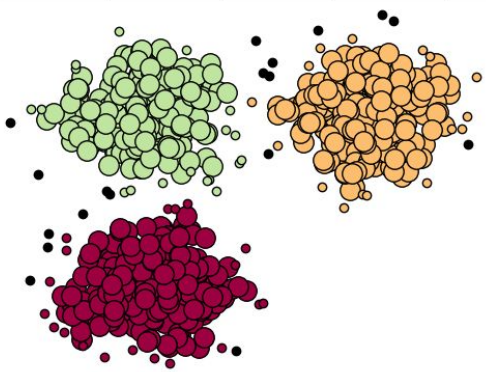
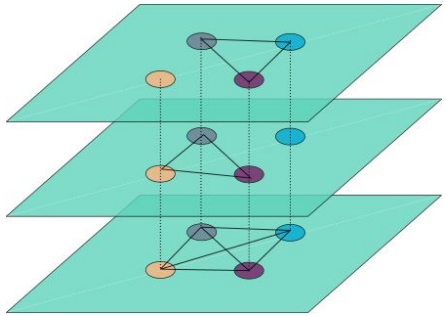




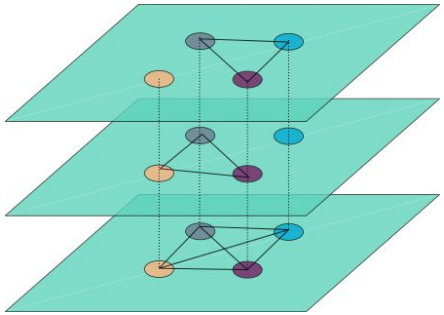
Multi-omics sequencing data integration (*in practice*)

Karolina Sienkiewicz

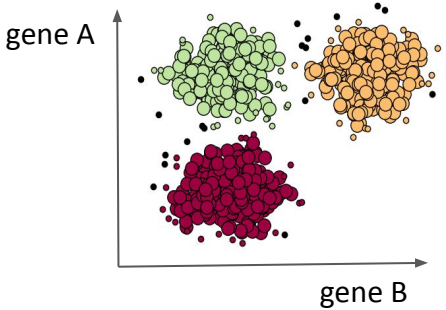
Mason Lab & Melnick Lab
Weill Cornell Medicine

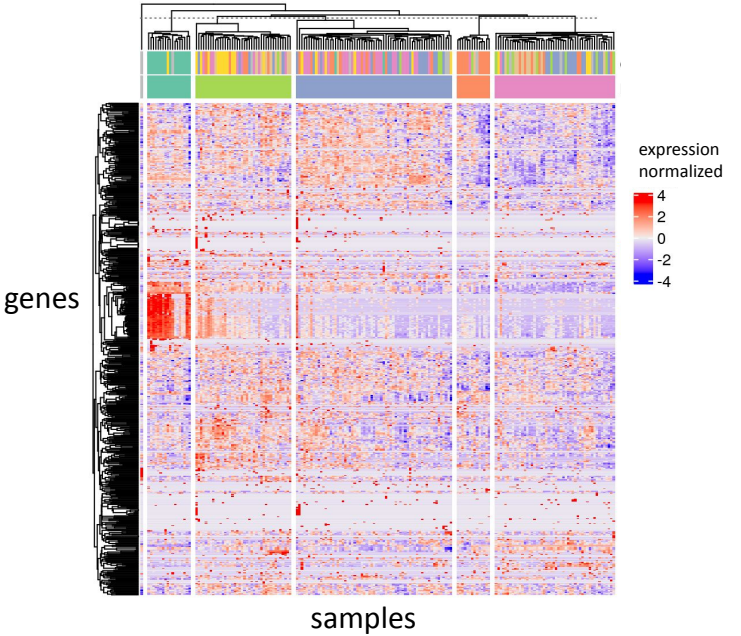
Single molecule sequencing class
14th May, 2024

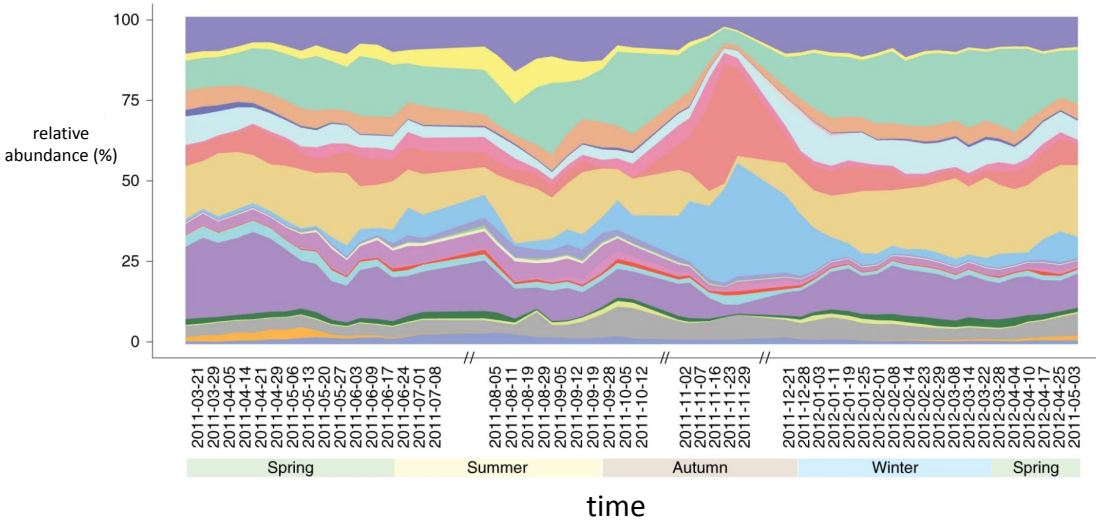
Biological datasets are complex



"omics"

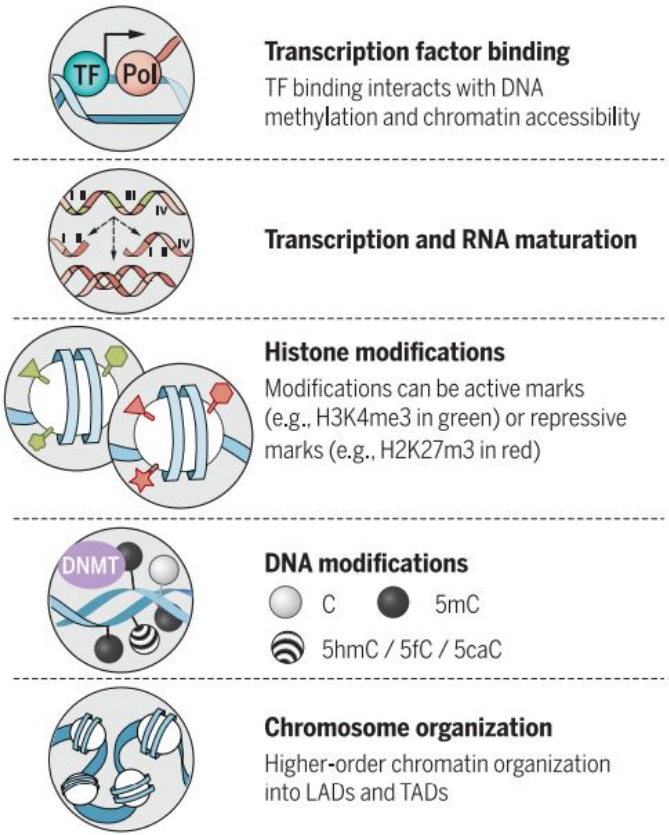


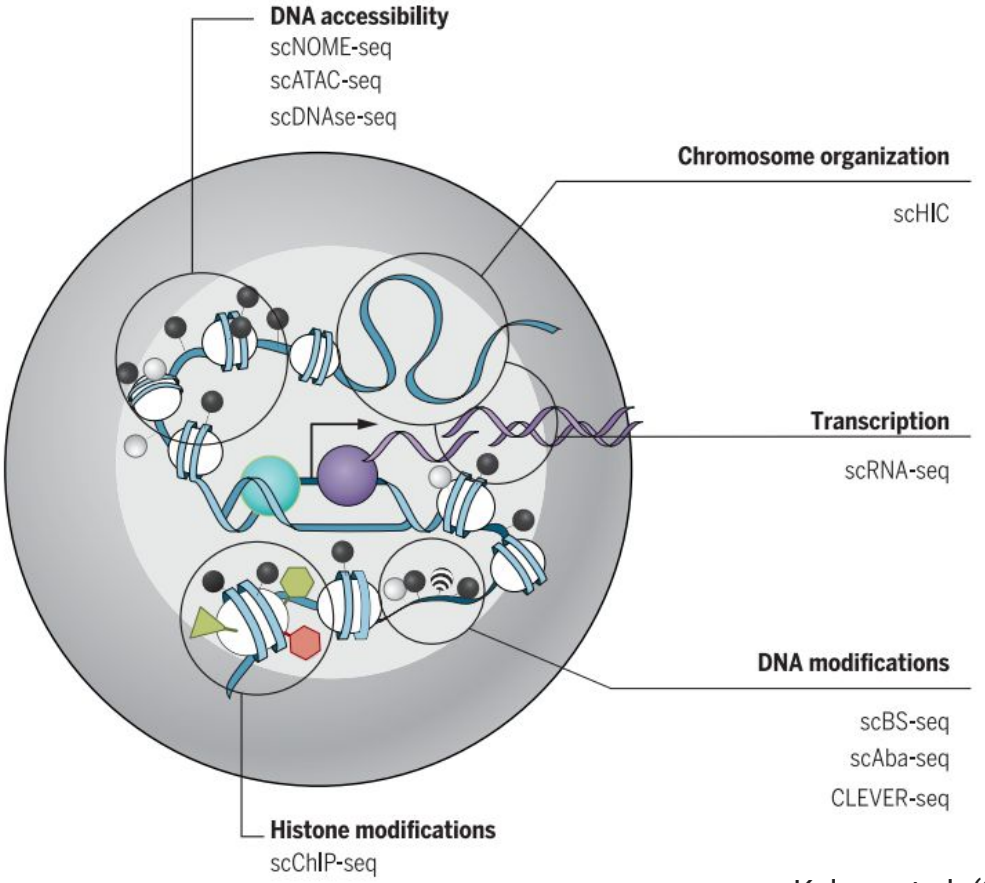




PMID: 33139880

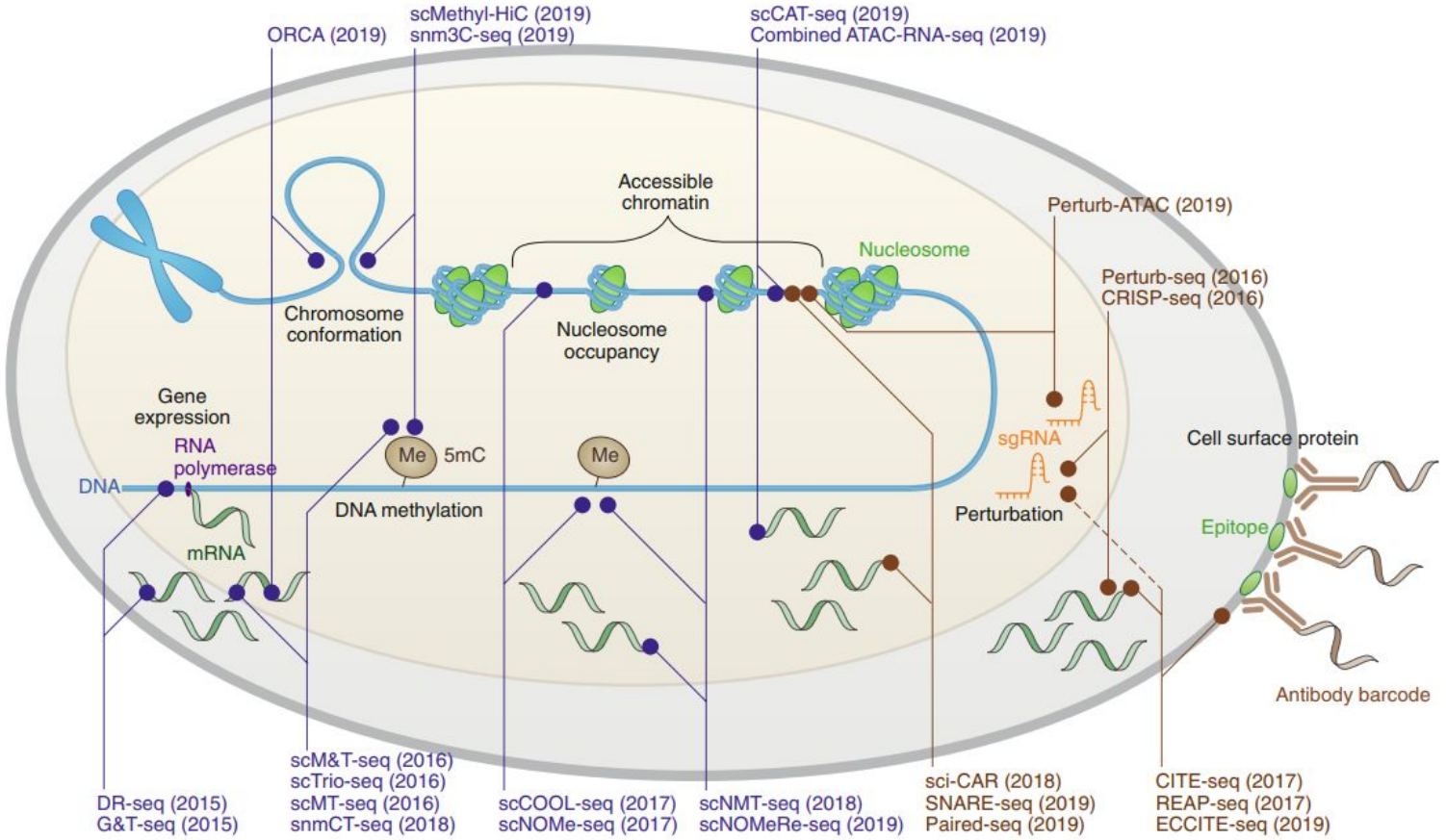
Single-cell epigenetic profiling





Kelsey et al. (2017)

Single-cell multimodal analysis



Zhu et al. (2020)

What are the challenges of multi-omic data integration?

