frontiers | Frontiers in Oncology

TYPE Methods
PUBLISHED 19 December 2022
DOI 10.3389/fonc.2022.964716

Check for updates

OPEN ACCESS

EDITED BY
Marco Cassano,
Lunaphore Technologies SA,
Switzerland

REVIEWED BY
Trevor David McKee,
Deciphex Inc., Ireland
Gary Zirpoli,
Boston University, United States

*CORRESPONDENCE
Henrik Failmezger
henrik.failmezger@astrazeneca.com

SPECIALTY SECTION
This article was submitted to
Cancer Immunity
and Immunotherapy,
a section of the journal
Frontiers in Oncology

RECEIVED 08 June 2022
ACCEPTED 23 November 2022
PUBLISHED 19 December 2022

CITATION
Failmezger H, Hessel H, Kapil A, Schmidt G and Harder N (2022) Spatial heterogeneity of cancer associated protein expression in immunohistochemically stained images as an improved prognostic biomarker. *Front. Oncol.* 12:964716. doi: 10.3389/fonc.2022.964716

Spatial heterogeneity of cancer associated protein expression in immunohistochemically stained images as an improved prognostic biomarker

Henrik Failmezger*, Harald Hessel, Ansh Kapil, Günter Schmidt and Nathalie Harder

Research & Development Early Oncology Translational Medicine, Computational Pathology, AstraZeneca Computational Pathology GmbH, Munich, Germany

The identification of new tumor biomarkers for patient stratification before therapy, for monitoring of disease progression, and for characterization of tumor biology plays a crucial role in cancer research. The status of these biomarkers is mostly scored manually by a pathologist and such scores typically, do not consider the spatial heterogeneity of the protein's expression in the tissue. Using advanced image analysis methods, marker expression can be determined quantitatively with high accuracy and reproducibility on a per-cell level. To aggregate such per-cell marker expressions on a patient level, the expression values for single cells are usually averaged for the whole tissue. However, averaging neglects the spatial heterogeneity of the marker expression in the tissue. We present two novel approaches for quantitative scoring of spatial marker expression heterogeneity. The first approach is based on a co-occurrence analysis of the marker expression in neighboring cells. The second approach accounts for the local variability of the protein's expression by tiling the tissue with a regular grid and assigning local spatial heterogeneity phenotypes per tile. We apply our novel scores to quantify the spatial expression of four different membrane markers, i.e., HER2, CMET, CD44, and EGFR in immunohistochemically (IHC) stained tissue sections of colorectal cancer patients. We evaluate the prognostic relevance of our spatial scores in this cohort and show that the spatial heterogeneity scores clearly outperform the marker expression average as a prognostic factor (CMET: p-value=0.01 vs. p-value=0.3).

KEYWORDS

spatial heterogeneity, immunohistochemistry, digital pathology, spatial statistics, prognostic biomarker

Introduction

In cancer research, tumor biomarkers play a crucial role for patient stratification in targeted therapies, for monitoring of disease progression, and for characterization of tumor biology. For example HER2 expression is an important biomarker for immunotherapies in breast cancer (1, 2). Applying advanced image analysis and machine learning methods, the quantitative expression of IHC markers can be determined for whole tissue sections with high accuracy. Such methods have been shown to be highly concordant to the manual scoring of a marker by pathologists (3, 4) and provide quantitative measurements of the marker's expression on a cellular level. To aggregate the cellular expression values into one score per sample, most often simple averaging is performed which does not consider the spatial heterogeneity of the marker expression (5). However, it has been shown that relevant prognostic information can be derived by analyzing the spatial composition of the tumor microenvironment (6–8), as well as more generally the spatial heterogeneity of protein expression. For example, intra-tumoral heterogeneity has been found to be associated with more aggressive tumors and unfavorable outcomes in many cancer indications (9).

To assess heterogeneity in marker expression, Cherenova et al. have transferred cell signal intensity levels of immunofluorescence images to marks in marked point patterns and defined scores based on the conditional mean and variance of this point pattern (5). They showed that their scores provide prognostic information in breast cancer.

In this technical proof of concept study, we present spatial scores to determine the expression heterogeneity of proteins in the tissue which turned out to have enormous potential to serve as future biomarkers in clinical studies.

In comparison to Cherenova et al. our approach does not view the tissue as a marked point pattern but is inspired by the well-known Haralick features (10) that are used for texture analysis in pixel images. Like the Haralick features, our scores are based on a co-occurrence analysis, however instead of analyzing grey value intensities of neighboring pixels, we examine the marker expressions of neighboring cells. Our method has the advantage that it can be applied to continuous measurements like marker expression intensities and that it enables the calculation of a variety of scores that capture different aspects of marker expression heterogeneity.

As the tumor microenvironment is highly complex and marker expression may vary locally in the tissue, we implemented a second methodology that is based on a tessellation of the tumor tissue. By overlaying the tissue with a grid, we quantify the local distribution of marker expression. The cells inside the grid tiles are classified into *marker-positive* and *marker-negative* by a range of intensity thresholds. Based on the *marker-positive* and *marker-negative* cell proportions, we calculate the Shannon entropy as a measure of local heterogeneity. We then consider different approaches to aggregate the local heterogeneity values into global per-sample scores.

We apply our novel scoring approaches to resections from a cohort of 34 colorectal cancer patients that have been stained by immunohistochemistry for HER2, CMET, CD44, and EGFR. We show that the spatial scores provide a considerable improvement in patient stratification as compared to the standard averaging of the marker values.

Methods

Data

The cohort included 34 patients of colorectal cancer (CRC). FFPE blocks (Indivumed) of tumor resections were stained for the membrane markers EGFR (3C6), HER2 (4B5), CMET (ECD), and CD44 (Figure 1C). Samples were scanned by Aperio with a magnification of 20x.

There were 34 samples available for each CMET, HER2, and EGFR, and 31 samples for CD44. All patients had late stage tumors (Stage IV = 33 patients, Stage II = 1 patient), 74% of the patients were treated by chemotherapy, for 15% therapy was unknown (Supplementary Figure S1). A subset of patients (41%) was also treated by targeted therapy (Cetuximab (Erbitux®)/Bevacizumab (Avastin®)). Overall survival data was available for all patients.

Image segmentation

In all whole slide tissue images epithelium was detected automatically and all epithelial cells were segmented into three compartments (nucleus, cytoplasm, membrane) using a dedicated image analysis approach described in more detail elsewhere (11). In short, the approach is based on two independent supervised convolutional neural network (CNN) models for semantic segmentation. The first model robustly detects epithelium regions while the second model predicts the cellular compartments for the whole image. In a postprocessing step the predictions of both CNNs are combined to compute the cell segmentation within the detected epithelium regions (Figure 1A). Both models have been trained on large amounts of IHC stained tissue images along with expert annotations including different indications and markers. Finally, based on the segmentation result the amount of brown DAB staining on the cellular membranes of all tumor cells was quantified by computing the average brown optical density (mean OD) across all pixels per cellular membrane. Note that tumor epithelium (vs. normal epithelium) was classified based on coarse tumor core region annotations provided by an expert pathologist. The performance of the segmentation approach on the images of this study has been assessed by correlating the image-based readouts with scoring performed by expert pathologists (Figure 1B).

