Self-supervised learning enables unbiased patient characterization from multiplexed cancer tissue microscopy images

Gantugs Atarsaikhan[1,2], Isabel Mogollon[1,2], Katja Välimäki[1], iCAN[2], Tuomas Mirtti[2,3,4,5], Teijo Pellinen[1,2] and Lassi Paavolainen[1,2]

1) Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Finland
2) iCAN Digital Precision Cancer Medicine Flagship, University of Helsinki, Finland
3) Department of Pathology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
4) Research Program in Systems Oncology, University of Helsinki, Helsinki, Finland
5) Finnish Cancer Institute, Helsinki, Finland

Correspondence: gantugs.atarsaikhan@helsinki.fi, lassi.paavolainen@helsinki.fi

Abstract

Multiplexed immunofluorescence microscopy provides detailed insights into the spatial architecture of cancer tissue. Classical analysis approaches focus on single-cell data but can be limited by segmentation accuracy and the representational power of extracted features, potentially overlooking crucial spatial interrelationships among cells. We developed a hierarchical self-supervised deep learning approach that learns feature representations from multiplexed microscopy images without expert annotations. The method encodes tissue samples at both the local (cellular) level and the global (tissue architecture) level. We applied our method to lung, prostate, and renal cancer tissue microarray cohorts to investigate whether self-supervised learning can recognize clinically meaningful protein marker patterns from multiplexed microscopy images. We observed that the learned features identified prognostically distinct patient groups, which show significant differences in survival outcomes. These patient groups matched earlier findings obtained with classical single-cell analysis using expert annotations. Moreover, attention maps extracted from these models highlighted crucial tissue regions that correlate with specific marker combinations. Overall, the approach effectively profiles complex multiplexed microscopy images, offering potential for improved biomarker discovery and more informed cancer treatment decisions.

Introduction

Understanding the heterogeneity of cancer cell states and their surrounding tumor microenvironment (TME) is essential for cancer patient diagnosis and treatment

optimization[1,2]. Interactions between cancer cells and the TME greatly influence cancer development, progression, and response to therapy[3–5]. Previous studies have examined various TME components, such as cancer associated fibroblasts (CAFs)[6–9] and tumor associated macrophages[10–12]. Due to the TME's complexity, studies suggest that targeting only certain parts of it is insufficient for developing optimal cancer treatments[3,13]. Therefore, a comprehensive understanding of cancer cells and their TME is needed to develop effective treatment strategies.

Multiplexed immunofluorescence (mIF) microscopy technique[14–16] is widely used to simultaneously visualize multiple proteins in tissues. This approach is especially useful for providing detailed spatial information on cell types and protein expression levels from complex cancer tissue structures[16–22]. Typical mIF microscopy data analysis relies on cell segmentation[23,24] and measurement of single-cell morphological and intensity features[25], followed by various downstream analyses tasks[16,26], such as identifying cell types, correlating with clinical information, and biomarker discovery. Using single-cell analyses, mIF images have been employed to identify correlations between CAF types and patient survival in lung cancer[19,27], classify cell types in head and neck squamous cell carcinoma[28], and discover a predictor cell types associated with favorable prognosis in breast cancer[29]. These methods rely on single-cell features, which can be limited by segmentation accuracy and sample preparation issues that yield partial or overlapping cells. Therefore, new methods are needed to fully leverage mIF images by incorporating cellular information, local and global spatial context, and cell–cell interactions.

Self-supervised learning (SSL) methods enable unbiased profiling of large image datasets without requiring expert annotations[30]. Recent advances in SSL methods for natural images have achieved impressive performance on downstream tasks[31–35]. However, models pre-trained on natural images are typically not directly applicable to fluorescent microscopy images because of differences in signal distribution and the number of channels. To employ SSL methods for fluorescence microscopy images, Doron et al.[36] and Pfaendler et al.[37] studied DINO SSL framework[31], and Kraus et al.[38] proposed a channel-agnostic SSL framework based on the Masked Autoencoder method[35] for single cell profiling. By successfully applying SSL to cell culture image data, these studies provide a proof-of-concept that SSL can be extended to mIF microscopy images of tissue sections, which we hypothesize will yield similarly valuable insights.

In this study, our goal is to determine whether SSL methods can uncover biologically meaningful marker patterns from large-scale mIF imaging data without *a priori* knowledge. We developed an SSL framework for profiling cancer tissue samples prepared as tissue microarrays (TMA) and imaged by cyclic mIF microscopy. The framework operates on two modular levels: local cellular neighborhoods and global tissue structures. This design is

inspired by the Hierarchical Image Pyramid Transformer (HIPT) method[39], which encodes whole-slide histopathology images into feature representations. We adapted HIPT for mIF microscopy and made each level modular, enabling the use of any SSL approach. We benchmarked four well-known SSL methods Self-DIstillation with NO Labels (DINO)[31], Masked Autoencoders (MAE)[35], A Simple Framework for Contrastive Learning of Visual Representations (SimCLR)[40], and Self-Supervised Learning of Local Visual Features (VICRegL)[32] using three in-house cancer cohorts from lung, prostate, and renal cancers.

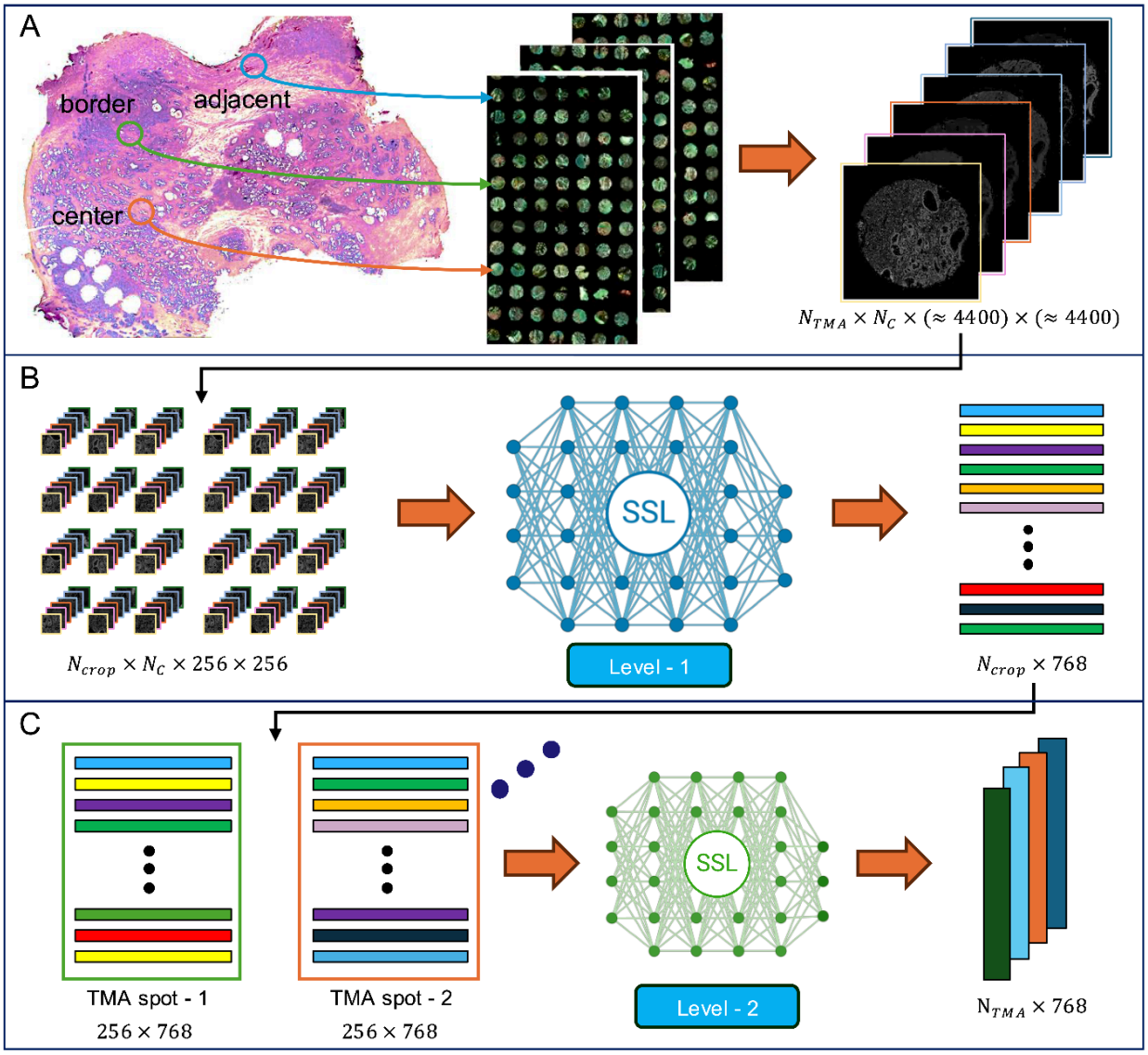

Figure 1. Overview of the proposed framework. A) The study cohorts consist of 7-10 TMA slides, each containing approximately 100 TMA spots. Each spot is cropped into non-overlapping patches (*256x256 px*) representing local cellular neighborhoods. **B)** At *Level 1*, a neural network model is trained using a self-supervised learning (SSL) approach

on these small patches to learn local marker patterns. **C)** At *Level 2*, another model is trained to learn global feature representations from the *Level 1* output features. This is achieved by using aggregated local feature representations to train the *Level 2* model.

