Statistical Methods for High-dimensional Biology

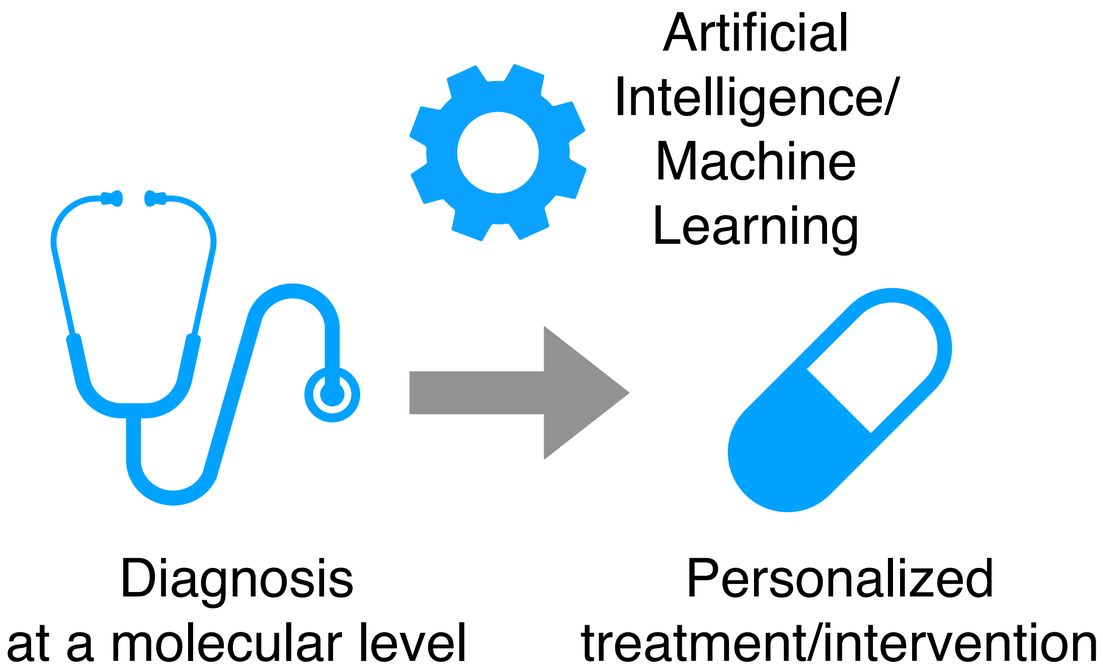
Multiomics data integration

Yongjin Park, UBC Path&Lab, STAT, BC Cancer

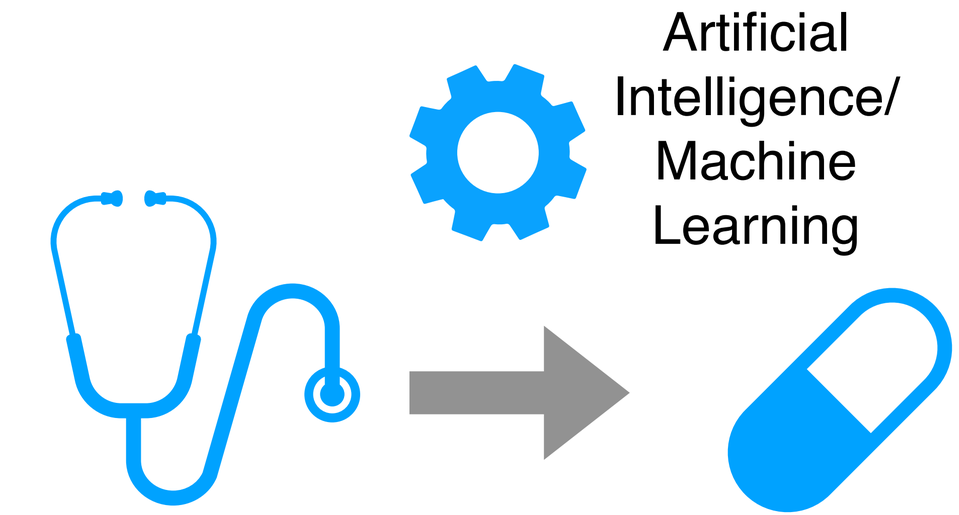
Today's lecture: Multiomics data integration

- **Why do we do multiomics data integration?**
  - view #1: borrowing information across modalities
  - view #2: efforts to provide mechanistic explanations
- **Global, unsupervised multiomics data integration**
  - Multiomics Factorization (and variants)
  - Network-based data integration
- **Local, linking between layers to understand mechanisms**
  - Deep dive into mechanisms of gene regulatory mechanisms

Genomic Medicine: Promises and Challenges



Genomic Medicine: Promises and Challenges



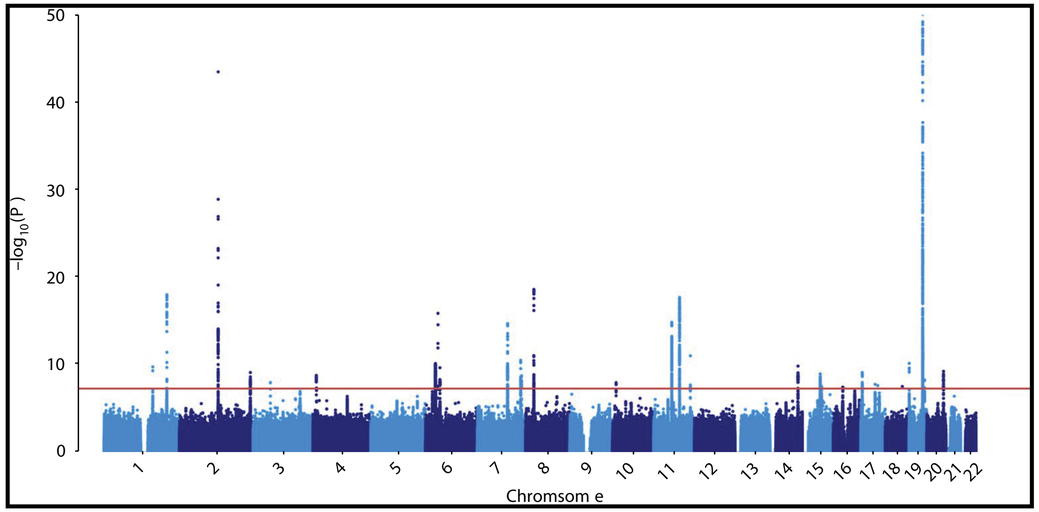
Artificial Intelligence/ Machine Learning

Diagnosis at a molecular level

Personalized treatment/intervention

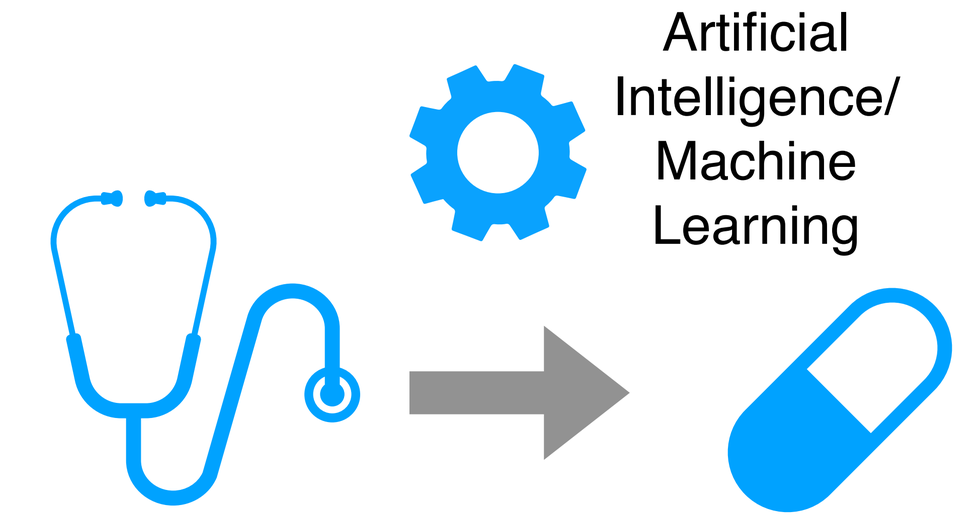
Genome-wide association studies

of complex disease traits



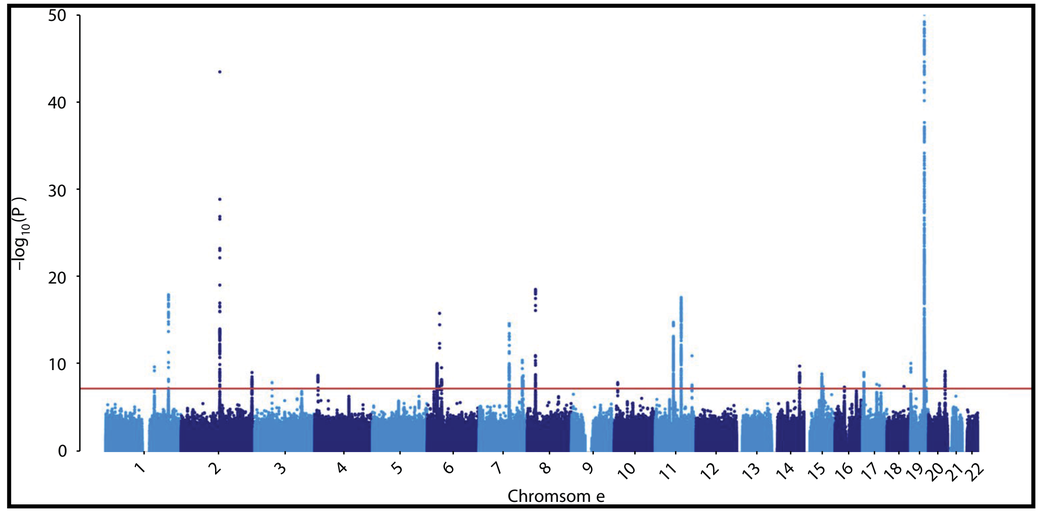
Lectures 13, 14, 15

Genomic Medicine: Promises and Challenges



Diagnosis
at a molecular level

Personalized
treatment/intervention



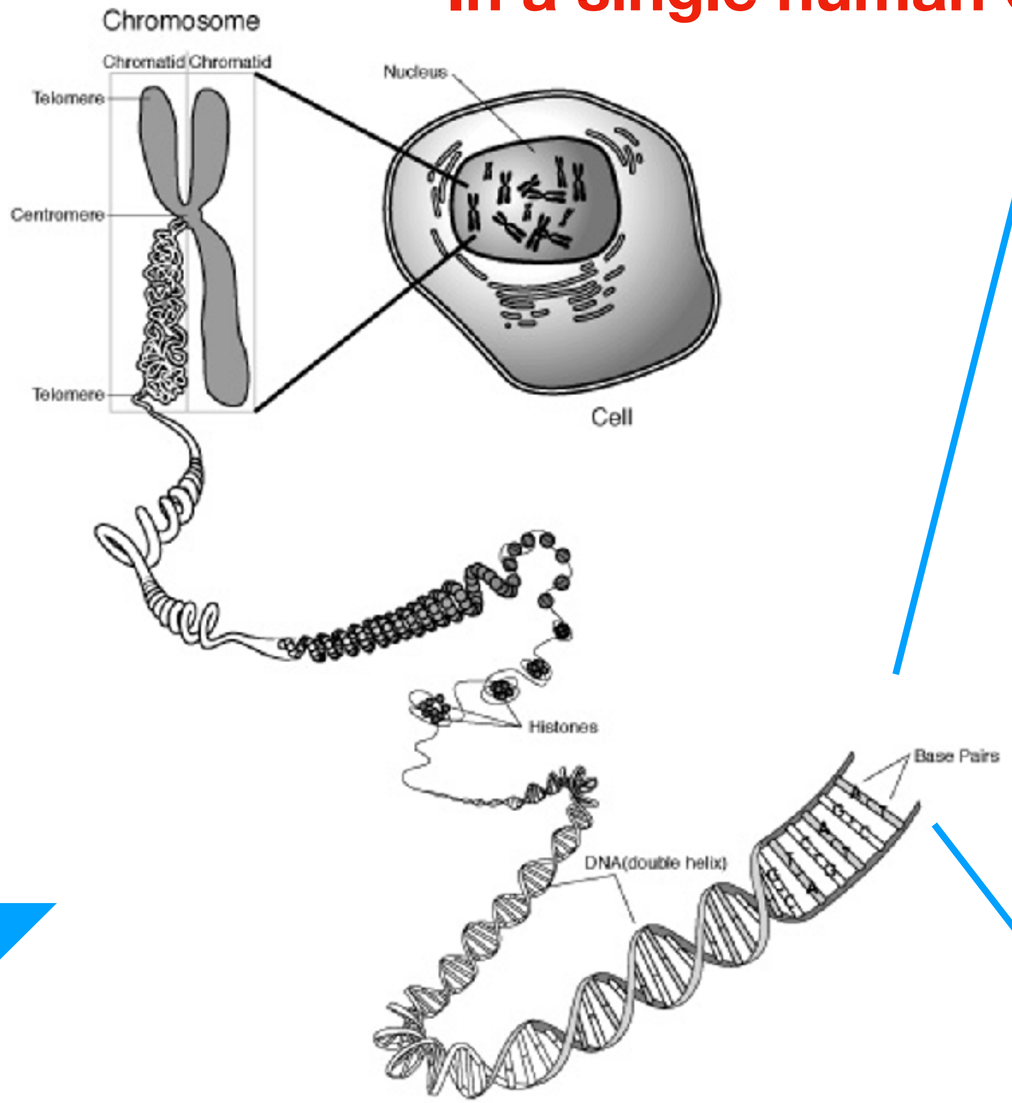
- 90% of disease hits are difficult to interpret.
- Causal variants & target genes, pathways are unknown.
- Target cell types are unknown.

- Not all disease mechanisms are observable.
- “Causal” interventions are often too invasive.
- Networks are highly inter-connected.
- Heterogeneity across individuals.

Where is the source of complexity?

In a single human cell (n=1)

zoom-in down to the DNA base pair level



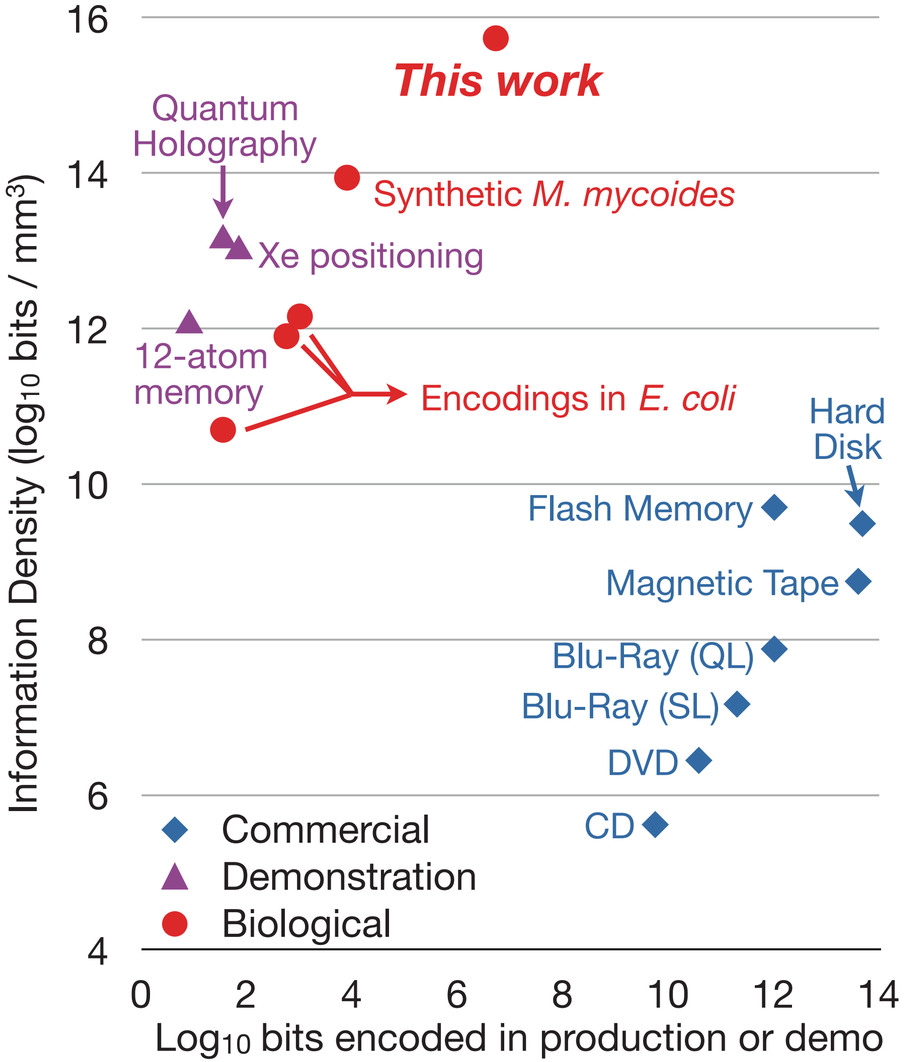
- **3.2B** base pairs (ATGC characters)
- **30k** mRNA/genes (protein coding) ± 5k to 100k non-coding
- **100+** Epigenetic modifications per gene regulatory element
- **4+** isoforms due to post-transcriptional modifications
- Post-translational modifications

DNA can store a lot of information

BREVIA

Next-Generation Digital Information Storage in DNA

George M. Church,[1,2] Yuan Gao,[3] Sriram Kosuri[1,2]*



Hard to write but easy to read for a cell

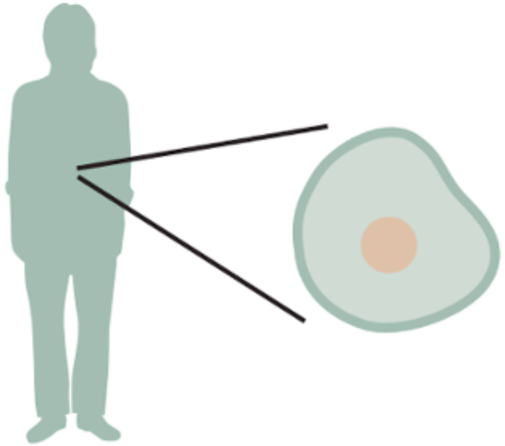
An efficient way to copy and edit

Highly parallelized operations

A long history of evolution

Trillions of cells

$M \approx 100\,kg$



$V \approx 1000 - 10{,}000\,\mu m^3$

$\Downarrow$

$m \approx 10^{-12} - 10^{-11}\,kg$

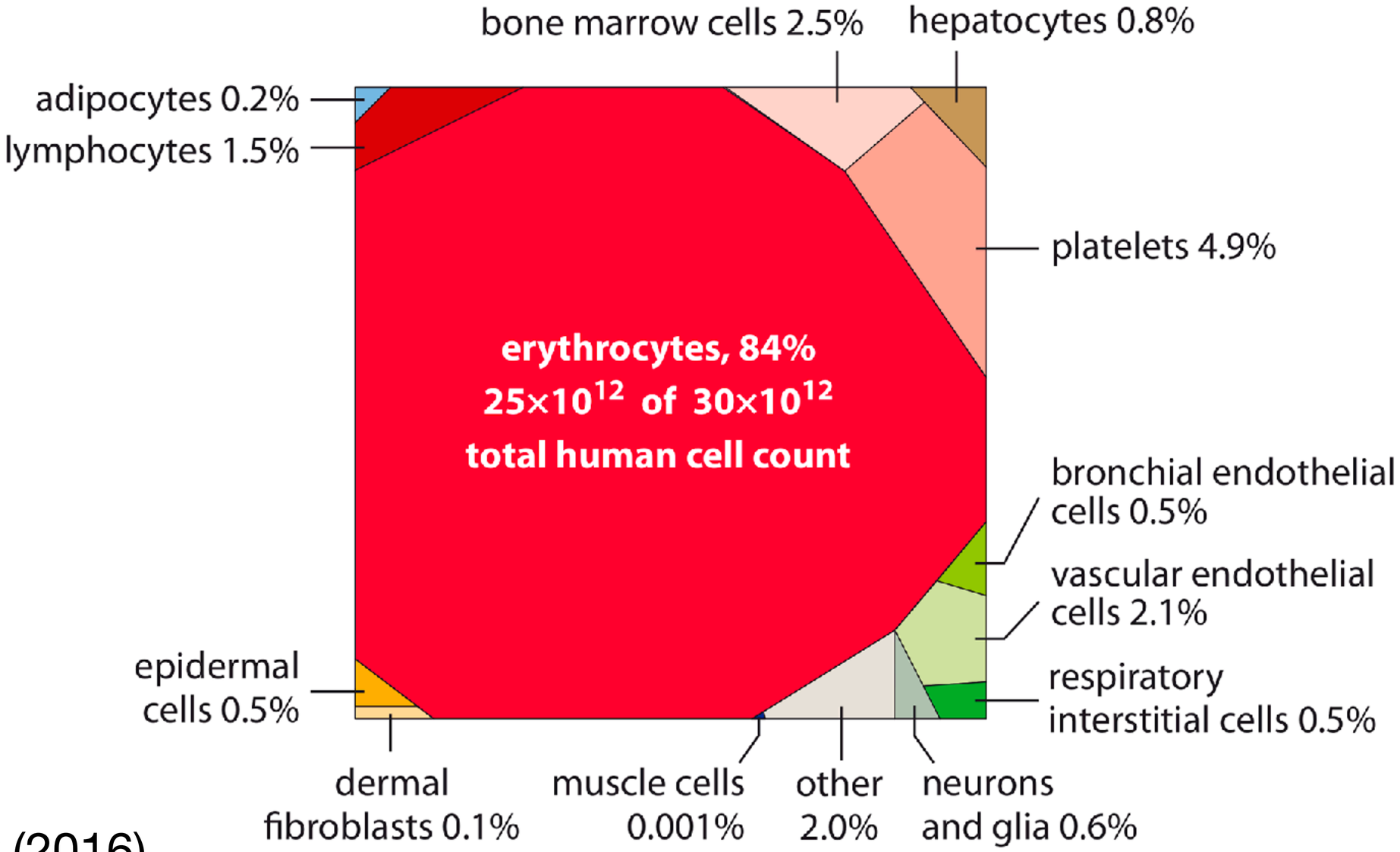
$\Downarrow$

$$N \approx \frac{M}{m} = \frac{100\,kg}{10^{-12} - 10^{-11}\,kg} = 10^{13} - 10^{14}$$

Sender *et al*.

Plos Biology (2016)

> 30 trillions of such complexity



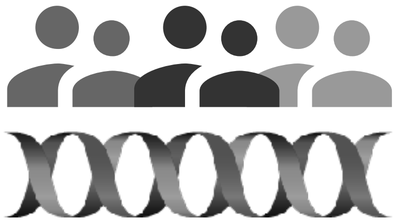
Sender *et al*.

Plos Biology (2016)

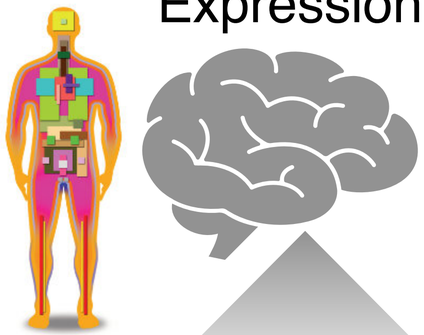

Where is
CCTTCTGTG
TCGGA

We need to understand how it works

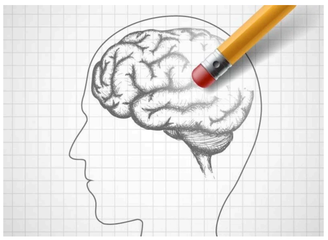
Genotype information (X)



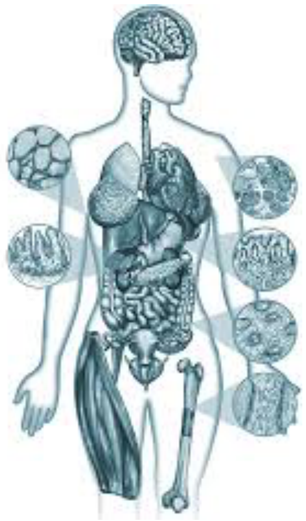
Expression (M)



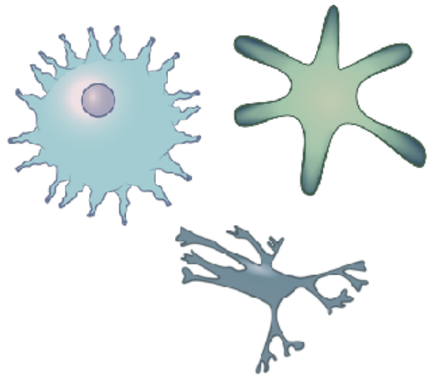
Phenotype (y)



interplay w/ epigenetic mediations

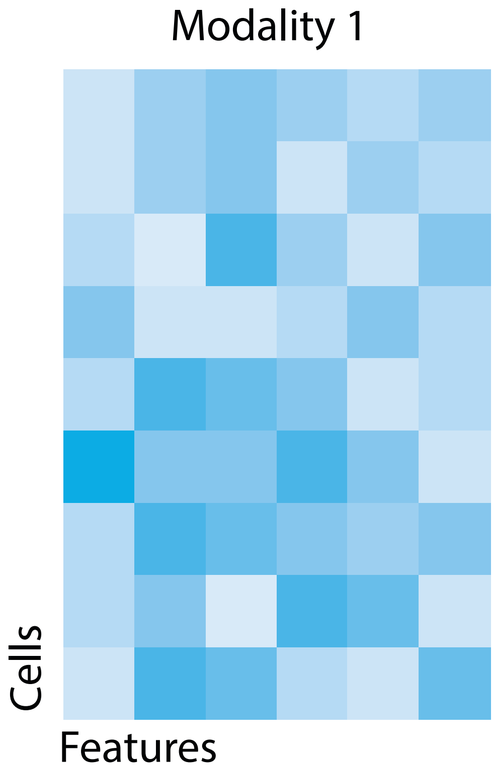


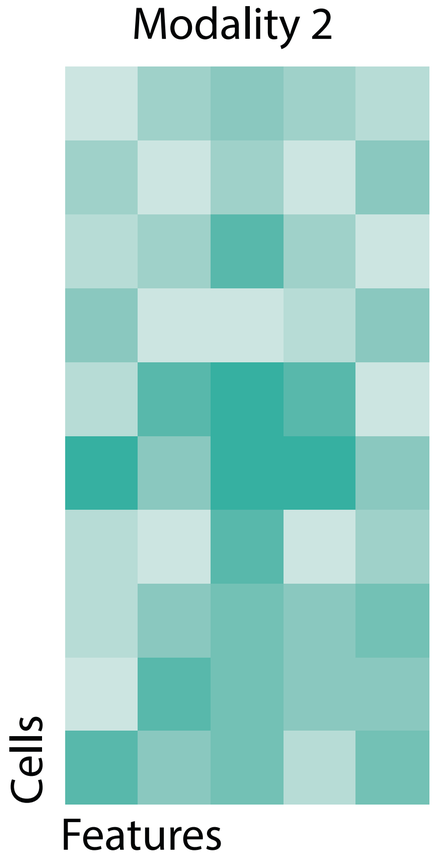
causal cell types



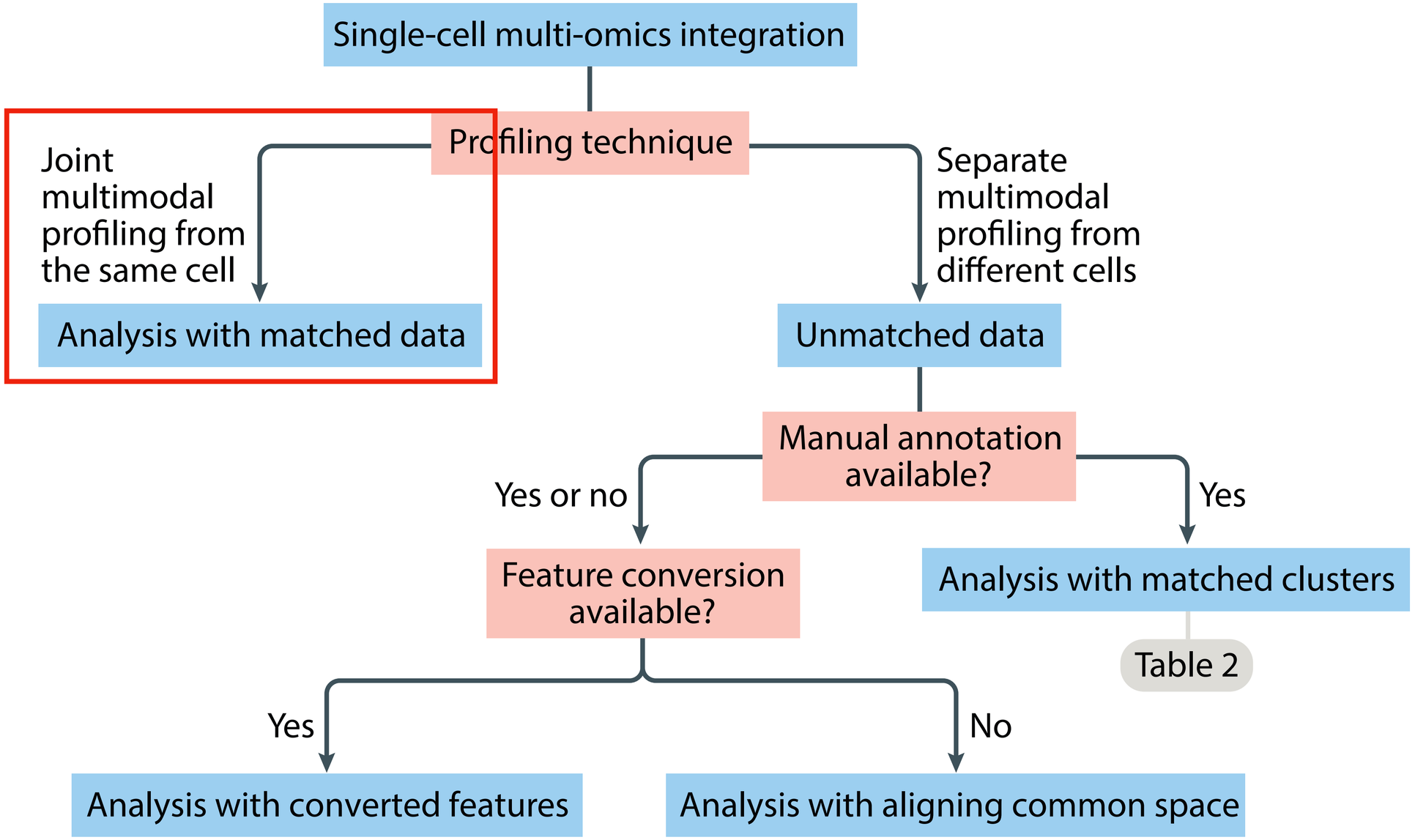
gene-gene interaction, gene-environment, trans-effect mediated by *cis*-effect (…)

Multimodal integration can go in several directions



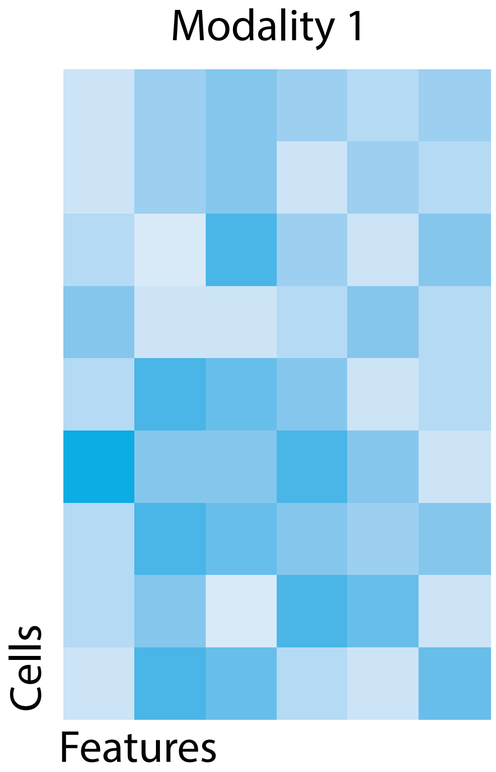


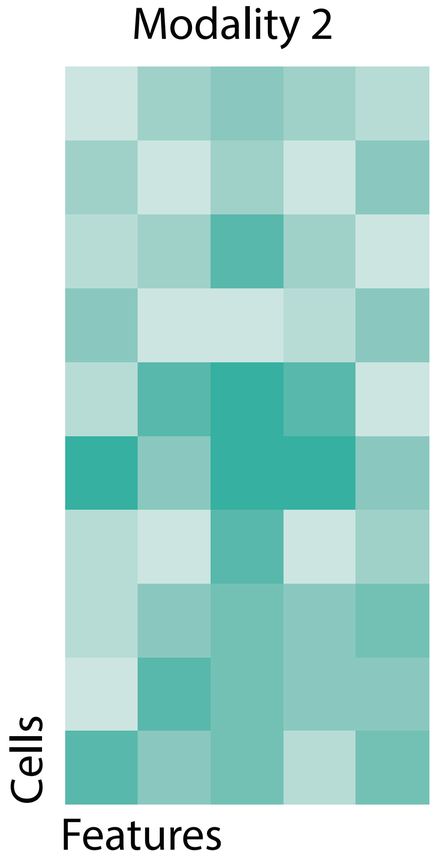
Miao .. Kim, Nat. Reviews Nephrology (2021)



Miao .. Kim, Nat. Reviews Nephrology (2021)

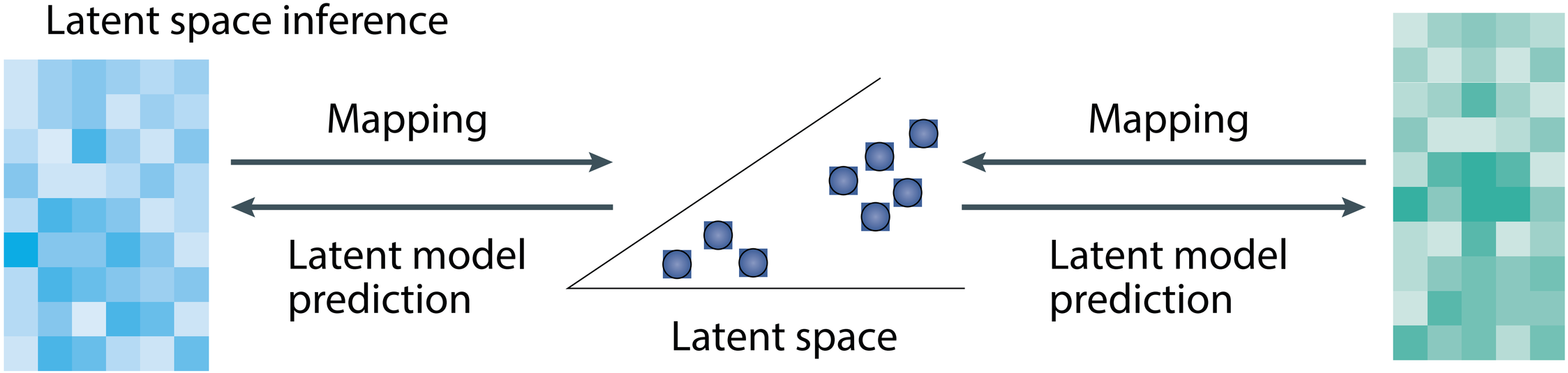
Multimodal integration can go in several directions





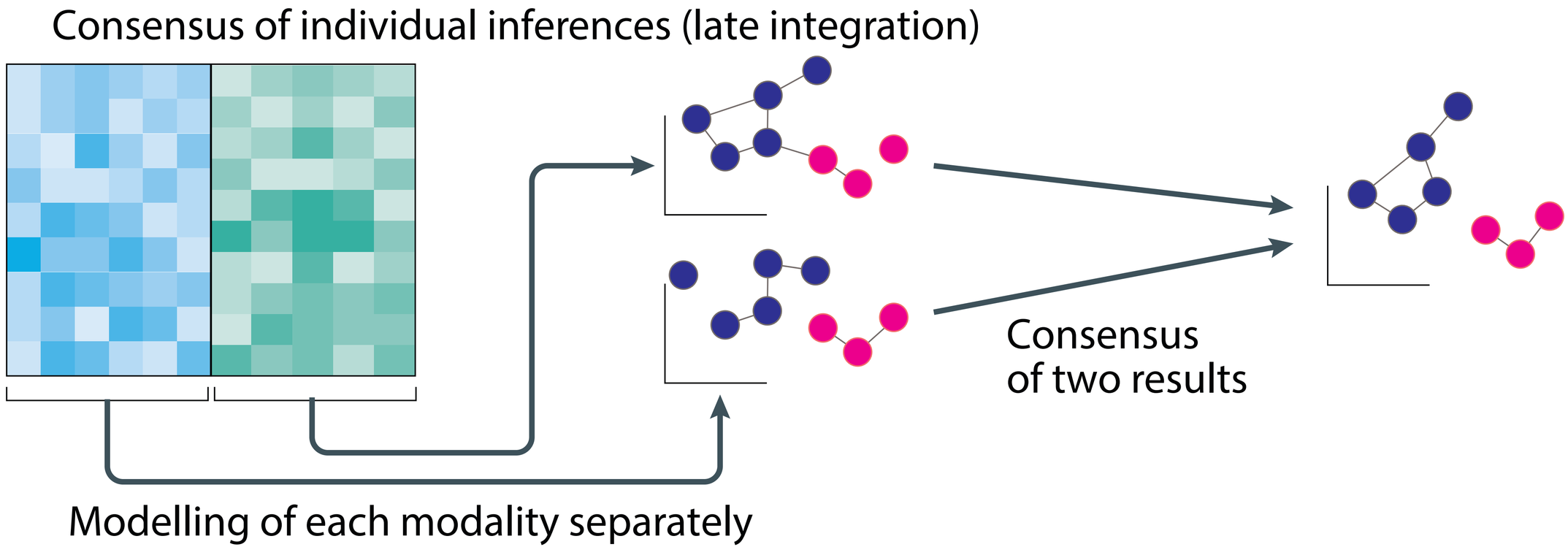
Miao .. Kim, Nat. Reviews Nephrology (2021)

