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

Running title: EMT heterogeneity of tumors and CTCs

Federico Bocci[1,#,a], Susmita Mandal[2], Tanishq Tejaswi[2,3], Mohit Kumar Jolly[2,#]

[1]Center for Theoretical Biological Physics, Rice University, Houston, TX; [2]Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore, India; [3]UG Programme, Indian Institute of Science, Bangalore, India;
[a]Current address: NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, CA

[#]Authors to whom correspondence may be addressed: fbocci@uci.edu (F.B.), mkjolly@iisc.ac.in (M.K.J.)

Abstract

Introduction: 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 a relationship between EMP associated heterogeneity of tumors and that of CTCs, we integrated transcriptomic analysis and biophysical modeling.
Methods: We apply three EMT (Epithelial-Mesenchymal Transition) 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.
Results: 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.
Conclusions: By integrating analysis of single cell gene expression and *in silico* modeling, 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: epithelial-mesenchymal transition, epithelial-mesenchymal plasticity, cancer, circulating tumor cells, RNA-sequencing

Introduction

Epithelial-Mesenchymal Transition (EMT) (now increasingly referred to as EMP (Epithelial-Mesenchymal Plasticity)) 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 apicobasal polarity, instead 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 [3–6].

Indeed, cellular heterogeneity is emerging as a hallmark of cancer, with cells with distinct EMT phenotypes often localized in distinct tumor regions [7,8]. Similarly, circulating tumor cells (CTCs) that launch into circulation to give rise to metastasis can exhibit a spectrum of EMT features [9]. Hybrid E/M cancer cells that partially conserve cell adhesion can travel the bloodstream collectively as highly metastatic CTC clusters [10]. 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 [11]. However, this heterogeneity seen in both primary tumors and CTCs sparks questions about the relationship between EMT progression in primary tumors and circulating tumor cells (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 poignant 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 then 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.

Materials and methods

1. Data analysis

We calculated EMT scores for multiple primary tumor and CTC datasets using three different EMT metrics – 76GS, KS, and MLR [12–14]. We also calculated correlations in the EMT scores of samples for a given dataset using Spearman's and Pearson's correlation coefficient values.

2. EMT model

This biophysical model focuses on cells at the periphery of a tumor that have the potential to undergo EMT and migrate as circulating tumor cells (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, cell in the more internal layers are not modeled explicitly since they lack the physical space to migrate. Starting from an epithelial state, cells in the invading edge undergo transitions with a rate (k) through a number of intermediate hybrid E/M states, and eventually to a mesenchymal state. The dynamics of cell fractions in different EMT states are described by a set of ordinary differential equations (**SI section 1.1**). Mesenchymal cells can migrate from the cell layer as individual cells; conversely, clusters of neighboring hybrid E/M cells can migrate together as a multicellular unit (**SI section 1.2-1.5**). The migration rate of clusters depends on i) the number of cells in the cluster (i.e. the cluster size), ii) the EMT state of neighboring cells as cell-cell adhesion bonds must be broken, and iii) a migration cooperativity parameter (c) that quantify the propensity to collective migration. While a low c favorites 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, cell(s) 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.

Results

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 [15–19]. Further, there has been a surge of high-throughput measurement such as RNA-seq of CTCs [20–22] as well as primary tumor [23–25], including investigations at a single-cell level. Given the heterogeneity of assessing EMT in multiple studies using diverse markers [26], such transciptomics-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 [27]. These three methods use

different gene lists and algorithms; 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 4 breast cancer CTC cell lines and its metastatic variants [28] (GSE112855) using the above-mentioned three metrics, displayed heterogeneity in their EMT-ness (**Fig 1A, S1A-E**). 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 [29] (PRJNA471754); however, the samples were overall shifted towards a more mesenchymal end of the spectrum as compared to breast cancer cell lines (GSE112855) (**Fig 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 [30] (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 EMTness (**Fig 1C-D, S2A-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, the EMT status of CTCs was also found to be different depending on cultured in petridish (2D) vs. in a 4D ex vivo model [31] (GSE50991) (**Fig 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 [32], and found CTCs to be relatively more mesenchymal than primary tumors (GSE111842) (**Fig 1E-F**). This observation prompted us to interrogate if the EMT status of primary tumor and CTCs can be informative of one another, i.e. can one predict the EMT status of primary tumor based on that of CTCs or *vice-versa*?

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 [33].

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 (**Fig. 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 (**Fig. 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 two main model's 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.

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 (**Fig. 2B**). Moreover, fixing k and increasing c leads 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 2 or more cells (**Fig. 2C**). More generally, varying both parameters triggers a transition from single cell migration to clustered cell migration, as seen for instance by inspecting the fraction of escape event of clusters of 2 or more cells, or cluster escape fraction (**Fig. 2D**), as well as changes in the steady state cell fractions (**Fig S3**).

