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

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

63

64 **RESULTS**

65 **Pan-cancer reconstruction of EMT trajectories from transcriptomics data**

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

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

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

105

106

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

108 To characterise the dominant EMT states governing the continuum of transcriptional activity
109 described above, we discretised the pseudotime trajectory based on expression values of
110 canonical EMT markers using a Hidden Markov Model approach and uncovered three macro-
111 states: epithelial (EPI), hybrid EMT (hEMT) and mesenchymal (MES) (Figure 1b-c,
112 Supplementary Figure S1b). These states were robust to varying levels of gene expression
113 noise (Supplementary Figure S1e). As expected, the probability for the cancer cells to switch
114 from the epithelial to the hEMT (0.32) state was higher than the probability to passage directly
115 into the mesenchymal state (0.09). The hEMT tumours tended to remain in the same state
116 42% of the times, suggesting this state could be more stable than anticipated – as previously
117 stipulated[22] and consistent with observations that a fully mesenchymal state is not always
118 observed[23].

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

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

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

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

139 The EMT classification was significantly correlated with the clinical cancer stage (Chi-square
140 test $p<<0.0001$, Supplementary Figure 2d), with transformed samples (hEMT/MES) found to
141 be 1.3-fold and 1.2-fold enriched in late-stage tumours, respectively, while the epithelial state
142 was 1.4-fold overrepresented in early-stage cancers (Figure S1i). In primary tumours, 12% of
143 the profiled samples were classified as fully transformed (MES), with the majority of them
144 (60%) annotated as late-stage tumours (Supplementary Table S2). Notably, metastatic
145 samples available from TCGA (n=343) were overwhelmingly classed as MES (94%,
146 Supplementary Figure 2e), suggesting that the transformed phenotype is more pronounced in
147 metastases than in primary tumours, as expected. While the correlation between cancer stage
148 and EMT state does not appear as strong as potentially anticipated in primary tumours, the
149 proportion of observed late-stage cancers increases as we move from EPI to hEMT and MES
150 cancers in most cancer tissues, with mesenchymal cholangiocarcinomas, esophageal and
151 kidney chromophobe cancers being entirely late stage (Supplementary Fig 2f). The fact that
152 some early stage cancers are classified as fully mesenchymal (5%) may suggest early
153 evidence for the phenotypic transformation required for metastasis. Indeed, multiple studies
154 have demonstrated the activation of the EMT transcriptional programme in the early stages of
155 cancer[10,25]. Even the hEMT phenotype was hypothesised to be sufficient for promoting
156 metastatic dissemination[26], although this is likely tissue-dependent.

157

158 **Tumour cell extrinsic hallmarks of EMT**

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

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

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

187 may play a greater role in attaining the hEMT state compared to the fixation of a fully
188 mesenchymal phenotype.
189
190 **Spatially-resolved EMT patterning reveals local microenvironmental effects**

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

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

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

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

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

241 **EMT diversity in single cell data**

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

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

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

269 cancers (Figure 4d). These observations, along with the spatial transcriptomics data, suggest
270 that the relation between EMT transformation of tumour cells and their interactions with the
271 TME is likely a complex one, highly tissue specific and driven by local spatial effects. Ideally,
272 single cell, spatially resolved longitudinal datasets would be needed to fully resolve such
273 heterogeneity.
274
275 **Tumour cell intrinsic hallmarks of EMT**

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

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

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

307

308 **Genomic driver events underlying the EMT transformation pan-cancer**

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

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

323 machine learning framework to identify those drivers able to discriminate between EPI, hEMT
324 and MES states across cancers, while accounting for tissue-specific effects (Methods, Figure
325 5e). The developed models were validated using several other machine learning approaches
326 and demonstrated remarkably high accuracies of 92-97% in distinguishing the fully
327 transformed state from either the hybrid or mesenchymal one (Supplementary Figures S6a-
328 b,d-e,g-h). Lower performance was obtained for the model discriminating between hEMT and
329 EPI (~62-73%, Supplementary Figures 6c,f,i), which is not surprising due to the intermediate,
330 hybrid nature of the former, but is still useful in understanding weaker effects on EMT
331 transformation.

332 Among the genomic biomarkers able to discriminate transformed tumours (hEMT, MES) from
333 the epithelial state, we identified genes that have been previously linked with cell migration,
334 invasion and EMT, such as *RB1*, *VHL*, *ERBB2, ARHGAP26*, *PRDM1, APC* (Figures 5f-g ,
335 Supplementary Tables S4, S5). *RB1*, a key cell cycle regulator, has been shown to promote
336 EMT in conjunction with p53 in triple negative breast cancer[43], while *VHL* alterations
337 contribute to EMT via regulation of hypoxia[44]. Larger scale events included deletions of the 4p,
338 6p and 17p chromosomal arms, all of which harboured cancer drivers which have been
339 previously linked with EMT, e.g. *FGFR3* on 4p[45], *DAXX* and *TRIM27* on 6p[46,47], *TP53* on 17p[12]
340 (Supplementary Table S4). Deletions of the 4p arm appeared in the majority of lung
341 squamous cell and esophageal carcinomas (58% and 50%, respectively), while 6p arm
342 deletions were most frequent in pancreatic, esophageal cancers and adrenocortical
343 carcinomas (>20% in each). 17p arm deletions were the most abundant, especially in ovarian
344 (76%) and kidney chromophobe cancers (76%), with an average of 37% of cases affected per
345 tissue. Therefore, no strong bias in terms of cancer type was observed for these large-scale
346 alterations. In addition to these, events less strongly linked with metastatic transformation
347 were also uncovered, such as mutations in *CHIC2*, encoding for a protein with a cysteine-rich
348 hydrophobic domain occasionally implicated in leukemia, or amplifications of the *ELL* gene, an
349 elongation factor for polymerase II. While the genomic hallmarks distinguishing the extremes

350 of mesenchymal transformation (MES versus EPI) were predominantly classical cancer
351 drivers involved in the most fundamental processes (e.g. cell cycle) (Supplementary Figure
352 6j), the ones distinguishing fully transformed from hybrid phenotypes were more clearly linked
353 with cell migration, including processes of cytoskeletal regulation, cell adhesion and T cell
354 signalling (Supplementary Figure 6k).

