# GERMLINE AND SOMATIC DETERMINANTS OF THE IMMUNE CONTEXTURE IN COLORECTAL CANCER

by

TORITSEJU OLUWAFUNMILAYO SILLO

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

Institute of Immunology and Immunotherapy
College of Medical and Dental Sciences
University of Birmingham
May 2021

## ABSTRACT

**Aims:** Microsatellite instability is a recognised marker for determining the efficacy of immunotherapy in colorectal cancer (CRC). However, other immunogenomic markers could also drive the anti-tumour immune response. This thesis explores the hypothesis that both germline and somatic factors are significant in shaping the immune microenvironment in CRC. Associations were studied between germline immune gene expression quantitative trait loci (eQTLs) and a phenotypic marker of the tumour immune environment – the Immunoscore®. The somatic component focused on the contributions of tumour mutational burden, neoantigen clonality and the microbiome to the immune microenvironment.

**Methods:** This *in silico* analysis utilised genomic data and tumour samples from 200 patients with CRC enrolled in the 100 000 Genomes Project. From germline data, a panel of eQTLs was correlated with the Immunoscore using logistic regression. Somatic whole genome sequencing data was used to correlate tumour mutational burden (TMB) and neoantigen clonality with the Immunoscore. Metagenomic analysis was also performed on somatic data. Finally, RNA sequencing and immunohistochemistry on formalin-fixed tumour tissue were performed to corroborate these results.

**Results:** eQTLs associated with differences in the Immunoscore included *TCF7, BCL11B, CCR1, CSK, IL19, IL23R* and *BCL10*. While TMB was not significantly associated with the Immunoscore, a combination of neoantigen burden and neoantigen clonality (intratumoral heterogeneity) was strongly associated with the Immunoscore and clinical outcomes, independent of microsatellite status.

RNA sequencing confirmed that the expression of major histocompatibility complex Class II genes, gut-bacteria-associated chemokines, and a T-helper centric metagene, were also all strongly associated with the Immunoscore.

**Conclusions:** Germline factors contribute to variability in the colorectal tumour immune contexture, particularly in MSS CRC. These effects are modulated by the contributions of somatic determinants, particularly the combination of neoantigen burden and clonality, which point to potential new biomarkers for determining the response to immunotherapy in colorectal cancer.

## DEDICATION

To my father, Dr Cyril Eyewu Sillo, always my inspiration; and to my mother,

Mrs Oluwatoyin Olufunke Sillo, my guide and champion.

## ACKNOWLEDGEMENTS

I would like to give massive thanks to my supervisors Professor Gary Middleton and Professor Andrew Beggs for their support, warmth and good humour through my time working with them on this project. I was a complete novice in this field when they took me on, and helped to build my skills, confidence and optimism. I cannot thank them enough for all they have done.

To my Mentor, Professor Dion Morton OBE, thank you so much for believing in me. Your kind words and advocacy took me through some of the most challenging times, and I would not have been able to get the most out of my research experience without you championing my cause.

To the Beggs group team – thank you all so much for being so welcoming, and for your patience and practical help during my time in the laboratory. I would also like to thank the Beggs group Clinical Research Fellows for being a great bunch to work with.

I would like to thank Professor Trevor Graham's group, particularly, Dr Eszter Lakatos, for her dedicated help with the neoantigen clonality analysis; Dr Nicky McGranahan for his insights into tumour heterogeneity analysis, and Professor Jerome Galon's team at HalioDx for their help with the Immunoscore. I would also like to thank Dr Phillipe Taniere for his help with the Class II immunohistochemistry.

To Pauline Goddard, Professor Middleton's PA, for her warmth, kindness and help with so many administrative tasks that could have been overwhelming.

To my family and friends for their love, care, and tolerance in putting up with all my whining during the most stressful times. To Mum, Teri, Bemi and Sisan, thank you for being the best family ever. Onwards and upwards! Special thanks to Sisan for being my lockdown chef and therapist. To my friends, particularly Damien, Meera, Deepa, Misha, Prinith, Jenny, Louise, Anna-Maria, Nic and Tehmi. Thank you all for believing in my ability to complete this successfully. To Amrit and Kathryn – you special women! I am grateful to have met you both.

To my church family – thank you for your prayers, and for encouraging my faith at difficult moments.

My acknowledgements also go to the Sandwell and West Birmingham NHS Trust General Surgery and University Hospitals Birmingham NHS Trust Renal Surgery departments for providing me with Research Fellow posts during this time. I am thankful to Birmingham Health Partners for providing me with a Research Fellowship, and the Bowel Disease Research Foundation and Cancer Research UK for providing an Open Grant to fund some of this work.

Finally, and most importantly, I would like to thank the patients and their loved ones who signed up to the 100 000 Genomes Project. This thesis was only possible due to their willingness to participate in research for the greater good.

## 100 000 GENOMES PROJECT STATEMENT

This research was made possible through access to the data and findings generated by the 100 000 Genomes Project. The 100 000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100 000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support.

## COVID-19 IMPACT STATEMENT

My time out of surgical training for research from October 2017 to October 2020 overlapped with the arrival of the COVID-19 pandemic. In March 2020, the University and Laboratories were closed at short notice, and I was re-drafted to perform clinical duties in the UHB NHS Trust Emergency General Surgery Department.

I was fortunate to be continue *in silico* research and data analysis on this project along with my clinical duties. Funding from the BDRF Open Grant enabled the University to supply computing resources to my house, and my supervisors continued intensive remote supervision. With these, I was able to complete the project and submit my thesis. I am immensely privileged to have had the opportunity to contribute to the Health Service at a time of crisis, and to have also produced a body of work worthy of being awarded a PhD.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## APPENDICES

**ABBREVIATIONS**

100KGP – 100 000 genomes project

ASCAT – allele-specific copy number analysis of tumors

ANICCA – a phase II trial assessing nivolumab in strong class II expressing microsatellite stable colorectal cancer

ANOVA – analysis of variance

APC – adenomatous polyposis coli

BAM – binary alignment map

BCL – B-cell lymphoma

bcl – binary base call

BEAR – Birmingham Environment for Academic Research

BRAF – v-raf murine sarcoma viral oncogene homolog B1

C5 – complement 5

CAR-T – chimeric antigen receptor T cell

CCL – c-c motif chemokine ligand

CCR5 – c-c chemokine receptor type 5

CD – cluster of differentiation

CEA-TCB – carcinoembryonic antigen T-cell bispecific antibody

CI – confidence interval

CIITA – Class II, major histocompatibility complex, transactivator

CIRC – Co-ordinate Immune Response Cluster

CMS – consensus molecular subtypes

CNA – copy number alteration

COX2 – cyclo-oxygenase 2

CRC – colorectal cancer

CSK – C- terminal src kinase

CT – centre of the tumour

CTLA-4 – cytotoxic T lymphocyte-associated protein 4

dMMR – deficient mismatch repair

DNA – deoxyribonucleic acid

DSS – disease-specific survival

DFS – disease-free survival

EMVI – extramural venous invasion

eQTL – expression quantitative trait loci

FAP – familial adenomatous polyposis

FoxP3 – forkhead box P3

FDR – false discovery rate

FMT – faecal mucosal transplantation

FFPE – formalin-fixed paraffin-embedded

FOLFOX - leucovorin, fluorouracil, and oxaliplatin

FOxTROT – Fluoropyrimidine, Oxaliplatin and Targeted Receptor Pre-Operative Therapy: A Controlled Trial in High-Risk Operable Colon Cancer

GeCIP – Genomics England Clinical Interpretation Partnership

GeL – Genomics England

GTEx – Genotype-Tissue Expression Project

GWAS – genome wide association studies

HBRC – Human Biomaterials Resource Centre

HLA – human leucocyte antigen

HNPCC – hereditary non-polyposis colon cancer

HNSCC – head and neck squamous cell carcinoma

HPC – high performance computing

HR – hazard ratio

ICB – immune checkpoint blockade

ICGC - The International Cancer Genome Consortium

IDEA - International Duration Evaluation of Adjuvant Therapy

IDO – indoleamine 2,3-dioxygenase

IHC – immunohistochemistry

IFNγ – interferon gamma

IL – interleukin

IM – invasive margin

Indel – insertion and deletion (of bases)

IS - Immunoscore®

ITH – intratumoral heterogeneity

KIR – killer immunoglobulin-like receptor

KRAS – Kirsten rat sarcoma virus oncogene

LAG3 – lymphocyte activation gene 3

LCL – lymphoblastoid cell line

LD – linkage disequilibrium

LEF – lymphoid enhancer-binding

MAD – median absolute deviation

MAF – minor allele frequency

MATH – mutant-allele tumour heterogeneity

Mbp – megabase pair

MDSC – myeloid derived suppressor cell

MHC I – major histocompatibility complex Class I

MHC II – major histocompatibility complex Class II

MSI – microsatellite instability

MuTHER – Multiple Tissue Human Expression Resource

mRNA – messenger RNA

NHS – National Health Service

NGS – next generation sequencing

NK – natural killer cell

NKG2D – natural killer G2D receptor

NSCLC – non-small cell lung cancer

ORR – objective response rate

OS – overall survival

OTU – operational taxonomic unit

PCA – principal components analysis

PCAWG – The International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) Pan-Cancer Analysis of Whole Genomes

PCR – polymerase chain reaction

PD-1 – programmed cell death protein 1

PD-L1 – programmed cell death ligand 1

PFS – progression-free survival

PIK3CA – phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha

pMMR – mismatch repair proficient

POLD1 – DNA polymerase delta

POLE – DNA polymerase epsilon

QA – quality alignment

QC – quality control

RFS – recurrence-free survival (used interchangeably with DFS)

RNA – ribonucleic acid

RIN – RNA integrity number

RNAseq – RNA sequencing

rRNA – ribosomal RNA

SB – strong binder

SITC – Society for Immunotherapy of Cancer

SMAD4 – Mothers against decapentaplegic homolog 4

SNP – single nucleotide polymorphism

SNV – single nucleotide variant

STAR – Spliced Transcripts Alignment to a Reference

TCF – transcription factor

TCGA – The Cancer Genome Atlas

TCR – T cell receptor complex

TGF-β – transforming growth factor beta

TIL – tumour-infiltrating lymphocytes

Tim-3 – T cell immunoglobulin and mucin domain-containing protein 3

TNF – tumour necrosis factor

TNM – tumour-node-metastasis

Treg – regulatory T cell

TSS – transcription start site

TMB – tumour mutational burden

TP53 – tumour protein 53

UBASH3A – ubiquitin-associated and SH3 domain-containing protein A

UHB – University Hospitals Birmingham

UICC – Union for International Cancer Control

UK – United Kingdom

UL16BP – UL16 binding proteins

UMI – unique molecular identifier

VAF – variant allele frequency

VCF – variant call format

VELIPI – venous emboli, lymphatic and perineural invasion of cancer

WB – weak binder

WES – whole exome sequencing

WMGMC – West Midlands Genomic Medicine Centre

WGS – whole genome sequencing

# Chapter 1: Introduction

## 1.1. Background

Colorectal cancer (CRC) is the fourth commonest cancer diagnosed but the second commonest cause of cancer death worldwide [1]. Considerable research has been done in understanding carcinogenesis [2], early detection [3] and risk factor modification [4, 5] in CRC. The progression from normal colorectal mucosa to adenoma and then carcinoma is well studied [2, 6, 7]. The first hit is usually a mutation in a key tumour suppressor gene (often the adenomatous polyposis coli (*APC*) gene, mutations of which are present in approximately 80% of CRCs [8]), which induce a selective growth advantage in a clone of cells. Subsequent accumulation of mutations in other driver genes (most notably the Kirsten ras sarcoma (*KRAS*) oncogene) drive carcinogenesis (Figure 1.1). Although a notably simplified paradigm, this is the accepted pathway for colorectal tumorigenesis in the majority of patients, as it has proved to be a useful model in understanding the pathogenesis of CRC. CRC has a significant genetic component, with an estimated 35% heritability [9, 10], aside from the hereditary forms including familial adenomatous polyposis, hereditary non-polyposis colon cancer (Lynch syndrome), juvenile polyposis and Peutz-Jeghers syndrome, which account for less than 5% of cases [11].

## 1.2. The treatment of colorectal cancer

The majority of patients undergo surgical treatment along with adjuvant or neo-adjuvant therapies (chemotherapy and radiotherapy) [12]. The Union for International Cancer Control (UICC) tumour-node-metastasis (TNM)

classification [13] (Table 1.1), provides key prognostic information and guides therapeutic decisions including the rationale for neo/adjuvant therapy [14]. However, there is significant variability in outcomes within disease stages [15], and the TNM classification does not provide sufficient information on response to therapies, or on additional risk factors that increase the risk of relapse after treatment.

There is some divergence in the treatment of tumours arising in the colon compared with those arising in the rectum. For both colon and rectal tumours, patients with early stage (Stage 1 and 2) tumors usually have surgery with curative intent. For locally advanced tumours (Stage 3), those with rectal cancer usually undergo neo-adjuvant chemo-radiotherapy, with the aim of downstaging the tumour, prior to surgery with curative intent. For those with colon tumours, surgery is offered, followed by adjuvant chemotherapy [16]. Some promising clinical trials are underway [17], analysing the viability and benefits of neo-adjuvant chemotherapy in this group, in particular, the Fluoropyrimidine, Oxaliplatin and Targeted Receptor Pre-Operative Therapy: a Controlled Trial in High-Risk Operable Colon Cancer (FOxTROT) trial [18, 19].

Despite advances in screening and early detection of CRC, about 21% of patients have metastatic (Stage 4) disease at initial presentation of which 80 to 90% are not curable with surgical resection. Furthermore, up to 50 to 60% of patients with earlier stage disease eventually develop metastases [20]. For these patients, prognosis is usually poor. Isolated and surgically-resectable metastatic lesions can be treated with curative intent, but for the majority the mainstay of treatment is the use of chemotherapy or targeted radiotherapy, with modest survival benefit

[21]. In the past decade, some biologic and targeted therapies have emerged, which add little survival benefit at high cost and toxicity [20] [22].

Immunotherapy has emerged in the past decade as a promising therapy. However, it is currently licensed for use only in patients with metastatic CRC due to deficient DNA mismatch repair (dMMR), also termed microsatellite instability (MSI)-high CRC, following failure of other systemic treatment. These patients only represent 3-4% of the total patient cohort [23], in comparison to those with mismatch repair proficient (pMMR) CRC who represent the majority. This severely limits the wider application and integration of immunotherapy into the standard treatment pathway for CRC.

A significant recent exploratory study, the Nivolumab, Ipilimumab and COX2-inhibition in Early Stage Colon Cancer (NICHE) trial, reported the results of neo-adjuvant immunotherapy in earlier stage (I to III) pMMR and dMMR CRC [24]. There were significant pathological responses to therapy both in patients with dMMR CRC and in some patients with pMMR CRC who had high levels of infiltration with effector T cells. Their results suggest that neo-adjuvant and adjuvant immunotherapy could have true potential in expanding the treatment options for patients with CRC at all disease stages.

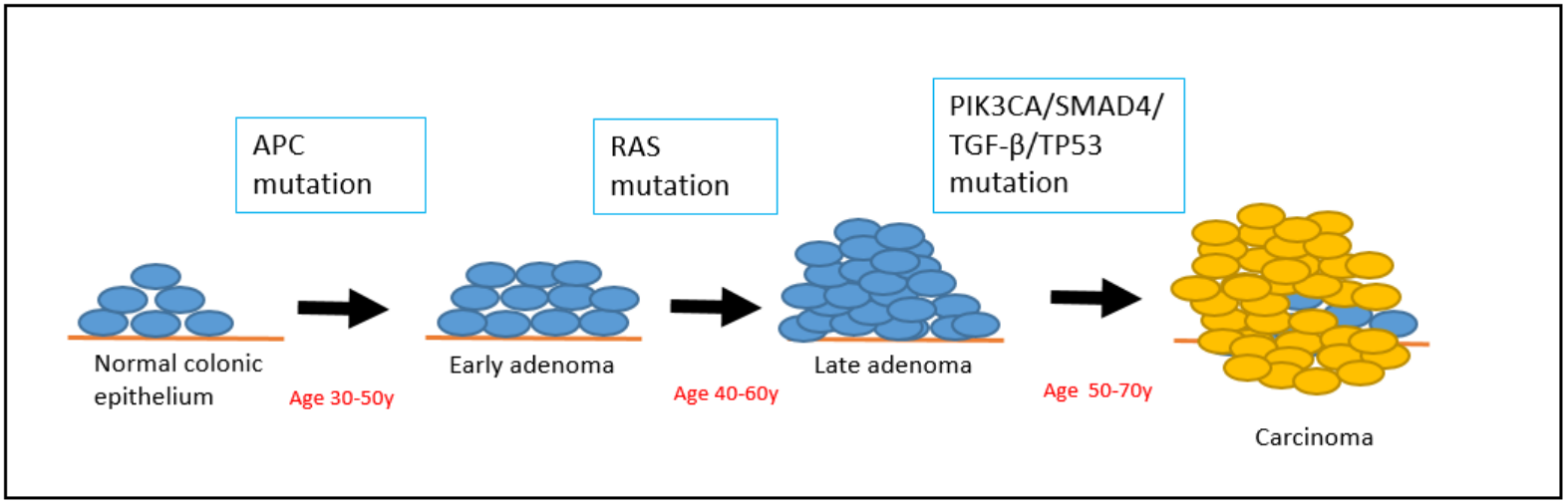



**Figure 1.1**: Simplified adenoma-carcinoma sequence illustrating progressive accumulation of mutations in colorectal cancer. APC – adenomatous polyposis coli. RAS – Kirsten ras oncogene. PIK3CA – phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha. SMAD4 – Mothers against decapentaplegic homolog 4 gene. TGF-β – transforming growth factor beta gene. Adapted from [7].

**Table 1.1**. The UICC TNM classification of CRC

| Definition | |
|---|---|
| **T stage** | |
| Tx | No information about local tumour infiltration available |
| T1 | Infiltration through lamina muscularis mucosae into submucosa, no infiltration of lamina muscularis propria |
| T2 | Infiltration into, but not beyond, lamina muscularis propria |
| T3 | Infiltration into subserosa or non-peritonealised pericolic or perirectal tissue, or both; no infiltration of serosa or neighbouring organs |
| T4a | Infiltration of the serosa |
| T4b | Infiltration of neighbouring tissues or organs |
| **N stage** | |
| Nx | No information about lymph node involvement available |
| N0 | No lymph node involvement |
| N1a | Cancer cells detectable in 1 regional lymph node |
| N1b | Cancer cells detectable in 2–3 regional lymph nodes |
| N1c | Tumour satellites in subserosa or pericolic or perirectal fat tissue, regional lymph nodes not involved |
| N2a | Cancer cells detectable in 4–6 regional lymph nodes |
| N2b | Cancer cells detectable in 7 or greater regional lymph nodes |
| **M stage** | |
| Mx | No information about distant metastases available |
| M0 | No distant metastases detectable |
| M1a | Metastasis to 1 distant organ or distant lymph nodes |
| M1b | Metastasis to more than 1 distant organ or set of distant lymph nodes |
| M1c | Peritoneal metastases with or without organ involvement |

Adapted from [13] and [14]. T – tumour, N – node, M – metastasis.

### 1.3. Colorectal cancer immuno-genomics

The immune response has been known to play a significant role in cancer outcomes for over a century. William Coley treated a range of sarcomas and carcinomas with a bacterial vaccine in the late 1800s [25]. Building on his observation in a patient with sarcoma whose tumour spontaneously regressed following development of a bacterial infection, he hypothesised that the infection caused tumour regression. He developed a vaccine called “Coley’s toxins”, containing two killed bacteria (*Streptococcus pyogenes* and *Serratia marcescens*). With this, he successfully treated hundreds of patients with advanced sarcomas. Coley’s toxins were also used to treat lymphomas, melanomas, myelomas and carcinomas.

However, in the early 20$^{th}$ century, this treatment fell out of favour. First, due to the widespread acceptance of aseptic techniques, cancer surgery became sterile, and deliberately inducing infection in patients in order to cure cancer was seen as regressive. Secondly, surgery, chemotherapy and radiotherapy emerged as the primary treatment modalities for most cancer types. They could be standardised, and had great promise as potential cures for cancer [25]. The potential for harnessing the immune system to target cancer treatment was overlooked.

However, in the past decade the emergence of immunotherapy with immune checkpoint blockade (ICB) agents has transformed the treatment landscape of some cancers, most strikingly in cutaneous melanoma [26] [27] and lung cancer

[28]. Primarily, their mode of action is by potentiating the immune response to tumour antigens [29].

Thus far, the role of immunotherapy in the treatment of CRC has been limited to the 3-4% of patients with metastatic disease whose tumours demonstrate microsatellite instability [23, 30]. Improved insights into the mechanisms underpinning the immune microenvironment in CRC are helping to develop the role of immunotherapy and suggest potential targeted approaches to its management in a wider patient cohort.

### *1.3.1. Targeted therapies in cancer immunotherapy*

Proposed advantages of targeted cancer immunotherapy include increased efficacy and specificity in targeting cancer rather than normal tissue, resulting in lower toxicity than current treatments. The focus of therapeutic targeting is either boosting the T cell response to tumour neoantigens, for example, adoptive cell transfer, chimeric antigen receptor T-cell (CAR-T) therapy and ICB, or altering the antigen landscape to favour the expression of those which are highly immunogenic [31]. The potential for use in solid tumours, such as breast cancer [32] is under exploration.

#### *1.3.1.1. Adoptive cell transfer and CAR-T therapy*

Adoptive cell transfer of T cells recognising certain tumour antigens has been shown to induce tumour regression in some trials, most notably in melanoma [33]. For lymphoid malignancies, particularly diffuse large B cell lymphoma and acute lymphoblastic lymphoma, CAR-T therapy has offered a new treatment option for those with refractory disease. The current therapeutic regimen involves the use

of the patient's autologous T cells, which are collected and genetically modified by using a lentivirus or retrovirus vector to transduce a chimeric antigen receptor fusion protein, which is specific for a tumour antigen (usually CD19). These modified T cells are expanded *ex vivo*, and then re-infused into the patient following a leukodepletion conditioning regimen [34]. Current CARs consist of a single-chain variable fragment, an antigen-recognition domain, a CD3-derived T-cell activation domain, and a co-stimulatory domain. The results have been impressive, with some trials showing complete responses in up to 80% of patients with acute lymphoblastic leukaemia [35, 36] and 40% to 60% of patients with aggressive lymphomas [34, 37]. In one study, high serum interleukin (IL)-15 levels correlated with peak CAR-T levels and remission of lymphoma [37].

However, these trials have had small patient numbers, and significant toxicity, particularly cytokine release syndrome and neurotoxicity are reported in between 10% and 50% of trial participants [34]. Cytokine release syndrome is likely due to the targeting of antigens which are also expressed to some degree on normal tissue as well as tumour tissues, and mimics a systemic inflammatory response syndrome with fever, haemodynamic instability and end-organ dysfunction [38]. There are also significant logistical challenges with delivery of adoptive cell therapy, not the least of which is the cost involved [39]. Finally, translation to solid malignancies has been hampered by therapeutic barriers, including challenges in determining which specific tumour antigens to target, variability in the trafficking of T cells into solid tumours and the fate and effectiveness of these cells within the tumour environment [33].

### *1.3.1.2. Immune checkpoint blockade*

ICB with cytotoxic T lymphocyte-associated protein 4 (CTLA-4), programmed cell death protein 1 (PD-1) and programmed cell death protein ligand 1 (PD-L1) antibodies have been shown to reactivate *in vivo* tumour infiltrating T cells leading to objective anti-tumour responses – specifically, tumour regression – in some cancers. This has been most marked in tumours with high mutation rates such as melanoma and lung cancer, but has also been seen in renal cancer (which has a low mutation rate) [40].

In CRC, initial Phase I studies showed poor or no objective clinical benefit in patients with advanced disease [41, 42]. However, when Le *et al.* compared outcomes in patients with or without dMMR who were given Pembrolizumab, an anti PD-1 antibody, there were immune-related objective response rates (ORR) and progression-free survival (PFS) of 40% and 78% respectively in patients with dMMR; compared to 0% and 11% respectively in patients without dMMR. This was associated with a significantly reduced hazard ratio (HR) for death or disease progression (HR = 0.10, $p < 0.001$) in the dMMR group, and a mean rate of 1782 somatic mutations per tumour compared with 73 per tumour in pMMR, or microsatellite stable (MSS) tumours. Furthermore, high levels of somatic mutations correlated with improved survival [43]. This provided the rationale for the licensing of Pembrolizumab for use in dMMR/MSI-high CRC. There are several ongoing clinical trials assessing checkpoint blockade agents. These are predominantly in advanced or metastatic disease (MSI-high and MSS) in patients who have been heavily pre-treated (Table 1.2).

Dual ICB blockade has some supporting evidence. The NICHE trial of neo-adjuvant immunotherapy in non-metastatic CRC used combined anti-CTLA-4 (ipilimumab) and anti-PD-1 (nivolumab) therapy [24]. In 20 of 20 patients with dMMR tumours, pathological response was observed, with major responses in 19 of 20. Follow-up at data cut-off was for a median of approximately 9 months in both groups. In pMMR tumours, 4 of 15 patients (27%) showed pathological responses. These patients' tumours showed higher levels of CD8+PD-1+ T cell infiltration but not increased tumour mutational burden (TMB), compared with non-responders. On the other hand, the Canadian Cancer Trials Group CO.26 study assessed the effect of combined immune checkpoint inhibition with anti-CTLA-4 (tremelimumab) and anti-PD-L1 (durvalumab) blockade compared with best supportive care alone in patients with advanced CRC [44]. Median overall survival (OS) was 6.6 months in those who had dual checkpoint blockade and 4.1 months in those who had best supportive care (HR 0.72; 90% CI, 0.54-0.97; $P = 0.07$). They found that the OS was significantly improved with dual checkpoint blockade in MSS CRC, particularly in those with TMB of 28 variants per megabase or more.

These apparently conflicting results suggest that other parameters separate from TMB are driving this response, which are prime targets for further investigation. Although long-term survival data is not yet available, this is encouraging data and further supports the rationale for neo-adjuvant immunotherapy not only in patients with dMMR CRC, but also selected patients with pMMR CRC, whose tumours show high markers of immune infiltration.

**Table 1.2.** Clinical trials of immunotherapy in CRC [45]

| Phase | Reference (Trial name) | Regimen | Subgroups | Outcomes | Follow-up duration |
|---|---|---|---|---|---|
| Phase II | Le *et al.* 2015 [43] | PD-1 inhibitor (pembrolizumab) | dMMR/MSI-high vs MSS CRC | Immune-related ORR<br>PFS | 20 weeks |
| Phase II | Overman *et al*. 2018 [23]<br>(CheckMate 142) | PD-1 inhibitor (nivolumab) +/- CTLA-4 inhibitor (ipilimumab) | Metastatic pre-treated dMMR/MSI-high CRC | Immune-related ORR<br>PFS<br>OS | 12 months |
| Phase II | Mettu *et al.* 2018 [46]<br>(BACCI) | Capecitabine/ bevacizumab +/- PD-L1 inh bitor (atezolizumab) | Metastatic CRC | PFS<br>OS | Ongoing |
| Phase III | Hoffmann-La Roche [47] (COTEZO Imblaze370) | Cobimetinib + PD-L1 inh bitor (atezolizumab) vs atezolizumab vs regorafenib | Heavily pre-treated locally advanced or metastatic CRC<br>(>95% MSS) | OS<br>PFS | 3 years |
| Phase III | Diaz *et al.* 2017 [48]<br>(KEYNOTE-177) | PD-1 inhibitor (pembrolizumab) vs standard chemotherapy | dMMR/MSI-high Stage 4 CRC | PFS<br>OS | 57 months |
| Phase III | Asan Medical Center [49] (POLE-M) | Standard 5-FU-based adjuvant chemotherapy +/-sequential PD-L1 inh bitor (avelumab) | Resected stage 3 dMMR/MSI-high or POLE-mutant colon cancer | DFS | 5 years |
| Phase III | Sinicrope *et al.* 2017 [50]<br>(ATOMIC, Alliance A021502) | Combined chemotherapy +/- PD-L1 inhibitor (atezolizumab) as monotherapy for additional 6 months | Resected stage 3 dMMR/MSI-high colon carcinomas | DFS<br>OS<br>Adverse events | 5 years |
| Phase I | Tabernero *et al.* 2017 [51] | CEA-TCB antibody +/- PD-L1 inhibitor (atezolizumab) | Heavily pre-treated metastatic CRC (majority MSS) | Adverse events<br>Anti-tumour activity (RECIST v1.1 criteria [52])<br>PFS | 40 months |
| Phase I (exploratory) | Chalabi *et al.* 2020 [24] (NICHE) | Combined PD-1 inhibitor (nivolumab), CTLA-4 inh bitor (ipilimumab) +/- COX2-inhibition | dMMR and pMMR CRC, neo-adjuvant, stage 1 to 3 disease only | Adverse events<br>Immune activating capacity<br>RFS | 3-5 years (ongoing) |
| Phase II | Antoniotti *et al.* 2020 [53] (AtezoTRIBE) | Combined 5-FU-based chemotherapy + bevacizumab + PD-L1 inh bitor (atezolizumab) vs combination treatment | Unresected and previously untreated metastatic CRC, irrespective of MMR status | PFS<br>Overall toxicity rate<br>ORR | 24 months (ongoing) |
| Phase II | Chen *et al.* 2020 [44] (Canadian Cancer Trials Group CO.26) | Combined PD-L1 (durvalumab) and CTLA-4 inh bitor (tremelimumab) with based supportive care vs best supportive care | Pre-treated metastatic dMMR and pMMR CRC | OS<br>PFS<br>ORR | 15.2 months |

5-FU – 5-fluorouracil. CEA-TCB – carcinoembryonic antigen-T cell bispecific. COX2 – cyclo-oxygenase 2. CRC – colorectal cancer. CTLA-4 - cytotoxic T-lymphocyte-associated protein 4. DFS – disease-free survival. dMMR / MSI-high – deficient mismatch repair / microsatellite instability high. MSS – microsatellite stable. PD-1 – programmed cell death protein 1. PD-L1 – programmed cell death protein ligand 1. PFS – progression-free survival. POLE-M – mutated DNA polymerase epsilon. RECIST – Response Evaluation Criteria in Solid Tumours. RFS – recurrence-free survival. ORR – objective response rate. OS – overall survival.

Mouse studies using cancer vaccines – injection of specially prepared candidate neoantigen peptides designed to stimulate an immunological response to cancer – show that the type of neoantigen is of importance. So far, the role of cancer vaccines in solid tumours in humans remains at the experimental stage. In a clinical trial of three patients with melanoma, WES was used to identify the highest binding epitope peptides and these patients were vaccinated with autologous dendritic cells, which had been pulsed with the top 7 highest binding peptides identified from each tumour [54]. This led to an increase in the breadth and diversity of neoantigen specific T cells from all patients, with no adverse autoimmune events. However, there was no demonstrated objective clinical benefit. It remains uncertain if cancer vaccination is potent enough to induce remission in solid tumours. One issue is that for a neoantigen to induce an immune response, the T cell receptor (TCR) repertoire of the patient needs to contain a TCR that specifically recognises the peptide bound to a specific human leucocyte antigen (HLA) allele. While the TCR repertoire in any individual is sufficiently diverse that, in theory, it should be capable of recognising virtually any pathogen, mutated cancer peptides typically differ from innate peptides only slightly, often by a single amino-acid [55]. Another significant limiting factor is that neoantigens may be polyclonal due to intratumoral heterogeneity, thus hindering their identification [56]. Other limiting factors include the potential high cost, and the possibility of significant adverse reactions.

Neoantigens represent ideal targets for cancer immunotherapy, as they are expressed only in tumour cells and so are less likely to induce either

immunological tolerance or toxicity from targeted therapy. However, *in silico* neoantigen prediction tools have some challenges to overcome.

### 1.3.2. The major histocompatibility complex (MHC)

Most prediction tools target major histocompatibility complex (MHC) Class I epitopes only, but the role of MHC Class II expression in the anti-cancer response is likely to be of great significance.

The MHC gene complex is found in all higher vertebrates, and encodes proteins that are expressed on the cell surface. In humans, it is known as the HLA complex. The HLA gene complex is found on chromosome 6, is comprised of an expanding number of genes (more than 200 have been discovered), and is highly polymorphic [57].

The MHC (or HLA) complex has a significant role in destroying pathogens, by presenting antigens to T cells to enable direct killing of virus-infected cells, activation of B cells to produce antibodies to neutralise extracellular pathogens and activation of macrophages to kill bacteria within their intracellular vesicles. The MHC complex is also critical in the priming the recognition of self-antigens within the thymus, and the prevention of targeting of these [58]. There is strong selection pressure in favour of pathogens and tumour antigens that have mutated in a way that they can evade presentation by MHC molecules [57]. MHC Class I proteins are found on the surface of nearly every cell, and can present antigens to CD8+ T lymphocytes for direct cytotoxic cell killing. However, Class II proteins are only found on specialised antigen-presenting cells such as dendritic cells, B

cells and macrophages. They present antigens to CD4+ T cells, which then activate B cell and macrophages responses to these antigens (Figure 1.2). Due to the highly polymorphic nature of the MHC gene complex, every person has a set of MHC molecules with wide and differing ranges of peptide-binding specificities.

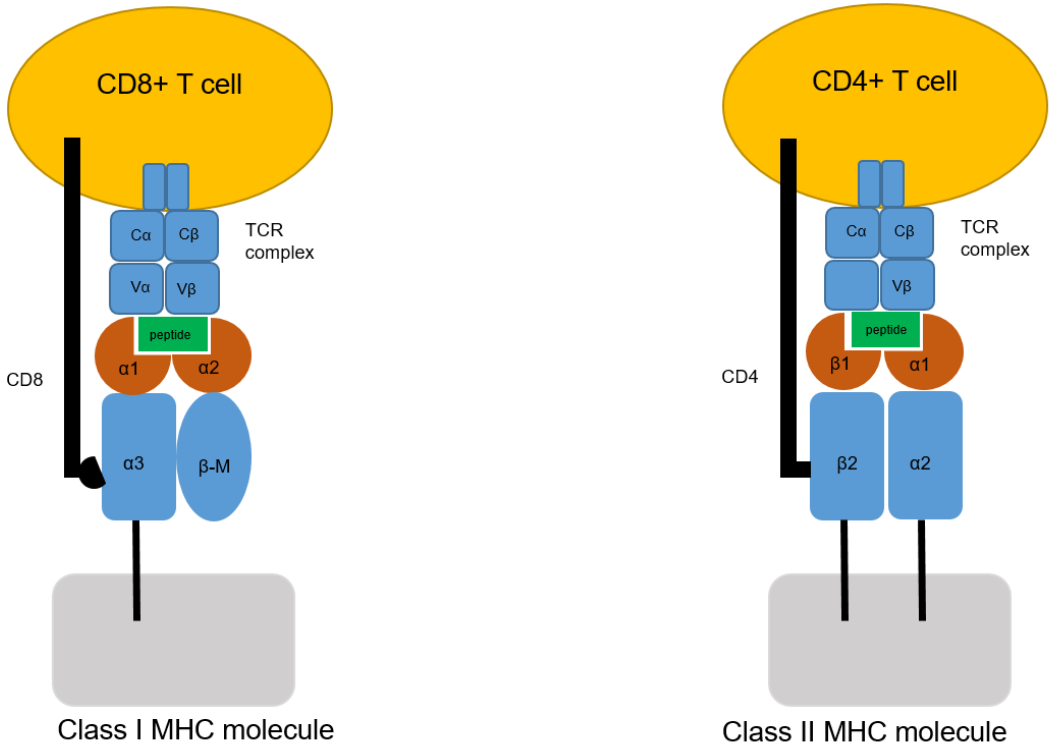



**Figure 1.2**. Illustration of the major histocompatibility complex (MHC) and its presentation of peptides to T cell receptors. The MHC Class I molecule has an α polypeptide and β2 microglobulin chain. It is expressed on all nucleated cells. The MHC Class II molecule is present only on antigen-presenting cells. It has an α and β polypeptide chain. MHC Class I molecules present antigens to CD8+ effector T cells, and MHC Class II molecules present antigens to CD4+ helper T cells.

#### *1.3.2.1. MHC Class II expression in colorectal cancer*

Class II expression is usually absent in normal colonic mucosal epithelium [59], however, it can be induced in other cell types, including tumours, particularly when stimulated by inflammatory signals such as interferon-gamma (IFN-γ).

In studies where CIITA, the master transcriptional activator of Class II, was transfected into poorly immunogenic Class II-negative murine adenocarcinoma cell lines, tumour elimination occurred [60, 61]. CIITA-transfected cancer cells developed robust antigen processing function, and CD4+ and CD8+ cells massively infiltrated CIITA-transfected tumours. In particular, the CD4+ cells took on the function of Th1 cells and produced IFNγ. Class II expression is seen in 25 to 50% of CRC [59, 62] . Class II expression is higher in well-differentiated and less invasive cancer. MHC Class II loss correlates with reduced TIL density and increased incidence of regional metastases [59].

In a study by Kreiter *et al.* using three different murine tumour models, mutated MHC Class II epitopes were more immunogenic than Class I epitopes [63]. Alspach *et al.* demonstrated in a mouse sarcoma model that co-expression of MHC Class I and II neoantigens were necessary for a response to ICB [64]. This was particularly striking in MHC Class II non-expressing tumours. In melanoma, tumour membrane Class II negative patients had lower response rate, PFS and OS when treated with PD-1/PD-L1 blockade than Class II positive patients. Class I expression and T cell density were not significantly predictive [65].

Abelin *et al.* [66] show that Class II-dependent antigen presentation depends mainly on autophagy i.e. phagocytosis of apoptotic tumour cells or absorption of

secreted tumour proteins, by antigen presenting cells. These peptides are then processed and presented on MHC molecules. CD4+ T cells appear to play mainly supportive roles in the TME, although direct tumour cell killing is possible. The mechanism of action is most likely through secretion of cytokines and chemokines that drive the trafficking and activation of other immune cells. This has the advantage of bypassing common tumour immune escape mechanisms, such as MHC Class I loss [67].

Prediction of MHC Class II epitopes is particularly challenging due to the high degree of polymorphism at this locus and length and variability of potential binding peptides compared with MHC Class I epitopes [68]. In addition, Class II epitopes appear to be rarely expressed directly on the surface of tumour cells, and are more typically expressed on antigen presenting cells infiltrating along with T lymphocytes [55]. In a mouse study, the Class II epitopes determined did not appear to be immunogenic [69].

Thus far, the finding of immunogenic neoantigens has rarely translated into tumour remission or clinical benefit. In both mouse and human studies, despite identified neoantigens eliciting strong T cell responses, this does not lead to complete or significant tumour rejection in the majority [55, 69]. One hypothesis is that MHC loss or downregulation in tumour cells reduces the likelihood of presentation of neoantigens to immune cells for recognition and killing. The other is that polyclonal tumours have different populations of neoantigens, and these undergo significant selection pressures that favour those tumour antigens that have escaped processing and presentation on the surface of MHC molecules to T cells.

### *1.3.3. Colorectal cancer sub-classification*

Although colon and rectal cancer are treated differently, data from The Cancer Genome Atlas (TCGA) show that they are genomically indistinguishable [8]. CRC is instead better sub-classified genomically by microsatellite status and consensus molecular subtypes.

#### *1.3.3.1. Microsatellite status*

Approximately 15% of patients with CRC have tumours that demonstrate microsatellite instability (MSI) secondary to mutations in DNA mismatch repair (dMMR) genes. MSI-high tumours are characterised by a high mutational burden and the generation of large numbers of neo-antigens, which are believed to trigger powerful anti-cancer host immune responses [70-72]. In contrast, the 85% of CRC that develops due to chromosomal instability, termed microsatellite stable (MSS) [73], usually has a much lower mutational burden and lower numbers of neoantigens.

Three variants of MSI-high CRC have been demonstrated [74, 75]. Hereditary non-polyposis colon cancer or Lynch syndrome is found in 3% of CRC. It is caused by an inactivating germline mutation of one or more of the MMR genes (*MLH1, MSH2, MSH6* and *PMS2*), with a second hit from a sporadic mutation, loss of heterozygosity or epigenetic silencing of a second MMR gene [75]. These patients have a 50-70% lifetime risk of CRC, as well as significant lifetime risks of endometrial cancer (in women), other intestinal and urothelial cancers [76]. More commonly, MSI-high tumours have no underlying germline mutations, and

arise as a consequence of epigenetic silencing of the MMR gene *MLH1* by hypermethylation of its promoter region [77]. Sporadic MSI-high CRC is frequently associated with the v-raf murine sarcoma viral oncogene homolog B1 *(BRAF)*-V600E mutation, through its association with the CpG island methylator phenotype.

BRAF is a downstream molecule in the Rat sarcoma protein-mitogen associated protein kinase (Ras-MAPK) signalling pathway, which is critical for cell survival and proliferation [78]. *BRAF* mutations are present in both sporadic MSI-high and MSS CRC but mostly absent in Lynch syndrome, and thus the presence of a *BRAF* mutation, in conjunction with *MLH1* methylation analysis, reliably distinguishes between sporadic MSI-high CRC and Lynch syndrome [79]. A third variant, Lynch-like syndrome is less well characterised. Lynch-like CRC tumours have no germline MMR gene mutations or hypermethylation of the *MLH1* promoter [80]. However, germline mutations in the DNA polymerase epsilon (POLE) and DNA polymerase delta (POLD1) genes also drive carcinogenesis in CRC [81] and have been linked to the Lynch-like syndrome [82]. POLE and POLD1 mutated tumours are associated with high TMB and also independently predict responses to immunotherapy [83, 84].

MSI-high CRC tumours have distinctive clinico-pathological features including an increased incidence in female patients, more proximal colonic location, high lymphocyte infiltration levels, lower incidence of metastasis, with better clinical prognosis at Stage 1 to 3 [71, 85]. A nationwide study of 6,692 patients by the Danish Colorectal Cancer Group revealed a reduced risk of synchronous metastases, specifically liver metastases, in patients with dMMR CRC (8.0% vs

15.8%, Odds Ratio = 0.54) [86]. There was also noted an inverse association between dMMR status and lymph node metastasis and venous invasion. The situation is reversed in metastatic (Stage 4) disease, where MSI appears to confer a worse prognosis [87].

MMR loss is associated with the rapid accumulation of mutations. MSI tumours typically have an order of magnitude more non-synonymous somatic mutations, compared with MSS tumours (with approximately 50-100 mutations per megabase for MSI CRC compared with 1-10 mutations per megabase for MSS CRC) [29, 88]. In addition to base substitutions, large numbers of insertions and deletions occur [79]. These may lead to frameshifts, which if occurring in tumour suppressor genes, can drive tumorigenesis. High mutation rates generate large numbers of neoantigens, which are not recognised as self and thus are strongly immunogenic. Neoantigens contribute to a better prognosis in MSI CRC due to the increased infiltration of effector cells (primarily effector T lymphocytes [89]) into the tumour environment [71, 85].

Patients with dMMR metastatic CRC have been shown to have significant clinical responses to immunotherapy with anti-programmed cell death 1 (PD-1)/anti-programmed cell death ligand 1 (PD-L1) treatment in Phase II trials [23, 43], in stark contrast to those in the MSS CRC subgroup where there is no objective response to immunotherapy [43]. Yarchoan *et al.* demonstrated in a study across a range of human cancer subtypes a strong correlation between tumour somatic mutation frequency (and therefore neoantigen burden) and the response to immunotherapy [29].

However, MSI-status and neoantigen burden do not sufficiently explain the variability in the colorectal tumour environment. The ORR for patients with refractory MSI-high CRC treated with immunotherapy (ICB) across studies ranges from about 30% to over 50%, compared with 0% to 5% for MSS CRC [90]. This illustrates that within the MSI-high subgroup about half of the patients do not benefit from ICB. This may be related to patient selection criteria, including variations in how microsatellite status is determined. These are currently undergoing standardisation, and involve one of either a polymerase chain reaction (PCR)–based assay of instability, or an immunohistochemical test to detect expression of the mismatch repair protein [91] . Furthermore, the majority are early phase trials, with patients with refractory heavily pre-treated disease. However, it is certain that other genomic and somatic factors are implicated in this resistance.

Immunogenic data show that approximately 20% of patients in the MSS CRC subgroup develop an immune signature similar to MSI-high CRC, despite low mutational burden [92]. This signature, termed the co-ordinate immune response cluster (CIRC) is strongly Th1 driven, with a high preponderance of major histocompatibility complex (MHC) Class II genes. There is evidence that activating mutations in Ras-MAPK pathway are associated with lower expression of this immune gene cluster and immune pathway downregulation [93-95]. Finally, lymphocytic infiltration, particularly of effector and memory T cells into the tumour, which is a key indicator of prognosis in CRC [89, 96] appears to be independent of MSI status [97].

### *1.3.3.2. Consensus molecular subtypes*

CRC can alternatively be divided into four consensus molecular subtypes, each with distinguishing pathological features [98]. The MSI-high group represents CMS1, showing hypermutation and strong immune activation. CMS2 ("canonical") shows chromosomal instability with marked *Wnt* and *myc* signalling; CMS3 ("epithelial") shows metabolic dysregulation; and CMS4 ("mesenchymal") shows prominent transforming growth factor β activation, stromal invasion, and angiogenesis. This subtype is also characterised by strong immune cell infiltration, most likely by mechanisms independent of neoantigen presentation [92].

In a study using a T helper-1 centric immune metagene as a marker of the immune contexture, 20% of patients in the MSS CRC subgroup had an immune signature very similar to MSI-high CRC, despite low numbers of mutations and fewer neoantigens [92]. This group segregated to the CMS4 subtype. The *KRAS* mutation, especially in the CMS2 and 3 subtypes, is associated with a downregulation of immune pathways and reduced immune cell infiltration [93]. *KRAS* mutation, apart from predicting non-response to anti-epidermal growth factor receptor (EGFR) chemotherapy, is independently associated with a worse prognosis in CRC [99].

Disappointingly, a recent trial of the use of bintrafusp alfa, a dual anti-PD-L1 antibody/TGFβ trap, in patients with metastatic CMS4 CRC was discontinued after the first stage due to futility. Of 13 patients, 2 showed stable disease and 11 had disease progression during treatment. Median PFS and OS were only 1.6 months and 5.0 months respectively [100]. However, the researchers noted that

in paired samples, treatment with bintrafusp alfa led to an increase in IFNγ expression signatures in non-irradiated metastatic tissue, which they state could provide a signal for refining potential therapeutic strategies. Further results are awaited.

### 1.4. The colorectal tumour immune microenvironment

A variety of mechanisms leads to immunosuppression of the tumour environment in CRC. Recruitment of immunoregulatory cells [101], upregulation of inhibitory molecules (including myeloid-derived suppressor cells (MDSCs), regulatory T (Treg) cells and type 2 macrophages amongst other cancer-associated cell-types [102-104]) and down-regulation of antigen presentation represent methods of immune evasion [105]. The alteration of metabolic pathways to favour glycolysis, even in the presence of sufficient oxygen has been well-studied and is termed the “Warburg effect” [106]. This, along with the upregulation of anabolic pathways which favour rapid tumour cell survival and proliferation, often leads to the generation of an environment that is hostile to T cells due to increased acidity, low oxygen levels, competition for nutrients and the generation of waste substrates [105, 107]. Immune infiltration in MSI-high tumours is often accompanied by the upregulation of immune checkpoint ligands [104]. T cell exhaustion, defined as the presence of T cells with decreased cytokine expression and effector function in the tumour environment, also occurs [108, 109].

The finding of selective upregulation of immune checkpoints (including PD-1, PD-L1, CTLA-4, LAG-3, Tim-3, IDO and others) in MSI-high tumours suggests that a

counterbalancing inhibitory process occurs in the presence of high levels of tumour infiltrating lymphocytes (TIL) and a highly active Th1 environment [110]. This may explain why MSI-high tumours are not naturally eliminated despite high immune activation, and why checkpoint blockade is effective in these tumours (Figure 1.3) [45].

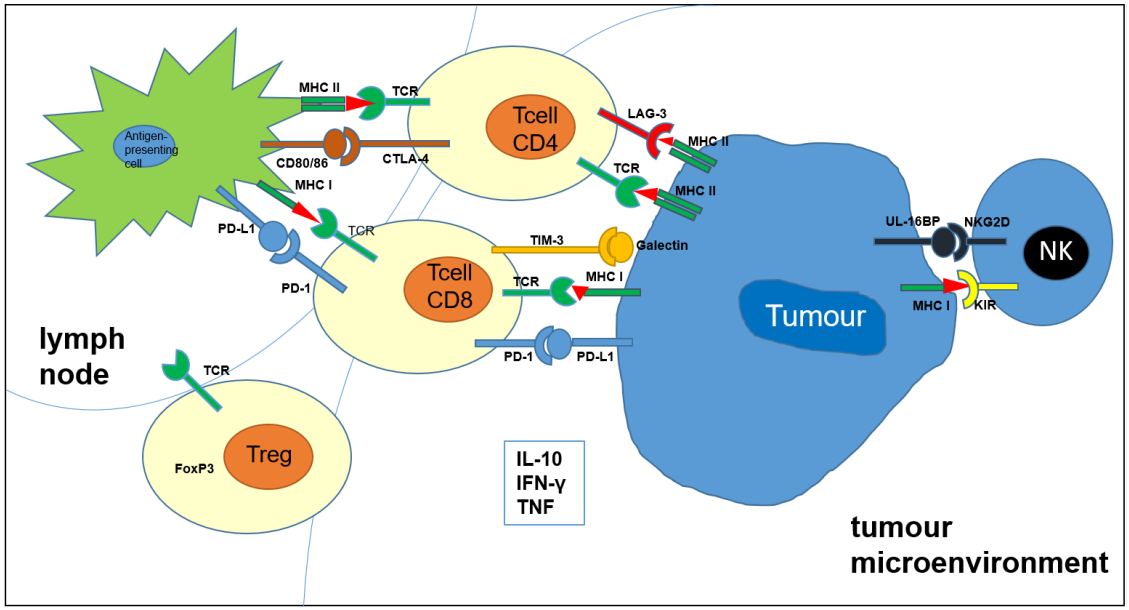



**Figure 1.3.** Immune regulatory pathways in the tumour microenvironment [45]. MHC Class I and II present on tumour cells and antigen presenting cells present antigens (red triangles) to T cells. Interactions with immune checkpoints and regulatory T cells modulate this process. CD – cluster of differentiation. CTLA-4 – cytotoxic T lymphocyte-associated protein 4. FoxP3 – forkhead box P3. IFN-γ – interferon γ. IL-10 – interleukin 10. KIR – killer immunoglobulin-like receptor. LAG-3 – lymphocyte -activation gene 3. MHC I – major histocompatibility complex Class I. MHC II – major histocompatibility complex Class II. NK – natural killer cell. NKG2D – natural killer G2D receptor. PD-1 – programmed cell death 1 protein. PD-L1 – programmed cell death 1 ligand. TCR – T cell receptor complex. TIM-3 – T cell immunoglobulin mucin 3. TNF – tumour necrosis factor. Treg – regulatory T cell. UL16BP – UL16 binding proteins.

In addition, other pathways may drive immunogenicity in MSI-high CRC tumours. Constitutive activation of the viral response cyclic GMP-AMP/synthase-stimulator of interferon genes (cGAS-STING) pathway with associated T cell infiltration in tumours is noted in DNA damage response-deficient breast cancers [111]. cGAS is activated by DNA damage and localises to micronuclei which form in the context of the genomic instability which occurs during tumorigenesis or autoimmunity [112]. This triggers a pro-inflammatory response, which is notably absent in STING-knockdown cell lines. STING knockdown mice also do not demonstrate the abscopal effect, which is tumour regression outside the irradiated field, usually observed following combined ionising radiation and immune checkpoint blockade therapy [113]. Deficiency in the DNA damage repair protein, MLH1, which is often mutated in dMMR CRC, has also been shown to be associated with deficient DNA double strand break repair and increased micronuclei formation [114], which may also trigger the cGAS-activated inflammatory response.

Emerging immunogenomic data show that the strength of the microenvironmental immune response, even in MSS CRC, is highly variable. Approximately 20% of MSS patients have similar levels of immune activation to MSI CRC and yet like the rest of the MSS cancer population have a low non-synonymous mutational burden and low neoantigen levels [92]. Thus, there are other factors that influence this immune response, which warrant further investigation.

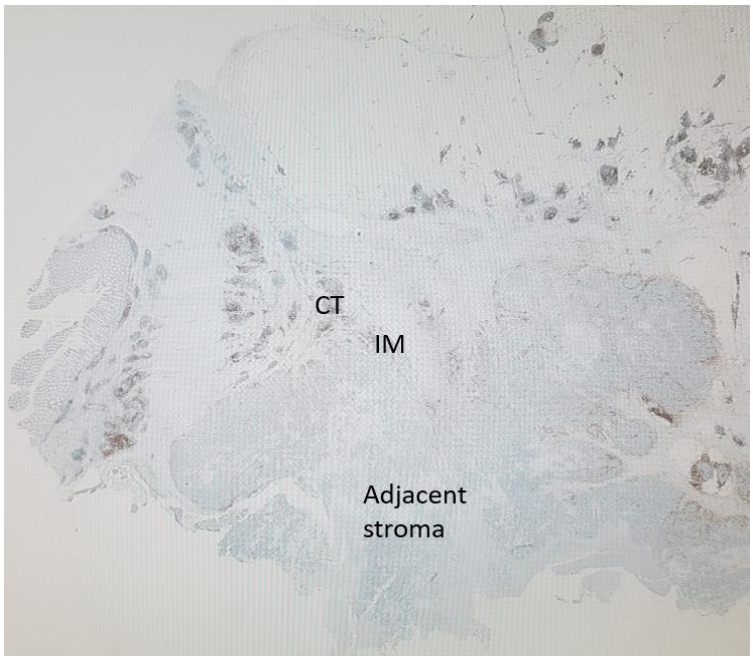
Hitherto unexplored germline and tumour-specific factors are likely to contribute significantly to these differences and form the basis of exploration in this thesis.

### *1.4.1. Quantification of the colorectal tumour immune microenvironment*

Galon *et al*. developed the Immunoscore®, an immunocytochemical score of the CRC immune microenvironment [115, 116]. This was based on the finding that the infiltration of cytotoxic (CD8+) and memory (CD45RO+) T cells was associated with increased expression of Th1 and CD8-cytotoxicity-related genes. The densities of CD45RO+ and CD8+ cells in the centre of the tumour (CT) and invasive margin (IM) could be used to stratify the patients into distinct populations with significantly different clinical outcomes at all disease stages [117]. In multivariate analysis, after adjusting for tumour invasion, differentiation, lymph node invasion and other tumour molecular biomarkers including MSI and BRAF mutation status, T cell infiltration ($CD3_{CT}/CD3_{IM}$) remained an independent prognostic factor in disease-free survival (DFS) analysis, and only $CD3_{CT}/CD3_{IM}$ density was an independent parameter associated with overall survival. This has been independently validated by the Society for Immunotherapy of Cancer (SITC)-supported worldwide consortium study [118] [119]. The Immunoscore is currently the only independently validated marker of the immune contexture in CRC, with key prognostic information.

The Immunoscore is calculated by counting two lymphocyte populations (CD3/CD45RO and either CD3/CD8 or CD8/CD45RO) in the centre of the tumour (CT) and the invasive margin (IM). This yields a score of 0 to 4, where 0 designates low densities of both populations in both areas of the cancer, and 4 designates high densities in both areas (Figure 1.4, Figure 1.5). I used the Immunoscore as the principal read-out of the immune contexture in this analysis for four main reasons.

a




b

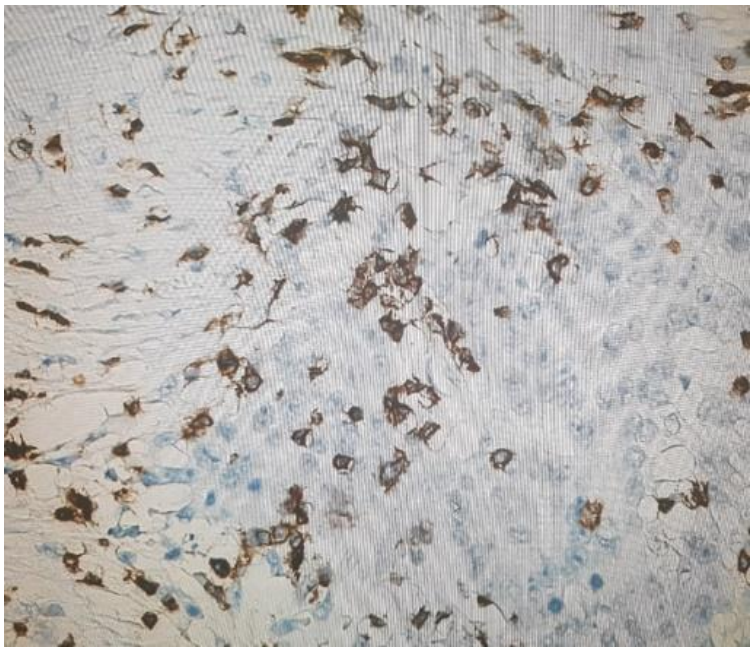


**Figure 1.4.** The Immunoscore®. Staining is performed for two markers (CD3 and CD8) in two regions (the centre of the tumour (CT) and invasive margin (IM)). Digital quantification of cell density is performed (cells/$mm^3$) and a score is assigned from 0 to 4 to delineate immune reactivity. (a). An example of CD3 stain on a formalin-fixed tissue slide. (b). A zoomed image of the same slide.

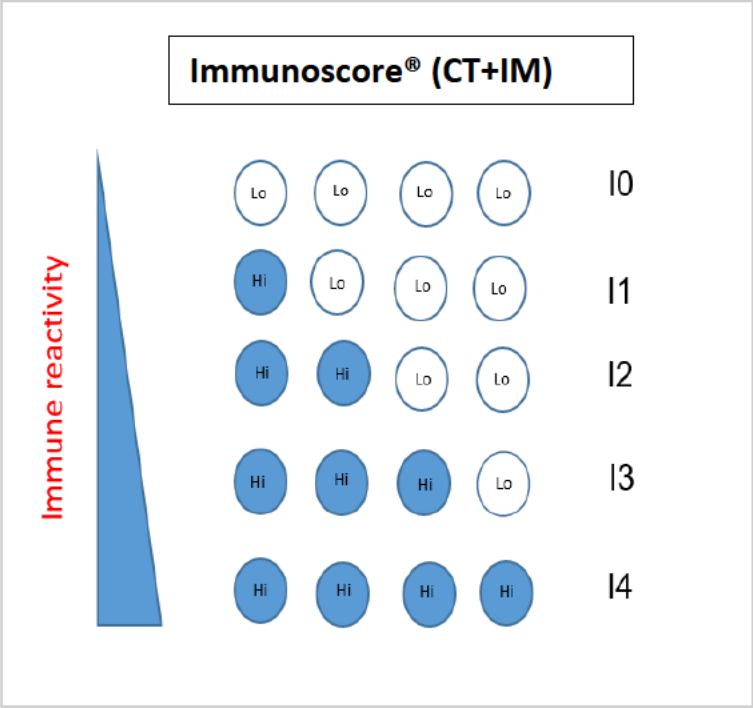



**Figure 1.5.** Numerical representation of the Immunoscore (adapted from [116]). CT = centre of the tumour. Hi = High. I = Immunoscore. IM = invasive margin. Lo = Low.

#### *1.4.1.1. The Immunoscore has optimal biomarker characteristics.*

The Immunoscore utilises a routine test performed by a pathologist using bright field microscopy. It requires only four formalin-fixed paraffin-embedded (FFPE) slides in total, making it feasible. It is simple and reproducible in that it uses automated digital immunohistochemistry (IHC) which removes inter-observer variation. It is quickly performed, robust and standardised.

#### *1.4.1.2. The Immunoscore yields a quantitative score of the immune contexture.*

Biological correlations are optimally performed with characteristics that can readily define distinct populations representing extremes of the biomarker in question. Approximately 21% of primary CRC are Im0/1 representing very low

levels of microenvironmental immunity and 30% are Im4, with a strong anti-cancer response [120]. The outcomes of these two groups are markedly different, with 31.6% and 52.8% 5-year DFS for Im0 and Im1 respectively and 85.4% for Im4 (HR = 0.19).

#### *1.4.1.3. The Immunoscore is clinically and biologically relevant.*

The Immunoscore is performed on either colorectal tumour biopsies or surgically resected specimens. Low Immunoscores are significantly over-represented in patients presenting with or subsequently developing metastatic disease, thus showing the crucial importance of primary immunobiology for the metastatic process [121]. In patients with metastatic disease, whilst Im4 patients have a 65% 5-year OS, all Im0 patients have died by 40 months. In a comprehensive multivariate analysis including stage, grade, venous emboli, lymphatic and perineural invasion of cancer (VELIPI), mucinous histology, perforation, obstruction and microsatellite status, only the Immunoscore was independently associated with all three endpoints of disease-specific survival (DSS), DFS and OS [97].

#### *1.4.1.4. The Immunoscore is fully validated.*

The SITC supported a large worldwide consortium study, in a multi-centre, multi-national context (14 centres from 13 countries). Over 3000 patients were recruited and 2681 samples were passed to final analysis, which confirmed that the Immunoscore is a reliable and reproducible prognostic biomarker in CRC [118]. This was confirmed by data from 559 patients with stage 3 CRC in the Phase III leucovorin, fluorouracil, and oxaliplatin (FOLFOX) +/- cetuximab trial [122, 123]. Low versus high Immunoscore was associated with lower DFS (HR =

1.69, 95% CI = 1.22 to 2.33, p = 0.001). In T1-T3 and N1 disease, the Immunoscore was the only statistically significant variable associated with DFS.

#### *1.4.1.5. The Immunoscore has predictive value.*

The Phase III International Duration Evaluation of Adjuvant Therapy (IDEA) France cohort study investigated 3 months versus 6 months of oxaliplatin-based adjuvant chemotherapy in Stage 3 CRC. It showed overall superiority of the 6-month compared with the 3-month regime, particularly in 'higher risk' (T4 and N2) groups.

Analysis of validated Immunoscore results in the modified intention to treat population showed that the Immunoscore was predictive for treatment duration for those with Intermediate and High Immunoscores (IS Int + High). [124]. For patients treated with the modified FOLFOX regime, 6 months of treatment was associated with significantly increased DFS compared with 3 months of treatment (HR = 0.53, 95% CI = 0.37 – 0.75, p = 0.0004). Of interest, this effect was present in both clinically low and high-risk Stage 3 disease. Conversely, those with low Immunoscore (representing 46.4% of the cohort) did not have significant benefit from the 6-month compared with the 3-month duration of therapy. The authors hypothesise that oxaliplatin-based agents drive immunogenic cell death, which is reduced or absent in IS Low tumours. 5-fluorouracil decreases MDSCs and increases cytotoxic T cell function. The potential benefits of this immunological boost are lost in the IS Low environment.

The main drawbacks to routine clinical use of the Immunoscore include cost and availability, as HalioDx®, based in Marseille, France, currently owns all rights to

perform the test in its laboratories, at a cost [125]. It would be beneficial to assess the predictive value of the Immunoscore in determining the effect of neo-/adjuvant immunotherapy, independent of MMR status.

### *1.4.2. Immune checkpoint expression in the tumour environment*

The role of expression of immune checkpoint molecules (such as PD-L1 as markers of probable response to ICB remains controversial. Strong PD-L1 expression (as determined by IHC) is noted in 37% of MSS (pMMR) and 29% of MSI-high (dMMR) CRC and correlates with better clinical outcomes [126]. Strong PD-L1 expression in pMMR tumours was associated with high CD8+ T cell infiltration into tumours. Early phase trials of nivolumab (an anti-PD-1 monoclonal antibodies) in solid tumours including melanoma, renal, prostate, lung cancer and CRC suggest that PD-L1 expression may serve as a marker for objective responses to ICB [41, 127], as patients with PD-L1 negative tumours showed no responses. However, the association is weak, and although PD-L1 expression appears to correlate with the infiltration of TILs, the link to clinical responses is borderline [128]. In these trials, microsatellite status was not universally reported for the CRC patients, who showed universally poor responses. In addition, there is a lack of standardised measures for PD-L1 expression, with different antibodies, diagnostic tools, scoring systems and cut-off expression values used in the studies.

In lung cancer, the association of PD-L1 expression with response to ICB appears more robust. In a meta-analysis of randomised controlled trials of anti-PD-1/anti-PD-L1 immunotherapy in non-small cell lung cancer (NSCLC), subgroup analyses showed greater OS with immunotherapy compared with chemotherapy

at cut-off levels of PD-L1 expression of ≥1%, ≥5%, ≥10% and ≥ 50%. ORR rates were greater with PD-L1 expression ≥50% compared with the <1% and 1-49% groups [28].

PD-L1 expression is different in tumour and immune cells. Valentini *et al*. [129], in a study of 63 CRC specimens from 61 patients, found differential expression of PD-L1 in tumour cells (NCs) and tumour-infiltrating immune cells (IICs). They obtained three cancer groups: group A (NCs-/ IICs-); group B (NCs-/ IICs+) and group C (NCs+/IICs+). Group A tumours were poorly immunogenic. Group B had more immunogenic CRCs but with upregulation of PD-L1 only on IICs, and group C was characterised by a large tumor neoantigen burden resulting in both lymphocytic infiltration and PD-L1 upregulation. Tumours with MSI-high status were more likely to be found in group C than either group A or group B.

A recent meta-analysis of the association of PD-L1 expression in CRC showed that high PD-L1 expression in tumour cells is associated with worse clinical outcomes, in particular reduced OS and DFS [130]. This study also found that PD-L1 expression was independent of tumour stage or microsatellite status, but high PD-LI expression is associated with right-sided, more poorly differentiated tumours. Similar findings have been noted with another immune checkpoint, Tim-3, the upregulation of which is associated with increased regional metastases and poorer prognosis in CRC [131, 132]. However, currently, no conclusions can be drawn on the use of expression of these immune checkpoints as biomarkers for determining the likely efficacy of ICB. Microsatellite status remains the only established biomarker for stratification of patients for immunotherapy in CRC.

## 1.5. Inherited differences in immune gene expression

Differences in gene expression levels are believed to be responsible for most of the phenotypic variation observed among individuals in natural populations [133]. Next generation sequencing (NGS) techniques have aided genome wide association studies (GWAS), which reveal that many of genetic variants associated with phenotypic variation (e.g. disease risk, protein expression) are found in non-coding or intronic parts of the genome and therefore are likely to be involved in gene regulation [134-136]. In addition, advances in RNA sequencing and expression microarrays have made it possible to quantify gene expression levels accurately [133, 137-140].

### *1.5.1. Expression quantitative trait loci*

Expression quantitative trait loci (eQTLs) are single nucleotide polymorphisms (SNPs) usually found in non-coding regions of the genome, which influence gene expression. These are *cis*-, (in close proximity to the genes they affect) or *trans*- (at a distance away from the genes they affect, or on separate chromosomes) [134, 141, 142] (Figure 1.6). *Trans*-eQTLs may exert their effects due to conformational changes, leading to differential transcription factor binding across different chromosomes that may be genomically distant but in close proximity due to the spatial organisation of DNA [143, 144]. *Cis*- variants are conventionally accepted as those within 1 megabase (Mb) of the transcription start site (TSS) on either side of the gene. Those more than 5 Mb on either side of the TSS or on another chromosome are considered *trans*- acting [145].

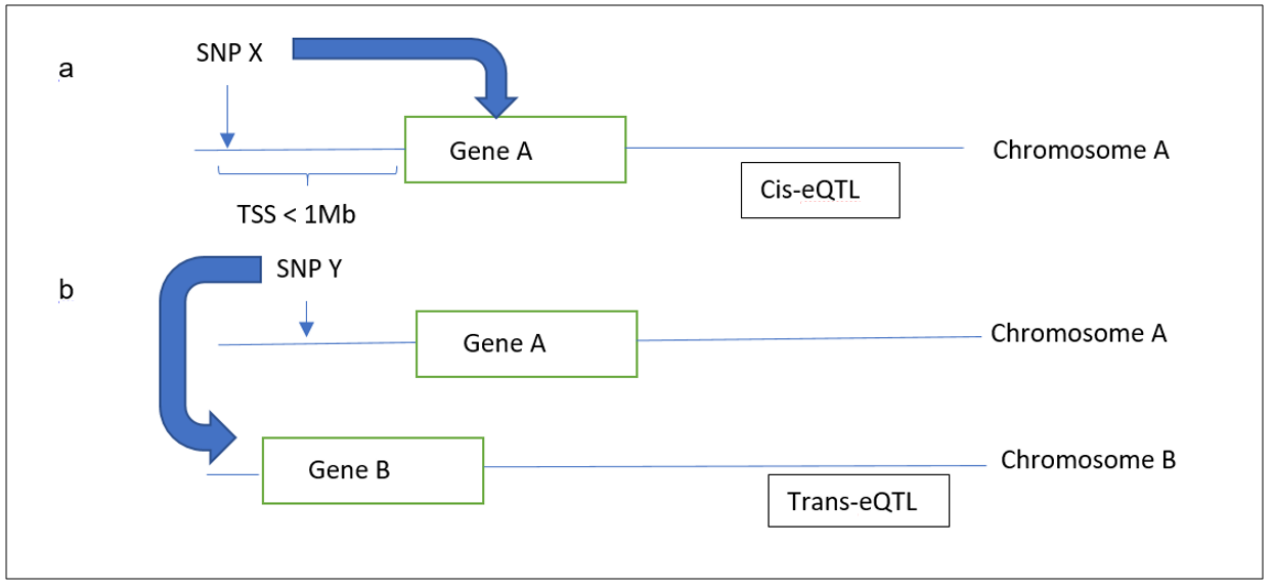



**Figure 1.6.** Schematic representation of cis- and trans-eQTL effects on targeted genes. (a) illustrates a cis-eQTL SNP effect on Gene A and (b) illustrates a trans-eQTL effect on Gene B on a different chromosome. eQTL – expression quantitative trait locus. SNP – single nucleotide polymorphism. TSS – transcription start site [45]

### *1.5.2. eQTL data repositories*

eQTL studies were first carried out in model organisms where the whole genome could be rapidly sequenced [146], and then expanded to human studies [141, 147]. Gene expression levels are typically measured in hundreds of people, and analysis involves an association test between markers of genetic variation [145]. The mainstay of eQTL analysis is robust statistical analysis, aiming for sufficient statistical power to detect the relevant variants while accounting for the multiple testing burden [148]. Correction for multiple testing is performed by one of several approaches [148-150]. The Bonferroni correction is widely used. However, it is conservative and does not account for linkage disequilibrium (LD) in the genome. LD refers to the non-random association of alleles at two or more loci in a population, which leads to haplotypes occurring at frequencies that are more or less than expected. Haplotypes are groups of SNPs in close proximity on a chromosome that tend to be inherited together, such as the HLA alleles in the MHC on chromosome 6. The Bonferroni method penalises those regions of the genome with strongly linked variants and reduces the statistical power to detect variants [148, 149].

The permutation test [149] accounts for LD but is cumbersome, computationally expensive and gives truncated *p* values to a level of significance determined by the number of permutations [148, 150]. The false discovery rate (FDR) method, devised by Benjamini and Hochberg [151, 152], is designed to control the expected proportion of false discoveries, that is, the expected proportion of rejected null hypotheses that are false. It provides greater statistical power to detect differences, albeit at a cost of increased Type 1 errors. The FDR method

has become a key method in eQTL studies, usually set at 0.05 [153]. SNPs with low minor allele frequency (MAF) (usually <5%) are filtered out from analysis [141, 154].

eQTL data repositories such as the Multiple Tissue Human Expression (MuTHER) study [155] and the Genotype-Tissue Expression Project (GTEx) [156, 157] help to provide insights into how eQTLs determine the expression of phenotypes of interest, including complex diseases and cancer (Table 1.3). Most human eQTL studies are performed using peripheral blood-derived cell lines, likely due to ease of sampling access. Data from studies assessing more than one tissue type show that regulation of gene expression is partially cell type-independent, with variable degrees of eQTL sharing across tissues [147, 155].

The GTEx pilot [147] reported *cis*-eQTL yields from 54 distinct body sites from 237 post-mortem donors. More than 50% of eQTLs were shared across all nine tissues studied. Only 7 – 21% of eQTLs were tissue-specific, with the effect strongest when closer to the TSS. The MuTHER study [154] compared *cis*-eQTLs in tissues from lymphoblastoid cell lines (LCLs), skin and adipose tissue in female adult twins from the United Kingdom (UK) Twins registry. 56 – 83% of *cis* eQTLs were calculated to be shared across the three tissues studied. However, the tissue-independent effect was stronger closer to the TSS in this analysis. In addition, several *trans*-eQTLs were discovered at a low false discovery rate (FDR). They were found to be mainly tissue-type dependent, with smaller effect sizes and often associated with multiple transcripts, suggesting their role as multiple-gene regulators.

**Table 1.3**. Human eQTL repositories

| Project name | Data repository | eQTL | Tissue subtypes | Sample size |
|---|---|---|---|---|
| MuTHER | http://www.muther.ac.uk/Data.html | *cis* | LCL, skin, adipose | 856 |
| GTEx | https://www.gtexportal.org/home/ | *cis* | multiple | 237 |
| Childhood asthma studies [90, 91] | http://csg.sph.umich.edu/liang/imputation/ | *cis* and *trans* | EBVL | 2642 |
| International HapMap Project [68] | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6536 | *cis* and *trans* | LCL | 270 |
| Gilad/ Pritchard Group | http://eqtl.uchicago.edu/Home.html | *cis* and *trans* | LCL, liver, brain | n/a |
| Pickrell Lab [92] | http://gwas-browser.nygenome.org | *cis* and *trans* | multiple | Combined sources |
| Geuvadis Project | https://www.ebi.ac.uk/Tools/geuvadis-das/ | *cis* | LCL | 465 |
| Blood eQTL [93] | https://genenetwork.nl/bloodeqtlbrowser/ | *cis* and *trans* | Peripheral blood | 5311 |

EBVL – Epstein-Barr virus-transformed cell lines. eQTL – expression quantitative trait locus. GTEx – Genotype-Tissue Expression Project. LCL – lymphoblastoid cell lines. MuTHER – Multiple Tissue Human Expression Resource Project. n/a – not applicable. Adapted from [45].

### *1.5.3. eQTLs and cancer genomics*

The contribution of eQTL SNPs in cancer susceptibility has been studied in many cancer types including breast [158], prostate [159] and ovarian cancer [160]. Although the study of germline determinants of gene expression in cancer is complex, due to the accumulation of mutations which increase the complexity of transcript regulation, this can be mitigated by using matched tumour and normal samples [158].

Vogelsang *et al.* [161] utilised data from the MuTHER project to identify immune gene eQTL SNPs and correlate these with outcomes in cutaneous melanoma, a strongly immunogenic cancer. Of the 382 immunomodulatory genes selected by interrogating the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases, SNP genotyping of the 50 top most significant *cis* eQTLs in the MuTHER LCL database was performed and the top 40 SNPs for which genotype-expression associations were obtained were correlated with outcome (DFS and OS).

Of interest, their gene list includes 22 of 28 genes included in the CIRC signature [92]. Sixteen of these had statistically significant sequence-based expression variation, including STAT1, a key Th1 module, IFN$\gamma$ and some MHC Class II genes [162]. Two SNPs identified were highly correlated with OS – one affecting IL-19 expression and the other BATF3 expression. Similarly, a study in breast cancer patients [163] revealed that the expression of MHC Class I and II genes was associated with SNPs in 100 genes. Comparison with a matched healthy cohort revealed specific associations with genes associated with immune system processes.

## 1.6. Neoantigens

Tumour-specific antigens (neoantigens) arise as a result of somatic mutations during tumour evolution. These may be driver mutations which cause the transformations required for tumorigenesis and tumour propagation, or passenger mutations which are passively acquired along with the driver mutations as by-products of the genomic instability that occurs during tumorigenesis [7]. Neoantigens generated from this process are less likely to undergo immunological tolerance as they are foreign to the individual. Mutations may be clonal (expressed in all tumour cells) or subclonal (expressed in a proportion of tumour cells) leading to the expansion of cell populations with different genomic and therefore phenotypic signatures [56, 164]. Tumour-associated antigens on the other hand, which are aberrantly expressed normal proteins, are less likely to be antigenic unless mechanisms of immunological tolerance are circumvented. Germline proteins which are usually tissue-restricted, may become antigenic when expressed in cancer cells [7].

