Phylogenies Derived from Matched Transcriptome Reveal the Evolution of Cell Populations and Temporal Order of Perturbed Pathways in 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, 15213, USA
[2] Joint Carnegie Mellon–University of Pittsburgh Ph.D. Program in Computational Biology, Pittsburgh PA, 15213, USA
[3] Department of Pharmacology and Chemical Biology, UPMC Hillman Cancer Center, Magee–Womens Research Institute, University of Pittsburgh, Pittsburgh PA, 15213, USA
[4] Department of Biological Sciences, Carnegie Mellon University, Pittsburgh PA, 15213, USA russells@andrew.cmu.edu

Abstract. Metastasis is the 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 results 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 temporal 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.[5]

Keywords: Breast cancer · Brain metastases · Phylogenetics · Deconvolution · Pathways · Gene modules.

[5] Algorithmic details, parameter settings, and proofs are provided in an Appendix with source code available at https://github.com/CMUSchwartzLab/BrM-Phylo.

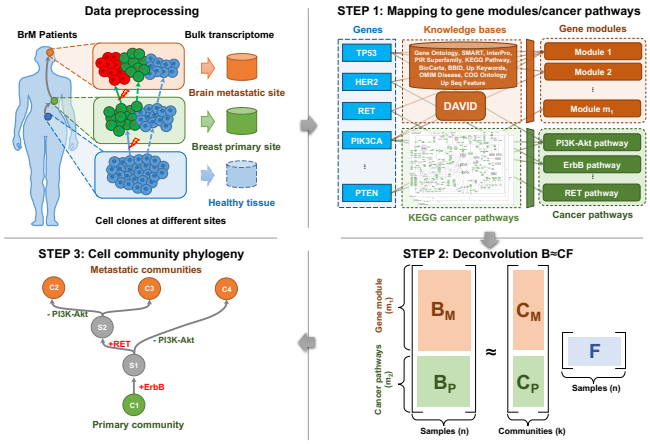

Fig. 1: The pipeline of BrM phylogenetics using matched bulk transcriptome.

1 Introduction

Metastatic disease is the primary mechanism by which cancer results in patient mortality [7,6]. By the time metastases have appeared, there are generally no viable treatment options [14]. Successful treatment thus depends on treating not just the primary tumor but 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 [13]. Phylogenetic studies of clonal populations have been inconclusive on the typical evolutionary relationships between primary and metastatic tumors [35] and 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 [5,10]. 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 [26]. 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 [41]. 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 [38,31]. Some past research has sought to infer phylogenetic models to explain the development of brain metastases based on somatic genomic alterations [4,21]. Such methods are challenged in drawing robust conclusions about recurrent progression processes, though, by the high heterogeneity both within single tumors and across progression stages and patients. 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, but what we dub "cell communities": collections of clones or other stromal cell types that persist as a group with similar proportions across samples (Sec. 2.4). We accomplish this via a novel genomic deconvolution approach designed to make use of multiple samples both within and between patients [36] while improving robustness to inter- and intra-tumor heterogeneity by integrating deconvolution with pathway-based analyses of expression variation [30].

2 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 Fig. 1. After the preprocessing of transcriptome data (Sec. 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, sparsity, and to provide markers of expression variation for the subsequent analysis (Sec. 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 (Sec. 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 (Sec. 2.5).

2.2 Transcriptome Data Preprocessing

We applied a series of preprocessing methods, including quantile normalization [1], to the raw bulk RNA-Sequencing data of 44 matched primary breast

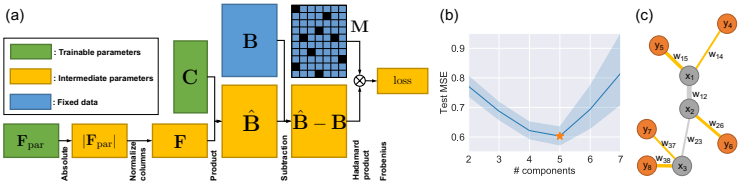

Fig. 2: Method details. (a) Neural network architecture of NND. (b) Test errors of NND using 20-fold CV. Errors in unit of mean square error (MSE). (c) Illustration of a phylogeny with five extant nodes and three Steiner nodes.

and metastatic brain tumors from 22 patients [31,38]. See Appendix Sec. A1 for detailed protocols of data preprocessing.

2.3 Mapping to Gene Modules and Cancer Pathways

The mapping step compresses the high dimensional data and provides markers of cancer-related biological processes (Fig. 1 Step 1). **Gene Modules:** Genes in the same "gene modules" [8,37] are usually affected by a common set of somatic alterations [30], and therefore are co-expressed in cells. We mapped the protein-coding gene expressions into gene modules using the DAVID tool and external knowledge bases [17,18]. The z-scores of $m_1 = 109$ gene modules in all the $n = 44$ samples were represented as a matrix $\mathbf{B}_M \in \mathbb{R}^{m_1 \times n}$. **Cancer Pathways:** We extracted the 23 cancer-related pathways from the KEGG database [19]. An additional recurrently gained *RET pathway* was added [38]. See y-axis of Fig. 3d for the complete list of pathways. z-scores of $m_2 = 24$ cancer pathways were represented as $\mathbf{B}_P \in \mathbb{R}^{m_2 \times n}$. In summary, the raw gene expressions were compressed into the gene module/pathway representation $\mathbf{B} = \left[\mathbf{B}_M^\intercal, \mathbf{B}_P^\intercal\right]^\intercal \in \mathbb{R}^{m \times n}$. The gene module serves for accurately deconvolving and unmixing the cell communities, while the pathway serves as markers/probes and for interpretation purpose. We will refer to the compressed representation containing both gene modules and pathways as "pathway representation" for brevity if not specified. See Appendix Sec. A2 for further details of the mapping.

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 (Fig. 1 Step 2) [22]. Note that common alternatives, such as principal components analysis (PCA) and non-negative matrix factorization (NMF) [23] are not amenable to this case, 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.

