# Genomic events shaping epithelial-to-mesenchymal trajectories in cancer

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

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

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

**ABSTRACT**

The epithelial to mesenchymal transition (EMT) is a key cellular process underlying cancer progression, with multiple intermediate states whose molecular hallmarks remain poorly characterized. To fill this gap, we explored EMT trajectories in 8,778 tumours of epithelial origin and identified three macro-states with prognostic and therapeutic value, attributable to epithelial, hybrid E/M (hEMT) and mesenchymal phenotypes. We show that the hEMT state is remarkably stable and linked with increased aneuploidy, APOBEC mutagenesis and hypoxia. Additionally, we provide an extensive catalogue of genomic events underlying distinct evolutionary constraints on EMT transformation, including novel pan-cancer dependencies of hEMT on driver genes *PRRX1, BCOR* and *CNOT3*, as well as links between full mesenchymal transformation and *REG3A* and *SHISA4* mutations in lung and breast cancers, respectively. This study sheds light on the aetiology of the lesser characterised hybrid E/M state in cancer progression and the broader genomic hallmarks shaping the mesenchymal transformation of primary tumours.

## INTRODUCTION

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

Elucidating the evolutionary trajectories that cells take to progress through these states is key to understanding metastatic spread and predicting cancer evolution.

The transcriptional changes accompanying EMT in cancer have been widely characterised and are governed by several transcription factors, including Snail, Slug, Twist and zinc fingers *ZEB1* and *ZEB2*[6,7,8]. EMT appears driven by waves of gene regulation underpinned by checkpoints, such as *KRAS* signalling driving the exit from an epithelial state, dependent upon *EGFR* and *MET* activation[9].

However, EMT progression is not only characterized by transcriptional alterations of regulatory circuits; the genetic background of the cell can also impact its capacity to undergo this transformation. Gain or loss of function mutations in a variety of genes, including *KRAS*[10], *BRCA1*[11], *STAG2*[12], *TP53*[13,14], as well as amplifications of chromosomes 5, 7 and 13 have been shown to promote EMT[15]. Several pan-cancer studies have also linked copy number alterations, miRNAs and immune checkpoints with EMT on a broader level[16,17]. Mathematical models have been developed to describe the switches between epithelial and mesenchymal states[4] but without considering any genomic dependencies.

Despite extensive efforts to study the dynamics of EMT, some aspects of this process remain poorly characterized. In particular, most of the studies mentioned considered EMT as a binary

52 switch and failed to capture evolutionary constraints that may change along the continuum of
53 EMT transformation. Single cell matched DNA- and RNA-seq datasets would ideally be
54 needed for this purpose, but they are scarce. We hypothesised that a pan-cancer survey of
55 EMT phenotypes across bulk sequenced samples should capture a broad spectrum of the
56 phenotypic variation one may expect to observe at single cell level, and this could be linked
57 with genomic changes accompanying EMT transformation. To explore this, we have
58 integrated data from the Cancer Genome Atlas (TCGA), MET500[18], MetMap[19], GENIE[20],
59 MSK-IMPACT[21], GDSC[22] and POG570[23] datasets to characterise and validate EMT and
60 linked genetic changes in cancer. By mapping 8,778 tumours of epithelial origin onto a
61 "timeline" of epithelial-to-mesenchymal transformation, we identified discrete EMT macro-
62 states and derived a catalogue of genomic hallmarks underlying evolutionary constraints of
63 these states. These genomic events shed light into the aetiology of the lesser characterised
64 hybrid E/M phenotype and could potentially act as early biomarkers of invasive cancer.

65

66 **RESULTS**

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

68 To explore the EMT process within bulk tumour samples, we employed a cohort of primary
69 tumours of epithelial origin (n = 8,778) spanning 25 cancer types from TCGA. The bulk RNA-
70 seq data from these tumours underlie multiple transcriptional programmes reflecting different
71 biological processes, including EMT. Inspired by McFaline-Figueroa et al[9], we quantified the
72 levels of EMT in these bulk tumours against a reference single cell RNA sequencing (scRNA-
73 seq) dataset derived from MCF10 breast cancer cells that have been profiled at different times
74 during the epithelial to mesenchymal transition *in vitro*. These data allowed us to reconstruct a
75 generic "pseudo-timeline" of spontaneous EMT transformation onto which we projected the
76 bulk sequenced samples from TCGA, positioning them within the continuum of EMT states
77 (Figure 1a). Since the bulk signal is arising from a mixture of cells, some of which would be

78 found in different stages of mesenchymal transformation, the overall signal would reflect the
79 average EMT state across the entire tumour cell population.

80 Using this approach, we reconstructed the EMT pseudotime trajectory across multiple cancer
81 tissues (Figure 1b, Supplementary Table S1). The expression of canonical epithelial and
82 mesenchymal markers was consistent with that observed in the scRNA-seq data and
83 expectations from the literature (Supplementary Figure S1a). Along the pseudotime, we
84 observed frequent co-expression of such markers, which could reflect a hybrid E/M state[24].

85 To characterise the dominant EMT states governing the continuum of transcriptional activity
86 described above, we discretised the pseudotime trajectory based on expression values of
87 canonical EMT markers using a Hidden Markov Model approach and uncovered three macro-
88 states: epithelial (EPI), hybrid EMT (hEMT) and mesenchymal (MES) (Figure 1b-c,
89 Supplementary Figure 1b). These states were robust to varying levels of gene expression
90 noise (Supplementary Figure S1c-d). As expected, the probability for the cancer cells to
91 switch from the epithelial to the hEMT (0.35) state was higher than the probability to passage
92 directly into the mesenchymal state (0.14). The hEMT tumours tended to remain in the same
93 state 44% of the times, suggesting this state could be more stable than anticipated – as
94 previously stipulated[25] and consistent with observations that a fully mesenchymal state is not
95 always reached[26].

96 The EMT scores progressively increased between the EPI, hEMT and MES states, as
97 expected (Figure 1d). Reassuringly, in an independent cohort of metastatic samples
98 (MET500), EMT levels were relatively elevated along the transformation timeline compared to
99 TCGA samples and were most abundantly falling within the hEMT state (Figure 1d,
100 Supplementary Figure S1e). Interestingly, we also observed possible cases of a reversion to
101 an epithelial state in metastatic samples, which is to be expected when colonizing a new
102 environmental niche.

103 We also applied our EMT scoring methodology to the MetMap resource, which has
104 catalogued the metastatic potential of 500 cancer cell lines across 21 cancer types. The

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

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

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

124 **Tumour cell extrinsic hallmarks of EMT**

125 Multiple microenvironmental factors, including tumour associated macrophages, secreted
126 molecules (IL-1, TNF-α) or hypoxia, have been extensively described to promote EMT[29,30,31,32].
127 However, their macro-state specificity is less well characterised. We confirmed that
128 transformed samples, particularly hEMT ones, exhibited higher infiltration by non-cancer cells
129 (Supplementary Figure S2a-c). Endothelial, cytotoxic and γδ T cells were progressively
130 enriched with increased stages of EMT transformation (Figure 2a-b, Supplementary Figure
131 S2d-e), suggesting that the fully mesenchymal state is most often linked with “immune hot”

132 tumours. In line with this hypothesis, these tumours also showed the highest exhaustion levels
133 (Figure 2c). In contrast, the hEMT samples displayed the highest enrichment of cancer-
134 associated fibroblasts (Figures 2a-b, d-e, Supplementary Figure S2f and Supplementary
135 Table S3), confirming previous reports[33]. To avoid confounding effects between fibroblast and
136 hEMT markers as highlighted by Tyler and Tirosh[34], we excluded EMT markers from the
137 employed fibroblast gene sets, but we acknowledge that part of the signal recovered may still
138 not be unambiguously attributed to either the cancer or microenvironmental component.

139 Additionally, samples with a transformed phenotype (MES, hEMT) presented significantly
140 elevated hypoxia levels (Figure 2f, Supplementary Figure S2g). Hypoxia has previously been
141 shown to promote EMT by modulating stemness properties[35]. We found that CD44, an
142 established cancer stem cell marker known to promote EMT[36,37], was most highly expressed
143 in the hEMT state across cancers (Figure 2g), and elevated levels of several other stemness
144 signatures most often accompanied the hEMT and MES macro-states (Figure S2h-i). Unlike
145 mesenchymal samples, the majority of hEMT tumours (35%) were characterized by both
146 hypoxia and CD44 expression (Supplementary Table S4). Thus, the interplay between
147 hypoxia and stemness may play a greater role in attaining the hEMT state compared to the
148 fixation of a fully mesenchymal phenotype.
149

150 **Tumour cell intrinsic hallmarks of EMT**

151 In addition to environmental factors, intrinsic cell properties such as increased proliferation,
152 mutational burden and aneuploidy would be expected along the EMT transformation
153 trajectory. Across distinct tissues, these changes were most pronounced in the hEMT state
154 (Figure 2h). Interestingly, this group also presented higher levels of centrosome amplification,
155 which have been linked with increased genomic instability[38,39] and poor prognosis[40].

156 Such alterations to the genomic integrity of the cells result from multiple mutational processes.
157 These processes leave recognizable patterns in the genome termed “mutational signatures”,
158 which in their simplest form constitute of trinucleotide substitutions and have been broadly

159 characterised across cancers[41]. However, their involvement in EMT transformation is poorly
160 understood. To investigate whether any neoplastic process introducing mutations in the
161 genomes was conditioned by EMT, we modelled the associations between mutational
162 signatures and EMT using linear mixed effects models while accounting for tissue effects
163 (Figure 2i-j, Supplementary Figure S3a, Supplementary Table S5). The ageing signature
164 SBS1 was significantly increased in non-mesenchymal tumours, while the APOBEC
165 mutagenesis signature SBS13 appeared specifically elevated in hEMT. It is worth noting that if
166 we did not account for the tissue effect, the signatures most strongly associated with EMT
167 progression and specifically increased in mesenchymal samples were SBS7a/b, linked with
168 UV light-induced damage (Supplementary Figure S3b-c). Samples exposed to these
169 carcinogens may progress to a mesenchymal phenotype earlier within the primary tumour, or
170 such mutations may be more easily fixed in transformed cells.
171

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

173 Beyond the broader hallmarks discussed above, we sought to identify specific genomic
174 changes creating a favourable environment for EMT transformation. We prioritised cancer
175 driver mutations, focal and arm-level copy number changes that may be linked with EMT, and
176 implemented a lasso-based machine learning framework to identify those drivers able to
177 discriminate between EPI, hEMT and MES states across cancers, while accounting for tissue-
178 specific effects (Methods, Figure 3a, Supplementary Table S6a). This model was validated
179 using several other machine learning approaches and demonstrated remarkably high
180 accuracies of 80-99% (Supplementary Figure S4a-n, Supplementary Table S6b).

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

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

190 The hEMT state-specific markers were mostly enriched in cell cycle, fate commitment and cell
191 adhesion pathways, hallmarks that are classically associated with cancer progression (Figure
192 3e, Supplementary Figure S4p). These alterations most often presented in a reduced cancer
193 cell fraction, concordant with subclonal diversification in line with phenotypic transformation
194 (Figure 3c). We identified several events that have not been previously linked with this
195 intermediate phenotype in the literature, including alterations of *PRRX1*, *BCOR*, *FAM135B,*
196 *CNOT3* and *ERCC3* (Figure 3e, Supplementary Figure S5a-b, Supplementary Table S6a).
197 *TCEA1* and *SPEN* alterations were specifically linked with high levels of aneuploidy,
198 stemness, centrosome amplification and hypoxia. While none of these events have been
199 specifically tied to an hEMT state before, *CNOT3* does not appear to have even been linked
200 more generally with invasion or metastasis. *CNOT3* is a translational repressor[42] upregulated
201 in non-small cell lung cancer[43].

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

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

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

217 A good fraction of the reported alterations was independently validated in the MET500, GENIE
218 and MSK-IMPACT studies: 63% of point mutations and 43% of copy number events
219 distinguishing MES from EPI, as well as 58% of somatic mutations and 70% of the copy
220 number events discriminating hEMT from EPI (Figure 4a-d). Acknowledging the limitations of
221 these resources to metastatic samples (MET500) and targeted sequencing (GENIE, MSK-
222 IMPACT), our analysis was nevertheless able to recapitulate genomic events necessary for
223 EMT transformation that are preserved during metastatic progression, but also uncovered
224 some events that may not be further maintained in the new metastatic niche. Furthermore,
225 31% and 57% of these alterations were confirmed to be linked with the metastatic potential of
226 cancer cell lines, at pan-cancer and tissue specific level, respectively (Figure 4e-f). 48% of the
227 copy number events were validated in at least one cell-line, e.g. *COX6AC* amplifications in the
228 pancreas (Figure 4h). Interestingly, the impact on metastatic potential varied between
229 increases and decreases, which can partly be explained by the heterogeneity and complex
230 dependencies of cell lines, and partly by our incomplete understanding of the metastatic
231 potential of hEMT cells. Suppression of most of these genes strongly impacted cell viability
232 (Figure 4i-j) and were targets of transcription factors regulating EMT (Figure 4k-l).
233 Knockdowns of the hEMT-linked genes *PRRX1* and *CREBBP* were also linked with a weak
234 EMT phenotype in murine epithelial cells[47].