Work on cutaneous melanoma shows that the most potent T cell responses are against neoantigens [165]. As the pattern of mutations is highly variable, and the cancer genome is unique to each individual, identification of these neoantigens was initially challenging. With the development of NGS techniques and bioinformatics strategies for *in* silico prediction, it is now possible to rapidly identify and filter neoantigens [40, 166, 167]. Whole exome sequencing (WES) and/or whole genome sequencing (WGS) of tumour samples allows identification of somatic mutations. Mutation calling is done by aligning the sequencing reads against the reference genome to identify variants, which are then compared

against data from matched normal tissue DNA to identify tumour-unique mutations. These are then modelled using a protein prediction algorithm such as the Variant Effect Predictor [168] and fed into an MHC-binding predictor to model the MHC binding capacity (Class I and II binding) [68, 169, 170]. On the other hand, structural variants (such as gene fusions) are more difficult to identify from WES data unless RNA sequencing data is available [31].

### *1.6.1. Neoantigen clonality*

Intratumoral heterogeneity (ITH) is of key significance in tumour immune escape mechanisms. Subclonal neoantigens are expressed in only a proportion of tumour cells. Consequently, non-expressing cells can avoid surveillance by antigen specific T cells and these sub-clones can multiply and metastasise. McGranahan *et al.*, in a series of lung cancers for which multi-region sequencing was available, showed that an average of 44% of neoantigens were found heterogeneously in a subset of regions [164]. The authors analysed clonality from both single and multi-region sample WES data from lung cancer samples in The Cancer Genome Atlas. The combination of neoantigen ITH and neoantigen burden was more predictive of outcome than either measure alone. High clonal neoantigen burden was characterised by an inflamed microenvironment (assessed by RNA expression).

This approach was also applied to WES data from a recent study demonstrating the predictive power of non-synonymous TMB for response to Pembrolizumab in lung cancer. The efficacy of PD-1 blockade was found to be dependent on clonal architecture. Tumours with similar numbers of neoantigens responded significantly more favourably if those neoantigens were clonal than if they were

subclonal [171]. These findings are supported by a study in melanoma in which ITH and TMB were uncoupled using a mouse model [172]. Mutations were introduced in a melanoma cell line using ultraviolet B irradiation. Mice inoculated with the heterogeneous cancer cell population showed more aggressive tumour growth compared with those inoculated with single cell clones. Rejection of the single cell clone was associated with higher infiltration of effector T cells into the immune environment. This effect was independent of mutational burden, and absent in immunocompromised mice.

Neoantigens are derived from proteins translated from non-synonymous mutations. These proteins undergo several steps prior to presentation at the cell surface on MHC Class I and II molecules, for presentation to T lymphocytes. Predicting binding affinity to MHC requires the determination of patient-specific HLA alleles. Determination of neoantigen clonality requires a series of *in silico* steps, including identification of non-synonymous variants by comparing matched tumour and normal WES or WGS data, determination of HLA haplotypes, peptide processing, MHC binding prediction, and cross-referencing of neoantigens with known epitopes [173, 174] (Figure 1.7). Following this, clonality can be determined from multi- or single-region WES or WGS data using mathematical modelling. Pipelines typically call Class I epitopes only, with high rates of accuracy [173]. Current Class II typing algorithms are less reliable, although there is significant development and improvement in progress [55, 175].

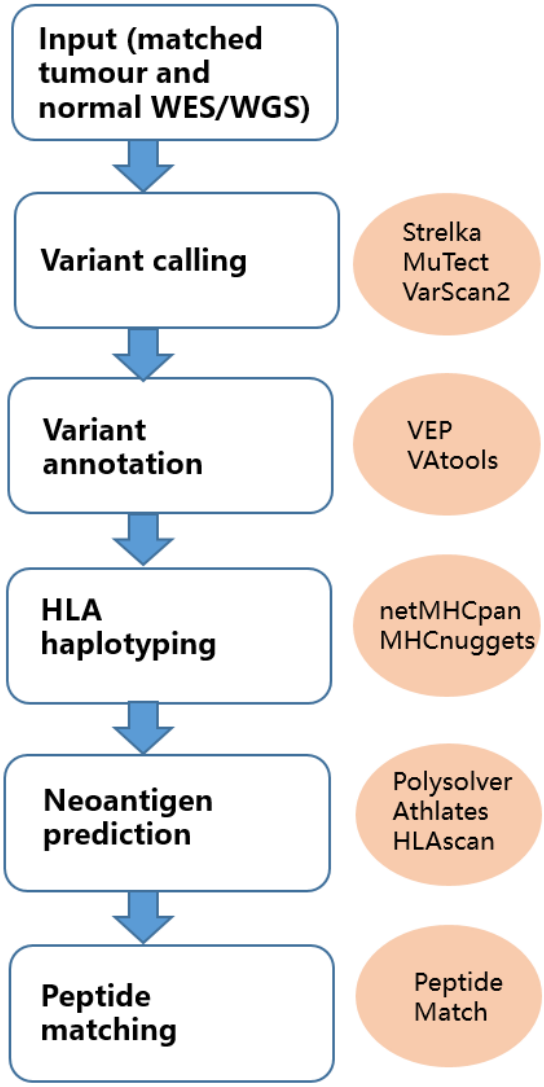



**Figure 1.7.** Bioinformatics pipelines for neoantigen prediction and examples of some tools devised for each stage. Adapted from [174] and [173] (Athlates [176], HLAscan [177], MHCnuggets [175], MuTect [178], netMHCpan [169], PeptideMatch [179], Polysolver [180], Strelka [181], VarScan2 [182], VAtools[183] and VEP [168]).

### *1.6.2. Determining neoantigen clonality*

Methods for determining neoantigen clonality are currently experimental. Ideally, multi-region sequencing data should be available to determine the prevalence of mutations in all sequenced regions of a tumour. However, this is not readily available in many patient data sets and extrapolation and modelling from single region sequencing data is more commonly used. All methods rely on a series of mathematical and genomic assumptions, and their reliability is variable.

For example, in a pan-cancer analysis of ITH using PyClone [184], Morris *et al.* showed marked associations between ITH and poorer survival outcomes in many cancer types, notably head and neck squamous cell carcinoma (HNSCC) and renal clear cell carcinoma [185]. However, there was no association noted in lung adenocarcinoma or squamous cell carcinoma, or bladder urothelial carcinoma. Colorectal carcinoma was also excluded from this analysis.

An analysis of six methods of estimating ITH using TCGA data from breast, urothelial and HNSCC samples showed marked differences in the estimation of ITH, depending on the method used [186]. Methods that employed both estimation of single nucleotide variants (SNVs), copy number alterations (CNAs), and tumour purity such as PyClone and PhyloWGS [187] had similar results in all three cancer types, compared with the mutant allele heterogeneity (MATH) score [188], which does not analyse CNAs, but is a single quantitative measure based on differences in the mutant allele fractions among mutated loci [184]. In the comparison of six methods, the MATH score had the best success rate (100%), which is defined as the fraction of samples for which the method produced an error-free output.

These methods thus far have not been convincingly been shown to be superior to other clinic-pathological markers for determining prognosis [186]. They also are dependent on the accuracy of variant calling, which can be challenging [189]. In addition, most methods require estimation of copy number and tumour purity. Three methods are discussed below and two were used in this thesis.

#### *1.6.2.1 Modified PyClone*

McGranahan *et al.* demonstrated techniques for determining neoantigen clonality on both multi-region and single-region WES [164]. For multi-region WES data, for each patient sample, a mutation was deemed clonal only if it was present in all tumour regions sequenced. For single-region data, the observed mutation copy number was calculated using a combination of the variant allele frequency (VAF), tumour purity and local copy number. The expected mutation copy number was calculated using the VAF and assigning a mutation to one of the possible copy numbers using maximum likelihood. Finally, the mutations were clustered using the PyClone Dirichlet process clustering, which allowed clustering to group clonal and subclonal mutations based on their cancer cell fraction estimates. ITH was set using thresholds of 0.00, 0.01 and 0.05, and in all cases, there was statistically significant increased survival in those patients with samples calculated to have low ITH (that is, those with clonal neoantigens).

#### *1.6.2.2 Neopredpipe*

In the Genomics England 100 000 Genomes Project [190], which provided the bulk of samples and data for this project, neoantigen burden and clonality were assessed using the Neopredpipe pipeline [174]. This calls single nucleotide variants (SNVs), as well as insertions-deletions (indels) and frameshift mutations.

HLA typing was performed for Class I alleles (HLA-A, -B and -C). ANNOVAR [191] was used to annotate variants to identify non-synonymous variants, with the prioritisation of exonic variants. HLA haplotypes were provided separately and netMHCpan 4.0 was used for primary neoantigen identification [169]. For each variant, all peptides of length 9 to 10 amino acids that contained mutated amino acids were evaluated. Strong binders (SB) and weak binders (WB) were reported, using netMHCpan 4.0's criterion of having a binding rank prediction of less than 2.0 (for WB), or less than 0.5 (for SB).

Clonality was determined by obtaining the cancer cell fraction and mutation clustering using DPClust [56] which uses a Dirichlet process based approach to estimate the number of mutation clusters in the data.

Clonality can then be estimated by filtering the DPClust output against the neoantigen data available from Neopredpipe. The most important caveat in this process is that Class II neoantigen data is not available for Neopredpipe, thus leaving out a potential substantial contribution to heterogeneity.

#### *1.6.2.3. The Mutant-Allele Heterogeneity score*

The MATH score [188] is another method for determining ITH, from mutant allele frequency data. For a tumour, the MATH score is a ratio of the width to the centre of the distribution of mutant-allele fractions, among mutated loci. This is calculated as a percentage of the ratio of the median absolute deviation (MAD) to the median of mutant allele fractions at tumour-specific mutated loci.

Mroz *et al*. [188], who described the MATH score, interrogated the relationship between the MATH score and prognosis in 74 HNSCC cases. They showed that

the MATH score was not related to overall mutation rate, and that the MATH score is higher in poorer outcome HNSCC (including those with disruptive tumour protein 53 (TP53) mutations). The authors note that MATH is calculated from single sample reads, and its precision depends on sampling of loci and of mutant versus reference alleles. The precision of the score is also higher in samples with higher mutation rates. They discuss the possibility that higher mutation rates could lead to higher MATH scores, but this was not the case in their data set.

They also showed that copy number differences did not have a significant impact on the MATH score. A linear comparison of the CNA-adjusted MATH score with the raw MATH showed an almost exact correlation between these two values.

The MATH score was used to quantify heterogeneity in a sample of over 2500 cancers from the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) Pan-Cancer Analysis of Whole Genomes (PCAWG) variant data set [192, 193]. In this data set, there was lower mutational heterogeneity among high-impact putative passenger mutations for both coding and non-coding regions. Greenbaum *et al.*[194], in a sample of 21 locally advanced rectal cancers, showed a correlation between higher MATH score and poorer responses to neo-adjuvant chemoradiotherapy (p=0.0039).

In summary, the MATH score provides a simple solution to the problem of determining Class I versus Class II neoantigens, and effectively determines ITH from TMB data.

### 1.7 Metagenomic determinants of the colorectal immune environment

The role of commensal microbiota in gut development, maintaining integrity, metabolism and immunity is critical [195]. There are over 1013 commensals in the human gut. Although the majority (99%) of species are bacterial, with *Bacteroides* and *Firmicutes* predominating [196], there are also viruses, archaea and eukarya. There is significant diversity in microbiota amongst healthy individuals, as well as between healthy people and those with some diseases, such as inflammatory bowel disease [197].

### *1.7.1. Gut microbiota and tumorigenesis*

There is convincing evidence that gut microbiota modulate colorectal tumorigenesis. In animal models, specific microbes associated with colonic inflammation can drive carcinogenesis [198]. *Bacteroides fragilis* rapidly induces colitis and colon tumours in mice heterozygous for the *APC* gene, with marked downregulation of effector T cell responses and upregulation of Treg responses [198] [199] . In humans, gut microbiota differ significantly between patients with CRC and healthy controls [200]. In particular, certain bacterial strains, most notably *Fusobacterium nucleatum*, *Escherichia coli*, *Bacteroides fragilis* and *Salmonella enterica,* are detected in human biopsies in gastrointestinal cancers, and could be considered high risk factors for carcinogenesis [201]. Moreover, there is a large degree of heterogeneity in microbiota composition in CRC patients, with differences between faecal and mucosal samples, and between proximal and distal tumours [200].

### *1.7.2. Gut microbiota and the anti-tumour immune response*

Routy *et al.* demonstrated that abnormal gut microbiome composition could be responsible for non-response to anti-PD-1 immunotherapy in patients with a

range of epithelial cancers, predominantly NSCLC and renal cell carcinoma [202]. They used quantitative metagenomics, with shotgun sequencing obtaining greater than 20 million short DNA reads per sample, which was analysed using a reference catalogue of the human microbiome genome [203]. In this cohort of 100 patients, administration of systemic antibiotic treatment just prior to commencing immunotherapy led to worsened PFS and OS than in a comparable non-treated group. This was postulated to be due to the alteration of the gut microbiome by antibiotic therapy. Responders to immunotherapy had differencing microbe profiles from non-responders, with the abundance of *Akkermansia municiphilia, Enterococcus hirae* and *Alistipes indistinctus* in responders. Furthermore, faecal mucosal transplantation (FMT) from responders into germ-free or antibiotic-treated mouse tumour models led to significant anti-tumour responses, with upregulation of dendritic cell and effector T cell responses. This did not occur with FMT from non-responders [202].

Using 16S ribosomal RNA (rRNA) gene amplicon sequencing, Matson *et al.* showed that stool samples from patients with metastatic melanoma who responded to immunotherapy had an abundance of certain bacterial species, notably *Bifidobacterium longum*, *Collinsella aerofaciens* and *Enterococcus faecium,* while non-responders had an abundance of *Ruminococcus obeum* and *Roseburia intestinalis* [204]. Flow cytometry and cytokine assays from patients showed that those with high abundance of favourable microbes (including *Clostridiales*, *Ruminococcaceae* and *Faecalibacterium*) had higher densities of effector T cells (CD4+ and CD8+) in the systemic circulation, while those with higher frequencies of *Bacteroidiales* species had more Tregs and

MDSCs. A similar germ-free mouse tumour model also demonstrated similar responses to FMT from responders [204].

Finally, exposure to bacterial species enriched in colonic tissues, notably *Fusobacterium nucleatum*, *Bacteroides fragilis* and *Escherichia coli,* stimulates chemokine production which drives T cell tracking into the tumour microenvironment both *in vitro* and in an *in vivo* CRC mouse model [205].

### *1.7.3. Metagenomic sequencing of tumour whole genome data*

Associations between differences in the gut microbiome and responses to immunotherapy in patients with CRC have been relatively underexplored and is an area of significant interest. Tumour WGS data can be used to explore the associations between metagenomic factors and the immune response in CRC, using Kraken2, a taxonomic sequence classifier which is able to assign taxonomic labels to DNA sequences [206]. WGS has been demonstrated to have many advantages over 16S rRNA sequencing, including increased detection of bacterial species, species diversity and increased accuracy of species detection, likely due to the longer read length with WGS compared with amplicon sequencing [207]. This information was available to explore the associations between the microbiome and the Immunoscore.

## 1.8. Project aims

Immunotherapy has revolutionised the treatment of some cancer types. However, the benefits are yet to be realised in CRC, the majority of which appears to be refractory to current immunotherapeutic regimens. For patients with relapsed or

metastatic disease there has been relatively little progress in the development of new therapeutic options.

A better understanding of the drivers of immunogenicity, particularly the germline, somatic and epigenetic factors, will provide the necessary breakthroughs to improve the efficacy of immunotherapy in CRC. It will generate novel biomarkers, which could expand the current pool of patients able to benefit from immunotherapy and provide novel options to harness the immune response to facilitate tumour rejection and better clinical outcomes.

This thesis explores germline, somatic and metagenomic determinants of the immune response in CRC. The germline component hypothesised that there is an association between immune gene eQTLs and the CRC immune environment, with the aim of establishing specific eQTL SNPs as potential biomarkers for immunotherapy. The somatic component hypothesised that neoantigen clonality (the inverse of intratumoral heterogeneity) is a stronger driver of the CRC immune environment than tumour mutational or neoantigen burden. This could serve as a more effective biomarker for targeting immunotherapy in CRC than microsatellite status. The final component explored the role of the gut microbiome in driving immune expression in CRC. These insights provide further support for targeted clinical trials exploring the role of neo-adjuvant and adjuvant immunotherapy in both dMMR and pMMR CRC at all disease stages.

I was fortunate to have access to the unique resources, tools and skills provided by the Genomics England 100 000 Genomes Project and the project collaborators, to drive this work [190].

# Chapter 2: Materials and Methods

## 2.1. Sample collection

### 2.1.1. Data and sample access

#### *2.1.1.1. The 100 000 Genomes Project*

Patient data and samples were obtained primarily from the 100 000 Genomes Project (100KGP) [190]. The 100KGP is a publicly funded genomics project with the aim of creating a genomic medicine service for the UK National Health Service (NHS). It is coordinated by Genomics England (GeL), which was set up by the UK Department of Health in July 2013. The 100KGP enrolled, collected and sequenced 100 000 genomes from 70 000 patients and their families, with cancer and rare diseases. The Genomics England Clinical Interpretation Partnerships (GeCIPs) co-ordinate the research activity. They are made of thousands of UK and international scientists and clinicians, organised into domains formed around related conditions. The Colorectal Cancer GeCIP domain studies data from patients with CRC. Clinical and WGS data from tumour tissue (resected specimens) and whole blood (germline DNA) are available within the GeL Research Environment. Sequencing of samples was reported to be at least to a read depth of 100x.

The West Midlands Genomic Medicine Centre (WMGMC), led by the University Hospitals Birmingham (UHB) NHS Foundation Trust was integral in recruiting and obtaining samples from patients in the West Midlands. The University of Birmingham's Human Biomaterials Resource Centre (HBRC) (Birmingham, UK) has tumour and normal colon samples for over 400 patients with CRC enrolled in the 100KGP in its Biorepository (Research Tissue Bank ethical approval from

National Research Ethics Service Committee North West – Haydock, reference number15/NW/0079). With appropriate permissions and ethical approval (see Appendix 2 and Appendix 3), WGS and clinical data, along with formalin fixed tissue blocks from resected specimens on these patients were available for performing the Immunoscore, RNA expression profiling and immunohistochemical analysis.

#### *2.1.1.2. Patient data sets*

Prior to obtaining data from the 100KGP, a pilot analysis was performed. The pilot analysis consisted of a local cohort of 50 patients with both colon and rectal tumours who had previously been recruited, and for whom germline and somatic WGS data and FFPE tumour samples were available [208].

Subsequently, genomic data and surgical resection specimens were obtained from 188 patients who had been recruited to the 100KGP. See Appendix 2 for the Intellectual Property Agreement from GeL. These patients had been treated at the UHB NHS Trust (Birmingham, UK) and formalin-fixed samples of their surgical resection specimens were available at the Biorepository. WGS data for the pilot cohort was accessed via the University of Birmingham's secure High Performance Computing (HPC) Linux-based interface, BlueBEAR (Birmingham Environment for Academic Research, http://www.birmingham.ac.uk/bear [209]). WGS data for the 100KGP cohort was accessed via the GeL Research Environment (https://re.extge.co.uk/ovd/). Both interfaces are user-restricted and encrypted.

Patients selected for inclusion included those with surgically resected CRC, with Stage 1 to 4 but predominantly Stage 2 and 3 tumours at the time of excision. Sampling was performed in real time as patients were being recruited to the 100KGP and their samples were made available in the Biorepository. Patients who had pre-operative radiotherapy were excluded, as this was shown to make the tumour tissue unsuitable for the Immunoscore. Radiotherapy induces significant tissue changes including fibrosis, mucus secretion and tumour regression, which preclude precise delineation of the tumour core and invasive margin, which are crucial for the Immunoscore [210]. Recruitment for this study stopped when samples for 208 patients were retrieved.

Data on clinical outcomes was available via the Clinical Portal at the UHB NHS Trust and corroborated with information available within the GeL Research Environment. Clinico-pathological criteria assessed included age at surgery, sex, ethnicity, primary tumour location, type of surgery performed, mucinous histology, tumour T stage, nodal status, extramural venous invasion (EMVI), disease stage, microsatellite status, *RAS* and *BRAF* mutation status, neo-/adjuvant treatment (predominantly chemotherapy and radiotherapy), recurrence-free survival (RFS) and overall survival (OS) at the time of analysis. All samples were anonymised by the HBRC to maintain non-traceability of patient information.

### 2.1.2. Tissue preparation

For each patient, FFPE specimen sections were prepared from tumour tissue blocks fixed in either Formal saline 10% (VWR International, Radnor, Penn, USA)

or Neutral Buffered Formalin (Fisher Scientific, Hampton, NH, USA) at the HBRC. For RNA extraction and 3' sequencing, 8µm thick specimen scrolls were prepared in Eppendorf tubes. Eight scrolls each of FFPE per patient specimen were available. For the Immunoscore, 4µm thick unstained slides were prepared according to specifications from HalioDx (see below). The FFPE scrolls were stored in Eppendorf tubes at room temperature, and fixed slides were stored in secure boxes with desiccants, at 4$^{o}$C.

## 2.2. Sample processing and sequencing

### 2.2.1. FFPE RNA extraction

RNA extraction was performed using the Covaris® total NA extraction protocol (Covaris, Inc., Woburn, MA, USA), using column-based purification. For each sample, two 8µm thick FFPE scrolls were emulsified in Tissue Lysis Buffer and proteinase K. Using adaptive focused ultrasonication with the Covaris E220 Evolution, each sample was processed for 300 seconds at 20$^{o}$C, with Peak Incident Power 175 Watts, at 200 cycles per Burst. Following incubation at 56°C for 30 minutes to release the RNA, and centrifugation to separate the RNA-containing supernatant from the DNA-containing tissue, the RNA-containing supernatant was de-crosslinked at 80°C and purified over a spin column. Removal of contaminant DNA was performed by incubation with the Invitrogen TURBO-DNA™ free kit (ThermoFisher Scientific, MA, USA) on the spin column for 30 minutes at room temperature. Following further column purification, 30µL of RNA was eluted for each specimen, and stored at -80$^{o}$C to minimise degradation. RNA quantification (ng/µL) was performed on the QuBit Fluorometer

(Thermofisher Scientific, MA, USA). The RNA integrity number (RIN) for each sample was assessed using the Agilent 4200 TapeStation Bioanalyzer system (Agilent Technologies, Santa Clara, USA).

### 2.2.2. 3' RNA library preparation

Following normalisation of samples to 20ng/µL, RNA library preparation was performed using the QuantSeq 3' mRNA-Seq Library Prep Kit FWD for Illumina by Lexogen® (Lexogen GmbH, Vienna, Austria). As this is a QuantSeq protocol using total RNA, no prior poly(A) enrichment or rRNA depletion was required. This protocol was chosen as it is able to generate Illumina-compatible libraries, even from very degraded or FFPE RNA. Library generation began with reverse transcription (with oligo (dT) priming containing the Illumina-specific Read 2 linker sequence). After first strand synthesis, the RNA was removed and second strand synthesis initiated by random priming and a DNA polymerase. The random primer contains the Illumina-specific Read 1 linker sequence. Second strand synthesis was followed by a magnetic bead-based purification step, after which the library was amplified by polymerase chain reaction (PCR) for 17-20 cycles each. External barcodes (with a combination of i5 and i7 indices) were introduced during the PCR amplification step. Following a second purification step, library quality control (QC) was performed on the Agilent TapeStation Bioanalyzer and $DV_{200}$ values, representing the percentage of fragments with greater than 200 nucleotides, were obtained for each sample. Samples were stored at -20$^{o}$C for sequencing. Library preparation was performed both manually and using the

Hamilton Microlab® Star™ robot (Hamilton Company, Reno, Nevada, USA), with equivalent results.

### 2.2.3. 3' RNA sequencing

RNA sequencing was performed in three batches due to the large sample size. The prepared libraries were pooled and quantified to 4nM. The libraries were subsequently denatured in 0.2 N sodium hydroxide (NaOH) (Sigma-Aldrich, St Louis, Missouri, USA) with Tris hydrochloric acid (HCl) at pH 7.0 (Sigma-Aldrich). Denatured libraries were diluted to 1.6pM concentration in Illumina hybridisation buffer (HT1; Illumina, San Diego, CA, USA) at 4$^{o}$C, and spiked with 1% Illumina PhiX control (Illumina) at 20pM. PhiX is derived from a bacteriophage genome. It is available as a concentrated library (10nM in 10µL), with an average size of 500 bp and a balanced base composition of approximately 45% G and C and 55% A and T. At a low concentration, it serves as sequencing control to monitor run quality (including cluster generation, sequencing, and alignment) [211].

Each prepared library was loaded into a separate Illumina NextSeq™ 500/550 Reagent Cartridge at room temperature. The Illumina NextSeq™ 500/550 High Output Flow Cell Cartridge and Reagent cartridge were loaded into the Illumina NextSeq™ 500 system. Sequencing was performed on the NextSeq to a read length of 75bp per sample. Reads were generated towards the poly(A) tail and directly correspond to the messenger RNA (mRNA) sequence. Information was stored in the Illumina BaseSpace Sequence Hub (https://basespace.illumina.com). Sequence data in bcl file format was transferred to the BlueBEAR Linux platform, where the files were converted to fastq format using the bcl2fastq application [212]. The fastq files were transferred

to the Partek Flow® (Partek Inc, St Louis, Missouri, USA) server for downstream analysis in our local server (http://fuggle.bham.ac.uk:8080/flow/login.xhtml).

### 2.2.4. Partek Analysis

The Lexogen QuantSeq unique molecular identifier (UMI) 0604 pipeline on Partek was used to align the reads to obtain both gene and transcript counts (Figure 2.1). First, the unaligned reads were trimmed and quality controlled. The transcript fragments in the Lexogen QuantSeq pipeline arise at the 3' end of mRNA sequence, so the Partek pipeline includes trimming of the poly(A) tails and low-quality adapter sequences. The fragments were then aligned using the STAR aligner [213], to the reference genome hg19/GrCh37 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/) and Ensembl release 75 (https://grch37.ensembl.org/index.html), with post-alignment QC statistics provided. The aligned reads were transferred for gene counting, using HTSeq-count [214] to count the number of reads per gene. Mapped reads were only counted if they uniquely mapped to the exons of the gene body.

Gene and transcript counts were normalised in the Partek suite (as counts per million). The normalisation equation involves multiplication of the raw read of each sample on each feature (gene) by 10e6 and division by the total mapped reads of that sample, and addition of 10e-04 to each result. The principal components analysis (PCA) node was used to visualise the samples. As there was a batch effect visible in the normalised reads, a correction for batch effect was performed by selecting the 'batch' and 'Immunoscore' factors and interactions, and the PCA node was once again used to determine that the batch effect had been eradicated.

To determine associations between the Immunoscore and gene counts, after filtering out ribosomal and mitochondrial genes, the Gene Set Analysis (GSA) module in the Differential Analysis node was used to compare patients with low and high Immunoscores. This generated statistical associations and a hierarchical clustering/heat map.

Gene and transcript count information for specific genes and groups was downloaded from the Partek Genomics suite in .txt format for more detailed analysis.

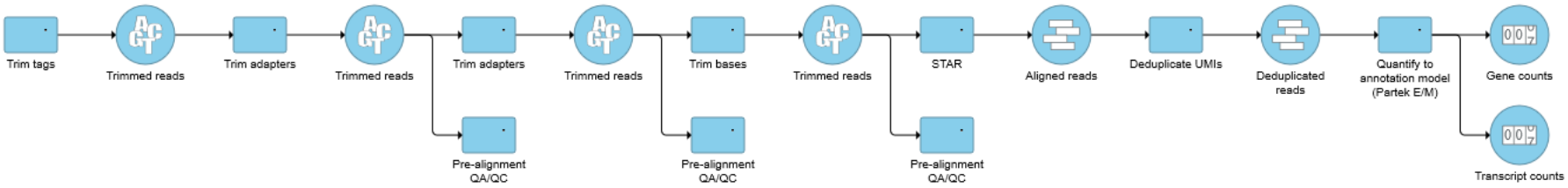



**Figure 2.1.** The Partek Lexogen QuantSeq pipeline. E/M – expectation/maximisation algorithm. QA – quality alignment. QC – quality control. STAR – Spliced Transcripts Alignment to a Reference [213]. UMI – unique molecular identifier.

## 2.3. Immunohistochemical analysis

### 2.3.1. The Immunoscore®

For the Immunoscore, from each formalin-fixed tumour tissue block, four unstained slides at 4µm thick were prepared according to instructions from HalioDx® (Luminy, Marseille,France)(https://www.immunoscore-colon.com/fileadmin/Documentation_Colon/EU_CLIA_et_CEIVD/Material_Requirements_Instructions_CE-IVD_MRS_MG.pdf).

Tumour tissue was fixed according to the standard HBRC laboratory protocol, cut to 4µm thick sections, mounted on positively charged glass slides and dried at 56°C overnight. The slides were placed in secured slide boxes and stored at 4°C, with desiccants, until transfer by shipping to HalioDx®. The Immunoscore was performed according to the internationally validated and patented protocol, which is described below [118]. All samples were shipped for analysis within twelve weeks of preparation.

The Immunoscore protocol is briefly described. Two prepared slides from each patient were processed. After overnight incubation at 56ºC, the slides were stained with monoclonal antibodies against CD3 and CD8 (rabbit anti-human CD3 (clone HDx2, HalioDx) and mouse anti-human CD8 (clone HDx1, HalioDx)), using the Ventana Benchmark XT (Roche Diagnostics, Basel, Switzerland). Revelation was with the Ultraview Universal DAB IHC Detection Kit (Ventana, Tucson, AZ, USA), and counterstaining with Mayer's haematoxylin. After incubation for 3.5hr and washing with ethanol and washing with de-ionised water, the slides were scanned on the Nanozoomer XR (Hamamatsu Photonics K.K.,

Hamamatsu City, Shizuoka, Japan). Digital images of the stained tissue sections were obtained at 20× magnification and 0·45 μm/pixel resolution. A specially developed Immunoscore module integrated into the image-analysis system of a Developer XD digital pathology software (Definiens, Munich, Germany) was used to determine the densities of CD3+ and CD8+ T cells in colon tumour and invasive margin regions.

The mean and distribution of the staining intensities were monitored to obtain an internal QC of each slide. Slides that failed QC were rejected and the staining protocol repeated with an alternate pair of slides. Samples that failed repeat staining and QC, or for which an insufficient margin was found were removed from further analysis.

For each case, CD3+ and CD8+ cell densities in the centre of the tumour (CT) and invasive margin (IM) regions were converted into percentiles. The mean of four percentiles (two markers, two regions) was calculated and converted into an Immunoscore value. For the three-group classification, a 0–25% density was scored as low, a density between 25% and 70% was scored as intermediate, and a density between 70% and 100% density was scored as high. For the five-group classification, the mean percentiles were scored as 0-10% (score 0), >10 to 25% (score 1), >25 to 70% (score 7), >70 to 90% (score 3) and >90 to 100% (score 4) [124].

I was privileged to collaborate with Professor Jérôme Galon, who pioneered the Immunoscore technique, and is a co-founder of HalioDx and the Chairman of its Scientific Advisory Board. I visited the HalioDx laboratory to learn the techniques

and participate in performing the Immunoscore on the first batch of samples that were analysed.

### 2.3.2. Antibody staining and expression analysis

MHC Class II staining was performed in collaboration with Dr Phillipe Taniere at the UHB NHS Trust Pathology Laboratories. The MHC class II (HLA DR + DP + DQ) antibody CR3/43 (ab17101)) (Abcam, Cambridge, UK), which is clinically validated for the ANICCA-Class II trial (a phase II trial assessing nivolumab in strong class II expressing microsatellite stable colorectal cancer) [215] was used. An expert pathologist (Dr Phillipe Taniere, at the UHB) reviewed the slides, and a Class II expression percentage score was given following an internal MHC Class II colorectal pathology interpretation guide.

## 2.4. Bioinformatics analyses

### 2.4.1. eQTL analysis

#### *2.4.1.1. Selection of candidate eQTL genes*

Vogelstein *et al.*[161] used data from the Multiple Tissue Human Expression (MuTHER) Project [155] (www.muther.ac.uk) and interrogated the Gene Ontology (www.geneontology.org) and Kyoto Encyclopedia of Genes and Genomes (www.genome.jp/kegg/) databases to derive a list of 385 immunomodulatory genes. They mined the list of top *cis*- eQTLs per probe in lymphoblastic cell lines (LCLs) from the MuTHER database for all the probes representing this panel of immunomodulatory genes. 50 SNPs with the most significant *cis*-eQTL activity (ranked with p values less than $4.46 \times 10^{-8}$) in cells of

the immune system were selected for genotyping. Confirmation genotype-expression associations for the 50 probe-SNP pairs was done using publicly available expression data from ArrayExpress (accession no. E-TABM-1140 [216]). Access to the genotype data set was obtained from the Department of Twin Research, King's College London. Twins (339 twin-pairs) from the same pair were separated into two twin sets and independent eQTL analyses were performed for each twin set using Spearman Rank Correlation. Genotype-expression correlations were assessed in 777 participants (including 339 twin-pairs) under three genetic models of inheritance (i.e. genotypic, dominant, and recessive) using Spearman Rank Correlation test. They successfully genotyped 40 SNPs which were passed to association analyses. The gene list is below (Table 2.1).

**Table 2.1.** List of top 40 eQTL SNPs (reproduced from Vogelsang *et al.* [161])

| | | | LCL-combined | |
|---|---|---|---|---|
| **SNP** | **GENE** | **PROBE** | **Beta** | **P value** |
| **rs4577037** | *IL16* | ILMN_2290628 | 0.65 | 2.14E-58 |
| **rs7574070** | *STAT4* | ILMN_1785202 | 0.48 | 2.19E-57 |
| **rs841718** | *STAT6* | ILMN_1763198 | 0.26 | 1.36E-53 |
| **rs8101605** | *LILRB1* | ILMN_1708248 | 0.44 | 2.13E-47 |
| **rs2071304** | *SPI1* | ILMN_1696463 | -0.20 | 2.24E-44 |
| **rs11569345** | *CD40* | ILMN_2367818 | 0.54 | 3.21E-38 |
| **rs17001247** | *CXCL10* | ILMN_1791759 | -0.77 | 3.78E-35 |
| **rs11919943** | *CCR1* | ILMN_1678833 | 0.34 | 9.39E-29 |
| **rs4500045** | *PAG1* | ILMN_1736806 | 0.17 | 3.25E-27 |
| **rs6673928** | *IL19* | ILMN_1799575 | 0.12 | 5.66E-23 |
| **rs10760142** | *C5* | ILMN_1746819 | 0.12 | 4.82E-22 |
| **rs859** | *IL16* | ILMN_1813572 | 0.15 | 1.09E-21 |
| **rs4500045** | *PAG1* | ILMN_2055156 | 0.20 | 6.45E-21 |
| **rs9921791** | *MLST8* | ILMN_1789240 | 0.17 | 2.52E-20 |
| **rs6692729** | *PSEN2* | ILMN_2404512 | -0.10 | 5.71E-20 |
| **rs7584870** | *SOCS5* | ILMN_2350970 | -0.08 | 2.50E-19 |
| **rs2701652** | *IRAK3* | ILMN_1661695 | 0.27 | 2.96E-19 |
| **rs4848306** | *IL1B* | ILMN_1775501 | 0.27 | 4.68E-19 |
| **rs1551565** | *CAMK4* | ILMN_1767168 | 0.10 | 2.78E-18 |
| **rs11203203** | *UBASH3A* | ILMN_2338348 | 0.14 | 3.16E-18 |
| **rs1049337** | *CAV1* | ILMN_1687583 | 0.06 | 6.31E-17 |
| **rs4808137** | *UBA52* | ILMN_2368576 | -0.12 | 5.69E-16 |
| **rs1149901** | *GATA3* | ILMN_2406656 | -0.12 | 1.17E-15 |
| **rs6692729** | *PSEN2* | ILMN_1714417 | -0.10 | 2.97E-15 |
| **rs7036417** | *SYK* | ILMN_2059549 | 0.14 | 1.75E-14 |
| **rs3807383** | *GIMAP5* | ILMN_1769383 | -0.27 | 2.38E-14 |
| **rs1378940** | *CSK* | ILMN_1754121 | 0.09 | 1.27E-13 |
| **rs12401573** | *SEMA4A* | ILMN_1702787 | -0.15 | 1.97E-13 |
| **rs9863627** | *PAK2* | ILMN_1659878 | 0.17 | 2.17E-13 |

| | | | LCL-combined | |
|---|---|---|---|---|
| **SNP** | **GENE** | **PROBE** | **Beta** | **P value** |
| **rs4500045** | *PAG1* | ILMN_1673640 | 0.07 | 4.81E-13 |
| **rs4402765** | *IL1A* | ILMN_1658483 | -0.22 | 3.31E-12 |
| **rs13331952** | *CKLF* | ILMN_2414027 | 0.37 | 7.64E-12 |
| **rs2291299** | *CCL5* | ILMN_2098126 | -0.20 | 1.35E-11 |
| **rs4796105** | *CCL5* | ILMN_1773352 | -0.21 | 2.44E-11 |
| **rs13331952** | *CKLF* | ILMN_1712389 | 0.47 | 3.83E-11 |
| **rs2295359** | *IL23R* | ILMN_1734937 | -0.14 | 2.26E-10 |
| **rs665241** | *FYB* | ILMN_1796537 | 0.09 | 5.89E-10 |
| **rs6695772** | *BATF3* | ILMN_1763207 | -0.16 | 6.93E-10 |
| **rs7720838** | *PTGER4* | ILMN_1795930 | -0.07 | 1.78E-09 |
| **rs2276645** | *ZAP70* | ILMN_1719756 | -0.15 | 6.38E-09 |
| **rs4469949** | *CD27* | ILMN_1688959 | -0.19 | 8.48E-09 |
| **rs10422141** | *TICAM1* | ILMN_1724863 | -0.08 | 1.24E-08 |
| **rs11161590** | *BCL10* | ILMN_1716446 | 0.08 | 3.67E-08 |
| **rs152112** | *ITK* | ILMN_1699160 | 0.05 | 4.26E-08 |

eQTL = expression quantitative trait loci, SNPs = single nucleotide polymorphisms

SNP chromosomal positions were confirmed using the UCSC Genome Browser (http://genome.ucsc.edu/) [217] with reference genome GRCh38/hg38 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39).

The extended eQTL SNP list from MuTHER was also interrogated and filtered for all LCL immune gene *cis*-eQTLs with expression p values <0.05. This provided a list of 385 eQTL SNPs for 269 genes (Appendix 4).

#### *2.4.1.2. Variant calling*

The computations described below were performed using the HPC service within the GeL Research Environment and the University of Birmingham's BlueBEAR Linux interface. Sequencing was performed by Genomics England using the Genomics England (GeL) v4 pipeline [218]. Downstream analyses by colleagues in the 100KGP showed that variant allele frequencies (VAFs) computed using the Isaac pipeline [219] are potentially biased, due to the preferential soft-clipping of semi-aligned reads [220]. The Isaac "–clip" parameter soft clips read ends until five consecutive bases are matched with the reference genome. This results in loss of support for alternate alleles occurring within five bases of each read end. FixVAF was developed and used to reduce the allelic bias, by soft clipping all reads by five bases at each end regardless of their status [220].

The Strelka germline and somatic variant caller [181] was used for variant calling. Germline and tumour whole genome sequencing Variant Call Format (vcf) files, in compressed format (.vcf.gz) were accessed within the GeL Research Environment's password protected LabKey server.

#### *2.4.1.3. SNP extraction/filtering*

SNP filtering was performed using the VCFtools module (vcftools.sourceforge.net) for each chromosome and SNP position. The RefSeq IDs of the top 40 eQTL SNPs explored by Vogelsang *et al.* were obtained and uploaded into the Bluebear Linux platform and the GeL Research Environment .txt format. Germline VCF files for each sample were filtered against these SNPs using VCFtools. The resultant .vcf.gz files were indexed using the samtools "tabix" function [221] and the genotypes were extracted in R. SNPs were coded as both additive (1,2, and 3 representing wild-type, heterozygous variant and homozygous variants) and dominant (1,2, and 2 representing wild-type, heterozygous variant and homozygous variants).

SNP-Immunoscore associations were performed using ordinal logistic regression with the "MASS", "caret" and "tidyverse" packages in R (https://www.r-project.org/). The main analysis was performed on the SNPs using the additive genotypic model. Supplementary analysis was also performed using the dominant model. The stepAIC function ("forward") was used for stepwise selection and subsequently the "train" function in the "caret" package was used to generate a predictive model with the best fit. Finally, p values were corrected for multiple testing using both the Bonferroni and false discovery rate (FDR) techniques. To examine the effect of gene expression of those SNPs in human cancer, the Human Protein Atlas (www.proteinatlas.org) [222] was consulted.

#### *2.4.1.3. a. Quality control and principal components analysis*

As a method of quality control, joint genotype files for each chromosome were interrogated to ensure that variants were called in all samples and in all chromosomes to an accuracy of ≥ 95%.

Finally, principal components analysis (PCA) of the significant SNPs by ethnicity was performed in R using the packages “stats” and “ggfortify”. The prevalence of each SNP by ethnicity was compared to determine any potential confounding impact of this on the SNP-Immunoscore associations.

### 2.4.2 Estimation of intratumoral heterogeneity

#### *2.4.2.1 Neoantigen prediction*

Neoantigen prediction was performed using the Neopredpipe pipeline [174] in collaboration with Professor Trevor Graham and Dr Eszter Lakatos at the Queen Mary University, London, United Kingdom. Neopredpipe predicts only Class I peptides but can generate both SNVs and indels. HLA typing was performed using HLAtyper [223] as part of Illumina’s sequencing panel procedure. Peptide binding prediction is incorporated into the pipeline using netMHCpan 4.0 [169], and neoantigen prediction with POLYSOLVER [180]. Predicted neoantigens were cross-referenced with normal peptides using PeptideMatch [179], which assesses for novelty of candidate epitopes by searching against a reference proteome, for example from Uniprot or Ensembl.

TMB data was obtained from LabKey within the Research Environment server and from the BlueBEAR interface. TMB is calculated per Mb by dividing the number of non-synonymous SNVs by genome size. For the Pilot samples, the

Illumina TruSight Oncology 500 (TSO500) panel was used to profile the samples and generate a TMB, using a panel size (genome size) of 1.85Mb [224] . For the 100KGP samples, the panel size was taken as the whole genome (3000Mb). We have shown that the estimation of TMB using the TSO500 panel is highly accurate and strongly correlated with the WGS TMB ($R^2 = 0.9$) [224]. Comparison plots of TMB and neoantigen and indel burden were created in R.

#### *2.4.2.2. Neoantigen clonality and intra-tumoral heterogeneity*

#### *2.4.2.2. a. Estimation of intratumoral heterogeneity using DPClust*

As only single region sequencing data was available for each sample, copy number alterations (CNAs), tumour purity and cancer cell fraction (CCF) estimates were required to determine the proportion of each mutation within the sample. CNA calculation was performed in the RE by Dr David Wedge's team at the University of Oxford, using ASCAT (allele-specific copy number analysis of tumors, version 2.2 [225]), as described by Bolli *et al.* [226]. ASCAT was shown, in samples from multiple myeloma, to reliably identify clonal and subclonal copy number changes in tumours using WES data [225].

Calculation of the cancer cell fraction and mutation clustering were performed with DPClust [56], which uses a Dirichlet process based approach to estimate the number of mutation clusters in the data. This method accounts for the effects of sample purity and copy number aberrations on allele frequency [226]. Class I neoantigen burden data was obtained from Neopredpipe, and filtered against the DPClust data, which selected out all neoantigens per sample. The proportion of each neoantigen in each tumour was calculated as a proportion from 0 to 1.

Some neoantigens were present at a proportion of greater than 1, for three main reasons. First, random sampling of reads can lead to substantial variation in allele frequency distributions. This is particularly so for mutations with low coverage and low purity samples. Secondly, some SNPs may be miscalled as somatic SNVs, and therefore have higher allele frequency than expected due to being present in both normal and tumour cells in the samples. Finally, copy number callers have small errors in copy number calling, and may miss small CNAs, leading to incorrect adjustment of allele frequencies in these regions.

Neoantigens present at a proportion of 1 (or above) were coded as 'clonal'. To determine the degree of intratumoral heterogeneity (ITH), the total number of subclonal neoantigens was divided by the total number of neoantigens, to derive a score as a proportion ranging from 0 to 1. A score closer to 1 represents a more heterogeneous tumour. This score was correlated with the Immunoscore.

#### *2.4.2.2. b. Estimation of intratumoral heterogeneity using the MATH score*

The MATH score [188] is a percentage score of the ratio of the width to the centre of the distribution of mutant allele fractions. As this incorporates all tumour mutations, and therefore includes both Class I and II neoantigens. For each tumour, it is calculated as a percentage of the median absolute deviation (MAD) and the median of its mutant allele fractions at tumour-specific mutated loci.

The equation is as follows: 100 * MAD / median.

The BCFtools "query" command [221] was used to filter out the per SNP read depth (the "DP" field) and the tumour mutant reads (the "ID/TAR") fields from the tumour WGS variant call files (vcf.gz). The median and MAD of these tumour-

specific mutated loci was calculated for each tumour and a MATH score was derived.

The authors of the MATH score note the possibility that tumours with high mutation rates could have greater heterogeneity. Their data did not show any association between number of mutations and the MATH score. In this data set, a comparison of the MATH score with the TMB was also performed to determine if any linear association existed.

### *2.4.3. Metagenomic analysis*

Metagenomic sequencing was performed using Kraken2 [206], and the results were visualised with Pavian [227]. Kraken is a fast and accurate program, which assigns taxonomic labels to metagenomics DNA sequences. It uses exact k-mer matches to a database, rather than inexact alignment of sequences, which vastly improves its accuracy [228]. The database of microbial genomes is large and growing, and the accuracy continues to improve.

First, the reads in each tumour WGS BAM (binary alignment map) file were extracted using samtools [221]. This generated bam files, of size ranging from 10GB to 20GB. The bam files were sorted using samtools. Sorted bam files were converted to compressed fastq files using picard tools (version 2.10.1-Java-1.8.0-131) [229]. Finally, the minikraken database was uploaded from the kraken2 website (http://ccb.jhu.edu/software/kraken/). The kraken2 command was used to match the compressed fastq files to the kraken2 database, generating out test and report files for each sample.

The report files were uploaded into the Pavian browser in R (pavian::runApp(port=5500)), where visualisation was performed for each sample. Pavian generates Sankey plots for each sample. Sankey diagrams display the flow of reads for each sample from the root of the taxonomy (the domain) to more specific ranks (the species). The width of each flow is proportional to the number of reads [227]. Information on the percentages of classified compared with unclassified reads, and percentages of reads by domain (chordate, bacterial, viral, fungal and protozoan) were also available to download in text csv format.

Finally, detailed data on the number of raw reads by domain, phylum, class, order, family, genus and species for each sample was generated and available to download in text csv format, where they could be filtered for downstream analysis.

## 2.5 Statistical analysis

### 2.5.1. Sample size considerations

Statistical analyses were performed using ordinal logistic regression to identify associations between genotypes for MuTHER eQTLs and the corresponding tumour Immunoscore levels. We adjusted for potential confounders including ethnicity, sex, and background genetic context using principal components analysis of common SNP genotype matrices.

Power calculations were performed using both R (R Foundation for Statistical Computing, Vienna, Austria, https://www.r-project.org/) and G*Power statistical packages (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/). In statistical analysis, effect sizes are conventionally grouped into small (0.1), medium (0.25) and large (0.4) [230].

For eQTL analysis, the “MASS” and “ordinal” packages in R were used, and simulations were run using a range of sample sizes from 50 to 300. Assuming a range of MAFs (0.1 to 0.4), a range of effect sizes (mean shift in Immunoscore, from 0.2 to 2.0) and an additive genomic model (1, 2 and 3), adequate statistical power (1-β, in the region of 0.8) could be obtained at a range of effect sizes with a sample size between 150 and 200 (see Results Chapter 3.2.1).

Naive whole genome scans are subject to significant high false discovery issues due to multiple testing; however, this was addressed by correcting for multiple testing using the Bonferroni and FDR method. As the data generated was unique, independent cohort replication was not possible (for example, using an openly available data set). Cross-validation testing was used to test robustness within our cohort.

For power analysis of the RNAseq data, the G*Power package was used. With three Immunoscore categories (Low, Intermediate and High) and a range of gene expression level effect sizes (0.2 to 1.0 in in increments of 0.05), *a priori* and *post-hoc* analyses of sample size requirements were performed using a one-way ANOVA with fixed effects.

### 2.5.2. Statistical tests

Statistical analysis was performed using R. Patient demographics and disease characteristics were compared across Immunoscore categories using the "ggpubr" package with the appropriate non-parametric tests (Kruskal-Wallis and Wilcoxon's tests). The Immunoscore ranked categories – "Low", "Intermediate" and "High" – were used for these comparisons.

Associations between the Immunoscore and eQTL SNPs were performed by ordinal logistic regression, using the "MASS" and "caret" packages in R. Survival data analysis (OS and RFS) was also performed in R using the "survival" and "survminer" packages.

Finally, metagenomic analysis was performed, after uploading the read counts in txt format from pavian, using both the "rcorr" and "ggpubr" packages in R for comparison across Immunoscore and microsatellite status categories, with correction for multiple testing using FDR. The results were presented in both tabular and figure formats.

## 2.6. Contributions of the author and collaborators to the study

Patient recruitment to the 100KGP was performed by the WMGMC under the directorship of Professor Dion Morton. With the Research Ethics Committee approval, patients diagnosed with CRC in the elective setting, and deemed suitable for surgical resection were counselled and recruited to the study. The author collected anonymised data on recruited patients, and added patients from the UHB site with tumour samples available at the HBRC to this project.

Germline and somatic WGS, and somatic TMB analysis were performed by the 100KGP research collaborative, and genomic data in the form of BAM and VCF file formats were uploaded to the Research Environment for the use of individual researchers. The author used these data to perform germline SNP extraction, PCA and logistic regression analysis.

DPClust estimation was performed by Dr Wedge and his team and the results were made available in the Research Environment. Neoantigen burden estimation was performed using Neopredpipe by Professor Trevor Graham's team in the Research Environment. The author used both pipelines to determine neoantigen clonality, and also replicated these pipelines on the local cohort samples which were outside the research environment. Close advice and supervision were provided by Dr Eszter Lakatos. Neoantigen clonality calculations with the MATH score were performed by the author, using the available somatic WGS data.

Class II immunohistochemistry staining and scoring was performed by Professor Phillipe Taniere and his team at the UHB Pathology Department.

Metagenomic analysis of somatic WGS data using kraken2 was performed by the author using the available bioinformatics tools within the Research Environment.

Finally, statistical power calculations were performed with support from Dr Christopher Yau at the University of Birmingham Centre for Computational Biology.

# Chapter 3: Clinico-pathological data results

## 3.1. Introduction

This chapter presents the clinical and pathological information derived from the patient data set used for the analyses in this thesis. It compares the data set with the wider CRC patient population. This is necessary to determine that the conclusions drawn are reliable, valid, potentially reproducible, and applicable to the wider patient population. It clarifies that the sample size provides sufficient statistical power to perform the analyses that follow. It accounts for any data attrition and discusses the possible consequences they may have for the validity of the analyses that follow.

Secondly, the Immunoscore results are displayed. The distribution of the Immunoscore results in this data set is shown to be statistically identical to that in the international validation study by Pages *et al.* [118]. Survival analysis confirms that the Immunoscore has significant prognostic value in the research patient population, and that it is significantly associated with other pathological makers including microsatellite status and disease stage at diagnosis.

## 3.2 Results

### 3.2.1. Statistical power calculations

Power calculations were performed using both R and G*Power. For the eQTL analysis, the “MASS” and “ordinal” packages in R were used, and simulations were run using a range of sample sizes from 50 to 300, and a range of effect sizes (defined as the mean shift in the numerical Immunoscore value, between 0 and 4). The results show sufficient statistical power at sample sizes between 150 and 200, with MAFs ranging from 0.1 to 0.4 (Table 3.1, Figure 3.1).

**Table 3.1.** Power calculations at a range of minor allele frequencies and sample sizes

| MAF | Sample size | Effect size | Power |
|---|---|---|---|
| 0.1 | 50 | 1.1359059 | 0.81 |
| 0.1 | 100 | 0.8070501 | 0.83 |
| 0.1 | 150 | 0.6932246 | 0.87 |
| 0.1 | 200 | 0.5782351 | 0.80 |
| 0.1 | 250 | 0.4625765 | 0.79 |
| 0.1 | 300 | 0.4625765 | 0.82 |
| 0.2 | 50 | 1.0633364 | 0.87 |
| 0.2 | 100 | 0.6740935 | 0.75 |
| 0.2 | 150 | 0.5397450 | 0.79 |
| 0.2 | 200 | 0.5397450 | 0.85 |
| 0.2 | 250 | 0.5397450 | 0.98 |
| 0.2 | 300 | 0.4046508 | 0.82 |
| 0.3 | 50 | 0.9196502 | 0.77 |
| 0.3 | 100 | 0.7695889 | 0.82 |
| 0.3 | 150 | 0.6168103 | 0.84 |
| 0.3 | 200 | 0.4626365 | 0.81 |
| 0.3 | 250 | 0.4626365 | 0.87 |
| 0.3 | 300 | 0.4626365 | 0.88 |
| 0.4 | 50 | 1.0315736 | 0.83 |
| 0.4 | 100 | 0.6937726 | 0.75 |
| 0.4 | 150 | 0.6937726 | 0.91 |
| 0.4 | 200 | 0.5206101 | 0.88 |
| 0.4 | 250 | 0.5206101 | 0.95 |
| 0.4 | 300 | 0.5206101 | 0.95 |

MAF = minor allele frequency. A range of power calculations for sample sizes 50 to 300 in increments of 50, showing the required effect sizes at MAFs 0.1 to 0.4 to give statistical power >0.75.

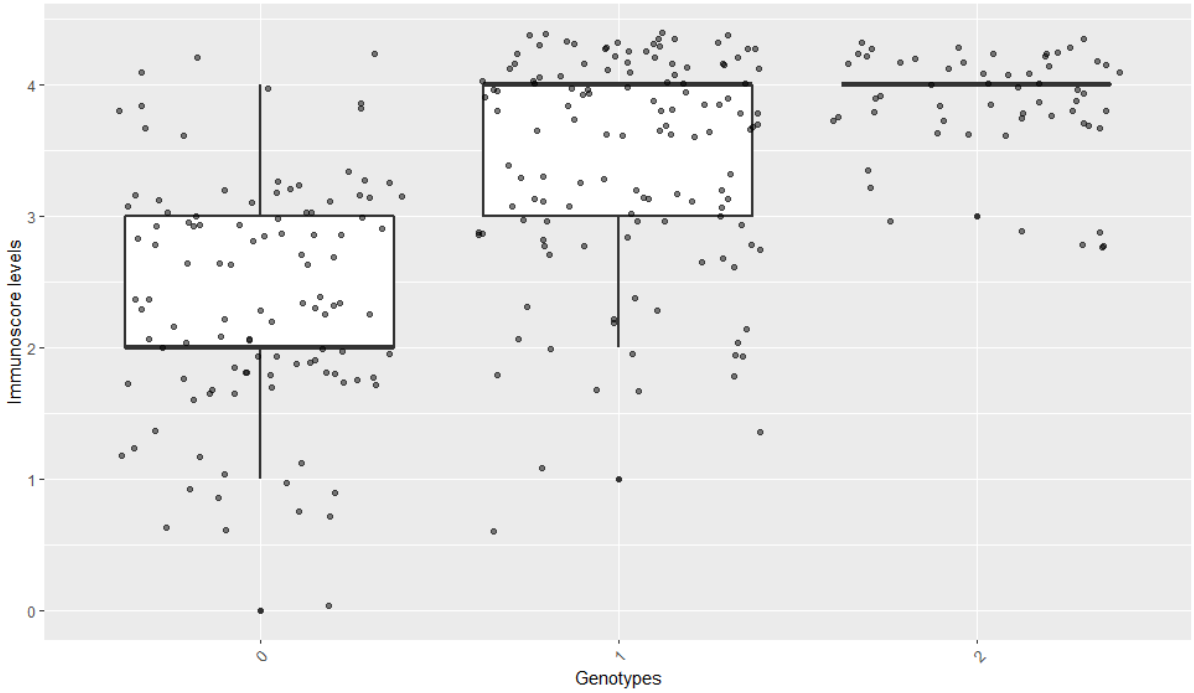



**Figure 3.1.** Boxplots illustrating the simulated differences in Immunoscore levels (0 to 4), with genotypes displayed as 0, 1 and 2. 0 = wild type. 1 = heterozygous mutant. 2 = homozygous mutant.

For power analysis of the RNAseq data, the G*Power package was used. With three Immunoscore categories (Low, Intermediate and High), an estimated range of gene expression level effect sizes (0.2 to 1.0 in in increments of 0.05), sufficient statistical power was achievable with a sample size of less than 200 (Figure 3.2).

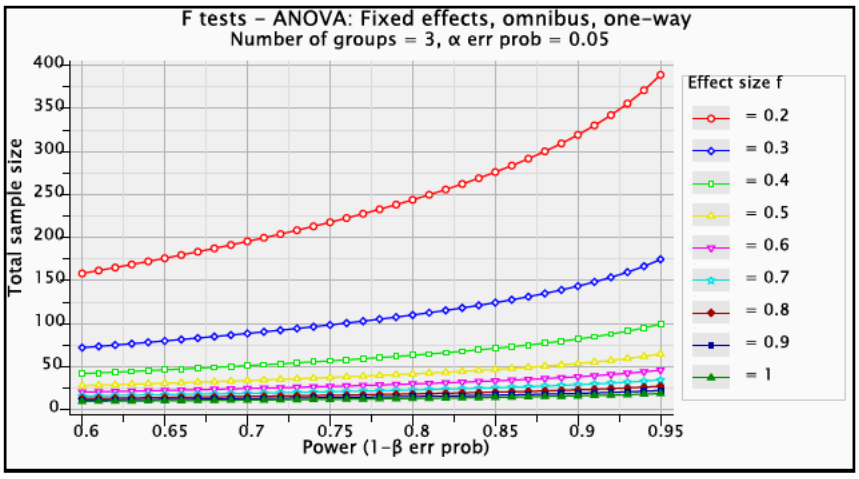



**Figure 3.2.** G*Power-derived plot illustrating the range of statistical power obtained at effect sizes ranging from 0.2 to 1.0. Statistical power at 0.8 and above is obtained at a sample size of less than 200 with an effect size of 0.3 or greater. One-way ANOVA with fixed effects was used. ANOVA = analysis of variance. β = Type II error rate.

*Post hoc* analysis of the RNAseq data was performed using the results from the analysis of the co-ordinate immune response cluster (CIRC) data set, as described in Chapter 5. The mean CIRC in each Immunoscore group was inputted into G*Power and an effect size was calculated (f=0.268). This was used to model the sample size required for adequate statistical power (β = 0.80). This was achieved at a sample size of n = 137 (Figure 3.3).

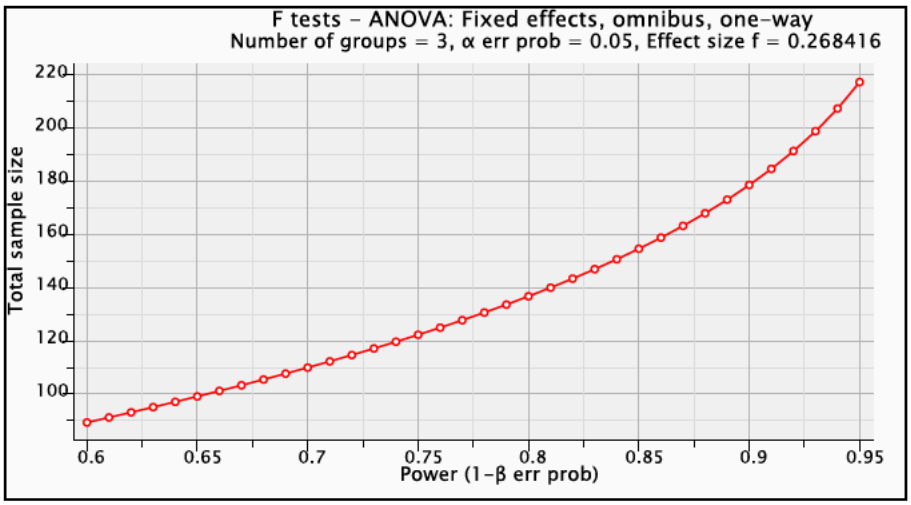



**Figure 3.3.** G*Power-derived plot illustrating the range of statistical power obtained at effect size of 0.268. Statistical power at 0.8 and above is obtained at a sample size of 137 or more.

### 3.2.2 Patient data set

Two hundred and thirty-eight (238) patients who had been recruited at the UHB Queen Elizabeth Hospital site were included in this study. This was to account for expected data attrition of approximately 10 to 20% of cases over the course of the study. Data and samples from both patient cohorts were analysed identically and combined for all analyses performed.

60.9% of patients were male (Table 3.2). UK national statistics from 2015-2017 show that 55.4% of new diagnoses are made in male patients [231]. Although there was a higher percentage of male patients in this data set, this difference was not statistically significant (Pearson's $\chi^2$ test, $p = 0.098$).

**Table 3.2.** Number and percentage of cases of colorectal cancer by sex. Comparison of national [231] and research data set

| | **Male** | **Female** |
|---|---|---|
| National data (n/%) | 22844 (55.4) | 18421 (44.6) |
| Research data (n/%) | 145 (60.9) | 93 (39.1) |

There is no statistically significant difference between the national and local data sets. Pearson's $\chi^2$ test, p = 0.098. n = number.

The mean and median age at diagnosis were 68 years and 69 years respectively, with a range from 31 to 88. In national data, mean age at diagnosis was 69.4 years [232], with peak incidence of diagnosis at ages 85 to 89 for men and women [231] (one-sample t-test, p = 0.031).

Disease stage was determined using the UICC-TNM classification of colorectal cancer [13]. Most patients had Stage 1 to 3 disease (92.9%) (Table 3.3). Patients in the national data set [233] had more advanced disease (Stage 4 = 23.3% compared with 7.1% in the research data set, Pearson's $\chi^2$ test, p <2.2e-16). As most patients with metastatic disease do not have surgical resection [20], this difference reflects the selection of patients deemed suitable for surgical resection of the primary tumour.

**Table 3.3.** Number and percentage of cases of colorectal cancer by stage at diagnosis. Comparison of national [233] and research data set

| | **Stage 1** | **Stage 2** | **Stage 3** | **Stage 4** | **Unknown** |
|---|---|---|---|---|---|
| National data (n/%) | 5782 (16.6) | 8041 (23.1) | 9490 (27.3) | 8122 (23.3) | 3390 (9.7) |
| Research data (n/%) | 27(11.3) | 99 (41.6) | 95 (39.9) | 17 (7.1) | 0 (0) |

The research data set was more likely to have earlier stage disease. $\chi^2$ test, p <2.2e-16. n = number.

Data on self-described ethnicity at diagnosis was collected for all patients and compared with available national data for patients diagnosed with CRC in 2015 [234]. Patients were grouped into five categories for simplicity of analysis. Most patients identified as White (87.8%) (Table 3.4). This was lower than for patients identified as White in the national data set (92.4%). This difference was statistically significant (Pearson's $\chi^2$ test, p = 2.2e-13). However, compared with the general West Midlands population, which has a higher Black and Minority Ethnic population than the wider UK population [235], there was a higher proportion of patients who identified as White recruited to this study (88% compared with 79.2%, Pearson's $\chi^2$ test, p = 0.001). Unfortunately, regional data on the ethnicity at diagnosis of patients with CRC was not available to make further comparisons.

**Table 3.4.** Number and percentage of cases of colorectal cancer by ethnicity at diagnosis. Comparison of national [234] and research data set

| | **Asian (all)** | **Black (all)** | **White** | **Other** | **Unknown** |
|---|---|---|---|---|---|
| National data (n/%) | 2366 (1.5) | 1803 (1.1) | 149149 (92.4) | 1201 (0.7) | 6899 (4.3) |
| Research data (n/%) | 11 (4.6) | 6 (2.5) | 209 (87.8) | 10 (4.2) | 2 (0.8) |

There was a lower percentage of self-described White patients in the research data set. Pearson's $\chi^2$ test, p = 2.2e-13. n = number.

50% of patients had left sided-tumours (n = 119), with 23.1% occurring in the rectum or recto-sigmoid junction. While this is a lower proportion than national data, in which 27.8% of cases occur in the rectum [231], it is not statistically significant (Pearson's $\chi^2$ test, p = 0.123). All tumours were adenocarcinomas.

Microsatellite status was determined by either immunohistochemical staining for DNA mismatch repair proteins (MLH1, MSH2, MSH6 and PMS2) or polymerase chain reaction (PCR) amplification. Data was available for 81.9% of patients, of which 24.1% were microsatellite unstable (MSI-high, defined as loss of expression of at least one mismatch repair protein or high expression of microsatellites by PCR). This proportion is higher than quoted in scientific literature (15%, Pearson's $\chi 2$ test, $p = 8.1e\text{-}06$) [71]. This discrepancy is most likely due to the recruitment of patients into the study who were having resections. These are more likely to have earlier stage disease, and MSI-high CRC is likely to be overrepresented in this cohort due to its generally more favourable prognosis [85]. Similarly, the overrepresentation of *BRAF*-mutated CRC in this dataset is most likely due to the association of the *BRAF V600E* mutation with sporadic MSI-high CRC, and possibly targeted BRAF testing in this cohort.

47.9% of patients had evidence of extramural venous invasion (EMVI). This was higher than figures reported in the literature, with an expected average of 30% [236, 237]. This is likely to reflect variations in the detection of venous invasion across different centres due to differences in the case mix, tissue sampling, use of special stains and the reporting pathologists' diligence [238]. EMVI under-reporting is more common outside specialist and research centres [239]. The mean number of lymph nodes retrieved per patient was 23.2, of which a mean of 1.6 nodes were positive. Other clinicopathological data is illustrated in Table 3.5.

**Table 3.5.** Clinico-pathological data available for research and national data sets

| Criterion | Research data | National data | p value |
|---|---|---|---|
| Age (years) | | | |
| - Range | 31 to 88 | 5 to 90+ | |
| - Median | 69 | n/a | n/a |
| - Mean | 68 | 69.4 | 0.031^* |
| Sex | | | |
| - M | 145 (60.9%) | 55.4% | |
| - F | 93 (39.1%) | 44.6% | 0.098$^{o}$ |
| Ethnicity (self-described) | | | |
| - Asian | 11 (4.6%) | 1.5% | |
| - Black | 6 (2.5%) | 1.1% | |
| - White | 209 (87.8%) | 92.4% | |
| - Other/unknown | 12 (5.0%) | 5.0% | 2.2e-13$^{o*}$ |
| Primary tumour location | | | |
| - Colon | 183 (76.9%) | 72.2% | |
| - Rectum | 55 (23.1%) | 27.8% | 0.123$^{o}$ |
| Pathological T stage | | | |
| - 1 | 6 (2.5%) | n/a | |
| - 2 | 32 (13.4%) | | |
| - 3 | 140 (58.8%) | | |
| - 4 | 60 (25.2%) | | |
| EMVI | | | |
| - Positive | 114 (47.9%) | Variable (~30%) | 0.014$^{o*}$ |
| - Negative | 124 (52.1%) | | |
| Disease stage | | | |
| - 1 | 27 (11.3%) | 16.6% | |
| - 2 | 99 (41.6%) | 23.1% | |
| - 3 | 95 (39.9%) | 27.3% | |
| - 4 | 17 (7.1%) | 23.3% | <2.2e-16$^{o*}$ |
| MMR status (available in 81.9%) | | | |
| - MSS | 148 (75.9%) | 85% | |
| - MSI-high | 47 (24.1%) | 15% | 8.1e-06$^{o*}$ |
| - n/a | 43 | | |
| *BRAF V600E* (available in 50.2%) | | | |
| - Mutant | 25.4% | 9.4% [8] | 0.005$^{o*}$ |
| *KRAS* (available in 82.4%) | | | |
| - Mutant | 36.7% | 42.0% [8] | 0.534$^{o}$ |

*BRAF* = v-raf murine sarcoma viral oncogene homolog B1. EMVI = extramural venous invasion. *KRAS* = Kirsten rat sarcoma virus oncogene. MMR = mismatch repair status. MSI-high = microsatellite instability high. MSS = microsatellite stable. n/a = not available. ^ = one-sample t-test. $^{o}$ = Pearson $\chi$-squared test. * = statistically significant p value.

Patients who had pre-operative radiotherapy to sites other than the primary tumour (for example, ablative radiotherapy for hepatic metastases), were included. 15 patients (6.3%) had neo-adjuvant therapy. Of these, 12 had chemotherapy only, and three (3) patients received chemotherapy and radiotherapy to metastases. There was no statistically significant difference in disease stage at diagnosis in patients who had neo-adjuvant treatment compared with those who did not (Wilcoxon rank sum test, $p = 0.1$).

109 (45.8%) patients had adjuvant treatment, of which the majority received chemotherapy only (n = 82, 34.5%). Patients who had adjuvant treatment were more likely to have advanced disease (Wilcoxon rank sum test, $p < 2.2e\text{-}16$).

Of the 17 patients with Stage 4 (metastatic) disease, one had MSI-high CRC. No patient received immunotherapy, and 9 patients received targeted therapy post-operatively (cetuximab or bevacizumab).

### 3.2.3 The Immunoscore

The Immunoscore was completed for 197 specimens (82.8%). Samples were excluded if they failed QC checks. 20 samples from the pilot set, and 21 from the 100KGP were excluded for this reason. The reasons for sample exclusion were predominantly due to poor cell detection (n=20) or lack of tumour detection (n=11) (Table 3.6).

**Table 3.6.** Reasons for sample exclusion from Immunoscore analysis

| | **Number of samples excluded** |
|---|---|
| QC failed – bad cell detection | 20 |
| QC failed – no invasive margin detected | 4 |
| QC failed – no centre of the tumour | 1 |
| QC failed – low staining intensity | 2 |
| Invasive margin too small | 3 |
| No invasive tumour detected | 11 |
| Total | **41** |

QC = quality control.

The exclusion of these samples leads to the possibility of bias in patient outcomes. Those samples with no tumour may represent patients with earlier stage or better prognosis disease, and those with poor cell detection may have represented those with necrotic or more poorly differentiated tumours and therefore likely poorer prognosis disease [240]. There was also the possibility that samples with failed Immunoscore could be overrepresented by patients who had a pathological response to neoadjuvant therapy. However, only three patients of the 41 excluded had neo-adjuvant therapy (7.1%). A comparison of demographic and clinico-pathological analysis in the total sample set when compared with those in which the Immunoscore was completed showed no significant differences between the data sets (Table 3.7).

**Table 3.7.** Comparison of clinico-pathological markers before and after sample exclusion for the Immunoscore

100KGP = 100 000 Genomes Project participants. EMVI = extramural venous invasion.

| Criterion | Full data set (n = 238) | Immunoscore complete (n = 197) | 100KGP set (n = 168) | p value |
|---|---|---|---|---|
| **Age (years)** | | | | |
| - Median | 69 | 69 | 69 | 0.973^ |
| - Mean | 68 | 67 | 68 | |
| **Sex (%)** | | | | |
| - M | 145 (60.9) | 119 (60.4) | 102 (60.7) | 0.994 º |
| **Ethnicity (n/%)** | | | | |
| - White | 209 (87.8) | 173 (87.8) | 146 (86.9) | 0.955 º |
| **Disease stage (n/%)** | 221 (92.9) | 183 (92.9) | 156 (92.9) | 0.98 º |
| - I-III | | | | |
| **Primary tumour side (n/%)** | 119 (50.0) | 96 (48.7) | 76 (45.2) | 0.911 º |
| - Left | 117 (49.2) | 99 (50.3) | 90 (53.6) | |
| - Right | | | | |
| **Primary tumour location** | 55 (23.8) | 43 (21.8) | 35 (20.8) | 0.858 º |
| - Rectum | | | | |
| **EMVI (n/%)** | | | | |
| - Positive | 114 (47.9) | 96 (48.7) | 85 (50.6) | 0.939º |
| **Nodes** | | | | |
| - Total | 23.2 | 23.3 | 22.6 | 0.805^ |
| - Positive | 1.6 | 1.7 | 1.6 | 0.927^ |
| **Microsatellite status (n/%)** | | | | |
| - MSI-high | 47 (19.7) | 39 (19.8) | 35 (20.8) | 0.064º |
| - N/A | 43 (18.1) | 28 (14.2) | 13 (7.7) | |
| **Neo-adjuvant treatment** | | | | |
| - Yes | 15 | 12 | 11 | 0.984º |
| - No | 223 | 185 | 157 | |

MSI-MSI-high = microsatellite instability high. ^ = Kruskal-Wallis test. º = Pearson $\chi$-squared test. n = number.

51.8% of patients had an Intermediate Immunoscore (IS2) (Figure 3.4). This Immunoscore distribution was very similar to that reported by Pages *et al.* [118] in the international validation study (Table 3.8) (Pearson's $\chi^2$ test, p = 0.536).

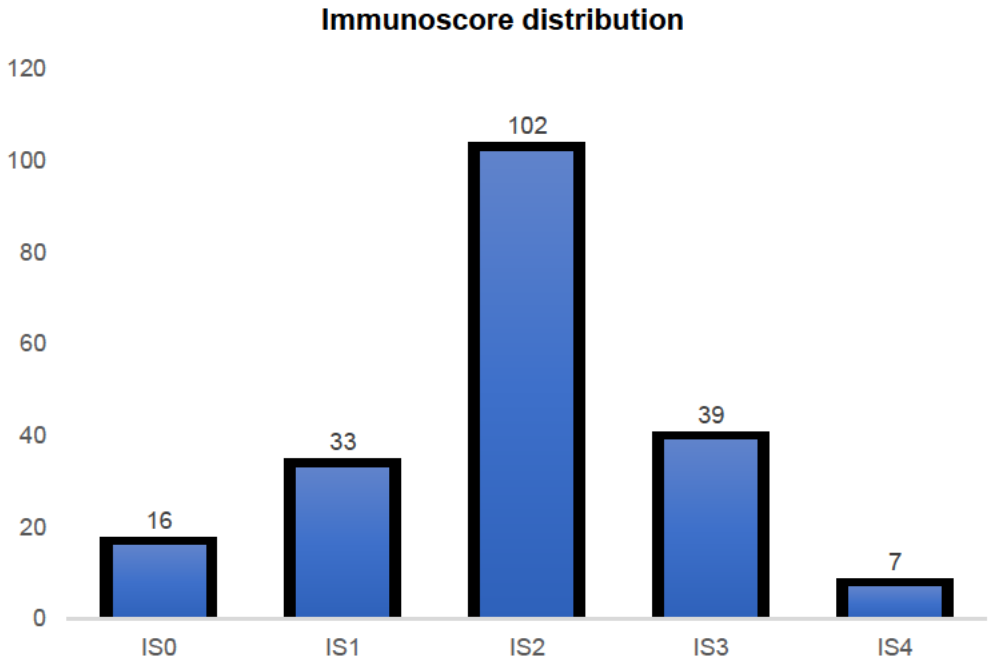



**Figure 3.4.** The Immunoscore distribution. The numbers of samples in each category are illustrated. The highest number of patients were IS2 and the lowest number were IS4. IS = Immunoscore.

