Original article

A neural network model delivers a highly prognostic protein signature in cancer stem cells that identifies relapse in stage III colorectal cancer patients.

Anna Sturrock[1,2], Sanghee Cho[3], Manuela Salvucci[1,2], Marc Sturrock[1,2], Joanna Fay[4], Tony O'Grady[4], Elizabeth McDonough[3], Christine Surrette[3], Jinru Shia[5], Canan Firat[5], Nil Urganci[5], Batuhan Kisakol[1,2], Emer P. O'Connell[1,2,6], John P. Burke[6,8], Niamh M. McCawley[6,8], Deborah A. McNamara[6,8], John F. Graf[3], Simon S. McDade[7], Mohammadreza Azimi[1,2], Daniel B. Longley[7], Fiona Ginty[3] and Jochen H. M. Prehn[1,2*].

[1] Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland University of Medicine and Health Sciences, 123 St. Stephen's Green, Dublin 2, Ireland.

[2] Centre of Systems Medicine, Royal College of Surgeons in Ireland University of Medicine and Health Sciences, 123 St. Stephen's Green, Dublin 2, Ireland.

[3] GE HealthCare Technology and Innovation Center, Niskayuna, NY, 12309, USA (formerly GE Research).

[4] Department of Pathology, Royal College of Surgeons in Ireland University of Medicine and Health Sciences, Beaumont Hospital, Beaumont Road, Dublin 9, Ireland.

[5] Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

[6] Department of Surgery, Royal College of Surgeons in Ireland University of Medicine and Health Sciences, 123 St. Stephen's Green, Dublin 2, Ireland.

[7] Centre for Cancer Research & Cell Biology, Queen's University Belfast, 97 Lisburn Road, Belfast BT9 7AE Northern Ireland, UK.

[8] Department of Colorectal Surgery, Royal College of Surgeons in Ireland University of Medicine and Health Sciences, Beaumont Hospital, Beaumont Road, Dublin 9, Ireland.

Corresponding author
Prof Jochen H. M. Prehn, PhD. Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland University of Medicine and Health Sciences, 123 St. Stephen's Green, Dublin 2, Ireland. Email: prehn@rcsi.ie. ORCID: 0000-0003-3479-7794.

Authors contribution

AS, SaCh, MS, MSt, BK, EMcD, EPO'C, JFG, SSMD and MA were involved in methodology, data validation, and curation. SaCh, EMcD, JFG, and FG performed the Cell DIVE processing as well as cell segmentation and single-cell quantification. AS performed cell type classification, statistically analysis and study investigation. JPB, NMcC, DAmcN, JS, CF, NU, JHMP, and DBL were involved in clinical sample acquisition and data collection. EMcD conducted the sample imaging and image processing. AS and JHMP wrote the manuscript. DBL, FG, and JHMP were involved in funding acquisition. JHMP supervised the study. All authors edited and revised the manuscript text.

Abstract

Background

Stage III colorectal cancer poses a significant threat of metastasis development, as tumour resection and adjuvant chemotherapy do not guarantee prolonged disease-free survival.

Objective

The spatial, quantitative, and qualitative characteristics of various cell types within tumour tissues could be key to developing accurate prognostic AI models.

Design

Tissue microarrays created from primary tumour tissues collected during surgical resection from a cohort of 466 stage III colorectal cancer (CRC) patients were analysed for 61 protein markers at the single-cell level using multiplexed immunofluorescence imaging via the Cell DIVE™ platform. Subsequent cell-type classification enabled quantitative cell-type analyses, co-localisation neighbourhood assessments, and cell-type-specific protein signature discoveries that distinguish between early and late/non-recurring patient samples.

Results

This study identifies a stem cell protein profile that drives tumour relapse. A deep neural network (DNN) model, based on a stem cell protein signature composed of BAX, MLKL, FLIP, GLUT1, and CDX2, provided accurate prognosis for stage III CRC patients in both discovery and validation cohorts and in an independent validation cohort. Nodal count-based metric further increased prognosis accuracy. Our study also revealed distinct spatial arrangements of immune, endothelial, and stem cells that were linked to early tumour recurrence.

Conclusion

Our findings propose a clinically promising prognostic tool based on a five-protein stem cell signature. These markers not only predict chemotherapy resistance in cancer stem cells but also suggest potential therapeutic strategies such as combinatorial treatments incorporating small molecule inhibitors targeting FLIP and GLUT1.

Key messages

What is already known on this topic

- More than 20% of stage III colorectal cancer patients will experience early tumour recurrence within the first 3 years post treatment that includes surgery and adjuvant 5-FU based chemotherapy treatment.
- Several studies pointed towards involvements of number of cell type specific spatial neighbourhoods in tumour progression where some immune tumour microenvironment promoting angiogenesis and intravasation events, some may provide immunosuppression.
- Cancer stem cells could be responsible for metastatic tumour spread, early recurrence and chemoresistance.

What this study adds

- Spatial single cell quantitative multiplex profiling of 45 cancer hallmark proteins and 15 cell identity markers in 466 stage III CRC patients tissue samples demonstrated significant

differences in cellular proximity neighbourhoods, cell type specific abundance and expression between the early and late recurrence samples.

- We discover that macrophages show spatial association with the blood vessels in early recurrence samples. Moreover, we observed conglomeration of B cells and macrophages with Tregulatory, Thelper and Tcytotoxic cells in association with early recurrences.

- We showed that stromal abundance of Tregulatory, Thelper, Tcytotoxic cells and monocytes are significantly in late, and no recurrence samples compared to early recurrence samples.

- The most differential expression profile that differentiates late and no recurrence samples from the early recurrence samples is related to the stem cell population. Particularly, we found overexpression of GLUT1, FLIP and downregulation of BAX, BAK, MLKL and CDX2 proteins in the cancer stem cell of early recurrence samples.

- We built a neural network based on the cancer stem cell protein signature (BAX, MLKL, FLIP, GLUT1 and CDX2 proteins) that delivers a high-performance prognostic classifier.

How this study might affect research, practice or policy

- Our results propose a clinically promising prognostic tool based on a five-protein stem cell signature that outperforms existing clinical and proposed transcriptomic based signatures for separation between risk groups.

- Moreover, our five-protein signature markers not only predict stem cell chemotherapy resistance and therefore tumour recurrence but also suggest potential therapeutic strategies. For instance, this approach could guide combinatorial treatments at high risk of chemoresistance, such as incorporating small molecule inhibitors targeting FLIP (currently in discovery phase) and GLUT1 (already under preclinical trial evaluation).

Introduction

Colorectal cancer (CRC) is the second leading cause of cancer-related death worldwide. The demographics of this disease have changed dramatically over the last decade, with surging numbers of patients being detected before the age of 40, resulting in an estimated 73% increase in CRC related deaths by 2040 according to the World Health Organization (2023, July 11).

The advanced stages of CRC are defined by the spread of cancer cells into lymph nodes and potentially other parts of the body. Stage III CRC is characterized by the spread of micrometastatic cells into lymph nodes surrounding the primary tumour site. This spread is believed to be initiated by a population of pluripotent cancer epithelial cells including cancer stem cells (CSCs) that are the most efficient “seeders” of metastasis due to their high clonogenic capacity [1]. In addition to metastasis, CSCs also contribute to tumour growth and chemotherapy resistance [2]. The current standard-of-care for stage III treatment involves surgical resection of the primary tumour and positive lymph nodes, followed by adjuvant oxaliplatin/fluoropyrimidine/leucovorin (5-fluorouracil (5FU), FOLFOX; or xeloda/capecitabine, XELOX) chemotherapy. However, around 24% of stage III patients will experience tumour recurrence within the first 3 years post-surgery [3]. It is believed that many of these

patients will progress because of the persistence of disseminated CSCs enabled by their resistance to adjuvant chemotherapy.

In this study, we investigated the spatial, quantitative and qualitative characteristics of CSCs and other cell types in treatment-naive stage III tumours resected from the patients followed for at least 3 years post-surgery. Tissue microarrays (TMAs) were constructed from tumour tissues collected during surgical resection of primary tumours in a cohort of 493 stage III CRC patients who underwent adjuvant chemotherapy (Figure 1). We quantified proteomic markers with multiplexed immunofluorescence imaging performed with the Cell DIVE™ platform followed by image processing, single cell quantification and cell type classification. We evaluated at the single cell level 61 protein markers, including markers for cell segmentation and identification, epithelia/stroma/endothelial segregation, cell proliferation and differentiation, microsatellite stability, epithelial mesenchymal transition (EMT), apoptosis and necroptosis, metabolism and immune signalling. We performed comparisons between frequencies of cell types, the expression of cell 'state' markers in individual cell types, and their spatial proximity and explored how these parameters correlate with early recurrence. Analysing separate discovery, validation and independent validation cohorts, we identify the existence of a highly prognostic CSCs specific protein signature. Based on this new signature, we developed a deep neural network (DNN) model that facilitates robust prognosis of recurrence.

Figure 1. *Schematic diagram representing the design of this study.*

Results

Robust cell type classification pipeline.

