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

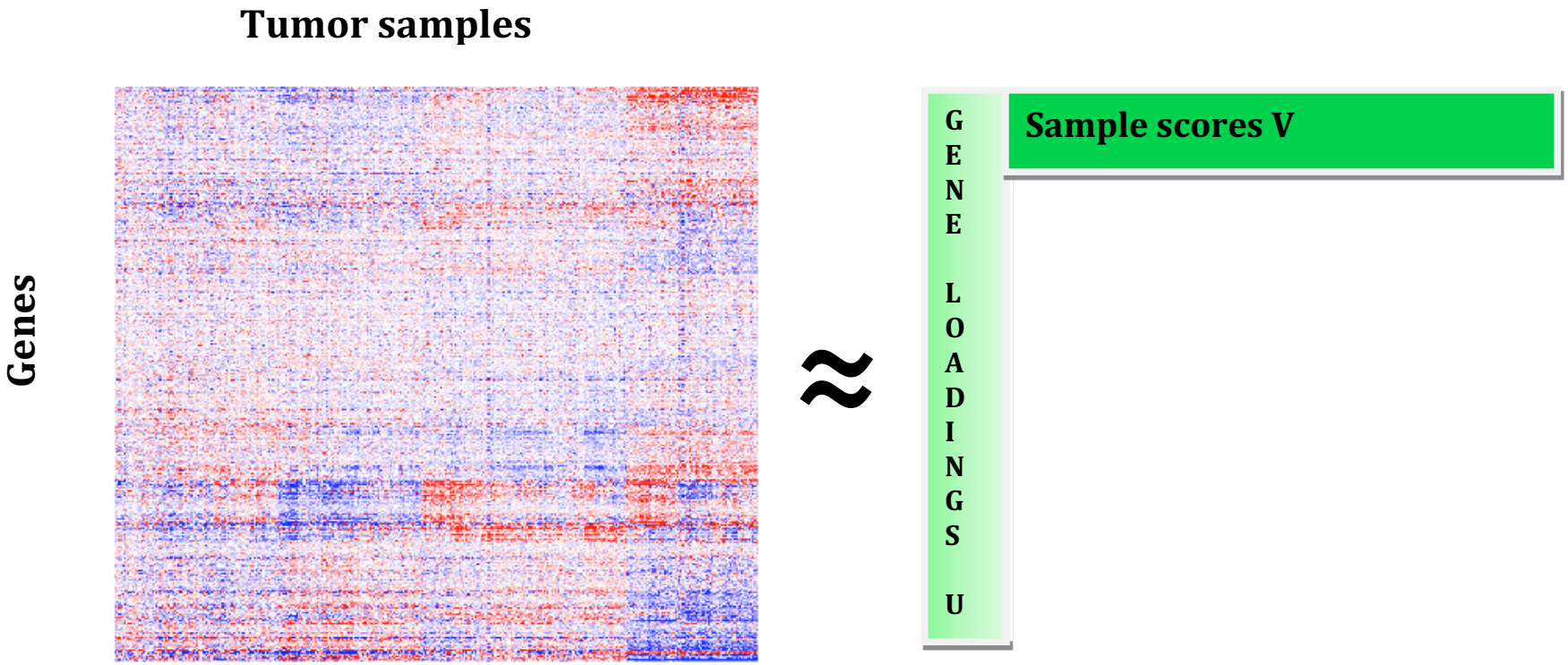
Eric F. Lock

University of Minnesota, Division of Biostatistics

JSM Virtual, 08/06/2020

Matrix factorization

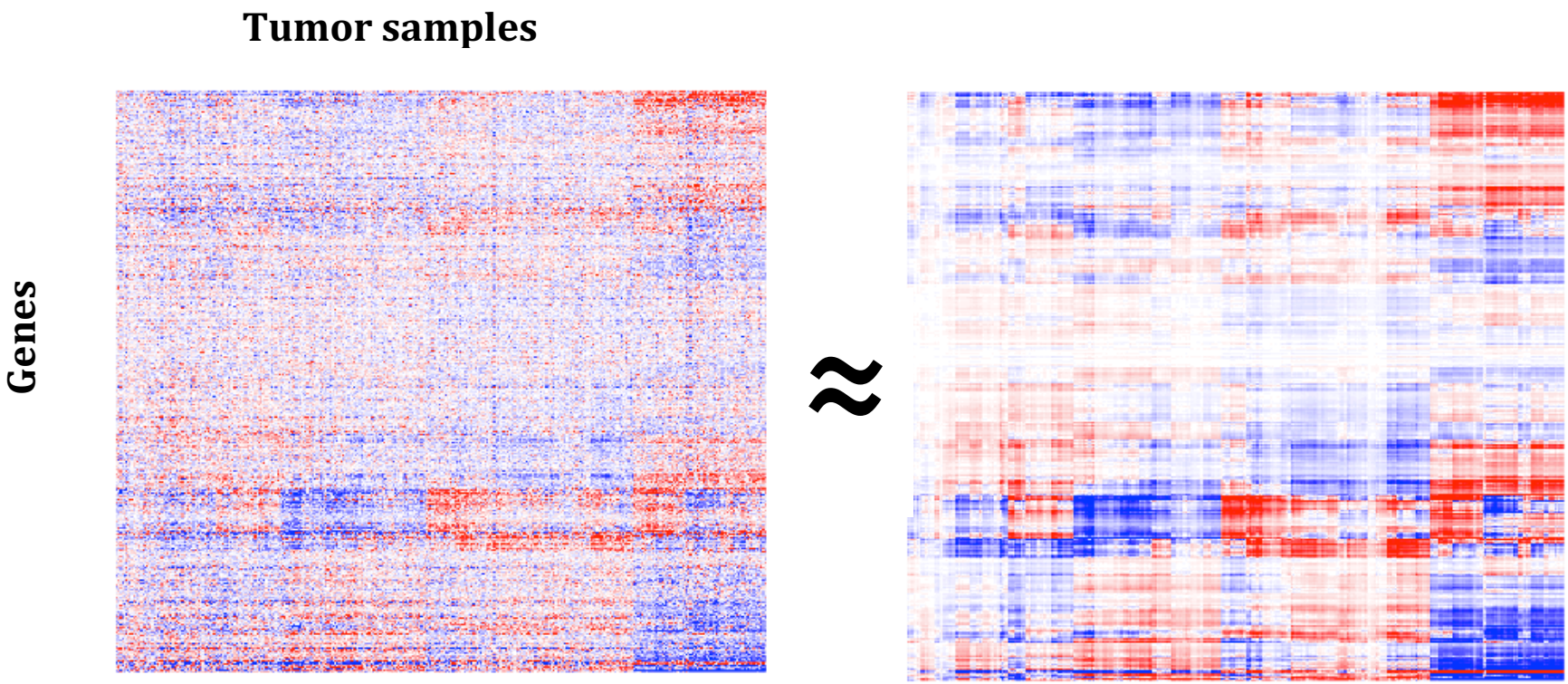
- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



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

Matrix factorization (r=3)

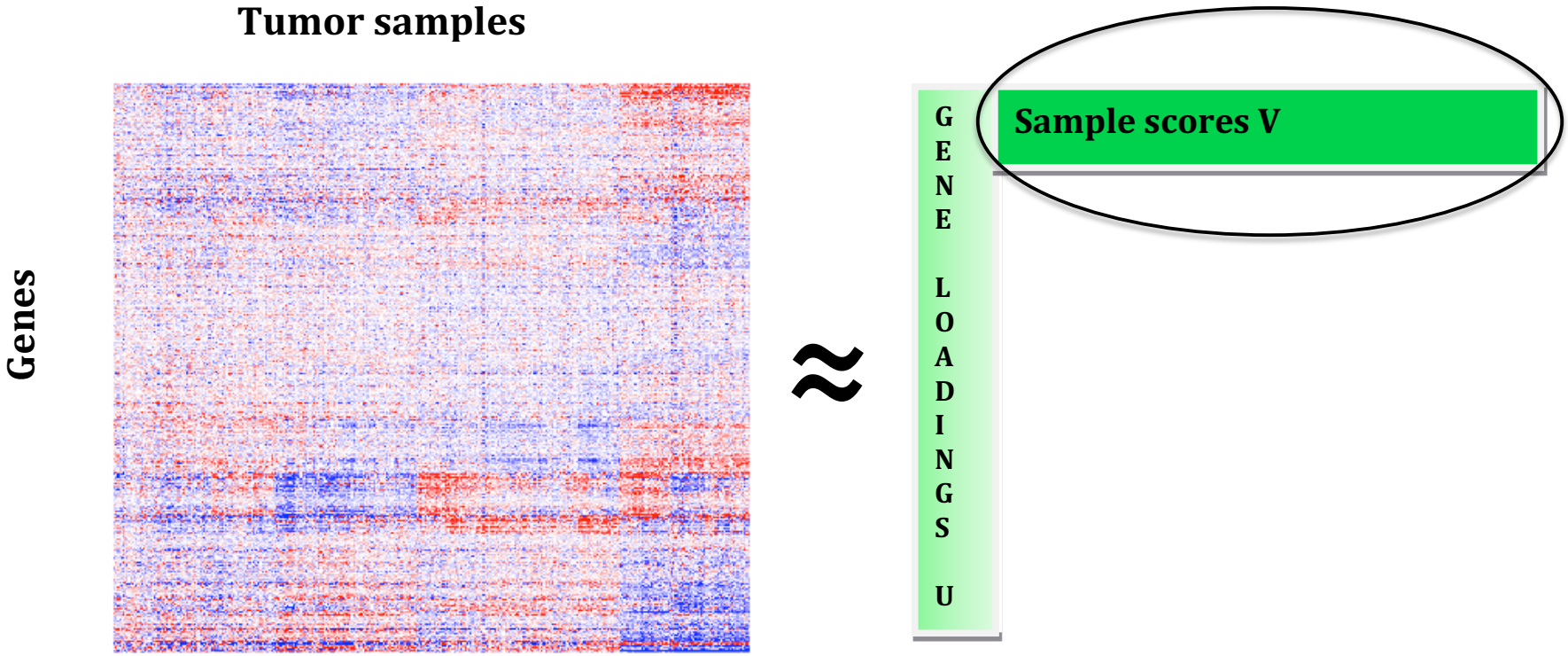
- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



- Low rank factorization: $X \approx UV$, $U : m \times 3$, $V : 3 \times n$.

Matrix factorization (r=3)

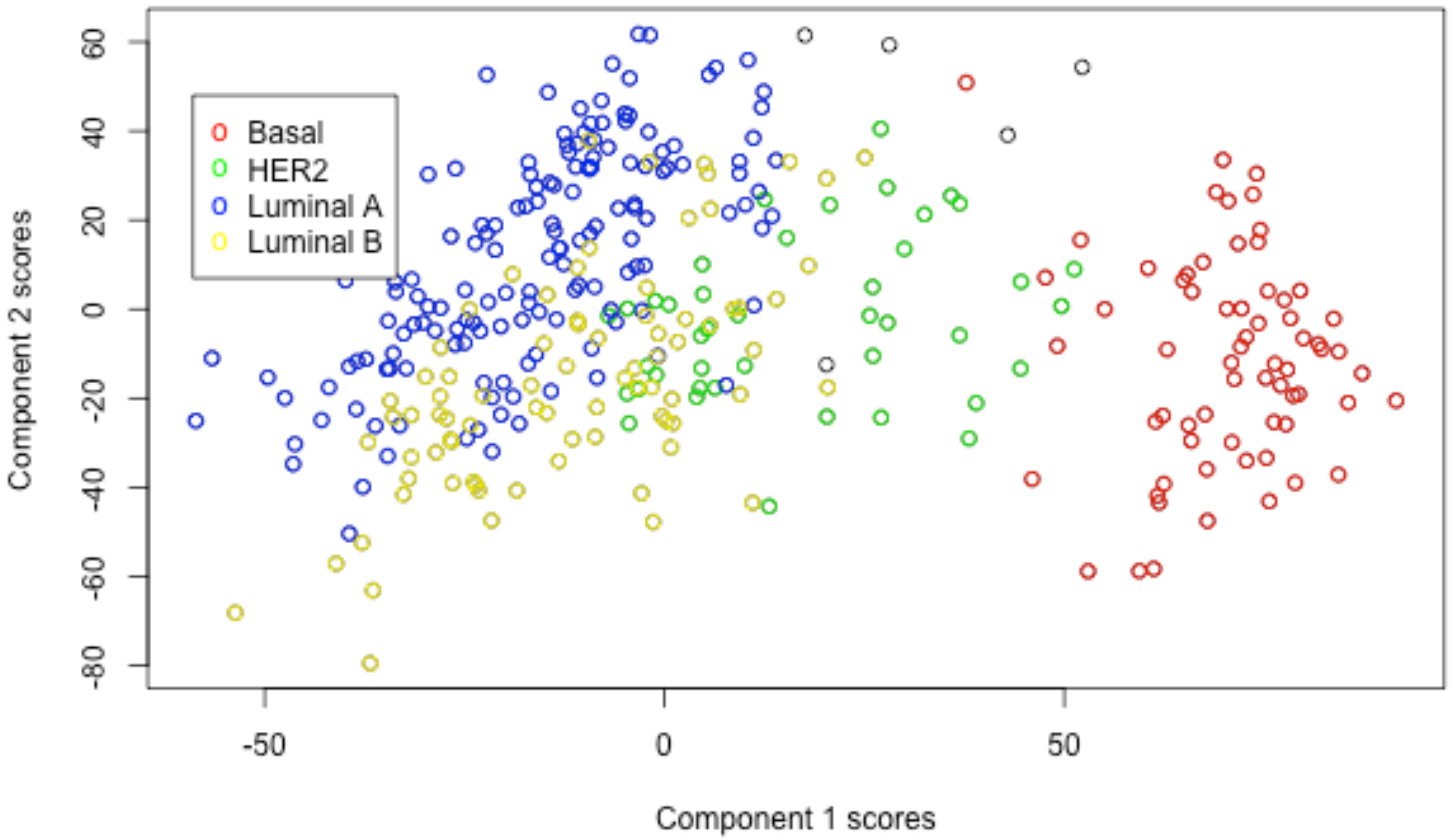
- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



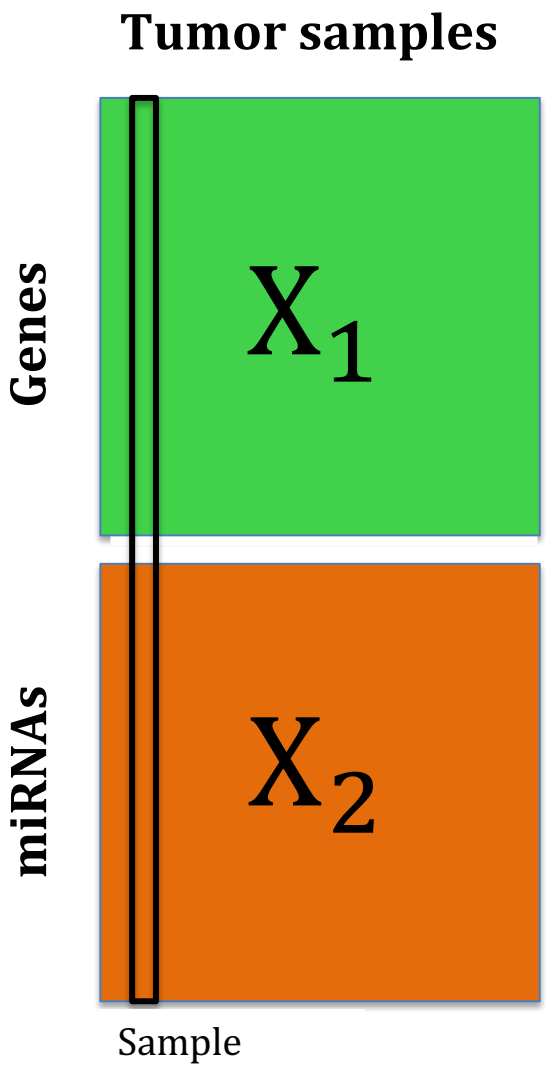
- Low rank factorization: $X \approx UV$, $U : m \times 3$, $V : 3 \times n$.

