Integrative Factorization of Bidimensionally Linked Matrices

Eric F. Lock

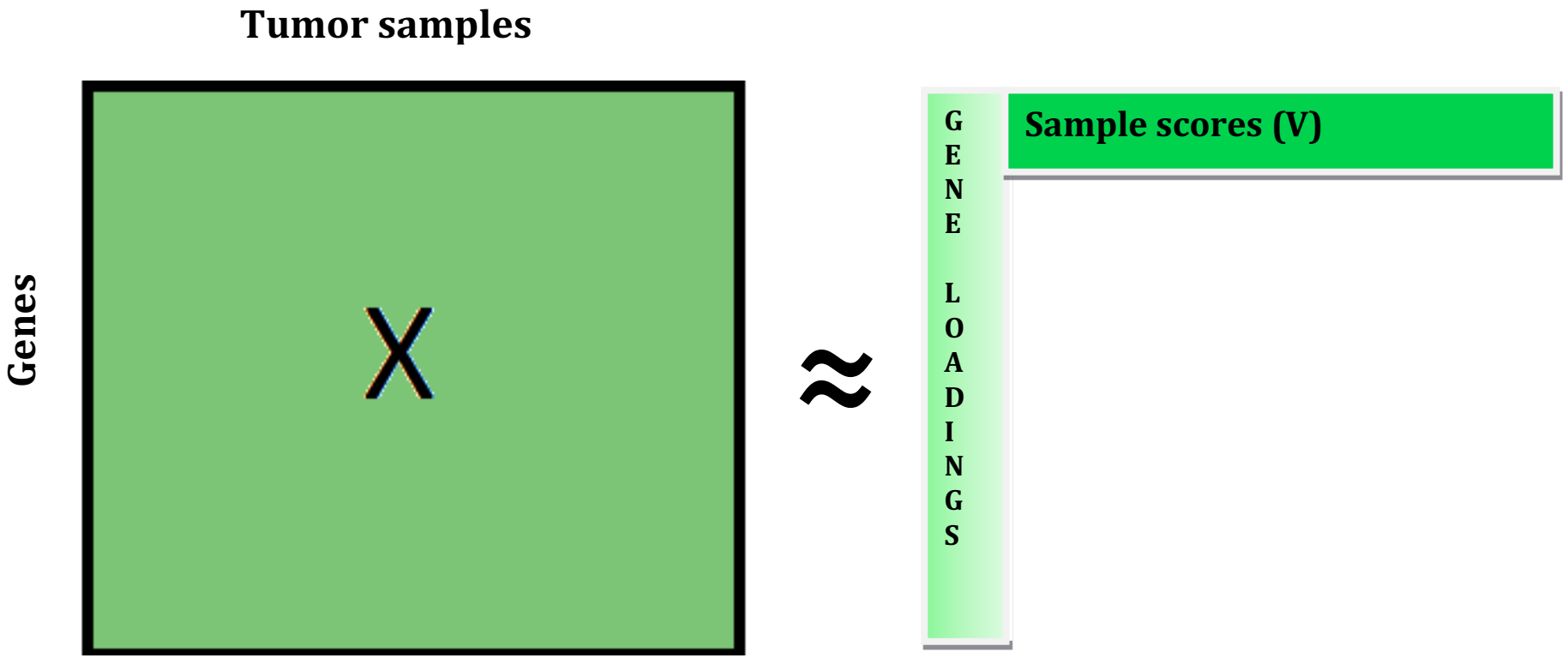
with **Jun Young Park**

University of Minnesota, Division of Biostatistics

JSM Denver, 07/30/2019

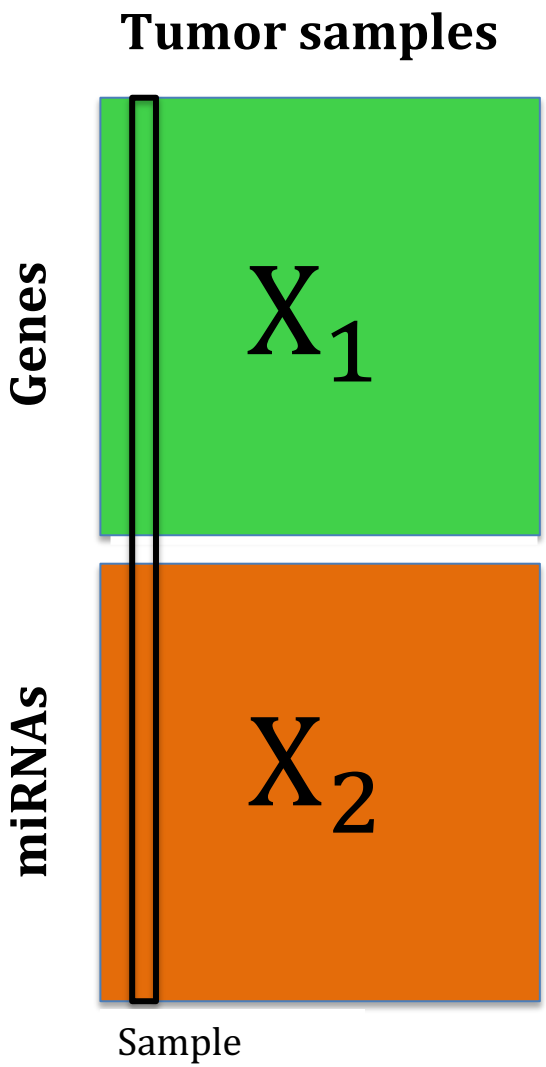
Matrix factorization

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

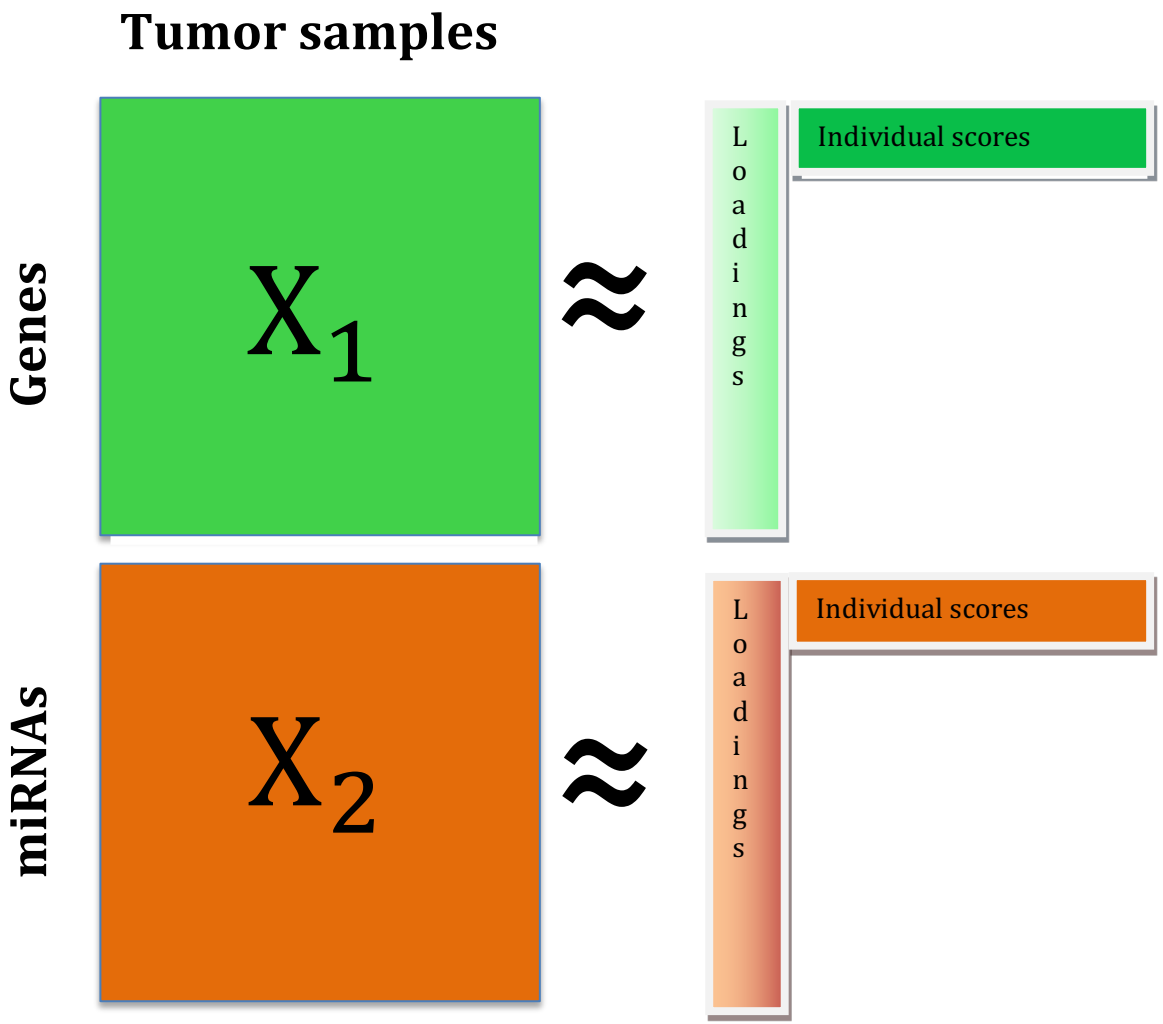


- Low rank factorization: $X \approx UV$, $U : m \times r$, $V : r \times n$.

