PROGNOSTIC INSIGHTS FROM MULTIPLEXED SPATIAL PROFILING OF THE TUMOUR MICROENVIRONMENT

A THESIS SUBMITTED TO
THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF BIOLOGY MEDICINE AND HEALTH

2021

ANNA-MARIA TSAKIROGLOU

SCHOOL OF MEDICAL SCIENCES
DIVISION OF CANCER SCIENCES

List of Contents

Word count: 41,635

List of Figures

List of Tables

List of Abbreviations

AF	Autofluorescence
AJCC	American Joint Committee on Cancer
AP	Average precision
ASCO/CAP	American Society of Clinical Oncology and College of American Pathologists
BM	Bone marrow
CAS	Computer assisted scoring
CEP17	Centromere of chromosome 17
CI	Confidence intervals
CISH	Chromogenic in situ hybridisation
CMP	Combinatorial molecular phenotypes
CNN	Convolutional neural network
CoV	Coefficient of variation
CyTOF	Cytometry by time of flight
DAB	3,3'-diaminiobenzidine
DAPI	4′, 6-diamidino-2-phenylindole
DFS	Disease free survival
DLBCL	Diffuse large B-cell lymphoma
ECOG	Eastern Cooperative Oncology Group performance status
EGFR	Epidermal growth factor receptor
ENS	Extra-Nodal Sites
ER	Oestrogen receptor-α
FDC	Follicular dendritic cells
FFPE	Formalin-fixed and paraffin embedded
FISH	Fluorescent in situ hybridisation
FL	Follicular lymphoma
FLIPI	Follicular lymphoma international prognostic index
FN	False negative
FP	False positive
H&E	Haematoxylin & eosin staining
Hb	Haemoglobin
HER2	Human epidermal growth factor receptor 2
HID	Hypothesized Interaction Distribution methodology
HPV	Human papillomavirus
HR	Hazard ratio
HRP	Horseradish peroxidase
IF	Immunofluorescence
IHC	Immunohistochemistry
IoU	Intersection over union
IPI	International prognostic index

LDH	Serum lactate dehydrogenase
LRC	Loco-regional control
MELC	Multi-epitope-ligand cartography
NMS	Non-maximum suppression
NS	Nodal sites
OPSCC	Oropharyngeal squamous cell carcinoma
OS	Overall survival
PD-1	Programmed cell death protein 1
PFS	Progression free survival
PH	Proportional hazards
PI3K	Phosphoinositide 3-kinase
POD24	Progression of disease within 24 months of starting treatment
R	Rituximab
R-CHOP	Rituximab, cyclophosphamide, doxorubicin hydrochloride (hydroxydaunorubicin), vincristine sulphate and prednisone
R-CVP	Rituximab, cyclophosphamide, vincristine sulphate, and prednisone
ROI	Regions of interest
SCCHN	Squamous cell carcinoma of the head and neck
T-regs	T regulatory cells
TAM	Tumour associated macrophages
TCR	T-cell receptor
T_{FH}	T follicular helper cells
TIL	Tumour infiltrating lymphocytes
TMA	Tissue microarray
TME	Tumour microenvironment
TMTV	Total metabolic tumour volume
TP	True positive
TR	Treatment response
TSA	Tyramide signal amplification
TTT	Time to transformation
WSI	Whole slide images
WW	Watchful waiting

Abstract

Fulfilling the promise of cancer immunotherapy would benefit from novel biomarkers to characterise the tumour microenvironment and risk-stratify patients. Multiplex immunofluorescence imaging methods are suitable for this task, enabling visualisation of multiple proteins on the same tissue section. This facilitates identification of multiple cell phenotypes based on the proteins they express, while preserving spatial context. The aim of this thesis was to explore how multiplexed and spatial profiling of the tumour microenvironment in two types of cancer, oropharyngeal squamous cell carcinoma (OPSCC) and follicular lymphoma (FL), can be utilised for biomarker development.

Computer assisted scoring tools are necessary to analyse multiplex images, as drawing conclusions from the large amount of information available is challenging. Thus, the first contribution of this work was to define design requirements of computer assisted scoring tools and assess their performance. Systematic review of the literature identified six design requirements: definability, accuracy, reproducibility, time-efficiency, interpretability and accurate confidence estimation. A meta-analysis of several HER2 and ER scoring studies established that automated scoring agreed with manual scoring, similar to how well pathologists usually agreed with each other.

The second contribution of this work was the introduction of a prognostic biomarker in OPSCC using multiplex immunofluorescence and a new computer assisted scoring system to observe spatial proximity between cell phenotypes. Frequent spatial proximity between cells known to interact during the PD-1/PD-L1 immune escape pathway was unfavourable in patients with HPV negative OPSCC.

The final contribution was the development and validation of a 7-plex immunofluorescent panel for FL and an automated scoring system to study the diversity of immune populations and their spatial relationships. Increased diversity of cell types and cell spatial interactions were favourable in multivariable analyses.

These findings underline the importance of the tumour microenvironment in prognosis of FL and OPSCC and merit further exploration in additional cohorts. The design requirements identified can be applied to guide further validation and establish clinical applicability of the proposed automated scoring systems.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Parts of this thesis have been published in the following:

- Tsakiroglou AM, Fergie M, Oguejiofor K, et al. Spatial proximity between T and PD-L1 expressing cells as a prognostic biomarker for oropharyngeal squamous cell carcinoma. *Br J Cancer*. 2019;(October):1-6. doi:10.1038/s41416-019-0634-z
- Tsakiroglou AM, West C, Astley S, Linton K, Fergie M, Byers R. Automated Multi-plex Immunofluorescence with TSA for CD4, CD8, FOXP3, CD21, PD1 and CD68 in Follicular Lymphoma. Published online 2019. doi:10.17504/protocols.io.49ygz7w (Experimental protocol)
- Tsakiroglou AM, Fergie M, West C, et al. Quantifying cell-type interactions and their spatial patterns as prognostic biomarkers in follicular lymphoma. In: *SPIE Proceedings Medical Imaging*. Vol 10581. ; 2018:15. doi:10.1117/12.2293572
- Tsakiroglou, Anna Maria Astley S, Dave M, Fergie M, et al. Tumour Infiltrating Lymphocytes in Follicular Lymphoma - additional data. Mendley data. doi:10.17632/274xbhc5rx.3 (Dataset)

Copyright

Acknowledgements

Completion of this work would not be possible without the support of my advisors; Dr. Sue Astley, Dr. Martin Fergie, Dr. Kim Linton, Prof. Catherine West and Dr. Richard Byers. I am grateful for your guidance and encouragement. You contributed to my research journey in a very substantial way, and I would like to express a heartfelt thank you.

I would also like to acknowledge and express my gratitude to:

- the histology and microscopy core facility staff of CRUK Manchester Centre for the generous support and training they provided.
- the Christie NHS Foundation Trust Leukaemia and Lymphoma team for their help during patient recruitment and sample collection.
- my collaborators, Dr. K. Oguejiofor, Dr. P. Stern, Dr. I. Peset-Martin, Prof. A. Martel, Dr. M. Dave, Dr. E. Harkness, A. Rosenberg and Dr. M. Sperrin for their support, exchange of ideas and critical feedback
- my colleagues and friends I made along the way, Georgia, Ethan, Luke, Raja, Luca, and the rest of the Stopford crew, for keeping me sane and grounded through the ups and downs of the doctorate journey.

Finally, I would like to thank my parents, brother and sister, and my partner, Panos, for always believing in me, even at times when I had given up. I feel truly blessed by your love and support. This thesis is dedicated to all of you.

Outline

The outline of the thesis is described below.

Chapter 1 is an introduction, outlining the clinical and technical background. Section 1.1 provides clinical background on two cancers, oropharyngeal squamous cell carcinoma and follicular lymphoma, and outlines the need and scope for biomarkers to characterise the tumour microenvironment. Section 1.2 provides the technical background on multiplex tissue imaging and image analysis technologies that can assist in the development of such biomarkers. Section 1.3 summarises the motivation, aims and objectives of this work.

Chapter 2 deals specifically with image analysis tools developed for histopathological scoring, i.e., quantifying the level of protein expression on the tissue and presents a systematic review of the literature.

Chapter 3 introduces a new biomarker in oropharyngeal squamous cell carcinoma, using multiplex and spatial analysis of the tumour microenvironment.

Chapter 4 introduces a new biomarker in follicular lymphoma, using multiplex, spatial and immune diversity analysis. It is split in two ways. Section 4.1 describes development and validation of a novel multiplex assay to observe the immune microenvironment in follicular lymphoma. Section 4.2 describes the image analysis pipeline and biomarker validation.

Chapter 5 presents concluding remarks and directions for future research.

1 Introduction

The paradigm shift in cancer treatment caused by the advent of immunotherapy has improved the lives of millions of people worldwide and has radically changed our understanding of how the disease develops. Cancer researchers now agree that "it takes a village"[1,2] for cancers to grow. That "village" signifies the tumour microenvironment and the dysfunctional community of host cells that infiltrate and surround the tumour. This environment includes immune cells, signalling molecules, extra-cellular matrix and blood vessels, forming a permissive ecosystem that promotes tumour growth and helps tumour cells escape immune detection.

Immune escape signalling pathways in the tumour microenvironment are the target of effective immunotherapies, that help the host immune system recognise and kill tumour cells. Fulfilling the potential of immunotherapy requires novel biomarkers to characterise the tumour microenvironment and assist in treatment selection. Such precision medicine approaches are needed as many new treatments become available. There is a pressing clinical need for biomarker led strategies that involve upfront risk-adapted therapy selection and/or subsequent response adapted therapy escalation/de-escalation to maximise efficacy and minimise toxicity.

Emerging highly multiplexed histopathological assays coupled with digital pathology present new opportunities for cancer biomarker development. This technology can offer an unprecedented amount of information on the tumour microenvironment. Multiplexed assays are the family of techniques that can concurrently visualise the expression of multiple protein targets (up to ~100) on single cells. Multiplexing presents an opportunity for precise recognition of cell phenotypes and functionality, while at the same time preserving the spatial context and producing an image of the tissue, similar to the conventional pathology workflow.

Drawing conclusions from the huge amount of information available from multiplex assays is challenging by manual pathologist assessment. Digital pathology, i.e., the use of digital image analysis for the assessment of tissue biopsies can mitigate this problem. Digital pathology enables new ways to assess multiplex images and draw quantitative conclusions in a way that previously has not been possible.

Thus, new insights into the tumour microenvironment are within reach, including a key aspect that has been generally understudied: the microenvironment's spatial

architecture. Spatial architecture refers to the overall spatial organisation that permeates the tissue, including the formation of tissue compartments with distinct phenotypic profiles, the location of high cell density areas ("hotspots"), and the positioning and spatial distribution of cell phenotypes relative to themselves and to each other.

Spatial architecture can reveal important insights on prognosis, disease development, and response to treatment.[3] Early studies[3] have shown links between the spatial proximity of certain cells and disease outcome. However, the implications of differential spatial architectures and the extent of their clinical usefulness are still not well understood. Furthermore, it is unclear whether spatial pattern is equally important for both solid and haematological tumours.

This thesis addresses the multiplexed and spatial profiling of the tumour microenvironment for novel cancer biomarker discovery. Using multiplexed immunofluorescence assays and automated image analysis, new methodologies to examine the spatial context are investigated and tested for their prognostic value as cancer biomarkers in oropharyngeal cancer and follicular lymphoma. Such methodologies that allow quantitative mapping of the spatial architecture can ultimately achieve a more comprehensive understanding of the heterogeneous tumour microenvironment.

1.1 Clinical background

In this section, the epidemiology, treatment pathways and current prognostic landscape of oropharyngeal squamous cell carcinoma and follicular lymphoma are outlined. Clinical background is provided to determine the unmet clinical need and potential for development of novel prognostic biomarkers based on the spatial architecture of tumours.

1.1.1 Oropharyngeal squamous cell carcinoma

Squamous cell carcinoma of the head and neck (SCCHN) is the most commonly encountered of the head and neck malignancies, affecting the thin, flat squamous cells. These cells form the lining of many organs in the human body, such as the oropharynx, mouth and nose and their functionality is primarily to enable and filter the transportation of molecules to and from these organs. Oropharyngeal squamous carcinoma (OPSCC), is the subset of SCCHN which includes the pharyngeal walls, base of tongue, soft palate and tonsils.[4]

1.1.1.1 Epidemiology, aetiology and treatment

In terms of prevalence, head and neck cancer is the 6th most common cancer globally[5] and 8th in the UK population.[6] The incidence rates in the UK have increased by approximately 40% since the early 1990s in 25-60 year olds.[6] In the UK, most patients are male (70%), while incidence increases with age (usual range 35-90+ years).[6] Prognosis in OPSCC is similar for men and women with a 1-year survival rate of 84% , 5-year survival of 66% and 10-year survival of 56%.[6]

The aetiology of SCCHN is quite well understood today; tobacco and excessive alcohol consumption have been established as the main risk contributors, while their carcinogenic effects seem to be synergistic. However, as not all people who consume tobacco and alcohol will necessarily develop cancer, genetic predisposition and immunosuppression also play a role.[4] Additionally, infection with high risk HPV sub types (primarily HPV-16) has been established as the cause of SCCHN in a subset of patients.[7] HPV^+ patients with SCCHN often have a favourable prognosis and present with different genetic and immune response characteristics. A meta-analysis of 12,263 patients from 44 countries showed that 31.5% of SCCHN patients and 45.8% of OPSCC patients were HPV^+ when assessed with DNA PCR

analysis.[8] Because of their marked differences, it is generally suggested that HPV^+ and HPV^- OPSCC should be considered separate entities.

Treatment for SCCHN usually depends on the stage at diagnosis and patient fitness. For early stages, radiotherapy with or without surgery is usually administered. For advanced stages, surgery may be combined or chemotherapy, radiation or immunotherapy. In patients with only locally advanced disease, platinum-based chemo-radiation is indicated as the standard treatment, while a combination of radiotherapy and cetuximab can be administered if chemotherapy is not tolerated. Finally, in patients with distant metastases or recurrent disease, currently the preferred option is a combination of platinum-based chemotherapy and cetuximab.[9] Cetuximab is a monoclonal antibody that gained FDA approval in 2006 and was designed to target the epidermal growth factor receptor (EGFR). EGFR is expressed in the surface of neoplastic cells in SCCHN and when activated by binding to ligands that normally exist in the body, such as EGF, it promotes proliferation and apoptosis evasion.[10] While cetuximab is established in clinical practice, many other immunotherapy treatments are in development,[11] such as DNA vaccines, adoptive T-cell transfer and immune checkpoint inhibitors (e.g. ipilimumab against CTLA-4 and pembrolizumab, nivolumab, durvalumab against the PD-1/ PD-L1 pathway of immune-escape[12]).

1.1.1.2 Prognostic factors

Historically, the clinically used prognostic indicator for SCCHN is TNM stage, which consists of three components; tumour size, local nodal involvement and presence of distant metastases.[4] However, as novel treatments based in specific targeting of immune escape mechanisms slowly emerge, there is a need for new biomarkers for treatment selection (or de-escalation) that would be tailored to patient specific immune characteristics. One such biomarker is the HPV status. Notably, HPV^+ patients are expected to have 28% lower risk of death than HPV^- OPSCC patients,[13] while HPV^+ OPSCC patients with a higher density of TILs as a whole have been shown to have significantly better disease specific survival.[14] While information on HPV status (p16 protein expression) is used to inform prognosis, no other biomarker is used routinely.

There are, however, numerous publications highlighting potential prognostic factors derived from the tumour microenvironment, ranging from hypoxia,[15] molecular

subtyping,[16] and densities of several tumour infiltrating lymphocyte (TIL) subsets. This thesis aimed to identify such key TIL subsets in the tumour microenvironment of SCCHN and study their spatial patterns. To this end, prior findings on the potential prognostic value of several TIL subsets in head and neck cancer are discussed below.

1.1.1.3 The role of the tumour microenvironment in prognosis

1.1.1.3.1 PD-1 and PD-L1

The PD-1 and PD-L1 signalling axis is considered a promising target for the development of novel immunotherapy treatments in several cancers. PD-1 (or programmed cell death protein 1) is a surface receptor expressed in T-cells and pro-B-cells.[17] PD-L1 is its ligand, expressed by normal tissue cells, which binds to PD-1^{+} T-cells, eventually leading them to apoptosis. This interaction constitutes an immune checkpoint that is in place to prevent autoimmune attacks. In cancer however this signalling axis is hypothesised to aid immune escape, as tumour cells overexpress PD-L1 to drive the PD-1 expressing T-cells to apoptosis or exhaustion.

In SCCHN anti-cancer drugs that act as immune checkpoint inhibitors against PD-1 or PD-L1 are being evaluated in clinical trials with promising results. Ran et al.[12] published a comprehensive review of immune checkpoint inhibitors for SCCHN in 2017, comparing their overall response rates and mechanisms of action. PD-L1 positivity (at varying levels) is commonly used as a criterion for entry in these studies, however on its own it seems unable to determine a favourable response to treatment. Thus, there emerges a need for predictive biomarkers that could adequately identify an immune suppression landscape and be used for treatment selection.

Several studies support the hypothesis of immune escape due to the PD-1/ PD-L1 checkpoint in SCCHN and have correlated high PD-L1 expression with a poor prognosis. Mattox et al.[18] showed that PD-1^{+} T-cells were often anergic when PD-L1 was positively expressed in the tumour, by observing the reduced co-localisation of Ki67 with these T-cells. Muller et al.[19] analysed PD-L1 expression in regards to overall survival (OS) in two cohorts (293 patients in total) of SCCHN and found a significant correlation between high PD-L1 levels and unfavourable outcome, which outperformed stage and the presence of distant metastasis in terms of prognostic power. This study included mostly OPSCC but did not discriminate patients by HPV status. Skinner et al.[20] recently established a correlation between high PD-

L1 expression in tumour cells of HPV$^-$ SCCHN (mostly oral sites) and early local treatment failure. Through proteomic, transcriptomic and immunohistochemistry analyses, they pointed to an AXL –– PI3 Kinase –– PD-L1 signalling axis related to a radiation resistant phenotype.

However, the correlation between PD-L1 or PD-1 expression and adverse outcome was not consistently observed, raising concerns about their use as biomarkers for treatment selection. Schneider et al.[21] examined PD-1 and PD-L1 expression in 129 SCCHN samples (58 OPSCC). They found that only HPV$^-$ OPSCC samples that were positive for PD-1$^+$ TILs had significantly improved OS and DFS (disease-free survival) while PD-L1 expression was prognostic for oral, but not oropharyngeal squamous cell carcinoma, regardless of HPV status. Oguejiofor et al. in 2017[22] analysed the expression of PD-L1 in a cohort of 124 OPSCC patients. They showed that the density of PD-L1$^+$ cells in the stroma correlated with unfavourable loco-regional control (LRC) and OS only in HPV$^-$ patients. The same study also assessed the population of CD8$^+$PD-1$^+$ TILs, and while a correlation was established between high densities in the stroma and favourable outcome in HPV$^+$ patients, no such effect was present for HPV$^-$ patients.

1.1.1.3.2 Tumour associated macrophages

The CD68 marker is used to identify tumour associated macrophages (TAM). In several solid tumours TAM have been shown to correlate with tumour progression and angiogenesis. Mattox et al.[18] recently reported that 14-32% of PD-L1 expression could be attributed to macrophages and that a co-localisation frequently occurred between CD68$^+$PD-L1$^+$ and CD4$^+$PD-1$^+$ cells, and somewhat less frequently between CD68$^+$PD-L1$^+$ and CD8$^+$PD-1$^+$ cells. Based on these results they postulated that TAM play a role in the PD-1/ PD-L1 immune escape pathway.

Ritta et al.[23] demonstrated a strong positive correlation between CD68 and Ki67 numbers using samples from 22 primary OPSCC tumours, regardless of HPV status. High Ki67 expression was an adverse prognostic factor in this study. In Russell et al.[24] CD68 was slightly increased in HPV$^+$ versus HPV$^-$, but no correlation with outcome was observed.[24] The absence of correlation between intra-tumoural CD68 and disease free survival was reiterated by Pretscher et al.[25] in a cohort of 33 SCCHN (12 OPSCC) patients, irrespective of HPV status. Oguejiofor et al.[22] concluded that CD68$^+$ and CD68$^+$PD-L1$^+$ expression in the stroma were associated with

adverse OS and LRC but only for HPV$^-$ tumours. They observed higher numbers of CD68$^+$ cells in HPV$^+$ than HPV$^-$ cells, in agreement with Russel et al.[24]

1.1.1.3.3 Cytotoxic T-cells

CD8 is used to identify the population of cytotoxic T-cells. These cells carry surface receptors that bind to specific antigens, like the ones expressed by tumour cells, infected or otherwise damaged cells. Once they bind to the antigens, the cytotoxic T-cells can then destroy these cells.[17] The numbers and proportion of CD8$^+$ cells has been shown to be significantly increased in HPV$^+$ compared to HPV$^-$ SCCHN.[14] Two additional studies[24,26] have confirmed this observation for CD8 cells in the central tumour, the invasive margin and the adjacent stroma.

Populations of CD8$^+$ T-cells are usually positively correlated with outcome in HPV$^+$ OPSCC. [14,27,28] Oguejiofor et al.[26] however demonstrated that this correlation was due to the stromal populations of CD8$^+$ cells, while intra-tumoural CD8$^+$ cells did not affect outcome.

In HPV$^-$ OPSCC the effect of CD8$^+$ infiltrate on survival is unclear. While some studies found that CD8$^+$ TILs were prognostic regardless of HPV status,[27,28] others showed no such effect in HPV$^-$ OPSCC.[14,22,26] The effect may also depend on the site of disease, as Feng et al.[29] recently published a study on oral squamous carcinoma where the numbers of CD8$^+$ TILs in the invasive margin correlated positively with improved OS in HPV$^-$.

1.1.1.3.4 Other microenvironment biomarkers

Other cell subsets of interest that have been previously studied for their contribution in disease development of head and neck cancer include fibroblasts, neutrophils, mast cells, and T regulatory cells (T-regs). A comprehensive review can be found in Peltanova et al.[30]

1.1.1.4 The importance of spatial architecture

There is evidence to support that beyond simple enumeration of the cell densities of TILs, observation of their infiltration patterns could offer meaningful information for biomarker development. For PD-L1 expression a study[31] recently observed two different patterns of infiltration; the lace-like induced pattern was found

in the invasive margin and mostly stained immunocytes, while the constitutive pattern was found in the central tumour, staining strongly and uniformly the neoplastic cells. No outcome endpoint was available for this data set, however it was shown that the induced pattern was most frequently found in HPV^+ OPSCC, while the constitutive pattern was equally encountered in both HPV^+ and HPV^- samples.

Another study[29] in HPV^- oral squamous carcinoma demonstrated that the co-localisation of TIL subsets could be prognostic for OS. Looking at the invasive margin, a high occurrence of $FOXP3^+$ T-cells or $PD\text{-}L1^+$ cells within 30 µm of $CD8^+$ correlated with worse outcome. Close proximity between cell types, can be an indication of interaction frequency between these cell types, indicating the presence of an immune-escape pathway. These results show that there is untapped potential in looking at the architectural pattern of infiltration of TILs and spatial interactions between their various subsets in SCCHN for novel biomarker development.

1.1.2 Follicular lymphoma

1.1.2.1 Epidemiology, aetiology and treatment

Non-Hodgkin's lymphoma is the 6th most common cancer in the UK.[32] Follicular lymphoma (FL) is the second most common Non-Hodgkin's lymphoma subtype, accounting for 5% of all haematological malignancy diagnoses in the UK, with an annual incidence rate of 3.1 per 100 000 (95% CI: 2.8-3.3).[33] Incidence increases with age,[34] with a median age at diagnosis of 64.6 years.[33] Slightly higher incidence have been observed for males and non-Hispanic white people.[34]

FL is a malignant B-cell lymphoma that generally follows an indolent and incurable clinical course. The 5 year relative survival rate in the US ranged from 80-90% for cases diagnosed between 2000-2016,[34] while in the UK, 5 year survival was 86.5% for cases diagnosed between 2004-2012.[35] Most patients with advanced stage disease respond to systemic treatment and achieve durable remissions lasting several years, but the majority will relapse and many eventually die of relapsed or refractory lymphoma. Around 20% experience early treatment failure, which is associated with an inferior overall survival.[36] Identifying those high-risk individuals with good precision at baseline is an important task to assist treatment selection and therefore the development of biomarkers for this purpose is critical.

FL is caused by failure of apoptosis in B-cells known as centrocytes and centroblasts. In the germinal centres of lymphoid tissue, such as lymph nodes or bone marrow, follicles structures will form that contain the FL B-cells. These follicles are characteristic nodular patterns in FL, which can vary in size. **Figure 1** shows an example of FL follicles.

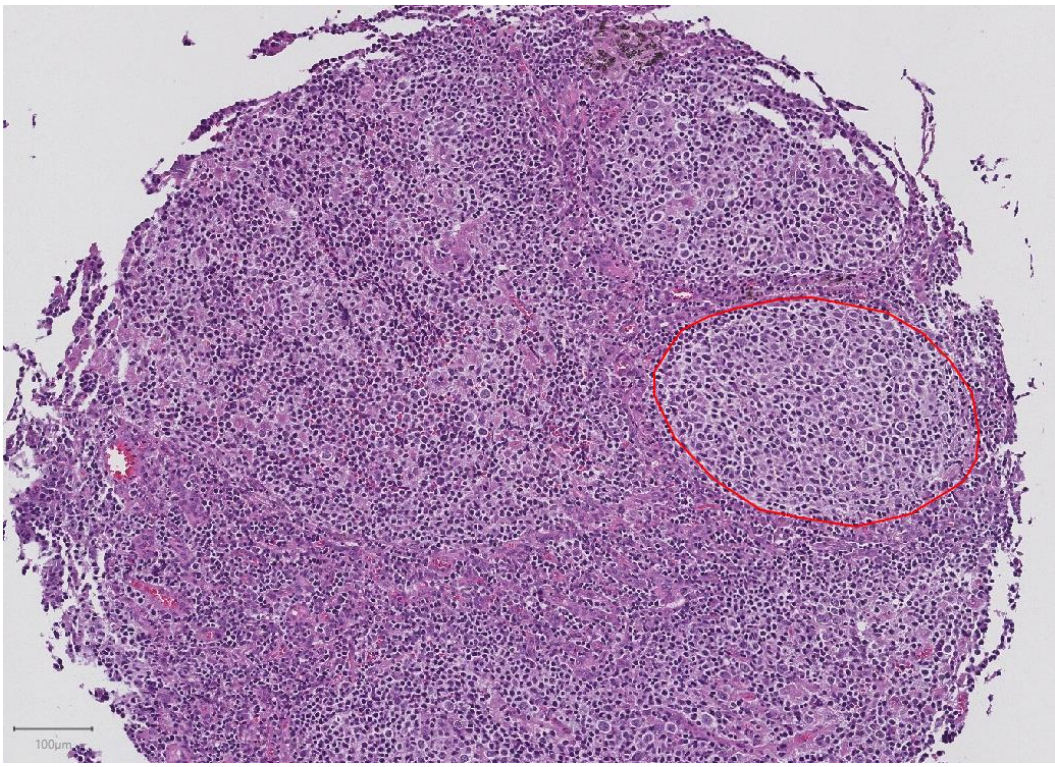

Figure 1 Follicle structures in a follicular lymphoma tissue microarray core shown following haematoxylin and eosin (H&E) staining. A follicle is annotated in red.

The B-cell t(14;18) translocation leading to over-expression of BCL-2 protein is a principle cause of FL development with a prevalence of >85% in FL patients.[34,37] This translocation is thought to be a necessary requirement, but not sufficient for the development of the disease, as it is observed in healthy individuals with a low rate of FL development. Further research is necessary to pinpoint risk factors, however family history of non-Hodgkin's lymphoma, certain genetic factors and some autoimmune disorders (e.g. Sjogren's syndrome, auto-immune haemolytic anaemia) have been previously associated with FL.[34] For most patients with FL, no causative factor can be identified.

Regarding treatment options, advanced stage FL is still considered incurable. However the introduction of monoclonal anti-CD20 antibody rituximab and an increasing number of available treatments has greatly improved survival outcomes in the

last 15 years.[38] Patient management of FL is presently defined based on the Lugano classification.[39] Patients presenting with grade 1, 2 and 3a disease are treated as FL, while patients with 3b disease are treated similarly to Diffuse Large B-cell Lymphoma (DLBCL).[40] The management of patients with grade 1, 2 or 3a FL is summarised below.

1.1.2.1.1 First-line treatment

Patients with stage I or contiguous stage 2 disease may successfully be treated with involved field radiotherapy. For patients with advanced stage low tumour burden, asymptomatic disease, as defined by the Groupe d'Etude des Lymphomes Folliculaires criteria,[41] watchful waiting or single agent rituximab are often recommended. Because of the indolent course of FL, systemic treatment of advanced stage disease is usually delayed until the patient is symptomatic or has high disease burden. When treatment is appropriate, chemoimmunotherapy is the systemic therapy of choice, where rituximab (or the newer anti-CD20 monoclonal antibody obinutuzumab) is administered in combination with chemotherapy (CVP, CHOP or bendamustine).[40] Relative benefits of these alternative first-line treatment options have been studied in the randomised StiL,[42] BRIGHT,[43] GALLIUM,[44] and RELEVANCE[45] trials. Maintenance rituximab may be used after chemoimmunotherapy treatment, to prolong remission.

1.1.2.1.2 Treatment at relapse

Most FL patients with advanced stage disease will eventually relapse, possibly multiple times.[40] At relapse, there is a risk of histologic transformation to high grade lymphoma (DLBCL) which requires a different therapeutic approach.[46] At first relapse, and if no transformation has taken place, the decision to treat is again based on symptoms and disease burden, and may include a further chemo-immunotherapy, a targeted therapy, or stem cell transplant.[40] After second relapse, the optimal treatment options are not well studied. Amongst others, PI3K inhibitors are a possible option.[40]

1.1.2.2 Prognostic factors

In the context of FL, given the many different available treatment options and the heterogeneous nature of the disease, the availability of precise prognostic information is important to help clinicians select the right treatment for the right patient.

Therefore, researchers in recent years have directed their efforts into developing clinically relevant, prognostic indices for FL.

1.1.2.2.1 International prognostic indices

The international prognostic index (IPI) was introduced for aggressive non-Hodgkin's lymphoma in 1993.[47] IPI combined five clinical variables into a single prognostic index: age, tumour stage, serum lactate dehydrogenase (LDH) concentration, performance status, and number of extra-nodal sites. IPI is a four-tier score (low risk, low intermediate, high intermediate and high risk). Its applicability was demonstrated for low grade non-Hodgkin's lymphomas and FL, however, it was not useful for clinical decision making, as it assigned only 11.2% of patients to the high risk group.[48]

The follicular lymphoma international prognostic index (FLIPI) is an adaptation of IPI developed exclusively for FL in 2004. FLIPI was developed using a cohort of 4167 FL patients from 27 centres or groups, diagnosed between 1985-1992, and tested on 919 patients treated before the routine addition of anti-CD20 monoclonal antibodies to standard chemotherapy.[49] FLIPI is calculated by summing five binary clinical risk indicators: age > 60 years, stage III or IV, haemoglobin level < 120 g/L, number of nodal sites > 4 and LDH above normal. FLIPI scores 0-1 are low risk, scores of 2 are intermediate risk, and scores 3-5 are high risk. The distribution of patients to the risk groups was more balanced than with IPI: low risk 37.6%, intermediate 34.8%, high risk 27.6%.[49] In 2006, FLIPI was also validated for use in cohorts treated with R-CHOP.[50]

FLIPI-2 was developed in 2009 in a prospective study and consists of an adaptation of FLIPI, more suitable for cohorts treated with rituximab immunochemotherapy.[51] Similar to FLIPI, it is calculated by summation of five binary risk factors: haemoglobin level < 120 g/L, age > 60 years, bone marrow involvement, nodal site diameter > 6cm, and elevated β2- microglobulin.

Finally, the PRIMA index was developed in 2018 in a cohort treated solely with immunochemotherapy and proposes a simplified version of FLIPI-2, where only two factors are considered: bone marrow involvement and β2- microglobulin higher than 3mg/L.[52]

1.1.2.2.2 M7-FLIPI

In 2015, a variation of the FLIPI that also took into account genetic factors was introduced for patients treated with R-CVP or R-CHOP, namely the m7-FLIPI. This prognostic index incorporated FLIPI, performance status, and the mutational status of the genes: CARD11, EZH2, MEF2B, EP300, ARID1A, FOXO1, and CREBBP.[53] High risk m7-FLIPI patients had 65% 5-year OS, compared to 90% in the low risk. This index has not yet been prospectively validated.

1.1.2.2.3 Post-treatment FDG-PET

For patients treated with rituximab immunochemotherapy, the response to treatment, as defined by observation of the FDG-PET scan post-treatment, has been shown to correlated favourably with overall survival and progression-free survival.[40] Even though response to treatment is an important index for prognosis, it cannot be predicted pre-treatment and therefore cannot be used for treatment selection.

1.1.2.2.4 Total metabolic tumour volume (TMTV)

Instead of observing the FDG-PET post-treatment, two studies have validated prognostic indices based on the baseline FDG-PET.[54,55] The total metabolic tumour volume (TMTV) can be calculated from FDG-PET images using dedicated software, and was shown to be negatively correlated with outcome.

1.1.2.2.5 Progression of disease

The duration of remission after systemic treatment is one of the most robust prognostic indicators in FL. Disease progression within 24 months of initiating treatment (POD24) is considered as a surrogate endpoint for OS and progression-free survival (PFS). A 2015 validation study including a discovery cohort of 2700 FL patients and a validation cohort of 588 patients, treated with first line R-CHOP, demonstrated a negative correlation between POD24 and outcome. The 20% of patients with POD24 had a 5-year survival of 50%, compared with 90% in the POD24 negative group.[56] Unfortunately, POD24 can only be assessed post-treatment and baseline indices, such as FLIPI, have proved unreliable predictors of POD24.[57]

1.1.2.3 The role of the tumour microenvironment in prognosis

Apart from the B-cell t(14;18) translocation, early gene expression studies[58–60] point to a large subset of additional differentially expressed genes related to FL development. Some of these genes regulate communication between various immune subsets (T-cells, B-cells, macrophages etc.). Furthermore, FL B-cells hardly survive on their own when cultured *in vitro*, unless incubated with an emulated tumour microenvironment,[61] indicating that the non-neoplastic immune infiltrate in FL can act as a growth support network for tumour cells, shielding them from apoptosis.[62] Many studies link the heterogeneity of TILs in the tumour microenvironment (TME) of follicular lymphoma to survival, even though there is not always consensus on the observed effects.[63,64] A number of prognostic studies carried out using multiple cohorts with distinct treatment arms have provided further evidence that the prognostic value of TIL subsets depends on the type of treatment, and varies drastically between cohorts treated with or without monoclonal anti-CD20 immunochemotherapy.[64,65]

1.1.2.3.1 CD4$^+$ T-cells

The CD4 molecule is found on the surface of monocytes, macrophages, dendritic cells and T-cells. In the context of FL, CD4 mostly identifies T-cells (co-expressing CD3), characterised as T-helper cells.[64,66] They are called “helper” cells, as one of their main roles is to send signals to other immune cells (such as CD8$^+$ T-cells) to coordinate the immune response. In FL these cells have been shown to upregulate PD-1, CTLA4 and TIGIT.[67] A higher population of CD4$^+$ non-neoplastic T-helper cells has been associated with favourable outcome in some studies[68] and unfavourable outcome in others.[69] In rituximab treated cohorts, the prognostic effect of CD4$^+$ cells has either been favourable[70] or insignificant.[65,71]

1.1.2.3.2 T-regulatory cells (T-regs)

T-regulatory cells (T-regs) are a subset of T-helper cells, expressing CD4, FOXP3 and CD25. FOXP3 is considered the principal lineage marker of these cells and is often used to identify them. The role of T-regs is to regulate immune responses and suppress any auto-immune attacks, by de-activation of T-cells. However, in the context of lymphoma, they can be related to unfavourable outcomes, by suppressing the host anti-tumour response.[62] In FL the effect of T-regs is even more complicated, as T-regs found in the germinal centre have been shown to suppress B-cell

expression through multiple pathways,[72,73] potentially inducing an anti-tumour effect. The prognostic significance of T-regs is therefore unclear. Higher populations of intra-follicular $CD4^+FOXP3^+$ T-regs have been found to correlate with favourable outcomes in some studies.[74,75] In other studies including rituximab treated patients, no effect was observed.[65,71,76]

The location of T-regs has also been tested as a potential prognostic biomarker. Farinha et al.[77] in 2010 identified two patterns of T-reg localisation: the diffuse, where T-regs could be found anywhere with equal probability, and the follicular, where T-regs formed dense hotspot clusters, either inside or outside the follicles. No rituximab treated patients were included in this study, which found that a diffuse distribution of T-regs was an independent favourable predictor of OS. The prognostic significance of T-reg localisation patterns has since been tested in two rituximab treated cohorts, where no correlation with outcome was observed.[71,76]

1.1.2.3.3 $CD8^+$ T-cells

Cells expressing CD8 are cytotoxic T-cells. Increased populations of $CD8^+$ cells correlated with favourable outcome in FL[78–80] and are therefore thought to play a role in control of the disease. The link between cytotoxic T-cells and outcome seems however to depend on the type of treatment received, and recent studies on rituximab treated cohorts did not observe a statistically significant correlation.[65,71]

1.1.2.3.4 PD-1^+ expressing cells

Cells expressing PD-1 in the microenvironment of FL were found to be mostly $CD4^+$ T follicular helper cells (T_{FH}).[81] Higher populations of intra-follicular PD-1^+ cells were associated with decreased survival in Richendollar et al.[82], but increased survival in other studies[83,84]. In 2015, Yang et al.[81] used a small rituximab-treated cohort of 32 patients and found that only $CD4^+PD\text{-}1^+_{low}$ or $CD8^+PD\text{-}1^+_{low}$ cells in interfollicular areas correlated with an unfavourable outcome, while $CD4^+PD\text{-}1^+_{high}$ cells inside the follicles were not significant. These findings, combined with functional analysis revealed two distinct PD-1 expressing cell populations: a) the PD-1^+_{high} T_{FH} cells found inside the follicles that actively supported FL B-cell growth, and b) the PD-1^+_{low} cells found outside the follicles, which represent exhausted T-cells and were unfavourably correlated with outcome. The PD-1^+ T_{FH} found within the follicles are known to express CD4 less strongly (30.7% lower CD4 intensity) compared with other $CD4^+$ cells in the interfollicular areas.[85] These intra-follicular

PD-1$^+$ T_{FH} were shown to be in close proximity or direct contact to actively proliferating (expressing CD20 and Ki67) FL B-cells, which could support the hypothesis that these cells promote tumour progression.[85]

1.1.2.3.5 CD68$^+$ macrophages

In FL microenvironment studies, the CD68 marker is used to identify tumour associated macrophages, although it can also be expressed by monocytes.[71,86–90] Tumour associated macrophages have been correlated with poor prognosis, but only in cohorts that underwent chemotherapy without rituximab.[89,90] However, in rituximab treated cohorts macrophages correlated with good outcome,[86] which can be attributed to one of the many mechanisms of action of the anti-CD20 antibody. Rituximab's "immune-mobilising" effect is known to alert macrophages to the presence of malignant B-cells, and to induce antibody-dependent phagocytosis.[91,92] Therefore, the presence of macrophages is potentially beneficial to patients treated with rituximab.

1.1.2.3.6 Follicular dendritic cells (FDC)

Follicular dendritic cells in FL are mostly arranged in distinct spherical mesh network patterns.[93] They are known to express CD21, CD23, S-100 and CD35, among other markers.[93] In healthy lymphoid tissue the functionality of FDC includes (i) long-lasting chronic retention and presentation of antigens, (ii) arranging the compartmentalisation of the lymphoid tissue, (iii) prevention of auto-immune attacks and (iv) facilitating apoptosis of B-cells that are auto-reactive or non-specific to the antigens presented by the FDC.[94] In the context of FL, in a rituximab treated cohort Blaker et al.[71] found that the presence of FDC networks at diagnosis was unfavourably correlated with outcome. They hypothesised that persistent antigen presentation by FDC, may prevent B-cell apoptosis and induce proliferation, a finding supported by the pro-tumoural effect of co-culturing FL B-cells with CD14$^+$ FDC.[95] Other studies have not confirmed this; Glass et al.[71] found the presence of intact CD21 meshwork pattern was favourable, while Shiozawa et al.[96] observed disappearance of CD21$^+$ FDC preceding transformation of FL to high grade lymphoma.

1.1.2.3.7 Other microenvironment biomarkers

Many other cell subsets have been studied for their prognostic potential in FL, notably neutrophils and mast cells,[86,88] MUM-1 expressing B-cells,[76] GrzB$^+$ cells,[80]

cells positive for TIA1, CD57 or PD-L1,[71] and cells expressing ICOS.[65] Comprehensive reviews on the tumour microenvironment of FL can be found in Carbone et al.[37] and De Jong et al.[64]

Given the prominent role of rituximab in current treatment practice, **Table 1** and **Table 2** provide an overview of studies that have assessed the prognostic effect of different microenvironment biomarkers in rituximab treated FL cohorts. The same studies are included in both tables; **Table 1** gives an overview of the observed effect of the biomarkers studied, while **Table 2** provides context on the study design and clinical endpoints used. Due to heterogeneity in study design, it is challenging to directly compare the findings of these studies. Better-quality evidence is expected from randomised clinical trial studies which are conducted in well controlled settings. However the two randomised clinical trials[65,70] did not use continuous variables in OS survival analyses, which could potentially have reduced their power of observation.

Table 1 Studies assessing tumour microenvironment biomarkers in rituximab treated FL cohorts: observed effects

Cell subset	Taskinen (2007, 2008)[86,88]	Canioni (2008)[89]	Sweetenham (2009)[76]	Laurent (2011)[80]	Wahlin (2011)[70]	Yang (2015)[81]	Blaker (2016)[71]	Xerri (2017)[65]
$CD68^+$	Favourable	No correlation					No correlation	
$CD3^+$	No correlation				Favourable		No correlation	Favourable
Neutrophils	No correlation							
Mast cells	Unfavourable							
$CD68^+$ intra-follicular		No correlation	No correlation				No correlation	
$CD68^+$ extra-follicular		No correlation	No correlation					
$FOXP3^+$ T-regs			No correlation				No correlation	No correlation
$FOXP3^+$ follicular pattern			No correlation				No correlation	
$MUM\text{-}1^+$ tumour cells			Unfavourable					
$CD8^+$				Favourable	Varies with treatment		No correlation	No correlation
$GrzB^+$				Favourable				
$CD4^+$					Favourable		No correlation	No correlation
$CD4^+PD\text{-}1^+_{high}$ intra-follicular						No correlation		
$CD4^+PD\text{-}1^+_{low}$ extra-follicular						Unfavourable		
$CD8^+PD\text{-}1^+_{low}$ extra-follicular						Unfavourable		
Presence of $CD21^+$ cells							Unfavourable	
$TIA1^+$							No correlation	
$CD57^+$							No correlation	
$PD\text{-}1^+$							No correlation	Favourable
$PD\text{-}L1^+$							No correlation	
$ICOS^+$								No correlation

Unfavourable	
Favourable	
Not studied	
No correlation	
Varies with treatment	

Table 2 Studies assessing tumour microenvironment biomarkers in rituximab treated FL cohorts: study design

Study	Patient selection	Patients	End-point	Treat-ment	Type of analysis	Biopsy time point
Taskinen (2007)[86]	continuous	96	PFS, OS	R-CHOP	continuous and dichotomised variables	pre-treat-ment, at di-agnosis or relapse
Taskinen (2008)[88]	continuous	98	PFS, OS	R-CHOP	continuous varia-bles	pre-treat-ment, at di-agnosis or relapse
Canioni (2008)[89]	high tu-mour bur-den	102	PFS, OS	R-CHVP-I	dichotomised variables	at diagnosis
Sweeten-ham (2009)[76]	advanced FL	77	OS	R-CHOP	dichotomised variables	previously untreated
Laurent (2011)[80]	continuous	80	PFS	R-chemo	dichotomised and categorical varia-bles	at diagnosis
Wahlin (2011)[70]	randomised trial	250	TR, OS	R-single R-IFN-α2a	continuous for TR, categorical for OS	pre-treat-ment, at di-agnosis or 1st relapse
Yang (2015)[81]	continuous	32	OS	R-chemo	dichotomised variables	at diagnosis
Blaker (2016)[71]	trans-formed vs. not trans-formed case con-trol	52 cases 40 con-trols	TTT, OS, PFS	R-chemo	continuous varia-bles	at diagnosis, at transfor-mation
Xerri (2017)[65]	randomised trial	>287	PFS	R-chemo	dichotomised and categorical varia-bles	at diagnosis

FL: follicular lymphoma, PFS: progression-free survival, OS: overall survival, TR: treatment response, TTT: time to transformation, R-chemo: rituximab immunochemotherapy, R-single: single agent rituximab, R-CHOP: rituximab, cyclophosphamide, doxorubicin (hydroxydauno-mycin), vincristine (oncovin), prednisolone (a steroid), R-CHVP-I: rituximab, CHVP (cyclo-phosphamide, adriamycin, etoposide, and prednisolone) plus interferon-α2a regimen, IFN-α2a: interferon α2a.

1.1.2.4 The importance of spatial architecture

Looking at studies searching for clinically useful tumour microenvironment bi-omarkers in follicular lymphoma, it becomes apparent that most use cell population numbers or density to quantify the heterogeneity, while only a handful[77,97] mention the potential of observing the spatial pattern as a biomarker. This approach is lim-ited because it does not account for the spatial distribution of cell populations. In FL, different spatial tissue compartments are known to exist (inter-follicular, intra-follicular, peri-follicular), with distinct phenotypic profiles. In addition, a complex

interplay is evident between different subsets of TILs in the tumour microenvironment. A logical hypothesis follows that a holistic and spatially aware assessment of the FL microenvironment would be more informative than examining isolated components.

Automated quantification of pattern is a promising field for biomarker discovery and various methods have been proposed.[98,99] In follicular lymphoma the "Hypothesized Interactions Distribution" (HID) method[97,100] has been applied to quantify the heterogeneity of cell type interactions in the tumour microenvironment, using multiplexed immunohistochemistry and machine learning. HID was used for overall survival prediction from right censored data in a follicular lymphoma data set stained for $CD3^+$, $CD69^+$ and $FOXP3^+$, which allowed them to observe interactions between $CD3^+FOXP3^+$ T-regs and other $CD3^+CD69^+$ activated T-cells.

1.1.3 Summary of tumour microenvironment populations

Sections 1.1.1 and 1.1.2 have provided clinical background and context on the importance of various tumour microenvironment cell populations in disease progression of OPSCC and FL. This section summarises several key tumour microenvironment populations that were discussed, lists the proteins they characteristically express, and juxtaposes their role in healthy tissue, OPSCC tumours and FL tumours (**Table 3**).

The microenvironment in the two cancers is quite different, and the same cell populations may participate in different pathways. As an example, the prominent role of the PD-1/PD-L1 immune escape pathway in OPSCC has not yet been established in FL, where only 5% of FL cells were found to express PD-L1.[101]

In both OPSCC and FL several tumour microenvironment populations have been shown to correlate with prognosis. Furthermore, spatial tissue compartments with distinct phenotypic profiles are shown to exist in both cancers, with varying levels of antigen expression. Therefore, spatially aware assessment of the tumour microenvironment in both cancers may hold additional unexplored prognostic information.

Table 3 A summary of tumour microenvironment populations discussed in this thesis for OPSCC and FL

Marker	*Phenotype*	*Role in healthy tissue*	*Role in OPSCC*	*Role in FL*
CD68^{+}	Monocytes, primarily macrophages	Detection and phagocytosis of bacteria and other harmful organisms/ cells, antigen presentation	Pro-tumour effect, angiogenesis	Depends on treatment, anti-tumour effect in rituximab treated
CD8^{+}	Cytotoxic T-cells	Recognise antigens presented by damaged cells and kill them	Anti-tumour effect, may depend on location/ HPV status	Depends on treatment, probable anti-tumour effect
CD4^{+} *CD68*$^{-}$	T-helper cells	Coordinate immune response	Probable anti-tumour effect[30]	Depends on treatment, probable anti-tumour effect
CD4^{+}*FOXP3*$^{+}$	T-regulatory cells (T-regs)	Immune-regulatory effect	Increased compared to healthy tissue, unclear prognostic significance[30]	Probable anti-tumour effect
CD21^{+} *CD23*$^{+}$ *CD14*$^{+}$	Primarily FDC (meshwork pattern)	FDC provide chronic antigen presentation, B-cell selection, immune-regulation, lymphoid tissue compartmentalisation	Absent	Probable pro-tumour effect, although they disappear preceding transformation
PD-1^{+}	Primarily T-cells	Immune checkpoint	Immune escape mechanism	Exhausted T-cells (anti-tumour effect) or T-follicular helper cells (pro-tumour)
PD-L1^{+}	Some tumour cells, macrophages, other	Immune checkpoint	Immune escape mechanism	Unclear, expressed by 5% FL cells, presence of immune escape still investigated[101]

1.2 Technical background

Multiplex immunofluorescence enables concurrent observation of multiple cell populations in the tumour microenvironment, such as the ones identified for their prognostic relevance in section 1.1. The following sections provide technical background on generating such multiplex images and analysing them to study the spatial architecture of the tumour microenvironment.

1.2.1 Multiplex immunofluorescence

1.2.1.1 General principles

Immunohistochemistry (IHC) and immunofluorescence (IF) assays, first implemented during the 1940s, are routinely used in clinical practice to detect antigens in tissue biopsies.[102] The underlying principle relies on antibodies, which are protein structures that can bind specifically to particular antigens of interest, present on the tissue. The antibodies are incubated on the tissue, often conjugated with dyes (brightfield IHC) or fluorophores (IF), and subsequently the binding sites can be visualised using a microscope.[102]

IHC and IF are applied on tissue sections. These sections of 3-5 μm are cut using a microtome, from tissue or cell blocks that have been either frozen or formalin-fixed and paraffin embedded (FFPE). This fixation aims to prevent loss of structure.[103] When using FFPE sections, de-paraffinisation through heating is required to remove the wax before further analysis. In some applications, to reduce the cost and increase throughput, sections from tissue microarrays (TMA) can also be used in IHC; TMA blocks are prepared by selecting and delineating relevant areas from multiple individual FFPE tissue blocks, extracting cores 0.6-2.0 mm in diameter and embedding them all on a new tissue block, in a grid formation (**Figure 2**).

IHC and IF analysis generally consist of two parts: the tissue pre-processing and detection phase. During tissue pre-processing, antigen retrieval is carried out, to restore the 3D structure of relevant antigens on the tissue that might have been compromised during fixations.[104] Antigen retrieval is done by means of enzymes (protease induced epitope retrieval) or heating (heat induced epitope retrieval).[104,105] Heating is carried out in high temperatures for a short period of time, in a microwave oven, steamer, or pressure cooker, while the sample is immersed in a buffer.[106]

During the detection phase, labels are conjugated to the antibodies. Labels are enzymes, or in the case of IF, fluorescent compounds. Afterwards, dyes (also called "tags") such as chromogens or fluorophores are added. These bind to the labels and can be observed using a microscope. Fluorophores are fluorescent chemical compounds that when excited at a wavelength range, emit light at a specific higher wavelength. The emission peaks are observed using excitation/ emission filter pairs in the fluorescent microscope, specific to each fluorophore.[107] When interpreting IF signals, the autofluorescence that is inherently emitted from the tissue needs to be taken into account. The detection mode can be direct or indirect. Direct detection is a one step process, where the label is attached directly to a primary antibody, and both are applied together on the sample. Indirect detection consists of two steps with a primary antibody applied first followed by a secondary label-conjugated antibody that binds specifically to the primary. Indirect labelling allows for several secondary antibody molecules to become attached to each primary antibody, amplifying the signal. Direct detection is faster, while indirect improves sensitivity. In both approaches, the amount of signal received is proportionate to the quantity of antigen in the sample.[108]

IF and IHC are complementary technologies,[109] however IF is better suited for quantitative analysis. The IF signal is linearly proportionate to the level of protein expression on the tissue, while the relationship for brightfield IHC is non-linear.[109]

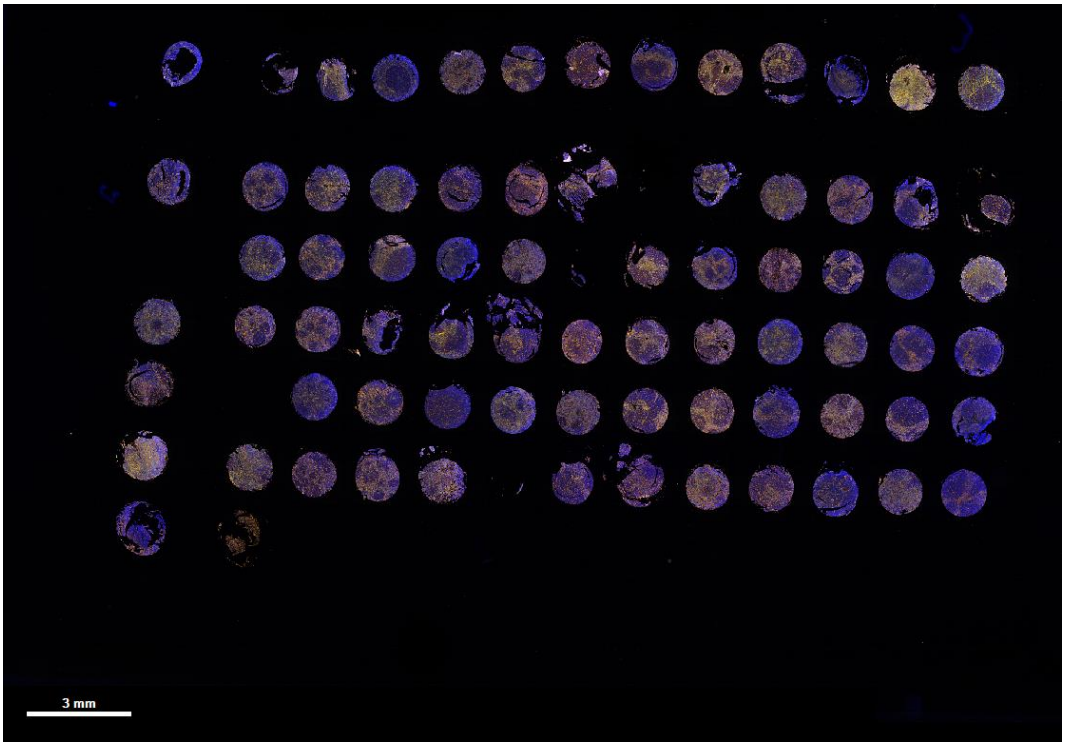

Figure 2 Human follicular lymphoma lymph node tissue microarray, stained with a multiplex immunofluorescence protocol[110] and scanned at 10x with the Vectra 3.5 microscope (only DAPI filter).

