Empirical Bayes Linked Matrix Decomposition

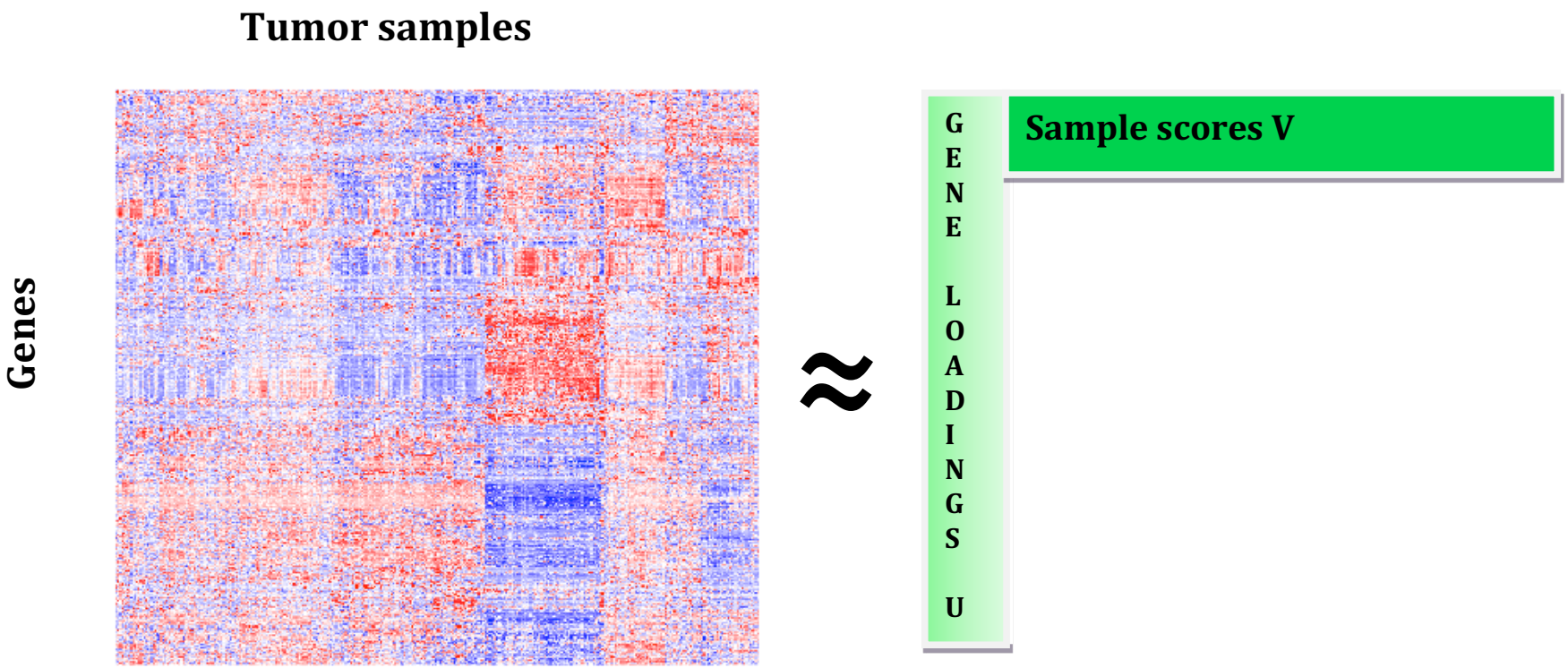
Eric F. Lock

University of Minnesota
Division of Biostatistics and Health Data Science

ENAR, New Orleans, 03/26/2024

Low-rank matrix approximation

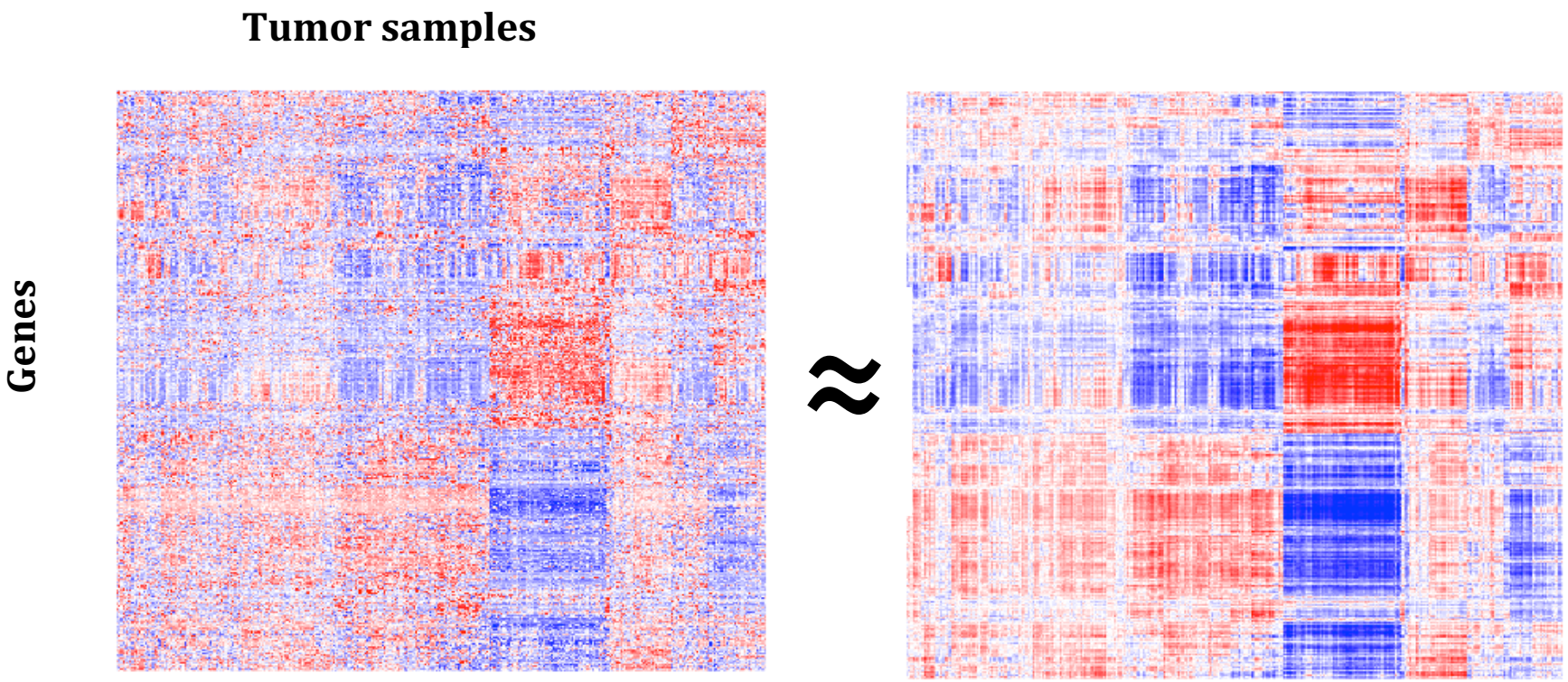
- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



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

Matrix factorization: nuclear norm

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



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

Low-rank matrix approximation

- $X = \mathbf{A} + E$ where rank($\mathbf{A}$)=r and noise E
- Singular value decomposition (SVD): $X = UDV^T$
  - D is diagonal with singular values $d_i = D[i, i]$
  - captures **signal** and noise:

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

Low-rank matrix approximation

- Approach 1: hard-thresholding
  - Minimize $\frac{1}{2}||X - \hat{X}||_F^2$ for $\text{rank}(\hat{X}) = r$
  - Then $\hat{X} = U\hat{D}V^T$ where $\hat{d}_i = \begin{cases} d_i \text{ for } i \leq r \\ 0 \text{ for } i > r \end{cases}$
  - Need to select r

- Over-fits! $\hat{D} = \text{diag}(\mathbf{a_1} + e_1, \cdots, \mathbf{a_r} + e_r, 0, \cdots)$

Low-rank matrix approximation

- Approach 2: soft-thresholding
  - Minimize $\frac{1}{2}||X - \hat{X}||_F^2 + \lambda||\hat{X}||_*$
  - $||\cdot||_*$ is the nuclear norm: $||\hat{X}||_* = \sum \hat{d}_i$
  - Then $\hat{X} = U\hat{D}V^T$ where $\hat{d}_i = max(d_i - \lambda, 0)$.
- Need to select λ
  - Assume $E : m \times n$ has iid $N(0, \sigma^2)$ entries
  - Then, largest singular value of $E \approx \sigma(\sqrt{m} + \sqrt{n})$
  - Set $\lambda = \hat{\sigma}(\sqrt{m} + \sqrt{n})$
- Over-shrinks! $\hat{D} = \text{diag}(\mathbf{a_1} + e_1 - \lambda, \cdots, \mathbf{a_r} + e_r - \lambda, 0, \cdots)$

Low-rank matrix approximation

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

$$||X - \tilde{U}\tilde{V}||_F^2 + \lambda(||\tilde{U}||_F^2 + ||\tilde{V}||_F^2)$$

for $\tilde{U} : m \times r'$ $\tilde{V} : r' \times n$ and r' sufficiently large

- Posterior mode with $N(0, \sigma^2/\lambda)$ priors on $\tilde{U}$ and $\tilde{V}$

- More flexible model:

$$\tilde{U}[:, r] \sim N(\mathbf{0}, \sigma\tau_u^2[r]\mathbf{I}), \;\; \tilde{V}[:, r] \sim N(\mathbf{0}, \sigma\tau_v^2[r]\mathbf{I})$$

Low-rank matrix approximation

- Approach 3: Empirical variational Bayes
  - Approximate posterior with $q(\tilde{U}, \tilde{V}) = q_u(\tilde{U})q_v(\tilde{V})$
  - Minimize free energy

$$E_q \log \frac{q_u(\tilde{U})q_v(\tilde{V})}{p(X \mid \tilde{U}, \tilde{V}, \sigma)p(\tilde{U} \mid \tau_U)p(\tilde{V} \mid \tau_V)}$$

