Interactive dynamics of reaction-diffusion and adhesion predict diverse invasion strategies of cancer cells in matrix-like microenvironments

Durjay Pramanik[1], Mohit Kumar Jolly[2] and Ramray Bhat[1]

1: Department of Molecular Reproduction, Development and Genetics, Indian Institute of Science, Bangalore, 560012 India

2: Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore 560012, India

Address for correspondence: ramray@iisc.ac.in, mkjolly@iisc.ac.in

Abstract:

The metastasis of malignant epithelial tumors begins with the egress of transformed cells from the confines of their basement membrane to their surrounding collagenous stroma. The invasion can be morphologically diverse, ranging from that of dispersed mesenchymal cells to multicellular collectives. When cancer cells are cultured within basement membrane-like matrix (BM), or Type 1 collagen, or a combination of both, they show collective-, dispersed mesenchymal-, and hybrid collective-dispersed (multiscale) invasion respectively. Here, we ask how distinct these invasive modes are with respect to the cellular and microenvironmental cues that drive them, and how can cancer cells transition between such states. A rigorous exploration of invasion was performed within an experimentally motivated Cellular Potts-based computational modeling environment. The model comprises of adhesive interactions between cancer cells, BM- and collagen-like extracellular matrix (ECM), and reaction-diffusion-based remodeling of ECM. The model output consisted of metrics cognate to dispersed- and collective- invasion. An exhaustive combination of input values gave rise to a spatial output distribution that comprised dispersed-, collective- and multiscale- invasion. K-means clustering of the output distribution followed by silhouette analysis revealed three optimal clusters: one signifying indolent invasion and two representing multiscale invasions, each of which encompass the purer dispersed- and collective invasions respectively. Principal component analyses of phenotypic outputs close to the separation of these two multiscale modes established specific input signatures: perturbing such signatures revealed the cues

that can allow phenotypic transitions among these three clusters. Our systems-level analysis provides quantitative insights into the biases and constraints cancer cells face during their invasion through tissue microenvironments with distinct physicochemical properties.

Introduction

The details of the beginnings of cancer progression determine not just the kinetics of its metastasis but also its response to therapeutic efforts (Haeger et al., 2020; Nieto et al., 2016). Epithelia from malignant tumors proliferate and breach their covering laminin-rich basement membrane (BM) barriers to migrate to the surrounding connective tissue consisting of fibroblasts, other resident cells, and extracellular matrix (Nelson and Bissell, 2005; Pally et al., 2019; Pickup et al., 2014). The latter is rich in Type 1 collagen and other fibrillar proteins as well as proteoglycans (Di Lullo et al., 2002; Hynes, 2009).

Bidirectional interactions including adhesion and degradation between cancer cells and their surrounding tissue microenvironment at biochemical and/or biophysical levels shape the nature of their migration (Anderson, 2005) (Nissen et al., 2019). Cancer cell migration can be broadly classified into unicellular or multicellular categories. Solitary cancer epithelia disperse and migrate through ECM in a fibroblast-like manner (Madsen and Sahai, 2010). Such a migration pattern often concurs with a series of changes in gene expression, intercellular adhesion and cell shape known as the epithelial to mesenchymal transition or EMT (van Zijl et al., 2011). The spindle-shaped ‘mesenchymal’ cells interact extensively with ECM in their process of migration (unlike the other unicellular migratory mode: amoeboid, where the adhesion with matrix is minimal) (Friedl et al., 2004; Gadea et al., 2008; Huang et al., 2014; Pankova et al., 2010; Sanz-Moreno et al., 2008). Collective cell invasion involves migration of cellular ensembles that retain adhesive and communicative contacts with each other through their movement (Cheung et al., 2013; Friedl et al., 2012). These diverse phenotypes are also reflected in blood-

borne dissemination of circulating tumor cells as individual cells or clusters (Bocci et al., 2019).

The migratory dynamics of cancer epithelia can also transition among the above-mentioned types: mesenchymal, amoeboid, and collective; such plasticity may further drive phenotypic heterogeneity and symbiotic behavior in cancer invasion patterns (Hecht et al., 2015; Huang et al., 2015; Lintz et al., 2017). As cancer cells migrate to newer microenvironments, transitions from mesenchymal to epithelial morphologies could appropriately render invasive single cells more adhesive to each other (Krakhmal et al., 2015). Collectively migrating strands of cells are typically thought of as hybrid epithelial/mesenchymal (E/M) phenotype (Nagai et al., 2020); the *in vitro*, *in vivo* and *in silico* evidence for existence, plasticity and aggressiveness of such hybrid E/M cells has been mounting (Jolly et al., 2019). Thus, it is not surprising to identify unique migratory behaviors of neoplastic cells that are phenotypically intermediate between their dispersed and collective counterparts; for instance, multicellular streaming of an amoeboid or mesenchymal nature (Friedl et al., 2012; Kedrin et al., 2008; Liu et al., 2019; Patsialou et al., 2013) driven by weak intercellular junction strength. The identification of leader versus follower cells in such collective migration and the conceptual and mechanistic overlaps of cancer progression with jamming-unjamming transition has also been an active area of investigation (Sadati et al., 2013). However, it still remains elusive how fundamentally distinct the dispersed, collective and multicellular streaming migratory modes are, or whether transitions between them requires modulating few or many intracellular/extracellular inputs. It is also ambiguous whether the streaming and other semi-collective migrations represent an intermediate phenotypic state between dispersed and collective transitions or are unique phenomena in themselves.

Using an experimental setup that mimicked the topographical arrangement of invasive cell clusters encapsulated within BM-like ECM and subsequently by Type 1 Collagen, we were able to observe the simultaneous occurrence of several spatially discernible migratory modes, a phenomenon we called multiscale invasion (Pally et al., 2019). These experimental observations were simulated using Compucell3D, a modeling framework based on the Cellular Potts model (Das et al., 2017; Glazier and Graner, 1993; Graner and Glazier, 1992; Swat et al., 2012; Zhang et al., 2011),

using which we established that regulating a specific subset of biophysical phenomena such as cell-cell and cell-ECM adhesion, and cell division and reaction-diffusion-based modulation of matrix proteolysis (computational equivalents of the ECM-degrading matrix metalloproteinases (MMPs) and their inhibitors (The inhibitors of matrix metalloproteinases or TIMPs))- can give rise to a set of invasive phenotypes that comprise dispersed, collective, and multiscale invasion categories. Here, we investigate how robust and distinct each of these phenotypes are, and what forces mediate shifts between them. Our multiparametric simulations followed by statistical analyses suggest to our surprise, that invasion is fundamentally of two modes: both are multiscale in nature and subsume exclusively dispersed mesenchymal and collective invasions respectively (and will be referred to henceforward as dispersed multiscale invasion: dispersed MSI, and collective MSI respectively). We show the inputs that guide transitions between invasive modes, which may provide insight into how migratory metastatic cancer cells can transition between distinct modes while negotiating everchanging tissue microenvironments.

Results

Construction of an invasion phenospace based on the dynamics of reaction-diffusion, and cell-cell and -matrix adhesion

In our previous paper, we were able to demonstrate, using a computational model, that an interplay between parametric combinations of cell-cell and -ECM (BM and Type 1 collagen) adhesion, and reaction-diffusion kinetics of MMP- and TIMP-based ECM remodeling were able to give rise to diverse phenotypes of cell invasion. Not just dispersed- and collective- invasion, but also phenotypes with concurrent presence of the latter two modes were also observed, which we termed multiscale invasion. Such simulations were calibrated with *in vitro* 3D culture experiments of breast cancer cells in combinations of appropriate ECMs (Figure 1A).

To better quantify the invasion phenotype and characterize the tendency of cells to switch among these phenotypes, here, we have constructed an invasion phenospace, wherein the horizontal and vertical axes are represented by metrics inherent to collective and dispersed cellular invasion respectively (Figure 1B and C).

Therefore, the X axis measures the size of the largest single multicellular mass (the starting ‘primary focus’) at the end of the simulation. The Y axis measures the number of objects (cells or tiny cellular clusters) that are spatially disconnected or dispersed from the primary focal mass at the end of the simulation. Outputs that are closer and parallel to X axis therefore represent phenotypes with exclusively collective migration of cells; those closer and parallel to Y axis represent dispersed cells migrating solitarily as a result of their disintegration from the primary focal multicellular mass. Outputs with a relatively higher magnitude of values on X and Y axes represent multiscale invasion (MSI) phenotype.

Based on previous simulation results, we took the following inputs to map the phenospace: the contact energies representing cell-cell, cell-BM and cell-collagen adhesion, cell division, diffusivity of MMP and the cooperativity between MMP and its diffusible inhibitor TIMP (at a given spatial point, the threshold ratio of MMP to TIMP concentration at a given spatial point that is required to degrade Type 1 collagen at that point is the cooperativity of MMP-TIMP and was shown to be an important regulator of multiscale invasion (Diambra et al., 2015; Pally et al., 2019)) (Figure 1A). Whereas this set was, by no means, an exhaustive set of possible inputs, they were shown by previous simulations to significantly determine the phenotypic outcome of invasion. We chose input values for which we observed exclusive dispersed, or collective invasion phenotypes and then progressively decreased them in discrete intervals to a minimum limit at which no invasion was observed. In total, 5 values of each input were chosen; outputs were computed for each combination of 5 values of 6 inputs; the computation was repeated 3 times (replicates) using CompuCell3D. The replicates aim to capture the variability in cellular responses enabled by the underlying stochastic framework of the Cellular Potts modeling framework. These outputs were mapped onto the 2D invasion phenospace (Figure 1D). The distribution of outputs indicated that multiscale invasion represented a continuous phenomenon that bridged the exclusive collection of dispersed- and collective invasions close to the respective cartesian axes.

Identification of optimal clustering of the invasive phenospace

We next investigated the optimal number of phenotypic clusters that can best classify the scatter of outputs in the two-dimensional phenospace. Such approaches are commonly used for gene expression studies to unravel hidden patterns in multi-variable high-dimensional spaces (Oyelade et al., 2016). K-means algorithm is a commonly used unsupervised clustering method which, given *a priori* an integer value of K, partitions the given dataset into K disjoint subsets (Macqueen, 1967). It selects a random initial seed point of preferred clusters which are used to cluster the remaining data points; thus, different runs of K-means can possibly give rise to distinct clusters and their arrangements.

We performed K-means clustering on the phenospace for K=2 to K=8 for multiple instances, and as expected, got different clusters across multiple (n=10) runs for a fixed value of K, except for K=3 (Figure 2A). For K=2, two predominant categorization patterns were seen. In the first pattern, which was obtained 8 out of 10 times, one of the clusters indicated a more indolent phenotype with relatively small number of dispersed objects and smaller area of largest disconnected mass (blue colored clusters in Fig 2A, top row, left panel) and the other cluster which denotes cases across dispersed, collective and multiscale invasion scenarios (yellow clusters in Fig 2A, top row, left panel). In the other categorization pattern for K=2, we observed a cluster denoting largely collective cell migration (blue cluster in Fig 2A, top row, right panel), and another one encompassing both single-cell and multiscale invasion (yellow cluster in Fig 2A, right panel). For K=3, the clustering pattern that emerged categorized the phenospace into an indolent phenotype (blue cluster in Fig 2A, middle row panel) and two multiscale invasive phenotypes (yellow and green cluster in Fig 2A), each of which included purer single cell- and collective invasion-representing outputs close and parallel to the Y and X axes respectively and will, therefore, be referred to as dispersed and collective MSI respectively. This pattern was obtained 10 out of 10 times for K-means clustering for K=3, suggesting the robustness of this categorization for further analysis. For values of K>3, we observed two or more cluster patterns (Supplementary figure 1). Clusters within such patterns could be subsumed within the two distinct multiscale invasion clusters seen for K=3 (see for e.g., cluster pattern #2 of K=5), or in other cases spatially independent of the latter (for e.g., cluster pattern #1 of K=5).

