Title: Disseminating cells in human oral tumours acquire an EMT cancer stem cell state that is predictive of metastasis

Authors: Gehad Youssef[1], Luke Gammon[1], Leah Ambler[1], Sophia Lunetto[1], Alice Scemama[1], Hannah Cottom[2], Kim Piper[2], Ian C. Mackenzie[1], Michael P. Philpott[1], Adrian Biddle[1]*

Affiliation and address of authors: [1]Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, UK. [2]Department of Cellular Pathology, Barts Health NHS Trust, London, UK.

*Corresponding author: Adrian Biddle, Centre for Cell Biology and Cutaneous Research, Blizard Institute, 4 Newark Street, London, E1 2AT, UK. E-mail: a.biddle@qmul.ac.uk

Abstract

Cancer stem cells (CSCs) undergo epithelial-mesenchymal transition (EMT) to drive metastatic dissemination in experimental cancer models. However, tumour cells undergoing EMT have not been observed disseminating into the tissue surrounding human tumour specimens, leaving the relevance to human cancer uncertain. We have previously identified both EpCAM and CD24 as markers of EMT CSCs with enhanced plasticity. This afforded the opportunity to investigate whether retention of EpCAM and CD24 alongside upregulation of the EMT marker Vimentin can identify disseminating EMT CSCs in human tumours. Examining disseminating tumour cells in over 12,000 imaging fields from 84 human oral cancer specimens, we see a significant enrichment of single EpCAM, CD24 and Vimentin co-stained cells disseminating beyond the tumour body in metastatic specimens. Through training an

artificial neural network, these predict metastasis with high accuracy (cross-validated accuracy of 87-89%). In this study, we have observed single disseminating EMT CSCs in human oral cancer specimens, and these are highly predictive of metastatic disease.

Introduction

In multiple types of carcinoma, cancer stem cells (CSCs) undergo epithelial-mesenchymal transition (EMT) to enable metastatic dissemination from the primary tumour (Biddle et al., 2011; Lawson et al., 2015; Liu et al., 2014; Ruscetti et al., 2016). This model of metastatic dissemination has been built from studies using murine models and human cancer cell line models. However, this process has not been observed in human tumours in the *in vivo* setting, leading to uncertainty over the relevance of these findings to human tumour metastasis (Bill and Christofori, 2015; Williams et al., 2019). A key complication with efforts to study metastatic processes in human tumours is the inability to trace cell lineage. As cancer cells exiting the tumour downregulate epithelial markers whilst undergoing EMT, they become indistinguishable from the mesenchymal non-tumour cells surrounding the tumour (Li and Kang, 2016). Therefore, once these cells detach from the tumour body and move away they are lost to analysis. Attempts have been made to use the retention of epithelial markers alongside acquisition of mesenchymal markers to identify cells undergoing EMT in human tumours (Bronsert et al., 2014; Jensen et al., 2015; Puram et al., 2017). However, these studies were limited to characterising cells undergoing the earliest stages of EMT whilst still attached to the cohesive body of the primary tumour.

EMT must be followed by the reverse process of mesenchymal-to-epithelial transition (MET) to enable new tumour growth at secondary sites, and therefore retained plasticity manifested as ability to revert to an epithelial phenotype is an important feature of metastatic CSCs (Ocana et al., 2012; Tsai et al.,

48 2012). We have previously demonstrated that a CD44^{high}EpCAM$^{low/-}$ EMT population can be separated
49 from the main CD44^{low}EpCAM^{high} epithelial population in flow cytometric analysis of oral squamous
50 cell carcinoma (OSCC) cell lines and fresh tumour specimens (Biddle et al., 2016; Biddle et al., 2011).
51 We identified retained cell surface expression of EpCAM (Biddle et al., 2011) and CD24 (Biddle et al.,
52 2016) in a minority of cells that have undergone a full morphological EMT. Both EpCAM and CD24
53 were individually associated with enhanced ability to undergo MET, and thus are markers of EMT CSCs
54 exhibiting retained plasticity. We therefore reasoned that retention of one or both of these markers
55 may identify an important population of tumour cells that have undergone EMT and disseminated
56 from the primary tumour in human tumour specimens, and are responsible for subsequent metastatic
57 seeding. Here, we characterise the combined role of EpCAM and CD24 in marking a population of
58 disseminating tumour cells in human OSCC specimens. Staining for EpCAM and CD24 alongside the
59 mesenchymal marker Vimentin in over 12,000 imaging fields from 84 human tumour specimens,
60 stratified on metastatic status, identifies cells that have undergone EMT and disseminated into the
61 stromal region surrounding metastatic primary tumours. Using a machine learning approach, we show
62 that the presence of these EMT CSCs in the tumour stroma is predictive of metastasis.

63

64 **Results**

65

66 **Identification of human tumour cells that have undergone an EMT and disseminated into the**
67 **surrounding stromal region**

68 The retention of EpCAM expression in a sub-population of tumour cells that have undergone EMT
69 raised the prospect that we may be able to identify these cells outside of the tumour body in human
70 tumour specimens, as EpCAM is a specific epithelial marker that would not normally be found in the
71 surrounding stromal region. In combination with EpCAM, we stained tumour specimens for CD24 as a

72 second marker of plastic EMT CSCs, and Vimentin as a mesenchymal marker to identify cells that have
73 undergone EMT. Notably, CD44 cannot be used as an EMT marker in the context of intact tissue as it
74 requires trypsin degradation in order to yield differential expression in EMT and epithelial populations
75 (Biddle et al., 2013; Mack and Gires, 2008). Vimentin, on the other hand, accurately distinguishes EMT
76 from epithelial tumour cells in immunofluorescent staining protocols (Biddle et al., 2016). By
77 combining EpCAM as a tumour lineage and EMT CSC marker, Vimentin as a mesenchymal marker, and
78 CD24 as a plastic EMT CSC marker, we aimed to identify tumour cells that have undergone EMT and
79 disseminated into the surrounding stromal region. For this, we developed a protocol for automated
80 4-colour (3 markers + nuclear stain) immunofluorescent imaging and analysis of entire
81 histopathological slide specimens, to test for co-localisation of the 3 markers in each individual cell
82 across each specimen.