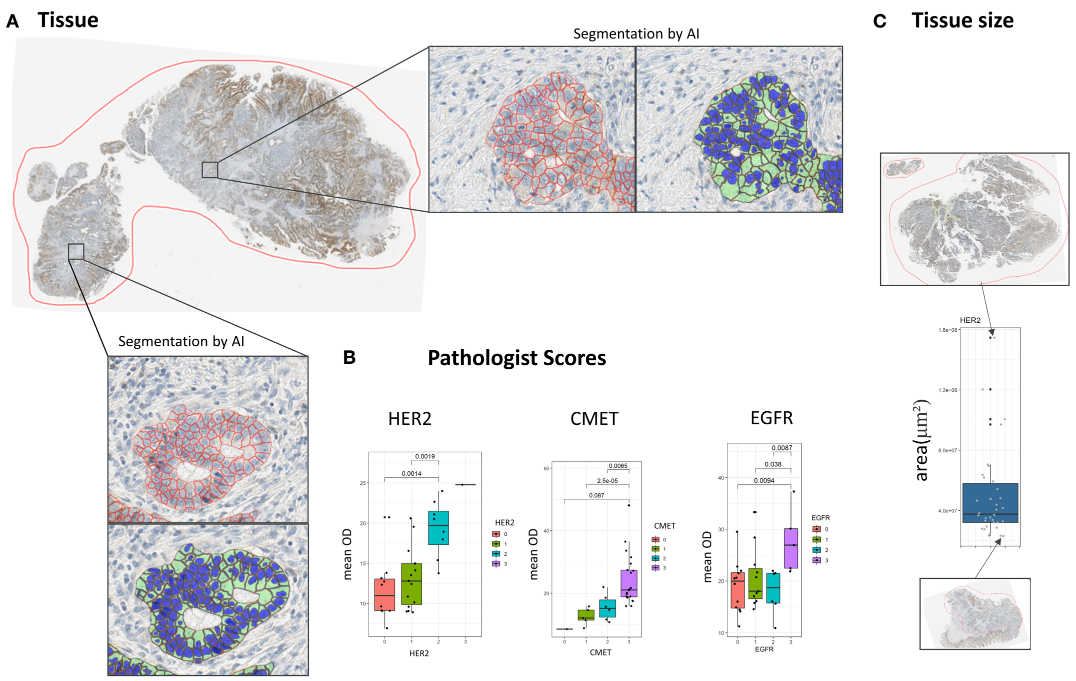


FIGURE 1
Cell segmentation pipeline. **(A)** The membrane of cells is segmented by an automated segmentation pipeline. **(B)** Concordance of scores resulting from the automatic image analysis pipeline with pathologist scores. **(C)** Distribution of tissue sizes in the dataset in μm^2.

Co-occurrence scores

The co-occurrence scores are inspired by the Haralick texture analysis (10). In comparison to the Haralick features, we do not create a pixel-based but an object-based co-occurrence matrix (12), where objects correspond to cells in the tissue. To this end, we first create a cell neighborhood graph, in which cells are considered as nodes and edges represent the neighborhood relation of cells based on a distance threshold (Figure 2A). Based on this cell graph a co-occurrence matrix is calculated as a 2D histogram of membrane optical density values for all pairs of cells sharing an edge in the neighborhood graph. The size of the resulting co-occurrence matrix is given by 255x255, which corresponds to the mapped OD values. Based on the co-occurrence matrix, we calculate a subset of the original aralick features, including *Homogeneity*, *Contrast*, *Correlation*, *Angular second moment* (10).

Let $p(i,j)$ be the normalized co-occurrence matrix, with i,j being the OD values of the cells. Let μ_x, μ_y and σ_x, σ_y be the marginal means and standard deviations of the rows or columns of the co-occurrence matrix. $Homogeneity = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$ is large if neighboring cells share similar OD values, whereas $Contrast = \sum_{i} \sum_{j} (i-j)^2 p(i,j)$ increases if neighboring cells have distinct OD values. $Correlation = \frac{\sum_i \sum_j (ij)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$ is closely related to the Pearson correlation statistic and measures the linear dependency of the expression values of neighboring cells. $Angular\ second\ moment = \sum_i \sum_j p(i,j)^2$ is a measure of general uniformity of OD values.

As the co-occurrence matrix is dependent on the choice of the distance threshold, we created separate co-occurrence matrices for a range of distances (10, 25, 50, 75 µm), resulting in 16 co-occurrence scores per tissue sample.

Tessellation scores

The second set of scores is based on tissue tessellation. We first calculate the relevant tissue region by the convex hull of all cell coordinates. This region is split into subregions based on a grid with a region size of 250x250 µm^2. Only tiles that include at least n=5 cells are considered for further analysis. The class *high expression* or *low expression* is assigned based on the marker expression of the cells (i.e., membrane mean OD) inside the grid tile using a fixed threshold. From the resulting spatial expression map (Figure 2B) a single score is calculated as the ratio of the expression classes. To make this score more independent from