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. Particularly when data is sparse and cell types are fit only approximately, as in the present work, 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.

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 Eq. (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 $\mathbf{F}$ (Appendix Fig. A2).

Neural Network Deconvolution. Although it is possible to build new algorithms for solving MF by adapting previous work [23], the additional but necessary constraints of Eq. (2-3) make the optimization much harder to solve. For the problem Eq. (1-3), one can prove that it does not generally guarantee convexity (Appendix Sec. A3.1). A slightly modified version of the algorithm to solve NMF with constraints may guarantee neither good fitting nor convergence [25]. Therefore, instead of revising existing MF algorithms, such as ALS-FunkSVD [3,12,22], 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 (Fig. 2a)

with PyTorch package (`https://pytorch.org/`) [20,34], based on the revised constraints:

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

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

where $|\mathbf{X}|$ applies element-wise absolute value, $\text{cwn}(\mathbf{X})$ is column-wise normalization, so that each column sums up to 1. The two operations of Eq. (5) naturally rephrase and remove the two constraints in Eq. (2-3), and meanwhile fit the framework of neural networks. This implementation is easy to adapt to a wide range of optimization scenarios with various constraints, and has the flexibility of allowing for cross-validation to prevent overfitting.

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 (Fig. 2b). 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.

See Appendix for details of NND, including architecture specifications (Sec. A3.2), hyperparameters (Sec. A3.3), evaluation of fitting ability (Sec. A3.4), sparsity of results (Sec. A3.5), and cross-validation implementation (Sec. A3.6).

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 [27] 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 (Fig. 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 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 to make clear the methodological similarity to more proper phylogenetic methods in wide use for analyzing mutational data in cancers [35].

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 [29], 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 [30].

Inference of Pathways. 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 $\mathcal{V}_S = \{1, 2, ..., k-2\}$, the indices of extant nodes $\mathcal{V}_C = \{k-1, k, ..., 2k-2\}$. For each edge $(u, v) \in \mathcal{E}$, where $1 \leq u < v \leq 2k-2$, the first node of edge $u \leq k-2$ is always a Steiner node. The second node v can be either a Steiner node ($v \leq k-2$) or extant node ($v \geq k-1$). 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 $\mathcal{V}_C$: $\mathbf{y} = [y_{k-1}, y_k, ..., y_{2k-2}]^\intercal = \mathbf{C}_{i\cdot}^\intercal$ and Steiner nodes $\mathcal{V}_S$: $\mathbf{x} = [x_1, x_2, ..., x_{k-2}]^\intercal$. Fig. 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 $U(\mathbf{x}, \mathbf{y}; \mathcal{W})$:

$$\min_{\mathbf{x}} \quad U(\mathbf{x}, \mathbf{y};\ \mathcal{W}) = \sum_{\substack{(u,v)\in\mathcal{E} \\ v\leq k-2}} \frac{1}{2} w_{uv}(x_u - x_v)^2 + \sum_{\substack{(u,v)\in\mathcal{E} \\ v\geq k-1}} \frac{1}{2} w_{uv}(x_u - y_v)^2, \qquad (6)$$

which can be further rephrased as a quadratic programming problem and solved easily. See Appendix Sec. A4 for the derivation and proof of this section.

3 Results

3.1 Gene Modules/Pathways Provide an Effective Representation

Gene expressions of samples were mapped into gene module and pathway space in order to reduce the noise of raw transcriptome data and reduce redundancy (Sec. 2.3). We verified that the 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.

Feature Space of the Pathway Representation. As one can see in Fig. 3a, the first principal component analysis (PCA) dimension of the 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 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 pathway and raw gene expression representations, we applied hierarchical clustering to the 44 samples using Ward's minimum variance method [39]. Two hierarchical trees were built based on the two different representations (Fig. 3b). The gene module/pathway features more

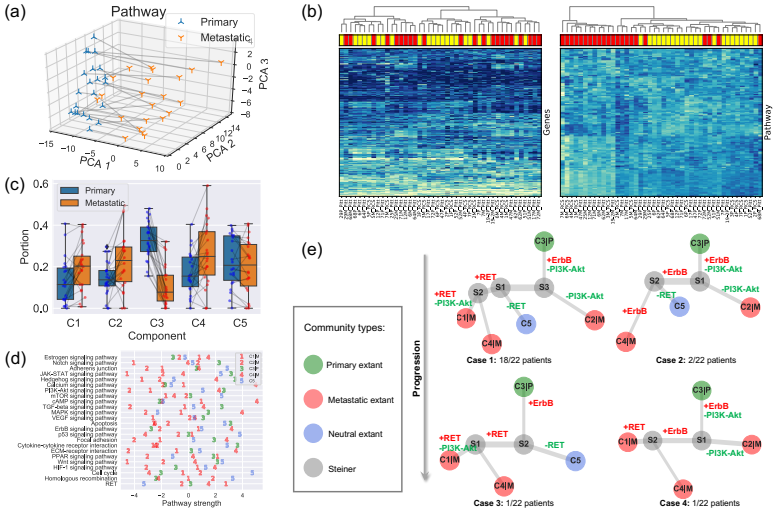

Fig. 3: Results and analysis. (a) First three pathway representation PCA dimensions of matched primary and metastatic samples. Matched samples are connected. (b) Hierarchical clustering of tumor samples based on raw gene expressions (left panel) and compressed gene module/pathway representation (right 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.

effectively separate the primary and metastatic samples into distinct clusters (Fig. 3b right panel) than do the raw gene expression values (Fig. 3a 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 over primary vs. metastatic features. Following previous work [30], 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 [11,40]. All the edge length 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.