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. 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 [12,13], the MLR score specifically focused on identifying a hybrid E/M signature [14]. Mimicking the MLR score, cells on the EMT spectrum are assigned weights ranging from 0 (epithelial) to 2 (mesenchymal). Thus, for a 5-state model, the weights for E, E-like, E/M, M-like, 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 (φ_{H1}, φ_{H2}, φ_{H3}, φ_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 ($k \ll 1$, **Fig 3A**, left), and continuously increases to hybrid E/M and mesenchymal for increasing k (**Fig 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 (**Fig 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 (**Fig 3B**). In the larger part of the (k, c) parameter space, CTCs have a larger EMT score than the tumor. Strikingly, a condition of fast EMT and high migration cooperativity (i.e. large k, c) gives rise to a switch where $S_{CTC} < S_T$ (**Fig 3C**). Overall, the relationship between S_T and S_{CTC} is not a fixed one but depends on (k, c),

i.e. more heterogeneous, thus making it difficult to predict one from the another (**Fig 3D**). The score inversion can be observed in models with variable number of intermediate states, indicating that it represents a robust feature of the model (**Fig S4**). 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.

4. Analysis of CTC cluster size distribution reveals EMT score heterogeneity

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 [34–40]. 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. 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 (**Fig 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.

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 many of the other variables can be predicted using our model (**Fig 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 score of a pancreatic ductal adenocarcinoma dataset [41], which is predicted to be hybrid E/M (1.13). 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 (**Fig 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 (**Fig 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 (**Fig 5B**, right).

Similarly, the CTC EMT score is not sufficient to fully constrain the model's parameters. For example, we find that the MLR CTC score of a breast cancer dataset [29] falls within the hybrid E/M range (1.087). The contour line at constant CTC EMT score, however, crosses parameter regions with variable tumor EMT scores (**Fig 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) that the CTCs; moreover, it can have variable cell fractions too (**Fig 5C**, right). Moreover, three other CTC datasets from prostate, myeloma and breast cancer [42–44] are mapped onto distinct model contour lines (Fig S5A).

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 [32]. 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 (**Fig 5D**, left); their intersection provides the (k, c) parameter combination that better reproduces the dataset. For this (k, c) combination, we can then predict the fraction of CTC clusters and their size distribution (**Fig 5D**, right) as well.

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 cases of tumor and CTC scores, however, a line of constant CTC fraction can be identified on the (k, c) parameter space (**Fig 5E**, right). For instance, CTC clusters isolated from a cohort of pancreatic cancer patients amount on average to 15% of the total CTCs [45]. 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 (**Fig 5E**, right). Similarly, 4 very different CTC cluster fraction measurements are mapped onto different contour lines (**Fig S5B**) [45–49].

Finally, the model's best fit 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 [36] is predicted to have a hybrid E/M tumor EMT score but a mesenchymal CTC EMT score (**Fig 5F**).

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.

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 [50]. Subsets of cells with varying EMT-ness (epithelial, hybrid E/M and mesenchymal states) have been observed in primary tumors, CTCs, and metastases [7,9,26].

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 [12–14]. 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 [33]. 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*.

EMT is connected to multiple axes of cancer progression, including invasion, stemness and immune response [51]. 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.

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

This work was supported by Ramanujan Fellowship awarded to MKJ by Science and Engineering Research Board (SERB), DST, Government of India (SB/S2/RJN-049/2018). FB was supported by the Center for Theoretical Biological Physics sponsored by the NSF

(Grant PHY-2019745), by the NSF grant DMS1763272 and a grant from the Simons Foundation (594598, QN).

Acknowledgement

We thank Josè Onuchic for sharing useful suggestions about the manuscript.

Author contribution

FB and MKJ designed the research; SM and TT performed EMT score calculations; FB developed *in silico* modeling; FB and MKJ wrote the manuscript; MKJ supervised the research.

