Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Semi-deconvolution of bulk and single-cell RNA-seq data with application to metastatic progression in breast cancer

Haoyun Lei [1,†], Xiaoyan A. Guo [1,2,†], Yifeng Tao [1], Kai Ding [3], Xuecong Fu [2], Steffi Oesterreich [3], Adrian V. Lee[3], and Russell Schwartz [1,2,*]

[1]Computational Biology Dept., Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2]Dept. of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3]Department of Pharmacology and Chemical Biology, UPMC Hillman Cancer Center, Magee-Womens Research Institute, Pittsburgh, 15213, USA

[†]Co-first authors; [*]To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Identifying cell types and their abundances and how these evolve during tumor progression is critical to understanding the mechanisms of metastasis and identifying predictors of metastatic potential that can guide the development of new diagnostics or therapeutics. Single-cell RNA sequencing (scRNA-seq) has been especially promising in resolving heterogeneity of expression programs at the single cell level, but is not always feasible, for example for large cohort studies or longitudinal analysis of archived samples. In such cases, clonal subpopulations may still be inferred via genomic deconvolution, but deconvolution methods have limited ability to resolve fine clonal structure and may require reference cell type profiles that are missing or imprecise. Prior methods can eliminate the need for reference profiles but show unstable performance when few bulk samples are available.

Results: In this work, we develop a new method using reference scRNA-seq to interpret sample collections for which only bulk RNA-seq is available for some samples, e.g., clonally resolving archived primary (PRM) tissues using scRNA-seq from metastases (METs). By integrating such information in a Quadratic Programming (QP) framework, our method can recover more accurate cell types and corresponding cell type abundances in bulk samples. Application to a breast tumor bone metastases dataset confirms the power of scRNA-seq data to improve cell type inference and quantification in same-patient bulk samples.

Availability: Source code is available on Github at https://github.com/CMUSchwartzLab/RADs

Contact: russells@andrew.cmu.edu

1 Introduction

Computational methods for resolving single-cell clonal evolutionary dynamics (Greaves and Maley, 2012) have become a central part of modern cancer genomics research as ever more powerful genomic tools have become available and as the role of tumor heterogeneity and clonal evolution in cancer progression have become more apparent (Beerenwinkel *et al.*, 2016). The fundamental goal of such methods is to characterize the genetics and genomics of tumor cells and various other cell types infiltrating them and understand how these populations of cells predict future tumor progression and evolve over its course. Numerous variations on this basic framework have been developed, for different kinds of genomic data (e.g., DNA-seq vs. RNA-seq, bulk vs. single-cell) or different research questions (e.g., understanding genetic vs. phenotypic evolution) (c.f., (Schwartz and Schäffer, 2017)). In the present work, we focus on one specific scenario: understanding RNA evolution in settings in which bulk and single-cell data are available for different time points, samples, or stages of progression.

Prior to the emergence of practical single-cell methods, most techniques for reconstructing clonal evolution depended on a strategy called *genomic deconvolution* (Lu *et al.*, 2003), in which one seeks to resolve clonal evolution by computationally inferring activities of homogeneous cell populations from mixtures of genomic data contained in bulk samples (Schwartz and Shackney, 2010). Methods for this problem can roughly be classified into two classes: partial deconvolutional algorithms, which interpret data in terms of a reference matrix of known cell types, and complete deconvolutional algorithms, which infer cell types *de novo* by comparison of multiple samples. Examples of partial deconvolution methods include DSA (Zhong *et al.*, 2013), which treats deconvolution as a matrix factorization problem; CIBERSORT, which makes use of a pre-defined LM22 signature matrix; and CIBERSORTx (Newman *et al.*, 2019), which derives a signature matrix for intepreting bulk data using reference single-cell data. See (Avila Cobos *et al.*, 2020) for a comparison of different partial deconvolution algorithms and the effects of various data transformation methods. Examples of complete deconvolution methods include Geometric Unmixing (Schwartz and Shackney, 2010), which proposed a strategy called archetype analysis based on geometric analysis of genomic point clouds; LinSeed (Zaitsev *et al.*, 2019), which identifies a set of anchor genes through linear correlation and uses the partial deconvolution algorithm DSA to solve for the non-anchor genes; NND (Tao *et al.*, 2020a), which poses partial deconvolution problem as a matrix factorization to be solved with gradient descent implemented through a neural network and RAD (Tao *et al.*, 2020b), which solved the formulation of NND using a hybrid optimizer with improved accuracy and speed.

Single-cell genomics has rapidly displaced deconvolutional bulk methods as the preferred practice for tumor genomic analysis, particularly for RNA-seq variants, as single-cell sequencing has become reliable and cost-effective (c.f., Kuipers *et al.* (2017)). Although single-cell sequencing introduces its own computational complications and data quality issues, having large numbers of direct single-cell measurements leads to substantially greater resolution for single-cell variation than is possible for deconvolutional methods even with high quality bulk data. Nonetheless, deconvolutional methods remain necessary in practice for a variety of real-world use cases. Single-cell data is not typically possible for older archived samples and the field still lacks large cohorts of single-cell tumor genomic data comparable to bulk resources such as the influential Cancer Genome Atlas (TCGA) (Chang *et al.*, 2013) or International Cancer Genome Consortium (ICGC) (Zhang *et al.*, 2011) datasets. The problem is particularly acute for current studies of patients being tracked longitudinally, for example in using clonal phylogenetics to understand metastatic progression (Naxerova and Jain, 2015). Patients being seen today for metastatic disease may have been first diagnosed years earlier and understanding the evolution of their tumors may require comparing recent samples, for which single-cell data is practical, with archived primary samples for which only bulk data is possible.

The present work was developed specifically to address scenarios such as this, in which one seeks to understand longitudinal progression of a cancer by comparing samples some of which may be amenable to single-cell methods (e.g., a recent metastases) and others of which can only be examined by bulk methods (e.g., an archived primary tumor biopsy). It accomplishes this by developing a hybrid algorithm to infer genomics and clonal frequencies in both bulk and single-cell samples, using single-cell data in some samples as partial references for bulk data in others. In this regard, it follows a strategy of mixed bulk and single-cell data previously applied in tumor evolutionary studies primarily to DNA-seq data (Malikic *et al.*, 2019b,a; Lei *et al.*, 2020). It poses the problem in terms of a matrix factorization formulation comparable to earlier bulk deconvolution methods, drawing on ideas from hybrid bulk/single-cell DNA methods to integrate the heterogeneous data sources. We demonstrate through simulated data and application to real paired primary and metastatic patient data that the methods are effective at resolving cell population dynamics across time points in comparison to more traditional deconvolution methods, providing novel insight into how cell population dynamics can underlie tumor progression.

2 Methods

Our method can roughly be divided into two steps: First, we develope a model to correctly infer a signature matrix $\mathbf{S}$ to represent the single-cell data from metastatic samples. Then, we apply a semi-deconvolution algorithm that differs from both the previous partial and complete deconvolution algorithms in optimizing to avoid problems of bias and variance that plagued prior approaches. The objective/loss function consists of two parts: a complete deconvolution part, which is less biased by the reference data but would be expect to suffer from high variance, and a partial deconvolution part, which would be expected to lead to results with smaller variance but higher bias particularly when using independent reference data for partial deconvolution. By combining these terms, we seek to mitigate the complementary weaknesses of each approach. The approach is particularly effective for characterizing how population frequencies of distinct cell types evolve over progression stages, for example via differential immune infiltration (c.f., Sturm *et al.* (2019)). Unlike other methods also using reference signatures, our method allows the inference of cell types not found in the single-cell samples by introducing a freedom component to balance the difference between bulk and single-cell samples while still keeping the variance low, a key consideration since we expect to see cell types at a primary site that are not found in metastatic sites. We call our method RAD with single-cells (RADs). The operation and a high level description of the method are summarized in Fig. 1.

