Bidimensional Linked Matrix Decomposition for Pan-Omics Pan-Cancer Analysis

Eric F. Lock

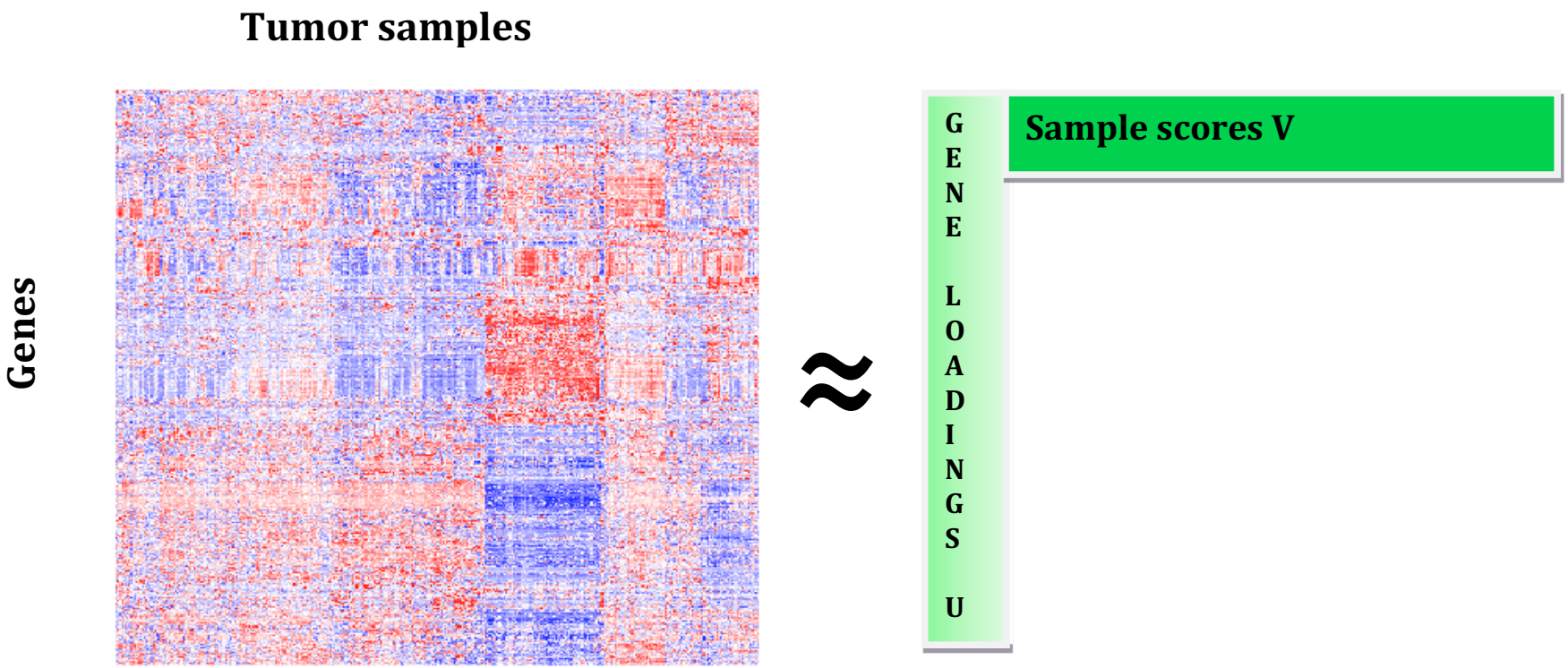
University of Minnesota, Division of Biostatistics

With **Jun Young Park**, University of Toronto
and **Katie Hoadley**, University of North Carolina

TAMU Statistics, 04/15/2022

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

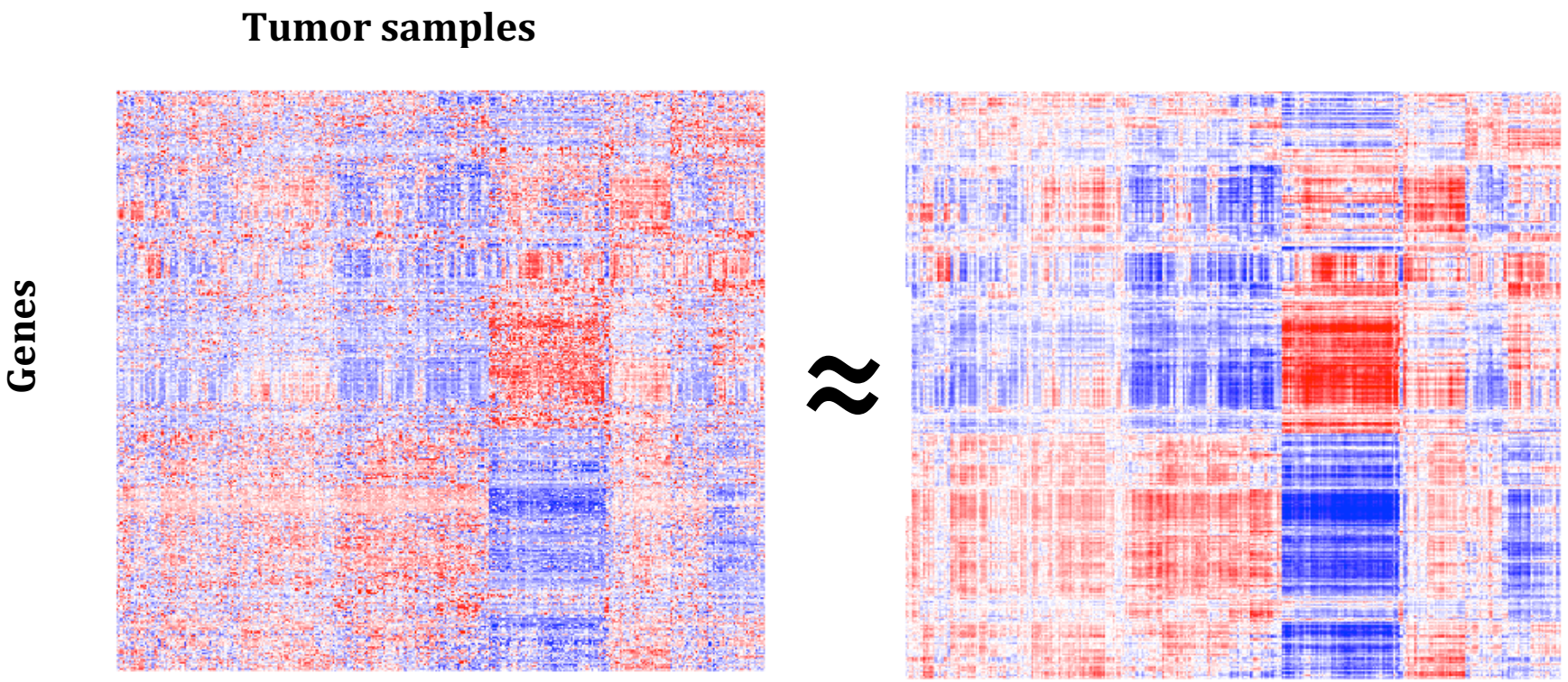
$$U = [\mathbf{u_1}, \cdots, \mathbf{u_r}, u_{r+1}, \cdots]$$
$$D = \text{diag}(\mathbf{a_1} + e_1, \cdots, \mathbf{a_r} + e_r, e_{r+1}, \cdots)$$
$$V = [\mathbf{v_1}, \cdots, \mathbf{v_r}, v_{r+1}, \cdots]$$

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

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

Matrix factorization

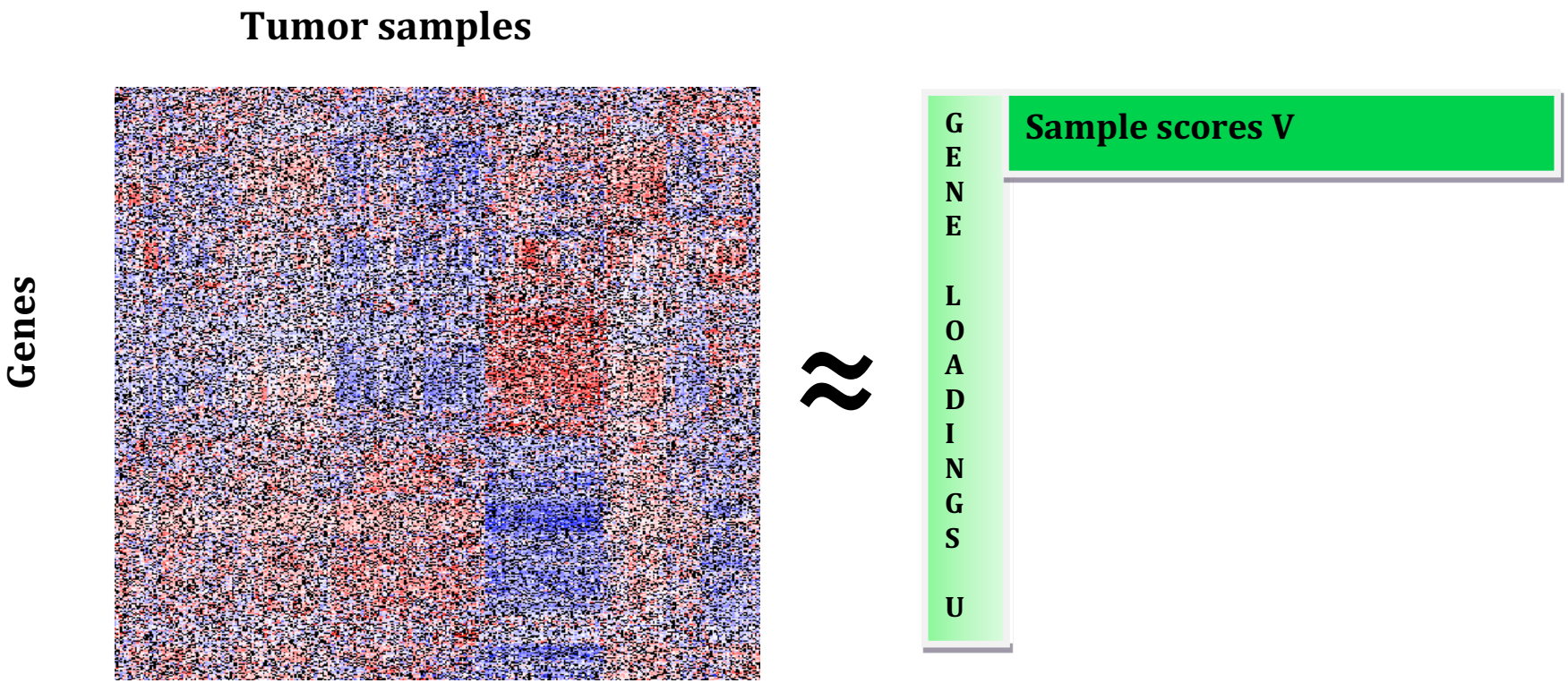
- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



- Rank 18 nuclear norm approximation.

Matrix factorization: missing data

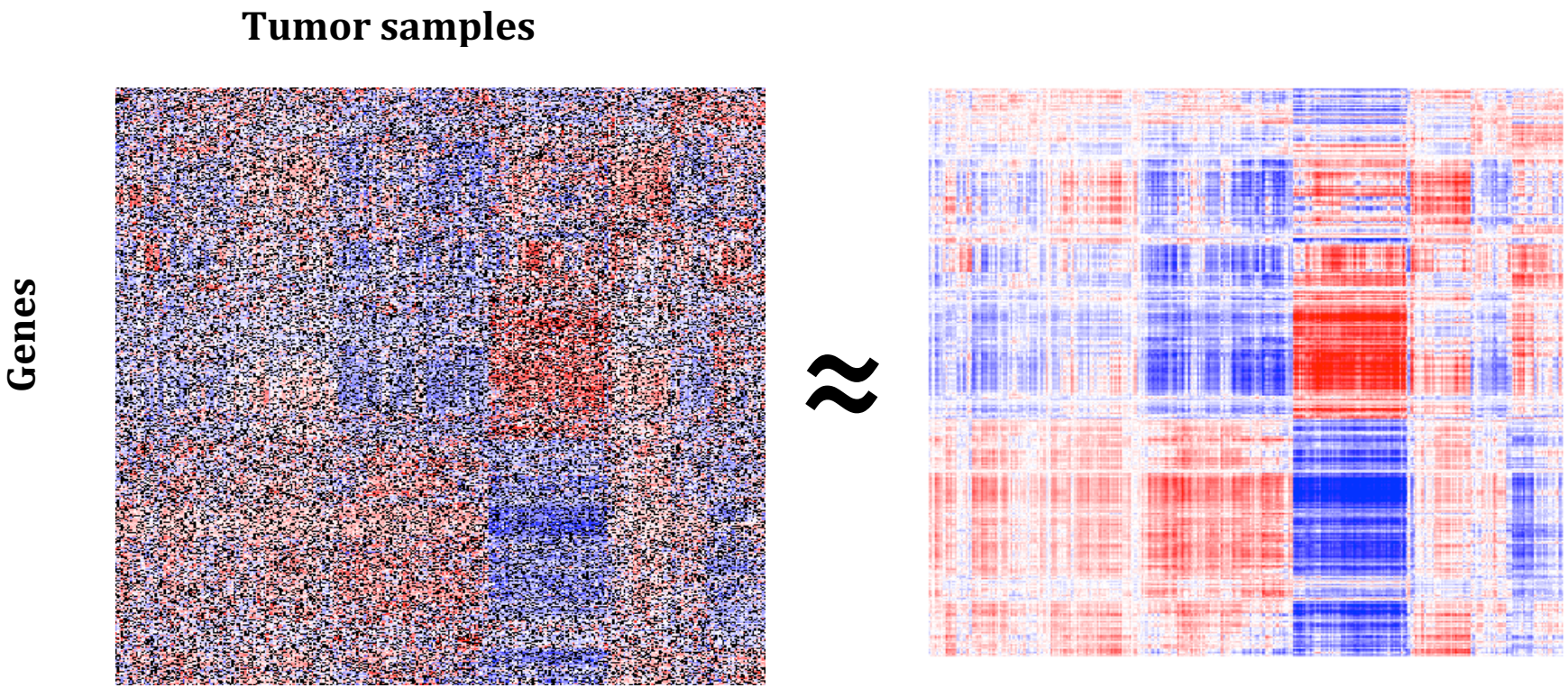
- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



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

Matrix factorization: missing data

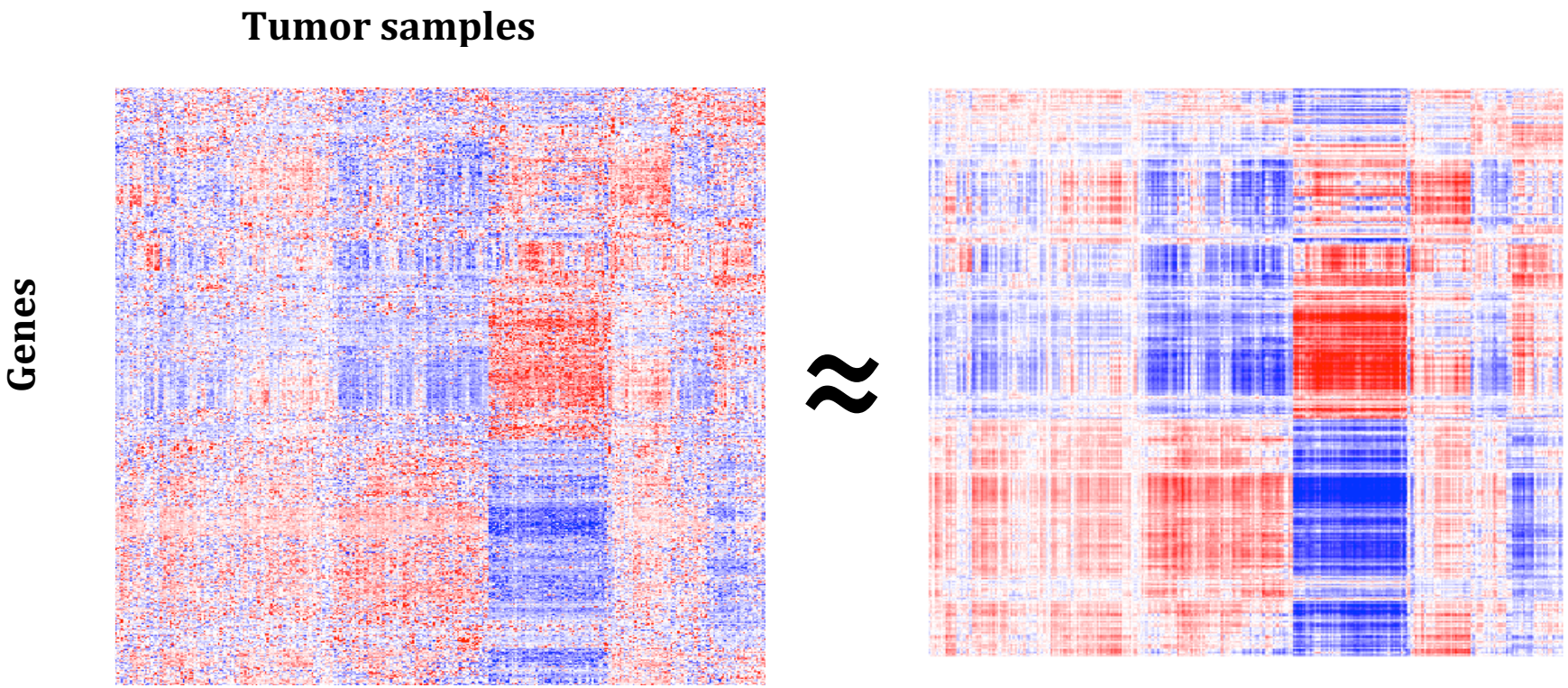
- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



