Mesenchymal-epithelial transition in lymph node metastases of oral squamous cell carcinoma is accompanied by ZEB1 expression

Kai Horny[1,2], Christoph Sproll[3], Lukas Peiffer[1,2], Frauke Furtmann[1,2,4], Patricia Gerhardt[1,2], Jan Gravemeyer[1,2], Nikolas H. Stoecklein[5], Ivelina Spassova[1,2,4], Jürgen C. Becker[1,2,4,*]

[1] Translational Skin Cancer Research, German Cancer Consortium (DKTK), 45141 Essen, Germany

[2] German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

[3] Department of Oral- and Maxillofacial Surgery, Medical Faculty, University Hospital of the Heinrich-Heine-University Düsseldorf, Germany

[4] Department of Dermatology, University Medicine Essen, 45141 Essen, Germany

[5] Department of General, Visceral and Pediatric Surgery, Medical Faculty, University Hospital of the Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

* Correspondence: j.becker@dkfz.de

Abstract

Background

Oral squamous cell carcinoma (OSCC), an HPV-negative head and neck cancer, frequently metastasizes to the regional lymph nodes but only occasionally beyond. Initial phases of metastasis are associated with an epithelial-mesenchymal transition (EMT), the consolidation phase is associated with mesenchymal-epithelial transition (MET). This dynamic is referred to as epithelial-mesenchymal plasticity (EMP). While it is known that EMP is essential for cancer cell invasion and metastatic spread, less is known about the heterogeneity of EMP states within a tumor and even less about the heterogeneity between the primary and metastatic lesions.

Methods

To capture heterogeneity of EMP states in OSCC, we performed single-cell RNA sequencing (scRNAseq) of 5 primary tumors and 9 matching lymph node metastases and re-analyzed publicly available scRNAseq data of 9 additional primary tumors. To account for possible bias in cell type compositions by scRNAseq, these were also deconvoluted from bulk transcriptome analyses. Protein expression of selected genes were confirmed by immunohistochemistry.

Results

From the 23 OSCC lesions the single cell transcriptome of a total of 7,263 carcinoma cells was available for in-depth analyses. We initially focused on one lesion to avoid inter-patient heterogeneity as a confounding factor and identified OSCC cells expressing genes characteristic of different epithelial and partial EMT stages, such as keratins and *SPRR1B* (cornifin B) or vimentin and matrix metallopeptidases. RNA velocity information together with the increase in inferred copy number variations indicated a progressive trajectory towards epithelial differentiation in this metastatic lesion. Extension to all samples revealed a less stringent but essentially similar pattern. Interestingly, cells undergoing MET show increased activity of the EMT activator ZEB1. Immunohistochemistry confirmed that ZEB1 was co-expressed with the epithelial marker cornifin B in individual tumor cells - more frequently in

lymph node metastases. The lack of E-cadherin mRNA expression suggests this is a partial MET.

Conclusions

This study reveals that EMP enables different partial EMT and epithelial phenotypes of OSCC cells, which are endowed with capabilities essential for the different stages of the metastatic process, including maintenance of cellular integrity. During MET, ZEB1 appears to be functionally active, indicating a more complex role of ZEB1 than mere induction of EMT.

Keywords: single cell RNA; oral cavity; squamous cell carcinoma; Epithelial-mesenchymal plasticity; EMT; MET; ZEB1; heterogeneity

Background

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cancer worldwide, with 890,000 new cases and 450,000 deaths in 2018 (1). The survival for HNSCC patients has improved modestly over the past decades; however, this improvement is partially attributable to the emergence of human papillomavirus (HPV)-associated HNSCC which has a better prognosis than HPV-negative tumors. One of the HPV-negative HNSCC subtypes is oral cavity squamous cell carcinoma (OSCC) which is mainly associated with tobacco and alcohol abuse (1). OSCCs are often diagnosed at an early stage owing to the patient's self-identification of the mass lesion and symptoms, still regional lymph node metastases are frequent; thus, surgical removal of primary tumor is accompanied by neck dissection and radiotherapy (2). Given the morbidity associated with this combined intervention, there is a need to identify molecular biomarkers to predict the presence of lymph node metastases and to prognosticate survival.

Invasion and metastasis becomes possible in many epithelial tumors through an epithelial-mesenchymal transition (EMT), i.e., a reactivation of an embryonic developmental program in which cells acquire migratory and invasive properties (3). In EMT, epigenetic, transcriptional, and post-translational changes cause epithelial cells to break down the strong homotypic cell-cell junctions and adopt a mesenchymal morphology (4). It is important to note that EMT should not be understood as a clearly defined process, but rather as many dynamic and complex processes, which may vary depending on tumor entity, stage, and microenvironment (4-6). Thus, expression of EMT-related genes and their regulating transcription factors is highly heterogeneous even within one cancer entity, between patients, different lesions from one patient and between individual cancer cells within one lesion (4, 7). Since it is a continuous, dynamic, and reversible process cancer cells can adopt a multitude of intermediate or partial states, e.g., epithelial to more mesenchymal or partial EMT (pEMT) states (5, 8-11). Therefore, it has recently been recommended that this EMT continuum should rather referred to as epithelial-mesenchymal plasticity (EMP) (4, 10).

77 To capture the EMP-associated heterogeneity of cancers, single-cell analyses appear to be
78 the most appropriate approach, however, to date most EMP-related single-cell studies are
79 based on controlled *in vitro* and *in vivo* experiments (5, 6, 9, 12, 13). Particularly in HNSCC
80 only few studies scrutinized EMP within freshly isolated tumor samples (8, 14, 15). Especially
81 noteworthy is the landmark paper by Puram *et al.* They characterized 2,215 malignant cells
82 from 18 patients and discovered several pEMT states with high variability in EMP-related gene
83 expression (8).

84 In the work presented here, we investigated the phenotypic heterogeneity of 5 primary and 9
85 regionally metastatic OSCC lesions isolated from 7 patients using multiplexed single-cell RNA
86 sequencing (scRNAseq). In addition, we used a recently published series of scRNAseq data
87 from primary HNSCC that included 9 OSCC tumors to put our observations on an even broader
88 data base (15). Our results not only confirm the EMP-associated heterogeneity of cancer cells
89 in primary and metastatic OSCC, but also demonstrate mesenchymal-epithelial transition
90 (MET) in established lymph node metastases. Surprisingly, we observed a high activity of
91 EMT-activator ZEB1 in metastatic OSCC cells with epithelial differentiation, a notion which was
92 confirmed by co-expression of ZEB1 and cornifin-B protein in individual tumor cells.

Methods

Tissue samples

19 tissue samples that included 5 primary tumors, 14 potentially affected lymph node metastases, from which 9 represent lymph node metastases as we detected carcinoma cells in histopathological examination and scRNAseq from 7 OSCC patients treated at the Department of Oral-, Maxillo- and Plastic Facial Surgery, University Hospital of the Heinrich-Heine-University Düsseldorf, Germany, were used for histology, IHC, immune fluorescence as well as bulk and single-cell transcriptome analyses. The clinical details are provided in table 1.

Histology and IHC

Hematoxylin and Eosin (H&E) and IHC were performed on 4 µm formalin-fixed, paraffin-embedded (FFPE) sections. H&E staining was performed using standard protocols (Supplementary Figure 1). Whole-slide imaging was performed using Zeiss Axioscan 7 and 10x magnification (Carl Zeiss Microscopy Deutschland GmbH, Oberkochen, Germany).

IHC using the rabbit polyclonal antibodies anti-SPRR1B (Cat. No.: SAB1301567-400UL, Sigma Aldrich, Darmstadt, Germany) anti-ZEB1 antibody (Cat. No. HPA027524-25UL, Sigma Aldrich) was performed as previously described (16). Briefly, after sections were deparaffinized for 60 min at 60°C and rehydrated, sections were incubated for 15 min in an inverter microwave oven with antigen retrieval buffer pH 9 for anti-SPRR1B and pH 6 for anti-ZEB1. After 3x2 min washes with Tris-buffered saline with 0.1% Tween (TBST) sections were incubated for 8 min with 3% peroxidase and after an additional washing step for 30 min with 3% bovine serum albumin (BSA) in TBST. For single staining sections were subjected for 1h to anti-SPRR1B in 1:600 dilution or overnight to anti-ZEB1 in a 1:500 dilution, both at room temperature. Afterwards, the secondary anti-rabbit HRP Polymer was applied for 30 min, followed by 1:20 diluted 3,3'-Diaminobenzidin (DAB) for 10 min and 1:10 diluted hematoxylin for 3 min with washing steps with TBST in between before a final washing with tap water for 3 min before fixation. For multiplexed antigen detection, the OpalTM chemistry system (Akoya Biosciences, Marlborough, MA, USA, Cat. No.: OP7TL4001KT) was used according to the manufacture's

120 description. Briefly, after deparaffinization and fixation, we processed the sections for 15 min
121 with retrieval buffers in an inverter microwave oven. Then, we incubated them with antibody
122 diluent for 10 min at room temperature, followed by incubation with the anti-SPRR1B antibody
123 for 30 min. Next, Opal Polymer horseradish peroxidase (HRP) secondary antibody solution
124 with the respective chromogen was applied for 10 min, antibodies were removed by microwave
125 treatment and the staining with anti-ZEB1 antibody was performed. Slides were lastly
126 incubated with 4′,6-diamidino-2-phenylindole (DAPI) for 5 min.

