# Integrative Regression and Factorization of Multi-Omics Multi-Cohort Data

Eric F. Lock

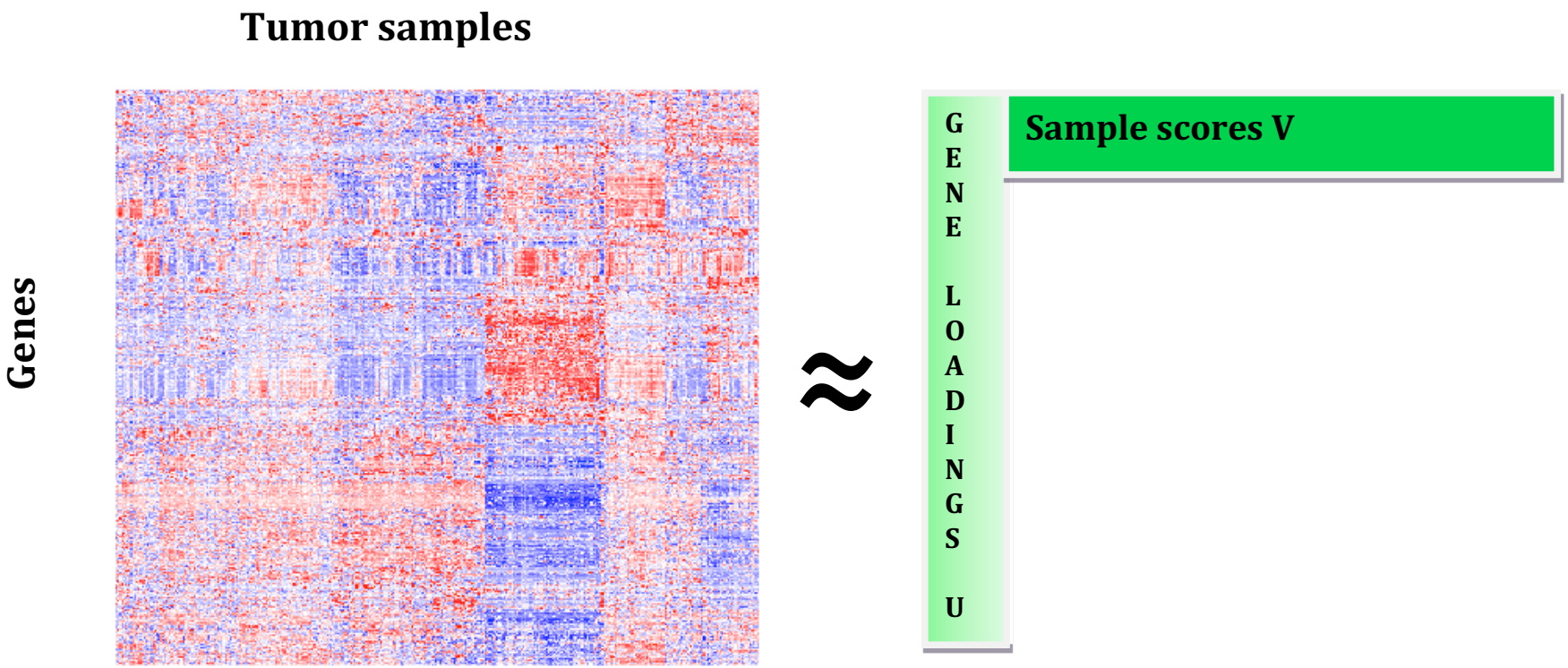
University of Minnesota, Division of Biostatistics

Collaborators:
**Jiuzhou Wang**, UMN
**Sarah Samorodnitsky**, UMN
**Jun Young Park**, University of Toronto
and **Katie Hoadley**, University of North Carolina

JSM 2022, 08/10/2022

# Matrix factorization

- Gene expression matrix $X : m \times n$
  - $m$ genes for $n$ breast cancer tumor samples




- Low rank factorization: $X \approx UV$, $U : m \times r$, $V : r \times n$.

# Matrix factorization: Nuclear norm

- Singular value decomposition (SVD): $X = UDV^T$
  - D is diagonal with singular values $D[i,i] = d_i$

- Minimize
$$\frac{1}{2}||X - \hat{X}||_F^2 + \lambda||\hat{X}||_*$$
where $||\cdot||_*$ defines the nuclear norm
$$\text{SVD}(\hat{X}) = \hat{U}\hat{D}\hat{V}^T \rightarrow ||\hat{X}||_* = \sum_{i=1}^{\min\{m,n\}} \hat{d}_i$$

- Then $\hat{X} = U\hat{D}V^T$ where $\hat{d}_i = max(d_i - \lambda, 0)$.

# Matrix factorization: Nuclear norm

- Consider $X = \mathbf{A} + E$ where rank($\mathbf{A}$)=r and $E \overset{indep}{\sim} N(0,1)$
- SVD $X = UDV$ where

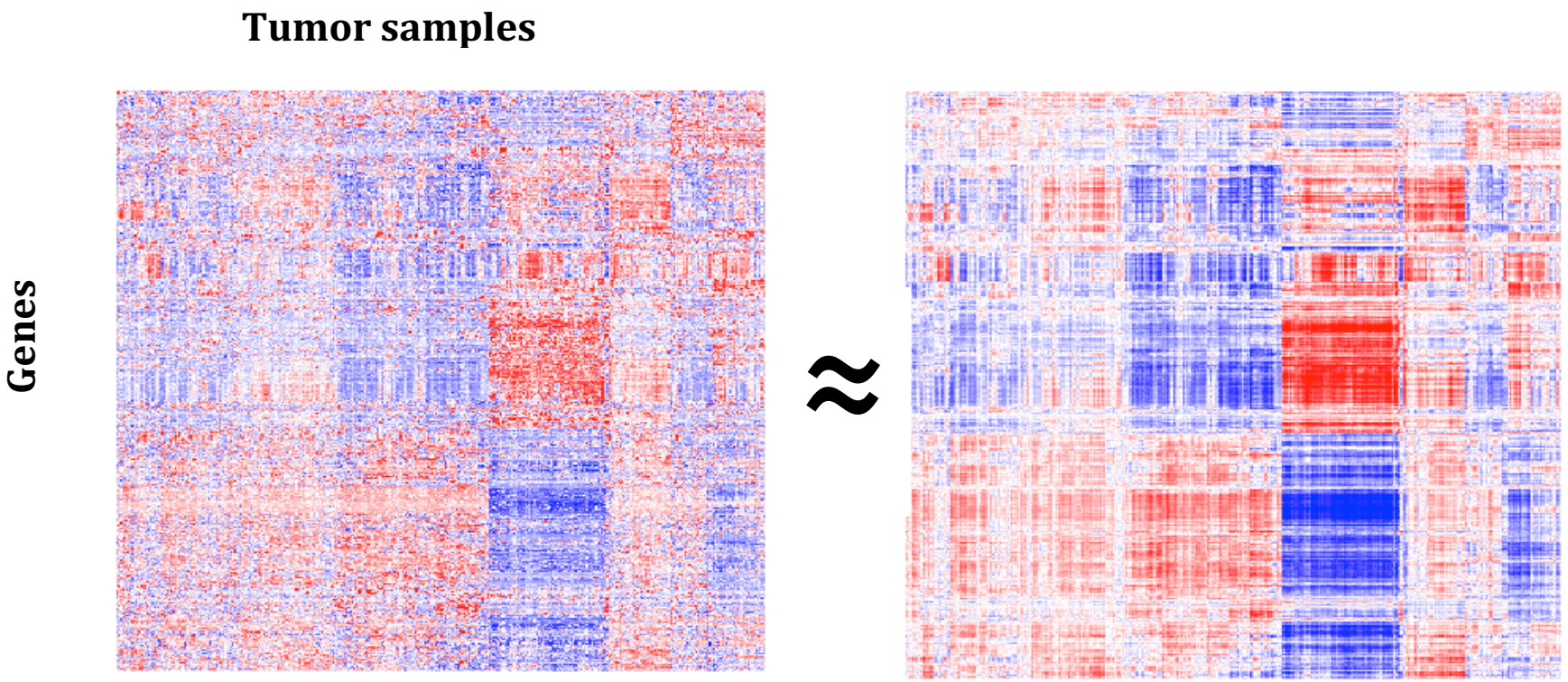
$$U = [\mathbf{u_1}, \cdots, \mathbf{u_r}, u_{r+1}, \cdots]$$
$$D = \text{diag}(\mathbf{a_1} + e_1, \cdots, \mathbf{a_r} + e_r, e_{r+1}, \cdots)$$
$$V = [\mathbf{v_1}, \cdots, \mathbf{v_r}, v_{r+1}, \cdots]$$

- The largest singular value of $E \approx \sqrt{m} + \sqrt{n}$
- Standardize $X$ to have error variance $\approx 1$ and set

$$\lambda = \sqrt{m} + \sqrt{n}$$

# Matrix factorization

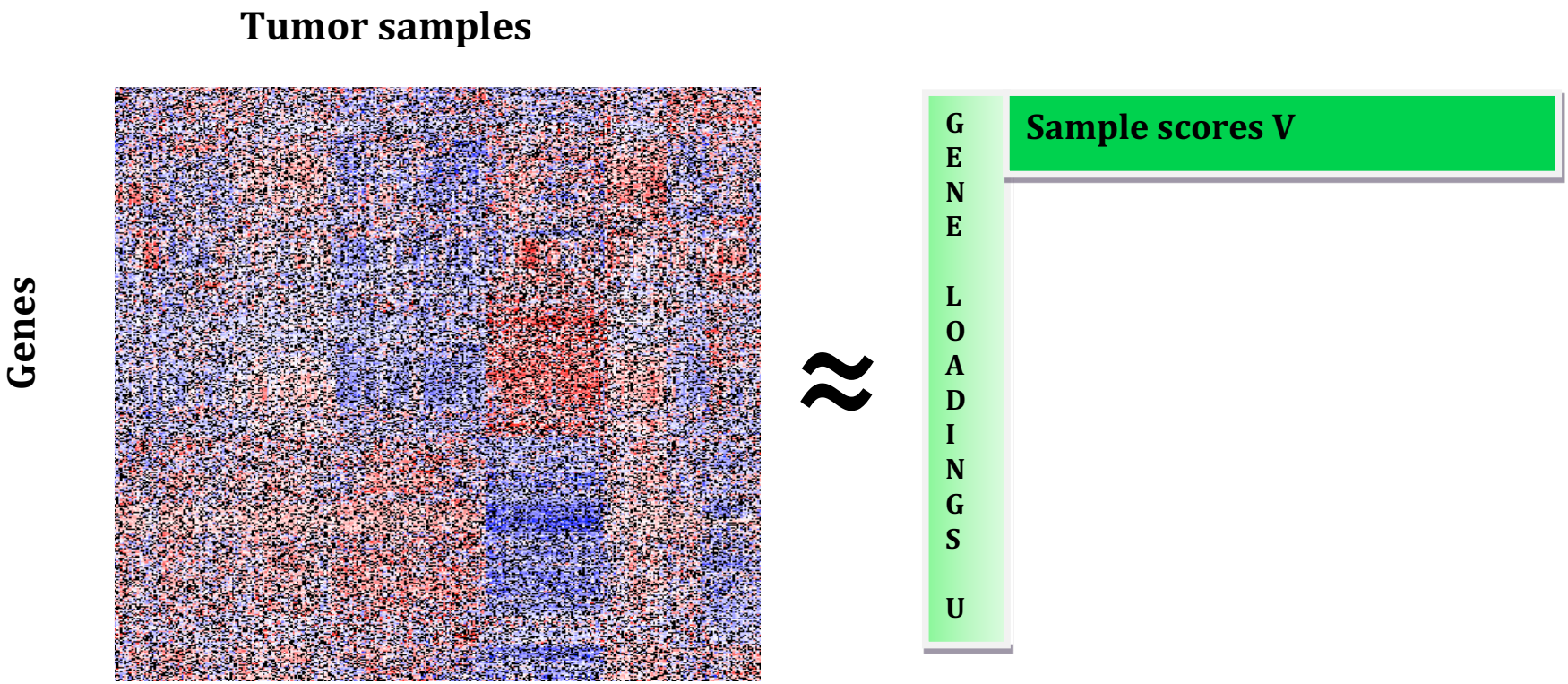
- Gene expression matrix $X : m \times n$
  - $m$ genes for $n$ breast cancer tumor samples




- Rank 18 nuclear norm approximation.

# Matrix factorization: missing data

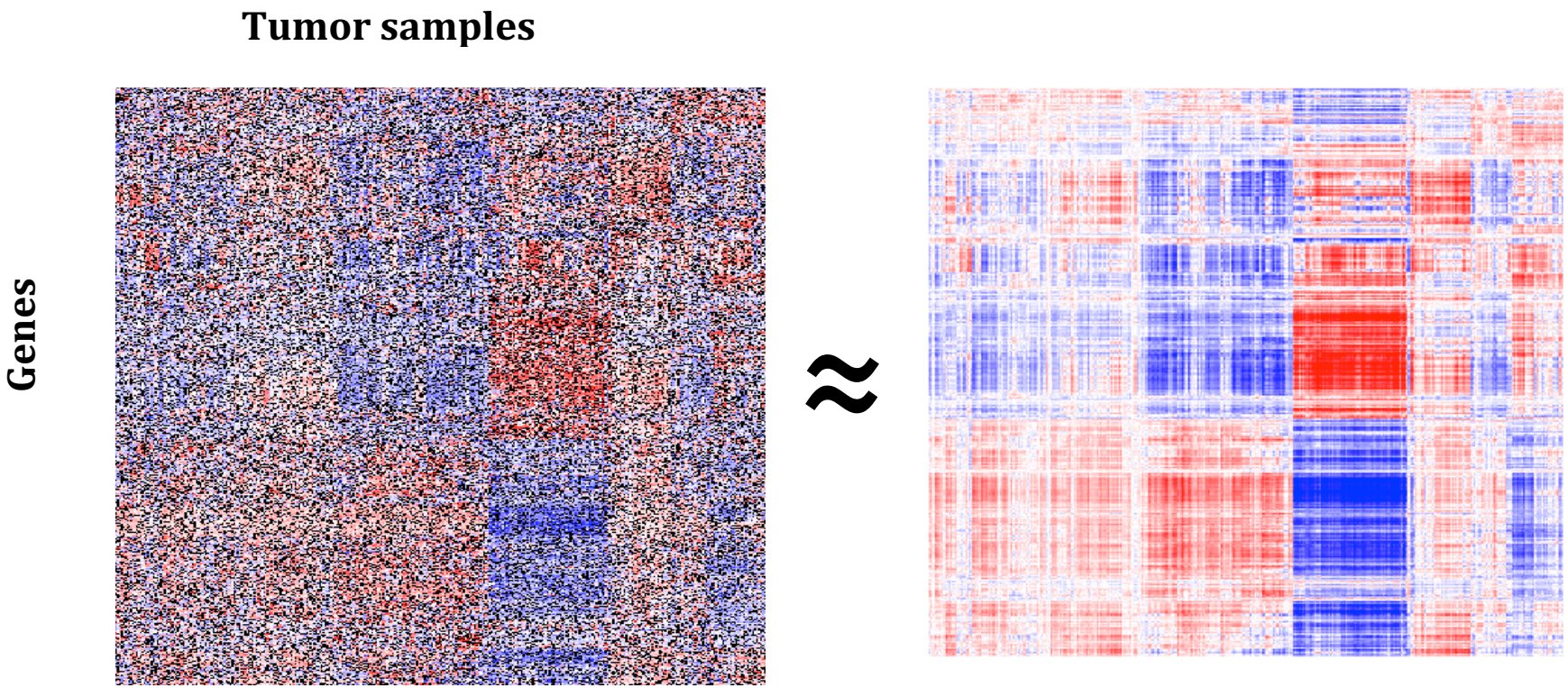
- Gene expression matrix $X : m \times n$
  - $m$ genes for $n$ breast cancer tumor samples