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

365

366 **Validation of genomic associations**

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

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

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

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

400 **Clinical relevance of EMT**

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

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

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

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

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

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

433

434 **DISCUSSION**

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

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

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

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

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

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

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

494

495 **METHODS**

496 **Data sources**

497 Bulk RNA-sequencing, copy number (segment file and focal alterations), somatic variants
498 (MuTect[59]), molecular subtypes and clinical data were retrieved for 8,778 primary tumours of
499 epithelial origin from the harmonized version of TCGA using the *TCGAbiolinks* R package[60].
500 Based on tumour purity estimates reported by Hoadley et al[61] samples with purity lower than
501 30% were removed leaving 7,180 samples. All other data sources employed for validation are
502 described below.

503 **Reconstruction of EMT trajectories in bulk data**

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

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

532 **Segmentation of the EMT trajectory and robustness evaluation**

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

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

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

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

559 **EMT pseudotime reconstruction with adjustment for TME contamination**

560 To account for confounding expression signals coming from non-tumour cells in the
561 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

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

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

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

602 **Tumour microenvironment quantification**

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

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

609 **Spatial transcriptomics data analysis**

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

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

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

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

634 *Gene module scores*

635 An EMT score was calculated per spot by adapting the method used to assign a score to the
636 TCGA samples, using only the breast cancer cell line scRNA-seq data. The EMT scRNA-seq
637 trajectory was mapped onto each spot within the spatial transcriptomic slide, and the mean of
638 the pseudotime values of the single cell samples detected by the kNN algorithm was used.
639 This was performed on multiple breast cancer cell lines and the average pseudotime across
640 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

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

675 **Genomic hallmark quantification**

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

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

689 **Mutational signature analysis**

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

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

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

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

709 $$EMT_score \sim SBS + (1|cancer)$$

710 **Prioritisation of genomic alterations in TCGA**

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

718 Copy number events were obtained using the *TCGAbiolinks* R package. Chromosomal arm-
719 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:

746 $$CCF_i = \left(2 + \frac{[purity * (CN_i - 2)]}{purity}\right) \cdot VAF_i ,$$

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

750 **Validation of genomic events linked with EMT**

751 Three large scale public siRNA screens were employed for experimental validation of the
752 proposed genomics associations with EMT. The first dataset from Koedoot et al[51] looked at
753 gene knockdown effects on cell migration abilities in the Hs578T (top panel) and MDA-MB-
754 231 breast cancer cell lines. The data were obtained from the associated publication and
755 contained detailed measurements of effects on cellular phenotype upon knockdown,
756 quantified as changes in cell net surface area, length of minor and major axes, axis ratio
757 (large/small: elongated cells, close to 1: round cells), perimeter score (larger – more
758 migration). Data from further phenotypic tests containing confirmed morphology (big/small
759 round cells) were also available on a subset of the genes.

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

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

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

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

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

779 **Drug response datasets**

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

785 **Gene ontology analysis**

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

788 **Survival analysis**

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

795

796 **Data visualization and basic statistics**

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

799 **Data availability**

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

806 **Code availability**

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

809

810 **ACKNOWLEDGEMENTS**

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

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

819

820 **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).