Next, we conducted Silhouette analysis on K-means to quantify the consistency in clustering, for varied values of K. Silhouette scores are indications of how similar a data-point is to others in its own clusters, compared to those in other clusters. Thus, Silhouette scores provide a measure of both the tightness of the clusters defined as well as the extent of separation among them. These values range from -1 to +1; the higher the value, the closer a given datapoint is to other points in those clusters and the farther it is from other clusters (Zhao et al., 2018). Thus, a higher average score would indicate a greater degree of cluster separation. Clustering for K=3 reported an average Silhouette score of 0.90 (Fig 2B and Supplementary Figure 1); which was the highest average Silhouette score among other values of K for which clustering was attempted (see graph in Figure 2B bottom right). Moreover, over 15 replicates of K-means tried over the same phenospace, the mean and variance of average Silhouette scores was calculated. K=3 showed the highest mean and the least variance (Fig 2B), further endorsing that three clusters in the pattern shown in Figure 2B is the most optimal clustering scenario for the phenospace.

The two multiscale invasion clusters exhibit distinct phenotypic patterns

Having identified the two multiscale invasion clusters, we asked whether the phenotypes associated with the outputs within each cluster, were visually distinct. In order to do so, we chose outputs within each cluster and examined the endpoint of simulation (600th Monte Carlo Step (MCS) of the simulations). We observed that the dispersed MSI did show an excess of cells that were dispersed from the starting cell mass, although the central mass had also grown in size, but it was less dysmorphic relative to its initial starting shape (Figure 3A). Interestingly, we also observed a variant of this more ‘classical’ dispersed MSI, wherein the occurrence of small oligocellular clusters that were disconnected from the central mass showing bulk invasion, was reminiscent of the phenomenon of tumor budding (Figure 3B; see Discussion).

An examination of the output within the collective MSI cluster showed a more connected distribution of cells that had extended from the central mass along with few dispersed cells (Figure 3C). A non-canonical variant phenotype observed was one in which several multicellular fingerlike projections emanated from the central

invasive mass, reminiscent of the Rayleigh-Taylor instability seen for immiscible fluids of distinct densities (Aref and Tryggvason, 1989; Mikaelian, 1990). Such a phenotype has been described as mesenchymal streaming and our analysis shows that it is observed within the collective MSI cluster, closer to its boundary with the indolent/noninvasive cluster (Figure 3D) suggesting that the interrelationships between the various invasive phenotypes do not constitute a linear spectrum but rather comprise more complex 2D terrain.

Determinants of the different phenotypes observed in invasive phenospace

Further, we investigated which, if any, combinatorial parametric combinations are overrepresented in a specific migration pattern or a particular cluster of outputs in our phenospace. Such analysis can be valuable in decoding how emergent behavior (different cellular invasion strategies) occurs as a function of the complex interplay among the different interconnected modules, such as adhesion and reaction-diffusion in our case. We employed Principal Component Analysis (PCA) – a classical method to reduce the dimensionality of the dataset by transforming the high-dimensional data into a new set of uncorrelated variables called as principal components (PCs) (for further details see appropriate methods section). These PCs are linear weighted combinations of some of the variables from the original dataset. The PCs are uncorrelated among themselves and are organized in decreasing order of the variability of the data they are able to capture within themselves. The kth PC can be considered as a direction that maximizes the variability in the projections of the datapoints such that it is orthogonal to all the remaining (k-1) PCs. Thus, choosing the right number and the right set of PCs is essential to extract any biologically relevant information from the PCA plots (Konishi, 2015; Yeung and Ruzzo, 2001).

A total of fourteen different areas were chosen from the two multiscale invasion phenotypes (D_{Violet}, D_{Yellow}, D_{Red} and D_{Blue} boxes from the dispersed multiscale cluster; C_{Violet}, C_{Yellow}, C_{Red} and C_{Blue} from the collective multiscale cluster are shown in Fig 4; for four more from dispersed and two more from collective cluster, see Supplementary Figure 2) The values of the six input parameters corresponding to the outputs lying within these eight areas were extracted, and a PCA was conducted

for each of these areas. PC1 in most cases explained over 60% of variability in the data (see 'scree plots' in corresponding eight cases). The composition of PC1 in terms of the linear weighted combinations of the six input parameters revealed striking patterns into the synergistic or antagonistic interplay among these input parameters in outputs corresponding to the two different clusters – dispersed MSI and collective MSI. In the PC1 composition for dispersed multiscale cases, the contributions of cell-cell and cell-BM contact energy parameters were positively correlated. The other four input parameters had a relatively weaker contribution. However, in the PC1 composition for collective multiscale invasion, the pattern obtained was different – the contributions of MMP diffusion rate and R-D cooperativity were positively correlated consistently; any contributions of remaining four parameters were not consistently observed (See Table 1). Put together, this analysis revealed the functional signatures/determinants that constitute different MSI phenotypes close to their boundary in the phenospace.

Transitions between the two multiscale invasions

The PCA provides insights into the input signatures that likely typify the individual multiscale invasion phenotypes. Therefore, we asked whether altering specific signature input values, guided by PCA (Table 1), can allow the phenotypes to transition between the three broad phenotypic clusters.

We first asked whether it was possible to transition *from* collective MSI *to* dispersed MSI by altering the following input parameters: 1. Increasing cell-cell contact energy (decreasing intercellular adhesion) 2. Increasing cell-BM contact energy (decreasing cell-BM adhesion) 3. Decreasing cooperativity of ECM remodeling by MMP/TIMP and 4. Decreasing diffusivity MMP. Of 6 randomly chosen outputs that originally showed collective MSI, upon decreasing cell-BM adhesion (input parameter #2), 4 of them transitioned to dispersed MSI cluster, and 2 of them transitioned to indolently invasive cluster (Figure 5A). Changing the other inputs (i.e., #1, #3 and #4) managed at best, cluster transition of only one output.

For six randomly chosen outputs that showed dispersed MSI, transition of all outputs to collective MSI cluster took place when either the cooperativity of MMP-TIMP was

decreased (input parameter #3) or the MMP diffusivity was increased (input parameter # 4) (Figure 5B). In fact, decreasing cell-BM adhesion (input parameter #2) allowed transit of 5 points to the indolently invasive cluster (Figure 5C). Thus, depending on different individual perturbations informed by the PCA, it is possible to witness transitions among these different phenotypes in a context-dependent manner.

Discussion

In an illustrative review, Friedl and coworkers provided an elegant classification of generic cell invasion: they enumerated about ten types of movements with solitary and bulk movements at the two ends of the phenotypic spectrum (Friedl et al., 2012). In addition, a three-dimensional graph shows how the three axes of leading-edge polarity, apicobasal polarity and cell-cell contacts can accommodate the different modes of phenotypes. This idea is based on an earlier review on the same topic where Friedl and Wolf argue for a “multiparameter tuning model of invasion wherein a combination of cues such as density, stiffness, and orientation of the extracellular matrix together with cell determinants—including cell–cell and cell–matrix adhesion, cytoskeletal polarity and stiffness, and pericellular proteolysis—interdependently control migration mode and efficiency” (Friedl and Wolf, 2008; Wolf et al., 2007). Our manuscript seeks to build on this idea by subjecting it to mathematical rigor and identify its underlying assumptions and mechanisms. While Friedl and others have acknowledged that distinct modes may concur spatially, their tuning model would attribute such co-occurrence to the heterogeneity in microenvironmental cues. However, does the co-occurrence of dispersed and collective modes of invasion (multiscale invasion or MSI) represent a distinct invasive mode, wrought through distinct signatures from a tunable combination of inputs? In this manuscript, we argue that such is the case. In fact, the input combinations give rise to not one but two types of MSI. The two types have higher relative representation of dispersed and collective cell invasion respectively and explain the exclusively solitary and bulk invasion respectively.

Surgical histopathological literature is replete with observations of multiscale behavior in cancer invasion. In the context of colorectal carcinoma, Prall and coworkers have made meticulous observations on ‘tumor budding’, wherein single or

small collectives of, cells seem to escape from growing edges of malignant neoplasms (Prall et al., 2005; Prall et al., 2009). Such multiscale behavior is associated with a more sinister kinetics of metastasis and poor prognosis (Hase et al., 1993; Prall et al., 2005; Prall and Ostwald, 2007). Tumor budding has been observed in breast and pancreatic cancers as well (Petrova et al., 2019; Salhia et al., 2015) and have been connected to a partial EMT phenotype (Bronsert et al., 2014; Grigore et al., 2016). Our previously published experiments, which were crucial to the development of the computational model we have explored here, described multiscale invasion, which is phenomenologically similar to tumor budding. Indeed, early investigations by others confirm our empirical and theoretical findings that surface proteins and ECM play key roles contributing to multiscale invasion (Graves et al., 2016; Masaki et al., 2003).

Our findings in this manuscript suggest that multiscale invasion may transition between the two modes in subtle ways. If this were to be true in the histopathological setting, changed kinetics of tumor budding and its consequences may be far more common that what is thought of at present, and needs to be studied in greater detail. Many phenotypes observed in this framework are reminiscent of fingering instabilities observed in epithelial sheet migration (Sepulveda et al., 2013). In fact cancer cells adopting either mode – dispersed or collective – can also contribute to the burden of circulating tumor cells occupying various positions on the epithelial-mesenchymal spectrum (Aceto et al., 2014; Jolly et al., 2017).

Our study indicates that the low cell-cell and cell-BM adhesion typifies mesenchymal MSI close to its boundary with collective MSI. This corroborates with observations from the rich literature on epithelial mesenchymal transitions in the context of cancer (Spaderna et al., 2006). On the other hand, collective MSI close to the boundary is characterized by signatures of high MMP diffusivity and cooperativity, which is corroborated literature on mammary gland development (Alcaraz et al., 2011) and wound healing (Theveneau and Linker, 2017). Our analysis allows us to make specific predictions on how cells may transition (or coaxed to do so) between different states – between the two MSI but also from MSI to an indolent phenotype. For example, decreasing cell-BM adhesion for dispersed MSI seems to inhibit their invasion. However, this is true only for phenotypes close to the boundary of the 2

MSI clusters. We also would like to add a cautionary note, based on our efforts that the transitioning between different states, despite being influenced by some signature inputs, is very contextual on the values of other seemingly less consistent outputs. In the future, we aim to perform a more exhaustive analysis of the transition dynamics between outputs in the center and peripheries of all clusters.