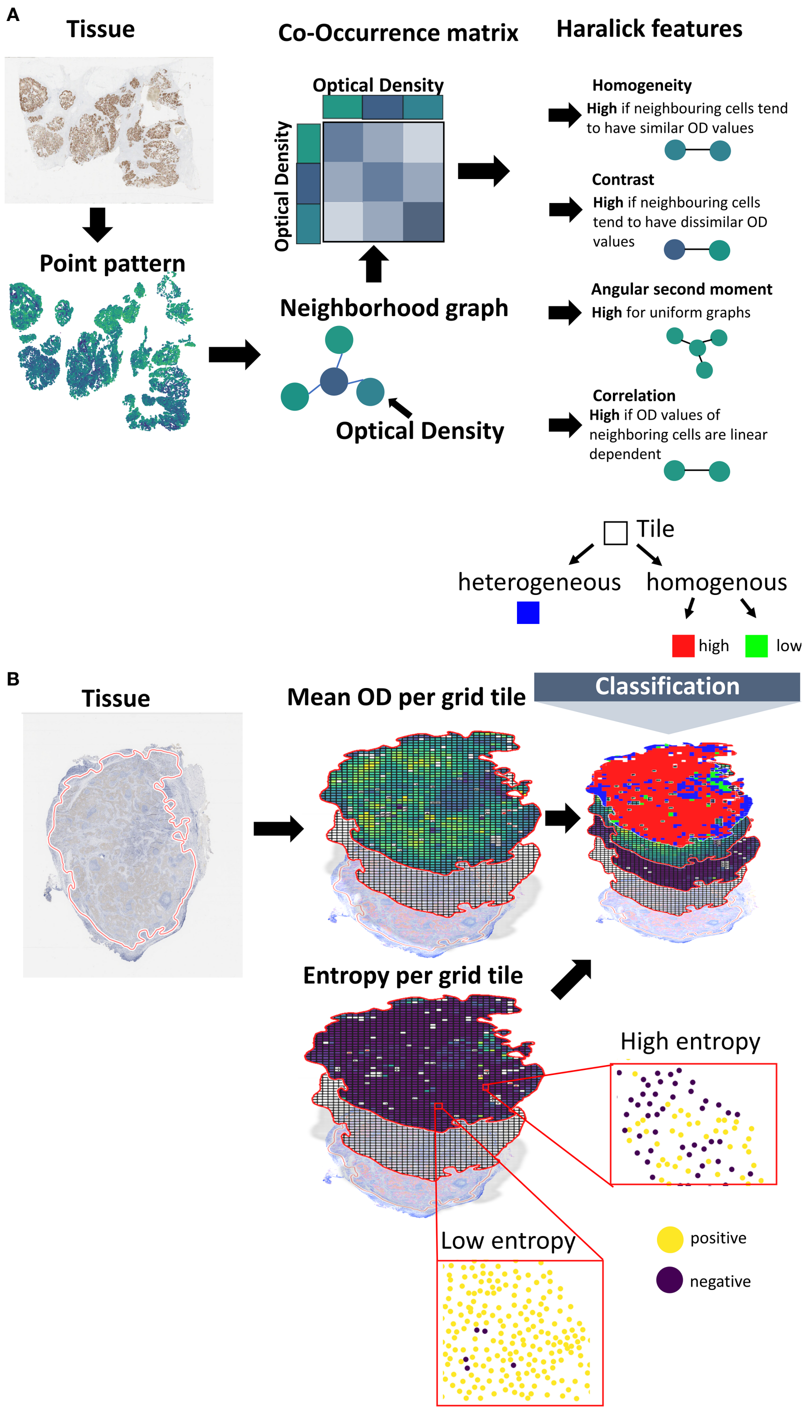


FIGURE 2
Calculation of spatial scores. **(A)** Spatial scores based on co-occurrence analysis. Using a machine learning pipeline, the cells are transferred into a point pattern. Center coordinates and marker membrane OD values are reported per cell. The point pattern is transformed into a neighborhood graph. Cells correspond to nodes and edges connect cells in spatial proximity (using a distance criterion). A co-occurrence matrix is created, where the entries of the matrix represent the frequencies in OD value combinations for pairs of neighboring cells. Scores are calculated from the co-occurrence matrix based on the Haralick texture features. **(B)** Tessellation scores. The tissue is overlaid with a grid. The averaged membrane OD values of the cells that fall inside a certain tile are calculated. Cells are divided into *marker-positive* and *marker-negative* based on a threshold. The Shannon entropy is calculated for the class proportions. Tiles are classified into *marker-homogenous* or *-heterogenous* based on the Shannon entropy. Homogenous tiles are classified into *marker-high* or *marker-low* based on a threshold.

the localization of the tiles, and thus, robust against tiling artefacts, we shift the grid by 25 µm in all 4 directions. The resulting score is the average of the scores of the individual shifts.

As an extension to the binary expression classes, we also assess the expression heterogeneity per tile. Therefore, we first classify the cells into *marker-positive* or *marker-negative* cells by a pre-defined threshold on the membrane mean OD values. We then calculate the Shannon entropy of the class proportions per tile. The per-tile entropy values are aggregated into a single score by taking their average and by calculating the so called, Ecosystem Diversity Index (7) (EDI). The EDI represents the number of components within the distribution of the per-tile entropy values. These components can be seen as the number of sub-populations in cellular expression heterogeneity. To identify the components, a range of Gaussian mixture models with different component numbers (1–5) are fitted to the distribution. The number of components of the best fitting Gaussian mixture model (w.r.t. Bayesian information criterion) defines the EDI of a sample (Supplementary Figure S2). Additionally, we calculate the EDI of all mean expression values and of the standard deviation of the expression values.

Next, we also incorporate the class *heterogenous* into the spatial expression map (Figure 2). Therefore, we apply a decision tree-like scheme, by first classifying cells inside the tile into *homogenous* or *heterogenous* expression based on the Shannon entropy values (Figures 2; S2). For this classification we use an entropy threshold of 0.61 that corresponds to maximum proportions of 70% positive and 30% negative cells (or vice versa). Tiles below this value are then classified as *homogenous-high* or *homogenous-low*. For the classification into *homogenous-high* or *homogenous-low*, we use the OD thresholds [8, 10, 15, 20, 25]. The corresponding range of threshold is used to calculate the Shannon entropy by classifying the individual cells into *marker-positive* vs. *marker-negative* as described above.

Feature selection and analysis

To reduce the overall number of scores, and thus avoid overfitting, spatial scores with a Pearson correlation larger than 0.9 were pruned. In the correlation filtering process, we applied univariate Cox regression to select the scores to keep, i.e. for a set of correlated scores we kept the score that had the best association with survival (i.e., lowest p-value). We then applied feature selection by *lasso* (least absolute shrinkage and selection operator) regression to the remaining set of uncorrelated scores using a stability selection approach (13). Only scores that had a selection probability of 0.20 in the stability selection were used for further analysis.

We used univariate and multiple Cox regression to determine the relationship of the scores with overall survival. In order to avoid multicolinearity in the multiple regression approach, we used only scores that had a Spearman correlation smaller than 0.9 with the average marker expression. We reduced the covariates by stepwise regression analysis using forward-backward selection (14).

Cross validation

To evaluate the robustness of the selected spatial scores we implemented an n-times repeated k-fold cross validation scheme. Therefore, we first implemented a k-fold cross validation scheme, in which the full data was partitioned into k equally-sized random splits. For a given split, all but one of the k-subsets were used for training an optimal cutoff. The cutoffs were then applied to the remaining unseen subset to test stratification performance. To evaluate log-rank performance, high and low scoring patients were grouped across all k test folds within one split to derive a cross-validated Kaplan-Meier curves. To test the statistical significance of the Kaplan-Meier curves, a permutation test (15) was applied to derive empirical p-values using m random permutations of the overall survival information. The above-described k-fold cross validation was then applied to each of the permutations to derive the distribution of the log-rank statistic under the Null hypothesis of independence of considered features and outcome. We used the parameters n=20, k=10, m=1000.

Software and implementation

The image analysis pipeline was implemented in Python and Tensorflow (16). The calculation of the spatial scores was implemented in Python. *Scikit-learn* was used for fitting the Gaussian mixture model (17), *scikit-image* was used to calculate the Haralick features.

Data analysis was conducted in R. The package *survminer* was used for the survival analysis. The package *lmtest* was used for the likelihood ratio test (18). The *stabs* package (19) was used for the stability selection approach. The package *MASS* was used for stepwise regression analysis (14).

Results

Mapping IHC marker expression to spatial scores

To calculate spatial heterogeneity scores for the expression of the four considered markers HER2, CMET, CD44, and EGFR the images were first processed by a machine learning pipeline that included region segmentation, cell segmentation and cell classification (11). We compared the results of the image analysis pipeline with pathologist scores for HER2, CMET, and EGFR (note that no pathologist scores were available for CD44) and found an overall good concordance between the average marker expression and *high/low* pathologist scores (Figure 1B).

