# SNeCT: Integrative cancer data analysis via large scale network constrained Tucker decomposition

Dongjin Choi and Lee Sael

XXXX

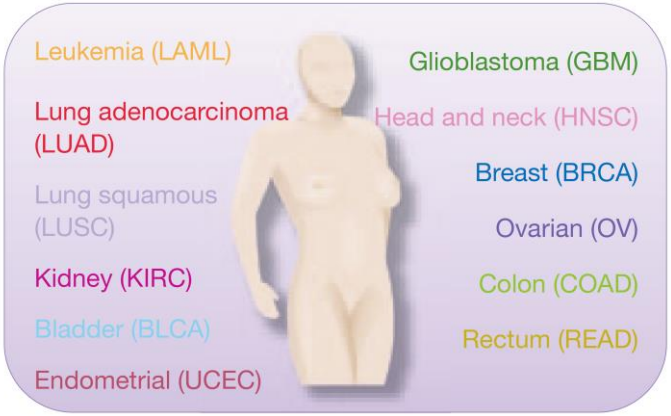
November xx, 2017

Lee Sael

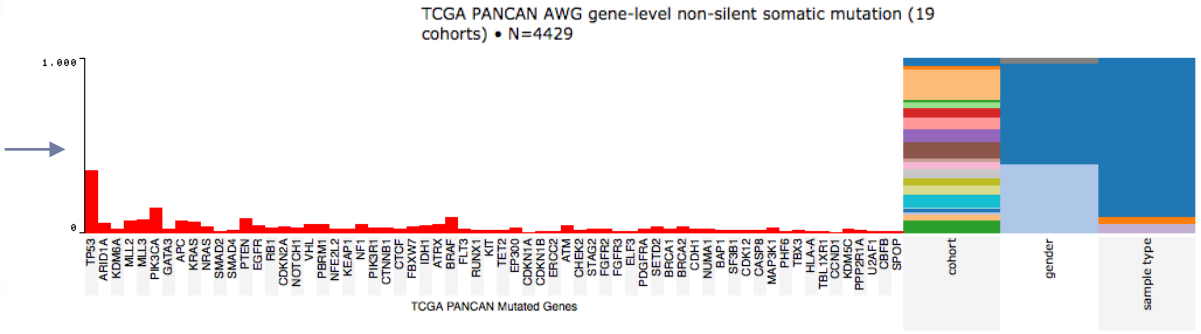
# Motivation

- Q: How can we characterize cancer patients?
  - A: The Cancer Genome Atlas (TCGA) Pan-Cancer data provide rich data across 12 tumor types

12 tumor types



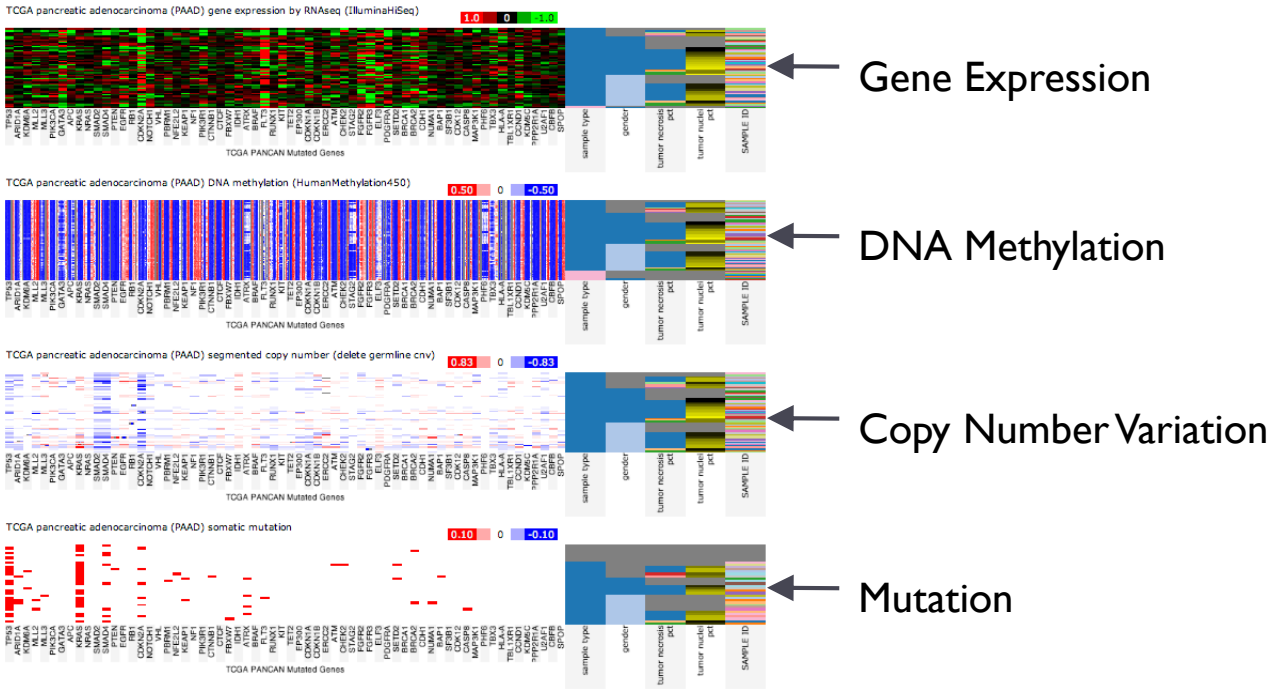

John N. Weinstein *et al. Nat Genet* 45(10), 1113-1120 (2013) doi:10.1038/ng.2764




