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

Authors: Gehad Youssef[1], Luke Gammon[1], Leah Ambler[1], Bethan Wicker[1], Swatisha Patel[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. Tel: +44 (0)20 7882 2348, E-mail: a.biddle@qmul.ac.uk

Abstract

Cancer stem cells 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. Here, we identify an EMT stem cell state that retains EpCAM and CD24 after undergoing EMT and exhibits 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 stem cells in human oral cancer specimens. Examining disseminating tumour cells in the stromal region of 3500 imaging fields from 24 human oral cancer specimens, evenly

divided into metastatic and non-metastatic specimens, we see a significant enrichment of EpCAM, CD24 and Vimentin co-stained cells in metastatic specimens. Through training an artificial neural network on the EpCAM, CD24 and Vimentin co-staining, we predict metastasis with high accuracy (F1 0.91; AUC 0.87). We have observed, for the first time, disseminating EMT stem cells in patient histological specimens and demonstrated their utility for predicting metastatic disease.

Introduction

In multiple types of carcinoma, cancer stem cells 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 cancer stem cells (Ocana et al., 2012; Tsai et al., 2012). We have previously demonstrated that a $CD44^{high}EpCAM^{low/-}$ EMT population can be separated from the main $CD44^{low}EpCAM^{high}$ epithelial population in flow cytometric analysis of oral squamous cell carcinoma (OSCC) cell lines and fresh tumour specimens (Biddle et al., 2016; Biddle et al., 2011). We identified retained cell surface expression of EpCAM (Biddle et al., 2011) and CD24 (Biddle et al., 2016) in a minority of cells that have undergone a full morphological EMT. Both EpCAM and CD24 were individually associated with enhanced ability to undergo MET, and we therefore reasoned that retention of one or both of these markers may identify an important population of tumour cells that have undergone EMT and disseminated from the primary tumour whilst retaining the plasticity required to enable subsequent MET and metastatic seeding. Here, we characterise the combined role of EpCAM and CD24 in marking a sub-population of tumour cells that has undergone EMT whilst retaining plasticity. Staining for EpCAM and CD24 alongside the mesenchymal marker Vimentin in 3500 imaging fields from 24 human tumour specimens, stratified on metastatic status, identifies cells that have undergone EMT and disseminated into the stromal region surrounding metastatic primary tumours. Using a machine learning approach, we show that the presence of disseminating EMT stem cells in the tumour stroma is predictive of metastasis. Finally, we identify a stem cell hierarchy within the population of cells that have undergone EMT, associated with a set of genes distinct from those governing epithelial/mesenchymal identity.

Results

EpCAM and CD24 in combination mark an EMT transition state that retains plasticity

In order to explore the transition states within the EMT population, we characterised the EpCAM and CD24 cell surface staining profile of the $CD44^{high}EpCAM^{low/-}$ EMT population in two OSCC cell lines and a panel of clonal sub-lines with varying levels of EMT and plasticity (Figure 1A). This showed that, whilst the majority of EMT cells are $EpCAM^{-}CD24^{-}$, a significant minority are split between

$EpCAM^+CD24^+$, $EpCAM^+CD24^-$ and $EpCAM^-CD24^+$ profiles (Figure 1B, C). The EMT population in the EMT-stem and EMT-restricted sub-lines displayed particularly marked differences in EpCAM and CD24 staining profile, with an average 28% of cells in the EMT-stem line being positive for at least one marker ($EpCAM^+CD24^+$, $EpCAM^+CD24^-$ and $EpCAM^-CD24^+$ profiles), compared to 2% in the EMT-restricted line (Figure 1C). Flow cytometric analysis of 6 fresh OSCC tumour specimens identified similar staining profiles, at a similar ratio, to those seen in cell lines (Figure 1D).

In order to test whether EMT cells staining for EpCAM and CD24 have enhanced ability to undergo MET, we FACS sorted the differentially stained sub-populations and cultured them separately for 7 days before re-analysing to determine the degree of MET in each culture. For the LUC4 cell line (Figure 2A) and EMT-stem sub-line (Figure 2B), the $EpCAM^+CD24^+$, $EpCAM^-CD24^+$ and $EpCAM^-CD24^-$ sub-populations were examined (in both these lines, the $EpCAM^+CD24^-$ sub-population was too small to be accurately sorted for re-culturing). In both lines, the $EpCAM^+CD24^+$ sub-population had the greatest ability to undergo MET, followed by the $EpCAM^-CD24^+$ sub-population, and finally the $EpCAM^-CD24^-$ sub-population which exhibited barely any MET. This was accompanied by a corresponding difference in ability to give rise to the other EMT states - the $EpCAM^+CD24^+$ sub-population could give rise to all the other EMT states, whereas the $EpCAM^-CD24^+$ sub-population was mostly limited to producing $EpCAM^-CD24^-$ cells, and the $EpCAM^-CD24^-$ sub-population could not give rise to any of the other EMT states. The CA1 cell line has a larger proportion of $EpCAM^+CD24^-$ cells than LUC4 or the EMT-stem sub-line, potentially allowing for these to be sorted too, but the very small size of the total EMT population in CA1 prevented accurate sorting for re-culturing. We therefore used the Epi-stem CA1 sub-line, as it has a larger EMT population than the parental CA1 line. We were able to accurately sort all 4 EMT sub-populations from the Epi-stem line, and re-analysed them for degree of MET after 7 days culture (Figure 2C). As with the other lines, the $EpCAM^+CD24^+$ sub-population had the greatest ability to undergo MET, followed by the $EpCAM^-CD24^+$ sub-population. The $EpCAM^+CD24^-$ and $EpCAM^-CD24^-$

sub-populations exhibited barely any MET. Compared to the other two lines, there did appear to be a greater degree of switching between the different EMT states in the Epi-stem line. Being a newly established EMT population from a recently clonally derived epithelial line, it is possible that the cellular hierarchies are still immature and thus more flexible than in other more established EMT populations.

These findings demonstrate that combined expression of EpCAM and CD24 marks EMT cells that possess enhanced ability to undergo MET and regenerate the epithelial tumour cell population, as required for metastasis. Interestingly, whilst cells expressing CD24 alone had some ability to undergo MET, those expressing EpCAM alone did not. This indicates a potential separation of the roles of CD24 and EpCAM, with CD24 being a marker of plasticity (and thus a putative EMT stem cell marker) and EpCAM being a retained epithelial marker. Notably, single cell cloning experiments showed no difference in self-renewal ability between the 4 EMT sub-populations (Supplementary Figure S1), demonstrating that the EMT stem cell hierarchy is related to plasticity and not self-renewal.

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

The retention of EpCAM expression in a sub-population of tumour cells that have undergone EMT raised the prospect that we may be able to identify these cells outside of the tumour body in human tumour specimens, as EpCAM is a specific epithelial marker that would not normally be found in the surrounding stromal region. In combination with EpCAM, we stained tumour specimens for CD24 as a marker of plasticity, and Vimentin as a mesenchymal marker to identify cells that have undergone EMT. Notably, CD44 cannot be used as an EMT marker in the context of intact tissue as it requires trypsin degradation in order to yield differential expression in EMT and epithelial populations (Biddle

et al., 2013; Mack and Gires, 2008). Vimentin, on the other hand, accurately distinguishes EMT from epithelial tumour cells in immunofluorescent staining protocols (Biddle et al., 2016). By combining EpCAM as a tumour lineage marker, Vimentin as a mesenchymal marker, and CD24 as a plasticity marker, we aimed to identify tumour cells that have undergone EMT and disseminated into the surrounding stromal region. For this, we developed a protocol for automated 4-colour (3 markers + nuclear stain) immunofluorescent imaging and analysis of entire histopathological slide specimens, to test for co-localisation of the 3 markers in each individual cell across each specimen.

To determine whether this marker combination identifies EMT stem cells, we initially tested the protocol on the CA1 cell line and its EMT-stem sub-line. $EpCAM^{+}Vim^{+}CD24^{+}$ cells were greatly enriched in the EMT-stem sub-line, comprising 41% of the population, compared to 2.1% in the CA1 line (Figure 3A, B, E). To test the specific role of EpCAM retention, we replaced EpCAM with a pan-keratin antibody against epithelial keratins. There was very little $Pan\text{-}keratin^{+}Vim^{+}CD24^{+}$ staining, and no enrichment for $Pan\text{-}keratin^{+}Vim^{+}CD24^{+}$ cells in the EMT-stem sub-line (Figure 3C, D, E). Therefore, whilst epithelial keratins are lost, EpCAM is retained in cells undergoing EMT and an $EpCAM^{+}Vim^{+}CD24^{+}$ staining profile can be used as a marker for EMT stem cells in immunofluorescent staining protocols.

