


ORIGINAL ARTICLE

Investigating epithelial-mesenchymal heterogeneity of tumors and circulating tumor cells with transcriptomic analysis and biophysical modeling

Federico Bocci^{1,4}  | Susmita Mandal² | Tanishq Tejaswi^{2,3} | Mohit Kumar Jolly²

¹ Center for Theoretical Biological Physics, Rice University, Houston, Texas, USA

² Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore, Karnataka, India

³ UG Programme, Indian Institute of Science, Bangalore, Karnataka, India

⁴ NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, California, USA

Correspondence

Federico Bocci, NSF Simons Center for Multiscale Cell Fate Research, University of California, Irvine, CA, USA.

Email: fbocci@uci.edu

Mohit Kumar Jolly, Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore, Karnataka, India.

Email: mkjolly@iisc.ac.in

Cellular heterogeneity along the epithelial-mesenchymal plasticity (EMP) spectrum is a paramount feature observed in tumors and circulating tumor cells (CTCs). High-throughput techniques now offer unprecedented details on this variability at a single-cell resolution. Yet, there is no current consensus about how EMP in tumors propagates to that in CTCs. To investigate the relationship between EMP-associated heterogeneity of tumors and that of CTCs, we integrated transcriptomic analysis and biophysical modeling. We apply three epithelial-mesenchymal transition (EMT) scoring metrics to multiple tumor samples and CTC datasets from several cancer types. Moreover, we develop a biophysical model that couples EMT-associated phenotypic switching in a primary tumor with cell migration. Finally, we integrate EMT transcriptomic analysis and in silico modeling to evaluate the predictive power of several measurements of tumor aggressiveness, including tumor EMT score, CTC EMT score, fraction of CTC clusters found in circulation, and CTC cluster size distribution. Analysis of high-throughput datasets reveals a pronounced heterogeneity without a well-defined relation between EMT traits in tumors and CTCs. Moreover, mathematical modeling predicts different phases where CTCs can be less, equally, or more mesenchymal than primary tumor depending on the dynamics of phenotypic transition and cell migration. Consistently, various datasets of CTC cluster size distribution from different cancer types are fitted onto different regimes of the model. By further constraining the model with experimental measurements of tumor EMT score, CTC EMT score, and fraction of CTC cluster in bloodstream, we show that none of these assays alone can provide sufficient information to predict the other variables. In conclusion, we propose that the relationship between EMT progression in tumors and CTCs can be variable, and in general, predicting one from the other may not be as straightforward as tacitly assumed.

KEYWORDS

cancer, circulating tumor cells, epithelial-mesenchymal plasticity, epithelial-mesenchymal transition, RNA-sequencing

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *Computational and Systems Oncology* published by Wiley Periodicals LLC

1 | INTRODUCTION

Epithelial-mesenchymal transition (EMT) (now increasingly referred to as epithelial-mesenchymal plasticity [EMP]) is a crucial axis of tumor progression that regulates motility, resistance to therapies and proliferation [1]. During EMT, epithelial cancer cells can partially or completely lose their E-cadherin-mediated cell-cell adhesion and apical-basal polarity, instead of becoming motile, mesenchymal cells [2]. Recent experimental and computational findings suggest that EMT is not necessarily a binary process; instead cells can manifest one or more hybrid E/M phenotype(s) with mixed traits of epithelial and mesenchymal cells along the EMT (or EMP) spectrum [3–8].

Indeed, cellular heterogeneity is emerging as a hallmark of cancer, with cells with distinct EMT phenotypes often localized in distinct tumor regions [9,10]. Similarly, circulating tumor cells (CTCs) that launch into circulation to give rise to metastasis can exhibit a spectrum of EMT features [11]. Hybrid E/M cancer cells that partially conserve cell adhesion can travel the bloodstream collectively as highly metastatic CTC clusters [12]. Recent advances in high-throughput techniques such as RNA-sequencing are providing insights into the multi-faceted dynamics of EMT, predicting the number of intermediate hybrid E/M states and the most probable EMT trajectories [7]. However, this heterogeneity sparks questions about the relationship between EMT progression in primary tumors and CTCs. Is it possible to reliably predict the composition of a tumor from gene expression measurements of CTCs, and/or vice versa, or does the relationship between EMT progression in tumors and CTCs depend on tumor type, subtype, or even patient-specific factors? Furthermore, how much can be said about EMT progression in tumors by measuring statistical properties of CTC migration, such as the fraction of CTC clusters or CTC cluster size distribution from blood samples?

Here, we tackle these open questions by integrating transcriptomic analysis and computational modeling. First, we apply three EMT scoring metrics to several tumor and CTC datasets; these scores, which correlate well with one another, demonstrate that the EMT traits of tumors and CTCs are highly heterogeneous, raising questions about how much can be predicted about the EMT score of CTCs from the primary tumor and vice versa. To further investigate this heterogeneity and interdependence of EMT in tumors and CTCs, we turn to *in silico* biophysical modeling that couples EMT in the primary tumor and cell migration. The model reveals different parameter regions in which CTCs can either be more mesenchymal or more epithelial than the primary tumor, depending on the rate of EMT

and migration dynamics (collective vs individual). Moreover, several CTC cluster size distribution datasets sampled from different tumors are mapped onto different parametric combinations in the model description, suggesting that the heterogeneous tumor-CTC EMT relation could be an important aspect *in vivo*. Finally, we integrate the EMT scoring and biophysical model in a single computational pipeline to investigate how much can be predicted by this biophysical model, in terms of tumor and CTC EMT score, CTC cluster fraction, and CTC cluster size distribution, when only one of these variables is provided as an input.

2 | MATERIALS AND METHODS

2.1 | Data analysis

We calculated EMT scores for multiple primary tumor and CTC datasets using three different EMT metrics – 76 gene signature (76GS), Kolmogorov-Smirnov (KS), and Multinomial Linear Regression (MLR) [13–15]. We also calculated correlations in the EMT scores of samples for a given dataset using Spearman's and Pearson's correlation coefficient values.

2.2 | EMT model

The biophysical model focuses on cells at the periphery of a tumor that have the potential to undergo EMT and migrate as CTCs individually or as a cluster. Therefore, cells in the model are arranged on a one-dimensional lattice that represents the tumor invading edge. Conversely, cells in the more internal layers are not modeled explicitly since they lack the physical space to migrate. Starting from an epithelial state, cells at the invading edge undergo transitions with a rate (k) through a number of intermediate hybrid E/M states, and eventually to a mesenchymal state.

Mesenchymal cells can migrate from the cell layer as individual cells; conversely, clusters of neighboring hybrid E/M cells can migrate together as multicellular units. The migration rate of clusters depends on 1. the number of cells in the cluster (i.e. the cluster size), 2. the EMT state of neighboring cells as cell-cell adhesion bonds must be broken, and 3. a migration cooperativity parameter (c) that quantifies the propensity to collective migration. While a low c favors individual migration, a large c promotes clustered migration if hybrid E/M cells are in contact. The migration is simply modeled as a discrete event, and

the physical motion of the migrating cells is not considered explicitly. When a single cell or cluster migrates, cells are instantaneously replaced by new epithelial cells. This process, which ensures a constant number of cells in the invading layer, considers the emergence of interior cells that become exposed to EMT-inducing signals once peripheral cells migrate.

