Genomic events shaping epithelial-to-mesenchymal trajectories in cancer

Guidantonio Tagliazucchi
University College London
Maria Secrier (m.secrier@ucl.ac.uk)
University College London https://orcid.org/0000-0003-2758-1741

Article

Keywords: epithelial to mesenchymal transition, hybrid E/M, cancer progression

Posted Date: August 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-754194/v1

Genomic events shaping epithelial-to-mesenchymal trajectories in cancer

Guidantonio Malagoli Tagliazucchi[1], Maria Secrier[1,*]

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

*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 8,778 tumours of epithelial origin and identified three macro-states with prognostic and therapeutic value, attributable to epithelial, hybrid E/M (hEMT) and mesenchymal phenotypes. We show that the hEMT state is remarkably stable and linked with increased aneuploidy, APOBEC mutagenesis and hypoxia. Additionally, we provide an extensive catalogue of genomic events underlying distinct evolutionary constraints on EMT transformation, including novel pan-cancer dependencies of hEMT on driver genes *PRRX1, BCOR* and *CNOT3*, as well as links between full mesenchymal transformation and *REG3A* and *SHISA4* mutations in lung and breast cancers, respectively. This study sheds light on the aetiology of the lesser characterised hybrid E/M state in cancer progression and the broader genomic 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*[6,7,8]. 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], *BRCA1*[11], *STAG2*[12], *TP53*[13,14], as well as amplifications of chromosomes 5, 7 and 13 have been shown to promote EMT[15]. Several pan-cancer studies have also linked copy number alterations, miRNAs and immune checkpoints with EMT on a broader level[16,17]. 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 evolutionary 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. 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 this, we have integrated data from the Cancer Genome Atlas (TCGA), MET500[18], MetMap[19], GENIE[20], MSK-IMPACT[21], GDSC[22] and POG570[23] datasets to characterise and validate EMT and linked genetic changes in cancer. By mapping 8,778 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 the lesser characterised hybrid E/M phenotype and could potentially act as early biomarkers of invasive cancer.

RESULTS

Pan-cancer reconstruction of EMT trajectories and states in bulk tumours

To explore the EMT process within bulk tumour samples, we employed a cohort of primary tumours of epithelial origin (n = 8,778) 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 reference single cell RNA sequencing (scRNA-seq) dataset derived from MCF10 breast cancer cells 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). Since the bulk signal is arising from a mixture of cells, some of which would be

found in different stages of mesenchymal transformation, the overall signal would 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[24].

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 1b). These states were robust to varying levels of gene expression noise (Supplementary Figure S1c-d). As expected, the probability for the cancer cells to switch from the epithelial to the hEMT (0.35) state was higher than the probability to passage directly into the mesenchymal state (0.14). The hEMT tumours tended to remain in the same state 44% of the times, suggesting this state could be more stable than anticipated – as previously stipulated[25] and consistent with observations that a fully mesenchymal state is not always reached[26].

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 S1e). 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

105 invasion potential of these cell lines increased along the pseudotime axis as expected (Figure
106 1e, Supplementary Figure S1f), Cell lines classified as MES by our HMM model were
107 predominantly metastatic, while hEMT cases had more varied invasion potential.

108 At tissue level, the proportion of samples in each EMT state was variable (Figure 1f,
109 Supplementary Figure S3g), with hEMT dominating in head and neck, oesophageal and
110 pancreatic carcinomas, while adenoid cystic, kidney carcinomas and melanomas were highly
111 mesenchymal.

112 As expected, the EMT classification was significantly correlated with the clinical cancer stage
113 (Chi-square test $p<<0.0001$), with transformed samples (hEMT/MES) found to be 1.6-fold and
114 3.3-fold enriched in late-stage tumours, respectively, while the epithelial state was 3-fold
115 overrepresented in early-stage cancers (Figure S1h). Overall, 18% of the profiled samples
116 were classified as fully transformed (MES) and only 8% of these were annotated as late-stage
117 tumours (Supplementary Table S2). Therefore, our method is able to detect an additional 10%
118 of cases presenting early evidence for the phenotypic transformation required for metastasis.
119 Indeed, multiple studies have demonstrated the activation of the EMT transcriptional
120 programme in the early stages of cancer[10,27]. Even the hEMT phenotype is hypothesised to be
121 sufficient for promoting metastatic dissemination[28], although our analysis of the MetMap
122 dataset suggests this may be tissue/context-specific.
123

124 **Tumour cell extrinsic hallmarks of EMT**

125 Multiple microenvironmental factors, including tumour associated macrophages, secreted
126 molecules (IL-1, TNF-α) or hypoxia, have been extensively described to promote EMT[29,30,31,32].
127 However, their macro-state specificity is less well characterised. We confirmed that
128 transformed samples, particularly hEMT ones, exhibited higher infiltration by non-cancer cells
129 (Supplementary Figure S2a-c). Endothelial, cytotoxic and γδ T cells were progressively
130 enriched with increased stages of EMT transformation (Figure 2a-b, Supplementary Figure
131 S2d-e), 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). In contrast, the hEMT samples displayed the highest enrichment of cancer-associated fibroblasts (Figures 2a-b, d-e, Supplementary Figure S2f and Supplementary Table S3), confirming previous reports[33]. To avoid confounding effects between fibroblast and hEMT markers as highlighted by Tyler and Tirosh[34], we excluded EMT markers from the employed fibroblast gene sets, but we acknowledge that part of the signal recovered may still not be unambiguously attributed to either the cancer or microenvironmental component.

Additionally, samples with a transformed phenotype (MES, hEMT) presented significantly elevated hypoxia levels (Figure 2f, Supplementary Figure S2g). Hypoxia has previously been shown to promote EMT by modulating stemness properties[35]. We found that CD44, an established cancer stem cell marker known to promote EMT[36,37], was most highly expressed in the hEMT state across cancers (Figure 2g), and elevated levels of several other stemness signatures most often accompanied the hEMT and MES macro-states (Figure S2h-i). Unlike mesenchymal samples, the majority of hEMT tumours (35%) were characterized by both hypoxia and CD44 expression (Supplementary Table S4). 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.

Tumour cell intrinsic hallmarks of EMT

In addition to environmental factors, intrinsic cell properties such as increased proliferation, mutational burden and aneuploidy would be expected along the EMT transformation trajectory. Across distinct tissues, these changes were most pronounced in the hEMT state (Figure 2h). Interestingly, this 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 while accounting for tissue effects (Figure 2i-j, Supplementary Figure S3a, Supplementary Table S5). The ageing signature SBS1 was significantly increased in non-mesenchymal tumours, while the APOBEC mutagenesis signature SBS13 appeared specifically elevated in hEMT. It is worth noting that if we did not account for the tissue effect, the signatures most strongly associated with EMT progression and specifically increased in mesenchymal samples were SBS7a/b, linked with UV light-induced damage (Supplementary Figure S3b-c). Samples exposed to these carcinogens may progress to a mesenchymal phenotype earlier within the primary tumour, or such mutations may be more easily fixed in transformed cells.

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. 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 3a, Supplementary Table S6a). This model was validated using several other machine learning approaches and demonstrated remarkably high accuracies of 80-99% (Supplementary Figure S4a-n, Supplementary Table S6b).

Among the genomic biomarkers able to discriminate transformed tumours (hEMT, MES) from the epithelial state (EPI), we identified well characterized driver genes such as *TP53*, *PIK3CA*, *KRAS*, along with chromosomal arm alterations at 1q, 8q, 17q, 20q, some of which have been previously linked with cell migration and invasion (Figure 3b, Supplementary Figure S4o, Supplementary Table S6c). Interestingly, the fraction of cancer cells harbouring *KRAS*, *RB1*

and *FGFR2* mutations was markedly increased at the mesenchymal level, suggesting that a clonal fixation of these events may be key for the establishment of a fully mesenchymal state (Figure 3c-d). This is in line with findings on *KRAS* reported by McFaline-Figueroa et al[9], but also pinpoints additional checkpoints.

The hEMT state-specific markers were mostly enriched in cell cycle, fate commitment and cell adhesion pathways, hallmarks that are classically associated with cancer progression (Figure 3e, Supplementary Figure S4p). These alterations most often presented in a reduced cancer cell fraction, concordant with subclonal diversification in line with phenotypic transformation (Figure 3c). We identified several events that have not been previously linked with this intermediate phenotype in the literature, including alterations of *PRRX1*, *BCOR*, *FAM135B,* *CNOT3* and *ERCC3* (Figure 3e, Supplementary Figure S5a-b, Supplementary Table S6a). *TCEA1* and *SPEN* alterations were specifically linked with high levels of aneuploidy, stemness, centrosome amplification and hypoxia. While none of these events have been specifically tied to an hEMT state before, *CNOT3* does not appear to have even been linked more generally with invasion or metastasis. *CNOT3* is a translational repressor[42] upregulated in non-small cell lung cancer[43].