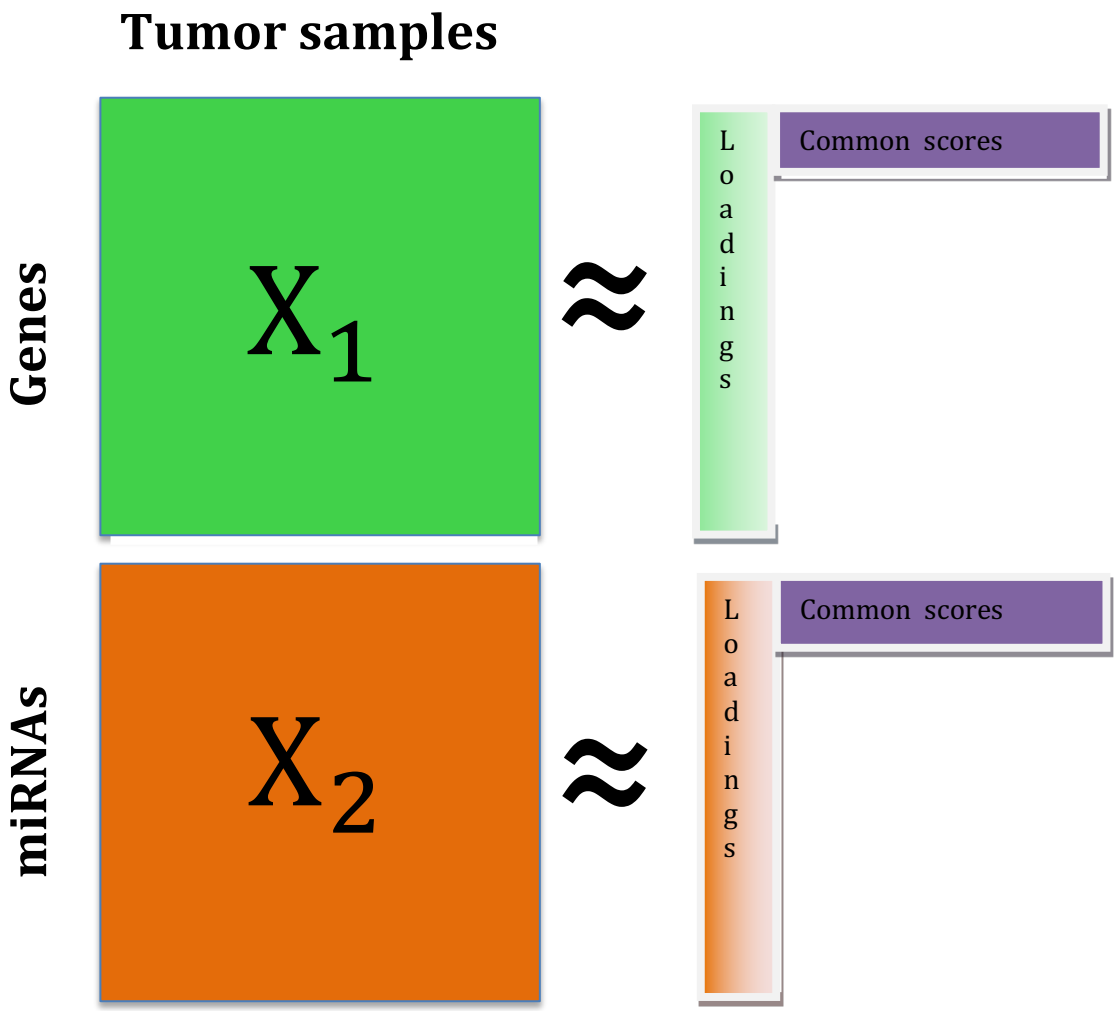
Vertically linked data



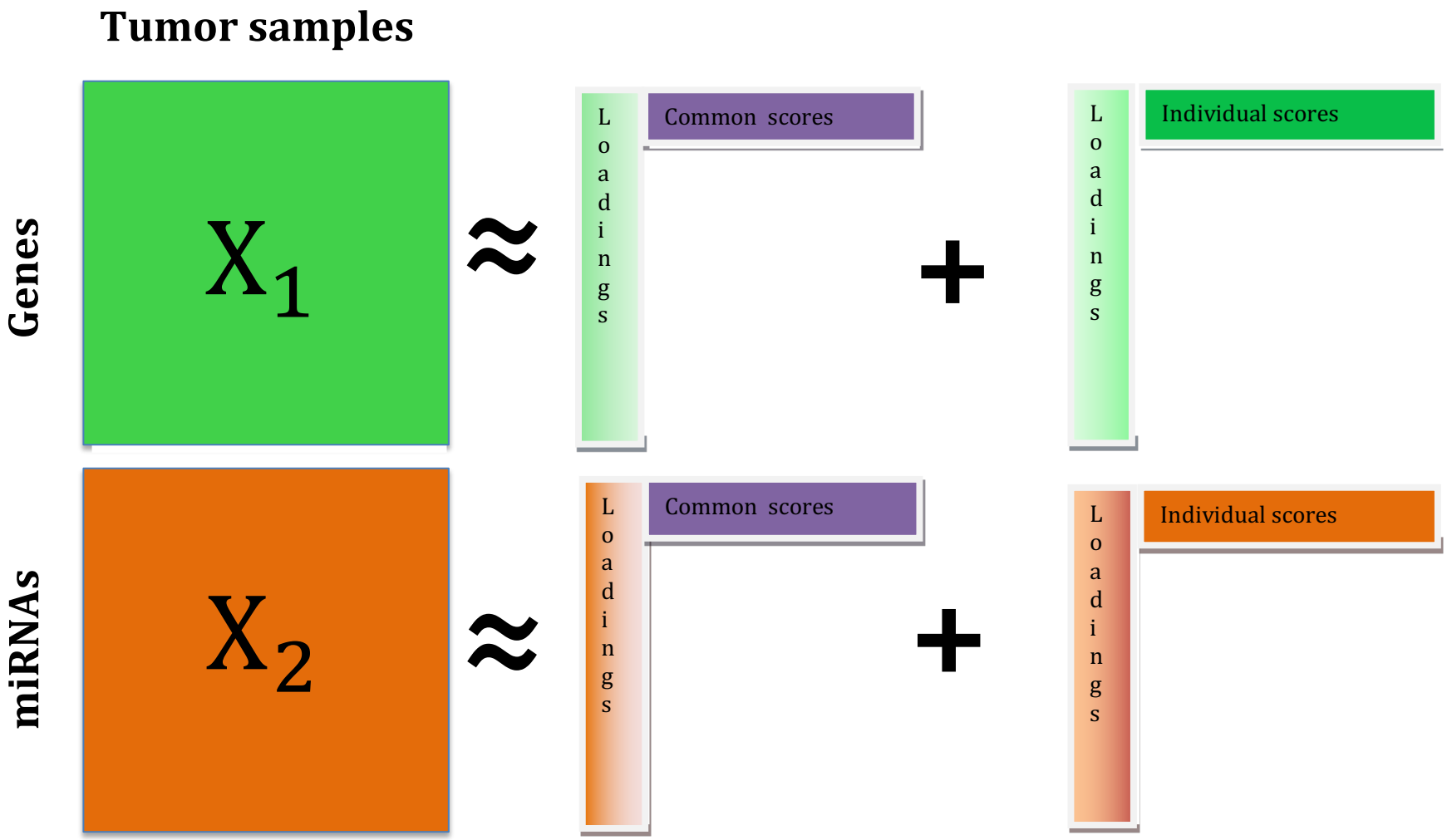
Vertically linked data: separate factorizations



Vertically linked data: joint factorization



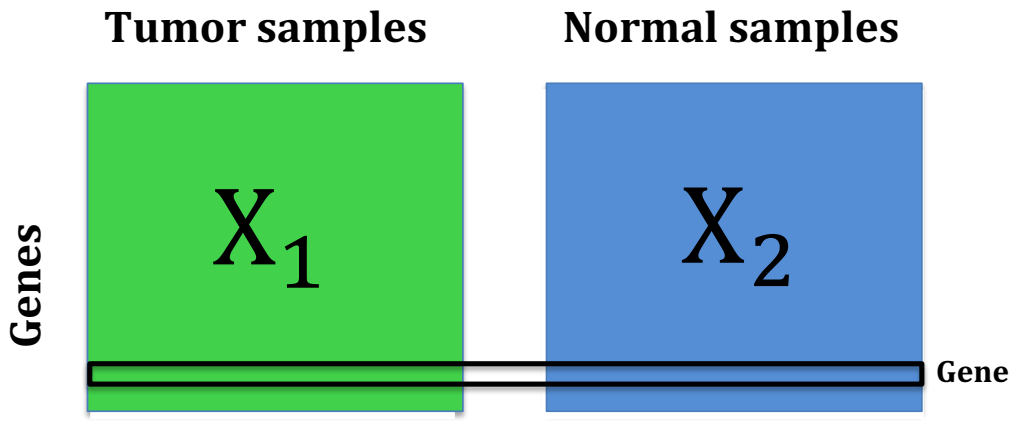
Vertically linked data: JIVE factorization