1.2.1.2 Multiplex tissue analysis *in situ*

Traditionally, IHC and IF can visualise only one antigen per tissue section. Serial tissue sections need to be cut from a tissue block to observe multiple antigens. Thus, two drawbacks are presented; tissue material is depleted faster, and after a few sections, the observed cells will not correspond exactly to the same tissue area, as sections are taken deeper into the tissue block.[107] Staining of sequential sections does not permit multiple protein expression profiling of each single cell. Multiplex tissue analysis overcomes this limitation and conserves tissue by visualising multiple antigens simultaneously on the same section.

There are several advantages to multiplexing. First, accurate identification of cell phenotypes often requires expression profiling of multiple proteins to fully comprehend their functional role. Second, sometimes multiplex stains can be used as landmarks to identify tissue areas relevant for further analysis. For example, a cytokeratin antibody could be used to identify tumour areas where the oestrogen receptor (ER) antibody should be quantified, to obtain a breast cancer subtype classi-

fication before selecting treatment. Third, clinical diagnosis and prognosis of complex diseases, such as lymphomas, require identifying multiple cell populations, each characterised by a unique protein expression profile. Last, expression levels of proteins relative to each other sometimes need to be assessed, rather than the absolute expression level of a single protein.

Flow cytometry[111] and mass cytometry[112] can achieve high levels of multiplexing with single-cell resolution. However, the technologies are typically destructive and do not preserve the spatial context of the tissue, making them unsuitable for live cell imaging. The following paragraphs describe highly multiplexed methods that preserve the spatial information and, therefore, can provide insight into protein localisation and tissue architecture. The list is not exhaustive, but rather a brief overview of the different approaches available.

1.2.1.2.1 Use of multiple tags

Multiple protein visualisation becomes possible in IHC and IF, by using a different tag for each antigen target. However, staining protocols need to account for cross reactivity that can occur between primary/secondary antibodies and labels. In brightfield IHC, even though a variety of chromogens is available (e.g. the brown 3,3'-diaminiobenzidine [DAB] or the purple Vector VIP), the number of discernible colours in the visible spectrum is limited.[107] Fluorescence microscopy permits higher multiplexing.[107] Microscope filters can be appropriately selected to isolate the signal of each fluorophore as a separate imaging channel, for up to 4-5 fluorophores.[113] The success of this approach depends on fluorophore properties; to obtain strong non-overlapping signals, the fluorophore excitation and emission spectra have to be as far away from each other as possible. This need for separation becomes impractical, as the number of fluorophores increases. Strategies to mitigate this problem include: (1) using emission filters to narrow the spectral range of detection for each fluorophore, (2) using a sub-optimal wavelength for exciting some fluorophores and (3) sequential excitation and detection of the fluorophores.[113] Another important parameter is to have good photo-stability of a fluorophore, i.e., resistance to gradual photo-bleaching due to exposure to fluorescent light.[107,113]

1.2.1.2.2 Sequential staining, scanning and bleaching

The sequential staining approach consists of cycles of staining with a fluorescent tag, imaging the sections using a slide scanner, and subsequent bleaching (inactivation) of the tag. Then the process is repeated for the next tag.

Multi-epitope-ligand cartography (MELC) is a sequential staining technique, first introduced in 2003 by Schubert.[114] MELC is an automatic iterative multiplexing method, where in every cycle two monoclonal antibodies are applied. In 2006, two publications[115,116] used MELC to multiplex up to 18 antibodies and introduced the concept of combinatorial molecular phenotypes (CMPs). CMPs are binary vectors assigned to each pixel, whose length equals the number of multiplexed proteins. The value of each position in the CMP vector is either 1 or 0, indicating the detection of a protein, or its absence, respectively. In this way, representative images were constructed, where a colour was assigned to each of the most prominent CMPs in the sample (toponome maps). The toponome imaging protocol was later detailed by Friedenberger et al.[117] Several applications of this method have been demonstrated,[118–123] including the achievement of multiplexing the detection of over 100 proteins in super-resolution.[124] Moreover, alternative interactive approaches for the visualisation of the CMPs were proposed, using graph theory,[125] and other methodologies.[126,127] Higher multiplicity is achieved for surface components, which can be easily bleached compared with intracellular components.[113]

In 2013, Gerdes et al.[128] demonstrated the use of another sequential staining protocol to achieve multiplexing of 61 markers on a single tissue section. An inactivation solution using alkaline oxidation chemistry was developed and patented. The solution reduced Cy3 and Cy5 fluorescent dye signal down to 2% of the original value, irreversibly, without altering the absorbency of DAPI (4′, 6-diamidino-2-phenylindole). DAPI is a fluorescent stain that binds to DNA in all cells, therefore illuminating the nucleus shape and position. The DAPI stained nuclei could subsequently be used as reference to register, i.e. align virtually, the images obtained from sequential scans. They also patented a protocol for de-waxing, rehydrating samples and implementing a two-step antigen retrieval process. This protocol was tested for loss of antigen epitopes and tissue integrity for a maximum of 100 reaction cycles.

1.2.1.2.3 Multispectral imaging and spectral unmixing

The cyclic imaging systems permit highly multiplexed analysis but scanning the section after each step can be time-consuming, unsuitable for live samples, and requires a specialised lab set-up. Multispectral imaging combined with spectral unmixing can overcome this issue, when fewer than ten antibodies are needed. A sequential staining approach is applied; each step applies one antibody-fluorophore combination and then strips away the antibodies, leaving only the fluorophores on the tissue. The next step repeats the process for a different antibody-fluorophore combination. When staining is completed all the fluorophores remain on the tissue. In this setting the sample is scanned only once after the staining is completed at multiple emission wavelengths, producing a multispectral λ stack.[129]

Confocal laser scanning microscopy, which operates with multiple lasers, allows quantitative multispectral analysis.[129–131] Otherwise, a rotating filter wheel or electronically tunable filters can be placed in front of a charge-coupled device camera, to capture images at fixed λ intervals.[132]

After acquiring the λ stack, linear spectral unmixing separates the contributions of individual fluorophores.[133] Since the signals of the fluorophores combine linearly to produce the observed mixed emission spectrum (**Figure 3**), linear regression is applied to separate them. The acquired image is considered a two-dimensional array of pixel elements, each defined by a height i and width j coordinate. Linear regression minimises the residuals at each pixel to obtain the contribution of each fluorophore:

$$arg\min_{\boldsymbol{w}_{i,j}} \left\| \boldsymbol{s}_{i,j} - (w_{i,j}^{a}\boldsymbol{a} + w_{i,j}^{b}\boldsymbol{b} + \ldots) \right\|_2 \quad (1)$$

, where $\boldsymbol{s}_{i,j}$ is the observed mixed signal vector across all λ intervals in pixel i, j; $\boldsymbol{a}, \boldsymbol{b}, \ldots$ are the base signal vectors of individual fluorophores across all λ; and $\boldsymbol{w}_{i,j} = (w_{i,j}^{a}, w_{i,j}^{b}, \ldots)^T$ the unknown weights indicating the contribution of each fluorophore in pixel i, j.[133] When solving the minimisation problem, a new image is constructed for each fluorophore, displaying the contribution $\boldsymbol{W}$ for that fluorophore on all pixel locations. Each of those images is stored as a separate imaging channel.

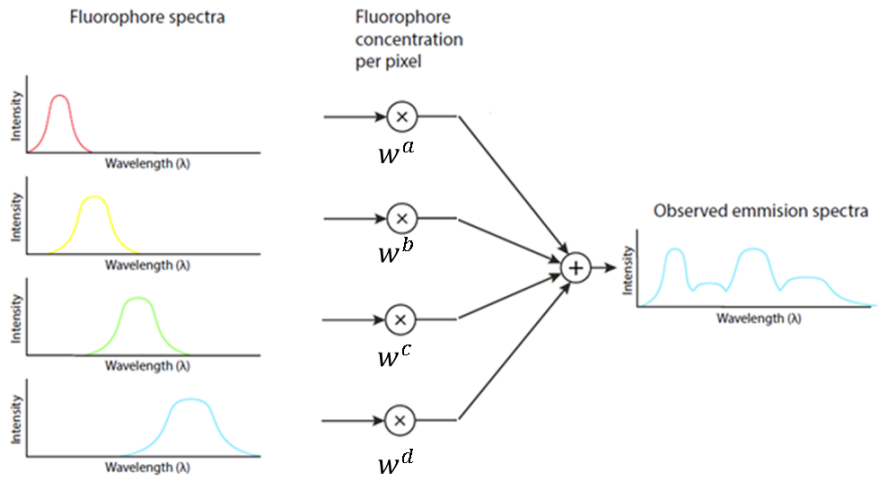

Figure 3 Demonstration of the convolution (mixing) of multiple fluorophores when applied on a tissue during multiplexing linearly to produce the observed emission spectrum.

To perform spectral unmixing, the base spectrum of each individual fluorophore $\boldsymbol{a}, \boldsymbol{b}$, ...must be known. These are acquired from singleplex experiments, where tissue is stained with only one fluorophore at a time and the spectrum is extracted and saved in a spectral library.[133,134] In the IF setting, the autofluorescence spectrum is also acquired and subtracted, using a section where no fluorophore has been applied.[133] Specialised microscope slide scanners, such as the Vectra (Akoya Biosciences, Marlborough, MA, USA) and Vectra Polaris (Akoya Biosciences, Marlborough, MA, USA) combined with software, such as the inForm (Akoya Biosciences, Marlborough, MA, USA) provide a user-friendly interface to build spectral libraries and unmix fluorophore signals.

1.2.1.2.4 Metal-based multiplexing

A competing approach to IHC that can achieve high levels of quantitative multiplexing while also preserving the spatial structure of the tissue is imaging mass cytometry.[135] Cytometry by time of flight (CyTOF) technology conjugates antibodies to heavy metal isotopes instead of conventional chromogens or fluorophores.[112] Accurate discrimination between these isotopes is achieved by measuring their mass-to-charge ratio in a time-of-flight mass spectrometer,[112] thus enabling multiplexing of up to 45 targets.[136,137] While originally mass cytometry did not allow quantification of protein expression *in situ*, imaging mass cytometry overcame this issue by use of a high resolution laser.[138] The laser performs successive ablation of

small tissue areas (1 μm^2 at a time), which are then analysed with CyTOF and combined to create a reconstructed image of the tissue.[135,138] Currently CyTOF is marketed by Fluidigm (South San Francisco, California, USA).

Another metal-based technology is Multiplexed Ion Beam Imaging, marketed by IonPath (Menlo Park, California, USA). The technology combines use of high energy beams and imaging mass cytometry to visualise *in situ* more than 40 targets.[139]

1.2.1.2.5 Oligonucleotide-based multiplexing

Another group of methodologies for multiplexing *in situ* that is gaining popularity quickly relies on oligonucleotide probes. DSP technology by NanoString (Seattle, WA, USA)[140,141] uses oligonucleotide "barcode" tags, attached to antibodies using photocleavable linkers. After antibody incubation, UV light is applied to photocleave and release the tags, which are then transferred onto a plate using microcapillary aspiration and digitally counted. This method does not produce a reconstructed image of protein expression on the tissue; however, the spatial component can be preserved by UV cleaving specific tissue areas at a time. Tissue area selection is guided by basic immunofluorescent staining.[140]

New competing platforms, based on oligonucleotide sequence probes that allow visualisation of all targets *in situ* include the InsituPlex (Ultivue, Cambridge, MA, USA)[142] for approximately 12 targets and the CODEX (Akoya Biosciences, Marlborough, MA, USA)[143] for up to 40 targets.

1.2.1.2.6 Virtual multiplexing

Finally, a few studies[144,145] have adopted a computational approach to multiplexing, by staining serial sections, scanning them and then aligning them virtually based on a counterstain marker. Virtual multiplexing can be combined with any other multiplex technology to achieve visualisation of even higher numbers of targets, using only 2-3 serial tissue sections.[146]

1.2.1.2.7 Choosing the appropriate multiplexing method

The field of quantitative multiplexing *in situ* is rapidly expanding with new technologies. Flow and mass cytometry are well validated methods that can achieve high levels of multiplexing; however, they are inadequate when spatial context is required. If concurrent observation of more than 40 targets is needed to answer

open-ended research questions and generate new hypotheses, cyclic immunofluorescence, metal-based multiplexing and oligonucleotide-based multiplexing provide this capability. However, multiplex Vectra or InsituPlex panels, visualising approximately 10 targets, are sufficient and more cost effective, when investigating pre-specified hypotheses in larger cohorts.[139]

Another important factor is how well validated and commercially available the methods are. Cyclic immunofluorescence is well validated, however not currently marketed commercially. Imaging mass cytometry, DPS and the Vectra systems are well supported by commercial companies, and well validated for research applications. However, CODEX and InsituPlex technologies, while well supported commercially, are relatively new and unexplored.[139] The level of optimisation required by the end-user also varies, and panels that require extensive tuning (e.g. the Vectra) are usually the most inexpensive. Finally, the Vectra and oligonucleotide-based technologies are better suited when sensitive or live samples are analysed.

As new competing technologies introduce rapid improvements in spatial resolution, scanning speed and number of targets, the optimal choice of multiplexing method will change dynamically. When choosing a method, the type and size of sample, cost, as well as the specific nature of the task and required level of multiplexing should be considered.

In this work, the well-established Vectra system was selected as a means to test pre-specified prognostic biomarker hypotheses, while considering the spatial context of the tumour microenvironment.

1.2.1.3 The Vectra multiplex protocol

Multiplexing with the Vectra system relies on sequential staining, using the OPAL fluorophores (Akoya Biosciences, Marlborough, MA, USA) and tyramide signal amplification (TSA). After the initial de-paraffinisation and antigen retrieval steps, staining is performed in iterative cycles, one for each antigen target (**Figure 4**).[107] In each cycle the primary antibody is incubated on the tissue, followed by a secondary antibody from the same species (rabbit or mouse) as the primary, conjugated to horseradish peroxidase (HRP). Then, an OPAL fluorophore is incubated, after titration with the TSA diluent. HRP acts as a catalyst, binding the fluorophores on the antigen epitopes with strong covalent bonds.[147] Thereafter, heat antigen retrieval

will remove the primary and secondary antibodies, while the fluorophores remain on the tissue. The next cycle repeats the process with a different primary antibody and fluorophore. This iteration of staining and stripping permits the use of the same secondary antibody for all targets. A blocking agent (e.g. DISCOVERY Inhibitor, Roche, Switzerland) is also applied every few cycles to neutralise endogenous HRP on the tissue.[148] As 2-3 hours may be required per cycle, this staining protocol is best performed using a robotic staining platform, such as the Ventana ULTRA Discovery (Roche, Switzerland) or the BOND Rx (Leica Biosystems, Germany). At last, the DAPI counterstain is applied directly on the section, which is then coverslipped and scanned using the Vectra microscope.

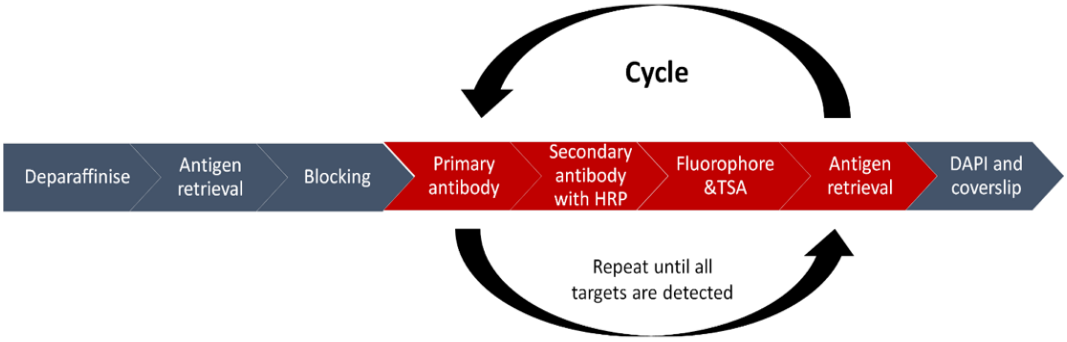

Figure 4 The Vectra multiplex staining protocol, using tyramide signal amplification (TSA). HRP: horseradish peroxidase and DAPI: 4′, 6-diamidino-2-phenylindole.

Multiplexing with the Vectra microscope uses multispectral scanning and spectral unmixing (see page 42). When optimising a new Vectra protocol, a spectral library is built, using singleplex stained control samples to extract the spectrum of each individual fluorophore. The spectral library can then unmix the signals of different fluorophores in the multiplex stained tissue sections.[149]

1.2.2 Multiplex immunofluorescent image analysis

Interpreting the rich information available from multiplex assays *in situ* proves challenging when done manually. When three markers or more are used, it becomes impractical for the human eye to accurately correlate staining patterns, recognise cell phenotypes, and observe co-localisations. Digital pathology overcomes this limitation by employing automated and quantitative image analysis solutions. Digital pathology tools can support a variety of tasks, such as background noise removal, relevant tissue area identification, cell detection and phenotyping, and lead

to extracting valuable prognostic and diagnostic information from patient biopsies.[150] In the following paragraphs, important steps for the analysis of multiplex images are described, along with frequently used algorithms in each step.

Typically, a pipeline for multiplex single-cell analysis will involve pre-processing steps for noise reduction and normalisation, nuclear and cell segmentation, and finally cell phenotyping (**Figure 5**).

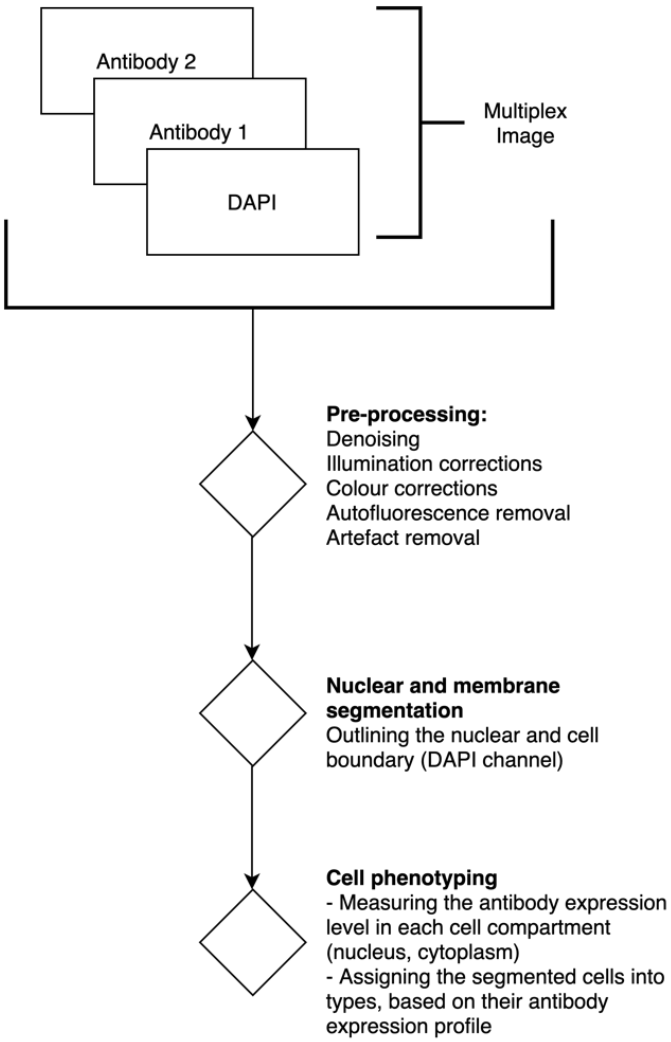

Figure 5 Overview of process needed to analyse a multiplex digital tissue image.

1.2.2.1 Pre-processing

A pre-processing step is usually necessary when analysing digital pathology images. In this step various corrections can be made to counter the effect of variable conditions and problems encountered during image acquisition.

There are many common problems encountered in digital pathology images. When a microscope's illumination field is uneven, illumination artefacts may be present. Noise may also be present originating from the imaging system (camera sensor noise). Stitching artefacts may occur, when the digital scanner combines multiple fields of view together to form the overall image. Furthermore, under or over-saturation is a known pitfall in microscopy[151] when too little or too much light is used during imaging, the image histogram shifts and a spike appears in the maximum (or minimum) permissible intensity value. This effect hinders quantitative assessment of protein expression and is impossible to correct post-acquisition. Additionally, autofluorescence is an issue in all fluorescent image analyses, where the unstained tissue will always emit a baseline signal that should be accounted for.

Other important sources of background signal that may confound digital pathology analyses are rooted in the staining process itself. Colour inconsistencies can be observed, caused by slight variations in staining or imaging conditions. Some staining protocols will produce non-specific binding of the antibody on cells that do not express the antigen of interest. The non-specific binding signal will usually appear in cellular compartments where the protein is not usually expressed and usually will have lower intensity than the true signal. Finally, several significant artefacts can be seen in these images, such as dirt, blurriness, bubbles of air that form between the tissue and the glass coverslip, folded tissue areas, and other small objects.

These problems are common, and therefore the pre-processing step is usually a part of all digital pathology analyses. Pathologists in the clinical setting can learn to recognise and work around these issues, however algorithms need to be adjusted appropriately to avoid reaching incorrect conclusions. A non-exhaustive list of pre-processing algorithms relevant to the multiplex immunofluorescent setting are discussed below.

1.2.2.1.1 Illumination and stitching corrections

To correct uneven illumination and any stitching artefacts, a blank-field correction can be applied,[152] where a blank image is captured and then subtracted from the sample images. Recent retrospective approaches, such as CIDRE[153] and BaSiC,[154] do not require capturing reference blank images and can derive the true uncorrupted signal by observing simultaneously multiple images acquired in the same way.

1.2.2.1.2 Techniques to remove noise caused by the imaging system

Filtering refers to a family of image processing operations, where the value of each pixel is modified, based on the value of its surrounding neighbourhood pixels. Examples include median filtering,[155] where the median neighbourhood value is assigned to each pixel, or Gaussian smoothing.[156] Filtering can sharpen or smooth the image and may be used to remove small artefacts and noise caused by the imaging system. Mathematical morphology can also be used for noise reduction. Specifically, by applying morphological erosion and dilation on the image small objects and sharp peaks are eliminated, or breaks and holes on the image are fused, depending on the order of application of the two operations.[157]

Other de-noising strategies may include masking out parts of the image on the Fourier domain to remove quasi-periodic noise,[158] or wavelet-based de-noising.[159]

1.2.2.1.3 Colour corrections

There are several methods to enhance image details and improve low contrast or achieve consistent colouring between different samples: histogram equalisation[160]; contrast stretching; matching the colour distribution to a histogram reference by linear transformation of each pixel[161]; and using index TMAs built from cell lines with negative, weak and strong expression.[162] However, some colour-related problems, such as over- or under-saturation during imaging cannot be corrected post-acquisition.

1.2.2.1.4 Thresholding

Thresholding is a widely used approach to remove unwanted background signal in digital tissue images. The presence of background could be caused either from the imaging system, or non-specific antibody staining. In thresholding, all the pixels that fall under a specific threshold (empirical or automatically calculated) of intensity in the image histogram are considered as background and ignored. Automated thresholding finds either a global image threshold (e.g. Otsu's algogrithm[163] or Rosin's algorithm for unimodal histograms such as seen in immunofluorescence[164]) or local thresholds (e.g. adaptive thresholding[165]) to account for illumination variations.

Some thresholding techniques were developed specifically for the multiplex setting, taking advantage of the co-localisation relationships between different stains to derive optimal cut-offs to separate signal from background. The Coste's algorithm,[166] selects the maximum threshold for each colour in the multiplex image, so that all pixels below the threshold are not statistically correlated. A similar approach was suggested by Baryenska et al.[167] A recent approach[168] automatically identifies unstained cells in each colour channel, based on *a priori* known co-localisation relationships between antigens, to infer the background distribution and remove it.

1.2.2.2 Cell and sub-cellular segmentation

A mainstay in the analysis of all multiplex images is outlining the boundaries (i.e. segmentation) of the cells and sub-cellular compartments (**Figure 6**). This task is relevant for innumerable life science applications. Cell images vary widely, depending on their origin, imaging and staining conditions, and thus, in the past 60 years, hundreds of algorithms have been introduced.[169–172]

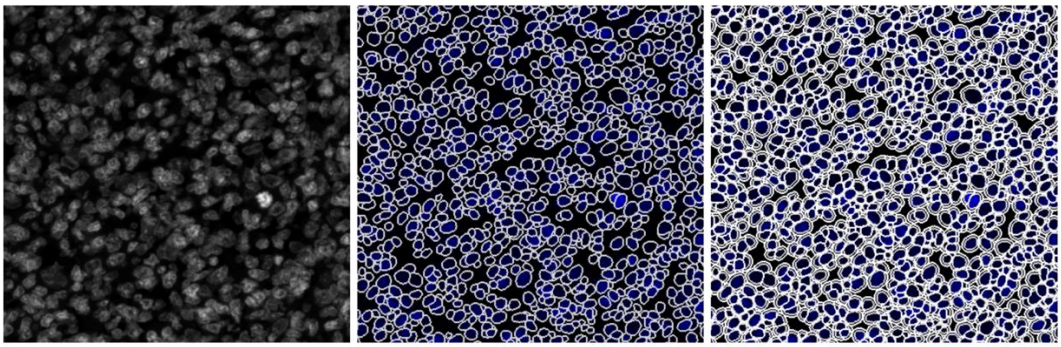

Figure 6 Cell and sub-cellular segmentation task illustration. Left: DAPI channel of a follicular lymphoma tissue, rendered with grey colourmap. Middle: After nuclear segmentation the shapes of the nuclei have been found. Right: Membranes have been simulated around the segmented nuclei, by growing out the nuclear regions by a few microns. Once the process is completed, the nuclear and cytoplasmic compartments of the cells have been defined.

1.2.2.2.1 Nuclear segmentation framework

In multiplex images, cell segmentation starts by outlining the nuclei shapes to pinpoint the cellular locations. Nuclear segmentation is based on the nuclear counterstain, which conveniently stains all cell nuclei. The goal is to define the shape of the nucleus and use it as a reference for the observation of staining patterns of all other markers.

In practice, algorithms developed for nuclear segmentation in brightfield microscopy can often transfer well to the immunofluorescent setting. In fluorescence, the nuclear counterstain is almost always DAPI (**Figure 7**). In brightfield, haematoxylin and eosin counterstains are often paired together ("H&E" stain). Haematoxylin renders the nuclei of the cells dark purple, and eosin renders the surrounding stroma pink. This section will provide a brief overview of nuclear segmentation methodologies for both fluorescence and brightfield microscopy, focusing on 2D, static images.

By conscious choice, nuclei detection algorithms that only locate the cells, without outlining their boundaries will not be discussed, as the overall goal of nuclear and cell segmentation in multiplex image analysis is to explicitly measure the expression of multiple antibodies (up to 100) in each sub-cellular compartment and, based on this information, to classify the cells into distinct subtypes. Therefore, it is necessary to obtain the outline of the nucleus.

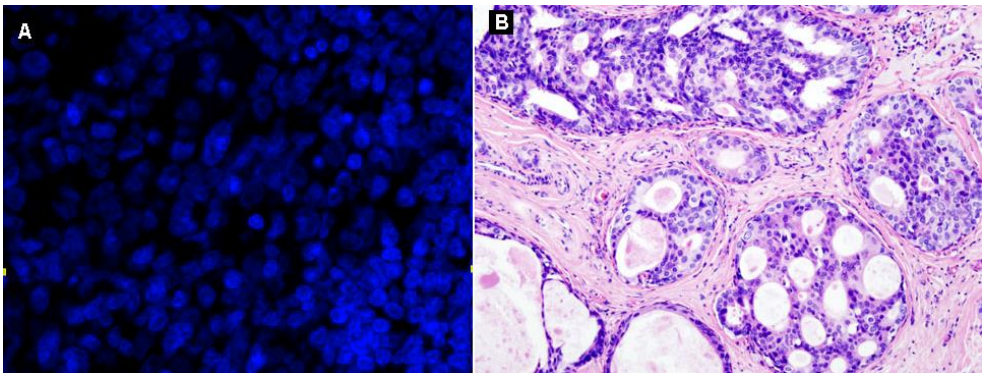

Figure 7 Nuclear counterstain in fluorescence and brightfield immunohistochemistry. A) DAPI stain. B) Haematoxylin and eosin stains.

When aiming to perform accurate single-cell quantitative observations, the algorithm should be able to treat multiple objects of the same class (e.g., nucleus) as separate objects/entities and separate them when they slightly overlap. Touching and overlapping nuclei can be seen in **Figure 7A**. This framework is called *instance segmentation*, and is commonly approached in one of three ways:[171]

1. <u>Binary pixel classification:</u> The algorithm first classifies each pixel as nucleus or background. Then post-processing ensues to separate the different nuclear objects, such as connected component labeling,[173] or watershed.[174]

2. Ternary pixel classification: The algorithm classifies each pixel in one of three classes; nucleus, background or nuclear boundary. This formulation simplifies post-processing, because after the classification, each nuclear object has already been separated from all others by a boundary.[171] This approach requires explicit annotation of the nuclear boundary.

3. Distance from nuclear centroid regression: Nuclear segmentation is formulated as a regression, rather than classification problem. After detecting the centroids of nuclei, a distance map is predicted from each centroid to the nuclear edge.[175,176]

Using these three different instance segmentation frameworks as reference, the next two sections will outline some unsupervised and supervised algorithms that have been previously applied for nuclear segmentation.

1.2.2.2.2 Unsupervised nuclear segmentation

Unsupervised segmentation is performed only in the context of the first framework described above. Early nuclear segmentation algorithms often used a combination of intensity thresholding (e.g. Otsu's, isodata, maximum entropy) to locate the nucleus and background pixels, and some type of post-processing, such as morphology transforms, watershed[174], graph cuts, [177] active contours[178], or Canny edge detection[179]. The performance of such algorithms depends substantially on image and staining quality.[180]

The watershed[174] approach works well for high signal to noise ratios, and requires no nuclear annotations, so it is still widely used today as a practical way to segment nuclei by many digital pathology software applications. Watershed methods are a family of algorithms that intuitively simulate the way water floods a basin to assign pixels to different nuclei. After thresholding, a distance transform is applied to the binary image (nucleus vs. background). Initial seed points are placed at local minima of the distance transform to represent the centre of each basin, which also corresponds to the centre of each nucleus. Then neighbouring pixels are added to each basin iteratively, the same way that water would flood the basin if it entered through the seed point. When two basins meet, a one-pixel wide ridge is formed allowing touching nuclei to be separated. Flooding stops when met with a pixel originally labelled as background.

Instead of thresholding, clustering (k-means[181], fuzzy c-means[182], or expectation-maximisation[183,184] algorithms) may also be used as the initial step, to obtain the binary pixel classification of nucleus vs. background. Clustering groups pixels into groups (in this case nucleus or background) based on their measured similarity. Pixel similarity can be assessed by generating various hand-crafted features (e.g. texture and intensity). Such label-free algorithms do not generalise well to variable staining conditions and usually underperform when segmenting an image with densely packed nuclei.

1.2.2.2.3 Supervised nuclear segmentation

Supervised machine learning can improve performance compared to unsupervised techniques by using human-generated nuclear annotations to train models (e.g. k-nearest neighbours[185] or support vector machines[186]). Supervised nuclear segmentation can potentially be applied in any of the three instance segmentation frameworks outlined above, however it is mostly used in the binary (nucleus vs. background) classification setting. Handcrafted features are selected and extracted as the image representation, including shape,[186] context[186] and colour texture features.[185] For supervised learning nuclear segmentation it is worth mentioning ilastik[187] software, which was designed to generate annotations and build models for instance segmentation of bio-image data. Comprehensive reviews of nuclear segmentation methods up to 2016 can be found in Irshad et al.[188] and Xing et al.[189]

1.2.2.2.4 Nuclear segmentation using deep learning

In the past five years, supervised deep learning methods have gained considerable popularity for nuclear segmentation tasks. They employ convolutional neural networks (CNN) to derive an image representation, bypassing the need of handcrafting custom features. The basic building block of a CNN is a convolution. Convolution is a type of mathematical operation or filtering, where the input image is transformed by iterative application of a kernel (usually a small 2D array, e.g. 3x3) on all of its pixels. When a kernel is applied on a pixel, elementwise multiplication of the kernel values is carried out with all neighbouring pixels and the sum is calculated and assigned as the new value of that central pixel (**Figure 8**). Complex, deep network architectures are constructed by stacking multiple convolutions, with intercalated non-linear activation functions (e.g. ReLU, [190] sigmoid[190]). During model training, the kernel weights are learned through backpropagation.[190]

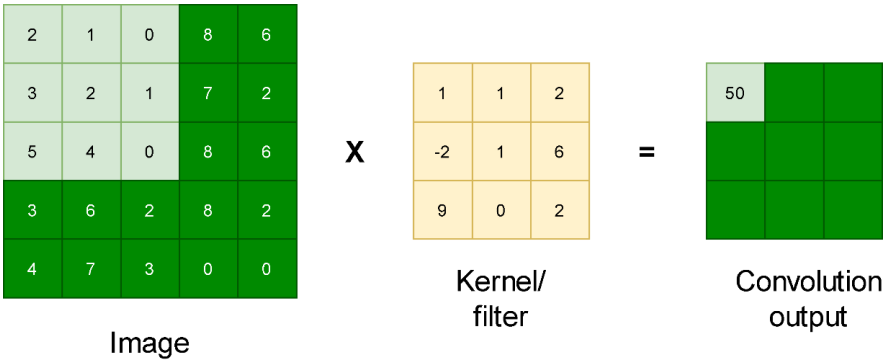

Figure 8 The convolution operation is shown for a 3x3 kernel. The kernel is applied by sliding over all pixels in the input image. In CNN the kernel weights are initialised randomly and then learned and updated iteratively during training.

In the nuclear segmentation Challenge (MoNuSeg 2018)[171] aiming to reveal the optimal algorithms for nuclear segmentation of H&E images from seven different human organs, deep learning was used by almost all participants. The most frequently used network architecture was the U-Net,[191] which follows an encoder-decoder U-shape. Other popular architectures were the VGGNet,[192] Mask R-CNN,[193] FCN,[194] DenseNet[195] and ResNet.[196] Deep learning can be used for the binary (nucleus vs. background), ternary classification task (nucleus vs. background vs. boundary), or in the distance regression setting. Post-processing may follow, using watershed, graph partitioning[197], morphology transforms and size filtering. To prevent false positives, non-maximum suppression[198] and h-minima [199,200] can also be applied.

Excessive data augmentation techniques were used by the top ten algorithms in MoNuSeg 2018.[171] Data augmentation is the practice of artificially increasing the number of training images available by applying transformations, such as rotations, flips, affine transformations, Gaussian noise addition, colour jitter, random HSV shifts and random brightness shifts. Using an augmented dataset reduces overfitting when training the deep learning models. Overfitting is a common problem in deep learning, where the model learns to memorise the exact examples in the training set and classify them correctly but is not able to generalise to new unseen images.[190] Because of the significant colour variations that can be seen in histopathology images, caused by slight changes in staining conditions, data augmentation techniques that randomly shift the colour space (colour jitter, HSV shifts and brightness shifts) are particularly successful ways to artificially produce additional realistic data.

Recent methods formulating nuclear segmentation as a deep regression problem,[175,176] require limited post-processing and have demonstrated excellent performance in brightfield and fluorescence microscopy images. The StarDist[175] algorithm proposes a nuclear detection approach, where a convex polygon is predicted for each nucleus. Convex polygons are adopted as a shape representation to better fit the nucleus shape. This way, nuclear detection and segmentation are carried out simultaneously. The StarDist CNN outputs a probability map of a pixel belonging to a nuclear object and predicts a star-convex polygon from each nuclear pixel, parameterised by a fixed number (typically 32) of radial distances. The object probability is formulated to represent the distance of a nuclear pixel to the nearest background pixel. Multiple polygons may be predicted for each nucleus, and non-maximum suppression (NMS) is applied to select the best one. NMS selects the polygon corresponding to the pixel with the highest object probability (i.e. the nuclear pixel that is the furthest from the background). The base architecture is a light-weight U-net (**Figure 9**).[191] StarDist performed very well in challenging fluorescent microscopy datasets,[175] and a StarDist ImageJ plugin[201] and QuPath script[202] are now available.

While supervised deep learning currently outperforms other nuclear segmentation methods by a significant margin, it requires large amounts of training data to work efficiently. Solutions to this problem can be found in approaches using interactive segmentation and weakly supervised learning. NuClick[203] provides an interactive nuclear segmentation framework, where a multi-scale encoder-decoder CNN generates the fully segmented outline of objects that a user has previously just clicked on once. This way new annotations to train nuclear segmentation models are rapidly generated. Mahmood et al.[204] proposed a conditional generative adversarial network[205] to create synthetic H&E datasets with perfect labelling for nuclear segmentation, initialised by a random map of polygon shapes. Haq et al.[206] suggested domain adversarial training to allow a cell segmentation CNN to generalise to another organ/ image acquisition setup without any labels from the new domain.

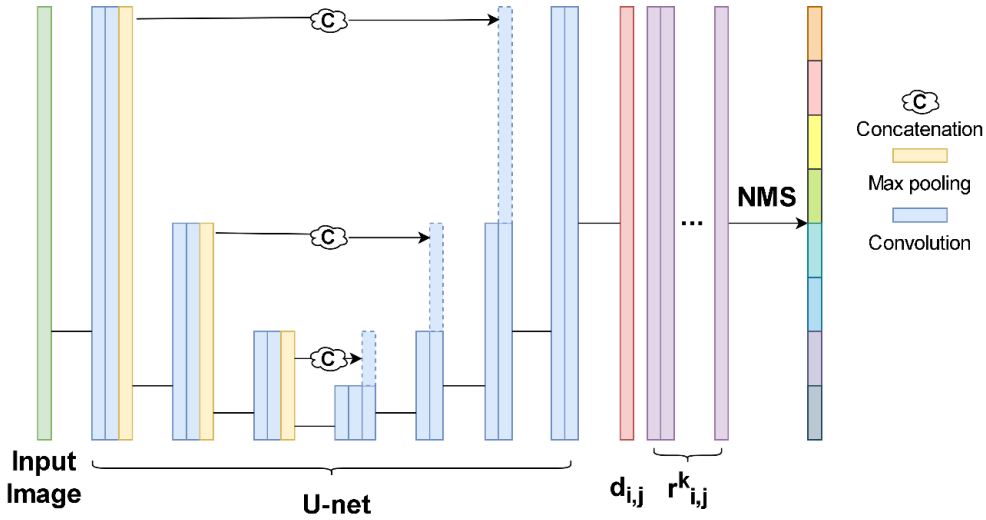

Figure 9 StarDist model architecture for nuclear segmentation. The input image is fed into a U-net type CNN, whose main building block consists of two convolutions, followed by a max pooling layer. The max pooling layer will reduce the dimensionality so that moving further into the network produces a lower dimensionality feature embedding. This part of the U-net is called the "encoder". The second half of the symmetric U-shaped network is the "decoder". In this part, up-sampling and concatenation with previous layers is used to increase the dimensionality of the embedding once again. In the end, the output layer has the dimensions of the original image. StarDist predicts two types of outputs: i) the $\boldsymbol{d_{i,j}}$ map gives the probability of each pixel $\boldsymbol{i,j}$ belonging to a nucleus. ii) the $\boldsymbol{r^k_{i,j}}$ gives the predicted distance to the nearest background pixel from each pixel $\boldsymbol{i,j}$. The distance is regressed for each $\boldsymbol{k = 1,..K}$ radial orientation, where usually $\boldsymbol{K = 32}$. For pixels with low probability of belonging to a nucleus, the regressed distance is set to zero. Finally, NMS is applied to examine overlapping polygons and keep only the ones originating from the most central nuclear pixels. Thus, the nuclear shapes are segmented.

1.2.2.2.5 Performance metrics

Several metrics have been introduced to assess the performance of nuclear segmentation. Simple approaches may compare only the numbers of true and predicted nuclei or use metrics (e.g. Jaccard or Dice coefficients) to assess if pixel labels were assigned correctly. However, these approaches cannot capture adequately an algorithm's ability to both locate each separate nucleus and predict the correct shape. The types of errors that may occur are shown in **Figure 10**.

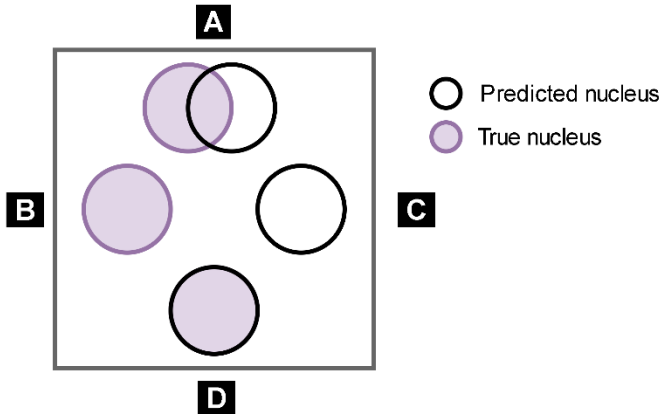

Figure 10 Types of errors in instance segmentation of cell nuclei.

A) The true nucleus is found by the algorithm; however, the shape is mismatched (true positive within a margin of error). B) Nucleus not found (false negative). C) A nucleus shape is predicted when none exist (false positive). D) Perfect prediction of nuclear shape (true positive).

Kumar et al.[171] introduced the aggregated Jaccard index to evaluate performance. This metric is an extension of the Jaccard index for instance segmentation. First, every ground truth (true) nucleus is matched to a predicted nucleus, by selecting the predicted nucleus to maximise the Jaccard index of the two shapes (i.e. the intersection over union). After matching, two metrics are calculated: the sum of all intersections C, and the sum of all unions U for matched nuclei. Unmatched nuclei are also added to the union sum. Finally, the AJI is calculated as the ratio: C/U. Another approach by Uwe et al.[175] adopted the average precision metric, calculated as:

$$AP = \frac{TP}{TP+FN+FP} \tag{2}$$

where TP are the true positive, FN the false negative and FP the false positive nuclei. These definitions follow the notation of **Figure 10**. True positive predictions are the ones that could be matched with a ground truth nucleus. Matching occurs when the intersection over union is higher than a pre-specified threshold $\tau \in [0,1]$. False negatives were the unmatched ground truth nuclei, whereas false positives were the unmatched predicted nuclei. Finally, a more radical approach was suggested by Al-Kofahi et al.[207] where a support vector machine was trained to classify whether a cell object had been correctly segmented.

1.2.2.2.6 Membrane and cell segmentation in multiplex fluorescence

Once the nuclei have been outlined, they are used to guide whole cell and membrane segmentation in multiplex fluorescent images. This task differs from the nuclear segmentation task described above, as a stain to universally visualise the membrane and cell boundaries has not yet been developed. Studies try to reconcile this limitation by combining multiple membrane markers (e.g. β-catenin, NaKATPase) to achieve delineation of all cell membranes.[208] However, this leads to a need for highly multiplexed panels of stains, which are not always practical and increase time and cost of assay.

In practice, the membrane boundaries of cells can be reasonably well approximated, once the nuclear shape is known. Algorithms have been developed to "grow" an estimated membrane shape around each nucleus by sliding band filters,[209] Voronoi tessellations[210–212] or a size-constrained watershed.[213,214] These approaches are practical in most applications but may fail for non-convex or elongated cell shapes. Finally, other methods will resort to only segmenting the membranes of cells that express a particular membrane stain of interest, and ignore all other cells.[215]

1.2.2.3 Cell phenotyping

Despite the inherent variability of antigen expression that renders each cell unique, biologically meaningful, and phenotypically coherent clusters of cells are known to exist, each performing a distinct function. Cell phenotyping is the task of assigning the cells into these phenotypically coherent clusters. Multiplexing could accurately discriminate between different cell phenotypes, through high dimensional protein expression profiling. Following nuclear and membrane segmentation, extensive data becomes available for each single cell (e.g., the fluorescent intensity for each marker in the sub-cellular compartments and cell shape morphology) that can be mined to successfully assign cells to phenotypes. This task is not only relevant for multiplex assays *in situ*, but also for all other methods that quantitatively observe protein expression at the single-cell level, such as flow cytometry and mass cytometry. Therefore, in this section describes cell phenotyping algorithms that could be applied either to the digital pathology, flow or mass cytometry setting alike.

To phenotype a single cell, one can examine whether it expresses a number of different protein markers. Determining the level of expression for a marker in a cell or in the overall tissue area is called *scoring*. Scoring constitutes an essential task in most pathology workflows and several digital pathology algorithms in the past 20

years have been developed to automatically carry out this task. Manual gating of protein signal is considered the gold standard in similar settings, such as flow cytometry.[216] However, when pathologists manually annotate which cells express a marker, this ground truth can be very unreliable.[217] In a typical clinical pathology workflow, pathologists usually provide semi-quantitative estimates for the level of staining expression, which can be subject to multiple biases, intra- and inter-observer variability. Thus, at this time, the most appropriate way to validate automated scoring algorithms remains unclear.

If the pathologists' manual annotations of cell phenotypes were to be considered the ground truth, several ways to approach automated cell phenotyping are presented below:

To determine whether a cell expresses a marker, one may adopt a binary on/off approach where a cell is either positive (expressing the marker) or negative. To this end, gating on the signal intensity is applied, where cut-offs are often determined manually by the annotator. If automated thresholding has been applied to separate true signal from background as described in page 49, then any cells where signal is present may be considered positive. Alternatively, Blom et al.[218] suggest using a cut-off equal to one standard deviation higher than the mean signal of all cells. Use of an index TMA with positive and negative control cell lines to derive a reasonable cut-off is also an option.[162]

When observing a high number of proteins simultaneously, not all of them might be relevant to define a phenotype. Some combinations of protein positivity could be considered equivalent. To negotiate such phenotypic relationships, one may use prior biological knowledge about possible protein co-localisations. A more data driven approach,[115] suggests that cell phenotypic groups or "motifs" can be defined by specifying the lead proteins (L), expressed in all cells of the group, the absent proteins (A), and some wildcard proteins (W), that may or may not be expressed. By this definition, a phenotype that expresses protein A, but not protein B is isomorphic to a phenotype that also expresses protein A, and only sometimes expresses protein B (a wildcard). Schubert et al[115] used this coding to represent graphs of inter-relationships between phenotypes.

Other approaches opt for continuous assessment of the expression level for phenotype identification. Clustering approaches are commonplace in the multiplex flow

cytometry setting. In the FlowCap[216] challenge of 2013, different algorithms competed in phenotyping cell populations assessed with flow cytometry. Manual gating was the gold standard in this challenge. The ADICyt (Adinis Ltd, Slovenia) commercial algorithm, using hierarchical clustering and entropy-based merging, ranked first and agreed the most with manual gating. For highly multiplexed mass cytometry data, Phenograph[219] suggests construction of a cell similarity graph to perform clustering and pinpoint cells likely to belong in the same phenotype. Each cell is represented by a vector of protein signals and similarity between cells is calculated as the Euclidean distance between vectors. Weighted graphs are constructed by connecting each cell with its k-nearest neighbours in that vector space.[219] A community detection method[220] is then applied to partition the graphs and cluster cells with similar phenotypic profiles. For toponome imaging systems data, the DisWOP approach suggests a way to cluster cells into phenotypes based on their protein co-dependence or anti-dependence profiles.[99]

Supervised learning can also be applied for phenotyping, where the phenotyping algorithm is trained by manual annotations of cells belonging to each phenotype. In an example implemented with QuPath, a random forest classifier was trained from manual annotations to identify cell phenotypes in multiplex immunofluorescent samples, using the protein signal intensities as features.[221] In FlowCap[216] supervised learning methods, such as use of radial SVM,[222] had similar performance to clustering.

1.2.2.4 Quantifying spatial patterns

Compared to flow cytometry and mass cytometry, multiplexing assays *in situ* offer the advantage of providing spatial context. Observing spatial gradients of protein expression, co-localisations and proximity relationships add new powerful tools to the arsenal of the pathologist, with potential to improve prognostic and diagnostic accuracy.[223] Spatial architecture in the tumour microenvironment can be assessed in many ways, for example by observing tissue compartments with distinct phenotypic profiles (e.g. tumour, stroma, invasive margin), the presence and location of tumour infiltrating lymphocyte hotspots and the positioning and spatial distribution of tumour associated phenotypes relative to themselves and to each other.

A fundamental hypothesis underpinning studies of spatial heterogeneity in the microenvironment of tumours is that cells found in close proximity are more likely to

interact. Cell-cell interactions could take place either by direct contact junctions or signalling molecule secretion, while each cell is estimated to be able to communicate within a maximum distance of ≈ 250 μm.[224] Image analysis algorithms provide reliable quantification of spatial patterns, with promising applications in cancer which are briefly reviewed below.

Many spatial heterogeneity studies have focused on immune populations infiltrating solid tumours. Huang at al.[225] observed that the tumour associated macrophage phenotype changed as a function of their distance to the nearest tumour cell in gastric cancer. They built an R package to calculate nucleus-nucleus distance in multiplex immunofluorescent images, namely ISAT (https://cran.r-project.org/web/packages/ISAT/index.html).[225] A second study in gastric cancer,[226] looked at the spatial distance between $CD8^+$ T-cells and T-regs, and noted that a distance of 30-110 μm was associated with favourable survival rates. In non-small cell lung cancer, Barua et al.[227] introduced the G-cross function to measure the spatial distribution of T-regs and other immune subsets around tumour cells and linked higher proximity between T-regs and tumour cells to poor survival.[227] A similar study in non-small cell lung cancer,[228] studied the distances between $CD8^+$ T-cells and tumour cells and showed that longer median distances were favourable. In colon cancer, Lazarus et al.[229] observed the mean distance between cytotoxic T-cells, epithelial cells and antigen presenting cells. Cytotoxic T-cells were considered "engaged" if found within 15 μm of either epithelial or antigen presenting cells, and the number of engagements was favourable for survival.[229] In oral squamous cell carcinoma, Feng et al.[29] showed that high numbers of co-localisations within 30 μm of PD-$L1^+$ cells/ T-regs and $CD8^+$ T-cells is associated with reduced overall survival. Such studies help identify cancer specific cell-cell interactions and introduce novel predictive and prognostic biomarkers for high-risk patient identification.

In breast cancer, several studies observed the spatial pattern of lymphocytic infiltration in H&E images to identify tumours shielded from immune attack. Cheikh et al.[230] used graph based mathematical morphology to identify immune hotspots in H&E images, i.e. areas with dense aggregates of immune cells. For the same purpose, Maley et al.[231] suggested use of the Morisita-Horn co-localisation index, a metric commonly used in ecology, to observe proximity between immune and tumour cells. Another robust proximity metric was suggested for triple negative breast cancer by Yuan et al.[232], which was calculated as the distance from each

lymphocyte to the centroid of the convex hull formed by the five nearest cancer cells.

Finally, metrics are being introduced to represent the overall spatial heterogeneity between multiple cell phenotype populations in highly multiplexed tissue images. Rose et al.[100] proposed one of the first statistical frameworks to observe spatial interactions, the Hypothesized Interaction Distribution (HID). To represent a tissue sample, a symmetrical $N \times N$ array was constructed, where N was the number of cell phenotypes found in the sample. Each element in this array corresponded to a pair of phenotypes. The number of times these two phenotypes co-localised within a pre-defined distance threshold d of each other was entered in the array. The HID array could be represented by a single summary statistic, such as the Shannon's entropy or the array's energy[100] and its prognostic value was demonstrated in follicular lymphoma.[97] Construction of a similar array was later suggested for breast cancer by Spagnolo et al.[98], where each element represents the pointwise mutual information for each pair of phenotypes, instead of the number of co-localisations. Comprehensive reviews of methods assessing spatial heterogeneity with automated image analysis can be found in Yuan et al.[223] and Heindl et al.[233]

1.2.2.5 Software platforms supporting end-to-end multiplex analysis

As a final remark on multiplex image analysis, it is worth mentioning some open source or commercial software applications that offer end-to-end pipelines and support data acquired by multiplex immunofluorescence or other multiplex *in situ* technologies. InForm (Akoya Biosciences, Marlborough, MA, USA) is a commercial software platform that is bundled with the Vectra microscope and supports multispectral scanning and spectral unmixing, along with other functions such as tissue segmentation, label-free nuclear and cell segmentation, and supervised cell classification. HALO (Indica labs, Albuquerque, NM, USA) also supports multiplex image formats, currently offering a larger suite of compatible algorithms, including spatial proximity analyses. It uses parallel computing to speed up the process considerably. Additionally, QuPath,[234] the open source pathology software built by Peter Bankhead and his associates at the University of Edinburgh, currently supports multiplex image analysis. QuPath supports the processing of large histopathology whole slide images, a functionality currently missing from the otherwise very powerful Im-

ageJ.[235] QuPath also provides the option for custom scripting that could enable extension of its functionality for specialised applications, such as spatial analysis. Oncotopix (Visiopharm, Hørsholm, Denmark) is another commercial software platform supporting multiplex image analysis that also offers deep learning capabilities for cell phenotyping. Finally, histoCAT[236] is an open source platform that, when combined with ilastik[187] and CellProfiler[237] for cell segmentation and phenotyping, allows multiplex and spatial interaction analysis.