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

240 (Supplementary Figures S6b,d-e). In contrast, proteins differentially expressed in a fully
241 mesenchymal state (Supplementary Figure S6c) were linked with mTOR signaling, response
242 to oxygen levels, and UV damage response (Supplementary Figure S6f-g), consistent with our
243 previous findings linking the MES state with hypoxia and the UV mutational signature SBS7.
244

245 **Tissue-specific EMT trajectories and genomic dependencies**

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

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

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

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

274 Most of the focal copy number alteration events in both lung and breast cancers were distinct
275 from the pan-cancer analysis (Supplementary Figure S7b-f, Supplementary Table S7e,j), and
276 several of them have been previously associated with EMT, albeit not in a micro-state specific
277 manner. Amplifications of the non-essential *PRDM2* gene in breast cancer and deletions of
278 the non-essential gene *CUX1* in lung cancer, among others, were linked with metastatic
279 potential in cell lines and transcription factor regulation of EMT (Figure 5i, Supplementary
280 Figure S7h,j).
281

282 **Clinical relevance of EMT**

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

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

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

294 Interestingly, the progression-free interval after oxaliplatin treatment was shorter in patients
295 with hEMT tumours (Supplementary Figure S8c), suggesting hEMT might be linked with poor
296 responses to this chemotherapy drug. To further explore potential links between EMT and
297 therapy responses, we investigated whether EMT progression might confer different levels of
298 sensitivity to individual cancer drugs using cell line data from GDSC[22]. We found 22
299 compounds whose IC50 values were significantly correlated with the EMT score (Figure 6e).
300 The strongest associations were observed with Acetalax, a drug used in the treatment of triple
301 negative breast cancers, and Sapitinib, an inhibitor of ErbB1/2/3[52]. Several hEMT and MES
302 biomarkers (e.g. *EP300*, *FAT1*, *NFE2L2*, *PTEN*) conferred increased sensitivity to multiple
303 drugs in a tissue specific manner (Supplementary Figure S8d-f), and suggested opportunities
304 for repurposing of non-oncology drugs (Supplementary Figure S8g-h).

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

314

315 **DISCUSSION**

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

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

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

342 While the exploration of EMT biomarkers is not new, most of the studies in this area have
343 been reliant on gene expression activity rather than mutational dependencies and they are
344 generally tissue-specific[17,30]. Pan-cancer studies generally consider EMT as a binary

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

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

360

361 **METHODS**

362 **Data sources**

363 Bulk RNA-sequencing, copy number (segment file and focal alterations), somatic
364 variants(MuTect[59]) and clinical data were retrieved for 8,7778 primary tumours of epithelial
365 origin from the harmonized version of TCGA using the *TCGAbiolinks* R package[60]. All other
366 data sources employed for validation are described below.

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

368 The reconstruction of the EMT trajectory of the TCGA samples was performed using a
369 procedure that allows to map bulk samples to single cell-derived expression programmes
370 inspired from McFaline-Figueroa et al[9]. The workflow of the analysis consists of several steps.

371 The analysis requires two gene expression matrices as input, corresponding to one bulk
372 sequenced dataset and a single cell dataset. In our case, we used as input the bulk RNA-seq
373 data from TCGA samples and scRNA-seq of a spontaneous EMT model derived in MCF10
374 cell lines[9] with the associated trajectory (P). In the first step of the analysis the matrices were
375 merged; then, in order to remove the batch effects originated by the two different platforms, a
376 correction was applied using ComBat[61]. In the second step, principal component analysis
377 (PCA) was performed on the merged matrix. The MCF10 derived EMT trajectory was then
378 mapped onto the TCGA data using an iterative process and a mapping strategy based on k
379 nearest neighbours (kNN). The number of iterations (i) is equal to the number of TCGA
380 samples. During each i-th step of iteration, a single TCGA sample and the MCF10 scRNA-seq
381 data were used as input for the kNN algorithm. The procedure computed the mean of the
382 pseudotime values of the single cell samples that have been detected by the kNN algorithm to
383 be associated with the i-th TCGA sample. The implementation of the kNN algorithm is based
384 on get.knnx() function from the *FNN* R package.

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

386 We used a Hidden Markov Model approach to identify of a discrete number of EMT states.
387 The input of this analysis was a matrix (M) where the rows were the TCGA samples (N) and
388 the columns the gene markers (G) of EMT (see the section "Computation of the EMT scores"
389 below for the list of genes). The original N columns were sorted for the t values of the
390 pseudotime (P). This matrix and P were provided as input for a lasso penalized regression. P
391 was used as response variable, the genes as the independent variables. The non-zero
392 coefficients obtained from this analysis were selected to create a sub-matrix of M that was
393 used as input for a Hidden Markov Model.

394 Different HMM models were tested while changing the number of states. After this tuning, and
395 through manual inspection, we determined that 3 states were most in line with biological
396 expectations. Each HMM state was assigned to a "biological group" (i.e. epithelial, hybrid
397 EMT, mesenchymal) by exploring the expression levels of known epithelial and mesenchymal

398 markers in each HMM state. The selection of the coefficients was performed with the R
399 package *glmnet*. The identification of the EMT states was done using the *deepmixS4* R
400 package.

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

412 **Computation of the EMT scores**

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

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

418 EMT scores for each TCGA sample were computed in a similar manner as described by Chae
419 et al[62]. Briefly, the average z-score transformed expression levels of the mesenchymal
420 markers were subtracted from the average z-score transformed expression levels of the
421 epithelial markers. To segment the EMT trajectory, along with the epithelial and mesenchymal
422 markers we have also considered markers of hybrid EMT[63,6]: *PDPN*, *ITGA5*, *ITGA6*, *TGFBI*,
423 *LAMC2*, *MMP10*, *LAMA3*, *CDH13*, *SERPINE1*, *P4HA2*, *TNC*, *MMP1*.

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

425 The RSEM gene-expression values of the Cancer Cell Line Encyclopedia[64] project were
426 retrieved from the CCLE Data Portal. We used the same procedure described above to map
427 the CCLE data onto the spontaneous MCF10 EMT trajectory. This allowed for the pseudotime
428 to be quantified for each CCLE sample. A segmentation using a HMM model was performed
429 to identify a discrete number of EMT states (n=3). The EMT scores were also computed for
430 each cell line. These results were referenced against the metastatic potential scores from
431 MetMap500[19]. The association between HMM states and experimentally measured metastatic
432 potential groups in cell lines (non-metastatic, weakly metastatic and metastatic) was assessed
433 using the *vcd* R package.

434 **Tumour microenvironment quantification**

435 The tumour purity values of TCGA samples were retrieved from Hoadley et al[65]. Immune
436 deconvolution was performed using the ConsensusTME R package[66] and the ssGSEA
437 method for cell enrichment analysis.

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

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

444 **Genomic hallmark quantification**

445 To characterize the aneuploidy and the centromeric amplification levels of the samples in
446 each EMT state we used the pre-computed values for TCGA from previous works [68],[40]. The
447 hypoxia levels were quantified as described by Bandhari et al[69]. Several hypoxia gene
448 signatures were considered, yielding similar results. Only the results obtained using the genes
449 from Buffa et al[70] were reported. Finally, to estimate the levels of stemness in each EMT state,

450 we considered a catalogue of stemness gene sets[71] and used them as input for gene set
451 enrichment analysis via the *GSVA* R package.

452 **Mutational signature analysis**

453 The identification of the mutational spectrum of the samples in each EMT state was performed
454 using a custom approach based on SigProfilerExtractor[41] and deconstructSigs[72].
455 SigProfilerExtractor was used for a de-novo identification of the mutational signatures. We
456 selected the solutions in which the minimal stability was greater than 0.4 and the sum of the
457 minimal stabilities across signatures was greater than 1. The cosine similarity with mutational
458 signatures catalogued in the COSMIC database was computed, and only the solutions with
459 non-redundant signatures were selected. Next, we independently ran deconstructSigs. To
460 ensure consistency with Alexandrov et al[41], we evaluated the presence of the ageing-linked
461 SBS1 and SBS5, which have been identified in all cancers. We employed the following steps
462 to obtain a final list of signatures and their exposures for each tissue individually:

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

468 To identify EMT-associated mutational processes we used a similar approach to the one
469 described in Bhandari et al.[69], based on linear mixed-effect models. Cancer type was
470 incorporated as a random effect in each model. An FDR adjustment was applied to the p-
471 values obtained from the analysis. The full model for a specific signature (SBS) is as follows:

472 $$EMT_score \sim SBS + (1|cancer)$$

473 **Identification of cancer driver genes**

474 Single nucleotide variants were obtained from TCGA using the *TCGAbiolinks* R package and
475 the Mutect pipeline. To identify putative driver events that are positively selected in

476 association with an EMT state, we employed dNdScv[73], which quantified the ratio of non-
477 synonymous and synonymous mutations (dN/dS) in each gene and state, by tissue. We only
478 considered protein coding genes in the analysis, and filtered out consecutive mutations in the
479 genome, as recommended by the authors. All the somatic driver events with a q-value less
480 than of 0.10 were considered for downstream analysis.

481 **Definition of somatic copy number events**

482 Focal and arm-level copy number events across TCGA samples were identified using
483 GISTIC2[74], with the following parameters: armpeel = 1, brlen 0.5, cap 1.5, conf 0.99 -ta -td =
484 0.3, -genegistic = 1, gcm = extreme, -maxseg = -2500, qvt = 0.1, rx = 0, save gene = 1, broad
485 = 1.

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

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

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

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

511 **Cancer cell fraction estimates**

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

514 $$CCF_i = \left(2 + \frac{[purity * (CN_i - 2)]}{purity}\right) \cdot VAF_i ,$$

515 where $CN_i$ stands for the absolute copy number of the segment spanning mutation *i* and $VAF_i$
516 is the variant allele frequency of the respective mutation. The purities of the TCGA samples
517 were obtained from Hoadley et al[65].

518 **Validation of the putative EMT biomarkers**

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

522 - **MET500**

523 Copy number regions encompassing 50% or more of the chromosomal arm were considered
524 as broad copy number events. Copy number regions below this threshold were classified as

525 focal events (based on hg19 cytoband positions and genomic coordinates obtained from
526 UCSC).

527 - **MSK-IMPACT**

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

538 - **GENIE**

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

545 - **MetMap**

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

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

553 **Gene essentiality evaluation using Project Achilles**

554 The viability of the cancer cell lines harbouring putative EMT biomarkers was evaluated based
555 on CRISPR screening data[75] conducted on 990 cell lines. CERES scores denoting gene
556 essentiality were downloaded from Project Achilles. Negative values of these scores indicate
557 that the depletion of a gene influences negatively the viability of a cell line. We only
558 considered genomic markers linked with the hEMT and MES states from our analysis and
559 assessed CERES scores for individual genes both pan-cancer and at tissue level.

560 **EMT-linked transcription factor network analysis**

561 Data from knockdown experiments of transcription factors regulating EMT was downloaded
562 from Meyer-Schaller et al[47]. Genes differentially expressed with an absolute log2 fold change
563 >1 and a q-value < 0.05 upon transcription factor knockdown were selected and intersected
564 with our list of putative EMT biomarkers.

565 **Proteomic biomarkers linked to EMT**

566 Level 4 RPPA proteomics data were retrieved from the Cancer Proteome Atlas (TCPAv3.0[76],
567 https://tcpaportal.org/tcpa/download.html). We identified proteins whose levels change
568 significantly between EMT states using logistic regression. Only the proteins with a q-value <
569 0.10 and odds ratio <-0.5 or >0.5 have been considered for downstream analysis.

570 **Tissue-specific EMT trajectory derivation**

571 To analyze the EMT in the context of several external stimuli we used the data generated by
572 Cook et al[49]. Specifically, we considered scRNA-seq data of MCF7 (breast cancer cell line)
573 and A549 (alveolar basal epithelial cells) under the stimulus of TNF, EGF and TGF-β1. For
574 each cell line, we integrated the expression data derived from all the conditions (TNF, EGF,
575 TGF-β1) using the FindIntegrationAnchors() and IntegrateData() functions from *Seurat*[77].
576 Next, we performed dimensionality reduction on the resulting data using UMAP (from the *uwot*

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

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

601 To identify the genes positively selected in each EMT tissue-specific cluster, we ran dNdScv[73]
602 in each cluster separately. To detect recurrent copy number alterations in lung
603 adenocarcinoma and breast cancer, we considered all genes with alterations in at least 10

604 patients and applied a Fisher's exact test to identify those that appeared enriched in a
605 particular EMT state. Only the genes with a q-value < 0.05 have been reported.

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

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

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

621 **Gene ontology analysis**

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

624 **Survival analysis**

625 Standardized clinical information for the TCGA cohort was obtained from Liu et al[81]. The
626 following end points were considered: overall survival, disease-specific survival and
627 progression-free interval. Patients were considered "lymph node positive" if they presented
628 infiltration in one or more lymph nodes. Cox proportional hazard models were used to model
629 survival based on variables of interest and to adjust for the following potential confounders:

630 tumour stage, age at diagnosis, gender and body mass index (BMI). Patients in clinical stages
631 I-II were denoted as having “early stage tumours”, while stages III-IV corresponded to “late
632 stage tumours”. The R packages *survival*, *survminer* and *ggforest* were used for data analysis
633 and visualization.

634 **Data visualization and basic statistics**

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

637 **Code**

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

640

641 **ACKNOWLEDGEMENTS**

642 GMT was supported by a Wellcome Trust Seed Award in Science (215296/Z/19/Z). MS was
643 supported by a UKRI Future Leaders Fellowship (MR/T042184/1) and an Academy of Medical
644 Science Springboard award (SBF004\1042).

645 The results published here are in part based upon data generated by the TCGA Research
646 Network: https://www.cancer.gov/tcga. The authors would like to acknowledge the American
647 Association for Cancer Research and its financial and material support in the development of
648 the AACR Project GENIE registry, as well as members of the consortium for their commitment
649 to data sharing. Interpretations are the responsibility of study authors.

650 **AUTHOR CONTRIBUTIONS**

651 MS designed the study and supervised the analyses. GMT conducted all the analyses.

652 **COMPETING INTEREST STATEMENT**

653 None declared.

654

655 **REFERENCES**


656 1. Thiery, J. P., Acloque, H., Huang, R. Y. J. & Nieto, M. A. Epithelial-mesenchymal transitions
657 in development and disease. *Cell* **139**, 871–890 (2009).
658 2. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J Clin Invest*
659 **119**, 1420–1428 (2009).
660 3. Pastushenko, I. & Blanpain, C. EMT Transition States during Tumor Progression and
661 Metastasis. *Trends in Cell Biology* **29**, 212–226 (2019).
662 4. Goetz, H., Melendez-Alvarez, J. R., Chen, L. & Tian, X.-J. A plausible accelerating function of
663 intermediate states in cancer metastasis. *PLOS Computational Biology* **16**, e1007682
664 (2020).
665 5. Pastushenko, I. *et al.* Identification of the tumour transition states occurring during EMT.
666 *Nature* **556**, 463–468 (2018).
667 6. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor
668 Ecosystems in Head and Neck Cancer. *Cell* **171**, 1611-1624.e24 (2017).
669 7. Karacosta, L. G. *et al.* Mapping lung cancer epithelial-mesenchymal transition states and
670 trajectories with single-cell resolution. *Nature Communications* **10**, 5587 (2019).
671 8. Stemmler, M. P., Eccles, R. L., Brabletz, S. & Brabletz, T. Non-redundant functions of EMT
672 transcription factors. *Nature Cell Biology* **21**, 102–112 (2019).
673 9. McFaline-Figueroa, J. L. *et al.* A pooled single-cell genetic screen identifies regulatory
674 checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nat Genet* **51**,
675 1389–1398 (2019).
676 10. Rhim, A. D. *et al.* EMT and Dissemination Precede Pancreatic Tumor Formation. *Cell*
677 **148**, 349–361 (2012).