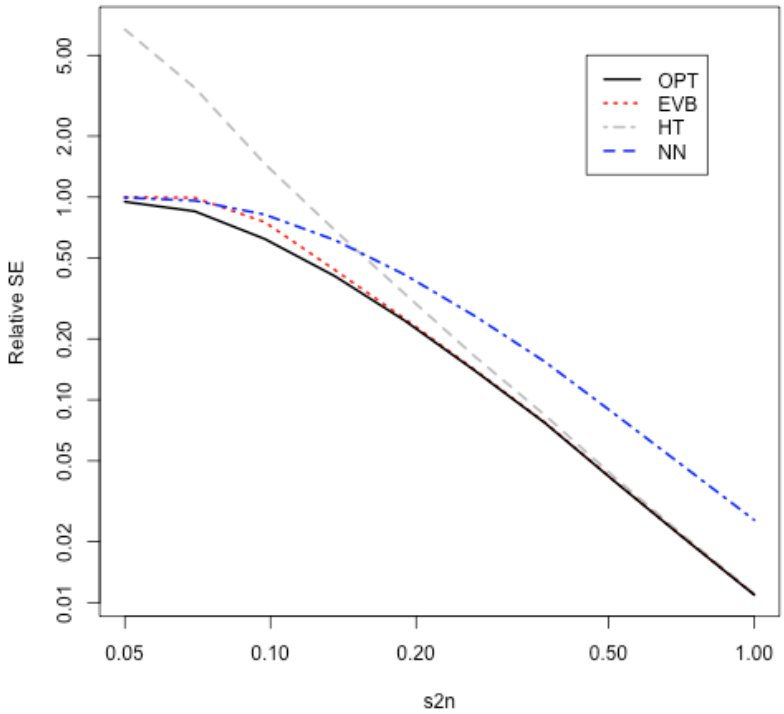
    over $\sigma, \tau_U, \tau_v, q_u, q_v$.

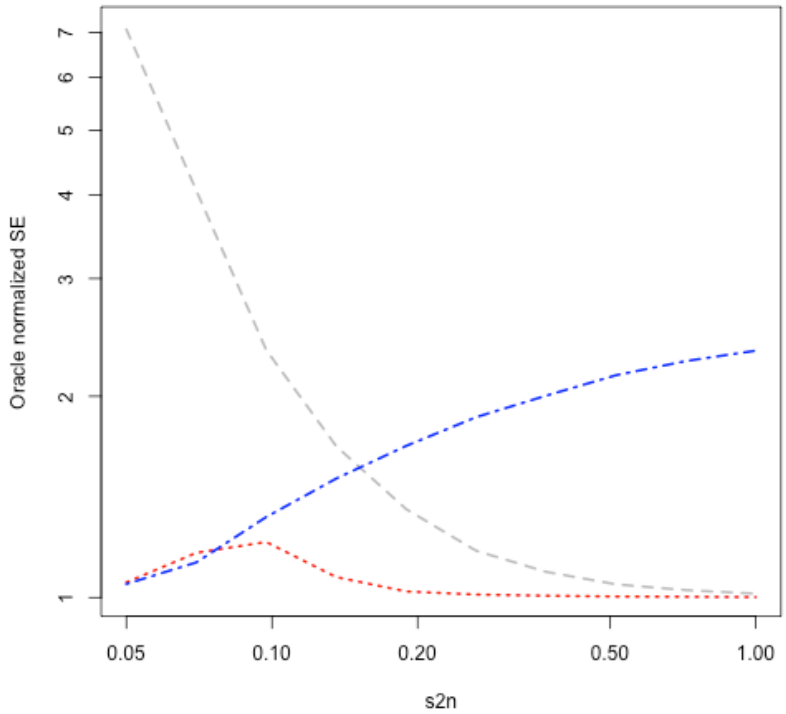
  - $\hat{X} = E_q(\tilde{U}\tilde{V}^T) = U\hat{D}V^T$ where $\hat{d}_i = f(d_i)$ for closed-form $f(\cdot)$

- No tuning parameters
- Just right! $\quad \hat{D} \approx \text{diag}(\mathbf{a_1}, \cdots, \mathbf{a_r}, 0, \cdots)$

Low-rank matrix approximation

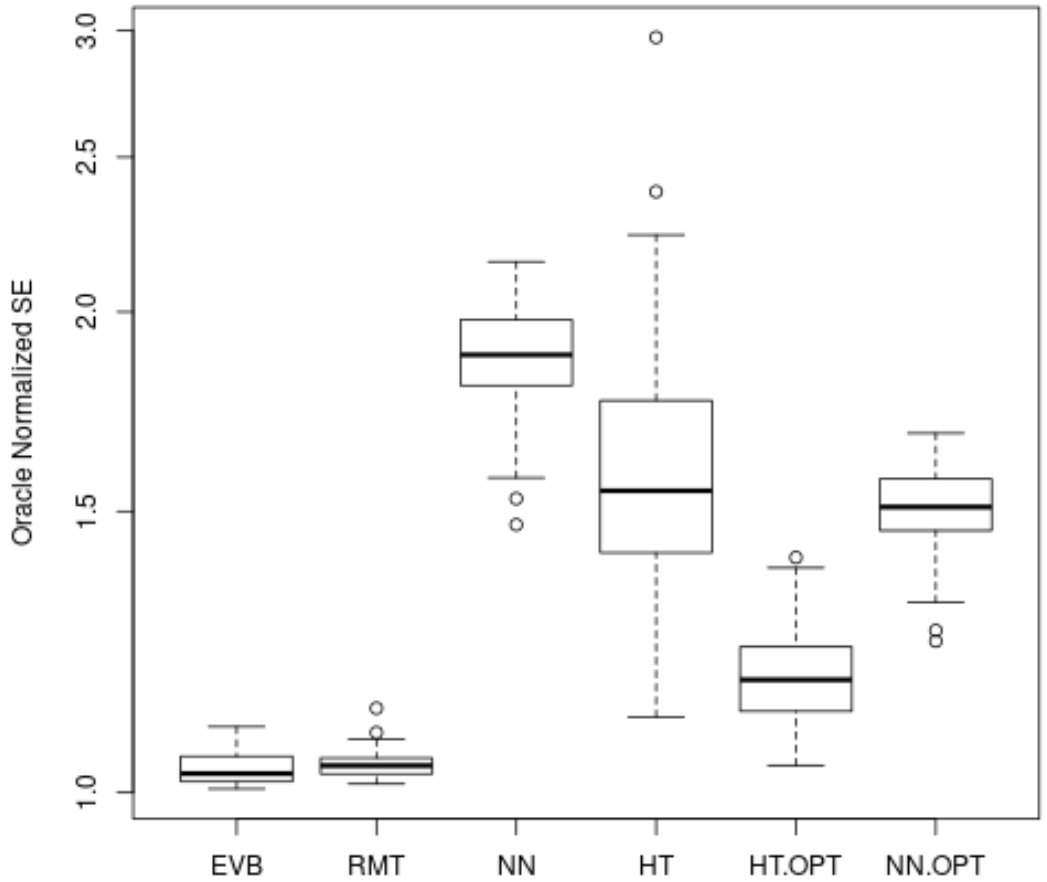
- Relative squared error (SE) and oracle normalized SE





