Genomic and local microenvironment effects shaping epithelial-to-mesenchymal trajectories in cancer

Guidantonio Malagoli Tagliazucchi[1,#], Anna J Wiecek[1,#], Eloise Withnell[1], Maria Secrier[1,*]

[1] UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, UK

[#]These authors contributed equally.

*To whom correspondence should be addressed (m.secrier@ucl.ac.uk)

ABSTRACT

The epithelial to mesenchymal transition (EMT) is a key cellular process underlying cancer progression, with multiple intermediate states whose molecular hallmarks remain poorly characterized. To fill this gap, we explored EMT trajectories in 7,180 tumours of epithelial origin and identified three macro-states with prognostic and therapeutic value, attributable to epithelial, hybrid E/M and mesenchymal phenotypes. We show that the hybrid state is remarkably stable and linked with increased aneuploidy and APOBEC mutagenesis. We further employed spatial transcriptomics and single cell datasets to show that local effects impact EMT transformation through the establishment of distinct interaction patterns with cytotoxic, NK cells and fibroblasts in the tumour microenvironment. Additionally, we provide an extensive catalogue of genomic events underlying distinct evolutionary constraints on EMT transformation. This study sheds light on the aetiology of distinct stages along the EMT trajectory, and highlights broader genomic and environmental hallmarks shaping the mesenchymal transformation of primary tumours.

INTRODUCTION

The epithelial to mesenchymal transition (EMT) is a cellular process in which polarized epithelial cells undergo multiple molecular and biochemical changes and lose their identity in order to acquire a mesenchymal phenotype[1]. EMT occurs during normal embryonic development, tissue regeneration, wound healing, but also in the context of disease[1,2]. In cancer, it promotes tumour progression with metastatic expansion[3]. Recent studies have uncovered that EMT is not a binary switch but rather a continuum of phenotypes, whereby multiple hybrid EMT states underly and drive the transition from fully epithelial to fully mesenchymal transformation[4,5]. Elucidating the evolutionary trajectories that cells take to progress through these states is key to understanding metastatic spread and predicting cancer evolution.

The transcriptional changes accompanying EMT in cancer have been widely characterised and are governed by several transcription factors, including Snail, Slug, Twist and zinc fingers *ZEB1* and *ZEB2*[6-8]. EMT appears driven by waves of gene regulation underpinned by checkpoints, such as *KRAS* signalling driving the exit from an epithelial state, dependent upon *EGFR* and *MET* activation[9].

However, EMT progression is not only characterized by transcriptional alterations of regulatory circuits; the genetic background of the cell can also impact its capacity to undergo this transformation. Gain or loss of function mutations in a variety of genes, including *KRAS*[10], *STAG2*[11], *TP53*[12], as well as amplifications of chromosomes 5, 7 and 13 have been shown to promote EMT[13]. Several pan-cancer studies have also linked copy number alterations, miRNAs and immune checkpoints with EMT on a broader level[14,15]. Mathematical models have been developed to describe the switches between epithelial and mesenchymal states[4] but without considering any genomic dependencies.

Despite extensive efforts to study the dynamics of EMT, some aspects of this process remain poorly characterized. In particular, most of the studies mentioned considered EMT as a binary switch and failed to capture intrinsic and local microenvironment constraints that may change

along the continuum of EMT transformation. Single cell matched DNA- and RNA-seq datasets would ideally be needed for this purpose, but they are scarce. To overcome this, we have integrated data from the Cancer Genome Atlas (TCGA), MET500[16], MetMap[17], GDSC[18], POG570[19], as well as orthogonal spatial transcriptomics and single cell datasets to characterise the EMT continuum, its spatial context and interactions established with cells in the tumour microenvironment. By mapping 7,180 tumours of epithelial origin onto a "timeline" of epithelial-to-mesenchymal transformation, we identified discrete EMT macro-states and derived a catalogue of genomic hallmarks underlying evolutionary constraints of these states. These genomic events shed light into the aetiology of hybrid E/M and fully mesenchymal phenotypes, and could potentially act as early biomarkers of invasive cancer.

RESULTS

Pan-cancer reconstruction of EMT trajectories from transcriptomics data

We hypothesised that a pan-cancer survey of EMT phenotypes across bulk sequenced samples should capture a broad spectrum of the phenotypic variation one may expect to observe at single cell level, and this could be linked with genomic changes accompanying EMT transformation. To explore the EMT process within bulk tumour samples, we employed a cohort of primary tumours of epithelial origin (n = 7,180) spanning 25 cancer types from TCGA. The bulk RNA-seq data from these tumours underlie multiple transcriptional programmes reflecting different biological processes, including EMT. Inspired by McFaline-Figueroa et al[9], we quantified the levels of EMT in these bulk tumours against a consensus reference single cell RNA sequencing (scRNA-seq) dataset derived from non-transformed epithelial cells as well as cancer cells from multiple tissues that have been profiled at different times during the epithelial to mesenchymal transition *in vitro*. These data allowed us to reconstruct a generic "pseudo-timeline" of spontaneous EMT transformation onto which we projected the bulk sequenced samples from TCGA, positioning them within the continuum of EMT states (Figure 1a). To account for signals from non-tumour cells in the

microenvironment, which have been recently shown by Tyler and Tirosh[20] to confound the EMT state inference in bulk data, we adjusted the expression of all genes based on the tumour purity inferred from matched DNA-sequencing (see Methods). These corrected expression profiles were then mapped to the single cell reference trajectories and an EMT pseudotimeline was reconstructed that accounted for potential tumour contamination. Moreover, the confounding effects highlighted by Tyler and Tirosh are prominent when using specific mesenchymal signatures that overlap with markers of cancer-associated fibroblasts (CAFs), but our approach should also be generally less prone to such biases as we employ a whole-transcriptome reference of single cells progressing through EMT rather than selected markers. Thus, the resulting signal should more reliably capture the transformation of epithelial cells rather than immune/stromal programmes, and is expected to reflect the average EMT state across the entire tumour cell population.

Using this approach, we reconstructed the EMT pseudotime trajectory across multiple cancer tissues (Figure 1b, Supplementary Table S1). The expression of canonical epithelial and mesenchymal markers was consistent with that observed in the scRNA-seq data and expectations from the literature (Supplementary Figure S1a). Along the pseudotime, we observed frequent co-expression of such markers, which could reflect a hybrid E/M state[21] (Supplementary Figure S1b). Importantly, when analysing cancer types individually by aligning against breast, lung and prostate reference cell lines rather than to a consensus reference, the pseudotime reconstruction and EMT scores obtained were strongly correlated with those from the pan-cancer analysis (Supplementary Figure S1c), thereby demonstrating that the pan-cancer methodology can broadly recapitulate phenotypes identified in individual cancers. Furthermore, the reconstructed pseudotimeline closely matched increasing levels of EMT transformation in independent cell line experiments from a variety of systems (Supplementary Figure S1d), further validating our approach experimentally.

EMT macro-states underlie the continuum of mesenchymal transformation in cancer

To characterise the dominant EMT states governing the continuum of transcriptional activity described above, we discretised the pseudotime trajectory based on expression values of canonical EMT markers using a Hidden Markov Model approach and uncovered three macro-states: epithelial (EPI), hybrid EMT (hEMT) and mesenchymal (MES) (Figure 1b-c, Supplementary Figure S1b). These states were robust to varying levels of gene expression noise (Supplementary Figure S1e). As expected, the probability for the cancer cells to switch from the epithelial to the hEMT (0.32) state was higher than the probability to passage directly into the mesenchymal state (0.09). The hEMT tumours tended to remain in the same state 42% of the times, suggesting this state could be more stable than anticipated – as previously stipulated[22] and consistent with observations that a fully mesenchymal state is not always observed[23].

The EMT scores progressively increased between the EPI, hEMT and MES states, as expected (Figure 1d). Reassuringly, in an independent cohort of metastatic samples (MET500), EMT levels were relatively elevated along the transformation timeline compared to TCGA samples and were most abundantly falling within the hEMT state (Figure 1d, Supplementary Figure S2a). Interestingly, we also observed possible cases of a reversion to an epithelial state in metastatic samples, which is to be expected when colonizing a new environmental niche.

We also applied our EMT scoring methodology to the MetMap resource, which has catalogued the metastatic potential of 500 cancer cell lines across 21 cancer types. The invasion potential of these cell lines increased with the EMT score as expected (Figure 1e, Supplementary Figure S2b). Cell lines classified as MES by our HMM model were predominantly metastatic, while hEMT cases had a weak invasion potential.

At tissue level, the proportion of samples in each EMT state was variable (Figure 1f), with hEMT dominating in head and neck, oesophageal, lung squamous and pancreatic carcinomas, while adenoid cystic, kidney carcinomas and melanomas were highly

mesenchymal. When investigating molecular subtypes already described for a variety of cancers, most of them did not show distinct distributions by EMT state (Supplementary Figure S2c). Nevertheless, the ovarian mesenchymal subtype was reassuringly enriched in hEMT and MES cancers, and the same could be observed for genomically stable gastric cancers, which have been linked with diffuse histology and enhanced invasiveness[24].

The EMT classification was significantly correlated with the clinical cancer stage (Chi-square test $p<<0.0001$, Supplementary Figure 2d), with transformed samples (hEMT/MES) found to be 1.3-fold and 1.2-fold enriched in late-stage tumours, respectively, while the epithelial state was 1.4-fold overrepresented in early-stage cancers (Figure S1i). In primary tumours, 12% of the profiled samples were classified as fully transformed (MES), with the majority of them (60%) annotated as late-stage tumours (Supplementary Table S2). Notably, metastatic samples available from TCGA (n=343) were overwhelmingly classed as MES (94%, Supplementary Figure 2e), suggesting that the transformed phenotype is more pronounced in metastases than in primary tumours, as expected. While the correlation between cancer stage and EMT state does not appear as strong as potentially anticipated in primary tumours, the proportion of observed late-stage cancers increases as we move from EPI to hEMT and MES cancers in most cancer tissues, with mesenchymal cholangiocarcinomas, esophageal and kidney chromophobe cancers being entirely late stage (Supplementary Fig 2f). The fact that some early stage cancers are classified as fully mesenchymal (5%) may suggest early evidence for the phenotypic transformation required for metastasis. Indeed, multiple studies have demonstrated the activation of the EMT transcriptional programme in the early stages of cancer[10,25]. Even the hEMT phenotype was hypothesised to be sufficient for promoting metastatic dissemination[26], although this is likely tissue-dependent.

Tumour cell extrinsic hallmarks of EMT

Multiple microenvironmental factors, including tumour associated macrophages, secreted molecules (IL-1, TNF-α) or hypoxia, have been extensively described to promote EMT[27-29].

However, their macro-state specificity is less well characterised. The three EMT macro-states we have described within TCGA cancers displayed no significant difference in tumour purity, confirming that non-tumour cell content did not play a significant part in assigning these states (Supplementary Figure S3a). This also made comparisons in tumour microenvironment compositions equitable across the three states. We observed that cytotoxic, γδ T and endothelial cells were progressively enriched with increased stages of EMT transformation (Figure 2a-b, Supplementary Figure S3b), suggesting that the fully mesenchymal state is most often linked with “immune hot” tumours. In line with this hypothesis, these tumours also showed the highest exhaustion levels (Figure 2c). The hEMT samples still displayed a relatively higher level of fibroblasts pan-cancer, despite the tumour purity correction, potentially suggesting a real biological association (Figures 2a-b). In fact, when examining these associations by cancer tissue, we noticed that active fibroblasts were often similarly enriched in hEMT and MES samples compared to epithelial ones (Supplementary Figure S3c), which would be expected with increased tumour progression. However, due to the confounding effects between fibroblast and hEMT markers as highlighted by Tyler and Tirosh[20], we acknowledge that part of the signal recovered may still not be unambiguously attributed to either the cancer or microenvironmental component despite our best efforts to correct for this.

Samples with a transformed phenotype (MES, hEMT) presented significantly elevated hypoxia levels in several cancer types (Figure 2d). Hypoxia has previously been shown to promote EMT by modulating stemness properties[30]. We found that CD44, an established cancer stem cell marker known to promote EMT[31,32], was most highly expressed in the hEMT state across cancers (Figure 2e), and elevated levels of several other stemness signatures most often accompanied the hEMT and MES macro-states (Figure S3d-e). Unlike mesenchymal samples, the majority of hEMT tumours (20%) were characterized by both hypoxia and CD44 expression (Supplementary Table S3). Thus, the interplay between hypoxia and stemness

may play a greater role in attaining the hEMT state compared to the fixation of a fully mesenchymal phenotype.