Imaging the tumour body and adjacent stroma in sections of human OSCC specimens, we detected single cells co-expressing EpCAM, Vimentin and CD24 in the stromal region surrounding the tumour (Figure 3F), confirming that these cells can be detected in human tumour specimens. We next stratified 24 human primary OSCC specimens into 12 tumours that had evidence of lymph node metastasis or perineural spread, and 12 that remained metastasis free (Supplementary Figure S2), and stained them for EpCAM, Vimentin and CD24. Single cells co-expressing EpCAM, Vimentin and CD24 were abundant in the stroma surrounding metastatic tumours. This was not the case in non-metastatic

tumours or normal epithelial regions (Figure 4, A-C). In contrast to EpCAM, pan-keratin staining did not identify cells in the stroma surrounding metastatic tumours (Figure 4D).

We developed an image segmentation protocol that separated the tumour body from the adjacent stroma, thus enabling each nucleated cell to be assigned to either the tumour or stromal region in automated image analysis (Figure 4E). Expression of EpCAM, Vimentin and CD24 was then analysed for every nucleated cell in every imaging field that included both tumour and stroma (3500 manually curated imaging fields across the 24 tumours). This enabled the proportion of each cell type in each region to be quantified (Figure 4F). $EpCAM^{+}Vim^{+}CD24^{+}$ cells were enriched in the stroma compared to the tumour body, and there was a much greater accumulation of $EpCAM^{+}Vim^{+}CD24^{+}$ cells in the stroma of metastatic tumours compared to non-metastatic tumours. Interestingly, this was not the case for $EpCAM^{+}Vim^{+}CD24^{-}$ cells, which were also enriched in the stroma but showed no difference between metastatic and non-metastatic tumours. $Pan\text{-}keratin^{+}Vim^{+}CD24^{+}$ cells were not detected.

These findings demonstrate that an $EpCAM^{+}Vim^{+}CD24^{+}$ staining profile marks tumour cells disseminating into the surrounding stroma, and that these cells are enriched specifically in metastatic tumours. The presence of disseminating tumour cells that express EpCAM but not CD24 did not correlate with metastasis. This highlights a requirement for retained plasticity, marked by CD24, in disseminating metastatic stem cells.

Identification of $EpCAM^{+}CD24^{+}Vim^{+}$ cancer stem cells enables clinical prediction using a machine learning approach

OSCC are an important health burden and one of the top ten cancers worldwide, with over 300,000 cases annually and a 50% 5-year survival rate. There is frequent metastatic spread to the lymph nodes

of the neck; this is the single most important predictor of outcome and an important factor in treatment decisions (Sano and Myers, 2007). If spread to the lymph nodes is suspected, OSCC resection is accompanied by neck dissection to remove the draining lymph nodes, a procedure with significant morbidity. At presentation it is currently very difficult to determine which tumours are metastatic and this results in sub-optimal tailoring of treatment decisions. Accurate prediction of metastasis would therefore have great potential to improve clinical management of the disease to reduce both mortality and treatment-related morbidity. We sought to determine whether the $EpCAM^{+}CD24^{+}Vim^{+}$ staining pattern could be predictive of metastasis.

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

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

We validated the predictive utility of the EMT stem cell state by deriving transcriptional signatures from our genome-wide microarray gene expression data for the CA1 cell line and the EMT-stem and EMT-restricted CA1 sub-lines (GSE74578). We investigated the utility of these transcriptional signatures in distinguishing progressing from progression-free tumours from the TCGA OSCC cohort. The top 60 genes specifically upregulated in each cell line were plotted as a ratio of the mean values in progressing and progression-free tumours (Supplementary Figure S3). Restricting the analysis to only those genes that were upregulated in progressing tumours yielded 24/60 genes for CA1, 37/60 genes for EMT-restricted, and 42/60 genes for EMT-stem. These genes were then tested for their predictive ability. Using a Support Vector Machine supervised learning classifier, multi-gene predictive signatures were determined for CA1, EMT-stem and EMT-restricted (Figure 6A). For EMT-stem, four genes (CLCA2, TCP1, IL1A and TNC) combined to give an optimum AUC of 70%. For EMT-restricted, the optimum AUC (from 5 genes) was 60%. For CA1, the optimum AUC (from 2 genes) was 59%. Restricting the analysis to the top 24 most upregulated genes for each cell line yielded the same outcome. These results show that, compared to the CA1 and EMT-restricted transcriptional signatures, the EMT-stem transcriptional signature has superior ability to distinguish progressing from progression-free tumours from the TCGA OSCC cohort.

A stem cell hierarchy exists within the EMT population, distinct from the epithelial/mesenchymal spectrum

We have shown that retention of EpCAM, an epithelial marker, is not on its own sufficient to confer an EMT stem cell state. We have also demonstrated high predictive utility of the EMT stem cell state, contrasting with the reported weak association with clinical survival metrics of an intermediate position on an epithelial/mesenchymal spectrum (George et al., 2017). These observations are at odds with the previously proposed concept that an intermediate position on an epithelial/mesenchymal spectrum confers an EMT stem cell state (George et al., 2017; Kroger et al., 2019; Pastushenko et al., 2018), and we therefore investigated this concept in our genome-wide gene expression dataset (GSE74578). Hierarchical clustering analysis of all differentially expressed genes, visualised as a heat map (Figure 6B), showed 10 distinct gene clusters based on expression across the 3 cell lines (CA1, EMT-stem, and EMT-restricted). Only a minority of differentially expressed genes showed a spectrum of expression from CA1 > EMT-stem > EMT-restricted, or vice versa (cluster 2 and cluster 6, 125 genes, 15% of total). This suggests that a model of an epithelial/mesenchymal spectrum governing plasticity of cells that have undergone EMT is not supported in OSCC. Rather, the major pattern was one of a switch in gene expression between the epithelial CA1 line and the mesenchymal EMT-stem/EMT-restricted lines (cluster 4 and cluster 8, 316 genes, 39% of total), and a separate set of genes that were specifically altered in the EMT-stem line compared to the CA1 and EMT-restricted lines (cluster 1 and cluster 9, 138 genes, 17% of total). A further group of genes supported this pattern, whilst being unclear whether they contributed to the epithelial/mesenchymal switch set or the EMT-stem set (cluster 3 and cluster 10, 180 genes, 22% of total). Key biological processes associated with the epithelial/mesenchymal switch gene set were epithelial differentiation (up in CA1) and chemotaxis (up in EMT-stem and EMT-restricted). Key biological processes associated with the EMT-stem gene set were downregulation of apoptosis and survival of oxidative stress (up in EMT-stem) and interferon signalling (up in CA1 and EMT-restricted) (Figure 6B and Supplementary Figure S4).

To further test the distinction between the epithelial/mesenchymal and EMT-stem gene sets, we looked at clustering of the three cell lines on two sets of differentially expressed genes – (1) those differentially expressed between the CA1 and EMT-restricted lines, as an epithelial/mesenchymal switch gene set and (2) those differentially expressed between the EMT-stem and EMT-restricted lines, as a plasticity (EMT-stem) gene set (Figure 6C). On set (1), EMT-stem clustered with EMT-restricted and away from CA1, providing further evidence for the fully EMT status of the EMT-stem line. On set (2), CA1 clustered with EMT-restricted and away from EMT-stem, further supporting a role for a distinct EMT-stem state that is not an intermediate on an epithelial/mesenchymal spectrum. This supports a model whereby cells undergo an epithelial/mesenchymal gene expression switch when undergoing a morphological EMT, and then come under the influence of separate transcriptional pathways that govern plasticity and generate a stem cell hierarchy within the EMT population.

To specifically investigate changes in key regulators of the EMT program, we FACS sorted the 4 EMT sub-populations ($EpCAM^+CD24^+$, $EpCAM^+CD24^-$, $EpCAM^-CD24^+$ and $EpCAM^-CD24^-$) from the CA1 and LUC4 OSCC cell lines, and performed quantitative RT-PCR analysis compared to the epithelial population, for 8 genes (Supplementary Figure S5A) and 2 miRNAs (Supplementary Figure S5B) that govern epithelial/mesenchymal identity. The mesenchymal genes Vimentin, Zeb1, Prrx1, Twist and FSP1 were upregulated in all EMT sub-populations compared to the epithelial population in both cell lines, with little differences in expression between the EMT sub-populations. The epithelial gene E-cadherin and epithelial miRNA MiR200c were downregulated in all EMT sub-populations compared to the epithelial population in both cell lines, also with little differences in expression between the EMT sub-populations. The mesenchymal gene Snail and epithelial miRNA MiR34a did not show any consistent pattern of differential expression. Only the transcription factor Slug showed a consistent pattern of differential expression between EMT sub-populations, being highest in both the $EpCAM^+CD24^+$ and $EpCAM^-CD24^+$ EMT sub-populations. However, Slug expression was not increased

in EMT sub-populations compared to the epithelial population. The lack of difference between EMT sub-populations in expression of classical EMT regulators demonstrates that plasticity of cells that have undergone EMT in OSCC is not governed by a spectrum of epithelial/mesenchymal regulators.