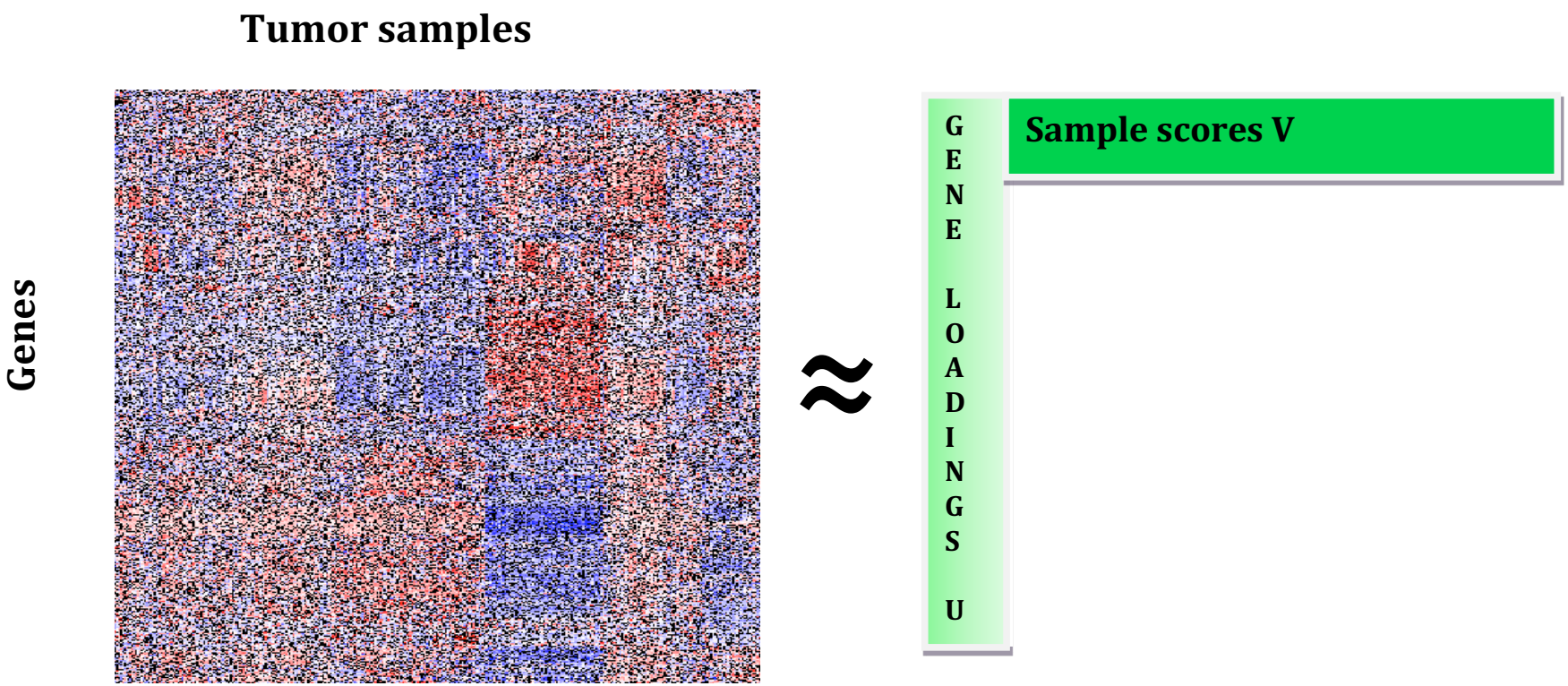
Low-rank matrix approximation

- Oracle normalized SE



Matrix factorization: missing data

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

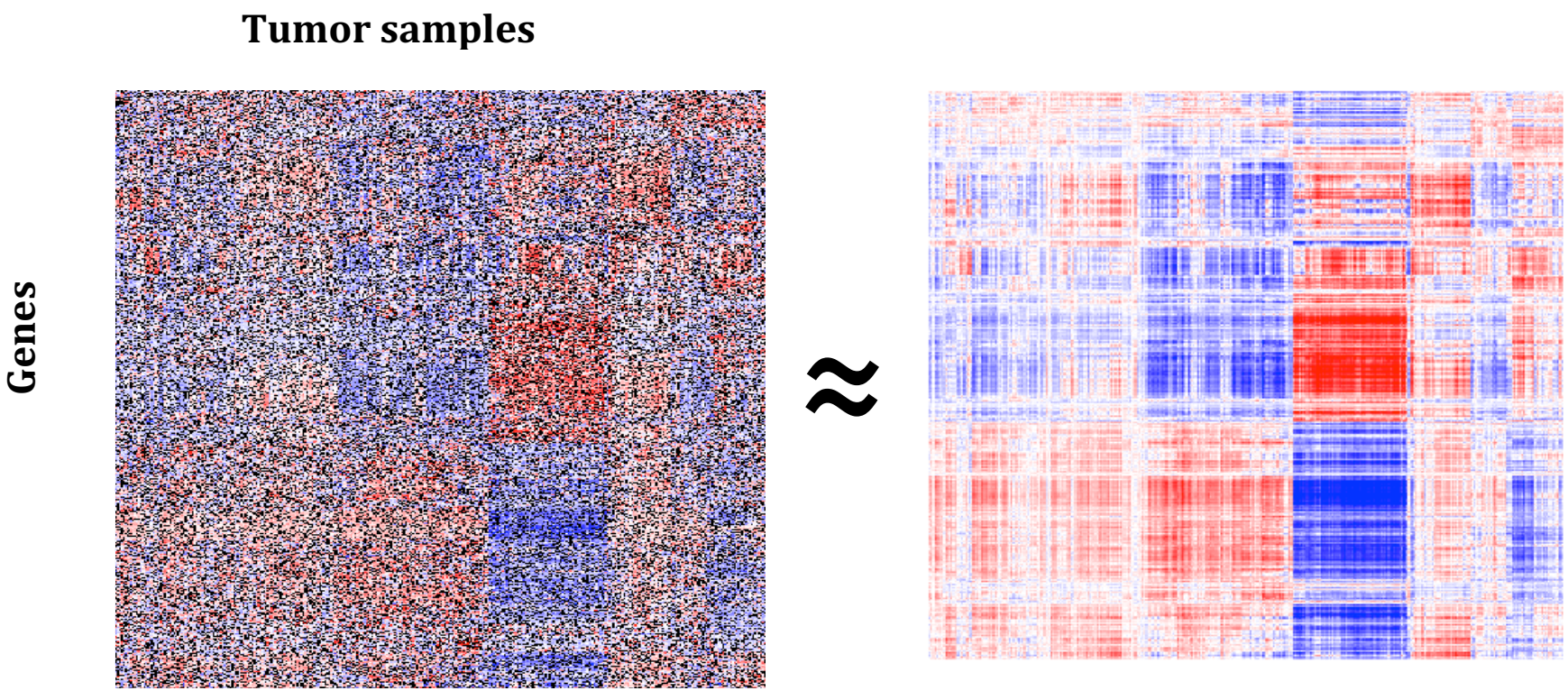


Matrix factorization: missing data

- Missing data $X_{\text{miss}} = \{X[i,j] : (i,j) \in \mathcal{M}\}$
- Minimize free energy over $\sigma, \tau_U, \tau_v, q_u, q_v$, and X_{miss}.
- EM-type approach:
  - Initialize X_{miss}
  - Update τ_U, τ_v, q_u, q_v given X_{miss}
  - Update $X_{\text{miss}} = E_q X_{\text{miss}}$
  - Repeat until convergence