Results

Hierarchical self-supervised learning framework

Our goal is to identify marker patterns in mIF-images of cancer tissue samples to discover clinically relevant associations with factors such as cancer grade or prognosis. As cyclic mIF methods can label tens of markers in tissue samples, systematically exploring all possible multimarker combinations is impractical, particularly in high-plex settings. Rather than replicating pathologists' work, we aim to enable novel discoveries by using self-supervised learning approaches to learn these marker patterns in a data-driven manner, without relying on tedious annotations.

We propose a two-level hierarchical SSL framework to extract marker patterns from mIF images (Fig. 1). This framework trains models to extract both local features, capturing cellular neighborhoods, and global features, representing entire TMA spots, without requiring annotations. Furthermore, the framework learns multi-channel relationships in mIF images, facilitating the discovery of marker patterns that may hold significant biological relevance.

We applied our framework to TMA images from lung, prostate, and renal cancer cohorts (Table 1). TMA spots (*diameter≈1mm*) were collected from three distinct regions: *tumor-center* (includes mostly cancer cells), *tumor-border* (includes both cancer and stromal cells), and *tumor-adjacent tissue* (mostly stromal cells). Each TMA spot image is approximately 4,400×4,400 pixels and was stained with six or seven markers (channels) using in-house designed CAF panel[18,19] as shown in Table 1. For local feature extraction, each TMA spot image was subdivided into 256×256 pixel (80×80 μm) patches, each containing several dozen cells. Next, we applied a logarithmic transformation to the patch intensities, compressing long-tailed histograms toward normal distributions to improve SSL efficiency.

The Level 1 model was trained on these patch images using an SSL approach with the aim of capturing hidden patterns within and between the markers. Once the Level 1 model was trained, we used it to extract local feature representations of the patch images. For Level 2, we trained another SSL model having the local feature representations of TMA spots as input feature matrices. The Level 2 model was used to extract global feature

representations from whole TMA spots. Finally, patient-level feature representations were created by averaging the global feature representations across TMA spots of the same region (tumor center, tumor border, or tumor adjacent) within each patient. Detailed description of the methodology is presented in the Methodology section, and the source code of the framework is available on GitHub (https://github.com/bioimage-profiling/SSL-Multiplexed-Imaging).

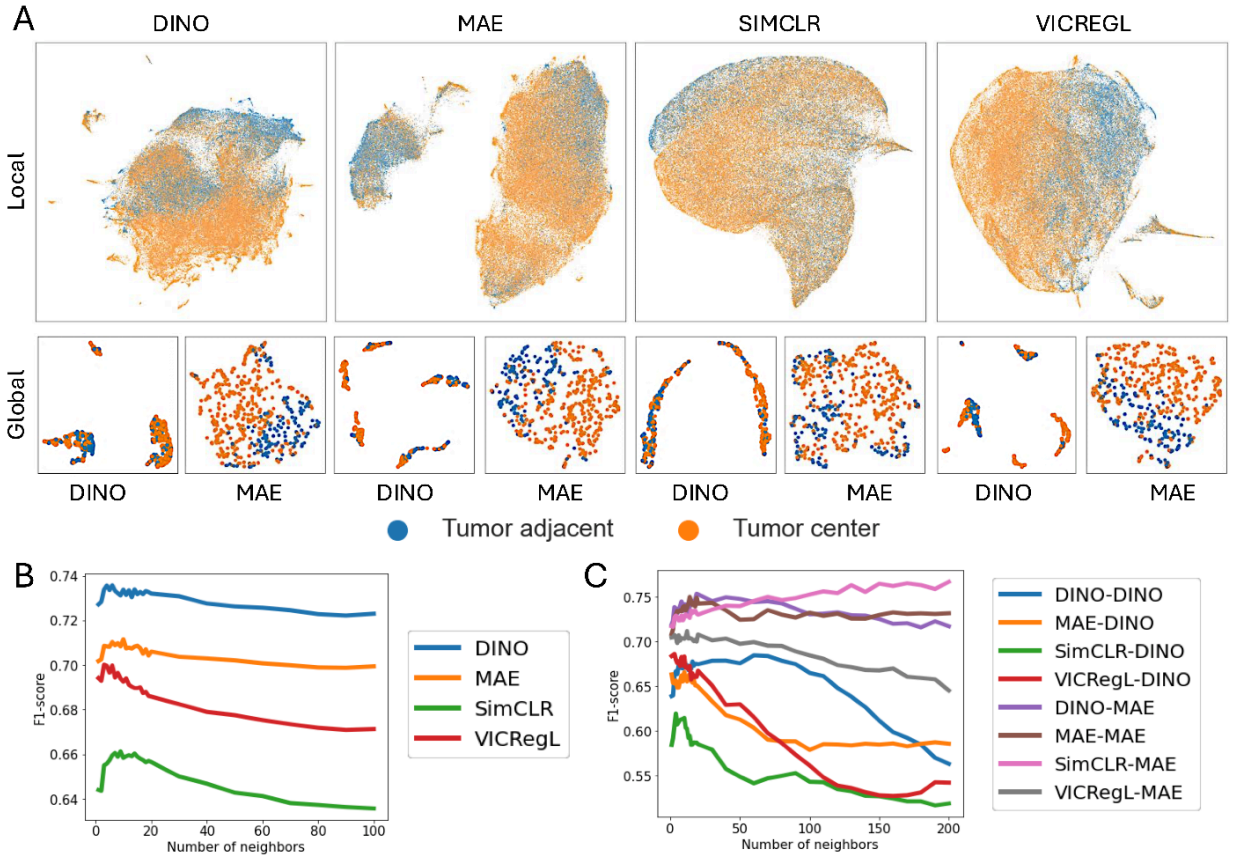

Figure 2. Comparison of self-supervised learning approaches. A) Top row:UMAPs show local feature representations of DINO, MAE, SimCLR, and VICRegL SSL methods when trained on the prostate cancer dataset. Each dot represents a patch from tumor-center (orange) or from tumor-adjacent (blue) TMA spot. Bottom row: UMAPs show global feature representations of combinations using DINO or MAE as a Level 2 method on the corresponding local feature representations. **B, C)** The graphs show the F1-scores for the tumor region classification task using k-nearest neighbor (kNN) as *k* is gradually increased. Tumor-center is considered as the positive category. We used a leave-one-patient-out strategy, meaning that the same patient was excluded from evaluation. The scores are calculated on equally sampled subsets from each class to mitigate class imbalance. **B)** shows F1-scores using local feature representations and **C)** shows F1-scores using global feature representations.

Self-supervised learning methods for multiplexed immunofluorescence microscopy

The proposed framework is designed as modular so that various SSL methods can be used in Level 1 and Level 2. We experimented with four SSL strategies on Level 1: DINO, MAE, SimCLR, and VICRegL; and with two SSL strategies on Level 2: DINO and MAE.

First, we compared local feature representations of the four approaches with the prostate cancer dataset. All of the approaches learn similar features that can differentiate between TMA spot regions, such as *tumor-adjacent* and *tumor-center*, although significant overlap exists between the regions (Fig. 2A). All methods result in relatively high F1-score (0.63–0.74) but lower specificity score (Suppl. Fig. S1A) when tested with class-balanced k-nearest neighbor (kNN) classification while setting the *tumor-center* as positive (Fig. 2B). We hypothesize that the models have high confidence in classifying *tumor-center* patches correctly since patches from those spots contain mostly cancer cells. However, the models are not as confident in classifying *tumor-adjacent* patches correctly due to similar cancer cells included in both regions. Moreover, F1-scores were larger with lower number of neighbors in kNN classifications, indicating that feature representations of patches from the same region are closer together in the feature dimension space. DINO and MAE result in the highest F1-scores, whereas SimCLR, arguably, results in the lowest F1-score and specificity (Fig. 2B; Suppl. Fig. S1A).

Next, we compared global feature representations of prostate cancer dataset using DINO and MAE as the Level 2 SSL methods. The feature representations show more distinct grouping when the Level 2 method is DINO rather than MAE (Fig. 2A). All eight combinations are still able to differentiate between TMA spot regions. Using MAE as Level 2 method, F1-scores are systematically similar with the increasing numbers of neighbors, indicating there are no clear differences between TMA spots that are captured (Fig. 2C). Contrary, when DINO is used as the Level 2 method, the kNN F1-scores are decreasing with increasing number of neighbors, indicating improved grouping of similar TMA spots. DINO-DINO combination (Level 1: DINO and Level 2: DINO) is able to distinguish between tumor-centers, however, it struggles to correctly classify tumor-adjacent spots with low specificity (Suppl. Fig. S1C). MAE-DINO and VICRegL-DINO combinations are better at classifying both tumor-center and tumor-adjacent spots.

We then looked into batch effects in the feature representations (Suppl. Fig. S1B, D). The prostate cancer dataset includes nine TMA slides with approximately hundred spots each. We defined each TMA slide as a separate batch and calculated kNN scores in two ways: leave-one-slide-out strategy, by ignoring patches or spots from the same slide when testing kNN, and without introducing any restrictions. We studied the difference between kNN

scores of leave-one-slide-out and no-restriction classifications. In Level 1, SimCLR local feature representations show the smallest batch effects, although it also had the lowest scores on classification tasks (Suppl. Fig. S1B). Other methods suffer similarly from batch effects, however, having higher scores on classification tasks. In Level 2, kNN score differences are small compared to Level 1 with VICRegL-DINO and MAE-DINO combinations having the smallest difference (Suppl. Fig. S1D). In general, the differences between restricted and non-restricted classifications are negligible, indicating that TMA slide level batch effect is minimal.