83

84 To determine whether this marker combination identifies EMT CSCs, we initially tested the protocol
85 on the CA1 OSCC cell line and an EMT CSC sub-line that is a derivative of this cell line (EMT-stem sub-
86 line) (Biddle et al., 2016). $EpCAM^{+}Vim^{+}CD24^{+}$ cells were greatly enriched in the EMT-stem sub-line,
87 comprising 41% of the population, compared to 2.1% in the CA1 line (Figure 1A, B, E). Cells with this
88 staining profile were absent from normal keratinocyte culture and cancer associated fibroblast culture
89 (Supplementary Figure S1). To test the specific role of EpCAM retention, we replaced EpCAM with a
90 pan-keratin antibody against epithelial keratins. There was very little $Pan\text{-}keratin^{+}Vim^{+}CD24^{+}$ staining,
91 and no enrichment for $Pan\text{-}keratin^{+}Vim^{+}CD24^{+}$ cells in the EMT-stem sub-line (Figure 1C, D, E).
92 Therefore, whilst epithelial keratins are lost, EpCAM is retained in cells undergoing EMT and an
93 $EpCAM^{+}Vim^{+}CD24^{+}$ staining profile can be used as a marker for EMT CSCs in immunofluorescent
94 staining protocols.

95

96 Imaging the tumour body and adjacent stroma in sections of human OSCC specimens, we detected
97 single cells co-expressing EpCAM, Vimentin and CD24 in the stromal region surrounding the tumour
98 (Figure 1F), confirming that these cells can be detected in human tumour specimens. We next
99 stratified 24 human primary OSCC specimens into 12 tumours that had evidence of lymph node
100 metastasis or perineural spread, and 12 that remained metastasis free (Supplementary Figure S2), and
101 stained them for EpCAM, Vimentin and CD24. Single cells co-expressing EpCAM, Vimentin and CD24
102 were abundant in the stroma surrounding metastatic tumours. This was not the case in non-metastatic
103 tumours or normal epithelial regions (Figure 2, A-C). In contrast to EpCAM, pan-keratin staining did
104 not identify cells in the stroma surrounding metastatic tumours (Figure 2D).

105

106 We developed an image segmentation protocol that separated the tumour body from the adjacent
107 stroma, thus enabling each nucleated cell to be assigned to either the tumour or stromal region in
108 automated image analysis (Figure 2E). Expression of EpCAM, Vimentin and CD24 was then analysed
109 for every nucleated cell in every imaging field that included both tumour and stroma (3500 manually
110 curated imaging fields across the 24 tumours). This enabled the proportion of each cell type in each
111 region to be quantified (Figure 2F). $EpCAM^{+}Vim^{+}CD24^{+}$ cells were enriched in the stroma compared to
112 the tumour body, and there was a much greater accumulation of $EpCAM^{+}Vim^{+}CD24^{+}$ cells in the
113 stroma of metastatic tumours compared to non-metastatic tumours. Interestingly, this was not the
114 case for $EpCAM^{+}Vim^{+}CD24^{-}$ cells, which were also enriched in the stroma but showed no difference
115 between metastatic and non-metastatic tumours. Pan-$keratin^{+}Vim^{+}CD24^{+}$ cells were not detected.

116

117 To extend this analysis, we stained and imaged a further 60 tumours, evenly stratified on the same
118 criteria. These displayed the same evidence of individual disseminating cells co-expressing EpCAM,
119 Vimentin and CD24 in metastatic tumours only (Figure 2G and Supplementary Figure S3F, G). For these
120 tumours, using a variation on the previous image segmentation protocol (Supplementary Figure S3,

121 A-D), the proportion of EpCAM$^+$Vim$^+$CD24$^+$ and EpCAM$^+$Vim$^+$CD24$^-$ cells was quantified for each cell in
122 over 9000 imaging fields at the tumour-stroma boundary (Supplementary Figure S3E). Consistent with
123 the previous set of tumours, only EpCAM$^+$Vim$^+$CD24$^+$ cells were specifically enriched in the stroma of
124 metastatic tumours.

125

126 To explore whether these EpCAM$^+$Vim$^+$CD24$^+$ cells in the stroma may in fact be non-tumour cell types,
127 we analysed a published scRNAseq dataset for human head and neck cancer (Puram et al., 2017). In
128 this dataset, tumour and non-tumour cells were separated using bioinformatic techniques (principally
129 inferred CNV and a 'tumour-epithelial' expression signature). Analysing this dataset for EpCAM,
130 Vimentin and CD24 co-expression, we found that 12% of tumour cells (267/2215) were
131 EpCAM$^+$Vim$^+$CD24$^+$. In the non-tumour cells, only 0.8% (29/3687) were EpCAM$^+$Vim$^+$CD24$^+$
132 (Supplementary Figure S4). Therefore, the observed EpCAM$^+$Vim$^+$CD24$^+$ cells in our tumour specimens
133 are highly likely to be a tumour cell population. Indeed, use of EpCAM as a tumour lineage marker is
134 specifically intended to exclude staining for stromal constituents. EpCAM is a specific epithelial
135 marker, that is not expressed in stromal or immune cells – it is expressed exclusively in epithelia and
136 epithelial-derived tumours (Keller et al., 2019).

137

138 These findings demonstrate that an EpCAM$^+$Vim$^+$CD24$^+$ staining profile marks tumour cells
139 disseminating into the surrounding stroma, and that these cells are enriched specifically in metastatic
140 tumours. The presence of disseminating tumour cells that express EpCAM but not CD24 did not
141 correlate with metastasis. This highlights a requirement for the plasticity marker CD24, when
142 identifying disseminating metastatic CSCs.