At this point of time, our study does not incorporate two salient aspects of invading cancer cells. The first is the change in shape of cells as they move through matrix microenvironments. This distinguishes amoeboid from mesenchymal single cell migration and is associated with shape-based asymmetries in remodeling of ECM intercellular adhesion (Pankova et al., 2010). Despite this lacuna, we believe our dispersed multiscale invasion is mesenchymal in nature as it is predicated on adhesion of cells to the ECM (which distinguishes it from amoeboid migration), and appropriate for the level of graining that is achieved in our analysis. However, cell shape dynamics will be implemented in our future efforts. Secondly, our computational environment does not incorporate the stromal cells that contribute in significant ways to the phenomenology of cancer cell invasion (Labernadie et al., 2017) due to limitations of the sheer computational power required to compute the dynamics of a multi-cell multi-ECM environment. We assume that the R-D dynamics of MMP-TIMP incorporated in the computational environment is a downstream effect of the cancer and stromal cell activities. This assumption will be tested in future efforts. Our effort serves to put multiscale migratory behavior in the fundamental center of the concept of cancer invasion. Understanding the process has potentially far reaching consequences for future therapeutic efforts.

We would conclude by drawing a parallel between our study and the multiscale modeling efforts employed for investigating developmental processes (Alber et al., 2006; Alber et al., 2004; Christley et al., 2007; Harrison et al., 2011; Izaguirre et al., 2004; Kiskowski et al., 2004; Merks and Glazier, 2006). While our model derives its motivation, initial conditions and tissue geometry (which have been shown to be a crucial determinant in morphogenetic pattern formation (Nelson et al., 2006) from breast cancerous contexts (Bhat and Bissell, 2014), one can as well interpret our model as a developmental system wherein inter-agent adhesion and reaction-diffusion based phenomena sculpt cellular patterns. In fact several of our

model inputs represent the biological physics associated with the framework of dynamical patterning modules, which has been employed to study the evolution and logic of developmental mechanisms (Newman and Bhat, 2008, 2009). In effect, this suggests that the redeployment of developmental and morphogenetic principles in the context of genomic aberrations may underlie the mechanisms behind discrete invasive behaviors of cancer cells and the transitions between them.

Materials and methods:

Modelling framework

Compucell3D (CC3D) is a problem-solving environment that combines the lattice-based GGH (Glazier–Graner–Hogeweg) model or CPM (Cellular Potts model) with PDE solvers and other models to allow for study and simulation of multiscale virtual biological processes (Swat et al., 2012). The software divides the whole simulation lattice into '*cell*s' (collection of pixels). A specific '*cell* type' (*cell* is italicised to distinguish simulated *cells* from biological cells) is assigned to each of them. Interaction parameters between *cell* types can be made to approximate biological constraints between components, similar to that of the original in vitro or in vivo biological system. Such constraints or parameters regulate the simulation through the effective energy or **Hamiltonian (H)** calculated at each **Monte Carlo Step (MCS)**. Calculation of H determines the allowed configuration and behavior of *cells* at each MCS.

$$H = \sum_{\substack{i,\ j \\ neighbours}} J(\tau(\sigma_i), \tau(\sigma_j))(1 - \delta(\sigma_i\ , \sigma_j)) + \sum_{\sigma}[\lambda_{vol}(\sigma)\big(v(\sigma) - V_t(\sigma)\big)^2]$$

A detailed description of the Hamiltonian may be found in Pally et al Frontiers Physiol. Briefly, at each Monte Carlo Step (MCS) the Hamiltonian is calculated by estimating two terms: the first term sums all neighboring pairs of lattice sites "I" and "j" with associated contact energies (J) between the pair of cells indexed at those i and j (σ signifies the *cell* ID, and τ signifies the *cell* type). Contact energy is inverse to the adhesion of two *cells*: greater the contact energy, weaker is the adhesion between those two cells and vice versa. The second term of the Hamiltonian determines the dynamics of cell volume (for *cell* σ, $\lambda_{vol}(\sigma)$ signifies the its inverse

compressibility, v(σ) is its volume, and $V_t(\sigma)$ is its target volume. At every MCS, the Hamiltonian calculation is done so as to move the system to a lower-energy configuration. The temporal progression of the simulation is determined by MCS value because iteratively, with increase in MCS the initial configuration of the simulation reaches its final spatial configuration. Associated simulation files are made up of steppables (python classes) and plugins (preset functions of the software) which contain instructions for all of *cell*'s behaviour executed through Hamiltonian (https://compucell3dreferencemanual.readthedocs.io/en/latest/index.html) (Swat et al., 2012)

Simulation lattice:

We use a 100*100*1-pixel square lattice with non-periodic boundary for all the simulations (the initial configuration of simulations resembles Fig 1(C) where no invasion of any type is observed).

Cell types:

There are total 6 different *cell* types used in the simulations-

Medium: all *cells* with unassigned *cell* type are medium cells. These cells act as free, uninterrupted space in the simulation space.

Cancer: these *cells* are initially situated at the centre grid surrounded by basement membrane (BM). Cancer *cells* start as rectangular objects of 16-unit volume (4 × 4 pixels), spanning a total 14 × 14 × 1 unit volume (16 cancer *cells*) at the center (*x, y* = 43:57) of the simulation grid without any intercellular space. Cancer is the only *cell* type that is allowed to grow and proliferate. To differentiate from non-cellular components, cancer *cells* have membrane fluctuation unlike others. The *cells* also have random motility associated with them. An 'ExternalPotential' plugin was used to enable 'cellMotility' to act on the cancer *cells* through the Hamiltonian.

$$\Delta \boldsymbol{H}_{\boldsymbol{AM}} = -\vec{\boldsymbol{F}}_{\boldsymbol{\sigma}(i)} \cdot \vec{\boldsymbol{r}}_{\boldsymbol{ij}}$$

for a given MCS step, when an attempt is made to copy a pixel from "I" to "j" for a *cell* $\sigma(i)$, the force vector is $\vec{F}_{\sigma(i)}$ and the distance vector between those pixels is $\vec{r}_{ij}$. Therefore, the product of these two vectors with correct direction alignment will satisfy the condition to minimize H and will determine the movement of the cell along

that direction. The steppable provides an external force on the center of mass of the *cells* which changes direction randomly every MCS, therefore mobilizing the *cells* in different directions and providing them with active motility like biological cells. The direction of $\vec{F}_{\sigma(i)}$ is randomly changed at any angle for each *cell* in each MCS; thereby favoring motility in random directions for all *cells* for a short period and changing direction again.

BM (laminin): This *cell* type surrounds the cancer *cells* in the initial configuration. 2 layers of tightly packed BM *cells* (x, y = 37:63; 3 × 3 pixels= 9 unit volume) separate the cancer *cells* from C1 *cells*. To approximate extracellular matrix (ECM) architecture surrounding luminal epithelial cells in mammary duct, BM *cells* are modelled as dense adhesive bloblike objects to mimic the lamina densa of basal lamina.

(Collagen) C1: Outside BM *cells,* fibrillar C1 *cells* mimics the interconnected fibres of collagen I. C1 *cells* span the remaining space of the simulation lattice (x, y = 63:100). Interfibrillar gaps are characteristic of the C1 *cell* region unlike BM. The gaps aid in non-proteolytic cancer *cell* motility through C1 layer. In the initial configuration of the simulation, elongated C1 *cells* are orientated in random directions with each *cells* having 4 x 2 =8 unit volume.

C_lysed: This *cell* type is used in an intermediate step during matrix degradation and regeneration. Reaction-diffusion dynamics of the chemicals secreted by cancer *cells* allows for degradation of BM and C1 matrix *cells*. Upon meeting certain criteria for degradation, the BM or C1 *cell* type of a particular *cell* becomes C_lysed, although retaining the shape and size of that *cell*. These *cells* track the MCS from their individual degradation event and transform into newly synthesised matrix *cells* after 10 MCS. They have properties specified as intermediates of C1 and medium *cell* types.

NC1: Designed to mimic the **‘**cancer matrisome’, the newly synthesized matrix *cells* are denoted as NC1. These cells are almost like C1 in their behavior and can undergo further degradation to become C_lysed and subsequently after 10 MCS

would become NC1 again. If undisturbed the C_lysed and NC1 *cell* type transformations keep the position, size and shape of the *cell* unchanged.

Contact energies:

As the whole simulation lattice is divided into *cells*, the contact energy is used to regulate the sorting of all the *cells*. Specific contact energy values are assigned to all pairs of interactions between different and same *cell* types. The contact energy is also *inversely proportional* to adhesion between the respective *cell* type pair or components of the system. As there are a total of 6 *cell* types in the simulation, there are total 21 contact energies that need to be assigned and the values determine the differential adhesion. The values of the contact energies were set using control and validation experiments from our previous publication (Pally et al, Frontiers, 2019). 3 among those 21 contact energy- or adhesion- parameters were selected as input variables, which would be changed during further analysis. These 3 contact energies or adhesions are associated with pairs of *cell* types of cancer *cells* (**c-c adh**), cancer *cells* and BM (**c-lam adh**), cancer *cells* and C1 (**c-c1 adh**) since in our previous efforts we had identified the combination of other contact energies (such as those between BM and Type 1 collagen that were permissive for invasion-based simulations.

Reaction diffusion:

CC3D allows chemical fields to determine specific spatiotemporal *cellular* behaviour during simulation. The fields contain the values of concentration of the chemical at each location of the simulation grid. Two chemicals, A and I, are used as activator and its inhibitor as per reaction diffusion dynamics and their concentrations are governed by the partial differential equations (PDE). The governing equations for these two fields are:

$$\frac{\partial[A]}{\partial t} = D_A \nabla^2 [A] + a - \delta_A [A] \qquad (1)$$

$$\frac{\partial[I]}{\partial t} = D_I \nabla^2 [I] + b - \delta_I [I] \qquad (2)$$

$$b = a = K - (c * [I] - d * [A]) \qquad (3)$$

Where, [A], [I]: concentration values for fields A and I.

D_A , D_I : diffusion constants of A and I

δ_A , δ_I : degradation rates of A and I

a, b : secretion rate of A and I

t ≡ MCS

Default parameterizations, D_A= input variable, D_I= 0.04, δ_A= δ_I= 0.003, K= 2.0, c= 4.0, d= 2.0

Here A is considered as the activated form of matrix metalloproteinases (MMPs) and I is considered as tissue inhibitor of matrix metalloproteinases (TIMP). The difference in diffusion constants of A and I formulates the pattern in Turing space, so altering one of them (in this case, D_A) can result in different RD dynamics for different simulations. Activation (or secretion of the activated form i.e. 'a') of A is assumed to be dependent on its inhibitors (inversely) and on its own concentration (autocatalysis). The cancer *cells* secrete both A and I when they come in contact with matrix cells such as C1, BM and NC1. Their concentrations are also calculated at the center of mass of the matrix cells.

If the ratio ([A]/[I]) of concentrations of A and I at the center of mass of any matrix cell is more than a threshold, then that matrix cell is degraded. The threshold value for the ratio is termed as RD cooperativity and is one of the input variables. The value of RD cooperativity signifies the inhibition effect of the inhibitor on the activator's activity (matrix degradation) during the course of a simulation. After degradation and conversion to C_lysed *cell* type, the *cell* becomes NC1 type in 10 MCS which undergoes same treatment as matrix *cells* again. This regeneration of matrix is essential to eliminate unnecessary free spaces formed as an artefact of matrix degradation which takes the computational model closer to its experimental counterpart. Volume of all the 'C_lysed' *cell* types are subjected to 0.1 unit volume decrease at each MCS to mimic dissipation of degraded matrix materials *in vivo*.