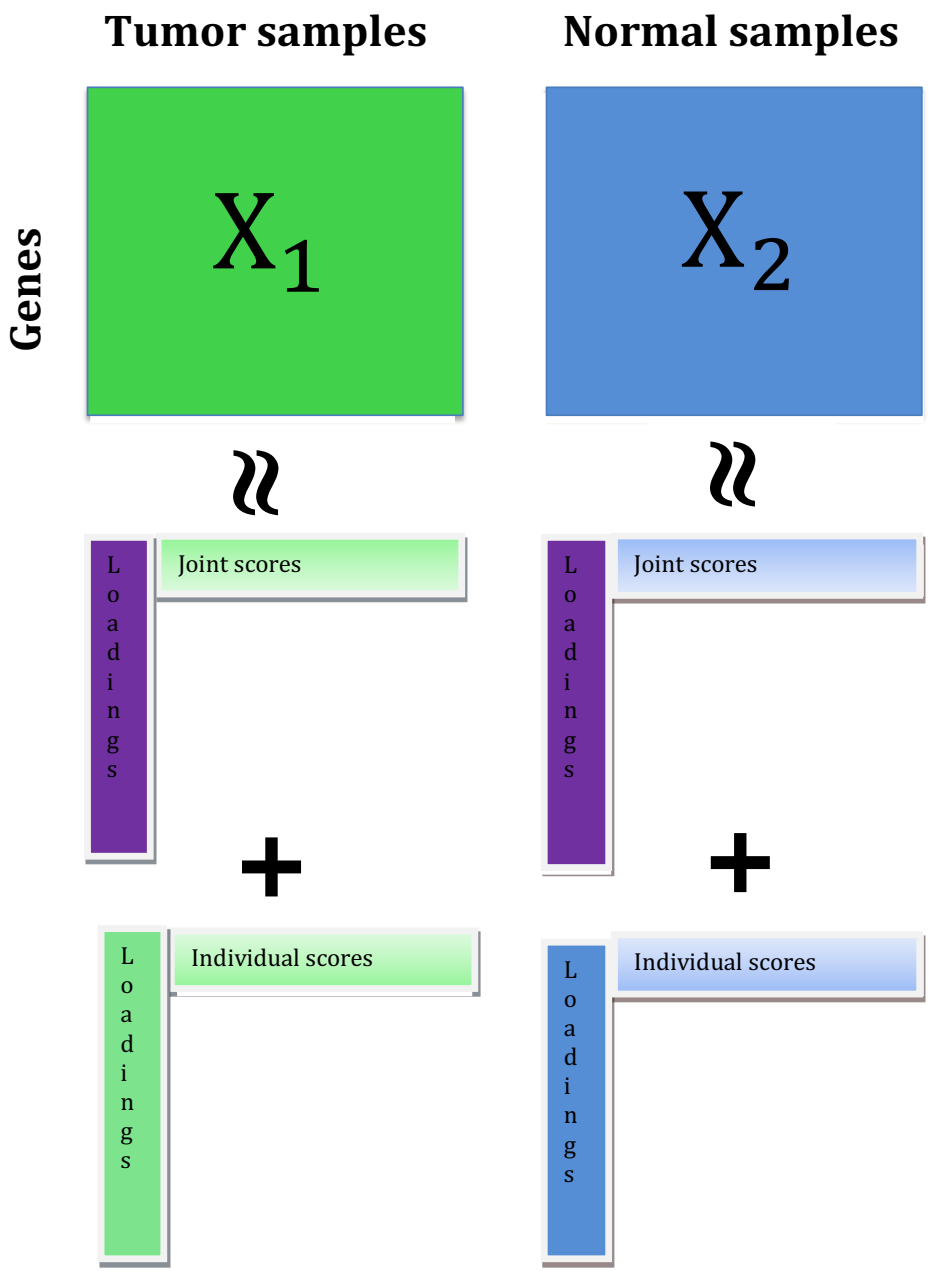
Joint + individual factorization methods

- JIVE [Lock, Hoadley, Marron, and Nobel, 2013]
- AJIVE [Feng, Jiang, Hannig and Marron, 2018]
- SLIDE [Gaynanova and Li, 2018]
- GIPCA [Zhu, Li, Lock, 2018]
  - See Session 598 (Wed 8:30-10:20 am)!
- COBE, SIFA, MOFA, & more!

Horizontally linked data



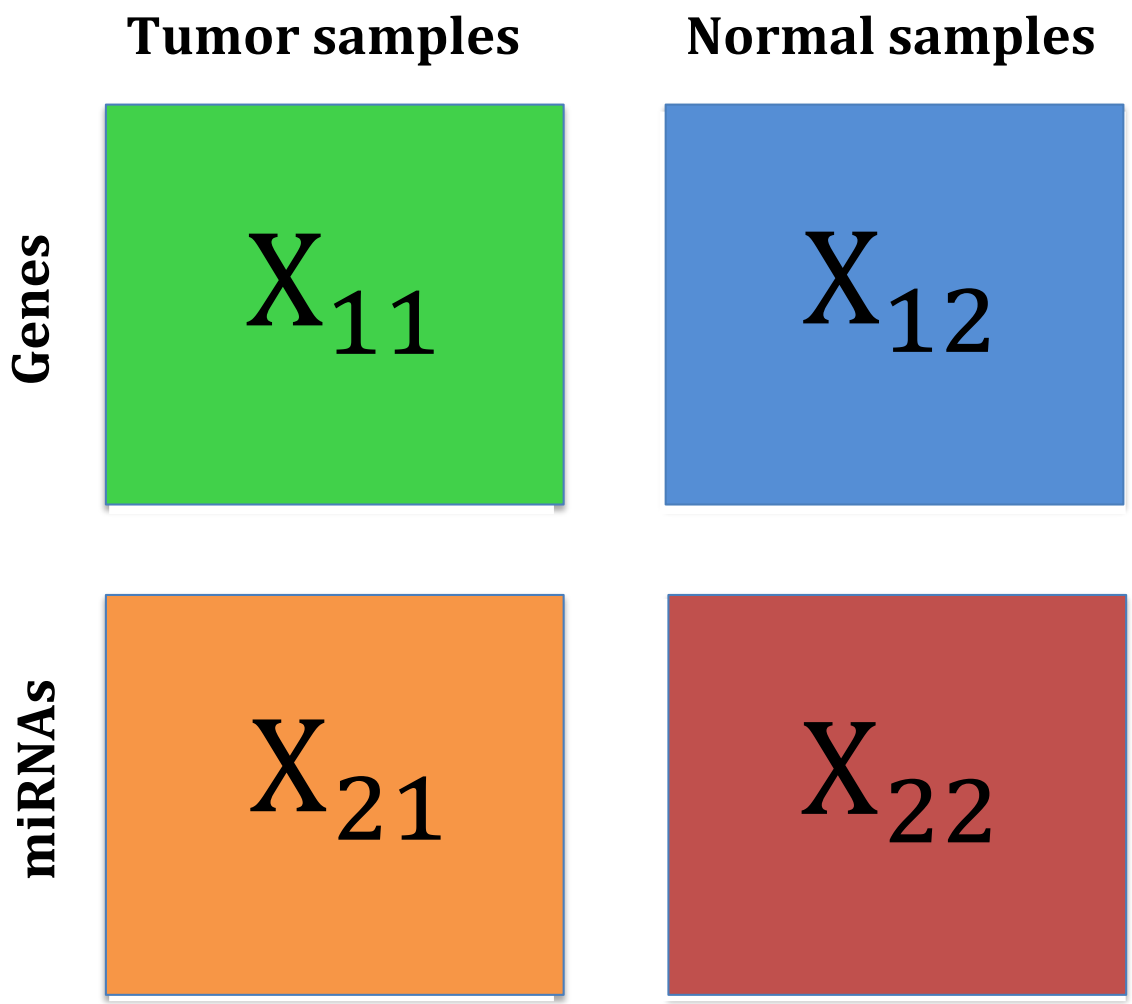
Horizontally linked data: JIVE factorization