The dynamics of cell fractions with a generic number (N) of hybrid E/M states are described by a set of ordinary differential equations:

$$\frac{d\rho_E}{dt} = -k\rho_E + \Theta(\rho_{H_1}, \rho_{H_2}, \dots, \rho_{H_N}) + \rho_M \quad (1)$$

$$\frac{d\rho_{H_i}}{dt} = +k\rho_{H_{i-1}} - k\rho_{H_i} - \frac{\rho_{H_i}}{\sum_{j=0}^N \rho_{H_j}} \Theta(\rho_{H_1}, \rho_{H_2}, \dots, \rho_{H_N}) \quad (2)$$

$$\frac{d\rho_M}{dt} = +k\rho_{H_N} - \rho_M \quad (3)$$

In Eqs 1–3, k is an EMT rate in arbitrary, dimensionless units. Eq. 2 represents the dynamics of the i -th hybrid E/M state cell population, which is increased due to transitions from the $i-1$ state ($+k\rho_{H_{i-1}}$) and decreased by transitions to the $i+1$ state ($-k\rho_{H_i}$). The supplementary section (SI section 1.1) shows the form of the equations for the case of three intermediate states (E-like, E/M, and M-like) that is used throughout most of the paper. Furthermore, the term $\Theta(\rho_{H_1}, \rho_{H_2}, \dots, \rho_{H_N})$ quantifies the loss of hybrid E/M cells due to migration out of the lattice (either as single cells or multicellular clusters); the explicit form of this term is derived in the supplementary information (SI section 1.2–1.5). The mesenchymal cell fraction (eq. 3) increases due to transition from the terminal hybrid E/M state ($+k\rho_{H_N}$) and decreases due to single cell migration ($-\rho_M$). Since migrating cells are replaced by new epithelial cells, the epithelial cell fraction (eq. 1) increased via an influx equal to the migration loss ($+\Theta(\rho_{H_1}, \rho_{H_2}, \dots, \rho_{H_N}) + \rho_M$). It is worth noting that all migration rates in Eqs 1–3 are not explicitly multiplied by a migration rate constant; since the model is dimensionless, time is rescaled so that this parameter is equal to 1. Therefore, the EMT rate (k) can be thought of as a ratio between the EMT rate and cell migration rate.

Custom-made python scripts to solve the model and reproduce the results are freely available at <https://github.com/federicobocci/Biophysical-model-of-EMT-heterogeneity>.

3 | RESULTS

3.1 | EMT scoring metrics analysis reveals heterogeneity in primary tumors and CTCs across cancer types

Recent approaches have investigated the varying degrees of EMT in CTCs using a handful of markers and their association with patient survival across cancer types [16–20]. Further, there has been a surge of high-throughput measurement such as RNA-seq of CTCs [21–23] as well as primary tumor [24–26], including investigations at a single-cell level. Given the heterogeneity of assessing EMT in multiple studies using diverse markers [27], such transcriptomics-based measurements can enable quantifying EMT in a more systematic manner using different scoring methods.

We calculated the EMT scores of multiple publicly available datasets associated with CTCs, using three different EMT metrics: 76GS, KS, and MLR [28]. These three methods use different gene lists and algorithms and have been developed based on pan-cancer signatures of EMT identified from preclinical (in vitro) and/or clinical data [13–15]. Thus, these scores can indicate the extent of EMT in a cell line, primary/metastatic tumor or CTC has undergone.

The more epithelial a sample is, the lower its KS score (on a scale of $[-1, 1]$) or MLR score (on a scale of $[0, 2]$) and the higher is its 76GS score (no a priori defined range of values). Thus, for a given dataset, while we expected a positive correlation between MLR and KS scores, we expected 76GS scores to correlate negatively with KS and MLR ones.

EMT scores of four breast cancer CTC cell lines and their metastatic variants [29] (GSE112855) using the above-mentioned three metrics, displayed heterogeneity in their EMT-ness (Figures 1A, S1A–E, S2A–C). However, none of these cell/sub lines could be classified as strongly mesenchymal, based on scores across the three metrics. Next, CTCs from various breast cancer patients exhibited heterogeneity [30] (PRJNA471754); however, the samples were overall shifted toward a more mesenchymal end of the spectrum as compared to breast cancer cell lines (GSE112855) (Figures 1B, S1F–J). Further, we examined the EMT-ness of individual CTCs and clusters of CTCs isolated from xenograft models as well as breast cancer patients [31] (GSE111065). Interestingly, while the EMT-ness of individual CTCs varied more along a spectrum, the EMT scores of CTC clusters followed a more bimodal distribution with a large difference in corresponding EMT-ness (Figures 1C and 1D, S2D–I, S3A–J). Put together, these results suggested that CTCs either freshly isolated from patients or established in culture as cell lines showed considerable heterogeneity in their EMT scores as assessed via these three independent EMT metrics. Interestingly,

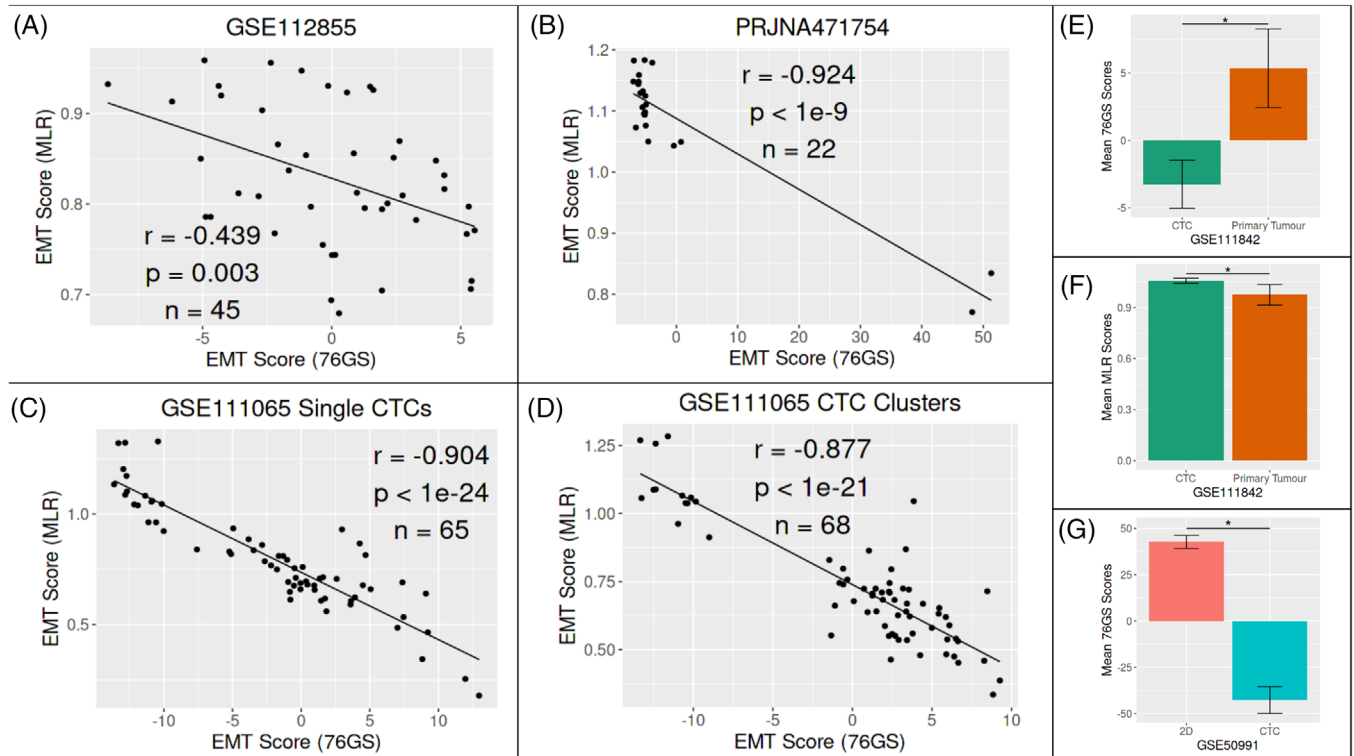


FIGURE 1 Heterogeneity in EMT scores of CTCs and primary tumors, calculated from corresponding transcriptomic data (publicly available microarray or RNA data). (A) Correlation between 76GS and MLR scores for GSE112855 (CTC-derived cell lines/sub lines in breast cancer) dataset. Each dot shows a sample in the dataset. Pearson's correlation values (r , p) are reported. n = number of samples. (B) Same as (A) but for individual CTCs in PRJNA471754. (C and D) Same as (A) but for single CTCs and CTC clusters separately in GSE111065. (E and F) Comparison of EMT scores (KS, MLR) for CTC and primary tumor for GSE111842 ($*p < 0.0001$). (G) Comparison of 76GS EMT scores for GSE50991 for 2D culture versus 4D culture of CTCs isolated from A549 lung cancer cells ($*p < 0.0001$). Students' two-tailed t -test was used in panels E, F, G

the EMT status of CTCs was also found to be different depending on cultured in petridish (2D) versus in a 4D ex vivo model for a lung cancer cell line [32] (GSE50991) (Figure 1G).