What are the challenges of multi-omic data integration?

Different number
of samples per data-type

Data heterogeneity

Normalization across
data types following
different distributions

Batch effect

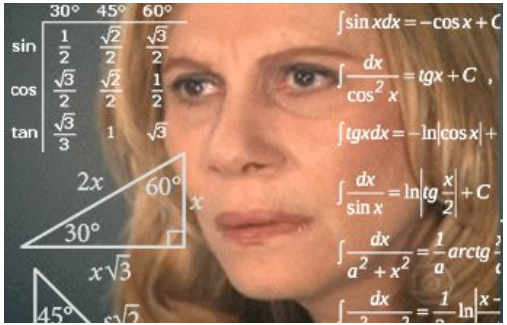
Missing values
in feature matrices

Feature selection

Sparse datasets

Batch effect = non-biological variation

Why is there a batch effect in my data?

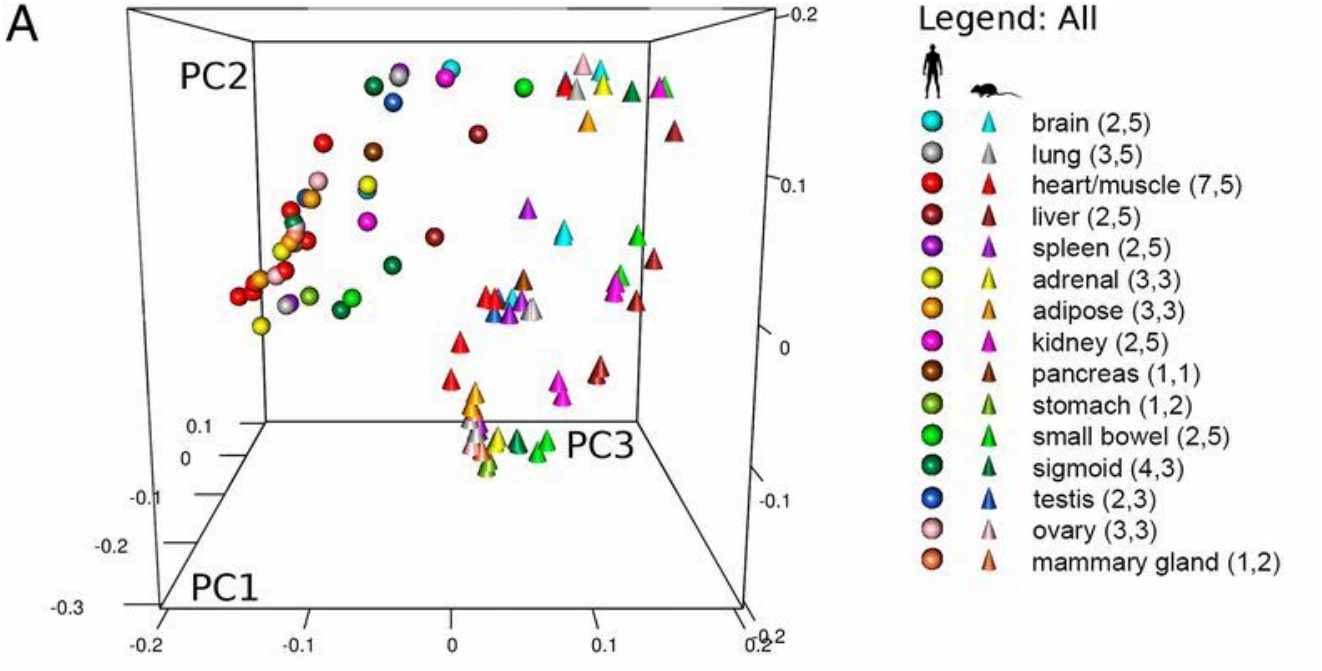


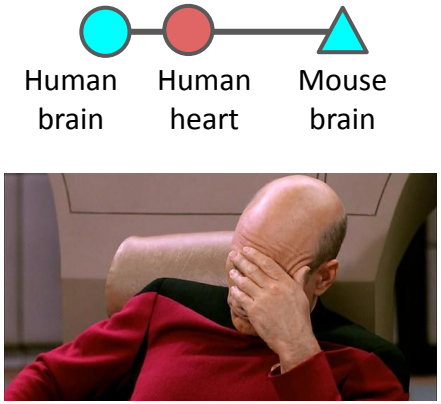
- Different days/months of the experiments
- Different reagents (enzymes, buffers)
- Different mice (from different companies)
- Different sequencers
- Lab protocol or experimenter/technicians

based on STAT115 2020 materials by Shirley Liu

Batch effect example

"Gene Expression Is More Similar Among Tissues Within a Species Than Between Corresponding Tissues of the Two Species"





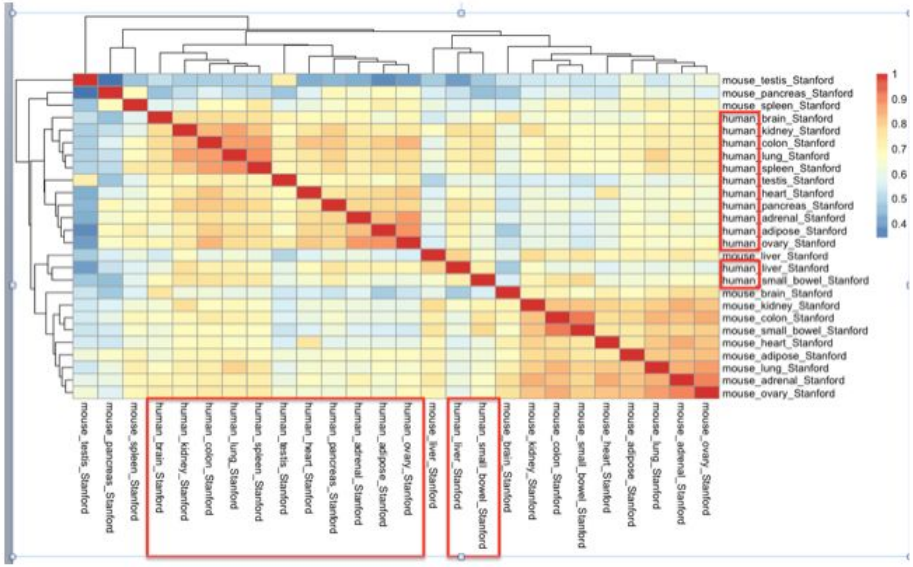
Lin et al (2014) *PNAS*

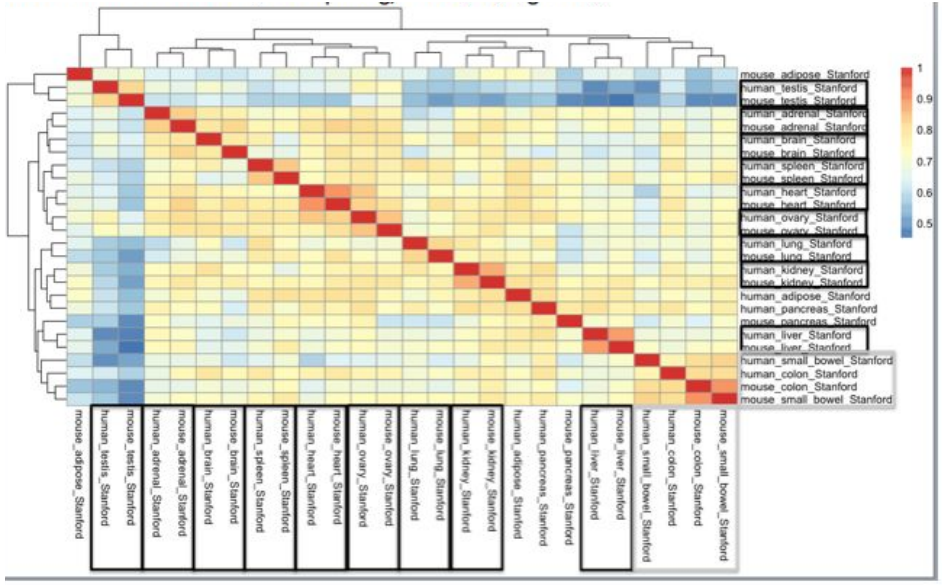
Yoav Gilad
@Y_Gilad

We reanalyzed the data from pnas.org/content/111/48... and found the following:

D87PMJN1 (run 253, lane 7)	D87PMJN1 (run 253, lane 8)	D4LHBFN1 (run 276, lane 4)	MONK (run 312, lane 6)	HWI-ST373 (run 375, lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	human
testis		pancreas		mouse

The original analysis in the paper, considering only the samples that were sequenced at Stanford (data cluster by species):



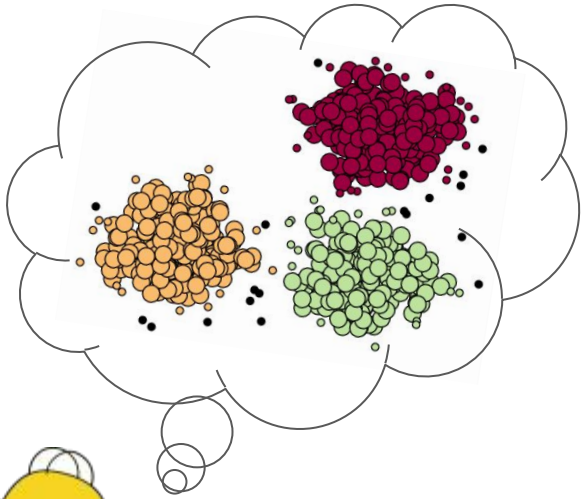


by Yoav Gilad & Orna Man

Good experimental design minimizes batch effects

consistent
pre-processing
of all samples

keeping track
of metadata



balancing
groups of interest

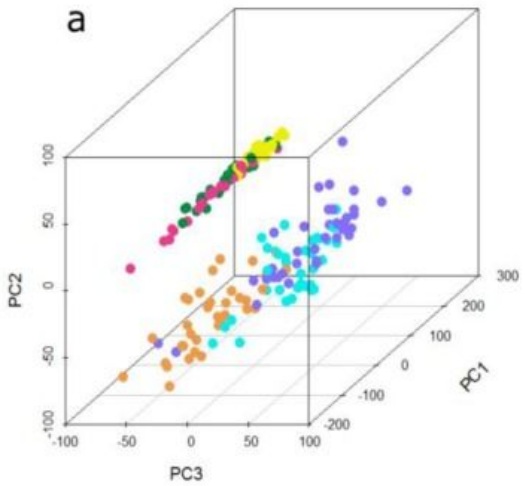
avoiding
“perfect confounding”