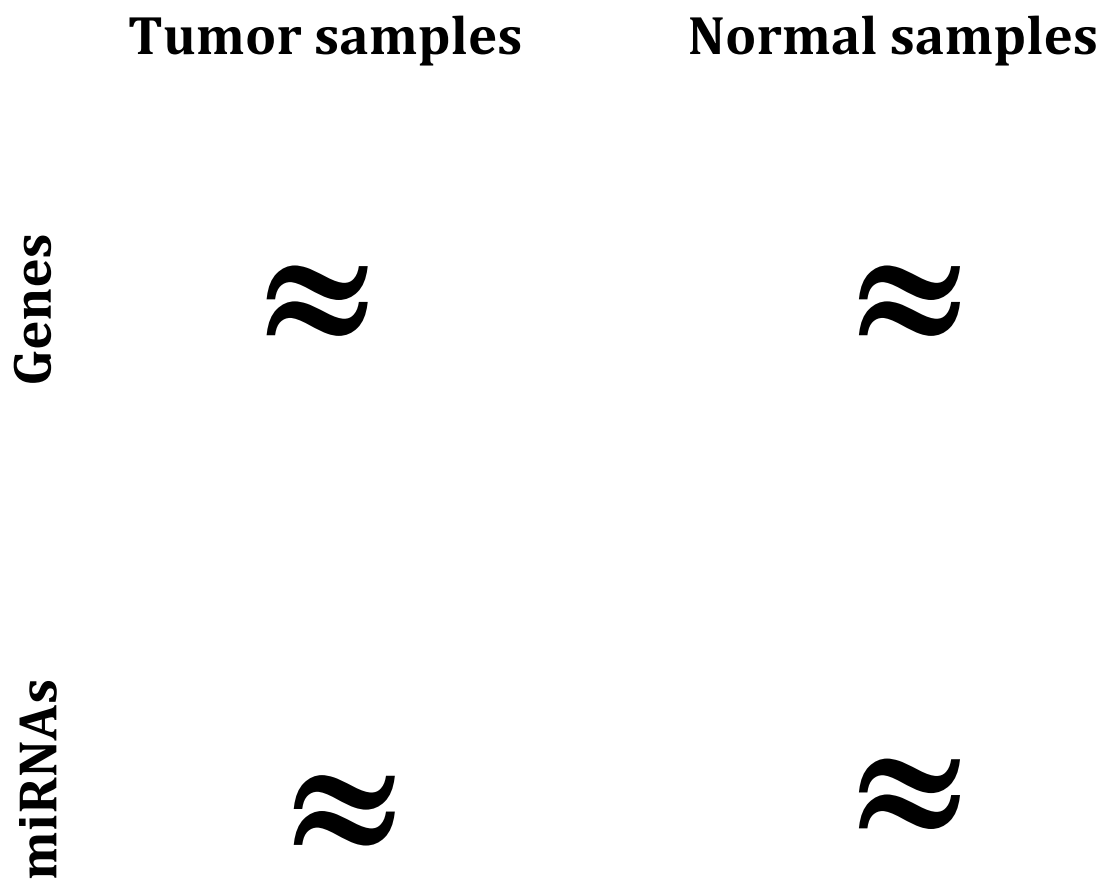
Motivating Data

- Gene expression and miRNA data for breast cancer tumor and normal tissue data from TCGA
  - 500 most variable genes
  - 500 most variable miRNA
  - 660 tumor samples
  - 86 independent normal samples