678 11. Xu, J. *et al.* A novel Ubc9 -dependent pathway regulates SIRT1- ER-α Axis and BRCA1-
679 associated TNBC lung metastasis. *Integr Mol Med* **4**, (2017).
680 12. Nie, Z. *et al.* STAG2 loss-of-function mutation induces PD-L1 expression in U2OS cells.
681 *Ann Transl Med* **7**, (2019).
682 13. Chang, C.-J. *et al.* p53 regulates epithelial-mesenchymal transition and stem cell
683 properties through modulating miRNAs. *Nat Cell Biol* **13**, 317–323 (2011).
684 14. Stein, Y., Rotter, V. & Aloni-Grinstein, R. Gain-of-Function Mutant p53: All the Roads
685 Lead to Tumorigenesis. *Int J Mol Sci* **20**, (2019).
686 15. Vasudevan, A. *et al.* Single-Chromosomal Gains Can Function as Metastasis
687 Suppressors and Promoters in Colon Cancer. *Dev Cell* **52**, 413-428.e6 (2020).
688 16. Zhao, M., Liu, Y. & Qu, H. Expression of epithelial-mesenchymal transition-related
689 genes increases with copy number in multiple cancer types. *Oncotarget* **7**, 24688–24699
690 (2016).
691 17. Mak, M. P. *et al.* A patient-derived, pan-cancer EMT signature identifies global
692 molecular alterations and immune target enrichment following epithelial to mesenchymal
693 transition. *Clin Cancer Res* **22**, 609–620 (2016).
694 18. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**,
695 297–303 (2017).
696 19. Jin, X. *et al.* A metastasis map of human cancer cell lines. *Nature* **588**, 331–336 (2020).
697 20. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine
698 through an International Consortium. *Cancer Discov* **7**, 818–831 (2017).
699 21. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective
700 clinical sequencing of 10,000 patients. *Nat Med* **23**, 703–713 (2017).

701 22. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–
702 754 (2016).
703 23. Rheinbay, E. The genomic landscape of advanced cancer. *Nat Cancer* **1**, 372–373
704 (2020).
705 24. Aiello, N. M. *et al.* EMT subtype influences epithelial plasticity and mode of cell
706 migration. *Dev Cell* **45**, 681-695.e4 (2018).
707 25. Jolly, M. K. *et al.* Implications of the Hybrid Epithelial/Mesenchymal Phenotype in
708 Metastasis. *Front Oncol* **5**, 155 (2015).
709 26. Plygawko, A. T., Kan, S. & Campbell, K. Epithelial–mesenchymal plasticity: emerging
710 parallels between tissue morphogenesis and cancer metastasis. *Philosophical Transactions*
711 *of the Royal Society B: Biological Sciences* **375**, 20200087 (2020).
712 27. Sabe, H. Cancer early dissemination: cancerous epithelial-mesenchymal
713 transdifferentiation and transforming growth factor β signalling. *J Biochem* **149**, 633–639
714 (2011).
715 28. Jolly, M. K., Ware, K. E., Gilja, S., Somarelli, J. A. & Levine, H. EMT and MET: necessary
716 or permissive for metastasis? *Mol Oncol* **11**, 755–769 (2017).
717 29. Jing, Y., Han, Z., Zhang, S., Liu, Y. & Wei, L. Epithelial-Mesenchymal Transition in tumor
718 microenvironment. *Cell Biosci* **1**, 29 (2011).
719 30. Gibbons, D. L. & Creighton, C. J. Pan-cancer survey of epithelial-mesenchymal
720 transition markers across the Cancer Genome Atlas. *Dev Dyn* **247**, 555–564 (2018).
721 31. Choi, B.-J., Park, S.-A., Lee, S.-Y., Cha, Y. N. & Surh, Y.-J. Hypoxia induces epithelial-
722 mesenchymal transition in colorectal cancer cells through ubiquitin-specific protease 47-
723 mediated stabilization of Snail: A potential role of Sox9. *Sci Rep* **7**, 15918 (2017).

724 32. Zhang, T., Suo, C., Zheng, C. & Zhang, H. Hypoxia and Metabolism in Metastasis. *Adv*
725 *Exp Med Biol* **1136**, 87–95 (2019).
726 33. Aggarwal, V., Montoya, C. A., Donnenberg, V. S. & Sant, S. Interplay between tumor
727 microenvironment and partial EMT as the driver of tumor progression. *iScience* **24**, 102113
728 (2021).
729 34. Tyler, M. & Tirosh, I. Decoupling epithelial-mesenchymal transitions from stromal
730 profiles by integrative expression analysis. *Nature Communications* **12**, 2592 (2021).
731 35. Emami Nejad, A. *et al.* The role of hypoxia in the tumor microenvironment and
732 development of cancer stem cell: a novel approach to developing treatment. *Cancer Cell*
733 *Int* **21**, 62 (2021).
734 36. San Juan, B. P., Garcia-Leon, M. J., Rangel, L., Goetz, J. G. & Chaffer, C. L. The
735 Complexities of Metastasis. *Cancers (Basel)* **11**, (2019).
736 37. Mani, S. A. *et al.* The epithelial-mesenchymal transition generates cells with properties
737 of stem cells. *Cell* **133**, 704–715 (2008).
738 38. Ganem, N. J., Godinho, S. A. & Pellman, D. A mechanism linking extra centrosomes to
739 chromosomal instability. *Nature* **460**, 278–282 (2009).
740 39. Lingle, W. L. *et al.* Centrosome amplification drives chromosomal instability in breast
741 tumor development. *Proc Natl Acad Sci U S A* **99**, 1978–1983 (2002).
742 40. de Almeida, B. P., Vieira, A. F., Paredes, J., Bettencourt-Dias, M. & Barbosa-Morais, N.
743 L. Pan-cancer association of a centrosome amplification gene expression signature with
744 genomic alterations and clinical outcome. *PLoS Comput Biol* **15**, e1006832 (2019).
745 41. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer.
746 *Nature* **578**, 94–101 (2020).

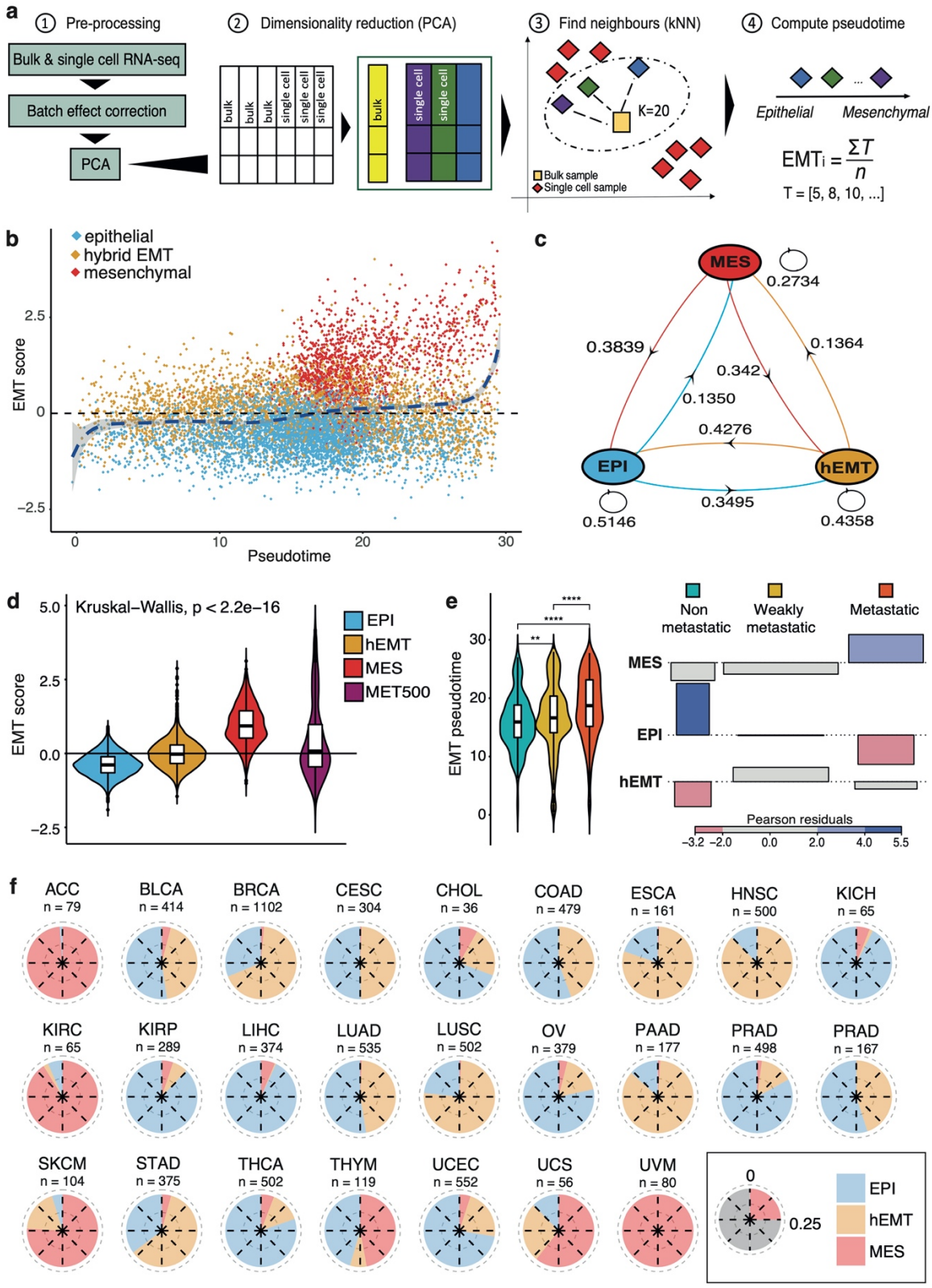
747 42. Jing, L. *et al.* CNOT3 contributes to cisplatin resistance in lung cancer through
748 inhibiting RIPK3 expression. *Apoptosis* **24**, 673–685 (2019).
749 43. Shirai, Y.-T. *et al.* CNOT3 targets negative cell cycle regulators in non-small cell lung
750 cancer development. *Oncogene* **38**, 2580–2594 (2019).
751 44. Zhang, S. *et al.* Loss of VHL expression contributes to epithelial-mesenchymal
752 transition in oral squamous cell carcinoma. *Oral Oncol* **50**, 809–817 (2014).
753 45. Duah, E. *et al.* Cysteinyl leukotriene 2 receptor promotes endothelial permeability,
754 tumor angiogenesis, and metastasis. *PNAS* **116**, 199–204 (2019).
755 46. Liu, W., Xin, H., Eckert, D. T., Brown, J. A. & Gnarra, J. R. Hypoxia and cell cycle
756 regulation of the von Hippel–Lindau tumor suppressor. *Oncogene* **30**, 21–31 (2011).
757 47. Meyer-Schaller, N. *et al.* A Hierarchical Regulatory Landscape during the Multiple
758 Stages of EMT. *Dev Cell* **48**, 539-553.e6 (2019).
759 48. Yamaguchi, H. & Taouk, G. M. A Potential Role of YAP/TAZ in the Interplay Between
760 Metastasis and Metabolic Alterations. *Front. Oncol.* **10**, (2020).
761 49. Cook, D. P. & Vanderhyden, B. C. Context specificity of the EMT transcriptional
762 response. *Nature Communications* **11**, 2142 (2020).
763 50. Zhang, M., Wang, J. & Guo, J. Role of Regenerating Islet-Derived Protein 3A in
764 Gastrointestinal Cancer. *Front. Oncol.* **0**, (2019).
765 51. Larue, L. & Bellacosa, A. Epithelial–mesenchymal transition in development and
766 cancer: role of phosphatidylinositol 3’ kinase/AKT pathways. *Oncogene* **24**, 7443–7454
767 (2005).
768 52. Mu, Z. *et al.* AZD8931, an equipotent, reversible inhibitor of signaling by epidermal
769 growth factor receptor (EGFR), HER2, and HER3: preclinical activity in HER2 non-amplified
770 inflammatory breast cancer models. *J Exp Clin Cancer Res* **33**, 47 (2014).