1.3 Aims and objectives

This section provides a brief synopsis of the clinical and technical background described in Sections 1.1-1.2 and states the aims and objectives of this thesis.

This thesis was motivated by the need for new cancer biomarkers that can accurately profile the microenvironment of tumours and assist in pre-treatment risk stratification for therapy selection, at the time of initial diagnosis. In Section 1.1 the potential for development of prognostic biomarkers based on the tumour microenvironment and its spatial architecture was identified in two different types of cancer: OPSCC (a solid tumour) and FL (a haematological cancer). Developing baseline prognostic biomarkers for risk stratification is a major area of research in these cancers, and in oncology generally, driven by a need to develop risk-adapted effective therapies capable of overcoming the poor outcomes of high-risk disease, and preventing over-treatment of low-risk disease.

In recent years, and in both cancers, novel immunotherapies (such as cetuximab and checkpoint inhibitors in OPSCC and rituximab in FL) have greatly improved patient outcomes. As treatments based on specific targeting of immune escape mechanisms slowly emerge, there is a need for novel biomarkers for treatment selection that would be tailored to specific host immune characteristics. The benefit of tumour microenvironment biomarkers, when compared with other prognostic indices (e.g., the FLIPI or stage), is that they allow a direct observation of interactions between relevant cell populations, permitting potential for discovery of new disease mechanisms, disease sub-types and novel therapy targets.

Multiplex *in situ* imaging was identified as a promising route to study the microenvironment of tumours (Section 1.2). Multiplex imaging, combined with automated image analysis, is capable of visualising concurrently a high number of antigens on biopsied tissue sections. Thus, extensive data becomes available for each single cell in the multiplex images and multiple cell phenotypes can be identified, while preserving the spatial context of the tissue.

Multiplexing *in situ* enables the investigation of a key property of the tumour microenvironment that is not well understood: its spatial architecture. Several studies have measured spatial architecture to gain valuable insights into the microenvironment of tumours. New biomarkers based on observing spatial distance and co-localisation of tumour infiltrating lymphocytes and tumour cells have demonstrated

predictive and prognostic value in gastric cancer, non-small cell lung cancer, head and neck cancer, breast cancer and follicular lymphoma, among others. These findings support further exploration of the tumour spatial context to improve precision in guiding treatment decisions in cancer.

Automated image analysis algorithms are required for the analysis of the complex information present in highly multiplexed images. Much progress has been made in recent years in the development of computer assisted scoring algorithms, able to assess antigen expression levels and derive scores for patient categorisation. The scoring task is an integral part of many routine analyses in anatomic pathology. However, the appropriate way to validate performance of computer assisted scoring algorithms is still an open research question, as the ground truth provided by pathologist manual scoring can lack reproducibility and be subject to a number of biases.[217] Thus, the strategy to validate computer assisted scoring algorithms and demonstrate comparable or superior performance to manual scoring remains unclear.

To explore the potential of multiplex spatial profiling of the tumour microenvironment for biomarker discovery in OPSCC and FL, the specific research objectives of this thesis were as follows:

i. To identify appropriate strategies for the validation of computer assisted scoring algorithms in anatomical pathology and design requirements to judge their performance.
ii. To develop and validate suitable multiplex assays and computer assisted scoring algorithms for the development of new biomarkers in OPSCC, based on the spatial proximity between cell types in the tumour microenvironment.
iii. To develop and validate suitable multiplex assays and computer assisted scoring algorithms for the development of new biomarkers in FL, based on concurrent observation of multiple immune populations and a spatially aware assessment of the FL microenvironment diversity as a whole.

2 Validation of computer assisted scoring approaches: a systematic review and meta-analysis

In anatomic pathology, upon routine examination of an immunohistochemically stained tissue biopsy, a pathologist will often provide a score to describe the level of antibody expression in relevant cell populations.[238] Scoring is a prerequisite for identifying cell phenotypes and analysing their spatial pattern in multiplex images. In recent years much progress has been made in the development of computer assisted scoring (CAS) tools to assist pathologists in scoring. Although many CAS algorithms have been introduced, the design requirements and reference methods for their validation have not been comprehensively considered. Furthermore, it is unclear whether CAS tools have achieved comparable performance to manual scoring.

Thus, before beginning the development of CAS tools for multiplex images in the next chapters, this chapter sought to clarify validation practices for CAS tools, by identifying design requirements, reference methods and assessing CAS performance. A systematic review of the literature on CAS tool validation for HER2, ER and three T-cell markers (CD3, CD4 and CD8) was carried out.

2.1 Introduction

Scoring is an integral part of histopathology with countless applications in supporting treatment decisions, predicting patient prognosis, or determining clinical trial design and enrolment. However, manual scores observed using a microscope are qualitative or semi-quantitative and so will always contain a degree of subjectivity, affected by a number of visual and cognitive biases.[217] In recent years, slide scanning technology lent itself to the development of automated image analysis tools, able to support the pathologist's workflow by performing scoring quantitatively and objectively. CAS tools aim to quantify the number and intensity of stained objects in tissue images, and offer a promising avenue towards better standardisation, reproducibility and throughput. Particularly in multiplex images, CAS is necessary for cell phenotyping, as correlating the information available from a high number of stains cannot be performed manually at a large scale.

Routine clinical adoption of CAS tools requires rigorous validation. Despite rapid advances in CAS technology, the appropriate way to validate performance of these

algorithms remains unclear. Typically, a reference method is needed as a comparator to establish analytical performance. Manual scoring is the most straightforward gold standard for CAS algorithm validation. Thus a paradox arises:[217] CAS is introduced as a more objective and reproducible alternative to manual scoring, however its performance is judged based on how well it agrees with the subjective manual scores. Furthermore, it is challenging to assess whether CAS has achieved superior performance to manual scoring, when good agreement of CAS with manual scoring is the only criterion.

To clarify the appropriate validation strategy for CAS systems, the requirements for clinical adoption, and their performance, this systematic review examined studies validating CAS tools for scoring the nuclear oestrogen receptor-α (ER), membranous human epidermal growth factor receptor-2 (HER2) and three T-cell markers (cluster of differentiation [CD] 3, CD4, CD8). Scoring of these antibodies is essential in many clinical and research applications in breast, lymphoma, head and neck and other types of cancer. While ER and HER2 expression levels are scored routinely in breast cancer to select adjuvant therapies,[239–241] T-cell populations in the tumour microenvironment are increasingly studied for their impact on cancer prognosis.[14,71,242,243] Multiple markers were included in this systematic review to provide insight on validation practices of CAS tools, irrespective of the markers they assessed. Even though the HER2 and ER markers are mostly relevant for breast cancer and not OPSCC or FL, a large majority of state-of-the-art CAS tools validated in recent years have focused on these markers due to their well-established role in routine clinical decision making. In contrast, not many CAS tools are currently available for the OPSCC and FL tumour microenvironment markers discussed in Chapter 1. By reviewing CAS tools for HER, ER, CD3, CD4 and CD8, a representative subset of this technology could be assessed.

While routine clinical scoring is typically carried out in the brightfield setting, research applications may also score samples acquired with immunofluorescent imaging. The scoring task in the two settings is identical, aiming to quantify protein expression levels visualised either with chromogens (brightfield) or fluorophores (immunofluorescence). This review retrieved CAS algorithms from the past 20 years validated in either setting.

2.1.1 Manual scoring systems

Before discussing design requirements for the validation of CAS, it is useful to describe the existing manual scoring systems. The scoring task is routinely carried out manually using a brightfield microscope. Several reference systems exist for manual scoring. The simplest systems only indicate if a sample is overall positive or negative for an antibody. To overcome the paucity of information available from this approach, other semi-quantitative systems have been adopted for ER, HER2 and T-cell scoring. These manual scoring systems also provide an estimate of the intensity of antibody expression in the sample.

2.1.1.1 ER scoring

ER is a nuclear marker. For ER, the American Society of Clinical Oncology (ASCO) and College of American Pathologists (CAP) guidelines[240] recommend scoring a sample as positive if >1% of tumour nuclei express the marker. In a clinical diagnostic setting, after reporting ER positivity status, pathologists may also supplement a semi-quantitative score incorporating the level of staining intensity, such as Allred or H-score. **Figure 11** demonstrates ER staining expression patterns in breast cancer.

In the Allred[244] (also known as “Quickscore”) system, a sample is assigned a score to indicate the percentage of positive tumour nuclei (0: none, 1: less than one tenth, 2: less than one third, 3: between one and two thirds, 4: more than two thirds), and a second subjective score (0-3) that indicates average staining intensity. The Allred is calculated by summation of those two scores and therefore and Allred of 0 is negative, 2-3 is weakly positive, 4-5 is intermediate positive and 6-7 is strongly positive. An Allred equal or greater than 2 corresponds to positive ER status.

The histochemical score (H-score)[245] may also be used, tumour nuclei are assigned a subjective level of intensity of staining (0: negative, 1: weak, 2: intermediate, 3: strongly positive). The overall H-score is calculated by multiplying each intensity score with the percentage of nuclei expressing it and then summing the results. The range of this continuous score spans from 0 to 300.

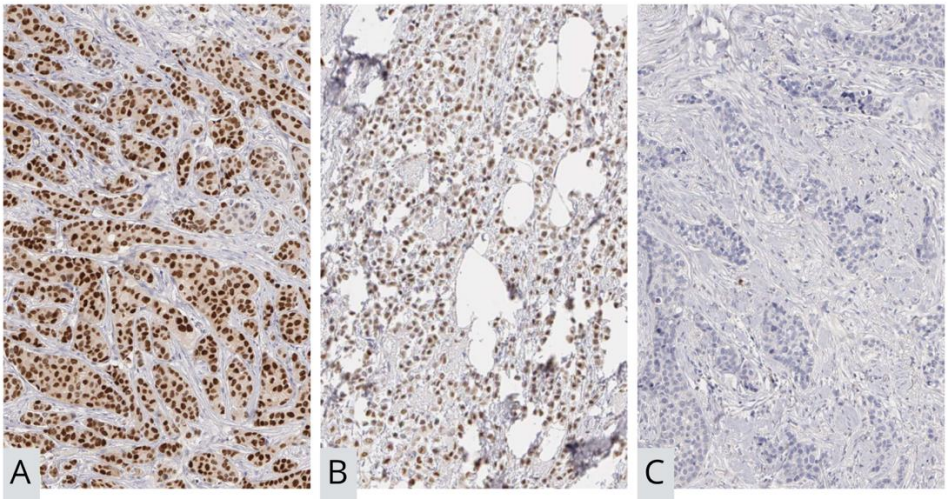

Figure 11 ER-alpha (ESR1 gene) staining expression levels in breast cancer, using HPA000449 antibody and brightfield immunohistochemistry. Image credit: Human Protein Atlas, available from v19.3.protein.atlas.org.[17,*] A: strong nuclear staining, B: moderate staining at >75% of tumour cells, C: negative staining.

2.1.1.2 HER2 scoring

True HER2 staining is localised on the tumour cell membrane. For HER2, the ASCO/CAP algorithm[246] uses a semi-quantitative HER2 score (0-3+) in breast cancer samples. To guide adjuvant treatment selection a definite decision is made on whether to consider a patient HER2 positive or negative. The ASCO/CAP algorithm is described in **Table 4**.

Table 4 Description of ASCO/CAP manual HER2 scoring algorithm

HER2 Immunohistochemical Staining	**HER2 score**
No stained tumour cells	0 negative
Incomplete, weak membrane staining in ≤ 10% of tumour cells	1+ negative
Complete staining in ≤ 10% of tumour cells or incomplete, weak staining in more than 10% of tumour cells	2+ equivocal
Complete and intense, circumferential membrane staining in more than 10% of tumour cells	3+ positive

[*]Online: https://www.proteinatlas.org/ENSG00000091831-ESR1/pathology/breast+cancer#, accessed 13/12/2020.

In the ASCO/CAP HER2 algorithm, 0-1+ patients are scored negative, 3+ patients are positive, and 2+ patients are equivocal and referred for supplementary fluorescent in-situ hybridisation (FISH) testing to clarify whether HER2 is amplified. **Figure 12** demonstrates HER2 staining patterns in breast cancer.

FISH is a powerful, yet costly, quantitative assay that preserves the spatial context of the tissue. Currently, a dual probe FISH assay is recommended,[246] where HER2 and centromere of chromosome 17 (CEP17) probes are counted concurrently. CEP17 can be used for HER2 probe count normalisation, as HER2 is known to be located on chromosome 17. Therefore, HER2 FISH status is determined based on the ratio of HER2 to CEP17 probes. Notably, some patients can still be classified as equivocal by FISH testing, in which case a repeat immunohistochemical or FISH assay is recommended to determine the final HER2 positivity status.

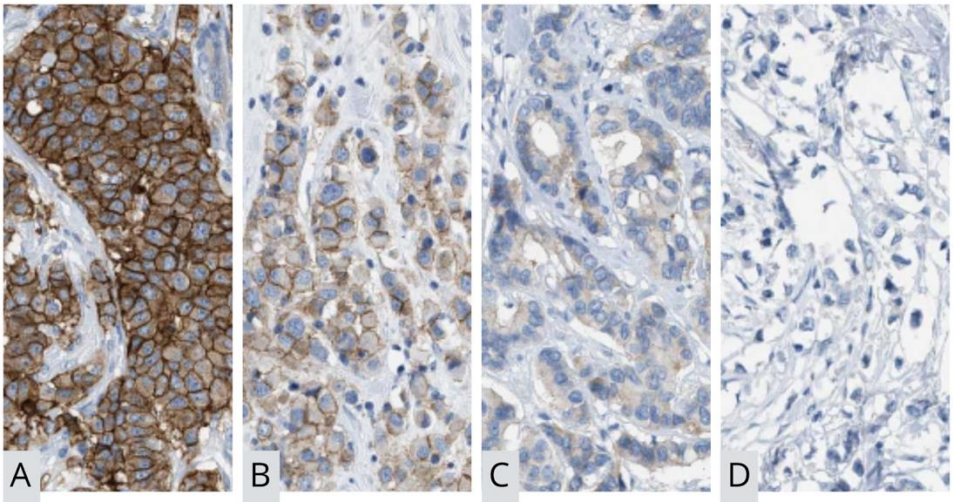

Figure 12 HER2 (ERBBR gene) staining expression levels in breast cancer, using CAB020416 antibody and brightfield immunohistochemistry. Image credit: Human Protein Atlas, available from v19.3.protein.atlas.org.[17,†] A: strong complete membranous staining, B: moderate staining, C: weak staining, D: negative staining.

2.1.1.3 T-cell marker scoring

The number or density of stained cells is usually reported for CD3, CD4 and CD8 during manual scoring. CD3 is a pan-T-cell marker, CD4 is predominantly expressed on the surface of T-helper cells, and CD8 is primarily found on the surface

[†] Online: https://www.proteinatlas.org/ENSG00000141736-ERBB2/pathology/breast+cancer#, accessed 13/12/2020.

of cytotoxic T-cells.[17] A cell is usually deemed positive for these markers based on cytoplasmic and/or membranous staining (**Figure 13**).

All manual scoring systems are potentially useful when used in a suitable context by trained experts and appropriately calibrated. The above-described semi-quantitative scores are the most popular in clinical and research settings and serve as the current gold standard in histopathological scoring.

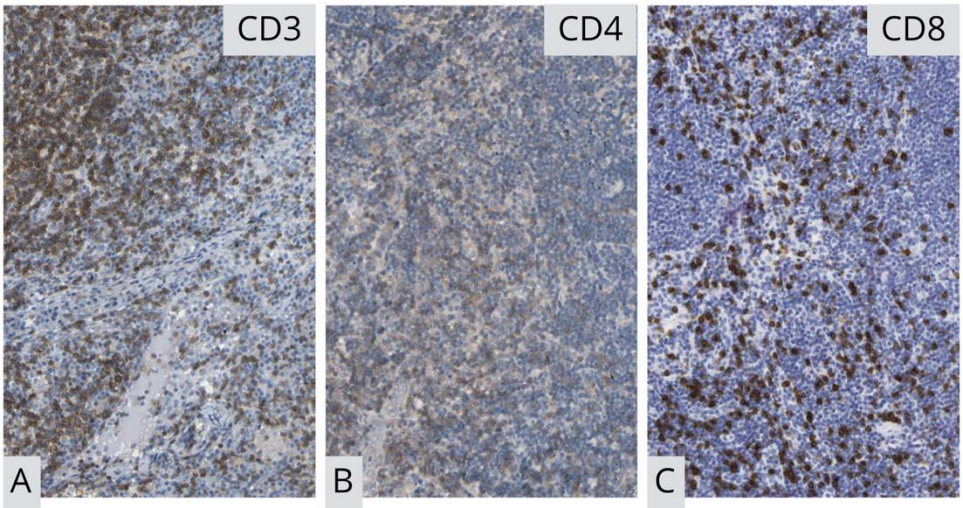

Figure 13 T-cell marker staining pattern in healthy lymph node tissue. Image credit: Human Protein Atlas, available from v19.3.protein.atlas.org.[17] A: CD3 staining, using CAB013055 antibody.[‡] B: CD4 staining, using HPA004252 antibody.[§] C: CD8 staining, using CAB000012 antibody.[**]

2.1.2 Review objectives

The strategy for validation and analytical performance of CAS algorithms for HER2, ER and T-cell markers were recorded. The objectives were to: i) outline design requirements fundamental for CAS systems, ii) describe how satisfaction of these requirements can be validated and iii) assess CAS performance. A key requirement of CAS systems is accuracy, often established through comparison with manual scoring. Performance of CAS in terms of agreement with manual scoring was described in a quantitative meta-analysis.

[‡] Online: https://www.proteinatlas.org/ENSG00000167286-CD3D/tissue/lymph+node#img, accessed 13/12/2020.
[§] Online: https://www.proteinatlas.org/ENSG00000010610-CD4/tissue/lymph+node#img, accessed 13/12/2020.
[**] Online: https://www.proteinatlas.org/ENSG00000153563-CD8A/tissue/lymph+node#img, accessed 13/12/2020.

2.2 Methods

The protocol is registered in the PROSPERO database of systematic reviews (no. CRD42019139688).[247] The review complies with PRISMA guidelines.[248] The following section describes how studies were selected for inclusion in the review and the data collected and plan for synthesis of findings.

2.2.1 Information sources and search strategy

The electronic bibliographic databases PubMed, Web of Science (Core Collection) and IEEE Xplore Digital Library were searched. Search strategy for PubMed was: (quantitat* OR automat*) AND (score OR scoring) AND (immunohistochem* OR immunofluorescen*) AND (("2000/01/01"[PDat] : "2019/12/31"[PDat]) AND Humans[MeSH]). The search terms were adapted for use with other bibliographic databases by including synonyms as necessary. The language was restricted to English. Filtering options were used when available to retrieve only human studies, peer-reviewed publications and exclude studies related to irrelevant domains.

2.2.2 Study eligibility criteria

CAS tools were reviewed for HER2, ER and T-cell markers (CD3, CD8, CD4) in the tumour microenvironment. Only peer-reviewed studies providing quantitative validation of CAS tools were included. If a study proposed a new CAS tool without quantitatively validating its performance, it was excluded. Studies validating algorithms against multiple markers without reporting performance for each individual marker were excluded; the premise that automated scoring performance is equivalent for different markers cannot be taken for granted *a priori*.

Selection criteria based on a pilot screening process and agreed by author consensus were formalin-fixed, paraffin-embedded (FFPE) or frozen tissue of human tumours or adjacent stroma, published between 1/1/2000 and 31/12/2019, with full text available. CAS tools developed either for immunohistochemical or immunofluorescent samples were included. Any staining protocol, detection system and scanning set-up was included, however, this information was recorded to provide context. Studies were excluded if they involved animal tissue, blood, cellular aspirates, cell lines or bony tissue.

2.2.3 Study selection

Studies retrieved according to the search strategy were pooled and duplicates removed. Retrieved studies were screened on their title and abstract. The selected studies were subsequently read in full and a final decision on inclusion made.

2.2.4 Data collection

A pre-piloted form was used to collect data (**Table 5)**. The form was trialled on ten studies by two investigators (A.M.T., A.M.). Collected data included information on sample preparation, algorithm description, and algorithm validation. Three trained investigators collected the data from the remaining studies (I.P.M., M.F., A.M.T., in duplicate or by independent reading).

2.2.5 Synthesis and meta-analysis methodology

Synthesis of findings listed the selected studies and categorised them based on antibody and imaging modality (immunofluorescence *vs* brightfield immunohistochemistry). Subsequently, the synthesis outlined the design requirements, i.e., the required attributes of CAS systems, and expanded on how these were validated in the reviewed studies.

A key design requirement of CAS systems is their accuracy, which is often established by comparing with manual scoring. To assess performance of CAS systems in terms of accuracy, a meta-analysis was performed. This meta-analysis quantified the agreement of CAS tools with manual scoring. Meta-analysis was limited to brightfield immunohistochemistry, because of the scarcity of standardised manual scoring systems for immunofluorescence. Agreement was probed by quantitative meta-analysis of Cohen's κ metric. Cohen's κ measures inter-rater variability, assessing whether the degree of agreement between two alternative forms of a test (here automated algorithm *vs* pathologist) is higher than expected by chance.

Studies were included in the meta-analysis if they reported data required for calculating Cohen's κ and its variance, as described by Sun.[249] For HER2, studies were included in the meta-analysis if they reported Cohen's κ using the 3-tier ASCO/CAP HER2 scoring system (0/1+, 2+, 3+). For ER, studies were included if they reported Cohen's κ for the dichotomised Allred score (≤ 2 = negative, > 2 =

positive). The limited number of T-cell studies prohibited quantitative meta-analysis.

Meta-analysis was performed with the *metafor* R package.[250] Heterogeneity was tested using the Higgin's I^2 statistic and scores >50% were considered heterogeneous. Forest plots were constructed, and combined effects were assessed by random effects meta-analysis because of the inherent variability in image preparation and algorithm components. Finally, sensitivity analysis was carried out by selecting different thresholds for minimum test size, minimum number of pathologists providing annotations and restricting to studies using an independent test set, and whole slide images.

Table 5 Pre-piloted form for data collection from reviewed studies

	Field	**Description**
General	Study details	Title, Author, Year published, Journal of Conference
	Marker	ER, HER2, CD3, CD4 or CD8
Image preparation	Compartment expressed	Membrane, Cytoplasm or Nucleus
	Number of patients	
	Type of images scanned	Regions of interest, whole sections or tissue micro-array cores
	Number of images	
	Reference scoring system	Either Allred, H-score, HER2 score, percentage of positive cells or other
	Scoring system details	
	Antibody clone	
	Detection system	Brightfield or fluorescence, single-plex or multiplex
	Type of tissue	Type of cancer, site of biopsy
	Fixation	Formalin fixed paraffin embedded or frozen samples
	Width of section (μm)	
	Scanner	
	Magnification	Magnification for scanning and processing
	Resolution (μm/pixel)	Resolution for scanning and processing
Algorithm	Software	Available software platform (commercial or open source)
	ROI selection	Method for region of interest selection
	Scoring algorithm description	
Validation setup & Performance	Type of ground truth (training)	Type of ground truth used to train or tune the algorithm
	Configuration of pathologists' annotations (training)	Single pathologist, multiple pathologists, consensus score or non-experts
	Number of annotations (training)	Number of training annotations
	Type of ground truth (testing)	Type of ground truth used to test the algorithm (does not matter if this is an independent test set or not)
	Configuration of pathologists' annotations (testing)	Single pathologist, multiple pathologists, consensus score or non-experts
	Number of annotations (testing)	Number of training annotations
	Independent test set	Indicate if the test set was independent from the training/ tuning dataset
	Number of pathologists	Number of pathologists that participated in the study overall
	Virtual reading	Ground truth by digital or microscope reading
	Notes on training testing	Additional information that may be relevant to describe the validation
	Agreement & Accuracy Metrics (vs human)	Values of all agreement metrics with pathologists' ground truth. For studies reporting Cohen's κ record or calculate confidence intervals.

	Additional performance metrics	List the metrics (e.g., time to process sample, inter-run agreement)
	Inter-observer agreement (automated)	
	Inter-observer agreement (manual)	
	Other modality used as ground truth	List the modalities (e.g., FISH)
	Agreement vs FISH (only for HER2)	Agreement with FISH for breast cancer studies, for manual and automated IHC scoring
	Nature of errors	As described in the study
	Comments	General comments
ROI indicates a region of interest; FISH indicates fluorescent in situ hybridisation; IHC indicates immunohistochemistry.		

2.2.6 Study quality

Individual study quality was assessed using the Hawker checklist,[251] where nine components (e.g. abstract, findings) are judged on a 0-3 scale. The sum of individual components provides an overall score, and studies were classified as low (0-9), medium (10-18) or high (19-27) quality. Initially, two reviewers assessed the quality of 12 studies, achieving good agreement (Spearman's rho = 0.94, $p < 10^{-5}$). One reviewer then assessed remaining studies. Differences in quality between studies were tested with the Kruskal-Wallis non-parametric test.

2.3 Results

2.3.1 Identified studies and their quality

Figure 14 shows the PRISMA flow chart[248] for the systematic review. Ninety-six studies were identified for qualitative synthesis (**Table 6**). A number of selected studies validated more than one algorithm; 65 algorithm validations are reported for HER2, 49 for ER and 13 for T-cell markers.

Table 6 Inventory of studies for each marker and imaging modality

	HER2	ER	CD3	CD4	CD8
Immunofluorescence	128,252,261,253–260	128,256,260,262–267	268	-	268
Immunohistochemistry	269,270,279–288,271,289–298,272,299–308,273,309–312,274–278	263,269,314–323,270,324–333,298,334–339,301,303,304,309,312,313	340–345	344	341,344–346

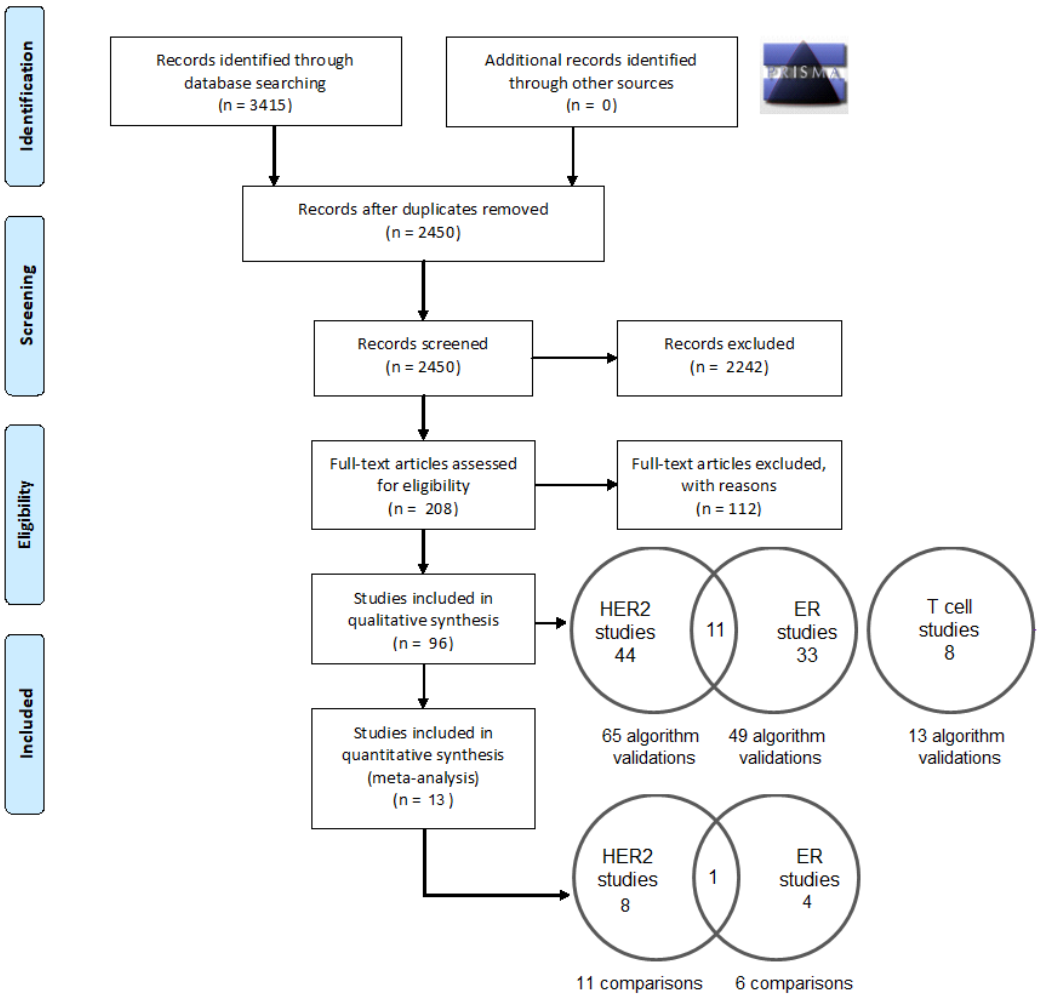

Figure 14 Adapted PRISMA (2009) flow chart[248] for study selection. After the initial screening 208 studies were fully read and assessed for eligibility based on the pre-defined selection criteria (see section 2.2.2). Ninety-six studies were finally included in the qualitative synthesis and 13 in the quantitative meta-analysis. This flow chart additionally indicates how many studies were reviewed per marker (HER2, ER and T-cell markers). Some studies validated more than one algorithm or validated algorithms for multiple markers (e.g., both ER and HER2). The number of algorithm validations ("comparisons") is also shown.

Most studies were medium to high quality (**Figure 15**), with no significant difference between markers (Kruskal-Wallis p=0.5).

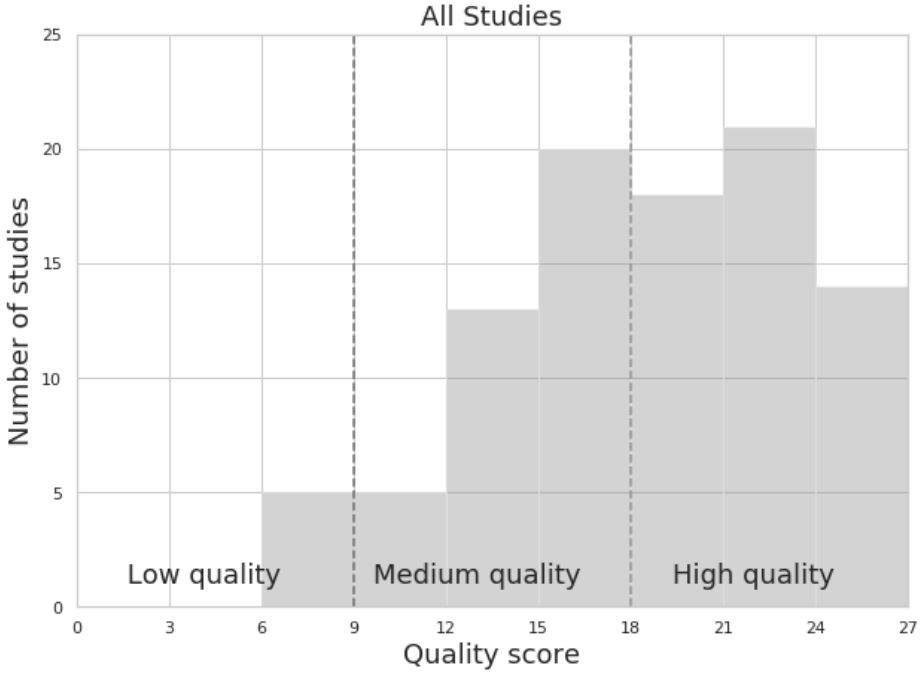

Figure 15 The distribution of quality scores obtained using the Hawker checklist for the 96 studies identified in the systematic review.

2.3.2 Validation of CAS design requirements

Validation of CAS determines how well they meet the design requirements for scoring systems. These design requirements were retrieved by review of studies validating CAS algorithms. They are similar for automated and manual systems; any scoring system should be definable, meaningful (accurate), reproducible, and time-efficient.[217,238] Particularly in the case of automated systems, additional design requirements may be introduced to ensure that algorithms can explain the basis of their decisions (interpretability) and indicate correctly when their predictions are uncertain, by providing a confidence estimation. An overview of the design requirements and how these were satisfied in reviewed studies is given in **Table 7.**

Table 7 Overview of computer assisted scoring (CAS) design requirement validation

Design requirement	***Description***	***Satisfied by***
Definability	Well defined algorithm and intended usage	Description of algorithm, sample preparation and validation setup
Accuracy	Accurate (meaningful) patient categorisation	Either of: 1. Agreement with previously validated manual scoring system 2. Agreement with previously validated orthogonal assay 3. Correlation with patient clinical endpoint
Reproducibility	Robustness to staining and imaging variability, consistent scores when different pathologists operate interactive CAS tools	1. Good intra/inter-lab agreement 2. Good intra/inter-observer agreement for interactive CAS tools
Time-efficiency	Time-efficient sample processing	Time efficiency comparable to manual scoring
Interpretability	Explaining why a score was assigned	Producing salient features or/ and representative image regions
Confidence estimation	Accurate indication of uncertainty about the assigned score	Accurate confidence estimates

2.3.2.1 Definability

Scoring systems should be definable and based on a predetermined set of rules. Definability can be demonstrated by describing the scoring algorithm, as well as its intended usage by providing information on sample preparation, staining and scanning, the setup for algorithm validation and the type of reference ground truth.

Error! Reference source not found. shows what percentage of the reviewed algorithm validations provided information on the type of images (whole slides, regions of interest or TMA core images), image acquisition setup, resolution, and described the validation setup. While general information on the image acquisition setup and type of images was generally well described, details on the image resolution were often lacking. When human annotations were used as reference ground truth, most

studies provided detailed information on how and by whom the annotations were produced.

CAS algorithms are usually well defined and rely on objective and measurable properties of the scanned slides to arrive at a decision. However, while some CAS systems are fully automated, others involve a degree of interaction by a pathologist.[303] The pathologist operates the automated tool, tunes its parameters appropriately,[310] selects regions for analysis,[278,328] accounts for predicted scores and other visual aids, and makes a decision. The intervention of a pathologist in the scoring process potentially introduces a degree of subjectivity. Despite this, scoring rules in CAS remain much better defined than the approximate intensity estimates of manual scoring systems, as the pathologist's decision is usually based on quantitative information of image properties.

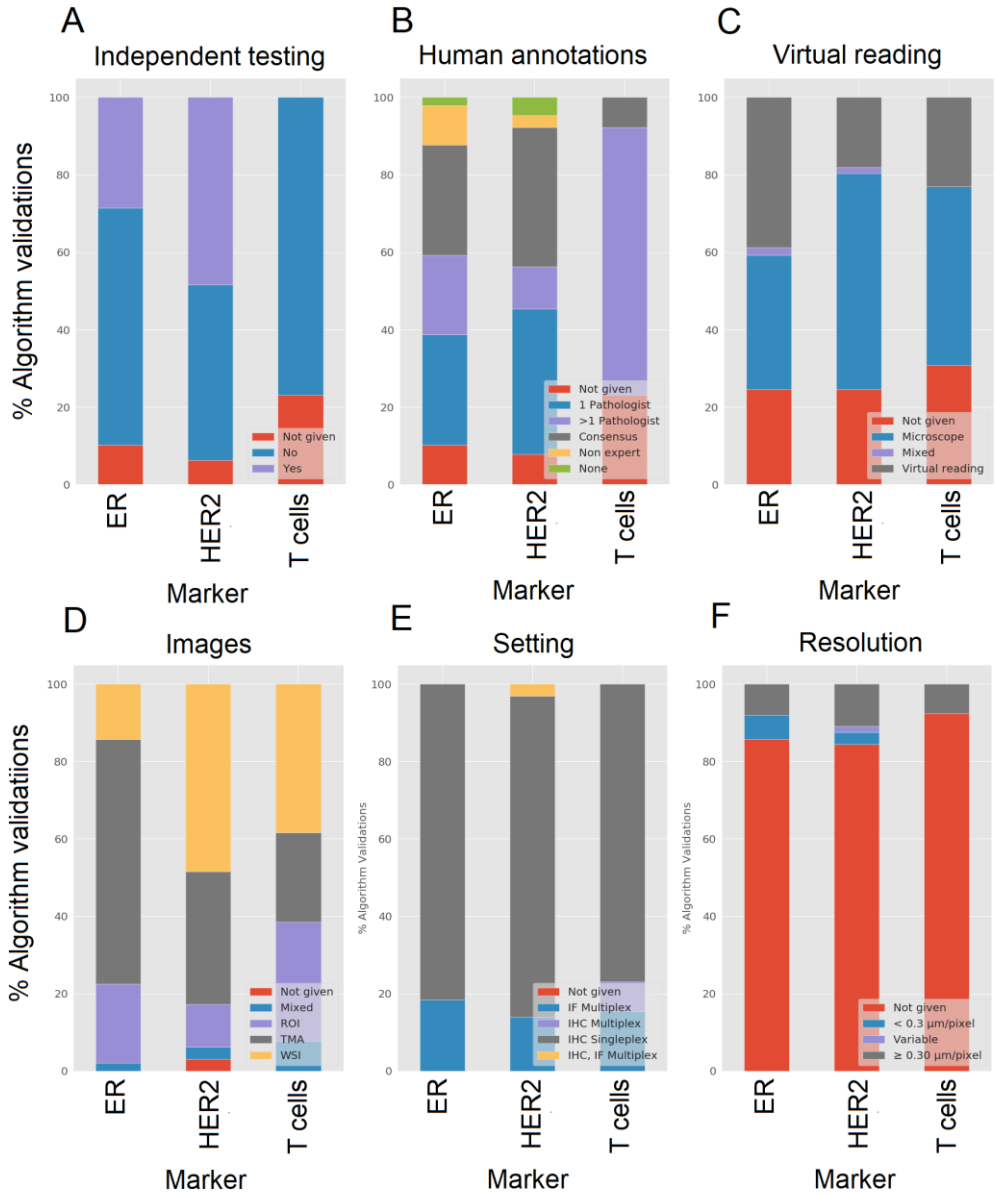

Figure 16 Details on image preparation, imaging setup, resolution and validation setup from the 96 reviewed studies. ROI indicates images of regions of interest; TMA, images of tissue microarray cores; WSI: whole slide image; IF, immunofluorescence; IHC; bright-field immunohistochemistry. A) Bar chart indicating whether an independent test set was used to assess performance during validation; B) Bar chart indicating who provided human annotations (pathologists *vs* non-experts) to be used as reference ground truth during algorithm validation. "Not given" indicates that human annotations were used but information on who provided them is not described. "Consensus" indicates that a consensus score by more than one pathologist was used as reference ground truth. "None" indicates that no human annotations were used during validation; C) Bar chart indicating whether the human annotations used as reference ground truth were produced by virtual reading (using a monitor) or microscope reading. D) Type of images used as input to the CAS system; E) Image acquisition setup; F) Resolution of images, if explicitly mentioned.

2.3.2.2 Accuracy

Scoring systems should be meaningful. This requirement translates to an accurate patient categorisation that can predict clinical outcome, select patients for treatment or enrolment in clinical trials and is related to relevant patient clinical characteristics. The accuracy of the scoring system was evaluated in one of three ways in the reviewed studies: CAS studies either validated agreement with an equivalent manual scoring system that had previously been proven to be clinically meaningful (Section 2.3.2.2.1), validated agreement to orthogonal quantitative assays (Section 2.3.2.2.2), or directly demonstrated correlation with patient outcome (Section 2.3.2.2.3).

2.3.2.2.1 Accessing accuracy via agreement with equivalent manual scoring system

Manual scoring is the most straightforward comparator for CAS and most commonly adopted by the reviewed studies. Even though concerns were raised[217] on the reliability of manual annotations as the gold standard, approximately half (29/65) of the reviewed HER2 CAS validations did not use any additional comparators to verify performance. For ER, over half the algorithm validations relied only on pathologist's annotations (36/49), while one study used crowdsourcing of non-expert manual scoring[339] as ground truth. Last, for T-cell markers all reviewed studies used pathologists' annotations as the gold standard. This approach to establishing CAS accuracy is potentially limited by the quality and accuracy of the reference manual scoring system.

The manual reference scoring systems adopted for CAS validation in the reviewed studies were heterogeneous. CAS tools for HER2 were validated 80% of the time against the reference ASCO/CAP HER2 score. Most algorithm validations (60/65) involved breast cancer samples. For ER, 16 studies adopted the Allred, eight the H-score, five the % of positive cells, and 20 in-house scoring systems. Most validations of automated scoring algorithms for ER were also carried out in breast cancer samples (48/49). Validation of CAS tools for T-cell markers involved multiple types of cancer. The percentage of manually detected positive cells was the comparator for all T-cell marker reviewed studies, either as a continuous or categorical score.

This study performed quantitative meta-analysis of agreement between CAS and manual scores for ASCO/CAP HER2 scoring and ER Allred, to establish the overall level of agreement when these reference scoring systems are used as ground truth. For quantitative meta-analysis, 9 HER2 studies (11 comparisons) and 5 ER studies (6 comparisons) provided sufficient information to calculate the Cohen's κ against pathologists' ground truth. Frequent reasons for exclusion were lack of automated methodology or quantitative validation, and markers other than the predefined.

2.3.2.2.1.1 Meta-analysis of HER2 studies

Table 8 details the characteristics of HER2 studies included in random effect meta-analysis. The results of the random effects meta-analysis placed the summary estimate of Cohen's κ for all HER2 CAS studies at 0.75 (95% CI: 0.70-0.81). However, high heterogeneity was present with Higgin's I^2=79.5% (95% CI: 54.2-93.6), suggesting 79.5% of the variability in performance is due to differences in study characteristics and only 20.5% due to chance. The random effects model is presented in **Figure 17**. In this analysis, the HER2 score was considered as: 0/1+ negative, 2+ equivocal and 3+ positive.

To judge whether this agreement between CAS and pathologists was satisfactory, how well pathologists usually agree with each other was investigated. As reference, the average inter-pathologist agreement from the studies of Bloom et al.[347] (κ = 0.60, 95% CI: 0.53-0.68) and Jefferson et al.[348] (κ = 0.77, 95% CI: 0.71-0.83) was plotted. Bloom et al.[347] reported inter-observer agreement between 10 pathologists for 126 whole slide images (WSI) while Jefferson et al.[348] provided inter-lab agreement between 17 laboratories for 36 tissue microarray (TMA) cores. Based on the raw data from these two human observer studies the agreement of each pathologist with the consensus score was calculated and the average Cohen's κ and 95% CI was reported as a reference. The overall performance was satisfactory; the automated HER2 scoring algorithms agreed with pathologists at least as well as pathologists agreed with each other.

Table 8 Studies included in meta-analysis of Cohen's κ agreement for HER2

Study	Cohen's κ [95% CI]	% Agreement	Images	Annotations	Data set	Independent test set
Khameneh et al. (2019)[272]	0.79 [0.64, 0.94]	87.0	WSI	≥ 1 pathologist consensus	52 patients	Yes, different cohort
Vandenberghe et al. (2017) [311]	0.69 [0.53, 0.85]	83.0	WSI	Single pathologist	71 patients	Yes, same cohort
Holten-Rossing et al. (2015) [288]	0.74 [0.69, 0.79]	90.5	TMA Cores	Single pathologist	904 cores	No
Micsik et al. (2015) [307]	0.87 [0.81, 0.94]		TMA Cores	> 2 pathologists' consensus	173 cores	No
Howat et al. (2014)[312] Rater 1	0.71 [0.63, 0.79]	93.7	TMA Cores	Single pathologist	716 cores	No
Howat et al. (2014)[312] Rater 2	0.62 [0.53, 0.71]	90.7	TMA Cores	Single pathologist	693 cores	No
Mohammed et al. (2012)[274]	0.81 [0.74, 0.88]	94.2	TMA Cores	2 pathologists scored separately or consensus	431 patients	No
Laurinaviciene et al. (2011)[286] Round 1	0.69 [0.55, 0.83]	89.4	TMA Cores	Single pathologist (scored twice)	161 patients	Yes, different cohort
Laurinaviciene et al. (2011)[286] Round 2	0.8 [0.69, 0.91]	92.5	TMA Cores	Single pathologist (scored twice)	161 patients	Yes, different cohort
Brügmann et al. (2011)[278]	0.86 [0.82, 0.91]	92.3	TMA Cores	5 pathologists' consensus	430 cores	Yes, same cohort
Minot et al. (2009)[296]	0.58 [0.43, 0.73]	84.3	WSI	Single pathologist	159 patients	Yes, different cohort
CI: Confidence intervals, WSI: whole slide images, TMA: tissue microarray						

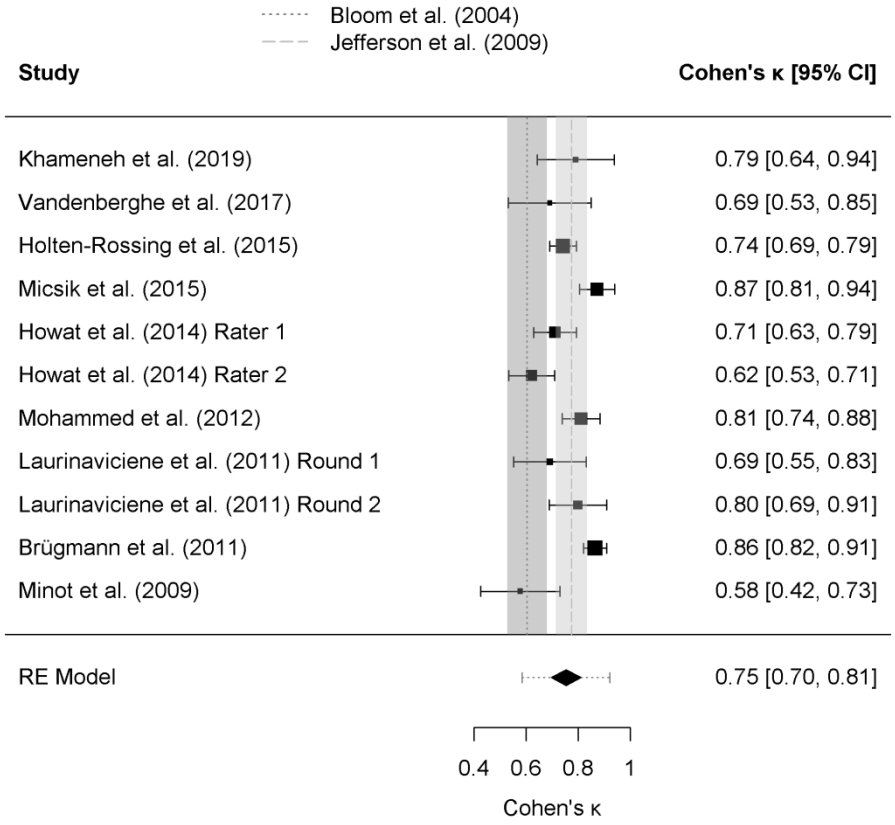

Figure 17 Random effects meta-analysis of Cohen's κ for HER2 scoring algorithm performance. The size of markers represents the size of the dataset for which performance is reported. All findings correspond to a three-class HER2 score (0 or 1+ as negative, 2+ equivocal, 3+ positive). As benchmark, the human inter-observer agreement for the HER2 scoring task is plotted for two studies; Bloom et al.[347] and Jefferson et al.[348]

2.3.2.2.1.2 Meta-analysis of ER studies

Table 9 details the characteristics of ER studies included in random effect meta-analysis. The results of meta-analysis placed the summarised agreement for all ER CAS algorithms with manual scoring at κ = 0.74, 95% CI: 0.66-0.83 (**Figure 18**). To assess agreement with manual scoring in this analysis, Cohen's κ was reviewed and the dichotomized Allred score (Allred ≤2: negative, Allred >2: positive) was selected as reference scoring system. Again, high performance heterogeneity was present with Higgin's I^2=91.0% (95% CI: 68.4-98.4).

To judge whether agreement of CAS with manual scoring was satisfactory, how well pathologists usually agree with each other for the same task was investigated. Although the overall agreement of CAS with pathologists was good, it fell short of the excellent inter-pathologist agreement reported[348] (κ = 0.96, 95% CI: 0.93-0.99) for manual dichotomized ER Allred scoring.

Table 9 Studies included in meta-analysis of Cohen's κ agreement for ER

Study	Cohen's κ. [95% CI]	% Agreement	Images	Annotations	Data set	Independent test set
Ali et al. (2013)[304]	0.82 [0.78, 0.85]	93.2	TMA cores	> 1 pathologist	1664 cores	No
Howat et al. (2014)[312]	0.62 [0.6, 0.64]	84.1	TMA cores	≥ 2 pathologists' consensus	6424 patients	Yes, different cohort
Sarikoc et al. (2013)[314] Rater 1	0.77 [0.52, 1.02]	88.9	ROI images	Single pathologist	27 ROIs	Yes, same cohort
Sarikoc et al. (2013)[314] Rater 2	0.84 [0.63, 1.05]	92.6	ROI images	Single pathologist	27 ROIs	Yes, same cohort
Bankhead et al. (2018)[301]	0.69 [0.6, 0.78]	84.3	TMA cores	Single pathologist	267 patients	No
Trahearn et al. (2017)[333]	0.85 [0.65, 1.05]	96.0	WSI	1st pathologist scored all, consensus with 2nd pathologist for cases discrepant with automated prediction	50 patients	No

TMA: tissue microarray, ROI: regions of interest, WSI: whole slide images.

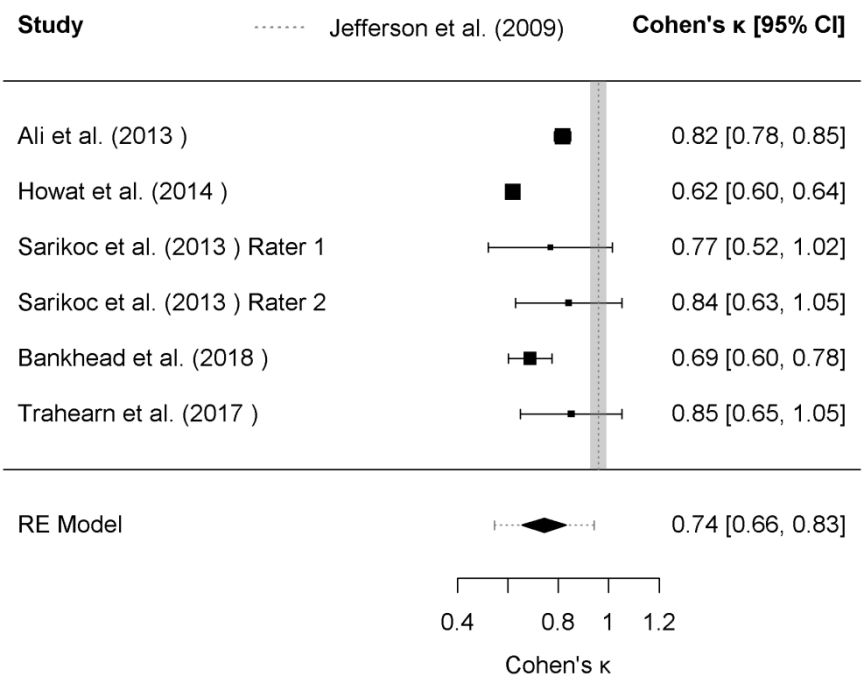

Figure 18 Random effects meta-analysis of Cohen's κ for ER scoring algorithm performance. The size of markers represents the size of the dataset for which performance is reported. All findings correspond to a dichotomised Allred score (≥3 positive, ≤2 negative).

As benchmark, the human inter-observer agreement for the same task is plotted, using data from the Jefferson et al.[348] study.

2.3.2.2.1.3 Sensitivity analysis of the meta-analyses

Varying the restrictions on cohort size, and whether this represented an independent test set, had no effect on the meta-analysis results (**Table 10**). When only studies reporting results for whole slide images were selected, this did not affect performance for HER2; this comparison was not possible for ER.