Among the chromosomal arm-level events, amplifications of the 16q arm appeared to confer the highest protective effect from transformation, with a >2-fold depletion in hEMT (Supplementary Figure S5c). Another notable event was the deletion of the 5q arm, associated with the EPI state. The trisomy of chromosome 5 has been previously linked with a hybrid E/M state in a colon cancer cell line[15], hence structural changes of chromosome 5 might impose a selective pressure on EMT transformation.

The genomic markers distinguishing mesenchymal from epithelial samples were enriched in oxidative stress and hypoxia response pathways (Figure 3f, Supplementary Figure S4q), suggesting an evolutionary adaptation to mesenchymal-promoting changes in the environment. All of the identified drivers have previously been linked to metastasis or EMT transformation[44,45] (Supplementary Table S6b-c), e.g. *VHL* mutations via regulation of

hypoxia[46]. Increases in copy number amplification were generally associated with a more mesenchymal state (Supplementary Figure S5d), as were deletions of the 6p chromosomal arm which harbours 16 cancer drivers, several of which have been previously linked with EMT (Supplementary Figure S5e, Supplementary Table S6a).

A good fraction of the reported alterations was independently validated in the MET500, GENIE and MSK-IMPACT studies: 63% of point mutations and 43% of copy number events distinguishing MES from EPI, as well as 58% of somatic mutations and 70% of the copy number events discriminating hEMT from EPI (Figure 4a-d). Acknowledging the limitations of these resources to metastatic samples (MET500) and targeted sequencing (GENIE, MSK-IMPACT), our analysis was nevertheless able to recapitulate genomic events necessary for EMT transformation that are preserved during metastatic progression, but also uncovered some events that may not be further maintained in the new metastatic niche. Furthermore, 31% and 57% of these alterations were confirmed to be linked with the metastatic potential of cancer cell lines, at pan-cancer and tissue specific level, respectively (Figure 4e-f). 48% of the copy number events were validated in at least one cell-line, e.g. *COX6AC* amplifications in the pancreas (Figure 4h). Interestingly, the impact on metastatic potential varied between increases and decreases, which can partly be explained by the heterogeneity and complex dependencies of cell lines, and partly by our incomplete understanding of the metastatic potential of hEMT cells. Suppression of most of these genes strongly impacted cell viability (Figure 4i-j) and were targets of transcription factors regulating EMT (Figure 4k-l). Knockdowns of the hEMT-linked genes *PRRX1* and *CREBBP* were also linked with a weak EMT phenotype in murine epithelial cells[47].

Finally, we confirmed downstream variations of the proteome accompanying EMT transformation for the key EMT markers E-cadherin, N-cadherin, serpin E1, fibronectin (Supplementary Figure S6a, Supplementary Table S6d). Proteins upregulated in hEMT included components of the Hippo pathway (YAP,TAZ), regulating cell adhesion and mechanical signals, with previously demonstrated roles in EMT transformation[48]

(Supplementary Figures S6b,d-e). In contrast, proteins differentially expressed in a fully mesenchymal state (Supplementary Figure S6c) were linked with mTOR signaling, response to oxygen levels, and UV damage response (Supplementary Figure S6f-g), consistent with our previous findings linking the MES state with hypoxia and the UV mutational signature SBS7.

Tissue-specific EMT trajectories and genomic dependencies

EMT is a ubiquitous program, however, it is strongly influenced by tissue type and external stimuli. Our pan-cancer reconstruction of the EMT pseudotime relied on data from single cells captured during spontaneous EMT transformation. To further refine EMT trajectories in a tissue-specific manner and to consider other sources of influence on the EMT transcriptional programme, we used scRNA-seq data from two cancer cell lines (the human adenocarcinoma alveolar basal epithelial cell line A549, and the breast cancer cell line MCF7) that were stimulated with different exogenous molecules (TGF-β1, EGF, TNF) in a time-course experiment[49]. As before, we mapped bulk lung and breast primary tumours from TCGA onto the EMT pseudo-timeline derived from the scRNA-seq data and observed a finer granularity of different activation states along the EMT trajectory in these cancers (Figure 5a-b, Supplementary Figure S7a,d). In both cancers we identified 5 states, with multiple epithelial, hEMT and mesenchymal-like phenotypes (Figure 5c-d).

We found evidence for positive selection in 17 genes across the 5 EMT stages in lung adenocarcinoma, 9 of which had also been identified in the pan-cancer analysis (Figure 5e, Supplementary Table S7a-e). Seven drivers (*RB1*, *MGA*, *ZIC1*, *NF1*, *REG3A*, *ARID2*, *ZFP36L1*) were only enriched in the mesenchymal-like cluster M2. Among these, *REG3A*, encoding for a secretory protein linked with inflammation-driven carcinogenesis and cell migration in gastrointestinal cancers[50], has not been previously linked with EMT in lung cancer and could potentially constitute a novel biomarker of mesenchymal transformation in this cancer type, where it tends to be more highly expressed when mutated. Indeed, cell lines carrying *REG3A* mutations show a strong increase in metastatic potential, particularly in lung

(Figure 5g-h). REG3A depletion does not impact cell viability in upper aerodigestive tissue (Supplementary Figure S7g), making it a putative pharmacological target.

In breast cancer, the most mesenchymal state M was again linked with *NF1* mutations, but also with alterations of *AKT1* and *SHISA4* (Figure 5f, Supplementary Table S7f-j). *AKT* has been shown to affect epithelial cell morphology and motility[51], while the role of *SHISA4* in EMT has not been characterised. Both genes are targets of transcription factors known to modulate EMT (Supplementary Figure S7i).

Most of the focal copy number alteration events in both lung and breast cancers were distinct from the pan-cancer analysis (Supplementary Figure S7b-f, Supplementary Table S7e,j), and several of them have been previously associated with EMT, albeit not in a micro-state specific manner. Amplifications of the non-essential *PRDM2* gene in breast cancer and deletions of the non-essential gene *CUX1* in lung cancer, among others, were linked with metastatic potential in cell lines and transcription factor regulation of EMT (Figure 5i, Supplementary Figure S7h,j).

Clinical relevance of EMT

Finally, we show that the defined EMT states have potential clinical utility. Patients with an hEMT phenotype had the worst overall survival outcomes (Figure 6a, Supplementary Table S8a), while those with mesenchymal tumours showed significantly decreased progression-free intervals (Figure 6b, Supplementary Table S8b) and disease-specific survival (Supplementary Figure S8a, Supplementary Table S8c). As expected, patients who presented transformed tumours and were lymph node positive had the worst prognosis (Supplementary Figure S8b).

Among the driver events that have been linked with EMT in this study, mutations in genes *CTNNB1* and *TSC2*, associated with a mesenchymal phenotype, and six other driver genes

associated with hEMT (*BCOR*, *RHOA*, *CDH10*, *SMAD2*, *SETD2*, *PTHCH1*) conferred worse prognosis (Figure 6c-d, Supplementary Table S9).

Interestingly, the progression-free interval after oxaliplatin treatment was shorter in patients with hEMT tumours (Supplementary Figure S8c), suggesting hEMT might be linked with poor responses to this chemotherapy drug. 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[22]. We found 22 compounds whose IC50 values were significantly correlated with the EMT score (Figure 6e). The strongest associations were observed with Acetalax, a drug used in the treatment of triple negative breast cancers, and Sapitinib, an inhibitor of ErbB1/2/3[52]. Several hEMT and MES biomarkers (e.g. *EP300*, *FAT1*, *NFE2L2*, *PTEN*) conferred increased sensitivity to multiple drugs in a tissue specific manner (Supplementary Figure S8d-f), and suggested opportunities for repurposing of non-oncology drugs (Supplementary Figure S8g-h).

Finally, we linked post-treatment EMT phenotypes with therapy responses using the POG570 dataset (Supplementary Figure S8i). The duration of treatment, used as a proxy for worsening outcome, varied based on EMT state for the aromatase inhibitors exemestane and letrozole, but also for the chemotherapeutics etoposide and temozolomide (Supplementary Figures S8j-n). Moreover, EMT potential was decreased upon chemotherapy treatment compared to unmatched treatment-naïve cases in EPI and hEMT samples, while MES potential increased (Figure 6f). There was a consistent drop in EMT potential in capecitabine and letrozole-treated tumours (Figure 6g-h, Supplementary Figure S8o-p). Hence, the level of EMT transformation may play a role in determining responses to a variety of cancer therapies.

DISCUSSION

Previous studies of the EMT process have suggested the existence of a phenotypic continuum characterised by multiple intermediate states[53]. We have shown that distinct EMT

318 trajectories in cancer are underpinned by three macro-states, reflecting both tumour cell
319 intrinsic as well as tumour microenvironment associated changes. The hybrid E/M state,
320 characterised by the co-expression of epithelial and mesenchymal markers, was surprisingly
321 frequent (38%). It presented traits linked with increased neoplastic aggressiveness, such as a
322 tumour-promoting microenvironment and worse clinical outcomes, also in certain
323 chemotherapy contexts. Indeed, it has been reported that cells with hEMT features give rise to
324 daughter cells that are either mesenchymal or epithelial[54,36], are more prone to migrate and
325 promote the formation of circulating tumour cells[37]. Moreover, stable late hEMT states have
326 been linked with maximal metastatic potential and worse overall survival[55], explaining the
327 variable metastatic potential we observed in cell lines in the context of hEMT. Undoubtedly,
328 the hEMT state can be further subdivided into sub-states, as shown by Goetz et al[4], Brown et
329 al[56] and also by us in the tissue-specific analyses. The true number of EMT intermediate
330 states is just beginning to be explored. However, the bulk sequencing data are limiting our
331 ability to capture them, and this will be best studied in single cell datasets.