We calculated four different co-occurrence scores based on the Haralick features for texture analysis: *Homogeneity*, *Contrast*, *Angular second moment (ASM)* and *Correlation* (Figures 2; 3C, Methods). To calculate co-occurrence scores for different ranges, we used the radii of 10, 25, 50, and 75 µm to define cellular neighborhood in the neighborhood graph construction (Figure 3C). For a radius of 10 µm we found cells to have a median of 2.3 neighbors, whereas for a radius of 75 µm cells had a median of 109.3 neighbors (Figure S3). We further found the *Homogeneity* as well as the *Correlation* feature to be larger for small radii and decreasing for larger radii, whereas the feature values for the other features stayed stable for different radii (Supplementary Figure S4).

For the tessellation scores, we overlaid the tissue with a quadratic grid of tile size 250x250 µm². We observed a large variety in per-tile marker expression over the whole tissue

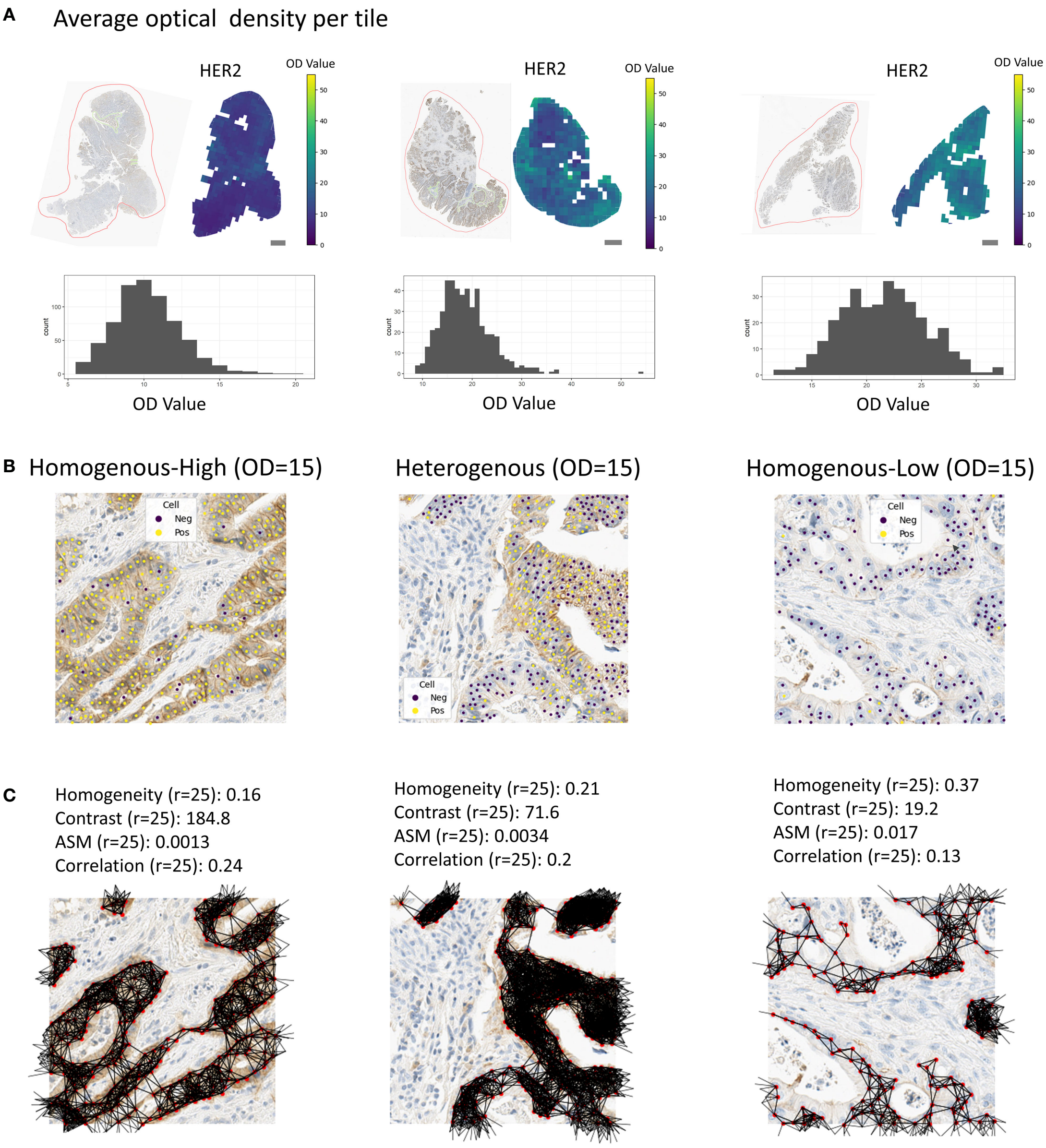


FIGURE 3
Tissue regions with associated spatial scores. **(A)** Examples of tissue samples with associated tile-wise optical densities (HER2). Scale bars correspond to *1mm*. **(B)** Examples of homogenous-high, heterogenous or homogenous-low tiles. **(C)** Associated co-occurrence features. Co-occurrence features were calculated on the cells in the tiles only.

(Figure 3A). In order to classify these tiles into the categories *homogenous-high* and *homogenous-low* (see Methods, Figure 3B), we used the OD thresholds of 8, 10, 15, 20, and 25. We used the same range of thresholds to calculate the entropy of *marker-positive* or *marker-negative* cells (Methods). We calculated the Ecosystem Diversity Index (EDI) for the average marker expression per tile and the standard deviation of this marker expression (Supplementary Figure S2). Furthermore, we calculated the EDI for the entropy of *marker-positive* or *marker-negative* cell proportions, resulting in 7 EDI scores for one sample as well as ratios for the numbers of *homogenous-high*, *homogenous-low*, and *heterogenous* tiles (Figure 3B). To assess the robustness of our tessellation scores with respect to the size and positioning of the tiles on the tissue we performed additional experiments. To account for the dependency of the scores on the tile location, we shifted the grid in all directions by 1/10 of the tile size (i.e., 25 μm) and computed the average score over all shifts as the final readout. To investigate the dependency of the tessellation scores on the tile size we increased and decreased the tile size by 50%, respectively, and compared the resulting scores for one feature. We found a very good concordance between the scores calculated on different tile sizes (Pearson correlation > 99%, Supplementary Figure S5) suggesting a solid robustness with respect to the choice of this parameter.

In total we computed 16 co-occurrence scores and 37 tessellation scores. We found high correlations between the tesselation scores and the co-occurrence scores. In general, scores based on *Correlation* and EDI were less correlated with the rest of the scores (Supplementary Figure S6).