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

Matrix factorization: missing data

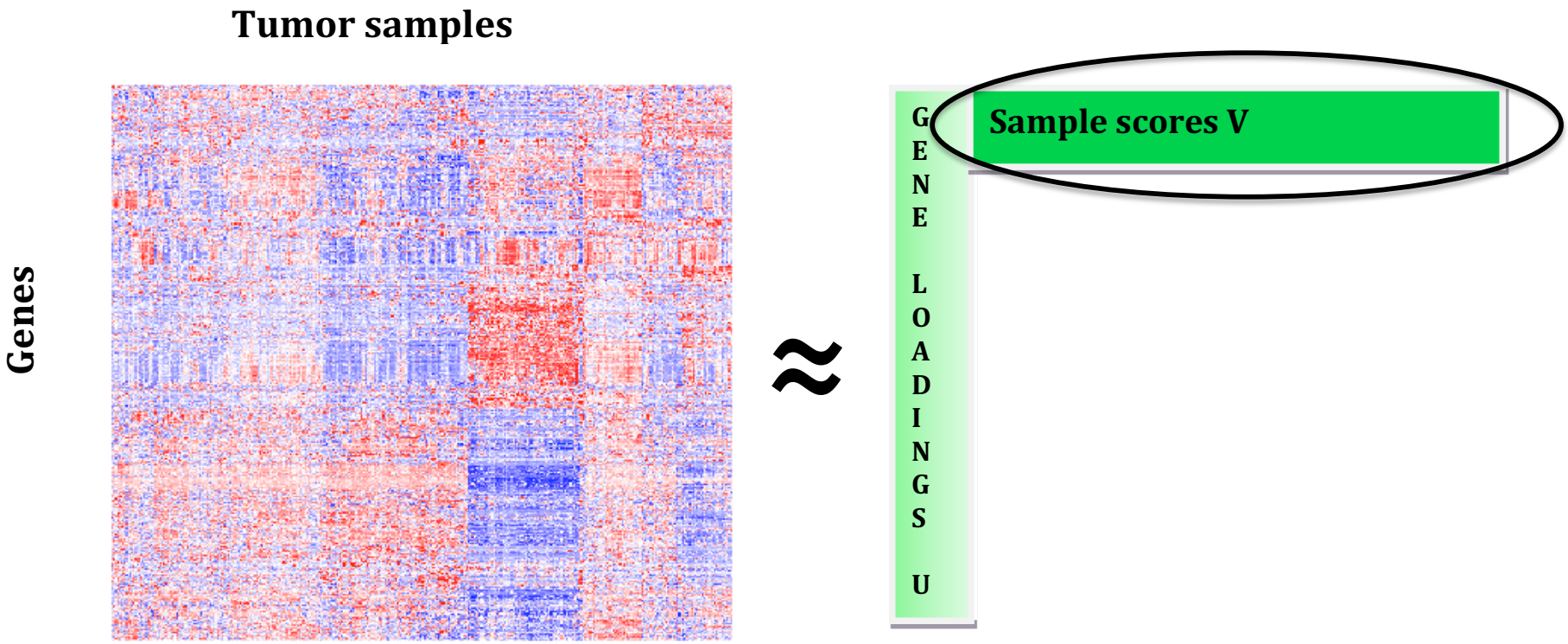
- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



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

Matrix factorization

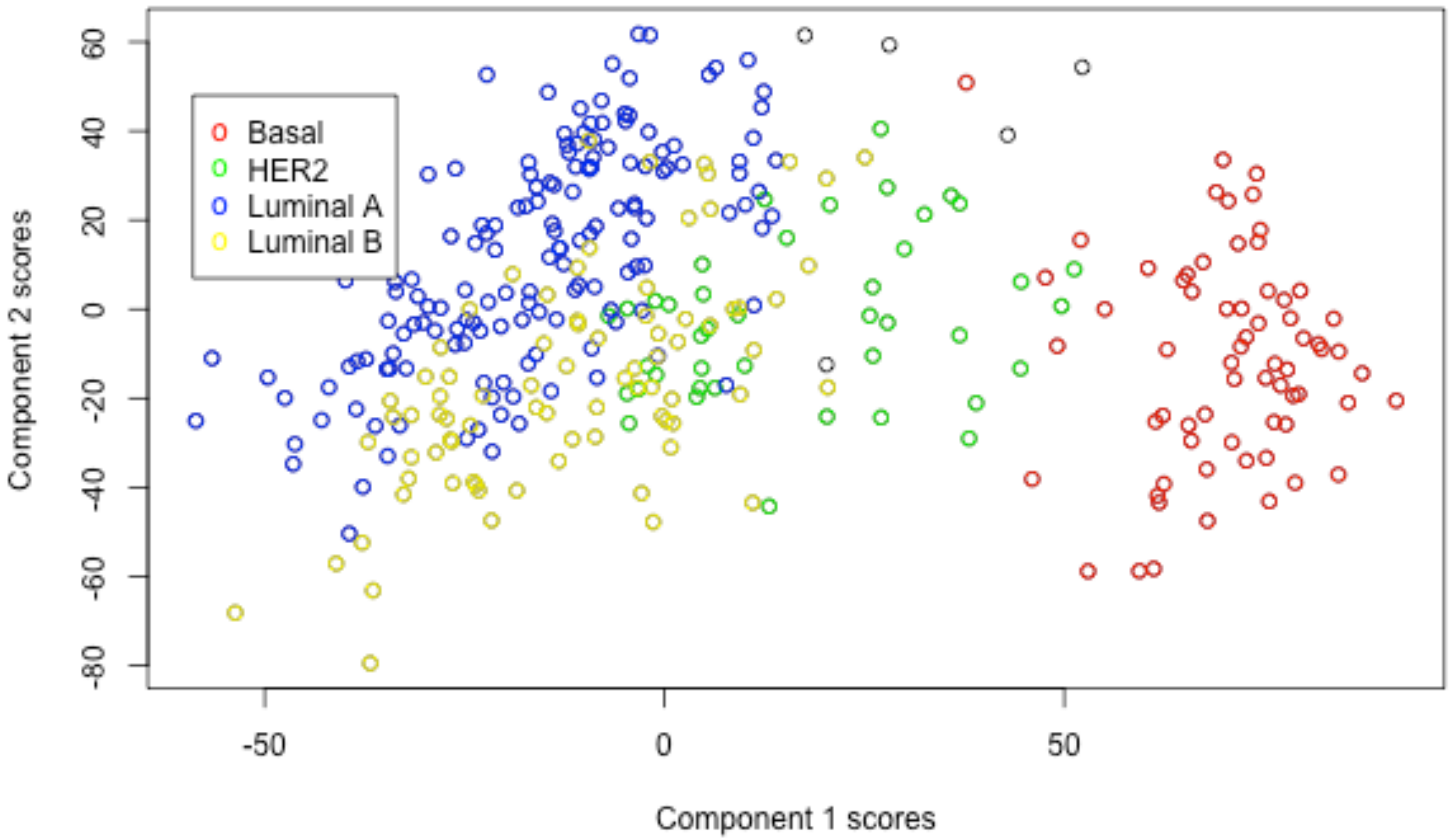
- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



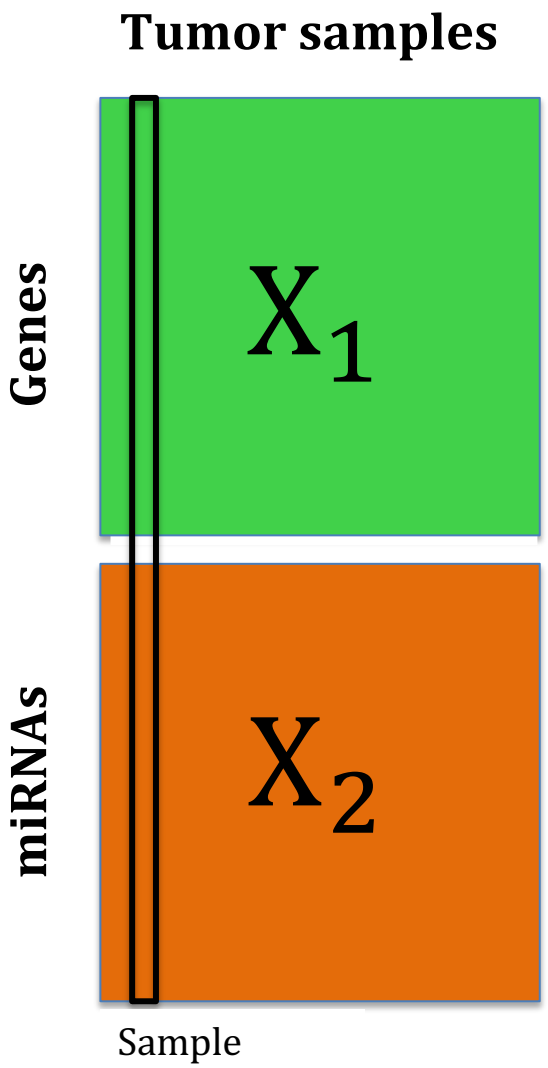
- Low rank factorization: $X \approx UV$, $U : m \times r$, $V : r \times n$.

Matrix factorization

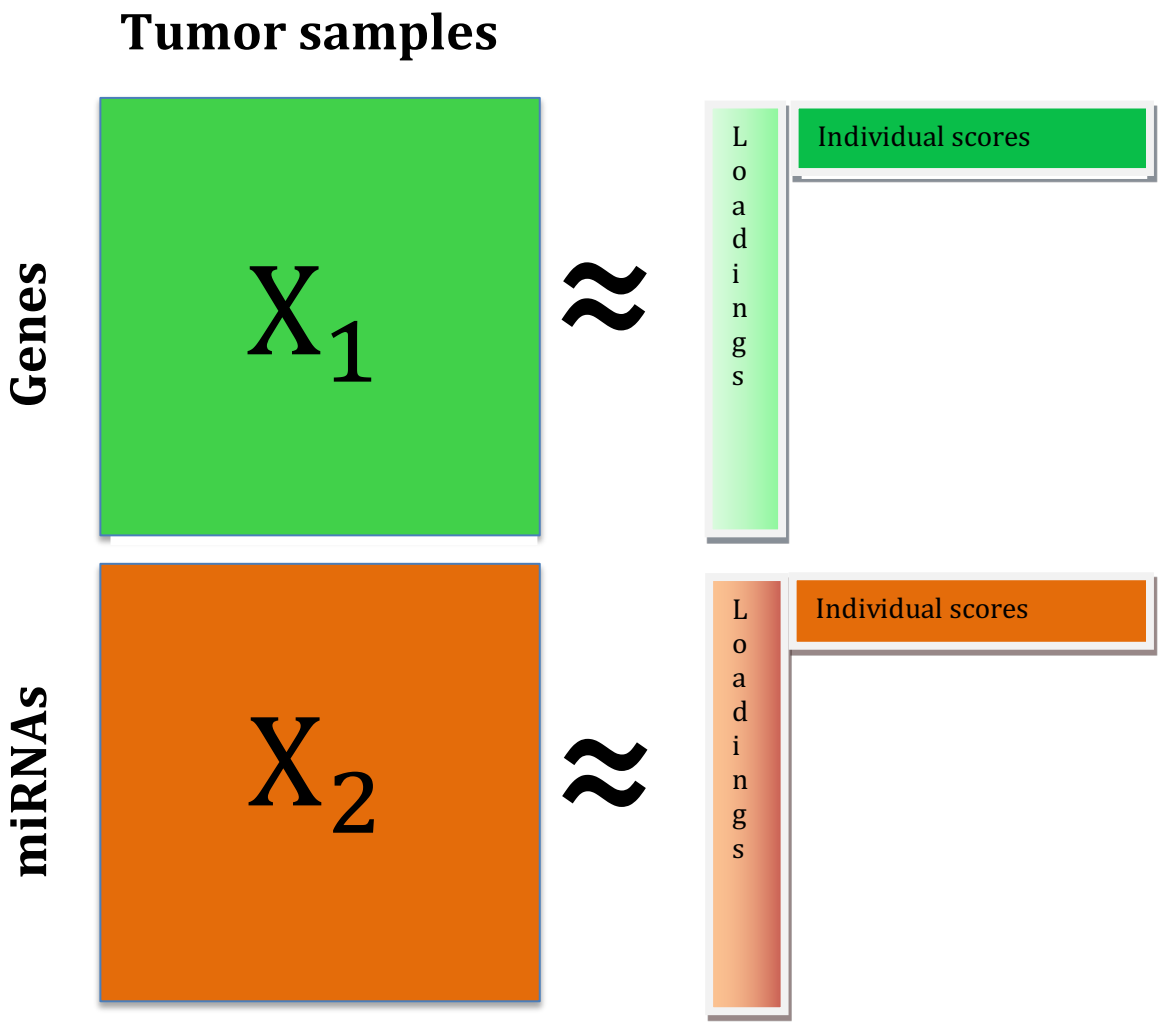
- First two principal component scores
  - Colored by breast tumor subtype



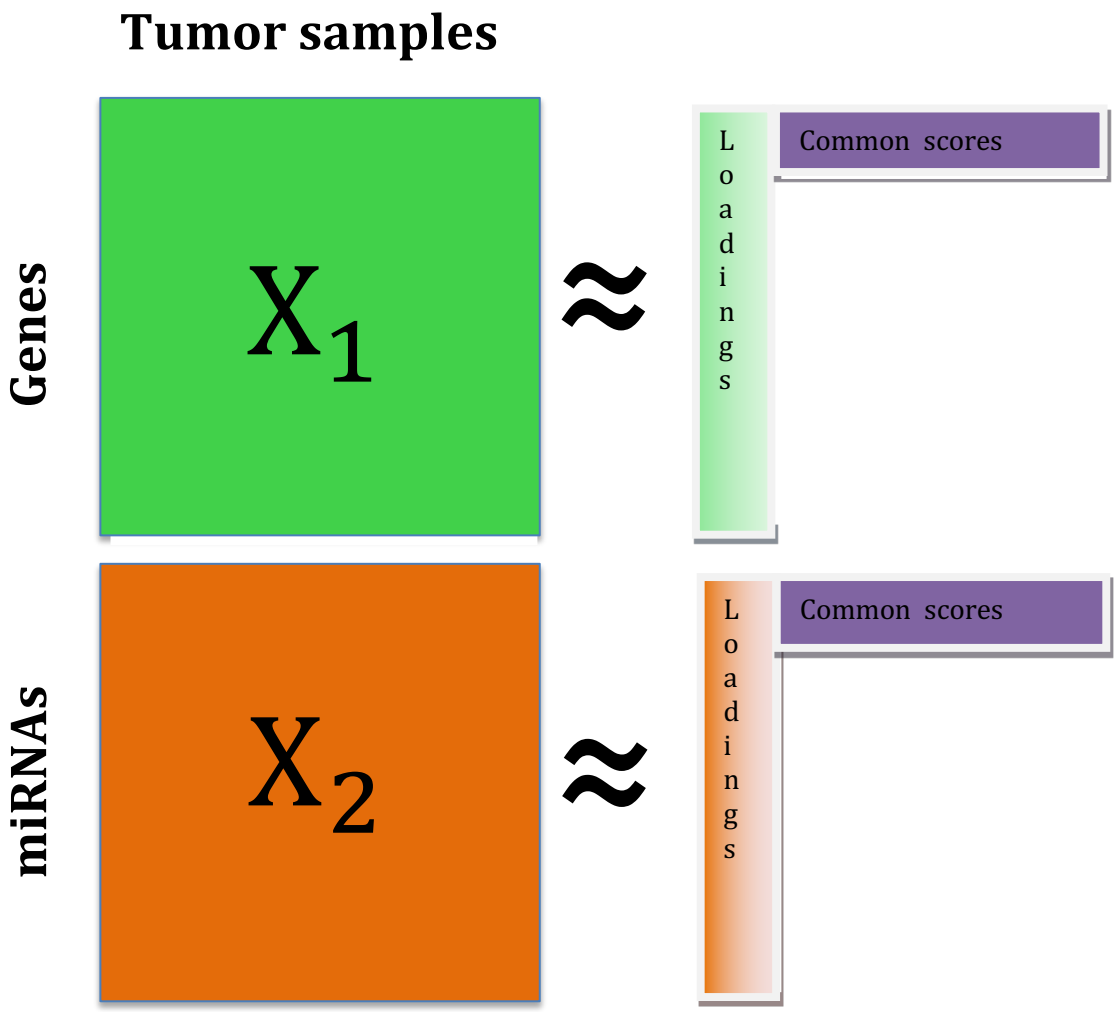
Vertically linked data



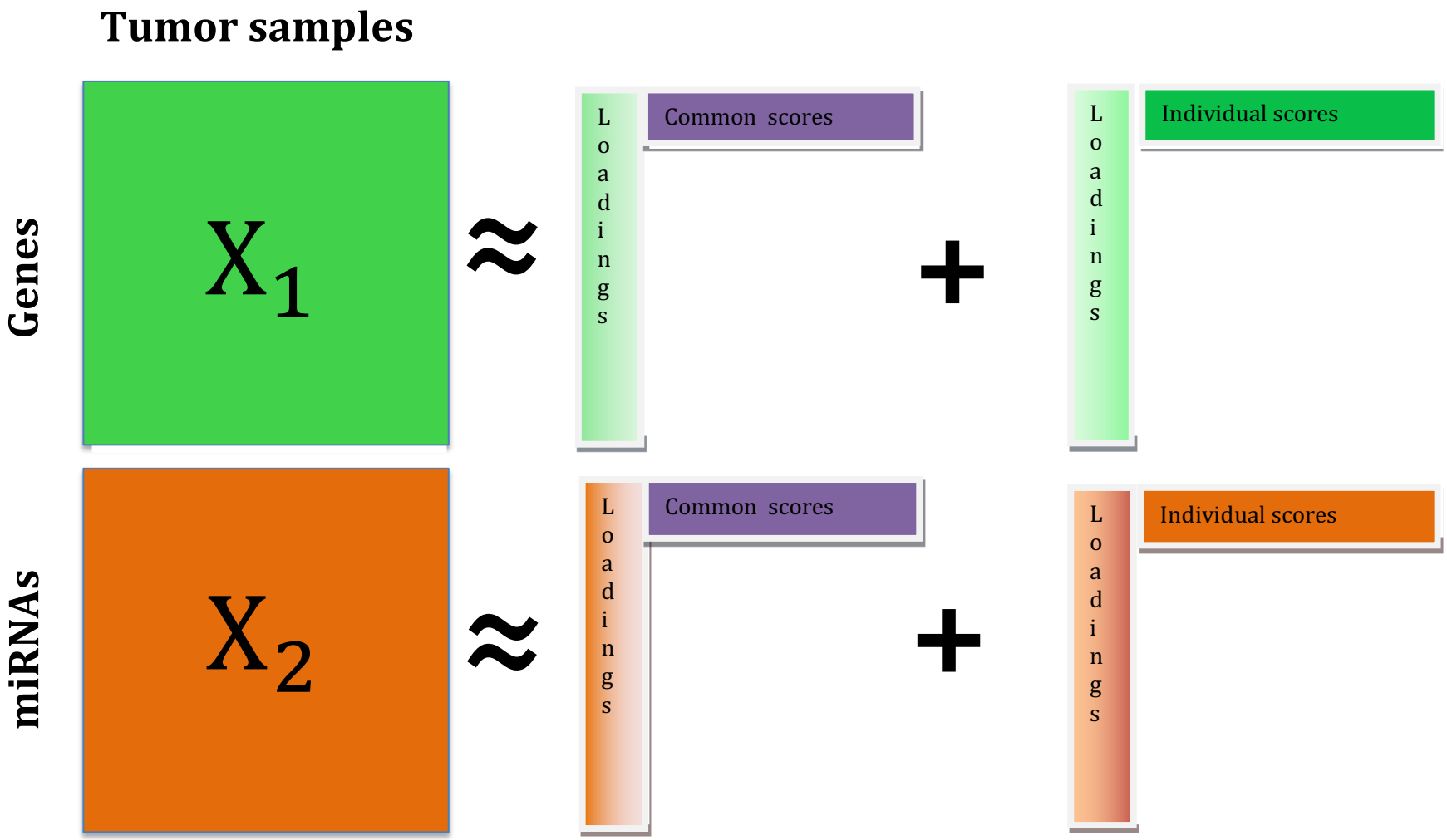
Vertically linked data: individual factorizations



Vertically linked data: joint factorization



Vertically linked data: joint and individual factorization



Joint + individual factorization methods

- JIVE [Lock, Hoadley, Marron, and Nobel, 2013]
  - "Joint and Individual Variation Explained"
- R.JIVE [O'Connell and Lock, 2016]
- AJIVE [Feng, Jiang, Hannig and Marron, 2018]
- SLIDE [Gaynanova and Li, 2018]
- GIPCA [Zhu, Li, Lock, 2018]
- COBE, SIFA, MOFA, & more!