Lastly, we determined the Level 1 model training epochs based on the UMAP visualizations of feature representations (Suppl. Fig. S2). Stopping the training too early (epoch=10) led to incomplete learning of features that appeared as a structure without any visual clusters. On the other hand, visualizations from later epochs (epoch=200) showed that patches from the same TMA spots were grouping together strongly which appeared as an extreme batch-effect. This indicates that the model learned features only specific to each TMA spot rather than meaningful representation of the whole dataset. We selected models after an epoch when the UMAP visualization started to overfit to individual TMA spots. Moreover, feature representations from these models produce more meaningful patient groups compared to those from the overfitted models, which do not create significantly different patient groups (Suppl. Fig. S9).

Overall, the local feature representations generated by the DINO method were the most expressive, followed by those from the MAE method. However, at Level 2, methods trained using MAE local feature representations produced more meaningful groupings. The Level 2 models trained on VICRegL local feature representations as inputs were also effective, although VICRegL local feature representations were less expressive. Thus, for further analysis, we used feature representations of the MAE-DINO combination.

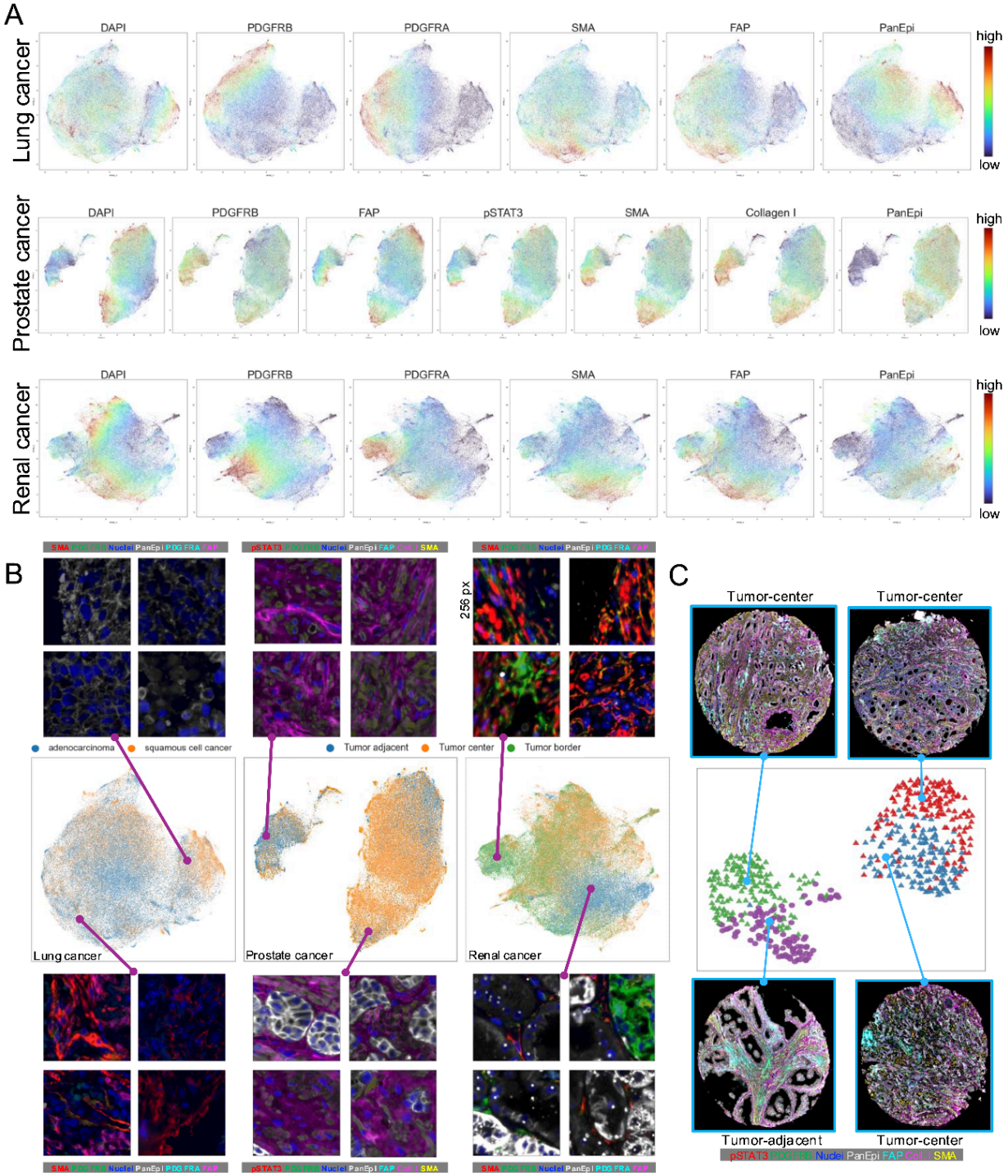

Figure 3. Local feature representations. A) UMAPs showing local (patch-image level) feature representations of three cancer cohorts by using MAE method. The colormaps on the UMAPs show the mean intensity value of each patch image for each channel. **B)** The same UMAPs are colored with tumor region in case of prostate and renal cancer, and histology in case of lung cancer. Selected sample images from each cohort are shown above

and below the UMAPs. **C)** Feature representations of all patches from one patient in the prostate cancer cohort. Feature representations are plotted in two groups by UMAP, each containing patches from two TMA spots.

Local neighborhood representations encode fluorescent marker intensities

The lung, prostate, and renal patch image datasets consist of around *100k*, *125k*, and *180k* images, respectively (Table 1). Lung and renal cancer patch images include six channels, while prostate cancer patch images include seven channels of fluorescent markers. Patch images have a size of *256 x 256* pixels (*80 x 80 μm*) and include a few dozen cells each. In Level 1 of our framework, we train a model on each cancer dataset separately, and extract local feature representation for each patch image with the aim of identifying distinctive multimarker patterns in cellular neighborhoods. We assess the quality of local feature representations using the MAE method.

First, we checked how marker intensities are encoded in the local feature representations. We created two dimensional UMAP embeddings from patch feature representations and studied the distribution of average marker intensities of the patches in these plots (Fig. 3A and Suppl. Fig. S3). Patches with similar average marker intensities (e.g., low DAPI, mid PDGFRB, high Collagen I, and low PanEpi) have similar local feature representations and are visualized together. Specifically on the PanEpi channel, the division between low average intensity and high average intensity patches is clear, although other channels do not follow this division. This indicates that the Level 1 model was able to learn patterns between multiple channels.

Next, using the same UMAP embeddings, we confirmed that patch feature representations are able to distinguish between TMA spot regions for the prostate and renal cancer datasets (Fig. 3B). In addition, patches from different regions share considerable overlap in the embedding space with low silhouette score (0.09 for the prostate cohort and 0.01 for the renal cohort). We hypothesize that tumor-center and tumor-adjacent tissue spots from the same patient share similar patterns. All TMA spots in the lung cancer cohort were sampled from the tumor tissue, thus we studied the patch embeddings by sample histological subtypes that are visibly divided in the UMAP.

Lastly, we checked all local feature representations from one randomly chosen patient with prostate cancer (Fig. 3C). The cohort includes one tumor-adjacent TMA spot and three tumor-center TMA spots from this patient. The UMAP embedding shows that local feature

representations from the same TMA spots are more similar than from other TMA spots. However, visual inspection of the TMA spots showed that two similar looking tumor-center TMA spots are plotted together. The third tumor-center TMA spot has a distinctive look with a high amount of epithelial cells. Overall, Level 1 models successfully encoded multi-channel patch images into patch feature representations showing distinct characteristics.