Discussion

The role of EMT in tumour dissemination has long been debated but, lacking evidence of cells undergoing EMT whilst disseminating from human tumours *in vivo*, this role has had to be inferred from mouse models and human cell line models. Here, through developing our understanding of the transition states that emerge as cells undergo EMT in human OSCC, we have identified a plastic EMT stem cell state that disseminates from the primary tumour in human pathological specimens. Importantly, the presence of these disseminating stem cells is strongly correlated with tumour metastasis. Using a machine learning approach, we have demonstrated the ability to predict metastasis through staining for this EMT stem cell state.

We propose a malignant stem cell model that is governed by a hierarchical relationship within the EMT population, as opposed to an epithelial/mesenchymal (E/M) spectrum (Figure 7). We term this hierarchy the stem/restricted (S/R) spectrum, as the primary readout of position on this spectrum is the plasticity required to undergo MET to regenerate the epithelial population for metastatic outgrowth. The S/R spectrum is separate from the E/M spectrum, and all EMT transition states in OSCC have undergone a full transcriptional EMT. Nevertheless, some retain the epithelial marker EpCAM, enabling them to be tracked beyond the primary tumour. CD24, on the other hand, is not epithelial-specific so cannot be used alone to track tumour cells, but in combination with EpCAM can mark disseminating tumour cells that have retained stem cell plasticity. Cells can sit at multiple positions along the S/R spectrum – we show, for example, that $CD44^{high}EpCAM^{+}CD24^{+}$ cells have greater

plasticity than $CD44^{high}EpCAM^{-}CD24^{+}$ cells, which in turn have greater plasticity than $CD44^{high}EpCAM^{+}CD24^{-}$ and $CD44^{high}EpCAM^{-}CD24^{-}$ cells. Position on the S/R spectrum, and thus ability to regenerate the epithelial tumour cell population, determines the ability of these disseminating cells to contribute to tumour spread and is predictive of outcome.

A partial EMT state has previously been identified in OSCC, which was correlated with nodal metastasis and adverse pathological features (Puram et al., 2017). However, this study restricted its analysis to tumour cells that retained epithelial keratins. The retention of keratins and lack of upregulation of classical EMT markers suggest that the partial EMT state identified by Puram et al. exists within the epithelial population, and indeed was restricted to the cohesive epithelial region of human OSCC tumours. The correlation of this partial EMT state with nodal metastasis may be due to its association with increased transitioning into EMT within the epithelial population, thus enhancing metastasis. In our study, we find that restricting analysis to cells that retain epithelial keratins results in inferior metastasis prediction.

In contrast, we have now identified a bona fide EMT stem cell state. Immunofluorescent antibody co-staining for EpCAM, CD24 and Vimentin identifies these EMT stem cells in human tumour specimens and is predictive of metastasis. The ability to separate disseminating tumour cells from the stromal content of human tumours, which has confounded attempts to develop a predictive EMT signature (Tan et al., 2014), is one important factor in this success. However, we also show that non-plastic $EpCAM^{+}CD24^{-}Vim^{+}$ tumour cells in the stroma do not correlate with metastasis, and therefore the clinically predictive utility of $EpCAM^{+}Vim^{+}$ cells can be isolated specifically to the $CD24^{+}$ EMT stem cells that retain the ability to undergo MET. Notably, the accuracy of clinical prediction was much higher with immunofluorescent antibody co-staining of OSCC specimens (AUC 0.87) than with using an EMT-stem transcriptional signature to probe the TCGA dataset (AUC 0.7). This highlights the value of using

techniques that give single cell resolution, enabling isolation of the signal to the specific cell type of interest within a highly heterogeneous cellular environment.

Our approach has enabled us to reconstruct the cellular hierarchy in OSCC, and demonstrate that the EMT stem cell state is not governed by the epithelial/mesenchymal spectrum that has been proposed to confer stemness and metastatic proclivity in previous studies. The majority of these studies looked at differences between, rather than within, cell lines (George et al., 2017; Huang et al., 2013). Cell lines with a more 'intermediate' character were more aggressive, but heterogeneity within lines was not assessed. These findings are compatible with the EMT stem cell hierarchy model proposed here, as a larger EMT-stem sub-population would result in greater epithelial/mesenchymal plasticity and therefore a more intermediate phenotype overall. Other studies have focussed on immortalised or experimentally transformed breast cell models, which may behave differently to *in vivo* tumours (Kroger et al., 2019; Zhang et al., 2014). Finally, a study of experimentally induced mouse cutaneous SCCs identified an intermediate state, but this was within a highly mesenchymal tumour model that did not depend on epithelial/mesenchymal plasticity for tumour dissemination (Pastushenko et al., 2018). This contrasts starkly with the highly epithelial character of human SCCs (Biddle et al., 2016; Toll et al., 2013), and is therefore of uncertain relevance to human tumours. An important strength of our study has been the ability to look at the single cell level in human tumours and spontaneously derived human cancer cell lines, enabling us to observe human tumour cells disseminating into the surrounding tissue and in the process identify a novel stem cell hierarchy within the EMT population.

Conflict of interest

The authors declare no conflicts of interest.

Acknowledgements

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

Methods

Cell culture

The CA1 and LUC4 cell lines were both previously derived in our laboratory, from biopsies of OSCC of the floor of the mouth. Cell culture was performed as previously described (Biddle et al., 2011). Cell removal from adherent culture was performed using 1x Trypsin-EDTA (Sigma, T3924) at 37ºC.

Flow cytometry and fluorescence-activated cell sorting (FACS)

Flow cytometry was performed as previously described (Biddle et al., 2011). Antibodies for cell line staining were CD44-PE (clone G44-26, BD Bioscience), CD24-FITC (clone ML5, BD Bioscience) and EpCAM-APC (clone HEA-125, Miltenyi Biotec). For fresh tumour cells, β4-integrin-PE (clone 439-9B, BD Bioscience) was added and CD44-PE was replaced with CD44-PerCP/Cy5.5 (clone G44-26, BD Bioscience). Single stained controls were performed for compensation, and negative controls were performed to set negative gating. Isolation of cells from human tumours for flow cytometry was performed as previously described (Biddle et al., 2016). Single cell plating for the self-renewal assay was performed using the single cell sorting function on the FACS Aria (Becton Dickinson). Flow cytometry data was analysed using FlowJo software, and statistical testing performed using 2-way ANOVA with multiple comparisons in GraphPad Prism software.

Immunofluorescent staining of cell lines and tumour tissue sections

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

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

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

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

Imaging of the stained slides was performed using the In Cell Analyzer 2200 (GE), a high content automated fluorescence microscope with four-colour imaging capability. The slides were imaged at x20 and x40 magnification. An image segmentation protocol was developed to extract grey level intensities corresponding to EpCAM, Vimentin and CD24 expression for every DAPI stained nucleated cell in the tumour body and the adjacent stroma separately. Segmentation was performed using the Developer Toolbook software (GE). As shown in figure 4E, an 'EpCAM dense cloud' was generated to isolate individual nucleated cells in the tumour body from the adjacent stroma and analyse them separately.

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

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

Machine learning for prognostic prediction using immunofluorescent staining data

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

The dataset was stratified into a training and validation cohort in a 70%:30% ratio using a random seed split. Supervised machine learning approaches were implemented using the skikit-learn Python 3.6 libraries (Pedregosa et al., 2011) and Tensorflow/Keras framework (https://www.tensorflow.org/api_docs/python/tf/keras/models). Hyper-parameter optimisation was performed by an exhaustive grid search and computed on Apocrita, a high performance cluster (HPC) facility at Queen Mary University of London (http://doi.org/10.5281/zenodo.438045). To further minimise overfitting, 10-fold cross-validation was performed and the mean accuracy metric, F1 score, was obtained for each learning iteration. Receiver-of-operator (ROC) curves and the area-under the-curve (AUC) were computed for the optimum supervised learning algorithm. Supervised approaches used were logistic regression, support vector machines (Smola and Scholkopf, 2004), Naïve Bayes (Zhang, 2005), K-Nearest Neighbours (Bentley, 1975), decision trees (Dumont et al., 2009), and artificial neural networks (Rumelhart et al., 1986).

RNA extraction, cDNA synthesis, and quantitative PCR

RNA extraction, cDNA synthesis and quantitative PCR were performed as previously described (Biddle et al., 2011), with the exception that the miRNeasy micro kit (Qiagen) was used for RNA extraction in order to preserve both mRNAs and miRNAs and, for miRNAs, cDNA synthesis and quantitative PCR were performed using the miScript PCR system (Qiagen). Primer sequences and miRNA assay codes are listed in the supplementary information. The ΔΔCt method was used to calculate fold-change in expression compared to the epithelial population, with GAPDH as a reference gene for mRNA analysis and RNU6-2 as a reference snRNA for miRNA analysis. Statistical testing was performed using paired t-tests.

