frontiers in Physiology

ORIGINAL RESEARCH
published: 04 September 2020
doi: 10.3389/fphys.2020.01055

Neural Network Deconvolution Method for Resolving Pathway-Level Progression of Tumor Clonal Expression Programs With Application to Breast Cancer Brain Metastases

Yifeng Tao[1,2], *Haoyun Lei*[1,2], *Adrian V. Lee*[3], *Jian Ma*[1] and *Russell Schwartz*[1,4]*

[1] Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, United States, [2] Joint Carnegie Mellon–University of Pittsburgh Ph.D. Program in Computational Biology, Pittsburgh, PA, United States, [3] Department of Pharmacology and Chemical Biology, UPMC Hillman Cancer Center, Magee-Womens Research Institute, University of Pittsburgh, Pittsburgh, PA, United States, [4] Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, United States

OPEN ACCESS

Edited by:
Katharina Jahn,
ETH Zürich, Switzerland

Reviewed by:
Zhihui Wang,
Houston Methodist Research Institute, United States
Yoshitaka Kimura,
Tohoku University, Japan

***Correspondence:**
Russell Schwartz
russells@andrew.cmu.edu

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *31 January 2020*
Accepted: *31 July 2020*
Published: *04 September 2020*

Citation:
Tao Y, Lei H, Lee AV, Ma J and Schwartz R (2020) Neural Network Deconvolution Method for Resolving Pathway-Level Progression of Tumor Clonal Expression Programs With Application to Breast Cancer Brain Metastases. Front. Physiol. 11:1055. doi: 10.3389/fphys.2020.01055

Metastasis is the primary mechanism by which cancer results in mortality and there are currently no reliable treatment options once it occurs, making the metastatic process a critical target for new diagnostics and therapeutics. Treating metastasis before it appears is challenging, however, in part because metastases may be quite distinct genomically from the primary tumors from which they presumably emerged. Phylogenetic studies of cancer development have suggested that changes in tumor genomics over stages of progression often result from shifts in the abundance of clonal cellular populations, as late stages of progression may derive from or select for clonal populations rare in the primary tumor. The present study develops computational methods to infer clonal heterogeneity and dynamics across progression stages via deconvolution and clonal phylogeny reconstruction of pathway-level expression signatures in order to reconstruct how these processes might influence average changes in genomic signatures over progression. We show, via application to a study of gene expression in a collection of matched breast primary tumor and metastatic samples, that the method can infer coarse-grained substructure and stromal infiltration across the metastatic transition. The results suggest that genomic changes observed in metastasis, such as gain of the *ErbB* signaling pathway, are likely caused by early events in clonal evolution followed by expansion of minor clonal populations in metastasis, a finding that may have translational implications for early detection or prevention of metastasis[1].

Keywords: breast cancer, brain metastases, phylogenetics, deconvolution, pathways, gene modules, transcriptome, matrix factorization

[1]Algorithmic details, parameter settings, and source code are available at https://github.com/CMUSchwartzLab/NND. Additional results and proofs are provided in the **Supplementary Material**.

1. INTRODUCTION

Metastatic disease is the primary mechanism by which cancer results in patient mortality (Chambers et al., 2002; Chaffer and Weinberg, 2011). By the time metastases have appeared, there are generally no viable treatment options (Guan, 2015). Successful treatment thus depends on treating not just the primary tumor but also the seeds of metastasis that may linger after a seemingly successful remission. Identifying successful treatment options for metastasis is problematic, however, since the genomics of primary and metastatic tumors may be quite different even in single patients and metastatic cell populations may be poorly responsive to therapies effective on the primary tumor. Studies of cell-to-cell variation in cancers have revealed often substantial clonal heterogeneity in single tumors, with clonal populations sometimes dramatically shifting across progression stages (Greaves and Maley, 2012). Phylogenetic studies of clonal populations have been inconclusive on the typical evolutionary relationships between primary and metastatic tumors (Schwartz and Schäffer, 2017). It remains a matter of debate whether changes in clonal composition occur primarily through ongoing clonal evolution, which results in novel clones with metastatic potential and resistance to therapy, or from selection on existing clonal heterogeneity already present at the time of first treatment (Ding et al., 2013; de Bruin et al., 2014). The degree to which either answer is true has important implications for prospects for early detection or prophylactic treatment of metastasis.

Brain metastases (BrMs) occur in around 10–30% of metastatic breast cancers cases (Lin et al., 2004). Although recent advances in the treatment of metastatic breast cancer have been able to achieve long-term overall survival, there are limited treatment options for BrMs and clinical prognoses are still disappointing (Witzel et al., 2016). Recent work examining transcriptomic changes between paired primary and BrM samples has demonstrated dramatic changes in expression programs over metastasis, including changes in tumor subtype with important implications for treatment options and prognosis (Priedigkeit et al., 2017; Vareslija et al., 2018). Some past research has sought to infer phylogenetic models to explain the development of brain metastases based on somatic genomic alterations (Brastianos et al., 2015; Körber et al., 2019). Such methods are challenged in drawing robust conclusions about recurrent progression processes, though, by the high heterogeneity within single tumors and across progression stages and patients. While single-cell methods are proving powerful for resolving such problems in other contexts (Qiu et al., 2011; Elyanow et al., 2020), such data is rarely available for studies of metastatic progression, which generally require working with samples archived years before metastases are discovered. Changes in the activity of particular genetic pathways or modules may provide a more robust measure of frequent genomic alterations across cancers.

In the present work, we develop a strategy for tumor phylogenetics to explore how changes in clonal composition, via both novel molecular evolution and shifts in population dynamics of tumor clones and associated stroma, influence changes in expression programs across such progression stages. Our methods make use of multi-site bulk transcriptomic data to profile changes evident in gene expression programs between clones and progression stages. We break from past work in this domain in that we seek to study not clones *per se*, as is typical in tumor phylogenetics (Eaton et al., 2018; Tao et al., 2019b), but what we dub "cell communities": collections of clones or other stromal cell types that persist as a group with similar proportions across samples (section 2.4). We accomplish this via a novel transcriptomic deconvolution approach designed to make use of multiple samples both within and between patients (Schwartz and Shackney, 2010; Zare et al., 2014) while improving robustness to inter- and intra-tumor heterogeneity by integrating deconvolution with pathway-based analyses of expression variation (Park et al., 2009).

2. MATERIALS AND METHODS

2.1. Overview

Cell populations evolve due to genomic perturbations that can result in changes in the activity of various functional pathways between clones. Our overall method for deriving coarse-grained portraits of cell community evolution at the pathway level is illustrated by **Figure 1**. After the preprocessing of transcriptome data (section 2.2), the overall workflow consists of three main steps: First, the bulk expression profiles are mapped into the gene module and pathway space using external knowledge bases to reduce redundancy, noise, and sparsity, and to provide markers of expression variation for the subsequent analysis (section 2.3). Second, a deconvolution step is implemented to resolve cell communities, i.e., coarse-grained mixtures of cell types presumed to represent an associated population of cancer clones and stromal cells, from the compressed pathway representation of samples (section 2.4). Third, phylogenies of these cell communities are built based on the deconvolved communities as well as inferred ancestral (Steiner) communities to reconstruct likely trajectories of evolutionary progression by which cell communities develop—through a combination of genetic mutations, expression changes, and changes in population distributions—as a tumor progresses from healthy tissue to primary and potentially metastatic tumor (section 2.5).

2.2. Transcriptome Data Preprocessing

We applied our methods to raw bulk RNA-Sequencing data of 44 matched primary breast and metastatic brain tumors from 22 patients (each patient gives two samples) (Priedigkeit et al., 2017; Vareslija et al., 2018), where six patients were from the Royal College of Surgeons (RCS) and sixteen patients from the University of Pittsburgh (Pitt). These data profiled the expression levels of ∼60,000 transcripts. These can be represented in the format of a matrix, with rows corresponding to genes and columns to the samples (primary tumors or metastases). We removed the genes that are not expressed in any sample. We also considered only protein-coding genes in the present study. Approximately 20,000 genes remain after the filter. We conducted quantile normalization across samples using the geometric mean to remove possible artifacts (Amaratunga and Cabrera, 2001). The top 2.5% and bottom