- Minimize $\frac{1}{2}||X[\text{observed}] - \hat{X}[\text{observed}]||_F^2 + \lambda||\hat{X}||_*$

# Matrix factorization: missing data

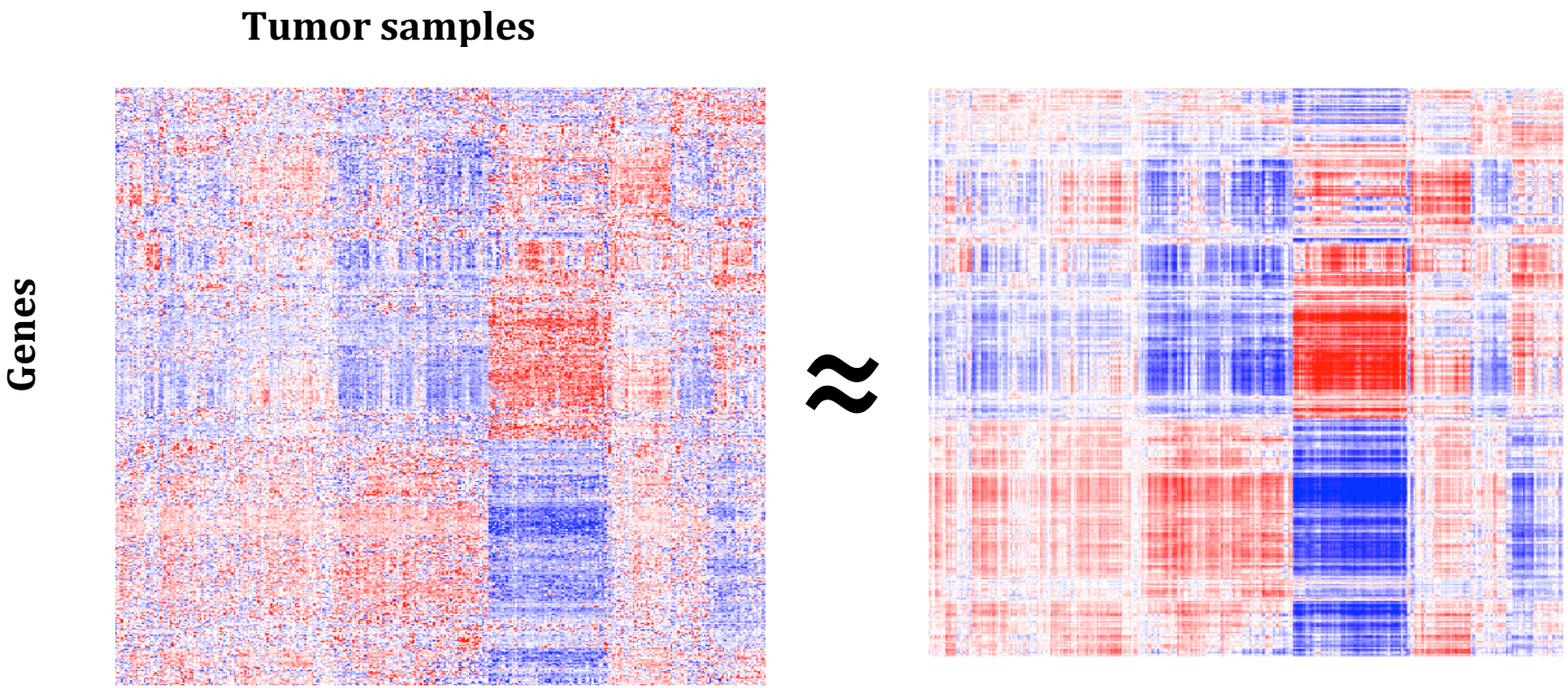
- Gene expression matrix $X : m \times n$
  - $m$ genes for $n$ breast cancer tumor samples




- Minimize $\frac{1}{2}||X[\text{observed}] - \hat{X}[\text{observed}]||_F^2 + \lambda||\hat{X}||_*$

# Matrix factorization: missing data

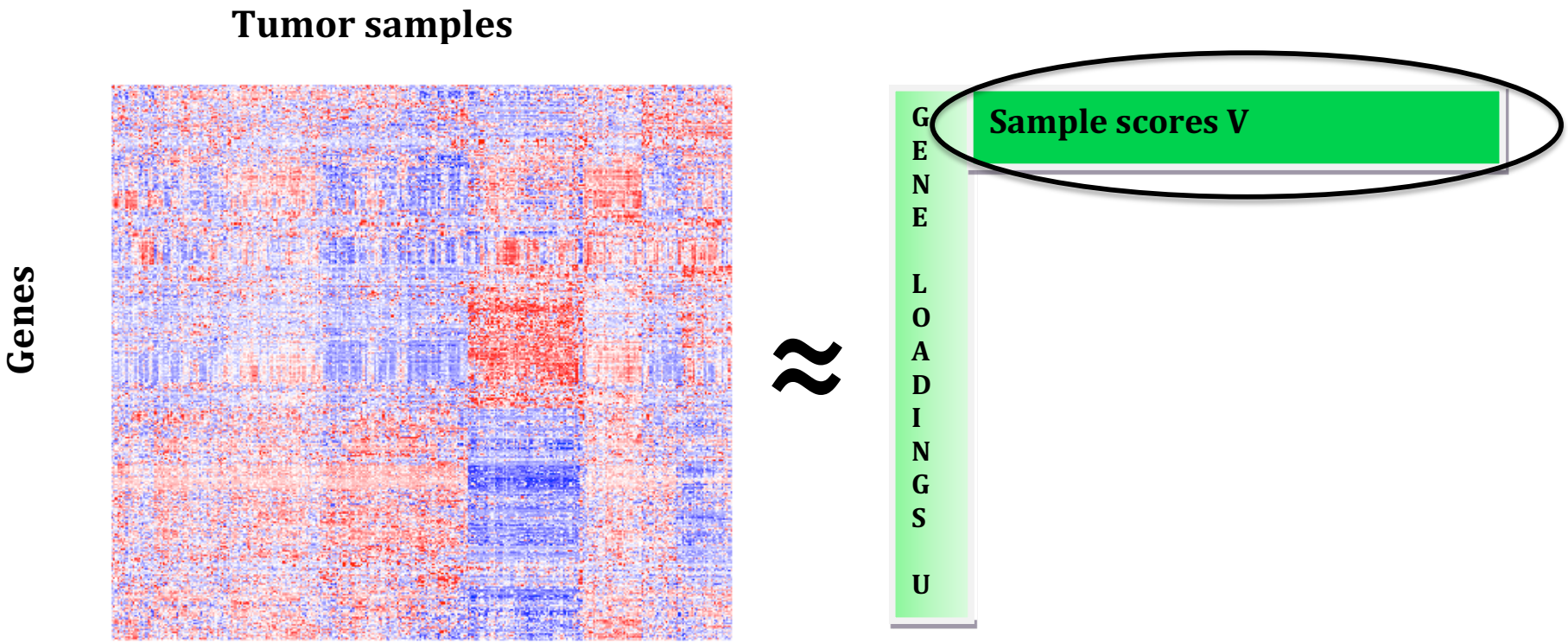
- Gene expression matrix $X : m \times n$
  - $m$ genes for $n$ breast cancer tumor samples




- Minimize $\frac{1}{2}||X[\text{observed}] - \hat{X}[\text{observed}]||_F^2 + \lambda||\hat{X}||_*$

# Matrix factorization

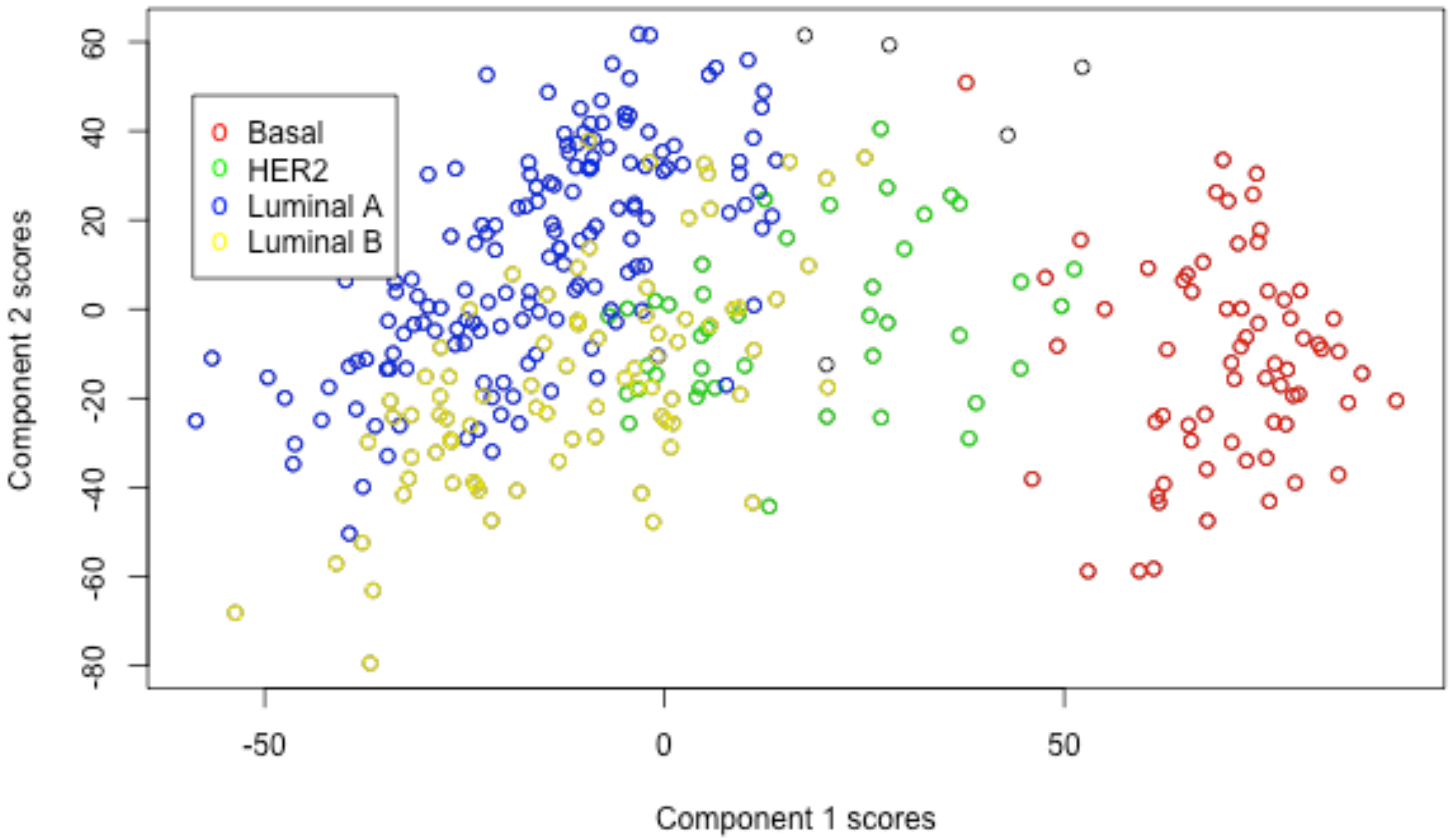
- Gene expression matrix $X : m \times n$
  - $m$ genes for $n$ breast cancer tumor samples



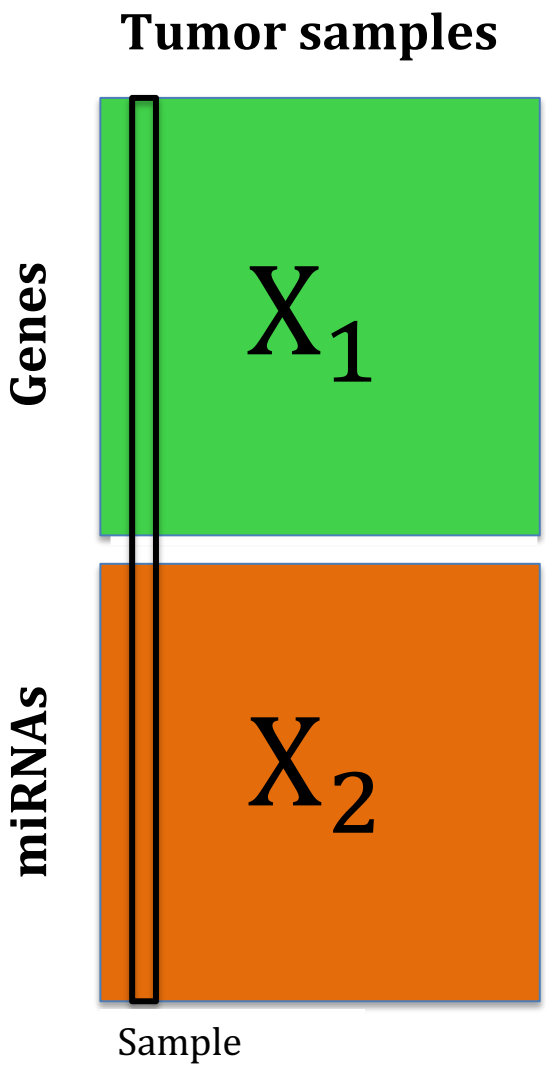

- Low rank factorization: $X \approx UV$, $U : m \times r$, $V : r \times n$.

