Robust and Accurate Deconvolution of Tumor Populations Uncovers Evolutionary Mechanisms of Breast Cancer Metastasis

Yifeng Tao[1], Haoyun Lei[1], Xuecong Fu[2], Adrian V. Lee[3], Jian Ma[1], Russell Schwartz[1,2]

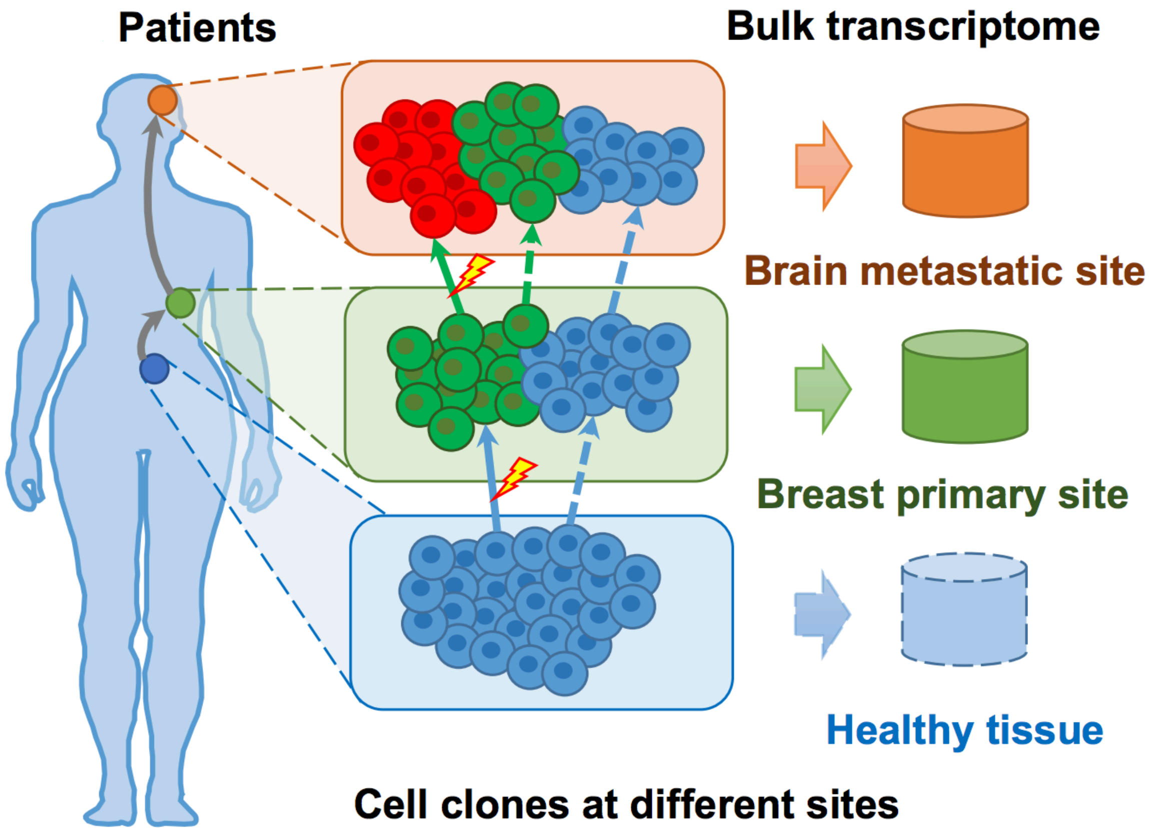
[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University

[2]Department of Biological Sciences, Carnegie Mellon University

[3]Department of Pharmacology and Chemical Biology, UPMC Hillman Cancer Center, Magee-Womens Research Institute

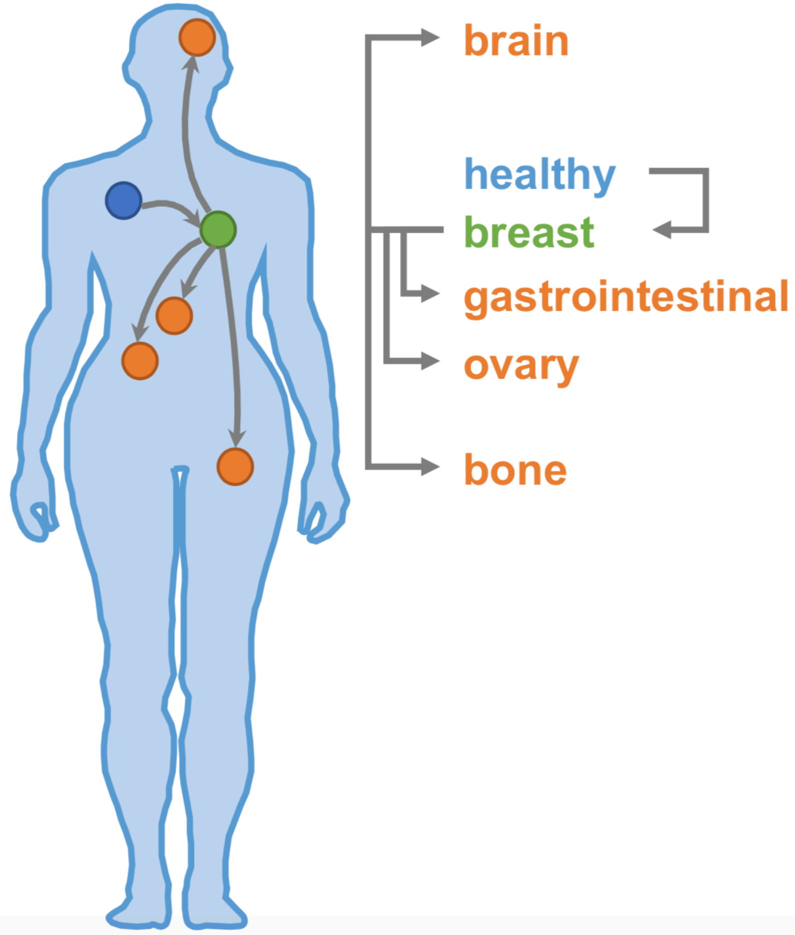
Background: cancer progression and metastasis