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; Appendix Table A1). The signaling pathways related to neurotransmitter and calcium homeostasis (*cAMP*, *Calcium* [16]) 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 [31,38]. Three pathways related to immune activity are under-expressed in metastatic samples (*Cytokine-cytokine receptor interaction* [24], *JAK-STAT* [24], *Notch* [2]), consistent with the previous inference of reduced immune cell expression in metastases in general and brain metastasis most prominently [44]. 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 signalling pathways (*Apoptosis* [42], *Wnt* [43], *Hedgehog* [15], *PI3K-Akt* [4], *TGF-beta* [28]) 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 co-expression pattern of genes [31,38], while our work uses external knowledge bases. Previous research also used somatic mutations or copy number variation to analyze perturbed genes [4,31], 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.

3.2 Landscape of Deconvolved Cell Communities in Tumors

We unmixed the bulk data $\mathbf{B}$ into five components using NND (Sec. 2.4). The deconvolution enables us to produce at least a coarse-grained landscape of major cell communities $\mathbf{C}$ and their distributions in primary and metastatic tumors $\mathbf{F}$.

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**

Community Distributions across Samples F. The portions of the 5 components in all the 44 samples are represented as the mixture fraction matrix $\mathbf{F} \in \mathbb{R}^{5\times 44}$ (Fig. 3c). 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 (Fig. 3c). 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 (Appendix Fig. A3).

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 (Fig. 3d), compared with that in Sec. 3.1, which is only able to distinguish the differentially expressed pathways in bulk samples. As noted in Sec. 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 versus 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 A1, it seems to be quite over-expressed in the metastatic communities $C1|M$ and $C4|M$ but not in the metastatic community $C2|M$.

3.3 Phylogenies of BrM Communities Reveal Common Temporal Order of Perturbed Pathways

We built phylogenies of cell communities and calculated the pathway representations of their Steiner nodes (Sec. 2.5). The phylogenies' topologies provide a way to infer a likely evolutionary history of cancer cell communities and thus their

constitutive cell types, while the perturbed pathways along their edges suggest the temporal order of genomic alterations or changes in community composition.

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 (Appendix Fig. A3). There are four different cases in general (Fig. 3e). 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 Case 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.

Common Temporal 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 (Appendix Table A2-A5). 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 [31,38,4], i.e., *ErbB*, *PI3K-Akt*, and *RET* (Fig. 3e). 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 Conclusions and Future Work

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 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. 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 stie 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. Validation is challenging as 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 [4,31,38]. Future work might compare to prior approaches for reconstruction of clonal evolution from expression data more generically [9,33,36] and seek replication on additional real or simulated expression data or artificial mixtures of different cell types [32] 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.

Funding

This work was supported in part by a grant from the Mario Lemieux Foundation, U.S. N.I.H. award R21CA216452, Pennsylvania Department of Health award 4100070287, Breast Cancer Alliance, Susan G. Komen for the Cure, and by a fellowship to Y.T. from the Center for Machine Learning and Healthcare at

Carnegie Mellon University. The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions.

References

1. Amaratunga, D., et al.: Analysis of data from viral DNA microchips. Journal of the American Statistical Association **96**(456), 1161–1170 (2001)
2. Aster, J.C., et al.: The varied roles of Notch in cancer. Annual Review of Pathology **12**, 245–275 (jan 2017)
3. Bell, R.M., et al.: Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In: Seventh IEEE International Conference on Data Mining (ICDM 2007). pp. 43–52 (2007)
4. Brastianos, P.K., et al.: Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. Cancer discovery **5**(11), 1164–1177 (nov 2015)
5. de Bruin, E.C., et al.: Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science (New York, N.Y.) **346**(6206), 251–256 (oct 2014)
6. Chaffer, C.L., et al.: A perspective on cancer cell metastasis. Science **331**(6024), 1559–1564 (2011)
7. Chambers, A.F., et al.: Dissemination and growth of cancer cells in metastatic sites. Nature Reviews Cancer **2**(8), 563–572 (2002)
8. Desmedt, C., et al.: Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clinical Cancer Research **14**(16), 5158–5165 (2008)
9. Desper, R., et al.: Tumor classification using phylogenetic methods on expression data. Journal of Theoretical Biology **228**(4), 477–496 (2004)
10. Ding, L., et al.: Advances for studying clonal evolution in cancer. Cancer letters **340**(2), 212–219 (nov 2013)
11. Floyd, R.W.: Algorithm 97: Shortest path. Communications of the ACM **5**(6), 344–348 (Jun 1962)
12. Funk, S.: Netflix update: Try this at home (2006)
13. Greaves, M., et al.: Clonal evolution in cancer. Nature **481**(7381), 306–313 (2012)
14. Guan, X.: Cancer metastases: Challenges and opportunities. Acta Pharmaceutica Sinica B **5**(5), 402–418 (sep 2015)
15. Gupta, S., et al.: Targeting the Hedgehog pathway in cancer. Therapeutic Advances in Medical Oncology **2**(4), 237–250 (jul 2010)
16. Hofer, A.M., et al.: Extracellular Calcium and cAMP: Second messengers as Third Messengers? Physiology **22**(5), 320–327 (oct 2007)
17. Hosack, D.A., et al.: Identifying biological themes within lists of genes with EASE. Genome Biology **4**(10), R70–R70 (2003)
18. Huang, D.W., et al.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols **4**(1), 44–57 (2009)
19. Kanehisa, M., et al.: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research **28**(1), 27–30 (jan 2000)
20. Kingma, D., et al.: Adam: A method for stochastic optimization. International Conference on Learning Representations (dec 2014)
21. Körber, V., et al.: Evolutionary trajectories of IDHWT glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. Cancer Cell (mar 2019)

