**Phenotypic heterogeneity driven by plasticity of the intermediate EMT state governs disease progression and metastasis in breast cancer**

Meredith S. Brown[1], Behnaz Abdollahi[2], Owen M. Wilkins[3], Hanxu Lu[1], Priyanka Chakraborty[4], Nevena B. Ognjenovic[1], Kristen E. Muller[5], Mohit Kumar Jolly[4], Brock C. Christensen[1,6], Saeed Hassanpour[2,3], Diwakar R. Pattabiraman[1,3]

[1] Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Hanover NH 03755 USA

[2] Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover NH 03755 USA

[3] Norris Cotton Cancer Center, Geisel School of Medicine, Lebanon, New Hampshire, USA.

[4] Centre for BioSystems Science and Engineering, Indian Institute of Science, Bengaluru, India 560012

[5] Department of Pathology, Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756 USA

[6] Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756 USA

E-mail: raman@dartmouth.edu

**Abstract**

The Epithelial-to-Mesenchymal Transition (EMT) is a dynamic cellular program that is co-opted by cancer cells to enhance, among others, migratory and invasive cell traits. It is a key contributor to heterogeneity, chemo-resistance, and metastasis in many carcinoma-types, with the intermediate or hybrid EMT state playing a critical role due to its increased tumor-initiating potential. Using the SUM149PT human breast cancer cell line, we isolate multiple distinct single cell clones across the EMT spectrum possessing distinct migratory, tumor-initiating, and metastatic qualities, including three that reside in unique intermediate states. We interrogate the transcriptional profiles and chromatin landscapes of these distinct EMT states, identifying CBFβ as a key factor regulating the maintenance of, and metastatic ability in, the intermediate state. In order to detect these distinct EMT states and quantify the resulting epithelial-mesenchymal heterogeneity within tumors, we develop a novel multiplexed immunostaining approach using SUM149-derived orthotopic tumors as a training set and find EMT state and epithelial-mesenchymal heterogeneity to be predictive of overall survival in a cohort of stage III breast cancer samples. Our model reveals novel insights into the complex EMT spectrum of cell states, the networks that regulate them, and sheds light on how epithelial-mesenchymal plasticity contributes to tumor heterogeneity in breast cancer.

**Teaser**

We uncover the roles of phenotypic heterogeneity and plasticity in breast cancer and interrogate the networks that propagate them

**Introduction**

The Epithelial-to-Mesenchymal Transition (EMT) is a developmental cellular program frequently coopted by cancer cells[1] and is a key contributor to intra-tumoral heterogeneity[2–4], chemo-resistance and metastasis[5,6]. Rather than being a switch from an epithelial to a mesenchymal state, increasing evidence points to the existence of intermediate EMT states, wherein cells co-express both epithelial and mesenchymal traits[7–13]. These robust stable and metastable transition states possess unique characteristics[7,8,14,15] and contribute to the complex heterogeneity of tumors and their overall metastatic behavior[16,17]. While much work has been carried out identifying and characterizing EMT-inducing transcription factors[18–21], the transcriptional and epigenetic networks responsible for the stability and maintenance of the midpoints along the EMT spectrum are poorly defined. In addition, there are currently no approaches to identifying and quantifying intermediate EMT subpopulations within patient tumors to evaluate their prognostic significance. Using single-cell clonally isolated derivatives of the SUM149PT breast cancer cell line, we systematically interrogate how each EMT state independently contributes to heterogeneity and influences metastatic progression, uncovering the role of CBFβ in stabilizing and maintaining metastatic capability in certain intermediate states. We develop an entropy-based model to quantify phenotypic heterogeneity and EMT status and primary patient tumors using SUM149PT-derived tumors stained with a panel of six EMT markers as a training set. The cell states captured in the SUM149PT model are represented in a cohort of patient tumors and are predictive of overall survival in these patients, laying the foundation for quantification of epithelial-mesenchymal heterogeneity (EMH) and understanding the role of the intermediate EMT state in tumor progression.

## Results

### *Generation of a Model to study EMT*

*Single cell clones reside in multiple distinct EMT states*

We derived single cell clones from the SUM149PT $ER^-/PR^-$ inflammatory breast cancer cell line[22,23] stratified by expression of CD44 and CD104 (ITGβ4)[14] (Figure S1A). Six single-cell derived clonal populations were isolated, ranging from epithelial-like (E) to mesenchymal (M1 and M2), including three distinct intermediate states (EM1, EM2, and EM3) – hereafter referred to as "EMT clones" (Figure 1A, Figure S1A). Briefly, the SUM149PT cell line was sorted into three populations stratifying the EMT spectrum (Figure S1A) and single cells were sorted into 96-well plates from which fourteen single-cell derived clonal populations were chosen and isolated based on morphological characteristics. Of these fourteen, six were chosen to best represent the spectrum of states within the SUM149PT parental cell line. These clones, which stably retained their EMT states *in vitro*, were ranked along the EMT spectrum relative to one another based on expression of hallmark epithelial and mesenchymal markers such as *Vim* (vimentin*)*, *CDH1* (e-cadherin), *ZEB1*, and *SNAI1* (snail) (Figure 1C & D), as well as by variable migratory and invasive characteristics *in vitro* (Figure 1B). While the mesenchymal clones exhibited greater migratory and invasive ability than the epithelial clone, the intermediate clones displayed 2-10-fold higher migratory and invasive potential than the mesenchymal clones. These data suggest that earlier EMT studies did not discern intermediate states from mesenchymal ones when assessing these specific characteristics. Classic flow cytometry approaches employing 2-3 markers were insufficient to distinguish between the intermediate states (Figure S1D). By other measures, however, the intermediate clones exhibit differences in cell migration (Figure 1B), EMT marker expression (Figure 1C & D), cell morphology (Figure 1E) and expression of vimentin and e-cadherin (Figure 1F). Interestingly, the three intermediate clones most closely resemble the characteristics of the parental line in migratory and invasive ability *in vitro* (Figure 1B) and co-expression of e-cadherin and vimentin (Figure 1F), as well as their overall transcriptional profiles determined by unsupervised hierarchical clustering of Pearson's correlation coefficients following RNA-sequencing (Figure S1E). Notably, there were no substantial genetic differences between clones E, EM1, EM2, EM3 and M1, with M2 exhibiting some *SNP* and *INDEL* variations (Figure S1B & C), indicating that the phenotypic and functional differences between these intermediate clones are likely driven by non-genetic mechanisms[24].

Various methods have been developed to quantify the extent to which cells undergo an EMT and determine an absolute comparable EMT score[25–27], which have been reviewed recently[28]. When applied to our model, all three methods - MLR[27], KS[26] and 76GS[25] - predict that E through EM3, i.e. the epithelial and three intermediate clones, fall within the intermediate state, while M1 and M2 are mesenchymal (Figure 1G, Figure S1F & G). When plotted amongst the 59 breast cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE), these clones fall along the intermediate and more mesenchymal end of the spectrum (Figure 1H, Figure S1F&G, Supplemental Table 1). This spread of EMT states is most relevant in the context of studying metastasis as many highly epithelial breast cancer cell lines such as those seen in this comparison (Figure 1H)

exhibit less migratory and invasive characteristics[29], which contribute significantly to metastatic potential. Notably, the parental line scored closer to the epithelial (E) clone in most methods (Figure 1G, Figure S1F & G).

Heterogeneity in breast cancer cell lines and presence of intermediate states has been validated across many breast cancer cell lines and EMT models[8,30]. To further corroborate these findings in the SUM149PT model, we validated the presence of multiple EMT states in four canonical breast cancer cell lines, which range from intermediate to mesenchymal in their EMT scores – HCC1143, HCC38, Cal-51 and MDA-MB-231 (MLR EMT score: 0.26, 0.76, 1.06, and 1.93, respectively; Figure 1H). All four cell lines are comprised of heterogeneous subpopulations along the epithelial-mesenchymal spectrum to varying degrees, as revealed by their co-expression of e-cadherin, vimentin, and KRT8, matching those observed in the SUM149PT parental cell line (Figure 1I, Figure S1H). Our model, thus, highlights the presence of multiple distinct intermediate EMT states that are also found in other breast cancer cell lines, validating the suitability of this model further investigation of each state's role in tumor development and metastasis.

### ***Characterizing in vivo roles of multiple EMT states***

#### *Intermediate clones possess high tumor-initiating cell frequencies*

*In vivo* tumor initiation and growth further highlighted the individuality of these EMT clones. Upon orthotopic injection, the parental cell line was able to initiate and form tumors more rapidly than the other clones (Figure 2A). Tumor growth analysis with a bimodal linear mixed model[31] revealed that the three intermediate clones were able to initiate tumors at the same rate as the parental line (Initial phase Holm adj. p-val <0.05) (Figure 2A & C), but exhibited a lag in growth (Exponential phase Holm adj. p-val <0.003). The epithelial (E) and two mesenchymal (M1 & M2) clones both failed to initiate tumors as readily or, in the case of the mesenchymal clones, grow as rapidly as the parental and intermediate clones (Holm adj. p-value < 0.01). Increased tumor growth corresponded with decreased survival, with the parental line exhibiting shortest survival post-injection, and the mesenchymal clones the longest (Figure 2B). Limiting dilution analyses revealed a high tumor-initiating cell frequency in the parental and intermediate clones, with all mice forming tumors by eight weeks (Figure 2C). The two mesenchymal (M1& M2) possessed the lowest tumor initiating cell frequencies at eight weeks (0 and 1/75,975, respectively) despite expressing high levels of CD44 expression, a marker frequently correlated with increased stemness and tumor initiating ability[11,12,32] (Figure S1D & S2B). All clones generated high-grade, poorly differentiated, invasive ductal carcinomas of no special type (ductal). Clone M1 tumors notably exhibited 50% squamous differentiation, and clone M2 showed abundant spindle cell morphology (Figure S2A). Tumor growth and TIC frequencies indicate that, while the intermediate EMT clones may represent a population with high tumor-initiating potential, a heterogeneous population such as the parental line is able to enter an exponential growth phase more rapidly. Notably, as with the clonal cell lines, flow cytometry was not able to distinguish between tumors of different clonal origin despite significant differences in growth and survival (Figure S2B).

*EMT state impacts outgrowth of metastasis*

For the purposes of monitoring metastatic outgrowth *in vivo,* derivatives of each EMT cell line stably expressing a Luciferase-IRES-ZsGreen construct were generated, which allowed for both live and post-mortem detection of tumor cells. As has been observed previously for SUM149PT[33], the parental line and all clones metastasized to the lung, with varying success, as seen by fluorescence (Figure 2D). No other metastatic lesions could be detected by luciferase or fluorescence.

To precisely delineate the propensity of the clones to form micro- and macro-metastasis, lungs from animals bearing orthotopic tumors were fixed and stained with H&E and counted for micro- (<10 adjacent cells) and macro- (>10 adjacent cells) metastases (Figure 2E & F). The three intermediate (EM1-3) clones seeded higher numbers of micro-metastatic lesions per lung, compared to the most epithelial (E) and most mesenchymal (M2) (p-value <0.05). Within this group, clones EM1 and EM3 seeded higher numbers of macro-metastases compared to EM2 (Figure 2E), although differences between the intermediate clones were not statistically significant. Interestingly, while exhibiting the highest rate of tumor growth and poorest survival, the parental line seeded fewer lung metastases than the intermediate clones, suggesting that other mechanisms could be contributing to mortality.

***Identification of transcriptional networks that sustain EMT states***

*Transcriptomic profiles reveal shared intermediate gene signature*

Given the lack of genetic differences between the clones, we hypothesized that the clonal variations in EMT state were driven by alterations in their transcriptional profiles. Using bulk RNA-sequencing (RNA-Seq) of each clone, alterations in the expression levels of various transcription factors were analyzed using the epithelial clone E as a benchmark. Principal Component Analysis (PCA) demonstrates clustering of the parental line, intermediate EM and E clones (Figure 3A & B), whereas the two mesenchymal clones (M1 & M2) share no overlap between themselves or any other cluster. This indicates that the EMT clones do not reside in a linear spectrum, but rather embark upon multiple distinct trajectories. Unsupervised hierarchical clustering of the 500 most variable genes across all clones reveals distinct transcriptional programs separating the three intermediate clones from the epithelial (E) and mesenchymal (M1 & M2) ones (Figure 3D). Within this intermediate cluster, EM2 and E are again distinct from the remaining intermediates and parental line (Figure 3B & D). Differential expression analysis revealed 581 shared genes were significantly differentially expressed (adjusted p<0.05) in all clones when compared to E, with 1881 genes shared between the two mesenchymal clones and 178 shared between the three intermediate clones (Figure 3C, Figure S3A, Supplemental Table 2). Indeed, in comparison to the epithelial baseline clone E, more differentially expressed genes are exclusive to each clone than are shared between two or more clones (Figure 3C), further corroborating the unique EMT states represented by this model. Gene set enrichment analysis (GSEA)[34] of the differentially expressed genes confirms activation of the Hallmark EMT gene set in all intermediate and mesenchymal clones, as well as other gene sets corresponding to cell division and chromatin remodeling, as expected (Figure S3B).