- Tumor phylogeny: tumor cells follow a clonal evolution process
- Metastasis: transfer from primary site to other sites
- Heterogeneous tumor populations/clones even from same tissue



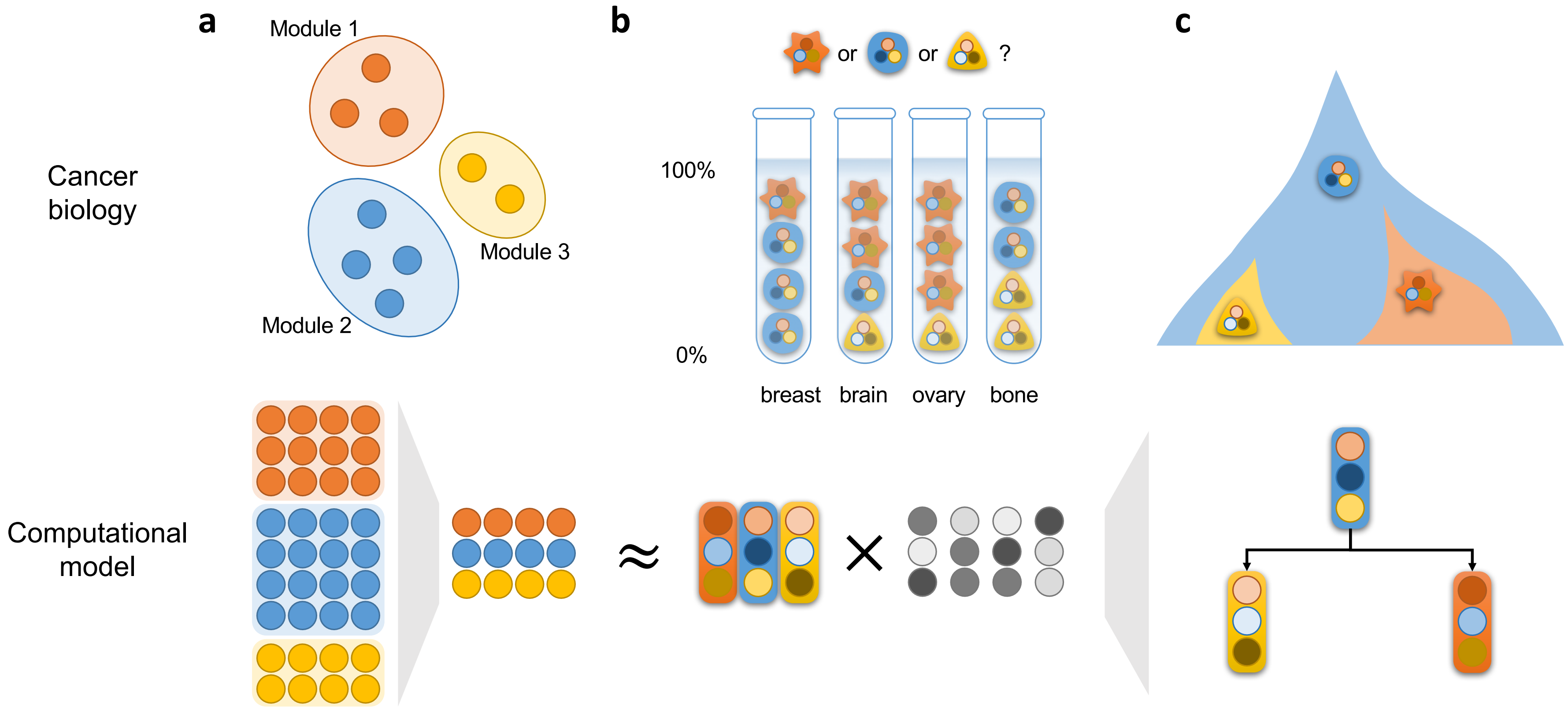
Background: breast cancer metastasis and bulk data

- Breast cancer: second common cause of death from cancer in women
- Breast cancer metastasis (BrM) causes majority of those deaths
- Mechanism of tumor progression during metastasis relies on phylogenetic analysis
- scRNA rarely available due to years between sample collection
- Robust and accurate deconvolution (RAD) of bulk tumor samples is essential



