Quantifying cell-type interactions and their spatial patterns as prognostic biomarkers in follicular lymphoma

DOI:
10.1117/12.2293572

Document Version
Final published version

PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Quantifying cell-type interactions and their spatial patterns as prognostic biomarkers in follicular lymphoma

Anna Maria Tsakiroglou, Sophie Fitzpatrick, Lilli Nelson, Catharine West, Kim Linton, et al.

Anna Maria Tsakiroglou, Sophie Fitzpatrick, Lilli Nelson, Catharine West, Kim Linton, Kang Zeng, Garry Ashton, Sue Astley, Richard Byers, Martin Fergie, "Quantifying cell-type interactions and their spatial patterns as prognostic biomarkers in follicular lymphoma," Proc. SPIE 10581, Medical Imaging 2018: Digital Pathology, 105810G (6 March 2018); doi: 10.1117/12.2293572

SPIE.

Event: SPIE Medical Imaging, 2018, Houston, Texas, United States

Quantifying cell-type interactions and their spatial patterns as prognostic biomarkers in follicular lymphoma

Anna Maria Tsakiroglou[a], Sophie Fitzpatrick[a], Lilly Nelson[a], Catharine West[a,b], Kim Linton[a,b], Kang Zeng[c,d], Garry Asthon[c,d], Sue Astley[a,b,c,e,f], Richard Byers[a,g], and Martin Fergie[a]

[a]University of Manchester, Oxford Rd M13 9PL, Manchester, UK
[b]The Christie NHS Foundation Trust, Wilmslow Rd M20 4BX, Manchester, UK
[c]Manchester Cancer Research Centre, Wilmslow Road M20 4GJ, Manchester, UK
[d]Cancer Research UK, Wilmslow Road M20 4BX, Manchester, UK
[e]Manchester Breast Centre, Wimslow Rd M20 4BX, Manchester, UK
[f]Nightingale Centre, Southmoor Rd Wythenshawe M23 9QZ, Manchester, UK
[g]Manchester University NHS Foundation Trust, Oxford Rd M13 9WL, Manchester, UK

ABSTRACT

Background: Observing the spatial pattern of tumour infiltrating lymphocytes in follicular lymphoma can lead to the development of promising novel biomarkers for survival prognosis. We have developed the "Hypothesised Interactions Distribution" (HID) analysis, to quantify the spatial heterogeneity of cell type interactions between lymphocytes in the tumour microenvironment. HID features were extracted to train a machine learning model for survival prediction and their performance was compared to other architectural biomarkers. Scalability of the method was examined by observing interactions between cell types that were identified using 6-plexed immunofluorescent staining. Methods: Two follicular lymphoma datasets were used in this study; a microarray with tissue cores from patients, stained with CD69, CD3 and FOXP3 using multiplexed brightfield immunohistochemistry and a second tissue microarray, stained with PD1, PDL1, CD4, FOXP3, CD68 and CD8 using immunofluorescence. Spectral deconvolution, nuclei segmentation and cell type classification was carried out, followed by extraction of features based on cell type interaction probabilities. Random Forest classifiers were built to assign patients into groups of different overall survival and the performance of HID features was assessed. Results: HID features constructed over a range of interaction distances were found to significantly predict overall survival in both datasets ($p = 0.0363$, $p = 0.0077$). Interactions of specific phenotype pairs, correlated with unfavourable prognosis, could be identified, such as the interactions between $CD3^+FOXP3^+$ cells and $CD3^+CD69^+$ cells. Conclusion: Further validation of HID demonstrates its potential for development of clinical biomarkers in follicular lymphoma.

Keywords: Follicular lymphoma, prognostic, biomarker, multiplexed immunofluorescence, Hypothesized Interaction Distribution (HID), architectural heterogeneity

1. INTRODUCTION

1.1 Background

Follicular lymphoma is the most common of the low grade Non-Hodgkins lymphomas (NHL), with a 5-year relative survival of 86.5% and a median survival of 12-16 years. It follows an indolent course[1] and most patients go through many periods of remittance and relapse, however it is considered an incurable disease.

A complex interplay is evident between the tumour infiltrating lymphocytes (TILs) in follicular lymphoma and the neoplastic B-cells. There have been many studies linking the heterogeneity of TILs in the tumour microenvironment (TME) of follicular lymphoma to survival. This link supports the potential for novel TME

Further author information: (Send correspondence to A.M.T.)
A.M.T.: E-mail: annamaria.tsakiroglou@postgrad.manchester.ac.uk

Medical Imaging 2018: Digital Pathology, edited by John E. Tomaszewski, Metin N. Gurcan,
Proc. of SPIE Vol. 10581, 105810G · © 2018 SPIE · CCC code:
1605-7422/18/$18 · doi: 10.1117/12.2293572

biomarkers for: survival prognosis; as a means to target end-of-spectrum cases for clinical trials; to select appropriate treatment in the age of personalised medicine; or, for companion diagnostics. TME biomarkers, when compared with conventional clinically used biomarkers such as the FLIPI score, allow a direct observation of the disease mechanisms and interactions between relevant cell types, permitting potential for discovery of new disease mechanisms, disease subtypes or new treatments. Looking at studies searching for clinically useful TME biomarkers in follicular lymphoma, it becomes apparent that the overwhelming majority use cell population numbers or density to quantify the heterogeneity, while only a handful[2,3] mention the potential of observing the spatial pattern as a biomarker. This approach is limited because it does not account for the spatial distribution of cell populations. Additionally, it ignores micro-heterogeneity,[4] the variation of protein expression within "homogeneous" populations of cells.