Mary Goldman. *UCSC Cancer Browser Workshop* (2015)

# Motivation

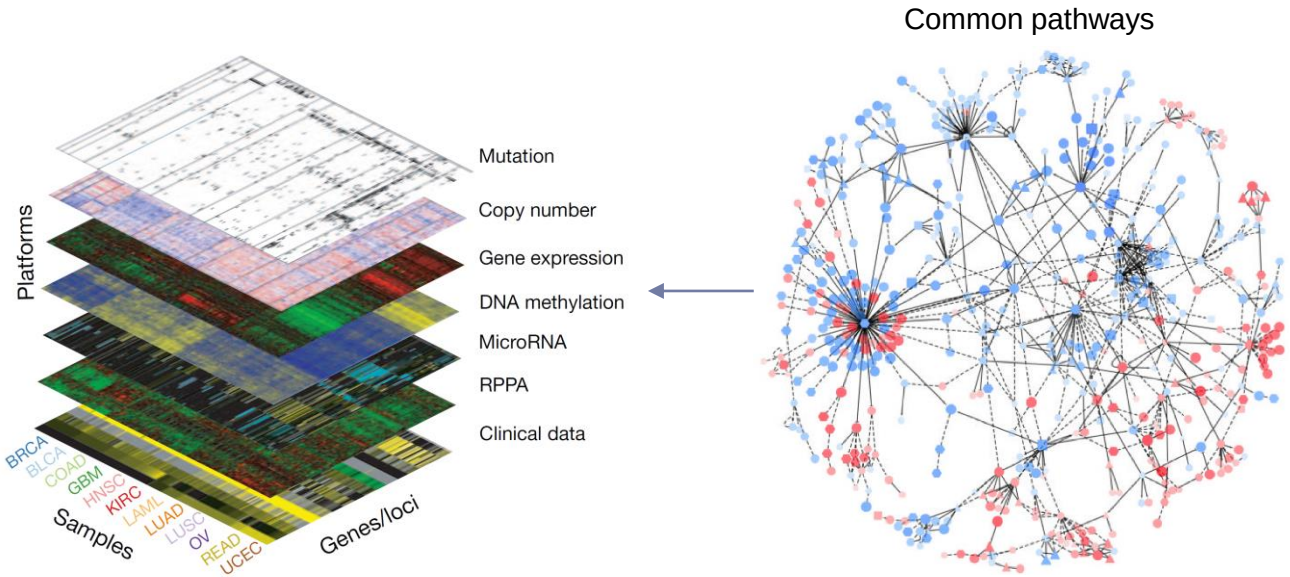
- How can we provide integrated analysis for multi-dimensional data?
  - Pan-Cancer12 data consist of multi-platform data




Mary Goldman. *UCSC Cancer Browser Workshop* (2015)

# Motivation

- How can we build a combined model exploiting gene networks?
  - Gene association networks provide gene similarity information