**Table 3.8.** Immunoscore distribution in research data set compared with the Pages *et al.* [118] data set

| | IS Low (0,1) | IS Int (2) | IS High (3,4) |
|---|---|---|---|
| Research data (n (%) | 49 (24.9) | 102 (51.8) | 46 (23.4) |
| International validation training set [118] (n/%) | 155 (22) | 357 (51) | 188 (27) |

No significant different is observed between the proportions in each Immunoscore category. Pearson's $\chi^2$ test, p =0.536. n=number. IS=Immunoscore. Int=Intermediate. n = number.

Analysis of the associations between clinico-pathological factors and the Immunoscore showed no associations between age (Figure 3.5), sex (Figure 3.6), primary tumour side (Figure 3.7), EMVI (Figure 3.8) and the Immunoscore. Patients with lower tumour T stage had higher Immunoscores (Pearson's $\chi^2$ test, p= 0.0003, Figure 3.9). There was a trend towards patients with lower disease stage having higher Immunoscores, but this was not statistically significant (Pearson's $\chi^2$ test, p= 0.096, Figure 3.10). Patients with MSI-high CRC had higher Immunoscores (Pearson's $\chi^2$ test p = 0.016, Figure 3.11).

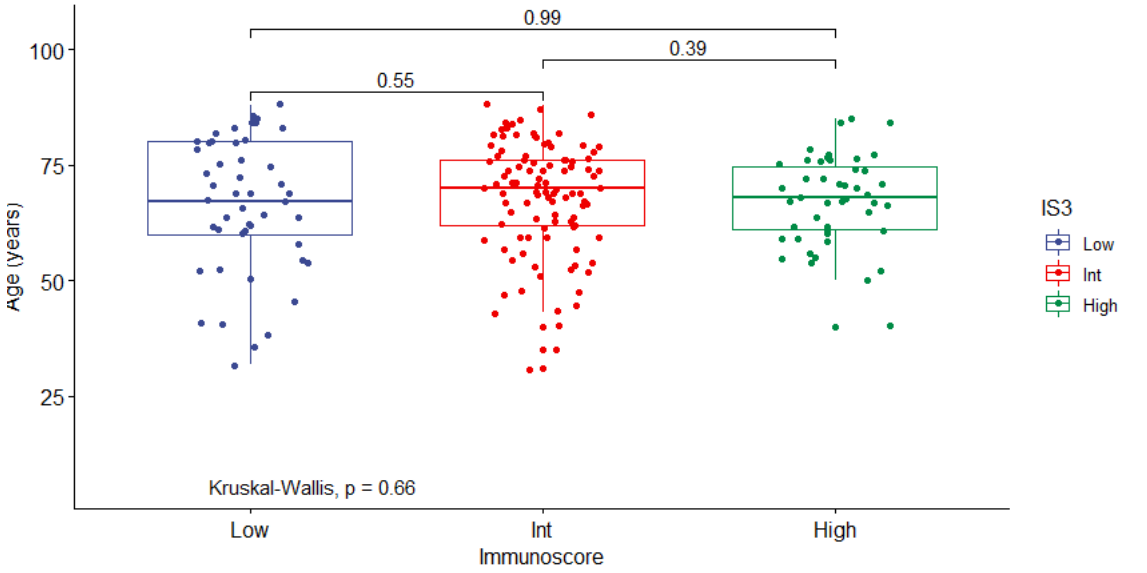



**Figure 3.5.** Boxplots illustrating the associations between age and the Immunoscore (Low, Int, High). There is no significant association between patient age at surgery and the Immunoscore. Kruskal-Wallis test, p = 0.66. Int = Intermediate.

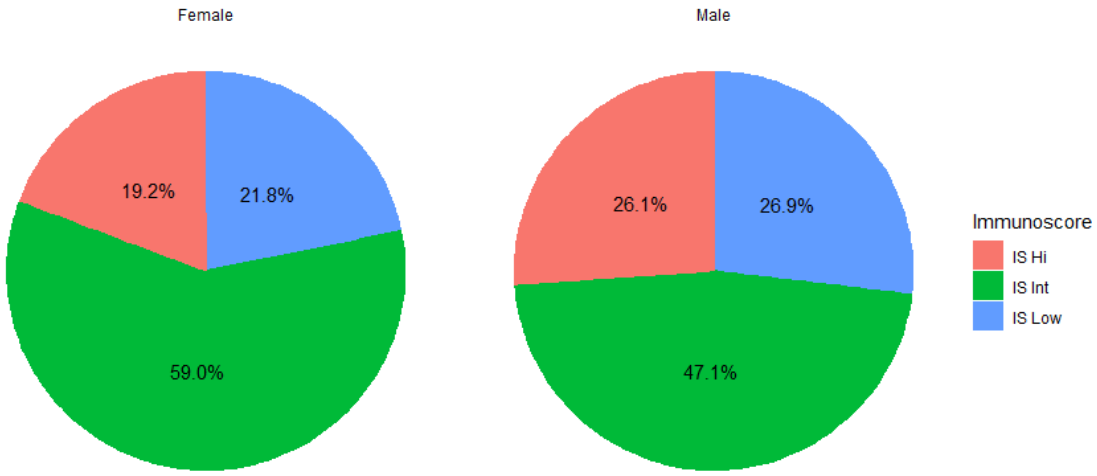



**Figure 3.6.** Pie charts illustrating the association between sex and the Immunoscore. There is no significant association between sex (female or male) and the Immunoscore. Pearson's $\chi^2$ test, p = 0.232. IS=Immunoscore. Int=Intermediate.

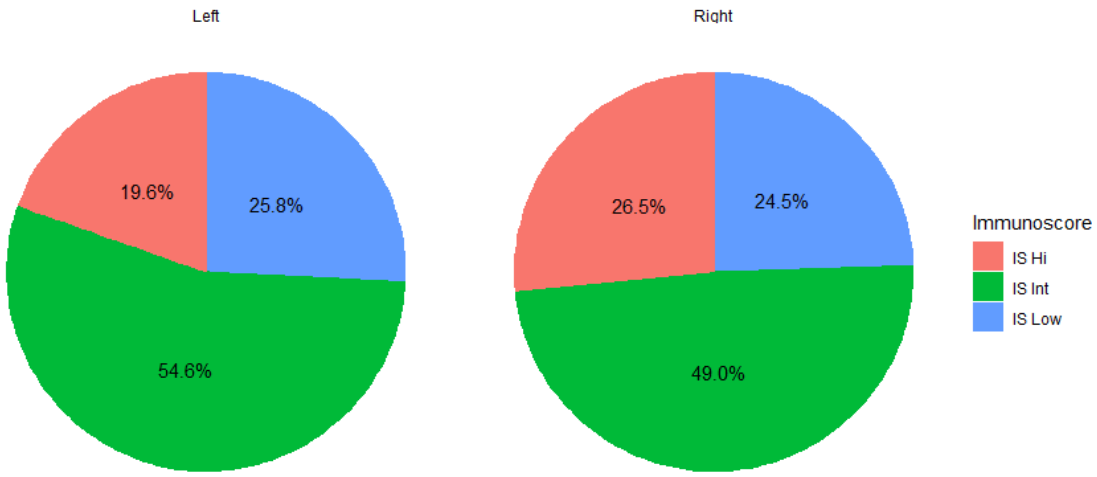



**Figure 3.7.** Pie charts illustrating the association between primary tumour side (left and right) and the Immunoscore. There is no significant association between primary tumour side and the Immunoscore. Pearson's $\chi^2$ test, p = 0.504. IS=Immunoscore. Int=Intermediate.

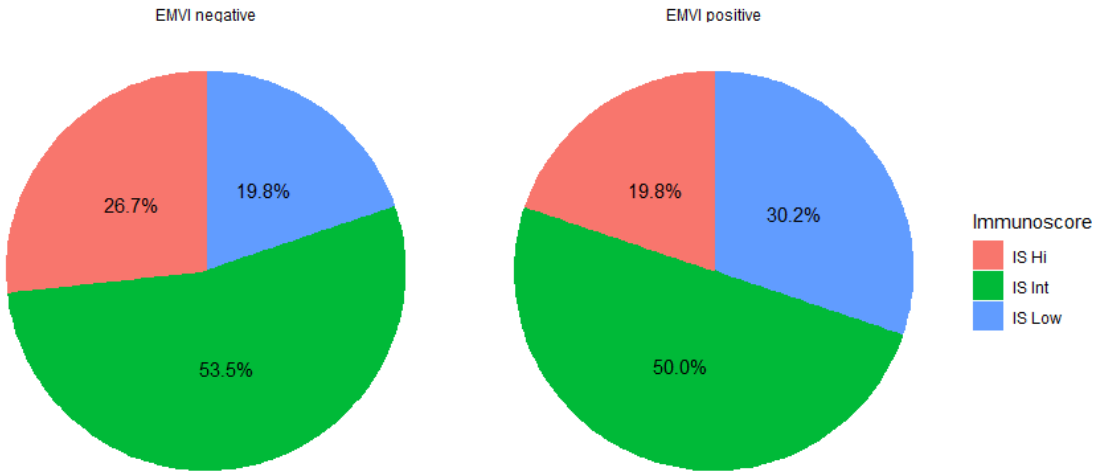



**Figure 3.8.** Pie charts illustrating the association between extramural venous invasion and the Immunoscore. There is no significant association between EMVI and the Immunoscore. Pearson's $\chi^2$ test, p = 0.192. EMVI = extramural venous invasion. IS=Immunoscore. Int=Intermediate.

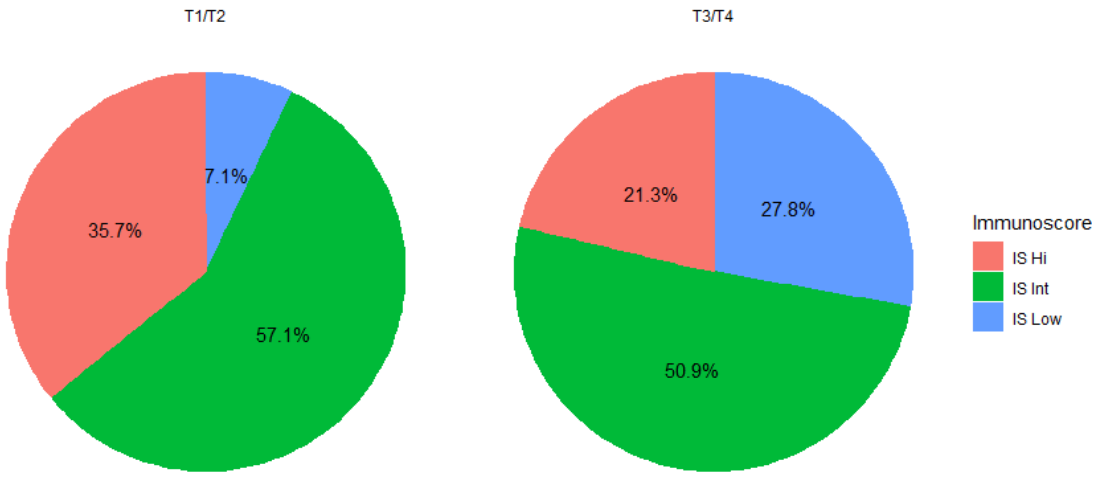



**Figure 3.9.** Pie charts illustrating the association between tumour T stage (T1/T2 and T3/T4) and the Immunoscore. Patients with lower tumour T stage (T1/T2) had higher Immunoscores. Pearson's $\chi^2$ test, p = 0.0003. IS=Immunoscore. Int=Intermediate.

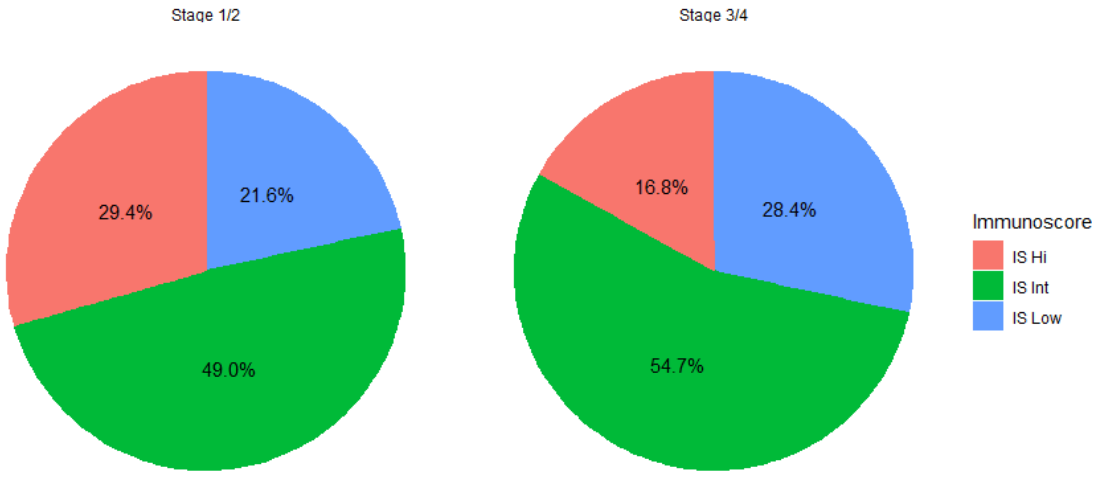



**Figure 3.10.** Pie charts illustrating the association between disease stage (1/2 and 3/4) and the Immunoscore. Patients with lower disease stage (1/2) had higher Immunoscores. Pearson's $\chi^2$ test, p = 0.097. IS=Immunoscore. Int=Intermediate.

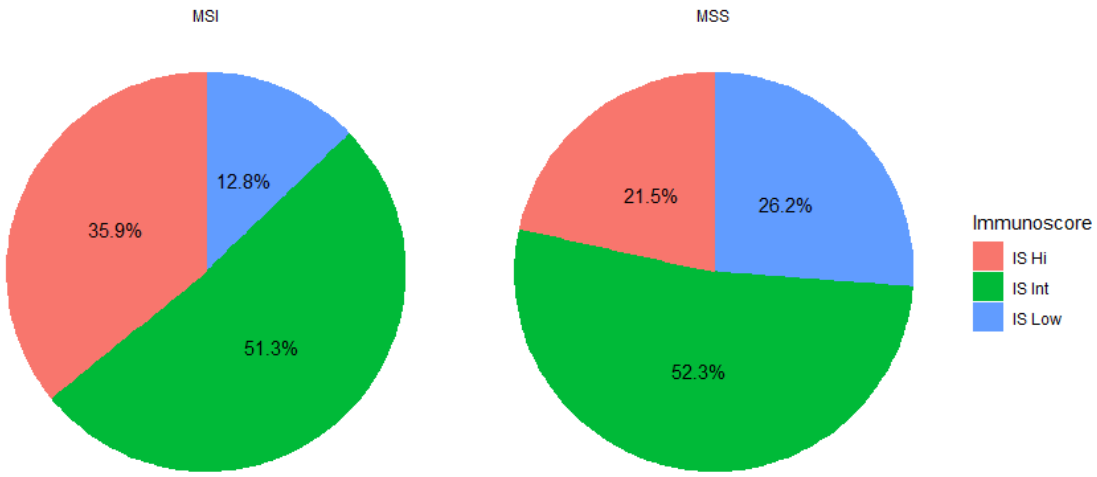



**Figure 3.11.** Pie charts illustrating the association between microsatellite status and the Immunoscore. Patients with MSI-high CRC had higher Immunoscores. Pearson's $\chi^2$ test, p = 0.016. CRC = colorectal cancer. IS=Immunoscore. Int=Intermediate. MSI = microsatellite instability-high. MSS = microsatellite stable.

### 3.2.4. Survival analyses

#### *3.2.4.1. Clinico-pathological markers and survival*

Survival analyses were performed on patients for whom the Immunoscore was successfully performed. The pilot data set consisted of patients who had surgery between October 2010 and October 2013, and the 100KGP data set had surgery between September 2011 and September 2018. Clinical data was not available for 1 patient, who was excluded from further analysis. Clinical follow-up for the data set was completed in April 2021, and thus the shortest duration of follow-up was 32.1 months (2.7 years). The median follow-up period for the pilot set was 113.8 months, and for the combined data set was 56.6 months. The median recurrence-free survival (RFS) was 46.5 months and median overall survival (OS) was 49.6 months.

Associations between clinical and pathological markers and OS and RFS were determined. OS and RFS were strongly correlated with disease stage (Figure 3.12, Figure 3.13), tumour T stage (Figure 3.14, Figure 3.15), and the presence of EMVI (Figure 3.16, Figure 3.17). OS, but not RFS was strongly correlated with age (Figure 3.18, Figure 3.19). RFS, but not OS was correlated with mismatch repair status (Figure 3.20, Figure 3.21), and ethnicity (with Black patients showing a trend towards lower RFS than other ethnic groups, hazard ratio (HR) = 3.9, 95% confidence interval (CI) = 0.9 – 16.4, p = 0.063, Figure 3.22, Figure 3.23). There was no association with either OS or RFS with anatomical sex (Figure 3.24) or disease side (Figure 3.25).

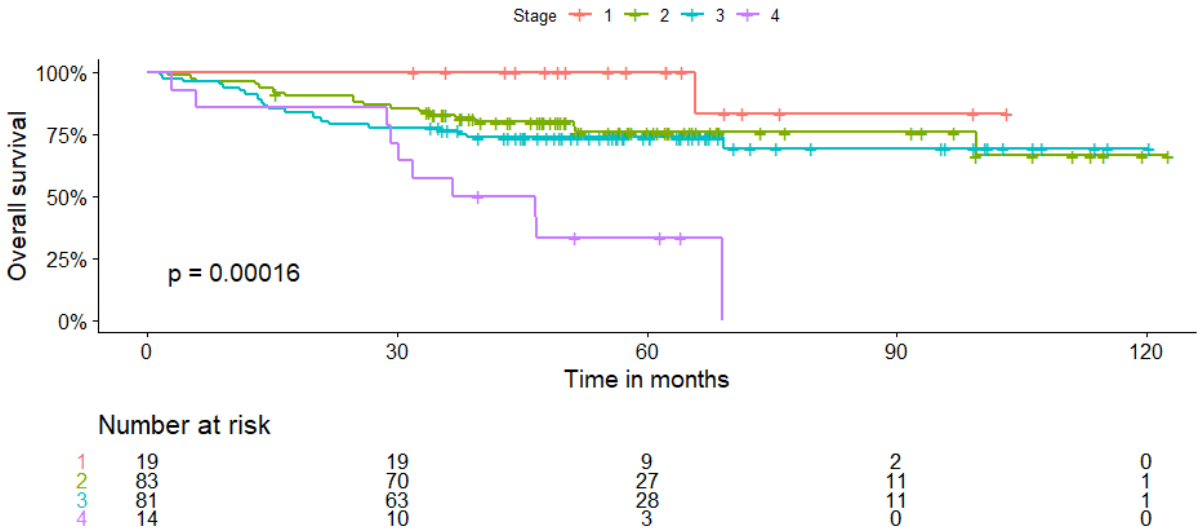



**Figure 3.12.** Kaplan-Meier estimate of overall survival (OS) stratified by disease stage for all patients. OS decreases with increasing disease stage. Hazard ratio Stage 1 vs Stage 4 disease = 1.9 (95% CI 1.3 – 2.8, p = 0.0004), Stage 1 vs Stage 3 disease = 1.6 (95% CI 0.96 – 2.6, p = 0.072).

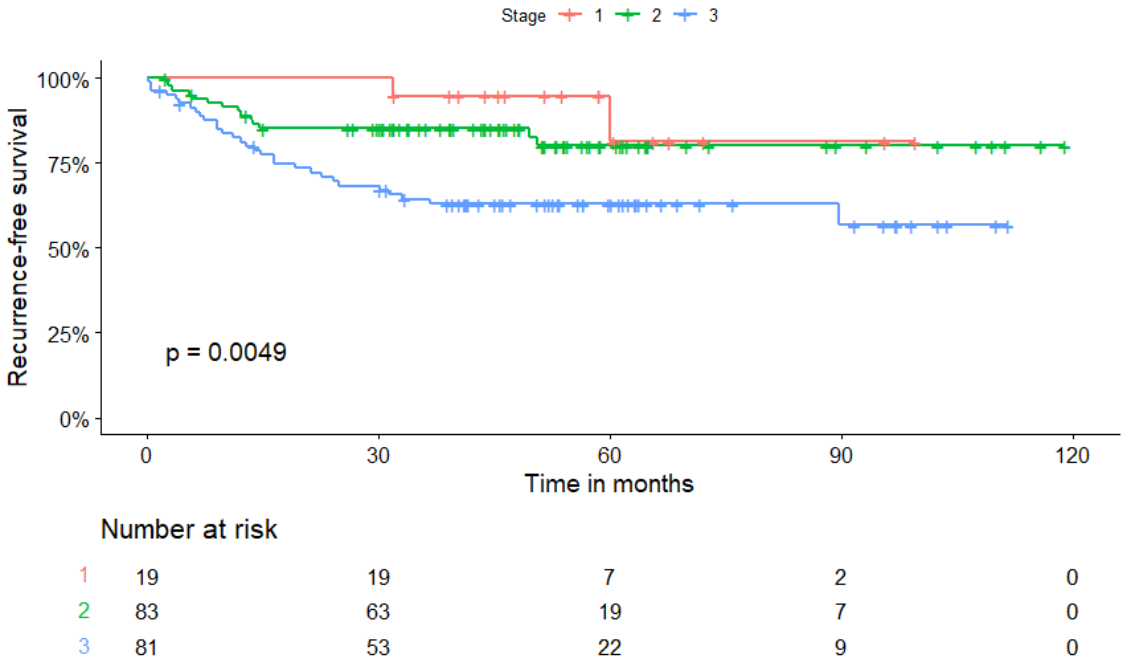



**Figure 3.13.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by disease stage for all patients. RFS decreases with increasing disease stage. Hazard ratio Stage 1 vs Stage 3 disease = 2.2 (95% CI 1.3 – 3.7, p = 0.001).

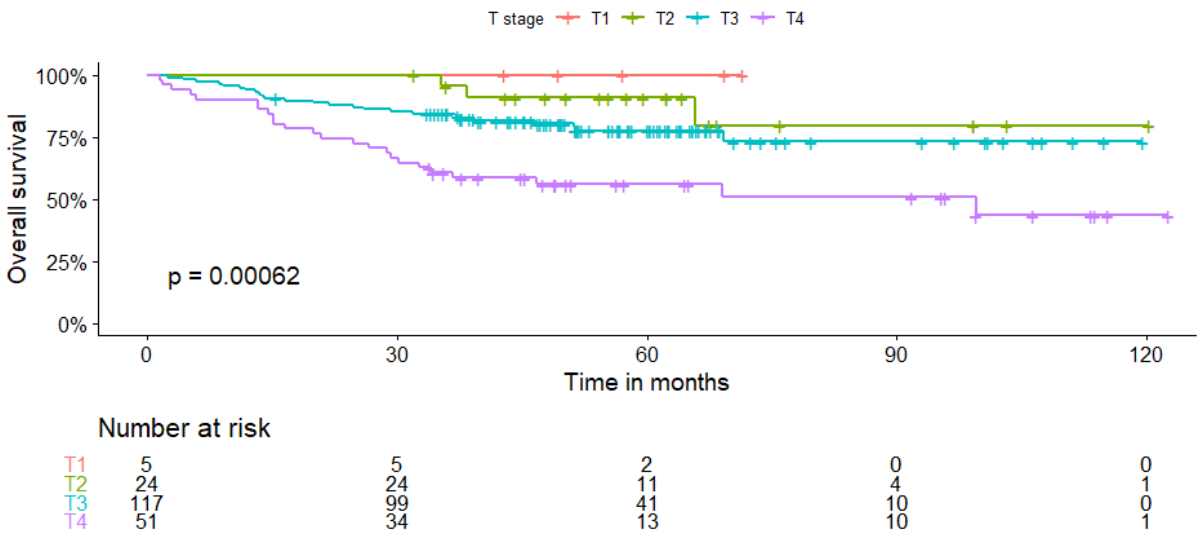



**Figure 3.14.** Kaplan-Meier estimate of overall survival (OS) stratified by pathological tumour stage for all patients. OS decreases with increasing T stage. Hazard ratio T1 vs T4 disease = 2.4 (95% CI 1.6 – 3.9, p = 0.0001).

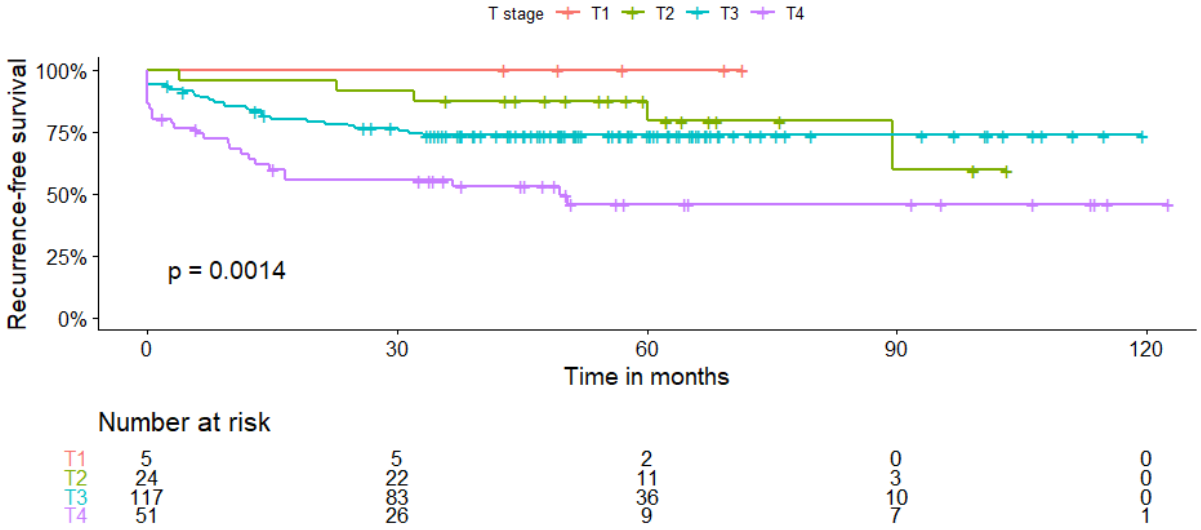



**Figure 3.15.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by pathological tumour stage for all patients. RFS decreases with increasing T stage. Hazard ratio T1 vs T4 disease = 2.1 (95% CI 1.4 – 3.3, p = 0.0003).

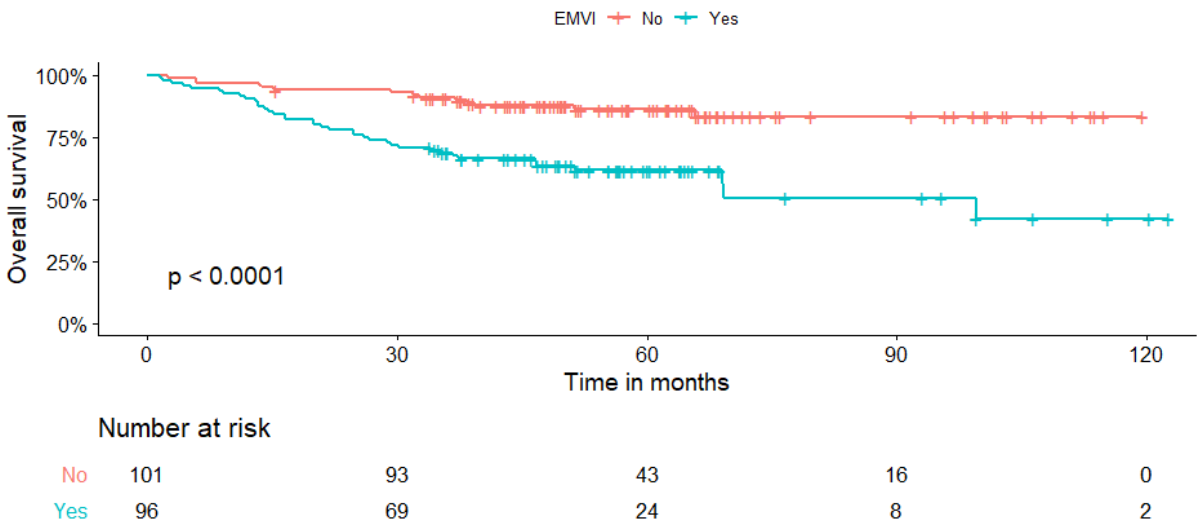



**Figure 3.16.** Kaplan-Meier estimate of overall survival (OS) stratified by the presence or absence of extramural venous invasion (EMVI). OS decreases where EMVI is present. Hazard ratio for EMVI positive disease = 3.5 (95% CI 1.9 – 6.5, p < 0.0001).

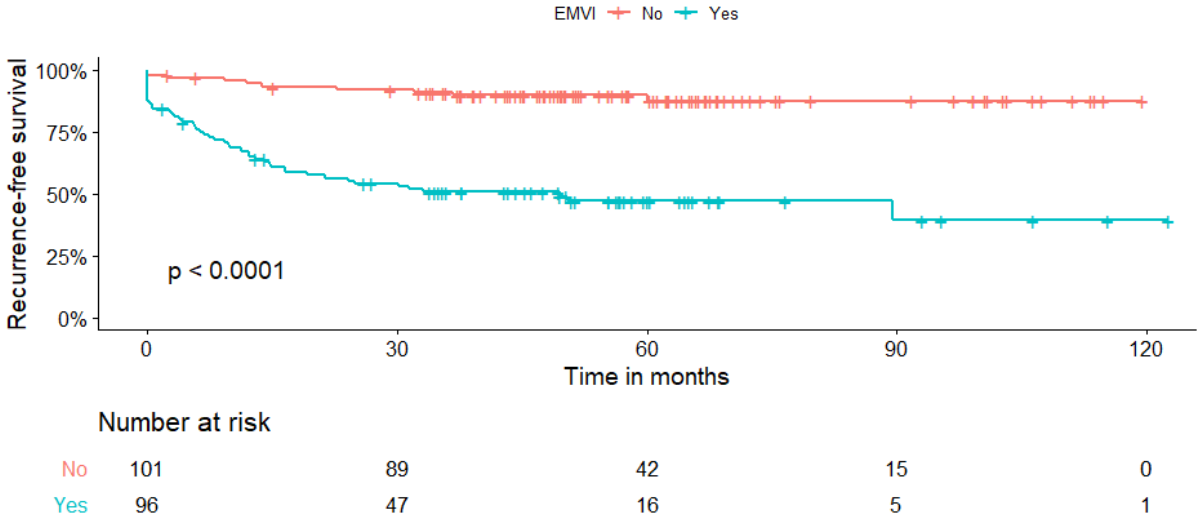



**Figure 3.17.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by the presence or absence of extramural venous invasion (EMVI). RFS decreases where EMVI is present. Hazard ratio for EMVI positive disease = 6.5 (95% CI 3.4 – 12.5, p < 0.0001).

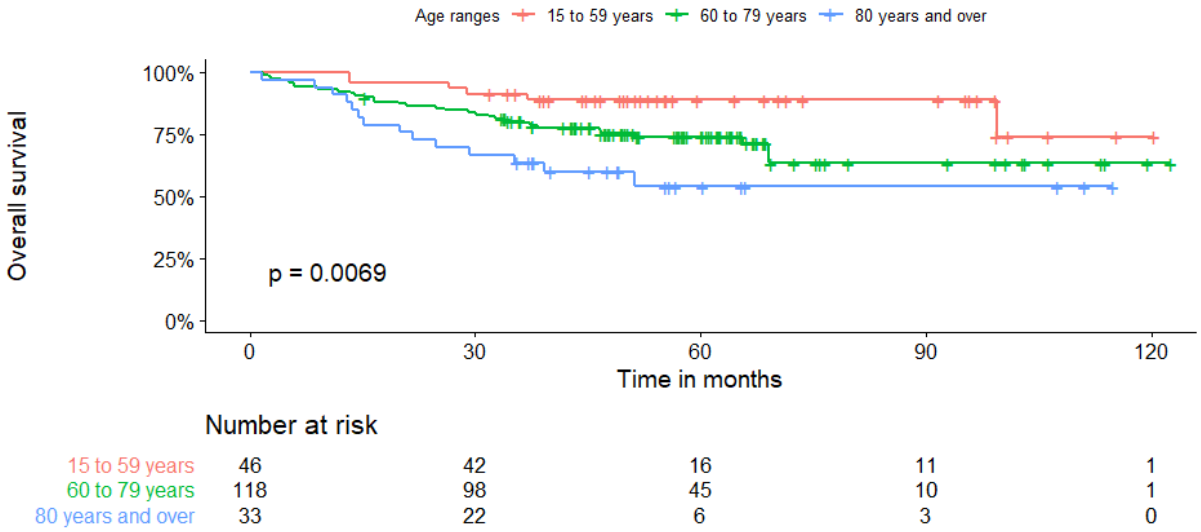



**Figure 3.18.** Kaplan-Meier estimate of overall survival (OS) stratified by age in three categories. OS is greatest in the lowest age categories. Hazard ratio 80 years and over compared with 15 to 59 years = 4.3 (95% CI = 1.6 – 11.1, p = 0.003).

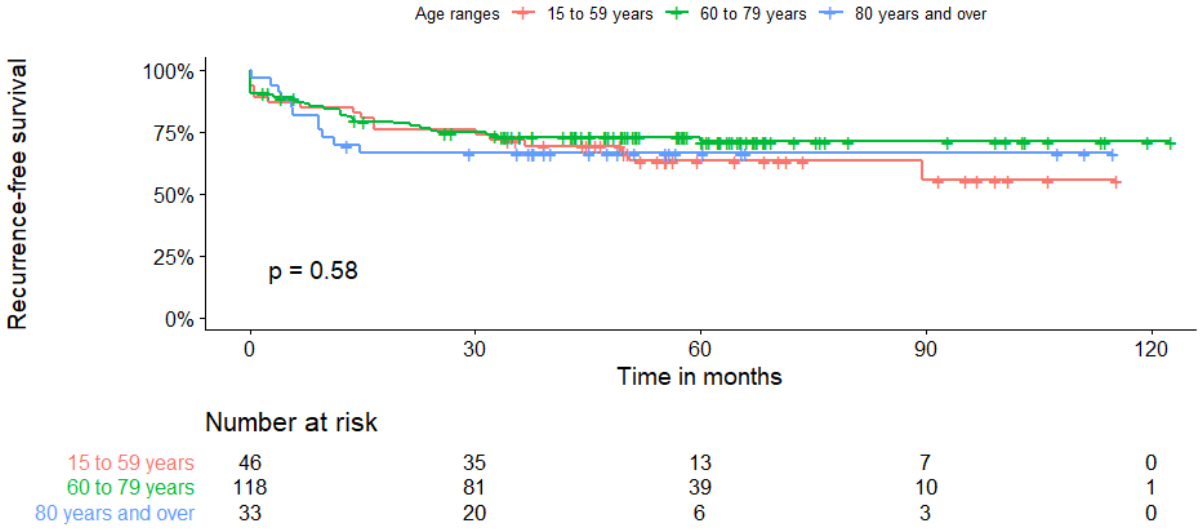



**Figure 3.19.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by age in three categories. No difference is observed across the three age categories.

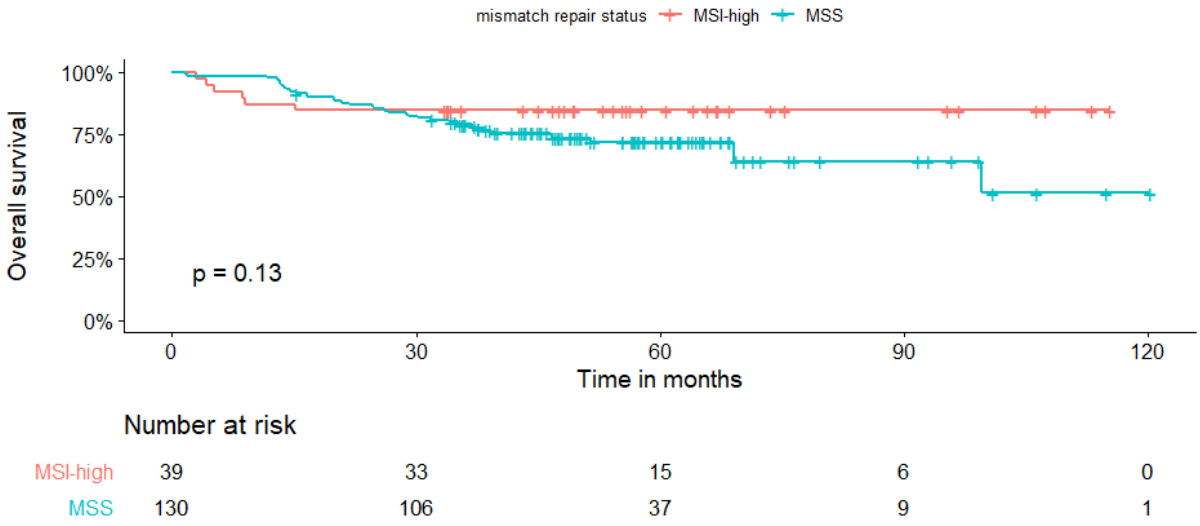



**Figure 3.20.** Kaplan-Meier estimate of overall survival (OS) stratified by DNA mismatch repair status, where information available. There is no difference in OS between the two groups, p = 0.13. MSI-high = microsatellite unstable. MSS = microsatellite stable.

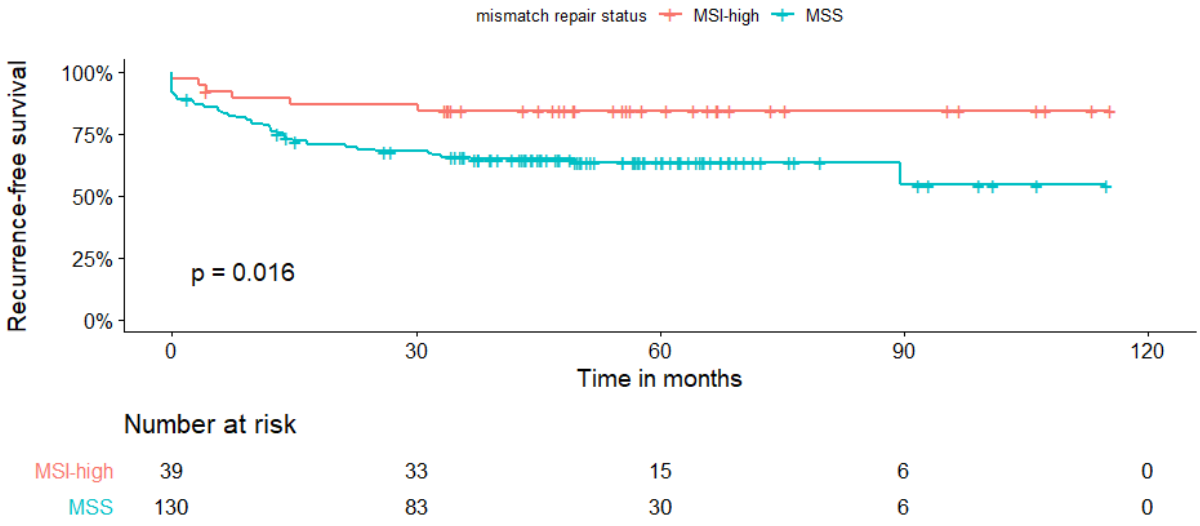



**Figure 3.21.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by DNA mismatch repair status, where information available. RFS is increased patients with MSI-high colorectal cancer. Hazard ratio MSS CRC 2.7 (95% CI = 1.2 – 6.4, p = 0.021). MSI-high = microsatellite unstable. MSS = microsatellite stable.

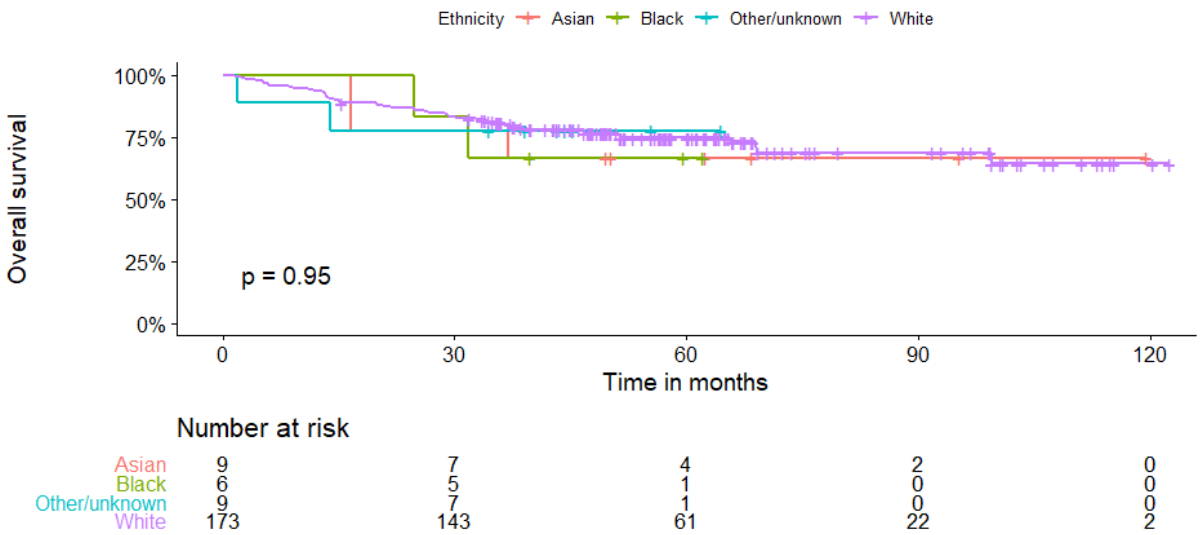



**Figure 3.22.** Kaplan-Meier estimate of overall survival (OS) stratified by ethnicity. There are no significant differences in OS between ethnic groups.

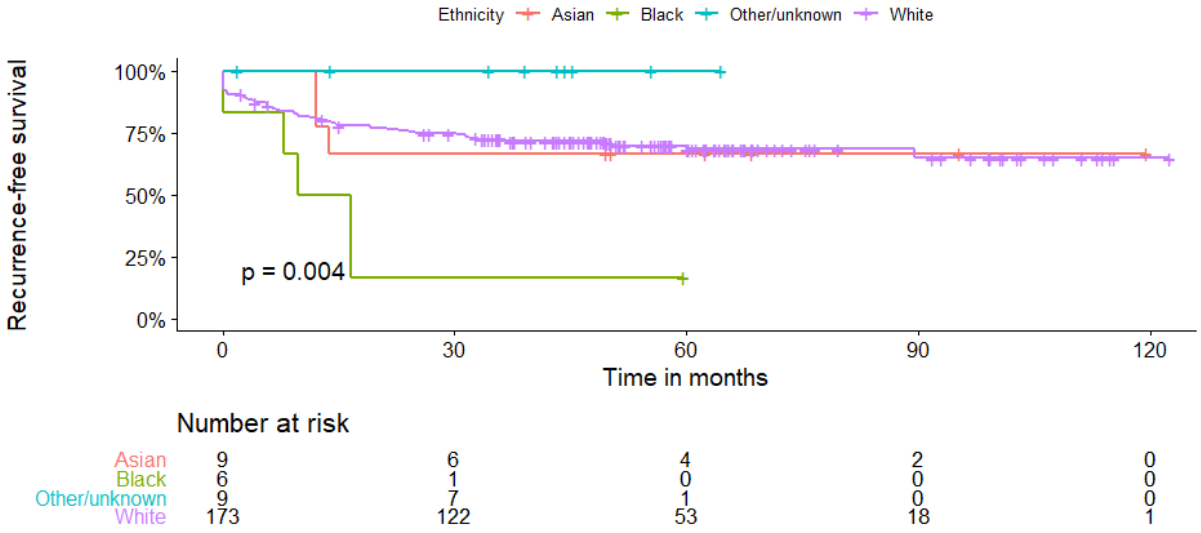



**Figure 3.23.** Kaplan-Meier estimate of recurrence-free survival (RFS) by ethnicity. RFS trends towards lower for patients identified as Black than other ethnic groups, hazard ratio = 3.9, 95% CI = 0.9 – 16.4, p = 0.063.

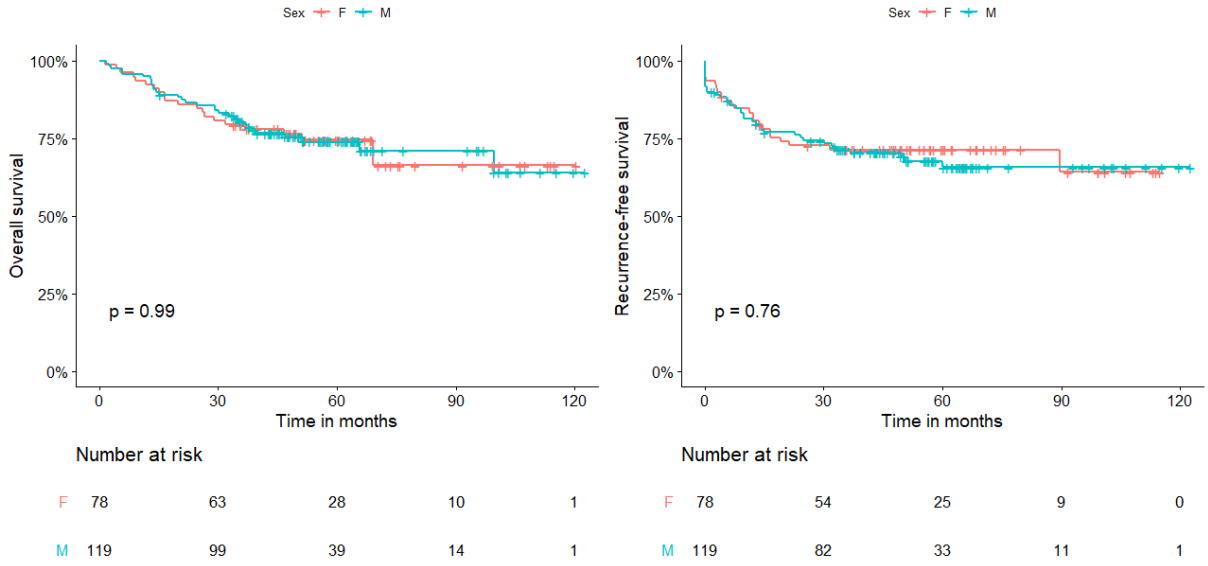



**Figure 3.24.** Kaplan-Meier estimate of overall survival (OS) and recurrence-free survival (RFS) stratified by sex. There is no difference in OS or RFS between male and female patients.

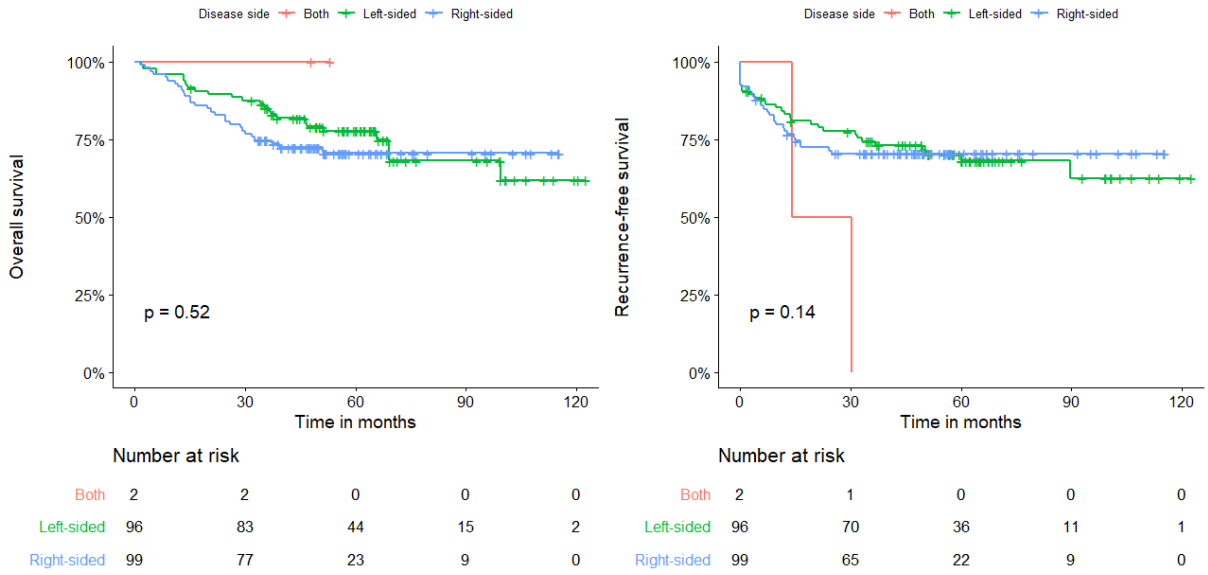



**Figure 3.25.** Kaplan-Meier estimate of overall survival (OS) and recurrence-free survival (RFS) stratified by side of primary tumour. There is no difference in OS or RFS between groups.

#### *3.2.4.2. Adjuvant treatment and survival*

Those who had adjuvant treatment had reduced RFS compared with those who did not (Figure 3.26, Figure 3.27). However, as expected, patients who had adjuvant therapy had higher stage disease (median Stage 3 compared with Stage 2 in those who had no adjuvant treatment, Wilcoxon test, p <2.2e-16). However, when the data was further stratified, patients with Stage 3 (that is, locally advanced disease) had increased OS and a trend to increased RFS with adjuvant treatment (Figure 3.28, Figure 3.29). Only 6.3% of patients had neo-adjuvant treatment. No difference in OS or RFS was seen with neo-adjuvant treatment (Figure 3.30). There appeared to be a trend towards lower RFS in those who received neo-adjuvant treatment. This could represent a higher disease stage in these patients. However, when compared, there was not a statistically significant difference in disease stage between the two groups (Wilcoxon test, p = 0.25). This is likely because only a small proportion received neo-adjuvant treatment (n = 12, 6.1%), with insufficient statistical power to detect a difference between the groups.

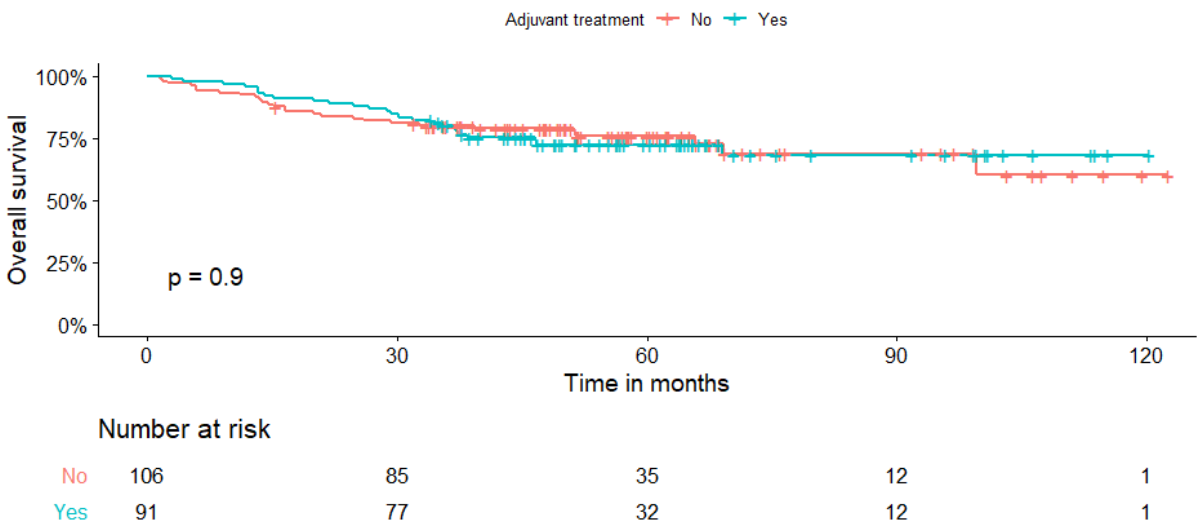



**Figure 3.26.** Kaplan-Meier estimate of overall survival (OS) stratified by adjuvant treatment. No difference in OS is observed between those who received and those who did not receive adjuvant treatment.

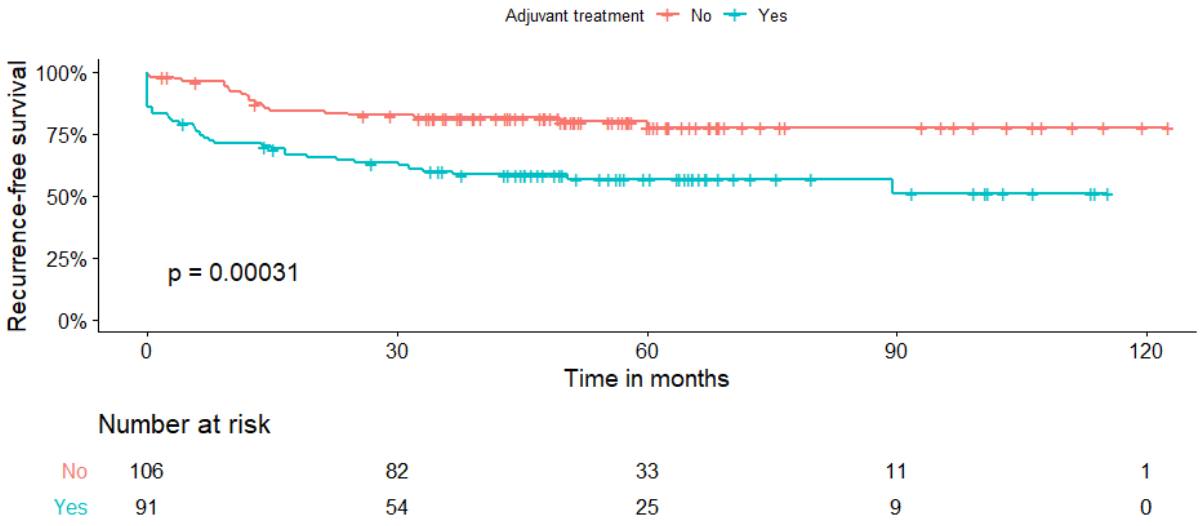



**Figure 3.27.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by adjuvant treatment. RFS is lower in the group who received adjuvant treatment.

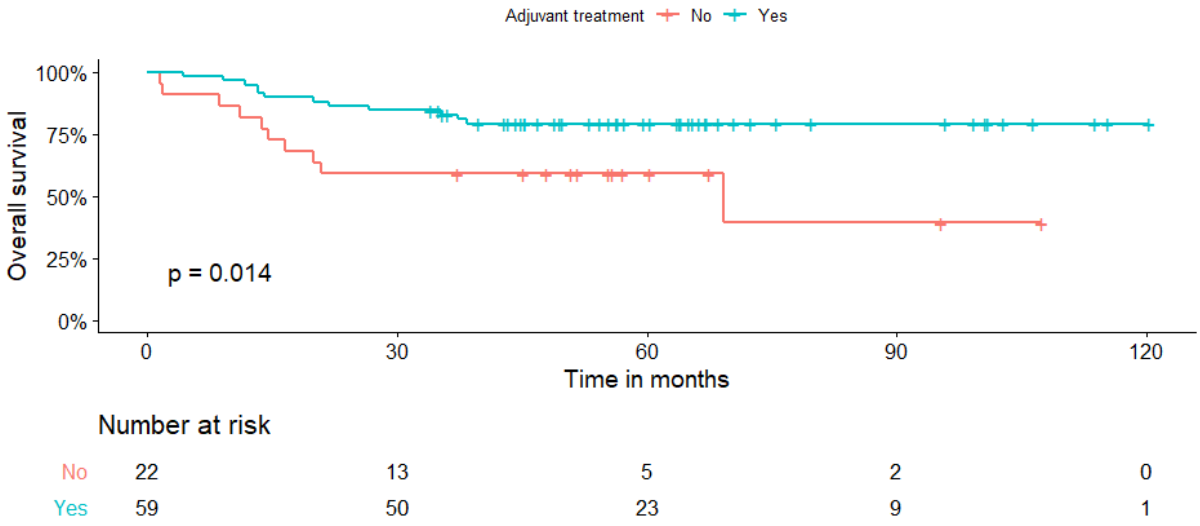



**Figure 3.28.** Kaplan-Meier estimate of overall survival (OS) stratified by adjuvant treatment in patients with Stage 3 disease. OS is higher in the group who received adjuvant treatment.

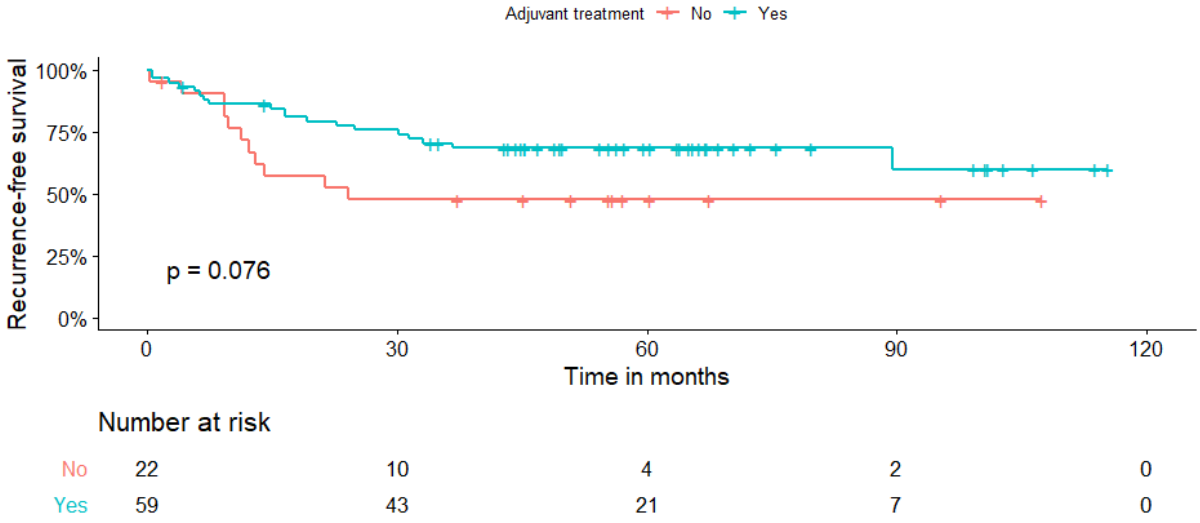



**Figure 3.29.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by adjuvant treatment in patients with Stage 3 disease. There is a trend towards higher RFS in the group who received adjuvant treatment.

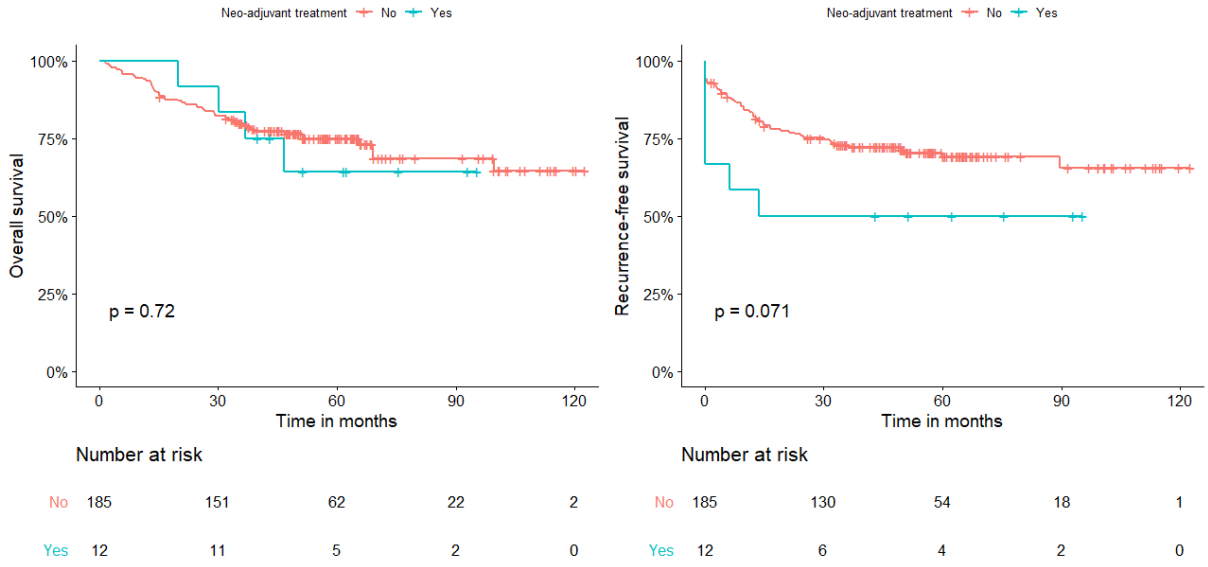



**Figure 3.30.** Kaplan-Meier estimate of overall survival (OS) and recurrence-free survival (RFS) stratified by neo-adjuvant treatment. There is no difference in OS and RFS between those who received and those who did not receive neo-adjuvant treatment.

#### *3.2.4.3. The Immunoscore and survival*

In line with data from Pages et al. [118], the Immunoscore was strongly prognostic of RFS, but not OS (Figure 3.31, Figure 3.32). The hazard ratio for Low vs High Immunoscore was 5.7 (95% CI 2.0 – 12.1, p = 0.0005). This difference in prognostic value remained when stratified by patients with Disease Stage 1 to 3 (that is, excluding those with metastatic disease) (Figure 3.33) and in patients with MSS CRC (Figure 3.34).

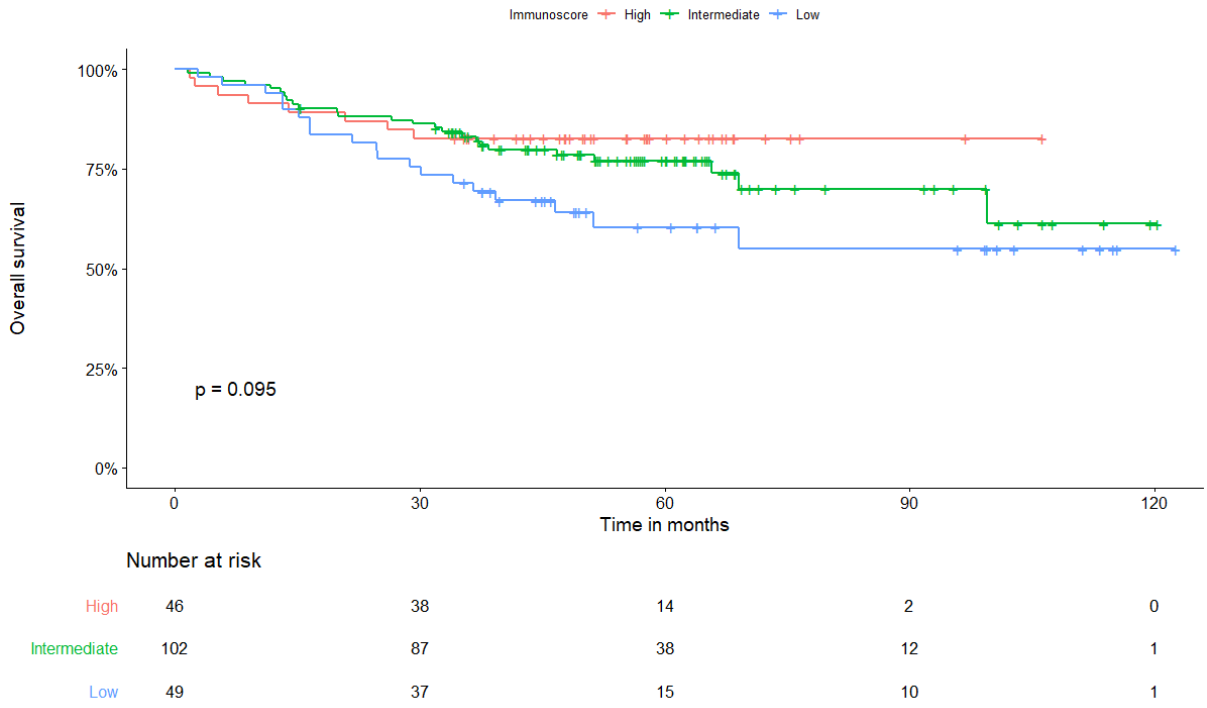



**Figure 3.31.** Kaplan-Meier estimate of overall survival (OS) stratified by Immunoscore treatment. Although there is a trend to increasing OS with Immunoscore "High", this is not statistically significant. p = 0.095. Int = Intermediate.

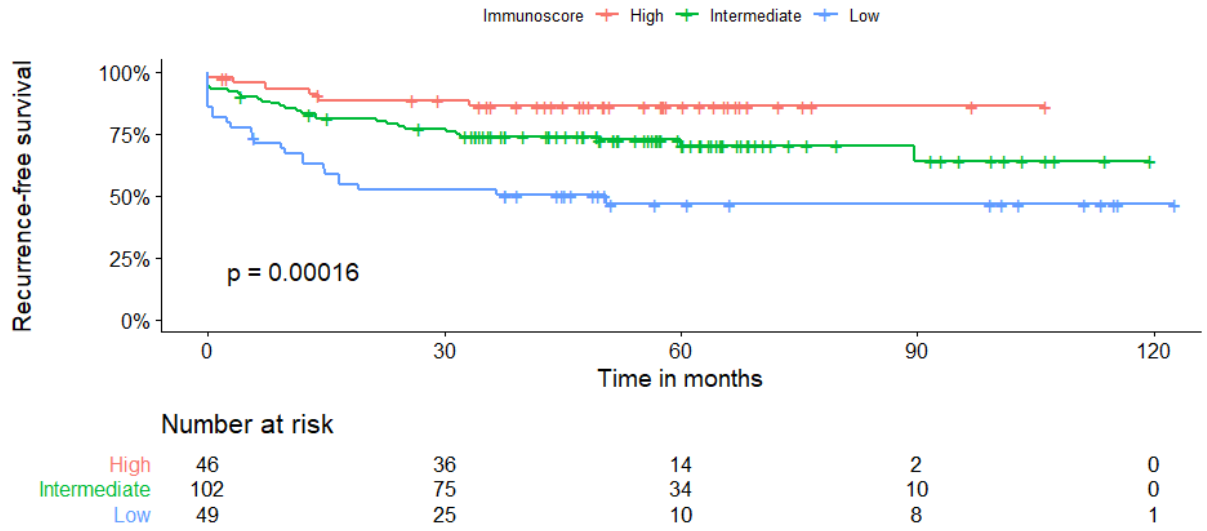



**Figure 3.32.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by Immunoscore. There is a clear association between the Immunoscore and RFS. Hazard ratio Low vs High Immunoscore = 4.9 (95% CI 2.0 – 12.1, p = 0.0005).

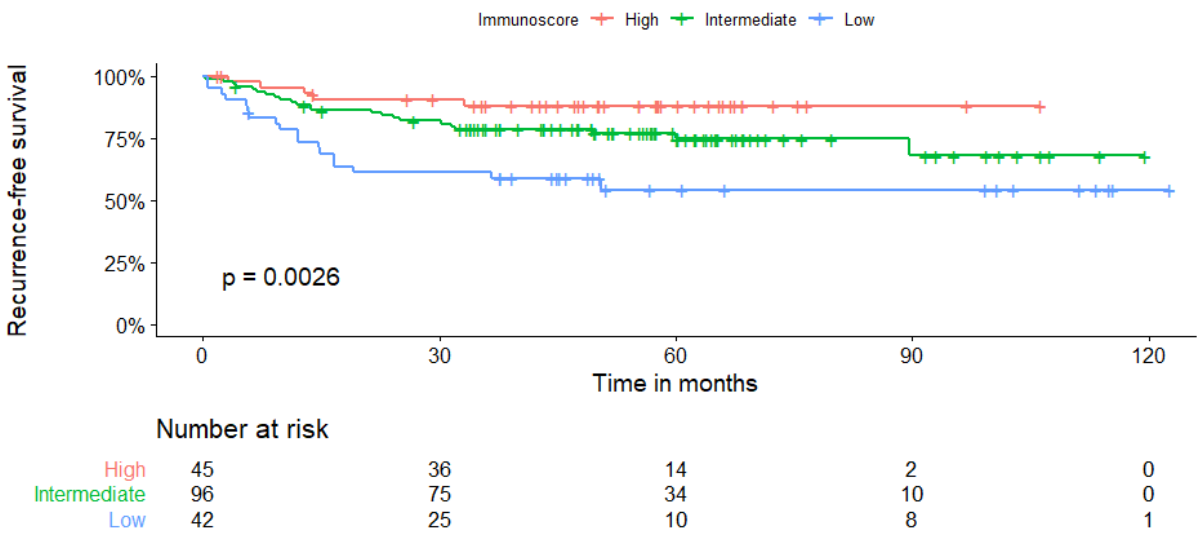



**Figure 3.33.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by Immunoscore, patients with Disease Stage 1 to 3. There is a clear association between the Immunoscore and RFS. Hazard ratio Low vs High Immunoscore = 4.5, 95% CI =1.7 to 12.1, p = 0.003).

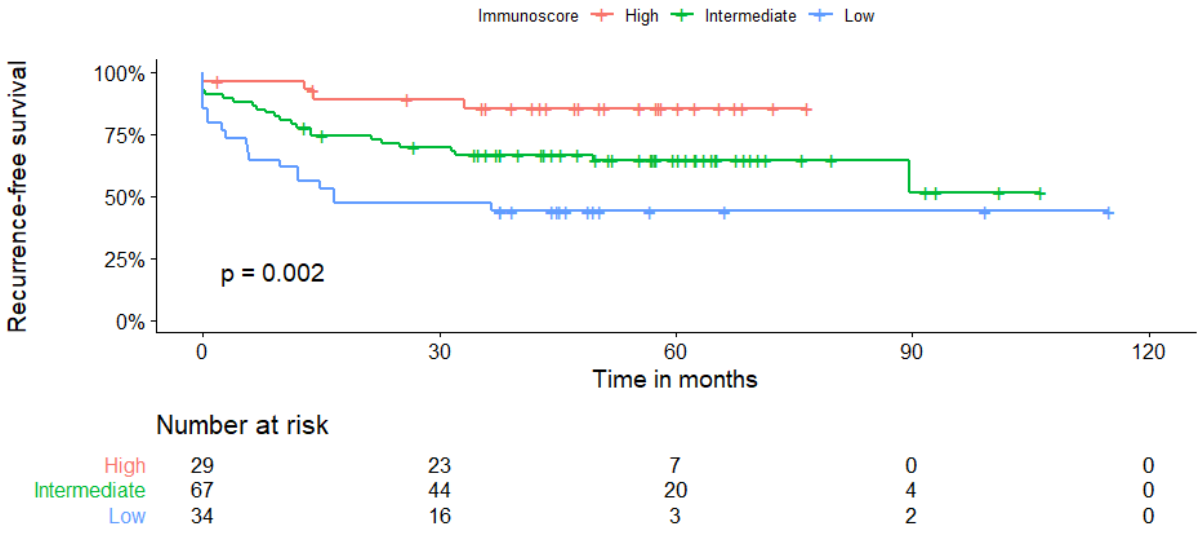



**Figure 3.34.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by Immunoscore, patients with microsatellite stable colorectal cancer only (n = 130). The clear association between the Immunoscore and RFS persists.

Combining the effects of mismatch repair status and the Immunoscore showed clear differences in RFS, with patients with MSS/Low Immunoscore having the lowest RFS, and those with MSI/High Immunoscore having the highest RFS (Figure 3.35).

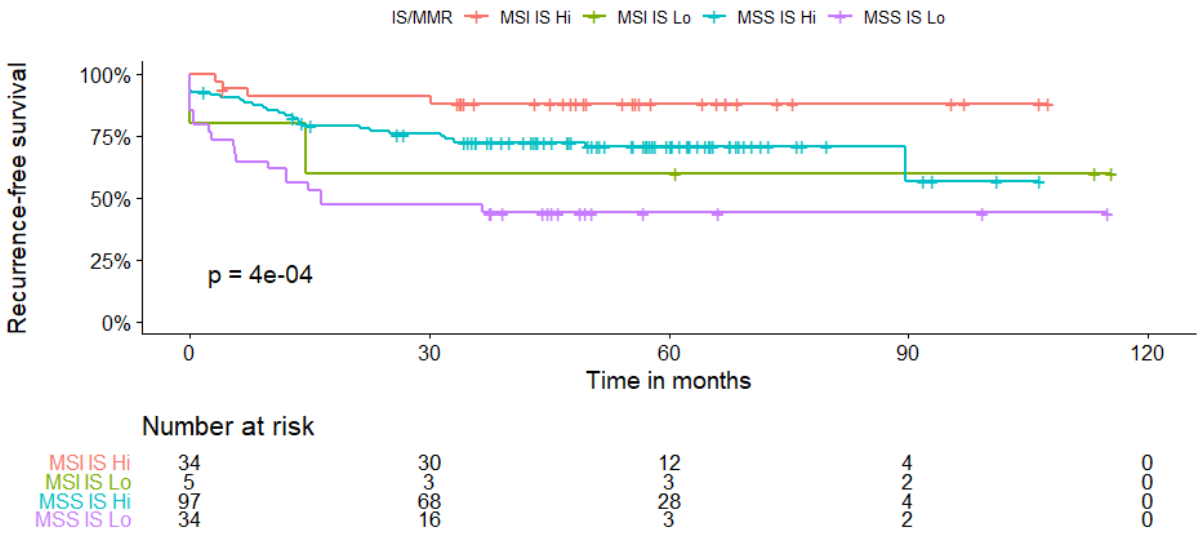



**Figure 3.35.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by Immunoscore. Intermediate and High Immunoscores combined into category IS Hi. Hazard ratio MSS IS Lo compared with MSI IS Hi = 6.6 (95% CI 2.3 to 19.6, p = 0.0006). IS = Immunoscore. Hi = High. Lo = Low. MSS = microsatellite stable. MSI = microsatellite instability high.

#### *3.2.4.4. Multivariate analysis*

A stratified multivariate Cox proportional hazards model was used to assess the associations between these factors RFS. The Immunoscore, disease stage and EMVI and were the strongest predictors of RFS (Figure 3.36). These associations remained when stratified for patients with UICC TNM Stage 1 to 3 disease (Figure 3.37).

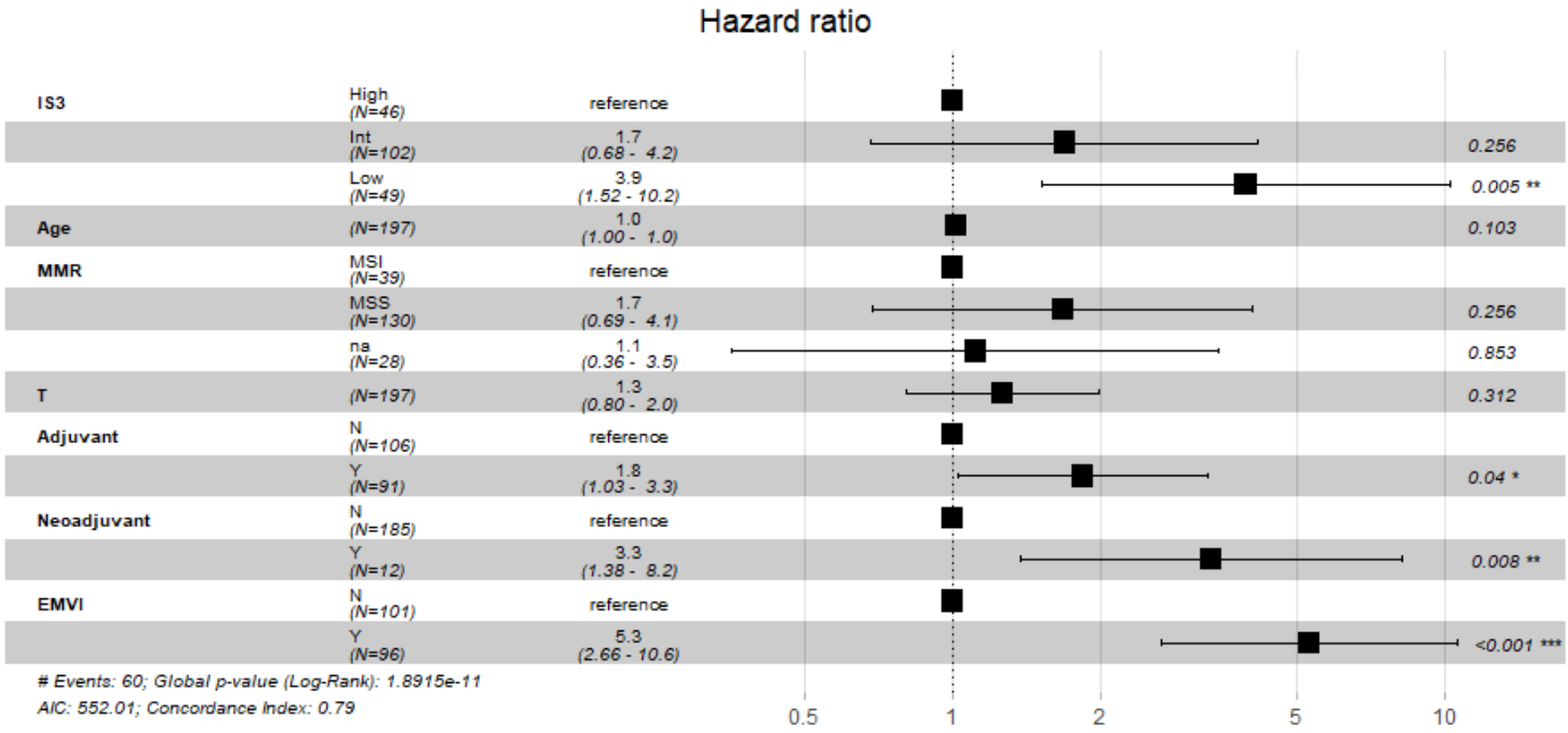



**Figure 3.36.** Forest plot showing hazard ratios for recurrence-free survival, all disease stages. Reference = Immunoscore High. EMVI = extramural venous invasion. IS3 = Three-category Immunoscore. MMR = mismatch repair status. MSI = microsatellite instability high. MSS = microsatellite stable. T = tumour T stage.