Spatially-resolved EMT patterning reveals local microenvironmental effects

The associations identified between EMT and tumour microenvironmental features, including fibroblasts and hypoxia, are interesting – but could potentially be confounded by averaged signals in bulk data. Indeed, bulk data is not able to capture the diversity of EMT states that may be comprised within an entire tumour, and may miss spatial effects on EMT transformation. To shed further light into these associations, we employed spatial transcriptomics data from three breast cancer slides from 10x genomics generated with the Visium platform, along with multi-region profiling of eight breast tumours generated using ST2K as described by Andersson et al[33] to explore the spatial heterogeneity of EMT and links with other phenotypes within the cancer tissue. We observe remarkable heterogeneity of EMT transformation across the tissue, with occasional clustering of EMT states within epithelial pockets (Figures 3a-b,i-j, Supplementary Figure S4a-b). Fibroblasts are generally observed to surround the epithelial neoplastic areas, with the amount of infiltration varying from patient to patient (Figures 3c,k, Supplementary Figure S4c). Furthermore, hypoxia was generally increased within areas presenting more advanced mesenchymal transformation, with strongest correlations observed either in hybrid or fully mesenchymal spots (Figures 3d,h,i,p, Supplementary Figure S4d,f).

We used clustering to identify areas within the tissue that present more homogeneous patterns of expression (see Methods, Figures 3e-f,m-n) and investigated the tumour microenvironment composition within these clusters in relation to EMT states. We confirm the associations between the MES state and CD8+/CD4+ T cell infiltration, monocytes and macrophages observed in bulk data (Figures 3g,o, Supplementary Figure 4e). We further uncover associations between transformed (hEMT/MES) areas and dendritic cells and polymorphonuclear leukocytes (PMNs). Fibroblast infiltration always appears more strongly

linked with highly transformed areas of the tumour, and occasionally with epithelial spots too. Within a larger dataset of multi-region spatial transcriptomics slides from multiple patients, we found that intermediate levels of transformation (hEMT) uniquely associated both with MSC/iCAF-like and myCAF-like cells, whereas the EPI state was only linked with the former and the MES state with the latter (Figure 3q). Furthermore, natural killer (NK) cells were the only cell type to solely associate with hEMT spots, potentially suggesting NK activation strategies may be effective against tumour cells in this hybrid state.

Overall, this analysis recapitulates some of the features observed in bulk tumours, while uncovering a fine-grained heterogeneous landscape of cell states and associations. Although some of the patterns are recurrent, there is a high degree of spatial and patient-to-patient variation in EMT and TME composition, suggesting that local spatial effects are likely important determinants of EMT progression. While we found a moderate association between hypoxia and EMT transformed cells in breast cancer tissues examined, the inter-patient and tissue-specific heterogeneity became clearer when examining spatial maps from prostate transcriptomes from Berglund et al[34] (Supplementary Figure S5a-b). Here, only one of two tissue sections showed a marked correlation between hypoxia and hEMT within the cancer areas, and not in normal or prostatic intraepithelial neoplasia (PIN). Furthermore, hypoxia was more strongly associated with high rather than moderate levels of EMT transformation in a 3D micro-tumour breast model where collective migration had been induced[35] (Supplementary Figure 5c). The associations in this experiment may nevertheless be overshadowed by the fact that early stages of transformation from ductal carcinoma in situ to invasive ductal carcinoma are being investigated. Hypoxia may be a clearer phenotype in more advanced cancers with hEMT or MES phenotypes, which is something we could not capture in these analyses as all tumours originated from stage I or II cancers.

Despite the large spatial variability, the continuum of EMT transformation is abundantly clear in spatially profiled slides, and stresses the importance of examining local effects to understand tumour progression and responses to treatment.

EMT diversity in single cell data

The analyses performed so far have been focused on datasets where the EMT signal is either measured in bulk across the entire tumour, or via spatial techniques within finer grained spots but still comprising multiple cells. This of course limits our ability to comprehend the true EMT heterogeneity of a tissue, as we lack single cell resolution of phenotypes. To further investigate this, we employed matched bulk and single cell data from the same cancer patients to test whether the EMT profiles estimated in bulk tissue might capture similar states as those seen at single cell level. Using breast cancer data from Chung et al[36], we were able to confirm a good correlation between the estimated EMT pseudotimeline in bulk and the average EMT signal captured from single cell data (Figure 4a). This provides some further reassurance that the bulk estimates, while fairly generic, do approximate the average signal across the tumour. Moreover, we investigated the interactions established between cells in different EMT states and other cell populations in the tumour microenvironment (Figure 4b). Within this dataset, the number of interactions with non-tumour cells increased with increasing EMT transformation, closely reflecting the observations in bulk tumours.

To further explore this in multiple cancer types, we investigated single cell data from breast, lung, colorectal and ovarian tumours as described by Qian et al[37]. We found that tumour cells in the EPI, hEMT and MES state formed distinct clusters that often reflected an EMT progression and were well separated from clusters of other cells in the microenvironment (Figure 4c). The majority of tumour cell clusters were clearly distanced from fibroblast clusters, confirming our premise that a whole-transcriptome reference would be better able to distinguish true malignant cells on the course of mesenchymal transformation from CAFs. Nevertheless, a minority of cells appear more similar to T cells (Figure 4c breast and lung panels) or fibroblasts (Figure 4c breast and colorectal panels), although they are not dispersed throughout these clusters but rather grouped at the extremity.

The cell-cell interaction landscape was quite diverse, with hEMT cells generally showing fewer interactions with the TME amongst the three states, while epithelial-fibroblast interactions were enhanced in lung and colon cancers, and mesenchymal-fibroblast interactions in ovarian

cancers (Figure 4d). These observations, along with the spatial transcriptomics data, suggest that the relation between EMT transformation of tumour cells and their interactions with the TME is likely a complex one, highly tissue specific and driven by local spatial effects. Ideally, single cell, spatially resolved longitudinal datasets would be needed to fully resolve such heterogeneity.

Tumour cell intrinsic hallmarks of EMT

Despite the insights into EMT spatial organisation and TME interactions observed previously, the genomic influence in these datasets cannot be measured. While lacking the granularity of single cell or spatially-resolved datasets, bulk sequenced datasets with matched genomics and transcriptomics measurements are still the main resource that allows us to glance into potential genomic determinants of cellular plasticity. We thus returned to the TCGA dataset to explore the genomic background that underlies EMT transformation. Intrinsic cell properties such as increased proliferation, mutational and copy number burden, as well as aneuploidy would be expected along the EMT trajectory. Across distinct tissues, these changes were most pronounced in the hEMT state (Figure 5a). While the clonality of tumours in the three states did not differ significantly (Kruskal-Wallis $p>0.05$), the number of clonal and subclonal mutations increased with the state of EMT transformation. Interestingly, the hEMT group also presented higher levels of centrosome amplification, which have been linked with increased genomic instability[38,39] and poor prognosis[40].

Such alterations to the genomic integrity of the cells result from multiple mutational processes. These processes leave recognizable patterns in the genome termed “mutational signatures”, which in their simplest form constitute of trinucleotide substitutions and have been broadly characterised across cancers[41]. However, their involvement in EMT transformation is poorly understood. To investigate whether any neoplastic process introducing mutations in the genomes was conditioned by EMT, we modelled the associations between mutational signatures and EMT using linear mixed effects models (Methods, Figure 5b). The mismatch

repair deficiency signature SBS6 and the smoking-linked SBS4 signature were significantly increased in hEMT tumours, while SBS39, of unknown aetiology, was most elevated in fully transformed tumours (Figure 5c). The APOBEC mutagenesis signatures SBS2 and SBS13 also appeared elevated in hEMT tumours, in line with observations that inflammation-induced upregulation of the activation-induced cytidine deaminase (AID) enzyme, a component of the APOBEC family, triggers EMT[42]. However, when taking the tissue effect into account in the modelling procedure, no pan-cancer tissue agnostic associations between mutational processes and EMT were identified – suggesting that the previously captured associations are likely tissue-restricted. Thus, while some influence may exist on EMT from tissue-specific mutational processes, there was no evidence of an overarching mutagen that might induce EMT.

Genomic driver events underlying the EMT transformation pan-cancer

Beyond the broader hallmarks discussed above, we sought to identify specific genomic changes creating a favourable environment for EMT transformation or imposing evolutionary constraints on its progression. We observed that subclonal diversification followed distinct routes according to the pattern of EMT transformation for several genes, including *BRAF*, *PMS1* and *FNBP1* (Figure 5d). The fraction of cancer cells harbouring *BRAF* mutations, frequently acquired in melanoma, was markedly increased in mesenchymal samples, suggesting that a clonal fixation of this event may be key for the establishment of a fully mesenchymal state, which is in line with the observed dominance of this phenotype in skin cancers (Figure 1f). Mutations in the mismatch repair gene *PMS1* and the actin cytoskeleton remodelling gene *FNBP1* were subclonally fixed in hEMT cancers, potentially suggesting that acquiring such alterations later during tumour evolution may benefit the establishment of a hybrid phenotype.

To further investigate such associations, we prioritised cancer driver mutations, focal and arm-level copy number changes that may be linked with EMT, and implemented a lasso-based

machine learning framework to identify those drivers able to discriminate between EPI, hEMT and MES states across cancers, while accounting for tissue-specific effects (Methods, Figure 5e). The developed models were validated using several other machine learning approaches and demonstrated remarkably high accuracies of 92-97% in distinguishing the fully transformed state from either the hybrid or mesenchymal one (Supplementary Figures S6a-b,d-e,g-h). Lower performance was obtained for the model discriminating between hEMT and EPI (~62-73%, Supplementary Figures 6c,f,i), which is not surprising due to the intermediate, hybrid nature of the former, but is still useful in understanding weaker effects on EMT transformation.

Among the genomic biomarkers able to discriminate transformed tumours (hEMT, MES) from the epithelial state, we identified genes that have been previously linked with cell migration, invasion and EMT, such as *RB1*, *VHL*, *ERBB2, ARHGAP26*, *PRDM1, APC* (Figures 5f-g , Supplementary Tables S4, S5). *RB1*, a key cell cycle regulator, has been shown to promote EMT in conjunction with p53 in triple negative breast cancer[43], while *VHL* alterations contribute to EMT via regulation of hypoxia[44]. Larger scale events included deletions of the 4p, 6p and 17p chromosomal arms, all of which harboured cancer drivers which have been previously linked with EMT, e.g. *FGFR3* on 4p[45], *DAXX* and *TRIM27* on 6p[46,47], *TP53* on 17p[12] (Supplementary Table S4). Deletions of the 4p arm appeared in the majority of lung squamous cell and esophageal carcinomas (58% and 50%, respectively), while 6p arm deletions were most frequent in pancreatic, esophageal cancers and adrenocortical carcinomas (>20% in each). 17p arm deletions were the most abundant, especially in ovarian (76%) and kidney chromophobe cancers (76%), with an average of 37% of cases affected per tissue. Therefore, no strong bias in terms of cancer type was observed for these large-scale alterations. In addition to these, events less strongly linked with metastatic transformation were also uncovered, such as mutations in *CHIC2*, encoding for a protein with a cysteine-rich hydrophobic domain occasionally implicated in leukemia, or amplifications of the *ELL* gene, an elongation factor for polymerase II. While the genomic hallmarks distinguishing the extremes

of mesenchymal transformation (MES versus EPI) were predominantly classical cancer drivers involved in the most fundamental processes (e.g. cell cycle) (Supplementary Figure 6j), the ones distinguishing fully transformed from hybrid phenotypes were more clearly linked with cell migration, including processes of cytoskeletal regulation, cell adhesion and T cell signalling (Supplementary Figure 6k).

The hEMT state-specific markers were mostly enriched in cell fate commitment and metabolic pathways (Figure 5h, Supplementary Figure S6l). Among the top events distinguishing this phenotype from the epithelial one was the disruption of *EPAS1* (*HIF2A)*, a well-known hypoxia regulator which has been previously implicated in EMT[48]*. SMAD4*, a suppressor of cell proliferation, was clearly linked with the switch between hEMT and EPI, with activating mutations contributing to an hEMT phenotype while deletions were prevalent in epithelial cancers. Indeed, *SMAD4* mutations have been shown to induce invasion and EMT marker upregulation in colorectal cancer[49]. Deletions of *FOXO3*, a gene involved in cell death and implicated in EMT[50], were specifically linked with high levels of aneuploidy, stemness and centrosome amplification.