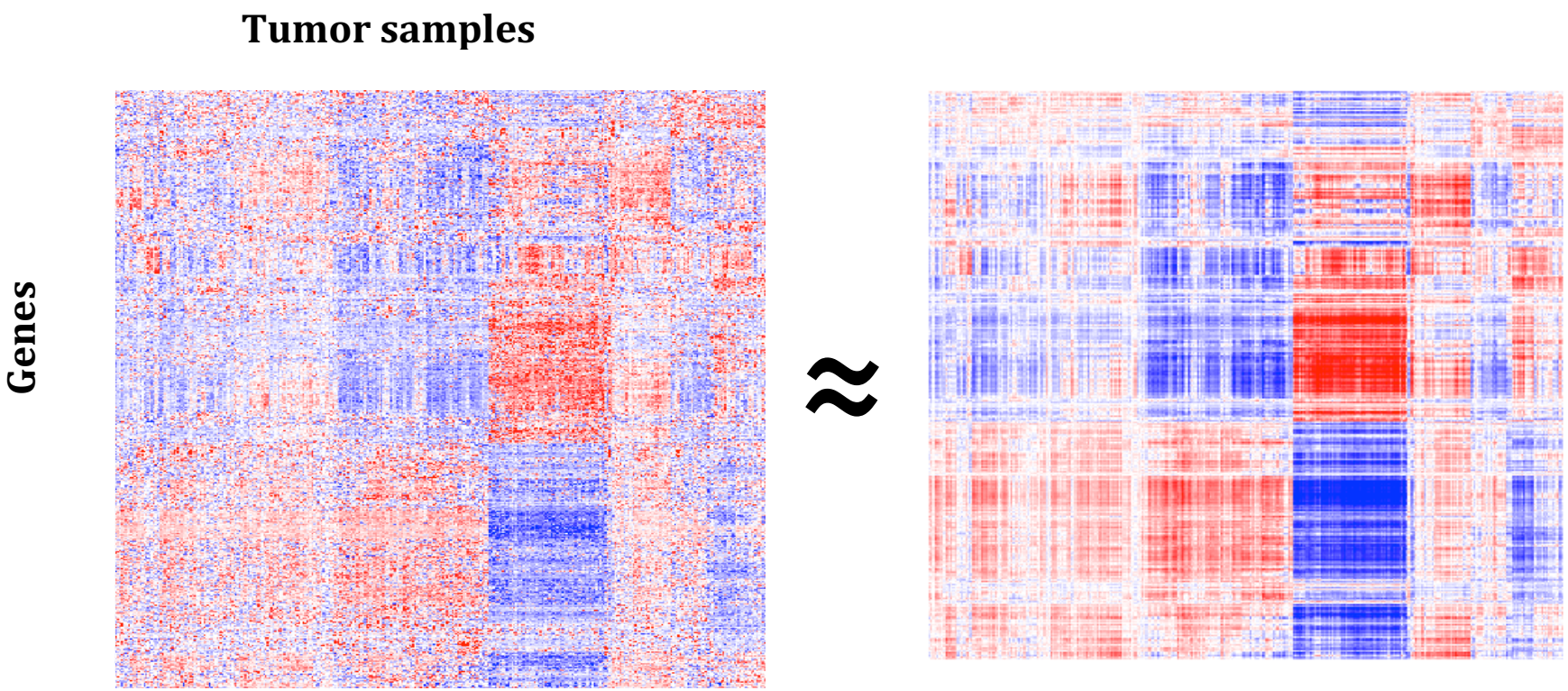
Matrix factorization: missing data

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

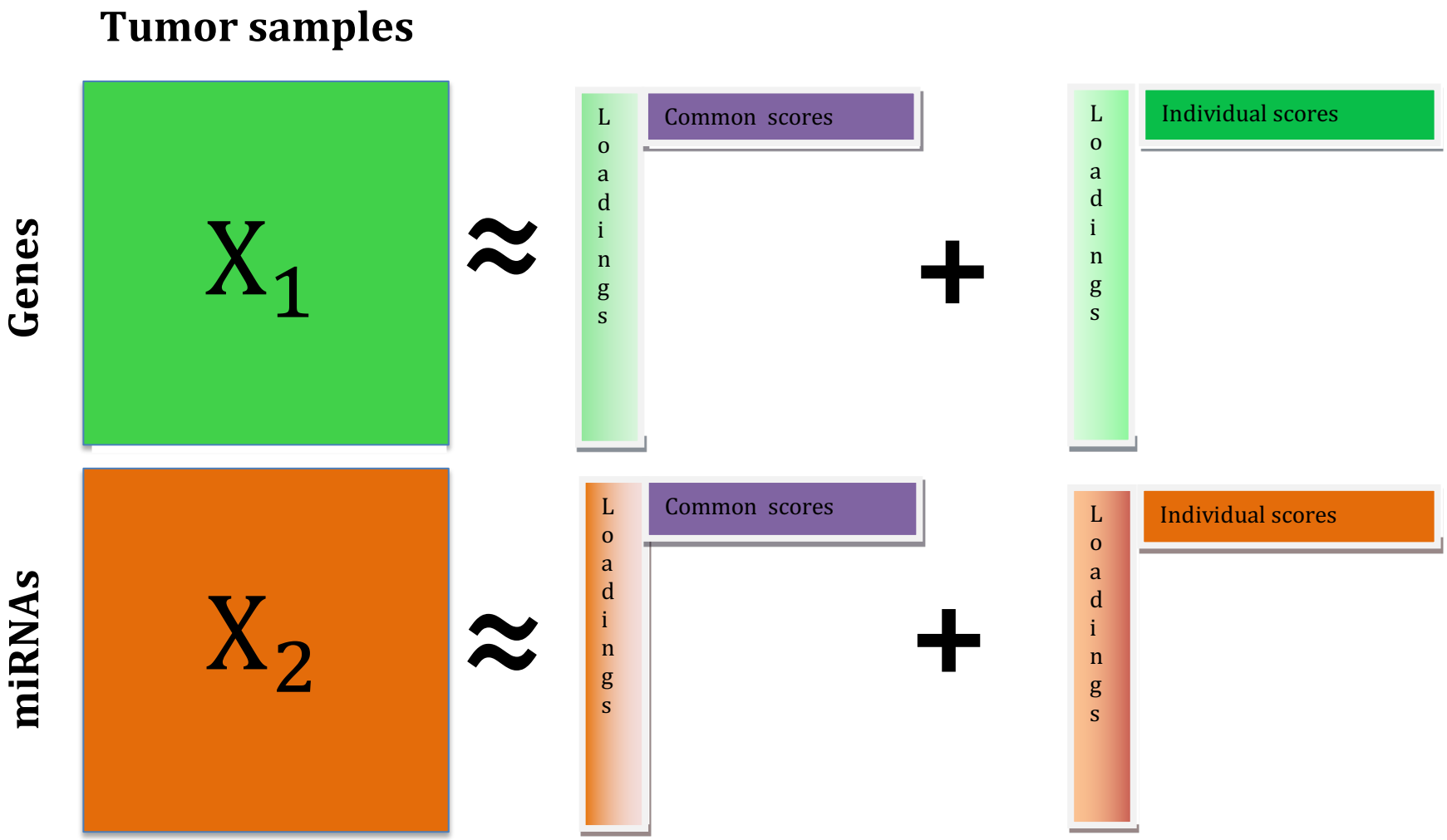


Matrix factorization: missing data

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



Vertically linked data: joint and individual factorization



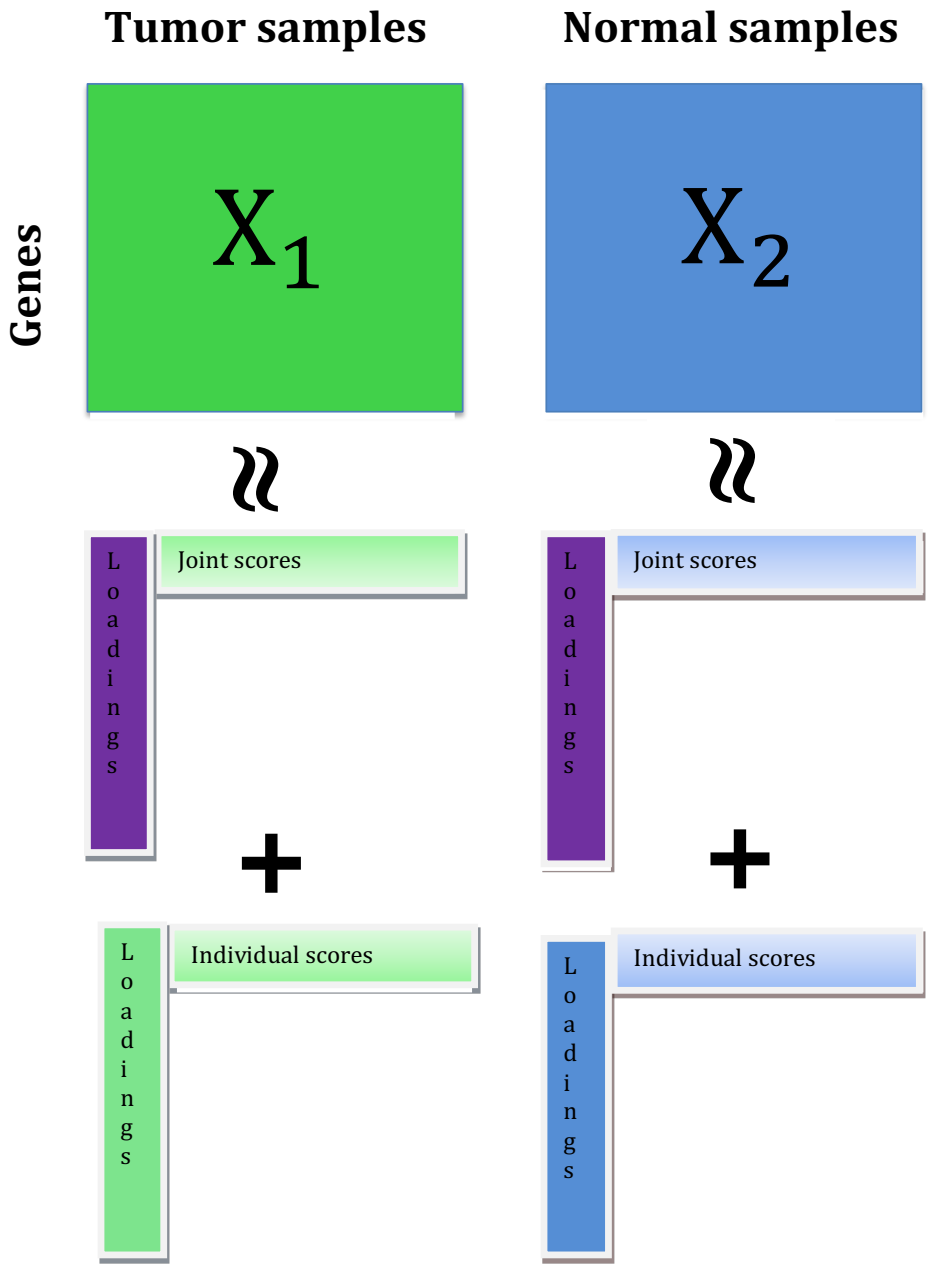
Joint + individual factorization methods

- 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]