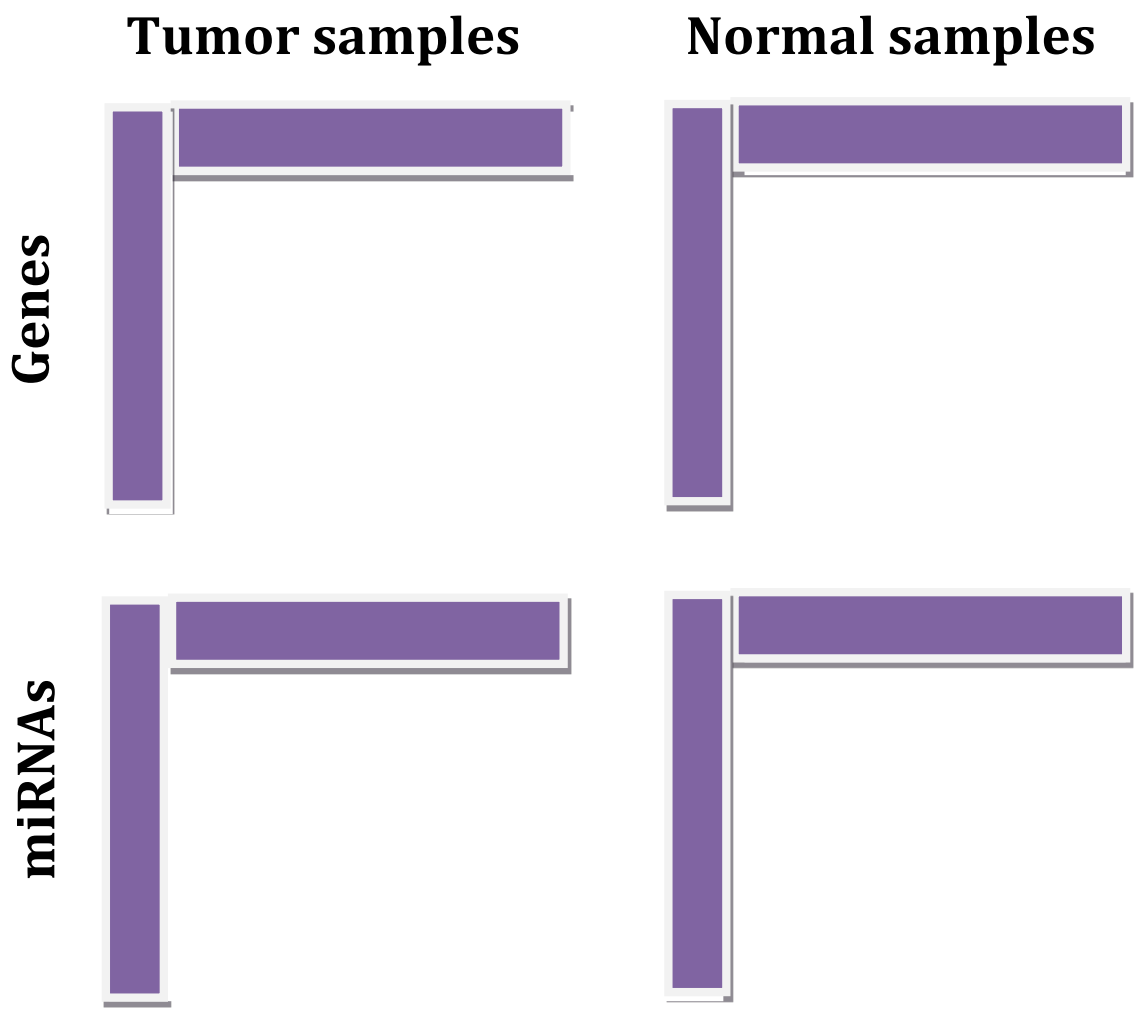
Bidimensionally linked data: BIDIFAC



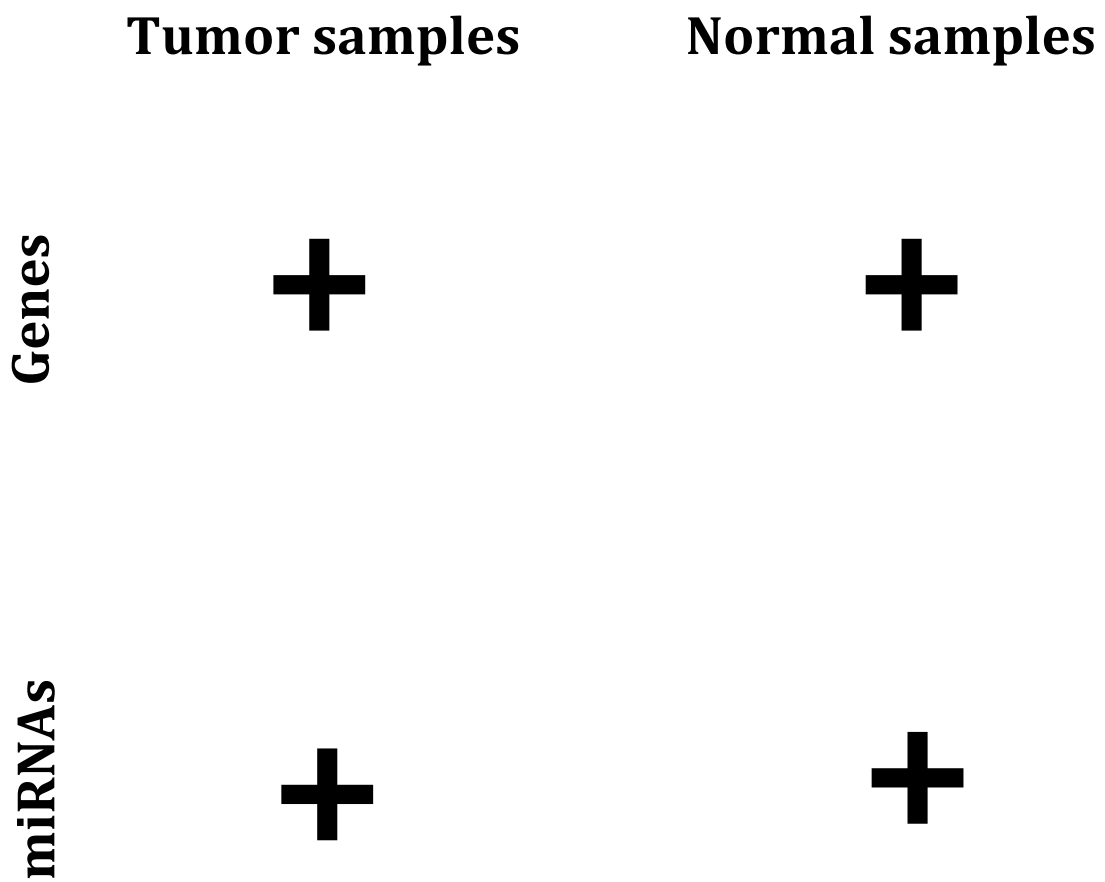
Bidimensionally linked data: BIDIFAC



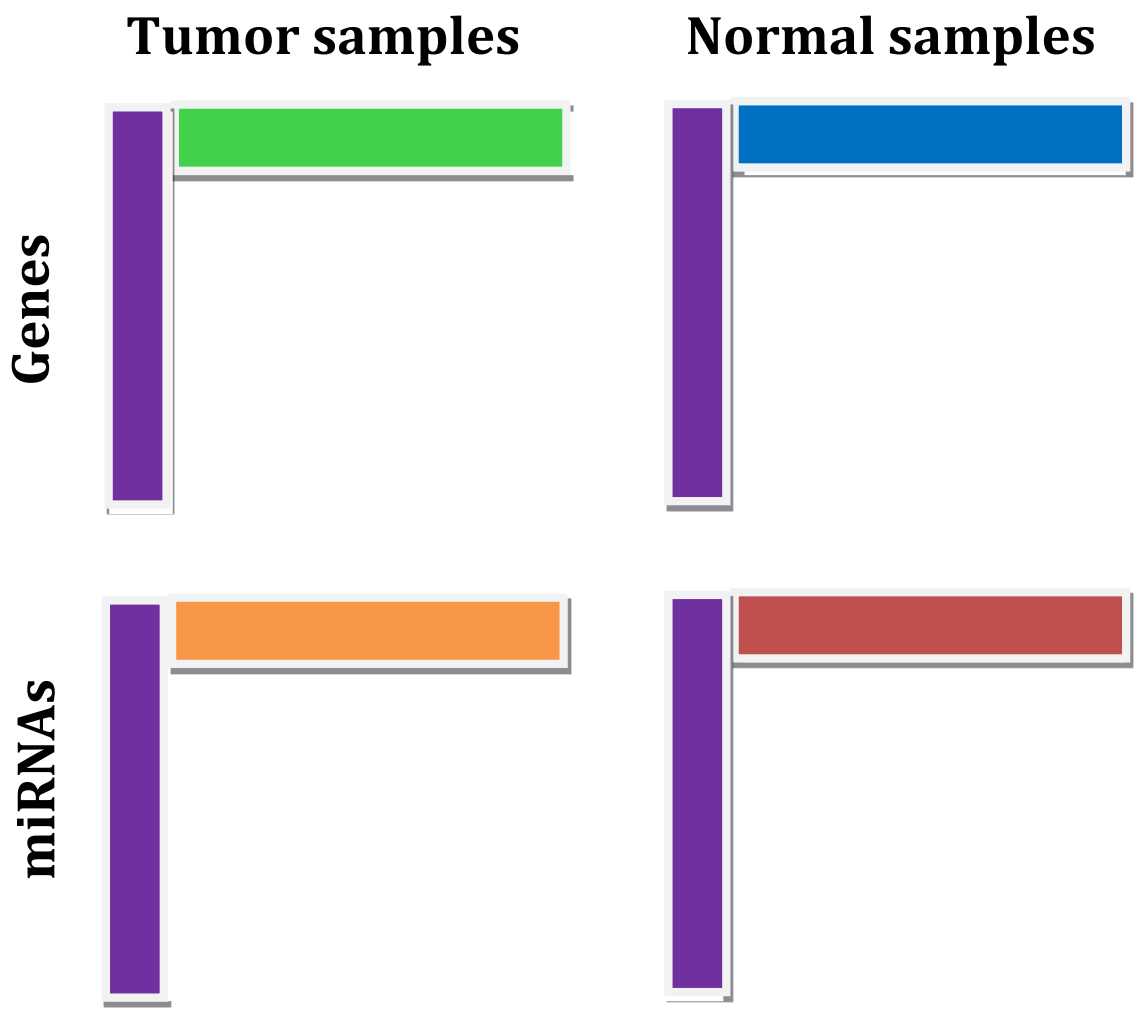
Bidimensionally linked data: BIDIFAC



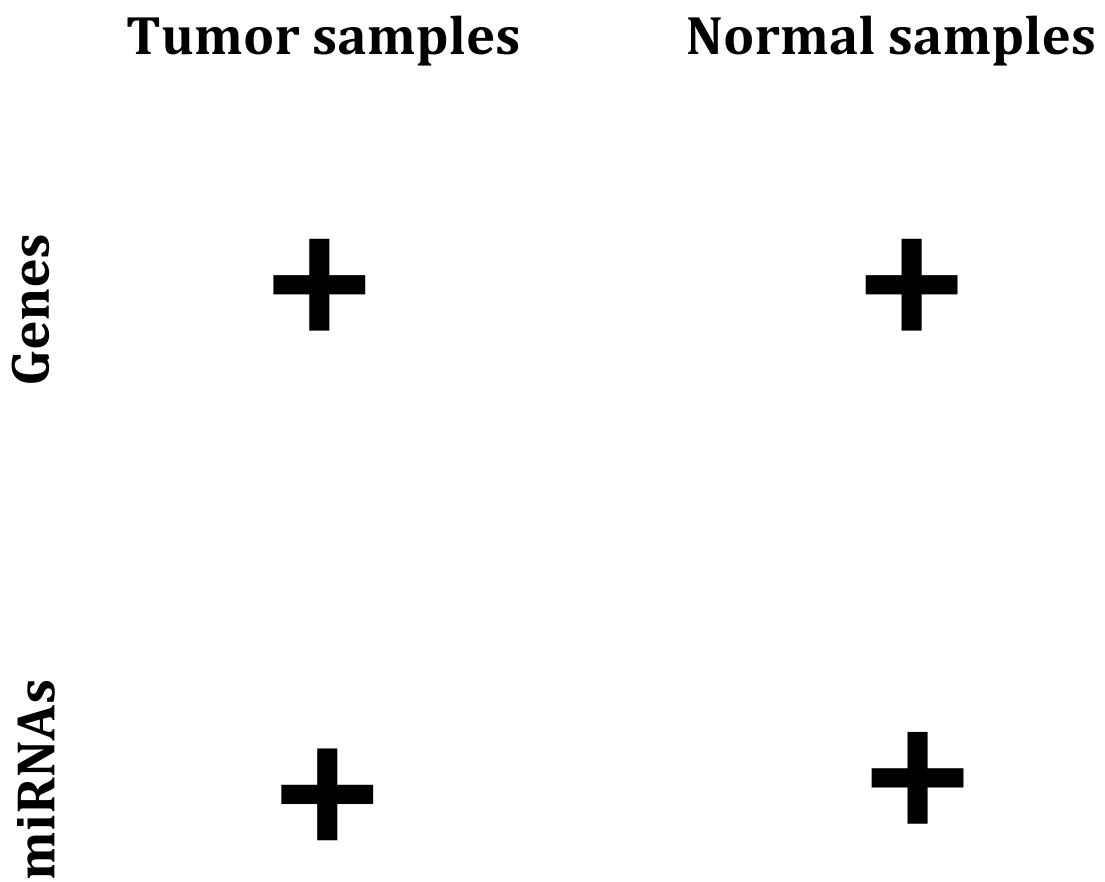
Bidimensionally linked data: BIDIFAC



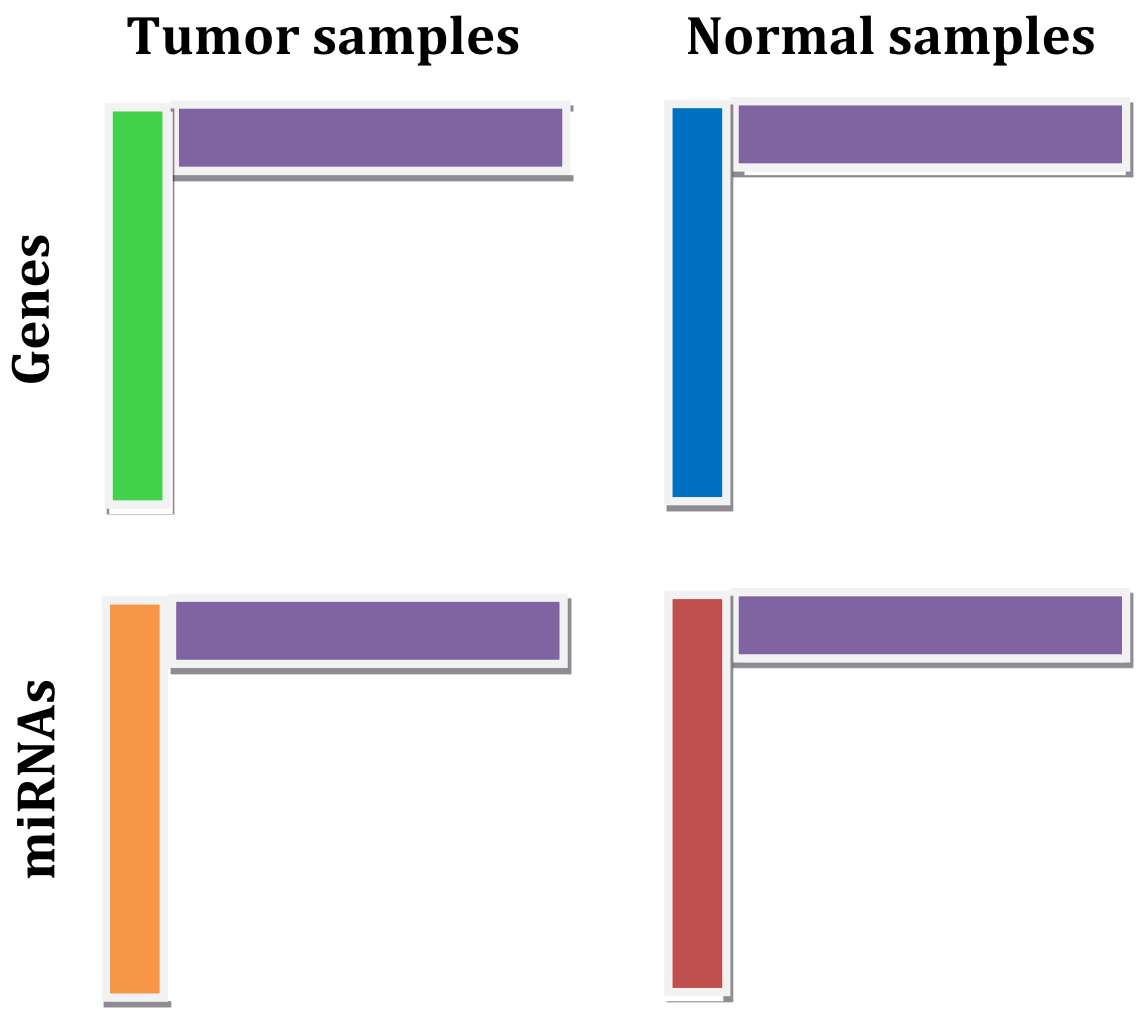
Bidimensionally linked data: BIDIFAC



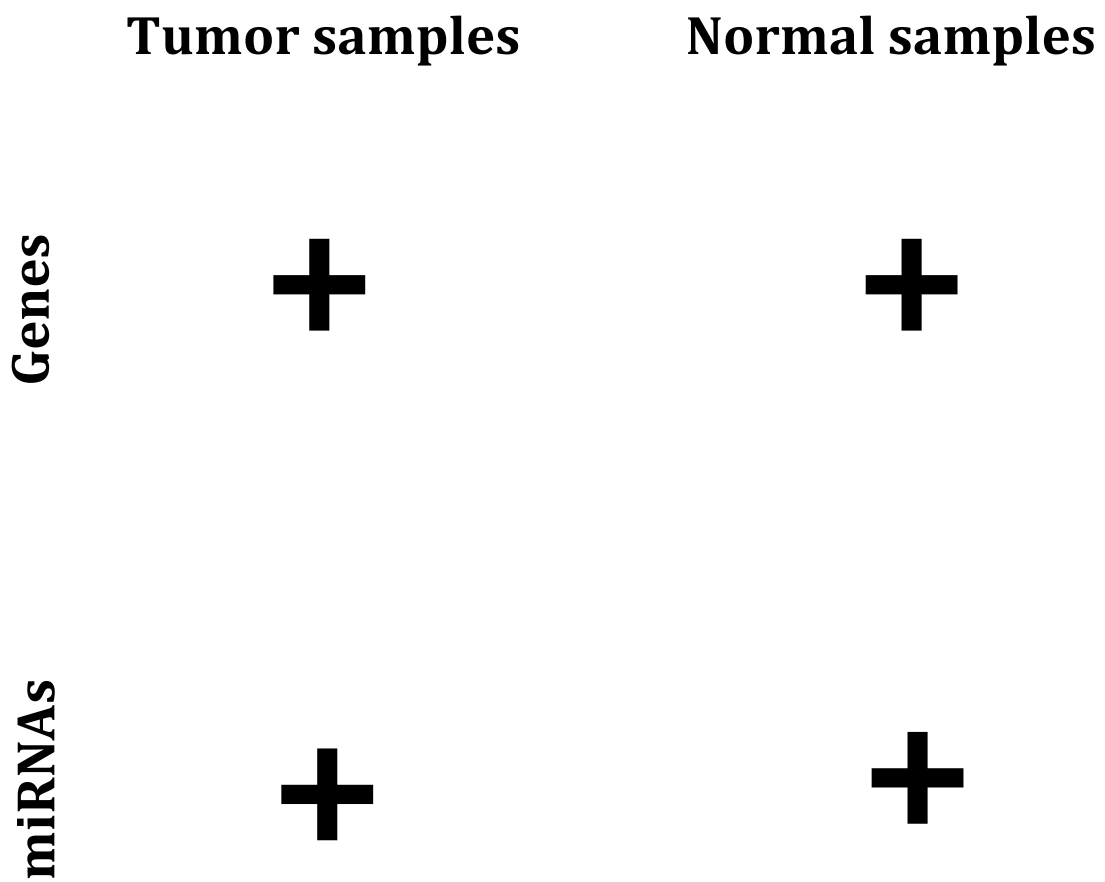
Bidimensionally linked data: BIDIFAC



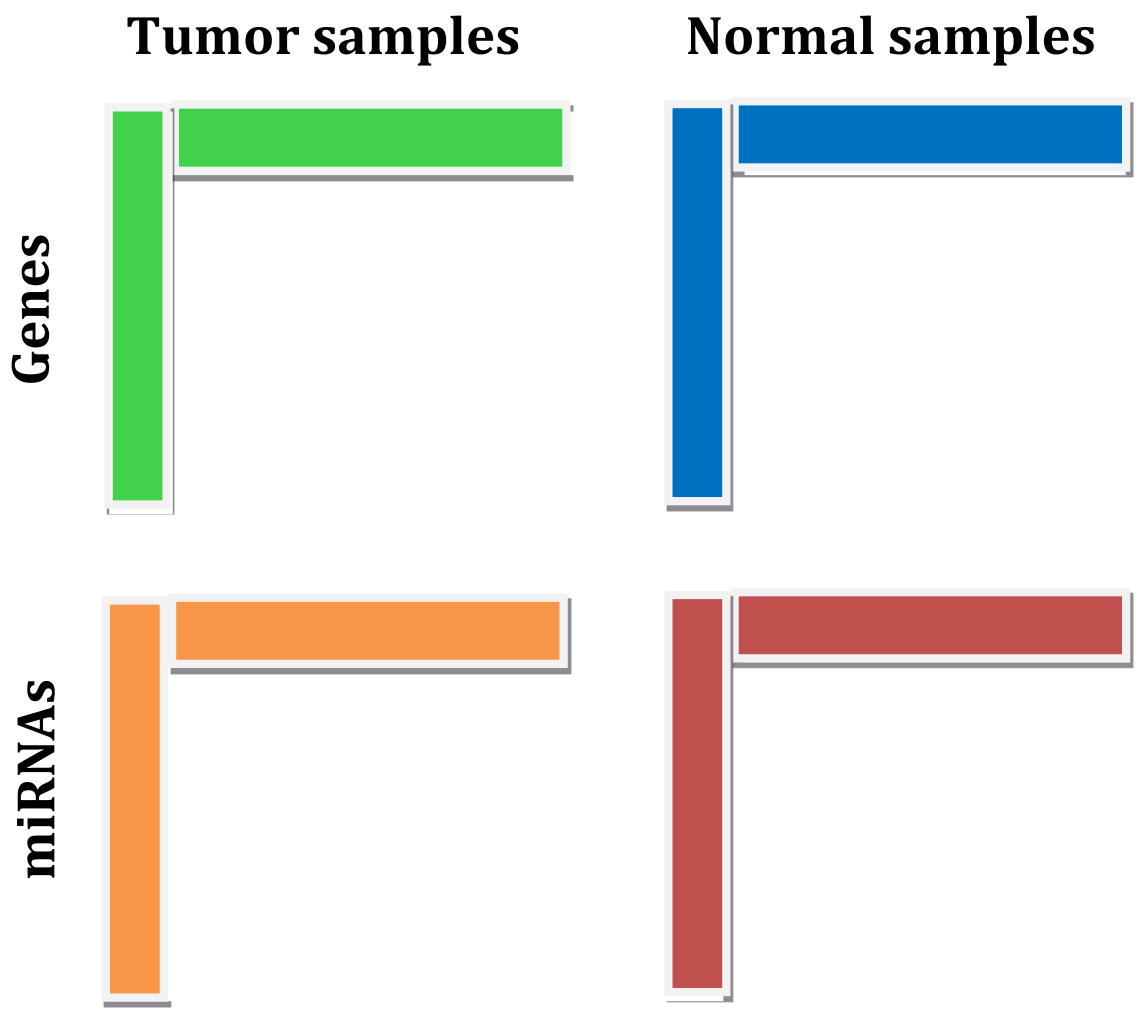
Bidimensionally linked data: BIDIFAC



Bidimensionally linked data: BIDIFAC



Bidimensionally linked data: BIDIFAC



BIDIFAC: 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

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

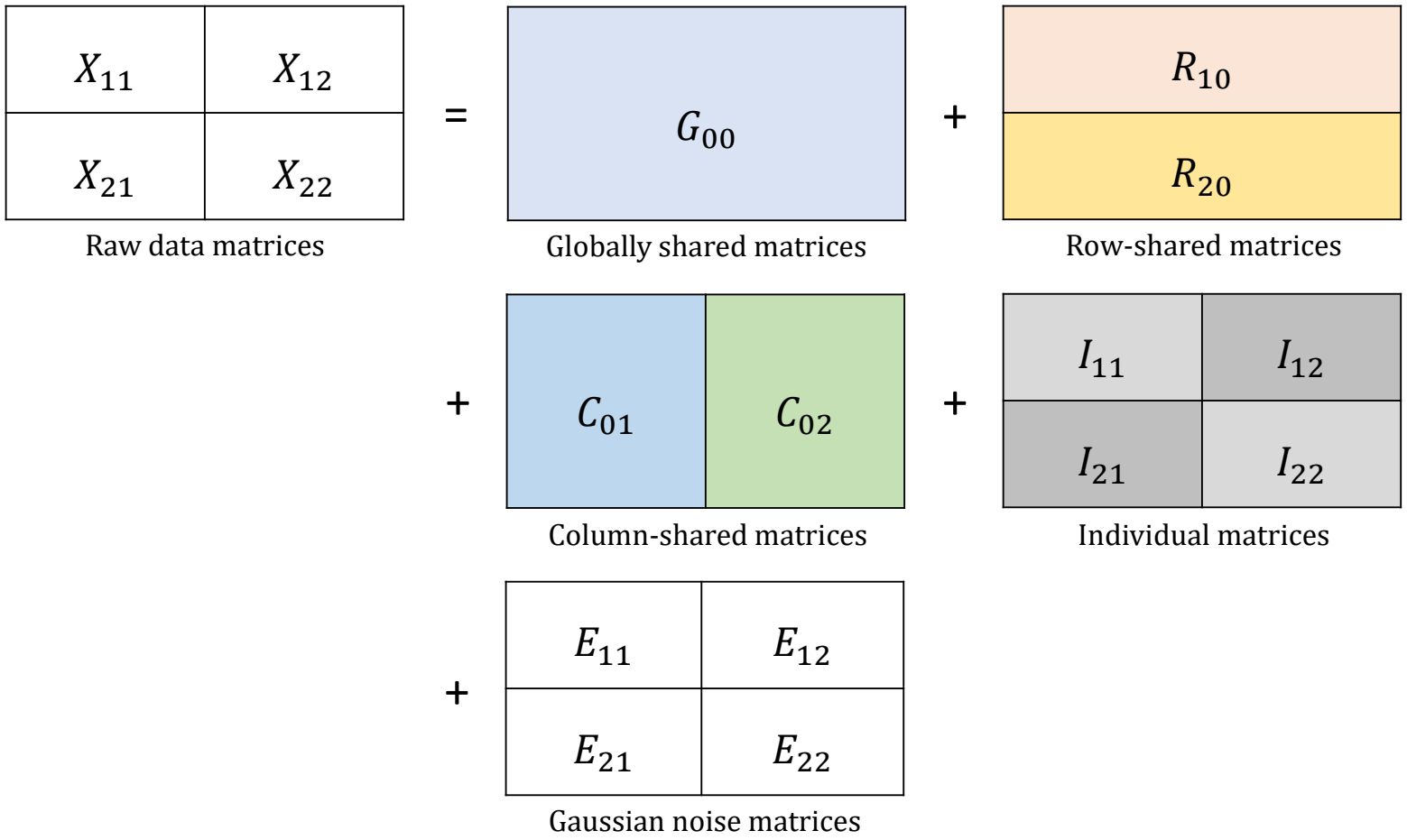
$$X_{i0} = [X_{i1}, \cdots, X_{iq}]$$

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

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

BIDIFAC: 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: objective

- Objective:

$$f_2(\{G_{ij}, R_{ij}, C_{ij}, I_{ij} \mid i = 1, \ldots, p, j = 1, \ldots, q\})$$
$$= \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{q} ||X_{ij} - G_{ij} - R_{ij} - C_{ij} - I_{ij}||_F^2$$