Matrix factorization

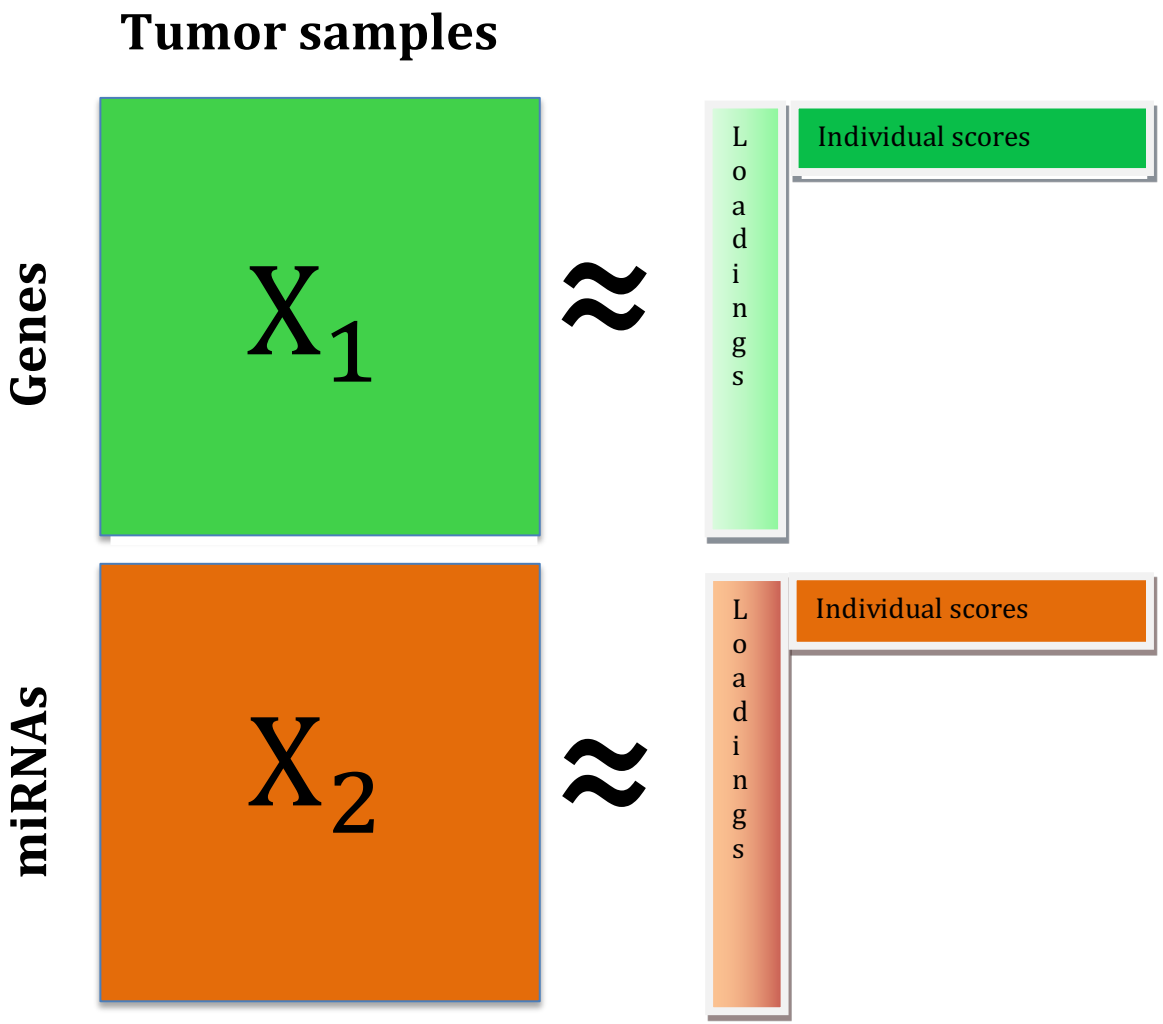
- First two principal component scores
  - Colored by breast tumor subtype



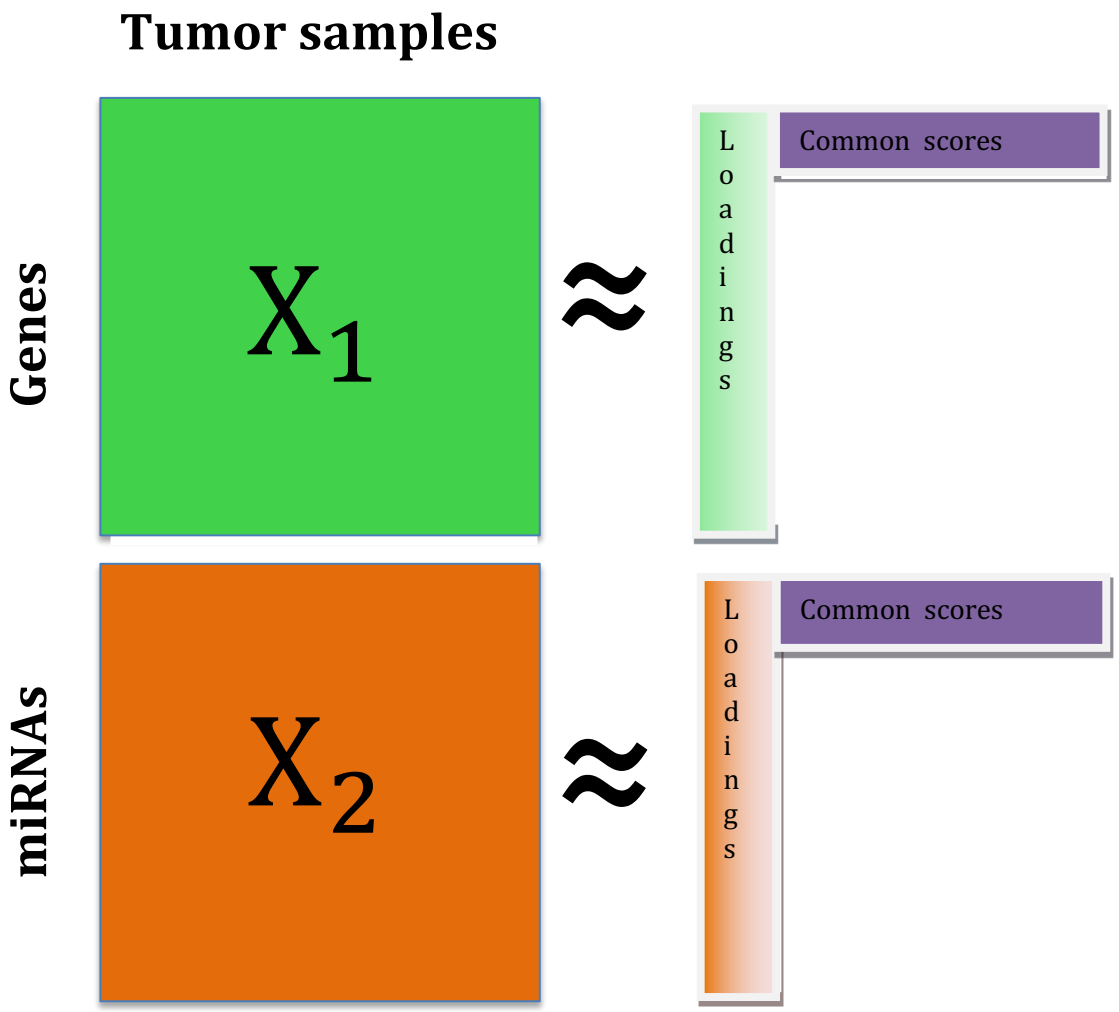
Vertically linked data



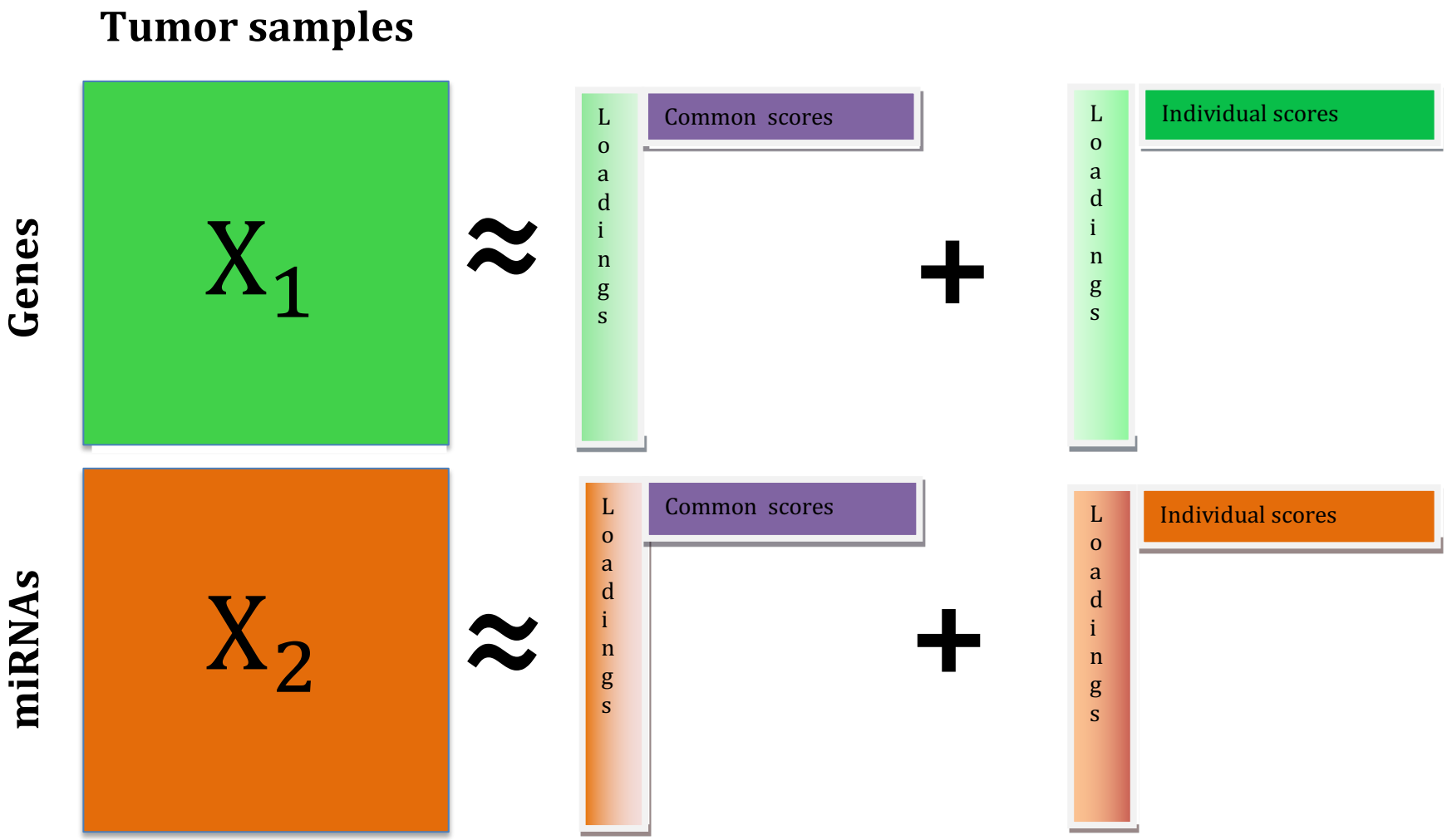
Vertically linked data: separate factorizations



Vertically linked data: joint factorization



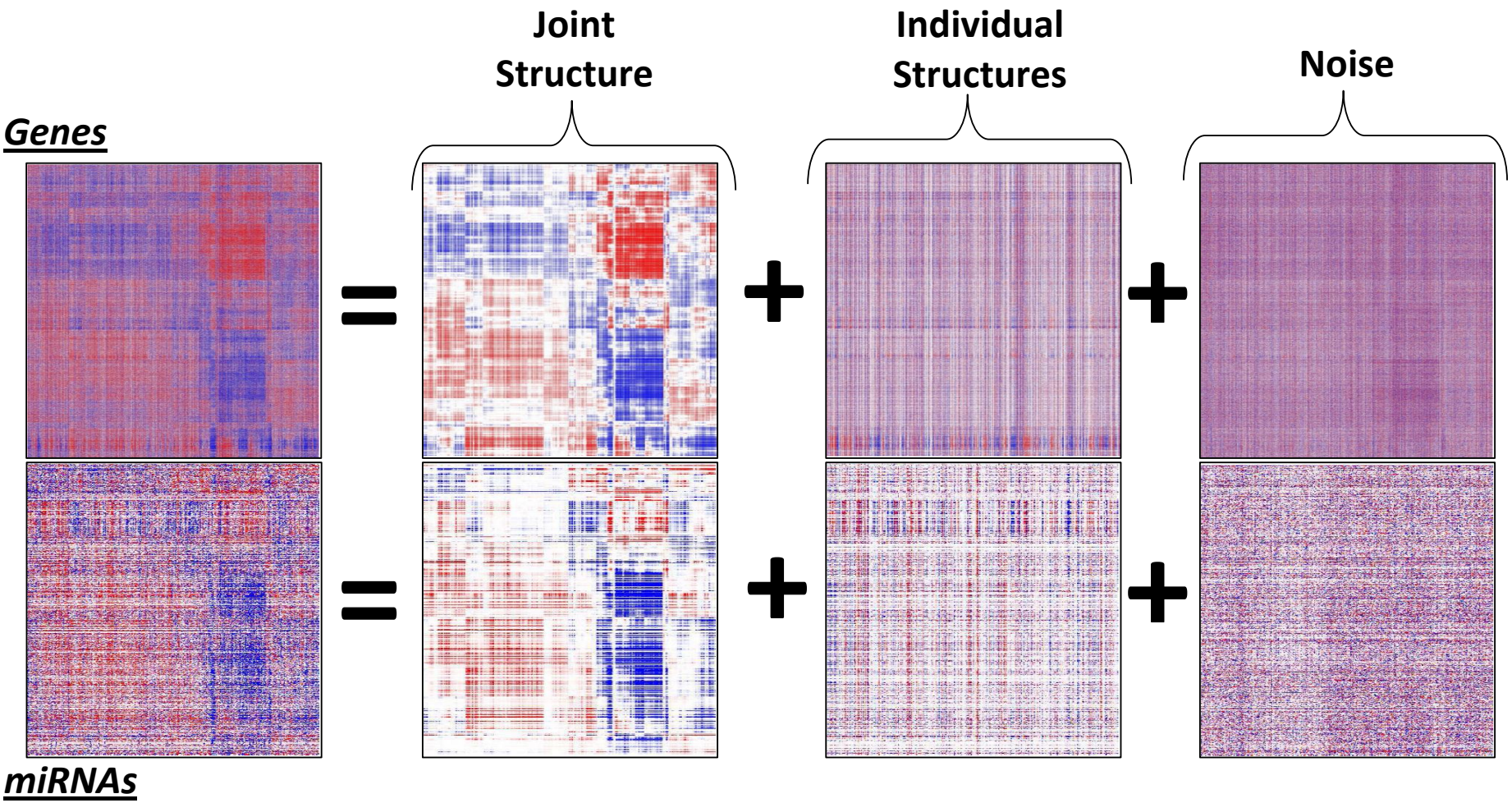
Vertically linked data: JIVE factorization



