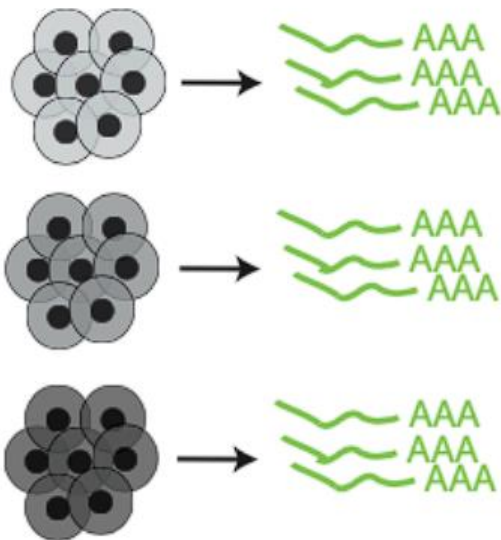




Data integration regimes

Horizontal integration (multiple batches)

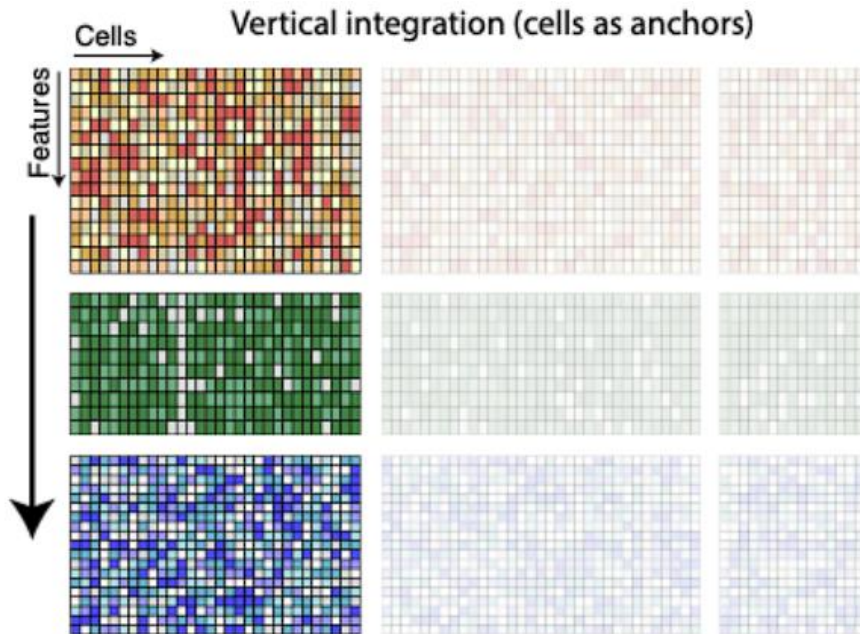
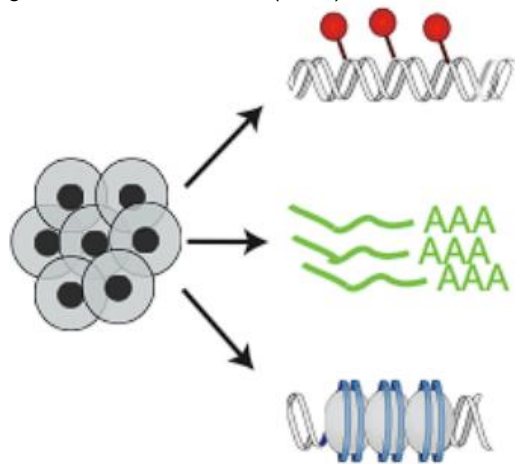
Argelaguet, R. EMBL-EBI training on multi-omics data integration and visualization (2025)



- This is what we are most familiar with

Vertical integration

Argelaguet, R. EMBL-EBI training on multi-omics data integration and visualization (2025)



- Omics data were acquired from the same cells / samples (split aliquot)
- Can directly interpret data

Vertical integration



- **Transcriptomics**
 - Increased gene X expression
- **Genomics**
 - Mutation in promoter / enhancer of gene X
 - Increased copy number of gene X
- **Epigenomics**
 - Increased TF occupancy of enhancer near gene X
 - Increased chromatin accessibility around gene X
 - Change in histone modification around gene X

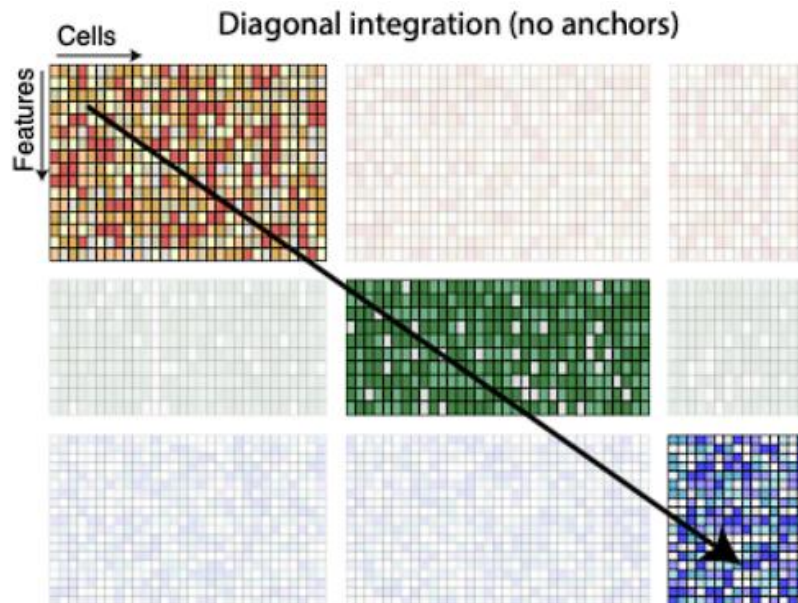
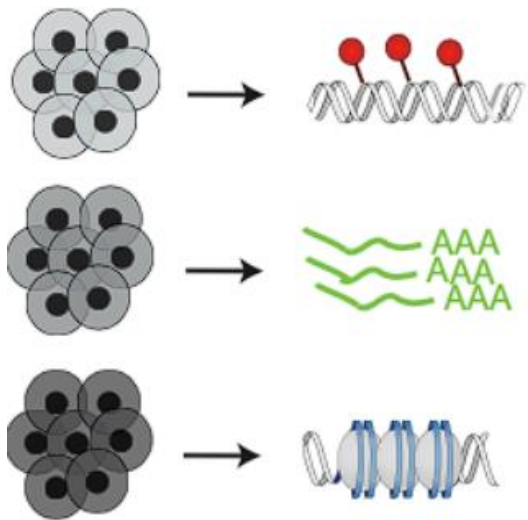
Vertical integration