$$+ \lambda_{00} ||G_{00}||_* + \sum_{i=1}^{p} \lambda_{i0} ||R_{i0}||_* + \sum_{j=1}^{q} \lambda_{0j} ||C_{0j}||_* + \sum_{i=1}^{p} \sum_{j=1}^{q} \lambda_{ij} ||I_{ij}||_*.$$

- Where $||\cdot||$ defines the nuclear norm

$$\text{SVD}(A) = UDV^T \text{ with singular values } D[i,i] = d_i$$
$$\rightarrow ||A||_* = \sum_{i=1}^{\min\{m,n\}} d_i$$

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

BIDIFAC: Tuning parameters

- $(1 + p + q + pq)$ λ_{ij} parameters need to be determined!
- Conditions are *necessary* to have nonzero $\widehat{G}_{00}, \widehat{R}_{i0}, \widehat{C}_{0j}, \widehat{I}_{ij}$.
  - $\max_j \lambda_{ij} < \lambda_{i0} < \sum_j \lambda_{ij}$
  - $\max_i \lambda_{ij} < \lambda_{0j} < \sum_i \lambda_{ij}$
  - $\max_j \lambda_{0j} < \lambda_{00} < \sum_j \lambda_{0j}$
  - $\max_i \lambda_{i0} < \lambda_{00} < \sum_i \lambda_{i0}$

BIDIFAC: Tuning parameters

- Random matrix theory to automatically determine λ's
  - If $E : m \times n$ has independent sub-Gaussian entries with variance 1, $\sqrt{m} + \sqrt{n}$ gives a tight upper bound on the largest singular value of E
- 1.) Estimate variance of error E_{ij} for each matrix ij (MAD)