22. Koren, Y., et al.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (aug 2009)
23. Lee, D.D., et al.: Algorithms for non-negative matrix factorization. In: Proceedings of the 13th International Conference on Neural Information Processing Systems. pp. 535–541. NIPS'00, MIT Press, Cambridge, MA, USA (2000)
24. Lee, S., et al.: Cytokines in cancer immunotherapy. Cancers **3**(4), 3856–3893 (oct 2011)
25. Lei, H., et al.: Tumor copy number deconvolution integrating bulk and single-cell sequencing data. In: Cowen, L.J. (ed.) Research in Computational Molecular Biology. pp. 174–189. Springer International Publishing, Cham (2019)
26. Lin, N.U., et al.: CNS metastases in breast cancer. Journal of Clinical Oncology **22**(17), 3608–3617 (sep 2004)
27. Lu, C.L., et al.: The full Steiner tree problem. Theoretical Computer Science **306**(1), 55–67 (sep 2003)
28. Massagué, J.: TGFβ in cancer. Cell **134**(2), 215–230 (2008)
29. Nei, M., et al.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4**(4), 406–425 (jul 1987)
30. Park, Y., et al.: Network-based inference of cancer progression from microarray data. IEEE/ACM Transactions on Computational Biology and Bioinformatics **6**(2), 200–212 (April 2009)
31. Priedigkeit, N., et al.: Intrinsic subtype switching and acquired ERBB2/HER2 amplifications and mutations in breast cancer brain metastases. JAMA oncology **3**(5), 666–671 (may 2017)
32. Qiu, P., et al.: Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nature Biotechnology **29**(10), 886–891 (2011)
33. Riester, M., et al.: A differentiation-based phylogeny of cancer subtypes. PLOS Computational Biology **6**(5), e1000777 (may 2010)
34. Rumelhart, D.E., et al.: Learning representations by back-propagating errors. Nature **323**, 533 (oct 1986)
35. Schwartz, R., et al.: The evolution of tumour phylogenetics: principles and practice. Nature Reviews Genetics **18**, 213 (feb 2017)
36. Schwartz, R., et al.: Applying unmixing to gene expression data for tumor phylogeny inference. BMC Bioinformatics **11**(1), 42 (jan 2010)
37. Tao, Y., et al.: From genome to phenome: Predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer. In: Pacific Symposium on Biocomputing (2020)
38. Vareslija, D., et al.: Transcriptome characterization of matched primary breast and brain metastatic tumors to detect novel actionable targets. Journal of the National Cancer Institute (jun 2018)
39. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**(301), 236–244 (mar 1963)
40. Warshall, S.: A theorem on boolean matrices. Journal of the ACM **9**(1), 11–12 (jan 1962)
41. Witzel, I., et al.: Breast cancer brain metastases: biology and new clinical perspectives. Breast Cancer Research **18**(1), 8 (jan 2016)
42. Wong, R.S.Y.: Apoptosis in cancer: from pathogenesis to treatment. Journal of Experimental & Clinical Cancer Research **30**(1), 87 (sep 2011)
43. Zhan, T., et al.: Wnt signaling in cancer. Oncogene **36**, 1461 (sep 2016)
44. Zhu, L., et al.: Metastatic breast cancers have reduced immune cell recruitment but harbor increased macrophages relative to their matched primary tumors. bioRxiv: 525071 (2019)

Appendix

A1 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) [31,38], where six patients are from the Royal College of Surgeons (RCS) and sixteen patients from the University of Pittsburgh (Pitt). These data profiled the expression levels of approximately 60,000 transcripts. We removed the genes that are not expressed in any sample. We also considered only protein-coding genes in the present study. We conducted quantile normalization across samples using the geometric mean to remove possible artifacts from different experiment batches [1]. The top 2.5% and bottom 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.

A2 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 [18], as well as the cancer pathways in KEGG database [19]. This step compresses the high dimensional data and provides markers of cancer-related biological processes (Fig. 1 Step 1).

Gene Modules. Functionally similar genes are usually affected by a common set of somatic alterations [30] and therefore are co-expressed in the cells. These genes are believed to belong to the same "gene modules" [8,37]. Inspired by the idea of gene modules, we fed a subset of 3,000 most informative genes out of the approximately 20,000 genes that have the largest variances into the DAVID tool for functional annotation clustering using several databases [18]. 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 [17]. 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.

Cancer Pathways. Although gene module representation is able to capture the variances across samples and reduce the redundancy of raw gene expressions, it has two disadvantages: First of all, 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. Secondly, the key perturbed cancer pathways or functions may not be always the ones that vary most across samples. For example, genes in cancer-related KEGG pathways (hsa05200) [19] 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 [38]. See y-axis of Fig. 3d 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.

A3 Deconvolution of Bulk Data

A3.1 Non-convexity of Deconvolution Problem

Theorem 1. *The deconvolution problem Eq. (1-3) below is not convex:*

$$\min_{\boldsymbol{C},\boldsymbol{F}} \quad f(\boldsymbol{C},\boldsymbol{F}) = \|\boldsymbol{B} - \boldsymbol{C}\boldsymbol{F}\|_{\mathrm{Fr}}^2, \tag{A1}$$

$$s.t. \quad \boldsymbol{F}_{lj} \geq 0, \qquad l = 1,...,k,\ j = 1,...,n, \tag{A2}$$

$$\sum_{l=1}^{k} \boldsymbol{F}_{lj} = 1, \qquad j = 1,...,n. \tag{A3}$$

Proof. If the problem is convex, we should have: $\forall \lambda \in (0,1)$, and $\forall \mathbf{C}_x, \mathbf{C}_y, \mathbf{F}_x, \mathbf{F}_y$ in the feasible domain, the following inequality always holds:

$$\lambda f(\mathbf{C}_x, \mathbf{F}_x) + (1-\lambda) f(\mathbf{C}_y, \mathbf{F}_y) \geq f(\lambda \mathbf{C}_x + (1-\lambda)\mathbf{C}_y, \lambda \mathbf{F}_x + (1-\lambda)\mathbf{F}_y). \tag{A4}$$