# Matrix factorization

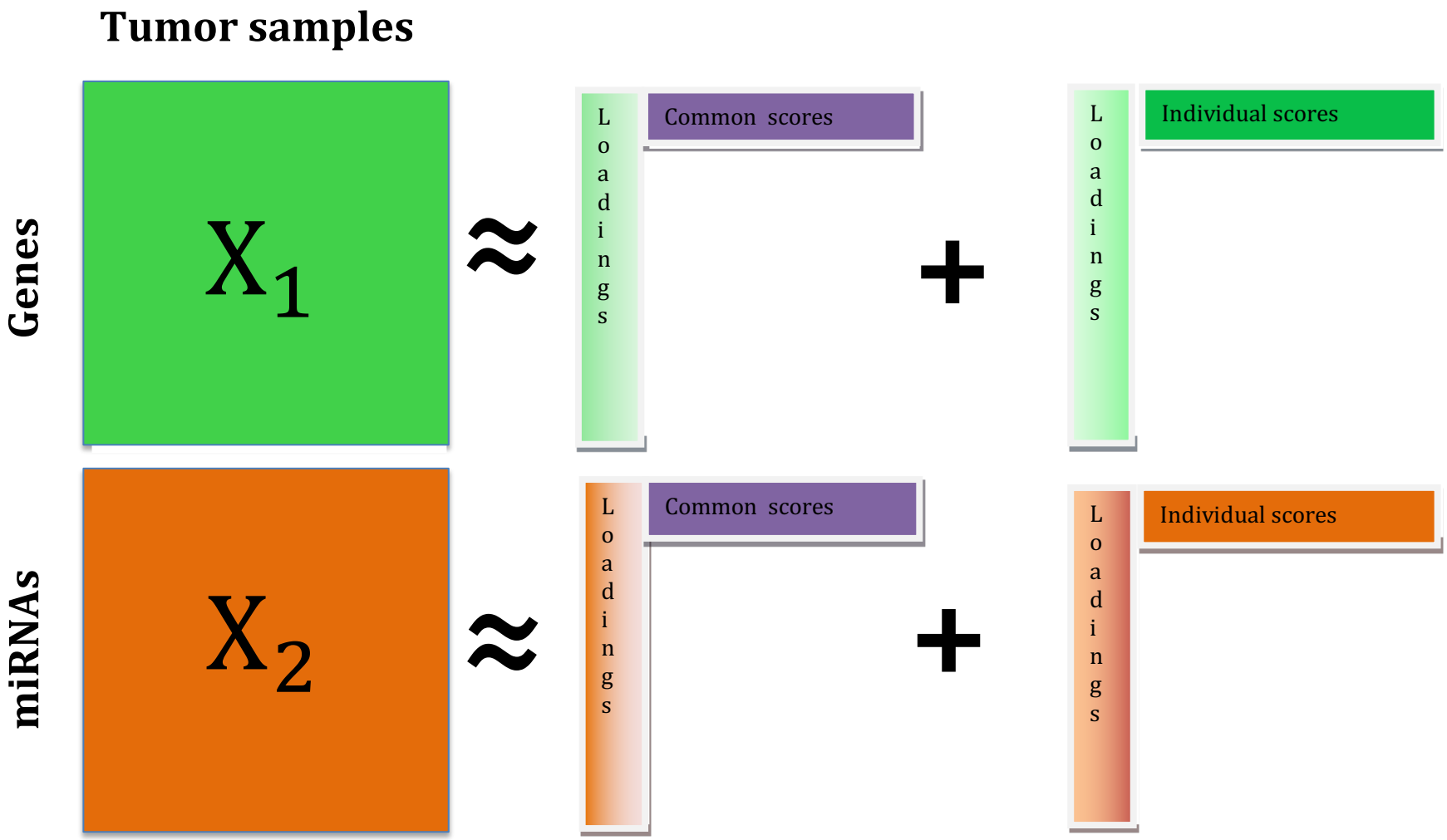
- First two principal component scores
  - Colored by breast tumor subtype



# Vertically linked data



# Vertically linked data: joint and individual factorization



# Joint + individual factorization methods

- JIVE [Lock et al., 2013]
  - "Joint and Individual Variation Explained"
- DISCO-SCA [Van Deun et al.,, 2013]
- AJIVE [Feng, Jiang, Hannig and Marron, 2018]
- SLIDE [Gaynanova and Li, 2018]
- GIPCA [Zhu, Li, Lock, 2020]

⋮

- The bi-factor method [Holzinger and Swineford, 1937]

# Structured nuclear norm penalty ("UNIFAC")

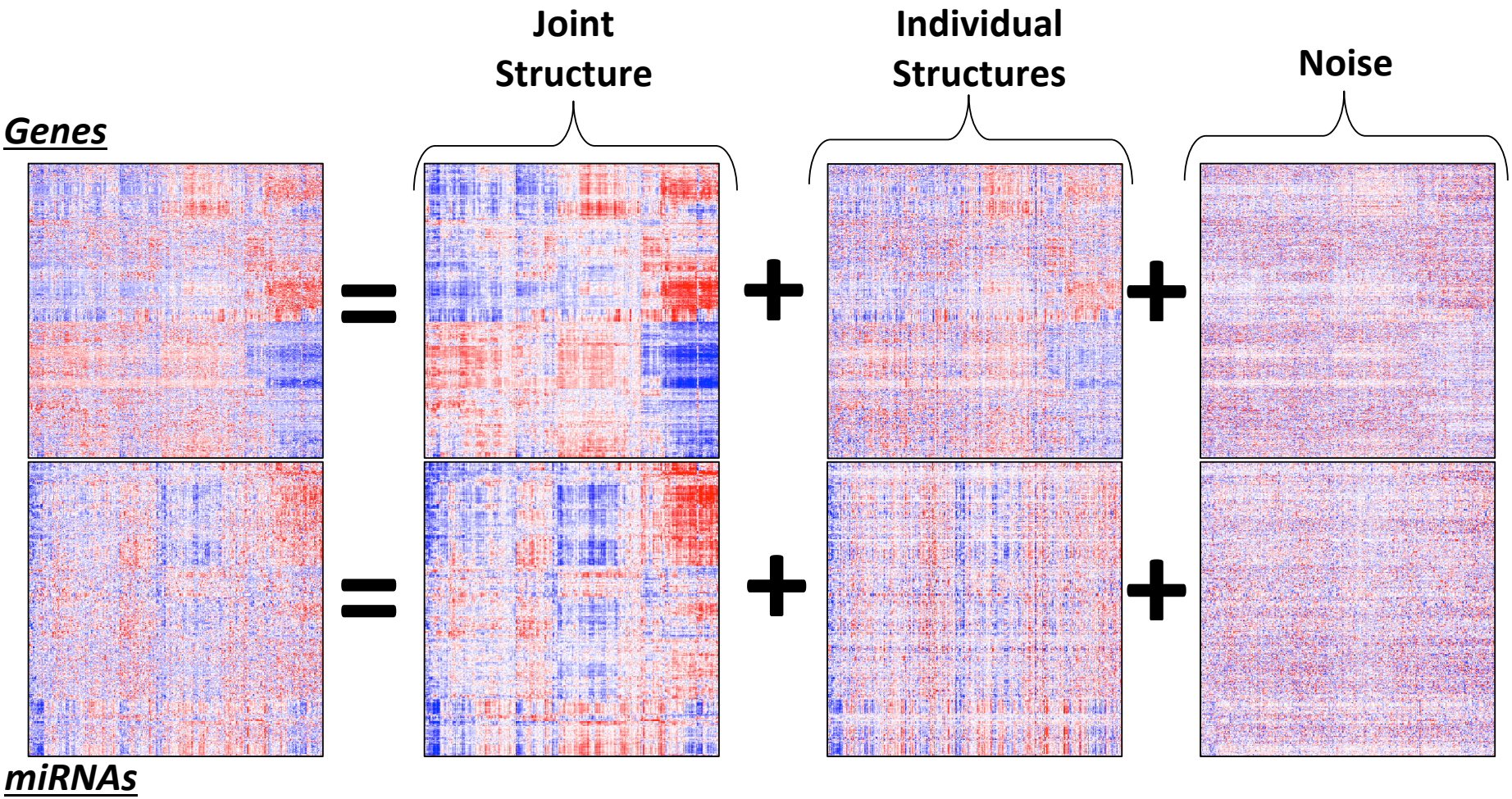
- $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ where $X_1 : m_1 \times n$, $X_2 : m_2 \times n$
- $X \approx J + A$ where $J = \begin{bmatrix} J_1 \\ J_2 \end{bmatrix}$ and $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$
- Minimize

$$\frac{1}{2}||X - J - A||_F^2 + \lambda_0||J||_* + \lambda_1||A_1||_* + \lambda_2||A_2||_*$$

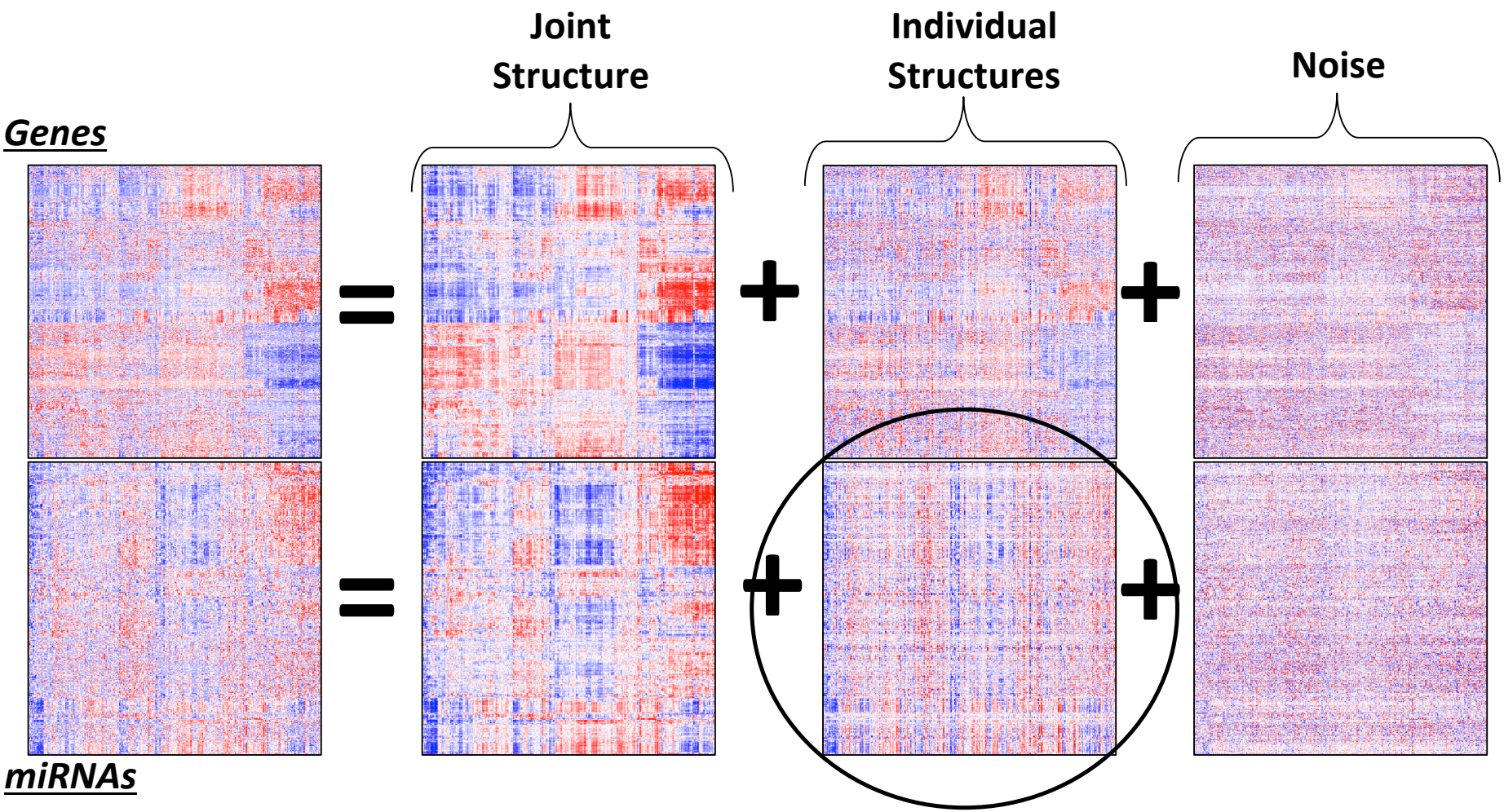
  where $\lambda_0 = \sqrt{n} + \sqrt{m_1 + m_2}, \quad \lambda_i = \sqrt{n} + \sqrt{m_i}$