Differential expression, hierarchical clustering and functional analysis of gene expression microarray data

Differentially expressed genes were filtered at $p<0.05$ in the Genome Studio software (Ilumina), using the Illumina error model and FDR multiple testing correction, with quantile normalisation. All genes differentially expressed between any two cell lines were included. Hierarchical clustering of these differentially expressed genes across the three cell lines was performed using ClustVis (Metsalu and Vilo, 2015) and Morpheus (Morpheus, https://software.broadinstitute.org/morpheus) online tools. Gene ontology analysis and functional clustering for each expression cluster was performed using the STRING online tool (String, https://string-db.org/).

Machine learning for prognostic prediction using gene expression microarray data

Normalised count data (FPKM-UQ) for OSCC RNA-seq were downloaded from the GDC portal for TCGA cases with available survival data (535 cases in total). The clinical data of each patient was further mined to only include cases which were either alive, metastasis-free and progression-free (denoted as 0) or dead, developed distant metastasis and the tumour progressed (denoted as 1). A

balanced cohort was obtained by including 100 cases from the progression-free group and 100 cases from the progressing group.

For the set of differentially expressed genes ($p<0.05$) from our gene expression microarray data, we used the average probe intensity for each individual gene (i) for CA1 (C), EMT-restricted (R) an EMT-stem (S) to produce the following ratios:

C_i/R_i and C_i/S_i for CA1

R_i/C_i and R_i/S_i for the EMT-restricted sub-line

S_i/C_i and S_i/R_i for the EMT-stem sub-line

We used these ratios to obtain a subset of genes that were uniquely upregulated in each cell line. For example, for EMT-stem, we took all genes that fulfilled this condition: $S_i/ C_i > 1.5$ and $S_i/R_i > 1.5$. We then ordered the list of genes to get the top 60 most upregulated for each cell line. These were used to interrogate our TCGA OSCC RNA-seq cohort.

We first determined whether each upregulated gene from our gene expression microarray dataset was congruent with upregulation in progressing tumours. The congruence of expression was determined by obtaining the mean FPMK-UQ values of each upregulated gene in the progression-free and progressing cases from the OSCC TCGA cohort. A mean ratio was calculated, and a ratio greater than 1 signified upregulation in progressing tumours. The results were visualised in a box plot using SPSS (version 25).

A supervised learning approach was used to identify the combination of genes from this final restricted gene list for each cell line (genes uniquely upregulated in the cell line AND upregulated in

progressing TCGA tumours) that could optimally classify the OSCC tumours into progression-free and progressing tumours using their FPKM-UQ values as inputs. Supervised learning was performed using Python's scikit-learn and the data was split and cross-validation of F1 and AUC scores performed in the same way as described in the immunofluorescence machine learning section. To obtain an optimum Support Vector Machine (SVM-RBF) based classifier an additive approach was taken to the cross-validated AUC score. The cross-validated AUC for each gene was calculated individually. The gene with the highest AUC was included and then the AUC for every other gene in the dataset was calculated to determine which second gene resulted in the highest increase of AUC. This was then repeated until no further increase of the AUC was possible. Note, this process was performed separately for the set of upregulated genes for each cell line. The results of the optimum classifier was depicted using an ROC and an adaptation of a Forest plots showing AUC instead of hazard ratios.

Data availability

The raw gene expression microarray data is available at https://www.ncbi.nlm.nih.gov/geo/ under accession code GSE74578.

References

Bentley, J. L. (1975). Multidimensional Binary Search Trees Used for Associative Searching. Commun Acm *18*, 509-517.

Biddle, A., Gammon, L., Fazil, B., and Mackenzie, I. C. (2013). CD44 staining of cancer stem-like cells is influenced by down-regulation of CD44 variant isoforms and up-regulation of the standard CD44 isoform in the population of cells that have undergone epithelial-to-mesenchymal transition. PloS one *8*, e57314.

Biddle, A., Gammon, L., Liang, X., Costea, D. E., and Mackenzie, I. C. (2016). Phenotypic Plasticity Determines Cancer Stem Cell Therapeutic Resistance in Oral Squamous Cell Carcinoma. EBioMedicine *4*, 138-145.

Biddle, A., Liang, X., Gammon, L., Fazil, B., Harper, L. J., Emich, H., Costea, D. E., and Mackenzie, I. C. (2011). Cancer stem cells in squamous cell carcinoma switch between two distinct phenotypes that are preferentially migratory or proliferative. Cancer Res *71*, 5317-5326.

Bill, R., and Christofori, G. (2015). The relevance of EMT in breast cancer metastasis: Correlation or causality? FEBS Lett *589*, 1577-1587.

Bronsert, P., Enderle-Ammour, K., Bader, M., Timme, S., Kuehs, M., Csanadi, A., Kayser, G., Kohler, I., Bausch, D., Hoeppner, J.*, et al.* (2014). Cancer cell invasion and EMT marker expression: a three-dimensional study of the human cancer-host interface. J Pathol *234*, 410-422.

Bur, A. M., Holcomb, A., Goodwin, S., Woodroof, J., Karadaghy, O., Shnayder, Y., Kakarala, K., Brant, J., and Shew, M. (2019). Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma. Oral Oncol *92*, 20-25.

Dumont, M., Maree, R., Wehenkel, L., and Geurts, P. (2009). Fast Multi-Class Image Annotation with Random Subwindows and Multiple Output Randomized Trees. Visapp 2009: Proceedings of the Fourth International Conference on Computer Vision Theory and Applications, Vol 2, 196-+.

George, J. T., Jolly, M. K., Xu, S., Somarelli, J. A., and Levine, H. (2017). Survival Outcomes in Cancer Patients Predicted by a Partial EMT Gene Expression Scoring Metric. Cancer Res *77*, 6415-6428.

Huang, R. Y., Wong, M. K., Tan, T. Z., Kuay, K. T., Ng, A. H., Chung, V. Y., Chu, Y. S., Matsumura, N., Lai, H. C., Lee, Y. F.*, et al.* (2013). An EMT spectrum defines an anoikis-resistant and spheroidogenic intermediate mesenchymal state that is sensitive to e-cadherin restoration by a src-kinase inhibitor, saracatinib (AZD0530). Cell Death Dis *4*, e915.

Jensen, D. H., Dabelsteen, E., Specht, L., Fiehn, A. M., Therkildsen, M. H., Jønson, L., Vikesaa, J., Nielsen, F. C., and von Buchwald, C. (2015). Molecular profiling of tumour budding implicates TGFβ-mediated epithelial-mesenchymal transition as a therapeutic target in oral squamous cell carcinoma. J Pathol *236*, 505-516.

Kroger, C., Afeyan, A., Mraz, J., Eaton, E. N., Reinhardt, F., Khodor, Y. L., Thiru, P., Bierie, B., Ye, X., Burge, C. B., and Weinberg, R. A. (2019). Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. Proc Natl Acad Sci U S A *116*, 7353-7362.

Lawson, D. A., Bhakta, N. R., Kessenbrock, K., Prummel, K. D., Yu, Y., Takai, K., Zhou, A., Eyob, H., Balakrishnan, S., Wang, C. Y., *et al.* (2015). Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature *526*, 131-135.

Li, W., and Kang, Y. (2016). Probing the Fifty Shades of EMT in Metastasis. Trends Cancer *2*, 65-67.

Liu, S., Cong, Y., Wang, D., Sun, Y., Deng, L., Liu, Y., Martin-Trevino, R., Shang, L., McDermott, S. P., Landis, M. D., *et al.* (2014). Breast Cancer Stem Cells Transition between Epithelial and Mesenchymal States Reflective of their Normal Counterparts. Stem cell reports *2*, 78-91.

Mack, B., and Gires, O. (2008). CD44s and CD44v6 expression in head and neck epithelia. PloS one *3*, e3360.

Metsalu, T., and Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res *43*, W566-570.

Ocana, O. H., Corcoles, R., Fabra, A., Moreno-Bueno, G., Acloque, H., Vega, S., Barrallo-Gimeno, A., Cano, A., and Nieto, M. A. (2012). Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1. Cancer cell *22*, 709-724.

Pastushenko, I., Brisebarre, A., Sifrim, A., Fioramonti, M., Revenco, T., Boumahdi, S., Van Keymeulen, A., Brown, D., Moers, V., Lemaire, S., *et al.* (2018). Identification of the tumour transition states occurring during EMT. Nature *556*, 463-468.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.* (2011). Scikit-learn: Machine Learning in Python. J Mach Learn Res *12*, 2825-2830.

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C. L., Mroz, E. A., Emerick, K. S., *et al.* (2017). Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell *171*, 1611-1624 e1624.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. Nature *323*, 533-536.

Ruscetti, M., Dadashian, E. L., Guo, W., Quach, B., Mulholland, D. J., Park, J. W., Tran, L. M., Kobayashi, N., Bianchi-Frias, D., Xing, Y., *et al.* (2016). HDAC inhibition impedes epithelial-mesenchymal plasticity and suppresses metastatic, castration-resistant prostate cancer. Oncogene *35*, 3781-3795.