Microscopy imaging of tissue sections stained with immunohistochemistry (IHC) markers in monoplex or multiplex can preserve the architectural information between different cell phenotypes and it can even localise proteins in a sub-cellular resolution.[5] This functionality makes IHC a good candidate to study spatial heterogeneity as contrary to other methods, such as flow cytometry or single cell sequencing, it preserves the spatial structure of the sample. Multiplexed IHC conjugates antibodies to dies or fluorophores and can be used to concurrently detect multiple antigens in a single frozen or paraffin-fixed tissue section.[5] It can be applied in whole tissue sections or tissue microarrays (TMA) to reduce the associated cost. However, IHC studies may be challenging to validate for a variety of reasons.[6] If the interpretation of the images is done manually, results may be inconsistent because of intra- and inter-observer variability. Choice of area and magnification on the microscope may also skew the results, as could different approaches in staining. These factors produce discrepancies in the reported prognostic significance of several patterns of follicle infiltration in the literature (eg. interfollicular, follicular, perifollicular, diffuse). As an example, in Farinha et al.[3] a diffuse pattern of FOXP3+ cells is shown to correlate with a favourable prognosis, while Lee et al.[7] report that a peri-follicular pattern of the same cells is a favourable indicator of patient outcome. The qualitative way in which these patterns are assessed makes it difficult to compare and reproduce these results. However, a computer automated approach for the quantification of architectural pattern could potentially remedy this problem, as well as aid in the interpretation of highly multiplexed images, where distinguishing the different protein signals by eye becomes impossible.

1.2 Related Work

Recently, several studies[8,9] have shifted their focus into developing computer vision systems specifically for interpreting highly multiplexed IHC, or similar types of images where multiple protein signals can be observed on the same tissue sample, such as the ones obtained using toponome imaging systems[10] (TIS). For multiplexed IHC, commercial solutions are available to spectrally unmix the multiple protein signals, segment the cells and perform further processing steps. Some examples include *inForm* (PerkinElmer), *Nuance* (PerkinElmer) and *Halo* (Indica Labs) and the *Definiens* imaging software. Once the signal from each protein has been isolated using spectral deconvolution or some other method, the cellular compartments are automatically segmented based on a nuclear counterstain. Each cell is then represented as a multidimensional vector of the various protein intensities in each of the sub-cellular components (nucleus, cytoplasm and membrane). This poses the problem of how we identify meaningful cell phenotypes from this stain intensity information that could be recognised by pathologists and linked to their biological functions (e.g. Tregs, natural killer cells).

A simplistic approach to phenotyping is assuming a cell can be either positive or negative for a specific stain, using a specified threshold for cut-off. However, finding the appropriate thresholds to perform this task is not straightforward. Barysenka et al.[11] proposed an information theoretic approach that calculates these thresholds based on the requirement that the intensity signals for different markers in one composite TIS image must be correlated. This method automatically accounts for the variation observed as part of the imaging process. In practice, an empirical approach is often adopted, where the thresholds for the cut-off are determined by a trained pathologist. Another approach to phenotyping proposes clustering, using a locality preserving, nonlinear embedding algorithm and the raw, entire cell, protein intensity vectors as features.[12] Clustering avoids the binarisation of the protein signals, minimising the information loss and allowing the discovery of sub-categories of cell populations and the observation of micro-heterogeneity. A similar approach was proposed Humaun et al.,[13] where phenotyping is carried out by unsupervised hierarchical clustering with a predefined number of 20 clusters. Affinity propagation clustering was also used by Kovacheva et al.[14] to phenotype colon cancer samples

stained with 12 cell markers, while unsupervised K-SVD clustering was used by Spagnolo et al.[15] Dress et al.[16] introduce a series of metrics that can be used in TIS/multivariate images to quantify similarity between pixels and assist on clustering for cell phenotyping. These metrics were used in an interactive visualisation tool, where selecting a pixel automatically highlights all other pixels with similar protein distributions.

Many studies have hypothesised that differential patterns of phenotype interactions have the potential to discriminate between disease sub-types and even predict outcome. The relevance of phenotype colocalisation for biomarker development was supported by a study[17] which found that the numbers of $FOXP3^+$ expressing Tregs are prognostic of good outcome in gastric cancer, only if located within a specified distance of $CD8^+$ T-cells. Spagnolo et al.[15] built networks where only the interacting cells of a sample were connected, where the two cells were considered to interact if found within an specified distance of each other. Based on these networks, the pointwise mutual information (PMI) of each phenotype pair was calculated, to quantify association between phenotypes. It was observed that differences in these associations could discriminate between different types of invasive ductal carcinoma. We have developed the Hypothesized Interactions Distribution (HID) method,[18] to quantify the heterogeneity of cell type interactions in the TME, using multiplexed immunohistochemistry and machine learning. We applied HID for overall survival prediction from right censored data in a follicular lymphoma data set stained for $CD3^+$, $CD69^+$ and $FOXP3^+$, which allowed us to observe interactions between $CD3^+FOXP3^+$ Tregs and other $CD3^+CD69^+$ activated T-cells. Methods that consider protein colocalisation instead of phenotype colocalisation have also been developed. These methods don't rely on the concept of a fixed distance, within which cells can interact. Such a method was introduced by Kovacheva et al.[14] to identify protein-pairs that are specific to cancer in colon tissue TIS images.