127 **Single-cell RNA sequencing**

128 Samples were processed immediately after surgery and temporarily stored for transport at 4°C
129 in tissue storage solution (Miltenyi Biotec, Bergisch Gladbach, Germany) as previously
130 described. Briefly, samples were dissociated into single-cell suspensions using the
131 gentleMACS Dissociator (Cat. No. 130-093-235, Miltenyi Biotec, Bergisch Gladbach,
132 Germany) with program “h_tumor_01”, followed by 2x program “h_tumor_02” in 4.7 ml RPMI
133 1640 (Cat. No. P04-16500, PAN-Biotech) and an enzyme mix consisting of 200µl Enzyme H,
134 100 µl Enzyme R and 25 µl Enzyme A (Cat. No. 130-095-929 Miltenyi Biotec). Afterwards,
135 single-cell suspensions were reconstituted and washed thrice with 0.05 % BSA phosphate-
136 buffered saline (PBS) and filtered through a 100 µl cell strainer.

137 In cases multiple samples of one patient had to be analyzed (Table 1), antibody hashing for
138 multiplexing of samples was performed according to manufacturer’s protocol. Briefly, 1 µg of
139 the respective TotalSeq anti-human hashtag antibody was used to incubate a maximum of ca.
140 2 million cells for 30 min at 4 °C (Cat. No. 394601, 394603, 394605 and 394661, 394663,
141 394665, respectively, Biolegend, San Diego, CA, USA). After 3 washes with PBS with 0.05 %
142 BSA respective cell suspensions were mixed prior to single-cell RNA library preparation. In
143 short, both unhashed and hashed single-cell suspensions were barcoded and processed with
144 the microfluidic system of 10x Genomics Chromium v2.0 platform as described in the
145 manufacturer’s protocols (10x Genomics, Leiden, Netherlands). Due to a change of system,
146 both the 3’ technology including Chromium Single Cell 3’ Library & Gel Bead Kit version 2 (Cat.

147 No. 120237), Chromium Single Cell A Chip Kit (Cat. No. 120236), Chromium i7 Multiplex Kit
148 (Cat. No. 120262) as well as the 5' technology including Chromium Single Cell 5' Library & Gel
149 Bead Kits version 2 (Cat. No. 1000263), Chromium Next GEM Chip K Single Cell Kit (Cat. No.
150 1000286), Dual Index Kit TT set A (Cat. No. 1000215) were used; for library construction the
151 Chromium Single Cell 3'/5' Library Construction Kit (Cat. No. 1000020) was applied. After
152 library preparation, the library from patient one was sequenced with an Illumina HiSeq 4000
153 (Illumina, Berlin, Germany) at the DKFZ Genomics and Proteomics Core Facility in Heidelberg
154 and all other libraries were sequenced on an Illumina NovaSeq 6000 (Illumina, Berlin,
155 Germany) in three runs (Run 1: patient 2,4,5; Run2: patient 3, Run3: Patient 6 and 7) at the
156 West German Genome Center in Cologne.

157 **Bulk transcriptome analysis**

158 The bulk transcriptome was analyzed using a quantitative nuclease protection assay from the
159 HTG Transcriptome Panel (HTP) according to the manufacturers protocol (Cat. No. HTG-001-
160 224, HTG Molecular Diagnostics, Tucson, AZ, USA). Briefly, the tumor areas were macro-
161 dissected as depicted in Supplementary Figure 1 from 4 µm FFPE sections and subjected to
162 Proteinase K and DNase digestion. Next, the quantitative nuclease protection assay was
163 performed using the HTG EdgeSeq Processors before adapters and sample tags were added
164 in a PCR amplification. The resulting libraries were sequenced using on an Illumina NextSeq
165 500/550 High Output Kit v2.5 (75 cycles) (Cat. No. 20024906, Illumina, Berlin, Germany).

166 The resulting FASTQ files were processed towards a gene expression count matrix using the
167 HTG EdgeSeq Reveal Software version 4.0.1. Quality Control, normalization, and principal
168 component analysis (PCA) were performed using R version 4.0.5. Sample 5 failed QC due to
169 low number of sequenced reads and was removed from the analysis. Deconvolution was
170 performed with web application of CIBERSORTx (https://cibersortx.stanford.edu/) using a
171 signature matrix derived from the gene expression count matrix of combined scRNAseq data
172 of the samples analyzed with HTP (17). We filtered for genes that are expressed less than 5%
173 within the given tumor phenotypes and randomly included only 75% of T- and B-Cells for better

174 performance. The resulting signature matrix was used for imputing cellular fractions from the
175 counts-per-million normalized HTP data without any batch correction or quantile normalization
176 and 500 permutations.

177 **Bioinformatic analysis of scRNAseq data**

178 ***Preprocessing***

179 Processing from FASTQ files towards the unfiltered count matrix (barcodes x genes) was
180 performed using Cellranger Software Suite version 3.1.0 and the human reference genome
181 build GRCh38, downloaded from 10x Genomics in version 3.0.0.

182 Cells were identified by evaluating quality criteria inspired by Luecken *et al.*, also see Additional
183 file 1 (18). Cells were defined by having more than 500 UMIs, less than 10 % mitochondrial
184 gene expression and additionally for patient 3 and 5 having more than 30 housekeeper genes
185 expressed. The filtered count matrices (cells x genes) were further processed using Seurat
186 version 4.0.1 and R version 4.0.5 (19). Demultiplexing of hashed libraries was performed
187 choosing manual threshold of hashtag oligo (HTO) expression based on quality assessments
188 described in Additional file 1. We removed doublets that were identified by demultiplexing of
189 HTO expression matrices from all analyses.

190 ***Normalization and dimensionality reduction***

191 Every time we performed analysis of a specific set of cells, e.g., only tumor cells or only a
192 specific patient, the respective set of cells is normalized using the SCtransform algorithm and
193 the 3000 most variable genes were selected for PCA (20). During normalization, we regressed
194 for cell cycle scores and percentage of mitochondrial gene expression. Cell cycle scores and
195 phases were determined in Seurat using log-normalized RNA counts and S- and G2M-Phase
196 genes defined by Tirosh *et al.* (21). When generating a uniform manifold approximation and
197 projection (UMAP) without patient-specific batch effect we used corrected PCs derived with
198 the harmony R package version 0.1.0 [54]. Based on the variance explained by each PC and
199 the respective ranked elbow plot we choose an appropriate number of PCs for UMAP

visualization and SNN clustering as implemented in Seurat. For deriving patient-specific clusters for calculating the intratumoral cosine similarity we used the same resolution parameter for comparability. UMAPs colored by specific gene expression were ordered by expression values.

Tumor cell identification and phenotyping

We annotated all cell types and identified tumor phenotypes by using a combination of methods: SNN clustering, differential gene expression, gene set enrichment analysis (GSEA), expression of literature-based marker genes and automated reference-based annotation with SingleR using the Monaco bulk-RNA Immune dataset (22, 23). Automated, reference-based annotation with SingleR version 1.4.1 was run on SNN clusters with resolution 100 yielding extremely small cluster including only few cells but increasing performance. Further, we excluded cells with ambiguous cell type marker expression. For example, within the tumor cells of the lymph node metastasis from patient 1, we observed few cells expressing genes typical for fibroblasts and DCs that were subsequently excluded from tumor-specific analysis. Similarly, we removed T-Cells, B-Cells, mast cells, fibroblasts, muscle cells, melanocytes, and other Immune cells from the tumor cell subsets. Malignant cells were first identified by high epithelial gene expression as for example cytokeratins and inferred CNVs (Supplementary Figure 2). CNVs were inferred using the R package inferCNV version 1.6.0 with the not normalized, filtered count matrix including all cells as input and algorithm run in “samples” mode . Inferred CNVs of mitochondrial genes were excluded. Differential gene expression was performed by calculating the log2 foldchanges between one cluster and all other cells from the subset based on log-normalized data using NormalizeData function and scale factor of 10,000. We filtered for genes with log2 foldchange greater than 0.25 and a minimum percentual expression of at least 10 % within the cluster or all other cells. For calculating the cosine similarity, we did not filter the log2 foldchanges. GSEA was performed using the “fgsea” R package version 1.16.0, log2 foldchanges from differential gene expression and gene ontology biological processes (GO:BP,C5 v7.1) as well as hallmark gene sets (H, v7.1) downloaded

227 from MSigDB database (24-26). Gene sets were included if they have at least 15 genes or at
228 maximum 500 genes within the gene set using 10,000 permutations.

229 For deriving epithelial differentiating and pEMT gene signatures of patient 1 we calculate
230 foldchanges between the “epi” and “pEMT” cluster and included genes with log2 foldchanges
231 greater or lesser than 1, respectively, and with at least 10 % of either epi or pEMT cells
232 expressing that gene. Gene set variation analysis (GSVA) was performed using the R package
233 GSVA version 1.38.2 with default settings, i.e., gaussian kernel (27). As input, EMTome
234 signatures, the pEMT and epithelial differentiation 1 and 2 signature from Puram *et al*. (8) and
235 the three EMT and the epithelial senescence signatures from Kinker *et al.* were used (12).