- **Genomics**
 - Heterozygous, non-synonymous mutation in coding sequence of gene Y
- **Transcriptomics and Proteomics**
 - Observed expression of mutant transcripts and proteins of Y
 - Decreased expression of wild-type form of Y
- **Metabolomics**
 - Change in concentration of metabolite(s) associated with activity of protein Y

Diagonal integration

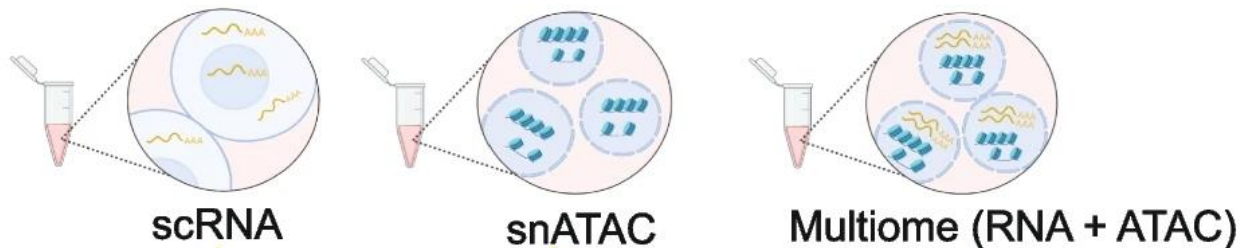
Argelaguet, R. EMBL-EBI training on multi-omics data integration and visualization (2025)



- Different omics data were acquired from different batches of samples
- **Assumption:** similar cell / sample distribution across batches

Vertical vs diagonal integration

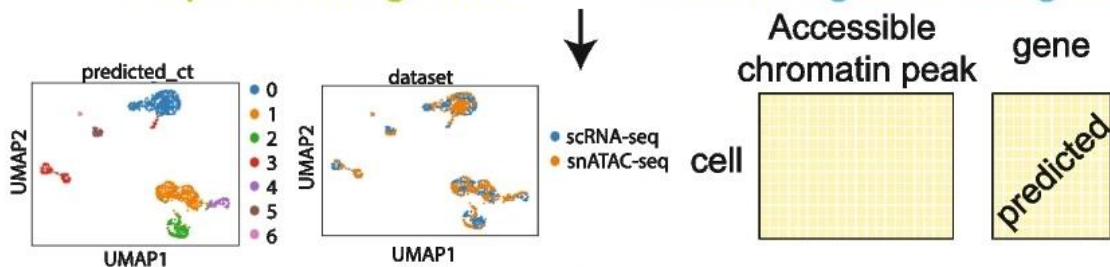
Input:



Lee, M.Y.Y. et al. Genome Biology 24:244 (2023)

Unpaired Integration VS. **Multiome-guided Integration**

Integration output:




- Diagonal integration relies on **cluster-based mapping**

Vertical vs diagonal integration



- Single cell RNA-seq and ATAC-seq from 10k-20k cells of known types
 - Collected as paired (vertical)
 - Simulate as unpaired (diagonal)
- Sample 1k-8k cells as paired or unpaired, analyze, and evaluate the accuracy of cell type assignment
 - Agreement with ground truth cell types
 - [Adjusted Rand Index \(ARI\)](#)
 - Normalized Mutual Information (NMI)

Adjusted Rand Index



Sample	Ground Truth	New Result
1	B cell	A
2	NK cell	A
3	B cell	A
4	B cell	A
5	B cell	B

$$RI = \frac{4}{10} \text{ (1, 3, 4) and (2, 5)}$$

ARI = 1 for perfect agreement

- **Rand Index** = proportion of sample pairs that are **consistently** assigned to either the same groups or different groups in both results
 - (1, 3) are assigned to the same group (B cell and A) in both results
 - (1, 2) are not consistently assigned
- **Adjusted RI** = $\frac{RI - \text{expected RI}}{\max RI - \text{expected RI}}$
considers the distribution of groups (high frequency of B cells and A)