Sano, D., and Myers, J. N. (2007). Metastasis of squamous cell carcinoma of the oral tongue. Cancer metastasis reviews *26*, 645-662.

Smola, A. J., and Scholkopf, B. (2004). A tutorial on support vector regression. Stat Comput *14*, 199-222.

Takamatsu, M., Yamamoto, N., Kawachi, H., Chino, A., Saito, S., Ueno, M., Ishikawa, Y., Takazawa, Y., and Takeuchi, K. (2019). Prediction of early colorectal cancer metastasis by machine learning using digital slide images. Comput Methods Programs Biomed *178*, 155-161.

Tan, T. Z., Miow, Q. H., Miki, Y., Noda, T., Mori, S., Huang, R. Y., and Thiery, J. P. (2014). Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. EMBO Mol Med *6*, 1279-1293.

Toll, A., Masferrer, E., Hernandez-Ruiz, M. E., Ferrandiz-Pulido, C., Yebenes, M., Jaka, A., Tuneu, A., Jucgla, A., Gimeno, J., Baro, T.*, et al.* (2013). Epithelial to mesenchymal transition markers are associated with an increased metastatic risk in primary cutaneous squamous cell carcinomas but are attenuated in lymph node metastases. J Dermatol Sci *72*, 93-102.

Tsai, J. H., Donaher, J. L., Murphy, D. A., Chau, S., and Yang, J. (2012). Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. Cancer cell *22*, 725-736.

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). Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. Int J Med Inform *128*, 79-86.

Williams, E. D., Gao, D., Redfern, A., and Thompson, E. W. (2019). Controversies around epithelial-mesenchymal plasticity in cancer metastasis. Nat Rev Cancer *19*, 716-732.

Zhang, H. (2005). Exploring conditions for the optimality of Naive bayes. Int J Pattern Recogn *19*, 183-198.

Zhang, J., Tian, X. J., Zhang, H., Teng, Y., Li, R., Bai, F., Elankumaran, S., and Xing, J. (2014). TGF-beta-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. Sci Signal *7*, ra91.

Figure 1

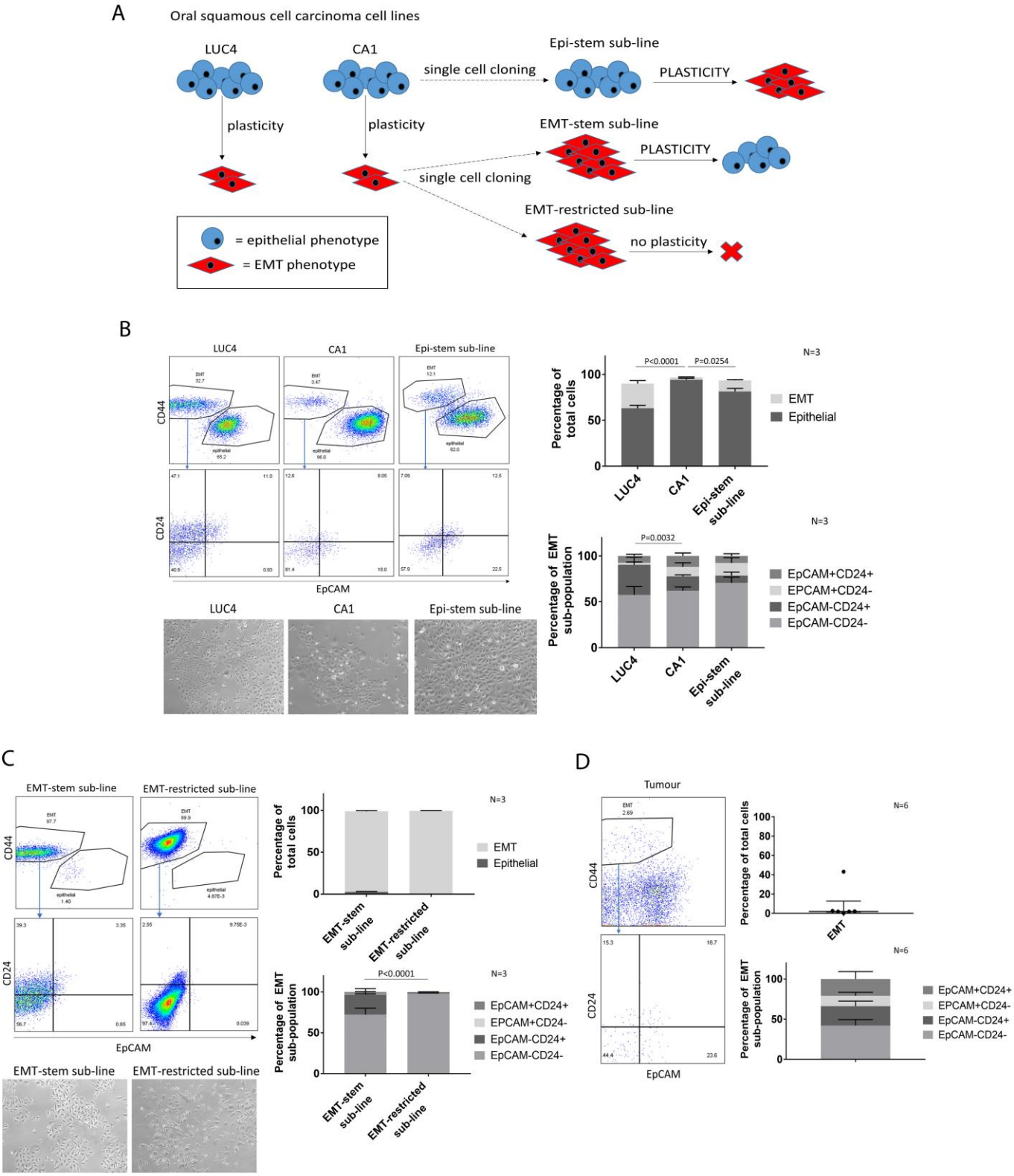


Figure 1 – EpCAM and CD24 in combination mark heterogeneous transition states within the EMT population. **A,** A schematic representation of the properties of the 5 cell lines used in this study (LUC4, CA1, Epi-stem sub-line, EMT-stem sub-line, EMT-restricted sub-line), showing their balance of epithelial and EMT populations and the derivation of the 3 CA1 sub-lines. The Epi-stem and EMT-stem sub-lines possess greater epithelial-mesenchymal plasticity than the CA1 cell line from which they were derived, denoted by capitalisation. **B, C,** Flow cytometric analysis of the epithelial phenotype cell lines (LUC4, CA1, and the Epi-stem CA1 sub-line) (B), and the EMT phenotype cell lines (EMT-stem and EMT-restricted CA1 sub-lines) (C), with accompanying brightfield images of cells in culture. The graphs show mean +/- SEM of the EMT ($CD44^{high}EpCAM^{low/-}$) and epithelial ($CD44^{low}EpCAM^{high}$) populations as a percentage of total cells (top) and the 4 EMT sub-populations (derived by separating the EMT population on both EpCAM and CD24 expression) as a percentage of the EMT population (bottom), with representative CD44/EpCAM and CD24/EpCAM flow cytometry plots for each cell line to the left. **D,** Flow cytometric analysis of 6 fresh oral squamous cell carcinoma specimens after enzymatic disaggregation. The top graph shows the EMT population as a percentage of total cells, with individual data points and median +/- IQR displayed due to the non-normal distribution of the data. The bottom graph shows mean +/- SEM of the 4 EMT sub-populations as a percentage of the EMT population. Representative CD44/EpCAM and CD24/EpCAM flow cytometry plots for one tumour are to the left. For each experiment, the number of biological repeats is indicated next to the graph. P-values were calculated using two-way ANOVA, and are displayed where the comparison showed a statistically significant difference.