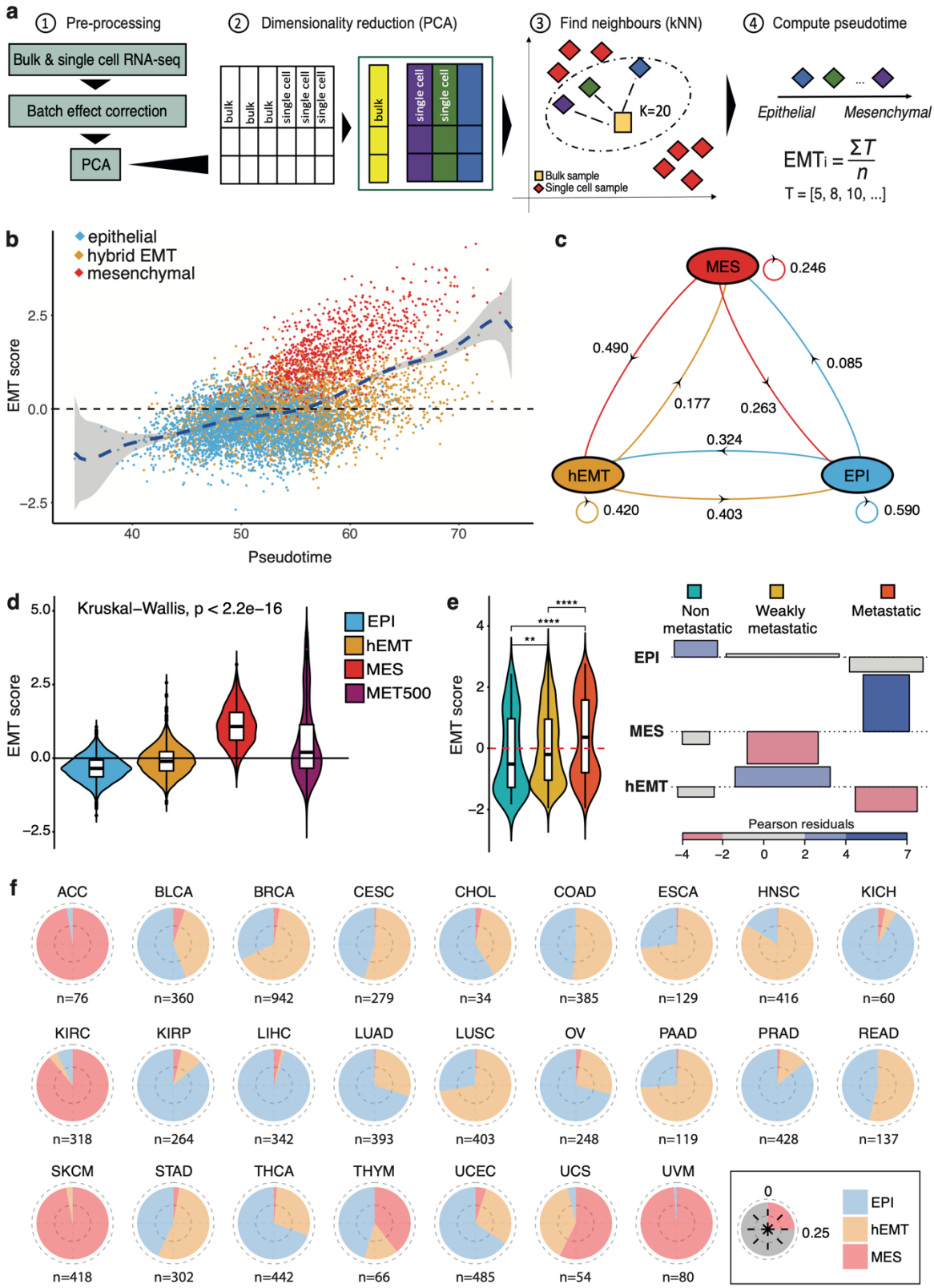
857 13. Vasudevan, A. *et al.* Single-Chromosomal Gains Can Function as Metastasis
858 Suppressors and Promoters in Colon Cancer. *Dev Cell* **52**, 413-428.e6 (2020).
859 14. Zhao, M., Liu, Y. & Qu, H. Expression of epithelial-mesenchymal transition-related
860 genes increases with copy number in multiple cancer types. *Oncotarget* **7**, 24688-99
861 (2016).
862 15. Mak, M.P. *et al.* A Patient-Derived, Pan-Cancer EMT Signature Identifies Global
863 Molecular Alterations and Immune Target Enrichment Following Epithelial-to-
864 Mesenchymal Transition. *Clin Cancer Res* **22**, 609-20 (2016).
865 16. Robinson, D.R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**,
866 297-303 (2017).
867 17. Jin, X. *et al.* A metastasis map of human cancer cell lines. *Nature* **588**, 331-336 (2020).
868 18. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-
869 754 (2016).
870 19. Rheinbay, E. The genomic landscape of advanced cancer. *Nat Cancer* **1**, 372-373
871 (2020).
872 20. Tyler, M. & Tirosh, I. Decoupling epithelial-mesenchymal transitions from stromal
873 profiles by integrative expression analysis. *Nat Commun* **12**, 2592 (2021).
874 21. Aiello, N.M. *et al.* EMT Subtype Influences Epithelial Plasticity and Mode of Cell
875 Migration. *Dev Cell* **45**, 681-695.e4 (2018).
876 22. Jolly, M.K. *et al.* Implications of the Hybrid Epithelial/Mesenchymal Phenotype in
877 Metastasis. *Front Oncol* **5**, 155 (2015).
878 23. Plygawko, A.T., Kan, S. & Campbell, K. Epithelial-mesenchymal plasticity: emerging
879 parallels between tissue morphogenesis and cancer metastasis. *Philos Trans R Soc*
880 *Lond B Biol Sci* **375**, 20200087 (2020).
881 24. Wang, Q., Liu, G. & Hu, C. Molecular Classification of Gastric Adenocarcinoma.
882 *Gastroenterology Res* **12**, 275-282 (2019).
883 25. Sabe, H. Cancer early dissemination: cancerous epithelial-mesenchymal
884 transdifferentiation and transforming growth factor β signalling. *J Biochem* **149**, 633-9
885 (2011).
886 26. Jolly, M.K., Ware, K.E., Gilja, S., Somarelli, J.A. & Levine, H. EMT and MET: necessary or
887 permissive for metastasis? *Mol Oncol* **11**, 755-769 (2017).
888 27. Jing, Y., Han, Z., Zhang, S., Liu, Y. & Wei, L. Epithelial-Mesenchymal Transition in tumor
889 microenvironment. *Cell Biosci* **1**, 29 (2011).
890 28. Gibbons, D.L. & Creighton, C.J. Pan-cancer survey of epithelial-mesenchymal transition
891 markers across the Cancer Genome Atlas. *Dev Dyn* **247**, 555-564 (2018).
892 29. Choi, B.J., Park, S.A., Lee, S.Y., Cha, Y.N. & Surh, Y.J. Hypoxia induces epithelial-
893 mesenchymal transition in colorectal cancer cells through ubiquitin-specific protease
894 47-mediated stabilization of Snail: A potential role of Sox9. *Sci Rep* **7**, 15918 (2017).
895 30. Emami Nejad, A. *et al.* The role of hypoxia in the tumor microenvironment and
896 development of cancer stem cell: a novel approach to developing treatment. *Cancer*
897 *Cell Int* **21**, 62 (2021).
898 31. San Juan, B.P., Garcia-Leon, M.J., Rangel, L., Goetz, J.G. & Chaffer, C.L. The
899 Complexities of Metastasis. *Cancers (Basel)* **11**(2019).
900 32. Mani, S.A. *et al.* The epithelial-mesenchymal transition generates cells with properties
901 of stem cells. *Cell* **133**, 704-15 (2008).
902 33. Andersson, A. *et al.* Spatial deconvolution of HER2-positive breast cancer delineates
903 tumor-associated cell type interactions. *Nat Commun* **12**, 6012 (2021).