References

1. Brabletz T, Kalluri R, Nieto MA, Weinberg RA. 2018 EMT in cancer. *Nat. Rev. Cancer* **18**, 128–134.
2. Nieto MA, Huang RY, Jackson RA, Thiery JP. 2016 EMT: 2016. *Cell* **166**, 21–45.
3. Pastushenko I *et al.* 2018 Identification of the tumour transition states occurring during EMT. *Nature* **556**, 463–468.
4. José L. McFaline-Figueroa, Andrew J. Hill, Xiaojie Qiu, Dana Jackson, Jay Shendure CT. 2019 A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nat. Genet.* **51**, pages1389–1398.
5. Tian XJ, Zhang H, Xing J. 2013 Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition. *Biophys. J.* **105**, 1079–1089. (doi:10.1016/j.bpj.2013.07.011)
6. Lu M, Jolly MK, Levine H, Onuchic JN, Ben-Jacob E. 2013 MicroRNA-based regulation of epithelial-hybrid-mesenchymal fate determination. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18174–9.
7. Liu S *et al.* 2014 Breast cancer stem cells transition between epithelial and mesenchymal states reflective of their normal counterparts. *Stem Cell Reports* **2**, 78–91.
8. Bocci F, Gaerhart-Serna L, Ribeiro M, Boareto M, Ben-Jacob E, Devi G, Levine H, Onuchic JN, Jolly MK. 2019 Towards understaning Cancer Stem Cell heterogeneity in the tumor microenvironment. *Proc. Natl. Acad. Sci.* **116**, 148–157.
9. Yu M *et al.* 2013 Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science (80-.).* **339**, 580–584.
10. Jolly MK, Mani SA, Levine H. 2018 Hybrid epithelial/mesenchymal phenotype(s): the "fittest" for metastasis? *BBA - Rev. Cancer* **1870**, 151–157.
11. Yutong Sha, Shuxiong Wang, Peijie Zhou QN. 2020 Inference and multiscale model of epithelial-to-mesenchymal transition via single-cell transcriptomic data. *Nucleic Acids Res.* **48**, 9505–9520.
12. Byers LA *et al.* 2013 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.* **19**, 279–90.
13. Tuan Zea Tan, Qing Hao Miow, Yoshio Miki, Tetsuo Noda, Seiichi Mori, Ruby Yun-Ju Huang JPT. 2014 Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol.*

Med. **6**, 1279–1293.
14. George JT, Jolly MK, Xu S, Somarelli JA, Levine H. 2017 Survival Outcomes in Cancer Patients Predicted by a Partial EMT Gene Expression Scoring Metric. *Cancer Res.* **77**, 6415–6428.
15. Ren Zhao, Zhen Cai2, Sheng Li, Yong Cheng, Hua Gao, Fang Liu, Shiyang Wu, Suyan Liu, Yan Dong, Lei Zheng, Wenbin Zhang, Xiaojun Wu XY. 2017 Expression and clinical relevance of epithelial and mesenchymal markers in circulating tumor cells from colorectal cancer. *Oncotarget* **8**, 9293–9302.
16. Anthony Genna, Aline M. Vanwynsberghe, Amélie V. Villard, Charles Pottier, Julien Ancel, Myriam Polette CG. 2020 EMT-Associated Heterogeneity in Circulating Tumor Cells: Sticky Friends on the Road to Metastasis. *Cancers (Basel).* **12**, 1632.
17. Lecharpentier A, Vielh P, Perez-Moreno P, Planchard D, Soria JC, Farace F. 2011 Detection of circulating tumour cells with a hybrid (epithelial/ mesenchymal) phenotype in patients with metastatic non-small cell lung cancer. *Br. J. Cancer* **105**, 1338–1341.
18. Saxena K, Subbalakshmi AR, Jolly MK. 2019 Phenotypic heterogeneity in circulating tumor cells and its prognostic value in metastasis and overall survival. *EBioMedicine* **46**, 4–5.
19. Tamasa De , Shina Goyal , Gowri Balachander , Kaushik Chatterjee, Prashant Kumar , Govind Babu K. AR. 2019 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.* **8**, 1473.
20. Min Yu, David T. Ting, Shannon L. Stott, Ben S. Wittner, Fatih Ozsolak, Suchismita Paul, Jordan C. Ciciliano, Malgorzata E. Smas, Daniel Winokur, Anna J. Gilman, Matthew J. Ulman, Kristina Xega, Gianmarco Contino, Brinda Alagesan, Brian W. Brannigan, Patr DAH. 2012 RNA sequencing of pancreatic circulating tumour cells implicates WNT signaling in metastasis. *Nature* **487**, 510–513.
21. Elisabetta Rossi RZ. 2019 Single-Cell Analysis of Circulating Tumor Cells: How Far Have We Come in the -Omics Era? *Front. Genet.* **10**, 958.
22. Lei Zhao, Xiaohong Wu, Tong Li, Jian Luo DD. 2020 ctcRbase: the gene expression database of circulating tumor cells and microemboli. *Database* **2020**, baaa020.
23. Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, Daniel G Deschler, Mark A Varvares, Ravi Mylvaganam, Orit Rozenblatt-Rosen, James W Rocco, BEB. 2017 Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611–1624.
24. Yu-Chih Chen, Saswat Sahoo, Riley Brien, Seungwon Jung, Brock Humphries, Woncheol Lee, Yu-Heng Cheng, Zhixiong Zhang, Kathryn E. Luker, Max S. Wicha, Gary D. Luker EY. 2019 Single-cell RNA-sequencing of migratory breast cancer cells: discovering genes associated with cancer metastasis. *Analyst* **144**, 7296–7309.
25. Ana Rita Lourenco, Yi Ban, Michael J. Crowley, Sharrell B. Lee, Divya Ramchandani, Wei Du, Olivier Elemento, Jason T. George, Mohit Kumar Jolly, Herbert Levine, Jianting Sheng, Stephen T. Wong NKA and DG. 2020 Differential Contributions of Pre- and Post-EMT Tumor Cells in Breast Cancer Metastasis. *Cancer Res.* **80**, 163–169.
26. Mohit Kumar Jolly, Jason A Somarelli, Maya Sheth, Adrian Biddle, Satyendra C Tripathi, Andrew J Armstrong, Samir M Hanash, Sharmila A Bapat, Annapoorni Rangarajan HL. 2019 Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacol. Ther.* **194**.
27. Priyanka Chakraborty, Jason T. George, Shubham Tripathi, Herbert Levine MKJ. 2020 Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum. *Front. Bioenginering Biotechnol.* **8**, 220.
28. Remi Klotz, Amal Thomas, Teng Teng, Sung Min Han, Oihana Iriondo, Lin Li, Sara Restrepo-Vassalli, Alan Wang, Negeen Izadian, Matthew MacKay, Byoung-San Moon, Kevin J. Liu, Sathish Kumar Ganesan, Grace Lee, Diane S. Kang, Charlotte S.