We developed a pipeline for cell type identification for high-plex single cell data analysis in stage III tumour tissue (*see Methods*). The pipeline begins with the recognition of epithelial $CDX2^+$, $PCK26^+$ and $AE1^+$ positive cells ('cancer cells'). Secondly, the pipeline identifies Monocytes ($CD3^-CD8^-CD4^+$), T regulatory cells ($CD3^+CD4^+FOXP3^+$), T helper cells ($CD3^+CD4^+FOXP3^-$), double negative T cells (T cell $CD3^+CD8^-CD4^-FOXP3^-$), cytotoxic T cells ($CD3^+CD8^+$) and double positive T cells ($CD3^+CD4^+CD8^+$) (**Figure 2A**). Thirdly, macrophages (Macrophage, CD4↑CD68↑), B cell-associated macrophages (Macroph+B1, CD4↑CD68↑CD20↑, Macroph+B2, CD4↑CD68↑modCD20) and endothelial cells (Endothelial_1, CD31↑CD34↑COLIV↑SMA↑, Endothelial _2, CD31↑CD34↓COLIV↑SMA↑) are classified (**see Supplementary Methods**). As the final step, the pipeline classifies the stem cell (ALDH1↑KI67↓AE1↓) population (**Figure 2B**).

Final cell type annotations (**Figure 2C**) established on the individual TMA slides ensured an accurate, unbiased and unsupervised classification, compared to, for example, a manually supervised annotation affected by a quantitatively limited training set, or deep learning-based classification pipelines that could be affected by model overfitting [4] or experimental batch effects.

Figure 2. Illustration of cell type classification pipeline*. (A) Automated single binary marker threshold detection algorithms determined epithelial and T cells over TMA slides. (B) Algorithm for multiple marker dependent detection of endothelial, macrophage and stem cells rely on hierarchical clustering following linear discriminant analysis (LDA) based classification model. (C) The example of final cell annotation illustrated on one of the cores.*

Stem, immune and endothelial cells proximity neighbourhoods differentiate early from late/non-recurring patients

Recent advances in tissue spatial proteome multiplex technologies (for example, Cell DIVE™, GeoMx® or CODEX [5] in conjunction with a range of spatial analysis computational tools [6–10] have accelerated the discovery of cellular neighbourhoods that dictate the course of CRC progression. Therefore, we first investigated the spatial cellular landscape and potential differences between early and late/non-recurring tumour samples from all four clinical cohorts (**Table S1**). We calculated the co-localization neighbourhood probabilities and correlated them between different cell types using the hoodscanR package [8]. These correlation coefficients were collated into the median coefficients over two groups of cores: early and late/non-recurrence. The final pairwise median coefficients were clustered to recognise the cell type-specific neighbourhoods that are potentially associated with propensity of recurrence (**Figure 3A, B**). We tested proximity neighbourhoods of different sizes and chose the optimal proximity of 20 nearest cells in the final co-localisation probabilities calculation that showed the most discriminant features between the two groups of cores.

Table S1. *Baseline table with characteristics of HV, MSK, RCSI and Taxonomy clinical cohorts.*

A first feature we detected was the proximity of macrophages to endothelial cell groups in the cores of patients with early recurrence (**Figure 3B, D**) compared to the group of cores from late/non-recurring patients (**Figure 3A, C**). This suggests that macrophages spatially associated with blood vessels may promote angiogenesis and be better positioned to induce more frequent intravasation events that lead to early metastatic recurrence [11].

Figure 3. Cell neighbourhood differences between early and late/non-recurring patients. *Pairwise spatial correlation of neighbourhood probabilities between different cell types in the group of late/non (A) and early (B) recurrent cores. (C, D) The example of cores from different recurrence groups that are demonstrating the differences in the macrophage proximity to the two types of endothelial cells. (E, F) The example of cores from different recurrence groups that are demonstrating the differences in integrity of epithelium and stem cells within it. (G, H) The example of cores from different recurrence groups that demonstrate the differences in composition of immune hot spots.*

The second feature that we found in early recurring cores were stroma-associated stem cells spatially separated from the epithelium (**Figure 3B, F**). The colon epithelium comprises epithelial cells expressing markers CDX2+ and PCK26+ (Epithelial_in_Epithelium cells) tightly clustered together. Typically, stem cells should remain within epithelium layer and not extend into stromal regions. In early recurring samples, stem cells were dispersed away from Epithelial_in_Epithelium cells. In contrast, in late/non-recurring samples, stem cells were spatially closer to Epithelial_in_Epithelium

cells (**Figure 3A**), suggesting better preservation of the epithelial crypt structure's integrity (as shown in **Figure 3E**) compared to early recurrence samples (**Figure 3F**).

The third feature of early recurring samples was a frequent colocalization of Tregulatory and Thelper with B cell-associated macrophages. These were also in proximity to Tcytotoxic, Tdouble positive and Tcell lymphocytes (**Figure 3H**). In contrast, we found that Tregulatory and Thelper cells in late/non-recurring samples were mostly excluded from this close proximity arrangement (**Figure 3G**). The conglomeration of B cells with Tregulatory, Thelper, Tcytotoxic and Tcell lymphocytes suggests the formation of immune 'hot spots', some of which possibly resemble tertiary lymphoid structures (TLS) that can play critical roles in antitumor immunity [12,13]. The identification of 'hot spots' also in early recurring patient samples suggests that factors other than antitumoral defence mechanism may play a more prominent role in a subset of early recurring patients.

Stroma enrichment with high numbers of T regulatory and T helper cells are prominent in tumour tissue of late/non-recurring patients

Next, we analysed the frequency of each cell type within stroma, epithelia and the entire core area across individual patients from all four clinical cohorts (**Table S1**) and how these relate to prognosis (**Figure 4A-C, Figure S2**). Our results showed a substantial, significant increase in the percentage of Tregulatory and Th cells in the late/non-recurring samples in both stroma (**Figure 4A**) and across the entire core (**Figure S2**). This finding additionally validated our previous discovery in a distinct, smaller patient cohort of 5-FU-treated stage III CRC patients where a Cox regression model predicted similar trends [14]. The percentage of epithelial cells dissociated from the epithelium and located in the stroma was also found to be higher in the early recurring samples (**Figure S2**), suggesting a higher propensity of epithelial cells for EMT is prognostic of early recurrence. We also observed that overall stem cell frequency was not indicative of recurrence (**Figure 4B**).

Figure 4. Prognostic cell state markers found in the stem cell population. *(A-C) Cell types frequencies are compared between early and late/non-recurrent groups of patients. Cores from individual patients are combined to calculate the percentage of each cell type over stroma and epithelium. Boxplot represents interquartile range (IQR) with the median, 25% (Q1) and 75% (Q3) quantiles as middle, lower and upper hinges, respectively. Lower and upper whiskers are Q1-1.5*IQR and Q3+1.5*IQR, respectively. P-values of the mean cell type frequency compared by the Wilcoxon test are indicated for significant difference between early and late/non-recurrent groups of patients (* P<0.05, ** P<0.005, *** P<0.0005, **** P<0.00005). Differential marker expression analysis in stem cell population (D) performed on the single core bases combining single cell expression into the median signal over the stem cell population of each core individually. Significance P-value cutoff at 0.05 is indicated by horizontal dashed line and log-fold change cutoffs at 0.25 are indicated by a vertical dashed line.*

Figure S2. Prognostic cell state markers found in the stem cell population. *Cell types frequencies are compared between early and late/non-recurrent groups of patients. Cores from individual patients are combined to calculate the percentage of each cell type over tissue core area. Boxplot represents interquartile range (IQR) with the median, 25% (Q1) and 75% (Q3) quantiles as middle, lower and upper hinges, respectively. Lower and upper whiskers are Q1-1.5*IQR and Q3+1.5*IQR, respectively. P-values of the mean cell type frequency compared by the Wilcoxon test are indicated for significant difference between early and late/non-recurrent groups of patients (* P<0.05, ** P<0.005, *** P<0.0005, **** P<0.00005).*

Stem cells in early vs. late/non-recurring samples show a prominent difference in protein profile

Even though the frequency of CSCs was not prognostic, we hypothesised that the specific biology of CSCs may predict their propensity as propagators of metastasis when resident in the primary tumour. To explore this hypothesis, we compared the expression profiles of CSC between early and late/non-recurring samples using a differential expression analysis of quantified protein markers. The median expression of cell 'state' markers of the stem cell population in individual cores were calculated for four cellular compartments (whole cell, cytoplasm, membrane and nucleus). In this analysis, we considered individual tissue cores (2-3 core per patient) rather than an aggregated (i.e. patient) level to account for intratumoral heterogeneity as every core was taken from different sites of the patient's tumour tissue. This analysis revealed 6 protein markers that were significantly differently expressed and could be considered to constitute a prognostic CSC signature (**Figure 4D**). We detected overexpression of two markers in stem cells of early recurring samples. One of these proteins is a marker of cell metabolism that was previously found to be associated with tumour progression. This protein, glucose transporter 1 (GLUT1), is known to contribute to the Warburg effect in cancer cells [15]. The second overexpressed protein was cellular-FLICE inhibitory protein (FLIP) that blocks extrinsic apoptosis by interfering with the activation of caspase-8 at death-inducing signalling complexes (DISCs) [16] upon death receptor stimulation by death ligand-expressing immune effector cells. Among down regulated proteins, we detected the pro-apoptotic proteins BAK and BAX that are the key components of the intrinsic apoptotic pathway that ensure mitochondrial outer membrane permeabilization (MOMP) [17–20], as well as mixed lineage kinase domain-like (MLKL) pseudokinase that controls necroptotic cell death resulting from plasma membrane permeabilization [21,22]. Finally, we detected the loss of caudal-related homeobox transcription factor 2 (CDX2), a protein that regulates homeostasis of intestinal epithelial cells and cell adhesion. Importantly, such profiles were not detected in epithelial 'non-stem' cancer cells where the differential protein profile lacks MLKL, GLUT1 and FLIP signatures (**Figure S3**). This analysis revealed that stem cells found in primary tumour samples from early recurring patients do not only show a different localisation pattern but are also primed to promote EMT and resist apoptotic and necroptotic cell death signalling as well as metabolic stress.