To further explore the epigenetic landscape of these clones, we employed ATAC-seq[35] to determine the differential chromatin accessibility across EMT clones in comparison to the epithelial clone E, which exhibits the most closed chromatin profile (Figure 3E). Notably, the chromatin landscape is similarly diverse between mesenchymal and intermediate clones (Figure S3C & D), as seen in the RNA-sequencing (Figure 3D). To identify TFs with a significant enrichment of motifs among peaks that were uniquely accessible in each clone (relative to E), presence of transcription factor binding motifs was scanned for using Motifmatchr[36], and tested for enrichment against the background set of all identified peaks (hypergeometric test, p-adj<0.05) (Figure 3F). The three intermediate and early mesenchymal clones were highly enriched for motifs of the RUNX family transcription factors (Figure 3F). The TFAP2A/B/C TF family also exhibit enriched binding accessibility in these intermediate and early mesenchymal clones, albeit less significantly than the RUNX family. All three RUNX transcription factors as well as their co-factor CBFβ have been previously implicated in various cancers[37–39] as well as in metastatic progression of lung adenocarcinoma[40] and triple negative breast cancer (TNBC)[41,42]. Enrichment of these RUNX TF motifs is unique to the intermediate (EM1, EM2, and EM3) and early mesenchymal (M1) clones, those that seeded the highest number of lung metastases (Figure 2E).

*Combined RNA-seq and ATAC-seq identifies EMT and MET networks*

To provide a more comprehensive picture of the epigenetic and regulatory landscape of EMT, we utilized a new multiomics approach, DiffTF[43], to quantify TF activity and regulatory state by integrating RNA-seq and ATAC-seq data for each clone. Similar to Motifmatchr, TF activity was inferred by computing the fold change in chromatin accessibility between each clone at each binding site of a given TF, once again using clone E as a baseline reference. Inferred TF activity was visualized against RNA-seq log2 fold-change for each transcription factor (Figure 4A, Figure S4A), facilitating simultaneous detection of TFs with increased expression and TF activity (upper right quadrant). Analysis of the directional changes of TF expression from RNA-seq also indicated whether the TF in question acts as an activator (green) or repressor (red). The two mesenchymal clones revealed much larger shifts in both fold-change expression and TF activity, indicative of larger changes in their overall transcriptional and chromatin profiles relative to E. In the three intermediate and early mesenchymal clones, RUNX2 and RUNX3 are both amongst the most highly expressed and exhibit the highest TF activity of all significant transcription factors (Figure 4A, Figure S4A). A list of all significant transcription factors, their relative TF activity, and fold-change expression are included in Supplemental Table 3. Additionally, of the transcription factors that are more accessible in the intermediate state, expression of RUNX2 positively and TFAP2C negatively correlated with CCLE breast line EMT scores with high significance (Figure S4B & C).

Unsupervised clustering analysis presents a clearer picture of how these TFs act in a transcriptional network, highlighting TFs which exhibit significant changes in their activity in at least one clone (adj. p-value $<1E^{-15}$) (Figure 4B). Cluster 1 is comprised of transcription factors that are upregulated and function as activators in the clones that exhibit increased metastatic potential (EM1-3 and M1) – the RUNX TF family and co-factor CBFβ. Cluster 2 includes factors that increase in activity with the progression of EMT e.g., ETS1 and related factors, NFκB, RELB, and FLI1. Cluster 3 identifies transcription factors that are activated and remain

consistently active after entrance into the EMT e.g., members and regulators of the AP1 complex. Multiple members of the KLF family of transcription factors as well as CTCF are found in cluster 4, which exhibit a consistent decrease in TF activity following entrance into an EMT. Cluster 5 highlights transcription factors that may be uniquely active in M2, while remaining inactive in the other EMT clones, including FOXJ2 and FOXD3, CEBP complex members, and ATF4. Finally, Cluster 6 delineates the transcription factors that have an overall and graded decrease in activity as the EMT progresses. This cluster contains multiple EMT TFs such as Snail, Slug, and ZEB1 as well as known MET promoting TFs such as GRHL1 and 2. Although these and other canonical EMT transcription factors, such as Twist and PRRX1 are expressed at significantly high levels ($p<0.05$), non-significant or decreased TF activity indicates no change in chromatin accessibility compared to clone E in response to this increase in expression (Figure S4A, Supplemental Table 3). Indeed, these TFs decrease in activity while their overall expression remains high (Figure S4A) likely resulting from strong repression of chromatin accessibility at their epithelial target genes, potentially through HDAC recruitment[44,45] or other mechanisms at ZEB1 target promoters *CDH1, EPCAM, GRHL2* (Figure 4C*).*

*CBFβ knockdown destabilizes EMT states*

RUNX2 and its co-activator CBFβ have been implicated in mammary development[46], mammary stemness[47], and increased metastatic capacity in breast cancer[41,48]. To further examine the roles that these transcription factors play in EMT and the propagation of a metastatic state, we tested the effects of CRISPR-Cas9 mediated knockout of RUNX2 and CBFβ, which are uniquely expressed and active in the intermediate and early mesenchymal clones, as well as a non-targeting LacZ control (Figure 3F, Figure 4A & B). Knockout of RUNX2 in the intermediate clones did not result in any significant alterations to the levels of canonical EMT markers, except a minor reduction in vimentin expression in EM2 (Figure 4D). On the other hand, knockout of CBFβ led to a downregulation of Snail and vimentin expression in EM2 and EM3, as well as a more subtle reduction in fibronectin levels, indicating a shift to a more epithelial state. Interestingly, the expression of all three RUNX TFs was reduced upon loss of CBFβ, likely due to its role as an essential co-factor (Figure 4D).

To determine the effect of destabilizing EMT on tumor formation and metastasis, clone EM2 and EM3 bearing CBFβ knockout or LacZ nontargeting guide RNAs were injected orthotopically into NSG mice. These cell lines express high levels of both RUNX2 and CBFβ in comparison to clone E (Figure 4B), and form predominantly either micro- or macro-metastases, respectively, in the lung (Figure 2E). Interestingly, in EM2, and to a lesser extent EM3, CBFβ knockout accelerated the timepoint at which tumors were first observed (Initial phase Holm adj. p-val $< 0.001$) and time to tumor burden (Figure 4E). However, these tumors had a near complete absence of metastasis to the lung (Figure 4F). In contrast, RUNX2 knockout did not appear to have any significant effects on tumor formation or metastasis, likely due to compensation from the other RUNX transcription factors. Thus, we conclude that loss of CBFβ destabilizes the intermediate EMT state in clones EM2 and EM3, leading them to acquire a more epithelial state that, despite being more proliferative lacks lung-metastatic capacity.

***Exploration of tumor heterogeneity in EMT clone-derived tumors***

To better understand the role that epithelial-mesenchymal (E-M) heterogeneity plays on tumor progression and metastasis, we utilized a multi-round, multiplexed tyramide signal amplification (TSA) approach[49] to assess protein levels of EMT markers, segregating out stromal cells that could obfuscate EMT scoring. To capture the full spectrum of EMT states, we designed a panel of six EMT markers, containing three intermediate filament proteins (KRT8, KRT14 and VIM), two EMT transcription factors (ZEB1 and Snail), and an adherens junction protein that serves as a hallmark epithelial marker (E-cadherin). Tumors from each of the clones were stained and divided into 50 regions of interest (ROI) and processed with the inForm® analysis software (Akoya Biosciences) to generate composite images (Figure 5A). inForm® image training algorithms were used to discern tumor and stromal composition, as well as conduct cell segmentation (Figure S5A). Normalized percentile distributions of the arbitrary fluorescent units (AFU) of each EMT marker, per cell, across all images provided an overview of the overall composition of these tumors (Figure S5B), indicating that no one unique marker signature defined any individual tumor. Overall, clone M2 tumors expressed less e-cadherin and more ZEB1, while clone E tumors contain less vimentin, indicative of their initial EMT states *in-vitro*. To determine the extent of E-M heterogeneity within the stained images, a scoring metric was developed using the SUM149PT clones as a training set. Tumor images were scored based on a rubric of low (one major cell trait with up to one minor trait), mid (two major cell traits with up to three minor traits), and high (three or more major cell traits present with two or more minor traits) (Figure 5C). Scoring based on these criteria revealed that the intermediate EMT clones form tumors that contain more regions of higher heterogeneity than the parental cell line (Figure 5C). Thus, increasing levels of heterogeneity may not linearly correlate with tumor growth or metastasis, but rather, there exists an optimal ratio of cell traits within the tumor that determines its growth and metastatic potential. Indeed, the requirement for this optimal ratio may explain the growth lag observed in the intermediate clones when compared to the parental line (Figure 2A), which likely results from constraints in the generation of a requisite level of heterogeneity from a homogenous cell culture.

*Composition of EMT Phenotypes and E-M Heterogeneity varies by EMT clone*

In a parallel method to assess tumor diversity and phenotypic composition, the inForm® analysis software was trained on a subset of tumor images to recognize eight distinct cell phenotypes encompassing the majority cell states present in all tumor images (Figure S5A). These phenotypes, spanning from most epithelial to most mesenchymal are: E-cadherin only, KRT8 & E-cadherin, KRT8 & KRT14 (KRT), Triple positive (KRT8+e-cadherin+vimentin), KRT8/14 & vimentin, Snail only, vimentin only, and vimentin & ZEB1. Phenotypes were validated manually and by fluorescent marker distribution (Figure S5C). Similar to EMT marker distribution, there were no phenotypes unique to any individual state, indicating that optimal tumor progression may be determined by global ratios of cell states rather than the presence or absence of an individual cell type. On either end of the spectrum, clone E and clone M2 derived tumors are made up of over 75% epithelial (E-cad only, KRT8+E-cad, or KRT14 only), or mesenchymal (Snail only, vimentin only, or vimentin+ZEB1) phenotypes, respectively. Tumors derived from the intermediate clone and the parental line, meanwhile, all contained a roughly equal distributions of epithelial and mesenchymal phenotypes (Figure 5C).

These data further support the notion that the ratio of EMT states (phenotypes), rather than the presence of any particular state, is more reflective of tumor growth and metastatic potential.

In order to develop a scoring algorithm to assess E-M heterogeneity, three feature extraction methods were tested using the segmented cell data files (Figure S5A). These three approaches sought to determine the best method to assess heterogeneity: (1) an entropy-based approach[50,51] using mean marker expression of cells per image, (2) a nearest neighbor analysis approach[52] using cell phenotypes as determined above (Figure 4C), and (3) a hybrid approach combining approaches 1 and 2 (Figure S5D). All three approaches were developed and evaluated using a randomized subset of clone tumor images and their corresponding heterogeneity scores according to the rubric outlined in Figure 5B as ground truths. Logistic regression performance on unseen test data indicated a 78% accuracy for entropy-based features of mean marker cell expressions. Five-fold cross-validation determined that the entropy-based approach 1 (F1-score= 0.78, Wilcoxon ranking test p-value=0.004) proved to be the most robust at correctly assessing tumor heterogeneity (Figure S5D). This shows that E and M2 clones consisted largely of areas of low or mid-level heterogeneity, whereas all intermediate clones were comprised of regions of high heterogeneity (Figure 5B).

In addition to scoring E-M heterogeneity, we sought to assign EMT scores to tumors based on the ratios of epithelial and mesenchymal traits they exhibit. EMT scores were generated by calculating the composition of cell phenotypes per image (50 images per clone) using the linearly weighted average of cell ratios expressing each phenotype from epithelial to mesenchymal. EMT scores range from 0 to 1, with zero being comprised of all epithelial phenotypes and one comprised of all mesenchymal phenotypes. Rather than attaining an equilibrated EMT state, clonal tumors held true to the EMT state of their starting populations, with clone E tumors being predominantly epithelial (mean=0.25), intermediate (EM1, EM2, EM3) and parental tumors maintaining an intermediate EMT score (mean=0.4-0.6), and the most mesenchymal (M2) scoring >0.7 (mean=0.7) (Figure 5D). Clone M1 tumors also scored as intermediate despite starting as a quasi-mesenchymal, further validating this clone as a late intermediate (Figure 5D). Interestingly, the variation in EMT scores between images in the intermediate clones (EM1, EM2, EM3, & M1) and parental line was highest of all of the groups, indicating higher intra-tumoral heterogeneity amongst tumor regions (Figure 5D). In exploring the connection between heterogeneity and EMT scores, we found a correlation between the two scores upon splitting the EMT score into tertiles (Epithelial = 0-0.29, Intermediate = 0.3-0.69, Mesenchymal = 0.7-1; chi-squared p-value = 1.1E-11). Low heterogeneity tumor regions correlated with more mesenchymal EMT scores, as can be seen in M2 tumors, and epithelial EMT scores correlate with mostly mid heterogeneity, seen in E tumors (Figure 5A, Figure S5E). Mid and high heterogeneity regions were characterized by intermediate EMT scores despite encompassing a more diverse array of possible EMT states (Figure 5E, Figure S5E).

*Knockout of CBF$\beta$ in intermediate SUM149 clones reduces E-M Heterogeneity and EMT score*

Upon developing a metric to calculate epithelial-mesenchymal heterogeneity and EMT score, we questioned whether the decrease in metastasis observed upon loss of CBFβ (Figure 4E & F) resulted from alterations to overall EMT state, or from a change in E-M heterogeneity. Indeed, when tumors generated from

EM2 and EM3 CBFβ knockout clones were stained with the multiplexed immunofluorescence panel, loss of CBFβ led to the formation of tumors with overall lower EMT scores (Fisher's p-val = $9.2E^{-11}$). Clone EM2 additionally had decreased E-M heterogeneity (Fisher's p-val = 0.05) (Figure 5F, Figure S5F & G). This reduction arises from an increased presence of more epithelial phenotypes such as e-cadherin only and KRT8 & e-cadherin co-expressing cells, which are associated with a more proliferative state[53,54], and concurrent decrease in mesenchymal phenotypes such as vimentin, which are more invasive (Figure S5H).