We also calculated the EMT scores of CTCs isolated from 16 patients in Stage II-III breast cancer and primary tumor available from 12 of them [33], and found CTCs to be relatively more mesenchymal than primary tumors (GSE111842) (Figures 1E and 1F). This observation prompted us to interrogate if the EMT status of primary tumor and CTCs can be informative of one another, that is can one predict the EMT status of primary tumor based on that of CTCs or vice-versa?

3.2 | A biophysical model to investigate EMT heterogeneity

Transcriptomic analysis of tumor samples and CTCs highlighted heterogeneity in EMT scoring. Therefore, we further sought to understand the relationship between EMT-ness of primary tumor and CTCs using a simple

coarse-grained biophysical model that couples phenotypic transitions driven by EMT with cell migration in the primary tumor [34].

In this model, cells are arranged on a lattice that represents the invading edge of a tumor. Based on the heterogeneity in EMT scores observed in CTCs, we considered a model with three intermediate states, E-like, E/M, M-like that are progressively less epithelial and more mesenchymal, in addition to the "pure" epithelial and mesenchymal ones (Figure 2A, top). Moreover, cells undergoing EMT can migrate from the lattice. While mesenchymal cells are assumed to migrate only as single cells, hybrid E/M cells in the E-like, E/M, and M-like states can migrate together as multicellular clusters if in spatial proximity (Figure 2A, bottom). The output of the model is the steady state fractions of cells in various EMT phenotypes in the tumor (E, E-like, E/M, M-like, M) as a function of the main model's two parameters: the EMT rate (k), which describes the speed of EMT transitions and migration cooperativity (c), which describes the propensity of hybrid E/M cells to migrate together as multicellular clusters.

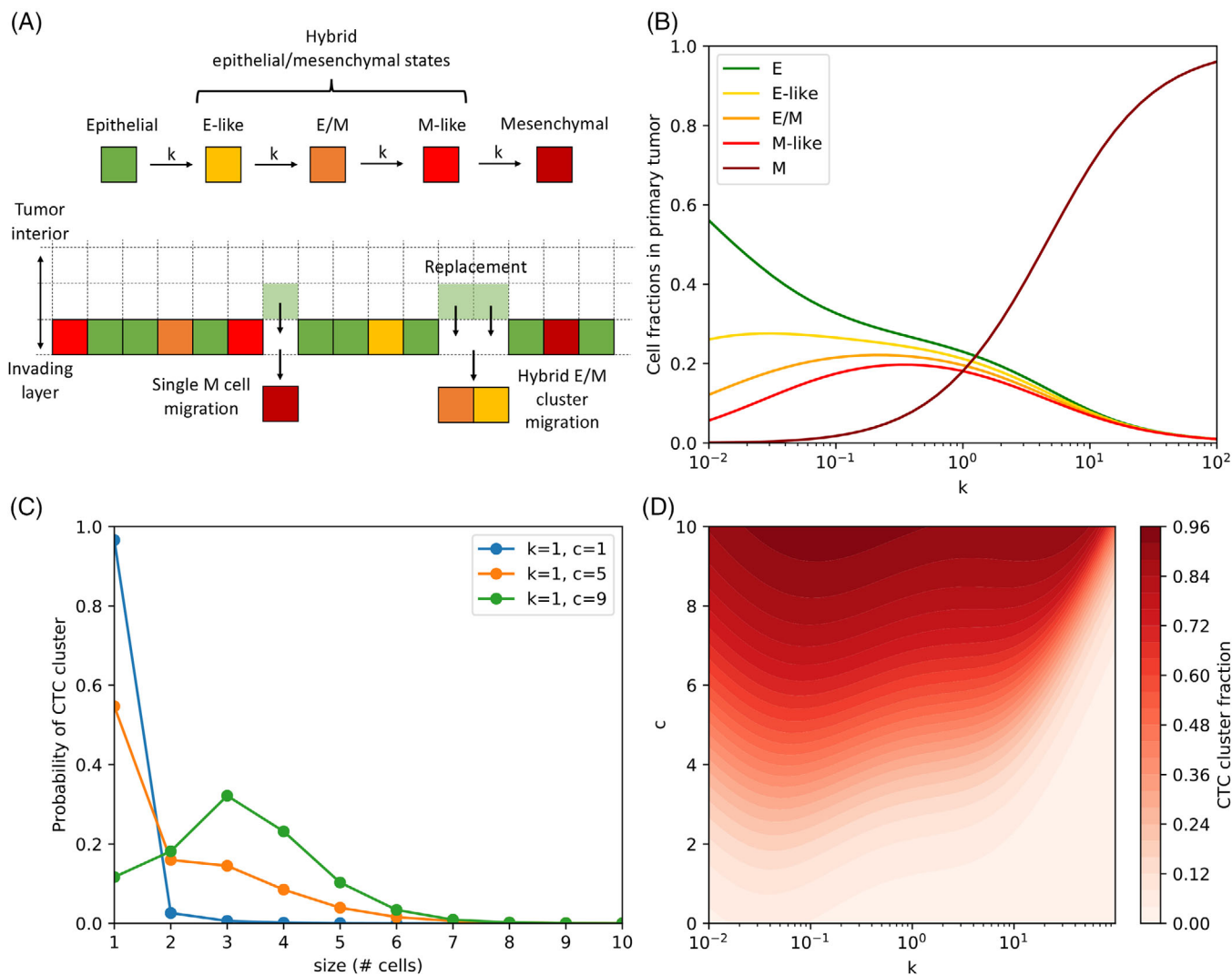


FIGURE 2 Single cell-cluster migration transition in the EMT-migration model. (A) Overview of the EMT-migration model. Top: cells undergo transitions through three hybrid E/M states with rate k . Bottom: mesenchymal cells migrate as single cells, while hybrid E/M cells can migrate together as multicellular clusters. Migrating cells are replaced by new Epithelial cells. (B) Fraction of Epithelial, E-like, E/M, M-like and Mesenchymal cells in the primary tumor as a function of the EMT rate (k) for a fixed value of $c = 2.5$. (C) Three predicted CTC cluster size distributions for a fixed EMT rate ($k = 1$) and increasing migration cooperativity (c). (D) Fraction of CTC clusters as a function of EMT rate (k) and migration cooperativity (c)

The relaxation to steady state depends on two timescales: the EMT rate (k) and the migration rate, which is fixed to unit value and never varied in the model (Methods – EMT model). Starting from an initially epithelial lattice, the model relaxes to its steady state with a speed that depends on the parameter choice (Figure S4). Considering that the timescales for EMT and cell migration are fast compared to the typical timescales of tumor progression [35], from now on we will always analyze the model at steady state.

We observed that varying the rate of EMT (k) at fixed cooperativity (c) can change the steady state fraction of cells with distinct EMT phenotypes in the tumor lattice: while cells are mostly epithelial for low values of k , more

and more cells convert to the E-like, E/M, and M-like states as the value of k increases. Eventually, most cells are mesenchymal for large k (Figure 2B). Moreover, fixing k and increasing c lead to a change in the migration strategy. At low c , the size distribution of escaping CTCs is dominated by single cells; conversely, larger values of c lead to an increased probability to observe multicellular clusters of two or more cells (Figure 2C). More generally, varying both parameters trigger a transition from single cell migration to clustered cell migration, as seen for instance by inspecting the fraction of escape event of clusters of two or more cells, or cluster escape fraction (Figure 2D), as well as changes in the steady state cell fractions (Figure S5).

Next, we investigated the implications of this model in the context of tumor-CTC EMT score relationship to provide a rationale for the observed EMT score heterogeneity in tumor and CTCs.