332 Our study confirmed previously established molecular hallmarks of EMT, including increased
333 genomic instability and hypoxia in hEMT, and cytotoxicity/exhaustion in mesenchymal
334 tumours[17], which could inform immunotherapy strategies. We also highlighted mutational
335 processes that have increased activity in specific EMT states. Beyond an expected
336 association with ageing-induced damage along the EMT progression axis, we also found that
337 the signature SBS13, linked with the activity of the APOBEC3A/B cytidine deaminases, was
338 prevalent in hEMT tumours. In mammary epithelial cells, inflammatory signals promote the
339 upregulation of the activation-induced cytidine deaminase (AID) enzyme, a component of the
340 APOBEC family, and this has been shown to trigger EMT[57]. Furthermore, cyclic hypoxia has
341 recently been proven to induce APOBEC activity[58], which could further explain the
342 convergence of these phenotypes in the hEMT state.

343 While the exploration of EMT biomarkers is not new, most of the studies in this area have
344 been reliant on gene expression activity rather than mutational dependencies and they are

generally tissue-specific[17,30]. Pan-cancer studies generally consider EMT as a binary switch[16,17,30]. In contrast, our study identified genomic hallmarks of three EMT macro-states derived from the integration of bulk and single cell datasets, which provided the opportunity to understand the evolutionary constraints on the lesser characterised hybrid E/M state. We uncovered novel putative hEMT drivers like *PRRX1* and *BCOR,* and events in genes *CNOT3*, *REG3A* and *SHISA4*, which have not been linked with hybrid/mesenchymal phenotypes or metastatic expansion in the studied cancers. A causal relationship between the acquisition of any of these genomic changes and EMT should be experimentally tested in the future. The EMT state by itself, as well as several putative EMT biomarkers, were linked to drug responses and could thus potentially be exploited for therapeutic benefit.

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 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]) and clinical data were retrieved for 8,7778 primary tumours of epithelial origin from the harmonized version of TCGA using the *TCGAbiolinks* R package[60]. 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 to map bulk samples to single cell-derived expression programmes

inspired from McFaline-Figueroa et al[9]. The workflow of the analysis consists of several steps. The analysis requires two gene expression matrices as input, corresponding to one bulk sequenced dataset and a single cell dataset. In our case, we used as input the bulk RNA-seq data from TCGA samples and scRNA-seq of a spontaneous EMT model derived in MCF10 cell lines[9] with the associated trajectory (P). 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[61]. In the second step, principal component analysis (PCA) was performed on the merged matrix. The MCF10 derived EMT trajectory was then mapped onto the TCGA 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 TCGA samples. During each i-th step of iteration, a single TCGA sample and the MCF10 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 TCGA sample. The implementation of the kNN algorithm is based on get.knnx() function from the *FNN* R package.

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.

Computation of the EMT scores

A list of epithelial and mesenchymal markers was compiled through manual curation of the literature[30,6,9], as follows:

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

EMT scores for each TCGA sample were computed in a similar manner as described by Chae et al[62]. Briefly, the average z-score transformed expression levels of the mesenchymal markers were subtracted from the average z-score transformed expression levels of the epithelial markers. To segment the EMT trajectory, along with the epithelial and mesenchymal

markers we have also considered markers of hybrid EMT[63,6]: *PDPN*, *ITGA5*, *ITGA6*, *TGFBI*, *LAMC2*, *MMP10*, *LAMA3*, *CDH13*, *SERPINE1*, *P4HA2*, *TNC*, *MMP1*.

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

The RSEM gene-expression values of the Cancer Cell Line Encyclopedia[64] project were retrieved from the CCLE Data Portal. We used the same procedure described above to map the CCLE data onto the spontaneous MCF10 EMT trajectory. 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[19]. 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[65]. Immune deconvolution was performed using the ConsensusTME R package[66] 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.

Fibroblasts related gene-sets were manually curated as described in Supplementary Table S3. Fibroblasts enrichments scores were calculated via ssGSEA using the *GSVA*[67] R package.

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 [68,40]. The hypoxia levels were quantified as described by Bandhari et al[69]. Several hypoxia gene

signatures were considered, yielding similar results. Only the results obtained using the genes from Buffa et al[70] were reported. Finally, to estimate the levels of stemness in each EMT state, we considered a catalogue of stemness gene sets[71] 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[72]. 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.[69], 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:

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

Identification of cancer driver genes

Single nucleotide variants were obtained from TCGA using the *TCGAbiolinks* R package and the Mutect pipeline. To identify putative driver events that are positively selected in association with an EMT state, we employed dNdScv[73], which quantified the ratio of non-synonymous and synonymous mutations (dN/dS) in each gene and state, by tissue. We only considered protein coding genes in the analysis, and filtered out consecutive mutations in the genome, as recommended by the authors. All the somatic driver events with a q-value less than of 0.10 were considered for downstream analysis.

Definition of somatic copy number events

Focal and arm-level copy number events across TCGA samples were identified using GISTIC2[74], with the following parameters: armpeel = 1, brlen 0.5, cap 1.5, conf 0.99 -ta -td = 0.3, -genegistic = 1, gcm = extreme, -maxseg = -2500, qvt = 0.1, rx = 0, save gene = 1, broad = 1.

Identification of genomic events linked with EMT

To search for genomic events linked with the described EMT macro-states, we considered all somatic mutations, focal and arm-level copy number events in driver genes from the COSMIC database that were obtained in the previous steps. Two parallel methodological approaches based on lasso and random forest were used to identify events that could be predictive of EMT transitions in a two-step process. First, feature selection was performed using a stability selection approach. We used the function createDataPartition() from the *caret* R package to generate an ensemble of vectors representing 1,000 randomly sampled training models. This is an iterative approach, in which at each iteration a lasso analysis is performed, and the non-negative coefficients computed by lasso are saved. This step was performed using the cv.glmnet() function from *glmnet*. The tissue source was included as potential confounder in the lasso model. The models were trained on 80% of the data. At the end of this stage, the variables that were selected in at least the 80% of the iterations were taken forward and

employed as predictors. A similar approach was employed for feature selection and model building with random forest.

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

Finally, to explore the relation between the outcomes (the EMT states) and the biomarkers we used partial dependence plots generated with the function variable_effect() from the R-package *DALEX*.

Cancer cell fraction estimates

The cancer cell fraction (CCF) of selected mutations was calculated using the following 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[65].

Validation of the putative EMT biomarkers

The following resources were used to validate the genomic associations with the EMT programme: MET500[18], MSK-IMPACT[21], GENIE (Version 8)[20] and MetMap[19]. Different strategies of pre-processing were employed as described below.

- **MET500**

Copy number regions encompassing 50% or more of the chromosomal arm were considered as broad copy number events. Copy number regions below this threshold were classified as focal events (based on hg19 cytoband positions and genomic coordinates obtained from UCSC).

- **MSK-IMPACT**

The MSK-IMPACT resource consists of metastatic (n= 4,048) and primary tumors (n= 6,262). First, we removed from the analysis all the samples where the metastatic site was 1) Blood vessel, 2) Brain, 3) Spinal Cord or 4) Not available. Copy number events were classified as broad or focal as described above for the MET500 data. Only regions where the absolute copy number was greater than 1 and which were overlapping driver genes as annotated in COSMIC were considered for downstream analysis. The frequency of each marker (e.g., TGFBR2_focal) was compared between the primary and metastatic tumors using a Chi-square test. The resulting p-values were adjusted using Benjamini-Hochberg multiple testing correction. Finally, only the markers with p-value<0.05 and odds ratio >1 were selected. The same approach was applied for point mutations.

- **GENIE**

The GENIE dataset consists of metastatic (n=22,243) and primary tumors (n= 55,742). First, we removed all the samples from 1) Glioma 2) Blood 3) Brain 4) Leukemia 5) Lymphoma 6) Sellar 7) Nervous 8) CNS 9) Neuro 10) Neoplasms 11) Lymphoproliferative, or 12) Nerve. The same approach described in the MSK-IMPACT section was used to analyze the GENIE dataset. Only the copy number markers with q-value<0.10 and odds ratio >1 were selected. Silent mutations were excluded.

- **MetMap**

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.

Gene essentiality evaluation using Project Achilles

The viability of the cancer cell lines harbouring putative EMT biomarkers was evaluated based on CRISPR screening data[75] 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.

EMT-linked transcription factor network analysis

Data from knockdown experiments of transcription factors regulating EMT was downloaded from Meyer-Schaller et al[47]. Genes differentially expressed with an absolute log2 fold change >1 and a q-value < 0.05 upon transcription factor knockdown were selected and intersected with our list of putative EMT biomarkers.