143

144 **Identification of EpCAM$^+$CD24$^+$Vim$^+$ CSCs enables clinical prediction using a machine learning**
145 **approach**

146 OSCC are an important health burden and one of the top ten cancers worldwide, with over 300,000
147 cases annually and a 50% 5-year survival rate. There is frequent metastatic spread to the lymph nodes
148 of the neck; this is the single most important predictor of outcome and an important factor in
149 treatment decisions (Sano and Myers, 2007). If spread to the lymph nodes is suspected, OSCC
150 resection is accompanied by neck dissection to remove the draining lymph nodes, a procedure with
151 significant morbidity. At presentation it is currently very difficult to determine which tumours are
152 metastatic and this results in sub-optimal tailoring of treatment decisions. Accurate prediction of
153 metastasis would therefore have great potential to improve clinical management of the disease to
154 reduce both mortality and treatment-related morbidity. We sought to determine whether the
155 EpCAM$^+$CD24$^+$Vim$^+$ staining pattern could be predictive of metastasis.

156

157 Starting with the EpCAM, Vimentin and CD24 immunofluorescence grey levels for each nucleated cell,
158 we used a supervised machine learning approach to predict whether an imaging field comes from a
159 metastatic or non-metastatic tumour (Figure 5A). As a benchmark we used the pan-keratin, Vimentin
160 and CD24 immunofluorescence grey levels, as we hypothesised that pan-keratin would provide an
161 inferior predictive value than EpCAM given that there was no dissemination of pan-keratin expressing
162 cells in the stroma. 3500 imaging fields containing 2,640,000 total nucleated cells from 24 tumour
163 specimens were used in the machine learning task. We compared the performance accuracy (10-fold
164 cross-validated F-score) of different machine learning classification algorithms. The best performing
165 classifiers for EpCAM, Vimentin and CD24 were the artificial neural network (ANN) and support vector
166 machine (SVM), with F1 accuracy scores of 91% and 87% respectfully (Figure 5B). For the ANN, the
167 area under the curve (AUC) accuracy score was 87%, with 94% sensitivity and 82% specificity. Training
168 with Pan-keratin, Vimentin and CD24 gave much worse prediction across all classifiers (Figure 5C).

169 These findings demonstrate that, utilising a machine learning algorithm, staining for EpCAM, Vimentin
170 and CD24 can predict metastatic status with high accuracy and may therefore have clinical utility.

171

172 To extend this analysis of utility for metastasis prediction, we stained and imaged a further 60
173 tumours, evenly stratified on the same criteria, for EpCAM, Vimentin and CD24. Over 9000 imaging
174 fields at the tumour-stroma boundary from 60 evenly stratified tumour specimens, containing over
175 8.5 million nucleated cells, were fed into an artificial neural network machine learning task. For this
176 task, we recorded the predictive accuracy from the training and validation sets after each training
177 epoch, which showed good alignment and an 89% accuracy score after 12 training epochs (Figure 5D).

178

179 To our knowledge, this is the first time immunofluorescent staining of human tumour tissue specimens
180 has been used in a machine learning pipeline for clinical prediction. Previous studies using cytokeratin
181 immunohistochemistry, clinicopathological data and serum biomarkers for clinical prediction via
182 machine learning have achieved AUCs of 75% in breast cancer (Tseng et al., 2019), 80% in OSCC (Bur
183 et al., 2019), and 82% in colorectal cancer (Takamatsu et al., 2019).

184

185 **Discussion**

186 The role of EMT in tumour dissemination has long been debated but, lacking evidence of cells
187 undergoing EMT whilst disseminating from human tumours *in vivo*, this role has had to be inferred
188 from mouse models and human cell line models. Here, through applying our understanding of EMT
189 cancer cell heterogeneity and markers for plastic EMT CSCs, we have identified EMT CSCs
190 disseminating from the primary tumour in human pathological specimens. Importantly, the presence
191 of these disseminating stem cells is strongly correlated with tumour metastasis. Using a machine

192 learning approach, we have demonstrated the ability to predict metastasis with high accuracy through
193 staining for these EMT CSCs.

194

195 A partial EMT state has previously been identified in an OSCC scRNAseq dataset; this state retained
196 epithelial gene expression alongside expression of mesenchymal genes, and was correlated with nodal
197 metastasis and adverse pathological features (Puram et al., 2017). Here, using immunofluorescent
198 staining for EMT CSCs that retain the epithelial marker EpCAM alongside the mesenchymal marker
199 Vimentin and the CSC plasticity marker CD24, we have identified single EMT CSCs disseminating into
200 the stroma surrounding oral tumours. However, epithelial keratins are not retained. We have also
201 shown that retention of EpCAM is not on its own sufficient alongside Vimentin to mark disseminating
202 EMT CSCs that correlate with metastasis. There is a requirement for CD24, which we have previously
203 shown to be a plasticity marker within the EMT population even when driven into full morphological
204 EMT under TGFβ treatment (Biddle et al., 2016). This suggests that the EMT CSC state may be more
205 complex than a simple coalescence of epithelial and mesenchymal characteristics.

206

207 We have identified an EMT CSC state that disseminates as single cells from human tumours and is
208 correlated with metastasis. Immunofluorescent antibody co-staining for EpCAM, CD24 and Vimentin
209 identifies these EMT CSCs in human tumour specimens and is predictive of metastasis. The ability of
210 this co-staining to separate disseminating tumour cells from the stromal content of human tumours,
211 which has confounded previous attempts to develop a predictive EMT signature (Tan et al., 2014), is
212 one important factor in this success. However, we also show that $EpCAM^{+}CD24^{-}Vim^{+}$ tumour cells in
213 the stroma do not correlate with metastasis, and therefore the clinically predictive utility of tumour
214 cell staining in the stroma can be isolated specifically to the $EpCAM^{+}CD24^{+}Vim^{+}$ EMT CSCs. This
215 highlights the value of using techniques that give single cell resolution, enabling isolation of the signal
216 to the specific cell type of interest within a highly heterogeneous cellular environment. An important

217 strength of our study has been the ability to look at the single cell level in an automated fashion across
218 thousands of fields of view from human tumours, enabling us to observe and quantify human tumour
219 cells disseminating into the surrounding tissue. In doing so, we have identified single disseminating
220 EMT CSCs that are predictive of metastasis.