Table 10 Sensitivity analyses based on the size of dataset, number of pathologists providing annotations, use of an independent test set and use of whole slide images

Sensitivity analysis	**HER2**		**ER**	
Criteria	**Cohen's kappa [95% CI]**	**Algorithms**	**Cohen's kappa [95% CI]**	**Algorithms**
All studies	0.75 [0.70, 0.81]	11	0.74 [0.66, 0.83]	6
Dataset ≥ 100 images	0.75 [0.69, 0.82]	9	0.71 [0.59, 0.83]	3
> 1 Pathologist	0.85 [0.82, 0.88]	4	0.75 [0.60, 0.90]	3
Independent test set	0.75 [0.66, 0.84]	6	0.71 [0.56, 0.86]	3
Whole slide images	0.69 [0.56, 0.81]	3	Not available	1

In the sensitivity analyses, the only factor shown to affect significantly the overall accuracy of CAS algorithms was the number of pathologists providing annotations. Improved Cohen's κ was observed for studies using more than one pathologist (κ = 0.85, 95% CI: 0.82-0.88) for HER2 CAS validation (**Table 10**), compared to all studies (κ = 0.75, 95% CI: 0.70-0.81). Since pathologists manually scoring HER2 agree only moderately well to each other,[347,348] the improved performance in this case could be attributed to better quality ground truth, acquired by using multiple pathologists as reference to reduce subjectivity. This effect was less pronounced in ER scoring, potentially because the inter-pathologist agreement for manual scoring is known to be excellent.[348] Therefore, using multiple pathologists in manual ER scoring would produce the same scoring results as using a single pathologist and would not necessarily improve the quality of reference ground truth.

2.3.2.2.1.4 Agreement of T-cell CAS with manual scoring

The limited number of T-cell studies and the heterogeneous accuracy metrics they adopted did not allow quantitative meta-analysis, as described above for HER2 and ER. However, **Table 11** reports the agreement between CAS and manual scoring

for each study separately. As reference, the manual inter-observer agreement between pathologists was sought for the task of scoring the % of positive cells; Singh et al.[341] reported Pearson's r=0.83 for CD8, and r=0.80 for CD3, while Halama et al.[343] reported Pearson's r=0.91 for CD3. Inter-observer agreement data for manual CD4 scoring was not available. Overall, the performance of CAS for CD3 and CD8 appears to vary across different studies and no study achieved good performance for CD4.

2.3.2.2.1.5 Limitations of manual scoring as a reference gold standard

Manual scoring can be a sub-optimal gold standard for CAS validation, as it is subject to a degree of subjectivity, visual and cognitive biases.[217] Particularly in scoring tasks that demonstrate high intra- and inter-pathologist disagreement, relying on annotations from a single pathologist can hinder the training of automated algorithms and produce biased CAS systems. In such case, using consensus scores from multiple pathologists may be beneficial.

Manual scoring can be an unsuitable reference method for CAS in immunofluorescence. Standardised manual scoring systems are scarce in the immunofluorescent setting and the manual inter-observer agreement is usually unknown.

Furthermore, CAS tools will usually produce continuous scores, where manual scoring is only provided as semi-quantitative categorical scores. This discrepancy makes comparisons less straightforward. When there are concerns that manual scoring is unreliable, orthogonal quantitative assays providing continuous assessment of protein expression levels may be a more suitable reference method. These are discussed in the next section.

Table 11 Agreement with pathologists' ground truth for automated scoring of T-cell markers (CD3, CD4, CD8)

Marker	Study	Performance*	Images	Annotations	Data set	Independent test set
CD3	Singh et al. (2018)[341]	Pearson's r: 0.83	WSI	Single pathologist	35 patients	No
	De Meulenaere et al. (2018)[345]	Spearman's rho: 0.63	ROI	>1 pathologist scored	75 patients	No
	Bouzin et al. (2015)[342]	Pearson's r: 0.90	WSI	>1 pathologist scored different samples	64 patients	No
	Sander et al. (2014)[344]	% Agreement: 32-43	TMA Cores	7 pathologists scored the same sample	54 cores	No
	Halama et al. (2009)[343]	Pearson's r: 0.96-0.97	WSI	2 pathologists scored the same sample	30 ROI	No
CD8	Singh et al. (2018)[341]	Pearson's r: 0.92	WSI	Single pathologist	35 patients	No
	Hartman et al. (2018)[346]	Pearson's r: 0.97	TMA Cores	Configuration not given	122 cores	No
	De Meulenaere et al. (2018)[345]	Spearman's rho: 0.65	ROI	>1 pathologist scored	75 patients	No
	Sander et al. (2014)[344]	% Agreement: 17-78	TMA Cores	7 pathologists scored the same sample	54 cores	No
CD4	Sander et al. (2014)[344]	% Agreement: 43-73	TMA Cores	7 pathologists scored the same sample	54 cores	No

*Pearson's r corresponds to ground truth provided as a continuous score representing the % of positive cells. Spearman's rho and % agreement correspond to ground truth provided as an ordinal score reflecting the % of positive cells. WSI: whole slide images, TMA: tissue microarray, ROI: regions of interest.

2.3.2.2.2 Assessing accuracy via agreement with orthogonal assays

Several studies adopted orthogonal protein expression assays as a reference gold standard in CAS validation.

2.3.2.2.2.1 Orthogonal assays for validating CAS of HER2 expression

The most frequently used orthogonal assay for HER2 CAS tool validation was chromogenic (CISH) or fluorescent (FISH) *in-situ* hybridisation (23 studies). FISH holds an important role in clinical HER2 scoring, as ASCO/CAP guidelines recommend it as a suitable alternative to immunohistochemistry.[246] Furthermore, patients deemed HER2 equivocal by manual immunohistochemical scoring are referred to supplementary FISH to confirm HER2 amplification. If the adoption of CAS in HER2 immunohistochemical assays could reduce the number of patients assigned as equivocal and demonstrate high concordance with FISH, then the need for and cost of confirmatory FISH testing would be reduced. To test this hypothesis, several studies evaluated agreement between immunohistochemistry and FISH in breast cancer, when CAS or manual scoring was applied (**Table 12**). Comparison between these studies is challenging, as ASCO/CAP criteria for FISH scoring have changed over the years.[279] From this analysis, no clear benefit was demonstrated when CAS is used instead of manual scoring. The numbers of patients assigned as equivocal (needing FISH) was similar for CAS and manual scoring, across all studies. Furthermore, CAS and manual scoring has similar negative and positive predictive value for patients designated 0, 1+ and 3+, when FISH was the reference gold standard.

Even though CAS cannot fully replace the need for FISH testing, FISH scoring can still provide a good orthogonal assay for HER2 CAS system validation. Other quantitative assays such as qRT-PCR[254,257] and RNA expression assays,[270] have been used for the same purpose. Additionally, the PAM50 genotypic categorisation[349] introduced in 2009, could be used as the reference to validate HER2 scoring.[298,309] This categorisation is used to classify patients into four intrinsic subtypes with different prognostic outcome: HER2-enriched, luminal A, luminal B, and basal-like (triple-negative). The HER2-enriched subtype is known to be HER2 positive and ER/PR negative. Luminal A is HER2 negative and ER/PR positive, with low Ki67 expression. Luminal B subtype is ER/PR positive and either HER2 positive or negative with high Ki67 expression. Finally, the basal-like subtype is negative for HER2 and ER/PR.

2.3.2.2.2.2 Orthogonal assays for validating CAS of ER expression

Similar to HER2, in ER CAS validation, some studies compared to PAM50 genotyping[298,339], dextran-coated charcoal ER biochemical assay[319] or qRT-PCR.[327] A synthetically generated data set, where the images were artificially simulated for ER positive and negative samples was also suggested for the initial tuning of the ER CAS algorithm[331]. FISH was not used for ER CAS tool validation in any of the reviewed studies.

2.3.2.2.2.3 Orthogonal assays for validating CAS of T-cells

For T-cell markers, only flow cytometry was used as an alternative comparator.[268,344] Even though the analytical validity of this method is well established, it does not preserve the spatial context of the tissue and requires fresh samples for evaluation, that can be difficult to acquire.

The orthogonal assays discussed in this section could be used as reference for the validation of CAS tools, independently or in addition to manual scoring annotations.

2.3.2.2.3 Assessing accuracy via agreement with patient outcome

CAS algorithms can be established as meaningful and accurate without the need for comparison with alternative reference scoring methods, if they are shown to significantly and robustly correlate with patient outcome or other relevant clinical characteristics. Several HER2 CAS studies[253,255,308,258,259,269,274,285,298,301,304] considered survival endpoints as the ultimate gold standard. For ER, again, several studies[255,263,269,298,301,304,319,321] used survival analysis as a comparator. For T-cell scoring, CAS tools were validated against pathological complete response[268,344] or overall survival[346]. Even though it directly assesses the relationship between scoring and patient outcome, survival analysis is challenging, requiring long patient follow-up and controlling for multiple confounding factors. In the cases where CAS is validated for tumour biomarker development, the REMARK guidelines[350] provide recommendations to report validation findings.

Table 12 Studies reporting agreement between HER2 immunohistochemistry (IHC) and FISH. IHC was scored using both manual scoring and CAS for comparison.

			% FISH Amplified (CAS)†			% FISH Amplified (Manual)†			CAS		Manual Scoring	
Study	**FISH Reporting**	**Dataset for FISH**	**0/1+**	**2+**	**3+**	**0/1+**	**2+**	**3+**	**2+ Cases**	**Agreement with FISH**	**2+ Cases**	**Agreement with FISH**
Holten-Rossing et al. (2015)[288]	ASCO/CAP 2013[351]	904 cores	0.4	2.5	88.5	0	6.2	100	41	Accuracy: 0.94 Cohen's κ: 0.76	127	
Micsik et al. (2015)[307]	ASCO/CAP 2007[352]	35 cores, only 2-3+	-	50	100	-	50	100	12		12	
Mohammed et al. (2012)[274]	Amplified if the ratio of HER2/CEP17 > 2.0	431 patients	3.7	64.7	91.8	2.5	78.3	96.4	17	ICCC: 0.92	23	ICCC: 0.95
Tuominen et al. (2012)[284]	Amplified if the ratio of HER2/CEP17 > 2.2, equivocal if 1.8-2.2	144 patients*	3	30	85.7				30		20	
Brügmann et al. (2011)[278]	Amplified if the ratio of HER2/CEP17 > 2.2, equivocal if 1.8-2.2	430 ROI (22 patients)*	0	13.5	98.7	0	27.3	98	37		44	
Atkinson et al. (2011)[279]	Amplified if the ratio of HER2/CEP17 > 2.2	997 WSI, 2+ excluded								Accuracy: 0.95 Cohen's κ: 0.86		Accuracy: 0.95 Cohen's κ: 0.85
Atkinson et al. (2011)[279]	Amplified if the ratio of HER2/CEP17 ≥ 2.0	997 WSI, 2+ excluded								Accuracy: 0.92 Cohen's κ: 0.79		Accuracy: 0.92 Cohen's κ: 0.78
Laurinaviciene et al. (2011)[286]	Amplified if the ratio of HER2/CEP17 > 2.0	152 patients*	4.4	15.8	80	2.5	13.3-25	82.4-88.2	19		8-15	
Cantaloni et al. (2011)[277]	Amplified if the ratio of HER2/CEP17 ≥ 2.2	292 patients, only 2+								ROC AUC: 0.81 (0.75-0.87)		ROC AUC: 0.79 (0.72-0.86)
Turashvili et al. (2009)[285]	Amplified if the ratio of HER2/CEP17 > 2.2, equivocal if 1.8-2.2	616 patients							2-3 times more than manual	Weighted κ; user 1: 0.67 (0.61–0.72), user 2: 0.54 (0.49–0.58)		Weighted κ; pathologist 1: 0.81 (0.77–0.86), pathologist 2: 0.76 (0.71–0.81)
Hall et al. (2008)[297]	Amplified if the ratio of HER2/CEP17 ≥ 2.3	99 patients										ROC AUC: 0.82

ICCC: Intra-class correlation coefficient, ROC AUC: Area under curve for receiver operating characteristic. Only studies that reported quantitative agreement with FISH both for CAS and manual scoring are shown. *Independent validation set. †The percentage of samples that were recorded as amplified by FISH, for each immunohistochemistry score (0/1+ vs 2+ vs 3+).

2.3.2.3 Reproducibility

Scoring systems should be reproducible. A fully automated CAS system should produce consistent scores when presented with the same image at different times. An interactive semi-automated CAS system, where pathologists are involved as users of the system, should produce consistent scores when the same sample is assessed by the same pathologist at a later time, or by different pathologists.

To gauge whether CAS guarantees agreement between different pathologists, studies have assessed inter-observer agreement in interactive CAS compared to manual scoring (**Table 13).** This analysis was not possible for T-cell studies, as none of the studies compared inter-observer agreement when using CAS *vs* manual scoring. Even though heterogeneous metrics were used, there was a trend for improved inter-observer agreement when CAS was used, compared to manual scoring.[287,303] This improvement was not always statistically significant, and in a few cases[279,319] CAS underperformed. Thus, the inter/intra-observer agreement for CAS systems are properties that should be explicitly validated.

Furthermore, scores obtained by CAS should be reproducible when the same sample is stained in the same lab at different times, by different operators, or in different labs. CAS is expected to cope at least as well as pathologists, when presented with variability stemming from experimental conditions. For this reason, several CAS validation studies assessed agreement between intra-run, inter-run and inter-lab experiments.[259,284,338,341] To satisfy this design requirement, CAS systems have been developed to automatically detect the image acquisition setting[276] and calibrate accordingly.[310]

Table 13 Comparison of inter-observer agreement in manual scoring and CAS

		Manual scoring	Computer assisted scoring
Marker	Study	Inter-observer agreement	Inter-observer agreement
HER2	Barnes et al. (2017)[303]	% agreement = 0.87 [0.82-0.92]	% agreement = 1.0 [0.97-1.0]
	Atkinson et al. (2011)[279]	Cohen's κ = 0.72 [SD 0.09]	Cohen's κ = 0.84 [SD = 0.05]
	Gavrielides et al. (2011)[287]	Kendall's τ = 0.77 [0.74-0.80]	Kendall's τ = 0.86 [0.84-0.89]
	Słodkowska et al. (2010)[292] ACIS	% agreement = 0.84	% agreement = 0.88
	Słodkowska et al. (2010)[292] APERIO	% agreement = 0.84	% agreement = 0.80
	Turashvili et al. (2009)[285]	Weighted κ = 0.93 [0.91-0.95]	Weighted κ = 0.81 [0.79-0.83]
ER	Barnes et al. (2017)[303]	% agreement = 0.95 [0.91-0.98]	% agreement = 0.98 [0.94-1.0]
	Zarrella et al. (2016)[263] AQUA	Pearson's r = 0.96	Pearson's r = 1.0
	Zarrella et al. (2016)[263] APERIO	Pearson's r = 0.97	Pearson's r = 0.98
	Nassar et al. (2011)[337]	% agreement = 0.94 [0.91-0.99]	% agreement = 0.98 [0.98-0.99]
	Turbin et al. (2007)[319]	Cohen's κ = 0.92 [0.90–0.93]	Cohen's κ = 0.91 [0.90–0.93]
Values in brackets [] represent 95% CI.			

2.3.2.4 Time-efficiency

A fundamental goal in anatomic pathology is to render a diagnosis in a timely fashion, in a way that is useful to the physician treating the patient.[353] Time-efficiency is a design requirement that should be satisfied by CAS, at least as well as by manual scoring. The challenge in this case lies in the size of whole slide images, that can typically be more than 100,000 x 100,000 pixels. Scanning and processing these images fully can be time-consuming. For this purpose, several CAS tools include sophisticated region selection algorithms, using supervised machine learning[301] or reinforcement learning,[302] to ensure that instead of processing the whole image, only relevant areas are assessed. In the reviewed studies, algorithm processing times were often not reported, but varied depending on the algorithm and computational resources (ranging from 3 seconds[277,282] to 40-45 min[311,342] per sample). CAS systems can potentially have great impact on the time-efficiency of the scoring task, by reducing processing time per sample and processing multiple samples in parallel.

2.3.2.5 Interpretability

Improvements in accuracy, reproducibility and time-efficiency are possible when pathologists act as users of the CAS system, instead of manually assigning a score. This however implies that the pathologist can trust the CAS system enough, to take into account the information it provides. Interpretability of CAS systems is important in this context, as understanding inspires trust. This requirement can be satisfied by systems designed to produce salient features and image areas explaining the basis of decisions.[334]

2.3.2.6 Confidence estimation

When a pathologist performs the scoring task manually, they are able to indicate when they are uncertain about a sample and request a second opinion or repeat staining/ biopsy. This ability is also critical for CAS systems. CAS tools should indicate when a sample is inadequate or significantly different from any sample they have previously encountered.

To this end, scoring systems have been developed to provide a confidence estimate for their output. Vandenberghe et al.[311] developed an algorithm to measure staining heterogeneity in HER2 slides. It was demonstrated that samples with high staining

heterogeneity were the ones that pathologists disagreed on the most. This algorithm was therefore able to automatically flag up heterogeneously stained and challenging samples. In 2017, an open HER2 scoring challenge contest was organised[299] as a means to identify the best algorithms for automated HER2 scoring. Manual scoring by two pathologists was the reference gold standard in this challenge. Algorithms participating in the challenge were ranked on their ability to predict an accurate score, as well as an accurate confidence estimate c. A weighted confidence metric was devised to evaluate the accuracy of the confidence estimate c, depending on whether the algorithm predicted the reference score correctly or not:

$$w_c = \begin{cases} \frac{2c - c^2}{2} & if\ correct\ prediction \\ \frac{1 - c^2}{2} & otherwise \end{cases} \tag{1}$$

Higher value of this metric indicated improved accuracy in confidence estimates. If the algorithm predicts a correct score with high confidence c, then it is rewarded with higher w_c. However, if the algorithm predicts an incorrect score with high confidence c, then it is penalised with lower w_c. Confidence estimation is particularly important for interactive CAS systems, as it can contribute to the pathologist's trust of the automated algorithms.

2.4 Conclusions

This review examined studies validating CAS systems for HER2, ER and T-cell marker scoring. CAS validation practices were retrieved for a large number of studies, in brightfield immunohistochemistry and immunofluorescence. Studies were overall medium to high quality. They covered a variety of immunohistochemical markers, and therefore, the validation practices that were identified are generalisable to various staining patterns.

This chapter sought to clarify appropriate design requirements for CAS validation. Six key design requirements guiding the validation of CAS systems were identified; CAS tools should be well defined, provide accurate patient categorisation, be reproducible, time-efficient, interpretable and able to provide an accurate confidence estimate for their predictions. A CAS system has the potential to safely and vastly improve upon the current clinical and research scoring practices; provided that it demonstrates equivalent or superior accuracy, reproducibility and time-efficiency compared to manual scoring, while at same time being well defined, interpretable and able to indicate uncertainty.

Furthermore, reference methods to validate CAS accuracy were sought. Accuracy of CAS is defined as accurate (meaningful) patient categorisation. It is most often established through agreement with equivalent manual scoring systems. Other strategies to establish analytical validity are available when there are concerns that manual scoring is unreliable (e.g., because of low manual scoring intra/inter-observer concordance). These include using a consensus score from multiple pathologists, or orthogonal assays as the reference gold standard. Last, a CAS system could be shown to be meaningful and accurate by directly demonstrating correlation to patient clinical characteristics. If reproducibly good correlation to patient clinical endpoints can be demonstrated, establishing agreement to an equivalent manual scoring method is not necessarily required.

Finally, the performance of existing CAS systems was assessed. Quantitative meta-analysis was used to determine how well existing HER2 and ER CAS systems agree with manual scoring. Moderately good agreement with manual scoring was observed for HER2 ($\kappa = 0.75$, 95% CI: 0.70-0.81) and ER algorithms ($\kappa = 0.74$, 95% CI: 0.66-0.83). CAS agreed with manual scoring, at a similar level to how well

pathologists agreed with each other in manual scoring. Furthermore, use of interactive CAS tools by pathologists was shown to sometimes improve inter-pathologist agreement. Through this quantitative analysis, it was shown that several CAS algorithms have achieved comparable performance to manual scoring in terms of accuracy and inter-observer agreement.

The meta-analysis was limited by use of a single metric, the Cohen's κ, to assess agreement with manual scoring. Consequently, studies that did not provide enough information for calculation of this metric could not be included in the random-effects model. Nevertheless, a sufficient number of studies were included to provide an estimate for the overall performance of this technology. Meta-analysis was only performed for CAS systems in brightfield immunohistochemistry, which is commonly applied in routine clinical practice and benefits from the existence of standardised manual scoring systems. For T-cell marker studies quantitative meta-analysis was not possible. As such, the performance for each study was recorded separately.

Clarifying design requirements, validation practices and performance of CAS systems will hopefully be useful for iterative improvements in this technology. Awareness of the design requirements provides a framework to judge CAS system performance, in an objective manner. Demonstrating satisfaction of each of these requirements would be necessary to consider a CAS system good enough to be used in clinical practice.

CAS systems for cell phenotype identification in the microenvironment of OPSCC and FL will be proposed in the next two chapters. The next two chapters describe the initial biomarker discovery phase, and attempt to validate primarily CAS accuracy, by comparing against clinical survival endpoints. Further validation steps will be required in future studies, to demonstrate satisfaction of all other design requirements identified in Chapter 2, to achieve clinical translation of this technology.

2.5 Summary

This section summarises Chapter 2 by repeating key findings of the systematic review discussed in this chapter. This review identified for the first time six design requirements needed to validate CAS systems: definability, accurate patient categorisation, reproducibility, time-efficiency, interpretability (particularly for interactive systems) and accurate confidence estimation. A CAS system able to demonstrate equivalent or superior performance to manual scoring on accuracy, reproducibility and time efficiency, while simultaneously being definable, interpretable and able to provide accurate confidence estimates, would be superior to manual scoring. A meta-analysis of several HER2 and ER CAS algorithms established the performance of CAS technology in terms of agreement with manual scoring. CAS agreed with manual scoring, similar to how well pathologists agreed with each other. Furthermore, use of interactive CAS tools by pathologists was shown to sometimes improve inter-pathologist agreement. These findings underline the potential of CAS systems. Clarifying validation practices and performance of CAS systems will hopefully be useful for iterative improvements in this technology.

3 Multiplex image analysis for biomarker discovery in oropharyngeal squamous cell carcinoma

Previous chapters have laid the groundwork on how multiplexed immunofluorescence and automated image analysis could be employed for the analysis of multiple cell types and their spatial relationships in the tumour microenvironment. This chapter introduces an application of this technology for biomarker discovery in oropharyngeal squamous cell carcinoma (OPSCC). A computer assisted scoring system (CAS) is developed to define a new biomarker based on observation of spatial patterns in the tumour microenvironment.[354]

3.1 Introduction

It is recognised that a plethora of immune regulatory factors in the tumour microenvironment (TME) contribute to the progression of cancers and limit their response to treatment.[355–357] An important class of inhibitory factors, designated immune checkpoints, have been associated with long-lasting response to treatment in a variety of cancers.[358,359] Many cancers engage the immune checkpoints to abrogate the host anti-tumour immune response, leading to T cell exhaustion, loss of immune surveillance and unchecked tumour proliferation. Therapeutic immune checkpoint blockade restores immune surveillance and re-engages an anti-tumour response. Checkpoint inhibitors have revolutionised the management of many solid and haematological cancers, underlining the direct and powerful role the host immune response and TME composition plays in the prognosis of many cancers.

The programmed cell death 1 (PD-1) receptor has emerged as a dominant negative regulator of anti-tumour effector function. Interaction with its ligand PD-L1 leads to PD-1 mediated T-cell exhaustion and inhibition of antitumour cytotoxic T-cells. The latter results from specific T-cells releasing interferon gamma (IFN-γ^+) after recognising their tumour associated antigens. IFN-γ^+ release leads to upregulation of PD-L1 on the local tumour and other cells, which in turn can compromise T-cell function through adaptive immune resistance. This state of local immune privilege can be reversed by blocking antibodies to PD-1 or PD-L1 and such single agent therapies are now licensed for the treatment of patients with multiple types of cancers.[12,358–363] Response rates can be as high as 90% for some tumour types but as low as 15% with others, but selection of patients likely to respond favourably to

such single agent therapy proves a challenge, as it requires an in depth understanding of immune interactions in the TME.[358,359]

Head and neck squamous cell carcinomas (SCCHN) develop as a consequence of either a persistent high risk HPV infection or through carcinogen exposure (e.g. smoking, alcohol).[364] In the subgroup of oropharyngeal squamous cell carcinomas (OPSCC), the HPV positive patients have a significantly better clinical outcome and this is linked to differences in tumour infiltrating lymphocyte (TIL) densities.[26] PD-L1 positivity within a tumour has been explored as a potential treatment biomarker but the results have not been consistent in predicting subsequent clinical responses.[19,21,22,365,366] The spectrum of conclusions may not be surprising considering the variability of tumour aetiology, the antibodies and detection methodologies used, the arbitrary cut-off levels defined and cellular diversity of cells expressing PD-L1. Moving beyond simple enumeration of cell densities, and observing the spatial organisation of the TME may provide further insight for the development of more informative biomarkers.[226,98] In the TME of SCCHN the existence of varying patterns of PD-L1 expression has been highlighted.[31] These qualitative results are useful pointers to further analysis but are not easily generalised, as the criteria for defining patterns are subjective. Quantitative, non-subjective, assessment of spatial organisation becomes possible using automated image analysis approaches[22,367] to minimise operator dependence and analysis time, and facilitate successful clinical application.

Here an automated analysis pipeline to quantify the potential of T-cells to interact with PD-L1 expressing cells in the TME is reported, which will reflect a key driving force for immune regulation. This pipeline discards artefacts and scanning errors, performs cell segmentation and accounts for the proximity between cell subsets. Using the Hypothesised Interaction Distribution (HID) method[100], it is assessed whether a high frequency of spatial interactions between $CD8^+$ or $PD\text{-}1^+$ and $PD\text{-}L1^+$ cells correlates with a poor prognosis in OPSCC, as previously observed in HPV^- OSCC.[367]

3.2 Materials and methods

This analysis used the multiplex immunofluorescent dataset prepared by the study of Oguejiofor et al.[22] Cohort characteristics and the multiplex immunofluorescent staining protocol that was employed to generate the images are described below.

3.2.1 Cohort characteristics

The dataset for this study derived from a retrospective collection of 218 OPSCC patients treated with radiotherapy alone or with concurrent chemotherapy at The Christie NHS Foundation Trust in Manchester, UK between January 2002 and December 2011 (REC reference: 03/TG/076). HPV status of these patients was assessed (p16 expression, in-situ hybridisation and human papillomavirus DNA PCR) as described elsewhere.[26] Within this cohort, 124 patients with concordant HPV status for all three assays had sufficient formalin fixed, paraffin embedded tissue available for multiplex immunofluorescence staining with antibodies against PD-L1, CD8, CD68 and PD-1.[22] Analyses were performed on randomly selected regions of interest (ROIs) from sections taken from pre-treatment diagnostic biopsies of OPSCC. The associated clinical data for grade, stage and comorbidities (alcohol and smoking) is described in Oguejiofor et al.[22] For the present analysis, updated overall survival (OS) information was obtained for 72 patients.

3.2.2 Ethics approval and consent to participate

The study was approved by the Tameside & Glossop Local Research Ethics Committee subsequently renamed the National Research Ethics Service Committee North West – Greater Manchester East (REC reference: 03/TG/076). Patients were not required to provide written consent as approved by the ethics committee due to researchers working on anonymised data. The study was performed according to the Declaration of Helsinki.

3.2.3 Multiplex staining and multispectral scanning

Table 14 Antibodies, titrations and fluorophores in the multiplex immune-fluorescent experiment

Order	Antibody	Dilution	Provider	Opal detection
1	Rabbit monoclonal against PD-L1	1:200	Cell Signalling, US	Cyanine 5.5
2	Mouse monoclonal against CD8 (clone C8/144B)	1:60	DAKO, Denmark	Cyanine 3
3	Mouse monoclonal against PD-1	1:50	Abcam, UK	Fluorescein
4	Mouse monoclonal against CD68	1:200	Abcam, UK	Cyanine 3.5
The order presented reflects the order in which the antibodies were placed on the tissue.				

Multiplex immunofluorescent staining was performed using the Ventana auto-staining platform (Ventana Medical Systems, Oro Valley, Arizona, United States) and the Opal detection system (PerkinElmer, Waltham, Massachusetts, United States) with tyramide signal amplification, as described elsewhere[22] and summarised in **Table 14**. Using TSA[148] and the Opal kit technology permits multiple repeated cycles of staining and stripping of anti-mouse or anti-rabbit antibodies, while the TSA conjugated fluorophores bind strongly to the epitopes and remain on the tissue. The auto-staining platform performed an initial deparaffinisation and epitope retrieval at pH 8.5. Subsequent staining cycles involved incubation with the primary antibody, the secondary antibody, and then the opal detection label. Each staining cycle was separated by a short denaturation at pH 6. After staining, slides were washed with EZ preparation (1:10) for 3 cycles of 5 min each and cover-slipped using the Prolong aqueous mounting agent (Thermo Fisher, Waltham, Massachusetts, United States) with DAPI for counter-staining. Imaging was performed using a Vectra microscope (PerkinElmer) and a 20x objective (0.495 µm per pixel). The Vectra microscope first scanned whole slides at low resolution to obtain the tissue grid using only the DAPI filter. Subsequently, 10 to 20 ROIs (1392×1040 pixels) were selected randomly for each slide from tissue areas for multispectral scanning at full resolution using all available filters (DAPI, FITC, Cy3, Texas Red and Cy 5).

3.2.4 Spectral unmixing

Linear spectral unmixing[134] was performed using the inForm software (PerkinElmer). For unmixing a spectral library was built comprising individual fluorophore spectra. Each spectrum was acquired from slides that were single stained for the different antibodies, using the same experimental parameters as in the multiplex experiment. A slide stained only with DAPI was also used to extract the DAPI spectrum. Finally, a slide that underwent all steps in the multiplex experiment without application of antibodies or fluorophores was used to extract the spectrum of tissue auto-fluorescence (AF). After spectral unmixing, the images had 6 channels ($1392 \times 1040 \times 6$ pixels), each containing the intensities of a different fluorophore (see **Figure 19**).

3.2.5 Deep learning for automated image quality check

After spectral unmixing, a quality assessment of images was needed to verify that only relevant areas of tissue were included in subsequent analyses. To discard artefacts and select areas of tissue suitable for analysis, supervised tissue segmentation using Support Vector Machines or CNN has previously been employed successfully.[368,369] It is shown that the CNN approach can also be used with immunofluorescence, where apart from blurring and artefacts, high auto-fluorescence in blood vessels and red blood cells cause problems. Immunofluorescence image artefacts include: bubbles created during cover-slipping; tissue folding; blurriness due to scanning errors; the presence of blood vessels with brightly auto-fluorescing red blood cells; and the presence of fatty tissue. Digital pathology datasets tend to be large, making manual checking of images to identify and exclude problematic areas slow and labour intensive.

To automate this essential pre-processing step, a deep CNN classifier was trained on a set of 3280 manually annotated image patches of size $128 \times 128 \times 6$ to discriminate at pixel level between problematic areas, useful tissue, or background. The image undergoes a series of transformations as it passes through the layers of the network and a predicted output label is generated for each pixel. This output is compared to the ground truth and the parameters of the network are updated during training to decrease the error. A variant of the U-Net network architecture,[191] popular in biomedical applications, was used as detailed in **Figure 20**.

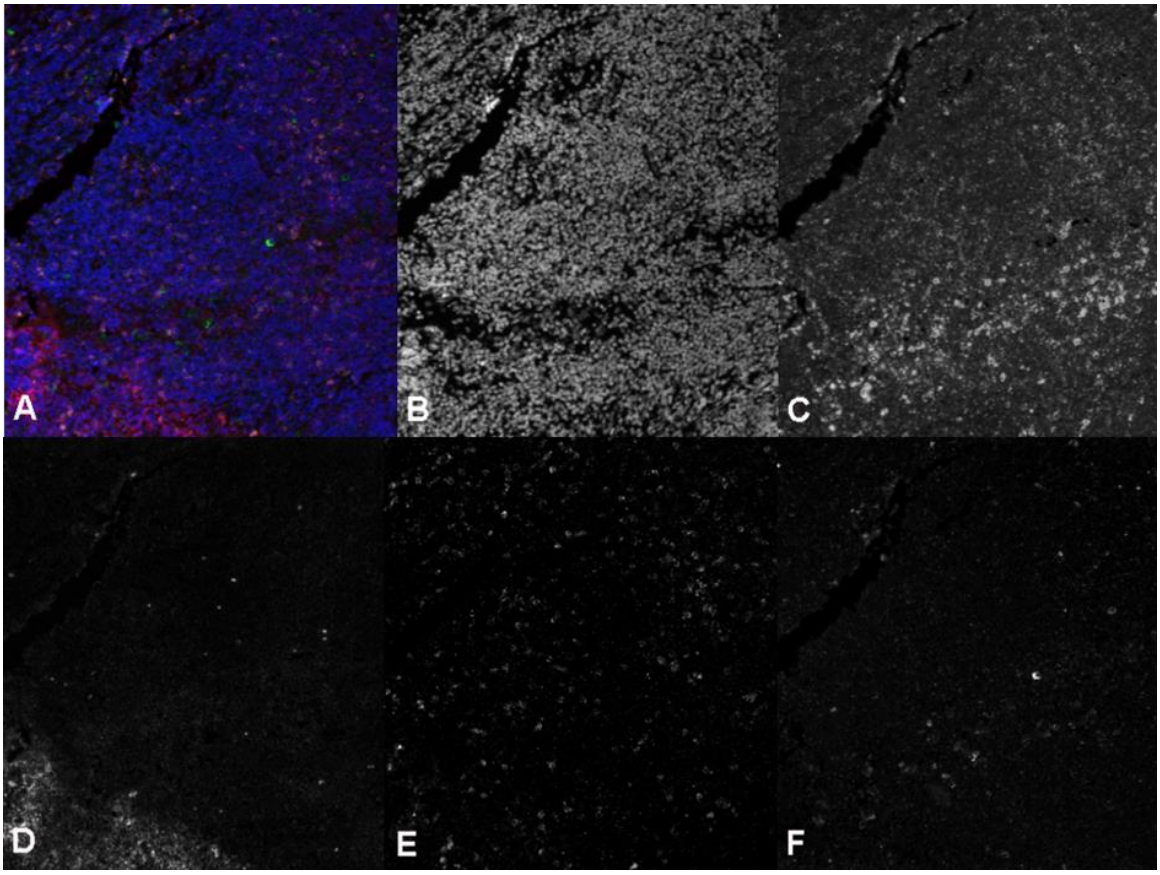

Figure 19 Example of an image in the data set, representing a single region of interest (1040 ×1392 pixels). A: Composite view where DAPI, CD8, CD68, PD-L1 and PD-1 were mapped to blue, green, yellow, red and magenta, respectively; B: DAPI nuclear counter-stain; C: PD-1; D: PD-L1; E: CD68 and F: CD8. The image also includes a channel with tissue auto-fluorescence (AF) signal - not shown here. AF has already been subtracted from all other channels during spectral unmixing.

Ground truth annotations were drawn using the open-source software QuPath and exported automatically using custom Groovy scripts, as this platform allows the user to easily implement their own algorithms to supplement the functionality of the main interface. This implementation allowed to keep the full resolution of the image and assign an integer label to each pixel based on the drawn annotation (0: background or fatty tissue, 1: useful tissue, 2: artefacts). The annotations were created by observing all the stains separately and also the composite image. The ROI images were tiled into patches of size $128 \times 128 \times 6$, run through the network and a map of the predicted pixel labels was generated for each patch. Example artefacts and corresponding segmentation results are shown in **Figure 21**. The predicted pixel labels from all patches were merged to create the map of the entire ROI. Tiling greatly improved the memory requirements for the task.

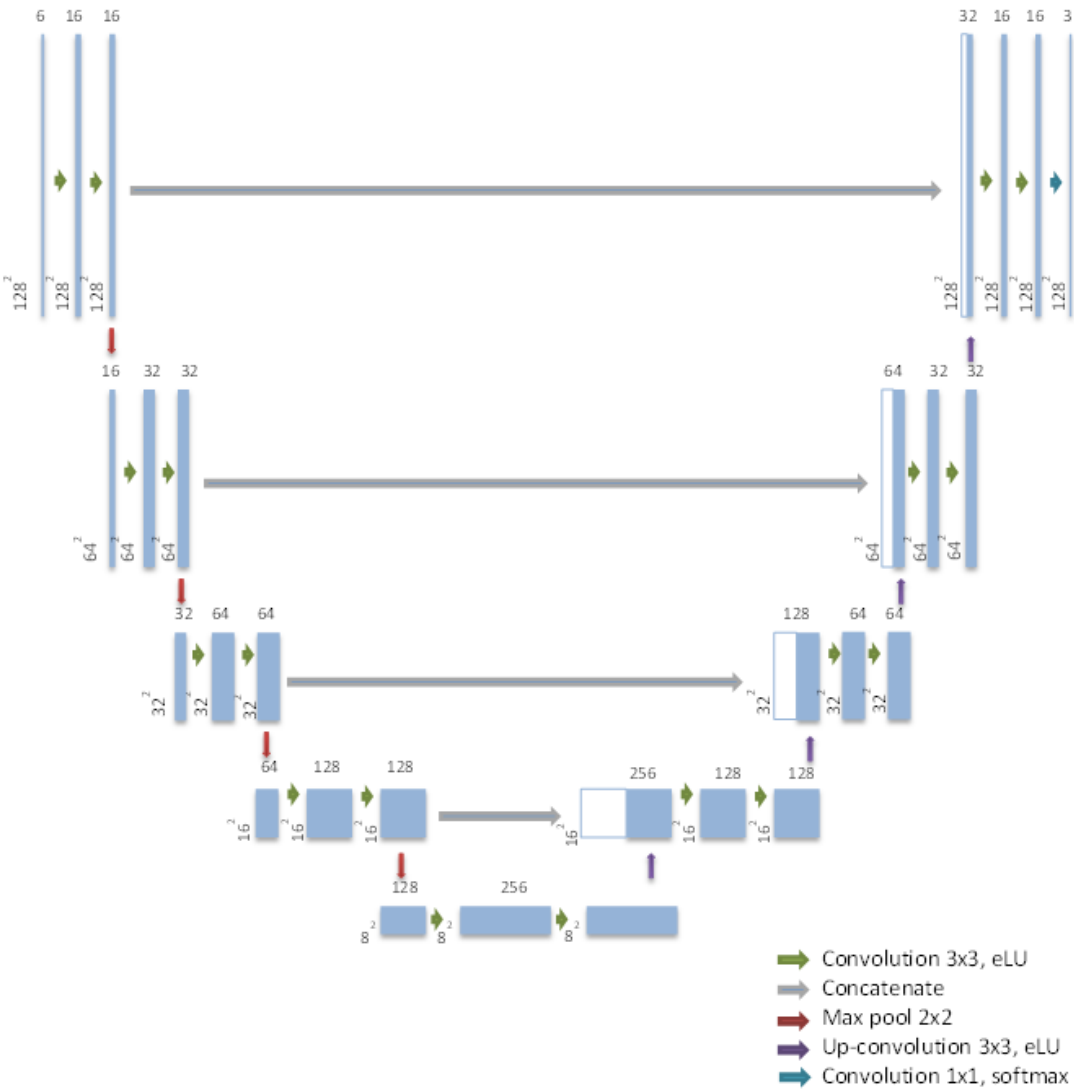

Figure 20 The architecture of the U-net segmentation model

Keras with a Tensforflow backend was used for the implementation and training was performed for 200 epochs and a batch size of 20. The Adaptive Moment Estimation (Adam) optimiser was used with the categorical cross-entropy loss and a learning rate parameter of 0.0005. All image channels were included during training. A separate validation set of 960 images (size 128×128×6) was used to tune empirically the hyper parameters of the network by observing the validation loss. The pixel-wise predicted accuracy on the validation set was 92.9%.

A test set of 640 images were used to assess performance. Pixel-wise accuracy was 88.3% when compared with manual annotations (**Figure 21** and **Table 15**).

Table 15 Normalised confusion matrix for the network predictions on the test set

U-net CNN Confusion Matrix

True label	Tissue	Artefact	Background
Tissue	0.92	0.06	0.02
Artefact	0.22	0.75	0.03
Background	0.01	0.10	0.89

Predicted label

This analysis compared the performance of the automated artefact finder using U-net to the image segmentation algorithm supplemented in *inForm 2.4* (Akoya Bio-sciences) software. For that purpose, the same set of manually drawn full image annotations previously used to train U-net, was also used to train a tissue segmentation module in *inForm*. The training was allowed to continue, until the training accuracy stabilised at 91%, while the segmentation resolution was set to coarse. All fluorophore channels and auto-fluorescence were used during training. To assess performance, the images of the test set were then processed using the trained *inForm* algorithm, resulting in pixel-wise accuracy of 81.2%, which was markedly worse compared to the 88.3% accuracy obtained by the U-net classifier for the same test set.

Therefore, the CNN was applied to remove artefacts and background. If there was <30% useful tissue identified by the CNN the ROI was excluded from subsequent analysis. After pre-processing, the dataset was reduced to 1620 images ($1392 \times 1040 \times 6$ pixels), and only cells with centroids located within the useful tissue areas were considered for subsequent analyses.

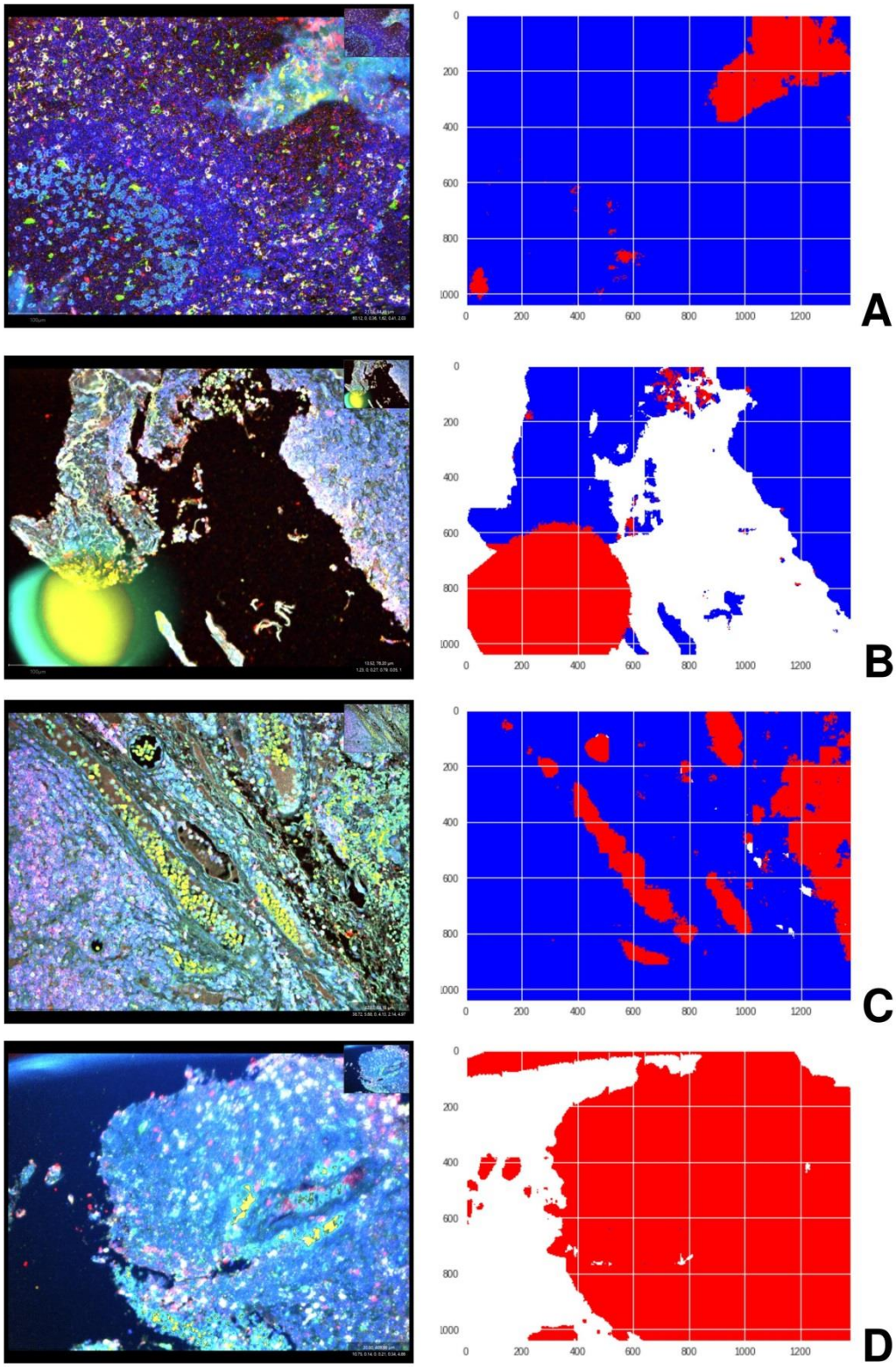

Figure 21 Problematic areas and predicted segmentation labels from the test set. A. Area with tissue fold (upper right corner). B. Bubble due to poor cover-slipping. C. Blood vessels filled with red blood cells. D. Whole region scanned out of focus. The colour-map of predicted segmentation masks (images on the right) as follows; red: problematic area, blue: normal tissue, white: background.

3.2.6 Cell segmentation and scoring

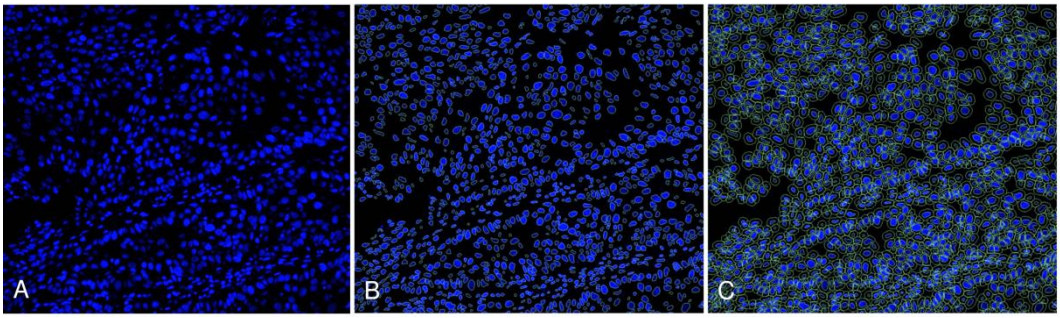

Figure 22 Cell segmentation was carried out in QuPath. A. DAPI channel view; B. Nucleus segmentation; C. Cytoplasm simulation by nucleus expansion. A Groovy batch script was written using QuPath's interface to segment the cells in all images and export the cell data.

Cell segmentation was carried out using the open source digital pathology software QuPath v0.1.3[234] (**Figure 22**). Nuclear detection was performed on the DAPI channel using an unsupervised watershed algorithm with parameters tuned on a validation set of 10 ROI. To assess the nuclear segmentation accuracy, a manually annotated independent test set of 5 ROI, containing 956 nuclei in total, was constructed from the head and neck data set. The performance of unsupervised nuclear segmentation in inForm 2.4 was compared to the unsupervised nuclear segmentation using a watershed algorithm in QuPath 0.1.3 open-source software. QuPath was used to generate and export the manual segmentation annotations as labelled integer masks by using custom Groovy scripting.

To compare these two algorithms the approach in Schmidt et al.[175] was adopted and the average precision $AP = \frac{TP}{TP+FP+FN}$ was calculated, where true positive (TP) predictions are defined as predicted nuclei which sufficiently overlap annotated ground truth nuclei. Overlap was measured as the intersection over union (IoU) between predicted and ground truth cells, but as "sufficient" overlap can be tricky to define, the performance assessment was repeated for various thresholds of the IoU metric. False positives (FP) were defined as the predicted nuclei with no corresponding ground truth nuclei, while false negative (FN) were the ground truth nuclei with no corresponding predicted nuclei.

Many different settings for algorithm implementation in inForm were qualitatively tested through trial and error, and the optimal settings were selected, as presented

in **Table 16**. The average precision results, calculated on a cell-wise basis, are presented in **Table 17.**

Table 16 Nuclear segmentation settings for inForm 2.4 and QuPath 0.1.3

Settings for inForm 2.4	
Algorithm	Adaptive cell segmentation
Component	DAPI
Relative intensity threshold	0.1
Nuclear staining quality	Mixed
Nuclear splitting settings threshold	0.442
Minimum nuclear size (pixels)	60
Fill nuclear holes smaller than (pixels)	50
Refining cell after segmentation	FALSE
Settings for QuPath 0.1.3	
Algorithm	Watershed cell detection
Component	DAPI
Requested pixel size	0.5
Background radius	8
Median filter radius	0.8
Sigma	1.2
Minimum area (μm^2)	5
Maximum area (μm^2)	200
Intensity threshold	2
Split by shape	TRUE
Smooth boundaries	TRUE

Table 17 Segmentation performance in the manually annotated test set

Intersection over union (IoU) threshold	10%	20%	30%	40%
AP inForm 2.4	0.845	0.755	0.65	0.495
AP QuPath 0.1.3	0.865	0.762	0.634	0.51
AP is the average precision for different values of the intersection over union threshold (IoU). As the IoU threshold increases the definition of a true positive cell becomes stricter (i.e., the shape of the predicted cell must match more closely the ground truth).				

While both software packages produce similar results as seen in **Figure 23**, QuPath was selected for this study as it is open-source software, with well-maintained documentation, version management and an active supportive community. Furthermore, it offers built-in capability of custom scripting, which facilitates quantitative validation of its algorithms' performance.

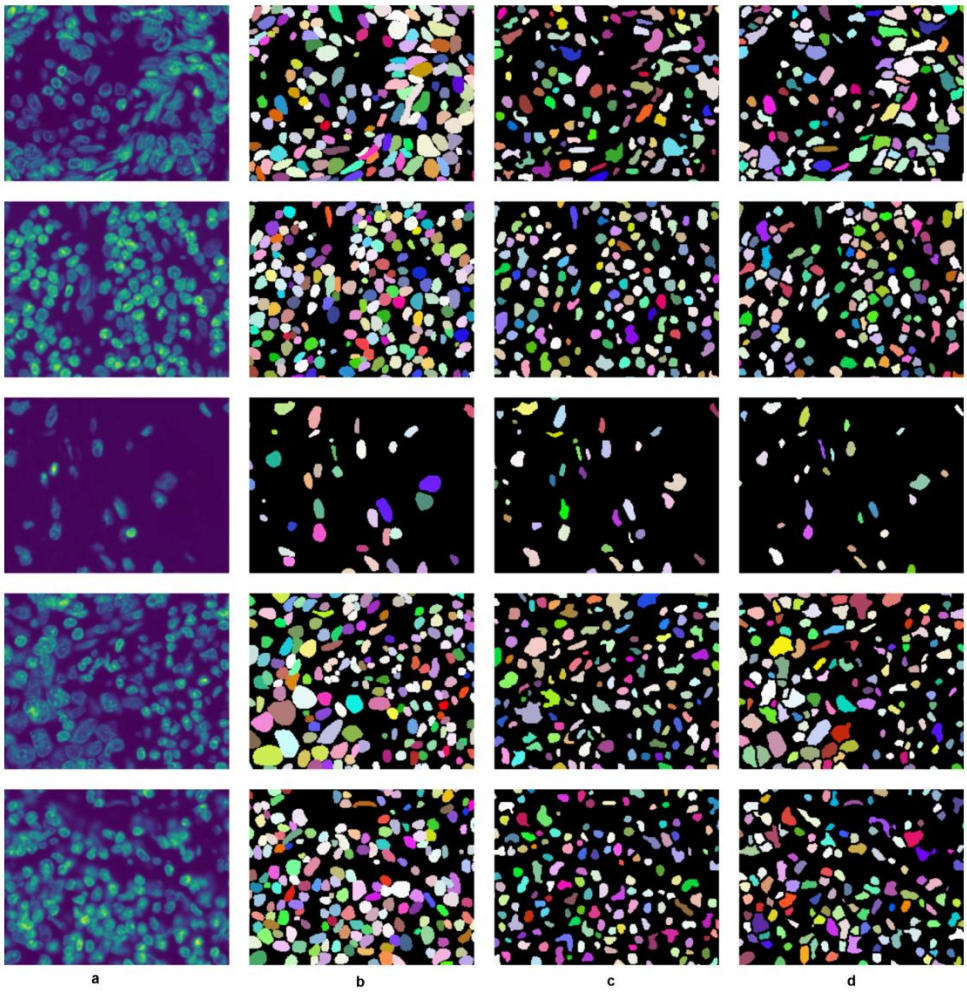

Figure 23 Nuclear segmentation comparison between inForm 2.4 and QuPath 0.1.3. a) DAPI component b) manual annotations c) QuPath segmentation results d) inForm segmentation results.

After nuclear detection, the cytoplasm around each nucleus was simulated by cell expansion of 2 µm and measurements generated for marker intensity in different compartments (mean, minimum, maximum and standard deviation of intensity in cytoplasm or nucleus). Details of this procedure are shown in **Figure 22.**

The intensity of each marker in the primary cell compartment where it is usually expressed was observed to determine whether a cell was positive for this marker. In this study, markers were cytoplasmic or membranous. Before cell scoring, the intensity of each marker was re-scaled onto a grey-scale colour map, with the brightest and darkest values corresponding to the 99% and 1% percentiles of the marker's pixel intensities in the entire data set. Having a consistent colour-map per marker ensured that the same intensity value was represented with equal brightness in all images.