*Lung metastases contain predominately late intermediate and mesenchymal phenotypes*

To further understand the roles of tumor cell E-M heterogeneity and EMT score on the metastatic cascade, we resected the primary tumors to allow for maximal metastatic outgrowth beyond the primary tumor burden. Briefly, intermediate and parental cell lines were orthotopically injected and allowed to grow to a volume of $1cm^3$ before surgical resection. Lungs were harvested and metastases quantified at 2.5 months post-resection, determined by burden of the relapsed ipsilateral tumor. Micro- and macro-metastasis trends were maintained in this later growth model, with EM1 forming a mixture of micro and macro metastases, EM2 forming predominantly micro-, and EM3 predominantly macro-metastases.

We subsequently used the multiplexed immunofluorescence approach to stain these lung metastases to determine their heterogeneity and EMT phenotypes (Figure 5G). To our surprise, all metastases, regardless of size or clone were predominantly mesenchymal in composition, expressing very little e-cadherin or KRT8 (Figure 5G & H). EMT and heterogeneity scores calculated from this staining revealed overall low heterogeneity and high EMT scores, features that were infrequently exhibited in the primary tumors (Figure 5I). These results suggest that, in contrast to trends observed in other models[55], the SUM149PT tumors give rise to metastases that are predominantly mesenchymal in nature.

*High Heterogeneity Score and Intermediate EMT states are predictive of poor survival outcomes in patients*

To assess whether the EMT phenotypes observed in our model were representative of those observed in human breast cancer specimens, we carried out TSA immunostaining of a tumor microarray of the CDP Breast Cancer Stage III Prognostic Tissue Microarray collected and indexed by the Cooperative Human Tissue Network (CHTN) (Figure 6A). Following staining with our 6-marker EMT panel, we recovered and analyze 124 cores from of both hormone positive (n=86) and hormone negative (n=38) patients with long term survival follow-up spanning 250 months. All eight phenotypes were reproducibly represented across all patient tumors (Figure 6A, Figure S6A) and ratios of epithelial and mesenchymal phenotypes follow similar trends to the SUM149PT clonal model (Figure S6B). These data serve to validate the SUM149PT clones as a model to study various states along the EMT spectrum.

Cores were phenotyped and analyzed as previously described to generate a heterogeneity score and EMT score, again segregated into tertiles, for each patient (Figure 6B, Figure S6C). Overall, high heterogeneity is associated with a more mesenchymal EMT score, while low heterogeneity is associated with an epithelial EMT score (Fisher's exact test p-value = $4.4xE^{-5}$, Figure 6B & C, Figure S6C). While this differs from trends seen in the SUM149PT model (Figure 5E, Figure S5E), the samples in this cohort better represent the

population distribution of stage III breast tumors. In survival models adjusting for patient age and hormone receptor status with all samples that passed QC (n=124), patients with a high heterogeneity score had significantly worse overall survival (HR: 5.2, 95% CI: 2.6-10.6, $P=5.2E^{-6}$), as did patients with intermediate EMT scores (HR:1.7, 95% CI:1.0-3.0, p=0.05) (Figure 6D & E). While high heterogeneity score on its own remained predictive of poor patient outcome, EMT score alone was not associated with survival (Figure S6D, HR(high):3.9, 95% CI: 2.00-7.61, $p=6.04E^{-5}$). As these metrics describe associated yet independent metrics of tumor complexity and EMT status, our results further demonstrate nuances of tumor heterogeneity in disease progression. When sub setting for hormone negative disease (n=36), which is known to exhibit increased intra-tumoral heterogeneity[2], high and low heterogeneity were both associated with significantly worse overall survival (HR (high): 16.7, 95% CI:4.2-66.6, $p=6.5E^{-5}$; HR (low): 7.9, 95% CI:2.3-27.1, p=0.001) (Figure 6F, Figure S6E), as was an intermediate EMT score (HR: 3.43, 95% CI: 1.04-11.3, p=0.042). This is likely due to a subset of cores in HR- patients that score as low heterogeneity/mesenchymal EMT (Figure S6F), such as those seen in our HR negative SUM149 clonal model (Figure S5E). Hormone positive disease showed similar trends to the overall model (Figure S6G). Overall, these analyses indicate an increased risk of death in patients with highly heterogeneous tumors that are comprised of intermediate EMT phenotypes.

**Discussion**

Our study encompasses a comprehensive analysis of the spectrum of EMT states represented within a breast cancer cell line to interrogate the nuances of each state, their respective contributions to tumor initiation and metastasis, and their epigenetic regulatory networks. We uncover the presence of multiple unique EMT states within the intermediate EMT category, as well as two distinct mesenchymal-like states, suggesting multiple, non-linear trajectories for EMT as has been previously shown[30,56]. These states were verified by flow cytometry (CD44 and CD104) and ranked by EMT scores among other CCLE breast cancer cell lines, where they fell between an early intermediate and mesenchymal. While this model does not span from one extreme state to the other, it represents the EMT spectrum within SUM149PT, a cell line that is classified the BL2 molecular subtype of TNBC[57], reflecting the heterogeneous nature of this subtype. We used the relative EMT states between the clones to study the epigenetic heterogeneity of metastatic breast cancer within an isogenic background and interrogate the fitness of each individual state. While the three intermediate EMT clones possess the highest migratory and invasive potential *in vitro*, they are outperformed in tumor-initiating ability and growth by the parental cell line, which, by RNA-seq clustering most closely resembles the epithelial clone.

We further confirm that the intermediate EMT state exhibits higher levels of cellular plasticity, manifesting in tumors that grow more quickly than solely epithelial or mesenchymal cell states, and exhibit high levels of E-M heterogeneity (EMH). Moreover, this plasticity-induced heterogeneity plays a key role in the metastatic propensity and tumor-initiating potential of these clones. However, no one individual EMT state is capable of recapitulating the aggressive growth and decrease in survival of the parental line. The intermediate clones, upon xenotransplantation, experience a growth lag followed by an overdiversification, resulting in high EMH-scored tumors compared to the parental line, indicating that their high levels of plasticity enable them to attain higher levels of heterogeneity that propel tumor growth and metastasis. In contrast, the extreme clones

exhibit less plasticity, taking longer periods of time to generate tumors that are less heterogeneous with weaker metastatic potential. The intermediate population in isolation appears to be the most potent tumor initiator, however, the presence of multiple states (i.e., heterogeneity) within the parental line imparts additional fitness that provides robust and exponential tumor growth. This suggests that the presence of heterogeneous subpopulations within a tumor confer a greater tumor growth advantage than the presence of more homogeneous subpopulations that exhibit higher levels of plasticity. A corollary assessment would be that a tumor benefits from harboring heterogeneous subpopulations, only a small subpopulation of which are required to exhibit high levels of plasticity. Future studies investigating the dynamics of expansion of this intra-tumoral heterogeneity may elucidate how EMT phenotypes work cooperatively to support tumor growth and progression to metastasis.

Through a multi-omics approach, we identify distinct transcriptional programs across the EMT spectrum. The intermediate state was found to be maintained and stabilized by a subset of transcription factors, including the RUNX family. Indeed, knockout of the co-activator of this RUNX family of transcription factors, CBFβ, results in decreased expression of all RUNX TFs in the intermediate EMT clones that leads to tumors that metastasize at lower rates as a result of reduced heterogeneity and increased presence of epithelial cells within the tumor. These observations are in line with other studies that have outlined a role for RUNX2 in metastasis[40], but provide additional granularity to this work by identifying specific cell states that benefit from the presence of active RUNX- CBFβ. This study also underscores the importance of combining the study of the transcriptional and chromatin state of cells as a means of uncovering their underlying regulatory networks, which particularly enabled the delineation between similar EMT states; the transcriptional analysis of these clones alone was not sensitive enough to identify differences in many of the canonical EMT transcription factors, let alone other less variable TFs. This is likely because TF gene expression is tightly regulated by multiple factors[58] and differences in TF activity would benefit from a higher resolution study at the chromatin level.

A concern when studying heterogeneous cell subpopulations is the retention of their initial EMT state following their isolation and culture. All six of our clones retained their morphological and phenotypic traits through multiple passages in culture, although we did observe clone E drifting to acquire a more spindle-shaped morphology upon extended periods of culture. Interestingly, all single cell clones retained their initial CD44/CD104 expression profiles except for the most mesenchymal clone, M2, which gained the expression of CD104 following its culture as an isolated single cell clone. Given that the maintenance of EMT state is regulated by a complex set of paracrine and autocrine signals[59], it is plausible, albeit speculative, that the loss of specific paracrine signals upon isolated culture could have resulted in an altered state in the M2 clone that allowed expression of CD104. Nevertheless, to ensure that all clones retain their original EMT states upon culture, their passage numbers were restricted to below 20.

Finally, we develop an approach to quantifying E-M heterogeneity (EMH) and “EMTness” within human tumors, the former showing promise in its ability to inform disease prognosis. Previous work has sought to elucidate phenotypic intra-tumoral heterogeneity in a manner of different contexts and analyses using an array of multiplexed staining methods[60,61] as well as through single cell approaches[62]. Here, we sought to specifically

delineate intra-tumoral heterogeneity (ITH) across the EMT spectrum, which has strong prognostic value for predicting invasion and metastasis in our model. A significant challenge in the quantification of EMH has been the ability to discern carcinoma cells that exhibit mesenchymal traits from stromal cell-types such as fibroblasts that express similar markers. Previous approaches to quantify EMH have considered morphological features[63] and analyzed gene expression profiles from publicly available datasets that identify cells that have undergone EMT[26,27]. These approaches, while being highly useful to study tumor cell EMT status, have been unable to segregate stromal infiltrates and their contributions to aggregate EMT scoring. Our multiplexed immunostaining approach, which employs a set of six EMT markers to assess EMH and EMT score, incorporates a segmentation step which ensures that the quantification excludes stromal elements. We were able to distinguish eight phenotypes within these EMT clone-derived tumors with high reproducibility, which were all present in a cohort of patient tumors, validating our approach to quantifying EMT in patient samples. Importantly, we found that patients with high heterogeneity and overall intermediate EMT state had significantly worse overall survival than any other group, independent of patient age and clinical subtype. Assessing these heterogeneity and EMT parameters within a patient tumor could, thus, hold value in predicting prognosis, enabling the clinical assessment of EMH as a tumor prognostic indicator and inform therapeutic decision-making.

## Materials and Methods

### Cell culture

The human derived SUM149PT cell line was obtained from the Weinberg lab, who in turn acquired it from Dr. Stephen P. Ethier (Michigan). All derivative cell lines were maintained according to ATCC. 293T cell line was maintained as per ATCC recommendations.

### Lentiviral Vectors

Lentivirus was made with 293T cells plated at 60% confluency in 10cm tissue culture treated plates and transfected using X-tremeGENE HP DNA Transfection Reagent with lentiviral vectors, psPAX2 (1.5ug), pCMV-VSV-G (1.5ug) and pcDNA3-eGFP (0.5ug) plus the lentiviral vector of interest (3ug). Supernatant collected at 48 and 72 hours post-transfection, concentrated using Lenti-X concentrator. See SI for more details

### Cell line Generation

To generate single cell clones, parental cell line SUM149PT was stained with CD44 -PeCy7 and CD104-APC. Florescence assisted cell sorting (FACS) was used to isolate $CD44^{lo}CD104^{hi}$ (epithelial), $CD44^{hi}CD104^{lo}$ (intermediate), and $CD44^{hi}CD104^{hi}$ (mesenchymal) cell populations into collection tubes and immediately plated at 0.5 cells per well into a 96-well plate to isolate single cell clonal populations. Single cell clones were then expanded and assessed for EMT characteristics. See SI for more details

*ZsGreen expressing cells*

All SUM149PT clones and parental line were infected with high titer pHIV-Luc-ZsGreen virus so as to generate ZsGreen and Luciferase expressing tumor cells for metastasis tracking in mouse. ZsGreen positive populations were isolated with fluorescence assisted cell sorting. See SI for more details

### RT-qPCR

RNA was harvested from 6-well plates of cells at confluency, extracted using the Qiagen RNeasy plus kit (Qiagen 74034) and quantified using a NanoDrop (Thermo Fisher Scientific - ND-2000-US-CAN). Reverse transcriptase PCR (AppliedBiosystems #4368814) was performed to generate cDNA and Power SYBR green PCR mastermix (Applied Biosystems) was used for qPCR. See Supplemental Information for RT primers.

**Western Blot**
Lysates were collected in 1x RIPA buffer with protease and phosphatase inhibitors. Quantified lysates were loaded at equal concentrations and stained for a variety of EMT markers. Please see SI for details and antibody references.

**Transwell Assay**
Transwell Assays were conducted using Costar transwell plates. Briefly, cells were plated at $2.5E^5$ above the membrane in 10% complete medium, with 100% complete medium beneath the transwell membrane. For invasion assays, transwells were coated with 5% Matrigel prior to seeding. After 16-18hrs at 37ºC, the bottom of the transwell was permeabilized with methanol and stained with 0.5% crystal violet to visualize and count migrated cells on the bottom of the membrane. See SI for details.