236 Trajectories were inferred using SlingShot version 1.8.0 with malignant cell clusters as shown
237 in Figure 2A of patient 1 and the first 20 PCs (28). RNA velocity was inferred using the VeloCyto
238 python and R package (version 0.6) (29). Creation of the loom file was done using default
239 options and gene annotations as used for Cellranger processing. Genes were filtered based
240 on the minimum average expression magnitude with a threshold of 0.05 for spliced and 0.02
241 for unspliced reads. Velocity estimates were calculated using as distance the inverse
242 correlation coefficient of the PC embedding correlation matrix, the top/bottom 2 % quantiles for
243 gamma fit, 50 neighboring cells projecting 1 deltaT into the future and projected on the UMAP
244 using 200 neighbors and 50 grid points.

245 Transcription factor activity was inferred using the VIPER algorithm (version 1.24.0) and
246 regulons from the DoRothEA database (version 1.2.2) (30-32). Hierarchical clustering in the
247 heatmaps was performed using Euclidean distance and ward.D2 method unless otherwise
248 noted. Visualization was performed using ggplot2 version 3.3.3 and ComplexHeatmap version
249 2.6.2 (33, 34).

250 **Reanalysis of HNSCC dataset from Kürten *et al*.**

251 From the publicly available scRNAseq data set on primary HNSCC tumors we downloaded the
252 FASTQ files of CD45-negative and HPV-negative libraries from the sequencing read archive
253 (SRA) under accession ID SRP301444 (15). The data sets were analyzed the same as

254 described above. HPV-negative samples were chosen for comparability to the OSCC dataset.
255 However, the HN07 tumor originated from the larynx, while all other samples indeed originated
256 from the oral cavity. We adjusted the cell identification thresholds based on our evaluation
257 criteria pooled for all libraries: cells were defined by having more than 175 genes expressed,
258 less than 10 % mitochondrial gene expression and more than 60 housekeeper genes
259 expressed (see Additional file 1).

Results

Single-cell gene expression signatures of tumor cells from a single metastasis show several predominant, but not necessarily exclusive, functional phenotypes.

To avoid inter-patient heterogeneity as a confounding factor, we first focused on the analysis of a single OSCC metastasis to develop hypotheses which would be subsequently tested in the entire cohort. For this, we chose a metachronous lymph node metastasis that was removed one year after the primary tumor, because we assumed that consolidation processes are particularly pronounced in this longer existing metastasis. Multiplexed scRNAseq recovered 4,121 cells that could be assigned to the following cell types: 1,906 (46.8 %) tumor cells, 1,186 (29.1 %) fibroblasts, 507 (12.4 %) dendritic cells (DCs), 375 (9.2 %) macrophages and 102 (2.5 %) endothelial cells (ECs) (Figure 1A). The absence of T- or B-cells was in line with the histology showing completely disrupted lymph node structures and only occasional tumor-infiltrating lymphocytes (Supplementary Figure 1). OSCC cells were identified both by the presence of copy number variations (CNVs) inferred from scRNAseq data as well as the expression of epithelial markers including *S100A2, S100A14*, cytokeratins (*KRT5*, *KRT14*, *KRT17)* and stratifin (*SFN*) (Supplementary Figure 2).

Detailed phenotyping of the cancer cells identified several clusters to which we could assign the following predominant functional phenotypes that differ in their EMT state, immunomodulatory capacity, as well as their response to hypoxia, stress, and metabolic constraints (Figure 1B-D). However, the predominance of a functional phenotype does not preclude additional traits. Specifically, 515 cells exhibited a pEMT phenotype by expressing a mixture of epithelial and mesenchymal genes such as matrix metallopeptidases (*MMPs*), caveolin-1 (*CAV1*) and *CXCL14* that were previously described in pEMT signatures and are enriched in the EMT hallmark gene set from the molecular signatures database (MSigDB) (Supplementary Figure 3A) (8, 25). In contrast, 385 cells showed higher expression of genes associated with epithelial differentiation such as *S100A7/A8/A9,* the keratinocyte envelope protein cornifin-B (*SPRR1B*), and various cytokeratins (e.g., *KRT6B* and *KRT16*). The EMP-

287 related gene expression patterns are correlating with established signatures from the EMTome
288 database (Supplementary Figure 3B-E) (7). Interestingly, both gene expression signatures
289 were present in a mixed cluster with 184 cells characterized by the high expression of cell-
290 cycle related genes (despite the fact that we applied cell cycle regression).

291 With respect to cell clusters whose gene expression was not predominately associated with
292 EMP, 268 cells exhibited an immune-regulatory phenotype enriched for genes associated with
293 cytokine-mediated responses and higher expression of the chemokines *CXCL1/2/3/8* and
294 *CCL20* and for 554 cells the gene expression pattern suggests adaptation to environmental
295 factors. In detail, 51 cells had higher expression of transcription factors *FOS* and *JUN*
296 suggesting a stress response. 310 cells can be assumed to respond to hypoxic conditions in
297 the tumor based on the higher expression of *NDRG1* and *EGLN3* both of which are regulated
298 by oxygen levels (35, 36): *NDRG1* regulates stress response and p53-mediate caspase
299 activation (35), and *EGLN3* has an important role in hypoxia-inducible factor 1 alpha (HIF1α)
300 regulation through prolyl hydroxylation (36). 193 cells expressed genes associated with amino
301 acid metabolism, starvation response and mTORC1 signaling, i.e., a regulator of mitochondrial
302 metabolism (37). The upregulated genes *ASNS*, *PSAT1* and *PHGDH* integrate the metabolites
303 of serine and glycine metabolism into glycolysis and therefore fuel glycolysis with amino acids
304 (38); hence, these cells appear to be adapted to low-glucose conditions.

305 **OSCC cells in lymph node metastases undergo mesenchymal-epithelial transition**

306 We next focused on possible dynamics within the predominantly EMP-related cancer cell
307 clusters using a higher resolved shared-nearest neighbor (SNN) clustering. We defined 4
308 pEMT clusters (pEMT-1 to 4), 4 clusters of more epithelial differentiated cells (epi-1 to 4) and
309 one cluster with mixed phenotypes (mix) (Figure 2A). pEMT-1 is enriched for genes involved
310 in coagulation and play a role in angiogenesis as *THBS1*, *CYR61* and *F3* (Figure 2B,
311 Supplementary Figure 4A). pEMT-2 and 3 showed higher expression of extracellular matrix
312 (ECM) remodeling genes, but pEMT-2 also higher expressed cytokeratin *KRT15* and
313 chemokine *CXCL14* while pEMT-3 higher expressed the serine protease inhibitor *SERPINA1*

314 and podoplanin (*PDPN*) which mediates efficient ECM degradation by controlling invadopodia
315 (39). Of the more epithelial differentiated cell clusters, epi-2's expression profile is closest to
316 the pEMT cluster: higher expression of *MMP1* and lower expression of *SPRR1B* and
317 *S100A8/A9* than epi-1, 3 and 4 (Figure 2B, E). Epi-3 showed higher expression of *S100A7* and
318 *KRTDAP*, whereas epi-4 has higher expression of kallikreins (*KLK6/7*), prostate stem cell
319 antigen (*PSCA*) and adipogenesis regulatory factor (*ADIRF*). Both *ADIRF* and *PSCA* play a
320 role in prostate cancer and *PSCA* is also reported as highly expressed in mucosal tissue, but
321 less in HNSCC (40-42).

322 To gain a better understanding on the gene expression dynamics we estimated RNA velocity,
323 which predicts the short-term future development in gene expression of individual cells using
324 the ratio of spliced and non-spliced mRNA counts (Figure 2C). This analysis revealed that
325 epithelial differentiated cells were strongly developing towards cluster epi-4, while most other
326 cells did not show a strong uniform developmental direction. Tracking the developmental
327 pathway within the metastasis by trajectory analysis across all EMP-related clusters revealed
328 a major developmental axis between pEMT and epithelial differentiated cells that is diversifying
329 within each end (Figure 2D, E, Supplementary Figure 4B). To confirm that we the found
330 progressive epithelial differentiation of the metastatic cells represents a MET, we inferred
331 CNVs from the scRNAseq data. This demonstrated an increased number of copy number gains
332 on chromosome 1, 8, 17, and 19 within epithelial differentiated cell clusters epi-1, epi-3, and
333 epi-4 (Figure 2F, Supplementary Figure 4C). It should be noted, however, that CNVs on
334 chromosomes 1 and 17 are associated with upregulated epithelial genes in close genomic
335 proximity; thus, these two copy number gains may not represent true genomic CNVs, but rather
336 reflect the high expression of these genes in epithelial differentiated cells (Supplementary
337 Figure 4D, E).