Proteomic biomarkers linked to EMT

Level 4 RPPA proteomics data were retrieved from the Cancer Proteome Atlas (TCPAv3.0[76], https://tcpaportal.org/tcpa/download.html). We identified proteins whose levels change significantly between EMT states using logistic regression. Only the proteins with a q-value < 0.10 and odds ratio $<$-0.5 or $>$0.5 have been considered for downstream analysis.

Tissue-specific EMT trajectory derivation

To analyze the EMT in the context of several external stimuli we used the data generated by Cook et al[49]. Specifically, we considered scRNA-seq data of MCF7 (breast cancer cell line) and A549 (alveolar basal epithelial cells) under the stimulus of TNF, EGF and TGF-β1. For each cell line, we integrated the expression data derived from all the conditions (TNF, EGF,

TGF-β1) using the FindIntegrationAnchors() and IntegrateData() functions from *Seurat*[77]. Next, we performed dimensionality reduction on the resulting data using UMAP (from the *uwot* R package) and identified single cell expression clusters using the *mclust* R package. To identify putative trajectories of EMT in each cell line, we used the R package *slingshot*[78] and the function getLineages(). We manually defined the starting cluster for the trajectory as the one with the lowest expression levels of mesenchymal markers. Using a similar bulk-to-single cell mapping approach as described before, we mapped the RNA-seq data of LUAD and BRCA tumours onto the trajectories derived from the single cell data (including batch effect removal using ComBat, PCA on 25 dimensions and kNN clustering). The EMT trajectories derived from the single cell experiments presented different ramifications, therefore a simple computation of the mean pseudotime values of the scRNA-seq samples associated with one TCGA sample was not feasible. To overcome this limitation, we computed the mean of the X and Y component values in the low dimensional matrix of the scRNA-seq data for those scRNA-seq samples that were associated with one TCGA sample. The averaged coordinates represent the position of the single cell samples in the UMAP space. The kNN analysis is based on the transcriptome level of the bulk and scRNA-seq experiments. Therefore, even if we are not able to compute pseudotime values of the TCGA samples with this analysis, we can calculate the distance from any TCGA sample to the closest single cell RNA-seq sample, and therefore obtain a "reference" for the single cell EMT trajectories. Using the "derived" X, Y values of the TCGA samples, we used *mclust* to determine the number of clusters in the TCGA cohort, then we performed a *de novo* reconstruction of the trajectory. The initial and final clusters that determine the path of the EMT trajectory were defined manually on the basis of increasing values of EMT scores (towards a mesenchymal phenotype) in each cluster. The attribution of phenotype (epithelial-like, hEMT-like, mesenchymal-like) was performed based on the lower and upper quartiles of the median EMT scores in each LUAD and BRCA cluster.

603 **Identification of the genomic events in the EMT tissue-specific analysis**

604 To identify the genes positively selected in each EMT tissue-specific cluster, we ran dNdScv[73]
605 in each cluster separately. To detect recurrent copy number alterations in lung
606 adenocarcinoma and breast cancer, we considered all genes with alterations in at least 10
607 patients and applied a Fisher's exact test to identify those that appeared enriched in a
608 particular EMT state. Only the genes with a q-value < 0.05 have been reported.

609 **Relation between EMT biomarkers and drug response**

610 We downloaded the drug sensitivity data from the Genomics of Drug Sensitivity in Cancer
611 database (GDSC)[22]. We considered only the mutated genes linked with EMT states and p-
612 value a cut-off <0.01. We also considered a second database, composed mainly of non-
613 oncology compounds[79]. For each compound, an ANOVA analysis was used to identify
614 significant changes in drug sensitivity between cell lines with or without a specific EMT
615 biomarker. Only the compounds with a p.value<0.01 were considered for downstream
616 analysis.

617 Finally, the POG570[23] dataset was used to study the relation between the EMT states and the
618 duration and effects of given cancer treatments. The EMT states in this dataset were inferred
619 similarly as described above using the kNN approach. To compare pre- and post-treatment
620 samples, we merged the TCGA and POG570 datasets and removed batch effects from the
621 two resources using ComBat. The resulting matrices were used to compute the EMT scores
622 all the samples and compare the levels of EMT between treatment-naïve and post-treatment
623 samples.

624 **Gene ontology analysis**

625 The characterization of the biological processes associated with the reported lists of genes
626 was performed using the R package *clusterProfiler*[80].

627

628

Survival analysis

Standardized clinical information for the TCGA cohort was obtained from Liu et al[81]. The following end points were considered: overall survival, disease-specific survival and progression-free interval. Patients were considered “lymph node positive” if they presented infiltration in one or more lymph nodes. 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.

Code

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). 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. The authors would like to acknowledge the American Association for Cancer Research and its financial and material support in the development of

the AACR Project GENIE registry, as well as members of the consortium for their commitment to data sharing. Interpretations are the responsibility of study authors.

AUTHOR CONTRIBUTIONS

MS designed the study and supervised the analyses. GMT conducted all the analyses. Both authors wrote the manuscript.

COMPETING INTEREST STATEMENT

None declared.

REFERENCES

1. Thiery, J. P., Acloque, H., Huang, R. Y. J. & Nieto, M. A. Epithelial-mesenchymal transitions in development and disease. *Cell* **139**, 871–890 (2009).
2. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J Clin Invest* **119**, 1420–1428 (2009).
3. Pastushenko, I. & Blanpain, C. EMT Transition States during Tumor Progression and Metastasis. *Trends in Cell Biology* **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 Computational Biology* **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. *Nature Communications* **10**, 5587 (2019).
8. Stemmler, M. P., Eccles, R. L., Brabletz, S. & Brabletz, T. Non-redundant functions of EMT transcription factors. *Nature Cell Biology* **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**, 349–361 (2012).
11. Xu, J. *et al.* A novel Ubc9 -dependent pathway regulates SIRT1- ER-α Axis and BRCA1-associated TNBC lung metastasis. *Integr Mol Med* **4**, (2017).
12. Nie, Z. *et al.* STAG2 loss-of-function mutation induces PD-L1 expression in U2OS cells. *Ann Transl Med* **7**, (2019).
13. Chang, C.-J. *et al.* p53 regulates epithelial-mesenchymal transition and stem cell properties through modulating miRNAs. *Nat Cell Biol* **13**, 317–323 (2011).
14. Stein, Y., Rotter, V. & Aloni-Grinstein, R. Gain-of-Function Mutant p53: All the Roads Lead to Tumorigenesis. *Int J Mol Sci* **20**, (2019).
15. Vasudevan, A. *et al.* Single-Chromosomal Gains Can Function as Metastasis Suppressors and Promoters in Colon Cancer. *Dev Cell* **52**, 413-428.e6 (2020).
16. Zhao, M., Liu, Y. & Qu, H. Expression of epithelial-mesenchymal transition-related genes increases with copy number in multiple cancer types. *Oncotarget* **7**, 24688–24699 (2016).

17. 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**, 609–620 (2016).
18. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).
19. Jin, X. *et al.* A metastasis map of human cancer cell lines. *Nature* **588**, 331–336 (2020).
20. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov* **7**, 818–831 (2017).
21. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med* **23**, 703–713 (2017).
22. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).
23. Rheinbay, E. The genomic landscape of advanced cancer. *Nat Cancer* **1**, 372–373 (2020).
24. Aiello, N. M. *et al.* EMT subtype influences epithelial plasticity and mode of cell migration. *Dev Cell* **45**, 681-695.e4 (2018).
25. Jolly, M. K. *et al.* Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Front Oncol* **5**, 155 (2015).
26. Plygawko, A. T., Kan, S. & Campbell, K. Epithelial–mesenchymal plasticity: emerging parallels between tissue morphogenesis and cancer metastasis. *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**, 20200087 (2020).
27. Sabe, H. Cancer early dissemination: cancerous epithelial-mesenchymal transdifferentiation and transforming growth factor β signalling. *J Biochem* **149**, 633–639 (2011).

28. 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).

29. Jing, Y., Han, Z., Zhang, S., Liu, Y. & Wei, L. Epithelial-Mesenchymal Transition in tumor microenvironment. *Cell Biosci* **1**, 29 (2011).

30. 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).

31. 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).

32. Zhang, T., Suo, C., Zheng, C. & Zhang, H. Hypoxia and Metabolism in Metastasis. *Adv Exp Med Biol* **1136**, 87–95 (2019).

33. Aggarwal, V., Montoya, C. A., Donnenberg, V. S. & Sant, S. Interplay between tumor microenvironment and partial EMT as the driver of tumor progression. *iScience* **24**, 102113 (2021).

34. Tyler, M. & Tirosh, I. Decoupling epithelial-mesenchymal transitions from stromal profiles by integrative expression analysis. *Nature Communications* **12**, 2592 (2021).

35. 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).

36. San Juan, B. P., Garcia-Leon, M. J., Rangel, L., Goetz, J. G. & Chaffer, C. L. The Complexities of Metastasis. *Cancers (Basel)* **11**, (2019).