**shRNA knockdown**
RUNX2 knockdown stable cell lines were created for all clones after infection with virus carrying tet-pLKO-puro with either one of two RUNX2 hairpins or LacZ negative control.
Cells were plates at $2x10^5$ cells per well in 6-well plates and treated with doxycycline (2ug/mL) for 72 hours before on-plate lysis for RNA of protein collection. See SI for details

**Cell Imaging**
*Bright Field*
Images were taken on a Nikon Eclipse TS100 under 20x magnification to determine cell morphology
*ImmunoFluorescence*
Cells were grown to 60% confluency in chamber slides (Falcon 354118) with standard media. Cells were fixed and permeabilized prior to staining with primary antibody (Vimentin CST # 1:100 and E-cadherin BD number 1:100) overnight followed by secondary (anti-rabbit ThermoFisher #31466 1:10000, anti-mouse ThermoFisher #31431 1:10000) for one hour. Slides were washed and mounted using ProLong Diamond (Invitrogen P36961) before imaging on Zeiss LSM 800 with Airyscan (63X)

***In Vivo* Studies**
Cell lines were resuspended in 30% Matrigel (VWR 47743-706) and injected in limiting dilutions (250k, 25k, and 2.5k cells per flank) orthotopically into the inguinal mammary fat pat (no. 4) of NOD-scid IL2R$\gamma^{null}$ (Stock No: 005557; Jackson Laboratory). Tumor growth was monitored weekly and tumor volume was measured along 3 axes with calipers (VWR 62379-531). Tumors and lungs were harvested at time of tumor burden (total tumor volume of $2cm^3$) and fixed overnight with 10% neutral buffered formalin. Tumor growth curves and survival statistically analyzed using TumGrowth[31]. Tumor initiating potential calculated with Extreme Limiting Dilution Analysis (ELDA)[64].

**Tumor and Lung staining**
Tumors and lungs were extracted at time of tumor burden (total tumor volume of $2cm^3$) and fixed overnight with 10% neutral buffered formalin. All samples were then processed and stained for Hematoxylin & Eosin (H&E) by the Dartmouth Hitchcock Pathology Shared Resource. Lungs bearing ZsGreen positive EMT clones were counted by eye on Nikon Eclipse TS100 and select images taken at 4x magnification.

**Metastasis counting**
H&E stained lung slides were scanned on a PerkinElmer Vectra3 slide scanner at 10x and counted by eye for micro (> 10 adjacent cells) and macro (10+ adjacent cells) metastatic tumors.

**Whole Exome Sequencing**
Whole Exome sequencing and subsequent SNP and INDEL alignment and discovery was performed by BGISeq on all EMT clones. SNPs and InDels for each clone were tested against the consensus set for clone E to determine possible genetic mutations with a Fishers Exact Test and plotted as Odds Ratios. See SI for further details.

**RNA-seq data processing and analysis**

RNA was collected as previously described and processed by the Dartmouth Bioinformatics core facility. Reads were aligned to human genome hg38 using STAR[65] (v 2.7.2b). Differential expression analysis was done using DESeq2[66], with clone E as the reference group in all comparisons. Genes with a Benjamini-Hochberg adjusted P-value < 0.05 (Wald-test) were considered statistically significant. Further details on RNA-processing and analysis can be found in the SI.
Gene Set Enrichment Analysis (GSEA) was performed on differential gene lists for each clone relative to clone E using the clusterProfiler[32] to determine overlaps with the Hallmark gene sets at a threshold of 0.05.

**ATAC-seq data processing**
Tagmented DNA and library prep for ATAC-sequencing was performed according to the protocol detailed in Buenrostro et. al. 2015[35]. ATAC peaks were annotated using ChIPseeker[67] package, R-package TxDb.Hsapiens.UCSC.hg38.knownGene was used to define gene models and coordinates of genomic features, and differential accessibility analysis of the consensus peak was also performed using the DESeq2 R-package, with clone E used as the reference group in all comparisons. Peaks with a Benjamini-Hochberg corrected P-value < 0.05 (Wald-test) were considered statistically significant. Unsupervised hierarchical clustering was performed with the ComplexHeatmap R-package, also using the 3000 most variable consensus peaks.
Differential TF activity between single-cell derived clones, as well as TF mode of action (i.e. activator, repressor) was estimated using diffTF[43]. This was used in conjunction with matched RNA-seq (classification mode) to classify each TF into one of the following modes of action (activator, repressor, not-expressed, undetermined) through correlation of TF expression levels with target site accessibility. For a full description of analysis and parameters, please see SI.

**Multiplexed TSA staining**
Tumors were selected from each EMT clone at approximately 1cm$^3$ and stained with (in order) Snail (CST #3895, 1:400), KRT8 (Invitrogen PA5-29607, 1:300), KRT14 (Invitrogen MA5-11599, 1:1000), Vimentin (CST #5741, 1:500), E-cadherin (BD #610182, 1:500), and ZEB1 (Invitrogen PA5-82982, 1:1000). Antibody optimization and multiplexed staining was done according to PerkinElmer OPAL Assay Development Guide (August 2017) and previous literature[49,68]. See SI for details.

*Image Analysis*
Whole slide scans were captured at 10x with the Akoya Biosciences Vectra®3 Slide Scanner and ~50 regions of interest (ROIs) per tumor chosen manually with Phenochart™ (Akoya Biosciences). ROIs were imaged at 20x resolution and imported into inForm® analysis software (Akoya Biosciences). Spectral unmixing single-stains and background fluorescence slides were generated from the Parental tumor according to the OPAL Assay Development Guide. ROIs were spectrally unmixed and assigned colors and exported as composite images (Figure 4A). Tissue (Trainable to 98% accuracy) and cell segmentation were performed (Nuclear compartment – DAPI, Cytoplasm – vimentin & KRT8, Membrane – e-cadherin) and cells were phenotyped based on expression of one or multiple markers (e-cadherin only, KRT8 & e-cadherin, KRT14 only, Triple positive (KRT8+e-cadherin+ vimentin), snail only, vimentin only, and vimentin+ZEB1) and validated by marker distribution (Figure S5A&B). Entire Cell Mean Fluorescent units extracted for each marker and normalized as a percentile of maximum and minimum fluorescence across all cells in all images.

*Heterogeneity and EMT Scores*
<u>Approach 1:</u> Heterogeneity scores were generated using penalized logistic regression based on entropies of mean marker cell expressions to identify markers and cellular compartments (Nucleus, Cytoplasm, and Membrane) that contributed most to the variability in the ranked tumor images (Figure S5C). In total 134 entropy-based features extracted, and 13 of them were selected by Recursive Feature Elimination (RFECV)[69] as most relevant. Logistic regression classified sample heterogeneity into levels mid, low and high. Ground truths were determined from the rubric: low (one major cell trait with up to one minor trait), mid (two major cell traits with up to three minor traits), and high (three or more major cell traits present with two or more minor traits). These were used to train and validate the algorithm using 70% training and 30% test images (n=409) in a five-fold cross validation.
<u>Approach 2:</u> Nearest Neighbor analysis was conducted with the scikit-learn python package[52] using cell phenotypes determined from inForm®. Similar feature selection methods were applied to nearest neighbors, with 26 out of 49 features selected.

Approach 3: A hybrid approach used combined the 134 and 49 features from approach 1 and 2 and selected 18 out of 183 features.
To generate the EMT score, the seven derived phenotypes were weighted from epithelial to mesenchymal (e-cadherin only -4, KRT8 & e-cadherin -3, KRT8/14 only -2, Triple positive -1, KRT + vimentin +1, snail only +2, vimentin only +3, and vimentin & zeb1 +4) and applied to a multivariate logistic regression.

**Raw Data**

Raw data from RNA-seq and ATAC-seq experiments that support the findings in this study have been deposited and can be accessed in GEO (accession number GSE172613)

**Research Animals**

All animal experiments were carried out under ethical regulations approved by the Dartmouth College IACUC

**References**


1. Thiery, J. P. Epithelial–mesenchymal transitions in tumour progression. *Nat. Rev. Cancer* **2**, 442–454 (2002).
2. Easwaran, H., Tsai, H. C. & Baylin, S. B. Cancer Epigenetics: Tumor Heterogeneity, Plasticity of Stem-like States, and Drug Resistance. *Molecular Cell* vol. 54 716–727 (2014).
3. Pattabiraman, D. R. & Weinberg, R. A. Tackling the cancer stem cells-what challenges do they pose? *Nat. Rev. Drug Discov.* **13**, 497–512 (2014).
4. Ognjenovic, N. B. *et al.* Limiting Self-Renewal of the Basal Compartment by PKA Activation Induces Differentiation and Alters the Evolution of Mammary Tumors. *Dev. Cell* **5**, (2020).
5. Shibue, T. & Weinberg, R. A. EMT, CSCs, and drug resistance: The mechanistic link and clinical implications. *Nat. Rev. Clin. Oncol.* **14**, 611–629 (2017).
6. Dongre, A. & Weinberg, R. A. New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. *Nature Reviews Molecular Cell Biology* vol. 20 69–84 (2019).
7. Huang, R. Y.-J. *et al.* An EMT spectrum defines an anoikis-resistant and spheroidogenic intermediate mesenchymal state that is sensitive to e-cadherin restoration by a src-kinase inhibitor, saracatinib (AZD0530). *Cell Death Dis.* **4**, e915–e915 (2013).
8. Pastushenko, I. *et al.* Identification of the tumour transition states occurring during EMT. *Nature* 1 (2018) doi:10.1038/s41586-018-0040-3.
9. Jolly, M. K. *et al.* Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Front. Oncol.* **5**, 155 (2015).
10. Hendrix, M. J. C., Seftor, E. A., Seftor, R. E. B. & Trevor, K. T. Experimental co-expression of vimentin and keratin intermediate filaments in human breast cancer cells results in phenotypic interconversion and increased invasive behavior. *Am. J. Pathol.* **150**, 483–495 (1997).
11. Mani, S. A. *et al.* The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. *Cell* **133**, 704–715 (2008).
12. Morel, A.-P. *et al.* Generation of Breast Cancer Stem Cells through Epithelial-Mesenchymal Transition. *PLoS One* **3**, e2888 (2008).
13. Grosse-Wilde, A. *et al.* Stemness of the hybrid Epithelial/Mesenchymal State in Breast Cancer and Its Association with Poor Survival. *PLoS One* **10**, e0126522 (2015).
14. Bierie, B. *et al.* Integrin-β4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E2337–E2346 (2017).
15. Jolly, M. K. *et al.* Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacol. Ther.* **194**, 161–184 (2019).
16. Kröger, C. *et al.* Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proc. Natl. Acad. Sci. U. S. A.* 201812876 (2019) doi:10.1073/pnas.1812876116.
17. Lüönd, F. *et al.* Distinct contributions of partial and full EMT to breast cancer malignancy. *Dev. Cell* **56**, 3203-3221.e11 (2021).
18. Krebs, A. M. *et al.* The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic cancer. *Nat. Cell Biol.* **19**, 518–529 (2017).
19. Grande, M. T. *et al.* Snail1-induced partial epithelial-to-mesenchymal transition drives renal fibrosis in mice and can be targeted to reverse established disease. *Nat. Med.* **21**, 989–997 (2015).
20. Guo, W. *et al.* Slug and Sox9 cooperatively determine the mammary stem cell state. *Cell* **148**, 1015–1028 (2012).
21. Li, Q.-Q. *et al.* Twist1-mediated adriamycin-induced epithelial-mesenchymal transition relates to multidrug resistance and invasive potential in breast cancer cells. *Clin. Cancer Res.* **15**, 2657–65 (2009).
22. Forozan, F. *et al.* Molecular cytogenetic analysis of 11 new breast cancer cell lines. *Br. J. Cancer* **81**, 1328–1334 (1999).
23. M Robertson, F. & Chu, K. Genomic Profiling of Pre-Clinical Models of Inflammatory Breast Cancer Identifies a