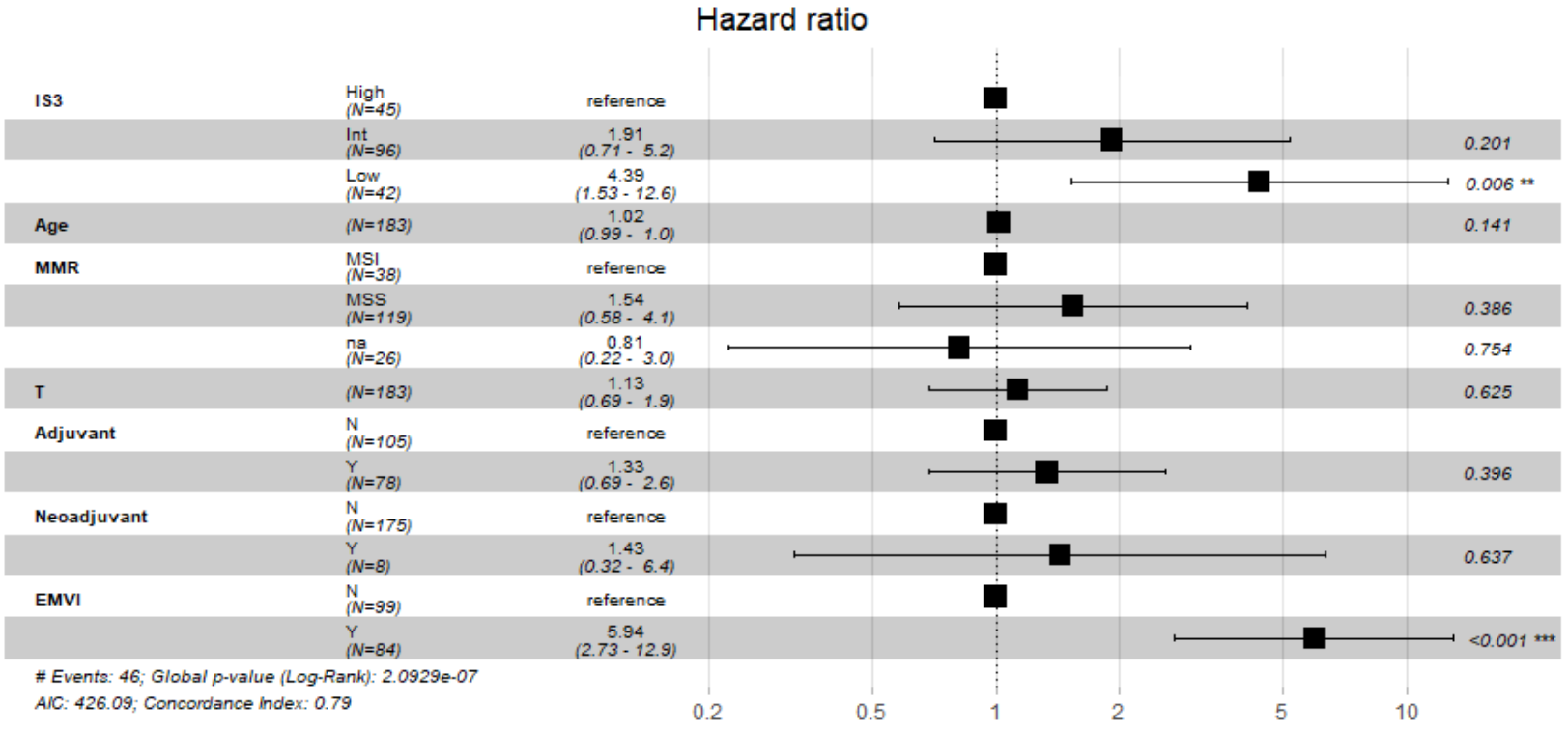



**Figure 3.37.** Forest plot showing hazard ratios for recurrence-free survival, UICC- TNM stage 1 to 3 disease. Reference = Immunoscore High. EMVI = extramural venous invasion. IS3 = Three-category Immunoscore. MMR = mismatch repair status. MSI = microsatellite instability high. MSS = microsatellite stable. T = tumour T stage.

## 3.3 Discussion

### 3.3.1. The sample population is clinicopathologically representative.

To determine the reliability and applicability of downstream analyses on this data set, it was important to determine the similarities between information available on these patients and those from nationwide cancer registries. The data set is shown to be representative of the wider population, but with two key differences. The first is that the results are skewed towards earlier stage disease, reflecting the sampling of patients having resections. The second is that the data set is skewed towards having an increased proportion of patients from a minority ethnic background. This will be discussed further below.

### 3.3.2. The Immunoscore is a valid marker of the colorectal immune environment.

The Immunoscore is used as the primary proxy marker of the immune contexture in this project. This preliminary analysis confirms its reliability, validity, and robustness as a key intermediate phenotype of the immune response to CRC in this patient set. The distribution of scores is shown to replicate the international validation study, and it strongly correlates with RFS.

### 3.3.3. The sample size is sufficiently powered for this study.

Determining the optimum sample size for a project of this nature required careful analysis and an awareness of the potential challenges and pitfalls. Some statistical and mathematical assumptions were made to make the computations possible. These included estimating MAF ranges for eQTL SNPs based on public databases, assuming no linkage or linkage disequilibrium, that is, that there are no non-random associations between alleles. Effect sizes were also assumed to be equally distributed. In addition, there can be significant differences in gene expression variation, both in *cis* and *trans,* between different population groups, as explored within the International HapMap3 project [141, 241]. The potential confounding effects of ethnicity on SNP MAFs is further explored in Chapter 2, and outliers in the data set which could confound the results have been excluded.

With adequate statistical power being obtained at a range of MAFs and effect sizes at sample sizes 150 to 200, patients were recruited in two cohorts with the aim to reach a sample size of at least 200. This also provided the advantage of providing cross-validation to test the robustness of data within the cohort. In future studies, increasing the sample size in each cohort will be important to verify the robustness of this data.

### 3.3.4. Potential sources of sample bias

#### *3.3.4.1. Sample attrition*

Due to the strict QC involved in generating the Immunoscore, and the requirement for completed WGS and RNA sequencing on the samples in which the Immunoscore was successfully completed, some sample attrition was inevitable. Although this represented a small fraction of the total, there were sufficient samples retained in each analysis to maintain statistical power to detect clinically meaningful effect sizes in *post hoc* analysis. However, this attrition may have introduced bias in the data set. This was addressed by performing a clinico-pathological analysis for the subset of samples in which the Immunoscore was successfully performed, and separately in the 100KGP samples. These showed no significant population differences.

#### *3.3.4.2. Recruitment bias*

The profile of patients who are recruited to clinical trials is often different from both the target population and the wider real-world patient population [242]. This is usually due to recruitment biases. Age biases in patient recruitment are well-recognised [243], as well as sex and ethnicity biases [244, 245]. This may have significant consequences in limiting the application of conclusions to the general patient population, particularly for randomised clinical trials.

Patient recruitment to the 100KGP and for this data set was done chronologically and opportunistically. Patients for whom tissue was not found in the Biorepository were excluded. The samples are also biased towards patients with earlier stage disease due to the recruitment of those having resections (usually curative), and the exclusion of those who had local field radiotherapy (which is reserved for locally advanced rectal cancer).

This data set does not appear significantly biased from national data sets regarding age or sex. However, the presence of a higher proportion of patients defined as being from a minority ethnic background could be a potential strength in this study given the highlighted long-standing and persistent deficiencies in recruitment of patients from these backgrounds to clinical trials.

#### *3.3.4.3. Reference bias*

In genomics research, reference bias is a pervasive issue. NGS techniques require mapping to a reference genome, which is itself being population-biased, and is being refined and continually updated [246]. Sequencing reads carrying alternate alleles will have mismatches when aligned against the reference genome and therefore lower mapping scores. This introduces problems with variant calling and can lead to alternative alleles being missed or wrongly called, influencing estimates of allele frequencies and heterozygosity [247]. Some of the strategies to overcome this including analysis of joint genotype VCFs and principal components analysis are explored in Chapter 2.

#### *3.3.4.4. Exclusion bias*

Finally, exclusion of samples that failed the Immunoscore QC checks could have introduced further bias by excluding patients either with smaller, earlier stage tumours, or necrotic, more poorly differentiated, or advanced cancers. The effect of these exclusions is more difficult to define. However, following further analysis of the data (Table 3.7), there were no significant alterations in the data profile after the Immunoscore was completed.

# Chapter 4: Immune gene expression quantitative trait loci (eQTL) single nucleotide polymorphism (SNP) analysis

## 4.1 Introduction

Genome-wide association studies (GWAS) have shown associations between germline differences in gene expression and cancer risk in many cancer types [158, 248]. Evidence of associations between germline differences in immune gene expression (*cis*- or *trans*-eQTLs) and cancer outcomes is growing. Notably, Vogelsang *et al.* observe strong correlations between the germline eQTL SNPs rs6673928 impacting *IL19* expression and rs6695772, impacting *BATF3* expression and overall survival in cutaneous melanoma [161]. In breast cancer, an enrichment of germline eQTL SNPs influencing the expression of MHC class I and II genes was observed in breast cancer survivors compared with healthy controls [163].

While somatic determinants of the immune response in CRC are being extensively studied, the role of germline determinants is significantly less so. This is partly due to the predominance of studies utilising WES databases which limits the exploration of eQTL associations and the bioinformatics complexity of eQTL studies, particularly *trans*-eQTLs.

However, it is reasonable to hypothesise that germline differences in immune gene expression contribute at least a part to the differences in the CRC immune environment. The WGS data provided by the 100KGP made investigating this feasible, and ready access to large eQTL databases such as MuTHER [155], and the initial identification of key immune gene SNPs by Vogelsang *et al.* provided helpful lynchpins to perform the requisite analysis.

## 4.2. eQTL analysis

### 4.2.1. Study population

eQTL SNP filtering was performed on patients for whom both the Immunoscore and germline VCFs were available in the GeL Research Environment and Bluebear. Data for 20 patients was not available within the Research Environment. 7 patients had been excluded due to failures in sample collection or processing. For 13 patients, data was not yet available in the Research Environment at the time of analysis in October 2020. In total, germline data was available for 177 patients. These comprised 30 from the pilot set and 147 from the 100KGP set. Analysis of the data was combined to obtain sufficient statistical power to detect a difference (Figure 4.1).

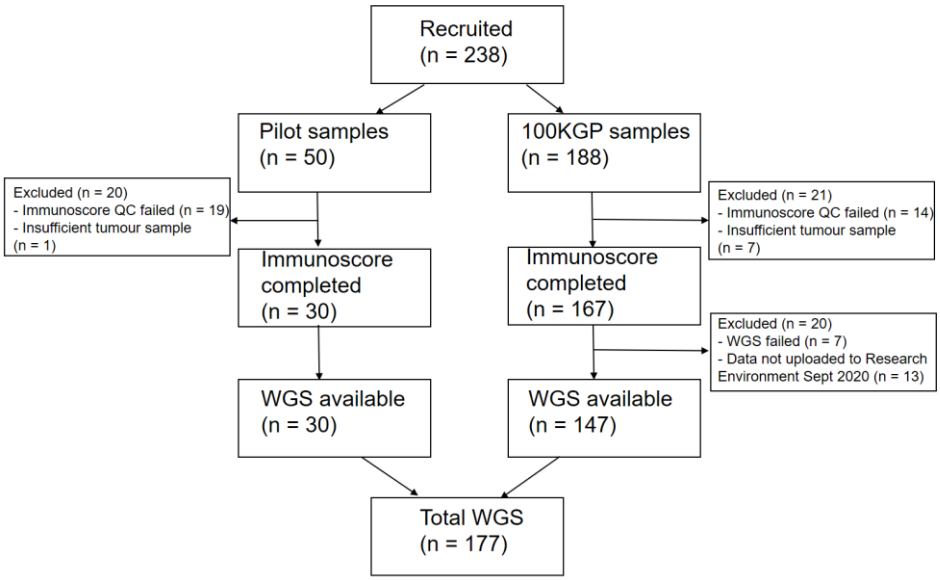



**Figure 4.1.** Flow diagram showing the number of patients recruited and included in the study population for the germline expression quantitative trait loci analysis. 238 patients were included, comprised of 50 in the pilot and 188 in the 100 000 Genomes Project. A total of 177 patients had both the Immunoscore and whole genome sequencing data available for analysis. 100KGP = 100000 Genomes Project, QC = quality control, WGS = whole genome sequencing.

### 4.2.2. eQTL SNP correlation with the Immunoscore

#### *4.2.2.1. The Vogelsang top 40 SNP panel*

The top 40 immune gene eQTL SNPs derived from Vogelsang *et al.* were correlated with the Immunoscore using ordinal logistic regression in R, using the additive genotypic model, with SNPs coded as 1, 2 and 3. Six SNPs had statistically significant correlations with the Immunoscore in initial analysis. These findings, as well as the SNP reference and alternative alleles and population minor allele frequencies from the Genome Aggregation Database (gnomAD) data set [249] are presented in Table 4.1.

**Table 4.1.** eQTL SNPs significantly associated with the Immunoscore

| RefSNP ID | Gene | Ref | Alt | MAF[249] | p | Odds ratio | 95% CI | Variant effect |
|---|---|---|---|---|---|---|---|---|
| rs6673928 | *IL19* | G | T | 0.20 | 0.037 | 2.41 | 1.06-5.56 | Increases Immunoscore |
| rs2295359 | *IL23R* | G | A | 0.32 | 0.007 | 0.19 | 0.05-0.63 | Decreases Immunoscore |
| rs11919943 | *CCR1* | T | C | 0.17 | 0.002 | 0.16 | 0.05-0.52 | Decreases Immunoscore |
| rs11161590 | *BCL10* | A | G | 0.40 | 0.020 | 0.54 | 0.32-0.90 | Decreases Immunoscore |
| rs11203203 | *UBASH3A* | G | A | 0.27 | 0.006 | 3.04 | 1.39-6.74 | Increases Immunoscore |
| rs10760142 | *C5* | T | C | 0.36 | 0.057 | 2.73 | 0.97-7.82 | Increases Immunoscore |

Alt = alternative allele. CI = confidence interval. eQTL = expression quantitative trait loci. Ref = reference allele. RefSNP ID = Reference SNP identity. MAF = minor allele frequency. p = statistical significance. SNP = single nucleotide polymorphism.

After stepwise selection with the stepAIC function (“forward”), the “train” function in the “caret” package was used to generate a predictive model with the best fit. The eQTL SNP rs11919943 (*CCR1*) had the strongest correlation with the Immunoscore.

#### *4.2.2.1.a. CCR1, IL23R and UBASH3A eQTL SNPs are moderately associated with the Immunoscore*

To correct for multiple testing, both the Bonferroni approach (which is conservative) and false discovery rate (FDR) and were used. As expected, the Bonferroni approach did not yield any significant results after correction for multiple testing. While the FDR correction similarly gave revised p values that were non-significant (greater than 0.05), there were three eQTL SNPs with revised p values closest to significant. These were rs22953593 (*IL23R*, p = 0.067), rs11919943 (*CCR1*, p = 0.067) and rs11203203 (*UBASH3A*, p = 0.067).

Using the dominant genotypic model, similar results were obtained. There were 6 eQTL SNPs most strongly associated with the Immunoscore. These were rs22953593 (*IL23R*, p = 0.007), rs11161590 (*BCL10*, p = 0.020), rs6673928 (*IL19*, p = 0.036), rs1378940 (CSK, p = 0.015), rs11919943 (*CCR1*, p = 0.002) and rs11203203 (*UBASH3A*, p = 0.006).

The results after correction for multiple testing were very similar to the additive model, with no SNPs remaining significant using the Bonferroni approach. Using the FDR approach, the three SNPs with revised p values closest to significant were rs22953593 (*IL23R*, p = 0.099), rs11919943 (*CCR1*, p = 0.086) and rs11203203 (*UBASH3A*, p = 0.099), as with the additive model. This is likely a

function of the statistical power of detection, as the relatively wide confidence intervals show.

#### *4.2.2.2. The extended MuTHER SNP panel*

The comparison was expanded to the wider list of immune gene eQTL SNPs with significant *cis*-eQTL activity from the MuTHER data, as this was likely to glean addition associations to this data set. The additive genotypic model was used. Of the 385 MuTHER eQTL SNPs analysed, 33 had significant associations with the Immunoscore (with p values less than 0.05, Table 4.2).

**Table 4.2.** MuTHER eQTL SNPs associated with the Immunoscore

| RefSNP ID | Gene | p value |
|---|---|---|
| rs2305740 | *IL12RB1* | 0.030 |
| rs2369006 | *AXL* | 0.045 |
| rs10422141 | *TICAM1* | 0.004 |
| rs9938225 | *NOD2* | 0.046 |
| rs9929191 | *IL17C* | 0.016 |
| rs9903464 | *NCOR1* | 0.025 |
| rs11650283 | *UBB* | 0.043 |
| rs159279 | *CCL7* | 0.021 |
| rs17558532 | *RARA* | 0.021 |
| rs7151065 | *IL25* | 0.012 |
| rs1951635 | *RNF31* | 0.016 |
| rs214267 | *PSEN1* | 0.019 |
| rs1152788 | *BCL11B* | 0.00005 |
| rs4145039 | *BCL11B* | 0.048 |
| rs2582559 | *AKT1* | 0.018 |
| rs4077582 | *CYP11A1* | 0.029 |

| | | |
|---|---|---|
| rs37831 | *PDPK1* | 0.027 |
| rs1642026 | *LAT* | 0.047 |
| rs13439094 | *STAR* | 0.006 |
| rs4963452 | *CD5* | 0.032 |
| rs9668139 | *PTPN6* | 0.022 |
| rs478829 | *KLRK1* | 0.004 |
| rs6581061 | *IL23A* | 0.010 |
| rs10473354 | *CCL28* | 0.004 |
| rs256208 | *TCF7* | 0.0002 |
| rs17517511 | *IL4* | 0.009 |
| rs889009 | *DOCK2* | 0.025 |
| rs2705777 | *CCL26* | 0.0003 |
| rs6978354 | *CAV1* | 0.048 |
| rs1519550 | *IL15* | 0.048 |
| rs13424201 | *CXCR1* | 0.012 |
| rs1381016 | *CXCL1* | 0.019 |
| rs2377856 | *LCK* | 0.031 |

eQTL = expression quantitative trait loci, ID = identification number, MuTHER = Multiple Tissue Human Expression Resource Project, SNP = single nucleotide polymorphism.

#### *4.2.2.2.a. The TCF7 eQTL SNP is the most significant gene associated with the Immunoscore*

When these results were corrected for multiple testing using the Bonferroni approach, only two SNPs remained strongly associated with the Immunoscore: rs256208 (*TCF7*) and rs1152788 (*BCL11B*). rs2705777 (*CCL26*) was marginally associated with the Immunoscore (p = 0.068). Using the FDR method, the same three SNPs remained strongly associated with the Immunoscore: rs256208 (TCF7), rs2705777 (CCL26) and rs1152788 (BCL11B) (Table 4.3).

**Table 4.3.** False discovery rate-corrected eQTL SNPs significantly associated with the Immunoscore

| RefSNP ID | Gene | Ref | Alt | MAF [249] | p | Bonferroni-corrected p | FDR-corrected p | Odds Ratio | 95% CI | Variant effect |
|---|---|---|---|---|---|---|---|---|---|---|
| rs256208 | *TCF7* | C | G | 0.28 | 0.0002 | 0.032 | 0.032 | 2.77 | 1.64-4.76 | Increases Immunoscore |
| rs2705777 | *CCL26* | A | G | 0.41 | 0.0003 | 0.068 | 0.034 | 0.39 | 0.23-0.64 | Decreases Immunoscore |
| rs1152788 | *BCL11B* | G | A | 0.26 | 0.00005 | 0.012 | 0.012 | 0.30 | 0.16-0.53 | Decreases Immunoscore |

Alt = alternative allele. CI = confidence interval. eQTL = expression quantitative trait loci. FDR = false discovery rate. Ref = reference allele. RefSNP ID = Reference SNP identity. P = statistical significance. SNP = single nucleotide polymorphism

These results were also corroborated using the dominant genotypic model. After correction for multiple testing using the FDR, only one SNP showed significant association with the Immunoscore: rs256208 (*TCF7*, p = 0.018). Two other SNPs showed associations with a trend toward significant values: rs17517511 (*IL4*, p = 0.071) and rs2705777 (*BCL11B*, p = 0.076).

The finding of the rs256208, the *TCF7* (Transcription Factor 7) SNP, as the most strongly associated SNP with the Immunoscore is particularly interesting, as TCF7 has an essential role in Wnt signalling through its interaction with β-catenin [250].

### 4.2.3. eQTL SNP correlation with gene expression in tumour tissue

Determining the effects of these variants on gene expression requires quantification of the expression of each immune gene. Ideally, total RNA from LCLs derived from whole blood would provide transcript levels to compare gene expression across the genotypes, as shown by Vogelsang *et al.*, who demonstrated a direct correlation between the rs6673928 alternative alleles (GG versus GT and TT) and IL19 mRNA expression in circulating CD4+ T cells [161].

I was unable to perform a similar analysis as whole blood and total RNA from the patients were not available at the time of this analysis. However, transcriptomic data from 3' RNA sequencing of fixed tumour tissue was available, and this was correlated with the genotypic data. The gene expression levels were normalised (as described in Chapter 3) and are stated as counts per million. Genotypes were correlated with RNA expression levels, using the genotypic model (with wild-type

genotypes labelled “1” and variants labelled “2” and “3”). Those eQTL SNPs correlated with the Immunoscore were compared with RNA expression levels. This approach is justified as the *cis*-eQTL SNP effects were found to be largely tissue-independent and shared across the three tissue types analysed in the MuTHER study [154].

The rs6673928 (IL19) and rs11919943 (CCR1) SNPs appeared to show increasing expression in tumour tissue when compared with the genotypes, however, these did not reach statistical significance (Figure 4.2).

The other SNPs compared (rs256208 (TCF7), rs2705777 (CCL26), rs1152788 (BCL11B), rs2295359 (IL23R), rs11161590 (BCL10) and rs10760142 (C5)) did not show correlations with RNA expression levels in tumour tissue (Figure 4.3).

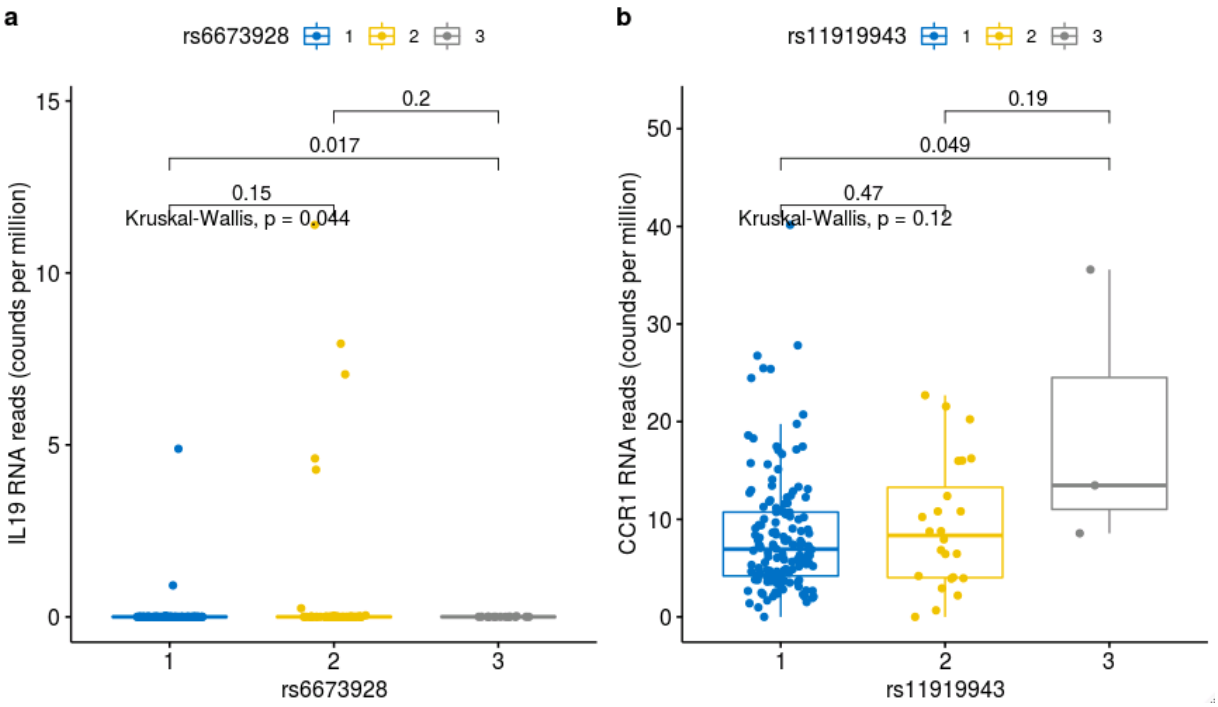



**Figure 4.2.** Boxplots illustrating the associations between germline genotypes and tumour RNA expression levels for (a) rs6673928 and IL19 and (b) rs11919943 and CCR1. Genotypes 1 = wild-type, 2 = heterozygous variant, 3 = homozygous variant. In (a) there is a rise in IL19 expression between the wild-type and heterozygous variant SNP, p = 0.017. In (b) there is an association with a rise in CCR1 expression between the wild-type and homozygous variant, p = 0.049.

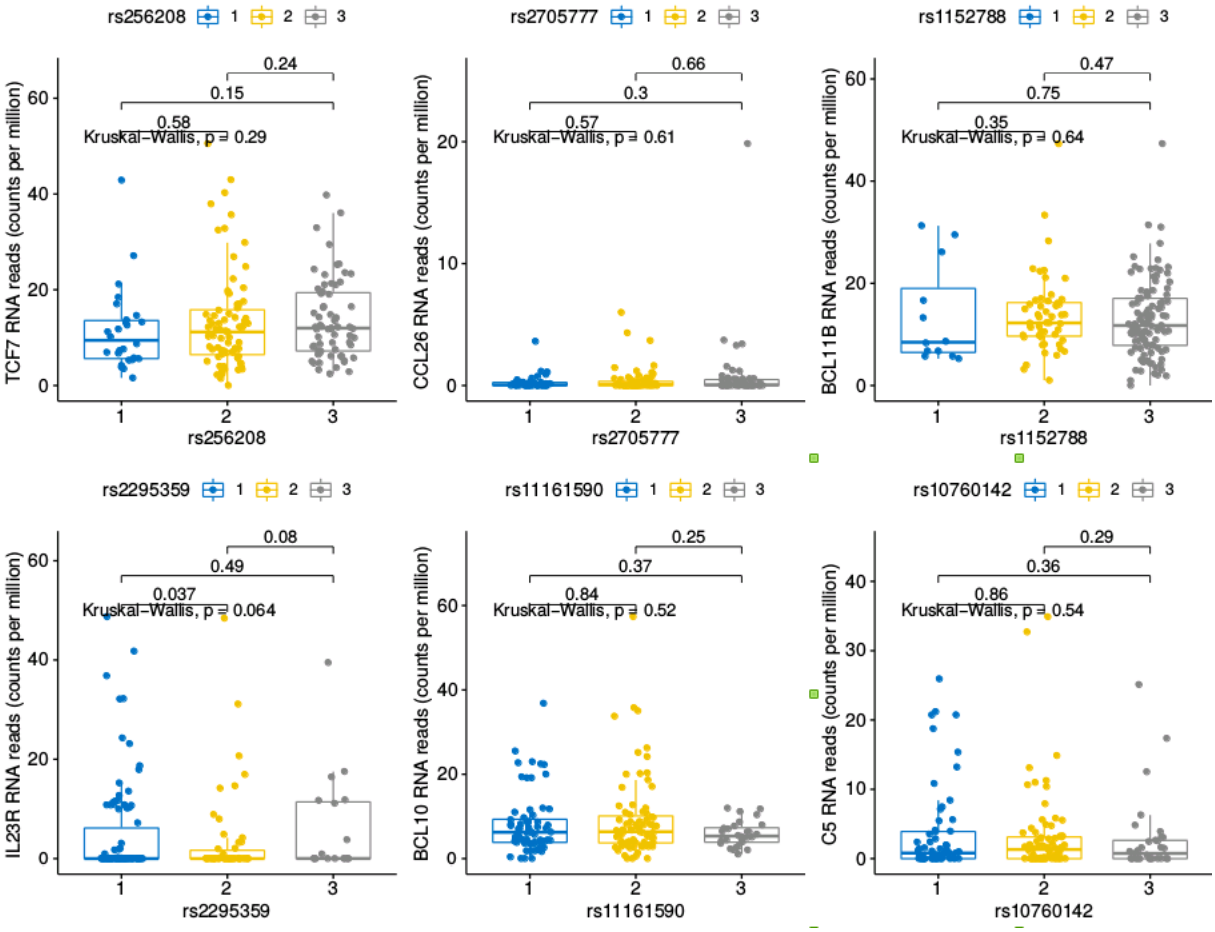



**Figure 4.3.** Boxplots illustrating the associations between germline genotypes and tumour RNA expression levels for six eQTL SNPs (TCF7, CCL26, BCL11B, IL23R, BCL10 and C5). Genotypes 1 = wild-type, 2 = heterozygous variant, 3 = homozygous variant. There are not significant associations between genotypes and tumour RNA expression levels.

These weak associations between germline genotypic variations and RNA expression in tumour tissue may suggest that there is more tissue specificity in these specific eQTLs than the MuTHER data suggest. Also, the tumour microenvironment is known to be extremely complex. Modulatory effects on immune cells may alter their ability to infiltrate and thrive within the tumour, which could obliterate any effects of increased expression of particular genes [105], while differential gene expression in tumour cells or adjacent cells may also be significantly modulated by metabolic suppressive effects [105, 107].

### 4.2.4. eQTL SNP association with patient survival

Further analysis was performed to determine if there was an association between these SNP variants and clinical outcomes. In particular, the associations with recurrence-free survival (RFS) were examined as this is more directly relevant to cancer outcomes than overall survival (OS).

SNPs were classified into wild-type and variant (combining both heterozygous and homozygous variants) to increase the size of each group and the likelihood of achieving statistical power. One sample was removed (from the pilot set) as clinical data was not available. Kaplan-Meier curves showed apparent survival differences for just the rs11203203 variant (*UBASH3A*), which was associated with both increased RFS and OS (Figure 4.4, Figure 4.5).

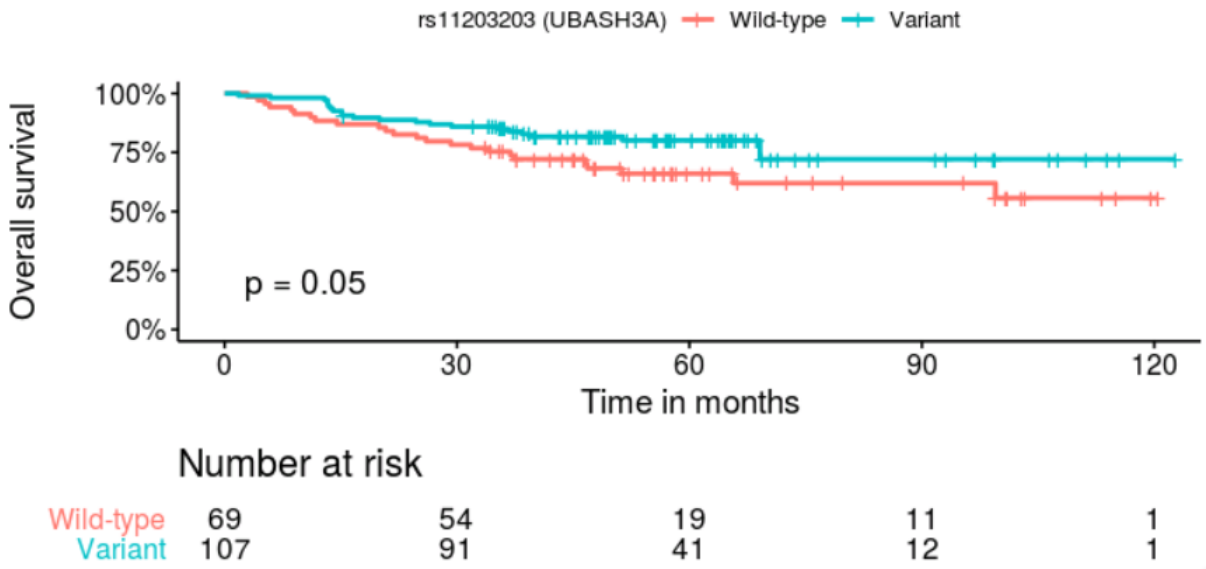



**Figure 4.4.** Kaplan-Meier estimate of overall survival (OS) stratified by rs11203203 (UBASH3A) eQTL SNP variant. There is increased survival with the variant allele (p = 0.017). Hazard ratio for variant allele = 0.57 (95% confidence interval = 0.32 – 1.00).

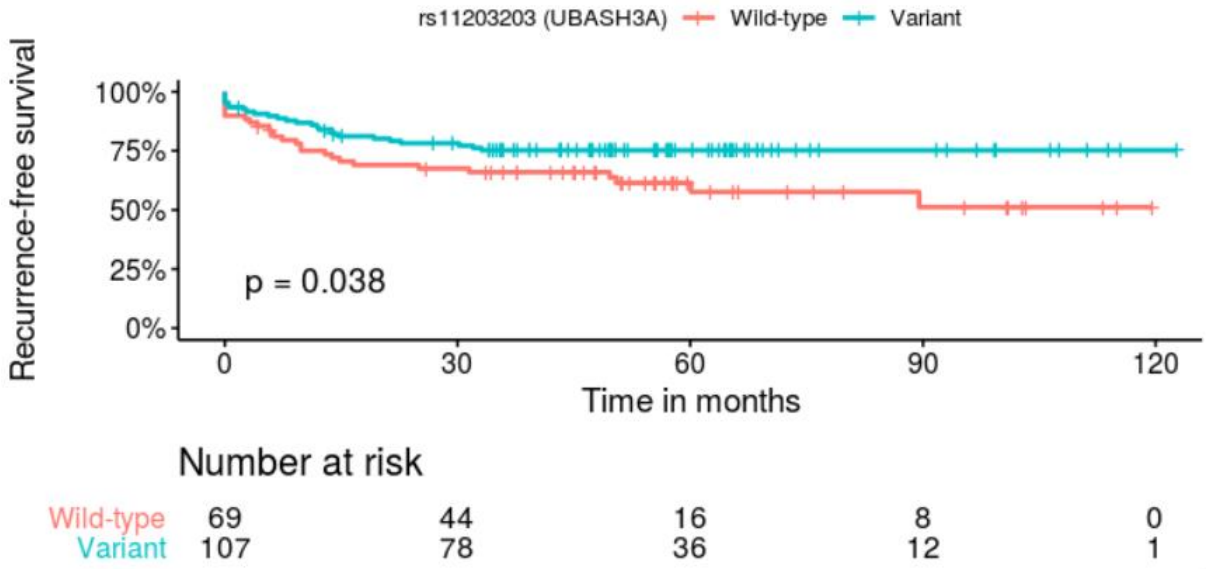



**Figure 4.5.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by rs11203203 (UBASH3A) eQTL SNP variant. There is increased survival with the variant allele (p = 0.016). Hazard ratio for variant allele = 0.57 (95% confidence interval = 0.33 – 0.97).

Despite a strong association with the Immunoscore, the rs256208 (*TCF7*) SNP was not associated with a difference in either OS or RFS (Figure 4.6, Figure 4.7).

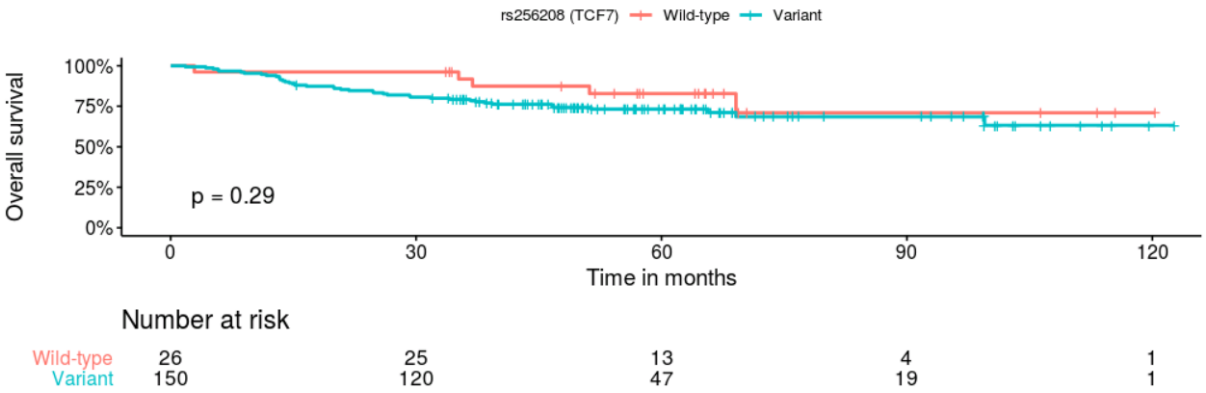



**Figure 4.6.** Kaplan-Meier estimate of overall survival (OS) stratified by rs256208 (TCF7) eQTL SNP variant. There is no significant difference in OS between the groups.

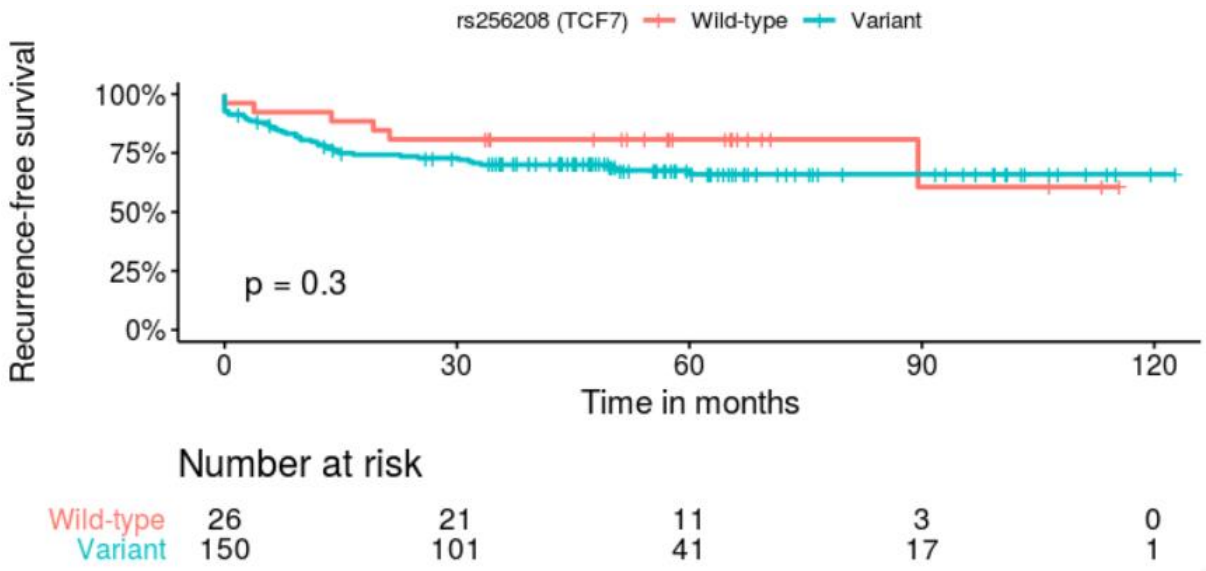



**Figure 4.7.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by rs256208 (TCF7) eQTL SNP variant. There is no significant difference in RFS between the groups.

These results suggest that the SNP effects may not be of sufficient magnitude to cause significant differences in clinical outcomes in a small data set. Further analysis showed no associations of these SNPs with other factors such as age, sex, primary tumour location, MMR status or disease stage (see Appendix 5). For the most SNPs there were also no associations with patient ethnicity, tumour T stage and EMVI. However, there was a significant association between the rs255208 SNP and patient ethnicity, which is discussed and interrogated further below. There were also associations between EMVI and the rs1152768 SNP, and tumour T stage and the rs250577 SNP.

### 4.2.5. Principal components analysis of SNP MAFs by ethnicity

SNP incidences are influenced by population demographics such as, for example, ethnicity. This could potentially skew the conclusions derived from comparing SNP differences. Data from the 100KGP suggest that self-declared ethnicity of participants in this project ties very closely with the haplotypes in population SNP panels, as determined by the HapMap Project [136]. To interrogate the possibility that the eQTL SNP differences could be influenced by ethnic differences in the patient cohort, the patient ethnicity data was compared with the genotypes. For simplicity of analysis, patients were divided into four ethnic groups. The majority were described as White British or Irish (taken as the “European” population), comprising 88% of participants.

Principal components analysis (PCA) of the top 40 immune gene SNPs was performed in R using the packages “stats” and “ggfortify”. The highest five proportions of the variances (eigenvalues) in the principal components were 5.78%, 5.38%, 5.15%, 4.73% and 4.33%, while the lowest value was 0.28%. The first 13 components contributed to 50.4% of the variances (Figure 4.8).

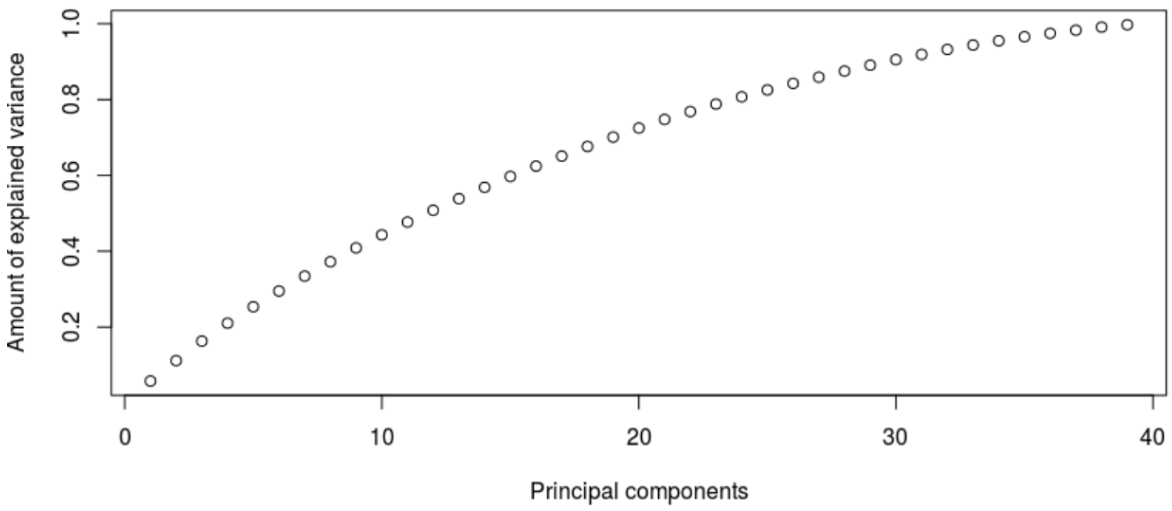



**Figure 4.8.** Cumulative variance plot of the principal components of the 40 individual SNPs. Most components contribute significantly to the data variability.

The first two components (PC1 and PC2) were plotted, with a focus on any outliers in the ethnicity distribution of the eQTL SNPs (Figure 4.9). These showed significant overlap in the distribution of the SNPs between all ethnic groups without significant outliers.

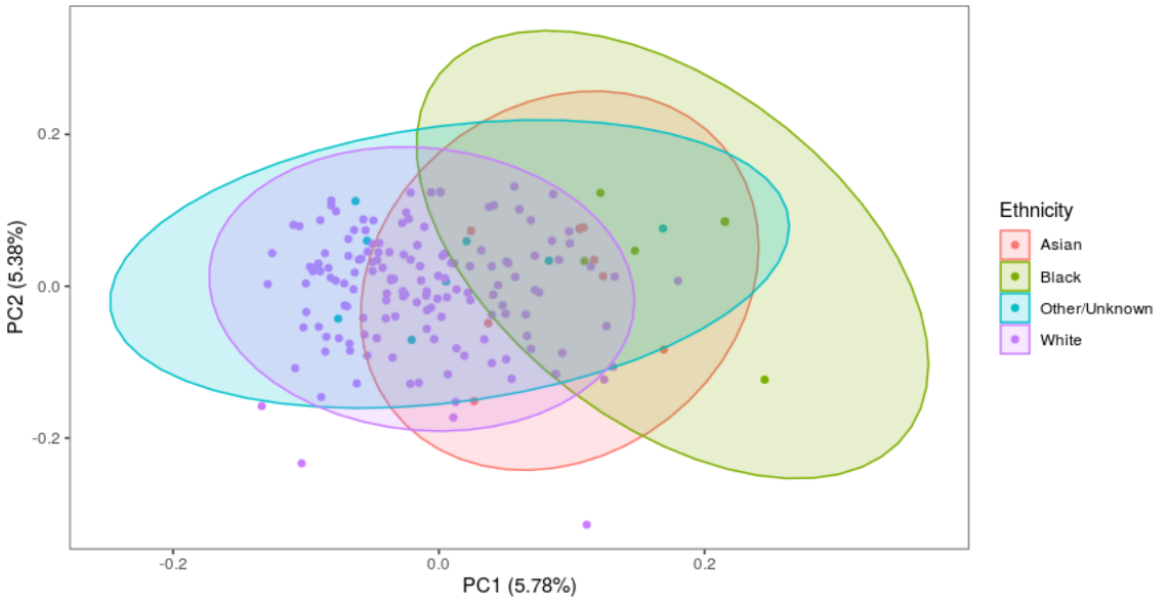



**Figure 4.9.** Principal components analysis of SNPs by ethnicity. PC1 = principal component 1, PC2 = principal component 2. The probability ellipses show clustering by ethnicity. There is significant overlap between all ethnicity categories.

However, when the PCA was expanded to include the extended SNP panel (385 SNPs), there was a clear difference in the distribution of SNPs amongst Black patients, in comparison to the other groups (Figure 4.10).

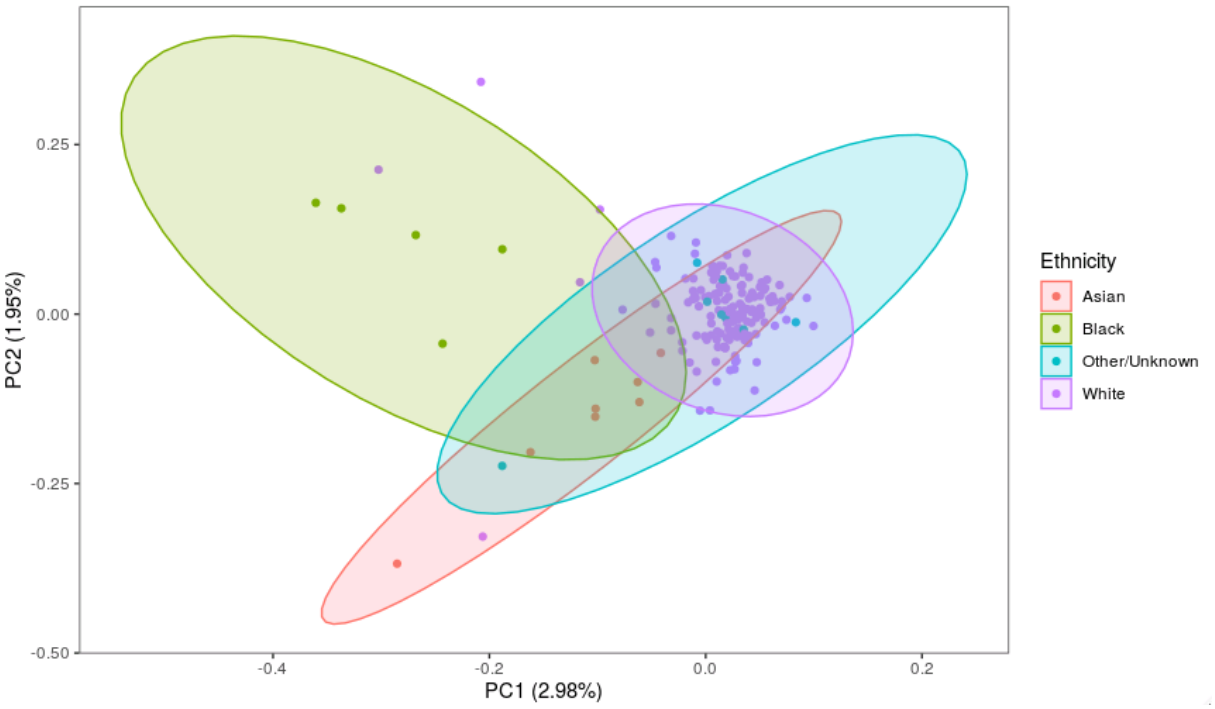



**Figure 4.10.** Principal components analysis of SNPs by ethnicity. PC1 = principal component 1, PC2 = principal component 2. The probability ellipses show clustering by ethnicity. There is significant discrepancy between the ellipses for Black patients compared with those of other ethnicities.

When the only top nine significant SNPs were compared, this discrepancy was less pronounced (Figure 4.11). The highest proportions of the variances (eigenvalues) in the principal components were 15.9% and 10.2%.

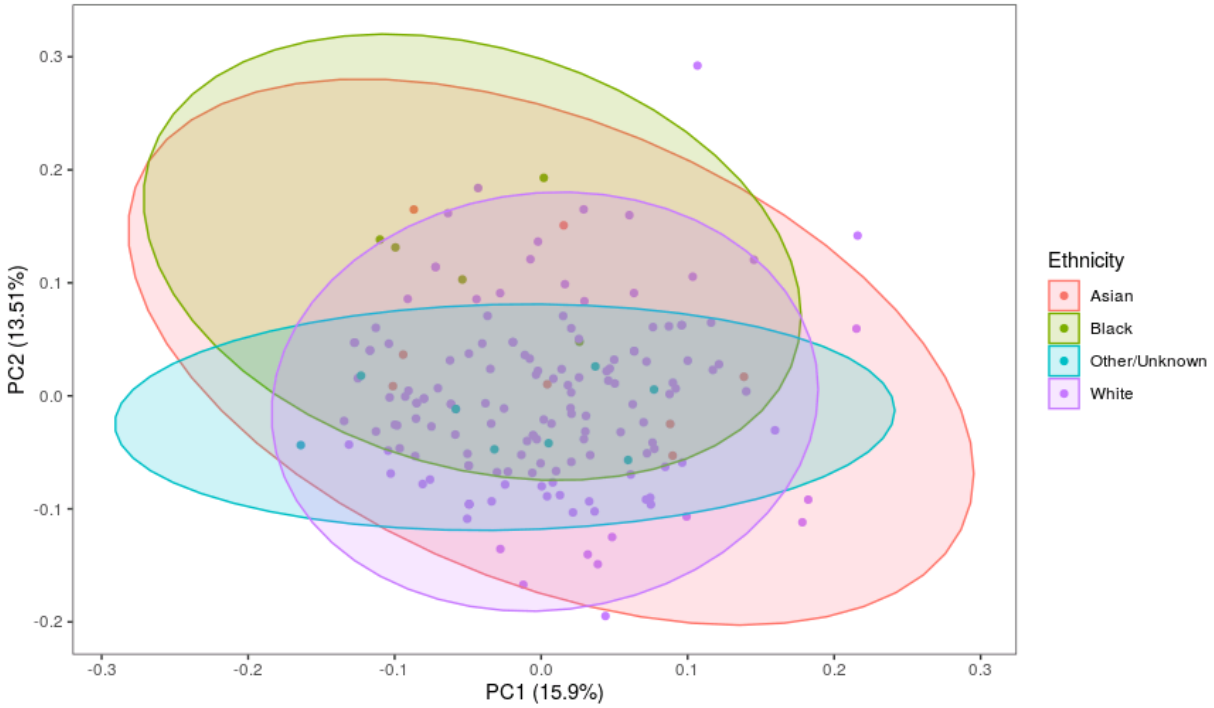



**Figure 4.11.** Principal components analysis of SNPs by ethnicity for nine SNPs (rs256208, rs2705777, rs17517511, rs6673928, rs2295359, rs11919943, rs11161590, rs11203203 and rs10761042). PC1 = principal component 1, PC2 = principal component 2. The probability ellipses show clustering by ethnicity. There is significant overlap between all ethnicity categories.

Analysis of differences in genotypes by ethnicity were performed individually for the significant nine SNPs. There were differences in the SNP genotypes primarily for Black patients compared with other patients. Particularly, differences were noted for Black patients compared with White patients with the SNPs rs256208 (Kruskal-Wallis test, p = 0.0057, Figure 4.12), rs2705777 (Kruskal-Wallis test, p = 0.034, Figure 4.12), rs11919943 (Kruskal-Wallis test, p = 0.002, Figure 4.13) and rs10760142 (Kruskal-Wallis test, p = 0.044, Figure 4.13). The discrepancies for Asian patients compared with other patients (primarily White patients) were only noted for one SNP, rs10760142 (Kruskal-Wallis test, p = 0.049, Figure 4.13).

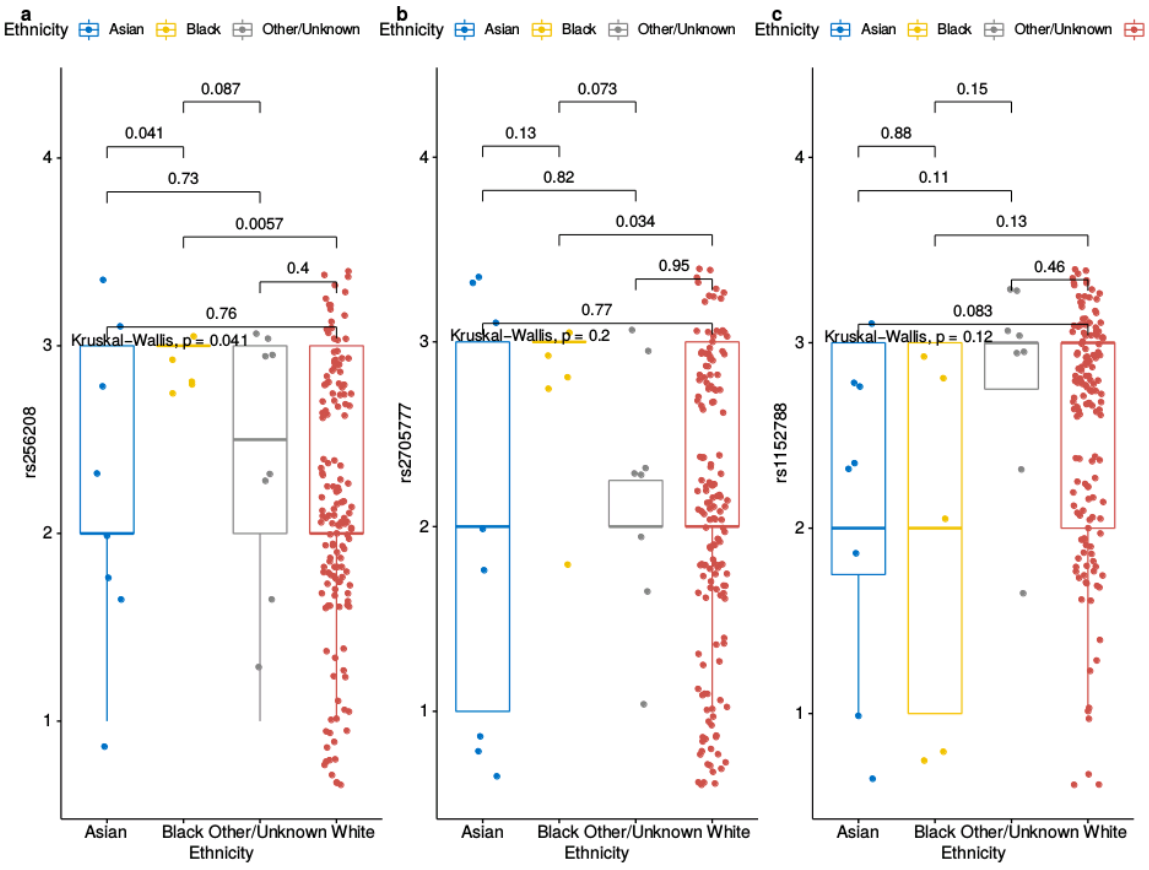



**Figure 4.12.** Boxplots illustrating the associations between the three SNP genotypes ((a) rs256208, (b) rs2705777 and (c) rs1152788) and patient ethnicity. Significant differences

between genotypes in Black patients compared with other patients are noted for rs256208 (a) and rs2705777 (b).

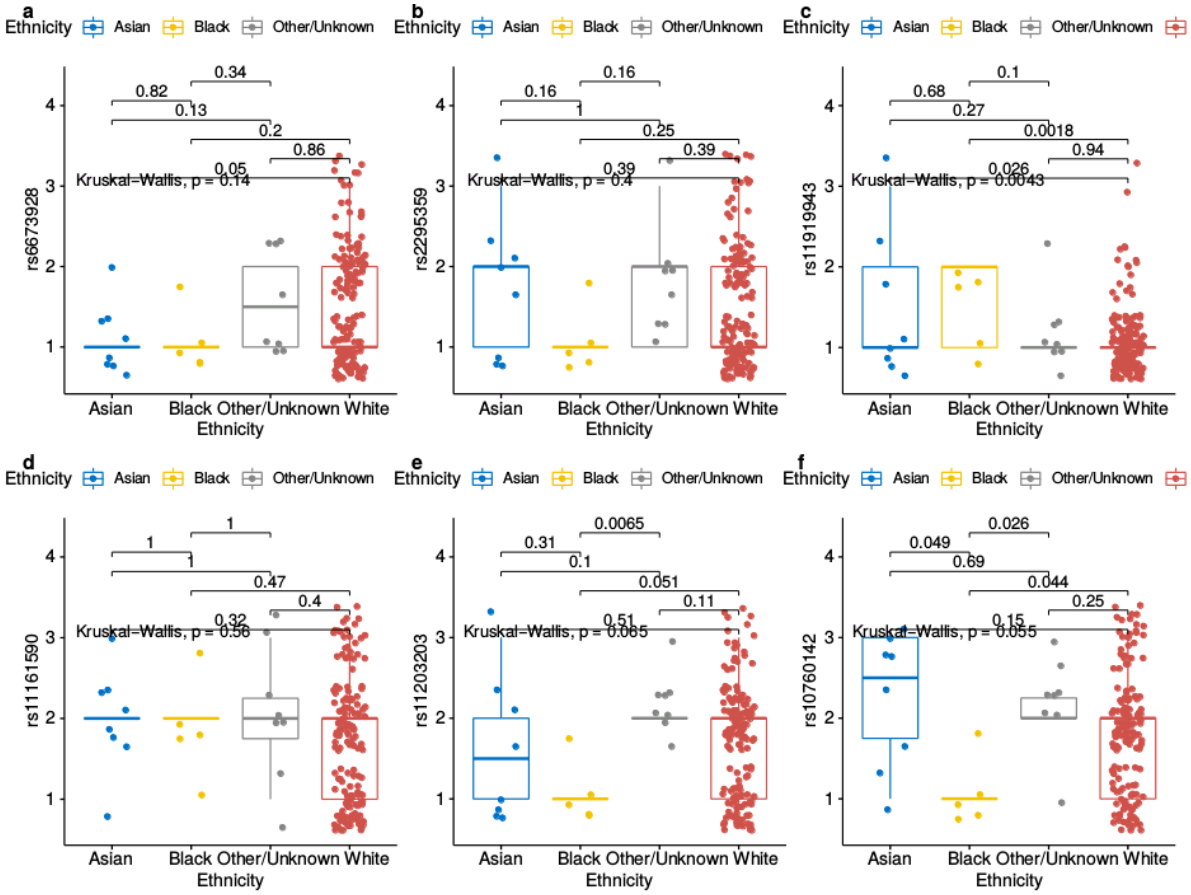



**Figure 4.13.** Boxplots illustrating the associations between six SNP genotypes ((a) rs6673928, (b) rs2295359, (c) rs11919943, (d) rs11161590, (e) rs11203203 and (f) rs10761042) and patient ethnicity. Significant differences between genotypes in Black patients compared with White patients are noted for rs11919943 (c) and rs10760142 (f), and significant differences between genotypes in Asian patients compared with White patients are noted for rs10760142 (f).

#### *4.2.5.1. Filtered eQTL-Immunoscore associations by ethnicity*

Black patients make up 3.5% of the study population in this assessment (6 of 177 patients). The eQTL SNP-genotype analysis was filtered by ethnicity and specifically details of the patients identified as Black were removed from the analysis. For the top 40 SNPs, following correction for multiple testing, using the Bonferroni approach there were no significant SNPs. Using the FDR approach, the three SNPs with p values closest to significant were almost identical to previous results: rs2295359 (*IL23R*, p = 0.092), rs11919943 (*CCR1*, p = 0.092) and rs11203203 (*UBASH3A*, p = 0.092). When this was similarly applied to the extended MUTHER SNP panel, once again, the results were similar (Table 4.4), confirming that the significant results seen are not influenced by the population SNP differences.

**Table 4.4.** False discovery rate-corrected eQTL SNPs significantly associated with the Immunoscore

| RefSNP ID | Gene | p value | Bonferroni p value | FDR p value |
|---|---|---|---|---|
| rs256208 | *TCF7* | 0.0002 | 0.088 | 0.044 |
| rs2705777 | *CCL26* | 0.0003 | 0.041 | 0.041 |
| rs1152788 | *BCL11B* | 0.00005 | 0.050 | 0.051 |

FDR = false discovery rate, ID = identification number, SNP = single nucleotide polymorphism.

## 4.4. Discussion

In this chapter, an exploratory, *in silico* analysis of potential germline drivers of the immune environment in colorectal cancer was made. I found that there are germline eQTL SNPs that are associated with differences in the immune environment in CRC. These warrant validation in a larger sample set and an exploration of the biological mechanisms underpinning their likely mechanisms of action. In addition, in this sample set, there were no patients who received immunotherapy, so an exploration of the predictive value of these eQTL SNPs in determining the response to immunotherapy was not possible.

### 4.4.1. There are associations between the key eQTL SNPs and RNA expression levels and survival

The results obtained reveal some complexities. The rs11203203 (*UBASH3A*) variant alleles are associated with both increasing Immunoscore and increased RFS, which suggests that patients with these alleles are more likely to have “hot” immune environments, and could potentially be ideal candidates for immunotherapy. However, the effects of these alleles on gene expression levels was not determined. On the other hand, while the rs256208 (TCF7) variants are associated with increasing Immunoscore, there were no associations with survival. The rs11919943 (*CCR1*) variants appear to be associated with both decreasing Immunoscore and potentially lower RFS.

These associations suggest that the SNP effects may not be of sufficient magnitude to cause significant differences in clinical outcomes in a small data set. It would be of particular interest to determine the predictive value of these

SNPs in determining responses to immunotherapy, but as no patients received immunotherapy, this was not possible in this data set.

### 4.4.2. Potential biological mechanisms

The genes found to be most significantly associated with the Immunoscore have plausible biological mechanisms of activity in colorectal cancer. Among the top 40 SNPs used in the initial analysis, exploration of the Human Protein Atlas [222] revealed that some of the relevant genes have expression levels associated with differential outcomes in different cancer types.

For example, BCL10 expression is a prognostic marker in colorectal cancer, with increased expression associated with improved survival. CCR1 expression is also known to be a prognostic marker in renal cancer, with increased expression of this marker linked to unfavourable outcomes. C5 expression is associated with favourable outcomes in liver cancer. The rs6673928 SNP driving IL19 expression has been shown to be linked with better survival in cutaneous melanoma [161].

Of the wider SNP panel, BCL11B expression is known to be a favourable prognostic marker in urothelial cancer. TCF7 expression is thought to be linked to worsened outcomes in gastric and a range of cancers including prostate, breast, adrenal and pancreatic cancer [251]. Its role in CRC is particularly interesting, as the TCF family of genes has a significant role in Wnt signalling.

#### *4.4.1.1. The TCF/LEF pathway in colorectal tumour biology*

Wnt/β-catenin signalling is critical in colorectal carcinogenesis. It is a global regulator of embryonic development, and subsequently is necessary for ongoing homeostatic tissue renewal. In intestinal crypt cells, pathway activity is necessary to maintain stem cells. Pathway mutations are the main drivers of colorectal carcinogenesis. Most notably, loss-of-function mutations in the *APC* gene drive over 80% of most sporadic colorectal tumours [8], by leading to β-catenin accumulation and subsequent activation of one of the TCF/ lymphoid enhancer-binding (LEF) family of transcriptional activators [252].

The TCF/LEF family has 4 members, with heterogeneous effects. All are subject to alternate splicing and their function is isoform-dependent, but mainly functioning as Wnt signalling effectors [251]. *TCF7* has both a high-mobility group DNA-binding domain and a β-catenin-binding domain. The role of TCF7 in colorectal cancer tumorigenesis and progression appears to be contradictory. For genes whose expression is induced by Wnt/β-catenin signalling, TCF/LEF appears to repress transcription in the absence of signalling, but is converted to an activator by association with β-catenin.

In CRC, loss-of-function mutations in TCFL2 are extremely common [8]. In addition, in a pre-clinical model of CRC, TCF7/TCF1 signalling was required for stimulating a CD8+ T cell effector response in the tumour environment in response to a combination of anti-Tim3 and anti-PD-1 blockade [253]. Tang *et al.* show that knockout of *TCF7* in mice produces adenomas in the intestine, whereas knockout of *TCF7* in colon cancer cell lines slows their growth [252]. This is likely explained by a switch in its isoform expression from a Wnt-opposing

dominant negative in normal cells, to a Wnt-promoting full-length isoform in cancer cells [250]. The positive association of *TCF7* variants with the Immunoscore in this data set suggest an overall tumour-suppressing role of TCF7 expression in CRC tumour biology.

### 4.4.3. Limitations and future development

Further validation of this data is warranted, requiring a larger sample set for which both germline WGS data and the Immunoscore (as the validated marker of the CRC immune environment) are available. The associations between these SNPs and gene expression also requires corroboration with RNA and protein expression from immune cells, best derived from patient whole blood. This requires access to these samples in large numbers in real-time.

Mechanistic explanations for these eQTL SNP effects are also not available. This requires exploration of the effects of inducing and downregulation of these genes in a CRC model to determine the effects on the immune contexture in these cells, and will form the basis for further work.

Finally, exploration of the predictive values of particular SNP, for example rs256208, requires identification of a population of patients who have undergone immune checkpoint blockade therapy, ideally in the neoadjuvant or early disease stage setting. Preliminary trials of immunotherapy in neoadjuvant settings are underway [24], and will provide greater clarity about germline markers of clinical efficacy.

# Chapter 5: Somatic determinants of the immune environment and the Immunoscore

## 5.1. Introduction

Tumour mutations and neoantigens are the main drivers of immune activation in the tumour environment of most cancer types that have been studied [7]. In particular, differences in tumour mutational burden (TMB) have been thought to explain the key differences in the immune environment observed in MSI-high versus MSS CRC [29]. In particular, there are suggestions that TMB may be predictive of the response to immunotherapy in MSI-high CRC [254]. However, this relationship does not appear to hold for MSS CRC [29].

It is increasingly clear that clonal neoantigens are more immunogenic than subclonal neoantigens [67, 185], and that the effect of clonality in determining outcomes and responses to immunotherapy may be more significant than the role of neoantigen burden [164]. Determining neoantigen burden involves prediction of MHC Class I (and increasingly Class II) epitopes. Further determination of clonality can be performed using a variety of strategies, of which the most promising are a modified Dirichlet process clustering approach [184] and the Mutant Allele Heterogeneity (MATH) score [188].

The first part of this chapter explores the associations between the Immunoscore and a series of somatic markers including the TMB, neoantigen burden and neoantigen clonality (as determined by the MATH and the modified DPClust approach). It also explores the associations between these somatic markers and clinical (survival) outcomes.

The second part of this chapter examines in detail the correlations between the expressions of several immune gene signatures, obtained from 3’ transcriptomic

sequencing data, and the Immunoscore. The most notable of this is the co-ordinate immune response cluster (CIRC), a Th-1 gene signature shown to be highly enriched in strongly immunogenic CRC immune environments [92]. Other markers and signatures studied were MHC Class II gene expression, which is known to be independently a marker of immune responsivity in CRC [61, 63, 64], gut bacteria chemokine-associated signatures [205], lymphangiogenic markers, which could promote the tracking of T cells into tumour tissue and paradoxically be associated with increased immunogenicity and better outcomes [255], and finally *wnt*-signalling-associated markers [256].

In the final part, immunohistochemical analysis of fixed tumour tissue was used to corroborate these results, to create a comprehensive picture of the effects of these somatic determinants on the CRC immune response.

### 5.2. Tumour mutational burden and the Immunoscore

Data on TMB, defined as the number of somatic coding non-synonymous mutations per megabase, was obtained from somatic WGS samples within the GeL Research Environment, and from the pilot data set. Of the 238 patients recruited, TMB data was available for 200 (29 from the pilot set and 171 from the 100KGP set). Of these, combined TMB and Immunoscore data were available for 177 patients.

TMB ranged from 0 mutations/Mb to 577.91 mutations/Mb (median = 4.39 mutations/Mb). Median TMB was significantly greater in MSI-high than MSS tumours (Wilcoxon test, p = 2.8e-15, Table 5.1).

**Table 5.1.** Comparison of tumour mutational burden in microsatellite stable and microsatellite unstable colorectal cancer

| | **Median TMB (mut/Mb)** |
|---|---|
| **MSS CRC (n = 121)** | 3.6 |
| **MSI-high CRC (n = 38)** | 68.3 |

Patients with mismatch repair status available, n = 159. Wilcoxon test, p = 2.8e-15. CRC = colorectal cancer, MSI-high= microsatellite instability high, MSS = microsatellite stable, Mut/Mb = somatic coding non-synonymous mutations per megabase, n = number, TMB = tumour mutational burden.

However, there was no significant association between TMB and the Immunoscore (IS) (Kruskal-Wallis test, p = 0.26, Figure 5.1). Median TMB across Immunoscore ranked categories was 3.74/ Mb for IS Low, 4.44/ Mb for IS Int, and 4.68 /Mb for IS High.

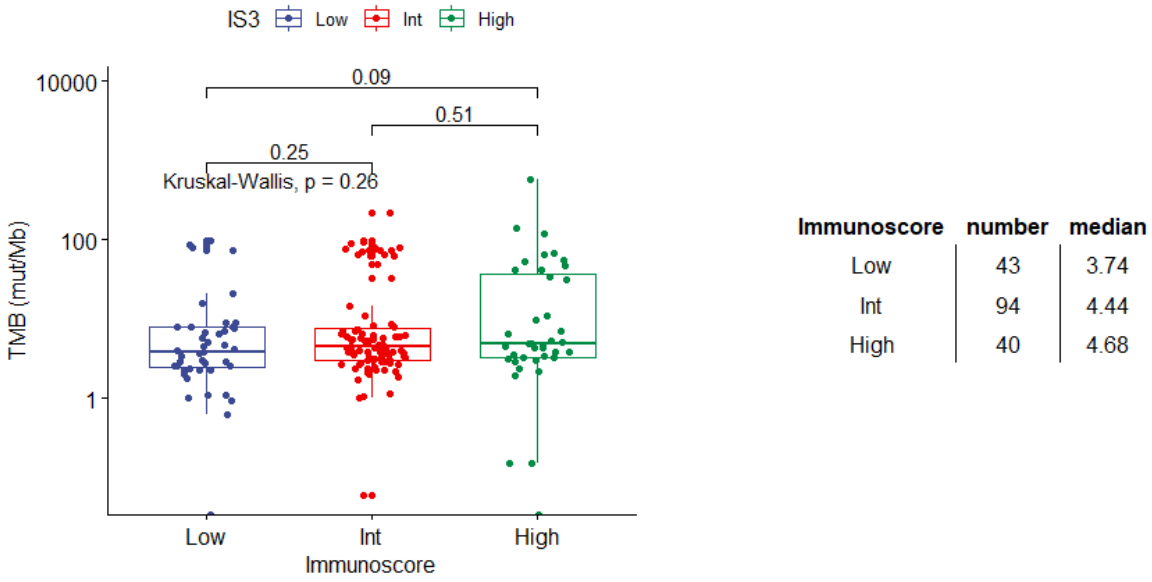



**Figure 5.1.** Comparison of TMB across Immunoscore categories. Int = Intermediate. IS3 = Immunoscore categories. mut/Mb = somatic coding non-synonymous mutations per megabase, TMB = tumour mutational burden.

As patients with MSI-high CRC had higher TMB scores, this could be a potential confounder. To assess this, analysis of TMB and Immunoscore in patients with MSS CRC only was performed. This confirmed a lack of a significant association between TMB and the Immunoscore (Figure 5.2).

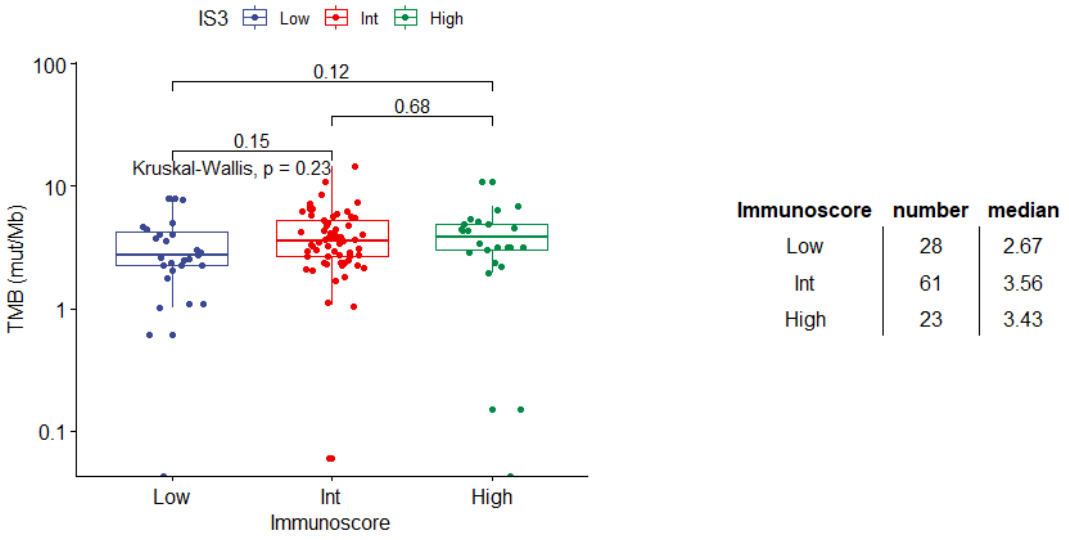



**Figure 5.2.** Comparison of TMB across Immunoscore categories (microsatellite stable colorectal cancer only). Int = Intermediate, IS3 = Immunoscore categories, mut/Mb = somatic coding non-synonymous mutations per megabase, TMB = tumour mutational burden.