Figure S3. Differential markers found in other cell populations. *The volcano plots show the results of differential marker expression analysis in all classified cell populations performed on a single-core basis, combining single-cell expression into the median signal over the given cell type population in each tissue core individually. Significance P-value cutoff at 0.05 is indicated by a horizontal dashed line and log-fold change cutoffs at 0.25 are indicated by vertical dashed lines*.

Limited differential signatures were detected in a few other prominent cell types (**Figure S3**). For example, Thelper, Macrophage and Tdouble positive CD4+CD8+ cells showed CASP8 downregulation in early recurring samples. This may indicate resistance to extrinsic apoptosis in these cells types. Th2 cells [23] and M2 macrophages [24] are known contributors to immunosuppression, and Th2 polarisation can be triggered by Tdouble positive CD4+CD8+ cells [25]. Thus, cell death resistance

in these cell types could be indispensable for tumour propagation. Collectively, however, the stem cell population offered the best potential for development of a prognostic signature.

A deep neural network delivers a highly prognostic model of recurrence in stage III CRC patients

Deep neural network (DNN) based approaches offer indispensable capabilities and solutions to the challenges in clinical oncology and medical data analysis. We next used DNN to build a prognostic model for early CRC recurrence based on the markers differentially expressed in stem cells in early recurring versus late/non-recurring patients. The successful DNN model was identified in Campion/Challenger approach based on 5 differentially expressed markers (cytoplasmic BAX, MLKL, FLIP, GLUT1 and whole cell CDX2). We trained our DNN with 700 patient cores from 3 discovery cohorts (HV, MSK, Taxonomy) (**Table S4**). During training, the data set was split between training and validation sets at patient level. In each model, variant hyperparameter optimization around the repeated k-fold method was employed to generate multiple balanced splits between training and validation data sets assisting the network's prediction accuracy by F1-score of early recurrent class of cores in the validation subset (**see Methods**). The champion network was able to capture significant differences in the recurrence dynamics between groups of early and late/non-recurring patient cores in the resulting training (**Figure 5A**) and validation cohorts (**Figure 5B**) that were subsequently verified in the independent validation cohort (RCSI, **Figure 5C**).

Table S4. *Baseline table with clinical characteristics for patients with identifiable stem cell signature from HV, MSK, RCSI and Taxonomy clinical cohorts used in the training and validation of the DNN model.*

Figure 5. Accurate prognostic model is established by DNN based on 5 protein markers. *High accuracy of DNN model based on 5 markers (cytoplasmic BAX, MLKL, FLIP, GLUT1 and whole cell CDX2) in the training (A), validation (B) and independent validation (C) rounds demonstrated in significant difference between early and late/non-recurrent groups of cores on Kaplan-Meier plots.*

Finally, clinicopathological data were incorporated to explore whether this discriminated further the recurrence groups. We performed univariate Cox regression analysis identifying clinical parameters in all four cohorts that associated with increased hazard ratio (**Table S5**) and tested them in a multivariate Cox regression model (**Figure 6A**) to account for potential covariates such as age and sex. We identified N staging as the key parameter that further improved prognostic accuracy when combined with predictions of our DNN stem cell model (**Figure 6B**), delivering a new and powerful concise protein signature of disease recurrence.

Table S5. *Univariate Cox proportional hazard analysis.*

Figure 6. Improved accuracy of prognostic model with inclusion of regional lymph nodes (N) stage. *(A) The Cox's proportional hazards regression analysis of clinical parameters based on multivariate and univariate model. (B) The improved accuracy in the prognosis with combination of DNN model and N stage information in the four cohorts combined.*

Discussion

The results of this study identify a stem cell-specific protein profile that correlates with tumour relapse, highlighting several key proteins that deliver accurate prognostic predictions from the primary tumours of patients diagnosed with stage III CRC and treated post-surgically with adjuvant chemotherapy. In addition, this study highlights several patterns of cell proximity arrangements, including immune, endothelial, and stem cells present in primary tumours that are associated with early tumour recurrence.

We first developed a cell classification pipeline that allowed unbiased identification of 15 cell types over the multiplexed and processed TMA slides. The cell type frequency analysis shows the association of higher stromal abundance of regulatory, helper, cytotoxic and double positive T cells with better prognosis. Next, we analysed cell neighbourhoods to find characteristic spatial proximity arrangements that specifically discriminate early from late/non-recurring patients. By including all cell types found in our neighbourhood analyses, the scope of our search was broad, ranging from identifying spatially organised antitumoral lymphocyte arrangements [26] and TLS-like structures [13,27] to cellular neighbourhoods that promote macrophage-fibroblast communities [28,29] and migration events of CSCs [30]. As a result, we identified three prominent features: Firstly, we found that early recurring cores more frequently include proximity arrangements of macrophages among groups of endothelial cells. This indicates the association of macrophages with blood vessels that, considering recent findings from other laboratories [11], could indicate the initiation of intravasation events that lead to early metastatic recurrence. In previous studies, such perivascular macrophages were found to be a cause of local transient hyperpermeability of tumour vasculature [31,32]. In turn, this permeability increase may lead to transendothelial migration of cancer cells during intravasation [33].

Secondly, we found immune ‘hot spots’ in some early recurring cores, specifically composed of Tregulatory, Thelper surrounded by Macroph+B1, Macroph+B2, Tcytotoxic, Tdouble positive and Tcell. Such ‘hot spots’ can participate in antitumor immunity [12,13] and reverse immunosuppression. Indeed, immunosuppression has already been associated with intratumoral immune ‘hot spots’ that include high content of Tregulatory together with B-cells; this occurrence, similar to our findings, was linked to unfavourable outcomes, for example, in lung squamous cell carcinoma (LUSC) [34]. Important to note that this effect did not correlate with the overall total content of Tregulatory or Thelper across stroma or entire tissue core. In fact, the high percentage of Tregulatory and Thelper cells across whole stroma or core areas was rather associated with late and non-recurring patient samples (**Figure 4A, Figure S2**).

Third, we established that the normal epithelial crypt structure is better preserved in the cores of late/non-recurring patients. Here, stem cells are better integrated within the crypt and do not extend into the stroma of the tissue. This may indicate that stem cells from late/non-recurring patients are less prone to EMT events. Moreover, we observed that the percentage of epithelial cells located in the stroma (Epithelial_in_Stroma, **Figure S2**) was higher in the early recurring samples. This may indicate that metastatic progression may involve both stem and differentiated epithelial cells as cells of origin [30,35].

Next, we established a proteomic signature that distinguished stem cells found in early recurring cores from the stem cells in late/non-recurring samples. The most overexpressed proteins in stem cells in early recurring core were GLUT1 glucose transporter and FLIP, a regulator of the extrinsic apoptotic pathway [16] and necroptosis [36]. Among downregulated proteins we found BAK and BAX proteins required for MOMP during intrinsic and BID-dependent extrinsic apoptosis [17–20], MLKL pseudokinase, an activator necroptotic cell death [21,22], and CDX2 which stimulates cell adhesion in intestinal epithelial cells. Thus, the stem cells from early recurring patients appear to have a unique survival advantage, making them resistant to the most prominent cell death pathways: death receptor-mediated/extrinsic apoptosis, intrinsic apoptosis and necroptosis. Through upregulation of GLUT1, these cells may also increase glucose influx rewiring their metabolism to promote growth and proliferation [15] and trigger cell migration and malignant transformation through loss of CDX2 which has previously been identified to be associated with poor prognosis [37].

Finally, we trained a DNN model to recognize this stem cell signature based on five differentially expressed proteins: BAX, MLKL, FLIP, GLUT1 and CDX2. Our model showed very high prognosis performance separating survival trends for the group of early and late/non-recurring samples, with an exceptionally high significance level, $p<0.0001$. Furthermore, the DNN model could be further enhanced by inclusion of N staging as clinicopathological parameter, delivering high hazard ratios in our patient cohort.

The ultimate goal for further clinical implementation of our prognostic signature is the establishment and calibration of the robust pipeline (**Figure S6**) compatible with the multiplex technologies available for clinical use, for example, such as ZEISS Axioscan 7 and Leica Aperio VERSA for imaging and quantification of 5 markers of the prediction signature as well as segmentation marker such as PCK26 and stem cell marker ALDH1 or other alternative stem markers such as leucine-rich repeat-containing G-protein-coupled receptor 5 (LGR5), CD44 or Musashi1 (MSI1). Additionally, the inclusion of the simultaneous brightfield image, as on ZEISS Axioscan 7 and Leica Aperio VERSA platforms, can illuminate the necessity of fluorescent staining for segmentation marker as the Hematoxylin and eosin staining (H&E) can provide accurate segmentation of the cytoplasm regions as well as epithelium with the appropriate image processing pipeline. Additionally, the use of larger tissue samples, currently relevant to the clinical setting, can further improve the accuracy of the prediction method.