Different types may implicate different aspects of data generation process



Miao .. Kim, Nat. Reviews Nephrology (2021)

If we have multiple types of data for each cell…
we can identify consensus groups/modules of cells

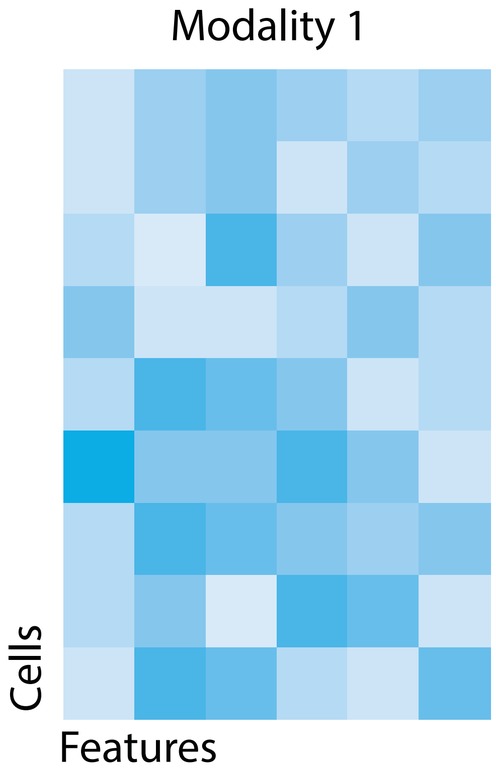


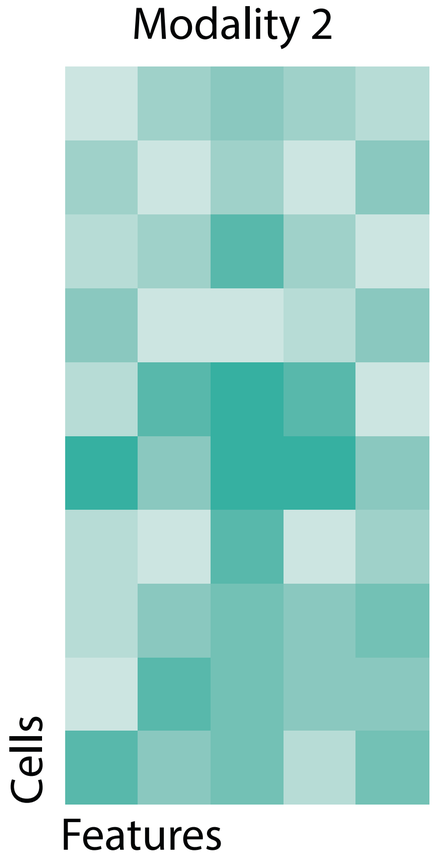
Miao .. Kim, Nat. Reviews Nephrology (2021)

Today’s lecture: Multiomics data integration

- **Why do we do multiomics data integration?**
  - view #1: borrowing information across modalities
  - view #2: efforts to provide mechanistic explanations
- **Global, unsupervised multiomics data integration**
  - Multiomics Factorization (and variants)
  - Network-based data integration
- **Local, linking between layers to understand mechanisms**
  - Deep dive into mechanisms of gene regulatory mechanisms

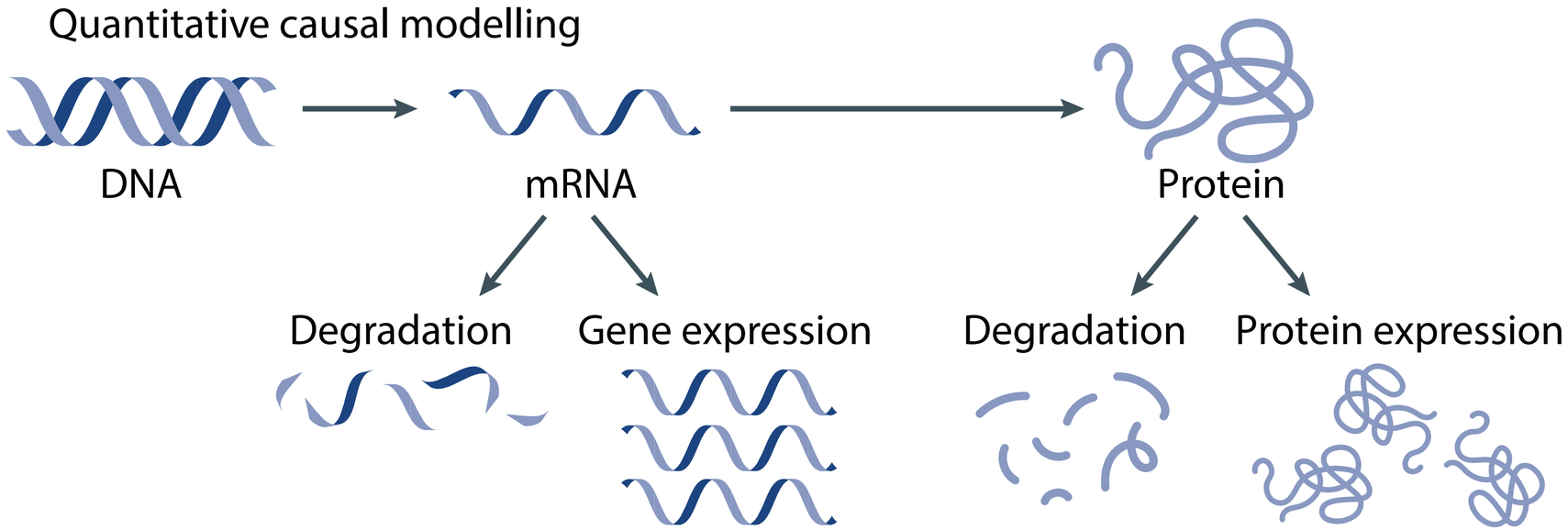
Multimodal integration can go in several directions





Miao .. Kim, Nat. Reviews Nephrology (2021)

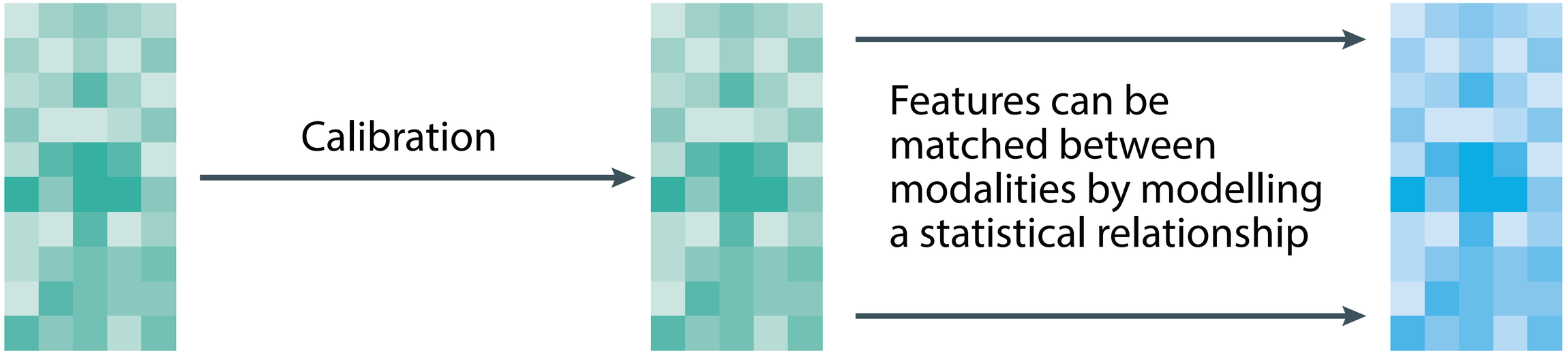
Ideally, we want to model causal relationships across multiple regulatory mechanisms



Miao .. Kim, Nat. Reviews Nephrology (2021)

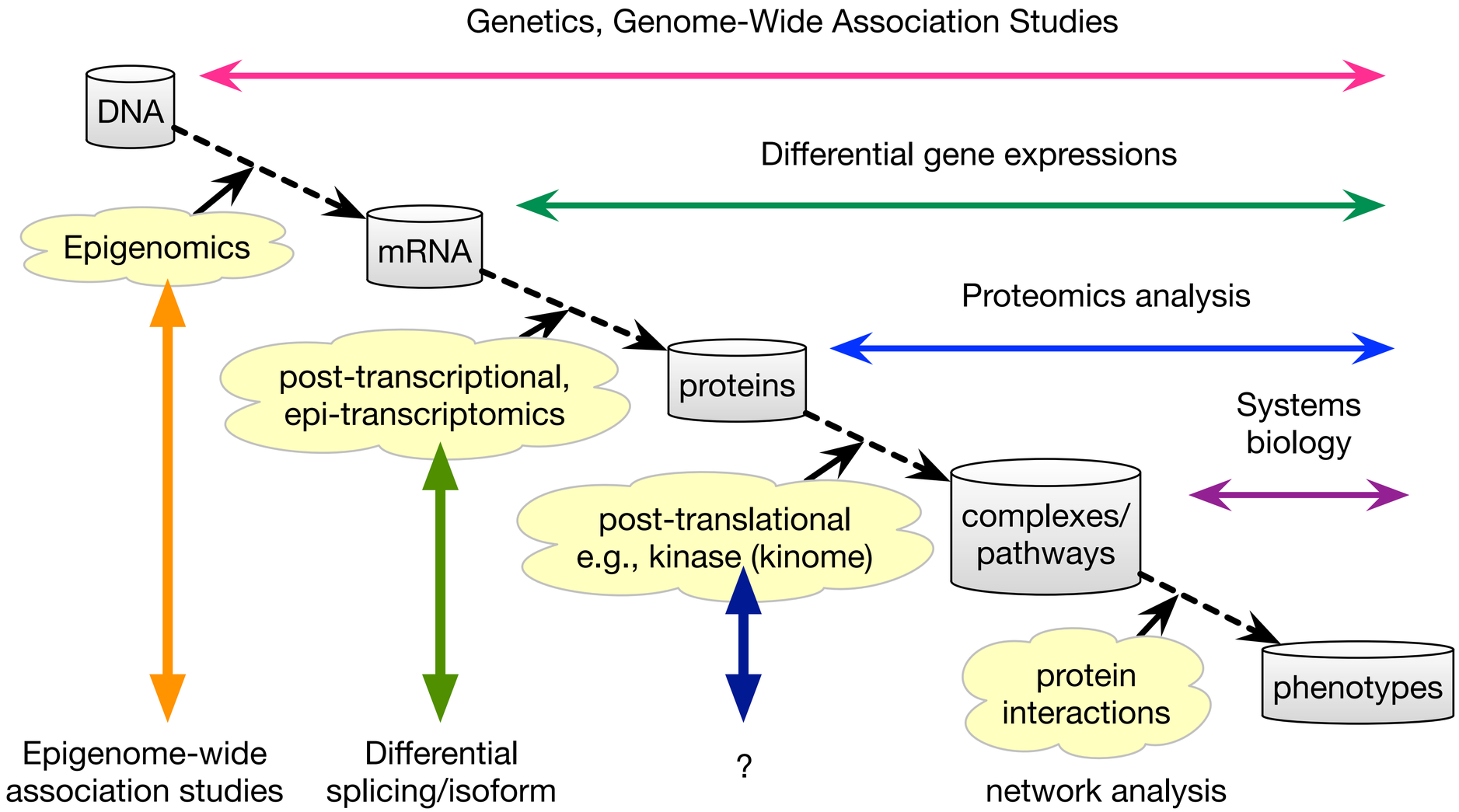
Can we simply match one feature to the other features?

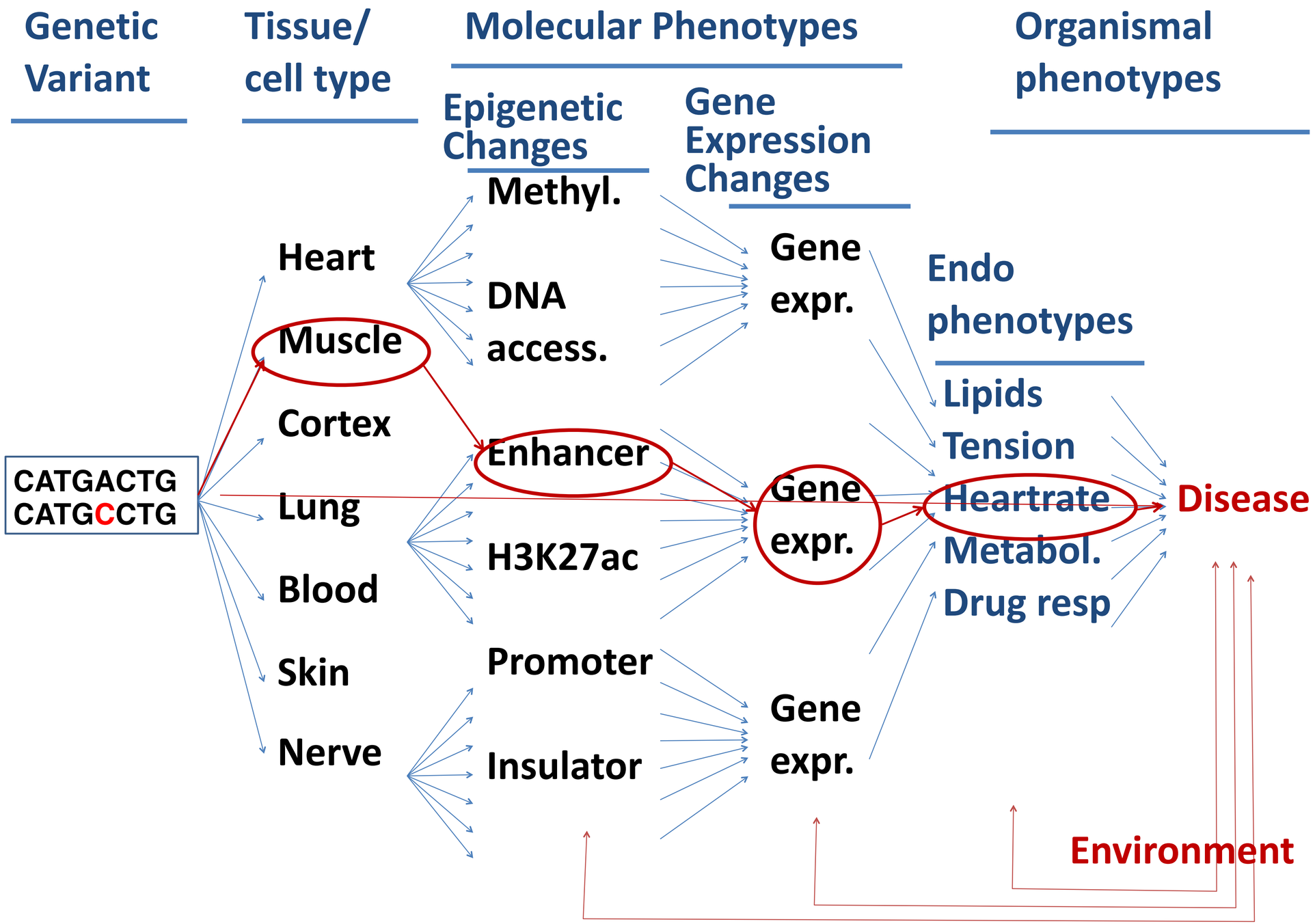
Statistical modelling between features



Miao .. Kim, Nat. Reviews Nephrology (2021)

Information flow and network across multiple layers of regulations



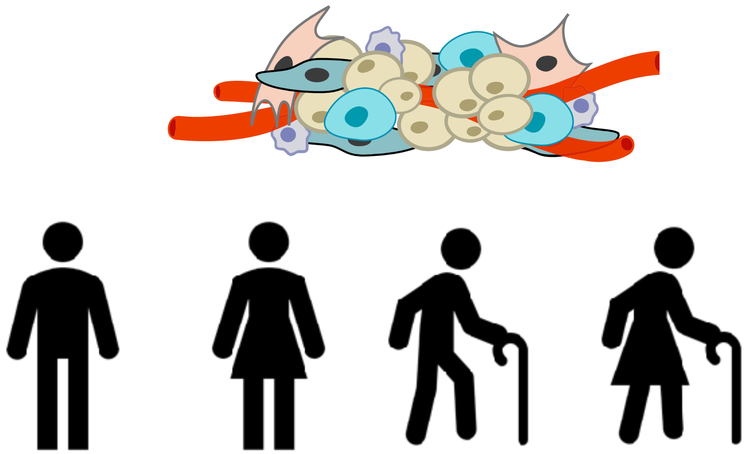


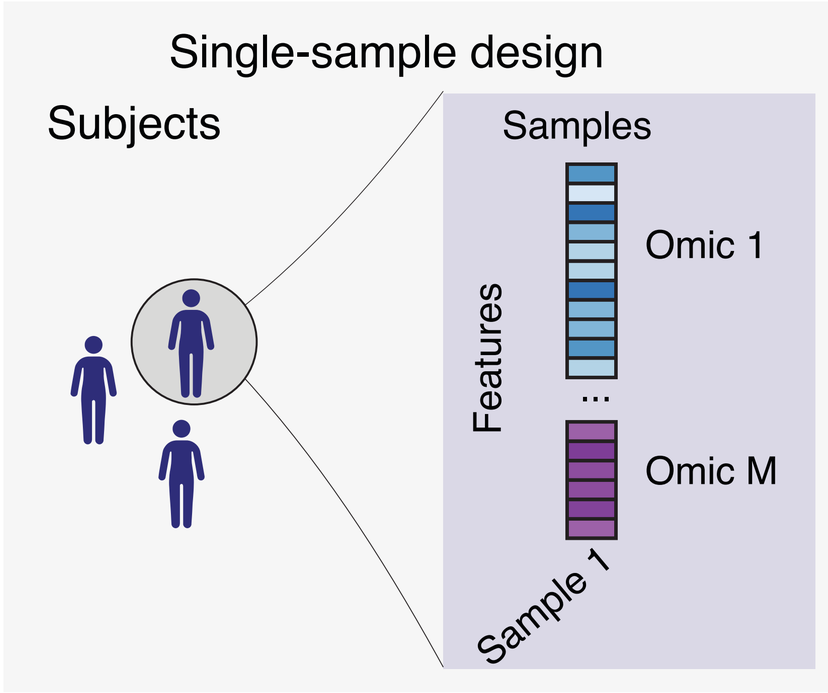
Slide credit: Manolis Kellis

Today's lecture: Multiomics data integration

- **Why do we do multiomics data integration?**
  - view #1: borrowing information across modalities
  - view #2: efforts to provide mechanistic explanations
- **Global, unsupervised multiomics data integration**
  - Multiomics Factorization (and variants)
  - Network-based data integration
- **Local, linking between layers to understand mechanisms**
  - Deep dive into mechanisms of gene regulatory mechanisms

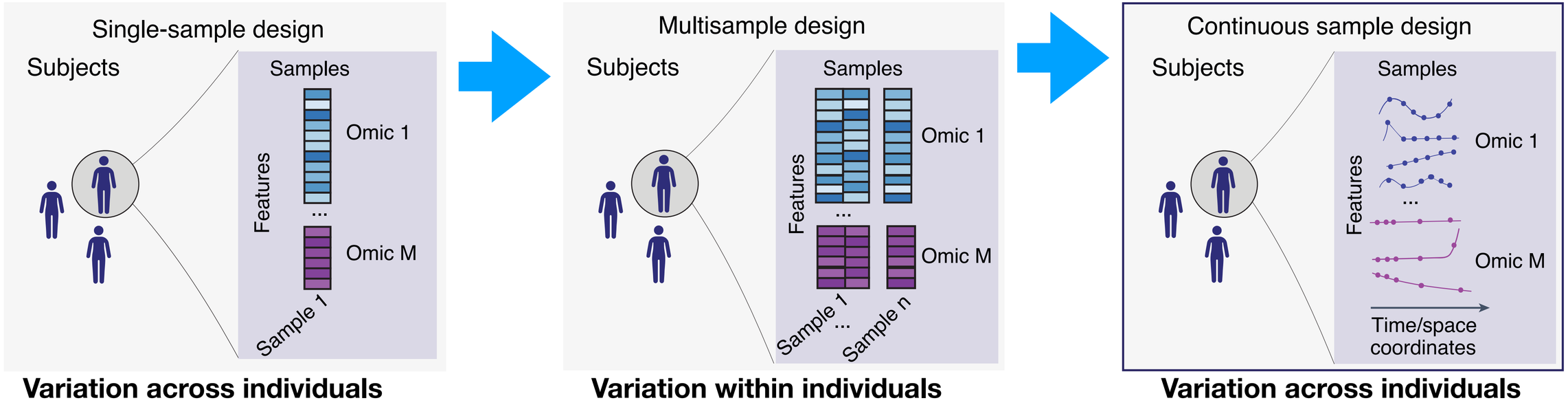
A traditional design for Omics studies: one sample per individual





Velten & Stegle, Nature Methods (2023)

A new experimental design: many samples per individual



Velten & Stegle, Nature Methods (2023)

Method

TRANSPARENT PROCESS

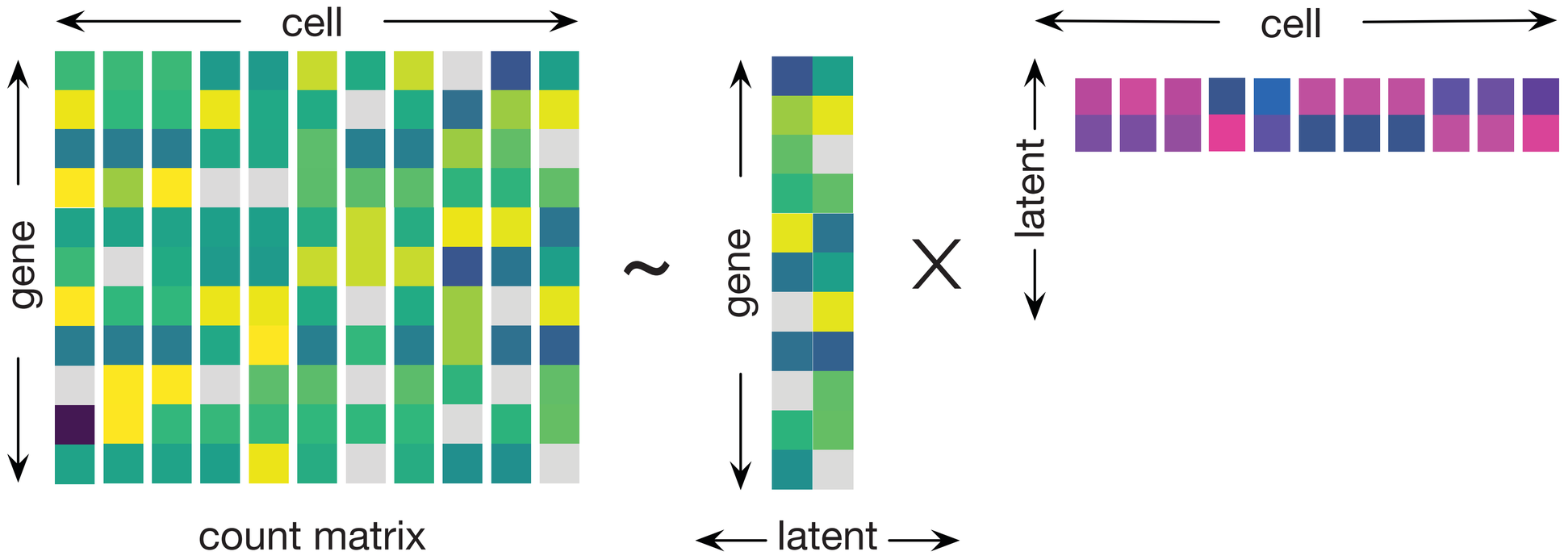
OPEN ACCESS

molecular systems biology

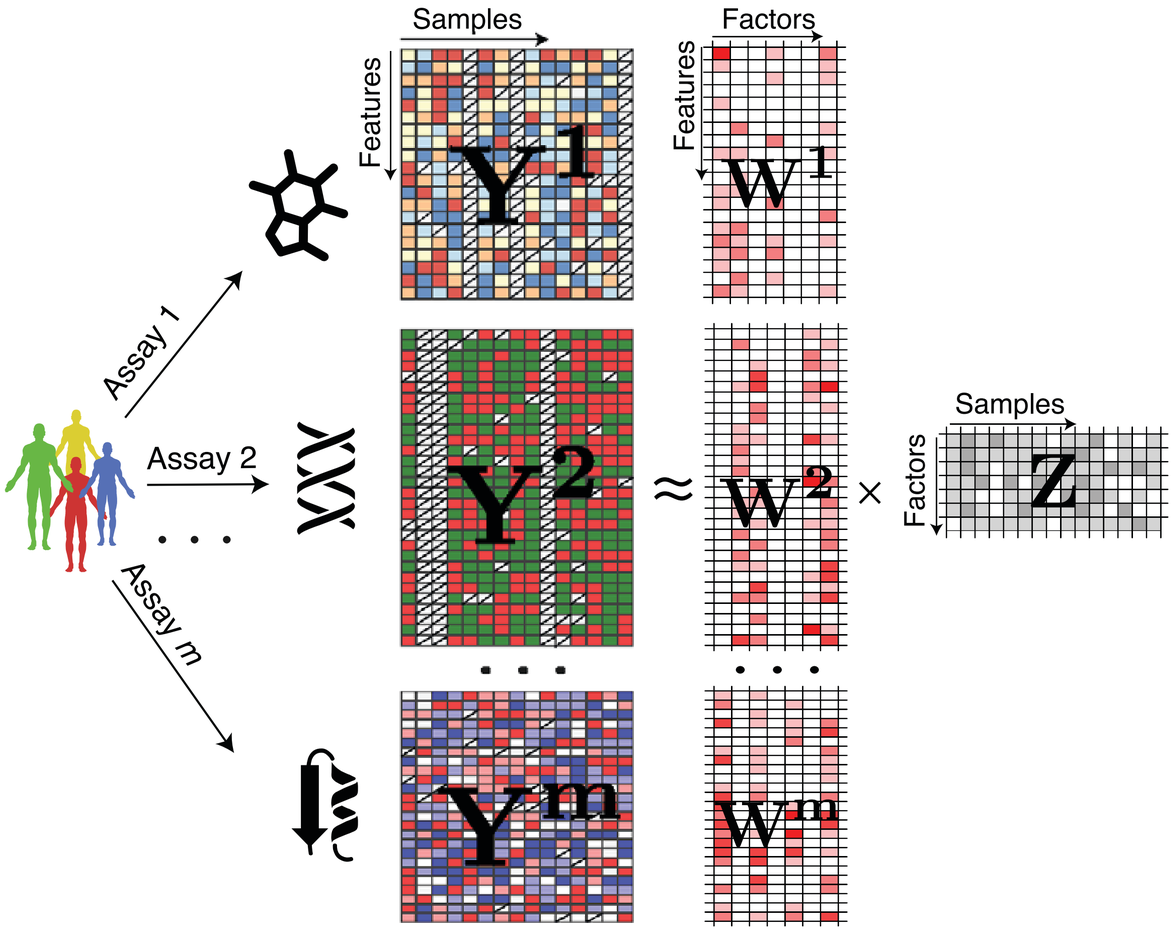
Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets

Ricard Argelaguet[1,†], Britta Velten[2,†], Damien Arnol[1], Sascha Dietrich[3], Thorsten Zenz[3,4,5], John C Marioni[1,6,7], Florian Buettner[1,8,*], Wolfgang Huber[2,**] & Oliver Stegle[1,2,***]

Factorization of a single data matrix



MOFA: Multi-Omics Factor Analysis



For each data modality *m*

$$\mathbf{Y}^{(m)} = ZW^{(m)^\top} + \epsilon^{(m)}$$

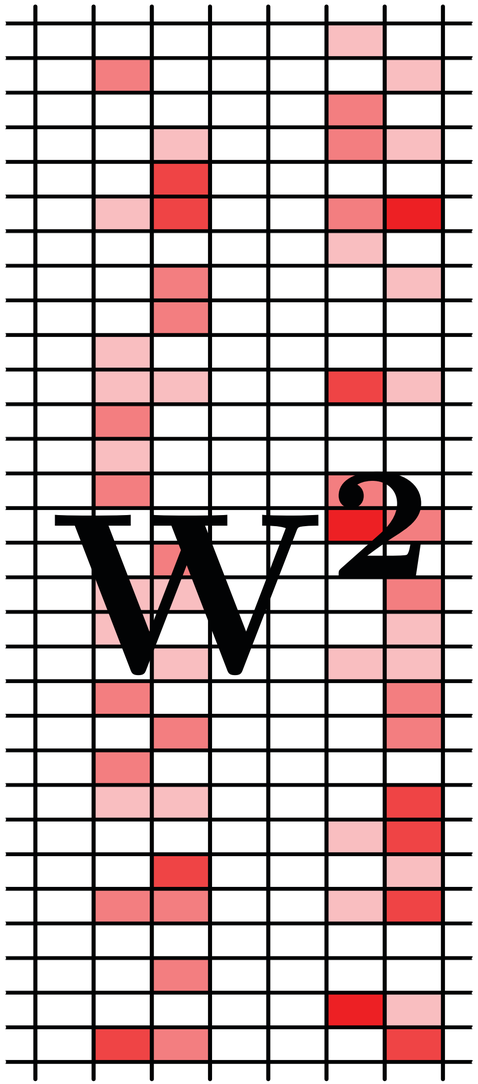
$$W = S\,\hat{W}$$

column-wise

spike (0/1 for each rank)

Argelaguet *et al.* (2018)

MOFA: Multi-Omics Factor Analysis



For each data modality *m*

$$\mathbf{Y}^{(m)} = ZW^{(m)^\top} + \epsilon^{(m)}$$

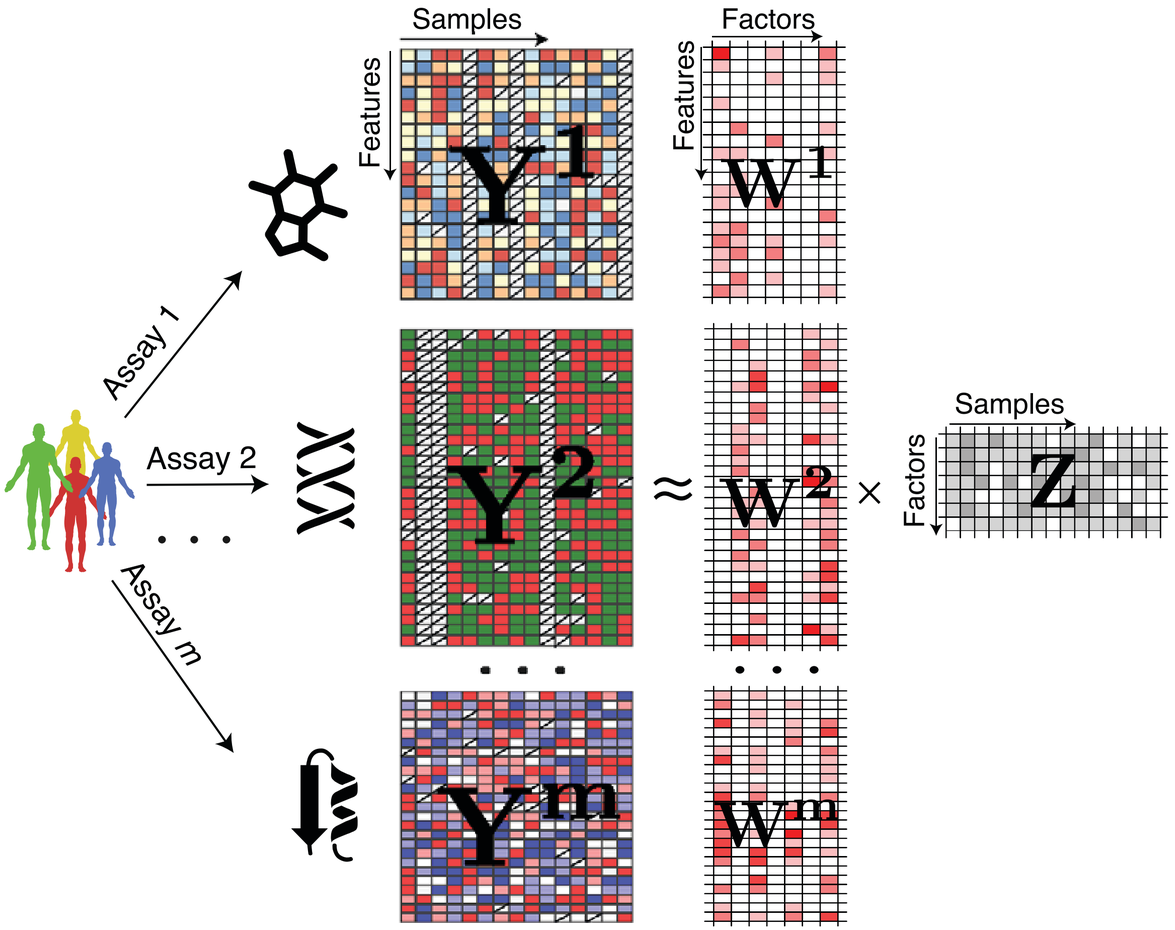
$$W = S\,\hat{W}$$

column-wise

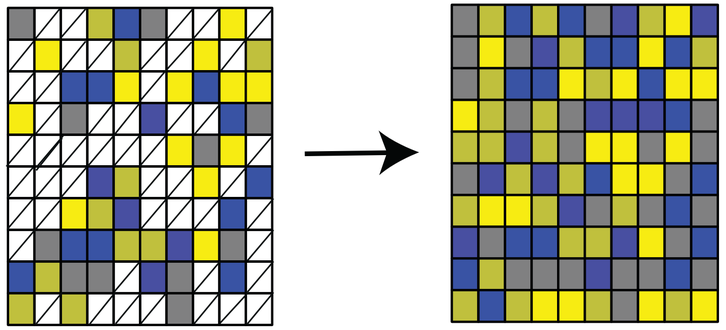
spike (0/1 for each rank)

Argelaguet *et al.* (2018)

When is MOFA useful?

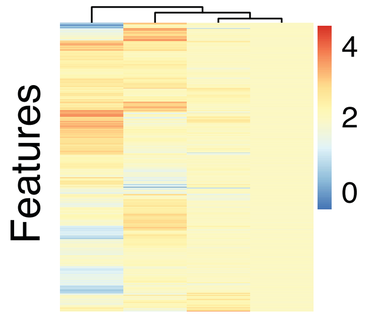


Imputation of missing values

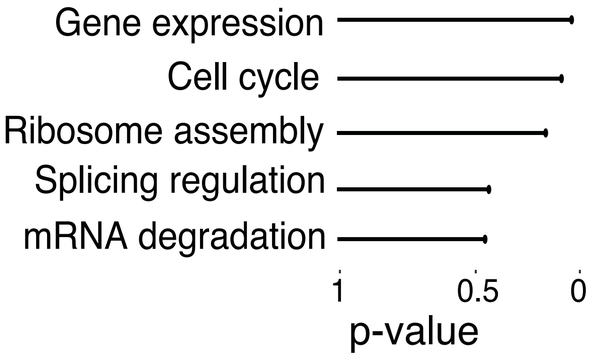


Annotation of factors

Inspection of loadings



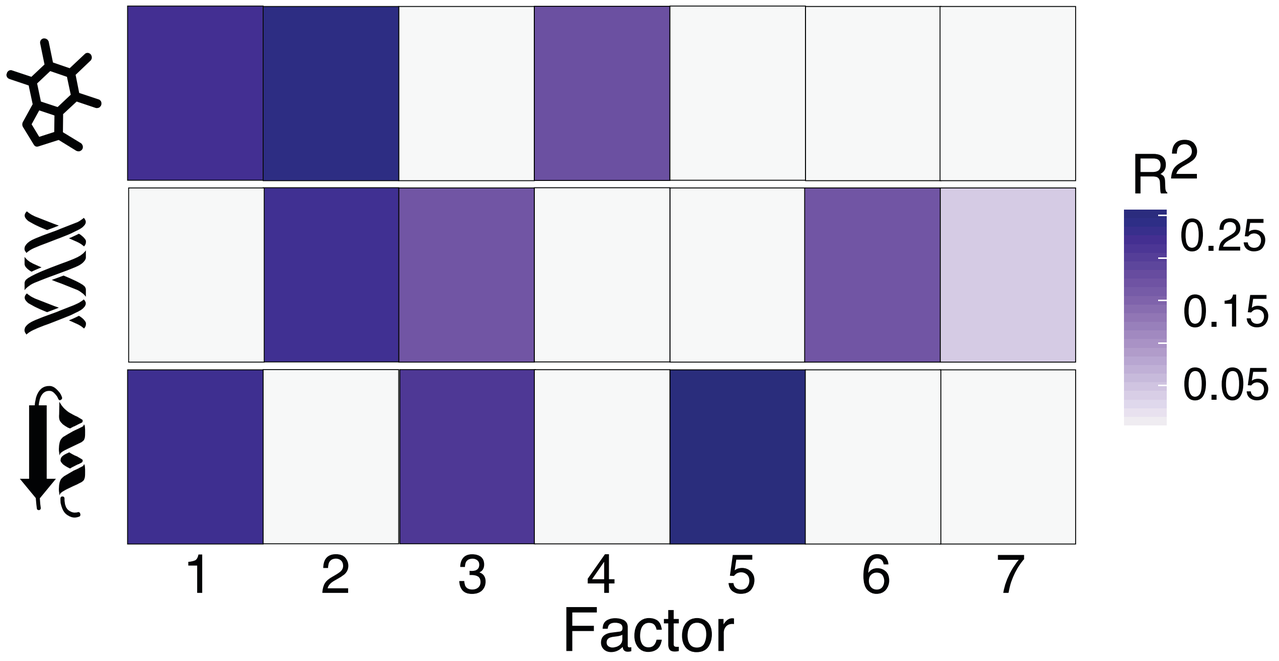
Feature set enrichment analysis



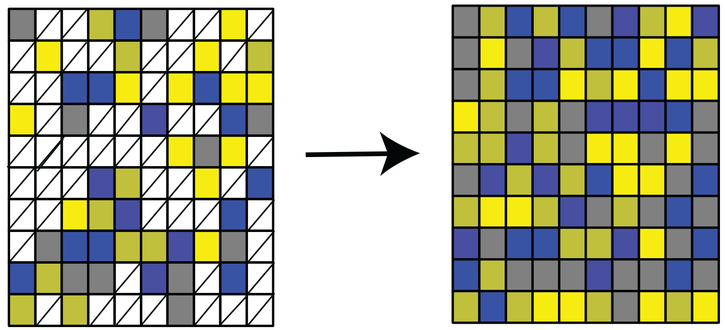
Argelaguet *et al.* (2018)

When is MOFA useful?