Signature of Epithelial Plasticity and Suppression of TGFβ Signaling. *J. Clin. Exp. Pathol.* **02**, (2012).
24. Jolly, M. K. & Celià-Terrassa, T. Dynamics of Phenotypic Heterogeneity Associated with EMT and Stemness during Cancer Progression. *J. Clin. Med.* **8**, 1542 (2019).
25. Byers, L. A. *et al.* An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clin. Cancer Res.* **19**, 279–290 (2013).
26. Tan, T. Z. *et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.* **6**, 1279–1293 (2014).
27. George, J. T., Jolly, M. K., Xu, S., Somarelli, J. A. & Levine, H. Survival Outcomes in Cancer Patients Predicted by a Partial EMT Gene Expression Scoring Metric. *Cancer Res.* **77**, 6415–6428 (2017).
28. Chakraborty, P., George, J. T., Tripathi, S., Levine, H. & Jolly, M. K. Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum. *Front. Bioeng. Biotechnol.* **8**, 1–13 (2020).
29. Blick, T. *et al.* Epithelial mesenchymal transition traits in human breast cancer cell lines. *Clinical and Experimental Metastasis* vol. 25 629–642 (2008).
30. Cook, D. P. & Vanderhyden, B. C. Context specificity of the EMT transcriptional response. *Nat. Commun.* **11**, 2142 (2020).
31. Enot, D. P., Vacchelli, E., Jacquelot, N., Zitvogel, L. & Kroemer, G. TumGrowth: An open-access web tool for the statistical analysis of tumor growth curves. *Oncoimmunology* **7**, (2018).
32. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3983–3988 (2003).
33. El-Gazzar, A. *et al.* Effects on tumor development and metastatic dissemination by the NKG2D lymphocyte receptor expressed on cancer cells. *Oncogene* **33**, 4932–4940 (2014).
34. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **16**, 284–287 (2012).
35. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **2015**, 21.29.1-21.29.9 (2015).
36. Schep, A. motifmatchr: Fast Motif Matching in R. R package version 1.12.0. (2020).
37. Ito, Y., Bae, S.-C., Shyue, L. & Chuang, H. The RUNX family: developmental regulators in cancer. *Nat. Publ. Gr.* **15**, 81 (2015).
38. Otálora-Otálora, B. A., Henríquez, B., López-Kleine, L. & Rojas, A. RUNX family: Oncogenes or tumor suppressors (Review). *Oncology Reports* vol. 42 3–19 (2019).
39. Blyth, K., Cameron, E. R. & Neil, J. C. The RUNX genes: Gain or loss of function in cancer. *Nature Reviews Cancer* vol. 5 376–387 (2005).
40. Lafave, L. M. *et al.* Epigenomic State Transitions Characterize Tumor Progression in Mouse Lung Adenocarcinoma. *Cancer Cell* **38**, 212–228 (2020).
41. Ran, R. *et al.* A role for CBFβ in maintaining the metastatic phenotype of breast cancer cells. *Oncogene* 1–14 (2020) doi:10.1038/s41388-020-1170-2.
42. Nagaraja, G. M. *et al.* Gene expression signatures and biomarkers of noninvasive and invasive breast cancer cells: Comprehensive profiles by representational difference analysis, microarrays and proteomics. *Oncogene* **25**, 2328–2338 (2006).
43. Berest, I. *et al.* Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF. *Cell Rep.* **29**, 3147-3159.e12 (2019).
44. Peinado, H., Ballestar, E., Esteller, M. & Cano, A. Snail Mediates E-Cadherin Repression by the Recruitment of the Sin3A/Histone Deacetylase 1 (HDAC1)/HDAC2 Complex. *Mol. Cell. Biol.* **24**, 306–319 (2004).
45. Aghdassi, A. *et al.* Recruitment of histone deacetylases HDAC1 and HDAC2 by the transcriptional repressor ZEB1 downregulates E-cadherin expression in pancreatic cancer. *Gut* **61**, 439–448 (2012).
46. McDonald, L. *et al.* RUNX2 correlates with subtype-specific breast cancer in a human tissue microarray, and ectopic expression of Runx2 perturbs differentiation in the mouse mammary gland. *DMM Dis. Model. Mech.* **7**, 525–534 (2014).
47. Fritz, A. J. *et al.* RUNX1 and RUNX2 transcription factors function in opposing roles to regulate breast cancer stem cells. *J. Cell. Physiol.* **235**, 7261–7272 (2020).
48. Chimge, N. O. *et al.* Regulation of breast cancer metastasis by Runx2 and estrogen signaling: The role of SNAI2. *Breast Cancer Res.* **13**, 1–13 (2011).
49. Roy, S., Axelrod, H. D., Valkenburg, K. C., Amend, S. & Pienta, K. J. Optimization of prostate cancer cell detection using multiplex tyramide signal amplification. *J. Cell. Biochem.* **120**, 4804–4812 (2019).
50. Liu, B. *et al.* An entropy-based metric for assessing the purity of single cell populations. *Nat. Commun.* **11**, 1–13 (2020).
51. Guo, M., Bao, E. L., Wagner, M., Whitsett, J. A. & Xu, Y. SLICE: Determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res.* **45**, (2017).
52. Pedregosa FABIANPEDREGOSA, F. *et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot*. *Journal of Machine Learning Research* vol. 12 http://scikit-learn.sourceforge.net. (2011).

53. Vega, S. *et al.* Snail blocks the cell cycle and confers resistance to cell death. *Genes Dev.* **18**, 1131–1143 (2004).
54. Mejlvang, J. *et al.* Direct repression of cyclin D1 by SIP1 attenuates cell cycle progression in cells undergoing an epithelial mesenchymal transition. *Mol. Biol. Cell* **18**, 4615–4624 (2007).
55. Tsai, J. H., Donaher, J. L., Murphy, D. A., Chau, S. & Yang, J. Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. *Cancer Cell* **22**, 725–36 (2012).
56. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729.e27 (2018).
57. Espinosa Fernandez, J. R. *et al.* Identification of triple-negative breast cancer cell lines classified under the same molecular subtype using different molecular characterization techniques: Implications for translational research. *PLoS One* **15**, (2020).
58. González, A. J., Setty, M. & Leslie, C. S. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nat. Genet.* **47**, 1249–1259 (2015).
59. Scheel, C. *et al.* Paracrine and autocrine signals induce and maintain mesenchymal and stem cell states in the breast. *Cell* **145**, 926–940 (2011).
60. Lin, J. R. *et al.* Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *Elife* **7**, (2018).
61. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* (2020) doi:10.1038/s41586-019-1876-x.
62. Karaayvaz, M. *et al.* Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* **9**, (2018).
63. Leggett, S. E. *et al.* Morphological single cell profiling of the epithelial-mesenchymal transition. *Integr. Biol. (United Kingdom)* **8**, 1133–1144 (2016).
64. Hu, Y. & Smyth, G. K. ELDA: Extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *J. Immunol. Methods* **347**, 70–78 (2009).
65. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
66. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
67. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
68. Lazarus, J. *et al.* Optimization, design and avoiding pitfalls in manual multiplex fluorescent immunohistochemistry. *J. Vis. Exp.* **2019**, (2019).
69. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).

**Acknowledgements**
We thank Jennifer Fields, Rebecca O'Meara, Dr. Radu V. Stan, Dr. Walburga Croteau and Dr. Fred W. Kolling IV for technical assistance; Molecular Biology Shared Resource at the Norris Cotton Center for assistance with RNA-Seq and ATAC-Seq assays; Dartlab (flow cytometry), microscopy and pathology shared resources at the Norris Cotton Cancer Center. Research reported in this publication was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health (under grant number P20GM104416), a Prouty Pilot Grant from Friends of the Norris Cotton Cancer Center, funding from The Elmer R. Pfefferkorn & Allan U. Munck Education and Research Fund at the Geisel School of Medicine at Dartmouth and an NCI Cancer Center Support Grant (5P30CA023108-40). This work was supported by a Ramanujan Fellowship (SB/S2/RJN-049/2018) awarded by Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India (to M.K.J.) and funding from the NIH 5R00CA201574-05 (to D.R.P.).

**Author contributions**
This project was conceived by M.S.B, N.B.O and D.R.P; N.B.O and M.S.B created the single cell EMT clones, which was the starting point for the project. M.S.B carried out all experiments except for generation of the entropy-based model for deriving heterogeneity scores and EMT scores, which was carried out by B.A and S.H, RNA-Seq and ATAC-Seq data analysis carried out by O.M.W and the EMT scoring by P.C and M.K.J. The manuscript was written by M.S.B and D.R.P with input and edits from all co-authors.

**Competing interests**
The authors declare no competing interests.

**Figure 1: Heterogeneous cell line SUM149PT contains multiple distinct EMT states that can be isolated as single cell clones**.
A) A schematic of the Flow cytometry method used to isolate single cell clones that present as an epithelial (E), three distinct intermediate (EM1, EM2, and EM3), and two mesenchymal (M1 and M2) EMT states. B) *In vitro* assessment of clonal migratory and invasive characteristics as measured in a standard transwell assay (n=3, SD, p-value ****<0.0001, ***<0.001, *<0.05). Canonical EMT marker expression levels as determined by (C) Quantitative RT-PCR (SD, n=4) or (D) immunoblotting to rank SUM149 clones along the EMT spectrum. E) Bright field and F) immunofluorescent images of EMT clones *in vitro* stained with Vimentin and E-cadherin displaying cell morphology and marker expression & localization, respectively. G) EMT signature of EMT clones and parental line generated from the ordinal Multinomial Logistic Regression method of gene scoring and H) distribution of EMT score of the EMT clones among other breast cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE). I) Immunofluorescent staining for E-cadherin (red), Vimentin (green), and KRT8 (white) of four triple negative breast cancer lines (two intermediate – HCC38 and Cal-51, one epithelial – HCC1143, and one mesenchymal – MDA-MB-231) from the CCLE displaying heterogeneous phenotypes.

**Figure 2: Differences in primary tumor growth and metastatic potential between EMT clones.**
A) Tumor growth curves measured weekly following orthotopic injection of clonal and parental cell lines at 2,500 cells exhibit exponential growth differences between EMT states (TumGrowth[30] piecewise regression model breakpoint = 6wks, ***Holm adj. p-value <0.0001, **Holm adj. p-value< 0.01, n=10). B) Survival curve of EMT clones and parental line displaying differences in survival across the parental line and early intermediate EMT states. Cox regression analysis (***Holm adj. p-value <0.0001, *Holm adj. p-value< 0.015). C) Tumor initiating cell frequency calculated by limiting dilution assay with cells injected at 250k, 25k, and 2.5k cells per flank. TIC calculated at 8 weeks post injection. D) Lung sections collected at time of maximum tumor burden ($2cm^3$), with GFP labeled tumor cells, following orthotopic injections as in (A). E) Lungs fixed and stained from (A) with H&E and enumeration of micro (<10 adjacent cells) and macro (10+ adjacent cells) metastatic regions (SD, n=5, micro p-val < 0.02 macro p-value = n.s.). F) Representative brightfield images of micro- and macro- metastases from one mouse lung (EM1).

**Figure 3: Identification of stabilizing transcription factors in the intermediate EMT state by transcriptional and chromatin analysis**
A) Principal Component Analysis (PCA) of the 500 most variable genes between all EMT clones and SUM149 Parental line and B) intermediate clones only from RNA-sequencing. C) UpSet Plot of all differentially expressed genes (referenced to clone E) that are shared and unique to each EMT clone. D) Unsupervised hierarchical clustering of the top 500 differentially expressed genes in all comparisons to clone E (p-value <0.05) E) ATAC-seq peak accessibility measured as counts per million (CPM) normalized Tn5 insertions surrounding consensus peaks, promoter-associated, and distal-intergenic peaks, respectively, for each EMT clone F) Unsupervised hierarchical clustering of transcription factor motif enrichment (-log10 adjusted p-value, hypergeometric test) among accessible chromatin peaks unique to each clone, relative to E. Presence of motifs obtained from the JASPAR database were identified using motifmatchr (P<0.05) and tested for enrichment against the background set of all peaks identified in the respective clone using the hypergeometric test. Asterisk indicates a -log10 p-value enrichment threshold greater than 20, scaled to fit.

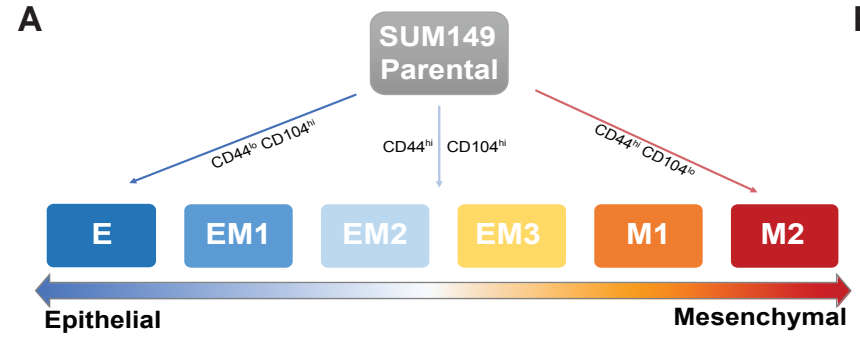
**Figure 4: Identification of stabilizing transcription factors in the intermediate EMT state by multi-omics analysis**
A) Advanced volcano plot of highly significant transcription factors, highlighting the RUNX family Transcription Factor activity, relative to clone E determined by diffTF from ATAC-seq along the X-axis (label cutoffs at 0.1, -0.1 TF activity), plotted against Log2 fold gene expression values of transcription factors on the y-axis (label cutoffs at 1, -1 log2 fold). Transcription Factor classification, determined by transcription factor expression, displayed in bubble color, and number of transcription factor binding sites used to determine TF activity plotted as bubble size. B) Unsupervised hierarchical clustering of TF activity scores (Z-score transformed) compared to clone E (adj. P-value $<1E^{-15}$, asterisk indicates n.s. in that comparison). Right hand side displays TF classification, determined by changes in TF expression (diffTF), as an activator, repressor, not expressed, or undetermined. C) Peak accessibility of Tn5 normalized, merged coverage of three canonical ZEB1 target genes, CDH1, EPCAM, and GRHL2 across all clones from ATAC-sequencing. ZEB1 TF motifs highlighted above signal tracks. D) Protein levels of canonical EMT markers determined western blot following CRISPR-Cas9 mediated knockout of LacZ, RUNX2, and CBFβ in late intermediate and early mesenchymal clones. E) Tumor growth curves measured weekly following orthotopic injection of CBFb knockout at 2,500 cells (TumGrowth[30] piecewise regression model breakpoint = 6wks, *** Initial phase Holm adj. p-value <0.0001 n=5). F) Lungs fixed and stained from (E) with H&E and enumeration of micro (<10 adjacent cells) and macro (10+ adjacent cells) metastatic regions (SD, n=5, EM2 p-val <0.001 EM3 p-value = n.s.)