However, for the following setting:

$$\mathbf{B} = \begin{bmatrix} -1.38 & 0.92 \\ 1.03 & -0.15 \end{bmatrix}, \tag{A5}$$

$$\mathbf{C}_x = \begin{bmatrix} -1.74 & 2.21 \\ 1.00 & -3.97 \end{bmatrix}, \qquad \mathbf{C}_y = \begin{bmatrix} 1.03 & -0.46 \\ -3.13 & 0.16 \end{bmatrix}, \tag{A6}$$

$$\mathbf{F}_x = \begin{bmatrix} 0.83 & 0.32 \\ 0.17 & 0.68 \end{bmatrix}, \qquad \mathbf{F}_y = \begin{bmatrix} 0.09 & 0.34 \\ 0.91 & 0.66 \end{bmatrix}, \tag{A7}$$

and $\lambda = 0.5$, we have

$$\lambda f(\mathbf{C}_x, \mathbf{F}_x) + (1-\lambda) f(\mathbf{C}_y, \mathbf{F}_y) = 4.86 < 11.74 = f(\lambda \mathbf{C}_x + (1-\lambda)\mathbf{C}_y, \lambda \mathbf{F}_x + (1-\lambda)\mathbf{F}_y). \tag{A8}$$

This is contradictory to Eq. (A4). □

A3.2 Architecture Specifications of NND

In the NND architecture, $|\mathbf{X}|$ applies element-wise absolute value, $\mathrm{cwn}(\mathbf{X})$ column-wisely normalizes $\mathbf{X}$, so that each column of the output sums up to 1. The two operations of Eq. (5) naturally rephrase and remove the two constraints in Eq. (2-3), and meanwhile fit the framework of neural networks. An alternative to the absolute value operation $|\mathbf{X}|$ might be rectified linear unit $\mathrm{ReLU}(\mathbf{X}) = \max(\mathbf{0}, \mathbf{X})$. 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 $\mathrm{cwn}(\mathbf{X})$ with softmax operation $\mathrm{softmax}(\mathbf{X})$. However, the nonlinearity introduced by softmax actually changes the original optimization problem Eq. (1-3) and the fitted $\mathbf{F}$ is therefore not sparse.

A3.3 Hyperparameters of NND

We used an Adam optimizer with default momentum parameters and learning rate of 1×10^{-5} [20]. 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 was run until convergence, when the relative decrease of training loss is smaller than $\epsilon = 1 \times 10^{-10}$ every 20,000 iterations.

A3.4 Fitting Ability of NND

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 $\hat{\mathbf{B}} = \mathbf{CF}$ (Fig. A1). One can easily see that NND provides good model fits to the data.

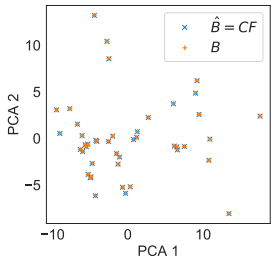

Fig. A1: PCA of pathway representation $\mathbf{B}$ and nnMF fitted $\hat{\mathbf{B}}$. Each dot represents the pathway values of a sample $\mathbf{B}_{\cdot j}$ or fitted $\hat{\mathbf{B}}_{\cdot j}$. The first two PCA dimensions of original data and fitted data are almost in the same positions, which indicates that NND is able to fit precisely in our application. The number of components is set to be $k = 5$ here.

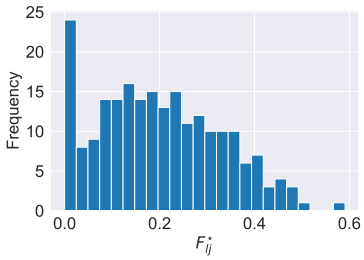

Fig. A2: Distribution of elements in fraction matrix $\mathbf{F}^{\star}$. Since each column of $\mathbf{F}$ is forced to sum up to be one, a Laplacian prior is applied to the elements of matrix $\mathbf{F}$. This leads to the sparsity of $\mathbf{F}^{\star}$: 24 out of its 220 elements ($k \times n = 5 \times 44$) are zeros (threshold set to 2.5×10^{-2}).

A3.5 Sparsity of NND Results

See Fig. A2 for distribution of fraction matrix in NND deconvolution results.

A3.6 Cross-validation of NND

In each fold of the CV, we used $\hat{\mathbf{B}} = \mathbf{CF}$ 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 Eq. (5), the optimization goal is:

$$\min_{\mathbf{C},\mathbf{F}_{\text{par}}} \|\mathbf{M}_{\text{train}} \odot (\mathbf{B} - \mathbf{CF})\|_{\text{Fr}}^2 \tag{A9}$$

where $\mathbf{X} \odot \mathbf{Y}$ is the Hadamard (element-wise) product. At the time of evaluation, given optimized $\mathbf{C}^\star$, $\mathbf{F}^\star_{\text{par}}$, and therefore optimized $\mathbf{F}^\star = \text{cwn}\left(\left|\mathbf{F}^\star_{\text{par}}\right|\right)$ for the optimization problem during training, the test error was calculated on the test set: $\|\mathbf{M}_{\text{test}} \odot (\mathbf{B} - \mathbf{C}^\star\mathbf{F}^\star)\|^2_{\text{Fr}}$. We used 20-fold cross-validation on the NND, so in each fold 95% positions of $\mathbf{M}_{\text{train}}$ and 5% positions of $\mathbf{M}_{\text{test}}$ were 1s.