Variance decomposition by factor

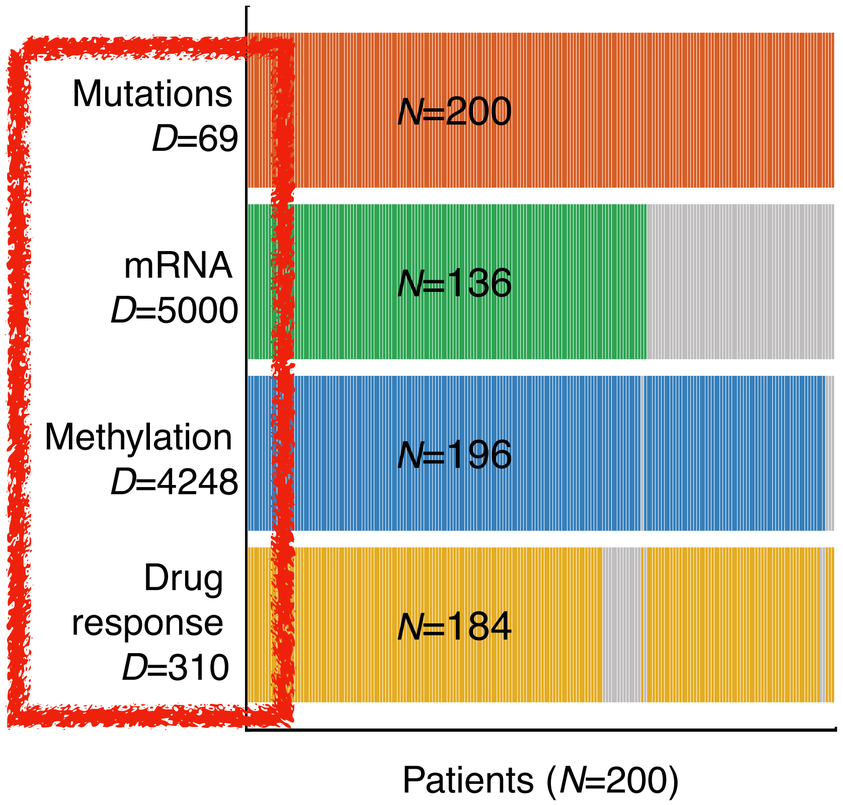


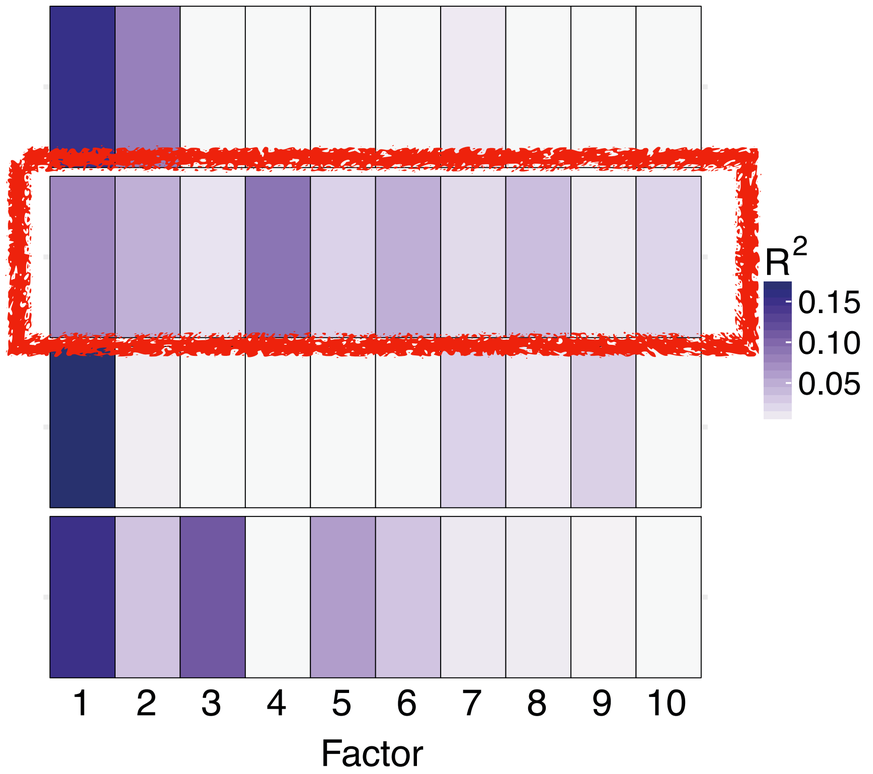
Imputation of missing values



Argelaguet *et al.* (2018)

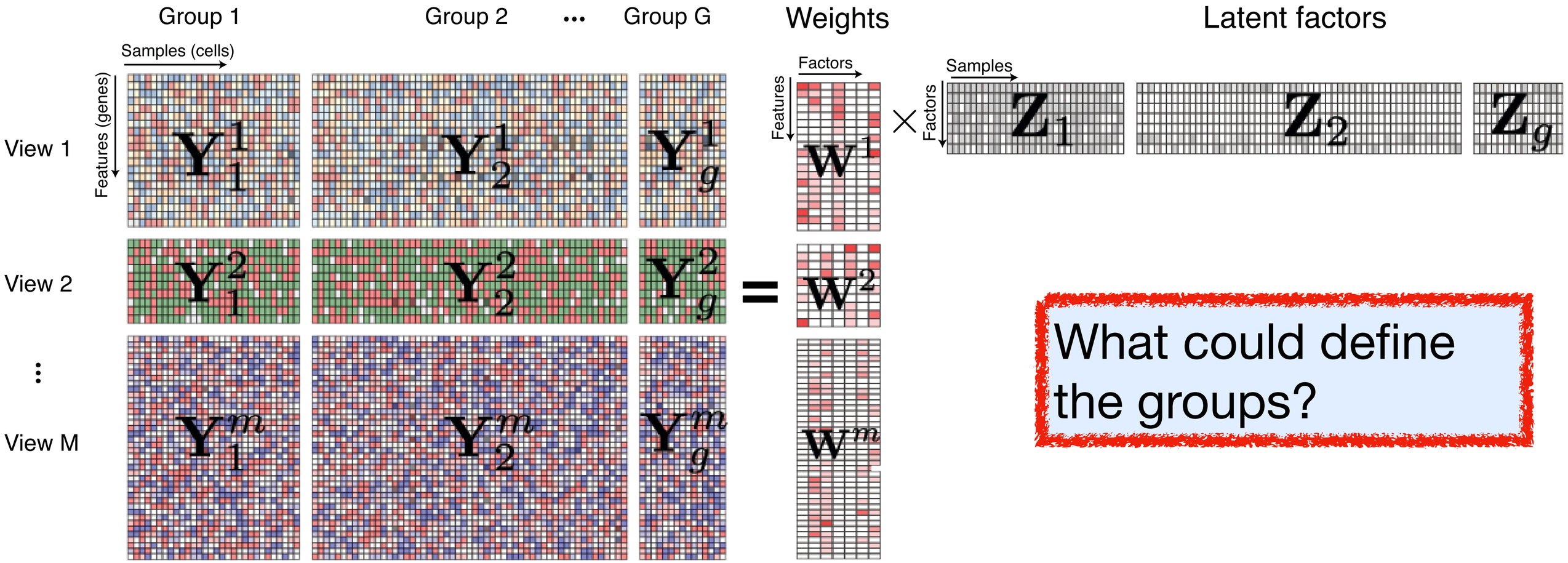
When is MOFA *not really* useful?





Argelaguet *et al.* (2018)

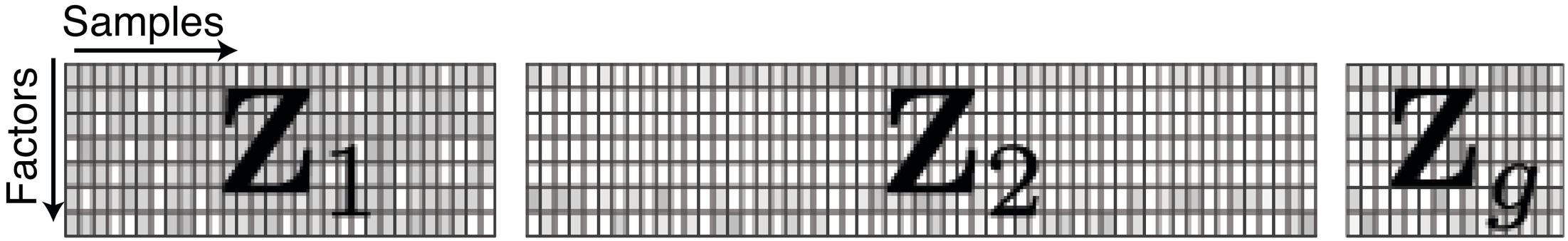
MOFA+ is multi-group and modality modelling



What could define the groups?

Argelaguet *et al.* (2020)

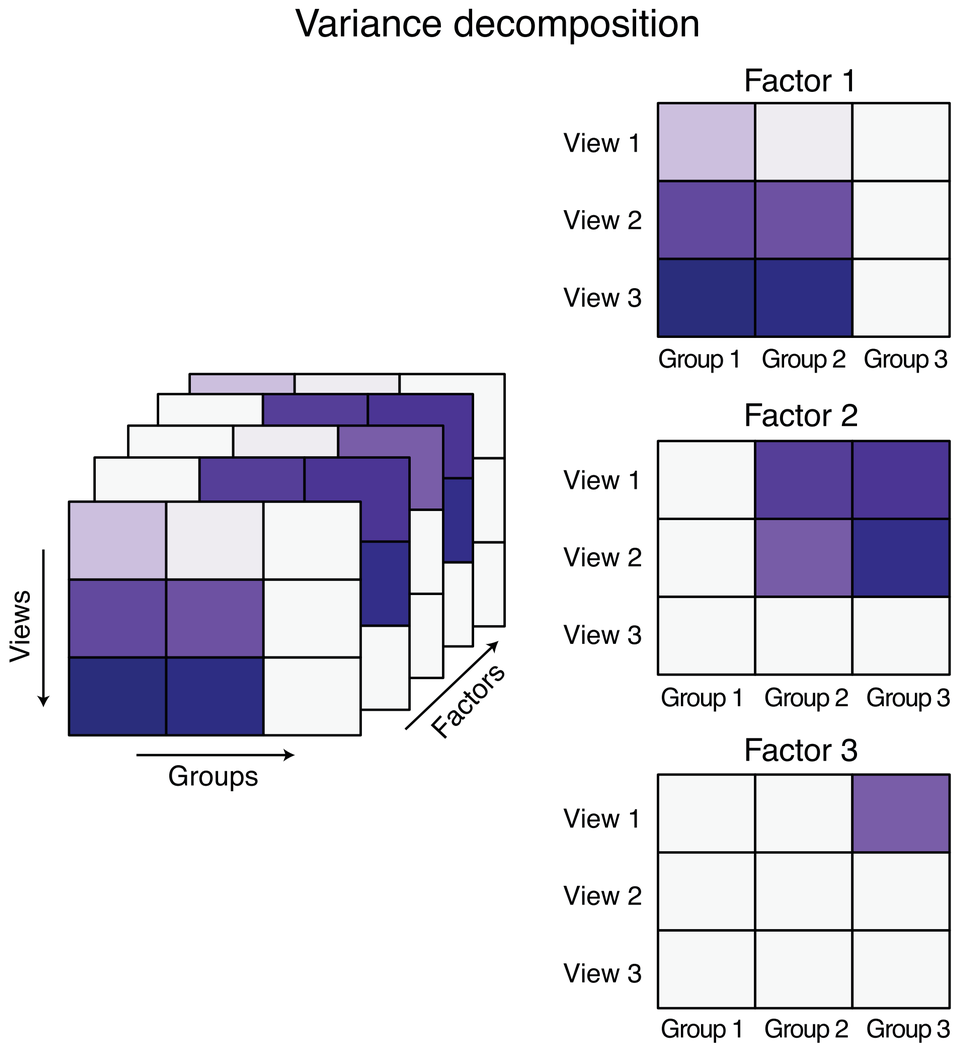
MOFA+ can model heterogeneity across groups



$$\mathbf{Y}^{(g,m)} = Z^{(g)} W^{(m)^\top} + \epsilon^{(g,m)}$$

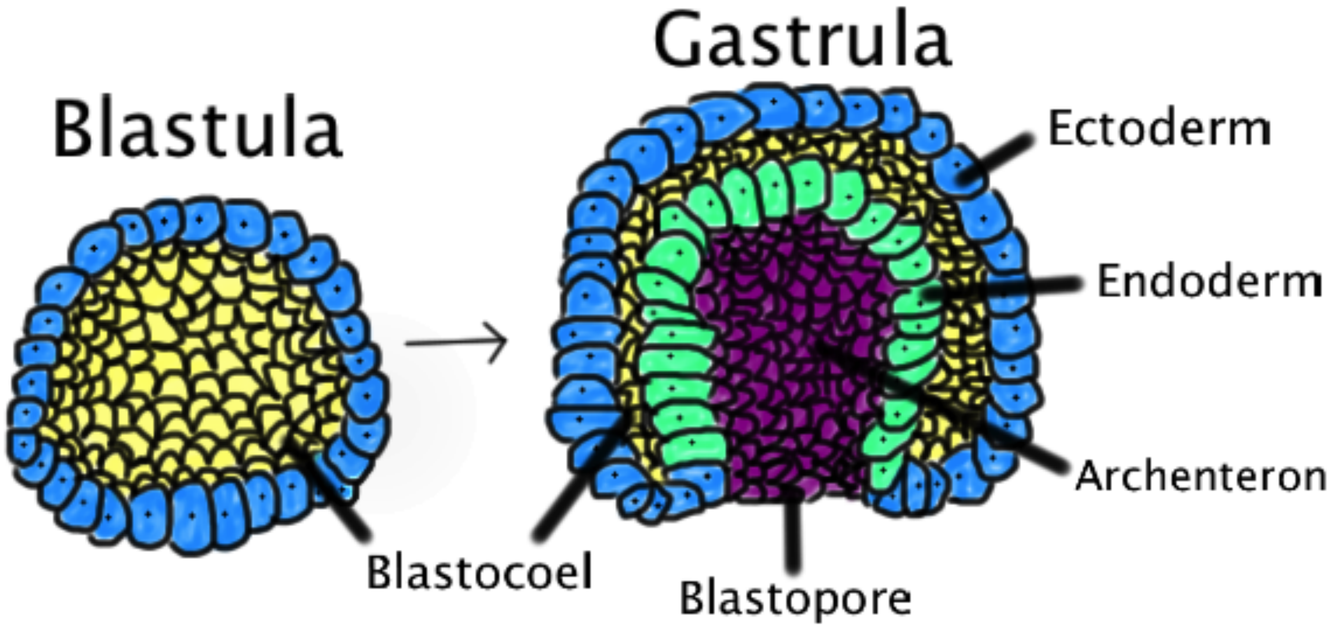
Argelaguet *et al.* (2020)

MOFA+ can factorize group and modality variance

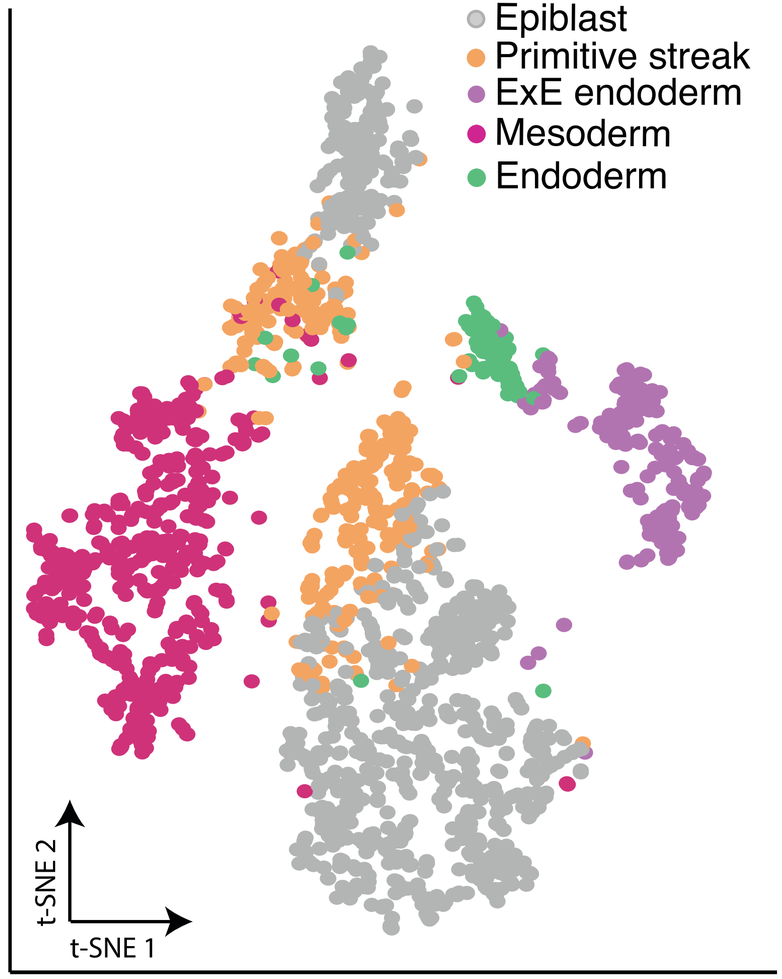


Argelaguet *et al.* (2020)

An example of mouse gastrulation

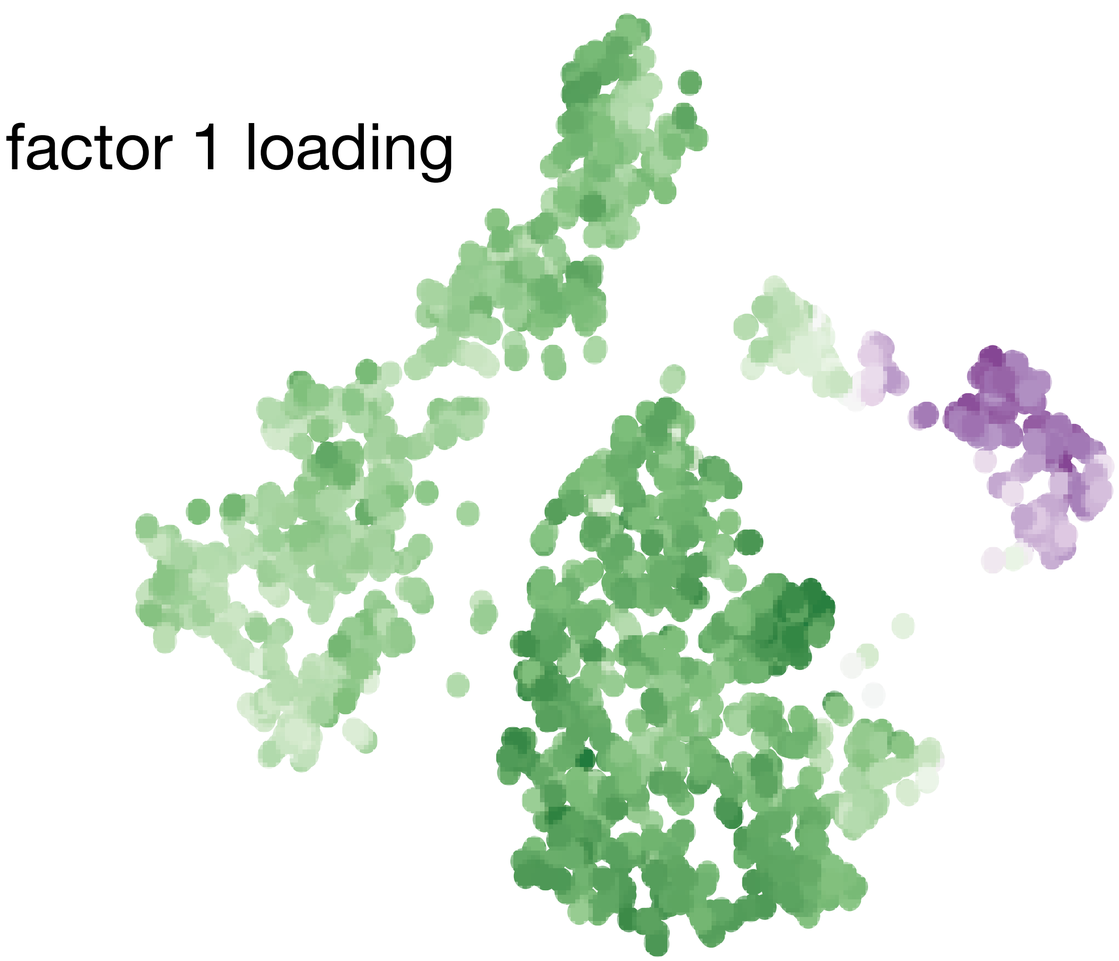


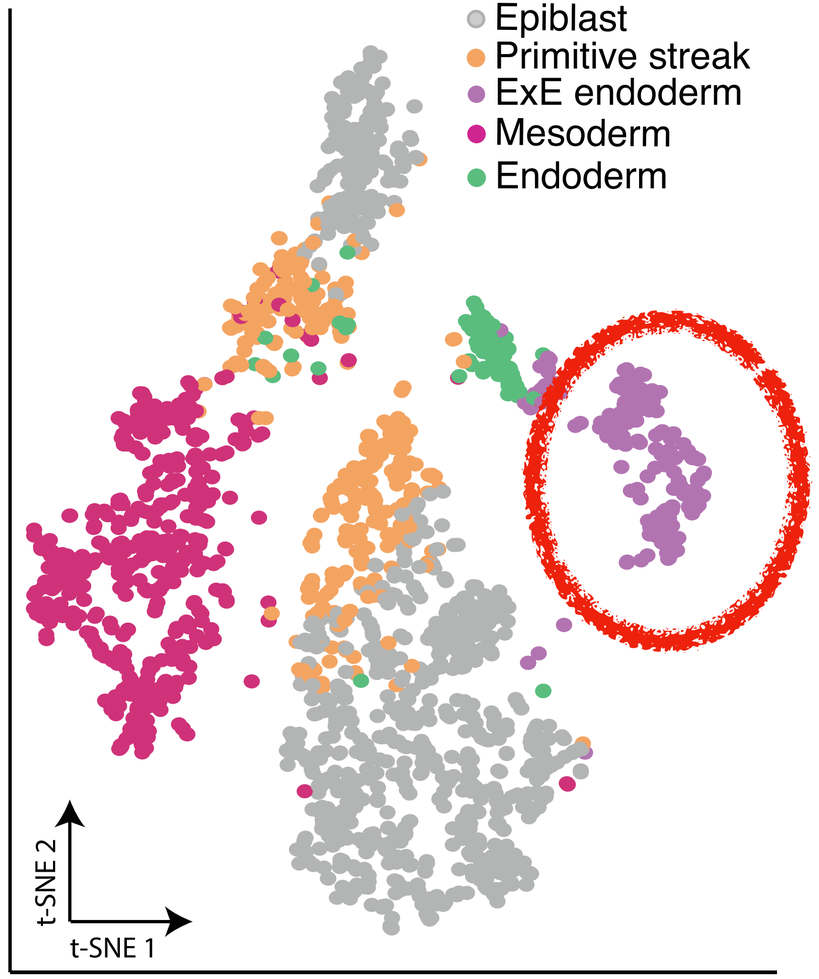
https://en.wikipedia.org/wiki/Gastrulation



Argelaguet *et al.* (2020)

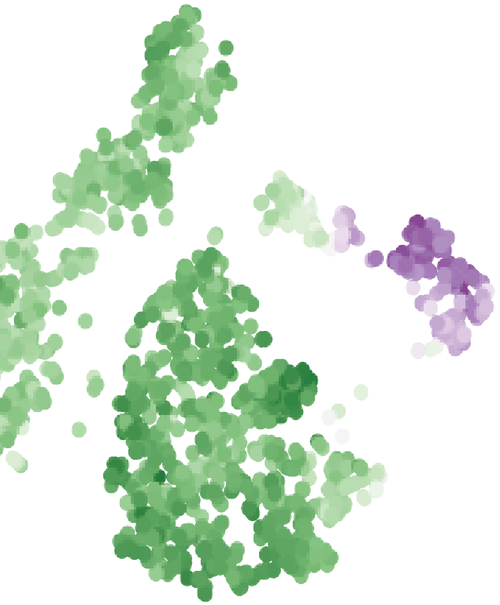
Factor 1 ≈ ExE endoderm



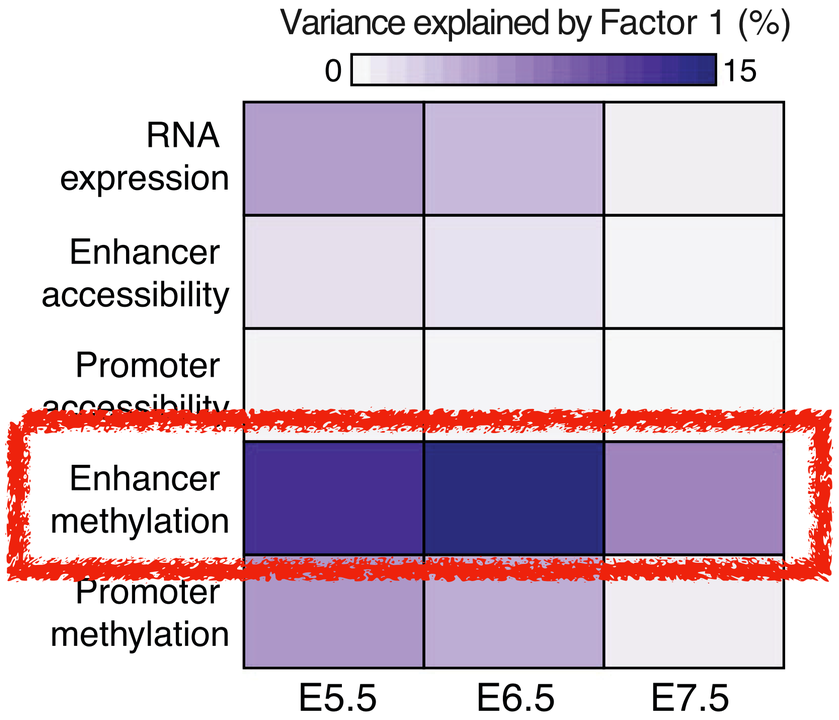


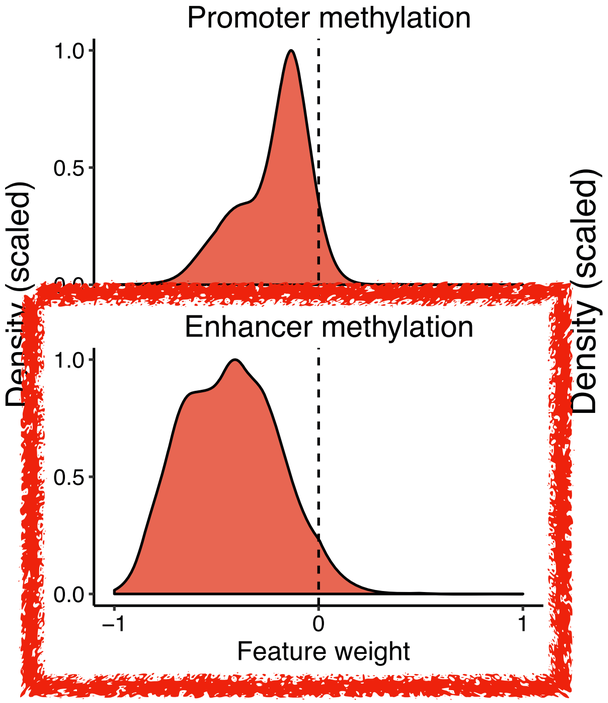
Argelaguet *et al.* (2020)

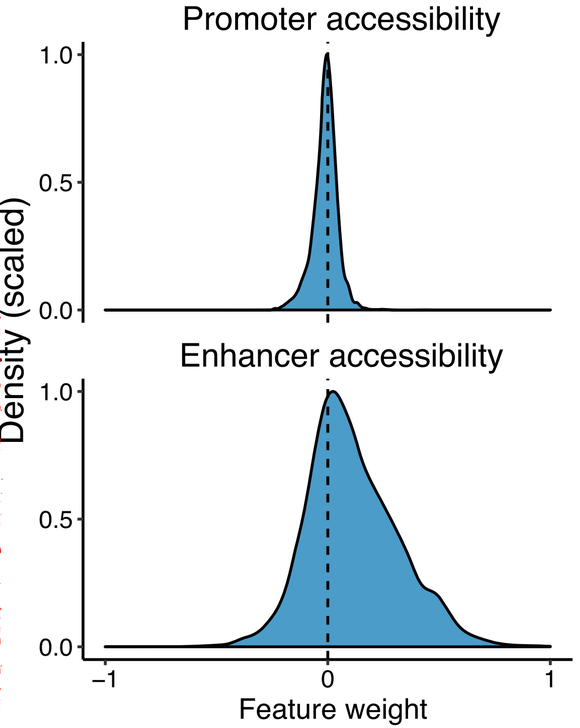
Factor 1 ≈ ExE endoderm + Enh methylation



factor 1 loading

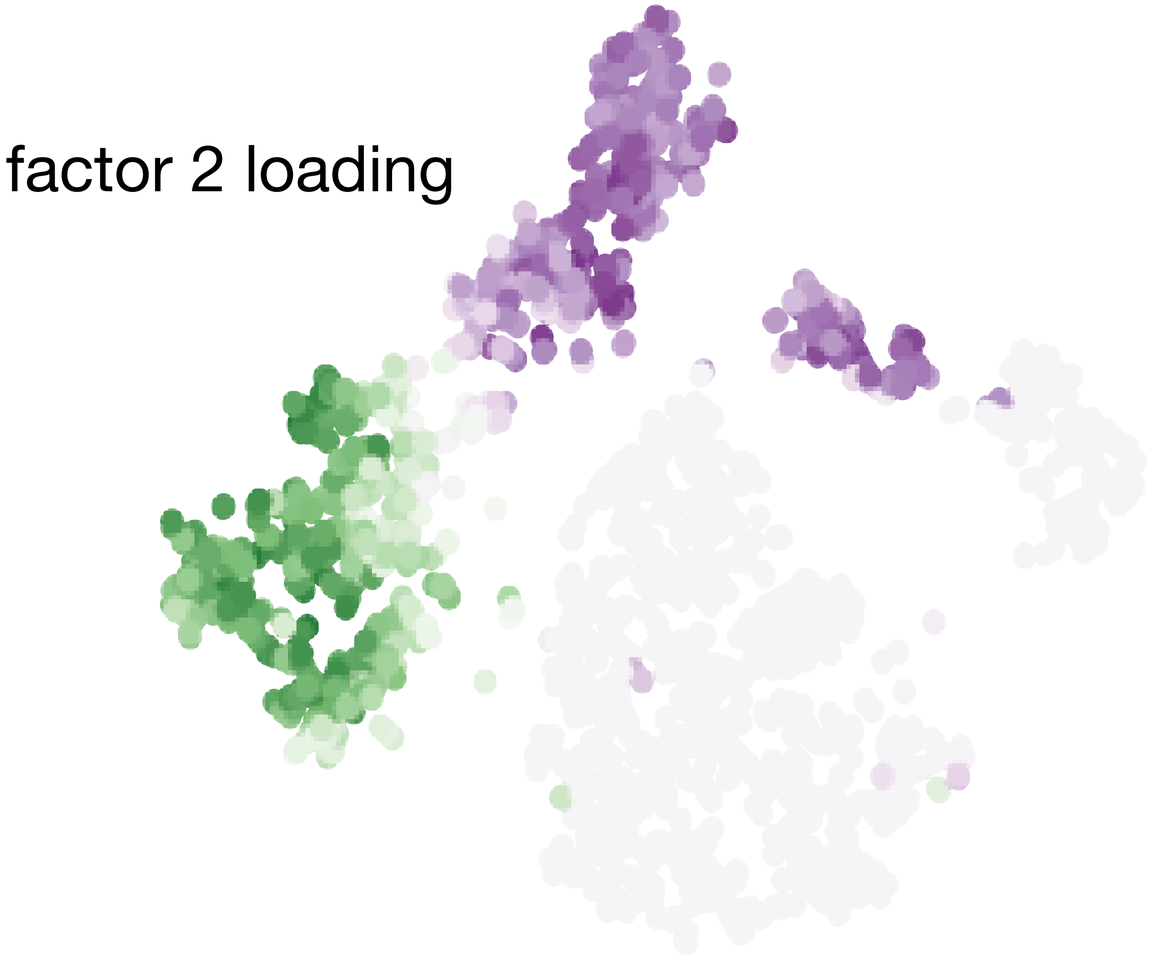


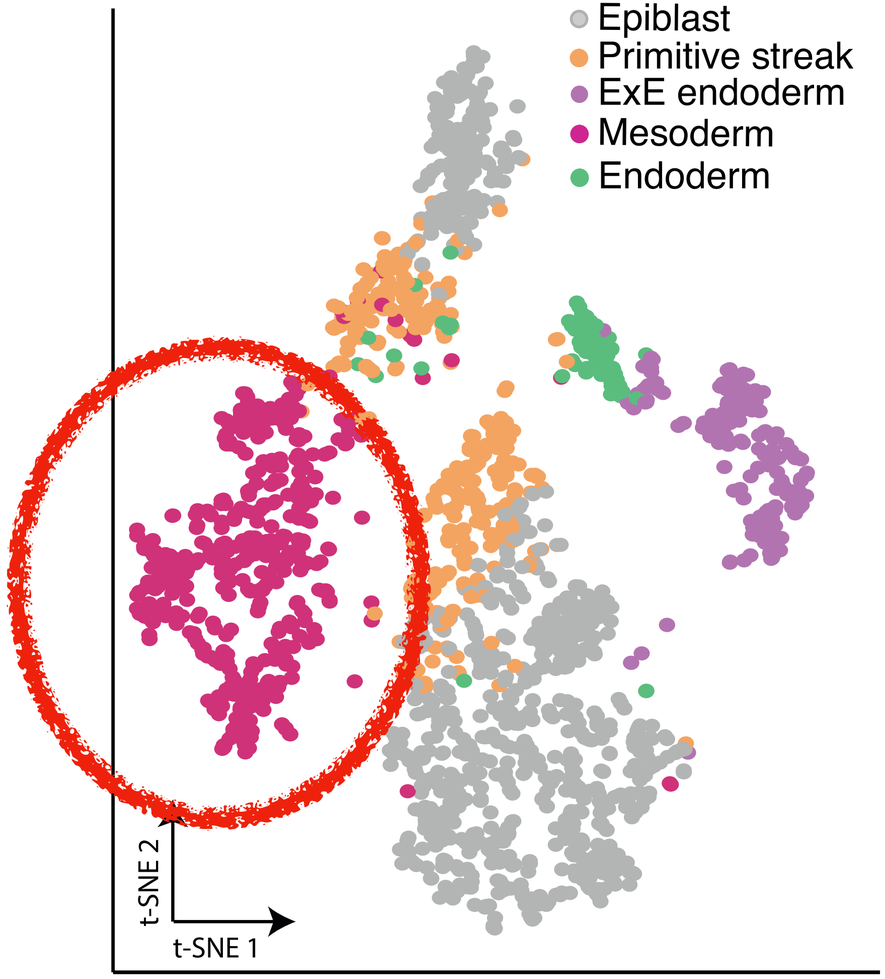




Argelaguet *et al.* (2020)