Approach: evolution inference of BrM from bulk RNA

- To boost RAD: knowledge-based gene module (DAVID; DW Huang et al. 2009)
- Core of RAD: bulk sample deconvolution
- Based on RAD-unmixed populations: phylogeny inference (MEP; Tao et al. 2019)

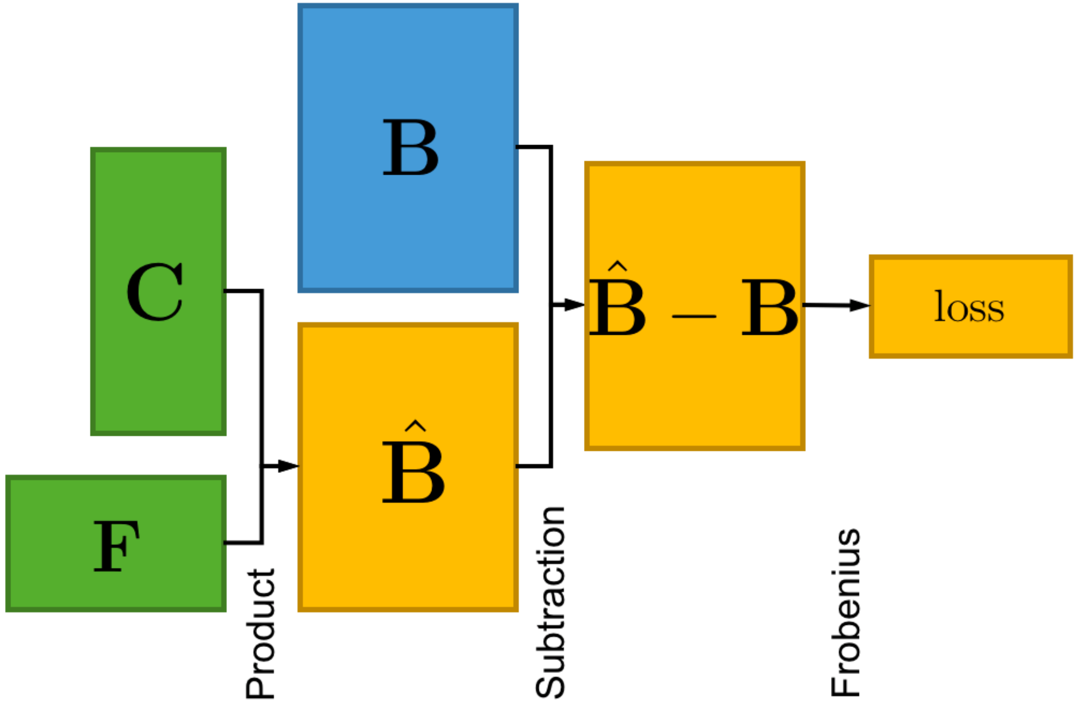


RAD formulation: biologically inspired NMF

- RAD formulated as non-negative matrix factorization (NMF)
  - B: bulk RNA of samples; C: RNA of populations; F: fractions of populations
  - Data noisy and correlated → gene module compression
  - Non-convex and no efficient optimizer → RAD three-phase optimizer
  - *k* not known in prior → cross-validation

$$\min_{\mathbf{C},\mathbf{F}} \quad \|\mathbf{B}-\mathbf{CF}\|^2_{\mathrm{Fr}} \, ,$$

$$\begin{aligned} \text{s.t.} \quad & \mathbf{C}_{il} \geq 0, \quad i=1,\ldots,m,\ l=1,\ldots,k, \\ & \mathbf{F}_{lj} \geq 0, \quad l=1,\ldots,k,\ j=1,\ldots,n, \\ & \textstyle\sum_{l=1}^{k} \mathbf{F}_{lj} = 1, \qquad j=1,\ldots,n \end{aligned}$$



RAD phase 1: multiplicative update warm-start

- Revised multiplicative update (MU) rules
  - Loop until objective stops decreasing

$$\mathbf{C} \leftarrow \mathbf{C} \odot (\mathbf{B}\mathbf{F}^{\intercal}) \oslash (\mathbf{C}\mathbf{F}\mathbf{F}^{\intercal})\,,$$

$$\mathbf{F} \leftarrow \mathbf{F} \odot (\mathbf{C}^{\intercal}\mathbf{B}) \oslash (\mathbf{C}^{\intercal}\mathbf{C}\mathbf{F})\,,$$

$$\boxed{\mathbf{F}_{lj} \leftarrow \mathbf{F}_{lj} \Big/ \sum_{l'=1}^{k} \mathbf{F}_{l'j}\ , \quad l = 1, ..., k,\ j = 1, ..., n}$$

  - MU is non-increasing objective only for general NMF problem (DD Lee et al. 2000)
  - Fast to converge to a reasonable solution