Figure 2

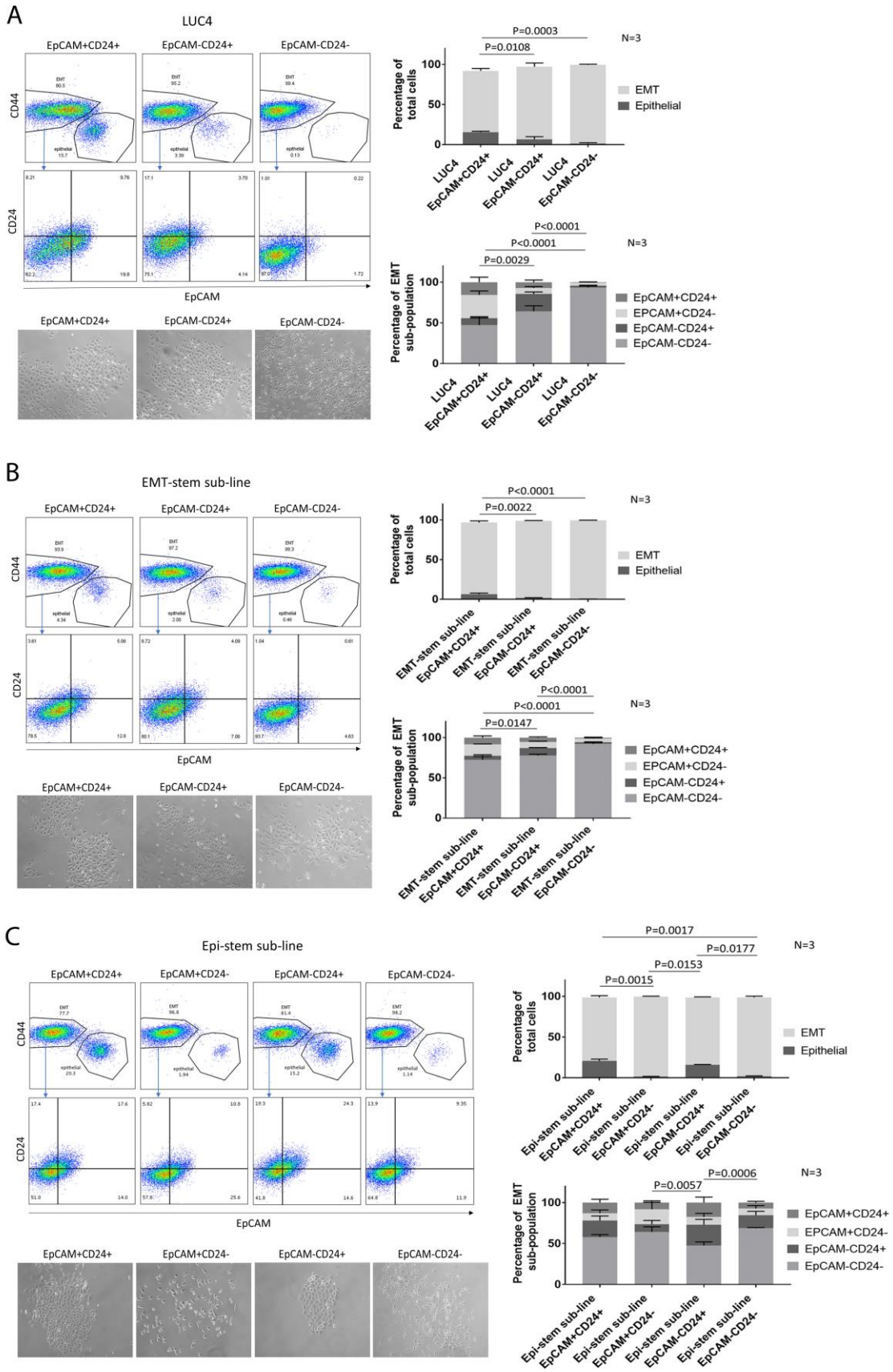


Figure 2 – A hierarchy of EMT sub-populations exists based on ability to re-population the epithelial tumour population. Flow cytometric analysis 7 days after sorting and re-plating the EMT sub-populations from LUC4 (A), the EMT-stem CA1 sub-line (B), and the Epi-stem CA1 sub-line (C), with accompanying brightfield images of the sorted sub-populations in culture. The graphs show mean +/- SEM of the EMT ($CD44^{high}EpCAM^{low/-}$) and epithelial ($CD44^{low}EpCAM^{high}$) populations as a percentage of total cells for each sorted sub-population (top) and the 4 EMT sub-populations (derived by separating the EMT population on both EpCAM and CD24 expression) as a percentage of the EMT population (bottom), with representative CD44/EpCAM and CD24/EpCAM flow cytometry plots for each sorted sub-population to the left. Note that the $EpCAM^{+}CD24^{-}$ sub-population could not be sorted from LUC4 or the EMT-stem sub-line, due to its small size in these lines. For each experiment, the number of biological repeats is indicated next to the graph. P-values were calculated using two-way ANOVA, and are displayed where the comparison showed a statistically significant difference.

Figure 3

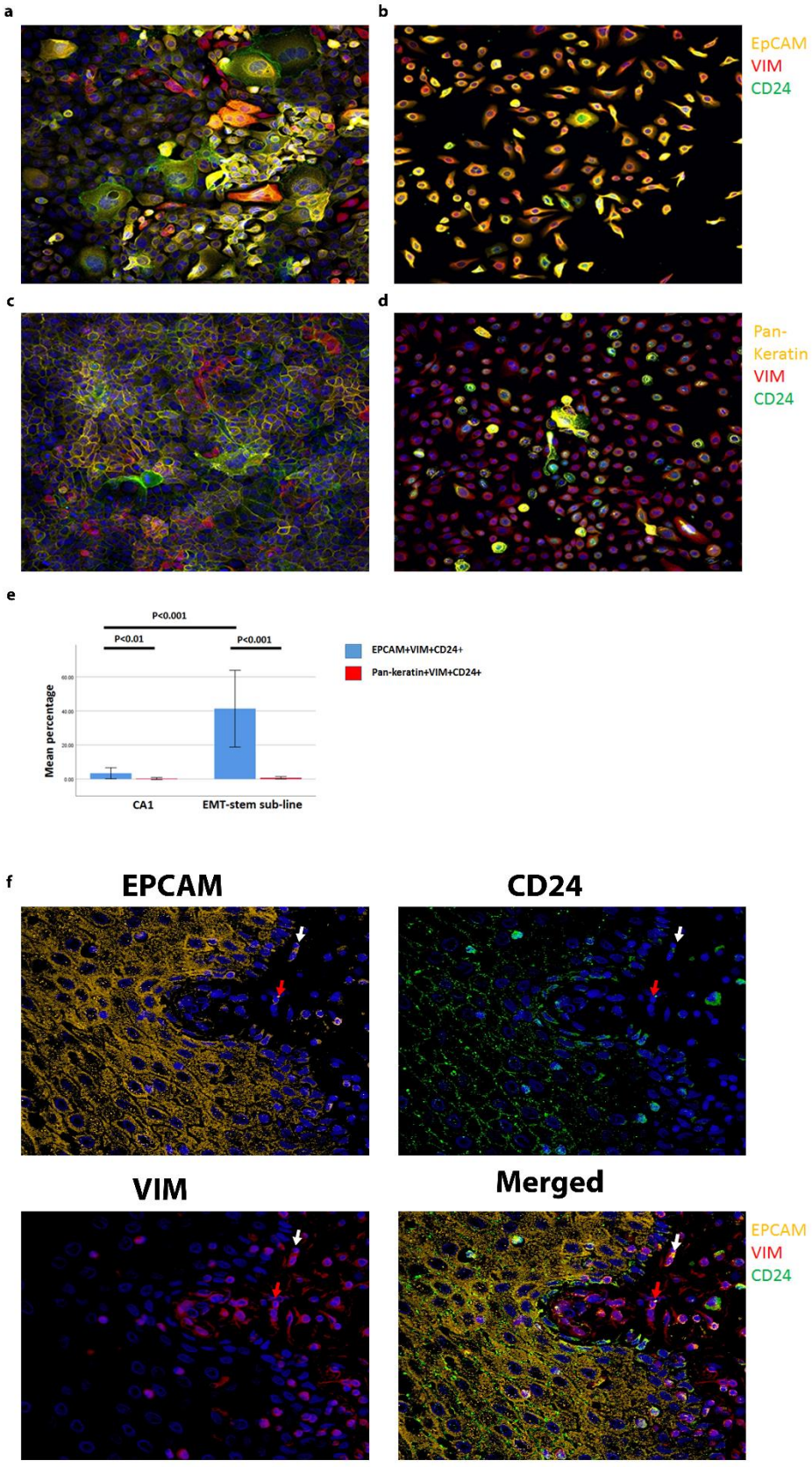


Figure 3 – Immunofluorescent co-staining for EpCAM, Vimentin and CD24 identifies the EMT stem cell state. **A-D,** Immunofluorescent staining for EpCAM, Vimentin and CD24 (A, B) and pan-keratin, Vimentin and CD24 (C, D) in the CA1 cell line (A, C) and the EMT-stem CA1 sub-line (B, D). **E,** Quantification of the percentage of $EpCAM^{+}Vim^{+}CD24^{+}$ and $pan\text{-}keratin^{+}Vim^{+}CD24^{+}$ cells in the CA1 cell line and EMT-stem sub-line. Significance is obtained from a two-tailed student t-test. The graph shows mean +/- 95% confidence interval. **F,** Detection of $EpCAM^{+}Vim^{+}CD24^{+}$ cells in the stroma surrounding an oral cancer tumour specimen. The white arrow highlights an $EpCAM^{+}Vim^{+}CD24^{+}$ cell in the stroma. The red arrow highlights an $EpCAM^{+}Vim^{+}CD24^{-}$ cell in the stroma. DAPI nuclear stain is blue.