- Update $J$, $A_1$, $A_2$ until convergence

# UNIFAC Estimates

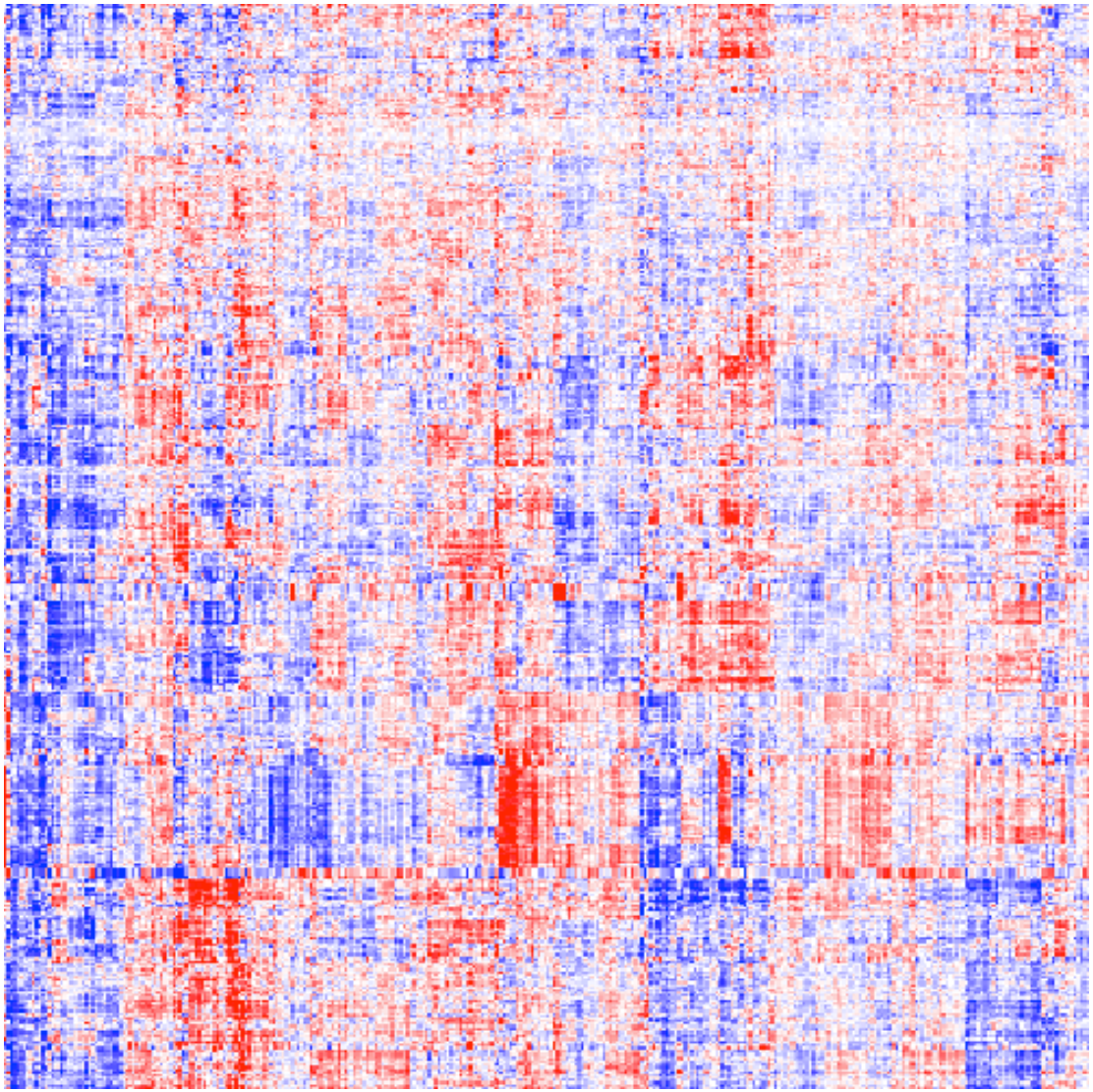


# UNIFAC Estimates

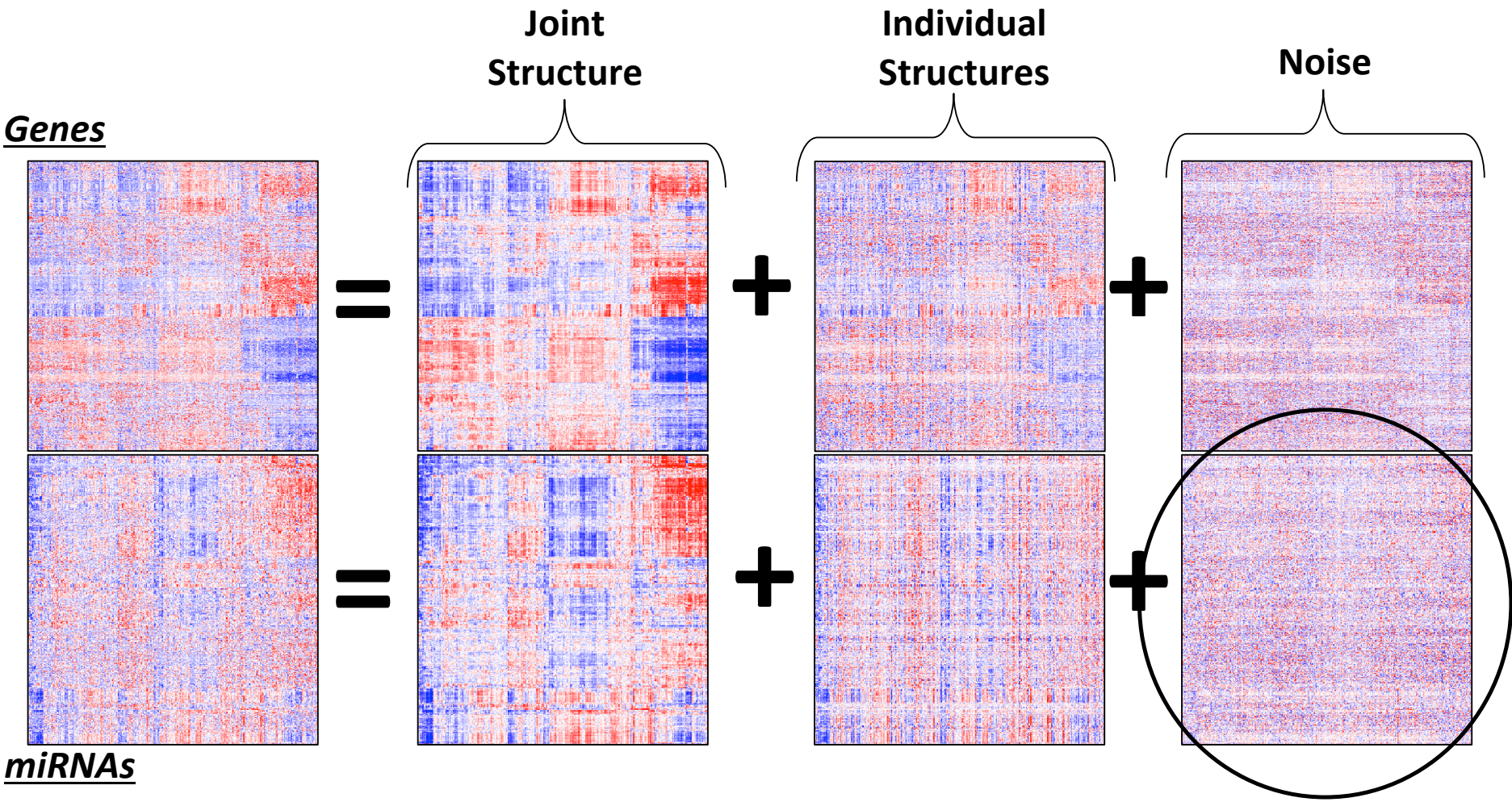


# UNIFAC Estimates

- miRNA individual (reorder rows and columns)

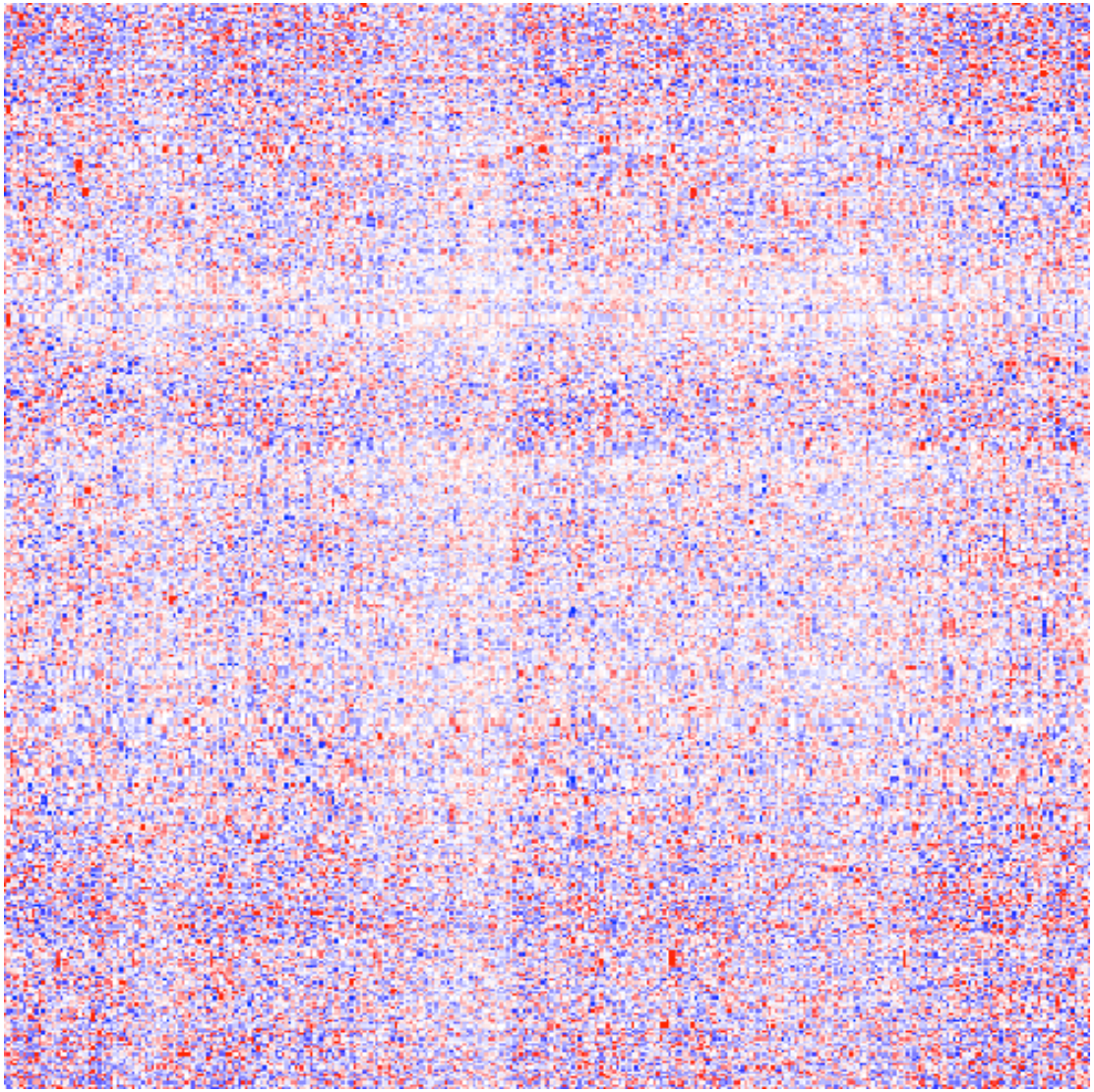


# UNIFAC Estimates

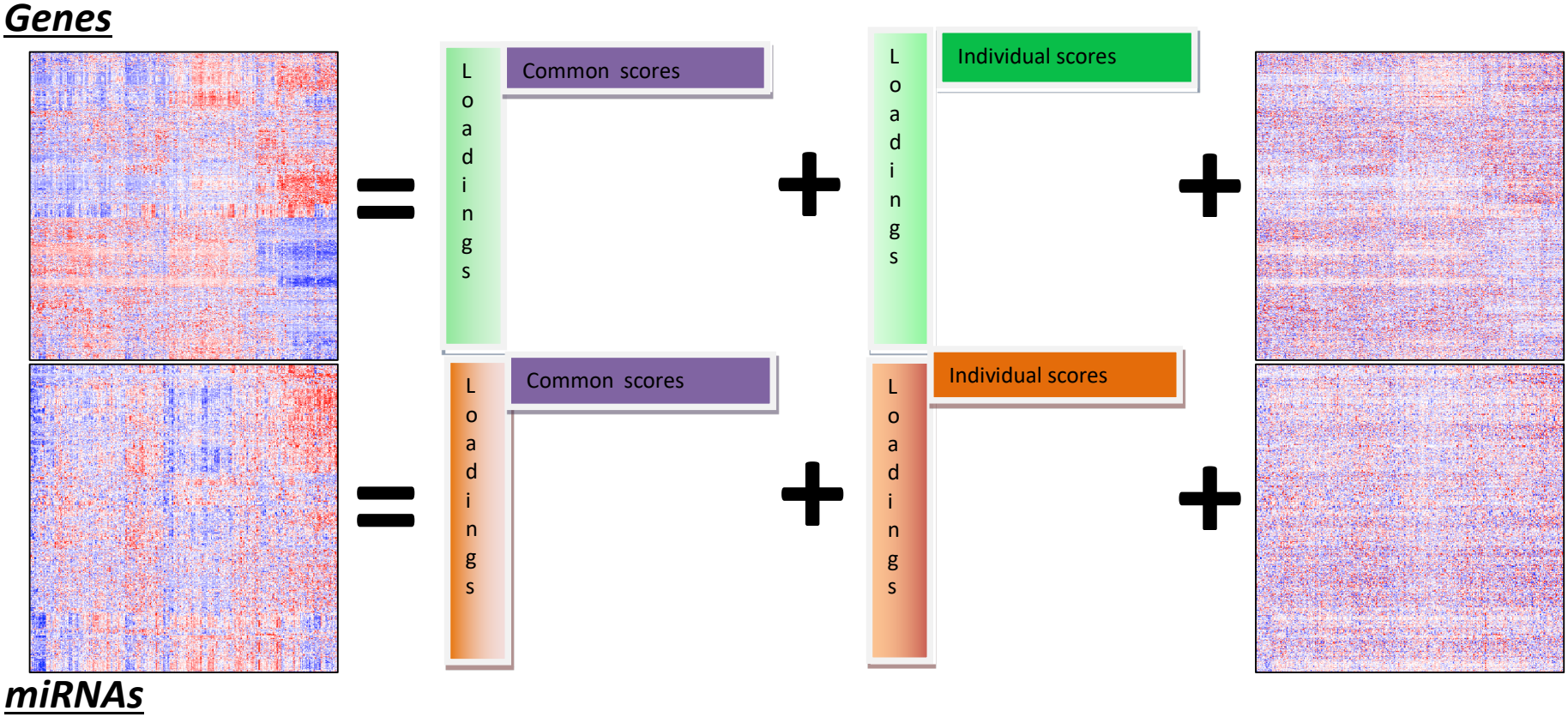


# UNIFAC Estimates

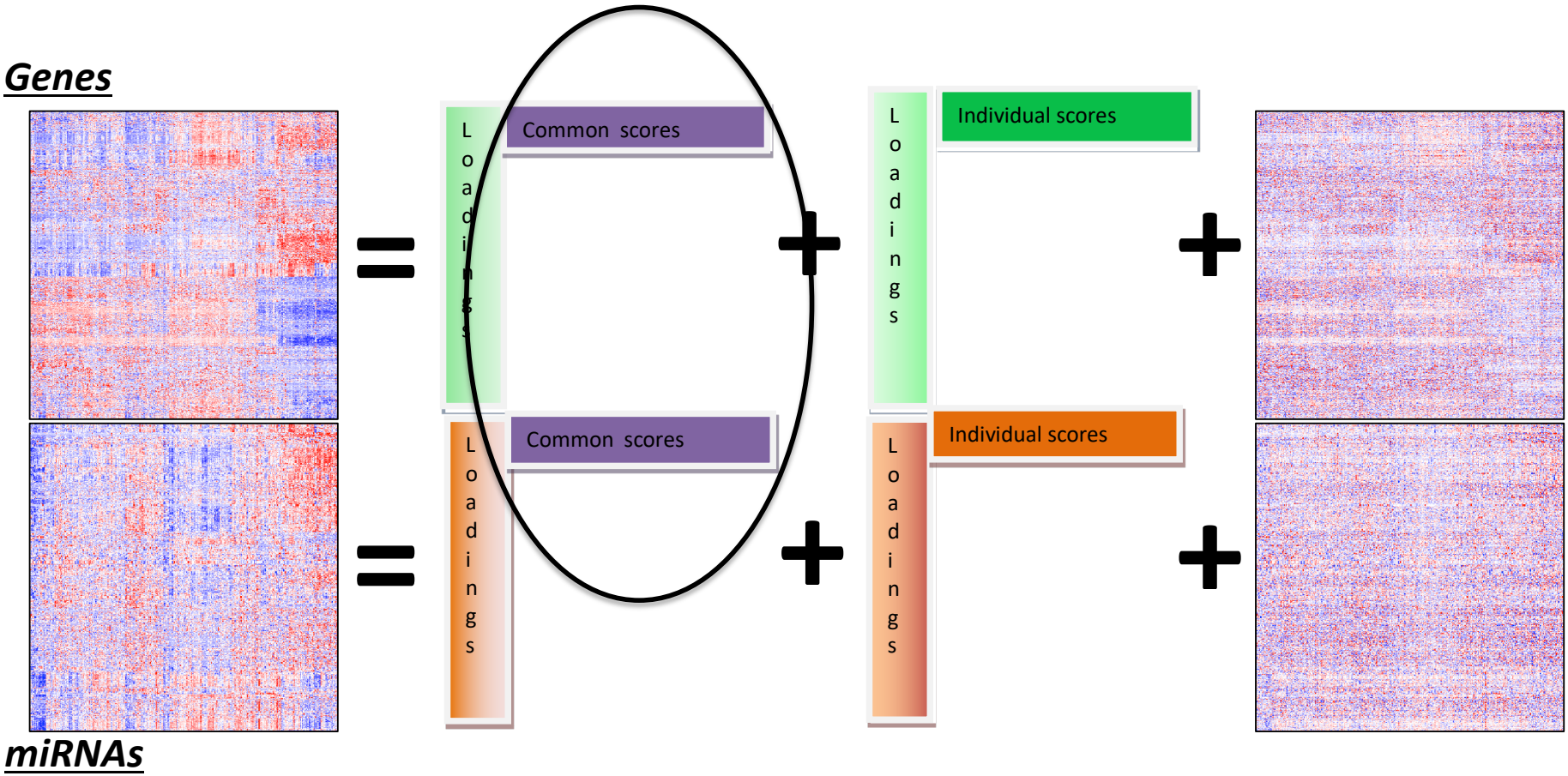
- miRNA error (reorder rows and columns)