1.3 Contribution

The current work studies further the spatial heterogeneity of cell interactions in follicular lymphoma by comparing the performance of the HID method against other methods that observe spatial interactions in the TME, and demonstrates its scalability in more highly multiplexed images, stained with PD1, PDL1, CD4, FOXP3, CD68 and CD8. The architectural relations between further subsets of TILs are examined in this way, and spatial patterns that correlate with overall survival are identified. Finally, a simple automated phenotyping scheme is introduced that relies on thresholding of the unimodal distribution of each stain intensity and can be applied on a per sample basis. The automatically selected thresholds were validated against thresholds manually selected by a pathologist.

2. MATERIALS AND METHODS

2.1 Data sets

For the retrospective analysis that follows we used formalin-fixed, paraffin embedded (FFPE) lymph node samples from a follicular lymphoma cohort of 44 patients from the study of Nelson et al.[2] These samples were obtained with informed consent, with ethical permission granted by the Central Manchester Multi-centre Research Ethical Committee (03/08/016). Overall survival data is available for these patients up to 171 months follow-up and 29.5% (13 patients) of the survival data is censored. None of the patients were treated with Rituximab and the median survival was 4.5 years. All the samples corresponded to initial pre-treatment biopsies at first presentation and were retrieved from the archives of the Christie Hospital, Manchester. After selection of regions of interest by a pathologist (R.J.B.), 56 cores of 1mm diameter and 3mm depth were selected, extracted and used to create tissue micro-arrays (TMA) for the multiplexed IHC experiment. Hereafter, this is referred to as the Nelson dataset, which was used to study interactions between $CD3^+$, $FOXP3^+$ and $CD69^+$ cells.

From the same cohort, a second TMA series were constructed, with the same process, containing 227 cores from 40 patients. This dataset, referred to as the Fitzpatrick dataset, was used to study interactions between cells positive for PD1, PDL1, CD4, FOXP3, CD68 and CD8.

Both TMA series contain multiple cores taken from each biopsy section, as well as multiple slides cut from each core, to account for inconsistencies in the staining and imaging process. Detailed information on the cohort's clinical data is available in table 1.

Table 1. The cohort's clinical data.

Total number of patients	53
Age	35-73 (mean 55) years
Male	54.72%
Female	45.28%
Stage	
I	3
II	5
III	10
IV	11
Unknown	24
FLIPI	
0	2
1	7
2	10
3	8
4	2
5	0
Unknown	24
Bone marrow involvement	
Yes	11
No	17
Unknown	25

2.2 Multiplexed Staining

The slides cut from the TMA of the Nelson dataset, underwent an indirect, bright-field, IHC 3-plex protocol on the Ventana Benchmark automated staining system (Ventana Systems, Arizona, USA). The primary antibodies used were CD3, FOXP3 and CD69, each paired with an appropriate secondary, raised against the host species of the primary. For illumination, chromogenic dyes (Dyomics Blue, Vector Red and BrightDAB, BrightVision plus kit) and a hematoxylin counterstain were applied. Antigen retrieval was heat induced. A detailed protocol for this experiment can be found in Nelson et al.[2] Once stained, the cores were scanned into individual images with .im3 spectral cube format and 20x magnification, using the Vectra slide scanner (PerkinElmer).

Slides cut from the TMA of the Fitzpatrick dataset underwent an indirect immunofluorescent IHC 6-plex protocol. The Ventana Benchmark stainer was employed to run repeated cycles of pre-treatment, blocking, incubation with a primary and a secondary antibody, and finally application of a fluorophore as a detection label. All primaries were raised in mouse, except PDL1, which was raised in rabbit. The use of Opal 7-Color Automation IHC kit (PerkinElmer), which relies on tyramide signal amplification (TSA) technology, was used. Table 2 contains detailed information on the protocol, fluorophore spectra and specifications of the reagents used. The slides were counterstained with DAPI to illuminate the nuclei of all cells and scanned with the Vectra slide scanner, in similar fashion to the Nelson dataset.

2.3 Pre-processing, Segmentation and Cell Phenotyping

Nelson dataset: Multi-spectral imaging in the visible spectrum was used for multiplexing and spectral unmixing was performed as described in Nelson et al.[2] *Nuance* (PerkinElmer) software was used to locate the centroid coordinates of the cells and mark them as positive or negative for each of the stains, using empirical thresholding by a pathologist. Thus, a binary vector of positivity for each stain was assigned to each cell. For N stains, $2^N - 1$ combinations of stain positivity are possible, since the cells negative for all stains cannot be observed. Each of these combinations was considered as a distinct cell phenotype. At the end of the pre-processing steps, for each sample, a list of cell coordinates and phenotype labels for each cell were available.

Fitzpatrick dataset: *inForm* (PerkinElmer) software was used for the spectral deconvolution and segmentation. The cells in these samples were densely packed and an object based segmentation algorithm was selected. The DAPI counterstain component was selected to segment the nuclear compartment, while cytoplasmic segmentation was based on two of the markers that illuminate the cytoplasm, namely PD1 and CD8. The area

Table 2. All steps in the 6-plex immunofluorescence staining protocol were carried out by the Ventana autostainer, except the final washing step and application of counterstain. At the time this protocol was performed, the Ventana software did not support a 6-plex configuration, thus the procedure was coded as three sequential 2-plex protocols. HIER: Heat induced epitope retrieval, CC: Cell Conditioning, HRP: Horseradish peroxidase, TSA: Tyramide signal amplification.