Figure 4

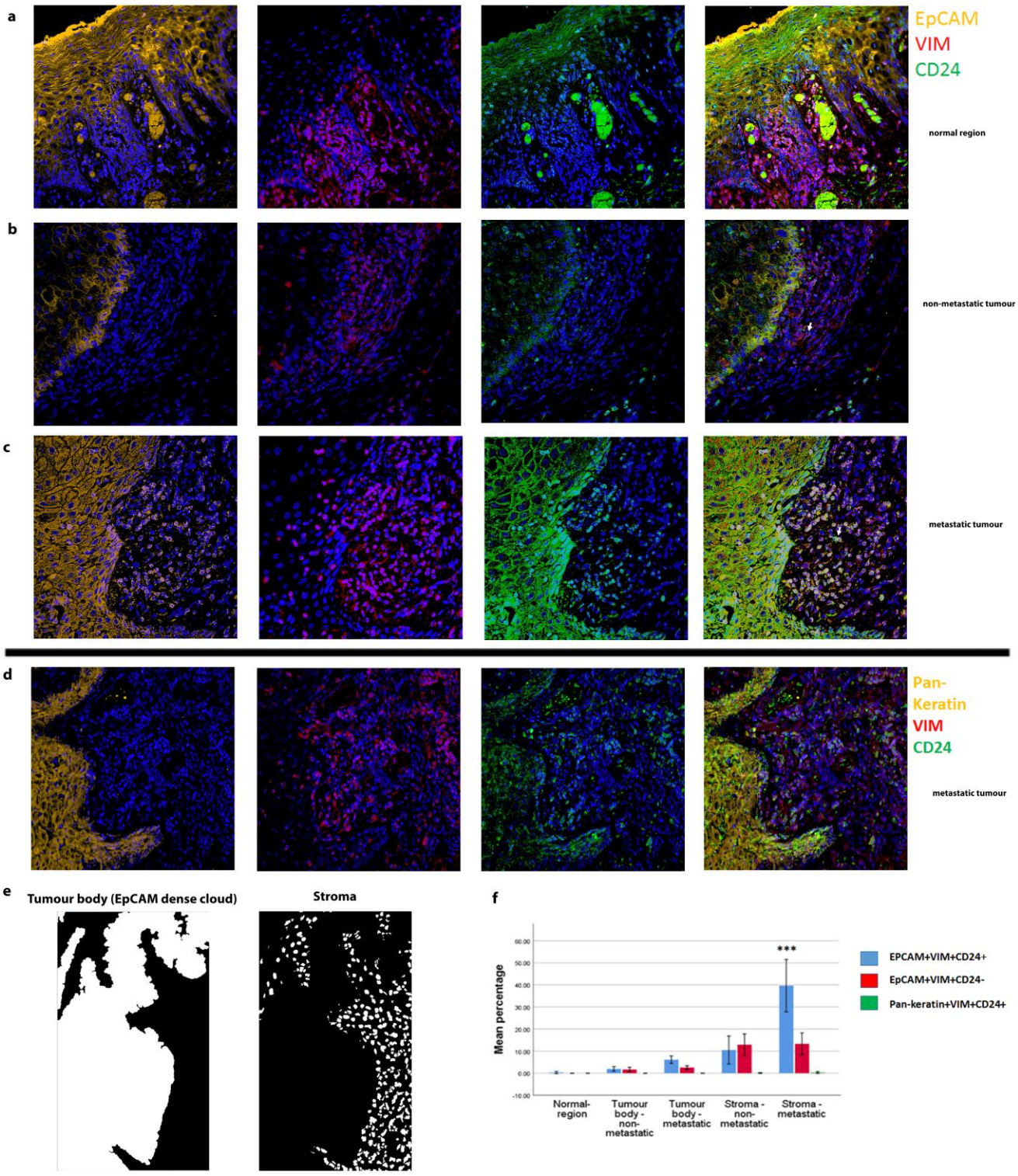
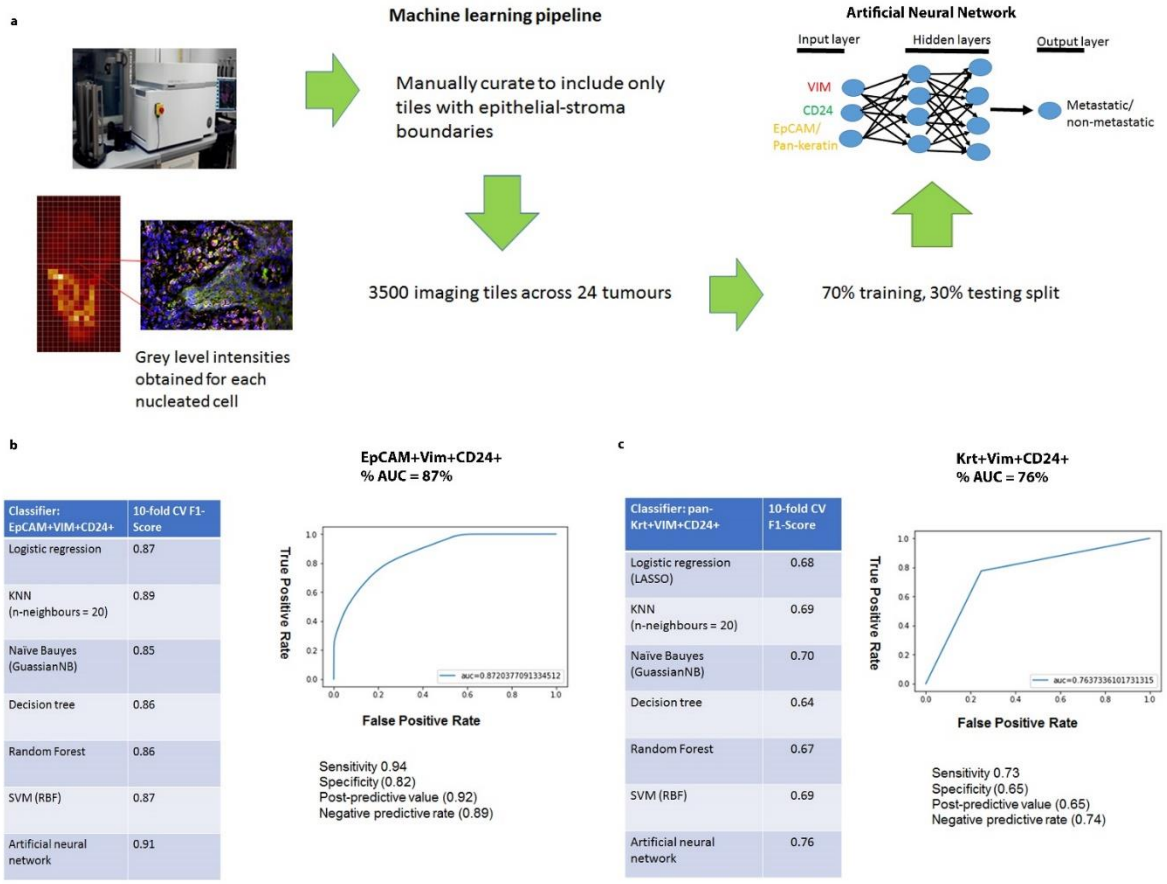


Figure 4 – Enrichment of $EpCAM^{+}Vim^{+}CD24^{+}$ cells in the stroma surrounding metastatic tumours. **A-C,** Immunofluorescent four-colour staining of oral tumour specimens for EpCAM (yellow), Vimentin (red) and CD24 (green) with DAPI nuclear stain (blue). Representative imaging fields from a normal epithelial region (A), a non-metastatic tumour (B) and a metastatic tumour (C). **D,** Staining of a metastatic tumour for pan-keratin, Vimentin and CD24. **E,** Image segmentation was performed, with generation of an 'EpCAM dense cloud' to distinguish the tumour body from the stroma. Grey level intensities for EpCAM, Vimentin and CD24 were obtained for every nucleated cell in each imaging field. **F,** Quantification of the percentage of $EpCAM^{+}Vim^{+}CD24^{+}$, $EpCAM^{+}Vim^{+}CD24^{-}$ and pan-$keratin^{+}Vim^{+}CD24^{+}$ cells in normal region (epithelium distant from the tumour), tumour body, and stromal region from metastatic and non-metastatic tumours. A student t-test was performed comparing the mean percentage of $EpCAM^{+}Vim^{+}CD24^{+}$ co-expressing cells in the metastatic stroma compared to the other fractions. *** signifies $p < 0.001$. The graph shows mean +/- 95% confidence interval.

Figure 5



Classifier: EpCAM+VIM+CD24+	10-fold CV F1-Score
Logistic regression	0.87
KNN (n-neighbours = 20)	0.89
Naïve Bauyes (GuassianNB)	0.85
Decision tree	0.86
Random Forest	0.86
SVM (RBF)	0.87
Artificial neural network	0.91

Classifier: pan-Krt+VIM+CD24+	10-fold CV F1-Score
Logistic regression (LASSO)	0.68
KNN (n-neighbours = 20)	0.69
Naïve Bauyes (GuassianNB)	0.70
Decision tree	0.64
Random Forest	0.67
SVM (RBF)	0.69
Artificial neural network	0.76

Figure 5 – Predicting metastasis using EpCAM, Vimentin and CD24 immunofluorescent staining and a supervised machine learning approach. **A,** Pipeline for machine learning based on grey level intensities for the three markers. The training tiles were classified as coming from a metastatic 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. 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.