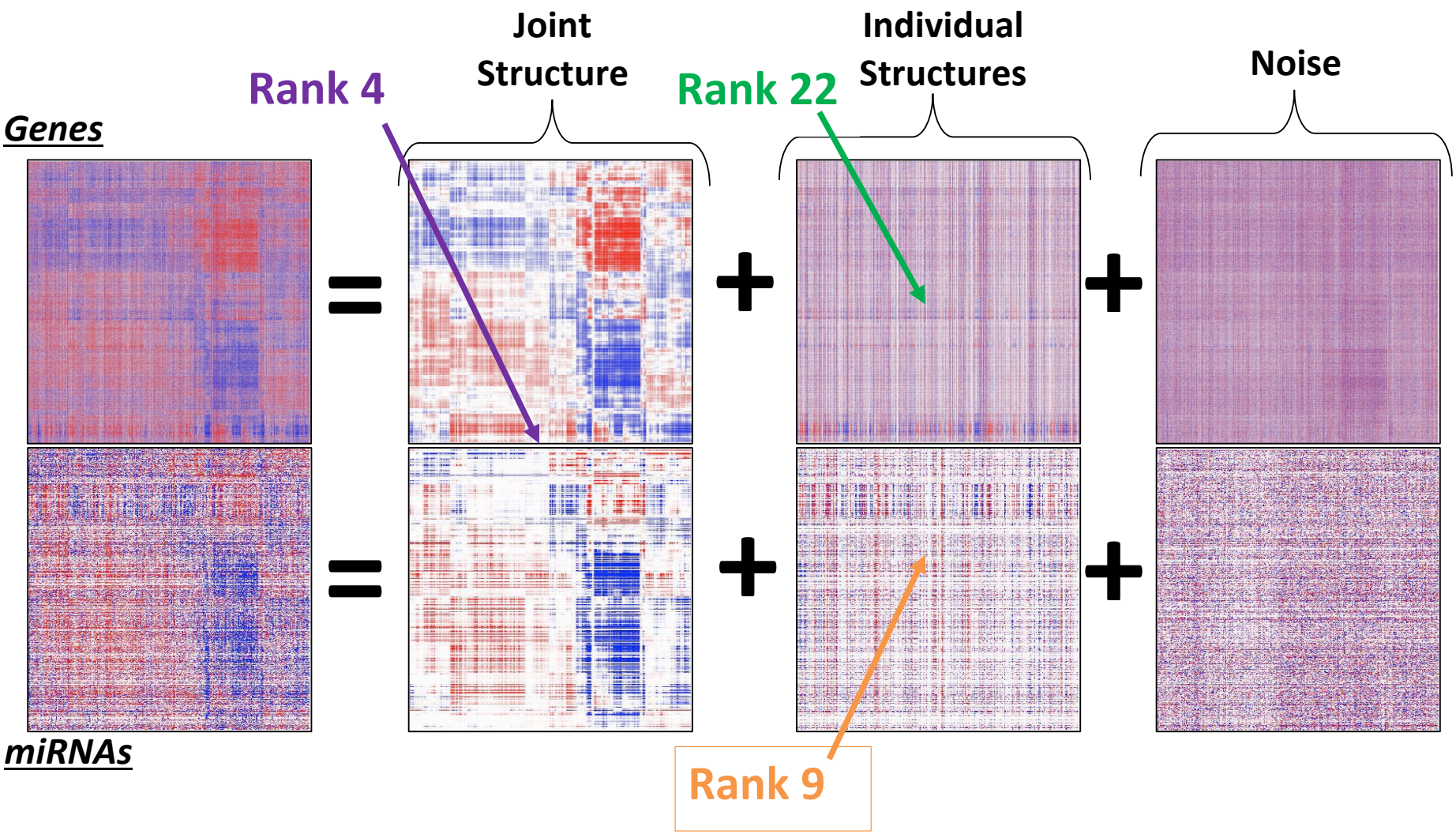
Joint + individual factorization methods

- 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!

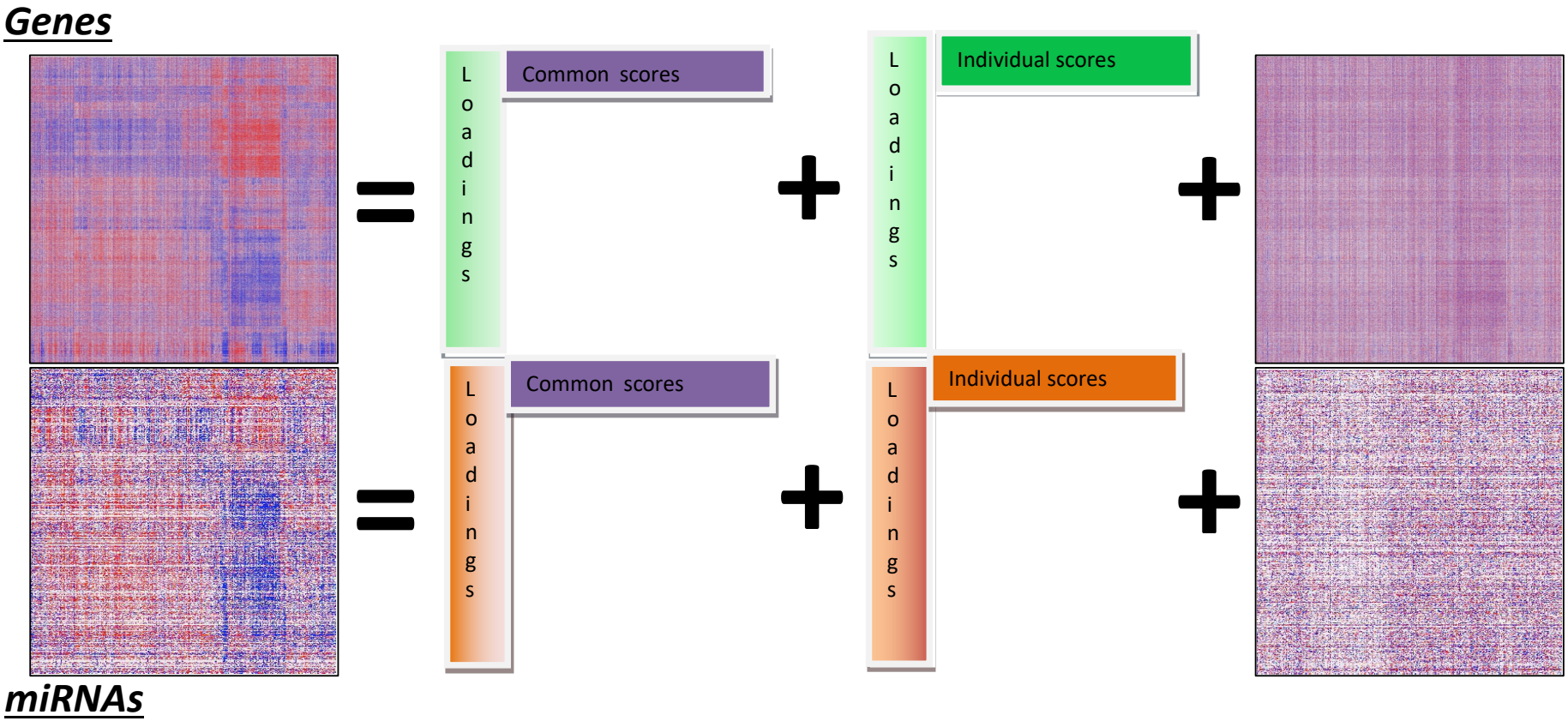
JIVE Estimates



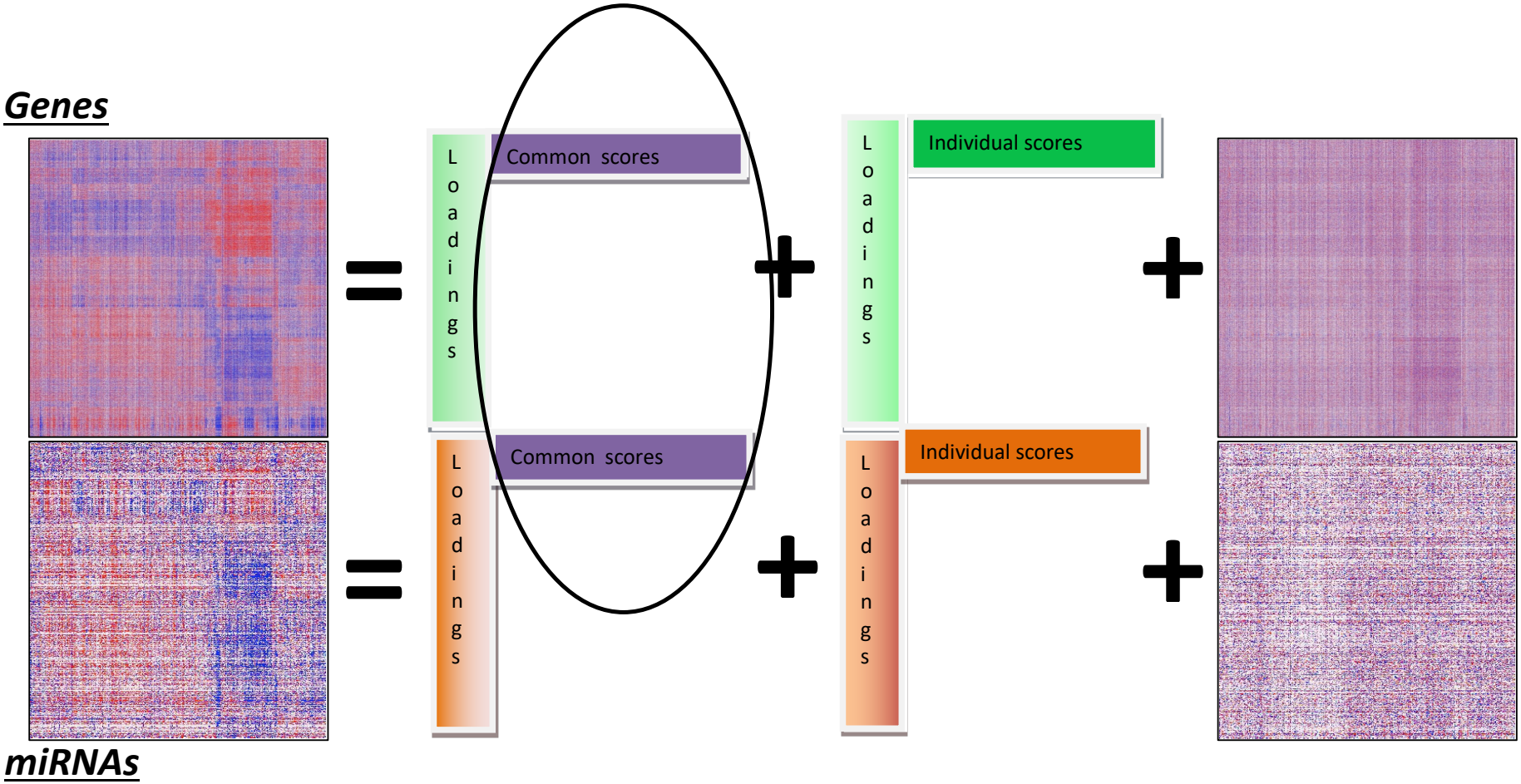
JIVE Estimates



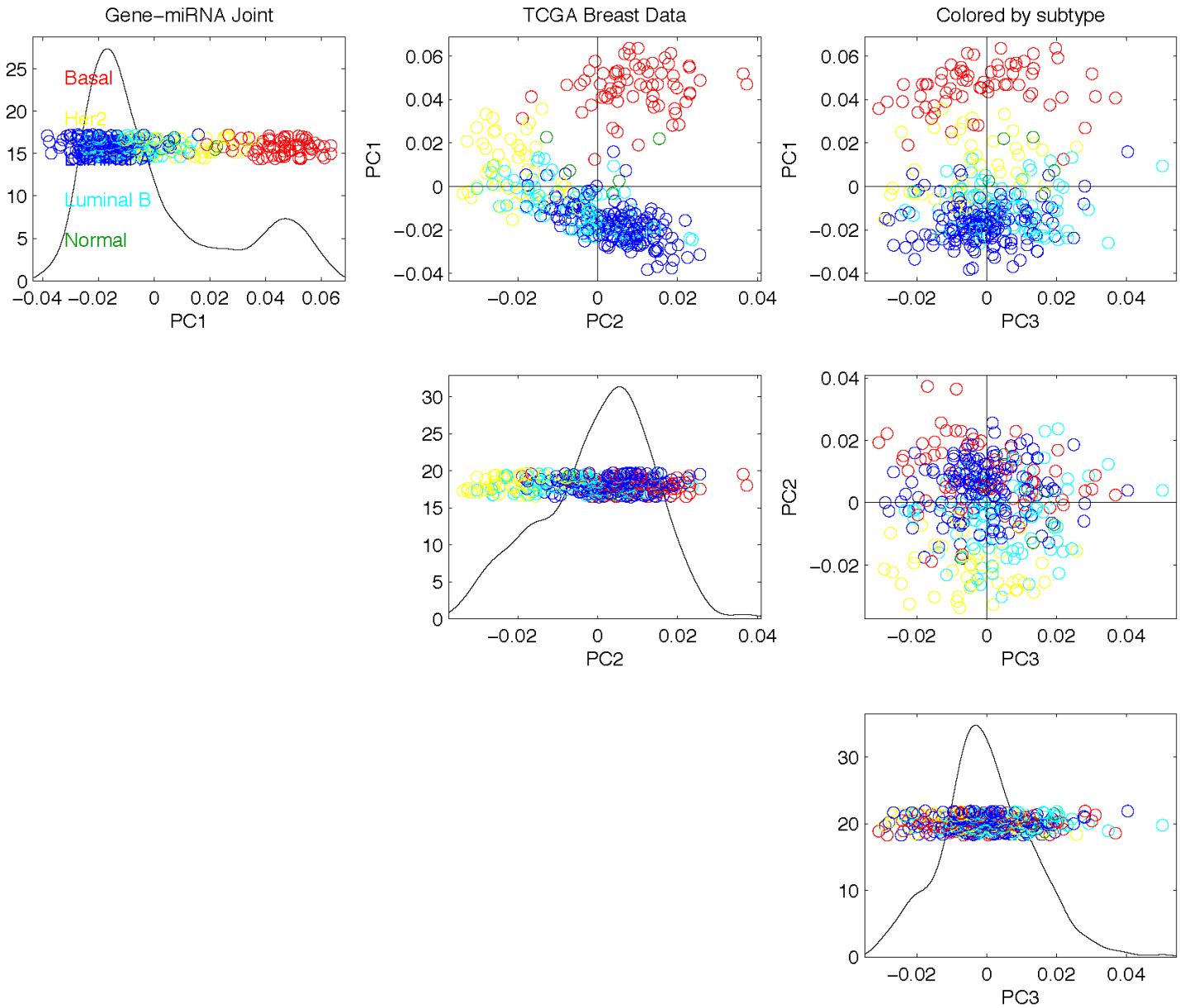
JIVE Estimates (factorized)