<u>Growth and proliferation:</u>

Cancer *cells* are designed to grow in linear combination of two processes.

$$\frac{dV}{dt} = G * [g * p + [GF] * q]$$

Where V = volume of cancer *cell*

g = measure of nutrient availability

[GF] = concentration of growth factor (GF) at center of mass of 'CELL'

p, q = constants

G= input variable

The common surface area of a cancer *cell* with its neighboring cancer cell (k) and the total *cell* surface area (s) is accessed to calculate *g* in this equation as *g=(s-k)/40.* The denominator in the calculation of *g* is due to 2D nature of the simulation as a *cell* can be surrounded by other *cells* only in xy plane and not in z axis. The scaling of that extra cell surface area without any neighboring cells in z axis is provided by the denominator. Another contributor of the growth function is *[GF]* which mimics the ECM-degradation dependence of growth and proliferation (Olivares et al., 2017). The 'C_lysed' *cell* type is programmed to secrete GF at each of its pixel location where the diffusion constant is kept low (0.02) to localize this growth signal to areas of matrix degradation. p (=1/12) and q (=1/21) constant values are set according to the assumed weightage of the two variables in growth equation. Value of G, then determines how much the resultant growth of the *cell* should be. The linear combination of the two contributors is multiplied by G, hence, G is considered as an input variable.

Cell division is incorporated into the cancer *cells* by a CC3D steppable called 'MitosisSteppable' with base function 'MitosisSteppableBase'. If any cancer *cell* reaches a threshold volume of 30 units then that *cell* will be divided in random orientation. The resultant two cells will have volumes half of its predecessor with all other properties kept same as the parent cell. In this model, growth rate is directly correlated to proliferation as it determines the volume of the cell to reach threshold for cell division.

<u>Parameter scan:</u>

CC3D's parameter scan feature was used to explore parameter space of the input variables. All the combinations of provided values of the parameters are considered for each individual simulation.

Here are the 5 values of the 6 input variables for a parameter scan:

RD cooperativity: 2, 3, 5, 7, 10

MMP (A) diffusion constant: 0.005, 0.01, 0.03, 0.04, 0.08

Cell-cell contact energy: 5, 15, 25, 35, 50

Cell-laminin contact energy: 5, 15, 25, 35, 50

Cell-C1 contact energy: 5, 15, 25, 35, 50

Cell Growth (G): 0, 0.2, 0.5, 0.7, 1

3 replicates of the parameter scan were performed.

Matlab-based analysis:

Quantification of invasion: [Fig1(B)]

Graphical representation of all the *cells* in the simulation grid is captured as images at different MCSs for each simulation. A certain 'timepoint' or MCS (=600) is chosen so that MATLAB accesses all the images of all simulations at MCS 600 and provides quantification. The choice of the 600th MCS is guided by the fact that this represents the stage where collective invasion can clearly be discerned from the dispersed cells. Beyond this stage, the dispersed cells move out of our framework window and the collectively invasive mass fills the same, disallowing distinctions to be made between invasions of different scales.

No. of dispersed objects: The collected simulation images are originally in RGB color scheme to differentiate *cell* types as displayed in Fig1(C). Individual images from all simulations are batch-processed to generate binary images with respect to cancer *cell* type (red color). The first set of binarized images undergo series of rudimentary Matlab image processing operations to generate different set of binary images for different analysis. From them, all the boundaries in each image are then identified to recognize areas which are enclosed by a continuous non-overlapping boundary. These areas without shared boundaries originally represent cancer *cell* clusters in the respective simulation grid. These clusters can contain single to multiple cancer *cells*. the number of these clusters denote the dispersiveness of the initial tumor mass in the simulation pertaining to certain combination of input variables, therefore considered as one of the output variables i.e. no. of dispersed objects. The values of this quantity pertaining to each simulation is plotted on the y-axis of the phenospace.

Area of largest cell cluster: From the sets of binarized images, areas of all dispersed objects are calculated by counting the pixels preset inside each object's

boundaries. The dispersed object having biggest area is considered the main collective cancer *cell* cluster of the respective simulation. The value of the area of such as object in the simulation image is quantified as 'area of largest cell cluster', the other output variable. The values of this quantity pertaining to each simulation is plotted on the x-axis of the phenospace.

K-means clustering:

Broad guidelines for performing the Silhouette analysis were adapted from the following link (https://in.mathworks.com/help/stats/kmeans.html#bues3lh). K-means clustering or Lloyd's algorithm is used for subsequent analysis of the phenospace using MATLAB (Lloyd, 1982). It is an iterative, data-partitioning algorithm that assigns all observations from simulations to exactly one of the K clusters defined by centroids. Being an unsupervised partition-based algorithm, value of K is chosen before the algorithm starts. 'kmeans' function in MATLAB uses a two-phase iterative algorithm to minimize the sum of point-to-centroid distances, summed over all k clusters. The first phase uses batch updates, where each iteration consists of reassigning points to their nearest cluster centroid, all at once, followed by recalculation of cluster centroids. In the second phase, points are individually reassigned, if doing so reduces the sum of distances, and cluster centroids are recomputed after each reassignment, by using feedback updates.

The x and y axis values from the phenospace are normalized (z-score) before the use of 'kmeans' function to provide same weightage to the axes of during clustering and at the end of the run, all points are assigned a cluster. The points are then colour-coded in the phenospace to differentiate hosting cluster identities.

Silhouette analysis:

Broad guidelines for performing the Silhouette analysis were adapted from the following link (https://in.mathworks.com/help/stats/silhouette.html#mw_9fe10316-522d-4a3a-9cdd-d39baea0c1ce). The silhouette value for each point is a measure of how similar that point is to points within its own cluster, when compared to points in other clusters. The silhouette value 'S_i' for the ith point is defined as

$S_i = (b_i - a_i) / \max(a_i, b_i)$

where a_i is the average distance from the ith point to the other points in the same cluster as i, and b_i is the minimum average distance from the ith point to points in a different cluster, minimized over clusters.

The silhouette value ranges from –1 to 1. A high silhouette value indicates that i is well matched to its own cluster, and poorly matched to other clusters. If most points have a high silhouette value, then the clustering solution is appropriate. If many points have a low or negative silhouette value, then the clustering solution might have too many or too few clusters. So, silhouette values are used as a clustering evaluation criterion for finding a clustering pattern which is invariable across replicates.

After a specific K-means clustering is performed on the points in the phenospace, the normalized matrix of x, y axes along with cluster identity matrix (cluster id of each point) is used to compute silhouette values for each point. The values pertaining to all the points in the simulation is represented in silhouette plot for convenience and the mean of the silhouette values in the plot is denoted with a dashed line. The silhouette method was performed on all the replicates of clustering with a constant k value. The mean of the silhouette values pertaining to each replicate is acquired. The mean and standard deviation of the mean silhouette values associated with replicates of different k values are plotted in a box plot.

Principal Component Analysis: Broad guidelines for performing the PCA were adapted from the following link (https://in.mathworks.com/help/stats/pca.html). Briefly, in order to analyse the points in the neighborhoods or patches (denoted by rectangles) in the phenospace, principal component analysis was used in MATLAB. The points within a specific rectangle of the phenospace are selected and their underlying simulation files are identified, and the input variable values associated with them are accessed. The whole input argument dataset for a rectangle's PCA is a matrix which consists all 6 input variable values for individual simulations relating to the points in that rectangle. The matrix is normalized for each variable. Normalization proved to be important as we are interested in relative contribution of the input variable and their changes with respect to other points (simulations), rather than their absolute values. PCA of each rectangle produces different linear combinations of input variables (forms principal component axes) for optimally fitting all points without overlapping in one of those linear relations. PCA quantifies how

much of the input data is ‘explained’ by each linear combination (or PC axis) and represented as a scree plot. The linear combination associated with highest y-axis value in the scree plot is identified as first principal component axis. Therefore, the relative contributions of the input variables for constructing PC1 can give insights into the convoluted dynamics of input variables governing the phenotype. The relative contribution of the input variables is plotted in bar graph on the right side of scree plots.

Codes: All the codes relating to the manuscript can be accessed through the following link: https://www.dropbox.com/s/29b841vg6gkaqjn/iScience_DP_codes_datafiles.zip?dl=0

Acknowledgments:

This work was supported by the DBT [BT/PR26526/GET/119/92/2017] SERB[ECR/2015/000280] and the Wellcome Trust/DBT India Alliance Fellowship/Grant [grant number IA/I/17/2/503312] awarded to RB. DP acknowledges KVPY for the scholarship. MKJ was supported by Ramanujan Fellowship awarded by SERB (SB/S2/RJN-049/2018), DST, Government of India.

References:

Aceto, N., Bardia, A., Miyamoto, D.T., Donaldson, M.C., Wittner, B.S., Spencer, J.A., Yu, M., Pely, A., Engstrom, A., Zhu, H., *et al.* (2014). Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. Cell *158*, 1110-1122.

Alber, M., Chen, N., Glimm, T., and Lushnikov, P.M. (2006). Multiscale dynamics of biological cells with chemotactic interactions: from a discrete stochastic model to a continuous description. Physical review E, Statistical, nonlinear, and soft matter physics *73*, 051901.

Alber, M.S., Kiskowski, M.A., and Jiang, Y. (2004). Two-stage aggregate formation via streams in myxobacteria. Physical review letters *93*, 068102.

Alcaraz, J., Mori, H., Ghajar, C.M., Brownfield, D., Galgoczy, R., and Bissell, M.J. (2011). Collective epithelial cell invasion overcomes mechanical barriers of collagenous extracellular matrix by a narrow tube-like geometry and MMP14-dependent local softening. Integrative biology : quantitative biosciences from nano to macro *3*, 1153-1166.