Validation of genomic associations

To gain further insight into the role of the putative genomic markers proposed by our pan-cancer model on EMT transformation, we validated some of these candidates and their effect on cell migration using several siRNA screens. First, using data from Koedoot et al[51], we found that knocking down 31 of the 61 targets resulted in significant changes in the surface area, perimeter and elongation/roundness of the cells in Hs578T and MDA-MBA-231 breast cancer cell lines, suggesting either an impairment or an enhancement of migratory properties (Figure 6a). *ETV6*, linked to EPI-hEMT transformation in our models, was shown in Koedoot et al to produce a big round cellular phenotype upon knockdown, with effects on cellular migration in line with expectations from the model. Indeed, *ETV6* disruption has been shown to promote TWIST1-dependent tumour progression[52], confirming our observations. Several

other genes also showed significant phenotypic effects upon knockdown, albeit to a lesser extent, and many of them, including *RB1*, *ELL* and *NCKIPSD* (involved in signal transduction) were confirmed in both cell lines. *RB1* also showed a low penetrance EMT microscopy phenotype upon knockdown in an independent transcription factor-focused siRNA screen from Meyer-Schaller et al[53], further confirming it as a mesenchymal marker.

Another gene with effects in the Hs578T cell line, *PRDM1*, a repressor of interferon activity which our model linked with the MES state, was also shown to alter multiple cellular properties associated with migration in an independent screen from Penalosa-Ruiz et al[54] (Figure 6b). In particular, PRDM1 knockdown increased the E-cadherin expression area and intensity, as did SETD2 knockdown. Among other MES-linked candidates from our models, knockdowns of the transcriptional regulators CDC73 and TRIM24 showed weaker phenotypes linked with migration, mostly related to homogeneity of textures observed under the microscope, again potentially related to a less transformed state. Overall, these analyses recapitulate many of the already described markers of EMT transformation, and also suggest that *ELL* and *NCKIPSD* mutations may affect the cancer cell's ability to undergo EMT transformation. Further experimental studies will be needed to clarify the mechanism by which this may occur.

A good fraction of the reported alterations (36%) were also confirmed to be linked with the metastatic potential of cancer cell lines at pan-cancer or tissue specific level (Supplementary Figures 7a-c). Among these *DEK*, a splicing regulator and putative hEMT biomarker, showed a particularly strong correlation. Suppression of several of these genes also strongly impacted cell viability (Supplementary Figure 7d-e), but *RB1*, *DEK, RGPD3*, *MN1*, *LMO1* and *ARHGAP26* were deemed non-essential and thus more likely to be promising targets for EMT manipulation.

Clinical relevance of EMT

Finally, we show that the defined EMT states have potential clinical utility. As expected, patients with a partially or fully transformed phenotype had worse overall survival outcomes

(Figure 7a, Supplementary Table S6). Furthermore, the EMT macro-state progression reflected a step-wise decrease in progression-free intervals (Figure 7b).

Among the driver events that have been linked with EMT in this study, alterations in genes *ERBB2*, *PRDM1*, *FLT4* and *TMPRSS2* associated with a mesenchymal phenotype, and ten other events associated with hEMT (including genome doubling, 3p/8p deletions, *EPAS1*, *NCKIPSD* mutations) were linked with worse prognosis (Figure 7c-d, Supplementary Table S7). Cases with mutations in *FNBP1* and *CHIC2* displayed better prognosis.

To further explore potential links between EMT and therapy responses, we investigated whether EMT progression might confer different levels of sensitivity to individual cancer drugs using cell line data from GDSC[18]. We found 22 compounds whose IC50 values were significantly correlated with the EMT score (Figure 7e). The strongest associations were observed with Sapitinib, an inhibitor of ErbB1/2/3[55], Osimertinib, a lung cancer EGFR inhibitor, and Acetalax, a drug used in the treatment of triple negative breast cancers. These observations reiterate the reported genomic links between events in the tyrosine kinase pathway and EMT transformation.

Finally, we investigated whether EMT transformation may be linked with different treatment outcomes in the clinic. Within TCGA, patients with higher EMT levels in the pre-treatment tumour showed progressively worse outcomes upon oxaliplatin treatment (Figure 7f), with complete responders significantly distinguished from patients with progressive disease. In fact, there was a two-fold enrichment in complete responders among patients with epithelial and hybrid tumours compared to mesenchymal ones (Fisher's exact test p=1.5e-05). We also linked post-treatment EMT phenotypes with therapy responses using the POG570 dataset (Supplementary Figure S8a). The EMT levels increased significantly in samples treated with temozolomide over progressively longer time frames, suggesting this drug may induce EMT transformation in cancer (Supplementary Figure S8b). The opposite effect was observed for rituximab, with tumours becoming more epithelial over the treatment course.

Overall, these analyses suggest that the level of EMT transformation may play a role in determining responses to some chemotherapies as well as targeted therapies. However, our insights into the exact context in which EMT matters is limited by the lack of longitudinal, spatially and microenvironmentally resolved datasets.

DISCUSSION

Previous studies of the EMT process have suggested the existence of a phenotypic continuum characterised by multiple intermediate states[56]. We have shown that distinct EMT trajectories in cancer are underpinned by three macro-states, reflecting both tumour cell intrinsic as well as tumour microenvironment associated changes. The hybrid E/M state, characterised by the co-expression of epithelial and mesenchymal markers, was surprisingly frequent (39%). It is clear that this state is distinct from epithelial tumours, presenting higher CAF infiltration and occasionally enhanced hypoxia and stemness. While it is likely this is an intermediate state in cancer progression along the EMT continuum, as suggested by the longitudinal datasets analysed and the intermediate progression-free intervals, it is also clearly heterogeneous and less genomically influenced than the extreme epithelial and mesenchymal states. Furthermore, the extent to which it is intrinsically rather than environmentally distinct cannot be determined in bulk datasets. It has been reported that cells with hybrid EMT features give rise to daughter cells that are either mesenchymal or epithelial and are more prone to migrate [31], which could explain some of the heterogeneity observed for this state. Undoubtedly, the hEMT state can be further subdivided into sub-states, as shown by Goetz et al[4] and Brown et al[57]. The true number of EMT intermediate states is just beginning to be explored. However, the noisy bulk sequencing data are limiting our ability to capture them, highlighting the need to complement these studies with spatially-resolved and single cell data.

Our spatial transcriptomics and single cell analyses demonstrate a heterogeneous EMT landscape, delineating clear spatial effects of the continuum of EMT transformation within the tissue. Fibroblasts and cytotoxic T cells often surround more mesenchymal neoplastic areas,

and these are occasionally accompanied by hypoxia. While some differential immune recognition is evidenced by co-localisation of MES with CD8/CD4+ T cell signals and hEMT with NK cell signals, partially or fully mesenchymal cells appear to interact less with the microenvironment, potentially due to evasion caused by neoantigen presentation in these more mutated cells[58]. The tumour cells in different EMT states are generally well distinguished from immune cells and the stroma in single cell datasets, with a minority of cells requiring improvement in discrimination methods. While this analysis is limited by our ability to capture a broad spectrum along the EMT transformation as the data are only sourced from early stage cancers, it does lay out a framework for future studies in this space. These should ideally integrate spatial and single cell transcriptomics for a better comprehension of the complex interplay between EMT and the tumour microenvironment.

Our study confirmed previously established molecular hallmarks of EMT, including increased chromosomal instability and hypoxia/stemness in hEMT, and cytotoxicity/exhaustion in mesenchymal tumours[15], along with several genomic dependencies of this process. While the exploration of EMT biomarkers is not new, most of the studies in this area have been reliant on gene expression activity rather than mutational dependencies and they are generally tissue-specific[15,28]. Pan-cancer studies generally consider EMT as a binary switch[14,15,28]. In contrast, our study identified genomic hallmarks of three EMT macro-states, providing further granularity into how genome-driven cancer evolution shapes EMT trajectories in a state-specific manner. Indeed, we show that distinct genes contribute to the establishment of a fully mesenchymal phenotype, e.g. *RB1* or *DEK*, while others such as *EPAS1*, *FNBP1* or *SMAD4* modulate switches between epithelial and hybrid phenotypes. Furthermore, the genomic distinction in the latter case was less strong than between the extremes of EMT transformation, suggesting that transcriptional or epigenetic alterations may play an increased role in the earlier stages of EMT, while genomic events may further promote and help establish a fully transformed phenotype, which was accurately predicted based solely on

genomic alterations. A causal relationship between the acquisition of any of these genomic changes and EMT should be further experimentally tested in the future.

The EMT process was also linked to responses to several targeted therapies as well as some chemotherapy drugs, with an expected reduction in response in more mesenchymal cancers. EMT could thus potentially be exploited for therapeutic benefit in certain contexts.

Overall, the results of this study demonstrate the complex intrinsic and microenvironmental mechanisms that shape the landscape of EMT transformation during cancer. We have not considered the role of chromosomal rearrangements or epigenetic changes in EMT, which could provide further explanations to the maintenance of an hEMT phenotype. Additional research is required to understand the biological role and spatial constraints of the identified biomarkers, their importance in a clinical setting, and to identify additional mechanisms that may promote EMT.

METHODS

Data sources

Bulk RNA-sequencing, copy number (segment file and focal alterations), somatic variants (MuTect[59]), molecular subtypes and clinical data were retrieved for 8,778 primary tumours of epithelial origin from the harmonized version of TCGA using the *TCGAbiolinks* R package[60]. Based on tumour purity estimates reported by Hoadley et al[61] samples with purity lower than 30% were removed leaving 7,180 samples. All other data sources employed for validation are described below.

Reconstruction of EMT trajectories in bulk data

The reconstruction of the EMT trajectory of the TCGA samples was performed using a procedure that allows the mapping of bulk-sequenced samples to single cell-derived expression programmes inspired from McFaline-Figueroa et al[9]. The workflow of the analysis consists of several steps. The first step of the analysis requires two gene expression matrices

as input, namely one bulk sequenced dataset, for which the EMT trajectory is to be determined, and one single cell reference dataset, for which the associated trajectory (P) of individual cells is known. In the first step of the analysis the matrices were merged; then, in order to remove the batch effects originated by the two different platforms, a correction was applied using ComBat[62]. In the second step, principal component analysis (PCA) was performed on the merged matrix. The single-cell derived EMT trajectory was then mapped onto the bulk data using an iterative process and a mapping strategy based on k nearest neighbours (kNN). The number of iterations (i) is equal to the number of bulk samples. During each i-th step of iteration, a single bulk-sequenced sample and the reference scRNA-seq data were used as input for the kNN algorithm. The procedure computed the mean of the pseudotime values of the single cell samples that have been detected by the kNN algorithm to be associated with the i-th bulk sample. The implementation of the kNN algorithm is based on get.knnx() function from the *FNN* R package. In our case, we used as input the bulk RNA-seq data from TCGA samples. scRNA-seq datasets from McFaline-Figueroa et al[9], as well as, Cook et al[63] were used as references. Overall, 10 different scRNA-seq datasets were used including A549, MCF7, DU145 and OVCA420 cell lines treated with TGFB1 or TNF. A spontaneous, as well as TGFB1 driven EMT model in MCF10 cell lines was also used. The procedure described above was repeated with with each of the 10 reference datasets as input along with the TCGA bulk expression data. This resulted in 10 separate pseudotime estimates for each TCGA bulk-sequenced sample, one based each one of the reference single-cell datasets. The average of the 10 pseudotimes was used to obtain the final pseudotime estimate. Because samples are projected individually along the consensus reference single cell data points, the pseudotime estimate only depends on the reference used and not on the specific cohort the sample is part of. Thus, the pseudotime estimates are cohort-independent.

Segmentation of the EMT trajectory and robustness evaluation

We used a Hidden Markov Model approach to identify of a discrete number of EMT states. The input of this analysis was a matrix (M) where the rows were the TCGA samples (N) and

the columns the gene markers (G) of EMT (see the section "Computation of the EMT scores" below for the list of genes). The original N columns were sorted for the t values of the pseudotime (P). This matrix and P were provided as input for a lasso penalized regression. P was used as response variable, the genes as the independent variables. The non-zero coefficients obtained from this analysis were selected to create a sub-matrix of M that was used as input for a Hidden Markov Model.

Different HMM models were tested while changing the number of states. After this tuning, and through manual inspection, we determined that 3 states were most in line with biological expectations. Each HMM state was assigned to a "biological group" (i.e. epithelial, hybrid EMT, mesenchymal) by exploring the expression levels of known epithelial and mesenchymal markers in each HMM state. The selection of the coefficients was performed with the R package *glmnet*. The identification of the EMT states was done using the *deepmixS4* R package.

To evaluate the "robustness" of the EMT states we applied the same procedure described above while increasing levels of expression noise in the original dataset. We used the *jitter* function in R to introduce a random amount of noise to the expression values of the genes (from the default parameter of the *jitter* function to noise levels of 5500). For each noise level, we repeated the analysis 100 times. We considered several metrics to measure the stability of the HMM-derived EMT states. We reasoned that increasing noise could result in classification mismatches of the samples compared to their originally assigned EMT state. Therefore, we evaluated two metrics to assess the correct assignment of the samples to the original EMT states. Firstly, for each level of noise added and at each iteration, we computed the change in number of samples categorised in the new states compared to the original EMT states. Second, we measured the assignment accuracy for the samples to the original EMT states.