Figure 6

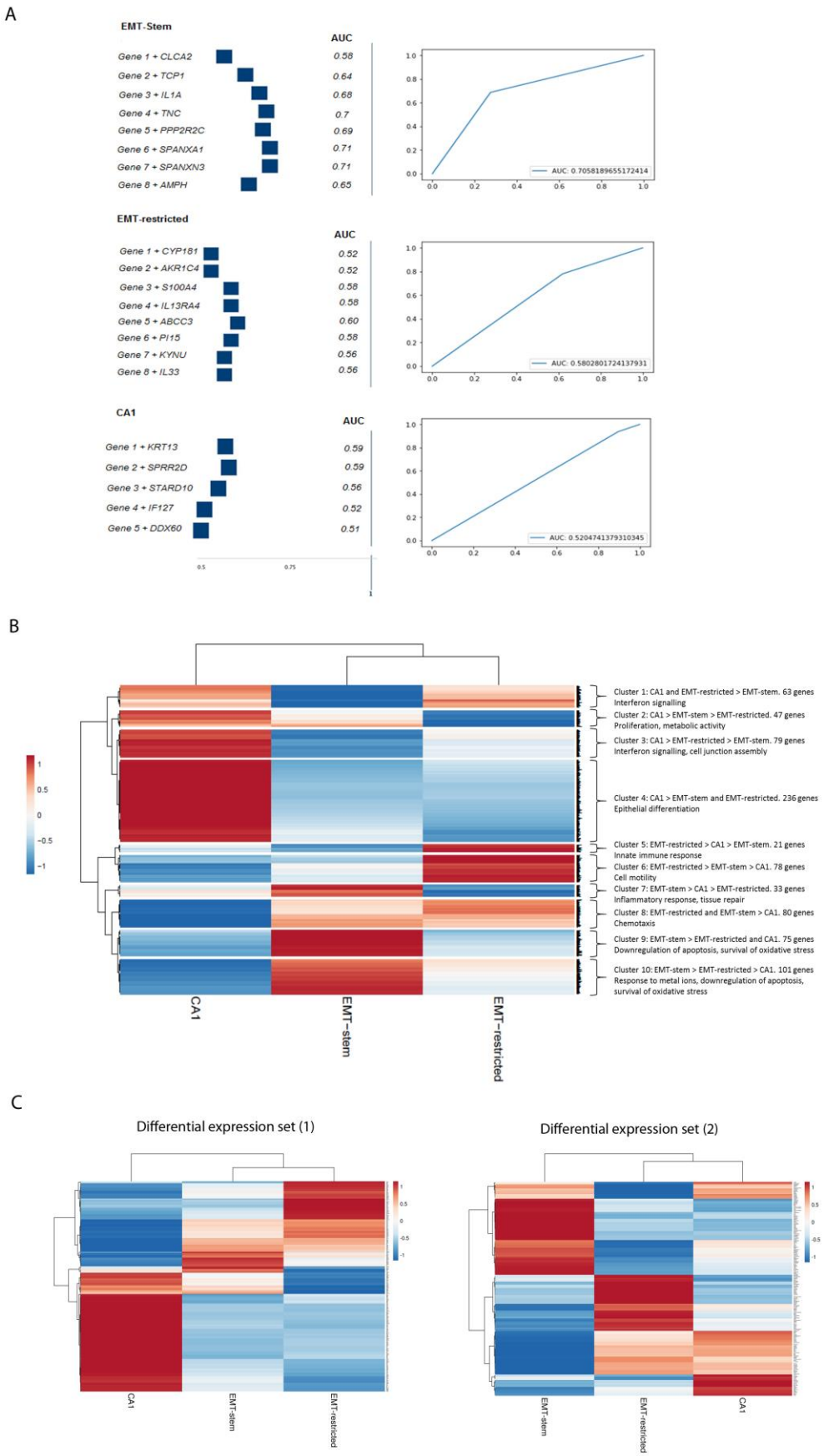


Figure 6 – An EMT-stem transcriptional signature is predictive of progressive disease and does not conform to an epithelial/mesenchymal spectrum model. Analysis of genome wide microarray gene expression data for the CA1 cell line and the EMT-stem and EMT-restricted CA1 sub-lines (GSE74578). **A,** Adapted Forest plots showing cumulative AUC values for the upregulated genes from each cell line that have the highest predictive ability against the TCGA dataset, highlighting the changes to the AUC value as the SVM classifier is built in an additive way. To the right is an ROC curve for the cumulative predictive ability of the top four genes in each analysis. **B,** Unsupervised hierarchical clustering of all genes that are differentially expressed between at least two of the lines ($p < 0.05$), displayed as a heatmap. 10 gene expression clusters are visualised, and annotated to the right. Biological processes associated with each cluster are included in the annotation. See supplementary figure S4 for further gene ontology analysis and functional clustering for each expression cluster. **C,** Unsupervised hierarchical clustering on the same dataset as in B, but with more restricted sets of differentially expressed genes: (1) those differentially expressed between CA1 and EMT-restricted (heatmap to left, $p < 0.05$), and (2) those differentially expressed between EMT-stem and EMT-restricted (heatmap to right, $p < 0.05$).

Figure 7

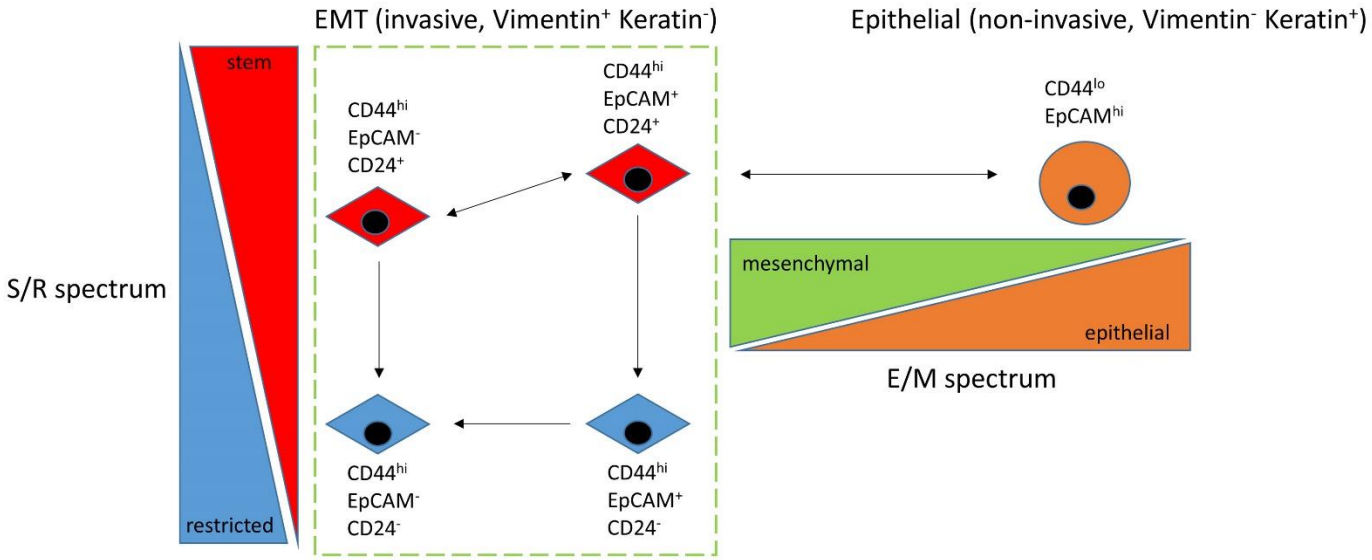


Figure 7 – Proposed model of the EMT sub-population hierarchy in oral cancer. Rather than being governed by an epithelial/mesenchymal (E/M) spectrum, plasticity within the population of tumour cells that have undergone a morphological EMT is governed by a separate stem/restricted (S/R) spectrum. The E/M spectrum and S/R spectrum have distinct transcriptional signatures. Position on the S/R spectrum, and thus ability to regenerate the epithelial tumour cell population, determines the ability of these invasive cells to contribute to tumour spread and is predictive of outcome. Cell surface marker profiles for each identified cell type are indicated, and arrows indicate ability to transition between cell types (uni-directional or bi-directional, as depicted by arrow heads). A green dashed box indicates all the cell sub-populations that exhibit a morphological and transcriptional EMT. A partial EMT, as previously described (Puram et al., 2017), is likely to exist within the cohesive epithelial population. The E/M spectrum is positioned to reflect this.