221

222 **Conflict of interest**

223 The authors declare no conflicts of interest.

224

225 **Acknowledgements**

226 We thank Ryan O'Shaughnessy, Sarah Marzi, and Jan Soetaert for technical assistance and discussion.
227 Gehad Youssef and Adrian Biddle were supported by Animal Free Research UK, as part of the Animal
228 Replacement Centre of Excellence at Queen Mary University of London. Leah Ambler was supported
229 by Oracle Cancer Trust.

230

231 **Methods**

232 **Cell culture**

233 The CA1 OSCC cell line and oral cancer associated fibroblasts were both previously derived in our
234 laboratory, from separate biopsies of OSCC of the floor of the mouth. The EMT-stem sub-line was
235 derived as a single cell clone from the CA1 cell line (Biddle et al., 2016). Normal keratinocytes were
236 the N/TERT hTERT-immortalised epidermal keratinocyte cell line (Smits et al., 2017). Cell culture was
237 performed as previously described (Biddle et al., 2011). Cell removal from adherent culture was
238 performed using 1x Trypsin-EDTA (Sigma, T3924) at 37ºC.

239

240

241 **Immunofluorescent staining of cell lines and tumour tissue sections**

242 Tumour specimens were obtained from the pathology department at Barts Health NHS Trust, with full
243 local ethical approval and patients' informed consent. Sections of formalin fixed paraffin embedded
244 (FFPE) archival specimens were dewaxed by clearing twice in xylene for 5 minutes then gradually
245 hydrating the specimens in an alcohol gradient (100%, 90%, 70%) for 3 minutes each. The sections
246 were then washed under running tap water before immersing the slides in Tris-EDTA pH9 for antigen
247 retrieval using a standard microwave at high power for 2 minutes and then 8 minutes at low power.

248

249 Four-colour immunofluorescent staining was performed by firstly staining the membranous proteins
250 prior to the permeabilisation and blocking steps. The sections were incubated with an IgG2a mouse
251 monoclonal CD24 antibody (clone ML5, BD Bioscience) and IgG rabbit recombinant monoclonal
252 EpCAM antibody (EPR20532-225, Abcam) in PBS overnight at 4°C (1/100 dilution). The sections were
253 then washed three times in PBS and incubated for 1 hour at room temperature with anti-mouse IgG2
254 Alexa Fluor 488 and anti-rabbit IgG Alexa Fluor 555 secondary antibodies (1/500 dilution). The sections
255 were then washed in PBS and permeabilised with 0.5% triton-X in PBS for 10 minutes followed by
256 blocking for 1 hour with blocking buffer (3% goat serum, 2% bovine serum albumin in PBS). The
257 sections were then incubated with an IgG1 mouse monoclonal Vimentin antibody (clone V9, Dako)
258 and (optionally, in place of EpCAM) IgG rabbit polyclonal wide spectrum cytokeratin antibody (ab9377,
259 Abcam) overnight at 4°C in blocking buffer (1/100 dilution). After washing with PBS, the sections were
260 incubated with anti-mouse IgG1 Alexa Fluor 647 antibody and (optionally) anti-rabbit IgG Alexa Fluor
261 555 for 1hr at 4°C (1/500 dilution). After washing three times with PBS, cell nuclei were stained with
262 DAPI (1/1000 dilution in PBS) for 10 minutes.

263

264 For cell line staining, cells were fixed in 4% PFA for 10 minutes then washed with PBS. Staining was
265 performed in the same manner as described above, however permeabilisation was performed with
266 0.25% Triton-X for 10 minutes and DAPI incubation was reduced to 1 minute.

267

268 **Quantifying the abundance of stained sub-populations in cell lines and tumour tissue sections**

269 Imaging of the stained slides was performed using the In Cell Analyzer 2200 (GE), a high content
270 automated fluorescence microscope with four-colour imaging capability. The slides were imaged at
271 x20 and x40 magnification. An image segmentation protocol was developed to extract grey level
272 intensities corresponding to EpCAM, Vimentin and CD24 expression for every DAPI stained nucleated
273 cell in the tumour body and the adjacent stroma separately. Segmentation was performed using the
274 Developer Toolbook software (GE). As shown in figure 2E and Supplementary Figure S3, an 'EpCAM
275 dense cloud' or 'Vimentin dense cloud' was generated to isolate individual nucleated cells in the
276 tumour body from the adjacent stroma and analyse them separately.

277

278 Grey level intensities obtained from the imaging analysis were processed in the following way. Firstly,
279 the median number of nucleated cells was calculated and imaging fields with fewer than 20% of the
280 median nucleated cells were excluded from the analysis pipeline. The folded edges of a specimen were
281 also excluded. The median grey level intensity of the FITC, CY3 and CY5 fluorescence channels
282 corresponding to CD24, EpCAM and Vimentin expression were computed for the negative control
283 stained slides. A nucleated cell was deemed to have positive CD24, EpCAM or Vimentin expression if
284 its grey level intensity exceeded the background threshold value (1.5 x median grey level intensity of
285 negative control slide) for the FITC, CY3 and CY5 channels respectively. If a nucleated cell surpassed
286 the background threshold for all three fluorescence channels it was termed a triple positive cell
287 ($CD24^{+}EpCAM^{+}Vim^{+}$) and denoted with 1 and if this criteria was not met the nucleated cell was

288 denoted with a 0. For $EpCAM^{+}Vim^{+}CD24^{-}$ cells (termed double positive), the nucleated cell must
289 exceed the background threshold for the CY3 and CY5 channels but not the FITC.

290

291 The scRNAseq dataset (Puram et al., 2017) was analysed using a threshold (median or quartile) using
292 the normalised count expression for EpCAM, CD24 and Vimentin for each cell.

293

294 **Machine learning for prognostic prediction using immunofluorescent staining data**

295 A dataset was created of a pool of 2,640,000 nucleated cells across 3500 imaging fields from 24 tumour
296 specimens (12 with lymph node metastasis or perineural spread, and 12 without) (batch 1) or
297 8,563,000 nucleated cells across 9,200 imaging fields from 60 tumour specimens (30 with lymph node
298 metastasis or perineural spread, and 30 without) (batch 2). The background threshold for the FITC,
299 CY3 and CY5 channels was subtracted from the grey level intensities for each nucleated cell. The
300 supervised machine learning task was to classify each imaging field into whether it belonged to a
301 metastatic or non-metastatic tumour.