Combating batch effect

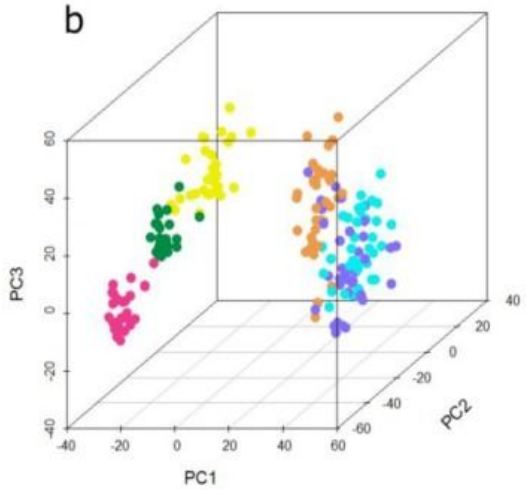
visualize/cluster → identify → correct

ComBat
Johnson et al (2007) *Biostatistics*
limma
complex (e.g. nested) batches

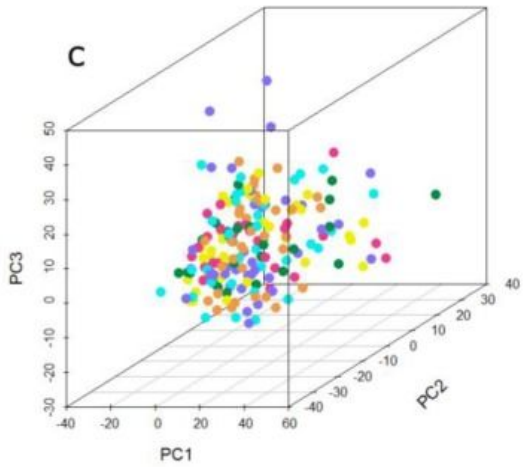
unnormalized data



normalized data



batch-corrected data



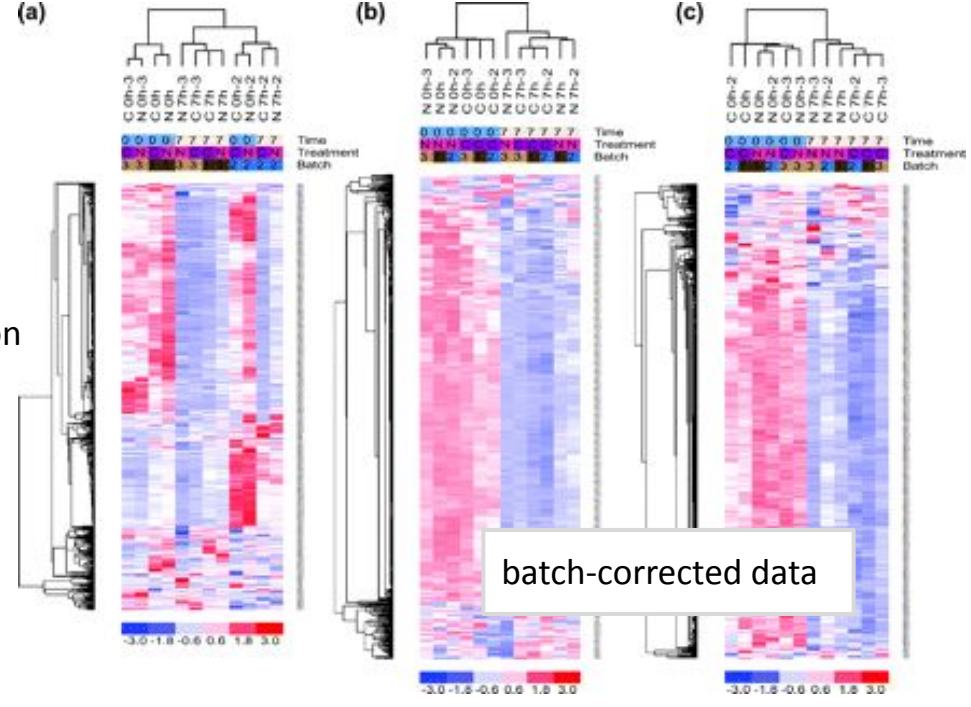
Brezina et al (2015) *Microarray*

Combating batch effect

visualize/cluster → identify → correct

ComBat
Johnson et al (2007) *Biostatistics*
limma
complex (e.g. nested) batches

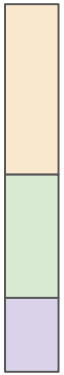
before batch correction

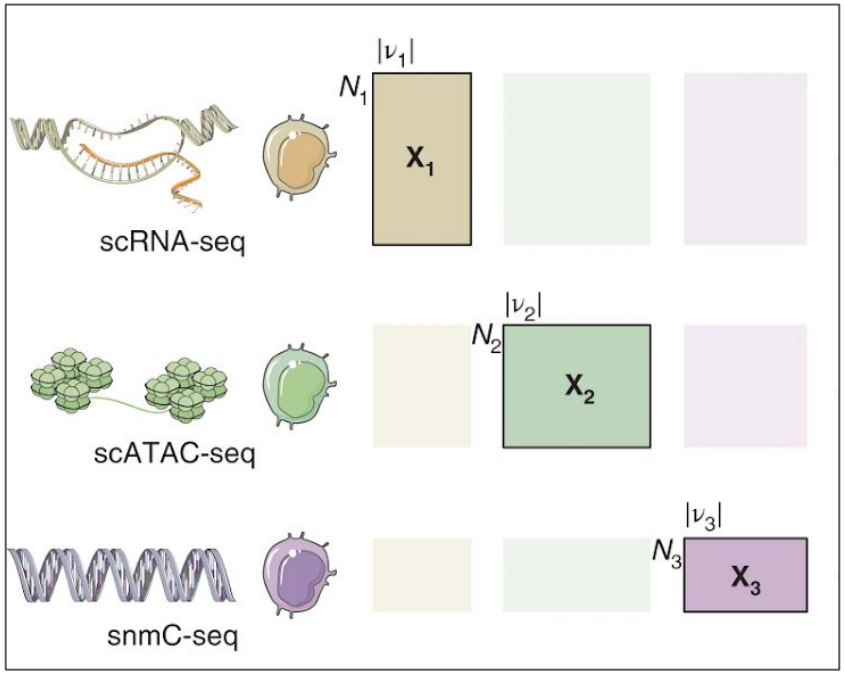


Johnson et al (2007) *Biostatistics*

Major obstacles of integrative modeling: distinct feature/sample space

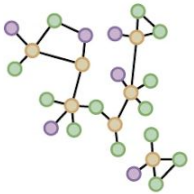
conversion to common feature space





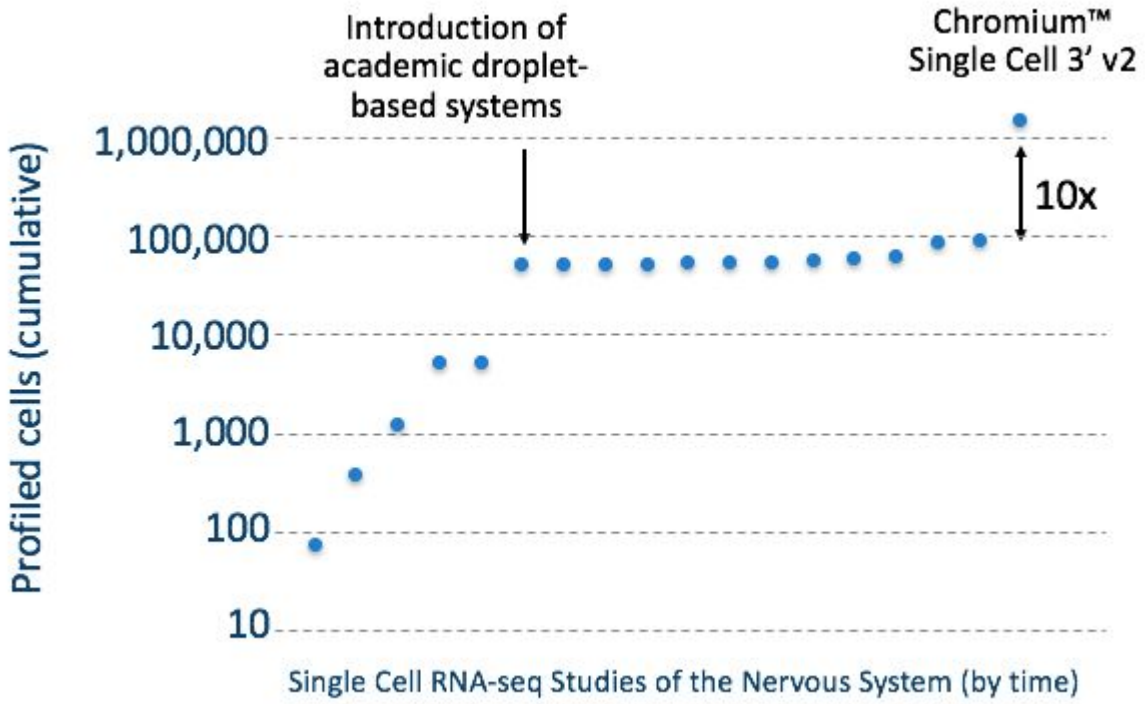
joint matrix factorization

cell matching



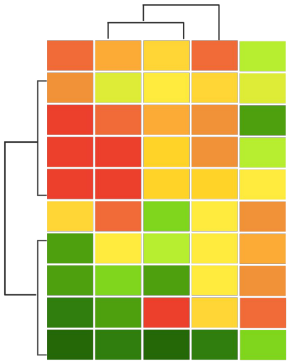
Zhi-Jie Cao & Ge Gao Nature Biotechnology (2022)

Major obstacles of integrative modeling: data scale



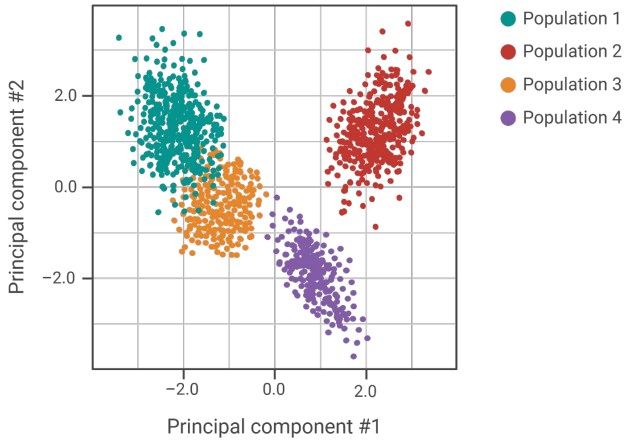
Approximate cumulative number of cells profiled in scRNA-seq literature (up to Oct 2016). *10x genomics*

Primer on dimensionality reduction



data representation (full)

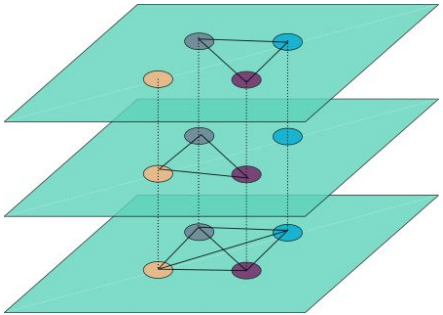
objective/cost function



data representation (reduced)

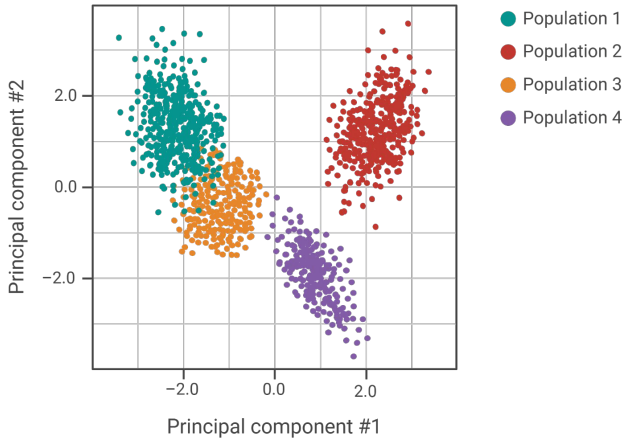
Recommended reading: "Multi-omics Analysis" chapter by Jonathan Ronen from the "Computational Genomics with R"

Primer on dimensionality reduction



data representation (full)

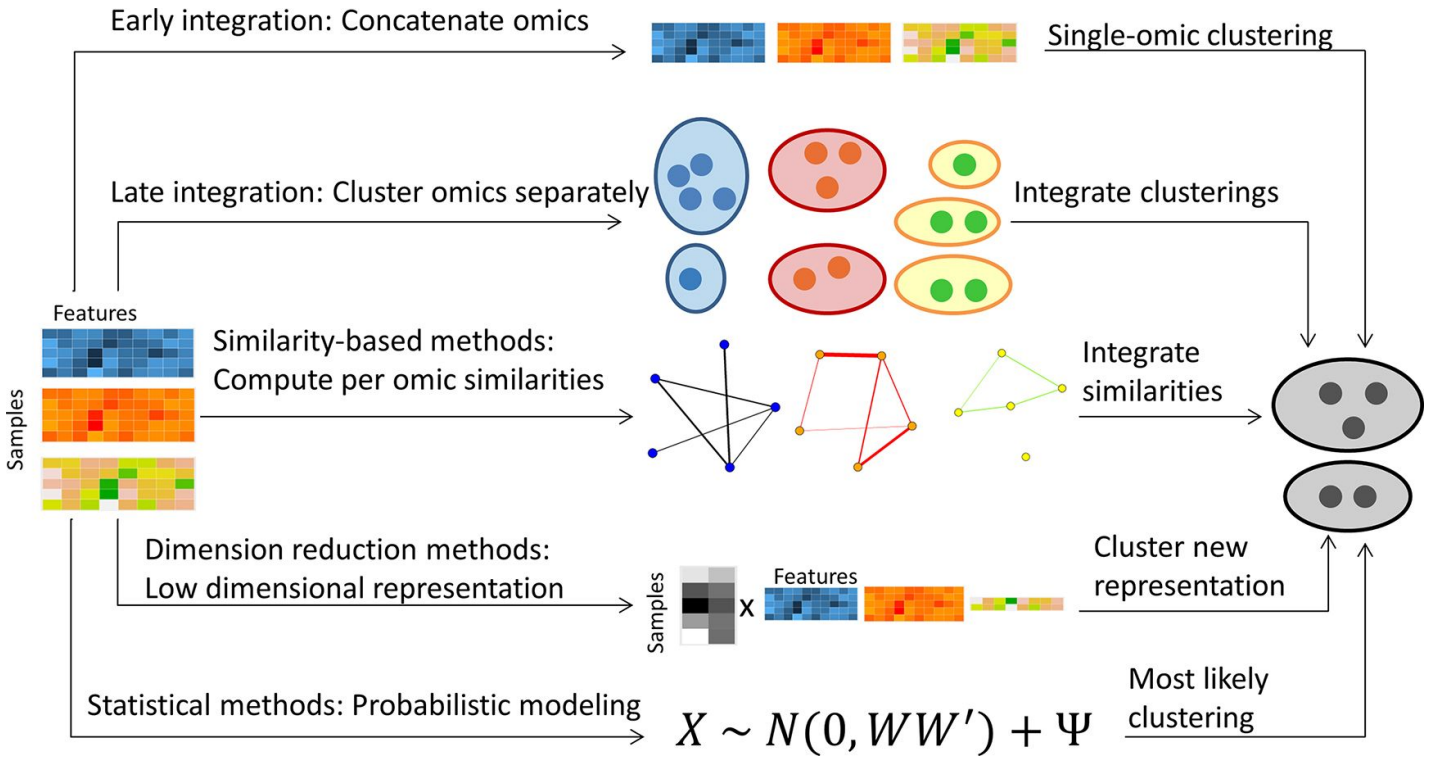
objective/cost function



data representation (reduced)