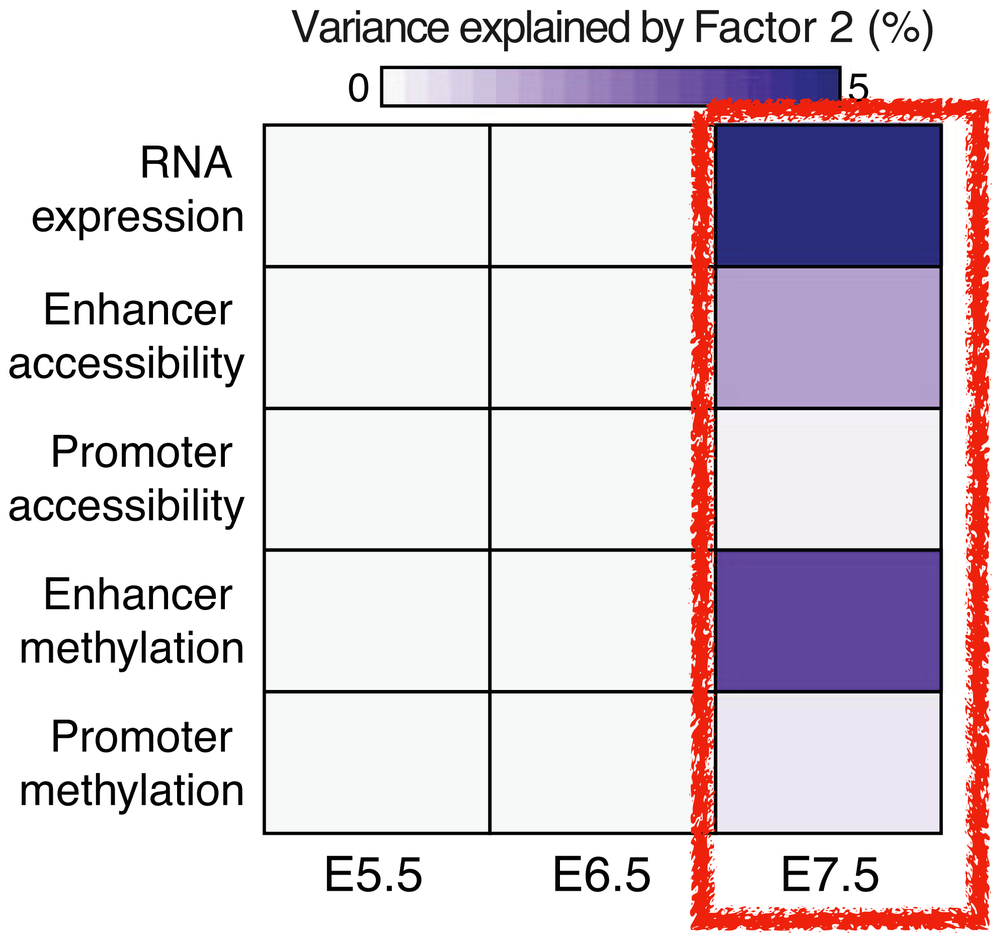
Factor 2 ≈ Mesoderm

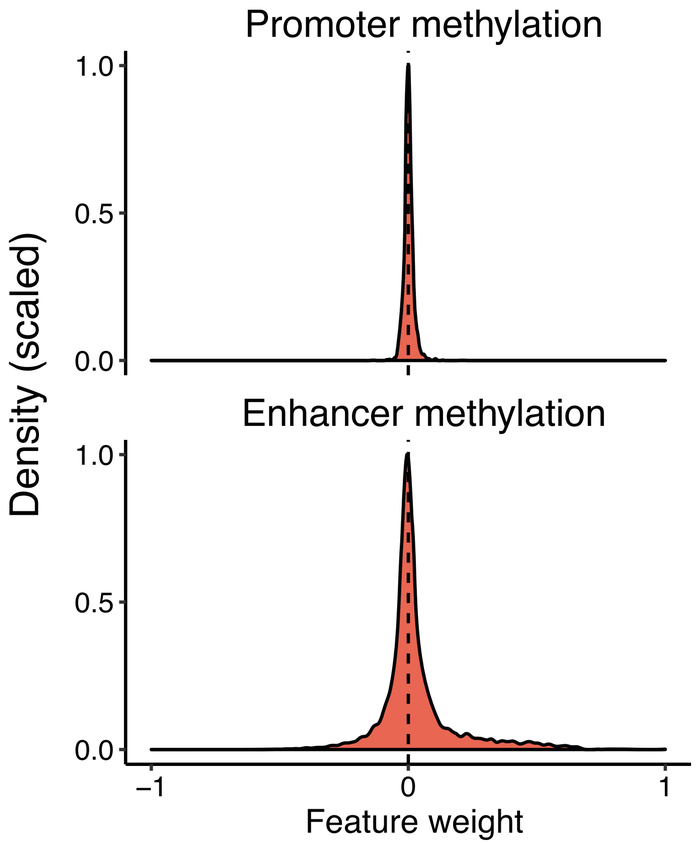


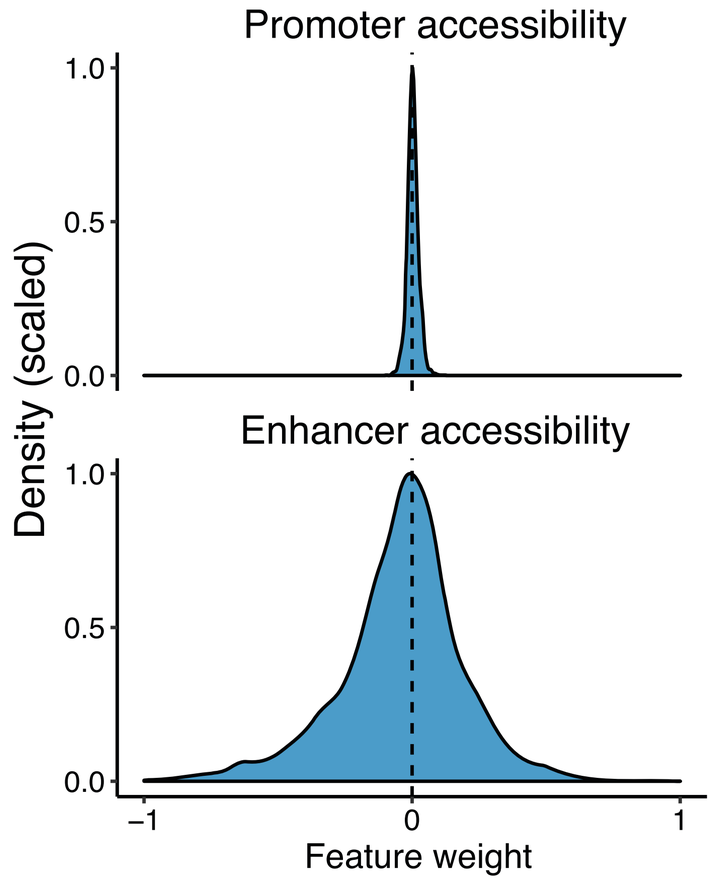


Argelaguet *et al.* (2020)

Factor 2 ≈ mesoderm @ late stage

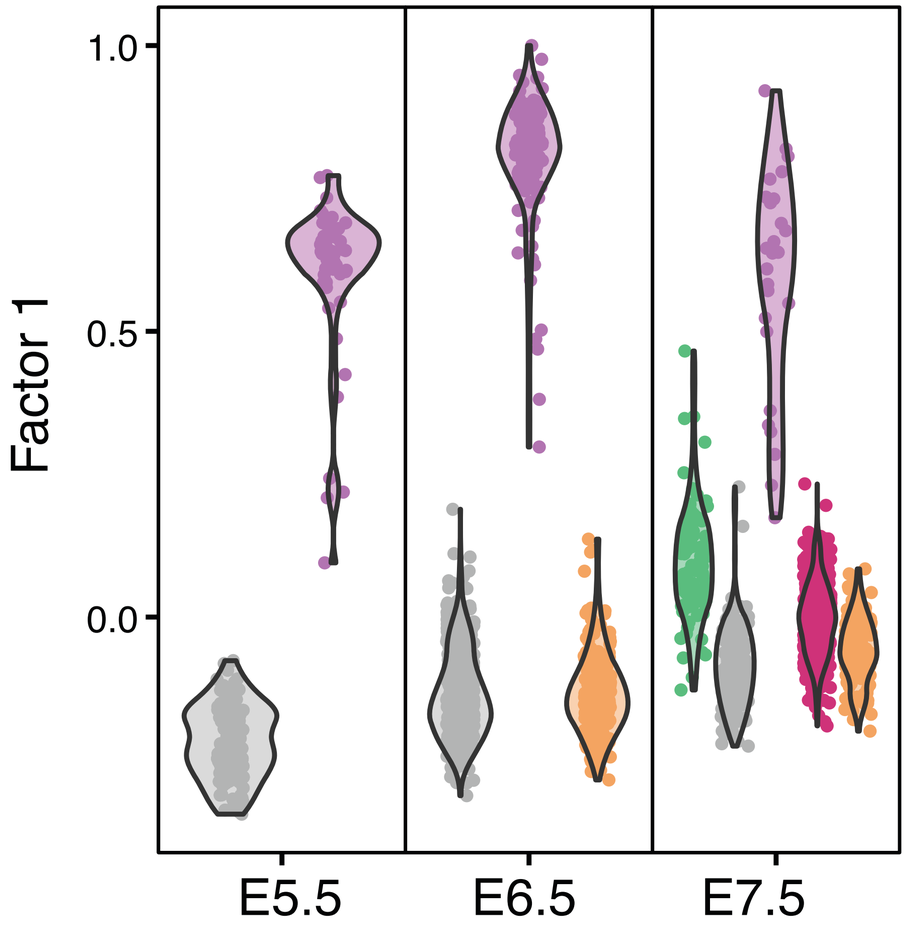


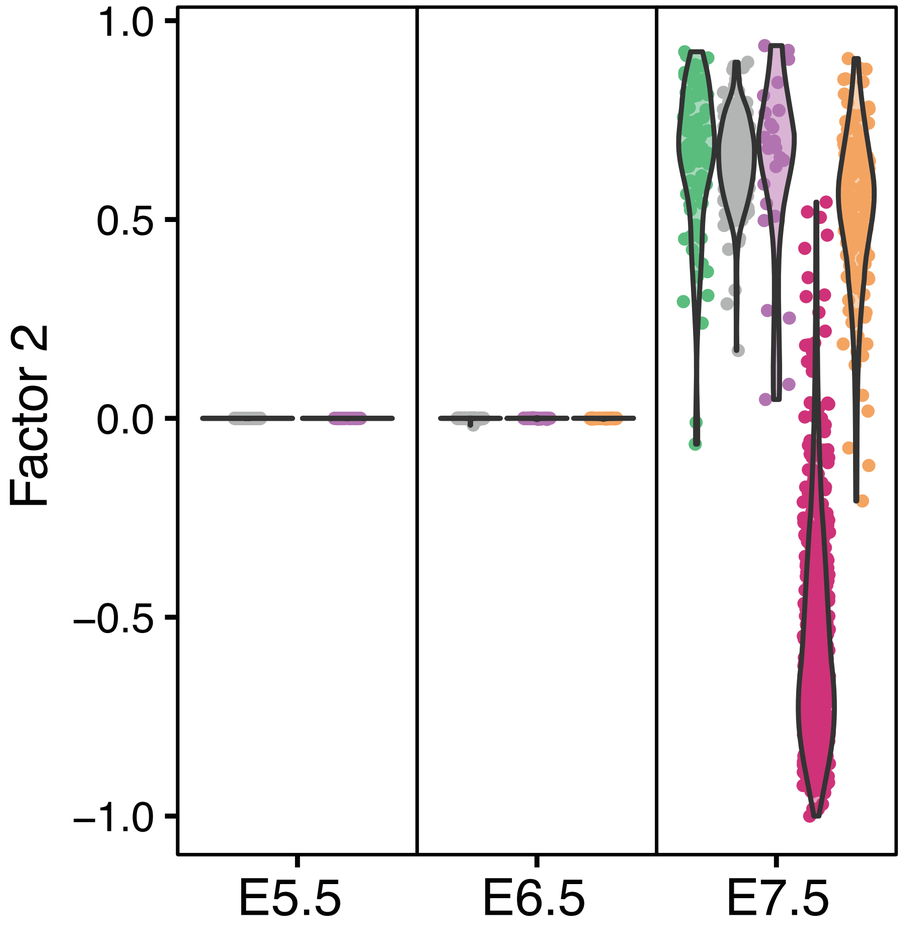




Argelaguet *et al.* (2020)

A bit ambiguous interpretations

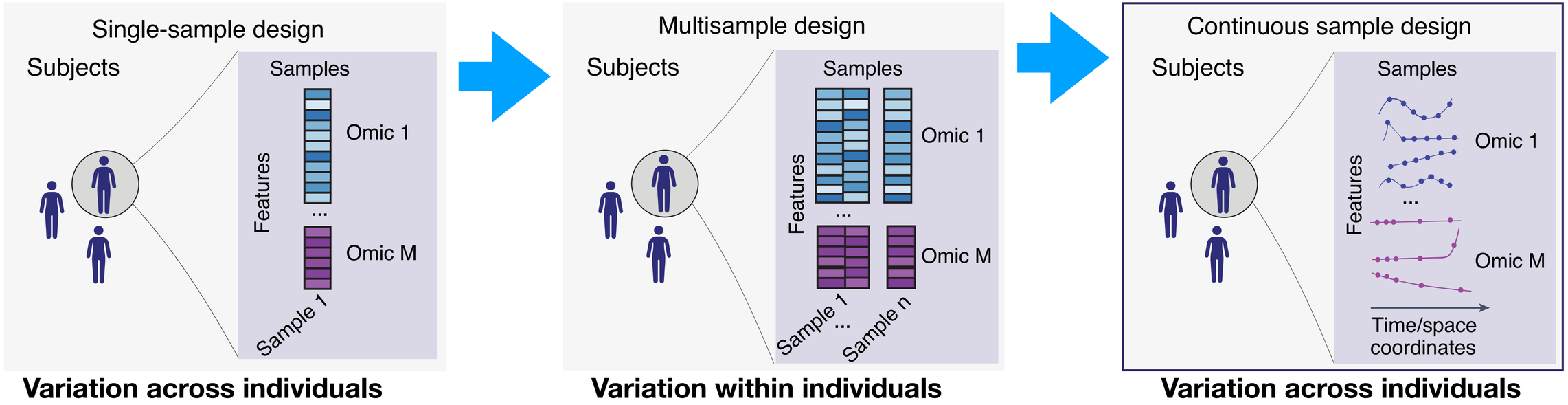




Argelaguet *et al.* (2020)

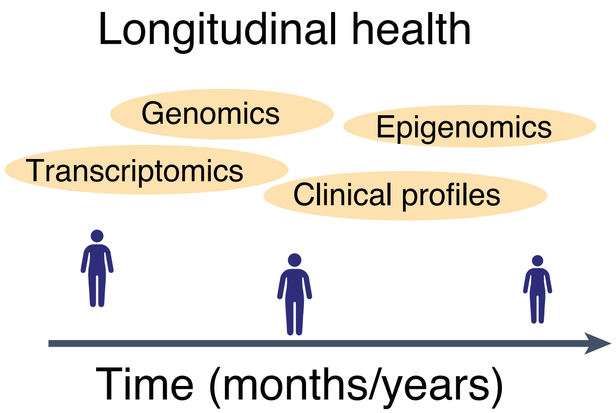
https://biofam.github.io/MOFA2/

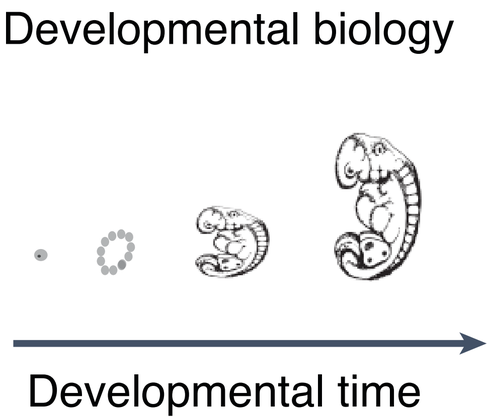
A new experimental design: many samples per individual

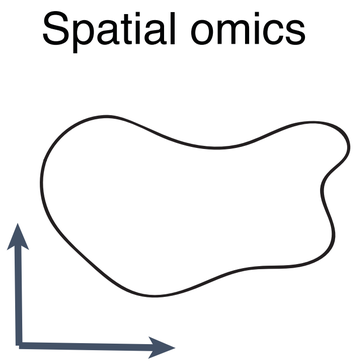


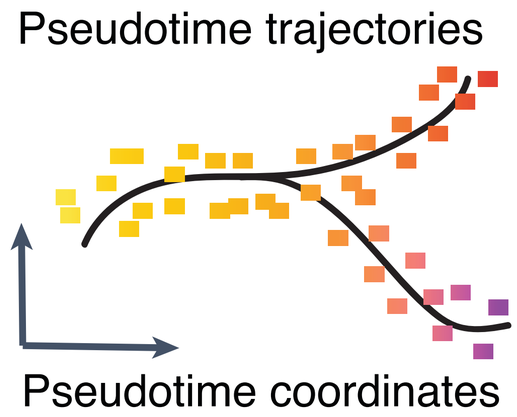
Velten & Stegle, Nature Methods (2023)

Spatiotemporal dependency structures are embedded in biological systems









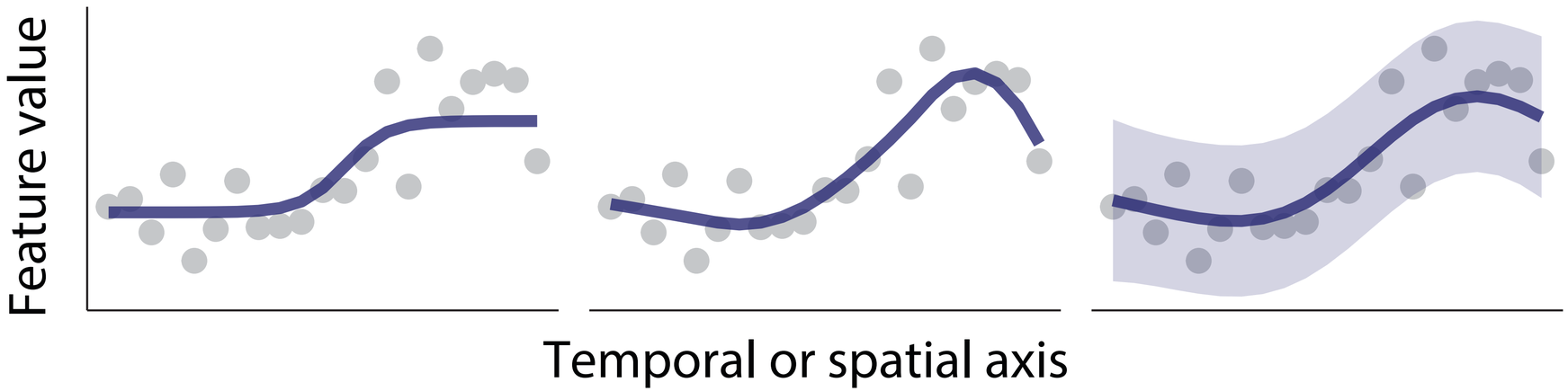
Velten & Stegle, Nature Methods (2023)

How do we model dependencies/relationships?

Regression-based modeling

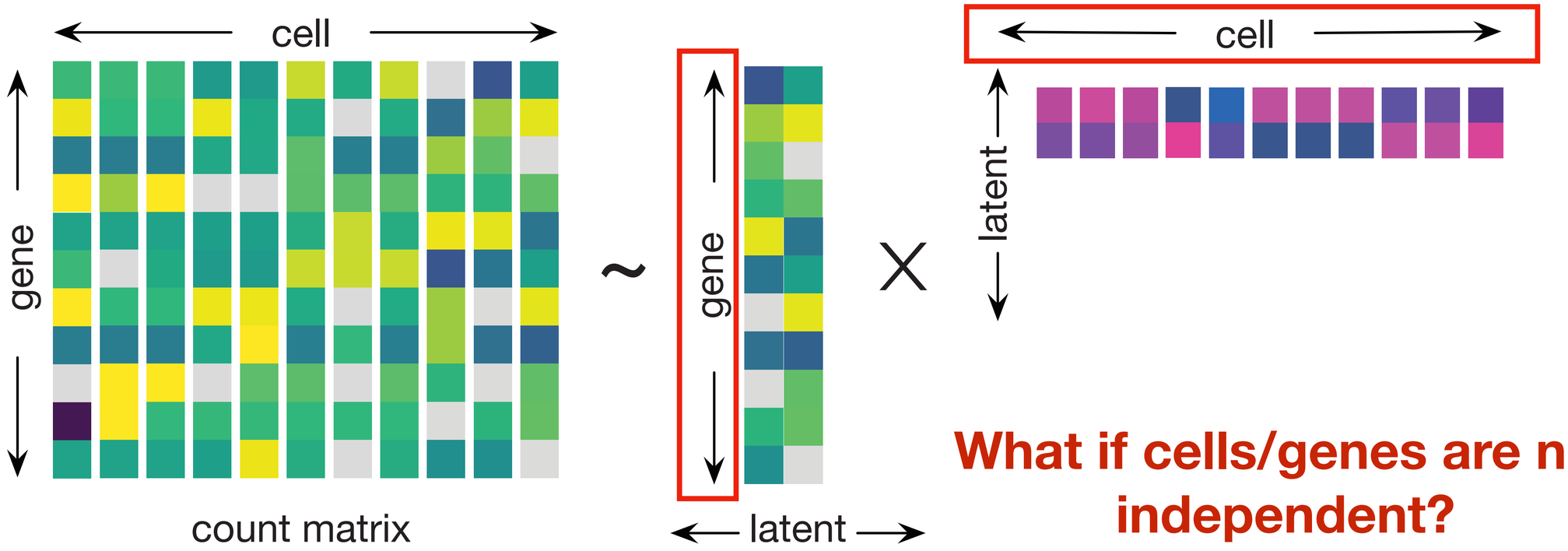
$z = f(t)$ or $z = f(x)$

Parametric regression	Splines	GP
e.g., $f(t) = (1 + e^{-t/a + b})^{-1}$	$f(t) = p_n(t),\ t \in [t_{n-1}, t_n)$	$f \approx GP(\mu, \kappa)$



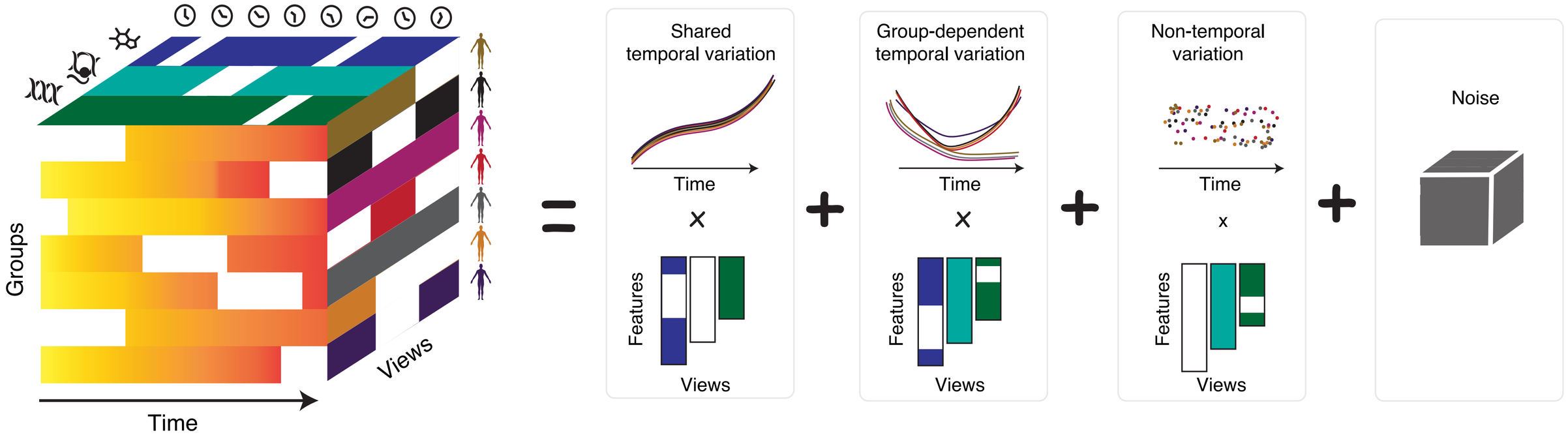
Velten & Stegle, Nature Methods (2023)

Can we incorporate “smooth” temporal models in factorization settings?



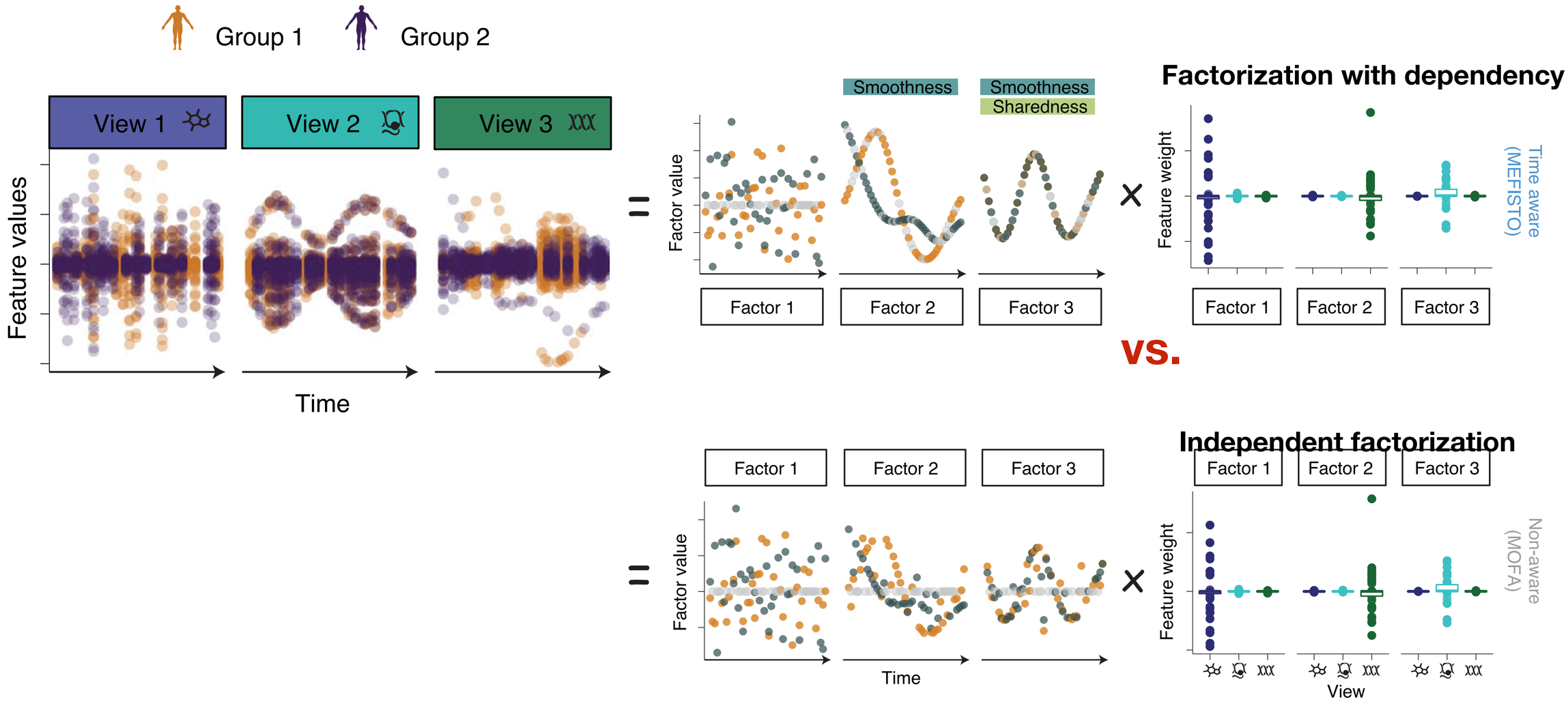
What if cells/genes are not independent?

MEFISTO: multi-view factor analysis + dependency modelling



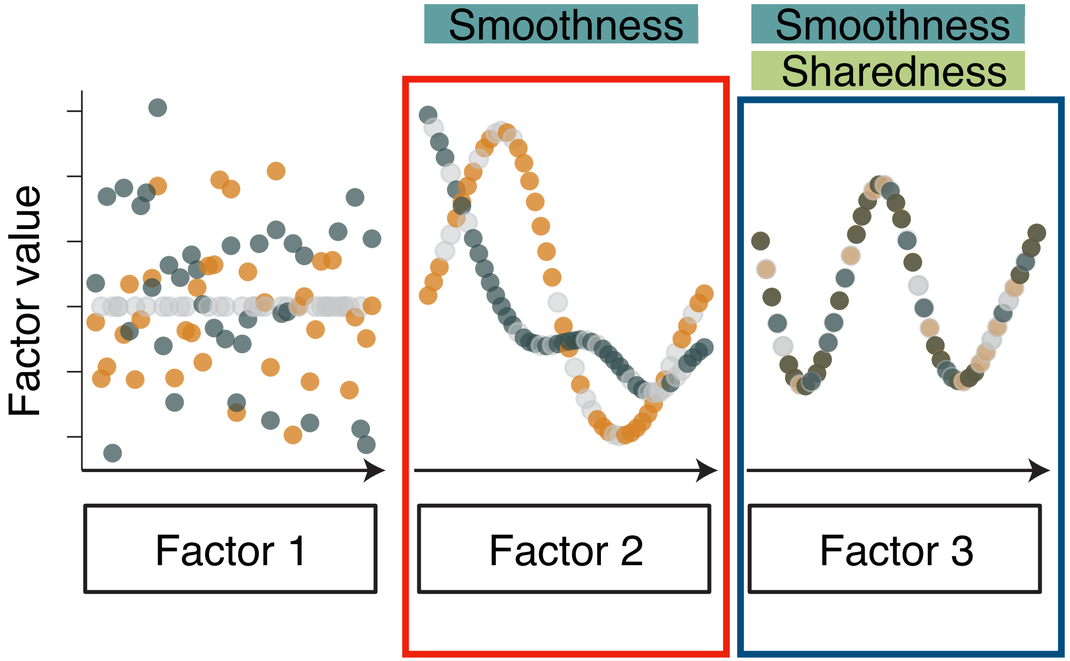
Velten .. Stegle, Nature Methods (2022)

Smoothness assumption helps recover true latent factors and temporal dependency structures

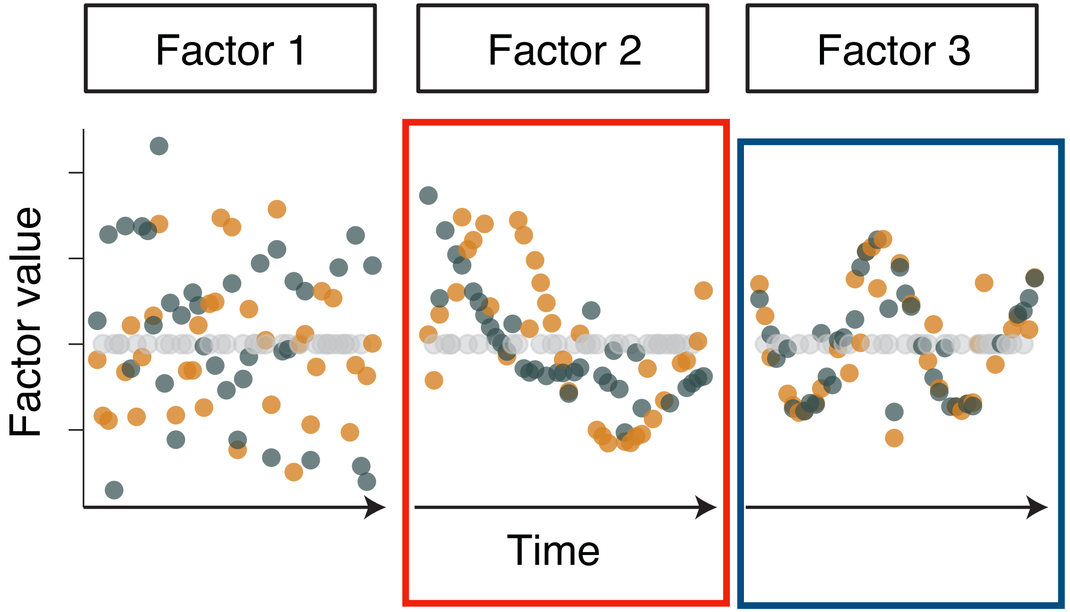


Velten .. Stegle, Nature Methods (2022)

Smoothness assumption helps recover true latent factors and temporal dependency structures



vs.



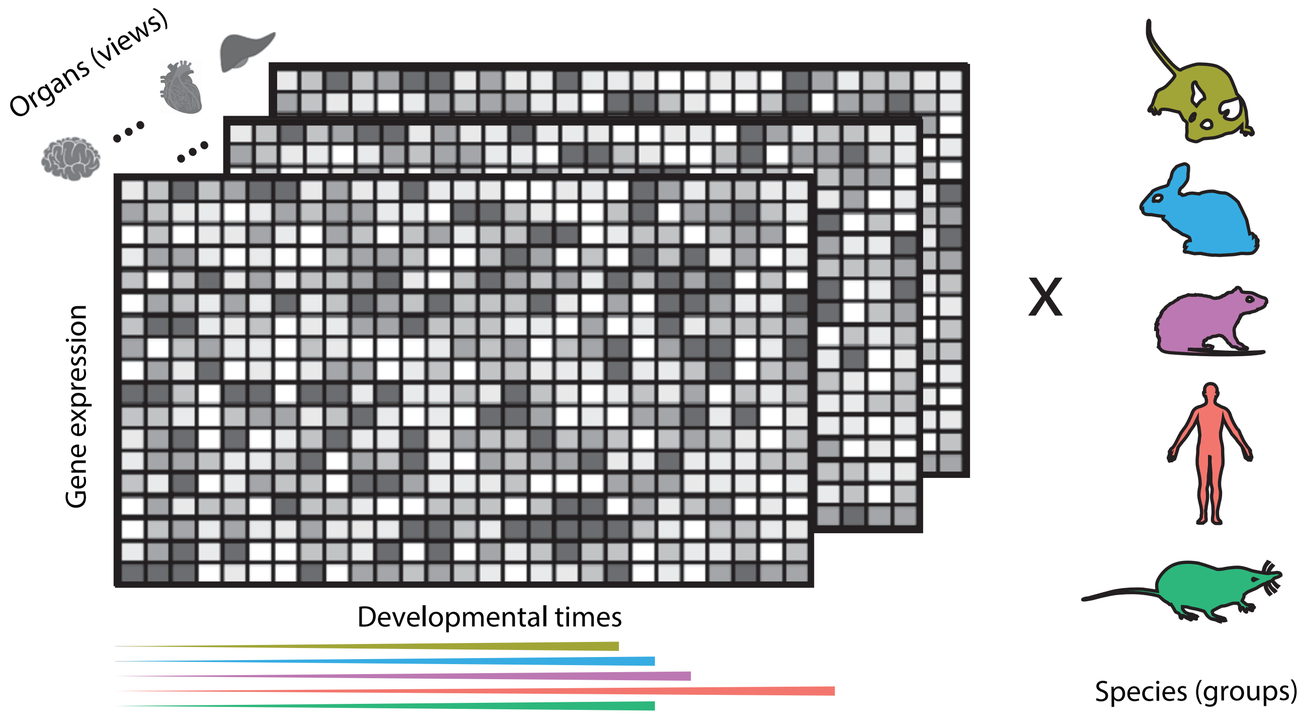
Factorization with dependency

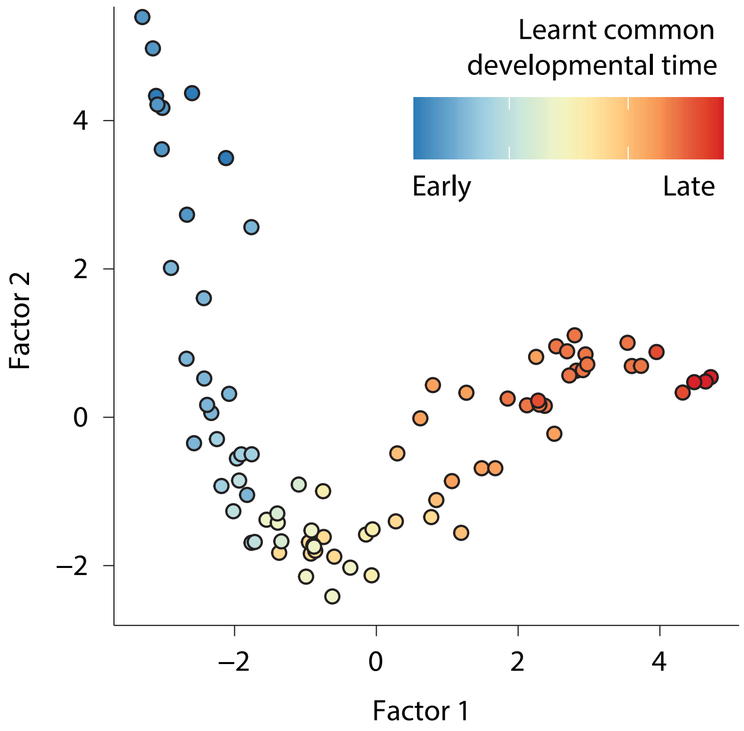
Independent factorization

Note: the grey dots are unobserved

Velten .. Stegle, Nature Methods (2022)

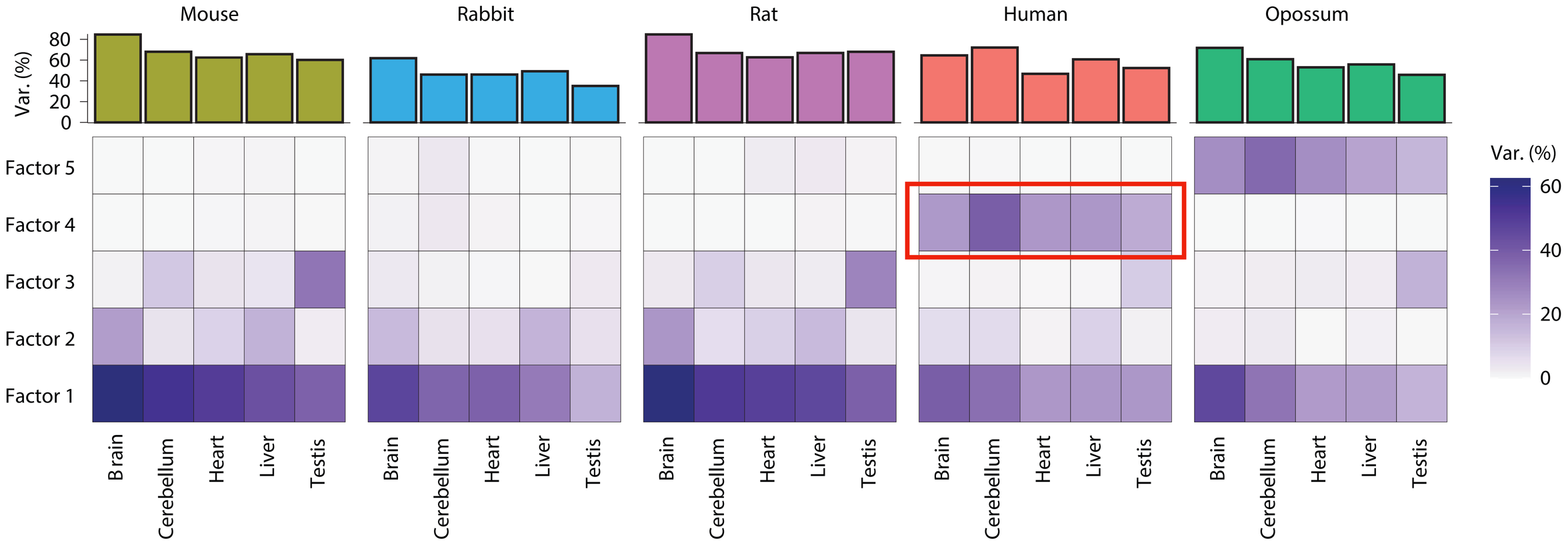
MEFISTO to infer evolutionary gene programs





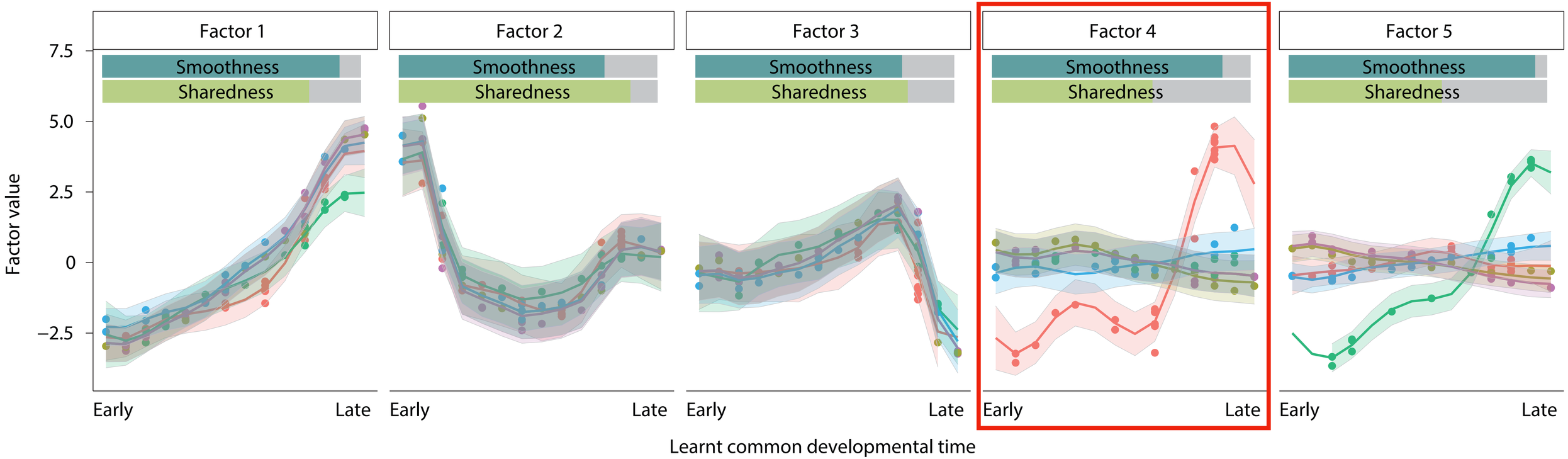
Velten .. Stegle, Nature Methods (2022)

What are human-specific gene factors/programs?