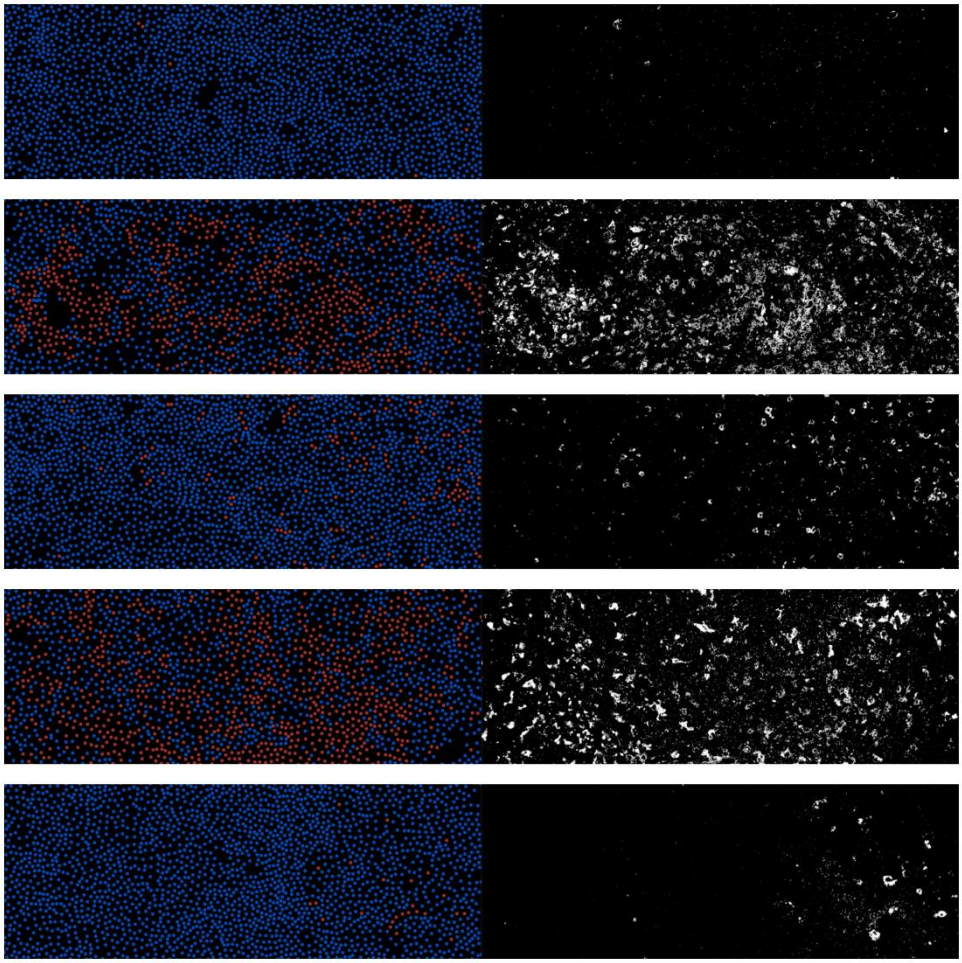

Figure 24 Results of scoring for five regions of interest (ROI) from different slides and patients for the CD8 marker. These images were part of the set used when selecting the thresholds. The colour map of the grey scale images is scaled to range between the minimum and maximum intensity value for CD8 in the entire data set.

Guided by a pathologist (R.B), a single threshold for each marker was selected as a cut-off to determine positivity across the entire data set. The threshold was identified by its ability to separate positive from negative cells in a set of 20 ROIs from 20 different patients (**Figure 24**). This cell scoring method was chosen for its simplicity but provided a non-optimal separation in some samples, possibly due to slight variations in fixation, staining, scanning or cell segmentation performance. For subsequent analysis these small variations were ignored, however their presence remains a challenge to overcome in order to improve the accuracy and robustness of the automated analysis pipeline.

3.2.7 Proximity analysis

To quantify the proximity relationships between cell phenotypes the Hypothesised Interaction Distribution (HID) analysis[100] was applied. For a pair of cell phenotypes i, j the HID is calculated to quantify how often these phenotypes occur close to each other in a sample. Let k, l be cells of phenotype i, j respectively. Then HID is computed as follows:

$$H(i,j) = \left|\left\{\left\{\boldsymbol{x}_i^k \in C^i, \boldsymbol{x}_j^l \in C^j\right\} \forall\, k, l,\ \boldsymbol{x}_i^k \neq \boldsymbol{x}_j^l\ s.t. \|\boldsymbol{x}_i^k - \boldsymbol{x}_j^l\|_2 < d\right\}\right| \quad (2)$$

where $\boldsymbol{x}$ represents the position of the centroid of a cell and d is the distance parameter that defines closeness. To construct HID, the algorithm iteratively examines the neighbourhood within a distance d around each cell of phenotype i and counts the number of occurrences of cells of phenotype j within that same neighbourhood. The distance parameter d is problem specific, as the size of the neighbourhood of interest depends on the type of cells, their mobility and mode of interaction (e.g., directly by contact or indirectly through secretion of cytokines).

The HID measure was normalised using the total number (N) of all cells, regardless of phenotype, in samples, as follows:

$$h(i,j) = \frac{H(i,j)}{N}) \quad (3)$$

The complete image analysis pipeline is presented in **Figure 25**.

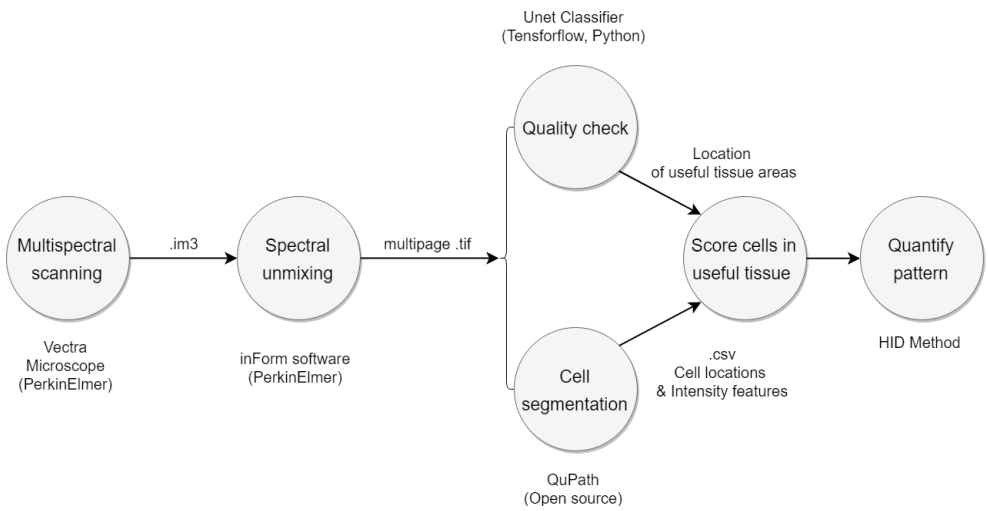

Figure 25 Diagram of image analysis pipeline

3.2.8 Statistical analysis

Kaplan-Meier and Proportional Hazards Cox Regression survival analyses for right censored data were performed using the Lifelines 0.18.1 library in Python. Statistical significance of differences between Kaplan-Meier curves was assessed using the Mantel-Haenszel log rank test. The variance of the Kaplan-Meier estimator plotted as error bars in the figures was derived using Greenwood's formula.[370] For comparisons of cell distributions between HPV positive and negative subgroups the Mann-Whitney one-sided U-test for unpaired data was used. This non-parametric test was selected as the observations did not satisfy the Kolmogorov-Smirnov (K-S) test of normality ($p < 0.005$). Significance is considered at a level $\alpha = 0.05$.

3.3 Results

3.3.1 Smoking and HPV status predict overall survival

The 72-patient cohort analysed in the current study had a minimum follow up of 7.1 years for the patients who were alive at the time at the time of data collection, and 43 observed events (40% censored data). The median overall survival (OS) of the 72 patients, observed and censored, was 86.8 months. Clinical data for HPV status, stage, alcohol consumption and smoking for the 72 patients are summarised in **Table 18**.

Table 18 Cohort characteristics

Characteristic		All	HPV Positive	HPV Negative
Patients [Events Observed OS]		72 [43]	41 [18]	31 [25]
Gender	Female	12	7	5
	Male	60	34	26
Age (years)	Median	58	56	59.5
AJCC Stage	I	0	0	0
	II	2	1	1
	III	10	4	6
	IV	23	14	9
	No data	37	22	15
Grade	Well differentiated	10	9	1
	Moderately differentiated	36	13	23
	Poorly differentiated	19	14	5
	No data	7	5	2
Alcohol	Never	8	5	3
	Moderate	35	19	16
	Excessive	27	16	11
	No data	2	1	1
Smoking	Never	14	12	2
	Ex-smoker	29	16	13
	Current smoker	27	12	15
	No data	2	1	1

Table 19 lists the findings from a univariable Cox regression analysis. As expected, negative HPV status was highly prognostic for poor OS (hazard ratio [HR] 3.30; $95\%\ CI\ 1.77 - 6.15$; $p = 0.0002$. Smoking also correlated with a worse outcome ($HR = 1.91, 95\%\ CI\ 1.05 - 3.48, p = 0.034$).

Table 19 Cox regression survival analysis (univariable) for clinical variables

	HR (CI 95%)	P value
HPV status	3.295 (1.767, 6.145)	*0.0002*
AJCC stage	0.790 (0.396, 1.579)	0.5052
Alcohol	0.982 (0.760, 1.269)	0.8896
Smoking	1.911 (1.049, 3.481)	*0.0342*
Grade	0.979 (0.636, 1.507)	0.9226
Alcohol was assessed as 0: never, 1: moderate and 2: excessive, while smoking was assessed as 0: never or ex-smoker and 1: currently smoking. Grade was given as 0: well differentiated, 1: moderately and 2: poorly and stage as I-IV according to staging criteria set by the American Joint Committee on Cancer (AJCC).		

3.3.2 Distribution and prognostic value of cell population densities

Table 20 Median population density expressed as a percentage of positive cells

Cell type	**All**	**HPV positive**	**HPV negative**	
T-cells	**Median Percentage of Positive Cells**			**P value***
$CD8^+$	6.60%	8.90%	4.90%	*0.034*
$CD8^+PD\text{-}1^+$	1.50%	1.70%	1.10%	0.111
Macrophages				
$CD68^+$	3.10%	6.00%	2.10%	*0.035*
$CD68^+PD\text{-}L1^+$	1.10%	2.20%	0.70%	0.058
PD-L1 and PD-1				
$PD\text{-}L1^+$	9.00%	9.00%	7.40%	0.356
$PD\text{-}1^+$	12.70%	13.50%	10.90%	0.284
Percentage cell expression was first assessed for individual ROIs, and the median expression from all ROI was selected to represent the patient.*P-value tests the difference between HPV positive and negative groups.				

Table 20 summarises the percent median cell expression of various cell phenotypes in the patient cohort of OPSCC. $CD8^+$ T-cells and $CD68^+$ macrophages were found in significantly greater numbers in HPV^+ OPSCC tumours. The $PD\text{-}1^+$ phenotype

outnumbered CD8$^+$ T-cells, which could be explained by PD-1 expression in different T-cell subsets, such as CD4$^+$ cells. Additionally, the PD-L1$^+$ category outnumbered CD68$^+$ macrophages, as PD-L1 expression is expected in immune related, as well as tumour cells. These marked populations did not significantly differ when stratified by HPV status. An up-to-date survival analysis using the median percent marker expression to define high and low expression levels is shown in **Table 21**. In this study only an increased detection of CD68$^+$ cells (macrophages) was significantly associated with improved outcome in the HPV negative patients. It is not possible to compare these results with those published previously as the current analysis did not distinguish the stromal versus tumour locations and a different methodological approach was applied to select ROIs, detect the cells, identify positives and report densities by normalising with total number of cells.[26,22]

Table 21 Univariable Cox Regression analysis of overall survival for patients stratified by median cell expression

	HPV Positive		**HPV Negative**		**All**	
Cell population	**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**
T-cells						
CD8$^+$	0.56 (0.21, 1.50)	0.25	1.03 (0.47, 2.26)	0.94	0.84 (0.46, 1.54)	0.57
CD8$^+$PD-1$^+$	0.76 (0.29, 1.99)	0.57	1.88 (0.85, 4.16)	0.12	1.16 (0.63, 2.14)	0.63
Macrophages						
CD68$^+$	1.34 (0.51, 3.53)	0.55	0.34 (0.14, 0.79)	*0.01*	0.58 (0.32, 1.07)	0.08
CD68$^+$PD-L1$^+$	1.33 (0.51, 3.45)	0.56	1.50 (0.67, 3.34)	0.32	1.34 (0.73, 2.47)	0.34
PD-L1 & PD-1						
PD-L1$^+$	1.36 (0.52, 3.52)	0.53	1.50 (0.67, 3.34)	0.32	1.42 (0.77, 2.60)	0.26
PD-1$^+$	0.60 (0.22, 1.60)	0.31	1.84 (0.83, 4.07)	0.14	1.06 (0.57, 1.94)	0.86

Variables stratified by the median to distinguish patients with high and low expression. Percentage cell expression was first assessed for individual ROIs, and the median expression from all ROI was selected to represent the patient.

3.3.3 Proximity analyses of T-cells with PD-L1$^+$ cells

Figure 26 illustrates an example of the methodology used to generate the HID measure reflecting potential cell interactions. An interaction is hypothesized to occur whenever a CD8$^+$ cell (yellow) occurs within 30 μm of a PD-L1$^+$ cell (pink). A connection is drawn (white) to represent each hypothesized interaction. Cells not expressing CD8 or PD-L1 are presented in blue.

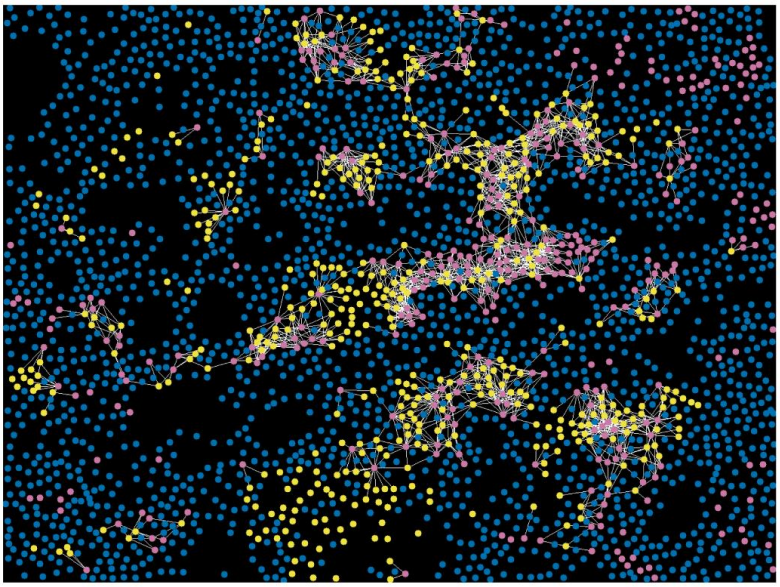

Figure 26 Illustrative HID interaction features for a region of interest. An interaction is hypothesized to occur whenever a CD8+ cell (yellow) occurs within 30 μm of a PD-L1+ cell (pink). A connection is drawn (white) to represent each hypothesized interaction. Cells not expressing CD8 or PD-L1 are presented in blue.

A pre-specified HID analysis was carried out for two pairs of interacting phenotypes co-localised within 30 μm of each other (CD8$^+$ and PD-L1$^+$ cells; PD-1$^+$ and PD-L1$^+$ cells). This distance was used by Feng et al[19] and represents a neighbourhood size of 2-3 cells. The mean +/- standard error of HID values are shown in **Table 22**. There was a larger number of CD8/PD-L1 and PD-1/PD-L1 proximal events in the HPV positive tumours, but the difference was not statistically significant. A univariable Cox regression analysis stratified patients by high versus low levels of co-localisation (percent mean). More frequent interactions between CD8$^+$ and PD-L1$^+$ or PD-1$^+$ and PD-L1$^+$ cells were prognostic for poor overall survival in

HPV$^-$ but not HPV$^+$ patients or the whole cohort (**Table 23, Figure 27**). When stratifying the HPV$^-$ patients by the mean value of PD-1$^+$ and PD-L1$^+$ HID interactions, 30% of the patients were assigned to the poor prognostic group. When grouping by CD8$^+$ and PD-L1$^+$ interactions 23% of patients were assigned to the poor prognostic group (**Figure 27** A, B).

Table 22 Distribution of HID features in all, HPV positive and HPV negative patients

Cell interactions	**All**	**HPV positive**	**HPV negative**	**P value***
CD8$^+$ within 30 µm of PD-L1$^+$	27.65 (± 6.86)	34.73 (± 10.68)	17.73 (± 6.69)	0.276
PD-1$^+$ within 30 µm of PD-L1$^+$	15.76 (± 7.32)	23.48 (± 12.41)	4.95 (± 1.72)	0.535
Data presented as mean HID · 10^3 (± standard error of the mean). *P value tests the difference between HPV positive and negative groups.				

Table 23 Univariable Cox Regression analysis of overall survival for patients stratified by mean HID proximity frequencies

	HPV Positive		**HPV Negative**		**All**	
Cell Interactions	**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**
CD8$^+$ within 30 µm of PD-L1$^+$	0.82 (0.26, 2.50)	0.73	2.95 (1.15, 7.56)	*0.02*	1.15 (0.58, 2.30)	0.68
PD-1$^+$ within 30 µm of PD-L1$^+$	0.59 (0.17, 2.06)	0.41	2.64 (1.04, 6.71)	*0.04*	1.15 (0.58, 2.29)	0.69

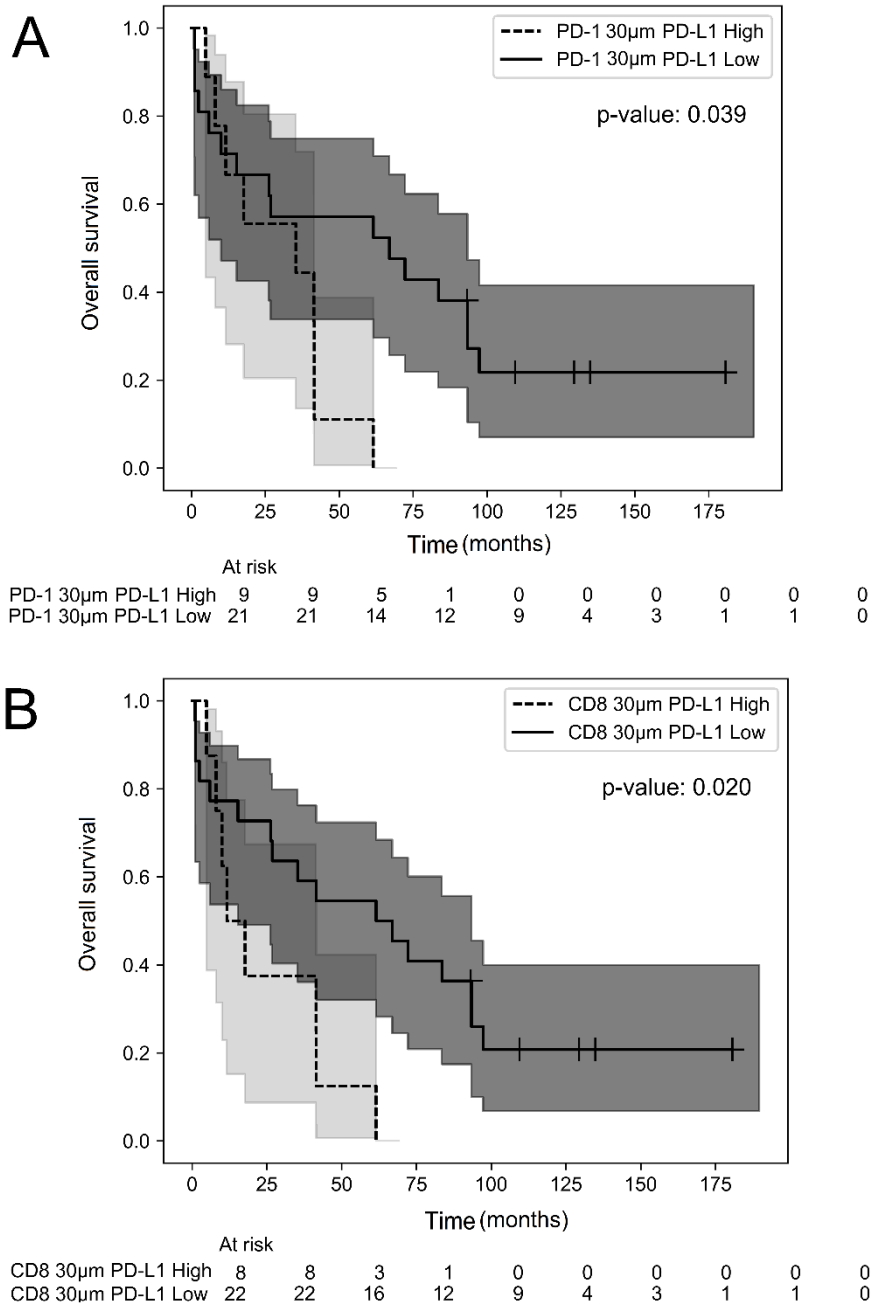

Figure 27 Kaplan-Meier analysis of the effect of HID interactions on prognosis in the HPV negative subgroup. Significance is considered using the log rank test. High and low co-localisations are considered by splitting the patients at the mean value. A: Interactions between PD-L1$^+$ and PD-1$^+$ cells. B: Interactions between PD-L1$^+$ and CD8$^+$ cells.

3.4 Discussion

This study introduces an automated pipeline for analysis of different biomarkers in the tumour microenvironment. In comparison with other automated image analysis studies, this pipeline used an automated quality check of scanned images and ROI selection prior to quantification of spatial interaction features. Checking image quality is a time consuming but essential part of any histopathological analysis. Blurred areas and artefacts (e.g., bubbles, tissue folds, presence of fatty tissue) lead to processing errors and consequently the samples are sent back for re-staining and scanning, increasing the time required for analysis. This automated selection of good quality ROIs decreases the need for input from a pathologist. A key component of this study is the use of HID methodology which can be used to assess the spatial relations (proximity) between particular cell phenotypes.[100] It has previously been used by us to analyse T-cell regulatory patterns in follicular lymphoma.[97,371]

This study provides novel evidence that the frequent proximity of PD-1^{+} and PD-L1^{+} cells is an adverse prognostic factor in HPV^{-} OPSCC. It is tempting to speculate this derives from the functional consequence of these interactions in the PD-1/PD-L1 pathway of immune escape. If the latter is correct, then quantifying the frequency of proximal cell-cell interactions using HID should be further explored as a secondary companion diagnostic potentially useful in directing checkpoint inhibitor treatment. Monitoring levels of PD-L1 expression alone, while biologically plausible, has shown inconsistent results, particularly in cases where expression levels are close to the cut-off threshold.[365] Interestingly, in this analysis no correlation between PD-L1 expression and overall survival was observed, regardless of HPV status for OPSCC. This result agrees with the observations from other studies.[21,367] However, previous analyses of the same cohort[22] demonstrated that PD-L1 expression was prognostic in HPV negative OPSCC but only if assessed in the stromal regions with a cut-off of 5%. The optimal manner of scoring PD-L1 is still being investigated, as the cut-off thresholds differ in lung, urothelial and head and neck cancer. Indeed, opinions differ on whether positivity should be assessed only for tumour cells or additionally for immune infiltrating cells.[372] An automated process to quantifying cell patterns promotes consistency and reproducibility and could facilitate its use to support the role of PD-L1 in personalised treatment strategies.

Interestingly, the correlation between overall survival and HID spatial interactions in the HPV positive subgroup was not significant. If this is a true effect, it would indicate reduced importance of T and PD-L1$^+$ cell interactions for the HPV$^+$ subgroup. This finding is not surprising as HPV related OPSCC is considered in many aspects different from HPV$^-$ OPSCC and is known to have a better prognosis,[373] more active anti-tumour immune response[26] and favourable response to treatment.[373] However, the nature of PD-L1$^+$ spatial interactions in HPV$^+$ OPSCC merits further investigation in larger cohorts, before their significance could be ruled out.

Due to the size of the cohort the power of the study is limited which increases the risk of false negative results. To avoid multiple testing, only two pre-determined hypotheses using HID in relation to overall survival were explored. Another possible limitation is the image analysis pipeline, which involved a single pathologist identifying positive cells by selecting a cut-off for each marker based on selected images with clear positive staining. However, variation was observed between the intensities of positive cells in different sections, which a simple on-off scoring approach cannot capture. Accuracy in cell phenotyping could be more reliable if it was carried out using cut-offs selected by multiple pathologists, or unsupervised machine learning for single-cell classification.

The automated pipeline developed in this chapter can be viewed as an interactive CAS system that enables accurate patient categorisation into risk stratified groups. The system is interactive, as a pathologist is required to set the stain intensity cut-off to identify positive cells. To ensure this CAS system's ability to translate into routine clinical practice, thorough validation would be necessary following the six design requirements identified in Chapter 2. This CAS system is well defined and interpretable as it relies on easily explainable features (i.e., the numbers of cells and cellular interactions). Furthermore, the accuracy of patient categorisation was validated against patient overall survival. Further work would be needed to establish reproducibility under variable staining conditions, time-efficiency and a methodology to identify samples with high uncertainty.

In summary, this study combined multiplex immunofluorescence and multispectral microscopy with an automated analysis pipeline for quality checking, spectral unmixing, cell segmentation, scoring and assessment of the spatial pattern of cell-cell interactions. In a cohort of OPSCC patients it was shown that frequent proximity of

$CD8^+$ or PD-1^+ and PD-$L1^+$ cells was prognostic for OS in patients with HPV^- tumours. This method is ready to be tested independently in additional, multicentre cohorts to validate its potential as a companion diagnostic for therapies targeting the PD-1/ PD-L1 pathway of immune escape.

3.5 Summary

Fulfilling the promise of cancer immunotherapy requires novel predictive biomarkers to characterise the host immune microenvironment. Deciphering the complexity of immune cell interactions requires an automated multiplex approach to histological analysis of tumour sections. A new automatic approach was tested to select tissue and quantify the frequencies of cell-cell spatial interactions occurring in the PD-1/PD-L1 pathway, hypothesised to reflect immune escape in oropharyngeal squamous cell carcinoma (OPSCC).

Single sections of diagnostic biopsies from 72 OPSCC patients, stained using multiplex immunofluorescence (CD8, PD-1, PD-L1, CD68) were retrieved from the study of Oguejiofor et al.[22]. Following multispectral scanning and automated regions-of-interest selection, the Hypothesised Interaction Distribution (HID) method quantified spatial proximity between cells. Method applicability was tested by investigating the prognostic significance of co-localised cells (within 30 μm) in patients stratified by HPV status.

High frequencies of proximal $CD8^+$ and $PD\text{-}L1^+$ (HR 2.95, $p = 0.025$) and $PD\text{-}1^+$ and $PD\text{-}L1^+$ (HR 2.64, $p = 0.042$) cells were prognostic for poor overall survival in patients with HPV negative OPSCC ($n = 31$).

The HID method can quantify spatial interactions considered to reflect immune escape and generate prognostic information in OPSCC. The new automated approach is ready to test in additional cohorts and its applicability should be explored in research and clinical studies.

4 Multiplex image analysis for biomarker discovery in follicular lymphoma

The spatial biomarker described in Chapter 3 for OPSCC observed cellular interactions taking place in PD-1/ PD-L1 pathway of immune escape. The presence and significance of this pathway has not been proven conclusively in FL,[101] therefore the biomarker developed in Chapter 3 was not considered for FL. Instead, in this chapter, the potential of observing the overall immune diversity in the tumour microenvironment of FL was investigated as an alternative approach to derive prognostic insights.

In follicular lymphoma (FL) there is a clinical need for upfront risk stratification, recognising the subset of patients (15-30%) with early progression of disease after first line therapy in need of more effective treatment to overcome poor outcomes ,[36,374,375] alongside the majority of patients with a more favourable prognosis who may do just as well with less therapy. Current FL prognostic indices such as the Follicular Lymphoma International Prognostic Index (FLIPI)[49,51] are well validated but lack the necessary precision for clinical decision making. The tumour microenvironment (TME) is known to affect disease progression and treatment response in many cancers,[355,356] and may hold the key to improving precision in patient risk stratification and development of rational risk adapted therapy. Heterogeneity of tumour infiltrating lymphocytes (TILs) has been associated with survival in FL but there is no consensus on the observed effect.[63,64] Consequently, TME derived biomarkers have not yet been developed for clinical use. Automated image analysis and multiplex immunofluorescence assays improve TME biomarker performance and reproducibility.[128,285,376] Use of such technology could therefore provide a reliable way of measuring the complex interactions in the TME of FL with the potential for improved risk prediction.

Many studies have examined the prognostic effects of TIL subsets, but with conflicting results particularly between cohorts treated before and after the introduction of rituximab.[36,63,377] Increased numbers of $CD68^+$ lymphoma associated macrophages was unfavourable in cohorts treated without rituximab[90], but either favourable[86] or not significant in cohorts treated with rituximab.[71,89] Additionally, $CD3^+$ T-cell[65,70,71,86] and $CD4^+$ T helper cell[65,70,71] densities in cohorts treated with rituximab were either favourable or not significant. The role of $CD8^+$ cytotoxic T-cells

is similarly controversial[65,71,80] and altered by the type of treatment.[70] Other TIL subsets of interest include $CD4^{+}FOXP3^{+}$ T regulatory cells (T-regs), [65,71,74,76,97] dendritic cells,[71] mast cells,[88] and PD-1 expressing T-cells.[65,71,81] However, as FL B-cells and TME cells engage in crosstalk through multiple pathways,[37] observing the diversity and spatial interactions between cells in the TME could be more informative than examining isolated components.

4.1 Developing a multiplex immune panel for follicular lymphoma

4.1.1 Motivation

A new 6-plex immunofluorescent assay using tyramide signal amplification, the Opal™ 7 colour Kit (Akoya Biosciences, Menlo Park, CA, USA) and the Ultra Discovery robot-stainer (Roche, Basel, Switzerland) was developed and validated, to enable concurrent visualisation on the same tissue sections of multiple immune subsets.

4.1.1.1.1 Panel selection

A panel of antibodies was selected to identify cell populations of non-neoplastic immune infiltrates:

- CD68 was selected to observe monocytic cells and particularly macrophages.[86,89] CD68 is expressed in the cell membrane.[17] Studies assessing macrophages in the microenvironment of FL[86,87,90,378] usually adopt CD68 as a macrophage specific marker. CD68 was adopted as a macrophage marker, to enable direct comparisons with previous studies.
- By inclusion of the CD4 marker, T-helper (T_{FH}) cells were identified.[377] CD4 is mostly expressed in the membrane of T-helper cells in the microenvironment of FL, although rare spots of staining may be seen in monocytes.[17] Thus, FL studies have often used CD4 to identify T-helper cell populations[85,379] and assess impact to survival endpoints. In this study, T-helper cells were identified as $CD4^{+}CD68^{-}$ cells to exclude $CD68^{+}$ macrophage subsets that might express CD4.
- CD8 was used to observe cytotoxic T-cells ($CD8^{+}$).[377] CD8 is a membrane marker often used to identify cytotoxic T-cells in the FL microenvironment.[62,80,380]
- FOXP3 was used to identified T regulatory cells (T-regs [$CD4^{+}FOXP3^{+}$]).[77] T-regulatory cells (T-regs) are a subset of T-helper cells, expressing CD4, FOXP3 and CD25. The FOXP3 nuclear marker is considered the principal lineage marker of these cells and is often used to identify them in FL.[74,381]
- The CD21 marker was added to observe the spatial arrangement of follicular dendritic cell (FDC)[17] meshwork areas.[95,380] CD21 may also be expressed

by mature B-cells, however it is expressed by FDC in a characteristic meshwork pattern that is easily recognisable.[93]

- PD-1 detected $CD4^+PD\text{-}1^+$ T follicular helper cells and $CD8^+PD\text{-}1^+$ lymphocytes.[81] Cells expressing PD-1 in the microenvironment of FL were found to be mostly $CD4^+$ $CD68^-$ T follicular helper cells, although some $CD8^+PD\text{-}1^+$ cytotoxic T-cells are also present. Recent studies[81] have shown that $PD\text{-}1^+$ cells do not necessarily represent exhausted cell phenotypes. Two distinct types of T_{FH} were found from functional analysis: a) the $PD\text{-}1^+_{high}$ T_{FH} cells found inside the follicles that actively supported FL B-cell growth, and b) the $PD\text{-}1^+_{low}$ cells found outside the follicles, which usually represent exhausted T-cells.
- Finally, DAPI (4′,6-diamidino-2-phenylindole) was the nuclear counterstain. No B-cell tumour marker was used, as the aim was to study the diversity of the non-neoplastic microenvironment.

The role and prognostic impact of these immune microenvironment cell types in FL has been extensively discussed in Section 1.1.2.3

4.1.2 Materials and methods

4.1.2.1 Dataset used for staining protocol development

Experimental assay development and validation was performed using sequential sections from a formalin-fixed paraffin-embedded (FFPE) tissue micro-array (TMA) constructed from 44 FL cases retrospectively collected from The Christie archives. The study was conducted with approval from the North-West Multi-centre Ethics Committee (03/08/016) and according to the Declaration of Helsinki. These patients were diagnosed in the 1980-1990s and treated using historical protocols.

4.1.2.2 Optimisation of a Vectra multiplex protocol

An overview of required steps for optimising a Vectra multiplex protocol is shown in **Figure 28**. The first step is to establish TSA immunofluorescence staining protocols for each antigen target separately (singleplex protocols).[149] At this stage, antibodies, antibody-fluorophore pairings, titration and incubation times, and optimal antigen retrieval strategies are selected, based on maximising the signal to noise

ratio. An acceptable range for the signal to noise ratio is above 10 to 1.[149] The singleplex optimisations could initially be carried out in brightfield, using DAB instead of fluorophores, where non-specific background staining is more easily distinguished from true signal. The suggested optimisation process for the singleplex protocols is 'trial-and-error', where the experiment is repeated for multiple antibody and fluorophore titrations, as well as antigen retrieval and incubation times, to determine an acceptable range of values.[149]

The optimised parameters from the singleplex assay of an antibody will also apply in the multiplex experiment (e.g., if anti-CD8 is titrated 1:200 in the singleplex and incubated for 12 minutes, the same will apply in the multiplex). The next important optimisation decision is the ordering of antibody staining cycles in the multiplex experiment. A good starting point is placing antibodies so that their order reflects the antigen retrieval time required for each antibody in the singleplex. Antibodies that require short retrieval prior to their application are placed first. After each staining cycle, additional heat antigen retrieval is carried out to strip away the previously applied antibodies. The heat retrieval has an additive effect; antigens that are detected last in the multiplex protocol will have been subjected to the longest retrieval times.

For example, suppose the CD8 and CD4 antibodies required 16 minutes and 24 minutes of antigen retrieval in their singleplex protocols, respectively. In the multiplex protocol, after the initial deparaffinisation and a short heat retrieval of 16 min, the CD8 antibody would be incubated first, followed by the secondary antibody and a fluorophore. Subsequently, additional heat retrieval of 8 minutes would follow, before incubating the CD4 antibody (for a total retrieval of 24 minutes). The second retrieval has a dual purpose; it removes all antibodies used in the CD8 staining cycle and provides the additional heating required to retrieve the 3-D structure of the CD4 antigen epitopes. However, the intermediate heat retrieval steps between staining cycles should be carefully tuned, as long heating times will also remove the fluorophores, which is an unwanted side effect. Furthermore, when multiple antigens require equal retrieval time, one of them would have to be detected at suboptimal conditions. This approach represents a heuristic that can be used as a good initial guess for the antibody ordering in the multiplex experiment. Subsequent refining of the order can be carried out by slightly changing the antibody ordering and observing the effects on the signal to noise ratio (trial-and-error).

In practice, the effect of changing the antibody order is tested by running the complete multiplex protocol (heat retrieval, washing and blocking steps for all cycles) multiple times, but each time only applying the reagents of a single staining cycle. No counterstain is applied. This setup is referred to as “single-stain multiplex”. Using this approach, the effects of fine-tuning the multiplex experiment can be quickly observed on each antibody independently. Thus, the optimal ordering and retrieval times can be found.

Another optimisation step deals with rebalancing the signals. To achieve accurate spectral unmixing, the signals from different fluorophores must have approximately equal intensity.[149] When the signals are imbalanced or too strong, bleed through might be observed, i.e. the signal of one fluorophore might show up in a second fluorophore’s channel. Therefore, the single-stain multiplex experiments are repeated with different fluorophore titrations, to establish the lowest concentration at which the signal-to-noise ratio remains acceptable.[149] This concentration is then adopted in the multiplex experiment. To achieve rebalancing, different antibody-fluorophore pairings might need to be selected, e.g. pairing bright fluorophores with weaker antibodies.

To check whether the antibodies are fully removed after the end of each staining cycle, the following stripping test can be performed; the multiplex experiment is run with only the reagents of a single staining cycle at a time (similar to a single-stain multiplex). After this cycle, heat retrieval follows. Then on the next staining cycle, only a different fluorophore is added. This setup is referred to as a “double-stain multiplex”. When antibody stripping is ineffective, both fluorophores will be seen on the sample, as the second fluorophore will bind onto the antibodies that remained on the tissue. To improve stripping, the retrieval times and antibody titrations are fine-tuned.[149]

Figure 28 Steps to set up a Vectra multiplex protocol

4.1.2.3 Building a spectral library

Once optimal parameters for all antibody staining cycles have been found, a single-stain multiplex is run for each of the antibodies to extract the representative fluorophore spectra. A section undergoing all multiplex staining steps without application of any fluorophore or antibody is used to extract the autofluorescence spectrum. Similarly, a section undergoing all the steps of the multiplex with only application of the DAPI counterstain is used to extract the DAPI spectrum.

The spectrum of a fluorophore is recorded using the Vectra microscope with all available filters. A multispectral snapshot of the section is captured, by auto-setting the exposure times on the relevant filters where the fluorophore is expected to emit. The default settings are used for all other filters. The inForm software stores the extracted spectra in a spectral library file (**Figure 29**).

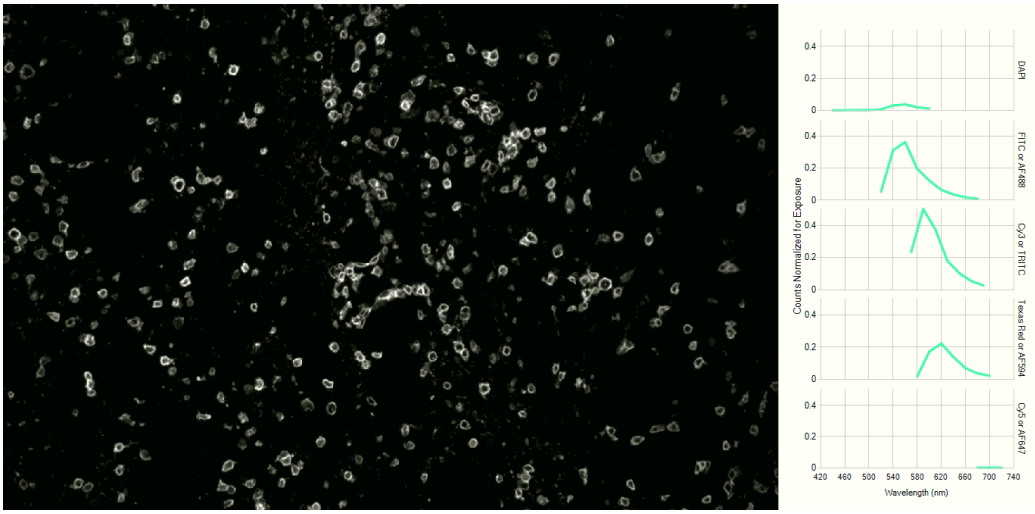

Figure 29 Adding a spectrum to a spectral library. Extracting the spectrum of fluorophore OPAL 540, paired with and anti-CD8 antibody in a single-stain multiplex experiment (snapshot of the inForm 3.5 software interface). The emission peaks on each of the 5 available filters are shown on the right panel.

Once the spectral library is built, the multiplex experiment is set. A section can then undergo all staining cycles with all reagents applied according to the optimal parameters. The multiplex-stained section will be scanned on the Vectra microscope multispectrally, by auto-setting the exposure times of all filters. The scanned image is then loaded in inForm software and unmixed into separate channels (one for each fluorophore) using the prepared spectral library (**Figure 30**).

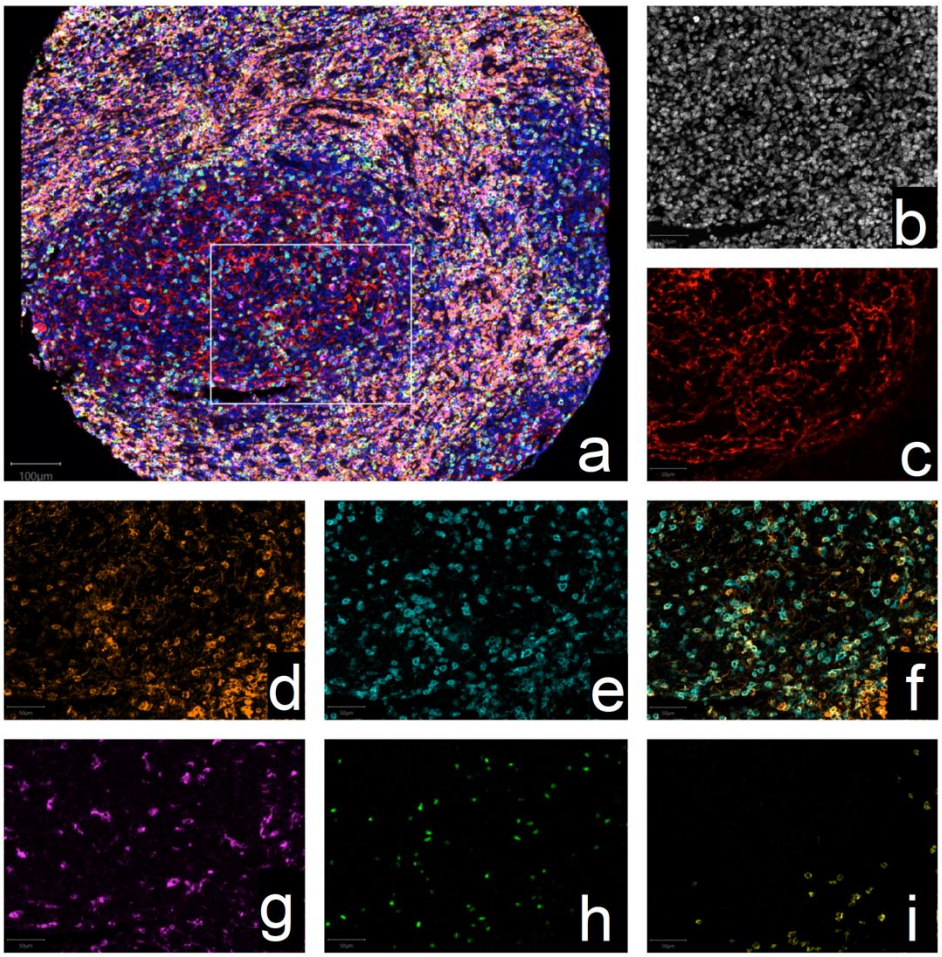

Figure 30 Human follicular lymphoma lymph node tissue, stained with the proposed 6-plex tyramide signal amplification protocol,[110] scanned multispectrally and unmixed using the Vectra 3.5 system. (a) Composite multiplex image displaying all stains together using pseudo-colours. In this composite image DAPI is blue, CD21 is red, CD4 is orange, PD-1 is cyan, CD8 is yellow, CD68 is magenta and FOXP3 is green. Panels from (b) to (i) correspond to the exact same tissue region, indicated as a white rectangle in (a) to demonstrate the process of spectral unmixing: (b) DAPI in grey; (c) CD21 in red; (d) CD4 in orange; (e) PD-1 in cyan; (f) CD4 (orange) and PD-1 (cyan) overlayed to show that PD-1 mostly almost always colocalised with CD4 in follicular regions; (g) CD68 in magenta; (h) FOXP3 in green; (i) CD8 in yellow.

4.1.2.4 Protocol validation

Agreement between singleplex and multiplex immunofluorescent assays was quantitatively validated by comparing the stained area in sequential sections of a FL FFPE TMA (**Figure 31**). DAPI was added in both the singleplex and multiplex experiments to quantify the whole tissue area in each core. Slides were scanned multispectrally on the Vectra microscope (Akoya Biosciences, software version

3.5) at 20x magnification, and the exposure times were set according to the observed signal strength of each filter. In the case of the singleplex experiments, exposure times were adjusted only for the relevant filters, while for the rest the default settings were applied; 40 ms for the overview and 150 ms for the multispectral scan. A spectral library was built and spectral unmixing of all sections was carried out in inForm 2.4 software (Akoya Biosciences).

Image analysis was subsequently performed in HALO software (Indica Labs, Albuquerque, NM, USA). Using the Multiplex Fluorescent Area Quantification module, automated thresholding of pixel intensities in each channel identified the percentage of stained area. This algorithm requires the user to specify minimum true signal intensity. These settings for the singleplex and multiplex sequential sections were chosen by the same user, leaving a "wash-out" period of 3 days between them. Cores with artefacts, such as bubbles and blood vessels were excluded from the analysis. In some cases, cores would be missing from one of the two sequential sections (tissue was broken or torn), and so these were excluded as well. A demonstration of automated area quantification is shown in **Figure 32**.

Agreement between multiplex and singleplex experiments was observed by constructing Bland-Altman plots and calculating Pearson's R^2.

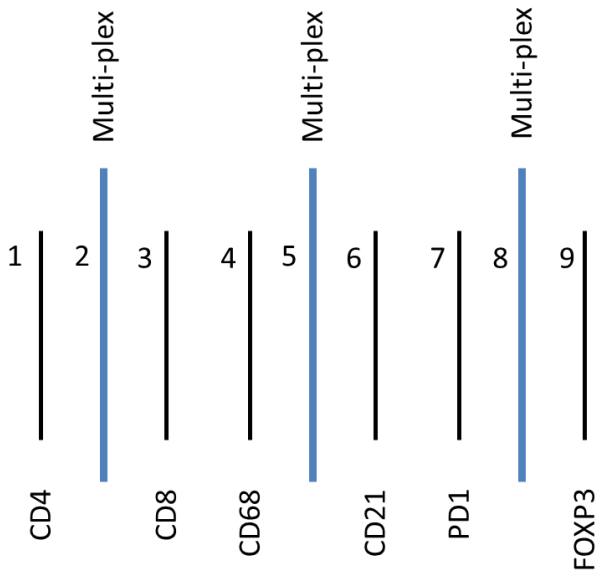

Figure 31 Sequential TMA sections setup for multiplex experiment validation. Each single-plex assay is compared to an adjacent multiplex assay.

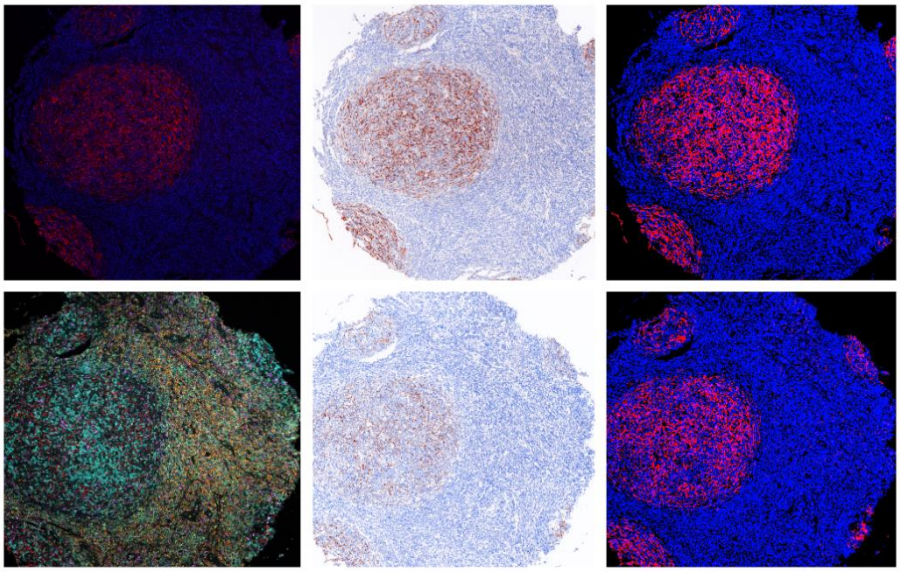

Figure 32 Area quantification in HALO for the CD21 antibody (570 fluorophore). Top row: single-plex. Bottom row: multiplex. Left: Unmixed composite image. Middle: simulated chromogenic view (inForm 2.4) for CD21. Right: Positive area quantification (HALO) where CD21 is rendered in red and DAPI in blue.

4.1.3 Results

The experimental protocol is publicly available online[110] and summarised in **Table 24**.

Table 24 Antibodies, titrations and fluorophores in the multiplex immunofluorescence protocol

Order	Antibody	Dilution	Provider	Opal detection
1	Anti-CD4 (SP35) Rabbit Monoclonal Primary Antibody	Pre-diluted	Roche, Switzerland	Opal 620
2	Anti-CD68 antibody mouse monoclonal [KP1] to CD68	1:40	Abcam, UK	Opal 650
3	Monoclonal Mouse Anti-Human CD8 Clone C8/144B	1:450	Agilent, Denmark	Opal 540
4	CD21 (2G9) Mouse Monoclonal Antibody	1:25	Cell Marque, USA	Opal 570
5	Anti-FOXP3 antibody mouse monoclonal [236A/E7]	1:60	Abcam, UK	Opal 520
6	Anti-PD-1 antibody [NAT105] (ab52587) Mouse monoclonal	1:150	Abcam, UK	Opal 690
The order reflects the order the antibodies are applied on the tissue.				

Comparisons between singleplex and multiplex experiments demonstrated satisfactory linear correlations as shown in **Figure 33**. For most markers, slightly lower

staining expression was observed in the multiplex compared to the singleplex experiments. Lower expression may derive from incomplete stripping in between staining cycles, which may lead to steric obstruction and slightly decreased antibody binding. This effect was however not significant, as seen in Bland-Altman plots (**Figure 34**). The mean difference of the two experiments was usually close to zero, with >94% of data points lying within the limits of agreement (mean ± 1.96 standard deviation) for all markers.

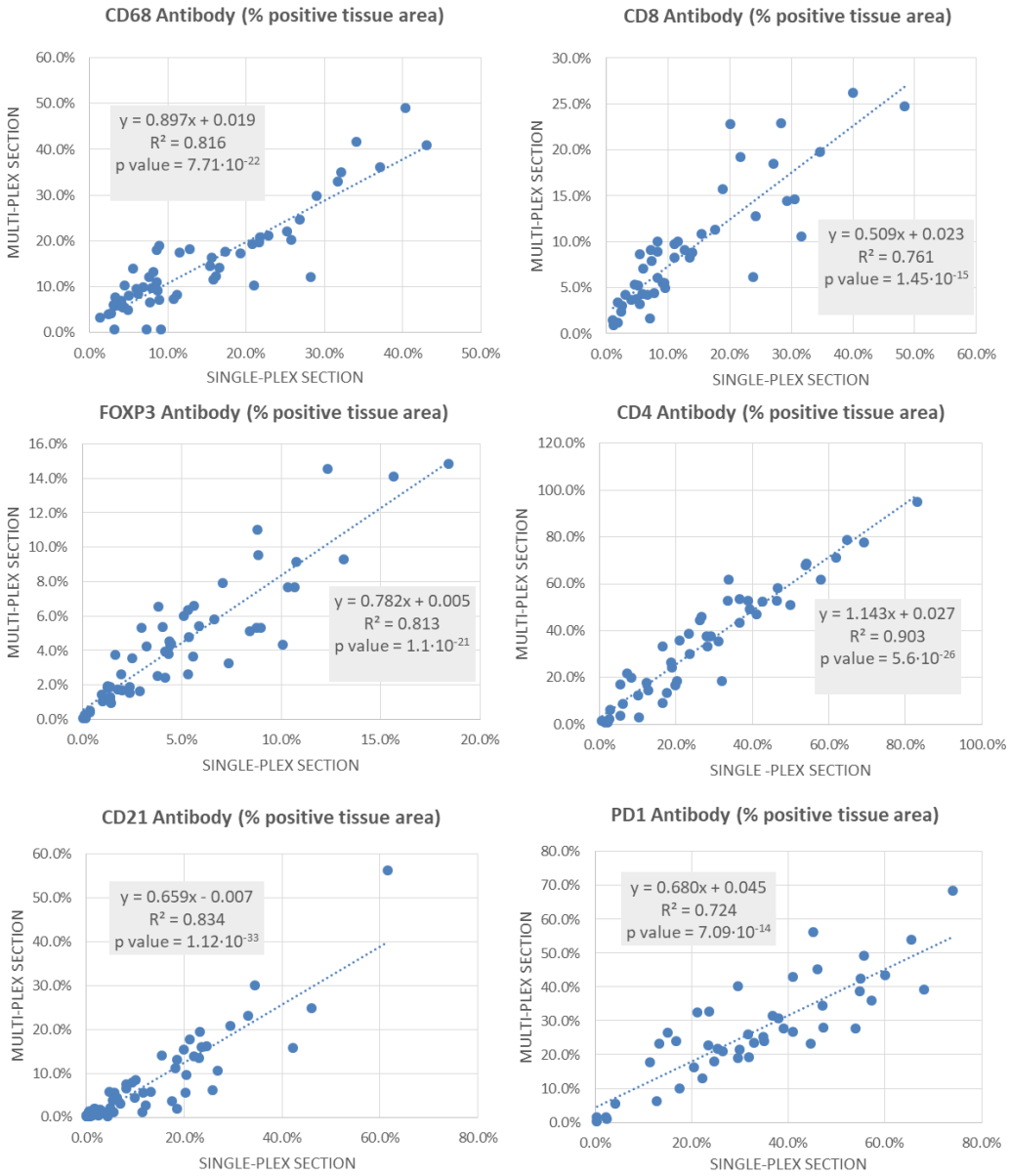

Figure 33 Comparison of % tissue area stained by each marker in two sequential 4µm TMA sections, a multiplex and a single-plex. The single-plex was also stained with DAPI and both sections were scanned multispectrally at 20x and unmixed with the same spectral library. Each point represents a TMA core.