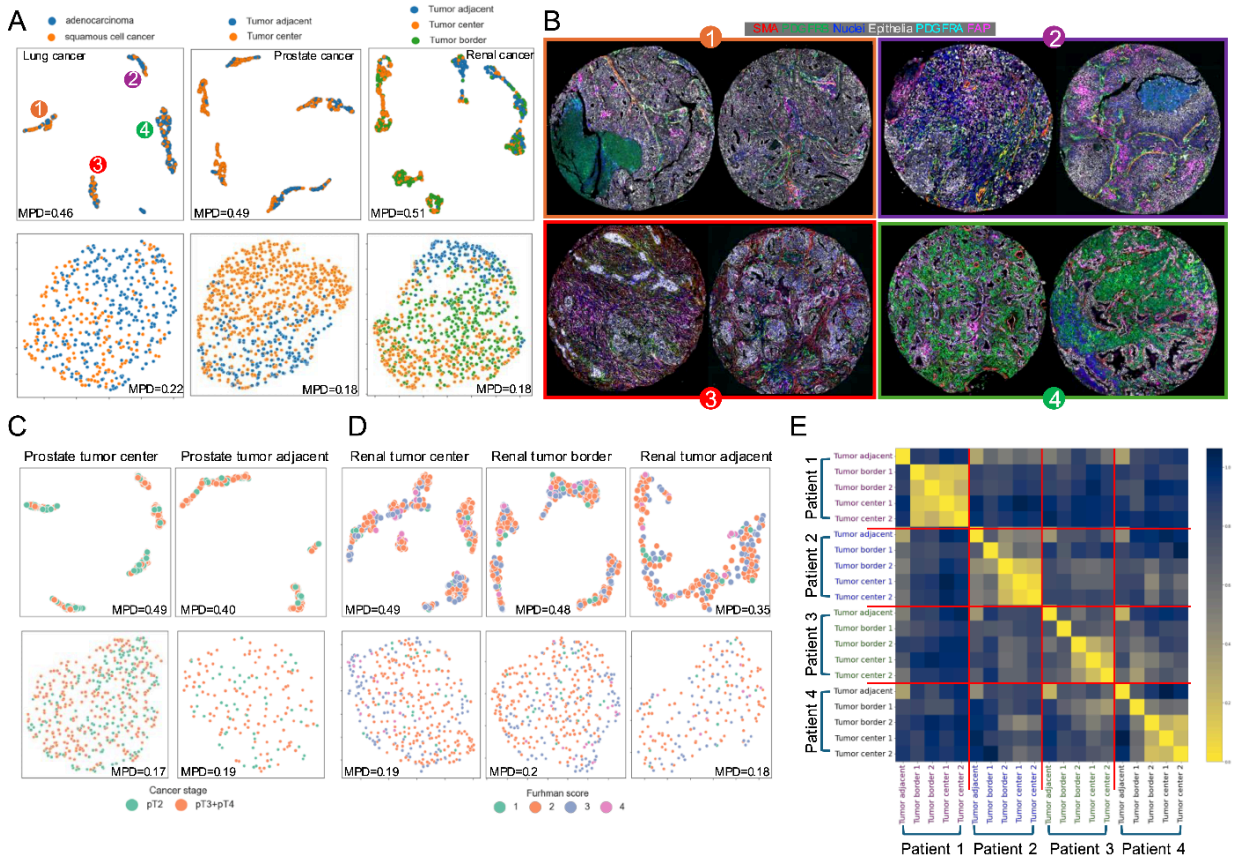

Figure 4. Global feature representations. For each representation, the normalized Mean Pairwise Distance (nMPD) score is calculated. The nMPD score is close to 0 when the embeddings are almost identical, and close to 1 when the embeddings are highly diverse and lack any formed clusters. **A)** Top row shows global feature representations (TMA spot-level) for each cancer type with MAE-DINO model. Bottom row shows global feature representations created by averaging all local feature representations of each TMA spot. UMAP embeddings are colored by histology for lung cancer, and by spot region for prostate and renal cancer. **B)** Example TMA spots from different clusters of global features from the lung cancer cohort. **C, D)** In the upper row, region-wise feature representations from the model are shown. Lower row shows region-wise averaged global feature representations for prostate and renal cancer. Dots in the UMAPs are colored by cancer staging for prostate cancer, and aggressiveness for renal cancer. **E)** Cosine distances between global feature

representations from four randomly chosen renal cancer patients. Low cosine distance means the features are similar to each other.

Global TMA spot feature representations encode sample types

Input to Level 2 model is aggregated feature-images created from local features for each TMA spot. The dataset contains approximately 500, 800 and 1200 TMA spots in lung, prostate, and renal cancer cohorts, respectively. We trained the DINO SSL approach at Level 2 using rearranged forms of local feature representations from the Level 1 MAE output (Fig. 4).

Even though Level 2 models are not trained with the original images, the models are able to learn meaningful feature representations for every TMA spots (Fig. 4A). In the case of lung cancer dataset, visually similar TMA spots are grouped together in the UMAP space (Fig. 4B). The model also encoded the majority of the adenocarcinoma samples with similar feature representations. Prostate and renal cancer feature representations are separated by the spot regions. We measured the overall separation of the feature representations using the normalized Mean Pairwise Distance (nMPD), which ranges from 0 to 1: 0 if the embeddings are identical and 1 if each embedding is a cluster. As a baseline result to compare with the Level 2 model outcome, we created TMA spot embeddings by averaging the local feature representations from Level 1 (bottom row of Fig. 4A). There were negligible distinctions between TMA spots using this approach with nMPD scores from 0.18 to 0.22 depending on the cohort. In contrast, the Level 2 SSL approach is able to learn and enhance minor distinctions in the data with nMPD scores from 0.46 to 0.51. To study regions independently, we created UMAP embeddings of global feature representations of each region separately (top row of Fig. 4C,D). In general, tumor-adjacent feature representations are less distinctive (nMPD=0.40 for prostate, nMPD=0.35 for renal) compared to those from other regions (nMPD=0.48–0.49, depending on the cohort and region), yet still more distinct than the averaged Level 1 embeddings (nMPD=0.2 at most). However, none of these groupings showed visually clear clusters when annotated with patients' clinical information (Suppl. Figs. S4, S5, S6).

Finally, we calculated cosine distance between global feature representations of randomly selected patients with renal cancer to study similarity of the feature representations (Fig. 4E). Mean distance (±standard deviation) between all pairs was 0.80±0.17. As expected, feature representations of tumor-center spots from the same patient were significantly more similar to each other (0.17±0.13) than to those from different patients (0.82±0.05). Tumor-adjacent spots were uniquely dissimilar to tumor-center (0.85±0.14) and tumor-border (0.82±0.15) TMA spots, even within the same patient, and were only loosely correlated with tumor-adjacent spots from other patients (0.61±0.23). The similarity of

tumor-border spots varied: in some patients, tumor-border spots were correlated with tumor-center spots (Patient 1 and 2 in Fig. 4E), while in others, they showed no correlation with any region (tumor border 1 in Patient 3 in Fig. 4E). Because of this heterogeneity between spot regions even from the same patients, we decided to calculate patient-level feature representations separately for each region.

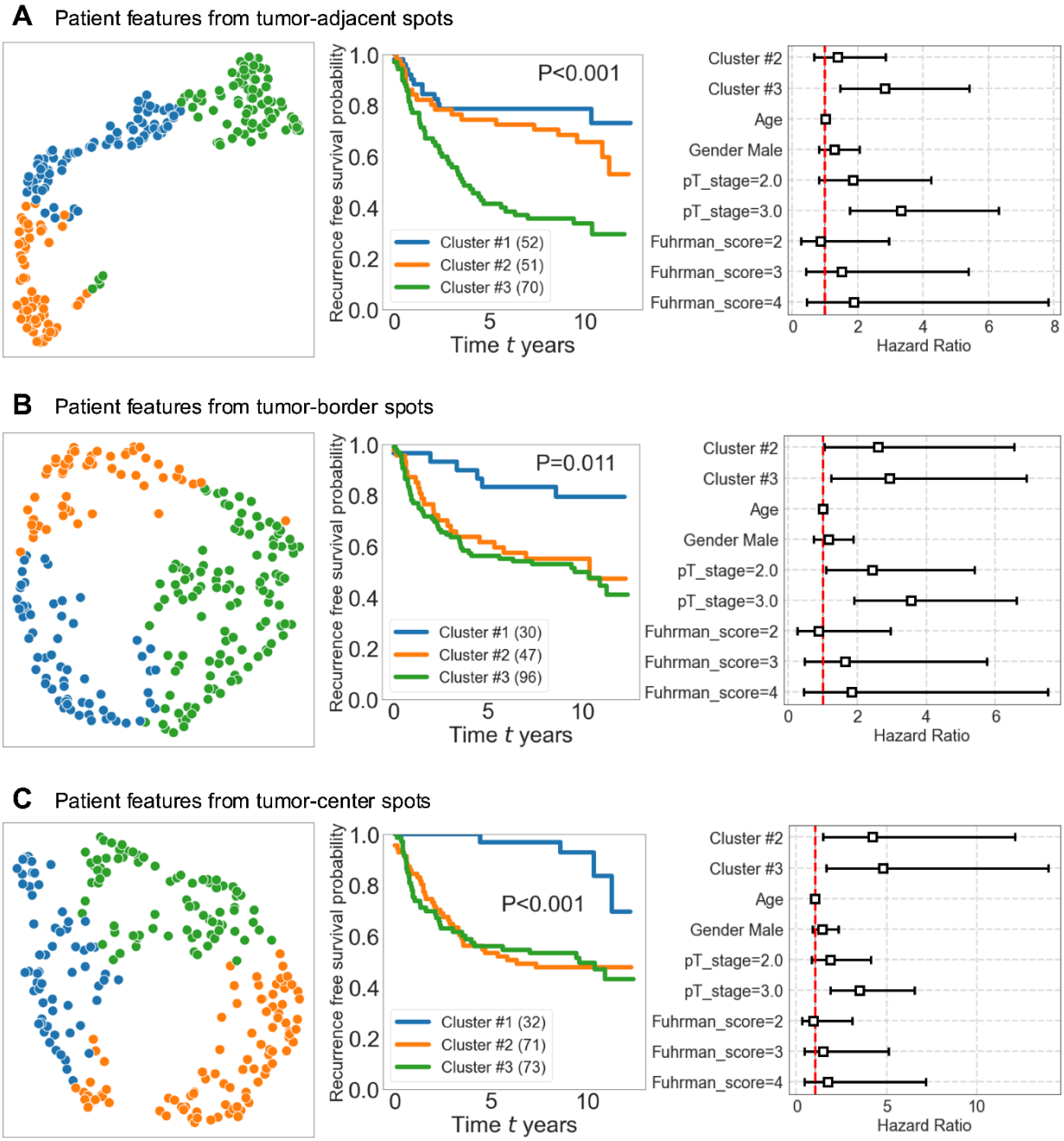

Figure 5. Cancer-free survival analysis of clear cell renal cell carcinoma stratified by patient-level feature representations. Patient-level feature representations were created individually by averaging all global level features from the same region: **A)** Tumor-adjacent

B) Tumor-border **C)** Tumor-center. We clustered the original high dimensional feature representations with k-means clustering using the elbow method to determine the number of clusters. P-values were calculated using the Multivariate Log-Rank Test. The color code in UMAP graphs presents the cluster ID. We studied patient survival or recurrence using Kaplan-Meier plot and calculated the hazard rate using Cox proportional hazard regression.