Velten .. Stegle, Nature Methods (2022)

Human-specific programs in the late developmental stages



Velten .. Stegle, Nature Methods (2022)

https://biofam.github.io/MOFA2/MEFISTO

MOFA vs. MEFISTO

$$\mathbf{Y}^{(g,m)} = Z^{(g)} W^{(m)^\top} + \epsilon^{(g,m)}$$

MEFISTO

$$Z_{nk}^{(g)} = f_k(\mathbf{c}_n^{(g)}) + \epsilon_{nk}^{(g)}$$

Gaussian Process (function approx)

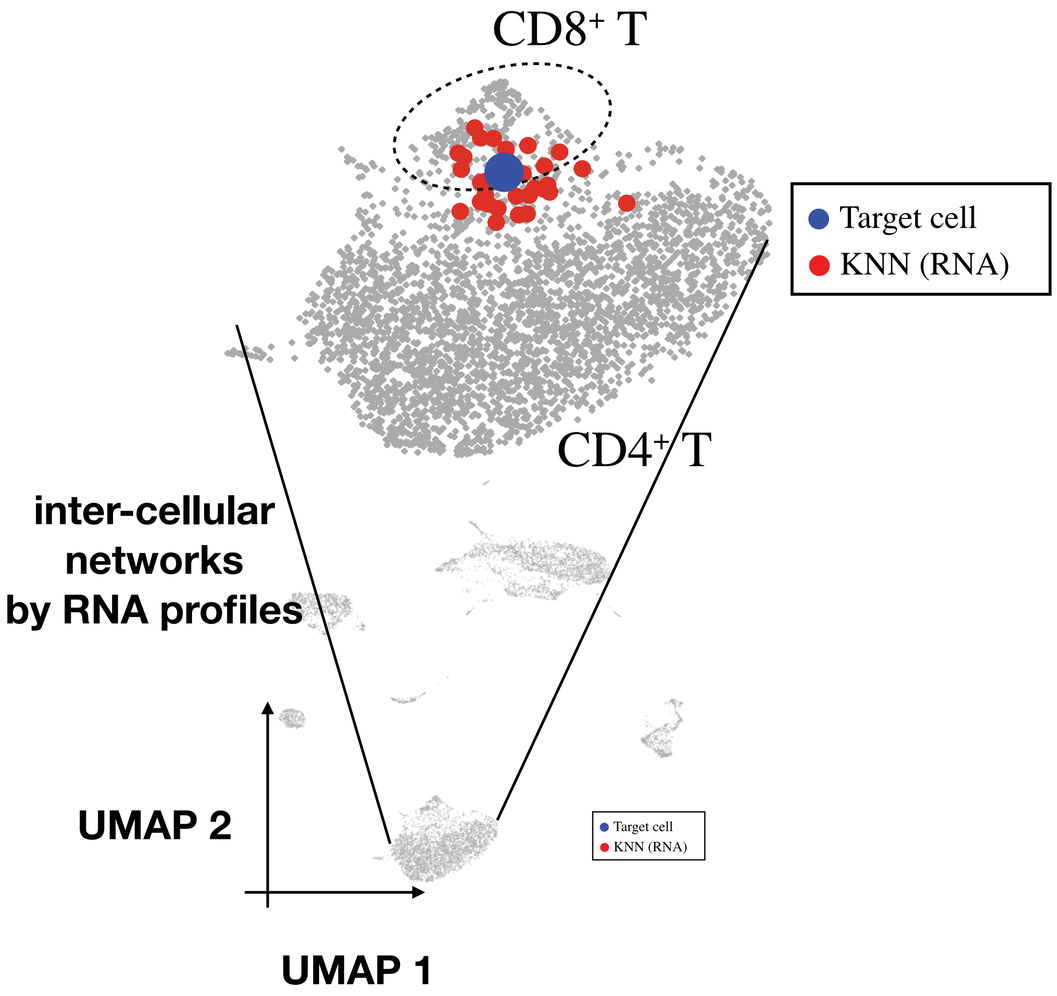
covariate manifesting the underlying dependency

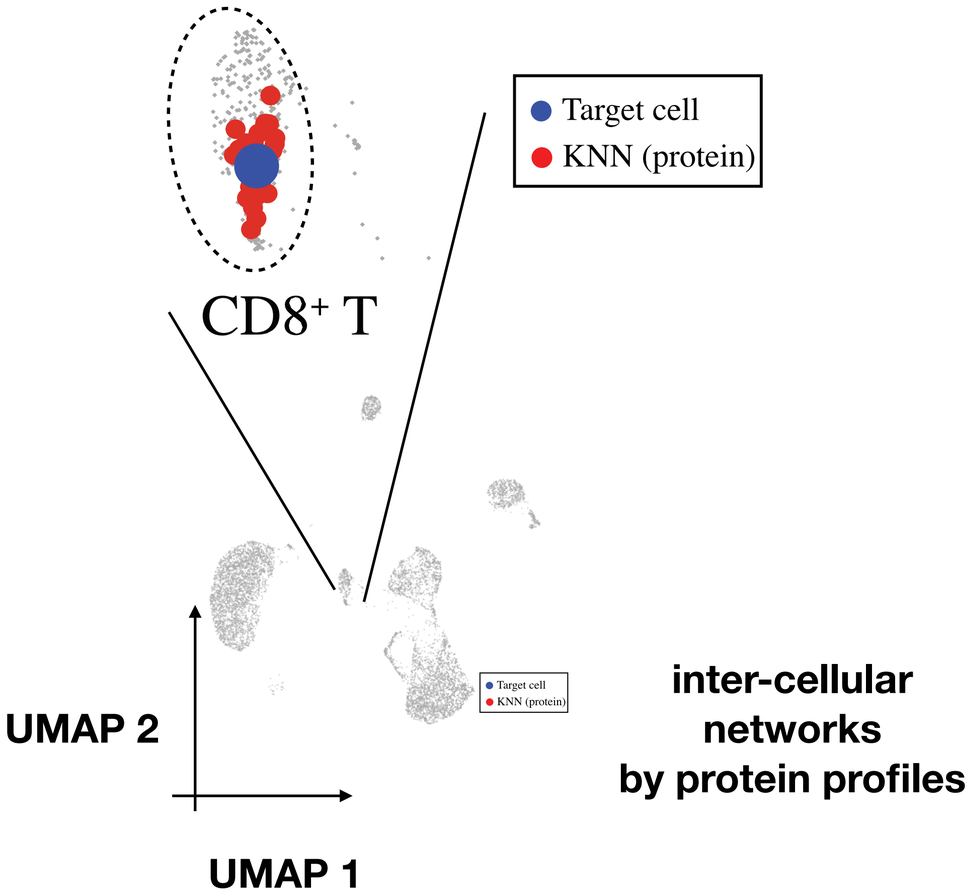
Discussion: Pros and cons for MOFA

Today's lecture: Multiomics data integration

- **Why do we do multiomics data integration?**
  - view #1: borrowing information across modalities
  - view #2: efforts to provide mechanistic explanations
- **Global, unsupervised multiomics data integration**
  - Multiomics Factorization (and variants)
  - Network-based data integration
- **Local, linking between layers to understand mechanisms**
  - Deep dive into mechanisms of gene regulatory mechanisms

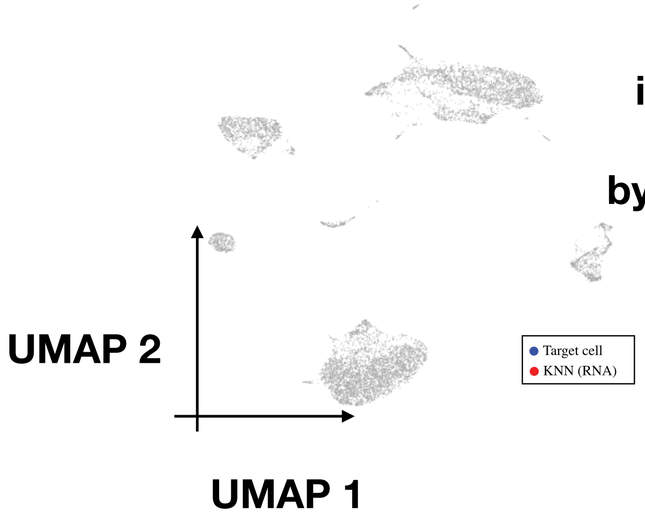
Can we integrate multiple types of data using multiple types of cell-cell networks?



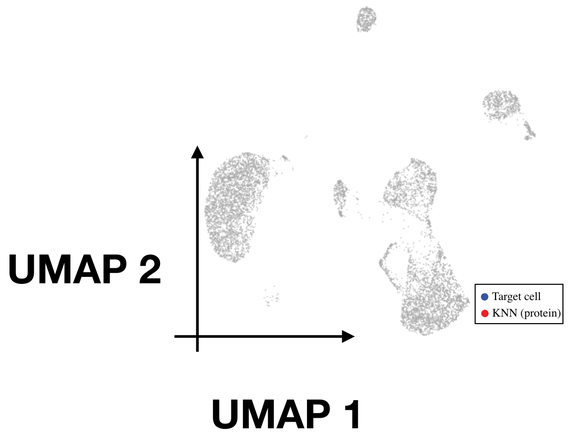


Hao .. Satija, Cell (2021)

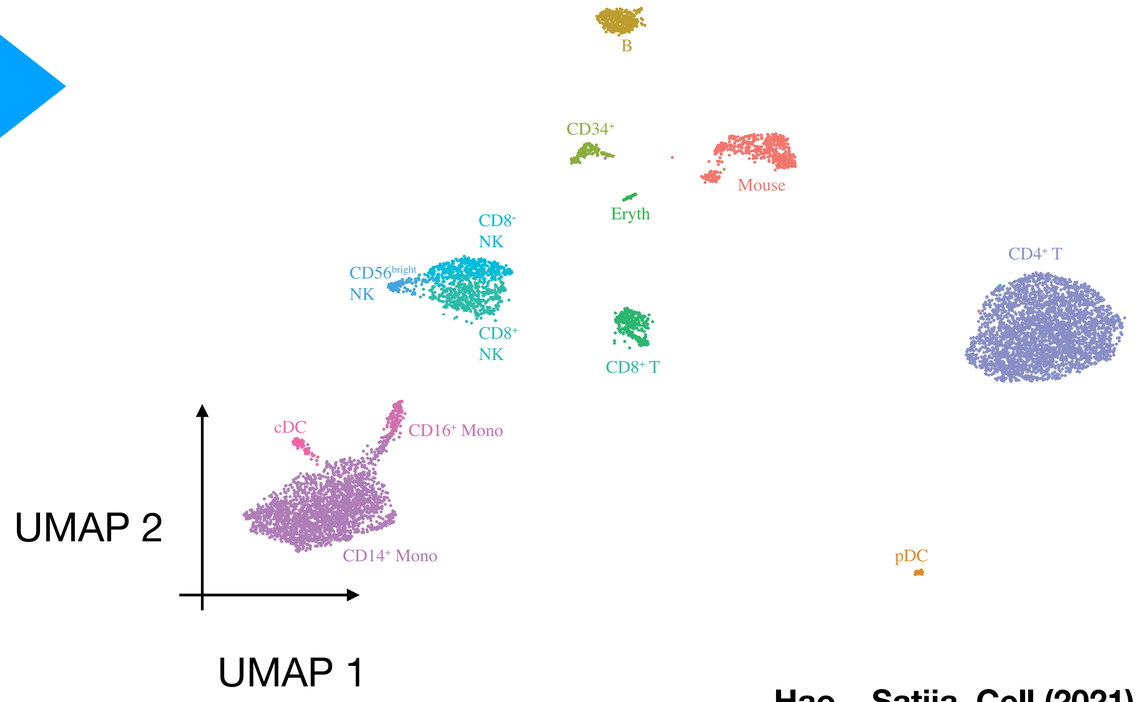
Reconcile two weighted networks to build one consensus network (borrow info from each other)



inter-cellular networks by RNA profiles

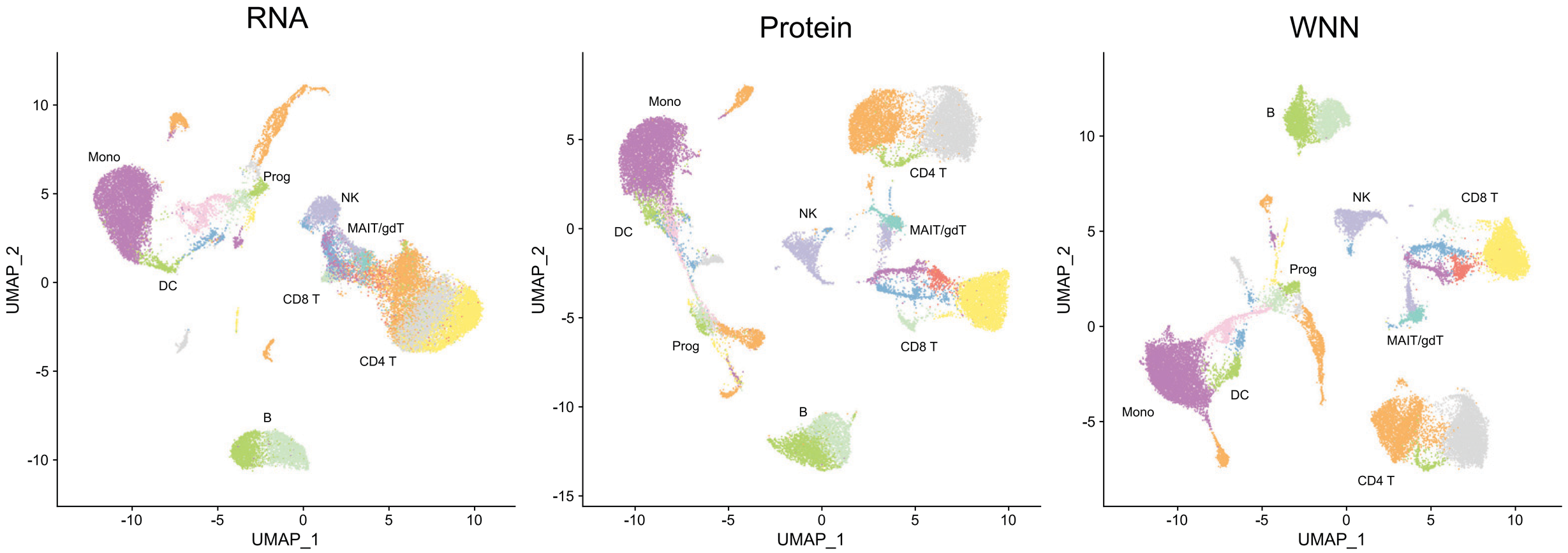


inter-cellular networks by protein profiles



Hao .. Satija, Cell (2021)

Network-based integration works well if the goal of the analysis is to identify cell groups

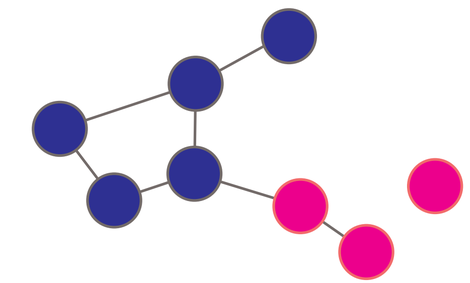


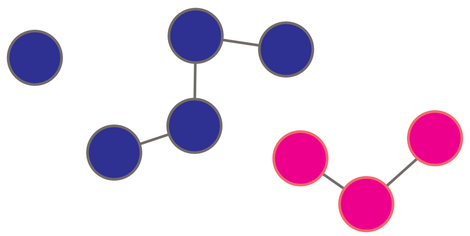
Hao .. Satija, Cell (2021)

What is the goal of this type of integration?

Goal:
1. Fill in missing info
2. Smooth out outliers

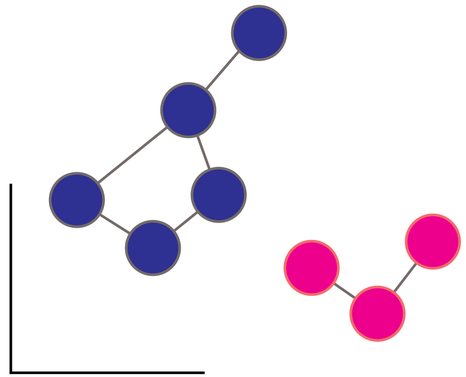
Cell-cell interaction network in data #1



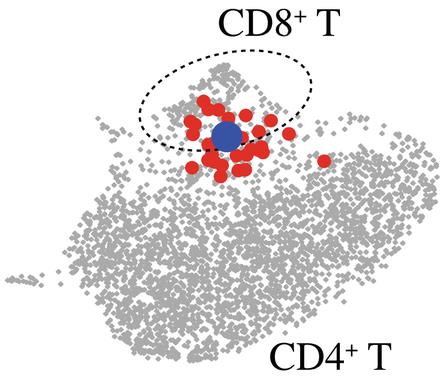


Cell-cell interaction network in data #2

Consensus of two results

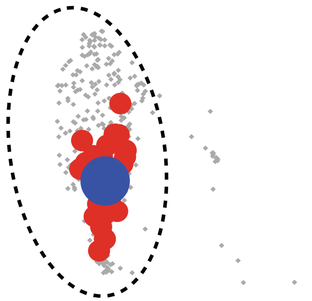


K-nearest neighbour can predict gene expressions and protein activities (in the same domain)



Target cell
KNN (RNA)

$$\hat{r}_{i,knn_r} = \frac{\sum_{j=1}^{k} r_{knn_{r,i,j}}}{k}$$: prediction of RNA profile for cell i, based on RNA neighbors

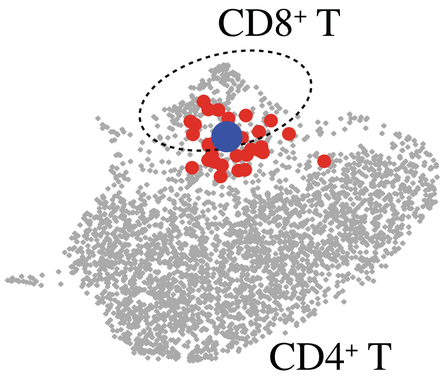


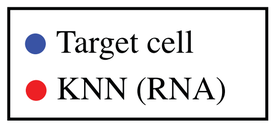
$$\hat{p}_{i,knn_p} = \frac{\sum_{j=1}^{k} p_{knn_{p,i,j}}}{k}$$: prediction of protein profile for cell i, based on protein neighbors

Target cell
KNN (protein)

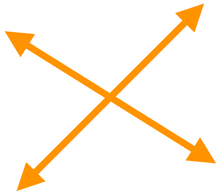
Hao .. Satija, Cell (2021)

If the interaction was consistent, we would share the same neighbourhoods

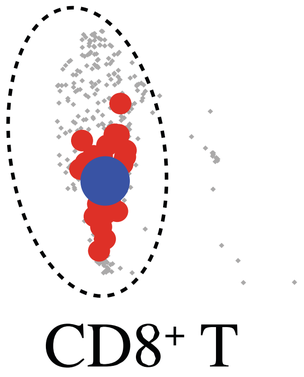


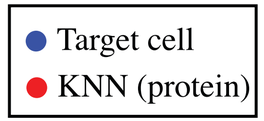


$$\hat{r}_{i,knn_p} = \frac{\sum_{j=1}^{k} r_{knn_{p,i,j}}}{k}$$: prediction of RNA profile for cell i, based on protein neighbors



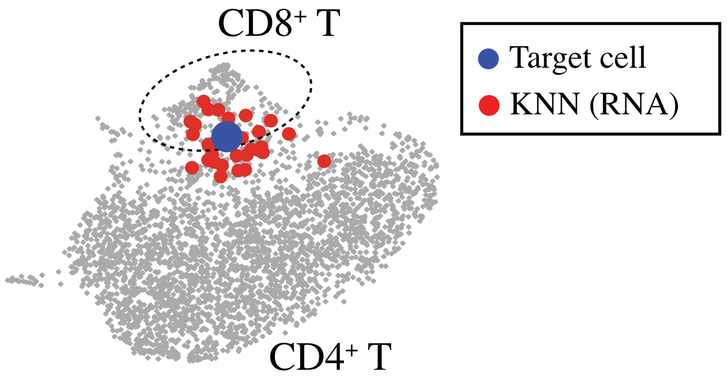
$$\hat{p}_{i,knn_r} = \frac{\sum_{j=1}^{k} p_{knn_{r,i,j}}}{k}$$: prediction of protein profile for cell i, based on RNA neighbors





Hao .. Satija, Cell (2021)

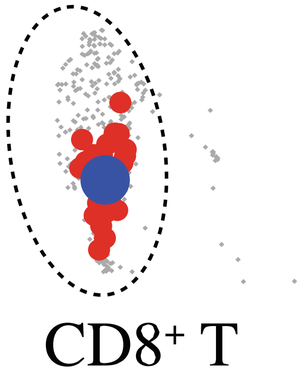
How do you put them together?
Take the weighted average over multiple types of edges



$$s_{rna}(i) = \frac{\theta_{rna}\left(r_i, \hat{r}_{i,knn_r}\right)}{\theta_{rna}\left(r_i, \hat{r}_{i,knn_p}\right) + \varepsilon}, \quad s_{protein}(i) = \frac{\theta_{protein}\left(p_i, \hat{p}_{i,knn_p}\right)}{\theta_{protein}\left(p_i, \hat{p}_{i,knn_r}\right) + \varepsilon}$$

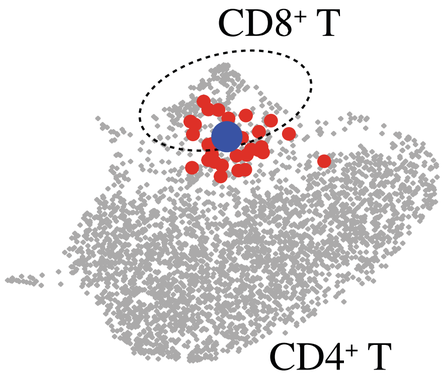
How much RNA weights can be explained by RNA or imputed by protein?

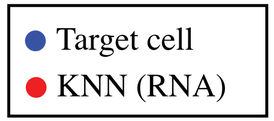
How much protein weights can be explained by protein or imputed by RNA?



Hao .. Satija, Cell (2021)

How do you put them together?
Take the weighted average over multiple types of edges

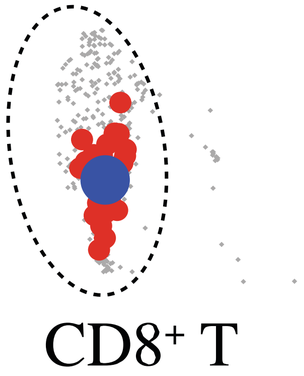


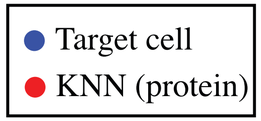


$$s_{rna}(i) = \frac{\theta_{rna}\left(r_i, \hat{r}_{i,knn_r}\right)}{\theta_{rna}\left(r_i, \hat{r}_{i,knn_p}\right) + \varepsilon}, \; s_{protein}(i) = \frac{\theta_{protein}\left(p_i, \hat{p}_{i,knn_p}\right)}{\theta_{protein}\left(p_i, \hat{p}_{i,knn_r}\right) + \varepsilon}$$

$$w_{rna}(i) = \frac{e^{s_{rna}(i)}}{e^{s_{rna}(i)} + e^{s_{protein}(i)}}, \; w_{protein}(i) = \frac{e^{s_{protein}(i)}}{e^{s_{rna}(i)} + e^{s_{protein}(i)}}$$

$$\theta_{weighted}(i,j) = w_{rna}(i)\theta_{rna}(r_i, r_j) + w_{protein}(i)\theta_{protein}(p_i, p_j)$$





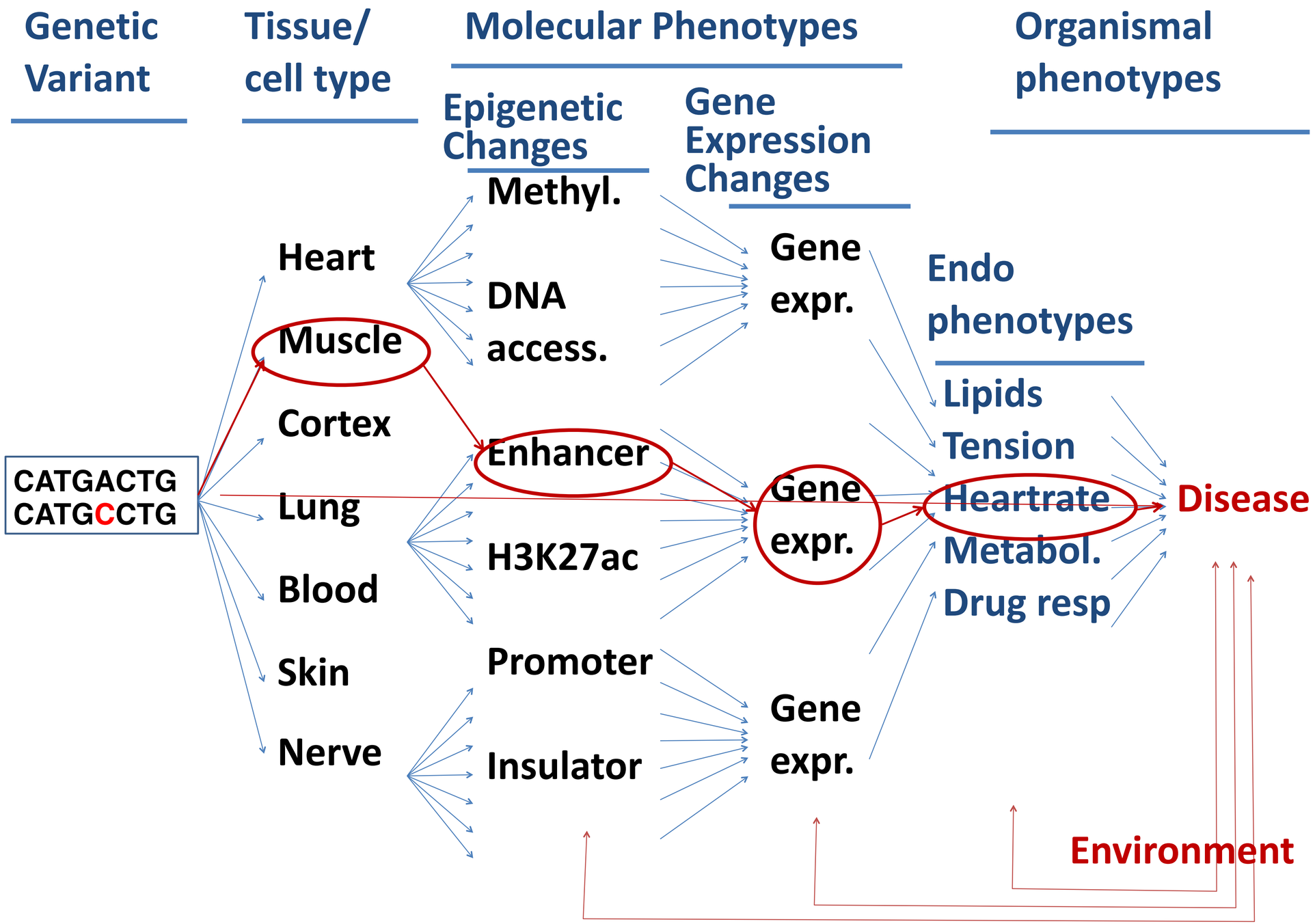
Hao .. Satija, Cell (2021)

Pros and cons for Weighted Network Approach

Today's lecture: Multiomics data integration

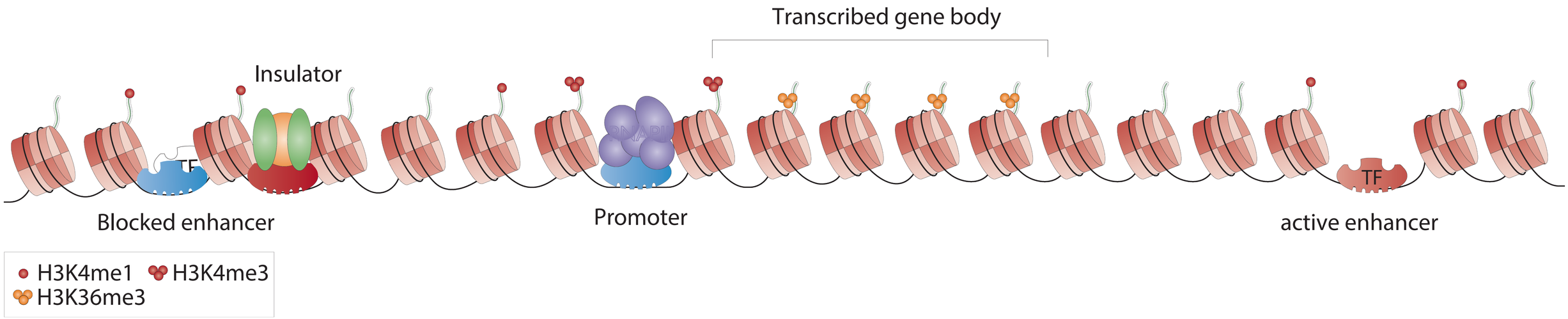
- **Why do we do multiomics data integration?**
  - view #1: borrowing information across modalities
  - view #2: efforts to provide mechanistic explanations
- **Global, unsupervised multiomics data integration**
  - Multiomics Factorization (and variants)
  - Network-based data integration
- **Local, linking between layers to understand mechanisms**
  - Deep dive into mechanisms of gene regulatory mechanisms


Where is
CCTTCTGTG
TCGGA



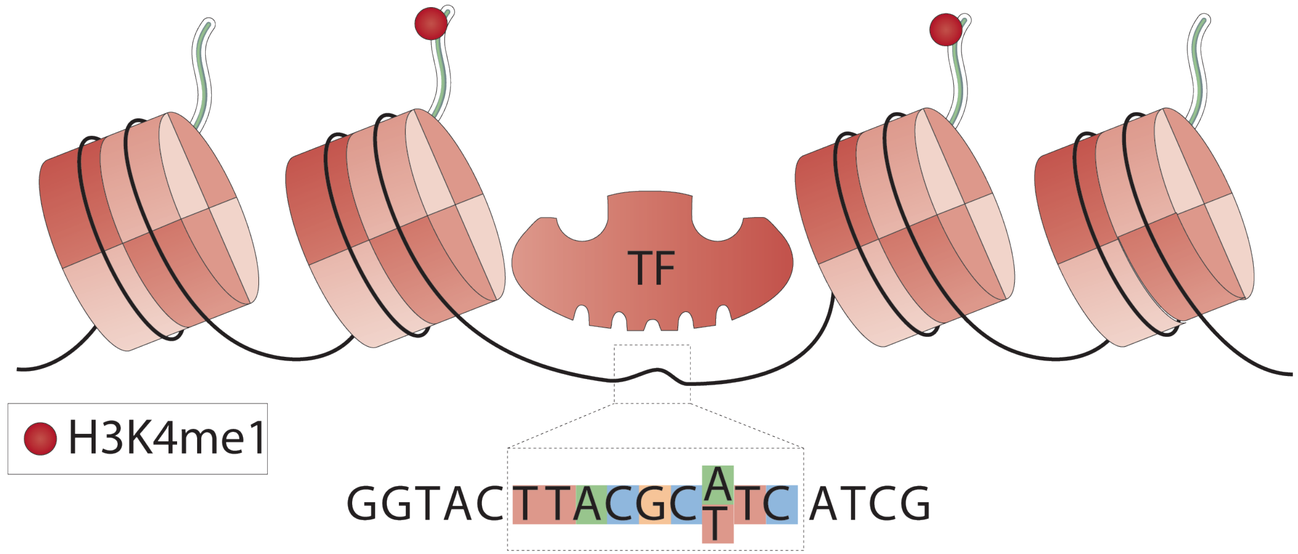
Slide credit: Manolis Kellis

Different histone code marks different types of regulatory elements



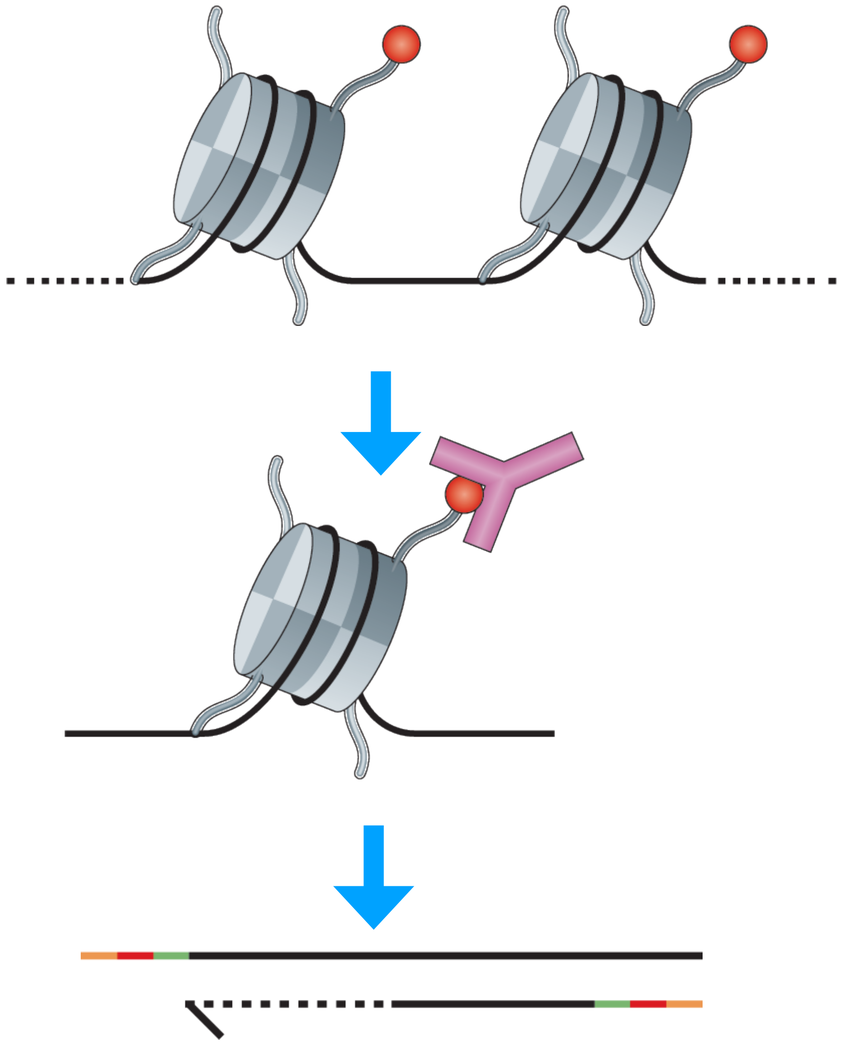
Hawkins, Hon, Ren, *Nature Review Genetics* (2010)

One of the goals is to understand the logic of non-coding DNA sequences



Hawkins, Hon, Ren, *Nature Review Genetics* (2010)