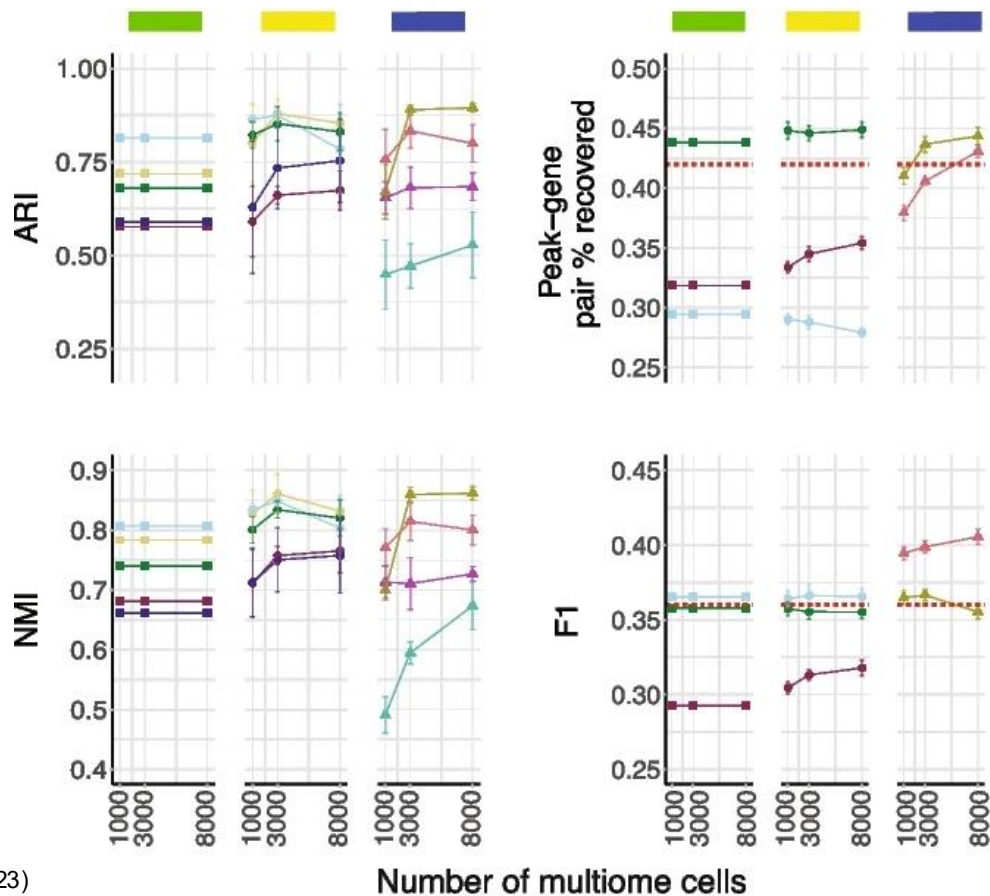
Vertical vs diagonal



- Better cell type identification with paired data (even when treated as unpaired)
- No difference in association peaks to genes (subject to bioinformatics tools used)

Method types

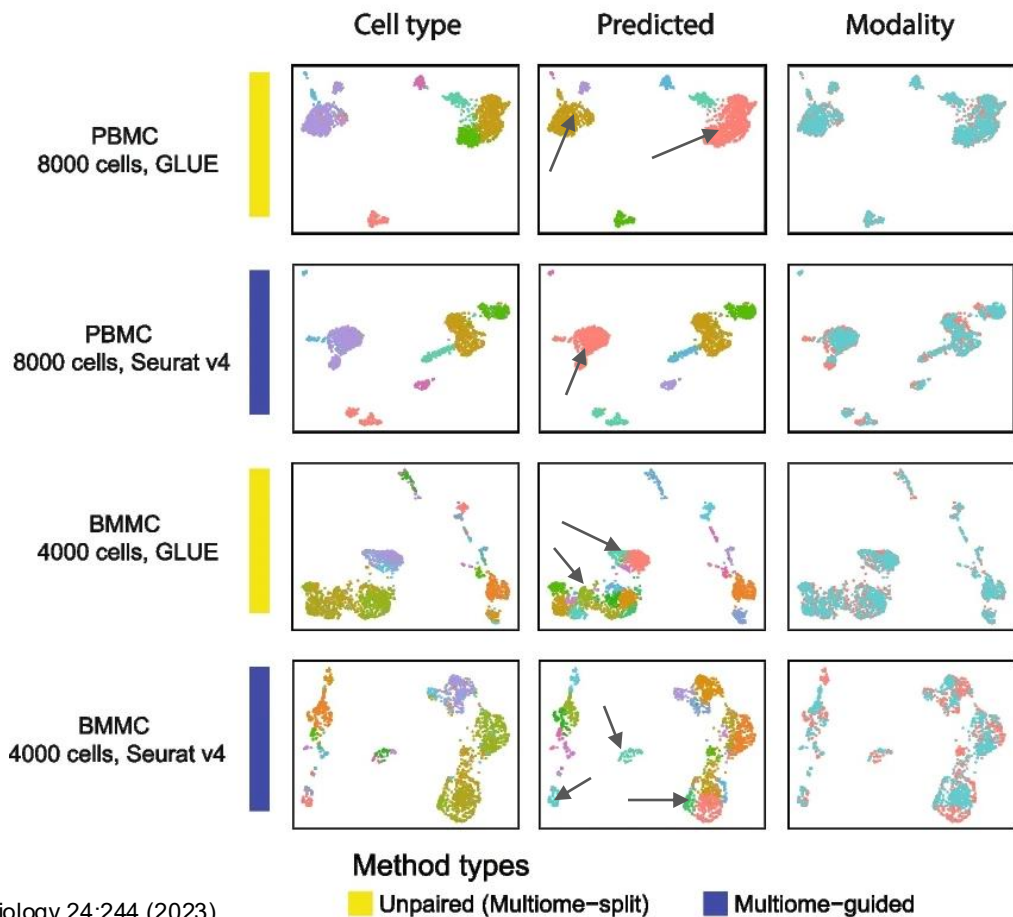
- Unpaired
- Unpaired (Multiome-split)
- Multiome-guided