Anderson, A.R. (2005). A hybrid mathematical model of solid tumour invasion: the importance of cell adhesion. Mathematical medicine and biology : a journal of the IMA *22*, 163-186.
Aref, H., and Tryggvason, G. (1989). Model of Rayleigh-Taylor instability. Physical review letters *62*, 749-752.
Bhat, R., and Bissell, M.J. (2014). Of plasticity and specificity: dialectics of the microenvironment and macroenvironment and the organ phenotype. Wiley interdisciplinary reviews Developmental biology *3*, 147-163.
Bocci, F., Kumar Jolly, M., and Onuchic, J.N. (2019). A Biophysical Model Uncovers the Size Distribution of Migrating Cell Clusters across Cancer Types. Cancer research *79*, 5527-5535.
Bronsert, P., Enderle-Ammour, K., Bader, M., Timme, S., Kuehs, M., Csanadi, A., Kayser, G., Kohler, I., Bausch, D., Hoeppner, J.*, et al.* (2014). Cancer cell invasion and EMT marker expression: a three-dimensional study of the human cancer-host interface. The Journal of pathology *234*, 410-422.
Cheung, K.J., Gabrielson, E., Werb, Z., and Ewald, A.J. (2013). Collective invasion in breast cancer requires a conserved basal epithelial program. Cell *155*, 1639-1651.
Christley, S., Alber, M.S., and Newman, S.A. (2007). Patterns of mesenchymal condensation in a multiscale, discrete stochastic model. PLoS computational biology *3*, e76.
Das, A., Monteiro, M., Barai, A., Kumar, S., and Sen, S. (2017). MMP proteolytic activity regulates cancer invasiveness by modulating integrins. Scientific reports *7*, 14219.
Di Lullo, G.A., Sweeney, S.M., Korkko, J., Ala-Kokko, L., and San Antonio, J.D. (2002). Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen. The Journal of biological chemistry *277*, 4223-4231.
Diambra, L., Senthivel, V.R., Menendez, D.B., and Isalan, M. (2015). Cooperativity to increase Turing pattern space for synthetic biology. ACS synthetic biology *4*, 177-186.
Friedl, P., Hegerfeldt, Y., and Tusch, M. (2004). Collective cell migration in morphogenesis and cancer. The International journal of developmental biology *48*, 441-449.
Friedl, P., Locker, J., Sahai, E., and Segall, J.E. (2012). Classifying collective cancer cell invasion. Nature cell biology *14*, 777-783.
Friedl, P., and Wolf, K. (2008). Tube travel: the role of proteases in individual and collective cancer cell invasion. Cancer research *68*, 7247-7249.
Gadea, G., Sanz-Moreno, V., Self, A., Godi, A., and Marshall, C.J. (2008). DOCK10-mediated Cdc42 activation is necessary for amoeboid invasion of melanoma cells. Current biology : CB *18*, 1456-1465.
Glazier, J.A., and Graner, F. (1993). Simulation of the differential adhesion driven rearrangement of biological cells. Physical review E, Statistical physics, plasmas, fluids, and related interdisciplinary topics *47*, 2128-2154.
Graner, F., and Glazier, J.A. (1992). Simulation of biological cell sorting using a two-dimensional extended Potts model. Physical review letters *69*, 2013-2016.
Graves, M.L., Cipollone, J.A., Austin, P., Bell, E.M., Nielsen, J.S., Gilks, C.B., McNagny, K.M., and Roskelley, C.D. (2016). The cell surface mucin podocalyxin regulates collective breast tumor budding. Breast cancer research : BCR *18*, 11.
Grigore, A.D., Jolly, M.K., Jia, D., Farach-Carson, M.C., and Levine, H. (2016). Tumor Budding: The Name is EMT. Partial EMT. Journal of clinical medicine *5*.

Haeger, A., Alexander, S., Vullings, M., Kaiser, F.M.P., Veelken, C., Flucke, U., Koehl, G.E., Hirschberg, M., Flentje, M., Hoffman, R.M.*, et al.* (2020). Collective cancer invasion forms an integrin-dependent radioresistant niche. J Exp Med *217*.
Harrison, N.C., Diez del Corral, R., and Vasiev, B. (2011). Coordination of cell differentiation and migration in mathematical models of caudal embryonic axis extension. PloS one *6*, e22700.
Hase, K., Shatney, C., Johnson, D., Trollope, M., and Vierra, M. (1993). Prognostic value of tumor "budding" in patients with colorectal cancer. Diseases of the colon and rectum *36*, 627-635.
Hecht, I., Bar-El, Y., Balmer, F., Natan, S., Tsarfaty, I., Schweitzer, F., and Ben-Jacob, E. (2015). Tumor invasion optimization by mesenchymal-amoeboid heterogeneity. Scientific reports *5*, 10622.
Huang, B., Jolly, M.K., Lu, M., Tsarfaty, I., Ben-Jacob, E., and Onuchic, J.N. (2015). Modeling the Transitions between Collective and Solitary Migration Phenotypes in Cancer Metastasis. Scientific reports *5*, 17379.
Huang, B., Lu, M., Jolly, M.K., Tsarfaty, I., Onuchic, J., and Ben-Jacob, E. (2014). The three-way switch operation of Rac1/RhoA GTPase-based circuit controlling amoeboid-hybrid-mesenchymal transition. Scientific reports *4*, 6449.
Hynes, R.O. (2009). The extracellular matrix: not just pretty fibrils. Science *326*, 1216-1219.
Izaguirre, J.A., Chaturvedi, R., Huang, C., Cickovski, T., Coffland, J., Thomas, G., Forgacs, G., Alber, M., Hentschel, G., Newman, S.A.*, et al.* (2004). CompuCell, a multi-model framework for simulation of morphogenesis. Bioinformatics *20*, 1129-1137.
Jolly, M.K., Boareto, M., Debeb, B.G., Aceto, N., Farach-Carson, M.C., Woodward, W.A., and Levine, H. (2017). Inflammatory breast cancer: a model for investigating cluster-based dissemination. NPJ breast cancer *3*, 21.
Jolly, M.K., Somarelli, J.A., Sheth, M., Biddle, A., Tripathi, S.C., Armstrong, A.J., Hanash, S.M., Bapat, S.A., Rangarajan, A., and Levine, H. (2019). Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. Pharmacology & therapeutics *194*, 161-184.
Kedrin, D., Gligorijevic, B., Wyckoff, J., Verkhusha, V.V., Condeelis, J., Segall, J.E., and van Rheenen, J. (2008). Intravital imaging of metastatic behavior through a mammary imaging window. Nature methods *5*, 1019-1021.
Kiskowski, M.A., Jiang, Y., and Alber, M.S. (2004). Role of streams in myxobacteria aggregate formation. Physical biology *1*, 173-183.
Konishi, T. (2015). Principal component analysis for designed experiments. BMC bioinformatics *16 Suppl 18*, S7.
Krakhmal, N.V., Zavyalova, M.V., Denisov, E.V., Vtorushin, S.V., and Perelmuter, V.M. (2015). Cancer Invasion: Patterns and Mechanisms. Acta naturae *7*, 17-28.
Labernadie, A., Kato, T., Brugues, A., Serra-Picamal, X., Derzsi, S., Arwert, E., Weston, A., Gonzalez-Tarrago, V., Elosegui-Artola, A., Albertazzi, L.*, et al.* (2017). A mechanically active heterotypic E-cadherin/N-cadherin adhesion enables fibroblasts to drive cancer cell invasion. Nature cell biology *19*, 224-237.
Lintz, M., Munoz, A., and Reinhart-King, C.A. (2017). The Mechanics of Single Cell and Collective Migration of Tumor Cells. Journal of biomechanical engineering *139*.
Liu, R., Song, K., Hu, Z., Cao, W., Shuai, J., Chen, S., Nan, H., Zheng, Y., Jiang, X., Zhang, H.*, et al.* (2019). Diversity of collective migration patterns of invasive breast cancer cells emerging during microtrack invasion. Physical review E *99*, 062403.

Lloyd, S.P. (1982). Least Squares Quantization in PCM. IEEE Transactions on Information Theory *28*, 129-137.
Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proc 5th Berkeley Symp Math Stat Probab *1965-66*, 281-297.
Madsen, C.D., and Sahai, E. (2010). Cancer dissemination--lessons from leukocytes. Developmental cell *19*, 13-26.
Masaki, T., Sugiyama, M., Matsuoka, H., Abe, N., Izumisato, Y., Sakamoto, A., and Atomi, Y. (2003). Coexpression of matrilysin and laminin-5 gamma2 chain may contribute to tumor cell migration in colorectal carcinomas. Digestive diseases and sciences *48*, 1262-1267.
Merks, R.M., and Glazier, J.A. (2006). Dynamic mechanisms of blood vessel growth. Nonlinearity *19*, C1-C10.
Mikaelian, K.O. (1990). Rayleigh-Taylor and Richtmyer-Meshkov instabilities in multilayer fluids with surface tension. Physical review A, Atomic, molecular, and optical physics *42*, 7211-7225.
Nagai, T., Ishikawa, T., Minami, Y., and Nishita, M. (2020). Tactics of cancer invasion: solitary and collective invasion. Journal of biochemistry *167*, 347-355.
Nelson, C.M., and Bissell, M.J. (2005). Modeling dynamic reciprocity: engineering three-dimensional culture models of breast architecture, function, and neoplastic transformation. Seminars in cancer biology *15*, 342-352.
Nelson, C.M., Vanduijn, M.M., Inman, J.L., Fletcher, D.A., and Bissell, M.J. (2006). Tissue geometry determines sites of mammary branching morphogenesis in organotypic cultures. Science *314*, 298-300.
Newman, S.A., and Bhat, R. (2008). Dynamical patterning modules: physico-genetic determinants of morphological development and evolution. Physical biology *5*, 015008.
Newman, S.A., and Bhat, R. (2009). Dynamical patterning modules: a "pattern language" for development and evolution of multicellular form. The International journal of developmental biology *53*, 693-705.
Nieto, M.A., Huang, R.Y., Jackson, R.A., and Thiery, J.P. (2016). Emt: 2016. Cell *166*, 21-45.
Nissen, N.I., Karsdal, M., and Willumsen, N. (2019). Collagens and Cancer associated fibroblasts in the reactive stroma and its relation to Cancer biology. Journal of experimental & clinical cancer research : CR *38*, 115.
Olivares, O., Mayers, J.R., Gouirand, V., Torrence, M.E., Gicquel, T., Borge, L., Lac, S., Roques, J., Lavaut, M.N., Berthezene, P.*, et al.* (2017). Collagen-derived proline promotes pancreatic ductal adenocarcinoma cell survival under nutrient limited conditions. Nature communications *8*, 16031.
Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., and Adebiyi, E. (2016). Clustering Algorithms: Their Application to Gene Expression Data. Bioinformatics and biology insights *10*, 237-253.
Pally, D., Pramanik, D., and Bhat, R. (2019). An Interplay Between Reaction-Diffusion and Cell-Matrix Adhesion Regulates Multiscale Invasion in Early Breast Carcinomatosis. Frontiers in physiology *10*, 790.
Pankova, K., Rosel, D., Novotny, M., and Brabek, J. (2010). The molecular mechanisms of transition between mesenchymal and amoeboid invasiveness in tumor cells. Cellular and molecular life sciences : CMLS *67*, 63-71.
Patsialou, A., Bravo-Cordero, J.J., Wang, Y., Entenberg, D., Liu, H., Clarke, M., and Condeelis, J.S. (2013). Intravital multiphoton imaging reveals multicellular streaming