Walmsley, Christopher ADS and MY. 2020 Circulating Tumor Cells Exhibit Metastatic Tropism and Reveal Brain Metastasis Drivers. *Cancer Discov.* **10**, 86–103.
29. Yu-Heng Cheng, Yu-Chih Chen, Eric Lin, Riley Brien, Seungwon Jung, Yu-Ting Chen, Woncheol Lee, Zhijian Hao, Saswat Sahoo, Hyun Min Kang, Jason Cong, Monika Burness, Sunitha Nagrath MSW& EY. 2019 Hydro-Seq enables contamination-free high-throughput single-cell RNA-sequencing for circulating tumor cells. *Nat. Commun.* **19**, 2163.
30. Sofia Gkountela, Francesc Castro-Giner, Barbara Maria Szczerba, Marcus Vetter, Julia Landin, Ramona Scherrer, Ilona Krol, Manuel C Scheidmann, Christian Beisel, Christian U Stirnimann, Christian Kurzeder, Viola Heinzelmann-Schwarz, Christoph Rochlitz, Wal NA. 2019 Circulating Tumor Cell Clustering Shapes DNA Methylation to Enable Metastasis Seeding. *Cell* **176**, 98–112.
31. Dhruva K Mishra, Chad J Creighton, Yiqun Zhang, Fengju Chen, Michael J Thrall MPK. 2015 Ex vivo four-dimensional lung cancer model mimics metastasis. *Ann Thorac Surg.* **99**, 1149–1156.
32. Lang JE, Ring A, Porras T, Kaur P, Forte VA, Mineyev N, Tripathy D, Press MF, Campo D. 2018 RNA-Seq of Circulating Tumor Cells in Stage II-III Breast Cancer. *Breast Oncol.* **25**, 2261–2270.
33. Bocci F, Jolly MK, Onuchic JN. 2019 A biophysical model uncovers the size distribution of circulating Tumor Cell Clusters across cancer types. *Cancer Res.* **79**, 5527–5535.
34. Sarioglu AF *et al.* 2015 A microfluidic device for label-free, physical capture of circulating tumor cell clusters. *Nat. Methods* **12**, 685–691.
35. Anne Meunier, Javier Alejandro Hernández-Castro, Sara Kheireddine, Sara Al Habyan, Benjamin Péant, Teodor Veres, Luke McCaffrey, Diane Provencher, Anne-Marie Mes-Masson DJ. 2019 Gravity-based microfiltration reveals unexpected prevalence of circulating tumor cell clusters in ovarian cancer. *bioRxiv* , 773507.
36. Kozminsky M, Fouladdel S, Chung J-S, Wang Y, Smith DC, Alva A, Azizi E, Morgan T, Nagrath S. 2018 Detection of CTC Clusters and a Dedifferentiated RNA-Expression Survival Signature in Prostate Cancer. *Adv. Sci.* , 1801254.
37. Bithi SS, Vanapalli SA. 2017 Microfluidic cell isolation technology for drug testing of single tumor cells and their clusters. *Sci. Rep.* **7**, 41707.
38. Krol I *et al.* 2018 Detection of circulating tumour cell clusters in human glioblastoma. *Br. J. Cancer* **119**, 487–491.
39. Cheung KJ *et al.* 2016 Polyclonal breast cancer metastases arise from collective dissemination of keratin 14-expressing tumor cell clusters. *Proc. Natl. Acad. Sci.* **113**, E854–E863.
40. Patil R, Tan X, Bartosik P, Detappe A, Runnels J, Ghobrial I, Lin CP, Niedre M. 2019 In Vivo Monitoring of Rare Circulating Tumor Cell and Cluster Dissemination in a Multiple Myeloma Xenograft Model. *J. Biomed. Opt.* **24**, 085004.
41. Spas Dimitrov-Markov, Javier Perales-Patón, Bruno Bockorny, Ana Dopazo, Manuel Muñoz, Natalia Baños, Victoria Bonilla, Camino Menendez, Yolanda Duran, Ling Huang, Sofia Perea, Senthil K. Muthuswamy, Fatima Al-Shahrour, Pedro P. Lopez-Casas MH. 2020 Discovery of New Targets to Control Metastasis in Pancreatic Cancer by Single-cell Transcriptomics Analysis of Circulating Tumor Cells. *Mol. Cancer Ther.* **19**, 1751–60.
42. Yen-Liang Liu, Aaron M Horning, Brandon Lieberman, Mirae Kim, Che-Kuang Lin, Chia-Nung Hung, Chih-Wei Chou, Chiou-Miin Wang, Chun-Lin Lin, Nameer B Kirma, Michael A Liss, Rohan Vasisht, Evan P Perillo, Katherine Blocher, Hannah Horng, Josephine A Taverna, C-LC. 2019 Spatial EGFR Dynamics and Metastatic Phenotypes Modulated by Upregulated EphB2 and Src Pathways in Advanced Prostate Cancer. *Cancers (Basel).* **11**, 1910.
43. Juan-Jose Garcés, Michal Simicek, Marco Vicari, Lucie Brozova, Leire Burgos, Renata Bezdekova, Diego Alignani, Maria-Jose Calasanz, Katerina Growkova, Ibai Goicoechea, Xabier Agirre, Ludek Pour, Felipe Prosper, Rafael Rios, Joaquin