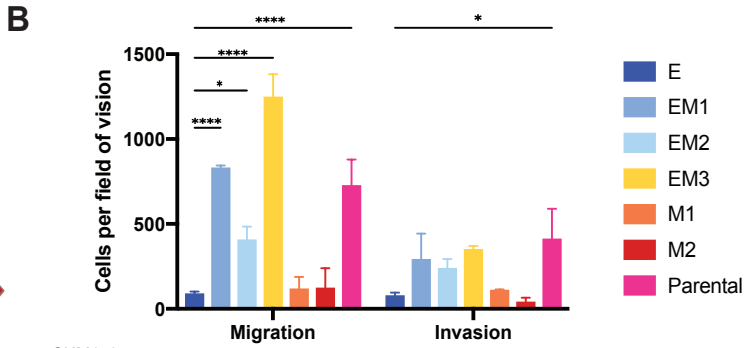
**Figure 5: Multiplexed staining of SUM149 tumors identifies phenotypes and quantifies tumor heterogeneity and overall EMT state**
A) EMT clone-derived tumors resected at $1.5cm^3$ and stained with six marker EMT panel using multiplexed immunostaining (n =~50 images per tumor) B) Empirically determined heterogeneity scores of EMT clone-derived tumors. Rubric: Low (one major cell trait with up to one minor trait), Mid (two major cell traits with up to three minor trait), and High (three or more major cell traits present with two or more minor traits). C) Boxplot of EMT phenotypes generated from inForm® cell phenotype analysis displaying the composition of each clonally derived tumor (n=~50 images per tumor). D) EMT score distribution in clonally derived tumors generated from weighted

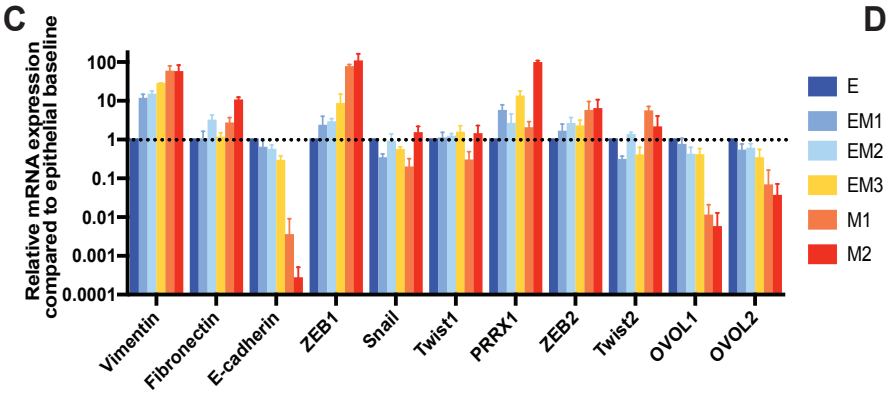
multivariable logistic regression of the phenotypes in C present in each tumor. E) Correlation of EMT (tertile; Epithelial: 0-0.29, Intermediate: 0.3-0.69, Mesenchymal: 0.7-1) and Heterogeneity score in EMT clone-derived tumor images (n=365, chi-square p-val=$1.1E^{-11}$). F) Distribution of EMT and Heterogeneity scores in EM2 and EM3 clone-derived tumors following CBFb knockout, compared to LacZ control tumors. G) Outgrowth lung metastases of EM1, EM2, EM3, and parental tumors stained with six marker EMT panel using multiplexed immunostaining to determine H) EMT phenotypes present and I) EMT and Heterogeneity scores.

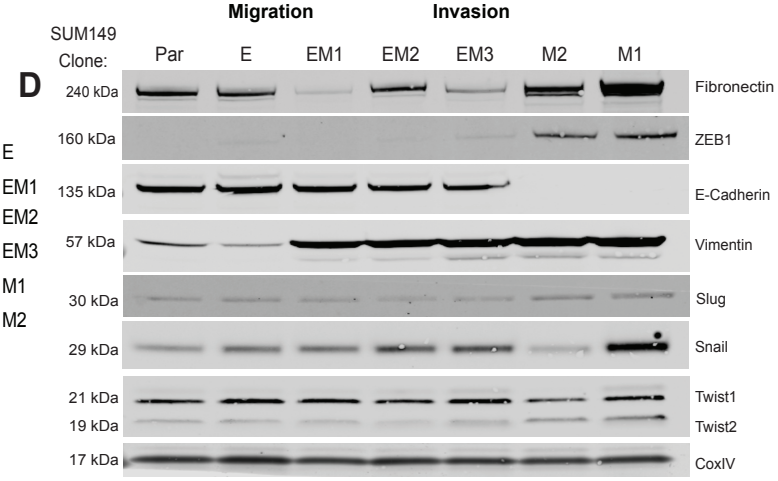
**Figure 6: Identification of shared phenotypes in multiplexed patient tumors validates SUM149 model and presents a novel diagnostic**

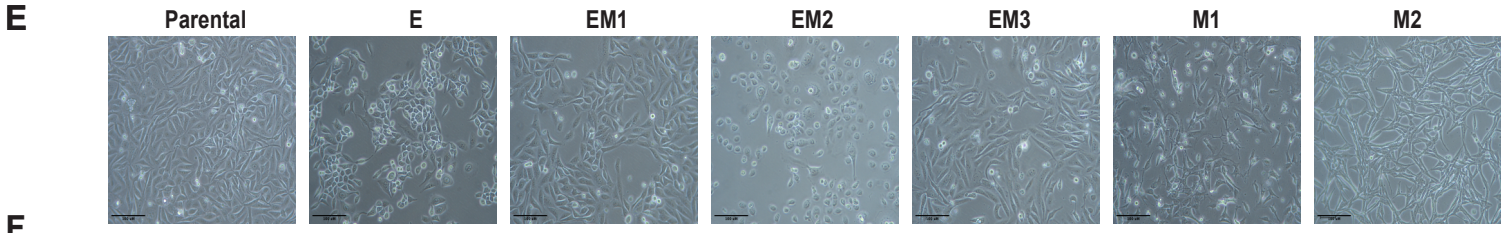
A) Stage III breast cancer patient tumors (n=124) stained with six marker EMT panel using multiplexed immunostaining B) Stage III breast cancer cohort tumor E-M Heterogeneity (EMH) score, determined by an entropy-based model of marker distribution at a single cell level per image, trained and validated by the SUM149 clone tumors, plotted with EMT score (tertile; Epithelial: 0-0.29, Intermediate: 0.3-0.69, Mesenchymal: 0.7-1) generated from weighted multivariable logistic regression of the phenotypes present in each tumor for each patient sample C) Correlation of EMT and Heterogeneity scores from B (Fisher's exact test $p=4.4E^{-5}$) D) Kaplan-Meier plot of overall survival stratified by heterogeneity score and E) forest plot of Multivariate Cox Proportional Hazard model of overall survival for Heterogeneity Score and EMT score accounting for age and subtype. F) Kaplan-Meier plot of overall survival in hormone negative patient samples stratified by heterogeneity score.

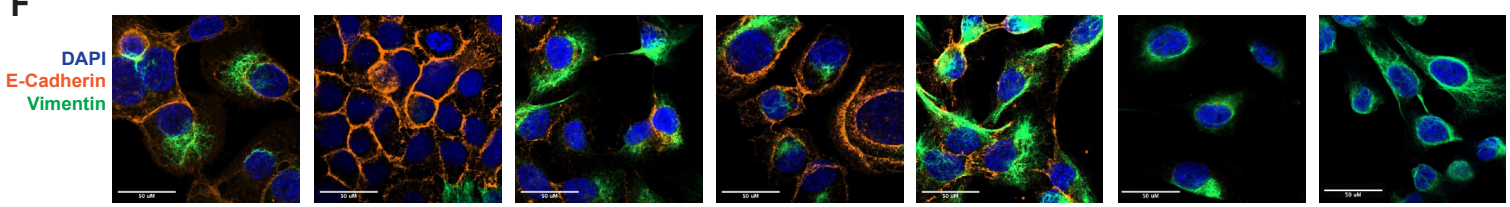

A
SUM149 Parental
CD44lo CD104hi
CD44hi CD104hi
CD44hi CD104lo
E
EM1
EM2
EM3
M1
M2
Epithelial
Mesenchymal

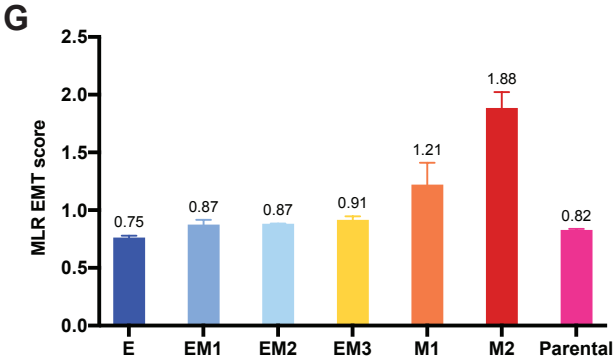

B
Cells per field of vision
1500
1000
500
0
****
*
Migration
Invasion
E
EM1
EM2
EM3
M1
M2
Parental

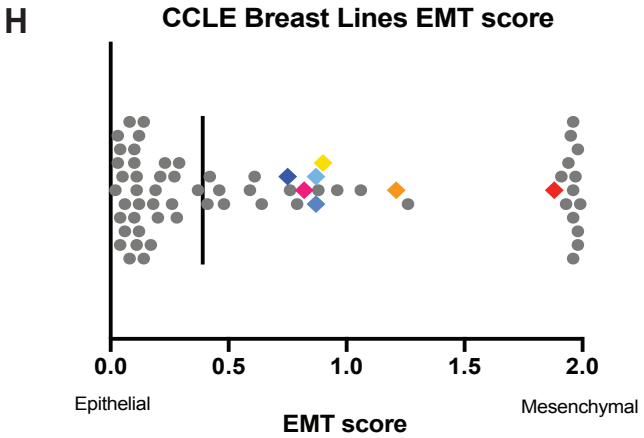

C
Relative mRNA expression compared to epithelial baseline
100
10
1
0.1
0.01
0.001
0.0001
Vimentin
Fibronectin
E-cadherin
ZEB1
Snail
Twist1
PRRX1
ZEB2
Twist2
OVOL1
OVOL2
E
EM1
EM2
EM3
M1
M2

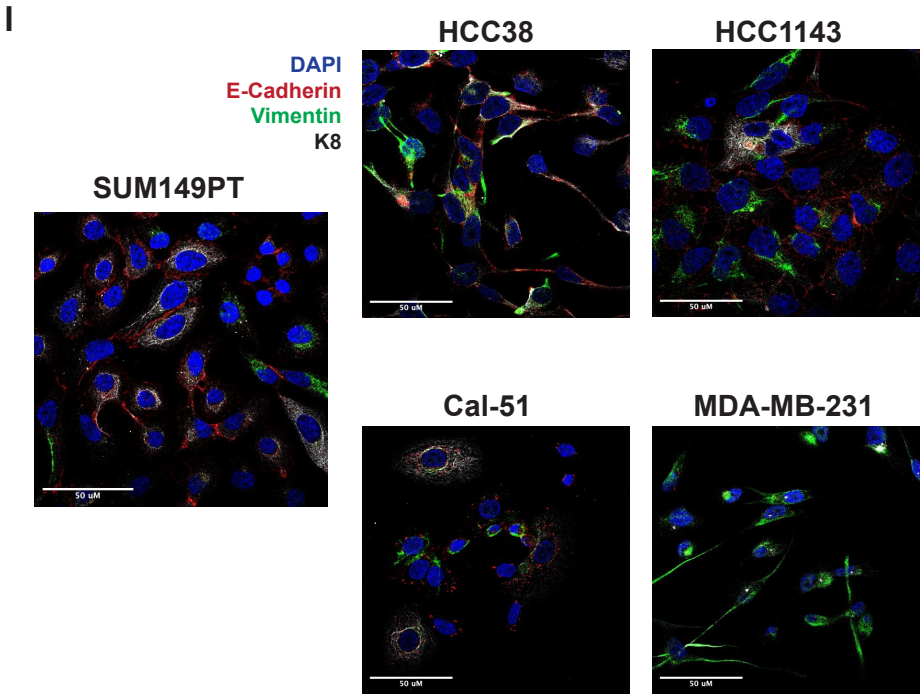

D
SUM149 Clone:
Par
E
EM1
EM2
EM3
M2
M1
240 kDa
Fibronectin
160 kDa
ZEB1
135 kDa
E-Cadherin
57 kDa
Vimentin
30 kDa
Slug
29 kDa
Snail
21 kDa
Twist1
19 kDa
Twist2
17 kDa
CoxIV

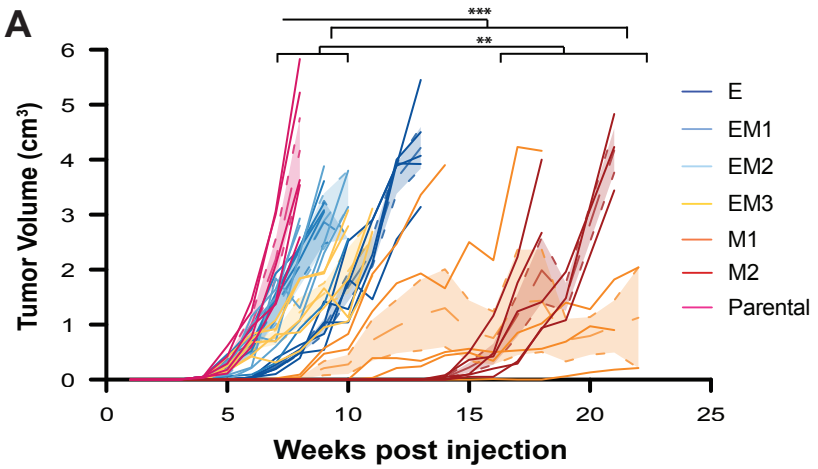

E
Parental
E
EM1
EM2
EM3
M1
M2

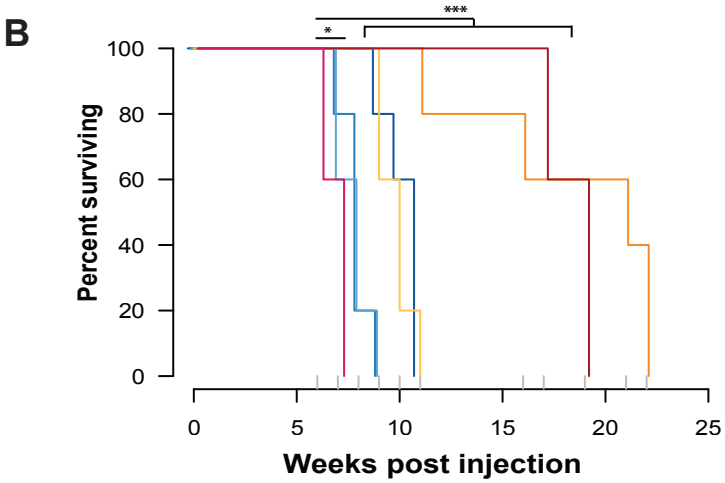

F
DAPI
E-Cadherin
Vimentin


G
MLR EMT score
2.5
2.0
1.5
1.0
0.5
0.0
0.75
0.87
0.87
0.91
1.21
1.88
0.82
E
EM1
EM2
EM3
M1
M2
Parental