771 53. Roche, J. The Epithelial-to-Mesenchymal Transition in Cancer. *Cancers (Basel)* **10**,
772 (2018).
773 54. Tripathi, S., Chakraborty, P., Levine, H. & Jolly, M. K. A mechanism for epithelial-
774 mesenchymal heterogeneity in a population of cancer cells. *PLoS Comput Biol* **16**,
775 e1007619 (2020).
776 55. Simeonov, K. P. *et al.* Single-cell lineage tracing of metastatic cancer reveals selection
777 of hybrid EMT states. *Cancer Cell* (2021) doi:10.1016/j.ccell.2021.05.005.
778 56. Brown, M. S. *et al. Dynamic plasticity within the EMT spectrum, rather than static*
779 *mesenchymal traits, drives tumor heterogeneity and metastatic progression of breast*
780 *cancers*. http://biorxiv.org/lookup/doi/10.1101/2021.03.17.434993 (2021)
781 doi:10.1101/2021.03.17.434993.
782 57. Muñoz, D. P. *et al.* Activation-induced cytidine deaminase (AID) is necessary for the
783 epithelial–mesenchymal transition in mammary epithelial cells. *PNAS* **110**, E2977–E2986
784 (2013).
785 58. Bader, S. B. *et al.* Replication catastrophe induced by cyclic hypoxia leads to increased
786 APOBEC3B activity. *Nucleic Acids Research* **49**, 7492–7506 (2021).
787 59. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and
788 heterogeneous cancer samples. *Nature Biotechnology* **31**, 213–219 (2013).
789 60. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of
790 TCGA data. *Nucleic Acids Research* **44**, e71–e71 (2016).
791 61. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression
792 data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

793 62. Chae, Y. K. *et al.* Epithelial-mesenchymal transition (EMT) signature is inversely
794 associated with T-cell infiltration in non-small cell lung cancer (NSCLC). *Scientific Reports* **8**,
795 2918 (2018).
796 63. Puram, S. V., Parikh, A. S. & Tirosh, I. Single cell RNA-seq highlights a role for a partial
797 EMT in head and neck cancer. *Mol Cell Oncol* **5**, e1448244 (2018).
798 64. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of
799 anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
800 65. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of
801 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).
802 66. Jiménez-Sánchez, A., Cast, O. & Miller, M. L. Comprehensive Benchmarking and
803 Integration of Tumor Microenvironment Cell Estimation Methods. *Cancer Res* **79**, 6238–
804 6246 (2019).
805 67. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for
806 microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
807 68. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer
808 Aneuploidy. *Cancer Cell* **33**, 676-689.e3 (2018).
809 69. Bhandari, V. *et al.* Molecular landmarks of tumor hypoxia across cancer types. *Nature*
810 *Genetics* **51**, 308–318 (2019).
811 70. Buffa, F. M., Harris, A. L., West, C. M. & Miller, C. J. Large meta-analysis of multiple
812 cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br J Cancer*
813 **102**, 428–435 (2010).
814 71. Miranda, A. *et al.* Cancer stemness, intratumoral heterogeneity, and immune response
815 across cancers. *Proc Natl Acad Sci U S A* **116**, 9020–9029 (2019).

816 72. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs:
817 delineating mutational processes in single tumors distinguishes DNA repair deficiencies and
818 patterns of carcinoma evolution. *Genome Biology* **17**, 31 (2016).
819 73. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues.
820 *Cell* **171**, 1029-1041.e21 (2017).
821 74. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the
822 targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41
823 (2011).
824 75. Dempster, J. M. *et al.* Extracting Biological Insights from the Project Achilles Genome-
825 Scale CRISPR Screens in Cancer Cell Lines. *bioRxiv* 720243 (2019) doi:10.1101/720243.
826 76. Chen, M.-J. M. *et al.* TCPA v3.0: An Integrative Platform to Explore the Pan-Cancer
827 Analysis of Functional Proteomic Data. *Mol Cell Proteomics* **18**, S15–S25 (2019).
828 77. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21
829 (2019).
830 78. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell
831 transcriptomics. *BMC Genomics* **19**, 477 (2018).
832 79. Corsello, S. M. *et al.* Discovering the anti-cancer potential of non-oncology drugs by
833 systematic viability profiling. *Nat Cancer* **1**, 235–248 (2020).
834 80. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters |
835 OMICS: A Journal of Integrative Biology.
836 https://www.liebertpub.com/doi/10.1089/omi.2011.0118.
837 81. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-
838 Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).
839

840 **FIGURES**




841

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

843 reconstructing the EMT trajectories of TCGA samples. 1: Bulk and single cell datasets are

844 combined and processed together to remove batch effects. 2: Dimensionality reduction using
845 PCA is performed. 3: A k-nearest neighbours (kNN) algorithm is used to map bulk RNA-
846 sequencing onto a reference EMT trajectory derived from scRNA-seq data. 4: Tumours are
847 sorted on the basis of their mesenchymal potential along an EMT “pseudotime” axis. (b)
848 Scatter plot of EMT scores along the pseudotime. Each dot corresponds to one bulk tumour
849 sample from TCGA. Samples are coloured according to the designated state by the HMM
850 model. (c) Diagram of the transition probabilities for switching from one EMT state to another,
851 as estimated by the HMM model. MES: fully mesenchymal state, hEMT: hybrid E/M, EPI:
852 epithelial state. (d) EMT scores compared across epithelial, hEMT, mesenchymal TCGA
853 samples, and the MET500 cohort. (e) Left: EMT pseudotime values compared between cell
854 lines from CCLE classified as “non metastatic” (aqua green), “weakly metastatic” (orange),”
855 metastatic” (red) according to the MetMap500 study. **$p<0.01$; ****$p<0.0001$. Right:
856 Association plot between the HMM-derived cell line states (rows) and their experimentally
857 measured metastatic potential (columns) ($p=2e-13$). (f) Distribution of the EMT states across
858 different cancer tissues. Each quarter of the pie corresponds to the 25% of the data. The
859 number of samples analysed is indicated for each tissue.

860

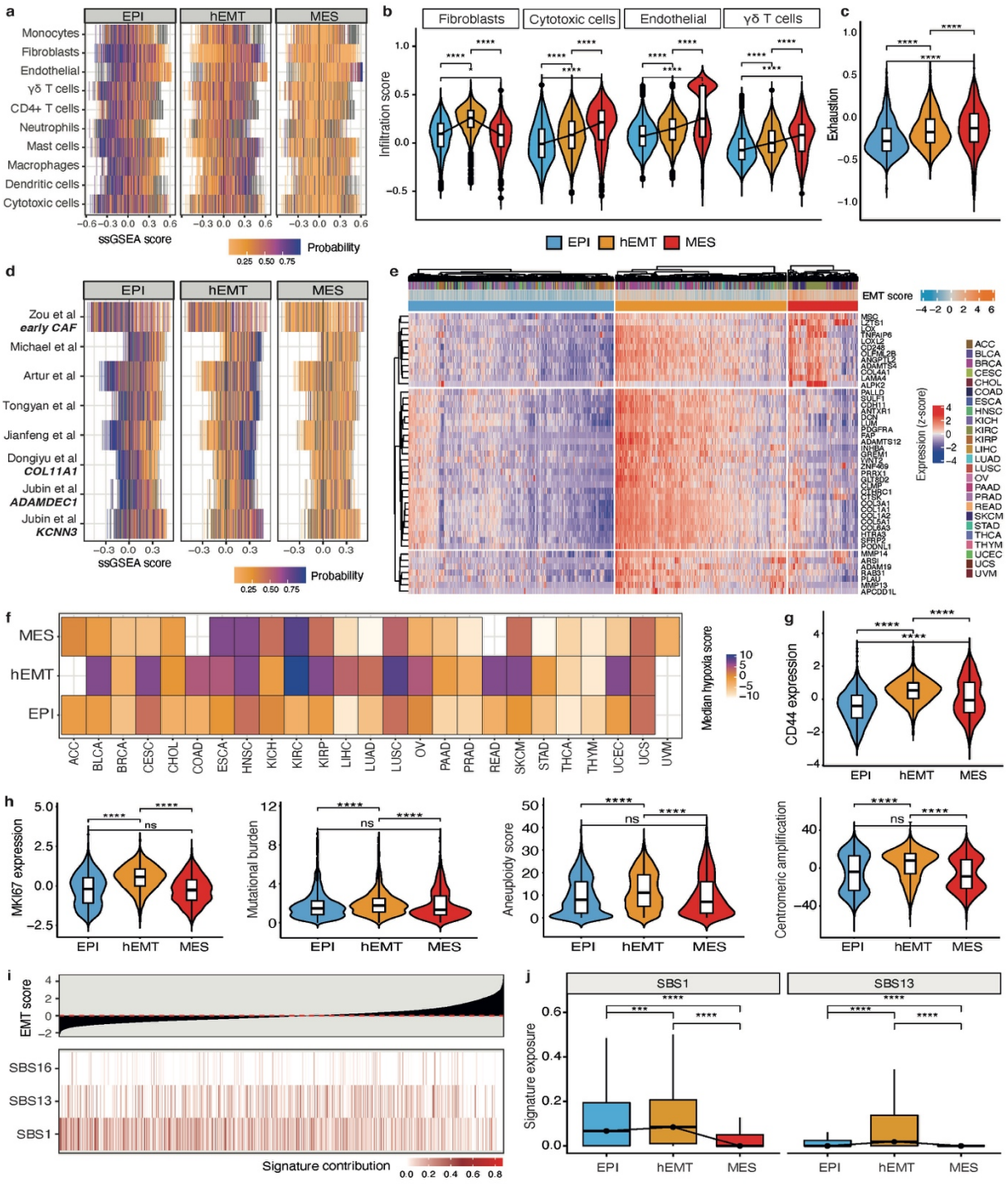



**Figure 2. Tumour extrinsic and intrinsic hallmarks of EMT.** (a) Heat map showcasing the results of a multinomial logistic regression model trained to predict EMT states based on cell infiltration in the microenvironment. Each row corresponds to a cell type and the corresponding per-sample infiltration 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

868 type. (b) Cell abundance compared across the EMT states for significantly predicted cell types
869 in the multinomial analysis. (c) Levels of exhaustion quantified across the three EMT states.
870 (d) Heat map displaying the results a multinomial logistic regression model predicting EMT
871 state using fibroblast signatures from multiple publications. (e) Heat map displaying the pan-
872 cancer expression of the fibroblast COL11A1_FS signature genes. Highest expression is
873 observed in the hEMT state. (f) Median hypoxia values in the three different EMT states
874 across tissues. (g) Gene expression levels of the stemness marker CD44 compared across
875 the three EMT states. (h) Expression of the proliferation marker Ki67, mutational burden,
876 aneuploidy, and centromeric amplification levels compared across the three EMT states. (i)
877 Mutational signature exposures across TCGA samples sorted by EMT score. Only mutational
878 signatures that were significantly linked with EMT from the linear mixed models are displayed.
879 The corresponding EMT scores are displayed above. (j) Signature contributions from SBS1
880 and SBS13 compared between the three EMT states.

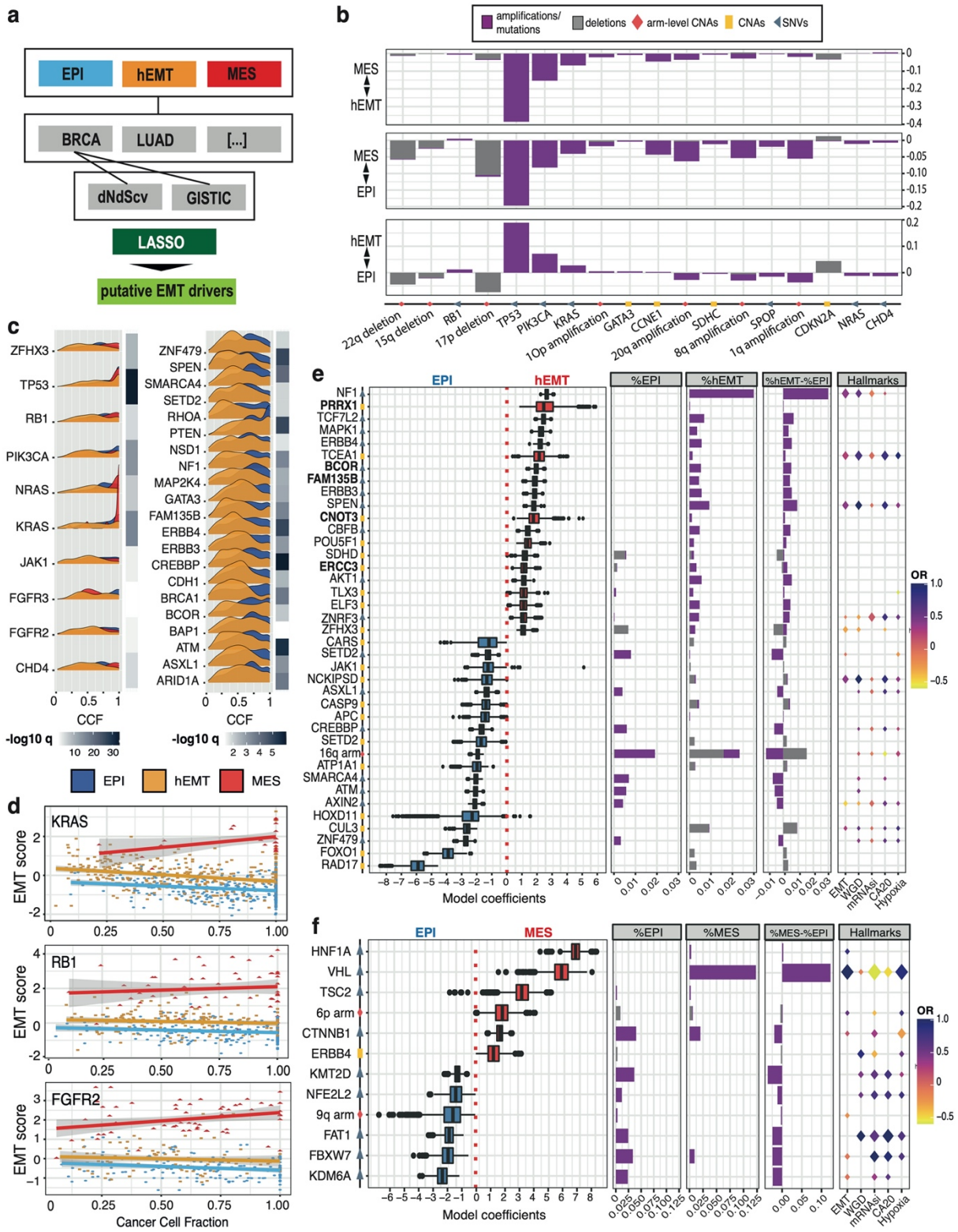



**Figure 3. Genomic driver events linked with EMT.** (a) The analytical workflow used to detect putative EMT biomarkers. For each state and cancer type, we used dNdScv and GISTIC to prioritise mutated genes and copy number events, respectively. These genomic