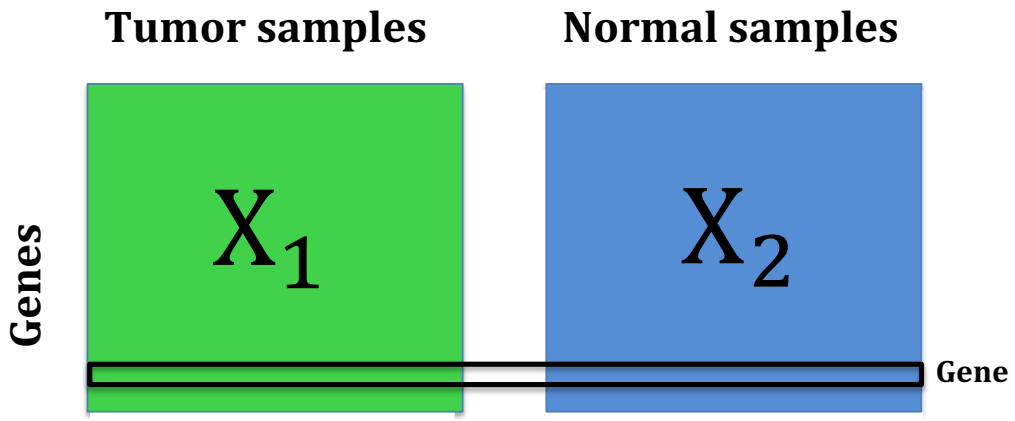
JIVE Estimates (factorized)



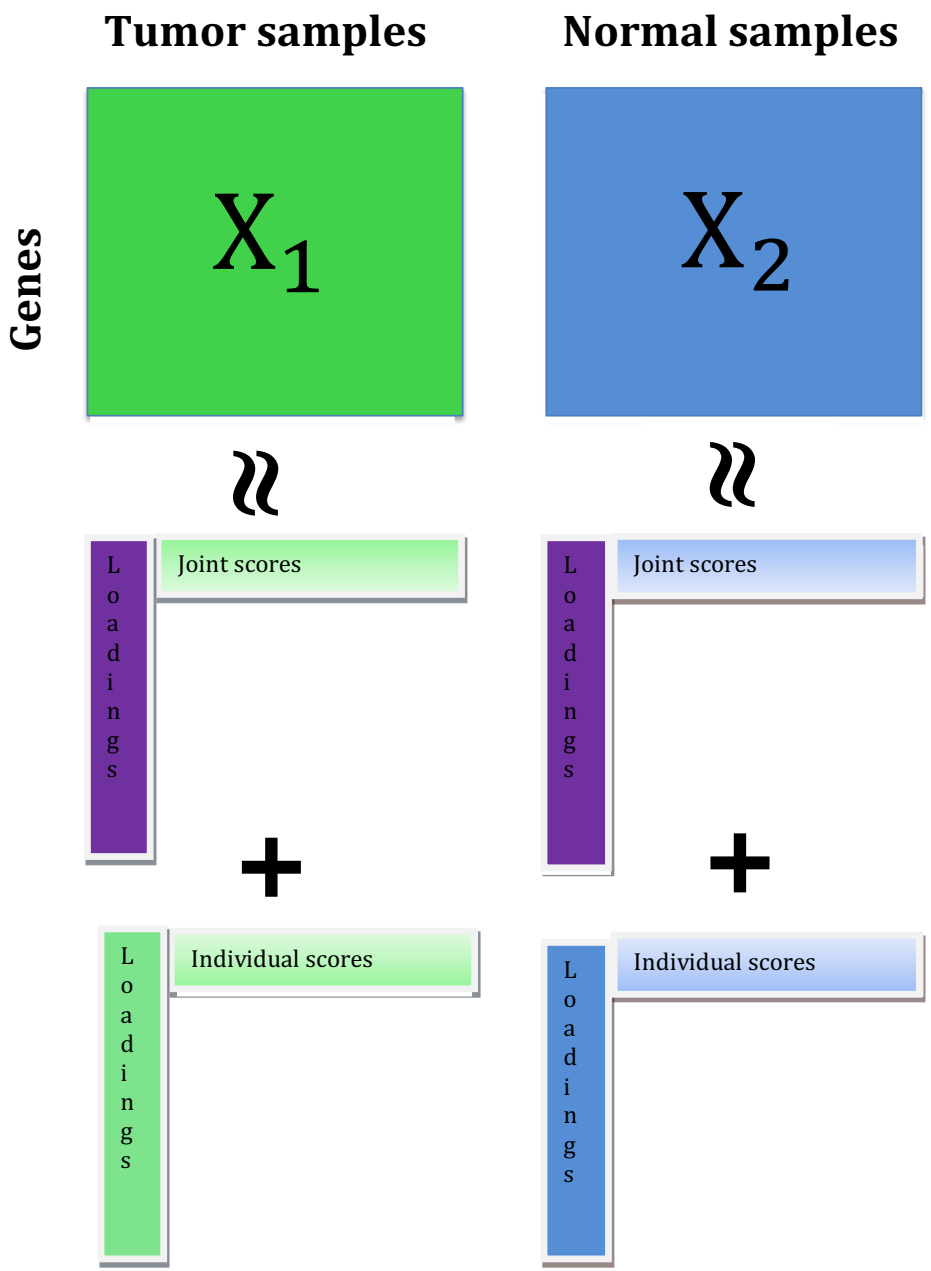
Joint PCs


Gene−miRNA Joint
Basal
Her2
Luminal A
Luminal B
Normal
PC1
TCGA Breast Data
PC1
PC2
Colored by subtype
PC1
PC3
PC2
PC2
PC3
PC3

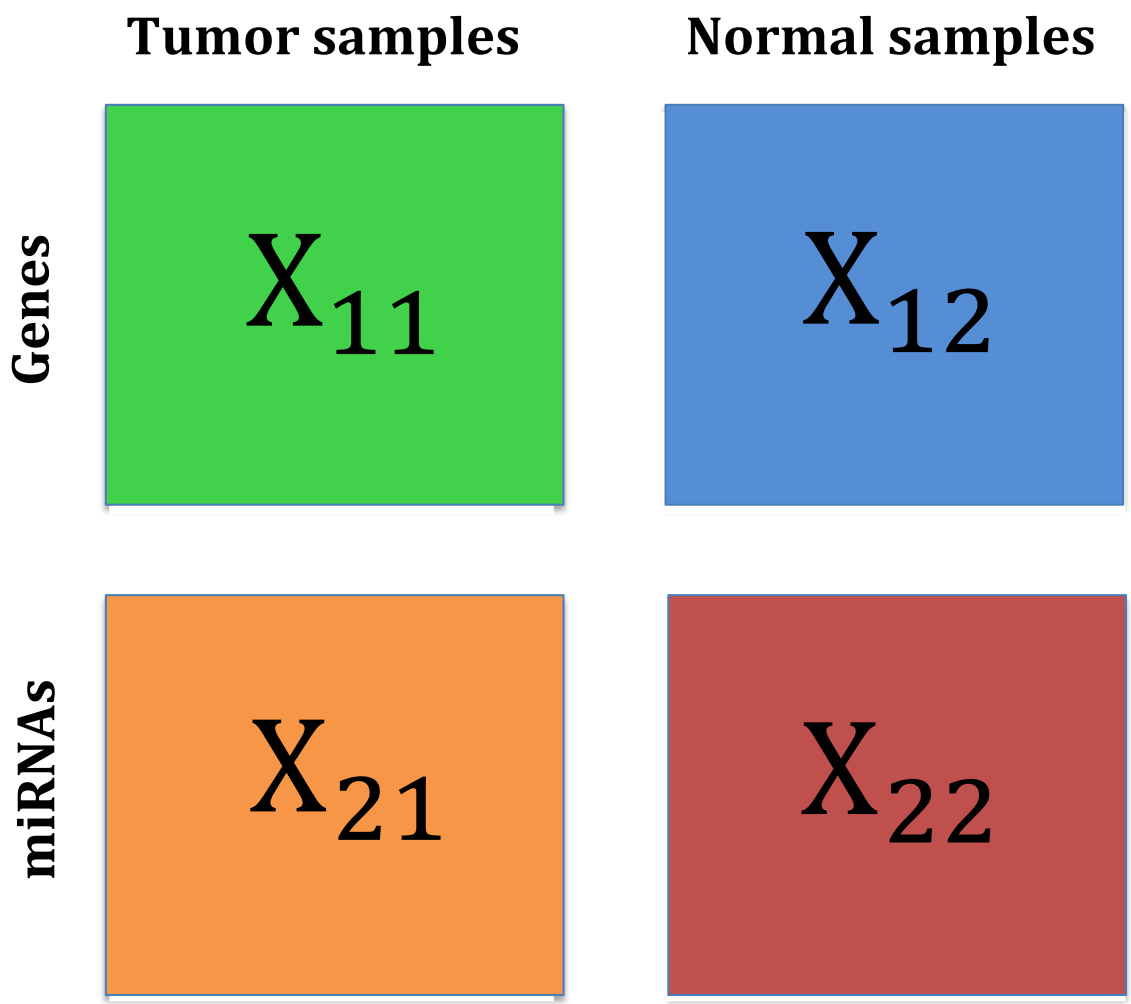
Horizontally linked data



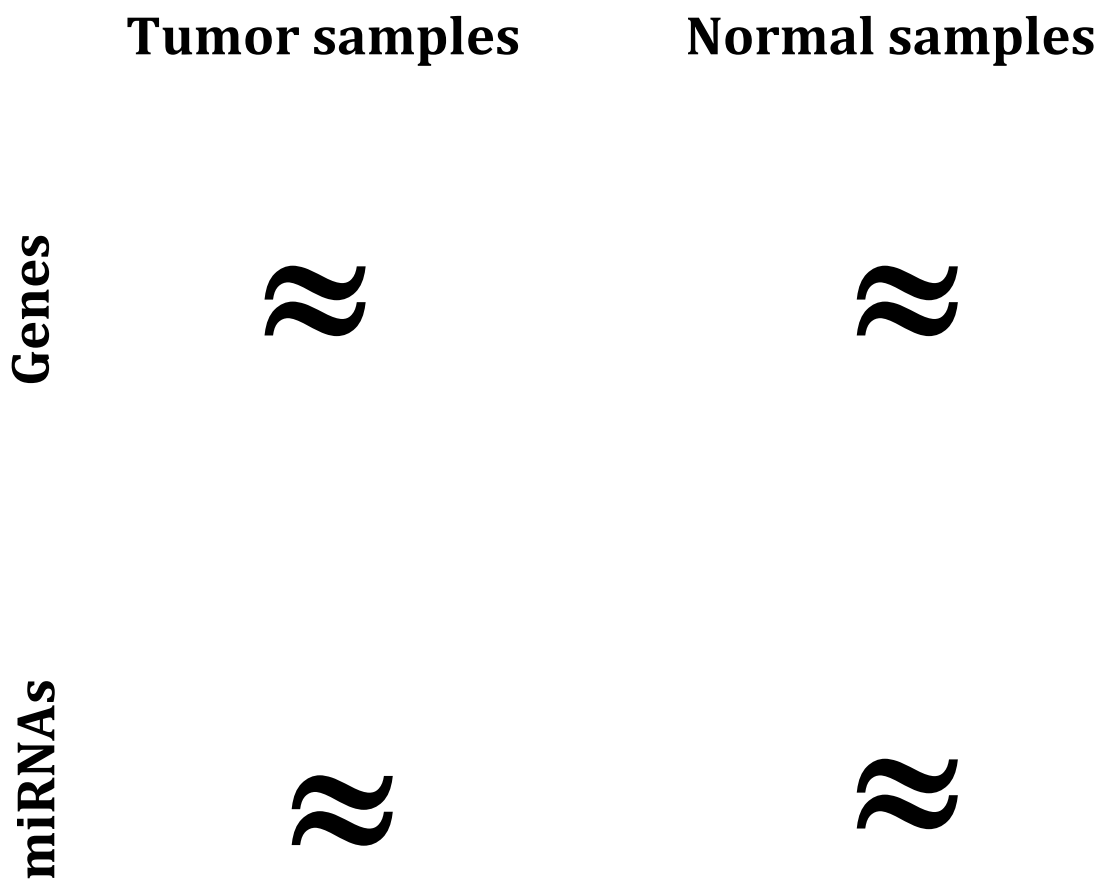
Horizontally linked data: JIVE factorization