Structured nuclear norm penalty

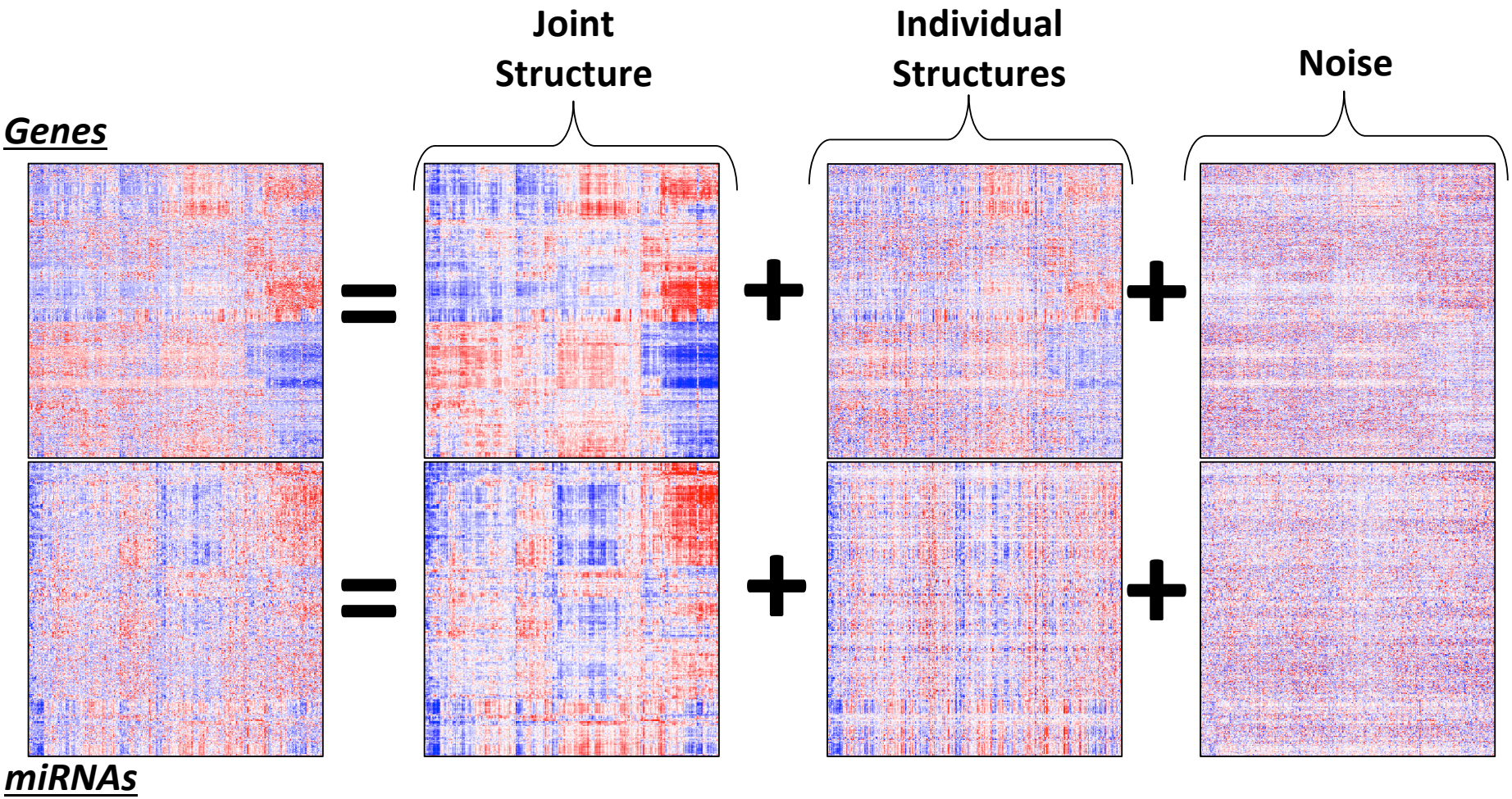
- $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ where $X_1 : m_1 \times n$, $X_2 : m_2 \times n$
- $X \approx J + A$ where $J = \begin{bmatrix} J_1 \\ J_2 \end{bmatrix}$ and $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$
- Minimize

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

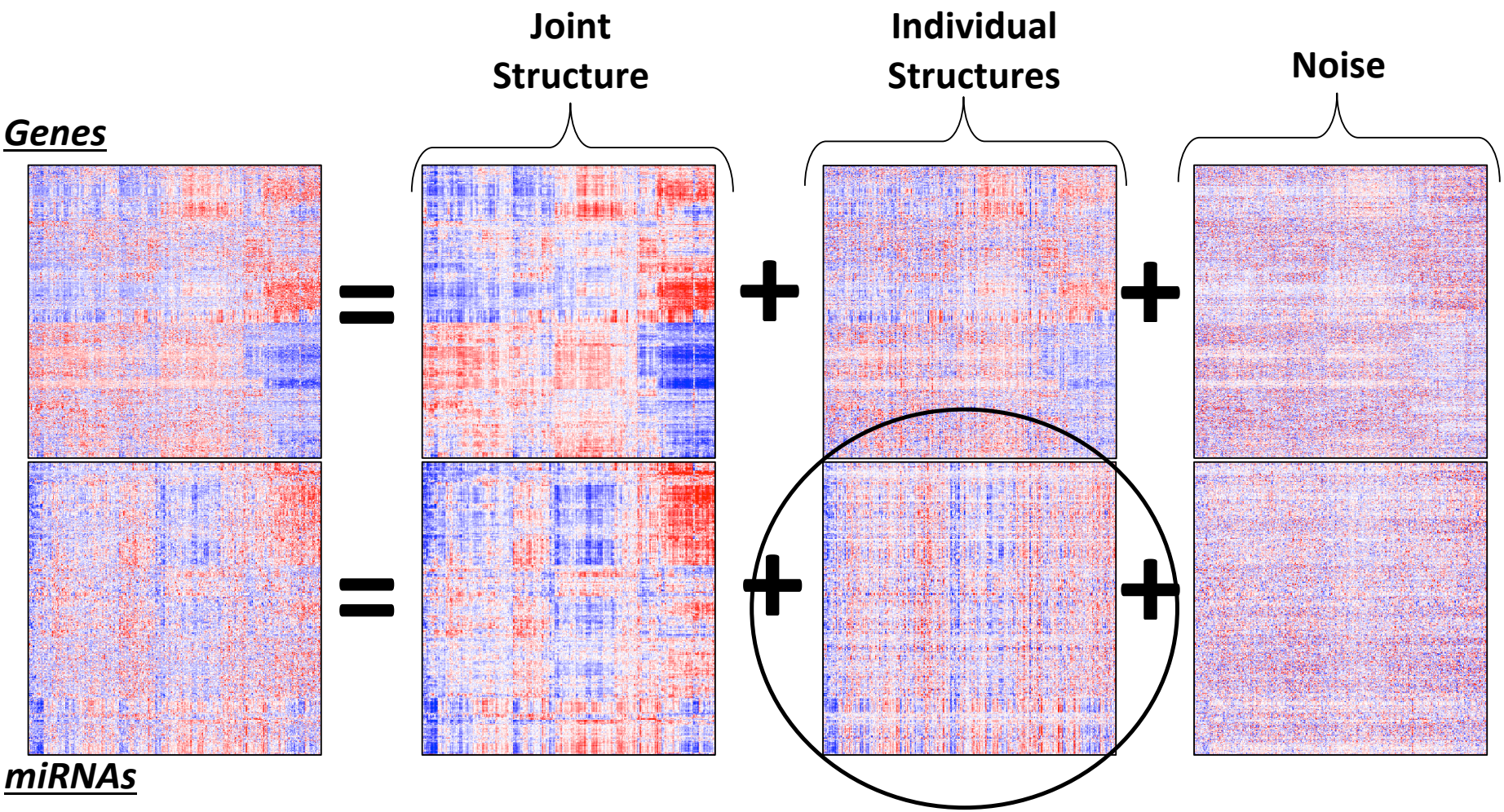
  where $\lambda_0 = \sqrt{n} + \sqrt{m_1 + m_2}, \quad \lambda_i = \sqrt{n} + \sqrt{m_i}$

- Update J, A_1, A_2 until convergence

JIVE Estimates

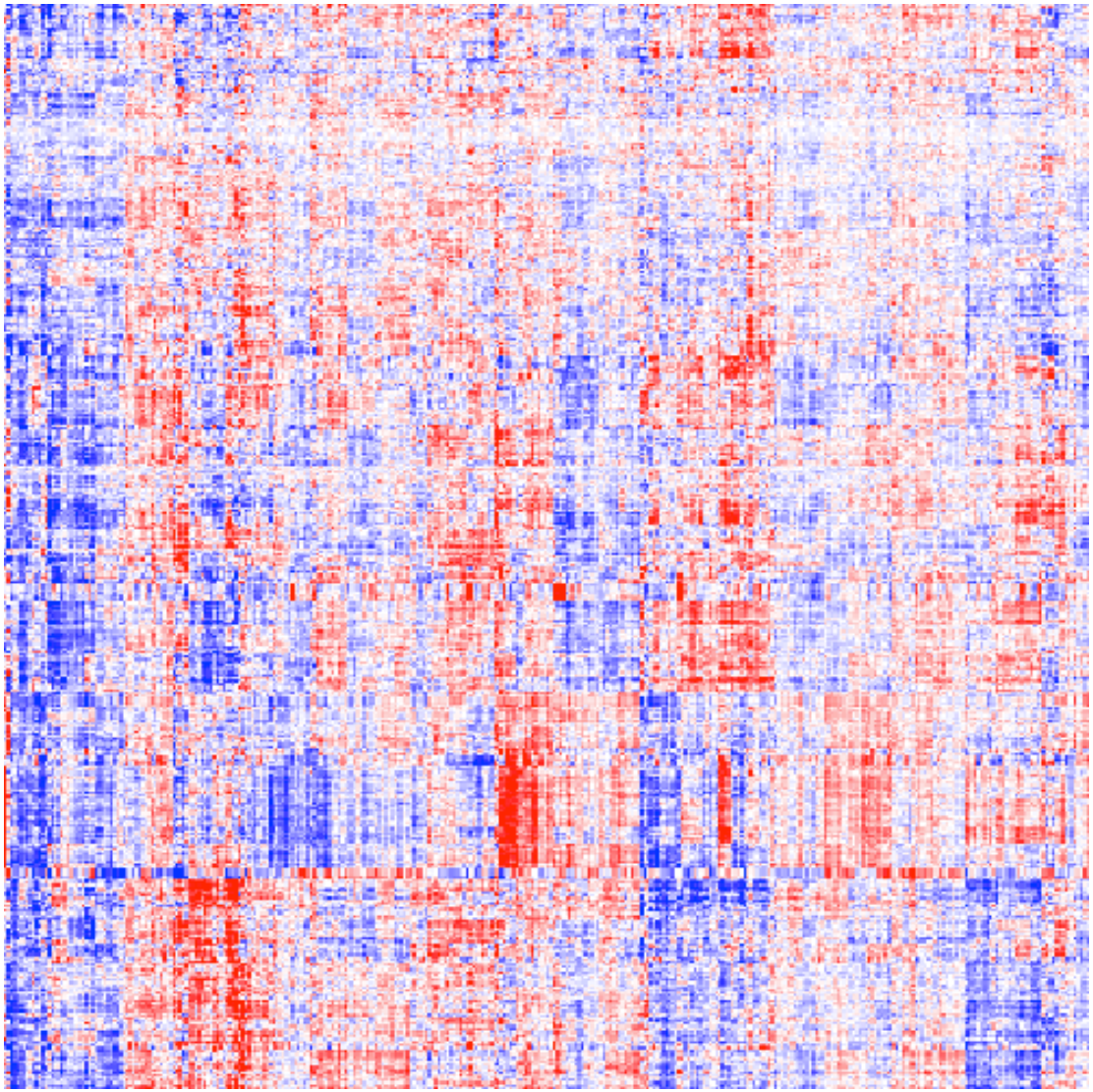


Estimates

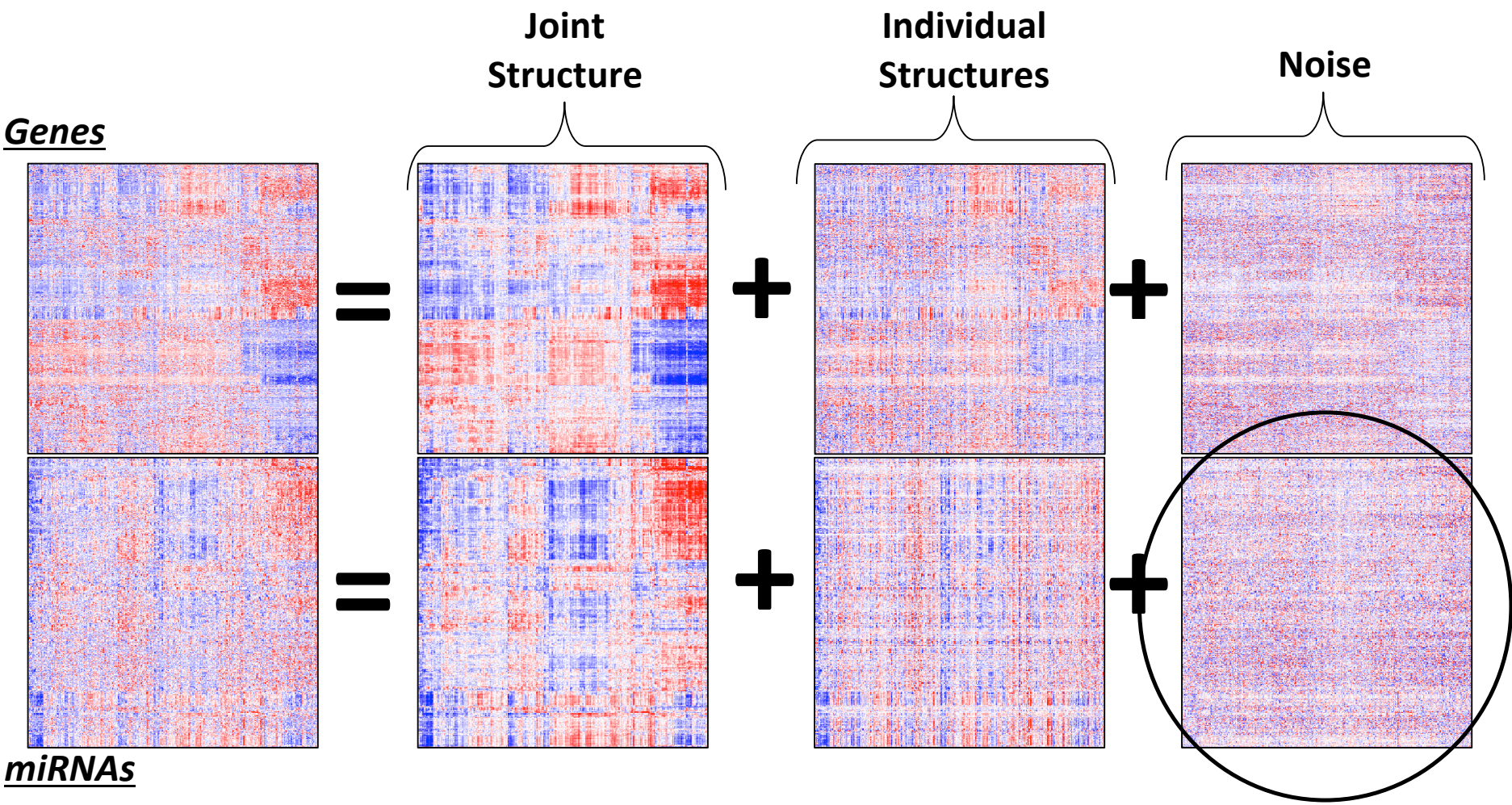

Joint Structure
Individual Structures
Noise
Genes
miRNAs

Estimates

- miRNA individual (reorder rows and columns)