904 34. Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an
905 unexplored landscape of heterogeneity. *Nat Commun* **9**, 2419 (2018).
906 35. Ardila, D.C. *et al.* Identifying Molecular Signatures of Distinct Modes of Collective
907 Migration in Response to the Microenvironment Using Three-Dimensional Breast
908 Cancer Models. *Cancers (Basel)* **13**(2021).
909 36. Chung, W. *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell
910 profiling in primary breast cancer. *Nat Commun* **8**, 15081 (2017).
911 37. Qian, J. *et al.* A pan-cancer blueprint of the heterogeneous tumor microenvironment
912 revealed by single-cell profiling. *Cell Res* **30**, 745-762 (2020).
913 38. Ganem, N.J., Godinho, S.A. & Pellman, D. A mechanism linking extra centrosomes to
914 chromosomal instability. *Nature* **460**, 278-82 (2009).
915 39. Lingle, W.L. *et al.* Centrosome amplification drives chromosomal instability in breast
916 tumor development. *Proc Natl Acad Sci U S A* **99**, 1978-83 (2002).
917 40. de Almeida, B.P., Vieira, A.F., Paredes, J., Bettencourt-Dias, M. & Barbosa-Morais, N.L.
918 Pan-cancer association of a centrosome amplification gene expression signature with
919 genomic alterations and clinical outcome. *PLoS Comput Biol* **15**, e1006832 (2019).
920 41. Alexandrov, L.B. *et al.* The repertoire of mutational signatures in human cancer.
921 *Nature* **578**, 94-101 (2020).
922 42. Muñoz, D.P. *et al.* Activation-induced cytidine deaminase (AID) is necessary for the
923 epithelial-mesenchymal transition in mammary epithelial cells. *Proc Natl Acad Sci U S*
924 *A* **110**, E2977-86 (2013).
925 43. Jiang, Z. *et al.* RB1 and p53 at the crossroad of EMT and triple-negative breast cancer.
926 *Cell Cycle* **10**, 1563-70 (2011).
927 44. Liu, W., Xin, H., Eckert, D.T., Brown, J.A. & Gnarra, J.R. Hypoxia and cell cycle
928 regulation of the von Hippel-Lindau tumor suppressor. *Oncogene* **30**, 21-31 (2011).
929 45. Li, L., Zhang, S., Li, H. & Chou, H. FGFR3 promotes the growth and malignancy of
930 melanoma by influencing EMT and the phosphorylation of ERK, AKT, and EGFR. *BMC*
931 *Cancer* **19**, 963 (2019).
932 46. Wu, C. *et al.* DAXX inhibits cancer stemness and epithelial-mesenchymal transition in
933 gastric cancer. *Br J Cancer* **122**, 1477-1485 (2020).
934 47. Zhang, Y. *et al.* TRIM27 functions as an oncogene by activating epithelial-mesenchymal
935 transition and p-AKT in colorectal cancer. *Int J Oncol* **53**, 620-632 (2018).
936 48. Yang, W. *et al.* MiR-182-5p promotes the Metastasis and Epithelial-mesenchymal
937 Transition in Non-small Cell Lung Cancer by Targeting EPAS1. *J Cancer* **12**, 7120-7129
938 (2021).
939 49. Frey, P. *et al.* SMAD4 mutations do not preclude epithelial–mesenchymal transition in
940 colorectal cancer. *Oncogene* **41**, 824-837 (2022).
941 50. Li, J. *et al.* Knockdown of FOXO3a induces epithelial-mesenchymal transition and
942 promotes metastasis of pancreatic ductal adenocarcinoma by activation of the β-
943 catenin/TCF4 pathway through SPRY2. *Journal of Experimental & Clinical Cancer*
944 *Research* **38**, 38 (2019).
945 51. Koedoot, E. *et al.* Uncovering the signaling landscape controlling breast cancer cell
946 migration identifies novel metastasis driver genes. *Nat Commun* **10**, 2983 (2019).
947 52. Tsai, Y.C. *et al.* Disruption of ETV6 leads to TWIST1-dependent progression and
948 resistance to epidermal growth factor receptor tyrosine kinase inhibitors in prostate
949 cancer. *Mol Cancer* **17**, 42 (2018).

950 53. Meyer-Schaller, N. *et al.* A Hierarchical Regulatory Landscape during the Multiple
951 Stages of EMT. *Dev Cell* **48**, 539-553.e6 (2019).
952 54. Peñalosa-Ruiz, G. *et al.* WDR5, BRCA1, and BARD1 Co-regulate the DNA Damage
953 Response and Modulate the Mesenchymal-to-Epithelial Transition during Early
954 Reprogramming. *Stem Cell Reports* **12**, 743-756 (2019).
955 55. Mu, Z. *et al.* AZD8931, an equipotent, reversible inhibitor of signaling by epidermal
956 growth factor receptor (EGFR), HER2, and HER3: preclinical activity in HER2 non-
957 amplified inflammatory breast cancer models. *J Exp Clin Cancer Res* **33**, 47 (2014).
958 56. Roche, J. The Epithelial-to-Mesenchymal Transition in Cancer. *Cancers (Basel)*
959 **10**(2018).
960 57. Brown, M.S. *et al.* Dynamic plasticity within the EMT spectrum, rather than static
961 mesenchymal traits, drives tumor heterogeneity and metastatic progression of breast
962 cancers. *bioRxiv*, 2021.03.17.434993 (2022).
963 58. Thelen, M. *et al.* Cancer-specific immune evasion and substantial heterogeneity within
964 cancer types provide evidence for personalized immunotherapy. *NPJ Precis Oncol* **5**,
965 52 (2021).
966 59. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and
967 heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-9 (2013).
968 60. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of
969 TCGA data. *Nucleic Acids Res* **44**, e71 (2016).
970 61. Hoadley, K.A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of
971 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).
972 62. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression
973 data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).
974 63. Cook, D.P. & Vanderhyden, B.C. Context specificity of the EMT transcriptional
975 response. *Nat Commun* **11**, 2142 (2020).
976 64. Chae, Y.K. *et al.* Epithelial-mesenchymal transition (EMT) signature is inversely
977 associated with T-cell infiltration in non-small cell lung cancer (NSCLC). *Sci Rep* **8**, 2918
978 (2018).
979 65. Puram, S.V., Parikh, A.S. & Tirosh, I. Single cell RNA-seq highlights a role for a partial
980 EMT in head and neck cancer. *Mol Cell Oncol* **5**, e1448244 (2018).
981 66. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of
982 anticancer drug sensitivity. *Nature* **483**, 603-7 (2012).
983 67. Jiménez-Sánchez, A., Cast, O. & Miller, M.L. Comprehensive Benchmarking and
984 Integration of Tumor Microenvironment Cell Estimation Methods. *Cancer Res* **79**,
985 6238-6246 (2019).
986 68. Rolong, A., Chen, B. & Lau, K.S. Deciphering the cancer microenvironment from bulk
987 data with EcoTyper. *Cell* **184**, 5306-5308 (2021).
988 69. Andersson, A. *et al.* Single-cell and spatial transcriptomics enables probabilistic
989 inference of cell type topography. *Commun Biol* **3**, 565 (2020).
990 70. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell
991 transcriptomic data across different conditions, technologies, and species. *Nat*
992 *Biotechnol* **36**, 411-420 (2018).
993 71. Buffa, F.M., Harris, A.L., West, C.M. & Miller, C.J. Large meta-analysis of multiple
994 cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br J*
995 *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).

1026 **FIGURES**



1027

1028 **Figure 1. Pan-cancer EMT trajectories and underlying macro-states.** (a) Workflow for

1029 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.