To select the most relevant subset of spatial scores based on their prognostic value we first pruned correlated scores, and afterwards applied Cox lasso stability selection (see Methods). We kept all scores that were selected in at least 20% of the stability selection runs. From the initial set of 53 spatial scores, 5 scores were selected for HER2, 4 scores for CMET, 4 scores for CD44, and 4 scores for EGFR (Table S1). For the selected spatial scores, we also checked the correlation with pathologist scores. We found only 5 spatial scores to be clearly correlated with pathologist scores (HER2: *Ratio of homogenous-low tiles* (OD=8); *Ratio of homogenous-high tiles* (OD=25); CMET: *Homogeneity* (r=75), Contrast (OD=10), *EDI for the entropy* (OD=20), Supplementary Figure S7), indicating that the spatial scores provide additional information, which is not represented by pathologist scores.

Univariate analysis

We checked if the selected spatial scores had an association with overall survival by univariate Cox regression analysis. *Homogeneity* (r=75, p-value=0.01, Figure 4) was associated with survival for CMET, whereas large *Homogeneity* reduced the patient's Hazard (HR=0.52). For CD44 *Correlation* (r=75) was significantly associated with survival and increased the patient's Hazard (p-value=0.03, HR=1.63, Figure 4). We did not find any other significant univariate associations of the spatial scores with overall survival. We additionally applied univariate cox regression for the average marker expression values, but also found no significant association to overall survival (Supplementary Figure S8).

Multiple regression analysis

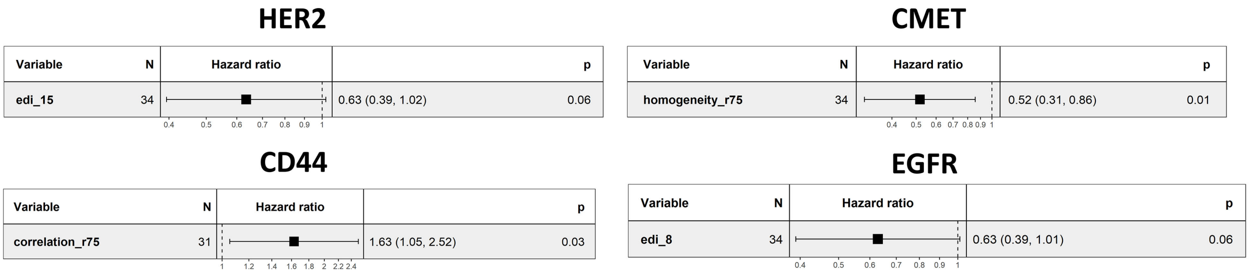
We checked if the spatial scores provide additional prognostic value compared to simple aggregation of the expression values by global averaging.

Therefore, we created baseline multiple regression models including the averaged marker expression values as well as the clinical variables age, gender, tumor grading, and the pathologist score for the marker expression (Supplementary Figure S9). We added the spatial scores selected by lasso regression and reduced the model by stepwise regression. We then compared the advanced model to the base model by a Likelihood-Ratio test to check whether the spatial scores provided additional prognostic value.

For HER2, the ratio of *homogenous-high* tiles (OD=25) and *homogenous-low* tiles (OD=8), the *EDI* for the entropy (OD=15), the *EDI* for the average expression, and the *EDI* for the standard deviation in expression (Methods) remained in the model after stepwise selection and added significant prognostic value to the baseline model (LR-test: p-value= 2.909e-06, Figure 5). For CMET, *Contrast* (r=10, Figure 5) was kept in the model and significantly improved its prognostic value (LR-test: p-value= 0.018, Figure 5). For CD44, *Correlation* (r=75), the *EDI* for the entropy (OD=15) and the ratio of *homogenous-low* tiles (OD=8), remained in the model (LR-test: p-value=0.003, Figure 5). For EGFR only the *EDI* for the entropy (OD=10) was kept in the model (LR-test: p-value=0.066, Figure 5).

Patient stratification

To test if the spatial scores can be used to stratify patients by overall survival, we tested the 25%, 50%, and 75% quantiles, respectively, as potential cutoffs for patient selection. We created survival curves for the patient groups and compared the difference of survival times by the logrank test. We found 8 features to be significant in the logrank test for the different markers. The *EDI* for the averaged OD per tile (75% quantile, p-value=0.04, Figure 6) and the *EDI* for the entropy (OD=15, 50% quantile, p-value=0.017, Figure 6) were significant for HER2. The *Contrast* (r=10, 75% quantile, p-value=0.002), the *Homogeneity* (r=75, 75% quantile, p-value=0.002, Figure 6), the *EDI* for the entropy (OD=20, 50% quantile, p-value=0.049), and *EDI* for the averaged OD per tile (average



HER2

Variable	N	Hazard ratio		p
edi_15	34		0.63 (0.39, 1.02)	0.06

CMET

Variable	N	Hazard ratio		p
homogeneity_r75	34		0.52 (0.31, 0.86)	0.01

CD44

Variable	N	Hazard ratio		p
correlation_r75	31		1.63 (1.05, 2.52)	0.03

EGFR

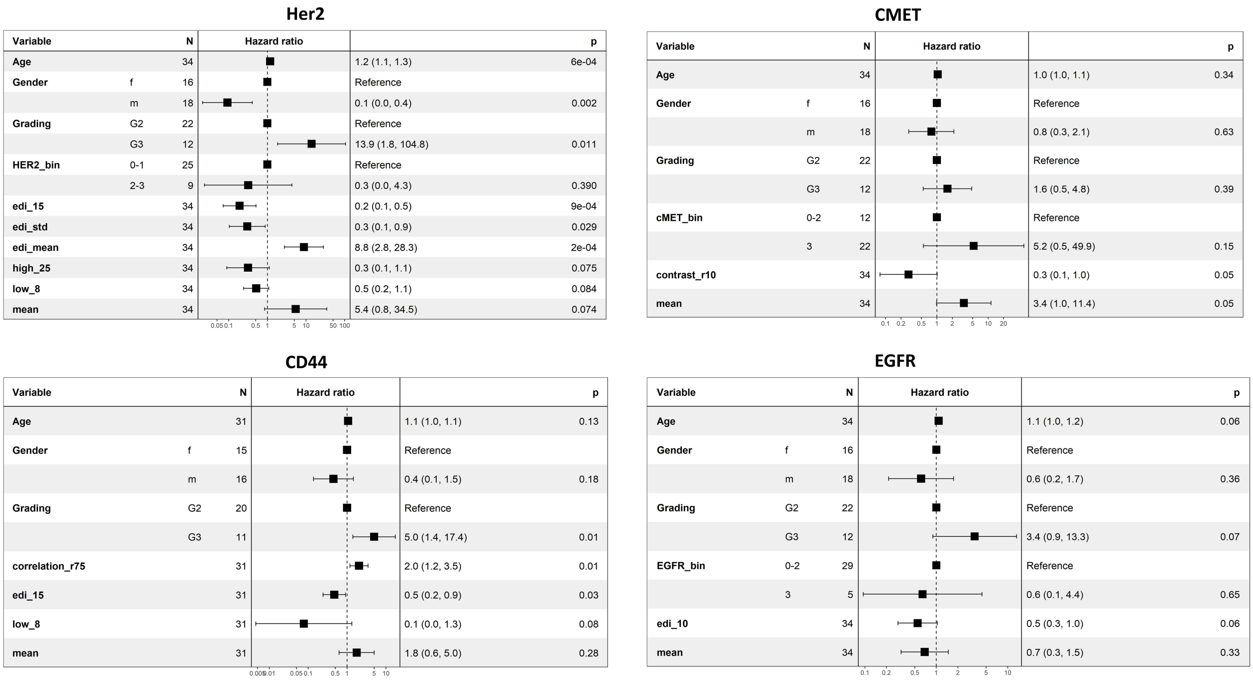
Variable	N	Hazard ratio		p
edi_8	34		0.63 (0.39, 1.01)	0.06

FIGURE 4
Univariate regression analysis for different markers. The scores with the lowest p-values are shown.