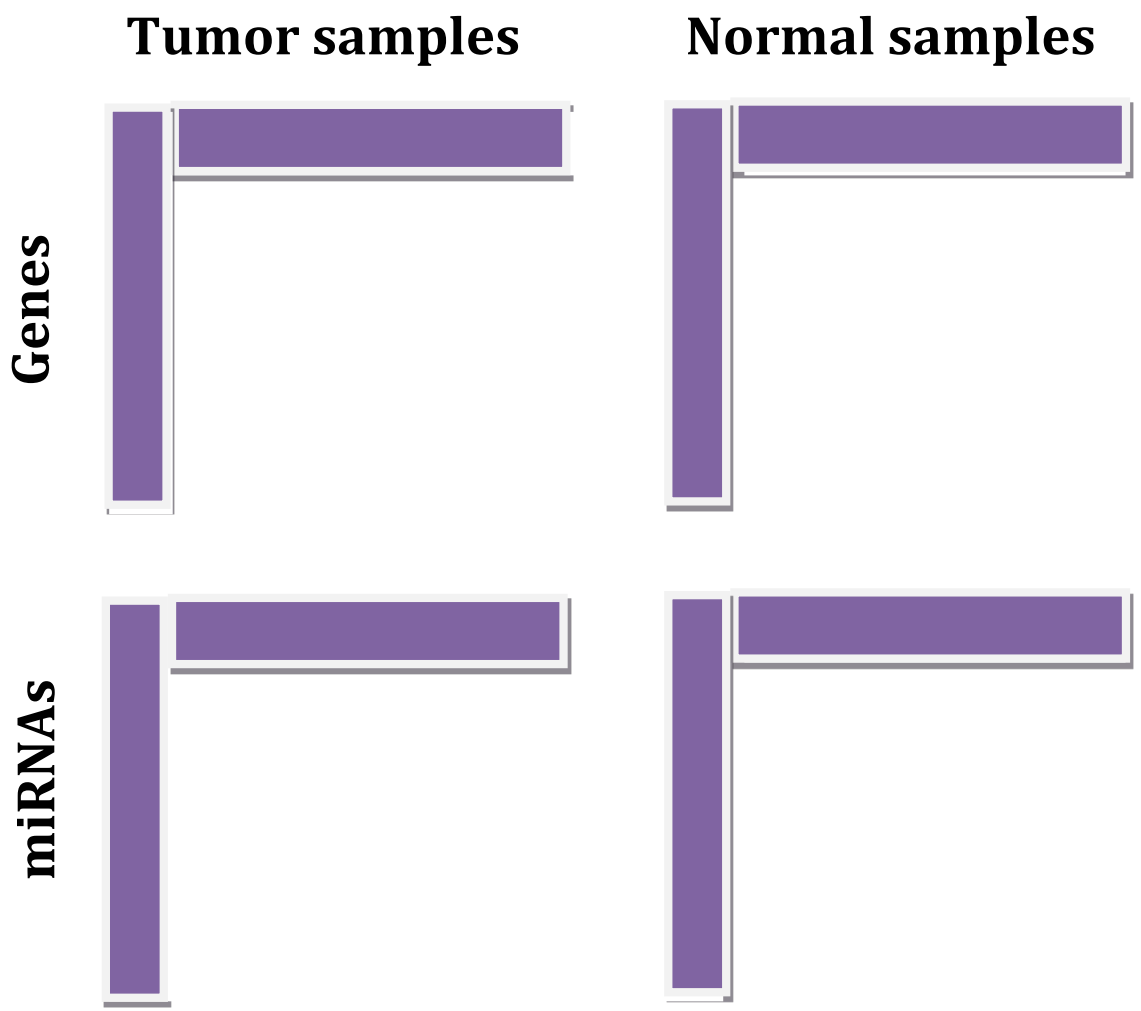
Bidimensionally linked data: BIDIFAC



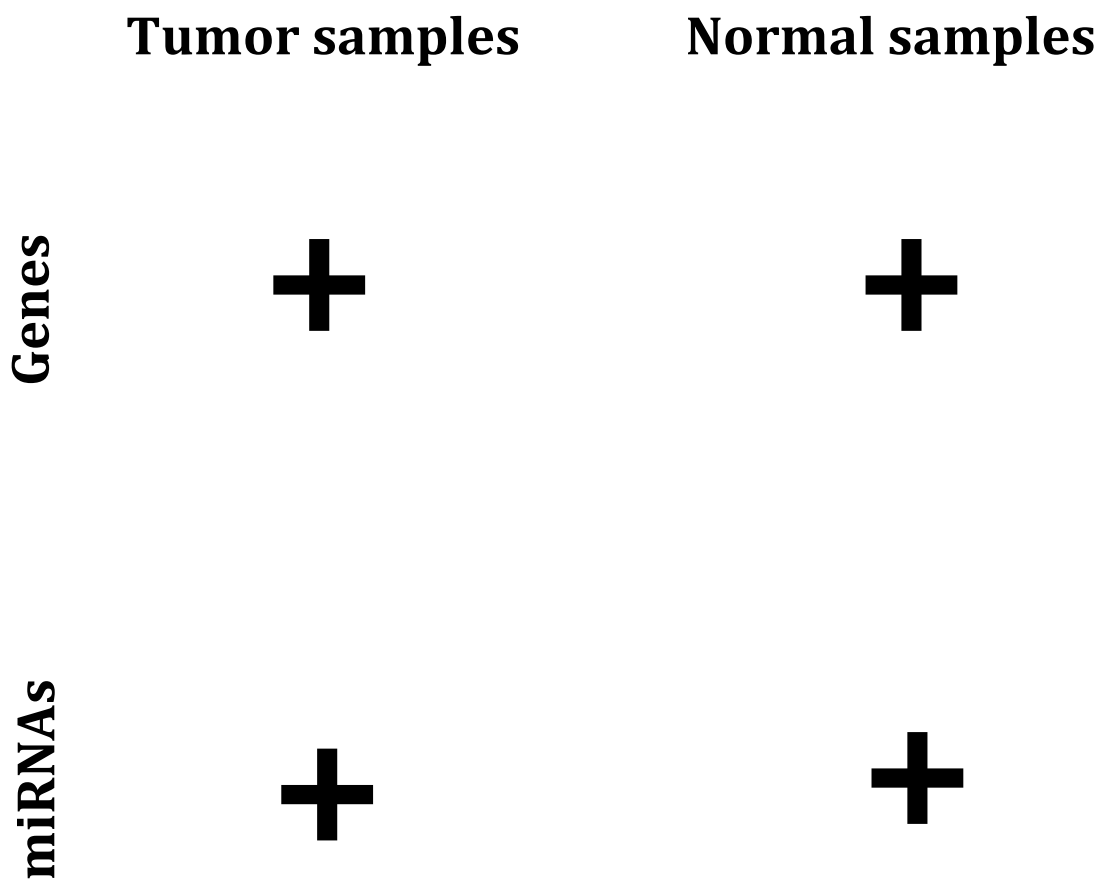
Bidimensionally linked data: BIDIFAC



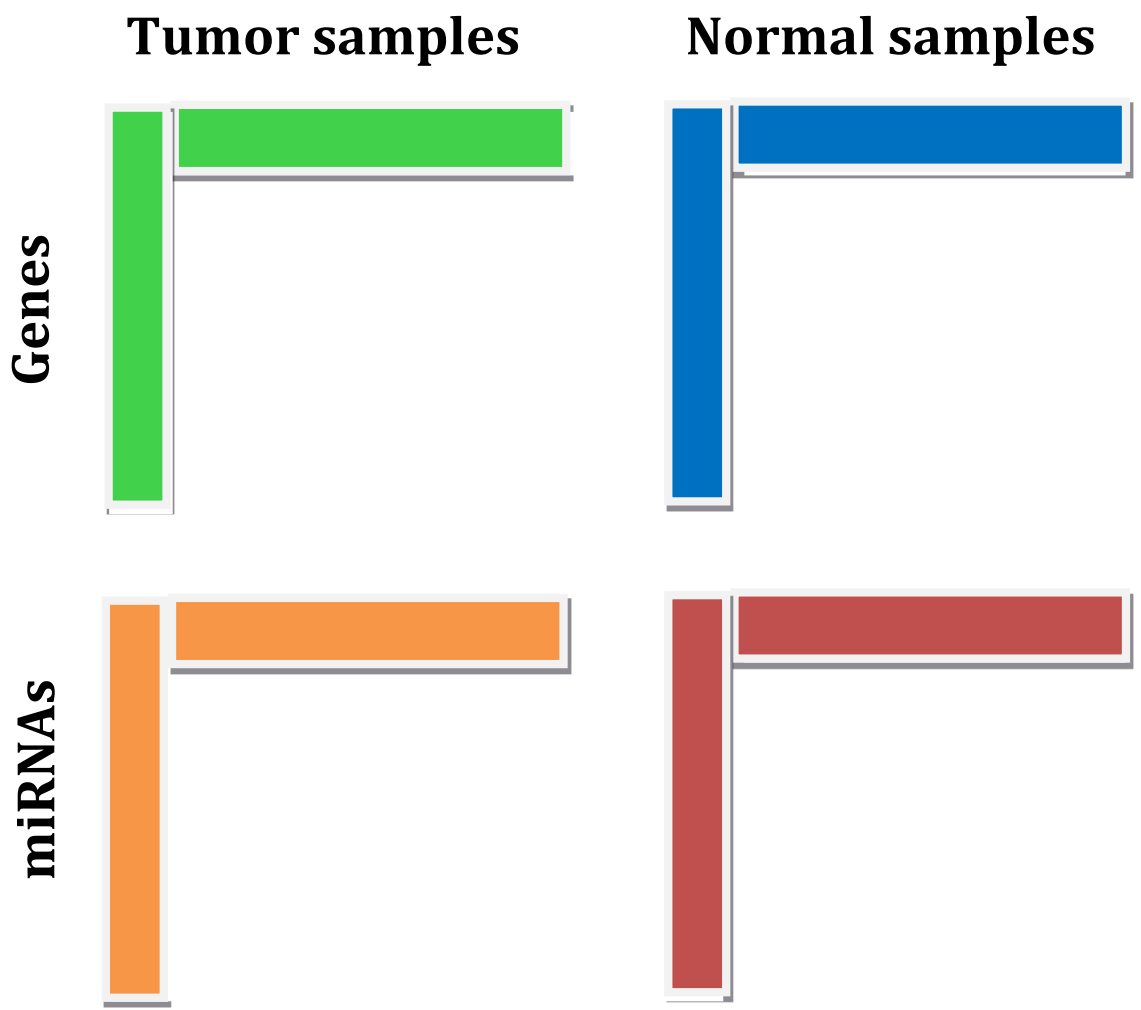
Bidimensionally linked data: BIDIFAC



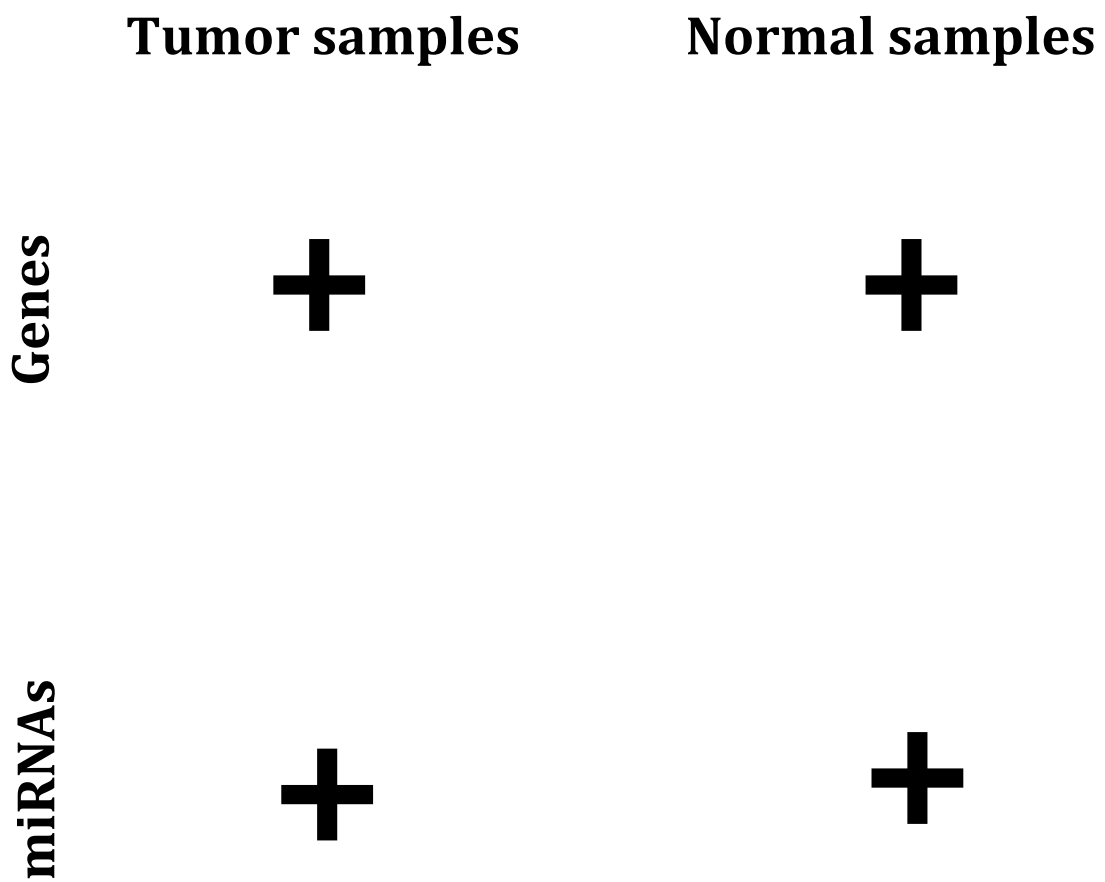
Bidimensionally linked data: BIDIFAC



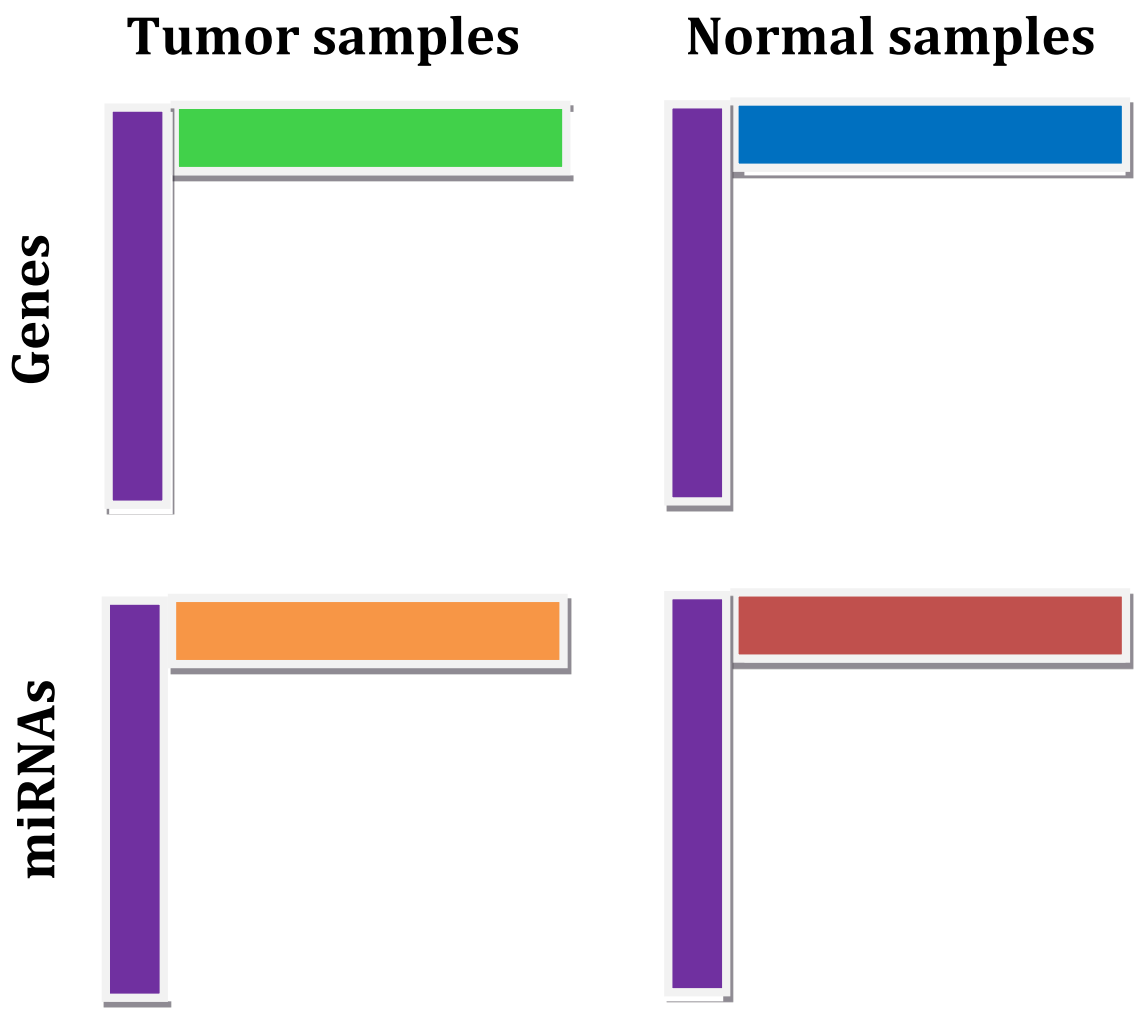
Bidimensionally linked data: BIDIFAC