RAD phase 2: coordinate descent

- Coordinate descent
  - Optimizes over C and F iteratively until convergence

$$
\begin{aligned}
\mathbf{C} \leftarrow \arg\min_{\mathbf{C}} \quad & \|\mathbf{B} - \mathbf{C}\mathbf{F}\|_{\mathrm{Fr}}^2 \,, \\
\text{s.t.} \quad & \mathbf{C}_{il} \geq 0, \quad i = 1, ..., m, \; l = 1, ..., k \\
\mathbf{F} \leftarrow \arg\min_{\mathbf{F}} \quad & \|\mathbf{B} - \mathbf{C}\mathbf{F}\|_{\mathrm{Fr}}^2 \,, \\
\text{s.t.} \quad & \mathbf{F}_{lj} \geq 0, \quad l = 1, ..., k, \; j = 1, ..., n, \\
& \textstyle\sum_{l=1}^{k} \mathbf{F}_{lj} = 1, \qquad\qquad j = 1, ..., n
\end{aligned}
$$

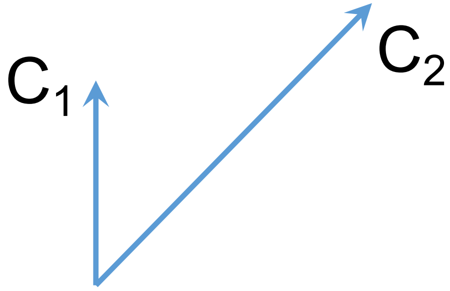
  - Subproblems solved as quadratic programming problems (MS Andersen et al. 2013)
  - Computationally expensive compared with MU warm-start
  - Further reduces loss by ~5-30%

RAD phase 3: minimum similarity selection

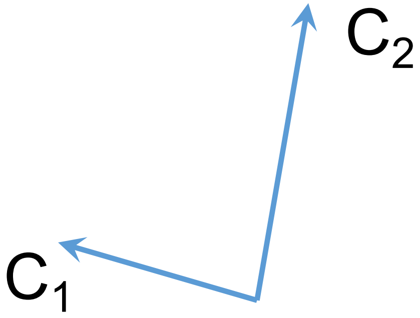
- Minimum similarity selection
  - Repeat random initialization, phase 1 and phase 2 for multiple (e.g., 10) times
  - Select solution with minimum similarity

$$\mathrm{cosim}(\mathbf{C}) = \sum_{l=1}^{k-1} \sum_{l'=l+1}^{k} \mathbf{C}_{\cdot l}^{\intercal} \mathbf{C}_{\cdot l'}$$

  - Better solution: components/populations orthogonal from each other



Solution 1: ✗



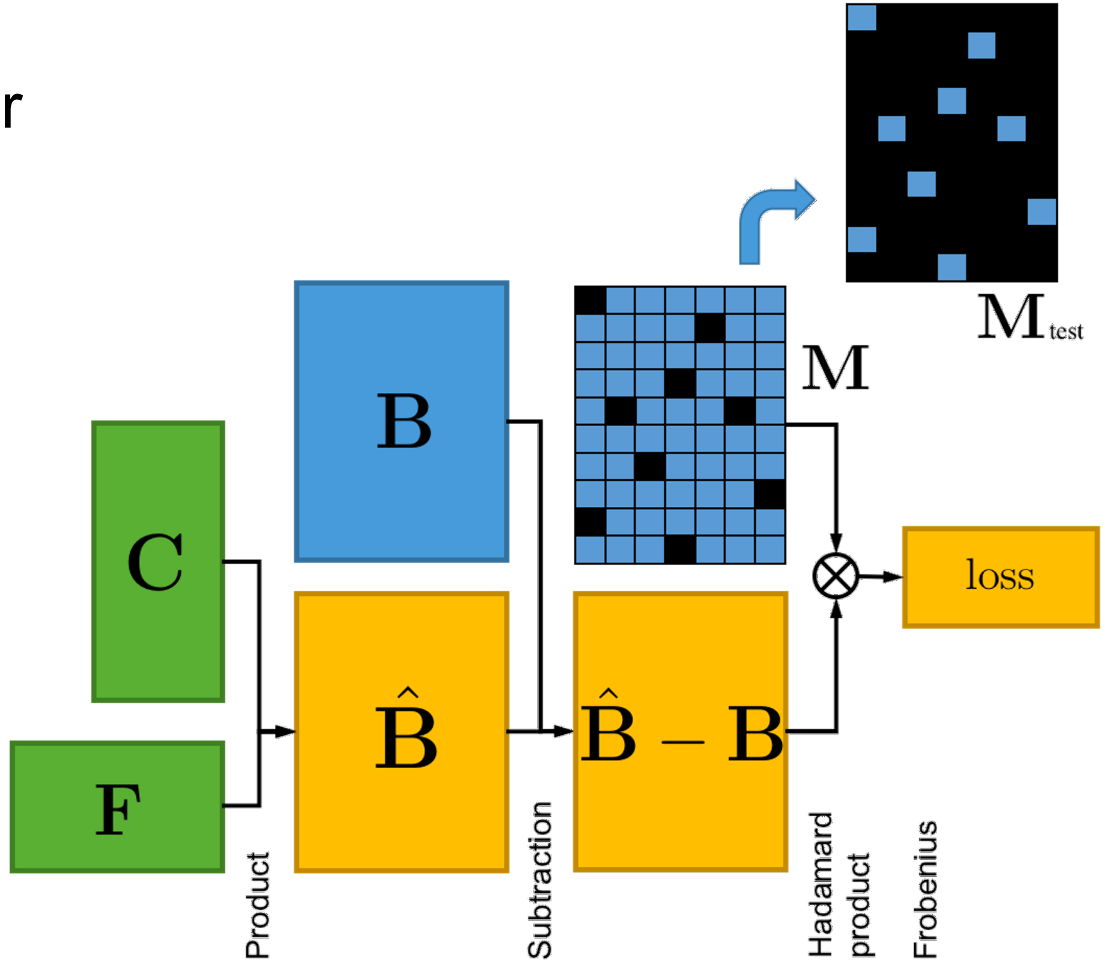
Solution 2:

Population number estimation via RAD

- Masking trick for cross-validation (CV)
- Select *k* that achieves minimum CV error
- Masked RAD algorithm exits!

$$\min_{\mathbf{C},\mathbf{F}} \quad \|\mathbf{M} \odot (\mathbf{B} - \mathbf{CF})\|_{\mathrm{Fr}}^2$$

$$\begin{aligned} \text{s.t.} \quad & \mathbf{C}_{il} \geq 0, \quad i = 1, \ldots, m,\ l = 1, \ldots, k, \\ & \mathbf{F}_{lj} \geq 0, \quad l = 1, \ldots, k,\ j = 1, \ldots, n, \\ & \textstyle\sum_{l=1}^{k} \mathbf{F}_{lj} = 1, \qquad j = 1, \ldots, n \end{aligned}$$

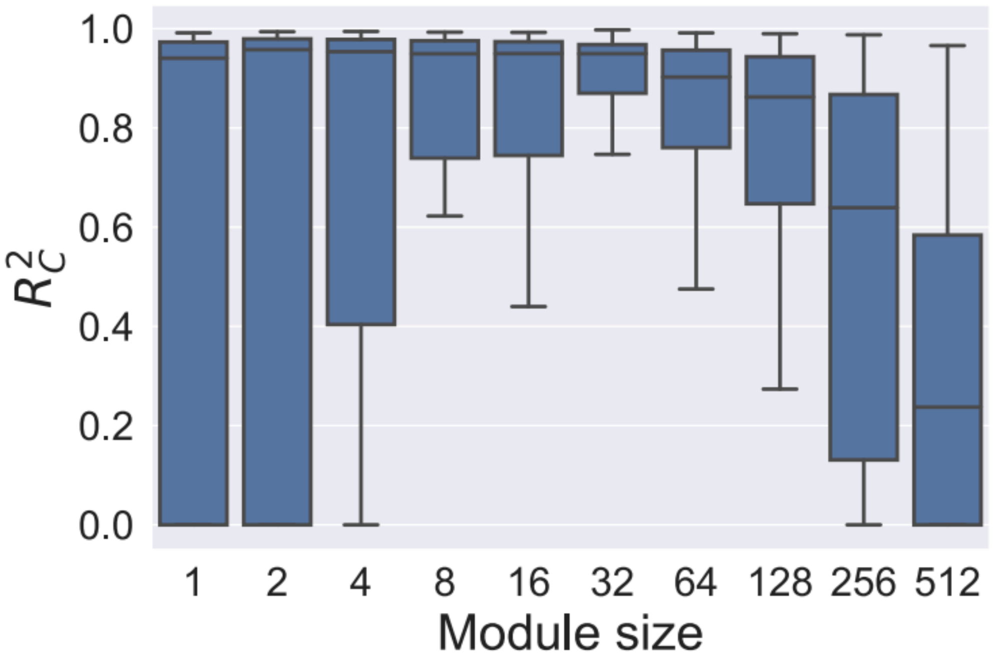


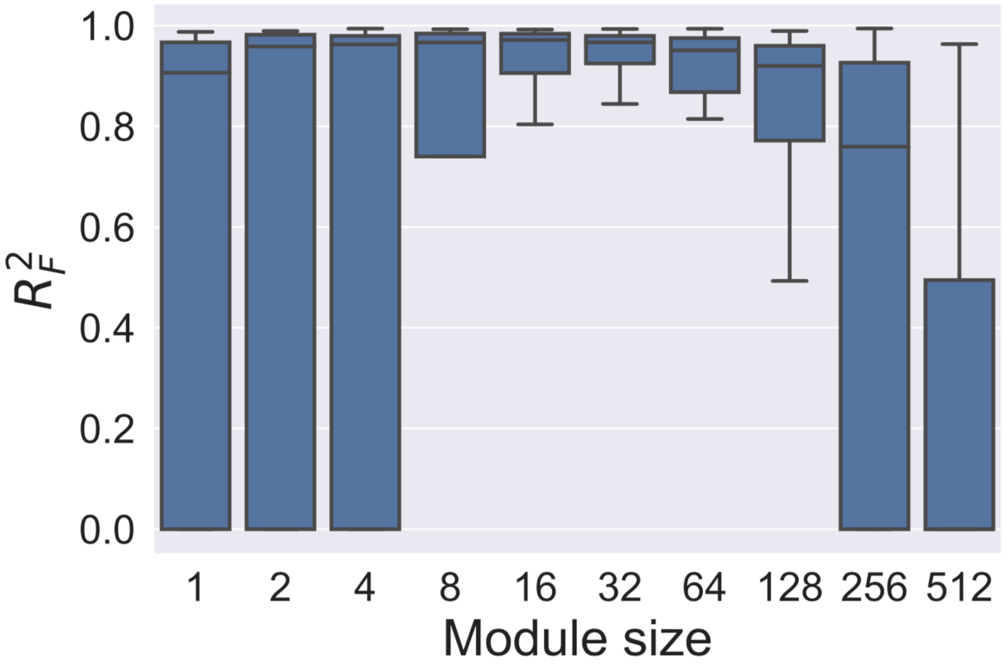
Datasets and experiment design