37. Mani, S. A. *et al.* The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **133**, 704–715 (2008).

746 38. Ganem, N. J., Godinho, S. A. & Pellman, D. A mechanism linking extra centrosomes to
747 chromosomal instability. *Nature* **460**, 278–282 (2009).
748 39. Lingle, W. L. *et al.* Centrosome amplification drives chromosomal instability in breast
749 tumor development. *Proc Natl Acad Sci U S A* **99**, 1978–1983 (2002).
750 40. de Almeida, B. P., Vieira, A. F., Paredes, J., Bettencourt-Dias, M. & Barbosa-Morais, N.
751 L. Pan-cancer association of a centrosome amplification gene expression signature with
752 genomic alterations and clinical outcome. *PLoS Comput Biol* **15**, e1006832 (2019).
753 41. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer.
754 *Nature* **578**, 94–101 (2020).
755 42. Jing, L. *et al.* CNOT3 contributes to cisplatin resistance in lung cancer through
756 inhibiting RIPK3 expression. *Apoptosis* **24**, 673–685 (2019).
757 43. Shirai, Y.-T. *et al.* CNOT3 targets negative cell cycle regulators in non-small cell lung
758 cancer development. *Oncogene* **38**, 2580–2594 (2019).
759 44. Zhang, S. *et al.* Loss of VHL expression contributes to epithelial-mesenchymal
760 transition in oral squamous cell carcinoma. *Oral Oncol* **50**, 809–817 (2014).
761 45. Duah, E. *et al.* Cysteinyl leukotriene 2 receptor promotes endothelial permeability,
762 tumor angiogenesis, and metastasis. *PNAS* **116**, 199–204 (2019).
763 46. Liu, W., Xin, H., Eckert, D. T., Brown, J. A. & Gnarra, J. R. Hypoxia and cell cycle
764 regulation of the von Hippel–Lindau tumor suppressor. *Oncogene* **30**, 21–31 (2011).
765 47. Meyer-Schaller, N. *et al.* A Hierarchical Regulatory Landscape during the Multiple
766 Stages of EMT. *Dev Cell* **48**, 539-553.e6 (2019).
767 48. Yamaguchi, H. & Taouk, G. M. A Potential Role of YAP/TAZ in the Interplay Between
768 Metastasis and Metabolic Alterations. *Front. Oncol.* **10**, (2020).

769 49. Cook, D. P. & Vanderhyden, B. C. Context specificity of the EMT transcriptional
770 response. *Nature Communications* **11**, 2142 (2020).
771 50. Zhang, M., Wang, J. & Guo, J. Role of Regenerating Islet-Derived Protein 3A in
772 Gastrointestinal Cancer. *Front. Oncol.* **0**, (2019).
773 51. Larue, L. & Bellacosa, A. Epithelial–mesenchymal transition in development and
774 cancer: role of phosphatidylinositol 3' kinase/AKT pathways. *Oncogene* **24**, 7443–7454
775 (2005).
776 52. Mu, Z. *et al.* AZD8931, an equipotent, reversible inhibitor of signaling by epidermal
777 growth factor receptor (EGFR), HER2, and HER3: preclinical activity in HER2 non-amplified
778 inflammatory breast cancer models. *J Exp Clin Cancer Res* **33**, 47 (2014).
779 53. Roche, J. The Epithelial-to-Mesenchymal Transition in Cancer. *Cancers (Basel)* **10**,
780 (2018).
781 54. Tripathi, S., Chakraborty, P., Levine, H. & Jolly, M. K. A mechanism for epithelial-
782 mesenchymal heterogeneity in a population of cancer cells. *PLoS Comput Biol* **16**,
783 e1007619 (2020).
784 55. Simeonov, K. P. *et al.* Single-cell lineage tracing of metastatic cancer reveals selection
785 of hybrid EMT states. *Cancer Cell* (2021) doi:10.1016/j.ccell.2021.05.005.
786 56. Brown, M. S. *et al. Dynamic plasticity within the EMT spectrum, rather than static*
787 *mesenchymal traits, drives tumor heterogeneity and metastatic progression of breast*
788 *cancers*. http://biorxiv.org/lookup/doi/10.1101/2021.03.17.434993 (2021)
789 doi:10.1101/2021.03.17.434993.
790 57. Muñoz, D. P. *et al.* Activation-induced cytidine deaminase (AID) is necessary for the
791 epithelial–mesenchymal transition in mammary epithelial cells. *PNAS* **110**, E2977–E2986
792 (2013).

58. Bader, S. B. *et al.* Replication catastrophe induced by cyclic hypoxia leads to increased APOBEC3B activity. *Nucleic Acids Research* **49**, 7492–7506 (2021).

59. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31**, 213–219 (2013).

60. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44**, e71–e71 (2016).

61. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

62. Chae, Y. K. *et al.* Epithelial-mesenchymal transition (EMT) signature is inversely associated with T-cell infiltration in non-small cell lung cancer (NSCLC). *Scientific Reports* **8**, 2918 (2018).

63. 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).

64. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

65. 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).

66. Jiménez-Sánchez, A., Cast, O. & Miller, M. L. Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods. *Cancer Res* **79**, 6238–6246 (2019).

67. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).

68. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676-689.e3 (2018).

69. Bhandari, V. *et al.* Molecular landmarks of tumor hypoxia across cancer types. *Nature Genetics* **51**, 308–318 (2019).

70. 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**, 428–435 (2010).

71. Miranda, A. *et al.* Cancer stemness, intratumoral heterogeneity, and immune response across cancers. *Proc Natl Acad Sci U S A* **116**, 9020–9029 (2019).

72. 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 Biology* **17**, 31 (2016).

73. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).

74. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).

75. Dempster, J. M. *et al.* Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines. *bioRxiv* 720243 (2019) doi:10.1101/720243.

76. Chen, M.-J. M. *et al.* TCPA v3.0: An Integrative Platform to Explore the Pan-Cancer Analysis of Functional Proteomic Data. *Mol Cell Proteomics* **18**, S15–S25 (2019).

77. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

78. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

79. Corsello, S. M. *et al.* Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer* **1**, 235–248 (2020).

80. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters | OMICS: A Journal of Integrative Biology. https://www.liebertpub.com/doi/10.1089/omi.2011.0118.

81. 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

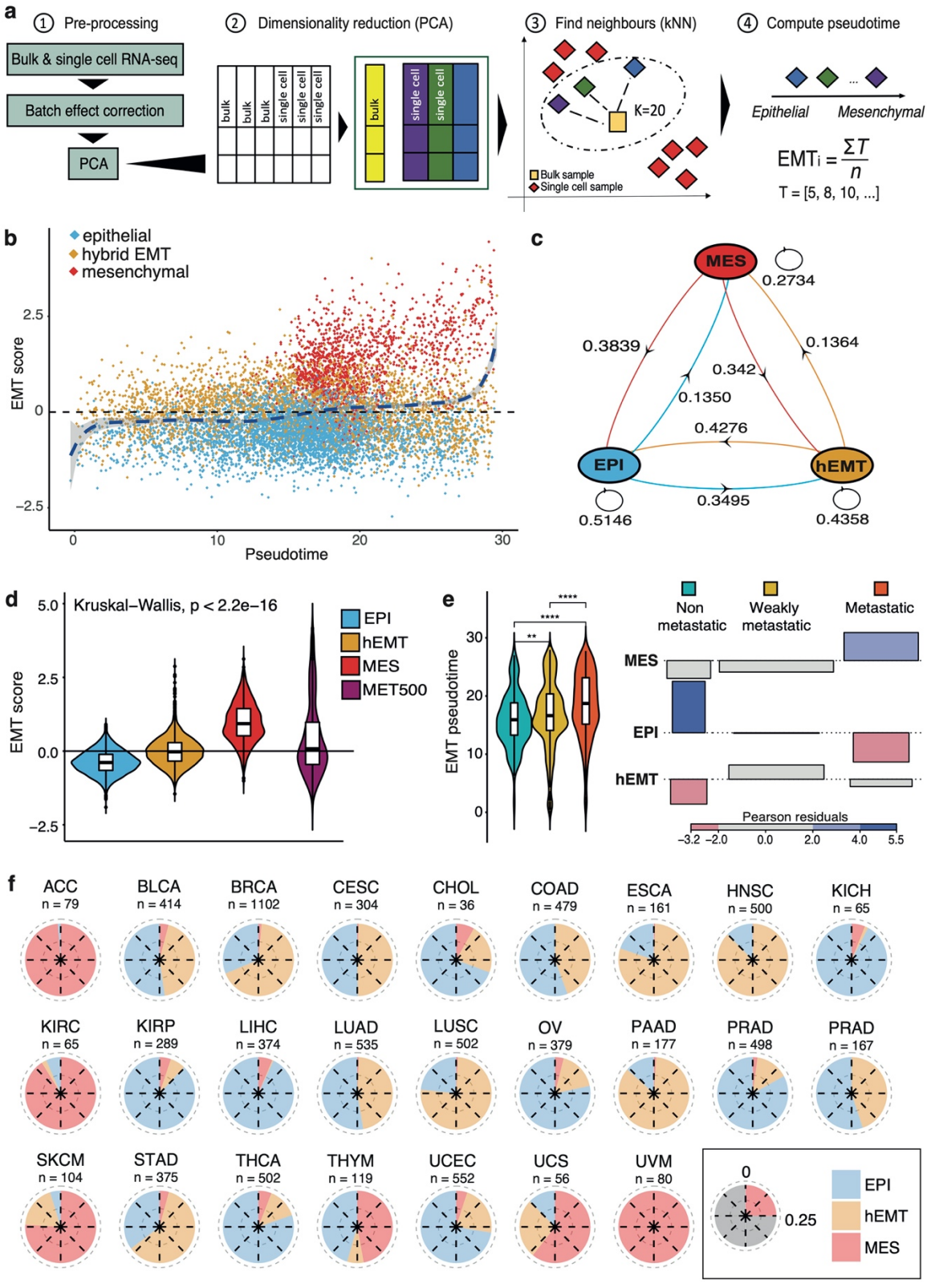


Figure 1. Pan-cancer EMT trajectories and underlying macro-states. (a) Workflow for reconstructing the EMT trajectories of TCGA samples. 1: Bulk and single cell datasets are