H
CCLE Breast Lines EMT score
0.0
0.5
1.0
1.5
2.0
Epithelial
Mesenchymal
EMT score


I
DAPI
E-Cadherin
Vimentin
K8
SUM149PT
HCC38
HCC1143
Cal-51
MDA-MB-231

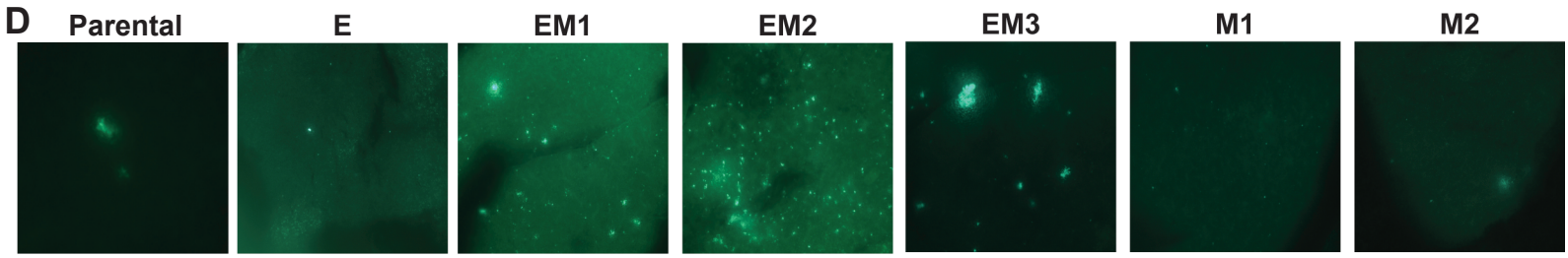


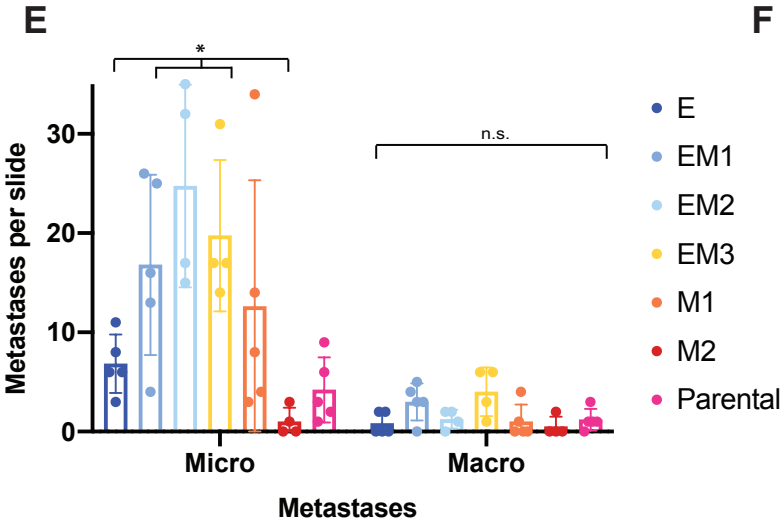



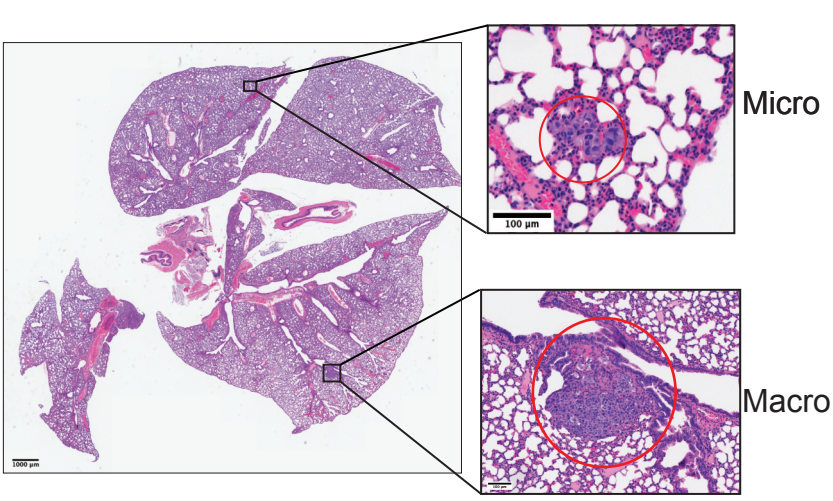

C

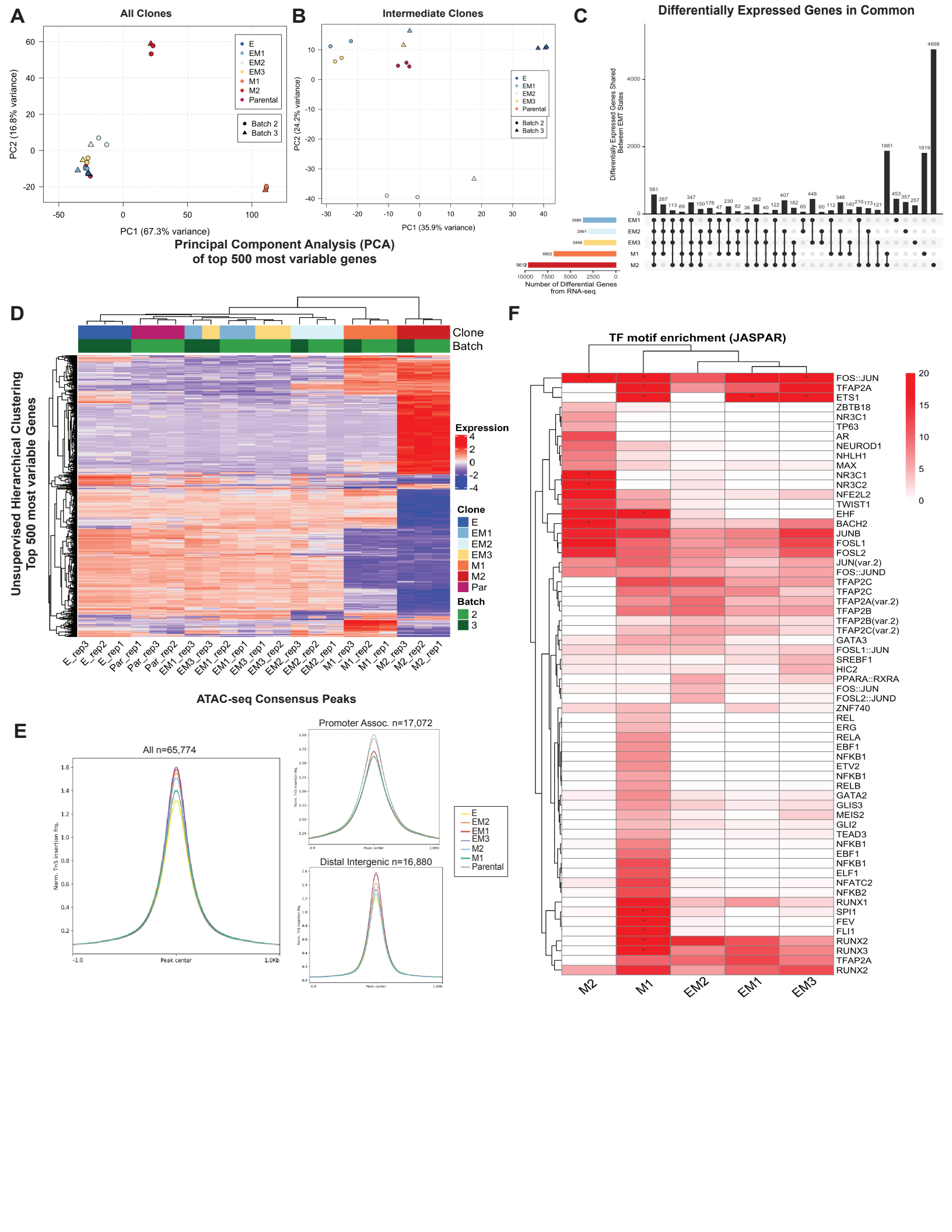
*at 8 weeks

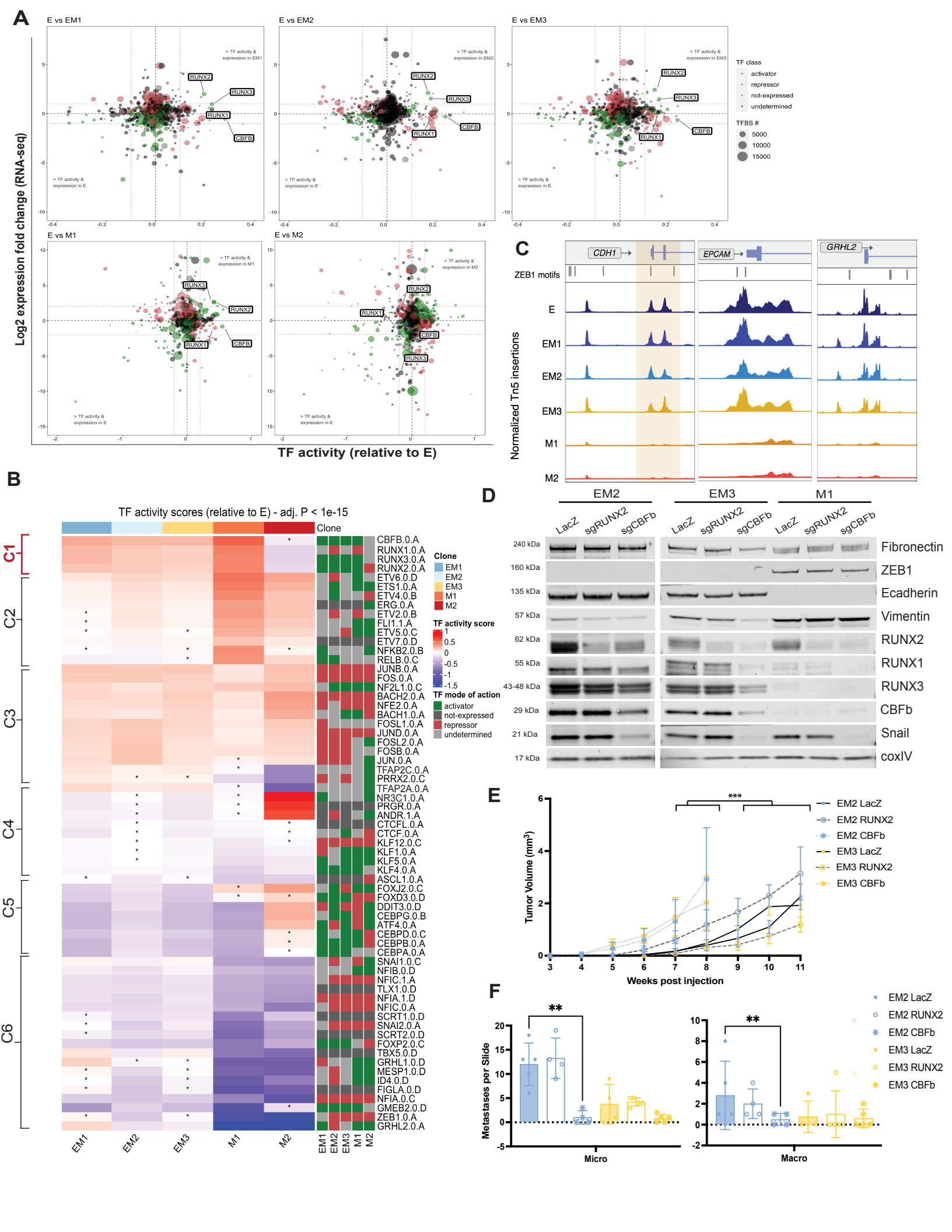
| No. Cells Injected | E | EM1 | EM2 | EM3 | M1 | M2 | Parental |
|---|---|---|---|---|---|---|---|
| 250,000 | 9/10 | 10/10 | 10/10 | 10/10 | 0/10 | 10/10 | 10/10 |
| 25,000 | 8/10 | 10/10 | 10/10 | 10/10 | 0/10 | 2/10 | 10/10 |
| 2,500 | 9/10 | 10/10 | 10/10 | 10/10 | 0/10 | 0/10 | 10/10 |
| TIC | 1/23291 | 1 | 1 | 1 | 0 | 1/75975 | 1 |