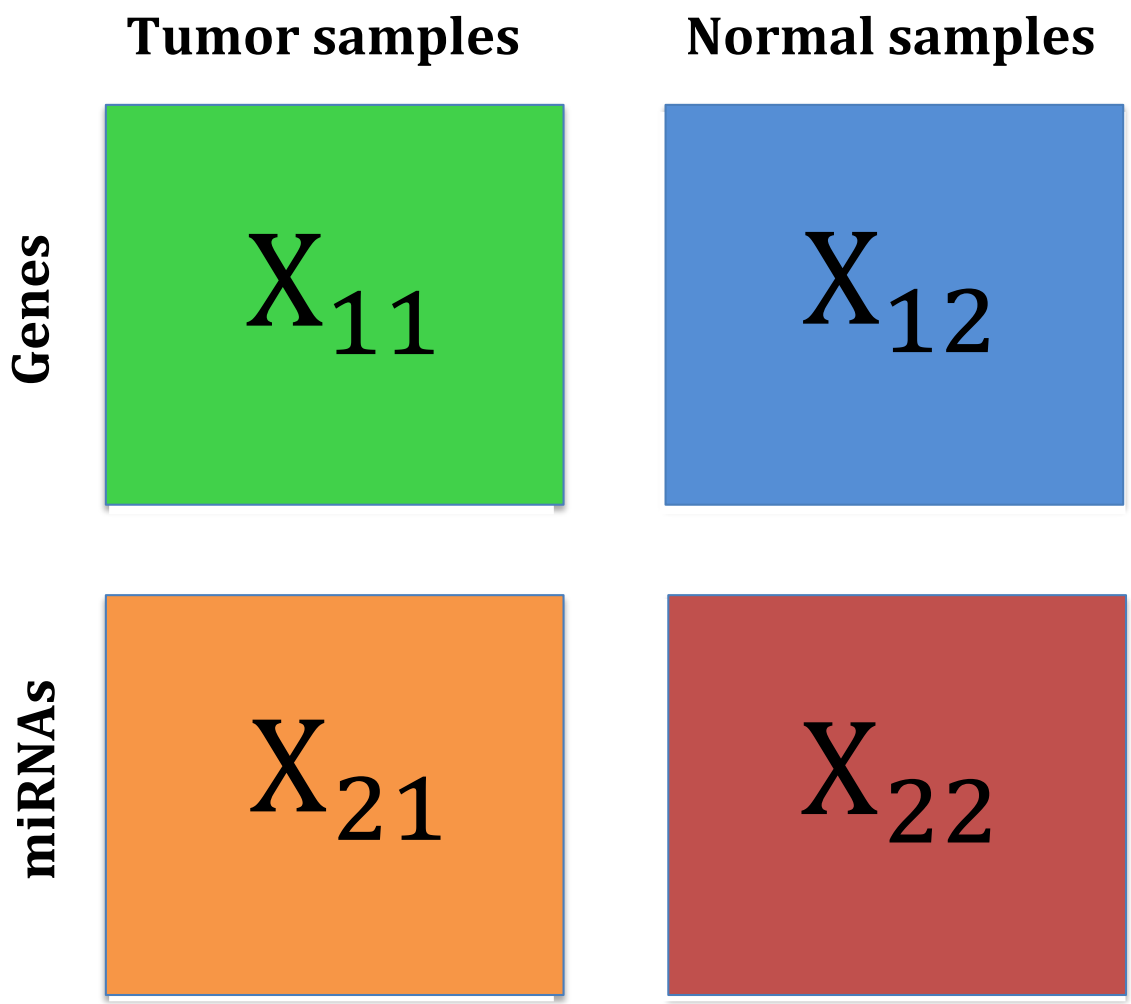
⋮

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

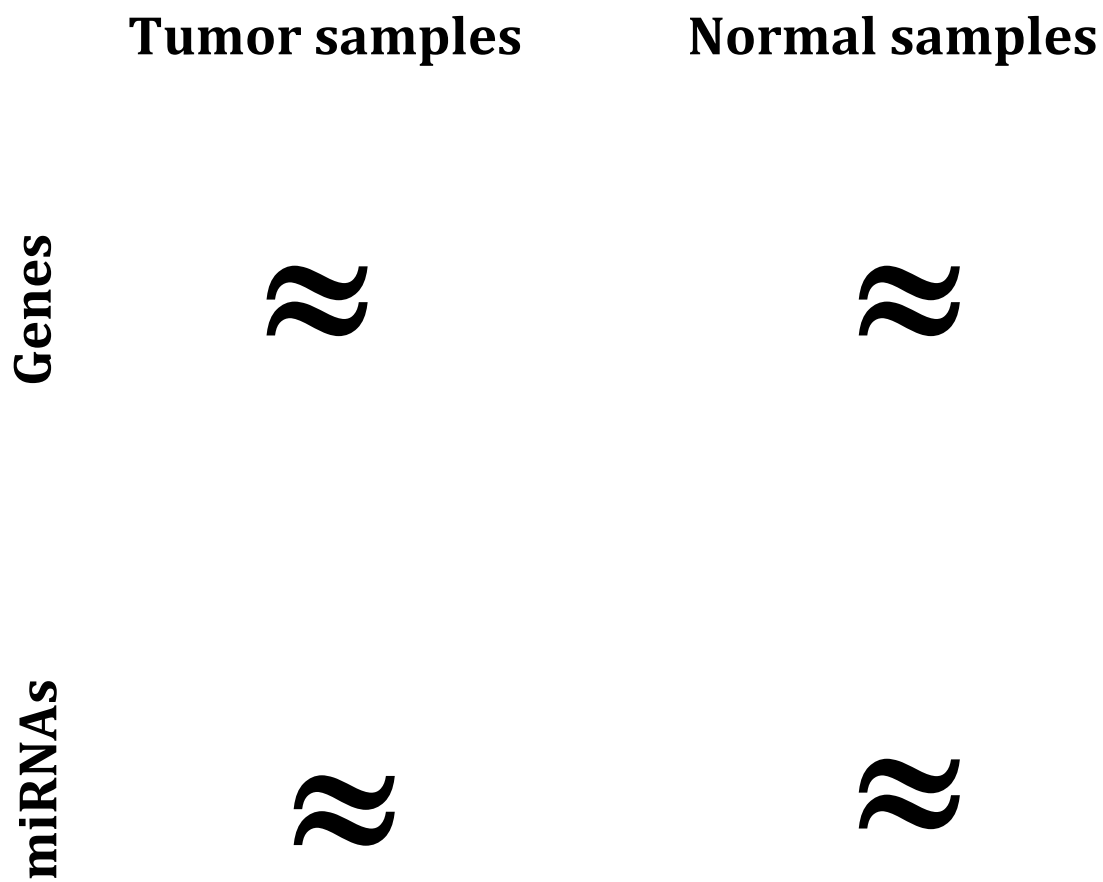
Horizontally linked data: Joint and individual