The TMB was averaged across the data set and the samples were divided into “high” and “low” TMB depending on whether the TMB was higher or lower than the median. Survival analysis did not show a significant association between the TMB rank and either OS or RFS (Figure 5.3).

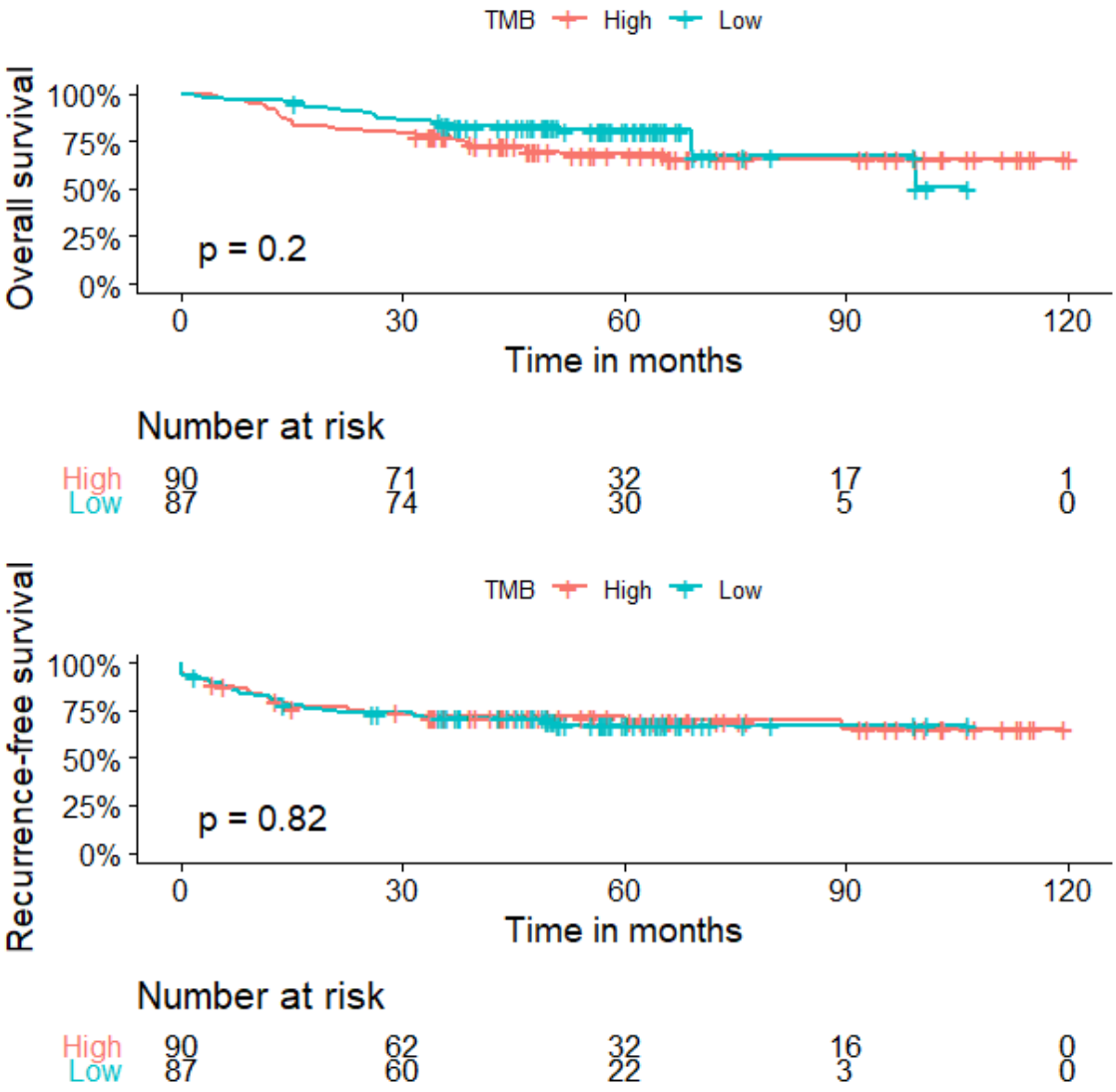



**Figure 5.3.** Kaplan-Meier estimate of overall survival (OS) and recurrence-free survival (RFS) stratified by tumour mutational burden. High = greater than median. Low = less than median. There is no difference in OS (Cox proportional hazards ratio, p = 0.39) or RFS (Cox proportional hazards ratio, p = 0.71) between the groups.

### 5.2.1. Neoantigen burden and the Immunoscore

Although TMB is not significantly associated with the Immunoscore in this data set, it is recognised that several mutations have no functional consequences, or do not lead to the generation of neoantigens, which are the primary targets of the T-cell-mediated anti-tumoral immune response [55].

The association between neoantigen burden and the Immunoscore was analysed to review specifically the impact of neoantigens on the colorectal tumour microenvironment. Neoantigen burden was calculated for tumours within the 100KGP RE using Neopredpipe [174]. The Neopredpipe pipeline uses netMHCpan to predict neoantigen binding, and as a result, only predicts MHC Class I neoantigens.

The Neopredpipe work was carried out in collaboration with Professor Graham's team. In Neopredpipe, peptides binding more than one HLA could be counted several times, and frame shift mutations could theoretically produce hundreds of neoantigens. As the aim was to count each peptide only once, in the pipeline, a filter for unique neoantigens, both single nucleotide variants (SNVs) and insertions and deletions (indels), was used to generate this information for each tumour. The pipeline was also completed on the pilot data set samples for 30 patients from the pilot set for whom the Immunoscore had been completed.

The 100KGP samples were not all available for analysis at the of completion of the pipeline, so that, in total, neoantigen burden data was available 113 of the 167 100KGP samples (67.7%) and 24 of the 30 pilot samples (80.0%). 137 samples were available for analysis (69.5%). This data attrition could introduce a

source of bias to the results and conclusions drawn, but the sample size is sufficiently large to support statistical comparisons, and the clinico-pathological features of this data set were not different from the larger data set (Appendix 6).

The identical pipeline was performed on both data sets. However, the number of unique neoantigen SNVs was significantly higher in the pilot data set than in the 100KGP data set (median 7119 vs 83 neoantigens per tumour, Wilcoxon test p <2.2e-16). The number of unique indels was also higher in the pilot data set (median 10841 vs 5 indels per tumour, Wilcoxon test p <2.2e-16). The reasons for this discrepancy are unclear.

For the 100KGP data set, the number of unique neoantigen SNVs ranged from 1 to 1713 per tumour, and for indels from 0 to 1040 per tumour. For the pilot set, SNVs ranged from 5130 to 13160 per tumour, and indels ranged from 6837 to 15984 per tumour. In both sets, the distribution of neoantigen values was positively skewed (Shapiro-Wilk test, p <2.2e-16). As with the TMB, MSI-high tumours had significantly higher numbers of neoantigens than MSS tumours (Table 5.2).

**Table 5.2.** Comparison of neoantigen burden in microsatellite stable and microsatellite unstable colorectal cancer

| | Median SNVs per tumour | Median indels per tumour |
|---|---|---|
| MSS CRC (n = 95) | 53 | 4 |
| MSI-high CRC (n = 29) | 5968 * | 400** |

MSI-high tumours have greater neoantigen burden than MSS tumours. Wilcoxon rank sum test. *p = 2.0e-12. **p=1.6e-12. CRC = colorectal cancer, Indels = insertions and deletions, MSI-high = microsatellite instability high, MSS = microsatellite stable, SNVs = single nucleotide variants.

As there were differences between the two sets, the 100KGP data set was analysed separately. Neoantigen burden (both for SNVs and indels) was highly correlated with TMB (R = 0.98, Figure 5.4). However, unlike the TMB, SNV neoantigen burden correlated with the Immunoscore for Low compared with Intermediate and High Immunoscores (Figure 5.5).

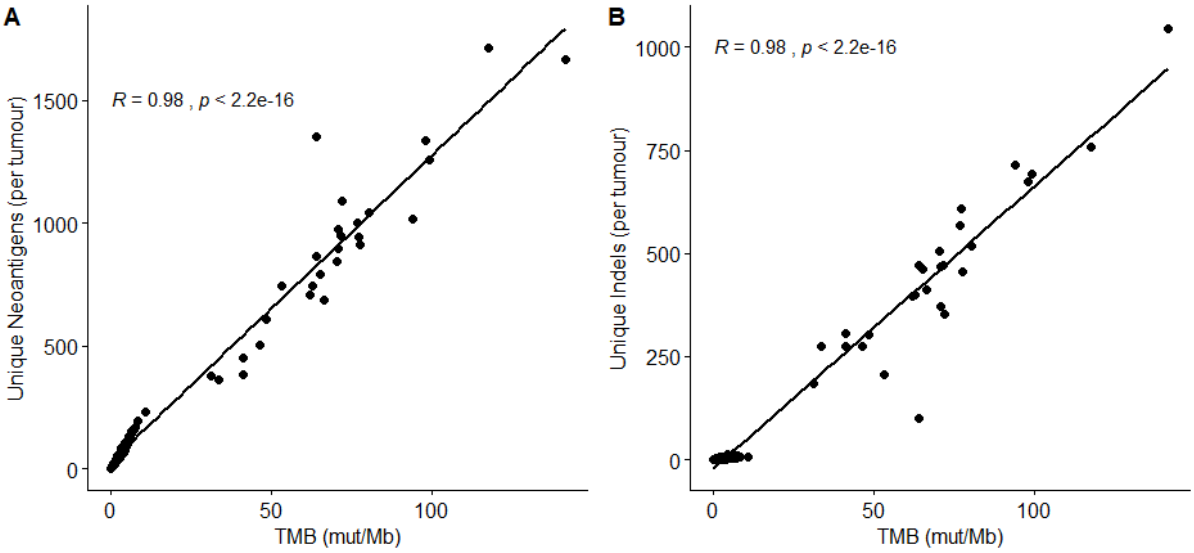



**Figure 5.4.** Scatterplots comparing tumour mutational burden with neoantigen burden for both single nucleotide variants (a, “Unique neoantigens”) and insertions and deletions (b, “Unique indels”). Unique neoantigens and indels are illustrated as some peptides bind to more than one HLA and may be counted more than once in Neopredpipe. TMB = tumour mutational burden, mut/Mb = number of mutations per megabase.

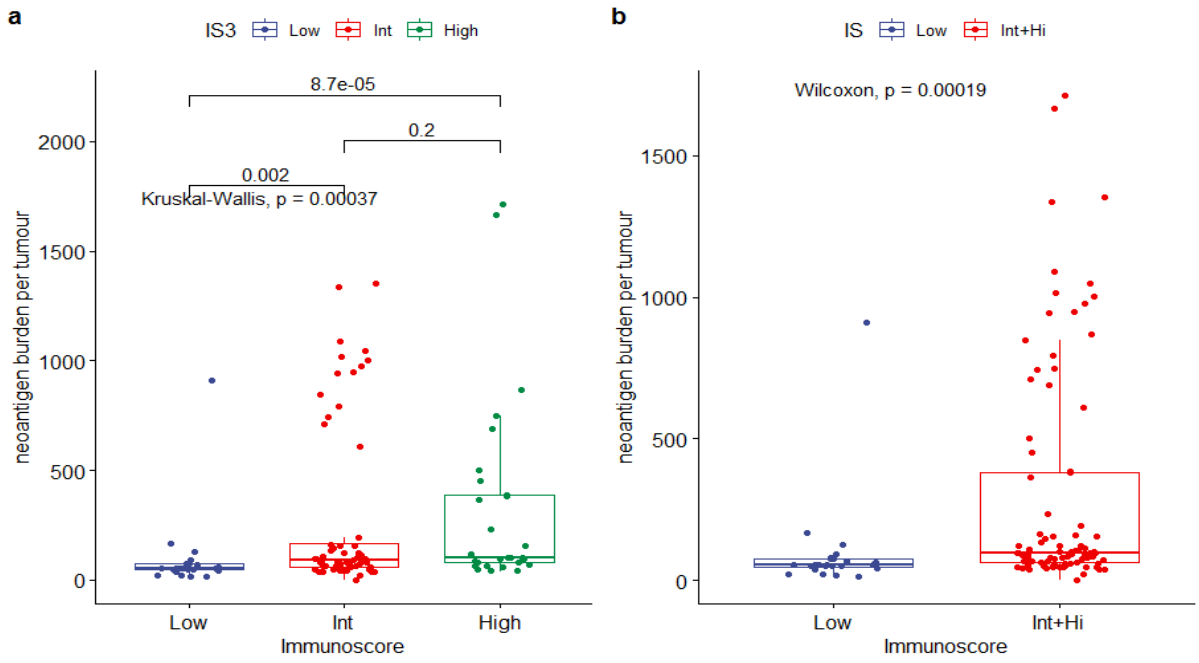



**Figure 5.5.** Boxplots illustrating the associations between neoantigen burden and the Immunoscore (Low, Int, High). a. SNVs are correlated with the Immunoscore, Kruskal-Wallis test IS Low vs High p = 8.7e-05 and Low vs Int Immunoscore p = 0.0002. b. SNVs are correlated with the Immunoscore, Wilcoxon test IS Low vs Int+Hi p = 0.0019. Hi = High, Int = Intermediate, IS = Immunoscore, IS3 = Immunoscore categories, SNV = single nucleotide variant.

Of key significance, these associations were not driven by microsatellite status, as they remained when patients with MSS CRC only were examined (Figure *5.6*).

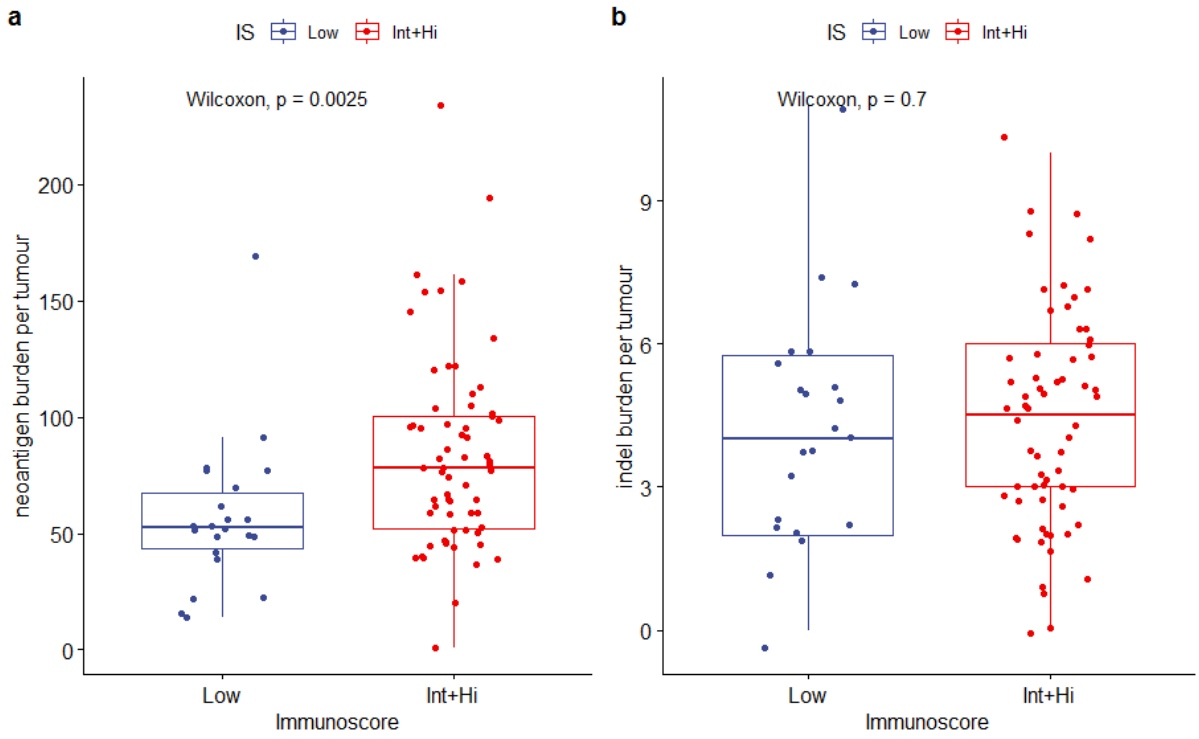



**Figure 5.6.** Boxplots illustrating the associations between neoantigen burden and the Immunoscore (Low, Int+Hi) for microsatellite stable cancers in the cohort. a. SNVs are correlated with the Immunoscore, Wilcoxon test Low vs High Immunoscore p = 0.0011 and Low vs Int Immunoscore = 0.016. b. Insertions and deletions show no significant correlation with the Immunoscore. Hi = High, Int = Intermediate. SNV = single nucleotide variant. IS3 = Immunoscore categories.

A possible explanation for this discrepancy could be the impact of structural variants (including indels and frameshift mutations) within the TMB, and SNVs do not account for this. Information on indel burden was also available from Neopredpipe, so this comparison was made with the Immunoscore. It was observed that for indel neoantigens, the association was weaker and not statistically significant (for Low compared with Intermediate and High Immunoscores, p = 0.057, Figure 5.7). The differences between the associations of SNVs and indels with the Immunoscore (Table 5.3) suggest that these structural variants are represented within the TMB.

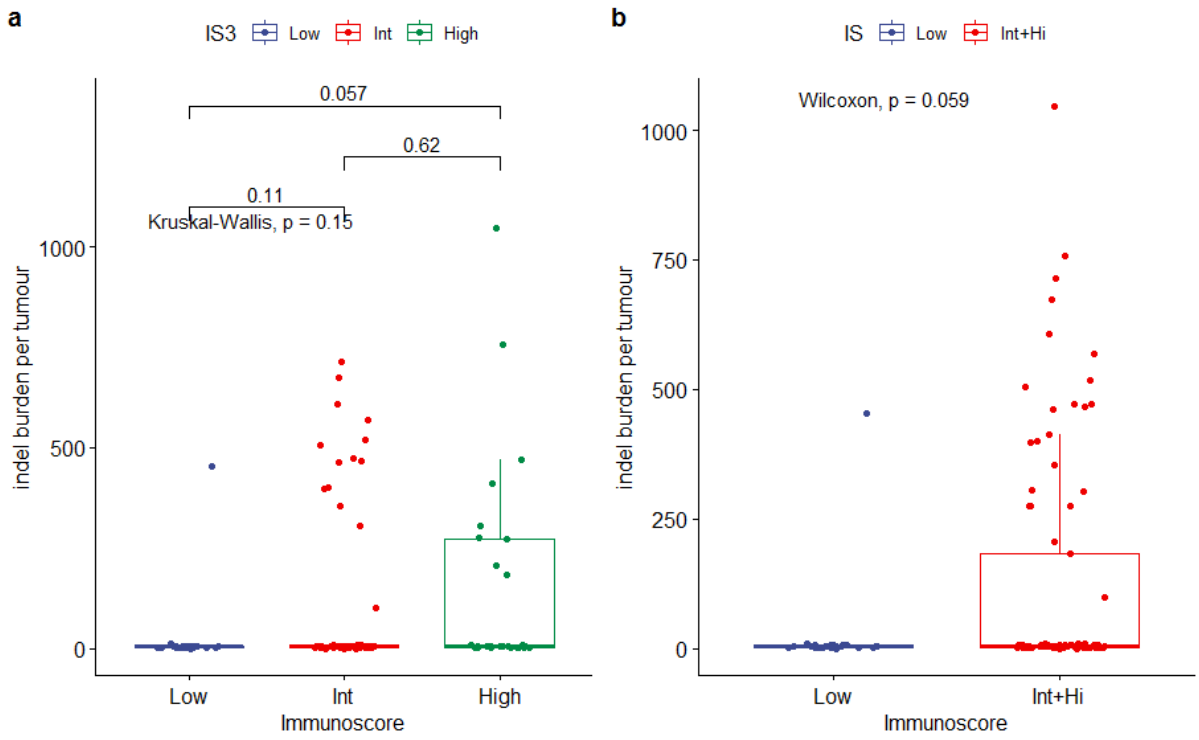



**Figure 5.7.** Boxplots illustrating the associations between indel burden and the Immunoscore (Low, Int, High). a. Indels are not significantly correlated with the Immunoscore, Kruskal-Wallis test IS Low vs High p = 0.057. b. Indels are not significantly correlated with the Immunoscore, Wilcoxon test IS Low vs Int+Hi p = 0.059. Hi = High, Int = Intermediate, IS = Immunoscore, IS3 = Immunoscore categories, Indels = insertions and deletions.

**Table 5.3.** Median neoantigens per tumour by Immunoscore category

| Immunoscore | Median (SNVs) | Median (Indels) |
|---|---|---|
| Low (n = 24) | 53 | 4 |
| Intermediate (n = 60) | 89 | 5 |
| High (n = 29) | 101 | 6 |

Summary of median neoantigens per tumour ranked by Immunoscore category. The difference in neoantigen burden (SNVs) between Immunoscore categories is significant (p=0.0037). The difference in Indel burden is not (p=0.15). Indel = insertion and deletion, SNV=single nucleotide variant.

As with TMB, there was no significant difference in OS in patients with high neoantigen burden (labelled as above the median of all values) compared with those with low neoantigen burden. There was a slight trend towards higher RFS in patients with high neoantigen burden, but this was not statistically significant (Figure *5.8*).

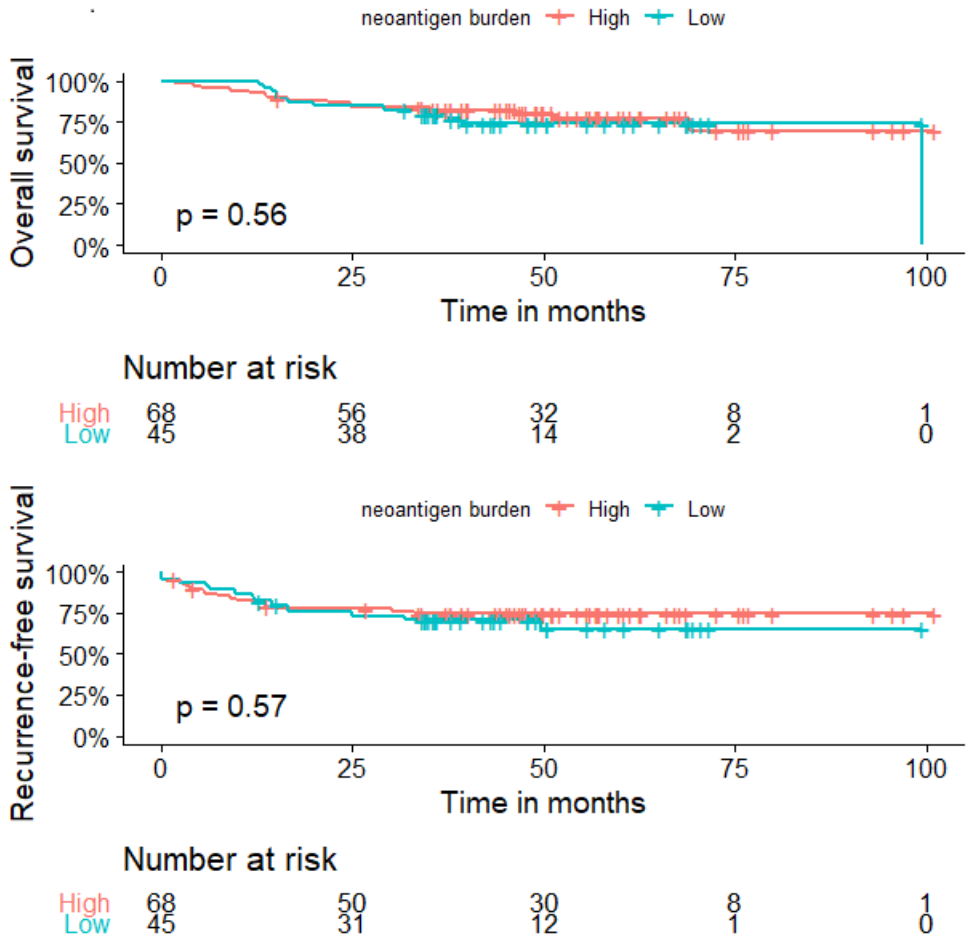



**Figure 5.8.** Kaplan-Meier estimate of overall survival (OS) and recurrence-free survival (RFS) stratified by neoantigen burden. High = greater than median, Low = less than median. There is no difference in OS or RFS between the groups.

## 5.3 Neoantigen clonality and the Immunoscore

The importance of neoantigen clonality in predicting responses to immunotherapy has been demonstrated, notably in a series of lung cancer samples by McGranahan *et al.* [164]. They showed that a combination of neoantigen burden and neoantigen clonality was most informative in determining the response to immunotherapy in lung cancer. In this analysis, TMB and neoantigen burden did not correlate reliably with the Immunoscore and clinical outcomes. Therefore, it was crucially important to compute neoantigen clonality in this patient set to determine its potential significance. Two methods were used for this. The first involved the use of a modified Dirichlet-based clustering approach (DPClust [56]). The second involved calculation of the MATH (mutant allele tumour heterogeneity) score. Associations between these results and the Immunoscore were performed.

### *5.3.1 DPClust*

Copy number analysis (CNA) was performed using DPClust [56]. Neoantigen burden was filtered against the DPClust data, which selected out all neoantigens per sample. The proportion of each neoantigen in each tumour was calculated as a proportion from 0 to 1. Intratumoral heterogeneity (ITH) was determined as the proportion of subclonal neoantigens compared with the total burden.

After filtering, ITH values were obtained for 120 patients in the 100KGP cohort. Ongoing clustering work is being performed on an additional 30 patients in the pilot cohort for whom the neoantigen burden is available and will be presented when available. For the 100KGP samples, values ranged from 0 (with every

neoantigen being clonal) to 1 (representing every neoantigen being subclonal). There was a significant skew towards high ITH, with a mean of 0.9 and a median of 1.0. Many samples had an ITH value of 1 (Figure 5.9). The median ITH for MSS CRC was 1.0 and for MSI-high CRC it was 0.98. This difference was small but significant (Wilcoxon rank sum test, p = 0.006).

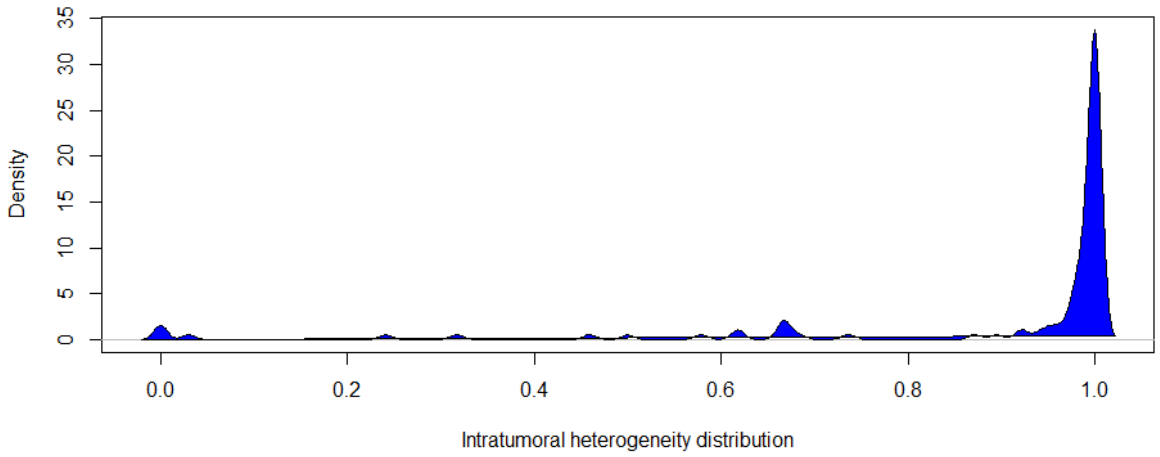



**Figure 5.9.** The distribution of ITH in the data set. The sample was markedly skewed towards a high ITH, with 54.7% of samples having an ITH of 1. Skewness = -2.71.

ITH was correlated with the Immunoscore. Due to calculation errors within the pipeline, results for 14 samples were unavailable, bringing the total number to 106 samples. For these, there was a trend towards an inverse correlation between ITH and the Immunoscore, with decreasing ITH as the Immunoscore increased. However, this was not statistically significant (linear regression R = -0.16, Pearson's product-moment correlation p = 0.102, Figure 5.10). When grouped into the Immunoscore Low and Intermediate combined with High categories, the median Immunoscore for median ITH in the Low group was 1.00

and in the Intermediate and High combined groups was 0.991 (Wilcoxon test, p = 0.31). MSS and MSI-high samples were analysed separately. An inverse association between the Immunoscore and ITH was noted for MSI-high samples but did not reach statistical significance for MSS CRC samples (Figure 5.11).

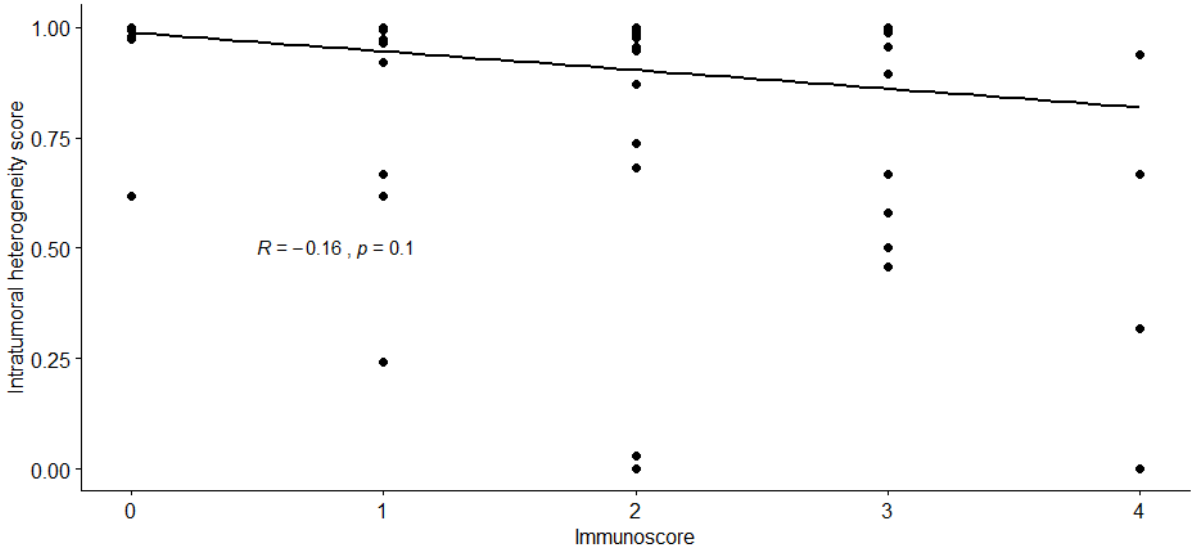



**Figure 5.10.** Distribution of intratumoral heterogeneity (ITH) by Immunoscore for samples in the 100KGP GeL environment (n=106). The Immunoscore is only slightly negatively correlated with ITH by linear regression analysis.

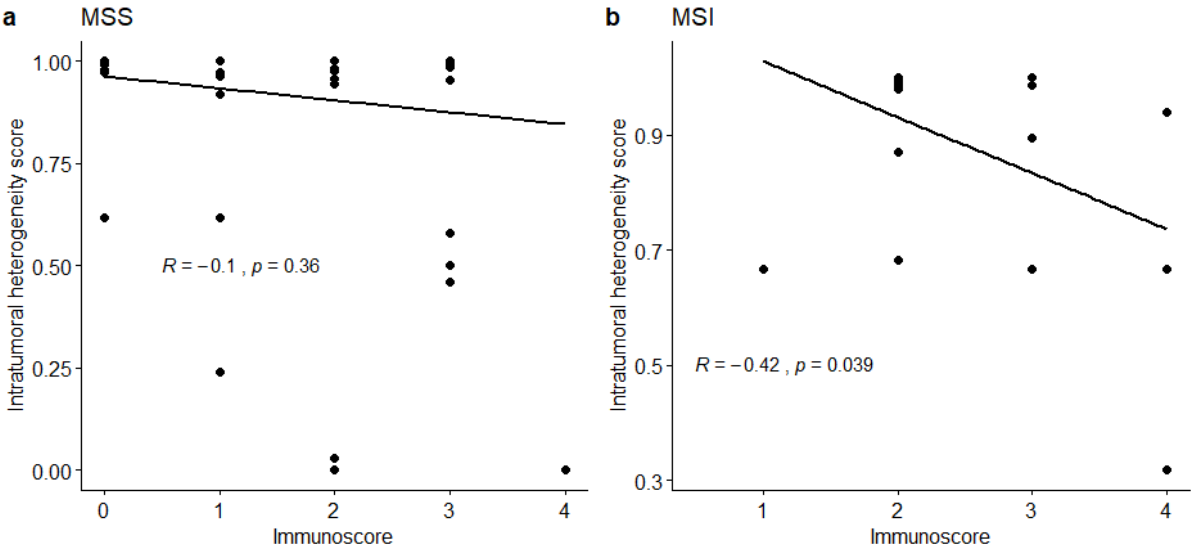



**Figure 5.11.** Distribution of intratumoral heterogeneity (ITH) by Immunoscore for MSS (a) and MSI-high (b) CRC. The Immunoscore is negatively correlated with ITH for MSI-high CRC by linear regression analysis. MSI = microsatellite instability. MSS = microsatellite stable.

A comparison of survival outcome by ITH (stratified into “Low” and “High” based on the average) was performed. No significant difference in either RFS or OS was observed (Figure 5.12). It is most likely that the significant skew towards high ITH reduced the ability to detect a difference between the groups.

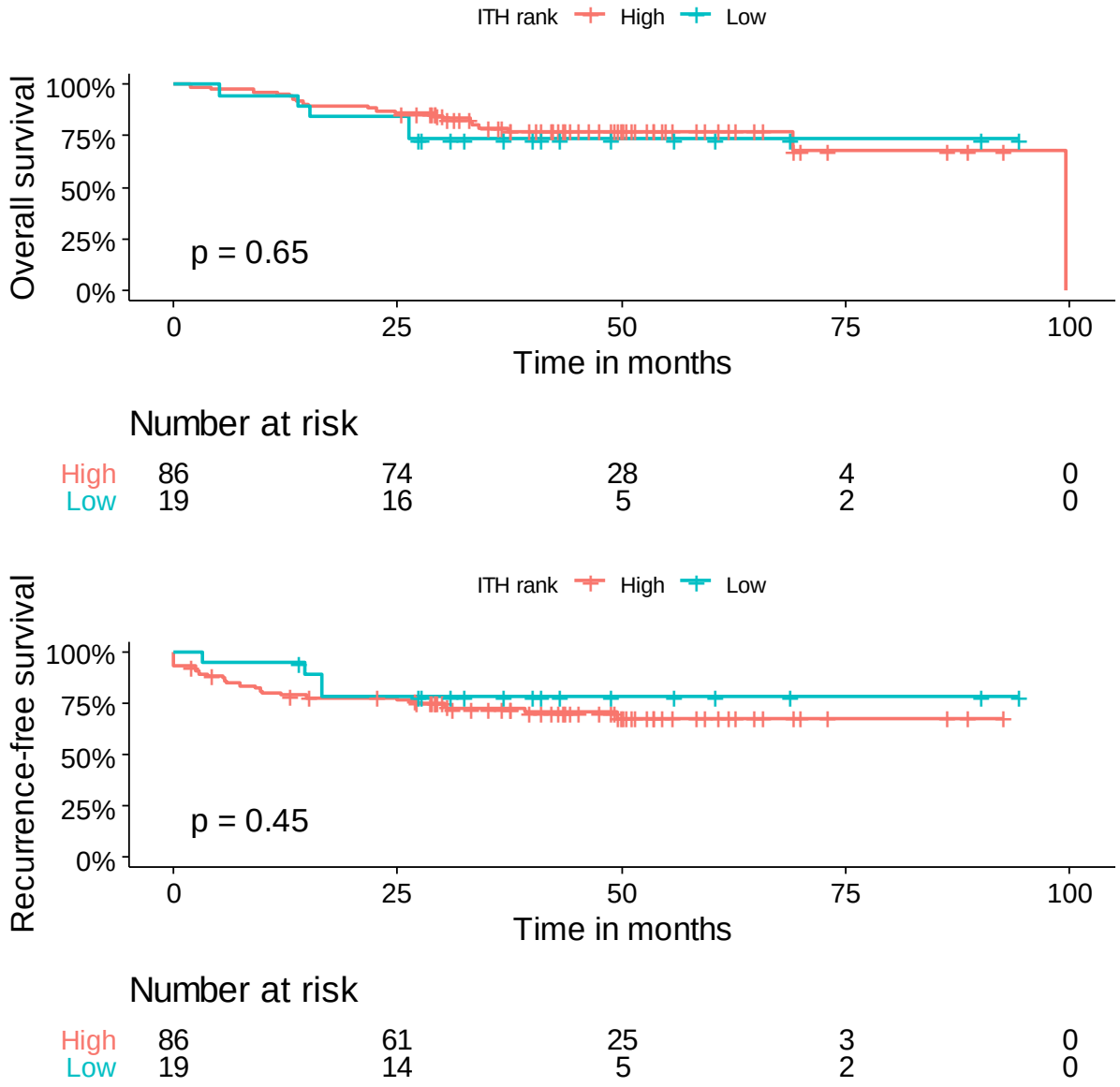



**Figure 5.12.** Kaplan-Meier estimate of overall survival (OS) and recurrence-free survival (RFS) stratified by intratumoral heterogeneity. High = greater than average, Low = less than average. There is no difference in OS between the groups. There is a trend towards a higher RFS in the low ITH burden group but this is not statistically significant. ITH = intratumoral heterogeneity.

### 5.3.2. Combined effects of neoantigen burden and ITH on the Immunoscore

A combination of neoantigen burden and neoantigen clonality was shown to be a better predictor of patient outcomes after immunotherapy in lung cancer [164]. To assess this in my patient group, a combination of ITH and neoantigen burden was performed in this study, and the patients were stratified into three groups. Those with High neoantigen burden and Low ITH were deemed to have a “Good” score. Those with either High neoantigen burden and High ITH, or Low neoantigen burden and Low ITH, were deemed to have an “Intermediate” score. Those with a “Low” neoantigen burden and “High” ITH were deemed to have a “Poor” score.

The Immunoscore was highly correlated with the combined rank (Kruskal-Wallis test, $p = 0.004$, Figure 5.13).

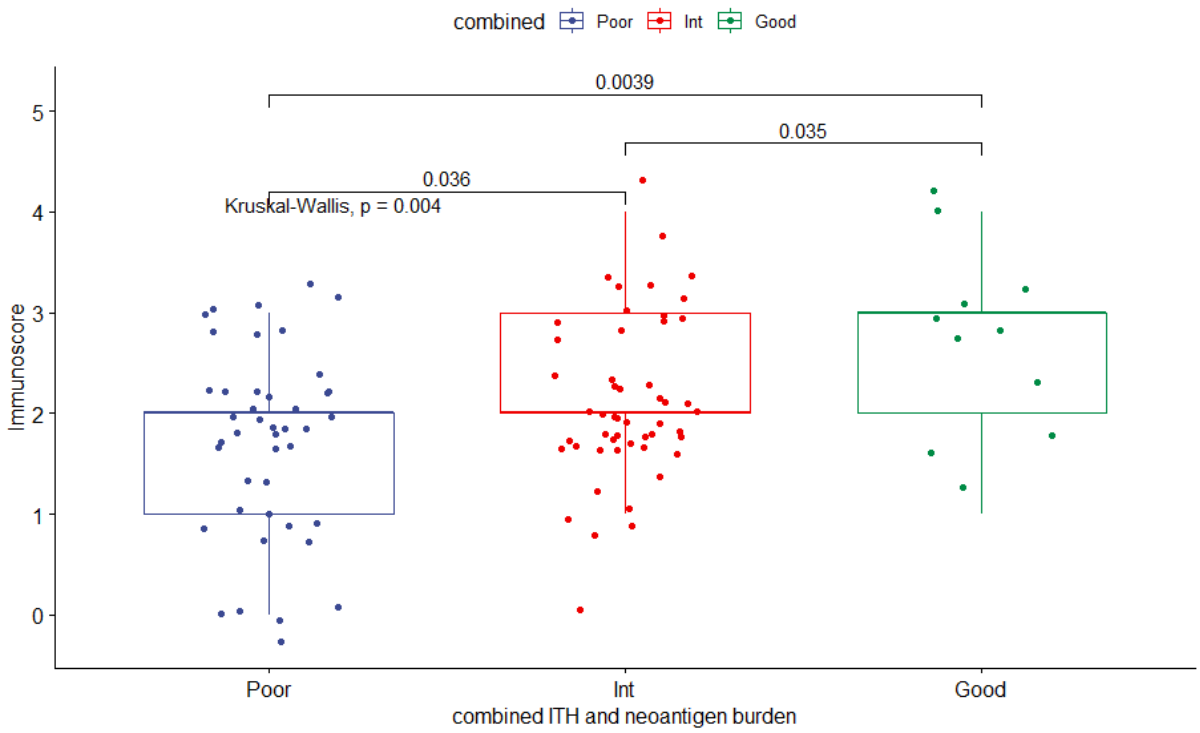



**Figure 5.13.** Boxplots illustrating the associations between the combined neoantigen burden and intratumoral heterogeneity, stratified into three groups, and the Immunoscore. The combined rank (stratified into "Poor", "Intermediate" and "Good") is highly correlated with the Immunoscore. Kruskal-Wallis test, p = 0.004. Int = Intermediate, ITH = intratumoral heterogeneity.

When this combined rank was compared with survival data, the association with RFS was not statistically significant (Figure 5.14). However, the trends seen imply that this is likely to be due to a small sample size and would likely be significant in a larger cohort.

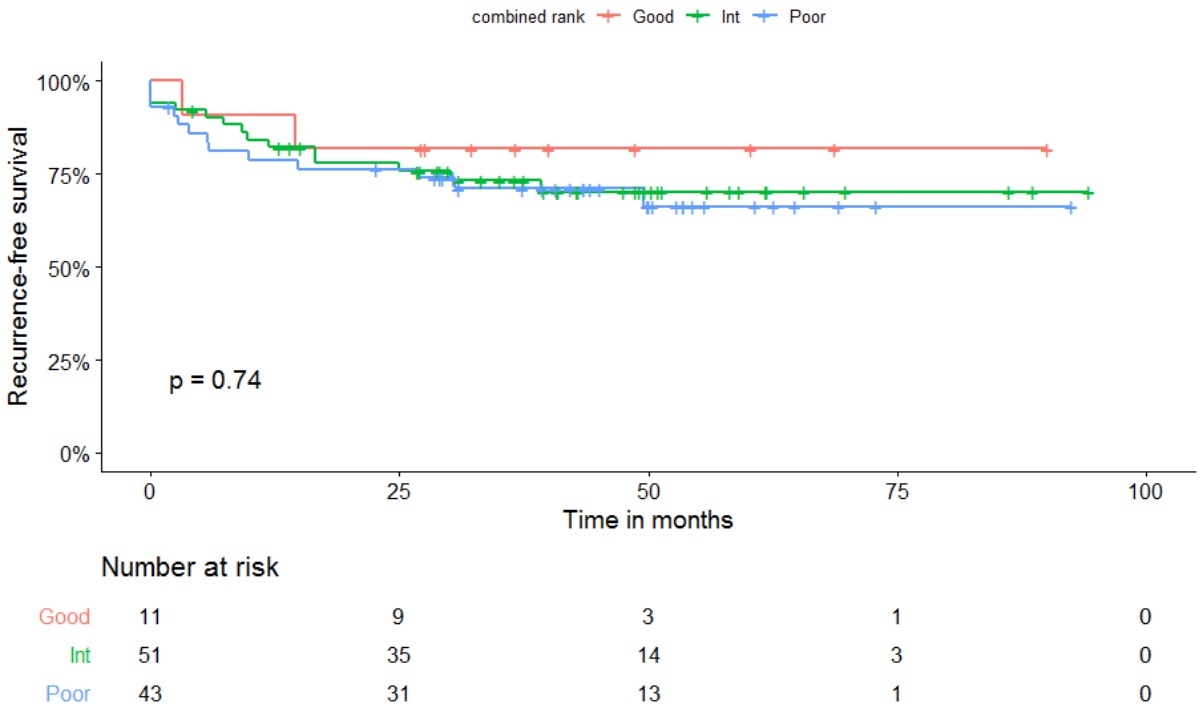



**Figure 5.14.** Kaplan-Meier estimates recurrence-free survival (RFS) stratified by combined ITH and neoantigen burden ranking for all patients. Those with “Poor” and “Int” scores appear to have lower RFS than those with “Good” scores, although this is not statistically significant.

While these results show moderate associations, the combination of clonality and neoantigen burden is highly compelling. It is probable that with cross-validation in a larger sample set, the results would be even more robust.

Furthermore, the DPClust approach, while meticulous, only examines the impact of Class I neoantigens. The role of Class II neoantigens and structural variants could also be highly significant, and a method to examine this was sought.

### 5.3.3. The MATH score

The data from the modified DPClust approach showed a trend towards a negative association between the Immunoscore and ITH. However, there were some weaknesses including data attrition and the compounding of errors in copy number and allele frequency estimations during the calculations.

Therefore, a second method was examined, using the mutant allele tumour heterogeneity (MATH) score, which measures the distribution of all mutant alleles by obtaining a percentage of the ratio of the median to the MAD. This incorporates all mutations and therefore both Class I and II neoantigens. The MATH score has the significant advantage of having demonstrated prognostic [188] and predictive value [194] in a variety of tumour subtypes.

MATH score results were obtained as described in the Methods (Chapter 2.4.2.2. b). Results were available for 145 samples, incorporating both the pilot and 100KGP data sets. The median MATH score was 31.3, with a range from 20.1 to 148.3 (Figure 5.15).

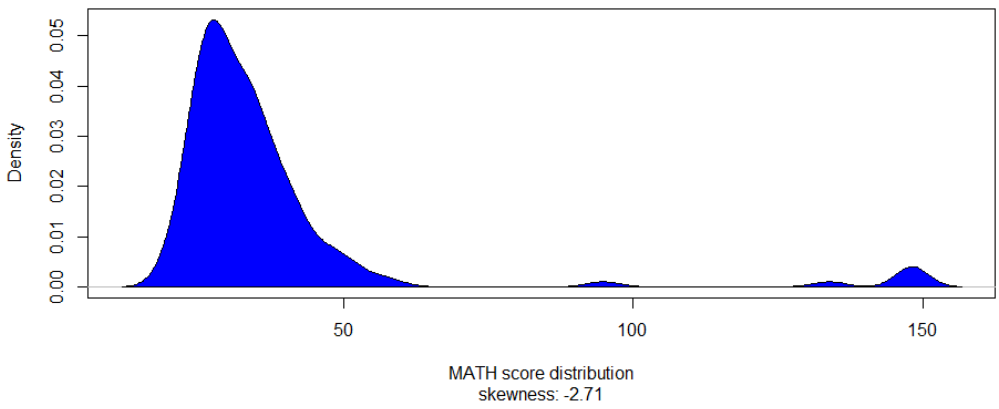



**Figure 5.15.** The MATH (mutant allele tumour heterogeneity score) distribution, which is negatively skewed.

The median MATH score was higher in the MSI-high CRC subset (36.8), compared with the MSS CRC subset (29.2, Wilcoxon rank sum test, p = 0.0005).

This raised the possibility that tumours with high mutation rates simply have greater heterogeneity. However, the MATH score was not significantly associated with the TMB, increasing confidence that the MATH score is not influenced by the number of mutations (R = 0.09, Pearson's product-moment correlation p = 0.3, Figure 5.16).

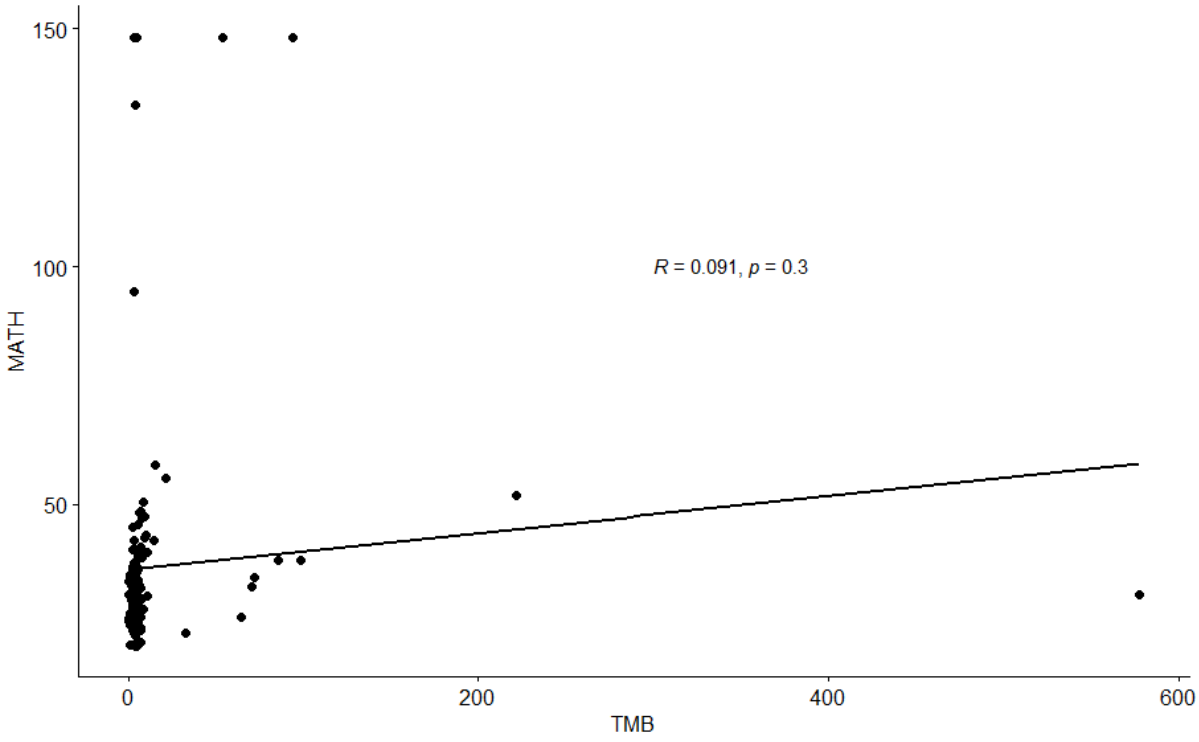



**Figure 5.16.** Comparison of the distribution of the tumour mutational burden (TMB, non-synonymous coding mutations per Mb) and the mutant allele heterogeneity (MATH) score. The MATH score is not influenced by tumour mutational burden.

When four outlying values were removed, the MATH score was significantly inversely correlated with the Immunoscore (R = -0.28, Pearson's product-moment correlation p = 0.0024, Figure 5.17).

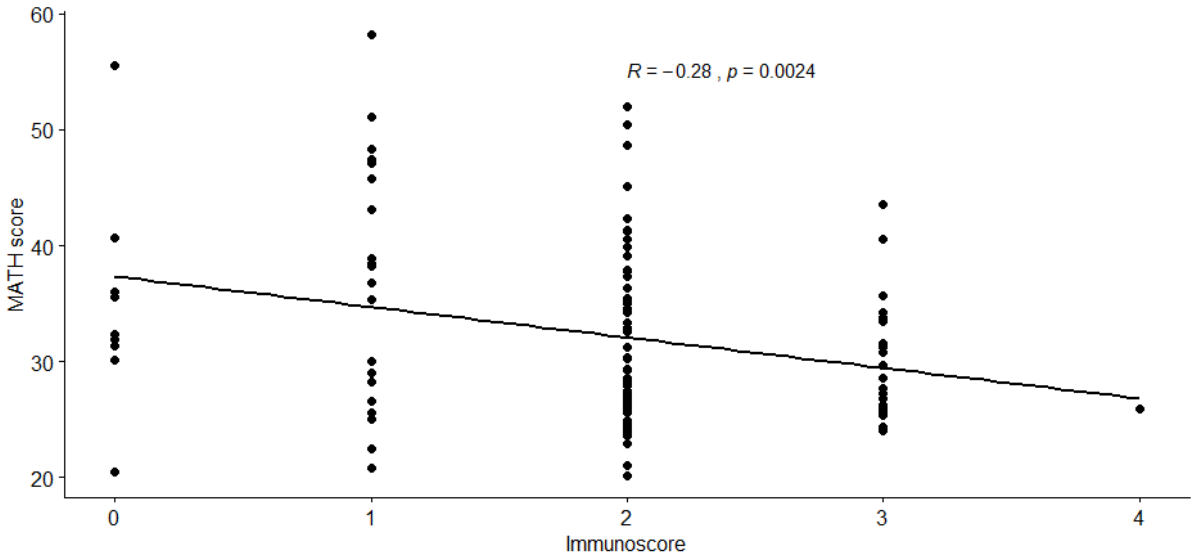



**Figure 5.17.** Distribution of the mutant allele heterogeneity (MATH) score by Immunoscore (outliers excluded). The Immunoscore is significantly negatively correlated with the MATH score by linear regression analysis, p = 0.0024.

This difference was striking when the Immunoscore samples were grouped into Low and combined Intermediate and High groups. The median MATH for IS Low was 36.0, and IS Int + High was 29.7 (Wilcoxon test, p = 0.015). Filtering out the four outlier samples further emphasised this association (Figure 5.18).

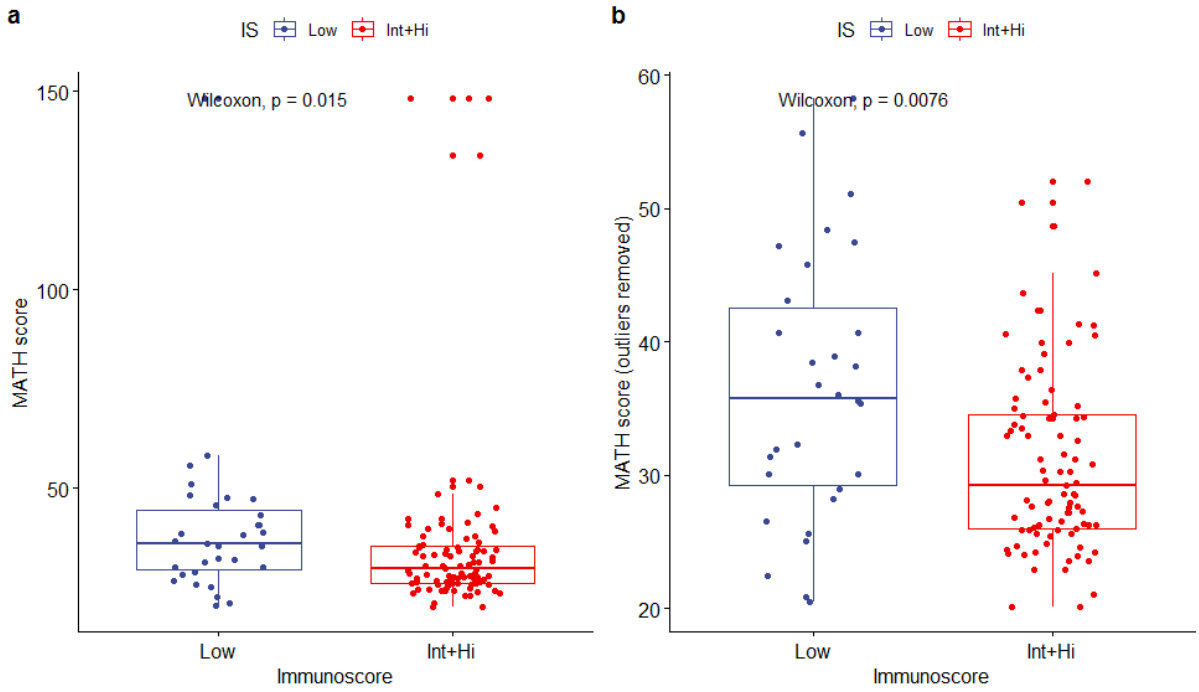



**Figure 5.18.** Comparison of the mutant allele heterogeneity (MATH) score by Immunoscore (IS Low vs IS Int+Hi). The MATH score is significantly inversely correlated with the Immunoscore, (a) for all samples and (b) with outlying samples removed (MATH score >130) excluded.

However, when MSS and MSI-high samples were analysed separately, this association was not significant. In the case of MSI-high CRC, this is most likely due to sample size reduction after filtering (Figure 5.19).

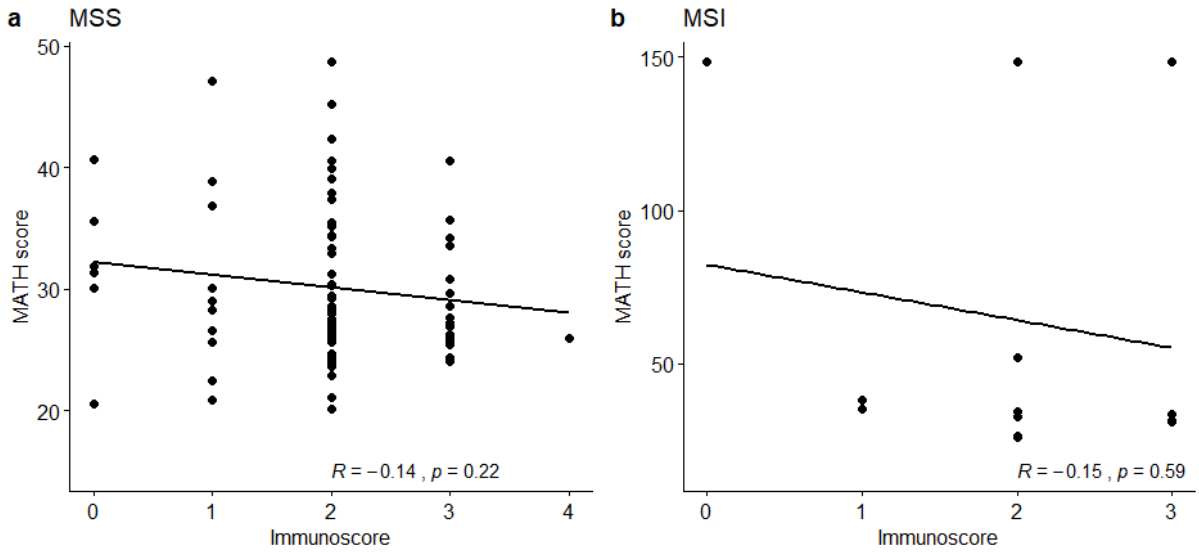



**Figure 5.19.** Distribution of the mutant allele heterogeneity (MATH) score by Immunoscore for MSS (a) and MSI-high (b) CRC. The Immunoscore is not significantly correlated with the MATH score when the data set is subdivided. MSI = microsatellite instability. MSS = microsatellite stable.

Associations with survival were plotted. The samples were categorised as either "High" or "Low" MATH scores, with a cut-off value taken as the median score, as performed by Mroz *et al.* [257]. The median MATH score was 30.4. There was no association seen between the MATH score ranking and either OS or RFS (Figure 5.20).

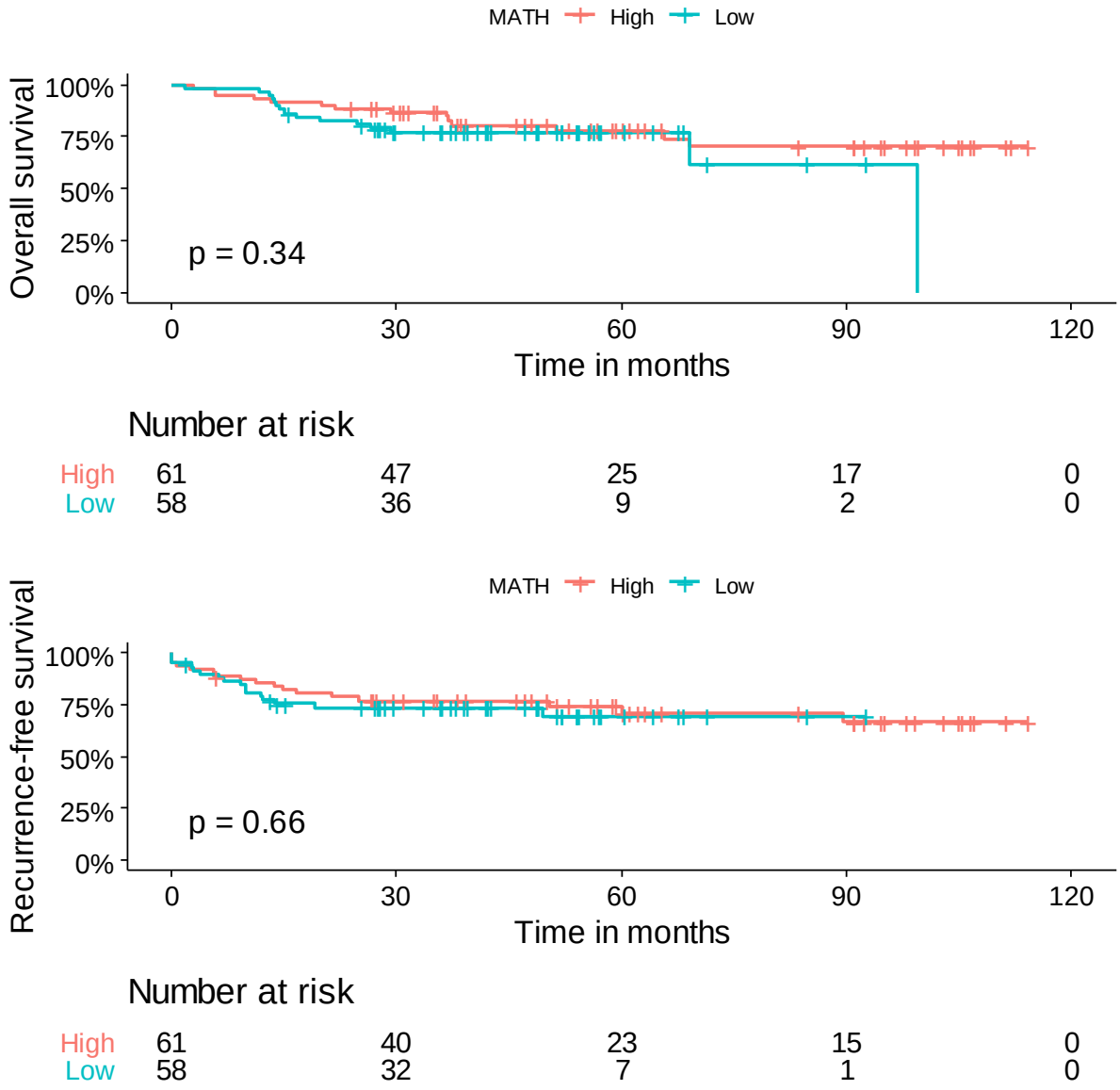



**Figure 5.20.** Kaplan-Meier estimates of (a) overall survival (OS) and (b) recurrence-free survival (RFS) stratified by MATH score ranking for all patients. There is no association between either OS or RFS and the MATH score in this data set. MATH = mutant allele tumour heterogeneity score.

This may have been influenced by other confounding parameters, for example, patient age. However, analysis of multiple confounders showed no significant association between the MATH score and patient age (Pearson's product-moment correlation, p = 0.44), gender (Wilcoxon test, p = 0.58), ethnicity (Kruskal-Wallis test, p = 0.82), disease stage (Kruskal-Wallis test, p = 0.9), tumour T stage (Kruskal-Wallis test, p = 0.67), presence of EMVI (Wilcoxon test, p = 0.34), disease site (Kruskal-Wallis test, p = 0.33), or whether they had neoadjuvant therapy (Wilcoxon test, p = 0.41).

Dissecting the MATH score to consider separately the numbers of clonal and non-clonal mutations per tumour showed strong association between the number of clonal mutations and the Immunoscore, and inverse association between the number of non-clonal mutations and the Immunoscore (Pearson's product-moment correlation for both, Figure 5.21).

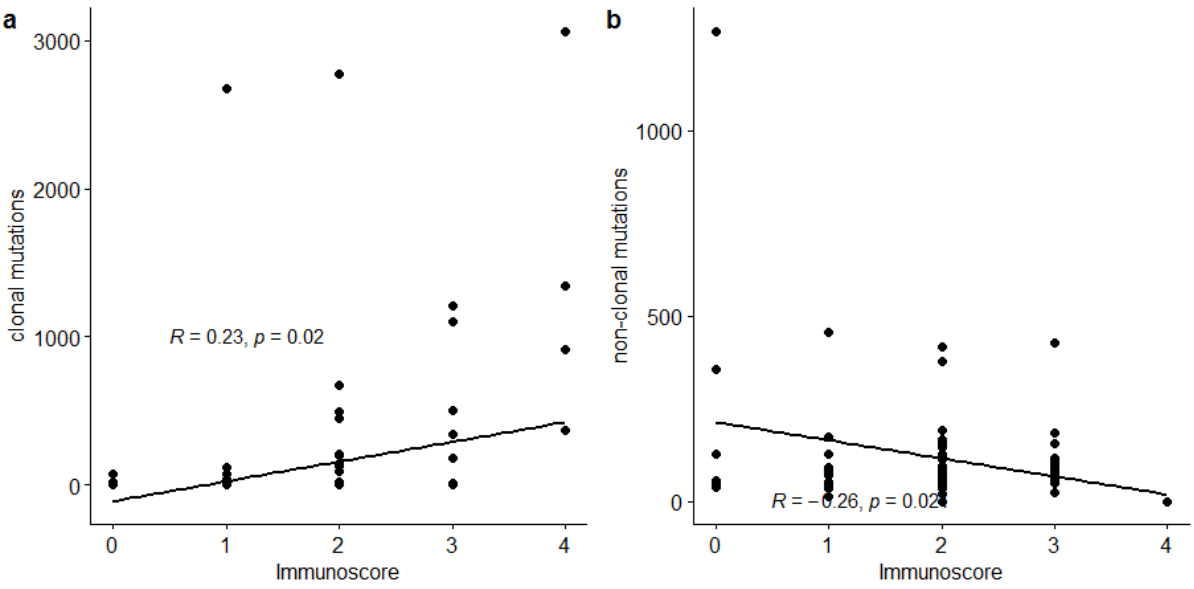



**Figure 5.21.** Comparison of number of mutations with the Immunoscore. (a) Number of clonal mutations is correlated with the Immunoscore, p = 0.02. (b) Number of non-clonal mutations is inversely associated with the Immunoscore, p = 0.02.

## 5.4 Somatic immune gene expression and the immune environment

It was important to corroborate the associations found with transcriptomic data. To this end, 3' RNA sequencing was performed using RNA extracted from formalin fixed tissue. Of the 197 patients, combined Immunoscore samples and RNAseq data were available for 190 patients. Analysis was performed in Partek® Flow©.

### 5.4.1. Gene set analysis

Mitochondrial and ribosomal genes were filtered out and a correction for batch effect was made using the Partek algorithm (Figure 5.22).

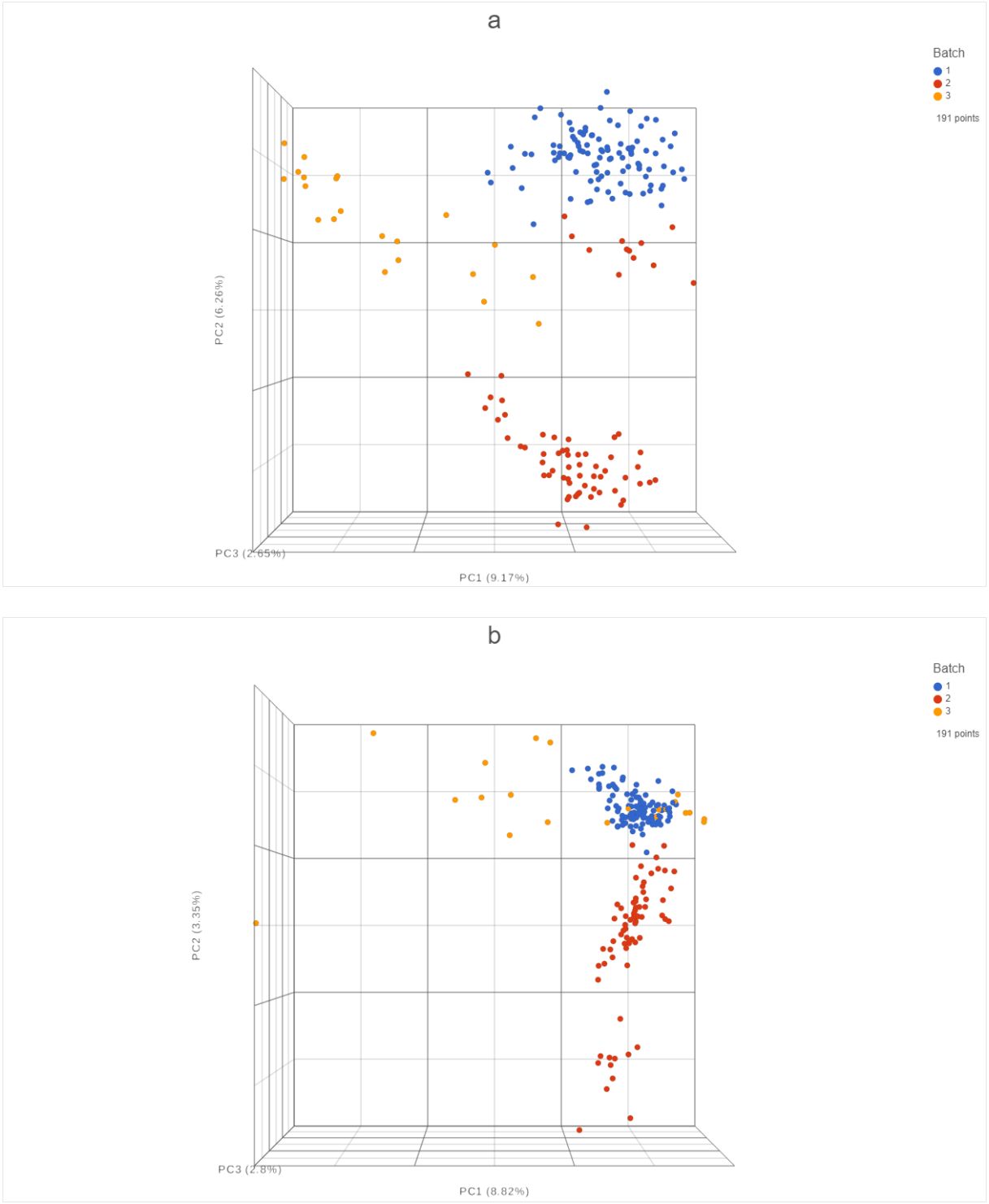
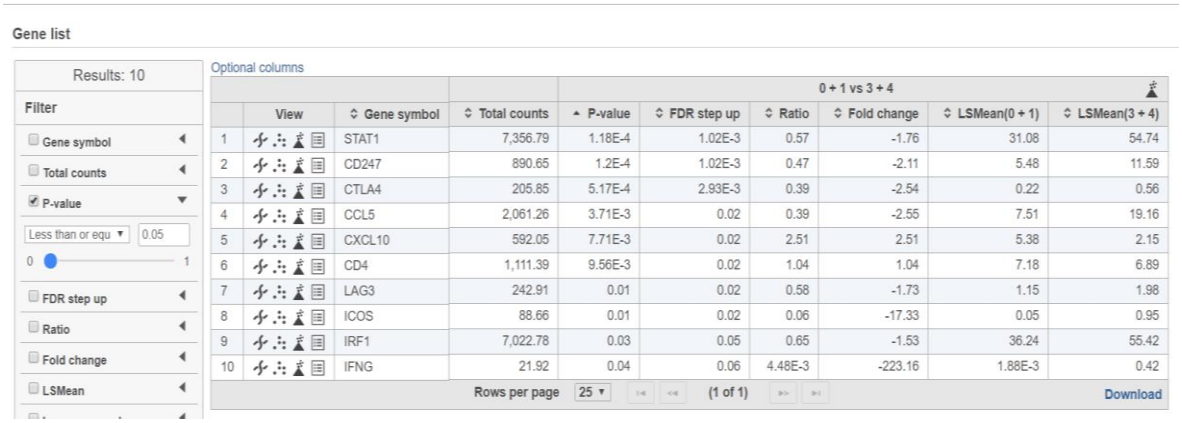



**Figure 5.22.** Principal components analysis of data nodes before and after adjustment for batch effect. (a) Raw data (b) following adjustment for Batch and Batch*Immunoscore.

Analysis of the association of gene expression against the Immunoscore by gene set analysis (GSA) showed several genes with expression differences most strongly correlated with the Immunoscore. A filter was applied to assess genes in the extended MuTHER eQTL list (Figure 5.23). When Low versus High Immunoscores were compared, the top gene associated was STAT-1, with increased expression associated with a higher Immunoscore (Kruskal-Wallis test, p = 8.37e-11).


Gene list

Results: 10

Filter

Gene symbol

Total counts

P-value

Less than or equ 0.05

0 1

FDR step up

Ratio

Fold change

LSMean

Optional columns

| | | | | 0 + 1 vs 3 + 4 | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | View | Gene symbol | Total counts | P-value | FDR step up | Ratio | Fold change | LSMean(0 + 1) | LSMean(3 + 4) |
| 1 | | STAT1 | 7,356.79 | 1.18E-4 | 1.02E-3 | 0.57 | -1.76 | 31.08 | 54.74 |
| 2 | | CD247 | 890.65 | 1.2E-4 | 1.02E-3 | 0.47 | -2.11 | 5.48 | 11.59 |
| 3 | | CTLA4 | 205.85 | 5.17E-4 | 2.93E-3 | 0.39 | -2.54 | 0.22 | 0.56 |
| 4 | | CCL5 | 2,061.26 | 3.71E-3 | 0.02 | 0.39 | -2.55 | 7.51 | 19.16 |
| 5 | | CXCL10 | 592.05 | 7.71E-3 | 0.02 | 2.51 | 2.51 | 5.38 | 2.15 |
| 6 | | CD4 | 1,111.39 | 9.56E-3 | 0.02 | 1.04 | 1.04 | 7.18 | 6.89 |
| 7 | | LAG3 | 242.91 | 0.01 | 0.02 | 0.58 | -1.73 | 1.15 | 1.98 |
| 8 | | ICOS | 88.66 | 0.01 | 0.02 | 0.06 | -17.33 | 0.05 | 0.95 |
| 9 | | IRF1 | 7,022.78 | 0.03 | 0.05 | 0.65 | -1.53 | 36.24 | 55.42 |
| 10 | | IFNG | 21.92 | 0.04 | 0.06 | 4.48E-3 | -223.16 | 1.88E-3 | 0.42 |

Rows per page 25 (1 of 1) Download

**Figure 5.23.** Gene set analysis of immune gene expression by the Immunoscore. The top gene associated with the Immunoscore was STAT1, with other genes, notably CD247, CTLA4, CCL5, CXCL10, CD4, LAG3, ICOS, IRF1 and IFNG also significantly associated with the Immunoscore.

Hierarchical clustering showed other genes showing strong associations with the Immunoscore, notably STAT-1, TGFβ1, TRAC, PARP14, GBP1, APOL6, PSMB9, CMC1, KHL18, CD3D, DCAF17, CD27, DUS2 and QSERP (Figure 5.24).