ChIP-seq: Quantifying histone modifications



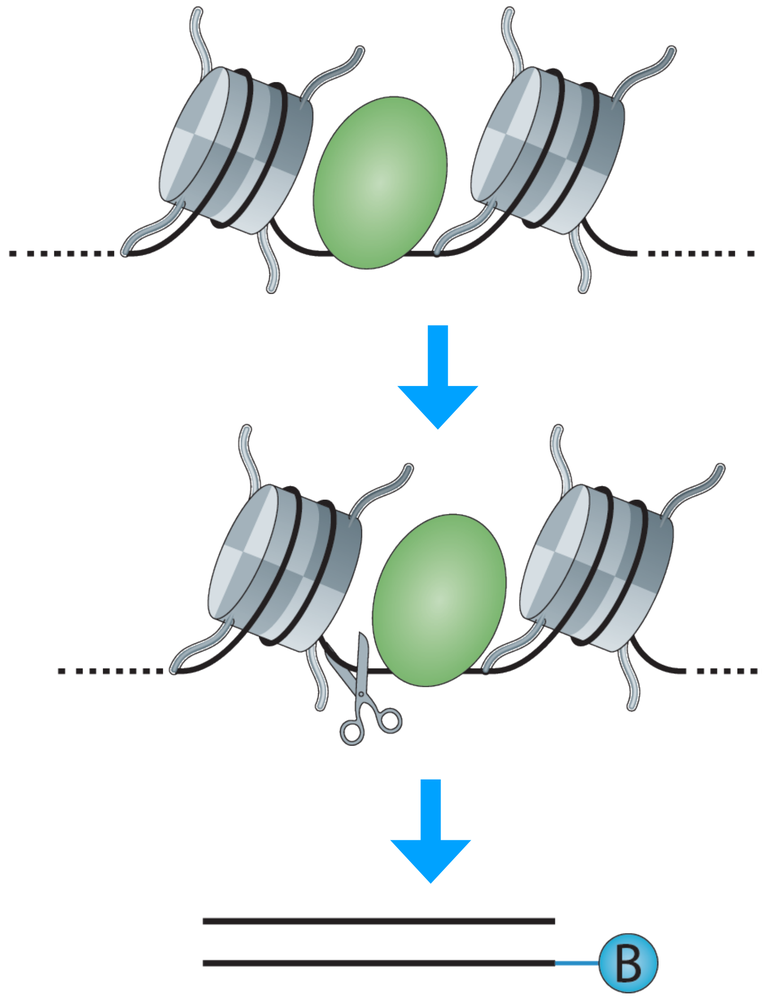
Crosslink proteins & DNA

Immunoprecipitate &
purify DNA bound

Amplify/duplicate
the extracted
DNA segments

Furey, *Nature Reviews Genetics* (2012)

DNase-seq: Quantifying DNA accessibility



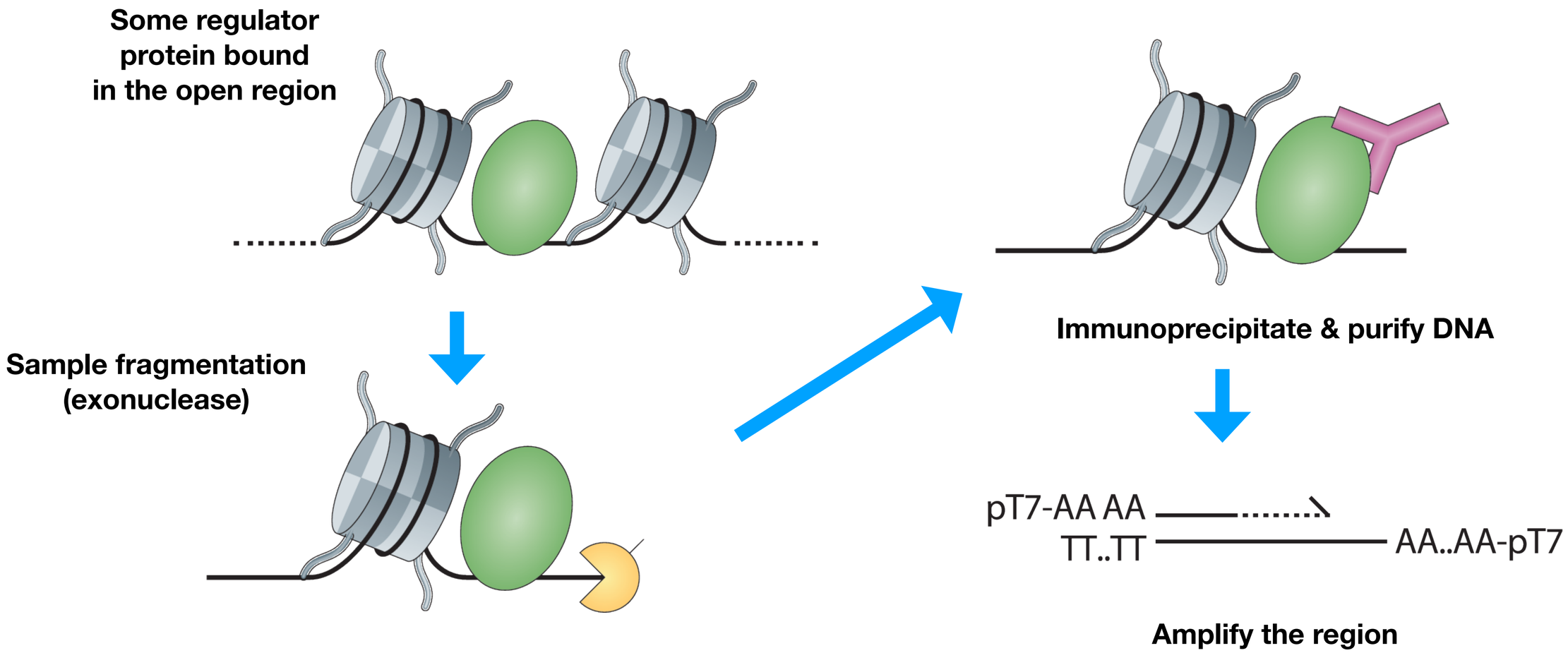
Some regulator protein bound
in the open region

Sample fragmentation
(DNaseI digestion)

Add biotynylated linkers
and extract

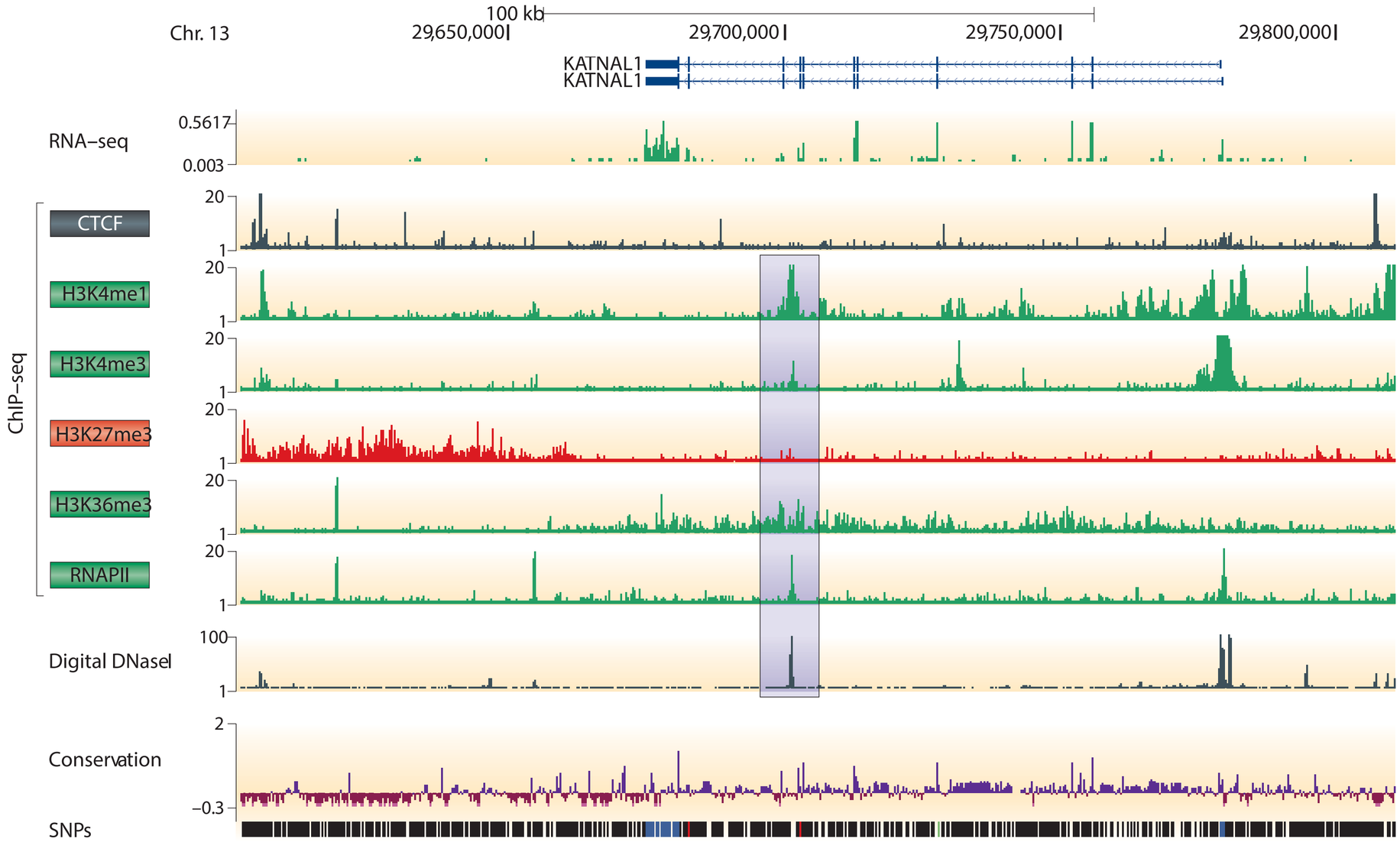
Furey, *Nature Reviews Genetics* (2012)

ChIP-seq: Quantifying regions bound by a particular protein



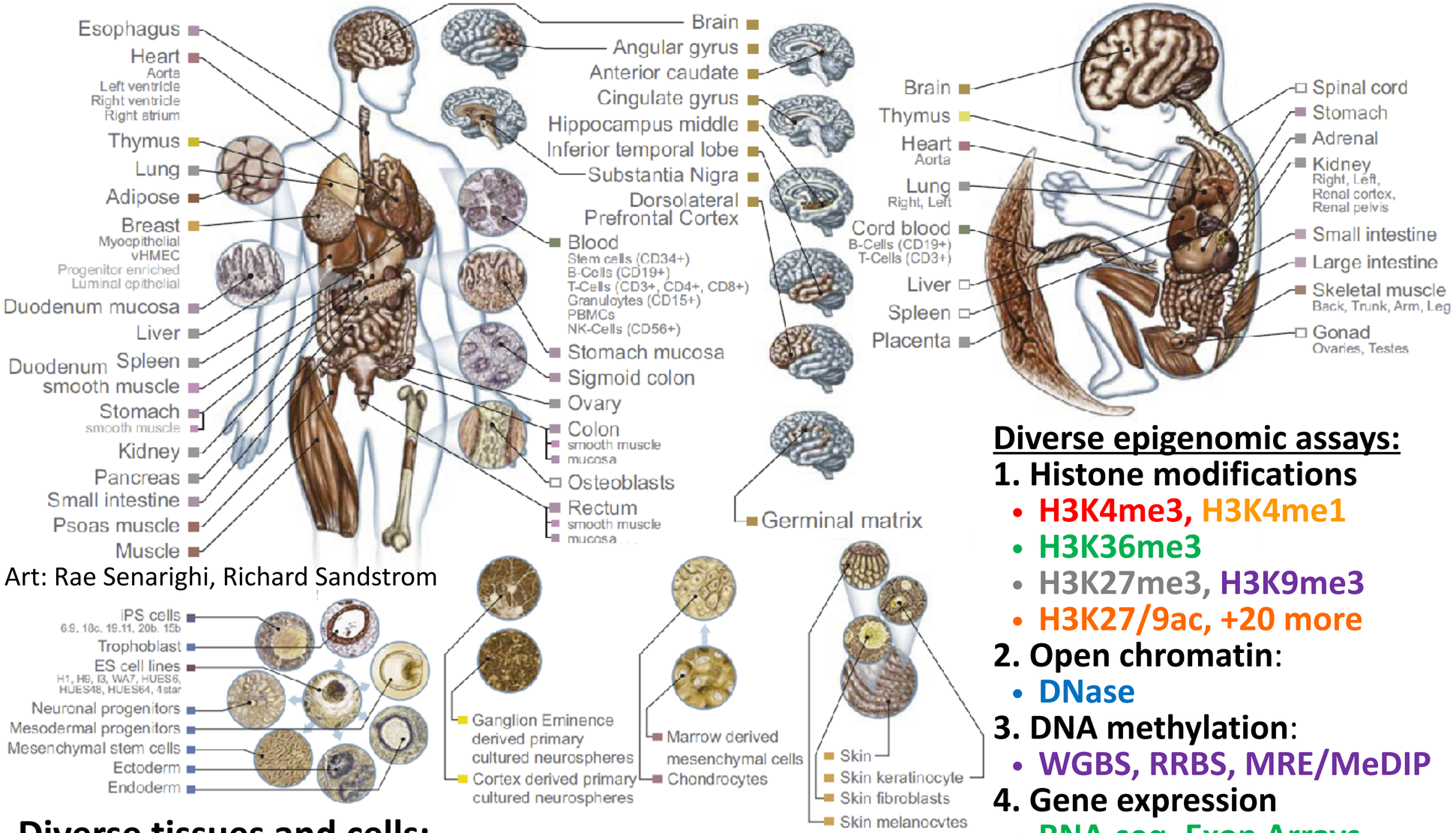
Furey, *Nature Reviews Genetics* (2012)

An example of multiple ChIP-seq tracks



Hawkins, Hon, Ren, *Nature Review Genetics* (2010)

Epigenomics Roadmap across 100+ tissues/cell types



Art: Rae Senarighi, Richard Sandstrom

Diverse epigenomic assays:

1. **Histone modifications**
   - **H3K4me3, H3K4me1**
   - **H3K36me3**
   - **H3K27me3, H3K9me3**
   - **H3K27/9ac, +20 more**
2. **Open chromatin**:
   - **DNase**
3. **DNA methylation**:
   - **WGBS, RRBS, MRE/MeDIP**
4. **Gene expression**
   - **RNA-seq, Exon Arrays**

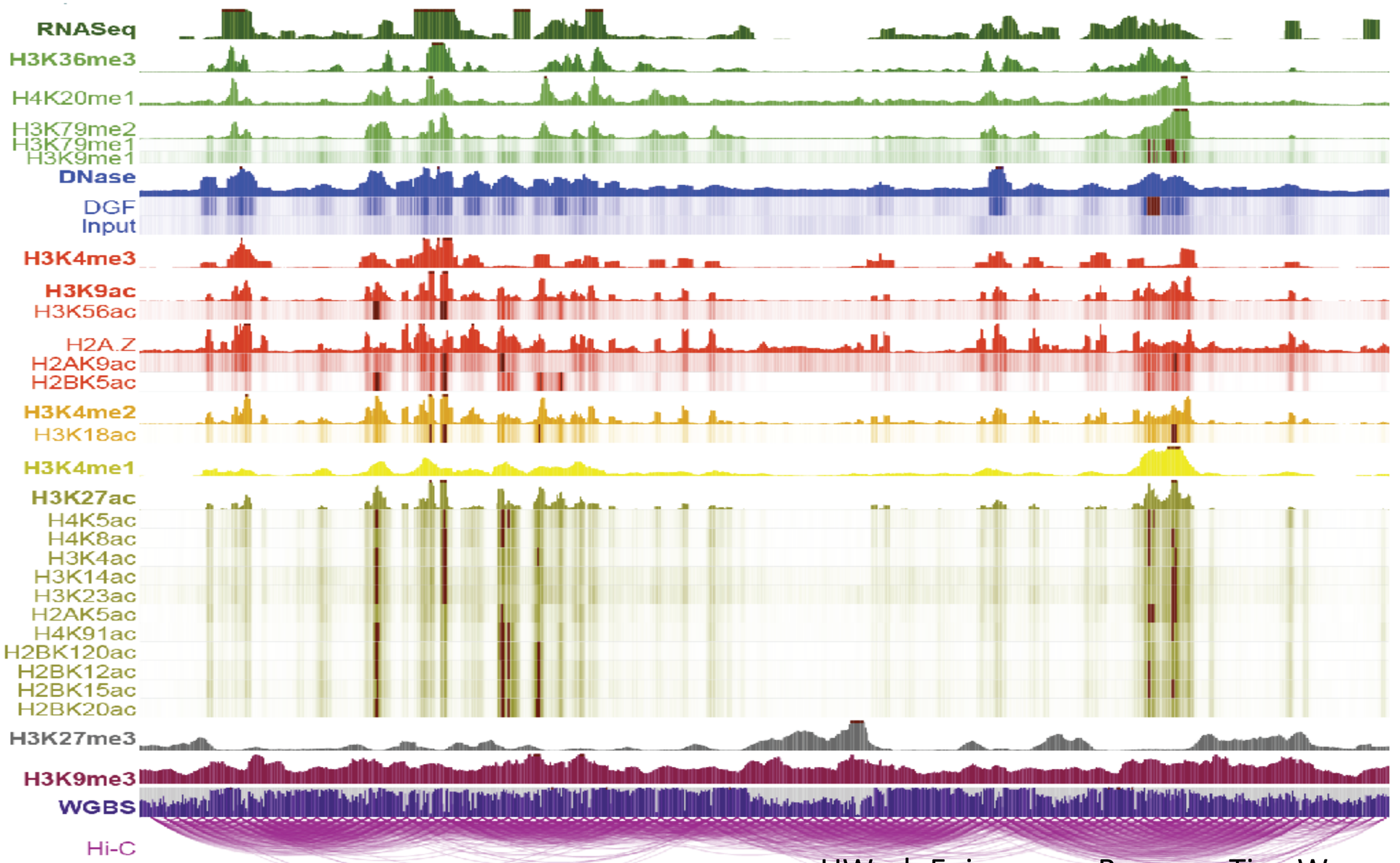
Slide: Manolis Kellis

Diverse tissues and cells:

1. **Adult tissues and cells** (brain, muscle, heart, digestive, skin, adipose, lung, blood…)
2. **Fetal tissues** (brain, skeletal muscle, heart, digestive, lung, cord blood…)
3. **ES cells, iPS, differentiated cells** (meso/endo/ectoderm, neural, mesench, trophobl)

Deep sampling of 9 reference epigenomes (e.g. IMR90)

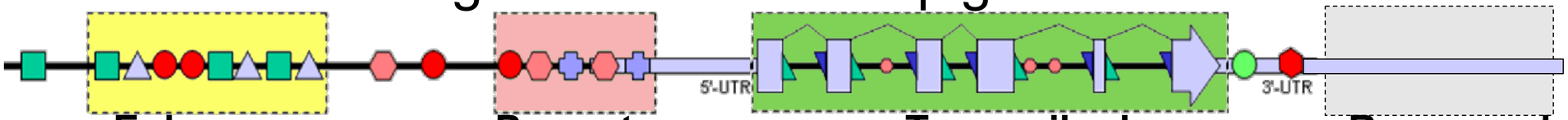
Slide: Manolis Kellis



UWash Epigenome Browser, Ting Wang

Chromatin state+RNA+DNAse+28 histone marks+WGBS+Hi-C

Diverse chromatin signatures encode epigenomic state

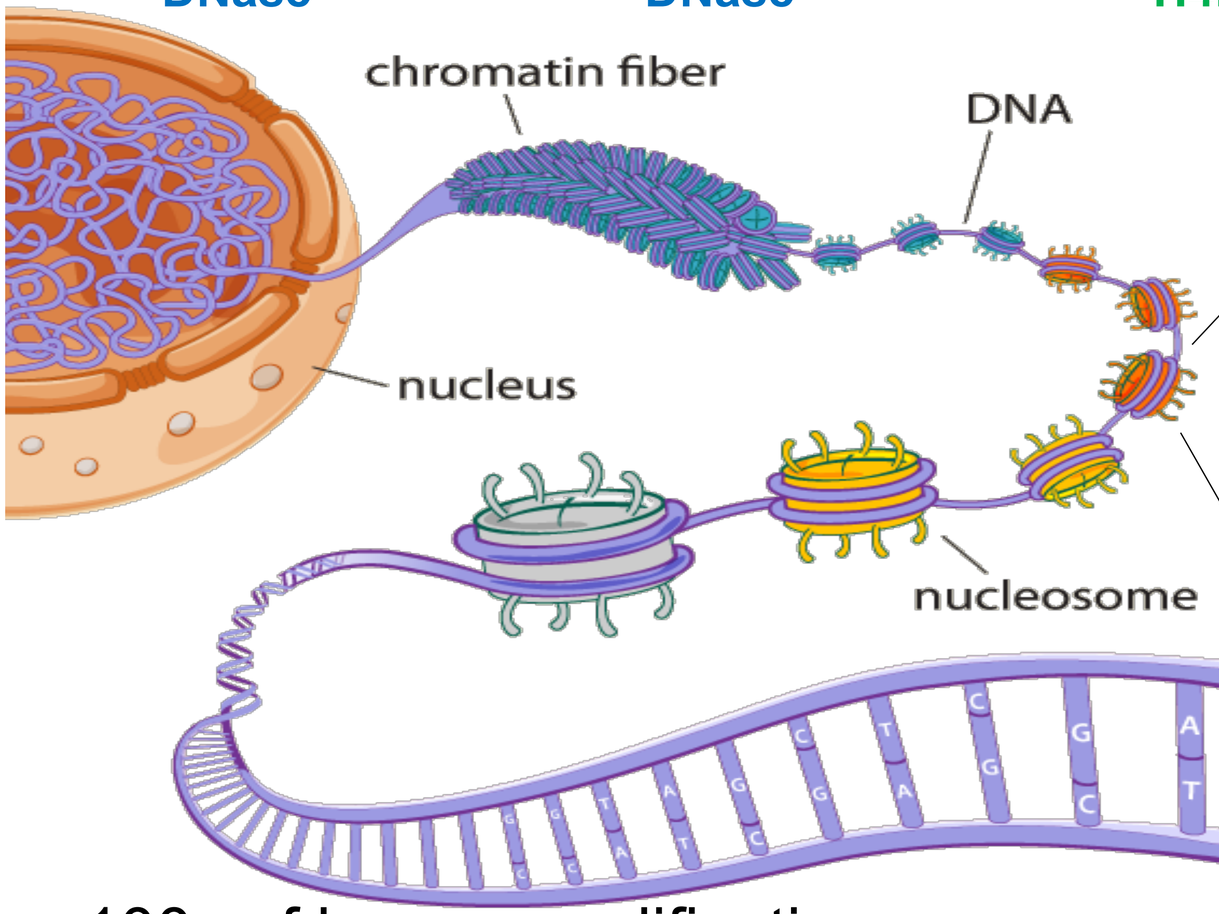


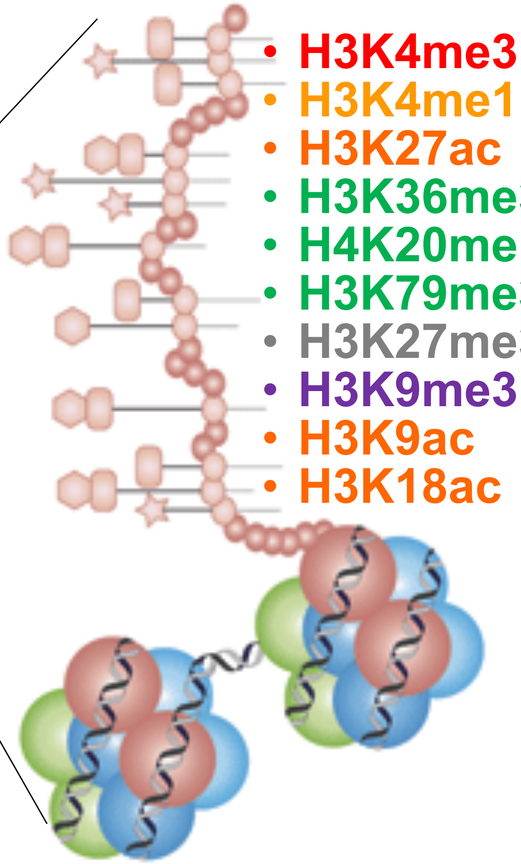
Enhancers
- H3K4me1
- H3K27ac
- DNase

Promoters
- H3K4me3
- H3K9ac
- DNase

Transcribed
- H3K36me3
- H3K79me2
- H4K20me1

Repressed
- H3K9me3
- H3K27me3
- DNAmethyl





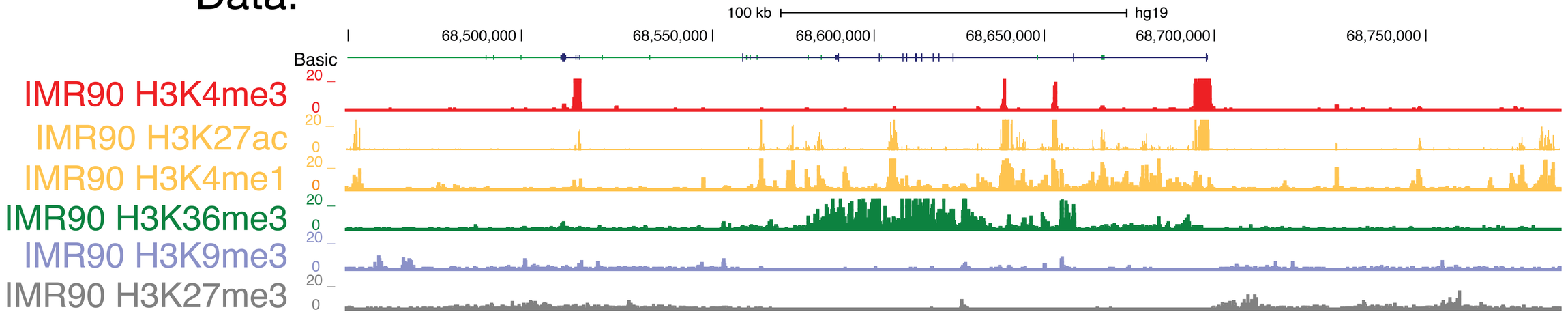
- H3K4me3
- H3K4me1
- H3K27ac
- H3K36me3
- H4K20me1
- H3K79me3
- H3K27me3
- H3K9me3
- H3K9ac
- H3K18ac

Slide: Manolis Kellis

- 100s of known modifications, many new still emerging
- Systematic mapping using ChIP-, Bisulfite-, DNase-Seq

Goal: aggregate multiple ChIP-seq tracks

Data:



Unknown hidden states (epigenetic status; colour)

Ernst & Kellis (2015)

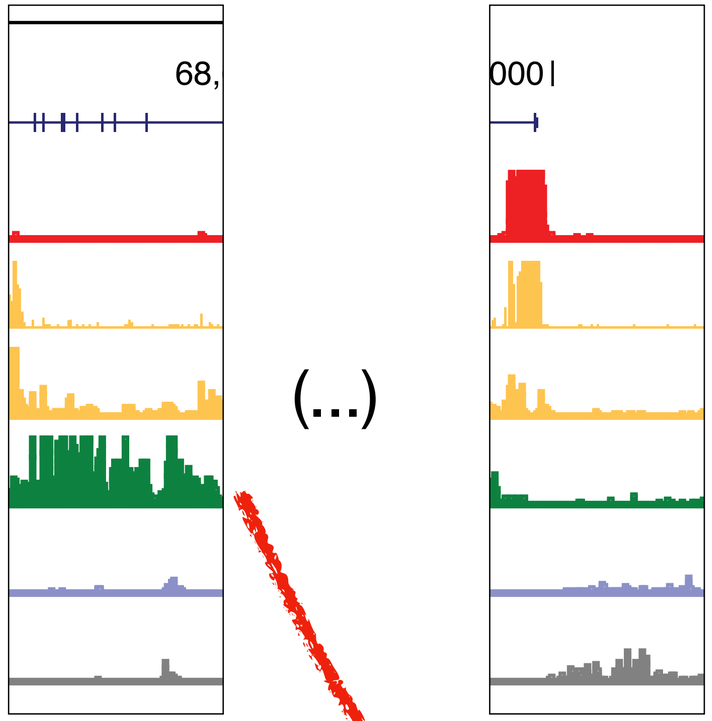
ChromHMM (hidden Markov model)

Data:

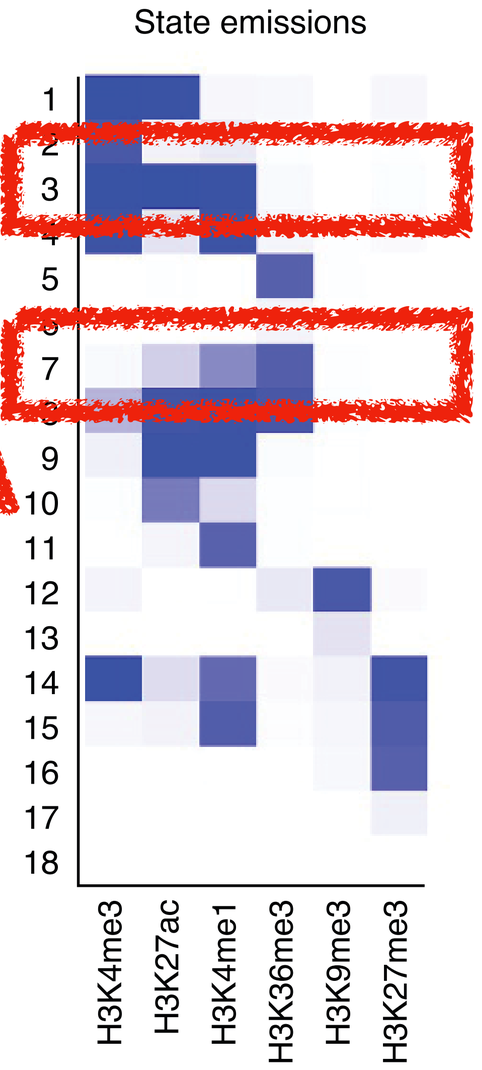
IMR90 H3K4me3
IMR90 H3K27ac
IMR90 H3K4me1
IMR90 H3K36me3
IMR90 H3K9me3
IMR90 H3K27me3

Systematic annotation:

(...)

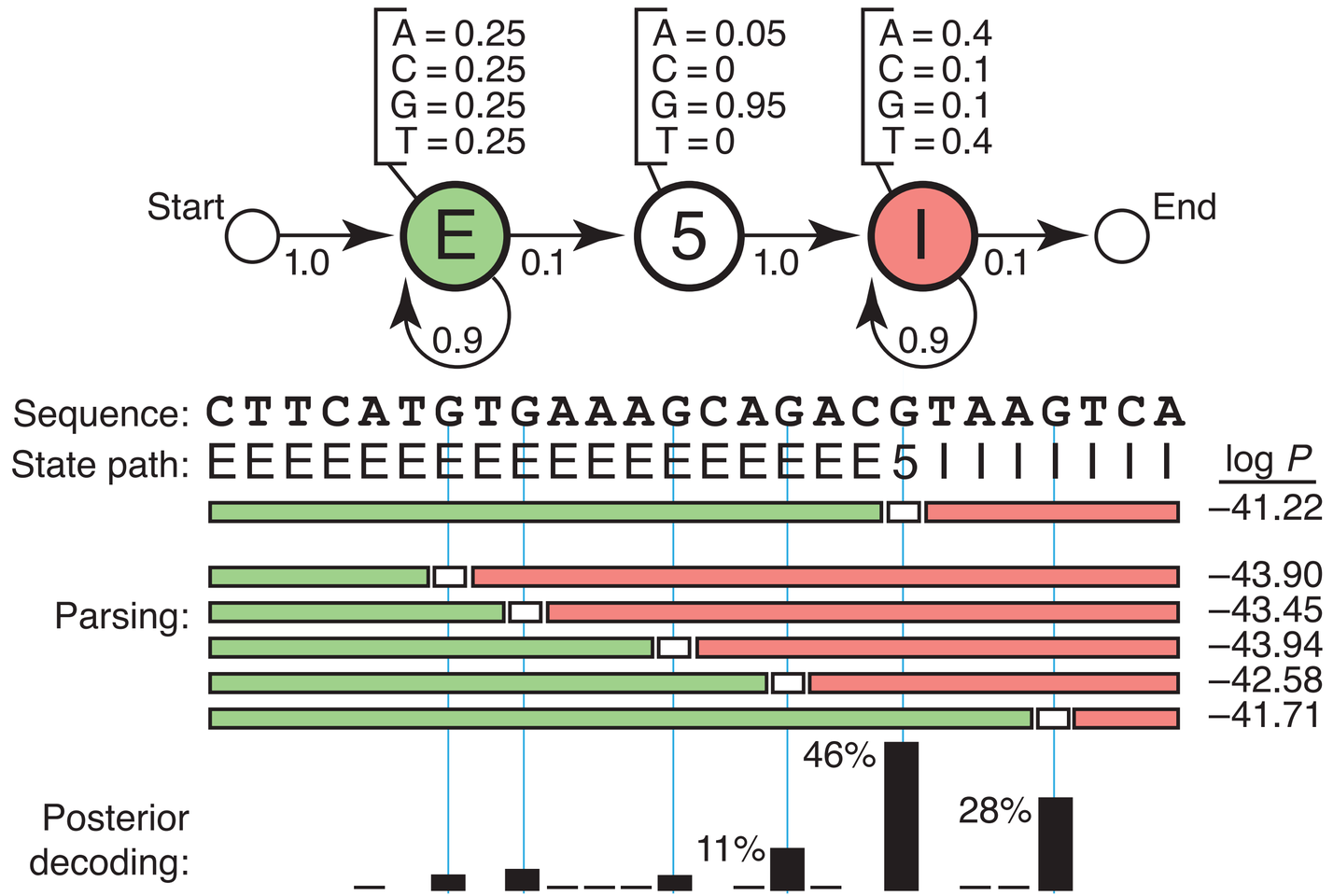


(...)