EMT pseudotime reconstruction with adjustment for TME contamination

To account for confounding expression signals coming from non-tumour cells in the microenvironment, we regressed the expression data on the tumour purity estimates obtained

562 from matched DNA-sequencing using the *MOFA* R package. The purity-adjusted expression
563 values were used as bulk input to the PCA projection for the pseudotime reconstruction.

564 **Computation of the EMT scores**

565 A list of epithelial and mesenchymal markers was compiled through manual curation of the
566 literature[6,9,28] , as follows:

567 • epithelial genes: *CDH1*, *DSP*, *OCLN*, *CRB3*
568 • mesenchymal genes: *VIM*, *CDH2*, *FOXC2*, *SNAI1*, *SNAI2*, *TWIST1*, *FN1*, *ITGB6*,
569 *MMP2*, *MMP3*, *MMP9*, *SOX10*, *GSC*, *ZEB1*, *ZEB2*, *TWIST2*

570 EMT scores for each TCGA sample were computed in a similar manner as described by Chae
571 et al[64]. Briefly, the average z-score transformed expression levels of the mesenchymal
572 markers were subtracted from the average z-score transformed expression levels of the
573 epithelial markers. To segment the EMT trajectory, along with the epithelial and mesenchymal
574 markers we have also considered markers of hybrid EMT[6,65]: *PDPN*, *ITGA5*, *ITGA6*, *TGFBI*,
575 *LAMC2*, *MMP10*, *LAMA3*, *CDH13*, *SERPINE1*, *P4HA2*, *TNC*, *MMP1*.

576 **Tissue-specific EMT trajectory derivation**

577 Using a similar bulk-to-single cell mapping approach as described before, we mapped the
578 RNA-seq data of BRCA, LUAD and PRAD tumours onto the trajectories derived from the
579 single cell data (including batch effect removal using ComBat, PCA on 25 dimensions and
580 kNN clustering). For the BRCA tumours, the final pseudotime estimates were averaged using
581 values calculated from the MCF10 and MCF7 scRNA-seq reference datasets only. Similarly,
582 for LUAD and PRAD bulk-sequenced samples only scRNA-seq references from A549 and
583 DU145 cell lines were used respectively.

584 **Longitudinal datasets of EMT transformation**

585 Longitudinal datasets for the validation of the EMT reconstruction method were obtained from
586 the Gene Expression Omnibus (GEO) database as follows: GSE17708, a time course
587 experiment of A549 lung adenocarcinoma lines treated with TGF-beta; GSE84135, a time

course EMT transition experiment in hSAEC airway epithelial cells; and GSE75487, a 7 day EMT transformation experiment in H358 non-small cell lung cancer cells under doxycycline treatment to induce Zeb1. EMT pseutime inference in these datasets was performed as described above.

EMT trajectory reconstruction of CCLE data and inference of the metastatic potential

The RSEM gene-expression values of the Cancer Cell Line Encyclopedia[66] project were retrieved from the CCLE Data Portal. We used the same procedure described above to map the CCLE data onto the 10 reference single-cell dataset EMT trajectories. This allowed for the pseudotime to be quantified for each CCLE sample. A segmentation using a HMM model was performed to identify a discrete number of EMT states (n=3). The EMT scores were also computed for each cell line. These results were referenced against the metastatic potential scores from MetMap500[17]. The association between HMM states and experimentally measured metastatic potential groups in cell lines (non-metastatic, weakly metastatic and metastatic) was assessed using the *vcd* R package.

Tumour microenvironment quantification

The tumour purity values of TCGA samples were retrieved from Hoadley et al[61]. Immune deconvolution was performed using the ConsensusTME R package[67] and the ssGSEA method for cell enrichment analysis.

The results of ConsesusTME were used as input for a multinomial logistic regression model. The function multinom() (from the *nnet* R package) was used to determine the probability of each sample belonging to a macro-EMT state based on the cellular content of the sample.

Spatial transcriptomics data analysis

Three breast cancer patient samples were downloaded from 10x genomics (https://support.10xgenomics.com/spatial-gene-expression/datasets). Patient 1 was AJCC Stage Group I, ER positive, PR positive and HER2 negative. Patient 2 was AJCC Stage Group IIA, ER positive, PR negative and Her2 positive. Patient 3 did not have molecular

details described. The output from the Space Ranger Visium pipeline was used for analysis. The *SCTransform* R package was used to normalise the data based on a regularised negative binomial regression method. Cell type and state proportions for each spot were estimated using EcoTyper[68] which was run using Docker. The cell types consisted of B cells, CD4 T cells, CD8 T cells, dendritic cells, endothelial cells, epithelial cells, fibroblasts, mast cells, monocytes/macrophages, NK cells, plasma cells and neutrophils.

ST2K (ST second generation, 2000 spots/array) datasets (9 patients with 3-5 repeats each) were downloaded from https://github.com/almaan/her2st. All samples were stained positive for HER2. Pre-processing steps were followed as described by the authors[33]. Briefly, this consisted of using *SCTransform* for normalisation and Non-Negative Matrix Factorisation (NMF) for dimensionality reduction. The factors that contained consistent patterns across the tissue replicates were kept for analysis. The *Stereoscope*[69] (v.0.2) R package was used for cell type deconvolution. The deconvolution data was downloaded from https://github.com/almaan/her2st. The major class consists of myeloid cells, T cells, B cells, epithelial cells, plasma cells, endothelial cells, cancer associated fibroblasts (CAFs), and perivascular-like cells (PVL cells). The minor tier contains finer partitioning of the major cell types, e.g., macrophages and CD8+ T cells. Further description of the deconvolution method is described by the authors[33].

The *Seurat*[70] R package was used for storing, manipulating and visualising the spatial transcriptomic data.

Gene module scores

An EMT score was calculated per spot by adapting the method used to assign a score to the TCGA samples, using only the breast cancer cell line scRNA-seq data. The EMT scRNA-seq trajectory was mapped onto each spot within the spatial transcriptomic slide, and the mean of the pseudotime values of the single cell samples detected by the kNN algorithm was used. This was performed on multiple breast cancer cell lines and the average pseudotime across the cell lines was used to calculate the EMT score. The pseudotime was split into three

intervals to define an epithelial-like, hybrid-like and mesenchymal-like state. The *AddModuleScore* Seurat R function, originally developed for single cell enrichment, was used to calculate the hypoxia gene set (Buffa et al[71]) score across the slides. It calculates the average expression levels for the gene set and subtracts the expression of a sample control feature set (here set to 200 controls/spot). The *SpatialFeaturePlot* Seurat R function was used to visualise the scores. Correlations for the EMT scores were calculated by filtering for the spots containing epithelial cells and using the *STUtility*[72] R package to calculate the 12 nearest neighbours for each epithelial spot. The proportions of cells within each spot were summed across the neighbours.

Cluster identification from spatial transcriptomics data

The spatial transcriptomic data was subsetted to include solely the epithelial, hybrid and mesenchymal genes. The *FindClusters* Seurat R package then clustered the gene expression data and assigned a cluster value to each barcode spot. This identified clusters by calculating the k-nearest neighbours (k-NN) and constructing a shared nearest neighbour graph. The EMT scores were averaged across the clusters. The results for each cluster were then binned so that 'low', 'medium' and 'high' groups (corresponding to EPI, hEMT, MES) were created. The cell type enrichment scores calculated per region were plotted using the enriched-region.py Python file from https://github.com/almaan/her2st.

Single cell data processing and analysis

Matched bulk and single-cell RNA-sequencing data from breast tumours described in Chung et al[36] were retrieved from the Gene Expression Omnibus using the GSE75688 accession code. Single cell sequencing data from breast, lung, colorectal and ovarian tumours as described by Qian et al[37] were obtained from an interactive web server provided by the authors (http://blueprint.lambrechtslab.org/). Quality control analysis and normalisation of the raw gene expression matrices provided by Qian et al[37] was performed using the *Seurat* R package[73]. Matrices were filtered by removing cells with < 200 and > 6000 expressed genes, as well as cells with > 15% of reads mapping to mitochondrial RNA. EMT pseudotime

estimates were calculated for tumour cells only as described previously using the scRNA-seq data references from McFaline-Figueroa et al[9] as well as Cook et al[63]. For each dataset, the cells were sorted according to their pseudotime and split into 3 equally-sized groups with low, medium or high mean pseudotime estimates, corresponding to EPI, hEMT and MES states. Cell-cell interaction analysis was performed using CellPhoneDB[74] using the normalised gene expression matrices as input, along with cell type and tumour cell pseudotime group annotation.

Genomic hallmark quantification

To characterize the aneuploidy and the centromeric amplification levels of the samples in each EMT state we used the pre-computed values for TCGA from previous works[40,75]. Copy number alterations and clonality estimates based on PhyloWGS were obtained from Raynaud et al[71]. The hypoxia levels were quantified as described by Bandhari et al[76]. Several hypoxia gene signatures were considered, yielding similar results. Only the results obtained using the genes from Buffa et al[71] were reported. The validation of hypoxia associations with EMT was performed using the spatial transcriptomics dataset from Berglund et al[34] and and Affymetrix profiled dataset GSE166211[35] downloaded from the GEO database using *GEOquery*. The EMT levels in these datasets were quantified via expression Z-scores using the GSVA package[77].

Finally, to estimate the levels of stemness in each EMT state, we considered a catalogue of stemness gene sets[78] and used them as input for gene set enrichment analysis via the *GSVA* R package.

Mutational signature analysis

The identification of the mutational spectrum of the samples in each EMT state was performed using a custom approach based on SigProfilerExtractor[41] and deconstructSigs[79]. SigProfilerExtractor was used for a de-novo identification of the mutational signatures. We selected the solutions in which the minimal stability was greater than 0.4 and the sum of the

minimal stabilities across signatures was greater than 1. The cosine similarity with mutational signatures catalogued in the COSMIC database was computed, and only the solutions with non-redundant signatures were selected. Next, we independently ran deconstructSigs. To ensure consistency with Alexandrov et al[41], we evaluated the presence of the ageing-linked SBS1 and SBS5, which have been identified in all cancers. We employed the following steps to obtain a final list of signatures and their exposures for each tissue individually:

(1) Considering the results obtained from deconstructSigs, the signatures with average contribution (across all samples) greater than 5% were taken forward in the analysis.
(2) We combined the signatures obtained in (1) and by SigProfiler to obtain a final list of signatures for the given tissue. If SBS1 and SB5 were not present, we added these signatures manually.

To identify EMT-associated mutational processes we used a similar approach to the one described in Bhandari et al[76], based on linear mixed-effect models. Cancer type was incorporated as a random effect in each model. An FDR adjustment was applied to the p-values obtained from the analysis. The full model for a specific signature (SBS) is as follows:

$$EMT\_score \sim SBS + (1|cancer)$$

Prioritisation of genomic alterations in TCGA

Single nucleotide variants were obtained from TCGA using the *TCGAbiolinks* R package and the Mutect pipeline. Cancer driver events harbouring nonsynonymous mutations were selected for further analysis. To identify putative driver events that are positively selected in association with an EMT state, we employed dNdScv[80], which quantified the ratio of non-synonymous and synonymous mutations (dN/dS) in each gene and state, by tissue. All the somatic driver events with a q-value less than of 0.10 were considered for downstream analysis.

Copy number events were obtained using the *TCGAbiolinks* R package. Chromosomal arm-level data were obtained from Taylor et al[75].

720 **Identification of genomic events linked with EMT**

721 To search for genomic events linked with the described EMT macro-states, we considered all
722 somatic mutations, focal and arm-level copy number events in driver genes from the COSMIC
723 database that were obtained in the previous steps. Two parallel methodological approaches
724 based on lasso and random forest were used to identify events that could be predictive of
725 EMT transitions in a two-step process. First, feature selection was performed using a stability
726 selection approach. We used the function createDataPartition() from the *caret* R package to
727 generate an ensemble of vectors representing 1,00 randomly sampled training models. This is
728 an iterative approach, in which at each iteration a lasso analysis is performed, and the non-
729 negative coefficients computed by lasso are saved. This step was performed using the
730 cv.glmnet() function from *glmnet*. The tissue source was included as potential confounder in
731 the lasso model. The models were trained on 80% of the data. At the end of this stage, the
732 variables that were selected in at least the 80% of the iterations were taken forward and
733 employed as predictors. Features selected in at least 50% of the iterations were also
734 considered for downstream validation. A similar approach was employed for feature selection
735 and model building with random forest.

736 In the second step, ROC curves were generated on the test dataset (20% of the data). In
737 addition, the predictors obtained from the two pipelines were also used as input for random
738 forest (*ranger* implementation), gradient boosting (*gbm*) and Naïve Bayes models. In these
739 cases, the trainControl() function (from the *caret* R package) was used in a 5-fold cross-
740 validation repeated 10 times. The function evalm() (from *MLeval* R package) was used to
741 compare the different machine learning methods. Only the features selected via the lasso
742 procedure were carried forward for downstream analysis.