A4 Derivation of Quadractic Programming, $\mathbf{P}(\mathcal{W})$, and $\mathbf{q}(\mathcal{W}, \mathbf{c})$

Recall Sec. 2.5, for the phylogeny $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the Steiner nodes are indexed as $\mathcal{V}_S = \{1, 2, ..., k-2\}$ ($|\mathcal{V}_S| = k-2$), the extant nodes are indexed as $\mathcal{V}_C = \{k-1, k, ..., 2k-2\}$ ($|\mathcal{V}_C| = k$). The i-th pathway values of Steiner nodes are denoted as $\mathbf{x} = [x_1, x_2, ..., x_{k-2}]^\intercal \in \mathbb{R}^{k-2}$, and values of extant nodes as $\mathbf{y} = [y_{k-1}, y_k, ..., y_{2k-2}]^\intercal \in \mathbb{R}^k$. Since we consider each pathway dimension separately here, the subscript i for $\mathbf{x}$ and $\mathbf{y}$ is omitted for brevity. The weight of edge $(u, v) \in \mathcal{E}$ connecting nodes u and v is w_{uv} ($1 \le u < v \le 2k-2$). Denote $\mathcal{W} = \{w_{uv} \mid (u, v) \in \mathcal{E}\}$. The inference of the i-th element in the pathway vector of the Steiner nodes can be formulated as minimizing the elastic potential energy $U(\mathbf{x}, \mathbf{y}; \mathcal{W})$ shown below:

$$\min_{\mathbf{x}} \quad U(\mathbf{x}, \mathbf{y};\, \mathcal{W}) = \sum_{\substack{(u,v)\in\mathcal{E} \\ v \le k-2}} \frac{1}{2} w_{uv}(x_u - x_v)^2 + \sum_{\substack{(u,v)\in\mathcal{E} \\ v \ge k-1}} \frac{1}{2} w_{uv}(x_u - y_v)^2, \quad \text{(A10)}$$

Theorem 2. *Equation (A10) can be further rephrased as a quadratic programming problem:*

$$\min_{\boldsymbol{x}} \quad \frac{1}{2}\boldsymbol{x}^\intercal \boldsymbol{P}(\mathcal{W})\boldsymbol{x} + \boldsymbol{q}(\mathcal{W}, \boldsymbol{y})^\intercal \boldsymbol{x}, \quad \text{(A11)}$$

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

Proof. Based on Eq. (A10), $U(\mathbf{x}, \mathbf{y}; \mathcal{W}) \ge 0$. Each term inside the first summation ($v \le k-2$) can be written as:

$$\frac{1}{2} w_{uv}(x_u - x_v)^2 = \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x}, \quad \text{(A12)}$$

where

$$\mathbf{P}(w_{uv}) = \begin{array}{c} \\ u\text{-th row} \\ \\ v\text{-th row} \\ \end{array} \begin{array}{c} \begin{array}{ccccc} & u\text{-th col} & & v\text{-th col} & \end{array} \\ \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & w_{uv} & \mathbf{0} & -w_{uv} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -w_{uv} & \mathbf{0} & w_{uv} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \end{array}. \quad \text{(A13)}$$

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

$$\frac{1}{2}w_{uv}(x_u - y_v)^2 = \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x} + \mathbf{q}(w_{uv}, y_v)^\intercal \mathbf{x} + C(w_{uv}, y_v), \tag{A14}$$

where

$$\mathbf{P}(w_{uv}) = \begin{matrix} \\ u\text{-th row} \end{matrix} \overset{\begin{matrix} u\text{-th col} \end{matrix}}{\begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0\,0\,0} \\ \mathbf{0} & w_{uv} & \mathbf{0\,0\,0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0\,0\,0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0\,0\,0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0\,0\,0} \end{bmatrix}}, \quad \mathbf{q}(w_{uv}, y_v) = \begin{matrix} \\ u\text{-th row} \end{matrix} \begin{bmatrix} \mathbf{0} \\ -w_{uv}y_v \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tag{A15}$$

and $C(w_{uv}, y_v) = \frac{1}{2}w_{uv}y_v^2$ is independent of $\mathbf{x}$. Therefore the optimization in Eq. (A10) can be calculated and written as below:

$$\min_{\mathbf{x}} \quad \sum_{\substack{(u,v)\in\mathcal{E} \\ v\leq k-2}} \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x} + \sum_{\substack{(u,v)\in\mathcal{E} \\ v\geq k-1}} \left(\frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(w_{uv})\mathbf{x} + \mathbf{q}(w_{uv}, y_v)^\intercal \mathbf{x}\right), \tag{A16}$$

$$\Leftrightarrow \min_{\mathbf{x}} \quad \frac{1}{2}\mathbf{x}^\intercal \left(\sum_{\substack{(u,v)\in\mathcal{E} \\ v\leq k-2}} \mathbf{P}(w_{uv}) + \sum_{\substack{(u,v)\in\mathcal{E} \\ v\geq k-1}} \mathbf{P}(w_{uv})\right)\mathbf{x} + \sum_{\substack{(u,v)\in\mathcal{E} \\ v\geq k-1}} \mathbf{q}(w_{uv}, y_v)^\intercal \mathbf{x}, \tag{A17}$$

$$\Leftrightarrow \min_{\mathbf{x}} \quad \frac{1}{2}\mathbf{x}^\intercal \mathbf{P}(\mathcal{W})\mathbf{x} + \mathbf{q}(\mathcal{W}, \mathbf{y})^\intercal \mathbf{x}. \square \tag{A18}$$

Remark 1. The optimal $\mathbf{x}^\star$ of the Eq. (A10), or the solution to the quadratic programming problem Eq. (A11) can be solved by setting the gradient to be $\mathbf{0}$:

$$\mathbf{P}(\mathcal{W})\mathbf{x}^\star + \mathbf{q}(\mathcal{W}, \mathbf{y}) = \mathbf{0}. \tag{A19}$$

Therefore,

$$\mathbf{x}^\star = -\mathbf{P}(\mathcal{W})^{-1}\mathbf{q}(\mathcal{W}, \mathbf{y}). \tag{A20}$$

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:

$$\mathbf{P} \leftarrow 0^{(k-2)\times(k-2)}, \quad \mathbf{q} \leftarrow 0^{k-2}. \tag{A21}$$

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 ($u \leq k-2$, $v \leq k-2$), we update $\mathbf{P}$ and keep $\mathbf{q}$ the same:

$$\mathbf{P}_{uu} \leftarrow \mathbf{P}_{uu} + w_{uv},\ \mathbf{P}_{vv} \leftarrow \mathbf{P}_{vv} + w_{uv},\ \mathbf{P}_{uv} \leftarrow \mathbf{P}_{uv} - w_{uv},\ \mathbf{P}_{vu} \leftarrow \mathbf{P}_{vu} - w_{uv}. \tag{A22}$$

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

$$\mathbf{P}_{uu} \leftarrow \mathbf{P}_{uu} + w_{uv}, \quad \mathbf{q}_u \leftarrow \mathbf{q}_u - y_v \cdot w_{uv}. \tag{A23}$$

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

A5 Differentially Expressed Cancer Pathways

Table A1 provides a list of the identified differentially expressed cancer pathways.