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

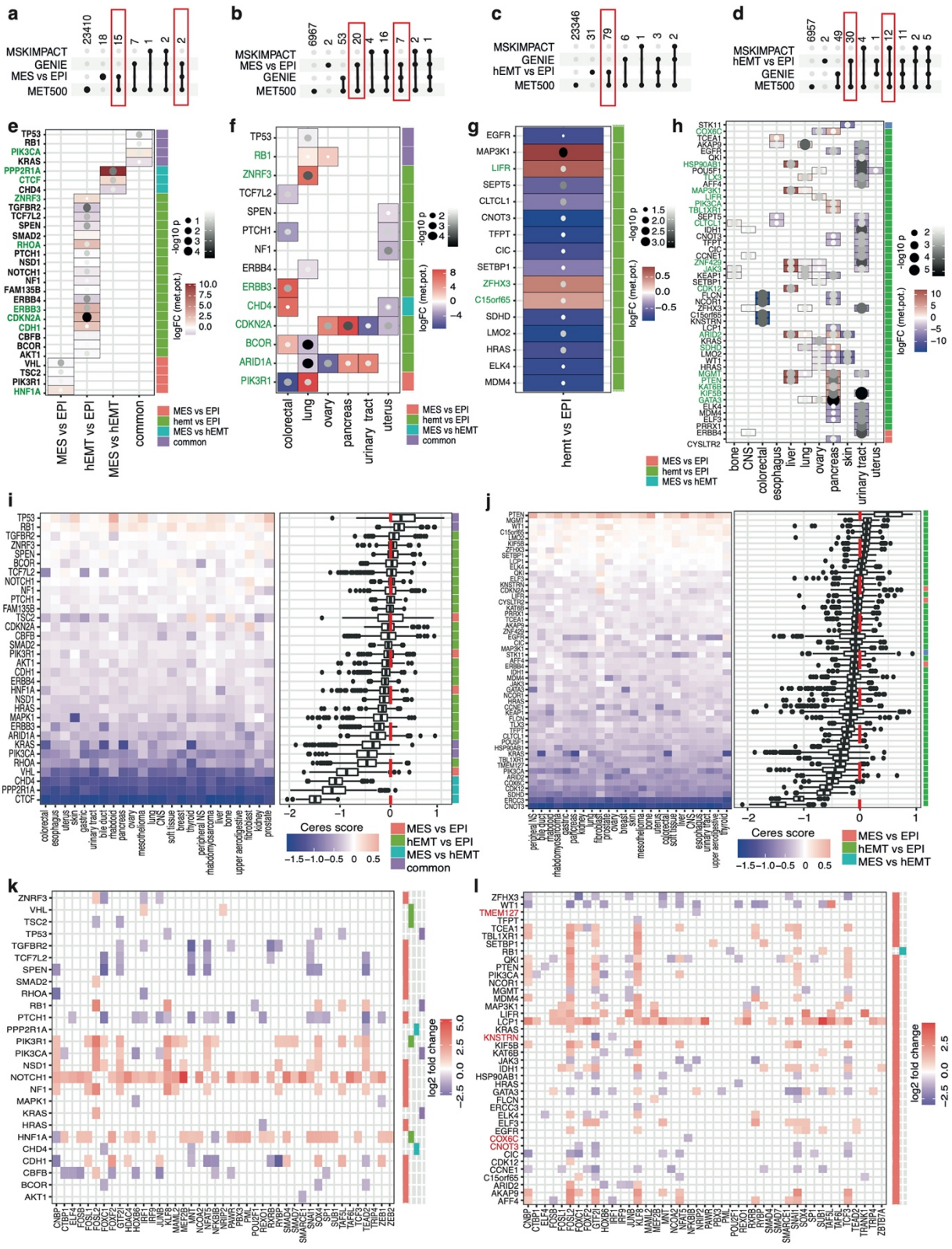



**Figure 4. Validation of the pan-cancer genomic associations with EMT in external datasets.** (a) UpSet chart displaying the overlap with external datasets for copy number events derived from the MES vs EPI comparison. (b) Overlap with external datasets for

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

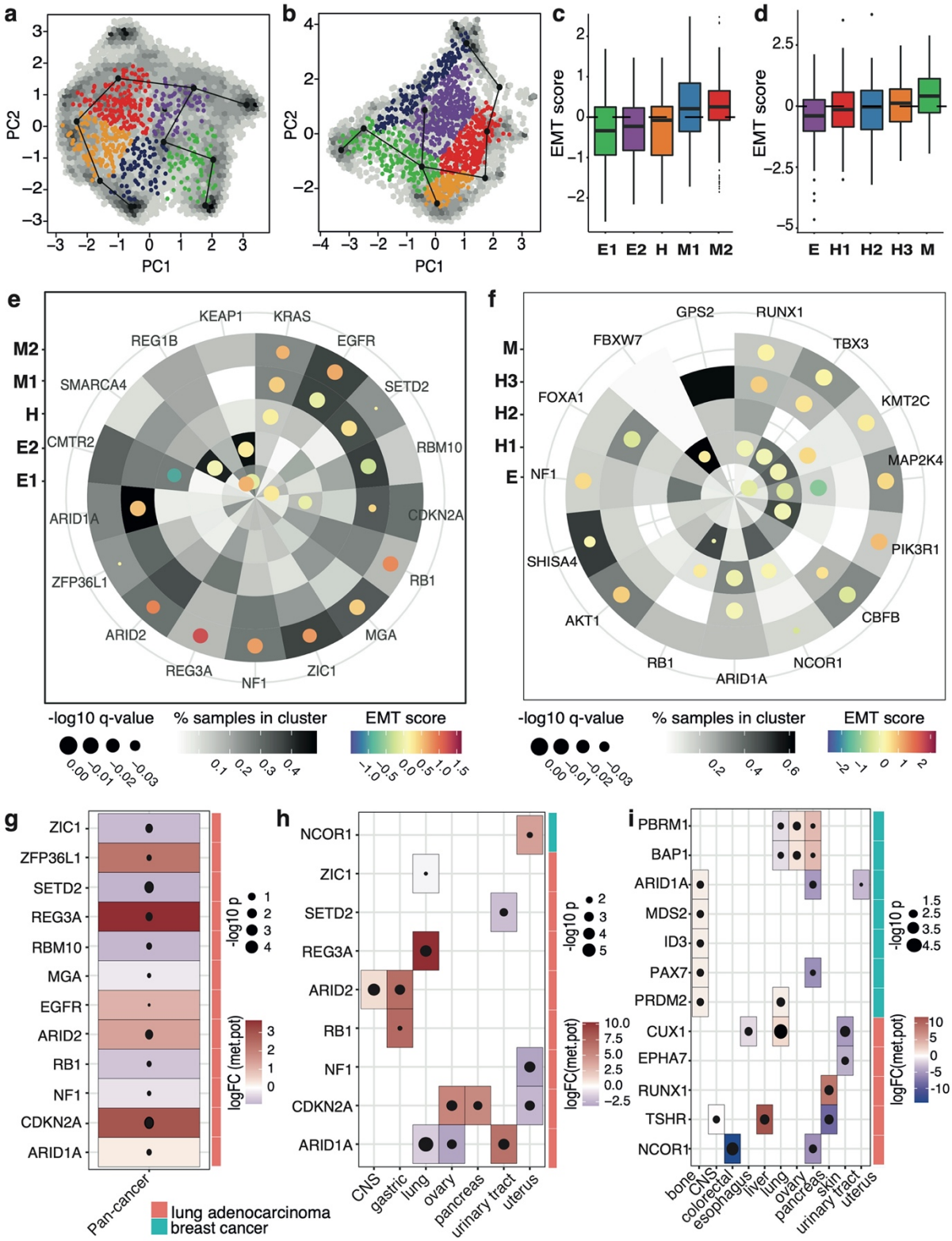



**Figure 5. EMT trajectories and associated genomic hallmarks in lung and breast cancer.** (a) Lung adenocarcinoma samples from TCGA mapped onto the single cell EMT trajectories derived in the A579 cell line. The first two principal components of the scRNA-seq

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

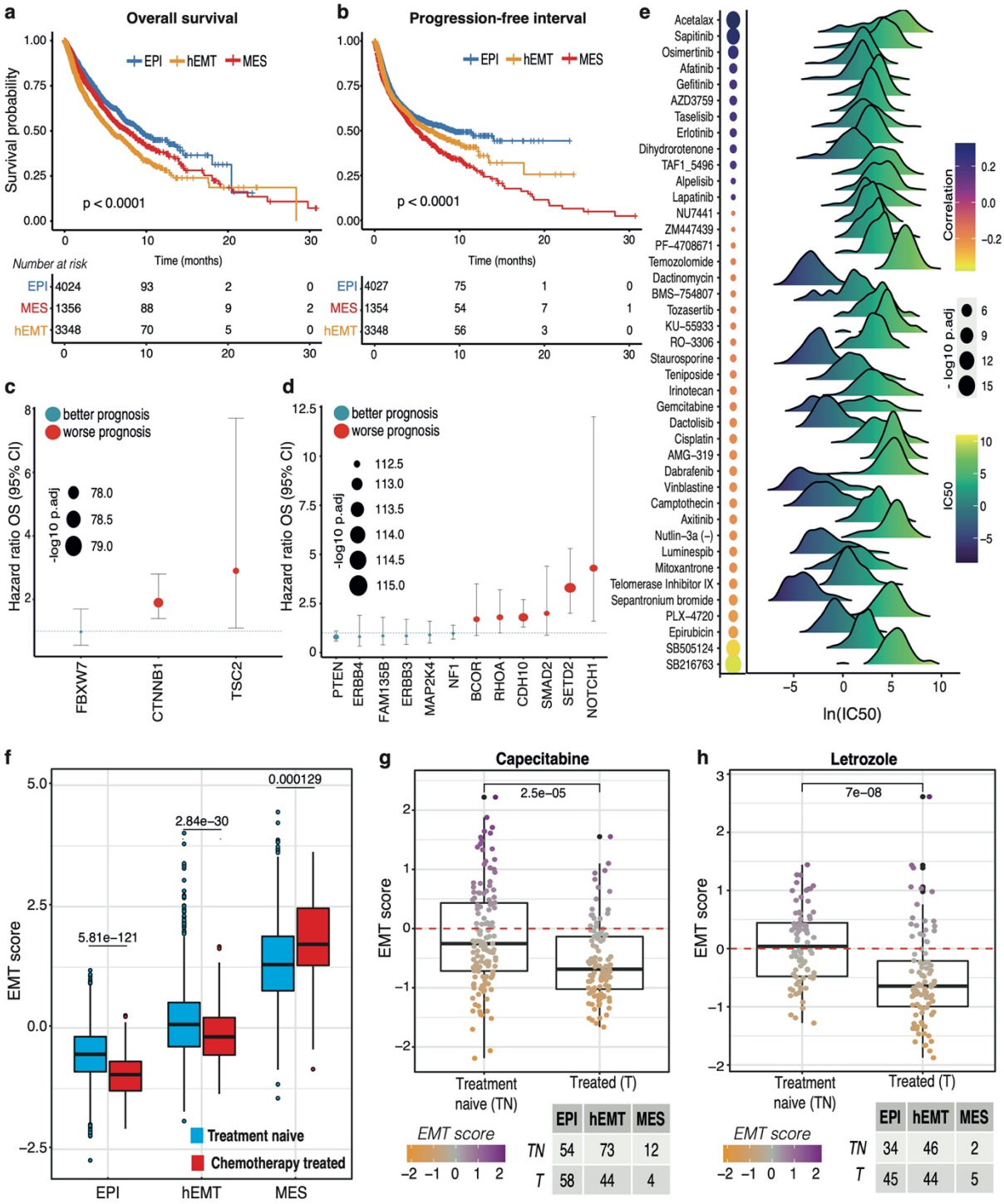
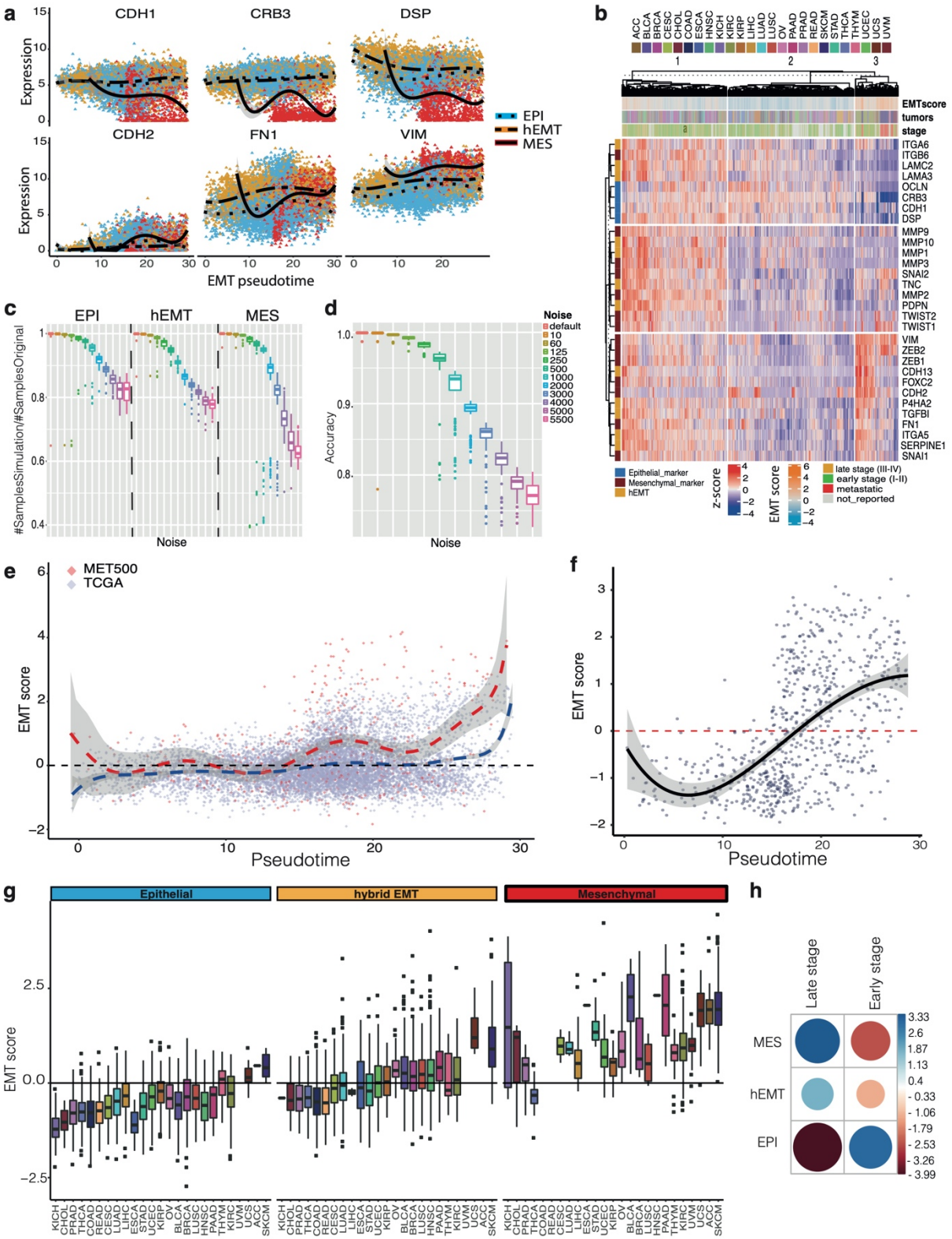