743 **Cancer cell fraction estimates**

744 The cancer cell fraction (CCF) of selected mutations was calculated using the following
745 formula:

$$CCF_i = \left(2 + \frac{[purity * (CN_i - 2)]}{purity}\right) \cdot VAF_i ,$$

where CN_i stands for the absolute copy number of the segment spanning mutation *i* and VAF_i is the variant allele frequency of the respective mutation. The purities of the TCGA samples were obtained from Hoadley et al[61].

Validation of genomic events linked with EMT

Three large scale public siRNA screens were employed for experimental validation of the proposed genomics associations with EMT. The first dataset from Koedoot et al[51] looked at gene knockdown effects on cell migration abilities in the Hs578T (top panel) and MDA-MB-231 breast cancer cell lines. The data were obtained from the associated publication and contained detailed measurements of effects on cellular phenotype upon knockdown, quantified as changes in cell net surface area, length of minor and major axes, axis ratio (large/small: elongated cells, close to 1: round cells), perimeter score (larger – more migration). Data from further phenotypic tests containing confirmed morphology (big/small round cells) were also available on a subset of the genes.

The second siRNA screen from Penalosa-Ruiz et al[54] quantified migration-related cell integrity in mouse embryonic fibroblasts through a variety of microscopic measurements of the cells upon gene knockdown across multiple replicates. The data were obtained from the corresponding publication.

The third screen from Meyer-Schaller et al[53] focused on the effect of transcription factor knockdown on cell migration in normal murine mammary gland epithelial cells.

To understand relevance of the hypothesised biomarkers to the metastatic dissemination of various cancer cell lines, we downloaded the experimentally measured metastatic potential levels for cancer cell lines from MetMap[19]. We compared metastatic potential between samples with and without a specific EMT marker event (mutations or copy number alterations), pan-cancer and by tissue. Only the markers that were linked with the hEMT or

MES states and that showed a statistically significant difference ($p < 0.05$) in metastatic potential between the two groups (with and without alteration) have been considered.

The viability of the cancer cell lines harbouring putative EMT biomarkers was evaluated based on CRISPR screening data[81] conducted on 990 cell lines. CERES scores denoting gene essentiality were downloaded from Project Achilles. Negative values of these scores indicate that the depletion of a gene influences negatively the viability of a cell line. We only considered genomic markers linked with the hEMT and MES states from our analysis and assessed CERES scores for individual genes both pan-cancer and at tissue level.

Drug response datasets

Cell line drug sensitivity data was obtained from the Genomics of Drug Sensitivity in Cancer database (GDSC)[22]. The treatment information for TCGA cancers was retrieved using the TCGA biolinks package. The POG570[23] dataset was used to study the relation between the EMT states and the duration and effects of given cancer treatments. The EMT states in this dataset were inferred similarly as described above using the kNN approach.

Gene ontology analysis

The characterization of the biological processes associated with the reported lists of genes was performed using the R package *pathfindR*[82].

Survival analysis

Standardized clinical information for the TCGA cohort was obtained from Liu et al[83]. Cox proportional hazard models were used to model survival based on variables of interest and to adjust for the following potential confounders: tumour stage, age at diagnosis, gender and body mass index (BMI). Patients in clinical stages I-II were denoted as having “early stage tumours”, while stages III-IV corresponded to “late stage tumours”. The R packages *survival*, *survminer* and *ggforest* were used for data analysis and visualization.

Data visualization and basic statistics

Graphs were generated using the *ggplot2*, *ggpubr* and *diagram* R packages. Groups were compared using the Student's t test, Wilcoxon rank-sum test or ANOVA, as appropriate.

Data availability

The results published here are based upon publicly available data generated by the TCGA Research Network (https://www.cancer.gov/tcga), MET500 (https://met500.path.med.umich.edu/), MetMap (https://depmap.org/metmap/), GDSC (https://www.cancerrxgene.org/) and POG570 (https://www.bcgsc.ca/downloads/POG570/). All data comply with ethical regulations, with approval and informed consent for collection and sharing already obtained by the relevant consortia.

Code availability

All code developed for the purpose of this analysis can be found at the following repository: https://github.com/secrierlab/EMT/tree/EMTquant.v1.1 .

ACKNOWLEDGEMENTS

GMT was supported by a Wellcome Trust Seed Award in Science (215296/Z/19/Z). AJW was supported by an MRC DTP grant (MR/N013867/1). EW acknowledges the receipt of a studentship award from the Health Data Research UK-The Alan Turing Institute Wellcome PhD Programme in Health Data Science (Grant Ref: 218529/Z/19/Z). MS was supported by a UKRI Future Leaders Fellowship (MR/T042184/1) and an Academy of Medical Science Springboard award (SBF004\1042).

The results published here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

AUTHOR CONTRIBUTIONS

MS designed the study, supervised the analyses and performed the validation of EMT genomic markers using public siRNA screens. GMT and AJW conducted the EMT reconstruction in bulk data, clinical and drug response correlations. GMT and MS correlated EMT states with genomic markers, tumour intrinsic and extrinsic features. AJW performed the single cell analysis. EW performed the analysis of the spatial transcriptomics data. MS and GMT wrote the manuscript. All authors read and approved the manuscript.

COMPETING INTEREST STATEMENT

None declared.

REFERENCES

1. Thiery, J.P., Acloque, H., Huang, R.Y. & Nieto, M.A. Epithelial-mesenchymal transitions in development and disease. *Cell* **139**, 871-90 (2009).
2. Kalluri, R. & Weinberg, R.A. The basics of epithelial-mesenchymal transition. *J Clin Invest* **119**, 1420-8 (2009).
3. Pastushenko, I. & Blanpain, C. EMT Transition States during Tumor Progression and Metastasis. *Trends Cell Biol* **29**, 212-226 (2019).
4. Goetz, H., Melendez-Alvarez, J.R., Chen, L. & Tian, X.J. A plausible accelerating function of intermediate states in cancer metastasis. *PLoS Comput Biol* **16**, e1007682 (2020).
5. Pastushenko, I. *et al.* Identification of the tumour transition states occurring during EMT. *Nature* **556**, 463-468 (2018).
6. Puram, S.V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).
7. Karacosta, L.G. *et al.* Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution. *Nat Commun* **10**, 5587 (2019).
8. Stemmler, M.P., Eccles, R.L., Brabletz, S. & Brabletz, T. Non-redundant functions of EMT transcription factors. *Nat Cell Biol* **21**, 102-112 (2019).
9. McFaline-Figueroa, J.L. *et al.* A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nat Genet* **51**, 1389-1398 (2019).
10. Rhim, A.D. *et al.* EMT and dissemination precede pancreatic tumor formation. *Cell* **148**, 349-61 (2012).
11. Nie, Z. *et al.* STAG2 loss-of-function mutation induces PD-L1 expression in U2OS cells. *Ann Transl Med* **7**, 127 (2019).
12. Chang, C.J. *et al.* p53 regulates epithelial-mesenchymal transition and stem cell properties through modulating miRNAs. *Nat Cell Biol* **13**, 317-23 (2011).

13. Vasudevan, A. *et al.* Single-Chromosomal Gains Can Function as Metastasis Suppressors and Promoters in Colon Cancer. *Dev Cell* **52**, 413-428.e6 (2020).
14. Zhao, M., Liu, Y. & Qu, H. Expression of epithelial-mesenchymal transition-related genes increases with copy number in multiple cancer types. *Oncotarget* **7**, 24688-99 (2016).
15. Mak, M.P. *et al.* A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. *Clin Cancer Res* **22**, 609-20 (2016).
16. Robinson, D.R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297-303 (2017).
17. Jin, X. *et al.* A metastasis map of human cancer cell lines. *Nature* **588**, 331-336 (2020).
18. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754 (2016).
19. Rheinbay, E. The genomic landscape of advanced cancer. *Nat Cancer* **1**, 372-373 (2020).
20. Tyler, M. & Tirosh, I. Decoupling epithelial-mesenchymal transitions from stromal profiles by integrative expression analysis. *Nat Commun* **12**, 2592 (2021).
21. Aiello, N.M. *et al.* EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration. *Dev Cell* **45**, 681-695.e4 (2018).
22. Jolly, M.K. *et al.* Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Front Oncol* **5**, 155 (2015).
23. Plygawko, A.T., Kan, S. & Campbell, K. Epithelial-mesenchymal plasticity: emerging parallels between tissue morphogenesis and cancer metastasis. *Philos Trans R Soc Lond B Biol Sci* **375**, 20200087 (2020).
24. Wang, Q., Liu, G. & Hu, C. Molecular Classification of Gastric Adenocarcinoma. *Gastroenterology Res* **12**, 275-282 (2019).
25. Sabe, H. Cancer early dissemination: cancerous epithelial-mesenchymal transdifferentiation and transforming growth factor β signalling. *J Biochem* **149**, 633-9 (2011).
26. Jolly, M.K., Ware, K.E., Gilja, S., Somarelli, J.A. & Levine, H. EMT and MET: necessary or permissive for metastasis? *Mol Oncol* **11**, 755-769 (2017).
27. Jing, Y., Han, Z., Zhang, S., Liu, Y. & Wei, L. Epithelial-Mesenchymal Transition in tumor microenvironment. *Cell Biosci* **1**, 29 (2011).
28. Gibbons, D.L. & Creighton, C.J. Pan-cancer survey of epithelial-mesenchymal transition markers across the Cancer Genome Atlas. *Dev Dyn* **247**, 555-564 (2018).
29. Choi, B.J., Park, S.A., Lee, S.Y., Cha, Y.N. & Surh, Y.J. Hypoxia induces epithelial-mesenchymal transition in colorectal cancer cells through ubiquitin-specific protease 47-mediated stabilization of Snail: A potential role of Sox9. *Sci Rep* **7**, 15918 (2017).
30. Emami Nejad, A. *et al.* The role of hypoxia in the tumor microenvironment and development of cancer stem cell: a novel approach to developing treatment. *Cancer Cell Int* **21**, 62 (2021).
31. San Juan, B.P., Garcia-Leon, M.J., Rangel, L., Goetz, J.G. & Chaffer, C.L. The Complexities of Metastasis. *Cancers (Basel)* **11**(2019).
32. Mani, S.A. *et al.* The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **133**, 704-15 (2008).
33. Andersson, A. *et al.* Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat Commun* **12**, 6012 (2021).

34. Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat Commun* **9**, 2419 (2018).
35. Ardila, D.C. *et al.* Identifying Molecular Signatures of Distinct Modes of Collective Migration in Response to the Microenvironment Using Three-Dimensional Breast Cancer Models. *Cancers (Basel)* **13**(2021).
36. Chung, W. *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun* **8**, 15081 (2017).
37. Qian, J. *et al.* A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Res* **30**, 745-762 (2020).
38. Ganem, N.J., Godinho, S.A. & Pellman, D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* **460**, 278-82 (2009).
39. Lingle, W.L. *et al.* Centrosome amplification drives chromosomal instability in breast tumor development. *Proc Natl Acad Sci U S A* **99**, 1978-83 (2002).
40. de Almeida, B.P., Vieira, A.F., Paredes, J., Bettencourt-Dias, M. & Barbosa-Morais, N.L. Pan-cancer association of a centrosome amplification gene expression signature with genomic alterations and clinical outcome. *PLoS Comput Biol* **15**, e1006832 (2019).
41. Alexandrov, L.B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
42. Muñoz, D.P. *et al.* Activation-induced cytidine deaminase (AID) is necessary for the epithelial-mesenchymal transition in mammary epithelial cells. *Proc Natl Acad Sci U S A* **110**, E2977-86 (2013).
43. Jiang, Z. *et al.* RB1 and p53 at the crossroad of EMT and triple-negative breast cancer. *Cell Cycle* **10**, 1563-70 (2011).
44. Liu, W., Xin, H., Eckert, D.T., Brown, J.A. & Gnarra, J.R. Hypoxia and cell cycle regulation of the von Hippel-Lindau tumor suppressor. *Oncogene* **30**, 21-31 (2011).
45. Li, L., Zhang, S., Li, H. & Chou, H. FGFR3 promotes the growth and malignancy of melanoma by influencing EMT and the phosphorylation of ERK, AKT, and EGFR. *BMC Cancer* **19**, 963 (2019).
46. Wu, C. *et al.* DAXX inhibits cancer stemness and epithelial-mesenchymal transition in gastric cancer. *Br J Cancer* **122**, 1477-1485 (2020).
47. Zhang, Y. *et al.* TRIM27 functions as an oncogene by activating epithelial-mesenchymal transition and p-AKT in colorectal cancer. *Int J Oncol* **53**, 620-632 (2018).
48. Yang, W. *et al.* MiR-182-5p promotes the Metastasis and Epithelial-mesenchymal Transition in Non-small Cell Lung Cancer by Targeting EPAS1. *J Cancer* **12**, 7120-7129 (2021).
49. Frey, P. *et al.* SMAD4 mutations do not preclude epithelial–mesenchymal transition in colorectal cancer. *Oncogene* **41**, 824-837 (2022).
50. Li, J. *et al.* Knockdown of FOXO3a induces epithelial-mesenchymal transition and promotes metastasis of pancreatic ductal adenocarcinoma by activation of the β-catenin/TCF4 pathway through SPRY2. *Journal of Experimental & Clinical Cancer Research* **38**, 38 (2019).
51. Koedoot, E. *et al.* Uncovering the signaling landscape controlling breast cancer cell migration identifies novel metastasis driver genes. *Nat Commun* **10**, 2983 (2019).
52. Tsai, Y.C. *et al.* Disruption of ETV6 leads to TWIST1-dependent progression and resistance to epidermal growth factor receptor tyrosine kinase inhibitors in prostate cancer. *Mol Cancer* **17**, 42 (2018).