as a crucial component of in vivo cell migration in human breast tumors. Intravital *2*, e25294.
Petrova, E., Zielinski, V., Bolm, L., Schreiber, C., Knief, J., Thorns, C., Bronsert, P., Timme-Bronsert, S., Bausch, D., Perner, S.*, et al.* (2019). Tumor budding as a prognostic factor in pancreatic ductal adenocarcinoma. Virchows Archiv : an international journal of pathology.
Pickup, M.W., Mouw, J.K., and Weaver, V.M. (2014). The extracellular matrix modulates the hallmarks of cancer. EMBO reports *15*, 1243-1253.
Prall, F. (2007). Tumour budding in colorectal carcinoma. Histopathology *50*, 151-162.
Prall, F., Nizze, H., and Barten, M. (2005). Tumour budding as prognostic factor in stage I/II colorectal carcinoma. Histopathology *47*, 17-24.
Prall, F., and Ostwald, C. (2007). High-degree tumor budding and podia-formation in sporadic colorectal carcinomas with K-ras gene mutations. Hum Pathol *38*, 1696-1702.
Prall, F., Ostwald, C., and Linnebacher, M. (2009). Tubular invasion and the morphogenesis of tumor budding in colorectal carcinoma. Hum Pathol *40*, 1510-1512.
Sadati, M., Taheri Qazvini, N., Krishnan, R., Park, C.Y., and Fredberg, J.J. (2013). Collective migration and cell jamming. Differentiation; research in biological diversity *86*, 121-125.
Salhia, B., Trippel, M., Pfaltz, K., Cihoric, N., Grogg, A., Ladrach, C., Zlobec, I., and Tapia, C. (2015). High tumor budding stratifies breast cancer with metastatic properties. Breast cancer research and treatment *150*, 363-371.
Sanz-Moreno, V., Gadea, G., Ahn, J., Paterson, H., Marra, P., Pinner, S., Sahai, E., and Marshall, C.J. (2008). Rac activation and inactivation control plasticity of tumor cell movement. Cell *135*, 510-523.
Sepulveda, N., Petitjean, L., Cochet, O., Grasland-Mongrain, E., Silberzan, P., and Hakim, V. (2013). Collective cell motion in an epithelial sheet can be quantitatively described by a stochastic interacting particle model. PLoS computational biology *9*, e1002944.
Spaderna, S., Schmalhofer, O., Hlubek, F., Berx, G., Eger, A., Merkel, S., Jung, A., Kirchner, T., and Brabletz, T. (2006). A transient, EMT-linked loss of basement membranes indicates metastasis and poor survival in colorectal cancer. Gastroenterology *131*, 830-840.
Swat, M.H., Thomas, G.L., Belmonte, J.M., Shirinifard, A., Hmeljak, D., and Glazier, J.A. (2012). Multi-scale modeling of tissues using CompuCell3D. Methods in cell biology *110*, 325-366.
Theveneau, E., and Linker, C. (2017). Leaders in collective migration: are front cells really endowed with a particular set of skills? F1000Research *6*, 1899.
van Zijl, F., Krupitza, G., and Mikulits, W. (2011). Initial steps of metastasis: cell invasion and endothelial transmigration. Mutation research *728*, 23-34.
Wolf, K., Wu, Y.I., Liu, Y., Geiger, J., Tam, E., Overall, C., Stack, M.S., and Friedl, P. (2007). Multi-step pericellular proteolysis controls the transition from individual to collective cancer cell invasion. Nature cell biology *9*, 893-904.
Yeung, K.Y., and Ruzzo, W.L. (2001). Principal component analysis for clustering gene expression data. Bioinformatics *17*, 763-774.
Zhang, Y., Thomas, G.L., Swat, M., Shirinifard, A., and Glazier, J.A. (2011). Computer simulations of cell sorting due to differential adhesion. PloS one *6*, e24999.

Zhao, S., Sun, J., Shimizu, K., and Kadota, K. (2018). Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results. Biological procedures online *20*, 5.

Figure Legends

Figure 1: Introduction to the computational framework, simulation inputs and output phenospace

(A) Schematic depiction of the input variables that have been deployed in the CompuCell 3D simulations. The result of the simulations was computed as two outputs that are representative of dispersed invasion and collective invasion.

(B) Depiction of computation of the invasion outputs: MATLAB® is used to quantify the morphology of the cancer cell mass in the simulations. 1. Images of all simulations are collected at the 600th Monte Carlo step (MCS). 2. The images are binarized to isolate the cancer cells. 3. All dispersed 'objects' (the objects share no common boundary with each other) are identified and counted and is called 'No. of dispersed objects'. 4. the mass having largest total covered area is isolated and the value of that highest area is called 'Area of Largest cell cluster'.

(C) Control simulations: Four simulations (red= transformed cancer cell, green= collagen I, blue= Basement membrane (BM)) represent control runs showing morphological variations with certain input variable values. In terms of invasiveness, they are characterized as **dispersed invasion** (blue), **collective invasion** (red), multiscale invasion (magenta) and **indolent or non-invasive phenotype** (green). The invasiveness of the controls interpreted through the 2 outputs show in 1B allows the construction of a phenospace with X axis measuring the 'Area of Largest cell cluster' and Y axis measuring the 'No. of dispersed objects'. Control simulations can therefore be mapped onto the phenospace of possible output points within the phenospace through various combinations of inputs deployed through simulations run till a similar endpoint (MCS:600).

(D) Distribution within the phenospace of phenotype outputs as a result of combination of five values for each of the six inputs mentioned in (A) and run

three times (replicates). The total number of simulations originating from this combination is 5^6X3=46875. Simulations points having x value more than around 36000 (~90% of the whole simulation lattice) is removed to omit the effect of spatial limitation of the simulation lattice.

Figure 2: The invasion phenotypic outputs can be optimally segregated into three clusters

(A) All the simulation outputs in the phenospace (in Figure 1D) are analyzed with K-means clustering using cluster number K=2-4 (leftmost column). Clustering was performed 15 times with random initial cluster centres. The cluster phenotype patterns are shown (middle column) along with their proportional representation within the 15 replicates (right column). (Cluster patterns for K=5-8 shown in supplementary figure X)

(B) The optimal value for K computed using the Silhouette method. Silhouette plots for K=2-4 are constructed with X axis signifying silhouette values and Y axis representing cluster numbers. (Silhouette plots for K=5-8 shown in supplementary figure 1)

(C) The average silhouette values and associated standard deviation for all 15 clustering replicates in different K-value groups are plotted in the box and whisker plot (bottom right). Statistical test was performed using unpaired ANOVA with Tukey's post hoc multiple comparison ($p<0.0001$) (for multiple comparisons data see Supplementary File 1)

Figure 3: Tumor budding is characteristic of dispersed MSI; mesenchymal streaming is characteristic of collective MSI

(A) The end point of a simulation output chosen within the dispersed MSI cluster close to the Y axis shows classical single cell-dispersed invasion phenotype with the primary focus surrounded by several single cells dispersed within the collagen microenvironment.

(B) For an output within the dispersed MSI cluster close to the indolent cluster, instead of single cells, smaller oligocellular clusters can be seen dispersed within collagen reminiscent of tumor budding (Prall, 2007).

(C) The end point of a simulation output chosen within the collective MSI cluster close to the X axis shows classical collective invasion phenotype with the primary focus growing centrifugally into the surrounding collagen microenvironment.

(D) For a given output within the collective MSI cluster close to the indolent cluster, along with the primary focal mass several finger-like projections can be well appreciated, that has been described as mesenchymal streaming (Friedl et al., 2012).

Figure 4: Principal component analysis of outputs near boundary between dispersed- and collective MSI clusters:

Principal component analysis is performed with outputs within 4 similar sized rectangles in the dispersed MSI cluster (depicted as D_{Red}, D_{Yellow}, D_{Blue} and D_{Violet}) and the collective MSI cluster (depicted as C_{Red}, C_{Yellow}, C_{Blue} and C_{Violet}) close to the boundary between the two clusters. Corresponding to data-points captured in each box, the extent of variance explained by principal components is mentioned. Given alongside are the coefficients of different input variables in the first principal component axis (PC1). The y-axis indicates the weight coefficient of respective variable.

Figure 5: Using the principal component analysis to induce transition between phenotypic clusters

A random simulation output is selected from each of the rectangles from the dispersed and collective MSI clusters created in Figure 3 and one of the input variables signatures cognate to the clusters is changed. The resultant transitioned outputs are identified and highlighted corresponding to each starting output. For each of the transition, we have depicted one specific transition from a beginning output to its transitioned end point. We also show the phenotypes at the end of the simulation (MCS=600) at the beginning and the end of the transition.

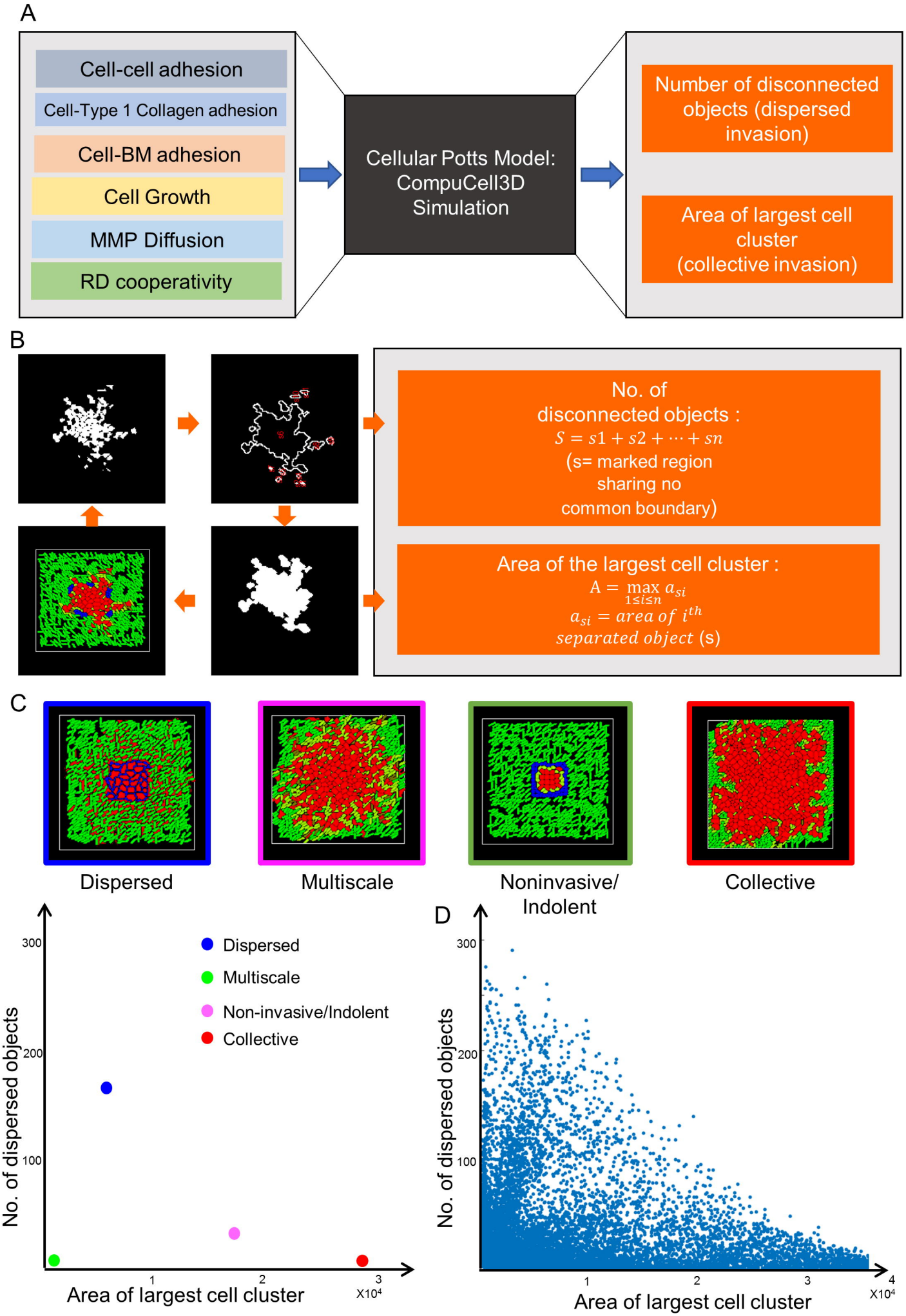
(A) shows the transition from a collective MSI to dispersed MSI when cell-BM adhesion is decreased (4 of 6 outputs showed successful transition; transition from red rectangle is highlighted).