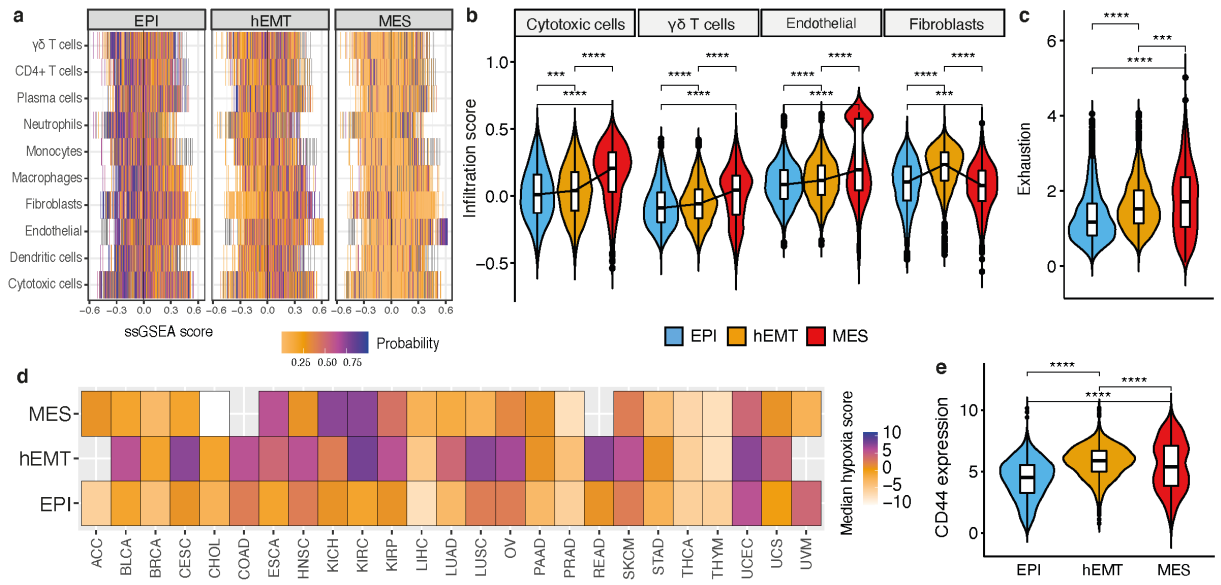


Figure 2. Tumour extrinsic and intrinsic hallmarks of EMT. (a) Heat map showcasing the results of a multinomial logistic regression model trained to predict EMT states based on cell infiltration in the microenvironment. Each row corresponds to a cell type and the corresponding per-sample infiltration 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 selected cell types. (c) Levels of exhaustion quantified across the three EMT states. (d) Median hypoxia values in the three different EMT states across tissues. (e) Gene expression levels of the stemness marker CD44 compared across the three EMT states.

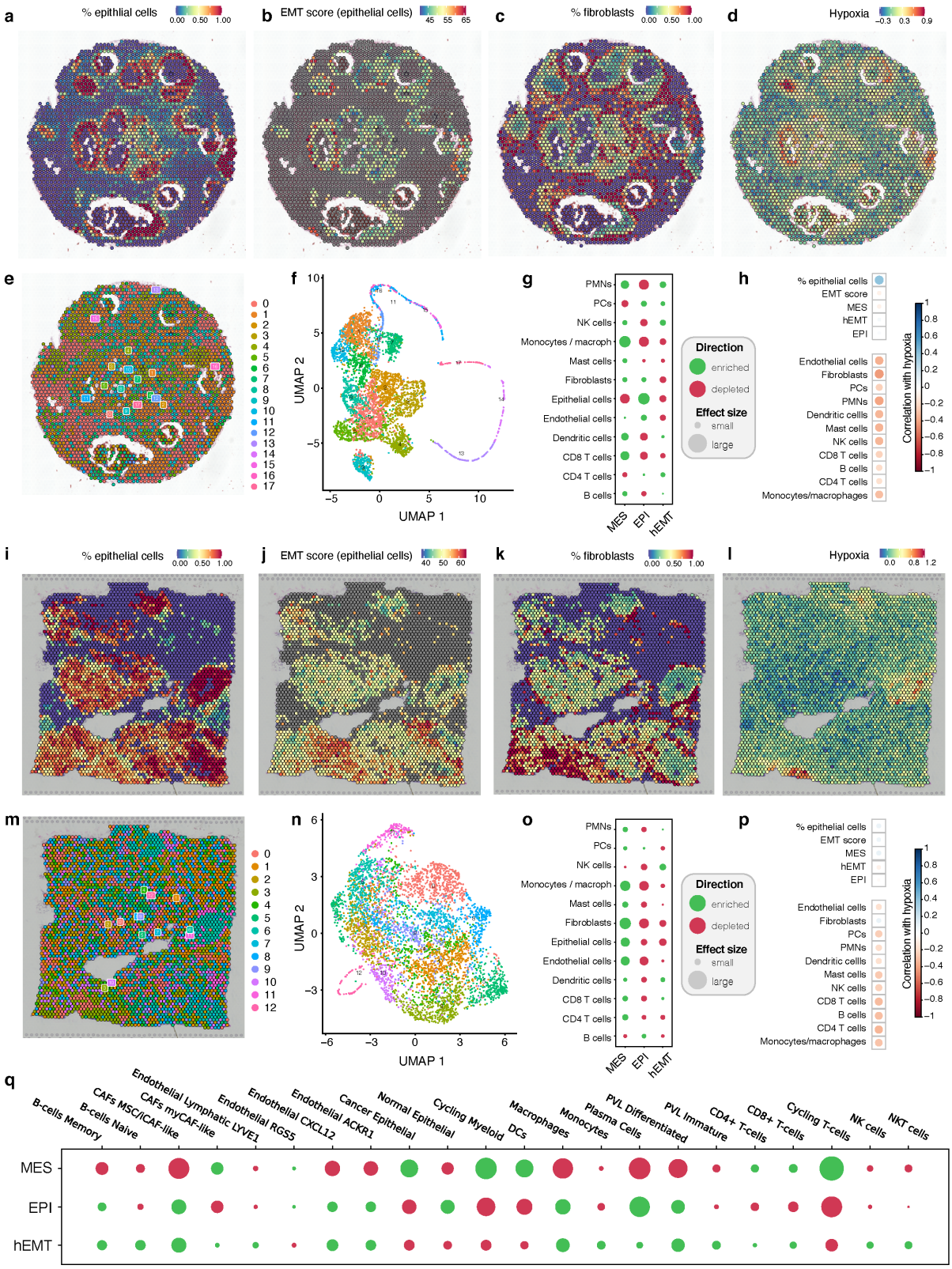


Figure 3. Spatial patterns of EMT. (a-d) Spot annotations of the fraction of epithelial cells (a), EMT scores across these epithelial spots (b), fraction of fibroblasts (c) and hypoxia (d) within individual spots profiled across the tissue in a selected breast cancer slide, derived from

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).