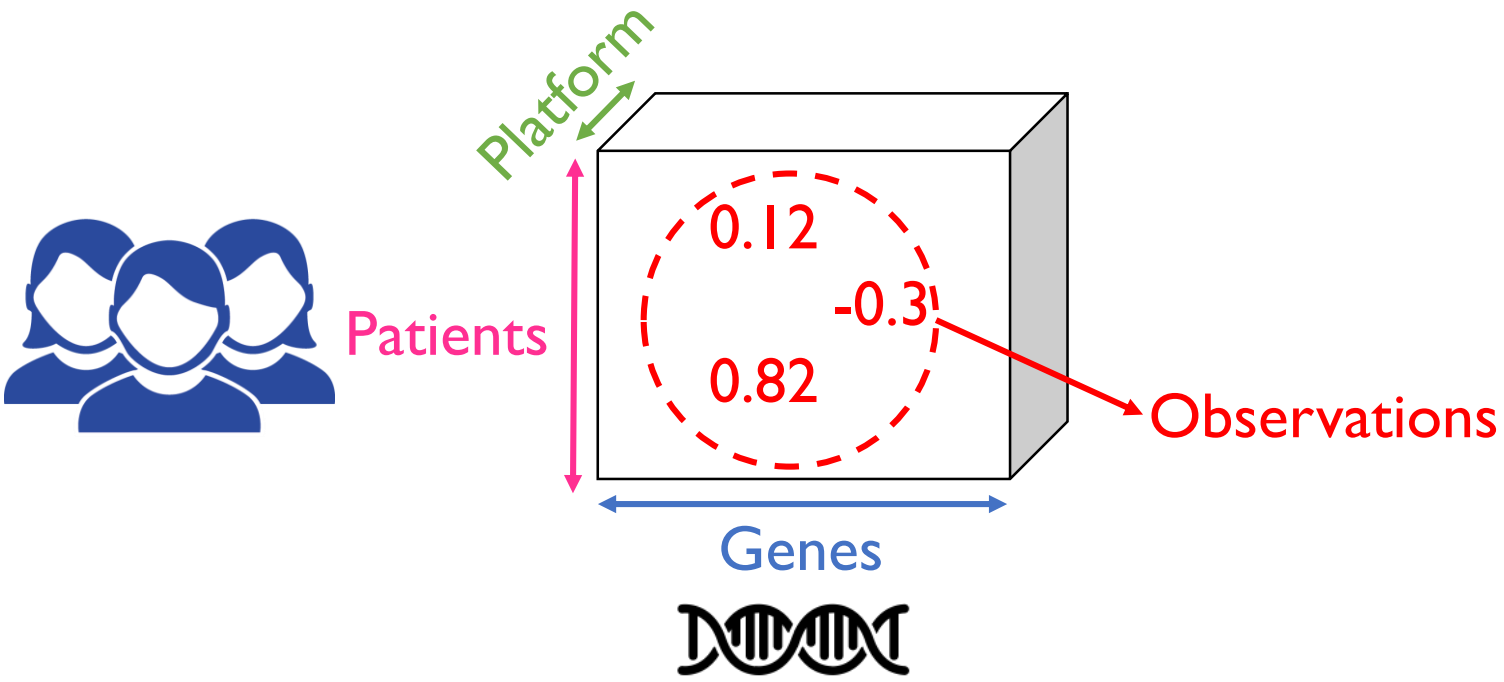

John N. Weinstein *et al. Nat Genet* 45(10), 1113-1120 (2013) doi:10.1038/ng.2764

# Overview

- Introduction
- Problem definition
- Proposed method
- Experiments
- Conclusion

# Tensor

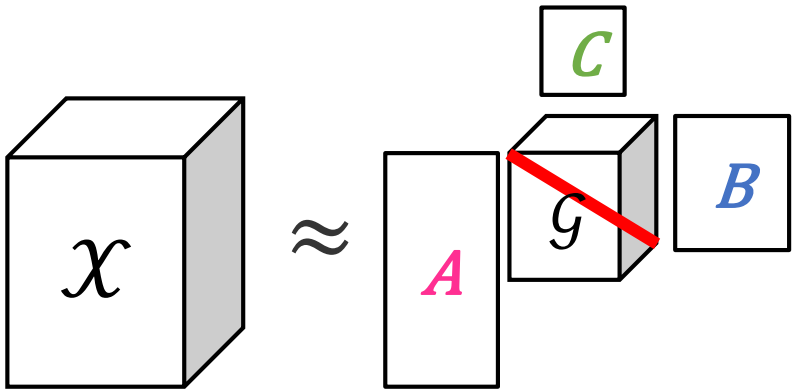
- A tensor is a multi-dimensional array
- Pan-can12 data are represented as a 3-D tensor



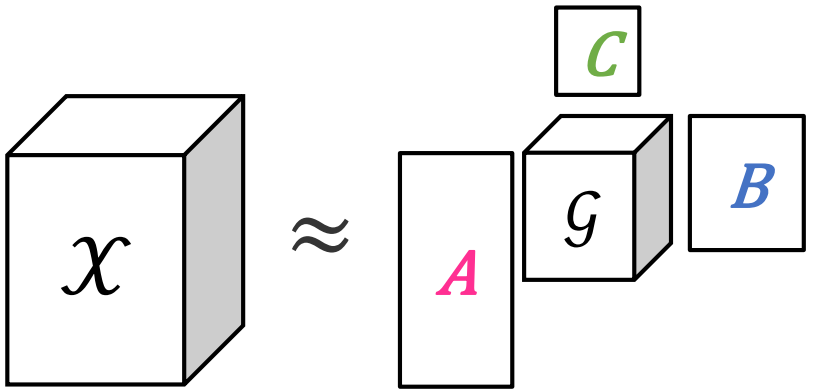
# Tensor Factorization

- Given a tensor, decompose the tensor into a core tensor and factor matrices whose product approximates the original tensor

**CP Decomposition**




**Tucker Decomposition (HOSVD)**

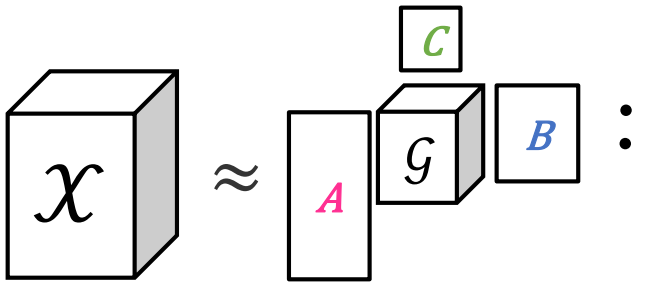


# Overview

- Introduction
- Problem definition
- Proposed method
- Experiments
- Conclusion

# Tucker Decomposition

- Tucker decomposition (Tucker, 1966)
  - Widely-used tensor factorization method
  - Given a tensor, Tucker decomposition factorizes the tensor into product of a core tensor and orthogonal factor matrices




:

$$\mathcal{X} \approx \widetilde{\mathcal{X}} = \mathcal{G} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C}$$

s.t. $\boldsymbol{A}^T\boldsymbol{A}=\boldsymbol{B}^T\boldsymbol{B}=\boldsymbol{C}^T\boldsymbol{C}=\boldsymbol{I}$

Elementwise,

$$x_{ijk} \approx \mathcal{G} \times_1 \boldsymbol{a}_i \times_2 \boldsymbol{b}_j \times_3 \boldsymbol{c}_k$$

$\boldsymbol{a}_i$: $i$-th row of $\boldsymbol{A}$
$\boldsymbol{b}_j$: $j$-th row of $\boldsymbol{B}$
$\boldsymbol{c}_k$: $k$-th row of $\boldsymbol{C}$