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 pseudotime values 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=2e-13). (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.

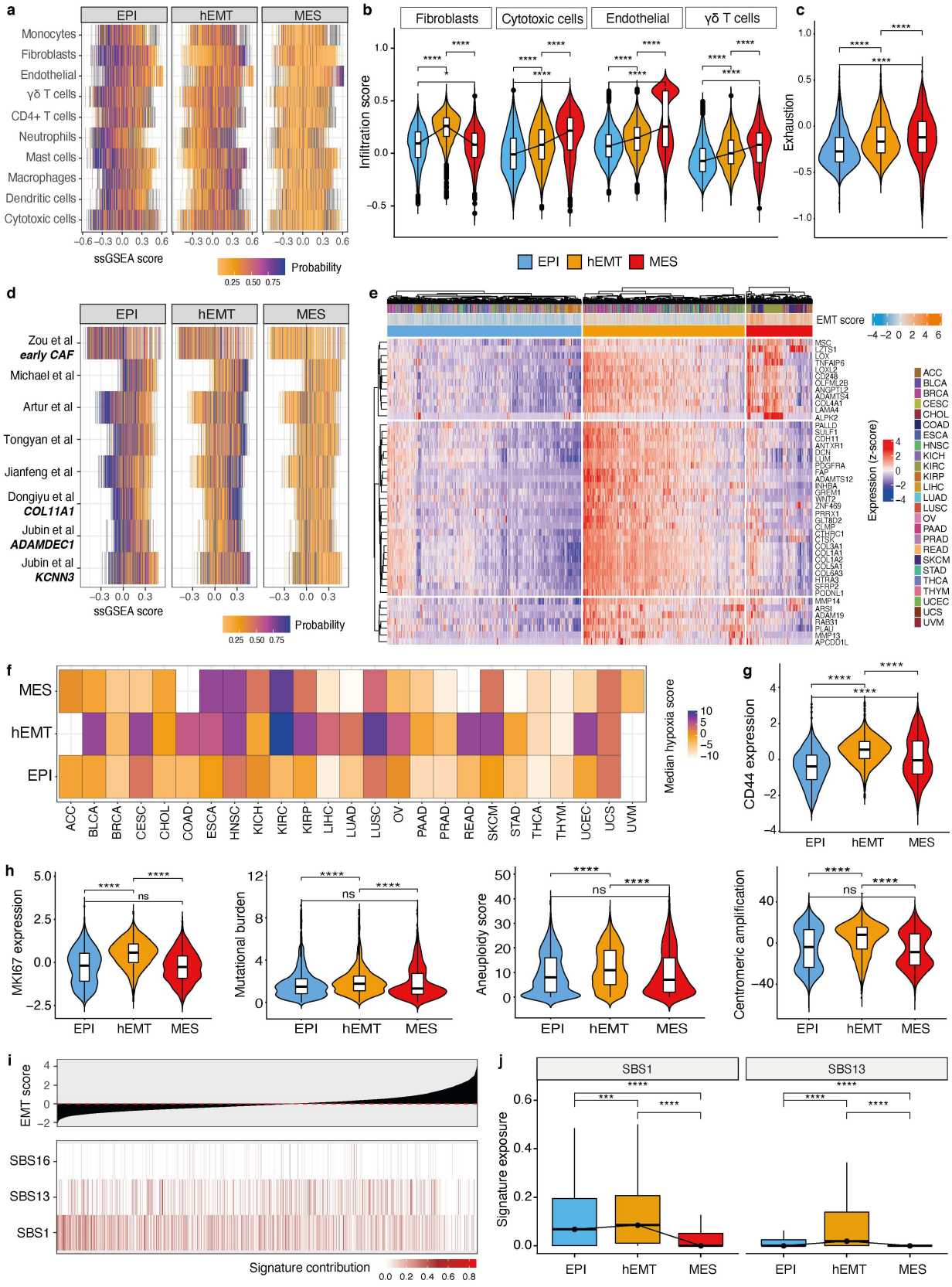


Figure 2. Tumour extrinsic and intrinsic hallmarks of EMT. (a) Heat map showcasing the results of a multinomial logistic regression model trained to predict EMT states based on cell

infiltration in the microenvironment. Each row corresponds to a cell type and the corresponding per-sample infiltration is highlighted via ssGSEA scores reported on the x axis. The values reported in the heat map are the probabilities that a sample should fall into the epithelial, hEMT or mesenchymal categories in relation to the ssGSEA score of a certain cell type. (b) Cell abundance compared across the EMT states for significantly predicted cell types in the multinomial analysis. (c) Levels of exhaustion quantified across the three EMT states. (d) Heat map displaying the results a multinomial logistic regression model predicting EMT state using fibroblast signatures from multiple publications. (e) Heat map displaying the pan-cancer expression of the fibroblast COL11A1_FS signature genes. Highest expression is observed in the hEMT state. (f) Median hypoxia values in the three different EMT states across tissues. (g) Gene expression levels of the stemness marker CD44 compared across the three EMT states. (h) Expression of the proliferation marker Ki67, mutational burden, aneuploidy, and centromeric amplification levels compared across the three EMT states. (i) 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. (j) Signature contributions from SBS1 and SBS13 compared between the three EMT states.