OD per tile, 75% quantile, p-value=0.023) were significant for CMET. Whereas only one score each were significant for CD44 (*ratio of homogenous-low tiles* (OD=8, 50% quantile, p-value=0.025, Figure 6), and for EGFR (*EDI* for the entropy, OD=10, 25% quantile, p-value=0.046, Figure 6). We found the average marker expression to be significant only in CMET (p-value=0.015, Figure 6).

We checked the robustness of our scores by a 20 x 10 repeated cross validation scheme (Methods). Three spatial scores had a significant median p-value in the repeated cross validation scheme (Supplementary Figure S10, CMET: *Homogeneity* (r=75), median p-value=0.033; *Contrast* (r=10), median p-value=0.009; *EDI* for the averaged OD, median p-value=0.043).

As the spatial scores do not indicate if the marker expression in the tissue in general is high or low, we combined the spatial scores with the average marker expression. We then grouped patients based on the categories high or low obtained for both feature values using the cutoff from the univariate analysis. For HER2, patients with low average expression and high *EDI* (entropy, OD=15) had the best prognosis (median survival



Her2

Variable		N	Hazard ratio		p
Age		34		1.2 (1.1, 1.3)	6e-04
Gender	f	16		Reference	
	m	18		0.1 (0.0, 0.4)	0.002
Grading	G2	22		Reference	
	G3	12		13.9 (1.8, 104.8)	0.011
HER2_bin	0-1	25		Reference	
	2-3	9		0.3 (0.0, 4.3)	0.390
edi_15		34		0.2 (0.1, 0.5)	9e-04
edi_std		34		0.3 (0.1, 0.9)	0.029
edi_mean		34		8.8 (2.8, 28.3)	2e-04
high_25		34		0.3 (0.1, 1.1)	0.075
low_8		34		0.5 (0.2, 1.1)	0.084
mean		34		5.4 (0.8, 34.5)	0.074

CMET

Variable		N	Hazard ratio		p
Age		34		1.0 (1.0, 1.1)	0.34
Gender	f	16		Reference	
	m	18		0.8 (0.3, 2.1)	0.63
Grading	G2	22		Reference	
	G3	12		1.6 (0.5, 4.8)	0.39
cMET_bin	0-2	12		Reference	
	3	22		5.2 (0.5, 49.9)	0.15
contrast_r10		34		0.3 (0.1, 1.0)	0.05
mean		34		3.4 (1.0, 11.4)	0.05

CD44

Variable		N	Hazard ratio		p
Age		31		1.1 (1.0, 1.1)	0.13
Gender	f	15		Reference	
	m	16		0.4 (0.1, 1.5)	0.18
Grading	G2	20		Reference	
	G3	11		5.0 (1.4, 17.4)	0.01
correlation_r75		31		2.0 (1.2, 3.5)	0.01
edi_15		31		0.5 (0.2, 0.9)	0.03
low_8		31		0.1 (0.0, 1.3)	0.08
mean		31		1.8 (0.6, 5.0)	0.28

EGFR

Variable		N	Hazard ratio		p
Age		34		1.1 (1.0, 1.2)	0.06
Gender	f	16		Reference	
	m	18		0.6 (0.2, 1.7)	0.36
Grading	G2	22		Reference	
	G3	12		3.4 (0.9, 13.3)	0.07
EGFR_bin	0-2	29		Reference	
	3	5		0.6 (0.1, 4.4)	0.65
edi_10		34		0.5 (0.3, 1.0)	0.06
mean		34		0.7 (0.3, 1.5)	0.33

FIGURE 5
Multiple regression analysis for the markers HER2, CMET, CD44, and EGFR. Spatial scores were added to a base model including age, gender, grading, pathologist scoring, and average marker expression, and afterwards reduced by stepwise regression analysis. The improvement of the advanced model to the base model was calculated by the Likelihood-Ratio test (HER2: p-value = 2.909e-06, CMET: p-value= 0.018, CD44: p-value=0.003, EGFR: p-value=0.066). Note that pathologist scores were not available for CD44.

time), whereas patients with low *EDI* (entropy, OD=15) and high average expression had the worst prognosis (Figure 6). We found similar group stratifications for EGFR, CMET and CD44 (Supplementary Figure S11).

Discussion

Aggregating cellular information from tissue slides is a challenging problem and applying straightforward methods such as simple averaging leads to a massive loss of information. The tumor microenvironment is a highly complex system, which can either trigger or downregulate protein expression. Finding appropriate scores that adequately capture the marker's expression heterogeneity in the tissue is crucial to better characterize and identify different patient populations.

We presented two novel scoring approaches that aggregate cellular protein expression values to per-sample scores while quantifying the spatial heterogeneity in protein