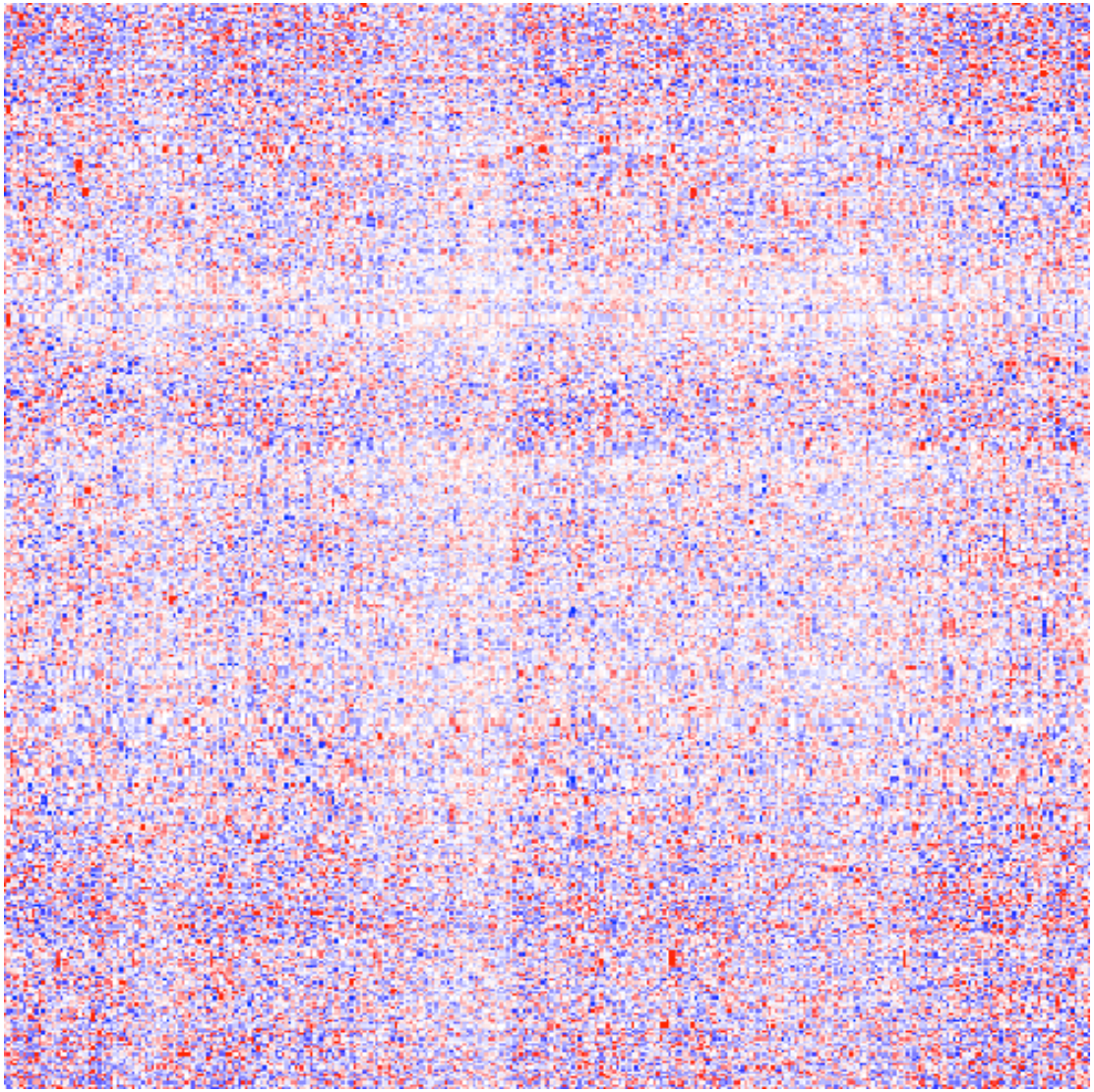


Estimates

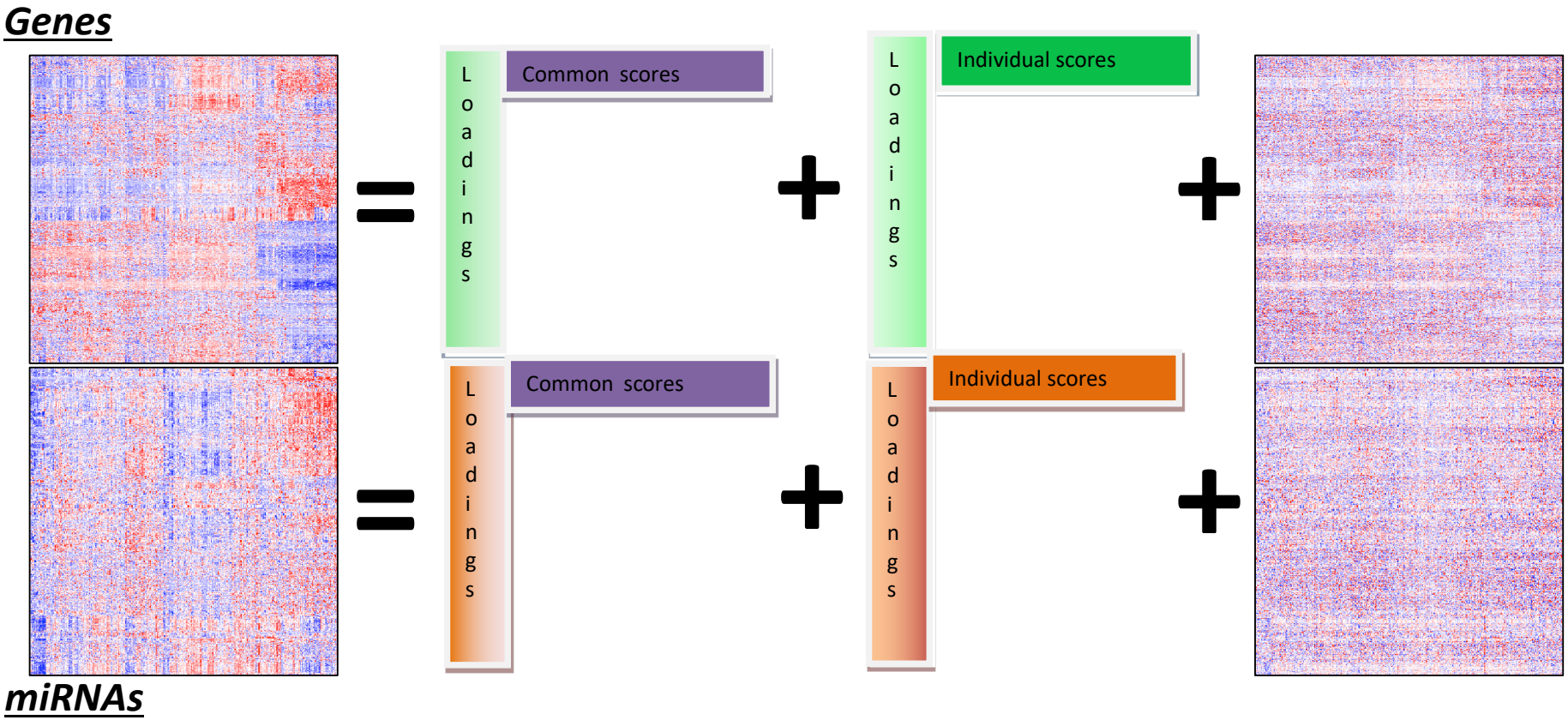

Joint Structure
Individual Structures
Noise
Genes
=
+
+
=
+
+
miRNAs

Estimates

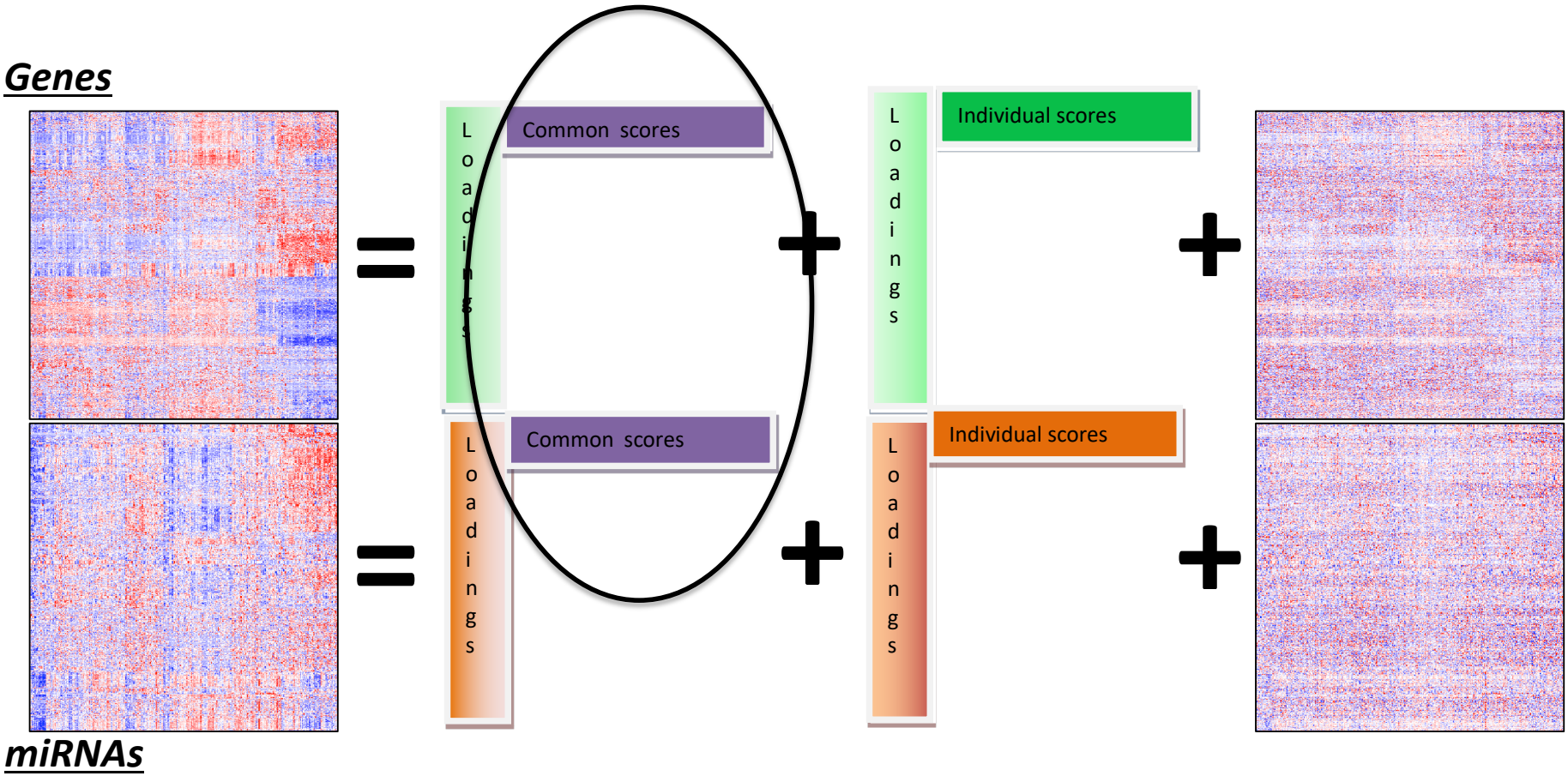
- miRNA error (reorder rows and columns)



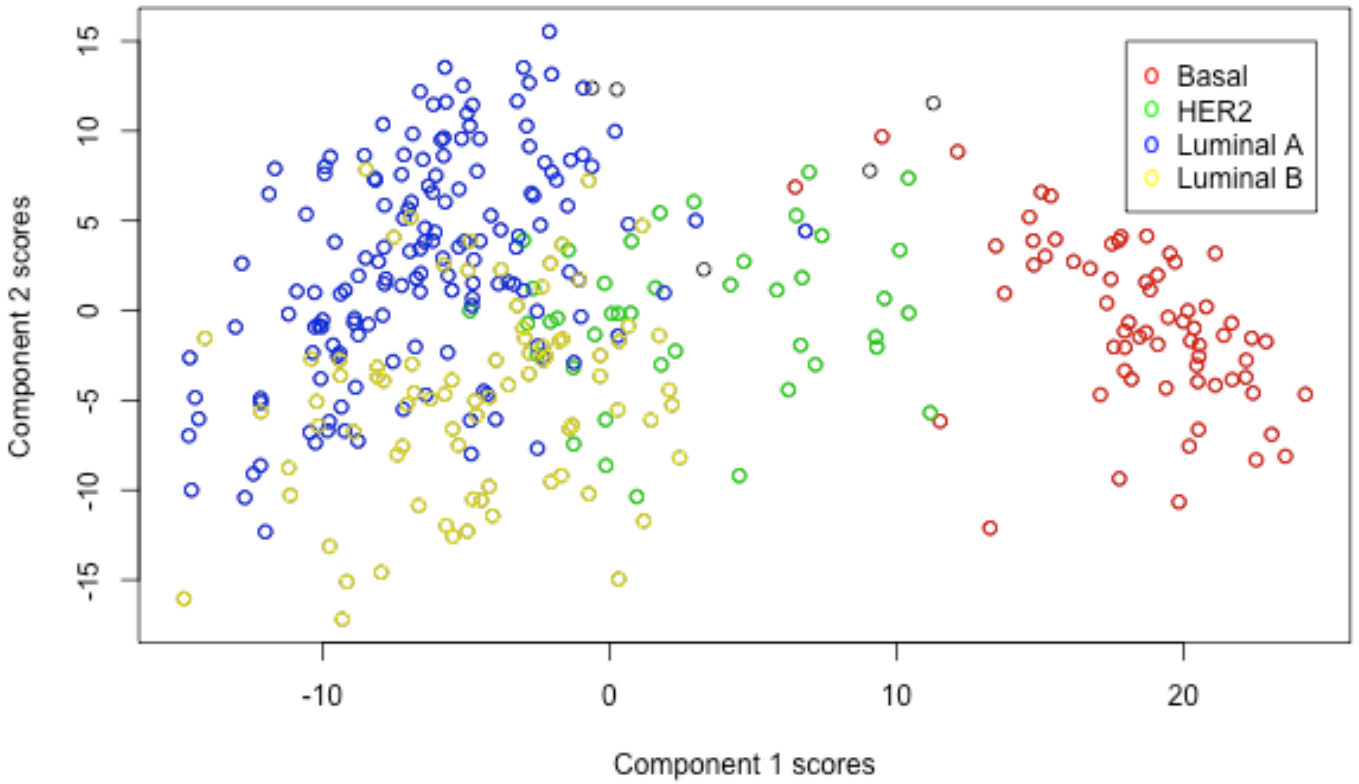
Estimates (factorized)



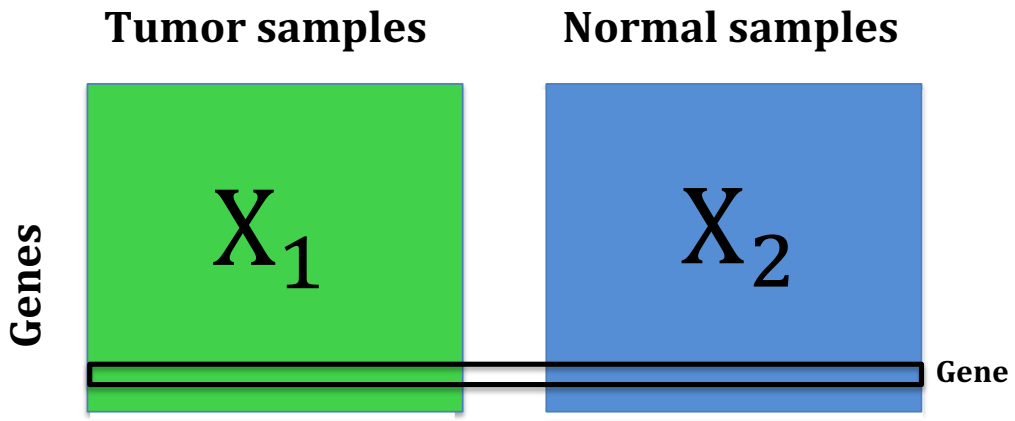
Estimates (factorized)