2.1 Datasets

2.1.1 Simulated datasets

It is not possible to establish with certainty the ground truth for any real deconvolution data set and so we rely partially on simulated data for validation. Our main strategy is to rely on true single-cell RNA-seq data from paired primary and metastatic breast cancer samples to generate artificial bulk data composed of mixtures of single cells, for which we would then have a known ground truth. We simulated bulk RNA-Seq of primary tumor tissue samples based on the following assumptions: 1) the underlying gene expression profile of each cell type in bulk RNA-Seq would be similar to the average expression values of cells belonging to the same type in single-cell RNA-Seq had dropout events in single-cell RNA-Seq not occurred and 2) the fractions of different cell types in metastatic tissue samples will be different from those in primary tumor tissue samples.

First, we inferred the cell type expression matrix of bulk RNA-Seq from the real single-cell RNA-Seq data of metastatic tumor tissue samples. The single-cell RNA-Seq results from two different samples were each normalized to CPM (Counts Per Million) before being aggregated. For each cell type, the aggregated cell expression values were averaged to obtain its gene expression profile. The gene expression profile was normalized to CPM again before being corrected for the effect of single-cell RNA-seq drop-out events. See section 2.3 for details of the correction.

The corrected gene expression profile then served as the ground truth cell type expression matrix for simulating the bulk RNA-Seq samples. 5000 randomly selected genes and all k cell types from the gene expression profile were $\log_2$ transformed and replicated n times to generate each of the n primary tumor tissue sample's component matrix. Inter-sample noise was added to the component matrices by replacing each of j^{th} gene's

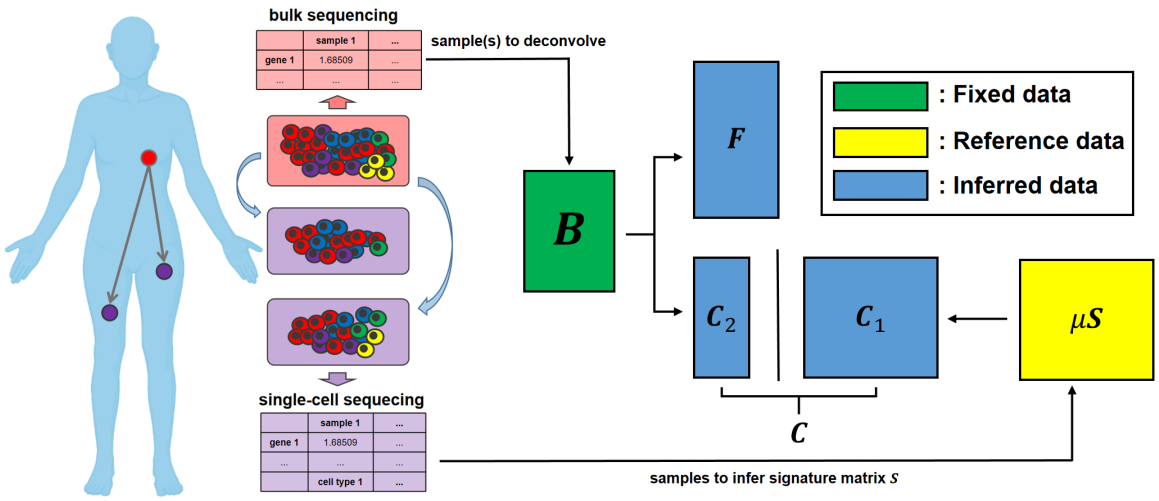

Fig. 1. Approach to semi-deconvolution using both bulk and single-cell RNA-Seq to uncover the cell type dynamics across progression stages. The left scheme shows an example of tumor sites that might lead to distinct bulk and single-cell RNA-seq samples. The right scheme shows the overall mathematical problem to solve using the bulk and single-cell RNA-seq. $\mathbf{B}$: bulk sample from primary tumor to deconvolve. $\mathbf{C_1}$: known cell types found in single-cell metastatic samples. $\mathbf{C_2}$: unknown cell types only in primary tumor. $\mathbf{C}$: horizontal stack of $\mathbf{C_1}$ and $\mathbf{C_2}$ to compose the total cell types in primary tumor. $\mathbf{F}$: corresponding fractions of cell types. $\mathbf{S}$: signature matrix inferred from single-cell metastatic samples.

expression value x with one instance drawn from the Gaussian distribution $N(x, \sigma_j/5)$, where σ_j is the standard deviation of j^{th} gene's expression values across cell types in the $\log_2$ space. The noisy component matrices were then projected back to the linear space. The fraction vector of the k cell types in these primary tumor samples was generated by perturbing the frequency of each cell type in the real single-cell RNA-Seq of metastatic tissue samples in $\log_2$ scale: each of the k frequencies f_i was replaced by with one instance drawn from the Gaussian distribution $N(f_i, \sigma_f/2)$, with σ_f being the standard deviation among the k cell types' frequencies. This fraction vector was then replicated n times before the addition of inter-sample noise in the same way as the cell type-gene expression component matrix. The fractions for all k cell types were then summed to normalize the matrix, resulting in a $n \times k$ fraction matrix where each of the n fraction vector sums to 1.

Finally, the n simulated bulk RNA-Seq of primary tumor tissue samples were generated by convolving each component matrix with its corresponding fraction vector.

2.1.2 Bulk and Single-cell RNA-Seq datasets

Statistics of bulk and single-cell RNA dataset used in the present work can be found in Table 1.

Dataset	PBT	BoM1	BoM2
Data type	Breast primary	Bone metastases at left acetabulum	Bone metastases at right tibia
# genes	57,557	18,386	18,386
# samples	1	4,649	5,505
# cell types	-	6	26
# fine cell types	-	6	24

Table 1. Match bulk RNA-Seq datasets used in the present study.

2.2 Constructing signature matrix from single-cell RNA-Seq data

The signature matrix $\mathbf{S}^{m \times k}$ can be derived by averaging the expression values of each measured gene across cells sharing the same cell type annotation in single-cell RNA-Seq. However, single-cell RNA-Seq data includes noise and bias from dropout events, which requires additional correction to meaningfully guide the deconvolution of dropout-free bulk RNA-Seq. In general, the smaller expression level for a cell/gene, the higher dropout rate it will have.

To correct for the effect of dropout events, we assume that the relationship between the j^{th} gene's average expression value μ^j and its dropout frequency follows a Michaelis-Menten function:

$$\mathbf{P}^j_{dropout} = 1 - \frac{\mu^j}{\mathbf{K}^j_M + \mu^j}, \tag{1}$$

where $\mathbf{K}^j_M$ is the Michaelis constant representing the mean expression value of j^{th} gene required for half of the cells to be detected. Therefore, we could fit the Michaelis-Menten function with mean expression values and percentages of values dropped from the real single-cell RNA-Seq data to estimate the $\mathbf{K}_M$ for each gene. Then given any average expression value for the cell type, it would be possible to estimate $\mathbf{P}_{dropout}$ using the Michaelis-Menten function. The average expression value for the j^{th} gene of bulk RNA-Seq data $\mathbf{C}_j$ without any dropout can then be inferred as: $\mathbf{C}_j = \frac{\mu^j}{1-\mathbf{P}^j_{dropout}}$.