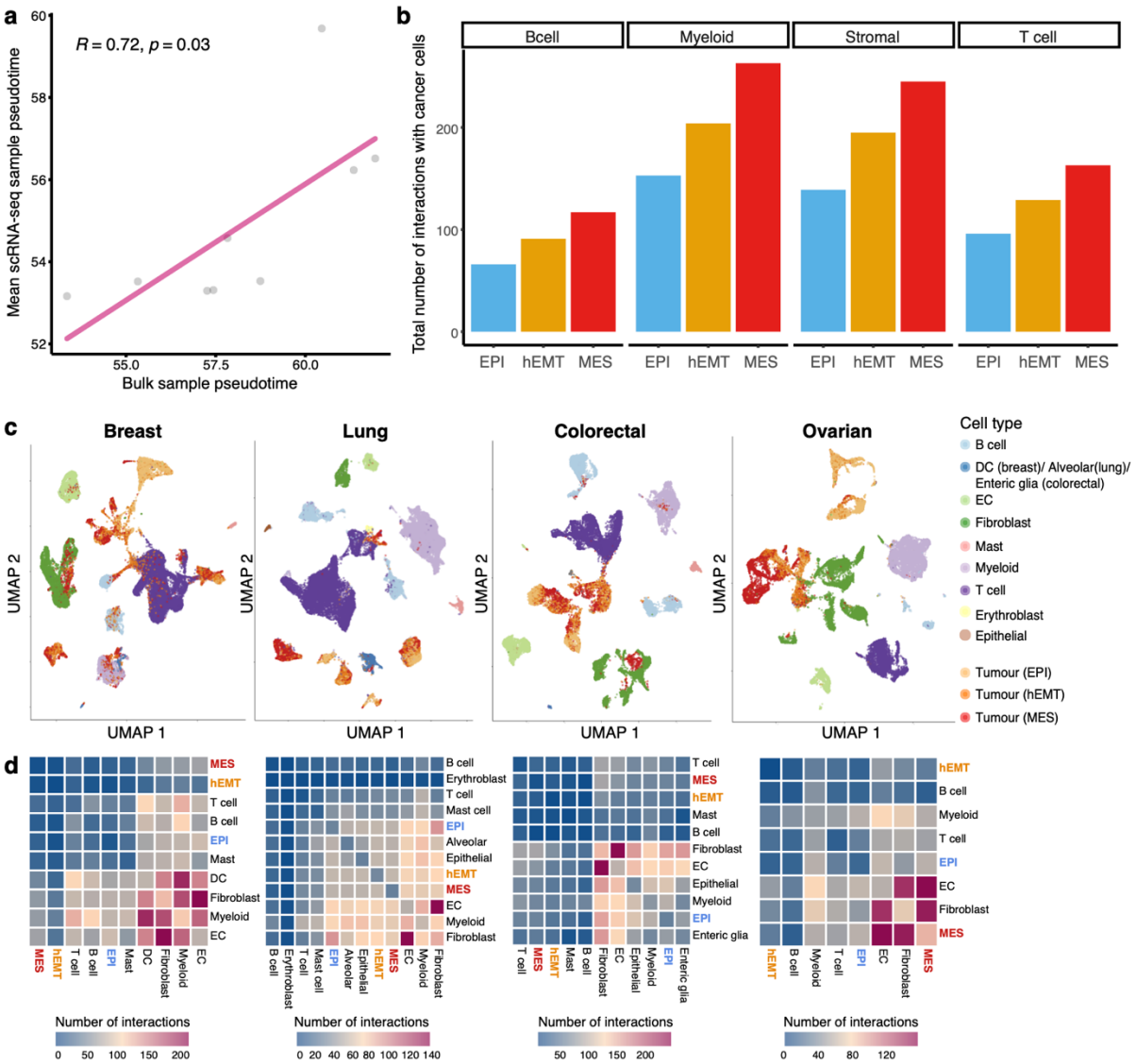
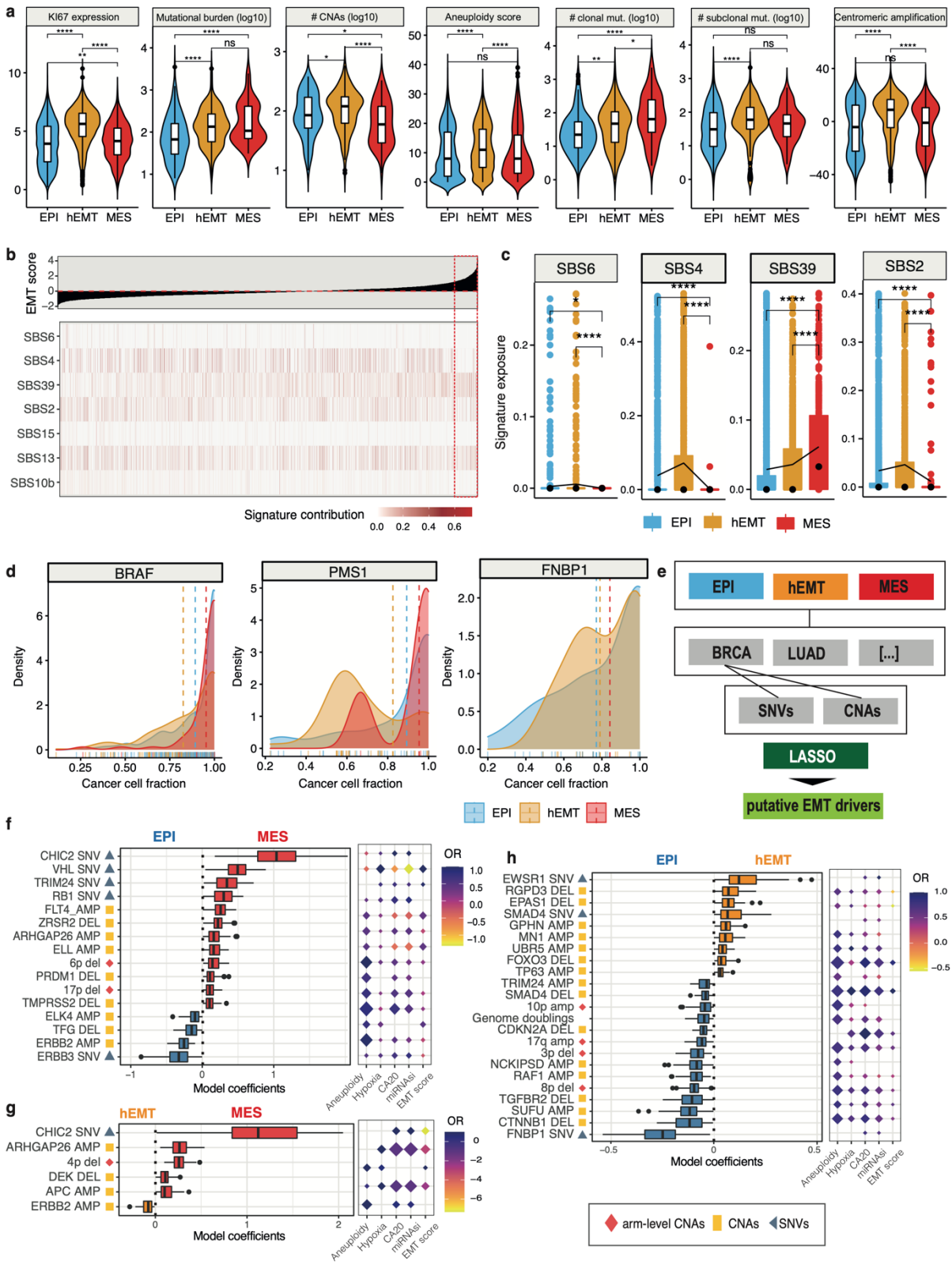