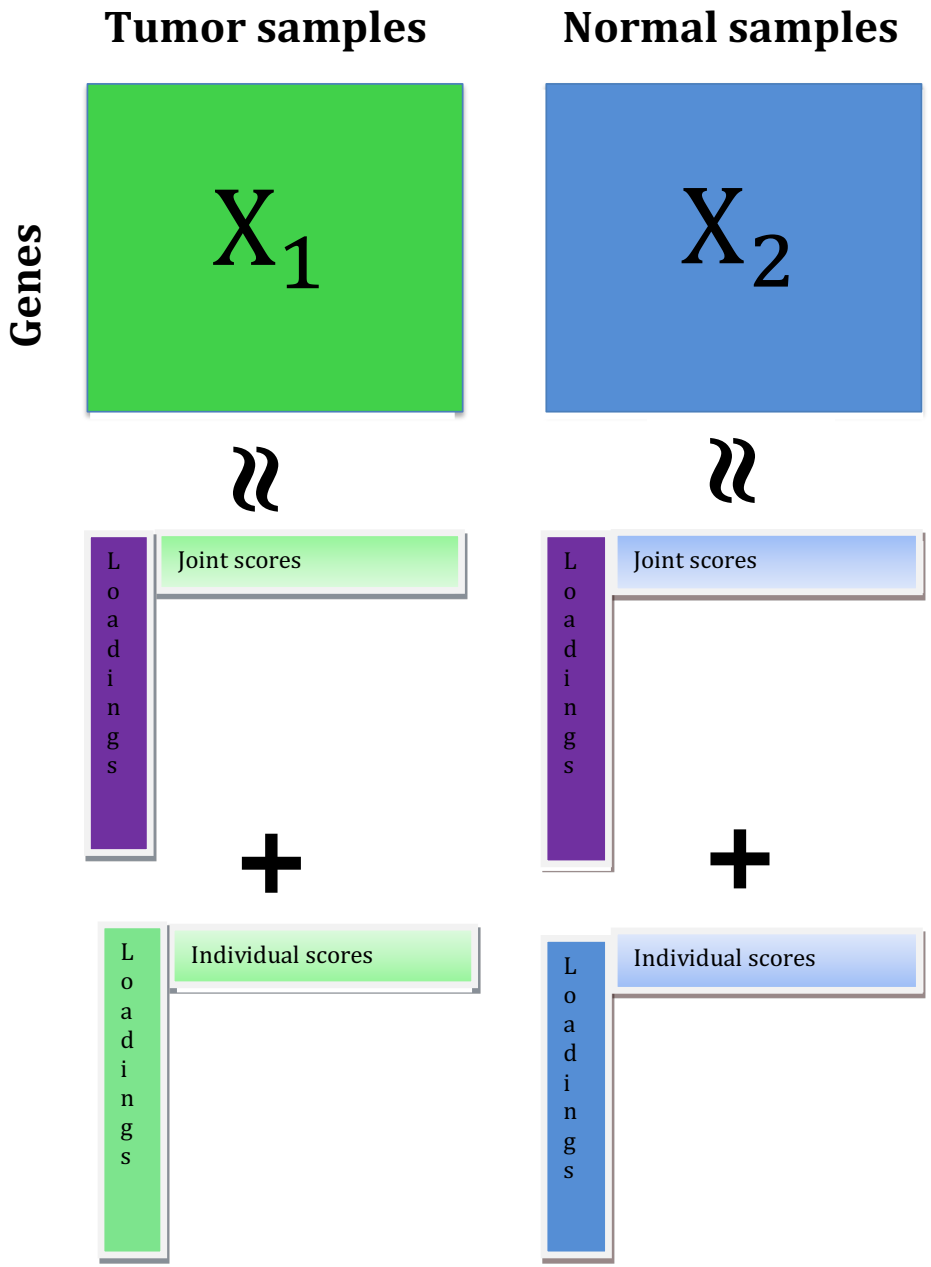
Joint PCs


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

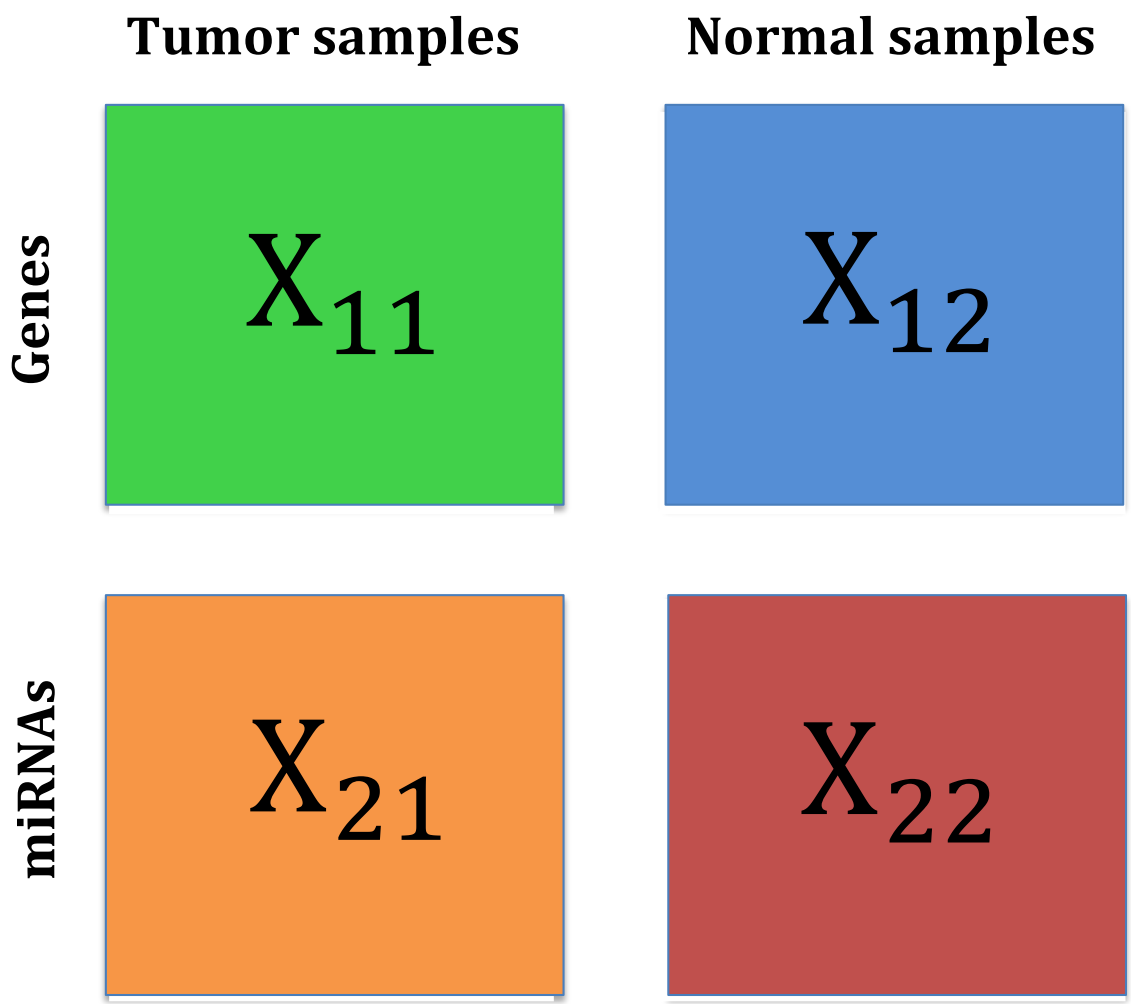
Horizontally linked data



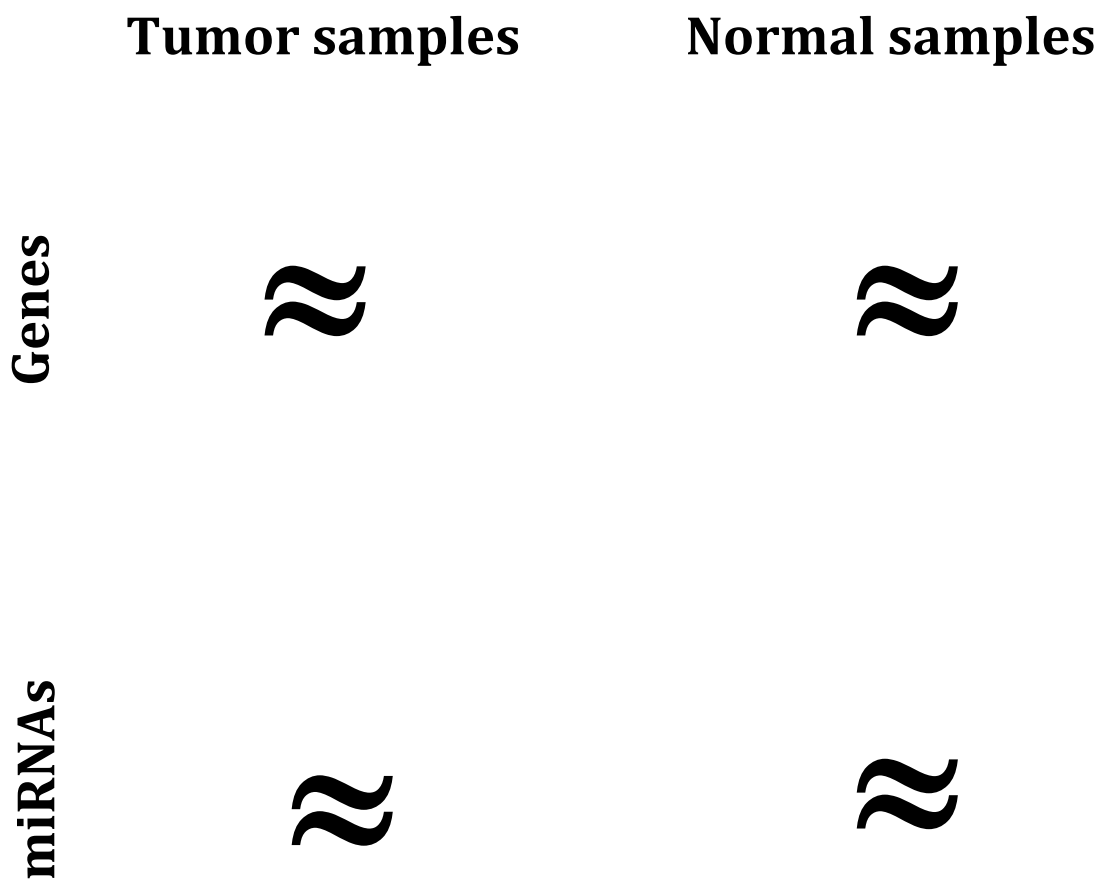
Horizontally linked data: JIVE factorization