Vertical vs diagonal



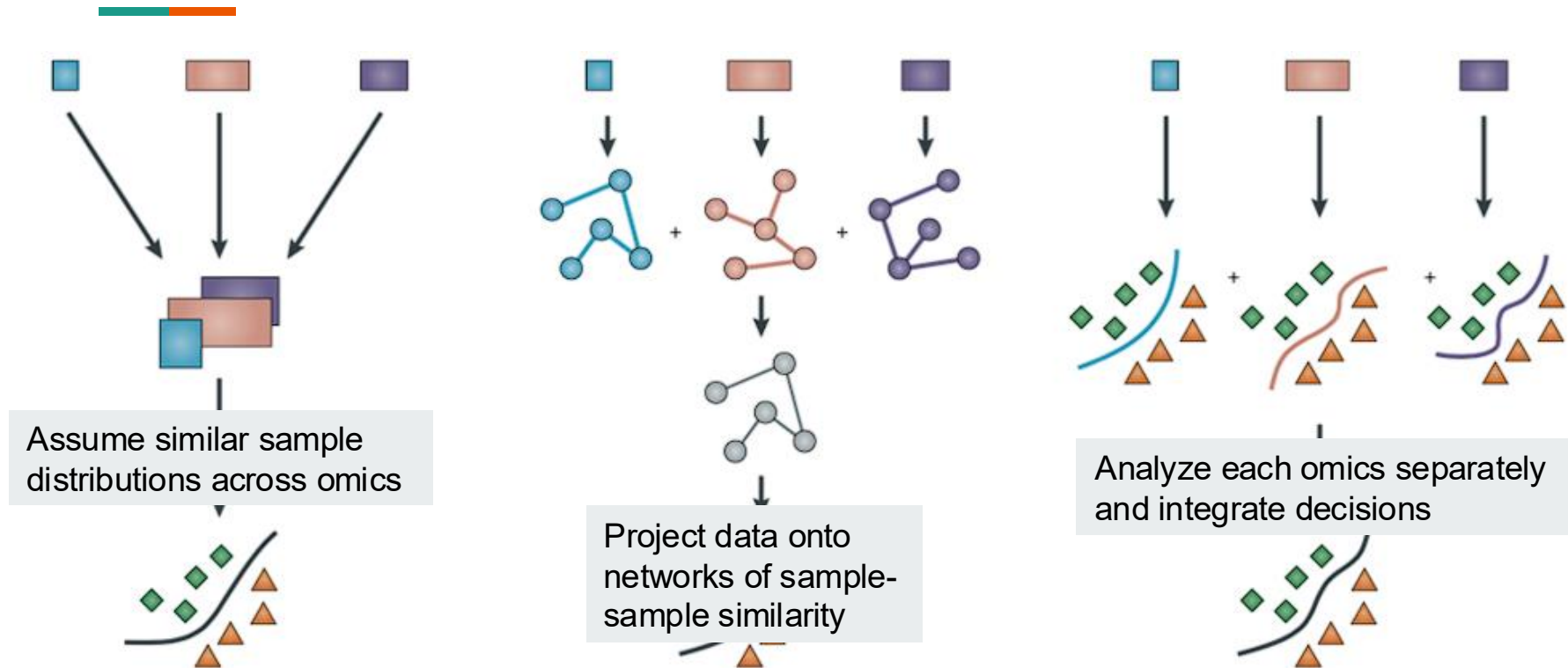
- Some tools can combine multi-omics data to guide cell type identification
- Not a significant difference between the two approach
- Unpaired technique is better at identifying unique cell types in a modality



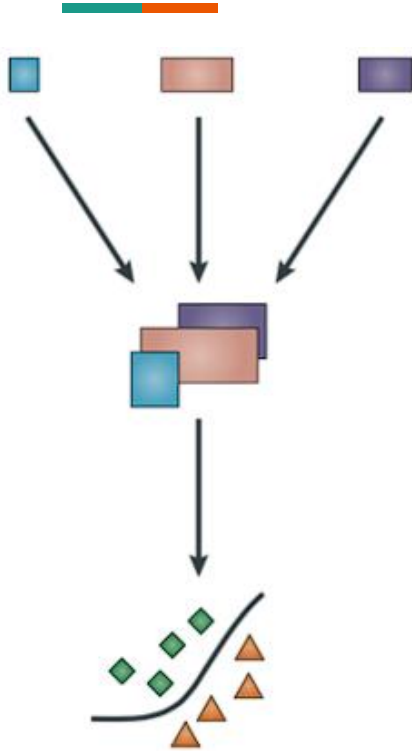


Diagonal data integration

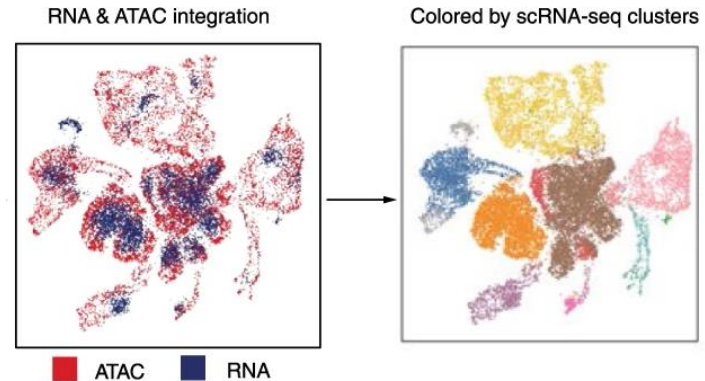
Frameworks for integrating diagonal multi-omics data