Table A1: 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

A6 Portions of Cell Communities in BrM Patients

Fig. A3 shows the inferred cell community portions across the BrM samples. The figure displays, for each patient, the proportion of each community in the primary and the metastatic sample.

A7 Perturbed Cancer Pathways along Phylogenies

Table A2-A5 provide a full list of perturbed pathways across the phylogenies for Case 1, 2, 3, and 4 in Fig. 3e.

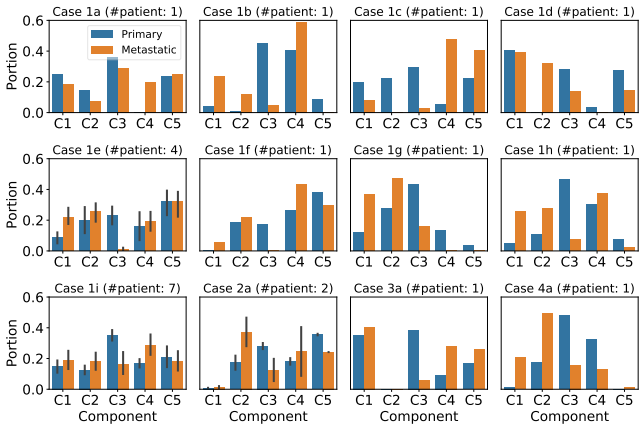

Fig. A3: Classification of BrM patients based on the consisted cell subcommunities in matched samples. There are 12 subcases of the 4 cases mentioned in Sec. 3.2. Specifically, there are 9 specific cases (Case 1a-i) in Case 1. Most patients (7) have all the five cell communities in both primary and metastatic samples (Case 1i). A few patients (4) have all communities in metastasis samples and all clones but community $C3|P$ in primary samples. The element $\mathbf{F}_{lj}$ is taken as 0 when it is smaller than a threshold 2.5×10^{-2}, and therefore the l-th community is missing in the j-th sample.

Table A2: Perturbed pathways during the evolution of cell communities in primary and metastatic tumors (Fig. 3e Case 1). The top five perturbed pathways whose gain or loss greater than 1.0 along each edge of phylogeny are shown. Clinically actionable perturbed cancer pathways during metastasis are shown in boldface, i.e., *ErbB*, *RET*, and *PI3K-Akt* [4,31,38].

Trajectory	Gain	Perturbed Pathways	Loss	Perturbed Pathways
$C3\|P \rightarrow S3$	+2.83	Homologous recombination	−3.76	Hedgehog signaling pathway
	+2.41	Cell cycle	−3.45	Cytokine-cytokine receptor interaction
	+1.86	**ErbB signaling pathway**	−3.39	**PI3K-Akt signaling pathway**
	+1.10	cAMP signaling pathway	−3.15	TGF-beta signaling pathway
			−3.14	JAK-STAT signaling pathway
$S3 \rightarrow S1$	< 1.0	$\emptyset$	< 1.0	$\emptyset$
$S1 \rightarrow S2$	+1.36	cAMP signaling pathway	−1.28	JAK-STAT signaling pathway
	+1.18	**RET**	−1.22	Apoptosis
			−1.21	Cytokine-cytokine receptor interaction
			−1.12	Wnt signaling pathway
			−1.04	Notch signaling pathway
$S2 \rightarrow C1\|M$	+1.90	**RET**	−3.25	Wnt signaling pathway
	+1.59	PPAR signaling pathway	−3.11	JAK-STAT signaling pathway
			−2.77	Notch signaling pathway
			−2.48	Hedgehog signaling pathway
			−2.18	**PI3K-Akt signaling pathway**
$S2 \rightarrow C4\|M$	+4.48	Calcium signaling pathway	−3.06	p53 signaling pathway
	+4.17	cAMP signaling pathway	−2.74	Cell cycle
	+3.83	MAPK signaling pathway	−2.21	Homologous recombination
	+3.35	ECM-receptor interaction	−1.40	Apoptosis
	+3.20	Focal adhesion	−1.33	Cytokine-cytokine receptor interaction
$S1 \rightarrow C5$	+3.91	Cell cycle	−3.00	**RET**
	+3.17	p53 signaling pathway	−1.58	MAPK signaling pathway
	+2.85	Adherens junction	−1.41	cAMP signaling pathway
	+2.76	Cytokine-cytokine receptor interaction		
	+2.68	Wnt signaling pathway		
$S3 \rightarrow C2\|M$	+1.39	Homologous recombination	−3.65	TGF-beta signaling pathway
			−3.61	**PI3K-Akt signaling pathway**
			−3.34	ECM-receptor interaction
			−3.20	Focal adhesion
			−2.60	PPAR signaling pathway

Table A3: Perturbed pathways during the evolution of cell communities in primary and metastatic tumors (Fig. 3e Case 2). The top five perturbed pathways whose gain or loss greater than 1.0 along each edge of phylogeny are shown.