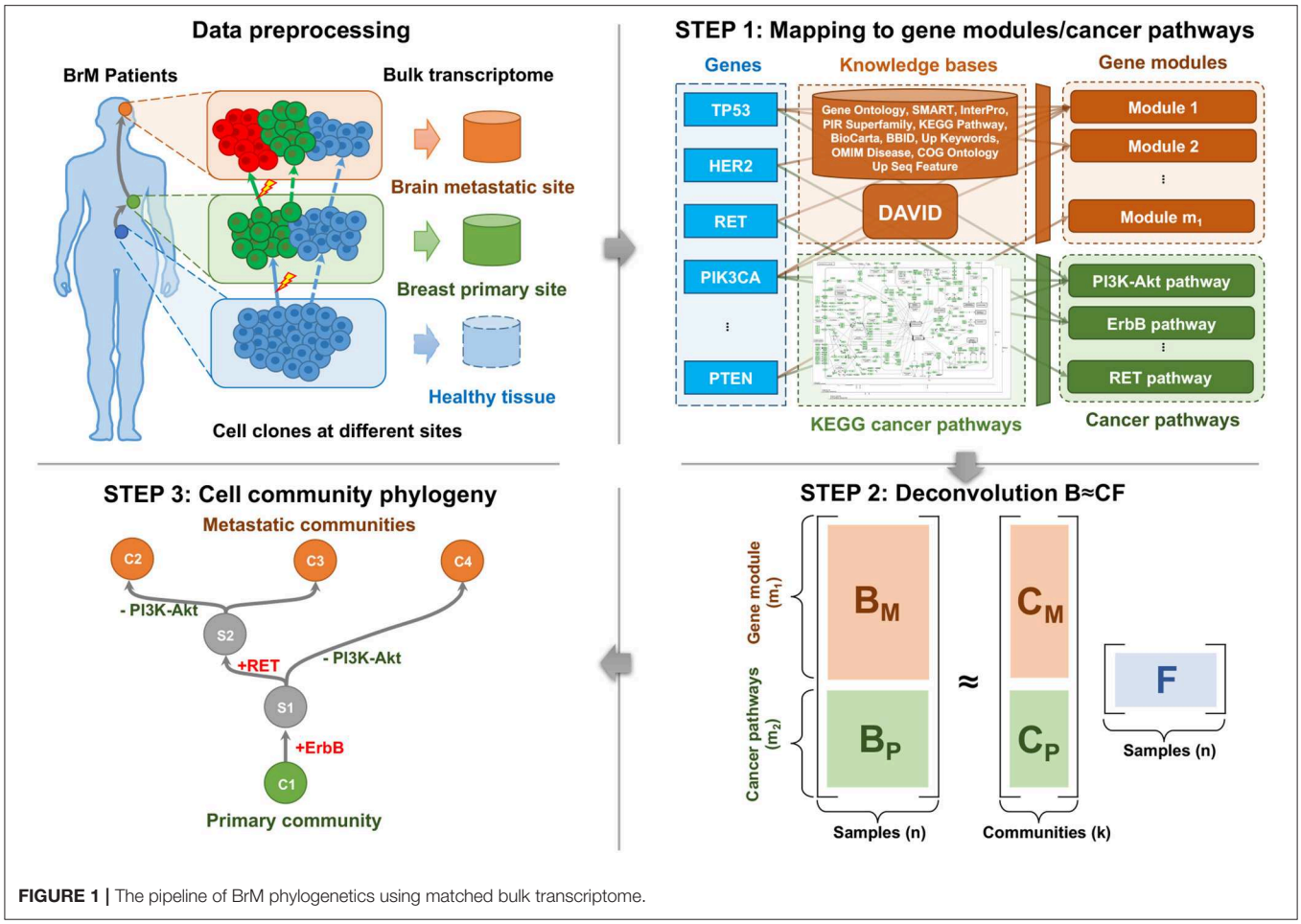

FIGURE 1 | The pipeline of BrM phylogenetics using matched bulk transcriptome.

2.5% of expressions were clipped to further reduce noise. Finally, we transformed the resulting bulk gene expression values into the log space and mapped those for each gene to the interval $[0, 1]$ by a linear transformation. The resulting preprocessed transcriptome data were used as the input of Step 1 (section 2.3).

2.3. Mapping to Gene Modules and Cancer Pathways

The protein-coding gene expressions were mapped into both perturbed gene modules and cancer pathways, using the DAVID tool and external knowledge bases (Huang et al., 2009), as well as the cancer pathways in the KEGG database (Kanehisa and Goto, 2000). This step compresses the high dimensional data and provides markers of cancer-related biological processes (**Figure 1**, Step 1). Note that although both gene module and cancer pathway representations capture recurrent features of metastatic progression, they serve different purposes in our analysis. Gene modules are an essential part of deconvolution in the following steps because they provide the major variance within the data. Cancer pathways serve primarily as probes for *post-hoc* interpretation of the unmixed communities, but are biased relative to the gene module space by the focus only on genes with known relevance to cancer.

2.3.1. Gene Modules

Functionally similar genes are usually affected by a common set of somatic alterations (Park et al., 2009) and therefore are co-expressed in the cells. These genes are believed to belong to the same "gene modules" (Desmedt et al., 2008; Tao et al., 2020). Inspired by the idea of gene modules, we fed a subset of 3,000 most informative genes out of the ∼20,000 genes that have the largest variances into the DAVID tool for functional annotation clustering using several databases (Huang et al., 2009). DAVID maps each gene to one or more modules. We did not force the genes to be mapped into disjunct modules because a gene may be involved in several biological functions and therefore more than one gene module. We removed gene modules that were not enriched (fold enrichment < 1.0) and kept the remaining $m_1 = 109$ modules (and the corresponding annotated functions), where fold enrichment is defined as the EASE score of the current module to the geometric mean of EASE scores in all modules (Hosack et al., 2003). The gene module values of all the $n = 44$ samples were represented as a gene module matrix $\mathbf{B}_M \in \mathbb{R}^{m_1 \times n}$. The i-th gene module value in j-th sample, $(\mathbf{B}_M)_{i,j}$, was calculated by taking the sum of expressions of all the genes in the i-th module. Then $\mathbf{B}_M$ was rescaled row-wise by taking the z-scores across samples to compensate for the effect of variable module sizes.

2.3.2. Cancer Pathways

Although the gene module representation is able to capture the variances across samples and reduce the redundancy of raw gene expressions, it has two disadvantages. The first is a lack of interpretability. Specifically, some annotations assigned by DAVID are not directly related to biological functions, and the annotations of different modules may substantially overlap. The second is that the key perturbed cancer pathways or functions may not always be the ones that vary most across samples. For example, genes in cancer-related KEGG pathways (hsa05200; Kanehisa and Goto, 2000) are not especially enriched in the top 3,000 genes with the largest expression variances. To make better use of prior knowledge on cancer-relevant pathways, we supplemented the generic DAVID pathway sets with a KEGG "cancer pathway" representation of samples $\mathbf{B}_P \in \mathbb{R}^{m_2 \times n}$, where the number of cancer pathways $m_2 = 24$. The cancer-related pathways in the KEGG database are cleaner and easier to explain, more orthogonal to each other, and contain critical signaling pathways to cancer development. We extracted the 23 cancer-related pathways from the following 3 KEGG pathway sets: *Pathways in cancer* (hsa05200), *Breast cancer* (hsa05224), and *Glioma* (hsa05214). An additional cancer pathway *RET pathway* was added, since it was found to be recurrently gained in the prior research (Vareslija et al., 2018). See *y*-axis of **Figure 4D** for the complete list of 24 cancer pathways. We considered all the ∼20,000 protein-coding genes other than top 3,000 genes. The following mapping of cancer pathways and transformation to *z*-scores were similar to that we did to map the gene modules.

Until this step, the raw gene expressions of n samples were transformed into the compressed gene module/pathway representation of samples $\mathbf{B} = \left[\mathbf{B}_M^\intercal, \mathbf{B}_P^\intercal\right]^\intercal \in \mathbb{R}^{m \times n}$, where $m = m_1 + m_2$. The gene module representation $\mathbf{B}_M$ serves for accurately deconvolving and unmixing the cell communities, while the pathway representation $\mathbf{B}_P$ serves as markers/probes and for interpretation purpose.

2.4. Deconvolution of Bulk Data

We applied a type of matrix factorization (MF) with constraints on the pathway-level expression signatures to deconvolve the communities/populations from primary and metastatic tumor samples (**Figure 1**, Step 2) (Koren et al., 2009). Note that common alternatives, such as principal components analysis (PCA) and non-negative matrix factorization (NMF) are not amenable to this case (Lee and Seung, 2000), since PCA does not provide a feasible solution to the constrained problem, and the NMF does not apply to our mixture data, which can be either positive or negative.

2.4.1. Cell Communities

We define a cell community to be a set of clones/clonal subpopulations and other cell types that propagate as a group during the evolution of a tumor. A community may be just a single subpopulation/clone, but is a more general concept in the sense that it usually involves multiple related clones and their associated stroma. For example, a set of immunogenic clones and the immune cells infiltrating them might collectively form a community that has a collective expression signature mixing signatures of the clones and associated immune cells, even if the individual cell types are not distinguishable from bulk expression data alone. While much work in this space has classically aimed to separate individual clones, or perhaps individual cell types more broadly defined, we note that deconvolution may be unable in principle to resolve distinct cell types if they are always co-located in similar proportions. It is particularly true when data is sparse and cell types are fit only approximately, as in the present work, that a model with large complexity to deconvolve the fine-grained populations is prone to overfit. The community concept is intended in part to better describe the results we expect to achieve from the kind of data examined here and in part because identifying these communities is itself of interest in understanding how tumor cells coevolve with their stroma during progression and metastasis. Single-cell methods may provide an alternative, but are not amenable to preserved samples, such as are needed when retrospectively studying primary tumors and metastases that may have been biopsied years apart.

2.4.2. Formulation of Deconvolution

With a matrix of bulk pathway values $\mathbf{B} \in \mathbb{R}^{m \times n}$, the deconvolution problem is to find a component matrix $\mathbf{C} = \left[\mathbf{C}_M^\intercal, \mathbf{C}_P^\intercal\right]^\intercal \in \mathbb{R}^{m \times k}$ that represents the inferred fundamental communities of tumors, and the corresponding set of mixture fractions $\mathbf{F} \in \mathbb{R}_+^{k \times n}$:

$$\min_{\mathbf{C},\mathbf{F}} \quad \|\mathbf{B} - \mathbf{C}\mathbf{F}\|_{\mathrm{Fr}}^2, \tag{1}$$

$$\text{s.t.} \quad \mathbf{F}_{lj} \geq 0, \qquad l = 1, ..., k,\ j = 1, ..., n, \tag{2}$$

$$\sum_{l=1}^{k} \mathbf{F}_{lj} = 1, \qquad j = 1, ..., n, \tag{3}$$

where $||\mathbf{X}||_{\mathrm{Fr}}$ is the Frobenius norm. The column-wise normalization in Equation (3) aims for recovering the biologically meaningful cell communities. In addition, they are equivalent to applying ℓ_1 regularizers and therefore enforce sparsity to the fraction matrix $\mathbf{F}$ (**Supplementary Material**).