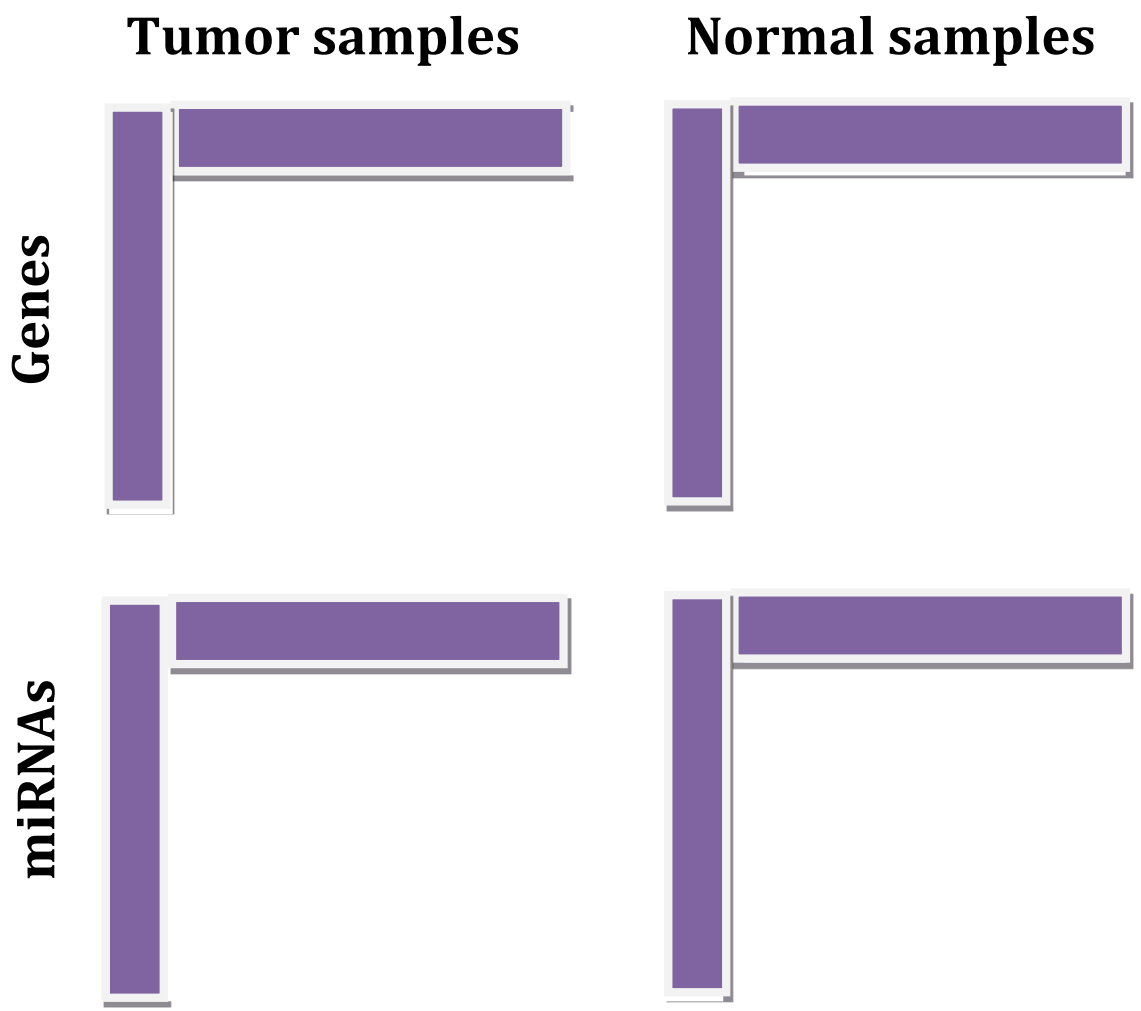
Bidimensionally linked data: BIDIFAC



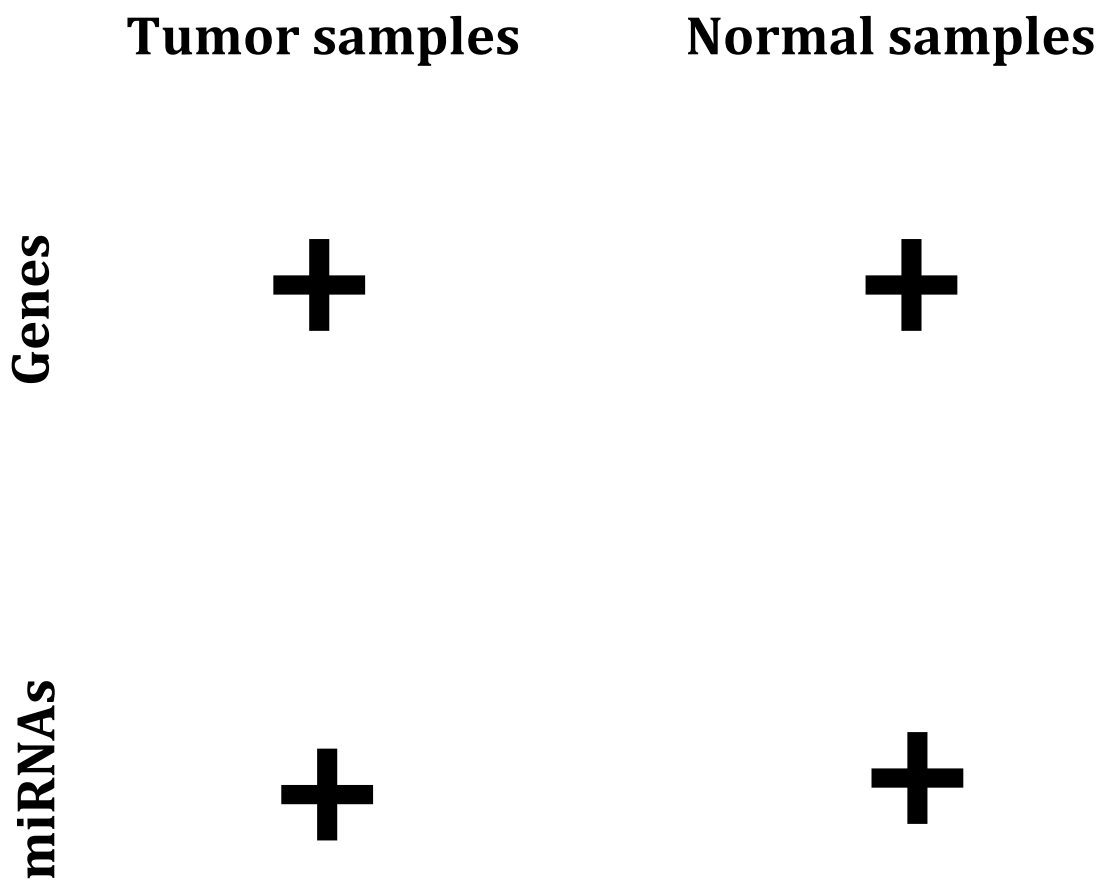
Bidimensionally linked data: BIDIFAC



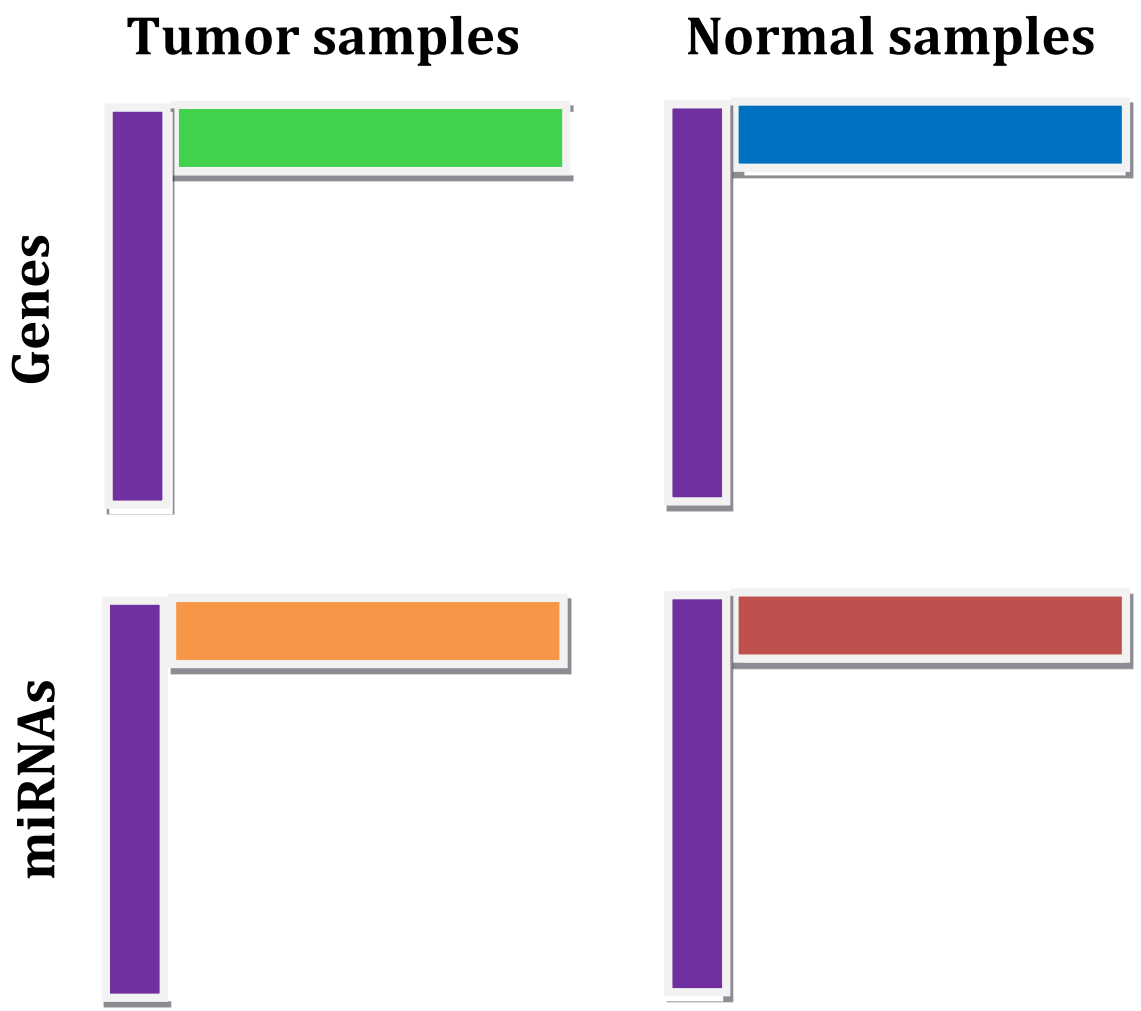
Bidimensionally linked data: BIDIFAC



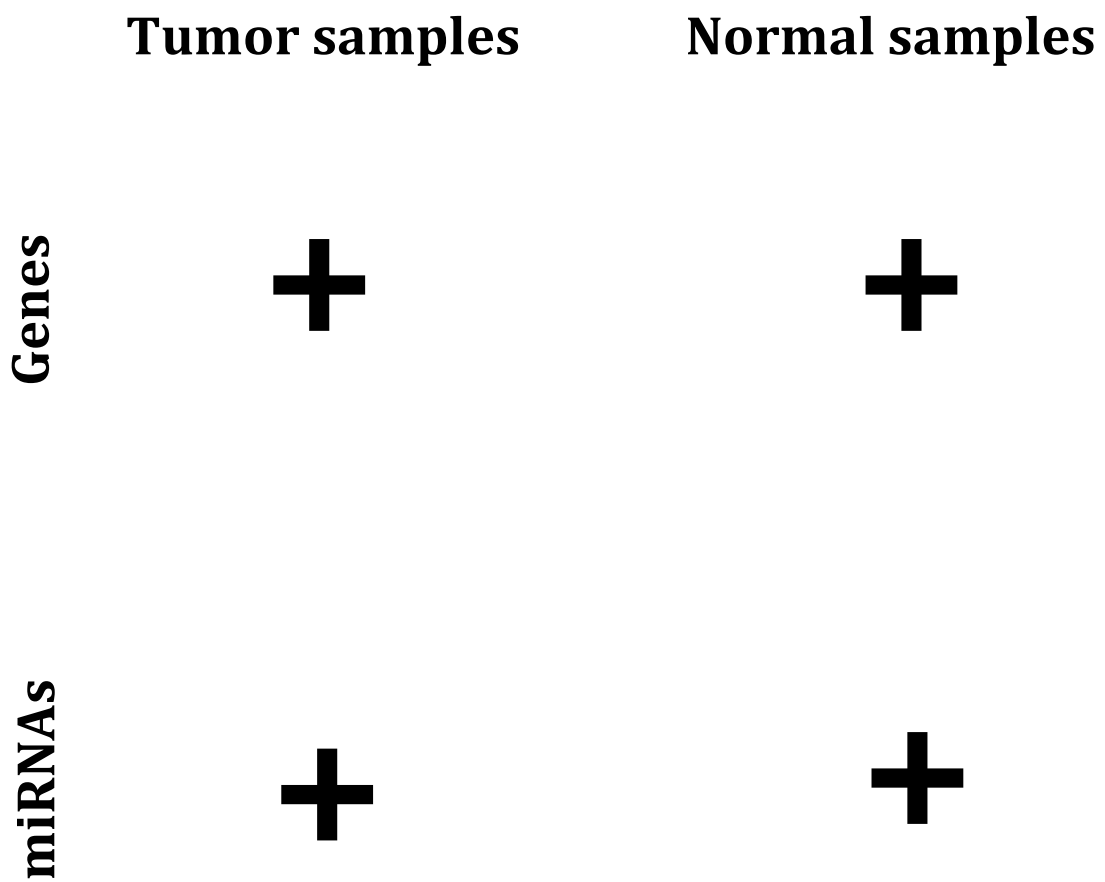
Bidimensionally linked data: BIDIFAC



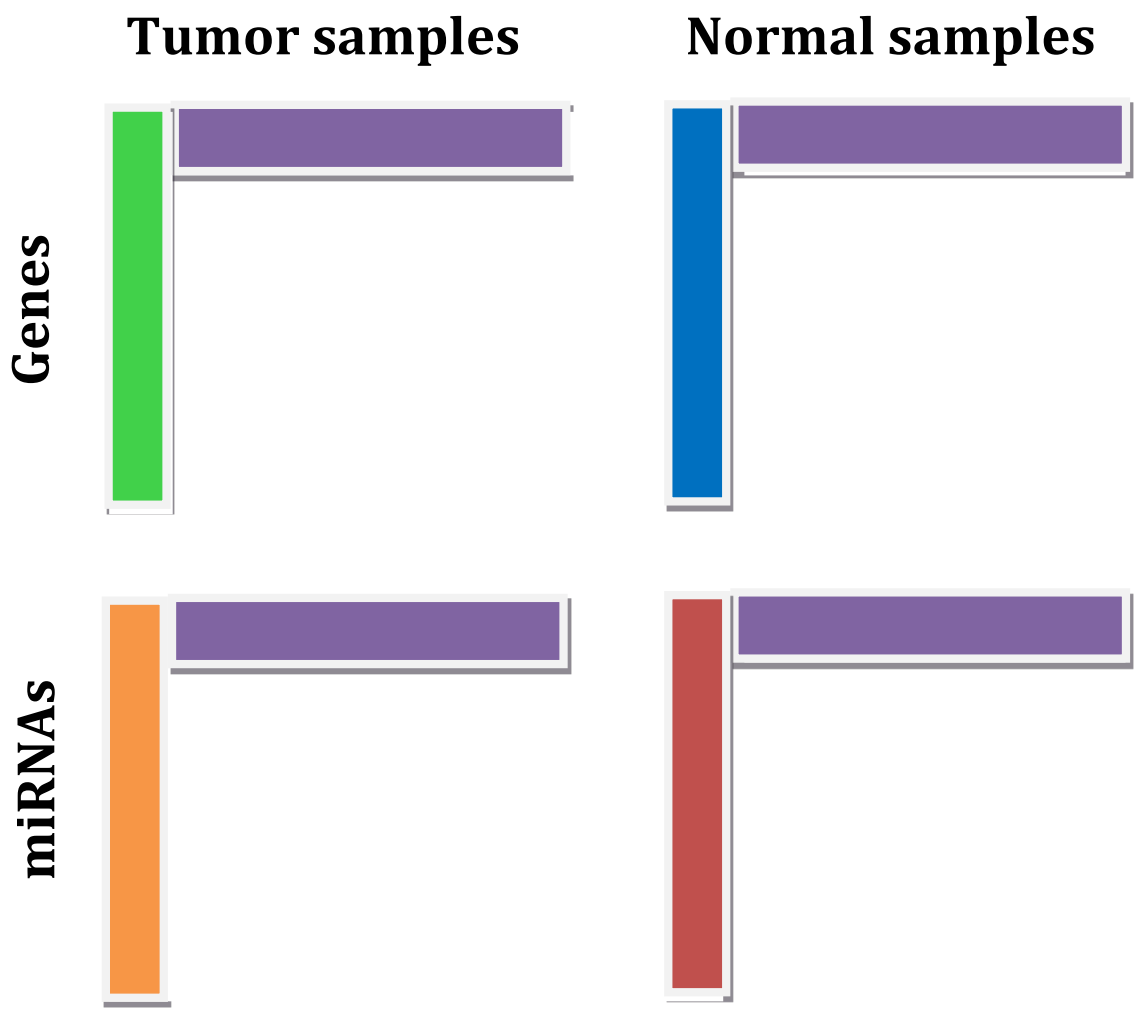
Bidimensionally linked data: BIDIFAC



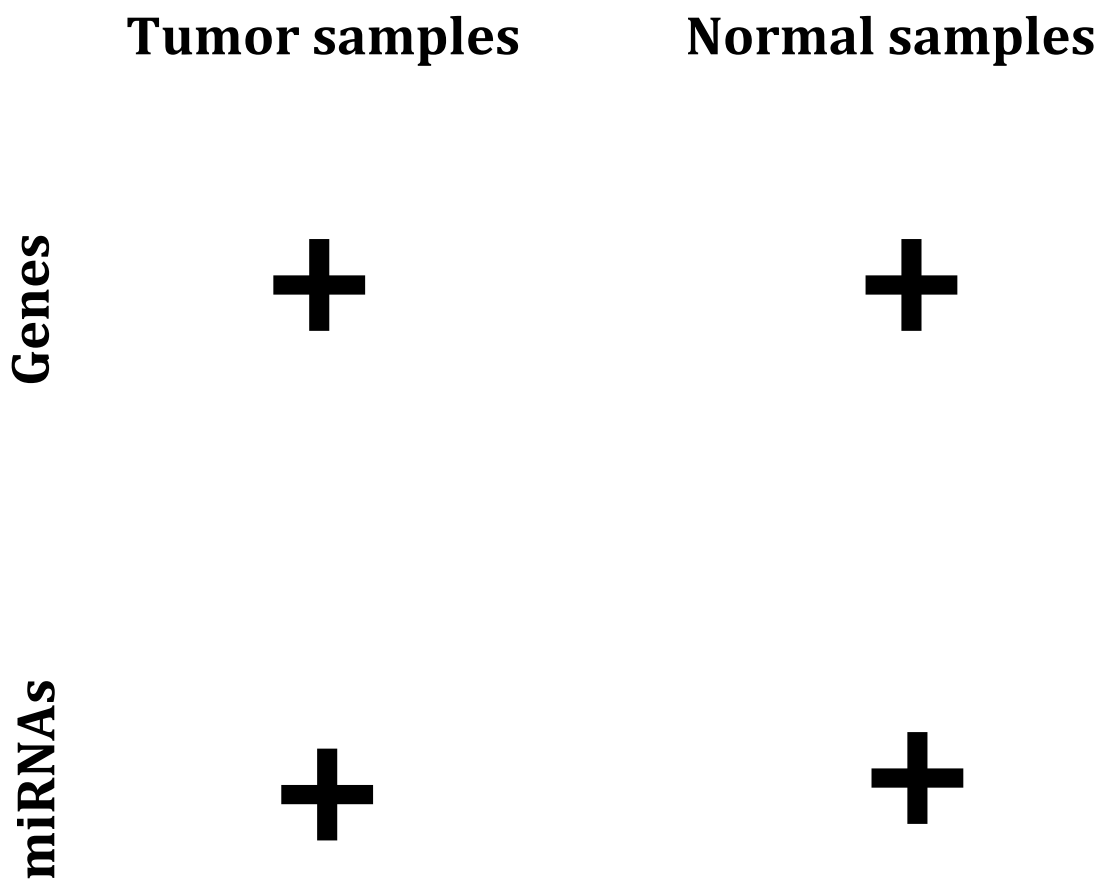
Bidimensionally linked data: BIDIFAC



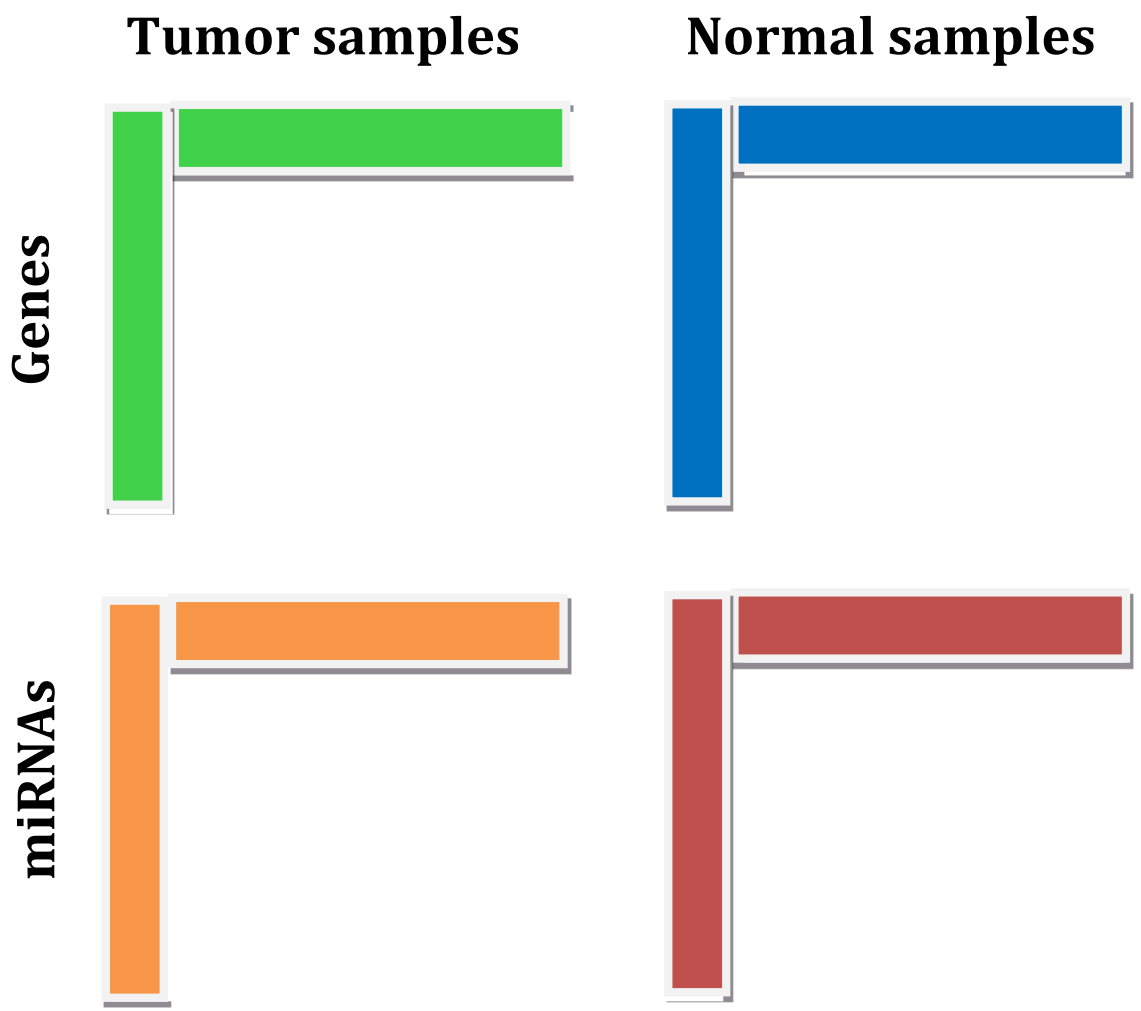
Bidimensionally linked data: BIDIFAC



Bidimensionally linked data: BIDIFAC

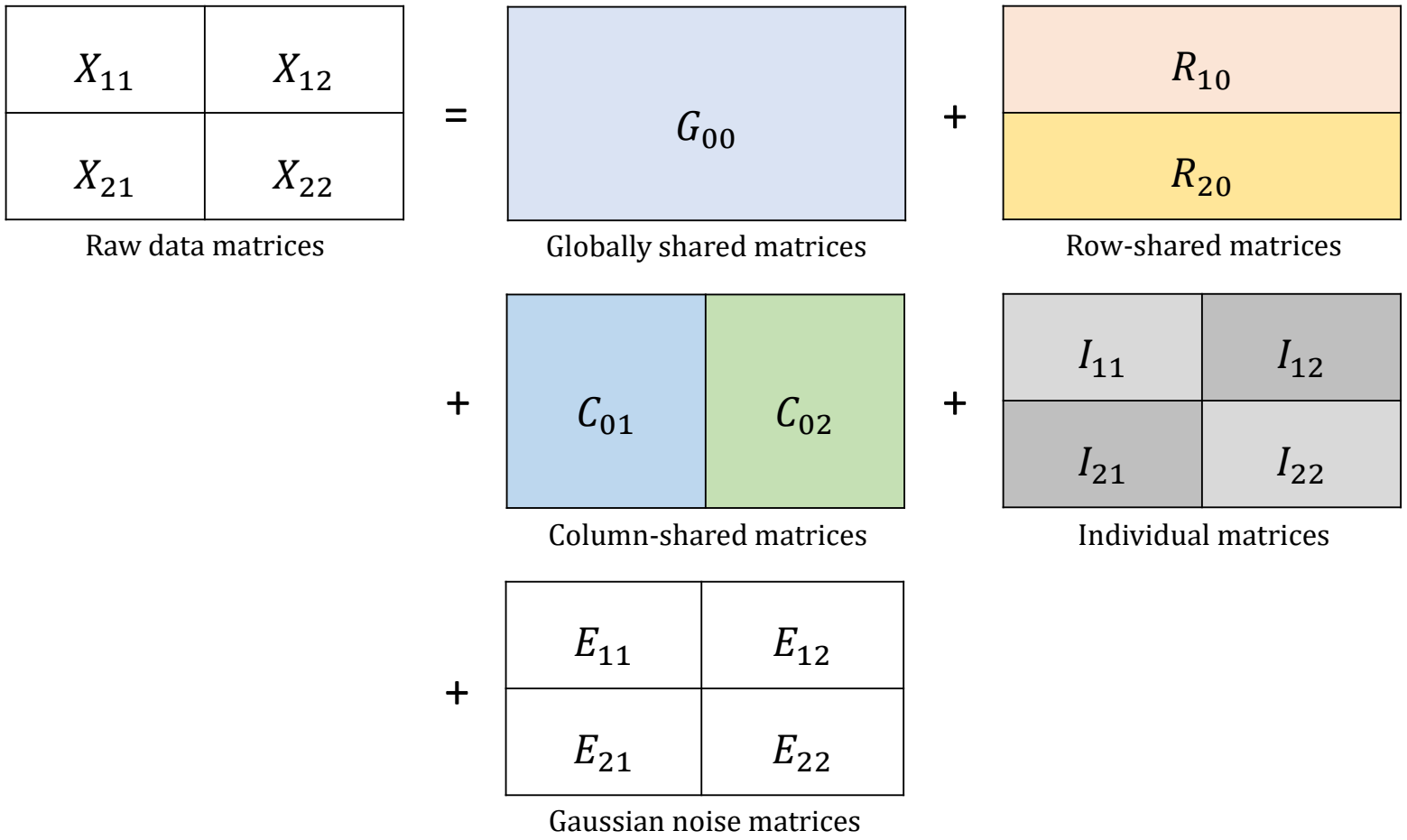


Bidimensionally linked data: BIDIFAC



BIDIFAC: 2×2

Suppose that $X_{ij} = G_{ij} + R_{ij} + C_{ij} + I_{ij} + E_{ij}$, where



G_{00} R_{i0}, C_{0j} and I_{ij} are low-rank.

BIDIFAC: general framework

Consider a set of pq matrices $\{X_{ij} : m_i \times n_j \mid i = 1, \ldots, p,\ j = 1, \ldots, q\}$, which may be concatenated to form the matrix

$$X_{00} = \begin{bmatrix} X_{11} & \cdots & X_{1q} \\ \vdots & \ddots & \vdots \\ X_{p1} & \cdots & X_{pq} \end{bmatrix}$$

$$X_{i0} = [X_{i1}, \cdots, X_{iq}]$$

$$X_{0j} = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{pj} \end{bmatrix}$$

Accordingly, let $m_0 = \sum_{i=1}^{p} m_i$ and $n_0 = \sum_{j=1}^{q} n_j$.

BIDIFAC: objective

- Objective:

$$f_2(\{G_{ij}, R_{ij}, C_{ij}, I_{ij} \mid i = 1, \ldots, p, j = 1, \ldots, q\})$$

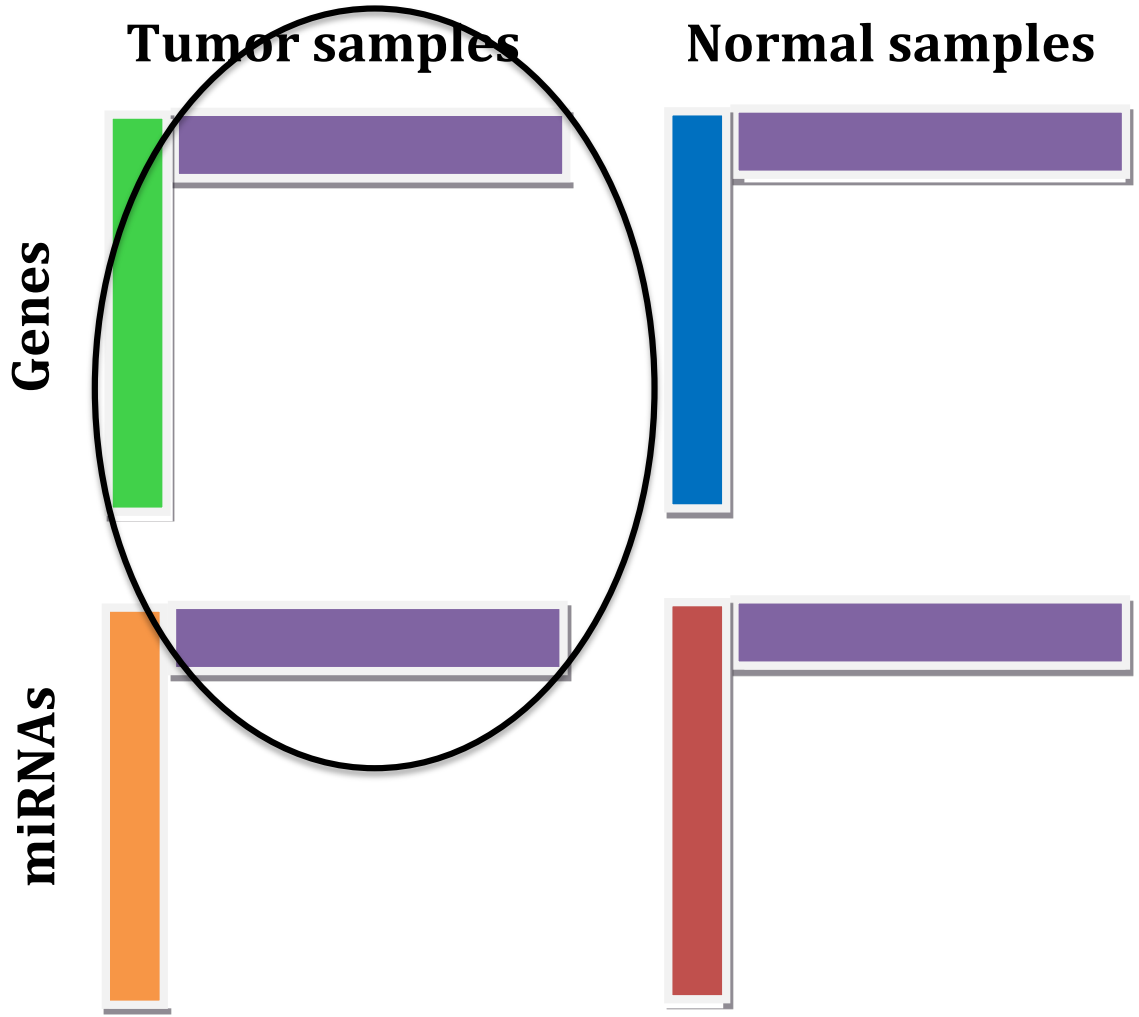
$$= \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{q} ||X_{ij} - G_{ij} - R_{ij} - C_{ij} - I_{ij}||_F^2$$

$$+ \lambda_{00} ||G_{00}||_* + \sum_{i=1}^{p} \lambda_{i0} ||R_{i0}||_* + \sum_{j=1}^{q} \lambda_{0j} ||C_{0j}||_* + \sum_{i=1}^{p} \sum_{j=1}^{q} \lambda_{ij} ||I_{ij}||_*.$$

- Update G_{00} R_{i0}, C_{0j} and I_{ij} until convergence

- Fix penalties
  - $\lambda_{00} = \sqrt{m_0} + \sqrt{n_0}$
  - $\lambda_{i0} = \sqrt{m_i} + \sqrt{n_0}$
  - $\lambda_{0j} = \sqrt{m_0} + \sqrt{n_j}$
  - $\lambda_{ij} = \sqrt{m_i} + \sqrt{n_j}$