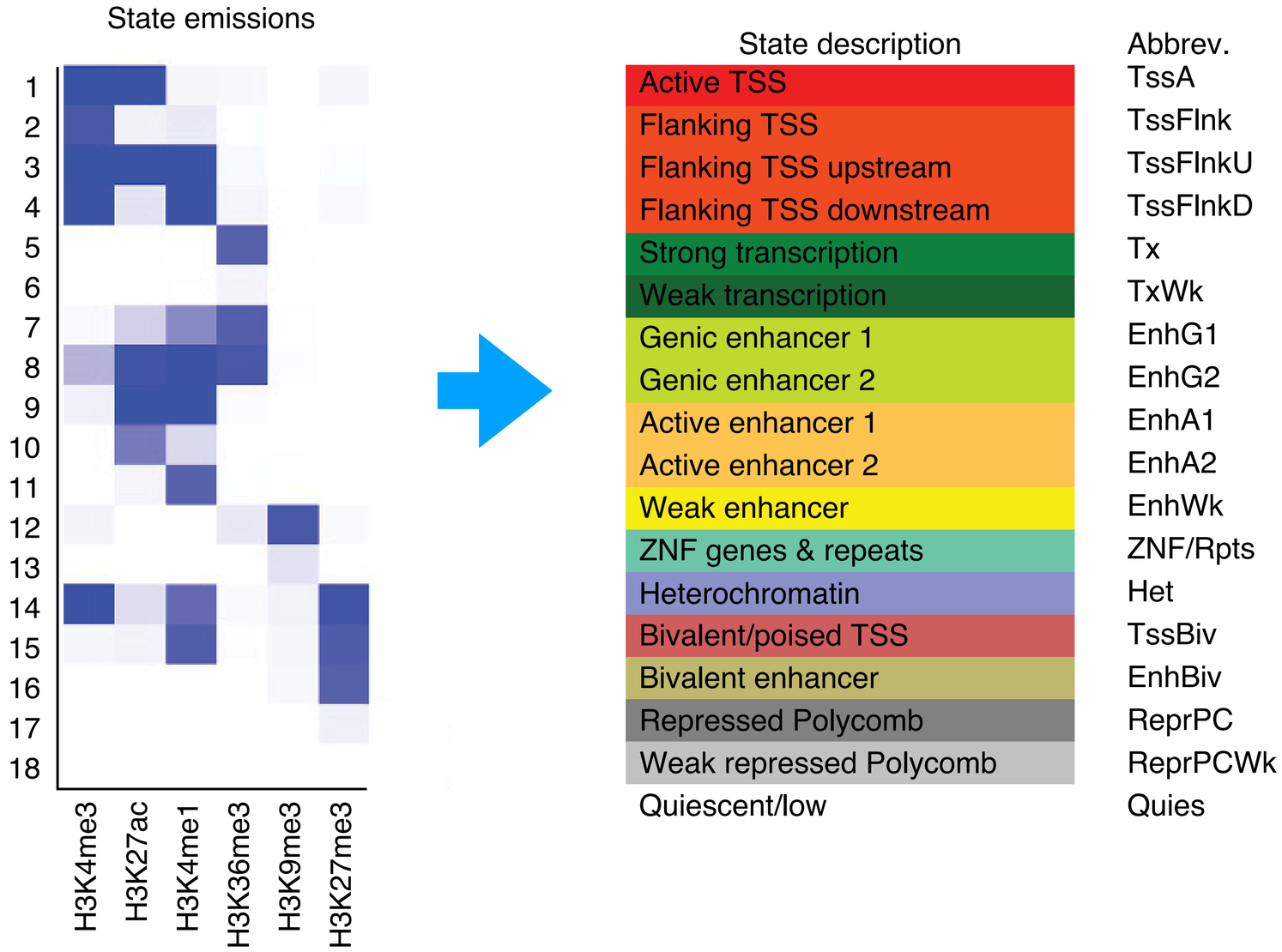
Ernst & Kellis (2015)

HMM has been used in multiple bioinformatics (especially DNA sequence analysis)



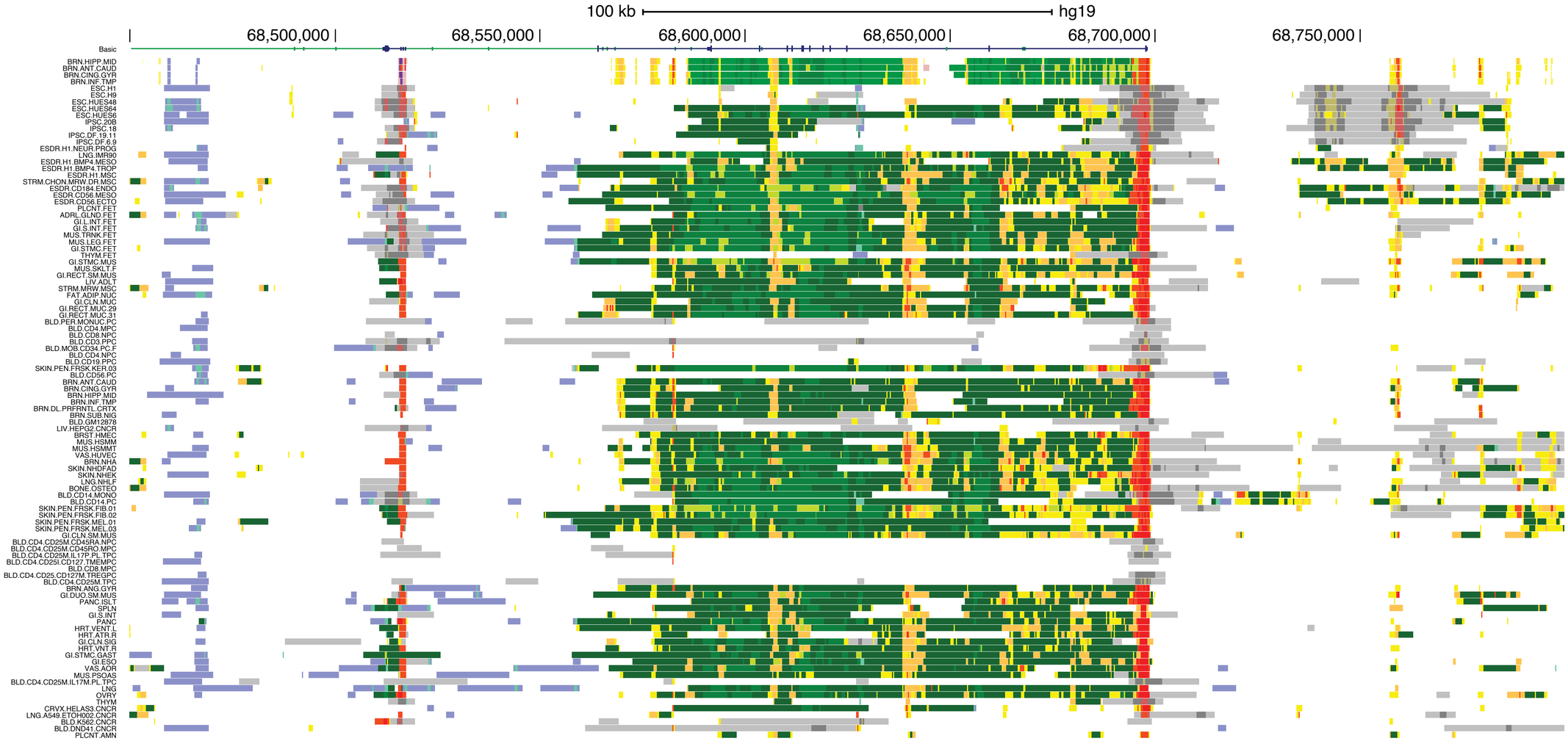
Eddy, *Nature Biotech.* (2004)

ChromHMM (hidden Markov model)



Ernst & Kellis (2015)

ChromHMM: multiple ChIP-seq signal tracks → systematic epigenome annotation



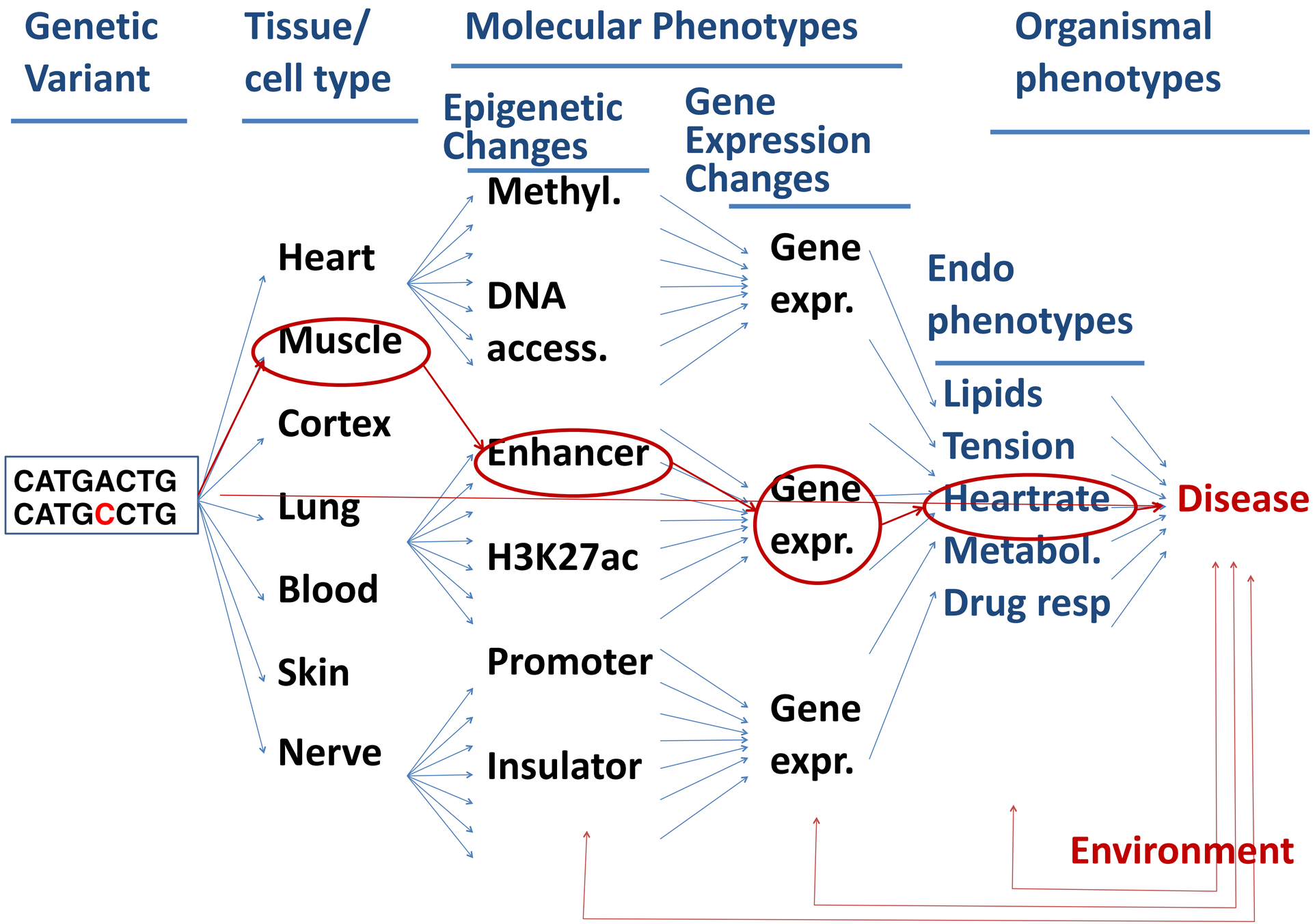
Ernst & Kellis (2015)

Discussion: ChromHMM

Do we have other alternatives?

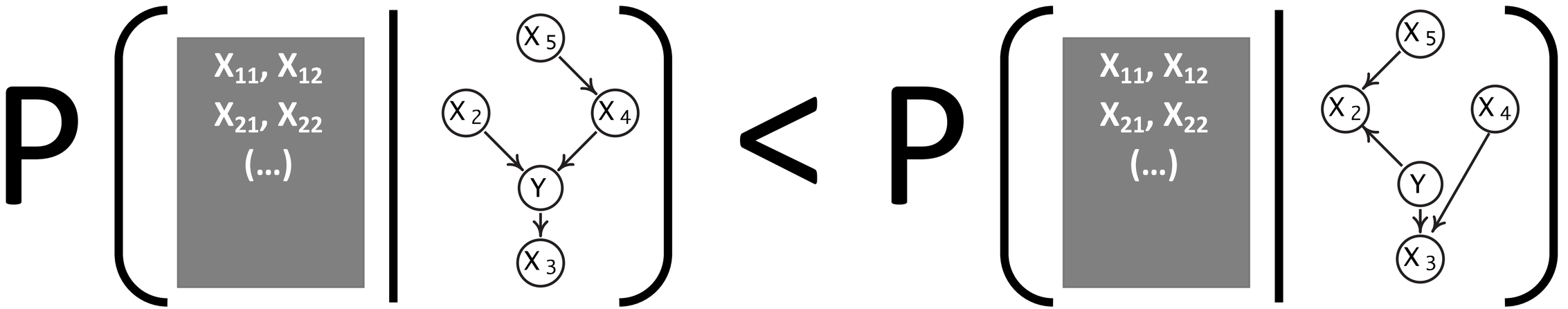
Today's lecture: Multiomics data integration

- **Why do we do multiomics data integration?**
  - view #1: borrowing information across modalities
  - view #2: efforts to provide mechanistic explanations
- **Global, unsupervised multiomics data integration**
  - Multiomics Factorization (and variants)
  - Network-based data integration
- **Local, linking between layers to understand mechanisms**
  - Deep dive into mechanisms of gene regulatory mechanisms



Slide credit: Manolis Kellis

Isn't it about finding the most probable structure?



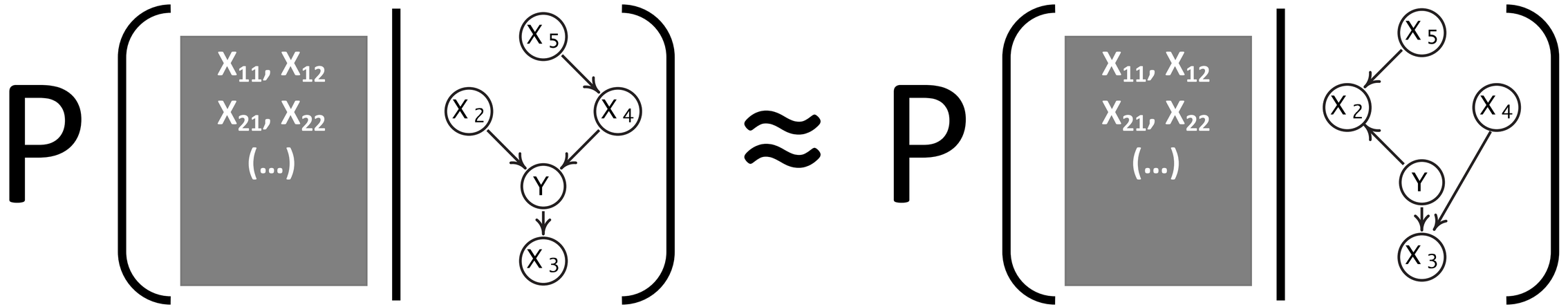
High-dimensional data matrix

Underlying causal model structure

Try a different structure.
Can we improve the data likelihood?

"Large-sample learning of Bayesian Network is NP-hard", Chickering .. Heckerman, UAI (2002)

A traditional structural learning may not achieve identifiablity … (Even if we increase N)

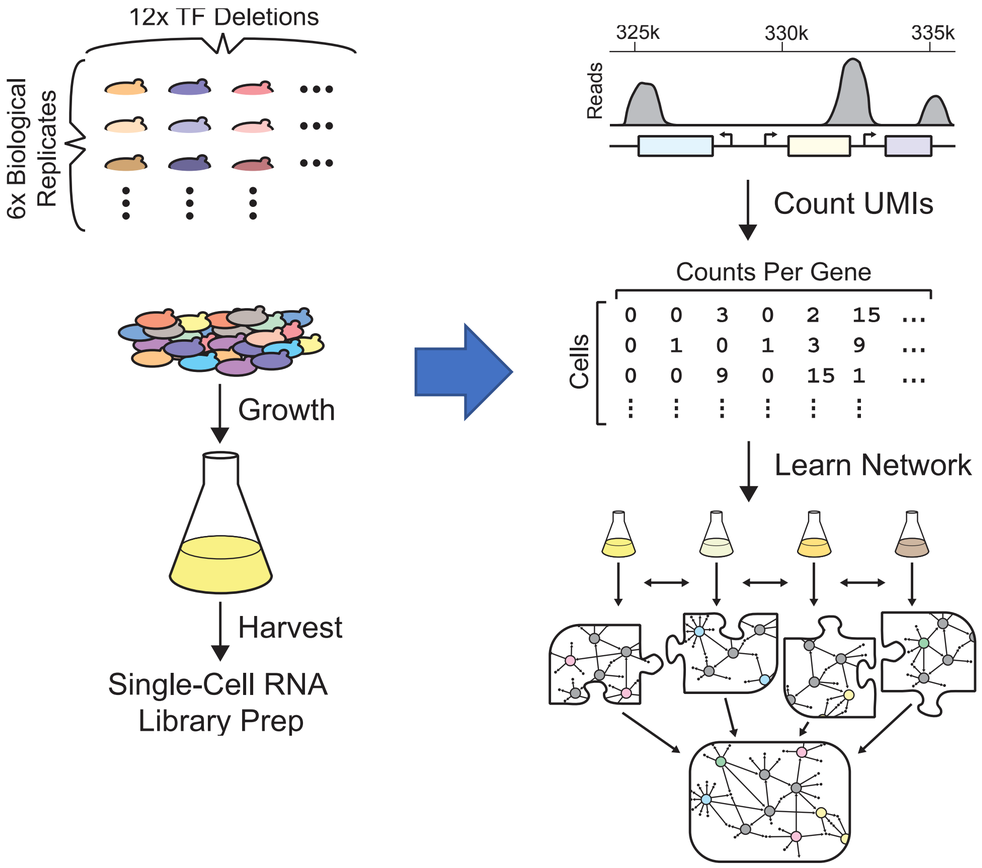


High-dimensional data matrices

Underlying causal model structure

What if our probability score is more or less the same for different structures?

Can we borrow the awesome power of genetics?



Jackson *et al.* (2020)

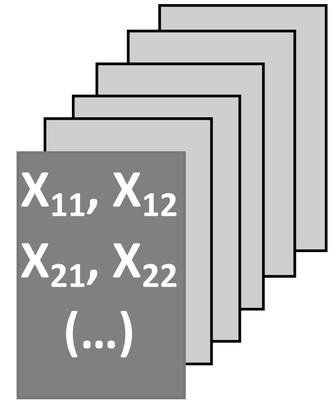
Early 2000's

X_{11}, X_{12}, X_{13}
X_{21}, X_{22}, X_{23}
(…)

A single or only a handful of conditions

Many computational biologists gave up on this NP-hard problem… learning BN from data

After perturb-seq

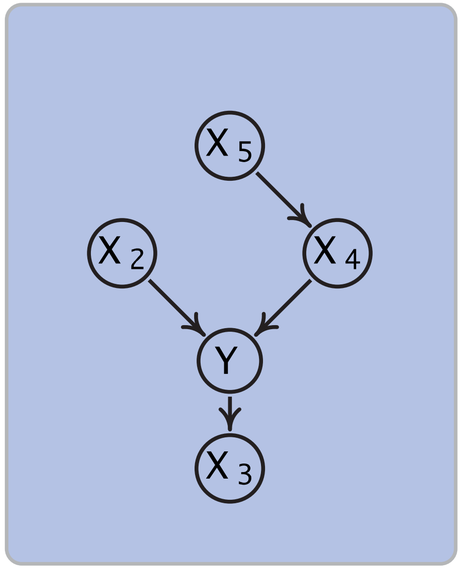


Massive (random) perturbation assays → hundreds of data matrices

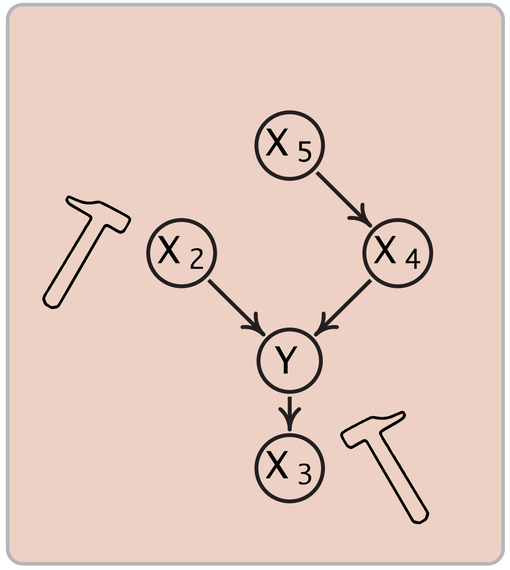
Causal Discovery by Invariant Predictions

The premise of Causation by Invariance:

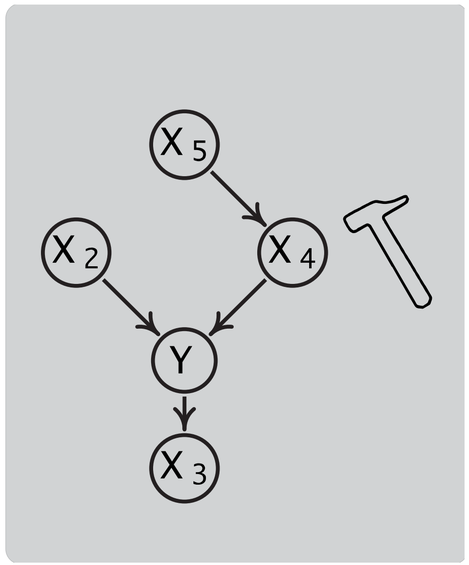
There is a causal structure that remains *invariant* across multiple experiments (perturbation-invariant)



No perturbation



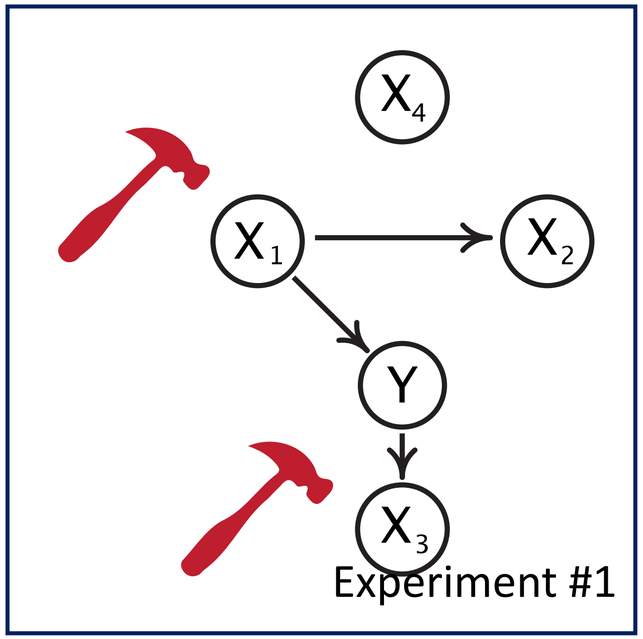
Gene knock-out #2, #3

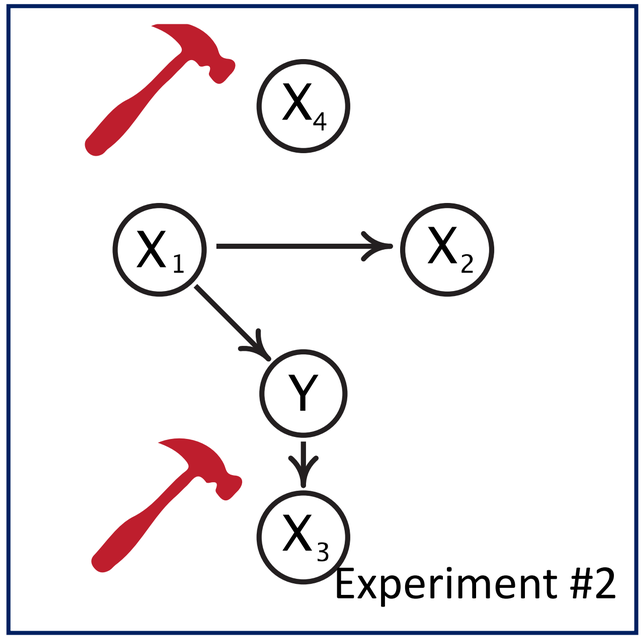


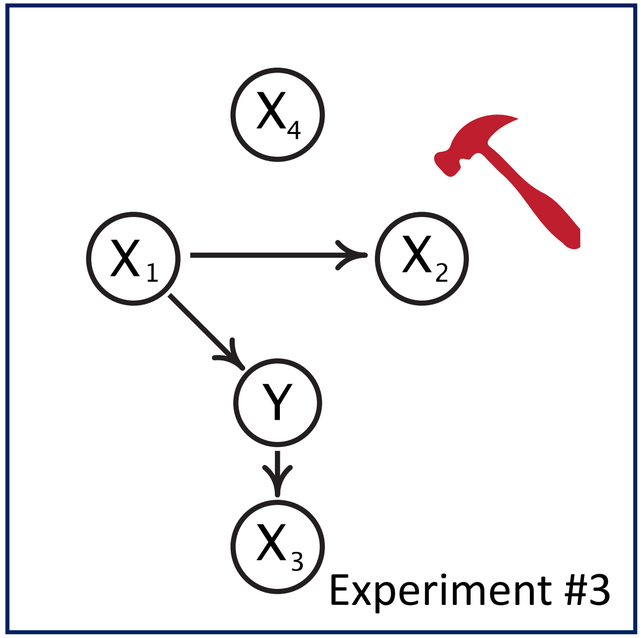
Gene knock-out #4

Peters, Buhlmann, Meinshausen, J.R. Statist. Soc. B (2016)

Causal Discovery by Invariance (toy example)





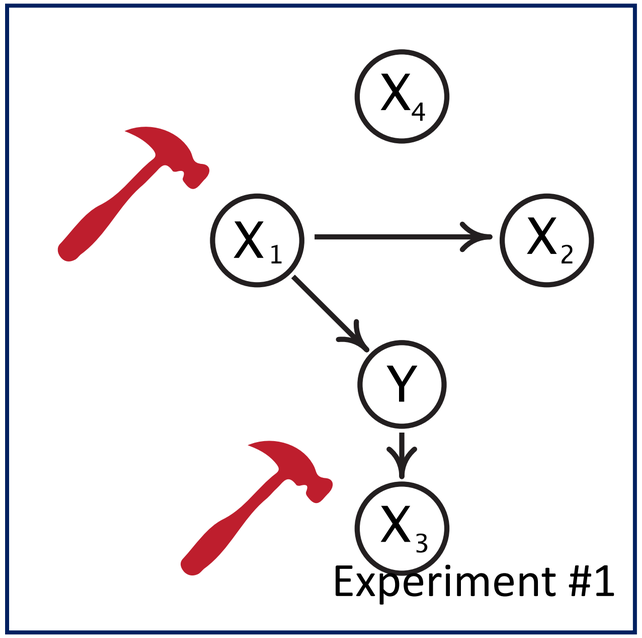


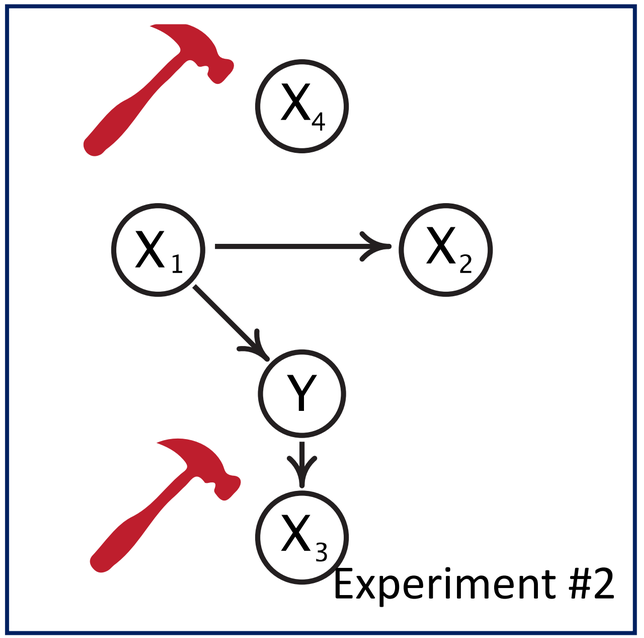
Knock-out do(X=0)

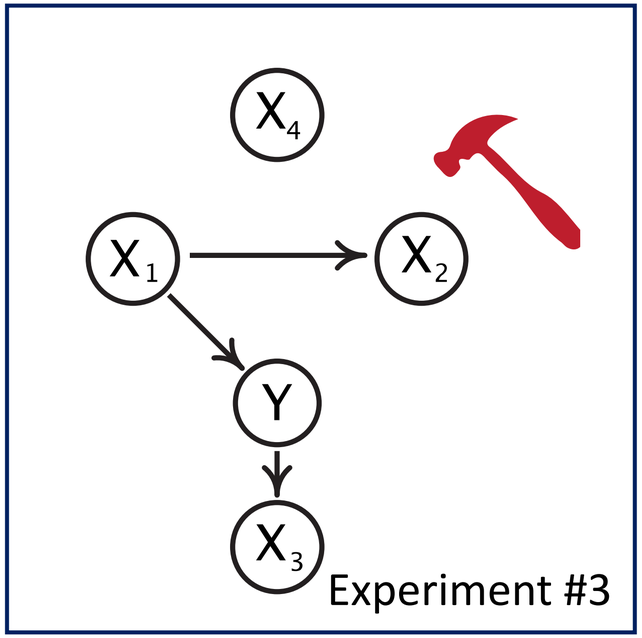
Causal Discovery by Invariance ≈ excluding non-parental variables

do(X=0)

Q. Does this model hold the invariance condition? $Y = X_1 + \varepsilon$







Observation:

$Y = \varepsilon_y$

$X_1 = 0$

$Y = X_1 + \varepsilon_y$

$X_1 = \varepsilon_1$

$Y = X_1 + \varepsilon_y$

$X_1 = \varepsilon_1$

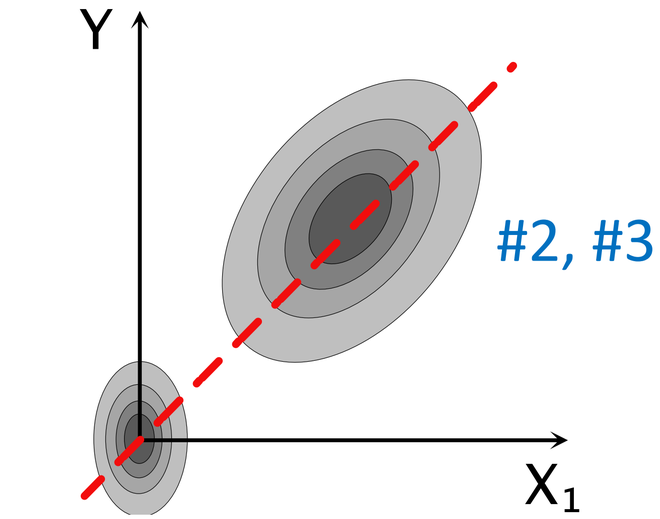
Does this model make sense in all the experiments?

Observation:

$Y = \varepsilon_y$ Experiment #1

$Y = X_1 + \varepsilon_y$ Experiment #2

$Y = X_1 + \varepsilon_y$ Experiment #3



Experiment #1

Regression Residuals:

$\varepsilon_y \sim 0$ #1

$X_1 + \varepsilon_y \sim X_1$ #2

$X_1 + \varepsilon_y \sim X_1$ #3

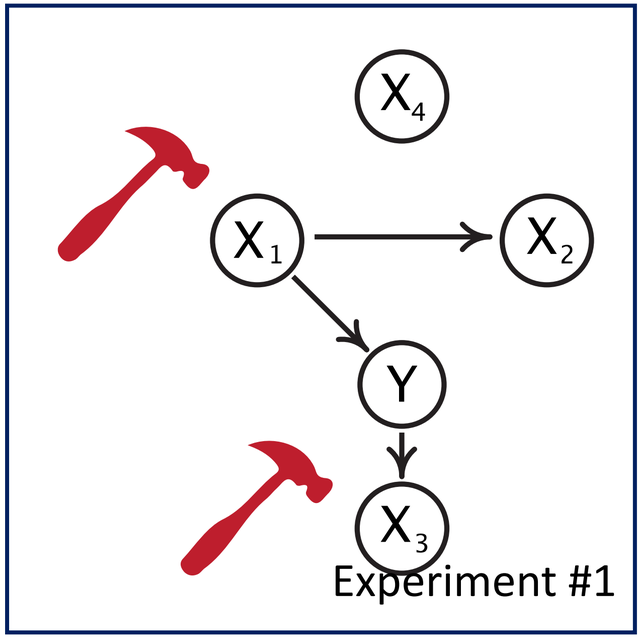
The same distribution of the residuals

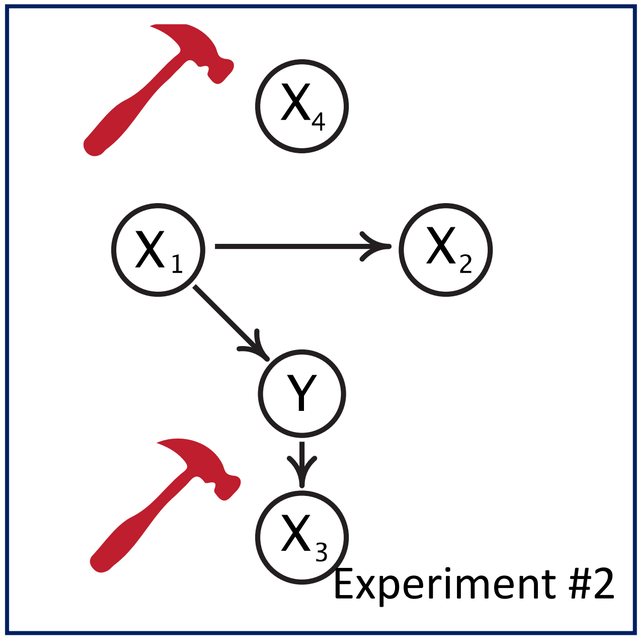
Cannot reject any experiments

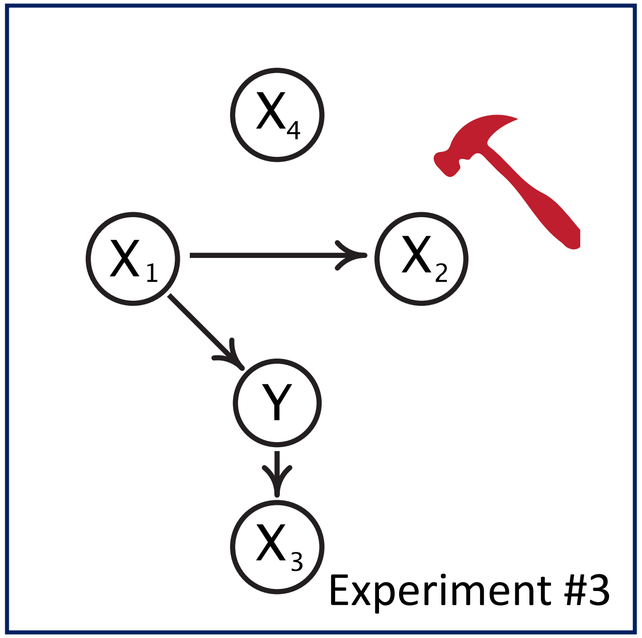
What if we specify a wrong predictor?

do(X=0)

Q. Does this model hold the invariance condition? $Y = X_1 + X_3 + \varepsilon$







Observation:

$Y = \varepsilon$

$X_1 = 0, X_3 = 0$

$Y = X_1 + \varepsilon$

$X_1 = \varepsilon_1, X_3 = 0$

$Y = X_1 + \varepsilon$

$X_1 = \varepsilon_1, X_3 = Y + \varepsilon_3$

Does this model make sense in all the experiments?

Observation:

#1 $Y = \varepsilon_y$ $X_3 = 0$

#2 $Y = X_1 + \varepsilon_y$ $X_3 = 0$

#3 $Y = X_1 + \varepsilon_y$ $X_3 = \varepsilon_3$

Regression Residuals:

$\varepsilon_y \sim 0$ #1

$X_1 + \varepsilon_y \sim X_1 + 0$ #2

$X_1 + \varepsilon_y \sim X_1 + Y + \varepsilon_3$ **#3**

$\sim X_1 + (X_1 + \varepsilon_y) + \varepsilon_3$

Invariance violated

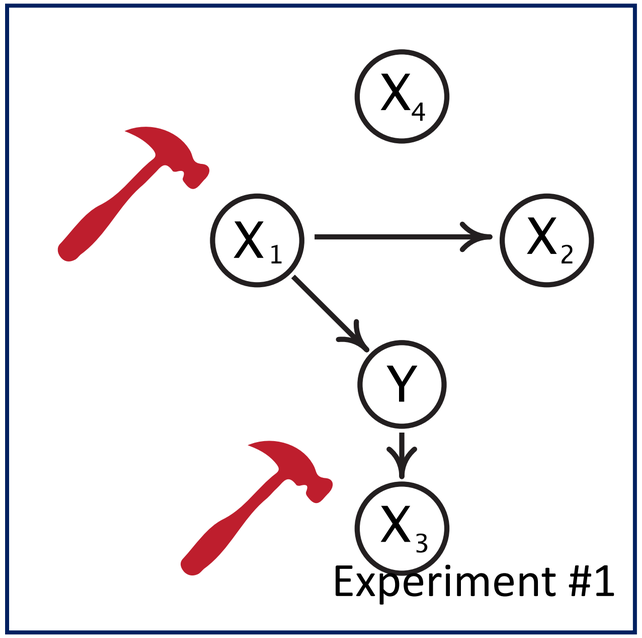
Reject the experiment #3

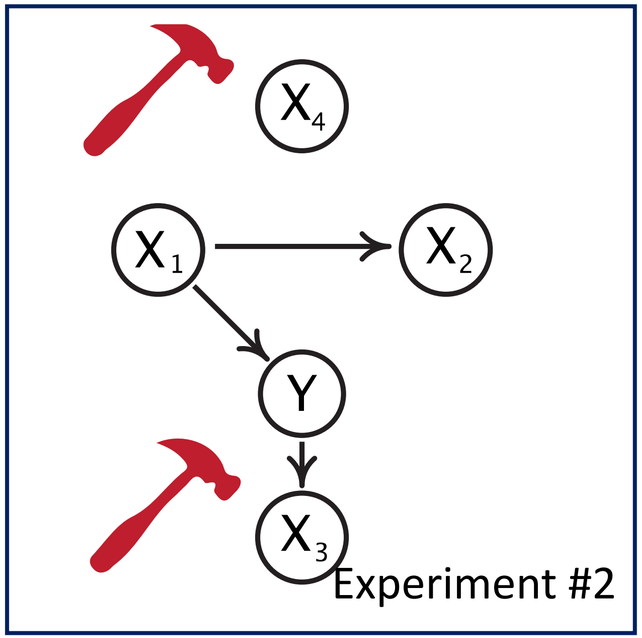
→ Reject the invariance of X_3

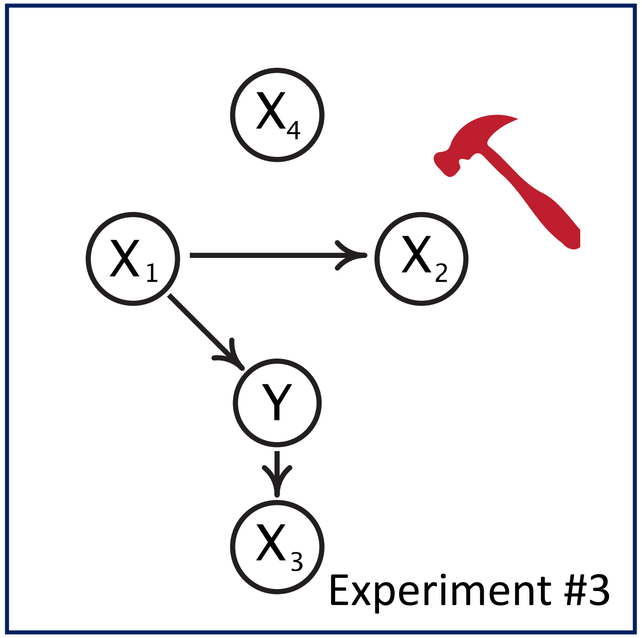
What if we specify a wrong predictor?

do(X=0)

Q. Does this model hold the invariance condition? $Y = X_2 + \varepsilon_y$







Observation:

$Y = \varepsilon$

$X_2 = 0$

$Y = X_1 + \varepsilon$

$X_2 = X_1 + \varepsilon_2$

$Y = X_1 + \varepsilon$

$X_2 = 0$

Does this model make sense in all the experiments?

Observation:

#1 $Y = \varepsilon_y$ $X_2 = \varepsilon_2$

#2 $Y = X_1 + \varepsilon_y$ $X_2 = X_1 + \varepsilon_2$

#3 $Y = X_1 + \varepsilon_y$ $X_2 = 0$

Regression Residuals:

$\varepsilon_y \sim \varepsilon_2$ #1

$\cancel{X_1} + \varepsilon_y \sim \cancel{X_1} + \varepsilon_2$ #2

$X_1 + \varepsilon_y \sim \varepsilon_1$ #3

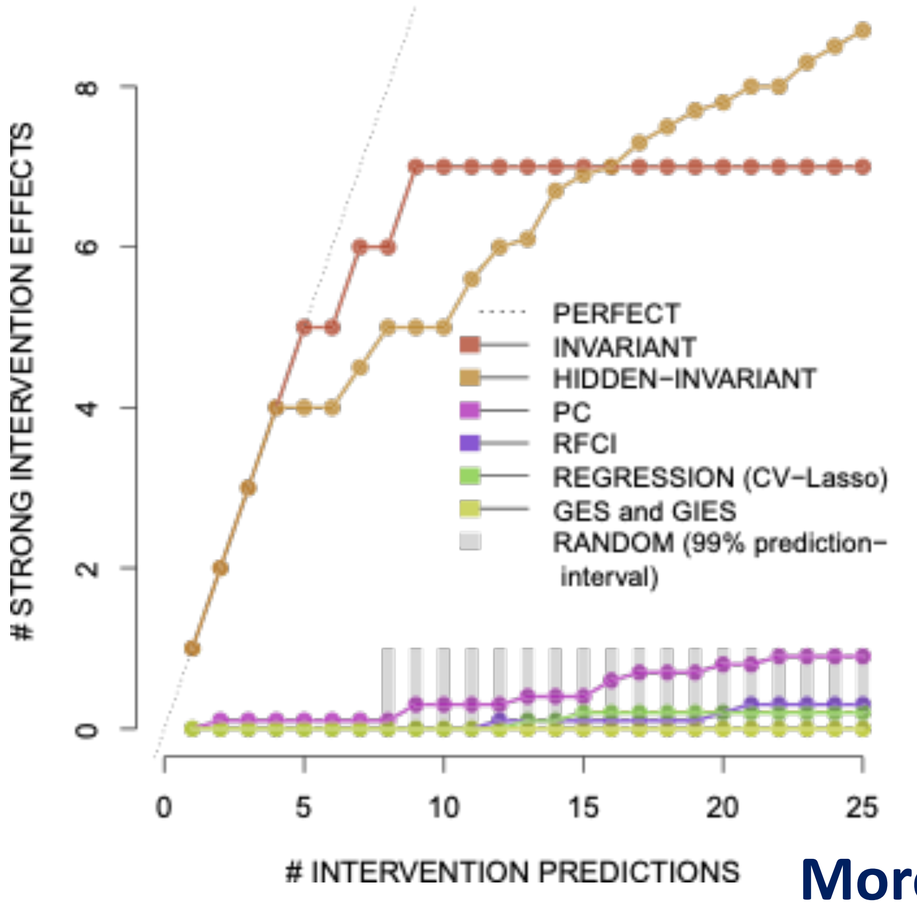
Invariance violated

Reject the experiment #3

→ Reject the invariance of X_2

Causal Model Discovery by Invariance outperforms regression-based methods

Evaluation of the actual Knock-out experiments



Causal regression

$$\mathbb{E}[y \mid do\left(X_j = x\right)]$$

Regression on observed data

$$\mathbb{E}[y \mid X = x]$$

More perturbation Exp.