2.4.3. Neural Network Deconvolution

Although it is possible to build new algorithms for solving MF by adapting previous work (Lee and Seung, 2000), the additional but necessary constraints of Equations (2) and (3) make the optimization much harder to solve. For the problem of Equations (1)–(3), one can prove that it does not generally guarantee convexity (**Supplementary Material**). A slightly modified version of the algorithm to solve NMF with constraints may guarantee neither good fitting nor convergence (Lei et al., 2019, 2020). Therefore, instead of revising existing MF algorithms, such as ALS-FunkSVD (Funk, 2006; Bell and Koren, 2007; Koren et al., 2009), we developed an algorithm which we call "neural network deconvolution" (NND) to solve the optimization problem using gradient descent. Specifically, the NND was implemented using backpropagation in the form of a neural network (**Figure 2A**) with the PyTorch package (https://pytorch.org/) (Rumelhart et al., 1986; Kingma and Ba,

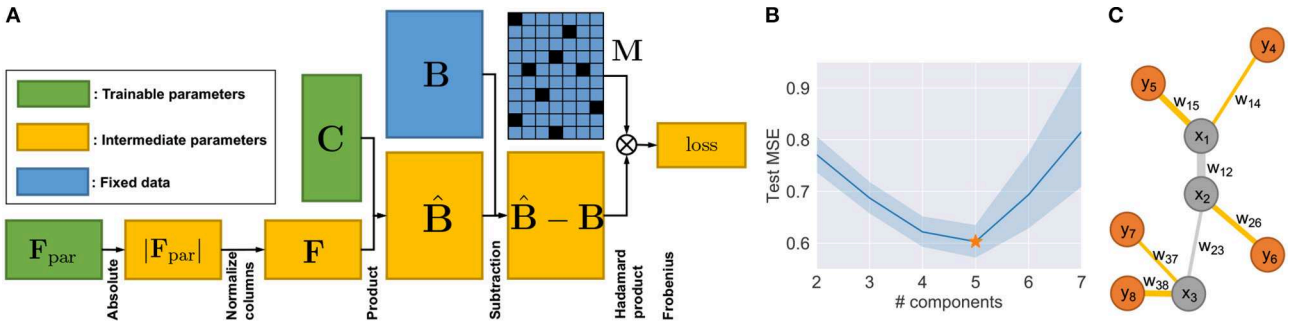

FIGURE 2 | Method details. **(A)** Neural network architecture of NND. **(B)** Test errors of NND using 20-fold CV. Errors in unit of mean square error (MSE). **(C)** Illustration of a phylogeny with five extant nodes and three Steiner nodes.

2014), based on the revised constraints:

$$\min_{\mathbf{C},\mathbf{F}_{\text{par}}} \quad \|\mathbf{B} - \mathbf{CF}\|_{\text{Fr}}^2, \tag{4}$$

$$\text{s.t.} \quad \mathbf{F} = \text{cwn}\left(|\mathbf{F}_{\text{par}}|\right), \tag{5}$$

where $|\mathbf{X}|$ applies element-wise absolute value and $\text{cwn}(\mathbf{X})$ is column-wise normalization, so that each column sums up to 1. The two operations of Equation (5) naturally rephrase and remove the two constraints in Equations (2) and (3), and meanwhile fit the framework of neural networks. An alternative to the absolute value operation $|\mathbf{X}|$ might be rectified linear unit $\text{ReLU}(\mathbf{X}) = \max(\mathbf{0}, \mathbf{X})$. However, this activation function is unstable and leads to inferior performance in our case, since $\mathbf{X}_{lj}$ will be fixed to zero once it becomes negative and will lose the chance to get updated in the following iterations. One may also want to replace the column-wise normalization $\text{cwn}(\mathbf{X})$ with softmax operation $\text{softmax}(\mathbf{X})$. However, the non-linearity introduced by softmax actually changes the original optimization problem (Equations 1–3) and the fitted $\mathbf{F}$ is therefore not sparse.

Based on the revised NND optimization problem (Equations 4 and 5), we built the neural network with the architecture shown in **Figure 2A**. An Adam optimizer other than vanilla gradient descent was used with default momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate of 1×10^{-5} (Kingma and Ba, 2014). The mini-batch technique is not required since the data size in our application is small enough not to require it ($\mathbf{B} \in \mathbb{R}^{m \times n}$, $m = 133$, $n = 44$). The training is run until convergence, which is defined as when the relative decrease of training loss is smaller than $\epsilon = 1 \times 10^{-10}$ every 20,000 iterations. This implementation has two main advantages: First, the method can be easily adapted to a wide range of optimization scenarios with various constraints, when existing methods do not or are hard to apply. Second, the NND has the flexibility of allowing for cross-validation, which is important for us in choosing the number of components k and preventing overfitting.

One might be suspicious whether the neural network fits precisely in practice, since it is based on a simple gradient descent optimization. To validate the fitting ability of NND, we plotted the PCA of original samples $\mathbf{B}$ and the fitted samples $\hat{\mathbf{B}}$ (**Supplementary Material**). One can easily see that NND provides a good fit to the data.

2.4.4. Cross-Validation of NND

In order to find the best tradeoff between model complexity and overfitting, we used cross-validation (CV) with the "masking" method to choose the optimal number of components/communities $k = 5$ that has the smallest test error (**Figure 2B**). In each fold of the CV, we used estimated $\hat{\mathbf{B}}$ to only fit some randomly selected elements of $\mathbf{B}$, and then the test error was calculated using the other elements of $\mathbf{B}$. This was implemented by introducing two additional mask matrices $\mathbf{M}_{\text{train}}, \mathbf{M}_{\text{test}} \in \{0,1\}^{m \times n}$, which are in the same shape of $\mathbf{B}$, and $\mathbf{M}_{\text{train}} + \mathbf{M}_{\text{test}} = 1^{m \times n}$. During the training time, with the same constraints in Equation (5), the optimization goal is:

$$\min_{\mathbf{C},\mathbf{F}_{\text{par}}} \|\mathbf{M}_{\text{train}} \odot (\mathbf{B} - \mathbf{CF})\|_{\text{Fr}}^2, \tag{6}$$

where $\mathbf{X} \odot \mathbf{Y}$ is the Hadamard (element-wise) product. At the time of evaluation, given optimized $\hat{\mathbf{C}}$, $\hat{\mathbf{F}}_{\text{par}}$, and therefore optimized $\hat{\mathbf{F}} = \text{cwn}\left(\left|\hat{\mathbf{F}}_{\text{par}}\right|\right)$ for the optimization problem during training, the test error was calculated on the test set: $\left\|\mathbf{M}_{\text{test}} \odot \left(\mathbf{B} - \hat{\mathbf{C}}\hat{\mathbf{F}}\right)\right\|_{\text{Fr}}^2$. We used 20-fold cross-validation on the NND, so in each fold 95% of positions of $\mathbf{M}_{\text{train}}$ and 5% of positions of $\mathbf{M}_{\text{test}}$ were 1s. Note that the actual number of cell populations is probably considerably larger than 5, and therefore each one of the five communities may contain multiple cell populations. Furthermore, it is likely that with sufficient numbers and precision of measurements, these communities could be more finely resolved into their constituent cell types. However $k = 5$ represents the largest hypothesis space of NND model that can be applied to the current dataset without severe overfitting.

2.5. Phylogeny of Inferred Cell Subcommunities and Pathway Inference of Steiner Nodes

We built "phylogenies" of cell subcommunities and estimated the pathway representation of unobserved (Steiner) nodes (Lu et al., 2003) inferred to be ancestral to them, with the goal of discovering critical communities that appear to be involved in the transition to metastasis and identifying the important changes of functions and expression pathways during this transition

(**Figure 1**, Step 3). Note that we are using the term "phylogeny" loosely here, as these trees are intended to capture evolution of populations of cells not just by accumulation of mutations from a single ancestral clone but also via changes in community structure, for example, due to generating or suppressing an immune response or migrating to a metastatic site. Although an abuse of terminology, we use the term phylogeny here due to the methodological similarity to more proper phylogenetic methods in wide use for analyzing mutational data in cancers (Schwartz and Schäffer, 2017).

2.5.1. Phylogeny of Communities

Given the pathway profiles of the extant communities at the time of collecting tumor samples $\mathbf{C} \in \mathbb{R}^{m\times k}$, a phylogeny of the k extant cell communities was built using the neighbor-joining (NJ) algorithm (Nei and Saitou, 1987), which inferred a tree that contains k extant nodes/leaves, $k-2$ unobserved Steiner nodes, and edges connecting two Steiner nodes or a Steiner node and an extant node. We estimated an evolutionary distance for any pair of two communities u, v as the input of NJ using the Euclidean distance between their pathway vectors $\|\mathbf{C}_{\cdot u} - \mathbf{C}_{\cdot v}\|_2$, similar to that in a prior work (Park et al., 2009).