TCGA Breast Cancer Data



Tumor-Specific Columns Shared PCs

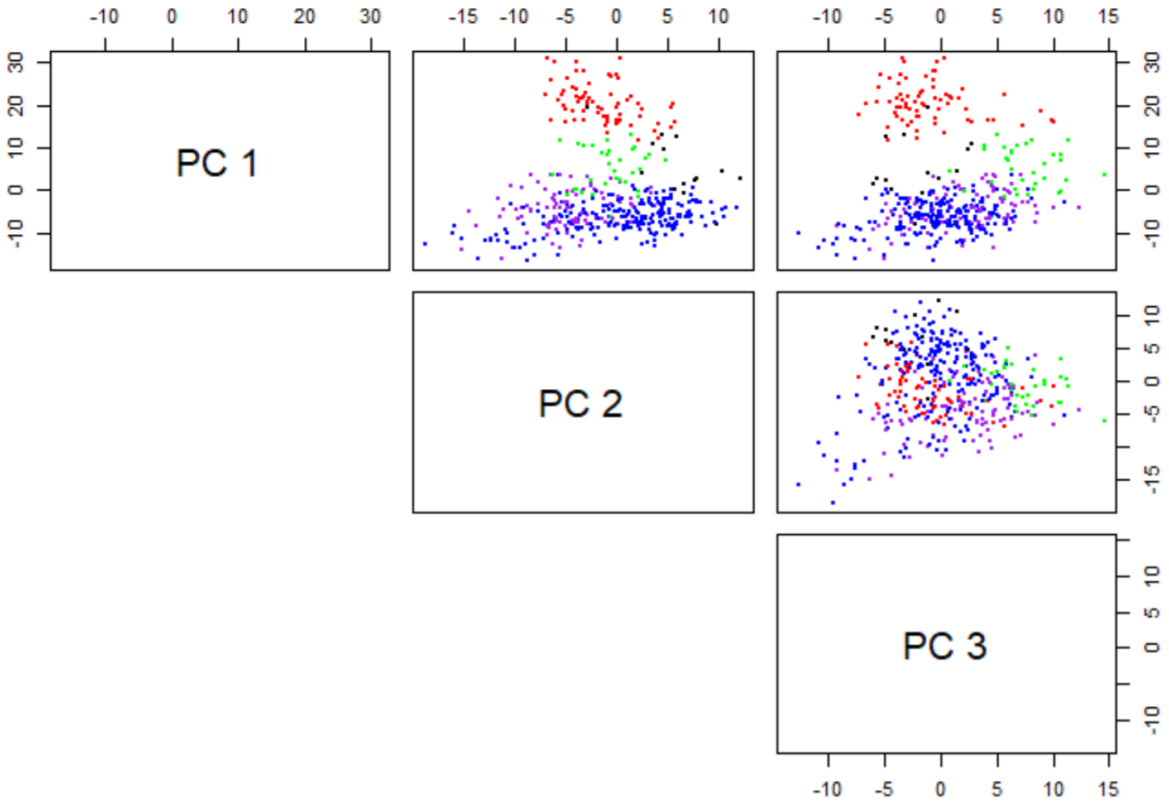
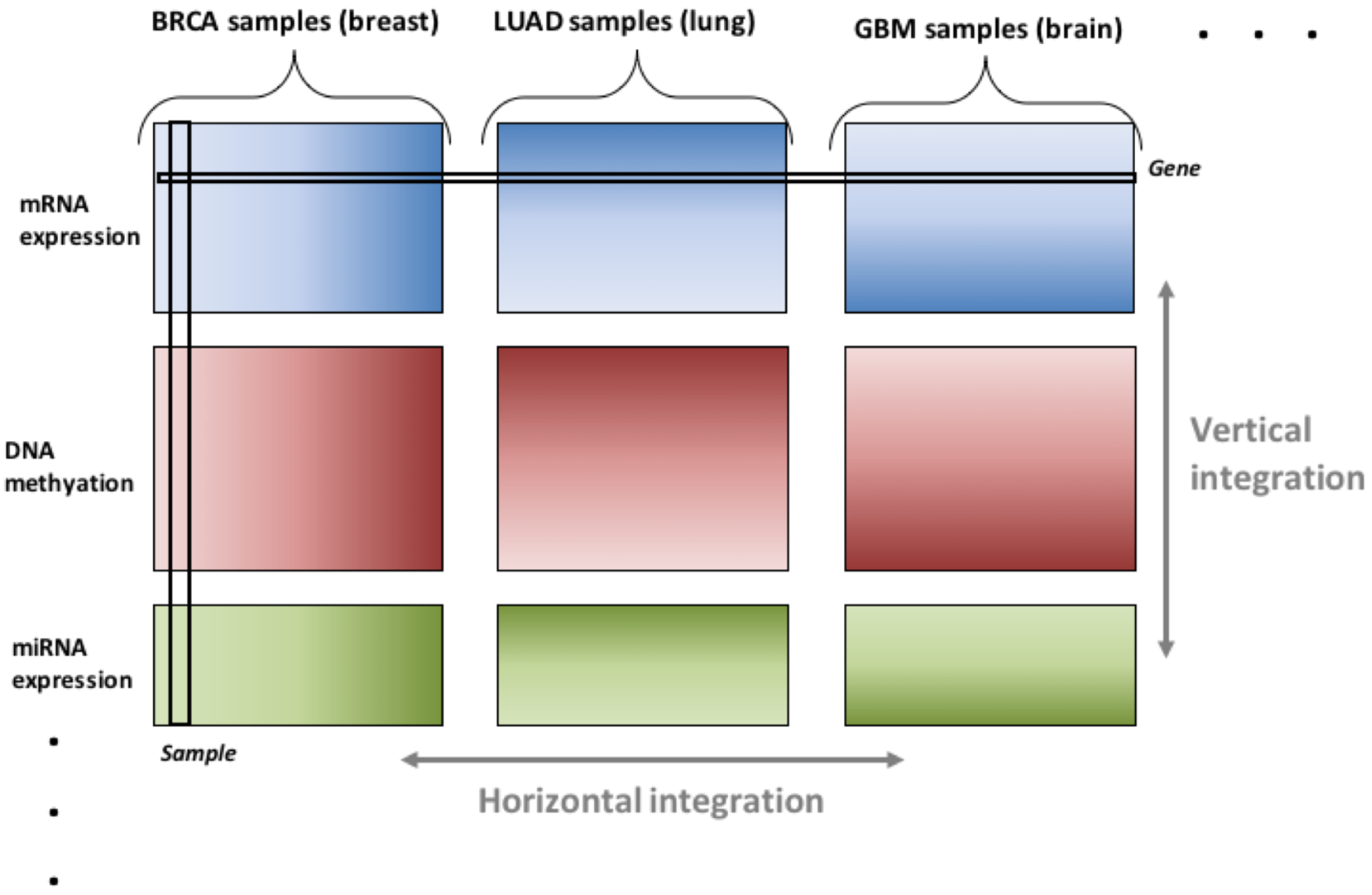


Figure: Principal components of the estimated column-shared structure, colored by subtype: Basal, HER2, Lum A, Lum B.

Pan-omics pan-cancer integration!



Pan-omics pan-cancer data

- Data from 6793 samples representing 29 cancer types:
  - ACC, BLCA, BRCA, CESC, CHOL, CORE, DLBC, ESCA, HNSC, KICH, KIRC, KIRP, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, and UCS.
- Data for 4 different 'omics platforms
  - Gene expression (mRNA), miRNA, DNA methylation, and protein abundance
- Possible shared structures:
  - mRNA+miRNA on BRCA+OV+UCEC
  - mRNA+methylation on KICH+KIRK+KIRP
  - etc..
- $(2^4 - 1) \cdot (2^{29} - 1) = 8053063665$ possible combinations!

BIDIFAC+: general framework

- Decompose X_{00} into structural *modules*:

$$X_{00} = \sum_{k=1}^{\kappa} S_{00}^{(k)} + E_{00}, \tag{1}$$

where

$$S_{00}^{(k)} = \begin{bmatrix} S_{11}^{(k)} & S_{12}^{(k)} & \cdots & S_{1q}^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ S_{p1}^{(k)} & S_{p2}^{(k)} & \cdots & S_{pq}^{(k)} \end{bmatrix}$$

and the presence of each $S_{ij}^{(k)}$ is determined by a binary matrix of row indicators $R : p \times \kappa$ and column indicators $C : q \times \kappa$:

$$S_{ij}^{(k)} = \begin{cases} 0_{M_i \times N_j} & \text{if } R[i,k] = 0 \text{ or } C[j,k] = 0 \\ U_i^{(k)} V_j^{(k)} & \text{if } R[i,k] = 1 \text{ and } C[j,k] = 1 \end{cases}.$$

BIDIFAC+: objective

- Minimize the following objective over R, C, and $\{S_{00}^{(k)}\}_{k=1}$:

$$||X_{00} - \sum_{k=1}^{\kappa} S_{00}^{(k)}||_F^2 + \sum_{k=1}^{\kappa} \lambda_k ||S_{00}^{(k)}||_*$$

where

$$\lambda_k = \sqrt{\sum_{i=1}^{p} R[i,k] m_i} + \sqrt{\sum_{j=1}^{q} C[j,k] n_j}$$

- In practice $\kappa < (2^p - 1) \cdot (2^q - 1)$
  - Only some possible modules are non-zero

BIDIFAC+: Uniqueness

- Let $\mathbb{S}_{\hat{X}}$ be the set of possible decompositions for $\hat{X}_{00}$:

$$\mathbb{S}_{\hat{X}} = \left\{ \{S_{00}^{(k)}\}_{k=1}^{K} \mid \hat{X}_{00} = \sum_{k=1}^{K} S_{00}^{(k)} \right\}.$$

Theorem

Consider $\{\hat{S}_{00}^{(k)}\}_{k=1}^{K} \in \mathbb{S}_{\hat{X}}$ *and let* $U_0^{(k)} \hat{D}^{(k)} V_0^{(k)T}$ *give the SVD of* $\hat{S}_{00}^{(k)}$. *The following three properties uniquely identify* $\{\hat{S}_{00}^{(k)}\}_{k=1}^{K}$.

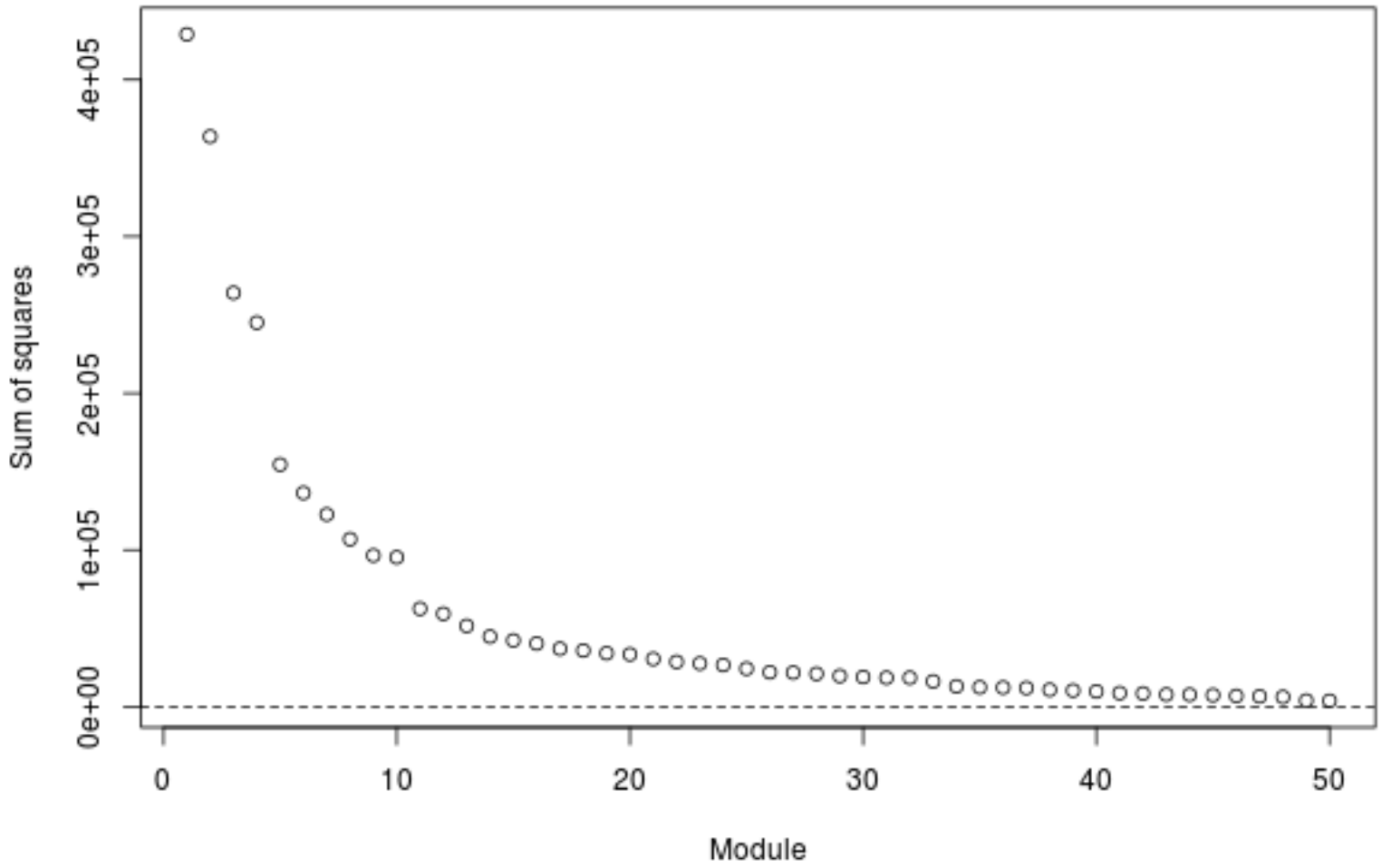
- $\{\hat{S}_{00}^{(k)}\}_{k=1}^{K}$ *minimizes* $\sum_{k=1}^{\kappa} \lambda_k ||S_{00}^{(k)}||_*$ *over* $\mathbb{S}_{\hat{X}}$,
- $\{\hat{U}_i^{(k)}[\bullet, r] : R[i,k] = 1 \text{ and } \hat{D}^{(k)}[r,r] > 0\}$ *are linearly independent for* $i = 1, \ldots p$,
- $\{\hat{V}_j^{(k)}[\bullet, r] : C[j,k] = 1 \text{ and } \hat{D}^{(k)}[r,r] > 0\}$ *are linearly independent for* $j = 1, \ldots, q$.

BIDIFAC: posterior mode

- Assume error entries E_{00} are iid $N(0, 1)$
- Entries of $U_i^{(k)}, V_j^{(k)}$ have iid $N(0, 1/\lambda_k)$ priors
- Then, the solution to the structured nuclear norm objective gives the posterior mode.

BIDIFAC+: results

- Variance explained for each structural module $S_{00}^{(k)}$:



BIDIFAC+: results

- Top structural modules, ranked by variance explained:

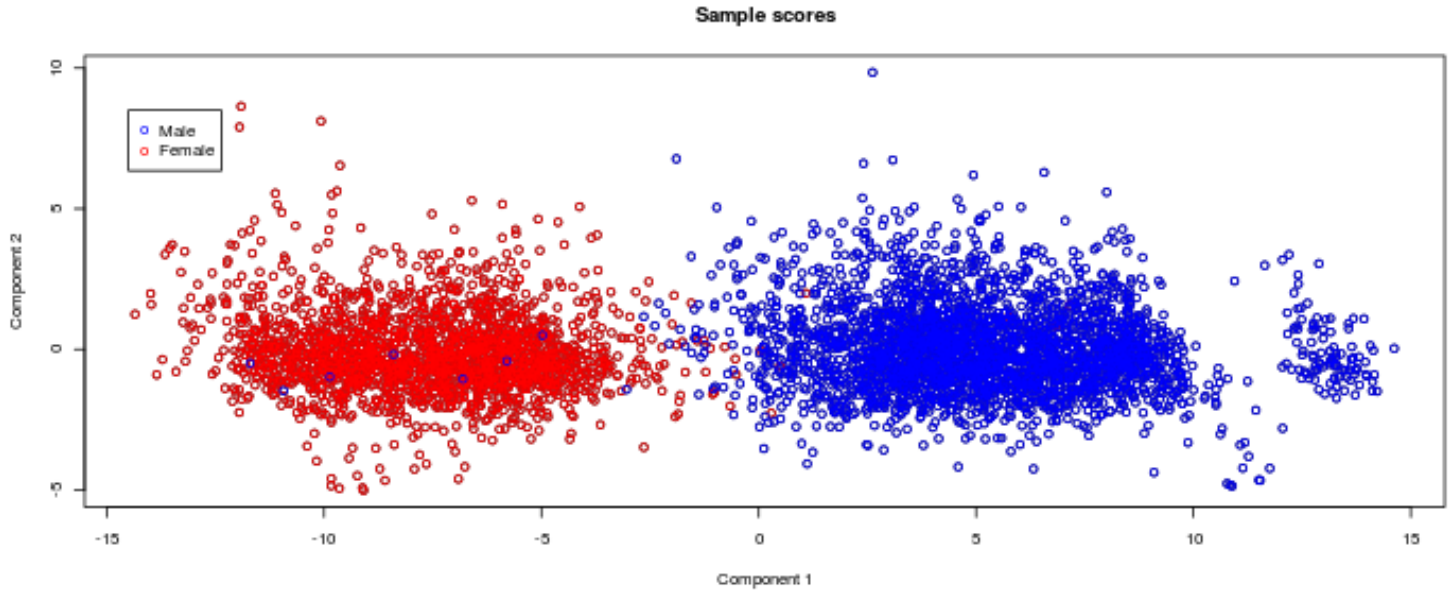
▶

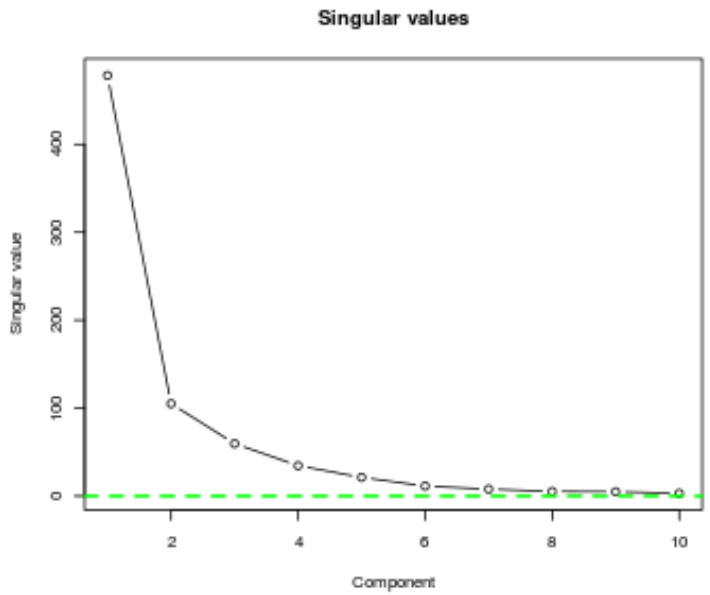
Module	Cancer types	Omics sources
1	All cancers	mRNA miRNA Meth Protein
2	All cancers	miRNA
3	BLCA BRCA CESC CHOL CORE DLBC ESCA HNSC LIHC LUAD LUSC OV PAAD PRAD SKCM STAD TGCT UCEC UCS	Meth
4	ACC BLCA CHOL CORE DLBC ESCA HNSC KICH KIRC KIRP LGG LIHC LUAD LUSC MESO PAAD PCPG SARC SKCM STAD THCA THYM	mRNA Meth
5	All cancers	mRNA
6	BRCA	mRNA miRNA Meth Protein
7	LGG	mRNA miRNA Protein
8	All cancers *but* LGG	Protein
9	THCA	mRNA miRNA Protein
10	All cancers *but* LGG and TGCT	miRNA
11	CHOL KIRC KIRP LIHC	mRNA miRNA Meth Protein
12	LGG	Meth
13	BLCA CESC CORE ESCA HNSC LUSC SARC STAD	mRNA miRNA Meth Protein
14	KICH KIRC KIRP	mRNA miRNA Protein
15	BLCA BRCA CESC CHOL ESCA HNSC LUAD LUSC PAAD PRAD SKCM STAD TGCT UCEC UCS	mRNA miRNA