338 **Intra-tumoral heterogeneity of OSCC is driven by EMP**

339 To test whether our observation that OSCC cells in one lymph node metastasis undergo a
340 mesenchymal-epithelial transition is generally valid we extended our analyses by adding 5

341 primary tumors and 8 matched lymph node metastases from 6 patients (Table 1). From the
342 publicly available scRNAseq data set on primary HNSCC tumors published by Kürten *et al.*
343 (15) we chose to include the 9 HPV-negative primary tumors, which were all but one originating
344 from the oral cavity (HN07 originated from the larynx) in our analysis. In total, we analyzed
345 7,263 cancer cells from 16 different patients (Figure 3A). Importantly, the frequency of cancer
346 cells was unevenly distributed across samples, which could not be explained by differences in
347 tumor cell content across samples as determined by histopathology (Supplementary Figure 1).
348 Since the recovery of cells by scRNAseq is also low, the stability of epithelial tumor cell
349 assemblies, which may not be sufficiently broken up by dissociation protocols, likely interfered
350 with the generation of OSCC single-cell suspensions. In addition, as expected from the inter-
351 patient heterogeneity, cancer cells were clustered based on their gene expression by patient
352 rather than functional phenotype (Figure 3A, B). Thus, we accounted for the patient-specific
353 effects with batch-corrected principal components (PCs) using the harmony R package which
354 indeed resulted in a clustering by functional phenotypes (43) (Figure 3C, Supplementary
355 Figure 5A-C). For initial annotation of the phenotype of each cluster, we were guided by the
356 gene signatures previously identified in the indicator sample, with additional predominantly
357 immunoregulatory phenotypes found. EMP-related phenotypes were present in all, but one
358 tumor sample with only one cell (Figure 3D). Nevertheless, direct comparison of the
359 frequencies of the predominant phenotypes in the pooled scRNAseq dataset proved difficult
360 to interpret due to differences in cell numbers, and patient-specific gene expression patterns.
361 To confirm the validity of our scRNAseq data, we examined the bulk transcriptome and
362 deconvoluted the respective cell types for our 18 OSCC samples (Supplementary Figure 6).
363 Using this approach, we observed a higher tumor and stroma content and less lymphocytes,
364 which was expected as scRNAseq is known to be biased towards leukocytes. Therefore, we
365 chose an alternative approach: to compare the EMP-related intrapatient heterogeneity of all
366 analyzed tumor cells between patients, we calculated the similarities of the differential gene
367 expression within each patient of all respective clusters of all patients. We first considered each
368 patient individually and performed clustering, annotation, inferCNV and differential expression

369 analysis. As exemplified for the primary OSCC of patient HN01, tumor cell clusters had the
370 same inferred CNVs and were annotated based on their phenotype using again the indicator
371 sample as a guide (Figure 3E, Supplementary Figure 5D, E). The cosine similarity between
372 patient-specific clusters demonstrated that within each patient, the heterogeneity in EMP is
373 most prominent and epithelial differentiated phenotypes are profoundly different from most
374 other, especially pEMT clusters (Figure 3F, Supplementary Figure 5F). Indeed, all pEMT
375 phenotypes are very similar to each other and show a large overlap of the gene expression
376 patterns with predominantly immune- and metabolic-related clusters. In epithelial differentiated
377 cells, the most upregulated genes are *S100A8* and *S100A9*, encoding calprotectin, and
378 *SPRR1B*, all members of the epidermal differentiation complex (44). Of note, hypoxia- and
379 stress-related heterogeneity is similar between patients suggesting a reactive response rather
380 than an aspect of tumor evolution.

381 **EMT-related transcription factor ZEB1 is highly active in metastatic epithelial**
382 **differentiated OSCC cells**

383 The transcription factors ZEB1/2, TWIST1/2, Snail (SNAI1) and Slug (SNAI2) are key for
384 regulation of EMP [7, 27]. While mRNA expression of *SNAI2* within single OSCC cells was
385 recently reported, the others were not detected (8). Here, we confirm this observation as we
386 detected *SNAI2* mRNA in almost half of the OSCC cells (Figure 4A). However, detection of
387 low expressed genes such as transcription factors by scRNAseq, especially in 10X genomics
388 technology, becomes unreliable due to dropout effects. Also, the activity of transcription factors
389 is often not reflected by the dynamics of their mRNA expression alone, as their activity depends
390 additionally on protein stability and posttranslational modifications; for example, ZEB1 protein
391 is more stable than Snail (45). To circumvent this problem, we inferred the activity of these
392 transcription factors based on the mRNA expression profile of their target genes using the
393 algorithm VIPER with regulons defined by DoRothEA database (30-32). Using this approach,
394 we were able to detect high activities of ZEB1, ZEB2, Snail and Slug in OSCC cells of different
395 patients with varying EMP phenotype (Figure 4B). Since, on the one hand, the observation that
396 epithelial differentiation in OSCC metastases is associated with higher activity of the EMT

397 activator ZEB1 was unexpected and, on the other hand, the method used to derive
398 transcription factor activities appears to maybe overestimate the activity for transcriptional
399 repressors, we addressed the plausibility of this observation (Figure 4C, Supplementary Figure
400 7). Consistent with the fact that one of the main functions of ZEB1 is the downregulation of E-
401 cadherin (46) (*CDH1*), OSCC cells generally showed low expression of *CDH1* (Figure 4B).
402 Next, we investigated the expression of ZEB1 protein by immunohistochemistry (IHC) in all 14
403 tumor lesions of our cohort. Since we observed nuclear ZEB1 expression in similar tumor areas
404 as high cytoplasmic cornifin-B expression, which served as a marker of epithelial differentiation
405 (Figure 4D, E), we validated the co-expression of ZEB-1 and cornifin-B in the same cell by
406 immunofluorescence double-staining (Figure 4F). In line with our scRNAseq data,
407 colocalization of both proteins was observed in a fraction of cancer cells in 9/14 (64 %)
408 samples, double positive cells were more frequently observed in lymph node metastases (7/9,
409 78 %; Table 1).

410 # Discussion

411 EMT represents the reactivation of an embryonic developmental program in which cells
412 acquire migratory and invasive properties, i.e., prerequisites for invasion and metastasis of
413 cancer (3, 47, 48). Thus, in early stages of metastasis, tumor cells undergo EMT, whereas in
414 established metastases the reverse process *aka* MET is also observed (49, 50). To assess the
415 EMP-associated heterogeneity among OSCC cells and gain some insight into the dynamics of
416 this process, we examined the transcriptomes of 7,263 individual carcinoma cells isolated from
417 primary or metastatic OSCC. Our results not only confirm the EMP-associated heterogeneity
418 of cancer cells in primary and metastatic OSCC, but also demonstrate progressive MET in
419 established lymph node metastases. Interestingly, the epithelial differentiation in OSCC
420 metastases is associated with higher activity of EMT-activator ZEB1, which was confirmed on
421 protein level by co-expression of ZEB1 and cornifin-B in individual tumor cells by
422 immunofluorescence staining.

423 EMP appears to be the main driver of cellular heterogeneity within OSCC: detailed phenotyping
424 of cancer cells identified several clusters whose predominant functional phenotypes
425 corresponded to different EMP states, ranging from a pEMT to an epithelial differentiated state.
426 However, particularly the pEMT phenotypes could be overlaid by features related to
427 angiogenesis, ECM remodeling, metabolic adaptations, stress, and interaction with the
428 immune system. Metabolic adaptations include response to environmental limitations as
429 hypoxia and low glucose. Low glucose conditions are counteracted with upregulation of genes
430 related to amino acid metabolism that fuel into glycolysis (38). While previous studies
431 suggested that the activity of specific metabolic pathways in OSCC varies widely among
432 patients (51), we observed that hypoxia- and stress-related gene expression patterns are
433 rather similar between patients supporting the notion of an reactive response rather than an
434 aspect of individual tumor evolution.

435 In terms of EMP dynamics, it is assumed that cells may transit from one EMP state to another
436 along a continuous spectrum of changes. Currently, however, it is also discussed if long-lived
437 phenotypes representing discrete EMP states prevail (3, 5, 6, 8, 9, 11-13). Most studies
438 supporting continuous transitions are based on *in vitro* or preclinical *in vivo* models that may
439 not fully reflect the complexity of the tumor and its microenvironments (5, 6, 9). Indeed, human
440 *in situ* or *ex vivo* studies suggested distinct EMP states; however, these approaches do not
441 fully capture cellular dynamics (8, 12, 48). Comparing the intratumoral heterogeneity we
442 demonstrated that within each patient, the EMP-driven heterogeneity is most prominent and
443 epithelial differentiated phenotypes are profoundly different from pEMT clusters suggesting
444 that these states may be rather static. Moreover, gene expression dynamics estimated by RNA
445 velocity demonstrated epithelial differentiated cells were strongly developing towards a more
446 pronounced epithelial differentiation with an increasing expression of genes of the epidermal
447 differentiation complex (44). Of note, OSCC cells with a pEMT phenotype did not show such a
448 uniform developmental direction. The assumption that epithelial differentiated metastatic cells
449 developed later than pEMT cells, i.e., undergoing MET, is supported by an increasing number

450 of inferred copy number gains towards increasing epithelial differentiation even if accounting
451 for the limitations of this approach.

452 However, we cannot conclude whether MET happened within the metastasis or primary
453 malignancy, as our data reflects the tumor heterogeneity within a specific timepoint of tumor
454 evolution. Multiple disseminated tumor cells reflecting the observed EMP heterogeneity might
455 have migrated collectively from the primary tumor that could be crucial for metastatic
456 consolidation, as we observed a similar EMP heterogeneity also within primary tumors.