Recommended reading: "Multi-omics Analysis" chapter by Jonathan Ronen from the "Computational Genomics with R"

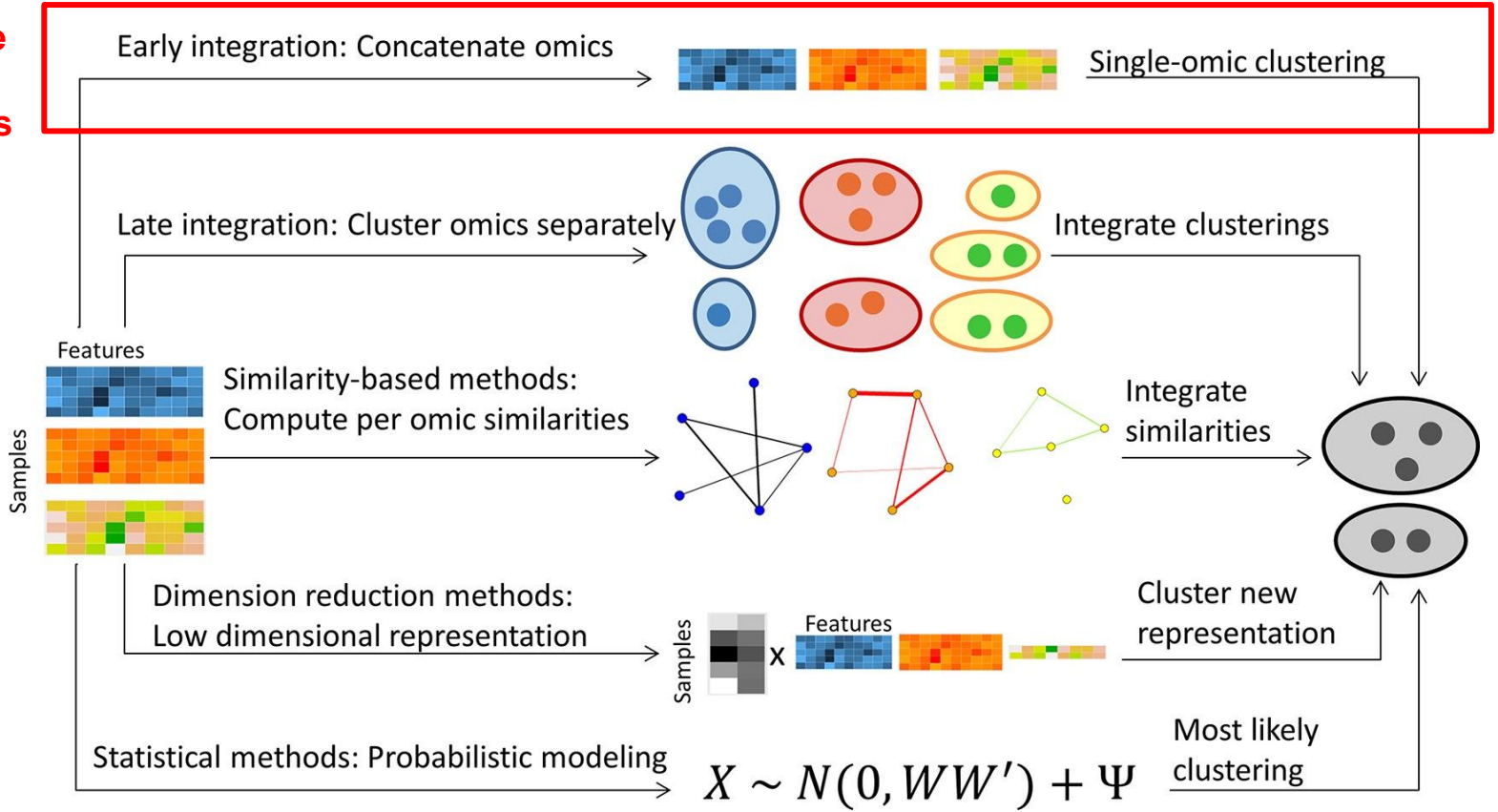
Approaches to multi-omic data integration



Rappoport & Shamir (2018) *Nucleic Acids Res.*

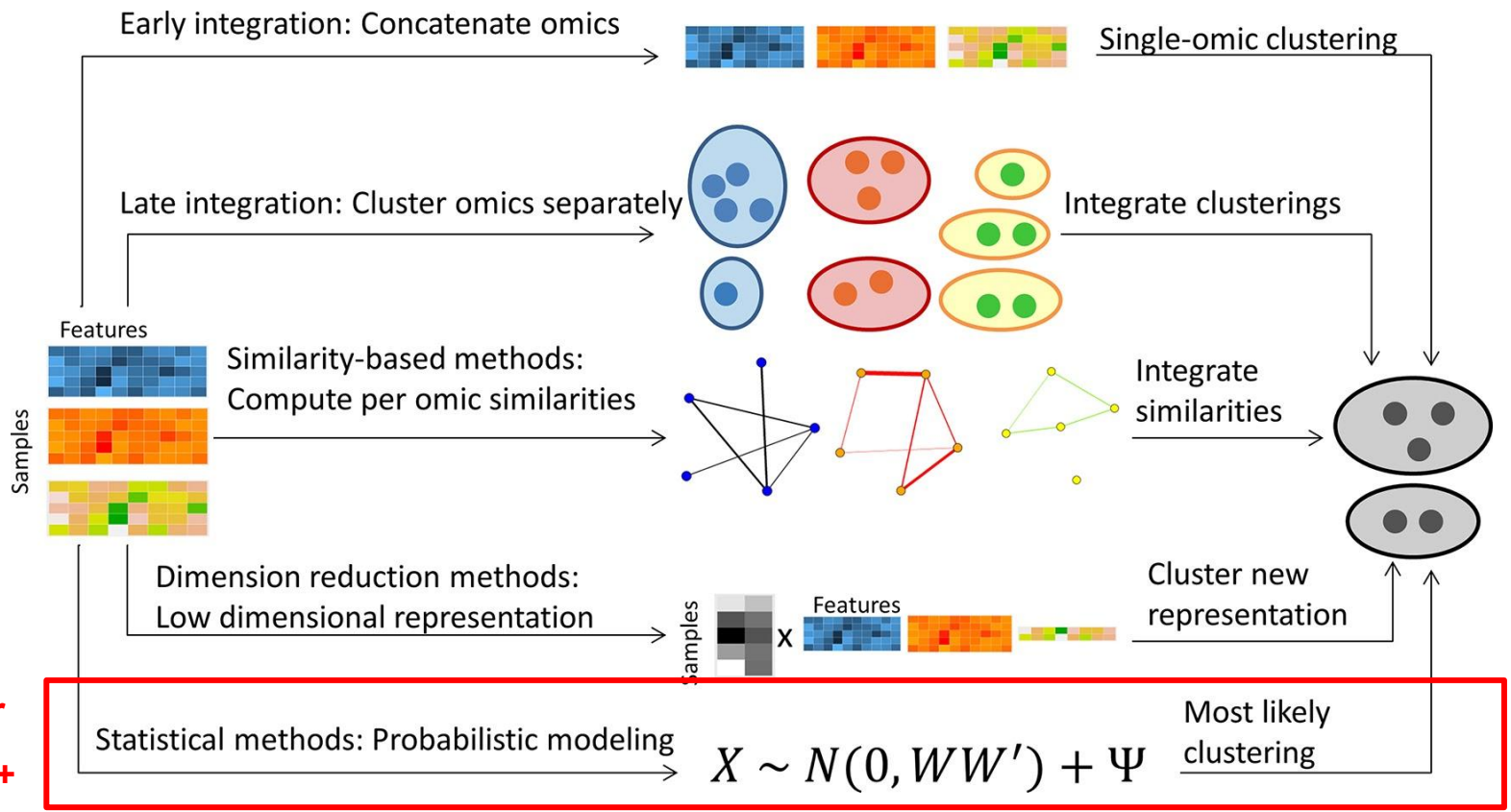
Methods for multi-omic data integration

Multiple Factor Analysis (MFA)