Figure S6*. *The overview of the plausible adaptation concept of the acquisition and application of the prognostic signature for the prediction of CRC recurrence in stage III CRC samples. *The pipeline begins with appropriate image segmentation to identify the epithelium and stem cell cytoplasm regions at the single cell level. As an example, we used PCK26 staining alone for epithelium and single cell detection, however, it can be complimented or substituted by H&E staining based segmentation pipeline. Next, stem cells are identified through the ALDH1 fluorescent signal, resulting in the Stem_mask generation. Finally, the overlay of single cell epithelial Cytoplasm_mask with Stem_mask delivers the single cell Stem_Cytoplasm_mask. This mask is used to acquire the mean pixel intensities of the fluorescent signal for 5 proteomic markers of the prognostic signature over the area of individual stem cells. These quantities are further agglomerated in the median protein signal across all stem cell populations for the individual core sample, thus creating the final core level 5 protein prediction signature. This signature input into our DNN prediction network that delivers the prognosis prediction.*

Overall, our results propose a clinically promising prognostic tool based on a five-protein stem cell signature. Compared to currently proposed transcriptomic based signatures [38] our prognostic signature exhibits superior discriminative ability which is demonstrated by a substantial separation between risk groups. Moreover, our five-protein signature markers not only predict stem cell chemotherapy resistance and therefore tumour recurrence but also suggest potential therapeutic strategies. For instance, this approach could guide combinatorial treatments at high risk of chemoresistance, such as incorporating small molecule inhibitors targeting FLIP (currently in discovery phase) and GLUT1 (already under preclinical trial evaluation).

Methods

Clinical cohorts description

This study was performed in accordance with ethical guidelines for clinical research with the approval of the ethics committee for each institution. Formalin-fixed, paraffin-embedded primary tumour tissue sections were obtained from four different cohorts. For “RCSI” cohort, tissues were provided by the Beaumont Hospital (RCSI, Ireland) Colorectal Cancer Biobank with written consent provided by all patients. Institutional ethical approval was granted by the Beaumont Hospital Research and Ethics Committee (References 08/62 and 19/46). For “Taxonomy” cohort, tissues were supplied by the Queen’s University Belfast (QUB, UK) Department of Pathology with written consent provided by all patients and institutional ethical approval granted (NIB12-0034). For “HV” cohort, samples were collected at Huntsville Clearview Cancer Center (CCI, US) under granted institutional ethical approval (IRB/CC1 Colon 01) and, for “MSK” cohort, tissue were provided by Memorial Sloan Kettering Cancer Center (MSK, US) under ethical agreement (IRB/WA0497-09). All patients included in the study underwent surgery for histologically proven stage III CRC cancer. Only patients with adjuvant and without neoadjuvant therapy were included, and only cases for which we had access to tissue blocks on the original resection and clinical follow-up information were included. Patients’ exclusion criteria were as follows: postoperative mortality within 30 days; a limited follow-up period of less than 3 years in cases without recurrence; synchronous multiple cancers; positive surgical margins; concomitant

inflammatory bowel disease; and familial adenomatous polyposis or previous malignancy within 5 years. A clinical summary of all cohorts can be found in **Table S1**.

The Patient and Public Involvement statement

Because this was a retrospective study of archived samples and data, patients or the public were not involved in the design, or conduct, or reporting or dissemination plans of our research.

Differential expression analysis

The median expression of cell 'state' markers of the stem cell population in individual cores were calculated for four cellular compartments (whole cell, cytoplasm, membrane and nucleus).

In this analysis we considered individual cores rather than patients to account for intratumoral heterogeneity considering that every core was taken from different sites of patients' tumour tissue. Despite our normalization pipeline [39] created to equalize the batch effects for every protein across several TMA slides, some proteins may remain affected. Therefore, to find robust protein markers which are independent of batch effects, and that differentiate early from late/non-recurring stem cell populations, the differential expression analysis was conducted with and without consideration of TMA slide related batch effects. Then the resulting strong significant difference in expression detected in the analysis without corrections and confirmed in the analysis with corrections were then considered for prognostic markers selection (**Figure 4D**). Thus, the initially normalised intensities for these selected markers can be confidently used further in the prediction model training that accounts for small variance that may be present due to the batch effects.

Deep neural network (DNN) training

We trained our DNN with 700 patient cores from three cohorts (HV, MSK, Taxonomy cohorts) (**Table S4**). During training the data set was split between training and validation sets at patient level. We used repeated 8-fold cross-validation method (20 repeats) to assess generalisation ability across multiple data splits. In each split the general proportional balance of early and late/non-recurring patients we preserved for training and validation sets. Each training event was started with initial learning rate that gradually decay over training steps progression (epoch) with the given decay rate. To avoid internal covariate shifts in the DNN layers the layer weights batch normalisation was applied. A class weight hyperparameter was applied to compensate for imbalance class representation (early and late/non-recurring cores) and was also subject for optimisation. Feature importance was assisted with L1-norm value in the first layer of the DNN. The validation performance was evaluated by calculating the F1-score for predicting early recurrence, accuracy, precision, specificity and recall. We optimised the DNN hyperparameters to maximise the F1-score in validation set using Champion/Challenger approach. The hyperparameters considered in Adam optimisation [40] procedure were: Initial learning rate, learning decay rate, class weight, activation function, number of hidden layers, size of each hidden layer, dropout rate, penalties for L1 and L2 regularisation. We implemented the early stopping patience principal to prevent over-training, which terminated the training when the validation F1-score was no longer improving for 500 consecutive epochs. The

early stopping was allowed only in the stable learning phase that begins at 20th epoch. Finally, the champion DNN was validated in the independent validation cohort (RCSI cohort) (**Table S4**). The neural network was implemented in Julia v1.9.3 [41], using Flux.jl v0.14.6 [42] for architecture design and training, whilst MLUtils v0.4.3 handled data preprocessing operations including observation shuffling and train-validation splitting whilst MKL v0.6.1 accelerates matrix operations through Intel's Math Kernel Library backend.

Statistics

Disease-free survival (DFS) was defined as the time from surgery to the date of local recurrence or relapse of CRC, distant metastasis, or death from an underlying disease. Survival characteristics were described using the Kaplan-Meier method and compared using the log-rank test. Hazard ratios were calculated using the Cox proportional hazard model. All conventional statistical analyses were performed in the R environment [43]. The level of statistical significance was set at $P < .05$. Wilcoxon tests were performed using ggpubr v0.3.0 package, differential expression analysis was done with the moderated t-test with Benjamini-Hochberg FDR adjustment in the *limma* package (v3.40.6) [44] and plots visualisation were performed with the ggplot2 v3.3.0 package [45]. The co-localization neighbourhood probabilities between different cell types were calculated with the hoodscanR package in R [8]. Hierarchical clustering was implemented in Julia v1.9.3 [41], using Clustering v0.15.8 where Ward's method with the Euclidean linkage was implode. MKL v0.6.1 were used for acceleration of matrix operations through Intel's Math Kernel Library backend. LDA runs were performed with R package MASS v7.3-51.5 [46].

Acknowledgement

We thank Dr Andreas Lindner for support with cell classification pipelines.

Funding

This work was supported by US-Ireland Tripartite R01 award from Research Ireland and the Health Research Board [16/US/3301 to J.H.M.P.]; the National Cancer Institute at the National Institutes of Health [R01CA208179 to F.G., P30 CA008748 to J.S., C.F. and N.U.]; a US-Ireland Tripartite R01 award from Health and Social Care Northern Ireland [STL/5715/15 to D.L. and S.M.D.]; Research Ireland through the Research Ireland Centre for Research Training in Genomics Data Science [18/CRT/6214 to B.K.]; EU's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant [H2020-MSCA-COFUND-2019-945385]; and an RCSI MD StAR fellowship [no grant number is applicable to E.P.O'C.].

Disclosure

JS is a consultant for Tempus AI. All other authors have no competing interests to declare.

References

1. Fumagalli A, Oost KC, Kester L, Morgner J, Bornes L, Bruens L, et al. Plasticity of Lgr5-Negative Cancer Cells Drives Metastasis in Colorectal Cancer. *Cell Stem Cell*. April 2020;26(4):569–578.e7.

2. Lei X, He Q, Li Z, Zou Q, Xu P, Yu H, et al. Cancer stem cells in colorectal cancer and the association with chemotherapy resistance. *Medical Oncology*. April 18, 2021;38(4):43.

3. Bouvier A-M, Jooste V, Lillini R, Marcos-Gragera R, Katalinic A, Giorgi Rossi P, et al. Differences in survival and recurrence of colorectal cancer by stage across population-based European registries. *Int J Cancer*. September 2024;155(5):807–815.

4. Amitay Y, Bussi Y, Feinstein B, Bagon S, Milo I, Keren L. CellSighter: a neural network to classify cells in highly multiplexed images. *Nat Commun*. July 2023;14(1):4302.

5. Goltsev Y, Samusik N, Kennedy-Darling J, Bhate S, Hale M, Vazquez G, et al. Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell*. August 2018;174(4):968–981.e15.

6. Chen Z, Soifer I, Hilton H, Keren L, Jojic V. Modeling Multiplexed Images with Spatial-LDA Reveals Novel Tissue Microenvironments. *J Comput Biol*. August 2020;27(8):1204–1218.