**Figure 6. Clinical relevance of the EMT states.** (a) Overall survival compared between MES, hEMT and EPI samples. (b) Progression free interval compared between the three groups. (c) Mutated markers of EMT pan-cancer with a significantly worse or improved outcome between the mesenchymal and epithelial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

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

953

954 **SUPPLEMENTARY MATERIAL**

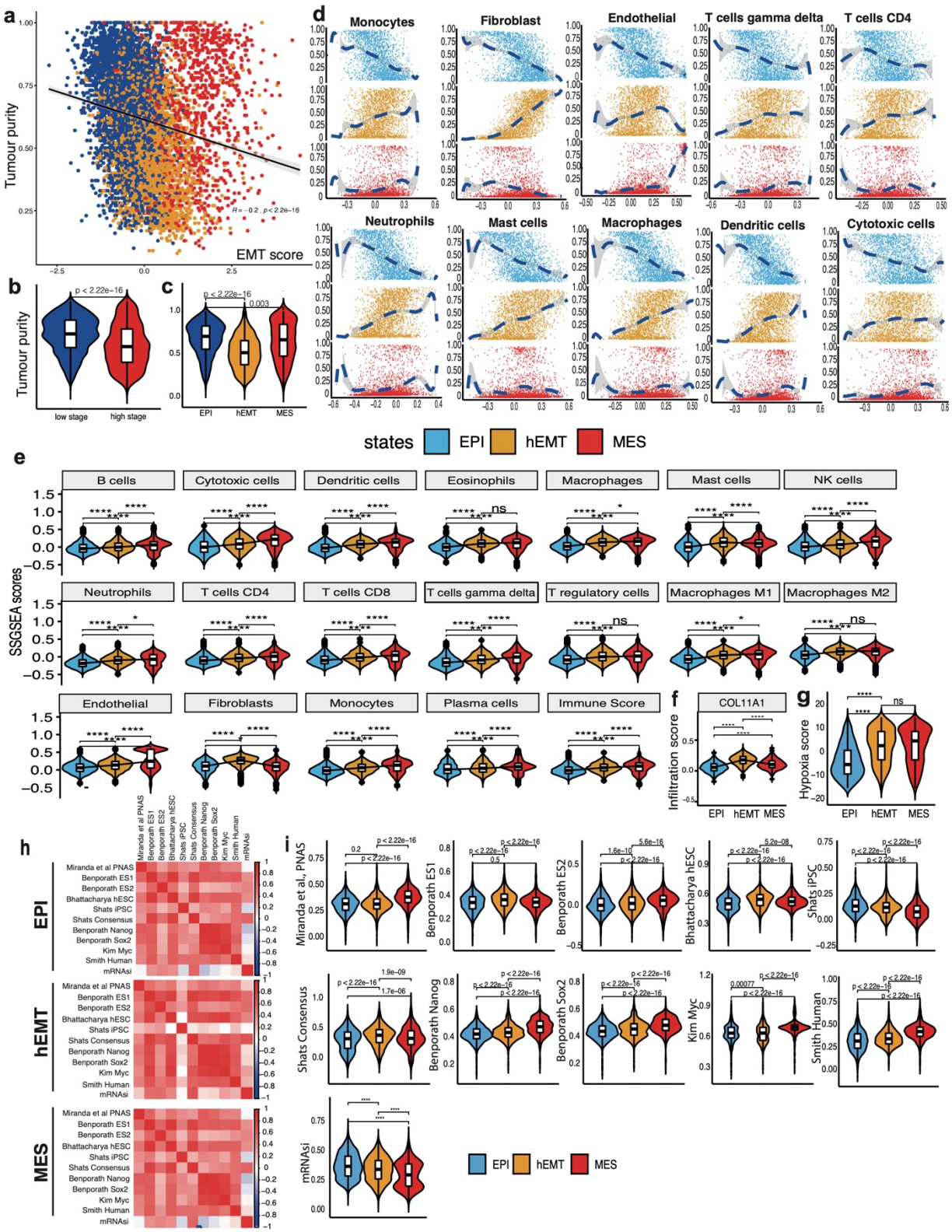



955

956 **Supplementary Figure S1. EMT trajectories and states in the TCGA and MET500**

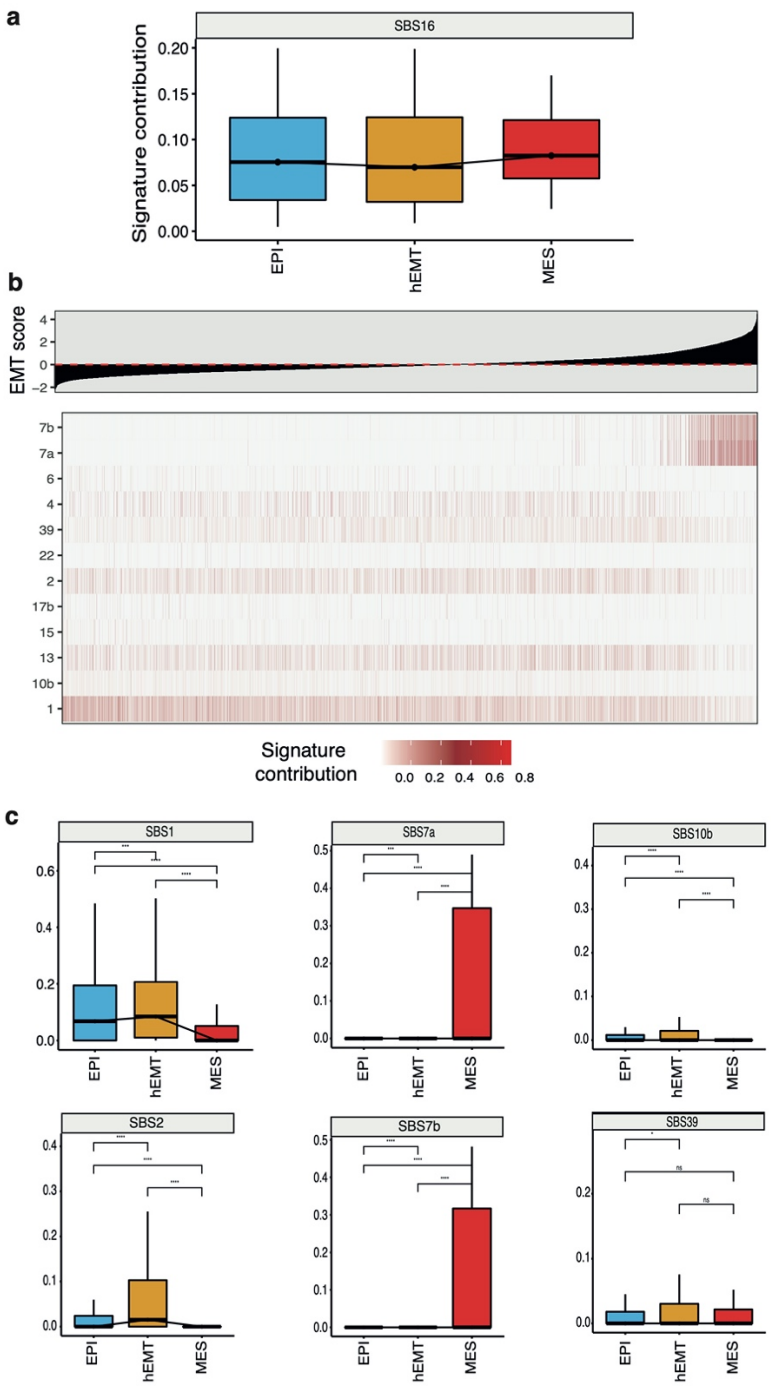
957 **datasets**. (a) Expression of epithelial/mesenchymal markers along the EMT pseudotime

958 derived from single cell data. Each dot represents a TCGA sample and is coloured according
959 to the assigned EMT state. (b) Heat map highlighting the pan-cancer expression of known
960 EMT markers (coloured by their state-specificity). (c) Fraction of samples correctly assigned to
961 the original HMM state with increasing levels of gene expression noise in the original data. (d)
962 Accuracy of predicting the original HMM states with increasing levels of gene expression
963 noise in the original data. (e) EMT scores of the MET500 (red points) and TCGA (blue points)
964 samples plotted along the EMT pseudotime derived from single cell data. (f) EMT scores of
965 the CCLE samples plotted along the EMT pseudotime. (g) EMT scores compared across
966 tissues and between EMT macro-states. (h) Relation between the EMT states and the clinical
967 cancer stage, computed using Chi-square statistics. The circles represent the associations
968 between EMT states and cancer stage (early/late). The colour and size define the strength of
969 the association.

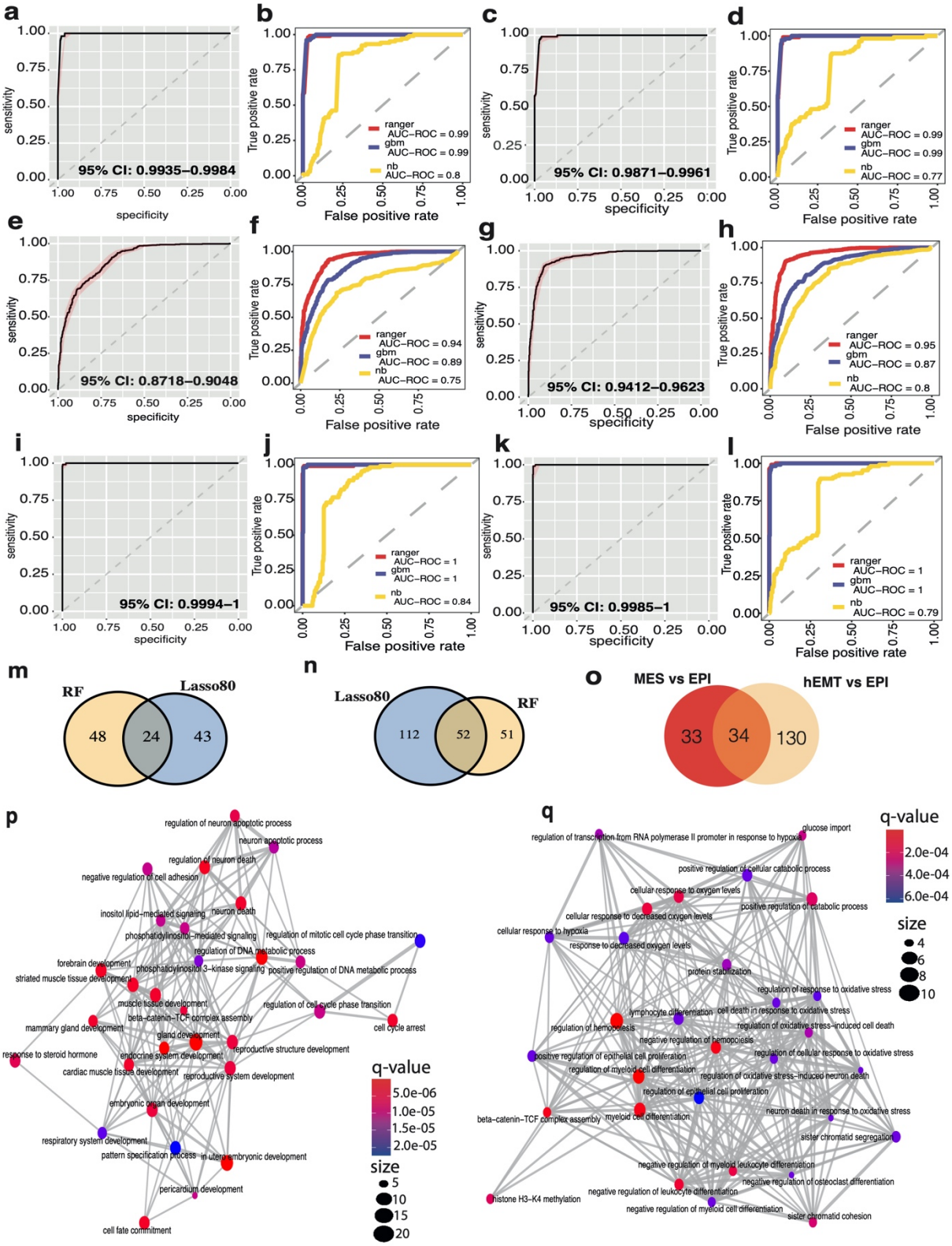



**Supplementary Figure S2. Tumour extrinsic properties in relation to EMT.** (a) Scatter plot of tumour purity values (y axis) against EMT scores (x axis). Every dot corresponds to a sample and is coloured according to the EMT macro-state. The Pearson correlation coefficient

and p-value of the association are also displayed. (b-c) Tumour purity compared between discrete states of low and high stage of EMT transformation pan-cancer (EMT score<0 – low stage; EMT score>0 – high stage). (c) Tumour purity compared between the three discrete EMT macro-states pan-cancer. (d) The results of the multinomial logistic regression models trained to predict EMT state based on non-cancer cell infiltration. The x-axis displays the ssGSEA scores of a specific microenvironmental component. The y-axis visualizes the probability of a sample being predicted to belong to a specific EMT state (EPI/hEMT/MES). Every dot is a TCGA sample. (e) Tumour microenvironment composition compared between the three EMT macro-states. (f) Active fibroblasts infiltration levels (quantified based on the COL11A1 signature) compared between the three EMT macro-states. (g) Hypoxia scores compared between the three EMT macro-states pan-cancer. (h) Correlation heat maps of the stemness scores assessed in this study. The strength of the Pearson correlation is highlighted by the colour gradient. (i) Stemness scores compared between each EMT state.