Martinez-Lopez, Pamela BP. 2020 Transcriptional profiling of circulating tumor cells in multiple myeloma: a new model to understand disease dissemination. *Leukemia* **34**, 589–603.
44. Arvind Iyer, Krishan Gupta, Shreya Sharma, Kishore Hari, Yi Fang Lee, Neevan Ramalingam, Yoon Sim Yap, Jay West, Ali Asgar Bhagat, Balaram Vishnu Subramani, Burhanuddin Sabuwala, Tuan Zea Tan, Jean Paul Thiery, Mohit Kumar Jolly, Naveen Ramalingam DS. 2020 Integrative Analysis and Machine Learning based Characterization of Single Circulating Tumor Cells. *J. Clin. Med.* **9**, 1206.
45. Maddipati R, Stanger BZ. 2015 Pancreatic cancer metastases harbor evidence of polyclonality. *Cancer Discov.* **5**, 1086–1097.
46. Aceto N *et al.* 2014 Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* **158**, 1110–1122.
47. Krebs MG *et al.* 2012 Analysis of circulating tumor cells in patients with non-small cell lung cancer using epithelial marker-dependent and independent approaches. *J. Thorac. Oncol.* **7**, 306–315.
48. Molnar B, Ladanyi A, Tanko L, Sreter L, Tulassay Z. 2001 Circulating Tumor Cell Clusters in the Peripheral Blood of Colorectal Cancer Patients. *Clin. Cancer Res.* **7**, 4080–4085.
49. Kulasinghe A, Schmidt H, Perry C, Whitfield B, Kenny L, Nelson C, Warkiani ME, Punyadeera C. 2018 A Collective Route to Head and Neck Cancer Metastasis. *Sci. Rep.* **8**, 746.
50. Mohit Kumar Jolly TC-T. 2019 Dynamics of Phenotypic Heterogeneity Associated with EMT and Stemness during Cancer Progression. *J. Clin. Med.* **25**, 1542.
51. Jia D, Li X, Bocci F, Tripathi S, Deng Y, Jolly MK, Onuchic JN, Levine H. 2019 Quantifying cancer epithelial-mesenchymal plasticity and its association with stemness and immune response. *J. Clin. Med.* **8**, 725.