7. Hu Y, Rong J, Xu Y, Xie R, Peng J, Gao L, et al. Unsupervised and supervised discovery of tissue cellular neighborhoods from cell phenotypes. *Nat Methods*. February 2024;21(2):267–278.

8. Liu N, Martin J, Bhuva DD, Chen J, Li M, Lee SC, et al. hoodscanR: profiling single-cell neighborhoods in spatial transcriptomics data. March 2024;

9. Mages S, Moriel N, Avraham-Davidi I, Murray E, Watter J, Chen F, et al. TACCO unifies annotation transfer and decomposition of cell identities for single-cell and spatial omics. *Nat Biotechnol*. October 2023;41(10):1465–1473.

10. Nirmal AJ, Sorger PK. SCIMAP: A Python Toolkit for Integrated Spatial Analysis of Multiplexed Imaging Data. *J Open Source Softw*. 2024;9(97).

11. Bied M, Ho WW, Ginhoux F, Blériot C. Roles of macrophages in tumor development: a spatiotemporal perspective. *Cell Mol Immunol*. July 10, 2023;20(9):983–992.

12. Fridman WH, Meylan M, Pupier G, Calvez A, Hernandez I, Sautès-Fridman C. Tertiary lymphoid structures and B cells: An intratumoral immunity cycle. *Immunity*. October 2023;56(10):2254–2269.

13. Karjula T, Niskakangas A, Mustonen O, Puro I, Elomaa H, Ahtiainen M, et al. Tertiary lymphoid structures in pulmonary metastases of microsatellite stable colorectal cancer. *Virchows Arch*. July 2023;483(1):21–32.

14. Stachtea X, Loughrey MB, Salvucci M, Lindner AU, Cho S, McDonough E, et al. Stratification of chemotherapy-treated stage III colorectal cancer patients using multiplexed imaging and single-cell analysis of T-cell populations. *Mod Pathol*. April 2022;35(4):564–576.

15. Ganapathy V, Thangaraju M, Prasad PD. Nutrient transporters in cancer: relevance to Warburg hypothesis and beyond. *Pharmacol Ther*. January 2009;121(1):29–40.

16. Dickens LS, Boyd RS, Jukes-Jones R, Hughes MA, Robinson GL, Fairall L, et al. A death effector domain chain DISC model reveals a crucial role for caspase-8 chain assembly in mediating apoptotic cell death. *Mol Cell*. July 2012;47(2):291–305.

17. Danial NN, Korsmeyer SJ. Cell death: critical control points. *Cell*. January 23, 2004;116(2):205–19.

18. Lindsten T, Ross AJ, King A, Zong WX, Rathmell JC, Shiels HA, et al. The combined functions of proapoptotic Bcl-2 family members bak and bax are essential for normal development of multiple tissues. *Mol Cell*. December 2000;6(6):1389–99.

19. Jürgensmeier JM, Xie Z, Deveraux Q, Ellerby L, Bredesen D, Reed JC. Bax directly induces release of cytochrome c from isolated mitochondria. *Proc Natl Acad Sci U S A*. April 28, 1998;95(9):4997–5002.

20. Wei MC, Lindsten T, Mootha VK, Weiler S, Gross A, Ashiya M, et al. tBID, a membrane-targeted death ligand, oligomerizes BAK to release cytochrome c. *Genes Dev*. August 15, 2000;14(16):2060–71.

21. Galluzzi L, Kepp O, Kroemer G. MLKL regulates necrotic plasma membrane permeabilization. *Cell Res*. February 2014;24(2):139–140.

22. Vandenabeele P, Galluzzi L, Berghe T Vanden, Kroemer G. Molecular mechanisms of necroptosis: an ordered cellular explosion. *Nat Rev Mol Cell Biol*. 2010;11(10):700.

23. Basu A, Ramamoorthi G, Albert G, Gallen C, Beyer A, Snyder C, et al. Differentiation and Regulation of TH Cells: A Balancing Act for Cancer Immunotherapy. *Front Immunol*. 2021;12:669474.

24. Väyrynen JP, Haruki K, Lau MC, Väyrynen SA, Zhong R, Dias Costa A, et al. The Prognostic Role of Macrophage Polarization in the Colorectal Cancer Microenvironment. *Cancer Immunol Res*. January 2021;9(1):8–19.

25. Bohner P, Chevalier MF, Cesson V, Rodrigues-Dias S-C, Dartiguenave F, Burruni R, et al. Double Positive CD4+CD8+ T Cells Are Enriched in Urological Cancers and Favor T Helper-2 Polarization. *Front Immunol*. 2019;10:622.

26. Schürch CM, Bhate SS, Barlow GL, Phillips DJ, Noti L, Zlobec I, et al. Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell*. September 2020;182(5):1341–1359.e19.

27. Wang Q, Zhong W, Shen X, Hao Z, Wan M, Yang X, et al. Tertiary lymphoid structures predict survival and response to neoadjuvant therapy in locally advanced rectal cancer. *NPJ Precis Oncol*. March 2024;8(1):61.

28. Liu X, Qin J, Nie J, Gao R, Hu S, Sun H, et al. ANGPTL2+cancer-associated fibroblasts and SPP1+macrophages are metastasis accelerators of colorectal cancer. *Front Immunol*. 2023;14:1185208.

29. Sathe A, Mason K, Grimes SM, Zhou Z, Lau BT, Bai X, et al. Colorectal Cancer Metastases in the Liver Establish Immunosuppressive Spatial Networking between Tumor-Associated SPP1+ Macrophages and Fibroblasts. *Clin Cancer Res*. January 2023;29(1):244–260.

30. Fumagalli A, Oost KC, Kester L, Morgner J, Bornes L, Bruens L, et al. Plasticity of Lgr5-Negative Cancer Cells Drives Metastasis in Colorectal Cancer. *Cell Stem Cell*. April 2020;26(4):569–578.e7.

31. Arwert EN, Harney AS, Entenberg D, Wang Y, Sahai E, Pollard JW, et al. A Unidirectional Transition from Migratory to Perivascular Macrophage Is Required for Tumor Cell Intravasation. *Cell Rep*. May 1, 2018;23(5):1239–1248.

32. Harney AS, Arwert EN, Entenberg D, Wang Y, Guo P, Qian B-Z, et al. Real-Time Imaging Reveals Local, Transient Vascular Permeability, and Tumor Cell Intravasation Stimulated by TIE2hi Macrophage-Derived VEGFA. *Cancer Discov*. September 2015;5(9):932–43.

33. Borriello L, Coste A, Traub B, Sharma VP, Karagiannis GS, Lin Y, et al. Primary tumor associated macrophages activate programs of invasion and dormancy in disseminating tumor cells. *Nat Commun*. February 2, 2022;13(1):626.

34. Zhang H, AbdulJabbar K, Moore DA, Akarca A, Enfield KSS, Jamal-Hanjani M, et al. Spatial Positioning of Immune Hotspots Reflects the Interplay between B and T Cells in Lung Squamous Cell Carcinoma. *Cancer Res*. May 2023;83(9):1410–1425.

35. Heinz MC, Peters NA, Oost KC, Lindeboom RGH, van Voorthuijsen L, Fumagalli A, et al. Liver Colonization by Colorectal Cancer Metastases Requires YAP-Controlled Plasticity at the Micrometastatic Stage. *Cancer Res*. May 2022;82(10):1953–1968.

36. Humphreys L, Espona-Fiedler M, Longley DB. FLIP as a therapeutic target in cancer. *FEBS J*. November 27, 2018;285(22):4104–4123.

37. Hansen TF, Kjær-Frifeldt S, Eriksen AC, Lindebjerg J, Jensen LH, Sørensen FB, et al. Prognostic impact of CDX2 in stage II colon cancer: results from two nationwide cohorts. *Br J Cancer*. November 2018;119(11):1367–1373.

38. Gallois C, Sroussi M, André T, Mouillet-Richard S, Agueeff N, Mulot C, et al. Prognostic Models From Transcriptomic Signatures of the Tumor Microenvironment and Cell Cycle in Stage III Colon Cancer From PETACC-8 and IDEA-France Trials. *Journal of Clinical Oncology*. May 20, 2025;43(15):1765–1776.

39. Graf J, Cho S, McDonough E, Corwin A, Sood A, Lindner A, et al. FLINO: a new method for immunofluorescence bioimage normalization. *Bioinformatics*. January 3, 2022;38(2):520–526.

40. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. January 30, 2017;

41. Bezanson J, Edelman A, Karpinski S, Shah VB. Julia: A Fresh Approach to Numerical Computing. *SIAM Review*. January 2017;59(1):65–98.

42. Innes M. Flux: Elegant machine learning with Julia. *J Open Source Softw*. May 3, 2018;3(25):602.

43. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria 2013.

44. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. April 2015;43(7):e47.