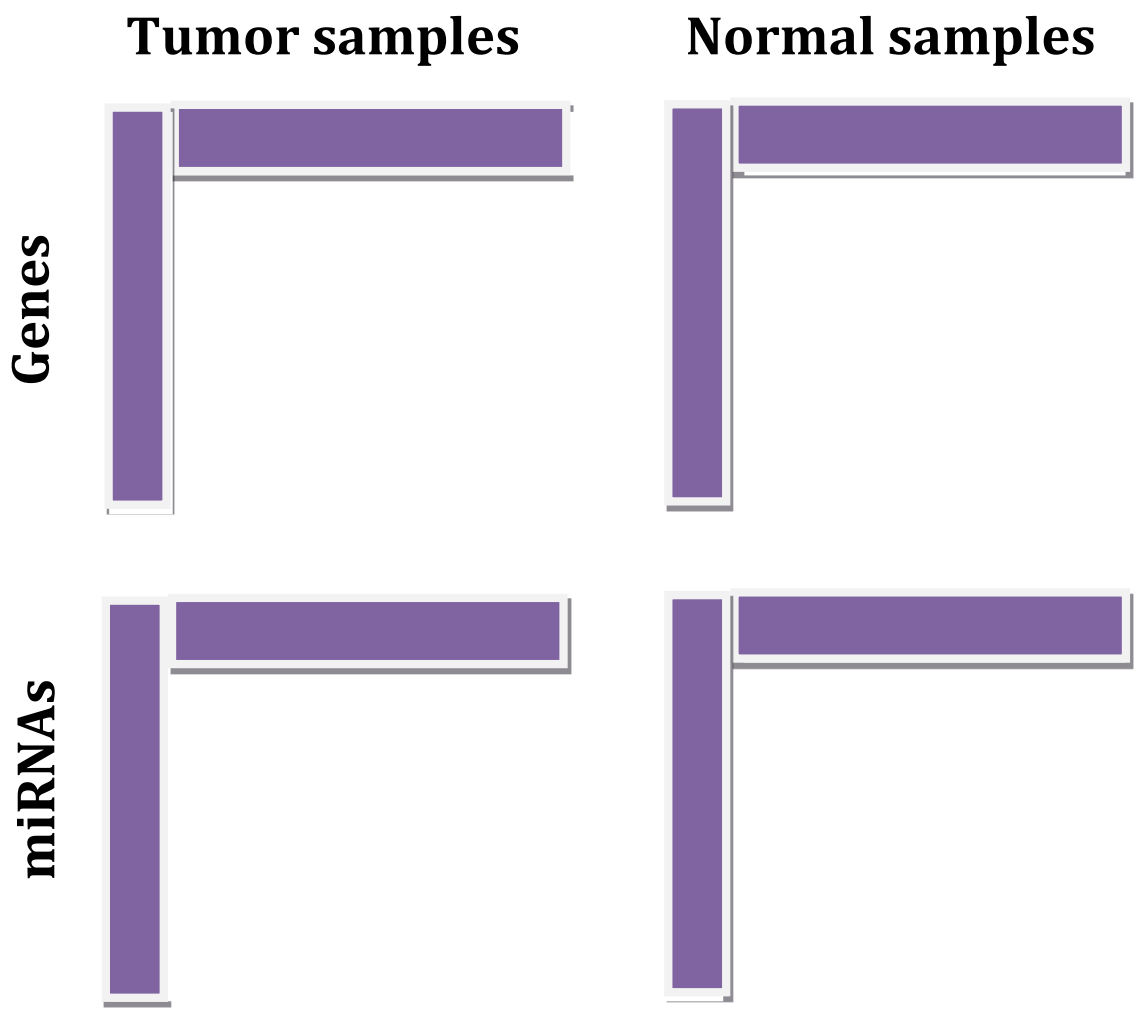
Bidimensionally linked data: BIDIFAC



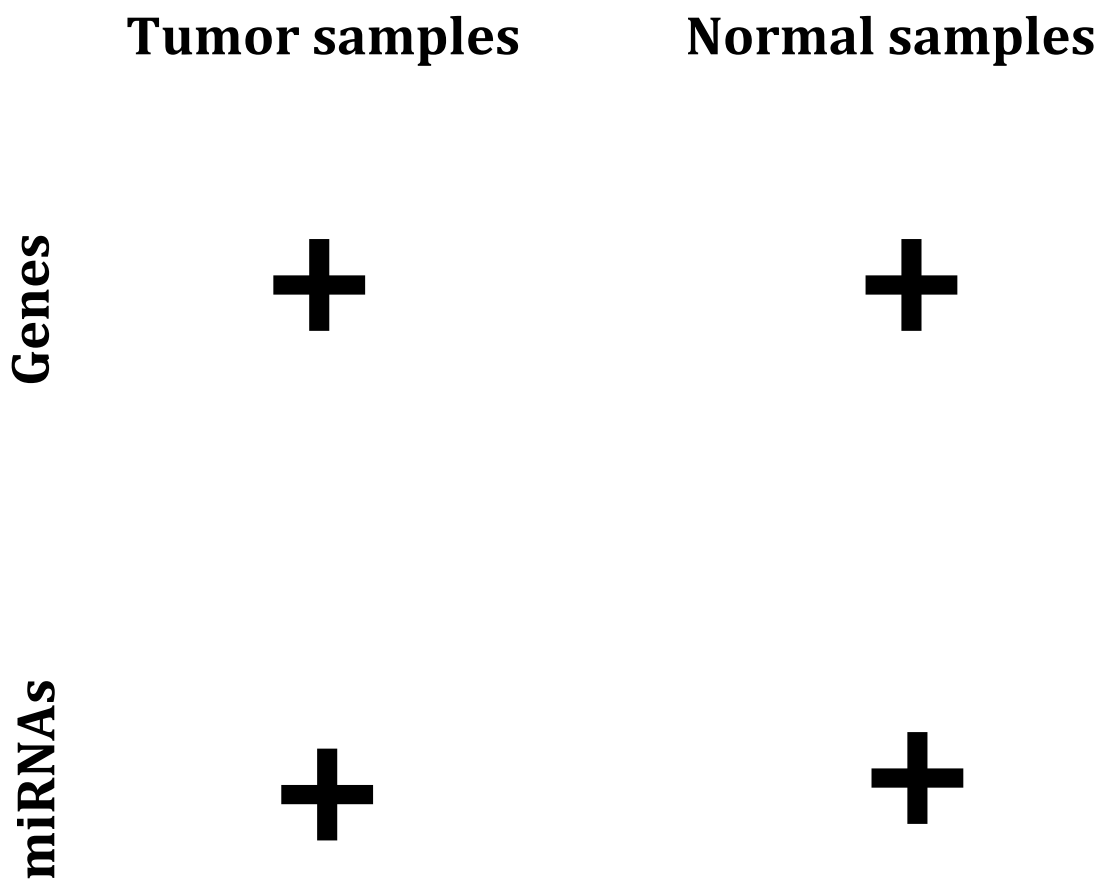
Bidimensionally linked data: BIDIFAC



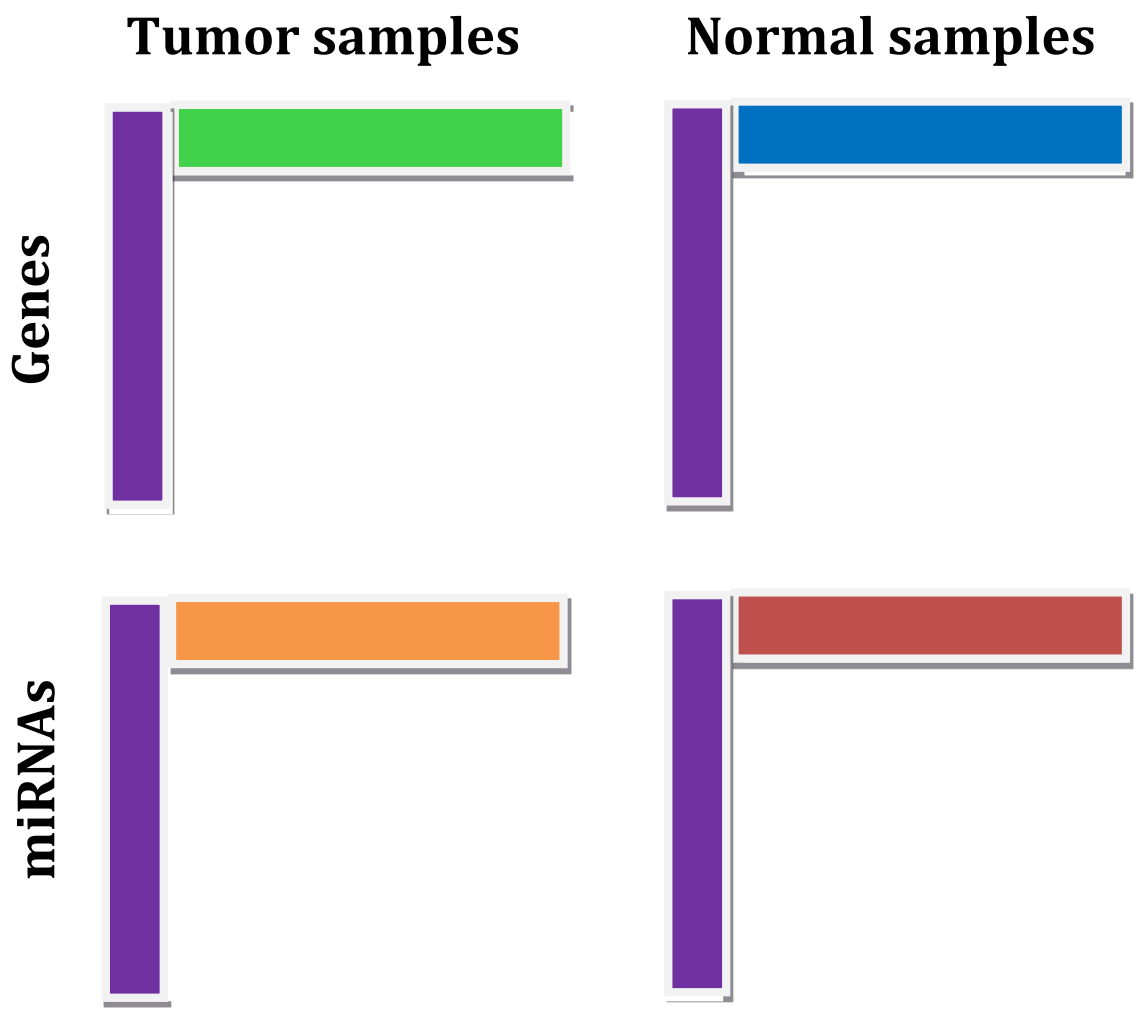
Bidimensionally linked data: BIDIFAC



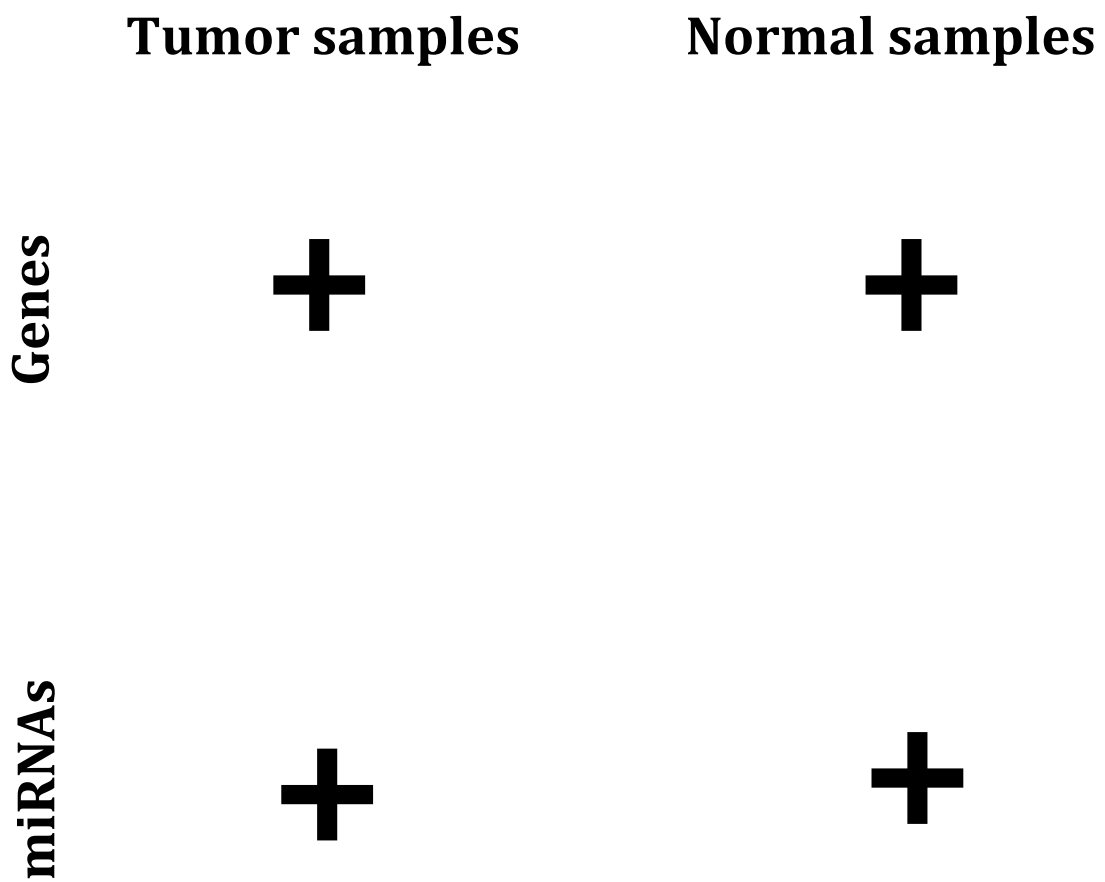
Bidimensionally linked data: BIDIFAC



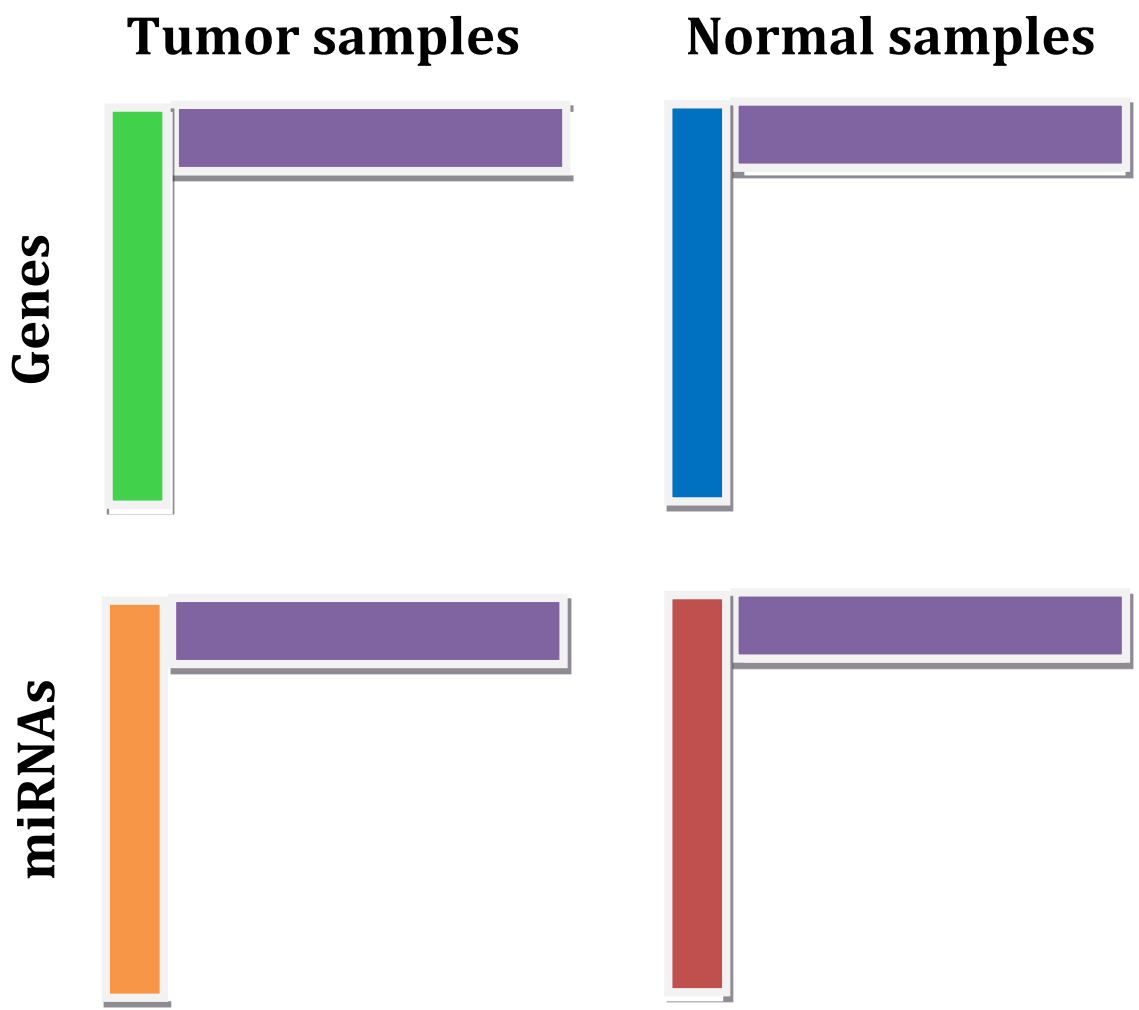
Bidimensionally linked data: BIDIFAC



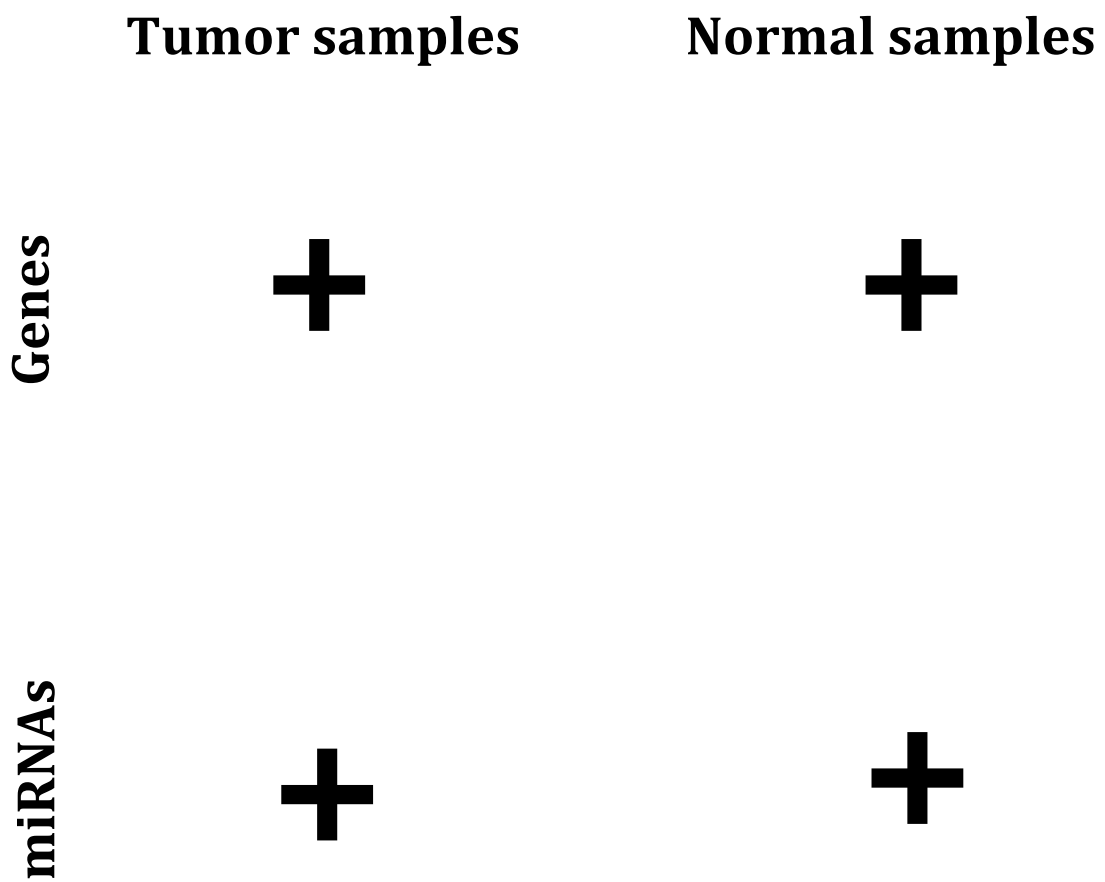
Bidimensionally linked data: BIDIFAC



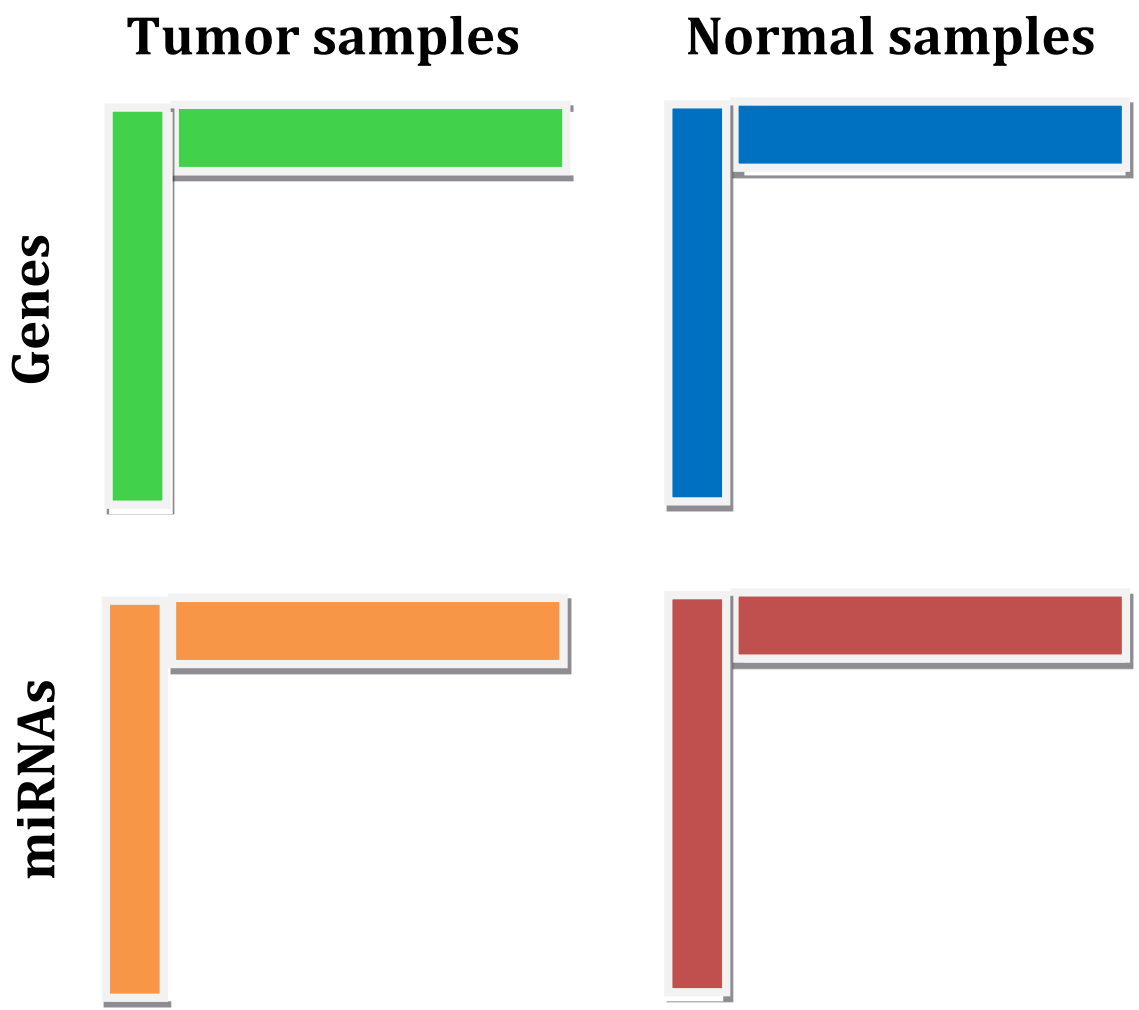
Bidimensionally linked data: BIDIFAC



Bidimensionally linked data: BIDIFAC



Bidimensionally linked data: BIDIFAC



BIDIFAC: general framework

- Linked matrices $\{X_{ij} : m_i \times n_j \mid i = 1, \ldots, I,\ j = 1, \ldots, J\}$:

$$X_{\cdot\cdot} = \begin{bmatrix} X_{11} & \ldots & X_{1J} \\ \vdots & \ddots & \vdots \\ X_{I1} & \ldots & X_{IJ} \end{bmatrix}$$

- Decompose $X_{\cdot\cdot}$ into structural *modules*:

$$X_{\cdot\cdot} = \sum_{k=1}^{K} S_{\cdot\cdot}^{(k)} + E_{\cdot\cdot},$$

where presence of submatrices $S_{ij}^{(k)}$ are determined by binary row indicators $R : I \times K$ and column indicators $C : J \times K$:

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

EV-BIDIFAC

- Model for $i = 1, \ldots I$ and $j = 1, \ldots, J$:

$$\mathbf{X}_{ij} = \sum_{k=1}^{K} \mathbf{U}_i^{(k)} \mathbf{V}_j^{(k)T} + \mathbf{E}_{ij} \text{ where } \mathbf{E}_{ij}[l, m] \overset{iid}{\sim} \text{Normal}(0, \sigma_{ij}^2),$$

$$\mathbf{U}_i^{(k)}[\cdot, r] = \mathbf{0} \text{ if } \mathbf{R}[i, k] = 0, \text{Normal}(\mathbf{0}, \sigma_{ij}\tau_u^2[k, r]\mathbf{I}) \text{ if } \mathbf{R}[i, k] = 1$$

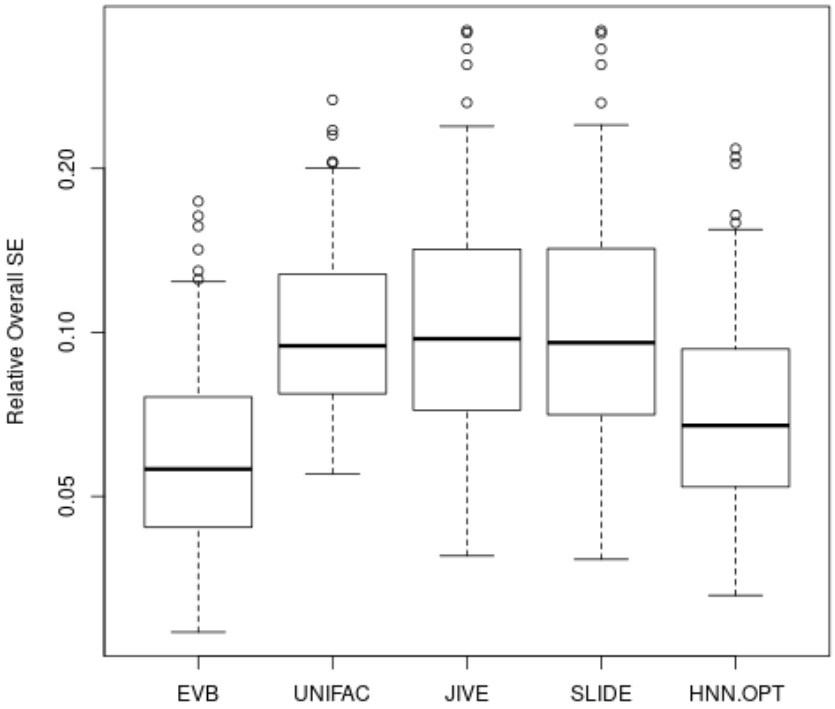
$$\mathbf{V}_j^{(k)}[\cdot, r] = \mathbf{0} \text{ if } \mathbf{C}[j, k] = 0, \text{Normal}(\mathbf{0}, \sigma_{ij}\tau_v^2[k, r]\mathbf{I}) \text{ if } \mathbf{C}[j, k] = 1$$

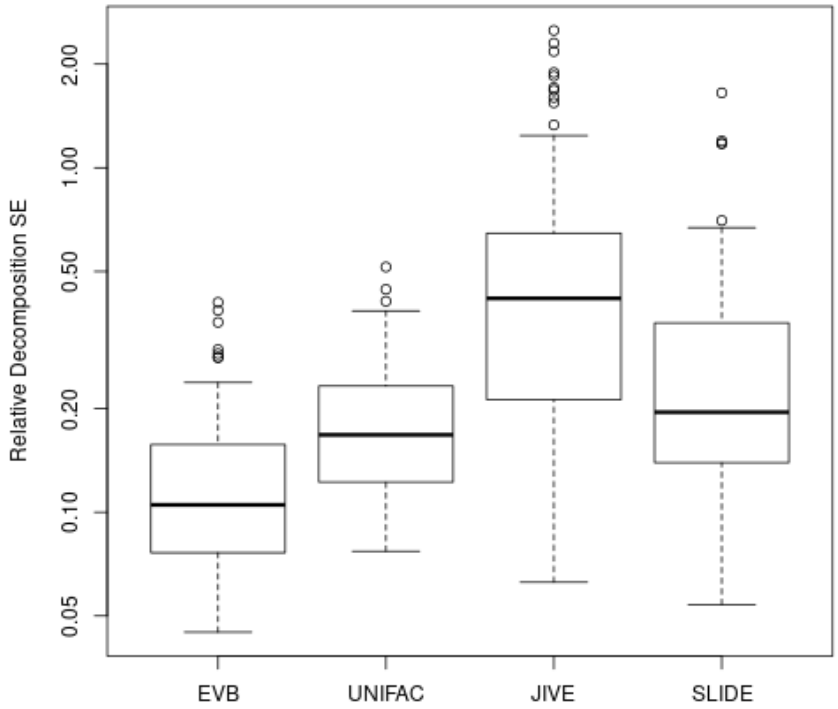
- Estimate under empirical variational Bayes framework

- Gives a unique decomposition under general conditions!

EV-BIDIFAC: Results

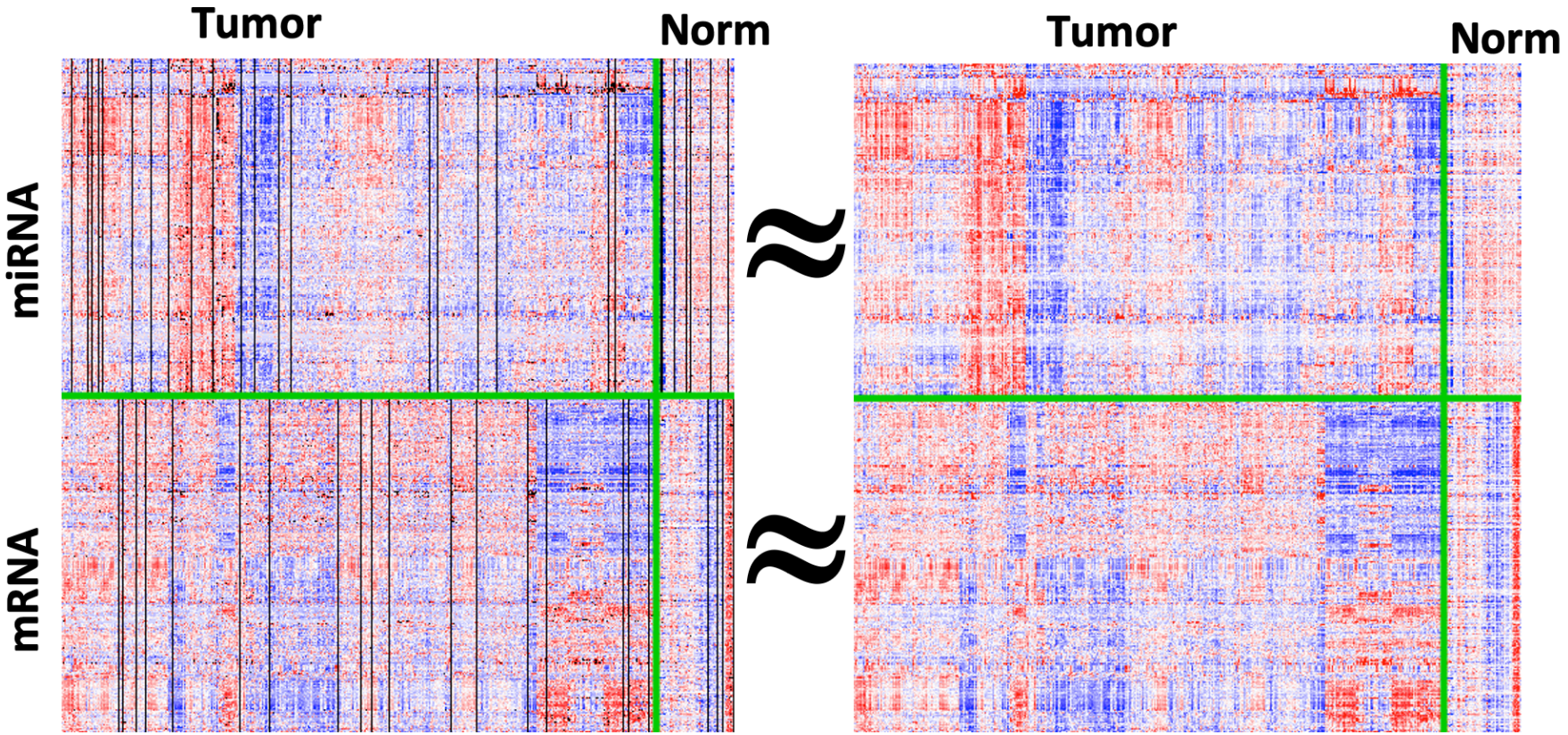
- Simulation with two linked matrices:





EV-BIDIFAC: Results

- Minimize free energy via EM for missing data



EV-BIDIFAC: results

- Imputation accuracy for BRCA (held-out data):

Method	Entry-missing	Col-missing	Row-missing	Overall
EV-BIDIFAC	**0.389**	**0.756**	**0.903**	**0.682**
BIDIFAC	0.526	0.870	**0.909**	0.768
EB-SEP	**0.381**	1.00	1.00	0.79
EB-JOINT	0.510	1.01	1.67	1.06
NN-SEP	0.536	1.00	1.00	0.845
NN-JOINT	0.614	0.891	**0.881**	0.796
HT-SEP	0.459	1.00	1.03	0.831
HT-JOINT	0.520	4.50	13.1	6.05
KNN	0.646	1.26	1.01	0.974
StructureMC	-	0.977	1.01	-

Thank you!

- Support: NIGMS grant R01-GM130622

- Reference:
  - **EV-BIDIFAC**: EF Lock. Empirical Bayes Linked Matrix Decomposition. *Machine Learning*, 113 (10): 7451-7477.

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