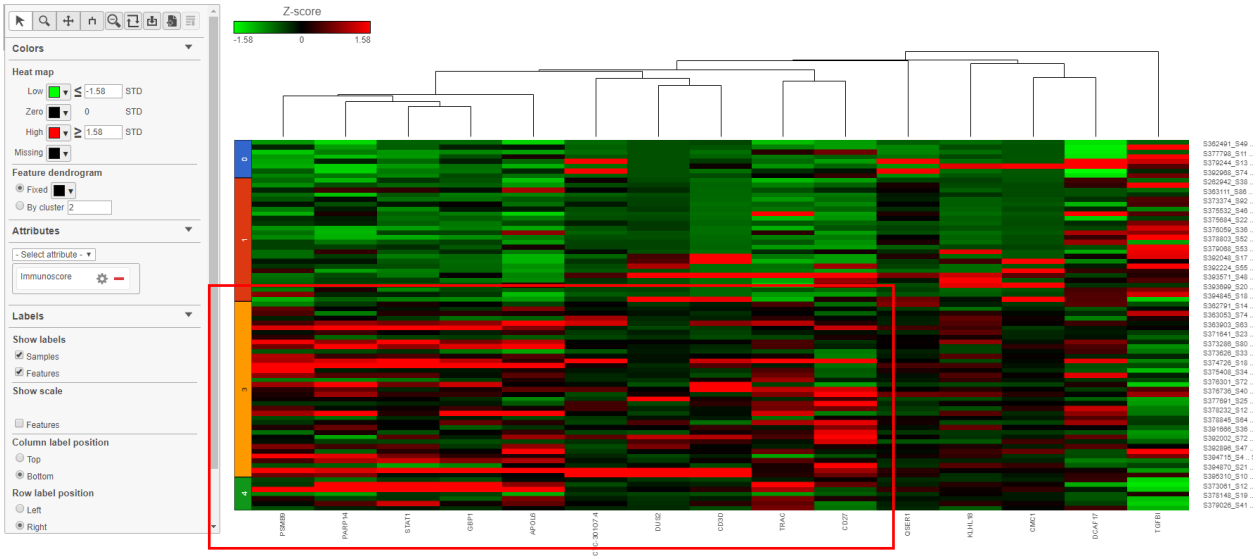



**Figure 5.24.** Heatmap showing differential gene expression by Immunoscore (y axis 0 to 4). Increased expression of PSMB9, PARP14, STAT1, GBP1, APOL6, DUS2, CD3D, TRAC and CD27 are seen at high Immunoscores (bottom left, emphasised by red square) while increased expression of TGFβ1 is seen at low Immunoscores (top left).

### 5.4.2. The co-ordinate immune response cluster (CIRC)

Data on expression of the 28 genes that make up the co-ordinate immune response cluster was collated (Table 5.4). The expression of each gene was log transformed, and the mean cluster expression z score was calculated for each sample as described by Lal et al. [92] (Figure 5.25).

**Table 5.4.** The co-ordinate immune response cluster genes

| Gene ID |
|---|
| CCL5 |
| CD247 |
| CD274 |
| CD4 |
| CD80 |
| CTLA4 |
| CXCL10 |
| CXCL9 |
| GNLY |
| HAVCR2 |
| HLA-DMA |
| HLA-DMB |
| HLA-DOA |
| HLA-DPA1 |
| HLA-DPB1 |
| HLA-DQA1 |
| HLA-DQA2 |
| HLA-DRA |
| HLA-DRB5 |
| ICAM1 |
| ICOS |
| IFNG |
| IL18RAP |
| IRF1 |
| LAG3 |
| PDCD1LG2 |
| STAT1 |
| TBX21 |

From Lal *et al.* 2015 [92].

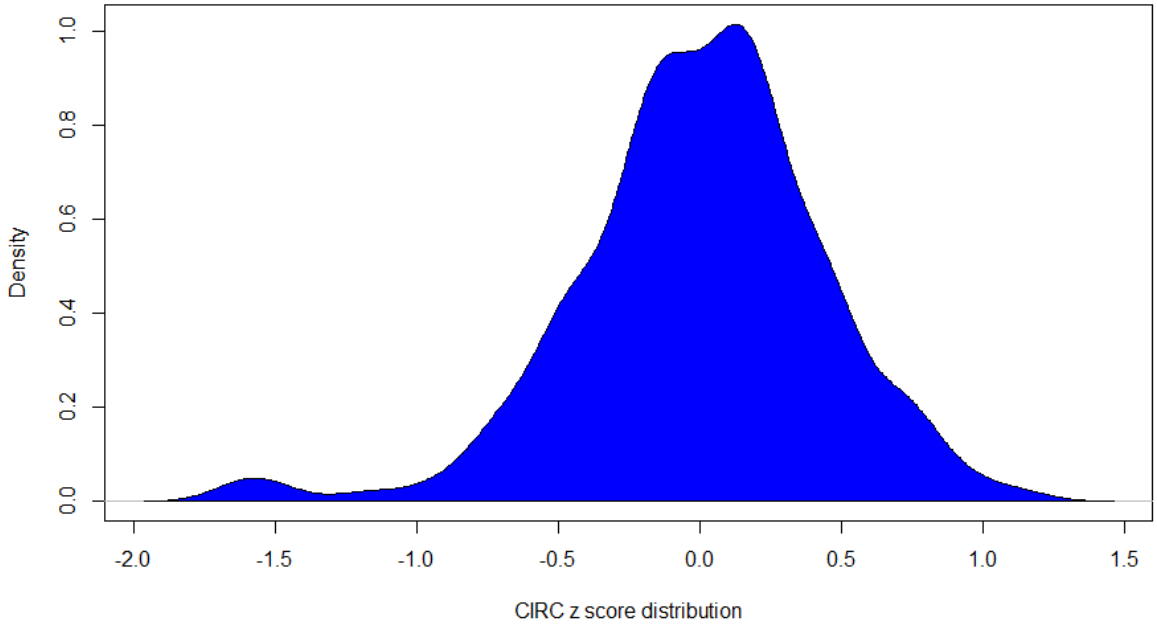



**Figure 5.25.** The distribution of the co-ordinate immune response cluster (CIRC) gene expression values for all samples. Log transformed and z scored, skewness = -0.62, Shapiro-Wilk normality test p = 0.0004.

The median CIRC score was higher for MSI (median z score = 0.35) than MSS tumours (median z score = -0.07, Wilcoxon rank sum test, p = 2.18e-06).

The CIRC score was correlated with the Immunoscore (Table 5.5). The median CIRC increased with Immunoscore category (Figure 5.26). When the IS Low was compared with the combined IS Int and Hi categories, this difference was markedly significant (Kruskal-Wallis test, p = 0.0006, Figure 5.27).

**Table 5.5.** Distribution of median and mean z scores for the CIRC by the Immunoscore

| Immunoscore (number) | Median | Interquartile range | Mean | Standard deviation |
|---|---|---|---|---|
| Low (n=45) | -0.11 | 0.52 | -0.22 | 0.49 |
| Intermediate (n=100) | 0.09 | 0.48 | 0.03 | 0.41 |
| High (n=45) | 0.12 | 0.51 | 0.13 | 0.36 |

The mean z score for the CIRC rises with each Immunoscore category. CIRC = co-ordinate immune response cluster

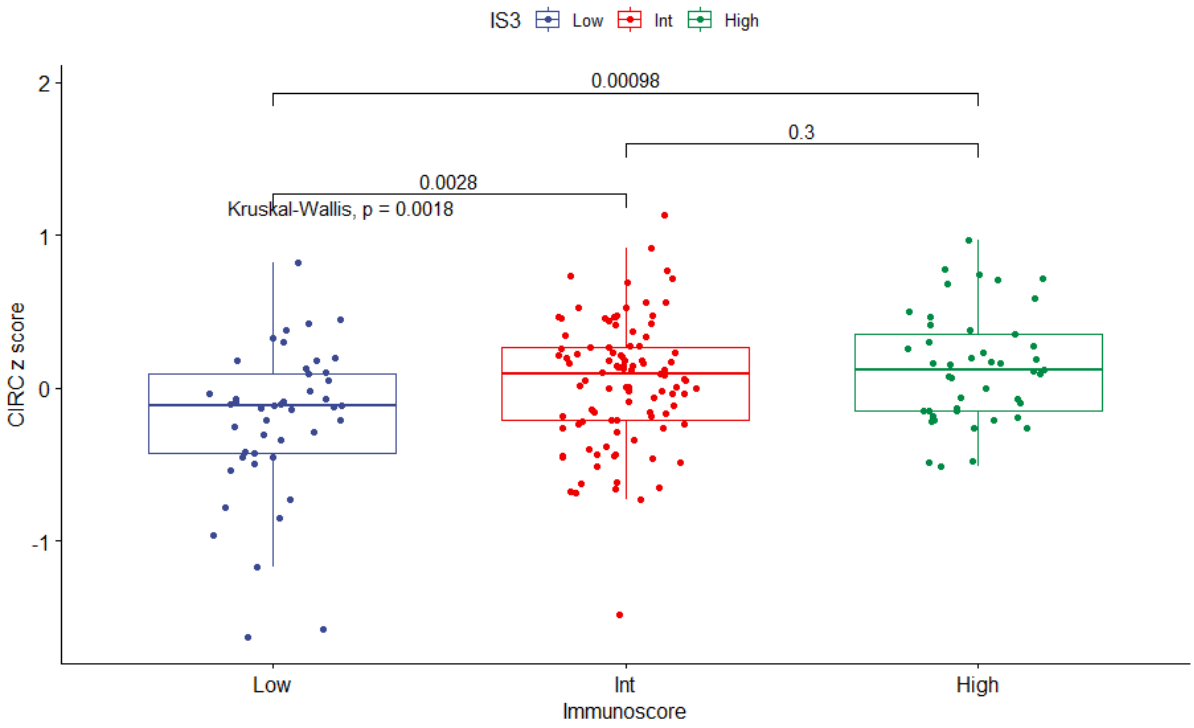



**Figure 5.26.** Boxplots illustrating the associations between the CIRC (co-ordinate immune response cluster) score and the Immunoscore (Low, Int, High). The CIRC score is correlated with the Immunoscore for Low compared with the Intermediate and High categories. Kruskal-Wallis test, p = 0.001. Int = Intermediate, IS3 = Immunoscore categories.

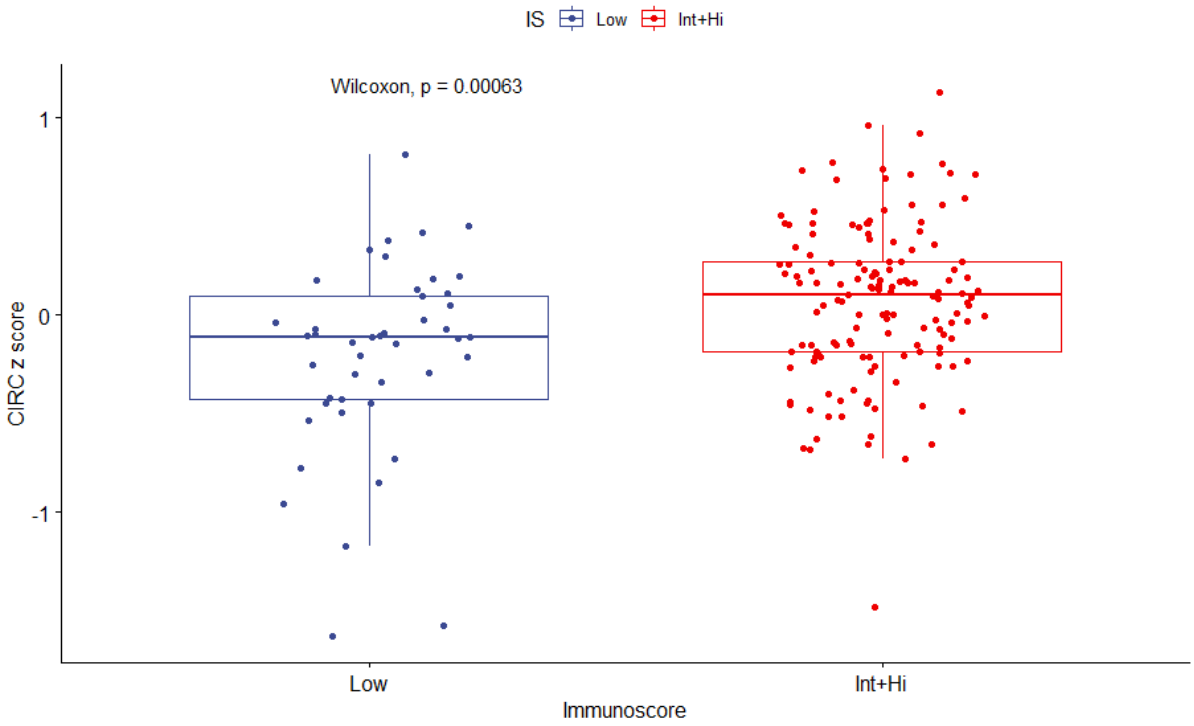



**Figure 5.27.** Boxplots illustrating the associations between the CIRC (co-ordinate immune response cluster) score and the Immunoscore (Low, Int + Hi). The CIRC score is correlated with the Immunoscore for Low compared with the Intermediate and High categories. Kruskal-Wallis test, p = 0.00063. Hi = High, Int = Intermediate, IS = Immunoscore categories

The CIRC score principally incorporates Th1- centric genes, such as TBX21, IFNG, IRF1 and STAT1, which are known to drive anti-cancer immunity. The score also incorporates immune checkpoint genes which are also upregulated in the presence of strong constitutive Th1 expression, so this is an expected finding.

The association between the CIRC score and the Immunoscore persisted even when filtered out for MSS tumours only (Figure 5.28).

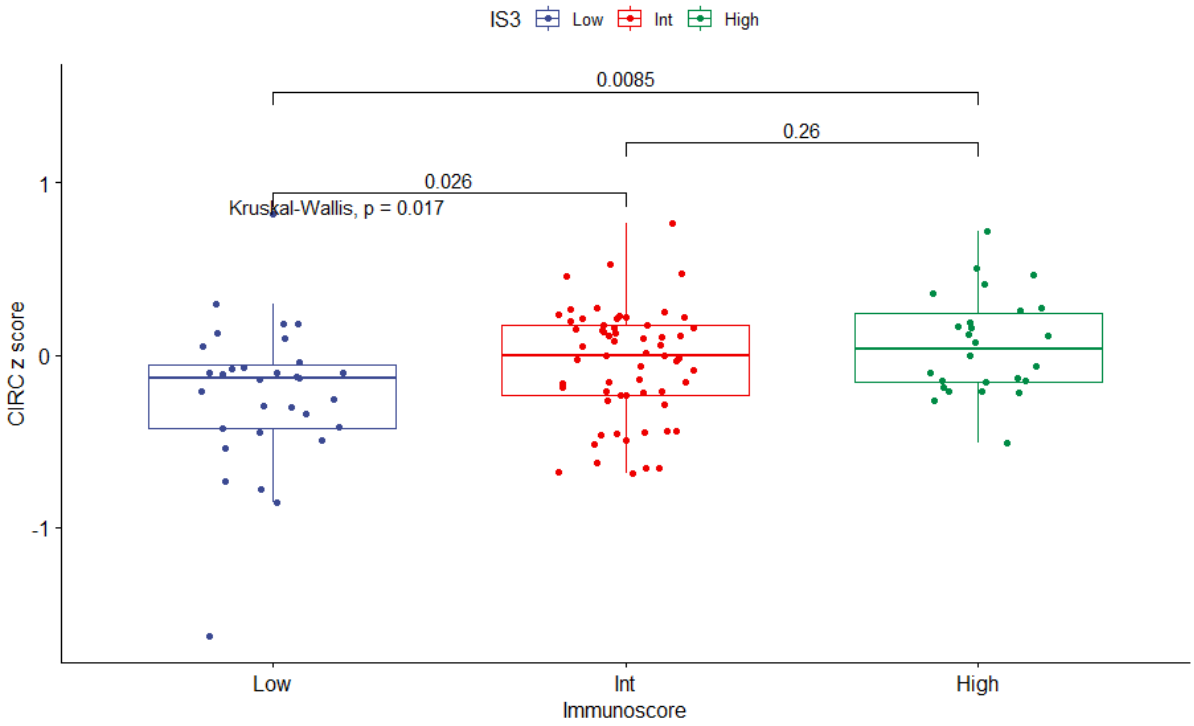



**Figure 5.28.** Boxplots illustrating the associations between the CIRC (co-ordinate immune response cluster) score and the Immunoscore (Low, Int, High) for microsatellite stable tumours only (n=118). The CIRC score is correlated with the Immunoscore for Low compared with the Intermediate and High categories. Kruskal-Wallis test, p = 0.017. Int = Intermediate, IS3 = Immunoscore categories.

The CIRC score was compared with survival outcomes in all patients (both MSS and MSI-high). The threshold for a 'High' CIRC score was defined as the median of the score in the MSI-high group. There was a trend towards an increased RFS in the group defined as a 'High' CRC, although this was not statistically significant (Cox proportional hazards model, p = 0.075, Figure 5.29). There was no association seen for OS (p = 0.202, Figure 5.30).

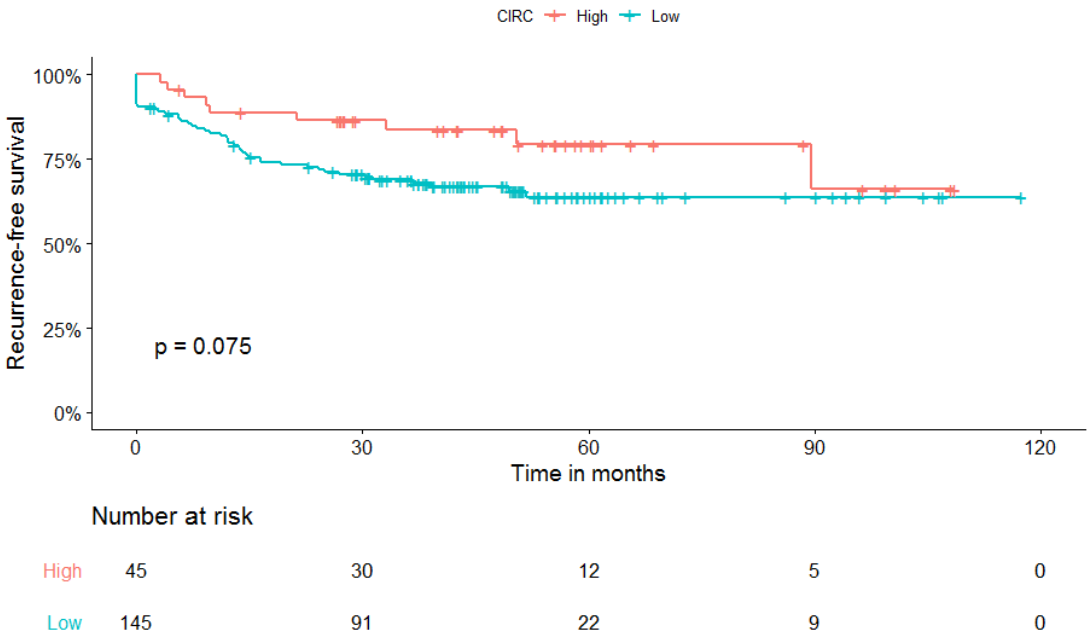



**Figure 5.29.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by CIRC z score (high or low) for all patients. There is a trend towards increased RFS with a High CIRC score, but this is not statistically significant. Cox proportional hazards, p = 0.075. Hazard ratio Low vs High CIRC = 1.90 (95% CI 0.93 – 3.90). CI = confidence interval, CIRC = co-ordinate immune response cluster.

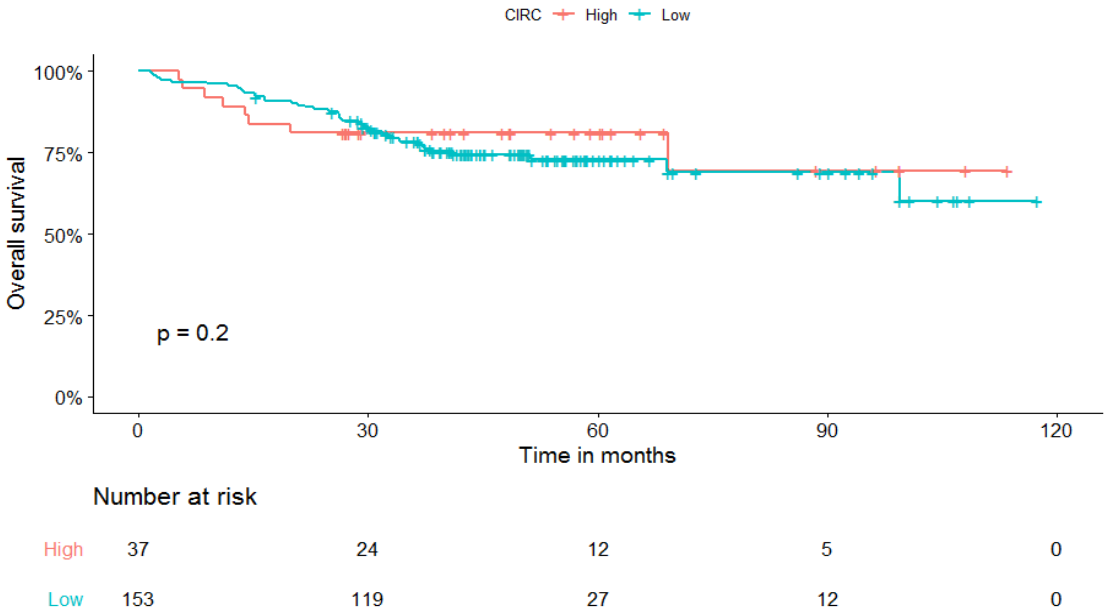



**Figure 5.30.** Kaplan-Meier estimate of overall survival (OS) stratified by CIRC z score (high or low). There is no significant difference in OS between those with High and Low CIRC scores. Cox proportional hazards, p = 0.202. CIRC = co-ordinate immune response cluster.

### 5.4.3. MHC Class II gene expression

MHC Class II (HLA-DP -DQ and -DR) expression signatures were also calculated, log transformed and normalised to z scores. The CIRC includes Class II genes amongst its other signatures. Therefore, as expected, Class II RNA expression correlated with the Immunoscore (Figure 5.31). Class II expression was higher in MSI-high tumours than MSS tumours, with median z score for MSI-high tumours 0.33 compared with median z score for MSS tumours of -0.04, Wilcoxon test p = 0.0006. This association however disappeared when filtered out for MSS tumours only (Figure 5.32).

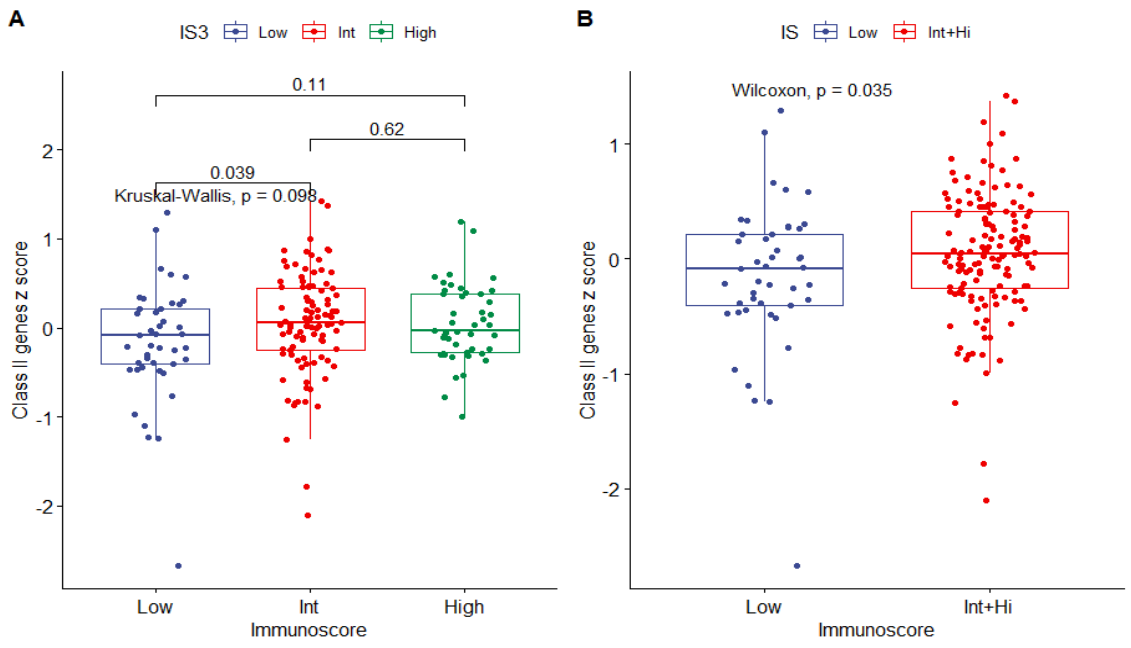



**Figure 5.31.** Boxplots illustrating the associations between the MHC Class II gene expression (HLA-DP, -DQ, and -DR) score and the Immunoscore (Low, Int, High). This is correlated with the Immunoscore for Low compared with High tumours, Kruskal-Wallis test, p = 0.035. Int = Intermediate. IS3 = Immunoscore categories.

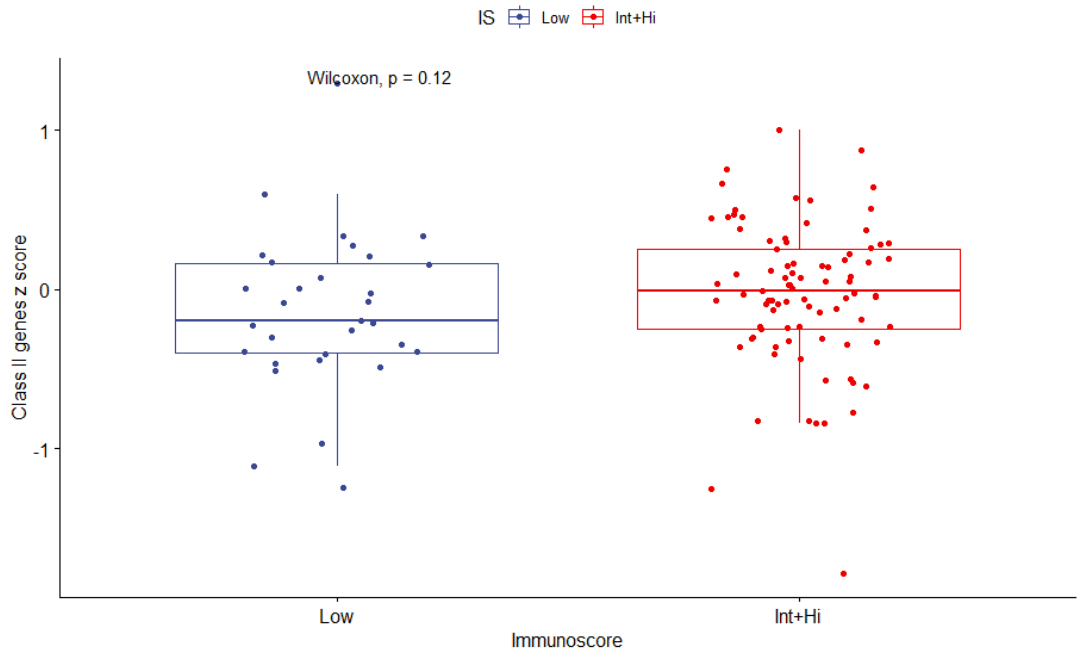



**Figure 5.32.** Boxplots illustrating the associations between the MHC Class II gene expression (HLA-DP, -DQ, and -DR) score and the Immunoscore (Low, Int, High) for microsatellite stable tumours only (n=118). There is not a statistically significant association with the Immunoscore. Wilcoxon test, p = 0.12. Hi = High. Int = Intermediate. IS = Immunoscore categories.

### 5.4.4. Chemokine expression

Other immune signatures and their correlation with the Immunoscore were explored. In particular, the signatures of a panel of gut bacteria-stimulated chemokines shown to drive T cell recruitment into tumour tissues [205] were analysed.

These chemokines were in four groups: those associated with cytotoxic/Th-1 function (CCL5, CXCL9 and CXCL10), those associated with regulatory T-cell/Th-1 functions (CCL17, CCL22 and CXCL12), those associated with follicular Th cells (CXCL13) and those associated with IL-17-producing Th cells (CCL20 and CCL17) [205].

Of these, the cytotoxic (Th-1)-associated signatures showed the strongest associations with the Immunoscore (Kruskal-Wallis $p = 0.00032$, Figure 5.33), with increased combined z score in the Intermediate and High combined Immunoscore group compared with the Low Immunoscore group.

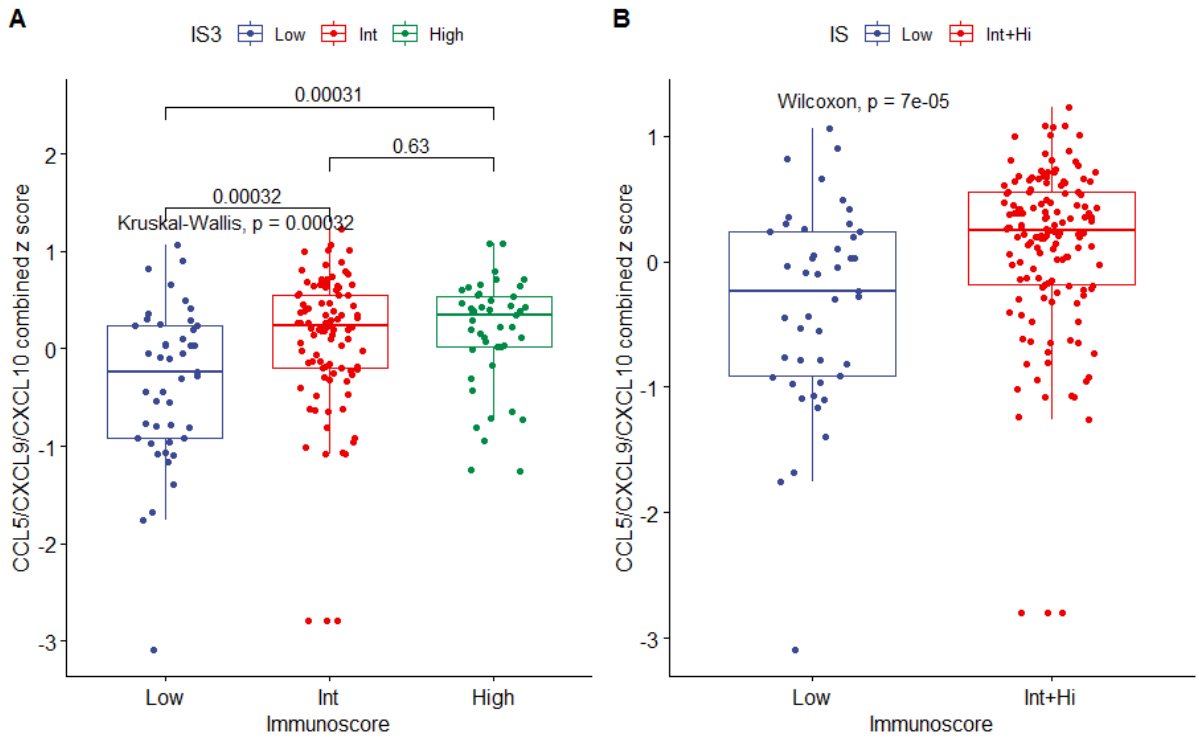



**Figure 5.33.** Boxplots illustrating the associations between Th-1 associated chemokines (the combined z score of CCL5, CXCL9 and CXCL10 expression) and the Immunoscore (Low, Int, High). (A) Shows the differences in gene expression scores when the Immunoscore categories are grouped into Low, Intermediate and High. (B) Shows the differences in gene expression scores when the Immunoscore categories are grouped into Low versus Intermediate and High. Kruskal-Wallis test p = 7e-05. Int = Intermediate. IS3 = Immunoscore categories. IS = Immunoscore.

The follicular T-cell-associated chemokines showed lower gene expression levels in the low Immunoscore group but no differences in gene expression between the Intermediate and High Immunoscore groups, suggesting that the differences were apparent only when comparing the completely ‘cold’ tumours (IS Low) with ‘warm’ or ‘hot’ tumours (IS Int and IS Hi) (Figure 5.34).

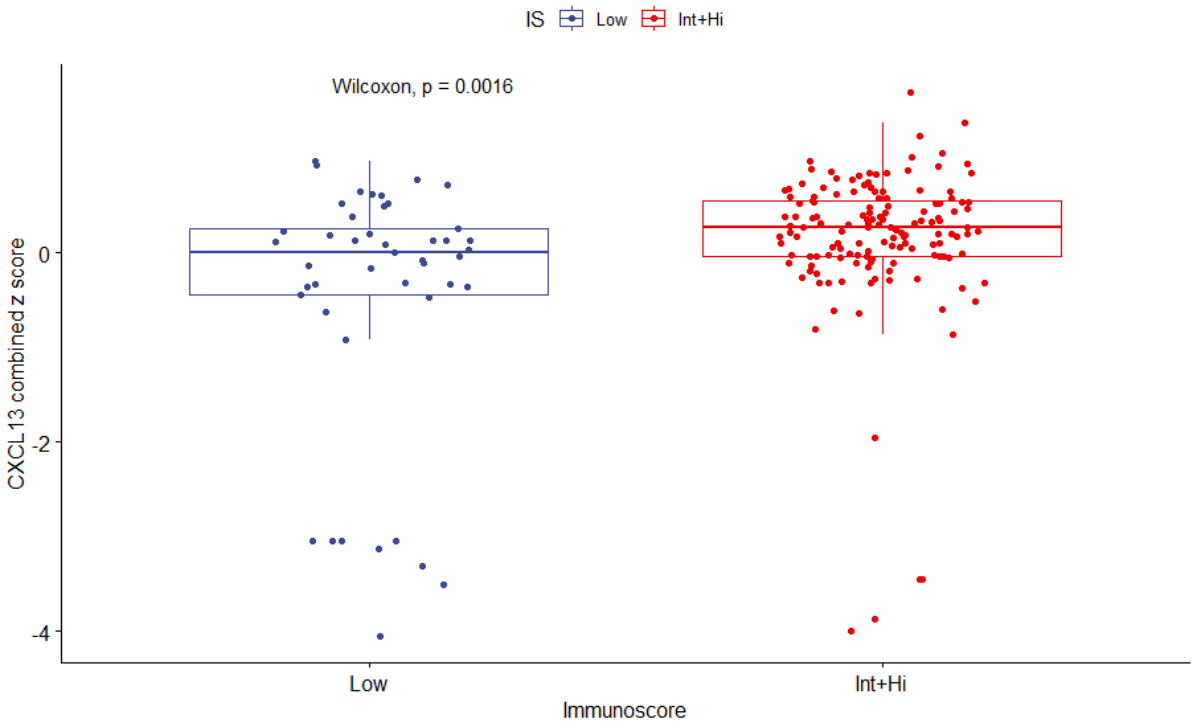



**Figure 5.34.** Boxplots illustrating the associations between follicular Th cell-associated chemokines (the CXCL13 z score) and the Immunoscore. Differences in expression are seen in the IS Low compared with Int and High categories for both, Wilcoxon test, p = 0.0016. Hi = High. Int = Intermediate. IS = Immunoscore categories. Treg = regulatory T cell.

In contrast, there were no apparent differences amongst Immunoscore categories in gene expression of the regulatory T cell or IL-17-associated chemokines (Figure 5.35).

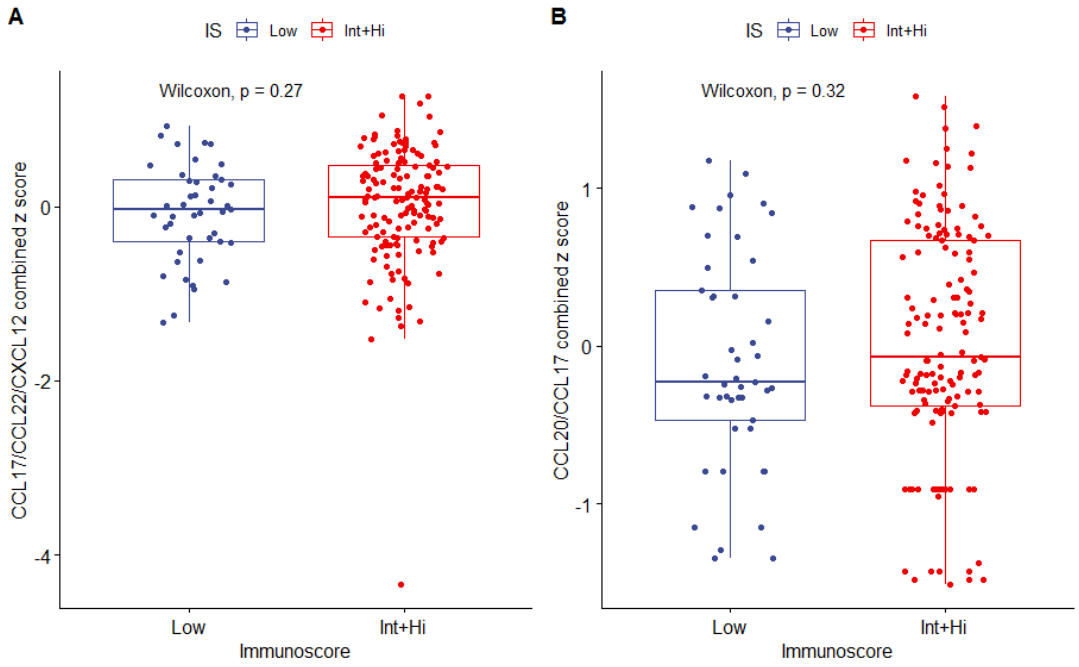



**Figure 5.35.** Boxplots illustrating the associations between (A) T-reg-associated chemokines (combined z score of CCL17, CCL22 and CXCL12 expression) and the Immunoscore. Wilcoxon test p = 0.27 and (B) IL17-associated chemokines (combined z score of CCL20 and CCL17 expression) and the Immunoscore, Wilcoxon test p = 0.32. Hi = High. Int = Intermediate. IS = Immunoscore categories. T-reg = regulatory T cells.

### 5.4.5. Wnt signalling pathway-associated markers

Data from Grasso *et al.* [256] showed that activated wnt/β-catenin signalling was correlated with the absence of T-cell infiltration in CRC. This is a key/canonical signalling pathway in CRC tumorigenesis. Inactivating mutations in *APC* drive wnt-signalling and APC loss was associated with reduced T cell infiltration in their study. They also found that nuclear β-catenin (CTNNB1) expression was inversely correlated with immune cell infiltration. Mouse melanoma studies have shown that activation of β-catenin intrinsic to tumour cells prevents tumour infiltration with T cells and causes tumour cell resistance to ICB [258]. Hypomethylation of AXIN2, a key wnt-signalling gene, leads to increased AXIN2 expression, which is inversely associated T cell infiltration in CRC [256].

However, in this data set, although APC2 expression was lower in IS Low tumours, there was no clear association between the Immunoscore and the expression of the other wnt signalling markers, including WNT7B, CTNNB1 and AXIN2 (Figure 5.36).

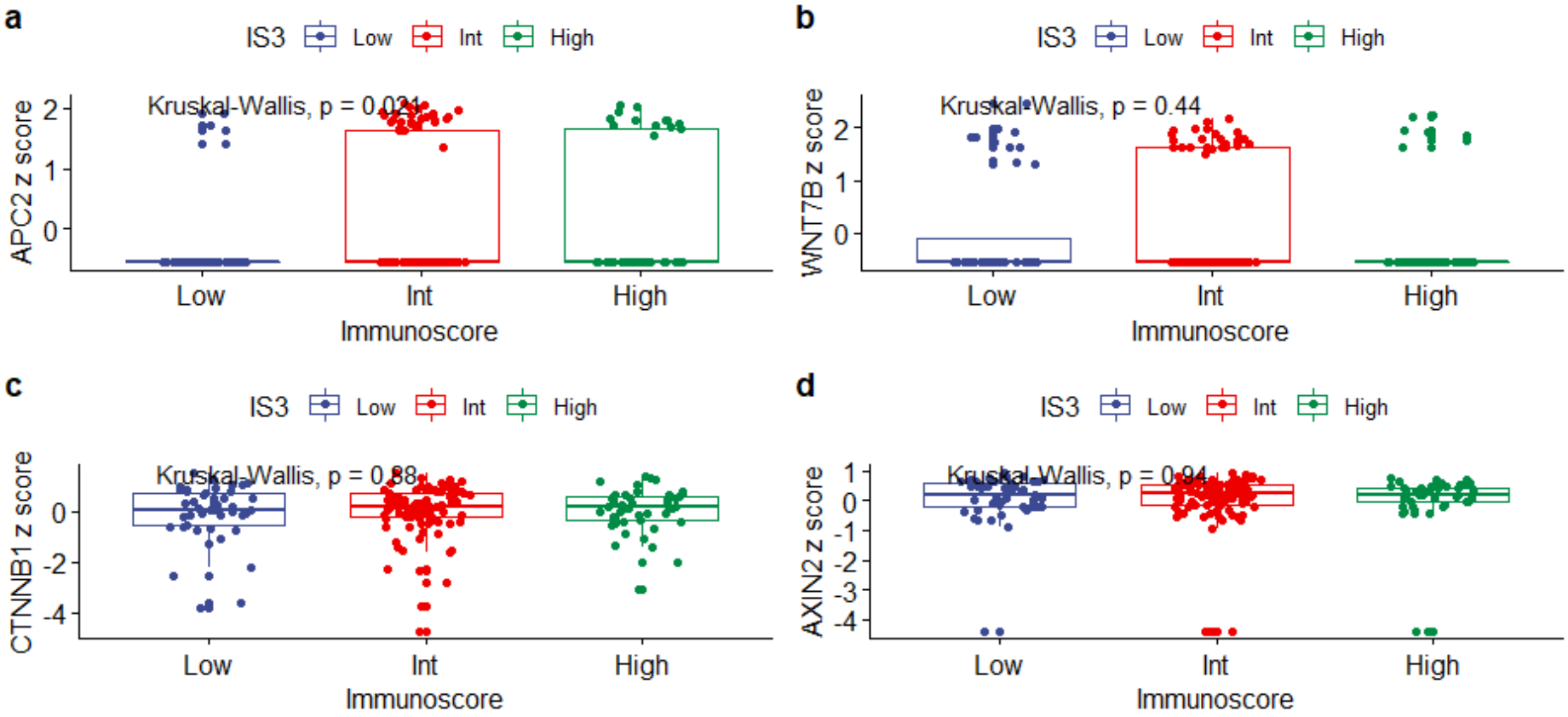



**Figure 5.36.** Box plots showing the association between wnt signalling markers and the Immunoscore (a) APC2 expression is lower in IS Low tumours, Kruskal-Wallis test, p = 0.021. (b) WNT7B expression is not correlated with the Immunoscore, p = 0.44. (c) CTNNB1 expression is not correlated with the Immunoscore, p = 0.88. (d) AXIN2 expression is not correlated with the Immunoscore, p = 0.94. IS = Immunoscore, IS3 = Immunoscore categories.

### 5.4.5. Lymphangiogenic markers

The associations between a panel of lymphangiogenic markers and the Immunoscore were also analysed. Tumour-associated lymphatic vessels may serve two opposing roles. They are potential routes for metastatic spread, but they may aid tracking of anti-tumour immune cells into the immune environment. Lymphatic endothelial cells may also release immunomodulatory cytokines and act as antigen presenting cells [259]. Fankhauser *et al.* showed that VEGFC signalling enhanced the response to immunotherapy in mouse melanoma models [260]. They also showed that, in human metastatic melanoma, VEGFC expression correlates strongly with T cell inflammation. VEGFC induces the upregulation of CCL21 on lymphatic endothelial cells, which tracks CCR7+ immune cells into immune environments.

The associations between VEGFC, CCL21 and CCR7 expression and the Immunoscore were assessed. The distribution of VEGFC z scores was bimodal. There was no difference in the distributions of MSS compared with MSI-high CRC (Wilcoxon test, $p = 0.537$, Figure 5.37).

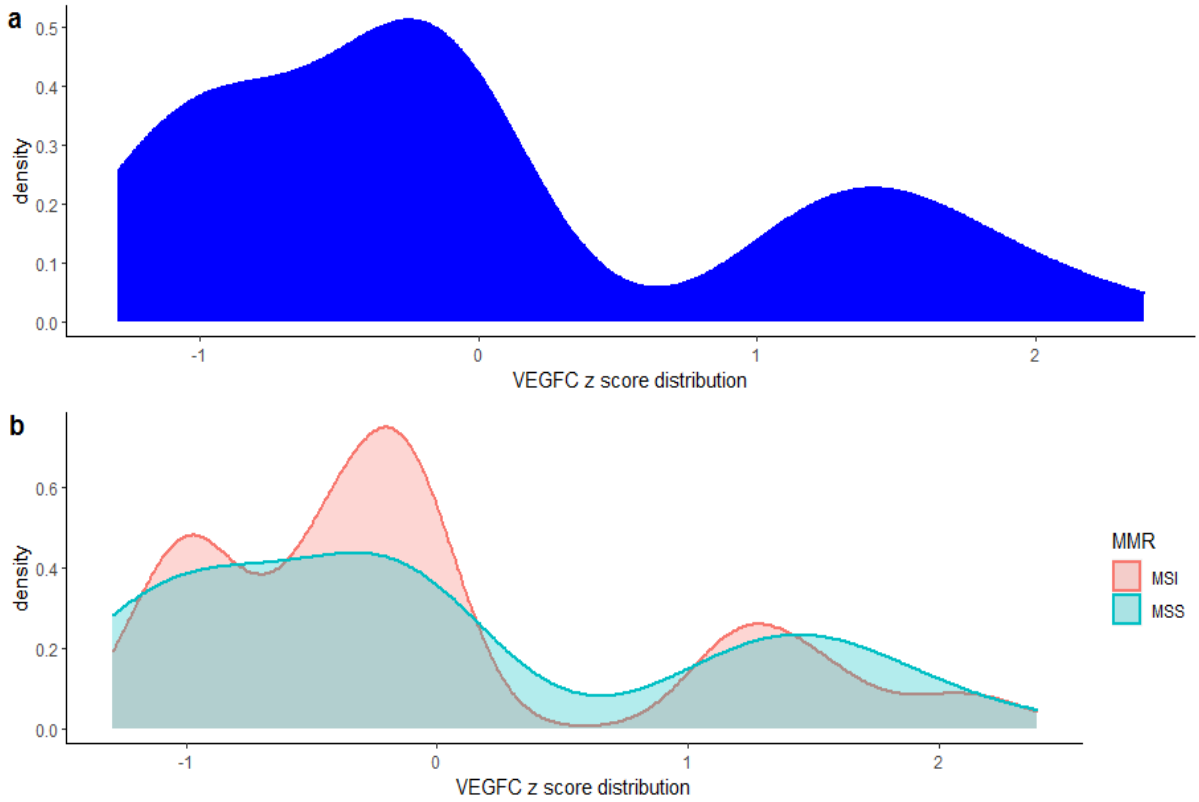



**Figure 5.37.** (a) Distribution of VEGFC expression across the sample data set, n = 190. This is bimodal. (b) This division is not due to differences in microsatellite status (MSS vs MSI-high CRC, Wilcoxon test, p = 0.531).

Both CCR7 and CCL21 expression were correlated with the Immunoscore, especially when IS Low samples were compared with IS Intermediate and High samples. However, VEGFC expression was not correlated with the Immunoscore in this data set (Figure 5.38).

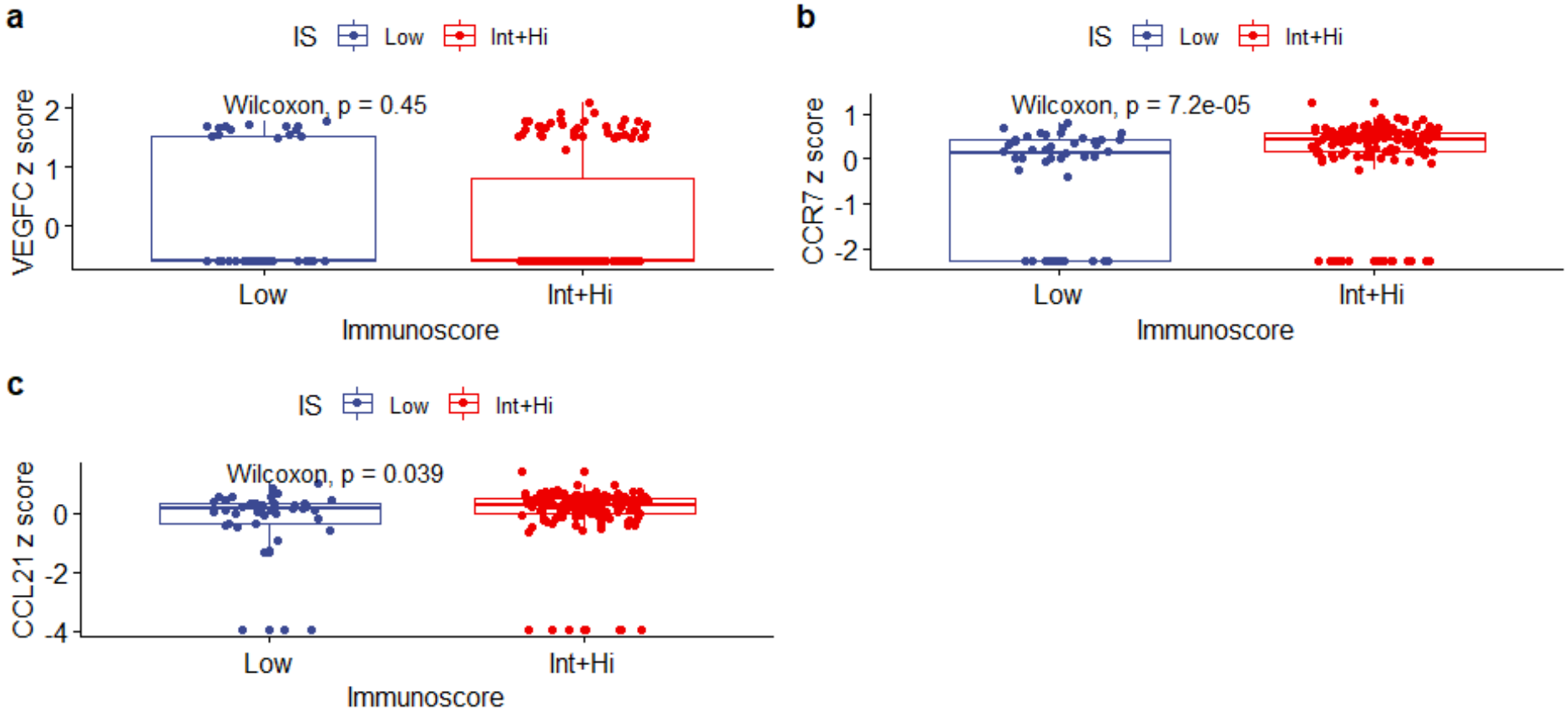



**Figure 5.38.** Boxplots comparing (a) VEGFC (b) CCR7 and (c) CCL21 expression z scores with the Immunoscore. Expression levels of CCR7 and CCL21 but not VEGFC are higher in the IS Int+Hi categories compared with the IS Low categories. Int+Hi = Intermediate and High categories. IS = Immunoscore.

VEGFC expression levels were low in most samples and may account for the lack of an association. No association of VEGFC expression levels with other confounding variables such as age (ANOVA, p= 0.89), disease stage (Kruskal-Wallis test, p = 0.74), or primary tumour location (Kruskal-Wallis test, p = 0.85) was present.

## 5.5 Immunohistochemical analysis

### 5.5.1. MHC Class II expression

Immunohistochemical tests were performed on fixed tumour slides. Staining for Class II expression was performed using the ANICCA Class II clinical trial protocol and technique [215], by Dr Phillipe Taniere at the UHB Department of Histopathology. Class II staining was completed in 186 samples. Class II expression was quantified as a percentage for each slide, using the Pathologist guide from the ANICCA trial.

Class II expression was generally very low in the samples, with only 26.3% showing any Class II expression (Table 5.6). Samples with expression less than 1% were considered Class II negative, while those with expression of 1% and greater were considered Class II positive.

**Table 5.6.** Distribution of Class II expression in the data set

| **Class II expression (%)** | **<1%** | **1-50%** | **>50%** |
|---|---|---|---|
| **Number (%)** | 137 (73.7) | 37 (19.9) | 12 (6.5) |

Class II = Major histocompatibility complex Class II antibody staining.

Class II IHC expression was correlated with the Immunoscore (Table 5.7). There was a greater proportion of samples with positive Class II expression in the Immunoscore “High” category compared with the “Intermediate” and “Low” categories (Pearson’s $\chi^2$ test, p = 0.041).

**Table 5.7.** Comparison of Class II expression by Immunoscore category

| | **IS Low** | **IS Int** | **IS high** |
|---|---|---|---|
| Class II expression <1% (n/%) | 37 (21.4) | 74 (54.0) | 26 (19.0) |
| Class II expression 1-100% (n/%) | 9 (19.6) | 23 (46.9) | 17 (34.7) |

Int = Intermediate. IS = Immunoscore. There was a greater proportion of Class II expressing samples in the Immunoscore High category, Pearson’s $\chi^2$ test, p = 0.041).

Class II expression was generally low (74% of samples had less than 1% expression). There was a positive association between the Immunoscore and Class II expression (Kruskal-Wallis test, p = 0.028, Figure 5.39).

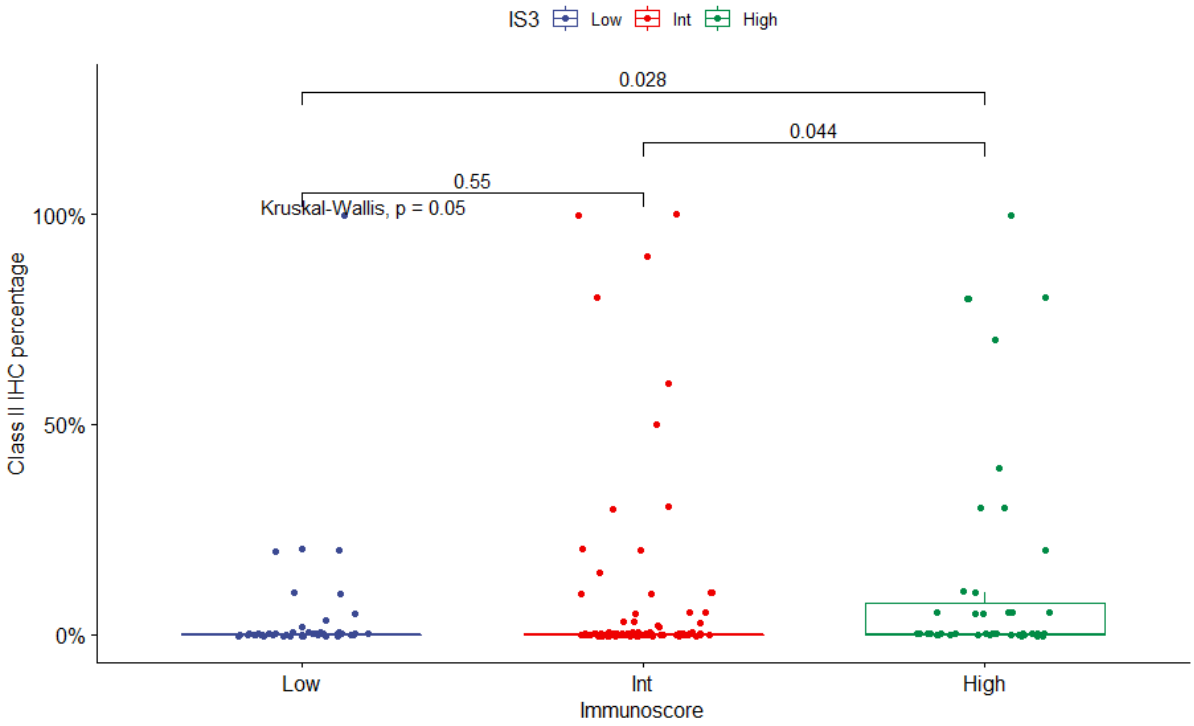



**Figure 5.39.** Boxplots illustrating the associations between MHC Class II percentage expression and the Immunoscore (Low, Int, High). Kruskal-Wallis difference in Immunoscore in Low vs High Immunoscore, p = 0.028. Int = Intermediate, IS3 = Immunoscore categories.

Class II IHC expression was also compared with Class II RNA expression to deduce the correlation between mRNA transcripts and protein expression in tissue. As most samples are Class II IHC negative, the association seen was small, but it was a positive and statistically significant correlation between Class II protein expression and the RNA z score (R = 0.15, p = 0.049, Figure 5.40).

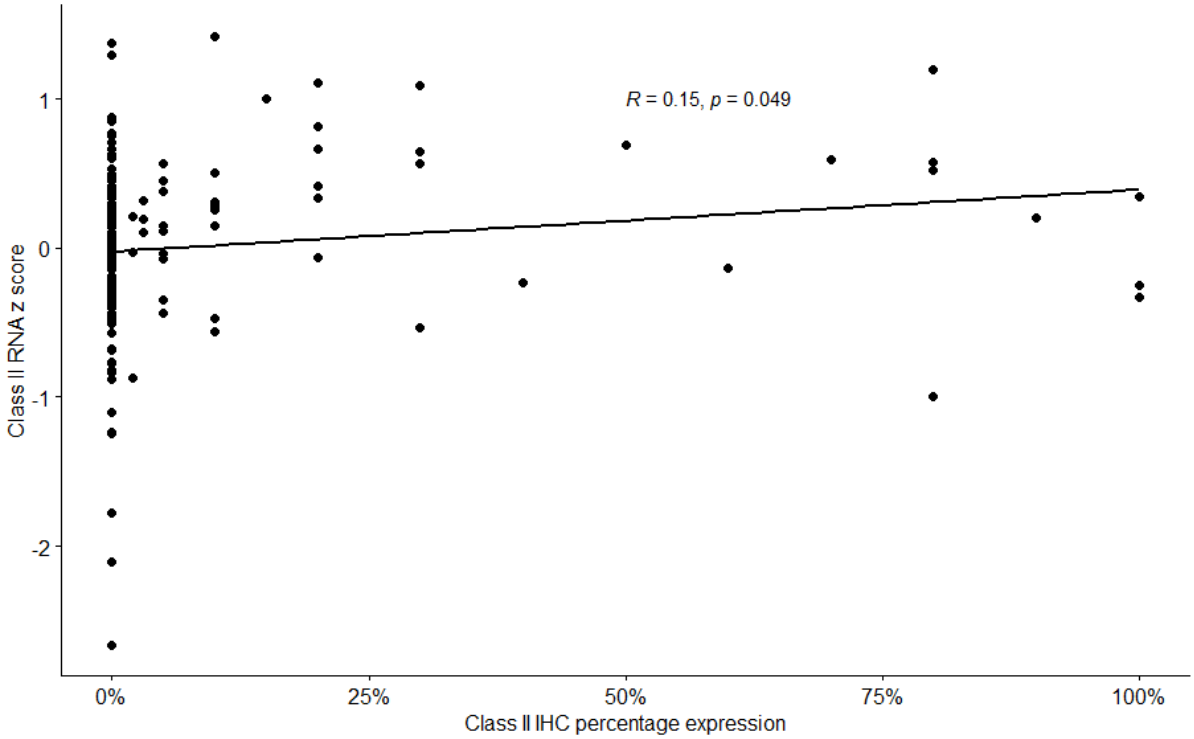



**Figure 5.40.** Comparison of MHC Class II protein expression (IHC percentage) and RNA expression in colorectal tumour samples. N = 180. IHC = immunohistochemistry.

Class II IHC expression was also compared with the CIRC z score. There was a significant positive association between the Class II IHC expression and the CIRC score (R = 0.22, p = 0.0027, Figure 5.41).

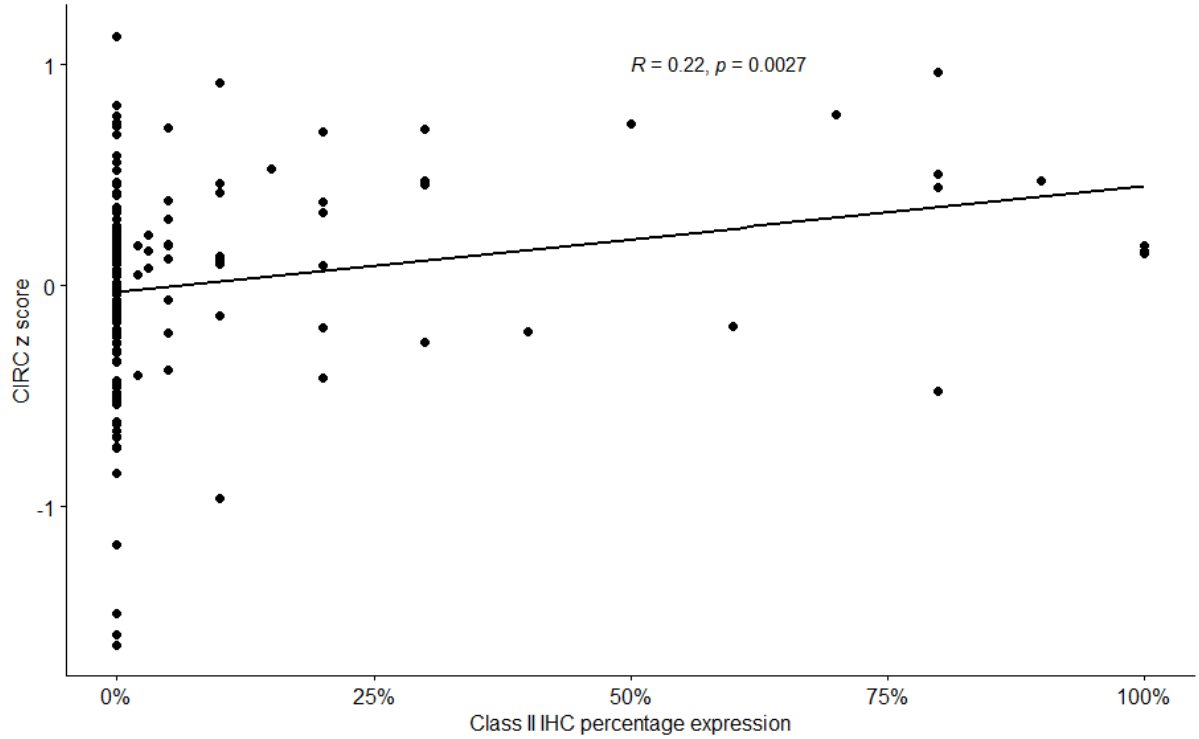



**Figure 5.41.** Comparison of MHC Class II protein expression (IHC percentage) and the CIRC score in colorectal tumour samples. N = 180, CIRC = co-ordinate immune response cluster. IHC = immunohistochemistry.

Associations between Class II expression and OS and RFS were also analysed. Samples were split into Class II negative (less than 1% expression) and Class II positive (greater than 1% expression). While there was no association between Class II expression and OS (Figure 5.42), there was a trend towards greater RFS in Class II positive patients (Figure 5.43).

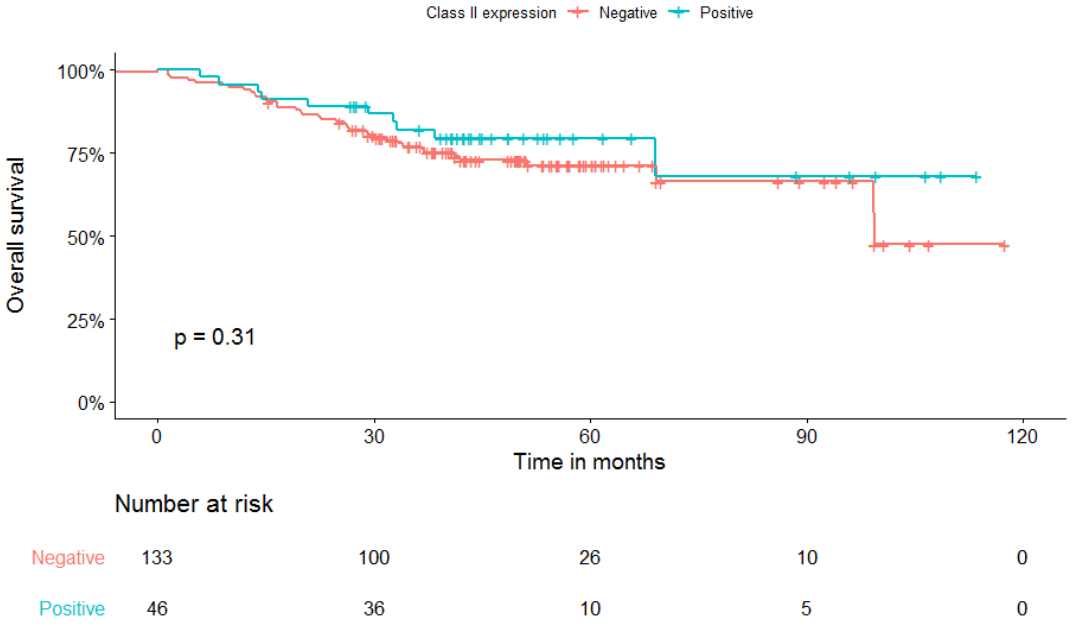



**Figure 5.42.** Kaplan-Meier estimate of overall survival (OS) stratified by MHC Class II expression in formalin-fixed colorectal tumour tissue. No difference is observed in Class II negative compared with Class II positive samples.

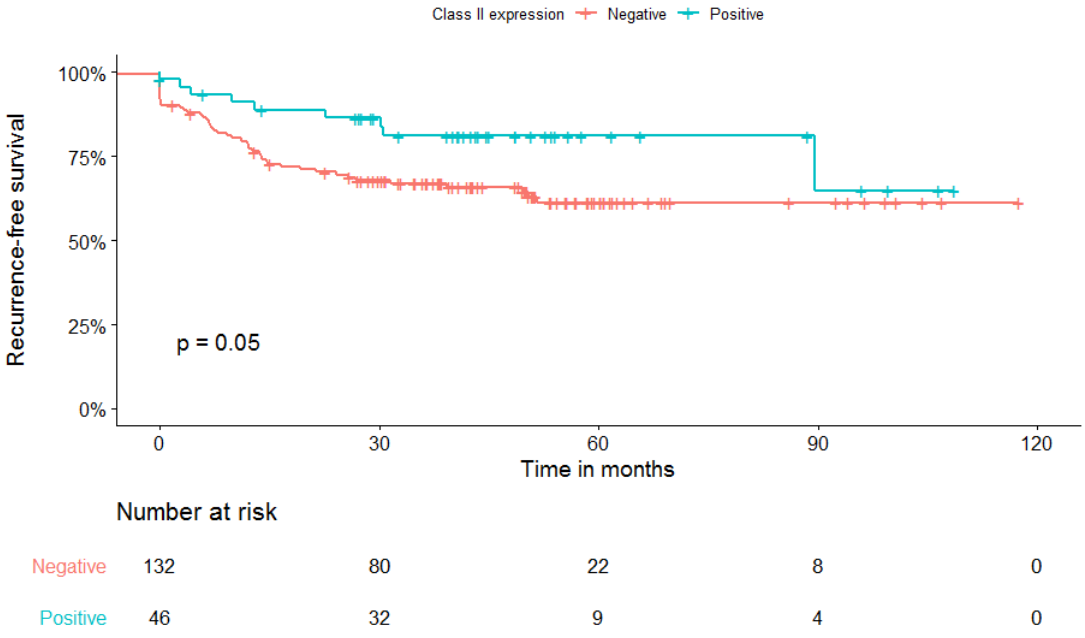



**Figure 5.43.** Kaplan-Meier estimate of recurrence-free survival (RFS) stratified by MHC Class II expression in formalin-fixed colorectal tumour tissue. There is a trend towards increased RFS is Class II positive patients compared with Class II negative patients. Cox proportional hazards model, p = Hazard ratio Class II positive versus Class II negative = 0.55 (95% CI = 0.27 – 1.1, p = 0.05).

## 5.6. Discussion

This chapter has analysed in significant detail the relative contributions of some somatic genomic determinants to the colorectal immune environment, using the Immunoscore as the key outcome. Comparisons with survival data have also been performed. In particular, the contributions of tumour mutational and neoantigen burden and intratumoral heterogeneity to the immune contexture and clinical outcomes were explored. Somatic immune gene expression analysis emphasised the key roles of MHC Class II expression and a Th-1 centric metric, the CIRC to the immune environment. These results are supported by immunohistochemical analysis of expression of a set of markers including Class II, PD-1/CD8 and lymphatic endothelial proteins.

### 5.6.1. Intratumoral heterogeneity is a greater determinant of the immune response in CRC than tumour mutational burden

TMB was found to have no association with the Immunoscore or patient survival. However, there was an association between neoantigen burden and the Immunoscore. TMB and neoantigen burden were highly correlated, so this discrepancy could be due to the other factors present within the TMB calculation, such as structural variants and Class II neoantigens. The impact of structural variants was partially analysed by the indel burden, which was shown to have no association with the Immunoscore. Class II neoantigens were not analysed in this data set due to difficulties with importing and utilising the available Class II

neoantigen prediction tools within the restricted and encrypted Research Environment.

On the other hand, intratumoral heterogeneity (as measured independently by the MATH score and modified DPClust approaches) is inversely associated with the Immunoscore. This correlation appears to be greater in the MSI-high samples, but due to a small sample size this could not be conclusively determined. However, a combination of neoantigen burden and ITH showed a stronger correlation with the Immunoscore and survival than either factor alone. This corresponds with lung cancer data by McGranahan *et al.* [164] and highlights the key role that clonal neoantigens also play in determining the immune response in CRC.

### 5.6.2. MHC Class II expression is strongly correlated with the immune response in CRC

Class II expression by mRNA and immunohistochemistry (protein) expression were associated with the Immunoscore. Class II expression is generally low in CRC (found in 26% of this data set), and may also identify a subset of patients with potentially potent anti-cancer immune responses. Antigen processing and presentation through both MHC Class I and Class II-mediated mechanisms is required for anti-cancer responses to clonal neoantigens. While Class II is constitutively expressed on professional APCs, these may be excluded from the tumour environment. Therefore, the induction of Class II expression on tumour cells becomes more significant. In CRC, the absence of Class II expression is

associated with reduced lymphocyte infiltration into tumours and a higher incidence of lymph node metastases [59]. Most established neoantigen determination pipelines only call Class I neoantigens [169], which is a significant draw back in assessing the role of Class II-restricted neoantigens in CRC. There are increasingly sophisticated Class II calling algorithms, including netMHCIIpan [261], MHCnuggets [175] and neonMHC2 [66], but was not possible to use these for the purposes of this thesis. When available, Class II neoantigen determination will further corroborate these results and consolidate the understanding of the mechanism through which clonal neoantigens induce the immune response.

### 5.6.3. Differential gut bacteria-derived chemokine expression correlates with the immune response in CRC

I found strong associations between Th1-centric gut bacteria-derived chemokines and the Immunoscore, which suggests a contribution of the gut microbiome to differential immune responses in cancer in addition to its established role in carcinogenesis [198]. Cremonesi *et al.* [205] demonstrated the induction of expression of these chemokines in tumour cells both *in vitro* and *in vivo* on exposure to gut bacterial cultures composed predominantly of *Fusobacterium nucleatum*, *Bacteroides fragilis* and *Escherichia coli*. 16s ribosomal RNA (rRNA) metagenomic sequencing showed associations between particular bacteria families and the expression of pro-inflammatory chemokines, particularly CCL5, CXCL9 and CXCL10, in tumour cells. In addition, variation in the gut microbiome composition appears to influence the response to immune checkpoint blockade in melanoma [202]. I was keen to explore the potential metagenomic implications

of these results, including the potential for confounding effects of germline and tumour-related factors on the microbiome in CRC. The CIRC, which is also a Th1-centric metagene, is also strongly associated with the Immunoscore, therefore it is possible that the role of gut microbiota may be secondary. Full metagenomic sequencing was performed on these samples, and the results are presented in Chapter 6.

### 5.6.4. There are no associations between wnt-driven markers or lymphangiogenic markers and the Immunoscore

A surprising finding was the lack of association between *wnt* signalling markers and the Immunoscore. This was contrary to other studies that suggest an inverse correlation between nuclear β-catenin expression by immunohistochemistry and T cell infiltration in colorectal tumours [256]. This may represent a discrepancy between β-catenin mRNA expression and immunohistochemical staining, or may be due to lower statistical power in this data set. Grasso *et al.* studied 1,150 samples of which the majority had high levels of β-catenin staining [256], while this data set has a smaller number (n = 190). Lymphangiogenic markers also did not show convincing associations with the Immunoscore. Lymphangiogenesis has potentially conflicting effects in cancer progression, by either facilitating metastatic progression through trafficking tumour cells to draining lymph nodes, or suppressing tumour progression by providing conduits through which anti-tumour immune cells can be more effectively trafficked to the tumour site [259]. It is possible that these effects negate each other and lead to no direct association with immune cell infiltration in CRC.

### 5.6.5. Limitations

This analysis supports the hypothesis presented, that neoantigen clonality has a stronger impact on the immune environment than mutational burden. Additional findings included the significant impacts of MHC Class II expression and the expression of Th-1 driven markers on the immune environment.

There were some limitations to this analysis. First, it was exploratory. Due to bioinformatics pipeline artefacts, there was sample loss during the data analysis process, thus potentially reducing statistical power to detect differences between patient groups. Secondly, although correlations were established between somatic factors and the Immunoscore, the potential causal mechanisms are unknown. There are good precedents from other solid tumours adding support to the conclusion that these associations are valid and causal. Finally, it was not possible to determine the predictive value of these somatic factors in determining the response to immune therapies, as no patient in this analysis received immunotherapy. It is of critical importance that subsequent clinical trials of immunotherapy in CRC also analyse neoantigen burden, clonality and metagenomic factors to assess their predictive value.

# Chapter 6: Metagenomic determinants and the Immunoscore

## 6.1. Introduction

The gut microbiome is hypothesised to play a crucial role in the development of colorectal cancer (CRC). In particular, *Fusobacterium nucleatum* is shown to be highly associated with the development of CRC [262], and has been shown to cause tumorigenesis in animal models [263]. Meta-analysis of faecal microbiota studies shows global microbial signatures associated with CRC, including genera including

*Fusobacterium*, *Porphyromonas*, *Parvimonas*, *Peptostreptococcus*, *Gemella*,

*Prevotella*, and *Solobacterium* amongst others [264].

Differences in gut microbiota have been shown to be associated with differential responses to anti-PD1 immunotherapy in some epithelial cancers, with *Akkermansia* and *Enterococcus* species predominating in responders [202]. In addition, stimulation of production of Th1-chemokines by gut microbiota including *Fusobacterium nucleatum, Bacteroides fragilis* and *Escherichia coli* appear to drive T cell tracking the colorectal cancer environment, in both in cancer cell models and mouse models [205].

Thus, it is important to explore the contribution of the gut microbiome to the immune contexture in our CRC cohort. Although most metagenomics analysis is performed using 16S rRNA amplicon sequencing of material from stool samples, whole genome sequencing (WGS) approaches have advantages including greater genomic diversity and less risk of biases associated with the PCR required for amplifying the marker genes in 16s rRNA sequencing [207, 265]. It is however, a slower process and analysis is more complex. This dataset had the

advantage of access to somatic WGS data directly from gut mucosa. Metagenomic sequencing was performed using the Kraken2 pipeline [206], which assigns taxonomic labels to DNA sequences. While stool metagenomic data was not available, mucosal somatic WGS provided a rich source of data for analysis and comparison with the Immunoscore and the other immunogenomic information available, to derive interesting conclusions about the role of the microbiome in differential immune responses in colorectal cancer.

### 6.2. Metagenomic data generation

Following extraction and sorting of the reads from somatic whole genome bam files within the Genomics England Research Environment, the kraken2 pipeline was used to generate taxonomic outputs for each sample. Tumour bam files were available within the Research Environment for 168 of 177 patients for whom the Immunoscore was available, and the pipeline was successfully completed in 164 patients. In the pilot data set, the same procedure was performed, but outside the Research Environment. The pipeline was successfully completed on 26 of 30 patients for whom the Immunoscore was available. Both data sets were combined to give a total number of 190 patients. Sequencing results were visualised in Pavian, which provides the data in both text csv format and in Sankey diagrams. Sankey diagrams display the flow of reads, with the width proportional to the number of reads [227]. The reason for non-availability of results in eight patients was due to inability of the pipeline to generate sorted bam files from the WGS inputs.