302

303 The dataset was stratified into a training and validation cohort in a 70%:30% ratio using a random seed
304 split. Supervised machine learning approaches were implemented using the skikit-learn Python 3.6
305 libraries (Pedregosa et al., 2011) and Tensorflow/Keras framework
306 (https://www.tensorflow.org/api_docs/python/tf/keras/models). Hyper-parameter optimisation was
307 performed by an exhaustive grid search and computed on Apocrita, a high performance cluster (HPC)
308 facility at Queen Mary University of London (http://doi.org/10.5281/zenodo.438045). To further
309 minimise overfitting, 10-fold cross-validation was performed and the mean accuracy metric, F1 score,
310 was obtained for each learning iteration. Receiver-of-operator (ROC) curves and the area-under the-
311 curve (AUC) were computed for the optimum supervised learning algorithm. Supervised approaches

312 used were logistic regression, support vector machines (Smola and Scholkopf, 2004), Naïve Bayes
313 (Zhang, 2005), K-Nearest Neighbours (Bentley, 1975), decision trees (Dumont et al., 2009), and
314 artificial neural networks (Rumelhart et al., 1986).

315

316 References

317

318 Bentley, J. L. (1975). Multidimensional Binary Search Trees Used for Associative Searching. Commun
319 Acm *18*, 509-517.
320 Biddle, A., Gammon, L., Fazil, B., and Mackenzie, I. C. (2013). CD44 staining of cancer stem-like cells is
321 influenced by down-regulation of CD44 variant isoforms and up-regulation of the standard CD44
322 isoform in the population of cells that have undergone epithelial-to-mesenchymal transition. PloS one
323 *8*, e57314.
324 Biddle, A., Gammon, L., Liang, X., Costea, D. E., and Mackenzie, I. C. (2016). Phenotypic Plasticity
325 Determines Cancer Stem Cell Therapeutic Resistance in Oral Squamous Cell Carcinoma. EBioMedicine
326 *4*, 138-145.
327 Biddle, A., Liang, X., Gammon, L., Fazil, B., Harper, L. J., Emich, H., Costea, D. E., and Mackenzie, I. C.
328 (2011). Cancer stem cells in squamous cell carcinoma switch between two distinct phenotypes that
329 are preferentially migratory or proliferative. Cancer Res *71*, 5317-5326.
330 Bill, R., and Christofori, G. (2015). The relevance of EMT in breast cancer metastasis: Correlation or
331 causality? FEBS Lett *589*, 1577-1587.
332 Bronsert, P., Enderle-Ammour, K., Bader, M., Timme, S., Kuehs, M., Csanadi, A., Kayser, G., Kohler, I.,
333 Bausch, D., Hoeppner, J.*, et al.* (2014). Cancer cell invasion and EMT marker expression: a three-
334 dimensional study of the human cancer-host interface. J Pathol *234*, 410-422.
335 Bur, A. M., Holcomb, A., Goodwin, S., Woodroof, J., Karadaghy, O., Shnayder, Y., Kakarala, K., Brant, J.,
336 and Shew, M. (2019). Machine learning to predict occult nodal metastasis in early oral squamous cell
337 carcinoma. Oral Oncol *92*, 20-25.
338 Dumont, M., Maree, R., Wehenkel, L., and Geurts, P. (2009). Fast Multi-Class Image Annotation with
339 Random Subwindows and Multiple Output Randomized Trees. Visapp 2009: Proceedings of the Fourth
340 International Conference on Computer Vision Theory and Applications, Vol 2, 196-+.
341 Jensen, D. H., Dabelsteen, E., Specht, L., Fiehn, A. M., Therkildsen, M. H., Jønson, L., Vikesaa, J., Nielsen,
342 F. C., and von Buchwald, C. (2015). Molecular profiling of tumour budding implicates TGFβ-mediated
343 epithelial-mesenchymal transition as a therapeutic target in oral squamous cell carcinoma. J Pathol
344 *236*, 505-516.
345 Keller, L., Werner, S., and Pantel, K. (2019). Biology and clinical relevance of EpCAM. Cell Stress *3*, 165-
346 180.
347 Lawson, D. A., Bhakta, N. R., Kessenbrock, K., Prummel, K. D., Yu, Y., Takai, K., Zhou, A., Eyob, H.,
348 Balakrishnan, S., Wang, C. Y.*, et al.* (2015). Single-cell analysis reveals a stem-cell program in human
349 metastatic breast cancer cells. Nature *526*, 131-135.
350 Li, W., and Kang, Y. (2016). Probing the Fifty Shades of EMT in Metastasis. Trends Cancer *2*, 65-67.
351 Liu, S., Cong, Y., Wang, D., Sun, Y., Deng, L., Liu, Y., Martin-Trevino, R., Shang, L., McDermott, S. P.,
352 Landis, M. D.*, et al.* (2014). Breast Cancer Stem Cells Transition between Epithelial and Mesenchymal
353 States Reflective of their Normal Counterparts. Stem cell reports *2*, 78-91.
354 Mack, B., and Gires, O. (2008). CD44s and CD44v6 expression in head and neck epithelia. PloS one *3*,
355 e3360.