45. Wickham H. ggplot2. Cham: Springer International Publishing 2016.

46. Venables WN, Ripley BD. Modern Applied Statistics with S. New York, NY: Springer New York 2002.

Figure 1

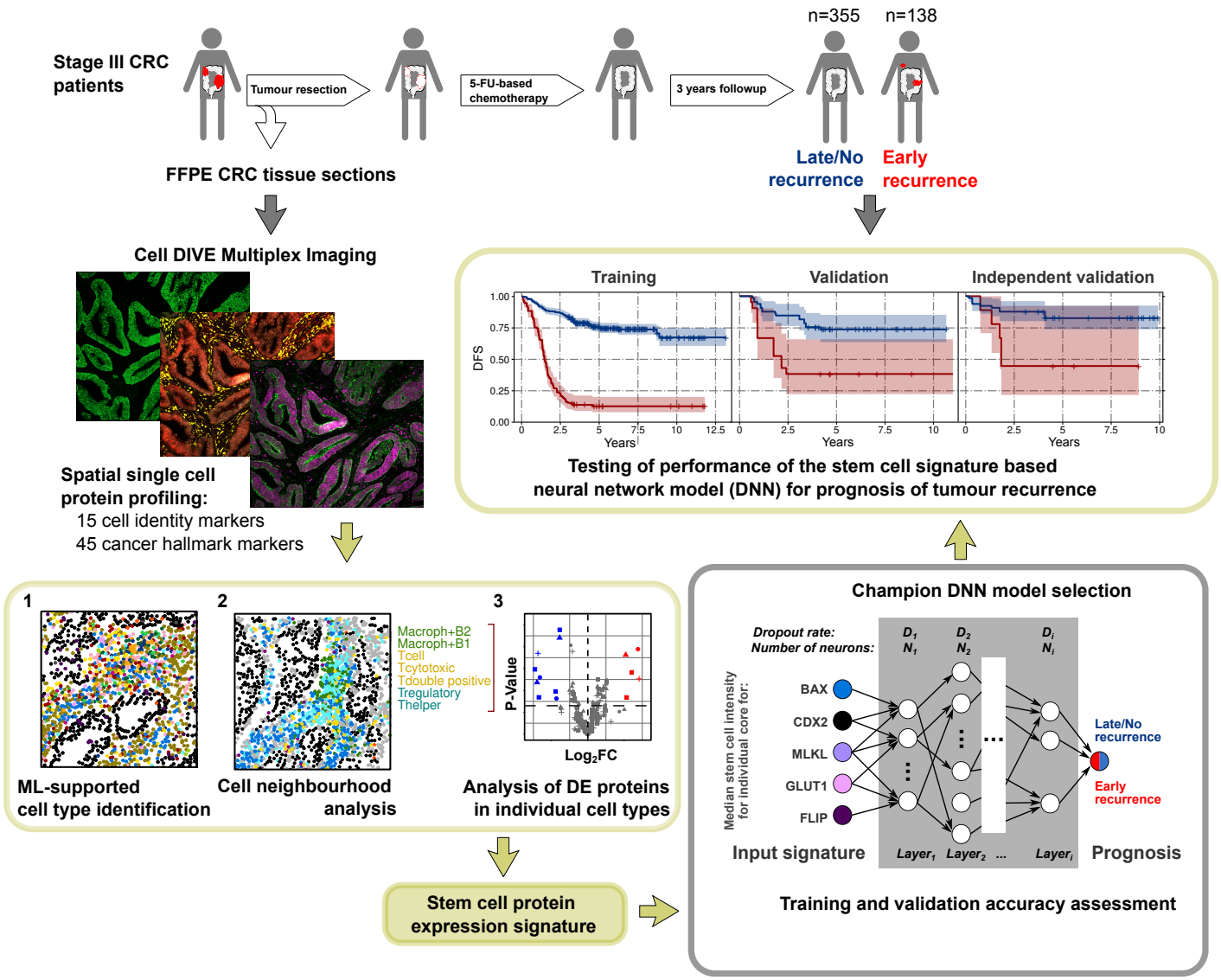

Stage III CRC patients
Tumour resection
5-FU-based chemotherapy
3 years followup
n=355
n=138
Late/No recurrence
Early recurrence
FFPE CRC tissue sections
Cell DIVE Multiplex Imaging
Spatial single cell protein profiling:
15 cell identity markers
45 cancer hallmark markers
Training
Validation
Independent validation
DFS
Years
Testing of performance of the stem cell signature based neural network model (DNN) for prognosis of tumour recurrence
1
2
3
Macroph+B2
Macroph+B1
Tcell
Tcytotoxic
Tdouble positive
Tregulatory
Thelper
P-Value
Log2FC
ML-supported cell type identification
Cell neighbourhood analysis
Analysis of DE proteins in individual cell types
Stem cell protein expression signature
Champion DNN model selection
Dropout rate:
Number of neurons:
Median stem cell intensity for individual core for:
BAX
CDX2
MLKL
GLUT1
FLIP
Input signature
Layer1
Layer2
Layeri
Prognosis
Late/No recurrence
Early recurrence
Training and validation accuracy assessment