Dataset	Gene module	Ground truth C and F	Purpose
Simulated (K Zaitsev et al. 2019)	Known	Known	• Evaluate effect of gene module
GSE19830 (SS Shen-Orr et al. 2010)	Knowledge base	Known	• Evaluate effect of gene module • Evaluate RAD accuracy on estimating C, F, and *k*
BrM (L Zhu et al. 2019)	Knowledge base	Unknown	• Understand breast cancer metastasis mechanism

Gene modules facilitate robust deconvolution

- Simulated datasets: gene module known
  - Too small module size → fragile deconvolution
  - Too large module size → worse estimation

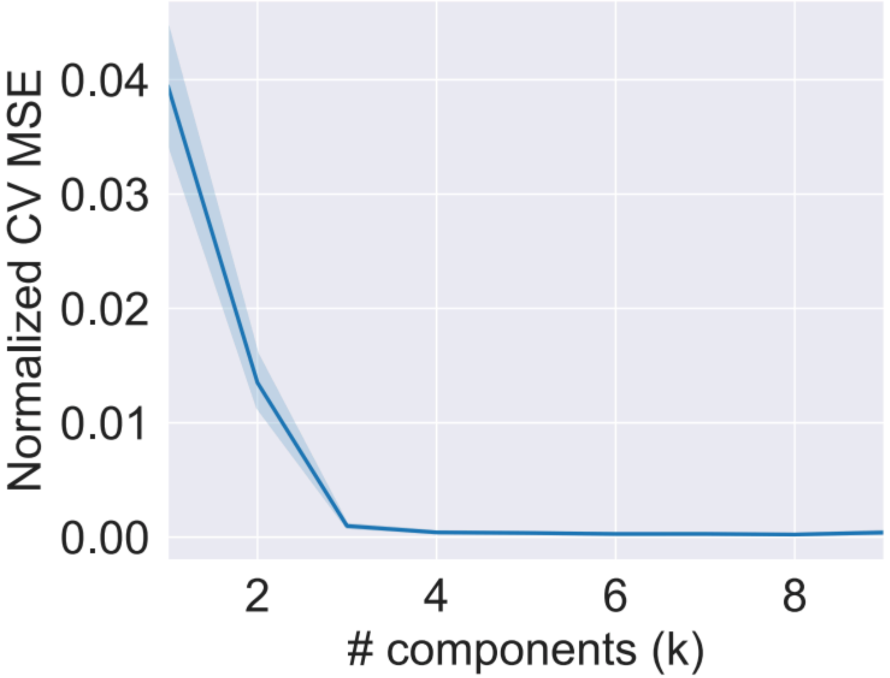




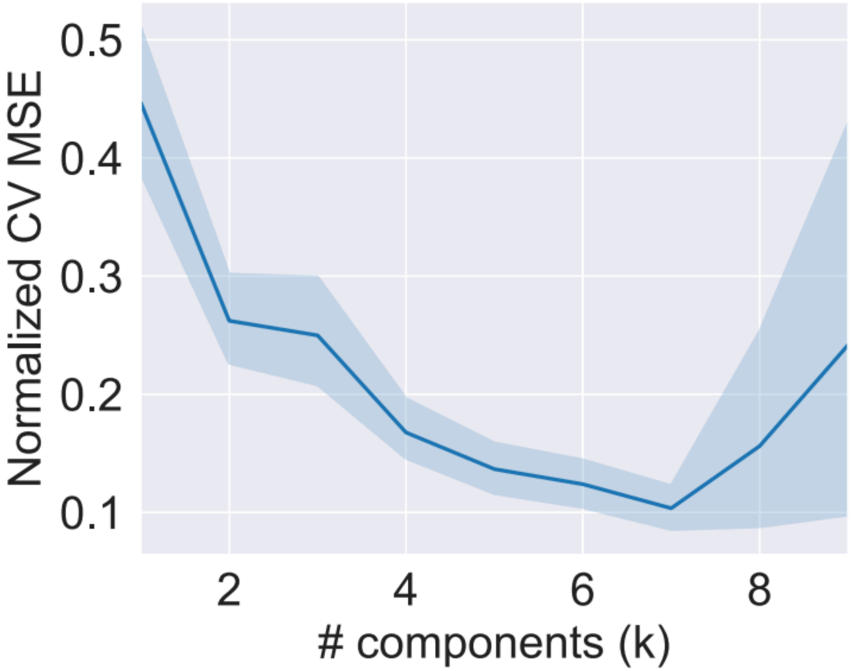
RAD detects correct number of cell components