2.3 Mathematical formulation of deconvolution

We aim to achieve better deconvolution algorithms by integrating the information from the noisy but informative single-cell RNA seq.

$$\min_{\mathbf{C},\mathbf{F},\mu} \quad \|\mathbf{B} - \mathbf{CF}\|^2_{\mathrm{Fr}} + \lambda\|\mathbf{C}_1 - \mu\mathbf{S}\|^2_{\mathrm{Fr}}, \tag{2}$$

$$\text{s.t.} \quad \mathbf{C}_{il} \geq 0,\ i = 1, ..., m,\ l = 1, ..., K, \tag{3}$$

$$\mathbf{F}_{lj} \geq 0,\ l = 1, ..., K,\ j = 1, ..., n, \tag{4}$$

$$\textstyle\sum_{l=1}^{K} \mathbf{F}_{lj} = 1,\ j = 1, ..., n, \tag{5}$$

where $\mathbf{B} \in \mathbb{R}_{\geq 0}^{m \times n}$ is the bulk RNA-Seq data, $\mathbf{S} \in \mathbb{R}_{\geq 0}^{m \times k}$ the matrix inferred from the single-cell reference data, λ is the penalty weight that suppresses the discrepancy between single-cell data and inferred expression profile matrix, μ to adjust the scale of inferred single-cell references.

Each row i is a gene, each column j is a bulk sample, k represent the k-th cell type/population.

Based on NND, we introduced additional parameters to optimize the problem above Eq. 2-5, and the meaning of all the necessary parameters are present in Table 2.

$\mathbf{B}^{m \times n}$	bulk samples (m gene $\times$ n samples) in primary tumor
$\mathbf{C_1}^{m \times k}$	known cell types (m gene $\times$ k known cell types) in primary tumor
$\mathbf{C_2}^{m \times y}$	possible unknown cell types (m gene $\times$ y unknown cell types) in primary tumor
$\mathbf{F}^{K \times n}$	fraction of cell types (K cell types $\times$ n samples) in primary tumor
$\mathbf{C}^{m \times K}$	total cell types (m gene $\times$ K cell types) in primary tumor; note $K = k + y$
$\mathbf{S}^{m \times k}$	representative reference (m gene $\times$ k cell types) from single-cell in metastasis
μ	scaling factor for $\mathbf{S}$
λ	penalty term to balance information from $\mathbf{S}$

Table 2. Parameters and constants notations

2.4 Solving the deconvolution problems

The original problem formulation Eq. 2-5 is non-negative. Inspired by previous work, we use a coordinate descent method to obtain a suboptimal solution to the problem. The coordinate descent algorithm iteratively repeats the following three phases until convergence. In each phase, the computational problem can be re-formulated to a well-known problem class, e.g., quadratic programming or linear regression, that can be easily solved.

Phase 1: Optimizing $\mathbf{F}$ At this phase, we fix $\mathbf{C}$ and μ to optimize $\mathbf{F}$, then the Eq. 2-5 is equivalent to:

$$\min_{\mathbf{F}} \quad \|\mathbf{B} - \mathbf{CF}\|_{\mathrm{Fr}}^2 , \tag{6}$$

$$\text{s.t.} \quad \mathbf{F}_{lj} \geq 0,\ l = 1, ..., k,\ j = 1, ..., n, \tag{7}$$

$$\sum_{l=1}^{k} \mathbf{F}_{lj} = 1,\ j = 1, ..., n, \tag{8}$$

Let $\mathbf{f} = \mathbf{F}_{\cdot j}$, which stands for each column in Eq. 6 that represents the frequency of each cell type/population in one bulk sample $\mathbf{B}_{\cdot j}$ (defined as $\mathbf{b}$). Then we can re-formulate it to minimize the error for each column of $\mathbf{B} - \mathbf{CF}$:

$$\min_{\mathbf{F}} \quad \|\mathbf{B} - \mathbf{CF}\|_{\mathrm{Fr}}^2 , \tag{9}$$

$$\iff \min_{\mathbf{F}} \quad \left[\sum_{j=1}^{n} \|\mathbf{B}_{\cdot j} - \mathbf{CF}_{\cdot j}\|_2^2 \right], \tag{10}$$

$$\iff \min_{\mathbf{F}_{\cdot j}} \quad \|\mathbf{B}_{\cdot j} - \mathbf{CF}_{\cdot j}\|_2^2 , \quad \forall j = 1, .., n \tag{11}$$

$$\iff \min_{\mathbf{f}} \quad \|\mathbf{b} - \mathbf{Cf}\|_2^2 , \tag{12}$$

(subscription j omitted for clarity)

$$\iff \min_{\mathbf{f}} \quad \mathbf{f}^\intercal \mathbf{C}^\intercal \mathbf{C} \mathbf{f} - 2\mathbf{b}^\intercal \mathbf{C} \mathbf{f} + \mathbf{b}^\intercal \mathbf{b}, \tag{13}$$

$$\iff \min_{\mathbf{f}} \quad \frac{1}{2} \mathbf{f}^\intercal \mathbf{C}^\intercal \mathbf{C} \mathbf{f} - \mathbf{b}^\intercal \mathbf{C} \mathbf{f} \tag{14}$$

It is not hard to see that $(\mathbf{C}^\intercal \mathbf{C})^\intercal = \mathbf{C}^\intercal (\mathbf{C}^\intercal)^\intercal = \mathbf{C}^\intercal \mathbf{C}$, so $\mathbf{C}^\intercal \mathbf{C}$ is symmetric matrix, and further that it will be semidefinite since $\mathbf{f}^\intercal \mathbf{C}^\intercal \mathbf{C} \mathbf{f} = (\mathbf{Cf})^\intercal (\mathbf{Cf}) \geq 0$, establishing that Eq. 14 can be solved a Quadratic Programming (QP) problem with the same constraints shown in Eq. 7 and 8.

Phase 2: Optimizing $\mathbf{C}$ At this phase, we fix $\mathbf{F}$ and μ to optimize $\mathbf{C}$, then the Eq. 2 equals to:

$$\min_{\mathbf{C}} \quad \|\mathbf{B} - \mathbf{CF}\|_{\mathrm{Fr}}^2 + \lambda \|\mathbf{C}_1 - \mu \mathbf{S}\|_{\mathrm{Fr}}^2, \tag{15}$$

$$\text{s.t.} \quad \mathbf{C}_{il} \geq 0,\ i = 1, ..., m,\ l = 1, ..., K, \tag{16}$$

According to the fact some of the primary cell types do not migrate to metastasis, it is possible that cell types included in single-cell samples in metastasis won't cover all the cell types in primary tumor, so we allow that primary tumor to have more cell types that might not be found in the metastasis by decomposing $\mathbf{C}$ to $\mathbf{C_1}$ representing cell types found in single-cell metastatic samples and an auxiliary matrix $\mathbf{C_2}$ representing the possible unknown cell types that are not present in single-cell metastatic samples (Table 2). Therefore $\mathbf{C} = [\mathbf{C_1}, \mathbf{C_2}]$ is the horizontal concatenation of $\mathbf{C_1}$ and $\mathbf{C_2}$:

$$\begin{bmatrix} c_{1,1} & \dots & c_{1,K} \\ . & \dots & . \\ . & \dots & . \\ . & \dots & . \\ c_{m,1} & \dots & c_{m,K} \end{bmatrix} = \left[\begin{array}{ccc|ccc} c_{1,1} & \dots & c_{1,k} & c_{1,k+1} & \dots & c_{1,K} \\ . & \dots & . & . & \dots & . \\ . & \dots & . & . & \dots & . \\ . & \dots & . & . & \dots & . \\ c_{m,1} & \dots & c_{m,k} & c_{m,k+1} & \dots & c_{m,K} \end{array}\right]$$

Then Eq. 15 can be rewritten as:

$$\min_{\mathbf{C_1}, \mathbf{C_2}} \quad \|\mathbf{B} - [\mathbf{C_1}, \mathbf{C_2}]\, \mathbf{F}\|_{\mathrm{Fr}}^2 + \lambda \|\mathbf{C_1} - \mu \mathbf{S}\|_{\mathrm{Fr}}^2, \tag{17}$$

$$\text{s.t.} \quad (\mathbf{C_1})_{il} \geq 0,\ i = 1, ..., m,\ l = 1, ..., k, \tag{18}$$

$$(\mathbf{C_2})_{ij} \geq 0,\ i = 1, ..., m,\ j = 1, ..., y, \tag{19}$$

We follow the idea of coordinate descent to optimize $\mathbf{C_1}$ and $\mathbf{C_2}$ in iterative steps:

Phase 2.1: Optimizing $\mathbf{C_1}$ Let $\mathbf{c_1} = \mathbf{C_1}_{i\cdot}^\intercal$, which represents the expression profile of one gene in each cell type/population. Similarly, we define $\mathbf{b} = \mathbf{B}_{i\cdot}^\intercal, \mathbf{s} = \mathbf{S}_{i\cdot}^\intercal$ to represent the transpose of one row of $\mathbf{B}$ and $\mathbf{S}$, respectively. Let $\mathbf{F} = \left[\mathbf{F}_1^\intercal, \mathbf{F}_2^\intercal\right]^\intercal$ be the vertical concatenation of $\mathbf{F}_1$ and $\mathbf{F}_2$. Then we can re-formulate Eq. 17 to a standard form for $\mathbf{C_1}$:

$$\min_{\mathbf{C_1}} \quad \|\mathbf{B} - [\mathbf{C_1}, \mathbf{C_2}]\,\mathbf{F}\|_{\mathrm{Fr}}^2 + \lambda\|\mathbf{C_1} - \mu\mathbf{S}\|_{\mathrm{Fr}}^2, \tag{20}$$

$$\iff \min_{\mathbf{C_1}} \quad \|\mathbf{B} - \mathbf{C_1}\mathbf{F}_1 - \mathbf{C_2}\mathbf{F}_2\|_{\mathrm{Fr}}^2 + \lambda\|\mathbf{C_1} - \mu\mathbf{S}\|_{\mathrm{Fr}}^2, \tag{21}$$

(let $\mathbf{B} = \mathbf{B} - \mathbf{C_2}\mathbf{F}_2$ for convenience)

$$\iff \min_{\mathbf{C_1}} \quad \|\mathbf{B} - \mathbf{C_1}\mathbf{F}_1\|_{\mathrm{Fr}}^2 + \lambda\|\mathbf{C_1} - \mu\mathbf{S}\|_{\mathrm{Fr}}^2, \tag{22}$$

$$\iff \min_{\mathbf{C_1}} \quad \left[\sum_{i=1}^{m} \left\|\mathbf{B}_{i\cdot} - (\mathbf{C_1})_{i\cdot}\,\mathbf{F}_1\right\|_2^2 + \lambda\left\|(\mathbf{C_1})_{i\cdot} - \mu\mathbf{S}_{i\cdot}\right\|_2^2\right], \tag{23}$$

$$\iff \min_{(\mathbf{C_1})_{i\cdot}} \quad \left\|\mathbf{B}_{i\cdot} - (\mathbf{C_1})_{i\cdot}\,\mathbf{F}_1\right\|_2^2 + \lambda\left\|(\mathbf{C_1})_{i\cdot} - \mu\mathbf{S}_{i\cdot}\right\|_2^2, \tag{24}$$

($\forall i = 1, ..., m$, we then omitted subscription i for clarity)

$$\iff \min_{\mathbf{c_1}^\intercal} \quad \left\|\mathbf{b}^\intercal - \mathbf{c_1}^\intercal\mathbf{F}_1\right\|_2^2 + \lambda\left\|\mathbf{c_1}^\intercal - \mu\mathbf{s}^\intercal\right\|_2^2, \tag{25}$$

$$\iff \min_{\mathbf{c_1}} \quad \frac{1}{2}\mathbf{c_1}^\intercal(\mathbf{F}_1\mathbf{F}_1^\intercal + \lambda\mathbf{I})\mathbf{c_1} - (\mathbf{F}_1\mathbf{b} + \lambda\mu\mathbf{s})^\intercal\mathbf{c_1} \tag{26}$$

with the same constraint as shown in Eq. 18, where $\mathbf{I}$ is an identity matrix, which implies $\mathbf{I}^\intercal = \mathbf{I}$. Let $\mathbf{Q} = \mathbf{F}_1\mathbf{F}_1^\intercal + \lambda\mathbf{I}$. We can show that $\mathbf{Q}^\intercal = (\mathbf{F}_1\mathbf{F}_1^\intercal + \lambda\mathbf{I})^\intercal = (\mathbf{F}_1\mathbf{F}_1^\intercal)^\intercal + \lambda\mathbf{I}^\intercal = \mathbf{F}_1\mathbf{F}_1^\intercal + \lambda\mathbf{I} = \mathbf{Q}$, so $\mathbf{Q}$ is a symmetric matrix, which indicates that Eq. 26 is a QP problem.

Phase 2.2: Optimizing $\mathbf{C_2}$ Let $\mathbf{c_2} = \mathbf{C_2}_{i\cdot}^\intercal$, which represents the expression profile of one gene in each cell type/population. Similarly, we define $\mathbf{b} = \mathbf{B}_{i\cdot}^\intercal$ to represent the transpose of one row of $\mathbf{B}$. Then we can re-formulate Eq. 17 to a standard form for $\mathbf{C_2}$:

$$\min_{\mathbf{C_2}} \quad \|\mathbf{B} - [\mathbf{C_1}, \mathbf{C_2}]\,\mathbf{F}\|_{\mathrm{Fr}}^2 + \lambda\|\mathbf{C_1} - \mu\mathbf{S}\|_{\mathrm{Fr}}^2, \tag{27}$$

$$\iff \min_{\mathbf{C_2}} \quad \|\mathbf{B} - \mathbf{C_1}\mathbf{F}_1 - \mathbf{C_2}\mathbf{F}_2\|_{\mathrm{Fr}}^2 + \lambda\|\mathbf{C_1} - \mu\mathbf{S}\|_{\mathrm{Fr}}^2, \tag{28}$$

(let $\mathbf{B} = \mathbf{B} - \mathbf{C_1}\mathbf{F}_1$ for convenience)

$$\iff \min_{\mathbf{C_2}} \quad \|\mathbf{B} - \mathbf{C_2}\mathbf{F}_2\|_{\mathrm{Fr}}^2, \tag{29}$$

$$\iff \min_{\mathbf{C_2}} \quad \left[\sum_{i=1}^{m} \left\|\mathbf{B}_{i\cdot} - (\mathbf{C_2})_{i\cdot}\,\mathbf{F}_2\right\|_2^2\right], \tag{30}$$

$$\iff \min_{(\mathbf{C_2})_{i\cdot}} \quad \left\|\mathbf{B}_{i\cdot} - (\mathbf{C_2})_{i\cdot}\,\mathbf{F}_2\right\|_2^2, \tag{31}$$

($\forall i = 1, ..., m$, we then omitted subscription i for clarity)

$$\iff \min_{\mathbf{c_2}^\intercal} \quad \left\|\mathbf{b}^\intercal - \mathbf{c_2}^\intercal\mathbf{F}_2\right\|_2^2, \tag{32}$$

$$\iff \min_{\mathbf{c_2}} \quad \frac{1}{2}\mathbf{c_2}^\intercal\mathbf{F}_2\mathbf{F}_2^\intercal\mathbf{c_2} - (\mathbf{F}_2\mathbf{b})^\intercal\mathbf{c_2} \tag{33}$$

We found that Eq. 33 has similar form as Eq. 14. It is also not hard to see that $(\mathbf{F}_2\mathbf{F}_2^\intercal)^\intercal = \mathbf{F}_2\mathbf{F}_2^\intercal$, so the term $\mathbf{F}_2\mathbf{F}_2^\intercal$ is symmetric, which also indicates that Eq. 33 is a Quadratic Programming problem with the same constraints of Eq. 19.