# Tucker Decomposition (cont.)

- Formal problem definition
  - Given a 3-D tensor $\mathcal{X}$ $(\in \mathbb{R}^{I \times J \times K})$ with observable entries $\{x_{ijk} | (i,j,k) \in \Omega_{\mathcal{X}}\}$, the rank-$[P,Q,R]$ factorization of $\mathcal{X}$ is to find the core tensor $\mathcal{G}$ and factor matrices $\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$ which minimizes the following loss function:

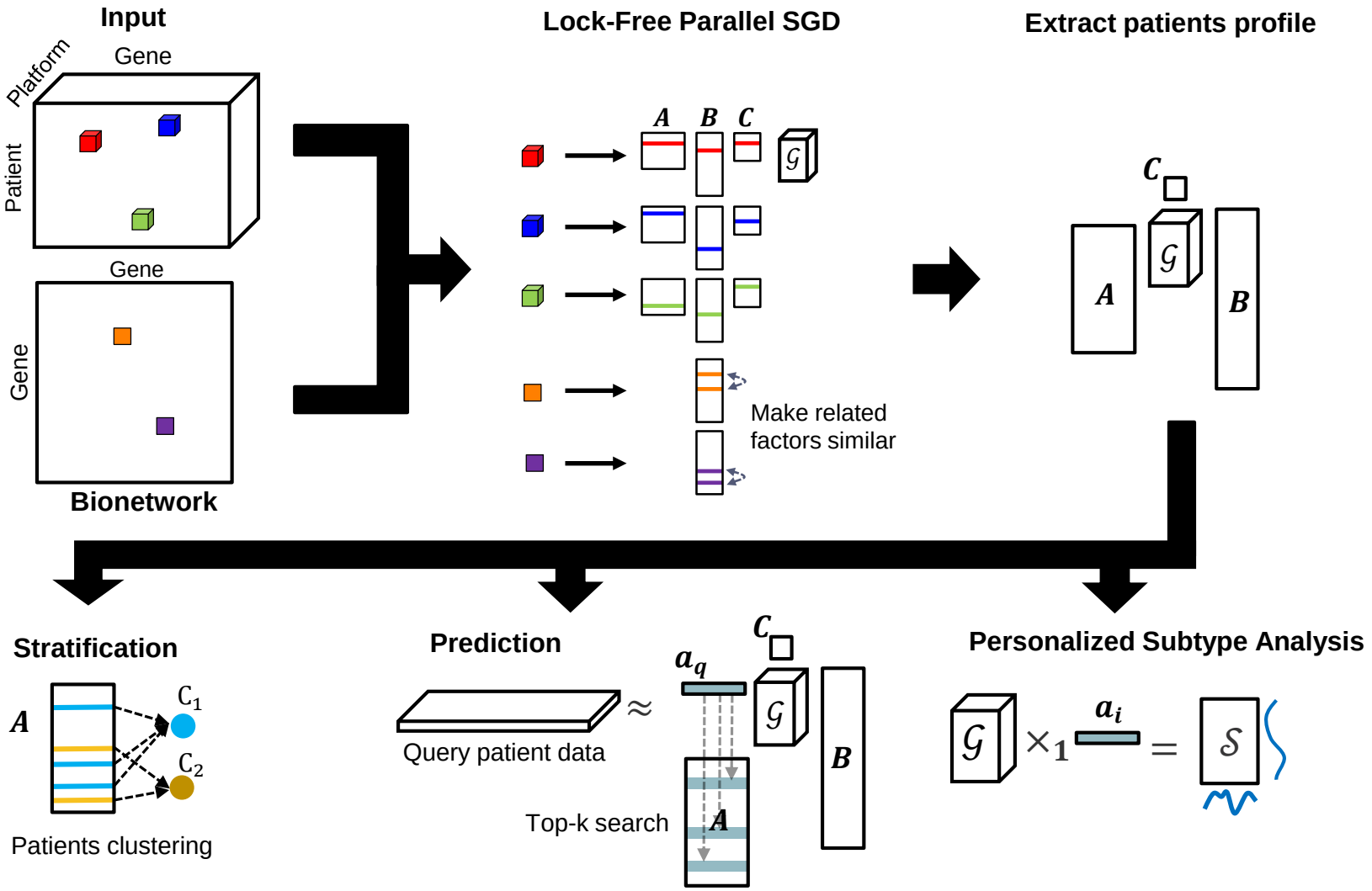
$$f(\mathcal{G}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \frac{1}{2} \left\| \mathcal{X} - \widetilde{\mathcal{X}} \right\|_F^2 + \frac{\lambda}{2} R(\mathcal{G}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$$

$$= \frac{1}{2} \sum_{(i,j,k) \in \Omega_{\mathcal{X}}} \left( x_{ijk} - \mathcal{G} \times_1 \boldsymbol{a}_i \times_2 \boldsymbol{b}_j \times_3 \boldsymbol{c}_k \right)^2 + \frac{\lambda}{2} R(\mathcal{G}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$$

# Overview

- Introduction
- Problem definition
- Proposed method
- Experiments
- Conclusion

# Scheme of SNeCT



# Proposed methods

- **SNeCT** enables integrative tensor factorization and analysis for tensor data with network constraint
  SNeCT = Scalable Network Constrained Tucker decomposition
- Method 1
  - Formulate SGD-amenable objective function
  - Iterative SGD update with lock-free parallel scheme
- Method 2
  - Personalized subtype analysis