457 The high transcriptional activity of EMT-activator ZEB1 in epithelial differentiated OSCC cells
458 in both primary and metastatic tumor lesions was unexpected, even considering that the
459 scRNAseq data inferred transcription factor activities are biased towards transcriptional
460 repressors. In the case of primary tumors this may be interpreted as the incipient EMT, but this
461 would not work for the metastatic lesions in which ZEB1 activity was associated with a
462 progressive epithelial differentiation. Indeed, in previous studies depletion of ZEB1 has driven
463 tumor cells from pEMT towards an epithelial phenotype (52, 53). However, depletion of Zeb1
464 also reduces phenotypic variability of cancer cells particular their phenotypic/metabolic
465 plasticity (52). While it is well established that ZEB1 together with microRNAs stabilizes EMT
466 through a feedforward loop, this loop could also induce epithelial differentiation based on
467 environmental factors (54). In addition to the transcriptional repressor activity, ZEB1 has been
468 demonstrated to induce the epithelial differentiation marker cornifin-B in response to IL-1β and
469 IFN-γ (55). We not only demonstrate the coincidence of *SPRR1B* mRNA expression and ZEB1
470 activity, but also the co-localization of cornifin-B and ZEB1 protein expression in OSCC lymph
471 node metastases. Thus, although ZEB1 activity is crucial for the induction of the pEMT state,
472 it does not seem to completely prevent partial epithelial differentiation. Remarkably, no relevant
473 differences in *CDH1* expression were detected between the different EMP states in the
474 metastatic OSCC lesions. Therefore, the epithelial differentiated phenotypes we observed
475 most closely correspond to a partial epithelial differentiation analogous with the observed
476 pEMT phenotypes. We speculate that the driving force behind this EMP-associated

477 heterogeneity of OSCC cells is to maintain cellular integrity. For example, ZEB1 is an ATM-
478 substrate linking ATM and CHK1, promoting homologous recombination-dependent DNA
479 repair and thereby protecting cells from genotoxic stress whereas expression of keratin
480 intermediate filaments helps to protect cells from stress associated apoptosis (56, 57).

481 # Conclusions

482 Single cell transcriptomics reveals that heterogeneity within OSCC cells is dominated by EMP
483 differences resulting in distinct partial EMT and epithelial differentiated phenotypes.
484 Particularly the partial EMT phenotypes can be accompanied by features related to metabolic
485 adaptations, stress, and interaction with the immune system. These EMP phenotypes likely
486 endow capabilities that are essential for the different stages of the metastatic process,
487 including maintenance and cellular integrity. This could be a possible additional function of
488 ZEB1, as it is also expressed during progressive epithelial differentiation in OSCC metastases.

489 **List of abbreviations**

490 CNV	Copy number variation
491 DC	Dendritic cell
492 DEG	Differentially expressed genes
493 EC	Endothelial cell
494 ECM	Extracellular matrix
495 EMP	Epithelial-mesenchymal plasticity
496 EMT	Epithelial-mesenchymal transition
497 FFPE	Formalin-fixed, paraffin-embedded
498 GSEA	Gene set enrichment analysis
499 GSVA	Gene set variation analysis
500 HNSCC	Head and neck squamous cell carcinomas
501 HPV	Human papillomavirus
502 HTO	Hashtag oligo
503 IHC	Immunohistochemistry
504 MET	Mesenchymal-epithelial transition
505 MSigDB	Molecular signature database
506 OSCC	Oral cavity squamous cell carcinoma
507 PC	Principal component
508 pEMT	Partial EMT
509 scRNAseq	Single-cell RNA sequencing
510 SNN	Shared-nearest neighbor
511 UMAP	Uniform manifold approximation and projection

512

Supplementary Information

Supplementary Figure 1 [.pdf]: Histology of OSCC primary and metastatic tumors.

Supplementary Figure 2 [.pdf]: Cell type identification by differential expression and inferred CNVs for a metachronous lymph node metastasis.

Supplementary Figure 3 [.pdf]: PEMT and epithelial differentiating gene expression signatures are comparable to previously published EMT signatures.

Supplementary Figure 4 [.pdf]: Extended analysis of the tumor phenotype characterization for the lymph node metastasis of patient 1.

Supplementary Figure 5 [.pdf]: Malignant phenotypes characterized across all analyzed patients.

Supplementary Figure 6 [.pdf]: Bulk transcriptomes reveal the cellular composition of OSCC.

Supplementary Figure 7 [.pdf]: Inferred transcription factor activity might be biased by activator or repressor function.

Table 1 [.xlsx]: Clinical and sequencing information from OSCC patients.

Additional file 1 [.pdf]: Supplementary Materials and Methods.

Declarations

Data availability

The processed datasets generated and analyzed during the current study are available in the GEO database with accession id GSE195655. The raw FASTQ files are available upon request due to privacy reasons. The high-resolution H&E Images are available through figshare under the DOI: https://doi.org/10.6084/m9.figshare.20905837.v1. The publicly available FASTQ files from Kürten *et al.* were downloaded from the sequencing read archive (SRA) accession id SRP301444.

536 ## Code availability

537 Code used for analysis is available at https://github.com/sci-kai/single_cell_EMP.

538 ## Acknowledgements

539 We thank Önder Bozdogan, Translational Skin Cancer Research, German Cancer Consortium
540 (DKTK), Essen, Germany, for his histological expertise and Linda Kubat, Department of
541 Dermatology, University Medicine Essen, Germany, for helping with microscopy. We thank
542 Christina H. Scheel, Skin Cancer Center, Department of Dermatology, Ruhr-University
543 Bochum, Germany, for her helpful remarks to the manuscript. We further thank Yvonne Krause
544 and Sophia Berger, both Department of Pathology, University Medicine Essen, Germany, for
545 their help and guidance with HTG EdgeSeq. We further thank the DKFZ Genomics and
546 Proteomics Core Facility and the West German Genome Center in Cologne for providing
547 Illumina sequencing and related services.

548 ## Funding

549 The project was funded by the German Cancer Consortium (DKTK) site budget OE 0460
550 ED003 and the federal ministry of education and research (03VP01062).

551 ## Author Contributions

552 All authors approved the final version and agreed to be accountable for all aspects of the work.
553 KH was responsible for data analysis, data curation, visualization, interpretation,
554 conceptualization, drafting the manuscript and updating revisions. IS, LP and PG contributed
555 by performing wet lab experiments including single-cell dissociation, scRNAseq and library
556 preparation. FF established and performed histological, IHC and bulk transcriptome analysis.
557 CS contributed with sample acquisition, interpretation and revising the manuscript. NS
558 contributed by manuscript reviewing and interpretation. JG further contributed to data analysis,
559 interpretation, and visualization. JCB was contributing by interpretation, analysis guidance,
560 immunofluorescence examination and manuscript editing. Further, he was responsible for
561 conceptualization and project administration including supervision and funding acquisition.

562 ## Ethics approval and consent to participate

563 Written informed consent was obtained from each patient and the Ethics Committee of the
564 Medical Faculty of the Heinrich-Heine-University Düsseldorf (#3090) approved the study. All
565 procedures involving human participants are in accordance with the ethical standards of the
566 institutional and research committee and with the Helsinki Declaration.

567 ## Consent for publication

568 Not applicable.

569 ## Competing Interests (COI)

570 JCB is receiving speaker's bureau honoraria from Amgen, Pfizer, MerckSerono, Recordati and
571 Sanofi, is a paid consultant/advisory board member/DSMB member for Boehringer Ingelheim,
572 InProTher, MerckSerono, Pfizer, 4SC, and Sanofi/Regeneron. His group receives research
573 grants from Bristol-Myers Squibb, Merck Serono, HTG, IQVIA, and Alcedis. None of these
574 activities are related to the present manuscript. The other authors including K.H., C.S., L.P.,
575 F.F., P.G., J.G., N.S., I.S. declare no conflict of interest.

576

577 ## References

578 1. Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and
579 neck squamous cell carcinoma. Nat Rev Dis Primers. 2020;6(1):92.
580 2. Cramer JD, Burtness B, Le QT, Ferris RL. The changing therapeutic landscape of head
581 and neck cancer. Nat Rev Clin Oncol. 2019;16(11):669-83.
582 3. Nieto MA, Huang Ruby Y-J, Jackson Rebecca A, Thiery Jean P. EMT: 2016. Cell.
583 2016;166(1):21-45.
584 4. Yang J, Antin P, Berx G, Blanpain C, Brabletz T, Bronner M, et al. Guidelines and
585 definitions for research on epithelial–mesenchymal transition. Nat Rev Mol Cell.
586 2020;21(6):341-52.
587 5. McFaline-Figueroa JL, Hill AJ, Qiu X, Jackson D, Shendure J, Trapnell C. A pooled
588 single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-
589 to-mesenchymal transition. Nat Genet. 2019;51(9):1389-98.
590 6. Cook DP, Vanderhyden BC. Context specificity of the EMT transcriptional response.
591 Nat Commun. 2020;11(1):2142.
592 7. Vasaikar SV, Deshmukh AP, den Hollander P, Addanki S, Kuburich NA, Kudaravalli S,
593 et al. EMTome: a resource for pan-cancer analysis of epithelial-mesenchymal transition genes
594 and signatures. Br J Cancer. 2021;124(1):259-69.
595 8. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell
596 Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck
597 Cancer. Cell. 2017;171(7):1611-24.e24.
598 9. Pastushenko I, Brisebarre A, Sifrim A, Fioramonti M, Revenco T, Boumahdi S, et al.
599 Identification of the tumour transition states occurring during EMT. Nature.
600 2018;556(7702):463-8.
601 10. Pal A, Barrett TF, Paolini R, Parikh A, Puram SV. Partial EMT in head and neck cancer
602 biology: a spectrum instead of a switch. Oncogene. 2021;40(32):5049-65.
603 11. Jolly MK, Tripathi SC, Jia D, Mooney SM, Celiktas M, Hanash SM, et al. Stability of the
604 hybrid epithelial/mesenchymal phenotype. Oncotarget. 2016;7(19).
605 12. Kinker GS, Greenwald AC, Tal R, Orlova Z, Cuoco MS, McFarland JM, et al. Pan-
606 cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. Nat Genet.
607 2020;52(11):1208-18.
608 13. Karacosta LG, Anchang B, Ignatiadis N, Kimmey SC, Benson JA, Shrager JB, et al.
609 Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell
610 resolution. Nat Commun. 2019;10(1):5587.
611 14. Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, et al. Multimodal
612 Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell.
613 2020;182(2):497-514.e22.