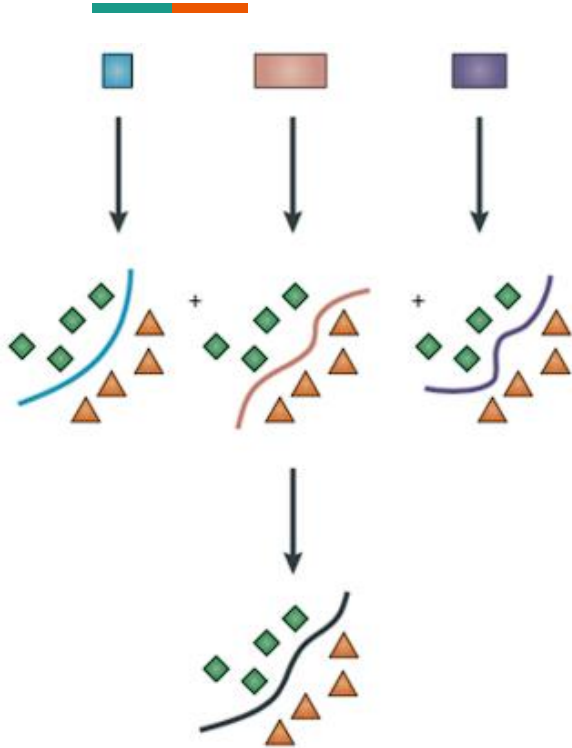
Concatenation



- **Assumption:** Different omics data were acquired from a similar distribution of cells or samples
- Map clusters of cells or samples across modalities

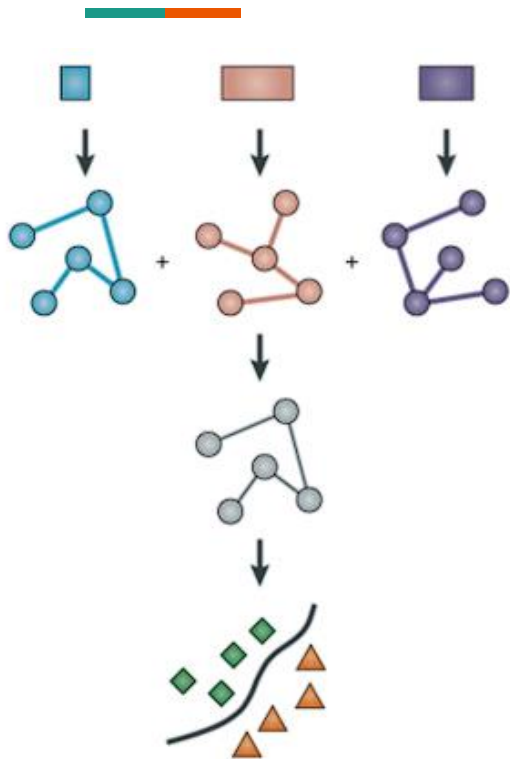


Separate modeling



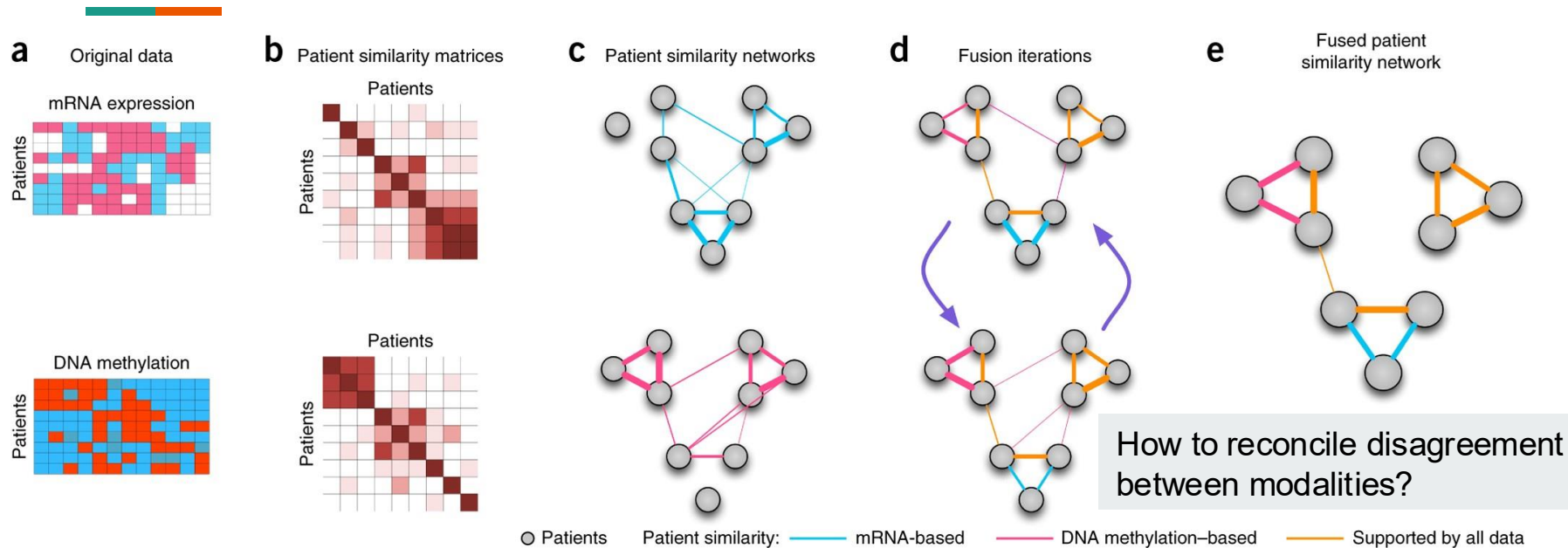
- **Assumption:** Affected pathways or functions should be consistently observed in each omics data
- **Example:**
 - Differential gene / protein expression
 - Differential TF binding, methylation, or histone modification
 - Differential metabolite abundance