3.3 | Collective cell migration can lead to non-trivial tumor-CTC score relationship

To investigate the relation between EMT progression in the primary tumor and CTCs, we define an EMT scoring metric in the biophysical model that can be directly compared to the scores computed in cancer datasets. Specifically, we define a metric (S) that can be directly compared to the MLR score. This choice is motivated by the observation that unlike the 76GS and KS metrics [13,14], the MLR score specifically focused on identifying a hybrid E/M signature [15]. Mimicking the MLR score, cells on the EMT spectrum are assigned weights ranging from 0 (epithelial) to 2 (mesenchymal). Thus, for a five-state model, the weights for E, E-like, E/M, M-like, and M states are $(w_E, w_{H1}, w_{H2}, w_{H3}, w_M) = (0, 0.5, 1, 1.5, 2)$. The tumor EMT score is defined as a weighted sum of the steady state fractions of cells: $S_T = w_E \rho_E + w_{H1} \rho_{H1} + w_{H2} \rho_{H2} + w_{H3} \rho_{H3} + w_M \rho_M$. Similarly, a score can be defined for the CTCs by considering the fractions of migrating cells with different EMT phenotypes $(\varphi_{H1}, \varphi_{H2}, \varphi_{H3}, \varphi_M)$: $S_{CTC} = w_{H1} \varphi_{H1} + w_{H2} \varphi_{H2} + w_{H3} \varphi_{H3} + w_M \varphi_M$. Notably, since E cells cannot migrate in our model formulation, S_{CTC} only considers three hybrid states and the mesenchymal state.

The tumor score (S_T) is very close to zero (i.e. strongly epithelial) for models with low EMT rate ($k \ll 1$, Figure 3A, left), and continuously increases to hybrid E/M and mesenchymal for increasing k (Figure 3A, left to right). Interestingly, a larger migration cooperativity (c) increases the propensity of hybrid E/M cells to undergo cluster-based migration before transitioning to a mesenchymal state, thus decreasing the tumor score (Figure 3A, bottom to top). Similarly, the CTC score (S_{CTC}) increases with k because more cells undergo a complete EMT before migrating and decreases with c because cells tend to migrate collectively as hybrid E/M rather than as single mesenchymal cells (Figure 3B). For most (k, c) parameter combinations, CTCs have a larger EMT score than the tumor (Figure 3C, red-shaded region). Strikingly, a condition of fast EMT and high migration cooperativity (i.e. large k, c) gives rise to a switch where $S_{CTC} < S_T$ (Figure 3C, blue-shaded region). Overall, the relationship between S_T and S_{CTC} is not a fixed one but depends on (k, c) , that is more heterogeneous, thus making it difficult to predict one from the another (Figure 3D). Generally speak-

ing, CTCs are more mesenchymal than primary tumor in models with low EMT rate; conversely, CTCs can either be more or less mesenchymal than primary tumor at high EMT rate depending on the value of the migration cooperativity (c). This dependence on the model's parameters can be observed in models with variable number of intermediate states, indicating that it represents a robust feature (Figure S6). From a clinical standpoint, this observation underscores the difficulty to fully characterize an invading tumor in terms of EMT when only considering samples from primary tumor or vice versa. These results suggest that different cancer types, subtypes, and potentially even patients, might lie in different regions of the score diagram shown below.

3.4 | Analysis of CTC cluster size distribution reveals variability of tumor and CTC scores across cancer types

Our EMT biophysical model predicts a heterogeneous relationship between the EMT score of primary tumors and that of CTCs. To investigate this prediction, we analyze several CTC cluster size distributions obtained experimentally through the lens of our model. In these datasets, which were obtained from different cancer types, single CTCs and CTC clusters were isolated with various techniques to obtain a frequency count or probability to observe CTC clusters with variable number of cells [36–42]. Specifically, we consider eight separate datasets isolated from either mouse models or patients from melanoma, glioblastoma, myeloma, ovarian, prostate, and breast cancer [36–42]. By fitting the model's CTC size distribution to the experimental distributions, we identify the parameter combinations (k, c) that can best fit corresponding experimental data (SI section 1.6). Plotting the position of the datasets on the (k, c) plane against the model's score diagram highlights a striking heterogeneity in terms of positioning of these datasets (Figure 4A). In three datasets measured from melanoma, ovarian, and prostate cancer, the CTCs are predicted to be considerably more mesenchymal than their corresponding tumor. However, in three other datasets from breast cancer and glioblastoma, this difference is less pronounced. Finally, two datasets from breast cancer and myeloma, respectively, are predicted to fall into the "inversion" region where CTCs are less mesenchymal than the tumor (Figure 4B). Therefore, the model predicts that the association between EMT-ness in a primary tumor and CTCs can depend on the tumor type. Intriguingly, even within same tumor, there seems to be no generic trend in terms of EMT scores of CTCs and primary tumors, as seen in the three datasets all from breast cancer models.

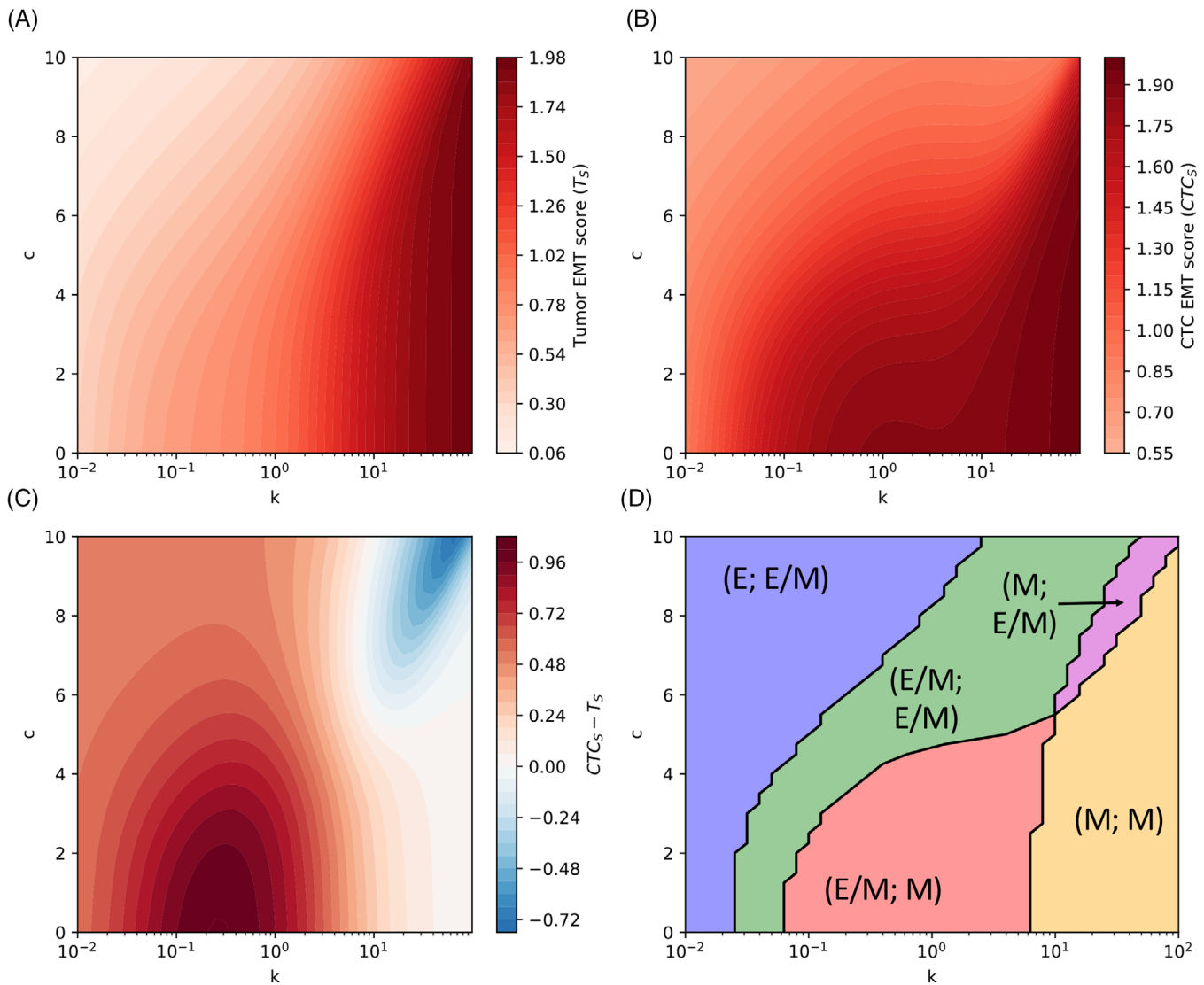


FIGURE 3 Inversions in tumor and CTC EMT scoring metrics. Predicted EMT score of primary tumor T_S (A), migrating cells, defined as both single CTCs and CTC clusters CTC_S (B), and difference between them $CTC_S - T_S$ (C) as a function of EMT rate (k) and migration cooperativity (c). (D) Possible combinations of EMT scores for tumor and CTCs. The scores are classified as epithelial ($s < 0.5$), hybrid E/M ($0.5 < s < 1.5$) or mesenchymal ($s > 1.5$). For instance, (E, E/M) indicates that the tumor has an epithelial score and the CTCs have a hybrid E/M score, and so forth

Adding to this complexity, some of these CTC cluster size distributions consider patients with different treatment regimes. For instance, an ovarian cancer dataset (Meunier and collaborators [37]) includes patients pre- and post-chemotherapy; a prostate cancer dataset (Kozminsky and collaborators [38]) contains data from patients exposed to several different hormone therapies; and a glioblastoma dataset includes patients treated with a microtubule inhibitor (Krol and collaborators [40]). Therefore, responses to different drugs could potentially represent an additional axis of variability in the EMT score relationship between tumors and CTCs.