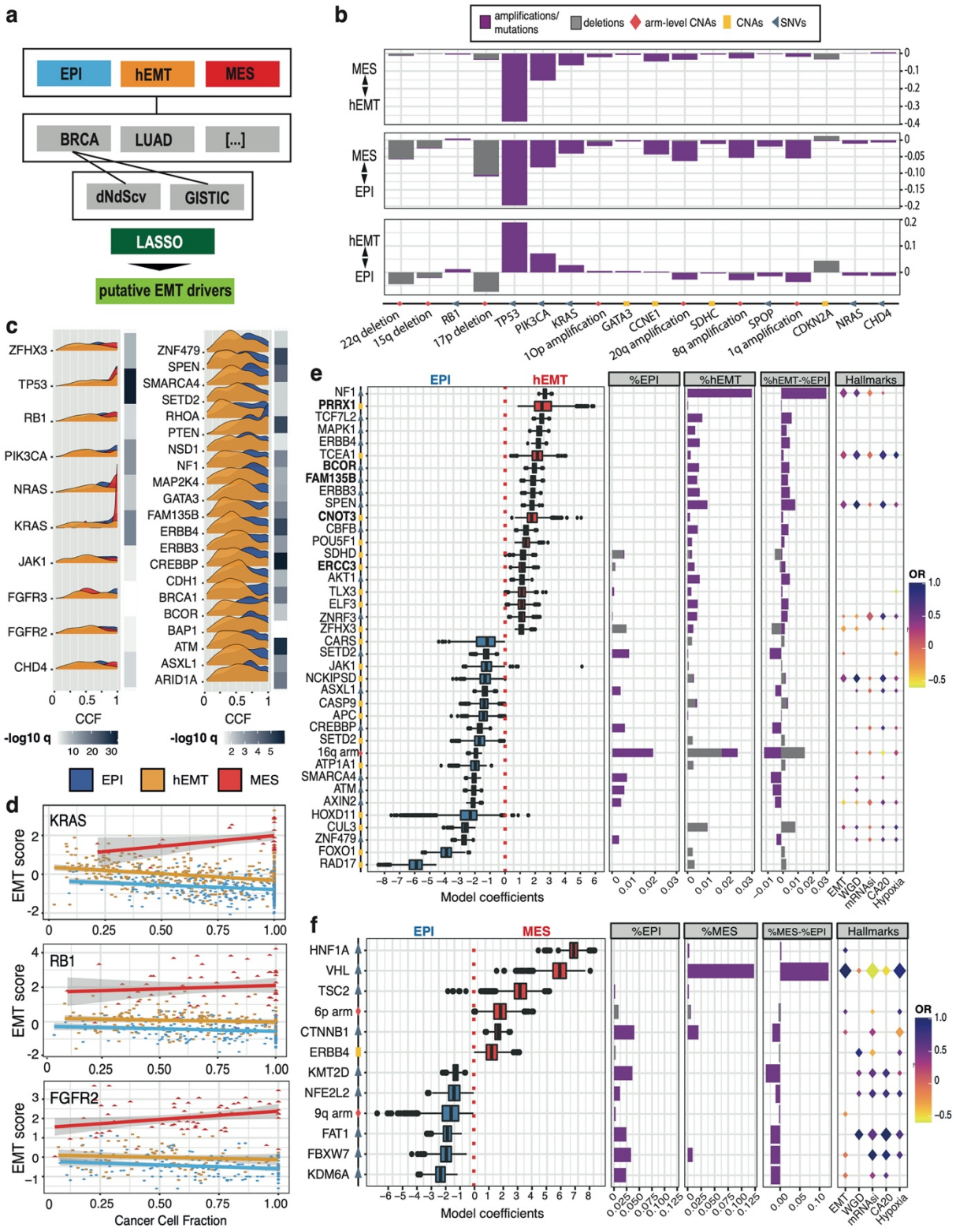
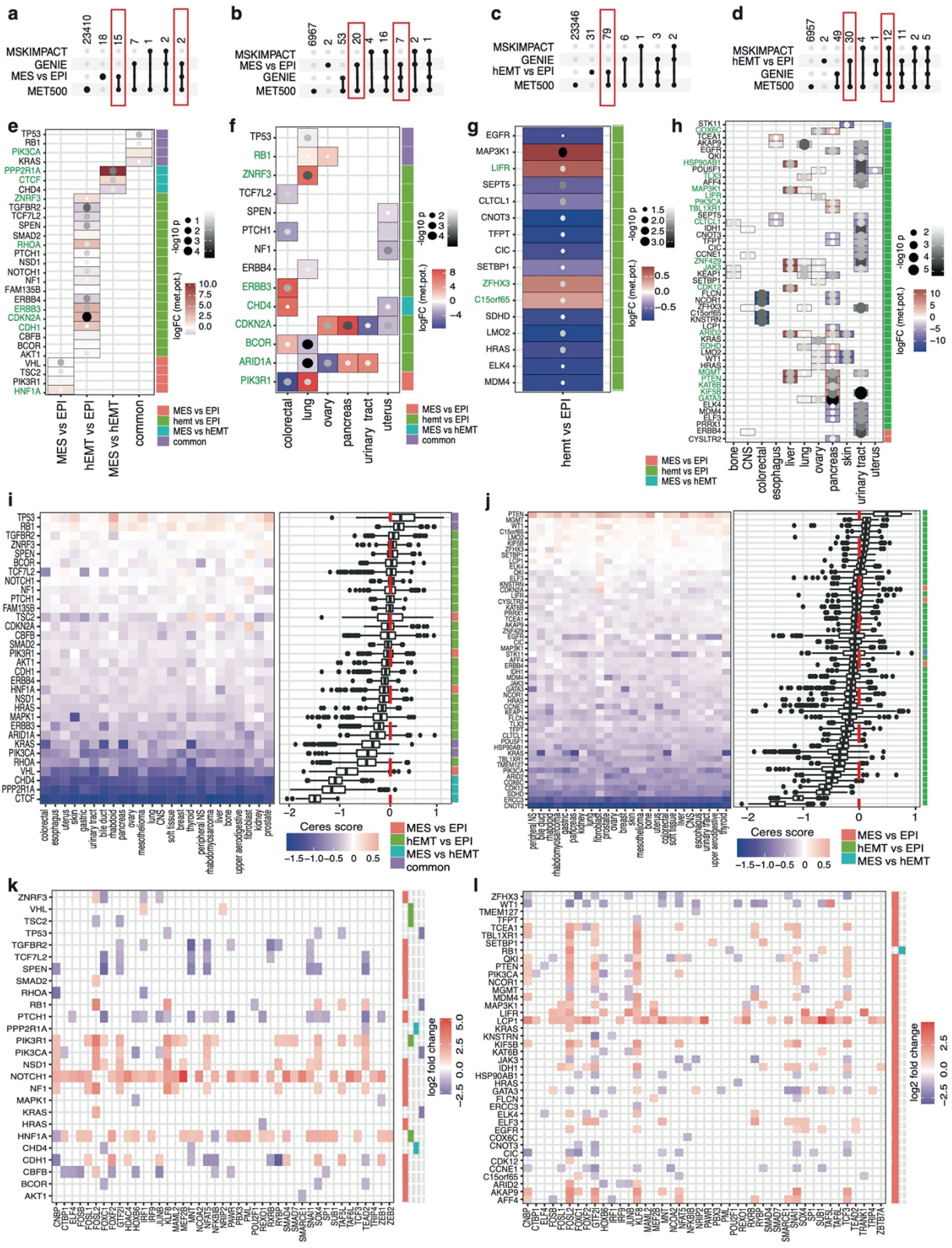


Figure 3. Genomic driver events linked with EMT. (a) The analytical workflow used to detect putative EMT biomarkers. For each state and cancer type, we used dNdScv and GISTIC to prioritise mutated genes and copy number events, respectively. These genomic

events were then employed as input for lasso modelling to classify EMT states. (b) Common genomic features distinguishing the transformed states (hEMT, mesenchymal) from epithelial. The individual bars depict the difference in the fraction of samples harbouring a specific genomic alteration between every pair of EMT states: MES vs hEMT, MES vs EPI, hEMT vs EPI. (c) Cancer cell fraction of common genomic markers between EMT states (left) and hEMT-specific markers (right) with significantly distinct distribution between the states (highlighted by the white-black colour gradient). (d) Scatter plot of EMT scores against the cancer cell fraction across TCGA for KRAS, RB1 and FGFR2. (e) Top-ranked genomic markers uniquely distinguishing hEMT from the epithelial state (notable ones highlighted in bold). The fraction of alterations in EPI and hEMT samples, as well as the difference, are displayed in the adjacent panels, The balloon chart on the right illustrates the association between each marker and EMT, genome doubling (WGD), stemness index (mRNAsi), centromeric amplification (CA20), and hypoxia. The size of the diamonds is proportional to the significance of association, the colours report the odds ratios. (f) List of the top-ranked genomic markers uniquely distinguishing the mesenchymal from the epithelial state and their associated hallmarks.



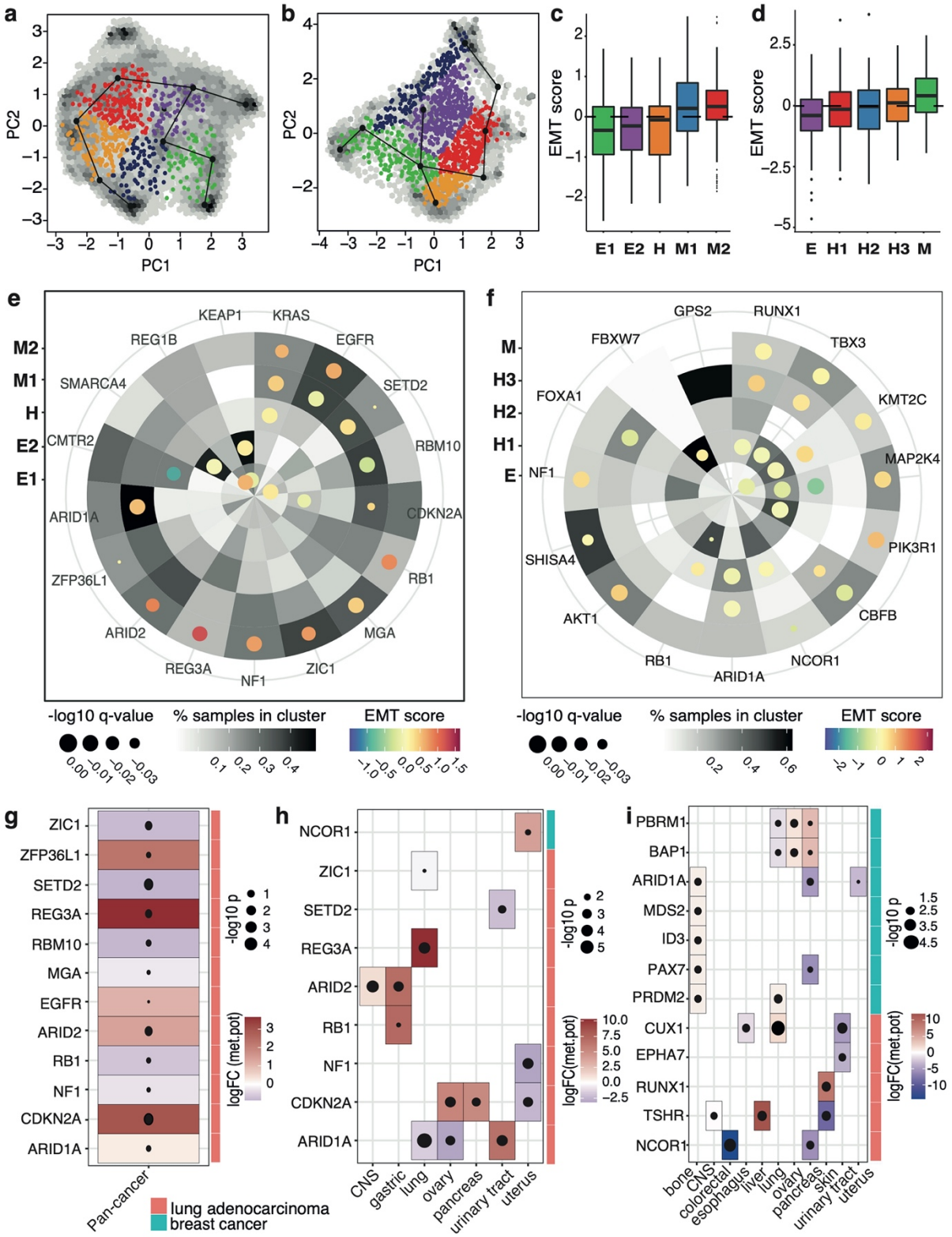
909

910 **Figure 4. Validation of the pan-cancer genomic associations with EMT in external**