Trajectory	Gain	Perturbed Pathways	Loss	Perturbed Pathways
$C3\|P \rightarrow S1$	+2.83	Homologous recombination	−3.22	Hedgehog signaling pathway
	+2.47	Cell cycle	−3.10	TGF-beta signaling pathway
	+1.81	**ErbB signaling pathway**	−3.08	Cytokine-cytokine receptor interaction
	+1.02	cAMP signaling pathway	−2.93	**PI3K-Akt signaling pathway**
			−2.64	PPAR signaling pathway
$S1 \rightarrow S2$	+1.08	ECM-receptor interaction		
	+1.08	**ErbB signaling pathway**		
$S2 \rightarrow C4\|M$	+5.51	cAMP signaling pathway	−3.97	Cell cycle
	+5.12	Calcium signaling pathway	−3.83	p53 signaling pathway
	+4.45	MAPK signaling pathway	−3.20	Apoptosis
	+3.37	ECM-receptor interaction	−3.15	Cytokine-cytokine receptor interaction
	+3.08	**ErbB signaling pathway**	−3.00	Homologous recombination
$S2 \rightarrow C5$	+3.68	Cell cycle	−2.25	**RET**
	+3.18	p53 signaling pathway	−1.81	MAPK signaling pathway
	+2.50	Homologous recombination	−1.43	cAMP signaling pathway
	+2.16	Adherens junction	−1.24	Hedgehog signaling pathway
	+2.15	Cytokine-cytokine receptor interaction	−1.13	Calcium signaling pathway
$S1 \rightarrow C2\|M$	+1.39	Homologous recombination	−4.06	**PI3K-Akt signaling pathway**
			−3.70	TGF-beta signaling pathway
			−3.55	Focal adhesion
			−3.52	ECM-receptor interaction
			−2.87	Adherens junction

Table A4: Perturbed pathways during the evolution of cell communities in primary and metastatic tumors (Fig. 3e Case 3). The top five perturbed pathways whose gain or loss greater than 1.0 along each edge of phylogeny are shown.

Trajectory	Gain	Perturbed Pathways	Loss	Perturbed Pathways
$C3\|P \rightarrow S2$	+3.10	Cell cycle	−3.51	Hedgehog signaling pathway
	+3.10	**ErbB signaling pathway**	−2.41	Notch signaling pathway
	+2.93	Homologous recombination	−2.39	Cytokine-cytokine receptor interaction
	+1.70	cAMP signaling pathway	−2.34	JAK-STAT signaling pathway
	+1.66	HIF-1 signaling pathway	−2.07	Apoptosis
$S2 \rightarrow S1$	+1.62	cAMP signaling pathway	−2.02	Cytokine-cytokine receptor interaction
	+1.54	**RET**	−1.98	JAK-STAT signaling pathway
	+1.14	Calcium signaling pathway	−1.91	Apoptosis
			−1.75	Wnt signaling pathway
			−1.32	Cell cycle
$S1 \rightarrow C1\|M$	+1.85	**RET**	−3.52	Wnt signaling pathway
	+1.19	PPAR signaling pathway	−3.38	JAK-STAT signaling pathway
			−2.78	**PI3K-Akt signaling pathway**
			−2.76	Hedgehog signaling pathway
			−2.68	Notch signaling pathway
$S1 \rightarrow C4\|M$	+4.20	Calcium signaling pathway	−3.18	p53 signaling pathway
	+3.89	cAMP signaling pathway	−2.65	Cell cycle
	+3.40	MAPK signaling pathway	−1.99	Homologous recombination
	+2.76	Hedgehog signaling pathway	−1.64	Cytokine-cytokine receptor interaction
	+2.72	ECM-receptor interaction	−1.61	Apoptosis
$S2 \rightarrow C5$	+3.67	Cell cycle	−2.69	**RET**
	+2.76	Homologous recombination	−2.08	MAPK signaling pathway
	+2.56	p53 signaling pathway	−1.59	PPAR signaling pathway
	+1.85	mTOR signaling pathway	−1.43	cAMP signaling pathway
	+1.79	Adherens junction	−1.02	Hedgehog signaling pathway

Table A5: Perturbed pathways during the evolution of cell communities in primary and metastatic tumors (Fig. 3e Case 4). The top five perturbed pathways whose gain or loss greater than 1.0 along each edge of phylogeny are shown.

Trajectory	Gain	Perturbed Pathways	Loss	Perturbed Pathways
$C3\|P \rightarrow S1$	+2.38	Homologous recombination	−4.49	Cytokine-cytokine receptor interaction
	+1.56	**ErbB signaling pathway**	−4.23	**PI3K-Akt signaling pathway**
	+1.54	Cell cycle	−4.10	JAK-STAT signaling pathway
	+1.41	cAMP signaling pathway	−3.97	Hedgehog signaling pathway
			−3.74	Apoptosis
$S1 \rightarrow S2$	+1.89	cAMP signaling pathway	−1.66	Notch signaling pathway
	+1.69	**ErbB signaling pathway**	−1.27	JAK-STAT signaling pathway
	+1.47	HIF-1 signaling pathway	−1.14	Apoptosis
	+1.47	ECM-receptor interaction	−1.01	Cytokine-cytokine receptor interaction
	+1.43	Calcium signaling pathway		
$S2 \rightarrow C1\|M$	+1.43	PPAR signaling pathway	−2.53	Notch signaling pathway
	+1.19	**RET**	−2.44	Wnt signaling pathway
	+1.09	p53 signaling pathway	−2.35	Hedgehog signaling pathway
			−2.32	JAK-STAT signaling pathway
			−1.66	VEGF signaling pathway
$S2 \rightarrow C4\|M$	+4.40	Calcium signaling pathway	−2.37	p53 signaling pathway
	+3.91	cAMP signaling pathway	−1.93	Cell cycle
	+3.81	ECM-receptor interaction	−1.74	Homologous recombination
	+3.64	MAPK signaling pathway		
	+3.62	Focal adhesion		
$S1 \rightarrow C2\|M$	+1.84	Homologous recombination	−3.07	TGF-beta signaling pathway
	+1.39	Cell cycle	−2.77	**PI3K-Akt signaling pathway**
			−2.69	ECM-receptor interaction
			−2.59	Focal adhesion
			−2.58	PPAR signaling pathway