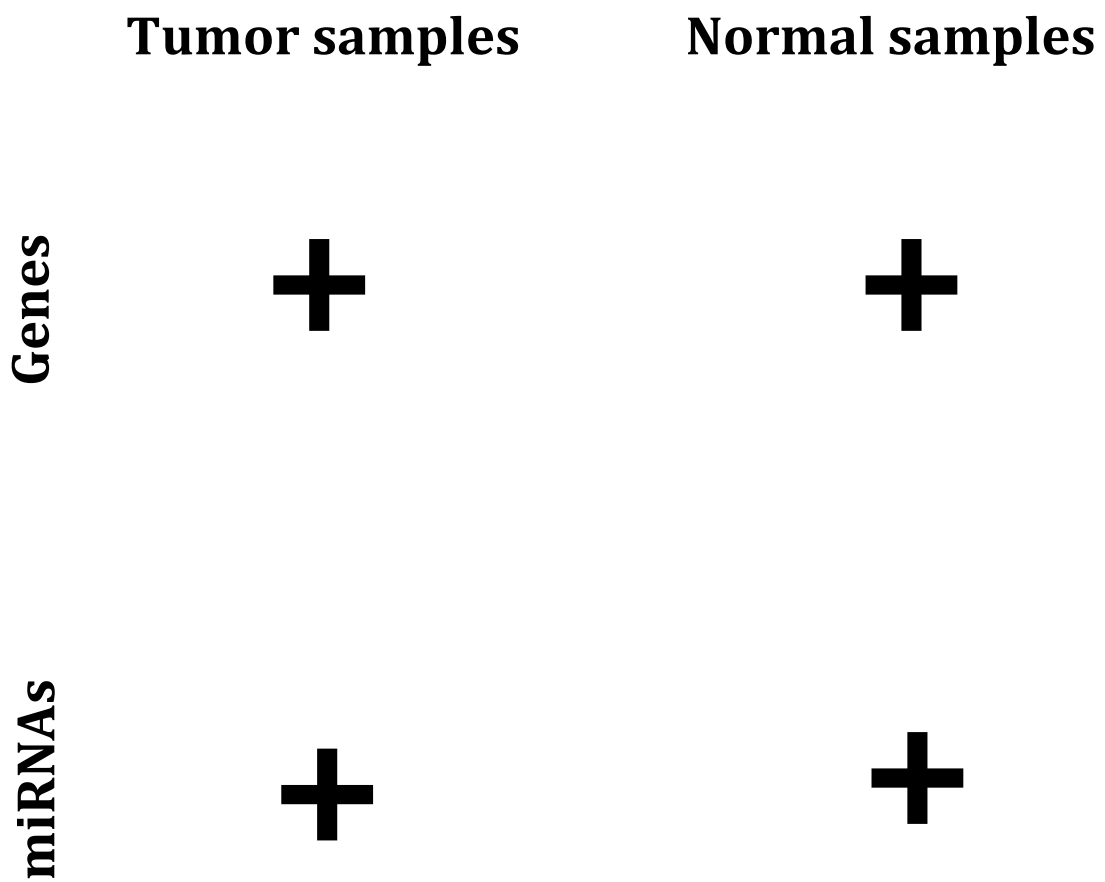
Bidimensionally linked data: BIDIFAC



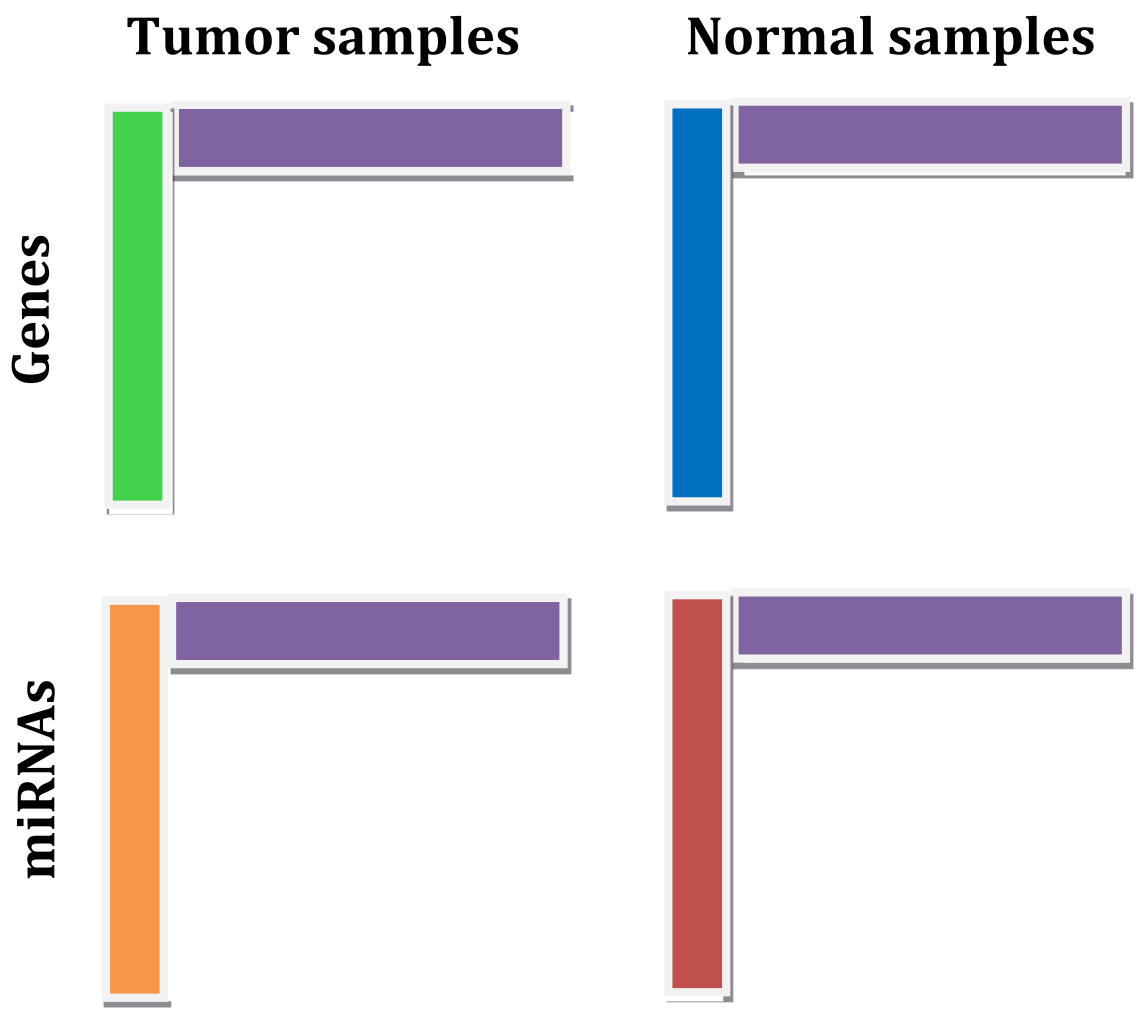
Bidimensionally linked data: BIDIFAC



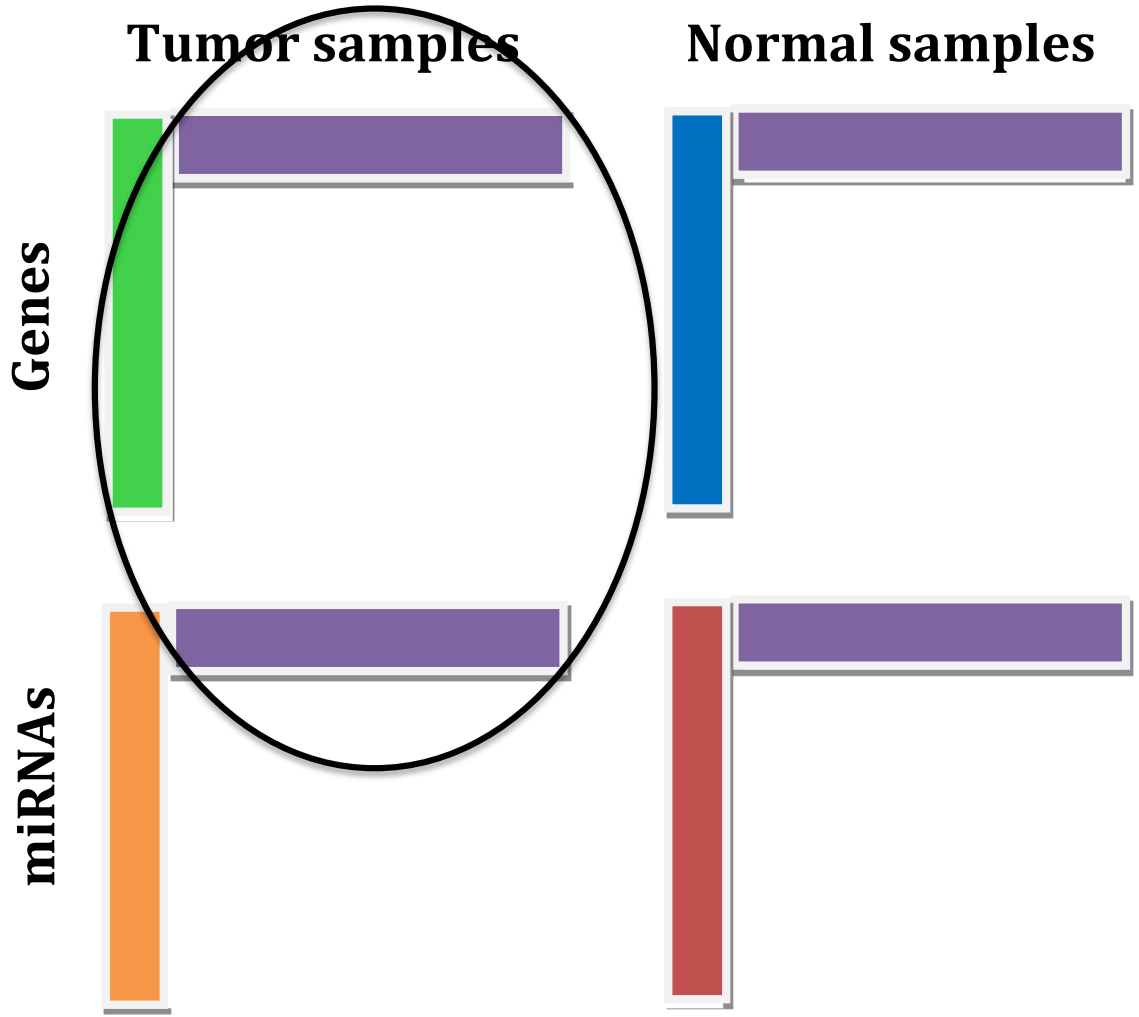
Bidimensionally linked data: BIDIFAC



Bidimensionally linked data: BIDIFAC



Bidimensionally linked data: BIDIFAC



Tumor-Specifc Columns Shared PCs

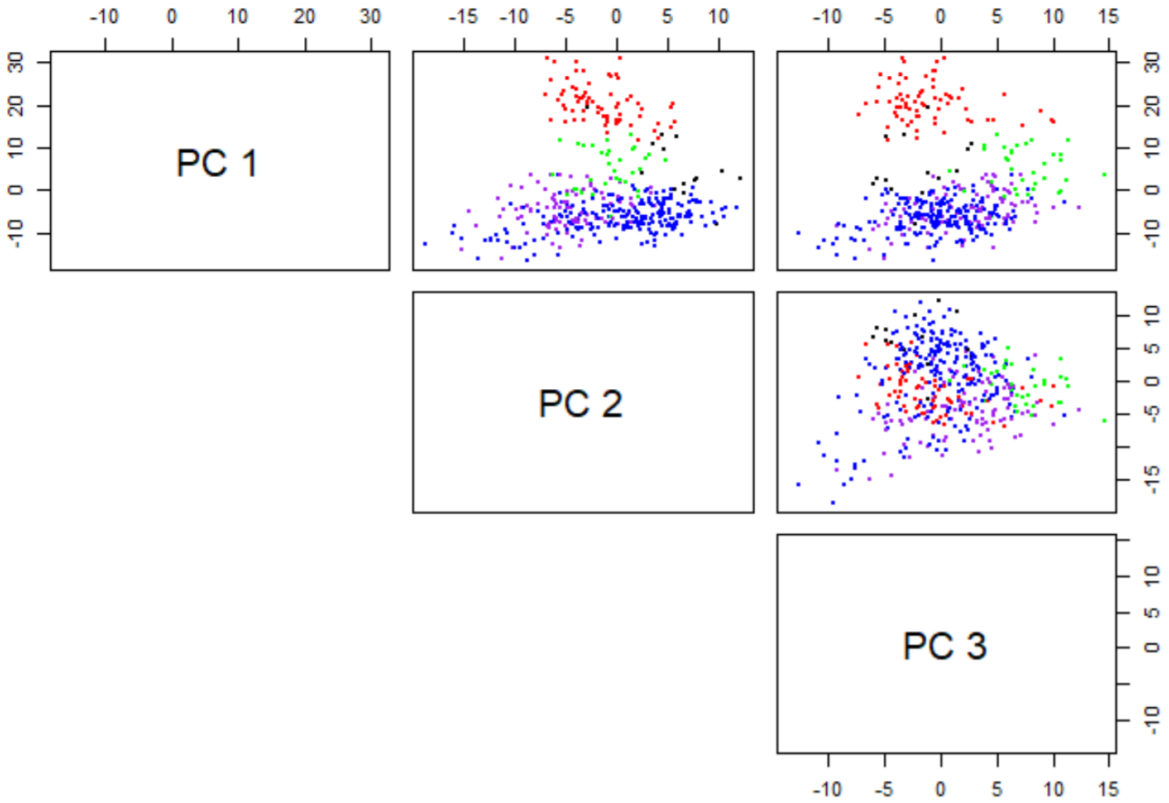
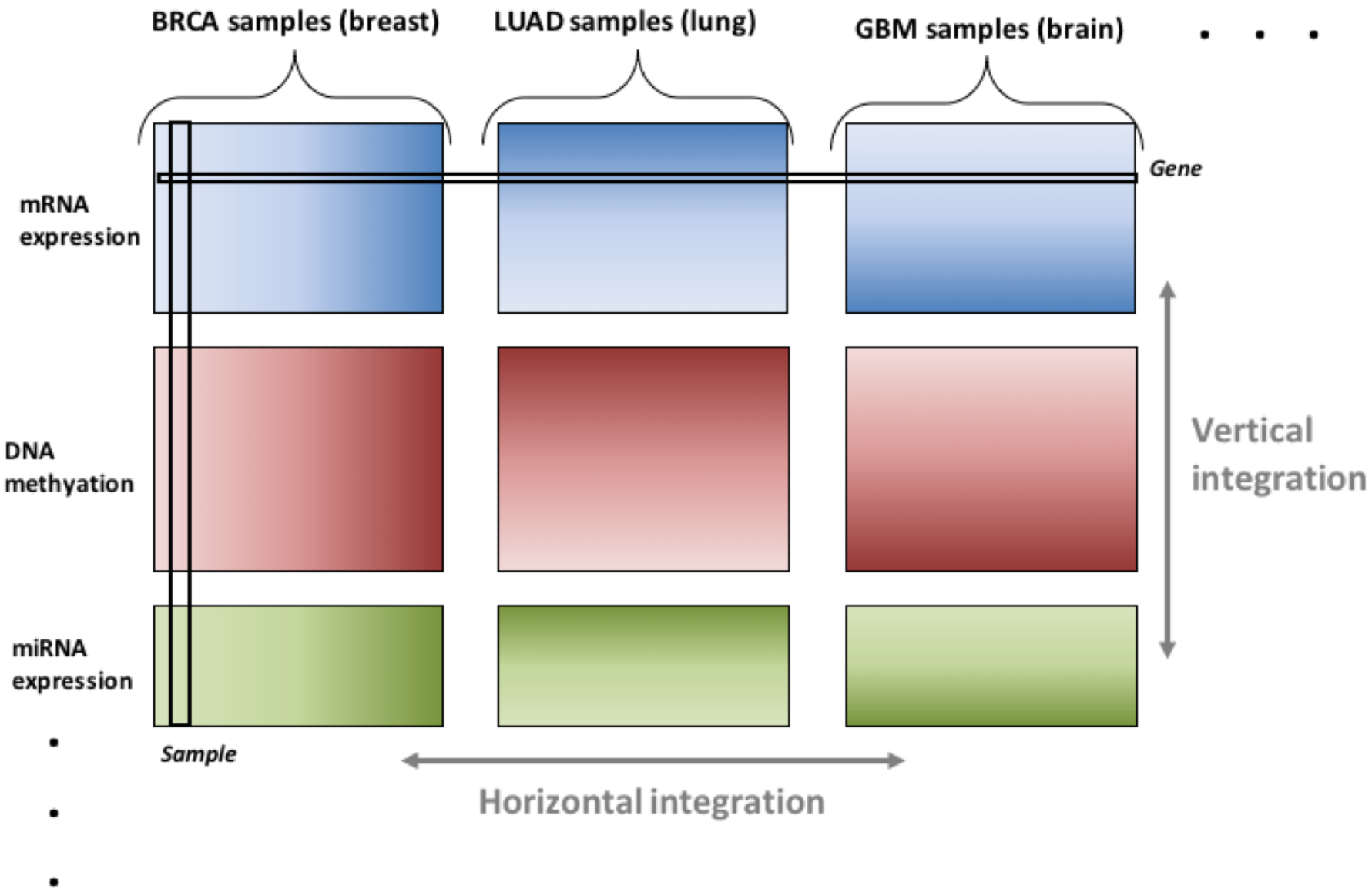