2.5.2. Inference of Pathways: Setting and Approach

Denote the phylogeny of cell subcommunities as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and $\mathcal{V} = \mathcal{V}_S \cup \mathcal{V}_C$, where the indices of Steiner node $\mathcal{V}_S = \{1, 2, ..., k-2\}$ ($|\mathcal{V}_S| = k-2$), the indices of extant nodes $\mathcal{V}_C = \{k-1, k, ..., 2k-2\}$ ($|\mathcal{V}_C| = k$). For each edge $(u, v) \in \mathcal{E}$, where $1 \le u < v \le 2k-2$, the first node of edge $u \le k-2$ is always a Steiner node. The second node v can be either a Steiner node ($v \le k-2$) or extant node ($v \ge k-1$). Denote the set of weights $\mathcal{W} = \{w_{uv} = 1/d_{uv} \mid (u, v) \in \mathcal{E}\}$ (inverse distance), where the edge length d_{uv} is the output of NJ. For each dimension i of the pathway vectors, we consider them independently and separately, so that each dimension of the Steiner nodes can be solved in the same way. Now let us consider the i-th dimension (and omit the subscript i for brevity) of extant nodes $\mathcal{V}_C$: $\mathbf{y} = \left[y_{k-1}, y_k, ..., y_{2k-2}\right]^\intercal = \mathbf{C}_{i\cdot}^\intercal \in \mathbb{R}^k$ and Steiner nodes $\mathcal{V}_S$: $\mathbf{x} = \left[x_1, x_2, ..., x_{k-2}\right]^\intercal \in \mathbb{R}^{k-2}$. **Figure 2C** illustrates a phylogeny where $k = 5$. The inference of the i-th element in the pathway vector of the Steiner nodes can be formulated as minimizing the following elastic potential energy $U(\mathbf{x}, \mathbf{y}; \mathcal{W})$:

$$\min_{\mathbf{x}} \quad U(\mathbf{x}, \mathbf{y};\ \mathcal{W}) = \sum_{\substack{(u,v)\in\mathcal{E}\\ v\le k-2}} \frac{1}{2} w_{uv}(x_u - x_v)^2 + \sum_{\substack{(u,v)\in\mathcal{E}\\ v\ge k-1}} \frac{1}{2} w_{uv}(x_u - y_v)^2, \tag{7}$$

which can be rephrased as a quadratic programming problem and solved easily, as we show below.

2.5.3. Inference of Pathways: Derivation of Quadratic Programming, P($\mathcal{W}$), and q($\mathcal{W}$, y)

Theorem 1. *Equation (7) can be further rephrased as a quadratic programming problem:*

$$\min_{\mathbf{x}} \quad \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(\mathcal{W})\mathbf{x} + \mathbf{q}(\mathcal{W}, \mathbf{y})^\intercal \mathbf{x}, \tag{8}$$

where $\mathbf{P}(\mathcal{W})$ *is a function that takes as input edge weights* $\mathcal{W}$ *and outputs a matrix* $\mathbf{P} \in \mathbb{R}^{(k-2)\times(k-2)}$, $\mathbf{q}(\mathcal{W}, \mathbf{y})$ *is a function that takes as input edge weights* $\mathcal{W}$ *and vector* $\mathbf{y}$ *and outputs a vector* $\mathbf{q} \in \mathbb{R}^{k-2}$.

Proof: Based on Equation (7), $U(\mathbf{x}, \mathbf{y}; \mathcal{W}) \ge 0$. Each term inside the first summation ($v \le k-2$) can be written as:

$$\frac{1}{2} w_{uv}(x_u - x_v)^2 = \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x}, \tag{9}$$

where

$$\mathbf{P}(w_{uv}) = \begin{array}{c} \\ u\text{-th row} \\ \\ v\text{-th row} \\ \end{array} \begin{array}{c} \begin{array}{ccccc} & & u\text{-th col} & & v\text{-th col} \end{array} \\ \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & w_{uv} & \mathbf{0} & -w_{uv} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -w_{uv} & \mathbf{0} & w_{uv} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \end{array}. \tag{10}$$

Each term ($v \ge k-1$) inside the second summation can be rephrased as:

$$\frac{1}{2} w_{uv}(x_u - y_v)^2 = \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x} + \mathbf{q}(w_{uv}, y_v)^\intercal \mathbf{x} + C(w_{uv}, y_v), \tag{11}$$

where

$$\mathbf{P}(w_{uv}) = \begin{array}{c} \\ u\text{-th row} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} & & u\text{-th col} & & \end{array} \\ \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & w_{uv} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \end{array},$$

$$\mathbf{q}(w_{uv}, y_v) = \begin{array}{c} \\ u\text{-th row} \\ \\ \\ \\ \end{array} \begin{array}{c} \\ \begin{bmatrix} \mathbf{0} \\ -w_{uv}y_v \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \end{array}, \tag{12}$$

and $C(w_{uv}, y_v) = \frac{1}{2} w_{uv} y_v^2$ is independent of $\mathbf{x}$. Therefore the optimization in Equation (7) can be calculated and written

as below:

$$\min_{\mathbf{x}} \sum_{\substack{(u,v)\in\mathcal{E}\\ v\le k-2}} \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x} + \sum_{\substack{(u,v)\in\mathcal{E}\\ v\ge k-1}} \left(\frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x} + \mathbf{q}(w_{uv}, y_v)^\intercal \mathbf{x}\right), \quad (13)$$

$$\Leftrightarrow \min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^\intercal \left(\sum_{\substack{(u,v)\in\mathcal{E}\\ v\le k-2}} \mathbf{P}(w_{uv}) + \sum_{\substack{(u,v)\in\mathcal{E}\\ v\ge k-1}} \mathbf{P}(w_{uv})\right)\mathbf{x} + \sum_{\substack{(u,v)\in\mathcal{E}\\ v\ge k-1}} \mathbf{q}(w_{uv}, y_v)^\intercal \mathbf{x}, \quad (14)$$

$$\Leftrightarrow \min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(\mathcal{W})\mathbf{x} + \mathbf{q}(\mathcal{W}, \mathbf{y})^\intercal \mathbf{x}. \quad \square \quad (15)$$

Remark 1. *The optimal* $\mathbf{x}^\star$ *of the Equation (7), or the solution to the quadratic programming problem Equation (8) can be solved by setting the gradient to be* $\mathbf{0}$*:*

$$\mathbf{P}(\mathcal{W})\mathbf{x}^\star + \mathbf{q}(\mathcal{W}, \mathbf{y}) = \mathbf{0}. \quad (16)$$

Therefore,

$$\mathbf{x}^\star = -\mathbf{P}(\mathcal{W})^{-1}\mathbf{q}(\mathcal{W}, \mathbf{y}). \quad (17)$$

Remark 2. *Based on the proof, we can derive how to calculate the matrix* $\mathbf{P}(\mathcal{W})$ *and vector* $\mathbf{q}(\mathcal{W}, \mathbf{y})$*.*

Initialize the matrix and vector with zeros:

$$\mathbf{P} \leftarrow 0^{(k-2)\times(k-2)}, \quad \mathbf{q} \leftarrow 0^{k-2}. \quad (18)$$

For each edge $(u, v) \in \mathcal{E}$ *with weight* w_{uv}*, there are two possibilities of nodes u and v: First, if both of them are Steiner nodes* ($u \le k-2$, $v \le k-2$)*, we update* $\mathbf{P}$ *and keep* $\mathbf{q}$ *the same:*

$$\mathbf{P}_{uu} \leftarrow \mathbf{P}_{uu} + w_{uv}, \mathbf{P}_{vv} \leftarrow \mathbf{P}_{vv} + w_{uv}, \quad \mathbf{P}_{uv} \leftarrow \mathbf{P}_{uv} - w_{uv}, \ \mathbf{P}_{vu} \leftarrow \mathbf{P}_{vu} - w_{uv}. \quad (19)$$

Second, if u is Steiner node and v is an extant node ($u \le k-2$, $v \ge k-1$)*, we update both* $\mathbf{P}$ *and* $\mathbf{q}$*:*

$$\mathbf{P}_{uu} \leftarrow \mathbf{P}_{uu} + w_{uv}, \quad \mathbf{q}_u \leftarrow \mathbf{q}_u - y_v \cdot w_{uv}. \quad (20)$$

We apply the same procedure to all dimension of pathways $i = 1, 2, ..., m$ *to get the full pathway values for each Steiner node.*

3. RESULTS

3.1. NND Deconvolves the Bulk RNA Accurately

Before we applied our deconvolution algorithm NND to the breast cancer brain metastatic samples, we first validated our algorithm on a semi-simulated dataset where the ground truth expressions and fractions of each cell clone in the mixture samples are known.