3.5 | Exploring the predictive power of several measurements that quantify tumor aggressiveness

Motivated by the non-trivial relation between EMT scores of tumor and CTCs predicted by the model, we reviewed various types of measurements typically used to estimate tumor aggressiveness. These include EMT scores of tumors and CTCs that can be computed from single cell gene expression measurements, CTC cluster fraction in circulation, and full CTC cluster size distribution. We constraint the model with each of these assays to investigate how

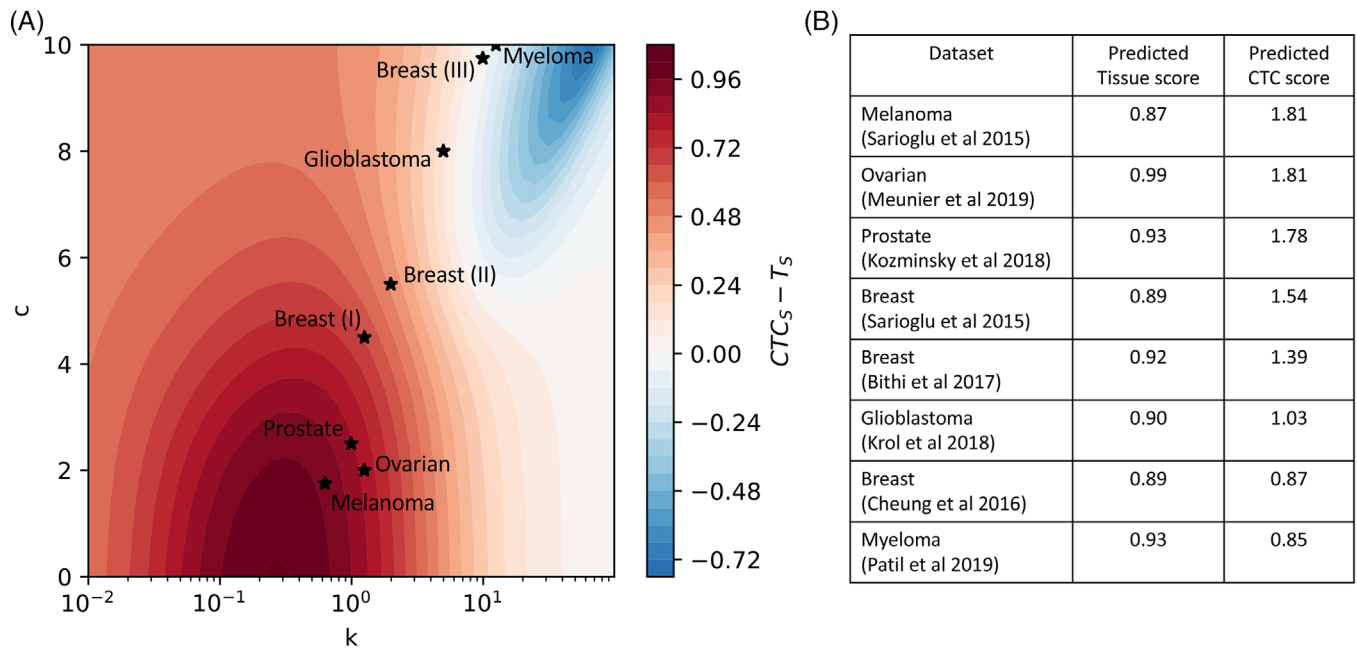


FIGURE 4 Heterogeneous tumor-CTC EMT score relationships in several CTC size distributions across cancer types. (A) best model fit on the (k,c) parameter space for various CTC cluster size distributions [36–42]. Each starred dot represents a CTC cluster size distribution that was fitted with the five-state model; the x- and y-coordinates indicate the (k,c) parameter combination yielding the best fit defined in terms of minimal root square distance between experimental distribution and model prediction. (B) predicted tumor and CTC scores based on model's fit

many of the other variables can be predicted using our model (Figure 5A).

First, we consider the case where the EMT score of the tumor is known; this is a typical scenario if gene expression from a tumor sample is analyzed, either at a single-cell or bulk resolution. For instance, we calculate the MLR scores of several samples from a pancreatic ductal patient-derived xenograft model [43]; on average, the samples are predicted to be hybrid E/M (1.13 ± 0.14). In the model, constraining the tumor EMT score to a fixed value is equivalent to selecting a contour line on the two-dimensional (k, c) parameter space (Figure 5B, left). Each parameter combination along the curve corresponds to a model where the tumor EMT score equals the EMT score for the given dataset. Models along the contour line exhibit variable CTC EMT score and CTC cluster fraction, as seen by overlapping the contour line onto the CTC score diagram (Figure 5B, left). Thus, information about tumor EMT score only is not sufficient to predict neither the EMT scores of CTCs nor cluster size distribution of CTCs (Figure 5B, right).

Similarly, the CTC EMT score is not sufficient to fully constrain the model's parameters. For example, we find that the average MLR score of CTCs from a breast cancer dataset [30] falls within the hybrid E/M range (1.087 ± 0.101). The contour line at constant CTC EMT score, however, crosses parameter regions with variable tumor

EMT scores (Figure 5C, left). Therefore, for the given CTC EMT scores, it is possible that the tumor has either a lower score (i.e. more epithelial) or higher score (i.e. more mesenchymal) than the CTCs; moreover, it can have variable cell fractions too (Figure 5C, right). Moreover, three other CTC datasets from prostate, myeloma, and breast cancer [44–46] are mapped onto distinct model contour lines (Figure S7A).

Interestingly, measuring both tumor and CTC MLR score allows to identify a single (k, c) parameter combination. To illustrate this scenario, we consider a cohort of stage II-III breast cancer patients comprising RNA-sequencing of both primary tumor and CTCs [33]. The MLR metric predicts hybrid E/M signature for both tumor (0.97) and CTCs (1.09). In the model, fixing both scores corresponds to identifying two contour lines (Figure 5D, left); their intersection provides the (k, c) parameter combination that better reproduces the dataset. With a unique (k, c) combination, the model is able to predict a CTC cluster size distribution characterized by almost 60% of CTC clusters with two or more cells (Figure 5D, right). Unfortunately, the lack of information on the CTC cluster size distribution for this dataset prevents a comparison between model and experiment.