53. Meyer-Schaller, N. *et al.* A Hierarchical Regulatory Landscape during the Multiple Stages of EMT. *Dev Cell* **48**, 539-553.e6 (2019).
54. Peñalosa-Ruiz, G. *et al.* WDR5, BRCA1, and BARD1 Co-regulate the DNA Damage Response and Modulate the Mesenchymal-to-Epithelial Transition during Early Reprogramming. *Stem Cell Reports* **12**, 743-756 (2019).
55. Mu, Z. *et al.* AZD8931, an equipotent, reversible inhibitor of signaling by epidermal growth factor receptor (EGFR), HER2, and HER3: preclinical activity in HER2 non-amplified inflammatory breast cancer models. *J Exp Clin Cancer Res* **33**, 47 (2014).
56. Roche, J. The Epithelial-to-Mesenchymal Transition in Cancer. *Cancers (Basel)* **10**(2018).
57. Brown, M.S. *et al.* Dynamic plasticity within the EMT spectrum, rather than static mesenchymal traits, drives tumor heterogeneity and metastatic progression of breast cancers. *bioRxiv*, 2021.03.17.434993 (2022).
58. Thelen, M. *et al.* Cancer-specific immune evasion and substantial heterogeneity within cancer types provide evidence for personalized immunotherapy. *NPJ Precis Oncol* **5**, 52 (2021).
59. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-9 (2013).
60. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**, e71 (2016).
61. Hoadley, K.A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).
62. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).
63. Cook, D.P. & Vanderhyden, B.C. Context specificity of the EMT transcriptional response. *Nat Commun* **11**, 2142 (2020).
64. Chae, Y.K. *et al.* Epithelial-mesenchymal transition (EMT) signature is inversely associated with T-cell infiltration in non-small cell lung cancer (NSCLC). *Sci Rep* **8**, 2918 (2018).
65. Puram, S.V., Parikh, A.S. & Tirosh, I. Single cell RNA-seq highlights a role for a partial EMT in head and neck cancer. *Mol Cell Oncol* **5**, e1448244 (2018).
66. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-7 (2012).
67. Jiménez-Sánchez, A., Cast, O. & Miller, M.L. Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods. *Cancer Res* **79**, 6238-6246 (2019).
68. Rolong, A., Chen, B. & Lau, K.S. Deciphering the cancer microenvironment from bulk data with EcoTyper. *Cell* **184**, 5306-5308 (2021).
69. Andersson, A. *et al.* Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol* **3**, 565 (2020).
70. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
71. Buffa, F.M., Harris, A.L., West, C.M. & Miller, C.J. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br J Cancer* **102**, 428-35 (2010).

72. Bergenstråhle, J., Larsson, L. & Lundeberg, J. Seamless integration of image and molecular analysis for spatial transcriptomics workflows. *BMC Genomics* **21**, 482 (2020).
73. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
74. Efremova, M., Vento-Tormo, M., Teichmann, S.A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc* **15**, 1484-1506 (2020).
75. Taylor, A.M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676-689.e3 (2018).
76. Bhandari, V. *et al.* Molecular landmarks of tumor hypoxia across cancer types. *Nat Genet* **51**, 308-318 (2019).
77. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
78. Miranda, A. *et al.* Cancer stemness, intratumoral heterogeneity, and immune response across cancers. *Proc Natl Acad Sci U S A* **116**, 9020-9029 (2019).
79. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* **17**, 31 (2016).
80. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
81. Dempster, J.M. *et al.* Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines. *bioRxiv*, 720243 (2019).
82. Ulgen, E., Ozisik, O. & Sezerman, O.U. pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks. *Front Genet* **10**, 858 (2019).
83. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).

FIGURES

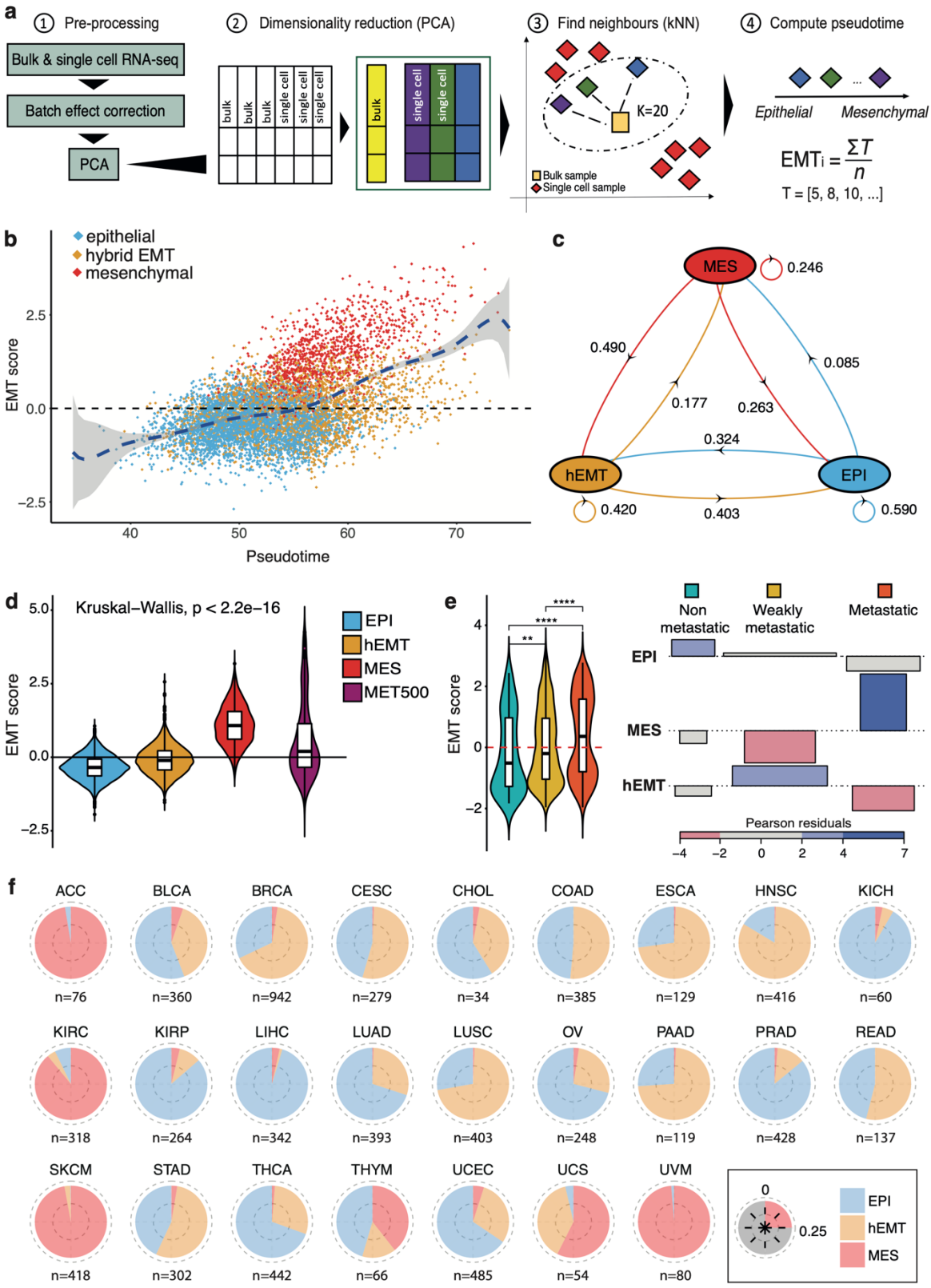
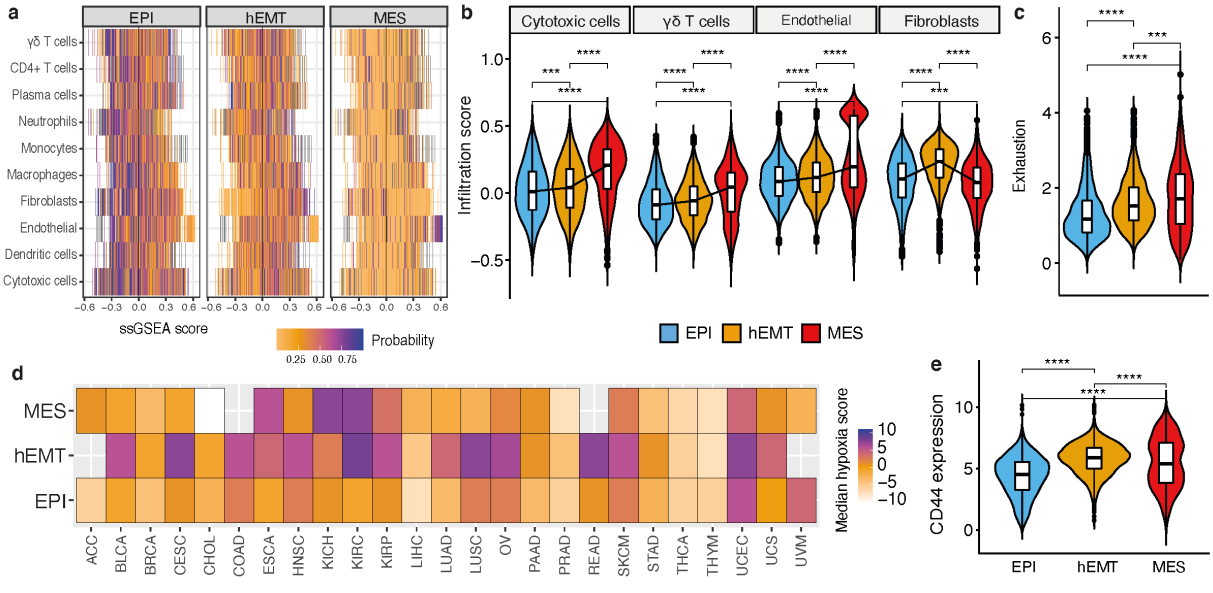


Figure 1. Pan-cancer EMT trajectories and underlying macro-states. (a) Workflow for reconstructing the EMT trajectories of TCGA samples. 1: Bulk and single cell datasets are

combined and processed together to remove batch effects. 2: Dimensionality reduction using PCA is performed. 3: A k-nearest neighbours (kNN) algorithm is used to map bulk RNA-sequencing onto a reference EMT trajectory derived from scRNA-seq data. 4: Tumours are sorted on the basis of their mesenchymal potential along an EMT "pseudotime" axis. (b) Scatter plot of EMT scores along the pseudotime. Each dot corresponds to one bulk tumour sample from TCGA. Samples are coloured according to the designated state by the HMM model. (c) Diagram of the transition probabilities for switching from one EMT state to another, as estimated by the HMM model. MES: fully mesenchymal state, hEMT: hybrid E/M, EPI: epithelial state. (d) EMT scores compared across epithelial, hEMT, mesenchymal TCGA samples, and the MET500 cohort. (e) Left: EMT scores compared between cell lines from CCLE classified as "non metastatic" (aqua green), "weakly metastatic" (orange)," metastatic" (red) according to the MetMap500 study. **p<0.01; ****p<0.0001. Right: Association plot between the HMM-derived cell line states (rows) and their experimentally measured metastatic potential (columns) (p=2.2e-16). (f) Distribution of the EMT states across different cancer tissues. Each quarter of the pie corresponds to the 25% of the data. The number of samples analysed is indicated for each tissue.