3.1.1. Semi-simulated GSE11103 Dataset

The semi-simulated dataset is based on the real data of pure clones from the GSE11103 dataset (Abbas et al., 2009; Barrett et al., 2013). Expression profiles of four different cells were measured using microarrays: Raji (B cell), IM-9 (B cell), THP-1 (monocyte), Jurkat (T cell). Each experiment was repeated three times. We took the average of the three replicates to get the expression data of the four pure cell clones. The top 300 genes that varied most across cell types were selected as the ground truth real data of pure cell clones: $\mathbf{C} \in \mathbb{R}_+^{300\times 4}$. We then created 100 mixture samples of the four pure clones *in silico* $\mathbf{B} \in \mathbb{R}_+^{300\times 100}$ by randomly generating the fraction matrix $\mathbf{F} \in \mathbb{R}_+^{4\times 100}$. The fraction matrix was generated in the following way:

$$\mathbf{F}_{lj} \leftarrow U(0,1), \qquad l = 1, ..., 4, \ j = 1, ..., 100, \quad (21)$$

$$\mathbf{F}_{lj} \leftarrow \frac{\mathbf{F}_{lj}}{\sum_{l'=1}^{4} \mathbf{F}_{l'j}}, \qquad j = 1, ..., 100, \quad (22)$$

where $U(0, 1)$ is a uniform distribution in the interval $[0, 1]$. The semi-simulated bulk expression matrix $\mathbf{B}$ was then generated from $\mathbf{C}$, $\mathbf{F}$, with a log-normal noise:

$$(\mathbf{B})_{ij} = (\mathbf{CF})_{ij} + 2^{\mathcal{N}(0,(s\sigma)^2)}, \quad i = 1, ..., 300, \ j = 1, ..., 100, \quad (23)$$

where $\mathcal{N}\left(0, (s\sigma)^2\right)$ is a Gaussian distribution; s controls the noise level, which we set to 0, 0.4, 0.9, and 1.3 for test; σ is the standard deviation of $\log_2$-transformed original GSE11103 data.

3.1.2. Performance Evaluation

Given the bulk matrix $\mathbf{B}$, we applied NND and other two algorithms to infer the estimated $\hat{\mathbf{C}}$, $\hat{\mathbf{F}}$ and $\hat{\mathbf{B}} = \hat{\mathbf{C}}\hat{\mathbf{F}}$, and compared the accuracy between estimated and actual values using the following metrics. For $\mathbf{C}$, we used L_1 loss (Zhu et al., 2018):

$$L_1 \text{ loss}(\mathbf{C}) = \frac{\|\hat{\mathbf{C}} - \mathbf{C}\|_1}{\|\mathbf{C}\|_1}. \quad (24)$$

For $\mathbf{F}$ and $\mathbf{B}$, we used root mean square error (RMSE):

$$\text{RMSE}(\mathbf{F}) = \sqrt{\|\hat{\mathbf{F}} - \mathbf{F}\|_{\text{Fr}}^2}, \quad (25)$$

$$\text{RMSE}(\mathbf{B}) = \sqrt{\frac{\|\hat{\mathbf{B}} - \mathbf{B}\|_{\text{Fr}}^2}{\|\mathbf{B}\|_{\text{Fr}}^2}} \quad (26)$$

Different levels of noise s were added to test the robustness of models and the performance of different models under different conditions. We repeated all the experiments for 10 times to get the boxplot.

3.1.3. Competing Algorithms

There are two competing algorithms for the deconvolution problem. Geometric unmixing is an algorithm that borrows the intuition from computational geometry (Schwartz and Shackney, 2010), which first identifies the corners of a simplex containing all the mixture sample points, and then infers the fraction matrix.

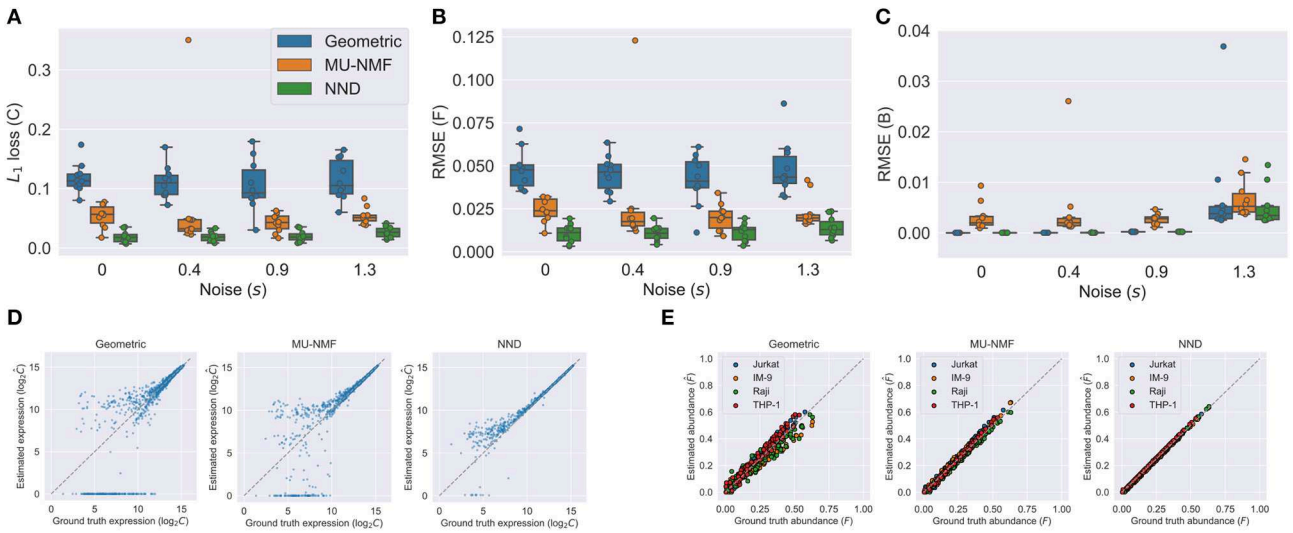

FIGURE 3 | Comparison of NND and other algorithms on the semi-simulated GSE11103 dataset shows the better accuracy and robustness of NND algorithm. **(A–C)** Accuracies in estimated **C**, **F**, and **B** with three different algorithms. We tested four noise levels and repeated the experiments for ten replicates. **(D)** Estimated $\hat{\mathbf{C}}$ and ground truth **C** using three different algorithms with noise level *s* of 1.3. **(E)** Estimated $\hat{\mathbf{F}}$ and ground truth **F** using three algorithms with noise level *s* of 1.3.

However, the algorithm does not directly optimize the problem (Equations 1–3). Another intuitive algorithm is based on the popular multiplicative update (MU) rule that solves general NMF problem (Lee and Seung, 2000): an additional update step of $\mathbf{F}_{lj} \leftarrow \frac{\mathbf{F}_{lj}}{\sum_{l'=1}^{k} \mathbf{F}_{l'j}}, j = 1, ..., n$ can be added to the loop. Although the original MU rule guarantees the non-increasing of the objective function, this additional update step can lead to an increasing objective and we need to stop the iteration once this happened. Since the two competing algorithms work on non-negative space, we adapted the NND by adding an element-wise absolute value operator after the **C** in the network (**Figure 2A**).

3.1.4. Superiority of NND

We show the results in **Figure 3**. **Figures 3A–C** show the accuracies of both **C**, **F**, and **B** using the three algorithms under various noise levels. One can easily see that NND achieves lower L_1 loss of **C**, RMSE of **F**, and RMSE of **B**. What is more, it is also much more robust than the geometric and MU-NMF algorithms, as there are fewer outliers that have huge errors. MU-NMF has a reasonable estimation accuracy of **C** and **F**. However, its overall fitting ability is limited due to its non-convergence-guaranteed MU optimization algorithm. We can also visualize the estimation accuracy by plotting the estimated values and ground truth values at a specific noise level, as is shown in **Figures 3D,E**. One can see the superiority of NND qualitatively over the other two algorithms in estimating expression profiles and fractions of individual pure clones.

3.2. Gene Modules/Pathways Provide an Effective Representation

Gene expressions of samples were mapped into the gene module and pathway space in order to reduce the noise of raw transcriptome data and reduce redundancy (section 2.3). We verified that the gene module/pathway representation is effective in the sense that it captures distinguishing features of primary/metastatic sites and individual samples well and is able to identify recurrently gained or lost pathways.

3.2.1. Feature Space of the Gene Module and Pathway Representation

As one can see in **Figure 4A**, the first principal component analysis (PCA) dimension of the gene module and pathway representation accounts for the difference between primary and metastatic samples, while the second and third PCA dimensions mainly capture variability between patients. This observation suggests the feasibility of using the gene module/pathway representation to distinguish recurrent features of metastatic progression across patients despite heterogeneity between patients. To make a direct comparison of the noise and redundancy between the gene module/pathway and raw gene expression representations, we applied hierarchical clustering to the 44 samples using Ward's minimum variance method (Ward, 1963). Two hierarchical trees were built based on the two different representations (**Figure 4B**). The gene module/pathway features more effectively separate the primary and metastatic samples into distinct clusters (**Figure 4B**, right panel) than do the raw gene expression values (**Figure 4B**, left panel). This is consistent with the PCA results that the largest mode of variance in the pathway representation distinguishes primary from metastatic samples. We do notice that in a few cases, matched primary and metastatic samples from the same patient are neighbors with pathway-based clustering. For example, 29P_Pitt:29M_Pitt and 51P_Pitt:51M_Pitt are grouped in the same clades using the pathway representation, showing that in a minority of cases, features of individual patients dominate