### 6.2.1. Classification and distribution of microbial reads

The number of microbial reads per sample varied from 29,043,075 to 250,107,348 (with a single outlying sample that only had 114,885 reads). The median read count was 97,533,904. There were differences in the numbers of reads and percentages of classified reads and proportion of bacterial reads per sample between the GeL dataset and the pilot dataset, most likely reflecting the two different populations sampled at different timepoints.

The median read count in the pilot data set was higher than in the GeL data set (99,842,676 compared with 88,267,042; Wilcoxon test p = 0.003). A median of 73.6% of reads per sample were classified in the GeL dataset. However, most of these reads were chordate reads, accounting for a median of 65.5% of reads per sample. These are not relevant for the metagenomics analysis and so were excluded, giving a median of 5.1% classified reads per sample. In contrast, in the pilot dataset, there was a median of 3.5% reads were classified per sample (Wilcoxon test, p = 4.04e-05).

The median percentage of bacterial reads per sample was 0.7% in the GeL dataset (ranging from 0.04% to 26.1%) and 3.0% in the pilot dataset (ranging from 2.2% to 20.5%) (Figure 6.1).

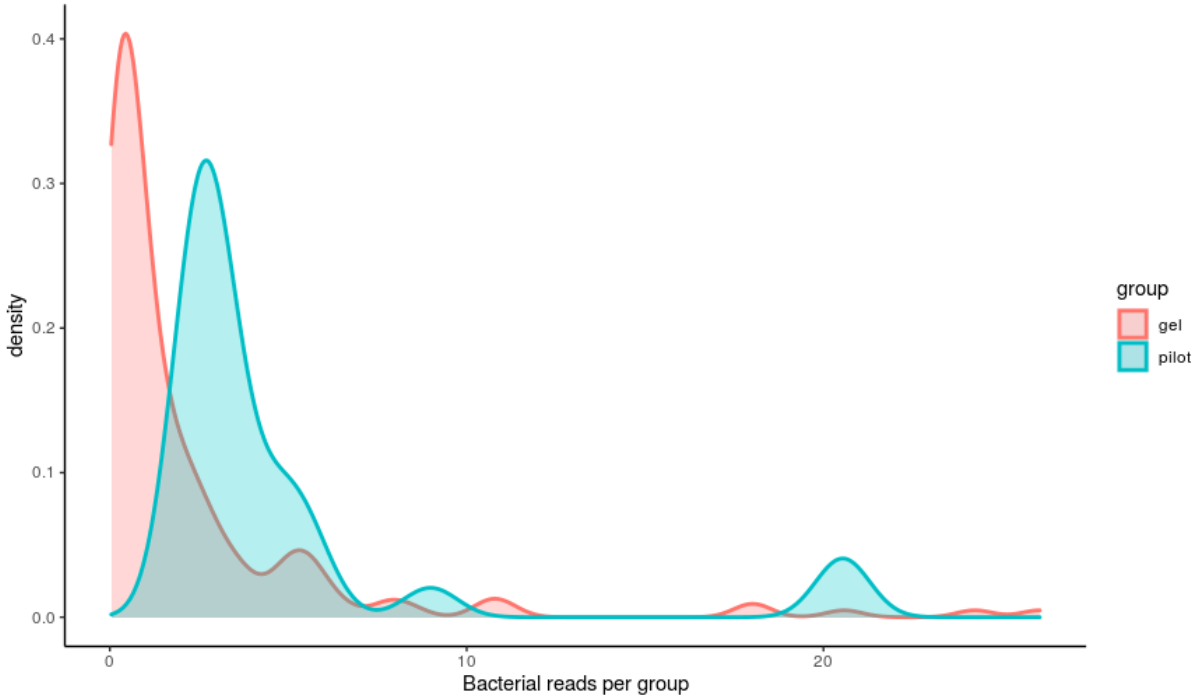



**Figure 6.1.** Median distribution of bacterial reads by percentage in the Genomics England dataset compared with the pilot dataset. The median percentage of reads is higher in the pilot data set (3.0% compared with 0.7%, Wilcoxon test p = 2.17e-07). geL = Genomics England sample set.

In order to maximise the sample size, the datasets were combined for further analysis. However, when analysing the associations between bacterial reads and the Immunoscore, sub-group testing was also performed to determine if the differences in bacterial reads between the pilot and GeL datasets could bias the conclusions drawn.

#### *6.2.1.1. Microsatellite status and read counts*

After exclusion of samples for which data on microsatellite status/DNA mismatch repair (MMR) status was not available, associations between microsatellite status and total read count and bacterial read percentages were studied. Data was

available for 151 patients. Median total read count per sample was higher in microsatellite unstable (MSI-high) when compared with microsatellite stable (MSS) CRC, however this difference was not significant (107,978,542 versus 97,533,904, Wilcoxon test, p = 0.12) (Figure 6.2).

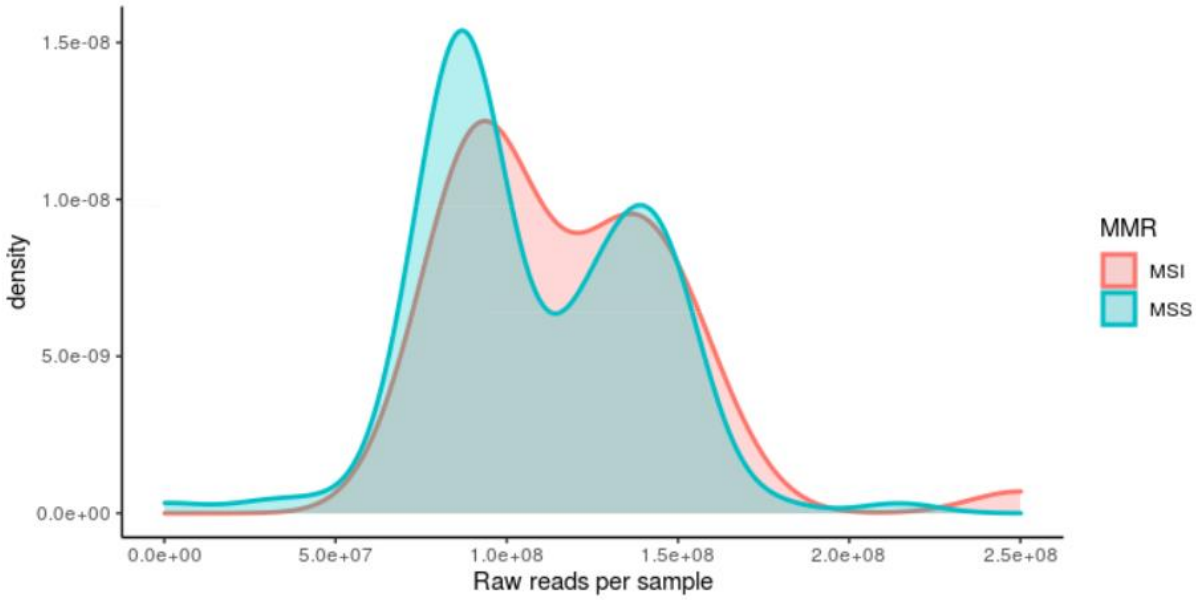



**Figure 6.2.** Comparison of number of raw reads per sample in microsatellite instability high (MSI) compared with microsatellite stable (MSS) colorectal samples. N = 151. The median number of reads is higher in MSI tumours but this is not statistically significant (Wilcoxon test, p = 0.12).

However, the percentage of bacterial reads was significantly higher in MSI-high samples compared with MSS CRC (2.44% versus 0.70%, Wilcoxon test, p = 0.0002, Figure 6.3).

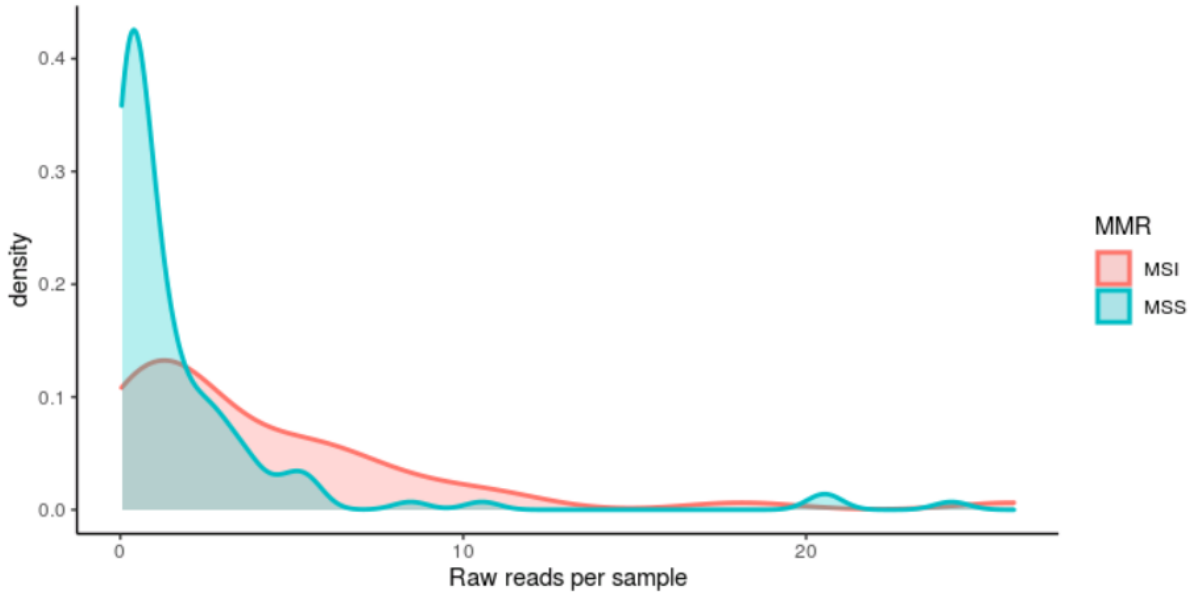



**Figure 6.3.** Comparison of number of bacterial reads per sample in microsatellite instability high (MSI) compared with microsatellite stable (MSS) colorectal samples. N = 151. The median number of reads is higher in MSI tumours (Wilcoxon test, p = 0.00016).

This raises the interesting possibility that microsatellite unstable tumours have a different microbiome from microsatellite stable tumours, and this may be one mechanism through which there are differential immune responses in the tumour environment.

### 6.2.2. Determination of bacterial operational taxonomic units

Metagenomics sequencing analysing small subunit 16S/18S rRNA marker gene sequence datasets utilise operational taxonomic units (OTUs), which are clusters of organisms grouped by DNA sequence similarity of specific taxonomic markers. This clustering is based on sequencing similarity thresholds and they have been shown to correspond roughly to microbial phylogeny [266].

To determine the taxonomic units for detailed analysis of these samples, the Sankey diagrams of the taxonomic outputs were examined. The Sankey diagrams for two individual samples are shown below for illustrative purposes. In Figure 6.4, this sample has a clear preponderance of *Bacteroides* reads in comparison to other genera such as *Pseudomonas* and *Prevotella*. In contrast, in Figure 6.5, Pseudomonas predominates while Bacteroides reads are relatively absent. Based on these outputs, bacterial genera were taken as the OTUs for further analysis.

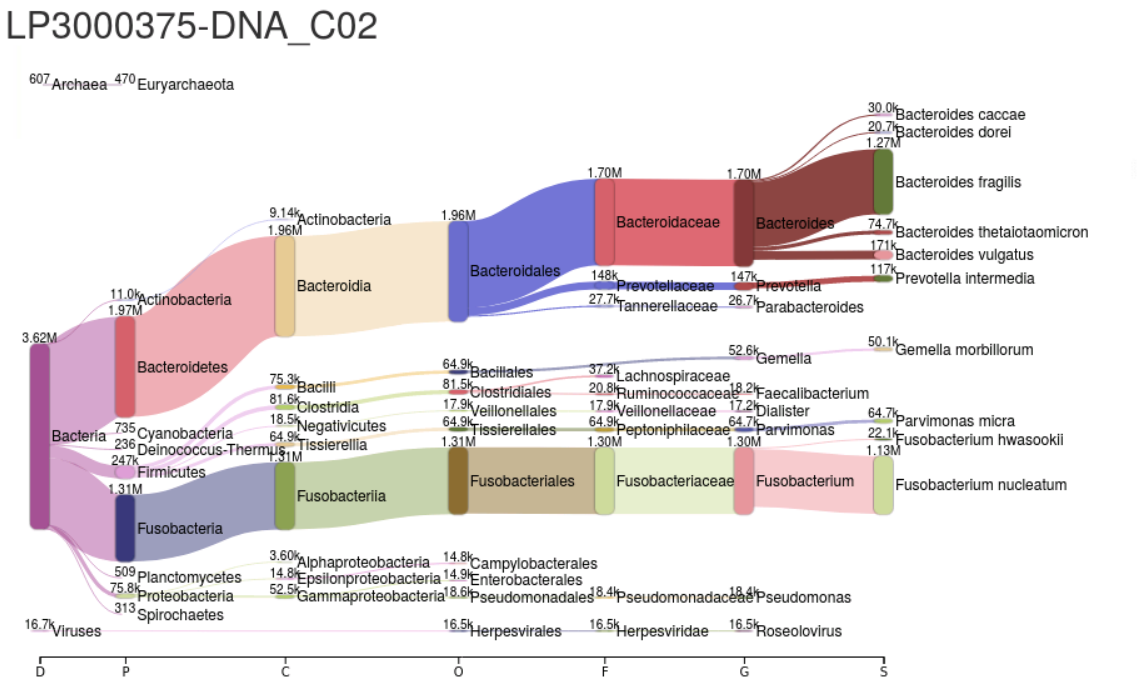



**Figure 6.4.** Sankey visualisation of metagenomic outputs from patient sample labelled LP3000375-DNA_C02. The flows show that there is clear preponderance of Bacteroides genera reads while other genera such as Pseudomonas are less represented. D = domain, P = phylum, C = class, O = order, F = family, G = genus, S = species.

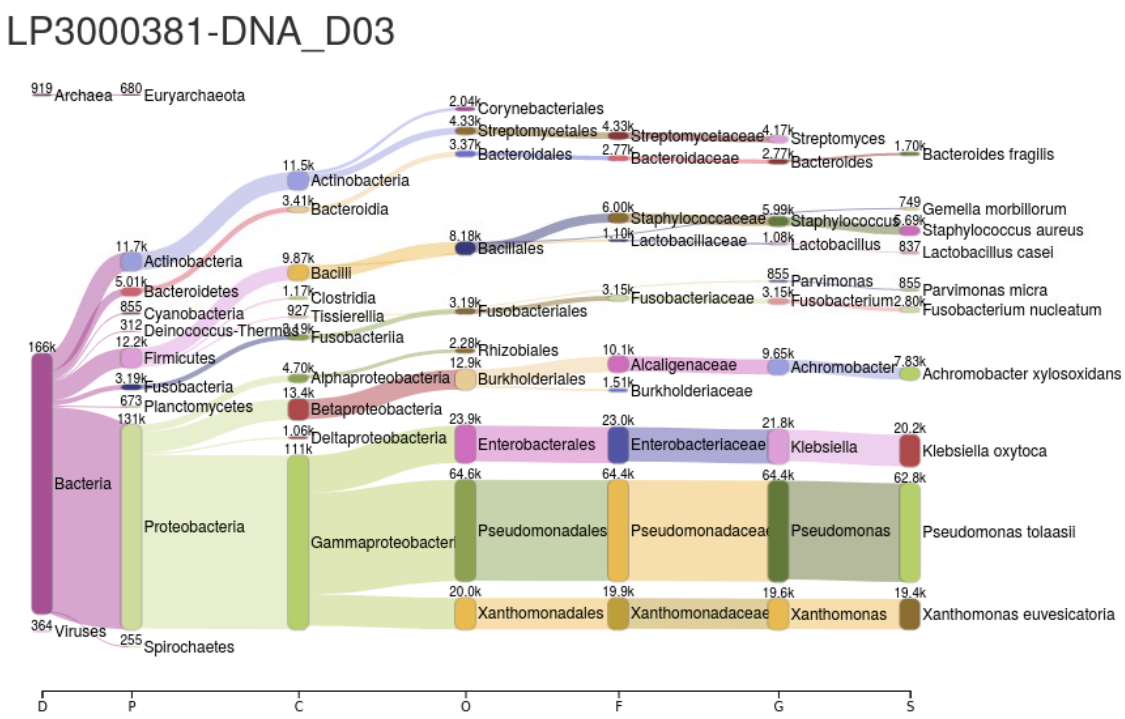



**Figure 6.5.** Sankey visualisation of metagenomic outputs from patient sample labelled LP3000381-DNA_D03. The flows show that there is clear preponderance of Pseudomonas genera reads while Bacteroides are clearly less represented. D = domain, P = phylum, C = class, O = order, F = family, G = genus, S = species.

### 6.3. Bacterial taxonomic unit association with the Immunoscore

For the combined dataset, Kraken2 generated 944 OTUs at genus level at the initial analysis. These were filtered to remove all non-bacterial genera (particularly viruses and fungi), leaving 606 bacterial genera.

Two separate analyses were performed using the Immunoscore result. In the first, the samples were divided into two groups. Those with IS 0 and 1 were coded as “Low” and those with IS 2, 3, and 3 were coded as “Int+Hi”. Differences in bacterial OTU reads between these two groups was determined. To generate correlation coefficients, these groups were given numerical values of 1 and 2. In the second analysis, samples with IS 0 and 1 (“Low”) were compared with

samples with IS 3 and 4 (“High”), and those with “Intermediate” scores were excluded. This reduced the sample size to 77.

### 6.3.1 Low compared with combined Intermediate and High Immunoscore

There were significant associations between the Immunoscore when divided into two groups (“Low” vs “Int+Hi”) and the number of reads per sample in 166 bacterial genera (OTUs). These included positive associations in groups such as *Halobacterium, Rhizobiales, Xanthomonas, Roseovarius, Corynebacterium and Nocardiopsis* amongst others (Table 6.1), but inverse associations between the Immunoscore and the number of reads per sample in genera such as *Klebsiella, Borrelia, Enterobacter, Neisseria, Pasteurella, Vibrio, Yersinia, Bacillus* and Legionella (Table 6.2).

**Table 6.1.** Positive associations between the number of bacterial reads per sample and the Immunoscore (Low vs Int+High) by bacterial operational taxonomic unit

| **Bacterial OTU** | **Correlation coefficient (r)** | **P value** |
|---|---|---|
| *Halobacterium* | 0.28 | 0.0002 |
| *Nitrosopumilus* | 0.26 | 0.0005 |
| *Streptosporangium* | 0.22 | 0.0035 |
| *Rathayibacter* | 0.22 | 0.0035 |
| *Rhizobiales* | 0.22 | 0.0043 |
| *Xanthomonas* | 0.22 | 0.0044 |
| *Salinimonas* | 0.21 | 0.0060 |
| *Kosmotoga* | 0.21 | 0.0067 |
| *Planctomycetaceae* | 0.21 | 0.0068 |
| *Actinotignum* | 0.21 | 0.0074 |
| *Gemmata* | 0.20 | 0.0080 |
| *Myxococcus* | 0.20 | 0.0087 |
| *Sulfitobacter* | 0.20 | 0.0089 |
| *Halorubrum* | 0.20 | 0.0091 |
| *DHEV2 Aciduliprofundum* | 0.20 | 0.010 |
| *Kitasatospora* | 0.20 | 0.010 |
| *Caulobacter* | 0.19 | 0.010 |
| *Dolichospermum* | 0.19 | 0.0132 |
| *Mycolicibacterium* | 0.19 | 0.0132 |
| *Tsukamurellacaeae* | 0.19 | 0.0147 |
| *Omnithinimicrobium* | 0.19 | 0.0152 |
| *Saccharothrix* | 0.19 | 0.0160 |
| *Ectothiorhodospira* | 0.18 | 0.0171 |
| *Massilia* | 0.18 | 0.0176 |
| *Leifsonia* | 0.18 | 0.0181 |
| *Actinomadura* | 0.18 | 0.0199 |
| *Pyrobaculum* | 0.18 | 0.0199 |
| *Methylibium* | 0.18 | 0.0203 |
| *Bosea* | 0.18 | 0.0206 |
| *Cellulomonas* | 0.18 | 0.0209 |
| *Nocardiopsis* | 0.18 | 0.0223 |
| *Desulfurococcaceae Aeropyrum* | 0.18 | 0.0229 |
| *Natronolimnobius* | 0.17 | 0.0238 |
| *Roseovarius* | 0.17 | 0.0240 |
| *Microbulbiferaceae* | 0.17 | 0.0241 |
| *Methanocellales* | 0.17 | 0.0242 |
| *Ketogulonicigenium* | 0.17 | 0.0246 |
| *Curtobacterium* | 0.17 | 0.0248 |
| *Celeribacter* | 0.17 | 0.0260 |
| *Thermococcus* | 0.17 | 0.0297 |
| *Brevundimonas* | 0.17 | 0.0323 |

| Bacterial OTU | Correlation coefficient (r) | P value |
|---|---|---|
| *Halorhabdus* | 0.16 | 0.0331 |
| *Dyella* | 0.16 | 0.0340 |
| *Haloplanus* | 0.16 | 0.0364 |
| *Actinoplanes* | 0.16 | 0.0364 |
| *Porphyrobacter* | 0.16 | 0.0374 |
| *Hydrogenophaga* | 0.16 | 0.0374 |
| *Natromonas* | 0.16 | 0.0392 |
| *Anaeromyxobactereceae* | 0.16 | 0.0414 |
| *Sphingomonas* | 0.16 | 0.0443 |
| *Serinicoccus* | 0.16 | 0.0443 |
| *Thermocrinis* | 0.16 | 0.0445 |
| *Desulfosarcina* | 0.15 | 0.0460 |
| *Methanoregulaceae* | 0.15 | 0.0463 |
| *Chloroflexineae* | 0.15 | 0.0468 |
| *Actinosynnema* | 0.15 | 0.0480 |
| *Brachybacterium* | 0.15 | 0.0481 |
| *Nocardioides* | 0.15 | 0.0491 |

OTU = operational taxonomic unit. Int+Hi = combined Intermediate and High

**Table 6.2.** Inverse associations between the number of bacterial reads per sample and the Immunoscore (Low vs Int+High) by bacterial operational taxonomic unit

| Bacterial OTU | Correlation coefficient (r) | P value |
|---|---|---|
| *Klebsiella* | -0.26 | 0.0005 |
| *Marinomonas* | -0.25 | 0.0009 |
| *Calothrix* | -0.25 | 0.0011 |
| *Cellulomonadaceae* | -0.24 | 0.0014 |
| *Gemmataceae* | -0.24 | 0.0014 |
| *Pectobacterium* | -0.24 | 0.0014 |
| *Borreliella* | -0.24 | 0.0015 |
| *Borrelia* | -0.23 | 0.0022 |
| *Xylella* | -0.23 | 0.0022 |
| *Anabaena* | -0.23 | 0.0022 |
| *Hydrogenothermaceae* | -0.23 | 0.0024 |
| *Enterobacter* | -0.23 | 0.0024 |
| *Dickeya* | -0.23 | 0.0024 |
| *Scytonemataceae* | -0.23 | 0.0024 |
| *Brachyspirales* | -0.23 | 0.0027 |
| *Neisseria* | -0.23 | 0.0027 |
| *Pasteurella* | -0.23 | 0.0029 |
| *Fervidobacteriaceae* | -0.23 | 0.0029 |
| *Fervidobacterium* | | |
| *Halorubraceae* | -0.23 | 0.0032 |
| *Liberibacter* | -0.22 | 0.0036 |
| *Bukholderia* | -0.22 | 0.0038 |
| *Prosthecochloris* | -0.22 | 0.0043 |
| *Aliivibrio* | -0.22 | 0.0043 |
| *Plantactinospora* | -0.22 | 0.0043 |
| *Rahnella* | -0.22 | 0.0044 |
| *Anaplasma* | -0.22 | 0.0044 |
| *Methanosarcina* | -0.22 | 0.0046 |
| *Vibrio* | -0.22 | 0.0046 |
| *Planococcus* | -0.22 | 0.0046 |
| *Oenococcus* | -0.22 | 0.0046 |
| *Halanaerobiaceae* | -0.22 | 0.0048 |
| *Sulfolobus* | -0.22 | 0.0048 |
| *Ehrlichia* | -0.22 | 0.0049 |
| *Thermodesulfobiaceae* | -0.22 | 0.0050 |
| *Yersinia* | -0.22 | 0.0050 |
| *Marinitoga* | -0.22 | 0.0050 |
| *Acidithiobacillia* | -0.21 | 0.0051 |
| *Chlamydiales* | -0.21 | 0.0052 |
| *Colwelliaceae* | -0.21 | 0.0056 |
| *Mycoplasma* | -0.21 | 0.0058 |
| *Glaesserella* | -0.21 | 0.0058 |
| *Helicobacter* | -0.21 | 0.0058 |
| *Acholeplasma* | -0.21 | 0.0061 |
| *Flavobacterium* | -0.21 | 0.0063 |
| *Luteimonas* | -0.21 | 0.0063 |
| *Bacillus* | -0.21 | 0.0064 |
| *Ureaplasma* | -0.21 | 0.0064 |
| *Microcoleaceae* | -0.21 | 0.0065 |

| Bacterial OTU | Correlation coefficient (r) | P value |
|---|---|---|
| *Leptospira* | -0.21 | 0.0069 |
| *Mycobacterium* | -0.21 | 0.0071 |
| *Micromonospora* | -0.20 | 0.0080 |
| *Listeriaceae* | -0.20 | 0.0080 |
| *Mycobacteroides* | -0.20 | 0.0081 |
| *Thermoanaerobacter* | -0.20 | 0.0083 |
| *Mesoplasma* | -0.20 | 0.0083 |
| *Frondihabitans* | -0.20 | 0.0084 |
| *Winogradskyella* | -0.20 | 0.0089 |
| *Spiroplasmataceae* | -0.20 | 0.0089 |
| *Cyanothecaceae* | -0.20 | 0.0090 |
| *Anoxybacillus* | -0.20 | 0.0090 |
| *Aquimarina* | -0.20 | 0.0090 |
| *Pseudoarcobacter* | -0.20 | 0.0097 |
| *Chroococcaceae* | -0.20 | 0.0097 |
| *Blattabacteriaceae* | -0.20 | 0.0099 |
| *Halomonas* | -0.20 | 0.0106 |
| *Methanococcus* | -0.20 | 0.0112 |
| *Bordetella* | -0.20 | 0.0112 |
| *Runella* | -0.20 | 0.0113 |
| *Paenibacillus* | -0.19 | 0.0116 |
| *Caldicellulosiruptor* | -0.19 | 0.0118 |
| *Shewanellaceae* | -0.19 | 0.0121 |
| *Fervidobacteriaceae* | -0.19 | 0.0121 |
| *Thermosipho* | | |
| *Geobacillus* | -0.19 | 0.0123 |
| *Moraxella* | -0.19 | 0.0144 |
| *Dokdonia* | -0.19 | 0.0146 |
| *Tolekusaellitidae* | -0.18 | 0.0167 |
| *Seratia* | -0.18 | 0.0180 |
| *Haloferax* | -0.18 | 0.0180 |
| *Sphingorhabdus* | -0.18 | 0.0183 |
| *Photorhabdus* | -0.18 | 0.0187 |
| *Rickettsia* | -0.18 | 0.0190 |
| *Francisella* | -0.18 | 0.0191 |
| *Entomoplasma* | -0.18 | 0.0191 |
| *Xenorhabdus* | -0.18 | 0.0200 |
| *Psychromonadaceae* | -0.18 | 0.0204 |
| *Herminiimonas* | -0.18 | 0.0204 |
| *Legionella* | -0.18 | 0.0209 |
| *Hydrogenobaculum* | -0.18 | 0.0214 |
| *Pandoraea* | -0.18 | 0.0221 |
| *Methanobacterium* | -0.17 | 0.0240 |
| *Hyphomonadaceae* | -0.17 | 0.0249 |
| *Aminobacter* | -0.17 | 0.0259 |
| *Streptomyces* | -0.17 | 0.0260 |
| *Frankiales* | -0.17 | 0.0270 |
| *Thermoanaerobacterium* | -0.17 | 0.0270 |
| *Vagococcus* | -0.17 | 0.0270 |
| *Nitrosopumilales* | -0.17 | 0.0292 |
| *Malacobacter* | -0.16 | 0.0328 |
| *Pleurocapsales* | -0.16 | 0.0402 |

| Bacterial OTU | Correlation coefficient (r) | P value |
|---|---|---|
| *Pedobacter* | -0.16 | 0.0406 |
| *Aliiarcobacter* | -0.16 | 0.0417 |
| *Phascolarctobacterium* | -0.15 | 0.0458 |
| *Methyloceanibacter* | -0.15 | 0.0471 |

OTU = operational taxonomic unit. Int+Hi = combined Intermediate and High

When correction for multiple testing was performed using both the Bonferroni and False Discovery Rate (FDR) approaches, no genus had an association with a p value less than 0.05. However, associations with significance (p) values less than 0.1 were present for 57 genera using the FDR method. These included Halobacterium and Salmonella amongst other genera (Table 6.3).

**Table 6.3.** FDR-corrected associations between the number of bacterial reads per sample and the Immunoscore (Low vs Int+High) by bacterial operational taxonomic unit

| **Bacterial OTU** | **p value** |
|---|---|
| *Acidaminococcales* | 0.080 |
| *Sphingorhabdus* | 0.080 |
| *Hyphomonadaceae* | 0.080 |
| *Calothrix* | 0.080 |
| *Cellulomonadaceae* | 0.080 |
| *Prevotellaceae* | 0.080 |
| *Verrucomicrobiaceae* | 0.080 |
| *Gemmataceae* | 0.080 |
| *Thermodesulfobiaceae* | 0.080 |
| *Colwelliceae* | 0.080 |
| *Runella* | 0.080 |
| *Frankiales* | 0.080 |
| *Fervidobacteriaceae* | 0.080 |
| *Fervidobacterium Hydrogenothermaeceae* | 0.080 |
| *Merismopediaceae* | 0.080 |
| *Scytonemataceae* | 0.080 |
| *Corynebacteriaceae* | 0.080 |
| *Frondihabitans* | 0.080 |
| *Salmonella* | 0.080 |
| *Winogradskyella* | 0.080 |
| *Plantactinospora* | 0.080 |
| *Halanaerobiaceae* | 0.080 |
| *Shewanellaceae* | 0.080 |
| *Sinorhizobium* | 0.080 |
| *Halorubraceae* | 0.080 |
| *Aquimarina* | 0.080 |
| *Herminiimonas* | 0.080 |
| *Oscillospiraceae* | 0.080 |
| *Pseudoarcobacter* | 0.080 |
| *Marinitoga* | 0.080 |
| *Luteimonas* | 0.080 |
| *Vibrio* | 0.080 |
| *Leptospira* | 0.080 |
| *Halobacterium* | 0.080 |
| *Spiroplasmataceae* | 0.080 |
| *Listeriaceae* | 0.080 |
| *Cyanothecaceae* | 0.080 |
| *Hydrogenobaculum* | 0.080 |
| *Yersinia* | 0.080 |
| *Prosthecochloris* | 0.080 |
| *Microcoleaceae* | 0.080 |

| Bacterial OTU | p value |
|---|---|
| *Xylella* | 0.080 |
| *Fervidobacteriaceae* | 0.080 |
| *Thermosipho* | |
| *Rhodoluna* | 0.080 |
| *Malacobacter* | 0.080 |
| *Streptomyces* | 0.080 |
| *Nitrosopumilales* | 0.083 |
| *Alliarcobacter* | 0.083 |
| *Aurantimonadaceae* | 0.083 |
| *Psychromonadaceae* | 0.088 |
| *Serratia* | 0.090 |
| *Thermoproteales* | 0.096 |
| *Melaminivora* | 0.096 |
| *Salegentibacter* | 0.096 |
| *Sphaerochaeta* | 0.096 |
| *Entomoplasma* | 0.096 |

FDR = false discovery rate. OTU = operational taxonomic unit

The top twelve positive associations and top twelve negative associations with the Immunoscore were visualised in a correlation matrix (Figure 6.6). This showed that the bacterial OTUs positively correlated with the Immunoscore and those inversely correlated with the Immunoscore were also inversely correlated with each other.

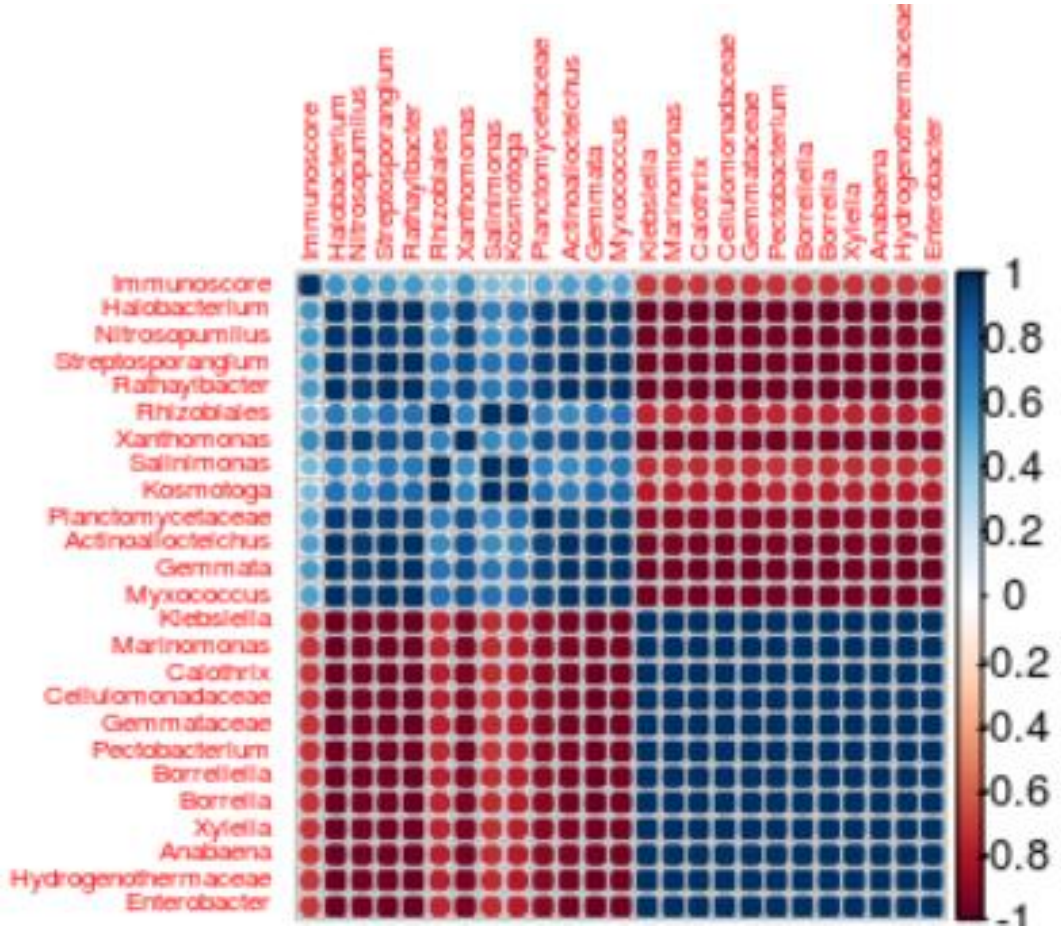



**Figure 6.6.** Correlation matrix showing correlations between bacterial operational taxonomic units and the Immunoscore (top and left). Blue = positive associations, red = inverse associations. Correlation gradient illustrated.

### 6.3.2. Low compared with High Immunoscore

The presence of the Intermediate Immunoscore group may also have reduced the ability to detect a difference between groups. To interrogate this, a sub-analysis of patients with either Low or High Immunoscore was performed. Although this reduced the sample size (n = 77), the effect size was hypothesised to be larger between the groups. The majority (n = 65) were within the 100KG dataset. Of these, there were 109 bacterial OTUs with significant associations with the Immunoscore. The top ten results are shown in Table 6.4.

**Table 6.4.** Associations between the Immunoscore and number of reads for each bacterial genus (Immunoscore Low versus High)

| **Bacterial genus** | **Association with Immunoscore (corr)** | **p value** |
|---|---|---|
| *Thaumarchaeota Candidatus* | +0.36 | 0.0009 |
| *Halobacterium* | +0.37 | 0.0024 |
| *Dolichospermum* | +0.32 | 0.0035 |
| *Microvirga* | +0.21 | 0.0040 |
| *Flaviflexus* | +0.09 | 0.0041 |
| *Pelosinus* | +0.11 | 0.0043 |
| *Borreliella* | -0.32 | 0.0045 |
| *Nitrosopumilus* | +0.28 | 0.0047 |
| *Cyanobium* | -0.09 | 0.0050 |
| *Streptosporangium* | +0.30 | 0.0064 |

100KG = 100 000 genomes. corr = correlation coefficient. + = positive correlation. - = inverse correlation.

These were modest associations, and after correction for multiple testing, none of these had a significant association, and there were no associations with a p

value <0.1. This suggests that the effect size remains modest even between Low and High Immunoscore samples, and simple size reduction reduced the power to detect a difference. Thus, further analysis was performed using the entire data set to maximise the sample size.

### 6.4. Bacterial OTU association with patient survival

Survival data was compared with the expression of bacterial reads per sample. There was no association seen between the median number of bacterial reads per sample and either OS or RFS (Figure 6.7).

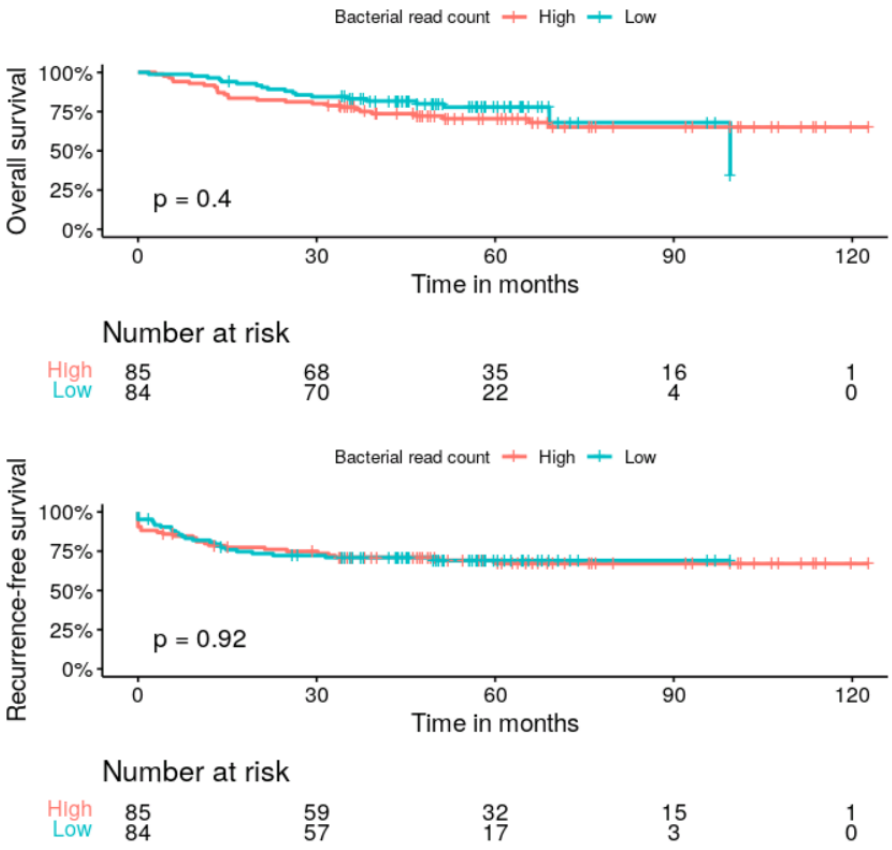



**Figure 6.7.** Kaplan-Meier estimate of overall survival (OS) and recurrence-free survival (RFS) stratified by bacterial read count. High = greater than median, Low = less than median. There is no difference in OS or RFS between the groups.

When analysis of specific bacterial taxonomic units was performed, one bacterial genus *Halobacterium*, showed significant associations with OS. Patients with a High median read count had greater OS (Figure 6.8).

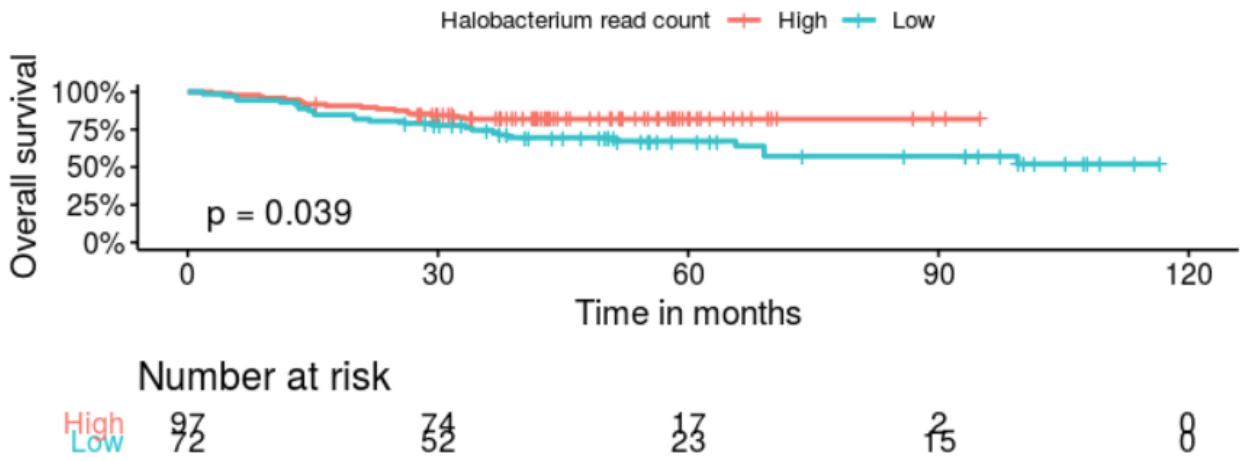



**Figure 6.8.** Kaplan-Meier estimate of overall survival (OS) Halobacterium read count. High = greater than median, Low = less than median. There is no difference in OS (Cox proportional hazards ratio, p = 0.45) or RFS (Cox proportional hazards ratio, p = 0.48) between the groups.

This leads to the conclusion that differences in bacterial OTU read counts are generally not significantly associated with survival differences.

This weak association with survival data despite the association with the Immunoscore may be due to small effect size of differences in bacterial read counts. They may also reflect the complexities in stratifying samples by read count, and in determining bacterial taxonomic units.

### 6.5. Bacterial orders and the Immunoscore

Grouping bacteria by order rather than genera also supported these findings. Using the Sankey classification to stratify results into read counts by order generated 79 orders. Analysis of their association with the Immunoscore, grouped into “Low” vs “Int+Hi” revealed that 13 of these had significant associations with the Immunoscore (Table 6.5). These were all inverse associations, with median counts higher in the “Low” group. However, after correcting for multiple testing using both the Bonferroni and FDR-approaches, none of these was significant, confirming that the effect size is likely to be small.

**Table 6.5.** Bacterial orders with significant associations with the Immunoscore

| **Bacterial order** | **p value** |
|---|---|
| *Enterobacterales* | 0.005 |
| *Desulfurococcales* | 0.006 |
| *Nostocales* | 0.009 |
| *Fusobacteria* | 0.010 |
| *Aquificae* | 0.010 |
| *Planctomycetia* | 0.010 |
| *Chloroflexia* | 0.013 |
| *Oscillatoriales* | 0.017 |
| *Pelagibacteriales* | 0.019 |
| *Myxococcales* | 0.022 |
| *Oceanospirillales* | 0.022 |
| *Lactobacillales* | 0.030 |
| *Corynebacteriales* | 0.049 |

### 6.6. Bacterial OTU associations with microsatellite status

It was previously noted that the median percentage of bacterial read counts was higher in MSI-high than in MSS tumours. Further analysis was performed to determine if particular bacterial genera were more predominantly expressed in MSI-high tumours. Of the 151 patients for whom MMR status was available, 39 were MSI-high (25.8%). Statistically significant differences in median read counts were seen between MSI-high and MSS tumours in 8 bacterial genera (Table 6.6).

**Table 6.6.** Bacterial genera with significant differences in median read counts per sample by mismatch repair status

| | **MSS CRC median read count** | **MSI-high CRC median read count** | **p value** |
|---|---|---|---|
| *Thermodesulfobiaceae* | 3.0 | 4.5 | 9.1e-05 |
| *Sinorhizobium* | 43.0 | 58.0 | 0.010 |
| *Vibrio* | 234.0 | 258.5 | 0.024 |
| *Halobacterium* | 0.0 | 0.0 | 0.026 |
| *Spiroplasmataceae* | 46.0 | 58.0 | 0.019 |
| *Yersinia* | 51.0 | 62.0 | 0.032 |
| *Serratia* | 99.0 | 134.5 | 0.028 |
| *Salengentibacter* | 5.0 | 8.5 | 0.010 |

MSS = microsatellite unstable. MSI-high = microsatellite instability high.

Of interest, six of these show positive associations with the Immunoscore (*Thermodesulfobiaceae, Sinorhizobium*, *Vibrio, Spiroplasmataceae, Yersinia* and *Serratia*), while *Halobacterium* read count is associated with overall survival. This interaction between microsatellite status and the microbiome is highly interesting and warrants further evaluation.

## 6.3. Discussion

### 6.6.1. Differential microbiome profiles are associated with differences in the colorectal immune environment

These results suggest that differential gut bacterial profiles are linked to differences in the colorectal tumour environment. Bacterial genera such as *Halobacterium* are enriched in High Immunoscore samples, whilst others including *Klebsiella*, *Enterobacter* and *Neisseria*, are enriched in Low Immunoscore samples. There is the possibility that these could distinguish immunotherapy responders and non-responders. It may be possible to boost the response to immunotherapy by either targeted anti-microbial therapy to deplete the metagenomic signatures associated with a more unfavourable tumour environment, or to boost the positively associated microbes through techniques including faecal mucosal transplantation (FMT) and selective microbial therapy [202].

While the bacterial genera found to be enriched in the Intermediate/High Immunoscore cohorts in these samples were not the same as those found to be associated with CRC risk (*Fusobacterium*) or non-response to immunotherapy in patients with lung cancer and renal cancer (including *Enterococcus, Staphylococcus and Corynebacterium*)[202], this may be due to specificity of microbiome signatures for different cancer types.

### 6.6.2. Microsatellite status appears to have an association with gut microbiome

MSI-high tumours had higher total read counts and bacterial percentages than MSS tumours. As MSI-high tumours generally have more lymphocyte infiltration into the tumour environment, this suggests that increased bacterial amounts (and possibly, by inference, diversity), is associated with a more inflamed tumour environment. It is not possible to determine the causal relationship between the microbiome and microsatellite status in this study. However, it is possible that microbiota assert their effects through epigenetic mechanisms. For example, it is known that the most sporadic MSI-high CRC tumours have evidence of epigenetic *MLH-1* silencing [77]. Data in Chapter 5 confirms that certain gut bacteria-derived chemokine signatures (particularly CCL5, CXCL9 and CXCL10 which are induced in a mouse model on exposure to cultures rich in *F. nucleatum*, *B. fragilis* and *E. coli*) [205] are positively associated with the Immunoscore and therefore a more inflamed tumour environment. A mechanism through which exposure to gut microbiota can induce epigenetic changes leading to microsatellite instability can similarly be explored, both in an *in vitro* colorectal tumour model or in induced tumours in a mouse model.

### 6.6.3 Implications and future study

This study utilised the availability of somatic WGS data to analyse the effects of differential microbiome signatures in the colorectal tumour immune microenvironment. There are associations between specific bacterial genera and the Immunoscore, which leads to the hypothesis that these bacteria may also

induce colorectal tumour immune activation or suppression through a different mechanism from the previously explored somatic (neoantigen-driven) or germline (differences in immune gene expression) mechanisms. Of particular note was the association between the microbiome density and microsatellite instability in the sample set.

A limitation of this data was the non-availability of genomic material from faecal samples in addition to somatic WGS data. There is a difference in faecal and mucosal microbiota composition [200], which may have implications for the conclusions reached. Comparison of the findings of this analysis with those in the literatures is also hampered by the use of varied techniques, with the majority of literature having used 16s rRNA amplicon sequencing approaches.

Thus far, metagenomic data has not been used in clinical practice to generate predictive biomarkers for treatment. This is largely due to the novel but rapidly increasing understanding of the role of the microbiome in tumorigenesis. Technical challenges with analysis of data generated and determination of the microbiome phylogeny of interest also leave the possibility of bias. This can be lessened with somatic WGS [207]. However, it requires significantly computational resources and bioinformatics skills for data analysis and interpretation.

Notwithstanding these challenges, the possibility of FMT as a method of increasing the response to certain immune therapies is an exciting one and worthy of future endeavour [202].

# Chapter 7: Discussion

This thesis has determined the contributions of germline and somatic immunogenomic factors to the immune contexture in colorectal cancer (CRC). The finding that there are immunogenomic differences in the immune response amongst patients has significant implications in determining the factors that predict a good response to immunotherapy. In particular, these findings provide the rationale for increasing the number of patients who can be recruited into clinical trials of immunotherapy in CRC, using stratifying markers beyond microsatellite status, which is currently the main marker in clinical use [267].

Emphasis has been placed on examining the determinants of the immune response in microsatellite stable CRC, which is currently not eligible for immunotherapy with immune checkpoint blockade agents, based on results from clinical trials of immunotherapy in metastatic disease. However, it is clear from my findings that MSS CRC is not uniformly poorly immunogenic as once thought. Using the Immunoscore as my primary marker of immune infiltration in colorectal tumours, there are significant variations in immunogenicity in both MSS and MSI-high CRC. I conclude that the principal determinants of immunogenicity are not limited to microsatellite status and tumour mutational burden as previously thought. I have shown that specific germline immune gene expression quantitative trait loci are associated with differences in the immune environment in both MSS and MSI-high CRC. I have also established a stronger relationship between a combination of neoantigen burden and neoantigen clonality and the immune response in CRC, than either marker alone. Despite the finding of lower neoantigen burden in MSS CRC, these tumours are not universally

immunologically “cold”, and therefore some of these patients could respond effectively to immune checkpoint blockade therapy.

A new important finding is the role of metagenomic factors, particularly the correlation of expression levels of gut microbiota with differential immune responses in CRC. These findings are persuasive in the context of the significant role the microbiome is known to play in the evolution and manifestation of many bowel disorders, notably inflammatory bowel disease and bowel cancer.

This therefore provides further support for expanding clinical trials of immunotherapy to patients with both MSS and MSI-high CRC, based on a panel of germline, somatic and epigenetic markers, such as in the NICHE trial [24]. In addition, they drive the potential for the use of immune therapies in earlier stage disease, as adjuncts to the established modalities (surgery, chemotherapy and radiotherapy).

### 7.1. Germline determinants of the colorectal cancer immune response

Most studies exploring SNP associations in cancer are genome wide association studies, targeted panel sequencing, or whole exome sequencing studies, such as in The Cancer Genome Atlas [8]. These have therefore not been able to explore the roles of intronic genomic regions of significance in determining clinical outcomes. In particular, the role of expression quantitative trait loci (eQTL) SNPs, which are found in non-coding regions of the genome, and which determine gene expression differences, had not previously been explored in colorectal cancer.

In this thesis, with access to high quality whole genome sequencing data, I have shown that germline eQTL SNPs contribute to differences in the immune response as determined by the Immunoscore. The eQTL SNPs discovered all have strong biological bases for their potential effects. Due to the large number of SNPs tested, some of these may represent statistical findings and may not reveal a causal link to the immune response in CRC. As this is a preliminary study, the effects of these eQTL SNPs can be further studied, first, by *in silico* analysis using larger datasets, with cross-validation. The biological effects of these can then be determined by performing *in vitro* assays to simulate the effects of each SNP on gene expression levels and the immune environment in a CRC model. These findings can finally be transferred to an early phase clinical trial of immunotherapy, targeting patients with favourable genotypes as determined by WGS.

### 7.1.1. Significant eQTL SNP correlations

The rs256208 eQTL SNP, which influences TCF7 expression, was most strongly associated with differences in the Immunoscore in this dataset. The variant alleles were associated with increasing Immunoscore. This SNP is common in the population, with an estimated frequency of 26.4% in the International HapMap Project [136] and highly prevalent in the dataset. Although the correlation of the germline variants with tumour mRNA transcript levels was not statistically significant, both the homozygous and heterozygous variants appeared to be associated with increased transcript counts, suggesting that the *TCF7* variants could be associated with a better immune response in CRC.

Inactivating mutations and fusions of *TCF7L1* and *TCFL2* are known to be common in CRC, with The Cancer Genome Atlas (TCGA) dataset reporting a frequency of *TCF7L2* mutations of 11.2% [8]. The TCF/LEF gene family is a critical part of the Wnt/β-catenin signalling pathway, mutations in which are canonical in colorectal tumorigenesis [268]. The overall function of TCF appears to be as a tumour suppressor [252]. It is reasonable to hypothesise that patients with the variant alleles could have better responses to immunotherapy.

Of the other eQTL SNPs which were associated with the Immunoscore, rs11161590, which determines *BCL10* expression was of interest, as tumour BCL10 expression has been found to be associated with a favourable prognosis in colorectal cancer in TCGA [222, 269]. The other significant SNPs are not yet known to have any specific outcomes in CRC, but increased CCR1 expression has an unfavourable association with outcomes in renal cancer [222]. In my dataset, the rs11919943 (*CCR1*) variant is shown to be inversely associated with the Immunoscore, and to have a not-statistically-significant trend towards increased CCR1 transcript counts, suggesting that CCR1 may have a similar role in CRC as in renal cancer. Finally, variants of the rs6673928 SNP, which drives IL19 expression, are associated with increased overall survival in cutaneous melanoma.

These are compelling findings that require further exploration, first by validation with a larger dataset for which the Immunoscore is available, and then by determination of a clear association between the immune gene eQTLs and mRNA transcript or protein expression of these genes, and finally by determination of the biological basis for the association. In an *in vitro* colorectal

model, this would potentially involve induction of these eQTLs using an expression vector, and establishing a co-culture model with autologous peripheral blood mononuclear cells or purified effector (CD8+ or CD4+) T cells and colorectal cancer organoids [270] to replicate the immune microenvironment.

## 7.2. Somatic determinants of the colorectal cancer immune response

### 7.2.1. A combination of neoantigen burden and neoantigen clonality correlates strongly with the colorectal immune response

Somatic whole genomic sequencing and 3' RNA sequencing of fixed tumour tissue yielded significant findings.

There was a positive association between neoantigen burden (single nucleotide variants) and the Immunoscore. Neoantigen clonality estimation involved *in silico* analysis, using both the mutant allele heterogeneity (MATH) score and a DPClust-based filtering algorithm, for the purposes of comparison. The MATH score is a whole mutation-based algorithm, which has been used to assess for the effects of intratumoral heterogeneity in head and neck tumours [188], while the DPClust algorithm which involves a modified Dirichlet clustering approach is much more computationally complex and relies on a number of assumptions, including that tumour mutational evolution is a linear process [56].

A combined categorisation of the samples, into 'good', 'intermediate' and 'poor', derived from uniting neoantigen burden and neoantigen clonality) yielded results. The combination was found to be better correlated both with the Immunoscore and recurrence-free survival than either neoantigen burden or neoantigen clonality separately. This corroborates the findings in lung adenocarcinoma [164], and confirms my hypothesis in CRC.

### 7.2.2. Several immune gene expression signatures are associated with the Immunoscore

My data showed striking associations between the coordinate immune response cluster (CIRC), a Th1-centric immune gene cluster, previously shown to distinguish CRC into four subsets, and the Immunoscore. A high CIRC expression is known to be correlated with microsatellite instability and strong immune expression, using TCGA whole exome data [92]. In my dataset, although the MSI-high CRC samples had higher CIRC signatures, the association of the CIRC score with the Immunoscore was independent of microsatellite status, which provides further evidence that microsatellite status is not the only driver of the immune response in CRC.

Other expression signatures that were found to be associated with the Immunoscore include MHC Class II gene expression (HLA-DP, -DQ and –DR), and cytotoxic T cell-associated gut bacteria-stimulated chemokine expression. While it has not been proved that these associations are causal, they are in keeping with previously published work. MHC Class II-dependent antigen presentation is known to be a driver of the anti-cancer immune response [66]. The chemokines CCL5, CXCL9 and CXCL10 are known to be associated with the trafficking of effector T cells into the CRC immune environment [205]. These particular chemokines are known to be upregulated by specific gut bacteria families including *Enterococcaceae, Lachnospiraceae, Methylobacteriaceae* and *Ruminococcaceae* [205]. These relationships were explored further by metagenomics analysis and discussed in section 7.3.

### 7.2.3. MHC Class II expression by immunohistochemistry is associated with the Immunoscore

The important role of MHC Class II expression in driving the immune response to CRC is supported by the finding of associations between MHC Class II expression by immunohistochemical analysis, and the Immunoscore. While Class II expression is generally low in CRC, where present, it was strikingly associated with Th1-associated immune markers, as well as the coordinate immune response cluster. Class II expression was also highly correlated with the Immunoscore. This further supports the accumulating evidence that Class II expression in CRC is a key determinant of patient outcomes.

Class II expression is induced by IFNγ in cancer cells, a process which is mediated by the transcriptional master regulator class II transactivator (CIITA). This leads to autophagy of cancer antigens and presentation to infiltrating T lymphocytes [271]. MHC Class II expression is rarely noted in metastatic CRC, and it is higher in well-differentiated than poorly differentiated tumours [59]. MHC Class II downregulation occurs as a mechanism of immune escape through a variety of mechanisms, including genomic alterations in *CIITA* and epigenetic silencing of MHC Class II induction pathways [272].

Class II expression in this cohort was 26%, which is slightly lower than has been observed in other datasets (quoted as ranging between 21% and 55%[62]). However, these results are reliable and robust, as the antibody staining and interpretation protocols were clinically validated for the ANICCA-Class II trial [215], with expert pathological interpretation of the slide results. The Class II expression results also showed good correlation with Class II mRNA expression.

The link between Class II expression and neoantigen clonality is less clear, primarily because this thesis did not explore Class II neoantigen burden and clonality. Class II neoantigen prediction is much more challenging than Class I prediction for several reasons. Class II endosomal peptide processing is complex and not well understood, predicting binding affinity of peptides is more complicated than in Class I, as the peptide-binding groove is open rather than closed, and Class II peptide-binding motifs have a longer amino acid length range (typically 11 to 20 amino acids) than Class I (usually 8 to 11 amino acids) [55]. Although *in silico* pipelines for determining Class II neoantigen burden are available, they are much more complex to manipulate than Class I pipelines [175]. This is a consideration for exploration in subsequent work.

### 7.3. Metagenomic associations with the colorectal cancer immune response

Analysis of the microbiome using somatic whole genome sequencing data also supported the findings from RNA expression of gut-bacteria-associated chemokines. In line with data from Cremonesi *et al.* [205] total bacterial read counts had no association with the Immunoscore or clinical outcomes, but specific bacterial genera did. In particular, there were clear differential expression patterns between MSI-high and MSS tumours. Specific taxonomic units were found to be more enriched in Immunoscore Intermediate/High (that is, more inflamed tumour environments) than in Immunoscore Low tumours.

The predictive value of these metagenomic signatures is yet to be fully elucidated, but can be performed readily as part of a clinical trial, distinguishing the signatures[202] of non-responders from responders.

Finally, the possibility that the microbiome could exert epigenetic selection pressures on the colorectal tumour environment leads to the prospect of modulation of the microenvironment by alteration of the microbiome to favour a more immunogenic one. This could be accomplished through simple, established means including faecal mucosal transplantation [273, 274] and selective antibiotic therapy.

### 7.4. Future directions

The findings of this thesis are compelling, and require corroboration using a larger dataset. A publicly available database such as the TCGA [192] could fulfil this role, although it would require the Immunoscore to have been performed on the samples. This is a significant financial challenge in the research setting due to the current test cost (a research cost of £200.00 per test at the time of the analysis), but it is anticipated there will be widening of this access to the Immunoscore, as there is a drive to incorporate it into clinical practice [124, 125].

In particular, the eQTL associations would be supported by work in a tissue model to interrogate the biological basis of the *in silico* associations found. Simulating the colorectal tumour microenvironment using a three-dimensional lymphocyte and colorectal organoid model has been successful in studying these interactions and generating tumour-specific T cells [270]. Induction of specific SNPs in a

colorectal model can be used to determine the downstream anti-tumour immune responses, and is an area of work which is currently being explored in our laboratory.

Finally, the aim of this thesis is to provide the rationale for early phase clinical trials of immunotherapy in CRC, expanding its current limited role. Although no patients in this study population received immunotherapy, there is convincing evidence to consider an early phase trial of neoadjuvant immunotherapy in patients with combined high neoantigen clonality and neoantigen burden, even in those with MSS CRC. Reverse translated studies, assessing immunogenomic and metagenomic features of responders and non-responders were crucial in determining the importance of microsatellite status in the early studies [267]. With improved access to WGS data, these will be able to further interrogate the interplay between metagenomic factors and the key germline and somatic factors found to be key determinants in this thesis.

## References


1. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018.
2. Fearon, E.R. and B. Vogelstein, *A genetic model for colorectal tumorigenesis.* Cell, 1990. **61**(5): p. 759-67.
3. Cairns, S.R., et al., *Guidelines for colorectal cancer screening and surveillance in moderate and high risk groups (update from 2002).* Gut, 2010. **59**(5): p. 666-89.
4. Durko, L. and E. Malecka-Panas, *Lifestyle Modifications and Colorectal Cancer.* Curr Colorectal Cancer Rep, 2014. **10**: p. 45-54.
5. Baena, R. and P. Salinas, *Diet and colorectal cancer.* Maturitas, 2015. **80**(3): p. 258-64.
6. Vogelstein, B., et al., *Genetic alterations during colorectal-tumor development.* N Engl J Med, 1988. **319**(9): p. 525-32.
7. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
8. Cancer Genome Atlas, N., *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.
9. Lichtenstein, P., et al., *Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland.* N Engl J Med, 2000. **343**(2): p. 78-85.
10. Dunlop, M.G., et al., *Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42,103 individuals.* Gut, 2013. **62**(6): p. 871-81.
11. Lung, M.S., et al., *Familial colorectal cancer.* Intern Med J, 2015. **45**(5): p. 482-91.
12. UK, N.C.R.A.S.a.C.R. *Chemotherapy, Radiotherapy and Tumour Resections in England: 2013-2014*. 2018 [cited 2018 23/10/2018]; Available from: http://www.ncin.org.uk/cancer_type_and_topic_specific_work/topic_specific_work/main_cancer_treatments.
13. Bertero, L., et al., *Eighth Edition of the UICC Classification of Malignant Tumours: an overview of the changes in the pathological TNM classification criteria-What has changed and why?* Virchows Arch, 2018. **472**(4): p. 519-531.
14. Brenner, H., M. Kloor, and C.P. Pox, *Colorectal cancer.* Lancet, 2014. **383**(9927): p. 1490-1502.
15. Whitehall, V.L., et al., *Morphological and molecular heterogeneity within nonmicrosatellite instability-high colorectal cancer.* Cancer Res, 2002. **62**(21): p. 6011-4.
16. CRUK. *About bowel cancer treatment* 2019 [cited 2019 16/12/2019]; Available from: https://www.cancerresearchuk.org/about-cancer/bowel-cancer/treatment.
17. Dehal, A., et al., *Neoadjuvant Chemotherapy Improves Survival in Patients with Clinical T4b Colon Cancer.* J Gastrointest Surg, 2018. **22**(2): p. 242-249.
18. Foxtrot Collaborative, G., *Feasibility of preoperative chemotherapy for locally advanced, operable colon cancer: the pilot phase of a randomised controlled trial.* Lancet Oncol, 2012. **13**(11): p. 1152-60.
19. Seymour, M.T., D. Morton, and o.b.o.t.I.F.T. Investigators, *FOxTROT: an international randomised controlled trial in 1052 patients (pts) evaluating neoadjuvant chemotherapy (NAC) for colon cancer.* Journal of Clinical Oncology, 2019. **37**(15_suppl): p. 3504-3504.
20. Zadlo, J., *Cost-effectiveness of new and emerging treatment options for the treatment of metastatic colorectal cancer.* Am J Manag Care, 2018. **24**(7 Suppl): p. S118-S124.
21. Petrelli, F., et al., *Stereotactic body radiotherapy for colorectal cancer liver metastases: A systematic review.* Radiother Oncol, 2018. **129**(3): p. 427-434.

22. Maughan, T.S., et al., *Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial.* Lancet, 2011. **377**(9783): p. 2103-14.
23. Overman, M.J., et al., *Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study.* Lancet Oncol, 2017. **18**(9): p. 1182-1191.
24. Chalabi, M., et al., *Neoadjuvant immunotherapy leads to pathological responses in MMR-proficient and MMR-deficient early-stage colon cancers.* Nat Med, 2020. **26**(4): p. 566-576.
25. Hoption Cann, S.A., J.P. van Netten, and C. van Netten, *Dr William Coley and tumour regression: a place in history or in the future.* Postgrad Med J, 2003. **79**(938): p. 672-80.
26. Sanlorenzo, M., et al., *Melanoma immunotherapy.* Cancer Biol Ther, 2014. **15**(6): p. 665-74.
27. Eggermont, A.M.M., M. Crittenden, and J. Wargo, *Combination Immunotherapy Development in Melanoma.* Am Soc Clin Oncol Educ Book, 2018(38): p. 197-207.
28. Xu, Y., et al., *The association of PD-L1 expression with the efficacy of anti-PD-1/PD-L1 immunotherapy and survival of non-small cell lung cancer patients: a meta-analysis of randomized controlled trials.* Transl Lung Cancer Res, 2019. **8**(4): p. 413-428.
29. Yarchoan, M., A. Hopkins, and E.M. Jaffee, *Tumor Mutational Burden and Response Rate to PD-1 Inhibition.* N Engl J Med, 2017. **377**(25): p. 2500-2501.
30. Overman, M.J., et al., *Durable Clinical Benefit With Nivolumab Plus Ipilimumab in DNA Mismatch Repair-Deficient/Microsatellite Instability-High Metastatic Colorectal Cancer.* J Clin Oncol, 2018. **36**(8): p. 773-779.
31. Schumacher, T.N. and R.D. Schreiber, *Neoantigens in cancer immunotherapy.* Science, 2015. **348**(6230): p. 69-74.
32. Lee, H.J., et al., *Expansion of tumor-infiltrating lymphocytes and their potential for application as adoptive cell transfer therapy in human breast cancer.* Oncotarget, 2017. **8**(69): p. 113345-113359.
33. Rosenberg, S.A. and N.P. Restifo, *Adoptive cell transfer as personalized immunotherapy for human cancer.* Science, 2015. **348**(6230): p. 62-8.
34. Sermer, D. and R. Brentjens, *CAR T-cell therapy: Full speed ahead.* Hematol Oncol, 2019. **37 Suppl 1**: p. 95-100.
35. Maude, S.L., et al., *Tisagenlecleucel in Children and Young Adults with B-Cell Lymphoblastic Leukemia.* N Engl J Med, 2018. **378**(5): p. 439-448.
36. Park, J.H., et al., *Long-Term Follow-up of CD19 CAR Therapy in Acute Lymphoblastic Leukemia.* N Engl J Med, 2018. **378**(5): p. 449-459.
37. Kochenderfer, J.N., et al., *Lymphoma Remissions Caused by Anti-CD19 Chimeric Antigen Receptor T Cells Are Associated With High Serum Interleukin-15 Levels.* J Clin Oncol, 2017. **35**(16): p. 1803-1813.
38. Jin, Z., et al., *The severe cytokine release syndrome in phase I trials of CD19-CAR-T cell therapy: a systematic review.* Ann Hematol, 2018. **97**(8): p. 1327-1335.
39. Lam, C., et al., *Systematic review protocol: an assessment of the post-approval challenges of autologous CAR-T therapy delivery.* BMJ Open, 2019. **9**(7): p. e026172.
40. Lu, Y.C. and P.F. Robbins, *Cancer immunotherapy targeting neoantigens.* Semin Immunol, 2016. **28**(1): p. 22-7.
41. Topalian, S.L., et al., *Safety, activity, and immune correlates of anti-PD-1 antibody in cancer.* N Engl J Med, 2012. **366**(26): p. 2443-54.
42. Brahmer, J.R., et al., *Safety and activity of anti-PD-L1 antibody in patients with advanced cancer.* N Engl J Med, 2012. **366**(26): p. 2455-65.

43. Le, D.T., et al., *PD-1 Blockade in Tumors with Mismatch-Repair Deficiency.* N Engl J Med, 2015. **372**(26): p. 2509-20.
44. Chen, E.X., et al., *Effect of Combined Immune Checkpoint Inhibition vs Best Supportive Care Alone in Patients With Advanced Colorectal Cancer: The Canadian Cancer Trials Group CO.26 Study.* JAMA Oncol, 2020. **6**(6): p. 831-838.
45. Sillo, T.O., et al., *Mechanisms of immunogenicity in colorectal cancer.* Br J Surg, 2019. **106**(10): p. 1283-1297.
46. Mettu, N.B., et al., *BACCI: A phase II randomized, double-blind, placebo-controlled study of capecitabine bevacizumab plus atezolizumab versus capecitabine bevacizumab plus placebo in patients with refractory metastatic colorectal cancer.* Journal of Clinical Oncology, 2018. **36**(4_suppl): p. TPS873-TPS873.
47. Roche, H.-L. *A Study to Investigate Efficacy and Safety of Cobimetinib Plus Atezolizumab and Atezolizumab Monotherapy Versus Regorafenib in Participants With Metastatic Colorectal Adenocarcinoma (COTEZO IMblaze370)*. 2016 June 12, 2018]; Available from: https://clinicaltrials.gov/ct2/show/NCT02788279.
48. Diaz, L.A., et al., *KEYNOTE-177: Randomized phase III study of pembrolizumab versus investigator-choice chemotherapy for mismatch repair-deficient or microsatellite instability-high metastatic colorectal carcinoma.* Journal of Clinical Oncology, 2017. **35**(4_suppl): p. TPS815-TPS815.
49. Kim, T.W. *Avelumab for MSI-H or POLE Mutated Metastatic Colorectal Cancer*. 2017 June 12, 2018]; Available from: https://clinicaltrials.gov/ct2/show/NCT03150706.
50. Sinicrope, F.A., et al., *Randomized trial of FOLFOX alone or combined with atezolizumab as adjuvant therapy for patients with stage III colon cancer and deficient DNA mismatch repair or microsatellite instability (ATOMIC, Alliance A021502).* Journal of Clinical Oncology, 2017. **35**(15_suppl): p. TPS3630-TPS3630.
51. Tabernero, J., et al., *Phase Ia and Ib studies of the novel carcinoembryonic antigen (CEA) T-cell bispecific (CEA CD3 TCB) antibody as a single agent and in combination with atezolizumab: Preliminary efficacy and safety in patients with metastatic colorectal cancer (mCRC).* Journal of Clinical Oncology, 2017. **35**(15_suppl): p. 3002-3002.
52. Eisenhauer, E.A., et al., *New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1).* Eur J Cancer, 2009. **45**(2): p. 228-47.
53. Antoniotti, C., et al., *AtezoTRIBE: a randomised phase II study of FOLFOXIRI plus bevacizumab alone or in combination with atezolizumab as initial therapy for patients with unresectable metastatic colorectal cancer.* BMC Cancer, 2020. **20**(1): p. 683.
54. Carreno, B.M., et al., *Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells.* Science, 2015. **348**(6236): p. 803-8.
55. Garcia-Garijo, A., C.A. Fajardo, and A. Gros, *Determinants for Neoantigen Identification.* Front Immunol, 2019. **10**: p. 1392.
56. Dentro, S.C., D.C. Wedge, and P. Van Loo, *Principles of Reconstructing the Subclonal Architecture of Cancers.* Cold Spring Harb Perspect Med, 2017. **7**(8).
57. Janeway CA Jr, T.P., Walport M, et al., *Immunobiology: The Immune System in Health and Disease. 5th edition. The major histocompatibility complex and its functions.* 2001.
58. Nedjic, J., et al., *Autophagy in thymic epithelium shapes the T-cell repertoire and is essential for tolerance.* Nature, 2008. **455**(7211): p. 396-400.
59. Warabi, M., M. Kitagawa, and K. Hirokawa, *Loss of MHC class II expression is associated with a decrease of tumor-infiltrating T cells and an increase of metastatic potential of colorectal cancer: immunohistological and histopathological analyses as compared with normal colonic mucosa and adenomas.* Pathol Res Pract, 2000. **196**(12): p. 807-15.

60. Meazza, R., et al., *Tumor rejection by gene transfer of the MHC class II transactivator in murine mammary adenocarcinoma cells.* Eur J Immunol, 2003. **33**(5): p. 1183-92.
61. Mortara, L., et al., *CIITA-induced MHC class II expression in mammary adenocarcinoma leads to a Th1 polarization of the tumor microenvironment, tumor rejection, and specific antitumor memory.* Clin Cancer Res, 2006. **12**(11 Pt 1): p. 3435-43.
62. Sconocchia, G., et al., *HLA class II antigen expression in colorectal carcinoma tumors as a favorable prognostic marker.* Neoplasia, 2014. **16**(1): p. 31-42.
63. Kreiter, S., et al., *Mutant MHC class II epitopes drive therapeutic immune responses to cancer.* Nature, 2015. **520**(7549): p. 692-6.
64. Alspach, E., et al., *MHC-II neoantigens shape tumour immunity and response to immunotherapy.* Nature, 2019. **574**(7780): p. 696-701.
65. Johnson, D.B., et al., *Melanoma-specific MHC-II expression represents a tumour-autonomous phenotype and predicts response to anti-PD-1/PD-L1 therapy.* Nat Commun, 2016. **7**: p. 10582.
66. Abelin, J.G., et al., *Defining HLA-II Ligand Processing and Binding Rules with Mass Spectrometry Enhances Cancer Epitope Prediction.* Immunity, 2019. **51**(4): p. 766-779.e17.
67. McGranahan, N., et al., *Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution.* Cell, 2017. **171**(6): p. 1259-1271.e11.
68. Jensen, K.K., et al., *Improved methods for predicting peptide binding affinity to MHC class II molecules.* Immunology, 2018.
69. Castle, J.C., et al., *Exploiting the mutanome for tumor vaccination.* Cancer Res, 2012. **72**(5): p. 1081-91.
70. Kim, H., et al., *Clinical and pathological characteristics of sporadic colorectal carcinomas with DNA replication errors in microsatellite sequences.* Am J Pathol, 1994. **145**(1): p. 148-56.
71. Boland, C.R. and A. Goel, *Microsatellite instability in colorectal cancer.* Gastroenterology, 2010. **138**(6): p. 2073-2087 e3.
72. Gryfe, R. and S. Gallinger, *Microsatellite instability, mismatch repair deficiency, and colorectal cancer.* Surgery, 2001. **130**(1): p. 17-20.
73. Lengauer, C., K.W. Kinzler, and B. Vogelstein, *Genetic instability in colorectal cancers.* Nature, 1997. **386**(6625): p. 623-7.
74. Cohen, R., et al., *Clinical and molecular characterisation of hereditary and sporadic metastatic colorectal cancers harbouring microsatellite instability/DNA mismatch repair deficiency.* Eur J Cancer, 2017. **86**: p. 266-274.
75. Carethers, J.M., *Differentiating Lynch-like from Lynch syndrome.* Gastroenterology, 2014. **146**(3): p. 602-4.
76. Sehgal, R., et al., *Lynch syndrome: an updated review.* Genes (Basel), 2014. **5**(3): p. 497-507.
77. Kane, M.F., et al., *Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines.* Cancer Res, 1997. **57**(5): p. 808-11.
78. Scaltriti, M. and J. Baselga, *The epidermal growth factor receptor pathway: a model for targeted therapy.* Clin Cancer Res, 2006. **12**(18): p. 5268-72.
79. Westdorp, H., et al., *Opportunities for immunotherapy in microsatellite instable colorectal cancer.* Cancer Immunol Immunother, 2016. **65**(10): p. 1249-59.
80. Rodriguez-Soler, M., et al., *Risk of cancer in cases of suspected lynch syndrome without germline mutation.* Gastroenterology, 2013. **144**(5): p. 926-932 e1; quiz e13-4.