911 **datasets.** (a) UpSet chart displaying the overlap with external datasets for copy number

912 events derived from the MES vs EPI comparison. (b) Overlap with external datasets for

913 mutated gene events derived from the MES vs EPI comparison. (c-d) Same as (a-b) but for
914 events derived from the hEMT vs EPI comparison. (e) Fold changes in metastatic potential
915 across all the cell lines from CCLE harbouring point mutations in markers genes of EMT,
916 compared to that of cell lines without the respective mutation. The size and the colours of
917 the dots highlight the significance of the association ($p<0.05$). Events which increase
918 metastatic potential are highlighted in green. (*f*) Similar to (*e*), but with the analysis performed
919 at tissue level, ($p<0.05$). (*g*-*h*) Similar to (*e*-*f*), but for biomarkers harbouring copy number
920 alterations. (*i*) CERES essentiality scores from DepMap in individual cell lineages for genes
921 harbouring mutations linked with EMT. Negative values indicate increased essentiality. The
922 boxplots on the right indicate the CERES score distribution across all lineages. (*j*) Similar to (*i*)
923 but considering the genes linked with EMT via copy number alteration. (*k*) Genes showing
924 pan-cancer associations with EMT (rows) that are dysregulated as a result of knocking down
925 transcription factors relevant for EMT (columns). The colour gradient reflects the log2 fold
926 change in expression of the gene upon transcription factor knockdown (adjusted $p<0.05$). (*l*)
927 Similar to (*k*) but considering putative EMT biomarkers with alterations in copy number.



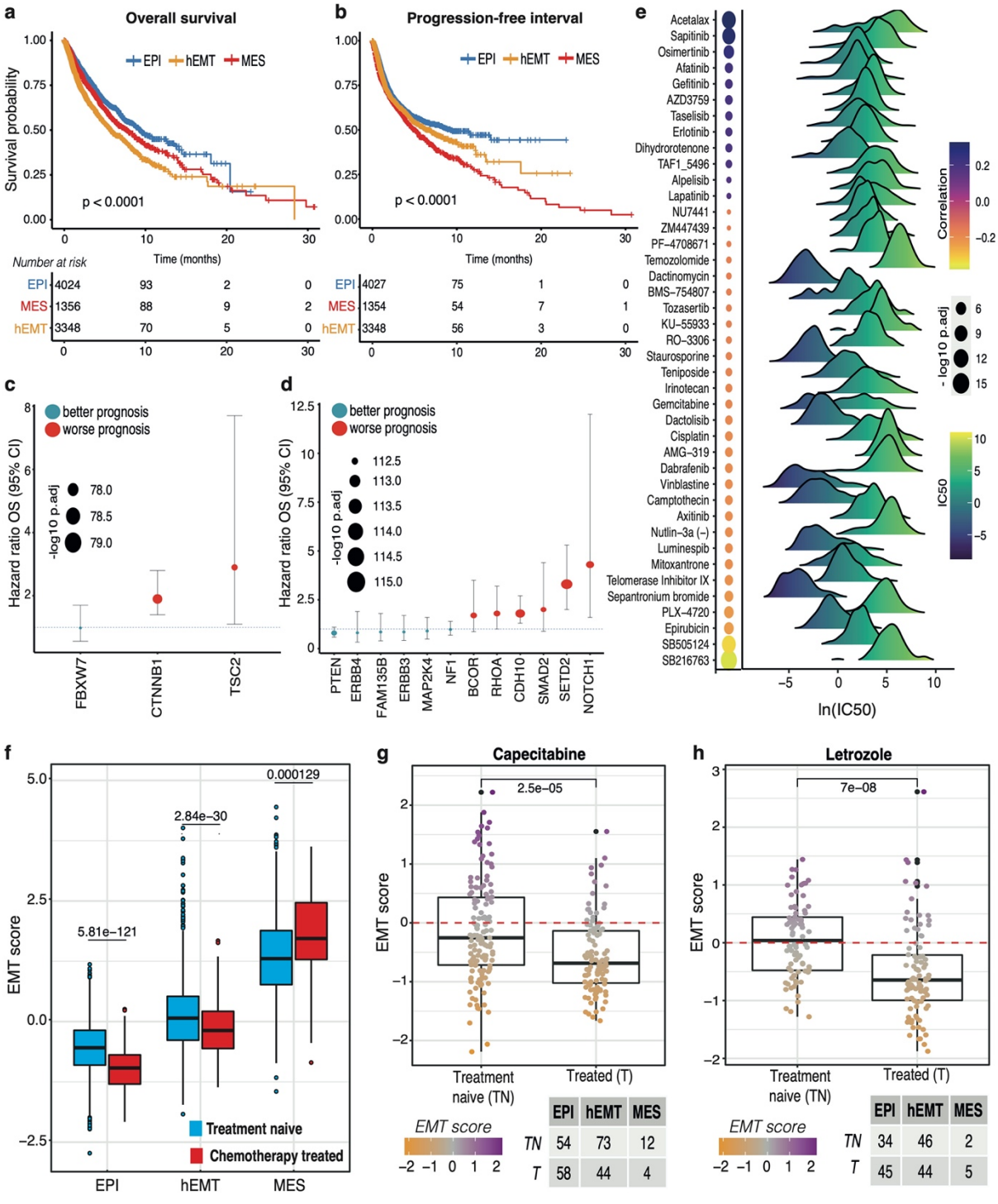
928

929 **Figure 5. EMT trajectories and associated genomic hallmarks in lung and breast**

930 **cancer.** (a) Lung adenocarcinoma samples from TCGA mapped onto the single cell EMT

931 trajectories derived in the A579 cell line. The first two principal components of the scRNA-seq

data are depicted. The hexagons represent the density of scRNA-seq samples, the TCGA samples are coloured according to their membership to an EMT cluster. The red EMT trajectory was computed on the scRNA-seq data. (b) Similar to (a), projecting breast cancer samples from TCGA onto scRNA-seq from the MCF7 cell line. (c-d) EMT scores compared across the clusters defined from (a-b) for lung and breast cancer, respectively. E1-3, H, M1-2 represent epithelial, hEMT and mesenchymal states in increasing order of transformation (sorted by median). (e) Radial plot highlighting positively selected genes in each EMT state in lung adenocarcinoma. Tile colours highlight the proportion of samples with a given mutated gene. Each circle depicts the EMT score (colour) and the level of significance of association between each driver gene and each cluster (size). (e) Similar to (f) but for the breast cancer samples. (g) Fold change in metastatic potential across all the cell lines harbouring mutations in marker genes of EMT in lung, compared with that of cell lines without the respective mutation. The size and the colours of the dots highlight the significance of the association ($p<0.05$). (h) Similar to (g), but with the analysis performed at tissue level ($p<0.05$). (i) Similar to (h), but for biomarkers harbouring copy number alterations.



947

948 **Figure 6. Clinical relevance of the EMT states.** (a) Overall survival compared between
949 MES, hEMT and EPI samples. (b) Progression free interval compared between the three
950 groups. (c) Mutated markers of EMT pan-cancer with a significantly worse or improved
951 outcome between the mesenchymal and epithelial states (q<0.001). (d) Mutated markers of
952 EMT pan-cancer with a significantly worse or improved outcome between the hEMT and

epithelial states. (e) Correlation between the EMT scores and IC50 values in cell lines from GDSC treated with various drugs. The balloon chart on the left illustrates the association between the IC50 for each compound and EMT. The size of the diamonds is proportional to the significance of association. The IC50 ranges for all cell lines are depicted by the density charts. (f) EMT scores compared between treatment naïve samples and those collected after chemotherapy, by EMT state. (g-h) EMT scores compared between (unmatched) samples before and after the treatment with capecitabine and letrozole, respectively. The tables indicate the number of samples in each category.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.pdf