# Proposed methods

- Formulate SGD-amenable objective function
  - Given the gene similarity matrix $\boldsymbol{Y}$ $(\in \mathbb{R}^{J\times J})$ with observable entries $\{y_{mn}|(m,n) \in \Omega_{\boldsymbol{Y}}\}$, network constraint is formulated to make similar genes have similar factors:

$$f_g(\boldsymbol{B},\boldsymbol{Y}) = \frac{1}{2}\sum_{l=1}^{Q}\left[\sum_{(m,n)\in\Omega_{\boldsymbol{Y}}} y_{mn}(b_{ml}-b_{nl})^2\right]$$
$$= \frac{1}{2}\sum_{(m,n)\in\Omega_{\boldsymbol{Y}}} y_{mn}\|\boldsymbol{b}_m-\boldsymbol{b}_n\|_F^2$$

# Proposed methods

- Formulate SGD-amenable objective function

$$f(\mathcal{G},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}) = \frac{1}{2}\sum_{(i,j,k)\in\Omega_{\mathcal{X}}}\left(x_{ijk}-\tilde{x}_{ijk}\right)^2 + \frac{\lambda}{2}R(\mathcal{G},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})$$

$$= \frac{1}{2}\sum_{(i,j,k)\in\Omega_{\mathcal{X}}}\left[\left(x_{ijk}-\tilde{x}_{ijk}\right)^2 + \frac{\lambda}{|\Omega_{\mathcal{X}}|}\|\mathcal{G}\|_F^2 + \lambda\left(\frac{\|\boldsymbol{a}_i\|_F^2}{|\Omega_{\mathcal{X}}^i|} + \frac{\|\boldsymbol{b}_j\|_F^2}{|\Omega_{\mathcal{X}}^j|} + \frac{\|\boldsymbol{c}_k\|_F^2}{|\Omega_{\mathcal{X}}^k|}\right)\right]$$

$$f_g(\boldsymbol{B},\boldsymbol{Y}) = \frac{1}{2}\sum_{(m,n)\in\Omega_{\boldsymbol{Y}}} y_{mn}\|\boldsymbol{b}_m-\boldsymbol{b}_n\|_F^2$$

- Integrate into single objective function

$$f_{opt} = f + \lambda_g f_g$$

# Proposed methods

- Calculate gradients of $f_{opt}$ with respect to the core tensor and factor matrices for a given data point $x_{\alpha=(ijk)}$ or $y_{\beta=(mn)}$

$$\left.\frac{\partial f_{opt}}{\partial \boldsymbol{a}_i}\right|_{\alpha} = -(x_\alpha - \tilde{x}_\alpha)\left[\mathcal{G} \times_2 \boldsymbol{b}_j \times_3 \boldsymbol{c}_k\right] + \frac{\lambda}{|\Omega_X^i|}\boldsymbol{a}_i$$

$$\left.\frac{\partial f_{opt}}{\partial \mathcal{G}}\right|_{\alpha} = -(x_\alpha - \tilde{x}_\alpha) \times_1 \boldsymbol{a}_i^T \times_2 \boldsymbol{b}_j^T \times_3 \boldsymbol{c}_k^T + \frac{\lambda}{|\Omega_X|}\mathcal{G}$$

$$\left.\frac{\partial f_{opt}}{\partial \boldsymbol{b}_m}\right|_{\beta} = \lambda_g y_\beta (\boldsymbol{b}_m - \boldsymbol{b}_n)$$

- $\left.\frac{\partial f_{opt}}{\partial \boldsymbol{b}_j}\right|_{\alpha}$, $\left.\frac{\partial f_{opt}}{\partial \boldsymbol{c}_k}\right|_{\alpha}$, and $\left.\frac{\partial f_{opt}}{\partial \boldsymbol{b}_n}\right|_{\beta}$ are calculated symmetrically

# Proposed methods

- Parallel update with calculated gradient
- SNeCT($\mathcal{X}, \boldsymbol{Y}, \lambda, \lambda_g, \eta$) ($\eta$: learning rate)

1. Initialize $\mathcal{G}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ randomly
2. **repeat**
3. **for** $\forall x_{(ijk)=\alpha} \in \mathcal{X}, \forall y_{(mn)=\beta} \in \boldsymbol{Y}$ in random order **in parallel**
4. **if** $x_{ijk} \in \mathcal{X}$ is picked **then**
5. $\boldsymbol{a}_i \leftarrow \boldsymbol{a}_i - \eta \frac{\partial f_{opt}}{\partial \boldsymbol{a}_i}\Big|_\alpha$, $\boldsymbol{b}_j \leftarrow \boldsymbol{b}_j - \eta \frac{\partial f_{opt}}{\partial \boldsymbol{b}_j}\Big|_\alpha$, $\boldsymbol{c}_k \leftarrow \boldsymbol{c}_k - \eta \frac{\partial f_{opt}}{\partial \boldsymbol{c}_k}\Big|_\alpha$
6. $\mathcal{G} \leftarrow \mathcal{G} - \eta \frac{\partial f_{opt}}{\partial \mathcal{G}}\Big|_\alpha$
7. **else if** $\forall y_{mn} \in \boldsymbol{Y}$ is picked **then**
8. $\boldsymbol{b}_m \leftarrow \boldsymbol{b}_m - \eta \frac{\partial f_{opt}}{\partial \boldsymbol{b}_m}\Big|_\beta$, $\boldsymbol{b}_n \leftarrow \boldsymbol{b}_n - \eta \frac{\partial f_{opt}}{\partial \boldsymbol{b}_n}\Big|_\beta$
9. **end if**
10. **end for**
11. **until** convergence condition satisfied
12. Orthogonalize $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ by QR decomposition
13. **return** $\mathcal{G}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$