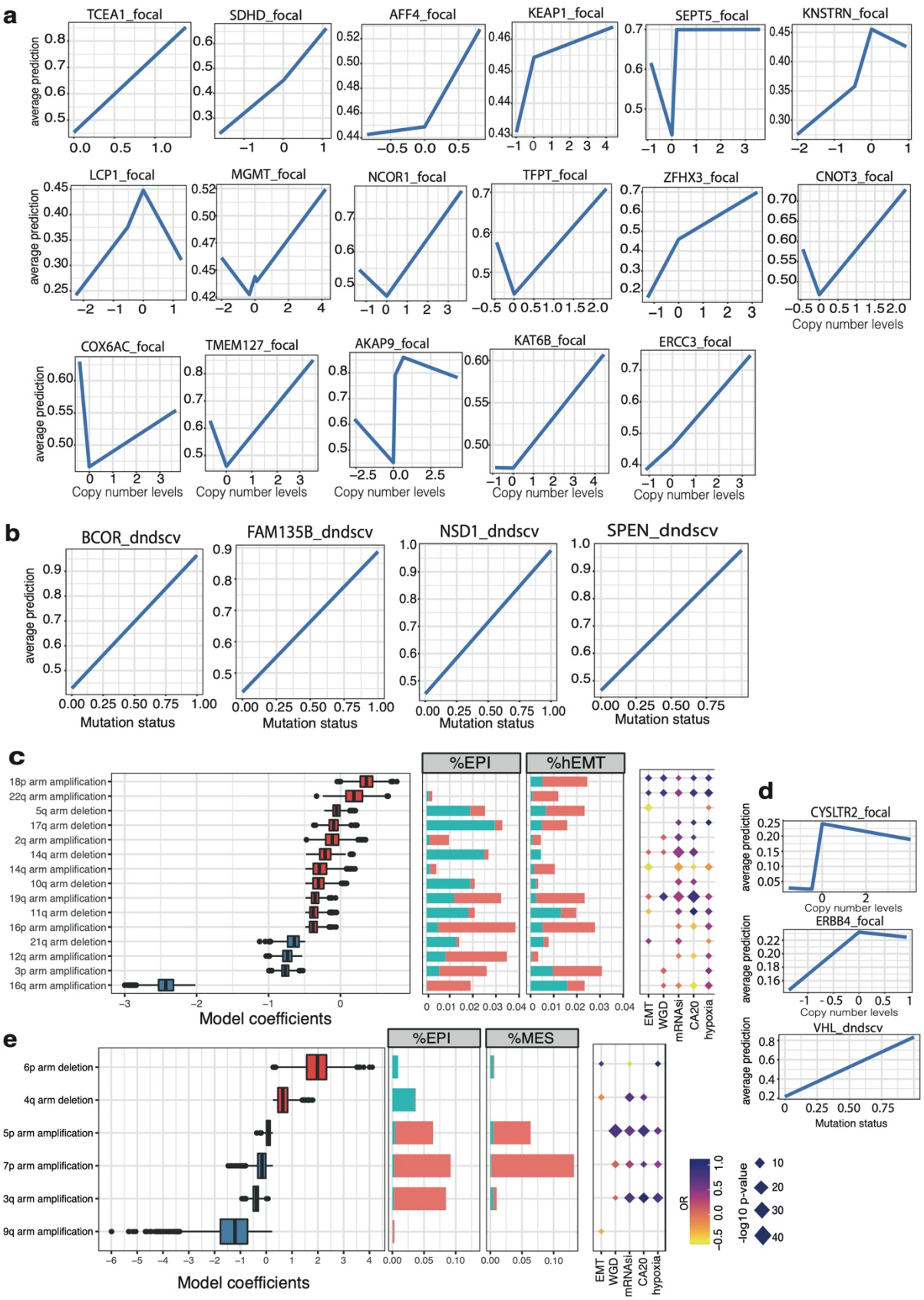



**Supplementary Figure S3. Mutational signatures across EMT states.** (a) Mutational contribution of SBS16 compared between EMT states. (b) Mutational contributions across TCGA samples for signatures derived as important in determining EMT by linear modelling without accounting for the tissue effect. EMT scores are reported in the top panel and samples are sorted according to these scores. Each value in the heatmap represents the contribution of a specific signature to the genome in one sample. (c) Mutational contributions compared between samples in the three EMT macro-states for selected signatures.



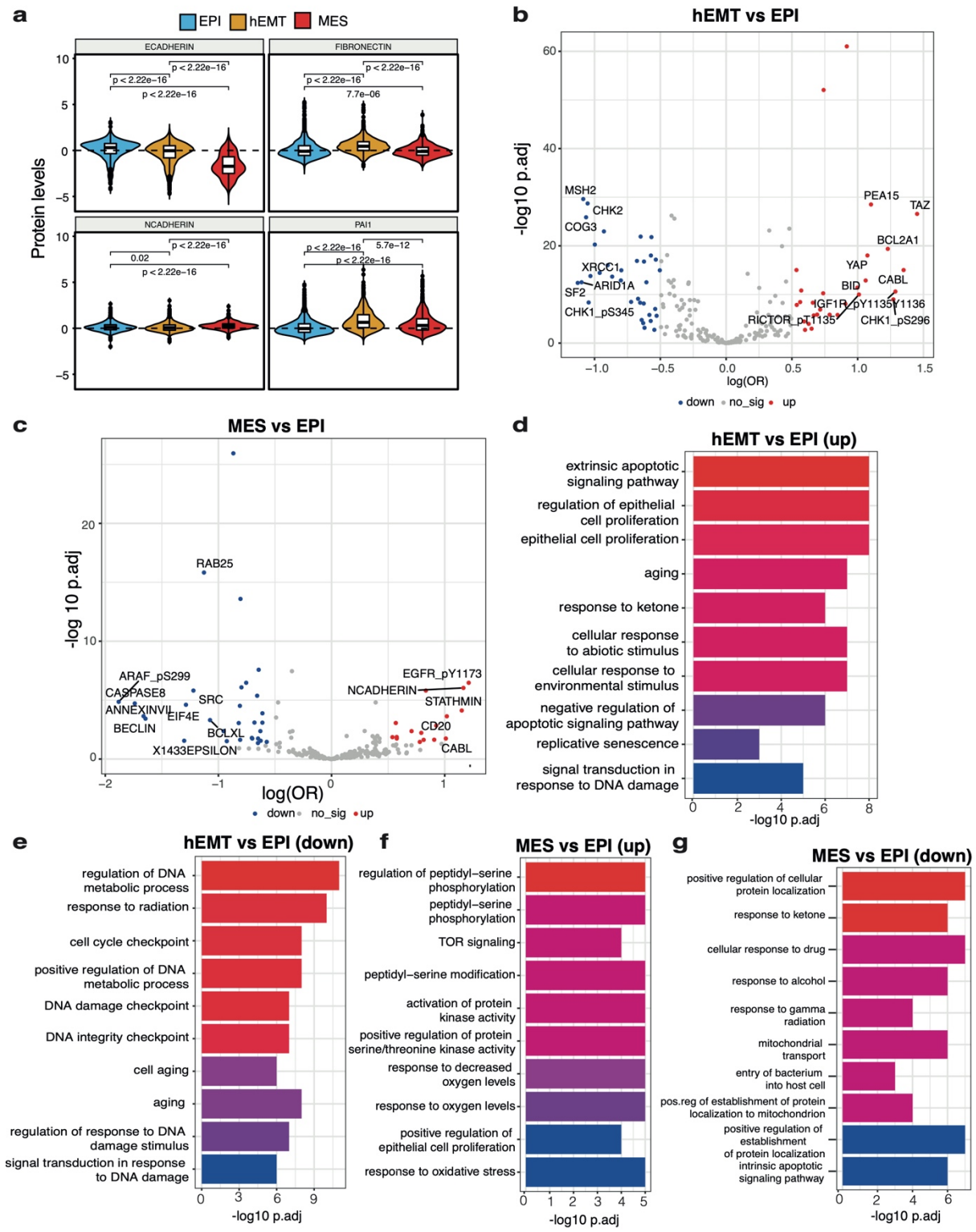

**Supplementary Figure S4. Discovery of genomic alterations linked with the EMT macro-states.** (a) ROC curves of the MES vs EPI lasso model predictions based on genomic markers obtained via the lasso procedure. (b) ROC curves of the MES vs EPI predictions based on genomic markers obtained with the lasso procedure and tested using different

1000 machine learning approaches. (c) ROC curves of the MES vs EPI lasso model predictions
1001 based on genomic markers obtained with the random forest procedure. (d) ROC curves of the
1002 MES vs EPI predictions based on genomic markers obtained with the random forest
1003 procedure and tested using different machine learning approaches. (e-h) Same as (a-d) but
1004 for the hEMT vs EPI model. (i-l) same as (a-d) but for the MES vs hEMT model. (m) Venn
1005 diagram highlighting common and specific genomic markers in the MES vs EPI comparison
1006 obtained using the lasso and random forest models, respectively. (n) Venn diagram
1007 highlighting common and specific genomic markers in the hEMT vs EPI comparison obtained
1008 using the lasso and random forest models, respectively. (o) Venn diagram highlighting
1009 common and specific genomic markers between the EMT state comparison models built using
1010 lasso. (p) Gene ontology terms enriched in the list of genes that distinguish the hEMT from the
1011 EPI group. (q) Gene ontology terms enriched in the list of genes that distinguish the MES from
1012 the EPI group.



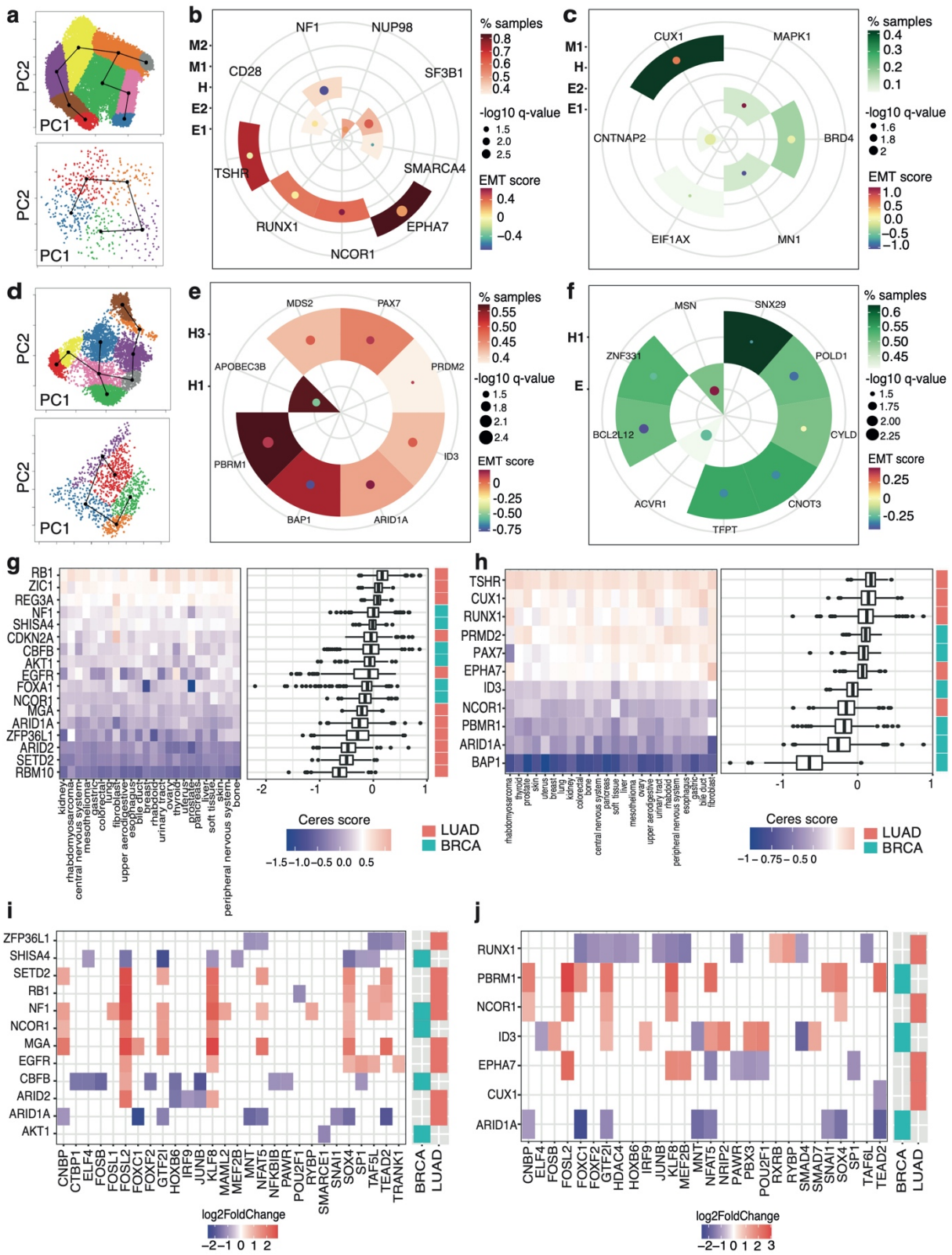

**Supplementary Figure S5. Characterization of the genomic alterations linked with the EMT macro-states.** (a-b) Partial dependence charts of selected gene markers from the hEMT vs EPI comparison model. The probability of a sample belonging to the hEMT group (y-axis) is

1017 plotted against (a) different levels of copy number (numbers <0 indicate losses; 0=copy
1018 number neutral; numbers >0 indicate gains) or (b) the presence of mutations (0/1:
1019 absent/present). (c) Chromosomal arm events distinguishing the hEMT and EPI states.
1020 Boxplots highlight the coefficients derived from the lasso analysis (hEMT vs EPI) of the
1021 chromosomal arm events. The horizontal bar plots in the middle panels indicate the pan-
1022 cancer percentage of samples with an amplification (red) or deletion (blue) of a specific
1023 chromosomal arm. The balloon chart on the right highlights the associations between cancer
1024 hallmarks and the chromosomal arm events. Only significant associations are displayed. (d)
1025 Partial dependence charts of the MES vs EPI gene markers. The probability of a sample
1026 belonging to the MES group (y-axis) is plotted against different levels of copy number
1027 (numbers <0 indicate losses; 0=copy number neutral; numbers >0 indicate gains) or the
1028 presence of mutations (0/1: absent/present). (e) Chromosomal arm events distinguishing the
1029 MES and EPI states. Boxplots depict the coefficients derived from the lasso analysis (MES vs
1030 EPI) of the chromosomal arm events. The horizontal bar plots in the middle panels indicate
1031 the pan-cancer percentage of samples with an amplification (red) or deletion (blue) of a
1032 specific chromosomal arm. The balloon chart on the right highlights the associations between
1033 cancer hallmarks and the chromosomal arm events. Only significant associations are
1034 displayed.