Sample-sample similarity



- **Assumption:** Cells or samples in similar states should exhibit similarity across broad omics data (not only in some modality)
- Improve the clustering confidence of cell into types or patients into molecular subtypes
- “Network fusion”

Network fusion algorithm sketch



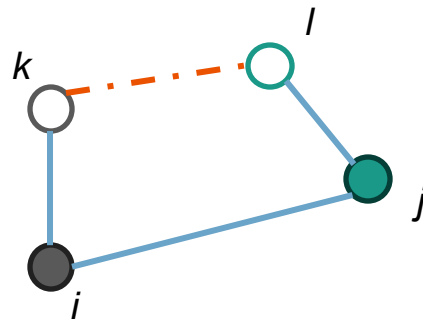
Wang, B. et al. Nature Methods 11:333-337 (2014)

- Group patients using similarities in gene expression and DNA methylation

Network fusion algorithm sketch



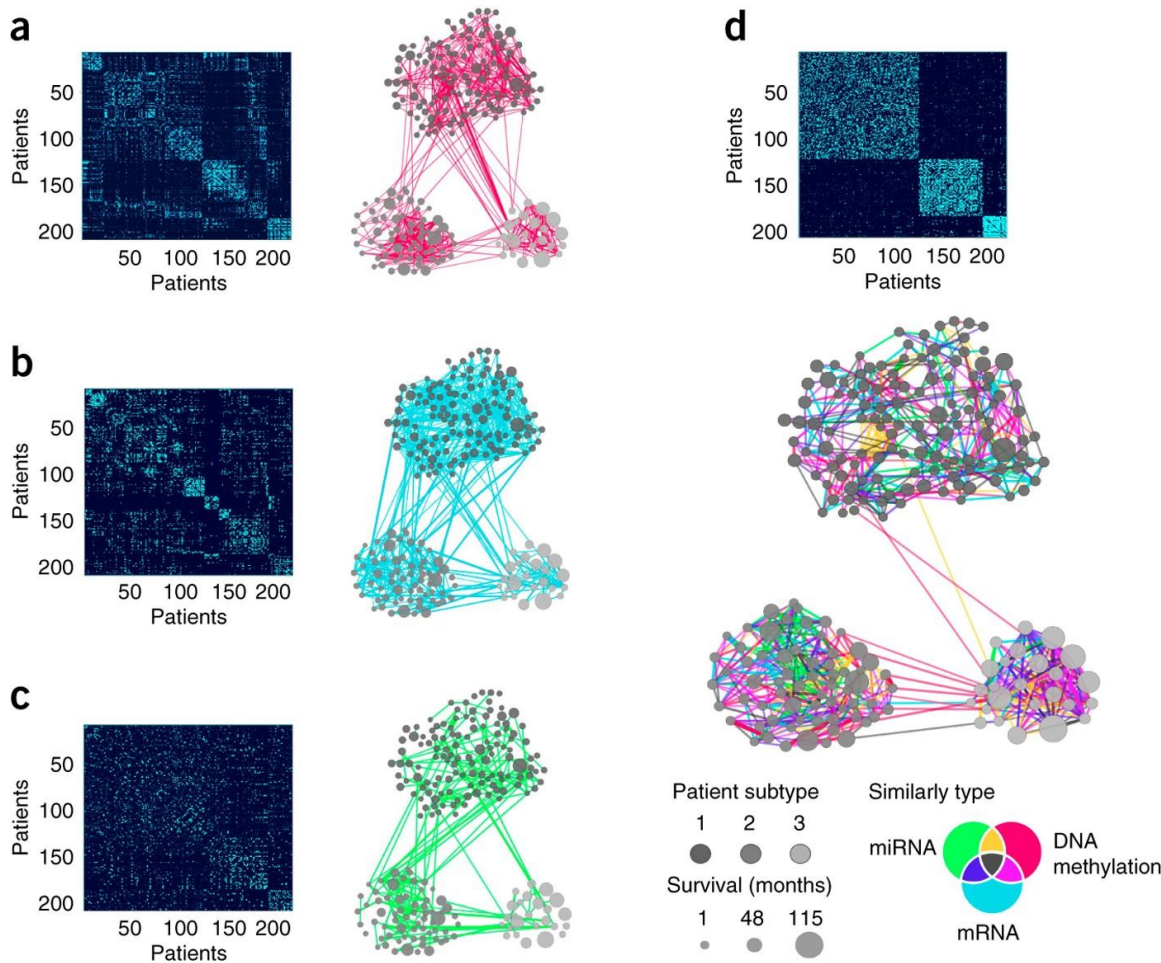
- Start with gene expression similarity
- For an edge (i, j) between patient i and j
- Consider patient k that is similar to i , and patient l that is similar to j (based on gene expression)
- Update edge (i, j) as the average of edge (k, l) (based on similarity in DNA methylation) with edge (i, k) and edge (j, l) as weights (based on similarity in gene expression)



Fused network



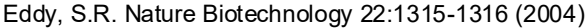
- DNA methylation, gene expression, and miRNA expression data
- Fused network combines evidence from all omics