- 2.) Scale each X_{ij} to have error variance 1
- 3.) Set penalties as follows
  - $\lambda_{00} = \sqrt{m_0} + \sqrt{n_0}$
  - $\lambda_{i0} = \sqrt{m_i} + \sqrt{n_0}$
  - $\lambda_{0j} = \sqrt{m_0} + \sqrt{n_j}$
  - $\lambda_{ij} = \sqrt{m_i} + \sqrt{n_j}$
- Guaranteed to satisfy necessary conditions for non-zero solution

BIDIFAC: dual L_2 objective

- Equivalent form of BIDIFAC objective:

$$f_1(\{\mathbf{U}_{ij}^{(\cdot)}, \mathbf{V}_{ij}^{(\cdot)} \mid i = 0, \ldots, p, j = 0, \ldots, q, (i,j) \neq (0,0)\})$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{q} ||\mathbf{X}_{ij} - \mathbf{U}_{i0}^{(G)}\mathbf{V}_{0j}^{(G)T} - \mathbf{U}_{i0}^{(R)}\mathbf{V}_{ij}^{(R)T} - \mathbf{U}_{ij}^{(C)}\mathbf{V}_{0j}^{(C)T} - \mathbf{U}_{ij}^{(I)}\mathbf{V}_{ij}^{(I)T}||_F^2$$

$$+\lambda_{00}(||\mathbf{U}_{00}^{(G)}||_F^2 + ||\mathbf{V}_{00}^{(G)}||_F^2) + \sum_{i=1}^{p}\lambda_{i0}(||\mathbf{U}_{i0}^{(R)}||_F^2 + ||\mathbf{V}_{i0}^{(R)}||_F^2)$$

$$+\sum_{j=1}^{q}\lambda_{0j}(||\mathbf{U}_{0j}^{(C)}||_F^2 + ||\mathbf{V}_{0j}^{(C)}||_F^2) + \sum_{i=1}^{p}\sum_{j=1}^{q}\lambda_{ij}(||\mathbf{U}_{ij}^{(I)}||_F^2 + ||\mathbf{V}_{ij}^{(I)}||_F^2)$$