- GSE19830: three cell types known in advance
- BrM: ground truth cell types unknown

GSE19830

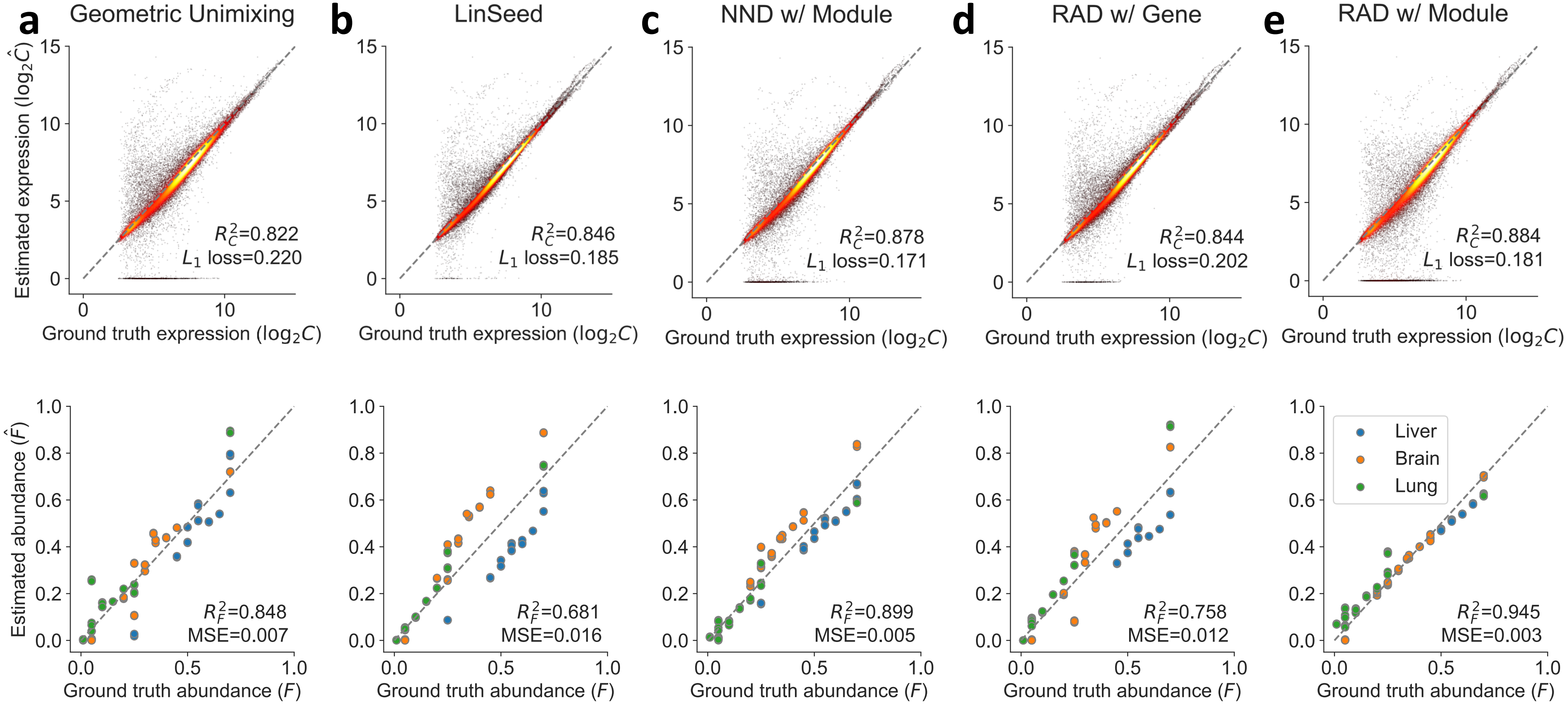


BrM



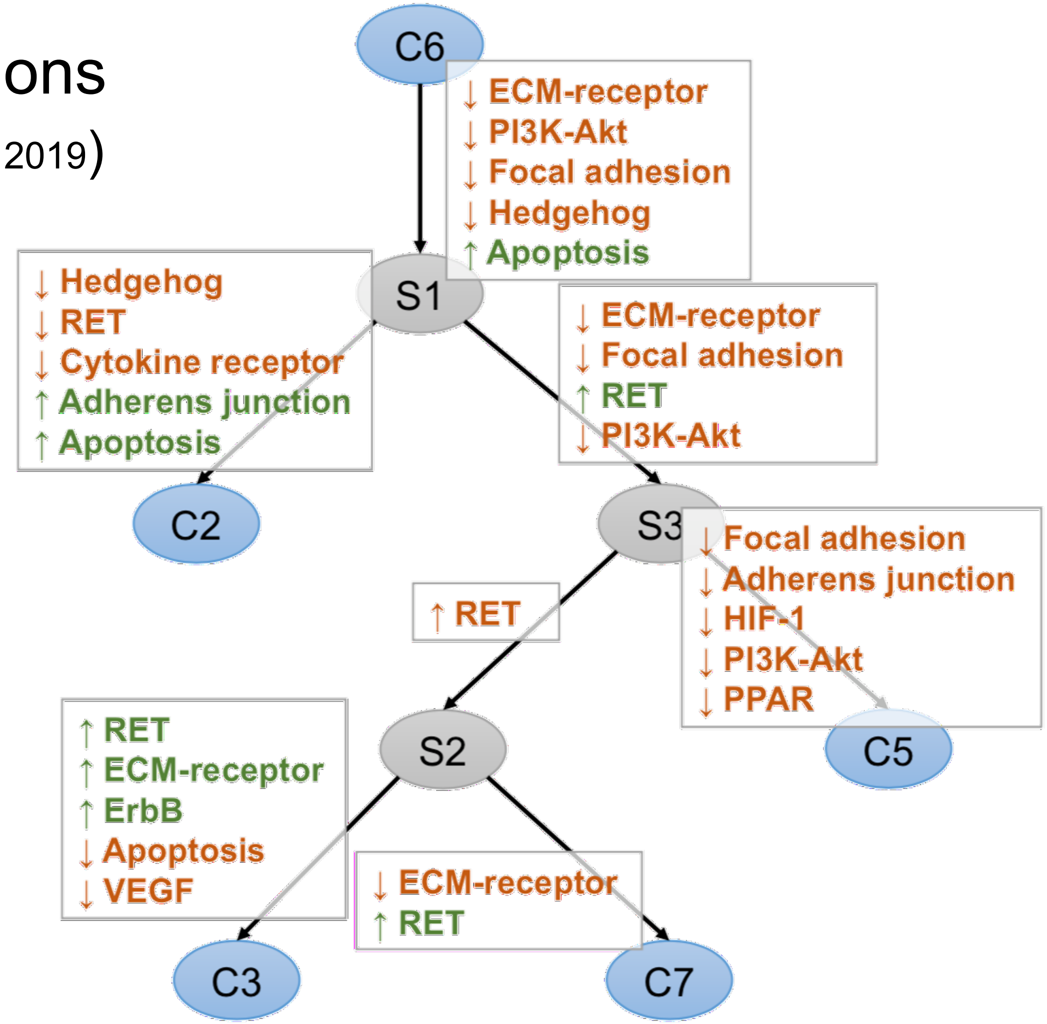
RAD estimates populations more accurately

- Outperforms three competing methods on GSE19830 dataset
- Gene module inferred from knowledge base improves RAD as well



Common evolutionary mechanisms of BrM

- Infer phylogenies from RAD-unmixed populations
  - Minimum elastic potential (MEP; Nei et al. 1987, Tao et al. 2019)
  - Four cases in total (one shown)
- Common early pathway-level events
  - ↓ PI3K-Akt (PK Brastianos et al. 2015)
  - ↓ Extracellular matrix (ECM)-receptor interaction
  - ↓ focal adhesion (M Nagano et al. 2012)



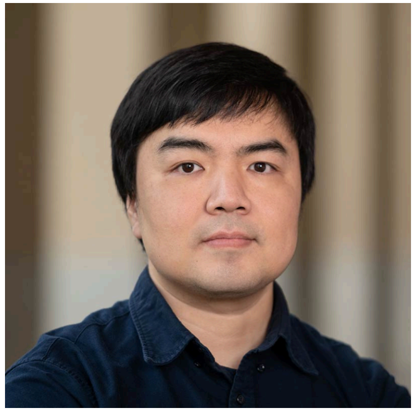
Conclusion and future work

- Deconvolution of bulk data is the key to understanding the BrM progression