STEP	Description	Time (min)
Deparaffinisation	3 cycles of 8 min tissue deparaffinisation at 69°C	24
HIER	HIER carried out at 91°C while the slides were incubated with CC solution (pH = 6)	32
Blocking	Incubation with DISCOVERY Inhibitor	16
Primary antibody: Anti-CD4	Incubation with primary antibody at 37°C	32
Secondary antibody	Incubation with DISCOVERY OmniMap Anti-mouse secondary, conjugated to HRP	16
Detection label: Opal 650	Incubation with Opal fluorophore, diluted at 1:75 with TSA solution	16
HIER	HIER carried out at 95°C while the slides were incubated with CC solution (pH = 9)	16
Blocking	Incubation with DISCOVERY Inhibitor	16
Primary antibody: Anti-CD8	Incubation with primary antibody at 37°C	32
Secondary antibody	Incubation with DISCOVERY OmniMap Anti-mouse secondary, conjugated to HRP	16
Detection label: Opal 540	Incubation with Opal fluorophore, diluted at 1:75 with TSA solution	16
HIER	HIER carried out at 95°C while the slides were incubated with CC solution (pH = 9)	16
Blocking	Incubation with DISCOVERY Inhibitor	16
Primary antibody: Anti-CD68	Incubation with primary antibody at 37°C	32
Secondary antibody	Incubation with DISCOVERY OmniMap Anti-mouse secondary, conjugated to HRP	16
Detection label: Opal 620	Incubation with Opal fluorophore, diluted at 1:75 with TSA solution	16
HIER	HIER carried out at 95°C while the slides were incubated with CC solution (pH = 9)	16
Blocking	Incubation with DISCOVERY Inhibitor	16
Primary antibody: Anti-FOXP3	Incubation with primary antibody at 37°C	32
Secondary antibody	Incubation with DISCOVERY OmniMap Anti-mouse secondary, conjugated to HRP	16
Detection label: Opal 570	Incubation with Opal fluorophore, diluted at 1:75 with TSA solution	16
HIER	HIER carried out at 95°C while the slides were incubated with CC solution (pH = 9)	16
Blocking	Incubation with DISCOVERY Inhibitor	16
Primary antibody: Anti-PD1	Incubation with primary antibody at 37°C	32
Secondary antibody	Incubation with DISCOVERY OmniMap Anti-mouse secondary, conjugated to HRP	16
Detection label: Opal 520	Incubation with Opal fluorophore, diluted at 1:75 with TSA solution	16
HIER	HIER carried out at 95°C while the slides were incubated with CC solution (pH = 9)	16
Blocking	Incubation with DISCOVERY Inhibitor	16
Primary antibody: Anti-PDL1	Incubation with primary antibody at 37°C	32
Secondary antibody	Incubation with DISCOVERY OmniMap Anti-rabbit secondary, conjugated to HRP	16
Detection label: Opal 690	Incubation with Opal fluorophore, diluted at 1:75 with TSA solution	16
Washing	Slides are submerged in EZ preparation solution for 3x5 min to remove oil coverslip	15
Washing	Slides are washed with water	5
Counterstain DAPI	Slides are coverslipped with aqueous Prolong Anti-Gold mounting agent with DAPI	4

of cells was determined to range between 70-700 pixels and refined splitting was applied after segmentation to break up clusters of closely packed cells that had been detected as one. The image of each core was exported from *inForm* as a collection of cell coordinates, and for each cell the mean intensity of the 6 markers in the cytoplasmic and the nuclear compartment were available.

High variance in the range of intensities for all markers between different slides and between different TMA cores within the same slide prohibited the use of a single set of thresholds for cell phenotyping. Instead, thresholds were determined in a per-sample basis from the histograms of mean stain intensities of the nuclear or cytoplasmic compartments. It was observed that in all samples these histograms followed unimodal distributions; a single peak close to zero would represent noise, i.e. all the cells with non specific background staining, while the true positive cells would be located at the right tail of the distribution. We employed the maximum deviation thresholding algorithm for unimodal images as proposed in Rosin et al.[19] to calculate a threshold for each sample as demonstrated in figure 1.

Table 3. Pearson's Coefficient to assess linear correlation between manually and automatically selected thresholds.

Antibody	Stain intensity	Pearson's R
PD1	Cytoplasm Opal 520 Mean	0.941
CD8	Cytoplasm Opal 540 Mean	0.782
FOXP3	Nucleus Opal 570 Mean	0.786
CD68	Cytoplasm Opal 620 Mean	0.871
CD4	Cytoplasm Opal 650 Mean	0.993
PDL1	Cytoplasm Opal 690 Mean	0.992

To validate the automatic threshold calculation, we manually selected thresholds for each stain for 20 samples from our dataset using the unmixed component views generated by the *inForm* software. We used RANSAC regression analysis to validate the automatic thresholds, and found that there was a strong linear correlation with the manually selected thresholds (see table 3). Therefore, to further improve the accuracy of phenotyping, we applied a linear correction to the automatic thresholds using the regression parameters.