Latent factor model

Hidden Markov Model



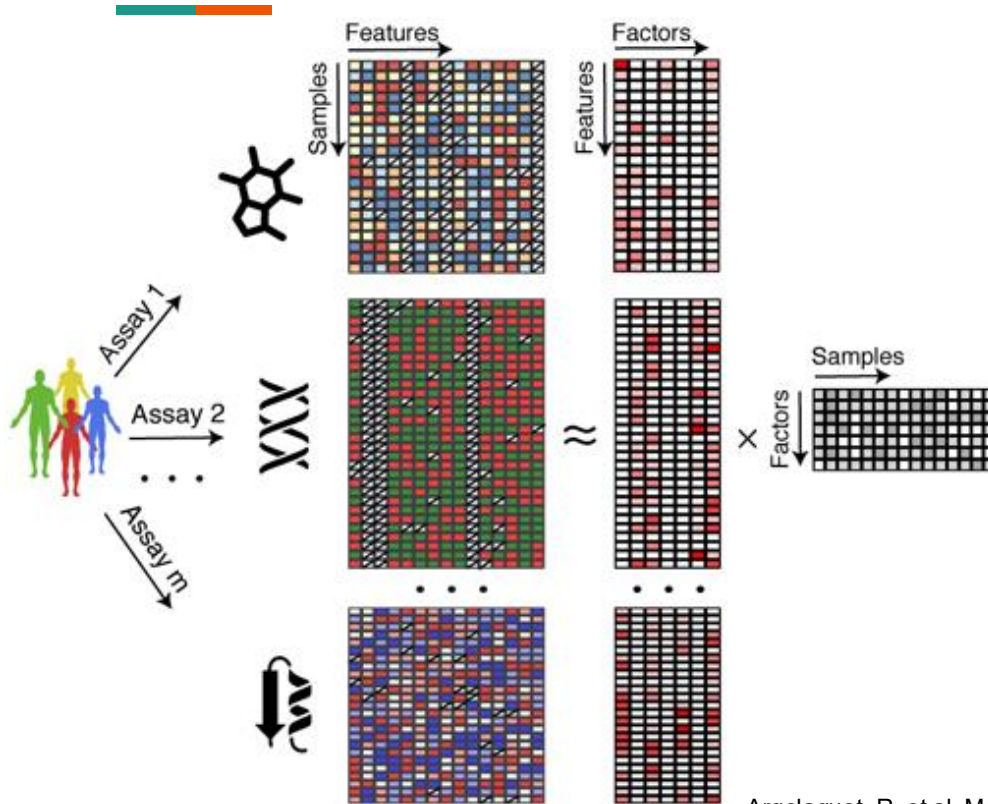
- Observe DNA sequence, which is determined by the exon-intron state
- Exon-intron states are **latent factors**
- Latent factors are (interpretable) characteristics that produce observations

Latent factors



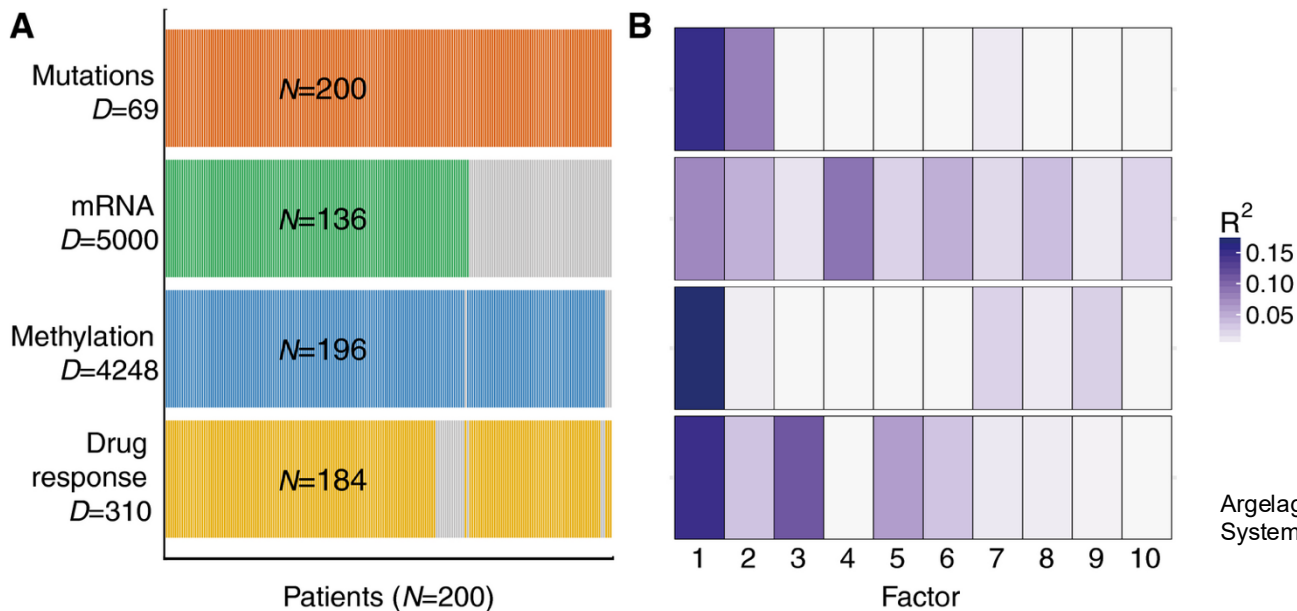
- Factor that mechanistically produce observations
 - **Cell type**: Produce observed gene expressions
 - **Disease state**: Produce observed phenotypes
- Used in statistical models to capture group effects
 - **Batch / Cluster ID**
 - Assign the same bias to cells currently assigned the same batches or clusters
 - Optimize batch / cluster assignment