614 15. Kürten CHL, Kulkarni A, Cillo AR, Santos PM, Roble AK, Onkar S, et al. Investigating
615 immune and non-immune cell interactions in head and neck tumors by single-cell RNA
616 sequencing. Nat Commun. 2021;12(1):7338.
617 16. Spassova I, Ugurel S, Terheyden P, Sucker A, Hassel JC, Ritter C, et al. Predominance
618 of Central Memory T Cells with High T-Cell Receptor Repertoire Diversity is Associated with
619 Response to PD-1/PD-L1 Inhibition in Merkel Cell Carcinoma. Clinical Cancer Research.
620 2020;26(9):2257-67.
621 17. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al.
622 Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat
623 Biotechnol. 2019;37(7):773-82.
624 18. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a
625 tutorial. Mol Syst Biol. 2019;15(6):e8746.
626 19. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated
627 analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87.e29.
628 20. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-
629 seq data using regularized negative binomial regression. Genome Biol. 2019;20(1):296.
630 21. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al.
631 Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.
632 Science. 2016;352(6282):189.
633 22. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of
634 lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol.
635 2019;20(2):163-72.
636 23. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq Signatures
637 Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell
638 Types. Cell Rep. 2019;26(6):1627-40.e7.
639 24. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene
640 set enrichment analysis. bioRxiv. 2021:060012.
641 25. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The
642 Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst.
643 2015;1(6):417-25.
644 26. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al.
645 Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide
646 expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545.
647 27. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray
648 and RNA-Seq data. BMC Bioinform. 2013;14(1):7.
649 28. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage
650 and pseudotime inference for single-cell transcriptomics. BMC Genom. 2018;19(1):477.

651 29. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA
652 velocity of single cells. Nature. 2018;560(7719):494-8.
653 30. Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. Benchmark and
654 integration of resources for the estimation of human transcription factor activities. Genome
655 Res. 2019;29(8):1363-75.
656 31. Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, et al.
657 Robustness and applicability of transcription factor and pathway analysis tools on single-cell
658 RNA-seq data. Genome Biol. 2020;21(1):36.
659 32. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, et al. Functional
660 characterization of somatic mutations in cancer using network-based inference of protein
661 activity. Nat Genet. 2016;48(8):838-47.
662 33. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
663 2016.
664 34. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in
665 multidimensional genomic data. Bioinformatics. 2016;32(18):2847-9.
666 35. Guo D-D, Xie K-F, Luo X-J. Hypoxia-induced elevated NDRG1 mediates apoptosis
667 through reprograming mitochondrial fission in HCC. Gene. 2020;741:144552.
668 36. Ivan M, Kaelin WG. The EGLN-HIF O2-Sensing System: Multiple Inputs and
669 Feedbacks. Mol Cell. 2017;66(6):772-9.
670 37. Laplante M, Sabatini DM. mTOR signaling at a glance. J Cell Sci. 2009;122(20):3589-
671 94.
672 38. DeBerardinis Ralph J. Serine Metabolism: Some Tumors Take the Road Less
673 Traveled. Cell Metab. 2011;14(3):285-6.
674 39. Martín-Villar E, Borda-d'Agua B, Carrasco-Ramirez P, Renart J, Parsons M, Quintanilla
675 M, et al. Podoplanin mediates ECM degradation by squamous carcinoma cells through control
676 of invadopodia stability. Oncogene. 2015;34(34):4531-44.
677 40. de Nooij-van Dalen AG, van Dongen GAMS, Smeets SJ, Nieuwenhuis EJC, Stigter-
678 van Walsum M, Snow GB, et al. Characterization of the human Ly-6 antigens, the newly
679 annotated member Ly-6K included, as molecular markers for head-and-neck squamous cell
680 carcinoma. Int J Cancer. 2003;103(6):768-74.
681 41. Reiter RE, Gu Z, Watabe T, Thomas G, Szigeti K, Davis E, et al. Prostate stem cell
682 antigen: A cell surface marker overexpressed in prostate cancer. Proc Natl Acad Sci USA.
683 1998;95(4):1735-40.
684 42. Zhou K, Arslanturk S, Craig DB, Heath E, Draghici S. Discovery of primary prostate
685 cancer biomarkers using cross cancer learning. Sci Rep. 2021;11(1):10433.
686 43. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and
687 accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289-96.

688 44. Oh IY, de Guzman Strong C. The Molecular Revolution in Cutaneous Biology: EDC
689 and Locus Control. J Invest Dermatol. 2017;137(5):e101-e4.
690 45. Stemmler MP, Eccles RL, Brabletz S, Brabletz T. Non-redundant functions of EMT
691 transcription factors. Nat Cell Biol. 2019;21(1):102-12.
692 46. Sánchez-Tilló E, Lázaro A, Torrent R, Cuatrecasas M, Vaquero EC, Castells A, et al.
693 ZEB1 represses E-cadherin and induces an EMT by recruiting the SWI/SNF chromatin-
694 remodeling protein BRG1. Oncogene. 2010;29(24):3490-500.
695 47. Williams ED, Gao D, Redfern A, Thompson EW. Controversies around epithelial–
696 mesenchymal plasticity in cancer metastasis. Nat Rev Cancer. 2019;19(12):716-32.
697 48. Derynck R, Weinberg RA. EMT and Cancer: More Than Meets the Eye. Dev Cell.
698 2019;49(3):313-6.
699 49. Dongre A, Weinberg RA. New insights into the mechanisms of epithelial–mesenchymal
700 transition and implications for cancer. Nat Rev Mol Cell. 2019;20(2):69-84.
701 50. Tsai JH, Donaher JL, Murphy DA, Chau S, Yang J. Spatiotemporal regulation of
702 epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. Cancer
703 Cell. 2012;22(6):725-36.
704 51. Xiao Z, Dai Z, Locasale JW. Metabolic landscape of the tumor microenvironment at
705 single cell resolution. Nat Commun. 2019;10(1):3763.
706 52. Krebs AM, Mitschke J, Lasierra Losada M, Schmalhofer O, Boerries M, Busch H, et al.
707 The EMT-activator Zeb1 is a key factor for cell plasticity and promotes metastasis in pancreatic
708 cancer. Nat Cell Biol. 2017;19(5):518-29.
709 53. Carstens JL, Yang S, Correa de Sampaio P, Zheng X, Barua S, McAndrews KM, et al.
710 Stabilized epithelial phenotype of cancer cells in primary tumors leads to increased
711 colonization of liver metastasis in pancreatic cancer. Cell Rep. 2021;35(2):108990.
712 54. Burk U, Schubert J, Wellner U, Schmalhofer O, Vincan E, Spaderna S, et al. A
713 reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and
714 invasion in cancer cells. EMBO reports. 2008;9(6):582-9.
715 55. Li S, Gallup M, Chen YT, McNamara NA. Molecular mechanism of proinflammatory
716 cytokine-mediated squamous metaplasia in human corneal epithelial cells. Invest Ophthalmol
717 Vis Sci. 2010;51(5):2466-75.
718 56. Zhang P, Wei Y, Wang L, Debeb BG, Yuan Y, Zhang J, et al. ATM-mediated
719 stabilization of ZEB1 promotes DNA damage response and radioresistance through CHK1.
720 Nat Cell Biol. 2014;16(9):864-75.
721 57. Oshima RG. Apoptosis and keratin intermediate filaments. Cell Death Differ.
722 2002;9(5):486-92.