The nine most frequent cell types that resulted from applying this phenotyping scheme, $PDL1^+$, $CD4^+PDL1^+$, $CD4^+$, $CD68^+$, $FOXP3^+$, $CD4^+FOXP3^+$, $CD8^+$, $PD1^+CD4^+$ and $PD1^+$ were considered in the analysis that follows. The rest had frequencies less than 2% and were considered artefacts which may occur because of the non-specific nature of cytoplasmic staining, bleed through from other components during spectral unmixing or suboptimal thresholds selected in the phenotyping stage.

2.4 Spatial Architecture Feature Extraction

HID and PMI features are extracted, as described in[18] and,[15] respectively. A cell interaction is hypothesized to occur whenever two cells fall within a distance threshold d of each other. If the position and phenotype label of each cell are represented as $\mathbf{x_i}$ and l_i , respectively, then each element of the HID feature vector is computed as follows:

$$H_{i,j} = |\{(i,j)|\mathbf{x_i} \in C^i, \mathbf{x_j} \in C^j, ||\mathbf{x_i} - \mathbf{x_j}||_2 < d\}| \,. \tag{1}$$

In Eq. (1) C^i and C^j are the sets of cells with phenotype label l_i and l_j respectively. Thus each element of the HID matrix contains the number of occurrences where a cell with phenotype i is less than d pixels from a cell of phenotype j. The symmetric HID matrix is normalised, so that its upper-triangle sums to one and its elements then represents the probabilities of observing an interaction between each pair of phenotypes.

Furthermore, a PMI feature vector can be calculated, where each element represents the pointwise mutual information for a pair of phenotypes:

$$PMI_{i,j} = log\frac{p_s(i,j)}{p_t(i)p_t(j)} \,, \tag{2}$$

where $p_s(i,j)$ is the probability of observing an interaction between phenotypes i and j in a single sample, whereas $p_t(i)$ and $p_t(j)$ refer to the probabilities of occurrence of phenotypes i and j in all samples. In the case of the PMI vector features, the interaction distance threshold is calculated from the mean distance of each cell from its 10 nearest neighbours.

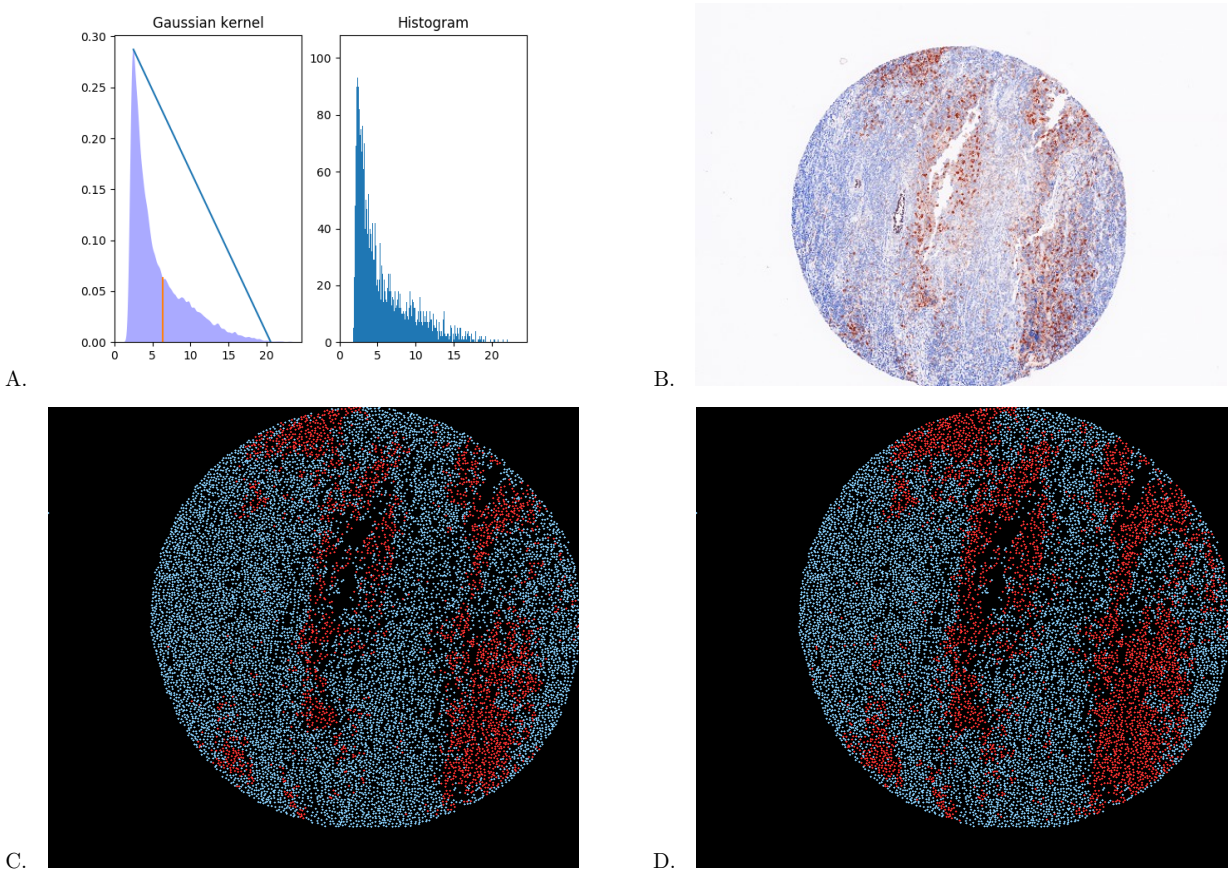

Figure 1. Automated threshold selection for cytoplasmic stain PD1 (Opal 520). A: First, a Gaussian kernel was applied to smooth the histogram and a line was drawn from the peak of the distribution to the end of the tail. Then, the appropriate threshold was selected as the point in the smoothed histogram that had the maximum perpendicular distance from the line. B: Unmixed component view in inForm. C: The result of phenotyping using the manually selected threshold of 8.5. Cells annotated in red are positive for PD1, while blue are stained with DAPI only. D: Phenotyping using the automatically selected threshold of 6.37.