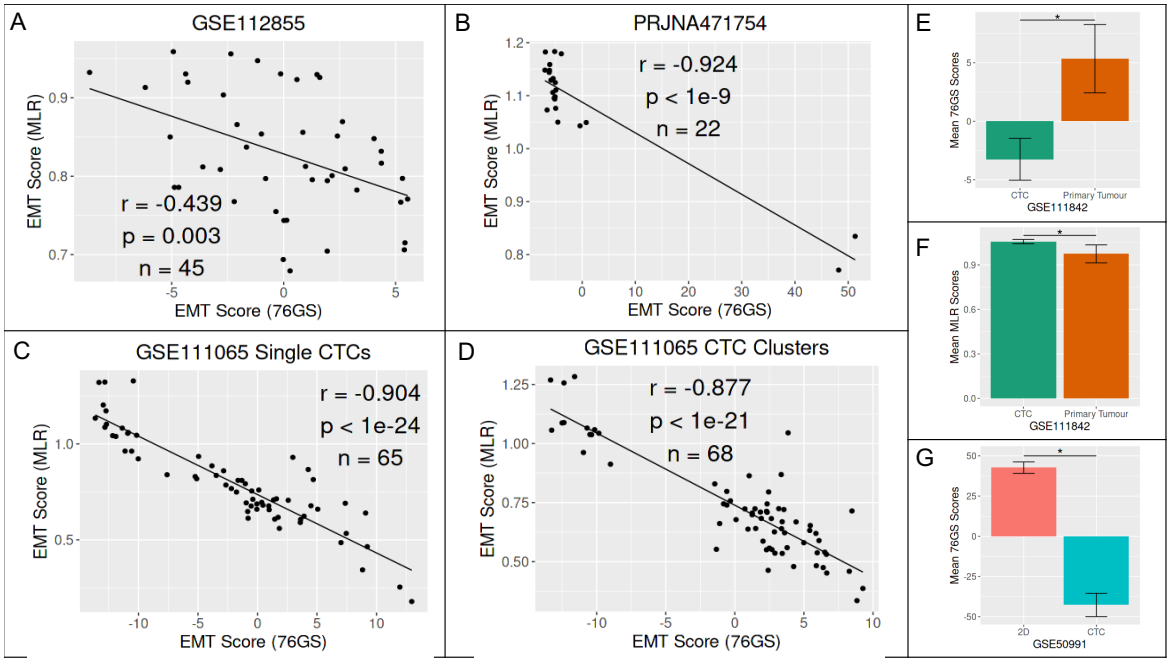


Figure 1. **Heterogeneity in EMT scores of CTCs and primary tumors**. **(A)** Correlation of 76GS and MLR scores for GSE112855 (CTC-derived cell lines/sub lines in breast cancer). Each dot shows a sample in the dataset. Pearson's correlation values (r, p) are reported. **(B)** Same as A) but for CTCs in PRJNA471754. **(C,D)** Same as A) but for for single CTCs and CTC clusters in GSE111065. **(E,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 (*: $p < 0.0001$).