Phase 3: Optimizing μ At this phase, we fix $\mathbf{F}$ and $\mathbf{C_1}$ to optimize μ. Then Eq. 2 equals to:

$$\min_{\mu} \quad \|\mathbf{C_1} - \mu\mathbf{S}\|_{\mathrm{Fr}}^2, \tag{34}$$

Since μ is a scalar and $\mathbf{C_1}$ and $\mathbf{S}$ are two known matrices with the same dimension of $m \times l$. Then the Eq. 34 can be viewed as a Linear Regression problem of which the goal is best fit the model regarding the independent variable $\mathbf{S}$ to the data $\mathbf{C_1}$ by using the least-squares method to find the optimal value of μ that minimizes the residual sum of squares:

$$\min_{\mu} RSS(\mu) = \sum_{i}^{m}(\mathbf{C_1}_{i\cdot} - \mu\mathbf{S}_{i\cdot})^2 = \sum_{i=1}^{m}\sum_{l=1}^{k}(\mathbf{C_1}_{il} - \mu\mathbf{S}_{il})^2. \tag{35}$$

In summary, we have shown that the deconvolution problem can be divided into 3 main phases. In each phase, we have formulated the sub-problem to be either a QP problem or a Linear Regression problem. With third-party software (e.g., CVXOPT), we iteratively solve each phase by using coordinate descent algorithm until the convergence to get optimal values for $\mathbf{F}$, $\mathbf{C_1}$, $\mathbf{C_2}$ and μ.

2.5 Evaluation

We evaluated the performance of our method by comparing the inferred expression profiles $\hat{\mathbf{C}}$ and corresponding fractions $\hat{\mathbf{F}}$ with the ground truth $\mathbf{C}$ and $\mathbf{F}$ by utilizing four different metrics: R_C^2 (Pearson coefficient of $\hat{\mathbf{C}}$ and $\mathbf{C}$), L_1 loss ($\|\hat{\mathbf{C}} - \mathbf{C}\|_1/\|\mathbf{C}\|_1$), R_F^2 (Pearson coefficient of $\hat{\mathbf{F}}$ and $\mathbf{F}$) and MSE (Mean Square Error of $\hat{\mathbf{F}}$ and $\mathbf{F}$).

3 Results

3.1 Our deconvolution is unbiased and robust on simulated data

First, we tested our method on the simulated data generated as discussed in Sec. 2.1.1. We generated a single-cell matrix $\mathbf{C}$ with size of $m = 5000$ random genes and $k = 5$ known cell types as well as a signature matrix $\mathbf{S}$ from the current scRNA dataset. We also allow $y = 1$ unknown cell type that only exists in the bulk sample. Noise has been introduced to bulk and single-cell samples to mimic the relatively low resolution in bulk sequencing and individual-level differences in single-cell RNA-seq data, respectively.

When the number of samples increases, the average performance of pure RAD becomes slightly better while the variance is still large. We also find that in most cases, adding a signature matrix $\mathbf{S}$ further improves the inference by reducing the L_1 loss (Fig. 2, (a)). One thing needed to be mentioned is that in the current experiment, we used $\mathbf{S}$ as the initialization for RAD. This is equivalent to feeding prior information to RAD, which is usually not the case for RAD since it assumes no single-cell information available at all and adding such priors would put the search at a good start point for faster and better convergence in some cases.

The inference of $\mathbf{F}$ shows a similar pattern to that for $\mathbf{C}$, although they are not sensitive to small number of samples (Fig. 2 (b)). This is not surprising since $\mathbf{F}$ has more constraints and the search space of both is much smaller than $\mathbf{C}$. The noise was also introduced (b_noise and s_noise are not 0), we can find that cell component deconvolution with $\mathbf{S}$ outperforms pure RAD and that of cell component deconvolution without $\mathbf{S}$ in all cases as well, and it is robust to the noise (Fig. 2). Based on the results, we can conclude that adding information from $\mathbf{C}$ (e.g., using signature matrix $\mathbf{S}$) can help in bulk sequencing data deconvolution particularly with small numbers of bulk samples.

3.2 Comparison with other methods

In the section, we compared our method to two other popular deconvolution methods: DSA, a complete deconvolution method requiring a list of highly-expressed genes corresponding to each cell type; and CIBERSORTx, a semi-deconvolution method with reference gene expression profiles from other tissues. There are some restrictions for these methods, for example, DSA cannot work when there is only 1 bulk sample and CIBERSORTx requires the number of bulk sample to be no less than the number of cell

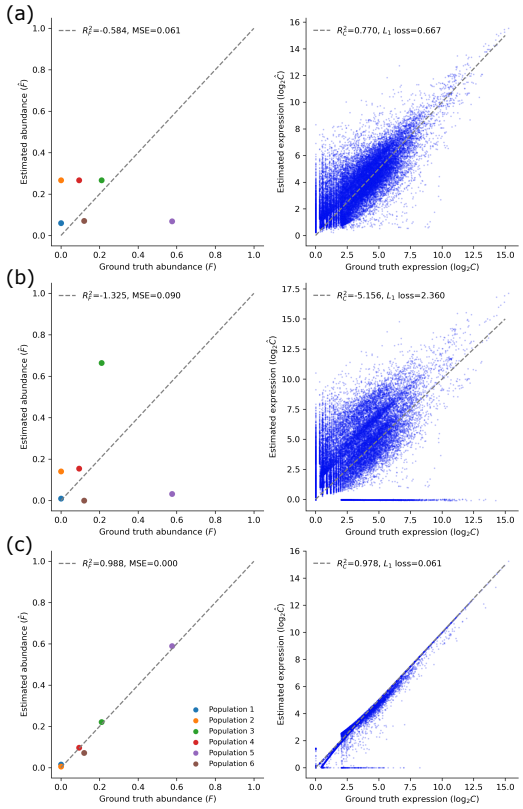

Fig. 2. Performance of deconvolution when bulk samples are limited. We show representative results of RAD, RADs without reference, RADs with reference for number of bulk sample $n = 1$ while number of cell type $K = 6$ (5 known cell types and 1 unknown cell types). We compared the performance both on estimated $\mathbf{C}$ and $\mathbf{F}$ using the evaluation metrics L_1 loss, R_C^2 for $\mathbf{C}$ and MSE, R_F^2 for $\mathbf{F}$, respectively. (a) RAD using $\mathbf{S}$ as initialization, (b) RADs without $\mathbf{S}$ ($\lambda = 0$) and (c) RADs with $\mathbf{S}$ ($\lambda = 0.1$)