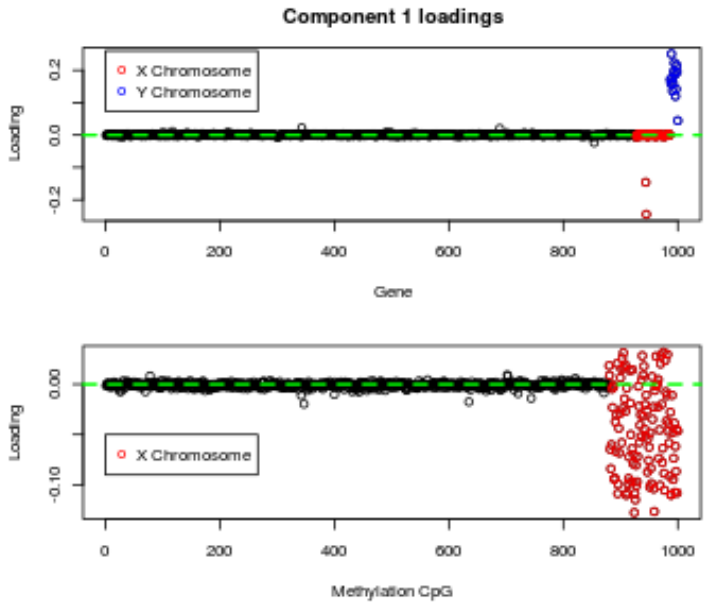
BIDIFAC+: results

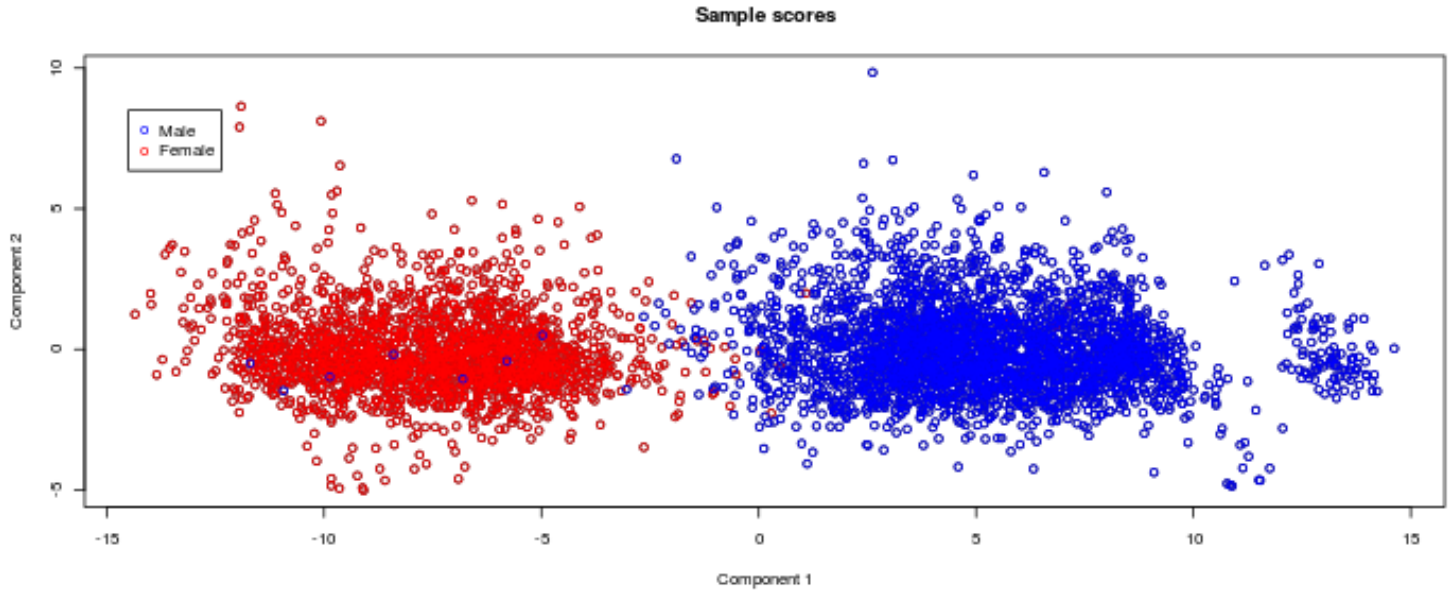
- Top structural modules, ranked by variance explained:

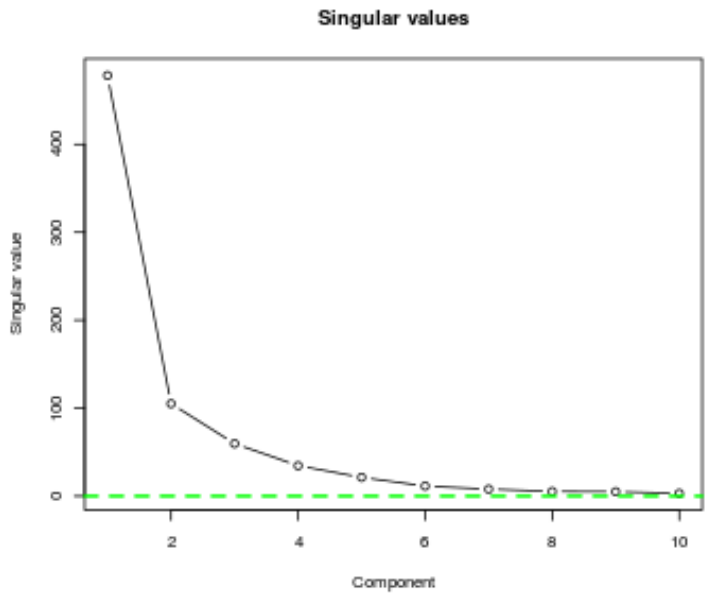
Module	Cancer types	Omics sources
1	All cancers	mRNA miRNA Meth Protein
2	All cancers	miRNA
3	BLCA BRCA CESC CHOL CORE DLBC ESCA HNSC LIHC LUAD LUSC OV PAAD PRAD SKCM STAD TGCT UCEC UCS	Meth
4	ACC BLCA CHOL CORE DLBC ESCA HNSC KICH KIRC KIRP LGG LIHC LUAD LUSC MESO PAAD PCPG SARC SKCM STAD THCA THYM	mRNA Meth
5	All cancers	mRNA
6	BRCA	mRNA miRNA Meth Protein
7	LGG	mRNA miRNA Protein
8	All cancers *but* LGG	Protein
9	THCA	mRNA miRNA Protein
10	All cancers *but* LGG and TGCT	miRNA
11	CHOL KIRC KIRP LIHC	mRNA miRNA Meth Protein
12	LGG	Meth
13	BLCA CESC CORE ESCA HNSC LUSC SARC STAD	mRNA miRNA Meth Protein
14	KICH KIRC KIRP	mRNA miRNA Protein
15	BLCA BRCA CESC CHOL ESCA HNSC LUAD LUSC PAAD PRAD SKCM STAD TGCT UCEC UCS	mRNA miRNA

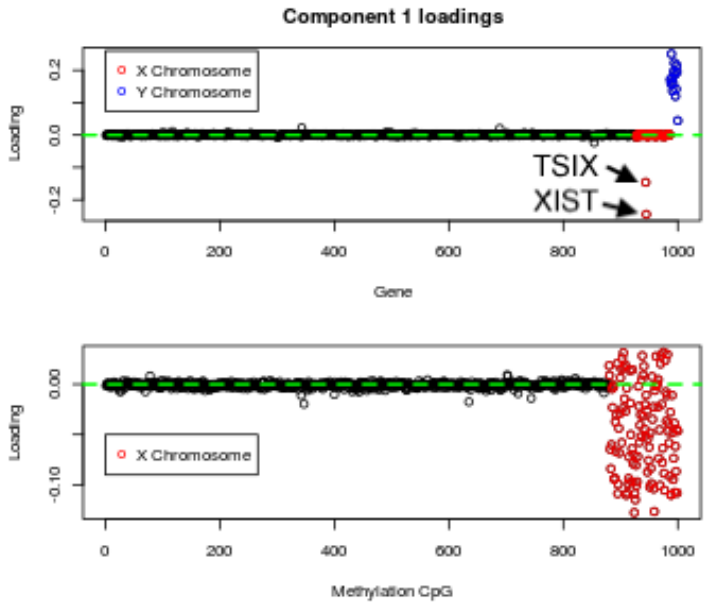

Sample scores
Male
Female
Component 2
Component 1


Singular values
Singular value
Component


Component 1 loadings
X Chromosome
Y Chromosome
Loading
Gene
X Chromosome
Loading
Methylation CpG


Sample scores
Male
Female
Component 2
Component 1


Singular values
Singular value
Component

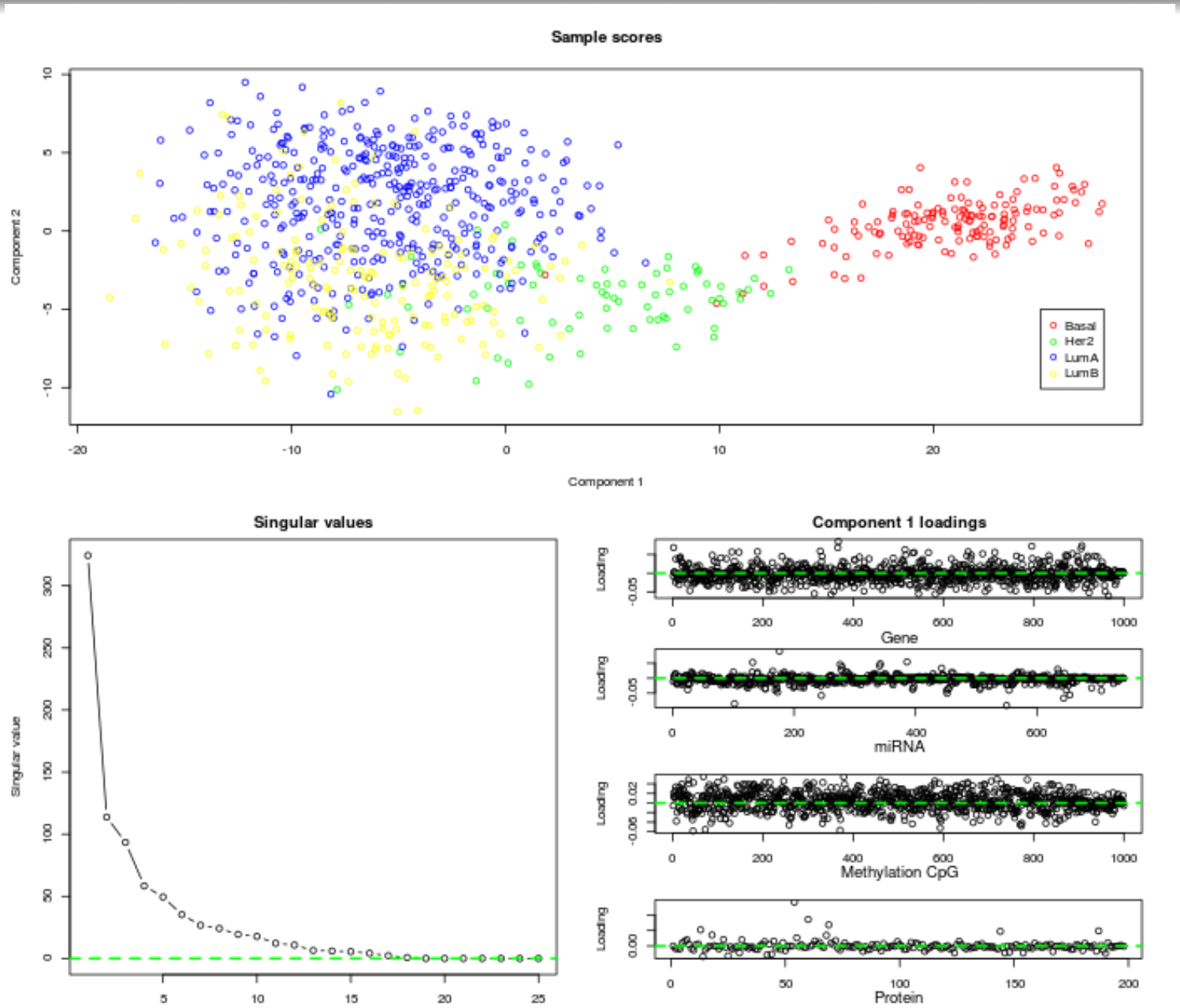

Component 1 loadings
X Chromosome
Y Chromosome
TSIX
XIST
Loading
Gene
Methylation CpG

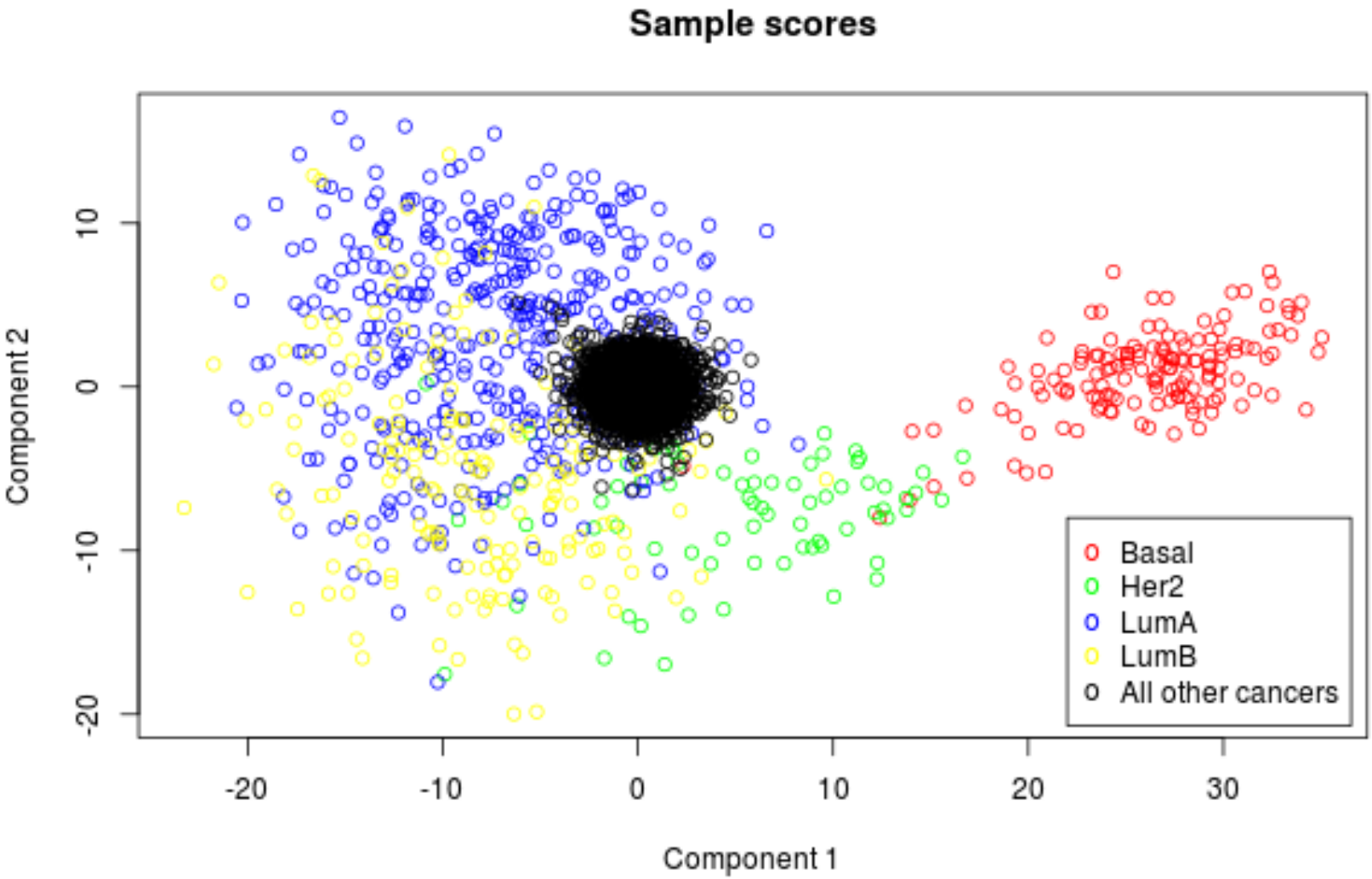
BIDIFAC+: results

- Top structural modules, ranked by variance explained:

▶

Module	Cancer types	Omics sources
1	All cancers	mRNA miRNA Meth Protein
2	All cancers	miRNA
3	BLCA BRCA CESC CHOL CORE DLBC ESCA HNSC LIHC LUAD LUSC OV PAAD PRAD SKCM STAD TGCT UCEC UCS	Meth
4	ACC BLCA CHOL CORE DLBC ESCA HNSC KICH KIRC KIRP LGG LIHC LUAD LUSC MESO PAAD PCPG SARC SKCM STAD THCA THYM	mRNA Meth
5	All cancers	mRNA
6	BRCA	mRNA miRNA Meth Protein
7	LGG	mRNA miRNA Protein
8	All cancers *but* LGG	Protein
9	THCA	mRNA miRNA Protein
10	All cancers *but* LGG and TGCT	miRNA
11	CHOL KIRC KIRP LIHC	mRNA miRNA Meth Protein
12	LGG	Meth
13	BLCA CESC CORE ESCA HNSC LUSC SARC STAD	mRNA miRNA Meth Protein
14	KICH KIRC KIRP	mRNA miRNA Protein
15	BLCA BRCA CESC CHOL ESCA HNSC LUAD LUSC PAAD PRAD SKCM STAD TGCT UCEC UCS	mRNA miRNA


Sample scores
Component 2
Component 1
Basal
Her2
LumA
LumB
Singular values
Singular value
Component 1 loadings
Loading
Gene
miRNA
Methylation CpG
Protein


Sample scores
Component 2
10
0
-10
-20
-20
-10
0
10
20
30
Component 1
Basal
Her2
LumA
LumB
All other cancers

Thank you!

- Support: NCI grant R21CA231214-01

- References:
  - **BIDIFAC**: J Park & EF Lock. Integrative Factorization of Bidimensionally Linked Matrices. *Biometrics*, 76 (1): 61-74 2020.
  - **BIDIFAC+**: EF Lock, J Park & KA Hoadley. Bidimensional linked matrix factorization for pan-omics pan-cancer analysis. *Annals of Applied Statistics*, 16 (1): 193-215, 2022.

- Code:
  - https://github.com/lockEF/bidifac