1047

1048 **Figure 2. Tumour extrinsic and intrinsic hallmarks of EMT.** (a) Heat map showcasing the
1049 results of a multinomial logistic regression model trained to predict EMT states based on cell
1050 infiltration in the microenvironment. Each row corresponds to a cell type and the
1051 corresponding per-sample infiltration is highlighted via ssGSEA scores reported on the x axis.
1052 The values reported in the heat map are the probabilities that a sample should fall into the
1053 epithelial, hEMT or mesenchymal categories in relation to the ssGSEA score of a certain cell
1054 type. (b) Cell abundance compared across the EMT states for selected cell types. (c) Levels
1055 of exhaustion quantified across the three EMT states. (d) Median hypoxia values in the three
1056 different EMT states across tissues. (e) Gene expression levels of the stemness marker CD44
1057 compared across the three EMT states.

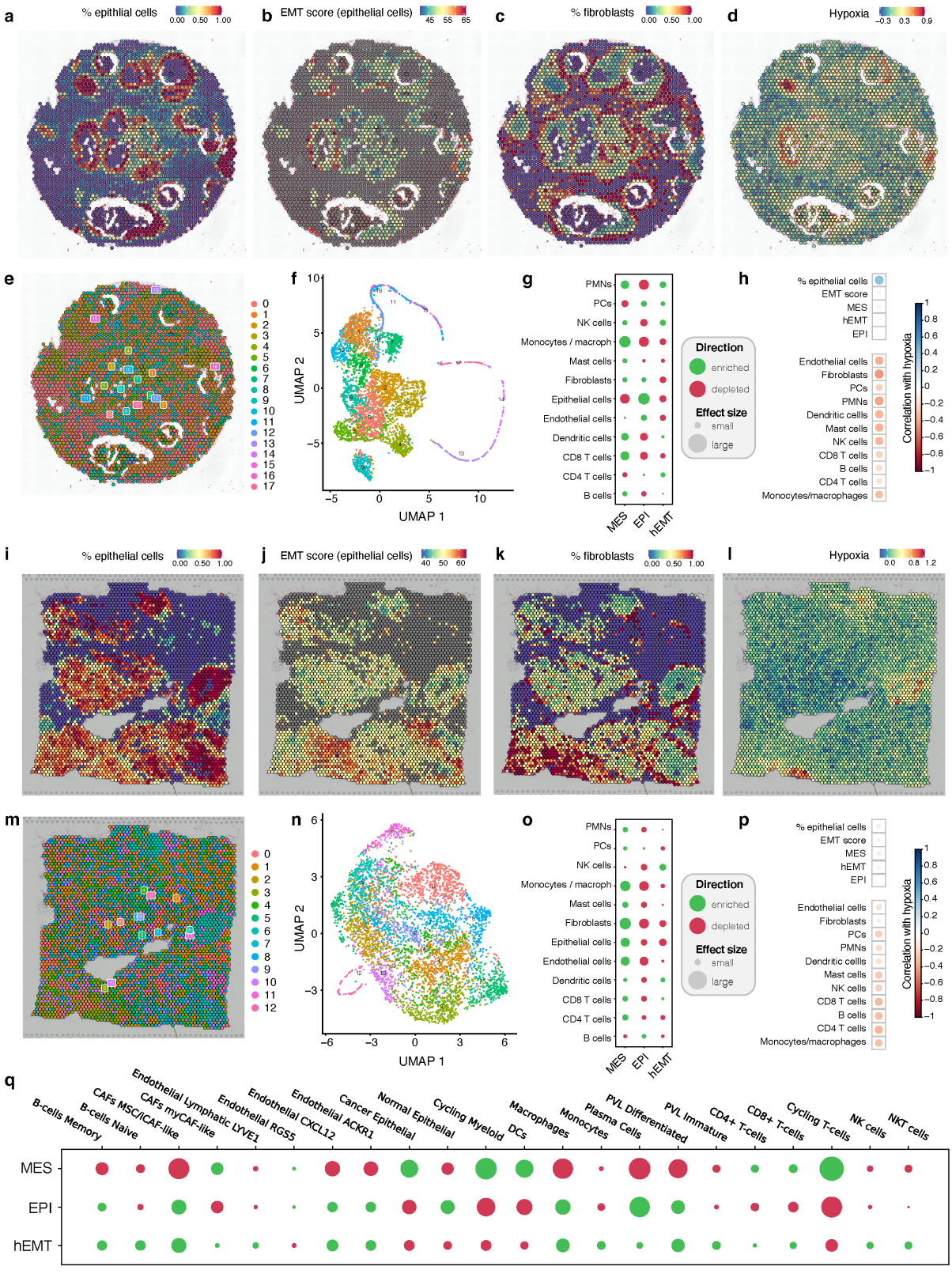


Figure 3. Spatial patterns of EMT. (a-d) Spot annotations of the fraction of epithelial cells (a), EMT scores across these epithelial spots (b), fraction of fibroblasts (c) and hypoxia (d) within individual spots profiled across the tissue in a selected breast cancer slide, derived from

spatial transcriptomics data (Patient 1). The blue to red gradient indicates increased expression of markers of the specific cell state or increased fraction of cell types. (b) Clusters of homogeneous expression profiles annotated within the spatially defined transcriptomic spots for the same slide. (c) Expression clusters visualised using UMAP dimensionality reduction. (d) Enrichment (green) and depletion (red) of cell types in each EMT-based cluster. The plots represent the difference between the average cell type proportion value per region, compared to a permuted spot value (calculated 10,000 times). The plot marker size corresponds to the absolute enrichment score, and the colour represents the enrichment sign (red for negative and green for positive). (h) Correlation between hypoxia and individual cell types and states. Blue indicates positive correlation, red indicates negative correlation, with the circle size being proportional to the correlation value. (i-p) The same annotations as above for a breast cancer sample from Patient 2. (q) Enrichment (green) and depletion (red) of cell types in EMT-based clusters derived from multi-region spatial transcriptomics slides from the ST2K cohort. Annotation as in (g) and (o).

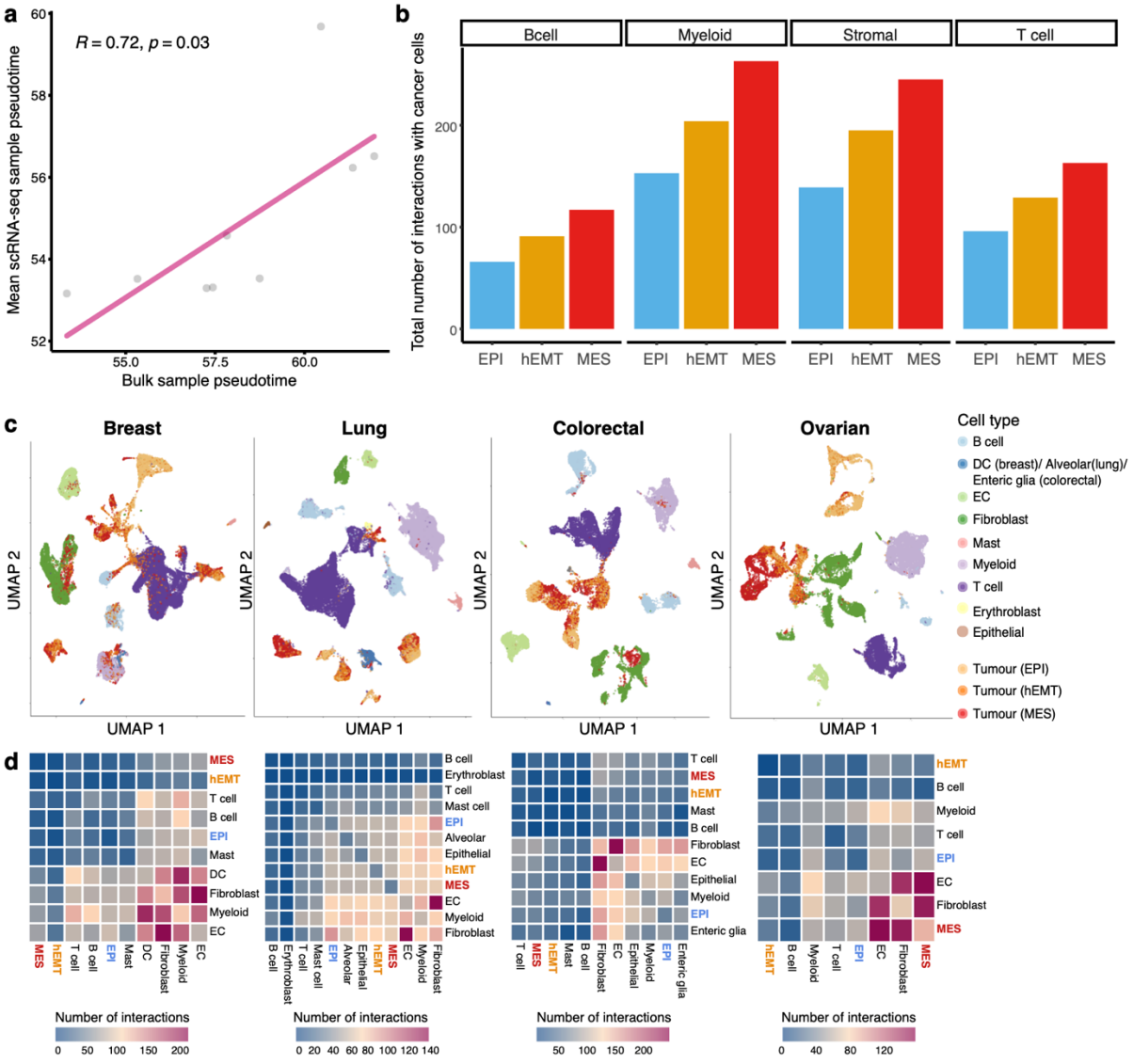
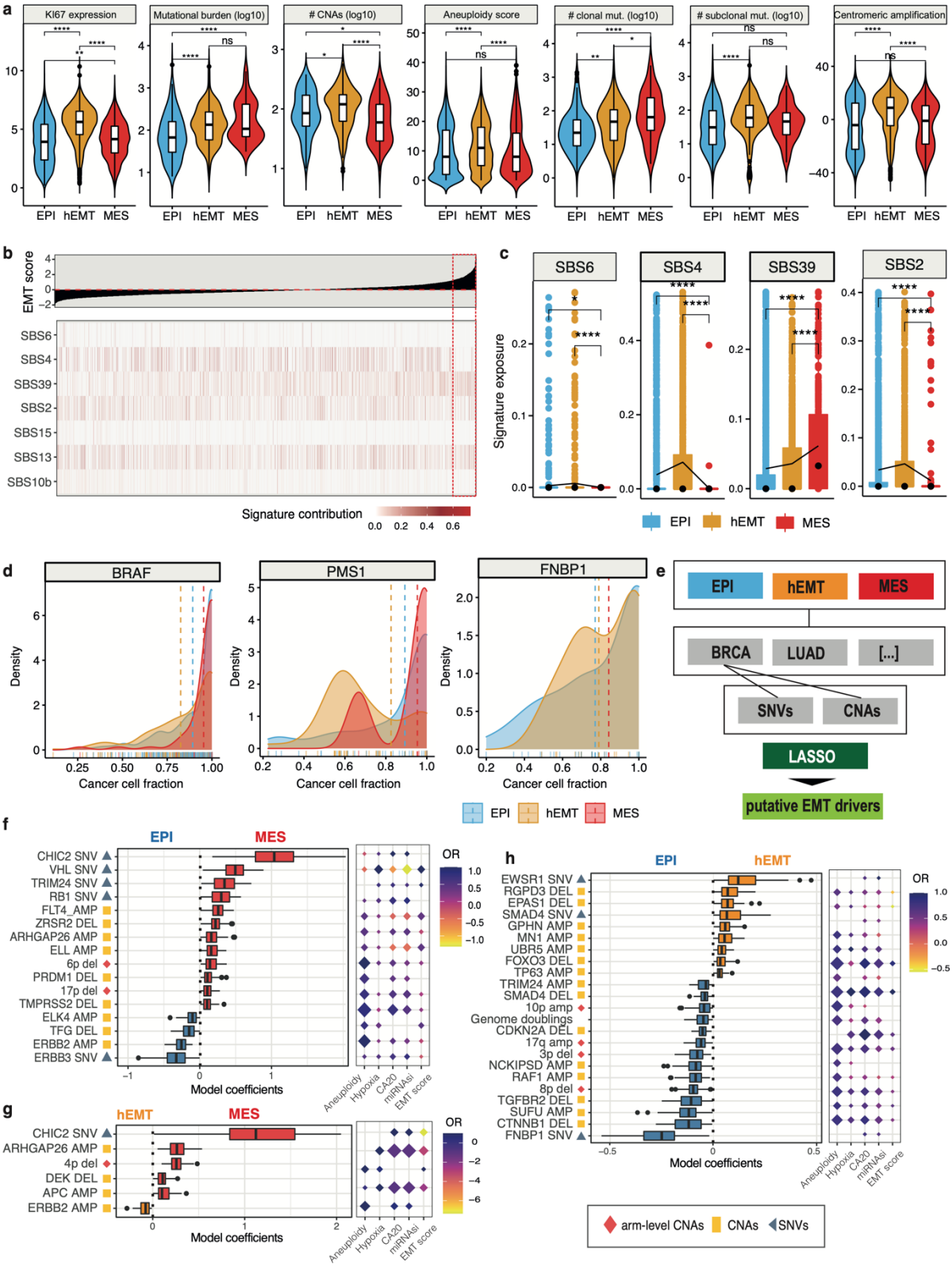


Figure 4. EMT diversity in single cell data. (a) Comparison between EMT pseudotime estimates in matched bulk and single cell samples from the same individuals. (b) Number of interactions established between tumour cells found in an EPI, hEMT or MES state and other cells in the tumour microenvironment in the Chung et al[36] dataset. (c) UMAP reconstruction of single cell expression profiles depicting the tumour and microenvironment landscape of breast, lung, colorectal and ovarian tumours from Qian et al[37]. Tumour cells are coloured according to their assigned EMT state (EPI/hEMT/MES). All other cells in the microenvironment are also depicted in different colours. DC – dendritic cells; EC – endothelial cells. (d) Heat maps depicting the total number of interactions established among all cell types

1087 in the same breast, lung, colorectal and ovarian datasets. The EPI, hEMT and MES tumour

1088 cells are highlighted in blue, orange and red, respectively.