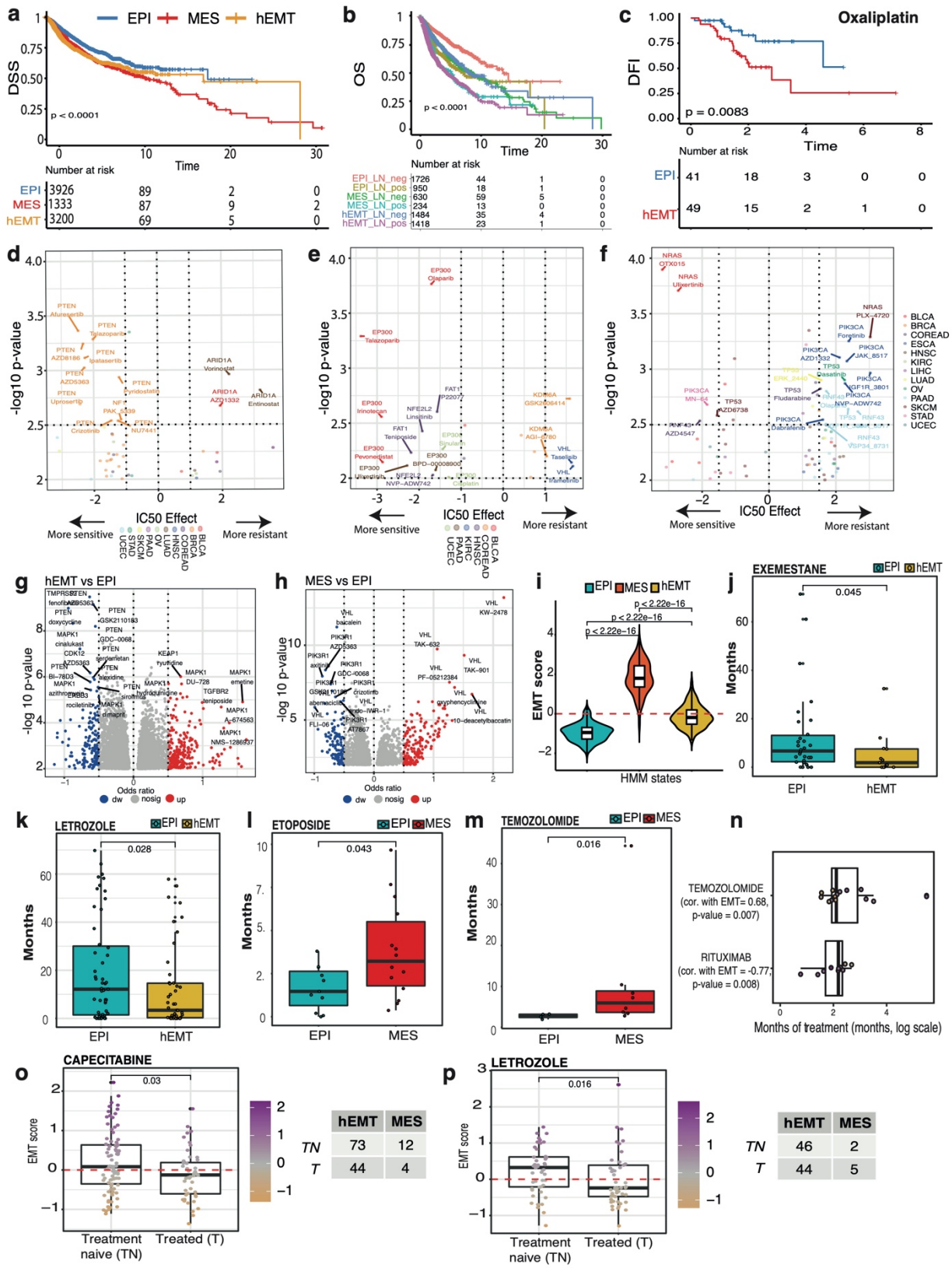

**Supplementary Figure S6. Proteomic profiling across the EMT states in TCGA.** (a) Protein expression levels of canonical EMT markers compared between EPI, hEMT and MES states. (b) Proteins with significantly different expression between hEMT and EPI states. The x-axis and y-axis contain the OR and p-values obtained from the logistic regression analysis, respectively. (c) Proteins with significantly different expression between MES and EPI states.

1041 The x-axis and y-axis contain the OR and p-values obtained from the logistic regression
1042 analysis, respectively. Panels (d-g) depict the results of gene ontology enrichment on: (d) up-
1043 regulated proteins between hEMT and EPI, (e) down-regulated proteins between hEMT and
1044 EPI, (f) up-regulated proteins between MES and EPI, (g) down-regulated proteins between
1045 MES and EPI.




**Supplementary Figure S7. EMT states and genomic features of lung and breast cancers.** (a) Top: results of slingshot analysis of EMT trajectories using the A549 cell line data. The samples in each cluster are highlighted using different colours. Bottom: results of

1050 slingshot analysis of EMT trajectories using the TCGA LUAD samples. Panels (b-c) display
1051 circular heat maps with copy number amplifications (b) and deletions (c) significantly
1052 associated EMT clusters in TCGA LUAD. The colour of each tile corresponds to the number of
1053 samples with alterations in the respective cluster. The circles inside the modules highlight the
1054 EMT scores of the TCGA samples in each cluster in the cases where the association between
1055 the genomic event and EMT cluster is significant. (d) Top: results of slingshot analysis of EMT
1056 trajectories using the MCF7 cell line data. Bottom: results of slingshot analysis of EMT
1057 trajectories using the TCGA BRCA samples. Panels (e-f) display circular heat maps with copy
1058 number amplifications (e) and deletions (f) significantly associated EMT clusters in TCGA
1059 BRCA. (g) Essentiality across cell lines for genes whose mutations are linked with EMT in
1060 lung or breast cancer. The heat map gradient indicates the CERES scores, with negative
1061 values indicating increased essentiality. The boxplots in the right hand-side panel display the
1062 CERES score distribution across all lineages. (h) Similar to (g) but considering the genes
1063 linked with EMT via copy number alteration. (i) Genes mutated in relation to EMT in
1064 breast/lung cancer that are dysregulated as a result of the knockdown of transcription factors
1065 relevant for EMT. The rows display the EMT genomic markers obtained from previous
1066 analyses, the columns indicate the transcription factors. The colour gradient reflects the log2
1067 fold change in expression of the genes upon transcription factor knockdown. Only significant
1068 changes are displayed (adjusted $p<0.05$). (j) Similar to (i) but considering the EMT markers
1069 with alterations in copy number. Only lung EMT biomarkers showed changes in expression as
1070 a result of transcription factor silencing.




**Supplementary Figure S8. Prognosis and therapeutic relevance of the EMT states.** (a) Disease-specific survival of patients in TCGA by EMT macro-state of their corresponding tumours. (b) Overall survival of patients in TCGA, stratified by the EMT macro-state of the

tumour and lymph node positivity. (c) Progression-free interval of the patients treated with Oxaliplatin, by EMT macro-state of the tumour. (d) Significance of association with various compounds plotted against the IC50 scores of cell lines from DepMap harbouring mutations in genes identified between hEMT and EPI. E.g. cell lines harbouring mutations in the hEMT marker ARID1A are more resistant to entinostat, vorinostat and AZD1332. (e) Significance of association with various compounds plotted against the IC50 scores of cell lines from DepMap harbouring mutations in genes identified between MES and EPI. (f) Significance of association with various compounds plotted against the IC50 scores of cell lines from DepMap harbouring mutations in genes identified commonly between transformed and EPI tumours. (g) Differential cell viability after chemical perturbation with various compounds from the PRISM drug repurposing resource for cell lines harbouring mutations identified by the hEMT vs EPI comparison model. E.g. cell lines with mutations in the hEMT marker MAPK1 show an increased cell viability after treatment with emetine. (h) Similar to (g), but only the genes mutated in the MES vs EPI comparison have been considered. (i) EMT scores are compared across distinct macro-states following EMT reconstruction of the post-treated samples from POG570. The differences observed confirm the classification. (j-n) Duration of treatment (as proxy for response) is compared between EMT macro-states for different drugs. (o-p) EMT scores compared between unmatched but batch effect-corrected TCGA samples (treatment-naïve) and POG570 samples after capecitabine and letrozole treatment, respectively. Significant decreases in EMT score can be observed in both instances. Only transformed samples (hEMT, MES) are considered for this analysis (numbers indicated in adjacent tables).

1105 **Supplementary Table S1. Pseudotime reconstruction of EMT trajectories in TCGA**
1106 **samples.** The results of the pan-cancer reconstruction of epithelial-to-mesenchymal
1107 trajectories is reported, along with the HMM state of the sample, and the corresponding EMT
1108 macro-state and score.

1109 **Supplementary Table S2.** Distribution of the TCGA samples in each EMT state across
1110 clinical cancer stages.

1111 **Supplementary Table S3. Catalogue of fibroblast gene-sets.** This table contains the gene
1112 sets used to quantify the fibroblast infiltration from bulk RNA-seq data of TCGA samples.

1113 **Supplementary Table S4.** Stratification of samples according to hypoxia score and CD44
1114 expression.

1115 **Supplementary Table S5**. **Mutational signatures linked with EMT states pan-cancer.** The
1116 results of the linear mixed model trained to predict EMT state pan-cancer based on mutational
1117 signature contributions while accounting for the tissue effect are reported. Only significant
1118 mutational signatures are displayed ($q < 0.05$). The columns report the false discovery rate
1119 (FDR), the marginal ($R^2M$) and conditional ($R^2C$) residuals.

1120 **Supplementary Table S6**. **Pan-cancer genomic markers linked with EMT states and**
1121 **protein analysis results.** (a) The results of the lasso models built to distinguish EMT states
1122 based on genomic markers. The first column indicates the comparison in which a biomarker
1123 (second column) has been identified. Additional details on the genes are provided. The
1124 coefficients from the lasso model (Median, Mean) along with the confidence intervals (Upper,
1125 Lower) are reported. The final two columns annotate if a gene is known to be relevant for
1126 EMT. (b) The results of the random forest models built to distinguish EMT states based on
1127 genomic markers. (c) Overlap between the genomic markers of EMT discovered via lasso
1128 modelling and a known list of genes with roles in cell migration from Novikov et al (Br J
1129 Cancer 2021). The cells marked with 1 highlight genes that were recapitulated by our

1130 analysis and are known markers of migration. (d) Differentially expressed proteins between
1131 EMT states.

1132 **Supplementary Table S7. Genomic dependencies identified in the tissue specific EMT**
1133 **trajectory analysis.** (a) Cancer drivers positively selected in each lung adenocarcinoma EMT
1134 cluster. The rows depict the genes detected by dNdScv, while columns highlight the derived
1135 EMT clusters obtained by projecting the bulk TCGA LUAD RNA-seq data onto the scRNA-seq
1136 of the A549 cell line. Genes expressed in the cohort are indicated in bold. The last column
1137 contains the overlap across the clusters for a given gene. (b) Number of samples containing
1138 mutations in the positively selected genes in EMT clusters from lung adenocarcinoma. The
1139 rows depict genes detected by dNdScv, while columns highlight the EMT clusters obtained by
1140 projecting of the TCGA LUAD RNA-seq data onto the scRNA-seq of A549 cell line. The cells
1141 contain the number of samples with a driver gene mutation in each cluster. (c) Overlap
1142 between mutation events linked with EMT derived from the pan-cancer and lung-specific
1143 analyses. Markers uniquely identified in LUAD are highlighted in bold, along corresponding
1144 clusters in brackets. (d) Gene ontology enrichment analysis for gene markers positively
1145 selected in different EMT states in lung adenocarcinoma. The table contains the top 10
1146 significant gene ontology terms obtained using EnrichR. (e) Overlap between focal copy
1147 number events linked with EMT derived from the pan-cancer and lung-specific analyses.
1148 Markers uniquely identified in LUAD are highlighted in bold, along corresponding clusters in
1149 brackets. The markers uniquely identified in the pan-cancer analysis have been removed. (f)
1150 Cancer drivers positively selected in each breast cancer EMT cluster from TCGA. (g) Number
1151 of samples containing mutations in the positively selected genes in EMT clusters from breast
1152 cancer. (h) Overlap between mutation events linked with EMT derived from the pan-cancer
1153 and breast-specific analyses. (i) Gene ontology enrichment analysis for gene markers
1154 positively selected in different EMT states in breast cancer. (j) Overlap between focal copy
1155 number events linked with EMT derived from the pan-cancer and breast-specific analyses.

1156 **Supplementary Table S8. Cox regression model results on clinical end points.** (a)
1157 Results of the multivariate Cox regression model for the overall survival. The mean and
1158 standard deviation are reported for each covariate along with the hazard ratios (HR) from the
1159 univariable and multivariable models. The p-value for the model is highly significant (p<1e-30).
1160 (b) Results of the multivariate Cox regression model for the progression free interval (PFI).
1161 The mean and standard deviation are reported for each covariate along with the hazard ratios
1162 (HR) from the univariable and multivariable models. The p-value for the model is highly
1163 significant (p<1e-10). (c) Results of the multivariate Cox regression model for the disease
1164 specific survival. The mean and standard deviation are reported for each covariate along with
1165 the hazard ratios (HR) from the univariable and multivariable models. The p-value for the
1166 model is highly significant (p<1e-33).

1167 **Supplementary Table S9. Mutation events impacting overall survival in TCGA.** The first
1168 column indicates the EMT comparison model the marker was derived from, the second
1169 column reports the name of the gene. Hazard ratios (log10) are reported along with
1170 confidence intervals, p-values and outcome (Out: 'positive' indicates better prognosis for
1171 patients harbouring the mutation in the respective gene, 'negative' indicates worse prognosis).