Figure 4. EMT diversity in single cell data. (a) Comparison between EMT pseudotime estimates in matched bulk and single cell samples from the same individuals. (b) Number of interactions established between tumour cells found in an EPI, hEMT or MES state and other cells in the tumour microenvironment 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

in the same breast, lung, colorectal and ovarian datasets. The EPI, hEMT and MES tumour cells are highlighted in blue, orange and red, respectively.



1091 **Figure 5. Genomic driver events linked with EMT.** (a) Expression of the proliferation
1092 marker Ki67, mutational and copy number aberration (CNA) burden, aneuploidy, number of
1093 clonal/subclonal mutations and centromeric amplification levels compared across the three
1094 EMT states. (b) Mutational signature exposures across TCGA samples sorted by EMT score.
1095 Only mutational signatures that were significantly linked with EMT from the linear mixed
1096 models are displayed. The corresponding EMT scores are displayed above. (c) Signature
1097 contributions from SBS6 (mismatch repair deficiency), SBS4 (smoking), SBS39 (unknown)
1098 and SBS2 (APOBEC) compared between the three EMT states. (d) Cancer cell fraction of
1099 genomic markers showing significantly distinct distribution between EMT states. (e) The
1100 analytical workflow used to detect genomic events linked with EMT. For each state and cancer
1101 type, we used dNdScv, SNV and copy number enrichment to prioritise mutated genes and
1102 copy number events, respectively. These genomic events were then employed as input for
1103 lasso modelling to classify EMT states. (f) Top-ranked genomic markers distinguishing the
1104 mesenchymal from the epithelial state. The balloon chart on the right illustrates the
1105 association between each marker and aneuploidy, hypoxia, centromeric amplification (CA20),
1106 stemness index (mRNAsi) and EMT score. The size of the diamonds is proportional to the
1107 significance of association, the colours report the odds ratios. (g) List of the top-ranked
1108 genomic markers distinguishing the hEMT from the MES state and their associated hallmarks.
1109 (h) List of the top-ranked genomic markers distinguishing the hEMT from the EPI state and
1110 their associated hallmarks.