# Overview

- Introduction
- Problem definition
- Proposed method
- Experiments
- Conclusion

# Experimental Settings

- Factorize data tensor with rank-[78,48,5]
- **Stratification**
  - Cluster analysis
  - Survival analysis
- **Prediction**
  - Top-k similarity search on clinical features
- **Personalized subtype analysis**
- **Performance**
  - Compare speed and convergence rate with competitor
  - Competitor: Narita *et al.* 2012

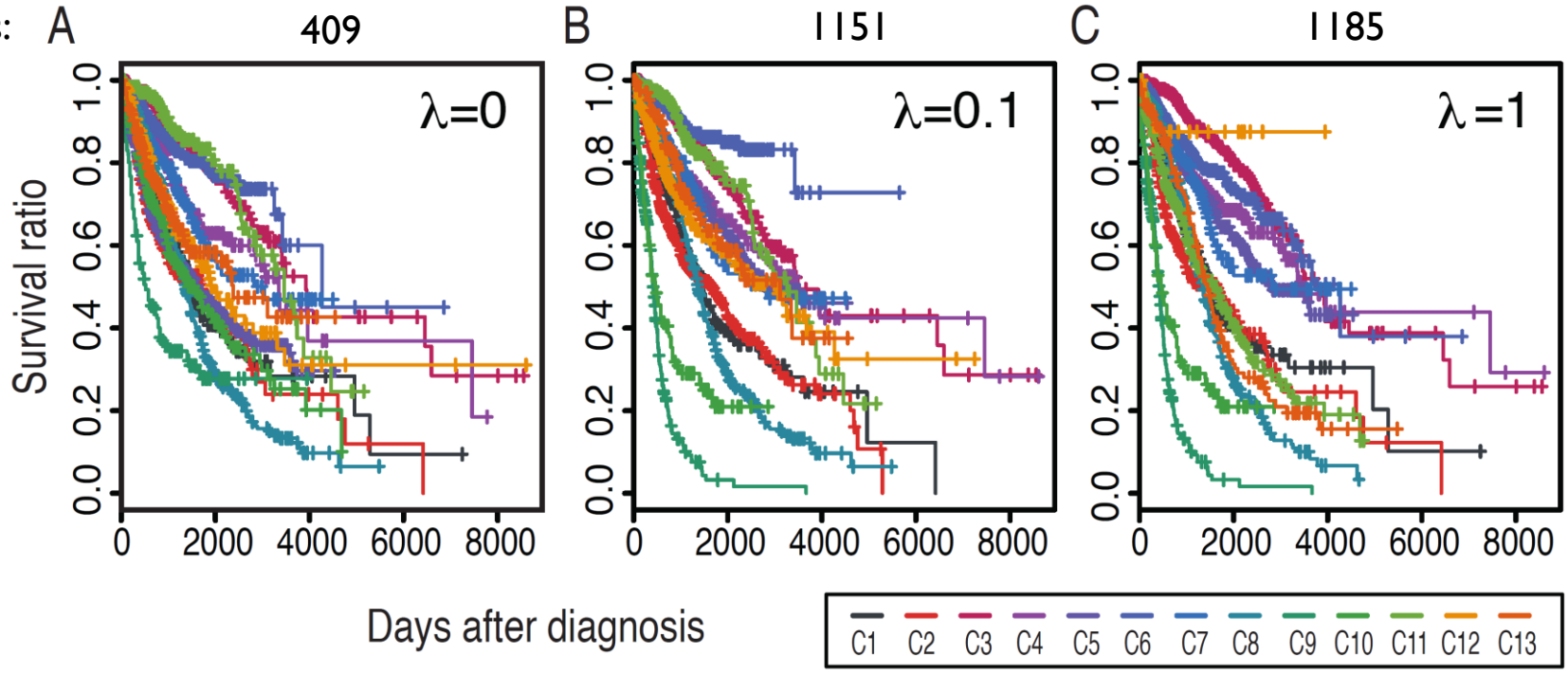
# Stratification – Cluster Analysis

| | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 | C13 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLCA | 16 | 32 | 2 | 19 | 0 | 22 | 3 | 0 | 0 | 0 | 32 | 0 | 0 | 126 |
| BRCA | 17 | 3 | 600 | 172 | 1 | 70 | 0 | 0 | 0 | 0 | 26 | 0 | 0 | 889 |
| COAD | 4 | 0 | 2 | 2 | 0 | 91 | 317 | 0 | 0 | 0 | 1 | 2 | 0 | 419 |
| GBM | 4 | 1 | 1 | 2 | 3 | 7 | 0 | 0 | 248 | 0 | 1 | 0 | 0 | 267 |
| HNSC | 0 | 242 | 1 | 6 | 0 | 1 | 0 | 0 | 0 | 0 | 60 | 0 | 0 | 310 |
| KIRC | 14 | 1 | 1 | 0 | 471 | 4 | 0 | 0 | 1 | 0 | 6 | 0 | 0 | 498 |
| LAML | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 188 | 0 | 0 | 0 | 197 |
| LUAD | 302 | 2 | 2 | 7 | 1 | 12 | 0 | 0 | 0 | 0 | 29 | 0 | 0 | 457 |
| LUSC | 26 | 32 | 0 | 29 | 0 | 7 | 0 | 0 | 0 | 0 | 246 | 0 | 0 | 340 |
| OV | 0 | 0 | 1 | 3 | 0 | 1 | 1 | 348 | 0 | 0 | 0 | 0 | 131 | 485 |
| READ | 1 | 1 | 0 | 5 | 0 | 9 | 145 | 0 | 0 | 0 | 1 | 1 | 0 | 163 |
| UCEC | 3 | 1 | 3 | 117 | 1 | 348 | 1 | 0 | 0 | 0 | 10 | 13 | 2 | 499 |
| Total | 387 | 315 | 613 | 362 | 477 | 581 | 467 | 348 | 249 | 188 | 412 | 17 | 134 | 4550 |