Meinshausen, .. Buhlmann, *PNAS* (2016)

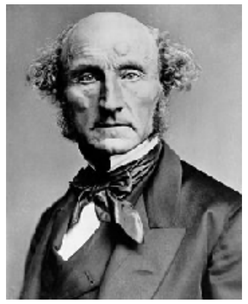
Causal triangulation to gain confidence in scientific explanations

Observation

1. The occurrence of X_1 and X_2 with Y_1 and Y_2
2. The occurrence of X_2 with Y_2

Induction (conclusion)

$X_1 \rightarrow Y_1$ or $X_1 \leftarrow Y_1$



John Stuart Mill

SECOND CANON.

If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon.

JS Mill, *A system of Logic* (1843)

Multiple lines of "orthogonal evidence"

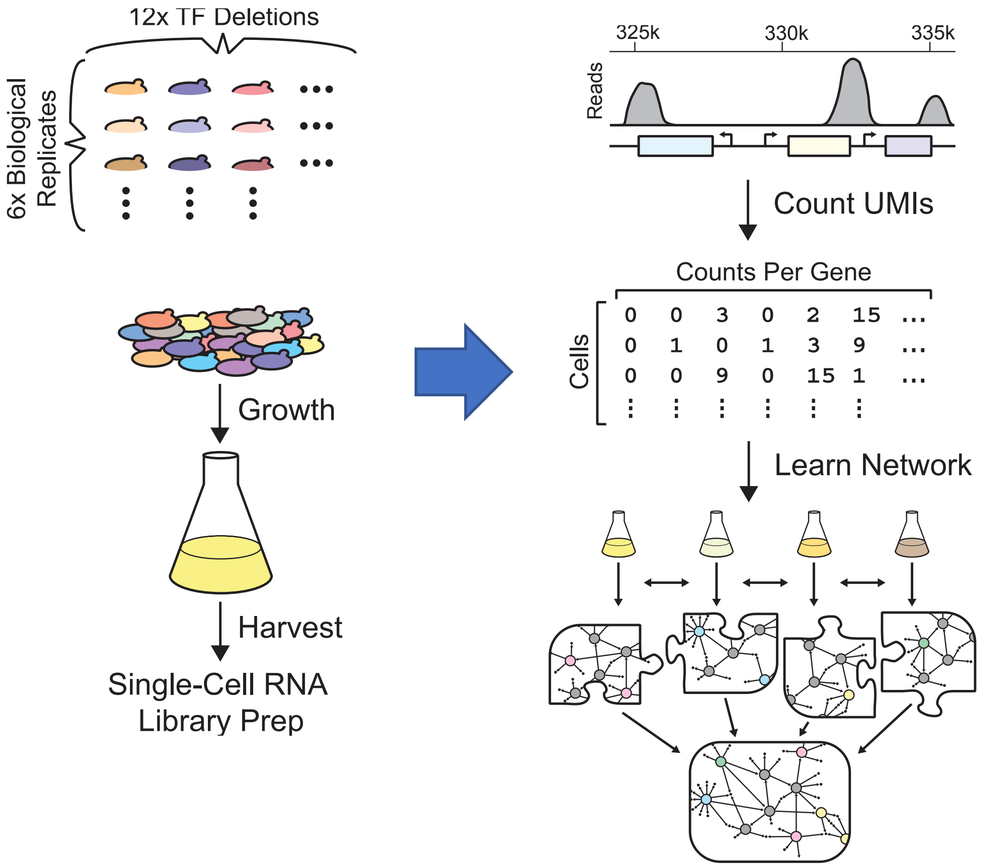
Munafo & Davey Smith, "Repeating experiments is not enough", Nature (2018)

Peter Lipton

Contrastive Explanation & causal triangulation, Philosophy of Science (1991)

George Davey Smith

Can we borrow the awesome power of genetics?



Jackson *et al.* (2020)

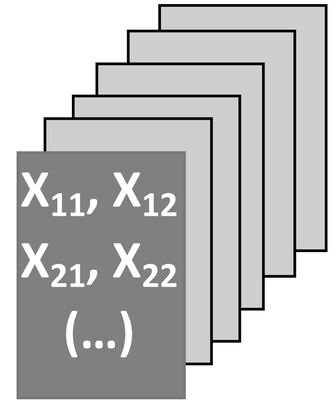
Early 2000's

X_{11}, X_{12}, X_{13}
X_{21}, X_{22}, X_{23}
(…)

A single or only a handful of conditions

Many computational biologists gave up on this NP-hard problem… learning BN from data

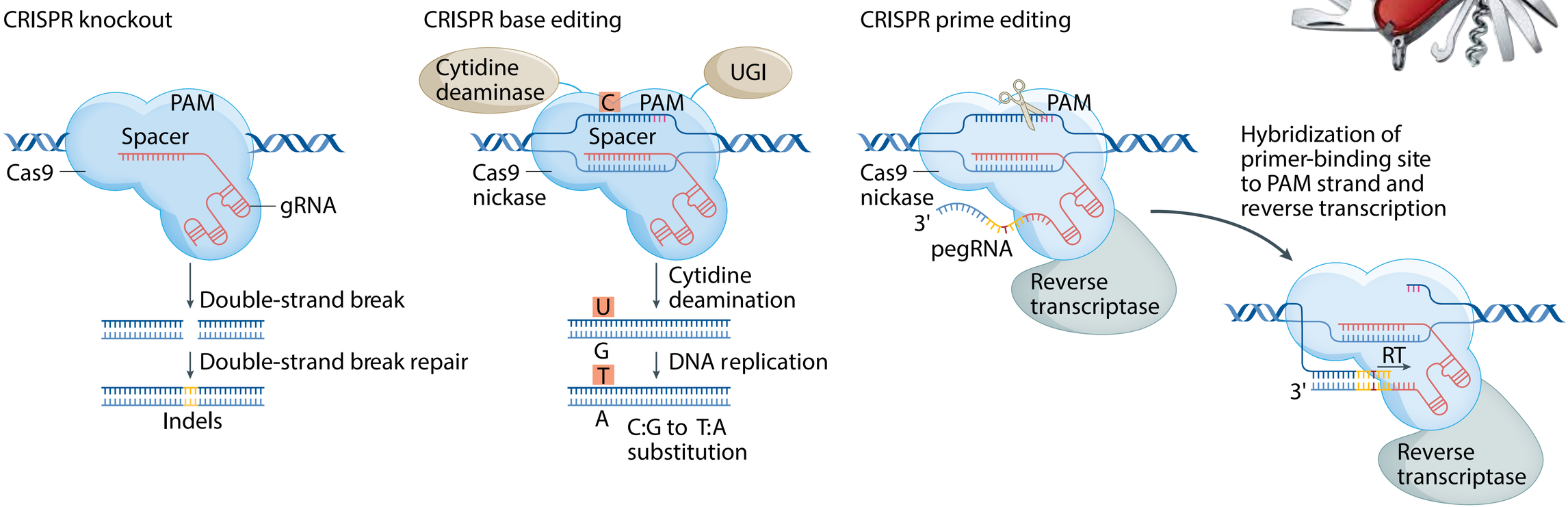
After perturb-seq



Massive (random) perturbation assays → hundreds of data matrices

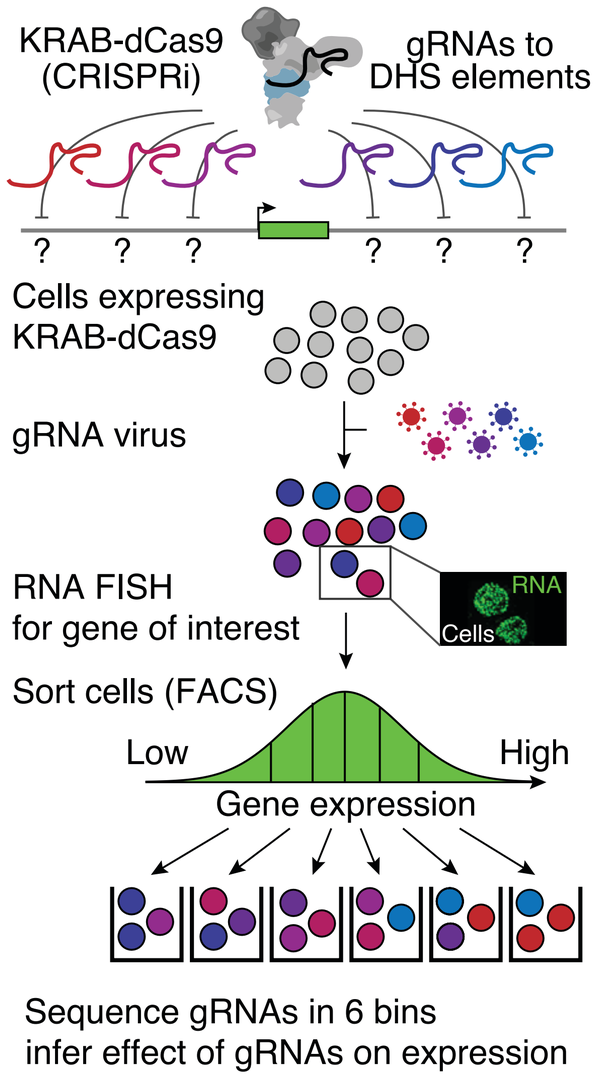
CRISPR-dCas9 system made genetic perturbation studies so easy and scalable

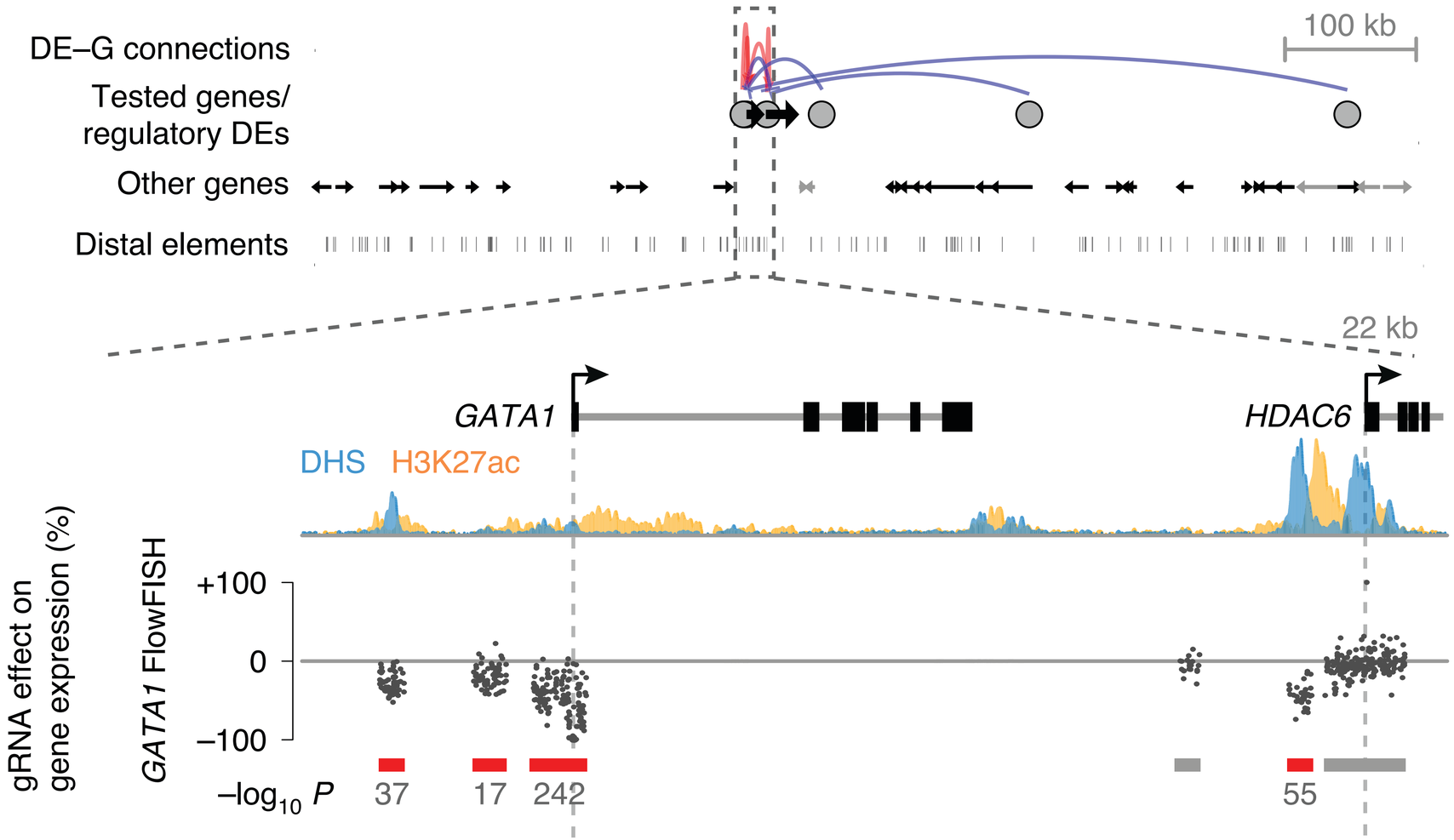
Molecular scissors for genome editing



Bock *et al.* Nature Reviews Primers (2022)

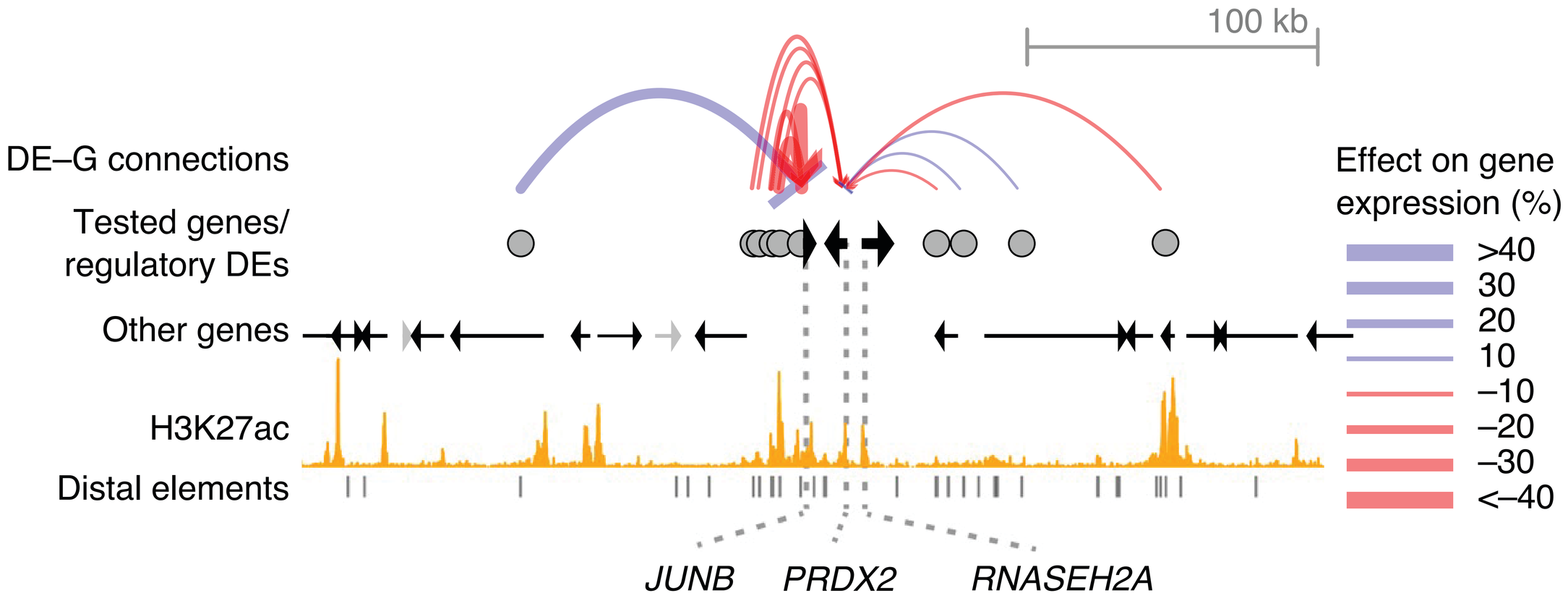
CRISPRi to disrupt gene regulatory elements of a target gene to interrogate connections between them



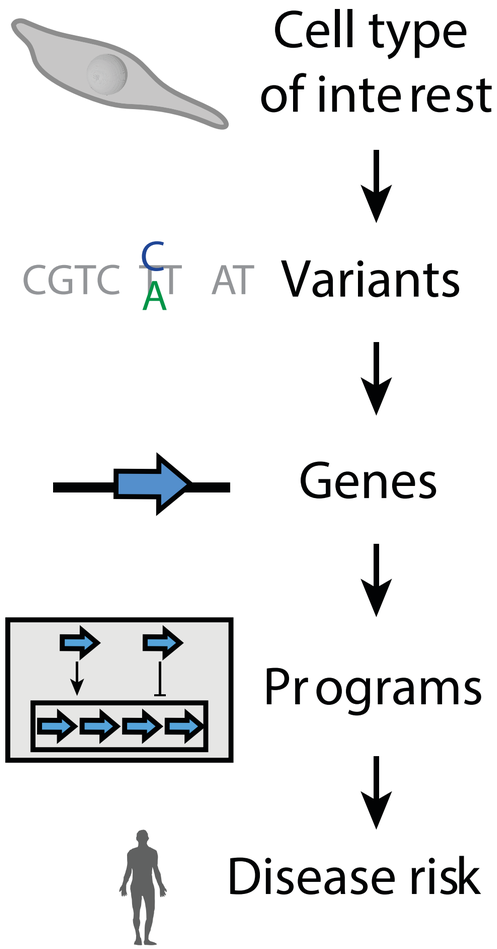


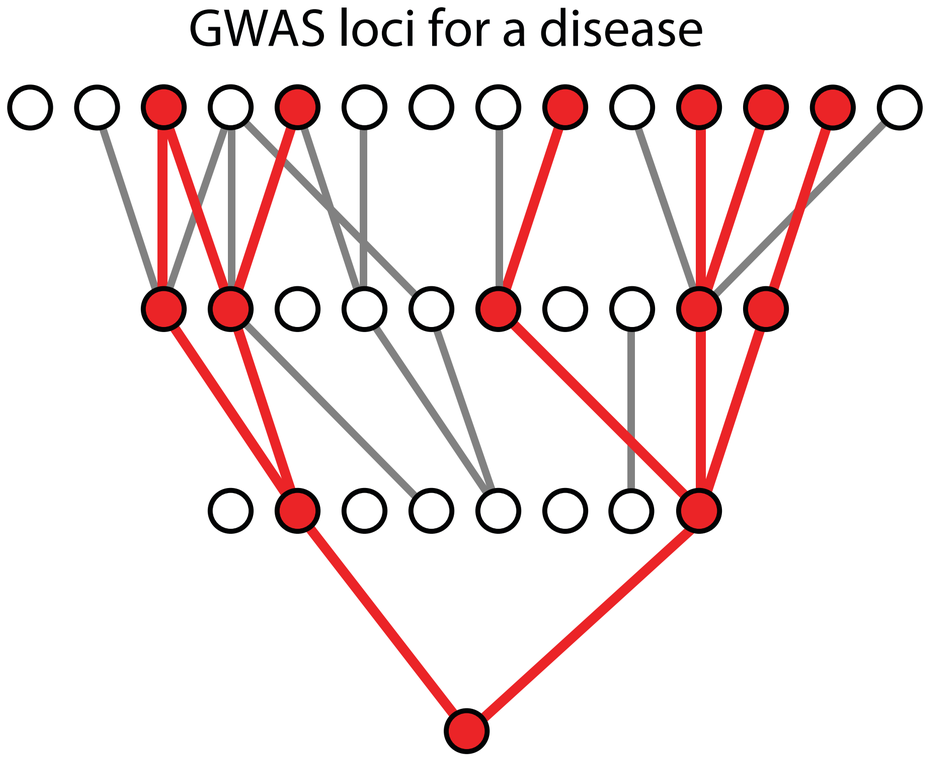
Fulco .. Engreitz *Nature Genetics* (2019)

We can systematically link epigenetic regions to target genes by experiments



GWAS variant to target gene



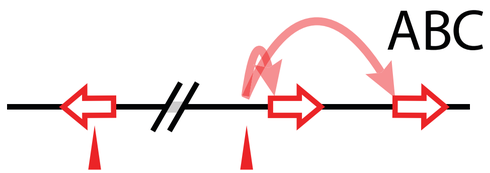


Schnitzler .. Engreitz, *Nature* (2024)

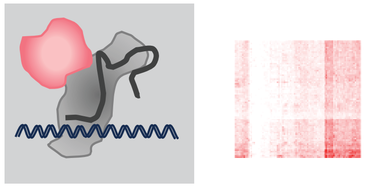
GWAS variant to gene to pathway

V2G2P enrichment test:

V2G map:

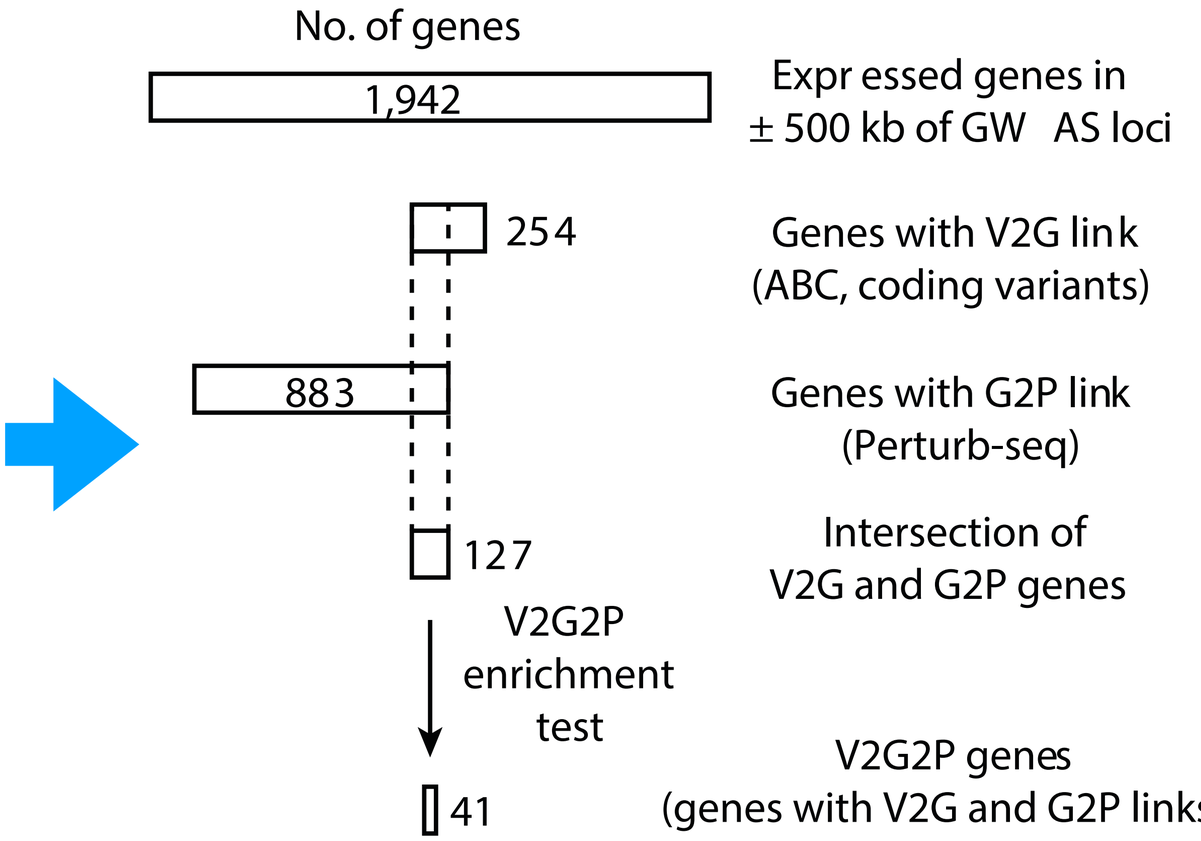


G2P map:



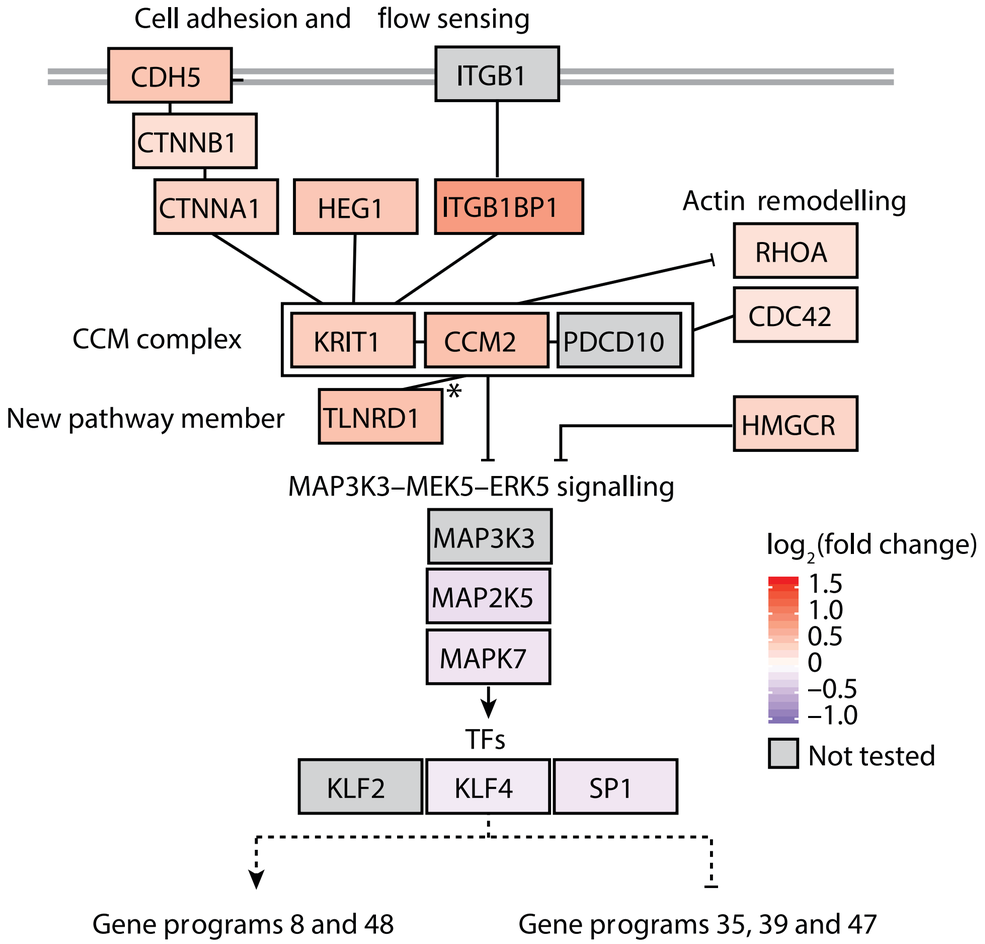
CRISPRi-Perturb-seq

Prioritize programs

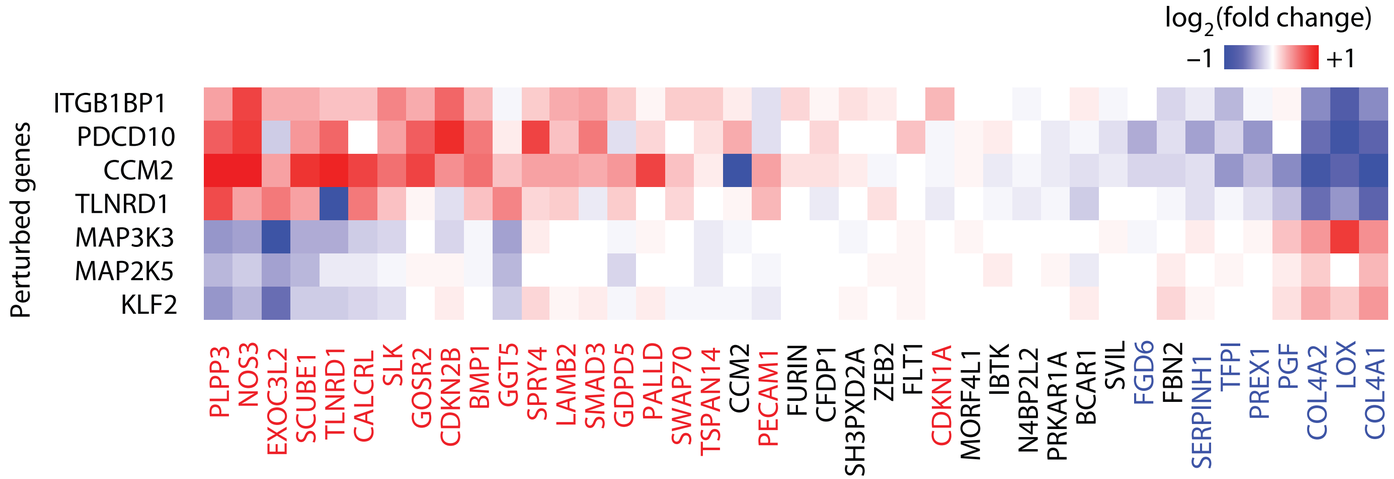


Schnitzler .. Engreitz, *Nature* (2024)

How can we prove Varaint to Gene to Pathway?



A series of Knock-down experiments



Schnitzler .. Engreitz, *Nature* (2024)

Multiple CRISPR screening can be combined

nature biotechnology

Article

https://doi.org/10.1038/s41587-023-01964-9

Scalable genetic screening for regulatory circuits using compressed Perturb-seq

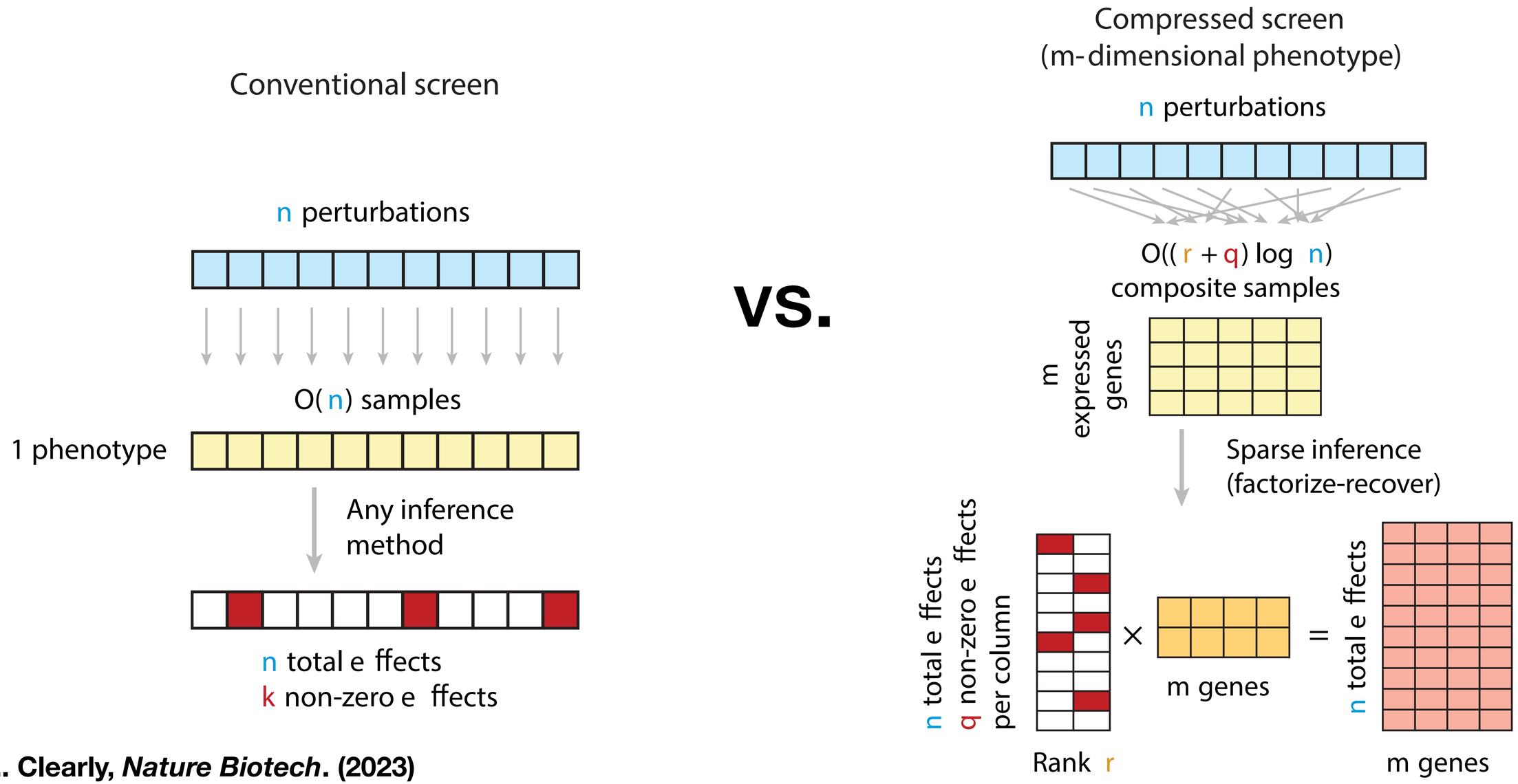
Received: 5 January 2023

Accepted: 22 August 2023

Published online: 23 October 2023

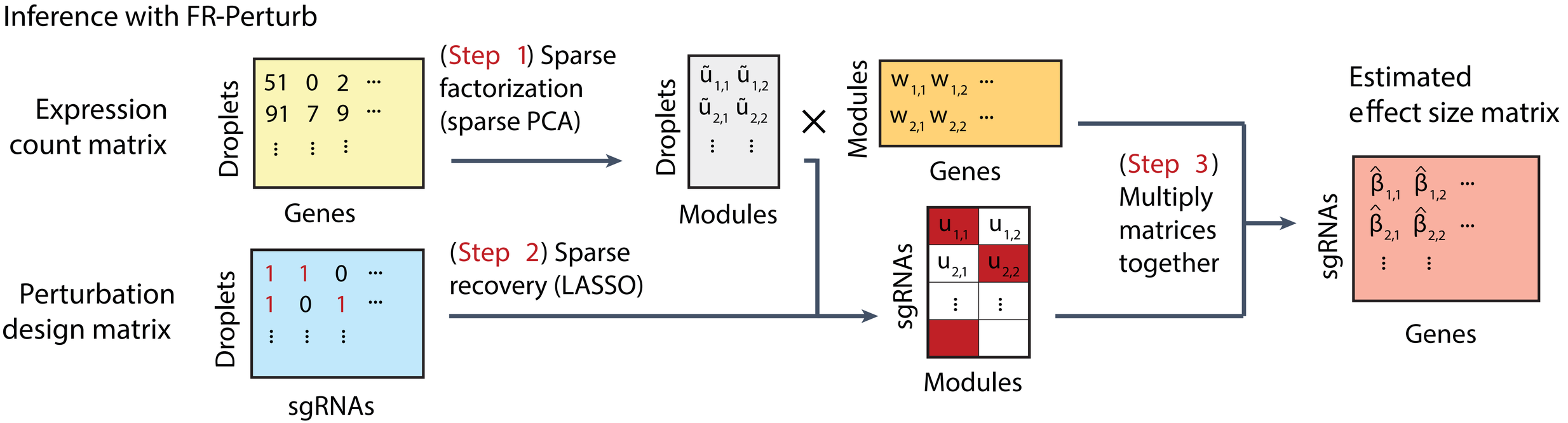
Douglas Yao[1], Loic Binan[2], Jon Bezney[2,13], Brooke Simonton[2], Jahanara Freedman[2], Chris J. Frangieh[2,3], Kushal Dey[4,14], Kathryn Geiger-Schuller[5], Basak Eraslan[5], Alexander Gusev[2,6,7,16], Aviv Regev[2,14,15] & Brian Cleary[8,9,10,11,12,16]

Multiple CRISPR screening can be combined



Yao .. Clearly, *Nature Biotech*. (2023)

Compressive sensing (sparse modelling) improves CRISPR experimental design



Yao .. Clearly, *Nature Biotech*. (2023)

Today's lecture: Multiomics data integration

- **Why do we do multiomics data integration?**
  - view #1: borrowing information across modalities
  - view #2: efforts to provide mechanistic explanations
- **Global, unsupervised multiomics data integration**
  - Multiomics Factorization (and variants)
  - Network-based data integration
- **Local, linking between layers to understand mechanisms**
  - Deep dive into mechanisms of gene regulatory mechanisms