Rappoport & Shamir (2018) *Nucleic Acids Res.*

Methods for multi-omic data integration



iCluster

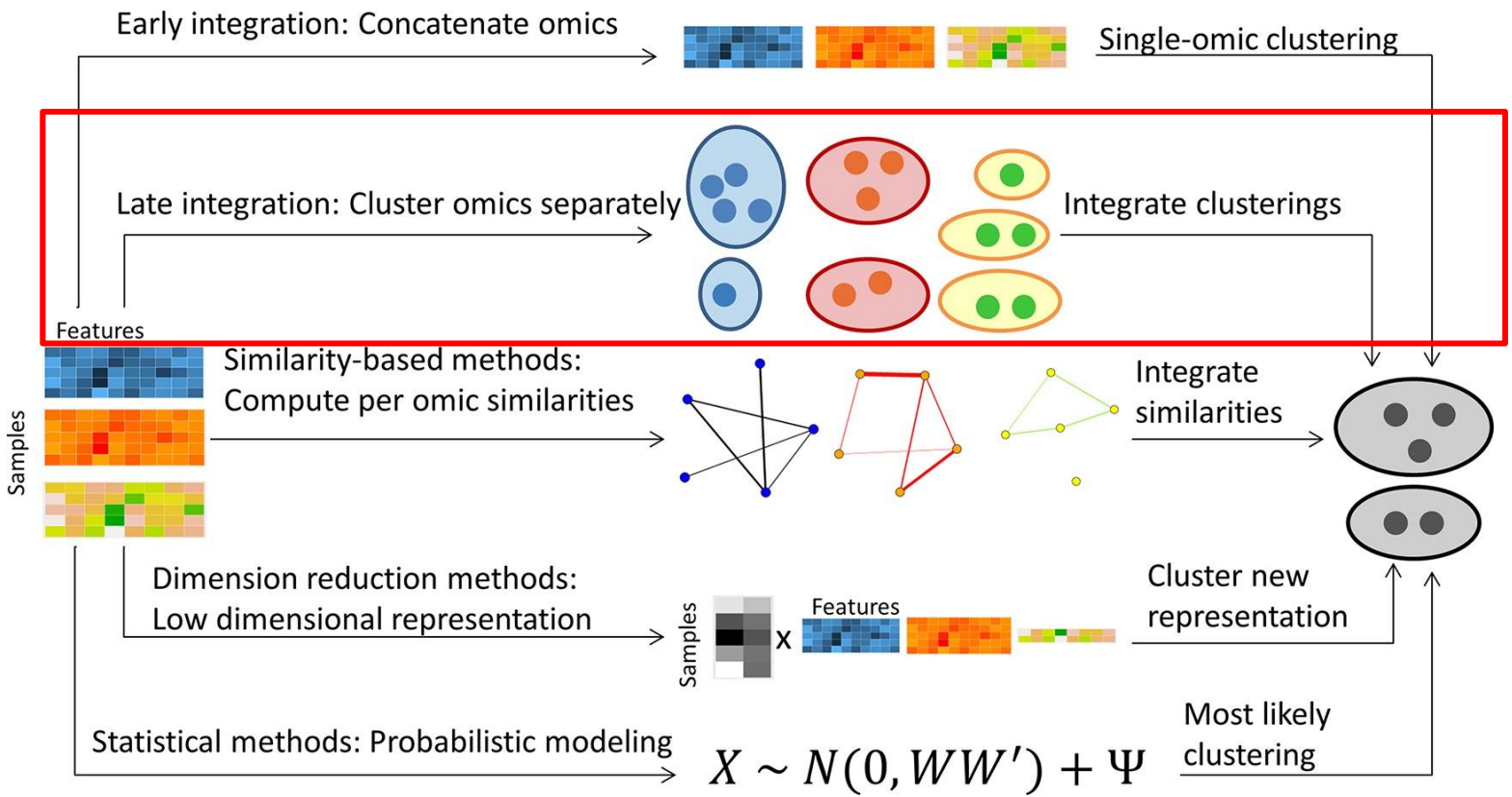
iCluster+

Rappoport & Shamir (2018) *Nucleic Acids Res.*

Methods for multi-omic data integration

PINS

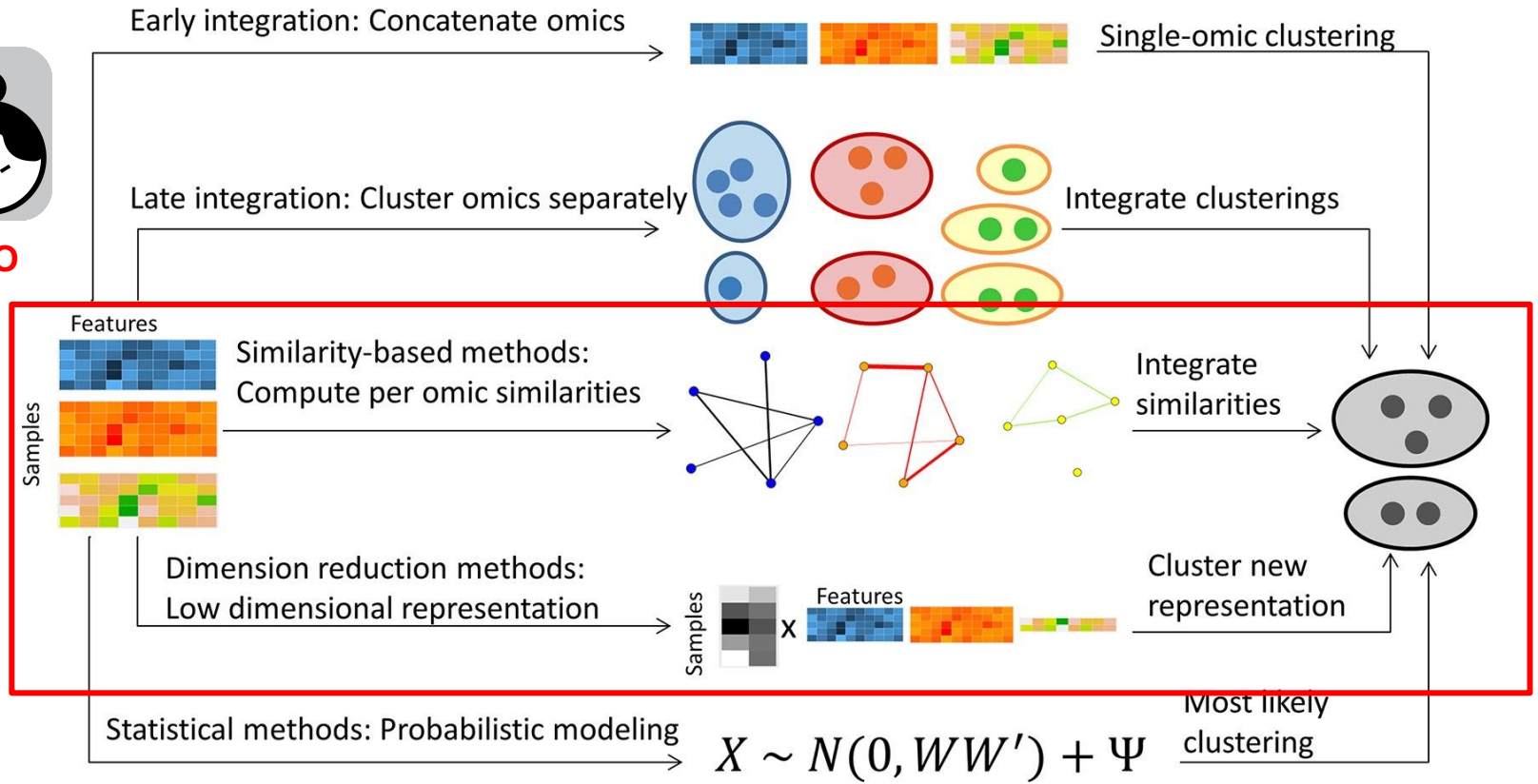
PINS+



Rappoport & Shamir (2018) *Nucleic Acids Res.*

Methods for multi-omic data integration

SUMO

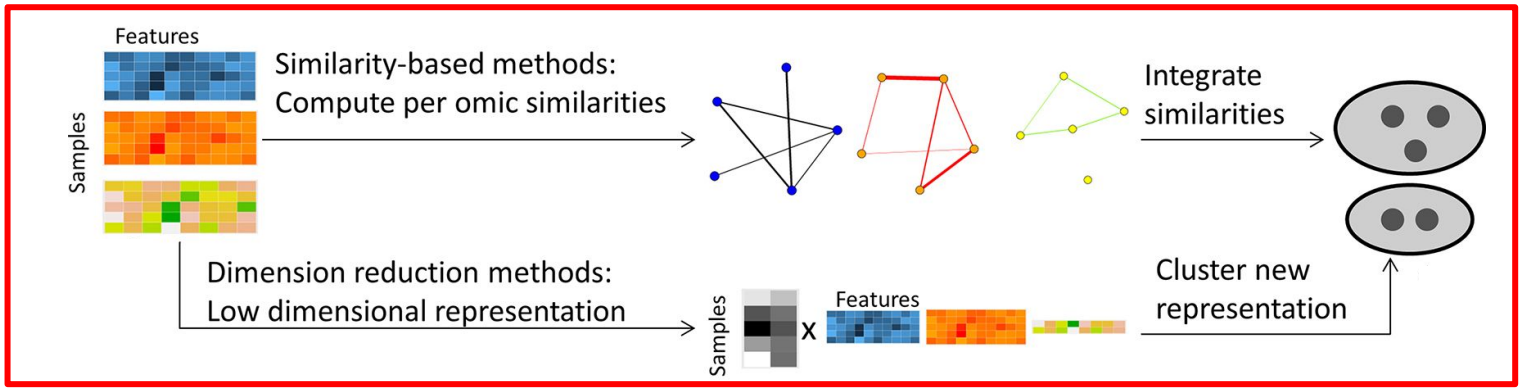


Rappoport & Shamir (2018) *Nucleic Acids Res.*

SUMO: **S**ubtyping **U**sing **M**ulti-**O**mic data

github.com/ratan-lab/sumo

pypi.org/project/python-sumo

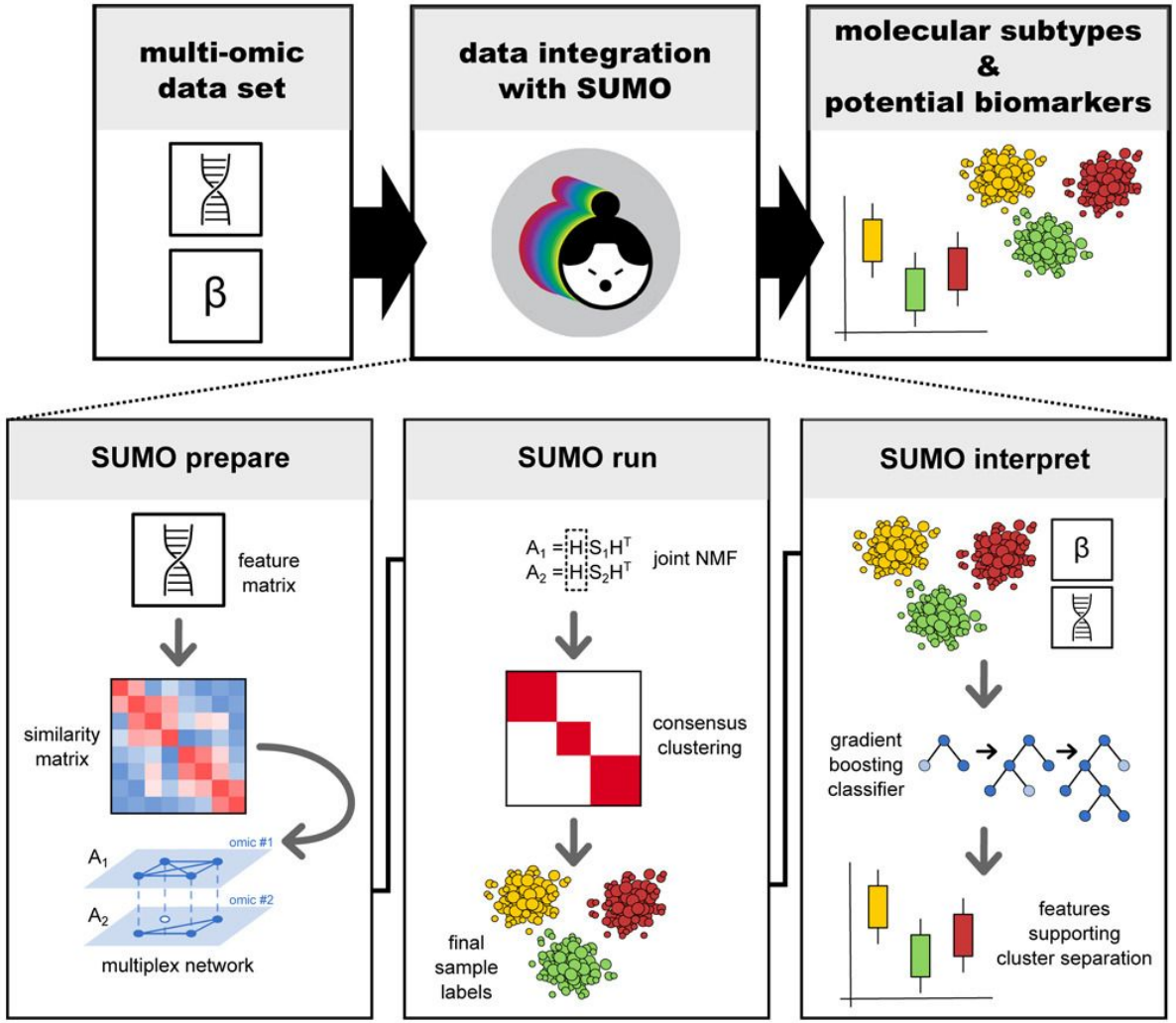


SUMO workflow

extendable & modular framework

does not require imputation

omic-specific similarity metrics:
- euclidean distance based
- cosine similarity
- correlation

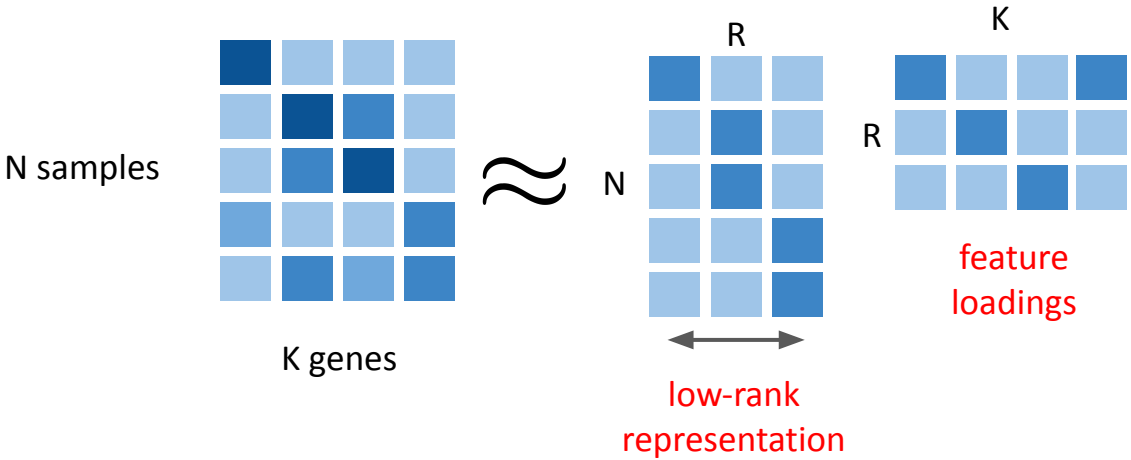


Sienkiewicz et al (2022)
Cell Reports Methods

Factorization basics

NMF	$X \approx WH$
Symmetric NMF	$A \approx HH^T$

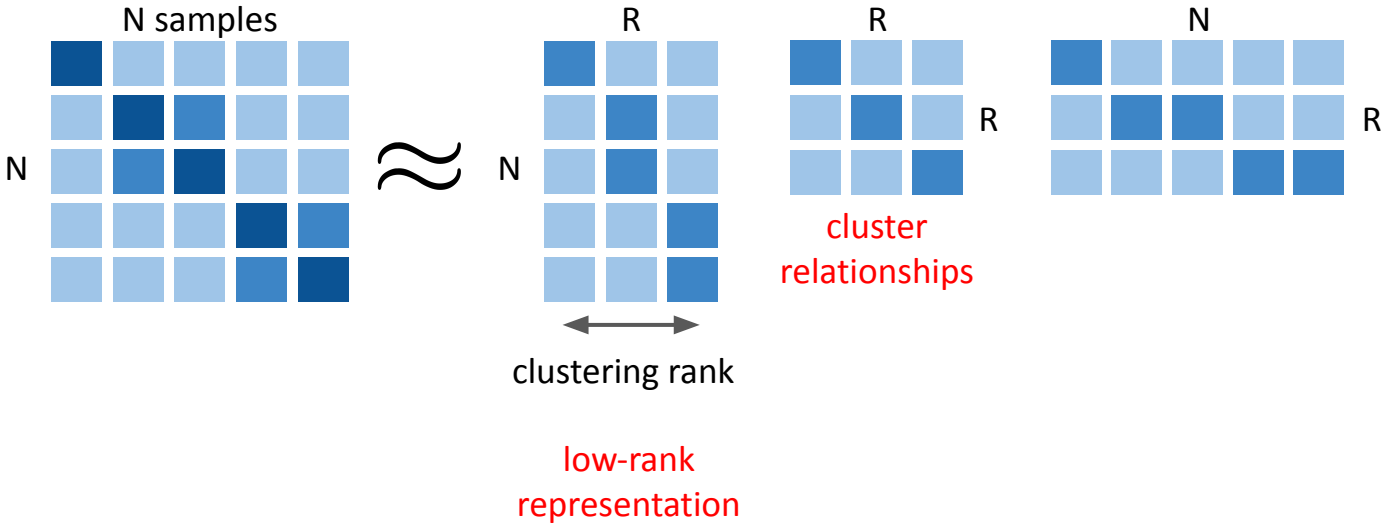
$$X \approx WH$$



SUMO: Factorization

NMF	$X \approx WH$
Symmetric NMF	$A \approx HH^T$

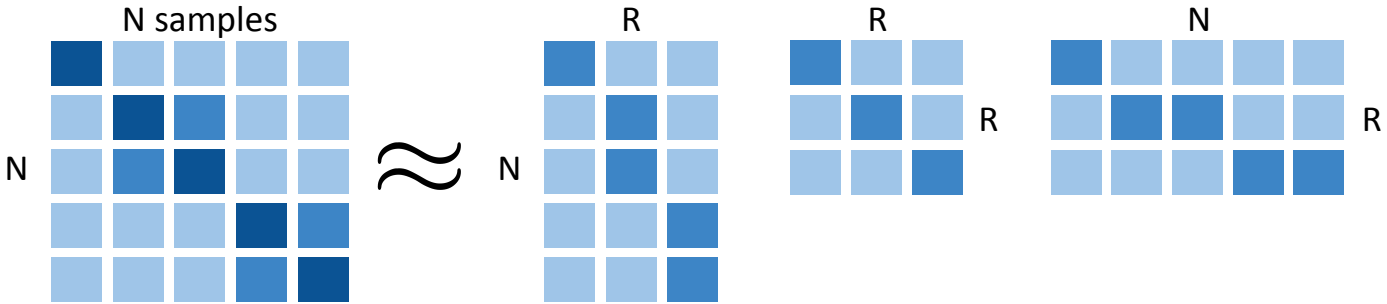
$$A^i \approx HS_iH^T$$



SUMO: Factorization

NMF	$X \approx WH$
Symmetric NMF	$A \approx HH^T$

$$A^i \approx HS_iH^T$$



$$\min_{H,\ S_i \geq 0} \sum_i \lambda_i \left|\left| W_i \circ (A_i - HS_iH^T) \right|\right|_F^2 + \eta ||H||_F^2$$