# Stratification – Survival Analysis

- Survival curves for clustered patients

log-rank statistics:



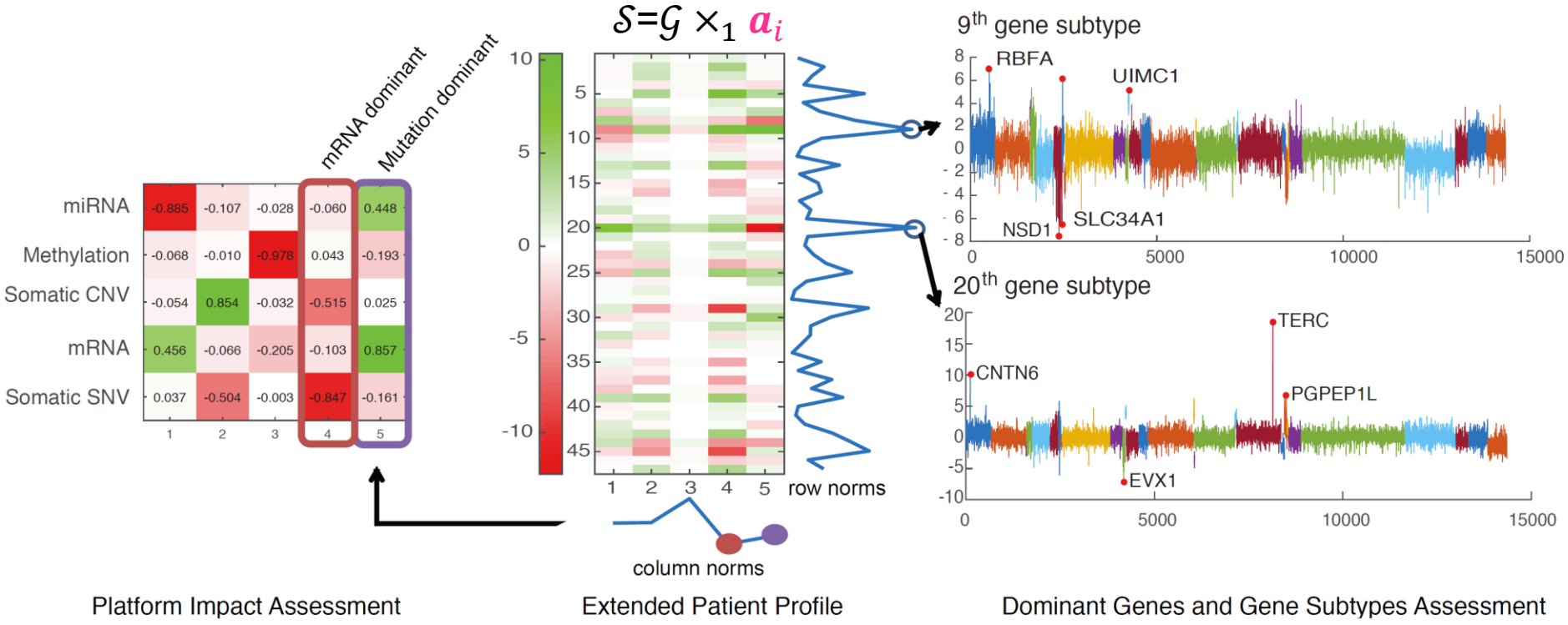
# Prediction – Top-k similarity search

- When a new query patient $q$ arrives with data $\mathcal{X}_q$, calculate factor $\boldsymbol{a}_q$ satisfying following equation: $\boldsymbol{a}_q = arg\min_{\boldsymbol{a}} \left\| \mathcal{X}_q - \mathcal{G} \times_1 \boldsymbol{a} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} \right\|$
- Find top-k similar patients to $q$ and compare

| Cohort | Clinical Features | Top 1 | Top 5 | Top 10 | Top R |
|---|---|---|---|---|---|
| BRCA | Estrogen receptor status | 0.72 | 0.85 | 0.86 | 0.81 |
| COAD | Braf gene analysis result | 1.00 | 0.80 | 0.70 | 0.92 |
| GBM | Histological type | 0.96 | 0.94 | 0.94 | 0.78 |
| HNSC | Hpv status by p16 testing | 0.78 | 0.78 | 0.77 | 0.73 |
| KIRC | Histological type | 1.00 | 0.99 | 0.99 | 0.73 |
| LAML | Calgb cytogenetics risk cat. | 0.85 | 0.84 | 0.81 | 0.65 |
| OV | Neoplasm histologic grade | 0.79 | 0.75 | 0.76 | 0.77 |
| READ | Braf gene analysis result | 1.00 | 1.00 | 1.00 | 1.00 |
| UCEC | Menopause status | 0.71 | 0.76 | 0.76 | 0.77 |