Another popular measurement is the overall fraction of CTC clusters, which can be obtained from blood samples by separating single CTCs and CTC clusters. Similar to the

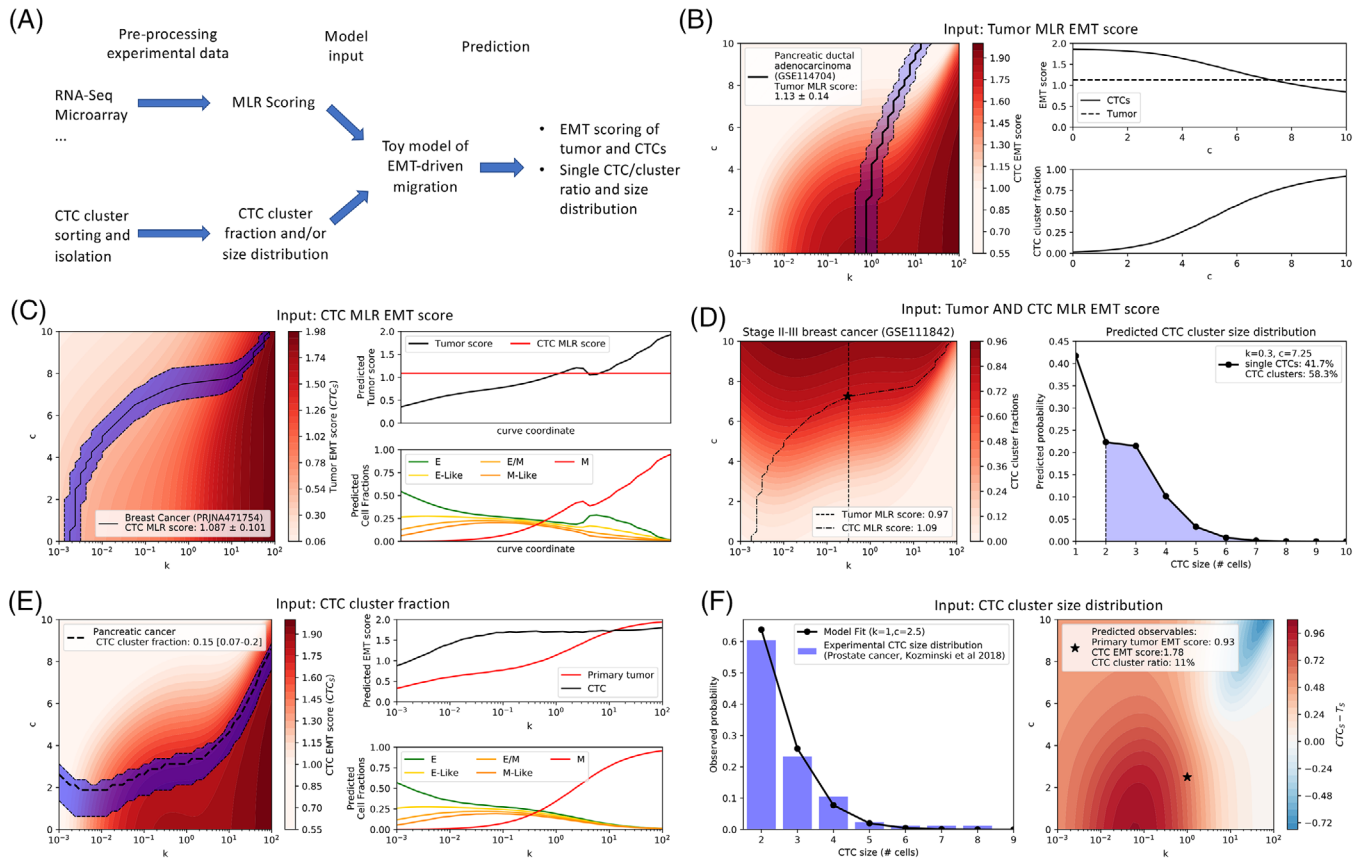


FIGURE 5 The model's predictive power for several popular measurements of cancer aggressiveness. (A) For single cell measurements of gene expression (RNA-seq, microarray, etc.), an EMT scoring technique such as MLR is used to provide an input to the model; conversely, for the case of CTC cluster isolation, the input is the CTC fraction or size distribution. (B) Left: The black line shows the contour line where the tumor EMT score is fixed and equal to the dataset's score (the blue shading shows the standard deviation of the measurement). Right: EMT scores (top) and CTC cluster fractions (bottom) moving along the contour line. (C) Left: The black line shows the contour line where the CTC EMT score is fixed and equal to the dataset's score (the blue shading shows the standard deviation of the measurement). Right: EMT scores (top) and cell fractions (bottom) moving along the contour line. (D) Left: The black lines show the contour line where the tumor and CTC EMT scores are fixed and equal to the dataset's scores. Right: Predicted CTC cluster size distribution and cluster fraction for the parameters identified in the left panel. (E) Left: The black line represents a model's contour line where the CTC cluster fraction equals the experimental measurement (the blue shading shows the extremal levels found in the experiment). Right: Models along the contour line exhibit variable EMT scores of tumor and CTCs (top) as well as variable cell phenotype fractions (bottom). (F) Left: Fitting a CTC cluster size distribution provides the optimal (k, c) parameter combination. Right: Predicted EMT scores and CTC cluster fraction from the model with optimal (k, c)

cases of tumor and CTC scores, however, a line of constant CTC fraction can be identified on the (k, c) parameter space (Figure 5E, left). For instance, CTC clusters isolated from a cohort of pancreatic cancer patients amount on average to 15% of the total CTCs [47]. Models along this CTC cluster fraction contour line, however, can exhibit variable tumor scores in the E, E/M, and M ranges, as well as E/M or M CTC cluster scores (Figure 5E, right). Similarly, four very different CTC cluster fraction measurements are mapped onto different contour lines (Figure S7B) [47–51].

Finally, the model's best fit MLR for a complete CTC cluster size distribution can identify a unique parameter combination (k, c), thus predicting the EMT score of tumor and CTCs. In this example, a distribution from prostate cancer

[38] is predicted to have a hybrid E/M tumor EMT score but a mesenchymal CTC EMT score (Figure 5F). Similar to the case of Figure 5D, where the CTC cluster distribution was predicted based on EMT scores, the lack of a complete dataset with knowledge of both EMT scores and cluster size distribution prevents a quantitative model validation.

Overall, our model is capable of characterizing a tumor in terms of its EMT-ness and the propensity to undergo collective cell migration when we are aware of either both the EMT score of tumor and CTCs or the full CTC cluster size distribution. Certainly, future experimental investigations capable to estimate all the three above-mentioned observables will offer a chance to test the model's prediction more quantitatively.

4 | DISCUSSION

Phenotypic heterogeneity in EMT is emerging as a hallmark of metastatic progression, and decoding its dynamics in a quantitative and predictive manner can lead to fundamental insights about metastasis [52]. Subsets of cells with varying EMT-ness (epithelial, hybrid E/M, and mesenchymal states) have been observed in primary tumors, CTCs, and metastases [9,11,27].

To quantify EMT phenotypic heterogeneity, we first analyzed data from multiple tumor and CTC samples with three different EMT scoring metrics that rely on different gene lists and algorithms [13–15]. Tumors and CTCs exhibited a strong EMT score heterogeneity, suggesting that there may not exist a specific region of the EMT spectrum that is uniquely associated with tumor progression and CTC migration.

Motivated by the variability across the EMT spectrum registered in tumor tissues and CTCs, we integrated the transcriptomic analysis with a mechanism-based biophysical model that couples EMT with cell migration [34]. This model predicts a heterogeneous relation between EMT status of primary tumor and CTCs; CTCs are more mesenchymal than the primary tumor in parameter regions where invasion is mostly carried by individual cells; a transition to multicellular, clustered cell migration, however, can give rise to CTCs equally or even less mesenchymal than their corresponding tumor. In other words, measuring the level of EMT progression in CTCs does not directly allow to predict the EMT phenotypic distribution of the corresponding tumor and vice versa. Specifically, in models with low EMT rate, CTCs are always more mesenchymal than primary tumor; conversely, in models with high EMT rate, CTCs can be either more or less mesenchymal than primary tumor based on the solitary or collective migration strategy, respectively. We have previously showed that high EMT rates describe well the EMT phenotype distribution of pre-treatment breast cancer patients, whereas lower EMT rates better describe the same patients after successful treatment [11,34]. Therefore, it is possible that the relation between EMT score of tumor and CTCs is not just specific to tumor type, but also evolves during cancer progression. This interesting prediction could be further explored with data from both tumor and CTCs at different stages of clinical treatment.