Feature representations uncover patients with clearly distinct prognosis

To link the feature representations with clinical information, we created patient-level feature representations by averaging the global features for each patient, separately for each region: tumor-adjacent, tumor-border, and tumor-center. Then, the k-means clustering method was used for patient categorization. Number of clusters was determined by the elbow method[41], which coincidentally gave *k=3* for all experiments. We associated the patient groups with overall survival for lung cancer, recurrence-free survival after radical prostatectomy for prostate cancer, and cancer-free survival for renal cancer. Survival probabilities were confirmed using Kaplan-Meier plots and validated with hazard rates from the Cox proportional hazards model (Fig. 5, Suppl. Fig. S7). Detailed clinical information on all patient clusters can be found in the supplementary materials (Suppl. Table. S2-S7).

We performed survival analyses for patient groups from clear cell renal cell carcinoma (ccRCC) (Fig. 5). Using tumor-adjacent spots of ccRCC, three distinct groups were identified by clustering the feature representations. Patients in the cluster#3 had the lowest survival probability. In this cluster, nearly 60% of the patients experienced relapse after initial treatment (19% in cluster#1 and 32% in cluster#2), of which 51% occurred in the lung (Suppl. Table S4). On the other hand, most patients in cluster#1 had a good prognosis with the smaller tumor without relapse. In similar manner, patients in tumor-border cluster#1 and tumor-center cluster#1 had a good prognosis. There was no difference between gender within the clusters, except in the tumor-center cluster#1, 69% were men. Even though there was no significant difference in cancer aggressiveness as indicated by the Fuhrman score, the model identified clusters with distinct prognoses demonstrating the capability of the proposed framework to extract information from mIF microscopy images not encoded in the Fuhrman score.

Lung cancer cohort was separately studied between the two main subtypes: adenocarcinoma and squamous cell carcinoma. Our framework identified three different clusters. Cluster#3 from lung adenocarcinoma patients showed significantly lower survival probability (Suppl. Fig. S7A,B). This cluster consists of more *current smokers* and have poorer *performance status scores* according to the World Health Organization (PS-WHO)[42]

criteria compared to the other two groups (Suppl. Table S2). In the lung squamous cell carcinoma patient subgroup, cluster#3 had the lowest and cluster#1 had the highest survival probability. Patients in these groups have similar clinical background, except cluster#3 patients tend to be younger (67 or lower), and more than half have the PS-WHO score of 1 which means fully ambulatory with some symptoms. In comparison, more patients did not have any symptoms (PS-WHO=0) in cluster#1 and cluster#2. At the last recorded time of follow-up, 87% of the cluster#3 patients were deceased compared to 60% of patients in the cluster#1.

For prostate cancer patients survival analysis, we considered time from radical prostatectomy (surgical removal of prostate) until status at the time of follow-up or year of cancer-related death. The framework again identified three distinct clusters from the feature representations of both tumor-center and tumor-adjacent regions. Survival probability using tumor-adjacent samples show a large difference between cluster#1 and cluster#3 patients (Suppl. Fig. S7C). More patients in tumor-adjacent cluster#3 have higher biopsy Gleason and RP/TURP Gleason scores, which may be associated with poor prognosis. The results from tumor-center samples show that patients from cluster#1 tend to have better survival probability than patients in other clusters, though the difference is not statistically significant (Suppl. Fig. S7D).

Furthermore, we examined how much the same patients cluster together in different regions. We calculated adjusted mutual information (AMI) scores for cluster labels from different regions of prostate and renal cancer (Suppl. Fig. S7E,F). The AMI score equals 1.0 if exactly the same patients are clustered together in two different experiments and is close to 0.0 if they are randomly distributed. Clusters from the tumor-adjacent and tumor-center region features include different sets of patients, leading to low AMI scores in both prostate (0.089) and renal cancer cohorts (0.064). The AMI score is the highest (0.23) between tumor-border and tumor-center feature representations in renal cancer, indicating that the features encode similar information and cluster the same set of patients together.

Lastly, we concatenated feature representations from all regions to check effectiveness of the collective feature representations (Suppl. Fig. S8). Concatenated patient-features of prostate cancer also clustered into three groups with the elbow method (Suppl. Table. S6). Patients in cluster with poor prognosis (cluster#3) had high Gleason scores as in the region-wise split features of the prostate cancer cohort. Concatenated renal cancer patient-features were also clustered into three groups with significant difference in survival probability between cluster#3 and other two clusters (Suppl. Fig. S8). In the poor prognosis cluster#3 of the renal cancer cohort, nearly half of the patients (49%) had relapsed in the lung compared to less than 15% in the other two clusters. Moreover, more than half of the

patients in cluster#3 had an invasive tumor (pT_simple=3). Patients with good prognosis (cluster#1) had smaller tumors without relapse.

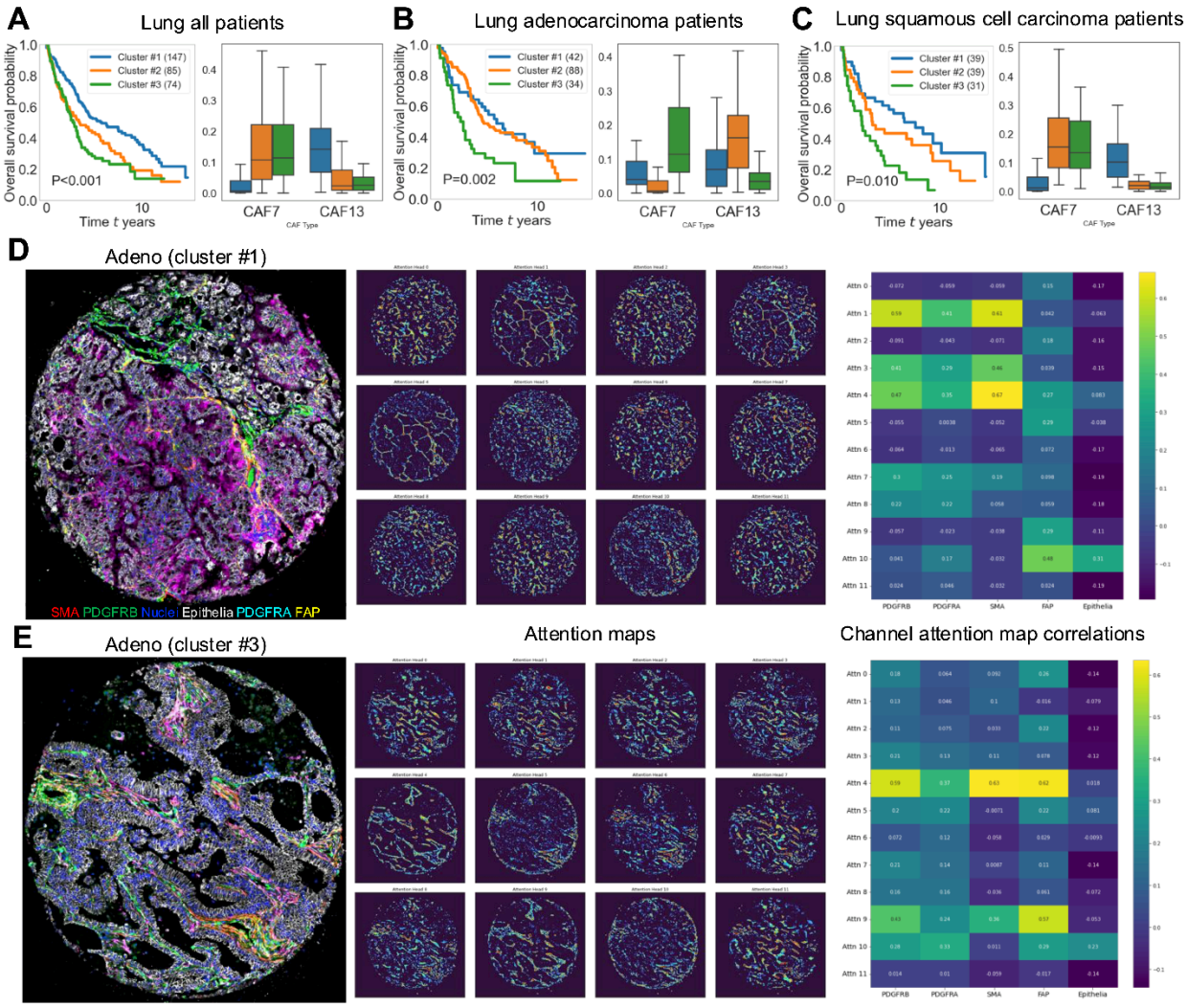

Figure 6. Associations between patient clusters and cancer associated fibroblasts (CAFs). A-C) Kaplan-Meier plots for patient clusters determined from patient-level feature representations, with the corresponding CAF ratio (ratio of CAF subtype and all CAFs) in each cluster for the lung cancer cohort. Subsets CAF7 (PDGFRB+, PDGFRA-, SMA+, FAP+) and CAF13 (PDGFRB+, PDGFRA+, SMA+, FAP-) were shown to be prognostic in an earlier study. **A)** Patient clusters from the entire lung cancer cohort, **B)** clusters from lung adenocarcinoma patients, and **C)** clusters from lung squamous cell carcinoma patients. **D)** and **E)** show the original composite images of spots, their corresponding attention maps, and the Pearson correlation between individual channels and the attention maps. Attention maps show regions that are concluded important by the model. Example in **D)** is a sample

from good prognosis patients, whereas example in **E)** is a sample from poor prognosis patients in **A)**.