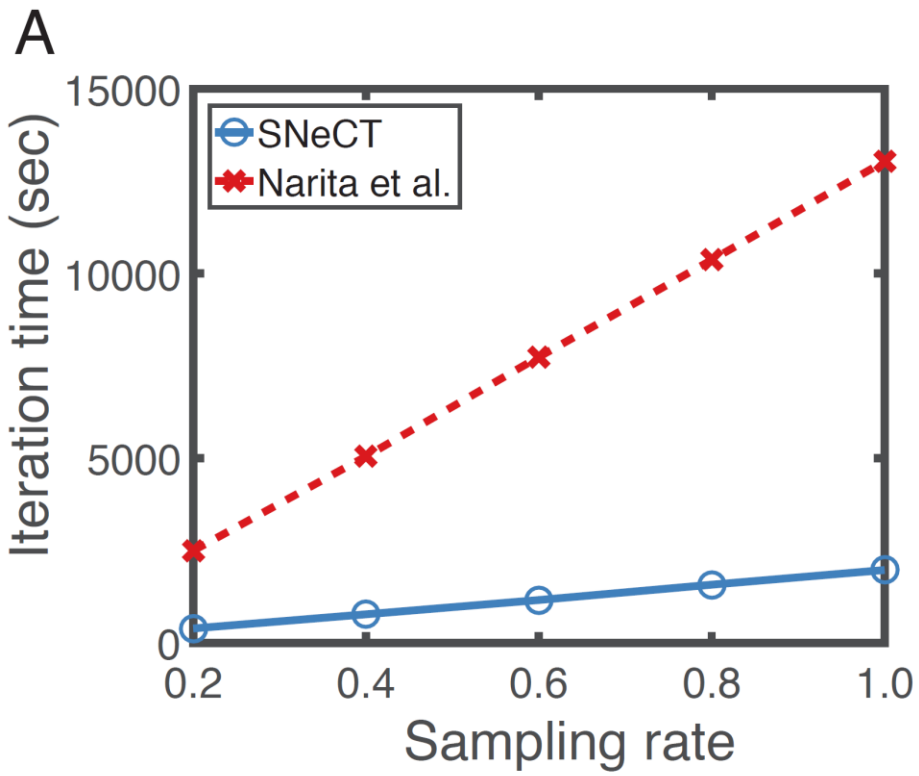
# Personalized subtype analysis

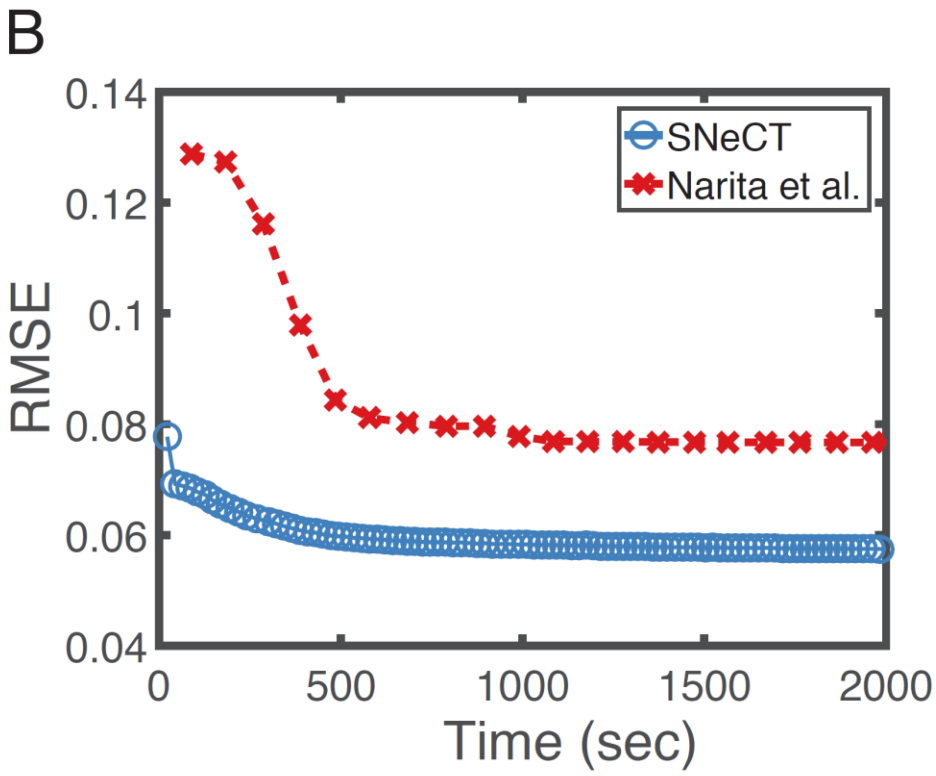
- To provide personalized interpretation for patient $i$, calculate $\mathcal{G} \times_1 \boldsymbol{a}_i = \mathcal{S} (\in R^{Q \times R})$
- Norms of rows represent gene subtype influence
- Norms of columns represent platform subtype influence



# Performance

- Comparison with another network-constrained tensor factorization method: Narita *et al.* 2012
  - **A. Speed**: Iteration time – measured on sampled data
  - **B. Accuracy**: Test RMSE



# Overview

- Introduction
- Problem definition
- Proposed method
- Experiments
- Conclusion

# Conclusion

- SNeCT
  - Parallel algorithms for network constrained tensor factorization
  - Solve tucker decomposition through parallel SGD update scheme
  - Engage common pathway gene network into Pan-Caner12 tensor
  - Utilize patient factor matrix on cluster analysis and survival analysis
  - Propose a personalized subtype analysis scenario

# Thank you!

Questions?