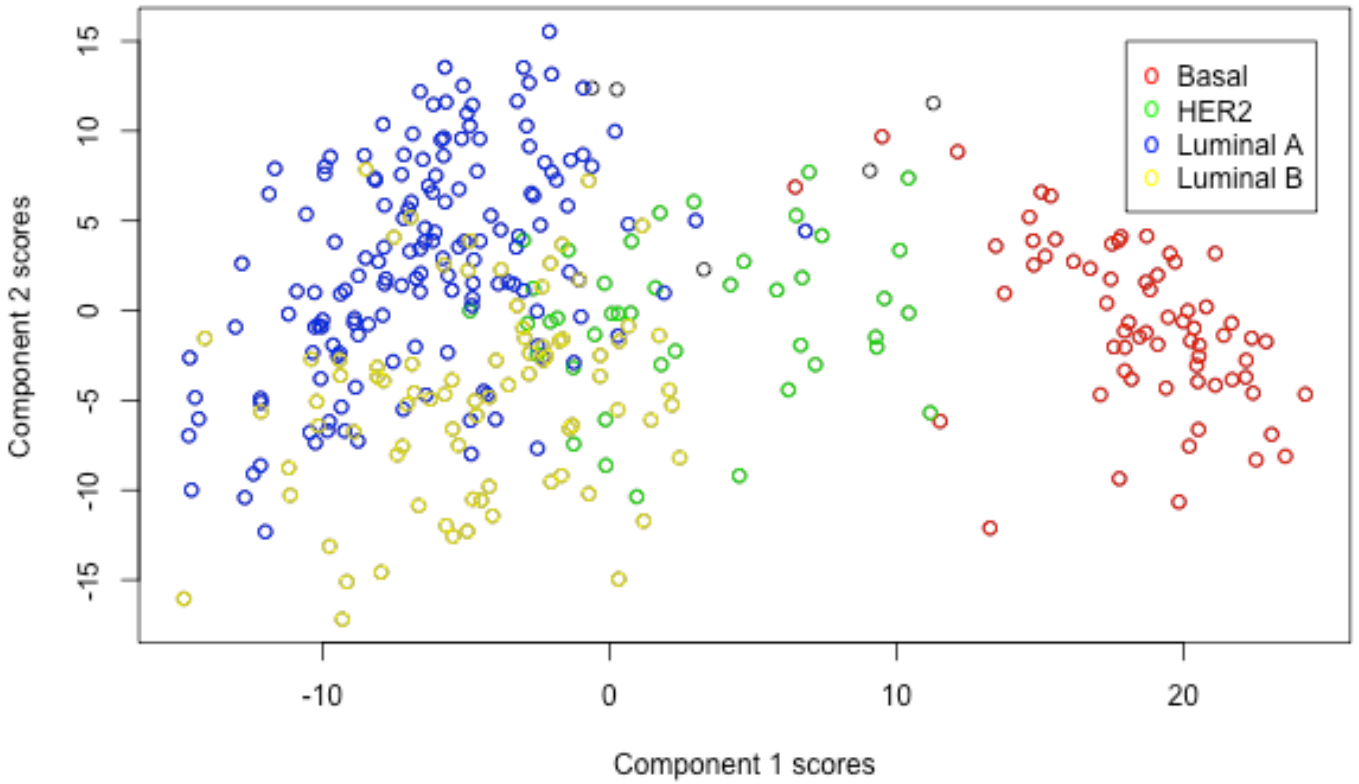
# Estimates (factorized)



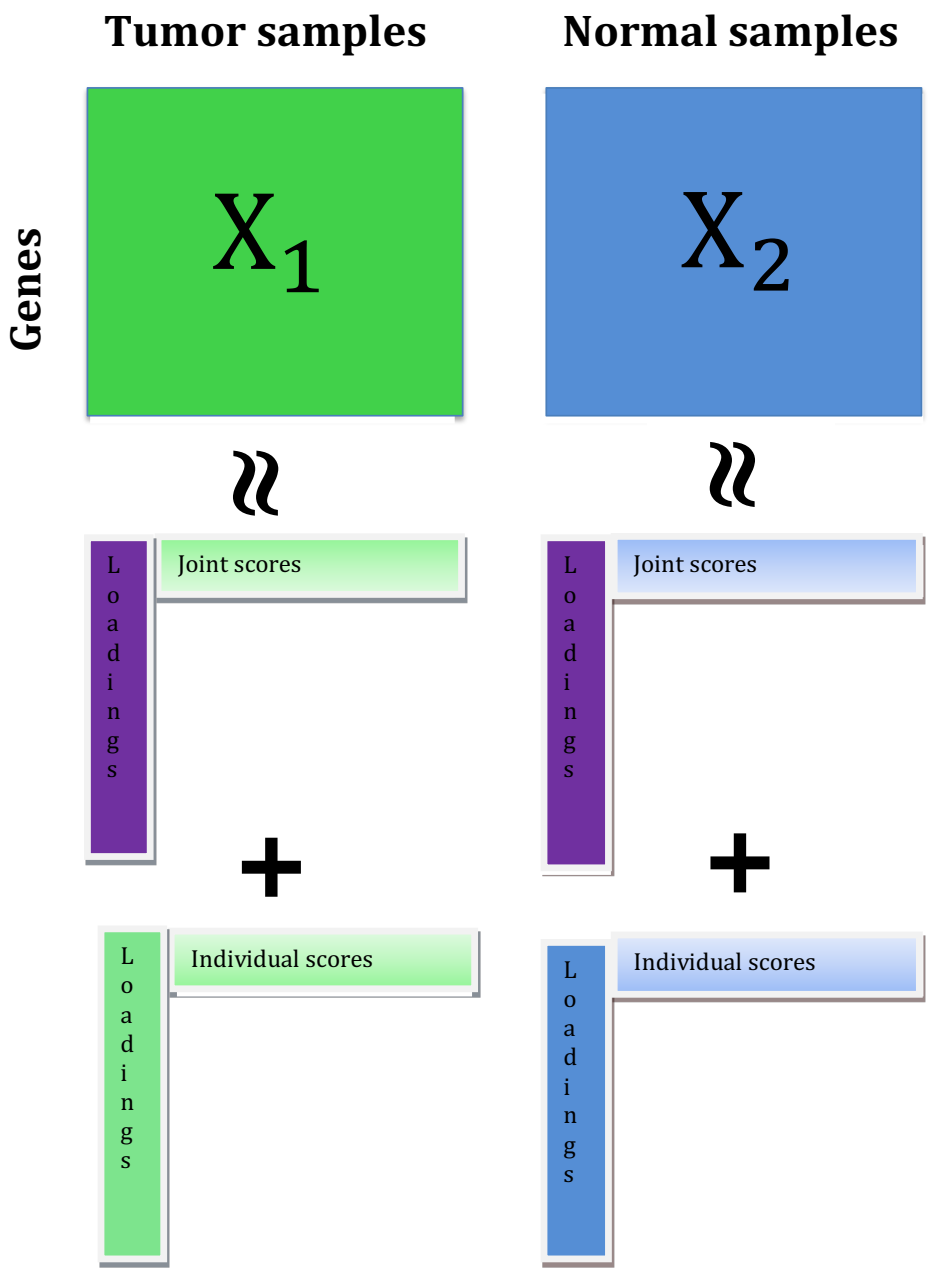
# Estimates (factorized)



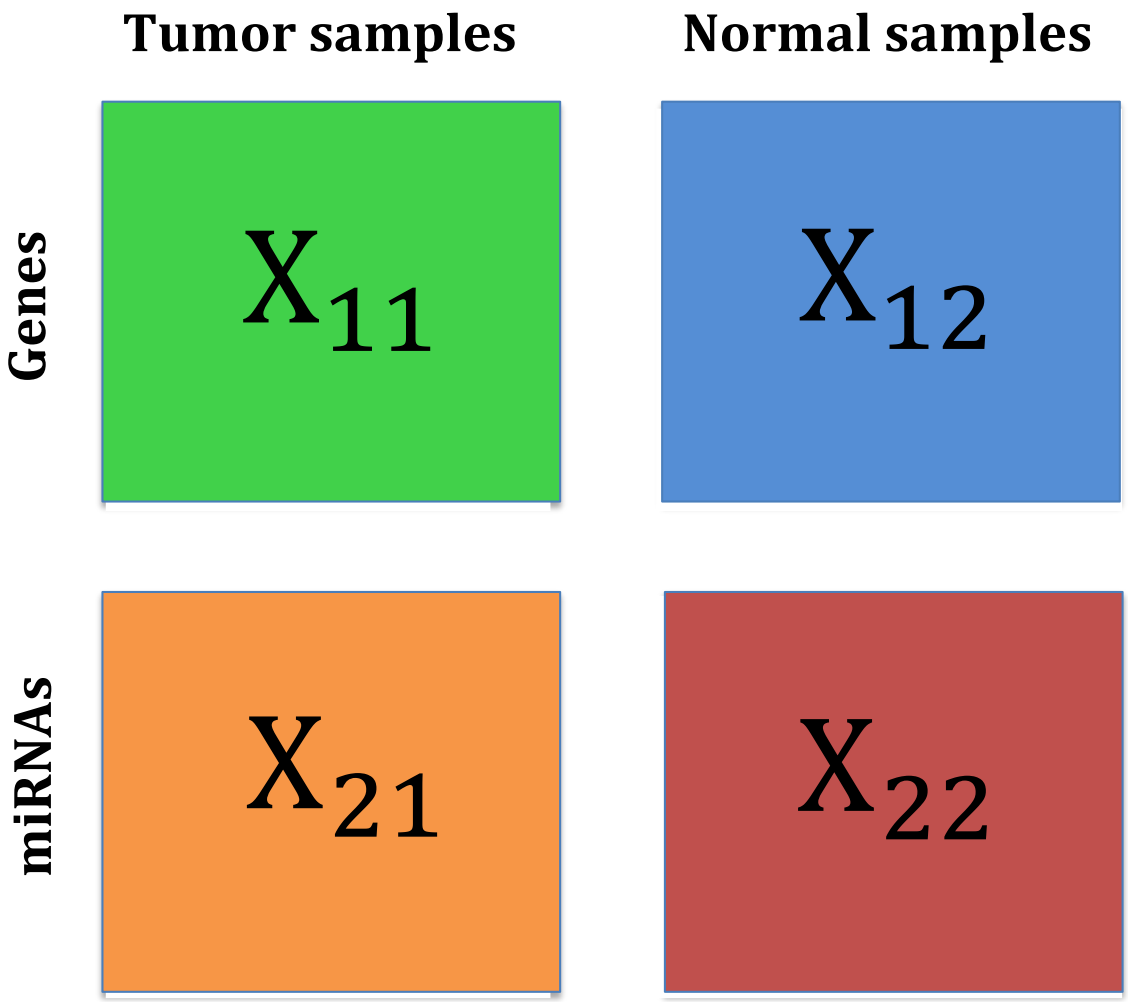
# Joint PCs


Basal
HER2
Luminal A
Luminal B
Component 2 scores
15
10
5
0
-5
-10
-15
-10
0
10
20
Component 1 scores