(B) shows the transition from dispersed MSI to collective MSI when MMP diffusivity is increased (all 6 outputs showed successful transition; transition from pink rectangle is highlighted)
(C) shows the transition from dispersed MSI to indolently invasive cluster when cell-BM adhesion is decreased (5 of 6 outputs showed successful transition; transition from red rectangle is highlighted).

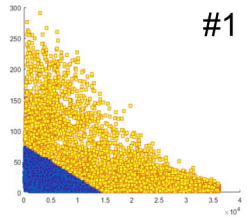
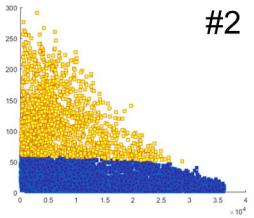
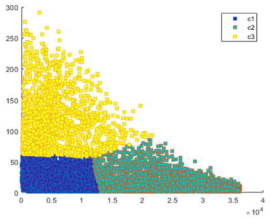
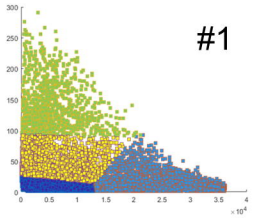
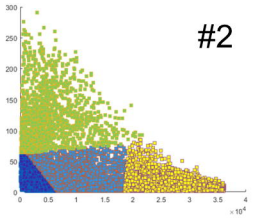
Table 1: Common patterns of input variable coefficients associated with the eight rectangles in Figure 4 and 6 rectangles in Supplementary Figure 2.

Supplementary Figure 1: (A) All the simulation outputs in the phenospace (in Figure 1D) are analyzed with K-means clustering using cluster number K=5-8 (leftmost column). Clustering was performed 15 times with random initial cluster centers. The cluster phenotype patterns are shown (middle column) along with their proportional representation within the 15 replicates (right column). (B) The optimal value for K computed using the Silhouette method. Silhouette plots for K=5-8 are constructed with X axis signifying silhouette values and Y axis representing cluster numbers.

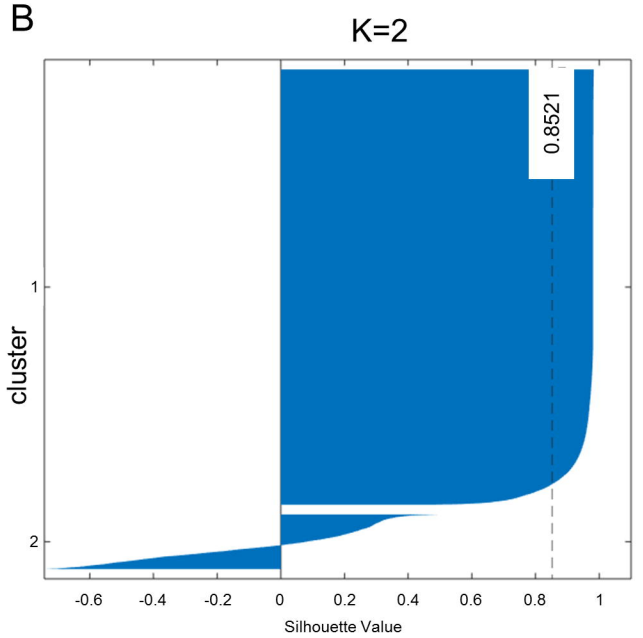
Supplementary Figure 2: Principal component analysis is performed with outputs within similar sized rectangles in the dispersed MSI cluster (depicted as D_{Green}, D_{Brown}, D_{Pink} and D_{Black}) and the collective MSI cluster (depicted as C_{Green}, and C_{Pink}) close to the boundary between the two clusters. Corresponding to data-points captured in each box, the extent of variance explained by principal components is mentioned. Given alongside are the coefficients of different input variables in the first principal component axis (PC1). The y-axis indicates the weight coefficient of respective variable.

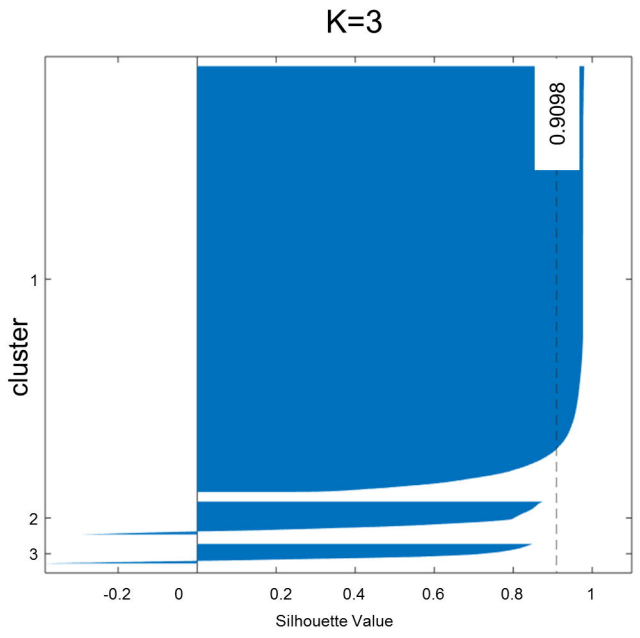

A
Cell-cell adhesion
Cell-Type 1 Collagen adhesion
Cell-BM adhesion
Cell Growth
MMP Diffusion
RD cooperativity
Cellular Potts Model: CompuCell3D Simulation
Number of disconnected objects (dispersed invasion)
Area of largest cell cluster (collective invasion)
B
No. of disconnected objects :
$S = s1 + s2 + \cdots + sn$
(s= marked region sharing no common boundary)
Area of the largest cell cluster :
$A = \max_{1 \le i \le n} a_{si}$
$a_{si} = area\ of\ i^{th}$ $separated\ object$ (s)
C
Dispersed
Multiscale
Noninvasive/ Indolent
Collective
Dispersed
Multiscale
Non-invasive/Indolent
Collective
No. of dispersed objects
Area of largest cell cluster
X10⁴
D
No. of dispersed objects
Area of largest cell cluster
X10⁴

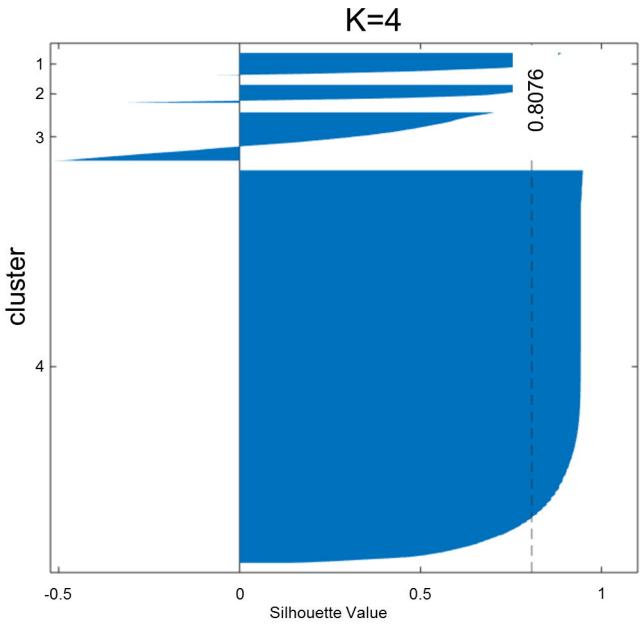
A

K value	Cluster Pattern		Proportional representation of each pattern (%)	
K=2	 #1	 #2	#1 80	#2 20
K=3			100	
K=4	 #1	 #2	#1 60	#2 40