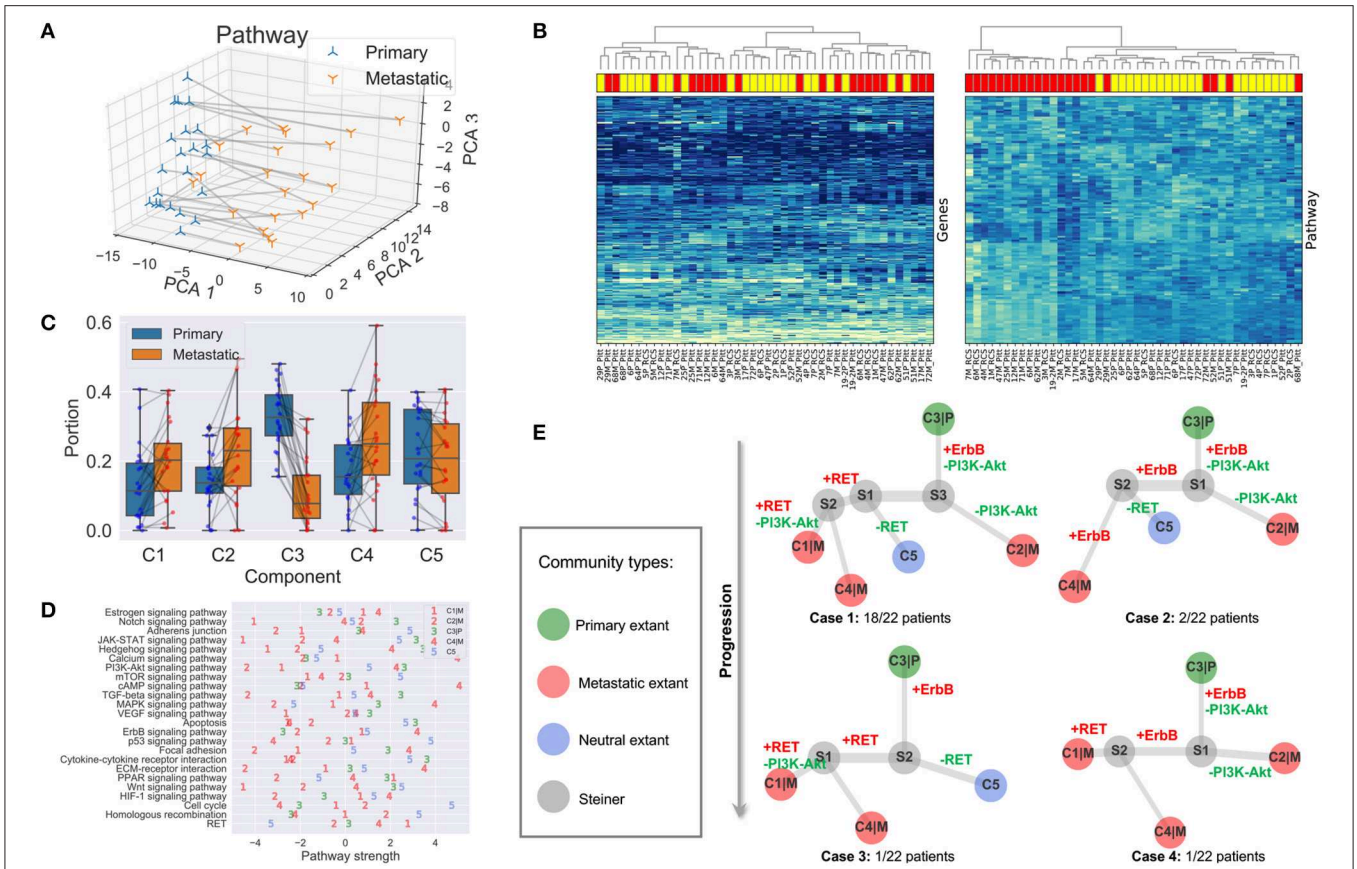

FIGURE 4 | Results and analysis. **(A)** First three gene module/pathway representation PCA dimensions of matched primary and metastatic samples. Matched samples are connected. **(B)** Hierarchical clustering of tumor samples based on raw gene expressions (left panel) and compressed gene module/pathway representation (right panel). Metastatic samples are shown in red rectangles and primary ones in yellow. **(C)** Portions and changes of the five communities in primary and metastatic sites. Each gray line connects the portions of a community in the primary site (blue node) and metastatic site (red node) from the same patient. **(D)** Pathway strengths across cell communities. **(E)** Phylogeny of cell subcommunities.

over primary vs. metastatic features. Following previous work (Park et al., 2009), we quantified the ability of the hierarchical tree to group the samples of the same labels using four metrics. (1) MSD: Mean square distance of edges that connect nodes of the same label (primary vs. metastatic). (2) z_{MSD}: The labels of all nodes were shuffled and the MSD is recalculated for 1,000 times to get the mean μ_{MSD} and standard deviation σ_{MSD}, which were used to get the z-score of the current assignment $z_{\text{MSD}} = (\text{MSD} - \mu_{\text{MSD}})/\sigma_{\text{MSD}}$. (3) rMSD: The ratio of MSD of edges that connect same label nodes and MSD of edges that connect distinct label nodes. (4) z_{rMSD}: as with MSD, a z-score of rMSD was calculated by shuffling labels for 1,000 times. Intuitively, the smaller values the MSD, z_{MSD}, rMSD, and z_{rMSD} are, the better is the feature representation at grouping same label samples together. The shortest paths and distances between all pairs of nodes were calculated using the Floyd-Warshall algorithm (Floyd, 1962; Warshall, 1962). All the edge lengths were considered as 1.0 to account for the different scales of pathway and gene representations. The pathway representation has significantly lower values for all four metrics (**Table 1**), indicating its strong grouping ability.

TABLE 1 | Quantitative performance of hierarchical clustering.

Feature representation	MSD	rMSD	z_{MSD}	z_{rMSD}
Gene expression	99.62	0.93	−2.60	−2.57
Gene module/pathway	86.23	0.66	−13.37	−11.42

3.2.2. Recurrently Perturbed Cancer Pathways

We next identified differentially expressed pathways in the primary and metastatic tumors using bulk data $\mathbf{B}_P \in \mathbb{R}^{24\times 44}$, prior to deconvolving cellular subcommunities. We conducted the Student's t-test followed by FDR correction on each of the 24 pathways. Eleven pathways are significantly different between the two sites (FDR < 0.05; **Table 2**). The signaling pathways related to neurotransmitter and calcium homeostasis, including *cAMP* and *Calcium* (Hofer and Lefkimmiatis, 2007), are enriched in metastatic samples, which we can suggest may reflect stromal contamination by neural cells in the brain metastatic samples. We also observed recurrent gains in *ErbB* pathway, as indicated by the primary studies (Priedigkeit et al., 2017; Vareslija et al.,

TABLE 2 | Differentially expressed cancer pathways between primary and metastatic samples (FDR < 0.05).

Gain/Loss after metastasis	Differentially expressed pathways	FDR
Relative gain	cAMP signaling pathway	6.88e-03
Relative gain	ErbB signaling pathway	2.09e-02
Relative gain	Calcium signaling pathway	4.39e-02
Relative loss	Cytokine-cytokine receptor interaction	4.37e-06
Relative loss	Apoptosis	8.53e-04
Relative loss	JAK-STAT signaling pathway	8.53e-04
Relative loss	Wnt signaling pathway	3.97e-03
Relative loss	Hedgehog signaling pathway	4.50e-03
Relative loss	PI3K-Akt signaling pathway	1.35e-02
Relative loss	TGF-beta signaling pathway	4.56e-02
Relative loss	Notch signaling pathway	4.56e-02

2018). Three pathways related to immune activity are under-expressed in metastatic samples, including *Cytokine-cytokine receptor interaction* (Lee and Margolin, 2011), *JAK-STAT* (Lee and Margolin, 2011), and *Notch* (Aster et al., 2017), consistent with the previous inference of reduced immune cell expression in metastases in general and brain metastasis most prominently (Zhu et al., 2019). We can suggest that this result similarly may reflect expression changes in infiltrating immune cells, due to the immunologically privileged environment of the brain, rather than expression changes in tumor cell populations. Five other signaling pathways, including *Apoptosis* (Wong, 2011), *Wnt* (Zhan et al., 2016), *Hedgehog* (Gupta et al., 2010), *PI3K-Akt* (Brastianos et al., 2015), and *TGF-beta* (Massagué, 2008), show reduction in metastatic samples and in each case, their loss or dysregulation has been reported to promote the tumor growth and brain metastasis. Note that the primary references for these data define pathways using the co-expression pattern of genes (Priedigkeit et al., 2017; Vareslija et al., 2018), while our work uses external knowledge bases. Previous research also used somatic mutations or copy number variation to analyze perturbed genes (Brastianos et al., 2015; Priedigkeit et al., 2017), while we focus exclusively on the transcriptome. Despite large differences in data types and pathway definitions, our observations are consistent with the prior analysis, especially with respect to variation in the *HER2/ErbB2* and *PI3K-Akt* pathways.

3.3. Landscape of Deconvolved Cell Communities in Tumors

We unmixed the bulk data **B** into five components using NND (section 2.4). The deconvolution enables us to produce at least a coarse-grained landscape of major cell communities **C** and their distributions in primary and metastatic tumors **F**. The number of components ($k = 5$) was chosen through 20-fold cross-validation (section 2.4; **Figure 2B**). Although the true heterogeneity of the samples may be much larger, we fit k to provide a balance between excessively coarse-grained communities if k is too small vs. excessively high variance and thus unstable deconvolution if k is too large.

3.3.1. Community Distributions Across Samples F

The portions of the five components in all the 44 samples are represented as the mixture fraction matrix $\mathbf{F} \in \mathbb{R}^{5\times 44}$ (**Figure 4C**). A primary or metastatic community is one inferred to change proportions substantially (magnitude > 0.05) in the tumor samples after metastasis, or perhaps to be entirely novel to or extinct in the metastatic sample (denoted by a $|P$ or $|M$ suffix). Otherwise, the component is classified as a neutral community. Three components ($C1|M$, $C2|M$, $C4|M$) are classified as metastatic communities; one ($C3|P$) as primary; and one ($C5$) as neutral (**Figure 4C**). Some components may be missing in both samples of some patients, e.g., $C1|M$, $C2|M$, $C5|M$ are absent in two, one, and one patient. We note that these five communities represent rough consensus clusters of cell populations inferred to occur frequently, but not universally, among the samples. Based on this rule, we can define four basic cases of patients in total. Twelve subcases can be found using a more detailed classification method based on the existence of communities in both primary and metastatic samples (**Supplementary Material**).