# Horizontally linked data: UNIFAC

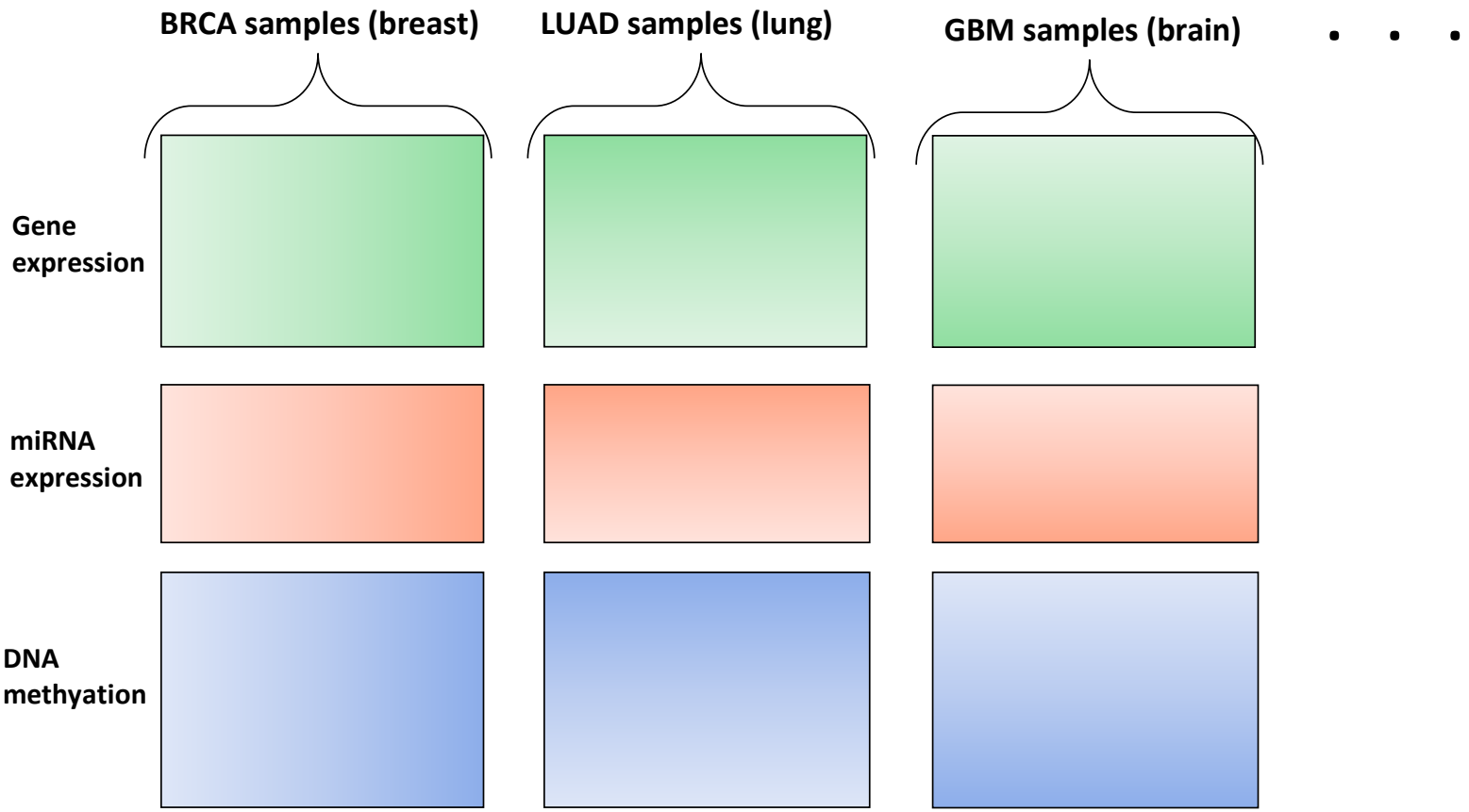


# BIDIFAC: Bidimensionally linked data



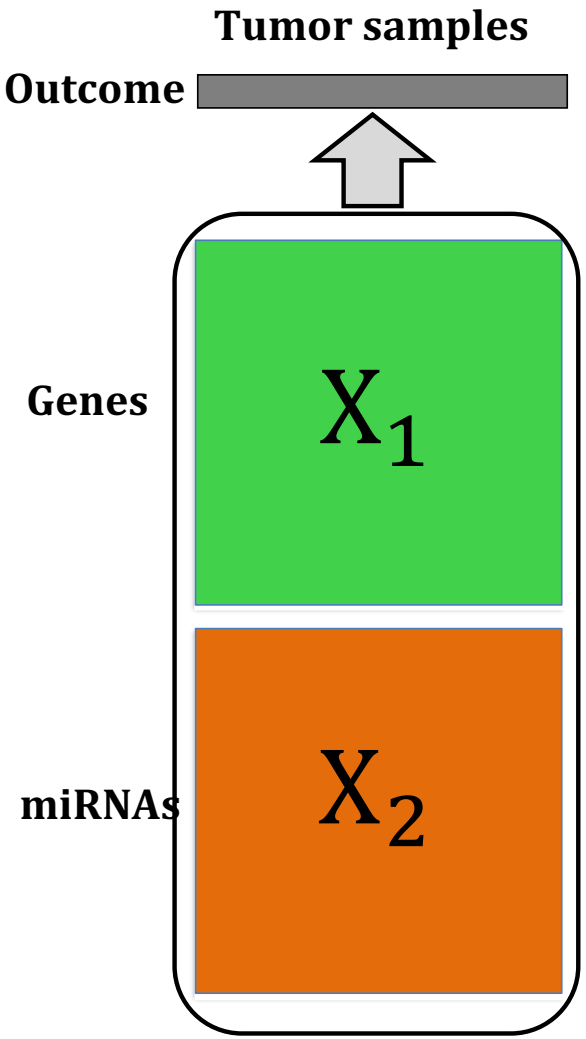

**Jun Young Park** and Eric F. Lock. Integrative Factorization of Bidimensionally Linked Matrices. *Biometrics*, 76 (1): 61-74, 2020.

# BIDIFAC+: Pan-omics pan-cancer integration

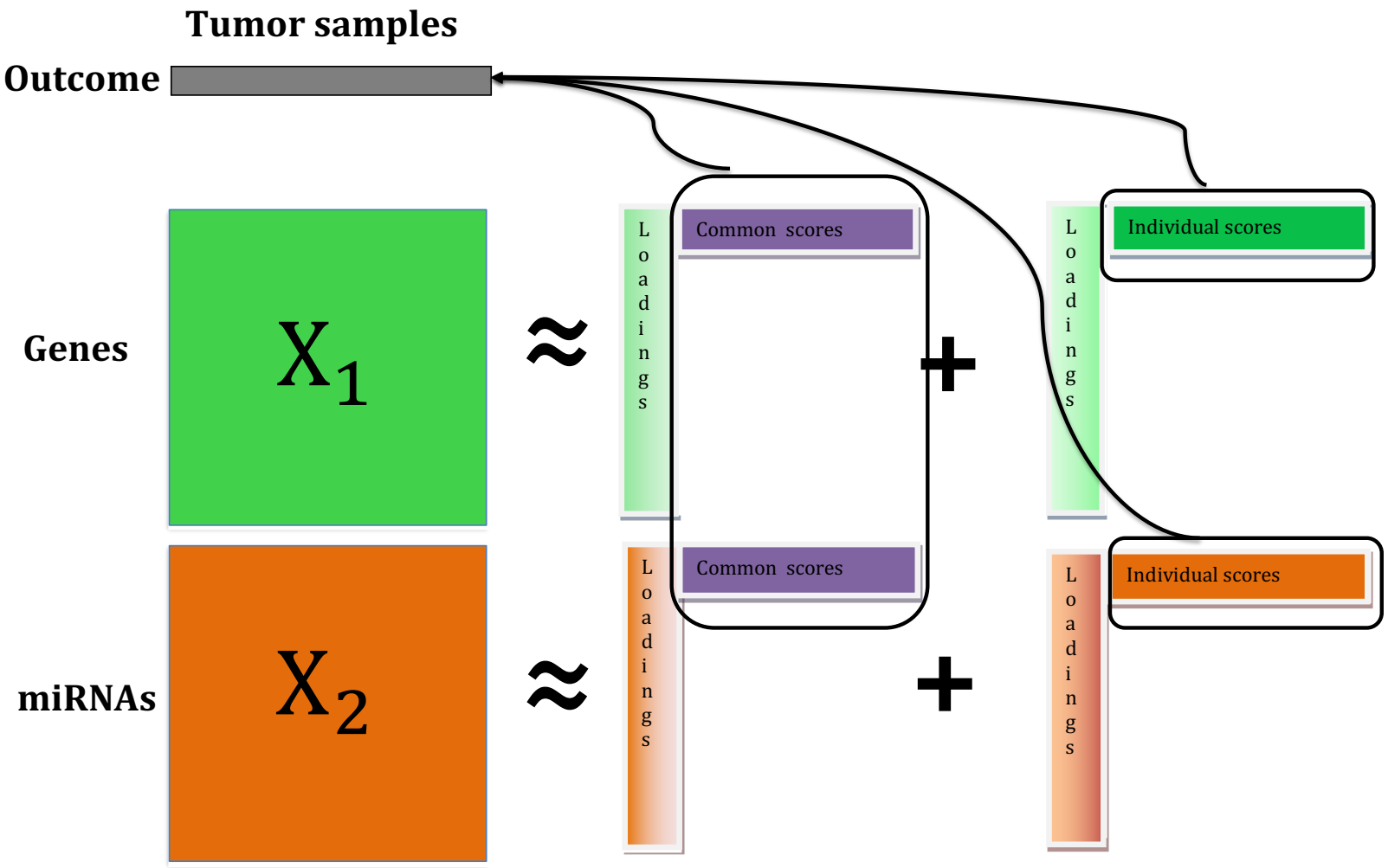



Eric F. Lock, **Jun Young Park**, **Katherine A. Hoadley**. Bidimensional linked matrix factorization for pan-omics pan-cancer analysis. *Annals of Applied Statistics*, 16 (1): 193-215, 2022

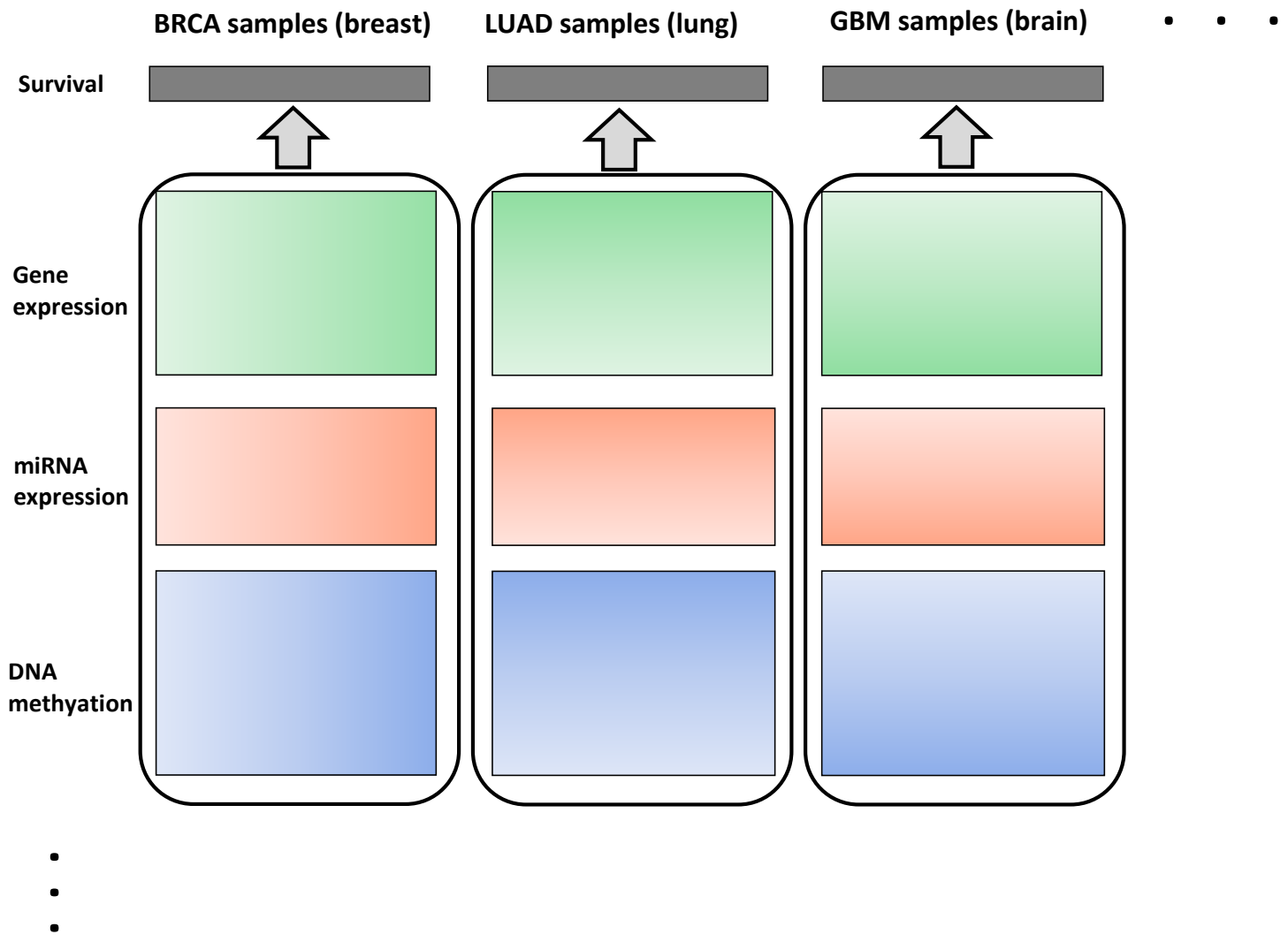
# Outcome prediction



# Outcome prediction: UNIFAC



# Outcome prediction: pan-omics pan-cancer




**Sarah Samorodnitsky**, **Katherine A. Hoadley**, Eric F. Lock. A Hierarchical Spike-and-Slab Model for Pan-Cancer Survival Using Pan-Omic Data. *BMC Bioinformatics*, 23: 235, 2022.

# Bayesian probabilistic matrix factorization (BPMF)

- Minimizing $\frac{1}{2}||X - \hat{X}||_F^2 + \lambda||\hat{X}||_*$ is equivalent to

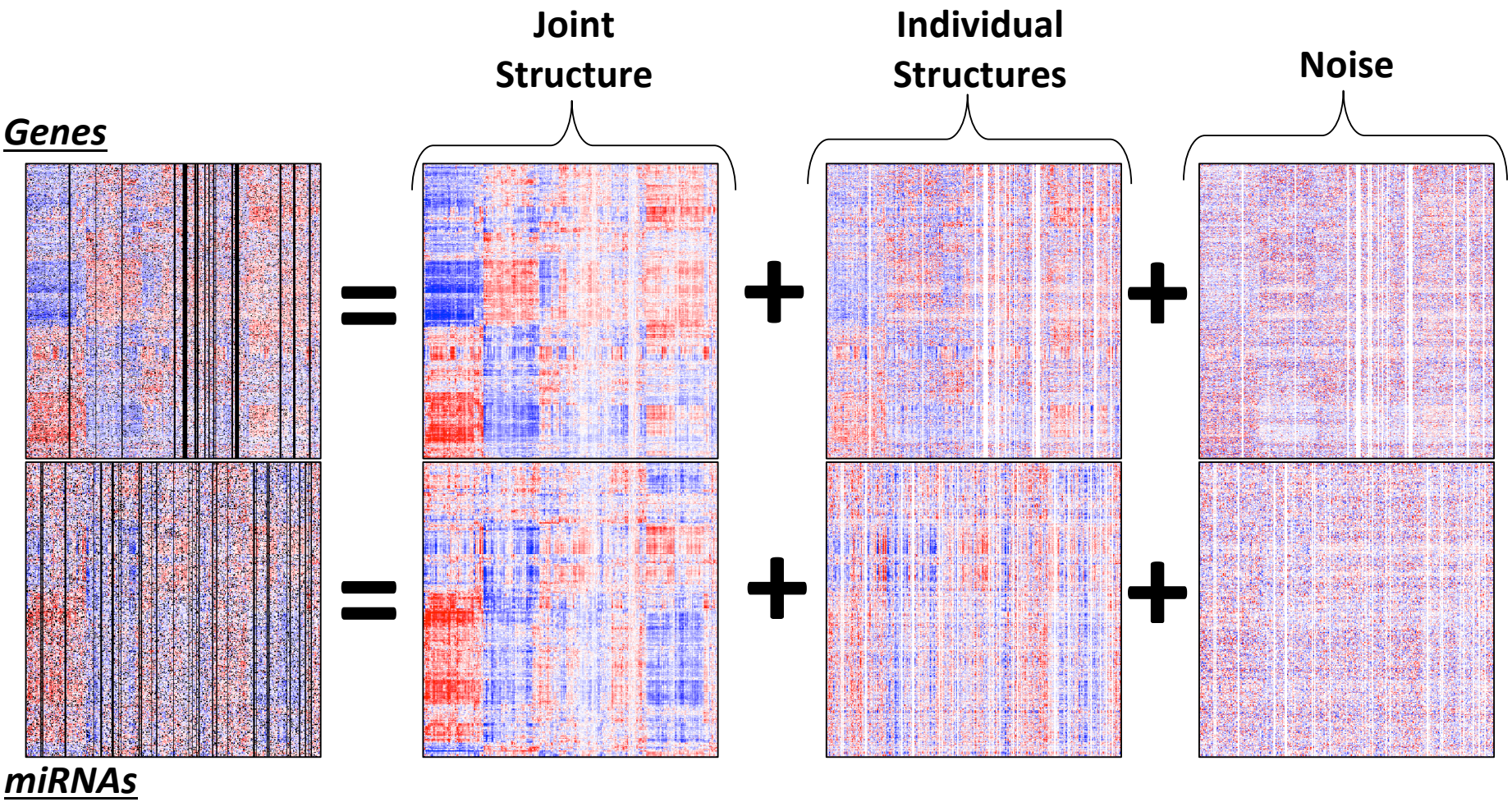
$$||X - UV||_F^2 + \lambda(||U||_F^2 + ||V||^2)$$

- Posterior mode with $N(0, 1/\lambda)$ priors on $U$ and $V$
- Analogous results for UNIFAC / BIDIFAC objectives
- Gibbs sample to infer full posterior!
- Useful for multiple imputation, etc.