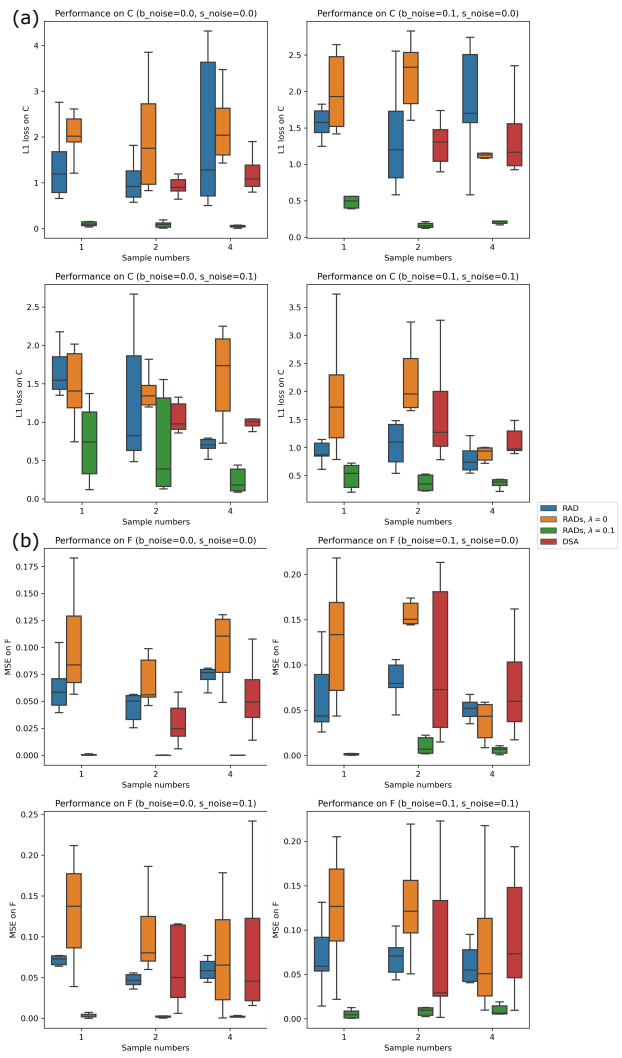

Fig. 3. Average performance of deconvolution in various noise (replicates=10). Average results of RAD, RADs without reference, RADs with reference for number of bulk samples $n = 1, 2, 4$ and number of cell types $K = 6$ (5 known cell types and 1 unknown cell types) with different levels of noise (b_noise=0.0, 0.1, s_noise=0.0, 0.1). Performance was calculated across 10 replicates, using L_1 loss and MSE for $\mathbf{C}$ and $\mathbf{F}$, respectively. (a) Performance on $\mathbf{C}$ inference, (b) Performance on $\mathbf{F}$ inference. Different boxes show RAD with $\mathbf{S}$ as initialization, RADs without $\mathbf{S}$, RADs with $\mathbf{S}$ and DSA (see Sec.3.2 for details, note that DSA cannot work then bulk sample $n = 1$), respectively.)

types plus 1. In addition, neither of these two methods is able to infer information regarding unknown cell types absent in the reference gene list or expression profiles.

In order to make the comparison reasonable, we ran two sets of experiments. In the first experiment, we ran DSA on small number of bulk samples (e.g., $n = 2, 4$). Note that DSA can only infer the cell types that exist in $\mathbf{S}$ so the L_1 loss and MSE were only calculated on the known cell types while our method includes both known and unknown cell types. We can find the DSA has similar performance as RAD and RADs without reference, which is worse than RADs with reference in cases without noise or with noise (Fig. 3, red boxes).

In the second experiment, we increased the number of simulated bulk samples to be 7, but varied the number of known and unknown cell types to be 5:1, 4:2 and 3:3, respectively. This setting is intended to align with requirement of CIBERSORTx and also allow us to investigate the effects of known information on the deconvolution performance. We find that DSA and CIBERSORTx perform worse than our method (Fig. 4 (d), different boxes). This might be due to the fact that DSA and CIBERSORTx can only infer the cell types in $\mathbf{S}$ and any unknown cell types then are included in the bulk sample profiles as noise rather than independent components in the deconvolution result, while our method considered both known and unknown cell types in the bulk samples and separated them in the deconvolution result. We also find that when there are more unknown cell types, our methods still works but the performance became worse (Fig. 4 (d), different X labels). This is not surprising since more unknown cell types means less useful information in the $\mathbf{S}$, which makes the penalty term in Eq. 2 less effective. This also indicates the importance of correct reference information in the bulk deconvolution.

3.3 Systematic changes in the breast cancer metastatic microenvironment

In this section, we applied our deconvolution method to the real data described in Sec. 2.1.2. We first retrieved genes that were deferentially expressed across different cell types in the single-cell RNA-Seq data. For each gene, its $\log_2$-scaled expression values in all cells annotated to be one cell type was compared to those in all other cells. A Wilcoxon rank-sum test was performed during the comparison to determine if the gene was significantly differentially expressed in the foreground cell type compared to the background. A Benjamini-Hochberg corrected p-value cutoff of

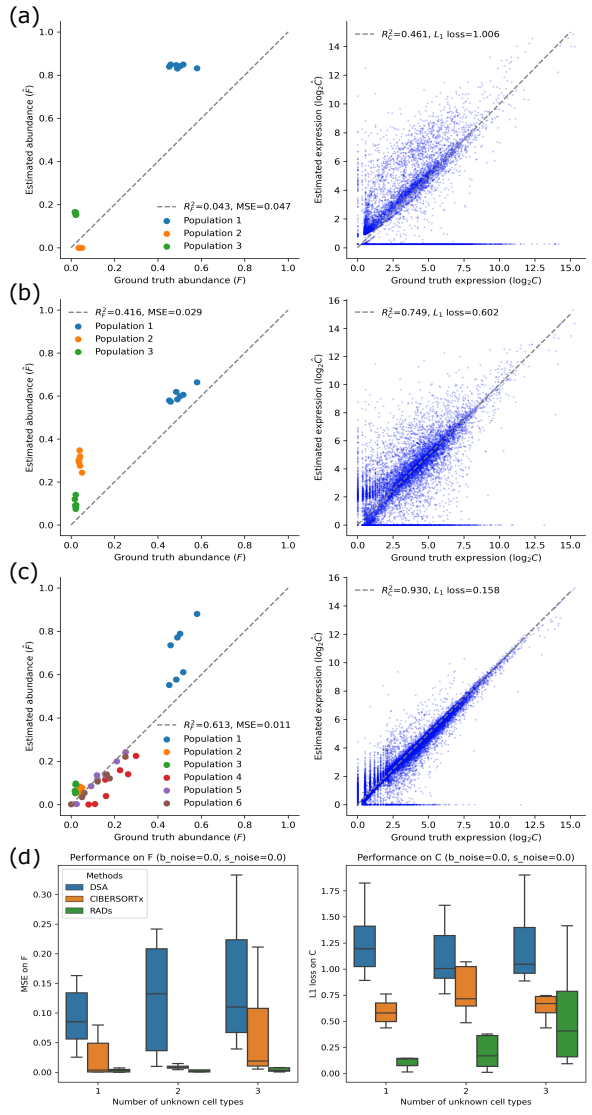

Fig. 4. Performance compared to DSA and CIBERSORTx (replicates=10). Deconvolution on 7 simulated bulk samples without noise using 6 cell types. Unknown cell type(s) were set to be 1, 2 and 3, respectively. (a)-(c) representative result for DSA, CIBERSORTx and RADs when there are 3 unknown cell types. (d) Average performance on $\mathbf{F}$ (left) and $\mathbf{S}$ (right) inference from the three methods. Metrics of DSA and CIBERSORTx were calculated only on the known cell types in the inferred results and ground truth while sRAD considered both known and unknown cell types.)

0.001 yielded a total of 1,226 genes differentially expressed across all six cell types annotated in the single-cell RNA-Seq. Then we checked with the bulk sample to make sure that same set of genes are available for both bulk and single-cell samples, which finally yielded 1,196 genes. The expression profiles of the 1,196 genes from the bulk sample were then retrieved to compose $\mathbf{B}$, and those from single-cell samples were retrieved to compose $\mathbf{C}_s$. The signature matrix $\mathbf{S}$ was inferred from $\mathbf{C}_s$ as described in Sec. 2.2. $\mathbf{B}$ and $\mathbf{S}$ are the input for Eq. (2). We also allow for 1 cell type ($\mathbf{C_2}$) to represent the cell type only found in primary tumor as well to balance the difference between bulk and single-cell sequencing data.