Nonetheless, we acknowledge multiple directions where the current model can be further improved. First, EMT transitions and migration are described with phenomenological parameters that are not explicitly connected with signaling and biophysical cellular processes. Developing models that more explicitly integrate the signaling and mechanical aspects of EMT and cell migration is a crucial

future challenge, which would make the modeling even more predictive [53]. Moreover, several assumptions were made about the dynamics of cancer cell migration in order to decrease the complexity of the model. First, it is assumed that only cells at the periphery of the tumor can undergo cell migration, even though it is possible for interior cells to intravasate [54,55]. A more detailed representation of the tumor structure could more correctly account for EMT heterogeneity not only at the tumor's invading edge, but also in more interior regions. Moreover, additional effects that could modulate migration and intravasation, such as cell death and/or breakup of multicellular clusters are not explicitly considered in the model [56]. Finally, EMP is not necessarily regulated in a cell-autonomous manner, but rather depends on communication with other cancer cells via contact-dependent signaling, such as Notch, or paracrine signaling, such as TGF- β [57,58]. *In silico* modeling of EMT and Notch underlying circuitry dynamics recently predicted that lateral induction between hybrid E/M cells can facilitate the formation of hybrid E/M multicellular clusters [59,60]. Similarly, reconstruction of TGF- β cell-cell communication networks recently suggested that hybrid E/M phenotypes can act as both senders and receivers of the signaling, thus facilitating EMT transitions in other cells [8]. Therefore, modeling the effect of cell-cell communication on EM plasticity and CTC cluster distributions could be another interesting future direction.

EMT is connected to multiple axes of cancer progression, including invasion, stemness, and immune response [61]. How the heterogeneity along the EMT axis propagates, and in turn depends, on other hallmarks of cancer progression, remains largely unknown. Future investigations through high-throughput single cell techniques and computational modeling will help answer these questions and identify the defining principles of dynamics of EMP.

5 | CONCLUSIONS

As our ability to perform quantitative measurements at the level of cell migration and single-cell gene expression during metastasis increases, we need to integrate these biochemical and biophysical aspects to decipher the hallmarks of metastasis-initiating cells. While these techniques enable us to inspect EMP at unprecedented resolution, how the EMP traits of tumors propagates to those of migrating CTCs remains poorly understood. Overall, our integrated transcriptomic and computational pipeline highlights that the relationship between "EMT-ness" in primary tumors and CTCs is likely heterogeneous, underscoring the need for a broader and multi-faceted approach to characterize tumor aggressiveness in the clinic.

FUNDING INFORMATION

This work was supported by Ramanujan Fellowship awarded to Mohit Kumar Jolly by Science and Engineering Research Board (SERB), DST, Government of India (SB/S2/RJN-049/2018). Federico Bocci was supported by the Center for Theoretical Biological Physics sponsored by the NSF (grant number: PHY-2019745), by the NSF grant number DMS1763272 and a grant from the Simons Foundation (grant number: 594598, QN).

ACKNOWLEDGMENT

We would like to thank Prof. Josè Onuchic (Rice University) for sharing his useful suggestions about the manuscript.

AUTHOR CONTRIBUTIONS

Federico Bocci and Mohit Kumar Jolly designed the research, Susmita Mandal and Tanishq Tejaswi performed EMT score calculations, Federico Bocci developed *in silico* modeling, Federico Bocci and Mohit Kumar Jolly wrote the manuscript, and Mohit Kumar Jolly supervised the research.

ORCID

Federico Bocci  <https://orcid.org/0000-0003-4302-9906>

REFERENCES

- Brabletz T, Kalluri R, Nieto MA, Weinberg RA. *EMT in cancer*. *Nat Rev Cancer*. 2018;**18**:128–134.
- Nieto MA, Huang RY, Jackson RA, Thiery JP. *EMT*: 2016. *Cell*. 2016;**166**(2016):21–45.
- Pastushenko I, Brisebarre A, Sifrim A, et al. *Identification of the tumour transition states occurring during EMT*. *Nature*. 2018;**556**:463–468.
- McFaline-Figueroa JL, Hill AJ, Qiu X, Jackson D, Shendure J, Trapnell C. *A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition*. *Nat Genet*. 2019;**51**:1389–1398.
- Tian XJ, Zhang H, Xing J. *Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition*. *Biophys J*. 2013;**105**(4):1079–1089.
- Lu M, Jolly MK, Levine H, Onuchic JN, Ben-Jacob E. *MicroRNA-based regulation of epithelial-hybrid-mesenchymal fate determination*. *Proc Natl Acad Sci*. 2013;**110**(45):18174–18179.
- Sha Y, Wang S, Zhou P, Nie Q. *Inference and multiscale model of epithelial-to-mesenchymal transition via single-cell transcriptomic data*. *Nucleic Acids Res*. 2020;**48**(17):9505–9520.
- Sha Y, Wang S, Bocci F, Zhou P, Nie Q. *Inference of Intercellular Communications and Multilayer Gene-Regulations of Epithelial–Mesenchymal Transition From Single-Cell Transcriptomic Data*. *Front Genet*. 2021;**11**:604585.
- Liu S, Cong Y, Wang D, et al. *Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts*. *Stem Cell Reports*. 2014;**2**(1):78–91.
- Bocci F, Gaerhart-Serna L, Ribeiro M, et al. *Towards understanding Cancer Stem Cell heterogeneity in the tumor microenvironment*. *Proc Natl Acad Sci*. 2019;**116**(1):148–157.
- Yu M, Bardia A, Wittner BS, et al. *Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition*. *Science* (80-). 2013;**339**(6119):580–584.
- Jolly MK, Mani SA, Levine H. *Hybrid epithelial/mesenchymal phenotype(s): the “fittest” for metastasis?* *BBA - Rev Cancer*. 2018;**1870**(2):151–157.
- Byers LA, Diao L, Wang J, et al. *An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance*. *Clin Cancer Res*. 2013;**19**(1):279–90.
- Tan TZ, Miow QH, Miki Y, et al. *Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients*. *EMBO Mol Med*. 2014;**6**(10):1279–1293.
- George JT, Jolly MK, Xu S, Somarelli JA, Levine H. *Survival Outcomes in Cancer Patients Predicted by a Partial EMT Gene Expression Scoring Metric*. *Cancer Res*. 2017;**77**(22):6415–6428.
- Zhao R, Cai Z, Li S, et al. *Expression and clinical relevance of epithelial and mesenchymal markers in circulating tumor cells from colorectal cancer*. *Oncotarget*. 2017;**8**:9293–9302.
- Genna A, Vanwynsberghe AM, Villard A V., et al. *EMT-Associated Heterogeneity in Circulating Tumor Cells: Sticky Friends on the Road to Metastasis*. *Cancers (Basel)*. 2020;**12**(6):1632.
- Lecharpentier A, Vielh P, Perez-Moreno P, Planchard D, Soria JC, Farace F. *Detection of circulating tumour cells with a hybrid (epithelial/mesenchymal) phenotype in patients with metastatic non-small cell lung cancer*. *Br J Cancer*. 2011;**105**(9):1338–1341.
- Saxena K, Subbalakshmi AR, Jolly MK. *Phenotypic heterogeneity in circulating tumor cells and its prognostic value in metastasis and overall survival*. *EBioMedicine*. 2019;**46**:4–5.
- De T, Goyal S, Balachander G, et al. *A Novel Ex Vivo System Using 3D Polymer Scaffold to Culture Circulating Tumor Cells from Breast Cancer Patients Exhibits Dynamic E-M Phenotypes*. *J Clin Med*. 2019;**8**(9):1473.
- Yu M, Ting DT, Stott SL, et al. *RNA sequencing of pancreatic circulating tumour cells implicates WNT signaling in metastasis*. *Nature*. 2012;**487**(7408):510–513.
- Elisabetta Rossi RZ. *Single-Cell Analysis of Circulating Tumor Cells: How Far Have We Come in the -Omics Era?* *Front Genet*. 2019;**10**:958.
- Lei Zhao, Xiaohong Wu, Tong Li, Jian Luo DD. *ctcRbase: the gene expression database of circulating tumor cells and microemboli*. *Database*. 2020;**2020**:baaa020.
- Puram S V, Tirosh I, Parkh AS, et al. *Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer*. *Cell*. 2017;**171**(7):1611–1624.
- Chen Y-C, Sahoo S, Brien R, et al. *Single-cell RNA-sequencing of migratory breast cancer cells: discovering genes associated with cancer metastasis*. *Analyst*. 2019;**144**:7296–7309.
- Lourenco AR, Ban Y, Crowley MJ, et al. *Differential Contributions of Pre- and Post-EMT Tumor Cells in Breast Cancer Metastasis*. *Cancer Res*. 2020;**80**(2):163–169.
- Jolly MK, Somarelli JA, Sheth M, et al. *Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas*. *Pharmacol Ther*. 2019;**194**(161–184).