81. Palles, C., et al., *Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas.* Nat Genet, 2013. **45**(2): p. 136-44.
82. Jansen, A.M., et al., *Combined mismatch repair and POLE/POLD1 defects explain unresolved suspected Lynch syndrome cancers.* Eur J Hum Genet, 2016. **24**(7): p. 1089-92.
83. Yao, J., et al., *Comprehensive analysis of POLE and POLD1 Gene Variations identifies cancer patients potentially benefit from immunotherapy in Chinese population.* Sci Rep, 2019. **9**(1): p. 15767.
84. Wang, F., et al., *Evaluation of POLE and POLD1 Mutations as Biomarkers for Immunotherapy Outcomes Across Multiple Cancer Types.* JAMA Oncol, 2019.
85. Popat, S., R. Hubner, and R.S. Houlston, *Systematic review of microsatellite instability and colorectal cancer prognosis.* J Clin Oncol, 2005. **23**(3): p. 609-18.
86. Nordholm-Carstensen, A., et al., *Mismatch repair status and synchronous metastases in colorectal cancer: A nationwide cohort study.* Int J Cancer, 2015. **137**(9): p. 2139-48.
87. Lochhead, P., et al., *Microsatellite instability and BRAF mutation testing in colorectal cancer prognostication.* J Natl Cancer Inst, 2013. **105**(15): p. 1151-6.
88. Timmermann, B., et al., *Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis.* PLoS One, 2010. **5**(12): p. e15661.
89. Galon, J., et al., *Type, density, and location of immune cells within human colorectal tumors predict clinical outcome.* Science, 2006. **313**(5795): p. 1960-4.
90. Boland, P.M. and W.W. Ma, *Immunotherapy for Colorectal Cancer.* Cancers (Basel), 2017. **9**(5).
91. Kawakami, H., A. Zaanan, and F.A. Sinicrope, *Microsatellite instability testing and its role in the management of colorectal cancer.* Curr Treat Options Oncol, 2015. **16**(7): p. 30.
92. Lal, N., et al., *An immunogenomic stratification of colorectal cancer: Implications for development of targeted immunotherapy.* Oncoimmunology, 2015. **4**(3): p. e976052.
93. Lal, N., et al., *KRAS Mutation and Consensus Molecular Subtypes 2 and 3 Are Independently Associated with Reduced Immune Infiltration and Reactivity in Colorectal Cancer.* Clin Cancer Res, 2018. **24**(1): p. 224-233.
94. Smeby, J., et al., *CMS-dependent prognostic impact of KRAS and BRAFV600E mutations in primary colorectal cancer.* Ann Oncol, 2018. **29**(5): p. 1227-1234.
95. Loi, S., et al., *RAS/MAPK Activation Is Associated with Reduced Tumor-Infiltrating Lymphocytes in Triple-Negative Breast Cancer: Therapeutic Cooperation Between MEK and PD-1/PD-L1 Immune Checkpoint Inhibitors.* Clin Cancer Res, 2016. **22**(6): p. 1499-509.
96. Nosho, K., et al., *Tumour-infiltrating T-cell subsets, molecular changes in colorectal cancer, and prognosis: cohort study and literature review.* J Pathol, 2010. **222**(4): p. 350-66.
97. Mlecnik, B., et al., *Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability.* Immunity, 2016. **44**(3): p. 698-711.
98. Guinney, J., et al., *The consensus molecular subtypes of colorectal cancer.* Nat Med, 2015. **21**(11): p. 1350-6.
99. Arrington, A.K., et al., *Prognostic and predictive roles of KRAS mutation in colorectal cancer.* Int J Mol Sci, 2012. **13**(10): p. 12153-68.

100. Sarshekeh, A.M., et al., *Consensus molecular subtype (CMS) as a novel integral biomarker in colorectal cancer: A phase II trial of bintrafusp alfa in CMS4 metastatic CRC.* Journal of Clinical Oncology, 2020. **38**(15_suppl): p. 4084-4084.
101. Burkholder, B., et al., *Tumor-induced perturbations of cytokines and immune cell networks.* Biochim Biophys Acta, 2014. **1845**(2): p. 182-201.
102. Siska, P.J. and J.C. Rathmell, *T cell metabolic fitness in antitumor immunity.* Trends Immunol, 2015. **36**(4): p. 257-64.
103. Keir, M.E., et al., *Tissue expression of PD-L1 mediates peripheral T cell tolerance.* J Exp Med, 2006. **203**(4): p. 883-95.
104. Postow, M.A., M.K. Callahan, and J.D. Wolchok, *Immune Checkpoint Blockade in Cancer Therapy.* J Clin Oncol, 2015. **33**(17): p. 1974-82.
105. Allison, K.E., B.L. Coomber, and B.W. Bridle, *Metabolic reprogramming in the tumour microenvironment: a hallmark shared by cancer cells and T lymphocytes.* Immunology, 2017. **152**(2): p. 175-184.
106. Warburg, O., F. Wind, and E. Negelein, *The Metabolism of Tumors in the Body.* J Gen Physiol, 1927. **8**(6): p. 519-30.
107. Cairns, R.A., I.S. Harris, and T.W. Mak, *Regulation of cancer cell metabolism.* Nat Rev Cancer, 2011. **11**(2): p. 85-95.
108. Jiang, Y., Y. Li, and B. Zhu, *T-cell exhaustion in the tumor microenvironment.* Cell Death Dis, 2015. **6**: p. e1792.
109. Zhang, Y. and H.C. Ertl, *Starved and Asphyxiated: How Can CD8(+) T Cells within a Tumor Microenvironment Prevent Tumor Progression.* Front Immunol, 2016. **7**: p. 32.
110. Llosa, N.J., et al., *The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints.* Cancer Discov, 2015. **5**(1): p. 43-51.
111. Parkes, E.E., et al., *Activation of STING-Dependent Innate Immune Signaling By S-Phase-Specific DNA Damage in Breast Cancer.* J Natl Cancer Inst, 2017. **109**(1).
112. Mackenzie, K.J., et al., *cGAS surveillance of micronuclei links genome instability to innate immunity.* Nature, 2017. **548**(7668): p. 461-465.
113. Harding, S.M., et al., *Mitotic progression following DNA damage enables pattern recognition within micronuclei.* Nature, 2017. **548**(7668): p. 466-470.
114. Zhang, Y., L.H. Rohde, and H. Wu, *Involvement of nucleotide excision and mismatch repair mechanisms in double strand break repair.* Curr Genomics, 2009. **10**(4): p. 250-8.
115. Galon, J., et al., *The immune score as a new possible approach for the classification of cancer.* J Transl Med, 2012. **10**: p. 1.
116. Galon, J., et al., *Towards the introduction of the 'Immunoscore' in the classification of malignant tumours.* J Pathol, 2014. **232**(2): p. 199-209.
117. Pages, F., et al., *In situ cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer.* J Clin Oncol, 2009. **27**(35): p. 5944-51.
118. Pages, F., et al., *International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study.* Lancet, 2018.
119. Galon, J., et al., *Cancer classification using the Immunoscore: a worldwide task force.* J Transl Med, 2012. **10**: p. 205.
120. Mlecnik, B., et al., *Histopathologic-based prognostic factors of colorectal cancers are associated with the state of the local immune reaction.* J Clin Oncol, 2011. **29**(6): p. 610-8.
121. Mlecnik, B., et al., *The tumor microenvironment and Immunoscore are critical determinants of dissemination to distant metastasis.* Sci Transl Med, 2016. **8**(327): p. 327ra26.

122. Alberts, S.R., et al., *Effect of oxaliplatin, fluorouracil, and leucovorin with or without cetuximab on survival among patients with resected stage III colon cancer: a randomized trial.* JAMA, 2012. **307**(13): p. 1383-93.
123. Sinicrope, F.A., et al., *Contribution of Immunoscore and Molecular Features to Survival Prediction in Stage III Colon Cancer.* JNCI Cancer Spectr, 2020. **4**(3): p. pkaa023.
124. Pages, F., et al., *Prognostic and predictive value of the Immunoscore in stage III colon cancer patients treated with oxaliplatin in the prospective IDEA France PRODIGE-GERCOR cohort study.* Ann Oncol, 2020. **31**(7): p. 921-929.
125. HalioDx. *IMMUNOSCORE®Scoring Immune Response in Colon Cancer*. 2020; Available from: www.haliodx.com.
126. Droeser, R.A., et al., *Clinical impact of programmed cell death ligand 1 expression in colorectal cancer.* Eur J Cancer, 2013. **49**(9): p. 2233-42.
127. Grosso, J., et al., *Association of tumor PD-L1 expression and immune biomarkers with clinical activity in patients (pts) with advanced solid tumors treated with nivolumab (anti-PD-1; BMS-936558; ONO-4538).* Journal of Clinical Oncology, 2013. **31**(15_suppl): p. 3016-3016.
128. Taube, J.M., et al., *Association of PD-1, PD-1 ligands, and other features of the tumor immune microenvironment with response to anti-PD-1 therapy.* Clin Cancer Res, 2014. **20**(19): p. 5064-74.
129. Valentini, A.M., et al., *PD-L1 expression in colorectal cancer defines three subsets of tumor immune microenvironments.* Oncotarget, 2018. **9**(9): p. 8584-8596.
130. Shen, Z., et al., *Clinicopathological and prognostic significance of PD-L1 expression in colorectal cancer: a systematic review and meta-analysis.* World J Surg Oncol, 2019. **17**(1): p. 4.
131. Zhou, E., et al., *Up-regulation of Tim-3 is associated with poor prognosis of patients with colon cancer.* Int J Clin Exp Pathol, 2015. **8**(7): p. 8018-27.
132. Zhang, Y., et al., *TIM-3 is a potential prognostic marker for patients with solid tumors: A systematic review and meta-analysis.* Oncotarget, 2017. **8**(19): p. 31705-31713.
133. Gilad, Y., S.A. Rifkin, and J.K. Pritchard, *Revealing the architecture of gene regulation: the promise of eQTL studies.* Trends Genet, 2008. **24**(8): p. 408-15.
134. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. **452**(7186): p. 423-8.
135. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
136. International HapMap, C., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
137. Li, M.J., et al., *GWASdb: a database for human genetic variants identified by genome-wide association studies.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1047-54.
138. Oshlack, A., M.D. Robinson, and M.D. Young, *From RNA-seq reads to differential expression results.* Genome Biol, 2010. **11**(12): p. 220.
139. Wolf, J.B., *Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial.* Mol Ecol Resour, 2013. **13**(4): p. 559-72.
140. Morley, M., et al., *Genetic analysis of genome-wide variation in human gene expression.* Nature, 2004. **430**(7001): p. 743-7.
141. Stranger, B.E., et al., *Population genomics of human gene expression.* Nat Genet, 2007. **39**(10): p. 1217-24.
142. Schadt, E.E., et al., *Mapping the genetic architecture of gene expression in human liver.* PLoS Biol, 2008. **6**(5): p. e107.

143. Wong, E.S., et al., *Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution.* Nat Commun, 2017. **8**(1): p. 1092.
144. Yao, L., B.P. Berman, and P.J. Farnham, *Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes.* Crit Rev Biochem Mol Biol, 2015. **50**(6): p. 550-73.
145. Nica, A.C. and E.T. Dermitzakis, *Expression quantitative trait loci: present and future.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1620): p. 20120362.
146. Brem, R.B. and L. Kruglyak, *The landscape of genetic complexity across 5,700 gene expression traits in yeast.* Proc Natl Acad Sci U S A, 2005. **102**(5): p. 1572-7.
147. Consortium, G.T., *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.
148. Sul, J.H., et al., *Accurate and fast multiple-testing correction in eQTL studies.* Am J Hum Genet, 2015. **96**(6): p. 857-68.
149. Conneely, K.N. and M. Boehnke, *So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests.* Am J Hum Genet, 2007. **81**(6): p. 1158-68.
150. Seaman, S.R. and B. Muller-Myhsok, *Rapid simulation of P values for product methods and multiple-testing adjustment in association studies.* Am J Hum Genet, 2005. **76**(3): p. 399-408.
151. Hochberg, Y. and Y. Benjamini, *More powerful procedures for multiple significance testing.* Stat Med, 1990. **9**(7): p. 811-8.
152. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.
153. Benjamini, Y. and D. Yekutieli, *Quantitative trait Loci analysis using the false discovery rate.* Genetics, 2005. **171**(2): p. 783-90.
154. Grundberg, E., et al., *Mapping cis- and trans-regulatory effects across multiple tissues in twins.* Nat Genet, 2012. **44**(10): p. 1084-9.
155. Nica, A.C., et al., *The architecture of gene regulatory variation across multiple human tissues: the MuTHER study.* PLoS Genet, 2011. **7**(2): p. e1002003.
156. Carithers, L.J. and H.M. Moore, *The Genotype-Tissue Expression (GTEx) Project.* Biopreserv Biobank, 2015. **13**(5): p. 307-8.
157. Bahcall, O.G., *Human genetics: GTEx pilot quantifies eQTL variation across tissues and individuals.* Nat Rev Genet, 2015. **16**(7): p. 375.
158. Li, Q., et al., *Integrative eQTL-based analyses reveal the biology of breast cancer risk loci.* Cell, 2013. **152**(3): p. 633-41.
159. Thibodeau, S.N., et al., *Identification of candidate genes for prostate cancer-risk SNPs utilizing a normal prostate tissue eQTL data set.* Nat Commun, 2015. **6**: p. 8653.
160. Lawrenson, K., et al., *Cis-eQTL analysis and functional validation of candidate susceptibility genes for high-grade serous ovarian cancer.* Nat Commun, 2015. **6**: p. 8234.
161. Vogelsang, M., et al., *The Expression Quantitative Trait Loci in Immune Pathways and their Effect on Cutaneous Melanoma Prognosis.* Clin Cancer Res, 2016. **22**(13): p. 3268-80.
162. Fagerberg, L., et al., *Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics.* Mol Cell Proteomics, 2014. **13**(2): p. 397-406.
163. Landmark-Hoyvik, H., et al., *Genome-wide association study in breast cancer survivors reveals SNPs associated with gene expression of genes belonging to MHC class I and II.* Genomics, 2013. **102**(4): p. 278-87.

164. McGranahan, N., et al., *Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade.* Science, 2016. **351**(6280): p. 1463-9.
165. Lennerz, V., et al., *The response of autologous T cells to a human melanoma is dominated by mutated neoantigens.* Proc Natl Acad Sci U S A, 2005. **102**(44): p. 16013-8.
166. Karasaki, T., et al., *Prediction and prioritization of neoantigens: integration of RNA sequencing data with whole-exome sequencing.* Cancer Sci, 2017. **108**(2): p. 170-177.
167. Vitiello, A. and M. Zanetti, *Neoantigen prediction and the need for validation.* Nat Biotechnol, 2017. **35**(9): p. 815-817.
168. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.
169. Andreatta, M. and M. Nielsen, *Gapped sequence alignment using artificial neural networks: application to the MHC class I system.* Bioinformatics, 2016. **32**(4): p. 511-7.
170. Nielsen, M., et al., *Reliable prediction of T-cell epitopes using neural networks with novel sequence representations.* Protein Sci, 2003. **12**(5): p. 1007-17.
171. Anagnostou, V., et al., *Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Non-Small Cell Lung Cancer.* Cancer Discov, 2017. **7**(3): p. 264-276.
172. Wolf, Y., et al., *UVB-Induced Tumor Heterogeneity Diminishes Immune Response in Melanoma.* Cell, 2019. **179**(1): p. 219-235 e21.
173. Richters, M.M., et al., *Best practices for bioinformatic characterization of neoantigens for clinical utility.* Genome Med, 2019. **11**(1): p. 56.
174. Schenck, R.O., et al., *NeoPredPipe: high-throughput neoantigen prediction and recognition potential pipeline.* BMC Bioinformatics, 2019. **20**(1): p. 264.
175. Shao, X.M., et al., *High-throughput prediction of MHC class I and class II neoantigens with MHCnuggets.* Cancer Immunol Res, 2019.
176. Liu, C., et al., *ATHLATES: accurate typing of human leukocyte antigen through exome sequencing.* Nucleic Acids Res, 2013. **41**(14): p. e142.
177. Ka, S., et al., *HLAscan: genotyping of the HLA region using next-generation sequencing data.* BMC Bioinformatics, 2017. **18**(1): p. 258.
178. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol, 2013. **31**(3): p. 213-9.
179. Chen, C., et al., *A fast Peptide Match service for UniProt Knowledgebase.* Bioinformatics, 2013. **29**(21): p. 2808-9.
180. Shukla, S.A., et al., *Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes.* Nat Biotechnol, 2015. **33**(11): p. 1152-8.
181. Saunders, C.T., et al., *Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs.* Bioinformatics, 2012. **28**(14): p. 1811-7.
182. Koboldt, D.C., et al., *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.* Genome Res, 2012. **22**(3): p. 568-76.
183. VAtools. *VCF Annotation Tools (VAtools)*. 2019 [cited 2019 31/12/2019]; Available from: https://vatools.readthedocs.io/en/latest/.
184. Roth, A., et al., *PyClone: statistical inference of clonal population structure in cancer.* Nat Methods, 2014. **11**(4): p. 396-8.
185. Morris, L.G., et al., *Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival.* Oncotarget, 2016. **7**(9): p. 10051-63.
186. Abécassis, J., et al., *Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data.* PLoS One, 2019. **14**(11): p. e0224143.
187. Deshwar, A.G., et al., *PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors.* Genome Biol, 2015. **16**: p. 35.

188. Mroz, E.A. and J.W. Rocco, *MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma.* Oral Oncol, 2013. **49**(3): p. 211-5.
189. Noorbakhsh, J., et al., *Distribution-based measures of tumor heterogeneity are sensitive to mutation calling and lack strong clinical predictive power.* Sci Rep, 2018. **8**(1): p. 11445.
190. England, G. *The 100,000 Genomes Project by numbers*. 2018 [cited 2018 2 August]; Available from: https://www.genomicsengland.co.uk/the-100000-genomes-project-by-numbers/.
191. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.
192. Consortium, I.T.P.-C.A.o.W.G., *Pan-cancer analysis of whole genomes.* Nature, 2020. **578**(7793): p. 82-93.
193. Kumar, S., et al., *Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences.* Cell, 2020. **180**(5): p. 915-927 e16.
194. Greenbaum, A., et al., *Tumor Heterogeneity as a Predictor of Response to Neoadjuvant Chemotherapy in Locally Advanced Rectal Cancer.* Clin Colorectal Cancer, 2019. **18**(2): p. 102-109.
195. Ng, C., et al., *Genomics and metagenomics of colorectal cancer.* J Gastrointest Oncol, 2019. **10**(6): p. 1164-1170.
196. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing.* Nature, 2010. **464**(7285): p. 59-65.
197. Manichanh, C., et al., *Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach.* Gut, 2006. **55**(2): p. 205-11.
198. Sears, C.L., A.L. Geis, and F. Housseau, *Bacteroides fragilis subverts mucosal biology: from symbiont to colon carcinogenesis.* J Clin Invest, 2014. **124**(10): p. 4166-72.
199. Wu, S., et al., *A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses.* Nat Med, 2009. **15**(9): p. 1016-22.
200. Flemer, B., et al., *The oral microbiota in colorectal cancer is distinctive and predictive.* Gut, 2018. **67**(8): p. 1454-1463.
201. Hernández-Luna, M.A., S. López-Briones, and R. Luria-Pérez, *The Four Horsemen in Colon Cancer.* J Oncol, 2019. **2019**: p. 5636272.
202. Routy, B., et al., *Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors.* Science, 2018. **359**(6371): p. 91-97.
203. Li, J., et al., *An integrated catalog of reference genes in the human gut microbiome.* Nat Biotechnol, 2014. **32**(8): p. 834-41.
204. Matson, V., et al., *The commensal microbiome is associated with anti-PD-1 efficacy in metastatic melanoma patients.* Science, 2018. **359**(6371): p. 104-+.
205. Cremonesi, E., et al., *Gut microbiota modulate T cell trafficking into human colorectal cancer.* Gut, 2018. **67**(11): p. 1984-1994.
206. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2.* Genome Biol, 2019. **20**(1): p. 257.
207. Ranjan, R., et al., *Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing.* Biochem Biophys Res Commun, 2016. **469**(4): p. 967-77.
208. Stodolna, A., et al., *Clinical-grade whole-genome sequencing and 3' transcriptome analysis of colorectal cancer patients.* Genome Med, 2021. **13**(1): p. 33.
209. BEAR. *Birmingham Environment for Academic Research*. 2020; Available from: http://www.birmingham.ac.uk/bear.

210. Anitei, M.G., et al., *Prognostic and predictive values of the immunoscore in patients with rectal cancer.* Clin Cancer Res, 2014. **20**(7): p. 1891-9.
211. Illumina. *What is the PhiX Control v3 Library and what is its function in Illumina Next Generation Sequencing?* 2020 [cited 2020 18 July]; Available from: https://emea.support.illumina.com/bulletins/2017/02/what-is-the-phix-control-v3-library-and-what-is-its-function-in-.html.
212. Illumina. *bcl2fastq*. 2020 [cited 2020 15 July]; Available from: https://emea.support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html.
213. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
214. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.
215. II, A.-C. *A phase II trial assessing nivolumab in strong class II expressing microsatellite stable colorectal cancer*. 2019; Available from: http://www.isrctn.com/ISRCTN40245896.
216. ArrayExpress. *E-TABM-1140 - Transcription profiling by array of subcutaneous fat, skin and LCLs dervied from 856 TwinsUK participants*. 2011; Available from: https://www.ebi.ac.uk/arrayexpress/experiments/E-TABM-1140/.
217. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
218. Illumina. *Isaac aligner*. 2018; Available from: https://github.com/Illumina/Isaac3.
219. Raczy, C., et al., *Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms.* Bioinformatics, 2013. **29**(16): p. 2041-3.
220. Cornish, A.J., et al., *Correcting reference bias from the Illumina Isaac aligner enables analysis of cancer genomes.* bioRxiv, 2019: p. 836171.
221. Li, H., *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.* Bioinformatics, 2011. **27**(21): p. 2987-93.
222. Atlas, T.H.P. 2020 [cited 2020 September 2015]; Available from: http://www.proteinatlas.org.
223. Bauer, D.C., et al., *Evaluation of computational programs to predict HLA genotypes from genomic sequencing data.* Brief Bioinform, 2018. **19**(2): p. 179-187.
224. Pestinger, V., et al., *Use of an Integrated Pan-Cancer Oncology Enrichment Next-Generation Sequencing Assay to Measure Tumour Mutational Burden and Detect Clinically Actionable Variants.* Mol Diagn Ther, 2020. **24**(3): p. 339-349.
225. Van Loo, P., et al., *Allele-specific copy number analysis of tumors.* Proc Natl Acad Sci U S A, 2010. **107**(39): p. 16910-5.
226. Bolli, N., et al., *Heterogeneity of genomic evolution and mutational profiles in multiple myeloma.* Nat Commun, 2014. **5**: p. 2997.
227. Breitwieser, F.P. and S.L. Salzberg, *Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification.* Bioinformatics, 2020. **36**(4): p. 1303-1304.
228. Wood, D.E. and S.L. Salzberg, *Kraken: ultrafast metagenomic sequence classification using exact alignments.* Genome Biol, 2014. **15**(3): p. R46.
229. Broad Institute, G.R. *Picard Toolkit*. 2019 [cited 2020 September 1]; Available from: http://broadinstitute.github.io/picard/.
230. G*Power. *G * Power 3.1 manual*. 2017 [cited 2020 August 4]; Available from: https://www.psychologie.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf.

231. UK, C.R. *Bowel cancer incidence statistics*. 2020 [cited 2020 26/03/2020]; Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence#heading-One.
232. Soriano, L.C., M. Soriano-Gabarro, and L.A. Garcia Rodriguez, *Trends in the contemporary incidence of colorectal cancer and patient characteristics in the United Kingdom: a population-based cohort study using The Health Improvement Network.* BMC Cancer, 2018. **18**(1): p. 402.
233. Network, N.C.I. *Survival by Stage*. 2017 [cited 2020 26/03/2020]; Available from: http://www.ncin.org.uk/publications/survival_by_stage.
234. Network, N.C.I. *Cancer Reports: Cancer and equality groups:key metrics 2015*. 2015 [cited 2020 26/03/2020]; Available from: http://www.ncin.org.uk/publications/reports/.
235. Government, U. *Regional ethnic diversity*. 2018 11 July 2019 [cited 2020 14 July]; Available from: https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/regional-ethnic-diversity/latest.
236. Quirke, P. and E. Morris, *Reporting colorectal cancer.* Histopathology, 2007. **50**(1): p. 103-12.
237. Parnaby, C.N., et al., *Prognostic value of lymph node ratio and extramural vascular invasion on survival for patients undergoing curative colon cancer resection.* Br J Cancer, 2015. **113**(2): p. 212-9.
238. Dawson, H., et al., *Optimizing the detection of venous invasion in colorectal cancer: the ontario, Canada, experience and beyond.* Front Oncol, 2014. **4**: p. 354.
239. Messenger, D.E., D.K. Driman, and R. Kirsch, *Developments in the assessment of venous invasion in colorectal cancer: implications for future practice and patient outcome.* Hum Pathol, 2012. **43**(7): p. 965-73.
240. Blenkinsopp, W.K., et al., *Histopathology reporting in large bowel cancer.* J Clin Pathol, 1981. **34**(5): p. 509-13.
241. Stranger, B.E., et al., *Patterns of cis regulatory variation in diverse human populations.* PLoS Genet, 2012. **8**(4): p. e1002639.
242. Weng, C., et al., *A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records.* Appl Clin Inform, 2014. **5**(2): p. 463-79.
243. Clegg, A., et al., *Improving recruitment of older people to clinical trials: use of the cohort multiple randomised controlled trial design.* Age Ageing, 2015. **44**(4): p. 547-50.
244. Holdcroft, A., *Gender bias in research: how does it affect evidence based medicine?* J R Soc Med, 2007. **100**(1): p. 2-3.
245. Niranjan, S.J., et al., *Bias and stereotyping among research and clinical professionals: Perspectives on minority recruitment for oncology clinical trials.* Cancer, 2020. **126**(9): p. 1958-1968.
246. Ballouz, S., A. Dobin, and J.A. Gillis, *Is it time to change the reference genome?* Genome Biol, 2019. **20**(1): p. 159.
247. Günther, T. and C. Nettelblad, *The presence and impact of reference bias on population genomic studies of prehistoric human populations.* PLoS Genet, 2019. **15**(7): p. e1008302.
248. Li, Q., et al., *Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types.* Hum Mol Genet, 2014. **23**(19): p. 5294-302.
249. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.
250. Cadigan, K.M. and M.L. Waterman, *TCF/LEFs and Wnt signaling in the nucleus.* Cold Spring Harb Perspect Biol, 2012. **4**(11).

251. Xu, X., et al., *Clinical Significance of Transcription Factor 7 (TCF7) as a Prognostic Factor in Gastric Cancer.* Med Sci Monit, 2019. **25**: p. 3957-3963.
252. Tang, W., et al., *A genome-wide RNAi screen for Wnt/beta-catenin pathway components identifies unexpected roles for TCF transcription factors in cancer.* Proc Natl Acad Sci U S A, 2008. **105**(28): p. 9697-702.
253. Kurtulus, S., et al., *Checkpoint Blockade Immunotherapy Induces Dynamic Changes in PD-1(-)CD8(+) Tumor-Infiltrating T Cells.* Immunity, 2019. **50**(1): p. 181-194 e6.
254. Schrock, A.B., et al., *Tumor mutational burden is predictive of response to immune checkpoint inhibitors in MSI-high metastatic colorectal cancer.* Ann Oncol, 2019. **30**(7): p. 1096-1103.
255. Yamanashi, T., et al., *Podoplanin expression identified in stromal fibroblasts as a favorable prognostic marker in patients with colorectal carcinoma.* Oncology, 2009. **77**(1): p. 53-62.
256. Grasso, C.S., et al., *Genetic Mechanisms of Immune Evasion in Colorectal Cancer.* Cancer Discov, 2018. **8**(6): p. 730-749.
257. Mroz, E.A., et al., *High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma.* Cancer, 2013. **119**(16): p. 3034-42.
258. Spranger, S., R. Bao, and T.F. Gajewski, *Melanoma-intrinsic beta-catenin signalling prevents anti-tumour immunity.* Nature, 2015. **523**(7559): p. 231-5.
259. Lund, A.W., et al., *Lymphatic vessels regulate immune microenvironments in human and murine melanoma.* J Clin Invest, 2016. **126**(9): p. 3389-402.
260. Fankhauser, M., et al., *Tumor lymphangiogenesis promotes T cell infiltration and potentiates immunotherapy in melanoma.* Sci Transl Med, 2017. **9**(407).
261. Reynisson, B., et al., *Improved Prediction of MHC II Antigen Presentation through Integration and Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data.* J Proteome Res, 2020. **19**(6): p. 2304-2315.
262. Kostic, A.D., et al., *Genomic analysis identifies association of Fusobacterium with colorectal carcinoma.* Genome Res, 2012. **22**(2): p. 292-8.
263. Thomas, A.M., et al., *Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation.* Nat Med, 2019. **25**(4): p. 667-678.
264. Wirbel, J., et al., *Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer.* Nat Med, 2019. **25**(4): p. 679-689.
265. Perez-Cobas, A.E., L. Gomez-Valero, and C. Buchrieser, *Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses.* Microb Genom, 2020. **6**(8).
266. Schmidt, T.S., J.F. Matias Rodrigues, and C. von Mering, *Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale.* PLoS Comput Biol, 2014. **10**(4): p. e1003594.
267. Le, D.T., et al., *Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade.* Science, 2017. **357**(6349): p. 409-413.
268. Schatoff, E.M., B.I. Leach, and L.E. Dow, *Wnt Signaling and Colorectal Cancer.* Curr Colorectal Cancer Rep, 2017. **13**(2): p. 101-110.
269. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
270. Dijkstra, K.K., et al., *Generation of Tumor-Reactive T Cells by Co-culture of Peripheral Blood Lymphocytes and Tumor Organoids.* Cell, 2018. **174**(6): p. 1586-1598 e12.
271. Crotzer, V.L. and J.S. Blum, *Autophagy and its role in MHC-mediated antigen presentation.* J Immunol, 2009. **182**(6): p. 3335-41.

272. Axelrod, M.L., et al., *Biological Consequences of MHC-II Expression by Tumor Cells in Cancer.* Clin Cancer Res, 2019. **25**(8): p. 2392-2402.
273. Quraishi, M.N., et al., *Systematic review with meta-analysis: the efficacy of faecal microbiota transplantation for the treatment of recurrent and refractory Clostridium difficile infection.* Aliment Pharmacol Ther, 2017. **46**(5): p. 479-493.
274. Paramsothy, S., et al., *Faecal Microbiota Transplantation for Inflammatory Bowel Disease: A Systematic Review and Meta-analysis.* J Crohns Colitis, 2017. **11**(10): p. 1180-1199.

## Appendix 1. Publications and presentations arising from this thesis

### PUBLICATIONS

- **Sillo TO**, Beggs AD, Morton DG, Middleton G. *Mechanisms of immunogenicity in colorectal cancer.* Br J Surg. 2019 Sep;106(10):1283-1297. Epub 2019 Jun 19.
- Pestinger V, Smith M, **Sillo T**, Findlay JM, Laes JF, Martin G, Middleton G, Taniere P, Beggs AD. *Use of an Integrated Pan-Cancer Oncology Enrichment Next-Generation Sequencing Assay to Measure Tumour Mutational Burden and Detect Clinically Actionable Variants.* Mol Diagn Ther. 2020 Jun;24(3):339-349.

### CONTRIBUTING PUBLICATION

- Stodolna A He M, Vasipalli M, Kingsbury Z, Becq J, Stockton JD, Dilworth MP, James, JD, **Sillo T,** Blakeway D, Ward ST, Ismail T, Ross MT, Beggs AD. Clinical-grade whole genome sequencing of colorectal cancer and 3-prime transcriptome analysis demonstrate targetable alterations in the majority of patients. Genome Med. 2021 Feb 25;13(1):33.

### ORAL PRESENTATIONS

- **Sillo T**, Bermingham H, Lal N, Morton D, Middleton G, Beggs A. *The role of the heritability of immune gene expression in determining the response to immunotherapy in colorectal cancer.* Oral presentation at the Association of Coloproctology of Great Britain and Ireland Annual Congress British Journal of Surgery Prize Session (2020).
- **TO Sillo**, AD Beggs, CW Yau, G Middleton, DG Morton. *The role of immune gene expression in determining the response to immunotherapy in colorectal cancer.* Oral presentation at the Society for Academic and Research Surgery, Royal College of Surgeons of England (2019).
- **Sillo T.** *Exploring germline and somatic derminants of the immune response in colorectal cancer.* University of Birmingham Annual Clinical Academic Training Meeting (2019).

## Appendix 2. Intellectual property agreement with Genomics England


Nick Maltby
c/o QMUL
Dawson Hall
Charterhouse Square
London EC1M 6BQ

12th April 2018

Dear Toritseju Sillo

**Re: The impact of inherited differences in immune gene expression and the clonality of antigenic neoepitopes on the immune system – Determination of Intellectual Property Issues**

We refer to your application entitled "The impact of inherited differences in immune gene expression and the clonality of antigenic neoepitopes on the immune system". Having reviewed your application I have set out below our determination regarding ownership and licensing of results and intellectual property arising from the research project described in your application (the "Research Project") for review by and discussion with University of Birmingham.

**Applicable Scenario**

The application referred to above falls under Scenario 2 as described in the Genomics England Intellectual Property Policy ("IP Policy"). In line with the IP Policy, we propose the following provisions for ownership and use of any intellectual property or results that arise out of the Research Project.

**Ownership of Intellectual Property**

- The results generated through the analysis of Genomics England's 100,000 Genomes Project dataset ("100KGP Dataset") within the GeCIP will be owned by Genomics England.
- Parts of the Research Project will be carried out outside of the GeCIP. The rights to the results from research performed outside of the GeCIP will be owned by University of Birmingham and/or the other academic institutions collaborating in the Research Project as appropriate.

Licensing Terms

- Genomics England will grant University of Birmingham a perpetual, irrevocable, royalty-free, non-exclusive licence to use the results that are generated through the analysis of the 100KGP Dataset within the GeCIP.
- University of Birmingham will grant Genomics England a perpetual, irrevocable, royalty-free, non-exclusive licence to incorporate the results of the Research Project into Genomics England's platform and/or the 100KGP Dataset and to use and sublicense the use of the results in accordance with Genomics England's IP Policy.

**Royalties/Revenue Sharing**

- If University of Birmingham receives any revenue from exploitation of any products or services that result from the analysis of the 100KGP Dataset and/or the other work in the Research Project carried out outside the GeCIP, then Genomics England should receive a fair and reasonable share of that revenue having regard to the contribution that the use of the Genomics England platform and the 100KGP Dataset made to the development of such products or services. No other royalties or revenue share will be payable.

**Further Written Agreement**

It is agreed that in the event that the results and/or intellectual property arising out of the Research Project have substantial commercial potential, Genomics England, University of Birmingham and any other academic institutions involved in the Research Project will negotiate in good faith a more detailed agreement to implement the ownership and, licensing and royalty provisions outlined above.

If University of Birmingham is in agreement with the above please sign the enclosed copy of this letter where indicated and return it to me at the above address.

Yours sincerely,

NICK MALTBY
GENERAL COUNSEL

We agree to the above terms and conditions.

Signed by..... 

Dated.....30.4.18.....
**For and on behalf of University of Birmingham**

## Appendix 3. Sample access agreement with the Human Biomaterials Resource Centre

UNIVERSITY OF BIRMINGHAM

**Human Biomaterials Resource Centre**
Advanced Therapies Facility
College of Medical and Dental Sciences
University of Birmingham
Edgbaston
Birmingham B15 2TT
United Kingdom

Tel (Office): [redacted]
Tel (Mobile): [redacted]
Email: [redacted]

10 January 2018

APPLICANT: TORITSEJU SILLO
PROPOSAL TITLE: IMPACT OF INHERITED DIFFERENCES AND OF THE CLONALITY OF ANTIGENIC NEOEPITOPES ON THE IMMUNE CONTEXTURE IN COLORECTAL CANCER

This research requires access to human tissue samples. I am writing to confirm that the University of Birmingham's Human Biomaterials Resource Centre can supply these samples and any associated clinical data, and can release them to this research group without the requirement for project-specific ethical approval. Samples and data will be released in an anonymised form in accordance with our ethical and R&D approvals.

The Human Biomaterials Resource Centre is licensed by the Human Tissue Authority (Research Licence Number 12358) and has received Research Tissue Bank ethical approval from NRES Committee North West - Haydock (Ref 15/NW/0079).

Please do not hesitate to contact me if you require further details.

Yours faithfully

[redacted]

pp Dr Jane C Steele
Director, Advanced Therapies Facility

## Appendix 4. Extended SNP list from www.muther.ac.uk

| Chromosome | Gene | PROBE | SNP | LCL p value |
|---|---|---|---|---|
| 12 | A2M | ILMN_1745607 | rs4883116 | 0.0124 |
| 20 | ADA | ILMN_1803686 | rs6031753 | 0.0023 |
| 4 | ADD1 | ILMN_1759252 | rs1203808 | 0.0038 |
| 4 | ADD1 | ILMN_2356786 | rs10454801 | 0.0037 |
| 14 | AKT1 | ILMN_1748661 | rs941475 | 0.0095 |
| 14 | AKT1 | ILMN_2388507 | rs879448 | 0.0184 |
| 14 | AKT1 | ILMN_2410909 | rs2582559 | 0.0057 |
| 1 | APOA2 | ILMN_1688543 | rs983494 | 0.0021 |
| 1 | ARF1 | ILMN_1661458 | rs4074668 | 7.68E-04 |
| 1 | ARF1 | ILMN_1802203 | rs10916180 | 0.0019 |
| 1 | ARF1 | ILMN_2330948 | rs4653503 | 0.0074 |
| 19 | AXL | ILMN_1701877 | rs268691 | 0.0016 |
| 19 | AXL | ILMN_2364521 | rs2369006 | 9.34E-04 |
| 14 | BATF | ILMN_1668822 | rs12147331 | 5.12E-04 |
| 11 | BATF2 | ILMN_1690241 | rs7115071 | 3.82E-04 |
| 1 | BATF3 | ILMN_1763207 | rs6695772 | 6.93E-10 |
| 16 | BCAR1 | ILMN_1672596 | rs7195938 | 0.0011 |
| 1 | BCL10 | ILMN_1716446 | rs11161590 | 3.67E-08 |
| 14 | BCL11B | ILMN_1665761 | rs1152788 | 0.0012 |
| 14 | BCL11B | ILMN_1667885 | rs4145039 | 0.0078 |
| 19 | BCL3 | ILMN_1710514 | rs2927488 | 2.95E-04 |
| 3 | BCL6 | ILMN_1737314 | rs3917109 | 0.001 |
| 3 | BCL6 | ILMN_1746053 | rs4686838 | 0.0025 |
| 19 | BST2 | ILMN_1723480 | rs7507441 | 0.0037 |
| 23 | BTK | ILMN_1662026 | rs7066006 | 0.0097 |
| 3 | BTLA | ILMN_1778536 | rs13079706 | 1.01E-04 |
| 3 | BTLA | ILMN_2099528 | rs1282731 | 0.0135 |
| 6 | BTN3A1 | ILMN_1802708 | rs4712990 | 2.00E-06 |
| 9 | C5 | ILMN_1746819 | rs10760142 | 4.82E-22 |
| 12 | CACNB3 | ILMN_2195482 | rs7975385 | 1.94E-16 |
| 2 | CACNB4 | ILMN_1673503 | rs10172230 | 0.0015 |
| 2 | CACNB4 | ILMN_1685164 | rs4664227 | 0.0039 |
| 2 | CACNB4 | ILMN_2257652 | rs11687663 | 0.0031 |
| 5 | CAMK4 | ILMN_1767168 | rs1551565 | 2.78E-18 |
| 5 | CAMK4 | ILMN_2166582 | rs10478077 | 0.0062 |
| 7 | CARD11 | ILMN_1721978 | rs1713920 | 0.0019 |
| 7 | CAV1 | ILMN_1687583 | rs1049337 | 6.31E-17 |
| 7 | CAV1 | ILMN_2149226 | rs6978354 | 2.55E-05 |
| 3 | CBLB | ILMN_1685580 | rs3772515 | 9.67E-08 |
| 3 | CCBP2 | ILMN_1763127 | rs2290973 | 0.0013 |
| 17 | CCL11 | ILMN_1725519 | rs16969946 | 0.0047 |
| 17 | CCL13 | ILMN_1783593 | rs9912511 | 0.0028 |
| 17 | CCL15 | ILMN_1669034 | rs11868806 | 9.50E-04 |

| | | | | |
|---|---|---|---|---|
| 17 | CCL15 | ILMN_1670658 | rs1634524 | 9.39E-04 |
| 17 | CCL15 | ILMN_1740609 | rs747979 | 0.0036 |
| 17 | CCL16 | ILMN_2045324 | rs8079109 | 0.0137 |
| 16 | CCL17 | ILMN_1710186 | rs247615 | 8.43E-04 |
| 17 | CCL18 | ILMN_1654411 | rs2239998 | 7.63E-04 |
| 17 | CCL2 | ILMN_1720048 | rs1982706 | 0.0025 |
| 2 | CCL20 | ILMN_1657234 | rs2063021 | 4.49E-06 |
| 9 | CCL21 | ILMN_1677505 | rs276671 | 0.0046 |
| 16 | CCL22 | ILMN_2160476 | rs170361 | 4.46E-08 |
| 17 | CCL23 | ILMN_1686109 | rs4796056 | 0.0113 |
| 17 | CCL23 | ILMN_1764030 | rs17571920 | 0.0072 |
| 7 | CCL24 | ILMN_1653766 | rs2302009 | 0.0099 |
| 19 | CCL25 | ILMN_1737817 | rs1104768 | 3.51E-08 |
| 19 | CCL25 | ILMN_1782596 | rs1104768 | 5.52E-05 |
| 7 | CCL26 | ILMN_1659601 | rs2705777 | 0.0832 |
| 5 | CCL28 | ILMN_1701347 | rs10473354 | 0.0125 |
| 5 | CCL28 | ILMN_1774087 | rs6860696 | 0.0125 |
| 17 | CCL5 | ILMN_1773352 | rs4796105 | 2.44E-11 |
| 17 | CCL5 | ILMN_2098126 | rs2291299 | 1.35E-11 |
| 17 | CCL7 | ILMN_1683456 | rs159279 | 0.0033 |
| 17 | CCL8 | ILMN_1772964 | rs1471616 | 0.0045 |
| 3 | CCR1 | ILMN_1678833 | rs11919943 | 9.39E-29 |
| 17 | CCR7 | ILMN_1715131 | rs4890093 | 1.13E-06 |
| 1 | CD247 | ILMN_1676924 | rs3108156 | 1.44E-05 |
| 12 | CD27 | ILMN_1688959 | rs4469949 | 8.48E-09 |
| 9 | CD274 | ILMN_1701914 | rs4740830 | 0.0044 |
| 7 | CD36 | ILMN_1665132 | rs17154948 | 0.0138 |
| 7 | CD36 | ILMN_1784863 | rs10486816 | 0.0029 |
| 11 | CD3D | ILMN_2261416 | rs7103514 | 0.001 |
| 11 | CD3D | ILMN_2325837 | rs12419365 | 0.0036 |
| 11 | CD3E | ILMN_1739794 | rs551662 | 0.0055 |
| 12 | CD4 | ILMN_1727284 | rs11064391 | 2.84E-06 |
| 20 | CD40 | ILMN_1779257 | rs2050111 | 0.0072 |
| 20 | CD40 | ILMN_2367818 | rs11569345 | 3.21E-38 |
| 11 | CD5 | ILMN_1753112 | rs4963452 | 0.0063 |
| 19 | CD70 | ILMN_1760247 | rs3763046 | 0.0143 |
| 5 | CD74 | ILMN_1736567 | rs2042249 | 0.006 |
| 5 | CD74 | ILMN_1761464 | rs375396 | 0.0056 |
| 5 | CD74 | ILMN_2379644 | rs4705094 | 0.0048 |
| 3 | CD80 | ILMN_1716736 | rs624035 | 1.34E-04 |
| 3 | CD86 | ILMN_1672097 | rs13058991 | 0.0056 |
| 3 | CD86 | ILMN_1714602 | rs11714406 | 1.89E-06 |
| 3 | CD86 | ILMN_1782560 | rs2681411 | 3.31E-04 |
| 2 | CD8A | ILMN_1760374 | rs10167259 | 0.0081 |
| 2 | CD8A | ILMN_1768482 | rs10167259 | 0.038 |
| 2 | CD8A | ILMN_2353732 | rs1518987 | 0.051 |

| | | | | |
|---|---|---|---|---|
| 2 | CD8B | ILMN_1669005 | rs17509413 | 0.0117 |
| 1 | CDC42 | ILMN_1675156 | rs2794278 | 0.0024 |
| 1 | CDC42 | ILMN_1696041 | rs876685 | 3.20E-04 |
| 1 | CDC42 | ILMN_2408139 | rs10917281 | 0.0069 |
| 10 | CHUK | ILMN_1677041 | rs11190688 | 0.0037 |
| 16 | CKLF | ILMN_1712389 | rs13331952 | 3.83E-11 |
| 16 | CKLF | ILMN_2298051 | rs2344574 | 0.0105 |
| 16 | CKLF | ILMN_2414027 | rs13331952 | 7.64E-12 |
| 5 | CSF2 | ILMN_1661861 | rs2158939 | 0.0052 |
| 15 | CSK | ILMN_1754121 | rs1378940 | 1.27E-13 |
| 2 | CTLA4 | ILMN_1763487 | rs1861764 | 0.0018 |
| 2 | CTLA4 | ILMN_2261627 | rs10932017 | 0.0114 |
| 16 | CX3CL1 | ILMN_1654072 | rs11866053 | 0.0055 |
| 4 | CXCL1 | ILMN_1787897 | rs1381016 | 0.0037 |
| 4 | CXCL10 | ILMN_1791759 | rs17001247 | 3.78E-35 |
| 4 | CXCL11 | ILMN_2067895 | rs4859956 | 0.0078 |
| 10 | CXCL12 | ILMN_1689111 | rs12772980 | 0.0034 |
| 10 | CXCL12 | ILMN_1791447 | rs7080655 | 0.0034 |
| 10 | CXCL12 | ILMN_1803825 | rs870957 | 0.0026 |
| 4 | CXCL13 | ILMN_1718552 | rs4859688 | 0.0062 |
| 5 | CXCL14 | ILMN_1748323 | rs2652085 | 0.001 |
| 17 | CXCL16 | ILMN_1672278 | rs1805429 | 1.96E-04 |
| 17 | CXCL16 | ILMN_1728478 | rs3744700 | 4.06E-06 |
| 4 | CXCL2 | ILMN_1682636 | rs7679277 | 0.0024 |
| 4 | CXCL3 | ILMN_1709350 | rs2091588 | 0.0021 |
| 4 | CXCL5 | ILMN_1752562 | rs872914 | 0.0012 |
| 4 | CXCL5 | ILMN_2171384 | rs12644965 | 0.0024 |
| 4 | CXCL6 | ILMN_1779234 | rs7658970 | 0.0307 |
| 4 | CXCL6 | ILMN_2161577 | rs2126207 | 0.0271 |
| 4 | CXCL9 | ILMN_1745356 | rs884304 | 1.87E-06 |
| 2 | CXCR1 | ILMN_1662524 | rs13424201 | 0.0084 |
| 23 | CXCR3 | ILMN_1797975 | rs4986622 | 0.0182 |
| 15 | CYP11A1 | ILMN_1768820 | rs4077582 | 2.35E-07 |
| 1 | DARC | ILMN_1723684 | rs11265248 | 2.89E-04 |
| 14 | DHRS2 | ILMN_1725726 | rs7157021 | 0.0029 |
| 14 | DHRS2 | ILMN_2384857 | rs222717 | 0.0215 |
| 5 | DOCK2 | ILMN_1799725 | rs889009 | 1.40E-04 |
| 2 | DPP4 | ILMN_1692535 | rs1861978 | 4.68E-05 |
| 17 | DUSP3 | ILMN_1797522 | rs1662744 | 0.0079 |
| 13 | ELF1 | ILMN_1664010 | rs2039281 | 0.0017 |
| 1 | ENAH | ILMN_1716552 | rs3219110 | 0.0021 |
| 1 | ENAH | ILMN_1727036 | rs10915993 | 0.015 |
| 1 | ENAH | ILMN_2370296 | rs7524430 | 0.002 |
| 23 | FLNA | ILMN_1687335 | rs11156600 | 0.0037 |
| 13 | FLT3 | ILMN_1766363 | rs1231051 | 0.0148 |
| 23 | FOXP3 | ILMN_1768049 | rs12559480 | 0.0073 |

| | | | | |
|---|---|---|---|---|
| 5 | FYB | ILMN_1796537 | rs665241 | 5.89E-10 |
| 5 | FYB | ILMN_2280548 | rs3849776 | 2.05E-05 |
| 6 | FYN | ILMN_1686555 | rs9487724 | 6.01E-06 |
| 6 | FYN | ILMN_1781207 | rs2182644 | 8.15E-05 |
| 6 | FYN | ILMN_2249920 | rs9487724 | 3.61E-05 |
| 6 | FYN | ILMN_2380801 | rs1409837 | 4.41E-07 |
| 13 | GAS6 | ILMN_1779558 | rs7338868 | 2.78E-04 |
| 13 | GAS6 | ILMN_1781614 | rs7997328 | 0.0154 |
| 13 | GAS6 | ILMN_1784749 | rs7338868 | 0.0025 |
| 23 | GATA1 | ILMN_1797251 | rs2977591 | 8.68E-04 |
| 10 | GATA3 | ILMN_2406656 | rs1149901 | 1.17E-15 |
| 7 | GIMAP5 | ILMN_1769383 | rs3807383 | 2.38E-14 |
| 22 | GRAP2 | ILMN_1778143 | rs3788560 | 0.0024 |
| 17 | GRB2 | ILMN_1742521 | rs4542691 | 8.22E-05 |
| 17 | GRB2 | ILMN_1748797 | rs939540 | 0.005 |
| 6 | HLA-DPA1 | ILMN_1772218 | rs2395309 | 1.70E-07 |
| 6 | HLA-DPB1 | ILMN_1749070 | rs7772134 | 3.97E-66 |
| 6 | HLA-DQB2 | ILMN_1741648 | rs9276024 | 4.03E-04 |
| 6 | HLA-DRA | ILMN_1689655 | rs13208583 | 2.77E-04 |
| 6 | HLA-G | ILMN_1656670 | rs2517681 | 7.35E-11 |
| 1 | HLX | ILMN_1686862 | rs17491176 | 0.0017 |
| 1 | HLX | ILMN_2087646 | rs796486 | 0.0012 |
| 22 | HMOX1 | ILMN_1800512 | rs929026 | 4.37E-04 |
| 2 | HSPD1 | ILMN_1774410 | rs1440086 | 0.0189 |
| 2 | HSPD1 | ILMN_1784367 | rs4349341 | 0.0046 |
| 19 | ICAM1 | ILMN_1812226 | rs7256672 | 0.008 |
| 2 | ICOS | ILMN_1669927 | rs10197319 | 8.35E-07 |
| 21 | ICOSLG | ILMN_1675671 | rs3737435 | 2.71E-05 |
| 9 | IFNA2 | ILMN_1698186 | rs10964734 | 1.91E-04 |
| 9 | IFNB1 | ILMN_1682245 | rs1379217 | 2.69E-04 |
| 12 | IFNG | ILMN_2207291 | rs2069727 | 0.0012 |
| 8 | IKBKB | ILMN_1727142 | rs9694574 | 3.49E-04 |
| 8 | IKBKB | ILMN_2172588 | rs2304297 | 0.002 |
| 1 | IL10 | ILMN_1674167 | rs6673928 | 1.88E-04 |
| 19 | IL11 | ILMN_1788107 | rs3745913 | 0.0019 |
| 5 | IL12B | ILMN_1681132 | rs17056092 | 0.0013 |
| 19 | IL12RB1 | ILMN_1699908 | rs2305740 | 1.27E-06 |
| 19 | IL12RB1 | ILMN_1815890 | rs436857 | 1.67E-05 |
| 4 | IL15 | ILMN_1724181 | rs1519550 | 3.53E-06 |
| 4 | IL15 | ILMN_1785312 | rs17464397 | 0.0019 |
| 4 | IL15 | ILMN_2273053 | rs17343501 | 0.0018 |
| 4 | IL15 | ILMN_2369221 | rs2874763 | 0.0016 |
| 15 | IL16 | ILMN_1813572 | rs859 | 1.09E-21 |
| 15 | IL16 | ILMN_2290628 | rs4577037 | 2.14E-58 |
| 6 | IL17A | ILMN_1774983 | rs667173 | 0.0011 |
| 5 | IL17B | ILMN_1766707 | rs10434720 | 8.70E-04 |

| | | | | |
|---|---|---|---|---|
| 16 | IL17C | ILMN_1788109 | rs9929191 | 0.0011 |
| 13 | IL17D | ILMN_1753823 | rs9509780 | 0.01 |
| 1 | IL19 | ILMN_1682592 | rs6673928 | 0.0018 |
| 1 | IL19 | ILMN_1799575 | rs6673928 | 5.66E-23 |
| 2 | IL1A | ILMN_1658483 | rs4402765 | 3.31E-12 |
| 2 | IL1B | ILMN_1775501 | rs4848306 | 4.68E-19 |
| 12 | IL22 | ILMN_1735208 | rs11177548 | 8.44E-04 |
| 12 | IL23A | ILMN_1715603 | rs6581061 | 0.001 |
| 1 | IL23R | ILMN_1734937 | rs2295359 | 2.26E-10 |
| 1 | IL24 | ILMN_1774685 | rs12119983 | 0.005 |
| 1 | IL24 | ILMN_2407799 | rs6701713 | 0.0109 |
| 14 | IL25 | ILMN_1720243 | rs445754 | 0.0018 |
| 14 | IL25 | ILMN_2401883 | rs7151065 | 0.0018 |
| 12 | IL26 | ILMN_2123182 | rs6581826 | 4.38E-05 |
| 16 | IL27 | ILMN_1753758 | rs11646047 | 0.0026 |
| 5 | IL3 | ILMN_1766320 | rs803054 | 0.0028 |
| 12 | IL31 | ILMN_2201866 | rs7964127 | 0.0131 |
| 16 | IL32 | ILMN_1778010 | rs10431961 | 5.08E-05 |
| 16 | IL32 | ILMN_2368530 | rs10431961 | 9.38E-05 |
| 9 | IL33 | ILMN_1809099 | rs10975728 | 0.0023 |
| 9 | IL33 | ILMN_2052924 | rs10739077 | 0.0012 |
| 16 | IL34 | ILMN_1713686 | rs7196917 | 0.0053 |
| 5 | IL4 | ILMN_1669174 | rs7702076 | 0.0018 |
| 5 | IL4 | ILMN_2389080 | rs17517511 | 6.59E-04 |
| 16 | IL4R | ILMN_1652185 | rs205413 | 8.37E-04 |
| 16 | IL4R | ILMN_1691881 | rs2520120 | 0.0059 |
| 5 | IL5 | ILMN_1709300 | rs4705959 | 6.52E-04 |
| 5 | IL5 | ILMN_2207190 | rs4705943 | 0.0144 |
| 7 | IL6 | ILMN_1699651 | rs4279506 | 8.27E-04 |
| 8 | IL7 | ILMN_2059744 | rs16907025 | 0.0019 |
| 4 | IL8 | ILMN_1666733 | rs11728915 | 0.007 |
| 4 | IL8 | ILMN_2184373 | rs2457996 | 0.0181 |
| 5 | IL9 | ILMN_1653704 | rs17716310 | 0.0089 |
| 2 | INPP5D | ILMN_1744212 | rs4511711 | 3.08E-04 |
| 12 | IRAK3 | ILMN_1661695 | rs2701652 | 2.96E-19 |
| 5 | IRF1 | ILMN_1708375 | rs154735 | 0.0021 |
| 16 | ITGAL | ILMN_1749591 | rs7196129 | 6.43E-04 |
| 5 | ITK | ILMN_1699160 | rs152112 | 4.26E-08 |
| 19 | JAK3 | ILMN_1739667 | rs2305767 | 0.0019 |
| 19 | KCNN4 | ILMN_1709937 | rs346064 | 0.0011 |
| 4 | KIT | ILMN_1790160 | rs12647373 | 5.76E-08 |
| 12 | KLRK1 | ILMN_2222443 | rs478829 | 0.004 |
| 12 | LAG3 | ILMN_1813338 | rs10431363 | 0.0033 |
| 16 | LAT | ILMN_1691539 | rs1364184 | 0.0075 |
| 16 | LAT | ILMN_1750188 | rs9923341 | 0.0292 |
| 16 | LAT | ILMN_2281320 | rs1641996 | 0.0493 |

| | | | | |
|---|---|---|---|---|
| 16 | LAT | ILMN_2404625 | rs1642026 | 0.0337 |
| 1 | LCK | ILMN_2277426 | rs10914471 | 0.0275 |
| 1 | LCK | ILMN_2279844 | rs942242 | 4.95E-04 |
| 1 | LCK | ILMN_2377109 | rs2377856 | 0.0023 |
| 5 | LCP2 | ILMN_1658962 | rs33368 | 0.0012 |
| 4 | LEF1 | ILMN_1679185 | rs7698367 | 0.0065 |
| 22 | LGALS1 | ILMN_1723978 | rs4820294 | 4.46E-08 |
| 1 | LGALS8 | ILMN_1669930 | rs602550 | 4.56E-05 |
| 1 | LGALS8 | ILMN_2353358 | rs12076055 | 0.0025 |
| 1 | LGALS8 | ILMN_2356654 | rs12076055 | 4.31E-05 |
| 13 | LIG4 | ILMN_1680714 | rs1224177 | 7.34E-04 |
| 13 | LIG4 | ILMN_1693758 | rs157014 | 0.0038 |
| 13 | LIG4 | ILMN_2373073 | rs11618532 | 0.0029 |
| 19 | LILRB1 | ILMN_1708248 | rs8101605 | 2.13E-47 |
| 19 | LILRB2 | ILMN_2312340 | rs7246537 | 3.05E-05 |
| 20 | LIME1 | ILMN_2183687 | rs6122248 | 0.0033 |
| 8 | LYN | ILMN_1781155 | rs10504214 | 7.60E-04 |
| 18 | MALT1 | ILMN_1730986 | rs6567030 | 4.23E-04 |
| 18 | MALT1 | ILMN_2387791 | rs17761871 | 0.0016 |
| 17 | MAP3K14 | ILMN_1724070 | rs4792847 | 4.45E-05 |
| 6 | MAP3K7 | ILMN_1810176 | rs711264 | 5.56E-04 |
| 10 | MAP3K8 | ILMN_1741159 | rs2247081 | 1.69E-04 |
| 22 | MAPK1 | ILMN_1706677 | rs2330029 | 9.28E-04 |
| 22 | MAPK1 | ILMN_1767320 | rs178255 | 0.0046 |
| 22 | MAPK1 | ILMN_2235283 | rs2298432 | 0.0014 |
| 9 | MAPKAP1 | ILMN_1691526 | rs12554306 | 0.0038 |
| 9 | MAPKAP1 | ILMN_2268068 | rs2416993 | 0.0094 |
| 9 | MAPKAP1 | ILMN_2360229 | rs2026133 | 0.0038 |
| 16 | MLST8 | ILMN_1789240 | rs9921791 | 2.52E-20 |
| 6 | MYB | ILMN_1711894 | rs6902048 | 0.0052 |
| 3 | MYD88 | ILMN_1738523 | rs11928949 | 3.56E-06 |
| 3 | NCK1 | ILMN_1698001 | rs3772388 | 0.0013 |
| 17 | NCOR1 | ILMN_2186369 | rs9903464 | 1.62E-04 |
| 16 | NOD2 | ILMN_1762594 | rs9938225 | 2.75E-04 |
| 8 | PAG1 | ILMN_1673640 | rs4500045 | 4.81E-13 |
| 8 | PAG1 | ILMN_1736806 | rs4500045 | 3.25E-27 |
| 8 | PAG1 | ILMN_2055156 | rs4500045 | 6.45E-21 |
| 11 | PAK1 | ILMN_1767365 | rs3758780 | 1.40E-04 |
| 3 | PAK2 | ILMN_1659878 | rs9863627 | 2.17E-13 |
| 3 | PAK2 | ILMN_1676385 | rs7623871 | 3.99E-04 |
| 12 | PAWR | ILMN_1806907 | rs17045871 | 0.0097 |
| 2 | PDK1 | ILMN_1670256 | rs7608097 | 0.0021 |
| 17 | PDK2 | ILMN_1705397 | rs2078864 | 2.44E-05 |
| 16 | PDPK1 | ILMN_1773758 | rs3810801 | 0.0171 |
| 16 | PDPK1 | ILMN_1810554 | rs37831 | 0.0031 |
| 4 | PF4V1 | ILMN_1745522 | rs4859662 | 0.0086 |

| | | | | |
|---|---|---|---|---|
| 9 | PHPT1 | ILMN_1676611 | rs3739943 | 3.75E-06 |
| 20 | PLCG1 | ILMN_1740160 | rs17179419 | 1.25E-11 |
| 20 | PLCG1 | ILMN_2382906 | rs5009525 | 0.0014 |
| 16 | PLCG2 | ILMN_1815719 | rs4888181 | 5.96E-07 |
| 23 | PLP2 | ILMN_1738767 | rs5906754 | 0.0022 |
| 10 | PRKCQ | ILMN_1733421 | rs693088 | 0.0013 |
| 1 | PRKCZ | ILMN_1662155 | rs10907174 | 3.22E-04 |
| 1 | PRKCZ | ILMN_2253286 | rs16824948 | 0.0153 |
| 1 | PRKCZ | ILMN_2386982 | rs12403214 | 0.002 |
| 19 | PRKD2 | ILMN_1753805 | rs2694542 | 0.0028 |
| 8 | PRKDC | ILMN_1769517 | rs1983000 | 0.0131 |
| 8 | PRKDC | ILMN_2253648 | rs3750259 | 0.0057 |
| 20 | PRNP | ILMN_1737988 | rs2422986 | 2.42E-04 |
| 20 | PRNP | ILMN_2360415 | rs2095639 | 0.0053 |
| 14 | PSEN1 | ILMN_1744267 | rs214267 | 7.67E-04 |
| 14 | PSEN1 | ILMN_1808548 | rs177392 | 0.0193 |
| 14 | PSEN1 | ILMN_1809193 | rs177378 | 8.99E-04 |
| 1 | PSEN2 | ILMN_1714417 | rs6692729 | 2.97E-15 |
| 1 | PSEN2 | ILMN_2404512 | rs6692729 | 5.71E-20 |
| 5 | PTGER4 | ILMN_1795930 | rs7720838 | 1.78E-09 |
| 1 | PTPN22 | ILMN_1695640 | rs4839348 | 0.0161 |
| 1 | PTPN22 | ILMN_1780108 | rs12730318 | 0.0052 |
| 12 | PTPN6 | ILMN_1664122 | rs9668139 | 0.0025 |
| 12 | PTPN6 | ILMN_1716578 | rs11064498 | 2.91E-04 |
| 12 | PTPN6 | ILMN_1738675 | rs10849479 | 0.004 |
| 1 | PTPRC | ILMN_1730842 | rs12087648 | 0.0057 |
| 1 | PTPRC | ILMN_1804279 | rs12085890 | 0.0042 |
| 11 | PTPRJ | ILMN_1731589 | rs4752829 | 0.0104 |
| 7 | RAC1 | ILMN_1652445 | rs2108783 | 0.0039 |
| 7 | RAC1 | ILMN_1761938 | rs17136059 | 0.0042 |
| 7 | RAC1 | ILMN_2359789 | rs12536544 | 8.95E-08 |
| 17 | RARA | ILMN_1659206 | rs2015561 | 0.0062 |
| 17 | RARA | ILMN_1677197 | rs4890100 | 2.84E-04 |
| 17 | RARA | ILMN_1716176 | rs907092 | 0.0108 |
| 17 | RARA | ILMN_1791902 | rs17558532 | 0.0063 |
| 11 | RELA | ILMN_1705266 | rs2285346 | 0.0068 |
| 19 | RELB | ILMN_1811258 | rs16979873 | 0.0015 |
| 5 | RICTOR | ILMN_1705828 | rs1428246 | 2.67E-04 |
| 8 | RIPK2 | ILMN_1758939 | rs10094579 | 0.0087 |
| 14 | RNF31 | ILMN_1758831 | rs1951635 | 6.19E-04 |
| 1 | RORC | ILMN_1651792 | rs3007684 | 0.0059 |
| 1 | RORC | ILMN_1734366 | rs12030667 | 2.98E-04 |
| 1 | RORC | ILMN_1771126 | rs3828054 | 0.0103 |
| 1 | RORC | ILMN_2275399 | rs771204 | 0.0177 |
| 2 | RSAD2 | ILMN_1657871 | rs6745308 | 8.92E-05 |
| 11 | SART1 | ILMN_1680145 | rs677740 | 0.0016 |

| | | | | |
|---|---|---|---|---|
| 23 | SASH3 | ILMN_1697554 | rs7882525 | 0.0318 |
| 1 | SEMA4A | ILMN_1702787 | rs12401573 | 1.97E-13 |
| 15 | SEMA7A | ILMN_1756312 | rs8041642 | 6.29E-04 |
| 17 | SKAP1 | ILMN_1751400 | rs9895554 | 9.29E-06 |
| 2 | SLC11A1 | ILMN_1735737 | rs3791978 | 0.0128 |
| 2 | SLC11A1 | ILMN_1741165 | rs7561119 | 0.0015 |
| 6 | SNX9 | ILMN_1726366 | rs12193949 | 0.0085 |
| 2 | SOCS5 | ILMN_1715584 | rs6720535 | 0.0014 |
| 2 | SOCS5 | ILMN_1785286 | rs6740102 | 0.0021 |
| 2 | SOCS5 | ILMN_2262749 | rs441327 | 0.004 |
| 2 | SOCS5 | ILMN_2350970 | rs7584870 | 2.50E-19 |
| 11 | SPI1 | ILMN_1696463 | rs2071304 | 2.24E-44 |
| 11 | SPI1 | ILMN_2392043 | rs2291119 | 6.34E-07 |
| 16 | SPN | ILMN_1658017 | rs1064524 | 0.018 |
| 16 | SPN | ILMN_1660315 | rs3764276 | 0.0135 |
| 16 | SPN | ILMN_1801040 | rs11574941 | 0.0087 |
| 20 | SRC | ILMN_1685898 | rs6097304 | 0.0039 |
| 20 | SRC | ILMN_1729987 | rs2144509 | 6.25E-04 |
| 20 | SRC | ILMN_1778253 | rs6017916 | 0.0086 |
| 8 | STAR | ILMN_1689702 | rs10958728 | 0.0076 |
| 8 | STAR | ILMN_2391176 | rs13439094 | 0.0251 |
| 2 | STAT1 | ILMN_1691364 | rs4853546 | 4.18E-04 |
| 2 | STAT1 | ILMN_1777325 | rs11695339 | 1.32E-04 |
| 12 | STAT2 | ILMN_1690921 | rs11171806 | 0.0016 |
| 17 | STAT3 | ILMN_1663618 | rs1474040 | 1.89E-04 |
| 17 | STAT3 | ILMN_2401978 | rs17500235 | 2.21E-04 |
| 17 | STAT3 | ILMN_2410986 | rs1474040 | 2.20E-04 |
| 2 | STAT4 | ILMN_1785202 | rs7574070 | 2.19E-57 |
| 12 | STAT6 | ILMN_1763198 | rs841718 | 1.36E-53 |
| 9 | STOML2 | ILMN_1663002 | rs950048 | 7.45E-04 |
| 9 | SYK | ILMN_2059549 | rs7036417 | 1.75E-14 |
| 5 | TCF7 | ILMN_1676470 | rs729800 | 2.93E-05 |
| 5 | TCF7 | ILMN_1677846 | rs651764 | 3.89E-04 |
| 5 | TCF7 | ILMN_1683986 | rs4958129 | 0.0061 |
| 5 | TCF7 | ILMN_1707005 | rs256208 | 2.02E-04 |
| 5 | TCF7 | ILMN_2367141 | rs152402 | 2.14E-04 |
| 4 | TEC | ILMN_1666969 | rs2071027 | 2.22E-04 |
| 19 | TGFB1 | ILMN_2129668 | rs2052080 | 0.0023 |
| 1 | TGFB2 | ILMN_1812526 | rs12760500 | 0.0039 |
| 3 | TGFBR2 | ILMN_1744862 | rs11129420 | 0.0039 |
| 3 | TGFBR2 | ILMN_2384241 | rs12495646 | 3.15E-06 |
| 11 | THY1 | ILMN_1779875 | rs1073636 | 6.87E-05 |
| 19 | TICAM1 | ILMN_1724863 | rs10422141 | 1.24E-08 |
| 19 | TICAM1 | ILMN_1815079 | rs11668223 | 2.10E-04 |
| 4 | TLR2 | ILMN_1772387 | rs6536024 | 5.14E-04 |
| 4 | TLR3 | ILMN_1689578 | rs13141650 | 1.09E-04 |

| | | | | |
|---|---|---|---|---|
| 4 | TLR3 | ILMN_2155708 | rs4862522 | 0.0033 |
| 9 | TLR4 | ILMN_1706217 | rs1927914 | 2.82E-05 |
| 7 | TMEM176A | ILMN_1791511 | rs10231216 | 4.59E-04 |
| 6 | TNF | ILMN_1728106 | rs3749946 | 0.0077 |
| 22 | TNFRSF13C | ILMN_1731742 | rs2269661 | 1.89E-05 |
| 1 | TNFRSF14 | ILMN_1697409 | rs2147905 | 0.0041 |
| 13 | TNFSF13B | ILMN_1758418 | rs9559216 | 0.0015 |
| 13 | TNFSF13B | ILMN_2066858 | rs1575476 | 0.0025 |
| 1 | TNFSF4 | ILMN_1746175 | rs6686744 | 1.56E-04 |
| 1 | TNFSF4 | ILMN_2089875 | rs2205960 | 3.75E-04 |
| 9 | TRAF2 | ILMN_1691487 | rs2784092 | 8.02E-04 |
| 11 | TRAF6 | ILMN_1700353 | rs262410 | 0.0058 |
| 11 | TRAF6 | ILMN_1783910 | rs596684 | 0.0048 |
| 11 | TRAF6 | ILMN_2392143 | rs996977 | 8.73E-05 |
| 3 | TRAT1 | ILMN_1684943 | rs9879707 | 0.0047 |
| 3 | TRAT1 | ILMN_2124833 | rs7640727 | 0.0401 |
| 7 | TRIL | ILMN_1778755 | rs6943689 | 0.0054 |
| 19 | TRPM4 | ILMN_1679401 | rs7250766 | 9.25E-04 |
| 4 | TXK | ILMN_1741143 | rs3805184 | 2.77E-11 |
| 19 | UBA52 | ILMN_2368576 | rs4808137 | 5.69E-16 |
| 21 | UBASH3A | ILMN_1684450 | rs915837 | 1.03E-04 |
| 21 | UBASH3A | ILMN_2338348 | rs11203203 | 3.16E-18 |
| 17 | UBB | ILMN_1762436 | rs11650283 | 0.0034 |
| 12 | UBC | ILMN_2252160 | rs10082832 | 0.0107 |
| 6 | UBD | ILMN_1678841 | rs2021723 | 2.80E-04 |
| 12 | UBE2N | ILMN_1793651 | rs832517 | 9.48E-04 |
| 20 | UBE2V1 | ILMN_1665862 | rs932905 | 1.08E-05 |
| 19 | VASP | ILMN_1743646 | rs10995 | 7.73E-55 |
| 19 | VAV1 | ILMN_1717334 | rs1422403 | 4.74E-05 |
| 3 | WNT5A | ILMN_1800317 | rs9849795 | 0.0046 |
| 2 | ZAP70 | ILMN_1674838 | rs12477450 | 0.0057 |
| 2 | ZAP70 | ILMN_1719756 | rs2276645 | 6.38E-09 |
| 19 | ZFP36 | ILMN_1720829 | rs12986299 | 0.0034 |
| 16 | ZFPM1 | ILMN_1651438 | rs10163412 | 2.75E-08 |

**Appendix 5. Associations between eQTL SNPs and clinico-pathological markers**

| RefSNP ID | Age | Sex | Primary tumour location | T stage | EMVI | Mismatch repair | Disease stage | Ethnicity |
|---|---|---|---|---|---|---|---|---|
| rs256208 | 0.154 | 0.450 | 0.432 | 0.759 | 0.150 | 0.580 | 0.290 | <0.001* |
| rs2505777 | 0.345 | 0.532 | 0.884 | 0.029* | 0.540 | 0.171 | 0.418 | 0.089 |
| rs1152788 | 0.933 | 0.120 | 0.797 | 0.899 | 0.033* | 0.313 | 0.288 | 0.508 |
| rs2295359 | 0.360 | 0.972 | 0.450 | 0.156 | 0.392 | 0.131 | 0.117 | 0.572 |
| rs11161590 | 0.144 | 0.652 | 0.257 | 0.224 | 0.368 | 0.326 | 0.605 | 0.114 |
| rs6673928 | 0.396 | 0.525 | 0.965 | 0.840 | 0.337 | 0.754 | 0.735 | 0.204 |
| rs10760142 | 0.527 | 0.944 | 0.436 | 0.936 | 0.058 | 0.067 | 0.564 | 0.639 |
| rs11919943 | 0.335 | 0.306 | 0.375 | 0.452 | 0.768 | 0.899 | 0.226 | 0.622 |
| rs11203203 | 0.171 | 0.369 | 0.058 | 0.727 | 0.981 | 0.405 | 0.484 | 0.941 |

Comparison of clinic-pathological data points with SNP frequencies using ordinal logistic regression. Sex = M vs F, Ethnicity = Black vs Other, Side = Right vs Left, EMVI = Yes vs No. EMVI = extramural venous invasion. P values, * denotes significant difference.

## Appendix 6. Comparison of clinico-pathological markers before and after neoantigen analysis

100KGP = 100 000 Genomes Project participants. EMVI = extramural venous invasion.

| Criterion | Full data set (n = 238) | Immunoscore complete (n = 197) | Neoantigen data (n = 137) | p value |
|---|---|---|---|---|
| **Age (years)** | | | | |
| - **Median** | 69 | 69 | 71 | 0.815^ |
| - **Mean** | 68 | 67 | 68 | |
| **Sex (%)** | | | | |
| - **M** | 145 (60.9) | 119 (60.4) | 82 (59.9) | 0.979 º |
| **Ethnicity (n/%)** | | | | |
| - **White** | 209 (87.8) | 173 (87.8) | 119 (86.9) | 0.997 º |
| **Disease stage (n/%)** | 221 (92.9) | 183 (92.9) | 130 (94.9) | 0.711 º |
| - **I-III** | | | | |
| **Primary tumour side (n/%)** | | | | |
| - **Left** | 119 (50.0) | 96 (48.7) | 67 (48.9) | 0.962 º |
| - **Right** | 117 (49.2) | 99 (50.3) | 69 (50.4) | |
| **Primary tumour location** | 55 (23.8) | 43 (21.8) | 27 (19.7) | 0.745 º |
| - **Rectum** | | | | |
| **EMVI (n/%)** | | | | |
| - **Positive** | 114 (47.9) | 96 (48.7) | 68 (49.6) | 0.948º |
| **Nodes** | | | | |
| - **Total** | 23.2 | 23.3 | 22.0 | 0.354^ |
| - **Positive** | 1.6 | 1.7 | 1.6 | 0.877^ |
| **Microsatellite status (n/%)** | | | | |
| - **MSI-high** | 47 (19.7) | 39 (19.8) | 26 (19.0) | 0.513º |
| - **N/A** | 43 (18.1) | 28 (14.2) | 16 (11.7) | |

MSI-MSI-high = microsatellite instability high. ^ = Kruskal-Wallis test. º = Pearson $\chi$-squared test. n = number.