The inferred gene expression profiles from bulk samples were found to match very well with the signature matrix $\mathbf{S}$ from single-cell samples. This indicates that the penalty term in Eq. 2 works with non-zero regularization term (Fig. $5(a), \lambda = 0.1$) in real data. The heat-map showing the gene expression also yields distinct expression patterns for different cell types (Fig. 5 (b)). However, fractions of cell types exhibit different patterns between primary site and metastatic sites. For example, the fibroblasts cell type takes a low proportion in the primary site, but makes up for a large one in the metastatic sites. Although the only bulk sample available might have an unpresentive low fibroblast content at the specific location where the tissue was sampled (e.g., high fibroblasts may be needed for the tumor to attach to the bone), the large proportion of fibroblasts in the bone metastatsis sample is consistent with previous studies claiming that cancer-associated fibroblasts (CAFs) contribute to tumor growth, invasion and metastatsis (Kalluri and Zeisberg, 2006; Joshi *et al.*, 2021), which leads to cancer malignancy in later stages. The fraction of lymphocytes also exhibits an interesting pattern. The primary sample shows a higher fraction than the average fraction in bone metastatic samples (Fig. 5, red bar vs grey bar) although BoM1 includes more lymphocytes than the primary sample while BoM2 includes fewer. However, the average of BoM1 and BoM2 shows a lower fraction than the primary sample (Fig. 5, blue bar). We also find that there are almost no myeloid cells in the primary tumor (fraction=1.3×10^{-4}) but some in bone metastases. A closer examination found that among the single-cell samples labeled as myeloid cell, macrophages occupy a lot of proportion (77.4%, fine cell type annotation on the same dataset, data not shown). All these findings are consistent with prior work, which concluded that the metastatic breast cancers show reduced immune cells but increased macrophages compared to the primary tumors (Zhu *et al.*, 2019). We also found epithelial cells relatively preserved in both primary and metastatic tumor. This is consistent with breast cancer origin from non-diseased epithelial tissue and transition to metastasis (Nguyen *et al.*, 2018; Wang and Zhou, 2011). The fractions of osteoclast can be used as negative controls since they are bone-tissue related cell types that should be rare in primary sites, and indeed they exhibit fractions close to zero in the primary tumor. The unknown cell types, meaning inferred cell types not found in the single-cell data, account for over 30% in the primary tumor. Although this unknown type is close to endothelial (Fig. 5 (c)) based on the expression distance, we would not assign any biological meaning to it until we have better evidence (e.g., comparing to single-cell data from the same primary tumor or other reliable public data). At this point, we summarize it as a free component that accounts for the difference between primary and metastatic tumor composition.

We then explored the inferences from the perspective of variations in inferred single-cell gene expression and their interpretation with respect to Gene Ontology (GO) enrichment pathways. We first show that $\mathbf{S}$ extracted from real data is a good representative of the true single-cell data. It has low average distance to the expression profiles of each of the single-cell sample and shows high correlation with such gene expression profiles (Fig. 6 (a), (b)). Although $\mathbf{S}$ is a good representative of the available single-cell samples while still allowing us to infer a reasonable $\mathbf{C}$ at the cell-type level, some variations were observed between inferred $\mathbf{C}$ for primary tumor tissues and the single cell RNA-Seq data measured from each of the two bone metastatic tumors at the individual gene expression level. This may suggest potential changes in gene expression at the single-cell level along the metastatic trajectory of the tumor. The top one percent most down-regulated and up-regulated genes in each metastatic sample, measured by their distances to the inferred expression values for the primary tumor sample, were selected for downstream GO enrichment analysis (Fig. 6 (c)). The top up-regulated genes in both metastatic samples showed significant enrichment for the receptor binding of several chemokines that were found to have a direct impact on cancer cell metastasis, including CCR5 and CCR1 (Mollica Poeta *et al.*, 2019). This is consistent with previous reports that an increase in CCR5 activities leads to an increase in homing behavior to metastatic sites of breast cancer (Jiao *et al.*, 2021). Similarly, the knock-down of CCR1 was experimentally shown to inhibit metastasis of breast

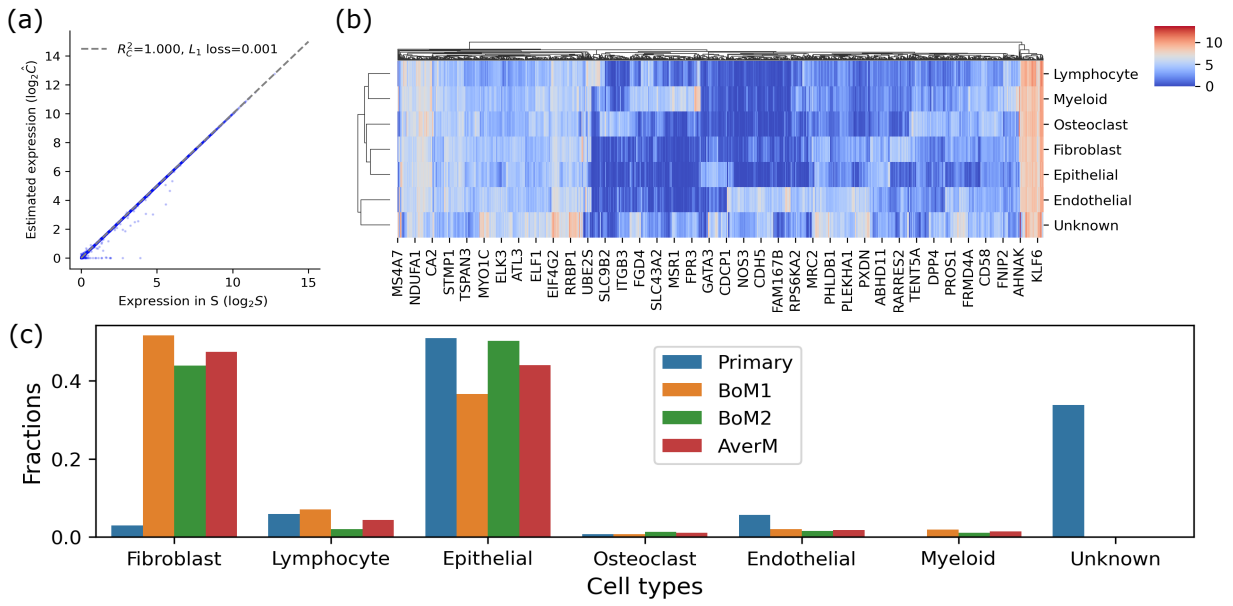

Fig. 5. Gene expression in 7 cell types and corresponding fractions in primary tumor and bone metastasis. (a) Comparison between inferred $\mathbf{C}$ and signature matrix $\mathbf{S}$, (b) gene expression profiles in inferred $\mathbf{C}$, (c) fractions of different cell types in primary tumor and two bone metastasis (BoM1 and BoM2). $AverM$ means the average fraction in BoM1 and BoM2.

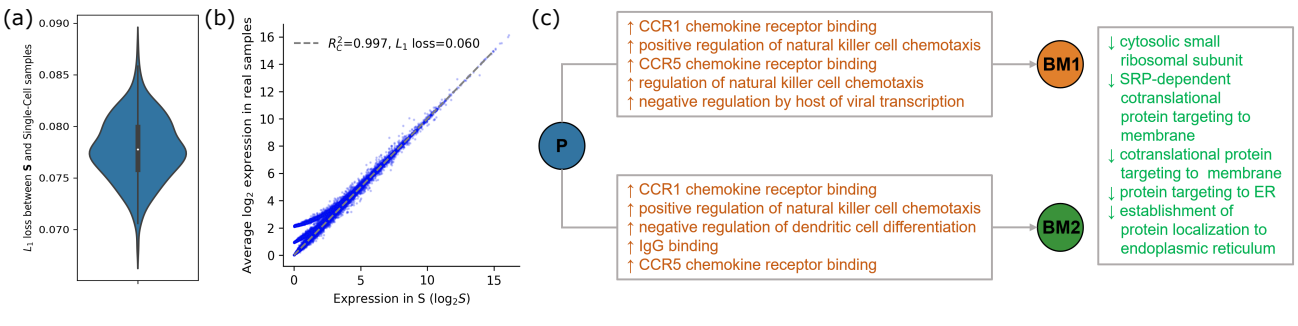

Fig. 6. Gene expression in $\mathbf{S}$ and real samples and the GO enrichment pathway in primary and metastases (a) Violin plot for L_1 loss of signature matrix $\mathbf{S}$ and single-cell samples, (b) correlation between $\mathbf{S}$ and average gene expression profiles ($\log_2$ scale) in single-cell samples, (c) Top GO enrichment pathway changes in primary (P) and two bone metastases (BM1 and BM2) with α = 0.05 for Holm-corrected p-values. Orange text indicates up-regulated pathways while green text indicates down-regulated pathways.

cancer (Shin *et al.*, 2017). The up-regulation in both metastatic samples could also have the effect of promoting chemotaxis of natural killer cells that may be recruited in response to increased activity of chemokines CCR1 and CCR5 (Aldinucci *et al.*, 2020). Additionally, an increase in expression of genes enriching the binding of IgG, such as is observed in BoM2, has been hypothesized to be a driver of breast tumor metastasis (Cui *et al.*, 2021).