Multi-omics factor analysis



- **Assumption:** Multi-omics data are determined by the states of the cells or samples
- Decompose omics data (**sample** x **feature**) into modality-specific mechanism (**factor** x **feature**) and shared sample states (**sample** x **factor**)
- Linear effect model

Tuning the number of factors

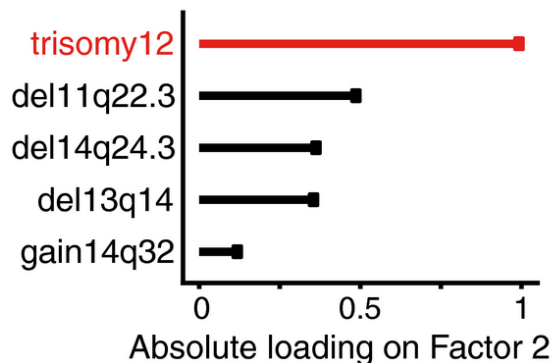
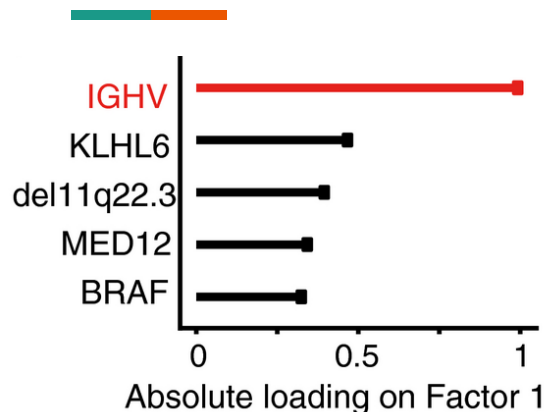


Argelaguet, R. et al. Molecular Systems Biology 14:e8124 (2018)

- Too few factors = cannot explain data
- Too many factors = redundant effects

Nested model, likelihood ratio, information criterion

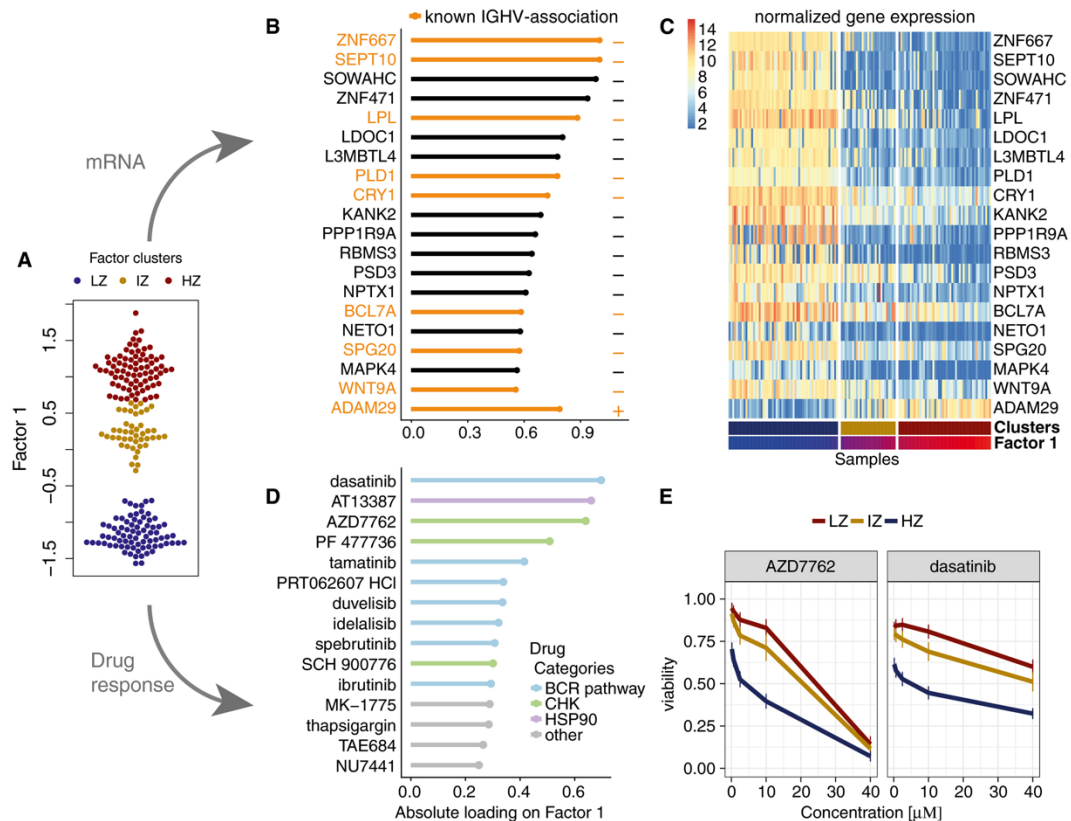
Linking latent factors to observed variables



- Based on coefficients in linear effect model
- **Identify important variables in each omics** for each latent factor
- **Link variables across omics** if they are associated with the same latent factor

Using latent factors to define sample groups

- Cluster samples based on latent factor profiles
- Identify omics signatures for each cluster
- Co-interpret across multi-omics data



Summary



- Different omics data provide distinct biological information
- We should combine multi-omics data to gain mechanistic insight and to weed out noises in individual omics
- Paired multi-omics is preferred (if possible)
- Several strategies for computationally integrating multi-omics data

Any question?



- See you next time