356 Ocana, O. H., Corcoles, R., Fabra, A., Moreno-Bueno, G., Acloque, H., Vega, S., Barrallo-Gimeno, A.,
357 Cano, A., and Nieto, M. A. (2012). Metastatic colonization requires the repression of the epithelial-
358 mesenchymal transition inducer Prrx1. Cancer cell *22*, 709-724.
359 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer,
360 P., Weiss, R., Dubourg, V.*, et al.* (2011). Scikit-learn: Machine Learning in Python. J Mach Learn Res *12*,
361 2825-2830.
362 Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C. L., Mroz,
363 E. A., Emerick, K. S.*, et al.* (2017). Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor
364 Ecosystems in Head and Neck Cancer. Cell *171*, 1611-1624 e1624.
365 Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning Representations by Back-
366 Propagating Errors. Nature *323*, 533-536.
367 Ruscetti, M., Dadashian, E. L., Guo, W., Quach, B., Mulholland, D. J., Park, J. W., Tran, L. M., Kobayashi,
368 N., Bianchi-Frias, D., Xing, Y.*, et al.* (2016). HDAC inhibition impedes epithelial-mesenchymal plasticity
369 and suppresses metastatic, castration-resistant prostate cancer. Oncogene *35*, 3781-3795.
370 Sano, D., and Myers, J. N. (2007). Metastasis of squamous cell carcinoma of the oral tongue. Cancer
371 metastasis reviews *26*, 645-662.
372 Smits, J. P. H., Niehues, H., Rikken, G., van Vlijmen-Willems, I. M. J. J., van de Zande, G. W. H. J. F.,
373 Zeeuwen, P. L. J. M., Schalkwijk, J., and van den Bogaard, E. H. (2017). Immortalized N/TERT
374 keratinocytes as an alternative cell source in 3D human epidermal models. Scientific Reports *7*, 11838.
375 Smola, A. J., and Scholkopf, B. (2004). A tutorial on support vector regression. Stat Comput *14*, 199-
376 222.
377 Takamatsu, M., Yamamoto, N., Kawachi, H., Chino, A., Saito, S., Ueno, M., Ishikawa, Y., Takazawa, Y.,
378 and Takeuchi, K. (2019). Prediction of early colorectal cancer metastasis by machine learning using
379 digital slide images. Comput Methods Programs Biomed *178*, 155-161.
380 Tan, T. Z., Miow, Q. H., Miki, Y., Noda, T., Mori, S., Huang, R. Y., and Thiery, J. P. (2014). Epithelial-
381 mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug
382 responses of cancer patients. EMBO Mol Med *6*, 1279-1293.
383 Tsai, J. H., Donaher, J. L., Murphy, D. A., Chau, S., and Yang, J. (2012). Spatiotemporal regulation of
384 epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. Cancer cell *22*,
385 725-736.
386 Tseng, Y. J., Huang, C. E., Wen, C. N., Lai, P. Y., Wu, M. H., Sun, Y. C., Wang, H. Y., and Lu, J. J. (2019).
387 Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with
388 machine learning technologies. Int J Med Inform *128*, 79-86.
389 Williams, E. D., Gao, D., Redfern, A., and Thompson, E. W. (2019). Controversies around epithelial-
390 mesenchymal plasticity in cancer metastasis. Nat Rev Cancer *19*, 716-732.
391 Zhang, H. (2005). Exploring conditions for the optimality of Naive bayes. Int J Pattern Recogn *19*, 183-
392 198.