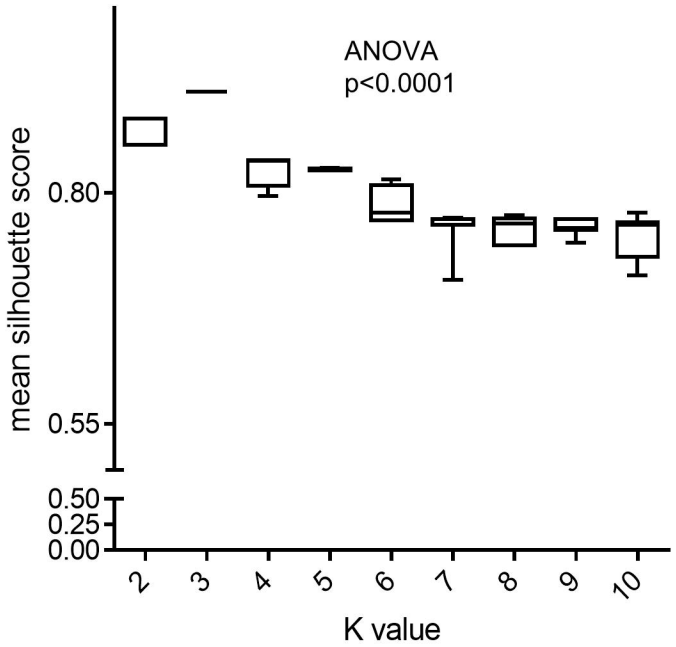
B

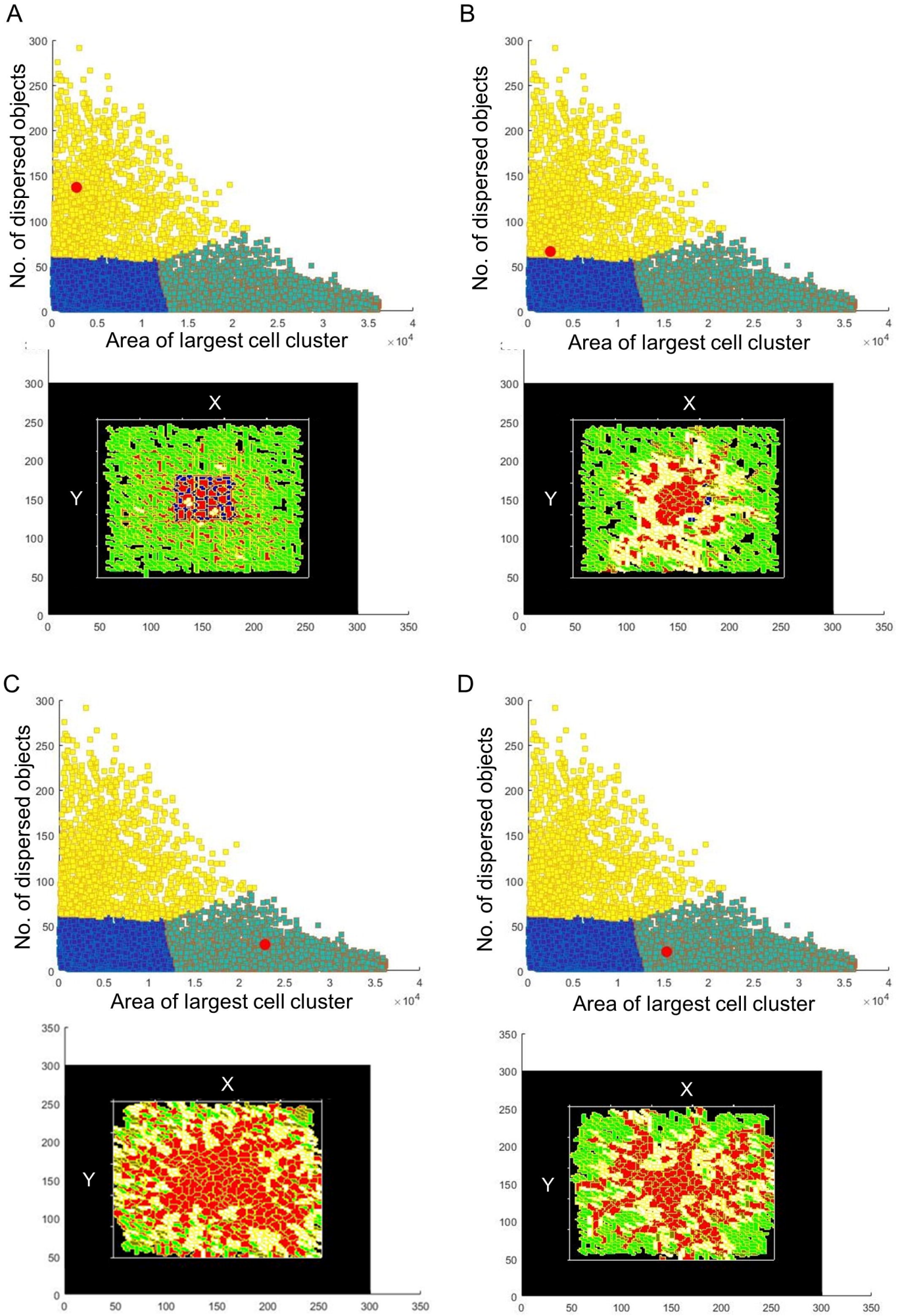


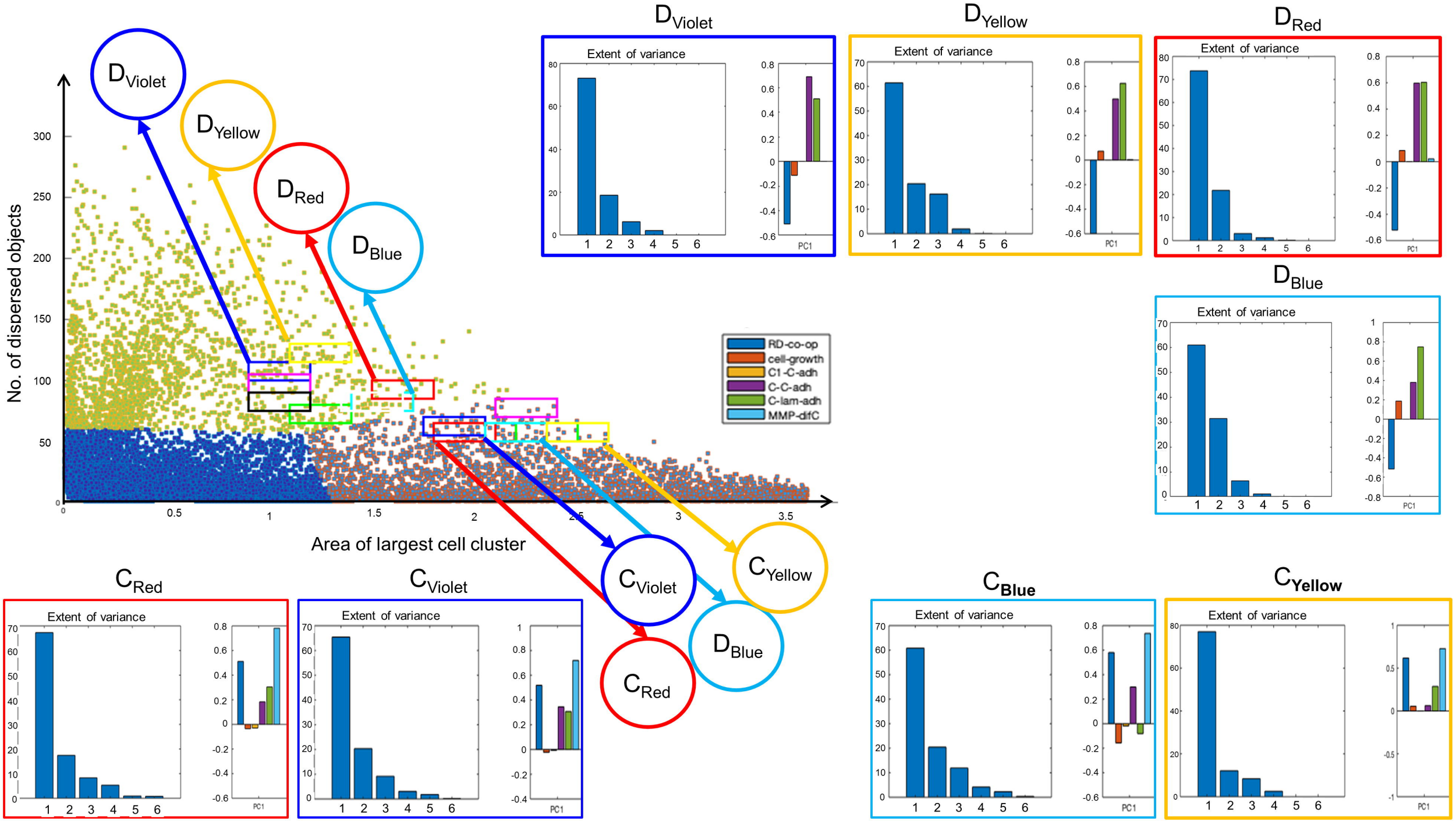


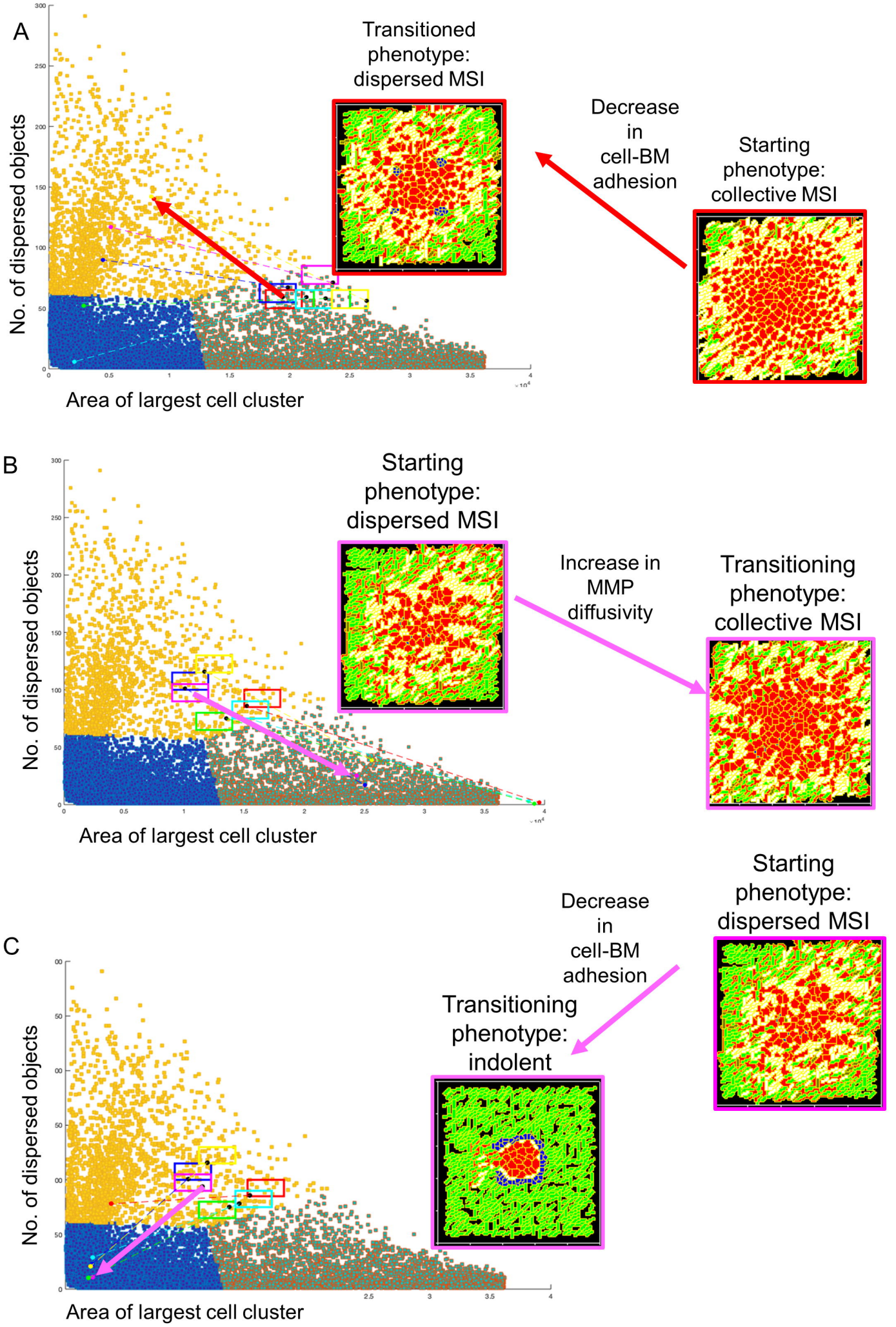


C




A
B
C
D
No. of dispersed objects
Area of largest cell cluster
×10^4
X
Y


D Violet
D Yellow
D Red
D Blue
No. of dispersed objects
Area of largest cell cluster
RD-co-op
cell-growth
C1-C-adh
C-C-adh
C-lam-adh
MMP-dfC
C Violet
C Yellow
C Red
C Blue
Extent of variance
PC1


A
Transitioned phenotype: dispersed MSI
Decrease in cell-BM adhesion
Starting phenotype: collective MSI
No. of dispersed objects
Area of largest cell cluster
B
Starting phenotype: dispersed MSI
Increase in MMP diffusivity
Transitioning phenotype: collective MSI
No. of dispersed objects
Area of largest cell cluster
C
Starting phenotype: dispersed MSI
Decrease in cell-BM adhesion
Transitioning phenotype: indolent
No. of dispersed objects
Area of largest cell cluster

Collective-MSI	RD cooperativity	Cell proliferation	Cell-Type 1 collagen adhesion	Cell-cell adhesion	Cell-BM adhesion	MMP diffusion rate
1	high2	median	median	median	high	high3
2	high2	median	median	high	median	high3
3	high2	median	median	high	high	high2
4	high2	median	median	high	median	high3
5	high3	median	median	low2	low3	median
6	high2	median	median	median	high	high3

Dispersed-MSI	RD cooperativity	Cell proliferation	Cell-Type 1 collagen adhesion	Cell-cell adhesion	Cell-BM adhesion	MMP diffusion rate
1	low2	median	median	high2	high2	median
2	high2	median	median	high	high	high3
3	low2	median	median	high3	high2	median
4	low2	median	median	high	high3	median
5	low	median	median	high3	high2	median
6	low3	median	median	high2	high3	median
7	high3	median	median	high2	median	high2
8	high2	median	median	high2	median	high3