While both the PMI and HID features operate on the basis of a single selected interaction distance, when looking at patterns between multiple pairs of phenotypes, it is possible that different patterns and therefore different interaction distances could be considered most informative per phenotype pair. A new set of HID type features is therefore proposed, the cumulative HID (cHID) to capture this information. For each combination of phenotypes, we build a cumulative histogram of the number of interactions between a combination of cell phenotypes over a range of distances. The number or interactions at any given distance are normalised by the total number of possible interactions contained within the sample.

2.5 Random Forests for Survival Prediction

Three Random Forest (RF) classifiers were trained to predict a binary risk score for the patients, using either HID, PMI or cHID feature vectors. As ground truth for training, target groups were defined by splitting the patient survival data at the median value. By attempting to classify each patient correctly into the short or long survival groups, the RF would learn to identify which features are most important in predicting overall survival. The median survival in the Nelson and Fitzpatrick datasets was 53 and 55 months respectively. The scoring function of the RF was classification accuracy and the criterion for selecting a threshold for the best split at each node was Gini impurity.

Training and scoring the performance of the RF was carried out on a per sample basis, and predictions were generated by leave one out cross validation. However, to avoid having the model over fit on patient specific characteristics during training, the cross validation scheme was stratified per patient. This way, in each fold all the samples for a patient would be found on either the train set or the test set, but not simultaneously in both. Predictions from all samples of a single patient were treated as equally weighted votes, and the aggregation of risk scores was carried out by selecting the most frequently occurring vote.

In the case of the HID vector features, the interaction distance threshold was chosen within the training fold during the cross validation procedure.

3. RESULTS

3.1 Risk prediction in the Nelson dataset

Figure 2 shows that the RF trained with HID features slightly out-performed the RF with PMI and cHID features when performing a Kaplan-Meier analysis on the predicted patient groups. In all cases RF with 100 trees were sufficient to discriminate well between high and low risk patients in the test set. The HID performed optimally for interaction distances of 65 microns, while the PMI selected a distance of 350 microns. The binary classification predictions were aggregated for all samples of a single patient. Thus a categorical, high or low risk index was generated. All three risk indices was were significant in univariate Cox proportional hazards regression analysis, presented in Table 4. The confidence intervals in the Kaplan-Meier plots were calculated based on Greenwoods estimator of the survival function's variance, with $\alpha = 95\%$. Representative samples for the long and short survival groups are presented in figure 3. Feature importance was assessed for the three random forests, and the features that are associated with the following interactions are found to have the most impact in the performance of all classifiers; (i) interactions between cells both positive for FOXP3 and (ii) interactions between cells positive for FOXP3 and positive for CD69. Favourable outcome was observed when type (i) interactions were infrequent and type (ii) interactions were frequent.

Single scores were also used to summarise the HID and PMI feature vectors in the Nelson dataset. These scores might lead to some information loss, associated with reducing the dimensions of the feature vectors, but are nevertheless useful, as they could be easily integrated into conventional Kaplan-Meier analysis. These scores represent measures of heterogeneity of the spatial interactions, overall. The Shannon entropy was used to summarise the HID features and the PMI HET[15] score to summarise the PMI features. The survival log rank test scores for the predicted splits were $p = 0.00004$, $\chi^2 = 16.59$ and $p = 0.0002$, $\chi^2 = 13.43$, respectively. Higher scores, corresponding to more heterogeneous interactions were associated with favourable outcome.

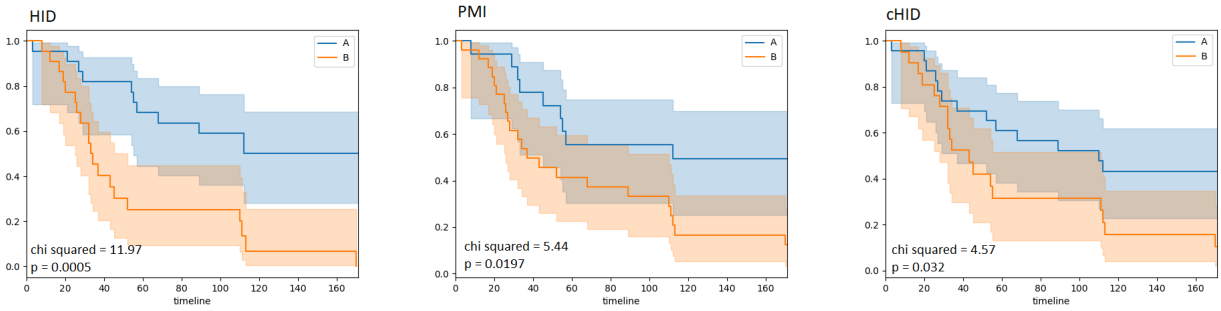

Figure 2. Kaplan-Meier curves and results for the survival log rank test between the two predicted groups of patients. The predictions were generated with leave one out cross-validation using the Random Forest with HID vector features, Random forest with PMI vector features and Random forest with cHID vector features. Group A corresponds to the positive class with the highest probability of survival. The timeline refers to months.