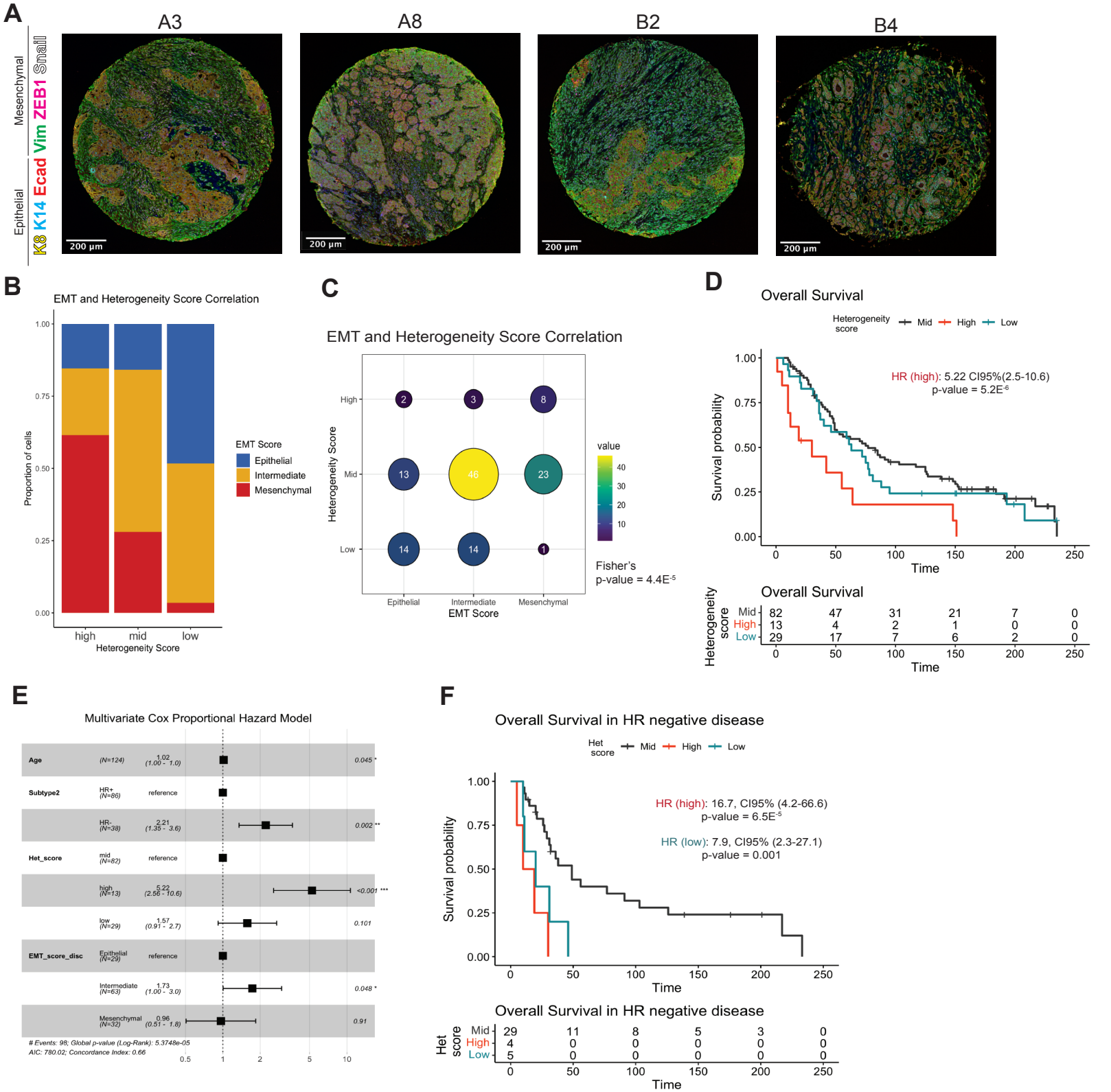


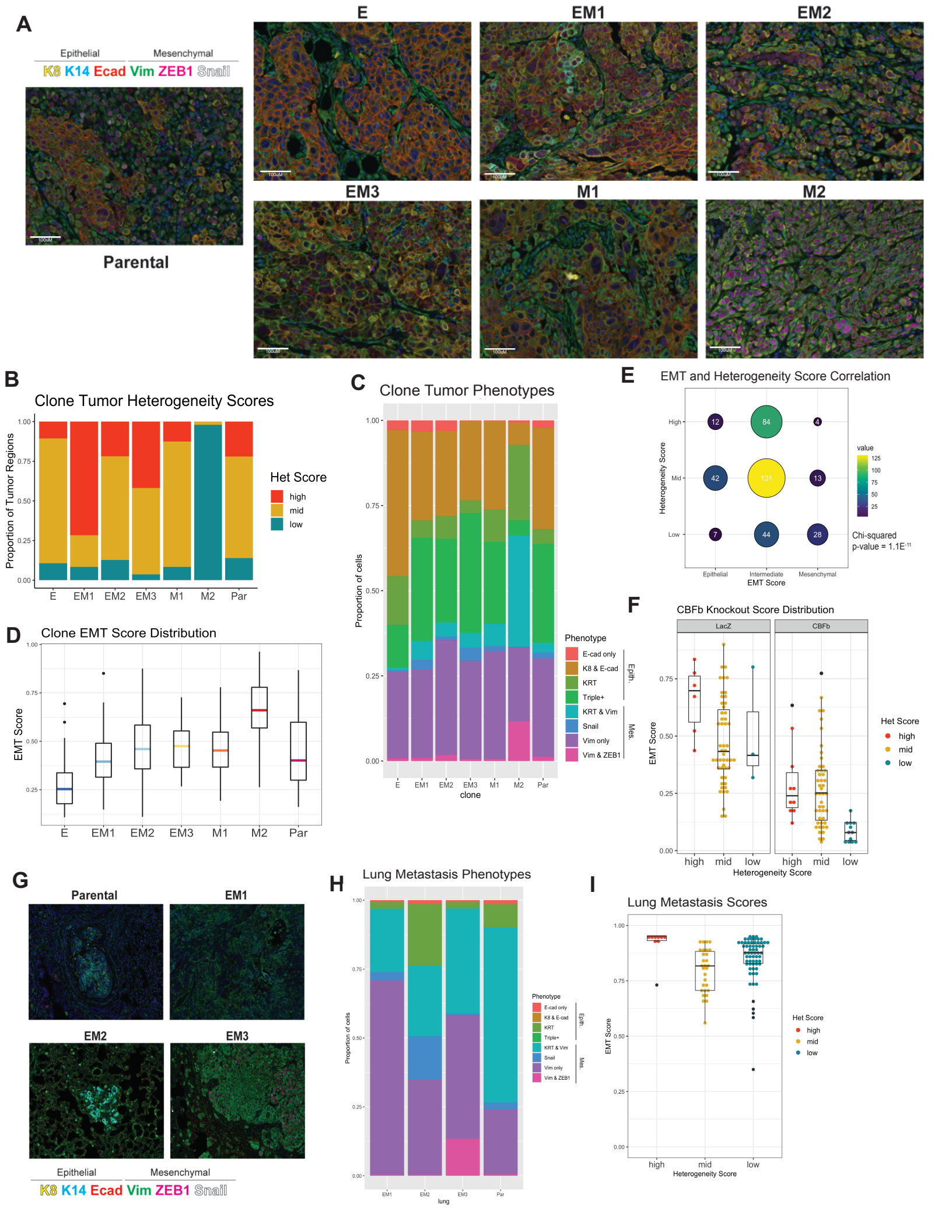




A
All Clones
PC2 (16.8% variance)
PC1 (67.3% variance)
E
EM1
EM2
EM3
M1
M2
Parental
Batch 2
Batch 3
B
Intermediate Clones
PC2 (24.2% variance)
PC1 (35.9% variance)
Principal Component Analysis (PCA)
of top 500 most variable genes
C
Differentially Expressed Genes in Common
Differentially Expressed Genes Shared Between EMT States
Number of Differential Genes from RNA-seq
D
Clone
Batch
Unsupervised Hierarchical Clustering
Top 500 most variable Genes
Expression
Clone
Batch
E_rep3
E_rep2
E_rep1
Par_rep1
Par_rep3
Par_rep2
EM1_rep3
EM3_rep3
EM1_rep2
EM1_rep1
EM3_rep1
EM3_rep2
EM2_rep3
EM2_rep2
EM2_rep1
M1_rep3
M1_rep2
M1_rep1
M2_rep3
M2_rep2
M2_rep1
ATAC-seq Consensus Peaks
E
All n=65,774
Promoter Assoc. n=17,072
Distal Intergenic n=16,880
Norm. Tn5 insertion frq.
Peak center
F
TF motif enrichment (JASPAR)
FOS::JUN
TFAP2A
ETS1
ZBTB18
NR3C1
TP63
AR
NEUROD1
NHLH1
MAX
NR3C1
NR3C2
NFE2L2
TWIST1
EHF
BACH2
JUNB
FOSL1
FOSL2
JUN(var.2)
FOS::JUND
TFAP2C
TFAP2C
TFAP2A(var.2)
TFAP2B
TFAP2B(var.2)
TFAP2C(var.2)
GATA3
FOSL1::JUN
SREBF1
HIC2
PPARA::RXRA
FOS::JUN
FOSL2::JUND
ZNF740
REL
ERG
RELA
EBF1
NFKB1
ETV2
NFKB1
RELB
GATA2
GLIS3
MEIS2
GLI2
TEAD3
NFKB1
EBF1
NFKB1
ELF1
NFATC2
NFKB2
RUNX1
SPI1
FEV
FLI1
RUNX2
RUNX3
TFAP2A
RUNX2
M2
M1
EM2
EM1
EM3


A
E vs EM1
E vs EM2
E vs EM3
E vs M1
E vs M2
Log2 expression fold change (RNA-seq)
TF activity (relative to E)
RUNX2
RUNX3
RUNX1
CBFB
TF class
activator
repressor
not-expressed
undetermined
TFBS #
5000
10000
15000
B
TF activity scores (relative to E) - adj. P < 1e-15
Clone
EM1
EM2
EM3
M1
M2
TF activity score
TF mode of action
activator
not-expressed
repressor
undetermined
C1
C2
C3
C4
C5
C6
C
CDH1
EPCAM
GRHL2
ZEB1 motifs
Normalized Tn5 insertions
D
LacZ
sgRUNX2
sgCBFb
Fibronectin
ZEB1
Ecadherin
Vimentin
RUNX2
RUNX1
RUNX3
CBFb
Snail
coxIV
E
Tumor Volume (mm³)
Weeks post injection
EM2 LacZ
EM2 RUNX2
EM2 CBFb
EM3 LacZ
EM3 RUNX2
EM3 CBFb
F
Metastases per Slide
Micro
Macro


A
A3
A8
B2
B4
Mesenchymal
Epithelial
K8 K14 Ecad Vim ZEB1 Snail
200 µm
B
EMT and Heterogeneity Score Correlation
Proportion of cells
Heterogeneity Score
high
mid
low
EMT Score
Epithelial
Intermediate
Mesenchymal
C
EMT and Heterogeneity Score Correlation
Heterogeneity Score
High
Mid
Low
EMT Score
value
Fisher's
p-value = 4.4E-5
D
Overall Survival
Heterogeneity score
Mid
High
Low
HR (high): 5.22 CI95%(2.5-10.6)
p-value = 5.2E-6
Survival probability
Time
Overall Survival
Heterogeneity score
Mid 82 47 31 21 7 0
High 13 4 2 1 0 0
Low 29 17 7 6 2 0
E
Multivariate Cox Proportional Hazard Model
Age (N=124) 1.02 (1.00 - 1.0) 0.045 *
Subtype2 HR+ (N=86) reference
HR- (N=38) 2.21 (1.35 - 3.6) 0.002 **
Het_score mid (N=82) reference
high (N=13) 5.22 (2.56 - 10.6) <0.001 ***
low (N=29) 1.57 (0.91 - 2.7) 0.101
EMT_score_disc Epithelial (N=29) reference
Intermediate (N=63) 1.73 (1.00 - 3.0) 0.048 *
Mesenchymal (N=32) 0.96 (0.51 - 1.8) 0.91
# Events: 98; Global p-value (Log-Rank): 5.3748e-05
AIC: 780.02; Concordance Index: 0.66
F
Overall Survival in HR negative disease
Het score
Mid
High
Low
HR (high): 16.7, CI95% (4.2-66.6)
p-value = 6.5E-5
HR (low): 7.9, CI95% (2.3-27.1)
p-value = 0.001
Survival probability
Time
Overall Survival in HR negative disease
Het score
Mid 29 11 8 5 3 0
High 4 0 0 0 0 0
Low 5 0 0 0 0 0


A
Epithelial
Mesenchymal
K8 K14 Ecad Vim ZEB1 Snail
Parental
E
EM1
EM2
EM3
M1
M2
B
Clone Tumor Heterogeneity Scores
Proportion of Tumor Regions
Het Score
high
mid
low
E EM1 EM2 EM3 M1 M2 Par
C
Clone Tumor Phenotypes
Proportion of cells
clone
Phenotype
E-cad only
K8 & E-cad
KRT
Triple+
KRT & Vim
Snail
Vim only
Vim & ZEB1
Epith.
Mes.
D
Clone EMT Score Distribution
EMT Score
E
EMT and Heterogeneity Score Correlation
Heterogeneity Score
High
Mid
Low
Epithelial
Intermediate
Mesenchymal
EMT Score
value
Chi-squared
p-value = 1.1E-11
F
CBFb Knockout Score Distribution
LacZ
CBFb
EMT Score
Heterogeneity Score
high mid low
Het Score
G
Parental
EM1
EM2
EM3
Epithelial
Mesenchymal
K8 K14 Ecad Vim ZEB1 Snail
H
Lung Metastasis Phenotypes
Proportion of cells
lung
EM1 EM2 EM3 Par
I
Lung Metastasis Scores
EMT Score
Heterogeneity Score
high mid low
Het Score