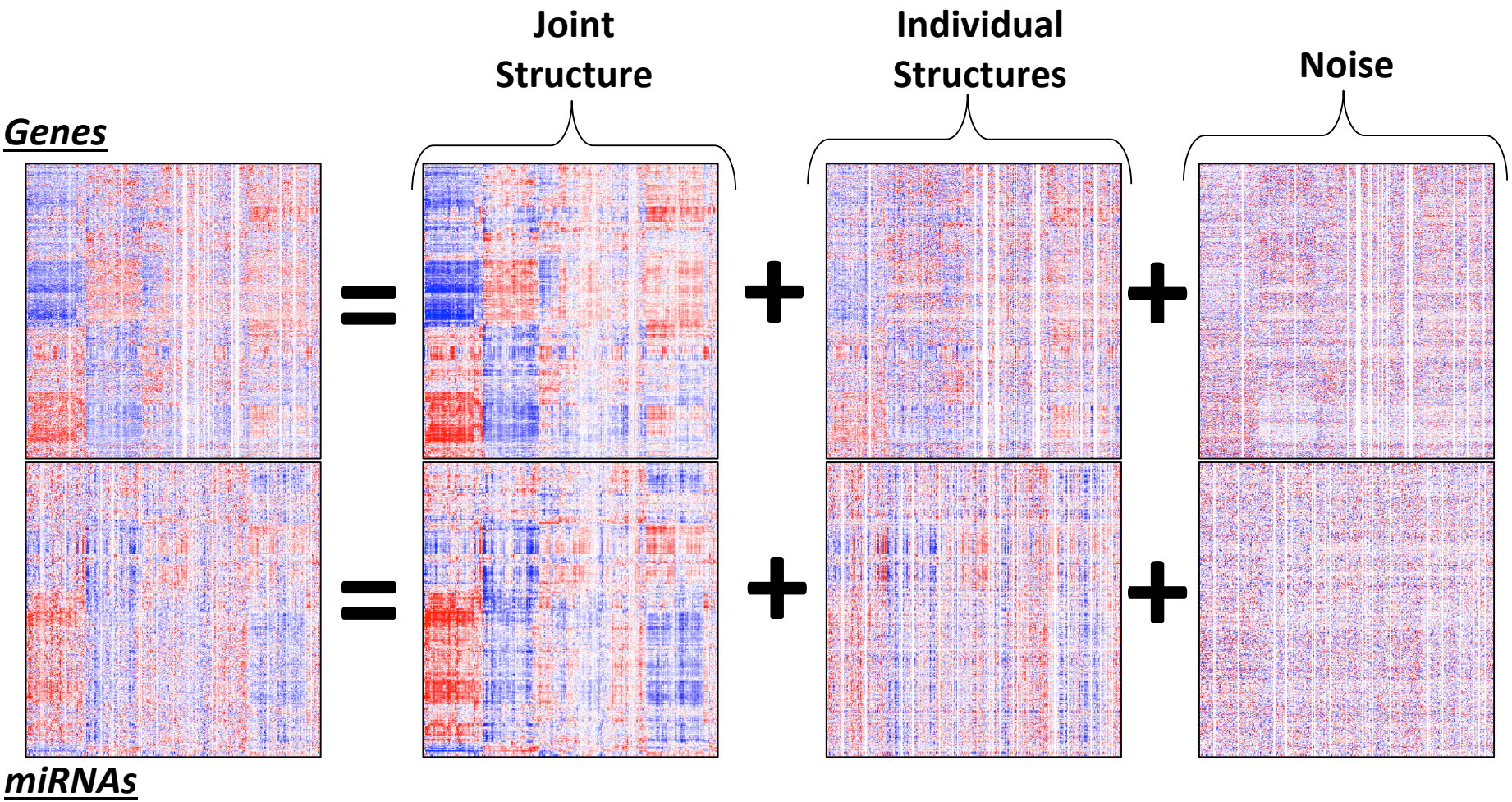
---

**Sarah Samorodnitsky** and Eric F. Lock. A Bayesian Approach to Simultaneous Factorization and Prediction Using Multi-Omic Data. In preparation.

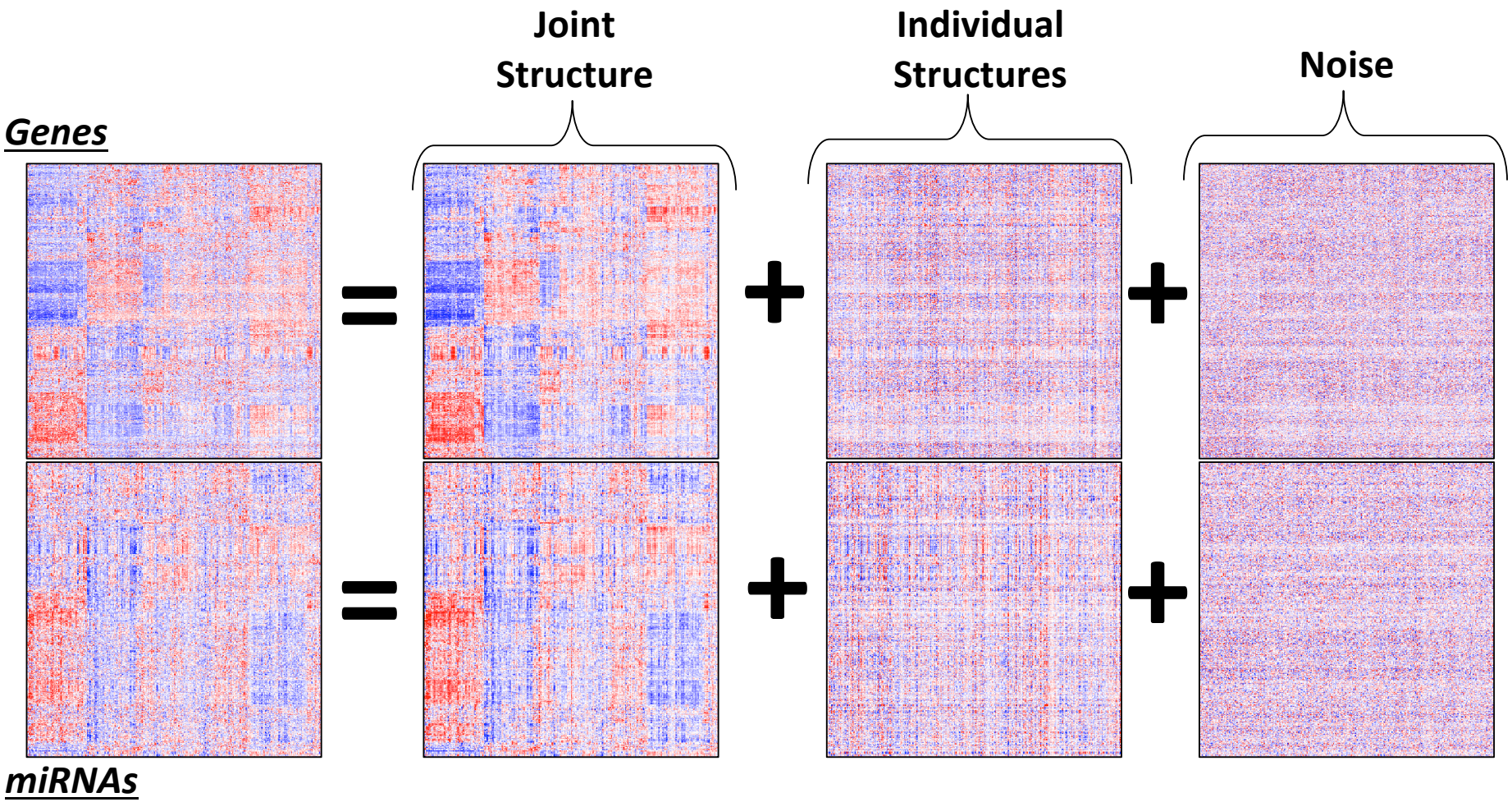
# Missing Data



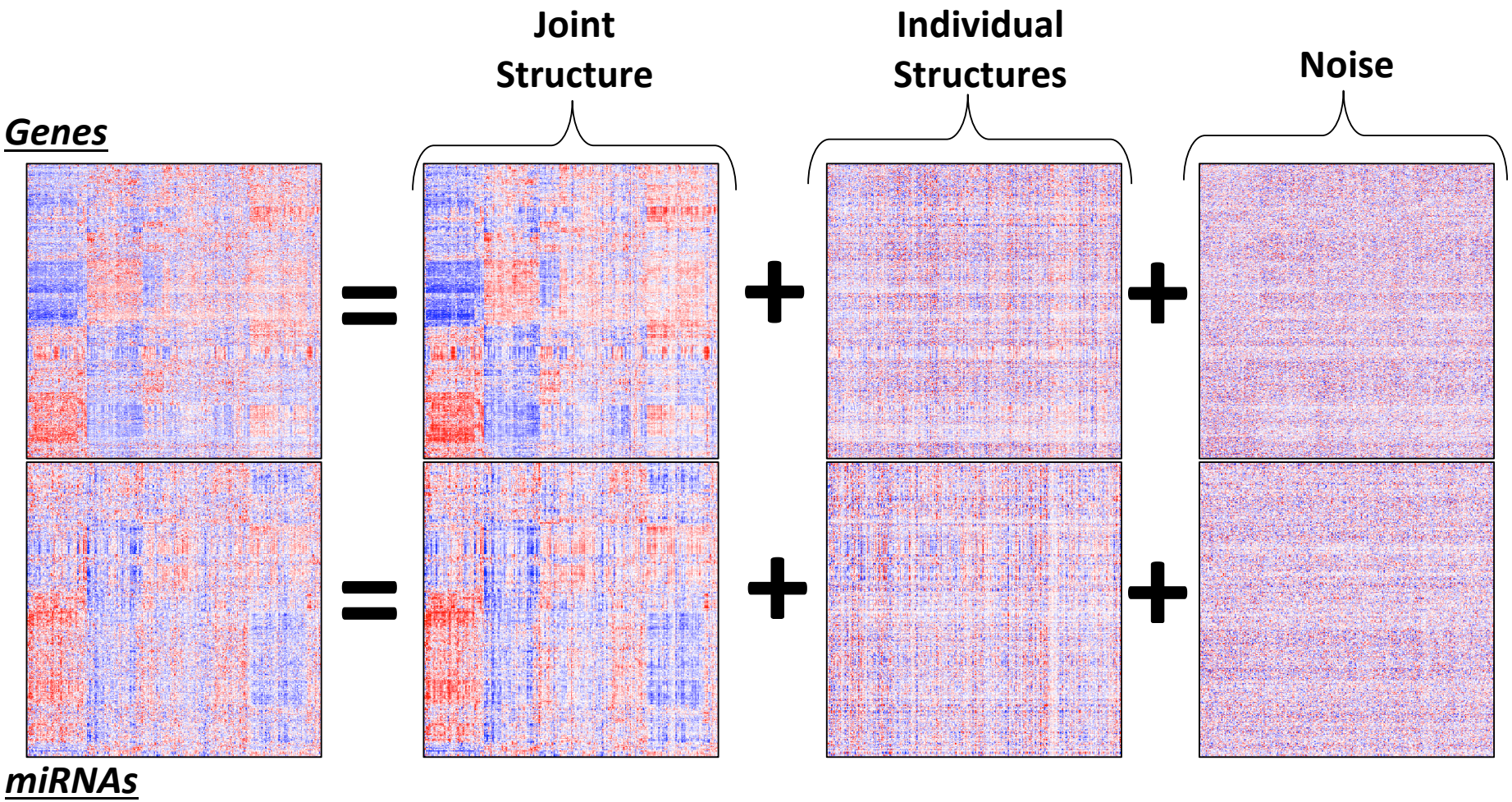
# Missing Data (Imputed)