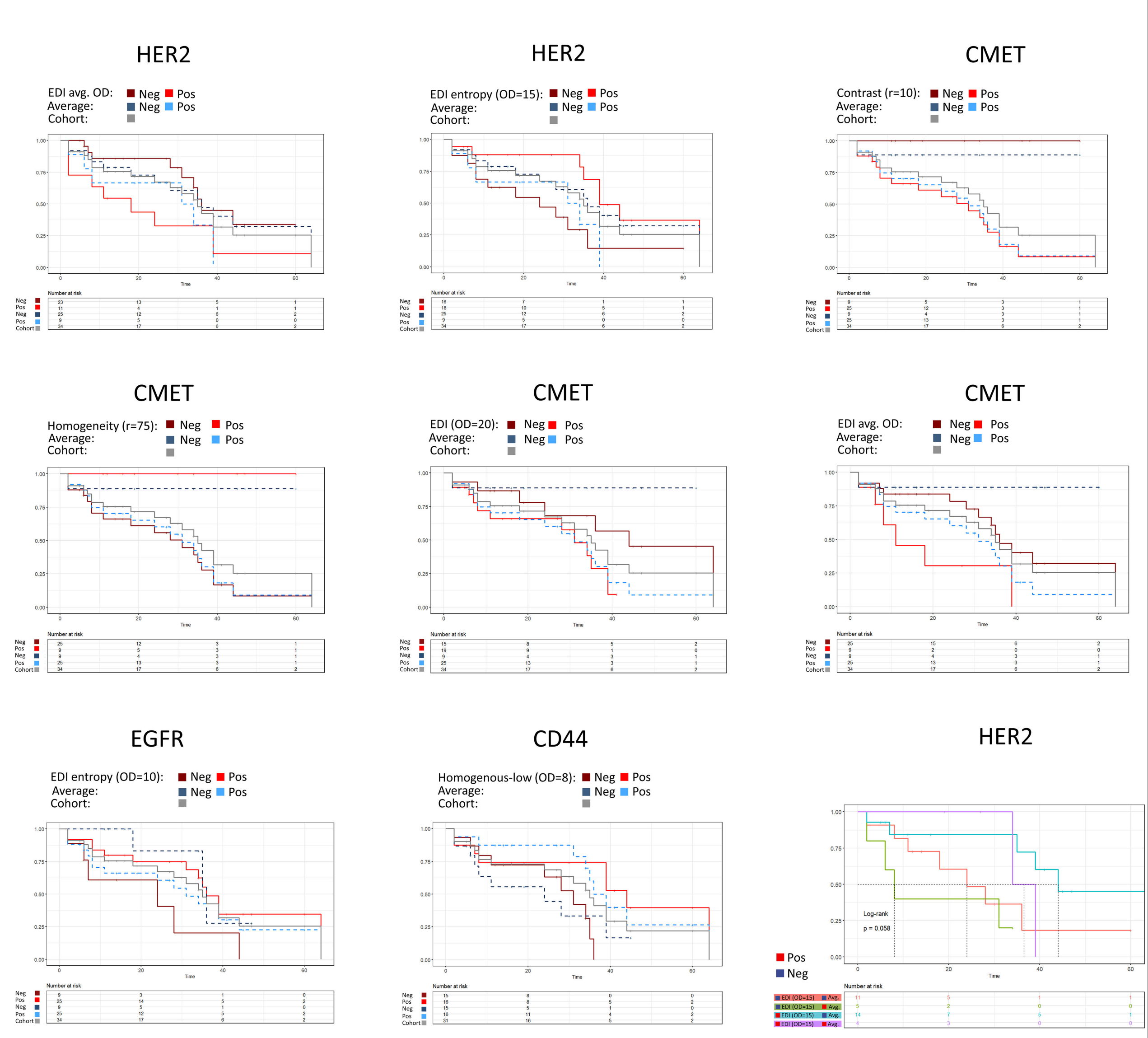


FIGURE 6
Patient stratification by spatial scores. The average expression is shown as dashed line. 25%, 50% or 75% quantiles of the score were used as stratification cutoffs, respectively, and the best cutoff was selected based on the p-value of the logrank test. HER2: *EDI* (average OD per tile, 75% quantile, p-value=0.04), *EDI* (entropy, OD=15, 50% quantile, p-value=0.017), average expression (75% quantile, p-value=0.23). CMET: *Contrast* (r=10, 75% quantile, p-value=0.002), Homogeneity (r=75, 75% quantile, p-value=0.002), *EDI* (OD=20, 50% quantile, p-value=0.049), *EDI* (average OD per tile, 75% quantile, p-value=0.023), average expression (25% quantile, p-value=0.015). EGFR: *EDI* (entropy, OD=10, 25% quantile, p-value=0.046), average expression (25% quantile, p-value=0.28). CD44: *Ratio of homogenous-low tiles* (OD=8, 50% quantile, p-value=0.025), average expression (50% quantile, p-value=0.07). In the bivariate analysis scores were grouped by the cutoff from the univariate analysis.

expression. Both approaches turned out to have immense potential for use as prognostic and potentially predictive biomarkers for patient stratification.

Our scoring approaches are widely applicable as they are generally independent of imaging modality or indication, they can be applied on any type of continuous measurements, and they can easily be adjusted to a different domain with an extremely small set of free parameters to be tuned. For the co-occurrence scores the only parameter to be specified is the distance criterion for setting up the cell neighborhood graph. The tessellation scores depend on three parameters that are (1) the tile size (2), the threshold on the continuous per-cell measurement to classify a tile as *homogenous-high* or *low* (in our case the membrane OD threshold), and (3) the threshold to classify a tile as *homogenous* or *heterogenous*. In our study we only varied parameter (2) the membrane OD threshold to compute a set of spatial scores corresponding to different resolutions of spatial dependency, and we used fixed values for the two other parameters. Finding an appropriate grid size for spatial analysis tasks in digital pathology is non-trivial and there exists no general solution (6, 7, 20). However, in our study we systematically investigated the influence of the tile size on the overall result, and we showed that our changes of 50% of the tile size had no relevant effect on the results (Supplementary Figure S5). However, larger variations of the tile sizes could certainly be employed to capture heterogeneity at different scales. The threshold to classify homogenous versus heterogenous tiles was fixed to an entropy value of 0.61, meaning that tiles were classified as heterogenous if less than 70% of the cells were either positive or negative. We did not optimize this value to avoid overfitting, however, adaptation of this value might be reasonable, for example, for other cancer indications.

In order to evaluate the generality of the scores, we tested them on a cohort of CRC patients for four different markers with known distinct expression patterns (21–23): HER2, CD44, CMET, and EGFR.

All four markers are well known to be associated with unfavorable prognosis in CRC (24–27). We applied a Cox lasso stability feature selection scheme and observed that tessellation scores as well as co-occurrence scores were selected as most relevant with respect to patient stratification for different markers (Supplementary Table S1). The spatial scores showed a higher prognostic value than the average cellular expression, underlining the importance of the spatial heterogeneity of the cellular expression values. This finding was confirmed by univariate and multiple cox regression. Overall, the spatial scores generally stratified patients better than averaging.

Interestingly, some scores were only found significant when analyzed together with the average expression in the multiple regression model (Figure 5). This potentially is due to suppression effects. Suppression effects in multiple regression happen, if a predictor (suppressor) that is uncorrelated with the output criterion, but is correlated with another predictor, improves the overall predictive power of a model (28). The reason for that is that the suppressor controls for variance in the other predictor. For the spatial scores this indicates a general challenge. The spatial scores identify spatially *homogenous* or *heterogenous* expression of the marker, however, they provide no information about the absolute value of the marker expression - meaning they do not indicate whether the marker is strongly or weakly expressed in the tissue. The combination of our spatial scores with the average expression considerably improved the prognostic value of the multiple regression model suggesting great potential for patient stratification. We showed this, for example, for HER2 where the combination of the spatial scores with the average expression resulted in a refined patient stratification into four groups with quite distinct prognosis (Figure 6).

The main limitation of this technical proof of concept study is the small dataset. In future work, the significance of our scores needs to be reproduced in a larger cohort, ideally including genomic and proteomic data.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Ethics statement

Samples used in this study were provided by Individumed GmbH, 20251 Hamburg, Germany. Written agreements for all relevant samples were signed, guaranteeing that appropriate ethics approvals were obtained. The patients/participants provided their written informed consent to participate in this study.

Author contributions

HF developed the scores, implemented the methods, analyzed the data, and wrote the manuscript. NH contributed to the development of the scores, to the image analysis and manuscript revision. AK contributed to the image analysis pipeline and provided technical advice. HH stained the tumor sections. GS provided the data and contributed to the development of the scores. All authors contributed to the article and approved the submitted version.