393

394

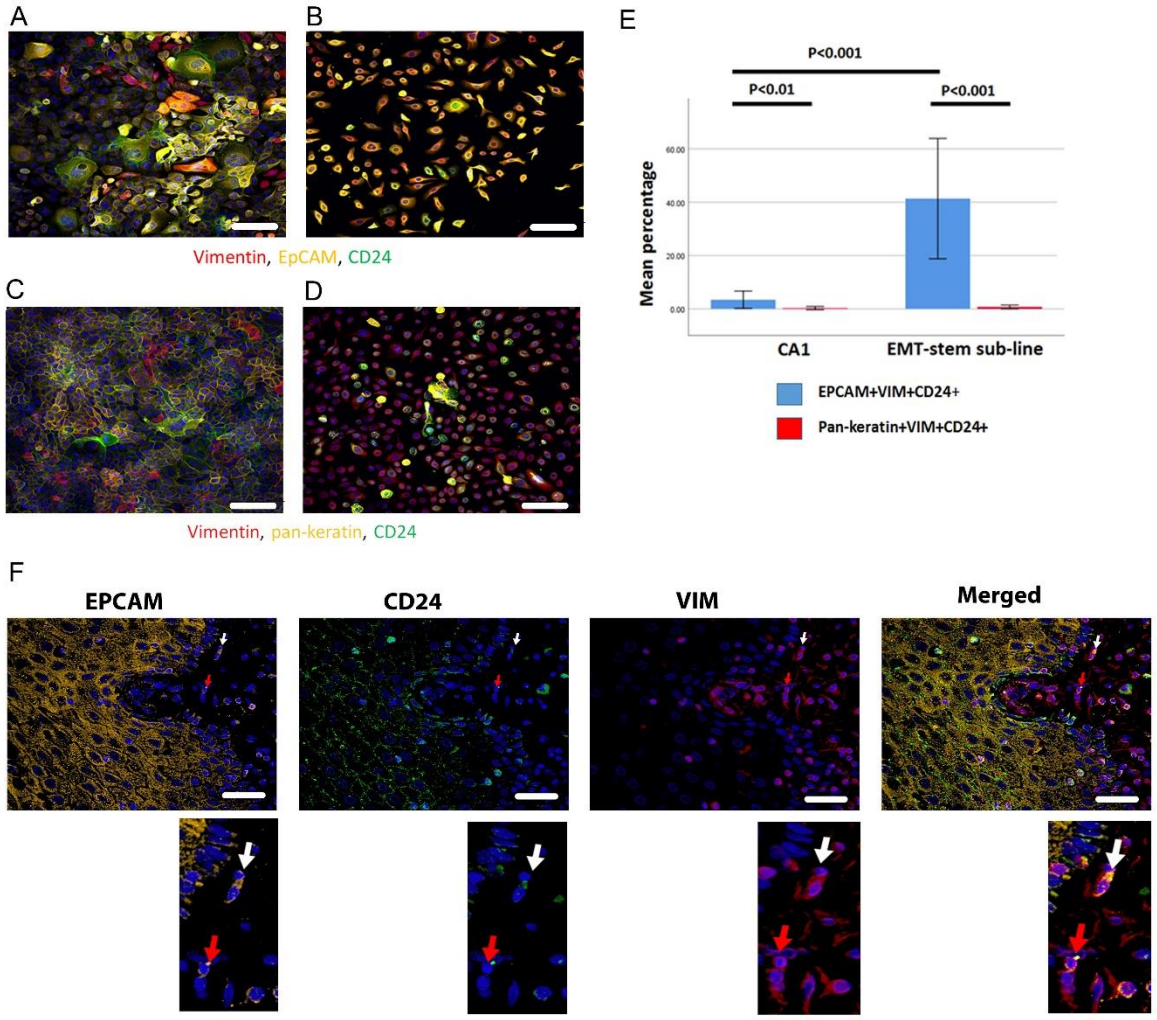
395

396

397

398

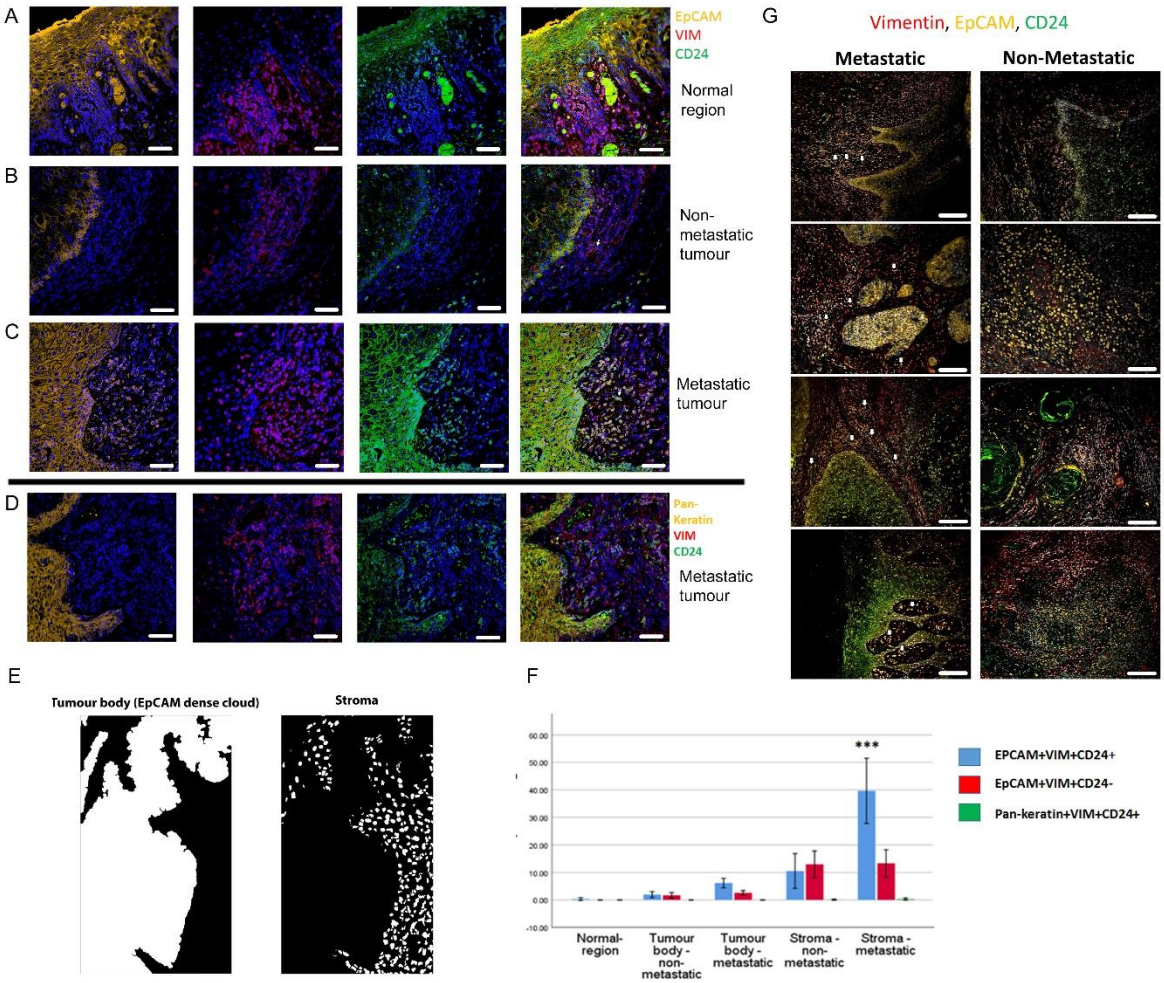
399 **Figure 1**



400

401 Figure 1 – Immunofluorescent co-staining for EpCAM, Vimentin and CD24 identifies the EMT stem cell
402 state. **A-D,** Immunofluorescent staining for EpCAM, Vimentin and CD24 (A, B) and pan-keratin,
403 Vimentin and CD24 (C, D) in the CA1 cell line (A, C) and the EMT-stem CA1 sub-line (B, D). **E,**
404 Quantification of the percentage of EpCAM^{+}Vim^{+}CD24^{+} and pan-keratin^{+}Vim^{+}CD24^{+} cells in the CA1
405 cell line and EMT-stem sub-line. Significance is obtained from a two-tailed student t-test. The graph
406 shows mean +/- 95% confidence interval. **F,** Detection of EpCAM^{+}Vim^{+}CD24^{+} cells in the stroma
407 surrounding an oral cancer tumour specimen. The white arrow highlights an EpCAM^{+}Vim^{+}CD24^{+} cell in
408 the stroma. The red arrow highlights an EpCAM^{+}Vim^{+}CD24^{-} cell in the stroma. DAPI nuclear stain is
409 blue. Below inset; enlargement of the highlighted cells for each marker. Scale bars = 100µm.

