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

Yifeng Tao¹, Haoyun Lei¹, Xuecong Fu², Adrian V. Lee³, Jian Ma¹, Russell Schwartz¹,²

¹Computational Biology Department, School of Computer Science, Carnegie Mellon University
²Department of Biological Sciences, Carnegie Mellon University
³Department 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 \( \rightarrow \) gene module compression
  - Non-convex and no efficient optimizer \( \rightarrow \) RAD three-phase optimizer
  - \( k \) not known in prior \( \rightarrow \) cross-validation

\[
\min_{C, F} \| B - CF \|_{Fr}^2,
\]

s.t. \( C_{i l} \geq 0, \quad i = 1, \ldots, m, \ l = 1, \ldots, k, \)
\[
F_{l j} \geq 0, \quad l = 1, \ldots, k, \ j = 1, \ldots, n,
\]
\[
\sum_{l=1}^{k} F_{l j} = 1, \quad j = 1, \ldots, n
\]
RAD phase 1: multiplicative update warm-start

- **Revised** multiplicative update (MU) rules
  - Loop until objective stops decreasing
    
    $$
    C \leftarrow C \odot (BF^T) \odot (CFF^T), \\
    F \leftarrow F \odot (C^TB) \odot (C^TCF), \\
    F_{lj} \leftarrow F_{lj} / \sum_{l' = 1}^{k} F_{lj}, \quad l = 1, \ldots, k, j = 1, \ldots, n
    $$

  - 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

\[
C \leftarrow \arg \min_C \| B - CF \|_{F_r}^2, \\
\text{s.t. } C_{il} \geq 0, \quad i = 1, \ldots, m, \quad l = 1, \ldots, k
\]

\[
F \leftarrow \arg \min_F \| B - CF \|_{F_r}^2, \\
\text{s.t. } F_{lj} \geq 0, \quad l = 1, \ldots, k, \quad j = 1, \ldots, n, \\
\sum_{l=1}^{k} F_{lj} = 1, \quad j = 1, \ldots, n
\]

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

$$\text{cosim}(C) = \sum_{l=1}^{k-1} \sum_{l'=l+1}^{k} C_{l}^T C_{l'}$$

- Better solution: components/populations orthogonal from each other

Solution 1: \(X\)

Solution 2:
Population number estimation via RAD

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

$$\min_{C,F} \| M \odot (B - CF') \|_F^2$$

s.t. $C_{il} \geq 0, \quad i = 1, \ldots, m, \ l = 1, \ldots, k,$

$F_{lj} \geq 0, \quad l = 1, \ldots, k, \ j = 1, \ldots, n,$

$$\sum_{l=1}^{k} F_{lj} = 1, \quad j = 1, \ldots, n$$
Datasets and experiment design

| Dataset                  | Gene module     | Ground truth C and F | Purpose                                                                 |
|--------------------------|-----------------|----------------------|-------------------------------------------------------------------------|
| Simulated (κ 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 $\rightarrow$ fragile deconvolution
  - Too large module size $\rightarrow$ worse estimation
RAD detects correct number of cell components

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