1089



1090

Figure 5. Genomic driver events linked with EMT. (a) Expression of the proliferation marker Ki67, mutational and copy number aberration (CNA) burden, aneuploidy, number of clonal/subclonal mutations and centromeric amplification levels compared across the three EMT states. (b) Mutational signature exposures across TCGA samples sorted by EMT score. Only mutational signatures that were significantly linked with EMT from the linear mixed models are displayed. The corresponding EMT scores are displayed above. (c) Signature contributions from SBS6 (mismatch repair deficiency), SBS4 (smoking), SBS39 (unknown) and SBS2 (APOBEC) compared between the three EMT states. (d) Cancer cell fraction of genomic markers showing significantly distinct distribution between EMT states. (e) The analytical workflow used to detect genomic events linked with EMT. For each state and cancer type, we used dNdScv, SNV and copy number enrichment to prioritise mutated genes and copy number events, respectively. These genomic events were then employed as input for lasso modelling to classify EMT states. (f) Top-ranked genomic markers distinguishing the mesenchymal from the epithelial state. The balloon chart on the right illustrates the association between each marker and aneuploidy, hypoxia, centromeric amplification (CA20), stemness index (mRNAsi) and EMT score. The size of the diamonds is proportional to the significance of association, the colours report the odds ratios. (g) List of the top-ranked genomic markers distinguishing the hEMT from the MES state and their associated hallmarks. (h) List of the top-ranked genomic markers distinguishing the hEMT from the EPI state and their associated hallmarks.

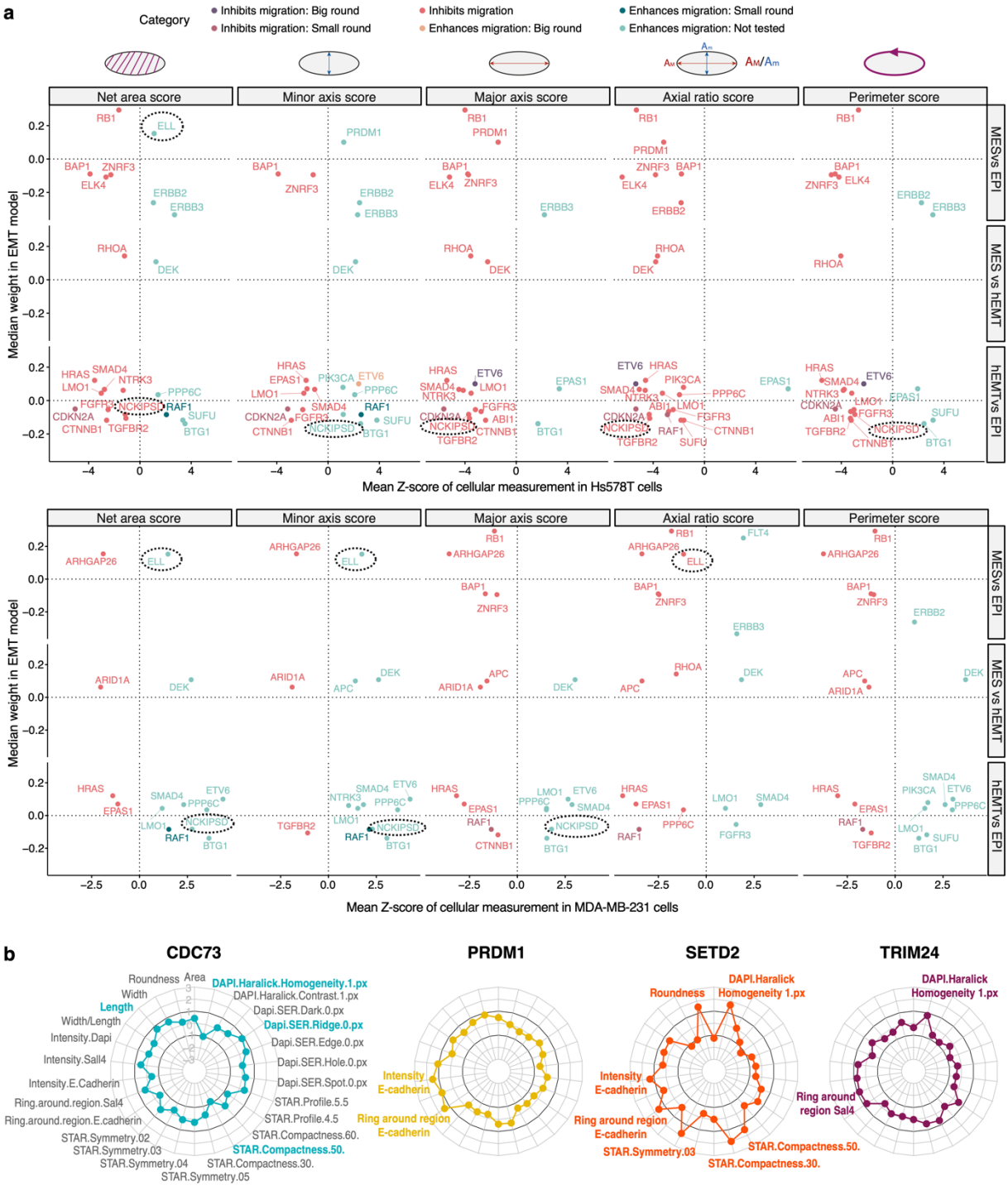


Figure 6. Validation of genomic associations with EMT using siRNA screens. (a) Gene knockdown effects on cell migration abilities in Hs578T (top panel) and MDA-MB-231 (bottom panel) cell lines (data from Koedoot et al[43]). The x axis depicts a change in the following measurements in the cells upon the knockdown: net surface area, length of minor and major axes, axis ratio (large/small: elongated cells, close to 1: round cells), perimeter score (larger –

more migration). The y axis depicts the median weight of the gene in the model distinguishing two different EMT states. Larger absolute weights indicate more confident associations with EMT. The genes are coloured according to the suggested phenotype by the respective cellular measurement. A few of the genes highlighted have undergone further phenotypic tests and this is indicated by the confirmed phenotype (big/small round). The rest of the genes were not further tested in the study ("Not tested"). Only candidates with a Z-score value of cellular measurement >1 or <-1 are shown. The genes ELL and NCKIPSD are highlighted with black dotted rectangles as they are less well characterised in the context of EMT and are found as hits in the screen shown in panel b too. (b) Gene knockdown effects on various measurements of migration-related cell integrity in mouse embryonic fibroblasts (data from Penalosa-Ruiz et al[45]). The radial plots show the mean z-score depicting the change in cell measurement across multiple knockdown replicates. Z-scores greater than 1 or less than -1 (above and below the corresponding black circles) suggest significant changes. All measurements are listed for the first gene only in grey text. Coloured text indicates significant changes in phenotype for each gene, e.g. knockdown of PRDM1 and SETD2 leads to an increase in E-cadherin expression intensity and area.

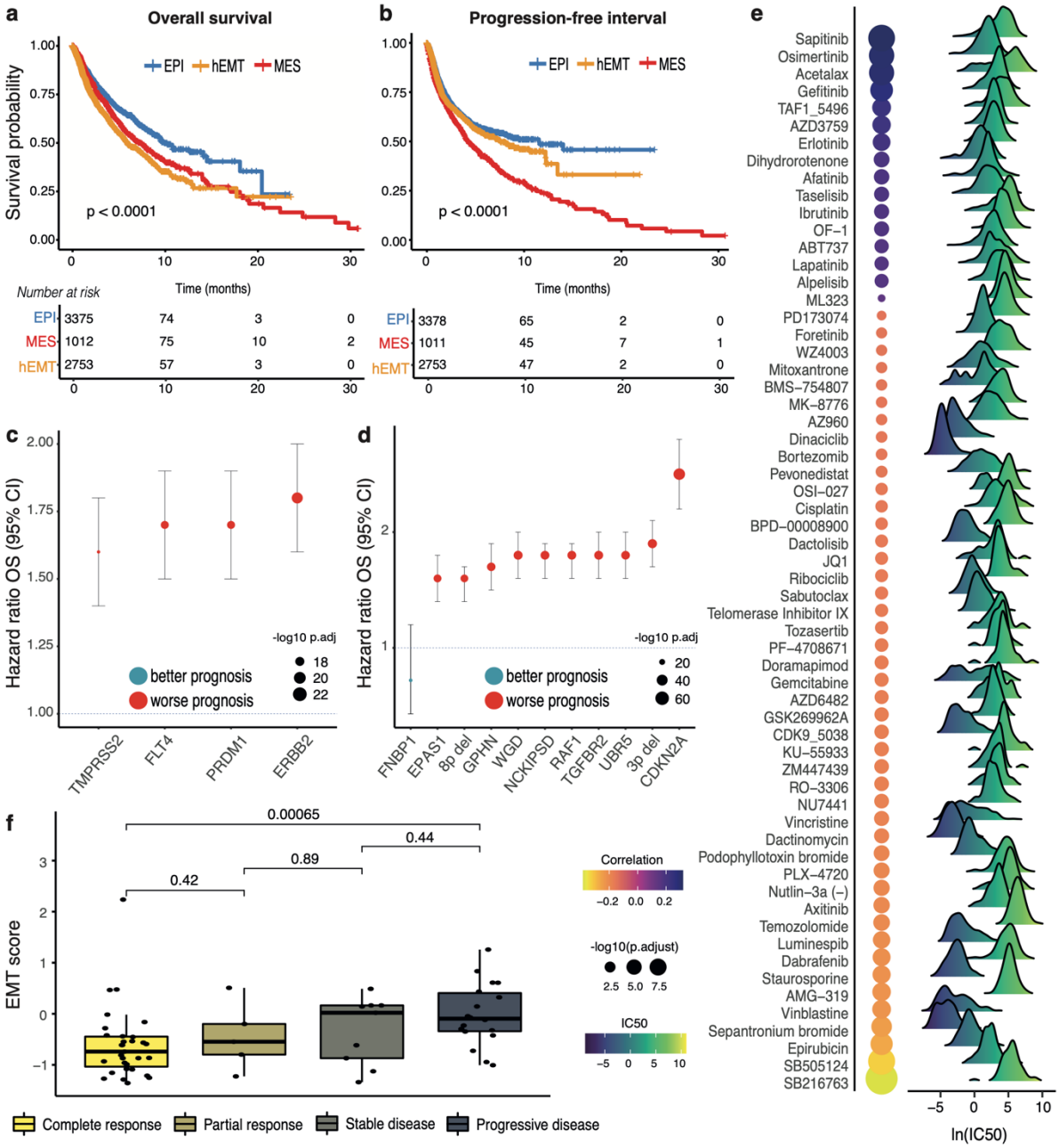


Figure 7. Clinical relevance of the EMT states. (a) Overall survival compared between MES, hEMT and EPI samples. (b) Progression free interval compared between the three groups. (c) Genomic markers distinguishing between mesenchymal and epithelial states with a significantly worse or improved outcome (q<0.001). (d) Genomic markers distinguishing between hybrid and epithelial states with a significantly worse or improved outcome. WGD = whole-genome doubling. (e) EMT scores compared between responders and non-responders to treatment with oxaliplatin. A gradual increase in EMT levels is observed with progressively worse outcomes. (f) Correlation between the EMT scores and IC50 values in cell lines treated

1144 with various drugs. The balloon chart on the left illustrates the association between the IC50
1145 for each compound and EMT. The size of the circles is proportional to the significance of
1146 association. The IC50 ranges for all cell lines are depicted by the density charts.

1147