723

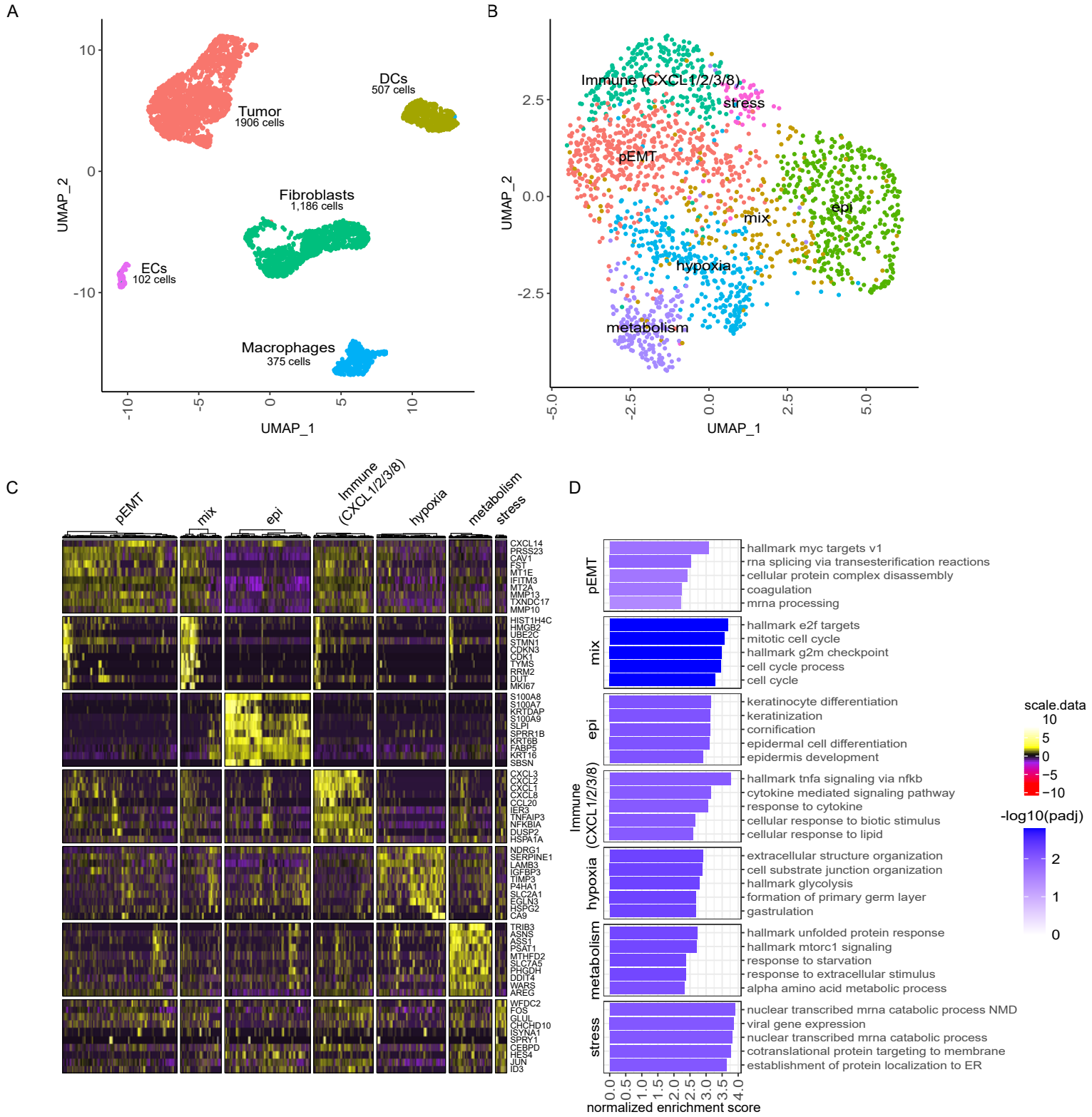

Figure 1 | Single-cell gene expression signatures in OSCC cells from a single metastasis reveals predominant functional phenotypes. (A) UMAP based on scRNAseq data of 4,076 cells isolated from a metachronous lymph node metastasis. Cells are annotated and summarized according to the presumed cell type. **(B)** UMAP of 1,906 OSCC cells depicted in A. Cells are annotated according to predominant functional phenotype. **(C)** Heatmap for scaled, log-normalized gene expression of tumor cells (columns) split by respective phenotype and the top 10 differentially expressed genes (DEGs) (rows) of the respective phenotype against all other tumor cells. DEGs are sorted from highest to lowest log2 foldchange. Row sections are ordered like column sections. **(D)** Top 5 enriched gene sets from log2 foldchanges of respective tumor phenotypes by normalized enrichment scores (x axis). Gene sets of respective phenotypes are sorted from highest to lowest enrichment. Bars are colored by the negative decadic logarithm of the Benjamini-Hochberg adjusted p-value (padj). DCs: dendritic cells. ECs: endothelial cells.

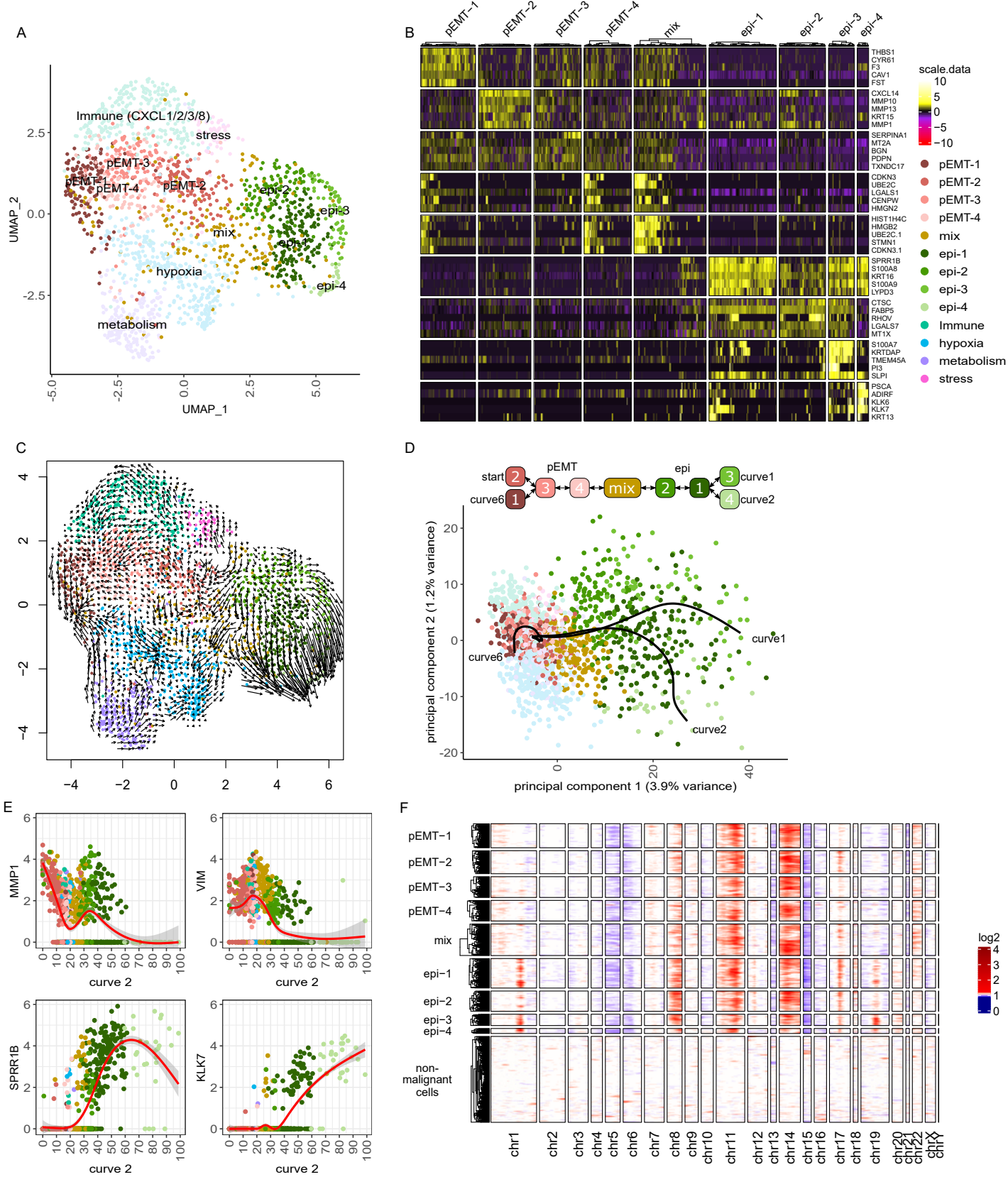

Figure 2 | Progressive epithelial differentiation, but no strong uniform direction of development in pEMT clusters. (A) UMAP of 1906 OSCC cells annotated based on SNN clustering, defining 4 pEMT (pEMT-1 to 4), 4 epithelial differentiated (epi-1 to 4) and one mixed (mix) cluster; clusters are numbered by size. **(B)** Heatmap for scaled, log-normalized gene expression in EMP-associated tumor cell phenotypes (columns) split by EMP cluster and their top 5 DEGs (rows) against all other EMP-related tumor cell phenotypes. DEGs are sorted from highest to lowest log2 foldchange. Row sections are ordered like column sections. **(C)** Projection of RNA velocity on the UMAP depicted in A. Arrows indicate the extrapolated direction of development and their length the strength of this future development. **(D)** First two principal components of OSCC cells with the three EMP-related principal curves that are derived from trajectory inference. Graph on top visualizes the relationship between EMP clusters described by the three principal curves forming a branching trajectory. **(E)** Log-normalized expression (y-axis) of MMP1, VIM, SPRR1B and KLK7 across pseudotime values (x-axis) of curve 2, color-coded by clusters. Red lines indicate smoothed expression values over the trajectory generated with a general additive model; 95 % confidence intervals are shaded gray. **(F)** Inferred CNVs across EMP-related tumor cells (rows) for all chromosomes (columns). Red indicates copy number gains, white diploid copy number and blue copy number loss. Columns show genes categorized in chromosomes and ordered by genome position; hence the size of the chromosome reflects the number of detected genes and not its nucleotide length. Mitochondrial genes were excluded.