4 Discussion

In this work, we have developed a novel tumor deconvolution method, RADs, designed for scenarios in which we have mixtures of bulk and single-cell data. We show that the method improves on standard bulk deconvolution or reference-based variants and works well even when there are limited numbers of bulk samples. Although our formulation of deconvolution using single-cell data from the same patient alleviates the bias caused by the single-cell reference signatures, the formulation still implicitly requires all potential populations be available in the single-cell data. The extension of $\mathbf{C}$ to unknown populations not available in single-cell data could further reduce bias of the deconvolution model. This heuristic idea not only has a biological meaning, such that some cell types in the primary tumor might not migrate to other organs or tissues, but also lends itself well to the coordinate descent algorithm. Combining the reference signature from single-cell while allowing unknown cell types in $\mathbf{C}$ helps to achieve robust performance without suffering too much from bias of the single-cell data. Comparing to other methods, our method directly takes advantage on the single-cell data from metastatic sites of the same patient, which is a more accurate reference than those from unrelated normal tissue or a panel of reference samples. In addition, our method can manage to work on a very limited number of bulk samples. We can still get reasonable results when the number of inferred cell types exceeds the number of bulk sample, which is usually a hard problem for deconvolution. The results on the real data align well with existing work on the breast cancer bone metastases as described in Sec. 3.3.

There are some limitations of our method, however. The method is still limited in its ability to infer new cell types with few bulk samples, making it difficult to characterize progression via novel evolution or complete loss of clones. Also, our method require matched bulk and single-cell

samples from the same patient, even if they can come from different sites and progression stages. While this is motivated by an important use case, the combination of data is still not common. Future work will explore how same-patient and third-party reference data may be synergistic in bypassing this limitation. Nevertheless, we believe our method builds a bridge between limited data and good deconvolution performance, and provides strategies for better leveraging heterogeneous data modalities that may have broader application in cancer research and other forms of single cell biology.

Acknowledgments

This work is partially supported by NIH awards R21CA216452 and R01HG010589, Pennsylvania Department of Health award #4100070287, Susan G. Komen for the Cure, the Mario Lemieux Foundation, and the Breast Cancer Alliance. It is also partially supported by the AWS Machine Learning Research Awards granted to J.M. and R.S., and by the Center for Machine Learning and Health Fellowship granted to Y.T. The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions.

Data Availability

Simulated data created for this study and additional results are provided with the RADs source code at https://github.com/CMUSchwartzLab/RADs The original tumor data on which the simulations are based is being released via the Gene Expression Omnibus (GEO) concurrent with the publication of the primary study from which it is derived.

References

Aldinucci, D. *et al.* (2020). The CCL5/CCR5 Axis in Cancer Progression. *Cancers (Basel)*, **12**(7).

Avila Cobos, F. *et al.* (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.*, **11**(1), 5650.

Beerenwinkel, N. *et al.* (2016). Computational cancer biology: an evolutionary perspective. *PLoS computational biology*, **12**(2), e1004717.

Chang, K. *et al.* (2013). The cancer genome atlas pan-cancer analysis project. *Nat Genet*, **45**(10), 1113–1120.

Cui, M. *et al.* (2021). Immunoglobulin Expression in Cancer Cells and Its Critical Roles in Tumorigenesis. *Front Immunol*, **12**, 613530.

Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, **481**(7381), 306–313.

Jiao, X. *et al.* (2021). Leronlimab, a humanized monoclonal antibody to CCR5, blocks breast cancer cellular metastasis and enhances cell death induced by DNA damaging chemotherapy. *Breast Cancer Res*, **23**(1), 11.

Joshi, R. S. *et al.* (2021). The role of cancer-associated fibroblasts in tumor progression. *Cancers*, **13**(6), 1399.

Kalluri, R. and Zeisberg, M. (2006). Fibroblasts in cancer. *Nature Reviews Cancer*, **6**(5), 392–401.

Kuipers, J. *et al.* (2017). Advances in understanding tumour evolution through single-cell sequencing. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, **1867**(2), 127–138.

Lei, H. *et al.* (2020). Tumor copy number deconvolution integrating bulk and single-cell sequencing data. *Journal of Computational Biology*, **27**(4), 565–598.

Lu, P. *et al.* (2003). Expression deconvolution: a reinterpretation of dna microarray data reveals dynamic changes in cell populations. *Proceedings of the National Academy of Sciences*, **100**(18), 10370–10375.

Malikic, S. *et al.* (2019a). Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature communications*, **10**(1), 1–12.

Malikic, S. *et al.* (2019b). Phiscs: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome research*, **29**(11), 1860–1877.

Mollica Poeta, V. *et al.* (2019). Chemokines and Chemokine Receptors: New Targets for Cancer Immunotherapy. *Front Immunol*, **10**, 379.

Naxerova, K. and Jain, R. K. (2015). Using tumour phylogenetics to identify the roots of metastasis in humans. *Nature reviews Clinical oncology*, **12**(5), 258–272.

Newman, A. M. *et al.* (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.*, **37**(7), 773–782.

Nguyen, Q. H. *et al.* (2018). Profiling human breast epithelial cells using single cell rna sequencing identifies cell diversity. *Nature communications*, **9**(1), 1–12.

Schwartz, R. and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, **18**(4), 213–229.

Schwartz, R. and Shackney, S. E. (2010). Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics*, **11**(1), 42.

Shin, S. Y. *et al.* (2017). C-C motif chemokine receptor 1 (CCR1) is a target of the EGF-AKT-mTOR-STAT3 signaling axis in breast cancer cells. *Oncotarget*, **8**(55), 94591–94605.

Sturm, G. *et al.* (2019). Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*, **35**(14), i436–i445.

Tao, Y. *et al.* (2020a). Neural network deconvolution method for resolving pathway-level progression of tumor clonal expression programs with application to breast cancer brain metastases. *Frontiers in Physiology*, **11**, 1055.

Tao, Y. *et al.* (2020b). Robust and accurate deconvolution of tumor populations uncovers evolutionary mechanisms of breast cancer metastasis. *Bioinformatics*, **36**, i407–i416.

Wang, Y. and Zhou, B. P. (2011). Epithelial-mesenchymal transition in breast cancer progression and metastasis. *Chinese journal of cancer*, **30**(9), 603.

Zaitsev, K. *et al.* (2019). Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nature Communications*, **10**(1), 2209.

Zhang, J. *et al.* (2011). International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database*, **2011**.

Zhong, Y. *et al.* (2013). Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, **14**(1), 89.

Zhu, L. *et al.* (2019). Metastatic breast cancers have reduced immune cell recruitment but harbor increased macrophages relative to their matched primary tumors. *Journal for ImmunoTherapy of Cancer*, **7**(1), 265.