28. Chakraborty P, George JT, Tripathi S, Levine H, Jolly MK. *Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum*. *Front Bioengineering Biotechnol.* 2020;**8**:220.
29. Klotz R, Thomas A, Teng T, et al. *Circulating Tumor Cells Exhibit Metastatic Tropism and Reveal Brain Metastasis Drivers*. *Cancer Discov.* 2020;**10**(1):86–103.
30. Cheng Y-H, Chen Y-C, Lin E, et al. *Hydro-Seq enables contamination-free high-throughput single-cell RNA-sequencing for circulating tumor cells*. *Nat Commun.* 2019;**19**:2163.
31. Gkountela S, Castro-Giner F, Szczerba BM, et al. *Circulating Tumor Cell Clustering Shapes DNA Methylation to Enable Metastasis Seeding*. *Cell.* 2019;**176**(1-2):98–112.
32. Mishra DK, Creighton CJ, Zhang Y, Chen F, Thrall MJ, Kim MP. *Ex vivo four-dimensional lung cancer model mimics metastasis*. *Ann Thorac Surg.* 2015;**99**(4):1149–1156.
33. Lang JE, Ring A, Porras T, et al. *RNA-Seq of Circulating Tumor Cells in Stage II-III Breast Cancer*. *Breast Oncol.* 2018;**25**:2261–2270.
34. Bocci F, Jolly MK, Onuchic JN. *A biophysical model uncovers the size distribution of circulating Tumor Cell Clusters across cancer types*. *Cancer Res.* 2019;**79**(21):5527–5535.
35. Tripathi S, Chakraborty P, Levine H, Jolly MK. *A mechanism for epithelial-mesenchymal heterogeneity in a population of cancer cells*. *PLoS Comput Biol.* 2020;**16**(2):e1007619.
36. Sarioglu AF, Aceto N, Kojic N, et al. *A microfluidic device for label-free, physical capture of circulating tumor cell clusters*. *Nat Methods.* 2015;**12**(7):685–691.
37. Meunier A, Hernández-Castro, Alejandro J, et al. *Gravity-based microfiltration reveals unexpected prevalence of circulating tumor cell clusters in ovarian cancer*. *bioRxiv.* 2019:773507.
38. Kozminsky M, Fouladdel S, Chung J-S, et al. *Detection of CTC Clusters and a Dedifferentiated RNA-Expression Survival Signature in Prostate Cancer*. *Adv Sci.* 2018:1801254.
39. Bithi SS, Vanapalli SA. *Microfluidic cell isolation technology for drug testing of single tumor cells and their clusters*. *Sci Rep.* 2017;**7**:41707.
40. Krol I, Castro-Giner F, Maurer M, et al. *Detection of circulating tumour cell clusters in human glioblastoma*. *Br J Cancer.* 2018;**119**:487–491.
41. Cheung KJ, Padmanaban V, Silvestri V, et al. *Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters*. *Proc Natl Acad Sci.* 2016;**113**(7):E854–E863.
42. Patil R, Tan X, Bartosik P, et al. *In Vivo Monitoring of Rare Circulating Tumor Cell and Cluster Dissemination in a Multiple Myeloma Xenograft Model*. *J Biomed Opt.* 2019;**24**(8):085004.
43. Dimitrov-Markov S, Perales-Patón J, Bockorny B, et al. *Discovery of New Targets to Control Metastasis in Pancreatic Cancer by Single-cell Transcriptomics Analysis of Circulating Tumor Cells*. *Mol Cancer Ther.* 2020;**19**(8):1751–60.
44. Liu Y-L, Horning AM, Lieberman B, et al. *Spatial EGFR Dynamics and Metastatic Phenotypes Modulated by Upregulated EphB2 and Src Pathways in Advanced Prostate Cancer*. *Cancers (Basel).* 2019;**11**(12):1910.
45. Juan-Jose Garcés M, Simicek C, Vicari M, et al. *Transcriptional profiling of circulating tumor cells in multiple myeloma: a new model to understand disease dissemination*. *Leukemia.* 2020;**34**(2):589–603.
46. Iyer A, Gupta K, Sharma S, et al. *Integrative Analysis and Machine Learning based Characterization of Single Circulating Tumor Cells*. *J Clin Med.* 2020;**9**(4):1206.
47. Maddipati R, Stanger BZ. *Pancreatic cancer metastases harbor evidence of polyclonality*. *Cancer Discov.* 2015;**5**(10):1086–1097.
48. Aceto N, Bardia A, Miyamoto DT, et al. *Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis*. *Cell.* 2014;**158**(5):1110–1122.
49. Krebs MG, Hou JM, Sloane R, et al. *Analysis of circulating tumor cells in patients with non-small cell lung cancer using epithelial marker-dependent and independent approaches*. *J Thorac Oncol.* 2012;**7**(2):306–315.
50. Molnar B, Ladanyi A, Tanko L, Sreter L, Tulassay Z. *Circulating Tumor Cell Clusters in the Peripheral Blood of Colorectal Cancer Patients*. *Clin Cancer Res.* 2001;**7**:4080–4085.
51. Kulasinghe A, Schmidt H, Perry C, et al. *A Collective Route to Head and Neck Cancer Metastasis*. *Sci Rep.* 2018;**8**(1):746.
52. Mohit Kumar Jolly TC-T. *Dynamics of Phenotypic Heterogeneity Associated with EMT and Stemness during Cancer Progression*. *J Clin Med.* 2019;**25**(8):1542.
53. Yang Y, Jolly MK, Levine H. *Computational Modeling of Collective Cell Migration: Mechanical and Biochemical Aspects*. *Adv Exp Med Biol.* 2019;**1146**:1–11.
54. Harney AS, Arwert EN, Entenberg D, et al. *Real-Time Imaging Reveals Local, Transient Vascular Permeability, and Tumor Cell Intravasation Stimulated by TIE2hi Macrophage-Derived VEGFA*. *Cancer Discov.* 2015;**5**(9):2159–8290.
55. Elena I Deryugina WBK. *Intratumoral Cancer Cell Intravasation Can Occur Independent of Invasion into the Adjacent Stroma*. *Cell Rep.* 2017;**19**(3):601–616.
56. Douglas S, Micalizzi, Shyamala Maheswaran DAH. *A conduit to metastasis: circulating tumor cell biology*. *Genes Dev.* 2017;**31**:1827–1840.
57. Bocci F, Onuchic JN, Jolly MK. *Understanding the principles of pattern formation driven by Notch signaling by integrating experiments and theoretical models*. *Front Physiol.* 2020;**11**:929.
58. Hao Y, Baker D, Dijke P ten. *TGF- β -Mediated Epithelial-Mesenchymal Transition and Cancer Metastasis*. *Int J Mol Sci.* 2019;**20**(11):2726.
59. Boareto M, Jolly MK, Goldman A, et al. *Notch-Jagged signalling can give rise to clusters of cells exhibiting a hybrid epithelial/mesenchymal phenotype*. *J R Soc Interface.* 2016;**13**(118):20151106.
60. Bocci F, Jolly MK, Tripathi SC, et al. *Numb prevents a complete epithelial – mesenchymal transition by modulating Notch signalling*. *J R Soc Interface.* 2017;**14**(136):20170512.
61. Jia D, Li X, Bocci F, et al. *Quantifying cancer epithelial-mesenchymal plasticity and its association with stemness and immune response*. *J Clin Med.* 2019;**8**(5):725.

How to cite this article: Bocci F, Mandal S, Tejaswi T, Jolly MK. Investigating epithelial-mesenchymal heterogeneity of tumors and circulating tumor cells with transcriptomic analysis and biophysical modeling. *Comp Sys Onco.* 2021;1:e1015. <https://doi.org/10.1002/cso2.1015>