Figure 2

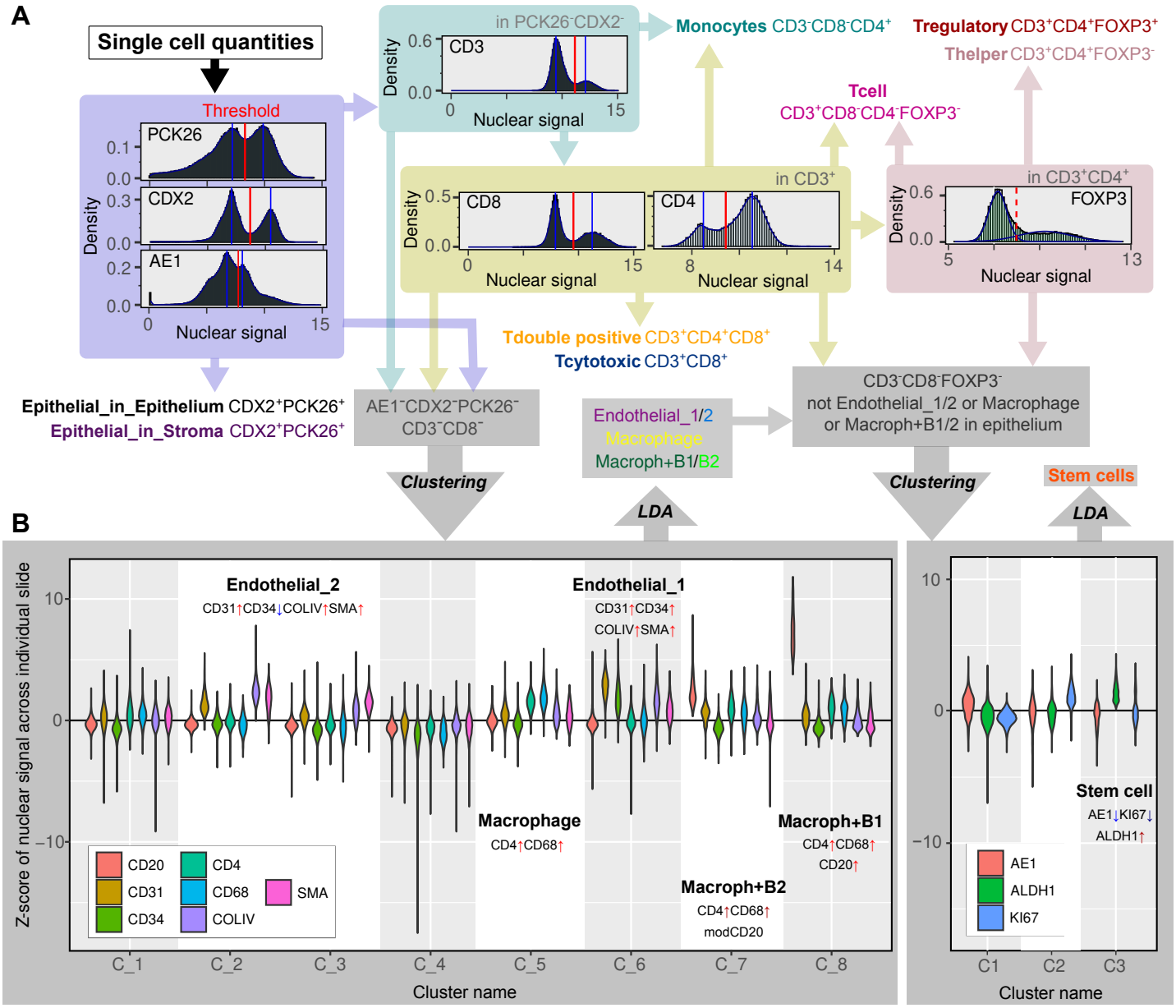

A
Single cell quantities
Threshold
PCK26
CDX2
AE1
Density
Nuclear signal
in PCK26⁻CDX2⁻
CD3
in CD3⁺
CD8
CD4
in CD3⁺CD4⁺
FOXP3
Monocytes CD3⁻CD8⁻CD4⁺
Tregulatory CD3⁺CD4⁺FOXP3⁺
Thelper CD3⁺CD4⁺FOXP3⁻
Tcell
CD3⁺CD8⁻CD4⁻FOXP3⁻
Tdouble positive CD3⁺CD4⁺CD8⁺
Tcytotoxic CD3⁺CD8⁺
Epithelial_in_Epithelium CDX2⁺PCK26⁺
Epithelial_in_Stroma CDX2⁺PCK26⁺
AE1⁻CDX2⁻PCK26⁻
CD3⁻CD8⁻
Endothelial_1/2
Macrophage
Macroph+B1/B2
CD3⁻CD8⁻FOXP3⁻
not Endothelial_1/2 or Macrophage
or Macroph+B1/2 in epithelium
Clustering
LDA
Clustering
Stem cells
LDA
B
Z-score of nuclear signal across individual slide
Endothelial_2
CD31↑CD34↓COLIV↑SMA↑
Endothelial_1
CD31↑CD34↑
COLIV↑SMA↑
Macrophage
CD4↑CD68↑
Macroph+B1
CD4↑CD68↑
CD20↑
Macroph+B2
CD4↑CD68↑
modCD20
CD20
CD31
CD34
CD4
CD68
COLIV
SMA
C_1
C_2
C_3
C_4
C_5
C_6
C_7
C_8
Cluster name
Stem cell
AE1↓KI67↓
ALDH1↑
AE1
ALDH1
KI67
C1
C2
C3
Cluster name
C
Cell annotation
Endothelial_1 (CD31↑CD34↑COLIV↑SMA↑)
Endothelial_2 (CD31↑CD34↓COLIV↑SMA↑)
Epithelial_in_Epithelium (CDX2⁺PCK26⁺ in epithelium)
Epithelial_in_Stroma (CDX2⁺PCK26⁺ in stroma)
Other_in_Epithelium
Macroph+B1 (CD4↑CD68↑CD20↑)
Macroph+B2 (CD4↑CD68↑modCD20)
Macrophage (CD4↑CD68↑)
Monocytes (CD4⁺)
Stem cell (AE1↓KI67↓ALDH1↑)
Stromal_in_Stroma (CDX2⁻PCK26⁻AE1⁻ in stroma)
Other_in_Stroma
Tcell (CD3⁺)
Tcytotoxic (CD3⁺CD8⁺)
Tdouble positive (CD3⁺CD4⁺CD8⁺)
Thelper (CD3⁺CD4⁺)
Tregulatory (CD3⁺CD4⁺FOXP3⁺)

Figure 3

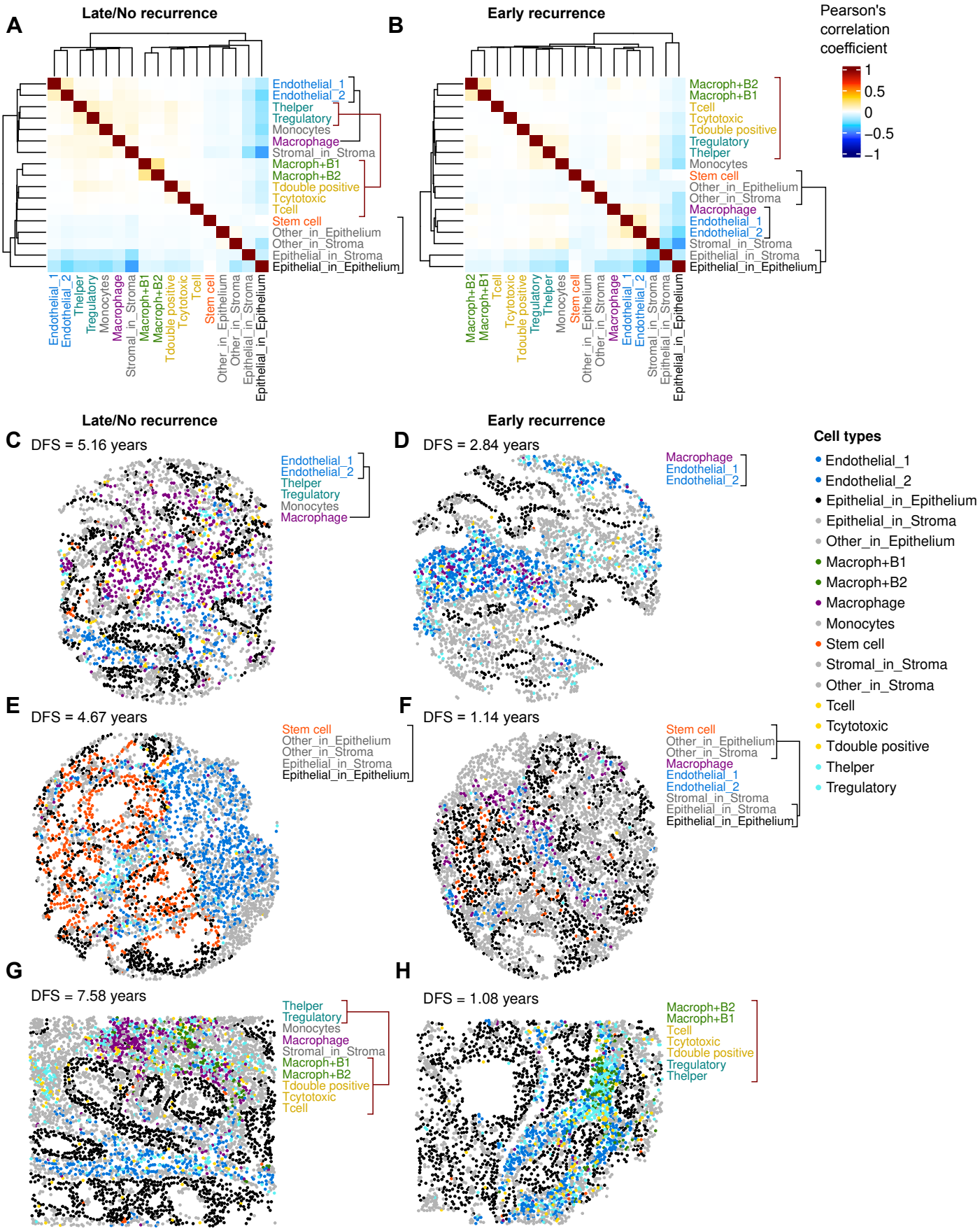

A
Late/No recurrence
B
Early recurrence
Pearson's correlation coefficient
1
0.5
0
−0.5
−1
Endothelial_1
Endothelial_2
Thelper
Tregulatory
Monocytes
Macrophage
Stromal_in_Stroma
Macroph+B1
Macroph+B2
Tdouble positive
Tcytotoxic
Tcell
Stem cell
Other_in_Epithelium
Other_in_Stroma
Epithelial_in_Stroma
Epithelial_in_Epithelium
C
Late/No recurrence
DFS = 5.16 years
D
Early recurrence
DFS = 2.84 years
E
DFS = 4.67 years
F
DFS = 1.14 years
G
DFS = 7.58 years
H
DFS = 1.08 years
Cell types
Endothelial_1
Endothelial_2
Epithelial_in_Epithelium
Epithelial_in_Stroma
Other_in_Epithelium
Macroph+B1
Macroph+B2
Macrophage
Monocytes
Stem cell
Stromal_in_Stroma
Other_in_Stroma
Tcell
Tcytotoxic
Tdouble positive
Thelper
Tregulatory