Figure: Principal components of the estimated column-shared structure, colored by subtype: Basal, HER2, Lum A, Lum B.

Pan-omics pan-cancer integration!



BIDIFAC: general framework

Consider matrices $\{\mathbf{X}_{ij} : M_i \times N_j \mid i = 1, \ldots, I,\ j = 1, \ldots, J\}$.

$$\mathbf{X}_{\cdot\cdot} = \begin{bmatrix} \mathbf{X}_{11} & \cdots & \mathbf{X}_{1I} \\ \vdots & \ddots & \vdots \\ \mathbf{X}_{J1} & \cdots & \mathbf{X}_{IJ} \end{bmatrix}$$

$$\mathbf{X}_{i\cdot} = [\mathbf{X}_{i1}, \cdots, \mathbf{X}_{iq}]$$

$$\mathbf{X}_{\cdot j} = \begin{bmatrix} \mathbf{X}_{1j} \\ \vdots \\ \mathbf{X}_{pj} \end{bmatrix}$$

BIDIFAC+: general framework

- Decompose $\mathbf{X}_{\cdot\cdot}$ into low-rank structural *modules*:

$$\mathbf{X}_{\cdot\cdot} = \sum_{k=1}^{\kappa} \mathbf{S}_{\cdot\cdot}^{(k)} + \mathbf{E}_{\cdot\cdot}, \tag{1}$$

where

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

presence of each $\mathbf{S}_{ij}^{(k)}$ is determined by $\mathbf{R} : I \times \kappa$ and $\mathbf{C} : J \times \kappa$:

$$\mathbf{S}_{ij}^{(k)} = \begin{cases} \mathbf{0}_{M_i \times N_j} & \text{if } \mathbf{R}[i,k] = 0 \text{ or } \mathbf{C}[j,k] = 0 \\ \mathbf{U}_i^{(k)} \mathbf{V}_j^{(k)} & \text{if } \mathbf{R}[i,k] = 1 \text{ and } \mathbf{C}[j,k] = 1 \end{cases}.$$

BIDIFAC+: objective

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

$$||\mathbf{X}_{\bullet\bullet} - \sum_{k=1}^{\kappa} \mathbf{S}_{\bullet\bullet}^{(k)}||_F^2 + \sum_{k=1}^{\kappa} \lambda_k ||\mathbf{S}_{\bullet\bullet}^{(k)}||_*$$

where $||\cdot||_*$ gives the nuclear norm

$$\text{SVD}(\mathbf{A}) = \mathbf{UDV}^T \rightarrow ||\mathbf{A}||_* = \sum \mathbf{D}[i,i].$$

- Choice of λ_k's
  - Determined by random matrix theory for singular values $\mathbf{D}$
  - Gaurantees each module $\mathbf{S}_{\bullet\bullet}^{(k)}$ is low-rank.
  - Any submatrix $(\mathbf{R}[,k], \mathbf{C}[,k])$ can have a non-zero module

BIDIFAC+: Identifiability

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

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

Theorem

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

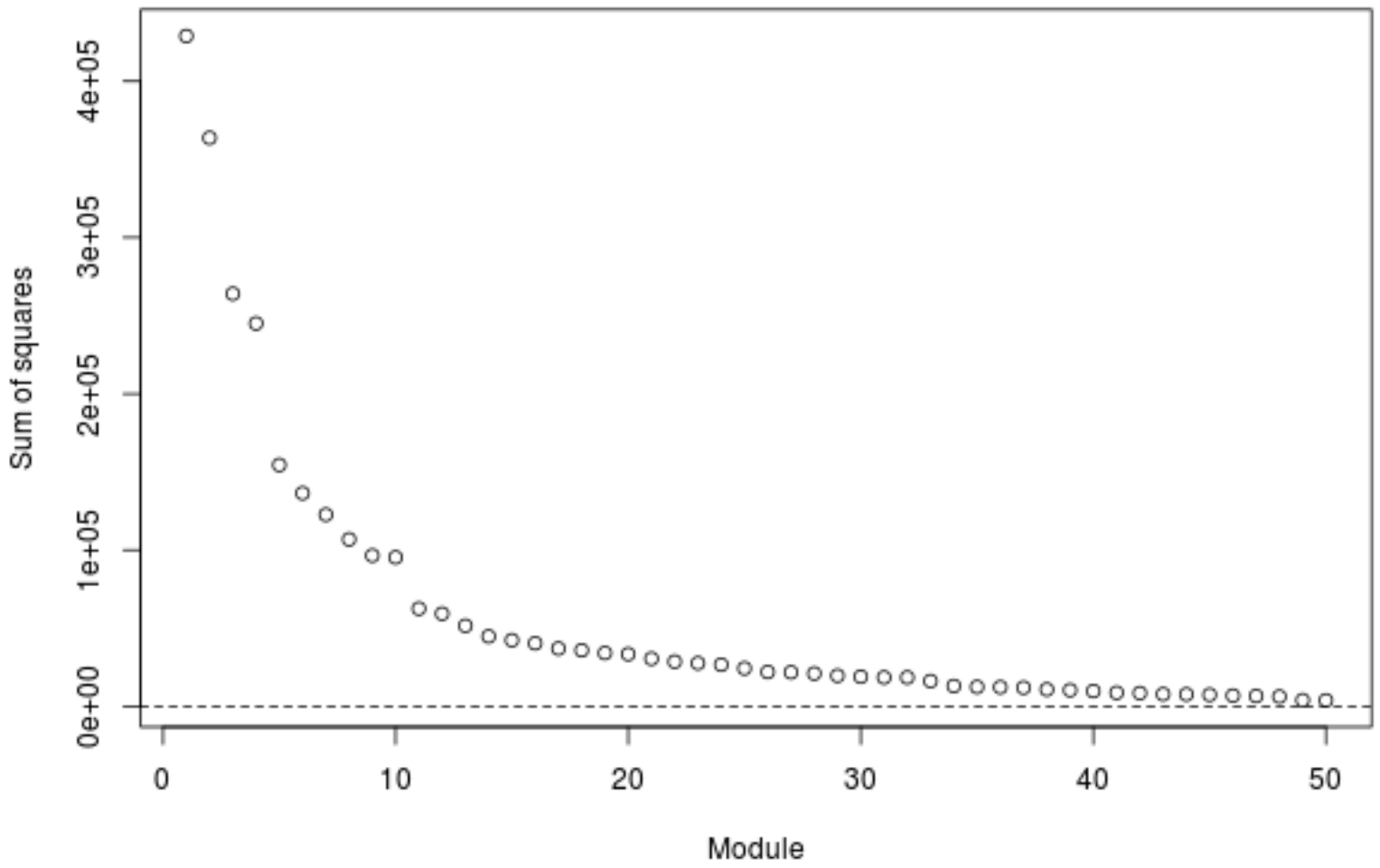
- $\{\hat{\mathbf{S}}_{\bullet\bullet}^{(k)}\}_{k=1}^{K}$ *minimizes* $\sum_{k=1}^{\kappa} \lambda_k ||\mathbf{S}_{\bullet\bullet}^{(k)}||_*$ *over* $\mathbb{S}_{\hat{X}}$,
- $\{\hat{\mathbf{U}}_i^{(k)}[\bullet, r] : \mathbf{R}[i,k] = 1$ *and* $\hat{\mathbf{D}}^{(k)}[r,r] > 0\}$ *are linearly independent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

Pan-omics pan-cancer data

- TCGA data for 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
- $(2^4 - 1) \cdot (2^{29} - 1) = 8053063665$ possible modules!

BIDIFAC+: results

- Variance explained for each module $\mathbf{S}_{\cdot\cdot}^{(k)}$ ($K = 50$):



BIDIFAC+: results

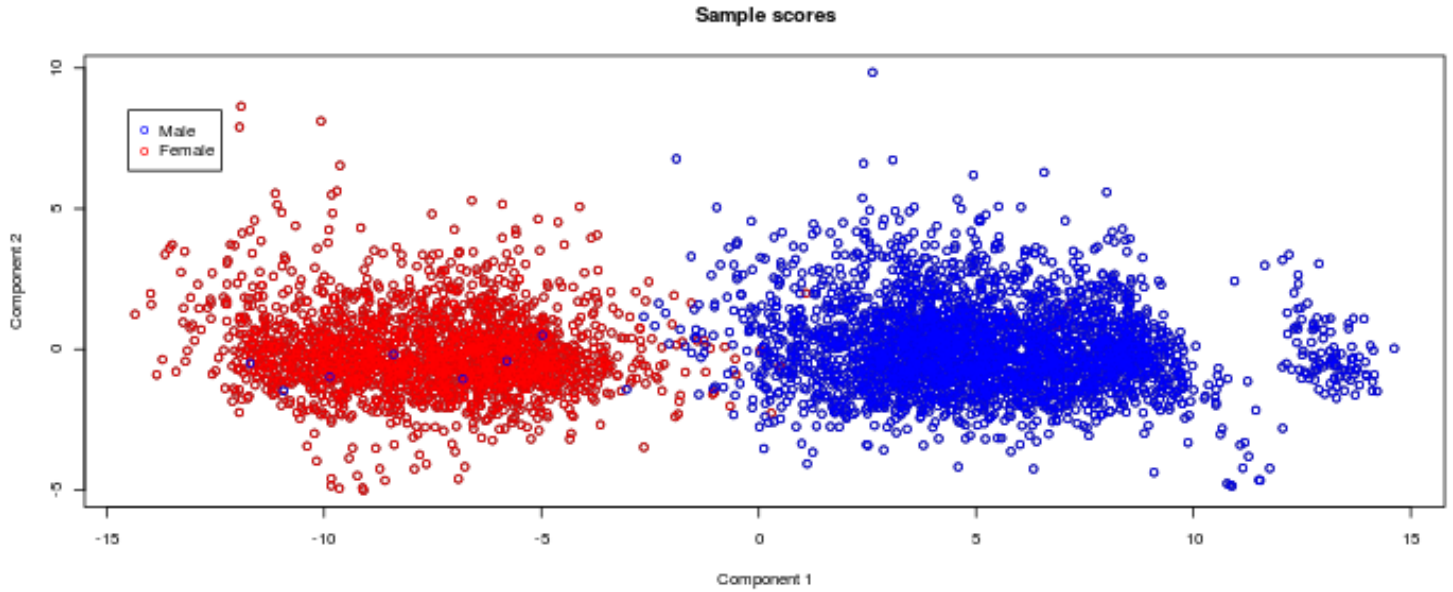
- Top structural modules, ranked by variance explained:

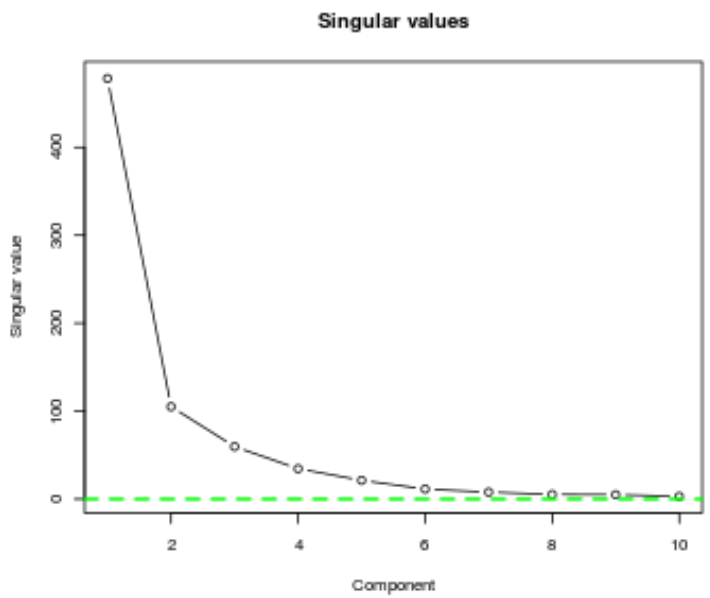
Module	Cancer types	Omics sources
1	All cancers	mRNA miRNA Meth Protein
2	All cancers	miRNA
3	BLCA BRCA CESC CHOL CORE DLBC ESCA HNSC LIHC LUAD LUSC OV PAAD PRAD SKCM STAD TGCT UCEC UCS	Meth
4	ACC BLCA CHOL CORE DLBC ESCA HNSC KICH KIRC KIRP LGG LIHC LUAD LUSC MESO PAAD PCPG SARC SKCM STAD THCA THYM	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