3.3.2. Pathway Values of Communities C

We are especially interested in the pathway part $\mathbf{C}_P$ of the cell community inferences, since it serves as the marker and provides results easier to interpret. The pathway values of five subcommunities using $\mathbf{C}_P$ provides a much more fine-grained description of samples (**Figure 4D**), compared with that in section 3.2, which is only able to distinguish the differentially expressed pathways in bulk samples. As noted in section 2.4, it is likely that true cellular heterogeneity is greater than the methods are able to discriminate and that communities inferred by our model may each conflate one or more distinct cell types and clones. We observe that the metastatic community $C4|M$ most prominently contributes to the enrichment for functions related to neurotransmitter and ion transport, since its strongest pathways (*cAMP*, *Calcium*) are greatly enriched relative to those of the other four communities. We might interpret this community as reflecting at least in part stromal contamination from neural cells specific to the metastatic site. $C4|M$ also contributes most to the gains of *ErbB* in brain samples. The metastatic subcommunity $C1|M$ is probably most closely related to the loss of immune response in metastatic samples as it has the lowest pathway values of *Notch*, *JAK-STAT*, and *Cytokine-cytokine receptor interaction*. This component might thus in part reflect the effect of relatively greater immune infiltration in the primary vs. the metastatic site. $C1|M$ also has the lowest pathway values of *Apoptosis*, *Wnt*, and *Hedgehog*. The metastatic community $C2|M$ is most responsible for the loss of *PI3K-Akt* and *TGF-beta* pathways. We also note that although *RET* does not show up in the list of **Table 2**, it seems to be quite over-expressed in the metastatic communities $C1|M$ and $C4|M$ but not in the metastatic community $C2|M$.

3.4. Phylogenies of BrM Communities Reveal Common Order of Perturbed Pathways

We built phylogenies of cell communities and calculated the pathway representations of their Steiner nodes (section 2.5). The phylogenies' topologies provide a way to infer a likely evolutionary history of cancer cell communities and thus their constitutive cell types. At the same time, the perturbed pathways along their edges suggest the order of genomic alterations or changes in community composition.

3.4.1. Topologically Similar BrM Phylogenies

All five cell components do not appear in each BrM patient. We analyze the distribution of communities in each patient based on whether the community is inferred to be present in the patient (**Supplementary Material**). There are four different cases in general (**Figure 4E**). Case 1: all five communities are found in the patient (majority; 18/22 patients). Case 2: only $C1|M$ missing (minority; 2/22). Case 3: only $C2|M$ missing (minority; 1/22). Case 4: only $C5$ missing (minority; 1/22). Although not all communities exist in Cases 2–4, the topologies are similar to that of Case 1 and can be seen as special cases of Case 1, representing some inferred common mechanisms of progression across all the BrM patients.

3.4.2. Common Order of Altered Cancer Pathways

After inferring the pathway values for Steiner nodes, the most perturbed pathways can also be found by subtracting the pathway vectors of nodes that share an edge. We focus on the top five gained or lost pathways along the evolutionary trajectories and the changes of magnitude larger than 1.0 (**Supplementary Material**). We further examine those perturbed cancer pathways that were specifically proposed in the study that generated the data examined here, as well as others that are clinically actionable (Brastianos et al., 2015; Priedigkeit et al., 2017; Vareslija et al., 2018), i.e., *ErbB*, *PI3K-Akt*, and *RET* (**Figure 4E**). As one may see from Case 1, the primary community $C3|P$ first evolves to community $S3$ by gaining expressions in *ErbB* and losing functions in *PI3K-Akt*. Then, if it continues to lose *PI3K-Akt* activity, it will evolve into the metastatic community $C2|M$. If it gains in *RET* activity, it will instead evolve into metastatic communities $C1|M$ and $C4|M$. The perturbed pathways along the trajectories of Cases 2–4 are similar to those of Case 1, with minor differences. We therefore draw to the conclusion that the evolution of BrMs follows a specific and common order of pathway perturbations. Specifically, the gain of *ErbB* reproducibly happens before the loss of *PI3K-Akt* and the gain of *RET*. Different subsequently perturbed pathways lead to different metastatic tumor cell communities. These inferences are consistent with the hypothesis that at least some major changes in expression programs between primary and metastatic communities occur by selecting for heterogeneity present early in tumor development rather than solely deriving from novel functional changes immediately prior to or after metastasis.

4. DISCUSSION

Cancer metastasis is usually a precursor to mortality with no successful treatment options. Better understanding mechanisms of metastasis provides a potential pathway to identify new diagnostics or therapeutic targets that might catch metastasis before it ensues, treat it prophylactically, or provide more effective treatment options once it occurs. The present work developed a computational approach intended to better reconstruct mechanisms of functional adaption from multisite RNA-Seq data to help us understand at the level of cancer pathways the mechanisms by which progression frequently proceeds across a patient cohort. Our method compresses expression data into a gene module/pathway representation using external knowledge bases, deconvolves the bulk data into putative cell communities where each community contains a set of associated cell types or subclones, and builds evolutionary trees of inferred communities with the goal of reconstructing how these communities evolve, adapt, and reconfigure their compositions across metastatic progression. Results on semi-simulated data show the method to yield improved accuracy in mixture deconvolution relative to prior deconvolution algorithms. We applied the pipeline to matched transcriptome data from 22 BrM patients and found that although there are slight differences of tumor communities across the cohort, most patients share a similar mechanism of tumor evolution at the pathway level. Specifically, the methods infer a fairly conserved mechanism of early gain of *ErbB* prior to metastasis, followed post-metastasis gain of *RET* or loss of *PI3K-Akt* resulting in intertumor heterogeneity between samples. Our methods provide a novel way of viewing the development of BrM with implications for basic research into metastatic processes and potential translational applications in finding markers or drug targets of metastasis-producing clones prior to the metastatic transition.

The results suggest several possible avenues for future development. In part, they suggest a need for better separating phylogenetically-related mixture components (i.e., distinct tumor cell clones) from unrelated infiltrating cell types (e.g., healthy stroma from the primary or metastatic site or infiltrating immune cells). The methods are likely finding only a small fraction of the true clonal heterogeneity of the tumors and stroma, and might benefit from algorithms capable of better resolution or from integration of multi-omics data (e.g., RNA-Seq, DNA-Seq, methylation) that might have complementary value in finer discrimination of cell types. The present methods are also using only a limited form of temporal constraint in considering a two-stage progression process and without use of quantitative time measurements. Models might be extended in future work to consider true time-series data, such as is becoming available through "liquid biopsy" technologies. In addition, we know of no data with known ground truth that models the kind of progression process studied here nor of other tools designed for modeling similar progression processes from expression data, leaving us reliant on validating based on consistency with prior research on brain metastasis (Brastianos et al., 2015; Priedigkeit et al., 2017; Vareslija et al., 2018). Future

work might compare to prior approaches for reconstruction of clonal evolution from expression data more generically (Desper et al., 2004; Riester et al., 2010; Schwartz and Shackney, 2010) and seek replication on additional real or simulated expression data or artificial mixtures of different cell types (Qiu et al., 2011) designed to mimic metastasis-like progression. The general approach might also have broader application than studying metastasis, for example in reconstructing mechanisms of other progression processes, such as pre-cancerous to cancerous, as well as to other tumor types or independent data sets. Finally, much remains to be done to exploit the translational potential of the method in better identifying diagnostic signatures and therapeutic targets, and what type of effective and safe clinical strategies can be taken to prevent metastasis at an early stage.

DATA AVAILABILITY STATEMENT

The breast cancer brain metastases dataset analyzed for this study can be found on Github: https://github.com/lizhu06/TILsComparison_PBTvsMET. The simulated dataset generated, and data for analysis for this study can be found on Github: https://github.com/CMUSchwartzLab/NND.

AUTHOR CONTRIBUTIONS

RS, JM, and AL contributed to the conceptualization. RS, JM, AL, YT, and HL contributed to the methodology. YT contributed to the software. RS, JM, AL, and YT contributed to the formal analysis, writing-review and editing, and funding acquisition. AL contributed to the resources. YT and HL contributed to the writing of the original draft. RS supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by a grant from the Mario Lemieux Foundation, U.S. N.I.H. awards R21CA216452 and R01HG010589, Pennsylvania Department of Health award 4100070287, Breast Cancer Alliance, Susan G. Komen for the Cure, and by a fellowship to YT from the Center for Machine Learning and Healthcare at Carnegie Mellon University. It was also supported in part by the AWS Machine Learning Research Awards. The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions.