different number of samples

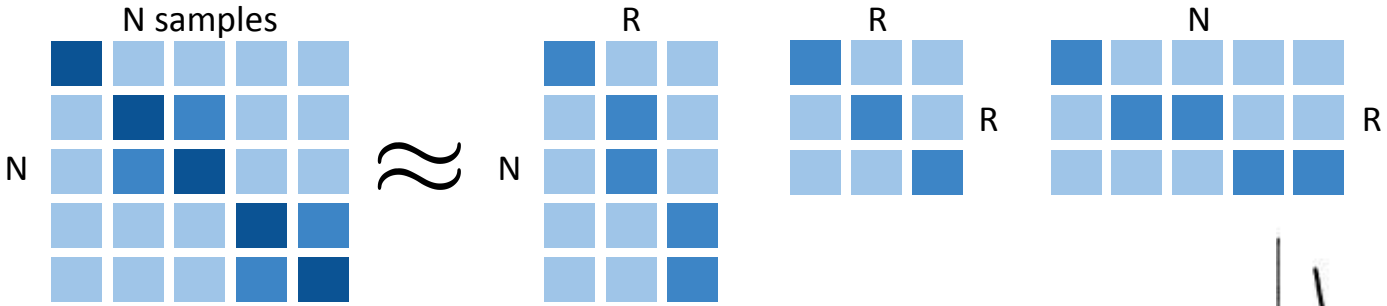
missing data

sparsity

SUMO: Factorization

NMF	$X \approx WH$
Symmetric NMF	$A \approx HH^T$

$$A^i \approx HS_iH^T$$

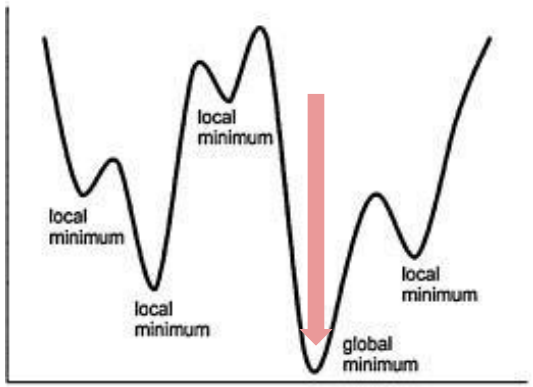


$$\min_{H,\ S_i \geq 0} \sum_i \lambda_i \left|\left| W_i \circ (A_i - HS_iH^T) \right|\right|_F^2 + \eta ||H||_F^2$$

different number of samples

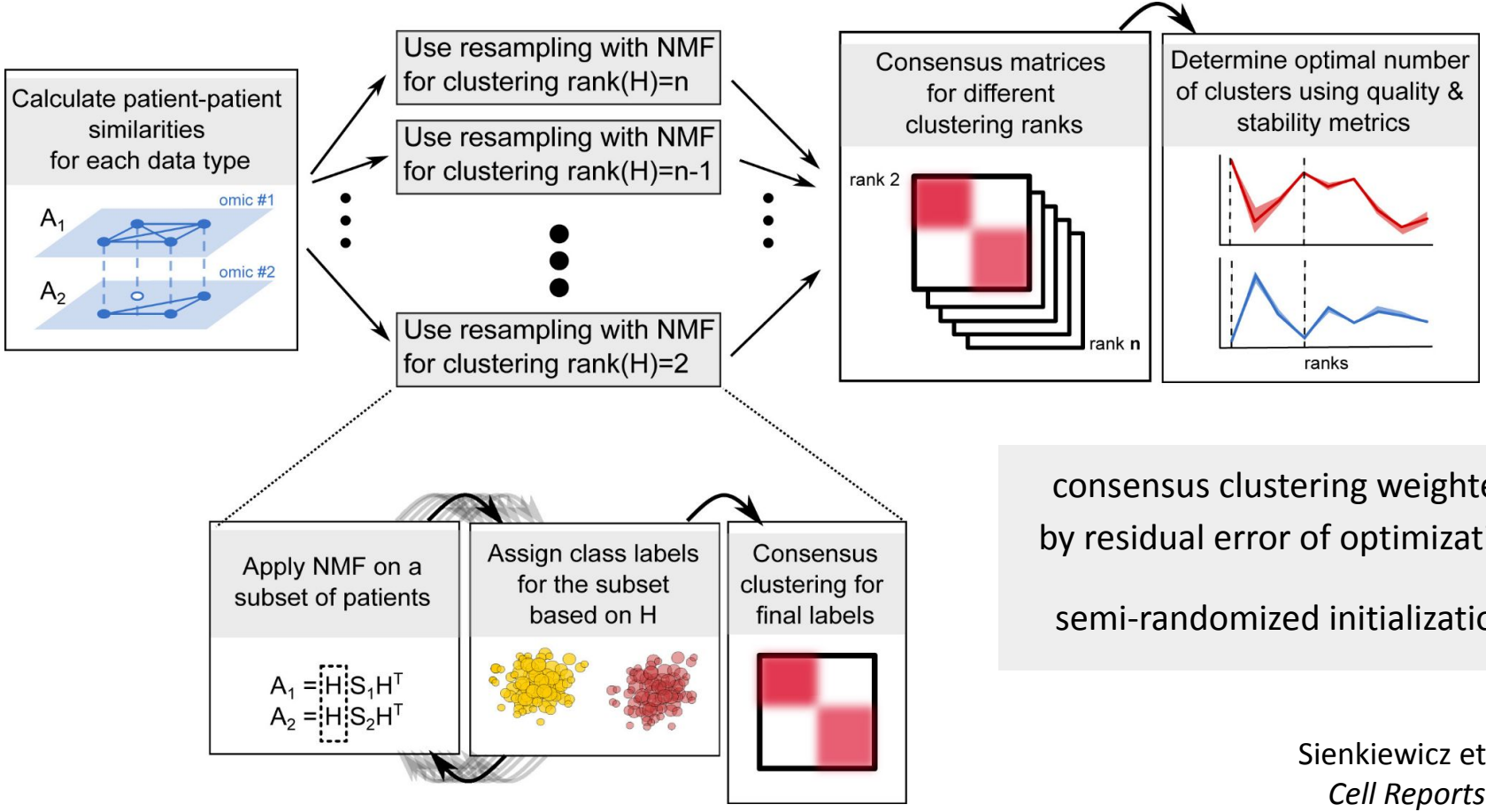
missing data

sparsity



using gradient descent

SUMO: Re-sampling based approach

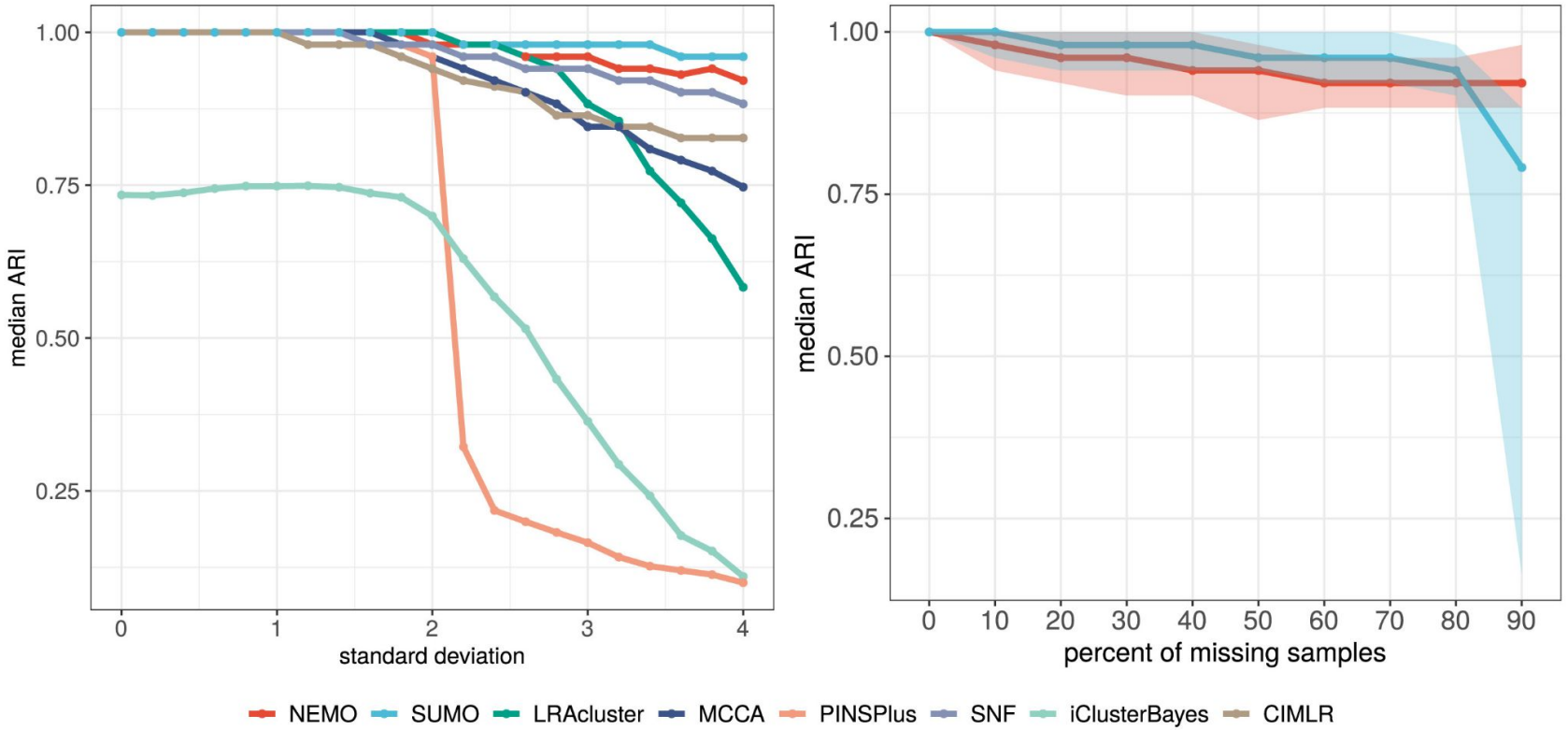


consensus clustering weighted by residual error of optimization

semi-randomized initialization

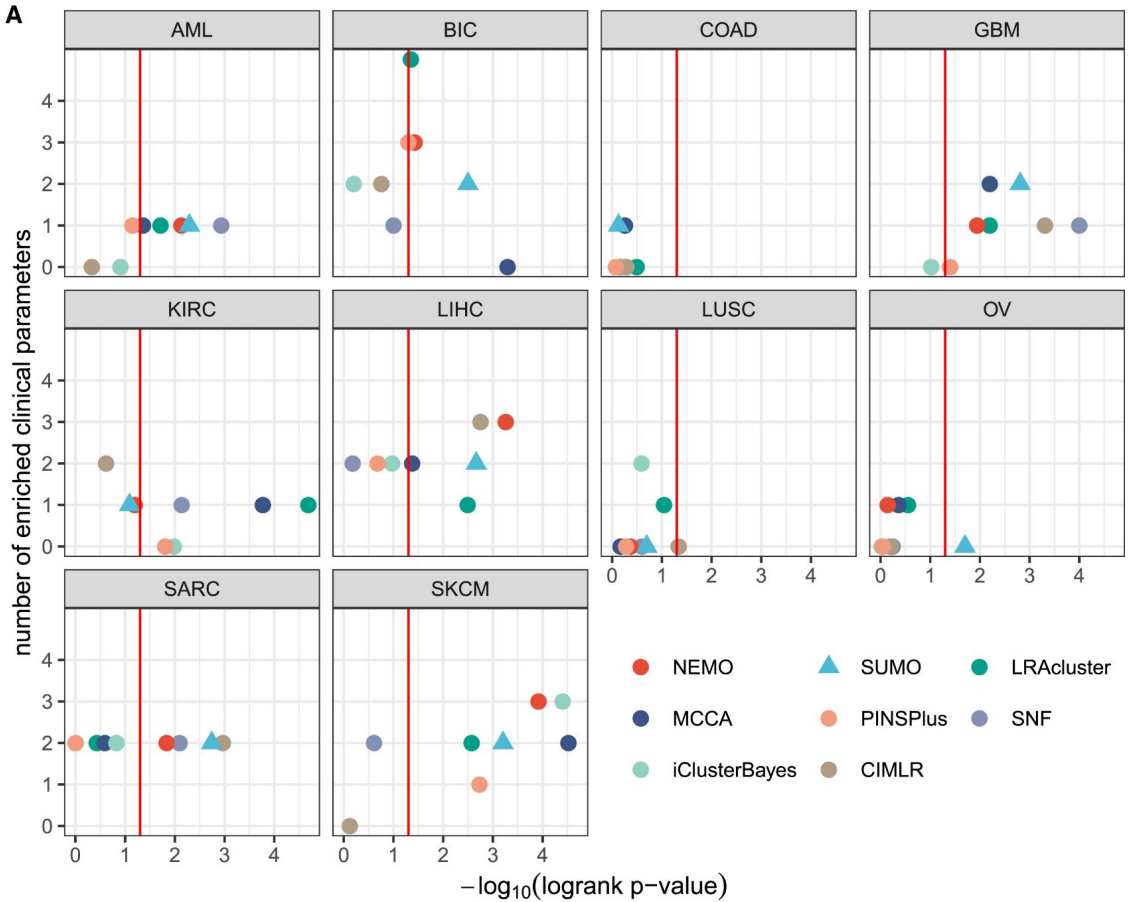
Sienkiewicz et al (2022)
Cell Reports Methods