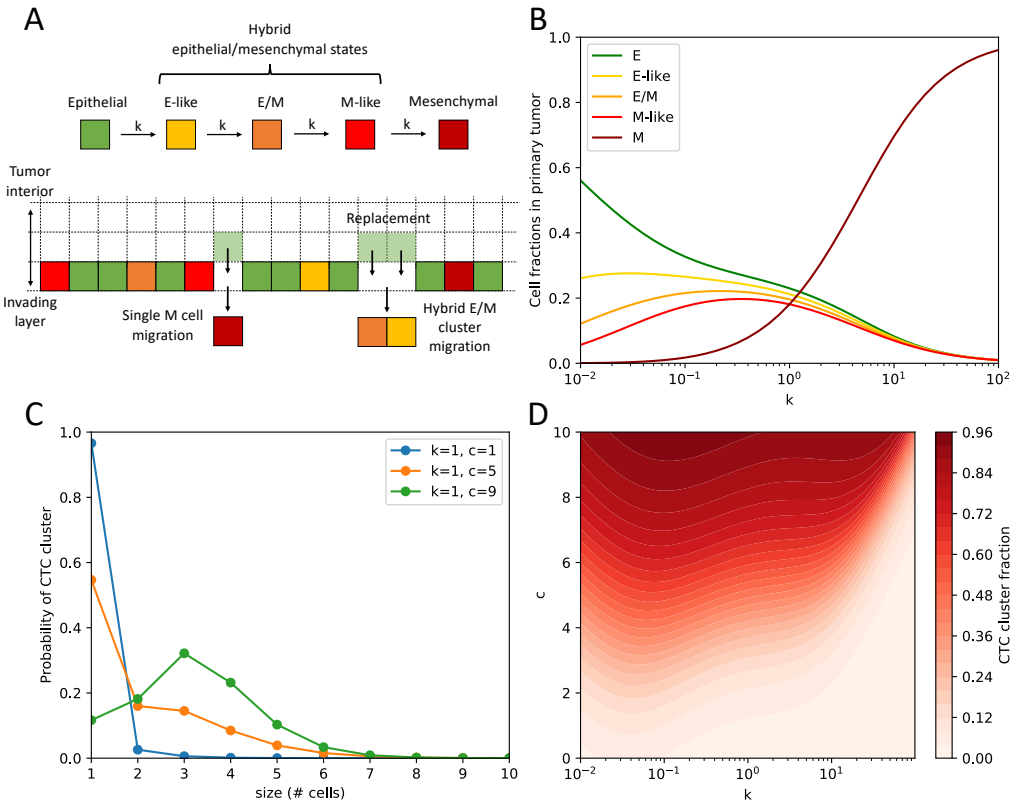


Figure 2. **Single cell-cluster migration transition in the EMT-migration model. (A)** Overview of the EMT-migration model. Top: cells undergo transitions through 3 hybrid E/M states with rate k. Bottom: mesenchymal cells migrate as single cells, while hybrid E/M cells can migrate together as multicellular 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).

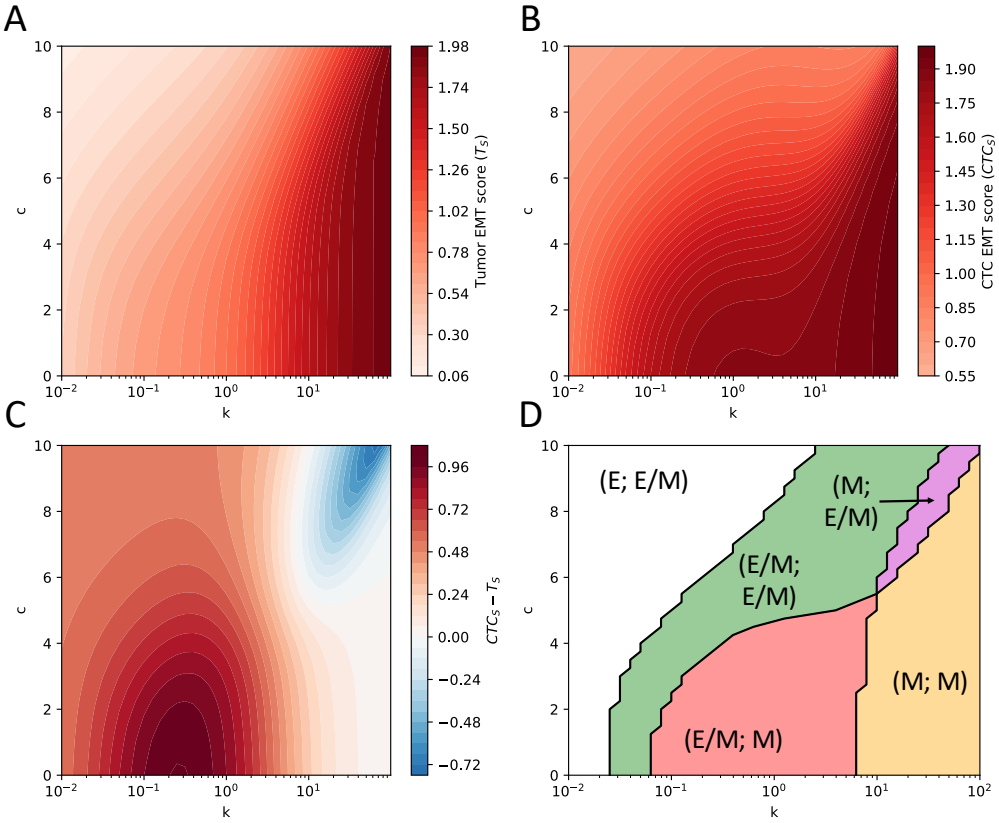
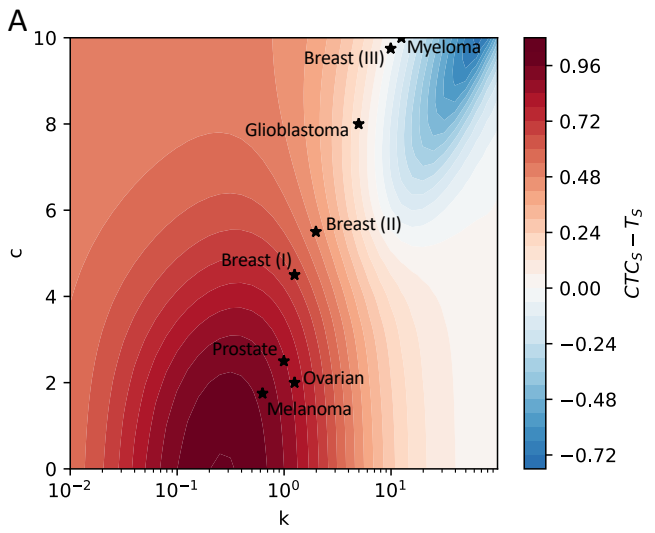


Figure 3. **Inversions in tumor and CTC EMT scoring metrics.** Predicted EMT score of primary tumor T_S **(A)**, migrating cells, defined as both single CTCs and CTC clusters CTC_S **(B)**, and difference between them $CTC_S - T_S$ **(C)** as a function of EMT rate (k) and migration cooperativity (c). **(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.