ACKNOWLEDGMENTS

An earlier version of this work was published in the International Symposium on Mathematical and Computational Oncology 2019 (Tao et al., 2019a). We would like to thank to the reviewers for their helpful suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2020.01055/full#supplementary-material

REFERENCES

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., and Clark, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE* 4:e6098. doi: 10.1371/journal.pone.0006098

Amaratunga, D., and Cabrera, J. (2001). Analysis of data from viral DNA microchips. *J. Am. Stat. Assoc.* 96, 1161–1170. doi: 10.1198/016214501753381814

Aster, J. C., Pear, W. S., and Blacklow, S. C. (2017). The varied roles of Notch in cancer. *Annu. Rev. Pathol.* 12, 245–275. doi: 10.1146/annurev-pathol-052016-100127

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bell, R. M., and Koren, Y. (2007). "Scalable collaborative filtering with jointly derived neighborhood interpolation weights," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (Omaha, NE), 43–52. doi: 10.1109/ICDM.2007.90

Brastianos, P. K., Carter, S. L., Santagata, S., Cahill, D. P., Taylor-Weiner, A., Jones, R. T., et al. (2015). Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discov.* 5, 1164–1177. doi: 10.1158/2159-8290.CD-15-0369

Chaffer, C. L., and Weinberg, R. A. (2011). A perspective on cancer cell metastasis. *Science* 331, 1559–1564. doi: 10.1126/science.1203543

Chambers, A. F., Groom, A. C., and MacDonald, I. C. (2002). Dissemination and growth of cancer cells in metastatic sites. *Nat. Rev. Cancer* 2, 563–572. doi: 10.1038/nrc865

de Bruin, E. C., McGranahan, N., Mitter, R., Salm, M., Wedge, D. C., Yates, L., et al. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* 346, 251–256. doi: 10.1126/science.1253462

Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., et al. (2008). Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res.* 14, 5158–5165. doi: 10.1158/1078-0432.CCR-07-4756

Desper, R., Khan, J., and Schäffer, A. A. (2004). Tumor classification using phylogenetic methods on expression data. *J. Theor. Biol.* 228, 477–496. doi: 10.1016/j.jtbi.2004.02.021

Ding, L., Raphael, B. J., Chen, F., and Wendl, M. C. (2013). Advances for studying clonal evolution in cancer. *Cancer Lett.* 340, 212–219. doi: 10.1016/j.canlet.2012.12.028

Eaton, J., Wang, J., and Schwartz, R. (2018). Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics* 34, i357–i365. doi: 10.1093/bioinformatics/bty270

Elyanow, R., Dumitrascu, B., Engelhardt, B. E., and Raphael, B. J. (2020). netNMF-sc: leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *Genome Res.* 30, 195–204. doi: 10.1101/gr.251603.119

Floyd, R. W. (1962). Algorithm 97: shortest path. *Commun. ACM* 5, 344–348. doi: 10.1145/367766.368166

Funk, S. (2006). *Netflix Update: Try This at Home*. Technical report.

Greaves, M., and Maley, C. C. (2012). Clonal evolution in cancer. *Nature* 481, 306–313. doi: 10.1038/nature10762

Guan, X. (2015). Cancer metastases: challenges and opportunities. *Acta Pharma. Sin. B* 5, 402–418. doi: 10.1016/j.apsb.2015.07.005

Gupta, S., Takebe, N., and Lorusso, P. (2010). Targeting the Hedgehog pathway in cancer. *Ther. Adv. Med. Oncol.* 2, 237–250. doi: 10.1177/1758834010366430

Hofer, A. M., and Lefkimmiatis, K. (2007). Extracellular calcium and cAMP: second messengers as "Third Messengers?" *Physiology* 22, 320–327. doi: 10.1152/physiol.00019.2007

Hosack, D. A., Dennis Jr, G., Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4:R70. doi: 10.1186/gb-2003-4-10-r70

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kingma, D., and Ba, J. (2014). "Adam: a method for stochastic optimization," in *International Conference on Learning Representations* (San Diego, CA).

Körber, V., Yang, J., Barah, P., Wu, Y., Stichel, D., Gu, Z., et al. (2019). Evolutionary trajectories of IDHWT glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. *Cancer Cell* 35, 692–704.e12. doi: 10.1016/j.ccell.2019.02.007

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer* 42, 30–37. doi: 10.1109/MC.2009.263

Lee, D. D., and Seung, H. S. (2000). "Algorithms for non-negative matrix factorization," in *Proceedings of the 13th International Conference on Neural Information Processing Systems, NIPS'00* (Cambridge, MA: MIT Press), 535–541.

Lee, S., and Margolin, K. (2011). Cytokines in cancer immunotherapy. *Cancers* 3, 3856–3893. doi: 10.3390/cancers3043856

Lei, H., Gertz, E. M., Schäffer, A. A., Fu, X., Tao, Y., Heselmeyer-Haddad, K., et al. (2020). Tumor heterogeneity assessed by sequencing and fluorescence *in situ* hybridization (fish) data. *bioRxiv*. doi: 10.1101/2020.02.29.970392

Lei, H., Lyu, B., Gertz, E. M., Schäffer, A. A., Shi, X., Wu, K., et al. (2019). "Tumor copy number deconvolution integrating bulk and single-cell sequencing data," in *Research in Computational Molecular Biology*, ed L. J. Cowen (Cham: Springer International Publishing), 174–189. doi: 10.1007/978-3-030-17083-7_11

Lin, N. U., Bellon, J. R., and Winer, E. P. (2004). CNS metastases in breast cancer. *J. Clin. Oncol.* 22, 3608–3617. doi: 10.1200/JCO.2004.01.175

Lu, C. L., Tang, C. Y., and Lee, R. C.-T. (2003). The full Steiner tree problem. *Theor. Comput. Sci.* 306, 55–67. doi: 10.1016/S0304-3975(03)00209-3

Massagué, J. (2008). TGFβ in cancer. *Cell* 134, 215–230. doi: 10.1016/j.cell.2008.07.001

Nei, M., and Saitou, N. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Park, Y., Shackney, S., and Schwartz, R. (2009). Network-based inference of cancer progression from microarray data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 6, 200–212. doi: 10.1109/TCBB.2008.126

Priedigkeit, N., Hartmaier, R. J., Chen, Y., Vareslija, D., Basudan, A., Watters, R. J., et al. (2017). Intrinsic subtype switching and acquired ERBB2/HER2 amplifications and mutations in breast cancer brain metastases. *JAMA Oncol.* 3, 666–671. doi: 10.1001/jamaoncol.2016.5630

Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., et al. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* 29, 886–891. doi: 10.1038/nbt.1991

Riester, M., Stephan-Otto Attolini, C., Downey, R. J., Singer, S., and Michor, F. (2010). A differentiation-based phylogeny of cancer subtypes. *PLoS Comput. Biol.* 6:e1000777. doi: 10.1371/journal.pcbi.1000777

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323:533. doi: 10.1038/323533a0

Schwartz, R., and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* 18:213. doi: 10.1038/nrg.2016.170

Schwartz, R., and Shackney, S. E. (2010). Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics* 11:42. doi: 10.1186/1471-2105-11-42

Tao, Y., Cai, C., Cohen, W. W., and Lu, X. (2020). "From genome to phenome: Predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer," in *Pacific Symposium on Biocomputing* (Hawaii).

Tao, Y., Lei, H., Lee, A. V., Ma, J., and Schwartz, R. (2019a). "Phylogenies derived from matched transcriptome reveal the evolution of cell populations and temporal order of perturbed pathways in breast cancer brain metastases," in *Mathematical and Computational Oncology*, eds G. Bebis, T. Benos, K. Chen, K. Jahn, and E. Lima (Cham: Springer International Publishing), 3–28. doi: 10.1007/978-3-030-35210-3_1

Tao, Y., Rajaraman, A., Cui, X., Cui, Z., Eaton, J., Kim, H., et al. (2019b). Improving personalized prediction of cancer prognoses with clonal evolution models. *bioRxiv*. doi: 10.1101/761510

Vareslija, D., Priedigkeit, N., Fagan, A., Purcell, S., Cosgrove, N., O'Halloran, P. J., et al. (2018). Transcriptome characterization of matched primary breast and brain metastatic tumors to detect novel actionable targets. *J. Natl. Cancer Inst.* 111, 388–398. doi: 10.1093/jnci/djy110

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244. doi: 10.1080/01621459.1963.10500845

Warshall, S. (1962). A theorem on boolean matrices. *J. ACM* 9, 11–12. doi: 10.1145/321105.321107

Witzel, I., Oliveira-Ferrer, L., Pantel, K., Muller, V., and Wikman, H. (2016). Breast cancer brain metastases: biology and new clinical perspectives. *Breast Cancer Res.* 18:8. doi: 10.1186/s13058-015-0665-1

Wong, R. S. Y. (2011). Apoptosis in cancer: from pathogenesis to treatment. *J. Exp. Clin. Cancer Res.* 30:87. doi: 10.1186/1756-9966-30-87

Zare, H., Wang, J., Hu, A., Weber, K., Smith, J., Nickerson, D., et al. (2014). Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* 10:e1003703. doi: 10.1371/journal.pcbi.1003703

Zhan, T., Rindtorff, N., and Boutros, M. (2016). Wnt signaling in cancer. *Oncogene* 36:1461. doi: 10.1038/onc.2016.304

Zhu, L., Lei, J., Devlin, B., and Roeder, K. (2018). A unified statistical framework for single cell and bulk RNA sequencing data. *Ann. Appl. Stat.* 12, 609–632. doi: 10.1214/17-AOAS1110

Zhu, L., Narloch, J. L., Onkar, S., Joy, M., Broadwater, G., Luedke, C., et al. (2019). Metastatic breast cancers have reduced immune cell recruitment but harbor increased macrophages relative to their matched primary tumors. *J. Immunother. Cancer* 7:265. doi: 10.1186/s40425-019-0755-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.