SUMO improves performance with noisy and incomplete data



Sienkiewicz et al (2022) *Cell Reports Methods*

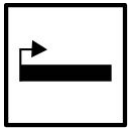
SUMO consistently identifies differential subtypes in TCGA data using an established benchmark



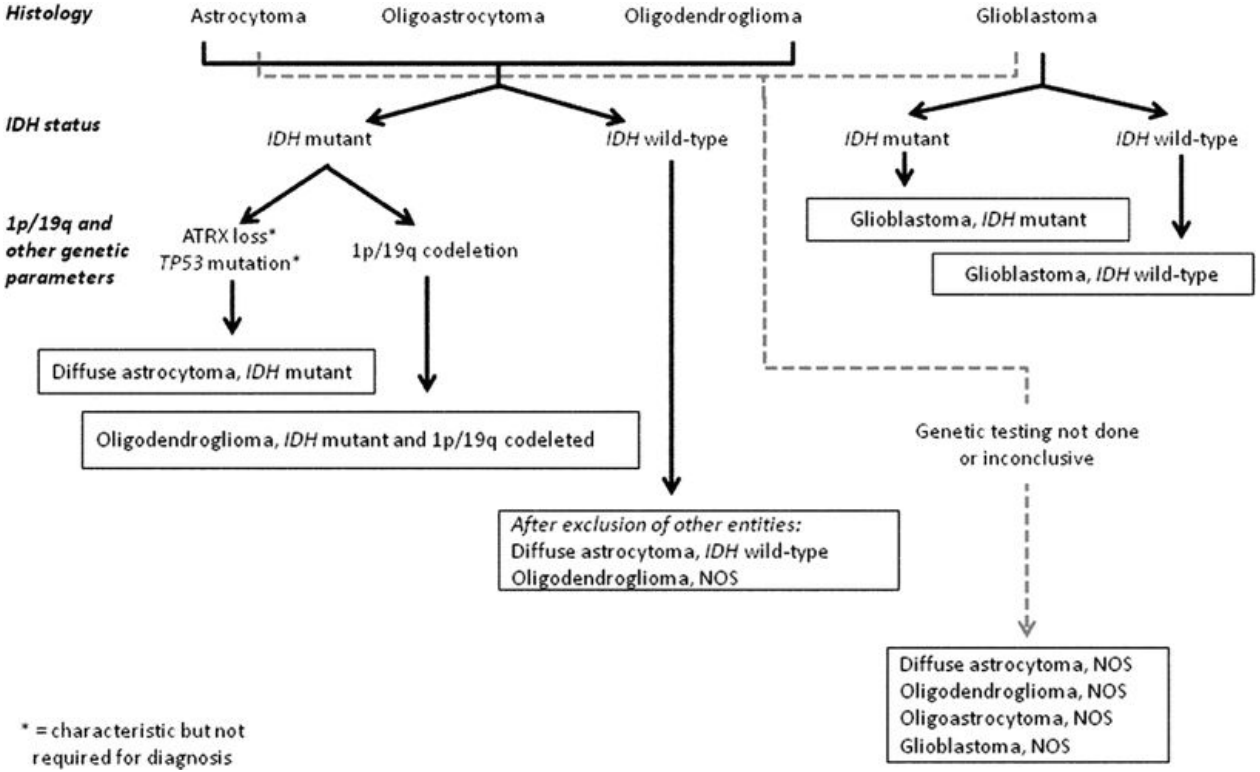
Sienkiewicz et al (2022) *Cell Reports Methods*

SUMO detects latent relationships between patients in Lower-Grade Gliomas

TCGA-LGG cohort
(a total of
556 primary tumors)

omic	# samples	tech
 RNA-seq	530	Illumina HiSeq
 DNA methylation	530	Illumina Infinium HumanMethylation450
 miRNA expression	524	Illumina HiSeq

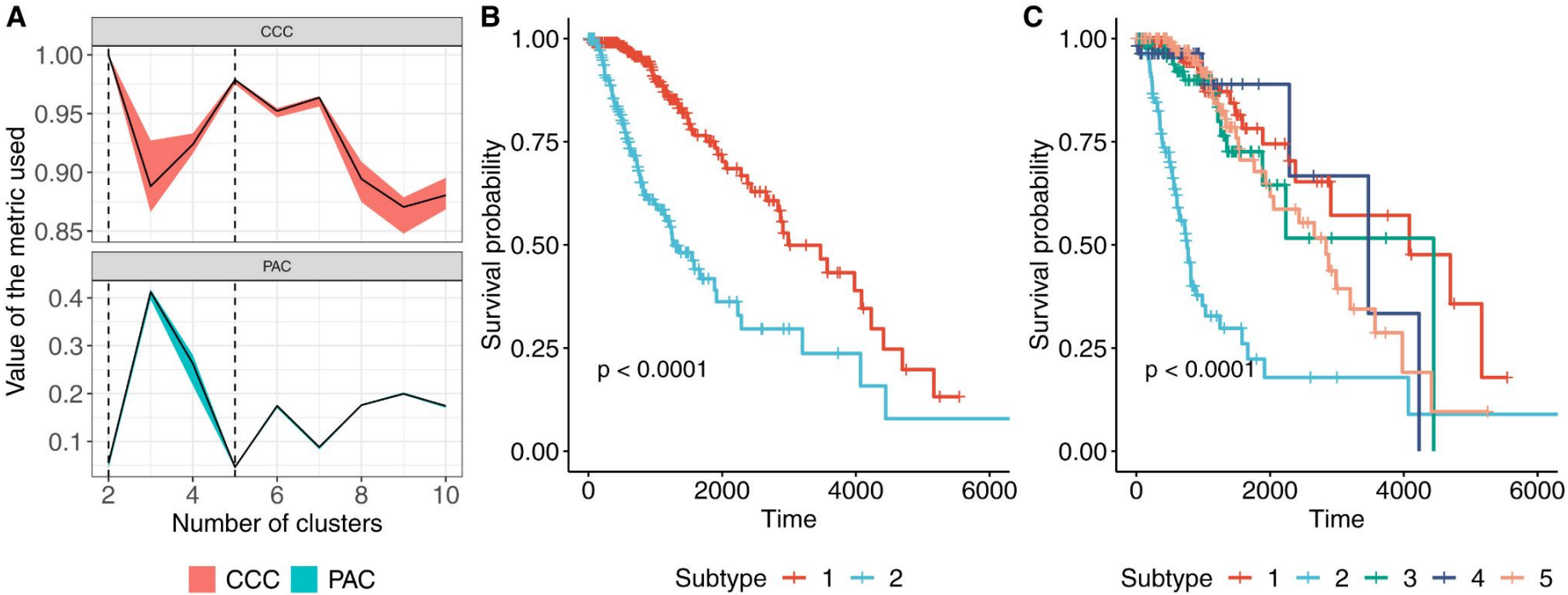
WHO guidelines for Low-Grade Gliomas



- Gliomas are neuroepithelial tumors originating from glial cells
- Postoperative RT remains the standard of care for adult grade 2/3
- ~80% have mutations in IDH1, 90% of those are R132H.
- ~50% have mutations in TP53, 40% have mutations in ATRX and 20% in CIC

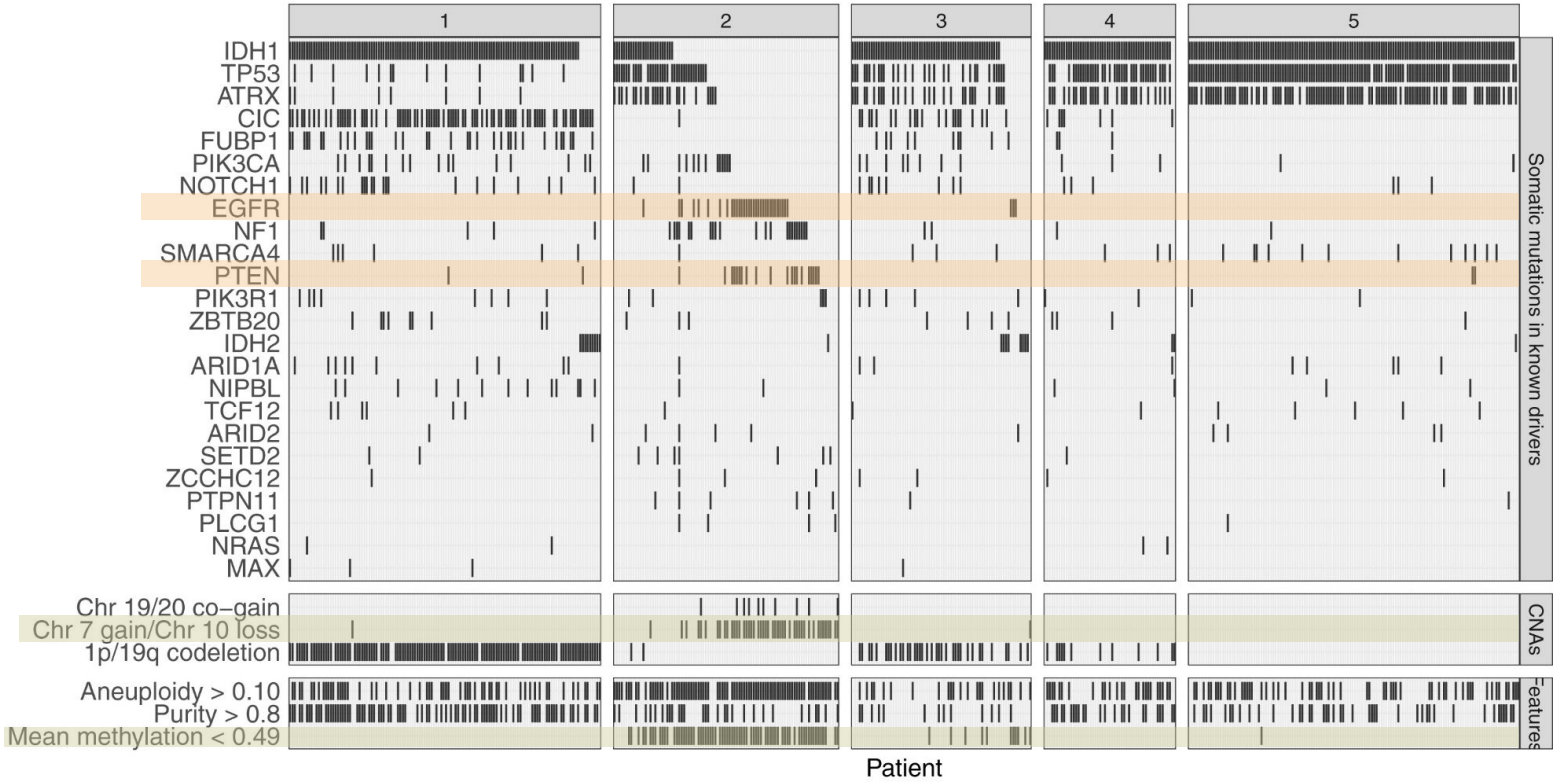
Source: The 2016 World Health Organization Classification of Tumors of the Central Nervous System (PMID: 27157931)