1111

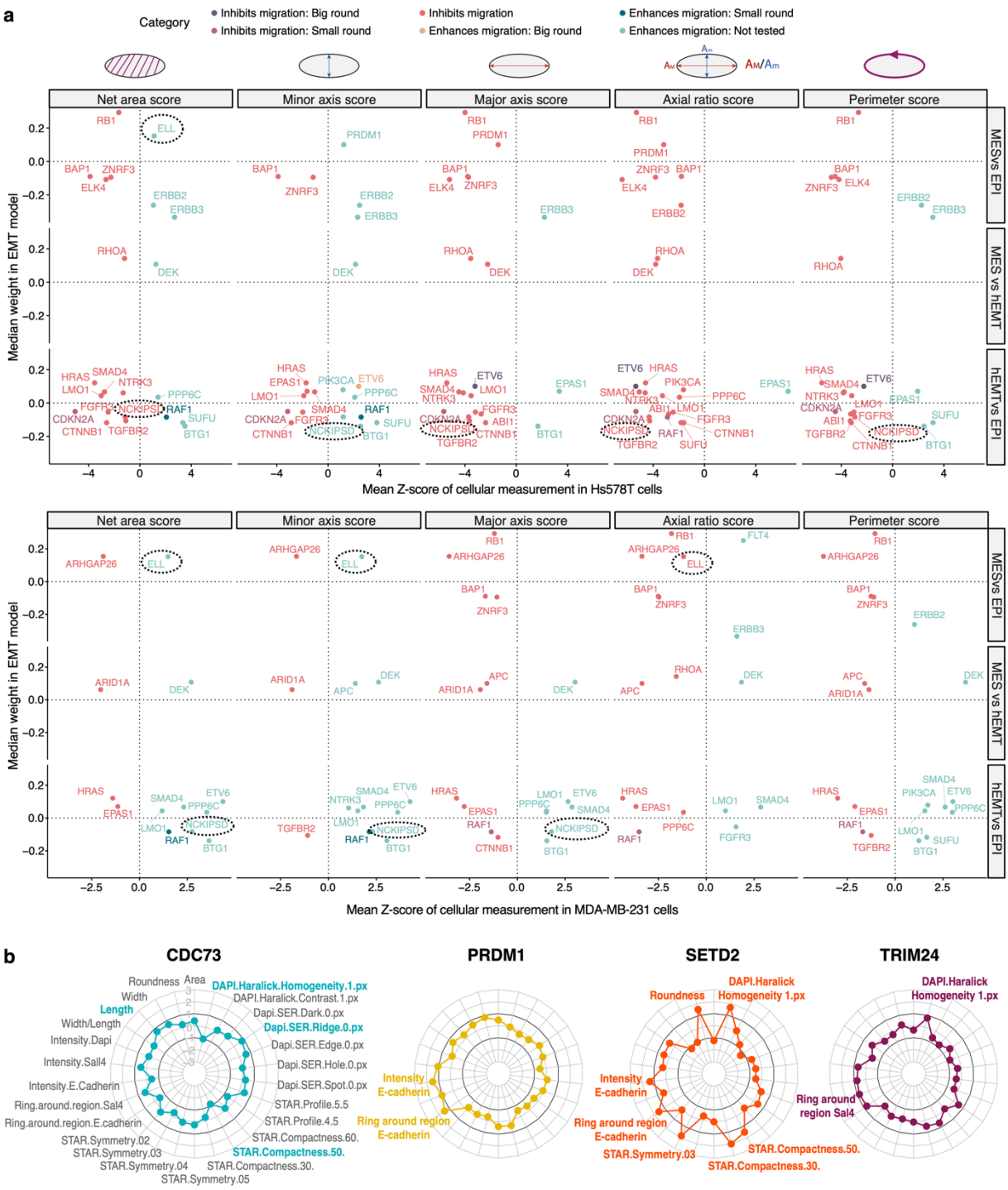


Figure 6. Validation of genomic associations with EMT using siRNA screens. (a) Gene knockdown effects on cell migration abilities in Hs578T (top panel) and MDA-MB-231 (bottom panel) cell lines (data from Koedoot et al[43]). The x axis depicts a change in the following measurements in the cells upon the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

1118 more migration). The y axis depicts the median weight of the gene in the model distinguishing
1119 two different EMT states. Larger absolute weights indicate more confident associations with
1120 EMT. The genes are coloured according to the suggested phenotype by the respective
1121 cellular measurement. A few of the genes highlighted have undergone further phenotypic tests
1122 and this is indicated by the confirmed phenotype (big/small round). The rest of the genes were
1123 not further tested in the study ("Not tested"). Only candidates with a Z-score value of cellular
1124 measurement >1 or <-1 are shown. The genes ELL and NCKIPSD are highlighted with black
1125 dotted rectangles as they are less well characterised in the context of EMT and are found as
1126 hits in the screen shown in panel b too. (b) Gene knockdown effects on various
1127 measurements of migration-related cell integrity in mouse embryonic fibroblasts (data from
1128 Penalosa-Ruiz et al[45]). The radial plots show the mean z-score depicting the change in cell
1129 measurement across multiple knockdown replicates. Z-scores greater than 1 or less than -1
1130 (above and below the corresponding black circles) suggest significant changes. All
1131 measurements are listed for the first gene only in grey text. Coloured text indicates significant
1132 changes in phenotype for each gene, e.g. knockdown of PRDM1 and SETD2 leads to an
1133 increase in E-cadherin expression intensity and area.

1134

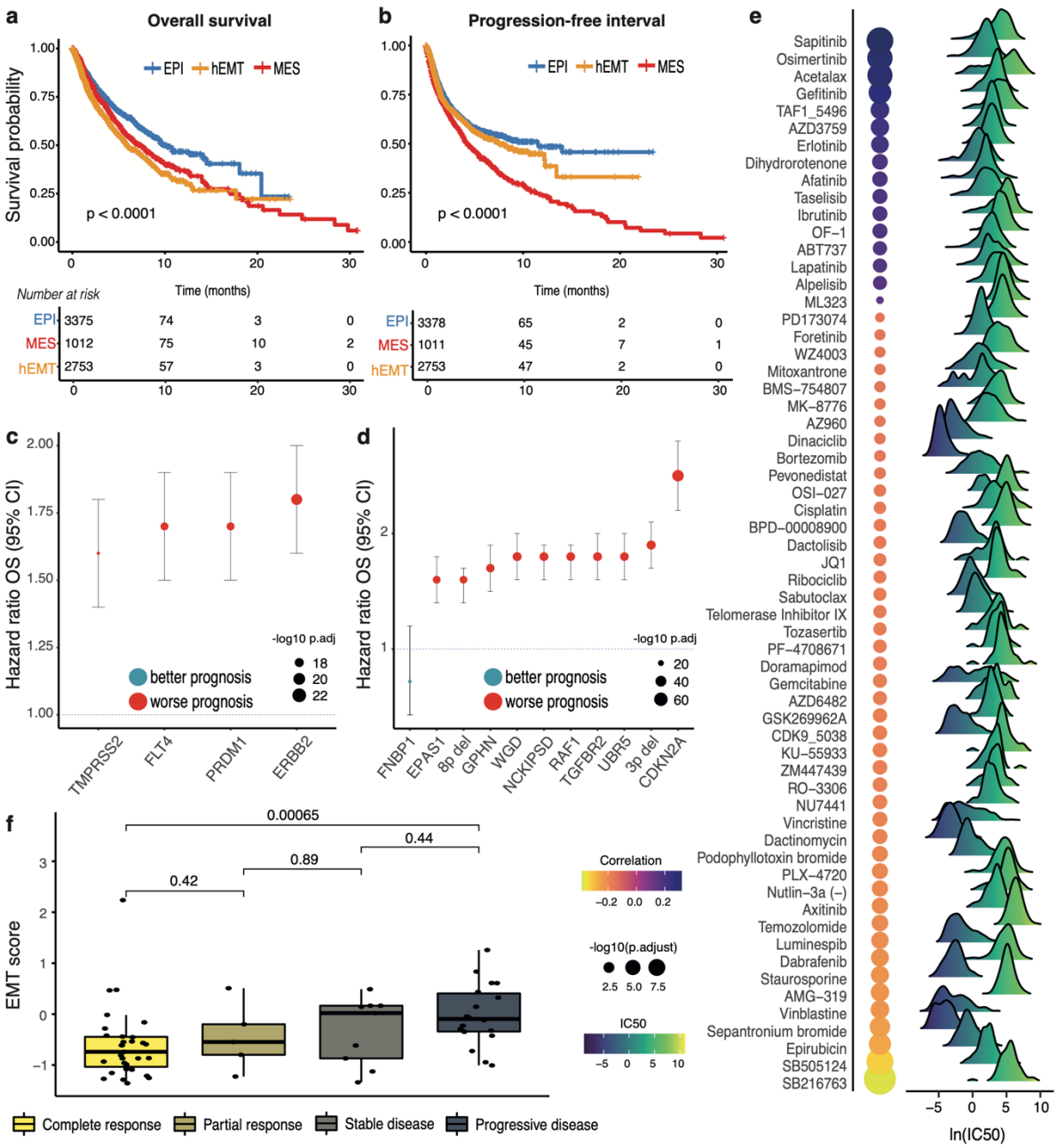


Figure 7. Clinical relevance of the EMT states. (a) Overall survival compared between MES, hEMT and EPI samples. (b) Progression free interval compared between the three groups. (c) Genomic markers distinguishing between mesenchymal and epithelial states with a significantly worse or improved outcome (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