Interpretation of identified patient clusters by CAF associations and attention maps

Lastly, we studied the patient clusters discovered by the hierarchical SSL method by comparing the results to previously published cancer-associated fibroblast (CAF) subsets (Suppl. Table. S1) determined using single-cell segmentation and expert-defined cellular positivity [18,19]. CAF enrichments were defined as the ratio of CAF type cell count to the total CAF cell count of any type. We found that CAF7 (PDGFRB+, PDGFRA-, SMA+, FAP+) is enriched in patient groups with poor prognosis, and CAF13 (PDGFRB+, PDGFRA+, SMA+, FAP-) is enriched in patient groups with good prognosis for both histologies in lung cancer cohort (Fig. 6A-C). These findings match perfectly with previously published single-cell results [18,19]. In the renal cancer cohort (Suppl. Fig. S10), CAF associations vary between spot regions, although CAF13 is associated with good prognosis regardless of the spot region. CAF8, CAF12, and CAF15 are also associated with good prognosis in tumor-border and tumor-center regions whereas CAF4 from the tumor-adjacent region is enriched in the patient group with poor prognosis. Some of the CAFs were inconclusive, such as CAF12 from tumor-adjacent region, which was enriched in both good and bad prognosis patient groups.

One advantage of using a ViT architecture is the ability to visualize attention maps, which highlight regions in the input image where the model focuses on. This feature is useful for interpreting the model's behavior. A ViT model can generate multiple attention maps depending on the number of self-attention mechanisms employed. We used ViT-base architecture with 12 attention heads in the last transformer layer. A sample from lung adenocarcinoma patients with high survival probability (Fig. 6D) shows attention maps 1 and 3 to be similar to the good prognosis associated-CAF13 subset, showing a higher correlation with PDGFRA, PDGFRB, and SMA, and a lower correlation with FAP and PanEpi. There is no similar pattern with poor prognosis associated-CAFs, except for attention map 4 that is similar to CAF15. An example from patients with low survival probability (Fig. 6E) associates attention maps 4 and 9 with poor prognosis associated-CAF7 with low PDGFRA correlation. Attention maps 5 and 6 are similar to CAF14, which is weakly associated with poor prognosis. Moreover, in both examples, only attention map 10 shows a clear correlation with the PanEpi marker, while the other attention maps show low correlation.

In the examples of lung squamous cell carcinoma (Suppl. Fig. S11), PanEpi marker has none to negative correlations with all attention maps, except for a moderate positive correlation with attention map 10. Attention maps 0 and 2 are similar to the bad prognosis

associated-CAF2 and have moderate correlation only with FAP marker, even in an example from the good prognosis group (Suppl Fig. S11A). On the other hand, attention maps 4 and 9 are still moderately correlated with the positive markers of CAF7 in the poor prognosis example (Suppl. Fig. S11B).

Discussion

We developed a hierarchical self-supervised learning framework to uncover unbiased representations from mIF images of cancer TMAs. This framework consists of modular Level 1 and Level 2 models that encode local and global features from mIF tissue images. The global features are further aggregated to patient level representations to study associations with clinical data. This framework enables explorative association analysis of tissue mIF microscopy images without any human bias or annotations.

We applied our framework to lung, prostate, and renal cancer cohorts. We showed that local features generally encode patterns of fluorescent marker intensities. This is beneficial since using all channels (6 or 7) as input to the model allows the model to automatically recognize multi-channel patterns. Our cohorts include markers for nuclei and epithelia (additionally stroma in prostate cancer cohort), which show the general tissue structure (Fig. 3A). The UMAPs of local features show that patches with high amounts of nuclei and epithelial cell markers are plotted together, whereas patches with high stromal markers were plotted separately from the high epithelial marker enabling distinction between epithelial and stromal cells without segmentation. Other markers exhibited more variability in patches, which provided various local marker patterns to explore on the global level.

Even though trained without any supervision, the models revealed clusters of patients that present significantly different prognosis in the survival analysis (Fig. 5 and Suppl. Fig. S7). For instance, cluster#1 discovered from the renal cancer cohort tumor center spots showed significantly good prognosis compared to the other two discovered clusters (Suppl. Fig. S10). Inversely, cluster#3 from tumor-adjacent spots in the same cohort showed significantly worse prognosis. In the prostate cancer cohort, identified cluster#1 and cluster#3 from both tumor-adjacent and tumor-center spots indicated different survival probabilities, although the difference was not statistically significant.

Our findings from the lung cancer cohort were in-line with previous work by Pellinen et al.[19]. They showed using classical single-cell analysis approaches that CAF7 and CAF13 subcategories were associated with poor and good prognosis, respectively. The clusters that are automatically discovered with the hierarchical SSL framework showed the good

prognosis patient group has low CAF7 and high CAF13 fractions whereas the poor prognosis patient groups have opposite CAF7 and CAF13 fractions. In the renal cancer cohort, we found that CAF8, CAF12, and CAF13 are associated with good prognosis confirming previous work on clear cell renal cell carcinoma study[18]. However, most of the other CAFs showed no clear association with either good or poor prognosis.

Moreover, the model focused on those specific CAFs through attention maps that highlight relevant areas within the TMA spot images, indicating where the model is focusing on. The ability to visualize attention maps from the ViT models greatly enhances the interpretability of our framework. In the examples (Fig. 6D,E and Suppl. Fig. S11, S13), attention maps do not correlate with the PanEpi marker. Moreover, for both adenocarcinoma and squamous cell carcinoma subtypes, attention maps 4 and 9 correlate with the bad prognosis CAFs, and attention map 13 correlates with the good prognosis CAFs. However, assignments of attention are specific to a model, and need to be studied separately for each model.

One limitation of our framework is that the models are trained on specific cancer cohorts and may not generalize well to unseen data. To analyze different datasets, the models should be trained from scratch with the new dataset. As the mIF images are extremely complex with practically an unlimited number of different fluorescent marker combinations, currently no fully generalized foundation model exists for mIF image data. Furthermore, the attention maps presented in the study are all from the Level 1 models that learn local representations from the images. It is difficult to connect the attention maps from Level 2 models to the images because the reshaping process of local features modifies the location information. One potential direction for development is to fine-tune the SSL trained models using specific clinical information, such as disease severity, to assess how the model and attention maps evolve in response to this additional supervision.

In conclusion, this study demonstrates the potential of self-supervised learning for the analyses of mIF images of cancer tissue cohorts. We have developed a framework that effectively captures complex fluorescent marker patterns in tumor samples without supervision, and enables associating these patterns with clinical information, such as survival probability.

Methods

Samples

Our multiplexed image dataset is prepared in-house using human cancer tissue samples. We used the following renal (kidney), lung, and prostate cancer cohorts in this study.

Renal cell carcinoma cohort: Formalin-fixed, paraffin-embedded (FFPE) surgical specimens from 236 treatment-naive patients with localized (N0M0) ccRCC who underwent partial or radical nephrectomy at Helsinki University Hospital between 2006 and 2013 were utilized in this study. Patients with distant metastases (M1), regional lymph node metastases (N1), a history of kidney cancer, or multiple kidney tumors at diagnosis were excluded. These specimens and the corresponding tissue microarrays (TMAs) had been previously collected and constructed by Pellinen et al.[18]. The TMAs included two cores from the central tumor (1.0 mm), two from the tumor border, and two from adjacent benign kidney tissue. Digital slides were scanned and annotated for TMA construction. Two pathologists selected representative tissue blocks. Clinical and pathological data—including TNM stage, tumor grade (4-tier Fuhrman), sarcomatoid/rhabdoid differentiation, necrosis, and survival outcomes—were collected from medical records. The clinical association analyses were carried out including only the clear cell carcinoma histology (n=179). The follow-up cutoff date was September 9, 2019.

Lung carcinoma cohort: The FFPE TMA cohort BOMI1 construction[43,44] and the mIF staining/imaging[19] has been previously described. Briefly, the cohort consisted of non-small cell lung carcinoma patients surgically resected at the Uppsala University between 1995 and 2005. Briefly, all specimens were reviewed by board-certified pathologists and representative areas were encircled on tissue slides before cores were taken from corresponding tissue blocks and incorporated into recipient blocks. All tumor samples were included in duplicates (2 × 1 mm tissue cores). Analysis was performed on 319 patient cases with high quality mIF image data produced by Pellinen et al.[19]. Tumor histologies represented adenocarcinoma (n=168), squamous cell carcinoma (n=111), and other (n=40).