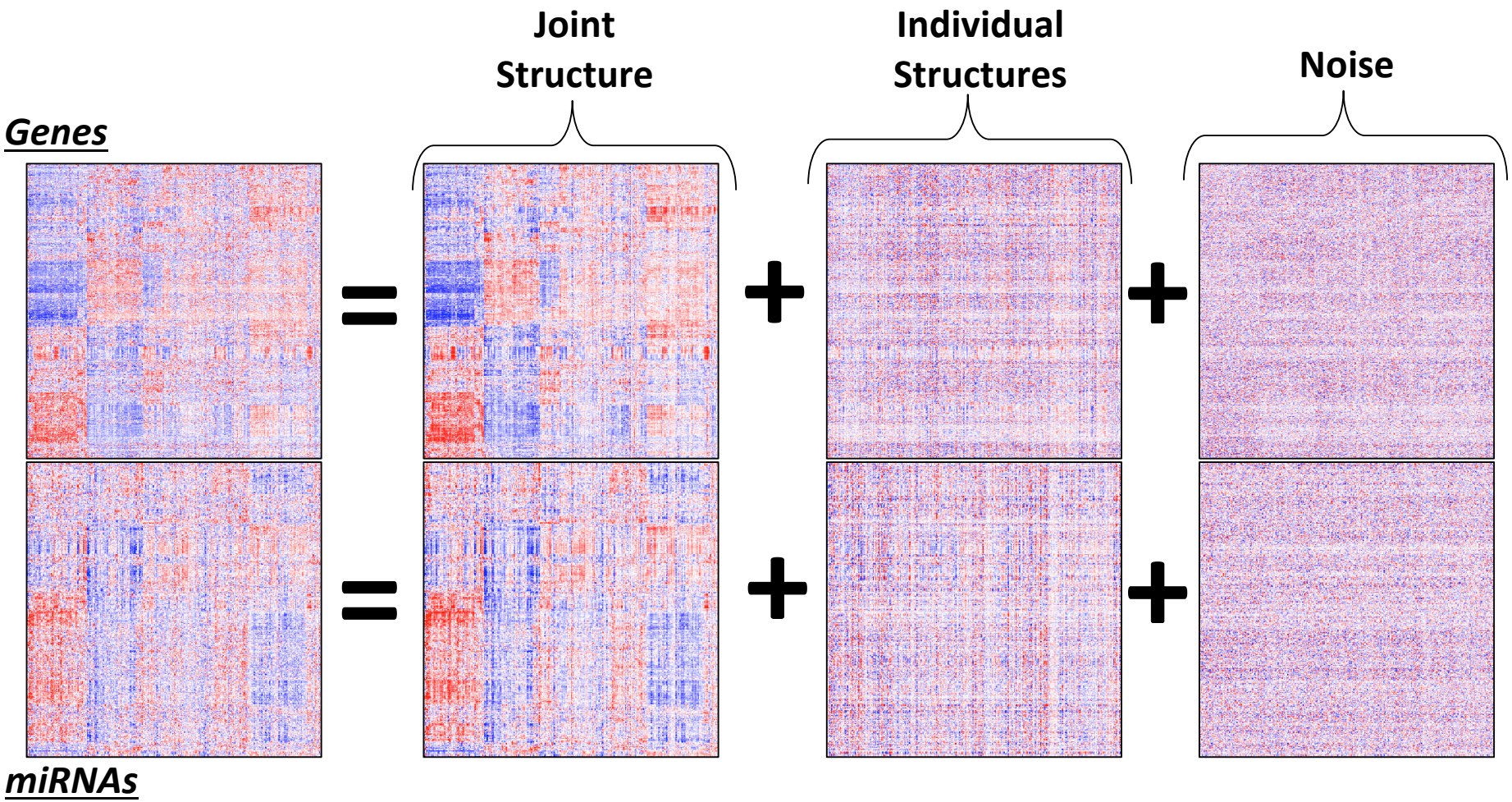
# Posterior Sample 1



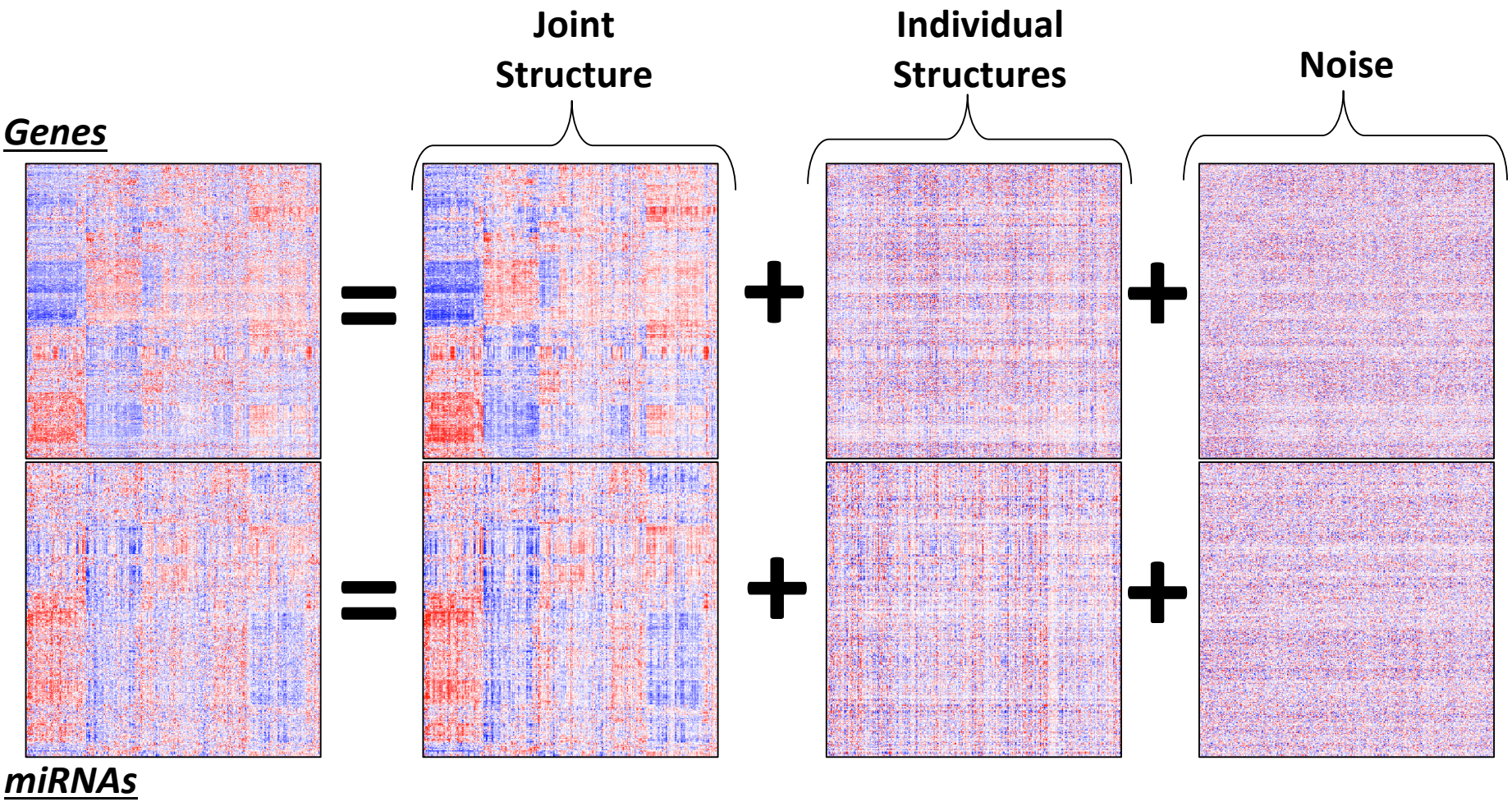
# Posterior Sample 2



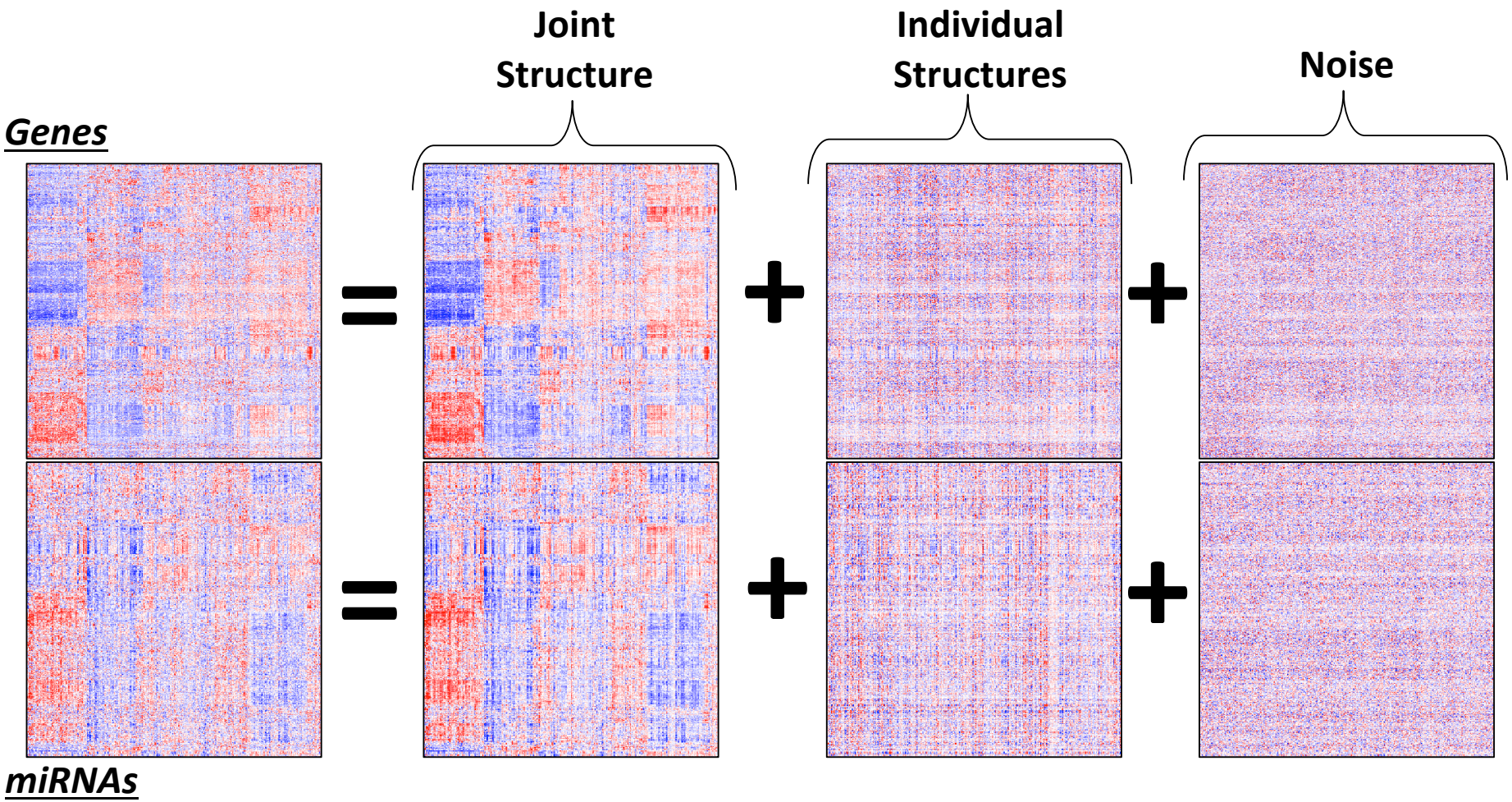
# Posterior Sample 3



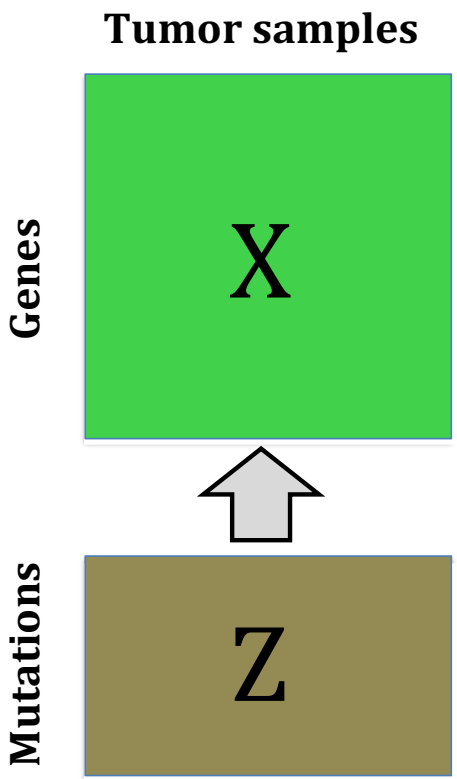
# Posterior Sample 4



# Posterior Sample 5

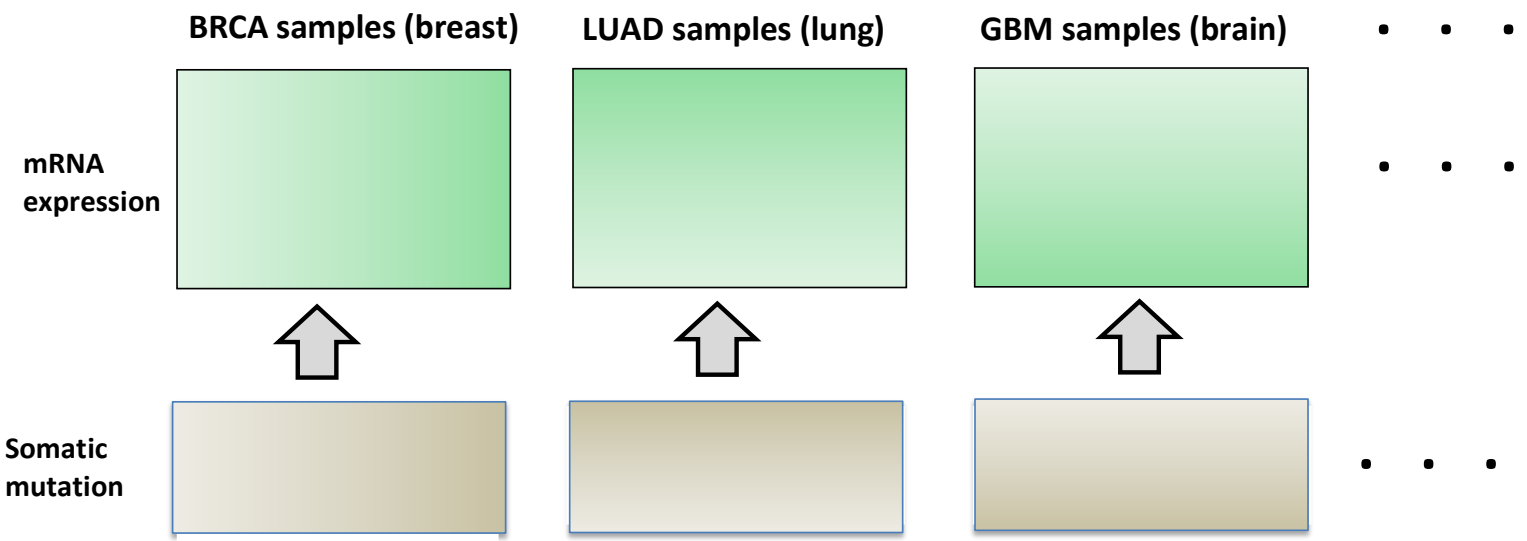


# Augmented reduced rank regression (aRRR)




- Minimize $||X - BZ - S||_F^2 + \lambda_B ||B||_* + \lambda_S ||S||_*$

# Multi-cohort augmented reduced rank regression (amRRR)




- Decompose $S$ and $B$ into shared and unshared components.

---

**Jiuzhou Wang** and Eric F. Lock. Multi-cohort augmented reduced rank regression. In preparation.

# Thank you!


- Support from:
  - NCI (R21-CA231214)
  - NIGMS (R01-GM130622)


- Slides available at www.ericfrazerlock.com/talks.