B

Dataset	Predicted Tissue score	Predicted CTC score
Melanoma (Sarioglu et al 2015)	0.87	1.81
Ovarian (Meunier et al 2019)	0.99	1.81
Prostate (Kozminsky et al 2018)	0.93	1.78
Breast (Sarioglu et al 2015)	0.89	1.54
Breast (Bithi et al 2017)	0.92	1.39
Glioblastoma (Krol et al 2018)	0.90	1.03
Breast (Cheung et al 2016)	0.89	0.87
Myeloma (Patil et al 2019)	0.93	0.85

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 [34–40]. Each starred dot represents a CTC cluster size distribution that was fitted with the 5-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.

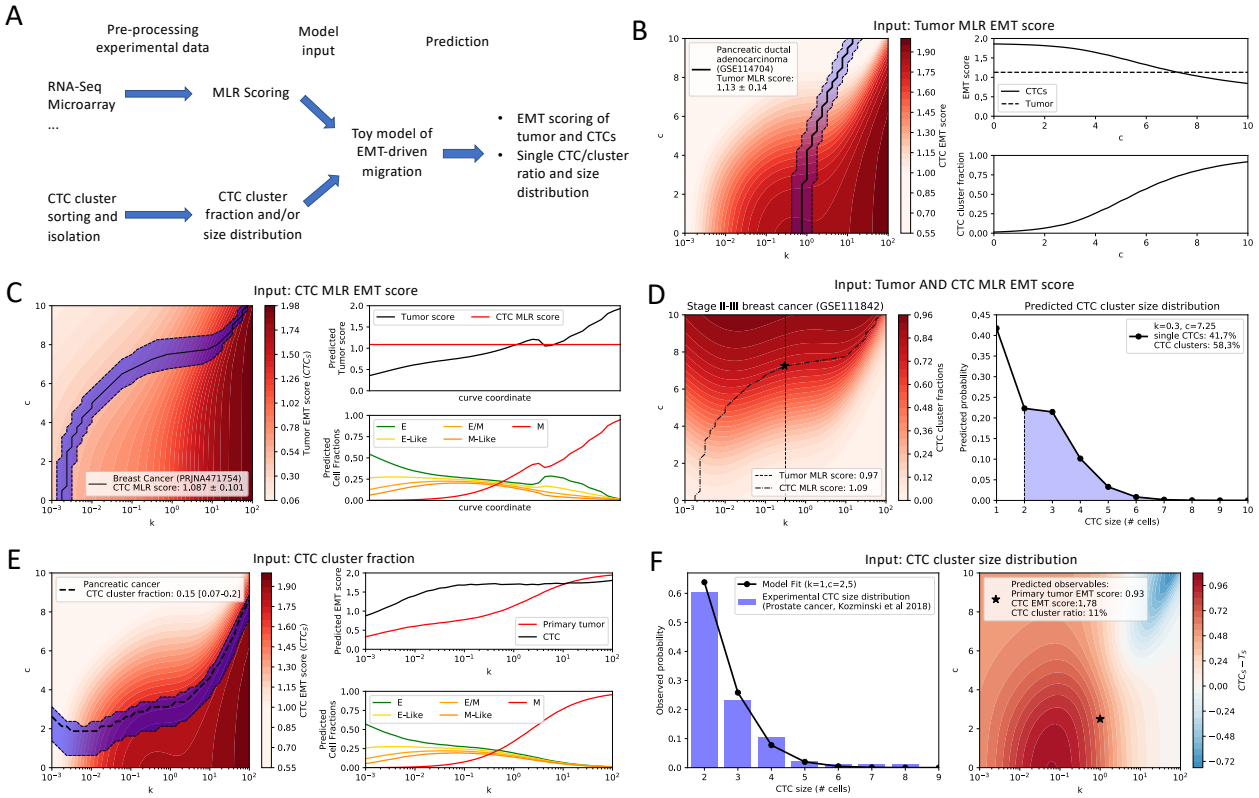


Figure 5. **The model's predictive power for several popular measurements of cancer aggressiveness. (A)** For single cell measurements of gene expression (RNA-seq, microarray, etc...), an EMT scoring technique such as MLR is used to provide an input to the model; conversely, for the case of CTC cluster 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).