410 **Figure 2**



411

412 Figure 2 – Enrichment of EpCAM^{+}Vim^{+}CD24^{+} cells in the stroma surrounding metastatic tumours. **A-C,**
413 Immunofluorescent four-colour staining of oral tumour specimens for EpCAM (yellow), Vimentin (red)
414 and CD24 (green) with DAPI nuclear stain (blue). Representative imaging fields from a normal
415 epithelial region (A), a non-metastatic tumour (B) and a metastatic tumour (C). **D,** Staining of a
416 metastatic tumour for pan-keratin, Vimentin and CD24. **E,** Image segmentation was performed, with
417 generation of an 'EpCAM dense cloud' to distinguish the tumour body from the stroma. Grey level
418 intensities for EpCAM, Vimentin and CD24 were obtained for every nucleated cell in each imaging
419 field. **F,** Quantification of the percentage of EpCAM^{+}Vim^{+}CD24^{+}, EpCAM^{+}Vim^{+}CD24^{-} and pan-
420 keratin^{+}Vim^{+}CD24^{+} cells in normal region (epithelium distant from the tumour), tumour body, and
421 stromal region from metastatic and non-metastatic tumours in the first batch of specimens. A student

422 t-test was performed comparing the mean percentage of $EpCAM^+Vim^+CD24^+$ co-expressing cells in the
423 metastatic stroma compared to the other fractions. *** signifies $p < 0.001$. The graph shows mean
424 +/- 95% confidence interval. **G,** Immunofluorescent four-colour staining of oral tumours from the
425 second batch of specimens, showing tumours with a range of invasive front presentations. White
426 arrows highlight single $EpCAM^+Vim^+CD24^+$ cells in the stroma. Scale bars = 100 µm.

427

428

429

430

431

432

433

434

435

436

437

438

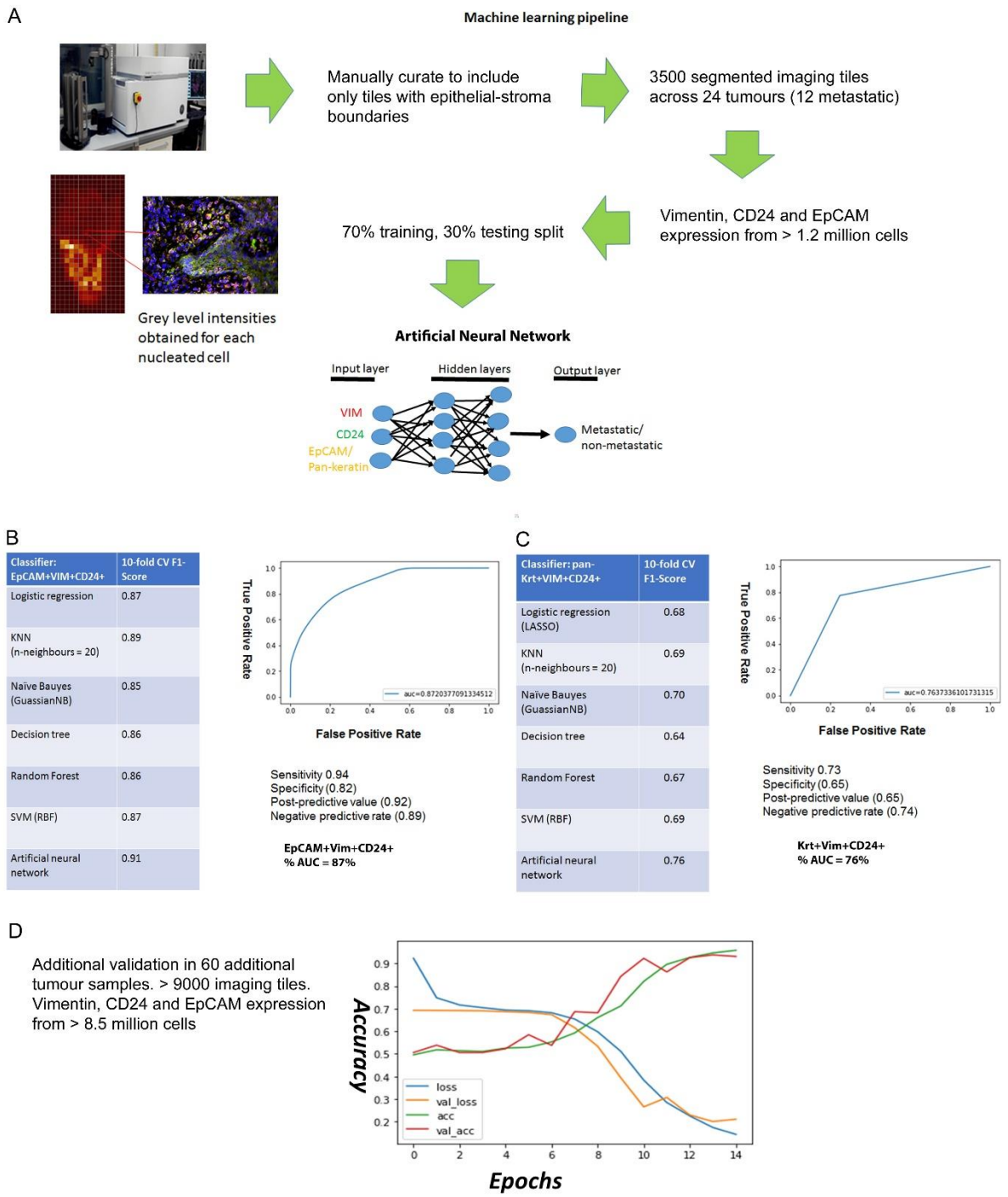
439

440

441

442

443 **Figure 3**



444

445 Figure 3 – Predicting metastasis using EpCAM, Vimentin and CD24 immunofluorescent staining and a

446 supervised machine learning approach. **A,** Pipeline for machine learning based on grey level intensities

447 for the three markers in tumour batch 1. The training tiles were classified as coming from a metastatic

448 or non-metastatic tumour. **B, C,** Performance of EpCAM, Vimentin and CD24 (B) and pan-keratin,

Vimentin and CD24 (C) in the supervised learning task on tumour batch 1. The tables show the 10-fold cross-validation F1 scores of different machine learning classification algorithms. To the right of each table is a receiver-of-operator curve (ROC) showing the area under the curve (AUC) of the artificial neural network (ANN) classifier. **D,** Performance of EpCAM, Vimentin and CD24 in the supervised learning task on tumour batch 2. An ANN classifier was trained and tested on batch 2, independently of tumour batch 1. Accuracy and loss scores are displayed for the training set (green and blue lines) and the validation set (red and yellow lines) drawn from within this batch, for 14 training epochs on the ANN classifier.