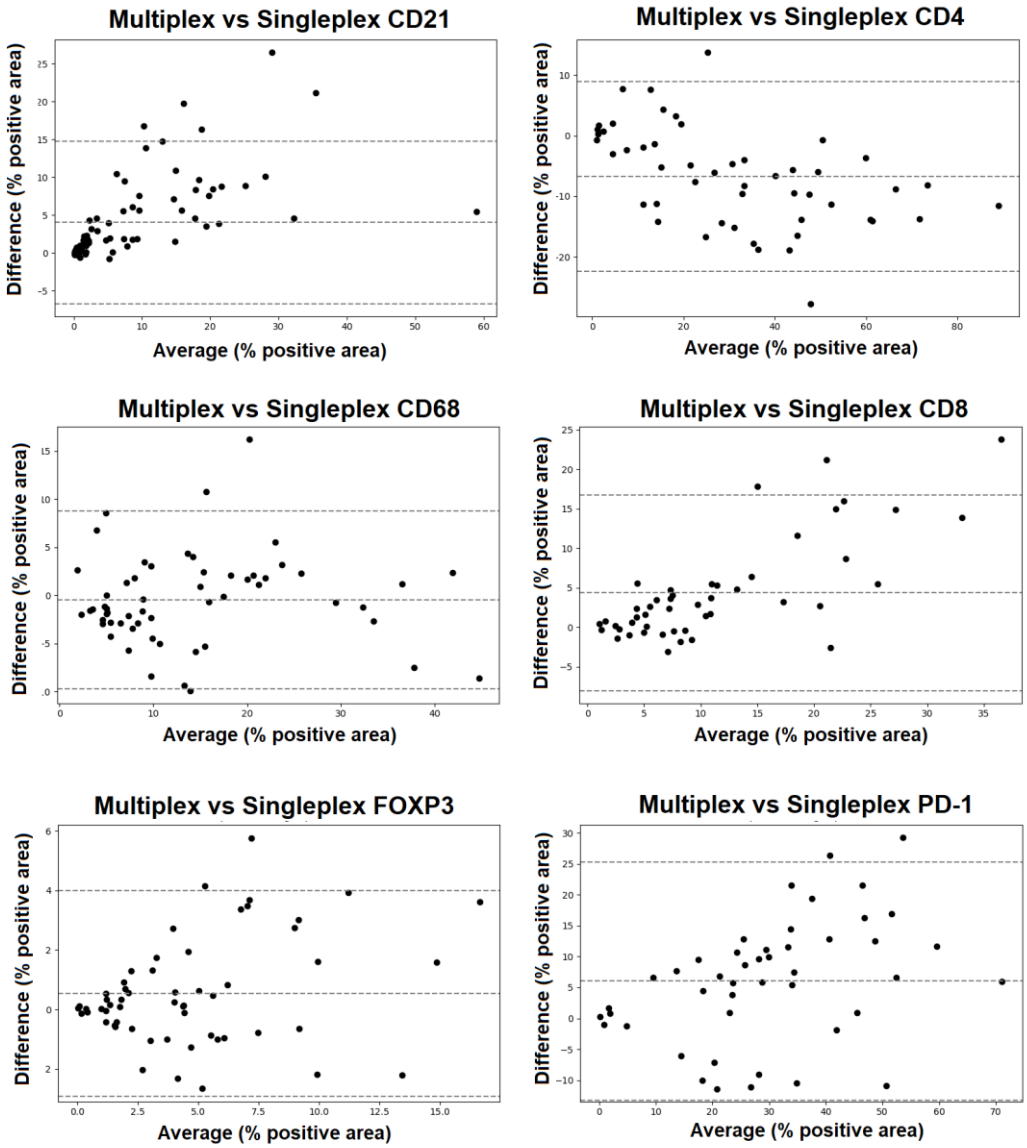

Figure 34 Bland-Altman plot comparisons between singleplex and multiplex immunofluorescent assays for each antibody. Antibody expression is measured as a percentage of the positively stained tissue area. Each point represents a tissue microarray core. The dotted lines represent the limits of agreement (± 1.96 standard deviation of difference).

4.1.4 Discussion

Through extensive optimisation, a 6-plex multiplex immunofluorescent protocol was developed using the Vectra platform. The protocol has been validated against adjacent singleplex assays and demonstrated sufficient agreement. Furthermore, it has been made publicly available in the protocols.io experimental platform.[110] This protocol, combined with automated analysis of the multiplex images can be used to study in depth the immune populations in the microenvironment of follicular lymphoma.

4.2 Follicular lymphoma biomarkers based on diversity of the immune microenvironment

4.2.1 Motivation

In ecological sciences several metrics can describe species diversity. The Shannon diversity (or entropy) index, a measure derived from information theory, quantifies biodiversity in terms of "evenness". For example, if three species are found in an area, and one accounts for 99% of the population, this community would be considered less diverse than one where the three species are found in approximately equal abundances. Shannon's entropy has found applications in histopathology to quantify heterogeneity of HER2 expression[382] and chromosome 8q24 copy number variation[383] in breast cancer. If each TIL phenotype is considered as a species, this metric can be applied to quantify the immune infiltrate diversity in the TME.

Similarly, it is possible to quantify the diversity of not only phenotypes but also their spatial interactions, which is recognised for its potential as a biomarker[3] for many tumour types including FL.[77,97] The Hypothesised Interactions Distribution (HID) method[100] can identify spatial interactions defined as co-localisation of different cell types within 30 μm. The diversity of these spatial interactions can also be investigated using Shannon's entropy.

The aim was to develop a methodology to quantitatively assess immune infiltrate diversity in the tumour microenvironment of FL and test its potential utility as a prognostic biomarker. To this end, an automated image analysis pipeline was developed and validated to simultaneously identify cells positive for CD4, CD68, CD8, CD21, FOXP3 and PD-1 in multiplex images produced with the protocol developed in section 4.1. Furthermore, the use of Shannon's entropy was tested as a means of quantifying overall and spatial diversity of TIL populations. It is shown that increased diversity of TIL populations and interactions are both associated with improved overall survival (OS) in a cohort of FL patients.

4.2.2 Materials and methods

4.2.2.1 Dataset

4.2.2.1.1 Cohort selection

The study was conducted with approval from the North-West Multi-centre Ethics Committee (03/08/016) and according to the Declaration of Helsinki. Patients with WHO 2008 histologically confirmed FL were identified from the archives of The Christie NHS Foundation Trust, Manchester, UK. An initial cohort of 1004 patients was retrieved from the electronic database by searching for the keywords "follicular lymphoma". Examination of the first 350 patients in a random order identified 262 patients meeting the inclusion criteria: adult patients with previously untreated FL; diagnosed from incisional or excisional biopsy; non primary cutaneous; and treated at first presentation with radiotherapy, watchful waiting or rituximab-based systemic therapy. Pre-treatment diagnostic biopsies were requested for 262 patients, of which 131 had sufficient tissue for analysis. The 131 patients included in this study were diagnosed between 1998 and 2015 and had a median follow-up of 114 months (range 3-199 months). A histological diagnosis of FL was re-confirmed by an expert haemato-pathologist (R.B). Regions of interest were identified by a haemato-pathologist (R.B) and cores were extracted in triplicate from formalin-fixed, paraffin-embedded (FFPE) blocks to construct five tissue microarrays (TMA). Follicular and extrafollicular regions were both selected for inclusion in the TMA. A section of each TMA stained with H&E is provided,[384] to demonstrate the morphology of selected regions. No cases of transformed high grade FL at baseline were present in this cohort. Three patients with FL grade 3b were excluded, as their disease progression and treatment pathways resemble more closely Diffuse Large B-cell Lymphoma and grade 3b FL is generally considered a separate disease entity.[385] After staining and scanning, some cores were excluded because of poor quality and artefacts, leaving 342 cores from 127 patients available for further analyses.

4.2.2.1.2 Clinical endpoints

OS was recorded for all patients. Progression-free survival (PFS) and disease progression within 24 months of starting treatment (POD24) were recorded only for patients treated with rituximab-containing immuno-chemotherapy at first presentation, as these endpoints are not well defined for patients assigned to watchful wait-

ing. Patients who received radiotherapy alone at first presentation were also excluded from PFS and POD24 analysis, as recent findings[65] have shown different effects of TILs in FL patients treated with and without rituximab. The events of disease progression and relapse were defined using the Lugano criteria.[39] PFS was calculated from diagnosis until the first observed progression event (or disease specific death) or, if no events were observed, until the date of last follow-up. POD24 was calculated from start of immuno-chemotherapy treatment. The patient flowchart for all survival analyses in this study is shown in **Figure 35**.

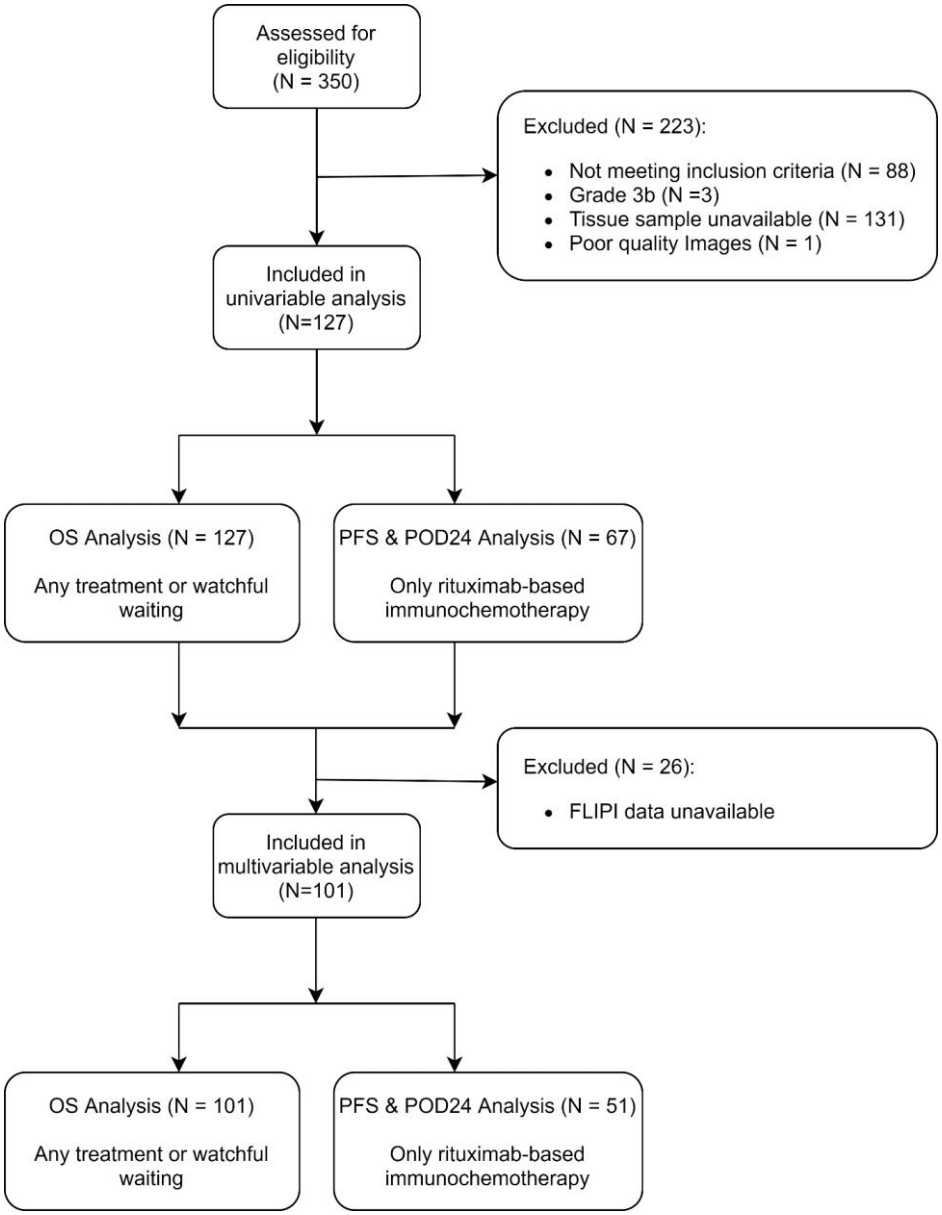

Figure 35 Patient flowchart in the follicular lymphoma study.

4.2.2.2 Multiplex immunofluorescence imaging

This section describes preparation of the image dataset using multiplex immunofluorescence and multispectral scanning.

4.2.2.2.1 Staining protocol

A single 4 µm section was cut from each of the five TMAs of the cohort. These sections were stained with the 6-plex immunofluorescent assay described in section 4.1, using tyramide signal amplification, the Opal 7 colour kit (Akoya Biosciences, CA, USA) and the Ultra Discovery auto-stainer (Roche, Switzerland).[110]

4.2.2.2.2 Image acquisition

Stained sections were scanned with the Vectra 3.5 microscope (Akoya Biosciences). Initially, a low-resolution scan (10x) was performed to manually annotate the TMA core locations. Then, a multispectral image of each core was acquired at 20x magnification (0.49 µm/ pixel). Spectral unmixing[134] was performed using inForm 2.4 software (Akoya Biosciences). To separate the fluorophore signals, a spectral library was pre-built; the individual spectrum of each fluorophore, DAPI and auto-fluorescence were acquired from single-plex controls. After unmixing, the images consisted of 6 channels, each containing the intensities of a different fluorophore, plus two channels for DAPI and auto-fluorescence ($2420 \times 2420 \times 8$ pixels). Areas containing artefacts were manually excluded.

All prepared images are publicly available[††].

4.2.2.3 Cell detection

Spectral unmixing was performed using inForm 2.4 software (Akoya Biosciences) and all images were manually examined to exclude any areas containing artefacts, such as folded tissue, bubbles and blood vessels. Nuclear detection in these images was challenging, because of densely packed and overlapping cells. A convolutional neural network (CNN) using the "StarDist" method[175] was trained for concurrent nuclear detection and segmentation. The model was trained from scratch on nuclear outline annotations drawn by a trained non-expert, under supervision from a pathologist (R.B.). The CNN was built with a U-net type[191] architecture and trained

[††] http://dx.doi.org/10.17632/274xbhc5rx.3

with two losses; the first penalised whether a pixel was correctly predicted as nucleus vs. background; the second penalised the predicted distance from each nuclear pixel to the background, which is used to obtain the shape of the nucleus.

To assess nuclear segmentation performance, the average precision $AP = \frac{TP}{TP+FP+FN}$ was considered, where true positive (TP) predictions are defined as predicted nuclei, for whom exist ground truth nuclei with sufficient overlap. Overlap was measured as intersection over union (IoU) > 30%. False positive (FP) were the unmatched predicted nuclei, while false negative (FN) were the unmatched ground truth nuclei. There were 3 ROI (883 nuclei) in the test set, 3 ROI (906 nuclei) in the validation set and 35 ROI (6791 nuclei) in the training set. The average precision for the testing set of nuclei was AP = 0.827. The worst image in the testing set is presented in **Figure 36** (AP=0.733) and in **Table 25** the AP for different threshold of the IoU is given for the test set.

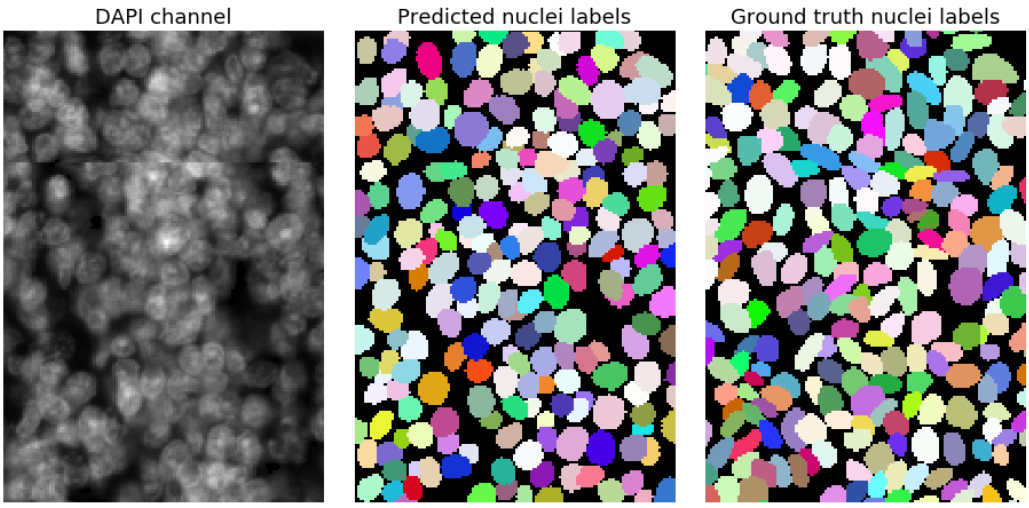

Figure 36 Worse performing image in test set for nuclear segmentation (AP=0.733). The amount of overlap between nuclei is challenging even for human annotators.

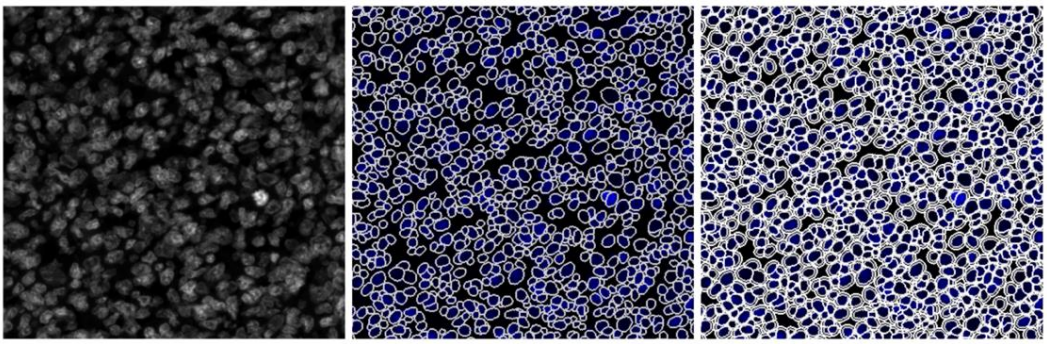

Figure 37 Growing membranes around detected nuclei.

After nuclear segmentation, simulated membranes are grown around the nuclei by maximum 1.5 μm to represent whole cells (**Figure 37**) and measurements are taken of the median intensity for all stains and each cell compartment (nucleus, membrane). All images in the dataset were manual examined and areas that presented artefacts because of folded tissue, bubbles or blood vessels were excluded from further analysis.

Table 25 Segmentation performance in the test set for different thresholds of the intersection over union (IoU) parameter

IoU threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AP	0.915	0.866	0.827	0.726	0.603	0.494	0.341	0.157	0.013
The test set included 3 ROI with 883 nuclei. AP: Average precision.									

4.2.2.4 Positive cell scoring

Cell scoring was performed using an interactive computer assisted scoring system, that required input from a human observer to select a positivity cut-off for each stain.

A validation set of 10 images, each containing a whole TMA core, was selected to determine a positivity cut-off, as follows; first, intensity scaling onto a consistent colour map across all images was carried out for each stain so that equal intensity levels were represented by equal brightness. Measurement of the median stain intensities in the relevant compartment (nuclear for FOXP3 and membrane for all the rest) were carried out. Then, a cut-off threshold was selected per image core and stain by two independent annotators (a non-expert [A.M.T] and a trainee pathologist

[M. D.]) to separate positive from negative cells. A single threshold was then selected as a positivity cut-off per stain by averaging all thresholds selected for the images in the validation set by both annotators.

Agreement between the two annotators that scored positive cells using the computer assisted scoring system is shown in **Table 26**. Agreement was assessed by using the selected cut-offs to classify the cells as positive or negative and calculating the f_1 score (harmonic mean of precision and recall) between the labels generated by different annotators. The fact that a single threshold across all images was mostly adequate to separate positive from negative cells ($0.68 \leq f_1$ score ≤ 0.92) indicates low staining variation across different patients and TMA blocks. These single cut-off thresholds were finally applied to phenotype cells in the entire dataset.

Table 26 Agreement for cell labels generated by selecting a positivity cut-off per image in the validation set.

FOXP3	Annotator 1	Trainee pathologist	Single threshold
Annotator 1	-	0.83	0.92
Trainee pathologist		-	0.90
Single threshold			-
CD8	Annotator 1	Trainee pathologist	Single threshold
Annotator 1	-	0.72	0.86
Trainee pathologist		-	0.85
Single threshold			-
CD4	Annotator 1	Trainee pathologist	Single threshold
Annotator 1	-	0.88	0.87
Trainee pathologist		-	0.88
Single threshold			-
CD68	Annotator 1	Trainee pathologist	Single threshold
Annotator 1	-	0.49	0.76
Trainee pathologist		-	0.68
Single threshold			-
PD-1	Annotator 1	Trainee pathologist	Single threshold
Annotator 1	-	0.69	0.76
Trainee pathologist		-	0.86
Single threshold			-
Agreement is calculated as the $f_1$score, representing the harmonic mean of precision and recall for the binary classification task of assigning a cell as positive or negative for each stain.			

This method was applied to identify cells positive for CD4, FOXP3, CD8, CD68 and PD-1. This approach was not adopted for CD21, as the staining pattern of

$CD21^+$ cells followed a non-convex meshwork pattern which would be challenging to simulate accurately by simply growing simulated membranes around the nuclei.

4.2.2.5 Cell density quantification

The tissue area was detected per core by Otsu's thresholding[163] of the DAPI channel. Cell density was subsequently measured for each cell phenotype of interest by dividing the number of positive cells by the tissue area. Thus, cell densities were calculated for cell types of interest, identified from prior studies on the FL microenvironment,[36,37,86,89,90,97,377,63,65,70,71,74,76,80,81] namely $CD4^+CD68^-$ T-helper cells, $CD4^+FOXP3^+$ T-regs, $CD8^+$ cytotoxic T-cells, $CD68^+$ macrophages, $CD4^+CD68^-$ $PD\text{-}1^+$ T_{FH} cells and $CD8^+PD\text{-}1^+$ T-cells.

Additionally, the total immune infiltrate was measured as the number of cells expressing any of the CD4, FOXP3, CD8, CD68 or PD-1 markers. The immune infiltrate ratio was subsequently calculated by dividing the immune infiltrate cells by the number of all non-immune cells that expressed only DAPI. This ratio can be used to represent the extent of total immune infiltration in the microenvironment of FL.

4.2.2.6 Identifying $CD21^+$ dendritic meshwork areas

To quantify the extent of $CD21^+$ dendritic meshwork areas, manual annotations were drawn around them for all samples (**Figure 38),** under supervision of a pathologist. Subsequently, the area covered by these meshwork patterns was measured, and expressed as a proportion of the total tissue area.

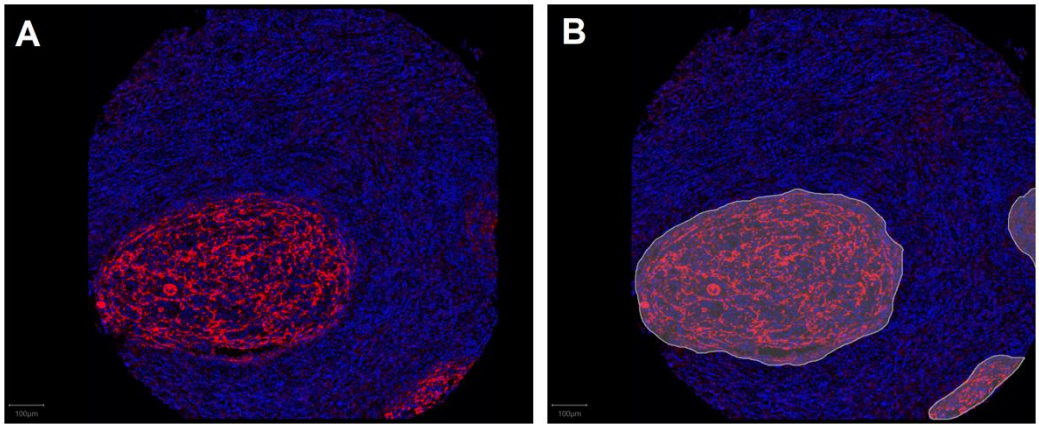

Figure 38 Dendritic meshwork areas were annotated manually by drawing around the CD21^{+} meshwork pattern regions. A) CD21 (red) and DAPI (blue) view of a multiplex TMA core image. B) Manual annotation of dendritic meshwork areas, overlayed in grey.

4.2.2.7 TME diversity quantification

TME diversity was assessed by computing the Shannon's entropy diversity index for non-neoplastic immune cell phenotypes. In this analysis a phenotype is defined as a combination of expression of five immune cell markers: CD4, CD8, CD68, PD-1 and FOXP3. For these five markers there are $n = 2^5 = 32$ potential combinations of stain positivity, each defining a separate cell phenotype. Cells expressing none of these markers were not included in the diversity analysis, as the aim was to assess the diversity of the non-neoplastic immune tumour microenvironment. Therefore, cells positive only for DAPI or only for CD21 were excluded as these could potentially include FL B-cell subsets. Shannon's entropy is calculated as:

$$Entropy = -\sum_{i}^{N} p_i \ln(p_i) \tag{4}$$

where N is the number of total species in the community and p_i the proportion of individuals belonging to the ith species.

4.2.2.8 Diversity of spatial interactions

The diversity of TIL spatial interactions in each sample was additionally quantified by applying the HID methodology[100]. HID performs a pair-wise examination of cell types identified during cell scoring and counts their spatial interactions, i.e. their frequency of co-occurrence within a pre-specified distance (see **Figure 39**). Further implementation details can be found in Rose et al[100]. The distance parameter was selected as 30 μm similar to chapter 3, which represents a neighbourhood of 3-4

cells. A co-localisation between each pair of phenotypes i and j was considered a unique type of spatial interaction. The proportion of all interactions belonging to this type $p_{i,j}$ could then be calculated. If each type of interaction is considered as a separate "species", Shannon's entropy diversity index for the distribution of interactions in a sample can be derived:

$$\text{Interaction entropy} = -\sum_{j=1}^{n}\sum_{i=\mathrm{j}}^{n} p_{i,j}\ln\left(p_{i,j}\right) \tag{5}$$

In the current study all n TIL phenotypes that were observed in the samples across the entire dataset were assessed, while cells expressing only DAPI were ignored. Intuitively, interaction entropy quantifies the diversity of co-localisations between immune subsets in the tumour microenvironment.

Figure 40 provides a summary of the methodology steps.

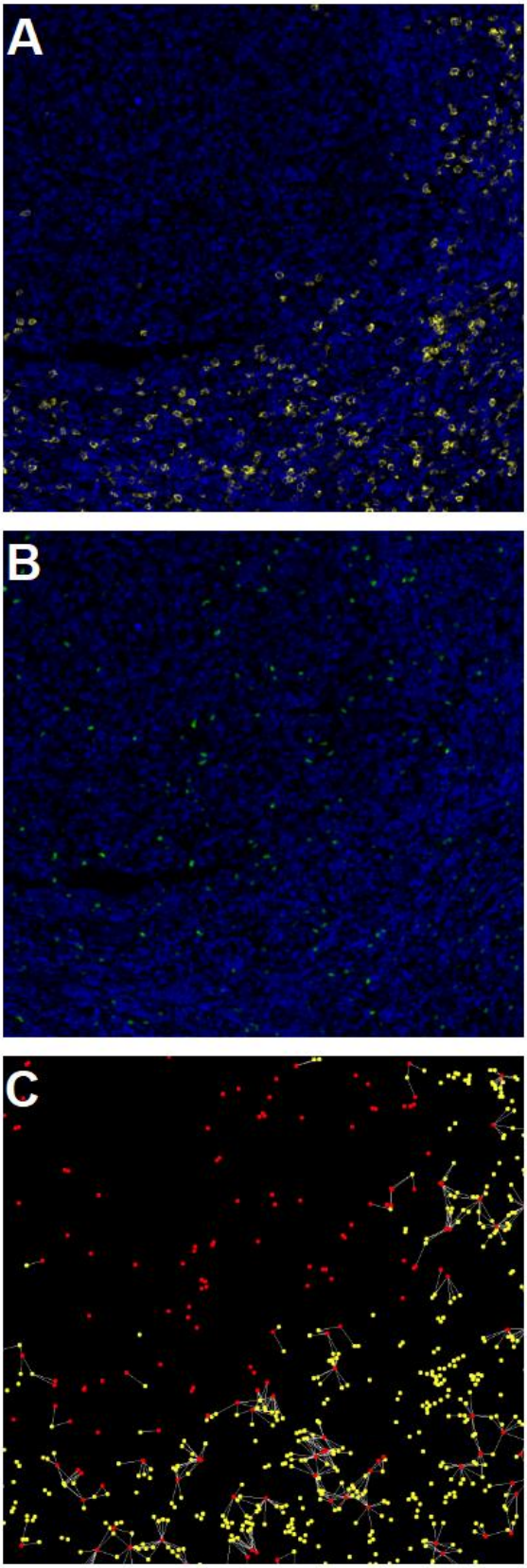

Figure 39 Demonstration of how spatial interactions are calculated. (A) CD8 (yellow) and DAPI (blue); (B) FOXP3 (green) and DAPI (blue); (m) spatial "interactions" between cells scored as $FOXP3^+$ (shown in red) and $CD8^+$ (shown in yellow) are plotted as connections (shown in white) between cells occurring within 30μm of each other.

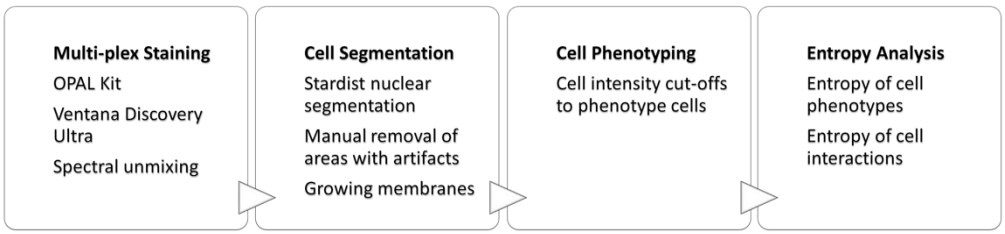

Figure 40 Summary of methodology for automated diversity analysis in the tumour micro-environment of FL.

4.2.2.9 Statistical analysis

Since cores were extracted in triplicate for each patient, the median feature value was used to represent the patient. Univariable analysis for OS and PFS was carried out using Cox regression models, where all features were treated as continuous variables. Multivariable analysis involved building Cox regression models to assess associations, independent of FLIPI. FLIPI was assessed as an ordinal score (0-5)[49]. Kaplan-Meier analysis for the diversity features was carried out by dichotomising the variables at the optimal cut-point and adjusting the estimated significance to account for bias using the Contal and O'Quigley method.[386] The findcut implementation in SAS 9.4 was used for optimal cut-point selection.[387] Univariable and multivariable logistic regression was also applied for POD24 prediction. All patients included in POD24 analyses had at least 24 months of follow-up.

Significance was assessed at a level $\alpha = 0.05$ and the Bonferroni correction was applied to account for multiple hypothesis testing. Statistical tests were performed using the lifelines v.0.14.6, statsmodels v.0.10.1 and scipy v.1.3.1 libraries in Python.

4.2.3 Results

Table 27 summarises patient characteristics at diagnosis. The 3- and 5- year OS rates were 97.7% (95% CI: 92.67, 99.22) and 94.46% (95% CI: 88.51, 97.14).

Table 27 Baseline characteristics of the 127-patient cohort

Characteristic	Value	No.	%
Median Age, years	59		
Age, years	≤ 60 > 60	69 58	54% 46%
Age Range, years	31-92		
Histologic Grading	1 1/2 2 2/3a 3a Unspecified	37 12 45 6 20 7	29% 9% 35% 5% 16% 6%
Serum LDH	> 549 IU/L ≤ 549 IU/L	12 96	11% 89%
Ann Arbor Stage	I-II III-IV	46 81	36% 64%
No. of Nodal Sites	0-4 > 4	79 35	70% 30%
Hb Level g/dL	≥ 12 < 12	91 20	82% 18%
BM Involvement	Presence Absence	45 73	38% 61%
ENS, Excluding BM	Presence Absence	37 88	30% 70%
ECOG Performance Status	0-1 > 1	85 3	97% 3%
FLIPI	0-1 2 3-5	45 33 23	44% 33% 23%
Initial Treatment	Watchful waiting (WW) Radiotherapy Rituximab regimens	35 25 67	28% 20% 52%
Rituximab Regimens	R-CVP R-CHOP R-Ibritumomab tiuxetan Rituximab single agent R-Bendamustine	44 9 10 3 1	66% 13% 15% 5% 1%

BM indicates bone marrow; ECOG, Eastern Cooperative Oncology Group (ECOG) Performance Status; ENS, Extra-Nodal Sites; Hb, haemoglobin; LDH, Lactic Acid Dehydrogenase; R, rituximab; R-CHOP, rituximab, cyclophosphamide, doxorubicin hydrochloride (hydroxydaunorubicin), vincristine sulphate and prednisolone; and R-CVP, rituximab, cyclophosphamide, vincristine sulphate, and prednisolone.

4.2.3.1 Prevalence of POD24

For POD24 analysis, 67 patients had a minimum of 2 years follow-up, and 14 had an observed progression event within 24 months of the initiation of immuno-chemotherapy (20.3%). POD24 was an indicator of unfavourable OS (p=0.001) and PFS (p<0.001; (**Figure 41**).

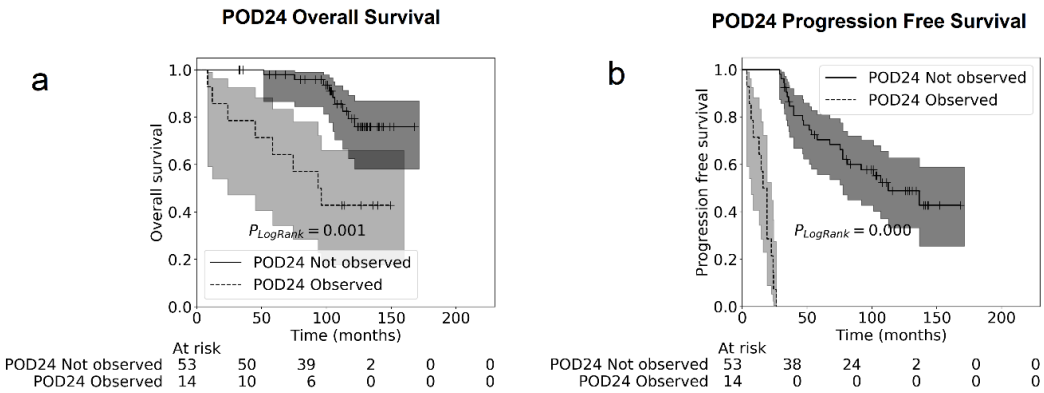

Figure 41 Kaplan-Meier analysis with POD24 in the rituximab treated subgroup to test associations to OS and PFS.

4.2.3.2 Prognostic value of clinical and biochemical characteristics

As a baseline for this cohort, the prognostic value of clinical and biochemical characteristics commonly used to assess patient risk (e.g., FLIPI) was tested in univariable Cox regression analysis.

4.2.3.2.1 FLIPI and extra-nodal site involvement predict OS

FLIPI considers age, stage, haemoglobin levels, Lactate Dehydrogenase (LDH) levels and number of nodal site involvement. FLIPI data was available at diagnosis for 101 patients, of which 51 were treated with rituximab-based regimens. The distribution of FLIPI index risk (low: 44%, intermediate: 33%, high: 23%) is similar to that reported by others[51]. FLIPI was prognostic for overall survival (HR=1.57, 95% CI 1.09, 2.26) in the 101-patient cohort, but not PFS (HR=1.30, 95% CI 0.95, 1.77) in the 51 rituximab treated patients. When examining individual FLIPI components, age, and haemoglobin were associated with OS (**Table 28**). Additionally, extra-nodal site (HR=3.81, 95% CI 1.68, 8.63) and bone marrow (HR=3.33, 95% CI 1.39, 8.01) involvement correlated to unfavourable OS.

4.2.3.2.2 Haemoglobin, ECOG status and stage predict early progression

Only low haemoglobin levels (HR=2.66, 95% CI 1.31, 5.43) and ECOG status (HR=6.83, 95% CI 1.49, 31.39) were associated with unfavourable PFS (**Table 28)**. Advanced stage at diagnosis was more commonly observed in patients who developed POD24 (p=0.041,**Table 28**).

Table 28 Survival and POD24 analysis for clinical variables

Adverse Factor	Cox PH Univariable OS			Cox PH Univariable PFS			POD24	
	All Patients			Rituximab Patients				
	HR (95% CI)	P*	N	HR (95% CI)	P*	N	P_{POD24}†	N
Age > 60 years	2.80 (1.2, 6.53)	**0.017**	127	0.77 (0.41, 1.45)	0.412	67	0.088	67
Grade 3a	0.98 (0.36, 2.63)	0.961	120	0.61 (0.23, 1.57)	0.305	61	0.308	61
LDH > 549 IU/L	0.95 (0.22, 4.11)	0.942	108	0.82 (0.31, 2.14)	0.688	58	0.315	58
Stage III or IV	2.69 (0.99, 7.3)	0.052	127	2.18 (0.77, 6.15)	0.140	67	**0.041**	67
NS > 4	0.95 (0.37, 2.45)	0.912	114	1.29 (0.65, 2.55)	0.469	56	0.085	56
Hb < 12 g/dL	3.13 (1.23, 7.97)	**0.017**	111	2.66 (1.31, 5.43)	**0.007**	59	0.106	59
BM Presence	3.33 (1.39, 8.01)	**0.007**	118	1.78 (0.88, 3.6)	0.107	60	0.122	60
ECOG > 1	6.05 (0.73, 49.97)	0.095	88	6.83 (1.49, 31.39)	**0.014**	46	0.090	46
ENS Presence	3.81 (1.68, 8.63)	**0.001**	125	1.26 (0.67, 2.37)	0.474	65	0.445	65
FLIPI 0-5	1.57 (1.09, 2.26)	**0.014**	101	1.30 (0.95, 1.77)	0.102	51	0.214	51

OS indicates overall survival; PFS indicates progression free survival; POD24, progression of disease within 24 months of treatment; CI indicates confidence intervals; HR, hazard ratio; PH, proportional hazards; BM indicates bone marrow; NS, nodal sites; ECOG, Eastern Cooperative Oncology Group (ECOG) Performance Status; ENS, Extra-Nodal Sites; Hb, haemoglobin; LDH, Lactic Acid Dehydrogenase; FLIPI, Follicular Lymphoma International Prognostic Index. *P value testing significance of the log rank test. †P value testing significance of the Mann-Whitney U statistic testing differences between POD24 positive and negative subgroups. P values < 0.05 are shown in bold.

4.2.3.3 Distribution of immune cell densities and diversity metrics

Table 29 provides the median, inter-quantile range, and intra-patient coefficient of variation (CoV) of cell populations and diversity metrics in the 127-patient cohort. The CoV measures intra-patient heterogeneity between different TMA cores for the same patient.

Table 29 Median and interquartile range for tumour microenvironment features in the data set

Features		**Feature Distribution (Median [Q25, Q75])**		
		Cohort (N=127)	**Rituximab (N=67)**	**CoV**
Cell Density, cells / mm^2	$CD4^+CD68^-$ T-helper cells	219.5 [110.9, 311.0]	170.6 [83.3, 275.9]	45.7%
	$CD4^+FOXP3^+$ T-regs	14.1 [5.8, 24.1]	11.5 [5.7, 23.8]	51.6%
	$CD8^+$ T-cells	72.8 [26.8, 125.5]	58.0 [22.8, 117.0]	37.4%
	$CD68^+$ cells	126.0 [77.6, 184.6]	121.2 [74.9, 171.8]	28.7%
	$CD4^+CD68^-PD\text{-}1^+$	26.6 [9.0, 58.3]	25.1 [6.7, 53.5]	61.3%
	$CD8^+PD\text{-}1^+$	10.3 [3.9, 23.0]	9.5 [3.9, 17.0]	58.3%
Cell Ratio	Immune infiltrate ratio†	0.4 [0.3, 0.7]	0.4 [0.2, 0.6]	32.4%
% Positive Area	$CD21^+$ dendritic meshwork area	0.3 [0.0, 0.4]	0.3 [0.1, 0.5]	73.5%
Diversity, natural digits	Phenotype entropy	1.9 [1.7, 2.1]	1.9 [1.8, 2.1]	8.3%
	Interaction entropy	4.0 [3.6, 4.4]	4.0 [3.7, 4.4]	7.7%

Q25 and Q75: 25th and 75th quantile, respectively. CoV: The average intra-patient coefficient of variation. †Immune infiltrate ratio is calculated as the total immune cells (positive for any marker) divided by the number of cells that expressed only DAPI.

The CoV of the diversity metrics was very low (7.7% and 8.3% for interaction and phenotype entropy, respectively), indicating that diversity in FL could be robustly measured by use of triplicate core samples. In contrast, all other features assessed showed higher disagreement between measurements from different cores (CoV >28 %), indicating that a higher number of cores may be necessary to obtain a representative value of these features on a patient level.

4.2.3.4 Cell population densities were not prognostic in multivariable analysis

In univariable Cox regression for OS, only the density of macrophages (HR=0.99, 95% CI 0.98, 1.0) was significant after the Bonferroni correction for multiple comparisons (**Table 30**). However, in univariable Cox regression for PFS (**Table 30**), POD24 logistic regression (**Table 32**), and all multivariable analyses (**Table 31, Table 32**), none of the cell population densities were statistically significant.

4.2.3.5 Immune infiltrate diversity analysis

Increased diversity of cell types (HR=0.22, 95% CI 0.07, 0.64) and diversity of spatial interactions (HR=0.47, 95% CI 0.27, 0.82) were favourable for OS in univariable Cox regression analysis (N=127, **Table 30**). Furthermore, in multivariable Cox regression analysis (N=101, **Table 31**), the diversity of phenotypes remained favourable for OS after the Bonferroni correction (HR=0.39, 95% CI 0.20, 0.75). Therefore, the immune diversity biomarker offers prognostic value, independent of FLIPI. This effect was not seen in PFS (**Table 31**) and POD24 (**Table 32**) regression analyses.

Kaplan-Meier analysis showed a trend towards increased diversities being favourable for OS (**Figure 42**), when stratified at the optimal cut-off. The optimal cut-off was selected using the Contal & O' Quigley[387] method, where all possible cut-offs are tested and the p-value is adjusted to account for the bias of multiple testing. Stratification of OS based on the diversity of phenotypes was significant (adjusted p = 0.032), assigning 45.6% of patients to the poor prognostic group.

Table 30 Univariable survival analysis for features derived from the tumour microenvironment

	Univariable Analysis	Cox PH Univariable OS All Patients, N=127, 27 Events		Cox PH Univariable PFS Rituximab Patients, N=67, 39 Events	
		HR (95% CI)	***P****	**HR (95% CI)**	***P****
Cell Density, cells / mm²	$CD4^+CD68^-$ T-helper cells	1 (1, 1)	0.264	1 (1, 1)	0.160
	$CD4^+FOXP3^+$ T-regs	0.96 (0.92, 0.99)	**0.023**	0.97 (0.95, 1)	**0.022**
	$CD8^+$ T-cells	0.99 (0.99, 1)	0.055	1 (0.99, 1)	0.211
	$CD68^+$ cells	0.99 (0.98, 1)	**0.002**	0.99 (0.99, 1)	**0.010**
	$CD4^+CD68^-PD\text{-}1^+$	0.99 (0.98, 1.01)	0.278	1 (0.99, 1.01)	0.467
	$CD8^+PD\text{-}1^+$	0.97 (0.94, 1)	0.084	0.99 (0.97, 1.01)	0.253
Cell Ratio	Immune infiltrate ratio†	0.21 (0.05, 0.92)	**0.039**	0.25 (0.08, 0.82)	**0.023**
% Positive Area	$CD21^+$ dendritic meshwork area	1.65 (0.31, 8.8)	0.556	1.35 (0.41, 4.48)	0.626
Diversity, natural digits	Phenotype entropy	0.22 (0.07, 0.64)	**0.006**	0.69 (0.3, 1.61)	0.393
	Interaction entropy	0.47 (0.27, 0.82)	**0.007**	0.81 (0.52, 1.27)	0.359

HR: hazard ratio; CI: confidence intervals; PH: proportional hazards; OS: overall survival; PFS: progression free survival. *The log rank test p value examines whether the null hypothesis of no effect (H_0: HR=1) can be rejected. †Immune infiltrate ratio is calculated as the total immune cells (positive for any marker) divided by the number of cells that expressed only DAPI. P values < 0.05 are shown in bold. All features were assessed as continuous variables. P values < 0.005 remain significant after the Bonferroni correction for multiple hypothesis testing.

Table 31 Multivariable survival analysis for features derived from the tumour microenvironment

	Multivariable Models with FLIPI	Cox PH Multivariable OS		Cox PH Multivariable PFS	
		All Patients, N=101, 20 events		Rituximab Patients, N=51, 29 events	
		HR (95% CI)	*P**	HR (95% CI)	*P**
Cell Density, cells / mm^2	$CD4^+CD68^-$ T-helper cells	0.872	0.872	1 (1, 1)	0.158
	$CD4^+FOXP3^+$ T-regs	0.96 (0.92, 1)	0.066	0.98 (0.95, 1)	0.109
	$CD8^+$ T-cells	1 (0.99, 1)	0.315	1 (0.99, 1)	0.561
	$CD68^+$ cells	0.99 (0.98, 1)	**0.013**	0.99 (0.99, 1)	**0.046**
	$CD4^+CD68^-PD\text{-}1^+$	1 (0.98, 1.01)	0.478	1 (0.99, 1.01)	0.907
	$CD8^+PD\text{-}1^+$	0.97 (0.94, 1.01)	0.137	1 (0.98, 1.01)	0.613
Cell Ratio	Immune infiltrate ratio†	0.37 (0.07, 2)	0.247	0.35 (0.09, 1.37)	0.131
% Positive Area	$CD21^+$ dendritic meshwork area	0.4 (0.09, 1.69)	0.2115	1.08 (0.25, 4.79)	0.915
Diversity, natural digits	Phenotype entropy	0.19 (0.06, 0.65)	**0.008**	0.85 (0.31, 2.31)	0.750
	Interaction entropy	0.39 (0.2, 0.75)	**0.005**	0.9 (0.53, 1.53)	0.700

Only subset of patients with available FLIPI data at diagnosis is included. HR: hazard ratio; CI: confidence intervals; PH: proportional hazards; OS: overall survival; PFS: progression free survival. *The log rank test p value examines whether the null hypothesis of no effect (H0: HR=1) can be rejected. Features are assessed as continuous variables. †Immune infiltrate ratio is calculated as the total immune cells (positive for any marker) divided by the number of cells that expressed only DAPI. P values < 0.05 are shown in bold. P values < 0.0056 remain significant after the Bonferroni correction for multiple hypothesis testing.

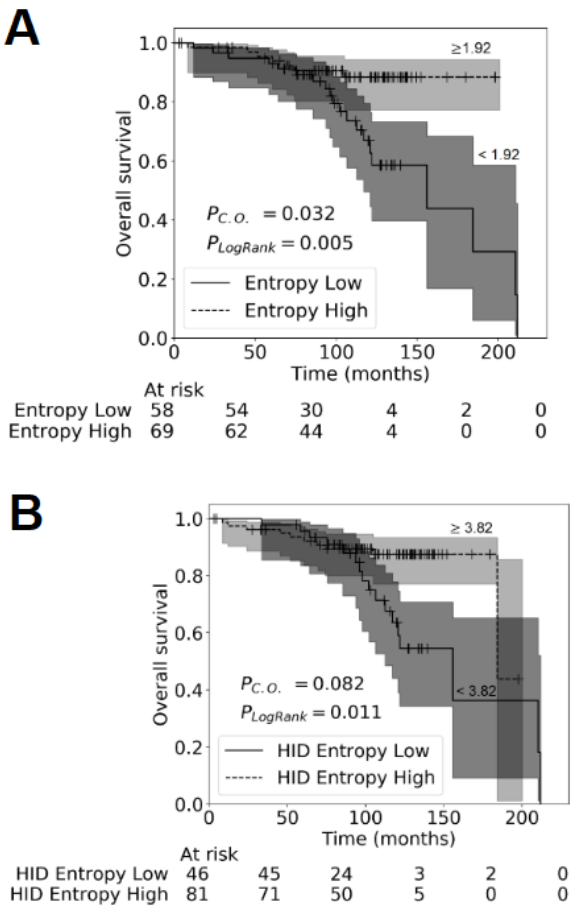

Figure 42 Kaplan-Meier survival analysis for the new diversity metrics. Analysis shown for overall survival (OS), where patients have been split into two groups based on the optimal cut-points, found using the Contal & O'Quigley method.[387] $P_{Log\ Rank}$: significance for the Log rank test and $P_{C.O.}$: significance for the Contal & O'Quigley test[387] adjusted for the fact that the optimal cut-point has been selected to maximise separation of patient groups. A) Effect of phenotype entropy (diversity) on OS. B) Effect of HID spatial "interaction" entropy (diversity) on OS.

Table 32 Logistic regression for POD24 prediction in the subset treated with rituximab containing regimens

	Logistic Regression for POD24	**Univariable (Rituximab patients, N=67 [14 events])**		**Multivariable with FLIPI (Rituximab patients, N=51 [8 events])**	
		OR (95% CI)	***P****	**OR (95% CI)**	***P****
Cell Density, cells / mm^2	$CD4^+CD68^-$ T-helper cells	0.99 (0.99, 1)	**0.027**	0.99 (0.98, 1)	**0.034**
	$CD4^+FOXP3^+$ T-regs	0.95 (0.9, 1)	0.066	0.95 (0.88, 1.02)	0.132
	$CD8^+$ T-cells	1 (0.99, 1.01)	0.465	1 (0.99, 1.01)	0.638
	$CD68^+$ cells	0.99 (0.98, 1)	0.051	0.98 (0.97, 1)	0.063
	$CD4^+CD68^-PD\text{-}1^+$	0.98 (0.96, 1.01)	0.116	0.96 (0.92, 1.01)	0.100
	$CD8^+PD\text{-}1^+$	0.98 (0.95, 1.02)	0.389	0.97 (0.92, 1.04)	0.410
Cell Ratio	Immune infiltrate ratio†	0.02 (0, 0.48)	**0.017**	0.01 (0, 1.23)	0.060
% Positive Area	$CD21^+$ dendritic meshwork area	0.28 (0.02, 3.65)	0.334	0.06 (0, 3.14)	0.166
Diversity, natural digits	Phenotype entropy	0.73 (0.13, 4.07)	0.718	0.64 (0.08, 5.07)	0.669
	Interaction entropy	0.82 (0.33, 2.02)	0.665	0.75 (0.25, 2.27)	0.610

Only subset of patients with available FLIPI data at diagnosis is included in multivariable analysis and features treated as continuous variables. *The log rank test p value examines whether the null hypothesis of no effect (H0: Odds ratio=1) can be rejected. †Immune infiltrate ratio is calculated as the total immune cells (positive for any marker) divided by the number of cells that expressed only DAPI. Features are assessed as continuous variables. P values < 0.05 are shown in bold. P values < 0.005 would remain significant after the Bonferroni correction for multiple hypothesis testing.

4.2.4 Discussion

This study introduced a 6-plex immunofluorescence protocol for concurrent observation of immune subsets and an image analysis pipeline to detect cell types and objectively measure tumour microenvironment diversity. This new approach provides a versatile and adaptable platform that could be extended to other tumour types. The proposed pipeline benefits from precise marker localisation as well as conservation of valuable tissue material through multiplexing. The improved accuracy and reliability of quantitative immunofluorescence compared to conventional immunohistochemistry, and its cost-effectiveness compared to *in-situ* hybridisation, provide scope and rationale for wider clinical adoption.

Developing baseline prognostic biomarkers for risk stratification is a major area of research in FL, driven by an urgent need to develop effective therapies capable of improving the outcomes, especially for patients with of high-risk disease. Using this pipeline, this analysis reports that increased diversity of immune infiltrate populations and interactions in FL are potential biomarkers of favourable OS. Diversity was quantified through a novel approach using Shannon's entropy, a metric describing species biodiversity in ecological sciences. The diversity of spatial interactions remained significant after Bonferroni correction for multiple comparisons in multivariable analysis of OS. Therefore, this biomarker could improve risk stratification, offering additional prognostic value to FLIPI. The diversity biomarkers also outperformed simple cell density measurements. Indeed, none of the immune infiltrate cell densities remained significantly associated with survival endpoints in multivariable analysis after Bonferroni correction (**Table 31**), similar to results reported by others[65] for rituximab treated patients. This evidence supports applicability of the diversity biomarker for risk stratification in FL. Furthermore, a favourable trend was observed between increased total immune infiltrate ratio and improved OS, PFS and fewer POD24 events in all univariable analyses, indicating that both the extent of the immune infiltrate and its diversity may affect FL prognosis.