Acknowledgments

We thank Alexander Buschle for handling the image data.

Conflict of interest

All authors are full time employees of AstraZeneca Computational Pathology GmbH.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2022.964716/full#supplementary-material

References

1. Greally M, Kelly CM, Cercek A. HER2: An emerging target in colorectal cancer. *Curr Probl Cancer* (2018) 42(6):560–71. doi: 10.1016/j.currproblcancer.2018.07.001

2. Iqbal N, Iqbal N. Human epidermal growth factor receptor 2 (HER2) in cancers: Overexpression and therapeutic implications. *Mol Biol Int* (2014) 2014:852748. doi: 10.1155/2014/852748

3. Bankhead P, Fernández JA, McArt DG, Boyle DP, Li G, Loughrey MB, et al. Integrated tumor identification and automated scoring minimizes pathologist involvement and provides new insights to key biomarkers in breast cancer. *Lab Invest* (2018) 98(1):15–26. doi: 10.1038/labinvest.2017.131

4. Zhang Z, Wang L, Wang S. Automated scoring system of HER2 in pathological images under the microscope (Accessed 2022 Jun 6). arxiv.org/abs/2110.12900.

5. Chervoneva I, Peck AR, Yi M, Freydin B, Rui H. Quantification of spatial tumor heterogeneity in immunohistochemistry staining images. *Bioinformatics* (2021) 37(10):1452–60. doi: 10.1093/bioinformatics/btaa965

6. Failmezger H, Zwing N, Tresch A, Korski K, Schmich F. Computational tumor infiltration phenotypes enable the spatial and genomic analysis of immune infiltration in colorectal cancer. *Front Oncol* (2021) 11:552331. doi: 10.3389/fonc.2021.552331

7. Natrajan R, Sailem H, Mardakheh FK, Arias Garcia M, Tape CJ, Dowsett M, et al. Microenvironmental heterogeneity parallels breast cancer progression: A histology–genomic integration analysis. *PloS Med* (2016) 13(2):e1001961. doi: 10.1371/journal.pmed.1001961

8. Yuan Y, Failmezger H, Rueda OM, Ali HR, Gräf S, Chin SF, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci Transl Med* (2012) 4(157):157ra143. doi: 10.1126/scitranslmed.3004330

9. Pekar G, Gere M, Tarjan M, Hellberg D, Tot T. Molecular phenotype of the foci in multifocal invasive breast carcinomas: intertumoral heterogeneity is related to shorter survival and may influence the choice of therapy. *Cancer* (2014) 120 (1):26–34. doi: 10.1002/cncr.28375

10. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern* (1973) SMC-3(6):610–21. doi: 10.1109/TSMC.1973.4309314

11. Schmidt G, Brieu N, Spitzmüller A, Kapil A. *SCORING METHOD FOR AN ANTI-HER2 ANTIBODY-DRUG CONJUGATE THERAPY*. WIPO PCT WO2022054009 (A2). (World Intellectual Property Organization).

12. Harder N, Athelogou M, Hessel H, Buchner A, Schönmeyer R, Schmidt G, et al. Co-Occurrence features characterizing gland distribution patterns as new prognostic markers in prostate cancer whole-slide images. In: *2016 IEEE 13th international symposium on biomedical imaging (ISBI)* (2016). p. 807–10. doi: 10.1109/ISBI.2016.7493389

13. Hofner B, Boccuto L, Göker M. Controlling false discoveries in high-dimensional situations: Boosting with stability selection. *BMC Bioinf* (2015) 16:144. doi: 10.1186/s12859-015-0575-3

14. Venables WN, Ripley BD. *Modern applied statistics with s [Internet]*. Fourth. (New York: Springer) (2002). Available at: https://www.stats.ox.ac.uk/pub/MASS4/.

15. Simon RM, Subramanian J, Li MC, Menezes S. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Brief Bioinform* (2011) 12(3):203–14. doi: 10.1093/bib/bbr001

16. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. (USA: USENIX Association) (2016). p. 265–83.

17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* (2011) 12:2825–30.

18. Zeileis A, Hothorn T. Diagnostic checking in regression relationships. *R News* (2002) 2(3):7–10.

19. Hofner B, Hothorn T. *Stabs: Stability selection with error control* (2021). Available at: https://CRAN.R-project.org/package=stabs.

20. Nawaz S, Heindl A, Koelble K, Yuan Y. Beyond immune density: critical role of spatial heterogeneity in estrogen receptor-negative breast cancer. *Mod Pathol* (2015) 28(6):766–77. doi: 10.1038/modpathol.2015.37

21. Ellina MI, Bouris P, Aletras AJ, Theocharis AD, Kletsas D, Karamanos NK. EGFR and HER2 exert distinct roles on colon cancer cell functional properties and expression of matrix macromolecules. *Biochim Biophys Acta BBA - Gen Subj* (2014) 1840(8):2651–61. doi: 10.1016/j.bbagen.2014.04.019

22. Wang J, Wang S, Sun J, Qiu L. Expression of c-MET, EGFR and HER-2 in gastric adenocarcinoma tissue and its relationship with clinicopathological characteristics. *Am J Transl Res* (2021) 13(9):10856–62. doi: 10.3389/fonc.2019.00309

23. Joosten SPJ, Spaargaren M, Clevers H, Pals ST. Hepatocyte growth factor/MET and CD44 in colorectal cancer: partners in tumorigenesis and therapy resistance. *Biochim Biophys Acta Rev Cancer* (2020) 1874(2):188437. doi: 10.1016/j.bbcan.2020.188437

24. Liu Y, Yu XF, Zou J, Luo ZH. Prognostic value of c-met in colorectal cancer: A meta-analysis. *World J Gastroenterol WJG* (2015) 21(12):3706–10. doi: 10.3748/wjg.v21.i12.3706

25. Li C, Liu D r, Ye L y, Huang L n, Jaiswal S, Li X w, et al. HER-2 overexpression and survival in colorectal cancer: a meta-analysis. *J Zhejiang Univ Sci B* (2014) 15 (6):582–9. doi: 10.1631/jzus.B1300258

26. Galizia G, Lieto E, Ferraraccio F, De Vita F, Castellano P, Orditura M, et al. Prognostic significance of epidermal growth factor receptor expression in colon cancer patients undergoing curative surgery. *Ann Surg Oncol* (2006) 13(6):823–35. doi: 10.1245/ASO.2006.05.052

27. Wang Z, Tang Y, Xie L, Huang A, Xue C, Gu Z, et al. The prognostic and clinical value of CD44 in colorectal cancer: A meta-analysis. *Front Oncol* (2019) 9:309. doi: 10.3389/fonc.2019.00309

28. Watson D, Clark LA, Chmielewski M, Kotov R. The value of suppressor effects in explicating the construct validity of symptom measures. *Psychol Assess* (2013) 25(3):929–41. doi: 10.1037/a0032781