- We propose RAD, a toolkit that accurately and robustly estimates the number of cell populations (*k*), expression profiles of cell populations (C), and fractions of populations (F)
- Through RAD, we find the loss of PI3K-Akt, ECM-receptor interaction, and focal adhesion emerge as the common early pathway-level events of BrM

- Integrate single cell data of metastatic samples to improve RAD performance

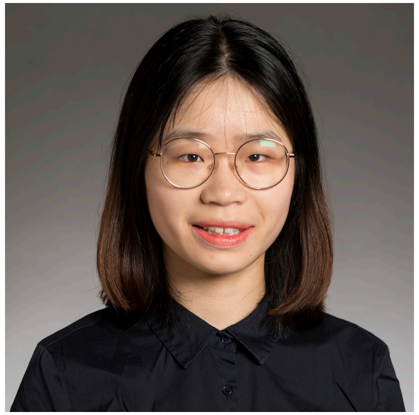
Acknowledgments

Dr. Russell Schwartz


Dr. Jian Ma

Dr. Adrian V. Lee

Haoyun Lei


Xuecong Fu

NIH National Institutes of Health
Turning Discovery Into Health

Mario Lemieux Foundation

pennsylvania
DEPARTMENT OF HEALTH

Susan G. Komen for the Cure

Pittsburgh Health Data Alliance
Carnegie Mellon University University of Pittsburgh UPMC
Center for Machine Learning and Health
Carnegie Mellon University

aws

BREAST CANCER ALLIANCE

CMUSchwartzLab/RAD

Follow @Yifeng_Tao