SUMO identifies a subtype of LGG with differential prognosis & GBM-like features



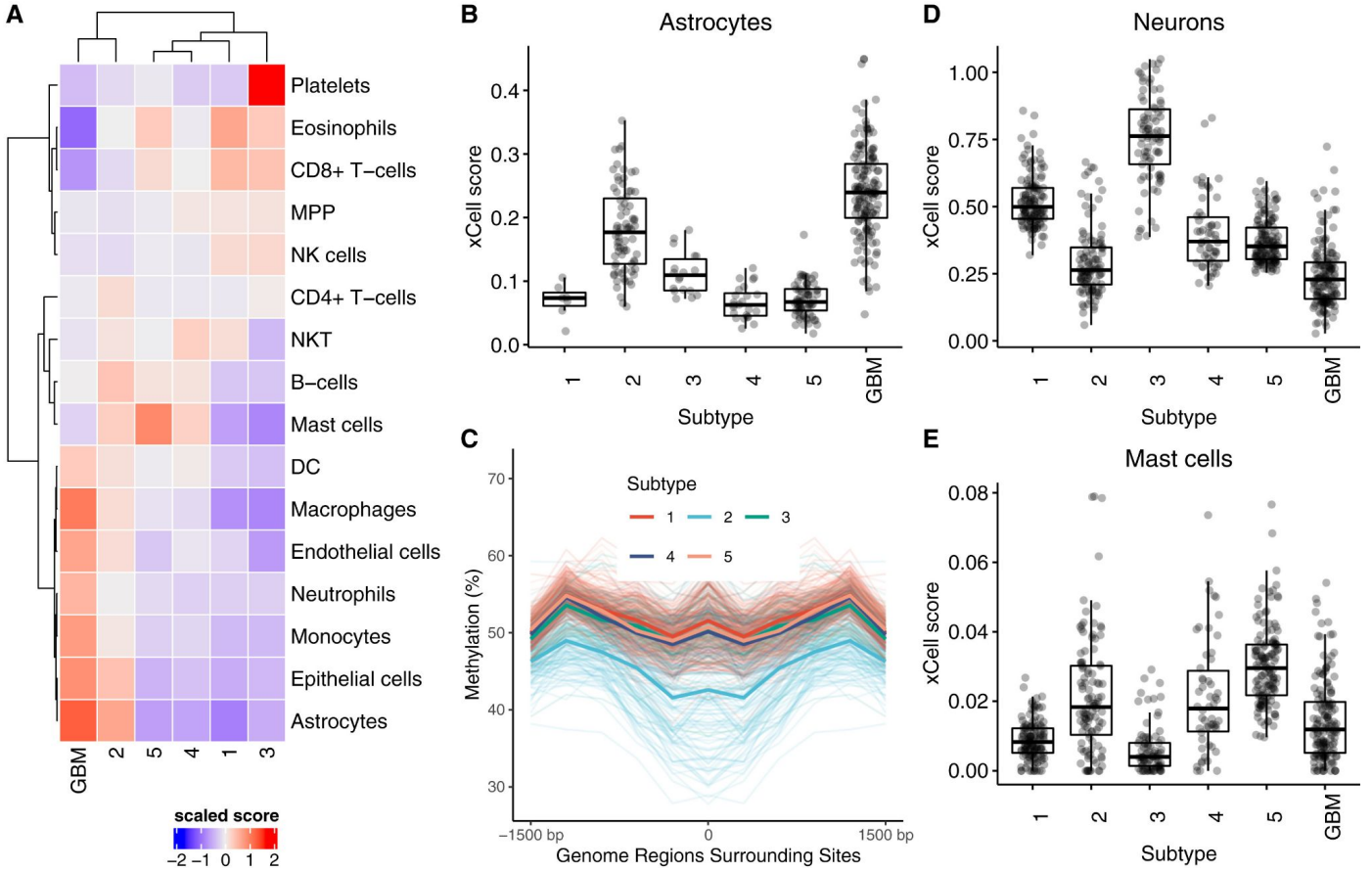
Sienkiewicz et al (2022) *Cell Reports Methods*

SUMO identifies a subtype of LGG with differential prognosis & GBM-like features



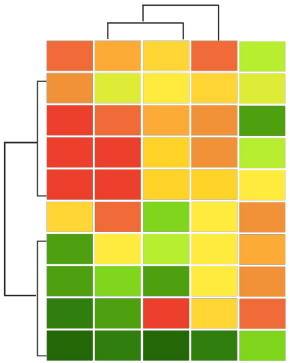
Sienkiewicz et al (2022) *Cell Reports Methods*

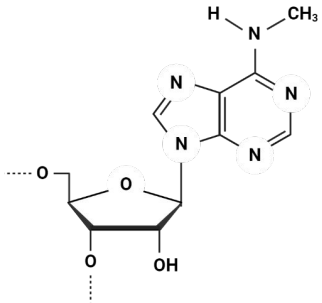
SUMO subtype 2 cellular profile is more similar to GBMs than other LGGs



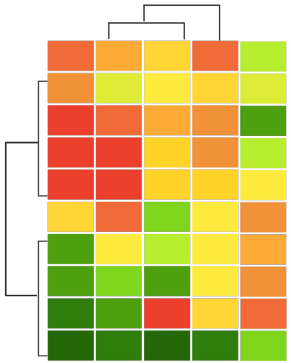
Sienkiewicz et al (2022) *Cell Reports Methods*

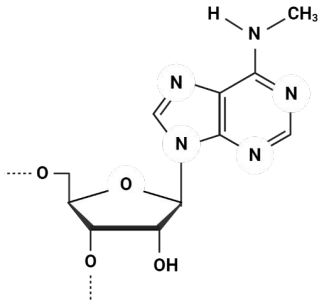
Single-molecule multi-omics?

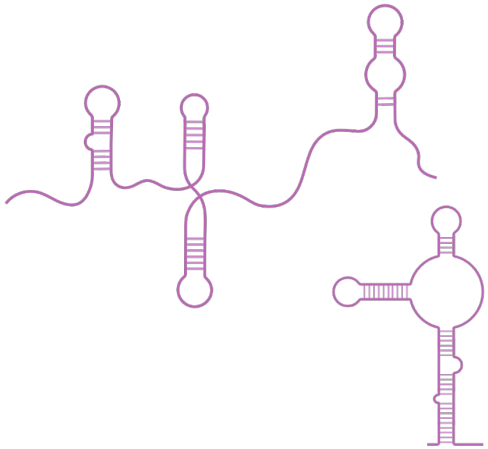




Single-molecule multi-omics?

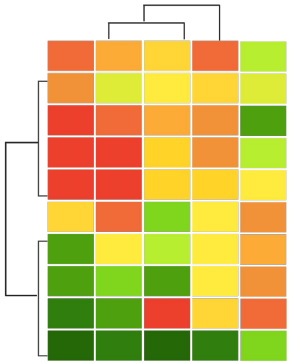


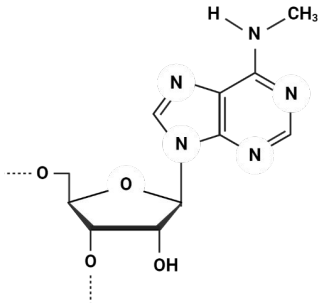


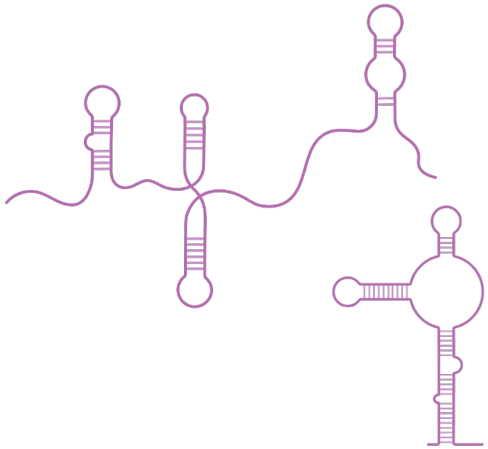


- matching transcript expression with base modifications
- multiple distinct base modification

Single-molecule multi-omics?



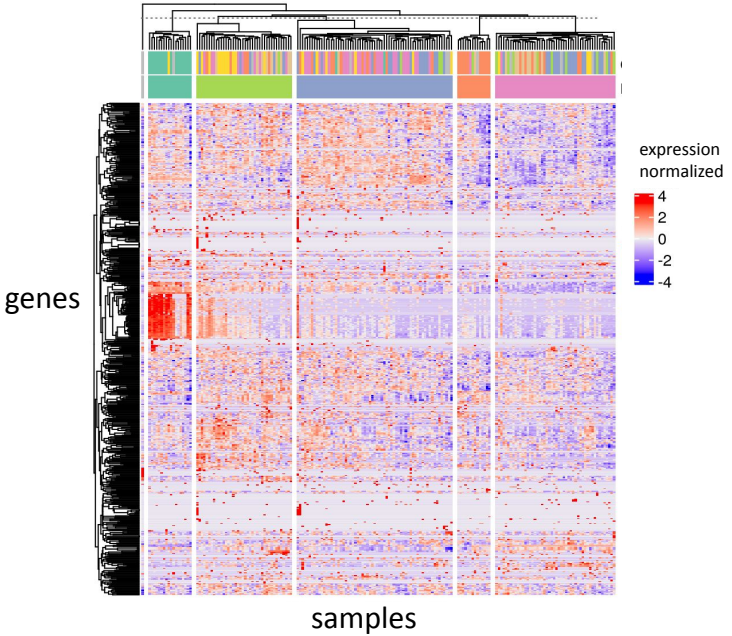




- matching transcript expression with base modifications
- multiple distinct base modification
- transcripts as features vs transcripts as samples
- secondary structure prediction and validation

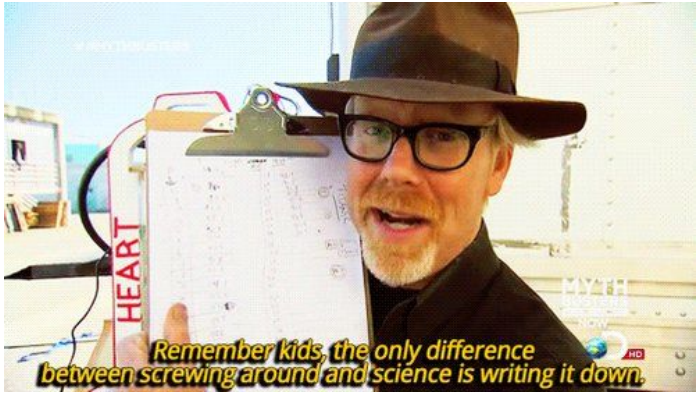
Conclusions

- Heatmaps are your friend



Conclusions

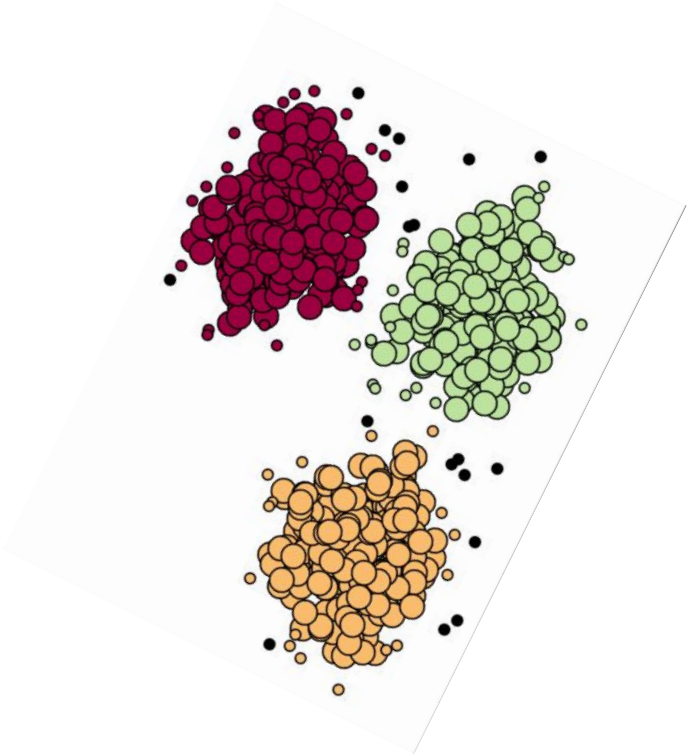
- Heatmaps are your friend
- Analysis-aware experimental design is essential
- Keep track of the metadata



Adam Savage

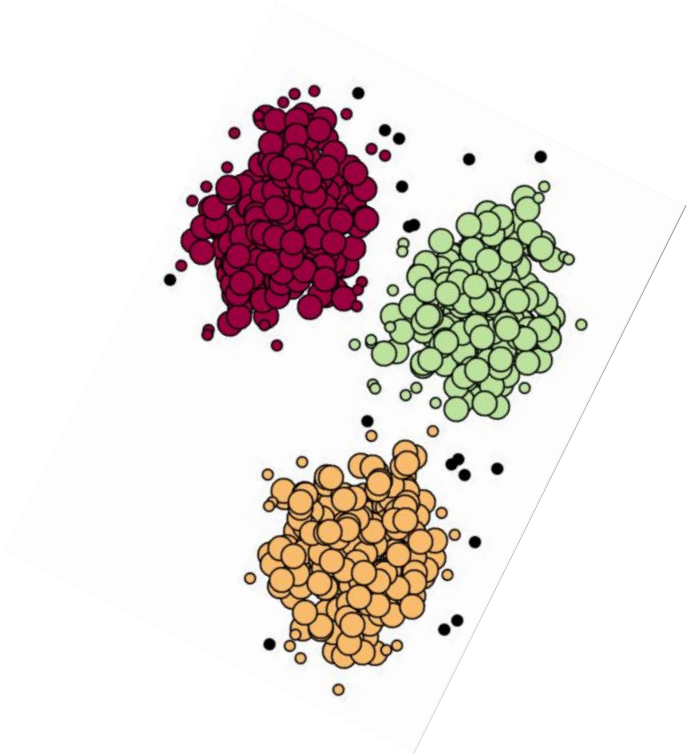
Conclusions

- Heatmaps are your friend
- Analysis-aware experimental design is essential
- Keep track of the metadata
- Each model has assumptions
- Don’t underestimate the value of simulated data



Conclusions

- Heatmaps are your friend
- Analysis-aware experimental design is essential
- Keep track of the metadata
- Each model has assumptions
- Don’t underestimate the value of simulated data
- Single-molecule omics are on the way!



Acknowledgements

SUMO Team

University of Virginia
Aakrosh Ratan
Ajay Chatrath
Nathan Sheffield
John Lawson

National University of Singapore
Jinyu Chen
Louxin Zhang

WCM
Mason Lab
Melnick Lab

SUMO Funding

Tri-Institutional PhD Program
Computational Biology & Medicine

Memorial Sloan Kettering
Cancer Center

Weill Cornell
Medicine