Prostate cancer: The prostate cancer cohort consisted of 274 patients with Grade group 2-4 localized cancer treated by open or robot-assisted-radical prostatectomy at Helsinki University Hospital (HUS) between the years 1992 and 2015. Tumor center and adjacent benign samples were from 274 and 220 patients, respectively. Clinical association analyses were carried out with 173 and 148 patients for tumor center and adjacent benign areas, respectively. This cohort's TMA design has been previously characterized by Lehto et al.[45].

Cyclic multiplexed immunofluorescence and imaging

The widefield microscopy imaging was done with two rounds of staining of fluorescent markers, and washing the TMA spots for the next staining round. In total, we used six fluorescent markers for renal and lung cancer TMAs, and seven fluorescent markers for prostate cancer TMAs. In addition, the nuclei marker is used in both rounds of imaging to enable image registration between cycles.

Dataset

TMA spot image dimensions (*~4000x4000 px*) do not fit into GPU memory during the training process when using a large batch size required by self-supervised learning methods. Thus, we cropped the TMA images into small patches of *256x256 px*, while preserving the number of channels. Small patches were treated as local cellular neighborhoods of a few dozen cells. Then, we converted the patches to logarithmic scale to manage the wide range of raw pixel values. In all our TMA spot images, the raw values are in the range of 0–65535 (16-bit). By converting to logarithmic scale, we avoided specifying any cut-off thresholds and were able to transform the long-tail histogram. We hypothesized that any unnecessary artifacts would have similar feature representations from the self-supervised learning phase, so it would be fairly simple to discard artifacts in post-processing, thus image artifacts were not removed beforehand. The details of the datasets are presented in Table 1.

Table 1. Details of TMA datasets.

Cancer Type	**Lung Cancer**	**Prostate Cancer**	**Renal Cancer**
TMA spot regions	Tumor-center	Tumor-center, Tumor-adjacent	Tumor-center, Tumor-border, Tumor-adjacent
Number of patients	319	299	237
Number of TMA spots	569	827 (585 / 242)	1225 (501 / 481 / 243)
Number of Patches	99,106	125,631	180,671
Number of Channels	6	7	6
Marker Names	DAPI, PDGFRB, PDGFRA, SMA, FAP,	DAPI, PDGFRB, FAP, pSTAT3, SMA,	DAPI, PDGFRB, PDGFRA, SMA, FAP,

	PanEpi	Collagen I, PanEpi	PanEpi

Level 1 training

Fig. 1 summarizes our proposed framework. Our framework consists of two levels of self-supervised training and aggregation to patient level features. In Level 1, we selected four approaches SimCLR, DINO, MAE, and VICRegL. SimCLR has two encoders to process two different transformations of the same input, each producing a feature representation. It then calculates contrastive loss that minimizes the distance with a weighted cross-entropy loss between feature representations, and updates the encoder model weights. DINO has teacher and student encoders, and the student encoder tries to learn to match the output of the teacher encoder, and updates the weights accordingly. The teacher encoder is updated with regard to the student encoder. VICRegL has three main objectives: increasing variance between images, maintaining invariance between feature representations of the same image, and minimizing covariance between feature dimensions to avoid redundant features. In MAE, a large portion of the input image is masked (70-90%), and only visible parts are fed into an encoder. Then it reconstructs the image with a decoder. The model weights are updated with a reconstruction loss between the original image and the reconstructed image.

In Level 1, we used ViT[46] models as encoders for DINO and MAE approaches. Specifically, we used the ViT-Base model architecture with 12 attention heads. Patch embedding module was modified to accept six channel images in case of lung and renal cancer datasets, and seven channel images in case of prostate cancer dataset. MAE uses a similar VIT decoder network that outputs the reconstructed image. In SimCLR and VICRegL approaches we used ResNet-50[47] architecture. ResNet-50 has 50 convolutional layers with residual connections between layer blocks. The first convolutional layer was modified for input images with 6 or 7 channels. The size of feature representations from ViT-Base models is 768, compared to 2048 from ResNet-50 models. In all our frameworks, our code is based on implementations of LightlySSL[48].

The original augmentations for the self-supervised approaches are designed for natural images, and are not optimized to fluorescence microscopy images. We removed all of the color jitter transformations (*random brightness, contrast, saturation, and hue*), and included horizontal flip, vertical flip, and rotation. Cropping and gaussian blurring were kept. All the augmentations were randomly applied with 0.5 probability. We kept the augmentations of original images as strict as possible to prevent unrealistic transformations in fluorescence microscopy images and to learn more from the existing data. After training the Level 1

model, we fed the patches once more with only center cropping augmentation and extracted the feature representations. Feature representations are class tokens from ViT models or flattened and pooled last layer features from ResNet-50 models. Lastly, feature representations from the same dataset and the same method were L2 scaled.

All approaches were trained using one node of the LUMI supercomputer (https://lumi-supercomputer.eu/) that has four AMD MI250x GPUs. The used batch size was 512 for all methods and datasets. For MAE, the masking parameter was set to 75% of the input. Model training time varied by the approach and dataset, although it took around 12 hours to train 500 epochs with the renal cancer dataset. However, we stopped the training once UMAP visualization of feature representations started showing signs of overtraining, that is patch images from each TMA started to group together.

Level 2 training

Level 2 begins with collecting local feature representations of patches cropped from the same TMA image. Then we created aggregated images for all TMA spots as follows: Suppose a TMA spot has N patches, and the local feature representations are $F_k \in \mathbb{R}^D$, where $k = 1, 2, \ldots, N$, where D is the length of the feature vector. We concatenate these feature representations into a single matrix $A = [F_1, F_2, \ldots, F_N] \in \mathbb{R}^{NxD}$. The height H and width W dimensions are created from the N, as $A \in \mathbb{R}^{NxD} \rightarrow A \in \mathbb{R}^{HxWxD}$ to serve as input for the Level 2 model. We chose $N = 256 = 16x16$ to create aggregated feature-images of $16x16xD$, applying zero padding if the patch count did not reach 256. This step ensured that the Level 2 model can learn relationships among the local feature representations to derive global features for the entire TMA spot. This is inspired by a similar technique proposed by Chen et al.[39] for H&E stained histopathological images. Using the aggregated images, we trained the ViT-base network with DINO and MAE approaches separately. We used 1x1 convolutions during patch embedding, effectively treating each pixel of the aggregated images as a separate token. The augmentations of the DINO approach were similar to that of Level 1 training, except global random crop size was set to 14 and local was 8. For MAE, the mask ratio was the same 0.75 as in Level 1. Both frameworks were trained for 100 epochs that took around an hour. After training, features representing each TMA spot were extracted and L2 normalized.

Patient-level features

In the renal cancer cohort, in total 3 to 6 TMA spots from tumor-adjacent, tumor-border, and tumor-center areas from each patient were cropped. In total 3 or 4 spots from tumor-adjacent and tumor-center regions were collected from each patient in the prostate

cancer cohort. In the lung cancer cohort, each patient has 1 to 2 spots from the tumor-center area. Spots from different regions are drastically different from each other even if the spots are from the same patient. Thus, we aggregated TMA spot feature representations to patient level by averaging each region separately. These patient feature representations were used to associate clinical information within each dataset to evaluate the overall methodology.

Patient groups were identified by k-Means clustering on patient-level feature representations. The optimal number of clusters was determined using the elbow method with inertia. Inertia is the sum of all element distances to the cluster center in one cluster. It decreases when the number of clusters is increased. The optimal number of clusters is the point where the inertia stops decreasing drastically. Then, we calculated the univariate Kaplan-Meier estimator and performed a log-rank test for each patient group. Moreover, in the multivariate Cox Proportional Hazards Model, the group with the best survival estimation in the Kaplan-Meier plot is used as baseline for verification. We used lifelines[49] library for the survival analyses.

Creating attention maps of TMA spots

Attention maps were generated by inputting an image into the ViT model, and extracting attention heads from different layers independently to capture the focus of the model on the image. To create TMA spot level attention maps, we used the Level 1 ViT model, as Level 2 input images are too small (*16x16*) to produce useful attention maps. Firstly, we cropped the input TMA spot image into overlapping small patches with size of *256x256,* and overlap of *128 pixels*. Then, we input those overlapped patches to the Level 1 ViT model, and extracted the attention maps. The attention maps were stitched back together based on the original positions of the respective input patches. Then, the values of the overlapping areas were averaged. Lastly, we masked the area outside of the true core with a loose mask created by dilating the DAPI channel. For visualizations in Fig. 6, we show only the top 10% of the values to highlight the most relevant regions inside the images.

Clustering metrics

We used multiple metrics on the feature representation space. All metrics are calculated with cosine distance.

Normalized Mean Pairwise Distance (nMPD): nMPD is the average distance (cosine in our case) between all pairs of points in the set divided by the largest distance. In the context of high dimensional datapoints, nMPD is close to 0 when the datapoints are similar to each other. On the other hand, nMPD is closer to 1 when the datapoints are sufficiently far away from all other datapoints.