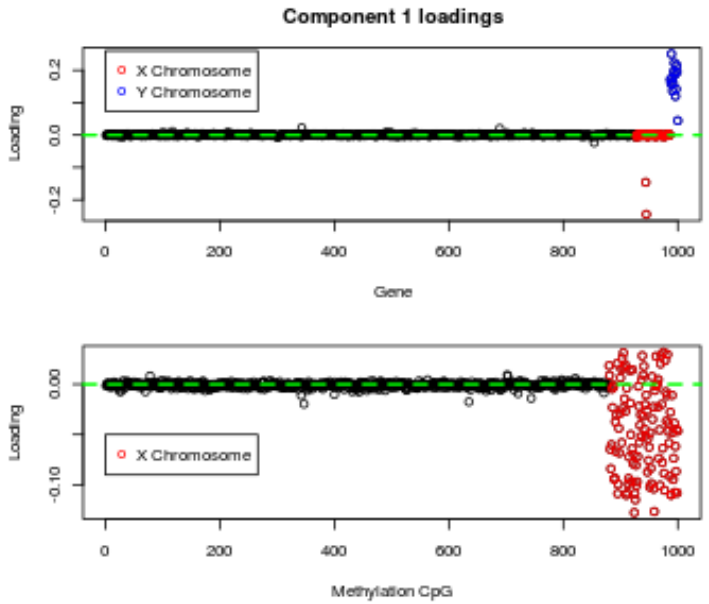
BIDIFAC+: results

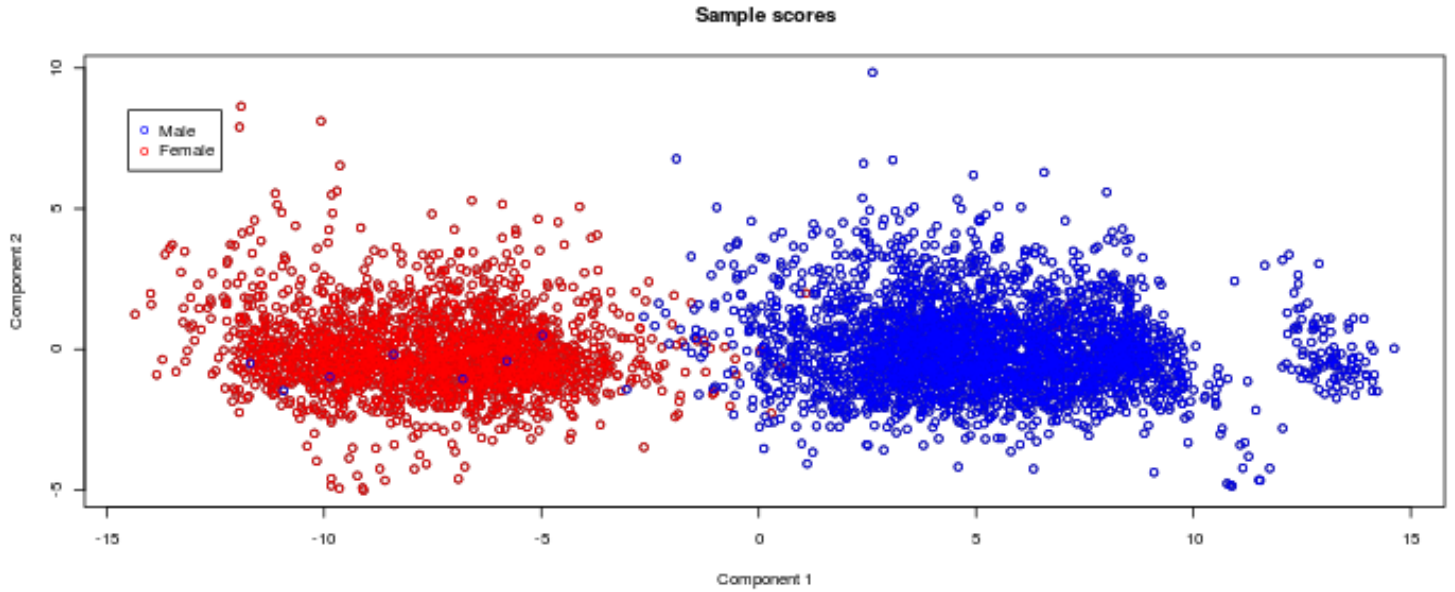
- Top structural modules, ranked by variance explained:

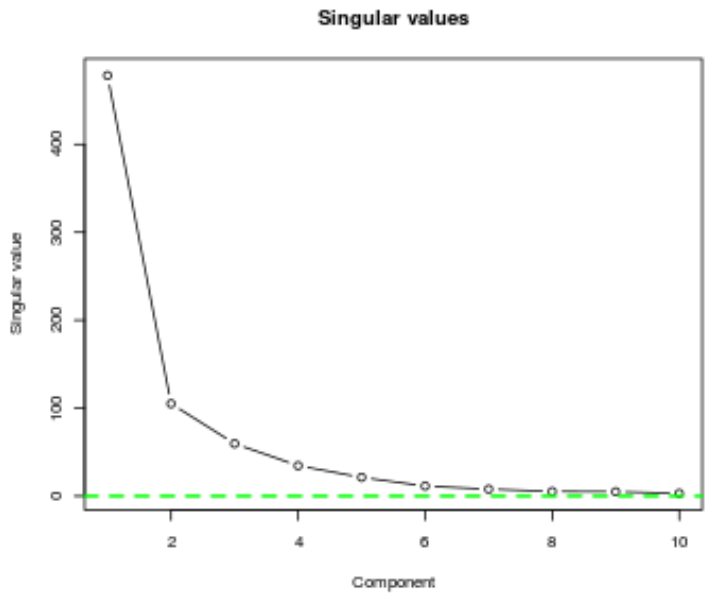
Module	Cancer types	Omics sources
1	All cancers	mRNA miRNA Meth Protein
2	All cancers	miRNA
3	BLCA BRCA CESC CHOL CORE DLBC ESCA HNSC LIHC LUAD LUSC OV PAAD PRAD SKCM STAD TGCT UCEC UCS	Meth
4	ACC BLCA CHOL CORE DLBC ESCA HNSC KICH KIRC KIRP LGG LIHC LUAD LUSC MESO PAAD PCPG SARC SKCM STAD THCA THYM	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

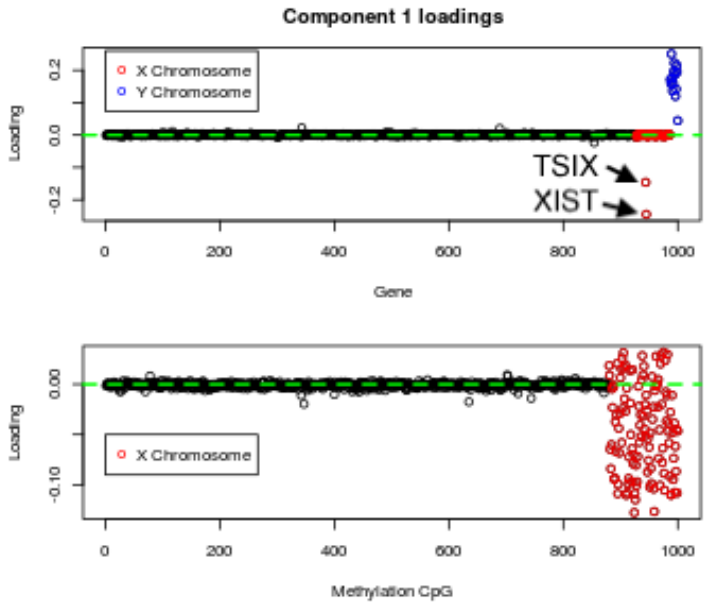

Sample scores
Male
Female
Component 2
Component 1


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

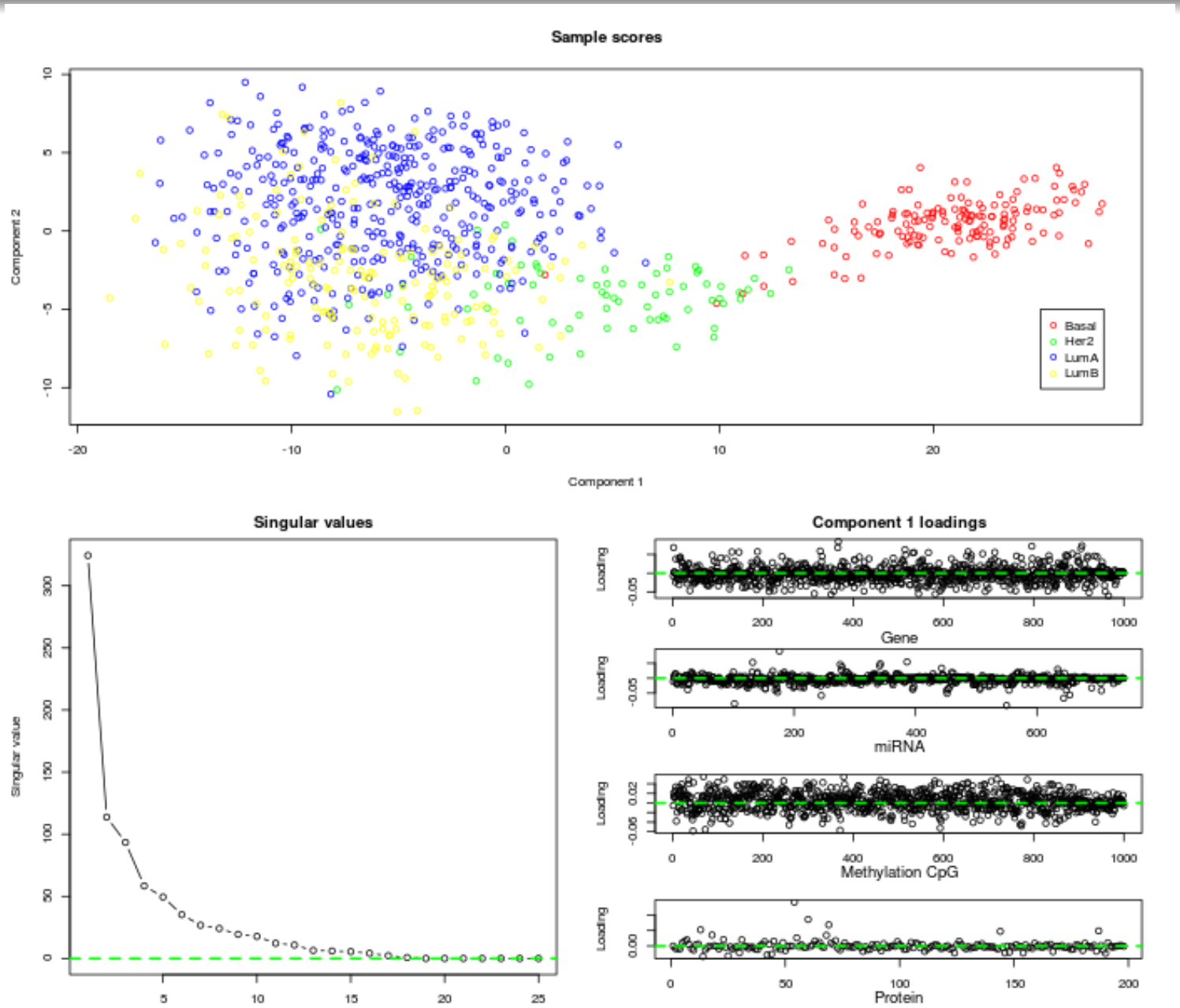

Singular values
Singular value
Component

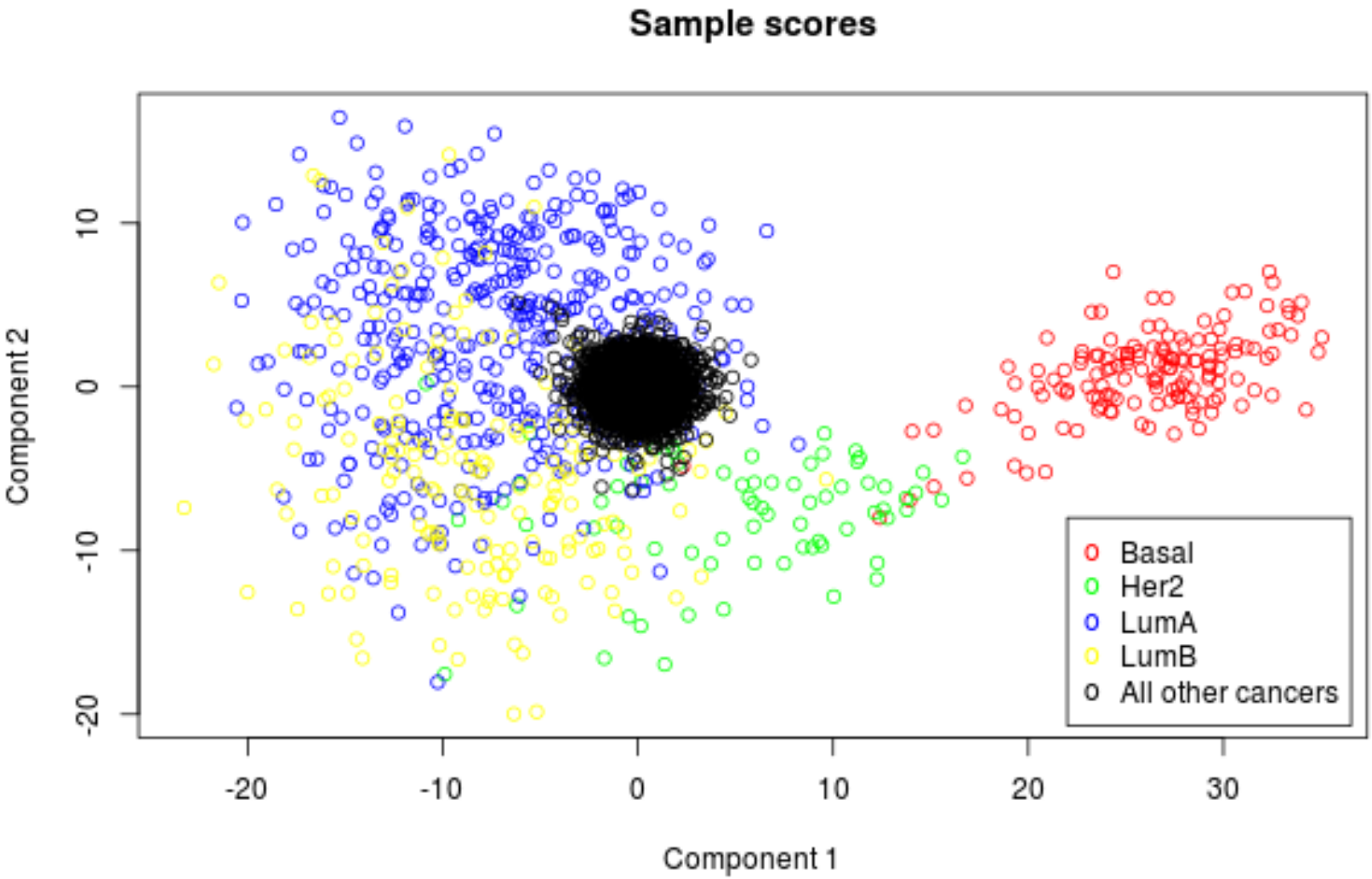

Component 1 loadings
X Chromosome
Y Chromosome
TSIX
XIST
Loading
Gene
X Chromosome
Loading
Methylation CpG

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. *Preprint*, arXiv:2002.0260, 2020.

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