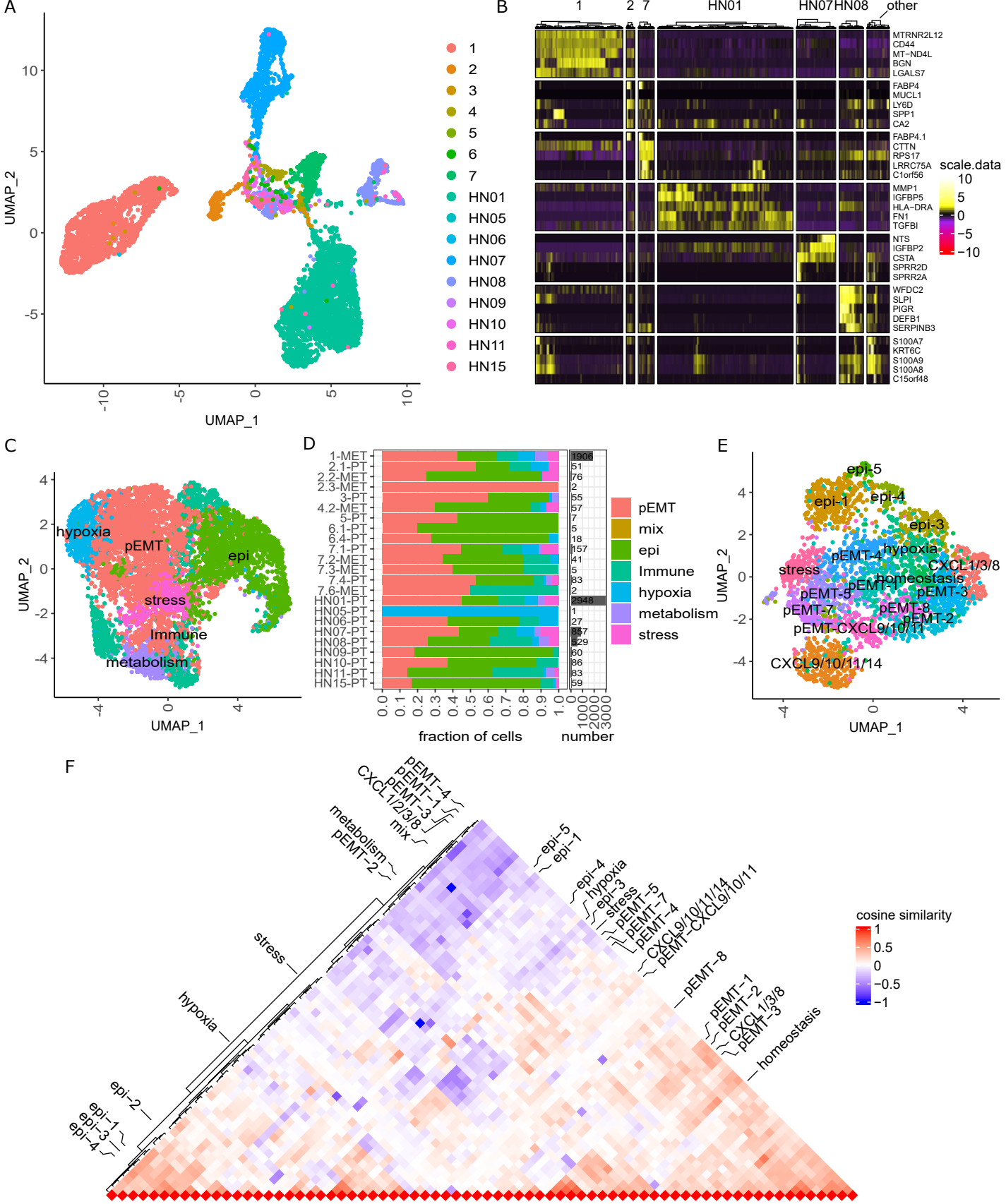

Figure 3 | Intra-tumoral heterogeneity of OSCC is driven by EMP. (A) UMAP based on scRNAseq data of 7,263 cancer cells from 16 different patients annotated by patient. **(B)** Heatmap for scaled, log-normalized gene expression of tumor cells split by patients and their top 5 DEGs (rows) against all other tumor cells. All patients with less than 100 cells are summarized in the 'other' column. DEGs are sorted from highest to lowest log2 foldchange. Row sections are ordered like column sections. **(C)** UMAP based on scRNAseq data depicted in A with PCs corrected for patient-specific effects using harmony. Cells are annotated according to their predominant phenotype. **(D)** Relative distribution of tumor cell phenotypes (left) and cancer cell abundance (right) across patients. The label on the y-axis shows the sample identification and tumor localization (primary tumor [PT] or lymph node metastasis [MET]). **(E)** UMAP based on scRNAseq data of 2,948 OSCC cells from patient HN01. Cells were annotated based on SNN clustering and the predominant phenotype. **(F)** Triangle heatmap of cosine similarity comparing the intratumoral heterogeneity across all patients. Cosine similarity is calculated between log2 fold changes from patient-specific clusters against all other tumor cells within the respective patient. Left side annotated are patient-specific clusters from patient 1 depicted in Figure 2A and right side from patient HN01 depicted in E. We included only patients with more than 50 tumor cells.

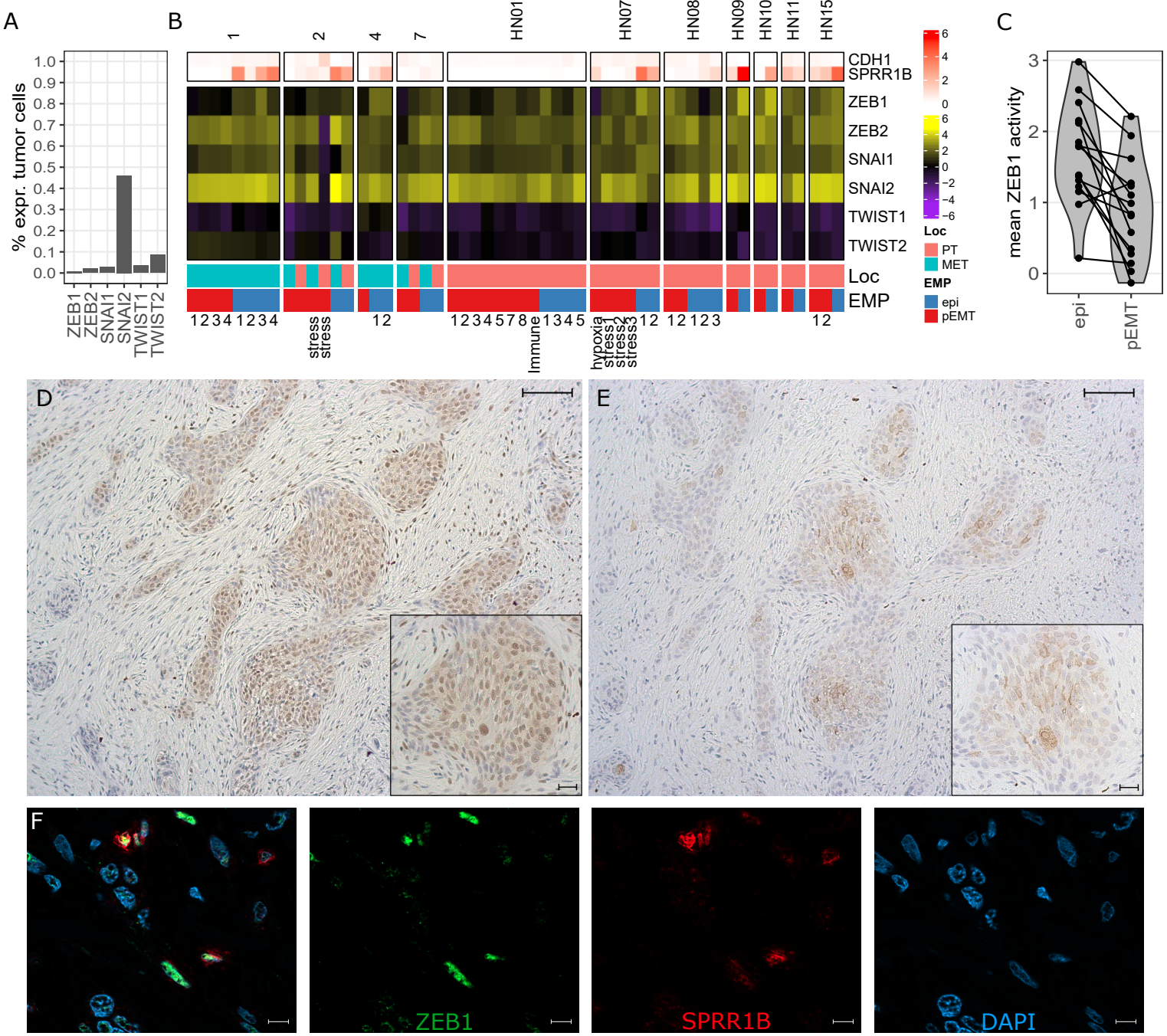

Figure 4 | ZEB1 is highly active in metastatic epithelial differentiated OSCC cells. (A) Percentage of tumor cells with detectable mRNA expression from scRNAseq (more than one UMI) encoding the indicted EMP-related transcription factors. **(B)** Mean inferred activity based on the target genes of the indicated transcription factors across tumor phenotypes from EMP-related patient-specific clusters. On top the log-normalized expression of *CDH1* and cornifin-B (*SPRR1B*) is shown, on bottom the localization (primary tumor [PT] or lymph node metastasis [MET]) and respective EMP phenotype of the cluster. **(C)** Mean activity of ZEB1 for epithelial differentiated and pEMT clusters of each patient, respectively. Connecting lines show dots belonging to the same patient. **(D, E)** ZEB1 (D) and cornifin-B (E) protein expression detected in serial sections by IHC of the primary tumor from patient 2; comparable areas are depicted. Scale bars equal 200 µm in overview and 100 µm in zoomed image. **(F)** Colocalized expression of ZEB1 (green) and cornifin-B (red) detected by double staining in the lymph node metastasis of patient 1 nuclei are stained in blue (DAPI), Scale bars equal 10 µm.