Increased diversity translates to diffuse and increased expression of multiple lymphocytic and myeloid cell subsets in the microenvironment of FL. Previous studies investigating tumour immune microenvironment diversity have demonstrated the importance of diversity in T-cell populations, as measured by T-cell receptor (TCR) Next Generation sequencing, in a way that is agnostic to the types of T-cells that

are quantified.[388] Increased TCR diversity has been associated with improved clinical benefit in metastatic melanoma,[389] and favourable overall survival in metastatic breast cancer.[390] Furthermore, clonal TCR diversity has been shown to increase after immunotherapy treatments (e.g., cryo-immunotherapy for breast cancer[389] and Sipuleucel-T immunotherapy for prostate cancer[391]) and is investigated as a potential endpoint for response to therapy.[390] A diverse T-cell repertoire is thought to increases the likelihood that a useful anti-tumour T-cell population is present,[389] leading to favourable outcomes. In this study the concept of diversity was expanded to include T-cells and macrophage subsets and propose that a diverse repertoire of immune cells in the microenvironment of FL would similarly increase the likelihood of relevant anti-tumour pathways being active.

In this study, $CD68^+$ macrophages were significantly correlated with favourable OS in univariable analysis, after Bonferroni correction. A favourable trend of increased $CD68^+$ density was observed for PFS and POD24. This effect could be attributed to one of the mechanisms of action of the anti-CD20 rituximab treatment, whose immune-mobilising effects include the induction of antibody-dependent cell phagocytosis.[392] Consequently, cells coated with rituximab are recognised by macrophages as targets and killed.[91] The favourable effect of macrophages has been previously demonstrated in a rituximab treated cohort.[86] However, this effect depends strongly on the type of treatment, as in cohorts treated without rituximab[89,90] increased numbers of tumour associated macrophages correlated with unfavourable outcome.

To ensure reproducibility of results, the staining assay and cell detection algorithms were quantitatively validated, and the image dataset is shared publicly[384]. The FL cohort included treatment pathways and prognostic outcomes reflective of current modern practice. The TMA technology employed is equivalent to whole section assessments in lymphomas[393], enabling rapid processing of large number of samples. Furthermore, the diversity metrics demonstrated low intra-patient heterogeneity (CoV = 7.7-8.3%), indicating robustness when assessed using triplicate TMA core samples.

A limitation of this study is the use of a single cut-off to score positive and negative cells for each stain. Robust cut-offs were selected by two different users of the computer assisted scoring system. However, this approach may sometimes underper-

form because of the inherent variation of staining intensities in positive cells. Notably, in FL two functionally different PD-1^{+} cell phenotypes have been observed,[81] characterised by different levels of PD-1 expression: PD-1$^{+}_{high}$ T follicular helper cells found inside the follicles actively support FL B-cell growth, while the PD-1$^{+}_{low}$ cells found outside the follicles represent exhausted T-cells. The PD-1^{+} T-helper cells found within the follicles are also known to express CD4 less strongly (30.7% lower CD4 intensity) compared to other CD4^{+} cells in the interfollicular areas.[85] The present study attempted to select single cut-offs able to pick up both the dim and bright positive cells. Use of multiple cut-offs was avoided as nuanced intensity variations can be challenging to capture using manual gating in a reproducible manner. An alternative scoring approach could adopt automated clustering[219] of cells based on their intensities, or rely on additional functional markers in the multiplex panel (e.g., TIM3 for exhausted phenotypes or CXCR5 for T follicular helper cells[81]) to differentiate between PD-1 subsets.

Analysis of PFS and POD24 in the rituximab-treated subset did not demonstrate significant prognostic value for any of the immune infiltrate biomarkers, after corrections for multiple hypothesis testing. However, the limited size and variable treatment increase the risk of false negative results. Therefore, the effect of tumour microenvironment diversity on early relapse merits further investigation before it could be ruled out.

Similar to the biomarker described in chapter 3, the automated pipeline to identify cell subsets and measure diversity in FL can also be considered as an interactive, definable and interpretable CAS system, whose accuracy in this study is validated against patient clinical endpoints (OS, PFS, POD24). Following the design requirements identified in chapter 2, further work before clinical adoption should validate reproducibility under variable staining conditions, time-efficiency and establish a method to identify samples that are challenging (i.e., have high uncertainty). Future work may also involve validation of diversity measurements using orthogonal assays, such as gene expression profiling.

In summary, automated assessment of immune infiltrate diversity, based on multiplex immunofluorescence, warrants further exploration as a prognostic biomarker in FL. This pipeline is ready to be tested in larger series, with the potential to significantly improve risk stratification and risk-adapted treatment for FL in the future.

4.3 Summary

Follicular lymphoma (FL) prognosis is influenced by the composition of the tumour microenvironment. An automated approach to quantitatively assess the phenotypic and spatial immune infiltrate diversity was tested as a prognostic biomarker for FL patients.

Diagnostic biopsies were collected from 127 FL patients initially treated with rituximab-based therapy (52%), radiotherapy (20%) or active surveillance (28%). Tissue microarrays were constructed and stained using multiplex immunofluorescence (CD4, CD8, FOXP3, CD21, PD-1, CD68 and DAPI). Subsequently, sections underwent automated cell scoring and analysis of spatial interactions, defined as cells co-localising within 30 μm. Shannon's entropy, a metric describing species biodiversity in ecological habitats, was applied to quantify immune infiltrate diversity of cell types and spatial interactions. Immune infiltrate diversity indices were tested in multivariable Cox regression and Kaplan-Meier analysis for overall (OS) and progression free survival (PFS).

Increased diversity of cell types (HR=0.19 95% CI 0.06-0.65, p=0.008) and cell spatial interactions (HR=0.39, 95% CI 0.20-0.75, p=0.005) were associated with favourable OS, independent of the Follicular Lymphoma International Prognostic Index. In the rituximab treated subset, the effect of diversity on PFS did not reach statistical significance.

Multiplex immunofluorescence and Shannon's entropy can objectively quantify immune infiltrate diversity and generate prognostic information in FL. This automated approach warrants validation in additional FL cohorts and its applicability as a pre-treatment biomarker to identify high risk patients should be further explored. The multiplex image dataset generated by this study is shared publicly to encourage further research on the FL microenvironment.

5 Conclusions

This thesis sought to establish multiplex assays and image analysis methodologies that would enable a clinically meaningful quantification of the tumour microenvironment. The key contributions are summarised below.

Computer assisted scoring (CAS) tools are essential for the analysis of multiplex images. The first contribution of this work was a comprehensive description of design requirements for CAS tools. Although requirements have been previously described for manual scoring systems,[217,238] to the best of the author's knowledge, the equivalent requirements for CAS are still unclear. These requirements and validation practices were identified through systematic review of CAS validation for HER2, ER and three T-cell marker assessment in the past 20 years. Scoring is a prerequisite for numerous routine histopathology analyses and could greatly benefit from the introduction of objective, computer assisted tools that increase reproducibility and throughput. The newly identified design requirements provide a guide to judge the performance of new algorithms and determine whether use of an automated scoring tool is equivalent or superior to the standard-of-care, manual scoring system. The requirements outline that CAS algorithms should be well defined, provide accurate patient categorisation, be reproducible, time-efficient, interpretable and able to accurately estimate confidence in their predictions. Furthermore, a meta-analysis of agreement between automated and manual scoring determined the overall accuracy of CAS technology for assessment of HER2 and ER. Assessment of these markers has significant clinical implications for selecting adjuvant treatments in breast cancer and determining prognosis. CAS agreement with manual scoring was similar to inter-observer agreement between pathologists in manual scoring.

The second contribution of this thesis was the development of CAS pipelines for the analysis of multiplex images in OPSCC and FL. CAS is required for the analysis and identification of cell phenotypes in such images, as the high number of stains visualised concurrently cannot be interpreted manually at large scale. A step-by-step comparison of the two pipelines developed for OPSCC and FL are presented in **Table 33**.

Table 33 Comparison of image analysis CAS pipelines developed for OPSCC and FL

Step	*OPSCC CAS tool*	*FL CAS tool*
Useful tissue area identification	Supervised classification (U-net) to identify background, artefacts and useful tissue	- Automated thresholding to find tissue area - Manual identification and removal of artefacts
Outlining nuclear shapes (segmentation)	Label-free watershed	Supervised nuclear segmentation (StarDist)
Outlining membranes	Size constrained region growing around the nucleus	Size constrained region growing around the nucleus
Cell classification	- Median intensity of each stain in the relevant cellular compartment was measured - Positivity cut-off for each stain manually selected by one user of the CAS system	- Median intensity of each stain in the relevant cellular compartment was measured - Positivity cut-off for each stain manually selected by two users of the CAS system

Both CAS tools are interactive, as they require some input from a trained expert to select positivity cut-offs and identify positive cells. They follow similar steps, with some noted differences. The OPSCC algorithm included a fully automated step for exclusion of artefactual and background image areas, while the FL algorithm relied on manual artefact removal. For analyses of subsequent FL cohorts at a larger scale, this step could be fully automated with the same approach used for OPSCC. Furthermore, the label-free nuclear segmentation approach used for OPSCC was not adopted for outlining the nuclear shapes in FL samples, where the nuclei were densely packed and often overlapping. Instead, a state-of-the-art supervised segmentation approach using a deep convolutional neural network (StarDist) was employed for FL samples.

In this work, the only design requirement (from the requirements identified in Chapter 2) that was explicitly validated for the new CAS tools was accuracy. In both CAS tools, the accuracy of the nuclear segmentation steps was validated against manual nuclear shape annotations. Furthermore, the analytical validity of these

CAS tools in terms of accuracy in patient categorisation was validated against patient clinical endpoints: overall survival in OPSCC and overall, progression-free survival and POD24 in FL.

Using the developed CAS system, the third contribution of this work was the discovery of a new biomarker in OPSCC, and analysis of cellular interactions in the tumour microenvironment using the HID methodology. Frequent spatial proximity between cell types that are known to interact as part of the PD-1/PD-L1 immune escape pathway was identified as an adverse prognostic biomarker in HPV negative OPSCC. This biomarker was able to successfully identify at the time of diagnosis the patients who will have poor overall survival.

The last contribution was the proposal of a new biomarker in FL, based on the phenotypic and spatial interaction diversity of the tumour microenvironment. By examining the follicular lymphoma microenvironment as a whole, this biomarker was able to accurately predict overall survival at baseline. The diversity biomarkers outperformed other microenvironment indicators and provided prognostic value independently of the standard-of-care prognostic index (FLIPI). Additional research outcomes from this work were the development and validation of a new multiplex immunofluorescent assay for the observation of six antigen targets concurrently on FL tissue, and an image analysis pipeline for characterisation tumour microenvironment diversity that could be extended for any type of cancer.

The new biomarkers based on spatial interaction analysis in OPSCC and diversity analysis of the tumour microenvironment in FL are ready to be validated in additional cohorts, offering potential for improved risk stratification of patients at baseline diagnosis. Compared to other clinical indicators (e.g., age, stage, FLIPI), biomarkers based on multiplexed spatial profiling of the tumour microenvironment offer additional insights and can capture information on significant cell interactions that affect prognosis, thus improving precision in clinical decision making. Other technologies to describe the tumour microenvironment, such as RNA signatures,[394] have sparked interest in recent years. Multiplex immunofluorescence methods using tissue microarray technology offer a cost-effective alternative to RNA assays, with the added benefit of preserving the spatial context.

Contrary to other tumour microenvironment assays, the adoption of multiplex spatial biomarkers is conditional upon the wider clinical adoption of digital pathology

and CAS tools in everyday practice. Even though this presents significant challenges, progress is steadily being made in this direction. Therefore, we may soon re-imagine the pathology workflow as fully digital, with information from multiple antigens available at once to the pathologist, along with several automatically calculated phenotype, spatial and diversity indicators to provide quantitative insights. This workflow could increase throughput and precision in clinical decision making across manifold applications.

5.1 CAS design requirements as a guide for further validation

Further validation of the tumour microenvironment biomarkers introduced in this thesis will be required for translation to clinically meaningful applications. The automated image analysis pipelines used to produce the biomarker scores can be seen as interactive CAS systems and therefore would need to demonstrate satisfaction of the design requirements outlined in Chapter 2. The systems developed in this thesis are well defined and interpretable, and accuracy in patient categorisation was validated against clinical survival endpoints. However further work is needed to a) validate accuracy of patient categorisation in independent cohorts and/or against orthogonal assays, b) assess reproducibility, c) time efficiency and d) develop accurate methodologies to estimate confidence. These steps are described below:

a) A limitation of the present analysis is the limited size of cohorts in OPSCC and FL, which may limit statistical power. The accuracy of patient categorisation based on the proposed biomarkers would need validation in larger, independent cohorts. Homogeneously treated cohorts would be necessary to validate these biomarkers as predictive, enabling their use for selection of appropriate treatments.

   Furthermore, accuracy of cell phenotyping may be improved by implementation of a single-cell classifier using label-free clustering approaches, instead of adopting a simplistic stain intensity cut-off to indicate positive/ negative cells. Accuracy of the cell density measurements could be further validated by demonstrating agreement with orthogonal assays, such as other multiplex *in situ* technologies (e.g., imaging mass cytometry), or alternative non *in situ* methods (e.g., flow cytometry). Accuracy may also be tested against manual scoring. The analytical validity and reproducibility of any

orthogonal methodologies used as reference would need to have been established in advance.

b) Reproducibility assessment would include demonstrating good intra and inter-observer agreement for different users of the CAS tools, as well as repeatability of scores despite variability in staining and slide scanning conditions. For this purpose, staining and scoring would be repeated for different runs and operators and/or in different labs.

c) Time efficiency would be judged based on the time window available to the physician to perform the scoring assessment. It would also depend on the computational resources available, and the time needed for staining and scanning the samples.

d) Finally, confidence estimation methodologies could be developed to identify challenging samples. Simple approaches may identify problematic samples with high numbers of artefacts or blurriness, indicate when staining intensities lie outside of an accepted range, or observe when there is high disagreement between different images of the same patient (different regions of interest or TMA cores). Alternatively, more complex Bayesian statistical frameworks could be employed during cell segmentation and classification, e.g., use of Bayesian deep neural networks[395] to provide an indication of uncertainty. Confidence estimation could assist in preventing "silent" failure of the automated image analysis pipeline. Achieving validation of these design requirements would facilitate clinical adoption of this technology.

5.2 Future work

The present work undertook initial exploration for tumour microenvironment biomarker discovery in OPSCC and FL. However, the OPSCC biomarker that observes proximity between cells interacting in the PD-1/PD-L1 pathway would be relevant for other cancers where this pathway of immune escape is present. Furthermore, the usefulness of the diversity biomarker could be explored in other lymphomas and solid tumours.

Overall, multiplexed spatial profiling of the tumour microenvironment merits further exploration to advance prognostic insights across multiple types of cancer.

Bibliography

1. Tabassum DP, Polyak K. Tumorigenesis: it takes a village. *Nat Rev Cancer*. 2015;15(8):473-483. doi:10.1038/nrc3971

2. Reilly KM, Van Dyke T. It takes a (dysfunctional) village to raise a tumor. *Cell*. 2008;135(3):408-410.

3. Heindl A, Nawaz S, Yuan Y. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Lab Investig*. 2015;95(4):377-384. doi:10.1038/labinvest.2014.155

4. Myers E, Suen J, Myers J, Hanna E. *Cancer of the Head and Neck*. 4th ed. Saunders; 1989.

5. Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. *Oral Oncol*. 2009;45(4-5):309-316. doi:10.1016/j.oraloncology.2008.06.002

6. Head and neck cancers statistics. Cancer Research UK. Accessed February 1, 2021. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/head-and-neck-cancers

7. Gillison ML, Chaturvedi AK, Anderson WF, Fakhry C. Epidemiology of human papillomavirus-positive head and neck squamous cell carcinoma. *J Clin Oncol*. 2015;33(29):3235-3242. doi:10.1200/JCO.2015.61.6995

8. Ndiaye C, Mena M, Alemany L, et al. HPV DNA, E6/E7 mRNA, and p16INK4adetection in head and neck cancers: A systematic review and meta-analysis. *Lancet Oncol*. 2014;15(12):1319-1331. doi:10.1016/S1470-2045(14)70471-1

9. Specenier P, Vermorken JB. Cetuximab: Its unique place in head and neck cancer treatment. *Biol Targets Ther*. 2013;7(1):77-90. doi:10.2147/BTT.S43628

10. Mehra R, Cohen RB, Burtness BA. The role of cetuximab for the treatment of squamous cell carcinoma of the head and neck. *Clin Adv Hematol Oncol*. 2008;6(10):742-750.

11. Economopoulou P, Perisanidis C, Giotakis EI, Psyrri A. The emerging role of immunotherapy in head and neck squamous cell carcinoma (HNSCC): anti-tumor immunity and clinical applications. *Ann Transl Med*. 2016;4(9):173. doi:10.21037/atm.2016.03.34

12. Ran X, Yang K. Inhibitors of the PD-1/PD-L1 axis for the treatment of head and neck cancer: Current status and future perspectives. *Drug Des Devel Ther*. 2017;11:2007-2014. doi:10.2147/DDDT.S140687

13. Ragin CCR, Taioli E. Survival of squamous cell carcinoma of the head and neck in relation to human papillomavirus infection: Review and meta-analysis. *Int J Cancer*. 2007;121(8):1813-1820. doi:10.1002/ijc.22851

14. Ward MJ, Thirdborough SM, Mellows T, et al. Tumour-infiltrating lymphocytes predict for outcome in HPV-positive oropharyngeal cancer. *Br J Cancer*. 2014;110(2):489-500. doi:10.1038/bjc.2013.639

15. Nordsmark M, Bentzen SM, Rudat V, et al. Prognostic value of tumor oxygenation in 397 head and neck tumors after primary radiation therapy. An international multi-center study. *Radiother Oncol*. 2005;77(1):18-24.

16. Alsahafi E, Begg K, Amelio I, et al. Clinical update on head and neck cancer: molecular biology and ongoing challenges. *Cell Death Dis*. 2019;10(8):1-17.

17. Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. *Science (80-)*. 2015;347(6220):1260419. www.proteinatlas.org

18. Mattox A, Lee J, Westra WH, et al. PD-1 expression in head and neck squamous cell carcinomas derives primarily from functionally anergic CD4+ TILs in the presence of PD-L1+ TAMs. *Cancer Res*. Published online 2017:canres.3453.2016. doi:10.1158/0008-5472.CAN-16-3453

19. Müller T, Braun M, Dietrich D, et al. PD-L1: a novel prognostic biomarker in head and neck squamous cell carcinoma. *Oncotarget*. 2017;8(32):52889-52900. doi:10.18632/oncotarget.17547

20. Skinner HD, Giri U, Yang LP, et al. Integrative analysis identifies a novel AXL-PI3 kinase-PD-L1 signaling axis associated with radiation resistance in head and neck cancer. *Clin Cancer Res*. 2017;23(11):2713-2722. doi:10.1158/1078-0432.CCR-16-2586

21. Schneider S, Kadletz L, Wiebringhaus R, et al. PD-1 and PD-L1 expression in HNSCC primary cancer and related lymph node metastasis - impact on clinical outcome. *Histopathology*. 2018;73(4):573-584. doi:10.1111/his.13646

22. Oguejiofor K, Galletta-Williams H, Dovedi SJ, Roberts DL, Stern PL, West CML. Distinct patterns of infiltrating CD8 + T cells in HPV + and CD68 macrophages in HPV - oropharyngeal squamous cell carcinomas are associated with better clinical outcome but PD-L1 expression is not prognostic. *Oncotarget*. 2017;8(9):14416-14427. doi:10.18632/oncotarget.14796

23. Ritta M, De Andrea M, Mondini M, et al. Cell cycle and viral and immunologic profiles of head and neck squamous cell carcinoma as predictable variables of tumor progression. *Head Neck*. 2009;31(3):318-327.

24. Russell S, Angell T, Lechner M, et al. Immune cell infiltration patterns and survival in head and neck squamous cell carcinoma. *Head Neck Oncol*. 2013;5(3):24. doi:10.1016/j.drugalcdep.2008.02.002.A

25. Pretscher D, Distel L V., Grabenbauer GG, Wittlinger M, Buettner M, Niedobitek G. Distribution of immune cells in head and neck cancer: CD8+T-cells and CD20+B-cells in metastatic lymph nodes are associated with favourable outcome in patients with oro- and hypopharyngeal carcinoma. *BMC Cancer*. 2009;9(1):292. doi:10.1186/1471-2407-9-292

26. Oguejiofor K, Hall J, Slater C, et al. Stromal infiltration of CD8 T cells is associated with improved clinical outcome in HPV-positive oropharyngeal squamous carcinoma. *Br J Cancer*. 2015;113(6):886-893. doi:10.1038/bjc.2015.277

27. Wansom D, Light E, Thomas D, et al. Infiltrating lymphocytes and human papillomavirus-16-associated oropharyngeal cancer. *Laryngoscope*. 2012;122(1):121-127. doi:10.1002/lary.22133

28. Näsman A, Romanitan M, Nordfors C, et al. Tumor infiltrating CD8 + and Foxp3 + Lymphocytes correlate to clinical outcome and human papillomavirus (HPV) status in Tonsillar cancer. *PLoS One*. 2012;7(6):1-8. doi:10.1371/journal.pone.0038711

29. Feng Z, Bethmann D, Kappler M, et al. Multiparametric immune profiling in HPV oral squamous cell cancer. *JCI Insight*. 2017;2(14). doi:10.1172/jci.insight.93652

30. Peltanova B, Raudenska M, Masarik M. Effect of tumor microenvironment on pathogenesis of the head and neck squamous cell carcinoma: A systematic review. *Mol Cancer*. 2019;18(1):1-24. doi:10.1186/s12943-019-0983-5

31. Scognamiglio T, Chen Y-T. Beyond the Percentages of PD-L1-Positive Tumor Cells: Induced Versus Constitutive PD-L1 Expression in Primary and Metastatic Head and Neck Squamous Cell Carcinoma. *Head Neck Pathol*. 2018;12(2):221-229. doi:10.1007/s12105-017-0857-3

32. Cancer Research UK - Non-Hodgkin lymphoma statistics. Accessed January 15, 2021. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/non-hodgkin-lymphoma#heading-Zero

33. Smith A, Howell D, Patmore R, Jack A, Roman E. Incidence of haematological malignancy by sub-type: A report from the Haematological Malignancy Research Network. *Br J Cancer*. 2011;105(11):1684-1692. doi:10.1038/bjc.2011.450

34. Cerhan JR. Epidemiology of Follicular Lymphoma. *Hematol Oncol Clin North Am*. 2020;34(4):631-646. doi:10.1016/j.hoc.2020.02.001

35. Smith A, Crouch S, Lax S, et al. Lymphoma incidence, survival and prevalence 2004-2014: Sub-type analyses from the UK's Haematological Malignancy Research Network. *Br J Cancer*. 2015;112(9):1575-1584. doi:10.1038/bjc.2015.94

36. Relander T, Johnson NA, Farinha P, Connors JM, Sehn LH, Gascoyne RD. Prognostic factors in follicular lymphoma. *J Clin Oncol*. 2010;28(17):2902-2913. doi:10.1200/JCO.2009.26.1693

37. Harris N, Ferry J, Carbone A, et al. Follicular lymphoma. *Nat Rev Dis Prim*. 2019;5(1):83. doi:10.1038/s41572-019-0132-x

38. Batlevi CL, Sha F, Alperovich A, et al. Follicular lymphoma in the modern era: survival, treatment outcomes, and identification of high-risk subgroups. *Blood Cancer J*. 2020;10(7). doi:10.1038/s41408-020-00340-z

39. Cheson BD, Fisher RI, Barrington SF, et al. Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-hodgkin lymphoma: The lugano classification. *J Clin Oncol*. 2014;32(27):3059-3067. doi:10.1200/JCO.2013.54.8800

40. Welaya K, Casulo C. Follicular Lymphoma: Redefining Prognosis, Current Treatment Options, and Unmet Needs. *Hematol Oncol Clin North Am*. 2019;33(4):627-638. doi:10.1016/j.hoc.2019.03.003

41. Brice P, Bastion Y, Lepage E, et al. Comparison in low-tumor-burden follicular lymphomas between an initial no-treatment policy, prednimustine, or interferon alfa: a randomized study from the Groupe d'Etude des Lymphomes Folliculaires. Groupe d'Etude des Lymphomes de l'Adulte. *J Clin Oncol*. 1997;15(3):1110-1117.

42. Rummel MJ, Niederle N, Maschmeyer G, et al. Bendamustine plus rituximab versus CHOP plus rituximab as first-line treatment for patients with indolent and mantle-cell lymphomas: An open-label, multicentre, randomised, phase 3 non-inferiority trial. *Lancet*. 2013;381(9873):1203-1210. doi:10.1016/S0140-6736(12)61763-2

43. Flinn IW, Van Der Jagt R, Kahl BS, et al. Open-label, randomized, noninferiority study of bendamustine-rituximab or R-CHOP/R-CVP in first-line treatment of advanced indolent NHL or MCL: the BRIGHT study. *Blood*. 2014;123(19):2944-2952.

44. Hiddemann W, Barbui AM, Canales MA, et al. Immunochemotherapy with

obinutuzumab or rituximab for previously untreated follicular lymphoma in the GALLIUM study: influence of chemotherapy on efficacy and safety. *J Clin Oncol*. 2018;36(23):2395-2404.

45. Morschhauser F, Fowler NH, Feugier P, et al. Rituximab plus lenalidomide in advanced untreated follicular lymphoma. *N Engl J Med*. 2018;379(10):934-947.

46. Al-Tourah AJ, Gill KK, Chhanabhai M, et al. Population-based analysis of incidence and outcome of transformed non-Hodgkin's lymphoma. *J Clin Oncol*. 2008;26(32):5165-5169.

47. Project IN-HLPF. A predictive model for aggressive non-Hodgkin's lymphoma. *N Engl J Med*. 1993;329(14):987-994.

48. Lopez-Guillermo A, Montserrat E, Bosch F, Terol MJ, Campo E, Rozman C. Applicability of the International Index for aggressive lymphomas to patients with low-grade lymphoma. *J Clin Oncol*. 1994;12(7):1343-1348.

49. Solal-Céligny P, Roy P, Colombat P, et al. FLIPI: Follicular Lymphoma International Prognostic Index. *Blood*. 2004;104(5):1258-1265. doi:10.1182/blood-2003-12-4434

50. Buske C, Hoster E, Dreyling M, Hasford J, Unterhalt M, Hiddemann W. The Follicular Lymphoma International Prognostic Index (FLIPI) separates high-risk from intermediate- or low-risk patients with advanced-stage follicular lymphoma treated front-line with rituximab and the combination of cyclophosphamide, doxorubicin, vinc. *Blood*. 2006;108(5):1504-1508. doi:10.1182/blood-2006-01-013367

51. Federico M, Bellei M, Marcheselli L, et al. Follicular Lymphoma International Prognostic Index 2: A New Prognostic Index for Follicular Lymphoma Developed by the International Follicular Lymphoma Prognostic Factor Project. *J Clin Oncol*. 2009;27(27):4555-4562. doi:10.1200/JCO.2008.21.3991

52. Bachy E, Maurer MJ, Habermann TM, et al. A simplified scoring system in de novo follicular lymphoma treated initially with immunochemotherapy. *Blood*. 2018;132(1):49-58.

53. Pastore A, Jurinovic V, Kridel R, et al. Integration of gene mutations in risk prognostication for patients receiving first-line immunochemotherapy for follicular lymphoma: a retrospective analysis of a prospective clinical trial and validation in a population-based registry. *Lancet Oncol*. 2015;16(9):1111-1122.

54. Meignan M, Cottereau AS, Versari A, et al. Baseline metabolic tumor volume predicts outcome in high–tumor-burden follicular lymphoma: a pooled analysis of three multicenter studies. *J Clin Oncol*. 2016;34(30):3618-3626.

55. Cottereau AS, Versari A, Luminari S, et al. Prognostic model for high-tumor-burden follicular lymphoma integrating baseline and end-induction PET: a LYSA/FIL study. *Blood*. 2018;131(22):2449-2453.

56. Casulo C, Byrtek M, Dawson KL, et al. Early relapse of follicular lymphoma after rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone defines patients at high risk for death: An analysis from the National LymphoCare Study. *J Clin Oncol*. 2015;33(23):2516-2522. doi:10.1200/JCO.2014.59.7534

57. Huet S, Szafer-Glusman E, Xerri L, et al. Evaluation of clinicogenetic risk models for outcome of follicular lymphoma patients in the PRIMA trial. *Hematol Oncol*. 2017;35:96-97.

58. Husson H, Carideo EG, Neuberg D, et al. Gene expression profiling of follicular lymphoma and normal germinal center B cells using cDNA arrays. *Blood*. 2002;99(1):282-289. doi:10.1182/blood.V99.1.282

59. Glas AM, Kersten MJ, Delahaye LJMJ, et al. Gene expression profiling in follicular lymphoma to assess clinical aggressiveness and to guide the choice of treatment. *Blood*. 2005;105(1):301-307. doi:10.1182/blood-2004-06-2298

60. Sandeep S. Dave, M.D., George Wright, Ph.D., Bruce Tan, M.D., Andreas Rosenwald MD, Randy D. Gascoyne, M.D., Wing C. Chan, M.D., Richard I. Fisher, M.D., Rita M. Braziel MD, Lisa M. Rimsza, M.D., Thomas M. Grogan, M.D., Thomas P. Miller, M.D., Michael LeBlanc PD, et al. Prediction of Survival in Follicular Lymphoma Based on Molecular Features of Tumor-Infiltrating Immune Cells. *N Engl J Med*. 2004;367(14):1287-1296. doi:10.1056/NEJMoa1208410

61. Johnson PWM, Watt SM, Betts DR, et al. Isolated follicular lymphoma cells are resistant to apoptosis and can be grown in vitro in the CD40/stromal cell system. *Blood*. 1993;82(6):1848-1857.

62. Amé-Thomas P, Tarte K. The yin and the yang of follicular lymphoma cell niches: Role of microenvironment heterogeneity and plasticity. *Semin Cancer Biol*. 2014;24:23-32. doi:10.1016/j.semcancer.2013.08.001

63. Solal-Céligny P, Cahu X, Cartron G. Follicular lymphoma prognostic factors in the modern era: What is clinically meaningful? *Int J Hematol*. 2010;92(2):246-254. doi:10.1007/s12185-010-0674-x

64. De Jong D, Fest T. The microenvironment in follicular lymphoma. *Best Pract Res Clin Haematol*. 2011;24(2):135-146. doi:10.1016/j.beha.2011.02.007

65. Xerri L, Huet S, Venstrom JM, et al. Rituximab treatment circumvents the prognostic impact of tumor-infiltrating T-cells in follicular lymphoma patients. *Hum Pathol*. 2017;64:128-136. doi:10.1016/j.humpath.2017.03.023

66. Carbone A, Gloghini A, Cabras A, Elia G. The Germinal centre-derived lymphomas seen through their cellular microenvironment. *Br J Haematol*. 2009;145(4):468-480. doi:10.1111/j.1365-2141.2009.07651.x

67. El Daker S, Gao Q, Roshal M, Dogan A. CD4 Lymphocytes and T Regulatory Cells Show Distinct Phenotypes in Follicular Lymphoma and Classical Hodgkin Lymphoma Which May Account for Differences in Responses to Checkpoint Therapy. Published online 2019.

68. Lee AM, Clear AJ, Calaminici M, et al. 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*. 2006;24(31):5052-5059. doi:10.1200/JCO.2006.06.4642

69. Byers RJ, Sakhinia E, Joseph P, et al. Clinical quantitation of immune signature in follicular lymphoma by RT-PCR-based gene expression profiling. *Blood*. 2008;111(9):4764-4770. doi:10.1182/blood-2007-10-115915

70. Wahlin BE, Sundström C, Holte H, et al. T cells in tumors and blood predict outcome in follicular lymphoma treated with rituximab. *Clin Cancer Res*. 2011;17(12):4136-4144. doi:10.1158/1078-0432.CCR-11-0264

71. Blaker YN, Spetalen S, Brodtkorb M, et al. The tumour microenvironment influences survival and time to transformation in follicular lymphoma in the rituximab era. *Br J Haematol*. 2016;175(1):102-114. doi:10.1111/bjh.14201

72. Lim HW, Hillsamer P, Banham AH, Kim CH. Cutting edge: direct suppression of B cells by CD4+ CD25+ regulatory T cells. *J Immunol*. 2005;175(7):4180-4183. doi:10.4049/jimmunol.175.7.4180

73. Lim HW, Hillsamer P, Kim CH. Regulatory T cells can migrate to follicles upon T cell activation and suppress GC-Th cells and GC-Th cell – driven B cell responses. *J Clin Invest*. 2004;114(11):1640-1649. doi:10.1172/JCI200422325.1640

74. Carreras J, López-Guillermo A, Fox BC, et al. High numbers of tumor-infiltrating FOXP3-positive regulatory T cells are associated with improved overall survival in follicular lymphoma. *Blood*. 2006;108(9):2957-2964. doi:10.1182/blood-2006-04-018218.E.C.

75. Tzankov A, Meier C, Hirschmann P, Went P, Pileri SA, Dirnhofer S. Correlation of high numbers of intratumoral FOXP3+ regulatory T cells with improved survival in germinal center-like diffuse large B-cell lymphoma, follicular lymphoma and classical Hodgkin's lymphoma. *Haematologica*. 2008;93(2):193-200. doi:10.3324/haematol.11702

76. Sweetenham JW, Goldman B, LeBlanc ML, et al. Prognostic value of regulatory T cells, lymphoma associated macrophages, and MUM-1 expression in follicular lymphoma treated before and after the introduction of monoclonal antibody therapy: A southwest oncology group study. *Ann Oncol*. 2009;21(6):1196-1202. doi:10.1093/annonc/mdp460

77. Farinha P, Al-Tourah A, Gill K, Klasa R, Connors JM, Gascoyne RD. The architectural pattern of FOXP3-positive T cells in follicular lymphoma is an independent predictor of survival and histologic transformation. *Blood*. 2010;115(2):289-295. doi:10.1182/blood-2009-07-235598

78. Álvaro-Naranjo T, Lejeune M, Salvadó MT, et al. Immunohistochemical patterns of reactive microenvironment are associated with clinicobiologic behavior in follicular lymphoma patients. *J Clin Oncol*. 2006;24(34):5350-5357. doi:10.1200/JCO.2006.06.4766

79. Wahlin BE, Sander B, Christensson B, Kimby E. CD8+ T-cell content in diagnostic lymph nodes measured by flow cytometry is a predictor of survival in follicular lymphoma. *Clin Cancer Res*. 2007;13(2 I):388-397. doi:10.1158/1078-0432.CCR-06-1734

80. Laurent C, Müller S, Do C, et al. Distribution, function, and prognostic value of cytotoxic T lymphocytes in follicular lymphoma: a 3-D tissue-imaging study. *Blood*. 2011;118(20):5371-5379. doi:10.1182/blood-2011-

81. Yang ZZ, Grote DM, Ziesmer SC, Xiu B, Novak AJ, Ansell SM. PD-1 expression defines two distinct T-cell sub-populations in follicular lymphoma that differentially impact patient survival. *Blood Cancer J*. 2015;5(2):e281-10. doi:10.1038/bcj.2015.1

82. Richendollar BG, Pohlman B, Elson P, Hsi ED. Follicular programmed death 1-positive lymphocytes in the tumor microenvironment are an independent prognostic factor in follicular lymphoma. *Hum Pathol*. 2011;42(4):552-557. doi:10.1016/j.humpath.2010.08.015

83. Carreras J, Lopez-Guillermo A, Roncador G, et al. High numbers of tumor-infiltrating programmed cell death 1-positive regulatory lymphocytes are associated with improved overall survival in follicular lymphoma. *J Clin Oncol*. 2009;27(9):1470-1476. doi:10.1200/JCO.2008.18.0513

84. Wahlin BE, Aggarwal M, Montes-Moreno S, et al. A unifying microenvironment model in follicular lymphoma: Outcome is predicted by programmed death-1-positive, regulatory, cytotoxic, and helper T cells and macrophages. *Clin Cancer Res*. 2010;16(2):637-650. doi:10.1158/1078-0432.CCR-09-2487

85. Townsend W, Pasikowska M, Yallop D, et al. The architecture of neoplastic follicles in follicular lymphoma; analysis of the relationship between the tumor and follicular helper T cells. *Haematologica*. 2020;105(6):1593-1603. doi:10.3324/haematol.2019.220160

86. Taskinen M, Karjalainen-Lindsberg ML, Nyman H, Eerola LM, Leppä S. A high tumor-associated macrophage content predicts favorable outcome in follicular lymphoma patients treated with rituximab and cyclophosphamide- doxorubicin-vincristine-prednisone. *Clin Cancer Res*. 2007;13(19):5784-5789. doi:10.1158/1078-0432.CCR-07-0778

87. De Jong D, Koster A, Hagenbeek A, et al. Impact of the tumor microenvironment on prognosis in follicular lymphoma is dependent on specific treatment protocols. *Haematologica*. 2009;94(1):70-77. doi:10.3324/haematol.13574

88. Taskinen M, Karjalainen-Lindsberg ML, Leppä S. Prognostic influence of tumor-infiltrating mast cells in patients with follicular lymphoma treated with rituximab and CHOP. *Blood*. 2008;111(9):4664-4667. doi:10.1182/blood-2007-11-125823

89. Canioni D, Salles G, Mounier N, et al. High numbers of tumor-associated macrophages have an adverse prognostic value that can be circumvented by rituximab in patients with follicular lymphoma enrolled onto the GELA-GOELAMS FL-2000 trial. *J Clin Oncol*. 2008;26(3):440-446. doi:10.1200/JCO.2007.12.8298

90. Farinha P, Masoudi H, Skinnider BF, et al. Analysis of multiple biomarkers shows that lymphoma-associated macrophage (LAM) content is an independent predictor of survival in follicular lymphoma (FL). *Blood*. 2005;106(6):2169-2174. doi:10.1182/blood-2005-04-1565

91. Cerny T, Borisch B, Introna M, Johnson P, Rose AL. Mechanism of action of rituximab. *Anticancer Drugs*. 2002;13(SUPPL. 2):S3-10. doi:10.1097/00001813-200211002-00002

92. Boross P, Leusen JHW. Mechanisms of action of CD20 antibodies. *Am J Cancer Res*. 2012;2(6):676-690.

93. Carbone A, Gloghini A. Follicular dendritic cell pattern in early lymphomas involving follicles. *Adv Anat Pathol*. 2014;21(4):260-269. doi:10.1097/PAP.0000000000000030

94. Aguzzi A, Krautler NJ. Characterizing follicular dendritic cells: a progress report. *Eur J Immunol*. 2010;40(8):2134-2138.

95. Smeltzer JP, Jones JM, Ziesmer SC, et al. Pattern of CD14+ follicular dendritic cells and PD1+ T cells independently predicts time to transformation in follicular lymphoma. *Clin Cancer Res*. 2014;20(11):2862-2872. doi:10.1158/1078-0432.CCR-13-2367

96. Shiozawa E, Yamochi-Onizuka T, Yamochi T, et al. Disappearance of CD21-positive follicular dendritic cells preceding the transformation of follicular lymphoma: immunohistological study of the transformation using CD21, p53, Ki-67, and P-glycoprotein. *Pathol Pract*. 2003;199(5):293-302.

97. Nelson LS, Mansfield JR, Lloyd R, et al. Automated prognostic pattern detection

shows favourable diffuse pattern of FOXP3+ Tregs in follicular lymphoma. *Br J Cancer*. 2015;113(October):1-9. doi:10.1038/bjc.2015.291

98. Spagnolo DM, Gyanchandani R, Al-Kofahi Y, et al. Pointwise mutual information quantifies intratumor heterogeneity in tissue sections labeled with multiple fluorescent biomarkers. *J Pathol Inform*. 2016;7:47. doi:10.4103/2153-3539.194839

99. Kovacheva VN, Khan AM, Khan M, Epstein DBA, Rajpoot NM. DiSWOP: A novel measure for cell-level protein network analysis in localized proteomics image data. *Bioinformatics*. 2014;30(3):420-427. doi:10.1093/bioinformatics/btt676

100. Rose CJ, Naidoo K, Clay V, Linton K, Radford JA, Byers RJ. A statistical framework for analyzing hypothesized interactions between cells imaged using multispectral microscopy and multiple immunohistochemical markers. *J Pathol Inform*. 2013;4(Suppl):S4. doi:10.4103/2153-3539.109856

101. Xu-Monette ZY, Zhou J, Young KH. PD-1 expression and clinical PD-1 blockade in B-cell lymphomas. *Blood, J Am Soc Hematol*. 2018;131(1):68-83.

102. Duraiyan J, Govindarajan R, Kaliyappan K, Palanisamy M. Applications of immunohistochemistry. *J Pharm Bioallied Sci*. 2012;4(Suppl 2):S307-9. doi:10.4103/0975-7406.100281

103. Werner M, Chott A, Fabiano A, Battifora H. Effect of formalin tissue fixation and processing on immunohistochemistry. *Am J Surg Pathol*. 2000;24(7):1016-1019.

104. Bogen SA, Vani K, Sompuram SR. Molecular mechanisms of antigen retrieval: antigen retrieval reverses steric interference caused by formalin-induced cross-links. *Biotech Histochem*. 2009;84(5):207-215. doi:10.3109/10520290903039078

105. Hayat MA. *Microscopy, Immunohistochemistry, and Antigen Retrieval Methods: For Light and Electron Microscopy*. Springer Science & Business Media; 2002.

106. Shi S-RR, Key ME, Kalra KL. Antigen retrieval in formalin-fixed, paraffin-embedded tissues: an enhancement method for immunohistochemical staining based on microwave oven heating of tissue sections. *J Histochem Cytochem*. 1991;39(6):741-748. doi:10.1177/39.6.1709656

107. Stack EC, Wang C, Roman KA, Hoyt CC. Multiplexed immunohistochemistry, imaging, and quantitation: A review, with an assessment of Tyramide signal amplification, multispectral imaging and multiplex analysis. *Methods*. 2014;70(1):46-58. doi:10.1016/j.ymeth.2014.08.016

108. Ramos-Vara JA. Technical aspects of immunohistochemistry. *Vet Pathol*. 2005;42(4):405-426. doi:10.1354/vp.42-4-405

109. Gustashaw KM, Najmabadi P, Potts SJ. Measuring protien expression in tissue: The complementary roles of brigtfield and flourescence in whole slide scanning. *Lab Med*. 2010;41(3):135-143. doi:10.1309/LM8QU4SBNV2VSAST

110. Tsakiroglou AM, West C, Astley S, Linton K, Fergie M, Byers R. Automated Multiplex Immunofluorescence with TSA for CD4, CD8, FOXP3, CD21, PD1 and CD68 in Follicular Lymphoma. Published online 2019. doi:10.17504/protocols.io.49ygz7w

111. Picot J, Guerin CL, Le Van Kim C, Boulanger CM. Flow cytometry: Retrospective, fundamentals and recent instrumentation. *Cytotechnology*. 2012;64(2):109-130. doi:10.1007/s10616-011-9415-0

112. Russi AE, Brown MA. Mass Cytometry: Single Cells, Many Features. *Cell*. 2016;165(2):255-269. doi:10.1016/j.trsl.2014.08.005.The

113. Grecco HE, Imtiaz S, Zamir E. Multiplexed imaging of intracellular protein networks. *Cytom Part A*. 2016;89(8):761-775. doi:10.1002/cyto.a.22876

114. Schubert W. Topological Proteomics, Toponomics, MELK-Technology. In: Hecker M, Müllner S, Cahill DJ, et al., eds. *Proteomics of Microorganisms: Fundamental Aspects and Application*. Springer Berlin Heidelberg; 2003:189-209. doi:10.1007/3-540-36459-5_8

115. Schubert W. A Three-Symbol Code for Organized Proteomes Based on Cyclical Imaging of Protein Locations. *Int Soc Anal Cytol*. 2006;69(A):659-676. doi:10.1002/cyto.a

116. Schubert W, Bonnekoh B, Pommer AJ, et al. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nat Biotechnol*. 2006;24(10):1270-1278. doi:10.1038/nbt1250

117. Friedenberger M, Bode M, Krusche A, Schubert W. Fluorescence detection of protein clusters in individual cells and tissue sections by using toponome imaging system: sample preparation and measuring procedures. *Nat Protoc*. 2007;2(9):2285-2294. doi:10.1038/nprot.2007.320

118. Bhattacharya S, Mathew G, Ruban E, et al. Toponome imaging system: In situ protein network mapping in normal and cancerous colon from the same patient reveals more than five-thousand cancer specific protein clusters and their subcellular annotation by using a three symbol code. *J Proteome Res*. 2010;9(12):6112-6125. doi:10.1021/pr100157p

119. Hillert R, Gieseler A, Krusche A, et al. Large molecular systems landscape uncovers T cell trapping in human skin cancer. *Sci Rep*. 2016;6(January):19012. doi:10.1038/srep19012

120. Schubert W. On the origin of cell functions encoded in the toponome. *J Biotechnol*. 2010;149(4):252-259. doi:10.1016/j.jbiotec.2010.03.009

121. Schubert W. Advances in toponomics drug discovery: Imaging cycler microscopy correctly predicts a therapy method of amyotrophic lateral sclerosis. *Cytometry A*. 2015;87(8):696-703. doi:10.1002/cyto.a.22671

122. Schubert W, Friedenberger M, Bode M, Krusche A, Hillert R. Functional architecture of the cell nucleus: Towards comprehensive toponome reference maps of apoptosis. *Biochim Biophys Acta - Mol Cell Res*. 2008;1783(11):2080-2088. doi:10.1016/j.bbamcr.2008.07.019

123. Schubert W, Gieseler A, Krusche A, Hillert R. Toponome mapping in prostate cancer: Detection of 2000 cell surface protein clusters in a single tissue section and cell type specific annotation by using a three symbol code. *J Proteome Res*. 2009;8(6):2696-2707. doi:10.1021/pr800944f

124. Schubert W, Gieseler A, Krusche A, Serocka P, Hillert R. Next-generation biomarkers based on 100-parameter functional super-resolution microscopy TIS. *N Biotechnol*. 2012;29(5):599-610. doi:10.1016/j.nbt.2011.12.004

125. Oeltze S, Freiler W, Hillert R, Doleisch H, Preim B, Schubert W. Interactive , Graph-Based Visual Analysis of High-Dimensional , Multi-Parameter Fluorescence Microscopy Data in Toponomics. *IEEE Trans Vis Comput Graph*. 2011;17(12):1882-1891.

126. Oeltze S, Klemm P, Hillert R, Preim B, Schubert W. Visualization and Exploration of 3D Toponome Data. *EurographicsWorkshop Vis Comput Biol Med*. Published online 2012:1-2. doi:10.2312/VCBM/VCBM12/115-122

127. Pieper F, Hillert R, Preim B, Schubert W. Interactive Labeling of Toponome Data. *Eurographics Work Vis Comput Biol Med*. Published online 2014.

128. Gerdes MJ, Sevinsky CJ, Sood A, Adak S, Bello MO. Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc Natl Acad Sci U S A*. 2013;110(29):11982-11987. doi:10.1073/pnas.1300136110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1300136110

129. Hiraoka Y, Shimi T, Haraguchi T. Multispectral imaging fluorescence microscopy for living cells. *Cell Struct Funct*. 2002;27(5):367-374. doi:10.1247/csf.27.367

130. Fountaine TJ, Wincovitch SM, Geho DH, Garfield SH, Pittaluga S. Multispectral imaging of clinically relevant cellular targets in tonsil and lymphoid tissue using semiconductor quantum dots. *Mod Pathol*. 2006;193800628(9):1181-1191. doi:10.1038/modpathol.3800628

131. Pauly JL, Allison EM, Hurley EL, Nwogu CE, Wallace PK, Paszkiewicz GM. Fluorescent human lung macrophages analyzed by spectral confocal laser scanning microscopy and multispectral cytometry. *Microsc Res Tech*. 2005;67(2):79-89. doi:10.1002/jemt.20191

132. Levenson RM, Lynch DT, Kobayashi H, Backer JM, Backer M V. Multiplexing with multispectral imaging: from mice to microscopy. *ILAR J*. 2008;49(1):78-88.

133. Mansfield JR, Hoyt C, Levenson RM. Visualization of microscopy-based spectral imaging data from multi-label tissue sections. *Curr Protoc Mol Biol*. 2008;(SUPPL. 84):1-15. doi:10.1002/0471142727.mb1419s84

134. Dickinson ME, Bearman G, Tille S, Lansford R, Fraser SE. Multi-spectral imaging and linear unmixing add a whole new dimension to laser scanning fluorescence microscopy. *Biotechniques*. 2001;31(6):1272-1278.

135. Ijsselsteijn ME, van der Breggen R, Sarasqueta AF, Koning F, de Miranda NFCC. A 40-marker panel for high dimensional characterization of cancer immune microenvironments by imaging mass cytometry. *Front Immunol*. 2019;10(OCT):1-8. doi:10.3389/fimmu.2019.02534

136. Bjornson ZB, Nolan GP, Fantl WJ. Single-cell mass cytometry for analysis of immune system functional states. *Curr Opin Immunol*. 2013;25(4):484-494.

137. Newell EW, Cheng Y. Mass cytometry: Blessed with the curse of dimensionality. *Nat Immunol*. 2016;17(8):890-895. doi:10.1038/ni.3485

138. Giesen C, Wang HAO, Schapiro D, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods*. 2014;11(4):417-422.

139. Tan WCC, Nerurkar SN, Cai HY, et al. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Commun*. 2020;40(4):135-153. doi:10.1002/cac2.12023

140. Merritt CR, Ong GT, Church SE, et al. Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nat Biotechnol*. 2020;38(5):586-599. doi:10.1038/s41587-020-0472-9

141. Reed AEM, Bennett J, Kutasovic JR, et al. Digital spatial profiling application in

breast cancer: a user's perspective.

142. O'Neil S, Huynh R, Hebert C, et al. Use of Ultivue InSituPlex (R) multiplex immunofluorescence to localize and quantify regulatory T lymphocytes in formalin-fixed paraffin-embedded human tissue sections. In: *Journal for Immunotherapy of Cancer*. Vol 7. BMC CAMPUS; 2019.

143. Braubach O, Nikulina N, Singh J, et al. CODEX ® Ultra-high Plex Immunofluorescence Imaging for Neuroscience. :34.

144. Chan RCK, Li JJX, Yeung W, Chan AWH. Virtual multiplex immunohistochemistry: Application on cell block of effusion and aspiration cytology. *Diagn Cytopathol*. 2020;48(5):417-423.

145. Trahearn N, Epstein D, Cree I, Snead D, Rajpoot N. Hyper-Stain Inspector: A Framework for Robust Registration and Localised Co-Expression Analysis of Multiple Whole-Slide Images of Serial Histology Sections. *Sci Rep*. 2017;7(1):1-13. doi:10.1038/s41598-017-05511-w

146. Visiopharm. Virtual 8-plex Multiplexing, Breast Cancer, TME. Published 2019. Accessed August 11, 2020. https://visiopharm.com/app-center/app/virtual-8-plex-multiplexing-breast-cancer-tme/

147. Toda Y, Kono K, Abiru H, et al. Application of tyramide signal amplification system to immunohistochemistry: A potent method to localize antigens that are not detectable by ordinary method. *Pathol Int*. 1999;49(5):479-483. doi:10.1046/j.1440-1827.1999.00875.x

148. Tóth ZE, Mezey É. Simultaneous visualization of multiple antigens with tyramide signal amplification using antibodies from the same species. *J Histochem Cytochem*. 2007;55(6):545-554. doi:10.1369/jhc.6A7134.2007

149. Optimization strategy for fluorescence multiplex immunohistochemistry tissue staining. (800):8401.

150. Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. *Histopathology*. 2012;61(1):1-9.

151. Brown CM. Fluorescence microscopy-avoiding the pitfalls. *J Cell Sci*. 2007;120(10):1703-1705.

152. Marty GDM. Blank-field correction for achieving a uniform white background in brightfield digital photomicrographs. *Biotechniques*. 2007;42(6):716-720. doi:10.2144/000112488

153. Smith K, Li Y, Piccinini F, et al. CIDRE : an illumination- correction method for optical microscopy. 2015;12(5). doi:10.1038/nmeth.3323

154. Peng T, Thorn K, Schroeder T, et al. A BaSiC tool for background and shading correction of optical microscopy images. *Nat Commun*. 2017;8(1):1-7.

155. Singh S, BRAY M, Jones TR, Carpenter AE. Pipeline for illumination correction of images for high-throughput microscopy. *J Microsc*. 2014;256(3):231-236.

156. Leong FJWM, Brady M, McGee JO. Correction of uneven illumination (vignetting) in digital microscopy images. *J Clin Pathol*. 2003;56(8):619-621.

157. Serra J. *Image Analysis and Mathematical Morphology*.; 1982.

158. Navarro L, Molimard J. Directional Denoising Using Fourier Spectrum Cloning. In:

Fourier Transforms-Century of Digitalization and Increasing Expectations. IntechOpen; 2019.

159. Cristobal G, Chagoyen M, Escalante-Ramirez B, Lopez JR. Wavelet-based denoising methods: A comparative study with applications in microscopy. In: *Wavelet Applications in Signal and Image Processing IV*. Vol 2825. International Society for Optics and Photonics; 1996:660-671.

160. Garg P, Jain T. A comparative study on histogram equalization and cumulative histogram equalization. *Int J New Technol Res*. 2017;3(9).

161. Reinhard E, Ashikhmin M, Gooch B, Shirley P. Color transfer between images. *IEEE Comput Graph Appl*. 2001;21(5):34-41. doi:10.1109/38.946629

162. Martinez-Morilla S, McGuire J, Gaule P, et al. Quantitative assessment of PD-L1 as an analyte in immunohistochemistry diagnostic assays using a standardized cell line tissue microarray. *Lab Investig*. Published online 2019:1-12.