- Gives posterior mode of Bayesian model where
  - Errors E_{ij} are iid $N(0,1)$
  - Entries of $\mathbf{U}_{ij}^{(\cdot)}, \mathbf{V}_{ij}^{(\cdot)}$ are iid $N(0, 1/\lambda_{ij})$
- Motivates MAP imputation for missing data

Data Analysis: TCGA Breast Cancer Data

Proportion of variance explained & (rank):

	Global	Global+Row	Global+Col	Global+Row+Col	Signal
Tumor mRNA	0.14 (34)	0.32 (68)	0.45 (93)	0.58 (127)	0.67 (173)
NAT mRNA	0.23 (34)	0.50 (68)	0.44 (41)	0.66 (75)	0.78 (83)
Tumor miRNA	0.09 (34)	0.46 (67)	0.30 (93)	0.63 (126)	0.76 (175)
NAT miRNA	0.13 (34)	0.66 (67)	0.24 (41)	0.75 (74)	0.76 (79)

Data Analysis: TCGA Breast Cancer Data

Model	Components	Rank	SWISS	p-value
BIDIFAC	Signal	173	0.54	0.002
	Global	34	0.69	0.046
	Row	34	0.75	0.085
	Col+Indiv	105	0.52	0.003
	Col	59	0.48	0.029
	Indiv	46	0.79	0.003

- SWISS score: normalized variability within clinical subtypes.
- p-value: tests if the set of factor scores of the estimated parameters are associated with patients' survival.

Data Analysis: TCGA Breast Cancer Data

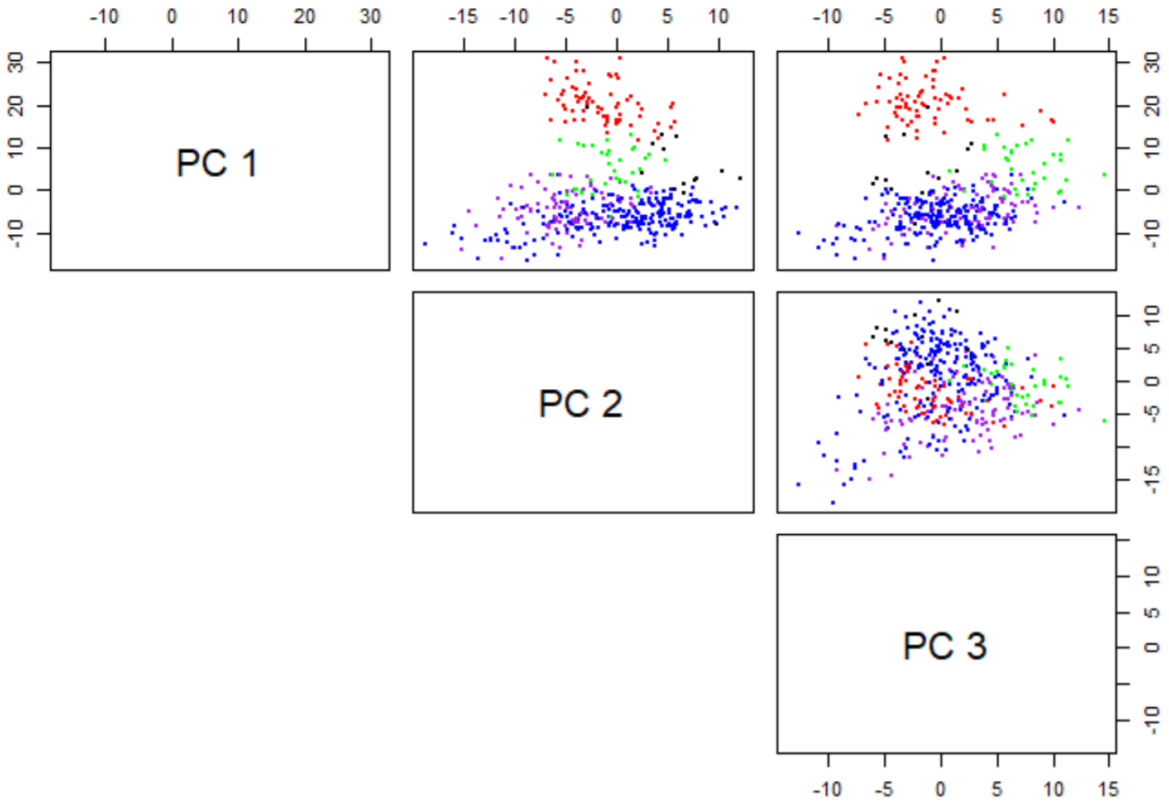
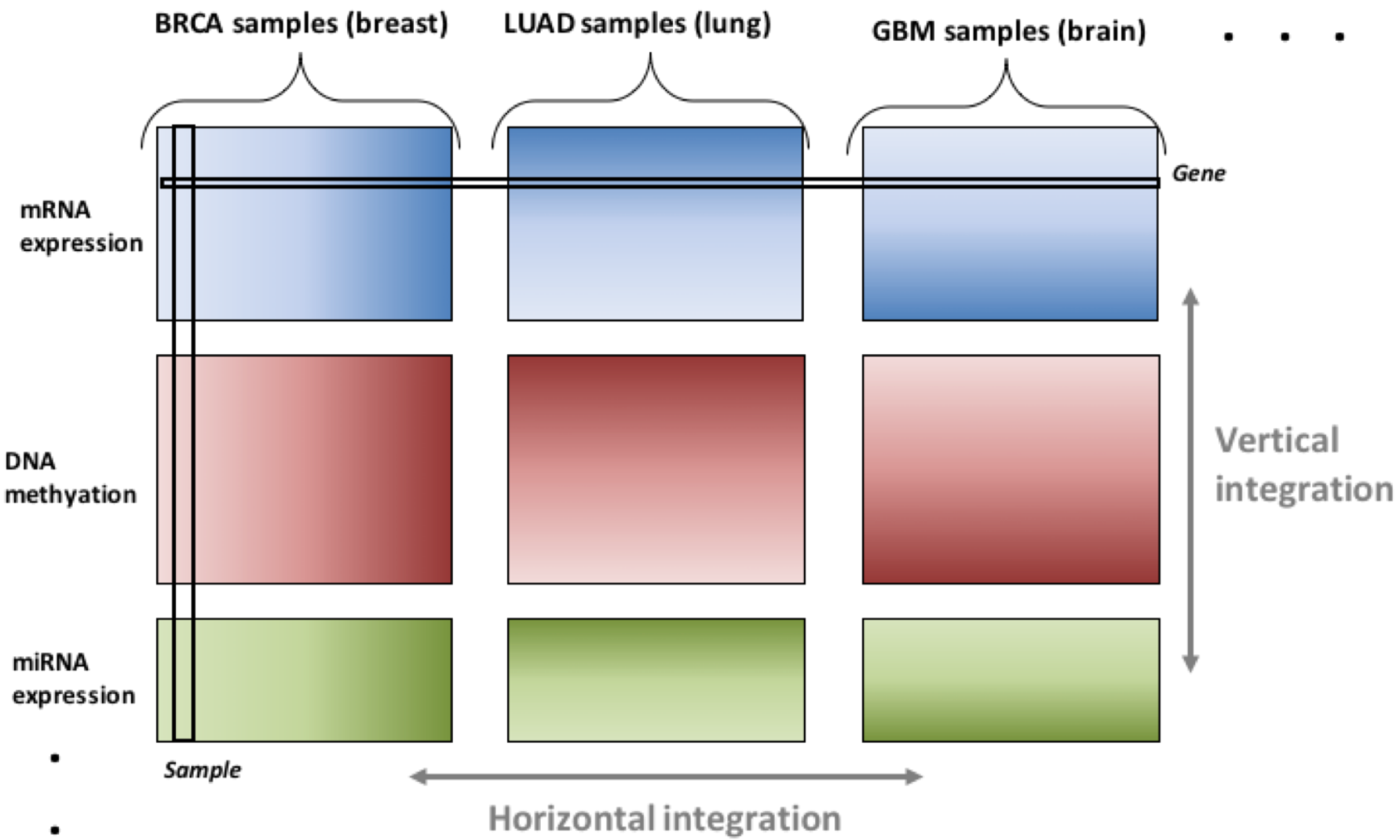


Figure: Principal components of the estimated column-shared structure, colored by subtype: Basal, HER2, Lum A, Lum B.

Future work

- Pan-omics pan-cancer integration!



Thank you!

- Support: NCI grant R21CA231214-01
- Slides: http://ericfrazerlock.com/Talks.html
- Code: https://github.com/lockEF/bidifac
- J Park and EF Lock. Integrative Factorization of Bidimensionally Linked Matrices. *arXiv*:1906.03722, 2019.
- **See also**: MJ O'Connell and EF Lock. Linked Matrix Factorization. *Biometrics*, doi: 10.1111/biom.13010, 2018.
  - Session 461, Wed 8:30-10:20am!