3.2 Risk prediction in the Fitzpatrick dataset

The RF were trained with HID, PMI or cHID features and 1000 trees on the Fitzpatrick dataset, where the phenotypes considered were nine. Predictions from multiple samples of the same patient were aggregated and

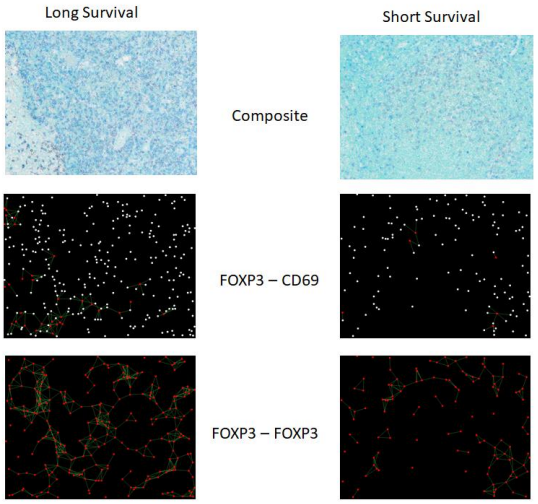

Figure 3. For representative patients from the long and short survival group, the composite images and the graphs of interactions between FOXP3$^+$-FOXP3$^+$ and FOXP3$^+$-CD69$^+$ cells are presented.

risk indices were generated, as in the Nelson dataset. However, in this dataset it was observed that HID and PMI features failed to correlate with overall survival in the univariate Cox regression, as seen in table 4. Only the RF trained with cHID features, could significantly predict survival. In this classifier, higher importance was given to features concerning interactions between cells positive for PD1, FOXP3, CD68 and double positive for FOXP3 and CD4.

Table 4. Univariate Cox Regression for the risk scores predicted by the Random Forests. For patients with multiple samples the risk scores have been aggregated by selecting the most frequently occurring prediction, before fitting the Cox model. The concordance index (CI) provides the equivalent of AUC estimate for survival analysis.

Fitzpatrick Dataset		Hazard Ratio	p	Lower 0.95	Upper 0.95	Mean CI (10-fold CV)
40 Patients	cHID score	0.54	0.00767	0.34	0.85	0.59
30 Events	HID score	0.78	0.21340	0.53	1.15	0.53
	PMI score	0.74	0.13650	0.50	1.10	0.57
Nelson Dataset		**Hazard Ratio**	**p**	**Lower 0.95**	**Upper 0.95**	**Mean CI (10-fold CV)**
44 Patients	cHID	0.68	0.03631	0.47	0.98	0.59
31 Events	HID score	0.50	0.00048	0.34	0.74	0.69
	PMI score	0.62	0.01746	0.42	0.92	0.65

4. DISCUSSION

The phenotyping method that we proposed, based on automated unimodal thresholding of stain intensity thresholds, offers many benefits. It is a simple, quick method which can be applied even in cases where staining is highly variable between samples in the dataset. Even if it does not offer the level of sophistication of clustering methods, it has the advantage that its results can be easily validated. The phenotypes produced using this method are furthermore easily linked to their biological function by a pathologist. When applied to the Fitzpatrick dataset only few phenotypes were observed in the resulting distribution with high frequencies, which points to the accuracy of the method.

The RF trained with HID, PMI and cHID features further validate that information encoded in the spatial interactions in the TME have the potential to predict outcome for follicular lymphoma. Even though the HID method has been applied for outcome prediction in Follicular lymphoma in the past, this is the first time, to our knowledge, that the PMI method has been used for the same application. Furthermore, the extended cumulative

HID method was introduced, which does not require a selection of interaction distance parameter, and was shown to perform robustly in both datasets, scaling successfully to a higher number of phenotypes. The accuracy of cell identification in highly multiplexed systems is often not optimal, which is to be expected, as the repeated cycles of staining wear the tissue and co-localisation of stains presents a challenge to spectral unmixing. The cHID could however overcome obstacles presented by the added noise in the data and significantly predict survival in a univariate Cox Regression for both dataset. It is worth mentioning that the hazard ratios in table 4 represent the death rate associated with the low risk group relative to the high risk group, as a higher predicted score by the RF corresponds to favourable outcome.

The features selected by the RF as most significant were all associated with Tregs, macrophages, and PD1 positive cells, which agrees to previous studies observing architecture as a biomarker in follicular lymphoma.[3,7] Further validation in larger datasets of follicular lymphoma patients would be necessary, however, in order to develop biomarkers that could be clinically relevant. Not having sufficient clinical data to compare the cHID with other clinically used risk scores, such as the FLIPI, presents a limitation for this analysis.

ACKNOWLEDGMENTS

Acknowledgments to Steve Bagley, Syed J. Islam, Chris Rose, and everyone who assisted with the collection and preparation of this dataset.

REFERENCES

[1] Ghielmini, M. and Montoto, S., [*Lymphomas Essentials for Clinicians*], ESMO Press, Viganello-Lugano (2012).

[2] Nelson, L. S., Mansfield, J., Lloyd, R., Oguejiofor, K., Salih, Z., Menasce, L. P., Linton, K. M., Rose, C. J., and Byers, R., "Automated prognostic pattern detection shows favourable diffuse pattern of FOXP3+ Tregs in follicular lymphoma," *Br. J. Cancer* **113**(8), 1197–1205 (2015).