Figure 4

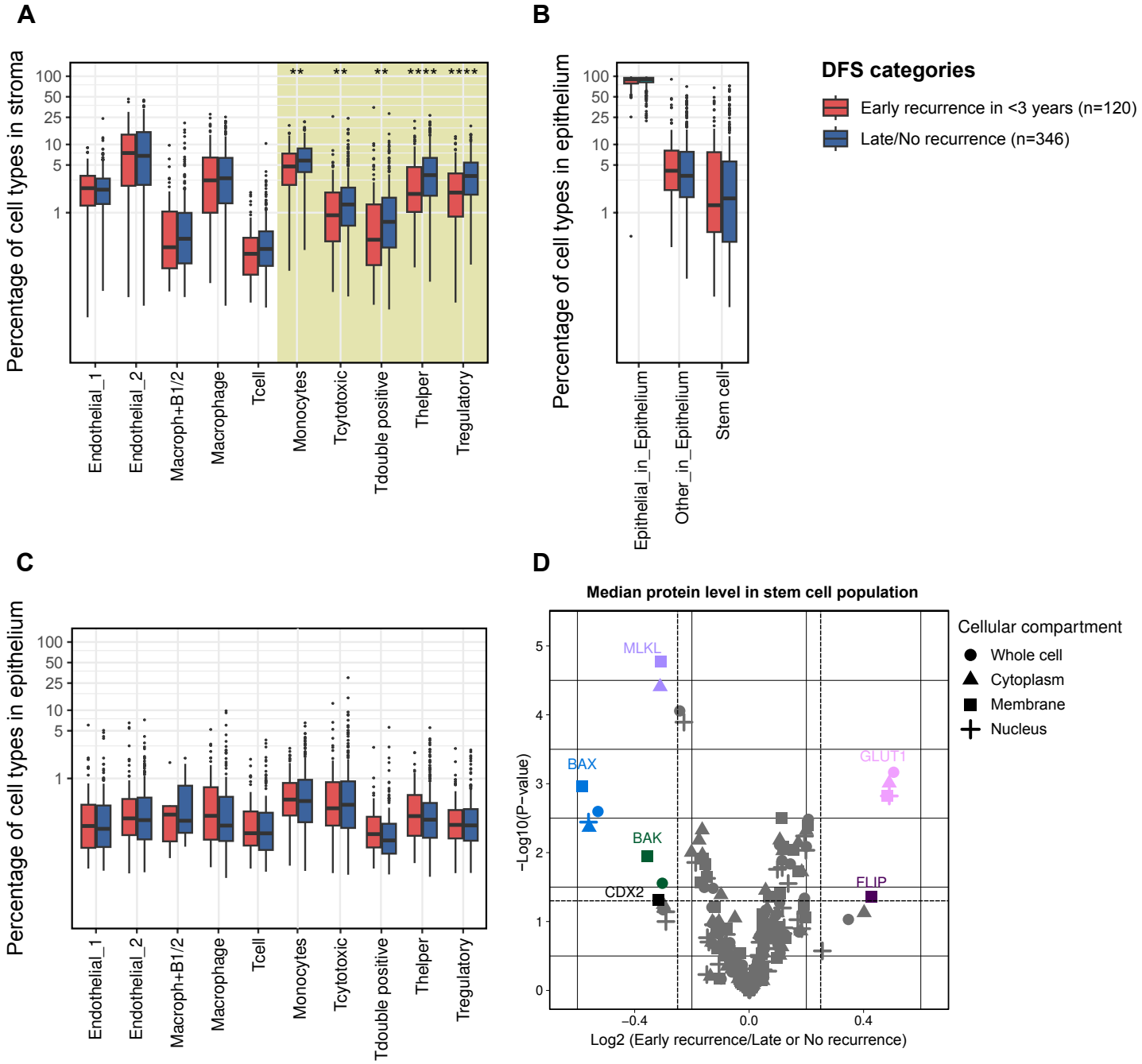

A
Percentage of cell types in stroma
100
50
25
10
5
1
**
**
**
****
****
Endothelial_1
Endothelial_2
Macroph+B1/2
Macrophage
Tcell
Monocytes
Tcytotoxic
Tdouble positive
Thelper
Tregulatory
B
Percentage of cell types in epithelium
Epithelial_in_Epithelium
Other_in_Epithelium
Stem cell
DFS categories
Early recurrence in <3 years (n=120)
Late/No recurrence (n=346)
C
Percentage of cell types in epithelium
D
Median protein level in stem cell population
−Log10(P−value)
Log2 (Early recurrence/Late or No recurrence)
−0.4
0.0
0.4
MLKL
BAX
BAK
CDX2
GLUT1
FLIP
Cellular compartment
Whole cell
Cytoplasm
Membrane
Nucleus

Figure 5

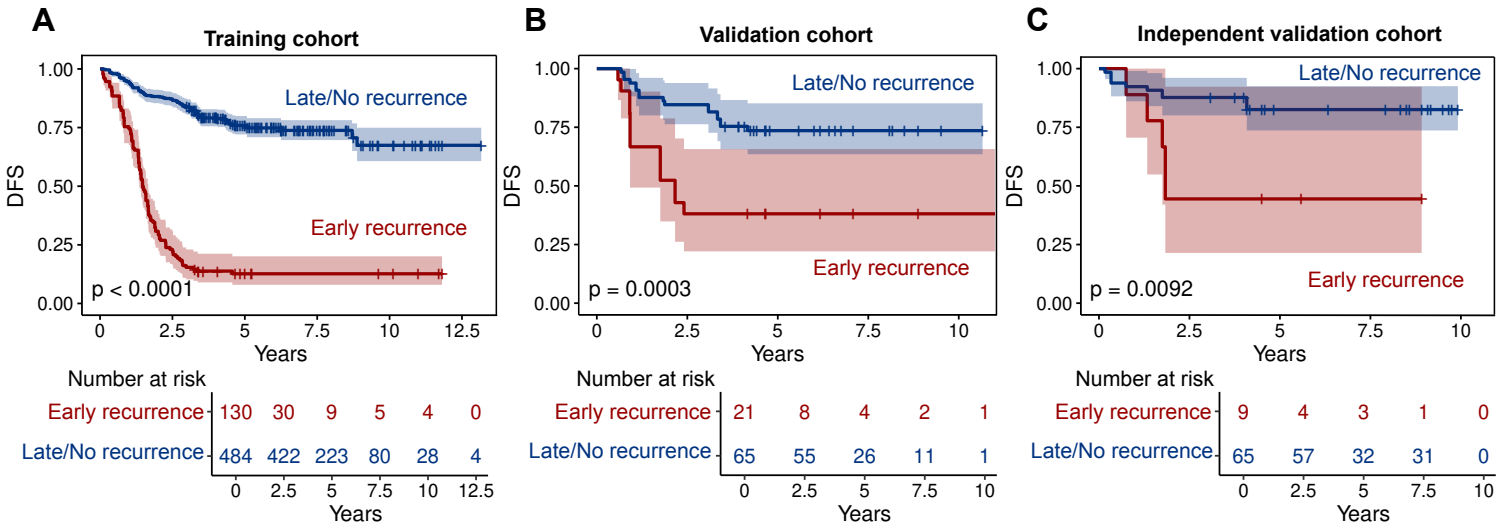
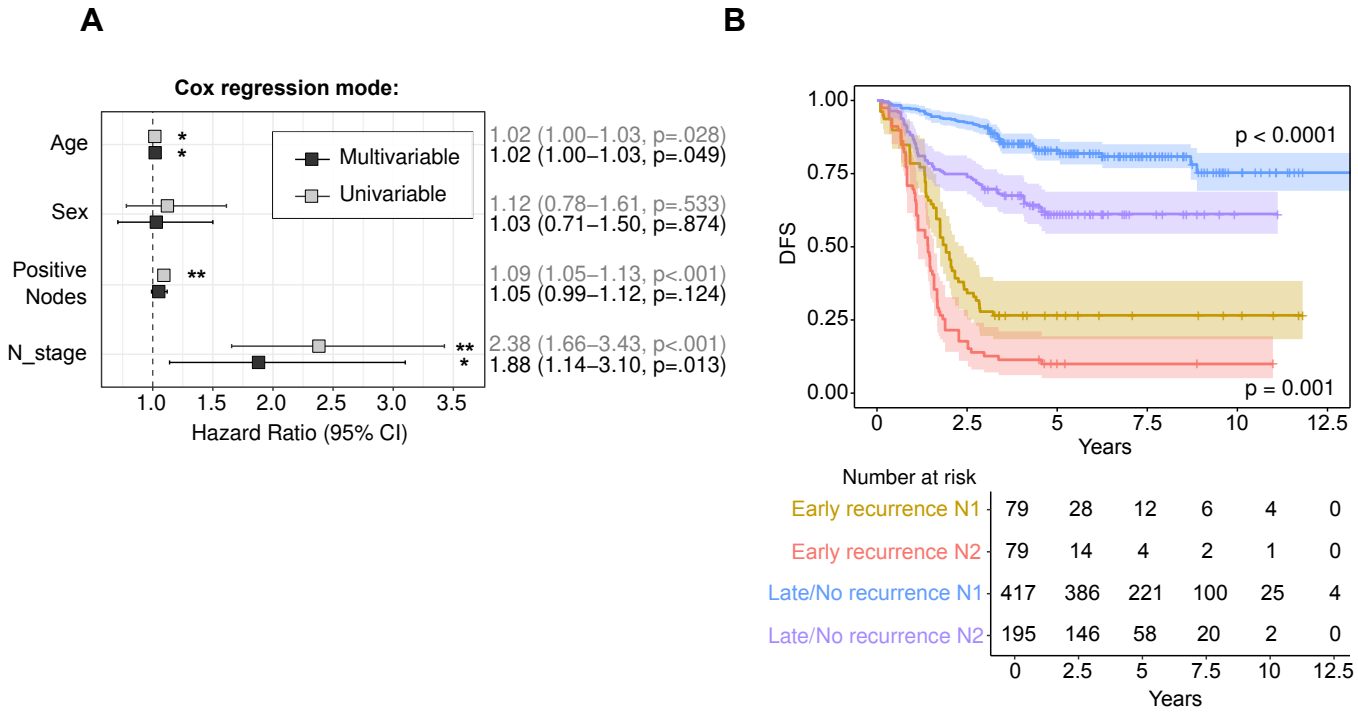

A
Training cohort
DFS
1.00
0.75
0.50
0.25
0.00
Late/No recurrence
Early recurrence
p < 0.0001
0
2.5
5
7.5
10
12.5
Years
Number at risk
Early recurrence 130 30 9 5 4 0
Late/No recurrence 484 422 223 80 28 4
B
Validation cohort
Late/No recurrence
Early recurrence
p = 0.0003
Number at risk
Early recurrence 21 8 4 2 1
Late/No recurrence 65 55 26 11 1
C
Independent validation cohort
Late/No recurrence
Early recurrence
p = 0.0092
Number at risk
Early recurrence 9 4 3 1 0
Late/No recurrence 65 57 32 31 0

Figure 6


A
Cox regression mode:
Age
Sex
Positive Nodes
N_stage
Multivariable
Univariable
1.02 (1.00–1.03, p=.028)
1.02 (1.00–1.03, p=.049)
1.12 (0.78–1.61, p=.533)
1.03 (0.71–1.50, p=.874)
1.09 (1.05–1.13, p<.001)
1.05 (0.99–1.12, p=.124)
2.38 (1.66–3.43, p<.001)
1.88 (1.14–3.10, p=.013)
1.0 1.5 2.0 2.5 3.0 3.5
Hazard Ratio (95% CI)
B
DFS
1.00 0.75 0.50 0.25 0.00
p < 0.0001
p = 0.001
0 2.5 5 7.5 10 12.5
Years
Number at risk
Early recurrence N1 79 28 12 6 4 0
Early recurrence N2 79 14 4 2 1 0
Late/No recurrence N1 417 386 221 100 25 4
Late/No recurrence N2 195 146 58 20 2 0
0 2.5 5 7.5 10 12.5
Years