Silhouette score: To evaluate separation of local features from different regions, we used average Silhouette score. Average Silhouette score of 0 or lower indicates poor clustering with complete overlap of clusters, and a score of 1 indicates perfect separation between clusters with regards to the label.
Classification scores: We calculated precision, recall, specificity, and F1 scores using kNN classification of tumor regions, while considering samples from the tumor-center region as positive. To evaluate the feature representations thoroughly, the classification scores are calculated with different numbers of neighbors for the kNN. Ideally, classification scores should be high with a smaller number of neighbors and gradually decrease as the number of neighbors increases. This should indicate that immediate neighboring points are similar to each other, while a larger neighborhood does not need to exhibit the same level of similarity.

Visualizations

TMA spot visualizations show values between the median and the 99th percentile for each channel. Values below a lower threshold are set to zero to remove the background, while values above a high threshold are set to the upper threshold value. However, inputs to the models are always logarithmic transformations of the raw images. Uniform Manifold Approximation and Projection (UMAP) for dimension reduction visualizations are created by first reducing the dimensionality to 50 using principal component analysis (PCA), then computing the two dimensional UMAP that are displayed in this paper.

Data availability

Source code is available on GitHub repository:
https://github.com/bioimage-profiling/SSL-Multiplexed-Imaging.
The trained Level 1 MAE models are available on Zenodo repository:
https://zenodo.org/records/15011375

Acknowledgement

- We acknowledge the Research Council of Finland (grants 340273, 346604, 359907, 363154 for L.P.; 363152 for T.P.; 322675 for T.M.), the Cancer Foundation Finland (220040 for L.P.; decision date 30/11/2022 for T.P.; 304667, 191118 for T.M.), Jane and Aatos Erkko Foundation (290520 for T.M.), Hospital District of Helsinki and Uusimaa (TYH2018214, TYH2018222, TYH2019235, TYH2019249 for T.M.), and Sigrid Juselius Foundation (230149, 240158 for T.P.; 8191 for T.M.)

- We want to acknowledge the participants and investigators of iCAN study. The iCAN project is funded by the Research Council of Finland (grant number 320185), University of Helsinki, Helsinki University Hospital and industry partner Boehringer Ingelheim.
- We acknowledge CSC–IT Center for Science, Finland for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium.
- Microscopic imaging of the samples was performed at the FIMM Digital Microscopy and Molecular Pathology Unit, supported by HiLIFE and Biocenter Finland.
- Johan Botling and Patrick Micke are acknowledged for lung cancer cohort clinical data.

References

1. Galli, F. *et al.* Relevance of immune cell and tumor microenvironment imaging in the new era of immunotherapy. *J. Exp. Clin. Cancer Res.* **39**, 89 (2020).
2. Sahai, E. *et al.* A framework for advancing our understanding of cancer-associated fibroblasts. *Nat. Rev. Cancer* **20**, 174–186 (2020).
3. Xiao, Y. & Yu, D. Tumor microenvironment as a therapeutic target in cancer. *Pharmacol. Ther.* **221**, 107753 (2021).
4. Sun, Y. Tumor microenvironment and cancer therapy resistance. *Cancer Lett.* **380**, 205–215 (2016).
5. Balkwill, F. R., Capasso, M. & Hagemann, T. The tumor microenvironment at a glance. *J. Cell Sci.* **125**, 5591–5596 (2012).
6. Olumi, A. F. *et al.* Carcinoma-associated fibroblasts direct tumor progression of initiated human prostatic epithelium. *Cancer Res.* **59**, 5002–5011 (1999).
7. Henriksson, M. L. *et al.* Colorectal cancer cells activate adjacent fibroblasts resulting in FGF1/FGFR3 signaling and increased invasion. *Am. J. Pathol.* **178**, 1387–1394 (2011).
8. Brennen, W. N., Isaacs, J. T. & Denmeade, S. R. Rationale behind targeting fibroblast activation protein-expressing carcinoma-associated fibroblasts as a novel chemotherapeutic strategy. *Mol.*

Cancer Ther. **11**, 257–266 (2012).

9. Chen, X. & Song, E. Turning foes to friends: targeting cancer-associated fibroblasts. *Nat. Rev. Drug Discov.* **18**, 99–115 (2019).
10. Cassetta, L. & Pollard, J. W. Targeting macrophages: therapeutic approaches in cancer. *Nat. Rev. Drug Discov.* **17**, 887–904 (2018).
11. DeNardo, D. G. & Ruffell, B. Macrophages as regulators of tumour immunity and immunotherapy. *Nat. Rev. Immunol.* **19**, 369–382 (2019).
12. Poh, A. R. & Ernst, M. Targeting macrophages in cancer: From bench to bedside. *Front. Oncol.* **8**, 49 (2018).
13. Sanegre, S. *et al.* Integrating the tumor microenvironment into cancer therapy. *Cancers* **12**, 1677 (2020).
14. Blom, S. *et al.* Systems pathology by multiplexed immunohistochemistry and whole-slide digital image analysis. *Sci. Rep.* **7**, 15580 (2017).
15. Lin, J.-R. *et al.* Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *Elife* **7**, (2018).
16. Windhager, J. *et al.* An end-to-end workflow for multiplexed image processing and analysis. *Nat. Protoc.* **18**, 3565–3613 (2023).
17. Bollhagen, A. & Bodenmiller, B. Highly multiplexed tissue imaging in precision oncology and translational cancer research. *Cancer Discov.* **14**, 2071–2088 (2024).
18. Pellinen, T. *et al.* Fibroblast activation protein defines an aggressive, immunosuppressive EMT-associated tumor subtype in highly inflamed localized clear cell renal cell carcinoma. *bioRxiv* (2024).
19. Pellinen, T. *et al.* Fibroblast subsets in non-small cell lung cancer: Associations with survival, mutations, and immune features. *J. Natl. Cancer Inst.* **115**, 71–82 (2023).
20. Färkkilä, A. *et al.* Immunogenomic profiling determines responses to combined PARP and PD-1

inhibition in ovarian cancer. *Nat. Commun.* **11**, 1459 (2020).

21. Liu, C. C. *et al.* Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering. *Nat. Commun.* **14**, 4618 (2023).
22. Kim, J. *et al.* Unsupervised discovery of tissue architecture in multiplexed imaging. *Nat. Methods* **19**, 1653–1661 (2022).
23. Hollandi, R. *et al.* NucleAIzer: A parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Syst.* **10**, 453–458.e6 (2020).
24. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
25. McQuin, C. *et al.* CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol.* **16**, e2005970 (2018).
26. Caicedo, J. C. *et al.* Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017).
27. Cords, L. *et al.* Cancer-associated fibroblast phenotypes are associated with patient outcome in non-small cell lung cancer. *Cancer Cell* **42**, 396–412.e5 (2024).
28. Zhang, W. *et al.* Identification of cell types in multiplexed in situ images by combining protein expression and spatial information using CELESTA. *Nat. Methods* **19**, 759–769 (2022).
29. Wang, J. *et al.* Multiplexed immunofluorescence identifies high stromal CD68+PD-L1+ macrophages as a predictor of improved survival in triple negative breast cancer. *Sci. Rep.* **11**, 21608 (2021).
30. Gui, J. *et al.* A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 9052–9071 (2024).
31. Caron, M. *et al.* Emerging properties in self-supervised vision transformers. in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 9650–9660 (2021).
32. Bardes, A., Ponce, J. & Le Cun, Y. VICRegL: Self-Supervised Learning of Local Visual Features. in

Advances in Neural Information Processing Systems vol. 35 8799–8810 (2022).

33. Bardes, A., Ponce, J. & LeCun, Y. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *arXiv:2105.04906* (2021).
34. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. Big Self-Supervised Models are Strong Semi-Supervised Learners. in *Advances in neural information processing systems* vol. 33 22243–22255 (2020).
35. He, K. *et al.* Masked Autoencoders Are Scalable Vision Learners. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16000–16009 (2022).
36. Doron, M. *et al.* Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv* (2023).
37. Pfaendler, R., Hanimann, J., Lee, S. & Snijder, B. Self-supervised vision transformers accurately decode cellular state heterogeneity. *bioRxiv* 2023.01.16.524226 (2023).
38. Kraus, O. *et al.* Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11757–11768 (2024).
39. Chen, R. J. *et al.* Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 16144–16155 (IEEE, 2022).
40. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. E. A simple framework for contrastive learning of visual representations. in *International Conference on Machine Learning* vol. abs/2002.05709 1597–1607 (PMLR, 2020).
41. Schubert, E. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *SIGKDD Explor.* **25**, 36–42 (2023).
42. Sørensen, J. B., Klee, M., Palshof, T. & Hansen, H. H. Performance status assessment in cancer patients. An inter-observer variability study. *Br. J. Cancer* **67**, 773–775 (1993).

43. Micke, P. *et al.* Aberrantly activated claudin 6 and 18.2 as potential therapy targets in non-small-cell lung cancer. *Int. J. Cancer* **135**, 2206–2214 (2014).

44. Micke, P. *et al.* The impact of the fourth edition of the WHO classification of Lung Tumours on histological classification of resected pulmonary NSCCs. *J. Thorac. Oncol.* **11**, 862–872 (2016).

45. Lehto, T.-P. K. *et al.* Transcript analysis of commercial prostate cancer risk stratification panels in hard-to-predict grade group 2-4 prostate cancers. *Prostate* **81**, 368–376 (2021).

46. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929* (2020).

47. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).

48. Susmelj, I., Heller, M., Wirth, P., Prescott, J. & Ebner, M. *Lightly*.

49. Davidson-Pilon, C. lifelines: survival analysis in Python. *J. Open Source Softw.* **4**, 1317 (2019).