163. Otsu N. A threshold selection method from gray-level histograms. *Automatica*. 1975;11(285-296):23-27.

164. Rosin PL. Unimodal thresholding. *Pattern Recognit*. 2001;34(11):2083-2096. doi:10.1016/S0031-3203(00)00136-9

165. Bradley D, Roth G. Adaptive Thresholding using the Integral Image. *J Graph Tools*. 2007;12(2):13-21. doi:10.1080/2151237X.2007.10129236

166. Costes S V., Daelemans D, Cho EH, Dobbin Z, Pavlakis G, Lockett S. Automatic and quantitative measurement of protein-protein colocalization in live cells. *Biophys J*. 2004;86(6):3993-4003. doi:10.1529/biophysj.103.038422

167. Barysenka A, Dress AWM, Schubert W. An information theoritic thresholding method for detecting protein colocalizations in stacks of fluorescent images. *J Biotechnol*. 2010;149:127-131.

168. Chang YH, Chin K, Thibault G, et al. RESTORE: Robust intEnSiTy nORmalization mEthod for multiplexed imaging. *Commun Biol*. 2020;3(1):1-9.

169. Meijering E. Cell segmentation: 50 Years down the road [life Sciences]. *IEEE Signal Process Mag*. 2012;29(5):140-145. doi:10.1109/MSP.2012.2204190

170. Kanade T, Yin Z, Bise R, et al. Cell image analysis: Algorithms, system and applications. *2011 IEEE Work Appl Comput Vision, WACV 2011*. Published online 2011:374-381. doi:10.1109/WACV.2011.5711528

171. Kumar N, Verma R, Anand D, et al. A Multi-Organ Nucleus Segmentation Challenge. 2020;39(5):1380-1391.

172. Guiet R, Burri O, Seitz A. Open source tools for biological image analysis. In: *Computer Optimized Microscopy*. Springer; 2019:23-37.

173. Shapiro L, Stockman G. *Computer Vision*. Prentice Hall; 2002. http://www.cse.msu.edu/~stockman/Book/2002/Chapters/ch3.pdf

174. Beucher S, Meyer F. The morphological approach to segmentation: the watershed transformation. *Math Morphol image Process*. 1993;34:433-481.

175. Schmidt U, Weigert M, Broaddus C, Myers G. Cell Detection with Star-convex Polygons. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2018:265-273.

176. Naylor P, Laé M, Reyal F, Walter T. Segmentation of Nuclei in Histopathology Images by Deep Regression of the Distance Map. *IEEE Trans Med Imaging*. 2019;38(2):448-459. doi:10.1109/TMI.2018.2865709

177. Malmberg F. Approximate Energy Minimization via Graph Cuts.

178. Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. *Int J Comput Vis*. 1988;1(4):321-331. doi:10.1007/BF00133570

179. Al-Hafiz F, Al-Megren S, Kurdi H. Red blood cell segmentation by thresholding and Canny detector. *Procedia Comput Sci*. 2018;141:327-334.

180. Dima AA, Elliott JT, Filliben JJ, et al. Comparison of segmentation algorithms for fluorescence microscopy images of cells. *Cytom Part A*. 2011;79 A(7):545-559. doi:10.1002/cyto.a.21079

181. Kothari S, Chaudry Q, Wang MD. Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques. *Proc - 2009 IEEE Int Symp Biomed Imaging From Nano to Macro, ISBI 2009*. 2009;2:795-798. doi:10.1109/ISBI.2009.5193169

182. Zhou X, Liu K-Y, Bradley P, Perrimon N, Wong STC. Towards automated cellular image segmentation for RNAi genome-wide screening. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2005:885-892.

183. Fatakdawala H, Xu J, Basavanhally A, et al. Expectation–maximization-driven geodesic active contour with overlap resolution (emagacor): Application to lymphocyte segmentation on breast cancer histopathology. *IEEE Trans Biomed Eng*. 2010;57(7):1676-1689.

184. Jung C, Kim C, Chae SW, Oh S. Unsupervised segmentation of overlapped nuclei using Bayesian classification. *IEEE Trans Biomed Eng*. 2010;57(12):2825-2832.

185. Kong H, Gurcan M, Belkacem-Boussaid K. Partitioning histopathological images: an integrated framework for supervised color-texture segmentation and cell splitting. *IEEE Trans Med Imaging*. 2011;30(9):1661-1677. doi:10.1109/TMI.2011.2141674

186. Cheng L, Ye N, Yu W, Cheah A. Discriminative segmentation of microscopic cellular images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2011:637-644.

187. Berg S, Kutra D, Kroeger T, et al. ilastik: interactive machine learning for (bio)image analysis. *Nat Methods*. Published online September 2019. doi:10.1038/s41592-019-0582-9

188. Irshad H, Veillard A, Roux L, Racoceanu D. Methods for nuclei detection, segmentation, and classification in digital histopathology: A review-current status and future potential. *IEEE Rev Biomed Eng*. 2014;7:97-114. doi:10.1109/RBME.2013.2295804

189. Xing F, Yang L. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review. *IEEE Rev Biomed Eng*. 2016;9:234-263. doi:10.1109/RBME.2016.2515127

190. Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep Learning*. Vol 1. MIT press Cambridge; 2016.

191. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical

image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. ; 2015:234-241.

192. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv Prepr arXiv14091556*. Published online 2014.

193. He K, Gkioxari G, Doll?r P, Girshick R. Mask R-CNN. Published online 2017.

194. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. ; 2015:3431-3440.

195. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. ; 2017:4700-4708.

196. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv Prepr arXiv160207261*. Published online 2016.

197. Song Y, Zhang L, Chen S, Ni D, Lei B, Wang T. Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. *IEEE Trans Biomed Eng*. 2015;62(10):2421-2433. doi:10.1109/TBME.2015.2430895

198. Neubeck A, Van Gool L. Efficient non-maximum suppression. In: *18th International Conference on Pattern Recognition (ICPR'06)*. Vol 3. IEEE; 2006:850-855.

199. Xing F, Member S, Xie Y, et al. An automatic learning-based framework for robust nucleus segmentation. *IEEE Trans Med Imaging*. 2016;35(2):550-566. doi:10.1109/TMI.2015.2481436

200. Jung C, Kim C. Segmenting clustered nuclei using H-minima transform-based marker extraction and contour parameterization. *IEEE Trans Biomed Eng*. 2010;57(10 PART 2):2600-2604. doi:10.1109/TBME.2010.2060336

201. StarDist for ImageJ. Accessed August 23, 2020. https://imagej.net/StarDist

202. Bankhead P. StarDist. Accessed August 23, 2020. https://qupath.readthedocs.io/en/latest/docs/advanced/stardist.html

203. Alemi Koohbanani N, Jahanifar M, Zamani Tajadin N, Rajpoot N. NuClick: A deep learning framework for interactive segmentation of microscopic images. *Med Image Anal*. 2020;65. doi:10.1016/j.media.2020.101771

204. Mahmood F, Borders D, Chen R, et al. Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images. *IEEE Trans Med Imaging*. Published online 2019:1-1. doi:10.1109/tmi.2019.2927182

205. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. ; 2014:2672-2680.

206. Haq MM, Huang J. Adversarial Domain Adaptation for Cell Segmentation. In: *Medical Imaging with Deep Learning*. ; 2020:1-11. http://ai.bu.edu/visda-2017/assets/attachments/VisDA_VLLab.pdf

207. Al-Kofahi Y, Fiona G. Image analytic algorithms for automated cell segmentation quality control. *Proc - Int Symp Biomed Imaging*. 2018;2018-April(Isbi):423-426. doi:10.1109/ISBI.2018.8363607

208. McKinley ET, Sui Y, Al-Kofahi Y, et al. Optimized multiplex immunofluorescence single-cell analysis reveals tuft cell heterogeneity. *JCI insight*. 2017;2(11).

209. Quelhas P, Marcuzzo M, Mendonça AM, Campilho A. Cell nuclei and cytoplasm joint segmentation using the sliding band filter. *IEEE Trans Med Imaging*. 2010;29(8):1463-1473. doi:10.1109/TMI.2010.2048253

210. Jones TR, Carpenter A, Golland P. Voronoi-based segmentation of cells on image manifolds. In: *International Workshop on Computer Vision for Biomedical Image Applications*. Springer; 2005:535-543.

211. Yu W, Lee HK, Hariharan S, Bu W, Ahmed S. Evolving generalized Voronoi diagrams for accurate cellular image segmentation. *Cytom Part A J Int Soc Adv Cytom*. 2010;77(4):379-386. doi:10.1002/cyto.a.20876

212. Han J, Chang H, Andarawewa K, Yaswen P, Barcellos-Hoff MH, Parvin B. Multidimensional profiling of cell surface proteins and nuclear markers. *IEEE/ACM Trans Comput Biol Bioinform*. 2010;7(1):80-90. doi:10.1109/TCBB.2008.134

213. Lin J-R, Izar B, Mei S, Wang S, Shah P, Sorger P. A simple open-source method for highly multiplexed imaging of single cells in tissues and tumours. *Elife*. 2017;7(e31657):1-46. doi:10.7554/eLife.31657

214. Schuffler P, Shapiro D, Giesen C, et al. Automatic Single Cell Segmentation on Highly Multiplexed Tissue Images. *Cytom Part A*. 2015;87(10):936-942. doi:10.1002/cyto.a.22702

215. Mckinley ET, Roland JT, Franklin JL, et al. Machine and deep learning single-cell segmentation and quantification of multi-dimensional tissue images. Published online 2019.

216. Aghaeepour N, Finak G, Hoos H, et al. Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods*. 2013;10(3):228-238. doi:10.1038/NMETH.2365

217. Aeffner F, Wilson K, Martin NT, et al. The Gold Standard Paradox in Digital Image Analysis: Manual Versus Automated Scoring as Ground Truth. *Arch Pathol Lab Med*. 2017;141(9):1267-1275. doi:10.5858/arpa.2016-0386-RA

218. Blom S, Paavolainen L, Bychkov D, et al. Systems pathology by multiplexed immunohistochemistry and whole-slide digital image analysis. *Sci Rep*. 2017;7(1):1-13. doi:10.1038/s41598-017-15798-4

219. Levine JH, Simonds EF, Bendall SC, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*. 2015;162(1):184-197. doi:10.1016/j.cell.2015.05.047

220. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech theory Exp*. 2008;2008(10):P10008.

221. Viratham Pulsawatdi A, Craig SG, Bingham V, et al. A robust multiplex immunofluorescence and digital pathology workflow for the characterisation of the tumour immune microenvironment. *Mol Oncol*. Published online 2020:0-1. doi:10.1002/1878-0261.12764

222. Quinn J, Fisher PW, Capocasale RJ, et al. A statistical pattern recognition approach for determining cellular viability and lineage phenotype in cultured cells and murine bone marrow. *Cytom Part A J Int Soc Anal Cytol*. 2007;71(8):612-624.

223. Yuan Y. Spatial Heterogeneity in the Tumor Microenvironment. *Cold Spring Harb*

Perspect Med. 2016;6(8). doi:10.1101/cshperspect.a026583

224. Francis K, Palsson BO. Effective intercellular communication distances are determined by the relative time constants for cyto/chemokine secretion and diffusion. *Proc Natl Acad Sci*. 1997;94(23):12258-12262. doi:10.1073/pnas.94.23.12258

225. Takko H, Pajanoja C, Kurtzeborn K, Hsin J, Kuure S, Kerosuo L. ShapeMetrics : A userfriendly pipeline for 3D cell segmentation and spatial tissue analysis. *Dev Biol*. 2020;462(1):7-19. doi:10.1016/j.ydbio.2020.02.003

226. Feichtenbeiner A, Haas M, Büttner M, Grabenbauer GG, Fietkau R, Distel L V. Critical role of spatial interaction between CD8+ and FOXP3^{+} cells in human gastric cancer: The distance matters. *Cancer Immunol Immunother*. 2014;63(2):111-119. doi:10.1007/s00262-013-1491-x

227. Barua S, Fang P, Sharma A, et al. Spatial interaction of tumor cells and regulatory T cells correlates with survival in non-small cell lung cancer. *Lung Cancer*. 2018;117(January):73-79. doi:10.1016/j.lungcan.2018.01.022

228. Mezheyeuski A, Bergsland CH, Backman M, et al. Multispectral imaging for quantitative and compartment-specific immune infiltrates reveals distinct immune profiles that classify lung cancer patients. *J Pathol*. 2018;244(4):421-431. doi:10.1002/path.5026

229. Lazarus J, Maj T, Smith JJ, et al. Spatial and phenotypic immune profiling of metastatic colon cancer. *JCI insight*. 2018;3(22):1-15. doi:10.1172/jci.insight.121932

230. Cheikh B Ben, Elie N, Plancoulaine B, Bor-Angelier C, Racoceanu D. Spatial interaction analysis with graph based mathematical morphology for histopathology. *ISBI, Melb*. Published online 2017:813-817. doi:10.1109/ISBI.2017.7950642

231. Maley CC, Koelble K, Natrajan R, Aktipis A, Yuan Y. An ecological measure of immune-cancer colocalization as a prognostic factor for breast cancer. *Breast Cancer Res*. 2015;17(1):131.

232. Yuan Y. Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. *J R Soc Interface*. 2015;12(103):20141153-. doi:10.1098/rsif.2014.1153

233. Mani NL, Schalper KA, Hatzis C, et al. Quantitative assessment of the spatial heterogeneity of tumor-infiltrating lymphocytes in breast cancer. *Breast Cancer Res*. 2016;18(1):78. doi:10.1186/s13058-016-0737-x

234. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath : Open source software for digital pathology image analysis. *Sci Rep*. 2017;7(1):16878. doi:https://doi.org/10.1101/099796

235. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods*. 2012;9(7):671-675.

236. Schapiro D, Jackson HW, Raghuraman S, et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat Methods*. 2017;14(9):873.

237. McQuin C, Goodman A, Chernyshev V, et al. CellProfiler 3.0: Next-generation image processing for biology. *PLoS Biol*. 2018;16(7):e2005970.

238. Gibson-Corley KN, Olivier AK, Meyerholz DK. Principles for valid

histopathologic scoring in research. *Vet Pathol*. 2013;50(6):1007-1015.

239. Payne SJL, Bowen RL, Jones JL, Wells CA. Predictive markers in breast cancer - The present. *Histopathology*. 2008;52(1):82-90. doi:10.1111/j.1365-2559.2007.02897.x

240. Allison KH, Hammond MEH, Dowsett M, et al. Estrogen and progesterone receptor testing in breast cancer: American society of clinical oncology/college of American pathologists guideline update. *Arch Pathol Lab Med*. 2020;144(5):545-563. doi:10.5858/arpa.2019-0904-SA

241. Nitta H, Kelly BD, Allred C, et al. The assessment of HER2 status in breast cancer: the past, the present, and the future. *Pathol Int*. 2016;66(6):313-324. doi:10.1111/pin.12407

242. Dahlin AM, Henriksson ML, Van Guelpen B, et al. Colorectal cancer prognosis depends on T-cell infiltration and molecular characteristics of the tumor. *Mod Pathol*. 2011;24(5):671-682. doi:10.1038/modpathol.2010.234

243. Nosho K, Baba Y, Tanaka N, et al. Tumour-infiltrating T-cell subsets, molecular changes in colorectal cancer, and prognosis: Cohort study and literature review. *J Pathol*. 2010;222(4):350-366. doi:10.1002/path.2774

244. Allred DC, Bustamante MA, Daniel CO, Gaskill H V, Cruz AB. Immunocytochemical analysis of estrogen receptors in human breast carcinomas: evaluation of 130 cases and review of the literature regarding concordance with biochemical assay and clinical relevance. *Arch Surg*. 1990;125(1):107-113.

245. McCarty KS, Szabo E, Flowers JL, et al. Use of a monoclonal anti-estrogen receptor antibody in the immunohistochemical evaluation of human tumors. *Cancer Res*. 1986;46(8 SUPPL.):4244-4249.

246. Wolff AC, McShane LM, Hammond MEH, et al. Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *Arch Pathol Lab Med*. 2018;142(11):1364-1382. doi:10.5858/arpa.2018-0902-SA

247. Tsakiroglou AM, Linton K, Peset-Martin I, et al. Automated scoring methods for quantitative interpretation of HER2, ER and T cell human pathology markers in the tumour microenvironment: a systematic review. PROSPERO International Prospective Register of Systematic Reviews. Published 2019. Accessed June 24, 2020. https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42019139688

248. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLOS Med*. 2009;6(7):1-28. doi:10.1371/journal.pmed.1000100

249. Sun S. Meta-analysis of Cohen's kappa. *Heal Serv Outcomes Res Methodol*. 2011;11(3-4):145-163.

250. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.

251. Hawker S, Payne S, Kerr C, Hardey M, Powell J. Appraising the evidence: reviewing disparate data systematically. *Qual Health Res*. 2002;12(9):1284-1299.

252. Giltnane JM, Murren JR, Rimm DL, King BL. AQUA and FISH analysis of HER-2/neu expression and amplification in a small cell lung carcinoma tissue microarray.

Histopathology. 2006;49(2):161-169. doi:10.1111/j.1365-2559.2006.02479.x

253. Giltnane JM, Molinaro A, Cheng H, et al. Comparison of Quantitative Immunofluorescence With Conventional Methods for HER2/neu Testing With Respect to Response to Trastuzumab Therapy in Metastatic Breast Cancer. *Arch Pathol Lab Med*. 2008;132(10):1635-1647.

254. Wasserman BE, Carvajal-Hausdorf DE, Ho K, et al. High concordance of a closed-system, RT-qPCR breast cancer assay for HER2 mRNA, compared to clinically determined immunohistochemistry, fluorescence in situ hybridization, and quantitative immunofluorescence. *Lab Invest*. 2017;97(12):1521-1526. doi:10.1038/labinvest.2017.93

255. Camp RL, Dolled-Filhart M, King BL, Rimm DL. Quantitative analysis of breast cancer tissue microarrays shows that both high and normal levels of HER2 expression are associated with poor outcome. *Cancer Res*. 2003;63(7):1445-1448.

256. Aitken SJ, Thomas JS, Langdon SP, Harrison DJ, Faratian D. Quantitative analysis of changes in ER, PR and HER2 expression in primary breast cancer and paired nodal metastases. *Ann Oncol Off J Eur Soc Med Oncol*. 2010;21(6):1254-1261. doi:10.1093/annonc/mdp427

257. Gupta S, Neumeister V, McGuire J, et al. Quantitative assessments and clinical outcomes in HER2 equivocal 2018 ASCO/CAP ISH group 4 breast cancer. *NPJ Breast Cancer*. 2019;5. doi:10.1038/s41523-019-0122-x

258. Cheng H, Bai Y, Sikov W, et al. Quantitative measurements of HER2 and phospho-HER2 expression: correlation with pathologic response to neoadjuvant chemotherapy and trastuzumab. *BMC Cancer*. 2014;14. doi:10.1186/1471-2407-14-326

259. Gustavson MD, Bourke-Martin B, Reilly D, et al. Standardization of HER2 Immunohistochemistry in Breast Cancer by Automated Quantitative Analysis. *Arch Pathol Lab Med*. 2009;133(9):1413-1419.

260. Sherman ME, Rimm DL, Yang XR, et al. Variation in breast cancer hormone receptor and HER2 levels by etiologic factors: a population-based analysis. *Int J cancer*. 2007;121(5):1079-1085. doi:10.1002/ijc.22812

261. Hashiguchi A, Hashimoto Y, Suzuki H, Sakamoto M. Using immunofluorescent digital slide technology to quantify protein expression in archival paraffin-embedded tissue sections. *Pathol Int*. 2010;60(11):720-725. doi:10.1111/j.1440-1827.2010.02590.x

262. Can A, Bello M, Cline HE, Tao X, Mendonca P, Gerdes M. A unified segmentation method for detecting subcellular compartments in immunofluroescently labeled tissue images. In: *International Workshop on Microscopic Image Analysis with Applications in Biology. Sept.* Vol 3. ; 2009.

263. Zarrella ER, Coulter M, Welsh AW, et al. Automated measurement of estrogen receptor in breast cancer: a comparison of fluorescent and chromogenic methods of measurement. *Lab Invest*. 2016;96(9):1016-1025. doi:10.1038/labinvest.2016.73

264. Camp RL, Chung GG, Rimm DL. Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nat Med*. 2002;473(11):1145-1152. doi:10.1038/nm

265. Escobar J, Klimowicz AC, Dean M, et al. Quantification of ER/PR expression in ovarian low-grade serous carcinoma. *Gynecol Oncol*. 2013;128(2):371-376. doi:10.1016/j.ygyno.2012.10.013

266. Can A, Bello MO, Gerdes MJ. Quantification of subcellular molecules in tissue microarray. In: *2010 20th International Conference on Pattern Recognition*. IEEE; 2010:2548-2551.

267. Chung GG, Zerkowski MP, Ghosh S, Camp RL, Rimm DL. Quantitative analysis of estrogen receptor heterogeneity in breast cancer. *Lab Investig*. 2007;87(7):662-669. doi:10.1038/labinvest.3700543

268. Brown JR, Wimberly H, Lannin DR, Nixon C, Rimm DL, Bossuyt V. Multiplexed quantitative analysis of CD3, CD8, and CD20 predicts response to neoadjuvant chemotherapy in breast cancer. *Clin Cancer Res*. 2014;20(23):5995-6005. doi:10.1158/1078-0432.CCR-14-1622

269. Charpin C, Secq V, Giusiano S, et al. A signature predictive of disease outcome in breast carcinomas, identified by quantitative immunocytochemical assays. *Int J cancer*. 2009;124(9):2124-2134. doi:10.1002/ijc.24177

270. Bolton KL, Garcia-Closas M, Pfeiffer RM, et al. Assessment of automated image analysis of breast cancer tissue microarrays for epidemiologic studies. *Cancer Epidemiol Biomarkers Prev*. 2010;19(4):992-999. doi:10.1158/1055-9965.EPI-09-1023

271. Cheong LS, Jean A, Tan TS, Kong W, Tan SY. Automated Segmentation and Measurement for Cancer Classification of HER2/neu Status in Breast Carcinomas. In: Qian, PY and Nghiem S, ed. *Biotechno 2011: The Third International Conference on Bioinformatics, Biocomputational Systems and Biothechnologies*. Iaria XPS Press; 2011:43-48.

272. Khameneh FD, Razavi S, Kamasak M. Automated segmentation of cell membranes to evaluate HER2 status in whole slide images using a modified deep learning network. *Comput Biol Med*. 2019;110:164-174. doi:10.1016/j.compbiomed.2019.05.020

273. Tawfik OW, Kimler BF, Davis M, et al. Comparison of immunohistochemistry by automated cellular imaging system (ACIS) versus fluorescence in-situ hybridization in the evaluation of HER-2/neu expression in primary breast carcinoma. *Histopathology*. 2006;48(3):258-267. doi:10.1111/j.1365-2559.2005.02322.x

274. Mohammed ZMA, Going JJ, McMillan DC, et al. Comparison of visual and automated assessment of HER2 status and their impact on outcome in primary operable invasive ductal breast cancer. *Histopathology*. 2012;61(4):675-684. doi:10.1111/j.1365-2559.2012.04280.x

275. Di Cataldo S, Ficarra E, Macii E. Computer-aided techniques for chromogenic immunohistochemistry: Status and directions. *Comput Biol Med*. 2012;42(10):1012-1025. doi:10.1016/j.compbiomed.2012.08.004

276. Wang J, Ruan J, He S, et al. Detection of Her2 Scores and Magnification from Whole Slide Images in Multi-Task Convolutional Network. In: *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*. Vol 01. ; 2018:7-10. doi:10.1109/ISCID.2018.00008

277. Cantaloni C, Tonini E, Eccher C, et al. Diagnostic Value of Automated Her2 Evaluation in Breast Cancer A Study on 272 Equivocal (score 2+) Her2 Immunoreactive Cases Using an FDA Approved System. *Appl Immunohistochem Mol Morphol*. 2011;19(4):306-312. doi:10.1097/PAI.0b013e318205b03a

278. Brugmann A, Eld M, Lelkaitis G, et al. Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains. *Breast Cancer Res Treat*.

2012;132(1):41-49. doi:10.1007/s10549-011-1514-2

279. Atkinson R, Mollerup J, Laenkholm A-V, et al. Effects of the change in cutoff values for human epidermal growth factor receptor 2 status by immunohistochemistry and fluorescence in situ hybridization: a study comparing conventional brightfield microscopy, image analysis-assisted microscopy, and inter. *Arch Pathol Lab Med*. 2011;135(8):1010-1016. doi:10.5858/2010-0462-OAR

280. Ciampa A, Xu B, Ayata G, et al. HER-2 status in breast cancer - Correlation of gene amplification by FISH with immunohistochemistry expression using advanced cellular imaging system. *Appl Immunohistochem Mol Morphol*. 2006;14(2):132-137. doi:10.1097/01.pai.0000150516.75567.13

281. Ellis CM, Dyson MJ, Stephenson TJ, Maltby EL. HER2 amplification status in breast cancer: a comparison between immunohistochemical staining and fluorescence in situ hybridisation using manual and automated quantitative image analysis scoring techniques. *J Clin Pathol*. 2005;58(7):710-714. doi:10.1136/jcp.2004.023424

282. Saha M, Chakraborty C. Her2Net: A Deep Framework for Semantic Segmentation and Classification of Cell Membranes and Nuclei in Breast Cancer Evaluation. *IEEE Trans Image Process*. 2018;27(5):2189-2200. doi:10.1109/TIP.2018.2795742

283. Ormenisan C, Wang J, Lawson D, Cohen C. Image cytometric HER2 in gastric carcinoma: is a new algorithm needed? *Appl Immunohistochem Mol Morphol AIMM*. 2013;21(5):414-419. doi:10.1097/PAI.0b013e31827955c8

284. Tuominen VJ, Tolonen TT, Isola J. ImmunoMembrane: a publicly available web application for digital image analysis of HER2 immunohistochemistry. *Histopathology*. 2012;60(5):758-767. doi:10.1111/j.1365-2559.2011.04142.x

285. Turashvili G, Leung S, Turbin D, et al. Inter-observer reproducibility of HER2 immunohistochemical assessment and concordance with fluorescent in situ hybridization (FISH): pathologist assessment compared to quantitative image analysis. *BMC Cancer*. 2009;9:165. doi:10.1186/1471-2407-9-165

286. Laurinaviciene A, Dasevicius D, Ostapenko V, Jarmalaite S, Lazutka J, Laurinavicius A. Membrane connectivity estimated by digital image analysis of HER2 immunohistochemistry is concordant with visual scoring and fluorescence in situ hybridization results: algorithm evaluation on breast cancer tissue microarrays. *Diagn Pathol*. 2011;6:87. doi:10.1186/1746-1596-6-87

287. Gavrielides MA, Gallas BD, Lenz P, Badano A, Hewitt SM. Observer Variability in the Interpretation of HER2/neu Immunohistochemical Expression With Unaided and Computer-Aided Digital Microscopy. *Arch Pathol Lab Med*. 2011;135(2):233-242.

288. Holten-Rossing H, Talman M-LM, Kristensson M, Vainer B. Optimizing HER2 assessment in breast cancer: application of automated image analysis. *Breast Cancer Res Treat*. 2015;152(2):367-375. doi:10.1007/s10549-015-3475-3

289. Jeung J, Patel R, Vila L, Wakefield D, Liu C. Quantitation of HER2/neu expression in primary gastroesophageal adenocarcinomas using conventional light microscopy and quantitative image analysis. *Arch Pathol Lab Med*. 2012;136(6):610-617. doi:10.5858/arpa.2011-0371-OA

290. Lehr HA, Jacobs TW, Yaziji H, Schnitt SJ, Gown AM. Quantitative evaluation of HER-2/neu status in breast cancer by fluorescence in situ hybridization and by

immunohistochemistry with image analysis. *Am J Clin Pathol*. 2001;115(6):814-822. doi:10.1309/AJ84-50AK-1X1B-1Q4C

291. Hatanaka Y, Hashizume K, Kamihara Y, et al. Quantitative immunohistochemical evaluation of HER2/neu expression with HercepTestTM in breast carcinoma by image analysis. *Pathol Int*. 2001;51(1):33-36.

292. Slodkowska J, Filas V, Buszkiewicz E, et al. Study on breast carcinoma Her2/neu and hormonal receptors status assessed by automated images analysis systems: ACIS III (Dako) and ScanScope (Aperio). *FOLIA Histochem Cytobiol*. 2010;48(1):19-25. doi:10.2478/v10042-010-0015-1

293. Arihiro K, Oda M, Ogawa K, et al. Utility of cytopathological specimens and an automated image analysis for the evaluation of HER2 status and intratumor heterogeneity in breast carcinoma. *Pathol Res Pract*. 2016;212(12):1126-1132. doi:10.1016/j.prp.2016.09.014

294. Singh P, Mukundan R. A Robust HER2 Neural Network Classification Algorithm Using Biomarker-Specific Feature Descriptors. In: *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*. ; 2018:1-5. doi:10.1109/MMSP.2018.8547043

295. Mukundan R. Analysis of Image Feature Characteristics for Automated Scoring of HER2 in Histology Slides. *J Imaging*. 2019;5(3). doi:10.3390/jimaging5030035

296. Minot DM, Kipp BR, Root RM, et al. Automated cellular imaging system III for assessing HER2 status in breast cancer specimens: development of a standardized scoring method that correlates with FISH. *Am J Clin Pathol*. 2009;132(1):133-138. doi:10.1309/AJCPJV0SKAF2PCMY

297. Hall BH, Ianosi-Irimie M, Javidian P, Chen W, Ganesan S, Foran DJ. Computer-assisted assessment of the Human Epidermal Growth Factor Receptor 2 immunohistochemical assay in imaged histologic sections using a membrane isolation algorithm and quantitative analysis of positive controls. *BMC Med Imaging*. 2008;8:1-13. doi:10.1186/1471-2342-8-11

298. Stalhammar G, Fuentes Martinez N, Lippert M, et al. Digital image analysis outperforms manual biomarker assessment in breast cancer. *Mod Pathol*. 2016;29(4):318-329. doi:10.1038/modpathol.2016.34

299. Qaiser T, Mukherjee A, Reddy Pb C, et al. HER2 challenge contest: a detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology*. 2018;72(2):227-238. doi:10.1111/his.13333

300. Mukundan R. Image Features Based on Characteristic Curves and Local Binary Patterns for Automated HER2 Scoring. *J Imaging*. 2018;4(2). doi:10.3390/jimaging4020035

301. Bankhead P, Fernandez JA, McArt DG, 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

302. Qaiser T, Rajpoot NM. Learning Where to See: A Novel Attention Model for Automated Immunohistochemical Scoring. *IEEE Trans Med Imaging*. 2019;38(11):2620-2631. doi:10.1109/TMI.2019.2907049

303. Barnes M, Srinivas C, Bai I, et al. Whole tumor section quantitative image analysis maximizes between-pathologists' reproducibility for clinical immunohistochemistry-based biomarkers. *Lab Invest*. 2017;97(12):1508-1515.

doi:10.1038/labinvest.2017.82

304. Ali HR, Irwin M, Morris L, et al. Astronomical algorithms for automated analysis of tissue protein expression in breast cancer. *Br J Cancer*. 2013;108(3):602-612. doi:10.1038/bjc.2012.558

305. Gavrielides MA, Masmoudi H, Petrick N, Myers KJ, Hewitt SM. Automated evaluation of HER-2/neu immunohistochemical expression in breast cancer using digital microscopy. In: *2008 IEEE INTERNATIONAL SYMPOSIUM ON BIOMEDICAL IMAGING: FROM NANO TO MACRO, VOLS 1-4*. IEEE International Symposium on Biomedical Imaging. IEEE; 2008:808+. doi:10.1109/ISBI.2008.4541119

306. Masmoudi H, Hewitt SM, Petrick N, Myers KJ, Gavrielides MA. Automated quantitative assessment of HER-2/neu immunohistochemical expression in breast cancer. *IEEE Trans Med Imaging*. 2009;28(6):916-925. doi:10.1109/TMI.2009.2012901

307. Micsik T, Kiszler G, Szabo D, et al. Computer Aided Semi-Automated Evaluation of HER2 Immunodetection--A Robust Solution for Supporting the Accuracy of Anti HER2 Therapy. *Pathol Oncol Res*. 2015;21(4):1005-1011. doi:10.1007/s12253-015-9927-6

308. Lee CM, Lee RJ, Hammond E, et al. Expression of HER2neu (c-erbB-2) and epidermal growth factor receptor in cervical cancer: prognostic correlation with clinical characteristics, and comparison of manual and automated imaging analysis. *Gynecol Oncol*. 2004;93(1):209-214. doi:10.1016/j.ygyno.2004.01.006

309. Allott EH, Cohen SM, Geradts J, et al. Performance of Three-Biomarker Immunohistochemistry for Intrinsic Breast Cancer Subtyping in the AMBER Consortium. *Cancer Epidemiol Biomarkers Prev*. 2016;25(3):470-478. doi:10.1158/1055-9965.EPI-15-0874

310. Keller B, Chen W, Gavrielides MA. Quantitative Assessment and Classification of Tissue-Based Biomarker Expression With Color Content Analysis. *Arch Pathol Lab Med*. 2012;136(5):539-550. doi:10.5858/arpa.2011-0195-OA

311. Vandenberghe ME, Scott MLJ, Scorer PW, et al. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. *Sci Rep*. 2017;7(1):45938. doi:10.1038/srep45938

312. Howat WJ, Blows FM, Provenzano E, et al. Performance of automated scoring of ER, PR, HER2, CK5/6 and EGFR in breast cancer tissue microarrays in the Breast Cancer Association Consortium. *J Pathol Clin Res*. 2015;1(1):18-32. doi:10.1002/cjp2.3

313. Vijayashree R, Aruthra P, Rao KR. A Comparison of Manual and Automated Methods of Quantitation of Oestrogen/Progesterone Receptor Expression in Breast Carcinoma. *J Clin Diagnostic Res*. 2015;9(3):EC01-EC05. doi:10.7860/JCDR/2015/12432.5628

314. Sarikoc F, Kalinli A, Akgun H, Ozturk F. An automated prognosis system for estrogen hormone status assessment in breast cancer tissue samples. *Turkish J Electr Eng Comput Sci*. 2013;21(4):1199-1221. doi:10.3906/elk-1111-10

315. Kostopoulos S, Cavouras D, Daskalakis A, et al. Assessing estrogen receptors' status by texture analysis of breast tissue specimens and pattern recognition methods. In: Kropatsch, WG and Kampel, M and Hanbury A, ed. *Computer Analysis of Images and Patterns, Proceedings*. Vol 4673. Lecture Notes in

Computer Science. Springer-Verlag Berlin; 2007:221-228.

316. Gokhale S, Rosen D, Sneige N, et al. Assessment of two automated imaging systems in evaluating estrogen receptor status in breast carcinoma. *Appl Immunohistochem Mol Morphol AIMM*. 2007;15(4):451-455. doi:10.1097/PAI.0b013e31802ee998

317. Chakraborty C, Chatterjee S, Arun I, Ahmed R, Tewary S. AutoIHC-scoring : a machine learning framework for automated Allred scoring of molecular expression in ER- and PR-stained breast cancer tissue. *J Microsc*. 2017;268(2):172-185. doi:10.1111/jmi.12596

318. Faratian D, Kay C, Robson T, et al. Automated image analysis for high-throughput quantitative detection of ER and PR expression levels in large-scale clinical studies: the TEAM Trial Experience. *Histopathology*. 2009;55(5):587-593. doi:10.1111/j.1365-2559.2009.03419.x

319. Turbin DA, Leung S, Cheang MCU, et al. Automated quantitative analysis of estrogen receptor expression in breast carcinoma does not differ from expert pathologist scoring: a tissue microarray study of 3,484 cases. *Breast Cancer Res Treat*. 2008;110(3):417-426. doi:10.1007/s10549-007-9736-z

320. Rocha RM, Miller K, Soares F, Schenka N, Vassallo J, Gobbi H. Biotin-free systems provide stronger immunohistochemical signal in oestrogen receptor evaluation of breast cancer. *J Clin Pathol*. 2009;62(8):699-704. doi:10.1136/jcp.2009.065326

321. Mohammed ZMA, Edwards J, Orange C, et al. Breast cancer outcomes by steroid hormone receptor status assessed visually and by computer image analysis. *Histopathology*. 2012;61(2):283-292. doi:10.1111/j.1365-2559.2012.04244.x

322. Kostopoulos S, Cavouras D, Daskalakis A, et al. Cascade pattern recognition structure for improving quantitative assessment of estrogen receptor status in breast tissue carcinomas. *Anal Quant Cytol Histol*. 2008;30(4):218-225.

323. Jamaluddin MF, Fauzi MFA, Abas FS, et al. Cell Classification in ER-Stained Whole Slide Breast Cancer Images Using Convolutional Neural Network. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. ; 2018:632-635. doi:10.1109/EMBC.2018.8512386

324. Kostopoulos S, Cavouras D, Daskalakis A, et al. Colour-texture based image analysis method for assessing the hormone receptors status in breast tissue sections. *Conf Proc . Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf*. 2007;2007:4985-4988. doi:10.1109/IEMBS.2007.4353459

325. Akbar S, Jordan LB, Purdie CA, Thompson AM, McKenna SJ. Comparing computer-generated and pathologist-generated tumour segmentations for immunohistochemical scoring of breast tissue microarrays. *Br J Cancer*. 2015;113(7):1075-1080. doi:10.1038/bjc.2015.309

326. Arihiro K, Oda M, Ogawa K, et al. Comparison of evaluations of hormone receptors in breast carcinoma by image-analysis using three automated immunohistochemical stainings. *Exp Ther Med*. 2010;1(6):927-932. doi:10.3892/etm.2010.142

327. Sinn H-P, Schneeweiss A, Keller M, et al. Comparison of immunohistochemistry with PCR for assessment of ER, PR, and Ki-67 and prediction of pathological complete response in breast cancer. *BMC Cancer*. 2017;17(1):124. doi:10.1186/s12885-017-3111-1

328. Ahern TP, Beck AH, Rosner BA, et al. Continuous measurement of breast tumour hormone receptor expression: a comparison of two computational pathology

platforms. *J Clin Pathol*. 2017;70(5):428-434. doi:10.1136/jclinpath-2016-204107

329. Hatanaka Y, Hashizume K, Nitta K, Kato T, Itoh I, Tani Y. Cytometrical image analysis for immunohistochemical hormone receptor status in breast carcinomas. *Pathol Int*. 2003;53(10):693-699. doi:10.1046/j.1440-1827.2003.01547.x

330. Rajan SS, Horgan K, Speirs V, Hanby AM. External validation of the ImmunoRatio image analysis application for ER alpha determination in breast cancer. *J Clin Pathol*. 2014;67(1):72-75. doi:10.1136/jclinpath-2013-201680

331. Yan D, Randolph T, Zou J, Gong P. Incorporating deep features in the analysis of tissue microarray images. *Stat Interface*. 2019;12(2):283-293.

332. Mungle T, Tewary S, Das DK, et al. MRF-ANN: a machine learning approach for automated ER scoring of breast cancer immunohistochemical images. *J Microsc*. 2017;267(2):117-129. doi:10.1111/jmi.12552

333. Trahearn N, Tsang YW, Cree IA, Snead D, Epstein D, Rajpoot N. Simultaneous automatic scoring and co-registration of hormone receptors in tumor areas in whole slide images of breast cancer tissue slides. *Cytom Part A*. 2017;91(6):585-594. doi:10.1002/cyto.a.23035

334. Yan D, Wang P, Linden M, Knudsen B, Randolph T. Statistical methods for tissue array images-algorithmic scoring and co-training. *Ann Appl Stat*. 2012;6(3):1280-1305. doi:10.1214/12-AOAS543

335. Ilic IR, Stojanovic NM, Radulovic NS, et al. The Quantitative ER Immunohistochemical Analysis in Breast Cancer: Detecting the 3+0, 4+0, and 5+0 Allred Score Cases. *Medicina-Lithuania*. 2019;55(8). doi:10.3390/medicina55080461

336. Andersen NL, Brugmann A, Lelkaitis G, et al. Virtual Double Staining: A Digital Approach to Immunohistochemical Quantification of Estrogen Receptor Protein in Breast Carcinoma Specimens. *Appl Immunohistochem Mol Morphol*. 2018;26(9):620-626. doi:10.1097/PAI.0000000000000502

337. Nassar A, Cohen C, Agersborg SS, et al. A new immunohistochemical ER/PR image analysis system: A multisite performance study. *Appl Immunohistochem Mol Morphol*. 2011;19(3):195-202. doi:10.1097/PAI.0b013e3181fe53cb

338. Cheung CC, Neufeld H, Lining LA, et al. The Laboratory Score/Reference Method Score Ratio (LSRSR) Is a Novel Tool for Monitoring Laboratory Performance in Immunohistochemistry Proficiency Testing of Hormone Receptors in Breast Cancer. *Am J Clin Pathol*. 2011;136(1):67-73. doi:10.1309/AJCPQ619GHJMCBEV

339. Irshad H, Oh E, Schmolze D, et al. Crowdsourcing scoring of immunohistochemistry images : Evaluating Performance of the Crowd and an Automated Computational Method. *Nat Publ Gr*. 2017;7(June 2016):1-10. doi:10.1038/srep43286

340. Buisseret L, Desmedt C, Garaud S, et al. Reliability of tumor-infiltrating lymphocyte and tertiary lymphoid structure assessment in human breast cancer. *Mod Pathol an Off J United States Can Acad Pathol Inc*. 2017;30(9):1204-1212. doi:10.1038/modpathol.2017.43

341. Singh U, Cui Y, Dimaano N, et al. Analytical validation of quantitative immunohistochemical assays of tumor infiltrating lymphocyte biomarkers. *Biotech Histochem*. 2018;93(6):411-423. doi:10.1080/10520295.2018.1445290

342. Bouzin C, Saini ML, Khaing K-K, et al. Digital pathology: elementary, rapid and reliable automated image analysis. *Histopathology*. 2016;68(6):888-896. doi:10.1111/his.12867

343. Halama N, Zoernig I, Spille A, et al. Estimation of immune cell densities in immune cell conglomerates: an approach for high-throughput quantification. *PLoS One*. 2009;4(11).

344. Sander B, de Jong D, Rosenwald A, et al. The reliability of immunohistochemical analysis of the tumor microenvironment in follicular lymphoma: a validation study from the Lunenburg Lymphoma Biomarker Consortium. *Haematologica*. 2014;99(4):715-725. doi:10.3324/haematol.2013.095257

345. De Meulenaere A, Vermassen T, Creytens D, et al. Importance of choice of materials and methods in PD-L1 and TIL assessment in oropharyngeal squamous cell carcinoma. *Histopathology*. 2018;73(3):500-509. doi:10.1111/his.13650

346. Hartman DJ, Ahmad F, Ferris RL, Rimm DL, Pantanowitz L. Utility of CD8 score by automated quantitative image analysis in head and neck squamous cell carcinoma. *Oral Oncol*. 2018;86:278-287. doi:10.1016/j.oraloncology.2018.10.005

347. Bloom K, Harrington D. Enhanced Accuracy and Reliability of HER-2/neu Immunohistochemical Scoring Using Digital Microscopy. *Am J Clin Pathol*. 2004;121(5):620-630. doi:10.1309/Y73U8X72B68TMGH5

348. Terry J, Torlakovic EE, Garratt J, et al. Implementation of a Canadian external quality assurance program for breast cancer biomarkers: An initiative of Canadian Quality Control in Immunohistochemistry (cIQc) and Canadian Association of Pathologists (CAP) national standards committee/immunohistoc. *Appl Immunohistochem Mol Morphol*. 2009;17(5):375-382. doi:10.1097/PAI.0b013e31819adacf

349. Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*. 2009;27(8):1160.

350. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. REporting recommendations for tumor MARKer prognostic studies (REMARK). *Breast Cancer Res Treat*. 2006;100(2):229-235.

351. Wolff AC, Hammond MEH, Hicks DG, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *Arch Pathol Lab Med*. 2014;138(2):241-256.

352. Wolff AC, Hammond MEH, Schwartz JN, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Arch Pathol Lab Med*. 2007;131(1):18-43.

353. Sirota RL. Defining error in anatomic pathology. *Arch Pathol Lab Med*. 2006;130(5):604-606.

354. Tsakiroglou AM, Fergie M, Oguejiofor K, et al. Spatial proximity between T and PD-L1 expressing cells as a prognostic biomarker for oropharyngeal squamous cell carcinoma. *Br J Cancer*. 2020;122(4):539-544. doi:10.1038/s41416-019-0634-z

355. Galon J, Mlecnik B, Bindea G, et al. Towards the introduction of the “Immunoscore” in the classification of malignant tumours. *J Pathol*. 2014;232(2):199-209. doi:10.1002/path.4287

356. Schreiber RD, Old LJ, Smyth MJ. Cancer immunoediting: Integrating the role of immunity in cancer suppression and promotion. *Science (80-)*. 2011;331(March):78.

357. Wang D, Du Bois RN. Immunosuppression associated with chronic inflammation in the tumor microenvironment. *Carcinogenesis*. 2015;36(10):1085-1093. doi:10.1093/carcin/bgv123

358. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer*. 2012;12(4):252.

359. Ribas A, Wolchok JD. Cancer immunotherapy using checkpoint blockade. *Science (80-)*. 2018;359(6382):1350-1355. doi:10.1126/science.aar4060

360. Topalian SL. Targeting immune checkpoints in cancer therapy. *Jama*. 2017;318(17):1647-1648.

361. Chae YK, Arya A, Iams W, et al. Current landscape and future of dual anti-CTLA4 and PD-1/PD-L1 blockade immunotherapy in cancer; lessons learned from clinical trials with melanoma and non-small cell lung cancer (NSCLC). *J Immunother cancer*. 2018;6(1):39.

362. Ghatalia P, Zibelman M, Geynisman DM, Plimack ER. Checkpoint inhibitors for the treatment of renal cell carcinoma. *Curr Treat Options Oncol*. 2017;18(1):7.

363. Zolkind P, Uppaluri R. Checkpoint immunotherapy in head and neck cancers. *Cancer Metastasis Rev*. 2017;36(3):475-489.

364. Solomon B, Young RJ, Rischin D. Head and neck squamous cell carcinoma: genomics and emerging biomarkers for immunomodulatory cancer treatments. *Semin Cancer Biol*. 2018;52:228-240.

365. Shen X, Zhao B. Efficacy of PD-1 or PD-L1 inhibitors and PD-L1 expression status in cancer: meta-analysis. *BMJ*. 2018;362:k3529.

366. Ou D, Adam J, Garberis I, et al. Clinical relevance of tumor infiltrating lymphocytes, PD-L1 expression and correlation with HPV/p16 in head and neck cancer treated with bio- or chemo-radiotherapy. *Oncoimmunology*. 2017;6(9):e1341030. doi:10.1080/2162402X.2017.1341030

367. Feng Z, Bethmann D, Kappler M, et al. Multiparametric immune profiling in HPV–oral squamous cell cancer. *JCI Insight*. 2017;2(14):pii: 93652. doi:10.1172/jci.insight.93652

368. Kohlberger T, Liu Y, Moran M, et al. Whole-Slide Image Focus Quality: Automatic Assessment and Impact on AI Cancer Detection. Accessed September 12, 2019. http://arxiv.org/abs/1901.04619

369. Hossain MS, Kimura F, Yagi Y, Yamaguchi M, Nakamura T. Practical image quality evaluation for whole slide imaging scanner. In: *SPIE Proceedings Biomedical Imaging and Sensing Conference*. ; 2018:60. doi:10.1117/12.2316764

370. Greenwood M. The natural duration of cancer (report on public health and medical subjects no 33). *London Station Off*. 1926;(33).

371. Tsakiroglou AMAM, Fergie M, West C, et al. Quantifying cell-type interactions and their spatial patterns as prognostic biomarkers in follicular lymphoma. In: *SPIE Proceedings Medical Imaging*. Vol 10581. ; 2018:15. doi:10.1117/12.2293572

372. Herbst RS, Soria JC, Kowanetz M, et al. Predictive correlates of response to the

anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature*. 2014;515(7528):563-567. doi:10.1038/nature14011

373. Kobayashi K, Hisamatsu K, Suzui N, Hara A, Tomita H, Miyazaki T. A Review of HPV-Related Head and Neck Cancer. *J Clin Med*. 2018;7(9):241. doi:10.3390/jcm7090241

374. Tobin JWD, Keane C, Gunawardana J, et al. Progression of disease within 24 months in follicular lymphoma is associated with reduced intratumoral immune infiltration. *J Clin Oncol*. Published online 2019:JCO.18.02365. doi:10.1200/jco.18.02365

375. Casulo C, Le-Rademacher J, Dixon J, et al. Validation of POD24 as a robust early clinical endpoint of poor survival in follicular lymphoma: results from the Follicular Lymphoma Analysis of Surrogacy Hypothesis (FLASH) investigation using individual data from 5,453 patients on 13 clinical trials. *Blood*. 2017;130(Supplement 1):412. doi:10.1182/blood.V130.Suppl_1.412.412

376. Lu S, Stein JE, Rimm DL, et al. Comparison of Biomarker Modalities for Predicting Response to PD-1/PD-L1 Checkpoint Blockade: A Systematic Review and Meta-analysis. *JAMA Oncol*. 2019;5(8):1195-1204. doi:10.1001/jamaoncol.2019.1549

377. Sugimoto T, Watanabe T. Follicular Lymphoma: The Role of the Tumor Microenvironment in Prognosis. *J Clin Exp Hematop*. 2016;56(1):1-19. doi:10.3960/jslrt.56.1

378. Greaves P, Clear A, Coutinho R, et al. Expression of FOXP3, CD68, and CD20 at diagnosis in the microenvironment of classical hodgkin lymphoma is predictive of outcome. *J Clin Oncol*. 2013;31(2):256-262. doi:10.1200/JCO.2011.39.9881

379. Amé-Thomas P, Le Priol J, Yssel H, et al. Characterization of intratumoral follicular helper T cells in follicular lymphoma: role in the survival of malignant B cells. *Leukemia*. 2012;26(5):1053-1063. doi:10.1038/leu.2011.301

380. Wahlin BE, Sander B, Christensson B, et al. Entourage: The immune microenvironment following follicular lymphoma. *Blood Cancer J*. 2012;2(1). doi:10.1038/bcj.2011.53

381. Ai WZ, Hou JZ, Zeiser R, Czerwinski D, Negrin RS, Levy R. Follicular lymphoma B cells induce the conversion of conventional CD4 + T cells to T-regulatory cells. *Int J Cancer*. 2009;124(1):239-244. doi:10.1002/ijc.23881

382. Potts SJ, Krueger JS, Landis ND, et al. Evaluating tumor heterogeneity in immunohistochemistry-stained breast cancer tissue. *Lab Invest*. 2012;92(9):1342-1357. doi:10.1038/labinvest.2012.91

383. Park SY, Gönen M, Kim HJ, Michor F, Polyak K. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *J Clin Invest*. 2010;120(2):636-644. doi:10.1172/JCI40724

384. Tsakiroglou, Anna Maria Astley S, Dave M, Fergie M, et al. Tumour Infiltrating Lymphocytes in Follicular Lymphoma - additional data H&E. Mendley data. doi:10.17632/sxxfr3hf78.1

385. Horn H, Schmelter C, Leich E, et al. Follicular lymphoma grade 3B is a distinct neoplasm according to cytogenetic and immunohistochemical profiles. *Haematologica*. 2011;96(9):1327-1334.

386. Contal C, O'Quigley J. An application of changepoint methods in studying the effect of age on survival in breast cancer. *Comput Stat Data Anal*. 1999;30(3):253-

270. doi:10.1016/S0167-9473(98)00096-6

387. Meyers JP, Mandrekar JN, Clinic M. Cutpoint Determination Methods in Survival Analysis using SAS ® : Updated % FINDCUT macro. 2015;(1997):1-8.

388. McNeel DG. TCR diversity–a universal cancer immunotherapy biomarker? *J Immunother cancer*. 2016;4(1):1-4.

389. Postow MA, Manuel M, Wong P, et al. Peripheral T cell receptor diversity is associated with clinical outcomes following ipilimumab treatment in metastatic melanoma. *J Immunother cancer*. 2015;3(1):23.

390. Manuel M, Trédan O, Bachelot T, et al. Lymphopenia combined with low TCR diversity (divpenia) predicts poor overall survival in metastatic breast cancer patients. *Oncoimmunology*. 2012;1(4):432-440.

391. Sheikh N, Cham J, Zhang L, et al. Clonotypic diversification of intratumoral T cells following sipuleucel-T treatment in prostate cancer subjects. *Cancer Res*. 2016;76(13):3711-3718.

392. Gül N, van Egmond M. Antibody-dependent phagocytosis of tumor cells by macrophages: a potent effector mechanism of monoclonal antibody therapy of cancer. *Cancer Res*. 2015;75(23):5008-5013.

393. Hedvat C V., Hegde A, Chaganti RSK, et al. Application of tissue microarray technology to the study of non-Hodgkin's and Hodgkin's lymphoma. *Hum Pathol*. 2002;33(10):968-974. doi:10.1053/hupa.2002.127438

394. Lau D, Bobe AM, Khan AA. RNA sequencing of the tumor microenvironment in precision cancer immunotherapy. *Trends in cancer*. 2019;5(3):149-156.

395. Titterington DM. Bayesian methods for neural networks and related models. *Stat Sci*. 2004;19(1):128-139.