[3] Farinha, P., Al-Tourah, A., Gill, K., Klasa, R., Connors, J. M., and Gascoyne, R. D., "The architectural pattern of FOXP3-positive T cells in follicular lymphoma is an independent predictor of survival and histologic transformation," *Blood* **115**(2), 289–295 (2010).

[4] Gough, A., Stern, A. M., Maier, J., Lezon, T., Shun, T.-Y., Chennubhotla, C., Schurdak, M. E., Haney, S. A., and Taylor, D. L., "Biologically relevant heterogeneity: Metrics and practical insights," *SLAS Discov.* **22**(3), 213–237 (2017).

[5] Stack, E. C., Wang, C., Roman, K. A., and Hoyt, C. C., "Multiplexed immunohistochemistry, imaging, and quantitation: A review, with an assessment of Tyramide signal amplification, multispectral imaging and multiplex analysis," *Methods* **70**(1), 46–58 (2014).

[6] Solal-Céligny, P., Cahu, X., and Cartron, G., "Follicular lymphoma prognostic factors in the modern era: What is clinically meaningful?," *Int. J. Hematol.* **92**(2), 246–254 (2010).

[7] Lee, A. M., Clear, A. J., Calaminici, M., Davies, A. J., Jordan, S., MacDougall, F., Matthews, J., Norton, A. J., Gribben, J. G., Lister, T. A., and Goff, L. K., "Number of CD4+ cells and location of forkhead box protein P3-positive cells in diagnostic follicular lymphoma tissue microarrays correlates with outcome.," *J Clin Oncol* **24**(31), 5052–9 (2006).

[8] Raza, S. E. A., Langenkämper, D., Sirinukunwattana, K., Epstein, D., Nattkemper, T. W., and Rajpoot, N. M., "Robust normalization protocols for multiplexed fluorescence bioimage analysis," *BioData Min.* **9**(1), 11 (2016).

[9] Schüffler, P. J., Schapiro, D., Giesen, C., Wang, H. A. O., Bodenmiller, B., and Buhmann, J. M., "Automatic single cell segmentation on highly multiplexed tissue images," *Cytometry Part A* **87**(10), 936–942 (2015).

[10] Schubert, W., Bonnekoh, B., Pommer, A. J., Philipsen, L., Böckelmann, R., Malykh, Y., Gollnick, H., Friedenberger, M., Bode, M., and Dress, A. W. M., "Analyzing proteome topology and function by automated multidimensional fluorescence microscopy.," *Nat. Biotechnol.* **24**(10), 1270–1278 (2006).

[11] Barysenka, A., Dress, A. W. M., and Schubert, W., "An information theoritic thresholding method for detecting protein colocalizations in stacks of fluorescent images," *J. Biotechnol.* **149**(3), 127–131 (2010).

[12] Khan, A. M., Humayun, A., Raza, S. E. A., Khan, M., and Rajpoot, N. M., "A novel paradigm for mining cell phenotypes in multi-tag bioimages using a locality preserving nonlinear embedding," in [*Neural Information Processing*], Huang, T., Zeng, Z., Li, C., and Leung, C. S., eds., *Lecture Notes in Computer Science* **7666**, 575–583 (2012).

[13] Humayun, A., Raza, S., Waddington, C., Abouna, S., Khan, M., and Rajpoot, N. M., "A novel framework for molecular co-expression pattern analysis in multi-channel toponome fluorescence images," *Proc. MIAAB 2011*, 109–112 (2011).

[14] Kovacheva, V. N., Khan, A. M., Khan, M., Epstein, D. B. A., and Rajpoot, N. M., "DiSWOP: A novel measure for cell-level protein network analysis in localized proteomics image data," *Bioinformatics* **30**(3), 420–427 (2014).

[15] Spagnolo, D. M., Gyanchandani, R., Al-Kofahi, Y., Stern, A. M., Lezon, T. R., Gough, A., Meyer, D. E., Ginty, F., Sarachan, B., Fine, J., Lee, A. V., Taylor, D. L., and Chennubhotla, S. C., "Pointwise mutual information quantifies intratumor heterogeneity in tissue sections labeled with multiple fluorescent biomarkers," *J Pathol Inform* **7**(47) (2016).

[16] Dress, A., Lokot, T., Schubert, W., and Serocka, P., "Two theorems about similarity maps," *Ann. Comb.* **12**(3), 279–290 (2008).

[17] Feichtenbeiner, A., Haas, M., Büttner, M., Grabenbauer, G. G., Fietkau, R., and Distel, L. V., "Critical role of spatial interaction between CD8+ and Foxp3 + cells in human gastric cancer: The distance matters," *Cancer Immunol. Immunother.* **63**(2), 111–119 (2014).

[18] Rose, C. J., Naidoo, K., Clay, V., Linton, K. M., Radford, J. A., and Byers, R. J., "A statistical framework for analyzing hypothesized interactions between cells imaged using multispectral microscopy and multiple immunohistochemical markers," *J Pathol Inform* **4**(4) (2013).

[19] Farinha, P., Al-Tourah, A., Gill, K., Klasa, R., Connors, J. M., and Gascoyne, R. D., "Unimodal thresholding," *Pattern Recognit.* **34**(11), 2083–2096 (2001).